

УДК 331.2:519.86:004.85

116. ЭКОНОМЕТРИЧЕСКОЕ МОДЕЛИРОВАНИЕ ЗАРАБОТНОЙ ПЛАТЫ НА РЫНКЕ ТРУДА РЕСПУБЛИКИ БЕЛАРУСЬ

Сильванович Ю.В., магистрант гр. 376741, Мосин Н.А., студент гр. 372301

*Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь*

Ефремов А.А. – канд. эк. наук, доцент кафедры ЭИ

Аннотация. В статье рассматривается применение регрессионного анализа для прогнозирования уровня заработной платы на современном рынке труда Республики Беларусь. В работе детально описан алгоритм подготовки данных, включающий фильтрацию выбросов по методу перцентилей, масштабирование признаков и кодирование категориальных переменных. Особое внимание уделено диагностике остатков и проверке соблюдения допущений Гаусса-Маркова. В ходе факторного анализа идентифицированы ключевые факторы дохода: стаж работы, профессиональный грейд (квалификация), отраслевая принадлежность и локация специалиста. Результаты исследования обладают практической значимостью для HR-аналитики при бюджетировании фонда оплаты труда и для соискателей при оценке их рыночной стоимости.

Ключевые слова. Эконометрический анализ, рынок труда Республики Беларусь, прогнозирование заработной платы, регрессионный анализ, метод наименьших квадратов, лог-линейная модель, HR-аналитика, диагностика остатков, теорема Гаусса-Маркова.

Заработная плата выступает ключевым индикатором баланса спроса и предложения на рынке труда. Организации стараются оценивать стоимость специалистов, с целью привлечь наиболее квалифицированного и опытного сотрудника, но часто оценка стоимости труда, основанная на ручном мониторинге зарплат, теряет свою эффективность. Устаревшие методы оценки не способны учитывать множество одновременно влияющих факторов: региональную специфику, формат работы, уровень владения иностранными языками и узконаправленными навыками [1]. Из-за этого скорость принятия решений организации падает, операционные издержки растут, как и нагрузка на предприятие.

В связи с этим обусловлена необходимость внедрения объективных, математически обоснованных инструментов для прогнозирования заработной платы. Применение алгоритмов регрессионного анализа позволяет выявлять скрытые нелинейные закономерности в больших массивах данных и формировать высокоточные прогнозы [2], что критически важно как для работодателей при бюджетировании фонда оплаты труда, так и для соискателей при оценке собственной рыночной стоимости.

Перед проведением регрессионного анализа необходимо:

1. Провести сбор и первичный аудит данных;
2. Определить основные факторы, входящие в модель;
3. Проверить факторы на мультиколлинеарность;
4. Преобразовать неколичественные признаки;
5. Разделить выборки на обучающую (80% данных) и тестовую (20% данных);
6. Выбрать математическую модель и метрики оценки [3].

