Практическая Работа Задание №1

Вступление: В этом отчете анализируется “swiss” набор данных, доступный в R, который предоставляет данные о социальных и экономических факторах в швейцарских кантонах. Основной целью этого анализа является изучение взаимосвязи между показателями младенческой смертности и двумя переменными: сельским хозяйством и частотой обследований.

Необходимо загрузить данные из указанного набора и произвести следующие действия.

Набор Данных: Swiss

Объясняемая переменная: Infant.Mortality

Регрессоры: Agriculture, Examination

**Код решение задачи 1**

library("lmtest")

library("GGally")

data = swiss

help(swiss)

data

summary(data)

ggpairs(data)

mean(swiss$Agriculture)

var(swiss$Agriculture)

sd(swiss$Agriculture)

mean(swiss$Examination)

var(swiss$Examination)

sd(swiss$Examination)

model1 <- lm(Infant.Mortality~Agriculture, data = swiss)

plot(model1)

ggpairs(model1)

model2 <- lm(Infant.Mortality~Examination, data = swiss)

plot(model2)

ggpairs(model2)

summary(model1)$r.squared

summary(model2)$r.squared

summary(model1)$coefficients

summary(model2)$coefficients

* Команды “library("lmtest")” и “library("GGally")” загружают пакеты, установленные на вашем компьютере.
* Команда "data = swiss" изменяет набор данных swiss на именованный "data".
* “Help(swiss)” показывает, какие данные хранятся в наборе данных “Swiss”.
* “mean(swiss$Agriculture)”, “var(swiss$Agriculture)” и “sd(swiss$Agriculture)” вычисляют среднее значение, дисперсию и стандартное отклонение для переменной сельского хозяйства

> mean(swiss$Agriculture)

[1] 50.65957

> var(swiss$Agriculture)

[1] 515.7994

> sd(swiss$Agriculture)

[1] 22.71122

* Те же вычисления выполняются для переменной Examination.

> mean(swiss$Examination)

[1] 16.48936

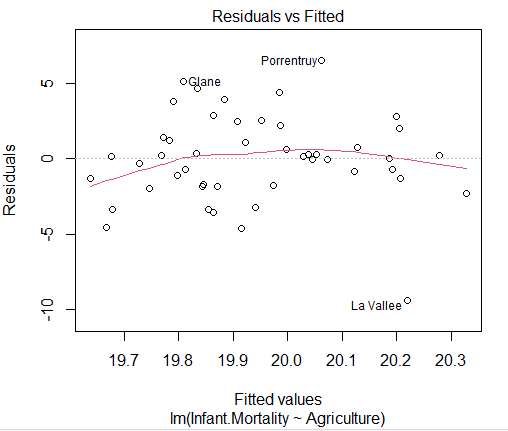
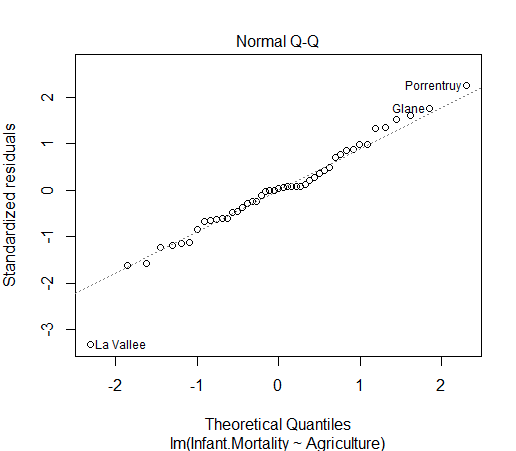
> var(swiss$Examination)

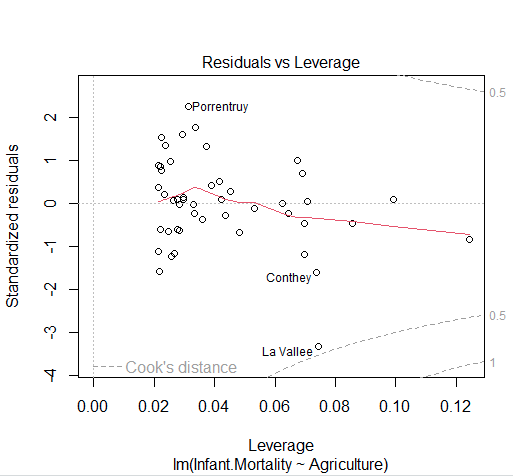
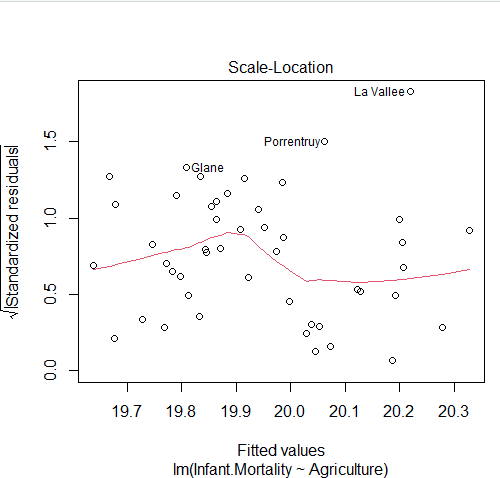
[1] 63.64662

