Федеральное государственное бюджетное образовательное учреждение высшего образования

«Российская академия народного хозяйства и государственной службы

при Президенте Российской Федерации»

Отчет по проектной работе по курсу эконометрика 1

По теме:

«Исследование гендерного разрыва в заработных платах»

Выполнили:

Равилова Диана

Анфимов Александр

Аникеев Никита

Кузнецов Данила

СП-20

Москва, декабрь 2022.

Оглавление

[1. Введение 2](#_Toc122348821)

[2. Обработка данных 2](#_Toc122348822)

[3. Базовая модель 3](#_Toc122348823)

# Введение

# Обработка данных

В нашей работе мы использовали данные Российского онеторинга экономического положения и здоровья населения НИУ ВШЭ. Для исследования были взяты данные за 2017 год для Санкт-Петербурга. Рассматривались совершеннолетние индивиды и их доходы за вычетом пенсий. Также для более точного исследования не рассматривались индивиды с доходом больше ста тысяч рублей и меньше десяти тысяч рублей, таким образом мы постарались избавиться от выбросов и нерелевантных данных.

Некоторые вопросы и ответы к ним были сокращены до бинарных. Так, вопрос о максимальной степени образования был сведен к тому, имеет или получает индивид высшее образование или нет. Так как зарплаты могут сильно отличаться у индивидов из-за сферы деятельности, показатель отрасли мы вынесли в отдельную переменную и он стал бинарным для каждой отрасли. То есть если человек работает в 10 отрасли, то sphere\_10 = 1, а все остальные 0.

Проанализируем равномерность нашей выборки. Мы имеем данные по 166 индивидам. Из них 75 мужчин и 91 женщин. Средняя зарплата женщин составляет 75% от мужской. При этом в области энергетики зарплата женщин превосходит на 193%, а в области торговли и ботового обслуживания, наоборот, меньше на 53%. Также большой разрыв в зарплатах наблюдается в области финансов (+184%), IT (+91%), армия и мвд (-46%). Однако так как в этих отраслях представлено малое количество работников, считать эти данные релевантными стоит с осторожностью. Средний возраст мужчин составил 42 года, а женщин – 44 года. Максимальная зарплата равна 90 000 рублей, минимальная – 10050 (данные приведены после удаления выбросов).

|  |  |  |  |
| --- | --- | --- | --- |
| Показатель |  | **Мужчины** | **Женщины** |
| брак | Численность, чел | 56 | 53 |
| Доля от численность работников этого пола | 75% | 58% |
| Употребление алкоголя | Численность, чел | 21 | 36% |
| Доля от численность работников этого пола | 28% | 40% |
| Наличие детей | Численность, чел | 47 | 68% |
| Доля от численность работников этого пола | 63% | 75% |
| Зарплата | Среднее значение, руб | 38350 | 28817 |
| Наличие проблем со здоровьем | Среднее значение | |  | | --- | | 0,773 | |  | | 0,703 |
| Уровень здоровья (от 1 до 5) | Среднее значение | 2,4 | 2,45 |
| Возраст | Среднее значение, лет | 41,9 | 43,9 |
| Опыт | Среднее значение, лет | 17,3 | 18,9 |
| Доля проработанных лет | 41% | 43% |

