Балашов Д. С. КМБО-03-20(Вариант 1)

Задача №3

Необходимо загрузить данные из указанного набора и произвести следующие действия.

Набор данных: Данные обследования РМЭЗ НИУ ВШЭ

Объясняемая переменная: *salary*

Регрессоры: *age, sex, higher\_educ, status2, dur, wed1,wed2,wed3.*

***№1***. Построить линейную регрессию зарплаты на все параметры, и оценить коэффициент вздутия дисперсии VIF.

Построим линейную регрессию зарплаты(*salary)* на все параметры, предварительно исключив из рассмотрения все строки с отсутствующими значениями(NA). После этого также исключим из рассмотрения все регрессоры с отсутствующими значениями, и все незначительные регрессоры (*см. Таблица 1*).

Построим линейную регрессию зарплаты(*salary)* на все оставшиеся параметры (*age, sex, dur).* R2 = 0.06639. Значение p-статистики хорошее у всех переменных(3 звезды, *см. Таблица 2*). VIF = 1 / (1- R2) = 1.004. VIF << 5, следовательно линейной зависимости между переменными нет.

***№2.***Введём в модель логарифмы и степени переменных:

***№2.1)*** Введём в модель логарифмы регрессоров, сравним модели и выберем наилучшую.

Рассмотрим модель (*salary* ~ *age*, *sex , dur, log(age), log(dur)*)*.*  R2 = 0.09061, следовательно, показатель увеличился на 0.024, по сравнению с моделью: "*salary* ~ *age*, *sex*, *dur* ". Значение p-статистики хорошее только у переменных *log(age)* и *sex* (3 звезды, *см. Таблица 3*).

Попробуем улучшить модель, убирая из неё параметры с наибольшими коэффициентами в VIF.

*Таблица 4*. Проверка на линейную зависимость между регрессорами *(age*, *sex , dur, log(age), log(dur)*) с помощью команды VIF.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | *age* | *sex* | *dur* | *log(age)* | *log(dur)* |
| VIF | 3.79 | 1.02 | 4.14 | 3.79 | 4.16 |

Рассмотрим модель (*salary* ~ *age*, *sex, dur, log(age)*)*.* R2 =  0.06722, следовательно, показатель увеличился на 0.008, по сравнению с моделью: "*salary* ~ *age*, *sex*, *dur*". Значение p-статистики очень хорошая(3 звезды, см. *Таблица 5*) у всех переменных, кроме *log(age)*и "Свободного Коэффициента".

*Таблица 6*. Проверка на линейную зависимость между регрессорами(*age*, *sex , dur, log(age)*)с помощью команды VIF.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Параметр \ Характеристики | *age* | *sex* | *dur* | *log(age)* |
| VIF | 3.64 | 1.04 | 1.05 | 3.61 |

Рассмотрим модель (*salary* ~ *sex , dur, log(age)*). R2 =  0.05907, следовательно, показатель уменьшился на 0.007, по сравнению с моделью: "*salary* ~ *age*, *sex*, *dur*". Значение p-статистики очень хорошее (3 звезды, см. *Таблица 7*) у всех переменных.

***№2.2)*** Введём в модель степени регрессоров, сравним модели и выберем наилучшую.

Рассмотрим модель (*salary* ~ *age*, *sex , dur, I(age0.1), I(dur0.1)*). R2 =  0.09058 - следовательно, показатель увеличился на 0.024, по сравнению с моделью: "*salary* ~ *age*, *sex*, *dur*". Значение p-статистики очень хорошее (3 звезды, см. *Таблица 8*) только у переменных *age* и *sex*.

*Таблица 9*. Проверка на линейную зависимость между регрессорами *(age*, *sex , dur, I(age0.1), I(dur0.1)*) с помощью команды VIF.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | *age* | *sex* | *dur* | *I(age0.1)* | *I(dur0.1)* |
| VIF | 4.835 | 1.02 | 5.01 | 4.829 | 5.03 |

Попробуем улучшить модель, убирая из неё параметры с наибольшими коэффициентами в VIF.