> sd(swiss$Examination)

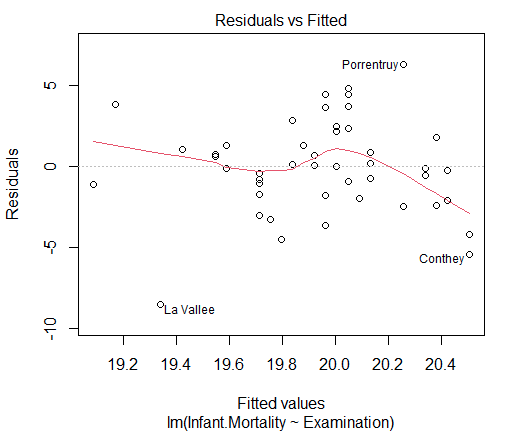
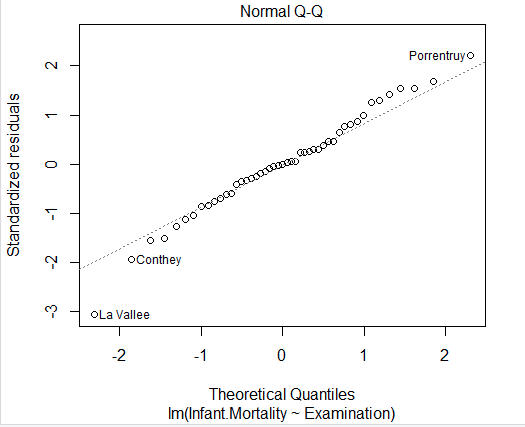
[1] 7.977883

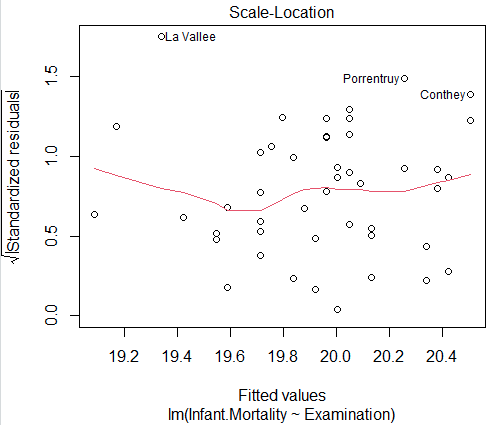
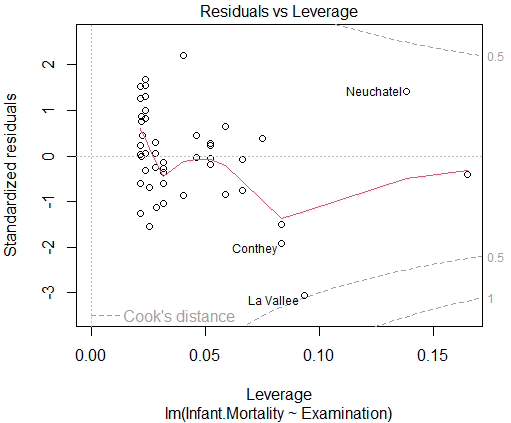
* “model1 <- lm(Infant.Mortality~Agriculture, data = swiss)” и “model2 <- lm(Infant.Mortality~Examination, data = swiss)” соответствуют моделям линейной регрессии для младенцев.Смертность в качестве зависимой переменной и сельское хозяйство и обследование в качестве предикторов, соответственно.
* “plot(model1)” и “plot(model2)” предоставляют различные диагностические графики для двух регрессионных моделей. Рисунки “plot(model1)”:





Рисунки “plot(model2)”:





* Функция ggpairs(data), ggpairs(model1) и ggpairs(model2) генерируют матрицу диаграмм рассеяния, показывающую взаимосвязи между всеми парами переменных в данном наборе данных. (На рисунках 1, 2 и 3)в
* summary(model1)$r.squared и summary(model2)$r.squared предоставляют значения R-квадрата для двух моделей, указывающие долю дисперсии зависимой переменной, которая предсказуема по независимой переменной.

1. R^2 < 30% Модель плоха. Возможно, нужно строить другую модель, а не редактировать эту. Зависимости нет.
2. R^2 > 80% Модель хороша. Лучше уже вряд ли получится. Линейная Регрессия – это упрощение реальной ситуации. За это упрощение расплачиваемся частью поведения объясняемой переменной, которую нельзя описать в упрощённой системе.

> summary(model1)$r.squared

[1] 0.00370377

> summary(model2)$r.squared

[1] 0.01300092

Обе модели плохи. Но на данный момент мы видим, что первый вариант лучше второго. Так что, если нам придется выбирать из этих 2-х, мы можем выбрать второй.

