Титульный лист

# **Оглавление**

[Оглавление 2](#_Toc1661764554)

[Задача №1 3](#_Toc368671184)

[Заключение к заданию №1 3](#_Toc1273370547)

[Задача №2 4](#_Toc1874483863)

[Заключение к заданию№2 6](#_Toc1184044308)

[Список литературы 9](#_Toc642353833)

[Приложение 9](#_Toc221359621)

[Приложение к задаче №1 9](#_Toc211496003)

[Приложение к задаче №2 12](#_Toc1525210417)

# **Задача №1**

Необходимо загрузить данные из указанного набора и произвести следующие действия.

Набор данных: Swiss.

Объясняемая переменная: *Infant.Mortality.*

Регрессоры: *Agriculture, Examination*.

1. Оценить среднее значение, дисперсию и СКО переменных.

Используя встроенную функцию *mean*, *var*, *sd*, получаем

Таблица 1. Среднее значение, дисперсия и СКО *Agriculture, Examination* и *Infant.Mortality.*

|  |  |  |  |
| --- | --- | --- | --- |
| Переменная | Среднее значение | Дисперсия | СКО |
| *Agriculture* | 50.65957 | 515.7994 | 22.71122 |
| *Examination* | 16.48936 | 63.64662 | 7.977883 |
| *Infant.Mortality* | 19.94255 | 8.483802 | 2.912697 |

1. Построить зависимости вида y = a + bx, где y – объясняемая переменная, x – регрессор. Оценить, насколько «хороша» модель по коэффициенту детерминации R2. Оценить, есть ли взаимосвязь между объясняемой переменной и объясняющей переменной

Строим модель *Infant.Mortality ~ Agriculture.* Получаем уравнение вида

*Infant*.*Mortality* = 20.338 - 0.008 \* *Agriculure.* R2 равен 0.003704 (0,4%), количество звездочек у регрессора равно нулю. Это говорит о том, что между регрессором и объясняемым значением нет зависимости.

Теперь построим модель *Infant.Mortality~Examination.* Получаем уравнение вида

*Infant.Mortality = 20.62899 -0.04163 \* Examination*. R2 равен 0.013 (1,3 %), количество звездочек у регрессора равно нулю. Это указывает на то, что модель также не объясняет нашу переменную

Код решения задачи и сведения о проверенных моделях приведены в [Приложении к задаче 1](#Bookmark1)

# **Заключение к заданию №1**

1. Построенные модели не отображают зависимости детской смертности от процента мужского населения, занимающегося сельским хозяйством, и от процента людей, получивших максимальный балл на экзаменах при поступлении на военную службу.
2. Улучшить модели невозможно -- нужно менять регрессоры.
3. Из модели мы можем понять, что причиной низкой или высокой детской смертности может быть что-то другое, а не данные параметры.

# **Задача №2**

Необходимо загрузить данные из указанного набора и произвести следующие действия.

Набор данных: *attitude.*

Объясняемая переменная: *rating.*

Регрессоры: *complaints, learning , raises.*

1. Проверить регрессоры на линейную зависимость, построить модели для проверки на линейную зависимость и исключить те регрессоры, R^2 моделей которых высокий.

Строим модели *complaints~learning, complaints~raises, raises~learning*

Таблица 2. проверка на линейную зависимость регрессоров путем перебора моделей.

|  |  |  |  |
| --- | --- | --- | --- |
| Регрессор 1 | Регрессор 2 | R^2 | Pr |
| *complaints* | *learning* | 0.3561 | 0.00050 |
| *complaints* | *raises* | 0.4478 | 5.27e-05 |
| *raises* | *learning* | 0.41 | 0.000138 |

Второй способ: используем функцию *vif* для модели *rating ~ complaints + learning + raises*

Таблица 3. результат функции *vif*

|  |  |  |
| --- | --- | --- |
| *complaints* | *learning* | *raises* |
| 1.983336 | 1.856191 | 2.164556 |

Устанавливаем, что между *complaints* и *raises* присутствует небольшая линейная зависимость, поэтому мы можем исключить *raises* из расмотрения

1. Построить модель линейной регрессии между объясняемой переменной и регрессорами. Построить модели с использованием логарифмов и произведений регрессоров.

