Аналитический отчет

Оглавление

Обработка данных

Немного об подготовке и обработке данных:

Я скачала файла в формате .dta с сайта о Данных обследования РМЭЗ НИУ ВШЭ. Согласно заданию, был взят файл 33 волны за 2024 год, репрезентативная выборка по индивидам. Дальнейшая работа с файлом выполнялась с помощью инструментов pandas: я считала файл, а затем очистила названия колонок от пробелов, привела их к нижнему регистру и в качестве разделителя во всех названиях точку поменяла на нижнее подчеркивание для единообразного представления и для того, чтобы дальнейший код нормально работал. С помощью документации к данным (Описание переменных) я нашла файл, в котором давалось содержательное пояснение для каждой переменной. Пользуясь поиском, я нашла все переменные, которые требовались в задании и переименовала колонки для дальнейшей работы с ними.

Таблица 1 – Соответствие оригинальных и переименованных названий колонок

Имя переменной в выборке	**Новое имя переменной**	**Содержательное значение**
ccj10	*wage*	Сколько денег в течение последних 30 дней Вы получили по основному месту работы после вычета налогов и отчислений?
cc_educ	*education*	ОБРАЗОВАНИЕ (ПОДРОБНО): старше 14 лет
cc_age	*age*	Возраст респондента
cc_marst	*family*	СЕМЕЙНОЕ ПОЛОЖЕНИЕ - 33 ВОЛНА
ccj72.172	*children*	Сколько всего у Вас детей?
ccj6.1a	*work_hours*	Сколько в среднем продолжается Ваш обычный рабочий день на этой работе? (часов)

cch5	*gender*	Пол респондента
region	*region*	Регион
cc_int_y	*year*	Год опроса
ccj161.3y	*exp*	Сколько полных лет и месяцев составляет Ваш официально оформленный общий трудовой стаж, не считая времени учебы на дневном отделении ВУЗа или техникума? (лет)

Далее было необходимо отфильтровать данные согласно варианту и создать новые категориальные и количественные переменные. Я делала выборку по мужчинам за 2024 год в Новосибирской области, Бердском районе, поэтому сразу отфильтровала свои данные по этим трем параметрам.

- Следующим шагом я отобрала респондентов в возрасте от 18 до 55 лет, избавилась от строк, в которых был хотя бы 1 пропуск.
- Создала новую переменную *exp2*, которая равняется стажу работы в квадрате.
- Сделала категориальную переменную *is_children*, которая принимает значение 1, если у респондента есть хотя бы 1 ребенок, и 0 в противном случае на основании переменной *children*.
- Сделала категориальную переменную *educ* на основе переменной *education* (0 если респондент учился только в школе, 1 если в ПТУ или техническом училище, 2 если респондент получал высшее образование).
- На основании переменной *family* была получена переменная *marriage* (1 если респондент живет с супругой (даже если брак гражданский), и 0 если нет).
- Обнаружилось, что по переменной *work_hours* есть респонденты, которые работают ненормированный рабочий день (99999996). Так как выборка небольшая, было принято решение не избавляться от таких значений, а заменить их на среднее по этому показателю.
- В конце были удалены промежуточные переменные *(children, education, gender, region, year, family)*.

В качестве дополнительной переменной я добавила переменную *marriage*. Я хочу проверить гипотезу о том, что женатые мужчины в среднем зарабатывают больше неженатых, так как как правило, на работающих мужчин возлагается обязательство обеспечивать семью.

Обзор основных переменных и связей между ними

Для всех величин были найдены описательные статистики: минимум, максимум, среднее, стандартное отклонение, размах. Посмотрим на самые интересные результаты.

1. Самому молодому респонденту 32 года, что почти в 1.8 раз больше минимальной границе, по которой мы фильтровали респондентов (от 18). Максимальный возраст респондента – 54.5, что совпадает с верхней границей отбора.
2. Размах 0 в показателе *is_children*, а также нулевое стандартное отклонение говорят о том, что у всех респондентов в выборке есть хотя бы 1 ребенок.

3. Максимальное значение 2 в *educ* говорит о том, что ни один из индивидов не учился в институте, университете или академии, то есть не получал высшее образование.
4. Среднее значение около 0.95 по *marriage* говорит о том, что 95% респондентов женаты/проживают в гражданском браке.