* “summary(model1)$coefficients” и “summary(model2)$coefficients” предоставляют оценочные коэффициенты для двух моделей, которые включают пересечения и наклоны для предикторов. Выходные данные также включают стандартные ошибки, t-значения и p-значения коэффициентов.

> summary(model1)$coefficients

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

(Intercept) 20.337954766 1.05754318 19.2313233 2.463854e-23

Agriculture -0.007805071 0.01908283 -0.4090101 6.844724e-01

> summary(model2)$coefficients

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

(Intercept) 20.62898680 0.98845610 20.8699068 8.941778e-25

Examination -0.04162888 0.05407041 -0.7699012 4.453814e-01

1. Coefficient estimates: Они представляют собой изменение переменной отклика при изменении предиктора на одну единицу, предполагая, что все остальные предикторы остаются постоянными.
2. Standard errors: Меньшие стандартные ошибки предполагают более точные (надежные) оценки коэффициентов.
3. t-values: Это оценки коэффициентов, деленные на их стандартные ошибки. Большие абсолютные значения t указывают на более веские доказательства против нулевой гипотезы о том, что истинный коэффициент равен нулю (учитывая, что все остальные предикторы присутствуют в модели).
4. p-values: Они представляют вероятность наблюдения такого экстремального t-значения, если бы нулевая гипотеза была верна. Меньшие значения p (обычно менее 0,05) свидетельствуют о веских доказательствах против нулевой гипотезы, указывая на то, что предиктор, скорее всего, значим.

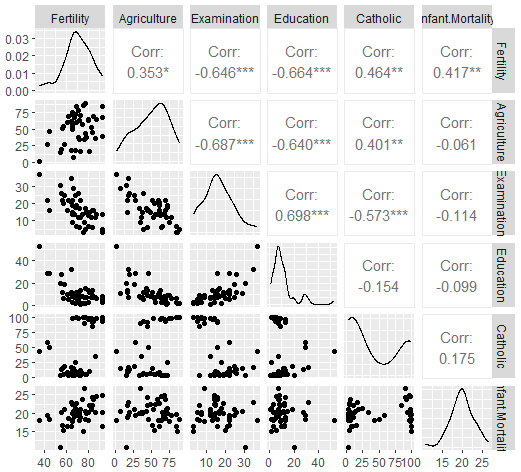


Рисунок 1. Результат работы команды ggpairs(data) – графики зависимостей между парами переменных наборе данных Swiss

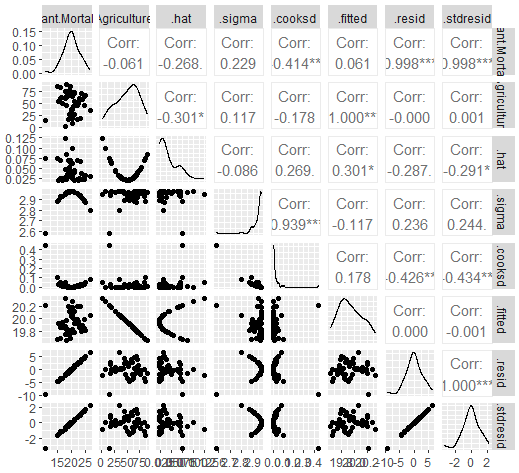


Рисунок 2. Результат работы команды ggpairs(model1) – графики зависимостей между парами Infant.Mortality и Agriculture

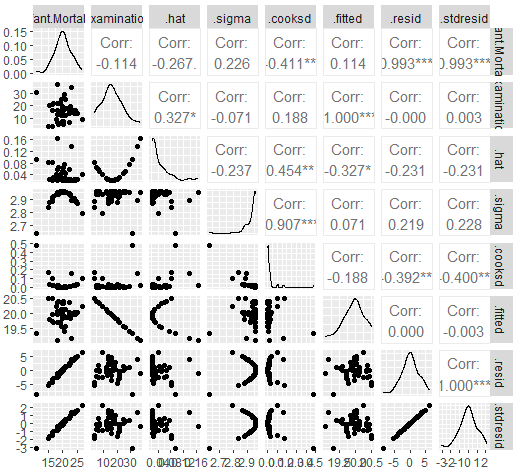


Рисунок 3. Результат работы команды ggpairs(model2) – графики зависимостей между парами Infant.Mortality и Examination

Практическая Работа Задание №2

Вступление: В этом отчете представлен анализ набора данных “attitude” в R, который содержит данные об отношении сотрудников к своей работе. Основная цель этого анализа - исследовать взаимосвязи между рейтингами должностей и тремя переменными: привилегиями, возможностями обучения и повышением заработной платы.

Необходимо загрузить данные из указанного набора и произвести следующие действия.

