Задача 3

Необходимо проанализировать данные волны мониторинга экономического положения и здоровья населения РФ (данные обследования РМЭЗ НИУ ВШЭ).

Прежде чем начать выполнение задачи, были выбраны следующие столбцы:

1. Размер заработной платы – предсказываемая переменная.
2. Пол.
3. Семейное положение.
4. Образование.
5. Возраст.
6. Тип населенного пункта.
7. Часы работы в неделю.
8. Регион.
9. Удовлетворенность трудом.
10. Является ли иностранная фирма владельцем предприятия.
11. Имеет ли человек несколько работ.
12. Курит ли человек.

Далее были удалены строки, которые содержали в себе пустые значения, после чего началась подготовка выборки к работе. Так, для начала столбцы с действительными числами были отнормализованы – это столбец с зарплатой, возрастом, продолжительностью рабочей недели. Остальные же столбцы были приведены к бинарному или категориальному признаку:

1. Пол – бинарный признак.
2. Семейное положение – разделено на категориальный признак.
3. Образование – бинарный признак (наличие высшего образования).
4. Населенный пункт – бинарный признак (проживает ли человек в городе).
5. Регион – бинарный признак (проживает ли человек в одном из крупнейших регионов).
6. Удовлетворенность трудом – бинарный признак.
7. Является ли иностранная фирма владельцем предприятия – бинарный признак.
8. Имеет ли человек несколько работ – бинарный признак.
9. Курит ли человек – бинарный признак.
10. Постройте линейную регрессию зарплаты на все параметры, которые Вы выделили из данных мониторинга. Не забудьте оценить коэффициент вздутия дисперсии VIF.

Построим линейную регрессию salary ~ sex + wed1 + wed2 + wed3 + higher\_educ + age + status + dur + region + satisfy + foreign\_owner + second\_job + smokes

Таблица 3.1. Характеристики полученной линейной регрессии зарплаты.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Переменная | Estimate | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | -0.86107 | 0.07095 | -12.136 | 2e-16 | \*\*\* |
| sex | 0.47431 | 0.03248 | 14.603 | 2e-16 | \*\*\* |
| wed1 | 0.01475 | -0.07566 | -0.195 | 0.845482 |  |
| wed2 | 0.11767 | 0.06907 | 1.704 | 0.088515 | . |
| wed3 | 0.06779 | 0.07917 | 0.856 | 0.391924 |  |
| higher\_educ | 0.41977 | 0.03273 | 12.824 | 2e-16 | \*\*\* |
| age | -0.05684 | 0.01531 | -3.712 | 0.000208 | \*\*\* |
| status | 0.26169 | 0.03196 | 8.189 | 3.43e-16 | \*\*\* |
| dur | 0.13843 | 0.01421 | 9.739 | 2e-16 | \*\*\* |
| region | 0.57806 | 0.03318 | 17.422 | 2e-16 | \*\*\* |
| satisfy | 0.21698 | 0.02834 | 7.656 | 2.35e-14 | \*\*\* |
| foreign\_owner | 0.50696 | 0.07096 | 7.144 | 1.06e-12 | \*\*\* |
| second\_job | -0.07166 | 0.06627 | -1.081 | 0.279592 |  |
| smokes | 0.03496 | 0.03243 | 1.078 | 0.281011 |  |

слабый (см. Таблица 3.1), 24%, однако действительно имеются значимые регрессоры, которые можно объяснить. Проверим VIF для каждого регрессора.

Таблица 3.2. VIF полученной модели.

|  |  |
| --- | --- |
| Параметр \ Характеристики | VIF |
| sex | 1.404549 |
| wed1 | 5.666387 |
| wed2 | 6.052101 |
| wed3 | 2.876117 |
| higher\_educ | 1.094895 |
| age | 1.255599 |
| status | 1.068236 |
| dur | 1.081754 |
| region | 1.075931 |
| satisfy | 1.063886 |
| foreign\_owner | 1.014206 |
| second\_job | 1.008588 |
| smokes | 1.362337 |

Так как wed1 wed2 и wed3 имеют очень большой VIF (см. Таблица 3.2), а сами они не важны (низкий уровень значимости), то от них можно избавиться.

