**Домашнее задание 3, Часть 3-4**

**Переверзев Виктор, Низоля Валерия, группа БЭК173**

Пайплайн для выбора регрессионной модели для индивидов в России в 2012 году (21 волна) по данным РЛМС (репрезентативная выборка).

1. Найти дескриптивные статистики (max, min, mean, sd)

2. Создать переменные

3. Построить гистограммы для переменных

4. Фильтрация (отбор) данных

5. Создание дамми-переменных для категориальных переменных

6. Диаграммы рассеяния. Похожа ли на линейную?

5. Оценить стартовую простую линейную модель со всеми созданными переменными

6. Работа с выбросами (использовать дамми для выбросов или исключить их совсем)

7. Выбрать функциональную модель (логарифмическая, линейная или полулогарифмическая)

8. Тест Рамсея на спецификацию модели

9. Проверка мультиколлинеарности

10. Метод исключения (в случае мультколлинеарности)

11. Тест Бройша-Пагана на гетероскедастичность

12. Qqnorm для остатков

13. Тест Шапиро-Уилка на нормальность остатков

14. Определить лучшую модель и проанализировать её с точки зрения статистики и экономики.

15. Проверка интересных гипотез (тест Вальда, тест правдоподобия)

16. Оценка бинарных моделей и проверка их качества

17. Предельные эффекты

База РЛМС – прекрасная возможность анализировать какие-либо модели на примере российской действительности. Очень важно, на наш взгляд, не погрязнуть в этом огромном количестве данных и выбрать наиболее релевантные для того или иного вопроса исследования. В этой домашней работе мы решили взять репрезентативную выборку по индивидам (21 волна) и попробовать построить адекватную регрессионную модель для размера среднемесячной зарплаты. Для стартовой модели мы использовали следующие переменные:

1. Salary - За последние 12 месяцев какова была Ваша среднемесячная зарплата на этом предприятии после вычета налогов - независимо от того, платят Вам ее вовремя или нет? (qj13.2)
2. Age - Количество полных лет (q\_age)
3. Sex (дамми-переменная) - Пол респондента (qh5)
4. Status - ТИП НАСЕЛЕННОГО ПУНКТА (status)
5. Diplom - ЗАКОНЧЕННОЕ ОБРАЗОВАНИЕ (ГРУППА) - 21 ВОЛНА (q\_diplom)
6. Job\_satisfaction - Насколько Вы удовлетворены или не удовлетворены Вашей работой в целом? (qj1.1.1)
7. Population -ЧИСЛЕННОСТЬ НАСЕЛЕНИЯ (popul)
8. Married - СЕМЕЙНОЕ ПОЛОЖЕНИЕ - 21 ВОЛНА (q\_marst)
9. Professional\_group - ПРОФЕССИОНАЛЬНАЯ ГРУППА - 21 ВОЛНА - по коду qj2cod08 (q\_occup08)
10. Children -  Сколько всего у Вас детей? (qj72.172)

Посмотрим на выбранные данные в том виде, как они загрузились в R

Observations: 4,872

Variables: 11

$ salary *<dbl>* 30000, 15000, 20000, 15000, 13000, 2000...

$ age *<dbl>* 39, 58, 55, 47, 54, 58, 59, 59, 57, 49,...

$ sex *<dbl>* 2, 2, 1, 2, 2, 1, 2, 2, 1, 2, 1, 2, 1, ...

$ status *<dbl>* 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, ...

$ diplom *<dbl>* 6, 2, 4, 2, 4, 4, 5, 3, 2, 4, 3, 6, 6, ...

$ job\_satisfaction *<dbl>* 1, 2, 3, 4, 2, 3, 2, 2, 2, 3, 4, 2, 2, ...

$ population *<dbl>* 12161, 12161, 12161, 3702, 2422, 2422, ...

$ married *<dbl>* 5, 2, 2, 2, 2, 2, 5, 2, 2, 2, 2, 2, 2, ...

$ professional\_group *<dbl>* 1, 9, 8, 8, 3, 8, 2, 5, 7, 9, 8, 1, 7, ...

$ children *<dbl>* 2, 3, 2, 2, 3, 3, 2, 3, 3, 3, 1, 1, 1, ...

Чтобы лучше ориентироваться в данных, рассмотрим дескриптивные статистики для переменных из датасета:

mean sd

salary 18353.56 15056.23

age 44.45 11.44

sex 1.57 0.50

status 2.12 1.20

diplom 4.70 1.12

job\_satisfaction 2.31 0.94

population 1464358.84 3340015.91

married 2.53 1.00

professional\_group 4.64 2.56

children 1.68 0.75

min max

salary 1000 280000

age 19 86

sex 1 2

status 1 4

diplom 1 6

job\_satisfaction 1 5

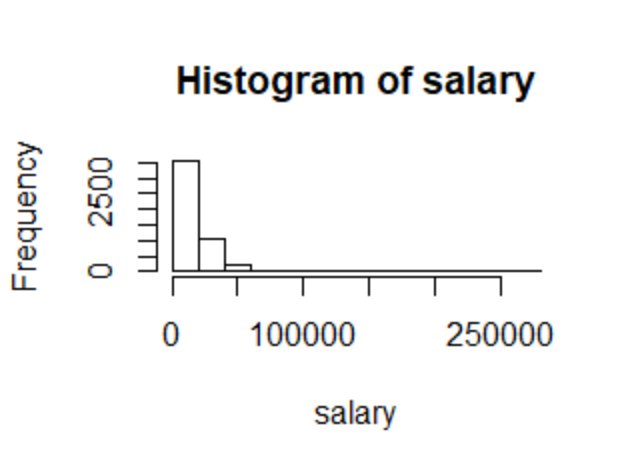
population 7 11503501

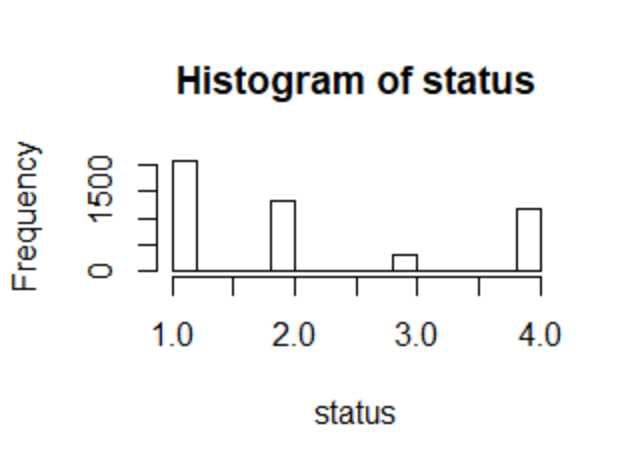
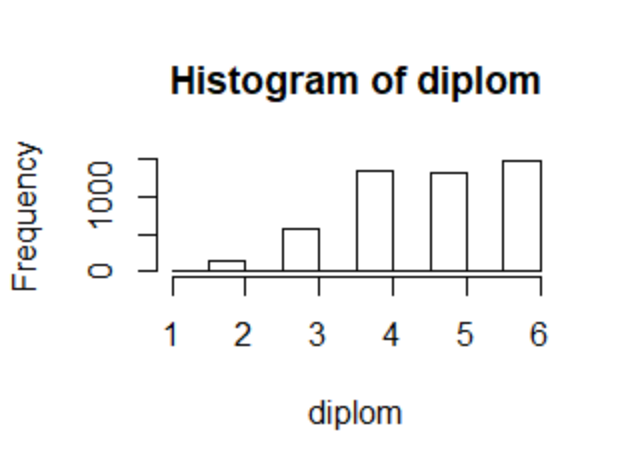
married 1 6

