Задача 3

Условие

Набор данных: r22i\_os26c.sav.

Объясняемая переменная: заработная плата – *salary.*

Регрессоры: возраст, пол, образование, населенный пункт, продолжительность рабочей недели, семейное положение (состоит ли в зарегистрированном браке/никогда не был в браке / разведён или вдовец /живет вместе, но не зарегистрированы /зарегистрированы, но не живут вместе), населенный пункт (областной центр/город), продолжительность рабочей недели, – *salary*, *age, sex, r\_educ, status1, dur, wed, wed2, wed3.*

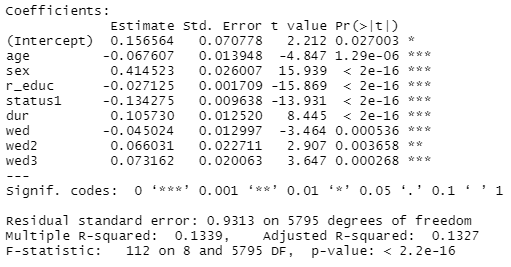
1. Постройте линейную регрессию зарплаты на все параметры, которые Вы выделили  
   из данных мониторинга. Не забудьте оценить коэффициент вздутия дисперсии VIF.
2. Поэкспериментируйте с функциями вещественных параметров: используйте  
   логарифм и степени (хотя бы от 0.1 до 2 с шагом 0.1).
3. Выделите 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 из файла данные исследования и представим их в удобном виде:

1. *salary* – переменная с нормализованной зарплатой (формула для нормализации значения: ((salary - mean(salary)) / sqrt(var(salary)) )
2. *age* ­– переменная с нормализованным возрастом.
3. Переменная *sex* принимает 1, если человеком был мужчина и 0, если человеком была женщина.
4. *r\_edu*c = 1, если у человека есть высшее образование.
5. *status1* = 1, если человек живёт в городе или в областном центре.
6. *dur* – продолжительность рабочей недели.
7. Семейное положение:
   * *wed* = 1, если человек никогда не был в браке или состоял в зарегистрированном браке.
   * *wed2* = 1, если человек разведён или вдовец
   * *wed3* = 1, если люди жили вместе, но не были зарегистрированы или официально были зарегистрированы, но вместе не проживали.
   * Проверим, что между *wed, wed2, wed3* нет линейной зависимости.
8. Построим линейную регрессию зарплаты на все параметры:

Строим модель salary~age + sex + r\_educ + status1 + dur + wed + wed2 + wed3

  
Рисунок 1. Характеристики model, где model = lm(data = data3, salary~age + sex + r\_educ + status1 + dur + wed + wed2 + wed3)

Оценим vif у модели:

  
Рисунок 2. Результат проверки нашей модели на линейную зависимость среди регрессоров.

Из рисунка 2 видим, что vif низкий – линейной зависимости между регрессорами нет.

1. Введём в модель степени и логарифмы.

Модели со степенями, где степень изменяется от 0.1 до 2, с шагом в 0.1:

Степень 0.1:

*model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(dur^0.1) + I(age^0.1))* - у это модели vif меньше 7, низкие показатели p-статистики и R^2 = 0.1691.

Степень 0.2:

model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(dur^0.2) + I(age^0.2)) - у это модели vif меньше 7, низкие показатели p-статистики и R^2 = 0.1695.

Степень 0.3:

model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(dur^0.3) + I(age^0.3)) - у это модели vif ~10, значит есть линейная зависимость.

Степень 0.4:

model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(age^0.4) + I(dur^0.4)) - у это модели vif ~13, значит есть линейная зависимость.

Так как при степень большей 0.3 появляется линейная зависимость, проверим модель сразу на высокой степени (2):

model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(age^2) + I(dur^2)) - у это модели vif меньше 7, неплохие показатели р-статистики и R^2 = 0.1488.

Модель с логарифмами:

model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(log(age)) + I(log(dur))) - у это модели vif меньше 7 и R^2 = 0.1687.

