Исследование «Исследование зависимости размера заработной платы от различных факторов»

Введение

Целью исследования является идентификация и анализ ключевых факторов, влияющих на уровень заработной платы среди людей разной возрастной категории.

Главной задачей исследования является определение параметров, влияющих на уровень заработной платы при помощи модели множественной линейной регрессии. Для формулировки и обоснования выдвигаемых гипотез проведем обзор научной литературы на схожие тематики. Далее проверим поставленные гипотезы при помощи построенной регрессии.

Актуальность данной темы обусловлена тем, что выявление факторов, влияющих на уровень заработной платы может помочь студентам и другим людям, ориентированным на создание своего карьерного пути, в понимании рынка труда и корректировке своей стратегии выхода на него, что может улучшить экономическую эффективность. Кроме того, мы можем посмотреть на то, действительно ли увеличение часов работы в неделю приводит исключительно к росту заработной платы и есть ли разница в зарплате при работе в отечественной и иностранной компаниях.

Описание данных

В работе используются данные 2022 года по индивидам РМЭЗ НИУ ВШЭ, Российский мониторинг экономического положения и здоровья населения НИУ-ВШЭ (RLMS-HSE)», проводимый Национальным исследовательским университетом "Высшая школа экономики" и ООО «Демоскоп» при участии Центра народонаселения Университета Северной Каролины в Чапел Хилле и Института социологии Федерального научно-исследовательского социологического центра РАН. (Сайты обследования RLMS-HSE: https://rlms-hse.cpc.unc.edu и http://www.hse.ru/rlms).

Перед началом эконометрической части работы, данные были обработаны. Были исключены индивиды, отказавшиеся от ответа, не работающие

индивиды, а также исключены наблюдения с выбросами в заработной плате. Количество наблюдений в полученной выборке составило 3071.

Экономическая модель

Для исследования влияния различных факторов на заработную плату после вычета налогов и отчислений (wage), измеряемую в рублях, был взят ряд объясняющих переменных:

- male – пол респондента; дамми-переменная, (0 – женский, 1 – мужской);
- age – количество полных лет;
- education – изначальная переменная, указывающая на уровень образования респондента, была разбита на несколько дамми-переменных
    - school – респондент имеет только школьное образование и/или какое-либо профессиональное образование без диплома (0 – нет, 1 – да);
    - tehnicum – респондент имеет только школьное образование и какое-либо профессиональное образование (техникум/ПТУ) с дипломом (0 – нет, 1 – да);
    - not_full_uni – респондент имеет, помимо школьного образования, неоконченное высшее образование, обучался в высшем учебном заведении от 1 и более лет (0 – нет, 1 – да);
    - bach – есть диплом о высшем образовании (0 – нет, 1 – да);
    - postgrad – аспирантура и т. п. с дипломом (0 – нет, 1 – да);
- hours_week – среднее количество рабочих часов в неделю (в часах);
- foreing_own – являются ли владельцами или совладельцами предприятия, организации, где работает респондент, иностранные фирмы или иностранные частные лица; дамми-переменная (0 – нет, 1 – да);
- danger – является ли производство, на котором работает респондент, вредным или опасным, т.е. дающим Вам право на досрочное назначение трудовой пенсии, на дополнительные выплаты или льготы; дамми-переменная (0 – нет, 1 – да).

При этом так как одна из дамми-переменных, отвечающих за уровень образования должна быть исключена (иначе Dummy trap), то в дальнейшую модель не будет включаться переменная school.

Именно эти переменные кажутся оказывающими наибольшее влияние на размер заработной платы. При выборе переменных учитывались факторы, которые использовали авторы научных исследований, а также другие переменные, которые гипотетически могут оказывать влияние на зарплату.

Анализируя влияние рассматриваемых объясняющих переменных на размер заработной платы, зарплаты у мужчин скорее всего будут больше, чем у женщин, что доказывается во многих исследованиях, например, Analysis of factors affecting earnings using Annual Survey of Hours and Earnings, 2016.

Зарплата положительно зависит от уровня образования, чем выше образование, тем больше размер зарплаты, что подчиняется логике, а также доказывается исследованиями, например, Worker Diversity and Wage Growth Since 1940, 2020.