professional\_group 0 9

children 1 8

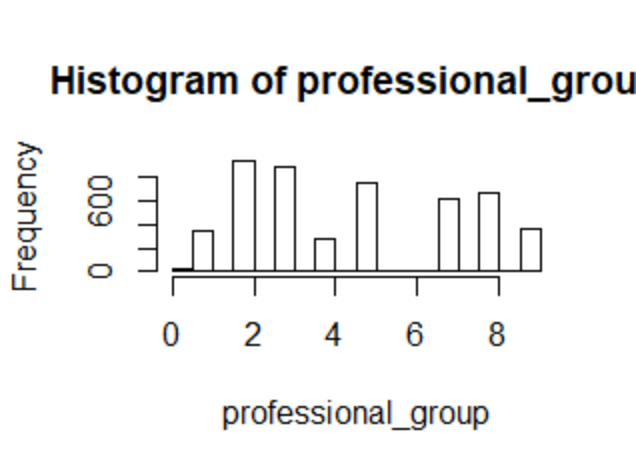
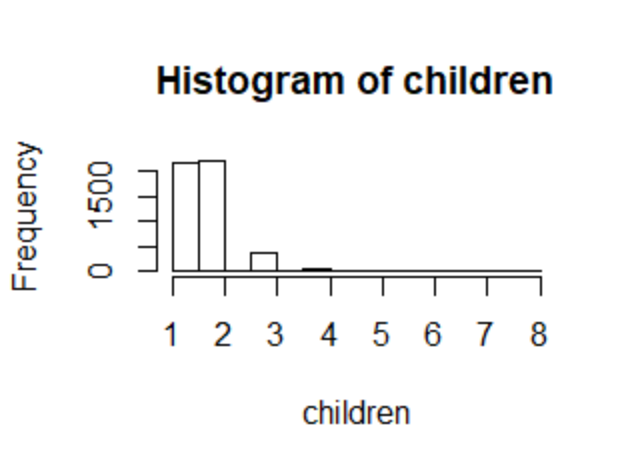
Теперь визуализируем их распределение с помощью гистограмм. Гистограммы для дамми-переменных не принесут новой информации после представленных дескриптивных статистик (выше уже понятно, к примеру, что мужчин и женщин примерно поравну, потому что mean примерно 1.57), поэтому не будем их изображать:

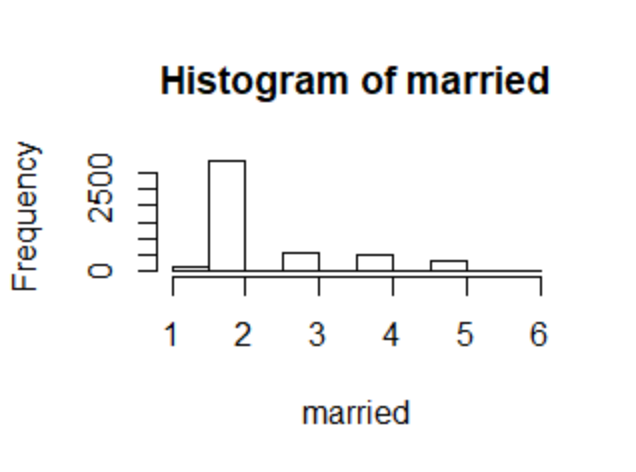










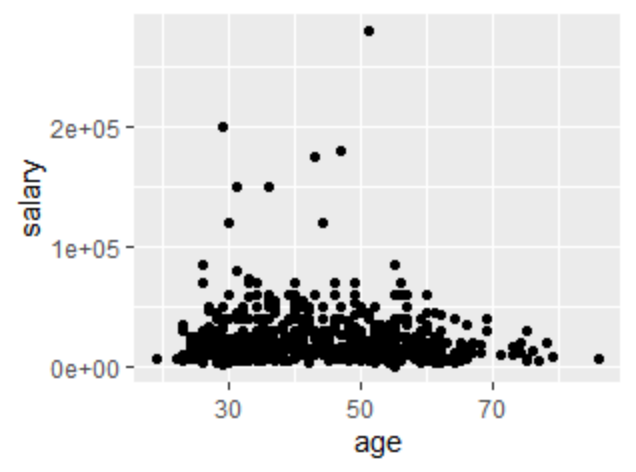
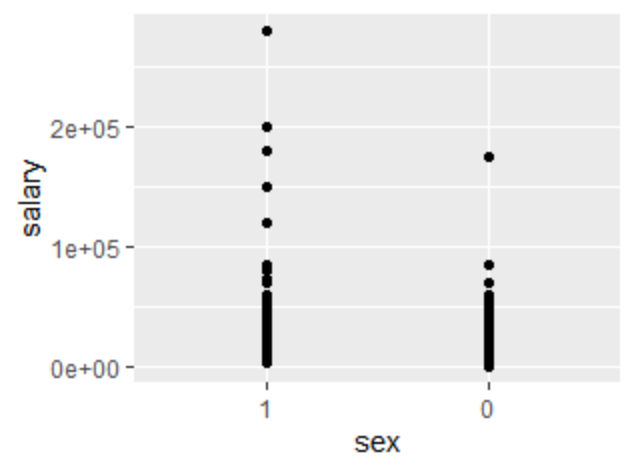
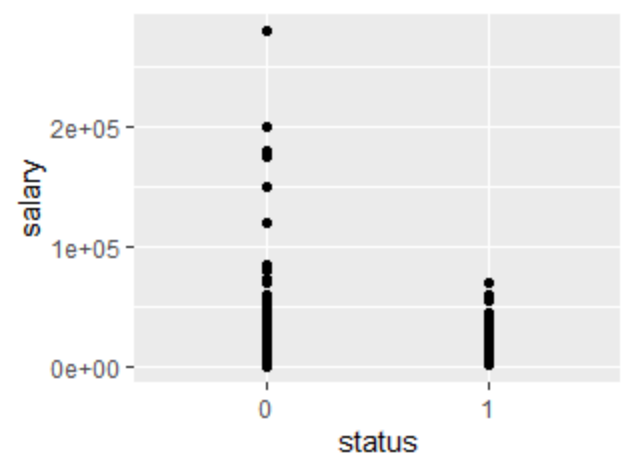


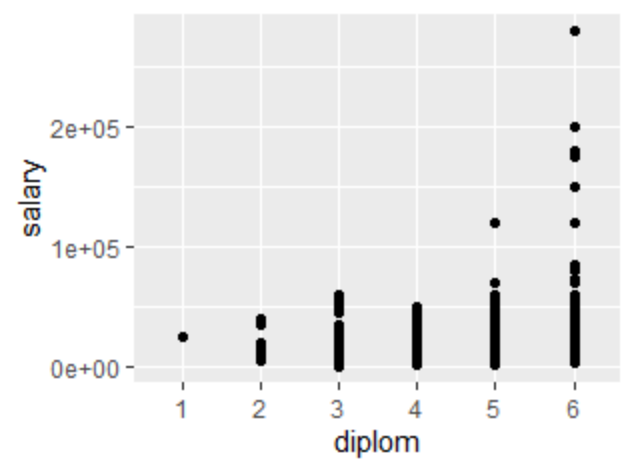
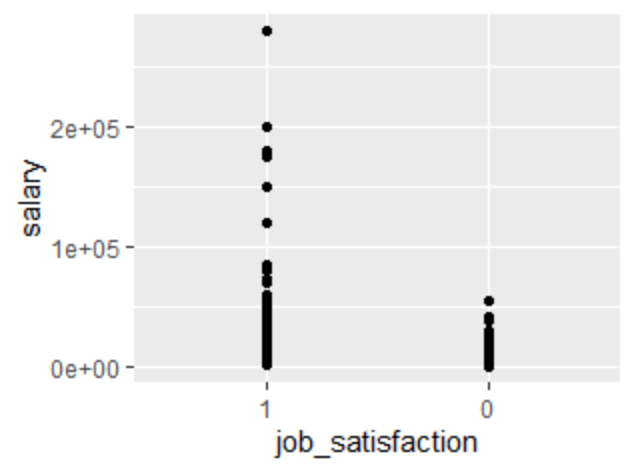
Для удобства дальнейшей работы и более доступной интерпретации модели будем использовать не все ответы респондентов, записанные в базе РЛМС, а только некоторые из них. Так, к примеру, конечно, степень удовлетворенности работой может быть разной, как представлено в данных, но мы решили взять только два противоположных значения по своей сути: удовлетворён или нет, чтобы создать в дальгейшем дамми-переменную job\_satisfaction. Аналогичная ситуация с местом жительства: проще взять только город/область и с наличием партнёров (зарегистрированы ли в браке? Проживаете ли вместе неофициально?)

