Задача 3

Условие

Набор данных: r23i\_os26c.sav – данные исследования RLMS-HSE

Объясняемая переменная: заработная плата за 30 дней - *salary*

Регрессоры:пол, возраст, семейное положение (состоит ли в зарегистрированном браке / разведён или вдовец / никогда не состоял в браке), наличие высшего образования, место проживания, среднее число рабочих часов в неделю – *sex, age, wed1, wed2, wed, higher\_educ, city\_status, working\_hours, salary.*

1. Постройте линейную регрессию зарплаты на все параметры, которые Вы выделили  
   из данных мониторинга. Не забудьте оценить коэффициент вздутия дисперсии VIF.
2. Поэкспериментируйте с функциями вещественных параметров: используйте  
   логарифм и степени (хотя бы от 0.1 до 2 с шагом 0.1).
3. Выделите наилучшие модели из построенных: по значимости параметров,  
   включённых в зависимости, и по объяснённому с помощью построенных  
   зависимостей разбросу adjusted R2 – R2adj.
4. Сделайте вывод о том, какие индивиды получают наибольшую зарплату.
5. Оцените регрессии для подмножества индивидов: а) городские жители, не состоявшие в браке; б) разведенные женщины, без высшего образования

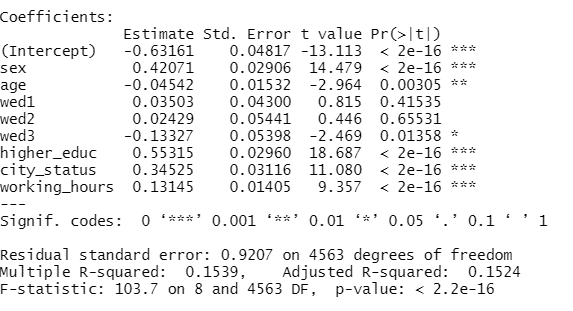
Решение

Считаем данные исследования из файла, уберём значения NA и представим их в удобном для исследования виде:

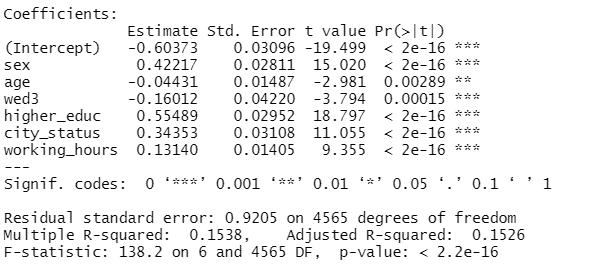
* Переменная *sex* принимает значение 1 для мужского пола, 0 – для женского
* *age* ­– переменная с нормализованным возрастом (формула для нормализации значения: (age - mean(age)) / sqrt(var(age)) )
* Семейное положение:
  + *wed1* = 1, если человек состоит в зарегистрированном браке, иначе 0
  + *wed2* = 1, если человек разведён или вдовец, иначе 0
  + *wed3* = 1, если человек никогда не был в браке, иначе 0
  + Проверим, что между *wed1, wed2, wed3* нет линейной зависимости
* *higher\_edu*c = 1, если у человека есть высшее образование, иначе 0
* *city\_status* = 1, если человек живёт в городе, иначе 0
* *working\_hours* – переменная с нормализованным числом рабочих часов в неделю (формула для нормализации значения: (working\_hours - mean(working\_hours)) / sqrt(var(working\_hours)) )
* *salary* – переменная с нормализованной зарплатой (формула для нормализации значения: ((salary - mean(salary)) / sqrt(var(salary)) )

1. Поострим линейную регрессию зарплаты на все параметры, оценим vif:

Модель строим командой model1 = lm(data = data2, salary ~ sex + age + wed1 + wed2 + wed3 + higher\_educ + city\_status + working\_hours)

  
Рисунок 1. Характеристики model1, где model1 = lm(data = data2, salary ~ sex + age + wed1 + wed2 + wed3 + higher\_educ + city\_status + working\_hours)

Из рисунка 1 видим, что переменные *wed1* и *wed2* имеют плохую p-статистику. Уберём их и посмотрим, как изменится R2:

  
Рисунок 2. Результат работы команды summary(model1), где model1 = lm(data = data2, salary ~ sex + age + wed3 + higher\_educ + city\_status + working\_hours)

Из рисунка 2 видим, что R2 изменился незначительно, зато p-статистика теперь хорошая для всех регрессоров. В дальнейшем будем работать с этой моделью.