Набор Данных: Attitude

Объясняемая переменная: Rating

Регрессоры: Privileges, Learning, Raises

**Код решение задачи №2**

library("lmtest")

library("GGAlly")

library("car")

data = attitude

#1 Checking that there is no linear dependence using R^2

#1 Проверка отсутствия линейной зависимости с использованием R ^ 2

model1 <- lm(rating ~ privileges+learning+raises, data=data)

vif(model1)#the value of vif is 1.37,1.86,1.76. It shows there are not linear dependent

#значение vif равно 1,37,1,86,1,76. Это показывает, что существуют не линейно зависимые

#2 Checking how good the model is

#2 Проверка того, насколько хороша модель

summary(model1)#if p-value < 0.05, it is considered good/ если p-значение < 0,05, то оно считается хорошим

#r-squared is 0.45/r-square равен 0,45

#3 enter the log of the regressors into the model

#3 введите журнал регрессоров в модель

model2 <- lm(rating ~ log(privileges)+log(learning)+log(raises), data=data)

summary(model1)$r.squared#checking which model is good/ проверяем, какая модель хороша

summary(model2)$r.squared

#model2 is worse than model1 (R^2"0.454<0.458")/ модель 2 хуже модели 1 (R ^ 2"0.454<0.458")

#4 All possible products of pairs of regressors(including squares)

#4 Все возможные произведения пар регрессоров (включая квадраты)

attitude$privileges\_learning <- attitude$privileges \* attitude$learning

attitude$privileges\_raises <- attitude$privileges \* attitude$raises

attitude$learning\_raises <- attitude$learning \* attitude$raises

attitude$privileges\_sq <- attitude$privileges^2

attitude$learning\_sq <- attitude$learning^2

attitude$raises\_sq <- attitude$raises^2

model3 <- lm(rating ~ privileges + learning + raises + privileges\_learning + privileges\_raises + learning\_raises + privileges\_sq + learning\_sq + raises\_sq, data = attitude)

summary(model3)$r.squared#This model is good/ Эта модель хороша

summary(model2)$r.squared

summary(model1)$r.squared

* Команды “library("car")” “library("lmtest")” и “library("GGally")” загружают пакеты, установленные на вашем компьютере.
* Команда "data = attitude" изменяет набор данных attitude на именованный "data".
* “model1 <- lm(rating ~ privileges+learning+raises, data=data)”, “model2 <- lm(rating ~ log(privileges)+log(learning)+log(raises), data=data)” и “model3 <- lm(rating ~ privileges + learning + raises + privileges\_learning + privileges\_raises + learning\_raises + privileges\_sq + learning\_sq + raises\_sq, data = attitude)” соответствуют моделям линейной регрессии для младенцев.Смертность в качестве зависимой переменной и сельское хозяйство и обследование в качестве предикторов, соответственно.
* “attitude$privileges\_learning <- attitude$privileges \* attitude$learning”, “attitude$privileges\_raises <- attitude$privileges \* attitude$raises”, “attitude$learning\_raises <- attitude$learning \* attitude$raises”, “attitude$privileges\_sq <- attitude$privileges^2”, “attitude$learning\_sq <- attitude$learning^2” и “attitude$raises\_sq <- attitude$raises^2”: Эти строки создают новую переменную. например, строка “attitude$privileges\_learning <- attitude$privileges \* attitude$learning” создает новую переменную, которая является произведением “privileges” и “attitude$learning”.
* vif(model1) - Он показывает значение vif между "privileges", "learning" и "raises”. Если один из регрессоров показывает значение, превышающее 5, это означает высокую мультиколлинеарность. На данный момент все они меньше 5, так что это означает, что мультиколлинеарность между ними низкая. Так что мы можем это использовать.

> vif(model1)

privileges learning raises

1.373324 1.865749 1.761178

* summary(model1)

Call:

lm(formula = rating ~ privileges + learning + raises, data = data)

Residuals:

Min 1Q Median 3Q Max

-18.968 -5.559 0.305 6.360 18.031

Coefficients:

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

(Intercept) 14.1672 11.5195 1.230 0.2298

privileges 0.1046 0.1682 0.622 0.5396

learning 0.3936 0.2044 1.926 0.0651 .

raises 0.3516 0.2242 1.568 0.1289

---

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

Residual standard error: 9.458 on 26 degrees of freedom

Multiple R-squared: 0.4587, Adjusted R-squared: 0.3963

F-statistic: 7.345 on 3 and 26 DF, p-value: 0.00101

* p-value, которое представлено здесь, равно 0,001. Он ниже 0,05, поэтому его можно считать хорошим и значимым.
* Сначала "summary(model1)$r.squared" и "summary(model2)$r.squared" показывают R^2 value.

> summary(model1)$r.squared

[1] 0.4587139

> summary(model2)$r.squared

[1] 0.4547352

исходя из этих значений, мы можем видеть, что первое из них лучше второго.

* И теперь мы добавляем "summary(model3)$r.squared", чтобы узнать, какая из этих моделей лучше.

> summary(model3)$r.squared

[1] 0.6483089

> summary(model2)$r.squared

[1] 0.4547352

> summary(model1)$r.squared

[1] 0.4587139

Из "summary(model3)$r.в квадрате" мы получили значение 0.4587139, которое лучше, чем первые 2.

Практическая Работа Задание №2.2

Набор данных “attitude”в R был использован для создания модели линейной регрессии (model1), предсказывающей рейтинг работы на основе привилегий, повышений и возможностей обучения. В этом отчете представлены 95% доверительные интервалы для коэффициентов этой модели, а также для одного прогноза.