Рассмотрим модель (*salary* ~ *age*, *sex , dur, I(age0.1)*). R2 =  0.09058 - следовательно, показатель увеличился на 0.024, по сравнению с моделью: "*salary* ~ *age*, *sex*, *dur*". Значение p-статистики очень хорошее (3 звезды) только у переменных *age* и *sex*.

*Таблица 10*. Проверка на линейную зависимость между регрессорами *(age*, *sex , dur, I(age0.1)*) с помощью команды VIF.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Параметр \ Характеристики | *age* | *sex* | *dur* | *I(age0.1)* |
| VIF | 4.6 | 4.58 | 1.24 | 1.05 |

Рассмотрим модель (*salary* ~ *sex , dur, I(age0.1)*). R2 =  0.06045 - следовательно, показатель увеличился на 0.0059, по сравнению с моделью: "*salary* ~ *age*, *sex*, *dur*". Значение p-статистики очень хорошее (3 звезды, см. *Таблица 11*) у всех переменных.

Попробуем улучшить модель, повышая степень (до степени "2", с шагом 0.1):

При решении этой задачи были проверены модели:

1. *salary* ~ *sex , dur, I(age0.2)*
2. *salary* ~ *sex , dur, I(age0.3)*
3. *salary* ~ *sex , dur, I(age0.4)*
4. *salary* ~ *sex , dur, I(age0.5)*

Поскольку при повышении степени от "0.1" до "0.5"(см. *Таблица 12*) R2 возрастает, то попробуем взять сразу степень равную "0.9":

1. *salary* ~ *sex , dur, I(age0.9)*
2. *salary* ~ *sex , dur, I(age1.1)*
3. *salary* ~ *sex , dur, I(age1.2)*
4. *salary* ~ *sex , dur, I(age1.3)*
5. *salary* ~ *sex , dur, I(age1.4)*

Поскольку при степени "1.4" значение R2 меньше, чем при степени "1.3"(см. *Таблица 12*), то попробуем сразу взять степень равную "1.9":

1. *salary* ~ *sex , dur, I(age1.9)*
2. *salary* ~ *sex , dur, I(age2)*

Построим сравнительную таблицу для всех проверенных моделей:

*Таблица 12*.Сравнение уровня значимости и значения R2 для одиннадцати рассматриваемых моделей.

|  |  |  |  |
| --- | --- | --- | --- |
| Рассматриваемая модель | Уровень значимости | Значение R2 | Насколько изменилось значение R2 по сравнению с моделью: "salary ~ age, sex, dur" |
| *salary* ~ *sex , dur, I(age0.2)* | Значение p-статистики очень хорошее (\*\*\*) у всех переменных | 0.06174 | уменьшился на 0.005 |
| *salary* ~ *sex , dur, I(age0.3)* | Значение p-статистики очень хорошее (\*\*\*) у всех переменных | 0.06291 | уменьшился на 0.003 |
| *salary* ~ *sex , dur, I(age0.4)* | Значение p-статистики очень хорошее (\*\*\*) у всех переменных, кроме "Свободного Коэффициента" | 0.06394 | уменьшился на 0.002 |
| *salary* ~ *sex , dur, I(age0.5)* | Значение p-статистики очень хорошее (\*\*\*) у всех переменных, кроме "Свободного Коэффициента" | 0.06394 | уменьшился на 0.001 |
| *salary* ~ *sex , dur, I(age0.9)* | Значение p-статистики очень хорошее (\*\*\*) у всех переменных, кроме "Свободного Коэффициента" | 0.06704 | увеличился на 0.0007 |
| *salary* ~ *sex , dur, I(age1.1)* | Значение p-статистики очень хорошее (\*\*\*) у всех переменных, кроме "Свободного Коэффициента" | 0.06749 | увеличился на 0.0011 |
| *salary* ~ *sex , dur, I(age1.2)* | Значение p-статистики очень хорошее (\*\*\*) у всех переменных, кроме "Свободного Коэффициента" | 0.06758 | увеличился на 0.00119 |
| *salary* ~ *sex , dur, I(age1.3)* | Значение p-статистики очень хорошее (\*\*\*) у всех переменных, кроме "Свободного Коэффициента" | 0.06759 | увеличился на 0.0012 |
| *salary* ~ *sex , dur, I(age1.4)* | Значение p-статистики очень хорошее (\*\*\*) у всех переменных, кроме "Свободного Коэффициента" | 0.06754 | увеличился на 0.00115 |
| *salary* ~ *sex , dur, I(age1.9)* | Значение p-статистики очень хорошее (\*\*\*) у всех переменных | 0.06643 | увеличился на 0.00004 |
| *salary* ~ *sex , dur, I(age2)* | Значение p-статистики очень хорошее (\*\*\*) у всех переменных | 0.07877 | увеличился на 0.012 |