Размер зарплаты зависит от возраста не только линейно, но и параболически (будет введена переменная age^2), это объясняется тем, что пик размера получаемой заработной платы приходится на средний возраст индивидов, после чего начинается снижение зарплаты, такая зависимость наблюдается и на данных (Analysis of factors affecting earnings using Annual Survey of Hours and Earnings, 2016).

Вероятнее всего увеличение количества рабочих часов положительно сказывается на размере зарплаты.

Можно предположить, что работа на вредном, опасном производстве, так же как и работа в организации, принадлежащей иностранным владельцам, увеличивает размер заработной платы работников таких предприятий.

Исходя из всего вышесказанного, рассмотрим следующие гипотезы:

- Действительно ли работа в иностранной компании увеличивает уровень зарплаты, поэтому коэффициент при переменной наличия в компании (месте работы респондента) иностранных владельцев (foreign_own) должен быть положительным.

- Верна ли гипотеза о том, что до определенного момента зарплата растет при увеличении часов работы, а затем начинает сокращаться. Для этого рассмотрим переменную hours*week^2 (кол-во рабочих часов в неделю в квадрате). Чтобы гипотеза подтвердилась, необходимо, чтобы парабола была ветвями вниз, соответственно, коэффициент перед новой переменной должен быть отрицательным.

Предварительный анализ данных

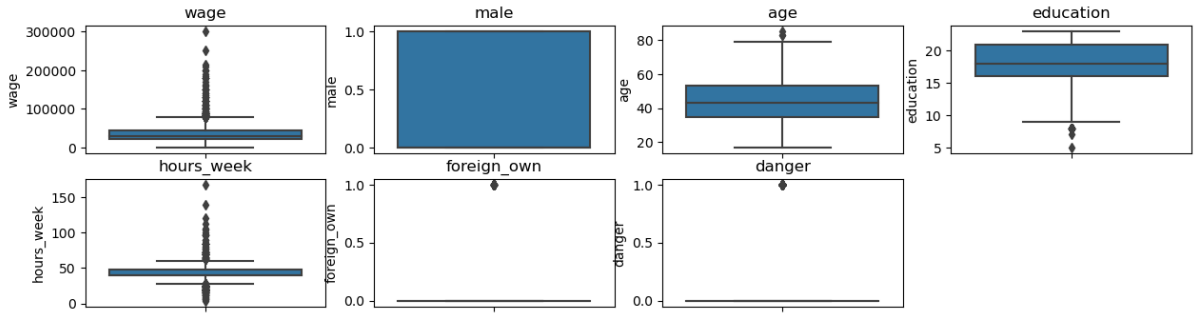
В рамках предварительного анализа данных, были выполнены следующие процедуры:
Первоначально, в наборе данных присутствовала переменная 'work' (обозначающему текущую занятость респондента), но после фильтрации записей по значению '1'(то есть респондент сейчас работает - то, что релевантно для нас) , данная переменная была исключена из дальнейшего рассмотрения, поскольку не содержала информации, необходимой для целей нашего исследования.

Далее был проведен анализ типов данных в датасете, в ходе которого было установлено, что переменные 'hours_week', 'foreign_own' и 'danger' имеют тип object, что предполагает возможность присутствия разнородных данных, включая строковые значения.

Column	Non-Null Count	Dtype
wage	4915	float64
male	4995	int64
age	4995	int64
education	4995	float64
hours_week	4995	object
foreign_own	4488	object
danger	4488	object