Необходимо загрузить данные из указанного набора и произвести следующие действия.

Набор Данных: Attitude

Объясняемая переменная: Rating

Регрессоры: Privileges, Learning, Raises

library("lmtest")

library("GGAlly")

data = attitude

#1 Confidence intervals for all coefficients in the model, p = 95%

#1 Доверительные интервалы для всех коэффициентов в модели, p = 95%

model1 <- lm(rating ~ privileges+raises+learning, data = data)

confint(model1, level = 0.95) #Get the confidence intervals for all coefficients

#2

summary(model1)

#3 Confidence interval for one forecast, p=95%

#3 Доверительный интервал для одного прогноза, p=95%

new.data = data.frame(privileges=8, raises=6, learning=4)

predict(model1, new.data, interval="confidence", level=0.95)

* Команды “library("lmtest")” и “library("GGally")” загружают пакеты, установленные на вашем компьютере.
* Команда "data = attitude" изменяет набор данных attitude на именованный "data".
* “model1 <- lm(rating ~ privileges+learning+raises, data=data)” соответствуют моделям линейной регрессии для младенцев.Смертность в качестве зависимой переменной и сельское хозяйство и обследование в качестве предикторов, соответственно.
* Функция confint(model1, level = 0.95) в R вычисляет доверительные интервалы для параметров подогнанной модели.

Когда вы передаете подогнанную модель и уровень достоверности в функцию confront(model1, level = 0.95), она вычисляет доверительные интервалы для каждого коэффициента в модели. Уровень достоверности определяет вероятность того, что интервал будет содержать истинное значение параметра.

Уровень достоверности 95% означает, что интервал содержит истинный коэффициент (вы можете быть уверены на 95%)

> confint(model1, level = 0.95)

2.5 % 97.5 %

(Intercept) -9.51142616 37.8457281

privileges -0.24121813 0.4503417

raises -0.10924298 0.8123659

learning -0.02651242 0.8137819

1. (Intercept): Расчетный коэффициент для перехвата находится в диапазоне от -9,51142616 до 37,8457281 с достоверностью 95%.
2. Privileges: Расчетный коэффициент для перехвата находится в диапазоне от -0.24121813 до 0.4503417 с достоверностью 95%.
3. Raises: Расчетный коэффициент для перехвата находится в диапазоне от -0.10924298 до 0.8123659 с достоверностью 95%.
4. Learning: Расчетный коэффициент для перехвата находится в диапазоне от -0.02651242 до 0.8137819 с достоверностью 95%.

* summary(model1)

Call:

lm(formula = rating ~ privileges + learning + raises, data = data)

Residuals:

Min 1Q Median 3Q Max

-18.968 -5.559 0.305 6.360 18.031

Coefficients:

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

(Intercept) 14.1672 11.5195 1.230 0.2298

privileges 0.1046 0.1682 0.622 0.5396

learning 0.3936 0.2044 1.926 0.0651 .

raises 0.3516 0.2242 1.568 0.1289

---

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

Residual standard error: 9.458 on 26 degrees of freedom

Multiple R-squared: 0.4587, Adjusted R-squared: 0.3963

F-statistic: 7.345 on 3 and 26 DF, p-value: 0.00101

* p-value, которое представлено здесь, равно 0,001. Он ниже 0,05, поэтому его можно считать хорошим и значимым.
* new.data = data.frame(privileges=8, raises=6, learning=4)

Он создает новый фрейм данных "new.data" с определенными значениями для "Privileges = 8", "Raises = 6" и "Learning = 4".

* predict(model1, new.data, interval="confidence", level=0.95)

Функция "predict()" в R используется для составления прогнозов на основе результатов статистической модели. В данном случае он используется с моделью линейной регрессии.

Код "predict(model1, new.data, interval="confidence", level=0,95)" предсказывает значение "rating" для точки "new.data", указанной в new.data, используя модель линейной регрессии "model1".

Аргумент "interval="confidence"" указывает функции вычислить доверительный интервал для прогноза. Аргумент "level=0.95" указывает, что это должен быть 95% доверительный интервал.

Проще говоря, функция оценивает, каким был бы "rating" для человека с "privileges =8", "raises = 6" и "learning =4" (как указано в "new.data"), основываясь на взаимосвязях в данных, которые были использованы для построения "model1". Затем функция выдает диапазон вероятных значений для этого прогноза (95% доверительный интервал), что означает, что мы можем быть на 95% уверены в том, что истинное значение "rating" находится в пределах этого диапазона.

> predict(model1, new.data, interval="confidence", level=0.95)

fit lwr upr

1 18.68755 -2.686112 40.06122

"fit": это точечная оценка или фактическое прогнозируемое значение зависимой переменной (в данном случае "rating") для заданных значений независимых переменных (privileges=8, raises =6 и learning=4).