***№3.***Выберем наилучшую модель среди построенных.

***3.1)*** Вывод: Среди моделей с логарифмами лучшей можно считать: " *salary* ~ *sex , dur, log(age)*", однако эта модель хуже первоначальной: "*salary* ~ *age*, *sex*, *dur*" (т.к. p-статистика хорошая в обоих случаях, а R2 в модели с логарифмом меньше, чем в первоначальной модели).  
Следовательно, l*og()* не дал ощутимых результатов.

***3.2)*** Вывод: Среди моделей со степенями лучшая - это модель с квадратом, поскольку в остальных случаях у нас ниже R2 и больше параметров с плохим значением p-статистики.

***№4***.Сделать вывод о том, какие индивиды получают наибольшую зарплату.

Вывод: Наибольшую зарплату получают мужчины более молодого возраста, живущие в городе, с высшим образованием, с большой продолжительностью рабочей недели и никогда не состоящие в браке.

***№5***.Оценить регрессии для подмножества индивидов:

***5.1)***  Городские жители, состоящие в браке;

***5.2)***  Разведенные, без высшего образования.

***5.1)*** Найдем подмножество городских жителей, состоящих в браке:

data\_3 = subset(data\_2, status2 == 1)  
data\_3  
  
data\_4 = subset(data\_3, wed1 == 1)  
data\_4

Рассмотрим модель (*salary* ~ *sex , dur, I(age2)*). R2 = 0.1173, Значение p-статистики хорошее у всех переменных, кроме "Свободного Коэффициента» (см. *Таблица 13*). Следовательно, можно сделать вывод, что модель довольно хорошая.

Вывод: Наибольшую зарплату получают мужчины, с высшим образованием.

***5.2)*** Найдем подмножество разведенных, без высшего образования:

data\_5 = subset(data\_2, wed2=1)  
data\_5  
  
data\_6 = subset(data\_5, higher\_educ==0)  
data\_6

Рассмотрим модель (*salary* ~ *sex , dur, I(age2)*). R2 = 0.09543, Значение p-статистики хорошее у всех переменных (см. *Таблица 14*). Следовательно, можно сделать вывод, что модель довольно хорошая.

Вывод: Наибольшую зарплату получают мужчины более молодого возраста, с большой продолжительностью рабочей недели.

*Приложение1*

install.packages("devtools")  
devtools::install\_github("bdemeshev/rlms")  
  
library("lmtest")  
library("rlms")  
library("dplyr")  
library("GGally")  
library("car")  
library("sandwich")  
  
data <- rlms\_read("C:\\Users\\dimab\\Downloads\\r24i\_os26c.sav")  
glimpse(data)  
data\_1 = select(data, tj13.2, t\_age, th5, t\_educ, status, tj6.2, t\_marst)  
  
[#Исключим](https://vk.com/im?sel=241682732&st=%23%D0%98%D1%81%D0%BA%D0%BB%D1%8E%D1%87%D0%B8%D0%BC) из рассмотрения строки с отсутствующими значениями NA  
data\_1 = na.omit(data\_1)  
glimpse(data\_1)  
  
[#Зарплата](https://vk.com/im?sel=241682732&st=%23%D0%97%D0%B0%D1%80%D0%BF%D0%BB%D0%B0%D1%82%D0%B0) c элементами нормализации  
data\_1$tj13.2  
sal = as.numeric(data\_1$tj13.2)  
sal1 = as.character(data\_1$tj13.2)  
sal2 = lapply(sal1, as.integer)  
sal = as.numeric(unlist(sal2))  
mean(sal)  
data\_1["salary"] = (sal - mean(sal)) / sqrt(var(sal))  
data\_1["salary"]  
  
