Министерство образования Республики Беларусь
Учреждение образования
«Белорусский государственный университет
информатики и радиоэлектроники»

Инженерно-экономический факультет

Кафедра экономики

**Ф. М. Файзрахманов**

# МАРКЕТИНГОВЫЕ ИССЛЕДОВАНИЯ. ЛАБОРАТОРНЫЙ ПРАКТИКУМ

*Рекомендовано УМО по образованию в области информатики и радиоэлектроники в качестве пособия для специальности 6-05-0611-07 «Цифровой маркетинг»*

В двух частях

Часть 2

Минск БГУИР 2026

УДК 339.138(076.5)
ББК 65.291.3я73
Ф17

Р е ц е н з е н т ы:

кафедра промышленного маркетинга и коммуникаций учреждения образования «Белорусский государственный экономический университет» (протокол № 8 от 26.02.2025);

доцент кафедры «Экономика и управление инновационными проектами в промышленности» Белорусского национального технического университета кандидат экономических наук, доцент Л. В. Гринцевич

**Файзрахманов, Ф. М.**

Ф17 Маркетинговые исследования. Лабораторный практикум : пособие : в 2 ч. Ч. 2 / Ф. М. Файзрахманов. – Минск : БГУИР, 2026. – 202 с. : ил.
ISBN 978-985-543-839-8 (ч. 2).

Представлен теоретический материал, условия и порядок выполнения с использованием программы IBM SPSS Statistics и приложения MS Excel лабораторных работ по основным темам дисциплины «Маркетинговые исследования» для студентов специальности 6-05-0611-07 «Цифровой маркетинг».

Первая часть пособия была опубликована в 2025 году.

**УДК 339.138(076.5)**
**ББК 65.291.3я73**

**ISBN 978-985-543-839-8 (ч. 2)**
**ISBN 978-985-543-837-4**

## СОДЕРЖАНИЕ

# ЛАБОРАТОРНАЯ РАБОТА № 5

## Расчет показателей описательной статистики по результатам выборочного наблюдения на рынке продукции компании

**Цель работы:** сформировать из созданной компанией основы выборочного наблюдения, состоящей из покупателей ее продукции, выборку требуемого объема и рассчитать для нее показатели описательной статистики.

**Задачи работы:** с использованием теоретических знаний, полученных при изучении тем № 6, 7, 10, 12, 13 и 14 дисциплины, а также изученных ранее курсов «Прикладной статистический анализ» и «Теория вероятностей и математическая статистика»:

– изучить порядок определения объема выборки требуемого объема с использованием метода доверительных интервалов;

– сформировать из созданной компанией основы выборочного наблюдения выборку требуемого объема;

– получить практические навыки в расчете показателей описательной статистики исследуемых в выборке признаков с помощью приложения MS Excel и программы IBM SPSS Statistics.

### 5.1 Теоретические сведения

#### 5.1.1 Основные термины

**Выборочное наблюдение** – это несплошное наблюдение, при котором статистическому обследованию (наблюдению) подвергаются единицы генеральной совокупности (основы выборочного наблюдения), отобранные случайным образом.

**Генеральная (изучаемая) совокупность** – это множество элементов или объектов (людей, организаций, продуктов и т. п.), обладающих информацией, которую желает получить исследователь и о которой нужно сделать заключение.

**Основа выборочного наблюдения (выборочная совокупность)** – это список (перечень) элементов, отображающих изучаемую генеральную совокупность и отобранных на основе установленных исследователем правил или процедур.

**Выборка** – это подмножество основы выборочного наблюдения, отобранное с целью изучения и анализа с помощью специальной процедуры, чтобы впоследствии обобщить полученные выводы на всю генеральную совокупность.

**Ошибка выборочного наблюдения** – разность между величиной параметра в генеральной совокупности и его величиной, вычисленной по результатам выборочного наблюдения.

**Единица выборки** – элемент основы выборочного наблюдения, выступающий в качестве единицы счета при различных процедурах формирования выборки.

**Параметр генеральной совокупности** – это показатель, вычисленный для всей генеральной совокупности. Является фиксированным числом, так как при его вычислении отсутствует случайность.

**Параметр выборки (выборочный параметр, статистика)** – это показатель, вычисленный на основе данных выборки. Является случайной величиной, так как в его основе лежат данные, полученные путем случайного отбора, который, в свою очередь, может рассматриваться как случайный эксперимент.

**Доверительный интервал** – это вычисленный на основе данных выборочного наблюдения интервал, который с заданной вероятностью включает интересующий исследователя параметр генеральной совокупности.

### 5.1.2 Определение объема простой случайной и систематической выборок методом доверительных интервалов

Простая случайная (собственно случайная) выборка (далее – простая случайная) заключается в отборе единиц из основы выборочного наблюдения наугад без каких-либо элементов системности. Все без исключения единицы основы выборочного наблюдения имеют абсолютно равную вероятность попадания в выборку. Технически простой случайный отбор проводят методом жеребьевки или с использованием таблицы (генератора) случайных чисел.

Простой случайный отбор может быть как повторным, так и бесповторным. При повторном отборе элемент основы выборочного наблюдения может попасть в выборку более одного раза. При бесповторном отборе любой элемент основы выборочного наблюдения может попасть в выборку только один раз. При использовании таблицы (генератора) случайных чисел бесповторность отбора достигается пропуском чисел в случае их повторения.

Систематическая (механическая) выборка (далее – систематическая) применяется в случаях, когда элементы основы выборочного наблюдения каким-то образом упорядочены (ранжированы) (например, по величине изучаемого или коррелирующего с ним признака).

В процессе проведения выборочного наблюдения имеют место два вида ошибок: регистрации и репрезентативности. Ошибки регистрации могут иметь случайный (непреднамеренный) или систематический (тенденциозный) характер и их можно избежать при правильной организации и проведении наблюдения. Ошибки репрезентативности органически присущи выборочному наблюдению и возникают по причине того, что выборочная совокупность не полностью воспроизводит генеральную. Избежать ошибок репрезентативности нельзя, но, пользуясь методами теории вероятностей, основанными на использовании предельных теорем и закона больших чисел, эти ошибки можно свести к минимальным значениям, границы которых устанавливаются с достаточно большой точностью.

Различают среднюю (стандартную) и предельную ошибки выборки.

Средняя (стандартная) ошибка выборки (далее – средняя) характеризует среднюю величину возможных расхождений выборочной и генеральной средней (доли).

При случайном отборе средняя ошибка выборочной средней определяется по формулам:

– при повторном отборе:

$$s_{\overline{x}} = \frac{s}{\sqrt{n}}, \tag{5.1}$$

где $s$ – стандартное отклонение изучаемого признака в выборке;

$n$ – объем (численность) выборки;

– при бесповторном отборе:

$$s_{\overline{x}} = \sqrt{\frac{s^2}{n}\left(1 - \frac{n}{N}\right)}, \tag{5.2}$$

где $s^2$ – дисперсия изучаемого признака в выборке;

$N$ – объем (численность) генеральной совокупности.

Если необходимо рассчитать среднюю ошибку выборочной биномиальной доли, то нужно использовать следующие формулы:

– при повторном отборе:

$$s_{\overline{x}} = \sqrt{\frac{p(1-p)}{n}}, \tag{5.3}$$

где $p$ – выборочная доля единиц выборки, обладающих изучаемым признаком;

– при бесповторном:

$$s_{\overline{x}} = \sqrt{\frac{p(1-p)}{n}\left(1 - \frac{n}{N}\right)}. \tag{5.4}$$

Предельная ошибка выборки представляет собой предельную величину разности между величиной параметра в генеральной совокупности и его величиной, вычисленной по результатам выборочного наблюдения. Она дает возможность выяснить, в каких пределах находится величина генеральной средней. Для ее расчета пользуются формулой

$$\varepsilon = \pm z s_{\overline{x}}, \tag{5.5}$$

где $z$ – коэффициент доверия, показатель, определяющий размер ошибки в зависимости от того, с какой вероятностью $P$ она находится.

Необходимый объем выборки при простой случайной и систематической выборках с помощью метода доверительных интервалов определяется в следующем порядке:

1 Задается степень точности (уровень надежности) результатов обследования респондентов (предельная ошибка ε).

2 Устанавливается необходимое значение вероятности *P* достоверности и в соответствии с ним из таблицы 5.1 выбирается коэффициент достоверности *z*.

Таблица 5.1 – Значения функции $P = \{|\overline{x} - \tilde{x}| \leq \varepsilon\} = \frac{1}{\sqrt{2\pi}} \int_{-z}^{z} e^{\frac{-z^2}{2}} dz$ для значений коэффициента доверия $0{,}00 \leq z \leq 4{,}99$

| *z* | 0,00 | 0,01 | 0,02 | 0,03 | 0,04 | 0,05 | 0,06 | 0,07 | 0,08 | 0,09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0,0 | 0,0000 | 0,0080 | 0,0160 | 0,0239 | 0,0319 | 0,0399 | 0,0478 | 0,0558 | 0,0638 | 0,0717 |
| 0,1 | 0,0797 | 0,0876 | 0,0955 | 0,1034 | 0,1113 | 0,1192 | 0,1271 | 0,1350 | 0,1428 | 0,1507 |
| 0,2 | 0,1585 | 0,1663 | 0,1741 | 0,1819 | 0,1897 | 0,1974 | 0,2051 | 0,2128 | 0,2205 | 0,2281 |
| 0,3 | 0,2358 | 0,2434 | 0,2510 | 0,2586 | 0,2661 | 0,2737 | 0,2812 | 0,2886 | 0,2960 | 0,3035 |
| 0,4 | 0,3108 | 0,3182 | 0,3255 | 0,3328 | 0,3401 | 0,3473 | 0,3545 | 0,3616 | 0,3688 | 0,3759 |
| 0,5 | 0,3829 | 0,3899 | 0,3969 | 0,4039 | 0,4108 | 0,4177 | 0,4245 | 0,4313 | 0,4381 | 0,4448 |
| 0,6 | 0,4515 | 0,4581 | 0,4647 | 0,4713 | 0,4778 | 0,4843 | 0,4907 | 0,4971 | 0,5035 | 0,5098 |
| 0,7 | 0,5161 | 0,5223 | 0,5285 | 0,5346 | 0,5407 | 0,5467 | 0,5527 | 0,5587 | 0,5646 | 0,5705 |
| 0,8 | 0,5763 | 0,5821 | 0,5878 | 0,5935 | 0,5991 | 0,6047 | 0,6102 | 0,6157 | 0,6211 | 0,6264 |
| 0,9 | 0,6319 | 0,6372 | 0,6424 | 0,6476 | 0,6528 | 0,6579 | 0,6629 | 0,6679 | 0,6729 | 0,6778 |
| 1,0 | 0,6827 | 0,6875 | 0,6923 | 0,6970 | 0,7017 | 0,7063 | 0,7109 | 0,7154 | 0,7199 | 0,7243 |
| 1,1 | 0,7287 | 0,7330 | 0,7373 | 0,7415 | 0,7457 | 0,7499 | 0,7540 | 0,7580 | 0,7620 | 0,7660 |
| 1,2 | 0,7699 | 0,7737 | 0,7775 | 0,7813 | 0,7850 | 0,7887 | 0,7923 | 0,7959 | 0,7994 | 0,8029 |
| 1,3 | 0,8064 | 0,8098 | 0,8132 | 0,8165 | 0,8198 | 0,8230 | 0,8262 | 0,8293 | 0,8324 | 0,8355 |
| 1,4 | 0,8385 | 0,8415 | 0,8444 | 0,8473 | 0,8501 | 0,8529 | 0,8557 | 0,8584 | 0,8611 | 0,8638 |
| 1,5 | 0,8664 | 0,8690 | 0,8715 | 0,8740 | 0,8764 | 0,8789 | 0,8812 | 0,8836 | 0,8859 | 0,8882 |
| 1,6 | 0,8904 | 0,8926 | 0,8948 | 0,8969 | 0,8990 | 0,9011 | 0,9031 | 0,9051 | 0,9070 | 0,9090 |
| 1,7 | 0,9109 | 0,9127 | 0,9146 | 0,9164 | 0,9181 | 0,9199 | 0,9216 | 0,9233 | 0,9249 | 0,9265 |
| 1,8 | 0,9281 | 0,9297 | 0,9312 | 0,9327 | 0,9342 | 0,9357 | 0,9371 | 0,9385 | 0,9399 | 0,9412 |
| 1,9 | 0,9426 | 0,9439 | 0,9451 | 0,9464 | 0,9476 | 0,9488 | 0,9500 | 0,9512 | 0,9523 | 0,9534 |
| 2,0 | 0,9545 | 0,9556 | 0,9566 | 0,9576 | 0,9586 | 0,9596 | 0,9606 | 0,9616 | 0,9625 | 0,9634 |
| 2,1 | 0,9643 | 0,9651 | 0,9660 | 0,9668 | 0,9676 | 0,9684 | 0,9682 | 0,9700 | 0,9707 | 0,9715 |
| 2,2 | 0,9722 | 0,9729 | 0,9736 | 0,9743 | 0,9749 | 0,9756 | 0,9762 | 0,9768 | 0,9774 | 0,9780 |
| 2,3 | 0,9786 | 0,9791 | 0,9797 | 0,9802 | 0,9807 | 0,9812 | 0,9817 | 0,9822 | 09827 | 0,9832 |
| 2,4 | 0,9836 | 0,9841 | 0,9845 | 0,9849 | 0,9853 | 0,9857 | 0,9861 | 0,9865 | 0,9869 | 0,9872 |
| 2,5 | 0,9876 | 0,9879 | 0,9883 | 0,9886 | 0,9889 | 0,9892 | 0,9895 | 0,9898 | 0,9901 | 0,9904 |
| 2,6 | 0,9907 | 0,9910 | 0,9912 | 0,9915 | 0,9917 | 0,9920 | 0,9922 | 0,9924 | 0,9926 | 0,9928 |
| 2,7 | 0,9931 | 0,9933 | 0,9935 | 0,9937 | 0,9939 | 0,9940 | 0,9942 | 0,9944 | 0,9946 | 0,9947 |
| 2,8 | 0,9949 | 0,9951 | 0,9952 | 0,9953 | 0,6655 | 0,9956 | 0,9958 | 0,9959 | 0,9960 | 0,9961 |
| 2,9 | 0,9963 | 0,9964 | 0,9965 | 0,9966 | 0,9967 | 0,9968 | 0,9969 | 0,9970 | 0,9971 | 0,9972 |
| 3,0 | 0,9973 | 0,9974 | 0,9975 | 0,9976 | 0,9976 | 0,9977 | 0,9978 | 0,9979 | 0,9979 | 0,9980 |
| 3,1 | 0,9981 | 0,9981 | 0,9982 | 0,9983 | 0,9983 | 0,9984 | 0,9984 | 0,9985 | 0,9985 | 0,9986 |
| 3,2 | 0,9986 | 0,9987 | 0,9987 | 0,9988 | 0,9988 | 0,9989 | 0,9989 | 0,9989 | 0,9990 | 0,9990 |
| 3,3 | 0,9990 | 0,9991 | 0,9991 | 0,9991 | 0,9992 | 0,9992 | 0,9992 | 0,9992 | 0,9993 | 0,9993 |
| 3,4 | 0,9993 | 0,9994 | 0,9994 | 0,9994 | 0,9994 | 0,9994 | 0,9995 | 0,9995 | 0,9995 | 0,9995 |

| $z$ | 0,00 | 0,01 | 0,02 | 0,03 | 0,04 | 0,05 | 0,06 | 0,07 | 0,08 | 0,09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 3,5 | 0,9995 | 0,9996 | 0,9996 | 0,9996 | 0,9996 | 0,9996 | 0,9996 | 0,9996 | 0,9997 | 0,9997 |
| 3,6 | 0,9997 | 0,9997 | 0,9997 | 0,9997 | 0,9997 | 0,9997 | 0,9997 | 0,9998 | 0,9998 | 0,9998 |
| 3,7 | 0,9998 | 0,9998 | 0,9998 | 0,9998 | 0,9998 | 0,9998 | 0,9998 | 0,9998 | 0,9998 | 0,9998 |
| 3,8 | 0,9999 | 0,9999 | 0,9999 | 0,9999 | 0,9999 | 0,9999 | 0,9999 | 0,9999 | 0,9999 | 0,9999 |
| 3,9 | 0,9999 | 0,9999 | 0,9999 | 0,9999 | 0,9999 | 0,9999 | 0,9999 | 0,9999 | 0,9999 | 0,9999 |
| 4,0 | 0,9999 | 0,9999 | 0,9999 | 0,9999 | 0,9999 | 0,9999 | 0,9999 | 0,9999 | 0,9999 | 0,9999 |

3 По итогам изучения вторичных источников или проведения пилотного исследования определяется значение стандартного отклонения среднего генеральной совокупности $s_{\tilde{x}}$. Если мера колеблемости признака неизвестна, то ее можно найти приближенно по величине предполагаемого размаха статистики (признака) $R$ по формуле

$$s_{\tilde{x}} = \frac{R}{6}. \tag{5.6}$$

4 Определяется объем выборки по одной из следующих формул:

– при определении среднего размера признака для простой случайной выборки с повторным отбором:

$$n = \frac{z^2 s_{\tilde{x}}^2}{\varepsilon^2}; \tag{5.7}$$

– при определении среднего размера признака для простой случайной выборки без повторного отбора:

$$n = \frac{z^2 s_{\tilde{x}}^2 N}{\varepsilon^2 N + z^2 s_{\tilde{x}}^2}; \tag{5.8}$$

– при определении среднего размера признака для систематической выборки с повторным отбором:

$$n = \frac{z^2 \overline{s}_{\tilde{x}}^2}{\varepsilon^2}; \tag{5.9}$$

– при определении среднего размера признака для систематической выборки без повторного отбора:

$$n = \frac{z^2 \overline{s}_{\tilde{x}}^2 N}{\varepsilon^2 N + z^2 \overline{s}_{\tilde{x}}^2}; \tag{5.10}$$

– при определении доли признака для простой случайной выборки с повторным отбором:

$$n = \frac{z^2 p(1-p)}{\varepsilon^2}; \tag{5.11}$$

– при определении доли признака для простой случайной выборки без повторного отбора:

$$n = \frac{z^2 p(1-p)N}{\varepsilon^2 N + z^2 p(1-p)}; \tag{5.12}$$

– при определении доли признака для систематической выборки с повторным отбором:

$$n = \frac{z^2 \overline{p(1-p)}}{\varepsilon^2}; \tag{5.13}$$

– при определении доли признака для систематической выборки без повторного отбора:

$$n = \frac{z^2 \overline{p(1-p)} N}{\varepsilon^2 N + z^2 \overline{p(1-p)}}. \tag{5.14}$$

Если рассчитанный объем выборки составит 10 % и больше от объема генеральной совокупности, то применяется окончательная коррекция объема первой. Необходимый объем выборки рассчитывается по формуле

$$n_{cor} = \frac{nN}{N + n - 1}. \tag{5.15}$$

Если изначально стандартное отклонение исследуемой совокупности $s_{\tilde{x}}$ было неизвестно и использовалось его предположительное (рассчитанное по формуле (5.6)) значение, то после получения выборки его следует повторно рассчитать по формуле

$$s_{\tilde{x}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \tilde{x})}{n}}. \tag{5.16}$$

### 5.1.3 Расчет показателей описательной статистики исследуемого в выборке признака

При статистической обработке данных, полученных в ходе маркетингового исследования, в обязательном порядке рассчитываются следующие статистики:

– средние значения и структурные характеристики вариационного ряда распределения (мода, медиана и квартили (при необходимости – децили и процентили));

– показатели вариации (размах, дисперсия, стандартное отклонение и коэффициент вариации);

– показатели, характеризующие форму распределения значений изучаемого признака (коэффициенты асимметрии и эксцесса).

Из средних величин всегда вычисляются средняя арифметическая простая (невзвешенная) или средняя арифметическая взвешенная:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \tag{5.17}$$

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i f_i}{n}, \tag{5.18}$$

где $x_i$ – $i$-е значение изучаемого признака;
$n$ – объем (размер) выборки;
$f_i$ – частота (вес) $i$-го значения изучаемого признака.

При расчете средней по интервальному вариационному ряду для выполнения необходимых вычислений от интервалов переходят к их серединам.

Мода (наиболее часто встречающееся значение признака у единиц выборки) для дискретного ряда определяется непосредственно как значение (вариант) $x$, имеющее наибольшую частоту или частость. Для интервального ряда она рассчитывается по формуле

$$Mo = x_{Mo} + h_{Mo} \frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})}, \tag{5.19}$$

где $x_{Mo}$ – начальная (нижняя, левая) граница модального интервала;
$h_{Mo}$ – величина (длина) модального интервала;
$f_{Mo}$ – частота модального интервала;
$f_{Mo-1}$ – частота интервала, предшествующего модальному;
$f_{Mo+1}$ – частота интервала, следующего за модальным.

Для нахождения медианы (значения переменной (признака) у средней единицы ранжированного ряда) сначала определяется ее порядковый номер, а затем по накопленным частотам определяется либо сама медиана (для дискретных рядов), либо медианный интервал (для интервальных рядов), в котором путем простой интерполяции рассчитывается среднее значение медианы по формуле

$$Me = x_{Me} + h_{Me} \frac{\frac{\sum_{i=1}^{n} f_i}{2} - S_{Me-1}}{f_{Me}}, \tag{5.20}$$

где $x_{Me}$ – начальная (нижняя, левая) граница медианного интервала;
$h_{Me}$ – величина медианного интервала;
$\frac{\sum_{i=1}^{n} f_i}{2}$ – порядковый номер медианы;
$S_{Me-1}$ – накопленная частота до медианного интервала;
$f_{Me}$ – частота медианного интервала.

Квартили представляют собой значения признака, делящие ранжированную совокупность на четыре равновеликие части. Различают квартиль нижний $Q_1$, отделяющий 1/4 часть совокупности с наименьшими значениями признака, и квартиль верхний $Q_3$, отсекающий ее 1/4 часть с наибольшими значениями признака. Средним квартилем $Q_2$ является медиана.

Для расчета квартилей по интервальному вариационному ряду используются формулы

$$Q_1 = x_{Q_1} + h\frac{\frac{1}{4}\sum_{i=1}^{n} f_i - S_{Q_1-1}}{f_{Q_1}}, \tag{5.21}$$

$$Q_3 = x_{Q_3} + h\frac{\frac{3}{4}\sum_{i=1}^{n} f_i - S_{Q_3-1}}{f_{Q_3}}, \tag{5.22}$$

где $x_{Q_1}$ – начальная (нижняя, левая) граница интервала, содержащего нижний квартиль (интервал определяется по накопленной частоте, первой превышающей 25 %);

$x_{Q_3}$ – начальная (нижняя, левая) граница интервала, содержащего верхний квартиль (интервал определяется по накопленной частоте, первой превышающей 75 %);

$h$ – величина интервала;

$S_{Q_1-1}$ – накопленная частота интервала, предшествующего интервалу, содержащему нижний квартиль;

$S_{Q_3-1}$ – накопленная частота интервала, предшествующего интервалу, содержащему верхний квартиль;

$f_{Q_1}$ – частота интервала, содержащего нижний квартиль;

$f_{Q_3}$ – частота интервала, содержащего верхний квартиль.

Размах выборки показывает, насколько велико различие между единицами выборки, имеющими самое маленькое и самое большое значение исследуемого признака:

$$R = x_{\max} - x_{\min}\ . \tag{5.23}$$

Дисперсия выборки представляет собой средний квадрат отклонений индивидуальных значений признака от их средней величины:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}. \tag{5.24}$$

Стандартное отклонение – это обобщающая характеристика размеров вариации признака в выборке:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}. \tag{5.25}$$

Коэффициент вариации (относительное стандартное отклонение) (далее – коэффициент вариации) – это мера относительного разброса значений случайной величины, которая показывает, какую долю среднего значения этой величины составляет ее средний разброс. Коэффициент выражается, как правило, в процентах, но иногда для его записи может использоваться десятичная дробь.

Для его расчета используется формула

$$V = \frac{s}{\overline{x}} \cdot 100\ \%\,. \tag{5.26}$$

Коэффициент асимметрии служит для характеристики асимметрии (скошенности) распределения значений исследуемого признака и определяется по формуле

$$A_S = \frac{n\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^3}{(n-1)(n-2)}\,. \tag{5.27}$$

Оценка существенности асимметрии производится на основе средней квадратичной ошибки коэффициента асимметрии, которая зависит от числа наблюдений и рассчитывается по формуле

$$S_{A_S} = \sqrt{\frac{6(n-1)}{(n+1)(n+3)}}\,. \tag{5.28}$$

В случае если $\frac{|A_S|}{S_{A_S}} > 3{,}0$, асимметрия считается существенной и распределение признака в генеральной совокупности несимметрично. В противном случае асимметрия несущественна и ее наличие может быть вызвано случайными обстоятельствами.

Коэффициент эксцесса используется для характеристики крутости (островершинности) распределения значений исследуемого признака в выборке. Он рассчитывается для симметричных (или близких к ним) распределений по формуле

$$\varepsilon_k = \frac{n(n+1)\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s}\right)^4}{(n-1)(n-2)(n-3)} - \frac{3(n-1)^2}{(n-2)(n-3)}\,. \tag{5.29}$$

Средняя квадратичная ошибка эксцесса рассчитывается по формуле

$$s_{\varepsilon_k} = \sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}}\,. \tag{5.30}$$

Если отношение $\frac{|\varepsilon_k|}{S_{\varepsilon_k}} > 3{,}0$, отклонение от нормального распределения нужно считать существенным; если $\frac{|\varepsilon_k|}{S_{\varepsilon_k}} < 3{,}0$, отклонение признается несущественным, а распределение признается приближенным к нормальному.

## 5.2 Вычисление показателей описательной статистики с использованием приложения MS Excel и программы IBM SPSS Statistics

Целевой сегмент (генеральная совокупность) покупателей продукции, предлагаемой ЧУП «Кэтнес» для рынка домашней мебели Витебской области, определен в размере примерно 12950 домашних хозяйств. В ходе выполненного ранее поискового маркетингового исследования было установлено что в этом сегменте значения среднегодовых расходов на покупку (обновление) мебели примерно равны: минимальное – 80 р., максимальное – 700 р., а среднее – 300 р.

Службой маркетинга ОАО «Крессида» совместно с сотрудниками отдела маркетинга ЧУП «Кэтнес» из генеральной совокупности для проведения запланированных описательного и причинно-следственного маркетинговых исследований была создана основа выборочного наблюдения объемом в 4320 домашних хозяйств. При ее создании были использованы данные, полученные не только в результате последних различного рода опросов, но и с помощью ИИ-ассистентов ОАО «Крессида» и ЧУП «Кэтнес», а их верификация была проведена с применением сервиса, разработанного на базе блокчейн-технологии.

Сведения об участниках основы выборочного наблюдения были собраны по следующим характеристикам (далее – переменным):

1) количество членов домашнего хозяйства;

2) возраст мужа;

3) возраст жены;

4) образование мужа;

5) образование жены;

6) семейный среднемесячный доход, р.;

7) количество автомобилей в семье, ед.;

8) наличие в квартире (доме) подключения к интернету;

9) вид телевидения, которым пользуется домохозяйство (кабельное или спутниковое);

10) вид жилья (квартира, отдельный дом);

11) площадь жилья, кв. м;

12) средний возраст мебели, лет;

13) стиль мебели («лофт», «кантри», «скандинавский», «классический», «неоклассический», «прованс»);

14) планируемая периодичность обновления мебели, лет;

15) примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели, р.

Необходимо:

– методом доверительных интервалов определить конечный и начальный объемы систематической бесповторной выборки домашних хозяйств, которые позволили бы с точностью ±5,00 р. определить величину примерных среднегодовых расходов на покупку (обновление) элементов мебели;

– из созданной основы выборочного наблюдения сформировать систематическую бесповторную выборку, статистический анализ которой позволил бы с требуемой точностью установить среднее значение планируемых среднегодовых расходов домашних хозяйств на покупку (обновление) элементов домашней мебели.

Все расчеты выполнять с точностью до сотых.

### 5.2.1 Расчет и формирование выборки необходимого объема

1 При уровне надежности 95 % рассчитать методом доверительных интервалов конечный объем систематической бесповторной выборки, который позволил бы с точностью ±5,00 р. установить среднее значение планируемых среднегодовых расходов домашних хозяйств на покупку (обновление) элементов мебели.

Для этого, используя отдельные ячейки на листе приложения MS Excel или калькулятор:

– по формуле (5.6) найти меру колеблемости признака (которая пока конкретно неизвестна):

$$s_{\tilde{x}} = \frac{700 - 80}{6} = 103{,}33;$$

– с использованием полученного значения определить левую и правую границы доверительного интервала:

$$300 - 1{,}96 \cdot 103{,}33 \leq \tilde{x} \leq 300 + 1{,}96 \cdot 103{,}33,$$

или

$$97{,}47 \leq \tilde{x} \leq 502{,}53,$$

и по формуле (5.8) конечный размер выборки:

$$n_{fin} = \frac{1{,}96^2 \cdot 103{,}33^2 \cdot 12950}{5^2 \cdot 12950 + 1{,}96^2 \cdot 103{,}33^2} = 1456 \text{ домашних хозяйств.}$$

Так как рассчитанный объем конечной выборки получился больше 10 % от объема генеральной совокупности, необходимо выполнить его коррекцию с использованием формулы (5.15):

$$n_{cor} = \frac{1456 \cdot 12950}{12950 + 1456 - 1} = 1309 \text{ домашних хозяйств.}$$

С учетом того, что в ходе ранее проведенных маркетинговых исследований анкеты примерно десятой части участников выборок заполнялись небрежно и неполно, принято решение начальный объем выборки принять в размере, превышающем на 10 % конечный:

$$n_{start} = 1309 \cdot 1{,}1 = 1440 \text{ домашних хозяйств.}$$

2 Рассчитать интервал выборки как отношение объема основы выборочного наблюдения к начальному объему выборки с округлением результата до ближайшего целого числа:

$$i = \frac{4320}{1440} = 3{,}0.$$

3 В приложении MS Excel открыть предоставленный преподавателем файл «05 Результаты выборочного наблюдения.xlsx», содержащий данные по созданной компанией основе выборочного наблюдения, и создать в нем, начиная с колонки **«R»** и заканчивая колонкой **«AG»**, новую таблицу, в которую будут внесены сведения по участникам выборки. Для этого:

– скопировать в первую и вторую строки этого интервала строки с пояснениями переменных (признаков) и шапкой создаваемой таблицы;

– предварительно отметив курсором ячейку **«A2»** и используя инструмент анализа **«Выборка»** (**«Данные»** – **«Анализ данных»** – **«Выборка»**), выделить из основы выборочного наблюдения выборку, состоящую из 1440 номеров домашних хозяйств с интервалом в 3 хозяйства (рисунок 5.1). В данном случае в исследуемую выборку будут выбраны домашние хозяйства, номера которых кратны трем;

– так как в этом поле будут находиться номера домохозяйств, поставить флажок напротив строки **«Метки в первой строке»**;

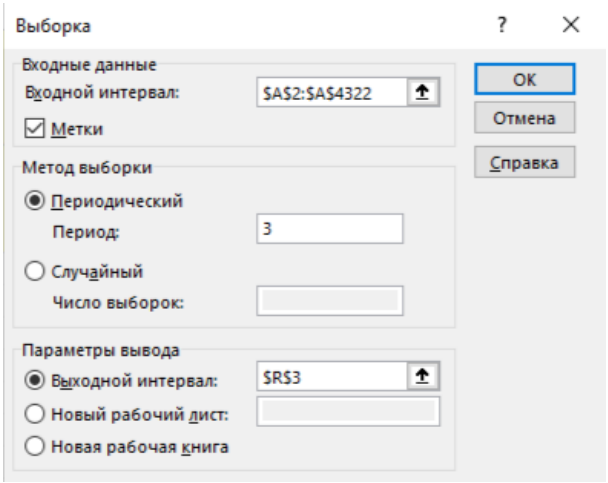


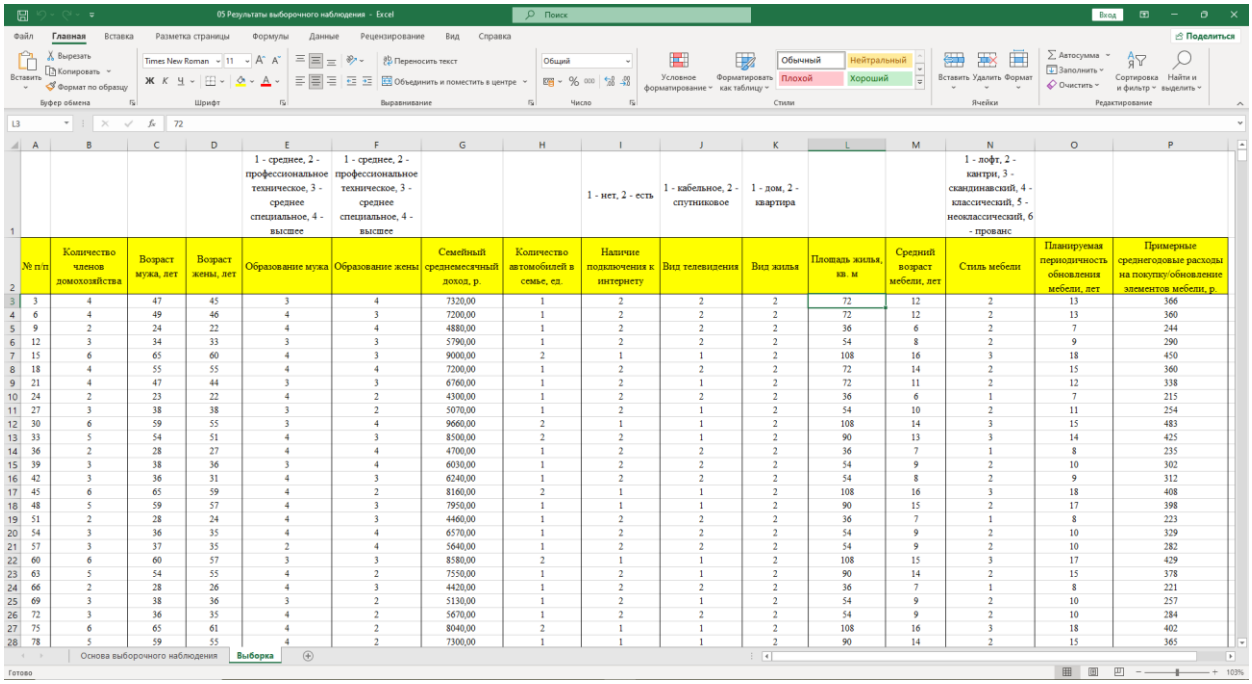

Рисунок 5.1 – Диалоговое окно инструмента анализа «Выборка» для выделения из основы выборочного наблюдения требуемого количества домашних хозяйств

– точно так же, последовательно отмечая в таблице с данными основы выборочного наблюдения ячейки с **«B2»** до **«P2»**, сделать выборки данных с шагом 3 по всем остальным показателям. Созданная таблица должна содержать данные для 1440 домашних хозяйств;

– создать в файле «05 Результаты выборочного наблюдения.xlsx» новый рабочий лист и присвоить ему название «Выборка»;

– вырезать (лучше скопировать) с листа «Основа выборочного наблюдения» созданную таблицу и перенести (вставить) ее в лист «Выборка».

Фрагмент таблицы с данными по полученной выборке представлен на рисунке 5.2.



| № п/п | Количество членов домохозяйства | Возраст мужа, лет | Возраст жены, лет | Образование мужа | Образование жены | Семейный среднемесячный доход, р. | Количество автомобилей в семье, ед. | Наличие подключения к интернету | Вид телевидения | Вид жилья | Площадь жилья, кв. м | Средний возраст мебели, лет | Стиль мебели | Планируемая периодичность обновления мебели, лет | Примерные среднегодовые расходы на покупку/обновление элементов мебели, р. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | 1 - среднее, 2 - профессиональное техническое, 3 - среднее специальное, 4 - высшее | 1 - среднее, 2 - профессиональное техническое, 3 - среднее специальное, 4 - высшее | | | 1 - нет, 2 - есть | 1 - кабельное, 2 - спутниковое | 1 - дом, 2 - квартира | | | 1 - лофт, 2 - кантри, 3 - скандинавский, 4 - классический, 5 - неоклассический, 6 - прованс | | |
| 3 | 4 | 47 | 45 | 3 | 4 | 7320,00 | 1 | 2 | 2 | 2 | 72 | 12 | 2 | 13 | 366 |
| 6 | 4 | 49 | 46 | 4 | 3 | 7200,00 | 1 | 2 | 2 | 2 | 72 | 12 | 2 | 13 | 360 |
| 9 | 2 | 24 | 22 | 4 | 4 | 4880,00 | 1 | 2 | 2 | 2 | 36 | 6 | 2 | 7 | 244 |
| 12 | 3 | 34 | 33 | 3 | 3 | 5790,00 | 1 | 2 | 2 | 2 | 54 | 8 | 2 | 9 | 290 |
| 15 | 6 | 65 | 60 | 4 | 3 | 9000,00 | 2 | 1 | 1 | 2 | 108 | 16 | 3 | 18 | 450 |
| 18 | 4 | 55 | 55 | 4 | 4 | 7200,00 | 1 | 2 | 2 | 2 | 72 | 14 | 2 | 15 | 360 |
| 21 | 4 | 47 | 44 | 3 | 3 | 6760,00 | 1 | 2 | 1 | 2 | 72 | 11 | 2 | 12 | 338 |
| 24 | 2 | 23 | 22 | 4 | 2 | 4300,00 | 1 | 2 | 2 | 2 | 36 | 6 | 1 | 7 | 215 |
| 27 | 3 | 38 | 38 | 3 | 2 | 5070,00 | 1 | 2 | 1 | 2 | 54 | 10 | 2 | 11 | 254 |
| 30 | 6 | 59 | 55 | 3 | 4 | 9660,00 | 2 | 1 | 1 | 2 | 108 | 14 | 3 | 15 | 483 |
| 33 | 5 | 54 | 51 | 4 | 3 | 8500,00 | 2 | 2 | 1 | 2 | 90 | 13 | 3 | 14 | 425 |
| 36 | 2 | 28 | 27 | 4 | 4 | 4700,00 | 1 | 2 | 2 | 2 | 36 | 7 | 1 | 8 | 235 |
| 39 | 3 | 38 | 36 | 3 | 4 | 6030,00 | 1 | 2 | 2 | 2 | 54 | 9 | 2 | 10 | 302 |
| 42 | 3 | 36 | 31 | 4 | 3 | 6240,00 | 1 | 2 | 2 | 2 | 54 | 8 | 2 | 9 | 312 |
| 45 | 6 | 65 | 59 | 4 | 2 | 8160,00 | 2 | 1 | 1 | 2 | 108 | 16 | 3 | 18 | 408 |
| 48 | 5 | 59 | 57 | 4 | 3 | 7950,00 | 1 | 1 | 1 | 2 | 90 | 15 | 2 | 17 | 398 |
| 51 | 2 | 28 | 24 | 4 | 3 | 4460,00 | 1 | 2 | 2 | 2 | 36 | 7 | 1 | 8 | 223 |
| 54 | 3 | 36 | 35 | 4 | 4 | 6570,00 | 1 | 2 | 2 | 2 | 54 | 9 | 2 | 10 | 329 |
| 57 | 3 | 37 | 35 | 2 | 4 | 5640,00 | 1 | 2 | 2 | 2 | 54 | 9 | 2 | 10 | 282 |
| 60 | 6 | 60 | 57 | 3 | 3 | 8580,00 | 2 | 1 | 1 | 2 | 108 | 15 | 3 | 17 | 429 |
| 63 | 5 | 54 | 55 | 4 | 2 | 7550,00 | 1 | 2 | 1 | 2 | 90 | 14 | 2 | 15 | 378 |
| 66 | 2 | 28 | 26 | 4 | 3 | 4420,00 | 1 | 2 | 2 | 2 | 36 | 7 | 1 | 8 | 221 |
| 69 | 3 | 38 | 36 | 3 | 2 | 5130,00 | 1 | 2 | 1 | 2 | 54 | 9 | 2 | 10 | 257 |
| 72 | 3 | 36 | 35 | 4 | 2 | 5670,00 | 1 | 2 | 2 | 2 | 54 | 9 | 2 | 10 | 284 |
| 75 | 6 | 65 | 61 | 4 | 2 | 8040,00 | 2 | 1 | 1 | 2 | 108 | 16 | 3 | 18 | 402 |
| 78 | 5 | 59 | 55 | 4 | 2 | 7300,00 | 1 | 1 | 1 | 2 | 90 | 14 | 2 | 15 | 365 |

Рисунок 5.2 – Фрагмент листа «Выборка» файла «05 Результаты выборочного наблюдения.xlsx» с данными участников созданной выборки

### 5.2.2 Расчет показателей описательной статистики с использованием приложения MS Excel

1 При уровне надежности 95 % рассчитать показатели описательной статистики для переменной «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели». Для этого (рисунок 5.3):

– в открытом файле «05 Результаты выборочного наблюдения.xlsx» выбрать лист «Выборка»;

– вызвать инструмент анализа **«Описательная статистика» («Данные» – «Анализ данных» – «Описательная статистика»)**;

– ввести в поле **«Входной интервал»** секции **«Входные данные»** значения примерных среднегодовых расходов на покупку (обновление) элементов домашней мебели (ячейки **«P2»** – **«P1442»**), группирование выбрать по столбцам. Так как в этом поле будет находиться название изучаемой переменной, поставить флажок напротив строки **«Метки в первой строке»**;

– в секции **«Параметры вывода»** выбрать **«Новый рабочий лист»** и присвоить ему название **«Среднегодовые расходы»** (можно просто пока оставить поле незаполненным, а после создания нового листа присвоить ему это название);

– поставить флажок напротив строки **«Итоговая статистика»**, уровень надежности оставить равным 95 % и нажать кнопку **«ОК»**;

– используя шрифт Times New Roman Cyr размером 12 пт, отформатировать полученные данные;

– вычисленные приложением значения представить в виде таблицы 5.2;

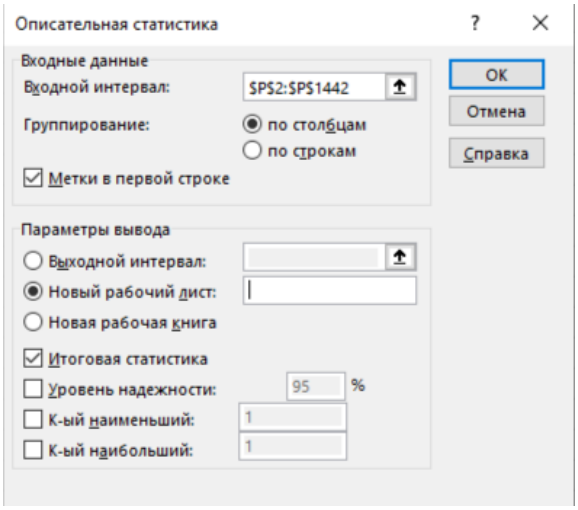



Рисунок 5.3 – Диалог инструмента анализа «Описательная статистика» с внесенными данными по выборке

Таблица 5.2 – Рассчитанные по выборке значения описательных статистик для переменной «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели»

| Статистики | Значения |
|---|---|
| Среднее | 336,62 |
| Стандартная ошибка | 3,35 |
| Медиана | 324,00 |
| Мода | 328,00 |
| Стандартное отклонение | 126,96 |
| Дисперсия выборки | 16118,69 |
| Эксцесс | –0,30 |
| Асимметричность | 0,51 |
| Интервал | 574,00 |
| Минимум | 116,00 |
| Максимум | 690,00 |
| Сумма | 484738,00 |
| Счет | 1440 |

– так как приложением средняя квадратичная ошибка асимметрии рассчитана не была, сделать это самостоятельно с использованием формулы (5.28):

$$S_{A_S} = \sqrt{\frac{6(1440-1)}{(1440+1)(1440+3)}} = 0{,}06.$$

Так как отношение показателя асимметрии к его средней квадратичной ошибке $\frac{|A_S|}{S_{A_S}} = \frac{0{,}51}{0{,}06} = 7{,}92$, что больше 3,0, то следует признать, что асимметрия переменной «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» существенна и распределение признака в выборке несимметрично;

– хотя распределение изучаемой переменной в выборке асимметрично существенно, в учебных целях по формуле (5.30) рассчитать среднюю квадратичную ошибку показателя эксцесса и оценить его существенность:

$$S_{\varepsilon_k} = \sqrt{\frac{24 \cdot 1440(1440-2)(1440-3)}{(1440+1)^2(1440+3)(1440+5)}} = 0{,}13.$$

Если бы распределение рассматриваемой переменной в выборке было симметричным, то на основе значения $\frac{|\varepsilon_k|}{S_{\varepsilon_k}} = \frac{|-0{,}30|}{0{,}13} = 2{,}31$, что меньше 3,0, его можно было бы считать приближенным к нормальному;

– коэффициент вариации рассматриваемой переменной рассчитать по формуле (5.26):

$$V_S = \frac{126{,}959}{336{,}624} \cdot 100\ \% = 37{,}70\ \%,$$

что больше 30 % и говорит о высокой колеблемости ее значений в выборке.

2 С использованием полученных значений построить гистограмму для переменной «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели». Для этого (рисунок 5.4):

– в ранее созданном листе «Среднегодовые расходы» справа от таблицы с вычисленными описательными статистиками с учетом того, что минимальное значение рассматриваемой переменной равно 116, а максимальное – 690, создать 12 интервалов группирования (карманов) с шагом в 50. Номера карманов с первого по двенадцатый разместить в ячейках **«D2»–«D13»**, а значения их правых границ – в ячейках **«E2»–«E13»**;

– перейти в лист «Выборка»;

– вызвать инструмент анализа **«Гистограмма» («Данные» – «Анализ данных» – «Гистограмма»**) и заполнить его диалоговое окно, внеся в него входной интервал данных по рассматриваемой переменной с листа «Выборка» (ячейки **«P2»–«P1442**») и интервал карманов с листа «Среднегодовые расходы» (ячейки **«E1»–«E13»**). Поставив флажок напротив строки **«Метки в первой строке»**, указав в качестве выходного интервала ячейку **«G1»** листа «Среднегодовые расходы» и поставив флажок напротив строки **«Вывод графика»**, нажать кнопку **«ОК»**;

– используя шрифт Times New Roman Cyr размером 12 пт, отформатировать появившуюся таблицу и убрать в ней строку **«Еще»**;

– используя этот же шрифт, отформатировать построенную гистограмму.

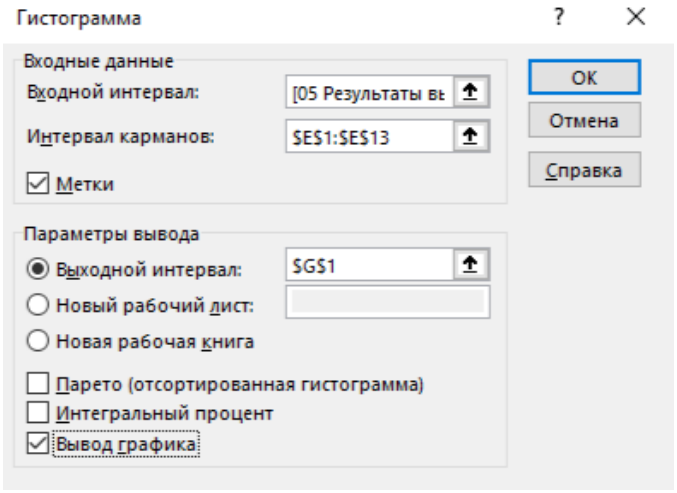



Рисунок 5.4 – Диалоговое окно инструмента анализа «Гистограмма» с внесенными данными по выборке

Итог выполненных действий представлен на рисунке 5.5.

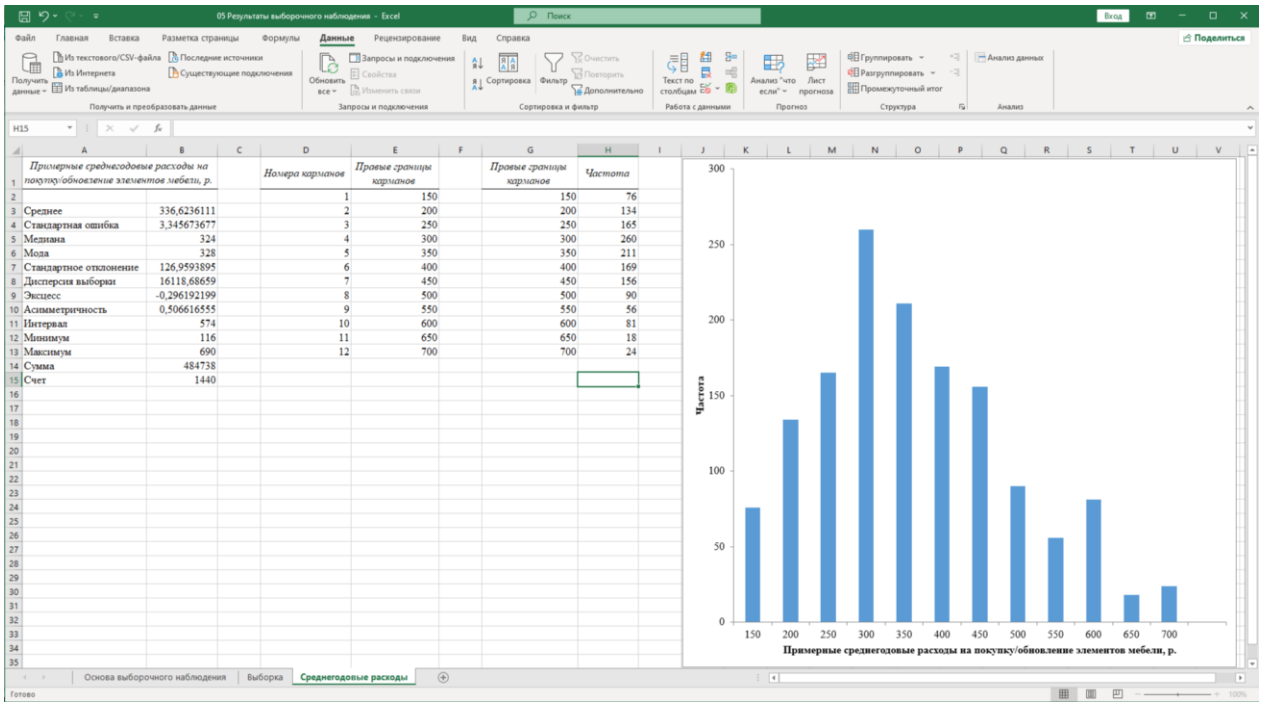



Рисунок 5.5 – Данные и выполненная с их использованием в приложении MS Excel гистограмма для переменной «Примерные среднегодовые расходы на покупку (обновление) элементов мебели»

### 5.2.3 Расчет показателей описательной статистики с использованием программы IBM SPSS Statistics

1 Открыть программу IBM SPSS Statistics. В стартовом диалоговом окне нажать кнопку **«Закрыть»**.

2 Сохранить создаваемый файл под именем «05 Результаты выборочного наблюдения.sav».

3 Внести во вкладку «Данные» редактора данных данные по сформированной выборке. Для этого скопировать (без номеров домашних хозяйств и шапки таблицы!) в ячейки вкладки «Данные» данные из ранее созданного листа «Выборка» файла «05 Результаты выборочного наблюдения.xlsx».

4 Нажав слева внизу кнопку **«Переменные»**, перейти в одноименную вкладку редактора данных и:

– присвоить имена переменным выборки;

– для всех переменных задать ширину в восемь символов без десятичных знаков после запятой, ширину колонки в десять (двенадцать) символов, выравнивание по центру и роль **«Входная»**;

– для переменных «Образование мужа» в колонке **«Значения»**, нажав список и поочередно нажимая кнопку **«Добавить»**, задать следующие значения и соответствующие им метки: «1 – среднее», «2 – профессиональное техническое», «3 – среднее специальное», «4 – высшее» (рисунок 5.6). В колонке **«Шкала»** установить **«Номинальные»**. Аналогично задать значения и для переменной «Образование жены»;

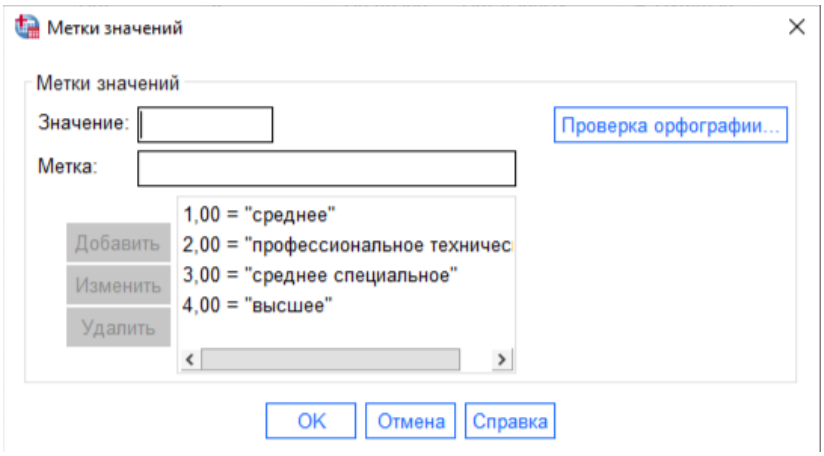
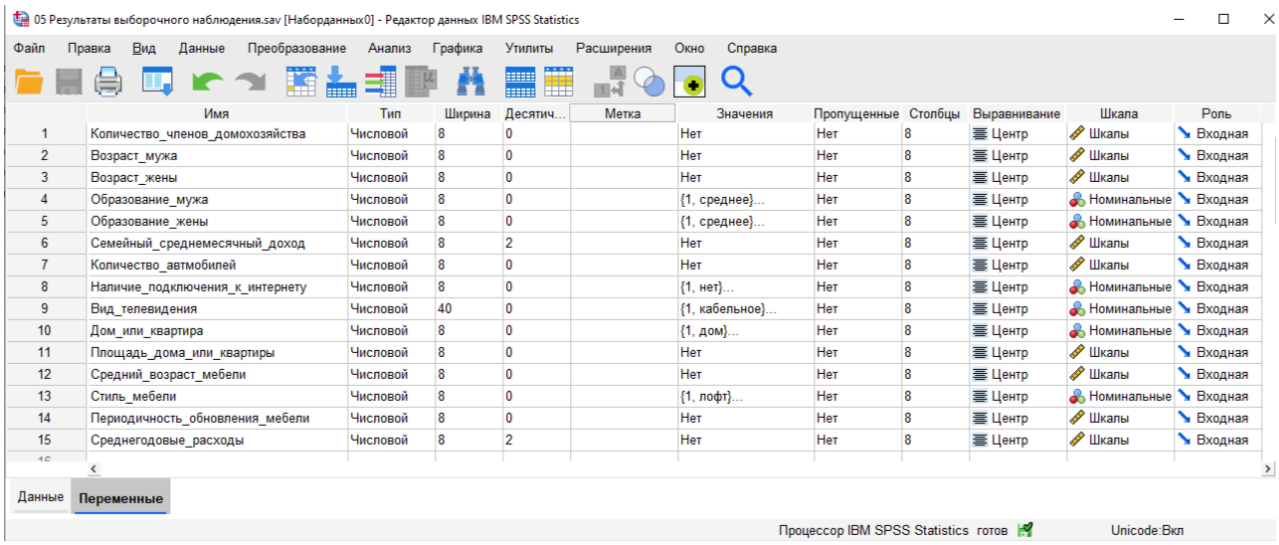



Рисунок 5.6 – Внесенные метки для колонки «Значения» в диалоговом окне «Данные» по переменной «Образование мужа»

– действуя подобным образом, задать значения и метки для переменных: «Наличие подключения к интернету» («1 – нет» и «2 – есть»), «Вид телевидения» («1 – кабельное» и «2 – спутниковое»), «Вид жилья» («1 – дом» и «2 – квартира»), «Стиль мебели» («1 – лофт», «2 – кантри», «3 – скандинавский», «4 – классический», «5 – неоклассический» и «6 – прованс»). В колонке **«Шкала»** для этих переменных установить **«Номинальные»**;

– для всех остальных переменных в колонке **«Шкала»** установить **«Шкалы»**;

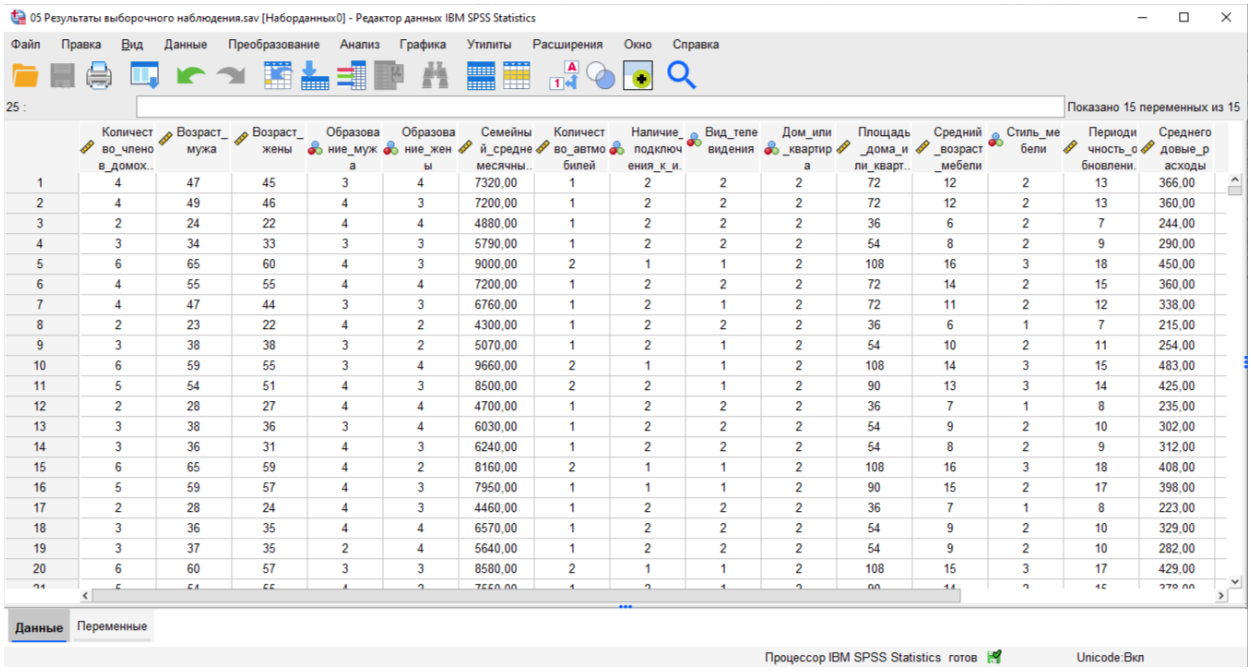
– выравнивание данных в колонках выбрать по центру (рисунок 5.7);




| | Имя | Тип | Ширина | Десятич... | Метка | Значения | Пропущенные | Столбцы | Выравнивание | Шкала | Роль |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Количество_членов_домохозяйства | Числовой | 8 | 0 | | Нет | Нет | 8 | Центр | Шкалы | Входная |
| 2 | Возраст_мужа | Числовой | 8 | 0 | | Нет | Нет | 8 | Центр | Шкалы | Входная |
| 3 | Возраст_жены | Числовой | 8 | 0 | | Нет | Нет | 8 | Центр | Шкалы | Входная |
| 4 | Образование_мужа | Числовой | 8 | 0 | | {1, среднее}... | Нет | 8 | Центр | Номинальные | Входная |
| 5 | Образование_жены | Числовой | 8 | 0 | | {1, среднее}... | Нет | 8 | Центр | Номинальные | Входная |
| 6 | Семейный_среднемесячный_доход | Числовой | 8 | 2 | | Нет | Нет | 8 | Центр | Шкалы | Входная |
| 7 | Количество_автмобилей | Числовой | 8 | 0 | | Нет | Нет | 8 | Центр | Шкалы | Входная |
| 8 | Наличие_подключения_к_интернету | Числовой | 8 | 0 | | {1, нет}... | Нет | 8 | Центр | Номинальные | Входная |
| 9 | Вид_телевидения | Числовой | 40 | 0 | | {1, кабельное}... | Нет | 8 | Центр | Номинальные | Входная |
| 10 | Дом_или_квартира | Числовой | 8 | 0 | | {1, дом}... | Нет | 8 | Центр | Номинальные | Входная |
| 11 | Площадь_дома_или_квартиры | Числовой | 8 | 0 | | Нет | Нет | 8 | Центр | Шкалы | Входная |
| 12 | Средний_возраст_мебели | Числовой | 8 | 0 | | Нет | Нет | 8 | Центр | Шкалы | Входная |
| 13 | Стиль_мебели | Числовой | 8 | 0 | | {1, лофт}... | Нет | 8 | Центр | Номинальные | Входная |
| 14 | Периодичность_обновления_мебели | Числовой | 8 | 0 | | Нет | Нет | 8 | Центр | Шкалы | Входная |
| 15 | Среднегодовые_расходы | Числовой | 8 | 2 | | Нет | Нет | 8 | Центр | Шкалы | Входная |

Рисунок 5.7 – Вкладка «Переменные» с внесенными метками и установленными шкалами для переменных, которые характеризуют участников выборки

– нажав кнопку **«Данные»**, переключиться в одноименную вкладку редактора данных (рисунок 5.8) и при необходимости изменить ширину колонок до удобной для восприятия.




| | Количество_членов_домох.. | Возраст_мужа | Возраст_жены | Образование_мужа | Образование_жены | Семейный_среднемесячны.. | Количество_автомобилей | Наличие_подключения_к_и.. | Вид_телевидения | Дом_или_квартира | Площадь_дома_или_кварт.. | Средний_возраст_мебели | Стиль_мебели | Периодичность_обновлени.. | Среднегодовые_расходы |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 4 | 47 | 45 | 3 | 4 | 7320,00 | 1 | 2 | 2 | 2 | 72 | 12 | 2 | 13 | 366,00 |
| 2 | 4 | 49 | 46 | 4 | 3 | 7200,00 | 1 | 2 | 2 | 2 | 72 | 12 | 2 | 13 | 360,00 |
| 3 | 2 | 24 | 22 | 4 | 4 | 4880,00 | 1 | 2 | 2 | 2 | 36 | 6 | 2 | 7 | 244,00 |
| 4 | 3 | 34 | 33 | 3 | 3 | 5790,00 | 1 | 2 | 2 | 2 | 54 | 8 | 2 | 9 | 290,00 |
| 5 | 6 | 65 | 60 | 4 | 3 | 9000,00 | 2 | 1 | 1 | 2 | 108 | 16 | 3 | 18 | 450,00 |
| 6 | 4 | 55 | 55 | 4 | 4 | 7200,00 | 1 | 2 | 2 | 2 | 72 | 14 | 2 | 15 | 360,00 |
| 7 | 4 | 47 | 44 | 3 | 3 | 6760,00 | 1 | 2 | 1 | 2 | 72 | 11 | 2 | 12 | 338,00 |
| 8 | 2 | 23 | 22 | 4 | 2 | 4300,00 | 1 | 2 | 2 | 2 | 36 | 6 | 1 | 7 | 215,00 |
| 9 | 3 | 38 | 38 | 3 | 2 | 5070,00 | 1 | 2 | 1 | 2 | 54 | 10 | 2 | 11 | 254,00 |
| 10 | 6 | 59 | 55 | 3 | 4 | 9660,00 | 2 | 1 | 1 | 2 | 108 | 14 | 3 | 15 | 483,00 |
| 11 | 5 | 54 | 51 | 4 | 3 | 8500,00 | 2 | 2 | 1 | 2 | 90 | 13 | 3 | 14 | 425,00 |
| 12 | 2 | 28 | 27 | 4 | 4 | 4700,00 | 1 | 2 | 2 | 2 | 36 | 7 | 1 | 8 | 235,00 |
| 13 | 3 | 38 | 36 | 3 | 4 | 6030,00 | 1 | 2 | 2 | 2 | 54 | 9 | 2 | 10 | 302,00 |
| 14 | 3 | 36 | 31 | 4 | 3 | 6240,00 | 1 | 2 | 2 | 2 | 54 | 8 | 2 | 9 | 312,00 |
| 15 | 6 | 65 | 59 | 4 | 2 | 8160,00 | 2 | 1 | 1 | 2 | 108 | 16 | 3 | 18 | 408,00 |
| 16 | 5 | 59 | 57 | 4 | 3 | 7950,00 | 1 | 1 | 1 | 2 | 90 | 15 | 2 | 17 | 398,00 |
| 17 | 2 | 28 | 24 | 4 | 3 | 4460,00 | 1 | 2 | 2 | 2 | 36 | 7 | 1 | 8 | 223,00 |
| 18 | 3 | 36 | 35 | 4 | 4 | 6570,00 | 1 | 2 | 2 | 2 | 54 | 9 | 2 | 10 | 329,00 |
| 19 | 3 | 37 | 35 | 2 | 4 | 5640,00 | 1 | 2 | 2 | 2 | 54 | 9 | 2 | 10 | 282,00 |
| 20 | 6 | 60 | 57 | 3 | 3 | 8580,00 | 2 | 1 | 1 | 2 | 108 | 15 | 3 | 17 | 429,00 |

Рисунок 5.8 – Вкладка «Данные» с частью данных по сформированной выборке

5 Рассчитать показатели описательной статистики для переменной «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели». Для этого:

– выбрать процедуру **«Частоты…» («Анализ» – «Описательные статистики» – «Частоты…»**);

– в открывшемся диалоговом окне в левой его части, где находятся переменные, которые характеризуют участников выборки, выделив «Среднегодовые расходы» и нажав стрелку, направленную вправо, перенести эту переменную в поле **«Переменные:»** (рисунок 5.9);

– нажать кнопку **«Статистики…»** и в открывшемся диалоге отметить флажками все переменные в группах **«Положение центра распределения»**, **«Распределение»** и **«Разброс»**, а в секции **«Значения процентилей»** только **«Квартили»**, после чего нажать кнопку **«Продолжить»** (рисунок 5.10);

– нажать кнопку **«Диаграммы…»** и в открывшемся диалоге выбрать график в виде гистограммы с показанной на ней кривой нормального распределения (рисунок 5.11), после чего нажать кнопку **«Продолжить»**;

– убрав флажок напротив строки **«Вывести частотные таблицы»** и нажав кнопку **«ОК»**, завершить процедуру.