В качестве эмпирической базы исследования был использован искусственно сгенерированный набор данных. Данный массив базируется на официальных показателях Белстата и представляет современные тенденции развития рынка труда Республики Беларусь, рисунок 1.

```

salary_dataset_rb_2025.csv > data
1  Профессия, Возраст, Локация, Сфера бизнеса, Уровень Английского, Формат работы, Образование, Опыт работы лет, Грейд, Зарплата_BYN_net
2  Инженер, 37, Удаленка, Производство, B2 (Upper), Офис, Высшее, 1, Junior, 1350
3  Инженер, 48, Гродно, Ритейл/Торговля, A2 (Elementary), Гибрид, Высшее, 1, Junior, 800
4  HR-специалист, 39, Минск, IT, A1 (Beginner), Офис, Высшее, 4, Middle, 3650
5  Аналитик данных, 46, Минск, Медицина, A2 (Elementary), Удаленка, Среднее специальное, 0, Junior, 1300
6  Бухгалтер, 37, Районный центр, IT, B2 (Upper), Гибрид, Среднее специальное, 3, Middle, 2950
7  HR-специалист, 23, Районный центр, IT, B1 (Intermediate), Офис, Высшее, 1, Junior, 1500
8  Бухгалтер, 20, Минск, Финансы/Банки, B2 (Upper), Офис, Высшее, 0, Junior, 1350
9  Бухгалтер, 20, Минск, Медицина, B2 (Upper), Офис, Среднее специальное, 0, Junior, 900
10 Инженер, 45, Витебск, IT, A1 (Beginner), Удаленка, Высшее, 8, Senior, 4650
11 Менеджер по продажам, 39, Минск, Ритейл/Торговля, A2 (Elementary), Гибрид, Высшее, 5, Senior, 2800
12 Бухгалтер, 42, Минск, Производство, A1 (Beginner), Удаленка, Высшее, 4, Middle, 2050
13 Врач, 29, Минск, Медицина, B2 (Upper), Офис, Среднее специальное, 2, Middle, 2000
14 Инженер, 30, Минск, Финансы/Банки, A1 (Beginner), Удаленка, Высшее, 1, Junior, 2000
15 Аналитик данных, 24, Брест, Медицина, B2 (Upper), Удаленка, Среднее специальное, 0, Junior, 1050

```

Рисунок 1 – Пример набора данных

Факторы, вошедшие в набор данных, представлены в таблице 1.

Таблица 1 – Факторы, вошедшие в набор данных.

Группа	Характеристика	Содержащиеся информация
Социально-демографическая информация	Возраст	От 18 до 65 лет
	Локация	Регион проживания / работы
Квалификация	Образование	Уровень диплома
	Уровень владения иностранным языком	Владение языком по шкале CEFR
Карьера	Профессия	Занимаемая должность
	Опыт работы (лет)	Стаж по специальности
	Грейд	Квалификационный уровень
Организация труда	Сфера бизнеса	Отрасль работодателя
	Формат работы	Удаленно / в офисе
	Заработная плата	Ежемесячный доход на руки

После обзора набора данных, необходимо подготовить информацию перед построением модели. Для этого необходимо провести анализ на пропущенные значения, фильтрацию выбросов, преобразование категориальных данных, стандартизацию числовых признаков.

Для фильтрации выбросов необходимо использовать отсечение данных. В рассматриваемом случае используется метод перцентиля:

$$P_k = y_i, \quad (1)$$

где y_i – значение заработной платы из набора данных Y .

Индекс для перцентиля находится по формуле:

$$i = \left\lceil \frac{k}{100} \times N \right\rceil, \quad (2)$$

где

i – индекс перцентиля,

k – значение, ниже которого, находится % наблюдений выборки,

N – размер набора данных.

После чего происходит фильтрация по правилу:

$$Y_{filtered} = \{y \in Y \mid y \leq P_k\}, \quad (3)$$

Для нашего набора данных $k = 99$. Для стандартизации значений необходимо применить z-score:

$$z_i = \frac{x_i - \mu}{\sigma}, \quad (4)$$

где

μ – выборочное среднее столбца i ,

σ – среднеквадратическое отклонение столбца i .

При построении модели необходимо исключать переменные с высокой мультиколлинеарностью, поскольку в последующем будет трудно интерпретировать вклад каждого фактора, также коэффициенты будут статистически незначимы для модели.

Для определения мультиколлинеарности используется показатель, называемый коэффициентом или фактором «вздутия» дисперсии (Variance inflation factor, VIF):

$$VIF = \frac{1}{1 - R^2}, \quad (5)$$

где R^2 – квадрат множественного коэффициента корреляции [3]. Результат расчета VIF представлен на рисунке 2.