Оценим vif у модели 1:

  
Рисунок 3. Результат работы команды vif(model1)

Из рисунка 3 видим, что vif низкий – линейной зависимости между регрессорами нет.

1. Введём в модель логарифмы и степени.

Логарифмы и степени имеет смысл вводить только для параметров *age* и *working\_hours*, так как остальные принимают только значения 0 или 1.

Модель с логарифмами: *model1 = lm(data = data2, salary ~ sex + working\_hours + age + wed3 + higher\_educ + city\_status + I(log(working\_hours)) + I(log(age)))* - у это модели плохой vif, попробуем убрать регрессоры с максимальным vif.

Модели с низким vif

model1 = lm(data = data2, salary ~ sex + working\_hours + age + wed3 + higher\_educ + city\_status + I(log(age)))

model2 = lm(data = data2, salary ~ sex + age + wed3 + higher\_educ + city\_status + I(log(working\_hours)) + I(log(age)))

model3 = lm(data = data2, salary ~ sex + higher\_educ + city\_status + I(log(working\_hours)) + I(log(age)))

Из них лучший R2 имеет вторая модель, но у неё плохая p-статистика для wed3 и *I(log(age))*. У модели 3 R2 чуть ниже, зато хорошая p-статистика для всех параметров

Построим модели со степенями в которых степень будет задаваться переменной *current\_pow,* меняющий значение от 0.1 до 2 с шагом 0.1:

current\_pow = 0.1

model1 = lm(data = data2, salary ~ sex + working\_hours + age + wed3 + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow)) – имеет плохие значения vif

Уберём из модели регрессоры working\_hours и age, как регрессоры с наибольшим vif

model1 = lm(data = data2, salary ~ sex + wed3 + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

Модель имеет R2 ~ 0.196 и плохую p-статистику у *wed3*. Уберём этот регрессор и посмотрим, как изменится R2

model1 = lm(data = data2, salary ~ sex+ higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

R2 ~ 0.196, везде хорошая p-статистика. Так как R2 изменился не сильно, *wed3* можно исключить как маловажный.

Если вместо *working\_hours* и *age* исключать из модели другие регрессоры или их комбинации, чтобы избавиться от линейной зависимости параметров, то R2 будет ниже.

Проделывая аналогичные действия для других степеней, заметим, что до степени 1.1 включительно R2 будет расти (исключая степень 1 – это будет исходная модель), а после этого будет снижаться. Vif и p-статистика у моделей при этом будет примерно одинаковой.

1. Выделим наилучшие модели из построенных: по значимости параметров, включённых в зависимости, и по объяснённому с помощью построенных зависимостей разбросу adjusted R2 – R2\_adj

Наилучшими по значению R2 из всех моделей без линейной зависимости регрессоров являются модели для степеней 0.9, 1.1, 1.2. Разброс R2 - R2\_adj у них одинаковый. Эти же модели являются лучшими по p-статистике у регрессоров. Из этих трёх моделей лучшей является модель model1 = lm(data = data2, salary ~ sex + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow)) для current\_pow = 1.1, которая имеет наивысший R2 = 0.21

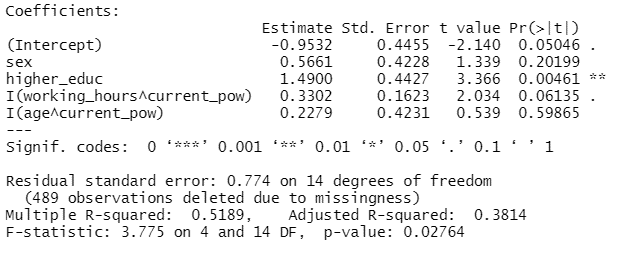
1. Согласно наилучшей модели больше всего зарабатывают молодые мужчины с высшим образованием, проживающие в городах, работающие большое число часов в неделю. Семейное положение, судя по всему, неважно.
2. Оценим регрессии для подмножества индивидов: а) Городские жители, не состоявшие в браке; б) разведенные женщины, без высшего образования

а) выделим множество городских жителей, не состоящих в браке:

data3 = subset(data2, city\_status == 1)

data3 = subset(data3, wed3 == 1)

Тогда имеем следующую модель:

  
Рисунок 4. Результат работы команды summary(model1), где model1 = lm(data = data3, salary ~ sex + higher\_educ + I(working\_hours^current\_pow) + I(age^current\_pow))

Все параметры, кроме *higher\_educ* - малозначимые. R2 ~ 0.5189. Согласно этой модели наибольшая зарплата у мужчин с высшим образованием старшего возраста, работающих много.