Попробуем ввести логарифм только для одного из двух регрессоров:

1. model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(log(dur))) - у это модели vif меньше 7, р-статистика хорошая, кроме dur и log(dur), но если мы уберем log(dur), то получим исходную модель и R^2 = 0.1407.
2. model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(log(age))) - у это модели vif меньше 7, р-статистика хорошая, кроме wed, wed2, wed3, но мы не можем их убрать, так как они важны для нашей модели; R^2 = 0.1596.

Из них лучший R2 имеет модель с log(dur), но у неё плохая p-статистика для dur и log(dur).

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

Из всех моделей, без линейной зависимости, лучшие по значению R^2 – это модели: со степенью 0.4 и логарифмом(log(dur), lof(age)). Разброс R2 - R2\_adj у них одинаковый. Из этих моделей лучшей является модель model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(log(age)) + I(log(dur))). Лучшей же моделью по р-статистике является наша исходная модель: model = lm(data = data3, salary~age + sex + r\_educ + status1 + dur + wed + wed2 + wed3).

1. Согласно наилучшей модели по р-статистике больше всего зарабатывают мужчины молодого возраста, без высшего образования, работающие в областном центре, перерабатывающие, не состоявшие в браке или вдовец или официально зарегистрированные, но не проживающие вместе.
2. Оценим регрессии для подмножества индивидов: а) Женщины не замужем б) Женщины, живущие в городе, разведенные

а) выделим множество женщин не замужем:

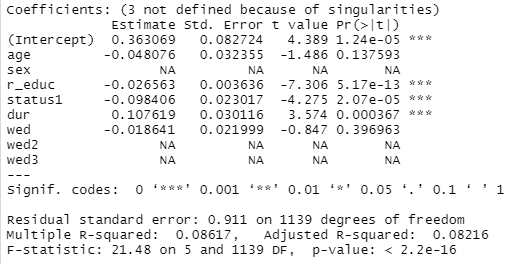
data4 = subset(data3, sex == 0)

data4

data5 = subset(data4, wed2 == 1)

data5

Имеем следующую модель:

  
Рисунок 3. Характеристики model\_subset, где model\_subset = lm(data = data5, salary~age + sex + r\_educ + status1 + dur + wed + wed2 + wed3)

R2 ~ 0.08617. Согласно этой модели наибольшая зарплата у мужчин, без высшего образования, работающих не в городе, и перерабатывающих.

б) выделим множество разведенных женщин без высшего образования:

data6 = subset(data3, sex == 0)

data6

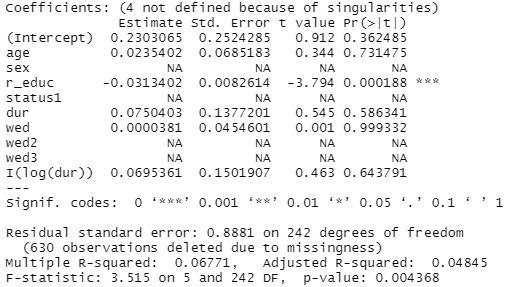
data7 = subset(data6, status1 == 1)

data7

data8 = subset(data7, wed2 == 1)

data8

Тогда имеем следующую модель:

   
Рисунок 4. Характеристики model\_subset, где model\_subset = lm(data = data8, salary~age + sex + r\_educ + status1 + dur + wed + wed2 + wed3)

R2 ~ 0.06771. Согласно этой модели наибольшая зарплата у мужчин, без высшего образования и перерабатывающих.

Полный код решения задачи приведён в Приложении 1.

Выводы

Из всех полученных данных наибольшую ЗП имеют мужчины молодого возраста, без высшего образования, работающие в областном центре, перерабатывающие, не состоявшие в браке или вдовец или официально зарегистрированные, но не проживающие вместе.

Среди женщин не замужем, наибольшая зарплата у мужчин с высшим образованием старшего возраста, работающих много.

Среди разведенных женщин без высшего образования наибольшая зарплата у мужчин, без высшего образования и перерабатывающих.