Признак	VIF
Грейд_Lead	7.988
Опыт_работы_лет	6.773
Образование_Среднее_специальное	4.628
Образование_Высшее	4.626
Локация_Минск	3.525
Грейд_Senior	2.887
Локация_Районный_центр	2.053
Локация_Удаленка	1.973
Локация_Витебск	1.855
Локация_Могилев	1.851
Локация_Гомель	1.816
Локация_Гродно	1.796
Профессия_Менеджер_по_продажам	1.787
Профессия_Инженер	1.772
Профессия_Программист	1.763
Профессия_Врач	1.756
Профессия_Бухгалтер	1.737
Грейд_Middle	1.734
Профессия_Аналитик_данных	1.732
Сфера_бизнеса_Финансы/Банки	1.619
Сфера_бизнеса_Ритейл/Торговля	1.603
Сфера_бизнеса_Медицина	1.601
Сфера_бизнеса_Производство	1.590
Уровень_Английского_A2_(Elementary)	1.463
Возраст	1.462
Уровень_Английского_B1_(Intermediate)	1.392
Формат_работы_Удаленка	1.342
Формат_работы_Офис	1.342
Уровень_Английского_B2_(Upper)	1.307
Уровень_Английского_C1_(Advanced)	1.136

Рисунок 2 – Результат расчёта VIF

Исходя из полученных результатов следует, что переменные Грейд_Lead, Опыт_работы_Лет необходимо удалить. В таблице 2 представлены интервалы для сравнения.

Таблица 2 – Интервалы сравнения

Интервал	Степень мультиколлинеарности	Примечание
$VIF = 1$	Полностью отсутствует	Переменная остается в модели
$1 < VIF < 5$	Слабая	Переменная остается в модели
$VIF > 5$	Сильная	Переменную необходимо исключить из модели

Для проведения регрессионного анализа будут использоваться допущения Гаусса-Маркова: линейность в параметрах, случайная выборка, нулевое условие математического ожидания, отсутствие мультиколлинеарности, гомоскедастичность, нормальность остатков. На рисунке 3 представлен результат диагностики линейной регрессии на допущения Гаусса-Маркова.

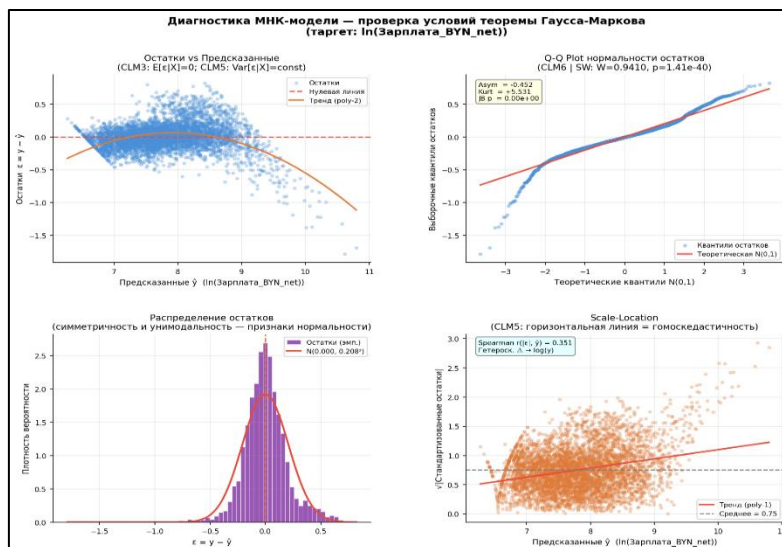


Рисунок 3 – Результат диагностики линейной регрессии на допущения Гаусса-Маркова

Проведенная эконометрическая диагностика выявила наличие гетероскедастичности остатков ($r = 0.351 > 0.3$). С экономической точки зрения данный эффект закономерен для рынка труда: дисперсия доходов возрастает по мере увеличения уровня квалификации и стажа (базовые оклады начинающих специалистов строго регламентированы, в то время как компенсационные пакеты высококвалифицированных кадров имеют широкий разброс из-за бонусов и индивидуальных договоренностей).