Таблица 3.3. Характеристики полученной модели после удаления wed1, wed2 и wed3.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Переменная | Estimate | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | -0.78582 | 0.03248 | -24.196 | 2e-16 | \*\*\* |
| sex | 0.48531 | 0.03160 | 15.357 | 2e-16 | \*\*\* |
| higher\_educ | 0.42763 | 0.03270 | 13.077 | 2e-16 | \*\*\* |
| age | -0.04380 | 0.01395 | -3.140 | 0.00017 | \*\* |
| status | 0.25938 | 0.03198 | 8.111 | 6.49e-16 | \*\*\* |
| dur | 0.13925 | 0.01423 | 9.758 | 2e-16 | \*\*\* |
| region | 0.57081 | 0.03318 | 17.203 | 2e-16 | \*\*\* |
| satisfy | 0.21437 | 0.02837 | 7.555 | 5.08e-14 | \*\*\* |
| foreign\_owner | 0.50685 | 0.07106 | 7.133 | 1.15e-12 | \*\*\* |
| second\_job | -0.07011 | 0.06637 | -1.056 | 0.2909 |  |
| smokes | 0.02842 | 0.03235 | 0.878 | 0.3797 |  |

новой модели практически не изменился (см. Таблица 3.3), но age стал менее значимым. Возможно, возраст респондента не сильно влияет на заработную плату.

1. Поэкспериментируйте с функциями вещественных параметров: используйте логарифм и степени (хотя бы от 0.1 до 2 с шагом 0.1).

Для нахождения наилучшей модели в среде R были использованы циклы. Были проверены следующие модели:

1. Степень age.
2. Степень dur.
3. Степень salary.
4. Степень age и dur.
5. Степень salary и dur.
6. Степень salary и age.
7. Все значения в степени.
8. Логарифм от age.
9. Логарифм от dur.
10. Логарифм от age и dur.
11. Логарифм от salary.

Наилучшей моделью оказалась та, которая имеет параметр и , 26%. Ввод еще одних логарифмов ни к чему не привел, а значит это и есть наилучшая модель.

Таблица 3.4. Характеристика новой модели со степенью 2 у age и dur.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Переменная | Estimate | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | -0.668916 | 0.034486 | -19.397 | 2e-16 | \*\*\* |
| sex | 0.506537 | 0.031378 | 16.143 | 2e-16 | \*\*\* |
| higher\_educ | 0.418646 | 0.032358 | 12.938 | 2e-16 | \*\*\* |
| age | -0.008808 | 0.014236 | -0.619 | 0.5362 |  |
|  | -0.119737 | 0.012373 | -9.677 | 2e-16 | \*\*\* |
| status | 0.264720 | 0.031648 | 8.364 | 2e-16 | \*\*\* |
| dur | 0.146109 | 0.017045 | 8.572 | 2e-16 | \*\*\* |
|  | -0.012027 | 0.006002 | -2.004 | 0.0452 | \* |
| region | 0.587792 | 0.032861 | 17.887 | 2e-16 | \*\*\* |
| satisfy | 0.235776 | 0.028144 | 8.378 | 2e-16 | \*\*\* |
| foreign\_owner | 0.485479 | 0.070311 | 6.905 | 5.76e-12 | \*\*\* |
| second\_job | -0.093516 | 0.065912 | -1.419 | 0.1560 |  |
| smokes | 0.007993 | 0.032061 | 0.249 | 0.8031 |  |

Проверим VIF полученной модели (см. Таблица 3.4).