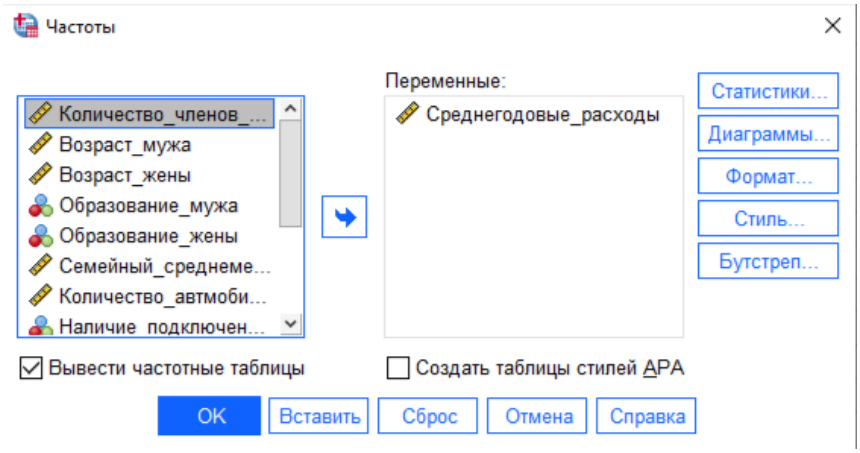



Рисунок 5.9 – Диалоговое окно процедуры «Частоты…» с выбранной для анализа переменной «Среднегодовые расходы на покупку и обновление мебели»

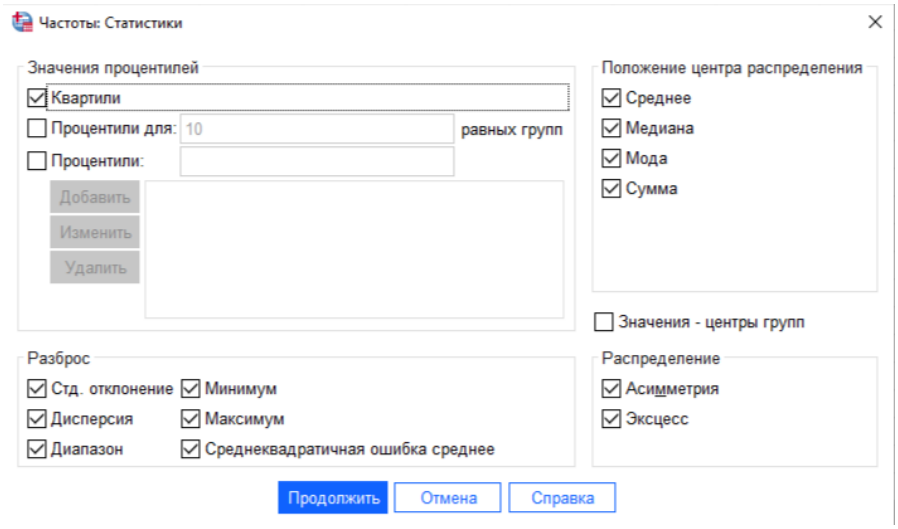



Рисунок 5.10 – Диалог «Частоты: Статистики» с выбранными для расчета статистиками по переменной «Среднегодовые расходы на покупку (обновление) элементов домашней мебели»

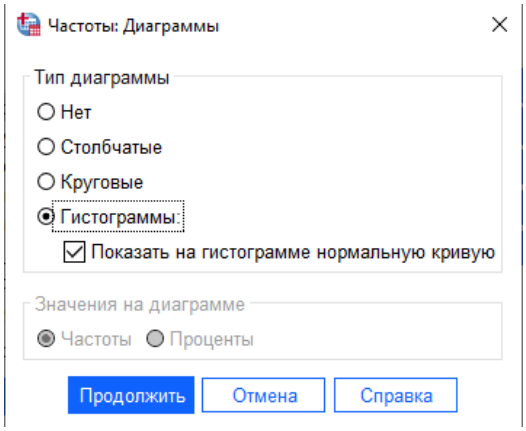



Рисунок 5.11 – Диалог «Частоты: Диаграммы» с выбранным типом диаграммы по переменной «Среднегодовые расходы на покупку (обновление) элементов домашней мебели»

Рассчитанные статистики представлены в таблице 5.3, а гистограмма – на рисунке 5.12.

Таблица 5.3 – Рассчитанные по выборке значения описательных статистик для переменной «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели»

| Статистики | | Значения |
|---|---|---|
| Количество значений | валидных | 1440 |
| | пропущенных | 0 |
| Среднее | | 336,62 |
| Стандартная ошибка среднего | | 3,35 |
| Медиана | | 324,00 |

| Статистики | | Значения |
|---|---|---|
| Мода | | 328,00 |
| Стандартное отклонение | | 126,96 |
| Дисперсия | | 16118,69 |
| Асимметрия | | 0,51 |
| Стандартная ошибка асимметрии | | 0,06 |
| Эксцесс | | –0,30 |
| Стандартная ошибка эксцесса | | 0,13 |
| Размах | | 574,00 |
| Минимум | | 116,00 |
| Максимум | | 690,00 |
| Сумма | | 484738,00 |
| Процентили | 25 | 246,00 |
| | 50 | 324,00 |
| | 75 | 413,00 |

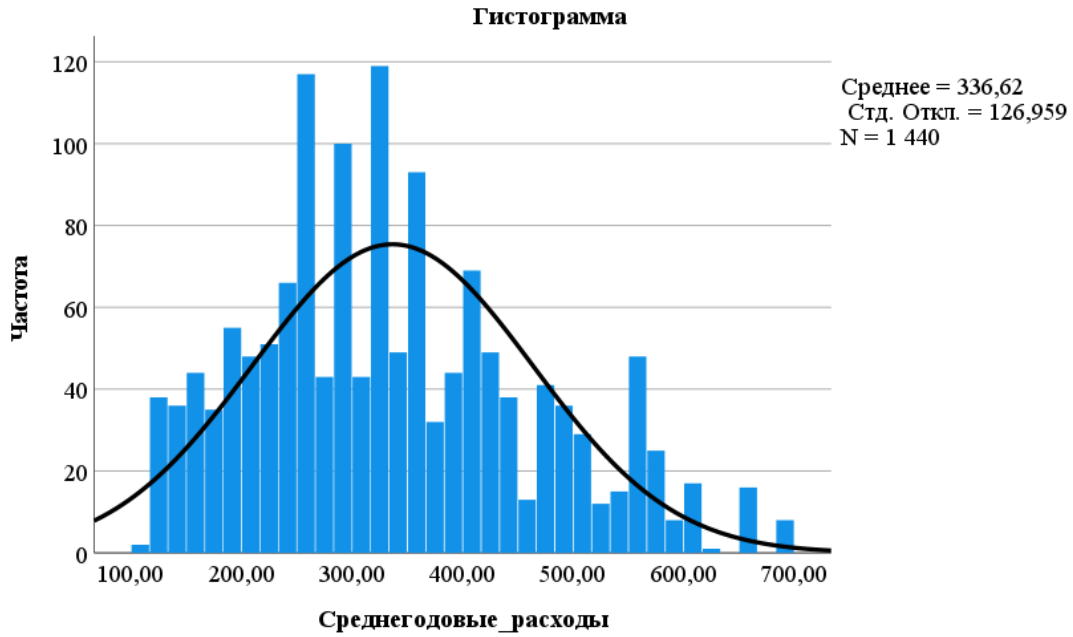



Рисунок 5.12 – Гистограмма распределения значений переменной «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели»

### 5.3 Задание для самостоятельного выполнения

Из основы выборочного наблюдения, представленной в файле «05 Результаты выборочного наблюдения.xlsx», сформировать указанный преподавателем объем простой случайной бесповторной выборки для характеристики «Планируемая периодичность обновления мебели», рассчитать для нее показатели описательной статистики и построить соответствующую гистограмму.

## 5.4 Вопросы для самоконтроля

1 Чем систематическая выборка отличается от простой случайной выборки?

2 Что является причинами возникновения ошибок регистрации в процессе выборочного наблюдения?

3 Что является причинами возникновения ошибок репрезентативности в процессе выборочного наблюдения?

4 Что представляет собой средняя ошибка выборки?

5 Что представляет собой предельная ошибка выборки?

6 Как определяется необходимый объем выборки методом доверительных интервалов?

7 Какие статистики относятся к структурным характеристикам вариационного ряда распределения?

8 Какие статистики характеризуют вариацию рассматриваемого признака в выборке?

9 Какие статистики характеризуют форму распределения значений изучаемого признака в выборке?

# ЛАБОРАТОРНАЯ РАБОТА № 6

## Дисперсионный анализ данных, полученных по выборке в ходе маркетингового эксперимента

**Цель работы:** выполнить дисперсионный анализ данных, полученных в результате оценки потребителями (частью участников выборки, сформированной в лабораторной работе № 5) предоставленных им сотрудниками маркетинговых подразделений ОАО «Крессида» и ЧУП «Кэтнес» вариантов телевизионной рекламы и материалов в местах продажи, и выбрать наиболее эффективную их комбинацию.

**Задачи работы:** с использованием теоретических знаний, полученных при изучении тем № 9–15 дисциплины, а также изученных ранее курсов «Прикладной статистический анализ» и «Теория вероятностей и математическая статистика»:

– изучить порядок проведения дисперсионного анализа данных, полученных в процессе маркетингового эксперимента;

– получить практические навыки в выполнении дисперсионного анализа данных, полученных в ходе маркетингового эксперимента, с использованием приложения MS Excel и программы IBM SPSS Statistics.

### 6.1 Теоретические сведения

#### 6.1.1 Основные термины

**Эксперимент** – это управляемый процесс изменения одной или нескольких независимых переменных (факторов) для измерения их влияния на одну или несколько зависимых переменных (признаков) при условии исключения влияния посторонних (искажающих) факторов.

**Фактор** – это категориальная переменная, соответствующая какому-то внешнему условию, влияющему на эксперимент.

**Факторный эксперимент (условия испытаний)** – это комбинация уровней факторов, действие которых значительно и поддается проверке.

**Дисперсионный анализ** – это статистический метод анализа результатов наблюдений, зависящих от различных одновременно действующих факторов, с последующим выбором наиболее важных из них и оценкой степени их влияния.

**Однофакторный дисперсионный анализ (ANOVA**) – это метод дисперсионного анализа, при котором исследуется влияние на зависимую переменную (признак) только одного фактора.

**Многофакторный дисперсионный анализ (MANOVA**) – это метод дисперсионного анализа, при котором исследуется влияние на зависимую переменную (признак) двух и более факторов.

**Гипотеза** – это недоказанное утверждение (предположение) относительно фактора (явления), интересующего исследователя. Она может представлять собой и возможный ответ на вопрос исследователя.

**Нулевая (прямая) гипотеза $H_0$** – это утверждение (предположение) (например, об отсутствии статистически значимой корреляционной связи между признаком и фактором), которое принимается, когда нет убедительных аргументов для его отклонения.

**Альтернативная (исследовательская, обратная) гипотеза $H_1$** – это утверждение (предположение) (например, о наличии статистически значимой корреляционной связи между признаком и фактором), которое принимается, когда есть убедительное статистическое доказательство, которое отвергает приемлемость нулевой гипотезы.

**Корреляционное отношение ($\eta^2$)** – это показатель, с помощью которого выражают степень влияния или силу эффекта независимой переменной (фактора) (независимых переменных (факторов)) на зависимую переменную. Его значение лежит в интервале от 0 до 1,0.

**Критерий Фишера (*F*-статистика, *F*-тест)** – это статистический критерий, с помощью которого проверяют нулевую гипотезу. В рамках дисперсионного анализа с его помощью проверяется нулевая гипотеза о равенстве категориальных средних в выборочных совокупностях.

### 6.1.2 Однофакторный дисперсионный анализ

Порядок выполнения однофакторного дисперсионного анализа включает этапы, показанные на рисунке 6.1.


Рисунок 6.1 – Порядок проведения однофакторного дисперсионного анализа

### 6.1.2.1 Определение зависимой и независимой переменных

Если $y$ – зависимая переменная, а $x$ – категориальная независимая переменная, имеющая $m$ категорий (уровней, групп), то для каждой категории $x$ может быть получено $n$ наблюдений $y$. Их результат можно представить, как это показано в таблице 6.1.

Таблица 6.1 – Матрица результатов наблюдений, полученных в ходе эксперимента

| Уровни $x$ | Результат наблюдения, $y$ | | | | Средние значения по уровням $x$ |
|---|---|---|---|---|---|
| | 1 | 2 | … | $n$ | |
| 1 | $y_{11}$ | $y_{12}$ | … | $y_{1n}$ | $\overline{y}_1 = \frac{\sum_{j=1}^{n} y_{1j}}{n}$ |
| 2 | $y_{21}$ | $y_{22}$ | … | $Y_{2n}$ | $\overline{y}_2 = \frac{\sum_{j=1}^{n} y_{2j}}{n}$ |
| … | … | … | … | … | … |
| $m$ | $y_{m1}$ | $y_{m2}$ | … | $y_{mn}$ | $\overline{y}_m = \frac{\sum_{j=1}^{n} y_{mj}}{n}$ |
| Среднее значение $y$ по всем уровням $x$ | | | | | $\overline{y}_{mn} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij}}{N}$ |

### 6.1.2.2 Разложение полной вариации

Полную вариацию $SS_y$ рассчитывают как сумму квадратов с поправкой на среднее (на число степеней свободы):

$$SS_y = \sum_{j=1}^{m}\sum_{i=1}^{n}\left(y_{ij} - \overline{y}\right)^2, \qquad (6.1)$$

где $y_{ij}$ – значение $i$-го наблюдения зависимой переменной при $j$-м уровне независимой переменной $x$;
$\overline{y}$ – средняя арифметическая всей совокупности наблюдений (значений зависимой переменной).

Ее также можно разложить на две составляющие:

$$SS_y = SS_{\text{между}} + SS_{\text{внутри}} \quad \text{или} \quad SS_y = SS_x + SS_{\text{ошибки}}\,, \qquad (6.2)$$

где $SS_{\text{между}}$ ($SS_x$) – сумма квадратов отклонений между категориями (рассеивание по факторам), характеризующая вариацию переменной $y$, связанную с различием средних между уровнями переменной $x$;
$SS_{внутри}$ ($SS_{ошибки}$) – сумма квадратов отклонений внутри уровня (остаточное рассеивание), характеризующая вариацию переменной внутри каждого уровня переменной $x$.

При этом

$$SS_x = \sum_{j=1}^{m} \left(\overline{y_j} - \overline{y}\right)^2, \tag{6.3}$$

$$SS_{\text{ошибки}} = \sum_{j=1}^{m} \sum_{i=1}^{n} \left(y_{ij-}\overline{y_j}\right)^2, \tag{6.4}$$

где $\overline{y_j}$ – среднее значение зависимой переменной при *j*-м уровне независимой переменной *x*.

#### 6.1.2.3 Оценка степени влияния независимой переменной на зависимую

Степень влияния (сила эффекта) независимой переменной *x* на зависимую *y* вычисляется по формуле

$$\eta^2 = \frac{SS_x}{SS_y} \quad \text{или} \quad \eta^2 = \frac{SS_y - SS_{\text{ошибки}}}{SS_y}. \tag{6.5}$$

Значения корреляционного отношения $\eta^2$ лежат в пределах от 0 до 1,0. Когда средние значения зависимой переменной *y* на всех уровнях независимой переменной *x* равны между собой, значение этого показателя равно нулю. Если внутри каждого уровня независимой переменной *x* значения зависимой переменной *y* равны между собой, но имеется некоторая изменчивость между уровнями независимой переменной *x*, значение $\eta^2$ равно 1,0.

#### 6.1.2.4 Проверка значимости нулевой гипотезы

Нулевая гипотеза при однофакторном дисперсионном анализе утверждает, что все средние значения зависимой переменной *y* на рассматриваемых уровнях *x* равны между собой:

$$H_0: \overline{y}_1 = \overline{y}_2 = \cdots = \overline{y}_m, \tag{6.6}$$

а альтернативная, наоборот, что они между собой различаются:

$$H_1: \overline{y}_1 \neq \overline{y}_2 \neq \cdots \neq \overline{y}_m. \tag{6.7}$$

Нулевую гипотезу проверяют с помощью *F*-статистики, в рамках однофакторного дисперсионного анализа рассчитываемой по формуле

$$F_{\text{расч}} = \frac{SS_x m(n-1)}{SS_{\text{ошибки}}(m-1)}. \tag{6.8}$$

### 6.1.2.5 Интерпретация результатов

Рассчитанное значение $F$-статистики сравнивается с ее критическим (табличным) значением. Если $F_{\text{расч}} < F_{\text{крит}}$, то принимают нулевую гипотезу, гласящую что независимая переменная $x$ не оказывает статистически значимого влияния на зависимую переменную $y$. Если же $F_{\text{расч}} > F_{\text{крит}}$, то нулевую гипотезу отклоняют и принимают альтернативную, согласно которой эффект независимой переменной $x$ на зависимую $y$ считается статистически значимым, т. е. среднее значение зависимой переменной $y$ различно для различных уровней независимой переменной $x$.

В том случае, если будет установлено существенное влияние фактора на изменчивость признака, исследуется значимость его средних значений на отдельных уровнях фактора в следующем порядке:

– рассчитывается разность средних оценок признака для каждой пары уровней факторов, например:

$$\Delta_{y_{x_1x_2}} = \overline{y_{x_1}} - \overline{y_{x_2}}\ ; \tag{6.9}$$

– рассчитывается внутригрупповая вариация оценок признака для всех вариантов фактора в выборке

$$SS_y = \frac{\sum_{j=1}^{m} S_{y_j}^2 \left(n_{x_j} - 1\right)}{n - m}; \tag{6.10}$$

– рассчитывается стандартная ошибка для разности средних оценок признака для каждой пары уровней факторов, например:

$$S_{y_{\overline{x_1x_2}}} = \sqrt{SS_y \left(\frac{1}{n_{y_{x_1}}} + \frac{1}{n_{y_{x_2}}}\right)}; \tag{6.11}$$

– для каждой пары уровней факторов рассчитывается доверительный интервал (как правило 95%-й) для разности средних оценок признака, например:

$$\Delta_{y_{x_1x_2}} - z \cdot S_{y_{\overline{x_1x_2}}} < \Delta_{y_{x_1x_2}} < \Delta_{y_{x_1x_2}} + z \cdot S_{y_{\overline{x_1x_2}}}\ ; \tag{6.12}$$

– с использованием $t$-критерия Стьюдента устанавливается статистическая значимость разницы оценок признака для каждой пары факторов, например:

$$t_{y_{x_1x_2}} = \frac{\Delta_{y_{x_1x_2}}}{S_{y_{\overline{x_1x_2}}}}\,. \tag{6.13}$$

### 6.1.3 Многофакторный дисперсионный анализ

Многофакторный дисперсионный анализ проводится в том же порядке, что и однофакторный. Главным отличием является наличие действия, направленного на оценку степени взаимодействия факторов, которое имеет место, когда эффекты одной независимой переменной, например, $x_1$ на зависимую переменную $y$ зависят от уровней других факторов, например, $x_2$, $x_k$, $x_l$, …, $x_p$.

Статистики, соответствующие многофакторному дисперсионному анализу, определяются аналогично определению статистик в однофакторном дисперсионном анализе.

#### 6.1.3.1 Двухфакторный дисперсионный анализ без повторений

##### 6.1.3.1.1 Определение зависимой и независимых переменных

Если $y$ – зависимая переменная, а $x_1$ и $x_2$ – категориальные независимые переменные, имеющие соответственно $m$ и $p$ категорий (уровней, групп), то по каждой из категорий $x_1$ *и* $x_2$ может быть получено $n$ наблюдений $y$. Их результат можно представить, как это показано в таблице 6.2.

##### 6.1.3.1.2 Разложение полной вариации

Полную вариацию $SS_y$ рассчитывают по формуле

$$SS_y = \sum_{j=1}^{m}\sum_{k=1}^{p}\sum_{i=1}^{n}\left(y_{ijp} - \overline{y}\right)^2, \qquad (6.14)$$

где $y_{jki}$ – значение $i$-го наблюдения зависимой переменной при $j$-м уровне независимой переменной $x_1$ и $k$-м уровне независимой переменной $x_2$.

Ее также можно разложить на четыре составляющие:

$$SS_y = SS_{x_1} + SS_{x_2} + SS_{x_1x_2} + SS_{\text{ошибки}}\,, \qquad (6.15)$$

где $SS_{x_1}$ и $SS_{x_2}$ – суммы квадратов отклонений между категориями (рассеивание по факторам), характеризующие вариацию переменной $y$, связанную с различием средних между уровнями соответственно переменных $x_1$ и $x_2$;
$SS_{x_1x_2}$ – сумма квадратов отклонений между категориями (рассеивание по факторам), характеризующая вариацию переменной $y$, связанную с взаимным влиянием друг на друга переменных $x_1$ и $x_2$;
$SS_{\text{ошибки}}$ – сумма квадратов отклонений внутри уровней (остаточное рассеивание), характеризующая вариацию переменной внутри каждого уровня переменных $x_1$ и $x_2$ соответственно.

Таблица 6.2 – Матрица результатов наблюдений, полученных в ходе эксперимента

| Уровни | | Результат наблюдения, $y$ | | | | Средние значения по уровням $x_1$ | Средние значения по уровням $x_2$ | Средние значения по уровням $x_1$ и $x_2$ |
|---|---|---|---|---|---|---|---|---|
| $x_1$ | $x_2$ | 1 | 2 | … | $n$ | | | |
| 1 | 1 | $y_{111}$ | $y_{112}$ | … | $y_{11n}$ | $\overline{y}_{1ki} = \frac{\sum_{k=1}^{p}\sum_{i=1}^{n} y_{1ki}}{pn}$ | $\overline{y}_{j1i} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n} y_{j1i}}{mn}$ | $\overline{y}_{jk1} = \frac{\sum_{j=1}^{m}\sum_{k=1}^{p} y_{jk1}}{n}$ |
| | 2 | $y_{121}$ | $y_{122}$ | … | $y_{12n}$ | | | |
| | … | … | … | … | … | | | |
| | $p$ | $y_{1k1}$ | $y_{1k2}$ | … | $y_{1kn}$ | | | |
| 2 | 1 | $y_{211}$ | $y_{212}$ | … | $y_{21n}$ | $\overline{y}_{2ki} = \frac{\sum_{k=1}^{p}\sum_{i=1}^{n} y_{2ki}}{pn}$ | $\overline{y}_{j2i} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n} y_{j2i}}{mn}$ | $\overline{y}_{jk2} = \frac{\sum_{j=1}^{m}\sum_{k=1}^{p} y_{jk2}}{n}$ |
| | 2 | $y_{221}$ | $y_{222}$ | … | $y_{22n}$ | | | |
| | … | … | … | … | … | | | |
| | $p$ | $y_{2k1}$ | $y_{2k2}$ | … | $y_{2kn}$ | | | |
| … | 1 | … | … | … | … | … | … | … |
| | 2 | … | … | … | … | | | |
| | … | … | … | … | … | | | |
| | $p$ | … | … | … | … | | | |
| $m$ | 1 | $y_{m11}$ | $y_{m12}$ | … | $y_{m1n}$ | $\overline{y}_{mki} = \frac{\sum_{k=1}^{p}\sum_{i=1}^{n} y_{mki}}{pn}$ | $\overline{y}_{jki} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n} y_{jki}}{mn}$ | $\overline{y}_{jkn} = \frac{\sum_{j=1}^{m}\sum_{k=1}^{p} y_{jkn}}{n}$ |
| | 2 | $y_{m21}$ | $y_{m22}$ | … | $y_{m2n}$ | | | |
| | … | … | … | … | … | | | |
| | $p$ | $y_{mk1}$ | $y_{mk2}$ | … | $y_{mkn}$ | | | |
| Среднее значение зависимой переменной по всей выборке $\overline{y}$ | | | | | | $\overline{y} = \frac{\sum_{j=1}^{m}\sum_{k=1}^{p}\sum_{i=1}^{n} y_{jki}}{mpn}$ | | |

При этом:

$$SS_{x_1} = pn\sum_{j=1}^{m}\left(\frac{\sum_{k=1}^{p}\sum_{i=1}^{n} y_{jki}}{pn} - \overline{y}\right)^2, \tag{6.16}$$

$$SS_{x_2} = mn\sum_{k=1}^{p}\left(\frac{\sum_{j=1}^{m}\sum_{i=1}^{n} y_{jki}}{mn} - \overline{y}\right)^2, \tag{6.17}$$

$$SS_{x_1x_2} = n\sum_{j=1}^{m}\sum_{k=1}^{p}\sum_{i=1}^{n}\left(\frac{y_{jki}}{n} - \overline{y}\right)^2 - SS_{x_1} - SS_{x_2}, \tag{6.18}$$

$$SS_{\text{ошибки}} = SS_y - SS_{x_1} - SS_{x_2} - SS_{x_1x_2}. \tag{6.19}$$

Большее влияние $x_1$ будет отражаться в большем отличии среднего в уровнях $x_1$ и более высоком значении $SS_{x_1}$. Это же касается и фактора $x_2$. Чем сильнее взаимодействие между факторами $x_1$ и $x_2$, тем больше значение $SS_{x_1x_2}$. С другой стороны, если $x_1$ и $x_2$ не зависят один от другого, то значение $SS_{x_1x_2}$ приближается к нулю.

#### 6.1.3.1.3 Оценка степени влияния независимой переменной на зависимую

Степень объединенного влияния (сила эффекта) двух факторов на зависимую переменную *y*, называемая полным эффектом или множественной корреляцией, вычисляется по формуле

$$\eta^2 = \frac{SS_{x_1} + SS_{x_2} + SS_{x_1x_2}}{SS_y} \text{ или } \eta^2 = \frac{SS_y - SS_{\text{ошибки}}}{SS_y}. \tag{6.20}$$

Значения корреляционного отношения $\eta^2$ лежат в пределах от 0 до 1,0.

#### 6.1.3.1.4 Проверка значимости нулевой гипотезы

Значимость полного эффекта проверяют с помощью *F*-статистики, рассчитываемой в рамках многофакторного дисперсионного анализа по формуле

$$F_{\text{расч}} = \frac{\left(SS_{x_1} + SS_{x_2} + SS_{x_1x_2}\right)mp(n-1)}{SS_{\text{ошибки}}(mp-1)}. \tag{6.21}$$

Если полный эффект оказывается статистически значимым, то после этого оценивают значимость эффекта взаимодействия. Если нулевая гипотеза утверждает, что взаимодействие между факторами отсутствует, то соответствующий *F*-критерий вычисляется по формуле

$$F_{x_1x_2} = \frac{SS_{x_1x_2} mp(n-1)}{SS_{\text{ошибки}}[(m-1)(p-1)]}. \quad (6.22)$$

Если окажется, что эффект взаимодействия статистически значимый, значит эффект $x_1$ зависит от $x_2$ и наоборот. Поскольку эффект (влияние) одного фактора не является однородным, а зависит от уровня другого фактора, то вообще бессмысленно проверять значимость главных эффектов. Однако имеет смысл проверить значимость главного эффекта каждого фактора, если эффект взаимодействия статистически незначимый.

Значимость главного эффекта каждого фактора может быть проверена следующим образом:

$$F_{x_1} = \frac{SS_{x_1} mp(n-1)}{SS_{\text{ошибки}}(m-1)}, \quad (6.23)$$

$$F_{x_2} = \frac{SS_{x_2} mp(n-1)}{SS_{\text{ошибки}}(p-1)}. \quad (6.24)$$

В том случае, если будет установлено существенное влияние факторов на изменчивость признака, исследуется значимость его средних на отдельных уровнях.

С этой целью для каждого фактора отдельно:

– рассчитывается значение исправленной (случайной) дисперсии зависимой переменной по формуле

$$s_z^2 = \frac{SS_{\text{ошибки}}}{mp(n-1)}; \quad (6.25)$$

– рассчитывается стандартная ошибка разности средних значений зависимой переменной для каждого фактора:

$$s_{\overline{x_1}} = \sqrt{\frac{2SS_{\text{ошибки}}^2}{np}} \quad \text{и} \quad s_{\overline{x_2}} = \sqrt{\frac{2SS_{\text{ошибки}}^2}{nm}}; \quad (6.26)$$

– с помощью таблицы *t*-распределения Стьюдента (или с использованием функции «СТЬЮДЕНТ.ОБР.2Х(α;*v*)» находится значение коэффициента доверия *z*, соответствующее заданной вероятности и степеням свободы ($v = mk(n-1)$), и вычисляется предельная ошибка разности средних по формулам

$$\varepsilon_1 = \pm z s_{\overline{x_1}} \quad \text{и} \quad \varepsilon_2 = \pm z s_{\overline{x_2}}; \quad (6.27)$$

– все средние значения признака попарно сравниваются. Если разности между рассматриваемой парой средних значений превышают предельную ошибку, то делается вывод о существенном влиянии фактора для этой пары.

При исследовании существенности влияния взаимодействия факторов выполняются следующие действия:

– записываются в виде таблицы (матрицы) средние значения признака для обоих факторов;

– по формуле (6.27) рассчитывается предельная ошибка разности средних;

– по всем уровням обоих факторов устанавливаются пары значений признака, разности которых больше предельной ошибки. Для таких пар взаимодействие факторов признается существенным;

– устанавливается сочетание факторов, при котором достигается наибольшее значение рассматриваемого признака.

### 6.1.3.2 Двухфакторный дисперсионный анализ с повторениями

#### 6.1.3.2.1 Разложение полной вариации

Полную вариацию $SS_y$ рассчитывают так же, как и для двухфакторного дисперсионного анализа без повторений по формуле (6.14).

Ее также разлагают на четыре составляющие по формуле (6.15). При этом $SS_{x_1}$, $SS_{x_2}$ и $SS_{\text{ошибки}}$ рассчитываются соответственно по формулам (6.16), (6.17) и (6.19) (таблица 6.3).

Но сумму квадратов отклонений между категориями (рассеивание по факторам), характеризующую вариацию переменной $y$, связанную с взаимным влиянием друг на друга переменных $x_1$ и $x_2$, рассчитывают по формуле

$$SS_{x_1x_2} = n\sum_{j=1}^{m}\sum_{k=1}^{p}\left[\left(\overline{y_{jk}} - \overline{y_j}\right) - \left(\overline{y_k} - \overline{y}\right)\right]^2, \qquad (6.28)$$

где $\overline{y_{jk}}$ – среднее значение переменной $y$ для $j$-го уровня переменной $x_1$ и для $k$-го уровня (повторения) переменной $x_2$;

$\overline{y_j}$ – среднее значение переменной $y$ для $j$-го уровня переменной $x_1$ и для всех уровней (повторений) переменной $x_2$;

$\overline{y_k}$ – среднее значение переменной $y$ для всех уровней переменной $x_1$ и для $k$-го уровня (повторения) переменной $x_2$.

#### 6.1.3.2.2 Оценка степени влияния независимой переменной на зависимую

Степень объединенного влияния (сила эффекта) двух факторов на зависимую переменную $y$, называемая полным эффектом или множественной корреляцией, вычисляется по формуле (6.20).

Таблица 6.3 – Матрица результатов наблюдений, полученных в ходе эксперимента

| Уровни фактора $x_1$ | Отдельные значения зависимой переменной $y$ по $n$ наблюдениям в зависимости от уровня фактора $x_2$ | | | | | | | | | | | | | Средние значения $y$ по уровням фактора $x_1$ для каждого из $p$ уровней фактора $x_2$ | | | | Среднее значение $y$ для всех уровней $x_2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | | | | 2 | | | | … | $p$ | | | | 1 | 2 | … | $p$ | |
| | 1 | 2 | … | $n$ | 1 | 2 | … | $n$ | | 1 | 2 | … | $n$ | | | | | |
| 1 | $y_{111}$ | $y_{112}$ | … | $y_{11n}$ | $y_{121}$ | $y_{122}$ | … | $y_{12n}$ | … | $y_{1p1}$ | $y_{1p2}$ | … | $y_{1pn}$ | $\overline{y_{11}} = \frac{\sum_{j=1}\sum_{i=1}^{n} y_{11n}}{n}$ | $\overline{y_{12}} = \frac{\sum_{j=1}\sum_{i=1}^{n} y_{12n}}{n}$ | … | $\overline{y_{1p}} = \frac{\sum_{j=1}\sum_{i=1}^{n} y_{1pn}}{n}$ | $\overline{y_1} = \frac{\sum_{j=1}\sum_{i=1}^{n} y_{1pn}}{pn}$ |
| 2 | $y_{211}$ | $y_{212}$ | … | $y_{21n}$ | $y_{221}$ | $y_{222}$ | … | $y_{22n}$ | … | $y_{2p1}$ | $y_{2p2}$ | … | $y_{2pn}$ | $\overline{y_{21}} = \frac{\sum_{j=2}\sum_{i=1}^{n} y_{21n}}{n}$ | $\overline{y_{22}} = \frac{\sum_{j=2}\sum_{i=1}^{n} y_{22n}}{n}$ | … | $\overline{y_{2p}} = \frac{\sum_{j=2}\sum_{i=1}^{n} y_{2pn}}{n}$ | $\overline{y_2} = \frac{\sum_{j=2}\sum_{i=2}^{n} y_{1pn}}{pn}$ |
| … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … | … |
| $m$ | $y_{m11}$ | $y_{m12}$ | … | $y_{m1n}$ | $y_{m21}$ | $y_{m22}$ | … | $y_{m2n}$ | | $y_{mp1}$ | $y_{mp2}$ | … | $y_{mpn}$ | $\overline{y_{m1}} = \frac{\sum_{j=m}\sum_{i=1}^{n} y_{m1n}}{n}$ | $\overline{y_{m2}} = \frac{\sum_{j=m}\sum_{i=1}^{n} y_{m2n}}{n}$ | … | $\overline{y_{mp}} = \frac{\sum_{j=m}\sum_{i=1}^{n} y_{mpn}}{n}$ | $\overline{y_m} = \frac{\sum_{j=m}\sum_{i=1}^{n} y_{1pn}}{pn}$ |
| Среднее значение $y$ для каждого из уровней фактора $x_1$ | $\overline{y_1} = \frac{\sum_{k=1}\sum_{i=1}^{n} y_{m1n}}{mn}$ | | | | $\overline{y_2} = \frac{\sum_{k=2}\sum_{i=1}^{n} y_{m2n}}{mn}$ | | | | … | $\overline{y_p} = \frac{\sum_{k=p}\sum_{i=1}^{n} y_{mpn}}{mn}$ | | | | Среднее значение по всей выборке: $\overline{y} = \frac{\sum_{j=1}^{m}\sum_{k=1}^{p}\sum_{i=1}^{n} y_{jki}}{mpn}$ | | | | |

#### 6.1.3.2.3 Проверка значимости нулевой гипотезы

Значимость полного эффекта проверяют с помощью $F$-статистики, рассчитываемой по формуле (6.21). Если полный эффект оказывается статистически значимым, то после этого оценивают значимость эффекта взаимодействия. Если нулевая гипотеза утверждает, что взаимодействие между факторами отсутствует, то соответствующий $F$-критерий вычисляется по формуле

$$F_{x_1x_2} = \frac{SS_{x_1x_2}mp(n-1)}{SS_{\text{ошибки}}(p-1)}. \quad (6.29)$$

Значимость главного эффекта каждого фактора может быть проверена по формулам (6.23) и (6.24) соответственно.

## 6.2 Выполнение однофакторного дисперсионного анализа с использованием приложения MS Excel и программы IBM SPSS Statistics

Рекламным подразделением службы маркетинга ОАО «Крессида» совместно с сотрудниками отдела маркетинга ЧУП «Кэтнес» по итогам маркетингового исследования, рассмотренного в лабораторной работе № 3, разработаны три варианта рекламно-информационных материалов для мест продажи (POS-материалов) и три варианта рекламы на областном телевидении для запланированных к производству и продаже журнальных столиков, характеристики которых были описаны в лабораторной работе № 4. При создании сценариев рекламы, содержания и оформления POS-материалов использовались возможности инструментов искусственного интеллекта.

Для оценки уровня коммуникационного эффекта, выбора наилучшей комбинации вариантов POS-материалов и рекламы были проведены три эксперимента, для которых были сформированы три однородные группы из выборки, которая рассматривалась в лабораторной работе № 5. Численность каждой группы при оценке коммуникационного эффекта отдельно материалов в местах продажи и телерекламы составила 30 человек, а при оценке их совместного эффекта – 15 человек.

В ходе первого эксперимента для каждой группы был продемонстрирован свой вариант POS-материалов, в ходе второго – свой вариант телерекламы. Третий эксперимент предполагал просмотр телерекламы с последующей демонстрацией POS-материалов.

В таблицах 6.4 и 6.5 представлены оценки, выставленные по десятибалльной шкале респондентами предложенным им вариантам соответственно POS-материалов и телерекламы, а в таблице 6.6 – оценки, отражающие степень готовности приобрести покупателями продукцию компании, после ознакомления с предложенными комбинациями вариантов POS-материалов и телерекламы.

Таблица 6.4 – Оценки, выставленные респондентами предложенным вариантам POS-материалов

| Номер респондента | Оценки за варианты POS-материалов | | |
|---|---|---|---|
| | 1-й | 2-й | 3-й |
| 1 | 1 | 8 | 7 |
| 2 | 2 | 7 | 6 |
| 3 | 3 | 6 | 7 |
| 4 | 2 | 9 | 5 |
| 5 | 2 | 5 | 7 |
| 6 | 3 | 9 | 5 |
| 7 | 2 | 5 | 4 |
| 8 | 5 | 9 | 7 |
| 9 | 2 | 10 | 6 |
| 10 | 1 | 7 | 5 |
| 11 | 1 | 5 | 4 |
| 12 | 1 | 7 | 5 |
| 13 | 2 | 10 | 5 |
| 14 | 5 | 10 | 4 |
| 15 | 2 | 5 | 5 |
| 16 | 1 | 6 | 7 |
| 17 | 2 | 8 | 4 |
| 18 | 3 | 8 | 4 |
| 19 | 5 | 9 | 4 |
| 20 | 1 | 10 | 5 |
| 21 | 3 | 8 | 7 |
| 22 | 3 | 5 | 5 |
| 23 | 5 | 7 | 6 |
| 24 | 4 | 5 | 5 |
| 25 | 3 | 10 | 7 |
| 26 | 5 | 7 | 6 |
| 27 | 4 | 8 | 5 |
| 28 | 1 | 9 | 5 |
| 29 | 3 | 10 | 7 |
| 30 | 1 | 5 | 4 |

Таблица 6.5 – Оценки, выставленные респондентами предложенным вариантам телерекламы

| Номер респондента | Оценки за варианты телерекламы | | |
|---|---|---|---|
| | 1-й | 2-й | 3-й |
| 1 | 5 | 4 | 10 |
| 2 | 4 | 8 | 6 |

| Номер респондента | Оценки за варианты телерекламы | | |
|---|---|---|---|
| | 1-й | 2-й | 3-й |
| 3 | 4 | 4 | 5 |
| 4 | 3 | 4 | 6 |
| 5 | 4 | 4 | 6 |
| 6 | 4 | 7 | 7 |
| 7 | 2 | 7 | 10 |
| 8 | 5 | 5 | 10 |
| 9 | 1 | 7 | 7 |
| 10 | 4 | 4 | 9 |
| 11 | 3 | 8 | 9 |
| 12 | 5 | 4 | 7 |
| 13 | 3 | 6 | 10 |
| 14 | 1 | 8 | 10 |
| 15 | 4 | 4 | 5 |
| 16 | 4 | 5 | 7 |
| 17 | 6 | 6 | 9 |
| 18 | 6 | 4 | 6 |
| 19 | 2 | 4 | 6 |
| 20 | 4 | 7 | 7 |
| 21 | 1 | 8 | 10 |
| 22 | 3 | 8 | 8 |
| 23 | 6 | 8 | 5 |
| 24 | 2 | 6 | 6 |
| 25 | 2 | 7 | 6 |
| 26 | 5 | 6 | 8 |
| 27 | 3 | 8 | 7 |
| 28 | 4 | 7 | 6 |
| 29 | 1 | 6 | 8 |
| 30 | 3 | 6 | 6 |

Таблица 6.6 – Оценки, отражающие степень готовности приобрести продукцию компании после ознакомления с предложенными комбинациями вариантов POS-материалов и телерекламы

| Номер респондента | Вариант POS-материалов | Вариант рекламы | Степень готовности |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 4 |
| 3 | 1 | 1 | 2 |
| 4 | 1 | 1 | 3 |
| 5 | 1 | 1 | 2 |
| 6 | 1 | 2 | 2 |

| Номер респондента | Вариант POS-материалов | Вариант рекламы | Степень готовности |
|---|---|---|---|
| 7 | 1 | 2 | 3 |
| 8 | 1 | 2 | 4 |
| 9 | 1 | 2 | 3 |
| 10 | 1 | 2 | 5 |
| 11 | 1 | 3 | 3 |
| 12 | 1 | 3 | 5 |
| 13 | 1 | 3 | 6 |
| 14 | 1 | 3 | 4 |
| 15 | 1 | 3 | 5 |
| 1 | 2 | 1 | 2 |
| 2 | 2 | 1 | 3 |
| 3 | 2 | 1 | 4 |
| 4 | 2 | 1 | 4 |
| 5 | 2 | 1 | 3 |
| 6 | 2 | 2 | 6 |
| 7 | 2 | 2 | 7 |
| 8 | 2 | 2 | 5 |
| 9 | 2 | 2 | 6 |
| 10 | 2 | 2 | 6 |
| 11 | 2 | 3 | 8 |
| 12 | 2 | 3 | 8 |
| 13 | 2 | 3 | 10 |
| 14 | 2 | 3 | 9 |
| 15 | 2 | 3 | 9 |
| 1 | 3 | 1 | 5 |
| 2 | 3 | 1 | 6 |
| 3 | 3 | 1 | 7 |
| 4 | 3 | 1 | 6 |
| 5 | 3 | 1 | 5 |
| 6 | 3 | 2 | 5 |
| 7 | 3 | 2 | 6 |
| 8 | 3 | 2 | 4 |
| 9 | 3 | 2 | 7 |
| 10 | 3 | 2 | 5 |
| 11 | 3 | 3 | 6 |
| 12 | 3 | 3 | 7 |
| 13 | 3 | 3 | 8 |
| 14 | 3 | 3 | 5 |
| 15 | 3 | 3 | 6 |

Необходимо установить:

– насколько высоко варианты POS-материалов и телерекламы оцениваются потенциальными покупателями;

– видят ли они различие в оформлении и содержании предложенных им вариантов POS-материалов и телерекламы и насколько статистически значимым оно является;

– наиболее эффективную комбинацию вариантов оформления и содержания POS-материалов и телерекламы, оказывающую наибольшее стимулирующее воздействие на потенциальных покупателей;

– значимость полного эффекта, оказываемого сочетанием этих факторов, значимость эффекта их взаимодействия, значимость главного эффекта каждого фактора, а также существенность влияния взаимодействия факторов.

Все расчеты представлять с точностью до сотых.

### 6.2.1 Оценка уровня и значимости различий в оформлении и содержании вариантов POS-материалов с использованием приложения MS Excel

1 Для решения необходимо выдвинуть следующие статистические гипотезы:

– нулевую, в соответствии с которой различий между вариантами оформления и содержания POS-материалов не существует или они существуют, но являются статистически незначимыми ($H_0$: $\overline{x_1} = \overline{x_2} = \overline{x_3}$; $F_{\text{расч}} < F_{\text{крит}}$);

– альтернативную, в соответствии с которой различия между вариантами оформления и содержания POS-материалов существуют и они являются статистически значимыми ($H_1$: $\overline{x_1} \neq \overline{x_2} \neq \overline{x_3}$; $F_{\text{расч}} > F_{\text{крит}}$).

2 Создать в приложении новый файл и присвоить ему имя «06 Дисперсионный анализ.xlsx».

3 В созданном файле присвоить листу название «POSM» и внести в него данные из таблицы 6.4 (рисунок 6.2).

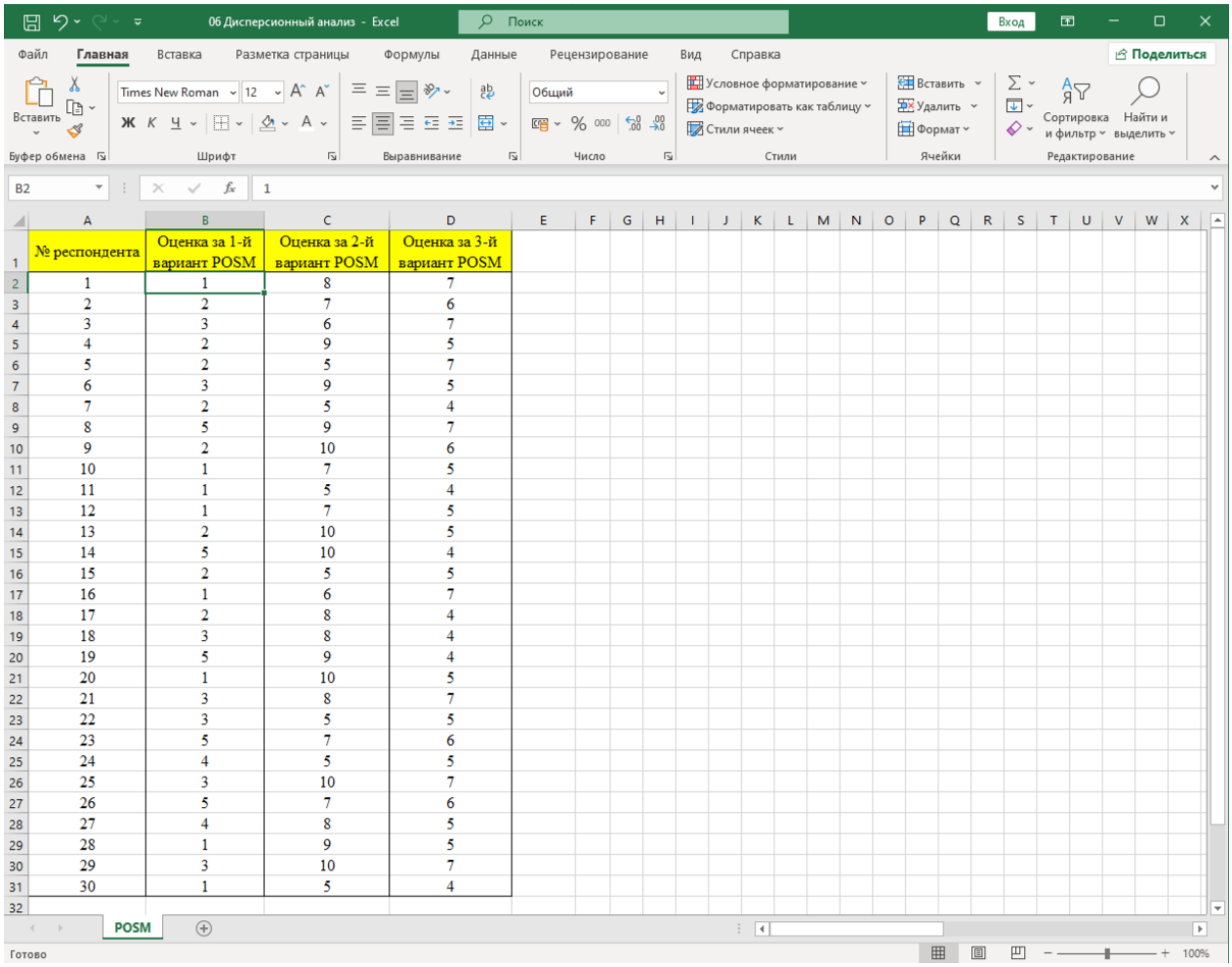
4 Выполнить однофакторный дисперсионный анализ с целью установления того, насколько высоко оцениваются потенциальными покупателями варианты предложенных им POS-материалов, видят ли они различия между их оформлением и содержанием, насколько статистически значимыми эти различия являются. Для этого (рисунок 6.3):

– выбрать инструмент анализа **«Однофакторный дисперсионный анализ»** (**«Данные»** – **«Анализ данных»** – **«Однофакторный дисперсионный анализ»**);

– ввести в поле **«Входной интервал:»** значения оценок вариантов POS-материалов (с указанием их вариантов) (ячейки **«F3»** – **«AJ5»**), группирование выбрать по строкам. Так как в столбце **«F»** будут находиться варианты POS-материалов, поставить флажок напротив строки **«Метки в первом столбце»**;

– в секции **«Параметры вывода»** выбрать **«Выходной интервал»**, в которой указать ячейку **«F2»**, и нажать кнопку «**ОК**»;

– используя шрифт Times New Roman Cyr размером 12 пт, отформатировать полученные данные.



| № респондента | Оценка за 1-й вариант POSM | Оценка за 2-й вариант POSM | Оценка за 3-й вариант POSM |
|---|---|---|---|
| 1 | 1 | 8 | 7 |
| 2 | 2 | 7 | 6 |
| 3 | 3 | 6 | 7 |
| 4 | 2 | 9 | 5 |
| 5 | 2 | 5 | 7 |
| 6 | 3 | 9 | 5 |
| 7 | 2 | 5 | 4 |
| 8 | 5 | 9 | 7 |
| 9 | 2 | 10 | 6 |
| 10 | 1 | 7 | 5 |
| 11 | 1 | 5 | 4 |
| 12 | 1 | 7 | 5 |
| 13 | 2 | 10 | 5 |
| 14 | 5 | 10 | 4 |
| 15 | 2 | 5 | 5 |
| 16 | 1 | 6 | 7 |
| 17 | 2 | 8 | 4 |
| 18 | 3 | 8 | 4 |
| 19 | 5 | 9 | 4 |
| 20 | 1 | 10 | 5 |
| 21 | 3 | 8 | 7 |
| 22 | 3 | 5 | 5 |
| 23 | 5 | 7 | 6 |
| 24 | 4 | 5 | 5 |
| 25 | 3 | 10 | 7 |
| 26 | 5 | 7 | 6 |
| 27 | 4 | 8 | 5 |
| 28 | 1 | 9 | 5 |
| 29 | 3 | 10 | 7 |
| 30 | 1 | 5 | 4 |

Рисунок 6.2 – Таблица с оценками, выставленными респондентами предложенным вариантам POS-материалов

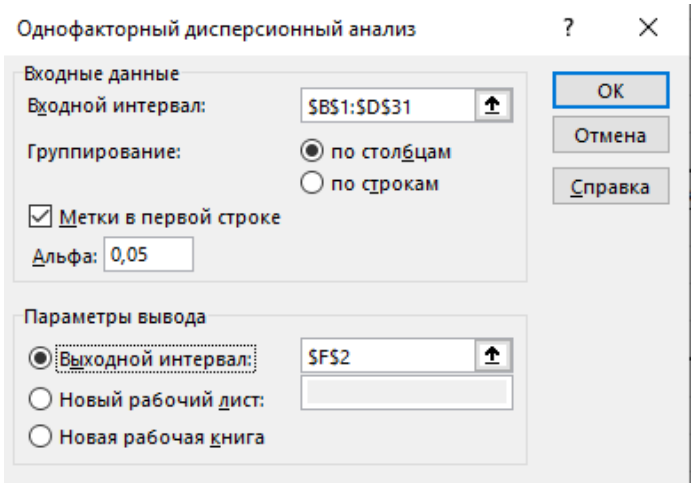



Рисунок 6.3 – Диалог «Однофакторный дисперсионный анализ» с внесенными данными по вариантам POS-материалов

Результаты расчетов, выполненных приложением, представлены в таблицах 6.7 и 6.8.

Таблица 6.7 – Результаты расчетов оценок, выставленных покупателями предложенным им вариантам POS-материалов

| Группы | Счет | Сумма | Среднее | Дисперсия |
|---|---|---|---|---|
| Оценка за 1-й вариант POSM | 30 | 78 | 2,60 | 1,97 |
| Оценка за 2-й вариант POSM | 30 | 227 | 7,57 | 3,43 |
| Оценка за 3-й вариант POSM | 30 | 163 | 5,43 | 1,29 |

Таблица 6.8 – Результаты однофакторного дисперсионного анализа оценок, выставленных покупателями предложенным им вариантам POS-материалов

| Источник вариации | Сумма квадратов, *SS* | Степени свободы, *df* | Средний квадрат, *MS* | Расчетное значение *F* | *P*-значение | Критическое значение *F* |
|---|---|---|---|---|---|---|
| Между группами | 372,47 | 2 | 186,23 | 83,55 | 0,00 | 3,10 |
| Внутри групп | 193,93 | 87 | 2,23 | | | |
| Итого | 566,40 | 89 | | | | |

Как видно из этих таблиц, наибольшую оценку получил второй вариант POS-материалов. Кроме этого, средние значения оценок, выставленных покупателями вариантам POS-материалов, различаются, а так как расчетное значение *F*-критерия превышает его критическое (что подтверждает и величина *P*-значения), то различия являются статистически значимыми. Таким образом, следует принять альтернативную гипотезу.

Для оценки степени влияния POS-материалов на готовность покупателей приобрести предлагаемые им компанией журнальные столики, рассчитать величину корреляционного отношения по формуле (6.5):

$$\eta^2 = \frac{SS_x}{SS_y} = \frac{372{,}47}{566{,}40} = 0{,}66.$$

Полученное значение показателя говорит о том, что на оформление и содержание POS-материалов приходиться примерно 66 % эффекта воздействия на принятие решения о покупке продукции, т. е. уровень влияния этих материалов является достаточно сильным. На остальные факторы (в том числе и на телерекламу), которые влияют на решение о приобретении журнальных столиков компании, приходятся остальные 34 %.

5 Приняв альтернативную гипотезу и оценив эффект воздействия POS-материалов на готовность потенциальными покупателями приобрести продукцию ЧУП «Кэтнес», установить статистическую значимость разниц средних оценок для всех вариантов их оформления и содержания.

Для этого:

– рассчитать по формуле (6.9) разность средних оценок для каждой пары вариантов POS-материалов:

$$\Delta_{y_{x_1x_2}} = 2{,}6 - 7{,}57 = -4{,}97,$$

$$\Delta_{y_{x_1x_3}} = 2{,}6 - 5{,}43 = -2{,}83,$$

$$\Delta_{y_{x_2x_3}} = 7{,}57 - 5{,}43 = 2{,}14;$$

– рассчитать по формуле (6.10) внутригрупповую вариацию оценок для всех вариантов POS-материалов

$$SS_y = \frac{1{,}97^2 \cdot (30-1) + 3{,}43^2 \cdot (30-1) + 1{,}29^2 \cdot (30-1)}{90-3} = 5{,}17;$$

– рассчитать по формуле (6.11) стандартную ошибку для разности каждой пары средних оценок. Так как количество наблюдений для каждого варианта POS-материалов одинаково и равно 30, то для каждой пары стандартная ошибка будет одинакова и равна ее значению для первой пары:

$$s_{y_{\overline{x_1x_2}}} = s_{y_{\overline{x_1x_3}}} = s_{y_{\overline{x_2x_3}}} = \sqrt{5{,}17 \cdot \left(\frac{1}{30} + \frac{1}{30}\right)} = 0{,}59.$$

Для первого и второго вариантов POS-материалов 95%-й доверительный интервал для разности средних оценок, рассчитанный по формуле (6.12), находится между

$$-4{,}97 - 1{,}96 \cdot 0{,}59 < \Delta_{y_{x_1x_2}} < -4{,}97 + 1{,}96 \cdot 0{,}59,$$

$$-6{,}13 < \Delta_{y_{x_1x_2}} < -3{,}80\ ,$$

а значение *t*-статистики, рассчитанное по формуле (6.13), равно

$$t_{x_1x_2} = \frac{\Delta_{y_{x_1x_2}}}{S_{y_{\overline{x_1x_2}}}} = \frac{-4{,}97}{0{,}59} = -8{,}46,$$

что больше ее критического значения (приблизительно 1,96) для 87 степеней свободы. Кроме этого, видно, что между нижней и верхней границами доверительного интервала не находится значение «ноль». Таким образом, разница средних оценок первого и второго вариантов POS-материалов является статистически значимой.

Аналогично для остальных двух сравнений пар POS-материалов:

– для первого и третьего:

$$-2{,}83 - 1{,}96 \cdot 0{,}59 < \Delta_{y_{x_1x_3}} < -2{,}83 + 1{,}96 \cdot 0{,}59,$$

$$-4{,}00 < \Delta_{y_{x_1x_3}} < -1{,}67\ ,$$

$$t_{x_1x_3} = \frac{\Delta_{y_{x_1x_3}}}{S_{y_{\overline{x_1x_3}}}} = \frac{-2{,}83}{0{,}39} = -4{,}83;$$

– для второго и третьего:

$$2{,}14 - 1{,}96 \cdot 0{,}59 < \Delta_{y_{x_2x_3}} < 2{,}14 + 1{,}96 \cdot 0{,}59,$$

$$0{,}98 < \Delta_{y_{x_2x_3}} < 3{,}30,$$

$$t_{x_2x_3} = \frac{\Delta_{y_{x_2x_3}}}{S_{y_{\overline{x_2x_3}}}} = \frac{2{,}14}{0{,}59} = 3{,}63\ .$$

Так как внутри границ всех трех доверительных интервалов нет значения «ноль», разницы средних оценок для всех пар вариантов POS-материалов являются статистически значимыми.

### 6.2.2 Оценка уровня и значимости различий в оформлении и содержании вариантов POS-материалов с использованием программы IBM SPSS Statistics

1 Создать в программе новый файл и дать ему название «06 Дисперсионный анализ.sav».

2 Во вкладке «Данные» редактора данных в первую колонку сначала внести по 30 раз «1» (1-й вариант POS-материалов), «2» (2-й вариант POS-материалов и «3» (3-й вариант POS-материалов). После этого для указанных вариантов во вторую колонку скопировать оценки потребителей.

3 Далее:

– нажав кнопку **«Переменные»** в окне редактора данных, перейти в одноименную вкладку и присвоить имена переменным – «Вариант POSM» и «Оценка варианта POSM»;

– для обеих переменных задать тип **«Числовой»**, установить ширину в восемь символов без десятичных знаков после запятой, ширину колонки в восемь символов, выравнивание по центру и роль **«Входная»**;

– для переменной «Вариант POSM» в колонке **«Шкала»** установить **«Номинальные»**, а для «Оценка вариантов POSM – **«Шкалы»**;

– нажав кнопку **«Данные»**, вернуться в одноименную вкладку и при необходимости изменить ширину колонок до удобной для восприятия (рисунок 6.4).

4 Выполнить однофакторный дисперсионный анализ для рассматриваемых вариантов POS-материалов. Для этого:

– выбрать процедуру **«Однофакторный дисперсионный анализ» («Анализ» – «Сравнение средних» – «Однофакторный дисперсионный анализ»**);

– в открывшемся диалоговом окне перенести в поле **«Список зависимых переменных»** переменные, связанные со всеми тремя вариантами POS-материалов, а в поле **«Фактор»** – переменную «Вариант POSM» (рисунок 6.5);

– нажав кнопку **«Параметры…»**, в наборе статистик выбрать **«Описательные»** и нажать кнопку **«Продолжить»**;

– нажать кнопку **«ОК»**.