Строим модели между *rating* и различными комбинациями регрессоров, их произведения и логарифмов.

Таблица 4. результаты перебора моделей

|  |  |
| --- | --- |
| зависимость | R^2 |
| *rating ~ learning + complaints + raises* | 0.7083 |
| *rating ~ complaints* | 0.6813 |
| *rating ~ learning + complaints* | 0.708 |
| *rating ~ complaints + raises* | 0.6839 |
| *rating ~ learning + raises* | 0.4507 |
| *rating ~ I(log(complaints)) + learning* | 0.7047 |
| *rating ~ I(log(learning)) + complaints* | 0.7018 |
| *rating ~ I(log(complaints + learning))* | 0.6496 |
| *rating ~ I(log(complaints \* learning))* | 0.6439 |
| *rating ~ I(log(complaints)) + complaints + learning* | 0.7088\* |
| *rating ~ I(complaints \* learning)* | 0.6626 |
| *rating ~ I(learning \* raises) + complaints* | 0.7022 |
| *rating ~ I(complaints\*raises\*learning)* | 0.6057 |

\*есть линейная зависимость между регрессорами

Лучшей моделью по значению R^2 является зависимость *rating ~ learning + complaints.* К тому же мы исключили *raises*, который был частично линейно зависим от *complaints.*

Таблица 5. модель *rating ~ learning + complaints.*

|  |  |  |  |
| --- | --- | --- | --- |
|  | *Estimate* | *Std. Error* | *Pr(>|t|)* |
| *Intercept* | 9.8709 | 7.0612 | 0.174 |
| *learning* | 0.2112 | 0.1344 | 0.128 |
| *complaints* | 0.6435 | 0.1185 | 9.57e-06 |

1. Построить доверительный интервал для каждого регрессора лучшей модели и проанализировать его. Посчитать интервал для одного прогноза.

Находим критерий Стьюдента при помощи встроенной функции *qt*

Получаем критерий Стьюдента равный 2.048

Умножаем значение стандартной ошибки регрессора на значение критерия Стьюдента и прибавляем / вычитаем из значения коэффициента.

Получаем 2 промежутка:

[-0.0628;0.4852] - для *learning* (0 входит в промежуток, поэтому мы не можем отвергнуть статистическую гипотезу о том, что коэффициент равен 0)

[0.399788;0.887212] - для *complaints* (0 не входит в промежуток, так что мы можем отвергнуть статистическую гипотезу о том, что коэффициент равен 0)

Теперь посчитаем доверительный интервал для прогноза. Возьмем *complaints* = 60, *learning* = 55

Таблица 6. доверительный интервал

|  |  |  |
| --- | --- | --- |
| нижняя граница | прогноз модели | верхняя граница |
| 57.17933 | 60.09749 | 63.01565 |

Код и решение задания приведены в [Приложении к заданию 2.](#mark2)

# **Заключение к заданию№2**

1. Построенные модели отображают зависимость рейтинга от работы с жалобами сотрудников. Скорее всего, чем больше финансовая организация работает с жалобами своих сотрудников, тем лучше ее оценивают. Соответственно, чем лучше организация слушает своих сотрудников, тем больше и повышений, это объясняет линейную зависимость между *complaints* и *raises.*
2. Для улучшения модели возможно необходимо использовать еще другие регрессоры или же попробовать иные функции от уже имеющихся регрессоров (степень, деление и т.д)
3. Построен доверительный интервал, позволяющие довольно точно предсказать рейтинг организации.