Приложение 1

Код решения:

#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\\vladb\\Downloads\\r22i\_os26c.sav")

glimpse(data)

data2 = select(data, rj13.2, r\_age, rh5, r\_educ, status, rj6.2, r\_marst)

#Убираем строки с NA

data2 = na.omit(data2)

glimpse(data2)

#зарплата c элементами нормализации

data2$rj13.2

sal = as.numeric(data2$rj13.2)

sal1 = as.character(data2$rj13.2)

sal2 = lapply(sal1, as.integer)

sal = as.numeric(unlist(sal2))

mean(sal)

data2["salary"] = (sal - mean(sal)) / sqrt(var(sal))

data2["salary"]

#возраст c элементами нормализации

age1 = as.character(data2$r\_age)

age2 = lapply(age1, as.integer)

age3 = as.numeric(unlist(age2))

data2["age"]= (age3 - mean(age3)) / sqrt(var(age3))

data2["age"]

#пол

#data2["sex"]=data2$rh5

#data2["sex"] = lapply(data2$rh5, as.character)

data2$sex = as.numeric(data2$rh5)

data2$sex[which(data2$sex!='1')] <- 0 #женский

data2$sex[which(data2$sex=='1')] <- 1 #мужской

#образование

#data2["r\_educ"] = data2$r\_educ

#data2["r\_educ"] = lapply(data2$r\_educ, as.character)

#data2["higher\_educ"] = data2$r\_educ

data2$r\_educ = as.numeric(data2$r\_educ)

data2$r\_educ[which(data2$r\_educ=='21')] <- 1 #есть диплом о высшем образовании

data2$r\_educ[which(data2$r\_educ=='22')] <- 1 #аспирантура и т.п. без диплома

data2$r\_educ[which(data2$r\_educ=='23')] <- 1 #аспирантура и т.п. с дипломом

#населенный пункт

#data2["status1"]=data2$status

#data2["status1"] = lapply(data2$status, as.character)

data2$status1 = as.numeric(data2$status)

data2$status1[which(data2$status1=='1')] <- 1 #областной центр

data2$status1[which(data2$status1=='2')] <- 1 #город

#продолжительность рабочей недели

dur1 = as.character(data2$rj6.2)

dur2 = lapply(dur1, as.integer)

dur3 = as.numeric(unlist(dur2))

data2["dur"] = (dur3 - mean(dur3)) / sqrt(var(dur3))

#семейное положение

#data2["wed"]= data2$r\_marst

#data2["wed"] = lapply(data2$r\_marst, as.character)

data2$wed = as.numeric(data2$r\_marst)

data2$wed[which(data2$wed=='1')] <- 1 #никогда не были в браке

data2$wed[which(data2$wed=='2')] <- 1 # Состоят в зарегистрированном браке

#data2["wed2"] = lapply(data2["wed"], as.character)

data2$wed2 = as.numeric(data2$r\_marst)

data2$wed2[which(data2$wed=='4')] <- 1 # Разведены

data2$wed2[which(data2$wed=='5')] <- 1 # Bдовец/вдова

#data2$wed2 = as.numeric(data2$wed2)

#data2["wed3"] = lapply(data2["wed"], as.character)

data2$wed3 = as.numeric(data2$r\_marst)

data2$wed3[which(data2$wed=='3')] <- 1 # Живете вместе, но не зарегистрированы

data2$wed3[which(data2$wed=='6')] <- 1 # ОФИЦИАЛЬНО ЗАРЕГИСТРИРОВАНЫ, НО ВМЕСТЕ НЕ ПРОЖИВАЮТ

data3 = select(data2, salary, age, sex, r\_educ, status1, dur, wed, wed2, wed3)

#построение зависимостей для данных

model = lm(data = data3, salary~age + sex + r\_educ + status1 + dur + wed + wed2 + wed3)

summary(model)

vif(model)

#R^2 = 0.1339 - очень низкий

#p-характеристика у всех значений (\*\*\*), кроме wed2 (\*\*)

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

model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(dur^0.1) + I(age^0.1))

summary(model\_test\_1)

vif(model\_test\_1)

#R^2 = 0.1691

#р - ухудшился

#vif <= 7

model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(dur^0.2) + I(age^0.2))

summary(model\_test\_1)

vif(model\_test\_1)