"lwr" и "upr": это нижняя и верхняя границы 95% доверительного интервала для прогноза. Это означает, что мы можем быть на 95% уверены в том, что истинное значение "rating" для человека с "privileges=8", "raises =6" и "learning =4" будет находиться между значениями "lwr" и "upr".

Практическая Работа Задание №3

Вступление: В настоящем отчете представлены результаты линейного регрессионного анализа, проведенного на основе набора данных RLMS (Russia Longitudinal Monitoring Survey). Цель анализа состояла в том, чтобы определить, какие люди получают самую высокую заработную плату на основе выбранных параметров, и оценить модели для подмножества людей, указанных в данных.

Необходимо загрузить данные из указанного набора и произвести следующие действия.

Номер Волны Выборки РМЭК НИУ ВШЭ: 21

Подмножества для пункта 5: Городские жители, не состоящие в браке, разведенные женщины, с высшим образованием

install.packages("devtools")

devtools::install\_github("https://github.com/bdemeshev/rlms")

library("lmtest")

library("rlms")

library("dplyr")

library("GGally")

library("car")

library("sandwich")

library("foreign")

library("haven")

library("devtools")

df <- read\_sav("r21i\_os24a.sav")

data = select(df, qj13.2, qh5, q\_educ, status, qj6.2, q\_marst, qm148\_a)

names(df)

data2 = na.omit(data)#remove missing values/ удалить пропущенные значения

#qh5-gender, q\_educ education, status city status, qj13.2 salary, qj6.2 work time in a week,q\_marst marital status, qm148\_a age

data$wed1 <- as.numeric(data$q\_marst == 2)

data$wed2 <- as.numeric(data$q\_marst %in% c(4, 5))

data$wed3 <- as.numeric(data$q\_marst == 1)

data$city\_status <- as.numeric(data$status == 2)

data2$status[which(data2$status!=2)] <- 0

data2$status[which(data2$status==2)] <- 1

data$sex <- as.numeric(data$qh5 == 1)

data2$qh5[which(data2$qh5!=1)] <- 0

data2$qh5[which(data2$qh5==1)] <- 1

data$higher\_education <- as.numeric(data$q\_educ >= 21)

#1 Build linear regression and adding VIF/ Построение линейной регрессии и добавление VIF

lm\_model <- lm(qj13.2 ~ qh5+q\_educ+status+qj6.2+q\_marst+qm148\_a, data = data2)

vif(lm\_model)

#2 Experiment with the functions of real parameters: use logarithms, powers/ Поэкспериментируйте с функциями реальных параметров: используйте логарифмы, степени

data$log\_age <- log(data$qm148\_a)

data$salary\_age <- data$qj13.2 \* data$qm148\_a

data$sqrt\_age <- sqrt(data$qm148\_a)