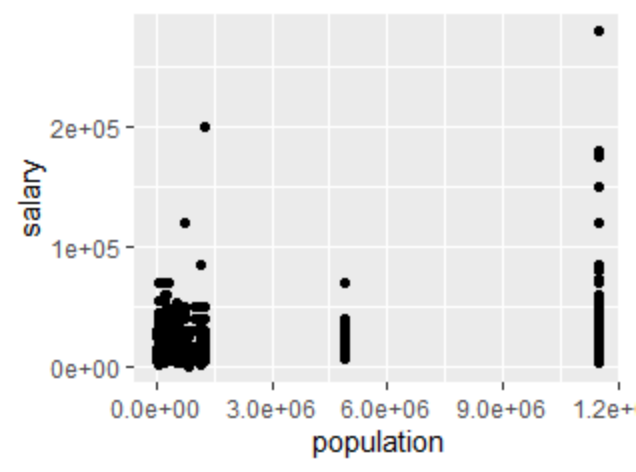
Теперь создаём дамми-переменные для категориальных признаков

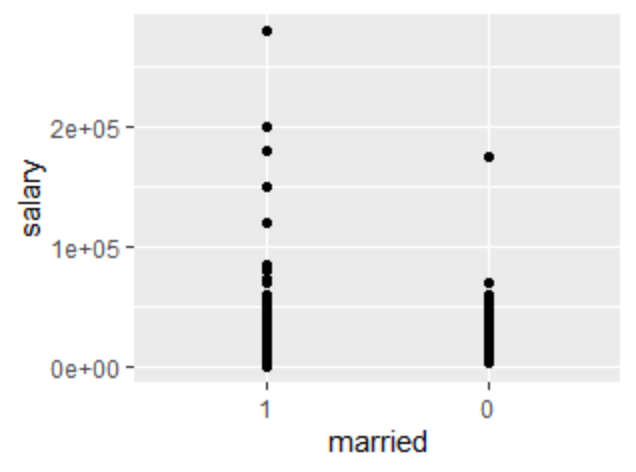
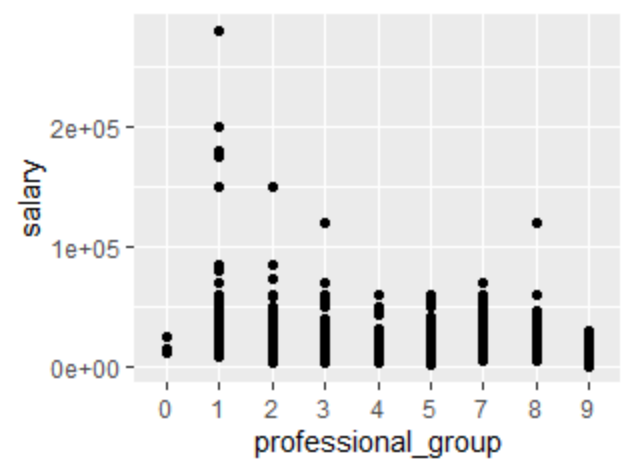
|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| название | Исп.пер. | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| status | status |  | 0 | 1 |  |  |  |  |  |  |  |
| Job\_sat | Job\_sat |  | 1 |  |  |  | 0 |  |  |  |  |
| sex | sex |  | 1 | 0 |  |  |  |  |  |  |  |
| Diplom0 | diplom |  | 1 | 1 | 1 | 0 | 0 | 0 |  |  |  |
| Diplom1 | diplom |  | 0 | 0 | 0 | 1 | 0 | 0 |  |  |  |
| Diplom2 | diplom |  | 0 | 0 | 0 | 0 | 1 | 0 |  |  |  |
| Diplom3 | diplom |  | 0 | 0 | 0 | 0 | 0 | 1 |  |  |  |
| married | married |  |  | 1 | 1 | 0 | 0 | 0 |  |  |  |
| Prof\_war | prof | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Prof\_rul | prof | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Prof\_spec | prof | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Prof\_ser | prof | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| Prof\_qual | pror | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| Prof\_nonqual | prof | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

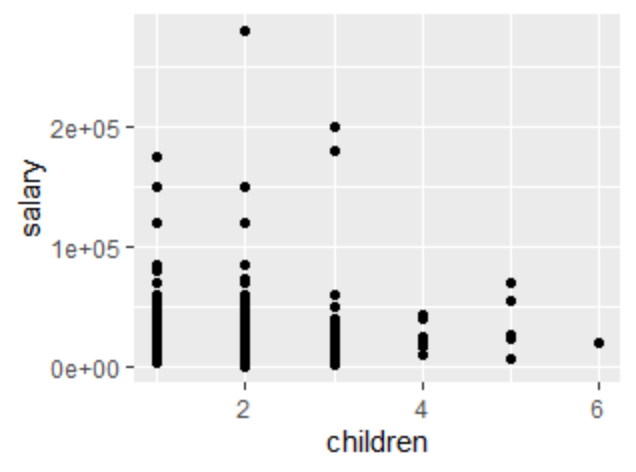
Посмотрим на зависимость целевой переменной от регрессоров иначе – с помощью диаграмм рассеяния:









Диаграммы рассеяния явно дают понять, что в данных присутствуют выбросы, которые портят статистику. Линейной сложно назвать даже зависимости без выбросов. Более того, у нас очень много категориальных переменных, что объясняет сложившуюся картину.

Немного познакомившись с данными, перейдём к анализу модели. Составляя пайплайн работы, мы приняли решение начать с простой модели, включающей в себя все выделенные переменные – *стартовой модели* – и постепенно преобразовывая её, прийти к модели значительно лучшего качества.

Call:

lm(formula = salary ~ sex + age + status + diplom\_1 + diplom\_2 +

diplom\_3 + population + married + prof\_war + prof\_qual +

prof\_service + prof\_spec + prof\_rul + children + job\_satisfaction,

data = df3)

Residuals:

Min 1Q Median 3Q Max

-39506 -8420 -1776 5829 214467

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.955e+04 1.151e+04 1.699 0.089851 .

sex0 -8.868e+03 1.738e+03 -5.101 4.45e-07 \*\*\*

age -2.957e+02 6.319e+01 -4.679 3.52e-06 \*\*\*

status1 -5.502e+02 1.752e+03 -0.314 0.753517

diplom\_10 -1.320e+03 2.542e+03 -0.519 0.603708

diplom\_20 3.681e+03 2.661e+03 1.384 0.166978

diplom\_30 9.509e+03 2.747e+03 3.462 0.000571 \*\*\*

population 1.889e-03 1.867e-04 10.116 < 2e-16 \*\*\*

married0 3.569e+03 1.978e+03 1.805 0.071616 .

prof\_war0 4.926e+03 1.133e+04 0.435 0.663967

prof\_qual1 5.906e+03 3.307e+03 1.786 0.074588 .