#R^2 = 0.1695

#р - ухудшился

#vif <= 7

model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(dur^0.3) + I(age^0.3))

summary(model\_test\_1)

vif(model\_test\_1)

#R^2 = 0.17

#р - ухудшился

#vif <= 10 ~ есть линейнай зависимость

model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(age^0.4) + I(dur^0.4))

summary(model\_test\_1)

vif(model\_test\_1)

#R^2 = 0.1704

#р - ухудшился

#vif <= 13 ~ есть линейнай зависимость

model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(dur^0.4))

summary(model\_test\_1)

vif(model\_test\_1)

#R^2 = 0.1413

#р - ухудшился

#vif <= 14 ~ есть линейнай зависимость

model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(age^0.4))

summary(model\_test\_1)

vif(model\_test\_1)

#R^2 = 0.1595

#р - ухудшился

#vif <= 11 ~ есть линейнай зависимость

model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(dur^0.4))

summary(model\_test\_1)

vif(model\_test\_1)

#R^2 = 0.1413

#р - ухудшился

#vif <= 14 ~ есть линейнай зависимость

model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(age^2) + I(dur^2))

summary(model\_test\_1)

vif(model\_test\_1)

#R^2 = 0.1488

#р - ухудшился

#vif <= 3

model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(dur^2))

summary(model\_test\_1)

vif(model\_test\_1)

#R^2 = 0.1345

#р - ухудшился

#vif <= 3

model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(age^2))

summary(model\_test\_1)

vif(model\_test\_1)

#R^2 = 0.1484

#р - чуть-чуть хуже, чем у исходной модели

#vif <= 3

model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(log(age)) + I(log(dur)))

summary(model\_test\_1)

vif(model\_test\_1)

#R^2 = 0.1687

#р - ухудшился

#vif <= 7

model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(log(dur)))

summary(model\_test\_1)

vif(model\_test\_1)

#R^2 = 0.1407

#р - хуже, чем у начальной модели, но лучше, чем у остальных моделей со степенями и логарифмами

#vif <= 5

model\_test\_1 = lm(data = data3, salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(log(age)))

summary(model\_test\_1)

vif(model\_test\_1)

#R^2 = 0.1596

#р - ухудшился

#vif <= 7

#Итог: лучшая модель это salary~ age + sex + r\_educ + status1 + dur + wed + wed2 + wed3 + I(log(dur))

#проверим модель на линейную зависимость параметров

modele\_1 = lm(dur~I(log(dur)), data3)

modele\_1

summary(modele\_1) # R^2 = 0.8063 < 0.8, значит нет линейной зависимости и можно использовать в одной модели

modele\_2 = lm(age~I(dur^2), data3)

modele\_2

summary(modele\_2) # R^2 = 3.001e-05 < 0.1 , значит нет линейной зависимости и можно использовать в одной модели

#Вывод: более высокую ЗП получают мужчины молодого возраста без высшего образования, работающие в областном центре , перерабатывающие,

# не состоявшие в браке или вдовец или официально зарегистрированные, но не проживающие вместе.

#(Intercept) age sex r\_educ status1 dur wed wed2 wed3 I(log(dur))

#0.009378 -0.100804 0.452144 -0.030679 -0.154806 0.068537 -0.086875 0.143205 0.152724 0.056648

#Ищем подмножества: "Женщины не замужем"

data4 = subset(data3, sex == 0)

data4

data5 = subset(data4, wed2 == 1)

data5

# Ищем подмножество: "женщины, живущие в городе, разведенные"

data6 = subset(data3, sex == 0)

data6

data7 = subset(data6, status1 == 1)

data7

data8 = subset(data7, wed2 == 1)

data8

model\_subset = lm(data = data5, salary~age + sex + r\_educ + status1 + dur + wed + wed2 + wed3)

summary(model\_subset)

# R^2 = 0.08617

# Наивысшую зарплату получают мужчины , без высшего образования, работающие не в городе, и перерабатывающие

model\_subset = lm(data = data8,salary~age + sex + r\_educ + status1 + dur + wed + wed2 + wed3)

summary(model\_subset)

# R^2 = 0.06726

# Наивысшую зарплату получают мужчины, без высшего образования и перерабатывающие