После выполнения расчетов созданному файлу дать имя «06 Дисперсионный анализ.spv», но его не закрывать.

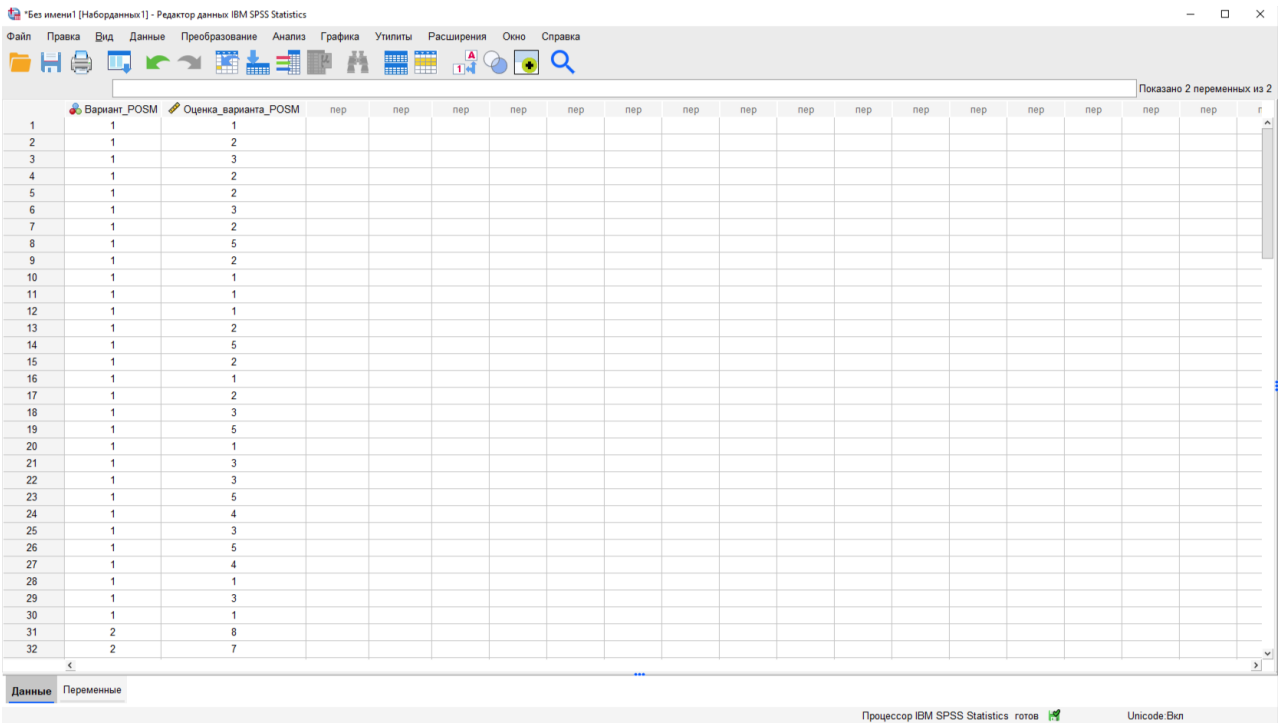



Рисунок 6.4 – Вкладка «Данные» с частью (32 из 90) оценок покупателей по вариантам POS-материалов

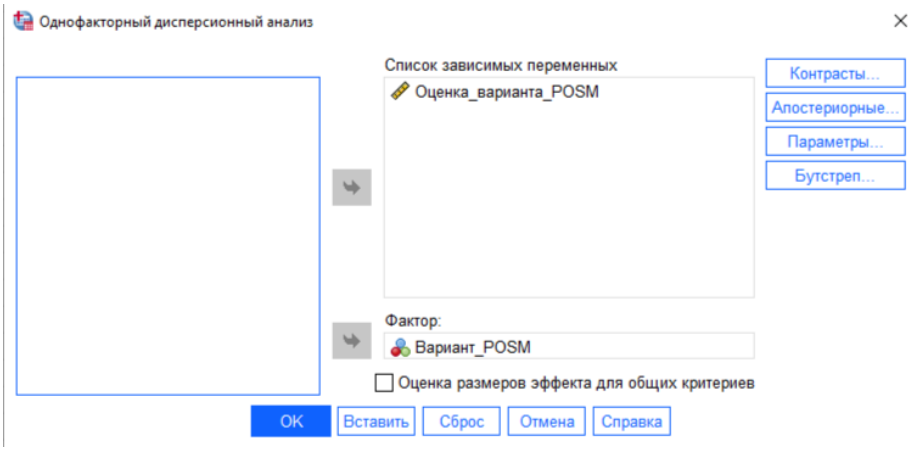



Рисунок 6.5 – Диалоговое окно «Однофакторный дисперсионный анализ» с внесенными переменными

В результате выполненных действий программа выведет таблицы, содержание которых представлено в таблицах 6.9 и 6.10. На основе их значений можно утверждать, что, так как расчетное значение *F*-критерия превышает его

критическое значение (что подтверждает и величина *P*-значения), средние значения оценок, выставленных покупателями вариантам POS-материалов, различаются и эти различия являются статистически значимыми. Таким образом, следует принять альтернативную гипотезу.

Таблица 6.9 – Результаты расчетов оценок, выставленных покупателями предложенным им вариантов POS-материалов

| Группы | Счет | Среднее | Стандартное отклонение | Стандартная ошибка | 95%-й доверительный интервал для среднего | | Минимум | Максимум |
|---|---|---|---|---|---|---|---|---|
| | | | | | нижняя граница | верхняя граница | | |
| 1 | 30 | 2,60 | 1,40 | 0,26 | 2,08 | 3,12 | 1 | 5 |
| 2 | 30 | 7,57 | 1,85 | 0,34 | 6,88 | 8,26 | 5 | 10 |
| 3 | 30 | 5,43 | 1,14 | 0,21 | 5,01 | 5,86 | 4 | 7 |
| Итого | 90 | 5,20 | 2,52 | 0,276 | 4,67 | 5,73 | 1 | 10 |

Таблица 6.10 – Результаты однофакторного дисперсионного анализа оценок, выставленных покупателями предложенным им вариантам POS-материалов

| Источник вариации | Сумма квадратов, *SS* | Степени свободы, *df* | Средний квадрат, *MS* | Расчетное значение *F* | *P*-значение |
|---|---|---|---|---|---|
| Между группами | 372,47 | 2 | 186,23 | 83,55 | 0,00 |
| Внутри групп | 193,93 | 87 | 2,23 | | |
| Итого | 566,40 | 89 | | | |

Величина корреляционного отношения получается точно такой же, как и по итогам расчета величин в приложении MS Excel:

$$\eta^2 = \frac{SS_x}{SS_y} = \frac{372{,}47}{566{,}40} = 0{,}66.$$

Исследование значимости разниц средних оценок для всех пар вариантов POS-материалов дает те же результаты, что и полученные по итогам использования приложения MS Excel.

### 6.2.3 Оценка уровня и значимости различий в содержании вариантов телерекламы с использованием приложения MS Excel

1 В приложении создать новый лист и дать ему название «Телереклама».

2 Для решения задачи необходимо выдвинуть следующие статистические гипотезы:

– нулевую, в соответствии с которой различий между вариантами содержания телерекламы не существует или они существуют, но являются статистически незначимыми ($H_0$: $\overline{x_1} = \overline{x_2} = \overline{x_3}$; $F_{\text{расч}} < F_{\text{крит}}$);

– альтернативную, в соответствии с которой различия между вариантами содержания телевизионной рекламы существуют и они являются статистически значимыми ($H_1$: $\overline{x_1} \neq \overline{x_2} \neq \overline{x_3}$; $F_{\text{расч}} > F_{\text{крит}}$).

3 Задачу решить в том же порядке, что и задачу по оценке вариантов POS-материалов.

Ниже представлены итоговые таблицы 6.11 и 6.12, полученные в результате решения задачи.

Таблица 6.11 – Результаты расчетов по оценке покупателями предложенным им вариантам телерекламы

| Группы | Счет | Сумма | Среднее | Дисперсия |
|---|---|---|---|---|
| Оценка за 1-й вариант рекламы | 30 | 104 | 3,47 | 2,26 |
| Оценка за 2-й вариант рекламы | 30 | 180 | 6,00 | 2,48 |
| Оценка за 3-й вариант рекламы | 30 | 222 | 7,40 | 2,94 |

Таблица 6.12 – Результаты однофакторного дисперсионного анализа оценок, выставленных покупателями предложенным им вариантам телерекламы

| Источник вариации | Сумма квадратов, *SS* | Степени свободы, *df* | Средний квадрат, *MS* | Расчетное значение *F* | *P*-значение | Критическое значение *F* |
|---|---|---|---|---|---|---|
| Между группами | 238,49 | 2 | 119,24 | 46,59 | 0,00 | 3,10 |
| Внутри групп | 222,67 | 87 | 2,56 | | | |
| Итого | 461,16 | 89 | | | | |

Как видно из этих таблиц, наибольшую оценку получил третий вариант телерекламы. Кроме этого, средние значения оценок, выставленных покупателями вариантам телерекламы, различаются, а так как расчетное значение *F*-критерия превышает его критическое значение (что подтверждает и величина *P*-значения), то различие является статистически значимым. Таким образом, следует принять альтернативную гипотезу.

4 Для оценки степени влияния телерекламы на готовность покупателями приобрести предлагаемые им компанией журнальные столики, рассчитать величину корреляционного отношения по формуле (6.5):

$$\eta^2 = \frac{SS_x}{SS_y} = \frac{238{,}49}{461{,}16} = 0{,}52.$$

Полученное значение показателя говорит о том, что на телерекламу приходиться примерно 52 % эффекта воздействия на принятие решения о покупке продукции, т. е. уровень ее влияния является относительно высоким. На остальные факторы, в том числе и на POS-материалы, которые влияют на решение приобрести журнальные столики ЧУП «Кэтнес», приходятся остальные 48 %.

5 Приняв альтернативную гипотезу и оценив эффект воздействия телерекламы, исследовать значимость разниц средних оценок для всех вариантов ее содержания.

Для этого:

– рассчитать по формуле (6.9) разность средних оценок для каждой пары вариантов телерекламы:

$$\Delta_{y_{x_1x_2}} = 3{,}47 - 6{,}00 = -2{,}53,$$

$$\Delta_{y_{x_1x_3}} = 3{,}47 - 7{,}40 = -3{,}93,$$

$$\Delta_{y_{x_2x_3}} = 6{,}00 - 7{,}40 = -1{,}40;$$

– рассчитать по формуле (6.10) внутригрупповую вариацию оценок для всех вариантов телерекламы

$$SS_y = \frac{2{,}26^2 \cdot (30-1) + 2{,}48^2 \cdot (30-1) + 2{,}94^2 \cdot (30-1)}{90-3} = 6{,}63;$$

– рассчитать по формуле (6.11) стандартную ошибку для разности каждой пары средних оценок. Так как количество наблюдений для каждого варианта телерекламы одинаково и равно 30, то для каждой пары стандартная ошибка будет одинакова и равна ее значению для первой пары:

$$s_{y_{\overline{x_1x_2}}} = s_{y_{\overline{x_1x_3}}} = s_{y_{\overline{x_2x_3}}} = \sqrt{SS_y\left(\frac{1}{n_{x_1}} + \frac{1}{n_{x_2}}\right)} = \sqrt{6{,}63 \cdot \left(\frac{1}{30} + \frac{1}{30}\right)} = 0{,}66.$$

Для первого и второго вариантов телерекламы 95%-й доверительный интервал для разности средних оценок, рассчитанный по формуле (6.12), находится между

$$-2{,}53 - 1{,}96 \cdot 0{,}66 < \Delta_{y_{x_1x_2}} < -2{,}53 + 1{,}96 \cdot 0{,}66,$$

$$-3{,}84 < \Delta_{y_{x_1x_2}} < -1{,}23,$$

а значение *t*-статистики равно

$$t_{y_{x_1x_2}} = \frac{\Delta_{y_{x_1x_2}}}{S_{y_{\overline{x_1x_2}}}} = \frac{-2{,}53}{0{,}66} = -3{,}81,$$

что больше ее критического значения (приблизительно 1,96) для 87 степеней свободы. Кроме этого, видно, что между нижней и верхней границами доверительного интервала не находится значение «ноль». Таким образом, разница средних оценок первого и второго вариантов телерекламы является статистически значимой.

Аналогично для остальных двух сравнений пар телевизионной рекламы:

– для первого и третьего:

$$-3{,}93 - 1{,}96 \cdot 0{,}66 < \Delta_{y_{x_1x_3}} < -3{,}93 + 1{,}96 \cdot 0{,}66,$$

$$-5{,}23 < \Delta_{y_{x_1x_3}} < -2{,}63,$$

$$t_{x_1x_3} = \frac{\Delta_{y_{x_1x_3}}}{S_{y_{\overline{x_1x_3}}}} = \frac{-3{,}93}{0{,}41} = -5{,}92;$$

– для второго и третьего:

$$-1{,}4 - 1{,}96 \cdot 0{,}66 < \Delta_{y_{x_2x_3}} < -1{,}4 + 1{,}96 \cdot 0{,}66,$$

$$-2{,}70 < \Delta_{y_{x_2x_3}} < -0{,}10,$$

$$t_{x_2x_3} = \frac{\Delta_{y_{x_2x_3}}}{S_{y_{\overline{x_2x_3}}}} = \frac{-1{,}40}{0{,}41} = -2{,}11.$$

Так как внутри границ всех трех доверительных интервалов нет значения «ноль», разницы средних оценок для всех пар вариантов телерекламы являются статистически значимыми.

### 6.2.4 Оценка уровня и значимости различий в оформлении и содержании вариантов телерекламы с использованием программы IBM SPSS Statistics

В программе IBM SPSS Statistics в той же вкладке **«Данные»**, которая использовалась при оценке вариантов POS-материалов, внести данные по вариантам телерекламы. Но теперь варианты телерекламы внести в третью колонку (их можно просто скопировать из первой колонки, так как количество вариантов POS-материалов равно количеству вариантов телерекламы), а оценки – в четвертую.

Оценку выполнить в том же порядке, что и задачу по оценке вариантов POS-материалов. До этого в процедуре не забыть удалить из диалога **«Список зависимых переменных»** и **«Фактор»** переменные, которые использовались для оценки их вариантов.

Ниже представлены итоговые таблицы 6.13 и 6.14, полученные в результате решения.

Таблица 6.13 – Результаты расчетов оценок, выставленных покупателями предложенным им вариантам телерекламы

| Группы | Счет | Среднее | Стандартное отклонение | Стандартная ошибка | 95%-й доверительный интервал для среднего | | Минимум | Максимум |
|---|---|---|---|---|---|---|---|---|
| | | | | | нижняя граница | верхняя граница | | |
| 1 | 30 | 3,47 | 1,50 | 0,27 | 2,91 | 4,03 | 1 | 6 |
| 2 | 30 | 6,00 | 1,58 | 0,29 | 5,41 | 6,59 | 4 | 8 |
| 3 | 30 | 7,40 | 1,71 | 0,31 | 6,76 | 8,04 | 5 | 10 |
| Итого | 90 | 5,62 | 2,28 | 0,24 | 5,15 | 6,10 | 1 | 10 |

Таблица 6.14 – Результаты однофакторного дисперсионного анализа оценок, выставленных покупателями предложенным им вариантам телерекламы

| Источник вариации | Сумма квадратов, *SS* | Степени свободы, *df* | Средний квадрат, *MS* | Расчетное значение *F* | *P*-значение |
|---|---|---|---|---|---|
| Между группами | 238,49 | 2 | 119,24 | 46,59 | 0,00 |
| Внутри групп | 222,67 | 87 | 2,56 | | |
| Итого | 461,16 | 89 | | | |

Величина корреляционного отношения получается абсолютно точно такой же, как и по итогам расчета величин в приложении MS Excel:

$$\eta^2 = \frac{SS_x}{SS_y} = \frac{238{,}49}{461{,}16} = 0{,}52.$$

Исследование значимости разниц средних оценок для всех пар вариантов телерекламы также дает те же результаты, что и по итогам использования программы MS Excel.

## 6.3 Выполнение двухфакторного дисперсионного анализа с использованием приложения MS Excel и программы IBM SPSS Statistics

### 6.3.1 Оценка уровня и значимости различий в воздействии на потребителя различных комбинаций вариантов POS-материалов и телерекламы с использованием приложения MS Excel

1 Для решения задачи необходимо выдвинуть следующие статистические гипотезы:

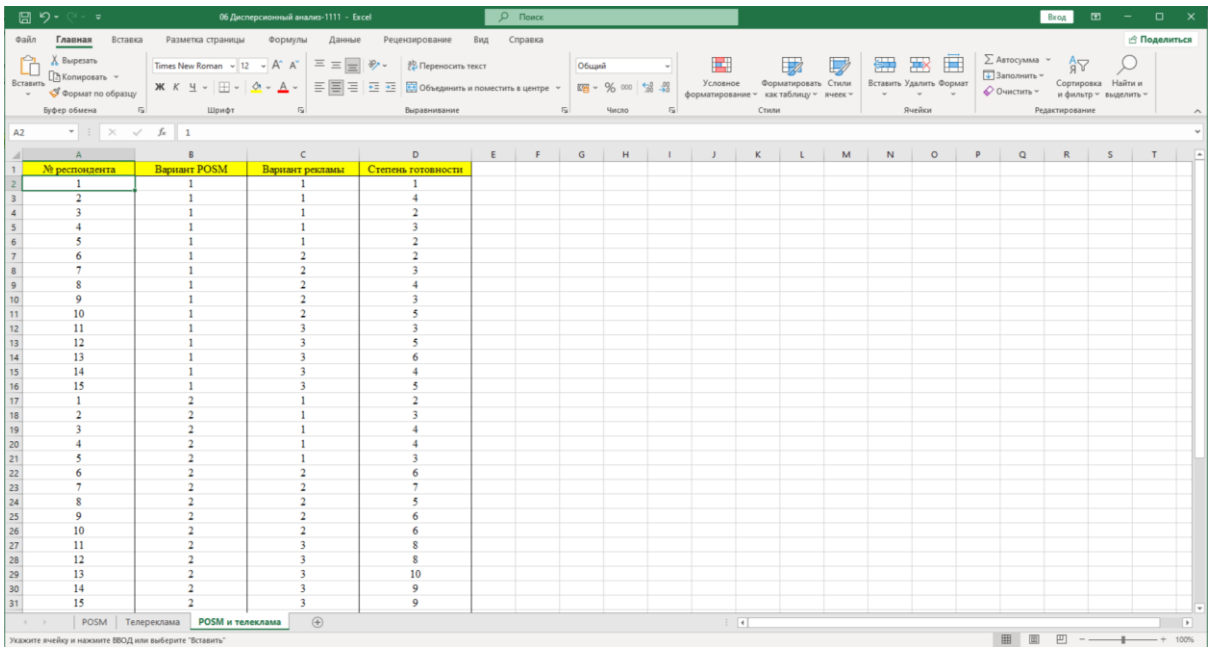
– нулевую, в соответствии с которой взаимодействие эффектов, производимых на покупателей оформлением и содержанием POS-материалов с одной стороны и телерекламой с другой, нет (эффект от POS-материалов на степень готовности к покупке не зависит от эффекта телевизионной рекламы и наоборот), но, если оно есть, то является статистически незначимым ($H_0$: $\eta^2 \neq 0$; $F_{\text{расч}} < F_{\text{крит}}$);

– альтернативную, в соответствии с которой взаимодействие эффектов, производимых на покупателей оформлением и содержанием POS-материалов с одной стороны и телерекламой с другой, имеет место (эффект от POS-материалов на степень готовности к покупке зависит от эффекта телерекламы и наоборот) и оно является статистически значимым ($H_1$: $\eta^2 \neq 0$; $F_{\text{расч}} > F_{\text{крит}}$).

2 В ранее созданном файле создать новый лист и присвоить ему имя «POSM и телереклама» и внести в него данные из таблицы 6.6 (рисунок 6.6).

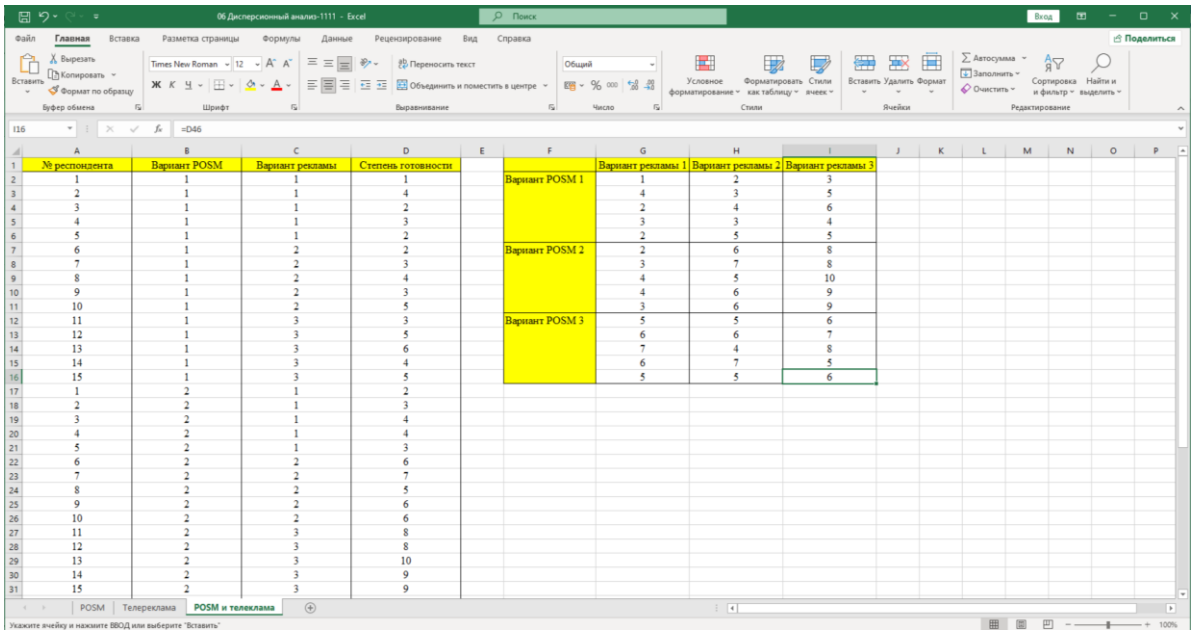
3 С использованием этих данных создать специальную таблицу для выполнения двухфакторного дисперсионного анализа (рисунок 6.7).

4 В связи с тем, что инструмент анализа «Двухфакторный дисперсионный анализ без повторений» приложения MS Excel в результате его использования не выводит значение суммы квадратов отклонений между категориями (рассеивание по факторам), характеризующее вариацию переменной *y*, связанную с взаимным влиянием друг на друга переменных $x_1$ и $x_2$ ($SS_{x_1x_2}$), выполнить двухфакторный дисперсионный анализ с повторениями с целью установления наличия взаимодействия эффектов, производимых на покупателей POS-материалами с одной стороны и телерекламой с другой (когда эффект от POS-материалов на степень готовности к покупке зависит от эффекта телерекламы и наоборот), и оценки его статистической значимости.



| № респондента | Вариант POSM | Вариант рекламы | Степень готовности |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 4 |
| 3 | 1 | 1 | 2 |
| 4 | 1 | 1 | 3 |
| 5 | 1 | 1 | 2 |
| 6 | 1 | 2 | 2 |
| 7 | 1 | 2 | 3 |
| 8 | 1 | 2 | 4 |
| 9 | 1 | 2 | 3 |
| 10 | 1 | 2 | 5 |
| 11 | 1 | 3 | 3 |
| 12 | 1 | 3 | 5 |
| 13 | 1 | 3 | 6 |
| 14 | 1 | 3 | 4 |
| 15 | 1 | 3 | 5 |
| 1 | 2 | 1 | 2 |
| 2 | 2 | 1 | 3 |
| 3 | 2 | 1 | 4 |
| 4 | 2 | 1 | 4 |
| 5 | 2 | 1 | 3 |
| 6 | 2 | 2 | 6 |
| 7 | 2 | 2 | 7 |
| 8 | 2 | 2 | 5 |
| 9 | 2 | 2 | 6 |
| 10 | 2 | 2 | 6 |
| 11 | 2 | 3 | 8 |
| 12 | 2 | 3 | 8 |
| 13 | 2 | 3 | 10 |
| 14 | 2 | 3 | 9 |
| 15 | 2 | 3 | 9 |

Рисунок 6.6 – Таблица с оценками степени готовности приобрести продукцию компании, выставленными респондентами после ознакомления с предложенными вариантами POS-материалов и телерекламы



| № респондента | Вариант POSM | Вариант рекламы | Степень готовности |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 4 |
| 3 | 1 | 1 | 2 |
| 4 | 1 | 1 | 3 |
| 5 | 1 | 1 | 2 |
| 6 | 1 | 2 | 2 |
| 7 | 1 | 2 | 3 |
| 8 | 1 | 2 | 4 |
| 9 | 1 | 2 | 3 |
| 10 | 1 | 2 | 5 |
| 11 | 1 | 3 | 3 |
| 12 | 1 | 3 | 5 |
| 13 | 1 | 3 | 6 |
| 14 | 1 | 3 | 4 |
| 15 | 1 | 3 | 5 |
| 1 | 2 | 1 | 2 |
| 2 | 2 | 1 | 3 |
| 3 | 2 | 1 | 4 |
| 4 | 2 | 1 | 4 |
| 5 | 2 | 1 | 3 |
| 6 | 2 | 2 | 6 |
| 7 | 2 | 2 | 7 |
| 8 | 2 | 2 | 5 |
| 9 | 2 | 2 | 6 |
| 10 | 2 | 2 | 6 |
| 11 | 2 | 3 | 8 |
| 12 | 2 | 3 | 8 |
| 13 | 2 | 3 | 10 |
| 14 | 2 | 3 | 9 |
| 15 | 2 | 3 | 9 |

| | Вариант рекламы 1 | Вариант рекламы 2 | Вариант рекламы 3 |
|---|---|---|---|
| Вариант POSM 1 | 1 | 2 | 3 |
| | 4 | 3 | 5 |
| | 2 | 4 | 6 |
| | 3 | 3 | 4 |
| | 2 | 5 | 5 |
| Вариант POSM 2 | 2 | 6 | 8 |
| | 3 | 7 | 8 |
| | 4 | 5 | 10 |
| | 4 | 6 | 9 |
| | 3 | 6 | 9 |
| Вариант POSM 3 | 5 | 5 | 6 |
| | 6 | 6 | 7 |
| | 7 | 4 | 8 |
| | 6 | 7 | 5 |
| | 5 | 5 | 6 |

Рисунок 6.7 – Исходная и специально созданная для проведения двухфакторного дисперсионного анализа таблица с оценками степени готовности приобрести продукцию компании

Для этого (рисунок 6.8):

– выбрать инструмент анализа **«Двухфакторный дисперсионный анализ с повторениями» («Данные» – «Анализ данных» – «Двухфакторный дисперсионный анализ с повторениями»)**;

– в появившемся диалоговом окне в поле **«Входной интервал:»** ввести значения оценок по совместному влиянию на готовность приобрести продукцию компании вариантов POS-материалов и телевизионной рекламы (с указанием их вариантов) (ячейки **«F1»** – **«I16»**), число строк по выборке установить 5 (равное числу оценок для каждого варианта телевизионной рекламы), значение α-критерия оставить равным 0,05;

– в поле **«Параметры вывода»** выбрать ячейку «**К2**» и нажать кнопку **«ОК»**;

– используя шрифт Times New Roman Cyr размером 12 пт, отформатировать полученные данные.

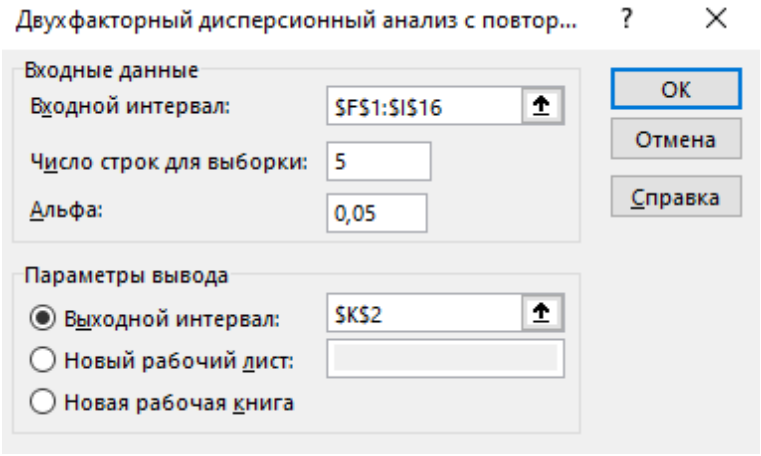



Рисунок 6.8 – Диалог «Двухфакторный дисперсионный анализ с повторениями» с внесенными данными из специально созданной таблицы

5 Вычисленные программой значения представлены в таблицах 6.15 и 6.16.

Таблица 6.15 – Результаты расчета средних значений оценок степени готовности приобрести продукцию компании, выставленных покупателями после ознакомления с предложенными им комбинациями вариантов POS-материалов и телерекламы

| Рассчитанные статистики | Вариант рекламы 1 | Вариант рекламы 2 | Вариант рекламы 3 | Итого |
|---|---|---|---|---|
| Вариант POS-материалов 1 | | | | |
| Счет | 5 | 5 | 5 | 15 |
| Сумма | 12,00 | 17,00 | 23,00 | 52,00 |
| Среднее | 2,40 | 3,40 | 4,60 | 3,47 |
| Дисперсия | 1,30 | 1,30 | 1,30 | 1,98 |
| Вариант POS-материалов 2 | | | | |
| Счет | 5 | 5 | | 15 |
| Сумма | 16,00 | 30,00 | 44,00 | 90,00 |
| Среднее | 3,20 | 6,0 | 8,80 | 6,00 |
| Дисперсия | 0,70 | 0,50 | 0,70 | 6,14 |

| Рассчитанные статистики | Вариант рекламы 1 | Вариант рекламы 2 | Вариант рекламы 3 | Итого |
|---|---|---|---|---|
| Вариант POS-материалов 3 | | | | |
| Счет | 5 | 5 | 5 | 15 |
| Сумма | 29,00 | 27,00 | 32,00 | 88,00 |
| Среднее | 5,80 | 5,40 | 6,40 | 5,87 |
| Дисперсия | 0,70 | 1,30 | 1,30 | 1,12 |
| Итого | | | | |
| Счет | 15 | 15 | 15 | |
| Сумма | 57,00 | 74,00 | 99,00 | |
| Среднее | 3,80 | 4,93 | 6,60 | |
| Дисперсия | 3,03 | 2,21 | 4,11 | |

Таблица 6.16 – Результаты двухфакторного дисперсионного анализа оценок степени готовности приобрести продукцию компании, выставленных покупателями после ознакомления с предложенными им комбинациями вариантов POS-материалов и телерекламы

| Источник вариации | Сумма квадратов, *SS* | Степени свободы, *df* | Средний квадрат, *MS* | Расчетное значение *F*-критерия | *P*-значение | Критическое значение *F*-критерия |
|---|---|---|---|---|---|---|
| Выборка | 60,98 | 2 | 30,49 | 30,15 | 0,00 | 3,26 |
| Столбцы | 59,51 | 2 | 29,76 | 29,43 | 0,00 | 3,26 |
| Взаимодействие | 33,56 | 4 | 8,39 | 8,30 | 0,00 | 2,63 |
| Внутри | 36,40 | 36 | 1,01 | | | |
| Итого | 190,44 | 44 | | | | |

Как видно из таблицы 6.15 и рисунка 6.9, наибольшую среднюю оценку получила комбинация из второго варианта POS-материалов и третьего варианта рекламы, а наименьшую – комбинация, состоящая из первых их вариантов.

Для оценки степени совместного влияния POS-материалов и телевизионной рекламы на готовность покупателями приобрести предлагаемые им компанией журнальные столики рассчитать величину корреляционного отношения по формуле (6.20):

$$\eta^2 = \frac{SS_{x_1} + SS_{x_2} + SS_{x_1x_2}}{SS_y} = \frac{60{,}98 + 59{,}51 + 33{,}56}{190{,}44} = 0{,}81.$$

Полученное значение показателя говорит о том, что на совместное влияние POS-материалов и телерекламы приходится примерно 81 % эффекта воздействия на принятие решения о покупке продукции компании, т. е. уровень влияния двух факторов является очень высоким. На остальные неучтенные факторы,

которые влияют на решение о приобретении журнальных столиков ЧУП «Кэтнес», приходятся остальные 19 %.

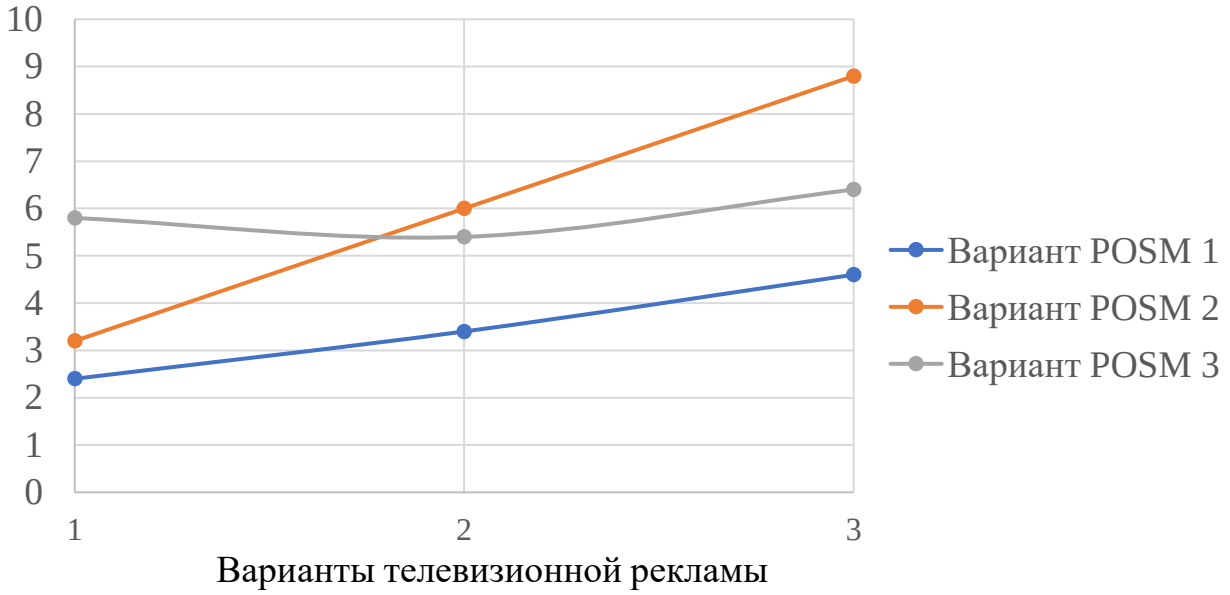



Рисунок 6.9 – Графики средних оценок степени готовности приобрести продукцию компании, выставленных респондентами после ознакомления с предложенными комбинациями вариантов POS-материалов и телерекламы

Значимость полного эффекта установить, сравнив расчетное и критическое значения $F$-критерия. Расчетное значение согласно формуле (6.21) будет равно

$$F_{\text{расч}} = \frac{(60{,}98 + 59{,}51 + 33{,}56) \cdot 3 \cdot 3 \cdot (15 - 1)}{36{,}40 \cdot (3 \cdot 3 - 1)} = \frac{154{,}05 \cdot 126}{36{,}40 \cdot 8} =$$

$$= \frac{19\,409{,}60}{291{,}20} = 66{,}65.$$

Критическое значение этого критерия при 8 и 126 степенях свободы примерно равно 2,02. Так как расчетное значение $F$-критерия существенно превышает критическое, можно утверждать, что полный эффект является статистически значимым.

Так как полный эффект оказался статистически значимым, то следует установить, является ли статистически значимым и эффект взаимодействия рассматриваемых факторов. Из таблицы 6.16 видно, что $P$-значение для $SS_{x_1x_2}$ равно примерно 0,00, что указывает на то, что расчетное значение $F$-критерия для этой статистики существенно превышает критическое и, следовательно, эффект взаимодействия является статистически значимым. Проверять значимость главных эффектов также не имеет смысла.

6 Исследовать значимость средних оценок степени готовности покупателями приобрести предлагаемые компанией журнальные столики при различных комбинациях вариантов POS-материалов и телерекламы.

Для этого:

– для каждого фактора отдельно рассчитать значения исправленной (случайной) дисперсии зависимой переменной. Так как количество уровней факторов одинаково и равно трем, а их комбинации были представлены 15 потребителям, то значение исправленной дисперсии оценок, рассчитанное по формуле (6.25), будет одинаково для обоих факторов:

$$s_z^2 = \frac{36{,}4}{3 \cdot 3 \cdot (15 - 1)} = 0{,}29;$$

– рассчитать стандартную ошибку разности средних значений зависимой переменной для каждого фактора. По причине равенства числа уровней факторов ее значение, рассчитанное по формуле (6.26), для них будет одинаковым:

$$s_{\overline{x_1}} = \sqrt{\frac{2 \cdot 0{,}29}{15 \cdot 3}} = 0{,}11;$$

– с помощью таблицы *t*-распределения Стьюдента (или с использованием функции **«СТЬЮДЕНТ.ОБР.2Х» («Формулы»** – **«Другие функции»** – **«Статистические»** – **«СТЬЮДЕНТ.ОБР.2Х»**) найти значение коэффициента доверия *z*, соответствующее заданной вероятности и степеням свободы ($v = mp(n - 1)$), и вычислить предельную ошибку разности средних. Критерий Стьюдента при $\alpha = 0{,}05$ и $v = 126$ равен 1,98, поэтому предельные ошибки разности средних для каждого из обоих факторов, рассчитанные по формуле (6.27), будут равны:

$$\varepsilon = \pm 1{,}98 \cdot 0{,}11 = \pm 0{,}22;$$

– попарно сравнить все средние значения оценок степени готовности покупателями приобрести продукцию и рассчитать их разницы (таблица 6.17).

Как видно из этой таблицы, разности средних оценок превышают предельную ошибку в размере 0,22 для всех пар комбинаций рассматриваемых факторов кроме пар, образованных:

– вторым вариантом POS-материалов совместно с первым вариантом телерекламы и первым вариантом POS-материалов совместно со вторым вариантом телерекламы;

– третьим вариантом POS-материалов совместно с первым вариантом телерекламы и вторым вариантом POS-материалов совместно со вторым вариантом телерекламы.

Следовательно, можно сделать вывод о существенном различии уровня влияния на степень готовности покупки продукции пяти из девяти комбинаций вариантов оформления и размещения POS-материалов и телерекламы.

Из этой же таблицы видно, что самой отличимой является и самую большую степень готовности приобрести продукцию компании вызывает комбинация из второго варианта POS-материалов и третьего варианта телерекламы.

Таблица 6.17 – Матрица абсолютных значений разниц оценок степени готовности покупателями приобрести продукцию компании для различных комбинаций вариантов POS-материалов и телерекламы

| Вариант POSM | | Вариант POSM | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | | | 2 | | | 3 | | |
| | вариант рекламы | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 |
| 1 | 1 | 0,00 | | | | | | | | |
| | 2 | 1,00 | 0,00 | | | | | | | |
| | 3 | 2,20 | 1,20 | 0,00 | | | | | | |
| 2 | 1 | 0,80 | 0,20 | 1,40 | 0,00 | | | | | |
| | 2 | 3,60 | 2,60 | 1,40 | 2,80 | 0,00 | | | | |
| | 3 | 6,40 | 5,40 | 4,20 | 5,60 | 2,80 | 0,00 | | | |
| 3 | 1 | 3,40 | 2,40 | 1,20 | 2,60 | 0,20 | 3,00 | 0,00 | | |
| | 2 | 3,00 | 2,00 | 0,80 | 2,20 | 0,60 | 3,40 | 0,40 | 0,00 | |
| | 3 | 4,00 | 3,00 | 1,80 | 3,20 | 0,40 | 2,40 | 0,60 | 1,00 | 0,00 |

### 6.3.2 Оценка уровня и значимости различий в воздействии на потребителя различных комбинаций вариантов POS-материалов и телерекламы с использованием программы IBM SPSS Statistics

1 В файле «06 Дисперсионный анализ.sav» вставить в пятую, шестую и седьмую колонки таблицы вкладки **«Данные»** данные из таблицы 6.6.

2 После этого:

– нажав кнопку **«Переменные»**, перейти в одноименную вкладку и присвоить имена переменным – «Вариант POSM-1», «Вариант рекламы-1» и «Степень готовности»;

– для всех переменных задать тип **«Числовой»**, ширину в восемь символов без десятичных знаков после запятой, ширину колонки в восемь символов, выравнивание по центру и роль **«Входная»**. Для переменных «Вариант POSM» и «Вариант рекламы» выбрать шкалу **«Номинальные»**, а для переменной «Степень готовности» – **«Шкалы»**;

– нажав кнопку **«Данные»**, вернуться в одноименную вкладку редактора данных и при необходимости изменить ширину колонок до удобной для восприятия.

3 Выполнить двухфакторный дисперсионный анализ для рассматриваемых комбинаций вариантов POS-материалов и телерекламы. Для этого:

– выбрать процедуру **«ОЛМ-одномерная» («Анализ»** – **«Общая линейная модель»** – **«ОЛМ-одномерная»**);

– в открывшемся диалоговом окне перенести в поле **«Зависимая переменная:»** переменную «Степень готовности», а в поле **«Фиксированные факторы:»** –

переменные «Вариант POSM-1» и «Вариант рекламы-1» (рисунок 6.10) («1» вписывается для того, чтобы отличить названия переменных от ранее введенных, которые находятся во вкладке «Данные»);

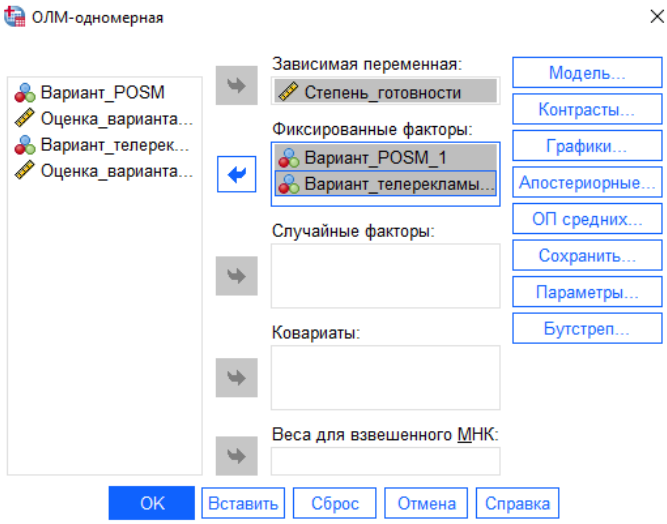



Рисунок 6.10 –Диалоговое окно «ОЛМ-одномерная» с внесенными переменными

– нажав кнопку **«Модель…»**, в открывшемся диалоге в поле **«Задать модель»** выбрать **«Полная факторная»**, сумму квадратов выбрать **«Тип III»**, убрать флажок напротив строки **«Включить в модель свободный член»** (рисунок 6.11) и, нажав кнопку **«Продолжить»**, вернуться в диалоговое окно выполняемой процедуры;

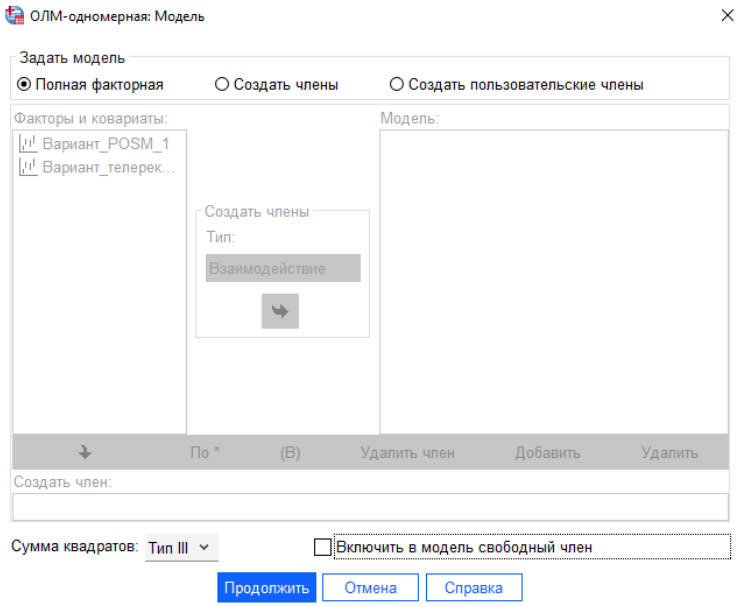



Рисунок 6.11 – Окно диалога «ОЛМ-одномерная: Модель» с установленными настройками

– нажав кнопку **«Контрасты…»** в секции **«Изменить контраст»**, выбрать (или оставить) **«Нет»** (рисунок 6.12) и, нажав кнопку **«Продолжить»**, вернуться в диалоговое окно процедуры;

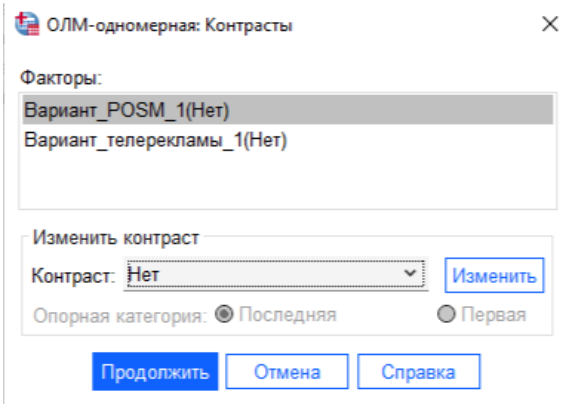



Рисунок 6.12 – Окно диалога «ОЛМ-одномерная: Контрасты»
с установленными настройками

– нажав кнопку **«Параметры…»**, в открывшемся диалоге в секции **«Вывести»** установить флажки напротив строк **«Описательные статистики»** и **«Оценки размера эффекта»** (рисунок 6.13) и, нажав кнопку **«Продолжить»**, вернуться в диалоговое окно процедуры;

– нажать кнопку **«ОК»** процедуры.

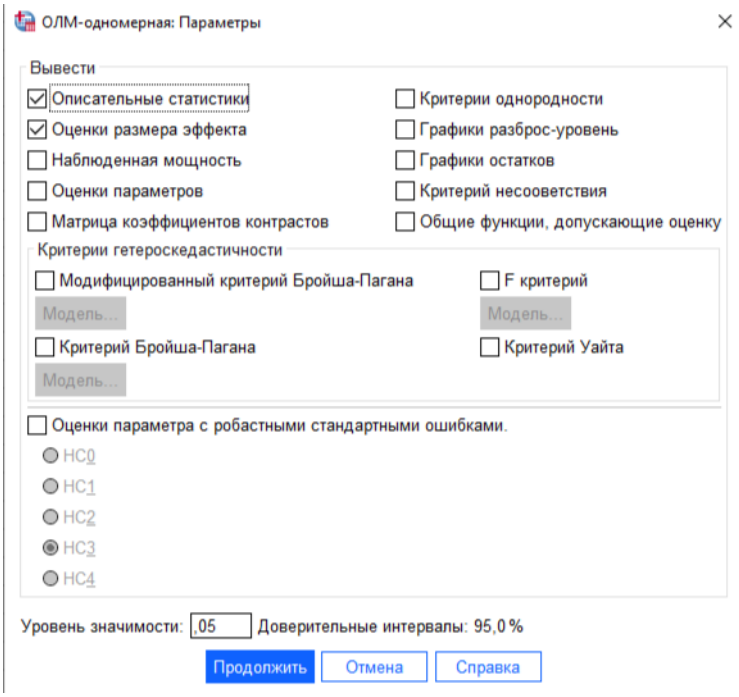



Рисунок 6.13 – Окно диалога «ОЛМ-одномерная: Параметры»
с установленными настройками

В результате выполненных действий программа выведет таблицы, содержание основных из которых представлено в таблицах 6.18 и 6.19.

Таблица 6.18 – Результаты расчетов описательных статистик, выполненных с помощью процедуры «ОЛМ-одномерная»

| Вариант оформления и размещения POS-материалов | Вариант телевизионной рекламы | Среднее значение оценки степени готовности приобрести продукцию | Стандартное отклонение оценки степени готовности приобрести продукцию | Количество наблюдений |
|---|---|---|---|---|
| 1 | 1 | 2,40 | 1,14 | 5 |
| | 2 | 3,40 | 1,14 | 5 |
| | 3 | 4,60 | 1,14 | 5 |
| | Всего | 3,47 | 1,41 | 15 |
| 2 | 1 | 3,20 | 0,84 | 5 |
| | 2 | 6,00 | 0,71 | 5 |
| | 3 | 8,80 | 0,84 | 5 |
| | Всего | 6,00 | 2,48 | 15 |
| 3 | 1 | 5,80 | 0,84 | 5 |
| | 2 | 5,40 | 1,14 | 5 |
| | 3 | 6,40 | 1,14 | 5 |
| | Всего | 5,87 | 1,06 | 15 |
| Всего | 1 | 3,80 | 1,74 | 15 |
| | 2 | 4,93 | 1,49 | 15 |
| | 3 | 6,60 | 2,03 | 15 |
| | Всего | 5,11 | 2,08 | 45 |

Таблица 6.19 – Результаты двухфакторного дисперсионного анализа оценок степени готовности приобрести продукцию компании, выставленных покупателями после ознакомления с предложенными им комбинациями вариантов POS-материалов и телерекламы

| Источник | Сумма квадратов типа III, *SS* | Степени свободы, *df* | Средний квадрат, *MS* | Расчетное значение *F*-критерия | *P*-значение | Частная $\eta^2$ |
|---|---|---|---|---|---|---|
| Модель | 1329,60 | 9 | 147,73 | 146,11 | 0,00 | 0,97 |
| Вариант POSM | 60,98 | 2 | 30,49 | 30,15 | 0,00 | 0,63 |
| Вариант рекламы | 59,51 | 2 | 29,76 | 29,43 | 0,00 | 0,62 |
| Вариант POSM и вариант рекламы | 33,56 | 4 | 8,39 | 8,30 | 0,00 | 0,48 |
| Ошибка | 36,40 | 36 | 1,01 | | | |
| Всего | 1366,00 | 45 | | | | |

Значение корреляционного отношения, рассчитанное с использованием данных таблицы 6.19 по формуле (6.20), полностью совпадает со значением, рассчитанным с использованием данных таблицы 6.14:

$$\eta^2 = \frac{60{,}98 + 59{,}51 + 33{,}56}{60{,}98 + 59{,}51 + 33{,}56 + 36{,}40} = 0{,}81.$$

Как видно из таблицы 6.18, наиболее отличимой из всех пар комбинаций является пара, образуемая вторым вариантом оформления и размещения POS-материалов и третьим вариантом телерекламы. Именно для нее наибольшим является среднее значение оценки степени готовности приобрести продукцию.

Из нее также можно установить, что разности средних оценок превышают предельную ошибку в размере 0,22 для всех пар комбинаций вариантов рассматриваемых факторов, кроме пар, образованных:

– вторым вариантом POS-материалов в сочетании с первым вариантом телерекламы и первым вариантом POS-материалов в сочетании со вторым вариантом телерекламы;

– третьим вариантом POS-материалов в сочетании с первым вариантом телерекламы и вторым вариантом POS-материалов в сочетании со вторым вариантом телерекламы.

Следовательно, можно сделать вывод о существенном различии уровня влияния на степень готовности покупки продукции пяти из девяти комбинаций вариантов оформления и размещения POS-материалов и телерекламы.

## 6.4 Задание для самостоятельного выполнения

Самостоятельно в приложении MS Excel создать файл, в котором на двух листах представить результаты использования мер по стимулированию покупателей и продвижению продукции компании, которая была рассмотрена в ходе выполнения самостоятельных заданий лабораторных работ № 1–5. Группы, состоящие из покупателей и (или) организаций розничной торговли, должны включать не менее 30 респондентов (субъектов). Оценку эффективности применения использованных мер выполнить с использованием одно- и двухфакторного дисперсионных анализов.

## 6.5 Вопросы для самоконтроля

1 Каков порядок проведения однофакторного дисперсионного анализа?

2 На основе значения какого показателя устанавливается эффект влияния независимой переменной на зависимую при проведении однофакторного дисперсионного анализа?

3 Какие статистические гипотезы формулируются в рамках однофакторного дисперсионного анализа и с помощью какой статистики они проверяются?

4 Каков порядок проведения двухфакторного дисперсионного анализа?

5 На основе значения какого показателя устанавливается эффект влияния независимых переменных на зависимую при проведении двухфакторного дисперсионного анализа?

6 Какие статистические гипотезы формулируются в рамках двухфакторного дисперсионного анализа и с помощью какой статистики они проверяются?

7 В каком порядке и с использованием каких статистик, в случае когда будет установлено существенное влияние факторов на изменчивость признака, исследуется значимость его средних на отдельных уровнях?

# ЛАБОРАТОРНАЯ РАБОТА № 7

## Парный (однофакторный) корреляционно-регрессионный анализ данных, полученных по выборке в процессе маркетингового исследования

**Цель работы:** выполнить парный (однофакторный) корреляционно-регрессионный анализ данных, характеризующих объекты исследования (домохозяйства), которые были включены в выборку, сформированную по итогам лабораторной работы № 5.

**Задачи работы:** с использованием теоретических знаний, полученных при изучении тем № 8, 10, 11, 13, 14 и 16 дисциплины, а также изученных ранее курсов «Прикладной статистический анализ» и «Теория вероятностей и математическая статистика»:

– изучить порядок выполнения парного (однофакторного) корреляционно-регрессионного анализа данных, полученных по выборке в процессе маркетингового исследования;

– получить практические навыки в выполнении парного (однофакторного) корреляционно-регрессионного анализа данных с использованием приложения MS Excel и программы IBM SPSS Statistics.

### 7.1 Теоретические сведения

#### 7.1.1 Основные термины

**Причинно-следственные отношения** – это связь явлений и процессов, когда изменение одного из них (причины) ведет к изменению другого (следствия).

**Причина** – это совокупность условий, обстоятельств, действие которых приводит к появлению следствия.

**Признак** – это основная отличительная черта, особенность изучаемого явления или процесса.

**Факторный признак** – это признак, обуславливающий изменения другого, связанного с ним результативного признака.

**Результативный признак** – это признак, изменяющийся под действием факторных признаков.

**Функциональная связь** – это связь, при которой определенному значению факторного признака соответствует одно и только одно значение результативного признака.

**Стохастическая связь** – это связь, которая проявляется не в каждом отдельном случае, а в общем, среднем или большем числе наблюдений.

**Корреляционная связь** – это связь, при которой изменение среднего значения результативного признака обуславливается изменением факторных признаков.

**Корреляция** – это статистическая зависимость между случайными величинами, которая не имеет строго функционального характера и при которой изменение одной из случайных величин приводит к изменению математического ожидания другой.

**Регрессионный анализ** – это аналитическое выражение связи, в которой изменение одной величины – результативного признака – обусловлено влиянием одного или нескольких факторных признаков, а множество всех остальных факторных признаков, также оказывающих влияние на зависимую величину, принимается за постоянные и средние значения.

**Автокорреляция** – это статистическая взаимосвязь между последовательными уровнями временного ряда. Может быть как положительной, так и отрицательной. Свидетельствует о постоянном действии неучтенных факторов на результат: положительная – об однонаправленном, отрицательная – о разнонаправленном.

**Критерий Стьюдента (*t*-статистика, *t*-тест)** – это общее название класса методов статистической проверки гипотез и статистических критериев, основанных на распределении Стьюдента.

**Тест Дарбина – Уотсона** – это тест на автокорреляцию, которая выражается в наличии систематических связей между остатками, которые представляют собой отклонения наблюдаемых значений от теоретически ожидаемых и вычисленных с помощью уравнения регрессии.

### 7.1.2 Парный (однофакторный) корреляционно-регрессионный анализ

#### 7.1.2.1 Парный (однофакторный) корреляционный анализ

Задачей парного корреляционного анализа является количественное определение тесноты связи между двумя признаками. Теснота связи количественно выражается величиной коэффициента корреляции, которая дает возможность определить «полезность» факторного признака при построении уравнения парной регрессии. Величина коэффициента корреляции служит также оценкой соответствия уравнения регрессии, выявленной в ходе описательного (дескриптивного) маркетингового исследования причинно-следственной связи.

Коэффициент корреляции (коэффициент корреляции Пирсона, простой (линейный) коэффициент корреляции) вычисляется по формулам:

$$r_{xy} = \hat{\rho}_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 (y_i - \overline{y})^2}} \quad \text{или} \quad r_{xy} = \hat{\rho}_{xy} = \frac{S_{xy}}{s_x s_y}, \qquad (7.1)$$

где $x_i$ – $i$-е значение факторного признака в выборке;

$y_i$ – значение результативного признака, соответствующее $i$-му значению факторного признака;

$\overline{x}$ – среднее значение факторного признака в выборке;

$\overline{y}$ – среднее значение результативного признака в выборке;
$S_{xy}$ – коэффициент ковариации между результативным и факторным признаками;
$s_x$ – стандартное отклонение факторного признака в выборке;
$s_y$ – стандартное отклонение результативного признака в выборке;
$n$ – число наблюдений в выборке.

По направлению выделяют связь прямую (положительную), при которой увеличение (уменьшение) значения факторного признака приводит к увеличению (уменьшению) значения результативного, и обратную (отрицательную), при которой наоборот увеличение (уменьшение) значения факторного признака приводит к уменьшению (увеличению) значения результативного.

По аналитическому выражению выделяют связь прямолинейную (линейную), когда статистическая связь может быть приближенно выражена уравнением прямой линии, и нелинейную (криволинейную), когда статистическая связь выражается уравнением какой-либо кривой линии (параболы, гиперболы, степенной, показательной, экспоненциальной и т. п.).

Для оценки степени тесноты связи используют значения коэффициента корреляции, приведенные в таблице 7.1.

Таблица 7.1 – Количественные критерии оценки тесноты корреляционной связи

| Величина коэффициента корреляции | Характер корреляционной связи |
|---|---|
| 0 – \| ±0,3 \| | Практически отсутствующая |
| \| ±0,3 \| – \| ±0,5 \| | Слабая |
| \| ±0,5 \| – \| ±0,7 \| | Умеренная |
| \| ±0,7 \| – \| ±1,0 \| | Сильная |

Статистическая значимость линейного коэффициента корреляции проверяется с использованием критерия Стьюдента:

$$t_{\text{расч}} = \frac{|r|}{\sqrt{1-r^2}}\sqrt{n-2}\,. \qquad (7.2)$$

Если расчетное значение критерия больше критического $t_{\text{расч}} > t_{\text{крит}}$, то нулевая гипотеза $H_0: r_{yx} = 0$ отвергается, что свидетельствует о статистической значимости линейного коэффициента корреляции и, следовательно, о статистической существенной зависимости между $x$ и $y$.

Коэффициент корреляции, возведенный в квадрат, называется коэффициентом детерминации. Он показывает, какая доля вариации зависимой переменной обусловлена вариацией независимой и определяется по формуле

$$d_{xy} = r_{xy}^2 \quad \text{или} \quad d_{xy} = \frac{SS_x}{SS_y}\,. \qquad (7.3)$$

При этом

$$SS_y = \sum_{i=1}^{n} (y_i - \overline{y})^2, \tag{7.4}$$

$$SS_x = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2, \tag{7.5}$$

где $SS_y$ – полная вариация результативного признака как сумма квадратов разниц между фактическими значениями зависимой переменной и ее средним значением в выборке;
$SS_x$ – полная вариация теоретических значений результативного признака, рассчитанная по уравнению регрессии как сумма квадратов разниц между расчетными и фактическими значениями результативного признака.

Проверка значимости коэффициента детерминации является еще одной равноценной проверкой значимости линейной зависимости между $x$ и $y$ (значимости коэффициента при независимой переменной $\beta_1$ в уравнении регрессии). В этом случае статистические гипотезы имеют вид

$$\begin{aligned} H_0&: d_{xy} = 0, \\ H_1&: d_{xy} \neq 0. \end{aligned} \tag{7.6}$$

Соответствующей статистикой, лежащей в основе проверки, является $F$-критерий, который представляет собой обобщенную форму критерия Стьюдента:

$$F = \frac{SS_y}{SS_y - SS_x}(n-2). \tag{7.7}$$

Как и в случае с критерием Стьюдента, если $F_{\text{расч}} > F_{\text{крит}}$, нулевая гипотеза $H_0: d_{xy} = 0$ отвергается, и наоборот, если $F_{\text{расч}} < F_{\text{крит}}$, нулевая гипотеза принимается.

#### 7.1.2.2 Парный (однофакторный) регрессионный анализ

Примерный порядок выполнения парного (однофакторного) регрессионного анализа показан на рисунке 7.1.

##### 7.1.2.2.1 Поле корреляции

Поле корреляции (рисунок 7.2) представляет собой графическое изображение точек с координатами, соответствующими значениям двух переменных для всех случаев. Обычно значения зависимой переменной откладывают по вертикальной оси, а значения независимой – по горизонтальной. Поле корреляции используется при определении формы зависимости между переменными. График

дает исследователю первое представление о полученных по итогам исследования данных и о возможных проблемах при проведении их анализа.


Рисунок 7.1 – Примерный порядок выполнения парного (однофакторного) регрессионного анализа

#### 7.1.2.2.2 Общая модель парной регрессии

В модели парной регрессии форма прямой линии выражается уравнением

$$y = \beta_0 + \beta_1 x \,, \tag{7.8}$$

где $y$ – зависимая (критериальная) переменная;
$x$ – независимая переменная (фактор, предиктор);
$\beta_0$ – отрезок прямой, отсекаемой по оси $y$;
$\beta_1$ – угловой коэффициент (тангенс угла наклона).