prof\_service1 1.080e+03 3.203e+03 0.337 0.736166

prof\_spec1 3.064e+03 3.182e+03 0.963 0.335865

prof\_rul1 2.130e+04 3.771e+03 5.649 2.43e-08 \*\*\*

children 1.800e+03 1.015e+03 1.774 0.076459 .

job\_satisfaction0 -4.039e+03 2.323e+03 -1.739 0.082568 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18700 on 641 degrees of freedom

Multiple R-squared: 0.3437, Adjusted R-squared: 0.3283

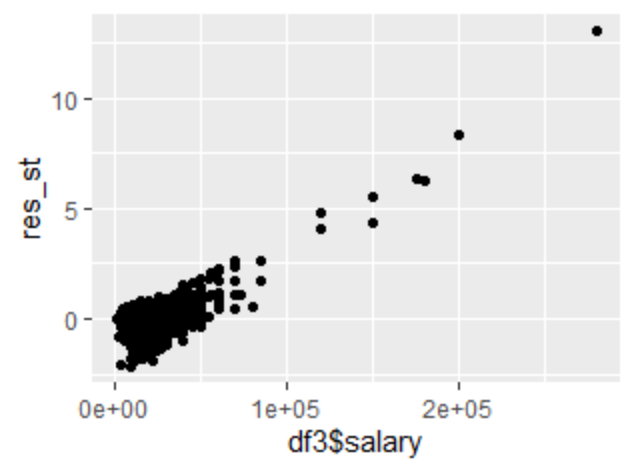
F-statistic: 22.38 on 15 and 641 DF, p-value: < 2.2e-16

В стартовой модели мы получили следующие значимые коэффициенты: sex, age, diplom\_3, population, married, prof\_qual, prof\_rul, children, job\_satisfaction. Каждый из них логично влияет на зарплату индивида в России. Мы не видим смысла более подробно анализировать стартовую модель, потому что далее будем её преобразовывать.

небольшой, однако нельзя сказать, что модель совсем плоха: ей объясняется почти 1/3 среднемесячной зарплаты, что уже достойно.

Начнём работать с моделью, а именно – решим вопрос с выбросами. Воспользуемся двумя методами (удаление выбросов и создание дамми-переменной для выбросов), сравним их и определим, какой лучше подходит для нашего случая.

Для начала с помощью стьюдентизированнных остатоков найдём выбросы в данных. Ниже изображен график стьюдентизированных остатков, по которому можно сделать вывод, что порог для определения выброса примерно 150 тысяч рублей (чуть меньше).



1. Посмотрим на модель без выбросов

Call:

lm(formula = salary\_f ~ sex\_f + age\_f + status\_f + diplom\_1\_f +

diplom\_2\_f + diplom\_3\_f + population\_f + married\_f + prof\_war\_f +

prof\_qual\_f + prof\_service\_f + prof\_spec\_f + prof\_rul\_f +

children\_f + job\_satisfaction\_f, data = df[-ind, ])

Residuals:

Min 1Q Median 3Q Max

-39546 -8392 -1776 5982 214592

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.991e+04 1.158e+04 1.720 0.085994 .

sex\_f0 -9.187e+03 1.768e+03 -5.196 2.76e-07 \*\*\*

age\_f -3.054e+02 6.438e+01 -4.743 2.61e-06 \*\*\*

status\_f1 -5.563e+02 1.785e+03 -0.312 0.755393

diplom\_1\_f0 -1.354e+03 2.580e+03 -0.525 0.599971

diplom\_2\_f0 3.869e+03 2.710e+03 1.428 0.153877

diplom\_3\_f0 9.605e+03 2.791e+03 3.442 0.000616 \*\*\*

population\_f 1.862e-03 1.893e-04 9.838 < 2e-16 \*\*\*

married\_f0 3.772e+03 2.011e+03 1.876 0.061179 .

prof\_war\_f0 5.086e+03 1.141e+04 0.446 0.655850

prof\_qual\_f1 5.965e+03 3.357e+03 1.777 0.076046 .

prof\_service\_f1 1.278e+03 3.253e+03 0.393 0.694548

prof\_spec\_f1 3.234e+03 3.222e+03 1.004 0.315880

prof\_rul\_f1 2.131e+04 3.813e+03 5.589 3.41e-08 \*\*\*

children\_f 1.824e+03 1.027e+03 1.776 0.076149 .

job\_satisfaction\_f0 -4.201e+03 2.370e+03 -1.772 0.076819 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18810 on 627 degrees of freedom

Multiple R-squared: 0.3424, Adjusted R-squared: 0.3267

F-statistic: 21.77 on 15 and 627 DF, p-value: < 2.2e-16

Как можно заметить, упал до 0.3267. Новая модель без выбросов стала только хуже.

1. Теперь попробуем создать дамми-переменную для выбросов.

Быть может, сформировав дамми-переменную для выбросов, мы улучшим качество модели. Всем индивидам со среднемесячной зарплатой больше 150 тыс.руб. была добавлена дамми-переменная special со значением = 1 (соответственно, 0 при зарплате меньше 150 тыс.руб.; порог определён на основе графика остатков выше).

Call:

lm(formula = salary ~ special + sex + age + status + diplom\_1 +

diplom\_2 + diplom\_3 + population + married + prof\_war + prof\_qual +

prof\_service + prof\_spec + prof\_rul + children + job\_satisfaction,

data = df3)

Residuals:

Min 1Q Median 3Q Max

-36273 -7841 -1931 5383 103668

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.129e+04 8.788e+03 2.423 0.015680 \*

special1 1.599e+05 7.464e+03 21.422 < 2e-16 \*\*\*

sex0 -7.534e+03 1.329e+03 -5.668 2.18e-08 \*\*\*

age -2.459e+02 4.832e+01 -5.090 4.72e-07 \*\*\*

status1 -5.902e+02 1.338e+03 -0.441 0.659216

diplom\_10 -9.709e+02 1.942e+03 -0.500 0.617209

diplom\_20 3.926e+03 2.032e+03 1.932 0.053813 .

diplom\_30 7.657e+03 2.099e+03 3.647 0.000287 \*\*\*

population 1.588e-03 1.433e-04 11.079 < 2e-16 \*\*\*

married0 1.894e+03 1.513e+03 1.252 0.211012

prof\_war0 2.795e+03 8.656e+03 0.323 0.746885

prof\_qual1 7.038e+03 2.526e+03 2.786 0.005494 \*\*

prof\_service1 1.815e+03 2.447e+03 0.742 0.458403

prof\_spec1 4.716e+03 2.431e+03 1.940 0.052857 .

prof\_rul1 1.375e+04 2.901e+03 4.739 2.65e-06 \*\*\*

children 4.612e+02 7.774e+02 0.593 0.553198

job\_satisfaction0 -4.034e+03 1.774e+03 -2.274 0.023287 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14280 on 640 degrees of freedom

Multiple R-squared: 0.6178, Adjusted R-squared: 0.6082

F-statistic: 64.65 on 16 and 640 DF, p-value: < 2.2e-16

Как мы видим, повысился до 0.6082. Заметим, что увеличился по сравнению с моделью без выбросов и стартовой моделью.

Теперь настало время выбрать наиболее подходящую функциональную форму модели. Создав логарифмированные переменные, оценим линейную в логарифмах и полулогарифмическую модели[[1]](#footnote-1).