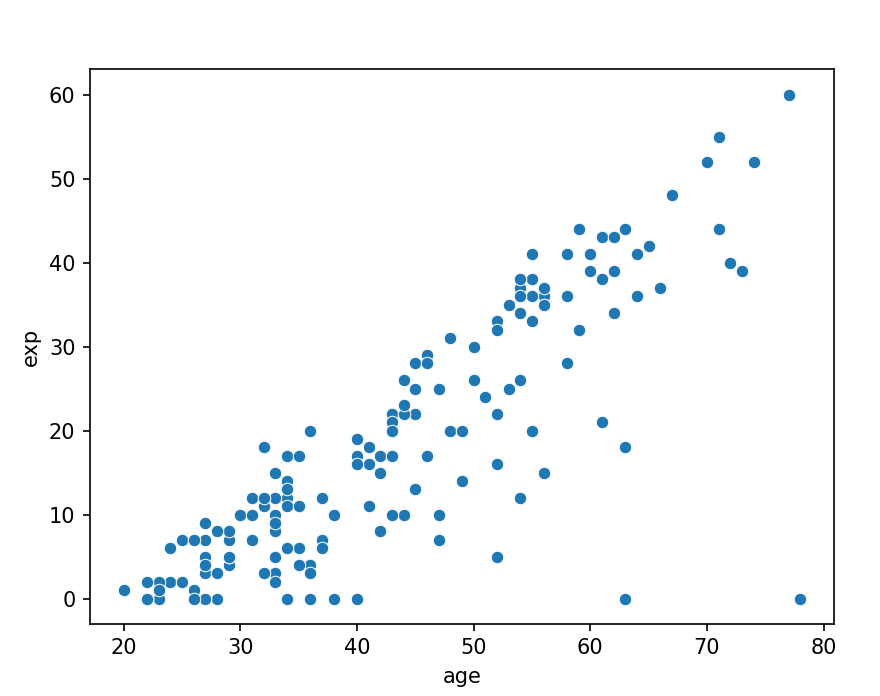
таблица. Описательная статистика.

# Базовая модель

Для исследования гендерного разрыва было решено взять в качестве зависимой переменной логарифм заработной платы. Наша проектная работа основывалась на многих известных работах в этой области. В особенности мы использовали работу И.И. Елисеева и М.П. Декина, которые строили модель на модификации уравнения Минцера.

Для того чтобы проверить, что мы правильно выбрали зависимую переменную, был проведен тест Зарембки. Он показал, что отличие логарифмической и линейной модели значимо. Была выбрана модель с меньшей RSS. Для линейной она равна 22.5988, а для логарифмической она оказалась равна 19.1010. Таким образом, модель с зависимой переменной в виде логарифма зарплаты использовать правильнее.

В качестве объясняющих переменных используются опыт работы в годах, квадрат опыта работы, наличие высшего образования, пол работника, отрасль, к которой относится его вид деятельности и показатель наличия подчиненных у работника. Стоит отметить, что на наших данных опыт работника и его возраст линейно зависимы(рис 1).



рисунок

Поэтому в качестве еще одной базовой модели мы оценили аналогичную модель, но вместо опыта взяли возраст. Такая модель дает похожие коэффициенты для всех переменных, но сами коэффициенты получаются более значимы при опыте. Кроме того, информационные критерии AIC и BIC показывают немного меньшие значения. Поэтому для в качестве основных показателей далее будет рассматриваться опыт и его квадрат, а не аналогичные переменные с возрастом.