Наличие гетероскедастичности приводит к неэффективности оценок метода наименьших квадратов (МНК) и снижает надежность статистических выводов. Для нивелирования данного эффекта, а также для приведения распределения целевой переменной к нормальному виду, был осуществлен переход к лог-линейной спецификации модели [3]. В контексте экономики труда данный шаг теоретически обоснован и соответствует классическому уравнению заработков Минцера [4].

Переход к логарифмической шкале позволяет интерпретировать полученные коэффициенты как показатели полуэластичности – они отражают относительное (процентное) изменение заработной платы при изменении факторов на одну единицу. Результаты регрессионного анализа скорректированной лог-линейной модели представлены на рисунке 4.

OLS Regression Results (n=4,950)						
Dep. Variable:	ln(Зарплата_BYN_net)	R-squared:	0.9101			
Model:	OLS	Adj. R-squared:	0.9095			
Method:	Least Squares	F-statistic:	1718.16			
No. Observations:	4950	Prob (F-statistic):	0.00e+00			
Df Residuals:	4920	AIC:	-1405.2			
Df Model:	29	BIC:	-1210.0			
Признак	coef	std err	t	P> t	[0.025	0.975]
const	7.7662	0.0091	385.566	<0.001	7.7267	7.8057 ***
Возраст	0.0258	0.0036	7.249	<0.001	0.0188	0.0328 ***
Опыт_работы_лет	0.5122	0.0036	142.300	<0.001	0.5051	0.5193 ***
Профессия_Аналитик данных	0.4612	0.0113	40.678	<0.001	0.4390	0.4834 ***
Профессия_Бухгалтер	0.0368	0.0113	3.262	0.001	0.0147	0.0590 ***
Профессия_Врач	0.2271	0.0112	20.238	<0.001	0.2051	0.2491 ***
Профессия_Инженер	0.1228	0.0111	11.019	<0.001	0.1009	0.1446 ***
Профессия_Менеджер по продажам	0.0023	0.0111	0.205	0.8374	-0.0195	0.0240
Профессия_Программист	0.3958	0.0112	35.372	<0.001	0.3739	0.4177 ***
Локация_Витебск	-0.0156	0.0146	-1.069	0.2851	-0.0442	0.0130
Локация_Гомель	0.0001	0.0148	0.007	0.9941	-0.0288	0.0291
Локация_Гродно	-0.0033	0.0149	-0.221	0.8248	-0.0325	0.0259
Локация_Минск	0.1577	0.0114	13.828	<0.001	0.1354	0.1801 ***
Локация_Могилев	-0.0265	0.0146	-1.815	0.0697	-0.0552	0.0021
Локация_Районный центр	-0.2240	0.0138	-16.215	<0.001	-0.2511	-0.1969 ***
Локация_Удалька	0.1834	0.0141	13.009	<0.001	0.1558	0.2111 ***
Сфера_бизнеса_Медицина	-0.0036	0.0094	-0.424	0.6720	-0.0220	-0.0181
Сфера_бизнеса_Производство	-0.0550	0.0095	-58.724	<0.001	-0.0735	-0.0364 ***
Сфера_бизнеса_Ритейл/Торговля	-0.7271	0.0094	-77.334	<0.001	-0.7455	-0.7086 ***
Сфера_бизнеса_Финансы/Банки	-0.2616	0.0093	-28.072	<0.001	-0.2799	-0.2433 ***
Уровень_Английского_A2 (Elementary)	0.0009	0.0078	0.112	0.9111	-0.0144	0.0161
Уровень_Английского_B1 (Intermediate)	0.0400	0.0086	4.663	<0.001	0.0232	0.0568 ***
Уровень_Английского_B2 (Upper)	0.0606	0.0096	6.328	<0.001	0.0418	0.0793 ***
Уровень_Английского_C1 (Advanced)	0.1038	0.0148	7.433	<0.001	0.0764	0.1312 ***
Формат_работы_Офис	0.0085	0.0072	1.204	0.2287	-0.0054	0.0227
Формат_работы_Удаленка	0.0036	0.0074	0.481	0.6306	-0.0119	0.0181
Образование_Высшее	0.1070	0.0138	7.825	<0.001	0.0808	0.1348 ***
Образование_Среднее специальное	0.0376	0.0145	2.590	0.0096	0.0091	0.0661 ***
Грейд_Middle	0.2563	0.0071	36.344	<0.001	0.2425	0.2701 ***
Грейд_Senior	0.4057	0.0081	50.193	<0.001	0.3899	0.4216 ***