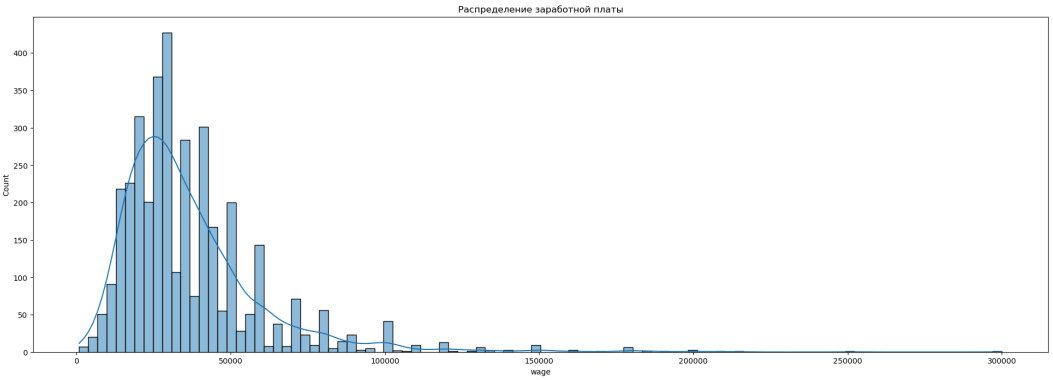
Специфические значения, такие как '99999997', '99999998', '99999999', обозначающие различные типы неразглашения информации респондентом (трудности с ответом, отказ от ответа и отсутствие ответа), потребовало их исключения из анализа путем замены на значения NaN и последующего удаления, чтобы обеспечить точность моделирования. В целях стандартизации данных переменные 'male', 'danger' и 'foreign_own' были преобразованы таким образом, чтобы их значения соответствовали формату 0 и 1, вместо 1 и 2.

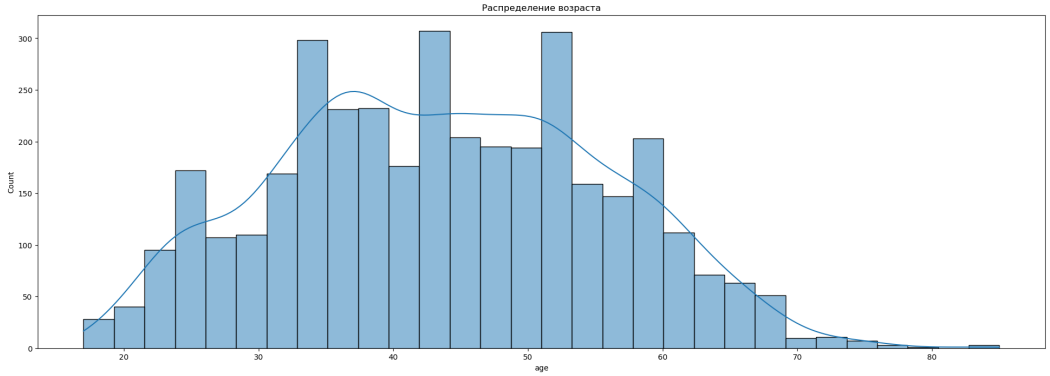
Для визуализации распределения данных и идентификации потенциальных выбросов были построены ящичковые диаграммы.



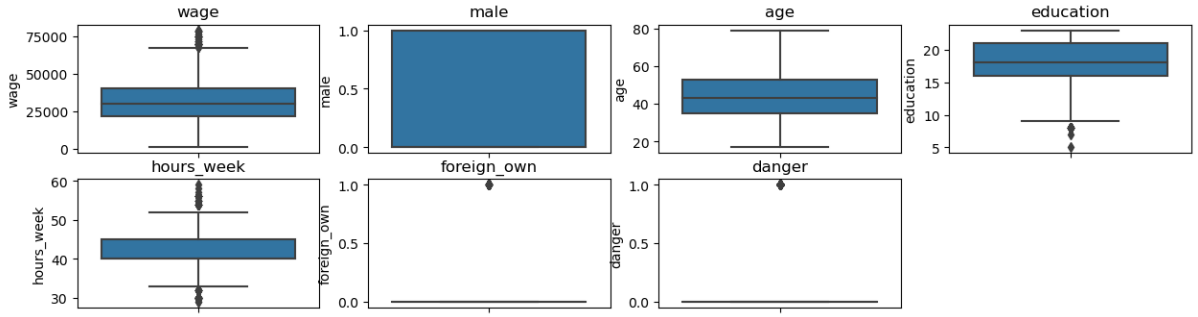
Ящичковые диаграммы в первоначальном виде

Кроме того, были изучены гистограммы распределения возраста и заработной платы.