Таким образом, мы получили МНК оценки коэффициентов в базовой модели для описанных показателей ( ).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***БАЗОВАЯ МОДЕЛЬ МНК С И ОПЫТОМ*** | | | | | | |
| Dep. Variable: | np.log(salary) | R-squared: | 0.428 |  |  |  |
| Model: | OLS | Adj. R-squared: | 0.321 |  |  |  |
| Method: | Least Squares | F-statistic: | 3.998 |  |  |  |
|  |  | Prob (F-statistic): | 5.18e-08 |  |  |  |
| No. Observations: | 166,00 | AIC: | 164.2 |  |  |  |
| Df Residuals: | 139,00 | BIC: | 248.3 |  |  |  |
| Df Model: | 26,00 |  |  |  |  |  |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| Intercept | 10,189 | 0,179 | 56,974 | 0,000 | 9,835 | 10,543 |
| C(sphere)[T.1] | 0,330 | 0,168 | 1,966 | 0,051 | -0,002 | 0,662 |
| C(sphere)[T.2] | 0,826 | 0,204 | 4,055 | 0,000 | 0,423 | 1,228 |
| C(sphere)[T.3] | 0,734 | 0,294 | 2,498 | 0,014 | 0,153 | 1,314 |
| C(sphere)[T.4] | 0,462 | 0,396 | 1,166 | 0,245 | -0,321 | 1,245 |
| C(sphere)[T.5] | 1,274 | 0,397 | 3,212 | 0,002 | 0,490 | 2,059 |
| C(sphere)[T.6] | 0,556 | 0,173 | 3,222 | 0,002 | 0,215 | 0,898 |
| C(sphere)[T.7] | 0,533 | 0,149 | 3,570 | 0,000 | 0,238 | 0,828 |
| C(sphere)[T.9] | 0,041 | 0,255 | 0,161 | 0,873 | -0,464 | 0,546 |
| C(sphere)[T.10] | 0,428 | 0,170 | 2,512 | 0,013 | 0,091 | 0,764 |
| C(sphere)[T.11] | 0,286 | 0,232 | 1,235 | 0,219 | -0,172 | 0,744 |
| C(sphere)[T.12] | 0,412 | 0,178 | 2,310 | 0,022 | 0,059 | 0,765 |
| C(sphere)[T.13] | -0,030 | 0,191 | -0,158 | 0,874 | -0,408 | 0,347 |
| C(sphere)[T.14] | 0,401 | 0,141 | 2,851 | 0,005 | 0,123 | 0,680 |
| C(sphere)[T.15] | 0,387 | 0,250 | 1,550 | 0,123 | -0,106 | 0,880 |
| C(sphere)[T.16] | 0,446 | 0,292 | 1,530 | 0,128 | -0,130 | 1,023 |
| C(sphere)[T.17] | 0,195 | 0,159 | 1,226 | 0,222 | -0,120 | 0,510 |
| C(sphere)[T.20] | 0,946 | 0,395 | 2,396 | 0,018 | 0,165 | 1,726 |
| C(sphere)[T.21] | 0,314 | 0,298 | 1,054 | 0,294 | -0,275 | 0,902 |
| C(sphere)[T.26] | -0,152 | 0,251 | -0,606 | 0,546 | -0,648 | 0,344 |
| C(sphere)[T.27] | 0,688 | 0,292 | 2,357 | 0,020 | 0,111 | 1,264 |
| C(sphere)[T.28] | -0,120 | 0,292 | -0,411 | 0,682 | -0,698 | 0,458 |
| gender | -0,194 | 0,069 | -2,816 | 0,006 | -0,331 | -0,058 |
| exp | 0,012 | 0,007 | 1,629 | 0,106 | -0,003 | 0,026 |
| I(exp \*\* 2) | 0,000 | 0,000 | -1,844 | 0,067 | -0,001 | 0,000 |
| degree | 0,124 | 0,069 | 1,786 | 0,076 | -0,013 | 0,260 |
| boss | -0,295 | 0,092 | -3,217 | 0,002 | -0,476 | -0,114 |