Полулогарифмическая модель:

Call:

lm(formula = salary ~ l\_age + l\_population + l\_children + sex +

status + diplom\_1 + diplom\_2 + diplom\_3 + married + prof\_war +

prof\_qual + prof\_service + prof\_spec + prof\_rul + job\_satisfaction +

special, data = df3)

Residuals:

Min 1Q Median 3Q Max

-32884 -8409 -2203 5207 107228

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -15397.8 13575.1 -1.134 0.257107

l\_age -8914.1 2166.0 -4.115 4.37e-05 \*\*\*

l\_population 4693.2 481.2 9.753 < 2e-16 \*\*\*

l\_children 565.0 1459.4 0.387 0.698774

sex0 -7512.7 1358.0 -5.532 4.61e-08 \*\*\*

status1 7455.3 1915.0 3.893 0.000109 \*\*\*

diplom\_10 -1154.2 1985.1 -0.581 0.561135

diplom\_20 4224.3 2076.7 2.034 0.042348 \*

diplom\_30 6973.4 2144.9 3.251 0.001210 \*\*

married0 1471.3 1542.7 0.954 0.340580

prof\_war0 1290.3 8848.7 0.146 0.884114

prof\_qual1 7523.5 2581.8 2.914 0.003692 \*\*

prof\_service1 2453.7 2500.9 0.981 0.326896

prof\_spec1 5546.3 2481.1 2.235 0.025734 \*

prof\_rul1 14292.8 2961.3 4.826 1.74e-06 \*\*\*

job\_satisfaction0 -3756.6 1814.4 -2.070 0.038816 \*

special1 162219.2 7621.3 21.285 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14600 on 640 degrees of freedom

Multiple R-squared: 0.6005, Adjusted R-squared: 0.5905

F-statistic: 60.12 on 16 and 640 DF, p-value: < 2.2e-16

Линейная в логарифмах модель:

Call:

lm(formula = l\_salary ~ l\_age + l\_population + l\_children + sex +

status + diplom\_1 + diplom\_2 + diplom\_3 + married + prof\_war +

prof\_qual + prof\_service + prof\_spec + prof\_rul + job\_satisfaction +

special, data = df3)

Residuals:

Min 1Q Median 3Q Max

-2.07162 -0.35621 -0.02301 0.36152 1.51636

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 8.722718 0.499063 17.478 < 2e-16 \*\*\*

l\_age -0.416221 0.079629 -5.227 2.34e-07 \*\*\*

l\_population 0.183963 0.017690 10.399 < 2e-16 \*\*\*

l\_children 0.026222 0.053651 0.489 0.625187

sex0 -0.336319 0.049924 -6.737 3.62e-11 \*\*\*

status1 0.261206 0.070403 3.710 0.000225 \*\*\*

diplom\_10 -0.009414 0.072977 -0.129 0.897401

diplom\_20 0.165061 0.076346 2.162 0.030988 \*

diplom\_30 0.257698 0.078854 3.268 0.001141 \*\*

married0 0.100998 0.056715 1.781 0.075417 .

prof\_war0 -0.351396 0.325307 -1.080 0.280461

prof\_qual1 0.631010 0.094914 6.648 6.36e-11 \*\*\*

prof\_service1 0.282742 0.091941 3.075 0.002193 \*\*

prof\_spec1 0.497715 0.091212 5.457 6.94e-08 \*\*\*

prof\_rul1 0.839473 0.108869 7.711 4.80e-14 \*\*\*

job\_satisfaction0 -0.280708 0.066705 -4.208 2.94e-05 \*\*\*

special1 1.464417 0.280184 5.227 2.34e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

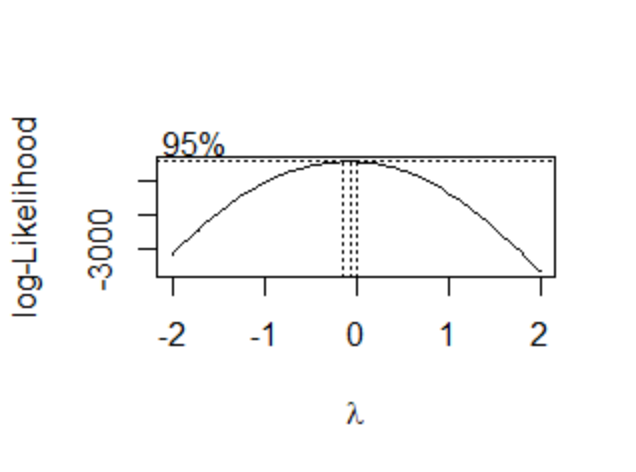
Residual standard error: 0.5369 on 640 degrees of freedom

Multiple R-squared: 0.4755, Adjusted R-squared: 0.4624

F-statistic: 36.27 on 16 and 640 DF, p-value: < 2.2e-16

Видно сразу, что полулогарифмическая лучше линейной в логарифмах по .

Однако нельзя сравнивать линейную модель с полулогарифмической или линейной в логарифмах по . Воспользуемся специальным тестом Бокса-Кокса:



По графику видно, что 0 попал в интервал, значит, линейная модель точно хуже, чем лог-модели. Между линейной в логарифмах и полулогарифмической мы выбираем лучшую модель на основе . Получается, что лучшая функциональная форма модели – это полулогарифмическая (model\_log), так как у неё больше.

Теперь проведём тест Рамсея на спецификацию модели, который помогает понять, есть ли у нас неучтённые переменные или же модель правильно специфирована.

RESET test

data: model\_log

RESET = 17.802, df1 = 2, df2 = 638, p-value = 2.997e-08

Нулевая гипотеза отвергается на уровне значимости в 5%, а значит можно сказать, что в модели присутствуют неучтённые переменные и модель неправильно специфирована. Попробуем это исправить, поэкспериментировав со степенями некоторых регрессоров, и посмотрим, как сильно поменяется статистика и p-value в reset-тесте.

Введём новые переменные:

l\_age2 <- log(df3$age)^2

l\_age3 <- log(df3$age)^3

l\_population2 <- log((df3$population)^2)

l\_population3 <- log(df3$population)^3

Аналогичным образом мы пробовали вводить children в степени (2 и 3 рассматривали, но это только ухудшало результат, если рассматривать модель без них).

Посмотрим на модель с новыми переменными и на reset-test:

Call:

lm(formula = salary ~ l\_age + age2 + age3 + l\_population + population2 +

population3 + l\_children + sex + status + diplom\_1 + diplom\_2 +

diplom\_3 + married + prof\_war + prof\_qual + prof\_service +

prof\_spec + prof\_rul + job\_satisfaction + special, data = df3)

Residuals:

Min 1Q Median 3Q Max

-35718 -7849 -1629 5299 101963

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.144e+05 6.171e+04 -1.854 0.064246 .

l\_age 4.069e+04 1.973e+04 2.062 0.039595 \*

age2 -2.676e+01 1.392e+01 -1.922 0.055014 .

age3 1.918e-01 1.261e-01 1.521 0.128799

l\_population 5.915e+02 1.120e+03 0.528 0.597632

population2 2.847e-10 2.563e-10 1.111 0.267066

population3 -1.452e-17 2.077e-17 -0.699 0.484707

l\_children 5.131e+02 1.428e+03 0.359 0.719554

sex0 -7.729e+03 1.331e+03 -5.806 1.01e-08 \*\*\*

status1 -4.217e+02 2.582e+03 -0.163 0.870322

diplom\_11 -9.800e+02 1.938e+03 -0.506 0.613307

diplom\_21 4.185e+03 2.033e+03 2.058 0.039966 \*

diplom\_31 7.927e+03 2.097e+03 3.779 0.000172 \*\*\*

married0 2.086e+03 1.514e+03 1.378 0.168772

prof\_war0 5.309e+03 8.659e+03 0.613 0.540052

prof\_qual1 6.230e+03 2.532e+03 2.461 0.014120 \*

prof\_service1 1.515e+03 2.441e+03 0.621 0.535118

prof\_spec1 3.953e+03 2.438e+03 1.621 0.105441

prof\_rul1 1.238e+04 2.926e+03 4.230 2.68e-05 \*\*\*

job\_satisfaction0 -4.202e+03 1.769e+03 -2.375 0.017852 \*

special1 1.597e+05 7.434e+03 21.481 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14220 on 636 degrees of freedom

Multiple R-squared: 0.6237, Adjusted R-squared: 0.6118

F-statistic: 52.7 on 20 and 636 DF, p-value: < 2.2e-16

RESET test

data: model\_log\_reset

RESET = 12.341, df1 = 2, df2 = 634, p-value = 5.52e-06

Как мы видим, добавление новых переменных значительно улучшило статистику reset-теста (уменьшило RESET и увеличило p-value). Более того, даже увеличился. Хотя нулевая гипотеза об отсутствии по-прежнему отвергается, можно сказать, что новые переменные сделали ситуацию чуть лучше. Кроме того, неудивительно, что гипотеза отвергается, потому что в модели очень много потенциальных объясняющих переменных и очень сложно учесть все факторы,влияющие на среднемесячную зарплату индивида.