data$qh5\_power\_0.1 <- data$qh5^0.1

data$q\_educ\_power\_0.1 <- data$q\_educ^0.1

data$qh5\_power\_0.2 <- data$qh5^0.2

data$q\_educ\_power\_0.2 <- data$q\_educ^0.2

data$qh5\_power\_0.3 <- data$qh5^0.3

data$q\_educ\_power\_0.3 <- data$q\_educ^0.3

data$qh5\_power\_0.4 <- data$qh5^0.4

data$q\_educ\_power\_0.4 <- data$q\_educ^0.4

data$qh5\_power\_0.5 <- data$qh5^0.5

data$q\_educ\_power\_0.5 <- data$q\_educ^0.5

data$qh5\_power\_0.6 <- data$qh5^0.6

data$q\_educ\_power\_0.6 <- data$q\_educ^0.6

data$qh5\_power\_0.7 <- data$qh5^0.7

data$q\_educ\_power\_0.7 <- data$q\_educ^0.7

data$qh5\_power\_0.8 <- data$qh5^0.8

data$q\_educ\_power\_0.8 <- data$q\_educ^0.8

data$qh5\_power\_0.9 <- data$qh5^0.9

data$q\_educ\_power\_0.9 <- data$q\_educ^0.9

data$qh5\_power\_1.0 <- data$qh5^1.0

data$q\_educ\_power\_1.0 <- data$q\_educ^1.0

data$qh5\_power\_1.1 <- data$qh5^1.1

data$q\_educ\_power\_1.1 <- data$q\_educ^1.1

data$qh5\_power\_1.2 <- data$qh5^1.2

data$q\_educ\_power\_1.2 <- data$q\_educ^1.2

data$qh5\_power\_1.3 <- data$qh5^1.3

data$q\_educ\_power\_1.3 <- data$q\_educ^1.3

data$qh5\_power\_1.4 <- data$qh5^1.4

data$q\_educ\_power\_1.4 <- data$q\_educ^1.4

data$qh5\_power\_1.5 <- data$qh5^1.5

data$q\_educ\_power\_1.5 <- data$q\_educ^1.5

data$qh5\_power\_1.6 <- data$qh5^1.6

data$q\_educ\_power\_1.6 <- data$q\_educ^1.6

data$qh5\_power\_1.7 <- data$qh5^1.7

data$q\_educ\_power\_1.7 <- data$q\_educ^1.7

data$qh5\_power\_1.8 <- data$qh5^1.8

data$q\_educ\_power\_1.8 <- data$q\_educ^1.8

data$qh5\_power\_1.9 <- data$qh5^1.9

data$q\_educ\_power\_1.9 <- data$q\_educ^1.9

data$qh5\_power\_2.0 <- data$qh5^2.0

data$q\_educ\_power\_2.0 <- data$q\_educ^2.0

lm\_model2 <- lm(qj13.2 ~ log\_age + qh5\_power\_0.1 + q\_educ\_power\_0.1 + q\_educ\_power\_0.2 + q\_educ\_power\_0.3 + q\_educ\_power\_0.4 + q\_educ\_power\_0.7 + q\_educ\_power\_1.2 + q\_educ\_power\_2.0, data = data)

vif(lm\_model2)

#3 Choose the best model from the last 2 models/ Выберите лучшую модель из последних 2-х моделей

summary(lm\_model)

summary(lm\_model2)#this model is better/Эта модель лучше

#4 Make a conclusion about which individuals receive the highest salary./ Сделайте вывод о том, какие люди получают самую высокую зарплату

coef(lm\_model)#If the coef of the lm\_model2 is positive, depending on education and age the salary increases./ Если коэффициент lm\_model2 положительный, то в зависимости от образования и возраста зарплата увеличивается.

#5 Subset the data for urban, unmarried/divorced women with higher education/ Подгруппируйте данные по городским незамужним/разведенным женщинам с высшим образованием

subset\_data <- filter(data, city\_status == 1 & wed1 == 0 & wed2 == 1 & higher\_education == 1)

lm\_subset\_data <- lm(qj13.2 ~ city\_status + q\_educ + q\_marst, data=subset\_data)

coef(lm\_subset\_data)

summary(lm\_subset\_data)

* “install.packages("devtools")

devtools::install\_github("https://github.com/bdemeshev/rlms")” - в данном случае мы устанавливаем пакет "devtools", чтобы загрузить и установить с "https://github.com/bdemeshev/rlms ".

* Команды “library("rlms")”, “library("dplyr")”, “library("car")”, “library("sandwich")”, “library("foreign")”, “library("devtools")”, “library("lmtest")” и “library("GGally")” загружают пакеты, установленные на вашем компьютере.

“df <- read\_sav("r21i\_os24a.sav")” - Первоначальный анализ данных включал загрузку данных из файла .sav, “data = select(df, qj13.2, qh5, q\_educ, status, qj6.2, q\_marst, qm148\_a)” - выбор соответствующих столбцов и “data2 = na.omit(data)” - удаление пропущенных значений.

* “data$wed1 <- as.numeric(data$q\_marst == 2)”

“data$wed2 <- as.numeric(data$q\_marst %in% c(4, 5))”

“data$wed3 <- as.numeric(data$q\_marst == 1)”

“data$city\_status <- as.numeric(data$status == 2)”

“data2$status[which(data2$status!=2)] <- 0”

“data2$status[which(data2$status==2)] <- 1”

“data$sex <- as.numeric(data$qh5 == 1)”

“data2$qh5[which(data2$qh5!=1)] <- 0”

“data2$qh5[which(data2$qh5==1)] <- 1”

“data$higher\_education <- as.numeric(data$q\_educ >= 21)” Двоичные переменные создаются для обозначения определенных условий, таких как семейное положение, статус города, пол и уровень высшего образования.

* “lm\_model <- lm(qj13.2 ~ qh5+q\_educ+status+qj6.2+q\_marst+qm148\_a, data = data2)” и “lm\_model2 <- lm(qj13.2 ~ log\_age + qh5\_power\_0.1 + q\_educ\_power\_0.1 + q\_educ\_power\_0.2 + q\_educ\_power\_0.3 + q\_educ\_power\_0.4 + q\_educ\_power\_0.7 + q\_educ\_power\_1.2 + q\_educ\_power\_2.0, data = data)” соответствуют моделям линейной регрессии для младенцев.Смертность в качестве зависимой переменной и сельское хозяйство и обследование в качестве предикторов, соответственно. Построены две модели линейной регрессии. Первая модель использует исходно выбранные параметры, и для обнаружения мультиколлинеарности вычисляется коэффициент инфляции дисперсии “vif(lm\_model)” (VIF). Вторая модель использует преобразования переменных (логарифмы, квадратные корни, степени) для изучения потенциальных нелинейных взаимосвязей между предикторами “vif(lm\_model 2)” и переменной отклика. Некоторые из новых управляемых переменных отсутствуют во 2-й модели. Поскольку некоторые из них являются "NA", из-за этого функция "vif" не работает. Вот почему некоторые из приведенных в действие переменных отсутствуют.