# **Список** **литературы**

1

# **Приложение**

# Приложение к задаче №1

library("lmtest")

data = swiss

#используем набор данных swiss

#переменные:

# объясняемая: Infant.Mortality

# объясняющие: Agriculture | Examination

#задания:

#1) Оценить среднее значение, дисперсию и СКО переменных

mean(data$Infant.Mortality)

# 19.94255

mean(data$Agriculture)

# 50.65957

mean(data$Examination)

# 16.48936 процент людей занимающихся фермерским хозяйством больше, чем получивших высокие баллы на экзамене

var(data$Infant.Mortality)

# 8.483802 дисперсия мала, отклонение от среднего не велико

var(data$Agriculture)

# 515.7994 большая (первышает 100) дисперсия, отклонение от среднего может быть велико

var(data$Examination)

# 63.64662 средняя (меньше 100, больше 10) дисперсия...

sd(data$Infant.Mortality)

# 2.912697 СКО малое...

sd(data$Agriculture)

# 22.71122 СКО большое...

sd(data$Examination)

# 7.977883 СКО не большое...

#2) Построить зависимости вида y = a + bx 3) и 4) анализ моделей, оценка зависимости объясняемой переменной от объясняющих

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

summary(model1)

#Infant.Mortality = 20.338 - 0.008 \* Agriculure

#зависимость отрицательная

#Pr(agriculture) = 0.684 (>0.005)

#причинно-следственной связи нет

#Pr(Coef) <2e-16

#есть зависимость от коэф-та

#p-value: 0.6845 > 0.05 (слишком большой)

#Multiple R-squared: 0.003704, Adjusted R-squared: -0.01844

#зависимости нет. модель не отображает действительности

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

summary(model2)

#Infant.Mortality = 20.62899 -0.04163 \* Examination

# зависимость отрицательная

#Pr(Examination) = 0.445 (>0.005)

#причинно-следственной связи нет

#Pr(Coef) <2e-16

#есть зависимость от коэф-та

#p-value: 0.4454 > 0.05 (слишком большой)

#Multiple R-squared: 0.013, Adjusted R-squared: -0.008932

#зависимости нет. модель не отображает действительности

#Вывод:

#у Infant.Mortality нет реальных зависимостей от Examination и Agriculture

#детская смертность абосолютно не зависит от того, как были сданы экзамены, и от того, аграрный район это или нет

#причиной этого может быть то, что уровень медицины не отличался для людей работающих аграрных регионах и в армии

#Результаты:

#мы не получили моделей, способных предсказывать значение Infant.Mortality

#полученные модели (model1 | model2) не отображают действительности

# Приложение к задаче №2

library("lmtest")

library(car)  
   
#пакет attitude  
data = attitude  
#help(attitude)  
   
#переменные:  
# объясняемая  
# rating  
# регрессор   
# complaints | learning | raises  
   
mean(data$rating) #64.63333  
mean(data$complaints)#66.6  
mean(data$learning)#56.36667  
mean(data$raises)#64.63333  
   
var(data$rating)#148.1713  
var(data$complaints)#177.2828  
var(data$learning)#137.7575  
var(data$raises)#108.1023  
   
sd(data$rating)#12.17256  
sd(data$complaints)#13.31476  
sd(data$learning)#11.73701  
sd(data$raises)#10.39723  
   
#средние значение близки друг к другу  
#разброс значений большой  
   
#задача:  
#1) Проверьте, что в наборе данных нет линейной зависимости

#1. способ  
 model\_test\_1 = lm(complaints~learning,data)  
 summary(model\_test\_1)  
 #y = 28.442 + 0.677 \* x  
 # зависимость положительная  
 #pr = 0.00050  
 #R^2 = 0.3561  
 #p-value: 5e-04 < 0.05  
 #зависимость между регрессорами присутствует  
 #   
 model\_test\_2 = lm(complaints~raises,data)  
 summary(model\_test\_2)  
 #y = 11.21 + 0.857\*x  
 # зависимость положительная  
 #pr = 5.27e-05  
 #R^2 = 0.4478  
 #p-value: 5.268e-05 < 0.05  
 # зависимость между регрессорами присутствует  
 #  
   
 model\_test\_3 = lm(raises~learning,data)  
 summary(model\_test\_3)  
 #y = 32.6609 + 0.5672 \* x  
 # зависимость положительная  
 #pr = 0.000138  
 #R^2 = 0.41  
 #p-value: 0.0001384 < 0.05  
 #зависимость между регрессорами присутствует  
 #  
   