Теперь проверим мультиколлинеарность, рассчитая VIFы для переменных.

Посчитаем VIFы для переменных:

vif\_age 🡨 1/(1-0.9905)= 105.2632

vif\_pop 🡨 1/(1-0.93) = 14.28571

vif\_child 🡨 1/(1-0.131) = 1.150748

Уже по первым двум коэффицентам понятно, что в модели присутствует мультиколлинеарность. Наличие мультиколлинеарности стала предполагаема, ещё когда мы боролись с неправильной спецификацией в модели. Теперь методом пошагового исключения переменных уберем незначимые коэффициенты:

1. уберём status
2. попробуем убрать l\_children
3. уберём diplom1
4. уберём prof\_war
5. population3
6. уберём prof\_service
7. уберём married
8. уберём age3

Теперь проверим модель на гетероскедастичность с помощью теста Бройша-Пагана:

studentized Breusch-Pagan test

data: model\_8

BP = 69.681, df = 12, p-value = 3.675e-10

Как можно заметить, гетероскедастичность есть, p-value слишком маленький, что говорит об отвержении нулевой гипотезы о гомоскедастичности.

Проверим данные тестом Голдфелда-Квандта на значимость гетероскедастичности.

Для начала отсортируем набор данных по признаку – в нашем случае логарифм возраста. Далее оценим параметры уравнения регрессии по первой и последней трети датасета, сохранив RSS. Посмотрим на результат:

Goldfeld-Quandt test

data: model\_8

GQ = 0.41338, df1 = 208, df2 = 207, p-value = 1

alternative hypothesis: variance increases from segment 1 to 2

Как показал тест Голдфелда-Квандта, гипотеза H0 не отвергается, гетероскедастичность данного вида признается незначимой.

Теперь предположим, что сортировать надо по признаку l\_population:

Goldfeld-Quandt test

data: model\_8

GQ = 2.8416, df1 = 208, df2 = 207, p-value = 9.968e-14

alternative hypothesis: variance increases from segment 1 to 2

К сожалению, в этом случае мы уже видим, что p-value очень маленькое, что не позволяет нам согласиться с гомоскедастичностью данных и отвергнуть нулевоую гипотезу.

Существует несколько способов борьбы с гетероскедастичностью. Один из них – это логарифмирование переменных в регрессии, что присутствует в нашей модели ещё с подбора наиболее оптимальной функциональной формы. Другим способом является взвешенный МНК. Попробуем реализовать его и посмотреть, как изменится ситуация с гетероскедастичностью.

Переходим ко взвешенному МНК. Взвешенный МНК требует знания структуры гетероскедастичности, что редко встречается на практике. Однако в тесте Голдфелда -Квандта мы видим, что гетероскедастичность выявляется при сортировке данных по l\_population.

Предположим, что в нашем случае Var(ϵi|lpopulation) = const⋅lpopulationi.

В R вектор весов weights должен быть обратно пропорционален дисперсиям, т.е.  wi ==

Call:

lm(formula = salary ~ l\_age + age2 + l\_population + population2 +

sex + diplom\_2 + diplom\_3 + prof\_qual + prof\_spec + prof\_rul +

job\_satisfaction + special, data = df3, weights = I(1/l\_population))

Weighted Residuals:

Min 1Q Median 3Q Max

-8857.0 -2088.6 -671.9 1349.8 25521.1

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.988e+04 2.197e+04 -1.360 0.17426

l\_age 1.185e+04 6.426e+03 1.844 0.06570 .

age2 -5.229e+00 1.589e+00 -3.291 0.00105 \*\*

l\_population 9.821e+02 4.557e+02 2.155 0.03150 \*

population2 1.077e-10 1.760e-11 6.117 1.65e-09 \*\*\*

sex0 -7.293e+03 1.235e+03 -5.906 5.67e-09 \*\*\*

diplom\_21 4.635e+03 1.431e+03 3.238 0.00126 \*\*

diplom\_31 7.917e+03 1.488e+03 5.322 1.42e-07 \*\*\*

prof\_qual1 5.354e+03 1.678e+03 3.190 0.00149 \*\*

prof\_spec1 3.163e+03 1.502e+03 2.106 0.03557 \*

prof\_rul1 1.111e+04 2.149e+03 5.173 3.08e-07 \*\*\*

job\_satisfaction0 -4.023e+03 1.666e+03 -2.415 0.01601 \*

special1 1.609e+05 7.729e+03 20.822 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3775 on 644 degrees of freedom

Multiple R-squared: 0.6034, Adjusted R-squared: 0.596

F-statistic: 81.65 on 12 and 644 DF, p-value: < 2.2e-16

Выше показан итог взвешенного МНК. Теперь снова посмотрим на один из тестов на гетероскедастичность:

Goldfeld-Quandt test

data: model\_mnk

GQ = 2.8416, df1 = 208, df2 = 207, p-value = 9.968e-14

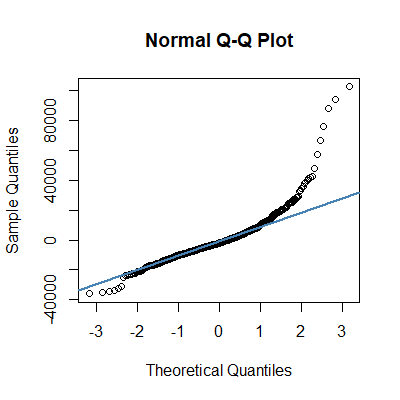
alternative hypothesis: variance increases from segment 1 to 2

Как мы видим, коррекция не принесла результатов, поэтому будем дальше пользоваться обычной моделью без весов, учитывая тот факт, что оценки не являются эффективными, но при этом их несмещенность сохраняется. Более того, надо помнить, что при использовании стандартных формул t-статистики не имеют t-распределения и доверительные интер-

валы и проверка гипотез по стандартным формулам даёт неверные результаты.

Наконец, проверим нормальность остатков. Ниже представлен график qqnorm, изобра-

жающий наше распределение и нормальное распределение:



Судя по графику, остатки не подчинены нормальному распределению. Чтобы подтвер-

дить данную гипотезу, мы воспользуемся тестом Шапиро-Уилка:

Shapiro-Wilk normality test

data: residuals

W = 0.8505, p-value < 2.2e-16

Как можно заметить, тест показал, что гипотеза H0 отвергается, а значит нельзя утверж-

дать о нормальности остатков. Ниже показана гистограмма распределения остатков регрессии.