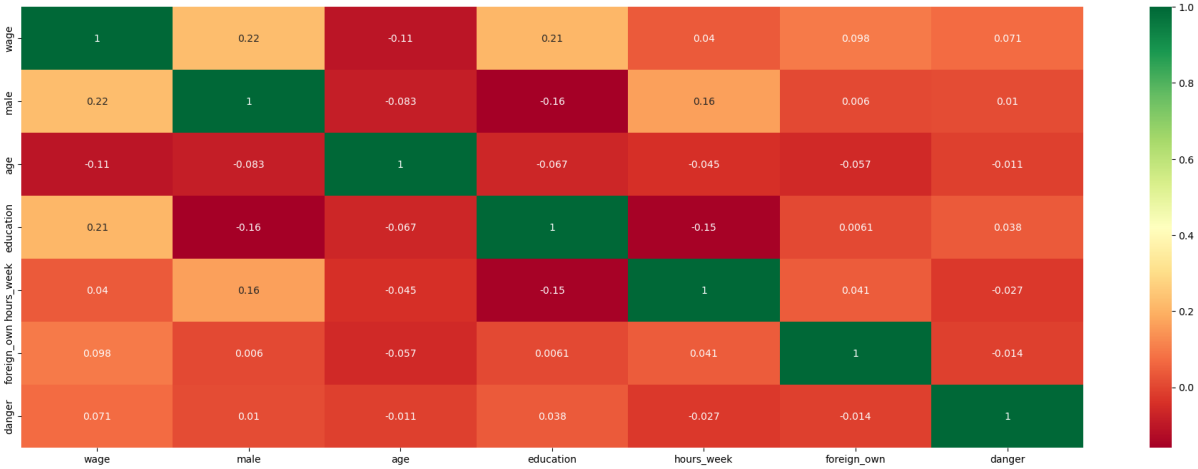




Используя метод межквартильного размаха (IQR), было выполнено удаление выбросов из набора данных для предотвращения искажения модели.



Завершающим этапом анализа стало использование матрицы корреляций для выявления взаимосвязей между переменными и оценки риска мультиколлинеарности.



Из полученной матрицы корреляций мы сформировали предварительные выводы о характере связей между изучаемыми переменными:

1. Wage - переменная

Наблюдается слабая положительная взаимосвязь между полом и величиной заработной платы и образованием и заработной платы, что может говорить о том, что мужчины в среднем имеют немного выше заработную плату, чем женщины и что более высокое образование положительно связано с уровнем заработной платы. Кроме того, наблюдается отрицательная корреляция между возрастом и заработной платой, что может говорить о том, что с возрастом величина заработной платы становится меньше.
Что касается переменных 'foreign_own' (работа в иностранной компании), 'danger' (опасность работы) и 'hours_week' (количество рабочих часов в неделю) статистически значимой связи с уровнем заработной платы обнаружено не было. Это может свидетельствовать о том, что указанные факторы не оказывают заметного влияния на величину дохода.

2. Male - переменная

   Наблюдается слабая положительная взаимосвязь между полом и количеством рабочих часов в неделю, что может говорить о том, что мужчины в среднем работают немного больше часов в неделю. Кроме того, наблюдается слабая отрицательная взаимосвязь между полом и образованием, что может говорить о том, что мужчины в среднем имеют более низкий уровень образования.
   Что касается переменных ‘age’ (возраст), foreign_own (работа в иностранной компании), danger (опасность работы) статистически значимой связи с ‘male’ (пол) обнаружено не было. Это может свидетельствовать о том, что указанные факторы не зависят от пола респондентов.