[#Возраст](https://vk.com/im?sel=241682732&st=%23%D0%92%D0%BE%D0%B7%D1%80%D0%B0%D1%81%D1%82) c элементами нормализации  
age1 = as.character(data\_1$t\_age)  
age2 = lapply(age1, as.integer)  
age3 = as.numeric(unlist(age2))  
data\_1["age"]= (age3 - mean(age3)) / sqrt(var(age3))  
data\_1["age"]  
  
[#Пол](https://vk.com/im?sel=241682732&st=%23%D0%9F%D0%BE%D0%BB)  
data\_1["sex"]=data\_1$th5  
[#data\_1](https://vk.com/im?sel=241682732&st=%23data_1)["sex"] = lapply(data\_1$th5, as.character)  
data\_1$sex[which(data\_1$sex!=1)] <- 0  
data\_1$sex[which(data\_1$sex==1)] <- 1  
data\_1$sex = as.numeric(data\_1$sex)

[#Образование](https://vk.com/im?sel=241682732&st=%23%D0%9E%D0%B1%D1%80%D0%B0%D0%B7%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5)  
data\_1["t\_educ"] = data\_1$t\_educ  
[#data\_1](https://vk.com/im?sel=241682732&st=%23data_1)["t\_educ"] = lappy(data\_1$t\_educ, as.character)  
data\_1["higher\_educ"] = data\_1$t\_educ  
data\_1["higher\_educ"] = 0 # Нет высшего образования  
data\_1$higher\_educ[which(data\_1$t\_educ==21)] <- 1  
data\_1$higher\_educ[which(data\_1$t\_educ==22)] <- 1  
data\_1$higher\_educ[which(data\_1$t\_educ==23)] <- 1  
  
[#Населенный](https://vk.com/im?sel=241682732&st=%23%D0%9D%D0%B0%D1%81%D0%B5%D0%BB%D0%B5%D0%BD%D0%BD%D1%8B%D0%B9) пункт  
data\_1["status1"]=data\_1$status  
[#data\_1](https://vk.com/im?sel=241682732&st=%23data_1)["status1"] = lapply(data\_1$status, as.character)  
data\_1["status2"] = 0  
data\_1$status2[which(data\_1$status1==1)] <- 1 # Область  
data\_1$status2[which(data\_1$status1==2)] <- 1 # Город  
data\_1$status2 = as.numeric(data\_1$status2)  
  
[#Продолжительность](https://vk.com/im?sel=241682732&st=%23%D0%9F%D1%80%D0%BE%D0%B4%D0%BE%D0%BB%D0%B6%D0%B8%D1%82%D0%B5%D0%BB%D1%8C%D0%BD%D0%BE%D1%81%D1%82%D1%8C) рабочей недели  
data\_1["dur123"]=data\_1$tj6.2  
[#data\_1](https://vk.com/im?sel=241682732&st=%23data_1)["dur123"] = lapply(data\_1$tj6.2, as.character)  
data\_1["dur123"]  
dur1 = as.character(data\_1$tj6.2)  
dur2 = lapply(dur1, as.integer)  
dur3 = as.numeric(unlist(dur2))  
mean(dur3)  
data\_1["dur"] = (dur3 - mean(dur3)) / sqrt(var(dur3))  
data\_1["dur"]  
  
[#Семейное](https://vk.com/im?sel=241682732&st=%23%D0%A1%D0%B5%D0%BC%D0%B5%D0%B9%D0%BD%D0%BE%D0%B5) положение:  
# 1)Состоите в зарегистрированном браке:  
data\_1["wed"]= data\_1$t\_marst  
[#data\_1](https://vk.com/im?sel=241682732&st=%23data_1)["wed"] = lapply(data\_1$t\_marst, as.character)  
data\_1$wed1 = 0  
data\_1$wed1[which(data\_1$wed==2)] <- 1 # Состоите в зарегистрированном браке  
data\_1$wed1 = as.numeric(data\_1$wed1)  
data\_1["wed1"]  
  