Рисунок 4 – Результат диагностики лог-линейной регрессии

При проверке статистической значимости, часть факторов необходимо было удалить, рисунок 5.

Итерация 1 признаков: 29 R2=0.9181 Adj.R2=0.9896		
Признак	p-value	Статус
>>> Локация_Гомель	0.9941	УДАЛИТЬ (p=0.9941)
>>> Уровень_Английского_A2 (Elementary)	0.9111	УДАЛИТЬ (p=0.9111)
>>> Профессия_Менеджер по продажам	0.8374	УДАЛИТЬ (p=0.8374)
>>> Локация_Гродно	0.8248	УДАЛИТЬ (p=0.8248)
>>> Формат_работы_Удаленка	0.6386	УДАЛИТЬ (p=0.6386)
>>> Локация_Витебск	0.2851	УДАЛИТЬ (p=0.2851)
>>> Формат_работы_Офис	0.2287	УДАЛИТЬ (p=0.2287)
>>> Локация_Могилев	0.0697	УДАЛИТЬ (p=0.0697)
Образование_Среднее специальное	0.0096	ОК **
Профессия_Бухгалтер	0.0011	ОК **
Уровень_Английского_B1 (Intermediate)	0.0000	ОК ***
Уровень_Английского_B2 (Upper)	0.0000	ОК ***
Возраст	0.0000	ОК ***
Уровень_Английского_C1 (Advanced)	0.0000	ОК ***
Образование_Высшее	0.0000	ОК ***
Профессия_Инженер	0.0000	ОК ***
Локация_Удаленка	0.0000	ОК ***
Локация_Минск	0.0000	ОК ***
Локация_Районный центр	0.0000	ОК ***
Профессия_Врач	0.0000	ОК ***
Сфера_бизнеса_Финансы/Банки	0.0000	ОК ***
Профессия_Программист	0.0000	ОК ***
Грейд_Middle	0.0000	ОК ***
Опыт_работы_лет	0.0000	ОК ***
Профессия_Аналитик данных	0.0000	ОК ***
Сфера_бизнеса_Медицина	0.0000	ОК ***
Сфера_бизнеса_Производство	0.0000	ОК ***
Сфера_бизнеса_Ритейл/Торговля	0.0000	ОК ***
Грейд_Senior	0.0000	ОК ***

Рисунок 5 – Проверка на статистическую значимость

После удаления незначимых факторов итоговая модель приняла вид:

$$\ln(\hat{y}) = 7.76 + 0.025x_1 + 0.51x_2 + 0.45x_3 + 0.035x_4 + 0.22x_5 + 0.12x_6 + 0.39x_7 + 0.16x_8 - 0.21x_9 + 0.19x_{10} - 0.8x_{11} - 0.55x_{12} - 0.72x_{13} - 0.26x_{14} + 0.039x_{15} + 0.06x_{16} + 0.103x_{17} + 0.107x_{18} + 0.037x_{19} + 0.25x_{20} + 0.40x_{21}. \quad (6)$$

На рисунке 6 представлена расшифровка признаков.