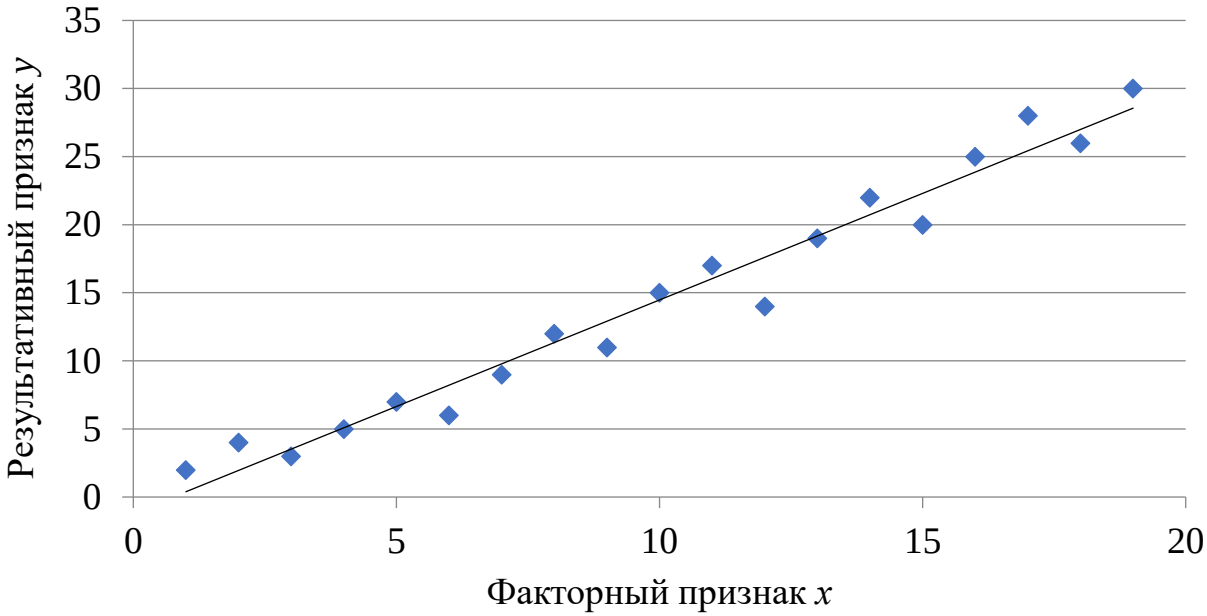



Рисунок 7.2 – Корреляционное поле

Эта модель исходит из того, что $y$ полностью определяется $x$. При известных значениях $\beta_0$ и $\beta_1$ можно предсказать значение $y$. Однако в маркетинговом исследовании немного связей между переменными четко детерминированы. Поэтому, для того чтобы учесть вероятностную природу связи, в регрессионное уравнение вводят ошибочный член. Базовое уравнение парной регрессии в этом случае принимает вид

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \,, \tag{7.9}$$

где $\varepsilon_i$ – член уравнения, характеризующий ошибку $i$-го наблюдения.

В большинстве случаев $\beta_0$ и $\beta_1$ неизвестны и их определяют (оценивают) исходя из имеющихся наблюдений по выборке с помощью следующего уравнения:

$$\hat{y}_i = \beta_0 + \beta_1 x_i \,, \tag{7.10}$$

где $\hat{y}_i$ – расчетное (теоретическое) значение $y_i$.

#### 7.1.2.2.3 Параметры уравнения парной регрессии

В уравнениях регрессии параметр $\beta_0$ соответствует усредненному влиянию на результативный признак неучтенных (не выделенных для исследования) факторов, параметр $\beta_1$ (а в уравнении, например, параболы и $\beta_2$) – коэффициент, который показывает, насколько изменится в среднем значение результативного признака при увеличении факторного признака на единицу собственного измерения. Значение углового коэффициента $\beta_1$ может быть вычислено через ковариацию между $x$ и $y$ и дисперсию $x$ по формуле

$$\beta_1 = \frac{S_{xy}}{s_x^2}. \quad (7.11)$$

А отрезок $\beta_0$, отсекаемый на оси ординат *y*, можно вычислить по формуле

$$\beta_0 = \overline{y} - \beta_1 \overline{x}\,. \quad (7.12)$$

Однако чаще всего оценка параметров уравнения парной регрессии осуществляется методом наименьших квадратов, в основе которого лежит предположение о независимости наблюдений исследуемой совокупности. Системы нормальных уравнений для нахождения параметров соответственно линейной и нелинейной парной регрессии имеют вид

$$\begin{cases} \beta_0 n + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}; \quad (7.13)$$

$$\begin{cases} \beta_0 n + \beta_1 \sum\limits_{i=1}^{n} x_i + \beta_2 \sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} y_i \\ \beta_0 \sum\limits_{i=1}^{n} x_i + \beta_1 \sum\limits_{i=1}^{n} x_i^2 + \beta_2 \sum\limits_{i=1}^{n} x_i^3 = \sum\limits_{i=1}^{n} x_i y_i \\ \beta_0 \sum\limits_{i=1}^{n} x_i^2 + \beta_1 \sum\limits_{i=1}^{n} x_i^3 + \beta_2 \sum\limits_{i=1}^{n} x_i^4 = \sum\limits_{i=1}^{n} x_i^2 y_i \end{cases}. \quad (7.14)$$

Сущность этого метода заключается в нахождении параметров модели парной регрессии, при которых минимизируется сумма квадратов отклонений фактических (эмпирических) значений результативного признака от расчетных (теоретических), полученных по выбранному уравнению регрессии:

$$SS_e = e^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_{x_i}\right)^2 \to \min. \quad (7.15)$$

### 7.1.2.2.4 Нормирование параметров уравнения парной регрессии

Нормирование представляет собой процедуру, посредством которой исходные данные преобразуют в новые переменные со значением средней, равным нулю, и дисперсией, равной 1,0. После нормирования данных, отрезок, отсекаемый на оси ординат *y*, принимает значение «ноль» (т. е. $\beta_0 = 0$). Нормированный коэффициент регрессии обозначают как «бета»-коэффициент или взвешенный «бета»-коэффициент. В этом случае угловой коэффициент регрессии *y* по *x*, обозначаемый $B_{xy}$, тот же, что и угловой коэффициент регрессии *x* по *y*, обозначаемый $B_{yx}$. Более того, каждый из этих коэффициентов регрессии равен простому (линейному) коэффициенту корреляции между *x* и *y*:

$$\mathrm{B}_{yx} = \mathrm{B}_{xy} = r_{xy}\ . \tag{7.16}$$

Существует простая связь между нормированным и ненормированным коэффициентами регрессии:

$$\mathrm{B}_{yx} = \beta_1 \left(\frac{S_x}{S_y}\right). \tag{7.17}$$

#### 7.1.2.2.5 Проверка значимости связи между зависимой и независимой переменными

Статистическую значимость линейной связи между *x* и *y* можно проверить, исследовав гипотезы:

$$\begin{aligned} H_0&: \beta_1 = 0, \\ H_1&: \beta_1 \neq 0. \end{aligned} \tag{7.18}$$

Нулевая гипотеза предполагает, что между *x* и *y* не существует линейной зависимости. Альтернативная гипотеза утверждает, что между *x* и *y* существует зависимость, либо положительная, либо отрицательная. Обычно проводят двустороннюю проверку. При этом можно использовать *t*-статистику с $(n-2)$ степенями свободы, где

$$t_{\text{расч.}} = \frac{\beta}{s_\beta}. \tag{7.19}$$

При этом

$$S_\beta = \frac{s_e}{s_x\sqrt{n-1}}, \tag{7.20}$$

$$s_e = s_y\sqrt{\left(1 - r_{xy}^2\right)\frac{n-1}{n-2}}. \tag{7.21}$$

#### 7.1.2.2.6 Оценка точности уравнения парной регрессии

Для оценки точности предсказанных (расчетных, теоретических) значений *y*, необходимо вычислить стандартную ошибку оценки уравнения регрессии *SEE*. Этот показатель (статистика) представляет собой стандартное отклонение фактических значений $y_i$ от рассчитанных значений $\hat{y}_i$:

$$SEE = \sqrt{\frac{SS_e}{n-2}}. \tag{7.22}$$

*SEE* можно интерпретировать как вид среднего значения остатка или среднюю ошибку предсказания *y* исходя из уравнения регрессии.

## 7.2 Выполнение парного (однофакторного) корреляционно-регрессионного анализа с использованием приложения MS Excel и программы IBM SPSS Statistics

В файле «05 Результаты выборочного наблюдения.xlsx» приведены данные о домохозяйствах, которые были включены в выборку, исследуемую совместно сотрудниками маркетинговых подразделений ОАО «Крессида» и ЧУП «Кэтнес».

Необходимо с использованием данных, полученных по исследуемой выборке, выполнить парный корреляционно-регрессионный анализ для всех выбранных переменных. При этом исходить из того, что переменная «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» является зависимой, а все остальные 14 – независимыми.

### 7.2.1 Выполнение парного (однофакторного) корреляционно-регрессионного анализа с использованием приложения MS Excel

1 Файл «05 Результаты выборочного наблюдения.xlsx», с помощью которого выполнялась лабораторная работа № 5, скопировать в папку, в которой будут находиться файлы текущей работы, и присвоить ему имя «07 Парный корреляционно-регрессионный анализ.xlsx». В созданном файле листы «Основа выборочного наблюдения» и «Среднегодовые расходы» рекомендуется удалить.

2 Выполнить парный корреляционно-регрессионный анализ между первой парой переменных, в которой переменная «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» является зависимой, а переменная «Количество членов домохозяйства» – независимой.

Это предполагает выдвижение следующих статистических гипотез:

– нулевой: зависимости между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и количеством членов домохозяйств не существует, а если она и существует, то является статистически незначимой ($H_0: r^2 = 0;\ d^2 = 0;\ \beta_1 = 0$);

– альтернативной: зависимость между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и количеством членов домохозяйств существует, она является не случайной и статистически значимой ($H_1: r^2 \neq 0;\ d^2 \neq 0;\ \beta_1 \neq 0$).

3 В созданном файле выполнить следующие действия (рисунок 7.3):

– выбрать инструмент корреляционно-регрессионного анализа **(«Данные»** – **«Анализ данных»** – **«Регрессия»**);

– ввести в поле **«Входной интервал Y»** значения примерных среднегодовых расходов на покупку (обновление) элементов домашней мебели по всей выборке (ячейки «P2» – «P1442» из листа «Выборка»);

– ввести в поле **«Входной интервал X»** значения количества членов домохозяйств по всей выборке (ячейки «B2» – «B1442» из листа «Выборка»);

– поставить флажок напротив строки **«Метки в первой строке»**, уровень надежности оставить равным 95 %;

– в параметрах вывода выбрать поле **«Новый рабочий лист»**;

– поставить флажок напротив строки **«Остатки»** и нажать кнопку **«ОК»**;

– созданный лист переместить правее листа «Выборка» и присвоить ему имя «Расходы и кол-во чл.дом.хоз»;

– в созданном листе для всех таблиц задать шрифт Times New Roman Cyr размером 12 пт и изменить ширину колонок так, чтобы названия всех параметров и величин были полностью видны. В случае, если в третьей таблице колонки **«Нижние 95%»** и **«Верхние 95 %»** повторились, их можно удалить.

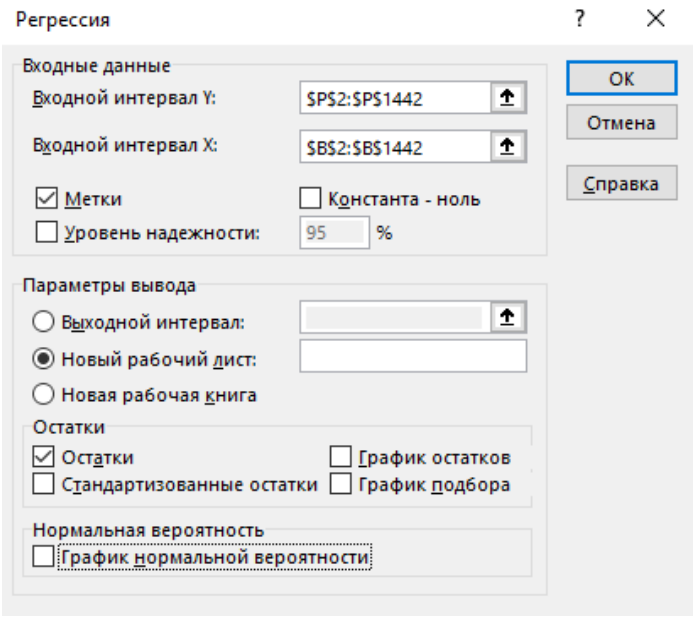



Рисунок 7.3 – Диалог инструмента анализа «Регрессия» с внесенными значениями переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Количество членов домохозяйства»

4 В результате выполнения корреляционно-регрессионного анализа будут получены значения, которые приведены в таблицах 7.2–7.4. Таблица с рассчитанными разницами (остатками) между вычисленными с использованием уравнения регрессии значениями примерных среднегодовых расходов на покупку (обновление) элементов домашней мебели и эмпирическими (фактическими по выборке) значениями по этой переменной здесь не приводится по причине большого количества в ней строк (1440).

Таблица 7.2 – Показатели регрессионной статистики для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Количество членов домохозяйства»

| Показатель | Значение |
|---|---|
| Парный коэффициент корреляции | 0,88 |
| Коэффициент детерминации | 0,78 |
| Нормированный коэффициент детерминации | 0,78 |
| Стандартная ошибка | 59,14 |
| Количество наблюдений | 1440 |

Таблица 7.3 – Результаты однофакторного дисперсионного анализа для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Количество членов домохозяйства»

| Показатель | Число степеней свободы, *df* | Сумма квадратов разностей, *SS* | Оценка дисперсий (средний квадрат), *MS* | Значение *F*-критерия | Значимость *F*-критерия |
|---|---|---|---|---|---|
| Регрессия | 1 | 18166118,08 | 18166118,08 | 5194,79 | 0,00 |
| Остаток | 1438 | 5028671,92 | 3496,99 | | |
| Итого | 1439 | 23194790,00 | | | |

Таблица 7.4 – Значения коэффициентов уравнения регрессии, оценок их статистической значимости и границ доверительных интервалов для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Количество членов домохозяйства»

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Свободный член, $\beta_0$ | 2,67 | 4,89 | 0,55 | 0,58 | –6,91 | 12,26 |
| Коэффициент при независимой переменной, $\beta_1$ | 83,72 | 1,16 | 72,07 | 0,00 | 81,44 | 86,00 |

Значения, приведенные в таблицах 7.2–7.4, позволяют заключить следующее:

– на основании величин парного коэффициента корреляции, значимости *F*-критерия (указывающего на то, что его расчетное значение превышает критическое) и *P*-значения для коэффициента при независимой переменной $\beta_1$ (указывающего на то, что расчетное значение *t*-критерия Стьюдента для него превышает критическое) следует принять альтернативную гипотезу;

– зависимость между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и количеством членов домохозяйств существует и является статистически значимой. Значение парного коэффициента корреляции, равное примерно 0,88, указывает на то, что связь по своему характеру является сильной, а знак коэффициента при независимой переменной – положительной. Значение коэффициента детерминации говорит о том, что примерно 78 % вариации зависимой переменной «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» объясняется вариацией независимой переменной «Количество членов домохозяйства». Остальные 22 % примерных среднегодовых расходов на покупку и (обновление) элементов домашней мебели зависят от других, не учтенных в данном случае переменных;

– уравнение регрессии имеет вид

$$y_x = 2{,}67 + 83{,}72x.$$

Из него видно, что увеличение значения независимой переменной «Количество членов домохозяйства» на одну единицу влечет увеличение значения зависимой переменной «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» примерно на 83 р. и 72 коп. и наоборот. Границы доверительного интервала для независимой переменной интерпретируются следующим образом: можно быть на 95 % уверенным в том, что примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели при изменении количества членов в домохозяйстве на одного человека будут находиться в пределах между 81,44 и 86,00 р. Что касается свободного члена, то его величина статистически незначима, о чем говорят не только значение критерия Стьюдента, превышающее критическое (что подтверждает *P*-значение большее 0,05), но и границы его доверительного интервала, внутри которого находится значение «ноль»;

– стандартная ошибка оценки уравнения регрессии *SEE* равна 59,14 р.

5 На листе «Расходы и кол-во чл.дом.хоз.» рассчитать описательные статистики для установленных разниц (остатков) между вычисленными на основе уравнения регрессии и эмпирическими (фактическими) по выборке значениями переменной «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели». Для этого, используя навыки, полученные при выполнении лабораторной работы № 5 (рисунок 7.4):

– выбрать инструмент анализа **«Описательная статистика» («Данные»** – **«Анализ данных»** – **«Описательная статистика»**);

– ввести в поле **«Входной интервал»** значения остатков (ячейки **«С24»** – **«С1464»**), группирование выбрать по столбцам. Так как в этом поле будет находиться название изучаемой характеристики, поставить флажок напротив строки **«Метки в первой строке»**;

– в поле **«Параметры вывода»** выбрать **«Выходной интервал»** и с использованием мыши ввести в поле ссылку на ячейку **«К1»**;

– поставить флажок напротив строки **«Итоговая статистика»**, уровень надежности оставить равным 95 % и нажать кнопку **«ОК»**;

– используя шрифт Times New Roman Cyr размером 12 пт, отформатировать полученные данные;

– вычисленные программой значения представлены в таблице 7.5.

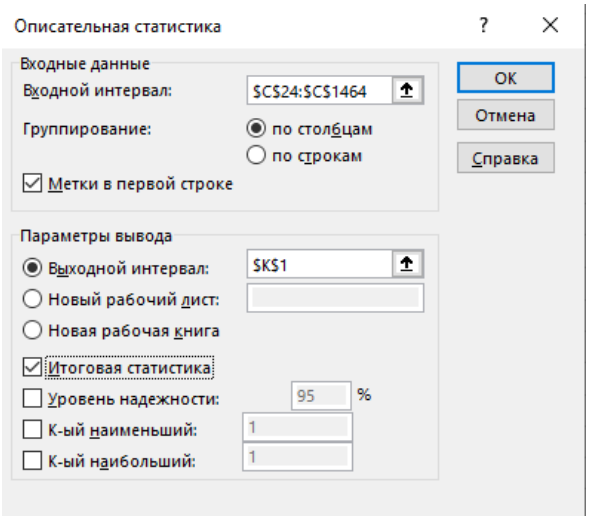



Рисунок 7.4 – Диалог инструмента анализа «Описательная статистика» с внесенными значениями остатков уравнения регрессии для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Количество членов домохозяйства»

Таблица 7.5 – Значения описательных статистик остатков уравнения регрессии для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Количество членов домохозяйства»

| Статистики | Значения |
|---|---|
| Среднее | 0,00 |
| Стандартная ошибка | 1,56 |
| Медиана | –1,11 |
| Мода | 3,17 |
| Стандартное отклонение | 59,11 |
| Дисперсия выборки | 3494,56 |
| Эксцесс | 0,47 |
| Асимметричность | 0,36 |
| Интервал | 346,28 |
| Минимум | –161,27 |
| Максимум | 185,01 |
| Сумма | 0,00 |
| Счет | 1440 |

Значение статистики «асимметричность» (независимо от знака), которое меньше, чем 0,5, позволяет сделать предположение о незначительной асимметрии гистограммы. Так как оценка существенности асимметрии зависит только от размера выборки, ее уже рассчитанную величину взять из решения задачи в лабораторной работе № 5:

$$S_{A_S} = \sqrt{\frac{6(1440-1)}{(1440+1)(1440+3)}} = 0{,}06.$$

Так как сейчас отношение значения асимметрии к ее средней квадратичной ошибке $\frac{|A_S|}{S_{A_S}} = \frac{0{,}36}{0{,}064} = 5{,}56$, что больше 3,0, асимметрию следует признать существенной, а распределение остатков не следует считать симметричным.

Положительное значение статистики «эксцесс» говорит об островершинности рассматриваемого ряда значений. Из той же лабораторной работы № 5 взять значение средней квадратичной ошибки показателя эксцесса, чтобы оценить его существенность:

$$S_{\varepsilon_k} = \sqrt{\frac{24 \cdot 1440(1440-2)(1440-3)}{(1440+1)^2(1440+3)(1440+5)}} = 0{,}13\,.$$

На основе значения $\frac{|\varepsilon_k|}{S_{\varepsilon_k}} = \frac{0{,}47}{0{,}13} = 3{,}64$, что больше 3,0, следует признать, что отклонение распределения остатков от нормального существенно и его нельзя считать нормальным или близким к нему.

6 С использованием полученных значений построить гистограмму для установленных остатков. Для этого:

– на ранее созданном листе «Расходы и кол-во чл.дом.хоз.» справа от таблицы с вычисленными описательными статистиками с учетом того, что минимальное значение остатков равно –161,27, а максимальное – 185,01, создать 20 интервалов группирования (карманов) с шагом в 20 р. Номера карманов с первого по двенадцатый разместить в ячейках **«N2»–«N21»**, а значения их правых границ – в ячейках **«O2»–«O21»**;

– выбрать инструмент анализа **«Гистограмма»** (**«Данные»** – **«Анализ данных»** – **«Гистограмма»**) и заполнить его диалоговое окно, внеся во входной интервал значения остатков (ячейки **«C24»–«C1464»**), а ниже – интервал карманов (ячейки **«O1»–«O21»**). Поставив флажок напротив строки **«Метки в первой строке»**, указав в качестве выходного интервала ячейку **«Q1»** и поставив флажок напротив строки **«Вывод графика»**, нажать кнопку **«ОК»**;

– используя шрифт Times New Roman Cyr размером 12 пт, отформатировать появившуюся таблицу и удалить в ней строку **«Еще»**;

– используя этот же шрифт, отформатировать построенную гистограмму;

– итог выполненных действий представлен на рисунке 7.5.

Как видно из гистограммы, распределение остатков не подпадает под нормальное. Так как свойства коэффициентов регрессии существенным образом

зависят от свойств остатков, а их распределение должно быть не только независимо от распределения переменных, но и нормально, то рассчитанное уравнение регрессии следует признать условно пригодным для прогнозирования расходов домохозяйств на приобретение (обновление) элементов домашней мебели в зависимости от количества их членов.

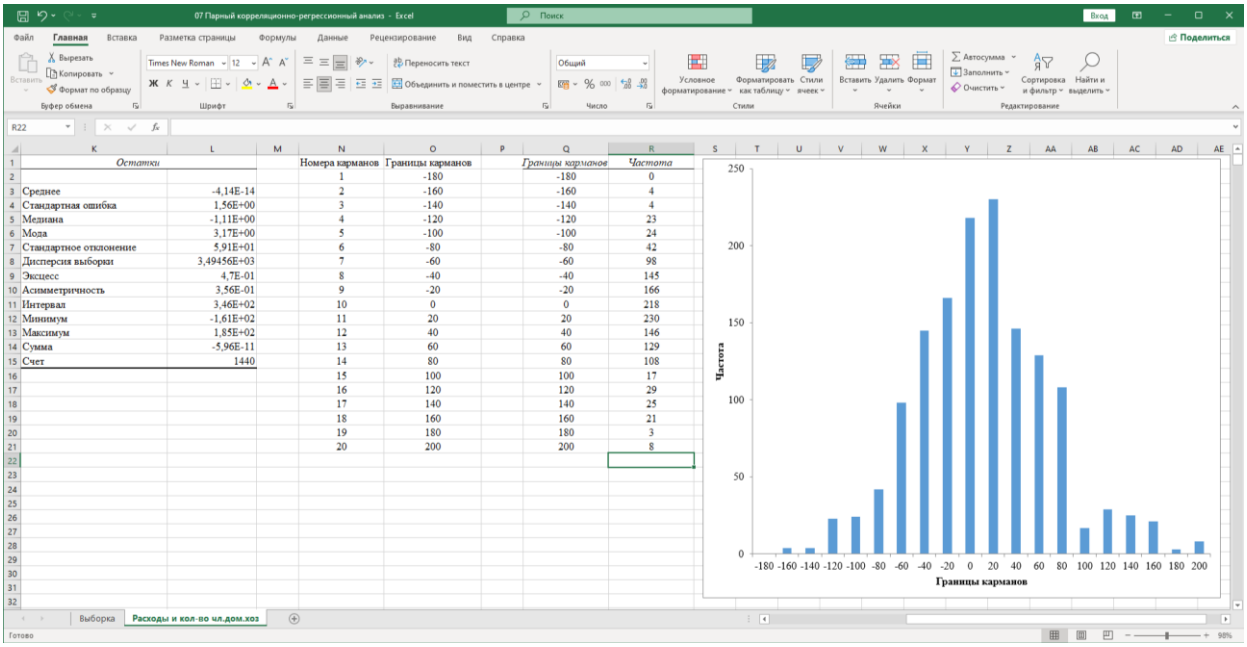



Рисунок 7.5 – Рассчитанные статистики и гистограмма остатков уравнения регрессии для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Количество членов домохозяйства»

7 Выполнить подобным образом парный корреляционно-регрессионный анализ для остальных пар переменных, в которых переменная «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» по-прежнему является зависимой, а все остальные переменные – независимыми. Для каждого анализа создать отдельные листы и дать им имена, соответствующие анализируемым парам переменных.

Выполнение работы каждый раз предполагает выдвижение следующих статистических гипотез:

– нулевой: зависимости между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и рассматриваемой независимой переменной не существует, а если она и существует, то является статистические незначимой ($H_0$: $r^2 = 0$; $d^2 = 0$; $\beta_1 = 0$);

– альтернативной: зависимость между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и рассматриваемой независимой переменной существует, она является неслучайной и статистически значимой ($H_1$: $r^2 \neq 0$; $d^2 \neq 0$; $\beta_1 \neq 0$).

Результаты выполнения анализов для каждой пары представлены ниже.

8 Значения для пар переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Возраст мужа» приведены в таблицах 7.6–7.9.

Таблица 7.6 – Показатели регрессионной статистики для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Возраст мужа»

| Показатель | Значение |
|---|---|
| Парный коэффициент корреляции | 0,31 |
| Коэффициент детерминации | 0,09 |
| Нормированный коэффициент детерминации | 0,09 |
| Стандартная ошибка | 120,92 |
| Количество наблюдений | 1440 |

Таблица 7.7 – Результаты однофакторного дисперсионного анализа для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Возраст мужа»

| Показатель | Число степеней свободы, *df* | Сумма квадратов разностей, *SS* | Оценка дисперсий (средний квадрат), *MS* | Значение *F*-критерия | Значимость *F*-критерия |
|---|---|---|---|---|---|
| Регрессия | 1 | 2170201,60 | 2170201,60 | 148,43 | 0,00 |
| Остаток | 1438 | 21024588,40 | 14620,72 | | |
| Итого | 1439 | 23194790,00 | | | |

Таблица 7.8 – Значения коэффициентов уравнения регрессии, оценок их статистической значимости и границ доверительных интервалов для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Возраст мужа»

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Свободный член, $\beta_0$ | 473,42 | 11,67 | 40,56 | 0,00 | 450,52 | 496,31 |
| Коэффициент при независимой переменной, $\beta_1$ | –2,98 | 0,24 | –12,18 | 0,00 | –3,46 | –2,50 |

Таблица 7.9 – Значения описательных статистик остатков уравнения регрессии для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Возраст мужа»

| Статистики | Значения |
|---|---|
| Среднее | 0,00 |
| Стандартная ошибка | 3,19 |
| Медиана | –9,21 |
| Мода | –103,24 |
| Стандартное отклонение | 120,87 |
| Дисперсия выборки | 14610,55 |
| Эксцесс | –0,72 |
| Асимметричность | 0,21 |
| Интервал | 505 |
| Минимум | –219,92 |
| Максимум | 285,08 |
| Сумма | 0,00 |
| Счет | 1440 |

Значения, приведенные в таблицах 7.6–7.9, позволяют заключить следующее:

– следует принять альтернативную гипотезу: корреляционная связь между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и возрастом мужа существует, является статистически значимой, отрицательной, но по силе – практически отсутствующей;

– примерно 9 % вариации зависимой переменной объясняется вариацией независимой переменной, а остальные 91 % зависят от других, не учтенных в данном случае переменных;

– уравнение регрессии имеет вид

$$y_x = 473{,}42 - 2{,}98x,$$

т. е. увеличение значения независимой переменной на одну единицу влечет уменьшение значения зависимой переменной примерно на 2 р. и 98 коп. и наоборот, величина его свободного члена статистически значима;

– можно быть на 95 % уверенным в том, что примерные среднегодовые расходы на покупку и (или) обновление элементов мебели при увеличении возраста мужа на один год будут уменьшаться на сумму в пределах между 2,50 и 3,46 р.;

– стандартная ошибка оценки уравнения регрессии *SEE* равна 120,92 р.;

– так как отношение значения асимметрии распределения остатков к ее средней квадратичной ошибке $\frac{|A_S|}{S_{A_S}} = \frac{0{,}21}{0{,}06} = 3{,}28$, что больше 3,0, асимметрию следует признать существенной, а распределение остатков не следует считать симметричным;

– так как отношение значения эксцесса распределения остатков к его средней квадратичной ошибке $\frac{|\varepsilon_k|}{S_{\varepsilon_k}} = \frac{|-0,72|}{0,13} = 5,60$, что больше 3,0, отклонение распределения остатков от нормального существенно и его нельзя считать нормальным или близким к нему;

– рассчитанное уравнение регрессии следует признать условно пригодным для прогнозирования расходов домохозяйств на приобретение (обновление) элементов домашней мебели в зависимости от возраста мужа.

9 Значения для пар переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Возраст жены» приведены в таблицах 7.10–7.13.

Таблица 7.10 – Показатели регрессионной статистики для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Возраст жены»

| Показатель | Значение |
|---|---|
| Парный коэффициент корреляции | 0,30 |
| Коэффициент детерминации | 0,09 |
| Нормированный коэффициент детерминации | 0,09 |
| Стандартная ошибка | 121,07 |
| Количество наблюдений | 1440 |

Таблица 7.11 – Результаты однофакторного дисперсионного анализа для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Возраст жены»

| Показатель | Число степеней свободы, *df* | Сумма квадратов разностей, *SS* | Оценка дисперсий (средний квадрат), *MS* | Значение *F*-критерия | Значимость *F*-критерия |
|---|---|---|---|---|---|
| Регрессия | 1 | 2115231,46 | 2115231,46 | 144,30 | 0,00 |
| Остаток | 1438 | 21079558,54 | 14658,94 | | |
| Итого | 1439 | 23194790 | | | |

Таблица 7.12 – Значения коэффициентов уравнения регрессии, оценок их статистической значимости и границ доверительных интервалов для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Возраст жены»

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Свободный член, $\beta_0$ | 470,60 | 11,60 | 40,57 | 0,00 | 447,85 | 493,36 |

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Коэффициент при независимой переменной, $\beta_1$ | –3,11 | 0,26 | –12,01 | 0,00 | –3,62 | –2,60 |

Таблица 7.13 – Значения описательных статистик остатков уравнения регрессии для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Возраст жены»

| Статистики | Значения |
|---|---|
| Среднее | 0,00 |
| Стандартная ошибка | 3,19 |
| Медиана | –6,64 |
| Мода | –101,63 |
| Стандартное отклонение | 121,03 |
| Дисперсия выборки | 14648,76 |
| Эксцесс | –0,70 |
| Асимметричность | 0,22 |
| Интервал | 505 |
| Минимум | –217,18 |
| Максимум | 287,82 |
| Сумма | 0,00 |
| Счет | 1440 |

Значения, приведенные в таблицах 7.10–7.13, позволяют заключить следующее:

– следует принять альтернативную гипотезу: корреляционная связь между примерными среднегодовыми расходами на покупку и (или) обновление элементов мебели и возрастом жены существует, является статистически значимой, отрицательной, но по силе – практически отсутствующей;

– примерно 9 % вариации зависимой переменной объясняется вариацией независимой переменной, а остальные 91 % зависят от других, не учтенных в данном случае переменных;

– уравнение регрессии имеет вид

$$y_x = 470{,}60 - 3{,}11x,$$

т. е. увеличение значения независимой переменной на одну единицу влечет уменьшение значения зависимой переменной примерно на 3 р. и 11 коп. и наоборот, величина его свободного члена статистически значима;

– можно быть на 95 % уверенным в том, что примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели при увеличении

возраста жены на один год будут уменьшаться на сумму в пределах между 2,60 и 3,62 р.;

– стандартная ошибка оценки уравнения регрессии *SEE* равна 121,07 р.;

– так как отношение значения асимметрии распределения остатков к ее средней квадратичной ошибке $\frac{|A_S|}{S_{A_S}} = \frac{0,22}{0,06} = 3,24$, что больше 3,0, асимметрию следует признать существенной, а распределение остатков не следует считать симметричным;

– так как отношение значения эксцесса распределения остатков к его средней квадратичной ошибке $\frac{|\varepsilon_k|}{S_{\varepsilon_k}} = \frac{|-0,70|}{0,13} = 5,50$, что больше 3,0, отклонение распределения остатков от нормального существенно и его нельзя считать нормальным или близким к нему;

– рассчитанное уравнение регрессии следует признать условно пригодным для прогнозирования расходов домохозяйств на приобретение (обновление) элементов домашней мебели в зависимости от возраста жены.

10 Значения для пар переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Образование мужа» приведены в таблицах 7.14–7.17.

Таблица 7.14 – Показатели регрессионной статистики для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Образование мужа»

| Показатель | Значение |
|---|---|
| Парный коэффициент корреляции | 0,15 |
| Коэффициент детерминации | 0,02 |
| Нормированный коэффициент детерминации | 0,02 |
| Стандартная ошибка | 125,59 |
| Количество наблюдений | 1440 |

Таблица 7.15 – Результаты однофакторного дисперсионного анализа для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Образование мужа»

| Показатель | Число степеней свободы, *df* | Сумма квадратов разностей, *SS* | Оценка дисперсий (средний квадрат), *MS* | Значение *F*-критерия | Значимость *F*-критерия |
|---|---|---|---|---|---|
| Регрессия | 1 | 513682,61 | 513682,61 | 32,57 | 0,00 |
| Остаток | 1438 | 22681107,39 | 15772,68 | | |
| Итого | 1439 | 23194790,00 | | | |

Таблица 7.16 – Значения коэффициентов уравнения регрессии, оценок их статистической значимости и границ доверительных интервалов для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Образование мужа»

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Свободный член, $\beta_0$ | 256,57 | 14,41 | 17,80 | 0,00 | 228,30 | 284,84 |
| Коэффициент при независимой переменной, $\beta_1$ | 26,52 | 4,65 | 5,71 | 0,00 | 17,40 | 35,63 |

Таблица 7.17 – Значения описательных статистик остатков уравнения регрессии для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Образование мужа»

| Статистики | Значения |
|---|---|
| Среднее | 0,00 |
| Стандартная ошибка | 3,31 |
| Медиана | –13,61 |
| Мода | –79,13 |
| Стандартное отклонение | 125,55 |
| Дисперсия выборки | 15761,71 |
| Эксцесс | –0,50 |
| Асимметричность | 0,47 |
| Интервал | 556 |
| Минимум | –228,64 |
| Максимум | 327,36 |
| Сумма | 0,00 |
| Счет | 1440 |

Значения, приведенные в таблицах 7.14–7.17, позволяют заключить следующее:

– следует принять альтернативную гипотезу: корреляционная связь между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и уровнем образования мужа существует, является статистически значимой, положительной, но по силе – практически отсутствующей;

– примерно 2 % вариации зависимой переменной объясняется вариацией независимой переменной, а остальные 98 % зависят от других, не учтенных в данном случае переменных;

– уравнение регрессии имеет вид

$$y_x = 256{,}57 + 26{,}52x,$$

т. е. увеличение значения независимой переменной на одну единицу влечет увеличение значения зависимой переменной примерно на 26 р. и 52 коп. и наоборот, величина его свободного члена статистически значима;

– можно быть на 95 % уверенным в том, что примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели при повышении образования мужа на один уровень будут увеличиваться на сумму в пределах между 17,40 и 35,63 р.;

– стандартная ошибка оценки уравнения регрессии *SEE* равна 125,59 р.;

– так как отношение значения асимметрии распределения остатков к ее средней квадратичной ошибке $\frac{|A_S|}{S_{A_S}} = \frac{0{,}47}{0{,}06} = 7{,}41$, что больше 3,0, асимметрию следует признать существенной, а распределение остатков не следует считать симметричным;

– так как отношение значения эксцесса распределения остатков к его средней квадратичной ошибке $\frac{|\varepsilon_k|}{S_{\varepsilon_k}} = \frac{|-0{,}50|}{0{,}13} = 3{,}94$, что больше 3,0, отклонение распределения остатков от нормального существенно и его нельзя считать нормальным или близким к нему;

– рассчитанное уравнение регрессии следует признать условно пригодным для прогнозирования расходов домохозяйств на покупку (обновление) элементов домашней мебели в зависимости от уровня образования мужа.

11 Значения для пар переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Образование жены» приведены в таблицах 7.18–7.21.

Таблица 7.18 – Показатели регрессионной статистики для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Образование жены»

| Показатель | Значение |
|---|---|
| Парный коэффициент корреляции | 0,20 |
| Коэффициент детерминации | 0,04 |
| Нормированный коэффициент детерминации | 0,04 |
| Стандартная ошибка | 124,36 |
| Количество наблюдений | 1440 |

Таблица 7.19 – Результаты однофакторного дисперсионного анализа для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Образование жены»

| Показатель | Число степеней свободы, *df* | Сумма квадратов разностей, *SS* | Оценка дисперсий (средний квадрат), *MS* | Значение *F*-критерия | Значимость *F*-критерия |
|---|---|---|---|---|---|
| Регрессия | 1 | 954994,75 | 954994,75 | 61,75 | 0,00 |
| Остаток | 1438 | 22239795,25 | 15465,78 | | |
| Итого | 1439 | 23194790,00 | | | |

Таблица 7.20 – Значения коэффициентов уравнения регрессии, оценок их статистической значимости и границ доверительных интервалов для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Образование жены»

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Свободный член, $\beta_0$ | 223,34 | 14,78 | 15,11 | 0,00 | 194,34 | 252,34 |
| Коэффициент при независимой переменной, $\beta_1$ | 39,58 | 5,04 | 7,86 | 0,00 | 29,70 | 49,46 |

Таблица 7.21 – Значения описательных статистик остатков уравнения регрессии для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Образование жены»

| Статистики | Значения |
|---|---|
| Среднее | 0,00 |
| Стандартная ошибка | 3,28 |
| Медиана | –14,10 |
| Мода | –14,10 |
| Стандартное отклонение | 124,32 |
| Дисперсия выборки | 15455,03 |
| Эксцесс | –0,20 |
| Асимметричность | 0,56 |
| Интервал | 559 |
| Минимум | –211,09 |
| Максимум | 347,91 |
| Сумма | 0 |
| Счет | 1440 |

Значения, приведенные в таблицах 7.18–7.21, позволяют заключить следующее:

– следует принять альтернативную гипотезу: корреляционная связь между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и уровнем образования жены существует, является статистически значимой, положительной, но по силе – практически отсутствующей;

– примерно 4 % вариации зависимой переменной объясняется вариацией независимой переменной, а остальные 96 % зависят от других, не учтенных в данном случае переменных;

– уравнение регрессии имеет вид

$$y_x = 223{,}34 + 39{,}58x,$$

т. е. увеличение значения независимой переменной на одну единицу влечет увеличение значения зависимой переменной примерно на 39 р. и 58 коп. и наоборот, величина его свободного члена статистически значима;

– можно быть на 95 % уверенным в том, что примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели при повышении образования жены на один уровень будут увеличиваться на сумму в пределах между 29,70 и 49,46 р.;

– стандартная ошибка оценки уравнения регрессии *SEE* равна 124,36 р.;

– так как отношение значения асимметрии распределения остатков к ее средней квадратичной ошибке $\frac{|A_S|}{S_{A_S}} = \frac{0{,}56}{0{,}06} = 8{,}80$, что существенно больше 3,0, асимметрию следует признать существенной, а распределение остатков не следует считать симметричным;

– так как отношение значения эксцесса распределения остатков к его средней квадратичной ошибке $\frac{|\varepsilon_k|}{S_{\varepsilon_k}} = \frac{|-0{,}20|}{0{,}13} = 1{,}57$, что меньше 3,0, при условии, что асимметрия несущественна, распределение остатков можно было бы считать близким к нормальному;

– рассчитанное уравнение регрессии следует признать условно пригодным для прогнозирования расходов домохозяйств на покупку (обновление) элементов домашней мебели в зависимости от уровня образования жены.

12 Значения для пар переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Семейный среднемесячный доход» приведены в таблицах 7.22–7.25.

Таблица 7.22 – Показатели регрессионной статистики для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Семейный среднемесячный доход»

| Показатель | Значение |
|---|---|
| Парный коэффициент корреляции | 0,999 |
| Коэффициент детерминации | 0,999 |

| Показатель | Значение |
|---|---|
| Нормированный коэффициент детерминации | 0,999 |
| Стандартная ошибка | 0,20 |
| Количество наблюдений | 1440 |

Таблица 7.23 – Результаты однофакторного дисперсионного анализа для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Семейный среднемесячный доход»

| Показатель | Число степеней свободы, *df* | Сумма квадратов разностей, *SS* | Оценка дисперсий (средний квадрат), *MS* | Значение *F*-критерия | Значимость *F*-критерия |
|---|---|---|---|---|---|
| Регрессия | 1 | 23194732,70 | 23194732,70 | 582106436,81 | 0,00 |
| Остаток | 1438 | 57,30 | 0,04 | | |
| Итого | 1439 | 23194790,00 | | | |

Таблица 7.24 – Значения рассчитанных коэффициентов уравнения регрессии, оценок их статистической значимости и границ доверительных интервалов для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Семейный среднемесячный доход»

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Свободный член, $\beta_0$ | 0,10 | 0,01 | 6,56 | 0,00 | 0,07 | 0,13 |
| Коэффициент при независимой переменной, $\beta_1$ | 0,05 | 0,00 | 24126,88 | 0,00 | 0,05 | 0,05 |

Таблица 7.25 – Значения описательных статистик остатков уравнения регрессии для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов мебели» и «Семейный среднемесячный доход»

| Статистики | Значения |
|---|---|
| Среднее | 0,00 |
| Стандартная ошибка | 0,01 |
| Медиана | –0,10 |
| Мода | 0,40 |
| Стандартное отклонение | 0,20 |
| Дисперсия выборки | 0,04 |
| Эксцесс | 0,29 |

| Статистики | Значения |
|---|---|
| Асимметричность | 1,51 |
| Интервал | 0,50 |
| Минимум | –0,10 |
| Максимум | 0,40 |
| Сумма | 0,00 |
| Счет | 1440 |

Значения, приведенные в таблицах 7.22–7.25, позволяют заключить следующее:

– следует принять альтернативную гипотезу;

– корреляционная связь между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и семейным среднемесячным доходом существует, является статистически значимой, положительной, а по силе – почти функциональной;

– почти 100 % вариации зависимой переменной объясняется вариацией независимой переменной;

– уравнение регрессии имеет вид

$$y_x = 0{,}01 + 0{,}05x,$$

т. е. увеличение значения независимой переменной на одну единицу влечет увеличение значения зависимой переменной примерно на 5 коп. и наоборот, величина его свободного члена статистически значима;

– можно быть на 95 % уверенным в том, что примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели при увеличении семейного среднемесячного дохода на 1 р. будут увеличиваться примерно на 5 коп.;

– стандартная ошибка оценки уравнения регрессии *SEE* равна 0,20 р.;

– так как отношение значения асимметрии распределения остатков к ее средней квадратичной ошибке $\frac{|A_S|}{S_{A_S}} = \frac{1{,}51}{0{,}06} = 23{,}6$, что значительно больше 3,0, асимметрию следует признать существенной, а распределение остатков не следует считать симметричным;

– так как отношение значения эксцесса распределения остатков к его средней квадратичной ошибке $\frac{|\varepsilon_k|}{S_{\varepsilon_k}} = \frac{0{,}29}{0{,}13} = 2{,}25$, что меньше 3,0, при условии, что асимметрия несущественна, распределение остатков можно было бы считать близким к нормальному;

– рассчитанное уравнение регрессии следует признать условно пригодным для прогнозирования расходов домохозяйств на покупку (обновление) элементов домашней мебели в зависимости от семейного среднемесячного дохода.

Однако на это уравнение необходимо обратить отдельное внимание по причине того, что оно является достаточно точным с точки зрения стандартной ошибки, которая по своей величине практически не отличается от таковых,

которые будут рассчитаны в лабораторной работе № 8 для всех трех уравнений множественной (многофакторной) регрессии. Кроме этого, построенная гистограмма остатков позволяет выдвинуть гипотезу о наличии на изучаемом рынке двух четко различимых сегментов, которая будет подтверждена по итогам множественного (многофакторного) корреляционно-регрессионного, кластерного и дискриминантного анализов в лабораторных работах № 8, 9 и 10.

13 Значения для пар переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Количество автомобилей в семье» приведены в таблицах 7.26–7.29.

Таблица 7.26 – Показатели регрессионной статистики для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Количество автомобилей в семье»

| Показатель | Значение |
|---|---|
| Парный коэффициент корреляции | 0,91 |
| Коэффициент детерминации | 0,83 |
| Нормированный коэффициент детерминации | 0,83 |
| Стандартная ошибка | 52,02 |
| Количество наблюдений | 1440 |

Таблица 7.27 – Результаты однофакторного дисперсионного анализа для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Количество автомобилей в семье»

| Показатель | Число степеней свободы, *df* | Сумма квадратов разностей, *SS* | Оценка дисперсий (средний квадрат), *MS* | Значение *F*-критерия | Значимость *F*-критерия |
|---|---|---|---|---|---|
| Регрессия | 1 | 19302722,96 | 19302722,96 | 7131,77 | 0,00 |
| Остаток | 1438 | 3892067,04 | 2706,58 | | |
| Итого | 1439 | 23194790,00 | | | |

Таблица 7.28 – Значения рассчитанных коэффициентов уравнения регрессии, оценок их статистической значимости и границ доверительных интервалов для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Количество автомобилей в семье»

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Свободный член, $\beta_0$ | 139,68 | 2,71 | 51,63 | 0,00 | 134,37 | 144,98 |

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Коэффициент при независимой переменной, $\beta_1$ | 166,04 | 1,97 | 84,45 | 0,00 | 162,19 | 169,90 |

Таблица 7.29 – Значения описательных статистик остатков уравнения регрессии для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Количество автомобилей в семье»

| Статистики | Значения |
|---|---|
| Среднее | 0,00 |
| Стандартная ошибка | 1,37 |
| Медиана | 1,24 |
| Мода | 22,28 |
| Стандартное отклонение | 52,01 |
| Дисперсия выборки | 2704,70 |
| Эксцесс | –0,87 |
| Асимметричность | 0,023 |
| Интервал | 221,96 |
| Минимум | –105,72 |
| Максимум | 116,24 |
| Сумма | 0,00 |
| Счет | 1440 |

Значения, приведенные в таблицах 7.26–7.29, позволяют заключить следующее:

– следует принять альтернативную гипотезу: корреляционная связь между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и количеством автомобилей в семье существует, является статистически значимой, положительной и сильной;

– примерно 83 % вариации зависимой переменной объясняется вариацией независимой переменной, а остальные 17 % зависят от других, не учтенных в данном случае переменных;

– уравнение регрессии имеет вид

$$y_x = 139{,}68 + 166{,}04x,$$

т. е. увеличение значения независимой переменной на одну единицу влечет увеличение значения зависимой переменной примерно на 166 р. и 4 коп. и наоборот, величина его свободного члена статистически значима;

– можно быть на 95 % уверенным в том, что примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели при увеличении количества автомобилей в семье на одну единицу будут увеличиваться на сумму в пределах между 162,19 и 169,90 р.;

– стандартная ошибка оценки уравнения регрессии *SEE* равна 52,02 р.;

– так как отношение значения асимметрии распределения остатков к ее средней квадратичной ошибке $\frac{|A_S|}{S_{A_S}} = \frac{0,02}{0,06} = 0,36$, что меньше 3,0, асимметрию следует признать несущественной, а распределение остатков следует считать близким к симметричному;

– так как отношение значения эксцесса распределения остатков к его средней квадратичной ошибке $\frac{|\varepsilon_k|}{S_{\varepsilon_k}} = \frac{|-0,87|}{0,13} = 6,81$, что больше 3,0, отклонение распределения остатков от нормального существенно и его нельзя считать нормальным или близким к нему;

– рассчитанное уравнение регрессии следует признать условно пригодным для прогнозирования расходов домохозяйств на приобретение (обновление) элементов домашней мебели в зависимости от количества автомобилей в семье.

14 Значения для пар переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Наличие подключения к интернету» приведены в таблицах 7.30–7.33.

Таблица 7.30 – Показатели регрессионной статистики для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Наличие подключения к интернету»

| Показатель | Значение |
|---|---|
| Парный коэффициент корреляции | 0,20 |
| Коэффициент детерминации | 0,04 |
| Нормированный коэффициент детерминации | 0,04 |
| Стандартная ошибка | 124,34 |
| Количество наблюдений | 1440 |

Таблица 7.31 – Результаты однофакторного дисперсионного анализа для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Наличие подключения к интернету»

| Показатель | Число степеней свободы, *df* | Сумма квадратов разностей, *SS* | Оценка дисперсий (средний квадрат), *MS* | Значение *F*-критерия | Значимость *F*-критерия |
|---|---|---|---|---|---|
| Регрессия | 1 | 962754,23 | 962754,23 | 62,27 | 0,00 |
| Остаток | 1438 | 22232035,76 | 15460,39 | | |
| Итого | 1439 | 23194790,00 | | | |

Таблица 7.32 – Значения рассчитанных коэффициентов уравнения регрессии, оценок их статистической значимости и границ доверительных интервалов для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Наличие подключения к интернету»

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Свободный член, $\beta_0$ | 235,27 | 13,25 | 17,75 | 0,00 | 209,27 | 261,27 |
| Коэффициент при независимой переменной, $\beta_1$ | 58,47 | 7,41 | 7,89 | 0,00 | 43,94 | 73,01 |

Таблица 7.33 – Значения описательных статистик остатков уравнения регрессии для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Наличие подключения к интернету»

| Статистики | Значения |
|---|---|
| Среднее | 0,00 |
| Стандартная ошибка | 3,28 |
| Медиана | –14,22 |
| Мода | -95,22 |
| Стандартное отклонение | 124,30 |
| Дисперсия выборки | 15449,64 |
| Эксцесс | –0,49 |
| Асимметричность | 0,42 |
| Интервал | 569 |
| Минимум | –231,22 |
| Максимум | 337,78 |
| Сумма | 0,00 |
| Счет | 1440 |

Значения, приведенные в таблицах 7.30–7.33, позволяют заключить следующее:

– следует принять альтернативную гипотезу: корреляционная связь между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и наличием подключения к интернету существует, является статистически значимой, положительной, но по силе – практически отсутствующей;

– примерно 4 % вариации зависимой переменной объясняется вариацией независимой переменной, а остальные 96 % зависят от других, не учтенных в данном случае переменных;

– уравнение регрессии имеет вид

$$y_x = 235{,}27 + 58{,}47x,$$

т. е. увеличение значения независимой переменной на одну единицу влечет увеличение значения зависимой переменной примерно на 58 р. и 47 коп. и наоборот, величина его свободного члена статистически значима;

– можно быть на 95 % уверенным в том, что примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели при подключении домохозяйства к интернету увеличатся на сумму в пределах между 43,94 и 73,01 р.;

– стандартная ошибка оценки уравнения регрессии *SEE* равна 124,34 р.;

– так как отношение значения асимметрии распределения остатков к ее средней квадратичной ошибке $\frac{|A_S|}{S_{A_S}} = \frac{0{,}42}{0{,}06} = 6{,}56$, что больше 3,0, асимметрию следует признать существенной, а распределение остатков не следует считать симметричным;

– так как отношение значения эксцесса распределения остатков к его средней квадратичной ошибке $\frac{|\varepsilon_k|}{S_{\varepsilon_k}} = \frac{|-0{,}49|}{0{,}13} = 3{,}79$, что больше 3,0, отклонение распределения остатков от нормального существенно и его нельзя считать нормальным или близким к нему;

– рассчитанное уравнение регрессии следует признать условно пригодным для прогнозирования расходов домохозяйств на покупку (обновление) элементов домашней мебели в зависимости от подключения домохозяйств к интернету.

15 Значения для пар переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Вид телевидения» приведены в таблицах 7.34–7.37.

Таблица 7.34 – Показатели регрессионной статистики для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Вид телевидения»

| Показатель | Значение |
|---|---|
| Парный коэффициент корреляции | 0,35 |
| Коэффициент детерминации | 0,12 |
| Нормированный коэффициент детерминации | 0,12 |
| Стандартная ошибка | 119,00 |
| Количество наблюдений | 1440 |

Таблица 7.35 – Результаты однофакторного дисперсионного анализа для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Вид телевидения»

| Показатель | Число степеней свободы, *df* | Сумма квадратов разностей, *SS* | Оценка дисперсий (средний квадрат), *MS* | Значение *F*-критерия | Значимость *F*-критерия |
|---|---|---|---|---|---|
| Регрессия | 1 | 2830479,07 | 2830479,07 | 199,87 | 0,00 |
| Остаток | 1438 | 20364310,92 | 14161,55 | | |
| Итого | 1439 | 23194790,00 | | | |

Таблица 7.36 – Значения рассчитанных коэффициентов уравнения регрессии, оценок их статистической значимости и границ доверительных интервалов для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Вид телевидения»

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Свободный член, $\beta_0$ | 211,17 | 9,41 | 22,44 | 0,00 | 192,71 | 229,63 |
| Коэффициент при независимой переменной, $\beta_1$ | 95,03 | 6,72 | 14,14 | 0,00 | 81,85 | 108,22 |

Таблица 7.37 – Значения описательных статистик остатков уравнения регрессии для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Вид телевидения»

| Статистики | Значения |
|---|---|
| Среднее | 0,00 |
| Стандартная ошибка | 3,13 |
| Медиана | –4,20 |
| Мода | 21,80 |
| Стандартное отклонение | 118,96 |
| Дисперсия выборки | 14151,71 |
| Эксцесс | –0,79 |
| Асимметричность | 0,15 |
| Интервал | 506 |
| Минимум | –217,23 |
| Максимум | 288,77 |
| Сумма | 0,00 |
| Счет | 1440 |

Значения, приведенные в таблицах 7.34–7.37, позволяют заключить следующее:

– следует принять альтернативную гипотезу: корреляционная связь между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и видом телевидения существует, является статистически значимой, положительной, но слабой;

– примерно 12 % вариации зависимой переменной объясняется вариацией независимой переменной, а остальные 88 % зависят от других, не учтенных в данном случае переменных;

– уравнение регрессии имеет вид

$$y_x = 211{,}17 + 95{,}03x,$$

т. е. увеличение значения независимой переменной на одну единицу влечет увеличение значения зависимой переменной примерно на 95 р. и 3 коп. и наоборот, величина его свободного члена статистически значима;

– можно быть на 95 % уверенным в том, что примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели при подключении домохозяйства к кабельному телевидению увеличатся на сумму в пределах между 81,85 и 108,22 р.;

– стандартная ошибка оценки уравнения регрессии *SEE* равна 119,00 р.;

– так как отношение значения асимметрии распределения остатков к ее средней квадратичной ошибке $\frac{|A_S|}{S_{A_S}} = \frac{0{,}15}{0{,}06} = 2{,}29$, что меньше 3,0, асимметрию следует признать несущественной, а распределение остатков следует считать близким к симметричному;

– так как отношение значения эксцесса распределения остатков к его средней квадратичной ошибке $\frac{|\varepsilon_k|}{S_{\varepsilon_k}} = \frac{|-0{,}79|}{0{,}13} = 6{,}14$, что больше 3,0, отклонение распределения остатков от нормального существенно и его нельзя считать нормальным или близким к нему;

– рассчитанное уравнение регрессии следует признать условно пригодным для прогнозирования расходов домохозяйств на покупку (обновление) элементов домашней мебели в зависимости от вида телевидения.

16 Значения для пар переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Вид жилья» приведены в таблицах 7.38–7.41.

Таблица 7.38 – Показатели регрессионной статистики для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Вид жилья»

| Показатель | Значение |
|---|---|
| Парный коэффициент корреляции | 0,70 |
| Коэффициент детерминации | 0,49 |

| Показатель | Значение |
|---|---|
| Нормированный коэффициент детерминации | 0,49 |
| Стандартная ошибка | 90,37 |
| Количество наблюдений | 1440 |

Таблица 7.39 – Результаты однофакторного дисперсионного анализа для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Вид жилья»

| Показатель | Число степеней свободы, *df* | Сумма квадратов разностей, *SS* | Оценка дисперсий (средний квадрат), *MS* | Значение *F*-критерия | Значимость *F*-критерия |
|---|---|---|---|---|---|
| Регрессия | 1 | 11450937,56 | 11450937,56 | 1402,13 | 0,00 |
| Остаток | 1438 | 11743852,44 | 8166,80 | | |
| Итого | 1439 | 23194790,00 | | | |

Таблица 7.40 – Значения рассчитанных коэффициентов уравнения регрессии, оценок их статистической значимости и границ доверительных интервалов для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Вид жилья»

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Свободный член, $\beta_0$ | 832,74 | 13,46 | 61,86 | 0,00 | 806,33 | 859,14 |
| Коэффициент при независимой переменной, $\beta_1$ | –265,28 | 7,08 | –37,45 | 0,00 | –279,18 | –251,38 |

Таблица 7.41 – Значения описательных статистик остатков уравнения регрессии для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Вид жилья»

| Статистики | Значения |
|---|---|
| Среднее | 0,00 |
| Стандартная ошибка | 2,38 |
| Медиана | –3,45 |
| Мода | 25,83 |
| Стандартное отклонение | 90,34 |
| Дисперсия выборки | 8161,12 |
| Эксцесс | –0,62 |
| Асимметричность | 0,06 |

| Статистики | Значения |
|---|---|
| Интервал | 377 |
| Минимум | –186,17 |
| Максимум | 190,83 |
| Сумма | 0,00 |
| Счет | 1440 |

Значения, приведенные в таблицах 7.38–7.41, позволяют заключить следующее:

– следует принять альтернативную гипотезу: корреляционная связь между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и видом жилья существует, является статистически значимой, отрицательной, а по силе – умеренной;

– примерно 49 % вариации зависимой переменной объясняется вариацией независимой переменной, а остальные 51 % зависят от других, не учтенных в данном случае переменных;

– уравнение регрессии имеет вид

$$y_x = 832{,}74 - 265{,}28x,$$

т. е. увеличение значения независимой переменной на одну единицу влечет уменьшение значения зависимой переменной примерно на 265 р. и 28 коп. и наоборот, величина его свободного члена статистически значима;

– можно быть на 95 % уверенным в том, что примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели у семей, живущих в квартирах, меньше, чем у семей, живущих в домах, на сумму в пределах между 251,38 и 279,18 р.;

– стандартная ошибка оценки уравнения регрессии *SEE* равна 90,37 р.;

– так как отношение значения асимметрии распределения остатков к ее средней квадратичной ошибке $\frac{|A_S|}{S_{A_S}} = \frac{0{,}06}{0{,}06} = 1{,}00$, что меньше 3,0, асимметрию следует признать несущественной, а распределение остатков следует считать симметричным;

– так как отношение значения эксцесса распределения остатков к его средней квадратичной ошибке $\frac{|\varepsilon_k|}{S_{\varepsilon_k}} = \frac{|-0{,}62|}{0{,}13} = 4{,}83$, что больше 3,0, отклонение распределения остатков от нормального существенно и его нельзя считать нормальным или близким к нему;

– рассчитанное уравнение регрессии следует признать условно пригодным для прогнозирования расходов домохозяйств на покупку (обновление) элементов домашней мебели в зависимости от вида жилья.

17 Значения для пар переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Площадь жилья» приведены в таблицах 7.42–7.45.

Таблица 7.42 – Показатели регрессионной статистики для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Площадь жилья»

| Показатель | Значение |
|---|---|
| Парный коэффициент корреляции | 0,86 |
| Коэффициент детерминации | 0,74 |
| Нормированный коэффициент детерминации | 0,74 |
| Стандартная ошибка | 64,18 |
| Количество наблюдений | 1440 |

Таблица 7.43 – Результаты однофакторного дисперсионного анализа для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Площадь жилья»

| Показатель | Число степеней свободы, *df* | Сумма квадратов разностей, *SS* | Оценка дисперсий (средний квадрат), *MS* | Значение *F*-критерия | Значимость *F*-критерия |
|---|---|---|---|---|---|
| Регрессия | 1 | 17271491,76 | 17271491,76 | 4193,00 | 0,00 |
| Остаток | 1438 | 5923298,24 | 4119,12 | | |
| Итого | 1439 | 23194790,00 | | | |

Таблица 7.44 – Значения рассчитанных коэффициентов уравнения регрессии, оценок их статистической значимости и границ доверительных интервалов для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Площадь жилья»

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Свободный член, $\beta_0$ | 182,99 | 2,91 | 62,80 | 0,00 | 177,28 | 188,71 |
| Коэффициент при независимой переменной, $\beta_1$ | 1,70 | 0,03 | 64,75 | 0,00 | 1,64 | 1,75 |

Таблица 7.45 – Значения описательных статистик остатков уравнения регрессии для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Площадь жилья»

| Статистики | Значения |
|---|---|
| Среднее | 0,00 |
| Стандартная ошибка | 1,69 |
| Медиана | –5,62 |

| Статистики | Значения |
|---|---|
| Мода | –17,58 |
| Стандартное отклонение | 64,16 |
| Дисперсия выборки | 4116,26 |
| Эксцесс | –0,41 |
| Асимметричность | 0,20 |
| Интервал | 285,41 |
| Минимум | –128,05 |
| Максимум | 157,35 |
| Сумма | 0,00 |
| Счет | 1440 |

Значения, приведенные в таблицах 7.42–7.45, позволяют заключить следующее:

– следует принять альтернативную гипотезу: корреляционная связь между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и площадью жилья существует, является статистически значимой, положительной и сильной;

– примерно 74 % вариации зависимой переменной объясняется вариацией независимой переменной, а остальные 26 % зависят от других, не учтенных в данном случае переменных;

– уравнение регрессии имеет вид

$$y_x = 182{,}99 + 1{,}70x,$$

т. е. увеличение значения независимой переменной на одну единицу влечет увеличение значения зависимой переменной примерно на 1 р. и 70 коп. и наоборот, величина его свободного члена статистически значима;

– можно быть на 95 % уверенным в том, что примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели при увеличении площади жилья на 1 кв. м будут увеличиваться на сумму в пределах между 1,64 и 1,75 р.;

– стандартная ошибка оценки уравнения регрессии *SEE* равна 64,18 р.;

– так как отношение значения асимметрии распределения остатков к ее средней квадратичной ошибке $\frac{|A_S|}{S_{A_S}} = \frac{0{,}21}{0{,}06} = 3{,}17$, что больше 3,0, асимметрию следует признать существенной, а распределение остатков не следует считать симметричным;

– так как отношение значения эксцесса распределения остатков к его средней квадратичной ошибке $\frac{|\varepsilon_k|}{S_{\varepsilon_k}} = \frac{|-0{,}42|}{0{,}13} = 3{,}24$, что больше 3,0, отклонение распределения остатков от нормального существенно и его нельзя считать нормальным или близким к нему;

– рассчитанное уравнение регрессии следует признать условно пригодным для прогнозирования расходов домохозяйств на покупку (обновление) элементов домашней мебели в зависимости от площади жилья.