  
# 2) Разведены и в браке не состоите/Bдовец (вдова):  
data\_1["wed2"] = lapply(data\_1["wed"], as.character)  
data\_1$wed2 = 0  
data\_1$wed2[which(data\_1$wed==4)] <- 1 # Разведены и в браке не состоите  
data\_1$wed2[which(data\_1$wed==5)] <- 1 # Bдовец (вдова)  
data\_1$wed2 = as.numeric(data\_1$wed2)  
data\_1["wed2"]

*Приложение2*

# 3)Никогда в браке не состояли:  
data\_1["wed3"]=data\_1$t\_marst  
data\_1$wed3 = 0  
data\_1$wed3[which(data\_1$wed==1)] <- 1 # Никогда в браке не состояли  
data\_1$wed3 = as.numeric(data\_1$wed3)  
data\_1["wed3"]

*Таблица 1*. Характеристики модели зависимости параметра *salary* от параметров *age, sex, higher\_educ, status2, dur, wed1,wed2,wed3* в наборе данных обследования РМЭЗ НИУ ВШЭ.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень  значимости | Можно ли исключить регрессор? |
| (Intercept) | -0.74 | 0.05 | -15.85 | <2e-16 | \*\*\* | Да |
| *age* | -0.08 | 0.02 | -5.09 | 3.79e-07 | \*\*\* | Нет |
| *sex* | 0.5 | 0.03 | 17.17 | <2e-16 | \*\*\* | Нет |
| *higher\_educ* | 0.56 | 0.03 | 18.6 | <2e-16 | \*\*\* | Да |
| *status2* | 0.36 | 0.03 | 11.53 | <2e-16 | \*\*\* | Да |
| *dur* | 0.11 | 0.01 | 7.94 | 2.60e-15 | \*\*\* | Нет |
| *wed1* | 0.11 | 0.04 | 2.56 | 0.01 | \* | Да |
| *wed2* | 0.13 | 0.05 | 2.38 | 0.02 | \* | Да |
| *wed3* | -0.1 | 0.05 | -1.82 | 0.07 | . | Да |

*Таблица 2*. Характеристики модели зависимости параметра *salary* от параметров *age, sex, dur* в наборе данных обследования РМЭЗ НИУ ВШЭ.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень  значимости |
| (Intercept) | -0.19 | 0.02 | -9.5 | <2e-16 | \*\*\* |
| *age* | -0.08 | 0.01 | -5.38 | 7.93e-08 | \*\*\* |
| *sex* | 0.42 | 0.03 | 14.08 | <2e-16 | \*\*\* |
| *dur* | 0.08 | 0.02 | 5.52 | 3.59e-08 | \*\*\* |

*Приложение3*

*Таблица 3*. Характеристики модели зависимости параметра *salary* от параметров *age, sex, dur, log(age), log(dur)* в наборе данных обследования РМЭЗ НИУ ВШЭ.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень  значимости |
| (Intercept) | 0.19 | 0.19 | 1.02 | 0.3 |  |
| *age* | -0.52 | 0.13 | -3.93 | 9.31e-05 | \*\*\* |
| *log(age)* | 009 | 0.07 | 1.29 | 0.2 |  |
| *sex* | 0.4 | 0.08 | 5.15 | 3.51e-07 | \*\*\* |
| *dur* | 0.07 | 0.07 | 1.06 | 0.29 |  |
| *log(dur)* | 0.01 | 0.09 | 0.13 | 0.9 |  |

*Таблица 5*. Характеристики модели зависимости параметра *salary* от параметров *age, sex, dur, log(age)* в наборе данных обследования РМЭЗ НИУ ВШЭ.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень  значимости |
| (Intercept) | 0.05 | 0.08 | 0.61 | 0.54 |  |
| *age* | -0.29 | 0.07 | -4.43 | 1.01e-05 | \*\*\* |
| *log(age)* | 0.032 | 004 | 0.89 | 0.37 |  |
| *sex* | 0.32 | 0.04 | 7.73 | 1.68e-14 | \*\*\* |
| *dur* | 0.09 | 0.02 | 4.76 | 2.05e-06 | \*\*\* |