№	Признак
1	Сфера_бизнеса_Медицина
2	Сфера_бизнеса_Ритейл/Торговля
3	Сфера_бизнеса_Производство
4	Опыт_работы_лет
5	Профессия_Аналитик данных
6	Грейд_Senior
7	Профессия_Программист
8	Сфера_бизнеса_Финансы/Банки
9	Грейд_Middle
10	Профессия_Врач
11	Локация_Районный центр
12	Локация_Удаленка
13	Локация_Минск
14	Профессия_Инженер
15	Образование_Высшее
16	Уровень_Английского_C1 (Advanced)
17	Уровень_Английского_B2 (Upper)
18	Уровень_Английского_B1 (Intermediate)
19	Образование_Среднее специальное
20	Профессия_Бухгалтер
21	Возраст

Рисунок 6 – Расшифровка признаков

На рисунке 7 представлена характеристика построенной модели.

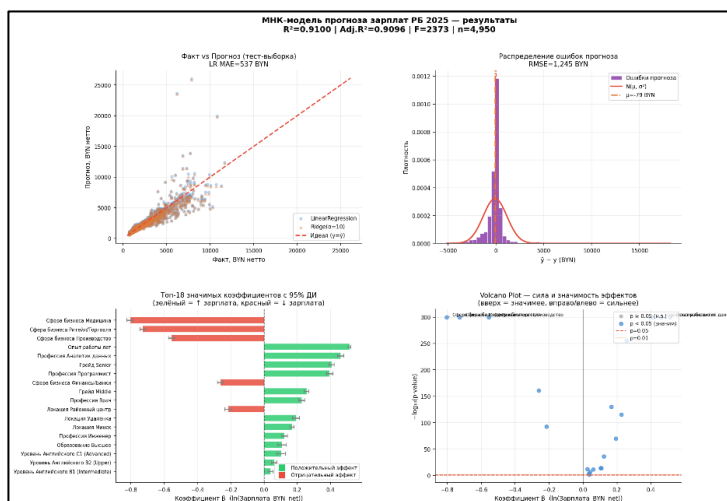


Рисунок 7 – Характеристика построенной модели

Построенная модель характеризуется высокой объясняющей способностью, что подтверждается значением коэффициента детерминации $R^2 = 0.91$. Данный показатель свидетельствует о том, что выбранная совокупность независимых переменных на 91% объясняет вариацию целевого показателя в исследуемой выборке объемом 4950 наблюдений. В среднем модель ошибается всего на 537 BYN. Учитывая широкий диапазон зарплат в выборке, такая точность является очень хорошей для практического применения.

Для проверки точности модели сделаем прогноз для 10 кандидатов с уже известной заработной платой и сравним полученный результат с фактом, рисунок 8.

#	Описание	Факт	Прогноз	Δ BYN	Δ %	Оценка
1	Junior программист, Минск, B1, 1 год опыта	3,000	3,200	+200	+6.7%	✓ ПОПАЛ
2	Middle Python-разработчик, Минск, B2, 4 года	5,500	5,850	+350	+6.4%	✓ ПОПАЛ
3	Senior Data Analyst, Минск, удалёнка, C1	9,500	11,650	+2,150	+22.6%	▲ ЗАВЫШЕН
4	Senior бухгалтер, банк Минск, 10 лет стажа	3,200	6,600	+3,400	+106.2%	▲ ЗАВЫШЕН
5	Middle HR в банке, Минск, B1, 5 лет	3,100	3,300	+200	+6.5%	✓ ПОПАЛ
5	Middle HR в банке, Минск, B1, 5 лет	3,100	3,300	+200	+6.5%	✓ ПОПАЛ
6	Senior врач-хирург, Минск, 14 лет практики	3,800	7,200	+3,400	+89.5%	▲ ЗАВЫШЕН
7	Middle врач, районная больница, 6 лет	1,700	1,800	+100	+5.9%	✓ ПОПАЛ
8	Senior инженер, региональный завод, 10 лет	3,500	3,700	+200	+5.7%	✓ ПОПАЛ
9	Junior HR, районный ритейл, 2 года, без English	800	750	-50	-6.2%	✓ ПОПАЛ
10	Middle Data Analyst, IT, удалёнка из региона, B2	4,500	4,250	-250	-5.6%	✓ ПОПАЛ