Таблица 3.5. VIF новой модели.

|  |  |
| --- | --- |
| Параметр \ Характеристики | VIF |
| sex | 1.336076 |
| higher\_educ | 1.090584 |
| age | 1.106094 |
|  | 1.106598 |
| status | 1.067869 |
| dur | 1.585656 |
|  | 1.470079 |
| region | 1.075590 |
| satisfy | 1.069220 |
| foreign\_owner | 1.014818 |
| second\_job | 1.016841 |
| smokes | 1.357221 |

Хоть VIF у всех параметров мал (см. Таблица 3.5), но можно избавиться от малозначимых и посмотреть, что будет. Тогда избавимся от age и .

Таблица 3.6. Характеристика новой модели после удаления age и .

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Переменная | Estimate | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | -0.68 | 0.034374 | -20.176 | 2e-16 | \*\*\* |
| sex | 0.50837 | 0.03133 | 16.224 | 2e-16 | \*\*\* |
| higher\_educ | 0.41656 | 0.03232 | 12.888 | 2e-16 | \*\*\* |
|  | -0.12272 | 0.01199 | -10.235 | 2e-16 | \*\*\* |
| status | 0.26646 | 0.03164 | 8.420 | 2e-16 | \*\*\* |
| dur | 0.12740 | 0.01412 | 9.021 | 2e-16 | \*\*\* |
| region | 0.58700 | 0.03284 | 17.874 | 2e-16 | \*\*\* |
| satisfy | 0.23557 | 0.02814 | 8.371 | 2e-16 | \*\*\* |
| foreign\_owner | 0.48884 | 0.07027 | 6.957 | 4e-12 | \*\*\* |
| second\_job | -0.10490 | 0.06571 | -1.596 | 0.110 |  |
| smokes | 0.01238 | 0.03183 | 0.389 | 0.697 |  |

практическине изменился, а значит модель (см. Таблица 3.6) можно оставить в данном виде.

1. Сделайте вывод о том, какие индивиды получают наибольшую зарплату.

Судя по обученным коэффициентам модели, можно сделать следующие выводы:

1. Пол, к сожалению, влияет на зарплату опрошенных сотрудников. Мужчина получает больше, чем женщина.
2. Наличие высшего образования определенно влияет на зарплату.
3. Возраст отрицательно влияет на зарплату - чем старше или моложе человек относительно среднего (для предоставленных данных) возраста, тем больше шанс, что платить будут меньше.
4. Статус населенного пункта сказывается на зарплату.
5. Чем больше человек работает, тем больше (но не сильно больше) он получит денег
6. Регион определённо влияет на прибыль - в крупных регионах получают больше.
7. Если человек доволен работой, значит, скорее всего, он получает достаточную прибыль.
8. Иностранное предприятие принесет большую прибыль.
9. Если у человека несколько работ, значит, скорее всего, он получает недостаточную прибыль, из-за чего ему и нужно несколько работ.
10. Курение не сильно влияет на прибыль. Возможно, сигареты тогда стоили меньше, и не могли повлиять на бюджет.

Значит, следующий индивид может получить наибольшую зарплату:

* Имеет мужской пол.
* Имеет высшее образование.
* Среднего (из предоставленных данных) возраста.
* Из крупного региона.
* Работает на иностранное предприятие.

1. Оцените регрессии для подмножества индивидов, указанных в варианте.

Вариант 18, номер волны выборки РМЭЗ – 14.

Подмножество I - женщины, не замужем.

Таблица 3.7. Характеристика модели, построенная на подмножестве I.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Переменная | Estimate | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | -0.58028 | 0.04981 | -11.650 | 2e-16 | \*\*\* |
| higher\_educ | 0.39864 | 0.04805 | 8.296 | 3.39e-16 | \*\*\* |
|  | -0.08970 | 0.01719 | -5.217 | 2.20e-07 | \*\*\* |
| status | 0.15423 | 0.04897 | 3.149 | 0.00168 | \*\* |
| dur | 0.07373 | 0.02285 | 3.227 | 0.00129 | \*\* |
| region | 0.49552 | 0.04821 | 10.279 | 2e-16 | \*\*\* |
| satisfy | 0.10850 | 0.04257 | 2.549 | 0.01095 | \* |
| foreign\_owner | 0.80938 | 0.11880 | 6.813 | 1.64e-11 | \*\*\* |
| second\_job | -0.10688 | 0.09113 | -1.173 | 0.24115 |  |
| smokes | 0.06116 | 0.04769 | 1.283 | 0.19991 |  |