*Таблица 7*. Характеристики модели зависимости параметра *salary* от параметров *sex, dur, log(age)* в наборе данных обследования РМЭЗ НИУ ВШЭ.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень  значимости |
| (Intercept) | -0.27 | 0.03 | -9.53 | <2e-16 | \*\*\* |
| *log(age)* | -0.1 | 0.02 | -5.4 | 7.18e-08 | \*\*\* |
| *sex* | 0.32 | 0.04 | 9.64 | 3.33-14 | \*\*\* |
| *dur* | 0.1 | 0.02 | 5.07 | 4.29e-07 | \*\*\* |

*Таблица 8*. Характеристики модели зависимости параметра *salary* от параметров *age, sex, dur, I(age0.1), I(dur0.1)* в наборе данных обследования РМЭЗ НИУ ВШЭ.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень  значимости |
| (Intercept) | -1.14 | 1.1 | -1.03 | 0.3 |  |
| *age* | 0.054 | 0.15 | -362 | 0.0003 | \*\*\* |
| *I(age0.1)* | 1.1 | 0.88 | 1.26 | 0.2 |  |
| *sex* | 0.4 | 0.08 | 5.1 | 3.68e-07 | \*\*\* |
| *dur* | 0.06 | 0.07 | 0.8 | 0.5 |  |
| *I(dur0.1)* | 0.26 | 0.96 | 0.27 | 0.79 |  |

*Приложение4*

*Таблица 11*. Характеристики модели зависимости параметра *salary* от параметров *sex, dur, I(age0.1)* в наборе данных обследования РМЭЗ НИУ ВШЭ.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень  значимости |
| (Intercept) | 0.9 | 0.2 | 4.57 | 5.29e-06 | \*\*\* |
| *I(age0.1)* | -1.2 | 0.21 | -5.69 | 1.44e-08 | \*\*\* |
| *sex* | 0.32 | 0.04 | 7.66 | 2.91e-14 | \*\*\* |
| *dur* | 0.1 | 0.02 | 5.02 | 5.67e-07 | \*\*\* |

*Таблица 13*. Характеристики модели зависимости параметра *salary* от параметров *sex, dur, I(age2)* для подмножества индивидов: “Городские жители, состоящие в браке” в наборе данных обследования РМЭЗ НИУ ВШЭ.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень  значимости |
| (Intercept) | 0.05 | 0.04 | 1.17 | 0.24 |  |
| *I(age2)* | 0.1 | 0.03 | 3.53 | 0.0004 | \*\*\* |
| *sex* | 0.65 | 0.05 | 12.43 | <2e-16 | \*\*\* |
| *dur* | -0.18 | 0.03 | -7.03 | 1.97e-12 | \*\*\* |

*Таблица 14*. Характеристики модели зависимости параметра *salary* от параметров *sex, dur, I(age2)* для подмножества индивидов: “ Разведенные, без высшего образования” в наборе данных обследования РМЭЗ НИУ ВШЭ.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Параметр \ Характеристики | Значение | Std. Error | t value | Pr(>|t|) | Уровень  значимости |
| (Intercept) | -0.26 | 0.04 | -5.9 | 7.22e-09 | \*\*\* |
| *I(age2)* | 0.03 | 0.03 | 2.78 | 0.0006 | \*\* |
| *sex* | 0.08 | 0.08 | 3.2 | 0.002 | \*\* |
| *dur* | -0.1 | 0.02 | -4.44 | 1.13e-15 | \*\*\* |

Заключение

В пункте ***№1*** была построена линейная регрессия зарплаты на все параметры. Оценив коэффициент вздутия дисперсии VIF, мы выявили, что линейной зависимости между переменными нет.

В пункте ***№2.1*** была попытка улучшить рассматриваемую модель, путём введения логарифмов регрессоров. Однако это не дало видимых результатов.