3. Age - переменная
   Кроме вышеуказанных переменных, остальные переменные не имеют статистической значимой связи с возрастом, что может говорить о том, что работа в иностранной фирме/ работа с повышенным уровнем опасности/образование/количество рабочих часов в неделю не связаны с возрастом человека
4. Education - переменная
   Наблюдается слабая отрицательная взаимосвязь между количеством рабочих часов в неделю и образованием, что может говорить о том,

что при более высоком образовании человек в среднем работает меньшее количество часов в неделю.
Кроме вышеуказанных переменных, остальные переменные практически не имеют статистически значимой связи с образованием, что может говорить о том, что работа в иностранной фирме/ работа с повышенным уровнем опасности не связаны с образованием человека

5. Hours_week - переменная
   Кроме вышеуказанных переменных, остальные переменные не имеют статистически значимой связи с количеством рабочих часов в неделю, что может говорить о том, работа в иностранной фирме/ работа с повышенным уровнем опасности не связаны с количеством рабочих часов в неделю
6. Foreign_own - переменная
   Кроме вышеуказанных переменных, danger не имеют статистически значимой связи с foreign_own, что может говорить о том, работа в иностранной фирме не связана с работой с повышенным уровнем опасности.

Оценка модели

Сравним несколько моделей. Первая - без логарифмов и без переменной hours_week^2.

$$\widehat{wage} = \hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 age^2 + \hat{\beta}_3 male + \hat{\beta}_4 technicum + \hat{\beta}_5 notfulluni + \hat{\beta}_6 bach + \hat{\beta}_7 postgrad + \hat{\beta}_8 hoursweek + \hat{\beta}_9 foreignown + \hat{\beta}_{10} danger$$

Вторая - с логарифмом на объясняемой переменной, но все еще без hours_week^2.

$$log(\widehat{wage}) = \hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 age^2 + \hat{\beta}_3 male + \hat{\beta}_4 technicum + \hat{\beta}_5 notfulluni + \hat{\beta}_6 bach + \hat{\beta}_7 postgrad + \hat{\beta}_8 hoursweek + \hat{\beta}_9 foreignown + \hat{\beta}_{10} danger$$

Третья - без логарифма, но с hours_week^2.

$$\widehat{wage} = \hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 age^2 + \hat{\beta}_3 male + \hat{\beta}_4 technicum + \hat{\beta}_5 notfulluni + \hat{\beta}_6 bach + \hat{\beta}_7 postgrad + \hat{\beta}_8 hoursweek + \hat{\beta}_9 foreignown + \hat{\beta}_{10} danger + \hat{\beta}_{11} hoursweek^2$$

Четвертая - с логарифмом на объясняемой переменной и с hours_week^2.

$$\widehat{log(wage)} = \hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 age^2 + \hat{\beta}_3 male + \hat{\beta}_4 technicum + \hat{\beta}_5 notfulluni + \hat{\beta}_6 bach + \hat{\beta}_7 postgrad + \hat{\beta}_8 hoursweek + \hat{\beta}_9 foreignown + \hat{\beta}_{10} danger + \hat{\beta}_{11} hoursweek^2$$

Для того, чтобы выявить из этих моделей ту, которая лучше описывает данные, мы сравнили их по R-squared adjusted. Он получился наибольшим в четвертой модели, поэтому в дальнейшем мы использовали ее для интерпретации.

Все коэффициенты оказались значимыми при любом разумном уровне значимости. Также гипотеза о неадекватности модели не подтверждается при любом разумном уровне значимости, следовательно, модель можно считать адекватной.

Все коэффициенты без квадрата положительно влияют на уровень зарплаты. Обе переменные с квадратом имеют форму параболы с ветвями вниз, так как их коэффициенты отрицательные.

	Dependent variable:			
	wage	log(wage)	wage	log(wage)
	MainModel1	MainModelLog1	MainModel2	MainModelLog2
	(1)	(2)	(3)	(4)
age	780.52***	0.03***	781.66***	0.03***
	(136.41)	(0.004)	(136.21)	(0.004)
I(age2)	-9.63***	-0.0003***	-9.63***	-0.0003***
	(1.53)	(0.0000)	(1.52)	(0.0000)
male	7,283.99***	0.23***	7,128.69***	0.22***
	(506.33)	(0.02)	(507.94)	(0.02)
tehnicum	2,017.05***	0.06**	1,980.53***	0.06**
	(734.65)	(0.02)	(733.65)	(0.02)
not_full_uni	4,911.23***	0.16***	4,762.50***	0.15***
	(1,351.39)	(0.04)	(1,350.20)	(0.04)
bach	9,601.32***	0.30***	9,482.79***	0.29***
	(770.81)	(0.02)	(770.58)	(0.02)
postgrad	21,000.03***	0.58***	21,351.68***	0.60***