18 Значения для пар переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Средний возраст мебели» приведены в таблицах 7.46–7.49.

Таблица 7.46 – Показатели регрессионной статистики для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Средний возраст мебели»

| Показатель | Значение |
|---|---|
| Парный коэффициент корреляции | 0,30 |
| Коэффициент детерминации | 0,09 |
| Нормированный коэффициент детерминации | 0,09 |
| Стандартная ошибка | 121,21 |
| Количество наблюдений | 1440 |

Таблица 7.47 – Результаты однофакторного дисперсионного анализа для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Средний возраст мебели»

| Показатель | Число степеней свободы, *df* | Сумма квадратов разностей, *SS* | Оценка дисперсий (средний квадрат), *MS* | Значение *F*-критерия | Значимость *F*-критерия |
|---|---|---|---|---|---|
| Регрессия | 1 | 2066218,19 | 2066218,19 | 140,63 | 0,00 |
| Остаток | 1438 | 21128571,81 | 14693,03 | | |
| Итого | 1439 | 23194790,00 | | | |

Таблица 7.48 – Значения рассчитанных коэффициентов уравнения регрессии, оценок их статистической значимости и границ доверительных интервалов для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Средний возраст мебели»

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Свободный член, $\beta_0$ | 469,37 | 11,64 | 40,32 | 0,00 | 446,53 | 492,20 |
| Коэффициент при независимой переменной, $\beta_1$ | –11,87 | 1,00 | –11,86 | 0,00 | –13,83 | –9,91 |

Таблица 7.49 – Значения описательных статистик остатков уравнения регрессии для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Средний возраст мебели»

| Статистики | Значения |
|---|---|
| Среднее | 0,00 |
| Стандартная ошибка | 3,19 |
| Медиана | –5,31 |
| Мода | –105,53 |
| Стандартное отклонение | 121,17 |
| Дисперсия выборки | 14682,82 |
| Эксцесс | –0,70 |
| Асимметричность | 0,241 |
| Интервал | 505 |
| Минимум | –213,15 |
| Максимум | 291,85 |
| Сумма | 0,00 |
| Счет | 1440 |

Значения, приведенные в таблицах 7.46–7.49, позволяют заключить следующее:

– следует принять альтернативную гипотезу: корреляционная связь между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и ее средним возрастом существует, является статистически значимой, отрицательной, но по силе – практически отсутствующей;

– примерно 9 % вариации зависимой переменной объясняется вариацией независимой переменной, а остальные 91 % зависят от других, не учтенных в данном случае переменных;

– уравнение регрессии имеет вид

$$y_x = 469{,}37 - 11{,}87x,$$

т. е. увеличение значения независимой переменной на одну единицу влечет уменьшение значения зависимой переменной примерно на 11 р. и 87 коп. и наоборот, величина его свободного члена статистически значима;

– можно быть на 95 % уверенным в том, что примерные среднегодовые расходы на покупку (обновление) элементов мебели при увеличении ее возраста на один год будут уменьшаться на сумму в пределах между 9,91 и 13,83 р.;

– стандартная ошибка оценки уравнения регрессии *SEE* равна 121,21 р.;

– так как отношение значения асимметрии распределения остатков к ее средней квадратичной ошибке $\frac{|A_S|}{S_{A_S}} = \frac{0{,}24}{0{,}06} = 3{,}76$, что больше 3,0, асимметрию следует признать существенной, а распределение остатков не следует считать симметричным;

– так как отношение значения эксцесса распределения остатков к его средней квадратичной ошибке $\frac{|\varepsilon_k|}{S_{\varepsilon_k}} = \frac{|-0{,}70|}{0{,}13} = 5{,}46$, что больше 3,0, отклонение распределения остатков от нормального существенно и его нельзя считать нормальным или близким к нему;

– рассчитанное уравнение регрессии следует признать условно пригодным для прогнозирования расходов домохозяйств на покупку (обновление) элементов домашней мебели в зависимости от возраста мебели.

19 Значения для пар переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Стиль мебели» приведены в таблицах 7.50–7.53.

Таблица 7.50 – Показатели регрессионной статистики для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Стиль мебели»

| Показатель | Значение |
|---|---|
| Парный коэффициент корреляции | 0,94 |
| Коэффициент детерминации | 0,88 |
| Нормированный коэффициент детерминации | 0,88 |
| Стандартная ошибка | 43,49 |
| Количество наблюдений | 1440 |

Таблица 7.51 – Результаты однофакторного дисперсионного анализа для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Стиль мебели»

| Показатель | Число степеней свободы, *df* | Сумма квадратов разностей, *SS* | Оценка дисперсий (средний квадрат), *MS* | Значение *F*-критерия | Значимость *F*-критерия |
|---|---|---|---|---|---|
| Регрессия | 1 | 20474712,90 | 20474712,90 | 10824,19 | 0,00 |
| Остаток | 1438 | 2720077,10 | 1891,57 | | |
| Итого | 1439 | 23194790,00 | | | |

Таблица 7.52 – Значения рассчитанных коэффициентов уравнения регрессии, оценок их статистической значимости и границ доверительных интервалов для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Стиль мебели»

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Свободный член, $\beta_0$ | 36,74 | 3,10 | 11,84 | 0,00 | 30,65 | 42,82 |

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Коэффициент при независимой переменной, $\beta_1$ | 140,75 | 1,35 | 104,04 | 0,00 | 138,10 | 143,41 |

Таблица 7.53 – Значения описательных статистик остатков уравнения регрессии для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Стиль мебели»

| Статистики | Значения |
|---|---|
| Среднее | 0,00 |
| Стандартная ошибка | 1,15 |
| Медиана | –0,25 |
| Мода | 9,75 |
| Стандартное отклонение | 43,48 |
| Дисперсия выборки | 1890,26 |
| Эксцесс | -0,99 |
| Асимметричность | 0,17 |
| Интервал | 171,25 |
| Минимум | -75,25 |
| Максимум | 965,00 |
| Сумма | 0,00 |
| Счет | 1440 |

Значения, приведенные в таблицах 7.50–7.53, позволяют заключить следующее:

– следует принять альтернативную гипотезу: корреляционная связь между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и стилем мебели существует, является статистически значимой, положительной и сильной;

– примерно 88 % вариации зависимой переменной объясняется вариацией независимой переменной, а остальные 12 % зависят от других, не учтенных в данном случае переменных;

– уравнение регрессии имеет вид

$$y_x = 36{,}74 + 140{,}75x,$$

т. е. увеличение значения независимой переменной на одну единицу влечет увеличение значения зависимой переменной примерно на 140 р. и 75 коп. и наоборот, величина его свободного члена статистически значима;

– можно быть на 95 % уверенным в том, что примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели при приобретении

мебели более модного стиля будут увеличиваться на сумму в пределах между 138,10 и 143,41 р.;

– стандартная ошибка оценки уравнения регрессии *SEE* равна 43,49 р.;

– так как отношение значения асимметрии распределения остатков к ее средней квадратичной ошибке $\frac{|A_S|}{S_{A_S}} = \frac{0,17}{0,06} = 2,64$, что меньше 3,0, асимметрию следует признать несущественной, а распределение остатков следует считать симметричным;

– так как отношение значения эксцесса распределения остатков к его средней квадратичной ошибке $\frac{|\varepsilon_k|}{S_{\varepsilon_k}} = \frac{|-0,99|}{0,13} = 7,70$, что больше 3,0, отклонение распределения остатков от нормального существенно и его нельзя считать нормальным или близким к нему;

– рассчитанное уравнение регрессии следует признать условно пригодным для прогнозирования расходов домохозяйств на покупку (обновление) элементов домашней мебели в зависимости от стиля мебели, к которому привержены члены домохозяйств.

20 Значения для пар переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Планируемая периодичность обновления мебели» приведены в таблицах 7.54–7.57.

Таблица 7.54 – Показатели регрессионной статистики для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Планируемая периодичность обновления мебели»

| Показатель | Значение |
|---|---|
| Парный коэффициент корреляции | 0,30 |
| Коэффициент детерминации | 0,09 |
| Нормированный коэффициент детерминации | 0,09 |
| Стандартная ошибка | 121,32 |
| Количество наблюдений | 1440 |

Таблица 7.55 – Результаты однофакторного дисперсионного анализа для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Планируемая периодичность обновления мебели»

| Показатель | Число степеней свободы, *df* | Сумма квадратов разностей, *SS* | Оценка дисперсий (средний квадрат), *MS* | Значение *F*-критерия | Значимость *F*-критерия |
|---|---|---|---|---|---|
| Регрессия | 1 | 2028837,83 | 2028837,83 | 137,84 | 0,00 |
| Остаток | 1438 | 21165952,16 | 14719,02 | | |
| Итого | 1439 | 23194790,00 | | | |

Таблица 7.56 – Значения рассчитанных коэффициентов уравнения регрессии, оценок их статистической значимости и границ доверительных интервалов для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Планируемая периодичность обновления мебели»

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Свободный член, $\beta_0$ | 471,59 | 11,93 | 39,52 | 0,00 | 448,19 | 495,00 |
| Коэффициент при независимой переменной, $\beta_1$ | –10,92 | 0,93 | –11,74 | 0,00 | –12,75 | –9,10 |

Таблица 7.57 – Значения описательных статистик остатков уравнения регрессии для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Планируемая периодичность обновления мебели»

| Статистики | Значения |
|---|---|
| Среднее | 0,00 |
| Стандартная ошибка | 3,20 |
| Медиана | –5,73 |
| Мода | –105,35 |
| Стандартное отклонение | 121,28 |
| Дисперсия выборки | 14708,79 |
| Эксцесс | –0,70 |
| Асимметричность | 0,25 |
| Интервал | 505 |
| Минимум | –210,12 |
| Максимум | 294,88 |
| Сумма | 0,00 |
| Счет | 1440 |

Значения, приведенные в таблицах 7.54–7.57, позволяют заключить следующее:

– следует принять альтернативную гипотезу: корреляционная связь между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и планируемой периодичностью обновления мебели существует, является статистически значимой, отрицательной, но по силе – практически отсутствующей;

– примерно 9 % вариации зависимой переменной объясняется вариацией независимой переменной, а остальные 91 % зависят от других, не учтенных в данном случае переменных;

– уравнение регрессии имеет вид

$$y_x = 471{,}59 - 10{,}92x,$$

т. е. увеличение значения независимой переменной на одну единицу влечет уменьшение значения зависимой переменной примерно на 10 р. и 92 коп. и наоборот, величина его свободного члена статистически значима;

– можно быть на 95 % уверенным в том, что примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели при увеличении периода обновления на один год будут уменьшаться на сумму в пределах между 9,10 и 12,75 р.;

– стандартная ошибка оценки уравнения регрессии *SEE* равна 121,32 р.;

– так как отношение значения асимметрии распределения остатков к ее средней квадратичной ошибке $\frac{|A_S|}{S_{A_S}} = \frac{0{,}25}{0{,}06} = 3{,}87$, что больше 3,0, асимметрию следует признать существенной, а распределение остатков не следует считать симметричным;

– так как отношение значения эксцесса распределения остатков к его средней квадратичной ошибке $\frac{|\varepsilon_k|}{S_{\varepsilon_k}} = \frac{|-0{,}70|}{0{,}13} = 5{,}47$, что больше 3,0, отклонение распределения остатков от нормального существенно и его нельзя считать нормальным или близким к нему;

– рассчитанное уравнение регрессии следует признать условно пригодным для прогнозирования расходов домохозяйств на покупку (обновление) элементов домашней мебели в зависимости от периодичности обновления мебели.

### 7.2.2 Выполнение парного (однофакторного) корреляционно-регрессионного анализа с использованием программы IBM SPSS Statistics

1 Создать в программе новый файл и скопировать в него все данные (без шапки таблицы) только из листа «Выборка» файла «07 Парный корреляционно-регрессионный анализ.xlsx».

2 Присвоить файлу при его сохранении название «07 Парный корреляционно-регрессионный анализ.sav».

3 В созданном файле:

– нажав в редакторе данных кнопку **«Переменные»**, перейти в одноименную вкладку и присвоить имена переменным созданной выборки (вместо пробелов использовать символ нижнего подчеркивания!);

– для всех переменных задать ширину в восемь символов без десятичных знаков после запятой, ширину колонки в восемь символов, выравнивание по центру и роль **«Входная»**;

– для всех переменных, кроме переменных, касающихся образования супругов, наличия подключения к интернету, вида телевидения, вида жилья и стиля мебели, задать тип шкалы **«Шкалы»**;

– для переменных, касающихся образования супругов, наличия подключения к интернету, вида телевидения, вида жилья и стиля мебели, задать шкалы так, как это было сделано в лабораторной работе № 5.

4 Нажав кнопку **«Данные»**, вернуться в одноименную вкладку редактора данных.

5 Выполнить парный корреляционно-регрессионный анализ для первой пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Количество членов домохозяйства», в которой первая является зависимой, а вторая – независимой. Для этого:

– выбрать процедуру корреляционно-регрессионного анализа **(«Анализ»** – **«Регрессия»** – **«Линейная…»**);

– в открывшемся диалоговом окне из поля, в котором представлены все переменные, нажав кнопку со стрелкой, направленной вправо, перенести в поле **«Зависимая переменная»** переменную «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели»;

– точно так же в поле **«Независимые переменные»** перенести переменную «Количество членов домохозяйства»;

– остальные поля не заполнять (рисунок 7.6);

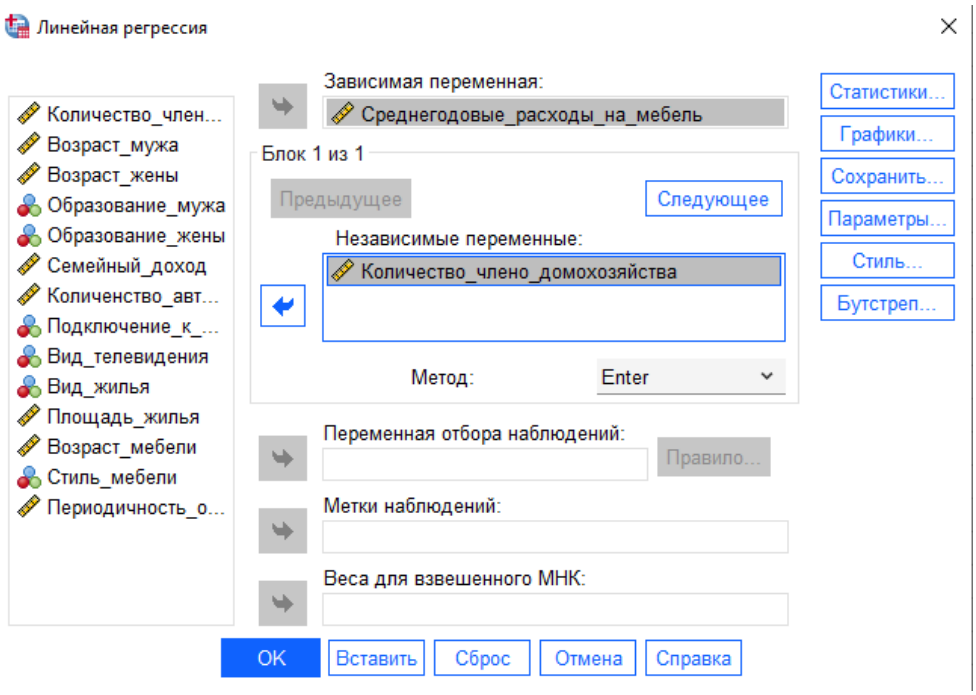



Рисунок 7.6 – Диалоговое окно «Линейная регрессия» с введенными именами зависимой и независимой переменных

– нажать кнопку **«Статистики…»** и в открывшемся диалоге выбрать расчет оценок коэффициентов регрессии, статистик согласия (множественного коэффициента корреляции и коэффициента детерминации, их скорректированных значений, таблицы дисперсионного анализа) и доверительных интервалов при уровне надежности 95 %, а также тест Дарбина – Уотсона (рисунок 7.7) и нажать кнопку **«Продолжить»**;

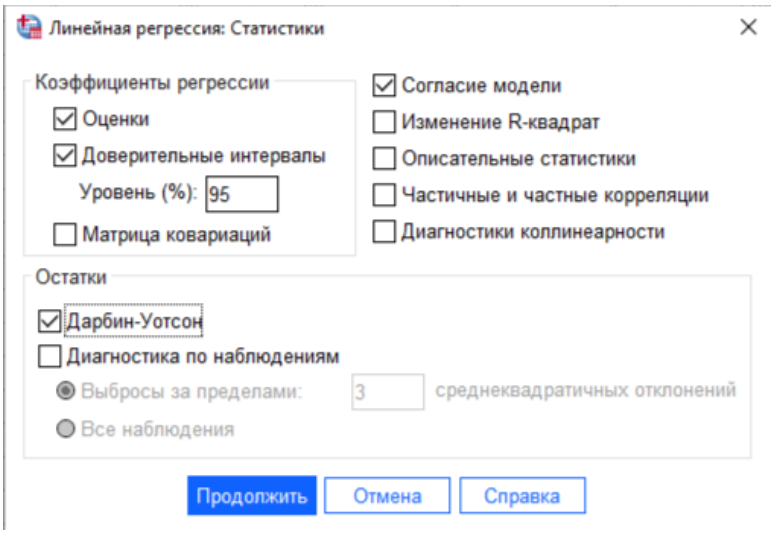



Рисунок 7.7 – Диалог «Линейная регрессия: Статистики» с заданными расчетами оценок коэффициентов регрессии и статистик согласия

– нажать кнопку **«Графики…»**, в открывшемся диалоге запросить вывод гистограммы с наложенным на нее графиком нормального распределения (рисунок 7.8) и нажать кнопку **«Продолжить»**;

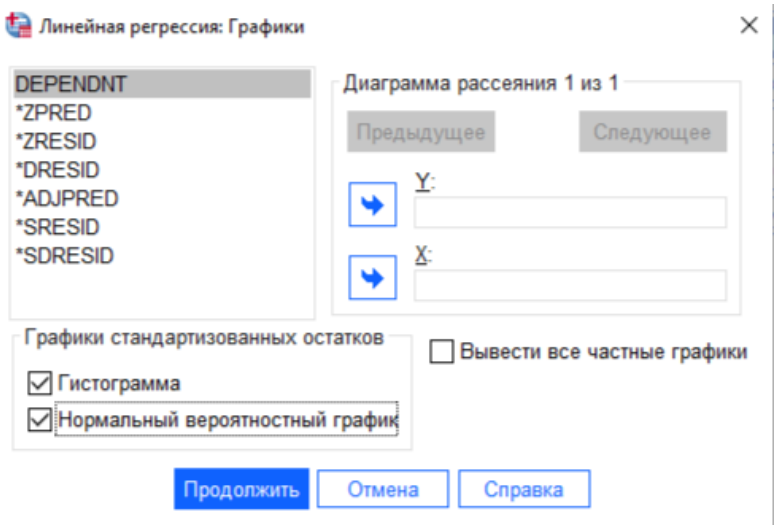



Рисунок 7.8 – Диалог «Линейная регрессия: Графики» с запросом вывести гистограмму и наложить на нее кривую нормального распределения

– нажать кнопку **«Сохранить…»**, в открывшемся диалоге запросить вывод нестандартизированных остатков (рисунок 7.9) и нажать кнопку **«Продолжить»**;

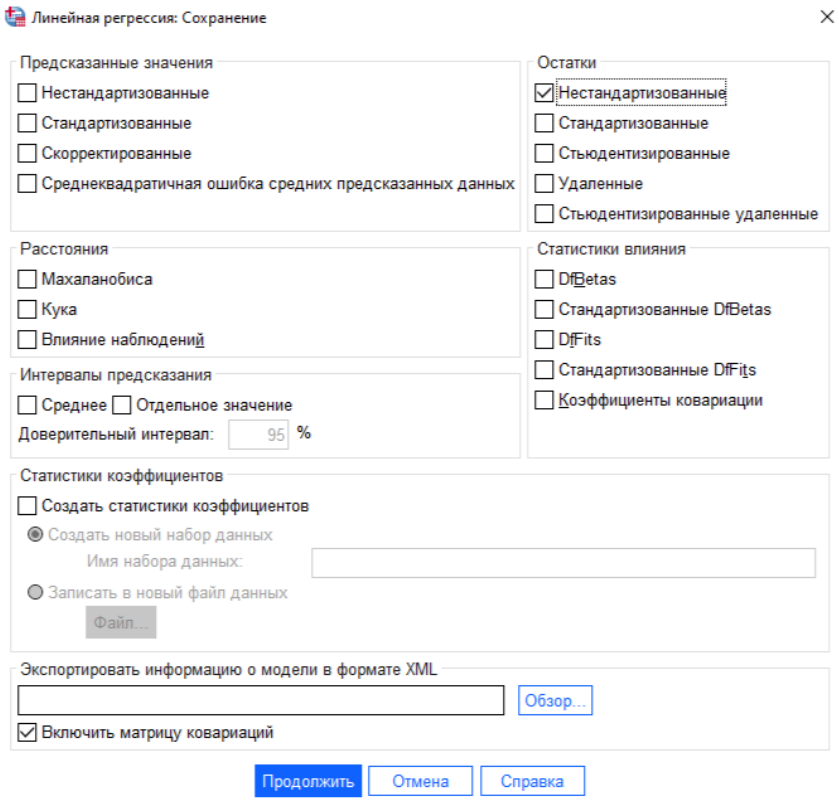



Рисунок 7.9 – Диалог «Линейная регрессия: Сохранить» с запросом вывода нестандартизированных остатков

– нажать кнопку **«Параметры»**, в открывшемся диалоге для критериев шагового отбора, оставив значимость *F*-критерия для порога включения переменной 0,05 и исключения 0,10, запросить включение в уравнение регрессии константы (свободного члена) (рисунок 7.10) и нажать кнопку **«Продолжить»**;

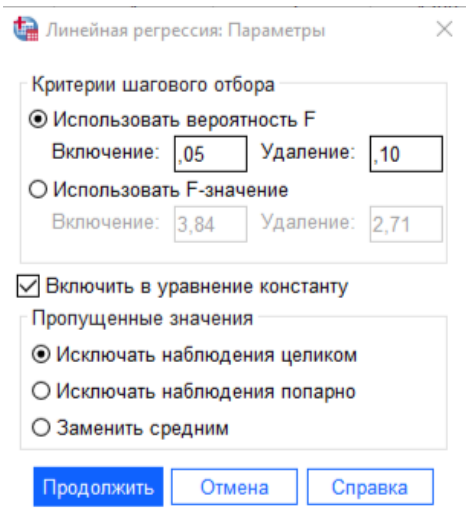



Рисунок 7.10 – Меню диалога «Линейная регрессия: Параметры» с заданными порогами для включения и исключения переменной, а также запросом включения в уравнение регрессии константы (свободного члена)

– в диалоговом окне **«Линейная регрессия»** нажать кнопку **«ОК»**.

В создавшемся файле с результатами выполненного анализа обратить внимание на таблицы «Сводка модели», «ANOVA» (дисперсионный анализ), «Коэффициенты» и гистограмму остатков с наложенной на нее кривой нормального распределения. Данные из этих таблиц сведены в таблицы 7.58–7.60 и полностью совпадают с данными из ранее выполненных таблиц 7.2–7.4.

Таблица 7.58 – Показатели регрессионной статистики для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Количество членов домохозяйства»

| Показатель | Значение |
|---|---|
| Парный коэффициент корреляции | 0,88 |
| Коэффициент детерминации | 0,78 |
| Скорректированный коэффициент детерминации | 0,78 |
| Стандартная ошибка оценки уравнения регрессии | 59,14 |

Значение теста Дарбина – Уотсона равно 1,75, что достаточно близко к 2,00 и говорит о возможности отсутствия автокорреляции, т. е. отклонения от теоретически возможных результатов (остатки) появляются случайным образом.

Таблица 7.59 – Результаты однофакторного дисперсионного анализа для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Количество членов домохозяйства»

| Показатель | Число степеней свободы, *df* | Сумма квадратов разностей, *SS* | Оценка дисперсий (средний квадрат), *MS* | Значение *F*-критерия | Значимость *F*-критерия |
|---|---|---|---|---|---|
| Регрессия | 1 | 18166118,08 | 18166118,08 | 5194,79 | 0,00 |
| Остаток | 1438 | 5028671,92 | 3496,99 | | |
| Итого | 1439 | 23194790,00 | | | |

Таблица 7.60 – Значения коэффициентов уравнения регрессии, оценок их статистической значимости и границ доверительных интервалов для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Количество членов домохозяйства»

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Свободный член, $\beta_0$ | 2,67 | 4,89 | 0,55 | 0,58 | –6,91 | 12,26 |

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Коэффициент при независимой переменной, $\beta_1$ | 83,72 | 1,16 | 72,07 | 0,00 | 81,44 | 86,00 |

Гистограмма остатков с наложенной на нее кривой нормального распределения (рисунок 7.11) снова показывает, что распределение остатков не подпадает под нормальное. Еще раз можно сделать вывод о том, что рассчитанное уравнение регрессии следует признать условно пригодным для прогнозирования расходов домохозяйств на покупку (обновление) элементов домашней мебели в зависимости от количества их членов.

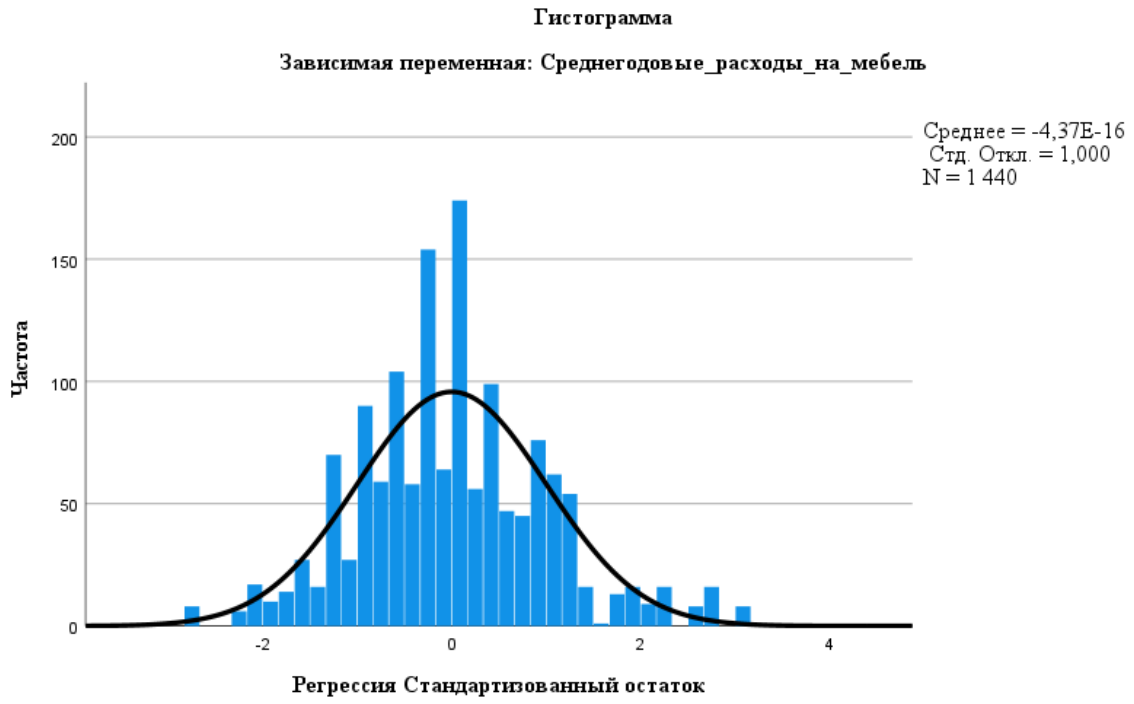



Рисунок 7.11 – Выполненная программой IBM SPSS Statistics гистограмма стандартизированных остатков с наложенной на нее кривой нормального распределения для пары переменных «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» и «Количество членов домохозяйства»

6 Вышеописанным образом выполнить парный корреляционно-регрессионный анализ для всех остальных пар переменных.

Результаты выполненных парных корреляционно-регрессионных анализов представлены в таблице 7.61.

Таблица 7.61 – Результаты парных корреляционно-регрессионных анализов, выполненных в приложении MS Excel и программе IBM SPSS Statistics

| Независимые переменные | Коэффициент корреляции | Уравнение регрессии | | |
|---|---|---|---|---|
| | | формула | статистическая значимость | стандартная ошибка оценки, *SEE*, р. |
| Количество членов домохозяйства | 0,88 | $y_x = 2{,}67 + 83{,}72x$ | да | 59,14 |
| Возраст мужа | 0,31 | $y_x = 473{,}42 - 2{,}98x$ | да | 120,92 |
| Возраст жены | 0,30 | $y_x = 470{,}60 - 3{,}11x$ | да | 121,07 |
| Образование мужа | 0,15 | $y_x = 256{,}57 + 26{,}52x$ | да | 125,59 |
| Образование жены | 0,20 | $y_x = 223{,}34 + 39{,}58x$ | да | 124,36 |
| Семейный среднемесячный доход | 0,999 | $y_x = 0{,}01 + 0{,}05x$ | да | 0,20 |
| Количество автомобилей в семье | 0,91 | $y_x = 139{,}68 + 166{,}04x$ | да | 52,02 |
| Наличие подключения к интернету | 0,20 | $y_x = 235{,}27 + 58{,}47x$ | да | 124,34 |
| Вид телевидения | 0,35 | $y_x = 211{,}17 + 95{,}03x$ | да | 119,00 |
| Вид жилья | 0,70 | $y_x = 832{,}74 - 265{,}28x$ | да | 90,37 |
| Площадь жилья | 0,86 | $y_x = 182{,}99 + 1{,}70x$ | да | 64,18 |
| Средний возраст мебели | 0,30 | $y_x = 469{,}37 - 11{,}87x$ | да | 121,21 |
| Стиль мебели | 0,94 | $y_x = 36{,}74 + 140{,}75x$ | да | 43,49 |
| Планируемая периодичность обновления мебели | 0,30 | $y_x = 471{,}60 - 10{,}92x$ | да | 121,32 |

Из таблицы 7.61 видно, что для рассмотренных зависимостей между значениями стандартной ошибки оценки и коэффициентом корреляции существует обратная связь: чем выше коэффициент корреляции, тем меньше стандартная ошибка.

## 7.3 Задание для самостоятельного выполнения

Исключив на листе «Расходы и ср.мес.доход» файла «07 Парный корреляционно-регрессионный анализ.xlsx» из списка 169 домохозяйств, для которых значения остатков оказались меньше –0,06, и 286 домохозяйств, для которых значения остатков оказались больше 0,38, выполнить парный корреляционно-регрессионный анализ для оставшихся 985 домохозяйств.

## 7.4 Вопросы для самоконтроля

1 Что представляет собой парный (однофакторный) корреляционно-регрессионный анализ и в каком порядке он проводится?

2 Что представляет собой корреляционное поле (поле корреляции)?

3 Как вычисляется простой (линейный) коэффициент корреляции и с помощью какого показателя оценивается его статистическая значимость?

4 Как вычисляется коэффициент детерминации и что он показывает?

5 Каким уравнением выражается форма прямой линии в модели парной (однофакторной) регрессии, в том числе и при учете вероятностной природы связи между переменными?

6 В чем заключается сущность метода наименьших квадратов, используемого при оценке параметров уравнения парной (однофакторной) регрессии?

7 Какие статистические гипотезы формулируются при проведении парного (однофакторного) корреляционно-регрессионного анализа?

8 Какой показатель используется для оценки точности значений зависимой переменной, предсказанных (расчетных, теоретических) с помощью уравнения парной (однофакторной) регрессии?

# ЛАБОРАТОРНАЯ РАБОТА № 8

## Множественный (многофакторный) корреляционно-регрессионный анализ данных, полученных по выборке в процессе маркетингового исследования

**Цель работы:** выполнить множественный (многофакторный) корреляционно-регрессионный анализ данных, характеризующих объекты исследования (домохозяйства), которые были включены в выборку, сформированную по итогам лабораторной работы № 5.

**Задачи работы:** с использованием теоретических знаний, полученных при изучении тем № 8, 10, 11, 13, 14 и 16 дисциплины, а также изученных ранее курсов «Прикладной статистический анализ» и «Теория вероятностей и математическая статистика»:

– изучить порядок выполнения множественного (многофакторного) корреляционно-регрессионного анализа данных, полученных по выборке в процессе маркетингового исследования;

– получить практические навыки в выполнении множественного (многофакторного) корреляционно-регрессионного анализа данных с использованием приложения MS Excel и программы IBM SPSS Statistics.

### 8.1 Теоретические сведения

#### 8.1.1 Основные термины

**Причинно-следственные отношения** – это связь явлений и процессов, когда изменение одного из них (причины) ведет к изменению другого (следствия).

**Причина** – это совокупность условий, обстоятельств, действие которых приводит к появлению следствия.

**Признак** – это основная отличительная черта, особенность изучаемого явления или процесса.

**Факторный признак** – это признак, обуславливающий изменения другого, связанного с ним результативного признака.

**Результативный признак** – это признак, изменяющийся под действием факторных признаков.

**Функциональная связь** – это связь, при которой определенному значению факторного признака соответствует одно и только одно значение результативного признака.

**Стохастическая связь** – это связь, которая проявляется не в каждом отдельном случае, а в общем, среднем или большем числе наблюдений.

**Корреляционная связь** – это связь, при которой изменение среднего значения результативного признака обуславливается изменением факторных признаков.

**Корреляция** – это статистическая зависимость между случайными величинами, которая не имеет строго функционального характера и при которой изменение одной из случайных величин приводит к изменению математического ожидания другой.

**Регрессионный анализ** – это аналитическое выражение связи, в которой изменение одной величины – результативного признака – обусловлено влиянием одного или нескольких факторных признаков, а множество всех остальных факторных признаков, также оказывающих влияние на зависимую величину, принимается за постоянные и средние значения.

**Мультиколлинеарность** – это тесная зависимость между факторными признаками, включенными в модель регрессии.

**Автокорреляция** – это статистическая взаимосвязь между последовательными уровнями временного ряда. Может быть как положительной, так и отрицательной. Свидетельствует о постоянном действии неучтенных факторов на результат: положительная – об однонаправленном, отрицательная – о разнонаправленном.

**Критерий Фишера – Снедекора (критерий Фишера**, ***F*-критерий**, ***F*-тест)** – это статистический критерий, тестовая статистика которого при выполнении нулевой гипотезы имеет распределение Фишера (*F*-распределение). Чтобы статистика имела распределение Фишера, необходимо, чтобы числитель и знаменатель были независимыми случайными величинами и соответствующие суммы квадратов имели распределение $\chi^2$.

**Критерий Стьюдента (*t*-статистика**, ***t*-тест)** – это общее название класса методов статистической проверки гипотез (статистических критериев), основанных на распределении Стьюдента.

**Тест Дарбина – Уотсона** – это тест на автокорреляцию, которая выражается в наличии систематических связей между остатками, которые представляют собой отклонения наблюдаемых значений от теоретически ожидаемых и вычисленных с помощью уравнения регрессии.

### 8.1.2 Множественный (многофакторный) корреляционно-регрессионный анализ

#### 8.1.2.1 Множественный (многофакторный) корреляционный анализ

Множественный (многофакторный) корреляционный анализ имеет своей задачей количественное определение тесноты связи между результативным и множеством (двумя и более) факторных признаков. Теснота связи количественно выражается величиной коэффициента множественной корреляции, которая дает возможность определить «полезность» факторных признаков при построении уравнения множественной регрессии. Величина коэффициента корреляции служит также оценкой соответствия уравнения множественной регрессии

причинно-следственным связям, выявленным в ходе описательного (дескриптивного) маркетингового исследования.

Множественный коэффициент корреляции рассчитывается при наличии линейной связи между результативным и несколькими факторными признаками, а также между каждой парой факторных признаков.

Множественный коэффициент корреляции рассчитывается по формуле

$$r_{y/x_1x_2\ldots x_k} = \sqrt{\frac{SS_{x_1x_2\ldots x_k}}{SS_y}}, \tag{8.1}$$

где $SS_y$ – полная вариация результативного признака как сумма квадратов разниц между фактическими значениями зависимой переменной и ее средним значением в выборке (см. формулу (7.4) лабораторной работы № 7);

$SS_{x_1x_2\ldots x_k}$ – полная вариация теоретических значений результативного признака, рассчитанная по уравнению регрессии как сумма квадратов разниц между фактическими и расчетными значениями результативного признака (см. формулу (7.5) лабораторной работы № 7).

При небольшом количестве наблюдений величина коэффициента множественной корреляции, как правило, завышается.

В случае если $\frac{n-k}{k} < 20$, величина коэффициента множественной корреляции корректируется с использованием следующего выражения:

$$\hat{r}_{y/x_1x_2\ldots x_k} = r_{y/x_1x_2\ldots x_k}\frac{n-1}{n-k-1}, \tag{8.2}$$

где $n$ – число наблюдений в выборке;

$k$ – число факторных признаков.

Соответственно, как и при парной корреляции, $d = r^2_{y/x_1x_2\ldots x_k}$.

Проверка значимости коэффициента множественной корреляции осуществляется на основе критерия Фишера – Снедекора, вычисляемого по формуле

$$F = \frac{(n-k)r^2_{y/x_1x_2\ldots x_k}}{(k-1)\left(1-r^2_{y/x_1x_2\ldots x_k}\right)}. \tag{8.3}$$

В случае если $F_{\text{расч}} > F_{\text{крит}}$, нулевая гипотеза $H_0: r_{y/x_1x_2\ldots x_k} = 0$ (или $H_0: d = 0$) отвергается и принимается альтернативная.

Частные коэффициенты корреляции характеризуют степень тесноты связи между двумя факторными признаками $x_1$ и $x_2$ при фиксированном значении других $(k-2)$ факторных признаков.

Коэффициент, в котором исключается влияние только одного факторного признака, называется коэффициентом частной корреляции первого порядка. В случае зависимости $y$ от двух факторных признаков $x_1$ и $x_2$ частные коэффициенты корреляции вычисляются следующим образом:

$$r_{yx_1/x_2} = \frac{r_{yx_1} - r_{x_1x_2}r_{yx_2}}{\sqrt{\left(1 - r_{x_2y}^2\right)\left(1 - r_{x_1x_2}^2\right)}}, \tag{8.4}$$

$$r_{yx_2/x_1} = \frac{r_{yx_2} - r_{x_1x_2}r_{yx_1}}{\sqrt{\left(1 - r_{x_1y}^2\right)\left(1 - r_{x_1x_2}^2\right)}}. \tag{8.5}$$

Проверка значимости коэффициентов частной корреляции проводится с использованием *F*-критерия:

$$F = \frac{SS_y}{SS_y - SS_x}(n - k)\,. \tag{8.6}$$

Частичный коэффициент корреляции представляет собой корреляцию между *y* и $x_1$, когда линейные эффекты других независимых переменных исключены из $x_1$, но не из *y*. В случае зависимости *y* от двух факторных признаков он вычисляется по формуле

$$r_{y(x_1x_2)} = \frac{r_{x_1y} - r_{x_1x_2}r_{yx_2}}{\sqrt{1 - r_{x_1x_2}^2}}\,. \tag{8.7}$$

#### 8.1.2.2 Множественный (многофакторный) регрессионный анализ

При исследовании зависимостей методами множественной регрессии задача формулируется так же, как и при использовании парной регрессии, т. е. требуется определить аналитическое выражение связи между результативным признаком *y* и факторными признаками $(x_1, x_2, \ldots, x_k)$, т. е. найти функцию

$$\hat{y} = f(x_1, x_2, \ldots, x_k). \tag{8.8}$$

Построение модели множественной регрессии включает этапы, показанные на рисунке 8.1.

##### 8.1.2.2.1 Выбор формы регрессионной связи

Выбор формы регрессионной связи затрудняется тем, что, используя математический аппарат, теоретическая зависимость между признаками может быть выражена большим числом различных уравнений. Уравнение регрессии строится главным образом для объяснения и количественного выражения взаимосвязей, оно должно хорошо отражать сложившиеся между исследуемыми признаками фактические связи.


Рисунок 8.1 – Примерный порядок построения модели множественной регрессии

Наиболее приемлемым способом определения вида исходного уравнения регрессии является достаточно трудоемкий метод перебора различных уравнений. Его сущность заключается в том, что достаточно большое число уравнений регрессии, отобранных для описания связей исследуемого маркетологами процесса (явления), реализуется с использованием алгоритма перебора и соответствующего программного обеспечения с последующей статистической проверкой, главным образом, на основе критериев Стьюдента и Фишера – Снедекора.

Почти все реально существующие зависимости между социально-экономическими явлениями можно описать, используя пять типов моделей:

1) линейную: $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$;

2) степенную: $\hat{y} = \beta_0 x_1^{\beta_1} x_2^{\beta_2}, \ldots x_k^{\beta_k}$;

3) показательную: $\hat{y} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}$;

4) параболическую: $\hat{y} = \beta_0 + \beta_1 x_1^2 + \beta_2 x_2^2 + \cdots + \beta_k x_k^2$;

5) гиперболическую: $\hat{y} = \beta_0 + \frac{\beta_1}{x_1} + \frac{\beta_2}{x_2} + \cdots + \frac{\beta_k}{x_k}$.

Основное значение имеют линейные модели в силу простоты и логичности их интерпретации. Нелинейные формы зависимости приводятся к линейным путем линеаризации.

Сложность формирования уравнения множественной регрессии во время отбора и последующего включения факторных признаков заключается в том, что почти все факторные признаки находятся в зависимости друг от друга. Определение оптимального числа факторных признаков для построения модели взаимосвязи может быть выполнено на основе эвристических (экспертных) оценок или многомерных статистических методов анализа (например, факторного или кластерного).

Анализ экспертной информации проводится на базе расчета и анализа непараметрических показателей связи: ранговых коэффициентов корреляции Спирмена и Кендалла, а также коэффициента конкордации.

#### 8.1.2.2.2 Отбор факторных признаков

Наиболее приемлемым способом отбора факторных признаков является шаговая регрессия (шаговый регрессионный анализ), которая предполагает последовательное включение (прямой метод) или исключение (обратный метод) факторов в уравнение регрессии с последующей проверкой их значимости.

При прямом методе факторы поочередно вводятся в уравнение регрессии, причем при проверке значимости каждого вводимого фактора определяется, насколько уменьшается сумма квадратов остатков $SS_e$ и увеличивается величина множественного коэффициента корреляции $r_{y/x_1x_2\ldots x_k}$. Также факторный признак признается существенным и его включение в уравнение регрессии является необходимым, если при этом величина множественного коэффициента корреляции $r_{y/x_1x_2\ldots x_k}$ увеличивается, а коэффициенты регрессии не изменяются (или меняются несущественно).

Обратный метод предполагает исключение из уравнения регрессии факторов, признанных незначимыми на основе значения критерия Стьюдента. Фактор также признается незначимым, если его включение в уравнение регрессии только изменяет значение коэффициентов регрессии, но не меняет значения суммы квадратов остатков $SS_e$. Также факторный признак признается нецелесообразным для включения в уравнение регрессии, если при этом коэффициенты регрессии меняют не только величину, но и знаки, а множественный коэффициент корреляции $r_{y/x_1x_2\ldots x_k}$ не возрастает.

Сложность и взаимное переплетение отдельных факторов, обуславливающих исследуемое явление (процесс), могут проявляться в так называемой мультиколлинеарности. Причинами возникновения мультиколлинеарности между факторными признаками являются:

1) изучаемые факторные признаки характеризуют одну и ту же сторону явления или процесса;

2) суммарное значение факторных показателей представляет собой постоянную величину;

3) факторные признаки являются составными элементами друг друга;

4) факторные признаки по экономическому смыслу дублируют друг друга.

Одним из индикаторов установления наличия мультиколлинеарности между факторными признаками является превышение парным коэффициентом корреляции $r_{x_ix_j}$ величины 0,8.

Устранение мультиколлинеарности может реализовываться через исключение из регрессионной модели одного или нескольких линейно-связанных факторных признаков или, как это будет показано в лабораторной работе № 11, через преобразование с использованием методов факторного анализа факторных признаков в новые, укрупненные.

Вопрос о том, какой из факторов следует отбросить, решается на основании качественного и логического анализов изучаемого явления (процесса).

Качество уравнения регрессии зависит от степени достоверности и надежности данных и объема совокупности. Маркетолог-исследователь должен стремиться к увеличению числа наблюдений, так как большой объем наблюдений является одной из предпосылок построения адекватных статистических моделей.

#### 8.1.2.2.3 Нормирование параметров уравнения регрессии

Как и в случае с парной регрессией, при множественном регрессионном анализе может применяться процедура нормирования, при которой исходные данные (значения факторных признаков) преобразуют в новые переменные со значением средней, равным нулю, и дисперсией, равной 1,0. Связано это с тем, что факторные признаки могут быть различны по своей сущности и иметь различные единицы измерения.

Бета-коэффициенты являются частными коэффициентами многомерной регрессии, полученными после того, как перед оценкой уравнения регрессии все переменные $y, x_1, x_2, \dots, x_k$ нормированы. Связь между нормированными и ненормированными коэффициентами та же, что и рассмотренная ранее в лабораторной работе № 7:

$$\mathrm{B}_{yx_1} = \beta_1 \left( \frac{S_{x_1}}{S_y} \right), \tag{8.9}$$

$$\mathrm{B}_{yx_k} = \beta_k \left( \frac{S_{x_k}}{S_y} \right). \tag{8.10}$$

Отрезок, отсекаемый на оси *y*, и частные коэффициенты регрессии определяют решением системы уравнений, выведенной дифференцированием и приравниванием к нулю частных производных с помощью различных компьютерных программ (например, MatLab, Mathematica, SciLab, MathCAD и др.). Но необходимо помнить, что уравнения нельзя решить, если размер выборки *n* меньше или равен числу независимых переменных *k*, или одна независимая переменная тесно связана с другой.

#### 8.1.2.2.4 Проверка значимости уравнения множественной регрессии

Проверка значимости уравнения множественной регрессии включает проверку значимости общего уравнения регрессии и конкретных частных коэффициентов регрессии. Нулевая гипотеза для проверки общего уравнения гласит, что коэффициент множественной детерминации для генеральной совокупности равен нулю:

$$H_0 : r^2_{y/x_1 x_2 \dots x_k} = 0. \tag{8.11}$$

Это эквивалентно следующей нулевой гипотезе:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0. \quad (8.12)$$

Общую проверку можно выполнить, используя *F*-критерий:

$$F = \frac{SS_x(n-k-1)}{SS_e k} \text{ или } F = \frac{r^2_{y/x_1x_2\ldots x_k}(n-k-1)}{\left(1-r^2_{y/x_1x_2\ldots x_k}\right)k}, \quad (8.13)$$

который имеет *F*-распределение с $k$ и $(n-k-1)$ степенями свободы.

Если общую нулевую гипотезу отклоняют, то один или несколько частных коэффициентов регрессии в совокупности имеют значение, отличное от нуля.

#### 8.1.2.2.5 Оценка точности уравнения регрессии

Для оценки точности предсказанных (теоретических) значений *y* также необходимо вычислить стандартную ошибку оценки уравнения регрессии *SEE*. При наличии $k$ независимых переменных этот показатель также представляет собой стандартное отклонение фактических значений *y* от предсказанных значений $\hat{y}_i$:

$$SEE = \sqrt{\frac{SS_e}{n-k-1}}. \quad (8.14)$$

*SEE* и здесь интерпретируется как вид среднего значения остатка или средняя ошибка предсказания *y* исходя из уравнения регрессии.

## 8.2 Выполнение множественного (многофакторного) корреляционно-регрессионного анализа с использованием приложения MS Excel и программы IBM SPSS Statistics

В файле «05 Результаты выборочного наблюдения.xlsx» приведены данные о домохозяйствах, которые были включены в выборку, исследуемую совместно сотрудниками маркетинговых подразделений ОАО «Крессида» и ЧУП «Кэтнес».

Необходимо с их использованием выполнить множественный корреляционно-регрессионный анализ для всех выбранных переменных. При этом исходить из того, что переменная «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» является зависимой, а все остальные 14 – независимыми.

### 8.2.1 Выполнение множественного (многофакторного) корреляционно-регрессионного анализа с использованием приложения MS Excel

1 Файл «05 Результаты выборочного наблюдения.xlsx», с помощью которого выполнялась лабораторная работа № 5, скопировать в папку, в которой будут находиться файлы текущей работы, и присвоить ему имя «08 Множественный корреляционно-регрессионный анализ.xlsx». В созданном файле листы «Основа выборочного наблюдения» и «Среднегодовые расходы» рекомендуется удалить.

2 Для выполнения работы необходимо выдвинуть следующие статистические гипотезы:

– нулевую, согласно которой зависимости между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и остальными 14 переменными не существует, а если она и существует, то является статистически незначимой ($H_0: r^2 = 0;\ d^2 = 0;\ \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_{14} = 0$);

– альтернативную, согласно которой зависимость между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и остальными 14 переменными существует, является не случайной и статистически значимой ($H_1: r^2 \neq 0;\ d^2 \neq 0;\ \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 \neq 0, \dots, \beta_{14} \neq 0,$).

3 В созданном файле выполнить следующие действия:

– выбрать инструмент корреляционно-регрессионного анализа **(«Данные»** – **«Анализ данных»** – **«Регрессия»**) (рисунок 8.2);

– ввести в поле **«Входной интервал Y»** значения примерных среднегодовых расходов на покупку (обновление) элементов домашней мебели по всей выборке (ячейки **«P2»**–**«P1442»** из листа «Выборка»);

– ввести в поле **«Входной интервал X»** значения всех 14 независимых переменных по всей выборке (ячейки **«B2»**–**«O1442**» из листа «Выборка»);

– поставить флажок напротив строки **«Метки»**, уровень надежности оставить равным 95 %;

– в секции **«Параметры вывода»** поставить метку напротив строки **«Новый рабочий лист»**;

– поставить флажок напротив строки **«Остатки»** и нажать кнопку **«ОК»**;

– созданный лист переместить правее листа «Выборка» и присвоить ему имя «Расходы и все переменные»;

– в созданном листе для всех таблиц задать шрифт Times New Roman Cyr размером 12 пт и изменить ширину колонок так, чтобы названия всех переменных и их величин были полностью видны. В случае, если в третьей таблице колонки **«Нижние 95%»** и **«Верхние 95 %»** повторились, их можно удалить.

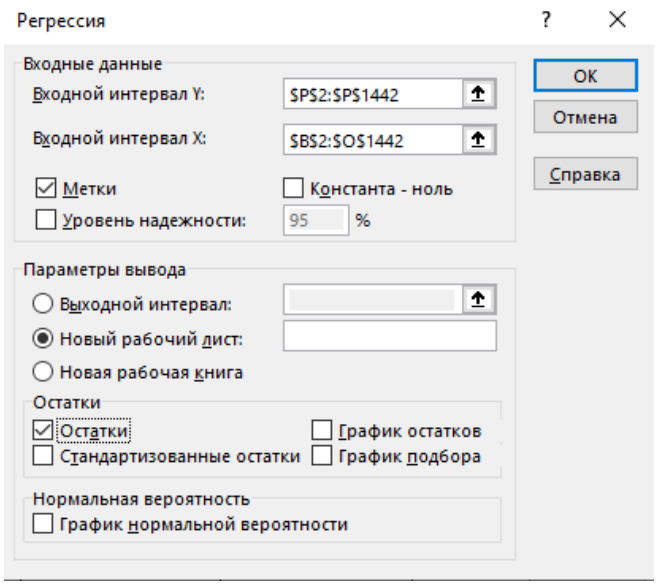



Рисунок 8.2 – Диалог «Регрессия» с внесенными значениями зависимой и независимых переменных по выборке

4 В результате выполнения корреляционно-регрессионного анализа будут получены значения, которые приведены в таблицах 8.1–8.3. Таблица с рассчитанными разницами (остатками) между вычисленными с использованием уравнения регрессии значениями примерных среднегодовых расходов на покупку (обновление) элементов домашней мебели и эмпирическими (фактическими по выборке) значениями по этой переменной здесь не приводится по причине большого количества в ней строк (1440).

Таблица 8.1 – Показатели регрессионной статистики для случая, когда в анализ включен весь набор независимых переменных

| Показатель | Значение |
|---|---|
| Множественный коэффициент корреляции | 0,999999 |
| Коэффициент детерминации | 0,999998 |
| Нормированный коэффициент детерминации | 0,999998 |
| Стандартная ошибка | 0,19 |
| Количество наблюдений | 1440 |

Таблица 8.2 – Результаты многофакторного дисперсионного анализа для случая, когда в анализ включен весь набор независимых переменных

| Показатель | Число степеней свободы, *df* | Сумма квадратов разностей, *SS* | Оценка дисперсий (средний квадрат), *MS* | Значение *F*-критерия | Значимость *F*-критерия |
|---|---|---|---|---|---|
| Регрессия | 14 | 23194738,01 | 1656767,00 | 45415525,62 | 0,00 |
| Остаток | 1425 | 51,98 | 0,04 | | |
| Итого | 1439 | 23194790,00 | | | |

Таблица 8.3 – Значения рассчитанных коэффициентов уравнения регрессии, оценок их статистической значимости и границ доверительных интервалов для случая, когда в анализ включен весь набор независимых переменных

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Свободный член, $\beta_0$ | 2,45 | 0,39 | 6,25 | 0,00 | 1,68 | 3,21 |
| Коэффициент при независимой переменной, $\beta_1$ | 0,08 | 0,04 | 2,21 | 0,03 | 0,01 | 0,15 |
| Коэффициент при независимой переменной, $\beta_2$ | –0,01 | 0,00 | –1,47 | 0,14 | –0,01 | 0,00 |
| Коэффициент при независимой переменной, $\beta_3$ | 0,02 | 0,00 | 5,03 | 0,00 | 0,01 | 0,03 |
| Коэффициент при независимой переменной, $\beta_4$ | –0,03 | 0,01 | –2,37 | 0,02 | –0,05 | 0,00 |
| Коэффициент при независимой переменной, $\beta_5$ | –0,04 | 0,01 | –3,21 | 0,00 | –0,06 | –0,02 |
| Коэффициент при независимой переменной, $\beta_6$ | 0,05 | 0,00 | 2960,69 | 0,00 | 0,05 | 0,05 |
| Коэффициент при независимой переменной, $\beta_7$ | 0,07 | 0,02 | 3,32 | 0,00 | 0,03 | 0,11 |
| Коэффициент при независимой переменной, $\beta_8$ | –0,07 | 0,02 | –3,04 | 0,00 | –0,11 | –0,02 |
| Коэффициент при независимой переменной, $\beta_9$ | 0,02 | 0,02 | 0,85 | 0,39 | –0,02 | 0,06 |
| Коэффициент при независимой переменной, $\beta_{10}$ | –1,04 | 0,19 | –5,35 | 0,00 | –1,42 | –0,66 |
| Коэффициент при независимой переменной, $\beta_{11}$ | –0,01 | 0,00 | –6,27 | 0,00 | –0,01 | –0,01 |

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Коэффициент при независимой переменной, $\beta_{12}$ | 0,01 | 0,04 | 0,27 | 0,79 | –0,07 | 0,09 |
| Коэффициент при независимой переменной, $\beta_{13}$ | –0,02 | 0,02 | –1,12 | 0,26 | –0,06 | 0,02 |
| Коэффициент при независимой переменной, $\beta_{14}$ | –0,07 | 0,02 | –3,00 | 0,00 | –0,12 | –0,02 |

Значения, приведенные в таблицах 8.1–8.3, позволяют заключить следующее:

– на основании величин множественного коэффициента корреляции, значимости *F*-критерия (указывающего на то, что его расчетное значение превышает табличное) следует принять альтернативную гипотезу;

– на основе *P*-значения можно заключить, что статистически значимыми являются коэффициенты при следующих 10 независимых переменных: «Количество членов домохозяйства», «Возраст жены», «Образование мужа», «Образование жены», «Семейный среднемесячный доход», «Количество автомобилей в семье», «Наличие подключения к интернету», «Вид жилья», «Площадь жилья» и «Планируемая периодичность обновления мебели». При остальных четырех независимых переменных – «Возраст мужа», «Вид телевидения», «Средний возраст мебели» и «Стиль мебели» – значения коэффициентов статистически незначимы;

– зависимость между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и рассматриваемыми 14 независимыми переменными существует и является статистически значимой. Значение множественного коэффициента корреляции, равное примерно 1,00, указывает на то, что связь по своему характеру является сильной и почти функциональной. Значение коэффициента детерминации, равное примерно 1,00, говорит о том, что примерно 100 % вариации зависимой переменной «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» объясняется вариацией рассматриваемых независимых переменных;

– уравнение регрессии имеет вид

$$y_x = 2{,}45 + 0{,}08x_1 - 0{,}01x_2 + 0{,}02x_3 - 0{,}03x_4 - 0{,}04x_5 + 0{,}05x_6 + \\ + 0{,}07x_7 - 0{,}07x_8 + 0{,}02x_9 - 1{,}04x_{10} - 0{,}01x_{11} + 0{,}01x_{12} - \\ - 0{,}02x_{13} - 0{,}07x_{14}.$$

Из него видно, что увеличение значения, например, независимой переменной «Количество членов домохозяйства» на одну единицу при условии, что значения остальных переменных остаются неизменными, влечет увеличение значения зависимой переменной «Примерные среднегодовые расходы на покупку (обновление) элементов мебели» примерно на 8 коп. и наоборот.

Значения границ доверительного интервала для каждой независимой переменной интерпретируются так же, как и при парном корреляционно-регрессионном анализе.

Что касается свободного члена, то его значение статистически значимо, о чем говорят не только значение критерия Стьюдента, превышающее табличное (что подтверждает *P*-значение, меньшее 0,05), но и границы его доверительного интервала, внутри которого отсутствует значение «ноль».

Как видно из таблицы 8.1, стандартная ошибка оценки уравнения регрессии *SEE* равна 0,19 р.

5 На листе «Расходы и все переменные» рассчитать описательные статистики для установленных разниц (остатков) между вычисленными на основе уравнения регрессии и эмпирическими (фактическими) по выборке значениями переменной «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели». Для этого, используя навыки, полученные при выполнении лабораторных работ № 5 и 7 (рисунок 8.3):

– выбрать инструмент анализа **«Описательная статистика» («Данные»** – **«Анализ данных»** – **«Описательная статистика»**);

– ввести в поле **«Входной интервал»** значения остатков (ячейки **«С37»**– **«С1477»**), группирование выбрать по столбцам. Так как в этом поле будет находиться название изучаемой характеристики, поставить флажок напротив строки **«Метки в первой строке»**;

– в секции **«Параметры вывода»** выбрать **«Выходной интервал»** и с использованием мыши ввести в поле ссылку на ячейку **«К1»**;

– поставить флажок напротив строки **«Итоговая статистика»**, уровень надежности оставить равным 95 % и нажать кнопку **«ОК»**;

– используя шрифт Times New Roman Cyr размером 12 пт, отформатировать полученные данные;

– вычисленные программой значения представлены в таблице 8.4.

Так как оценки существенности асимметрии и эксцесса зависят только от размера выборки (а она остается такой же), их уже рассчитанные величины взять из решения задачи парного корреляционно-регрессионного анализа в лабораторной работе № 7.

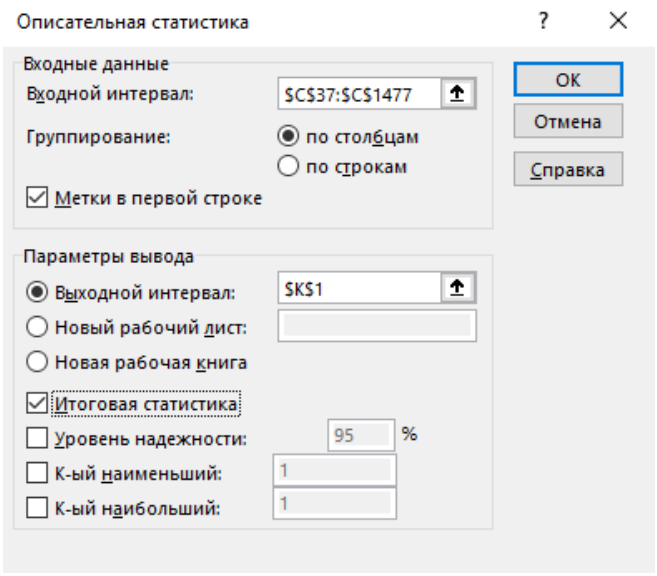



Рисунок 8.3 – Диалог инструмента анализа «Описательная статистика» с внесенными значениями остатков для переменной «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели»

Таблица 8.4 – Значения описательных статистик остатков уравнения регрессии для случая, когда в него включен весь набор независимых переменных

| Статистики | Значения |
|---|---|
| Среднее | 0,00 |
| Стандартная ошибка | 0,01 |
| Медиана | –0,08 |
| Мода | 0,36 |
| Стандартное отклонение | 0,19 |
| Дисперсия выборки | 0,04 |
| Эксцесс | 0,20 |
| Асимметричность | 1,31 |
| Интервал | 0,76 |
| Минимум | –0,25 |
| Максимум | 0,51 |
| Сумма | 0,00 |
| Счет | 1440 |

Так как отношение показателя асимметрии к ее средней квадратичной ошибке $\frac{|A_S|}{S_{A_S}} = \frac{1{,}31}{0{,}06} = 20{,}43$, что намного превосходит 3,0, то следует признать, что асимметрия распределения остатков существенна и распределение признака в генеральной совокупности несимметрично.

Положительное значение статистики эксцесса говорит об островершинности рассматриваемого ряда значений. Значение $\frac{|\varepsilon_k|}{S_{\varepsilon_k}} = \frac{0{,}20}{0{,}13} = 1{,}58$, что почти в два

раза меньше 3,0. Если бы рассматриваемое распределение остатков было бы симметрично или близко к нему, то его можно было бы признать нормальным или близким к нему.

6 С использованием полученных значений построить гистограмму распределения остатков. Для этого:

– в ранее созданном листе «Расходы и все переменные» справа от таблицы с вычисленными описательными статистиками, с учетом того, что минимальное значение остатков равно –0,25, а максимальное – 0,51, создать 18 интервалов группирования (карманов) с шагом в 0,05 р. Номера карманов с первого по восемнадцатый разместить в ячейках **«N2»–«N19»**, а значения их правых границ – в ячейках **«O2»–«O19»**;

– выбрать инструмент анализа **«Гистограмма»** (**«Данные»** – **«Анализ данных»** – **«Гистограмма»**) и заполнить его диалоговое окно, внеся во входной интервал значения остатков (ячейки **«C37»–«C1477»**), а ниже – карманов (ячейки **«O1»–«O19»**). Поставив флажок напротив строки **«Метки в первой строке»**, указав в качестве выходного интервала ячейку **«Q1»** и поставив флажок напротив строки **«Вывод графика»**, нажать кнопку **«ОК»**;

– используя шрифт Times New Roman Cyr размером 12 пт, отформатировать появившуюся таблицу и убрать в ней строку **«Еще»**;

– используя этот же шрифт, отформатировать построенную гистограмму;

– итог выполненных действий представлен на рисунке 8.4.