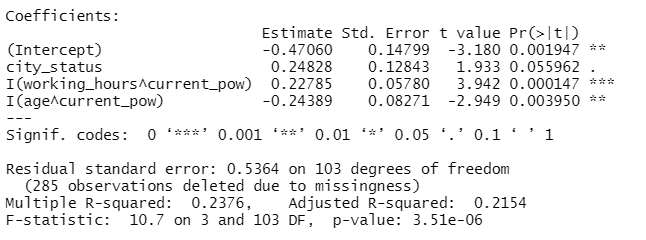
б) выделим множество разведенных женщин без высшего образования:

data3 = subset(data2, sex == 0)

data3 = subset(data3, wed2 == 1)

data3 = subset(data3, higher\_educ == 0)

Тогда имеем следующую модель:

  
Рисунок 5. Результат работы команды summary(model1), где model1 = lm(data = data3, salary ~ city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

Все параметры, кроме *city\_status*, значимые, R2 ~ 0.2376

Согласно этой модели наибольшая зарплата у проживающих в городе, работающих много, молодого возраста.

Полный код решения задачи приведён в Приложении 1.

Выводы

Из всей выборки больше всего зарабатывают молодые мужчины с высшим образованием, проживающие в городах, работающие большое число часов в неделю. Семейное положение, судя по всему, неважно.

Среди городских жителей, не состоящих в браке, наибольшая зарплата у мужчин с высшим образованием старшего возраста, работающих много.

Среди разведенных женщин без высшего образования наибольшая зарплата у проживающих в городе, работающих много, молодого возраста.

Приложение 1

Код решения задачи:

library("lmtest")

library("GGally")

library(car) # функция vif()

library("rlms") # Пакет предназначен для работы с данными исследования RLMS

library("dplyr") # Пакет для различных манипуляций с данными

library(sandwich) # Пакет для оценки стандартной ошибки

**data** <- rlms\_read("r23i\_os26c.sav") # Считываю данные исследования для своего варианта

"

sh5 Пол респондента

1 МУЖСКОЙ

2 ЖЕНСКИЙ

age Количество полных лет

s\_marst СЕМЕЙНОЕ ПОЛОЖЕНИЕ

1 Никогда в браке не состояли

2 Состоите в зарегистрированном браке

3 Живете вместе, но не зарегистрированы

4 Разведены и в браке не состоите

5 Bдовец (вдова)

6 ОФИЦИАЛЬНО ЗАРЕГИСТРИРОВАНЫ, НО ВМЕСТЕ НЕ ПРОЖИВАЮТ

s\_diplom ЗАКОНЧЕННОЕ ОБРАЗОВАНИЕ (ГРУППА)

1 окончил 0 - 6 классов

2 незаконч среднее образование (7 - 8 кл)

3 незаконч среднее образование (7 - 8 кл) + что-то еще

4 законч среднее образование

5 законч среднее специальное образование

6 законч высшее образование и выше

status ТИП НАСЕЛЕННОГО ПУНКТА

1 областной центр

2 город

3 ПГТ

4 село

sj13.2 За последние 12 месяцев какова была Ваша среднемесячная зарплата на этом

предприятии после вычета налогов - независимо от того, платят Вам ее вовремя или

нет?

sj6.2 Сколько часов в среднем продолжается Ваша обычная рабочая неделя?

"

**data** = select(**data**, sh5, s\_age, s\_marst, s\_diplom, status, sj13.2, sj6.2)

**data** = na.omit(**data**)

glimpse(**data**)

**data**2 = select(**data**,) #Новая база данных для нормализованных значений

#Пол

**data**2["sex"] = 0

**data**2$sex[which(**data**$sh5 == 1)] <- 1

#Возраст

age = **data**$s\_age

**data**2["age"] = (age - mean(age)) / sqrt(var(age))

#Семейное положение:

#Состоит ли в зарегестрированном браке?