	(3,361.37)	(0.11)	(3,358.22)	(0.11)
hours_week	120.05**	0.005***	1,981.83***	0.09***
	(52.27)	(0.002)	(588.74)	(0.02)
foreign_own	8,937.15***	0.25***	8,897.50***	0.25***
	(1,611.19)	(0.05)	(1,608.85)	(0.05)
danger	2,649.97***	0.09***	2,751.57***	0.09***
	(742.08)	(0.02)	(741.67)	(0.02)
I(hours_week2)			-21.95***	-0.001***
			(6.91)	(0.0002)
Constant	5,272.19	9.36***	-33,608.83***	7.68***
	(3,716.49)	(0.12)	(12,796.56)	(0.41)
Observations	3,071	3,071	3,071	3,071
R^2	0.16	0.15	0.16	0.16
Adjusted R^2	0.15	0.15	0.16	0.16
Residual Std. Error	13,562.63 (df = 3060)	0.44 (df = 3060)	13,542.55 (df = 3059)	0.44 (df = 3059)
F Statistic	56.40*** (df = 10; 3060)	55.94*** (df = 10; 3060)	52.34*** (df = 11; 3059)	52.80*** (df = 11; 3059)
Note:				*p**p***p<0.01

Выводы

Исходя из полученных коэффициентов модели, можем сделать следующие выводы о влиянии факторов, рассмотренных в гипотезах, на уровень заработной платы:

1) В компаниях с иностранным владельцем зарплаты выше на 25%. Коэффициент больше 0, соответственно, гипотеза о том, что при работе в подобных компаниях зарплаты увеличиваются верна. Значит, иностранные компании склонны к большим затратам на труд. Это может быть связано с тем, что такие компании имеют большие финансовые возможности и/или чаще относятся к отраслям, в которых труд дает большую отдачу от масштаба.
2) Зависимость зарплаты от количества рабочих часов в неделю имеет вид параболы ветвями вниз (так как коэффициент отрицателен). Следовательно, гипотеза о том, что зарплата сначала растет при

увеличении рабочих часов до 45 часов в неделю (вершина параболы), а затем начинает снижаться, верна. Можно предположить, что это связано с тем, что, если человек начинает работать слишком много, его производительность падает и/или с тем, что низкоквалифицированный труд, который чаще всего оплачивается ниже, затрачивает больше времени.

Также из анализа регрессий можно сделать выводы о других переменных, оказывающих влияние на размер зарплаты. При прочих равных:

3) Уровень образования положительно влияет на заработную плату. Так как была исключена, принята за базовую, дамми-переменная, отвечающая за наличие только школьного образования (school), то остальные ступени образования будут сравниваться со школьным. Из результатов логарифмической модели при получении индивидом профессионального образования в техникуме или ПТУ (technicum) зарплата увеличивается на 6% в отличии от индивида, имеющего только школьное образование; при неоконченном высшем образовании (not_full_uni) - на 15%; при наличии диплома о высшем образовании (bach) зарплата увеличивается на 29% относительно наличия школьного образования; при оконченной аспирантуре и более высоких уровнях образования - на 60%. Полученные результаты коэффициентов регрессии свидетельствуют о том, что при более высоком уровне образования зарплата выше.
4) Зарплаты мужчин на 22% выше чем у женщин. Это объясняется различными социально-демографическими факторами и рассматривается во многих научных работах,например, в статье лауреата Нобелевской премии Клаудии Голдин “A Grand Gender Convergence: Its Last Chapter”.
5) В местах работы с повышенной опасностью зарплаты выше на 9%. Предложение труда в отраслях, связанных с рисками для жизни меньше, поэтому необходим дополнительный стимул в виде более высокой заработной для компенсации рисков и возможных последствий работы на подобных предприятиях.
6) Зарплата параболически зависит от возраста. До определенного возраста зарплата растет, а затем начинает постепенно снижаться. Скорее всего, это связано со снижением производительности труда.