Изображение выглядит как снимок экрана

Автоматически созданное описание

Таким образом, мы выявили, на наш взгляд, наиболее оптимальную модель для объяснения среднемесячной зарплаты индивида в 2012 году:

Call:

lm(formula = salary ~ l\_age + age2 + l\_population + population2 +

sex + diplom\_2 + diplom\_3 + prof\_qual + prof\_spec + prof\_rul +

job\_satisfaction + special, data = df3, weights = I(1/l\_population))

Weighted Residuals:

Min 1Q Median 3Q Max

-8857.0 -2088.6 -671.9 1349.8 25521.1

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.988e+04 2.197e+04 -1.360 0.17426

l\_age 1.185e+04 6.426e+03 1.844 0.06570 .

age2 -5.229e+00 1.589e+00 -3.291 0.00105 \*\*

l\_population 9.821e+02 4.557e+02 2.155 0.03150 \*

population2 1.077e-10 1.760e-11 6.117 1.65e-09 \*\*\*

sex0 -7.293e+03 1.235e+03 -5.906 5.67e-09 \*\*\*

diplom\_21 4.635e+03 1.431e+03 3.238 0.00126 \*\*

diplom\_31 7.917e+03 1.488e+03 5.322 1.42e-07 \*\*\*

prof\_qual1 5.354e+03 1.678e+03 3.190 0.00149 \*\*

prof\_spec1 3.163e+03 1.502e+03 2.106 0.03557 \*

prof\_rul1 1.111e+04 2.149e+03 5.173 3.08e-07 \*\*\*

job\_satisfaction0 -4.023e+03 1.666e+03 -2.415 0.01601 \*

special1 1.609e+05 7.729e+03 20.822 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3775 on 644 degrees of freedom

Multiple R-squared: 0.6034, Adjusted R-squared: 0.596

F-statistic: 81.65 on 12 and 644 DF, p-value: < 2.2e-16

мы получили достаточно большой, если учитывать, что это настоящие данные по России и для начала пришлось выбросить много пропусков. Почти 60% объясняется построенной моделью. Все коэффициенты значимы, правда, на разных уровнях значимости. Что касается интерпретации, то размер зарплаты возрастает на 1.111e+04, если человек работает крупным чиновником; на 5.354e+03, если индивид является квалифицированным рабочим и на 3.163e+03, если индивид специалист. Мужской пол отрицательно влияет на размер зарплаты, что в принципе неудивительно (если женщина, то зарплата меньше на 7.293e+03). Здесь надо быть осторожнее, потому что при кодировании мы делали так, чтобы женщина была 0, мужчина 1, но при этом у нас создалась переменная sex0, что значит, что теперь sex0 = 1 – это женищина, а sex0 = 0 – мужчина. Так же надо быть внимательными при job\_satisfaction0: равно 1, если не удовлетворен работой -> у этого человека зарплата будет меньше на -4.023e+03. Если у нас выброс, то зарплата больше на 1.609e+05.Образование тоже логично влияет: если у индивила высшее образование, то зарплата больше на 7.917e+03, а если на ступень ниже (законченное среднее специальное), то на 4.635e+03 зарплата больше, чем у людей с незакоченным средним образованием.

Теперь рассмотрим нелинейные зависимости:

Чем больше населенный пункт, тем, как правило, больше зарплата индивида. Это хорошо объясняет сложившиеся зависимости. Чем больше город, тем развитее его экономика и больше высокооплачиваемых мест работы.

Пик заработной платы приходится на 30-34 года, затем зарплата плавно снижается. Как можно заметить, работодатели больше всего ценят труд 30-34летних работников, у которых уже достаточно опыта и еще полно сил для работы. Людей младше оценивают меньше из-за отсутствия необходимого опыта, людям старше платят меньше, так как работодатель считает, что у них меньше сил, хуже способность быстро обучаться.

Теперь проверим гипотезы о линейных ограничениях. Мы попытаемся понять, влияет ли пол на размер заработка. Для этого мы воспользуемся тестом Вальда. Мы взяли две модели: одну с указанием пола, другую же без. Посмотрим на результаты теста:

Wald test

Model 1: salary ~ l\_age + age2 + l\_population + population2 + diplom\_2 +

diplom\_3 + prof\_qual + prof\_spec + prof\_rul + job\_satisfaction +

special

Model 2: salary ~ l\_age + age2 + l\_population + population2 + sex + diplom\_2 +

diplom\_3 + prof\_qual + prof\_spec + prof\_rul + job\_satisfaction +

special

Res.Df Df F Pr(>F)

1 645

2 644 1 33.572 1.075e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Дополнительно с ошибками,скорректированным под гетероскедастичность

Wald test

Model 1: salary ~ l\_age + age2 + l\_population + population2 + diplom\_2 +

diplom\_3 + prof\_qual + prof\_spec + prof\_rul + job\_satisfaction +

special

Model 2: salary ~ l\_age + age2 + l\_population + population2 + sex + diplom\_2 +

diplom\_3 + prof\_qual + prof\_spec + prof\_rul + job\_satisfaction +

special

Res.Df Df F Pr(>F)

1 645

2 644 1 28.058 1.619e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

P-value меньше уровней значимости, следовательно, гипотеза H0 (верна ограниченная модель без пола) отвергается. Необходимо учитывать пол как объясняющий фактор. Для полной уверенности воспользуемся тестом отношения правдоподобия:

Likelihood ratio test

Model 1: salary ~ l\_age + age2 + l\_population + population2 + diplom\_2 +

diplom\_3 + prof\_qual + prof\_spec + prof\_rul + job\_satisfaction +

special

Model 2: salary ~ l\_age + age2 + l\_population + population2 + sex + diplom\_2 +

diplom\_3 + prof\_qual + prof\_spec + prof\_rul + job\_satisfaction +

special

#Df LogLik Df Chisq Pr(>Chisq)

1 13 -7202.3

2 14 -7207.3 1 9.8804 0.001671 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Мы опять получили низкое P-value, поэтому гипотеза о том, что пол не влияет на среднемесячную зарплату, отвергается. Существует разница между размером оплаты труда и гендером.

**Бинарные модели**

В качестве зависимой переменной мы выбрали факт удовлетворения работой среди индивидов (job\_satisfaction). Оценим бинарные модели:

Логит. модель:

Call:

glm(formula = job\_satisfaction ~ salary + age + population +

children + sex + status + diplom\_1 + diplom\_2 + diplom\_3 +

married + prof\_war + prof\_qual + prof\_service + prof\_spec +

prof\_rul, family = binomial(link = "logit"), data = df3,

x = TRUE)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.2084 -0.5348 -0.3455 -0.1962 2.8957

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -1.385e+01 8.287e+02 -0.017 0.98667

salary -5.409e-05 1.699e-05 -3.184 0.00145 \*\*

age -1.313e-02 1.108e-02 -1.184 0.23631

population 1.214e-08 4.154e-08 0.292 0.77001

children -1.000e-01 1.901e-01 -0.526 0.59875

sex0 2.940e-01 3.226e-01 0.911 0.36210

status1 1.341e-01 2.866e-01 0.468 0.63992

diplom\_10 -2.406e-01 3.654e-01 -0.658 0.51026

diplom\_20 -4.424e-01 4.054e-01 -1.091 0.27514

diplom\_30 -3.588e-01 4.505e-01 -0.797 0.42567

married0 -2.497e-01 3.502e-01 -0.713 0.47591

prof\_war0 1.460e+01 8.287e+02 0.018 0.98594

prof\_qual1 -7.910e-01 4.524e-01 -1.748 0.08039 .

prof\_service1 -3.089e-01 3.851e-01 -0.802 0.42243

prof\_spec1 -1.973e+00 4.737e-01 -4.164 3.12e-05 \*\*\*

prof\_rul1 -1.225e+00 6.579e-01 -1.863 0.06249 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 498.47 on 656 degrees of freedom

Residual deviance: 418.83 on 641 degrees of freedom

AIC: 450.83

Number of Fisher Scoring iterations: 14

Таким образом, можно выделить следующие значимые переменные в логит-модели: salary, prof\_qual, prof\_spec, prof\_rul.