не сильно изменился (см. Таблица 3.7), но p-значение значительно ухудшилось - status, durability и satisfy объясняют хуже. Наибольшую зарплату получат женщины, которые имеют высшее образование, из крупного региона и работают на заграничное предприятие.

Подмножество II - женщины, живущие в городе, разведённые.

Таблица 3.8. Характеристика модели, построенная на подмножестве II.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Переменная | Estimate | Std. Error | t value | Pr(>|t|) | Уровень значимости |
| (Intercept) | -0.44978 | 0.07382 | -6.093 | 4.16e-09 | \*\*\* |
| higher\_educ | 0.24607 | 0.09556 | 2.575 | 0.0106 | \* |
|  | -0.10810 | 0.04410 | 2.451 | -0.0149 | \* |
| dur | 0.04055 | 0.05151 | 0.787 | 0.4319 |  |
| region | 0.53363 | 0.09858 | 5.413 | 1.45e-07 | \*\*\* |
| satisfy | 0.22352 | 0.08935 | 2.502 | 0.0130 | \* |
| foreign\_owner | 1.70699 | 0.26484 | 6.445 | 5.89e-10 | \*\*\* |
| second\_job | 0.11733 | 0.16337 | 0.718 | 0.4733 |  |
| smokes | 0.03601 | 0.09335 | 0.386 | 0.7000 |  |

стал лишь немного лучше, но p-значение осталось хорошим только у region и foreign owner (см. Таблица 3.8), из-за чего нельзя сделать адекватных выводов из полученной модели.

Вывод

Данные из 14 выборки были приведены к «хорошему» виду – строки с пустыми значениями были удалены, столбцы с действительными числами были отнормализованы, а остальные столбцы были приведены к бинарному или категориальному признаку. Далее, с помощью выбранных столбцов, которые содержат в себе информацию о поле, семейном положении, образовании, возрасте, типе населенного пункта, часов работы, регионе, удовлетворенности трудом, владельце предприятия, информации о имеет ли человек несколько работ и курит ли он была построена модель, которая предсказывает зарплату респондента (см. Таблица 3.1), однако данная модель имеет малый 24%, а некоторые регрессоры имеют большой VIF. Тогда с помощью вводов нелинейных переменных и с помощью избавления от переменных с большим VIF была получена новая модель (см. Таблица 3.8), которая имеет больший , а ее регрессоры можно адекватно описать. Из полученной модели следует, что в 2004 году наибольшую прибыль мог бы получить мужчина с высшим образованием, который имеет средний возраст и проживает в крупном регионе, а сам он работает на иностранное предприятие.

Затем с помощью тех же данных из 14 волны были выделены 2 подвыборки:

1. женщины, не замужем (см. Таблица 3.7);
2. женщины, живущие в городе, разведённые (см. Таблица 3.8):

Выборку I можно адекватно описать, так как большинство регрессоров являются значимыми. По полученной модели видно, что наибольшую зарплату получат женщины, которые имеют высшее образование, проживающие в крупном регионе и работают на заграничное предприятие. У выборки I имеется всего два значимых регрессора, а значит эту модель тяжело адекватно описать.

При этом стоит заметить, что во всех полученных моделях наиболее «сильными» критериями получения большей зарплаты, помимо пола, являются заграничный работодатель и регион, в котором проживает человек.