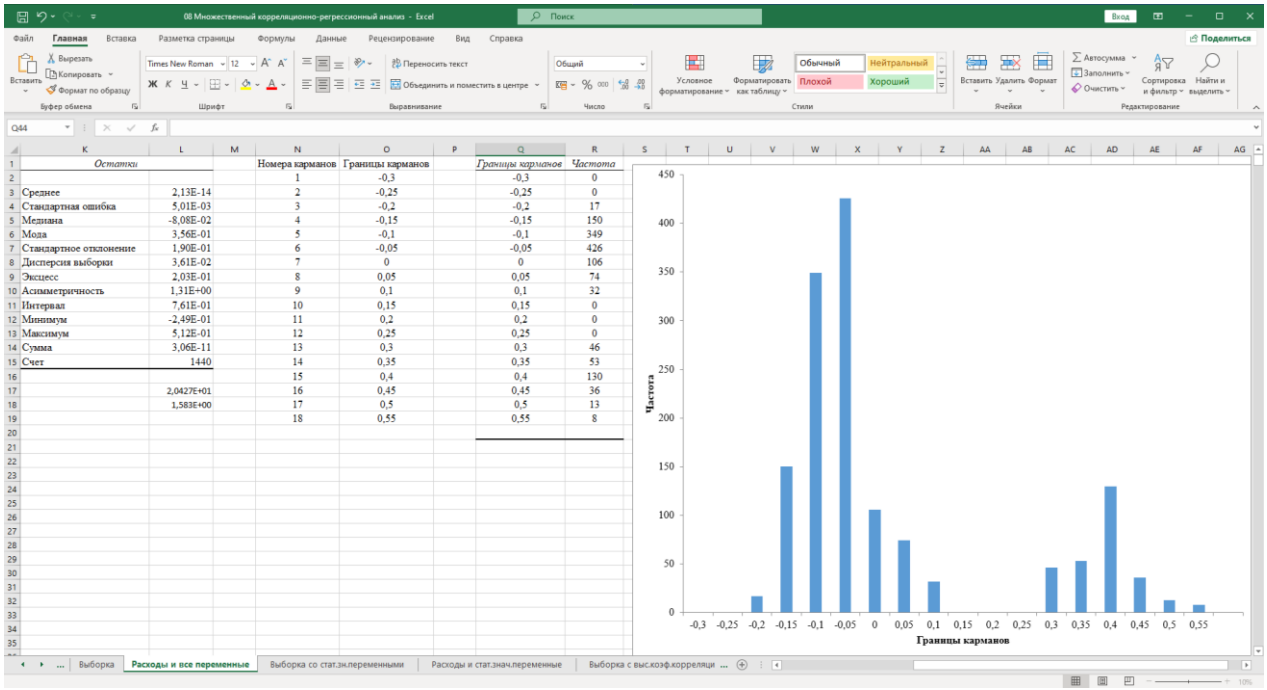



Рисунок 8.4 – Рассчитанные статистики и гистограмма остатков уравнения регрессии для случая, когда в анализ включен весь набор независимых переменных

Как видно из гистограммы, распределение остатков не подпадает под нормальное, а сама гистограмма может быть признана двухвершинной. Так как свойства коэффициентов регрессии существенным образом зависят от свойств остатков, а их распределение должно быть не только независимо от распределения переменных, но и нормально, рассчитанное уравнение регрессии следует признать условно пригодным для прогнозирования расходов домохозяйств на покупку (обновление) элементов домашней мебели в зависимости от количества их членов.

7 Выполнить множественный корреляционно-регрессионный анализ для случая, когда в набор независимых переменных включены только те, у которых коэффициент при них оказался статистически значимым (*P*-значение в этом случае меньше 0,05). Для этого:

– создать в файле «08 Множественный корреляционно-регрессионный анализ.xlsx» лист с именем «Расходы и стат.знач.переменные»;

– скопировать в него все данные из листа «Выборка» (ячейки **«A1»**–**«P1442»**);

– в созданном листе удалить столбцы с переменными, коэффициенты при которых являются статистические незначимыми («Возраст мужа», «Вид телевидения», «Средний возраст мебели» и «Стиль мебели»);

– выполнить множественный корреляционно-регрессионный анализ, введя в поля его диалога значения зависимой и оставшихся независимых переменных в описанном ранее порядке.

Результаты выполненного анализа приведены в таблицах 8.5–8.7.

Таблица 8.5 – Показатели регрессионной статистики для случая, когда в анализ включены независимые переменные, имеющие статистически значимые коэффициенты

| Показатель | Значение |
|---|---|
| Множественный коэффициент корреляции | 0,999999 |
| Коэффициент детерминации | 0,999998 |
| Нормированный коэффициент детерминации | 0,999998 |
| Стандартная ошибка | 0,19 |
| Количество наблюдений | 1440 |

Таблица 8.6 – Результаты многофакторного дисперсионного анализа для случая, когда в анализ включены независимые переменные, имеющие статистически значимые коэффициенты

| Показатель | Число степеней свободы, *df* | Сумма квадратов разностей, *SS* | Оценка дисперсий (средний квадрат), *MS* | Значение *F*-критерия | Значимость *F*-критерия |
|---|---|---|---|---|---|
| Регрессия | 10 | 23194737,79 | 2319473,78 | 63492346,65 | 0,00 |
| Остаток | 1429 | 52,20 | 0,04 | | |
| Итого | 1439 | 23194790,00 | | | |

Таблица 8.7 – Значения рассчитанных коэффициентов уравнения регрессии, оценок их статистической значимости и границ доверительных интервалов для случая, когда в анализ включены независимые переменные, имеющие статистически значимые коэффициенты

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Свободный член, $\beta_0$ | 2,45 | 0,39 | 6,30 | 0,00 | 1,69 | 3,21 |
| Коэффициент при независимой переменной, $\beta_1$ | 0,07 | 0,03 | 2,14 | 0,03 | 0,01 | 0,14 |
| Коэффициент при независимой переменной, $\beta_2$ | 0,02 | 0,00 | 5,97 | 0,00 | 0,01 | 0,03 |
| Коэффициент при независимой переменной, $\beta_3$ | –0,02 | 0,01 | –2,11 | 0,03 | –0,04 | 0,00 |
| Коэффициент при независимой переменной, $\beta_4$ | –0,04 | 0,01 | –3,32 | 0,00 | –0,06 | –0,01 |
| Коэффициент при независимой переменной, $\beta_5$ | 0,05 | 0,00 | 3254,72 | 0,00 | 0,05 | 0,05 |
| Коэффициент при независимой переменной, $\beta_6$ | 0,06 | 0,02 | 2,96 | 0,00 | 0,02 | 0,09 |
| Коэффициент при независимой переменной, $\beta_7$ | –0,07 | 0,02 | –3,47 | 0,00 | –0,10 | –0,03 |
| Коэффициент при независимой переменной, $\beta_8$ | 1,04 | 0,19 | –5,38 | 0,00 | –1,42 | –0,66 |
| Коэффициент при независимой переменной, $\beta_9$ | –0,01 | 0,00 | –6,29 | 0,00 | –0,01 | –0,01 |
| Коэффициент при независимой переменной, $\beta_{10}$ | –0,07 | 0,01 | –5,61 | 0,00 | –0,09 | –0,04 |

Таблица 8.8 – Значения описательных статистик остатков уравнения регрессии для случая, когда в анализ включены независимые переменные, имеющие статистически значимые коэффициенты

| Статистики | Значения |
|---|---|
| Среднее | 0,00 |
| Стандартная ошибка | 0,01 |
| Медиана | –0,08 |
| Мода | 0,34 |
| Стандартное отклонение | 0,19 |
| Дисперсия выборки | 0,04 |
| Эксцесс | 0,22 |
| Асимметричность | 1,32 |
| Интервал | 0,74 |
| Минимум | –0,24 |
| Максимум | 0,51 |
| Сумма | 0,00 |
| Счет | 1440 |

Значения, приведенные в таблицах 8.5–8.8, позволяют заключить следующее:

– следует принять альтернативную гипотезу: связь между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и рассматриваемыми 10 независимыми переменными является почти функциональной и статистически значимой;

– примерно 100 % вариации зависимой переменной объясняется вариацией независимых переменных;

– уравнение регрессии теперь имеет вид

$$y_x = 2{,}45 + 0{,}07x_1 + 0{,}02x_2 - 0{,}02x_3 - 0{,}04x_4 + 0{,}05x_5 + 0{,}06x_6 - \\ - 0{,}07x_7 + 1{,}04x_8 - 0{,}01x_9 - 0{,}07x_{10};$$

– величина свободного члена уравнения регрессии статистически значима;

– увеличение значения независимой переменной, например, «Количество членов домохозяйства» на одну единицу при условии, что значения остальных переменных остаются неизменными, влечет увеличение значения зависимой переменной «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» примерно на 7 коп. и наоборот;

– можно быть на 95 % уверенным в том, что примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели при увеличении, например, членов домохозяйства на одного человека будут увеличиваться на сумму в пределах между 0,01 и 0,14 р.;

– стандартная ошибка оценки уравнения регрессии *SEE* равна 0,19 р.;

– так как отношение значения асимметрии распределения остатков к ее средней квадратичной ошибке $\frac{|A_S|}{S_{A_S}} = \frac{1{,}32}{0{,}06} = 20{,}61$, что значительно больше 3,0, асимметрию следует признать существенной, а распределение остатков не следует считать симметричным;

– так как отношение значения эксцесса распределения остатков к его средней квадратичной ошибке $\frac{|\varepsilon_k|}{S_{\varepsilon_k}} = \frac{0{,}22}{0{,}13} = 1{,}72$, что меньше 3,0, при условии, что асимметрия несущественна, распределение остатков можно было бы считать близким к нормальному;

– рассчитанное уравнение регрессии следует признать условно пригодным для прогнозирования расходов домохозяйств на приобретение (обновление) домашней мебели в зависимости от выбранных 10 независимых переменных.

8 В последующих лабораторных работах, связанных с кластерным, дискриминантным и факторным анализами, для сокращения времени на их выполнение целесообразно использовать меньшее чем 10 количество независимых переменных. Для этого в учебных целях возможно из общего их числа выбрать, например, только те, у которых по итогам парных корреляционно-регрессионных анализов в лабораторной работе № 7 связь с зависимой переменной оказалась сильной (значение коэффициента корреляции по модулю превышает 0,7) и статистически значимой. Таких переменных шесть: «Количество членов домохозяйства», «Семейный среднемесячный доход», «Количество автомобилей в семье», «Вид жилья», «Площадь жилья» и «Стиль мебели».

Выполнить множественный корреляционно-регрессионный анализ для данного набора переменных. Для этого:

– создать в файле «08 Множественный корреляционно-регрессионный анализ.xlsx» лист с именем «Выборка с выс.коэф.корреляции»;

– скопировать в него все данные из листа «Выборка» (ячейки **«A1»**–**«P1442»**);

– в созданном листе удалить столбцы с переменными, для которых коэффициенты корреляции по абсолютному значению меньше 0,7 («Возраст мужа», «Возраст жены», «Образование мужа», «Образование жены», «Наличие подключения к интернету», «Вид телевидения», «Средний возраст мебели» и «Планируемая периодичность обновления мебели»);

– выполнить множественный корреляционно-регрессионный анализ, введя в поля его диалога значения зависимой и оставшихся независимых переменных в описанном выше порядке.

Результаты выполненного анализа приведены в таблицах 8.9–8.12.

Таблица 8.9 – Показатели регрессионной статистики для случая, когда в анализ были включены независимые переменные, у которых связь с зависимой переменной является сильной и статистически значимой

| Показатель | Значение |
|---|---|
| Множественный коэффициент корреляции | 0,999999 |
| Коэффициент детерминации | 0,999998 |
| Нормированный коэффициент детерминации | 0,999998 |
| Стандартная ошибка | 0,19 |
| Количество наблюдений | 1440 |

Таблица 8.10 – Результаты многофакторного дисперсионного анализа для случая, когда в анализ были включены независимые переменные, у которых связь с независимой переменной является сильной и статистически значимой

| Показатель | Число степеней свободы, *df* | Сумма квадратов разностей, *SS* | Оценка дисперсий (средний квадрат), *MS* | Значение *F*-критерия | Значимость *F*-критерия |
|---|---|---|---|---|---|
| Регрессия | 6 | 23194736,01 | 3865789,34 | 102612402,69 | 0,00 |
| Остаток | 1433 | 53,99 | 0,04 | | |
| Итого | 1439 | 23194790,00 | | | |

Таблица 8.11 – Значения рассчитанных коэффициентов уравнения регрессии, оценок их статистической значимости и границ доверительных интервалов для случая, когда в анализ были включены независимые переменные, у которых связь с зависимой переменной является сильной и статистически значимой

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Свободный член, $\beta_0$ | 1,94 | 0,39 | 5,01 | 0,00 | 1,18 | 1,94 |
| Коэффициент при независимой переменной, $\beta_1$ | 0,10 | 0,03 | 3,92 | 0,00 | 0,05 | 0,10 |
| Коэффициент при независимой переменной, $\beta_2$ | 0,05 | 0,00 | 5870,13 | 0,00 | 0,05 | 0,05 |
| Коэффициент при независимой переменной, $\beta_3$ | 0,06 | 0,02 | 2,89 | 0,00 | 0,02 | 0,06 |

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Коэффициент при независимой переменной, $\beta_4$ | –0,93 | 0,19 | –4,81 | 0,00 | –1,31 | –0,93 |
| Коэффициент при независимой переменной, $\beta_5$ | –0,01 | 0,00 | –5,57 | 0,00 | –0,01 | –0,01 |
| Коэффициент при независимой переменной, $\beta_6$ | –0,01 | 0,02 | –0,38 | 0,71 | –0,04 | 0,03 |

Таблица 8.12 – Значения описательных статистик остатков уравнения регрессии для случая, когда в анализ были включены независимые переменные, у которых связь с зависимой переменной является сильной и статистически значимой

| Статистики | Значения |
|---|---|
| Среднее | 0,00 |
| Стандартная ошибка | 0,01 |
| Медиана | –0,09 |
| Мода | 0,39 |
| Стандартное отклонение | 0,19 |
| Дисперсия выборки | 0,04 |
| Эксцесс | 0,13 |
| Асимметричность | 1,36 |
| Интервал | 0,65 |
| Минимум | –0,19 |
| Максимум | 0,46 |
| Сумма | 0,00 |
| Счет | 1440 |

Значения, приведенные в таблицах 8.9–8.12, позволяют заключить следующее:

– следует принять альтернативную гипотезу: между примерными среднегодовыми расходами на покупку (обновление) элементов домашней мебели и рассматриваемыми 6 независимыми переменными является почти функциональной и статистически значимой;

– примерно 100 % вариации зависимой переменной объясняется вариацией независимых переменных;

– уравнение регрессии имеет вид

$$y_x = 1{,}94 + 0{,}10x_1 + 0{,}05x_2 + 0{,}06x_3 - 0{,}93x_4 - 0{,}01x_5 - 0{,}01x_6;$$

– величина свободного члена уравнения регрессии статистически значима;

– величина коэффициента при переменной «Стиль мебели» теперь оказалась статистически незначимой, на что указывают значения критерия Стьюдента (подтвержденные *P*-значением) и границы его доверительного интервала, в пределах которого находится значение «ноль»;

– увеличение значения независимой переменной, например, «Количество членов домохозяйства» на одну единицу при условии, что значения остальных переменных остаются неизменными, влечет увеличение значения зависимой переменной «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» примерно на 0,10 р. и наоборот;

– можно быть на 95 % уверенным в том, что примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели при увеличении, например, членов домохозяйства на одного человека будут увеличиваться на сумму в пределах между 0,05 и 0,10 р.;

– стандартная ошибка оценки уравнения регрессии *SEE* равна 0,19 р.;

– так как отношение значения асимметрии распределения остатков к ее средней квадратичной ошибке $\frac{|A_S|}{S_{A_S}} = \frac{1{,}36}{0{,}06} = 21{,}21$, что значительно больше 3,0, асимметрию следует признать существенной, а распределение остатков не следует считать симметричным;

– так как отношение значения эксцесса распределения остатков к его средней квадратичной ошибке $\frac{|\varepsilon_k|}{S_{\varepsilon_k}} = \frac{0{,}14}{0{,}13} = 1{,}05$, что меньше 3,0, при условии, что асимметрия несущественна, распределение остатков можно было бы считать близким к нормальному;

– рассчитанное уравнение регрессии следует признать условно пригодным для прогнозирования расходов домохозяйств на приобретение (обновление) элементов домашней мебели в зависимости от выбранных переменных.

### 8.2.2 Множественный (многофакторный) корреляционно-регрессионный анализ с использованием программы IBM SPSS Statistics

1 Создать в программе новый файл и скопировать в него все данные только из листа «Выборка» файла «08 Множественный корреляционно-регрессионный анализ.xlsx».

2 Присвоить файлу при его сохранении название «08 Множественный корреляционно-регрессионный анализ.sav».

3 В редакторе данных созданного файла, нажав кнопку **«Переменные»**, перейти в одноименное окно и присвоить имена переменным созданной выборки так же, как это было сделано при выполнении лабораторных работ № 5 и 7.

4 Нажав кнопку **«Данные»**, перейти в одноименную вкладку редактора данных.

5 Выполнить множественный корреляционно-регрессионный анализ для всех переменных, среди которых переменная «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» является зависимой, а остальные 14 – независимыми. Для этого:

– выбрать процедуру корреляционно-регрессионного анализа (**«Анализ»** – **«Регрессия»** – **«Линейная…»**);

– в открывшемся диалоговом окне из левого поля, где представлены все переменные, нажав кнопку со стрелкой, направленной вправо, перенести в поле **«Зависимая переменная»** переменную «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели»;

– точно так же в поле **«Независимые переменные:»** перенести все независимые переменные;

– остальные поля не заполнять (рисунок 8.5);

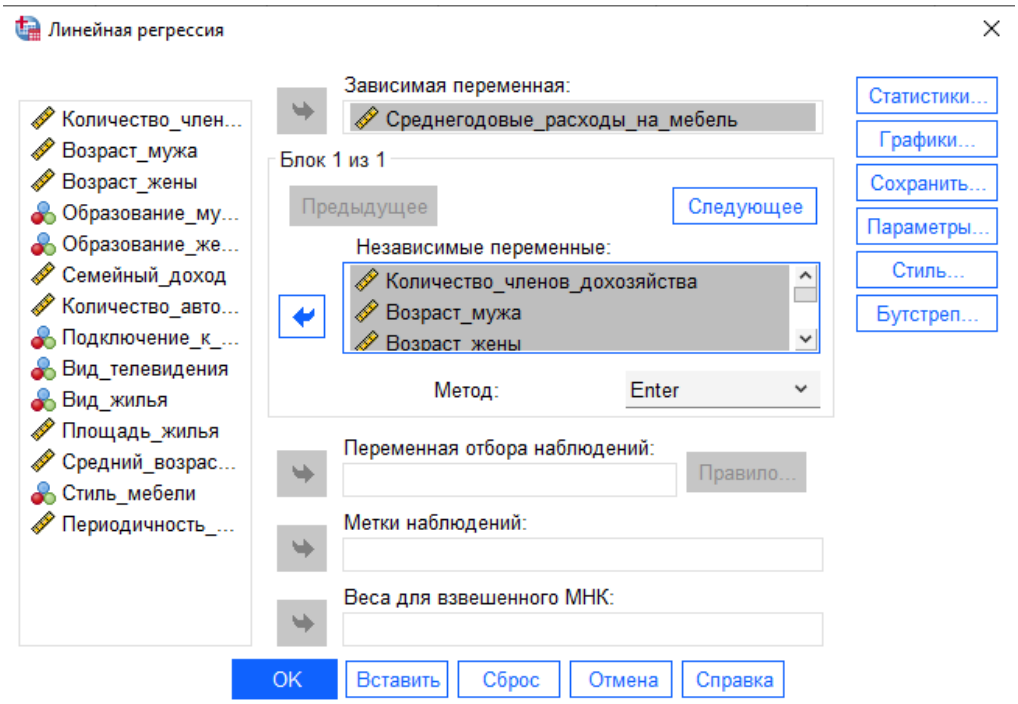



Рисунок 8.5 – Диалоговое окно «Линейная регрессия» с введенными именами зависимой и независимых переменных

– нажать кнопку **«Статистики…»** и в открывшемся диалоге выбрать расчет оценок коэффициентов регрессии, статистик согласия (множественного коэффициента корреляции и коэффициента детерминации, их скорректированных значений, таблицы дисперсионного анализа) и доверительных интервалов при вероятности достоверности 95 % (рисунок 8.6), а также тест Дарбина – Уотсона и нажать кнопку **«Продолжить»**;

– нажать кнопку **«Графики…»**, в открывшемся диалоге запросить вывод гистограммы и графика нормального распределения и нажать кнопку **«Продолжить»**;

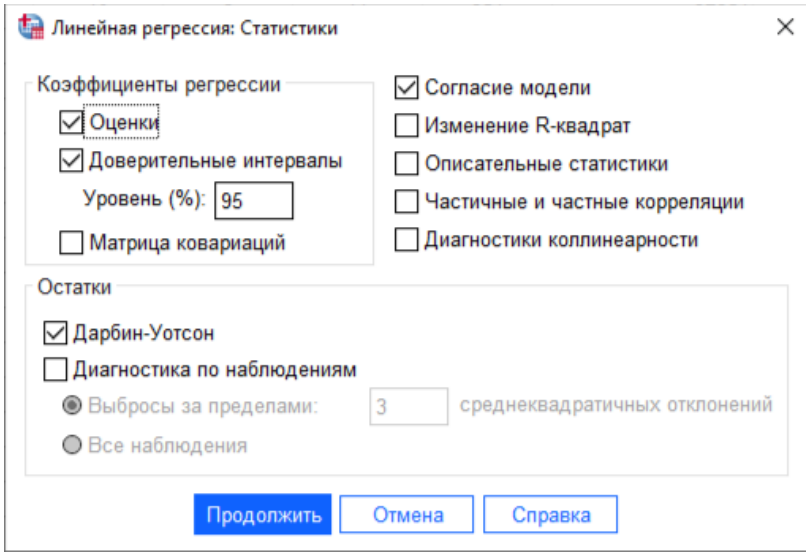



Рисунок 8.6 – Диалог «Линейная регрессия: Статистики» с выбранными расчетами оценок коэффициентов регрессии и статистик согласия

– нажать кнопку **«Сохранить…»**, в открывшемся диалоге запросить вывод нестандартизированных остатков (рисунок 8.7), после чего нажать кнопку **«Продолжить»**;

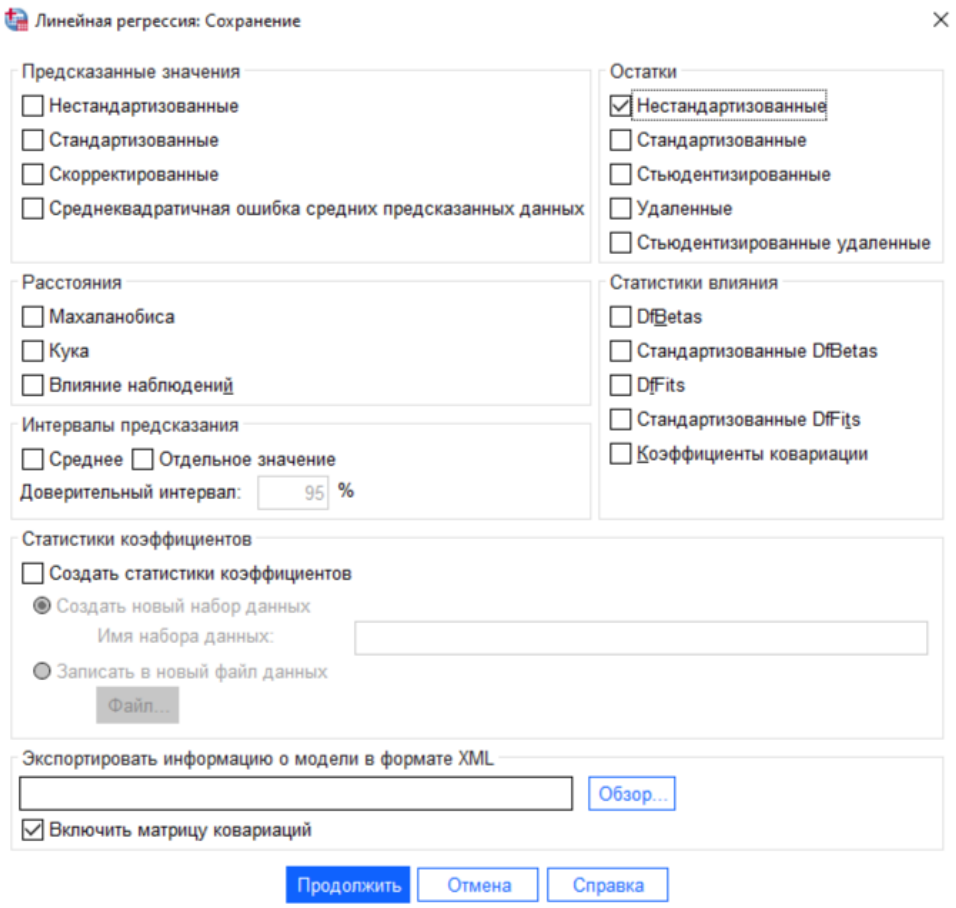



Рисунок 8.7 – Диалог «Линейная регрессия: Сохранение» с выбранными расчетами оценок коэффициентов регрессии и статистик согласия

– нажать кнопку **«Параметры…»** и в открывшемся диалоге для критериев шагового отбора оставить значимость *F*-критерия для порога включения переменной 0,05 и исключения 0,10, запросить включение в уравнение регрессии константы (свободного члена) (рисунок 8.8) и нажать кнопку **«Продолжить»**;

– в диалоге **«Линейная регрессия»** нажать кнопку **«ОК»**.

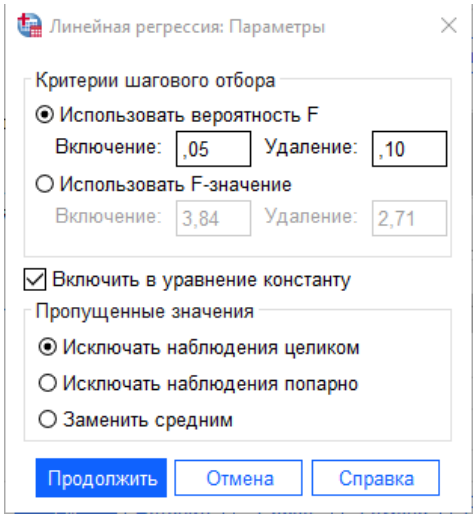



Рисунок 8.8 – Диалог «Линейная регрессия: Параметры» с заданными порогами для включения и исключения переменной и запросом включения в уравнение регрессии константы (свободного члена)

В открывшемся файле с результатами выполненного анализа обратить внимание на таблицы «Сводка модели», «Дисперсионный анализ», «Коэффициенты» и на гистограмму остатков с наложенной на нее кривой нормального распределения. Значения из этих таблиц сведены в таблицы 8.13–8.15 и полностью совпадают со значениями из таблиц 8.1–8.3.

Таблица 8.13 – Показатели регрессионной статистики, рассчитанные для случая, когда в анализ включен весь набор независимых переменных

| Показатель | Значение |
|---|---|
| Множественный коэффициент корреляции | 0,999999 |
| Коэффициент детерминации | 0,999998 |
| Скорректированный коэффициент детерминации | 0,999998 |
| Стандартная ошибка оценки уравнения регрессии | 0,19 |

Значение теста Дарбина – Уотсона равно 2,19, что относительно близко к 2,00 и говорит о возможности отсутствия автокорреляции, т. е. отклонения от теоретически возможных результатов (остатки) появляются случайным образом.

Таблица 8.14 – Результаты многофакторного дисперсионного анализа для случая, когда в анализ включен весь набор независимых переменных

| Показатель | Число степеней свободы, *df* | Сумма квадратов разностей, *SS* | Оценка дисперсий (средний квадрат), *MS* | Значение *F*-критерия | Значимость *F*-критерия |
|---|---|---|---|---|---|
| Регрессия | 14 | 23194738,01 | 1656767,00 | 45415525,45 | 0,00 |
| Остаток | 1425 | 51,98 | 0,036 | | |
| Итого | 1439 | 23194790,00 | | | |

Таблица 8.15 – Значения рассчитанных коэффициентов уравнения регрессии для случая, когда в анализ включен весь набор независимых переменных

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Свободный член, $\beta_0$ | 2,45 | 0,39 | 6,25 | 0,00 | 1,68 | 3,21 |
| Коэффициент при независимой переменной, $\beta_1$ | 0,08 | 0,04 | 2,21 | 0,03 | 0,01 | 0,15 |
| Коэффициент при независимой переменной, $\beta_2$ | –0,01 | 0,00 | –1,47 | 0,14 | –0,01 | 0,00 |
| Коэффициент при независимой переменной, $\beta_3$ | 0,02 | 0,00 | 5,03 | 0,00 | 0,01 | 0,03 |
| Коэффициент при независимой переменной, $\beta_4$ | –0,03 | 0,01 | –2,37 | 0,02 | –0,05 | 0,00 |
| Коэффициент при независимой переменной, $\beta_5$ | –0,04 | 0,01 | –3,21 | 0,00 | –0,06 | –0,02 |
| Коэффициент при независимой переменной, $\beta_6$ | 0,05 | 0,00 | 2960,69 | 0,00 | 0,05 | 0,05 |
| Коэффициент при независимой переменной, $\beta_7$ | 0,07 | 0,02 | 3,32 | 0,00 | 0,03 | 0,11 |
| Коэффициент при независимой переменной, $\beta_8$ | –0,07 | 0,02 | –3,04 | 0,00 | –0,11 | –0,02 |

| Коэффициенты уравнения регрессии | Значения коэффициентов | Стандартная ошибка | Значение *t*-статистики | *P*-значение | Границы доверительного интервала | |
|---|---|---|---|---|---|---|
| | | | | | нижняя | верхняя |
| Коэффициент при независимой переменной, $\beta_9$ | 0,02 | 0,02 | 0,85 | 0,39 | –0,02 | 0,06 |
| Коэффициент при независимой переменной, $\beta_{10}$ | –1,04 | 0,19 | –5,35 | 0,00 | –1,42 | –0,66 |
| Коэффициент при независимой переменной, $\beta_{11}$ | –0,01 | 0,00 | –6,27 | 0,00 | –0,01 | –0,01 |
| Коэффициент при независимой переменной, $\beta_{12}$ | 0,01 | 0,04 | 0,27 | 0,79 | –0,07 | 0,09 |
| Коэффициент при независимой переменной, $\beta_{13}$ | –0,02 | 0,02 | –1,12 | 0,26 | –0,06 | 0,02 |
| Коэффициент при независимой переменной, $\beta_{14}$ | –0,07 | 0,02 | –3,00 | 0,00 | –0,12 | –0,02 |

Гистограмма остатков с наложенной на нее кривой нормального распределения (рисунок 8.9) еще раз показывает, что распределение остатков не подпадает под нормальное. Следует сделать вывод о том, что рассчитанное уравнение регрессии является условно пригодным для прогнозирования расходов домохозяйств на покупку (обновление) элементов домашней мебели в зависимости от всех рассмотренных переменных. Кроме этого, в генеральной совокупности (на целевом рынке продукции ЧУП «Кэтнес») можно предполагать наличие двух значимо различающихся групп (сегментов рынка).

6 Вышеописанным образом выполнить множественный корреляционно-регрессионный анализ для статистически значимых независимых переменных и для статистически значимых независимых переменных, которые с зависимой переменной имеют абсолютное значение коэффициента корреляции 0,7 и выше.

7 Сводные результаты выполненных множественных корреляционно-регрессионных анализов представлены в таблице 8.16. При этом множественный коэффициент корреляции для всех трех случаев почти равен 1,0, результаты анализа статистически значимы, а стандартная ошибка оценки *SEE* равна 0,19 р.

8 Необходимо отметить, что построенные гистограммы остатков для всех трех уравнений множественной регрессии подтверждают гипотезу, выдвинутую в лабораторной работе № 7 в результате парного корреляционно-регрессионного анализа с использованием переменной «Семейный среднемесячный доход», о

наличии на целевом рынке продукции ЧУП «Кэтнес» двух четко различимых сегментов.

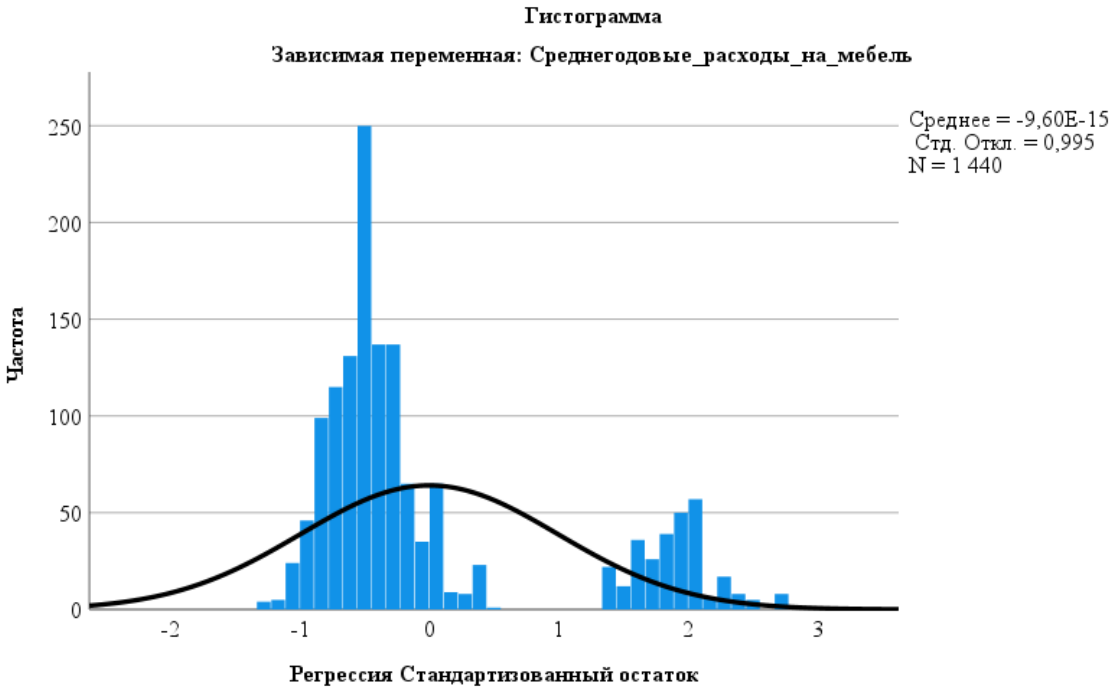



Рисунок 8.9 – Выполненная программой IBM SPSS Statistics гистограмма стандартизированных остатков с наложенной на нее кривой нормального распределения для случая, когда в анализ включен весь набор независимых переменных

Таблица 8.16 – Сводные результаты выполнения парного корреляционно-регрессионного анализа

| Независимые переменные | Уравнение регрессии |
|---|---|
| Включены все | $y_x = 2{,}45 + 0{,}08x_1 - 0{,}01x_2 + 0{,}02x_3 - 0{,}03x_4 - $ $- 0{,}04x_5 + 0{,}05x_6 + 0{,}07x_7 - 0{,}07x_8 +$ $+ 0{,}02x_9 - 1{,}04x_{10} - 0{,}01x_{11} + 0{,}01x_{12} -$ $- 0{,}02x_{13} - 0{,}07x_{14}$ |
| Включены только те, коэффициенты при которых являются статистически значимыми | $y_x = 2{,}45 + 0{,}07x_1 + 0{,}02x_2 - 0{,}02x_3 - 0{,}04x_4 +$ $+ 0{,}05x_5 + 0{,}06x_6 - 0{,}07x_7 + 1{,}04x_8 -$ $- 0{,}01x_9 - 0{,}07x_{10}$ |
| Включены только те, у которых коэффициенты корреляции с зависимой переменной превышают по модулю значение 0,70 | $y_x = 1{,}94 + 0{,}10x_1 + 0{,}05x_2 + 0{,}06x_3 - 0{,}93x_4 -$ $- 0{,}01x_5 - 0{,}01x_6$ |

## 8.3 Задание для самостоятельного выполнения

Исключив на листе «Расходы и выс.коррел.переменные» файла «08 Парный корреляционно-регрессионный анализ.xlsx» из списка 286 домохозяйств, для которых значения остатков оказались больше 0,30, выполнить парный корреляционно-регрессионный анализ для оставшихся 1154 домохозяйств.

## 8.4 Вопросы для самоконтроля

1 Что представляют собой множественный (многофакторный) корреляционно-регрессионный анализ и в каком порядке он проводится?

2 Как вычисляются множественный, частные и частичные коэффициенты корреляции?

3 Какие способы используются для отбора факторных признаков при множественном (многофакторном) регрессионном анализе?

4 Что является причинами мультиколлинеарности между факторными признаками? Как она может быть устранена?

5 Какие статистические гипотезы формулируются при проведении множественного (многофакторного) корреляционно-регрессионного анализа?

6 Какой критерий используется для общей проверки статистических гипотез при множественном (многофакторном) регрессионном анализе?

# ЛАБОРАТОРНАЯ РАБОТА № 9

## Кластерный анализ данных, полученных по выборке в процессе маркетингового исследования

**Цель работы:** выполнить кластерный анализ объектов исследования (домохозяйств), которые были включены в выборку, сформированную по итогам лабораторной работы № 8 и в которой коэффициенты корреляции между зависимой и независимыми переменными равны или превышают по модулю 0,70.

**Задачи работы:** с использованием теоретических знаний, полученных при изучении тем № 14 и 17, а также изученного ранее курса «Прикладной статистический анализ»:

– изучить порядок выполнения кластерного анализа данных, полученных по выборке в процессе маркетингового исследования;

– получить практические навыки в выполнении кластерного анализа данных, полученных в ходе маркетингового исследования, с использованием программы IBM SPSS Statistics.

### 9.1 Теоретические сведения

#### 9.1.1 Основные термины

**Кластерный анализ** – это одно из направлений статистического исследования, которое включает в себя совокупность методов, позволяющих классифицировать многомерные наблюдения, каждое из которых описывается набором исходных переменных $x_1, x_2, \ldots, x_n$.

**Кластер** – это подмножество объектов статистической совокупности (множества), однородных по своим признакам. Кластер может рассматриваться как самостоятельная единица исследования, обладающая определенным набором свойств.

**Иерархическая кластеризация** – это совокупность алгоритмов упорядочивания данных, направленных на создание иерархии (дерева) вложенных кластеров.

**Дендрограмма (древовидная диаграмма)** – это визуализатор, используемый для представления результатов иерархической кластеризации. Она показывает степень близости отдельных объектов и кластеров, а также наглядно демонстрирует в графическом виде последовательность их объединения или разделения. Количество уровней дендрограммы соответствует числу шагов слияния или разделения кластеров.

**Кластерный центроид** – это место типичных наблюдений для данного класса (подмножества), который можно использовать как для описания различий между классами (подмножествами), так и для определения принадлежности «неизвестных» объектов к одному из классов (подмножеств).

***T*-тест для независимых выборок** – это метод статистической проверки гипотезы о равенстве средних величин рассматриваемого признака в двух группах (подмножествах) объектов наблюдения. При этом каждый объект наблюдения должен быть только в одной из групп (подмножеств).

**Тест Ливиня** – это тест на равенство дисперсий рассматриваемой переменной (признака) в исследуемых выборках.

### 9.1.2 Порядок проведения кластерного анализа

Порядок выполнения кластерного анализа в ходе маркетингового исследования показан на рисунке 9.1. При этом необходимо помнить, что различия между зависимыми и независимыми переменными не проводят и проверяются взаимозависимые связи всего набора переменных.

#### 9.1.2.1 Формулирование проблемы кластерного анализа

На этапе формулирования проблемы кластеризации задача состоит в том, чтобы выбранный набор переменных смог описать сходство между объектами с точки зрения признаков, имеющих отношение к данной проблеме маркетингового исследования, и не включал посторонние переменные, которые могут исказить результаты кластеризации.


Рисунок 9.1 – Порядок выполнения кластерного анализа

### 9.1.2.2 Оценка сходства объектов кластеризации

Для оценки сходства объектов в кластерах могут использоваться различные метрики, наиболее часто используемыми из которых являются:

– евклидово расстояние:

$$d_{ij} = \sqrt{\sum_{k=1}^{l} \left(x_{ik} - x_{jk}\right)^2}, \tag{9.1}$$

где $d_{ij}$ – расстояние между $i$-м и $j$-м объектами;
$x_{ik}$ и $x_{jk}$ – значения $k$-й переменной соответственно у $i$-го и $j$-го объектов;

– взвешенное евклидово расстояние:

$$d_{ij} = \sqrt{\sum_{k=1}^{l} \alpha_k \left(x_{ik} - x_{jk}\right)^2}, \tag{9.2}$$

где $\alpha_k$ – вес, приписываемый $k$-й переменной;

– расстояние Хэмминга (расстояние city-block):

$$d_{ij} = \sum_{k=1}^{l} \left|x_{ik} - x_{jk}\right| ; \tag{9.3}$$

– расстояние Минковского ($Lp$-норма):

$$d_{ij} = \left(\sum_{k=1}^{l} \left|x_{ik} - x_{jk}\right|^p\right)^{\frac{1}{p}}, \tag{9.4}$$

где $p$ – количество признаков, характеризующих каждый объект;

$$d_{ij} = \left(X_i - X_j\right)' S_*^{-1} \left(X_i - X_j\right), \tag{9.5}$$

где $X_i$ и $X_J$ – векторы значений переменных у $i$-го и $j$-го объектов;
$S_*^{-1}$ – общая ковариационная матрица.

Оценка сходства между объектами сильно зависит от абсолютного значения признака и от степени его вариации в совокупности. Чтобы устранить подобное влияние на процедуру классификации, значения исходных переменных можно нормировать с использованием одной из следующих формул:

$$z_{ij} = \frac{x_{ik} - \overline{x}}{s_k}, \tag{9.6}$$

$$z_{ij} = \frac{x_{ik}}{x_{\max_k}}, \tag{9.7}$$

$$z_{ij} = \frac{x_{ik}}{x_{\min_k}}, \tag{9.8}$$

$$z_{ij} = \frac{x_{ik}}{\overline{x}_k}, \tag{9.9}$$

где $x_{ik}$ – $i$-е значение $k$-го признака объектов;
$x_{\max_k}$, $x_{\min_k}$, и $\overline{x}_k$ – соответственно максимальное, минимальное и среднее значение $k$-го признака объектов;
$s_k$ – стандартное отклонение значение $k$-й переменной

### 9.1.2.3 Методы кластеризации объектов

Для кластеризации объектов в ходе маркетингового исследования используются методы, представленные на рисунке 9.2.

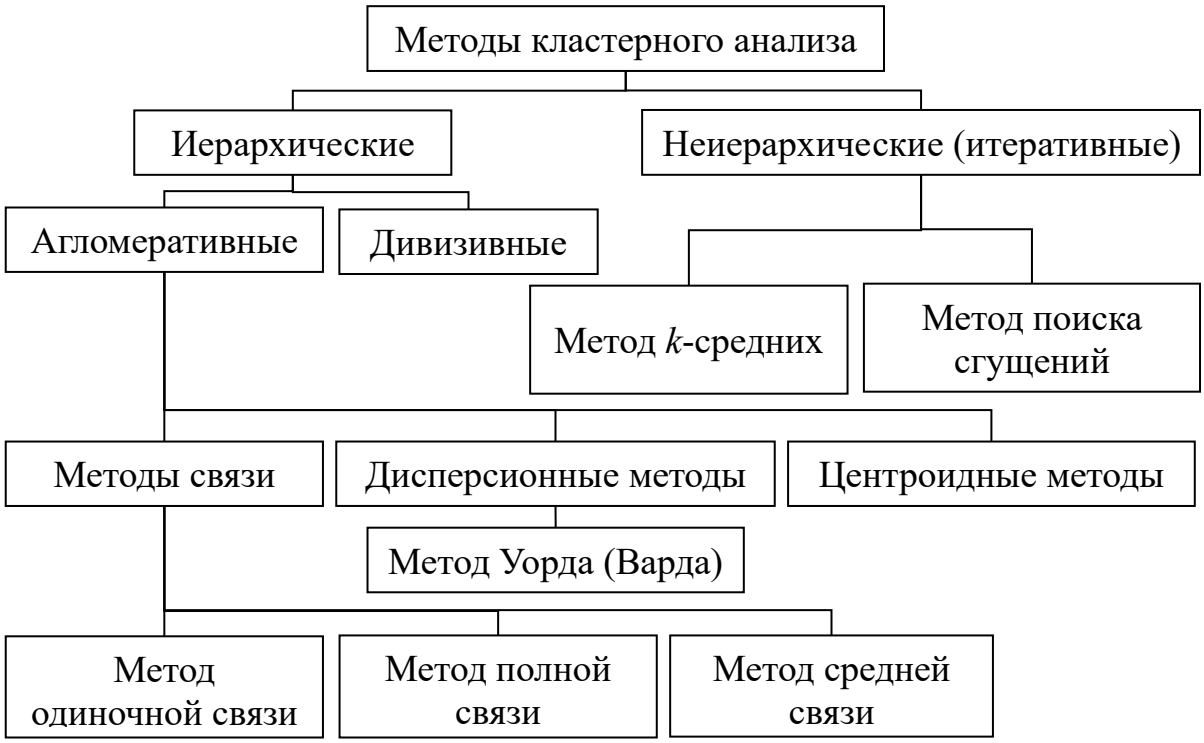



Рисунок 9.2 – Классификация методов кластерного анализа

Иерархические методы могут быть агломеративными (объединительными) и дивизивными. Агломеративная кластеризация начинается с каждого объекта в отдельном кластере. Затем кластеры объединяют, группируя объекты каждый раз во все более и более крупные кластеры. Этот процесс продолжают до тех пор, пока все объекты не станут членами одного-единственного кластера.

Дивизивная, или разделяющая, кластеризация начинается со всех объектов, сгруппированных в единственном кластере. Кластеры делят (расщепляют) до тех пор, пока каждый объект не окажется в отдельном кластере.

Обычно в маркетинговых исследованиях используют иерархические агломеративные методы. Порядок выполнения иерархического агломеративного кластерного анализа представлен на рисунке 9.3.

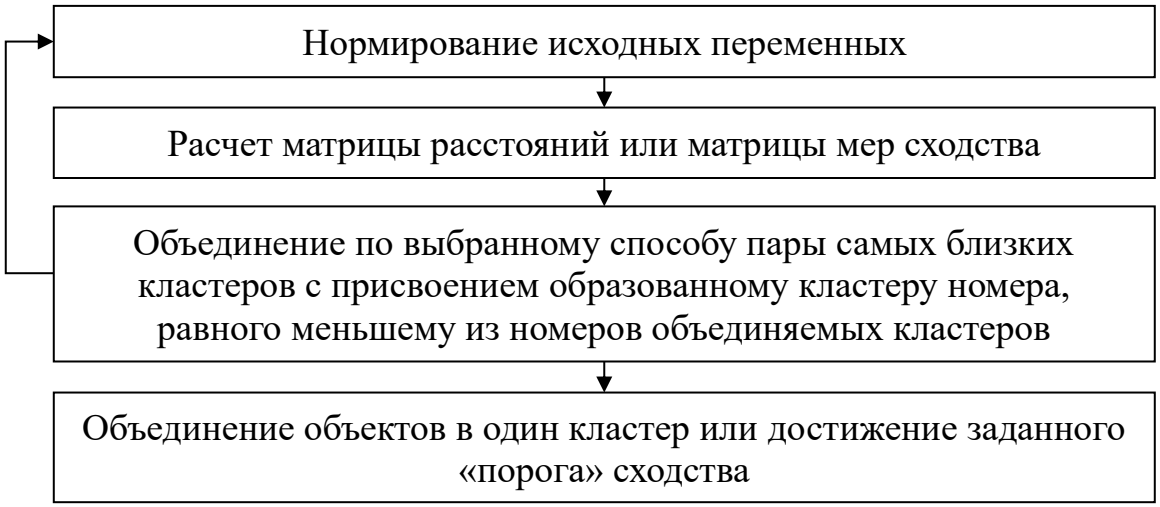



Рисунок 9.3 – Алгоритм иерархического агломеративного кластерного анализа

Мера сходства для объединения двух кластеров на третьем шаге определяется четырьмя методами:

– методом «ближайшего соседа», когда оценивается степень сходства между наиболее схожими (ближайшими) объектами кластеров;

– методом «дальнего соседа», когда оценивается степень сходства между наиболее отдаленными (несхожими) объектами кластеров;

– методом средней связи, когда оценивается средняя величина степеней сходства между объектами кластеров;

– методом медианной связи, когда расстояние между любым существующим кластером и новым кластером, который, в свою очередь, получился в результате объединения двух других кластеров, определяется как расстояние от центра существующего кластера до середины отрезка, соединяющего центры этих двух кластеров.

Основной исходной посылкой дивизивных методов является то, что первоначально все объекты принадлежат одному кластеру (множеству, классу). В процессе классификации по определенным правилам постепенно от этого кластера отделяются группы схожих между собой объектов. Таким образом, на каждом шаге количество кластеров возрастает, а мера расстояния между кластерами уменьшается.

Дивизивный алгоритм не требует пересчета матрицы расстояний на каждом шаге классификации в отличие от агломеративных методов, что способствует снижению трудоемкости расчетов.

Наряду с иерархическими методами классификации существует многочисленная группа итеративных методов кластерного анализа. Сущность их заключается в том, что процесс классификации начинается с задания некоторых начальных условий (количество образуемых кластеров, порог завершения процесса классификации и т. д.).

Порядок выполнения кластерного анализа с использованием итеративных методов представлен на рисунке 9.4.

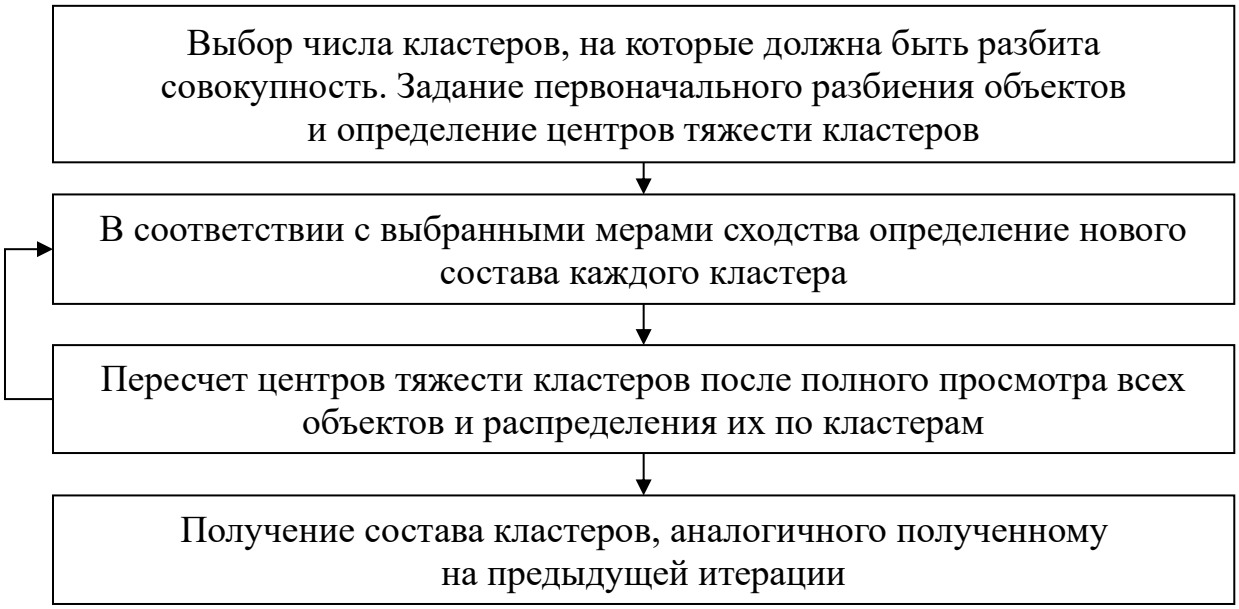



Рисунок 9.4 – Алгоритм вычислительных процедур итеративных методов классификации

### 9.1.2.4 Принятие решения о количестве кластеров, их интерпретация и профилирование

Для определения количества кластеров можно руководствоваться следующими правилами:

– опираться на теоретические и практические соображения;

– в иерархической кластеризации в качестве критерия можно использовать расстояния, при которых объединяют кластеры;

– в неиерархической кластеризации выполнять график зависимости отношения суммарной внутригрупповой дисперсии к межгрупповой дисперсии от числа кластеров. При этом точка, в которой наблюдается изгиб или резкий поворот, указывает на приемлемое количество кластеров, а увеличение числа кластеров за эту точку обычно безрезультативно;

– относительные размеры кластеров должны быть достаточно выразительными.

Интерпретация и профилирование кластеров включает проверку кластерных центроидов (центров). Центроиды (центры) представляют собой средние

значения по каждой из переменных тех объектов, которые содержатся в кластере. Они позволяют описывать каждый кластер, если присвоить ему номер или метку.

#### 9.1.2.5 Оценка надежности и достоверности кластеризации

Имея несколько умозаключений, выведенных из кластерного анализа, не следует принимать никакого решения по кластеризации, не выполнив оценку надежности и достоверности этого решения. Следующие процедуры могут обеспечить адекватную проверку качества кластерного анализа:

– выполнить кластерный анализ на основании одних и тех же данных, но с использованием различных способов измерения расстояния, после чего сравнить результаты, полученные на основе разных мер расстояния, чтобы определить, насколько совпадают полученные результаты;

– использовать разные методы кластерного анализа и сравнивать полученные результаты;

– разбить выборку на две равные части случайным образом и выполнить кластерный анализ отдельно для каждой половины, после чего сравнить кластерные центроиды двух подвыборок;

– случайным образом удалить некоторые переменные и выполнить кластерный анализ по сокращенному набору переменных, после чего сравнить результаты с полученными на основе полного набора переменных;

– в связи с тем, что в неиерархической кластеризации решение может зависеть от порядка случаев в наборе данных, выполнить анализ несколько раз, меняя порядок случаев до получения стабильного решения.

### 9.2 Выполнение кластерного анализа с использованием программы IBM SPSS Statistics

По итогам выполнения корреляционно-регрессионных анализов в лабораторных работах № 7 и 8 были установлены пять статистически значимых независимых переменных, у которых коэффициент корреляции с зависимой переменной «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» превышает значение по модулю 0,70: «Количество членов домохозяйства», «Семейный среднемесячный доход», «Количество автомобилей в семье», «Вид жилья» и «Площадь жилья».

Необходимо с использованием этих независимых и зависимой переменных выполнить кластерный анализ объектов наблюдения (домохозяйств) в выборке и сделать обоснованные выводы о возможной структуре генеральной совокупности (целевого рынка продукции ЧУП «Кэтнес»).

Работу выполнить в следующем порядке:

1 Создать в программе новый файл и скопировать в него из листа «Выборка с выс.коэф.корреляции» файла «08 Множественный корреляционно-регрессионный анализ.xlsx» данные по этим пяти переменным, а также номера

домохозяйств и значения примерных среднегодовых расходов на покупку (обновление) элементов мебели.

2 Присвоить файлу при его сохранении название «09 Кластерный анализ.sav».

3 В созданном файле, нажав в редакторе данных кнопку **«Переменные»**, перейти в одноименное окно и, как это было сделано в ранее выполненных лабораторных работах № 5, 7 и 8, присвоить имена переменным созданной выборки, для всех переменных задать тип **«Числовой»**, ширину колонки в восемь символов без десятичных знаков после запятой, выравнивание по центру, роль **«Входная»** и задать соответствующие шкалы. Для переменной «Номер домохозяйства» задать тип **«Строка»** и шкалу **«Номинальные»**.

4 Нажав кнопку **«Данные»**, вернуться во вкладку с данными для анализа.

5 Выполнить кластерный анализ. Для этого:

– выбрать процедуру кластерного анализа (**«Анализ»** – **«Классификация»** – **«Иерархическая кластеризация…»**);

– в открывшемся диалоговом окне, в котором представлены используемые метрические переменные, выделить их и, нажав кнопку со стрелкой, направленной вправо, перенести в поле **«Переменные:»**, а переменную «Номер домохозяйства» аналогичным образом перенести в поле **«Метить значениями»**. В секции **«Кластер»** выбрать **«Наблюдения»**, а в секции **«Вывести»** поставить флажки напротив строк **«Статистики»** и **«Графики»** (рисунок 9.5);

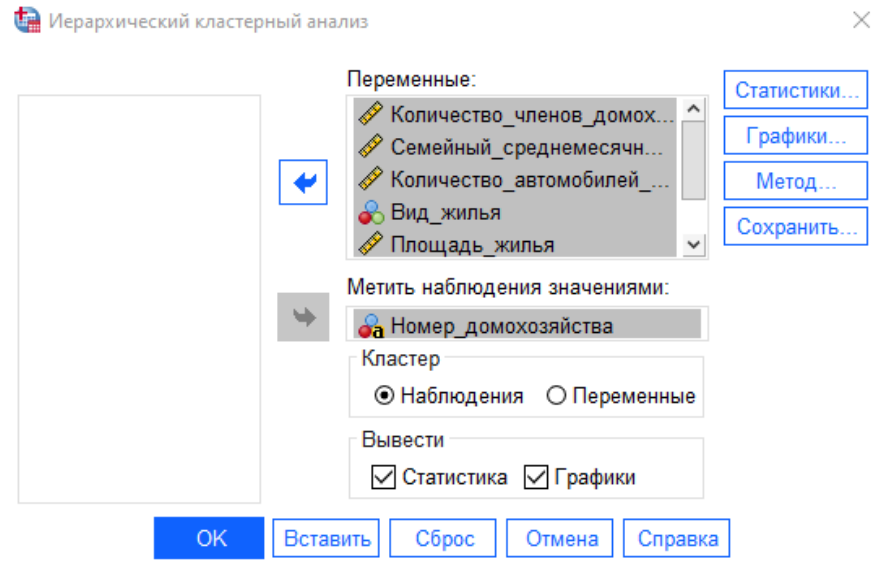



Рисунок 9.5 – Диалоговое окно «Иерархический кластерный анализ» с введенными переменными

– нажать кнопку **«Статистики…»**, в появившемся диалоге поставить флажок напротив строки **«Порядок агломерации»**, в секции **«Принадлежность к кластерам»** выбрать **«Нет»** (рисунок 9.6) и нажать кнопку **«Продолжить»**;

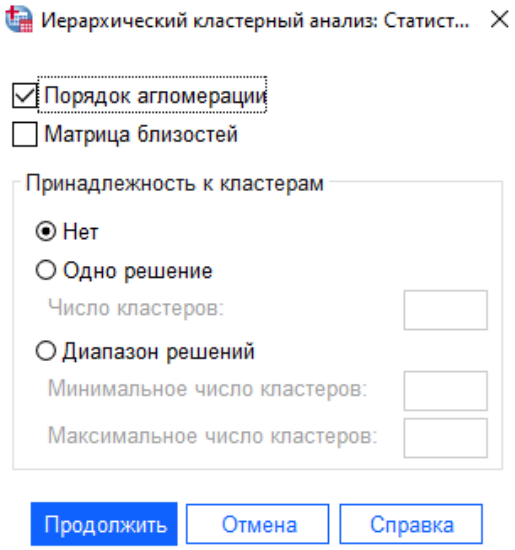



Рисунок 9.6 – Диалог «Иерархический кластерный анализ: Статистики» с установленным первоначальным порядком агломерации

– нажать кнопку **«Графики…»** и в появившемся диалоге поставить флажок напротив строки «**Дендрограмма»**, в секции **«Сосульчатая диаграмма»** запросить вывод всех кластеров, а в секции **«Ориентация»** установить горизонтальную ориентацию дендрограммы (рисунок 9.7) и нажать кнопку **«Продолжить»**;

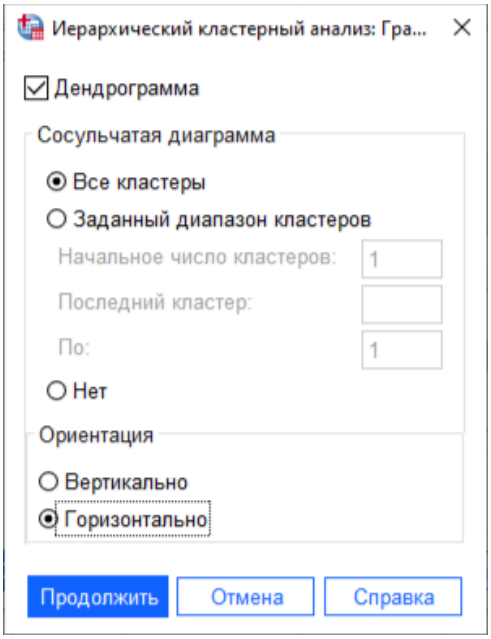



Рисунок 9.7 – Диалог «Иерархический кластерный анализ: Графики» с выбранным видом дендрограммы

– нажать кнопку **«Метод…»**, в появившемся диалоге (рисунок 9.8) в списке **«Метод:»** выбрать **«Метод Уорда»**, в списке **«Интервальная»** секции **«Шкала»**

выбрать **«Квадрат Евклидовой»**, в списке **«Стандартизация»** секции **«Преобразовать значения»** выбрать **«*Z*-оценки»** и нажать кнопку **«Продолжить»**;

– нажать кнопку **«ОК»** в диалоговом окне **«Иерархический кластерный анализ»**.

6 В создавшемся файле с результатами выполненного анализа обратить внимание на таблицу «Порядок агломерации (кластеров)» (в таблице 9.1 приведены начальные, промежуточные и конечные шаги (этапы)) и дендрограмму, построенную с использованием метода Уорда (Варда) (рисунок 9.9).