Оценим качество модели с помощью R^2 МакФаддена:

> loglik <- 0.5\*(-m\_logit$aic+2\*length(m\_logit$coefficients))

> prob0 <- glm(data=df3, job\_satisfaction~1, family = binomial(link='logit'))

> loglik0 <- 0.5\*(-prob0$aic+2\*length(prob0$coefficients))

> MFrsq <- 1-(loglik/loglik0)

> MFrsq

[1] 0.1597708

Качество получилось около 16%, что достаточно высоко для бинарных моделей.

Проинтерпретируем значимые коэффициенты. Коэффициент при salary равен -5.409e-05, что значит, что если среднемесячная зарплата индивида возрастёт на 1 ед., то логарифм отношения шансов того, что индивид удовлетворён своей работой к тому, что он не удовлетворён, упадёт на -5.409e-05\*100%. Теперь проинтерпретируем дискретные коэффициенты. Начнём с prof\_qual, который равен 1, если индивид выполняет квалифицированный труд на работе, и 0 в противном случае. При переходе от неквалифициованного труда к квалифицированному отношение шансов быть удовлетворённым своей работой падает в exp(7.910e-01) раз. Аналогично интерпретируются и prof\_spec, prof\_rul. Однако, такая интепретация сложна для понимания, поэтому зачастую оценивают предельные эффекты, когда говорят о бинарных моделях.

maBina(m\_logit, digits = 6)

effect error t.value p.value

(Intercept) -0.916546 51.867507 -0.017671 0.985907

salary -0.000004 0.000012 -0.307236 0.758763

age -0.000869 0.002911 -0.298463 0.765447

population 0.000000 0.000000 0.212485 0.831796

children -0.006620 0.024901 -0.265850 0.790440

sex0 0.019006 0.065218 0.291418 0.770826

status1 0.009054 0.035338 0.256215 0.797867

diplom\_11 -0.015175 0.054305 -0.279440 0.779997

diplom\_21 -0.026647 0.090499 -0.294450 0.768509

diplom\_31 -0.022921 0.079796 -0.287248 0.774016

married0 -0.015529 0.054895 -0.282888 0.777354

prof\_war0 0.075799 0.013458 5.632159 0.000000

prof\_qual1 -0.044032 0.147800 -0.297918 0.765862

prof\_service1 -0.018887 0.065869 -0.286737 0.774407

prof\_spec1 -0.121291 0.383588 -0.316201 0.751953

prof\_rul1 -0.055186 0.188104 -0.293380 0.769327

Выше представлены предельные эффекты в логит-модели для среднестатического индивида (индивид среднего пола, возраста, профессиональной группы и т.д.). Предельный эффект при age показывает насколько при увеличении возраста индивида на 1 год вероятность быть удовлетворённым работой снижается на 0.000869.Для дискретных переменных R автоматически счиает разницу вероятностей, а не производных. Если индивид занимается квалифицированным трудом, то вероятность того, что он удовлетворён своей работой в отличие от неквалифицированного труда, меньше на 0.04403, если специалист, то на 0.121291, а если работает в гос.структурах/управлении, то на 0.055. Похожим образом можно найти предельные эффекты для каждого индивда и только потом усреднить.

Логит-модель не многим отличается от пробит-модели, и только в силу масштабирования скрытой переменной коэффициенты первой отличаются от второй примерно в 1.6 раза.

Рассмотрим пробит-модель:

Call:

glm(formula = job\_satisfaction ~ salary + age + population +

children + sex + status + diplom\_1 + diplom\_2 + diplom\_3 +

married + prof\_war + prof\_qual + prof\_service + prof\_spec +

prof\_rul, family = binomial(link = "probit"), data = df3,

x = TRUE)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.1786 -0.5496 -0.3542 -0.1854 2.9187

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -4.108e+00 1.332e+02 -0.031 0.97539

salary -2.558e-05 8.213e-06 -3.115 0.00184 \*\*

age -7.738e-03 6.046e-03 -1.280 0.20058

population 7.690e-09 2.159e-08 0.356 0.72165

children -6.322e-02 1.010e-01 -0.626 0.53126

sex0 1.288e-01 1.730e-01 0.744 0.45671

status1 7.299e-02 1.574e-01 0.464 0.64275

diplom\_11 -1.137e-01 2.061e-01 -0.552 0.58114

diplom\_21 -2.208e-01 2.259e-01 -0.978 0.32819

diplom\_31 -2.371e-01 2.449e-01 -0.968 0.33307

married0 -1.724e-01 1.913e-01 -0.901 0.36747

prof\_war0 4.507e+00 1.332e+02 0.034 0.97300

prof\_qual1 -5.320e-01 2.590e-01 -2.054 0.03999 \*

prof\_service1 -2.067e-01 2.307e-01 -0.896 0.37020

prof\_spec1 -1.074e+00 2.560e-01 -4.195 2.73e-05 \*\*\*

prof\_rul1 -7.194e-01 3.458e-01 -2.081 0.03748 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 498.47 on 656 degrees of freedom

Residual deviance: 419.80 on 641 degrees of freedom

AIC: 451.8

Number of Fisher Scoring iterations: 13

Если индивид военный, то вероятность того, что он удовлетворён своей работой в отличие от неквалифицированного труда, больше на 0.075799.

По умолчанию пакет R не рассчитывает R^2 МакФалдена, поэтому рассчитаем его по определению

> #оценим качество

> loglik <- 0.5\*(-m\_probit$aic+2\*length(m\_probit$coefficients))

> prob0 <- glm(data=df3, job\_satisfaction~1, family = binomial(link='probit'))

> loglik0 <- 0.5\*(-prob0$aic+2\*length(prob0$coefficients))

> MFrsq <- 1-(loglik/loglik0)

> MFrsq

[1] 0.1578289

Качество получается больше 15%, что, на первый взгляд, не очень много, но когда речь идёт о бинарных моделях, это действительно достойный результат.

Значимые коэффициенты такие же, как и в логит-модели: salary, prof\_qual, prof\_spec, prof\_rul.

maBina(m\_probit)

effect error t.value p.value

(Intercept) -0.607 19.156 -0.032 0.975

salary 0.000 0.000 -1.109 0.268

age -0.001 0.001 -0.866 0.387

population 0.000 0.000 0.342 0.733

children -0.009 0.017 -0.551 0.582

sex0 0.019 0.030 0.631 0.528

status1 0.011 0.026 0.424 0.671

diplom\_11 -0.016 0.032 -0.512 0.609

diplom\_21 -0.030 0.039 -0.770 0.442

diplom\_31 -0.034 0.045 -0.755 0.450

married0 -0.024 0.032 -0.734 0.463

prof\_war0 0.082 0.014 6.087 0.000

prof\_qual1 -0.064 0.065 -0.983 0.326

prof\_service1 -0.028 0.038 -0.733 0.464

prof\_spec1 -0.143 0.126 -1.142 0.254

prof\_rul1 -0.071 0.074 -0.961 0.337

Если смотреть на значимость только предельных эффектов, а не коэффициентов в модели, то значим только prof\_war, то есть вероятность того, что военные удовлетворены своей работой больше неквалифицированных рабочих больше на 0.082

Выше представлены предельные эффекты в пробит-модели для среднестатического ндивида (индивид среднего пола, возраста, профессиональной группы и т.д.). Предельный эффект при age показывает насколько при увеличении возраста индивида на 1 год вероятность быть удовлетворённым работой снижается на 0.001. Для дискретных переменных R автоматически считает разницу вероятностей, а не производных. Если индивид занимается квалифицированным трудом, то вероятность того, что он удовлетворён своей работой в отличие от занятий неквалифицированным труда, меньше на 0.04403, если же он специалист, то вероятность быть удовлетворённы своей работой меньше, чем у неквалифицированного на 0.143, а если же он работает в гос. структурах / в управлении, то она меньше на 0.071.

1. Выбросы уже учтены как дамми-переменная [↑](#footnote-ref-1)