 #2. способ  
 model1 = lm(rating ~ complaints + learning + raises, data) #!  
 summary(model1)  
 vif(model1)  
 #rating = 10.523 + 0.653 \* complaints + 0.221 \* learning - 0.029 \* raises  
 #pr (complaints) = 5.82e-05 rating зависит от данной переменной  
 #pr (learning) = 0.152 rating не зависит от данной переменной  
 #pr (raises) = 0.876 rating не зависит от данной переменной  
 #p-value: 3.957e-07 < 0.05 | R^2 = 0.7083  
 #vif ~~ 2 для каждого регрессорра (есть зависимость)  
 #  
 #зависимость между регрессорами есть  
 ##регрессоры линейно выражаются друг через друга  
   
#2) Постройте линейную модель  
#mode1 показала, что rating лучше всего выражается через complaints  
#попробуем поменять местами регрессоры и/или убрать линейно выражающиеся  
#  
   
#уже найдена модель, отображающая зависимость rating от регрессоров (model1), проверим есть ли более точная модель  
   
model2 = lm(rating ~ learning + complaints + raises, data)  
summary(model2)  
#результат идентичный  
#rating лучше всего выражается через complaints  
#убираем линейновыражающиеся переменные  
   
model3 = lm(rating ~ complaints, data)  
summary(model3)  
#R^2 = 0.6813  
#p-value: 1.988e-08 < 0.05  
#модель стала менее точной (судя по R^2)  
#  
   
model4 = lm(rating ~ learning + complaints, data) #!  
summary(model4)  
# rating = 9.8709 + 0.2112 \* learning + 0.6435 \* complaints  
#зависимость положительная для обоих регрессоров  
#pr(learning) = 0.128 > 0.05  
#pr(complaints) = 9.57e-06 < 0.05  
#R^2 = 0.708  
#p-value: 6.058e-08  
#модель стала лучше по R^2, но p-value увеличился  
   
model5 = lm(rating ~ complaints + raises, data)  
summary(model5)  
#R^2 = 0.6839  
#p-value: 1.769e-07  
#модель хуже (судя по R^2)  
#модель значимо не отличается от модели 3, в будущем исключим аргумент raises  
   
modeljoke = lm(rating ~ learning + raises, data)  
summary(modeljoke)  
#R^2 = 0.4507  
#pr(learning) = 0.0333 < 0.05  
#pr(raises) = 0.0930 > 0.05  
#нет зависимости от данных переменных (что странно, так как у них есть связь с complaints)  
   
#3) Введем в модель логарифмы регрессоров  
   
#исключим raise, так как с этим аргументом модели практически не меняются  
model\_1 = lm(rating ~ I(log(complaints)) + learning, data)  
summary(model\_1)  
#R^2 = 0.7047 > 0.7  
#p-value: 7.06e-08 < 0.05  
#pr(I) = 1.12e-05 < 0.05  
#зависимость описывает rating  
#модель приемлима  
   
model\_2 = lm(rating ~ I(log(learning)) + complaints, data)  
summary(model\_2)  
#R^2 = 0.7018 > 0.7  
#p-value:8.034e-08 < 0.05  
#pr(I) = 0.184 > 0.05  
#зависимость описывает rating  
#модель приемлима, но log не улучшил результат  
   