Заработная плата – *wage*

- Среднее – 68671.5
- Медиана – 70000
- Мода – 70000
- Среднее < Медиана = Мода

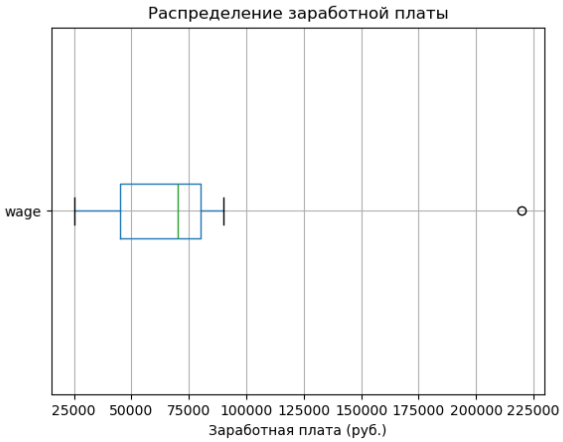

Рисунок 1 – Распределение заработной платы

Видно, что в данных есть 1 выброс – человек с очень большой зарплатой в 220 тыс.руб.

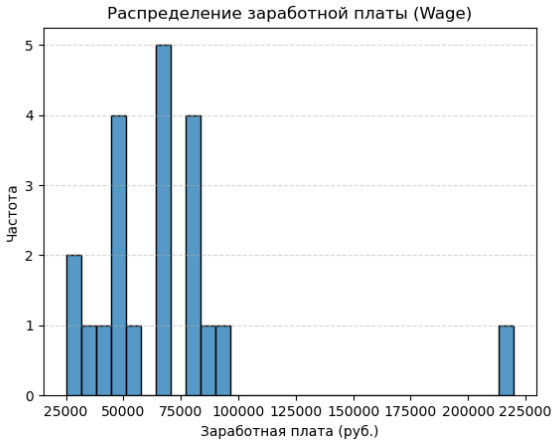

Рисунок 2 – Гистограмма распределения по заработной плате

- Коэффициент асимметрии – 2.86
- Коэффициент эксцесса – 10.97

Коэффициент асимметрии положителен, левосторонняя асимметрия. Коэффициент эксцесса положителен и значительно больше 3, что говорит об островершинном распределении.

Таблица 2 – Корреляционная таблица

	wage	age	work_hours	exp	exp2	is_children	educ	marriage

wage	1	-0,015	-0,004	-0,494	-0,389		-0,051	0,252
age	-0,015	1	-0,404	0,773	0,849		-0,073	0,085
work_hours	-0,004	-0,404	1	-0,295	-0,363		-0,13	0,060
exp	-0,494	0,773	-0,295	1	0,98		0,092	0,182
exp2	-0,389	0,849	-0,363	0,98	1		0,063	0,2
is_children								
educ	-0,051	-0,073	-0,13	0,092	0,063		1	0,194
marriage	0,252	0,085	0,060	0,182	0,2		0,194	1

Построив корреляционную матрицу, можно заметить несколько интересных вещей:

- Отсутствует корреляция всех коэффициентов с параметром *is_children.* Это связано с тем, что абсолютно все респонденты в нашей выборке имеют хотя бы одного ребенка, поэтому для всех этот параметр равен 1, а значит, неинформативен для построения регрессии.
- Зарплата отрицательно коррелирует со всеми параметрами, кроме *marriage*, который отвечает за семейное положение. Однако, корреляция умеренная, ближе к слабой.
- Удивительным образом, наиболее сильная отрицательная корреляция между зарплатой и опытом работы (-0.494).
- Возраст сильно положительно коррелирует с опытом работы, что неудивительно.
- С возрастом среднее количество рабочих часов сокращается, о чем говорит отрицательный коэффициент корреляции между *age* и *work_hours* (-0.404).

Регрессионный анализ

Согласно уравнению Минсера логарифм заработной платы зависит от количества лет обучения, а также от трудового стажа и трудового стажа в квадрате. Возьмем за основу эту же модель и добавим к ней другие параметры, которые есть в нашей выборке: продолжительность рабочего дня, возраст и семейное положение. Не будем использовать фактор, показывающий наличие детей, потому что как было сказано выше, для каждого респондента в выборке дети есть, а значит на выборке включение этого фактора не даст никакой информации. Таким образом, регрессионное уравнение принимает вид:

$$ln(w_i) = \beta_1 + \beta_2 \times age_i + \beta_3 \times work\ hours_i + \beta_4 \times exp_i + \beta_5 \times exp^2{}_i + \beta_6 \times educ_i + \beta_7 \times marriage_i + \varepsilon_i$$

$$ln(w_i) = 9.197 + 0.045 \times age_i - 0.032 \times work\ hours_i - 0.037 \times exp_i - \\ - 0.001 \times exp^2{}_i + 0.072 \times educ_i + 1.245 \times marriage_i$$

R^2 = 0.607

R_{adj}^2 = 0.439

60.7% вариации заработной платы объясняется зависимыми переменными. Если скорректировать показатель R^2, то 43.9% вариации заработной платы будет объяснено зависимыми переменными.