В пункте ***№2.2*** в модель были введены всевозможные произведения пар регрессоров, и была выявлена одна наилучшая модель по доле объяснённого разброса в данных R2.

Исходя из предоставленных данных, было выявлено, что наибольшую зарплату получают мужчины более молодого возраста, живущие в городе, с высшим образованием, с большой продолжительностью рабочей недели и никогда не состоящие в браке.

В пункте ***№5.1*** была построена зависимость параметра *salary* от параметров *sex, dur, I(age2)* для подмножества индивидов: “Городские жители, состоящие в браке”, и было выявлено, что наибольшую зарплату получают мужчины, с высшим образованием.

В пункте ***№5.2*** была построена зависимость параметра *salary* от параметров *sex, dur, I(age2)* для подмножества индивидов: “Разведенные, без высшего образования”, и было выявлено, что наибольшую зарплату получают мужчины более молодого возраста, с большой продолжительностью рабочей недели.

*Приложение*

Код решения задачи:

data\_2 = select(data\_1, salary, age, sex, higher\_educ, status2, dur, wed1,wed2,wed3)

data\_2 = na.omit(data\_2)  
glimpse(data\_2)

model1 = lm(data = data\_2, salary~age + sex + higher\_educ + status2 + dur + wed1 + wed2 + wed3)  
summary(model1)  
vif(model1)  
  
model2 = lm(data = data\_2, salary~age + sex + dur)   
summary(model2)   
vif(model2)   
  
model3 = lm(salary~age+log(age) + sex + dur + log(dur), data = data\_2)    
summary(model3)    
vif(model3)

model4 = lm(salary~age + log(age) + sex + dur, data = data\_2)    
summary(model4)    
vif(model4)

model5 = lm(salary~log(age) + sex + dur, data = data\_2)    
summary(model5)    
vif(model5)   
  
model6 = lm(salary~age + I(age^0.1) + sex + dur+ I(dur^0.1),data = data\_2)

summary(model6)

vif(model6)

model7 = lm(salary~age + I(age^0.1) + sex + dur, data = data\_2)    
summary(model7)    
vif(model7)   
  
model8 = lm(salary~I(age^0.1) + sex + dur, data = data\_2)    
summary(model8)    
vif(model8)   
  
model9 = lm(salary~I(age^0.2) + sex + dur, data = data\_2)    
summary(model9)    
vif(model9)   
  
model10 = lm(salary~I(age^0.3) + sex + dur, data = data\_2)    
summary(model10)    
vif(model10)

model11 = lm(salary~I(age^0.4) + sex + dur, data = data\_2)    
summary(model11)

vif(model11)   
  
model12 = lm(salary~I(age^0.5) + sex + dur, data = data\_2)    
summary(model12)    
vif(model12)

model13 = lm(salary~I(age^0.9) + sex + dur, data = data\_2)    
summary(model13)    
vif(model13)   
  
model14 = lm(salary~I(age^1.1) + sex + dur, data = data\_2)    
summary(model14)    
vif(model14)

model15 = lm(salary~I(age^1.2) + sex + dur, data = data\_2)    
summary(model15)    
vif(model15)   
  
model16 = lm(salary~I(age^1.3) + sex + dur, data = data\_2)    
summary(model16)    
vif(model16)   
  
model17 = lm(salary~I(age^1.4) + sex + dur, data = data\_2)    
summary(model17)    
vif(model7)

model18 = lm(salary~I(age^1.9) + sex + dur, data = data\_2)    
summary(model18)    
vif(model18)

model19 = lm(salary~I(age^2) + sex + dur, data = data\_2)    
summary(model19)    
vif(model19)

data\_3 = subset(data\_2, status2 == 1)  
data\_3  
data\_4 = subset(data\_3, wed1 == 1)  
data\_4  
  
data\_5 = subset(data\_2, wed2 == 1)  
data\_5  
data\_6 = subset(data\_5, higher\_educ == 0)  
data\_6

model\_subset1 = lm(data = data\_4, salary~dur + sex + I(age^2))  
summary(model\_subset1)  
  
model\_subset2 = lm(data = data\_6, salary~dur + sex + I(age^2))

summary(model\_subset2)