Приложение

install.packages("devtools")

devtools::install\_github("bdemeshev/rlms")

library("lmtest")

library("rlms")

library("dplyr")

library("GGally")

library("car")

library("sandwich")

"""

jj13.2 - зп ! check

jh5 - пол ! check

j\_marst - семейное положение ! check

j\_educ - ВО ! check

j\_age - возраст ! check

status - тип населенного пункта ! check

jj6.2 - Часов работы ! check

psu - регион check

jj1.1.1 - удовлетворенность трудом check

jj24 - являются ли иностранные фирмы владельцем check

jj32.1 - несколько работ check

jm71 - курит ли человек check

"""

o\_data <- rlms\_read("C:/Users/ruthh/Documents/workspace/data/r14iall\_32.sav")

glimpse(o\_data)

data = select(o\_data, jj13.2, jh5, j\_educ, j\_age, status, jj6.2, psu, jj1.1.2, jj24, jj32, jm71, j\_marst)

data = na.omit(data)

data

#зарплата c элементами нормализации

sal = as.numeric(data$jj13.2)

sal1 = as.character(data$jj13.2)

sal2 = lapply(sal1, as.integer)

sal = as.numeric(unlist(sal2))

mean(sal)

data["salary"] = (sal - mean(sal)) / sqrt(var(sal))

data["salary"] = data["salary"]\*1.034

#пол

data["sex"]=data$jh5

data$sex[which(data$sex!=1)] <- 0

data$sex[which(data$sex==1)] <- 1

#семейное положение

data["wed"]= data$j\_marst

data$wed1 = 0

data$wed1[which(data$wed==1)] <- 1

data$wed1[which(data$wed==3)] <- 1

data$wed2 = 0

data$wed2[which(data$wed==2)] <- 1

data["wed3"]=data$j\_marst

data$wed3 = 0

data$wed3[which(data$wed==4)] <- 1

data["wed4"]=data$j\_marst

data$wed4 = 0

data$wed4[which(data$wed==5)] <- 1

#образование

data["h\_educ"] = data$j\_educ

data["higher\_educ"] = 0

data$higher\_educ[which(data$h\_educ==21)] <- 1

data$higher\_educ[which(data$h\_educ==22)] <- 1

data$higher\_educ[which(data$h\_educ==23)] <- 1

#возраст c элементами нормализации

age = as.character(data$j\_age)

age = lapply(age, as.integer)

age = as.numeric(unlist(age))

data["age"]= (age - mean(age)) / sqrt(var(age))

data["age"]

#населенный пункт

data["status1"]=data$status

data["status2"] = 0

data$status2[which(data$status1==1)] <- 1

data$status2[which(data$status1==2)] <- 1

data["status"] = data$status2

#продолжительность рабочей недели

dur1 = as.character(data$jj6.2)

dur2 = lapply(dur1, as.integer)

dur3 = as.numeric(unlist(dur2))

data["dur"] = (dur3 - mean(dur3)) / sqrt(var(dur3))

########## ДАЛЕЕ ИДУТ ВЫБРАННЫЕ СТОЛБЦЫ

#region

data["region"]=0

data$region[which(data$psu==1)] <- 1

data$region[which(data$psu==2)] <- 1

data$region[which(data$psu==3)] <- 1

data$region[which(data$psu==16)] <- 1

data$region[which(data$psu==32)] <- 1

#удовлетворенность

data["sat"]=data$jj1.1.2

data["satisfy"] = 0

data$satisfy[which(data$sat==1)] <- 1

data$satisfy[which(data$sat==2)] <- 1

#другое государство - совладелец предприятия

data["foreign"] = data$jj24

data["foreign\_owner"] = 0

data$foreign\_owner[which(data$foreign==1)] <- 1

#есть приработок, вторая, n-я работа

data["sj"] = data$jj32

data["second\_job"] = 0

data$second\_job[which(data$sj==1)] <- 1

#Курит ли человек

data["sp"] = data$jm71

data["smokes"] = 0

data$smokes[which(data$sp==1)] <- 1

data = na.omit(data)

data = select(data, salary, sex, wed1,wed2,wed3,wed4, higher\_educ, age, status, dur, region,

satisfy, foreign\_owner, second\_job, smokes)