**data**2$wed1 = 0

**data**2$wed1[which(**data**$s\_marst==2)] <- 1

**data**2$wed1[which(**data**$s\_marst==6)] <- 1

#Разведён или вдовец?

**data**2$wed2 = 0

**data**2$wed2[which(**data**$s\_marst==4)] <- 1

**data**2$wed2[which(**data**$s\_marst==5)] <- 1

#Никогда не состоял в браке?

**data**2$wed3 = 0

**data**2$wed3[which(**data**$s\_marst==1)] <- 1

# Проверим, что отсутствует линейная зависимость между симейными положениями

vif(lm(**data**$sj13.2 ~ **data**2$wed1 + **data**2$wed2 + **data**2$wed3))

#Наличие высшего образования

**data**2$higher\_educ = 0

**data**2$higher\_educ[which(**data**$s\_diplom==6)] <- 1

#Живёт в городе?

**data**2$city\_status = 0

**data**2$city\_status[which(**data**$status==1)] <- 1

**data**2$city\_status[which(**data**$status==2)] <- 1

#Нормализованное среднее число рабочих часов в неделю

working\_hours = **data**$sj6.2

**data**2$working\_hours = (working\_hours - mean(working\_hours)) / sqrt(var(working\_hours))

#Нормализованная средняя зарплата

salary = **data**$sj13.2

**data**2$salary = (salary - mean(salary)) / sqrt(var(salary))

# 1. Постройте линейную регрессию зарплаты на все параметры, которые Вы выделили из данных мониторинга. Не забудьте оценить коэффициент вздутия дисперсии VIF.

model1 = lm(**data** = **data**2, salary ~ sex + age + wed1 + wed2 + wed3 + higher\_educ + city\_status + working\_hours)

vif(model1)

summary(model1)

model1 = lm(**data** = **data**2, salary ~ sex + age + wed3 + higher\_educ + city\_status + working\_hours)

vif(model1)

summary(model1)

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

# с логарифмами:

model1 = lm(**data** = **data**2, salary ~ sex + working\_hours + age + wed3 + higher\_educ + city\_status + I(log(working\_hours)) + I(log(age)))

vif(model1)

summary(model1)

#R^2 ~ 0.21

# Не самые хорошие p-статистика и vif, увеличился R^2

model1 = lm(**data** = **data**2, salary ~ sex + working\_hours + age + wed3 + higher\_educ + city\_status + I(log(age)))

vif(model1)

summary(model1)

# R^2 ~ 0.179

model1 = lm(**data** = **data**2, salary ~ sex + age + wed3 + higher\_educ + city\_status + I(log(working\_hours)) + I(log(age)))

vif(model1)

summary(model1)

# R^2 ~ 0.2082

model1 = lm(**data** = **data**2, salary ~ sex + higher\_educ + city\_status + I(log(working\_hours)) + I(log(age)))

vif(model1)

summary(model1)

# R^2 ~ 0.1924, хорошая p-статистика

#Со степенями:

current\_pow = 0.1

model1 = lm(**data** = **data**2, salary ~ sex + working\_hours + age + wed3 + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.213

# Плохие значения vif

model1 = lm(**data** = **data**2, salary ~ sex + wed3 + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.196

# Плохая p-статистика у wed3

model1 = lm(**data** = **data**2, salary ~ sex+ higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.196

# Везде хорошая p-статистика

current\_pow = 0.2

model1 = lm(**data** = **data**2, salary ~ sex + working\_hours + age + wed3 + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.213

# Плохие значения vif

model1 = lm(**data** = **data**2, salary ~ sex + age + wed3 + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.21

# Плохие значения vif

model1 = lm(**data** = **data**2, salary ~ sex + wed3 + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.199

# Плохая p-статистика у wed3

model1 = lm(**data** = **data**2, salary ~ sex + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.199

# Везде хорошая p-статистика

current\_pow = 0.3

model1 = lm(**data** = **data**2, salary ~ sex + working\_hours + age + wed3 + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.213

# Плохие значения vif

model1 = lm(**data** = **data**2, salary ~ sex + wed3 + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.20

# Плохая p-статистика у wed3

model1 = lm(**data** = **data**2, salary ~ sex + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.202

# Везде хорошая p-статистика

current\_pow = 0.4

model1 = lm(**data** = **data**2, salary ~ sex + working\_hours + age + wed3 + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.213