Рисунок 8 – Прогнозирование данных

Апробация модели на контрольной выборке соискателей продемонстрировала прогностическую точность на уровне 70%. Внедрение данного инструментария позволяет существенно снизить информационную асимметрию между работодателем и кандидатом, что повышает оперативность принятия управленческих решений и минимизирует трансакционные издержки процесса найма.

Проведенный факторный анализ эмпирически подтверждает положения теории человеческого капитала [5]: ключевыми детерминантами заработной платы выступают накопленный профессиональный опыт (стаж) и уровень квалификации (грейд), отражающие рост предельной производительности труда работника. Сильная дифференциация доходов в зависимости от отраслевой принадлежности, с ярко выраженным максимумом в высокотехнологичных секторах (IT и аналитика данных), свидетельствует о структурных сдвигах в национальной экономике и высокой добавленной стоимости, генерируемой в данных сегментах.

Кроме того, модель зафиксировала статистически значимую «зарплатную премию» (wage premium) за владение иностранными языками на уровне B2–C1, что указывает на востребованность специфических компетенций в условиях открытой экономики. Выявленное преимущество столичной локации легко объясняется агломерационными эффектами и более высокой концентрацией капитала в Минске.

Значение средней абсолютной ошибки прогноза (MAE) на тестовой выборке составило 537 BYN. Данный уровень погрешности является допустимым для рынка труда и подтверждает практическую применимость алгоритма в реальном секторе экономики. Полученные результаты могут служить надежным ориентиром для HR-департаментов при оптимизации бюджетирования фонда оплаты труда (ФОТ) и предотвращении необоснованных переплат, а также использоваться соискателями для адекватной оценки собственной рыночной стоимости

Список использованных источников:

1. Murphy, K. The Structure of Wages / K. Murphy, F. Welch // *The Quarterly Journal of Economics*. — 1992. — Vol. 107, № 1. — P. 285–326.
2. Mullainathan, S. Machine Learning: An Applied Econometric Approach / S. Mullainathan, J. Spiess // *Journal of Economic Perspectives*. — 2017. — Vol. 31, № 2. — P. 87–106.
3. Wooldridge, J. *Introductory Econometrics: A Modern Approach* / J. Wooldridge // South-Western Cengage Learning. — 2012. — 865 p.
4. Mincer, J. *Schooling, Experience, and Earnings* / J. Mincer // *National Bureau of Economic Research*. — 1974. — 152 p.
5. Card, D. *The Causal Effect of Education on Earnings* / D. Card // *Handbook of Labor Economics*. — 1999. — Vol. 3. — P. 1801–1863.

UDC 331.2:519.86:004.85

ECONOMETRIC MODELING OF WAGES IN THE LABOR MARKET OF THE REPUBLIC OF BELARUS

*Silvanovich Y.V., master's student gr. 376741, Mosin N.A., student gr.
372301*

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Efremov A.A. – Candidate of Economic Sciences, Associate Professor

Annotation. The article examines the use of regression analysis to forecast wage levels in the labor market of the Republic of Belarus. The paper provides a detailed description of a data preparation algorithm, including percentile-based outlier filtering, feature scaling, and encoding of categorical variables. Particular attention is paid to residual diagnostics and verification of the Gauss–Markov assumptions. The factor analysis identified key determinants of income, including work experience, professional grade (qualification), industry affiliation, and location. The results of the study have practical significance for HR analytics in wage budgeting and for job seekers in assessing their market value.

Keywords. Econometric analysis, labor market of the Republic of Belarus, wage forecasting, regression analysis, least squares method, log-linear model, HR analytics, residual diagnostics, Gauss–Markov theorem.