#среднее, медиана и мода для каждого параметра

uniqv <- unique(data$salary)

u1 = uniqv[which.max(tabulate(match(data$salary, uniqv)))]

table\_salary = data.frame(Value = "salary", Mean = mean(data$salary,na.rm = TRUE), Median = median(data$salary,na.rm = TRUE), Mode = u1)

table\_salary

mean(data$sex)

mean(data$wed1)

mean(data$wed2)

mean(data$wed3)

mean(data$wed4)

mean(data$higher\_educ)

uniqv <- unique(data$age)

u1 = uniqv[which.max(tabulate(match(data$age, uniqv)))]

table\_age = data.frame(Value = "age", Mean = mean(data$age,na.rm = TRUE), Median = median(data$age,na.rm = TRUE), Mode = u1)

table\_age

mean(data$status)

uniqv <- unique(data$dur)

u1 = uniqv[which.max(tabulate(match(data$dur, uniqv)))]

table\_dur = data.frame(Value = "dur", Mean = mean(data$dur,na.rm = TRUE), Median = median(data$dur,na.rm = TRUE), Mode = u1)

table\_dur

mean(data$region)

mean(data$satisfy)

mean(data$foreign\_owner)

mean(data$second\_job)

mean(data$smokes)

#графики парных зависимостей

ggpairs(data)

#уровни факторных переменных

qplot(data = data, salary)

qplot(data = data, age)

qplot(data = data, dur)

#построение зависимостей

model = lm(data = data, salary ~ sex + wed1 + wed2 + wed3 + higher\_educ + age + status + dur + region +

satisfy + foreign\_owner + second\_job + smokes)

summary(model)

vif(model)

# wed1 wed2 и wed3 имеют очень большой vif, как и p-значение - они не важны.

model2 = lm(data = data, salary ~ sex + higher\_educ + age + status + dur + region +

satisfy + foreign\_owner + second\_job + smokes)

summary(model2)

vif(model2)

waldtest(model2)

waldtest(model)

waldtest(model2, model)

bptest(model2)

gqtest(model2, order.by = ~salary, data = data, fraction = 0.2)

#наблюдается гетероскедастичность по тесту Бройша-Паганаи и по Голдфельду-Квандту.

#оценка коэффициентов и их ошибки. Доверительные интервалы: коэффициент +- 1.96\*ошибка

conftable = coeftest(model1, vcov.=vcovHC(model2))

conftable[,1]

conftable[,2]

#Введем нелинейные регрессоры

# Найдем наилучшую модель с помощью перебора

best\_model = model2

summary(best\_model)

# Выделим следующие случаи:

# i. степень для age

for(i in seq(0.1, 2, 0.1)) {

#print(i)

i\_model = lm(data = data, salary ~ sex + higher\_educ + age + I(age^i) + status + dur + region +

satisfy + foreign\_owner + second\_job + smokes)

if(summary(i\_model)$r.squared > summary(best\_model)$r.squared){

best\_model = i\_model

print(i)

print("стал лучшим коэффициентом для age")

}

}

summary(best\_model)

# ii. степень для dur

for(i in seq(0.1, 2, 0.1)) {

#print(i)

i\_model = lm(data = data, salary ~ sex + higher\_educ + age + status + dur + I(dur^i) + region +

satisfy + foreign\_owner + second\_job + smokes)

if(summary(i\_model)$r.squared > summary(best\_model)$r.squared){

best\_model = i\_model

print(i)

print("стал лучшим коэффициентом для dur")

}

}

summary(best\_model)

# iii. степень для salary

for(i in seq(0.1, 2, 0.1)) {

#print(i)

i\_model = lm(data = data, salary^i ~ sex + higher\_educ + age + status + dur + region +

satisfy + foreign\_owner + second\_job + smokes)

if(summary(i\_model)$r.squared > summary(best\_model)$r.squared){

best\_model = i\_model

print(i)

print("стал лучшим коэффициентом для salary")