# Плохие значения vif

model1 = lm(**data** = **data**2, salary ~ sex + wed3 + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.205

# Плохая p-статистика у wed3

model1 = lm(**data** = **data**2, salary ~ sex + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.205

# Везде хорошая p-статистика

#Заметим, что с увеличением current\_pow немного увеличивается R^2, перейдём сразу к степени 0.9

current\_pow = 0.9

model1 = lm(**data** = **data**2, salary ~ sex + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.2117

# Везде хорошая p-статистика

# Для степени 1 результат мы уже имеем

current\_pow = 1.1

model1 = lm(**data** = **data**2, salary ~ sex + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.2118

# Везде хорошая p-статистика

current\_pow = 1.2

model1 = lm(**data** = **data**2, salary ~ sex + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.2114

# Везде хорошая p-статистика

current\_pow = 1.3

model1 = lm(**data** = **data**2, salary ~ sex + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.2108

# Везде хорошая p-статистика

current\_pow = 1.4

model1 = lm(**data** = **data**2, salary ~ sex + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.21

# Везде хорошая p-статистика

current\_pow = 1.5

model1 = lm(**data** = **data**2, salary ~ sex + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.209

# Везде хорошая p-статистика

# и далее R^2 уменьшается

current\_pow = 1.7

model1 = lm(**data** = **data**2, salary ~ sex + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.206

# Везде хорошая p-статистика

current\_pow = 1.9

model1 = lm(**data** = **data**2, salary ~ sex + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.203

# Везде хорошая p-статистика

current\_pow = 2

model1 = lm(**data** = **data**2, salary ~ sex + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# R^2 ~ 0.15

# Плохая p-статистика у I(working\_hours^current\_pow)

# 3.Выделите наилучшие модели из построенных: по значимости параметров,включённых в зависимости, и по объяснённому с помощью построенных зависимостей разбросу adjusted R2 - R2adj.

# Сравним лучшие модели:

current\_pow = 0.9

model1 = lm(**data** = **data**2, salary ~ sex + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

#Multiple R-squared: 0.2117, Adjusted R-squared: 0.2061

current\_pow = 1.1

model1 = lm(**data** = **data**2, salary ~ sex + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

#Multiple R-squared: 0.2118, Adjusted R-squared: 0.2062

current\_pow = 1.2

model1 = lm(**data** = **data**2, salary ~ sex + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

#Multiple R-squared: 0.2114, Adjusted R-squared: 0.2058

# Разброс R2 - R2\_adj везде одинаковый, а R^2 больше для степени 1.1

# Итого, среди моделей без линейной зависимости параметров с хорошими показателями p-статистики у регрессоров лучшей по R^2 оказалась модель для степени 1.1:

current\_pow = 1.1

model1 = lm(**data** = **data**2, salary ~ sex + higher\_educ + city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

vif(model1)

summary(model1)

# с R^2 ~ 0.2118

# 4. Сделайте вывод о том, какие индивиды получают наибольшую зарплату.

#Согласно этой модели больше всего зарабатывают молодые мужчины с высшим образованием, проживающие в городах, работающие большое число часов в неделю.

#Семейное положение, судя по всему, неважно

# 5. Оцените регрессии для подмножества индивидов: а) Городские жители, не состоявшие в браке; б )разведенные женщины, без высшего образования

current\_pow = 1.1

#Городские жители, не состоявшие в браке

**data**3 = subset(**data**2, city\_status == 1)

**data**3 = subset(**data**3, wed3 == 1)

model1 = lm(**data** = **data**3, salary ~ sex + higher\_educ + I(working\_hours^current\_pow) + I(age^current\_pow))

summary(model1)

# Параметры, кроме higher\_educ - малозначимые. R^2 ~ 0.5189

# Наибольшая зарплата у мужчин с высшим образованием старшего возраста, работающих много

#разведенные женщины, без высшего образования

**data**3 = subset(**data**2, sex == 0)

**data**3 = subset(**data**3, wed2 == 1)

**data**3 = subset(**data**3, higher\_educ == 0)

model1 = lm(**data** = **data**3, salary ~ city\_status + I(working\_hours^current\_pow) + I(age^current\_pow))

summary(model1)

# почти все параметры значимые, R^2 ~ 0.2376

# Наибольшая зарплата у проживающих в городе, работающих много, молодого возраста.