Логарифм зарплаты отрицательно зависит от количества рабочих часов, опыта работы и квадрата опыта работы: при увеличении продолжительности рабочего дня на 1 час логарифм зарплаты падает на 0.032, при увеличении опыта работы на 1 год логарифм зарплаты падает на 0.037 (при прочих равных).

Зависимая переменная положительно зависит от возраста, уровня образования и семейного положения.

Свободный коэффициент модели не имеет содержательного экономического объяснения.

Можно заметить, что результат оценки модели и значимости отдельных коэффициентов получился совершенно контринтуитивным: зарплата отрицательно (пусть и слабо) зависит от продолжительности рабочего дня, от опыта. Более того, эти коэффициенты оказались еще и не значимыми. Такой результат можно объяснить довольно специфическим набором данных: человек с самой высокой зарплатой в выборке имеет самый маленький опыт работы среди всех, и наоборот, люди с очень большим, почти максимальным уровнем работы, имеют почти минимальную зарплату по выборке. Эти особенности могли бы быть скорректированы большим объемом выборки, однако общее количество респондентов всего 21, что сильно влияет на корректность оценки коэффициентов.

Гипотезы о значимости отдельных коэффициентов и модели в целом

Гипотеза о значимости модели в целом на 5%-ом уровне значимости:

$$H_0: \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$

$$H_A: \exists\, i: \beta_i \neq 0$$

$F \sim F_{6;14\,;5\%}$

Критическая область: $[2.85;\ +\infty]$

$F_{H_0} = 3.611$

$P - value = Prob(F - statistics) = 0.0224$

p-value = 0.0224 < 0.05, а расчетное значение F-статистики попадает в критическую область, поэтому мы не принимаем H_0, модель в целом значима.

Проверим отдельно гипотезу о значимости коэффициента на уровне значимости 1% перед параметром *marriage*, который отвечает за семейное положение респондента.

$$H_0: \beta_7 = 0$$

$$H_A: \beta_7 \neq 0$$

$t \sim t_{14,1\%}$

$t_{14;1\%} = 2.977 \rightarrow t_{\text{крит}} \in (-\infty;\ -2.977] \cup [2.977;\ +\infty)$

$t_{H_0} = 3.165$

Значение t-статистики попадает в критическую область, нулевая гипотеза не принимается. Значит, коэффициент статистически значим на уровне 1%.

Ограничения модели

Проверим наличие мультиколлинеарности двумя способами.

1 способ: обратимся к корреляционной матрице. Мы видим сильную связь между *age* и *exp* (0.773), между *age* и *exp2* (0.849), а также почти линейную зависимость между *exp* и *exp2* (0.98). Это логично. Очевидно, что переменная и она же в квадрате будут сильно коррелировать. Также понятно, что возрастом опыт работы растет.

2 способ: посчитаем VIF.

Коэффициент	VIF
age	5.675
work_hours	1.399
exp	42.985
exp2	63.190
educ	1.148
marriage	1.199

Значение VIF больше 10 (значительно) у exp и exp2 говорит о наличии сильной мультиколлинеарности.

Результаты двух способов проверки на мультиколлинеарность согласуются. Удалим переменную, отвечающую за опыт работы в квадрате. В классической модели Минсера она есть, однако при моем маленьком размере выборки не удается компенсировать почти линейную зависимость с помощью объема выборки, поэтому целесообразно избавиться от этой переменной.

Проведем тест Бройша-Пагана (уровень значимости – 5%):

H_0: гетероскедастичность отсутствует (дисперсии ошибок постоянны)

H_A: есть гетероскедастичность

$LM \sim X - \text{квадрат}_{6\,;5\%}$

Критическая область: $[18.548;\ +\infty]$

$X_{H_0} = 5.87$

$P - value = 0.437$

Нулевая гипотеза не отвергается, а значит, в данных нет гетероскедастичности.

Проведем так же тест Уайта (уровень значимости – 5%):

H_0: гетероскедастичность отсутствует (дисперсии ошибок постоянны)

H_A: есть гетероскедастичность

$LM \sim X - \text{квадрат}_{19\,;5\%}$

Критическая область: $[38.582;\ +\infty]$

$X_{H_0} = 20.06$

$P - value = 0.39$

Значение статистики не попадает в критическую область. Нет оснований отвергать нулевую гипотезу. Можно считать дисперсии ошибок постоянными.

Выводы

Основной проблемой в ходе выполнения всего задания была маленькая выборка. При такой маленькой выборке попытки бороться с мультиколлинеарностью, которая, как показали тесты, была, довольно бесполезны. Мультиколлинеарность перестает быть большой проблемой на больших объемах, а на маленьких остается заметна. Так что да, для устранения большинства проблем нужно просто брать выборку больше.