}

}

# iv. age и dur

for(i in seq(0.1, 2, 0.1)) {

#print(i)

i\_model = lm(data = data, salary ~ sex + higher\_educ + age + I(age^i) + status + dur + I(dur^i) + region +

satisfy + foreign\_owner + second\_job + smokes)

#print(summary(i\_model)$r.squared)

if(summary(i\_model)$r.squared > summary(best\_model)$r.squared){

best\_model = i\_model

print(i)

print("стал лучшим коэффициентом для age and dur")

}

}

summary(best\_model)

# v. salary и dur

for(i in seq(0.1, 2, 0.1)) {

#print(i)

i\_model = lm(data = data, salary^i ~ sex + higher\_educ + age + status + dur + I(dur^i) + region +

satisfy + foreign\_owner + second\_job + smokes)

#print(summary(i\_model)$r.squared)

if(summary(i\_model)$r.squared > summary(best\_model)$r.squared){

best\_model = i\_model

print(i)

print("стал лучшим коэффициентом для age and dur")

}

}

# vi. salary и age

for(i in seq(0.1, 2, 0.1)) {

#print(i)

i\_model = lm(data = data, salary^i ~ sex + higher\_educ + age + I(age^i) + status + dur + region +

satisfy + foreign\_owner + second\_job + smokes)

#print(summary(i\_model)$r.squared)

if(summary(i\_model)$r.squared > summary(best\_model)$r.squared){

best\_model = i\_model

print(i)

print("стал лучшим коэффициентом для age and dur")

}

}

# vii. все

for(i in seq(0.1, 2, 0.1)) {

#print(i)

i\_model = lm(data = data, salary^i ~ sex + higher\_educ + age + I(age^i) + status + dur + I(age^i) + region +

satisfy + foreign\_owner + second\_job + smokes)

#print(summary(i\_model)$r.squared)

if(summary(i\_model)$r.squared > summary(best\_model)$r.squared){

best\_model = i\_model

print(i)

print("стал лучшим коэффициентом для age and dur")

}

}

# Проверим то же для логарифма

# i. логарифм для age

for(i in seq(0.1, 2, 0.1)) {

#print(i)

i\_model = lm(data = data, salary ~ sex + higher\_educ + age + I(log(i, base=age)) + status + dur + region +

satisfy + foreign\_owner + second\_job + smokes)

if(summary(i\_model)$r.squared > summary(best\_model)$r.squared){

best\_model = i\_model

print(i)

print("стал лучшим коэффициентом для age")

}

}

summary(best\_model)

# ii. степень для dur

for(i in seq(0.1, 2, 0.1)) {

#print(i)

i\_model = lm(data = data, salary ~ sex + higher\_educ + age + status + dur + I(log(i, dur)) + region +

satisfy + foreign\_owner + second\_job + smokes)

if(summary(i\_model)$r.squared > summary(best\_model)$r.squared){

best\_model = i\_model

print(i)

print("стал лучшим коэффициентом для dur")

}

}

summary(best\_model)

# iii. age и dur

for(i in seq(0.1, 2, 0.1)) {

#print(i)

i\_model = lm(data = data, salary ~ sex + higher\_educ + age + I(log(i, age)) + status + dur + I(log(i, dur)) + region +

satisfy + foreign\_owner + second\_job + smokes)

#print(summary(i\_model)$r.squared)

if(summary(i\_model)$r.squared > summary(best\_model)$r.squared){

best\_model = i\_model

print(i)

print("стал лучшим коэффициентом для age and dur")

}

}

# Как итог, самой лучшей моделью стала модель с нелинейным коэффициентом age^2

# Попробуем дополнительно ввести логарифмы

summary(best\_model)

for(i in seq(0.1, 2, 0.1)) {

#print(i)

i\_model = lm(data = data, salary ~ sex + higher\_educ + age + I(age^i) + status + dur + region +

satisfy + foreign\_owner + second\_job + smokes)

if(summary(i\_model)$r.squared > summary(best\_model)$r.squared){

best\_model = i\_model

print(i)

print("стал лучшим коэффициентом для log age")