model\_3 = lm(rating ~ I(log(complaints + learning)), data)  
summary(model\_3)  
#R^2 = 0.6496 < 0.7  
#p-value:7.204 7.68e-08 < 0.05  
#pr(I) = 7.676e-08 < 0.05  
#зависимость хуже описывает rating  
#модель приемлима, гораздо хуже предыдущих  
   
model\_4 = lm(rating ~ I(log(complaints \* learning)), data)  
summary(model\_4)  
#R^2 = 0.6439 < 0.7  
#p-value:9.636e-08 < 0.05  
#pr(I) = 9.64e-08 < 0.05  
#зависимость хуже описывает rating  
#модель приемлима, но такая же, как и \_3  
   
model\_5 = lm(rating ~ I(log(complaints)) + complaints + learning, data)  
summary(model\_5)  
#R^2 = 0.7088> 0.7  
#p-value: 3.866e-07 < 0.05  
#pr(I) = 0.791 < 0.05  
#зависимость приемлимо описывает rating  
#в модели нет высого уровня зависимости от переменных...  
#модель "плохая"  
   
#вывод: log от переменных не позволиил прийти к лучшей модели  
#  
   
#4) Введите в модель всевозможные произведения пар регрессоров.Найдите одну или несколько наилучших моделей по доле объяснённого разброса в данных R^2  
   
   
а  
   
model\_\_1 = lm(rating ~ I(complaints \* learning), data)  
summary(model\_\_1)  
#R^2 = 0.6626 < 0.7  
#p-value: 4.476e-08< 0.05  
#pr(I) = 4.48e-08 < 0.05  
#зависимость хуже описывает rating  
   
model\_\_2 = lm(rating ~ I(learning \* raises) + complaints , data)  
summary(model\_\_2)  
#R^2 = 0.7022 > 0.7  
#p-value: 7.896e-08 < 0.05  
#pr(I) = 0.1798 > 0.05  
#введение I не влияет на результат  
   
   
model\_\_3 = lm(rating ~ I(complaints\*raises\*learning), data)  
summary(model\_\_3)  
#R^2 = 0.6057 < 0.7  
#p-value:4.135e-07 < 0.05  
#pr(I) = 4.13e-07 < 0.05  
#зависимость плохо описывает rating  
#худшая модель с complaints  
   
#изучая зависимость rating от complaints, raises и learning, можно прийти к выводу  
#rating больше всего зависит от complaints и не зависит от raises  
#логарифмическая йункция и произведение регрессовров не улучшили показатели переменных  
#  
#лучшие модели - model1 и model4  
#R^2 > 0.7  
   
   
#теперь найдем доверительный интервал для model4 при p = 95%  
#p = 95%; p1 = p + (100 - p)/2 | p1 = 97.5%  
#p2 = 0.975  
#число замеров 30 | число регрессоров 2 => 30 - 2 = 28  
#найдем t-критерия Стьюдента  
t = qt(0.975,28)  
#t = 2.048 не лучший, но не худший результат  
   
#стандартная ошибка learning\_q = 0.134  
#стандартная ошибка complaints\_q = 0.119  
   
   
#доверительный интервал для learning = [0.2112 - 0.274;0.2112 + 0.274] = [-0.0628;0.4852]  
#0 принадлежит интервалу, так что мы не можем отвергнуть статистическую гипотезу о том, что коэффициент равен 0  
   
#доверительный интервал для complaints = [0.6435 - 0.119 \* 2.048; 0.6435 + 0.119 \* 2.048] = [0.399788;0.887212]  
#0 не принадлежит интервалу, так что мы можем отвергнуть статистическую гипотезу о том, что коэффициент равен 0  
   
#теперь посчитаем доверительный интервал для прогноза  
#возьмем complaints = 60, learning = 55  
new.data = data.frame(complaints = 60, learning = 55)  
predict(model4,new.data,interval = "confidence")  
#прогноз модели 60.09749  
#нижняя граница 57.17933  
#верхняя граница 63.01565

Приложение к задаче №3

1