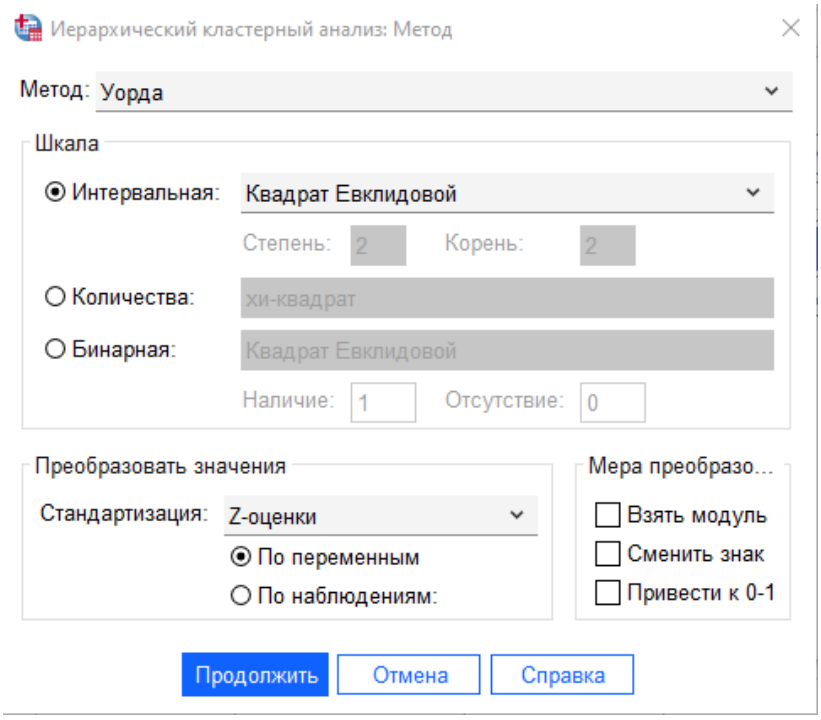



Рисунок 9.8 – Диалог «Иерархический кластерный анализ: Метод» с установленным методом кластеризации объектов

Таблица 9.1 – Фрагменты таблицы, показывающей порядок включения рассматриваемых домохозяйств в кластеры

| Этап | Кластер объединен с | | Коэффициенты | Этап первого появления | | Следующий этап |
|---|---|---|---|---|---|---|
| | кластером 1 | кластером 2 | | кластера 1 | кластера 2 | |
| 1 | 1269 | 1440 | 0,00 | 0 | 0 | 162 |
| 2 | 1259 | 1439 | 0,00 | 0 | 0 | 172 |
| 3 | 1078 | 1438 | 0,00 | 0 | 0 | 349 |
| 4 | 1257 | 1437 | 0,00 | 0 | 0 | 174 |
| 5 | 1408 | 1435 | 0,00 | 0 | 0 | 30 |
| 6 | 1254 | 1434 | 0,00 | 0 | 0 | 176 |
| 7 | 1073 | 1433 | 0,00 | 0 | 0 | 354 |
| 8 | 1306 | 1432 | 0,00 | 0 | 0 | 128 |
| 9 | 1251 | 1431 | 0,00 | 0 | 0 | 179 |

| Этап | Кластер объединен с | | Коэффициенты | Этап первого появления | | Следующий этап |
|---|---|---|---|---|---|---|
| | кластером 1 | кластером 2 | | кластера 1 | кластера 2 | |
| 10 | 1379 | 1430 | 0,00 | 0 | 0 | 59 |
| … | … | … | … | … | … | … |
| 960 | 404 | 440 | 0,00 | 0 | 829 | 996 |
| 961 | 388 | 439 | 0,00 | 0 | 843 | 1011 |
| 962 | 387 | 438 | 0,00 | 0 | 897 | 1012 |
| 963 | 401 | 437 | 0,00 | 0 | 881 | 999 |
| 964 | 76 | 436 | 0,00 | 0 | 625 | 1309 |
| 965 | 419 | 435 | 0,00 | 0 | 846 | 981 |
| 966 | 398 | 434 | 0,00 | 0 | 662 | 1002 |
| 967 | 253 | 433 | 0,00 | 0 | 835 | 1129 |
| 968 | 367 | 432 | 0,00 | 0 | 861 | 1032 |
| 969 | 251 | 431 | 0,00 | 0 | 919 | 1130 |
| 970 | 304 | 430 | 0,00 | 0 | 920 | 1087 |
| … | … | … | … | … | … | … |
| 1430 | 30 | 50 | 210,61 | 1421 | 1415 | 1435 |
| 1431 | 1 | 27 | 247,86 | 1413 | 1425 | 1434 |
| 1432 | 3 | 4 | 294,50 | 1404 | 1429 | 1436 |
| 1433 | 5 | 11 | 352,00 | 1420 | 1427 | 1438 |
| 1434 | 1 | 16 | 441,11 | 1431 | 1428 | 1436 |
| 1435 | 30 | 49 | 554,32 | 1430 | 1418 | 1439 |
| 1436 | 1 | 3 | 976,27 | 1434 | 1432 | 1437 |
| 1437 | 1 | 28 | 1924,35 | 1436 | 1426 | 1438 |
| 1438 | 1 | 5 | 3654,18 | 1437 | 1433 | 1439 |
| 1439 | 1 | 30 | 8634,00 | 1438 | 1435 | 0 |

7 Как видно из таблицы «Порядок агломерации (кластеров)» в файле с результатами анализа и из таблицы 9.1, первый наибольший скачок в значениях колонки «Коэффициенты» (4979,82) происходит после 1438 шага (этапа). Поэтому, следуя правилу, что оптимальное количество кластеров определяется как разность между количеством объектов наблюдения и номером шага (этапа), после которого наблюдается первый самый большой скачок в значениях этого показателя, следует выделить два кластера (1440 – 1438 = 2). Такое заключение подтверждает и построенная дендрограмма.

В соответствии с результатами кластерного анализа можно согласиться с высказанной по итогам лабораторной работы № 8 гипотезой о наличии на целевом рынке компании двух основных сегментов покупателей. Причем в каждом из них, как это по итогам лабораторной работы № 10 будет показано на примере второго сегмента, после продолжения применения процедуры кластеризации могут быть выделены свои субсегменты.

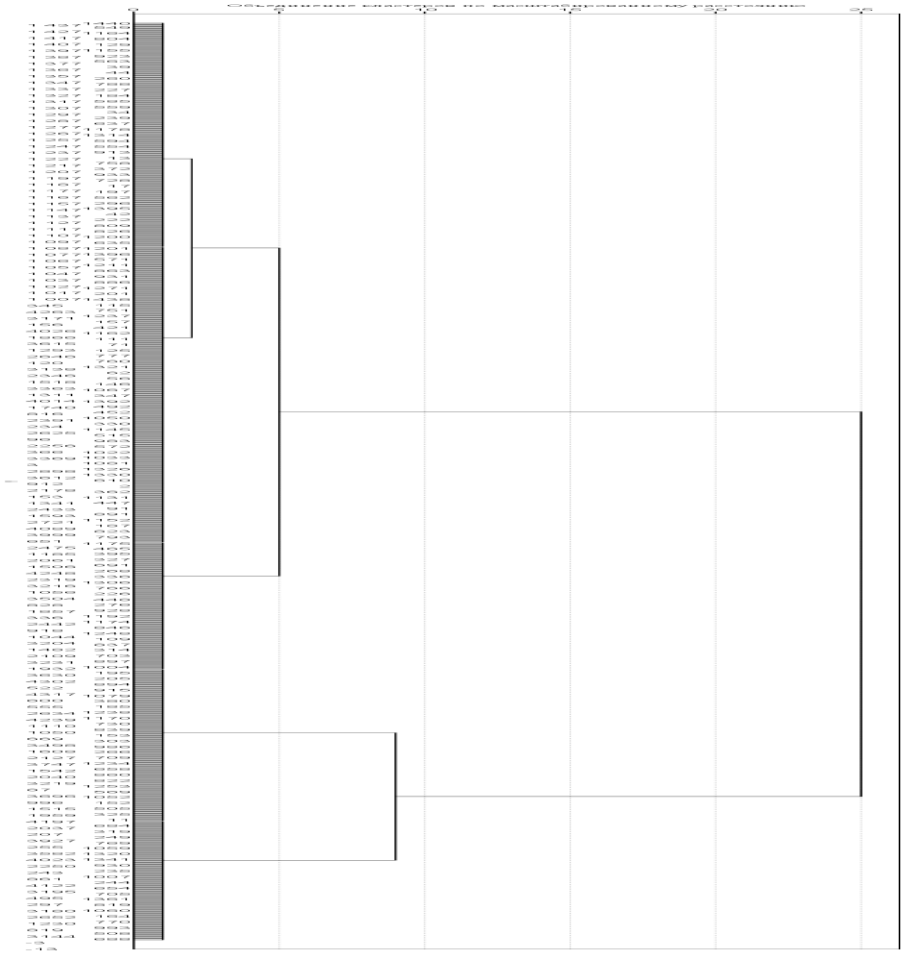


Рисунок 9.9 – Дендрограмма, выполненная по итогам кластерного анализа с использованием метода Уорда (Варда)

8 Еще раз выполнить кластерный анализ, но теперь после нажатия кнопки **«Статистики»** в секции **«Принадлежность к кластерам»** выбрать **«Одно решение»** и указать число кластеров **«2»** (рисунок 9.10) и нажать кнопку **«Продолжить»**.

Введенную информацию проверить, нажав кнопку **«Сохранить»**. Если количество кластеров в диалоге **«Иерархический кластерный анализ: Сохранить»** не появилось, самостоятельно ввести их число в секции **«Принадлежность к кластерам»** (рисунок 9.11) и нажать кнопку **«Продолжить»**.

После этого нажать кнопку «**ОК**» диалога **«Иерархический кластерный анализ»**.

Таблица «Порядок агломерации (кластеров)» и рисунок «Дендрограмма с использованием метода Варда» по своему составу абсолютно идентичны тем, которые были получены, когда количество кластеров точно известно не было.

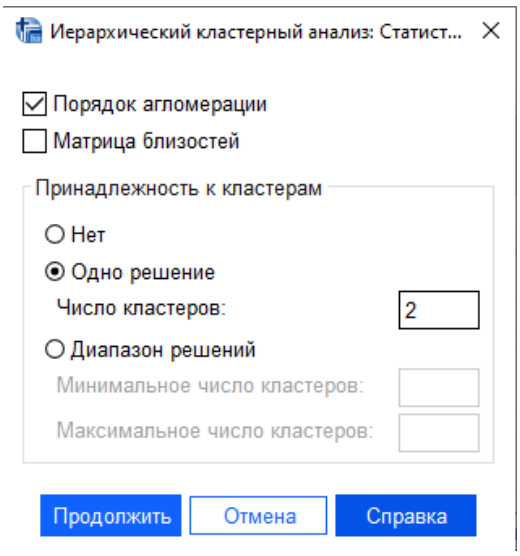



Рисунок 9.10 – Диалог «Иерархический кластерный анализ: Статистики» с введенным числом кластеров

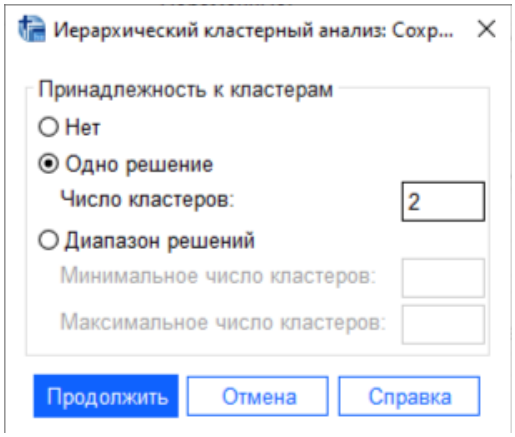



Рисунок 9.11 – Диалог «Иерархический кластерный анализ: Сохранить» с введенным числом кластеров

Таблица 9.2 «Принадлежность к кластерам» (в ней снова приведены начальные, промежуточные и конечные значения) показывает, на каком шаге (этапе) конкретное домохозяйство было отнесено к тому или иному кластеру.

Таблица 9.2 – Фрагменты таблицы, показывающей порядок включения рассматриваемых домохозяйств в выделенные кластеры

| Наблюдение | Номер домохозяйства | Номер кластера |
|---|---|---|
| 1 | 3 | 1 |
| 2 | 6 | 1 |
| 3 | 9 | 1 |
| 4 | 12 | 1 |
| 5 | 15 | 2 |
| 6 | 18 | 1 |

| Наблюдение | Номер домохозяйства | Номер кластера |
|---|---|---|
| 7 | 21 | 1 |
| 8 | 24 | 1 |
| 9 | 27 | 1 |
| 10 | 30 | 1 |
| … | … | … |
| 960 | 2880 | 2 |
| 961 | 2883 | 1 |
| 962 | 2886 | 1 |
| 963 | 2889 | 1 |
| 964 | 2892 | 2 |
| 965 | 2895 | 1 |
| 966 | 2898 | 1 |
| 967 | 2901 | 1 |
| 968 | 2904 | 1 |
| 969 | 2907 | 2 |
| 970 | 2910 | 1 |
| … | … | … |
| 1430 | 4290 | 1 |
| 1431 | 4293 | 1 |
| 1432 | 4296 | 1 |
| 1433 | 4299 | 1 |
| 1434 | 4302 | 1 |
| 1435 | 4305 | 2 |
| 1436 | 4308 | 1 |
| 1437 | 4311 | 1 |
| 1438 | 4314 | 1 |
| 1439 | 4317 | 1 |
| 1440 | 4320 | 1 |

9 Определить количество домохозяйств, входящих в каждый из двух кластеров, путем их сортировки. Для этого:

– выбрать инструмент сортировки (**«Данные»** – **«Сортировать наблюдения…»**);

– в открывшемся диалоге в качестве параметра для сортировки выбрать **«Ward Method»** и, нажав кнопку со стрелкой, направленной вправо, перенести его в поле **«Сортировать по:»**. В секции **«Порядок сортировки»** выбрать **«По возрастанию»**, а в секции **«Сохранить отсортированные данные»** поставить флажок напротив строки **«Сохранить файл с отсортированными данными»** (рисунок 9.12). После того как кнопка **«Файл…»** станет активной, нажать ее и выбрать папку, в которой находятся файлы с лабораторной работой, дать ему название «09 Выделенные кластеры.sav» и нажать кнопку **«Сохранить»**;

– нажать кнопку «**ОК**» диалога **«Сортировка наблюдений»**.

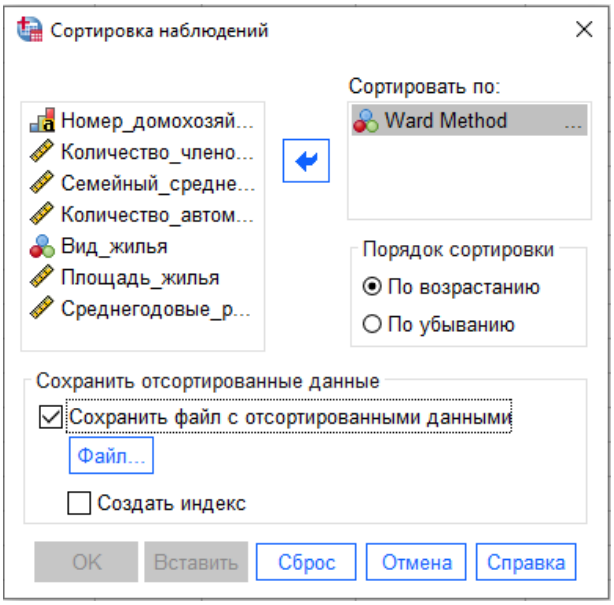



Рисунок 9.12 – Диалоговое окно «Сортировка наблюдений» с заданными условиями сортировки домохозяйств

В созданном файле видно, что в состав первого кластера включено 1253 домохозяйства (примерно 87,0 % от их рассмотренного числа), а в состав второго кластера включено 187 домохозяйств (примерно 13 %).

Таблица 9.3 – Фрагменты таблицы с результатами классификации рассмотренных домохозяйств по выделенным кластерам

| № п/п | Номер домохозяйства | Количество членов домохозяйства | Семейный среднемесячный доход | Количество автомобилей в семье | Вид жилья | Площадь жилья | Ср. год. расходы на мебель | Кластер |
|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 4 | 7320 | 1 | 2 | 72 | 366 | 1 |
| 2 | 6 | 4 | 7200 | 1 | 2 | 72 | 360 | 1 |
| 3 | 9 | 2 | 4880 | 1 | 2 | 36 | 244 | 1 |
| 4 | 12 | 3 | 5790 | 1 | 2 | 54 | 290 | 1 |
| 5 | 15 | 6 | 9000 | 2 | 2 | 108 | 450 | 1 |
| 6 | 18 | 4 | 7200 | 1 | 2 | 72 | 360 | 1 |
| 7 | 21 | 4 | 6760 | 1 | 2 | 72 | 338 | 1 |
| 8 | 24 | 2 | 4300 | 1 | 2 | 36 | 215 | 1 |
| 9 | 27 | 3 | 5070 | 1 | 2 | 54 | 254 | 1 |
| 10 | 30 | 6 | 9660 | 2 | 2 | 108 | 483 | 1 |
|  | … | … | … | … | … | … | … | … |
| 1251 | 4314 | 5 | 5250 | 1 | 2 | 90 | 263 | 1 |
| 1252 | 4317 | 6 | 8640 | 2 | 2 | 108 | 432 | 1 |
| 1253 | 4320 | 3 | 5070 | 1 | 2 | 54 | 254 | 1 |

| № п/п | Номер домохозяйства | Количество членов домохозяйства | Семейный среднемесячный доход | Количество автомобилей в семье | Вид жилья | Площадь жилья | Ср. год. расходы на мебель | Кластер |
|---|---|---|---|---|---|---|---|---|
| 1254 | 90 | 6 | 11280 | 2 | 1 | 259 | 564 | 2 |
| 1255 | 147 | 6 | 13260 | 3 | 1 | 259 | 663 | 2 |
| 1256 | 150 | 5 | 10100 | 2 | 1 | 216 | 505 | 2 |
| 1257 | 165 | 6 | 11280 | 2 | 1 | 259 | 564 | 2 |
| 1258 | 180 | 6 | 11340 | 2 | 1 | 259 | 567 | 2 |
| 1259 | 192 | 6 | 11700 | 2 | 1 | 259 | 585 | 2 |
| 1260 | 207 | 6 | 13800 | 3 | 1 | 259 | 690 | 2 |
|  | … | … | … | … | … | … | … | … |
| 1431 | 4122 | 6 | 10440 | 2 | 1 | 259 | 522 | 2 |
| 1432 | 4197 | 6 | 12240 | 3 | 1 | 259 | 612 | 2 |
| 1433 | 4212 | 6 | 12240 | 3 | 1 | 259 | 612 | 2 |
| 1434 | 4224 | 5 | 11100 | 2 | 1 | 216 | 555 | 2 |
| 1435 | 4227 | 6 | 10260 | 2 | 1 | 259 | 513 | 2 |
| 1436 | 4257 | 5 | 11450 | 2 | 1 | 216 | 573 | 2 |
| 1437 | 4260 | 6 | 12540 | 3 | 1 | 259 | 627 | 2 |
| 1438 | 4272 | 5 | 10000 | 2 | 1 | 216 | 500 | 2 |
| 1439 | 4275 | 6 | 10260 | 2 | 1 | 259 | 513 | 2 |
| 1440 | 4305 | 5 | 11100 | 2 | 1 | 216 | 555 | 2 |

10 В созданном файле «09 Выделенные кластеры.sav» в редакторе данных нажать кнопку **«Переменные»**, переменную «CLU2_1» переименовать в «Кластеры» и выбрать для нее номинальную шкалу. Если в колонке **«Метка»** осталась запись **«Ward Method»**, удалить ее.

11 Сохранить выполненные изменения в файле «09 Выделенные кластеры.sav».

12 По итогам выполненного кластерного анализа, применив *t*-тест для независимых выборок, охарактеризовать выделенные кластеры по переменной «Количество членов домохозяйства». Для этого:

– выбрать процедуру *t*-теста для независимых выборок (**«Анализ»** – **«Сравнение средних»** – **«Т-критерий для независимых выборок…»**);

– в открывшемся диалоговом окне из левого поля, нажав стрелку, направленную вправо, перенести в поле **«Проверяемые переменные:»** самую первую из них – «Количество членов домохозяйства»;

– таким же образом из левого поля в поле **«Группировать по:»** перенести переменную «Кластеры» (рисунок 9.13);

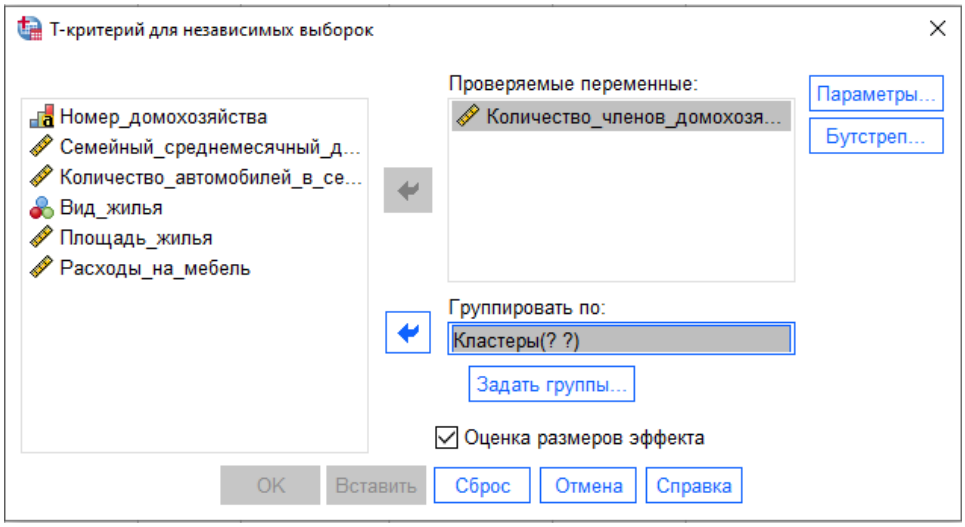



Рисунок 9.13 – Диалоговое окно «Т-критерий для независимых выборок» с введенной характеристикой для выполнения *t*-теста для двух независимых выборок

– нажать кнопку **«Задать группы…»**, в появившемся диалоге убедиться, что для теста выбраны именно две группы (два кластера) (рисунок 9.14) и нажать кнопку **«Продолжить»**. Содержание диалога станет таким, как это показано на рисунке 9.15;

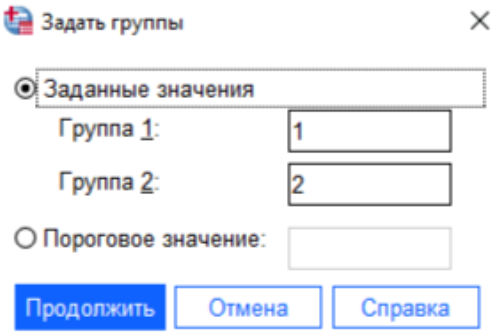



Рисунок 9.14 – Диалог «Задать группы» с указанным количеством групп для выполнения *t*-теста

– нажать кнопку **«Параметры»**, в открывшемся диалоге убедиться, что доверительный интервал допускает вероятность ошибки 95 % в случае отклонения нулевой гипотезы о равенстве средних в рассматриваемых выборках (кластерах) (рисунок 9.16) и нажать кнопку **«Продолжить»**;

– нажать кнопку **«ОК»** диалога **«Т-критерий для независимых выборок»**.

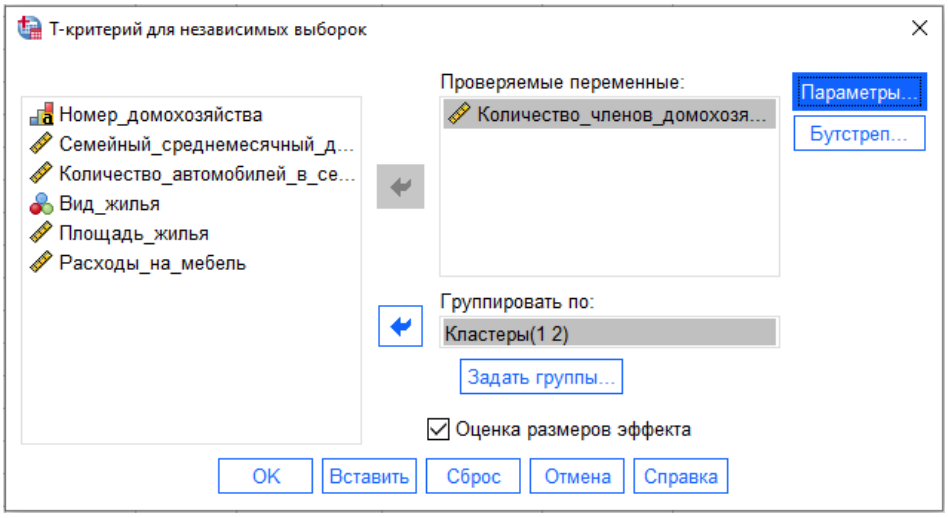



Рисунок 9.15 – Диалоговое окно «Т-критерий для независимых выборок» с введенными характеристикой и количеством кластеров для выполнения *t*-теста для двух независимых выборок

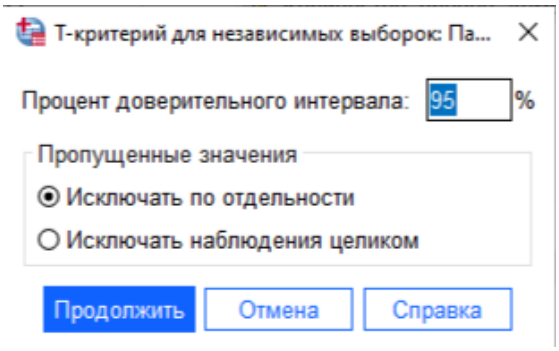



Рисунок 9.16 – Диалог «Т-критерий для независимых выборок: Параметры» с допущением вероятности ошибки 95 % в случае отклонения нулевой гипотезы о равенстве средних в рассматриваемых выборках

Результаты теста представлены в таблицах 9.4 и 9.5.

На основании значений, приведенных в таблицах 9.4 и 9.5, можно сделать следующие выводы:

– среднее количество членов в одном домохозяйстве первого кластера равно примерно 3,73, второго – 5,74;

– так как значимость критерия равенства дисперсий Ливиня примерно равна 0,00, что меньше 0,05, гипотеза о равенстве дисперсий переменной «Количество членов домохозяйства» в кластерах отвергается;

– так как значимость *t*-критерия также не превышает 0,05, нулевая гипотеза о равенстве среднего числа членов одного домохозяйства в каждом из кластеров должна быть отклонена.

Следовательно, выделенные два кластера различаются между собой по значению характеристики «Количество членов домохозяйства» и это различие является статистически значимым.

Таблица 9.4 – Таблица со значениями показателей описательной статистики в кластерах для переменной «Количество членов домохозяйства»

| Номер кластера | Количество домохозяйств | Количество членов, чел. | Стандартное отклонение | Средн. квадр. ошибка среднего |
|---|---|---|---|---|
| 1 | 1253 | 3,73 | 1,23 | 0,04 |
| 2 | 187 | 5,74 | 0,44 | 0,03 |

Таблица 9.5 – Таблица со значениями критериев для независимых выборок в кластерах для переменной «Количество членов домохозяйства»

| Равные дисперсии предполагаются | Критерий равенства дисперсий Ливиня | | *t*-критерий для равенства средних | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *F* | знч. | *t* | степ. своб. | знч. (2-стор.) | средн. разность | средн. квадр. ошибка разности | 95%-й доверит. интервал разности | |
| | | | | | | | | нижн. | верхн. |
| Да | 167,81 | 0,00 | –22,20 | 1438 | 0,00 | –2,02 | 0,09 | –2,19 | –1,84 |
| Нет | | | –42,66 | 731,23 | 0,00 | –2,02 | 0,05 | –2,11 | –1,92 |

13 Выполнить данный *t*-тест с использованием остальных пяти переменных. Его результаты представлены в таблицах 9.6–9.18.

Таблица 9.6 – Таблица со значениями показателей описательной статистики в кластерах для переменной «Семейный среднемесячный доход»

| Номер кластера | Количество домохозяйств | Семейный среднемесячный доход, р. | Стандартное отклонение | Средн. квадр. ошибка среднего |
|---|---|---|---|---|
| 1 | 1253 | 6041,34 | 1896,07 | 53,57 |
| 2 | 187 | 11348,13 | 1022,22 | 74,75 |

Таблица 9.7 – Таблица со значениями критериев для независимых выборок в кластерах для переменной «Семейный среднемесячный доход»

| Равные дисперсии предполагаются | Критерий равенства дисперсий Ливиня | | *t*-критерий для равенства средних | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *F* | знч. | *t* | степ. своб. | знч. (2-стор.) | средн. разность | средн. квадр. ошибка разности | 95%-й доверит. интервал разности | |
| | | | | | | | | нижн. | верхн. |
| Да | 101,46 | 0,00 | –37,46 | 1438 | 0,00 | –5307 | 141,66 | –5585 | –5029 |
| Нет | | | –57,71 | 409,99 | 0,00 | –5307 | 91,96 | –5488 | –5126 |

На основании значений, приведенных в таблицах 9.6 и 9.7, можно сделать следующие выводы:

– семейный среднемесячный доход одного домохозяйства первого кластера равен примерно 6041,34 р., второго – 11348,13 р.;

– так как значимость критерия равенства дисперсий Ливиня примерно равна 0,00, что меньше 0,05, гипотеза о равенстве дисперсий переменной «Семейный среднемесячный доход» в кластерах отвергается;

– так как значимость *t*-критерия не превышает 0,05, нулевая гипотеза о равенстве среднего значения семейного среднемесячного дохода домохозяйств в каждом из кластеров должна быть отклонена.

Следовательно, выделенные два кластера различаются между собой по значению характеристики «Семейный среднемесячный доход» и это различие является статистически значимым.

Таблица 9.8 – Таблица со значениями показателей описательной статистики в кластерах для переменной «Количество автомобилей в семье»

| Номер кластера | Количество домохозяйств | Количество автомобилей | Стандартное отклонение | Средн. квадр. ошибка среднего |
|---|---|---|---|---|
| 1 | 1253 | 1,03 | 0,59 | 0,02 |
| 2 | 187 | 2,22 | 0,42 | 0,03 |

Таблица 9.9 – Таблица со значениями критериев для независимых выборок в кластерах для переменной «Количество автомобилей в семье»

| Равные дисперсии предполагаются | Критерий равенства дисперсий Ливиня | | *t*-критерий для равенства средних | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *F* | знч. | *t* | степ. своб. | знч. (2-стор.) | средн. разность | средн. квадр. ошибка разности | 95%-й доверит. интервал разности | |
| | | | | | | | | нижн. | верхн. |
| Да | 0,33 | 0,57 | –26,68 | 1438 | 0,00 | –1,19 | 0,05 | –1,28 | –1,11 |
| Нет | | | –34,25 | 308,70 | 0,00 | –1,19 | 0,04 | –1,26 | –1,13 |

На основании значений, приведенных в таблицах 9.8 и 9.9, можно сделать следующие выводы:

– среднее количество автомобилей в одном домохозяйстве первого кластера равно примерно 1,03, второго – 2,22;

– так как значимость критерия равенства дисперсий Ливиня примерно равна 0,57, что больше 0,05, гипотеза о равенстве дисперсий переменной «Количество автомобилей в семье» в кластерах принимается;

– так как значимость *t*-критерия не превышает 0,05, нулевая гипотеза о равенстве среднего числа количества автомобилей в домохозяйствах в каждом из кластеров должна быть отклонена.

Следовательно, выделенные два кластера различаются между собой по значению характеристики «Количество автомобилей в семье» и это различие является статистически значимым.

Таблица 9.10 – Таблица со значениями показателей описательной статистики в кластерах для переменной «Вид жилья»

| Номер кластера | Количество домохозяйств | Вид жилья | Стандартное отклонение | Средн. квадр. ошибка среднего |
|---|---|---|---|---|
| 1 | 1253 | 2,00 | 0,00 | 0,00 |
| 2 | 187 | 1,00 | 0,00 | 0,00 |

Таблица со значениями критериев для независимых выборок в кластерах для переменной «Вид жилья» программой IBM SPSS Statistics не была создана, так как среднеквадратичные отклонения обеих групп равны 0,00.

На основании значений, приведенных в таблице 9.10, можно сделать вывод о том, что кластеры отличаются между собой по виду жилья, в котором проживают их участники. Представители первого кластера проживают в квартирах, представители второго – в домах. Выделенные кластеры различаются между собой по значению характеристики «Вид жилья» и это различие является статистически значимым.

Таблица 9.11 – Таблица со значениями показателей описательной статистики в кластерах для переменной «Площадь жилья»

| Номер кластера | Количество домохозяйств | Площадь жилья, кв. м | Стандартное отклонение | Средн. квадр. ошибка среднего |
|---|---|---|---|---|
| 1 | 1253 | 67,09 | 22,14 | 0,63 |
| 2 | 187 | 247,96 | 18,83 | 1,38 |

Таблица 9.12 – Таблица со значениями критериев для независимых выборок в кластерах для переменной «Площадь жилья»

| Равные дисперсии предполагаются | Критерий равенства дисперсий Ливиня | | *t*-критерий для равенства средних | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *F* | знч. | *t* | степ. своб. | знч. (2-стор.) | средн. разность | средн. квадр. ошибка разности | 95%-й доверит. интервал разности | |
| | | | | | | | | нижн. | верхн. |
| Да | 5,03 | 0,03 | –106 | 1438 | 0,00 | –180,88 | 1,71 | –184,2 | –177,5 |
| Нет | | | –120 | 269 | 0,00 | –180,88 | 1,51 | –183,9 | –177,9 |

На основании значений, приведенных в таблицах 9.11 и 9.12, можно сделать следующие выводы:

– средняя площадь жилья домохозяйств первого кластера равна примерно 67,09 кв. м, второго – 247,96 кв. м;

– так как значимость критерия равенства дисперсий Ливиня примерно равна 0,03, что меньше 0,05, гипотеза о равенстве дисперсий переменной «Площадь жилья» в кластерах отвергается;

– так как значимость *t*-критерия также не превышает 0,05, нулевая гипотеза о равенстве средней площади жилья домохозяйств в каждом из кластеров должна быть отклонена.

Следовательно, выделенные два кластера различаются между собой по значению характеристики «Площадь жилья» и это различие является статистически значимым.

Таблица 9.13 – Таблица со значениями показателей описательной статистики в кластерах для переменной «Среднегодовые расходы на приобретение (обновление) элементов домашней мебели»

| Номер кластера | Количество домохозяйств | Среднегодовые расходы, р. | Стандартное отклонение | Средн. квадр. ошибка среднего |
|---|---|---|---|---|
| 1 | 1253 | 302,17 | 94,83 | 2,68 |
| 2 | 187 | 567,45 | 51,10 | 3,74 |

Таблица 9.14 – Таблица со значениями критериев для независимых выборок в кластерах для переменной «Среднегодовые расходы на приобретение (обновление) элементов домашней мебели»

| Равные дисперсии предполагаются | Критерий равенства дисперсий Ливиня | | *t*-критерий для равенства средних | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | *F* | знч. | *t* | степ. своб. | знч. (2-стор.) | средн. разность | средн. квадр. ошибка разности | 95%-й доверит. интервал разности | |
| | | | | | | | | нижн. | верхн. |
| Да | 101,38 | 0,00 | –37,45 | 1438 | 0,00 | –265 | 7,09 | –279,2 | –251,4 |
| Нет | | | –57,70 | 410,27 | 0,00 | –265 | 4,60 | –274,3 | –256,2 |

На основании значений, приведенных в таблицах 9.13 и 9.14, можно сделать следующие выводы:

– в домашних хозяйствах первого кластера среднегодовые расходы на приобретение (обновление) элементов домашней мебели равны 302,17 р., второго – 567,45;

– так как значимость критерия равенства дисперсий Ливиня примерно равна 0,00, что меньше 0,05, гипотеза о равенстве дисперсий переменной «Среднегодовые расходы на приобретение (обновление) элементов домашней мебели» в кластерах отвергается;

– так как значимость *t*-критерия также не превышает 0,05, нулевая гипотеза о равенстве средних значений переменной «Среднегодовые расходы на

приобретение (обновление) элементов домашней мебели» в кластерах должна быть отклонена.

Следовательно, выделенные два кластера различаются между собой по значению переменной «Среднегодовые расходы на приобретение (обновление) элементов домашней мебели» и это различие является статистически значимым.

14 Итоговые результаты выполнения *t*-тестов представлены в таблице 9.15.

Таблица 9.15 – Итоговые результаты выполненных *t*-тестов сравнения средних значений характеристик в выделенных кластерах

| Переменная | Кластер 1 | Кластер 2 | Различие статистически значимо |
|---|---|---|---|
| Количество семей | 1253 | 187 | |
| Среднее количество членов домохозяйства, чел. | 3,73 | 5,74 | да |
| Семейный среднемесячный доход, р. | 6041,34 | 11348,13 | да |
| Среднее количество автомобилей в семье | 1,03 | 2,22 | да |
| Вид жилья | квартира | дом | да |
| Средняя площадь жилья, кв. м | 67,09 | 247,96 | да |
| Среднегодовые расходы на покупку (обновление) элементов домашней мебели, р. | 302,17 | 567,45 | да |

Таким образом, в результате кластерного анализа гипотеза, выдвинутая ранее по итогам корреляционно-регрессионных анализов, выполненных в лабораторных работах № 7 (с использованием независимой переменной «Семейный среднемесячный доход») и № 8 (со всеми независимыми переменными), о наличии на целевом рынке продукции ЧУП «Кэтнес» двух значимо различающихся сегментов еще раз подтверждена.

Первый сегмент (примерно 87,0 % рынка) составляют семьи, средний состав которых равен примерно 3,73 чел., имеющие среднемесячный доход в размере примерно 6041,34 р., проживающие в квартирах, средняя площадь которых равна примерно 67,09 кв. м, в основном обладающие примерно одним автомобилем, расходующие на приобретение (обновление) элементов домашней мебели примерно 302,17 р. в год.

Второй сегмент (примерно 13,0 % рынка) составляют семьи, средний состав которых равен примерно 5,74 чел., имеющие среднемесячный доход в размере примерно 11348,13 р., проживающие в домах, в основном обладающие более чем двумя автомобилями и расходующие на приобретение (обновление) элементов домашней мебели примерно 567,45 р. в год.

## 9.3 Задание для самостоятельного выполнения

В созданном файле «09 Кластерный анализ», используя в качестве меры квадрат расстояния Евклида, выполнить кластерный анализ методами межгрупповой и внутригрупповой связи, «ближайшего соседа» и «наиболее удаленного соседа».

## 9.4 Вопросы для самоконтроля

1 Что представляет собой кластерный анализ и в каком порядке он проводится?

2 Какие метрики могут использоваться для оценки сходства объектов при проведении кластерного анализа?

3 Какие методы кластерного анализа относятся к иерархическим?

4 Какие методы кластерного анализа относятся к неиерархическим (итеративным)?

5 Каков порядок проведения иерархического кластерного анализа?

6 Какие меры сходства объектов используются для объединения двух кластеров?

7 Каков порядок выполнения вычислительных процедур при использовании итеративных методов классификации кластерного анализа?

# ЛАБОРАТОРНАЯ РАБОТА № 10

## Дискриминантный анализ данных, полученных по выборке в процессе маркетингового исследования

**Цель работы:** выполнить дискриминантный анализ для кластеров (сегментов рынка продукции ЧУП «Кэтнес»), полученных по итогам выполнения лабораторной работы № 9.

**Задачи работы:** с использований теоретических знаний, полученных при изучении тем № 17 и 18, а также изученного ранее курса «Прикладной статистический анализ»:

– изучить порядок выполнения дискриминантного анализа данных, полученных по выборке в процессе маркетингового исследования;

– получить практические навыки в выполнении дискриминантного анализа данных, полученных в ходе маркетингового исследования, с использованием программы IBM SPSS Statistics.

### 10.1 Теоретические сведения

#### 10.1.1 Основные термины

**Дискриминантный анализ** – это одно из направлений математической статистики, содержанием которого является разработка методов решения задач разделения (дискриминации) неоднородной совокупности (множества) объектов наблюдения на однородные группы (подмножества, классы) по определенным признакам.

**Дискриминантные переменные** – это признаки, которые используются для того, чтобы отличать одну группу (подмножество, класс) объектов наблюдения от другого.

**Канонические дискриминантные функции** – это ортогональные оси, в максимальной степени различающие центроиды групп (подмножеств, классов).

**Структурные коэффициенты канонических дискриминантных функций** – это коэффициенты, отражающие связь (корреляцию) дискриминантных переменных с каноническими дискриминантными функциями и показывающие вклад каждой дискриминантной переменной в различительную способность соответствующей функции.

**λ Уилкса** – это критерий, используемый при проведении теста и показывающий, значимо ли различаются между собой средние значения дискриминантной функции в исследуемых классах (подмножествах).

### 10.1.2 Порядок проведения дискриминантного анализа

Порядок проведения дискриминантного анализа показан на рисунке 10.1.


Рисунок 10.1 – Порядок проведения дискриминантного анализа

#### 10.1.2.1 Формулирование проблемы дискриминантного анализа

Формулирование проблемы путем определения целей анализа, зависимой и независимых переменных является первым шагом дискриминантного анализа.

Зависимая переменная должна состоять из двух или больше взаимоисключающих и взаимно исчерпывающих категорий. Если зависимая переменная измерена с помощью интервальной или относительной шкалы, ее следует в первую очередь перевести в статус категориальной.

Предикторы (факторы) следует выбирать исходя из теоретической модели или ранее проведенного описательного исследования или, в случае поискового исследования, из интуиции и опыта исследователя.

#### 10.1.2.2 Определение коэффициентов дискриминантной функции при наличии двух групп (подмножеств, классов)

На рисунке 10.2 изображено пять объектов, которые принадлежат двум различным группам (далее – подмножествам) $M_1$ и $M_2$. Каждый объект в рассматриваемом случае характеризуется двумя переменными $x_1$ и $x_2$. Если рассмотреть проекции этих объектов (точек) на каждую ось, то можно увидеть, что эти подмножества пересекаются, так как по каждой переменной отдельно некоторые их объекты имеют близкие значения.

Чтобы наилучшим образом разделить рассматриваемые подмножества, нужно построить соответствующую линейную комбинацию переменных $x_1$ и $x_2$. Для двумерного пространства эта задача сводится к определению новой системы координат. Причем новые оси, например, $L$ и $C$, должны быть расположены таким образом, чтобы проекции объектов, принадлежащих разным подмножествам на

ось $L$, были максимально отличимы. Ось $C$ является перпендикулярной оси $L$ и разделяет два подмножества точек наилучшим образом, т. е. в этом случае подмножества оказались по разные стороны от этой прямой.

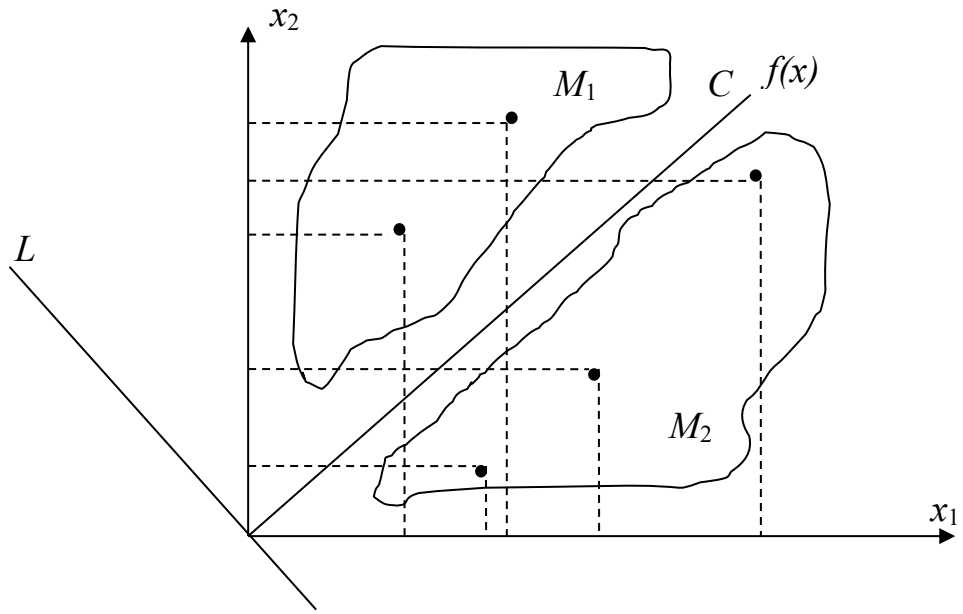



Рисунок 10.2 – Геометрическая интерпретация дискриминантной функции и дискриминантных переменных

При этом вероятность ошибки классификации должна быть минимальной. Сформулированные условия должны быть учтены при определении коэффициентов $\beta_1$ и $\beta_2$ функции, которая называется канонической дискриминантной:

$$f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \,. \qquad (10.1)$$

Если обозначить через $\overline{x}_{ij}$ среднее значение $j$-го признака у объектов $i$-го подмножества, то для подмножеств $M_1$ и $M_2$ средние значения их функций будут равны:

$$\overline{f}_1(x) = \beta_{0_1} + \beta_1 \overline{x}_{11} + \beta_2 \overline{x}_{12} \,. \qquad (10.2)$$

$$\overline{f}_2(x) = \beta_{0_2} + \beta_1 \overline{x}_{21} + \beta_2 \overline{x}_{22} \,. \qquad (10.3)$$

Геометрическая интерпретация этих функций представляет собой две параллельные прямые, проходящие через центры подмножеств (рисунок 10.3).

Для вычисления коэффициентов дискриминантной функции используется один из двух методов:

– прямой метод, который предполагает одновременное введение всех переменных. В этом случае учитывается каждая из них, но дискриминирующая ее сила при этом не принимается во внимание. Этот метод применяется тогда, когда, исходя из результатов предыдущего исследования, целесообразно, чтобы в основе различения лежали все переменные;

– пошаговый метод, когда переменные вводят последовательно, исходя из их способности различить (дискриминировать) подмножества. Этот метод рекомендуется применять в ситуации, когда стоит задача отобрать переменные для включения их в дискриминатную функцию. Коэффициенты дискриминантной функции $\beta_i$ определяются таким образом, чтобы $\overline{f_1}(x)$ и $\overline{f_2}(x)$ как можно больше различались между собой, т. е. чтобы для двух подмножеств выполнялось условие

$$\overline{f}_1(x) - \overline{f}_2(x) = \sum_{i=1}^{n_1} \beta_i x_{1i} - \sum_{i=1}^{n_2} \beta_i x_{2i} \to \max. \qquad (10.4)$$

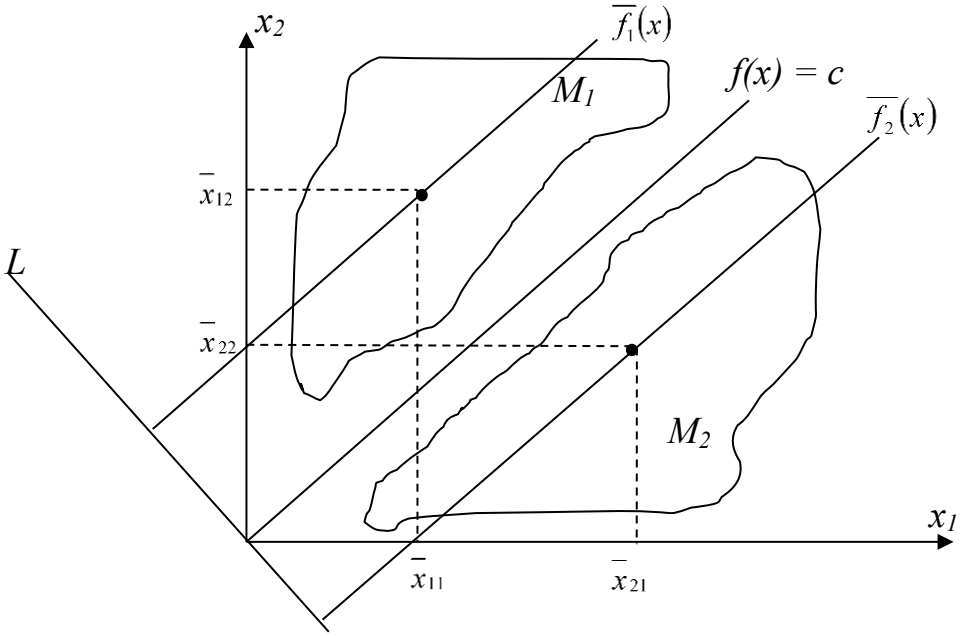



Рисунок 10.3 – Центры разделяемых подмножеств и константа дискриминации

При этом для каждого наблюдения *i* в *k*-й группе можно записать следующее выражение:

$$f_{ki} - \overline{f}_k(x) = \\ = \beta_1(x_{1ki} - \overline{x}_{1k}) + \beta_2(x_{2ki} - \overline{x}_{2k}) + \cdots + \beta_p\left(x_{pki} - \overline{x}_{pk}\right) \to \min. \qquad (10.5)$$

где $k$ – номер группы;
$i$ – номер наблюдения;
$p$ – число переменных, характеризующих каждое наблюдение.

Полученные коэффициенты подставляют в формулу канонической дискриминантной функции, для каждого объекта в обоих подмножествах вычисляют дискриминантные функции, а затем находят среднее значение для каждого подмножества. Таким образом, каждое *i*-е наблюдение, которое изначально

описывалось $p$ переменными, будет как бы помещено в одномерное пространство, т. е. ему будет соответствовать одно значение дискриминантной функции.

Прежде чем приступить непосредственно к классификации объектов, следует определить границу, разделяющую рассматриваемые подмножества. В качестве такой величины, называемой «константой дискриминации», может быть значение функции, равноудаленное от $\overline{f}_1(x)$ и $\overline{f}_2(x)$, т. е.

$$C = \frac{\overline{f}_1(x)+\overline{f}_2(x)}{2}. \quad (10.6)$$

#### 10.1.2.3 Установление значимости дискриминантной функции

Для установления достоверности дискриминантного анализа следует выполнить статистическую проверку нулевой гипотезы о равенстве средних всех дискриминантных функций во всех группах рассматриваемого множества:

$$H_0: \overline{f}_1(x) = \overline{f}_2(x) = \cdots = \overline{f}_k(x). \quad (10.7)$$

Проверка выполняется с использованием значения λ Уилкса, которое может находиться в интервале от нуля до единицы. Этот критерий используется для оценки достоверности различения подмножеств при помощи конкретного набора переменных и в качестве меры остаточной дискриминантной способности переменных при учете данного набора канонических функций. Считается, что чем ближе значение λ Уилкса к нулю, тем лучше данная каноническая функция (или весь их набор) различает объекты.

#### 10.1.2.4 Интерпретация полученных результатов

Интерпретация дискриминантных коэффициентов аналогична интерпретации коэффициентов во множественном регрессионном анализе. Значение коэффициента для конкретной переменной зависит от других переменных, включенных в дискриминантную функцию. Знаки коэффициентов условны, но они указывают, какие значения переменной приводят к большим и маленьким значениям функции, и связывают их с конкретными группами.

При наличии мультиколлинеарности между независимыми переменными однозначной меры относительной важности предикторов для дискриминации между группами не существует. Помня об этом предостережении, можно получить некоторое представление об относительной важности переменных, изучив абсолютные значения нормированных коэффициентов дискриминантной функции. Как правило, переменные с относительно большими нормированными коэффициентами вносят больший вклад в дискриминирующую мощность функции по сравнению с переменными, имеющими меньшие коэффициенты.

Некоторое представление об относительной важности переменных можно также получить, изучив структурные коэффициенты корреляции, которые также

называют «каноническими» или «дискриминантными нагрузками». Эти линейные коэффициенты корреляции между каждой из переменных и дискриминантной функцией представляют дисперсию, которую переменная делит вместе с функцией. Как и нормированные коэффициенты, эти коэффициенты корреляции следует использовать осторожно.

При интерпретации результатов дискриминантного анализа также может помочь разработка характеристической структуры для каждой группы посредством описания каждой группы через групповые средние для переменных.

Если важные переменные установлены, то сравнение групповых средних по ним может помочь понять межгрупповые различия. Однако, прежде чем интерпретировать какие-либо факты, необходимо убедиться в их достоверности.

#### 10.1.2.5 Оценка достоверности дискриминантного анализа

Для оценки достоверности дискриминантного анализа, выполненного в рамках маркетингового исследования, выборку разбивают случайным образом на две подвыборки. Анализируемую часть выборки используют для вычисления дискриминантной функции, а проверочную – для построения классификационной матрицы. Дискриминантные веса, определенные анализируемой выборкой, умножают на значения независимых переменных в проверочной выборке, чтобы получить дискриминантные показатели для случаев в этой выборке. Затем случаи распределяют по группам исходя из дискриминантных показателей и соответствующего правила принятия решения. Например, при дискриминантном анализе двух групп случай может быть отнесен к группе с самым близким по значению центроидом. Сложив элементы, лежащие на диагонали матрицы, и разделив полученную сумму на общее количество случаев, можно определить коэффициент результативности или процент верно классифицированных случаев.

Полезно сравнить процент случаев, верно классифицированных с помощью дискриминантного анализа, с процентом случаев, который можно получить случайным образом. Для равных по размеру групп процент случайной классификации равен частному от деления единицы на количество групп.

Большинство программ, позволяющих выполнить дискриминантный анализ, также определяют классификационную матрицу исходя из анализируемой выборки. Поскольку программы учитывают даже случайные вариации в данных, то полученные результаты всегда точнее, чем классификация данных на основе проверочной выборки.

## 10.2 Выполнение дискриминантного анализа с использованием программы IBM SPSS Statistics

По итогам лабораторной работы № 9 в анализируемой выборке были выделены два кластера (сегмента потребителей продукции ЧУП «Кэтнес»), которые статистически значимо различаются между собой по пяти независимым

переменным, которые с зависимой переменной имеют сильную и статистически значимую корреляционную связь. Для кластеризации использовалась и сама зависимая переменная.

Теперь по итогам кластерного анализа необходимо рассчитать каноническую дискриминантную функцию, которую целесообразно использовать для разделения объектов наблюдения (домохозяйств) на отдельные непересекающиеся кластеры (сегменты рынка).

Работу выполнить в следующем порядке:

1 Создать копию файла «09 Выделенные кластеры.sav» и присвоить ему имя «10 Дискриминантный анализ.sav».

2 Выполнить в созданном файле дискриминантный анализ. Для этого:

– выбрать процедуру дискриминантного анализа (**«Анализ»** – **«Классификация»** – **«Дискриминантный анализ…»**);

– в открывшемся диалоговом окне в левой части выделить переменную «Кластеры» и, нажав кнопку со стрелкой, направленной вправо, перенести ее в поле **«Группировать по:»** (рисунок 10.4);

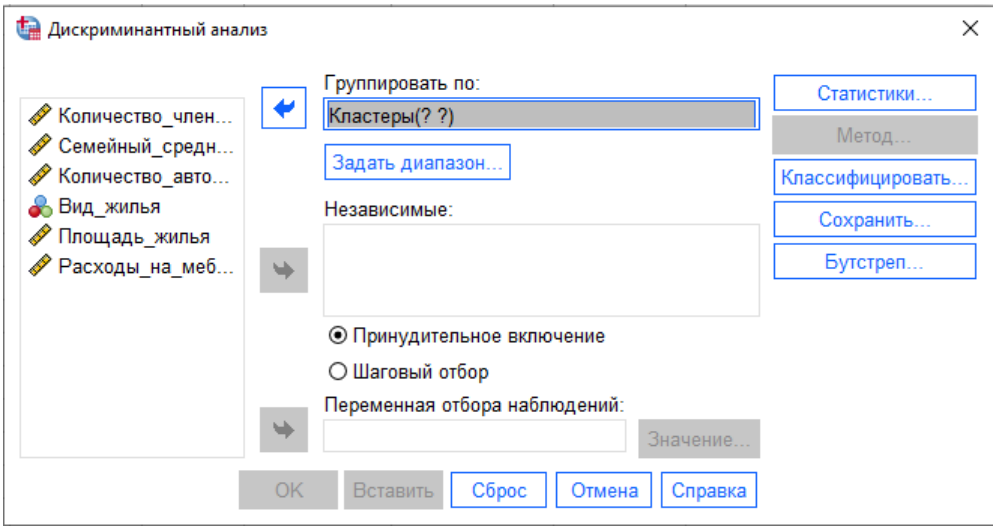



Рисунок 10.4 – Диалоговое окно «Дискриминантный анализ» с введенной группирующей переменной (принадлежностью к кластерам)

– нажать кнопку **«Задать диапазон…»**, задать минимальное и максимальное количество кластеров, равное соответственно «1» и «2» (рисунок 10.5), и нажать кнопку **«Продолжить»**;

– в левом поле диалогового окна **«Дискриминантный анализ»** выделить оставшиеся переменные и, нажав кнопку со стрелкой, направленной вправо, перенести их в поле **«Независимые»**;

– в диалоговом окне **«Дискриминантный анализ»** установить **«Принудительное включение»**;

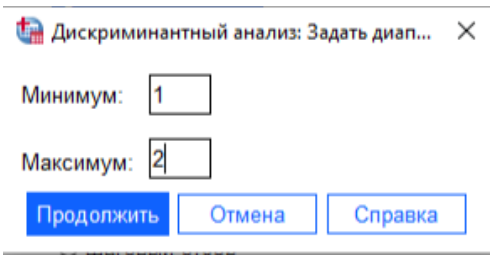



Рисунок 10.5 – Диалог «Дискриминантный анализ: Задать диапазон» с заданным диапазоном кластеров

– нажать кнопку **«Статистики…»**, после чего в появившемся диалоге в секции **«Описательные статистики»** установить флажки напротив строк **«Средние»** и **«Однофакторный дисперсионный анализ»**, в секции **«Матрицы»** – напротив строки **«Внутригрупповая корреляция»**, а в секции **«Коэффициенты функции»** – напротив строки **«Нестандартизованные»** (рисунок 10.6) и нажать кнопку **«Продолжить»**;

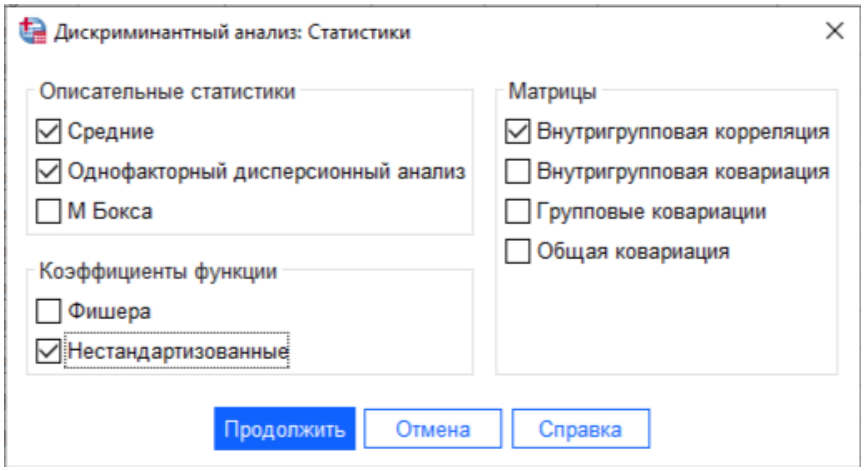



Рисунок 10.6 – Диалог «Дискриминантный анализ: Статистики» с запрошенными показателями

– нажать кнопку **«Классифицировать…»**, в появившемся диалоге в секции **«Априорные вероятности»** установить **«Все группы равны»**, в секции – **«Ковариационная матрица»** – **«Внутригрупповая»**, в секции **«Вывести»** – установить флажки напротив строк **«Поточечные результаты»** и **«Итоговая таблица»**, в секции **«Графики»** – **«Отдельно по группам»** (рисунок 10.7), нажать кнопку **«Продолжить»**;

– нажать кнопку **«Сохранить…»**, после чего в открывшемся диалоге установить флажок напротив строки **«Предсказанная принадлежность к группе»** (рисунок 10.8) и нажать кнопку **«Продолжить»**;

– нажать кнопку **«ОК»** диалогового окна **«Дискриминантный анализ»**.

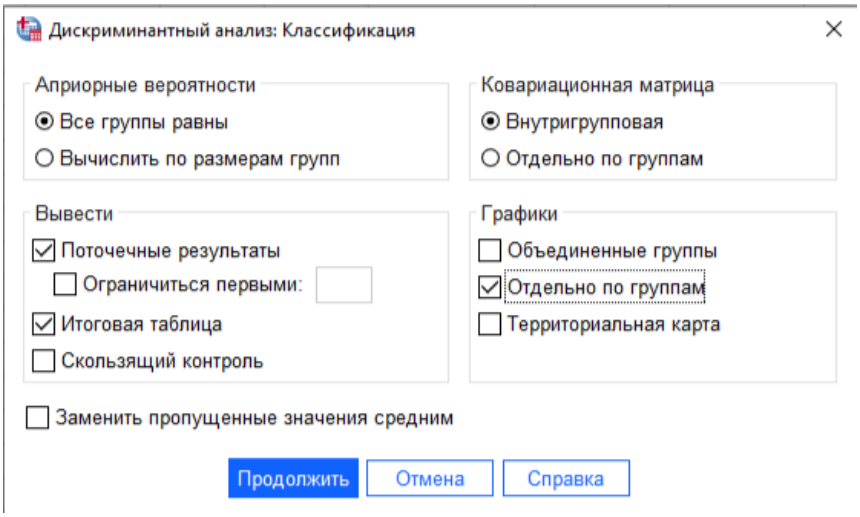



Рисунок 10.7 – Диалог «Дискриминантный анализ: Классификация» с установленными условиями для проведения дискриминантного анализа

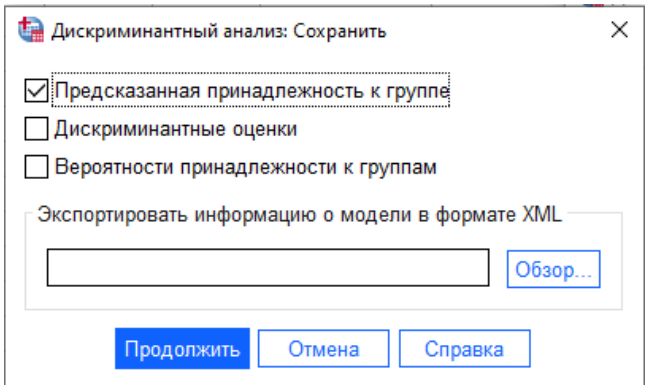



Рисунок 10.8 – Диалог «Дискриминантный анализ: Сохранить» с установленными условиями для проведения дискриминантного анализа

3 В открывшемся файле с результатами дискриминантного анализа обратить внимание прежде всего на таблицы с названиями «Критерии равенства групповых средних» (таблица 10.1), «Собственные значения» (таблица 10.2), «λ Уилкса» (таблица 10.3), «Коэффициенты стандартизированной канонической дискриминантной функции» (таблица 10.4), «Коэффициенты канонической дискриминантной функции» (таблица 10.5), «Статистика по наблюдениям» (таблица 10.6), «Результаты классификации» (таблица 10.7), а также на графики разброса значений дискриминантной функции в выделенных кластерах (рисунки 10.9 и 10.10).