}

}

for(i in seq(0.1, 2, 0.1)) {

#print(i)

i\_model = lm(data = data, salary ~ sex + higher\_educ + age + I(age^i) + status + dur + I(log(i, dur)) + region +

satisfy + foreign\_owner + second\_job + smokes)

if(summary(i\_model)$r.squared > summary(best\_model)$r.squared){

best\_model = i\_model

print(i)

print("стал лучшим коэффициентом для age")

}

}

model3 = lm(data = data, salary ~ sex + higher\_educ + age + I(age^2) + status + dur + I(dur^2) + region +

satisfy + foreign\_owner + second\_job + smokes)

summary(model3)

vif(model3)

# У age крайне малый p. Посмотрим что будет, если удалить его

model3 = lm(data = data, salary ~ sex + higher\_educ + I(age^2) + status + dur + I(dur^2) + region +

satisfy + foreign\_owner + second\_job + smokes)

# Сделаем то же самое для dur^2

model3 = lm(data = data, salary ~ sex + higher\_educ + I(age^2) + status + dur + region +

satisfy + foreign\_owner + second\_job + smokes)

summary(model3)

vif(model3)

# R^2 изменился не критично, а вздутие уменьшилось. Можно оставить в таком виде

# Судя по обученным коэффициентам модели, можно сделать следующие выводы:

# 1. Пол, к сожалению, влияет на зарплату опрошенных сотрудников. Мужчина получает больше, чем женщина.

# 2. Наличие высшего образования влияет на зарплату в некоторой мере.

# 3. Возраст отрицательно влияет на зп - чем старше или моложе человек относительно среднего (для предоставленных данных) возраста, тем больше шанс, что платить будут меньше

# 4. Статус населенного пункта сказывается на зп

# 5. Чем больше человек работает, тем больше (но не сильно больше) он получит денег

# 6. Регион определеленно влияет на прибыль - в крупных регионах получают больше

# 7. Если человек доволен работой, значит, скорее всего, он получает достатучную прибыль

# 8. Иностранное предприятие принесет большую прибыль

# 9. Если у человека несколько работ, значит, скорее всего, он получает недостатучную прибыль, из-за чего ему и нужно несколько работ

# 10. Курение не зависит на прибыль. Возможно, сигареты тогда стоили меньше, и не могли повлиять на бюджет

# Значит, индивид, который:

# Имеет мужской пол

# Имеет высшее образование

# Среднего (из предоставленных данных) возраста

# Из крупного региона

# Работающий на иностранное предприятие

# Получит наибольшую прибыль.

data

# Оценим полученную модель на выборке Разведённыъ незамужних женщин (), которые живущит в городе.

data <-filter(data, sex==0)

data

data\_a <-filter(data, data$wed1==1 | data$wed3==1 | data$wed4==1)

data\_a

data\_b <- filter(data, data$status==1 & data$wed3==1)

data\_b

model\_a = lm(data = data\_a, salary ~ higher\_educ + I(age^2) + status + dur + region +

satisfy + foreign\_owner + second\_job + smokes)

summary(model\_a)

# R^2 не сильно изменился, но p-значение значительно ухудшилось - status, durability и satisfy объясняют хуже

# наибольшую зп получат женщины, которые имеют высшее образование, из крупного региона, работают на заграничное предприятие.

model\_b = lm(data = data\_b, salary ~ higher\_educ + I(age^2) + dur + region +

satisfy + foreign\_owner + second\_job + smokes)

summary(model\_b)

# R^2 стал лучше, но p-значение осталось хорошим только у region и foreign owner

# наибольшую зп получат женщины, которые проживают в крупном регионе и работают на иностранное предприятие.