Vif(lm\_model)

qh5 q\_educ status qj6.2 q\_marst qm148\_a

1.059426 1.070041 1.083244 1.060705 1.078872 1.087967

Vif(lm\_model2)

log\_age qh5\_power\_0.1 q\_educ\_power\_0.1 q\_educ\_power\_0.2

2.623653e+00 1.737337e+00 -2.091677e+15 1.353989e+16

q\_educ\_power\_0.3 q\_educ\_power\_0.4 q\_educ\_power\_0.7 q\_educ\_power\_1.2

1.510617e+16 -1.635553e+16 -1.762447e+15 -4.379912e+13

q\_educ\_power\_2.0

-2.797968e+11

* Summary(lm\_model)

Call:

lm(formula = qj13.2 ~ qh5 + q\_educ + status + qj6.2 + q\_marst +

qm148\_a, data = data2)

Residuals:

Min 1Q Median 3Q Max

-29063544 -18459451 -10574973 -5013414 89407032

Coefficients:

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

(Intercept) 1.934e+06 3.416e+07 0.057 0.955

qh5 -3.320e+06 1.141e+07 -0.291 0.773

q\_educ 4.504e+05 1.518e+06 0.297 0.768

status 2.521e+06 5.458e+06 0.462 0.647

qj6.2 -1.713e-01 2.218e-01 -0.772 0.445

q\_marst -1.682e-01 3.786e-01 -0.444 0.659

qm148\_a 1.092e-01 1.121e-01 0.974 0.336

Residual standard error: 36040000 on 38 degrees of freedom

Multiple R-squared: 0.0503, Adjusted R-squared: -0.09965

F-statistic: 0.3354 on 6 and 38 DF, p-value: 0.9139

Summary(lm\_model2)

Call:

lm(formula = qj13.2 ~ log\_age + qh5\_power\_0.1 + q\_educ\_power\_0.1 +

q\_educ\_power\_0.2 + q\_educ\_power\_0.3 + q\_educ\_power\_0.4 +

q\_educ\_power\_0.7 + q\_educ\_power\_1.2 + q\_educ\_power\_2.0, data = data)

Residuals:

Min 1Q Median 3Q Max

-52674383 -16686654 -11101157 -420839 87484137

Coefficients:

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

(Intercept) 5.490e+16 1.020e+17 0.538 0.594

log\_age 4.582e+05 7.869e+05 0.582 0.564

qh5\_power\_0.1 -8.060e+07 1.624e+08 -0.496 0.623

q\_educ\_power\_0.1 -2.253e+17 4.205e+17 -0.536 0.595

q\_educ\_power\_0.2 3.606e+17 6.759e+17 0.533 0.597

q\_educ\_power\_0.3 -2.706e+17 5.096e+17 -0.531 0.599

q\_educ\_power\_0.4 8.253e+16 1.561e+17 0.529 0.600

q\_educ\_power\_0.7 -2.079e+15 3.986e+15 -0.522 0.605

q\_educ\_power\_1.2 2.335e+13 4.580e+13 0.510 0.613

q\_educ\_power\_2.0 -5.279e+10 1.073e+11 -0.492 0.626

Residual standard error: 36650000 on 35 degrees of freedom

(17045 observations deleted due to missingness)

Multiple R-squared: 0.09548, Adjusted R-squared: -0.1371

F-statistic: 0.4105 on 9 and 35 DF, p-value: 0.9209

Исходя из этого, мы можем видеть, что 2-я модель лучше. 0.0503 < 0.09548

* “coef(lm\_model)”

(Intercept) qh5 q\_educ status qj6.2

1.933764e+06 -3.320440e+06 4.504230e+05 2.520769e+06 -1.713234e-01

q\_marst qm148\_a

-1.682453e-01 1.091827e-01

Исходя из этого, мы можем видеть, что некоторые из них являются либо негативными, либо позитивными. Это означает, что если он положительный, то, исходя из этой переменной, зарплата будет выше.

* “subset\_data <- filter(data, city\_status == 1 & wed1 == 0 & wed2 == 1 & higher\_education == 1)” - используя эту функцию, мы фильтруем данные, необходимые нам для выполнения этой задачи.
* lm\_subset\_data <- lm(qj13.2 ~ city\_status + q\_educ + q\_marst, data=subset\_data)

coef(lm\_subset\_data)

(Intercept) city\_status q\_educ q\_marst

5.979899e+07 NA -1.169833e-01 -1.109719e+07

summary(lm\_subset\_data)

Call:

lm(formula = qj13.2 ~ city\_status + q\_educ + q\_marst, data = subset\_data)

Residuals:

Min 1Q Median 3Q Max

-15405127 -15394229 -15340229 -4293040 95686957

Coefficients: (1 not defined because of singularities)

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

(Intercept) 5.980e+07 3.359e+07 1.780 0.0794 .

city\_status NA NA NA NA

q\_educ -1.170e-01 1.848e-01 -0.633 0.5288

q\_marst -1.110e+07 7.661e+06 -1.449 0.1519

---

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

Residual standard error: 31340000 on 70 degrees of freedom

(54 observations deleted due to missingness)

Multiple R-squared: 0.03423, Adjusted R-squared: 0.006637

F-statistic: 1.241 on 2 and 70 DF, p-value: 0.2955