таблица. Оценка базовой модели с опытом.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***БАЗОВАЯ МОДЕЛЬ МНК С ВОЗРАСТОМ*** | | | | | | |
| Dep. Variable: | np.log(salary) | R-squared: | 0.435 |  |  |  |
| Model: | OLS | Adj. R-squared: | 0.329 |  |  |  |
| Method: | Least Squares | F-statistic: | 4.109 |  |  |  |
|  |  | Prob (F-statistic): | 2.73e-08 |  |  |  |
| No. Observations: | 166,00 | AIC: | 166.2 |  |  |  |
| Df Residuals: | 139,00 | BIC: | 250.2 |  |  |  |
| Df Model: | 26,00 |  |  |  |  |  |
| Covariance Type: | nonrobust |  |  |  |  |  |
|  | coef | std err | t | P>|t| | [0.025 | 0.975] |
| Intercept | 9,970 | 0,351 | 28,397 | 0,000 | 9,276 | 10,664 |
| C(sphere)[T.1] | 0,290 | 0,168 | 1,723 | 0,087 | -0,043 | 0,623 |
| C(sphere)[T.2] | 0,801 | 0,202 | 3,961 | 0,000 | 0,401 | 1,201 |
| C(sphere)[T.3] | 0,722 | 0,291 | 2,482 | 0,014 | 0,147 | 1,296 |
| C(sphere)[T.4] | 0,436 | 0,394 | 1,106 | 0,271 | -0,344 | 1,216 |
| C(sphere)[T.5] | 1,261 | 0,394 | 3,197 | 0,002 | 0,481 | 2,041 |
| C(sphere)[T.6] | 0,519 | 0,173 | 3,009 | 0,003 | 0,178 | 0,861 |
| C(sphere)[T.7] | 0,488 | 0,150 | 3,266 | 0,001 | 0,193 | 0,784 |
| C(sphere)[T.9] | 0,039 | 0,251 | 0,153 | 0,878 | -0,457 | 0,534 |
| C(sphere)[T.10] | 0,376 | 0,170 | 2,208 | 0,029 | 0,039 | 0,713 |
| C(sphere)[T.11] | 0,267 | 0,229 | 1,165 | 0,246 | -0,186 | 0,721 |
| C(sphere)[T.12] | 0,389 | 0,177 | 2,200 | 0,029 | 0,039 | 0,739 |
| C(sphere)[T.13] | -0,065 | 0,191 | -0,338 | 0,736 | -0,442 | 0,313 |
| C(sphere)[T.14] | 0,354 | 0,141 | 2,517 | 0,013 | 0,076 | 0,633 |
| C(sphere)[T.15] | 0,346 | 0,248 | 1,395 | 0,165 | -0,145 | 0,837 |
| C(sphere)[T.16] | 0,411 | 0,290 | 1,415 | 0,159 | -0,163 | 0,985 |
| C(sphere)[T.17] | 0,184 | 0,157 | 1,171 | 0,244 | -0,127 | 0,495 |
| C(sphere)[T.20] | 0,880 | 0,394 | 2,230 | 0,027 | 0,100 | 1,659 |
| C(sphere)[T.21] | 0,217 | 0,296 | 0,731 | 0,466 | -0,369 | 0,802 |
| C(sphere)[T.26] | -0,168 | 0,249 | -0,674 | 0,502 | -0,661 | 0,325 |
| C(sphere)[T.27] | 0,624 | 0,292 | 2,141 | 0,034 | 0,048 | 1,201 |
| C(sphere)[T.28] | -0,080 | 0,290 | -0,277 | 0,782 | -0,654 | 0,493 |
| gender | -0,189 | 0,069 | -2,742 | 0,007 | -0,325 | -0,053 |
| age | 0,019 | 0,014 | 1,332 | 0,185 | -0,009 | 0,046 |
| I(age \*\* 2) | 0,000 | 0,000 | -1,617 | 0,108 | -0,001 | 0,000 |
| degree | 0,154 | 0,067 | 2,293 | 0,023 | 0,021 | 0,286 |
| boss | -0,295 | 0,091 | -3,246 | 0,001 | -0,475 | -0,115 |

таблица. Оценка базовой модели с возрастом.

При целевом показателе gender значение примерно равно -0,19. То есть при всех остальных равных показателях зарплата мужчин и женщин отличается на 19%, и у мужчин она больше. Также стоит отметить, что этот коэффициент значим даже на 1% уровне значимости.

# Проверка условий Гаусса-Маркова

* 1. Гетероскедостичность

Для всех рассматриваемых и оценённых моделей проводились тесты Глейзера, Голдфелда-Квандта и Бройша-Пагана. Такое количество тестов позволяет с разных сторон оценить возможную гетерескедостичность, хотя и ожидается, что они дадут похожие результаты.

Для теста Голдфелда-Квандта нулевая гипотеза состоит в том, что у нас нет гетероскедостичности. Приведем здесь только полученные пи-значения (таблица