Величины *P*-значения λ Уилкса для каждой переменной меньше 0,05, поэтому ни одну из них не следует исключать из анализа.

Далее важно обратить внимание на таблицу «Переменные, не соответствующие критерию допуска», согласно которой переменная «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» не может быть использована для расчета коэффициентов канонической дискриминантной функции из-за того, что уровень ее допуска меньше 0,001.

Таблица 10.1 – Критерии равенства групповых средних

| Переменные | λ Уилкса | Значение *F*-критерия | Степени свободы 1 | Степени свободы 2 | Значимость (*P*-значение) |
|---|---|---|---|---|---|
| Количество членов домохозяйства | 0,75 | 492,78 | 1 | 1438 | 0,00 |
| Семейный среднемесячный доход | 0,51 | 1403,39 | 1 | 1438 | 0,00 |
| Количество автомобилей в семье | 0,70 | 711,61 | 1 | 1438 | 0,00 |
| Вид жилья | – | – | – | – | – |
| Площадь жилья | 0,11 | 11260,13 | 1 | 1438 | 0,00 |
| Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели | 0,51 | 1402,13 | 1 | 1438 | 0,00 |
| *Примечание* – Статистики для переменной «Вид жилья» программой не вычислены, так как ее значения (либо «1» («квартира»), либо «2» («дом»)) являются одинаковыми для каждого кластера | | | | | |

Таблица 10.2 – Таблица с собственными значениями канонической дискриминантной функции

| Собственное значение | Процент объясненной дисперсии | Кумулятивный процент | Каноническая корреляция |
|---|---|---|---|
| 159,23 | 100,00 | 100,00 | 0,997 |

Мерой качества разделения выделенных кластеров является λ Уилкса. Его значение, равное 0,006, говорит о достаточно высоком качестве разделения.

Таблица 10.3 – Таблица с результатами теста λ Уилкса

| λ Уилкса | Критерий $\chi^2$ Пирсона | Степени свободы | Значимость (*P*-значение) |
|---|---|---|---|
| 0,006 | 7289,98 | 4 | 0,00 |

Величина значимости λ Уилкса подтверждает гипотезу о том, что средние значения канонической дискриминантной функции в обоих кластерах различаются и различие является статистически значимым.

Таблица 10.4 – Таблица со значениями коэффициентов стандартизированной канонической дискриминантной функции

| Переменные | Коэффициенты при переменных |
|---|---|
| Количество членов домохозяйства | –5,74 |
| Семейный среднемесячный доход | 0,48 |
| Количество автомобилей в семье | –0,16 |
| Площадь жилья | 5,58 |
| *Примечание* – Переменные «Вид жилья» и «Среднегодовые расходы на покупку (обновление) элементов домашней мебели» не соответствуют минимальному критерию допуска, равному 0,001 | |

Таблица 10.5 – Таблица со значениями коэффициентов канонической дискриминантной функции

| Переменные | Коэффициенты при переменных |
|---|---|
| Количество членов домохозяйства | –4,95 |
| Семейный среднемесячный доход | 0,00 |
| Количество автомобилей в семье | –0,28 |
| Площадь жилья | 0,26 |
| Свободный член (константа) | –4,94 |
| *Примечание* – Переменные «Вид жилья» и «Среднегодовые расходы на покупку (обновление) элементов домашней мебели» не соответствуют минимальному критерию допуска, равному 0,001 | |

Если теперь независимые переменные, прошедшие проверку критерия допуска, обозначить как $x_1$ («Количество членов домохозяйства»), $x_2$ («Семейный среднемесячный доход»), $x_3$ («Количество автомобилей в семье») и $x_4$ («Площадь жилья»), то каноническая дискриминантная функция, разделяющая оба кластера (сегмента рынка), будет выглядеть следующим образом:

– со стандартизированными коэффициентами:

$$y_x = -5{,}74x_1 + 0{,}48x_2 - 0{,}16x_3 + 5{,}58x_4;$$

– с нестандартизованными коэффициентами:

$$y_x = -4{,}94 - 4{,}95x_1 - 0{,}28x_3 + 0{,}26x_4.$$

Значения функции в центроидах выделенных в ходе лабораторной работы № 9 кластеров, вычисленные с использованием нестандартизованных коэффициентов, равны соответственно –4,87 и 32,64.

Разброс значений дискриминантной функции в выделенных кластерах показан на рисунках 10.9 и 10.10. Как видно из них, области разброса значений функций между собой не пересекаются, что говорит о достаточно четком различии состава кластеров. Однако во втором кластере можно предположить наличие двух групп, состоящих из 42 и 145 домохозяйств (составляющих 3 и 10 % от

размера изучаемой выборки), которые возможно статистически значимо различаются по всем переменным, указанным в таблицах 10.4 и 10.5.

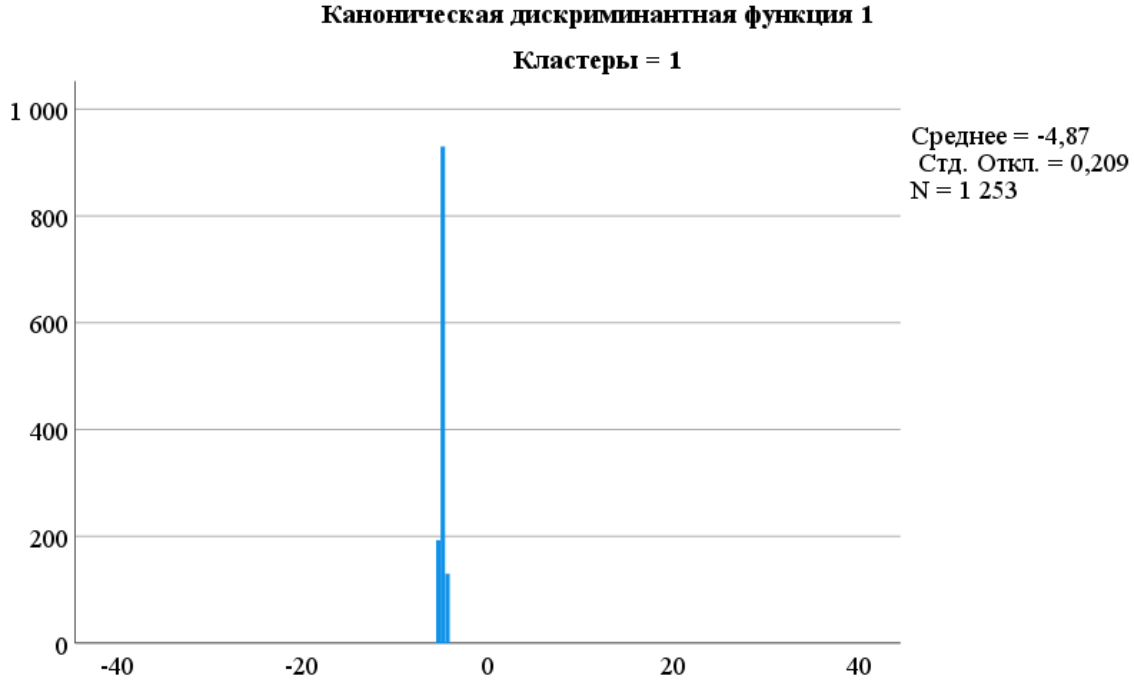



Рисунок 10.9 – Распределение значений дискриминантной функции для первого кластера


Рисунок 10.10 – Распределение значений дискриминантной функции для второго кластера

Содержание таблицы «Статистика по наблюдениям» показывает, что номера групп, к которым домохозяйства принадлежат фактически по итогам кластерного анализа, оказались равными их номерам, предсказанным по итогам дискриминантного анализа (таблица 10.6).

Оценка точности прогноза по итогам дискриминантного анализа представлена в таблице 10.7. По сути, дискриминантный анализ подтвердил результаты кластерного анализа, выполненного в лабораторной работе № 9.

Таблица 10.6 – Фрагмент таблицы с распределением домашних хозяйств по фактической и предсказанной принадлежности к кластерам

| Номер домохозяйства по итогам кластерного анализа | Принадлежность к кластеру | | Дискриминационные баллы (значение дискриминантной функции) |
|---|---|---|---|
| | фактическая | предсказанная | |
| 1 | 1 | 1 | –4,62 |
| 2 | 1 | 1 | –4,65 |
| 3 | 1 | 1 | –4,60 |
| 4 | 1 | 1 | –4,69 |
| 5 | 1 | 1 | –5,11 |
| 6 | 1 | 1 | –4,65 |
| 7 | 1 | 1 | –4,76 |
| 8 | 1 | 1 | –4,75 |
| 9 | 1 | 1 | –4,88 |
| 10 | 1 | 1 | –4,94 |
| … | … | … | … |
| 1251 | 1 | 1 | –5,50 |
| 1252 | 1 | 1 | –5,21 |
| 1253 | 1 | 1 | –4,88 |
| 1254 | 2 | 2 | 34,25 |
| 1255 | 2 | 2 | 34,49 |
| 1256 | 2 | 2 | 27,85 |
| 1257 | 2 | 2 | 34,25 |
| 1258 | 2 | 2 | 34,26 |
| 1259 | 2 | 2 | 34,36 |
| 1260 | 2 | 2 | 34,64 |
| … | … | … | … |
| 1431 | 2 | 2 | 34,02 |
| 1432 | 2 | 2 | 34,22 |
| 1433 | 2 | 2 | 34,22 |
| 1444 | 2 | 2 | 28,12 |
| 1435 | 2 | 2 | 33,98 |
| 1436 | 2 | 2 | 28,21 |

| Номер домохозяйства по итогам кластерного анализа | Принадлежность к кластеру | | Дискриминационные баллы (значение дискриминантной функции) |
|---|---|---|---|
| | фактическая | предсказанная | |
| 1437 | 2 | 2 | 34,30 |
| 1438 | 2 | 2 | 27,82 |
| 1439 | 2 | 2 | 33,98 |
| 1440 | 2 | 2 | 28,12 |

Таблица 10.7 – Результаты распределения домохозяйств по выделенным кластерам

| Параметр | Кластеры | Предсказанная принадлежность к кластеру | | Всего |
|---|---|---|---|---|
| | | 1 | 2 | |
| Количество | 1 | 1253 | 0 | 1253 |
| | 2 | 0 | 187 | 187 |
| Процент | 1 | 100,00 | 0,00 | 100,00 |
| | 2 | 0,00 | 100,00 | 100,00 |

## 10.3 Задание для самостоятельного выполнения

В созданном файле «09 Выделенные кластеры.sav» выполнить дискриминантный анализ с пошаговым отбором независимых переменных, интерпретировать его результаты и сравнить их с результатами, полученными при выполнении лабораторной работы.

## 10.4 Вопросы для самоконтроля

1 Каковы цель и порядок проведения дискриминантного анализа?

2 Как в двумерном пространстве выглядит геометрическая интерпретация результата дискриминантного анализа?

3 Что представляет собой геометрическая интерпретация в двумерном пространстве дискриминационных функций?

4 Какие методы используются для вычисления коэффициентов дискриминационной функции?

5 С использованием какого статистического критерия проводится проверка достоверности дискриминантного анализа?

# ЛАБОРАТОРНАЯ РАБОТА № 11

## Факторный анализ данных, полученных по выборке в процессе маркетингового исследования

**Цель:** выполнить факторный анализ независимых переменных, характеризующих объекты исследования (домохозяйства, являющиеся покупателями продукции ЧУП «Кэтнес»), для сокращения их числа (обобщения).

**Задачи работы:** с использованием теоретических знаний, полученных при изучении тем № 17–19, а также изученного ранее курса «Прикладной статистический анализ»:

– изучить порядок выполнения факторного анализа данных, полученных по выборке в процессе маркетингового исследования;

– получить практические навыки в выполнении факторного анализа данных, полученных в ходе маркетингового исследования, с использованием программы IBM SPSS Statistics.

### 11.1 Теоретические сведения

#### 11.1.1 Основные термины

**Факторный анализ** – это совокупность методов, которые на основе реально существующих связей признаков (или объектов) позволяют выявлять латентные обобщающие характеристики организационной структуры и механизма развития изучаемых явлений и процессов.

**Факторы** – это гипотетические, непосредственно не измеряемые, скрытые (латентные) переменные, в той или иной мере связанные с измеряемыми характеристиками и их проявлениями. Представляют собой искусственные статистические показатели, возникающие в результате специальных преобразований таблицы коэффициентов корреляции между изучаемыми признаками или матрицы интеркорреляций.

**Латентность** – это свойство объектов или процессов находиться в скрытом состоянии, не проявляя себя явным образом. В факторном анализе латентность – это неявность характеристик, раскрываемых с его помощью.

**Матрица интеркорреляций** – это создаваемая в ходе факторного анализа квадратная матрица типа «признак/признак», оцифрованная только номерами признаков и содержащая коэффициенты корреляции между ними.

**Факторизация матрицы интеркорреляций** – это процедура извлечения факторов из матрицы интеркорреляций.

**Матрица факторных нагрузок (факторная матрица)** – это способ представления результатов факторного анализа (метода главных компонент). Строки матрицы соответствуют исходным переменным, а столбцы – факторам (главным

компонентам). На пересечении строки и столбца указывается значение факторной нагрузки.

**Факторная нагрузка (вес)** – это коэффициент корреляции между измеряемой переменной и латентным фактором.

**Критерий сферичности Бартлетта** – это статистика, проверяющая нулевую гипотезу о том, что переменные в генеральной совокупности не коррелируют между собой.

**Критерий адекватности выборки Кайзера – Мейера – Олкина (КМО)** – это показатель, используемый для проверки целесообразности выполнения факторного анализа.

### 11.1.2 Порядок проведения факторного анализа

Порядок проведения факторного анализа данных, полученных в ходе маркетингового исследования, приведен на рисунке 11.1.

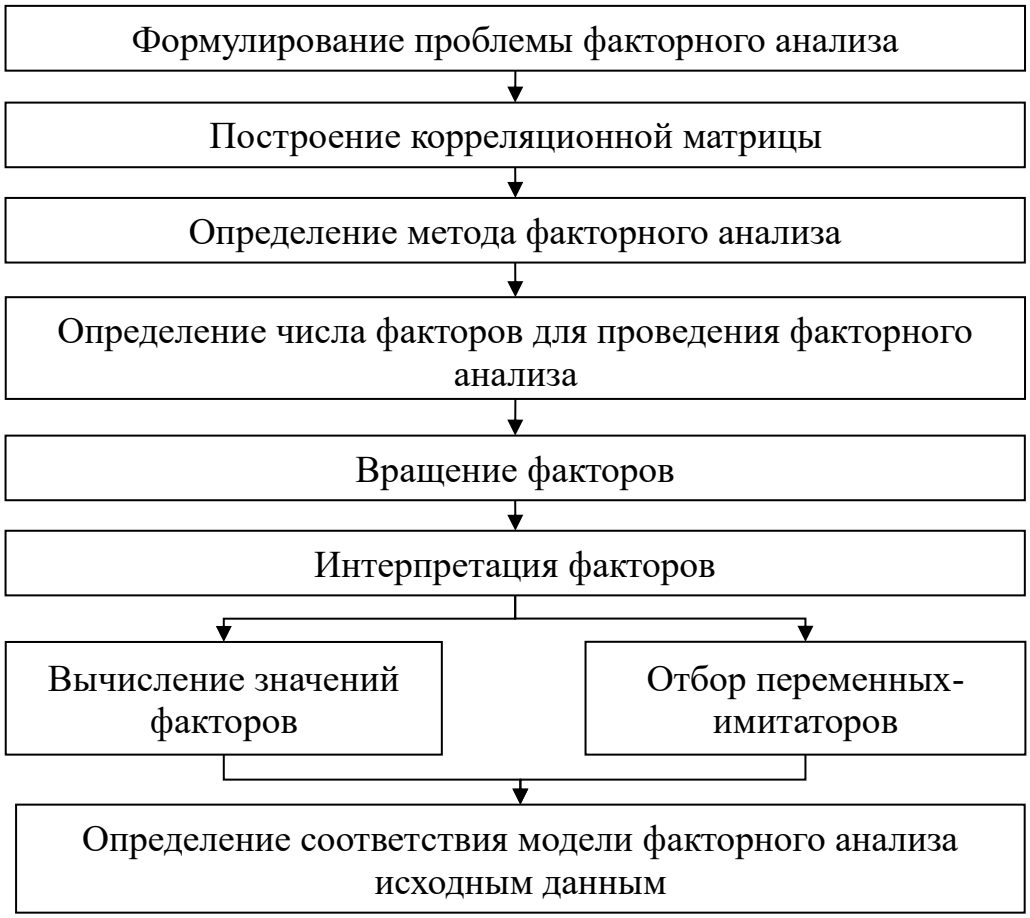



Рисунок 11.1 – Порядок выполнения факторного анализа

#### 11.1.2.1 Определение проблемы факторного анализа

Формулировка проблемы факторного анализа при проведении маркетингового исследования предполагает соблюдение следующих условий:

– четкое определение целей факторного анализа;

– переменные, подвергаемые факторному анализу, задаются исходя из прошлых исследований, теоретических выкладок и по усмотрению маркетолога-исследователя (важно, чтобы переменные измерялись в интервальной или относительной шкалах);

– выборка должна быть подходящего размера (рекомендуется брать выборку по крайней мере в четыре или пять раз больше, чем число переменных).

#### 11.1.2.2 Интеркорреляционная матрица

Факторному анализу подвергают матрицы интеркорреляций (корреляционные матрицы), которые содержат коэффициенты корреляции Пирсона, вычисленные для переменных (признаков), включенных в исследование (таблица 11.1).

Таблица 11.1 – Пример матрицы интеркорреляций

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ |
|---|---|---|---|---|---|---|
| $x_1$ | 1,000 | | | | | |
| $x_2$ | –0,041 | 1,000 | | | | |
| $x_3$ | 0,873 | –0,143 | 1,000 | | | |
| $x_4$ | –0,086 | 0,590 | –0,248 | 1,000 | | |
| $x_5$ | –0,858 | –0,007 | –0,778 | –0,007 | 1,000 | |
| $x_6$ | 0,004 | 0,713 | –0,018 | 0,640 | –0,136 | 1,000 |

#### 11.1.2.3 Методы факторного анализа

Общая классификация методов факторного анализа представлена на рисунке 11.2.

#### 11.1.2.4 Определение числа факторов

В результате специальных преобразований (факторизации) матрицы интеркорреляций возникают искусственные статистические показатели, называемые факторами. Из нее может быть извлечено разное количество факторов, вплоть до числа, равного количеству исходных переменных. Однако факторы, выделяемые в результате факторизации, как правило, неравноценны по своему назначению.

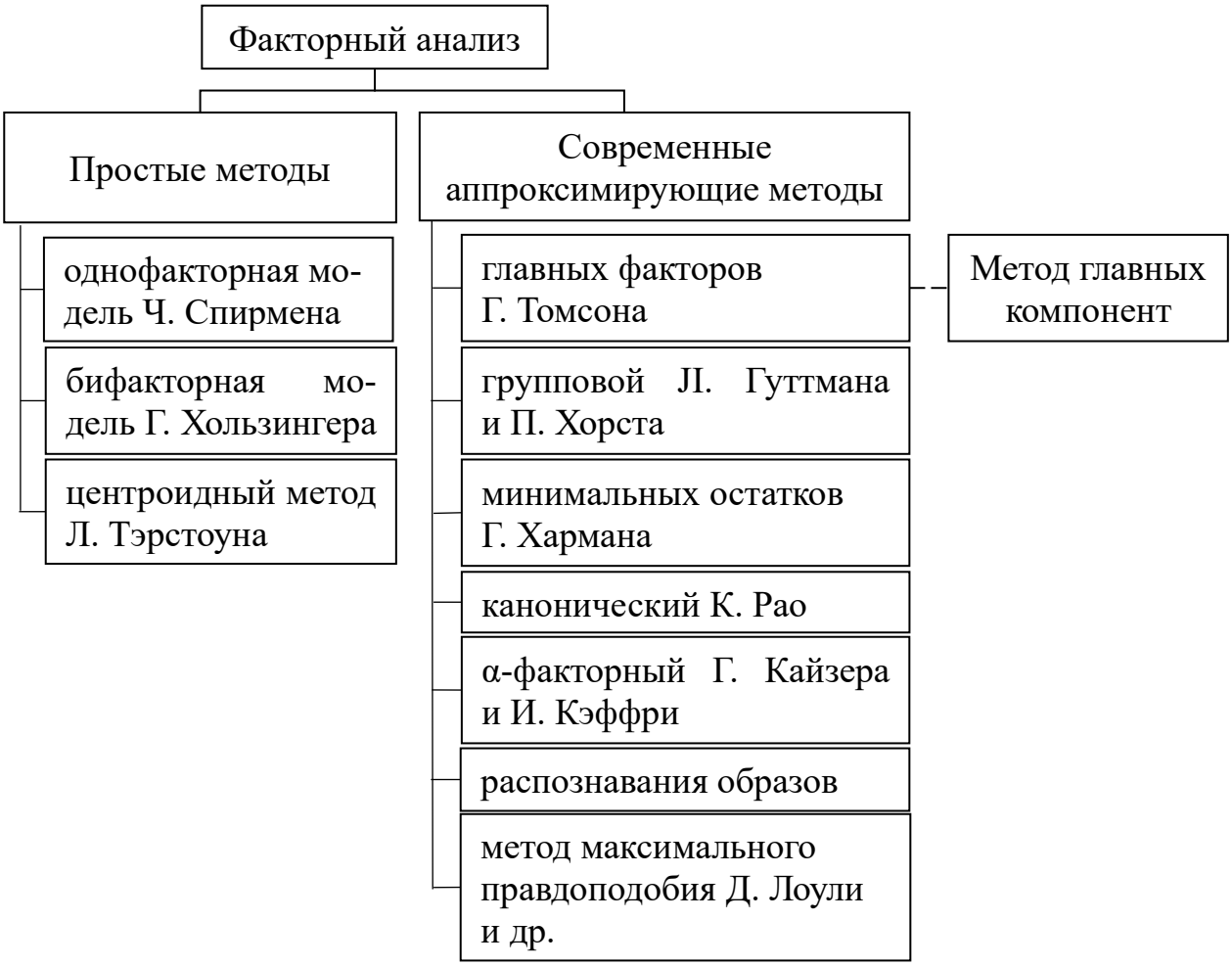



Рисунок 11.2 – Общая классификация методов факторного анализа

Факторная матрица (таблица 11.2) показывает, какие переменные образуют каждый фактор. Это связано прежде всего с уровнем значимости факторных весов. Традиционно минимальный уровень значимости коэффициентов корреляции в факторном анализе берется равным по абсолютной величине 0,4 или даже 0,3.

Таблица 11.2 – Пример факторной матрицы, полученной с использованием метода главных компонент

| Признаки | Главные факторы и нагрузки входящих в них переменных | |
|---|---|---|
| | 0,927 | 0,259 |
| 1 | –0,293 | 0,826 |
| 2 | 0,935 | 0,139 |
| 3 | –0,348 | 0,778 |
| 4 | –0,865 | –0,359 |
| 5 | –0,189 | 0,884 |
| 6 | 0,927 | 0,259 |

Выделенный в результате факторизации фактор представляет собой совокупность тех переменных из числа включенных в анализ, которые имеют значимые нагрузки. Однако нередко случается, что в фактор входит только одна переменная со значимым факторным весом, а остальные имеют незначимую факторную нагрузку. В этом случае фактор будет определяться по названию единственной значимой переменной.

Поскольку каждый фактор является своего рода переменной величиной, то сами факторы также могут коррелировать между собой. Здесь возможны два случая:

– корреляция между факторами равна нулю, и в таком случае факторы считаются независимыми (ортогональными);

– корреляция между факторами больше нуля, и в таком случае факторы считаются зависимыми (облическими).

Ортогональные факторы в отличие от облических дают более простые варианты взаимодействий внутри факторной матрицы.

Определение числа выделяемых факторов из корреляционной матрицы является одним из наиболее важных решений, которое маркетологу-исследователю необходимо будет принять при проведении факторного анализа, так как неверное решение может привести к бессмысленным результатам при обработке самого четкого набора данных по выборке.

Для определения числа факторов разработаны следующие процедуры:

– определение, основанное на предварительной информации, когда, руководствуясь ею, маркетолог знает, сколько факторов можно ожидать и, таким образом, может заранее определить сколько их будет извлечено. После извлечения желаемого числа факторов их выделение прекращают;

– определение, основанное на собственных значениях факторов, когда учитывают только те, собственные значения которых выше 1,0, а остальные факторы в модель не включают;

– определение, основанное на критерии «каменистая осыпь» (рисунок 11.3), графическое изображение которого представляет собой график зависимости собственных значений факторов от их номеров в порядке выделения;

– определение на основе процента объясненной дисперсии, при котором число выделяемых факторов определяют так, чтобы кумулятивный процент дисперсии, объясняемой факторами, достиг удовлетворительного уровня. Рекомендуется выделять такое число факторов, которое объясняет по крайней мере 60 % дисперсии;

– определение, основанное на оценке надежности, выполняемой расщеплением, при котором выборку делят на две равные части и факторный анализ выполняют для каждой половины. При этом оставляют только факторы с высокой степенью соответствия факторных нагрузок в двух подвыборках;

– определение, основанное на критериях значимости, когда можно определить статистическую значимость отдельных собственных значений и оставить только статистически значимые факторы.

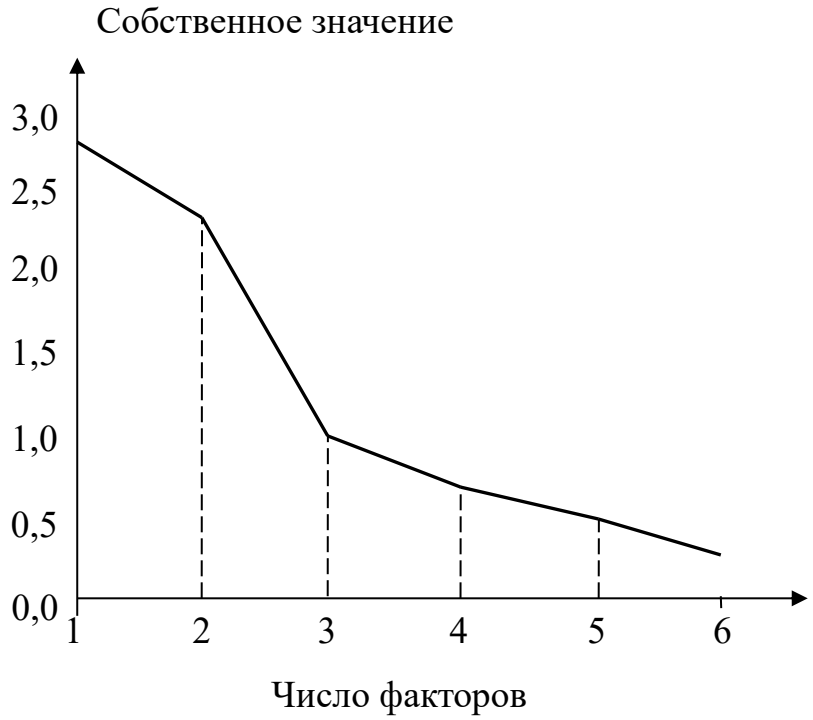



Рисунок 11.3 – Графическая иллюстрация критерия «каменистая осыпь»

Несмотря на то что матрица исходных (неповернутых) факторов указывает на их взаимосвязь с отдельными переменными, она редко приводит к факторам, которые можно интерпретировать, поскольку факторы коррелируют со многими переменными. Поэтому вращением матрицу факторных коэффициентов преобразуют в более простую, которую легче интерпретировать.

#### 11.1.2.5 Вращение факторов

Вращение факторов перемещает их относительно переменных таким образом, чтобы каждый из них начинал обладать несколькими существенными нагрузками и несколькими нагрузками, близкими к нулю. Таким образом, цель вращения заключается в том, чтобы получилась простая структура, в которой каждый фактор имел бы некоторое количество больших нагрузок и некоторое количество маленьких и, подобно этому, каждая переменная имела бы существенные нагрузки только по некоторым факторам (таблица 11.3).

При вращении факторов желательно, чтобы каждый фактор имел ненулевые или значимые нагрузки только для небольшого числа переменных. Аналогично желательно, чтобы каждая переменная имела ненулевые или значимые нагрузки с небольшим числом факторов, а если можно, то только с одним фактором. Если несколько факторов имеют высокие значения факторных нагрузок с одной и той же переменной, то их трудно интерпретировать. Вращение не влияет на общности и процент объясненной полной дисперсии. Однако процент дисперсии, обусловленной влиянием каждого фактора, изменяется.

Таблица 11.3 – Пример результатов факторного анализа до и после вращения пространства главных факторов

| Переменная, $x_j$ | Общие факторы | | | |
|---|---|---|---|---|
| | до вращения | | после вращения | |
| | $F_1$ | $F_2$ | $W_1$ | $W_2$ |
| Обеспеченность долгосрочными активами в расчете на среднегодовую численность персонала компании, $x_1$ | 0,46 | 0,80 | 0,92 | 0,09 |
| Уровень энерговооруженности труда, $x_2$ | 0,61 | 0,41 | 0,89 | 0,30 |
| Отдача долгосрочных активов, $x_3$ | 0,50 | 0,55 | 0,41 | 0,75 |
| Рентабельность выручки от реализации продукции, $x_4$ | 0,57 | 0,38 | 0,09 | 0,87 |
| Среднегодовая выработка в расчете на одного работника, $x_5$ | 0,72 | 0,67 | 0,27 | 0,57 |
| Уровень реализуемости товарной продукции, $x_6$ | 0,58 | 0,51 | 0,37 | 0,46 |

Вращение общих факторов может быть:

– ортогональным, при котором сохраняется прямоугольная система координат и взаимодействие факторов исключается;

– косоугольным, порождающим корреляционные связи латентных факторов, при котором прямоугольная система координат не сохраняется.

Применение любого из критериев качества структуры общих факторов означает, что после каждого поворота факторного пространства проводятся расчет критерия и соответствующая оценка качества структуры факторов. Вращение завершается, когда критерий, достигнув максимального (минимального) значения, на следующем шаге алгоритма начинает отклоняться от оптимума.

### 11.1.2.6 Интерпретация факторов

Для интерпретации факторов необходимо определить переменные, которые имеют высокие значения нагрузок по одному и тому же фактору. Затем этот фактор следует проанализировать с учетом этих переменных.

После интерпретации факторов необходимо вычислить их значения. Фактор представляет собой линейную комбинацию исходных переменных. Значение для $i$-го фактора можно вычислить по формуле

$$F_i = \alpha_{i1}x_1 + \alpha_{i2}x_2 + \cdots + \alpha_{ik}x_k \, . \qquad (11.1)$$

Веса или коэффициенты значений факторов $\alpha_{ij}$, используемые для объединения нормированных переменных, получают из матрицы коэффициентов значения фактора.

Иногда вместо вычисления значений факторов маркетолог-исследователь может выбрать переменные-заменители. Выбор переменных-заменителей

заключается в выделении нескольких из исходных переменных для использования их в последующем анализе. Это позволяет выполнить последующий анализ и интерпретировать результаты с точки зрения исходных переменных, а не значения факторов. Из матрицы факторных коэффициентов можно выбрать для каждого фактора переменную с наивысшим значением нагрузки на данный фактор. Затем эту переменную используют в качестве переменной-заменителя для соответствующего фактора. Этот процесс протекает без проблем, если одна из факторных нагрузок переменной значительно выше остальных. Однако сделать выбор не так легко, если нагрузки двух или больше переменных одинаково высокие. В таком случае выбор осуществляют исходя из теоретических предпосылок. Например, теоретически предполагают, что переменная с несколько большей нагрузкой важнее, чем переменная с несколько меньшей нагрузкой. Но если переменная имеет несколько меньшую, но более точно измеренную нагрузку, то в качестве переменной-имитатора рекомендуется использовать именно ее.

#### 11.1.2.7 Определение соответствия модели факторного анализа исходным данным

Последняя стадия факторного анализа заключается в определении соответствия модели факторного анализа исходным данным, т. е. степени ее подгонки. Основное допущение, лежащее в основе факторного анализа, состоит в том, что наблюдаемая корреляция между переменными может быть свойственна общим факторам. Следовательно, корреляции между переменными можно вывести или воспроизвести из определенных корреляций между переменными и факторами. Изучив разности между наблюдаемыми корреляциями (данными в исходной корреляционной матрице) и вычисленными корреляциями (определенными из матрицы факторных нагрузок), можно определить соответствие модели исходным данным. Эти разности называют «остатками». Если получено много остатков с большими значениями, то можно считать, что факторная модель не обеспечивает хорошее соответствие данным и требует пересмотра.

## 11.2 Выполнение факторного анализа с использованием программы IBM SPSS Statistics

По итогам выполнения дискриминантного анализа в лабораторной работе № 10 были установлены четыре переменные, которые позволили достаточно четко различить в составе изучаемой выборки два кластера (сегмента рынка).

Необходимо с использованием факторного анализа рассмотреть возможность сократить до минимума количество переменных для последующей общей характеристики домохозяйств в генеральной совокупности (на целевом рынке продукции ЧУП «Кэтнес»).

Работу выполнить в следующем порядке:

1 Скопировать файл «10 Дискриминантный анализ.sav» в папку, в которой будут сохранены результаты лабораторной работы, и присвоить ему имя «11 Факторный анализ.sav».

2 В созданном файле удалить:

– колонки со значениями переменных «Вид жилья» и «Среднегодовые расходы на покупку (обновление) элементов домашней мебели», которые в ходе дискриминантного анализа программой IBM SPSS Statistics были признаны не соответствующими минимальному критерию допуска, равному 0,001;

– колонки «Кластеры» и «Dis_1» с номерами кластеров, к которым по итогам кластерного и дискриминантного анализов были отнесены домашние хозяйства);

– колонку с номерами домашних хозяйств.

3 После этого выполнить факторный анализ. Для этого:

– выбрать процедуру факторного анализа (**«Анализ»** – **«Снижение размерности»** – **«Факторный анализ»**);

– в открывшемся диалоговом окне, в котором представлены используемые переменные, выделить их и, нажав кнопку со стрелкой, направленной вправо, перенести в поле **«Переменные:»** (рисунок 11.4);

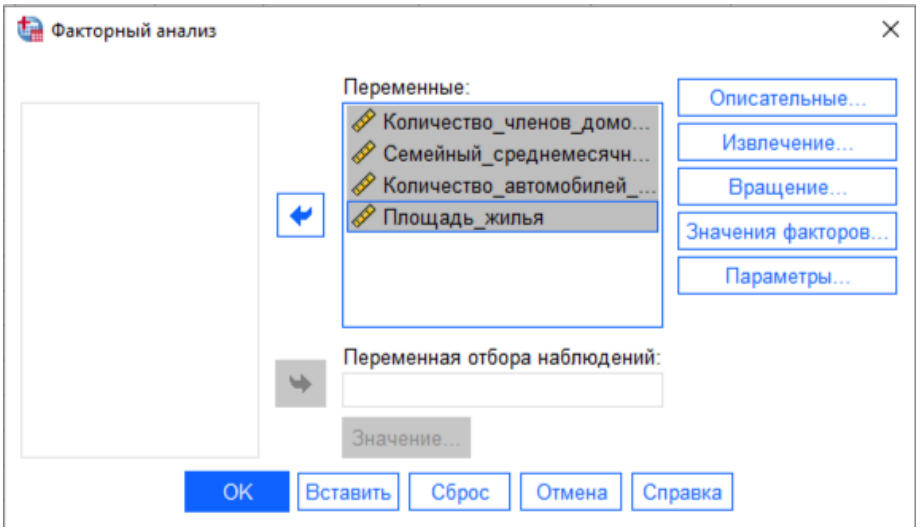



Рисунок 11.4 – Диалог «Факторный анализ» с введенными переменными

– нажать кнопку **«Описательные…»**, в открывшемся диалоге в секции **«Статистика»** поставить флажок напротив строки **«Начальное решение»**, а в секции **«Корреляционная матрица»** – напротив строк **«Коэффициенты»** и **«КМО и критерий сферичности Бартлетта»** (рисунок 11.5), после чего нажать кнопку **«Продолжить»**;

– нажать кнопку **«Извлечение…»** и в открывшемся диалоге выбрать метод главных компонент. В секции **«Анализ»** запросить матрицу корреляций, в секции **«Вывести»** установить флажки напротив строк **«Неповернутое решение»** и **«График собственных значений»**, а в секции **«Выделить»** указать, что должны быть выделены только те компоненты, собственное значение которых превышает

1,0 (рисунок 11.6). Установив количество итераций до сходимости 25, нажать кнопку **«Продолжить»**;

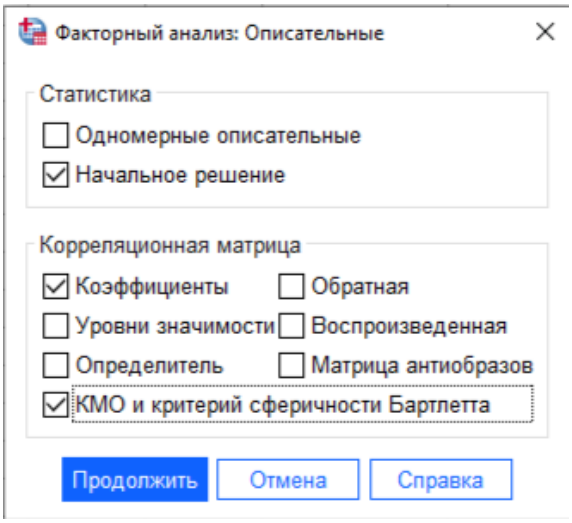



Рисунок 11.5 – Диалог «Факторный анализ: Описательные» с заданными условиями проверки пригодности данных

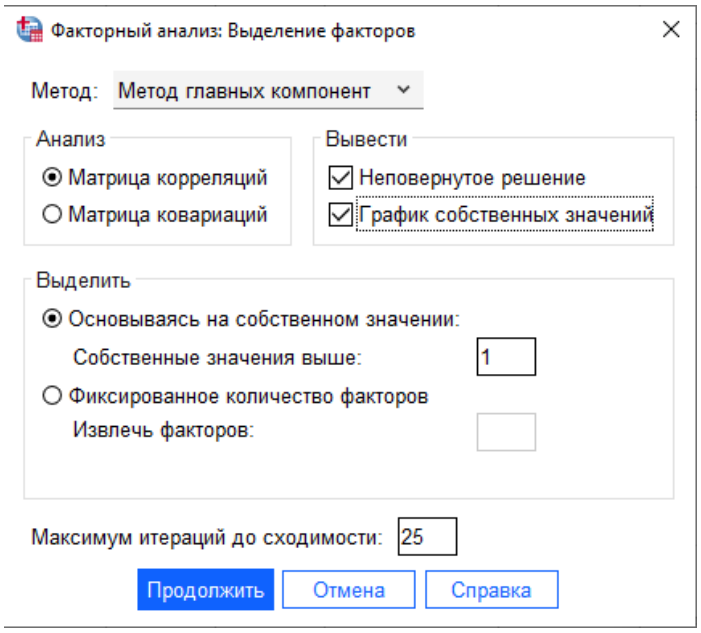



Рисунок 11.6 – Диалог «Факторный анализ: Выделение факторов» с заданными условиями выделения количества факторов

– нажать кнопку **«Значения факторов…»**, в открывшемся диалоге поставить флажки напротив строк **«Сохранить как переменные»** и **«Вывести матрицу коэффициентов значений факторов»**, в секции **«Метод»** выбрать **«Регрессия»** (рисунок 11.7) и нажать кнопку **«Продолжить»**;

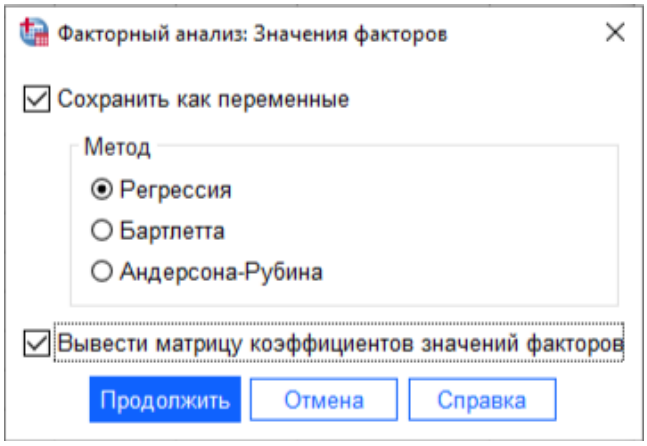



Рисунок 11.7 – Диалог «Факторный анализ: Значения факторов» с запросом создания новых переменных

– нажать кнопку **«Параметры…»**. В секции **«Пропущенные значения»** выбрать **«Исключать наблюдения целиком»**, в секции **«Формат вывода коэффициентов»** поставить флажок напротив строки **«Отсортировать по величине»** (рисунок 11.8) и нажать кнопку **«Продолжить»**;

– нажать кнопку «**ОК**» диалогового окна **«Факторный анализ»**.

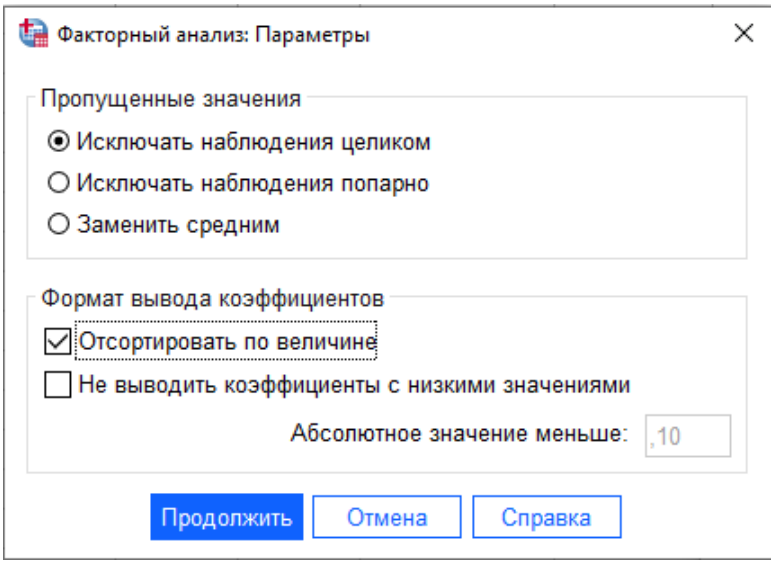



Рисунок 11.8 – Диалог «Факторный анализ: Параметры» с указанием исключать наблюдения с пропущенными значениями переменных и выводить коэффициенты при них в зависимости от величины

На листе с результатами выполненного анализа обратить внимание на таблицы «Корреляционная матрица», «КМО и критерий Бартлетта», «Общности», «Объясненная совокупная дисперсия», «Матрица компонентов» (таблицы 11.4–11.8) и график собственных значений (рисунок 11.9).

Таблица 11.4 – Матрица корреляций, полученная по итогам факторного анализа методом главных компонент без вращения факторов

| | Количество членов домохозяйства | Семейный среднемесячный доход | Количество автомобилей в семье | Площадь дома или квартиры |
|---|---|---|---|---|
| Количество членов домохозяйства | 1,00 | 0,89 | 0,82 | 0,76 |
| Семейный среднемесячный доход | 0,89 | 1,00 | 0,91 | 0,86 |
| Количество автомобилей в семье | 0,82 | 0,91 | 1,00 | 0,74 |
| Площадь жилья | 0,76 | 0,86 | 0,74 | 1,00 |

Таблица 11.5 – Значения критерия адекватности выборки Кайзера – Мейера – Олкина (КМО) и критерия Бартлетта, полученные по итогам факторного анализа методом главных компонент без вращения факторов

| Критерии | | Значения |
|---|---|---|
| Мера адекватности выборки Кайзера – Мейера – Олкина (КМО) | | 0,79 |
| Критерий сферичности Бартлетта | примерное значение $\chi^2$ | 6795,04 |
| | количество степеней свободы | 6 |
| | значимость | 0,00 |

Таблица 11.6 – Значения общностей факторов, рассчитанные по итогам факторного анализа методом главных компонент без вращения факторов

| Переменные | Начальные | Извлеченные |
|---|---|---|
| Количество членов домохозяйства | 1,00 | 0,86 |
| Семейный среднемесячный доход | 1,00 | 0,96 |
| Количество автомобилей в семье | 1,00 | 0,87 |
| Площадь жилья | 1,00 | 0,81 |

Таблица 11.7 – Значения полной объясненной дисперсии, рассчитанные по итогам факторного анализа методом главных компонент без вращения факторов

| Компонента | Начальные собственные значения | | | Суммы квадратов нагрузок извлечения | | |
|---|---|---|---|---|---|---|
| | итого | процент дисперсии | кумулятивный процент | итого | процент дисперсии | кумулятивный процент |
| 1 | 3,49 | 87,36 | 87,36 | 3,49 | 87,36 | 87,36 |
| 2 | 0,27 | 6,77 | 94,13 | | | |

| Компонента | Начальные собственные значения | | | Суммы квадратов нагрузок извлечения | | |
|---|---|---|---|---|---|---|
| | итого | процент дисперсии | кумулятивный процент | итого | процент дисперсии | кумулятивный процент |
| 3 | 0,18 | 4,60 | 98,73 | | | |
| 4 | 0,05 | 1,27 | 100,00 | | | |

Таблица 11.8 – Матрица с факторными нагрузками, рассчитанная по итогам факторного анализа методом главных компонент без вращения факторов

| Переменные | Компонента |
|---|---|
| Семейный среднемесячный доход | 0,98 |
| Количество автомобилей в семье | 0,93 |
| Количество членов домохозяйства | 0,93 |
| Площадь жилья | 0,90 |

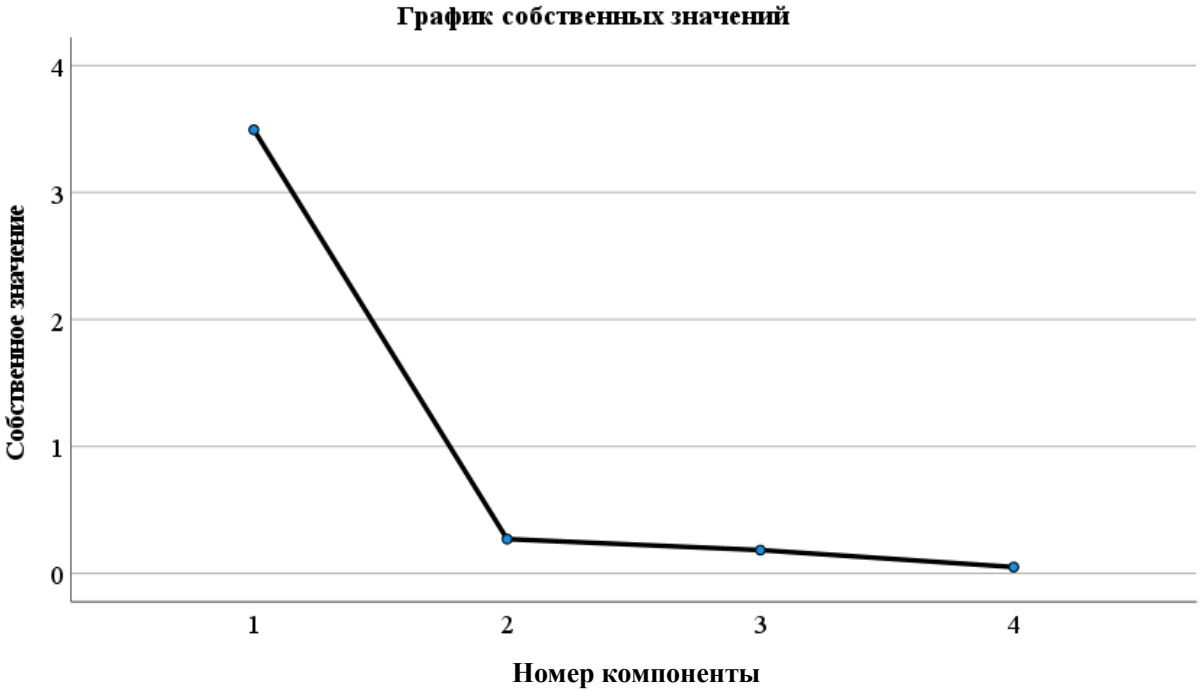



Рисунок 11.9 – График собственных значений («Каменистая осыпь»), выполненный по итогам факторного анализа методом главных компонент без вращения факторов

Как видно из таблицы 11.5, значение КМО превышает 0,5, что подтверждает целесообразность факторного анализа. Значение же критерия Бартлетта равно 6795,04 и статистически значимо, значит, нулевую гипотезу о том, что переменные в генеральной совокупности не коррелируют между собой, следует отвергнуть.

4 Таким образом, проведенный факторный анализ позволяет выделить только одну переменную, которую можно условно назвать, например, «Размер домохозяйства», включающую в себя все четыре использованные переменные.

5 Рассчитать матрицы компонент с факторными нагрузками, применив доступные в программе методы вращения факторов «Варимакс», «Прямой облимин», «Квартимакс», «Эквимакс» и «Промакс».

Для этого, например, для метода вращения «Варимакс»:

– выбрать процедуру факторного анализа, в диалоговом окне которого нажать кнопку **«Вращение…»**;

– в появившемся диалоге все опции оставить неизменными, но в секции **«Метод»** выбрать **«Варимакс»**, а в секции **«Вывести»** поставить флажки напротив строк **«Повернутое решение»** и **«График(и) нагрузок»** (рисунок 11.10), после чего нажать кнопку «**Продолжить»**;

– нажать кнопку **«ОК»** диалогового окна **«Факторный анализ»**.

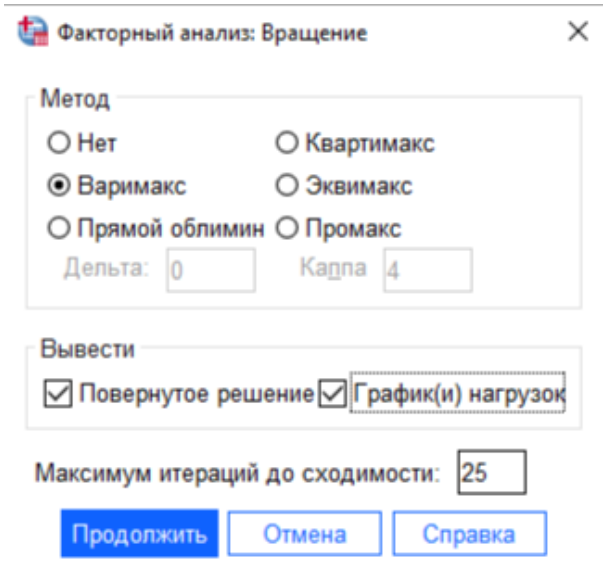



Рисунок 11.10 – Меню диалога «Факторный анализ: Вращение» с выбранным методом вращения факторов

Результаты применения методов окажутся абсолютно такими же, какие были получены до этого без применения вращения факторов.

## 11.3 Задание для самостоятельного выполнения

В созданном файле «11 Факторный анализ.sav» выполнить факторный анализ методом «Невзвешенный МНК» без вращения факторов и с их вращением. Полученные результаты сравнить с теми, которые были получены во время выполнения лабораторной работы.

## 11.4 Вопросы для самоконтроля

1 Что такое факторный анализ как метод статистического анализа данных и в каком порядке он проводится?

2 Какие условия должны быть соблюдены при формулировании проблемы факторного анализа?

3 Что представляет собой фактор, выделенный в результате факторизации матрицы интеркорреляций?

4 Какие существуют процедуры для определения числа факторов в ходе факторного анализа?

5 Какие виды вращений общих факторов применяются в факторном анализе для получения простой структуры факторов, которую легче интерпретировать?

## ЗАКЛЮЧЕНИЕ

По итогам выполненных лабораторных работ можно сделать следующие выводы:

1 ЧУП «Кэтнес», являющееся одной из СБЕ «ОАО «Крессида» и работающее на рынке мебели Витебской области, занимает позицию, которая в соответствии с концепцией матрицы BCG характеризуется как «Трудный ребенок», а в соответствии с концепцией матрицы GE\McKinsey – «Знак вопроса», и является перспективной с точки зрения вложения инвестиций в развитие ее бизнеса.

2 Зависимость между значениями выручки от реализации продукции и ее рентабельностью для ЧУП «Кэтнес» является положительной. Экономическая рациональность ассортимента продукции компании на основании значений коэффициентов ранговой корреляции Спирмена и Кендалла, равных соответственно 0,62 и 0,43, охарактеризована как средняя. С целью повышения ее уровня компании необходимо постепенно свернуть производство и продажу стульев и кресел-качалок как пользующихся слабым спросом и отнесенных по итогам совмещенного ABC- и XYZ-анализа к категории «CZ», а освободившиеся производственные мощности начать использовать для производства журнальных столиков.

3 Плановая длительность описательного маркетингового исследования ЧУП «Кэтнес», предполагающего использование наряду с традиционными и инструментов искусственного интеллекта с целью прогноза рыночного спроса и установления основных характеристик журнальных столиков, оцениваемых домашними хозяйствами при принятии решения об их покупке, составила 47 рабочих дней, а затраты на его проведение были запланированы в размере 28011,20 р.

4 Последующее причинно-следственное исследование, проведенное в виде пробного маркетинга на стандартном рынке компании, показало, что представленные образцы журнальных столиков потенциальными покупателями были оценены достаточно высоко и подтвердили их конкурентоспособность на рынке домашней мебели Витебской области.

5 Дисперсионный анализ материалов, подготовленных для продвижения на рынке журнальных столиков, показал, что наибольшую степень готовности их приобрести у покупателей вызывает комбинация из второго варианта POS-материалов и третьего варианта телерекламы.

6 С использованием данных по выборке, состоящей из 1440 домашних хозяйств Витебской области, были рассчитаны показатели описательной статистики по пятнадцати первоначально установленным их характеристикам. Для проведения последующих парного (однофакторного) и множественного (многофакторного) корреляционно-регрессионных анализов в качестве зависимой переменной была выбрана «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели».

7 Выполненные корреляционно-регрессионные анализы показали, что среди шести независимых переменных, имеющих коэффициент корреляции с

зависимой переменной, превышающий по абсолютному значению 0,7, статистически значимыми являются только пять: «Количество членов домохозяйства», «Семейный среднемесячный доход», «Количество автомобилей в семье», «Вид жилья» и «Площадь жилья»». Именно они, а также зависимая переменная «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» были выбраны для кластеризации участников сформированной выборки.

8 Множественный корреляционно-регрессионный и кластерный анализы участников выборки с использованием указанных шести переменных подтвердили гипотезу, выдвинутую в результате парного корреляционно-регрессионного анализа с использованием переменной «Семейный среднемесячный доход» о том, что на рынке домашней мебели Витебской области можно выделить два четко различающихся сегмента. Большую часть рынка (примерно 87 %) представляет сегмент, состоящий из семей, которые можно охарактеризовать как небольшие и проживающие в квартирах. Остальные 13 % семей, отнесенные ко второму сегменту рынка, можно охарактеризовать как относительно большие и проживающие в собственных домах.

9 Дискриминантный анализ в отношении выделенных сегментов показал статистически значимое различие между ними по четырем из шести вышеуказанных переменных (две переменные «Вид жилья» и «Примерные среднегодовые расходы на покупку (обновление) элементов домашней мебели» оказались не соответствующими критерию допуска) и позволил рассчитать каноническую дискриминантную функцию, четко их разделяющую.

По итогам дискриминантного анализа во втором сегменте можно предположить наличие двух субсегментов, составляющих 3 и 10 % от размера изучаемого рынка, которые возможно статистически значимо различаются по четырем соответствующим критериям допуска переменных.

10 Выполненный в завершение факторный анализ подтвердил оправданность объединения соответствующих согласно дискриминантному анализу критериев допуска четырех переменных и создания на их основе одной интегральной, характеризующей размер домохозяйства. Выделенную интегральную переменную в последующем целесообразно использовать как критерий для начальной (предварительной) сегментации рынка покупателей домашней мебели Витебской области.

11 Сотрудникам маркетинговых подразделений ОАО «Крессида» и входящего в его структуру ЧУП «Кэтнес» рекомендуется:

– с использованием полученной по итогам факторного анализа интегральной переменной выполнить рассмотренные виды статистического анализа и оценить ее пригодность для описания и сегментации рынка покупателей домашней мебели Витебской области;

– изменить (дополнить, сформировать новый) (в том числе и на основе итогов персонализации покупателей, полученных с использованием инструментов искусственного интеллекта в ходе кампании по продвижению продукции) набор

переменных и выполнить рассмотренные виды статистического анализа с целью определения тех из них, с применением которых окажется возможным получить уравнение регрессии, пригодное (достаточно пригодное) для прогнозирования расходов домашних хозяйств на покупку (обновление) элементов домашней мебели на рынке Витебской области. С использованием этого набора выполнить кластерный, дискриминантный и факторный анализы и оценить полученные результаты с точки зрения их соответствия рыночной ситуации;

– опыт, полученный в результате определения характера конкурентной позиции ЧУП «Кэтнес», оценки уровня конкурентоспособности продукции и экономической рациональности ее ассортимента, разработки планов поискового (разведочного), описательного (дескриптивного) и причинно-следственного (каузального) маркетинговых исследований, а также применения методов многомерного статистического анализа данных, полученных в ходе их проведения, тщательно изучить и предложить для использования сотрудникам маркетинговых подразделений других СБЕ, входящих в состав ОАО «Крессида».

*ПРИЛОЖЕНИЕ А*
*(обязательное)*

**Образец оформления титульного листа отчета по лабораторной работе**

Учреждение образования «Белорусский государственный университет информатики и радиоэлектроники»

Инженерно-экономический факультет

Кафедра экономики

Дисциплина: Маркетинговые исследования

ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ

по теме

«Парный (однофакторный) корреляционно-регрессионный анализ данных, полученных в ходе маркетингового исследования»

| | | |
|---|---|---|
| Студент 3-го курса группы 474001 специальности «Цифровой маркетинг» | _____ *(подпись и дата)* | Андронов Николай Иванович |
| Преподаватель: | _____ *(подпись и дата)* | Файзрахманов Фаниль Мударисович, д. ф. э. |

Минск, 2026

## СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ

1 Винстон, У. Бизнес-моделирование и анализ данных. Решение актуальных задач с помощью Microsoft Excel / У. Винстон. – СПб. : Питер, 2019. – 864 с.

2 Галицкий, Е. Б. Маркетинговые исследования. Теория и практика : учеб. для вузов / Е. Б. Галицкий, Е. Г. Галицкая. – М. : Юрайт, 2019. – 570 с.

3 Доугерти, К. Введение в эконометрику / К. Доугерти. – М. : ИНФРА-М, 1999. – 402 с.

4 Кинг, К. Искусственный интеллект в маркетинге. Как использовать ИИ и быть на шаг впереди / К. Кинг. – М. : АСТ, 2024. – 256 с.

5 Козлов, А. Ю. Статистический анализ данных в MS Excel : учеб. пособие / А. Ю. Козлов, В. С. Мхитарян, В. Ф. Шишков. – М. : ИНФРА-М, 2019. – 320 с.

6 Котлер, Ф. Маркетинг менеджмент / Ф. Котлер, Л. Келлер. – СПб. : Питер, 2020. – 848 с.

7 Малхотра, Н. К. Маркетинговые исследования. Практическое руководство / Н. К. Малхотра. – М. : Вильямс, 2016. – 1184 с.

8 Многомерный статистический анализ в экономике : учеб. пособие для вузов / под ред. проф. В. Н. Тамашевича. – М. : ЮНИТИ-ДАНА, 1999. – 598 с.

9 Могайар, У. Блокчейн для бизнеса / У. Могайар. – М. : Бомбора, 2018. – 224 с.

10 Моосмюллер, Г. Маркетинговые исследования с SPSS : учеб. пособие / Г. Моосмюллер, Н. Н. Ребик. – М. : ИНФРА-М, 2024. – 200 с.

11 Сафронова, Н. Б. Маркетинговые исследования : учеб. пособие / Н. Б. Сафронова, И. Е. Корнеева. – 2-е изд., стер. – М. : Дашков и К, 2019. – 294 с.

12 Сигел, Э. Ф. Практическая бизнес-статистика / Э. Ф. Сигел. – М. : Вильямс, 2008. – 1051 с.

13 Сошникова, Л. А. Многомерный статистический анализ. Практикум : учеб. пособие / Л. А. Сошникова, Е. Е. Шарилова. – Минск : БГЭУ, 2024. – 230 с.

14 Токарев, Б. Е. Маркетинговые исследования : учеб. / Б. Е. Токарев. – 2-е изд., перераб. и доп. – М. : Магистр : ИНФРА-М, 2019. – 512 с.

15 Факторный анализ как метод исследования бренда : монография / К. А. Аржанова, Г. В. Довжик, П. О. Щукина [и др.]; отв. ред. С. Г. Бычкова. – М. : РУСАЙНС, 2025. – 126 с.

16 Хацкевич, Г. А. Эконометрика : учеб. / Г. А. Хацкевич, Т. В. Русилко. – Минск : РИВШ, 2021. – 452 с.

17 Яковлев, В. Б. Статистика. Расчеты в Microsoft Excel / В. Б. Яковлев. – 2-е изд., испр. и доп. – М. : Юрайт, 2022. – 353 с.

Св. план 2025, поз. 44

*Учебное издание*

**Файзрахманов** Фаниль Мударисович

# МАРКЕТИНГОВЫЕ ИССЛЕДОВАНИЯ. ЛАБОРАТОРНЫЙ ПРАКТИКУМ

ПОСОБИЕ

В двух частях

Часть 2

Редактор *Л. И. Артёмова*
Корректор *Е. Н. Батурчик*
Компьютерная правка, оригинал-макет *В. А. Долгая*

Подписано в печать 24.02.2026. Формат 60×84 1/16. Бумага офсетная. Гарнитура «Таймс».
Отпечатано на ризографе. Усл. печ. л. 11,97. Уч.-изд. л. 12,5. Тираж 80 экз. Заказ 143.

Издатель и полиграфическое исполнение: учреждение образования
«Белорусский государственный университет информатики и радиоэлектроники».
Свидетельство о государственной регистрации издателя, изготовителя,
распространителя печатных изданий №1/238 от 24.03.2014,
№2/113 от 07.04.2014, №3/615 от 07.04.2014.
Ул. П. Бровки, 6, 220013, г. Минск