Задача 1

Условие

Набор данных: swiss.

Объясняемая переменная: *Agriculture*.

Регрессоры: *Examination, Catholic.*

1. Оцените среднее значение, дисперсию и СКО переменных, указанных во втором и третьем столбце.
2. Постройте зависимости вида y = a + bx, где y – объясняемая переменная, x – регрессор (для каждого варианта по две зависимости).
3. Оцените, насколько «хороша» модель по коэффициенту детерминации R2?
4. Оцените, есть ли взаимосвязь между объясняемой переменной и объясняющей переменной (по значению p-статистики, «количеству звездочек» у регрессора в модели).

Решение

1. Для оценки среднего значения будем использовать команду mean, для оценки дисперсии – команду var, для оценки СКО – команду sd.   
   В результате выполнения команд имеем:

* Среднее значение *Agriculture* = 50.66
* Среднее значение *Examination* = 16.49
* Дисперсия *Agriculture* = 515.80
* Дисперсия *Examination* = 63.65
* СКО *Agriculture* = 22.71
* СКО *Examination* = 7.98

1. Для построения линейной зависимости используем команду lm. В результате выполнения команды для первой и второй модели имеем:
   * y = 67.243 - 1.511x для *Agriculture~Education*
   * y = 41.6728 + 0.2184x *Agriculture~Catholic*
2. Чтобы посмотреть R2 воспользуемся командой summary. В результате её выполнения видим, что R2 у первой модели = 0.41 – это значит, что модель 1 объясняет 41% колебаний переменной *Agriculture* – меньше половины, но довольно неплохо для одного регрессора. Для второй модели R2 = 0.16 – совсем плохой показатель.
3. P-статистика также показывается при выполнении команды summary. Для первой модели имеем очень низкие показатели p-статистики (3 звёздочки у каждого из параметров), что означает наличие сильной взаимосвязи между параметрами и объясняемой переменной. У второй модели регрессор *Catholic* имеет только 2 звёздочки (большее значение p-статистики) – это значит, что взаимосвязь между ним и *Agriculture* не такая сильная, но она всё равно есть.

Выводы

В первой модели есть причинно-следственная связь между поведением переменных, но она нелинейная и/или требует дополнительных регрессоров. То же самое, но в меньшей степени можно сказать и про вторую модель.

Приложение 1

Код решения задачи:

#КМБО-01-20, Гребнев Никита, вариант 5

#практическое задание 1

**library**("lmtest")

**library**("GGally")

data = swiss

#выводим данные

data

#для многострочных комментариев будут использоваться кавычки ""

"Пункт 1. Оцените среднее значение, дисперсию и СКО переменных, указанных во втором и

третьем столбце (Agriculture и Examination)"

#среднее значение:

print(paste("среднее значение Agriculture =", mean(data$Agriculture)))

print(paste("среднее значение Examination =", mean(data$Examination)))

#дисперсия

print(paste("дисперсия Agriculture =", var(data$Agriculture)))

print(paste("дисперсия Examination =", var(data$Examination)))

#СКО

print(paste("СКО Agriculture =", sd(data$Agriculture)))

print(paste("СКО Examination =", sd(data$Examination)))

"Пункт 2. Постройте зависимости вида y = a + bx, где y – объясняемая переменная, x –

регрессор."

model1 = lm(Agriculture~Education, data)

model2 = lm(Agriculture~Catholic, data)

model1 # y = 67.243 - 1.511x

model2 # y = 41.6728 + 0.2184x

"Пункт 3. Оцените, насколько «хороша» модель по коэффициенту детерминации R^2"

summary(model1) #R^2 = 0.409. R^2 относительно низкий, модель плохая. (Модель объясняет 40% колебаний Agriculture)

summary(model2) #R^2 = 0.1609 - Так как R^2 очень низкий, модель очень плоха. (Модель объясняет только 16% колебаний Agriculture)

"Пункт 4. Оцените, есть ли взаимосвязь между объясняемой переменной и объясняющей

переменной"

# В model1 у параметров по 3 звёздочки. Взаимосвязь x и y очень сильная.

# В model2 у параметров 2 и 3 звёздочки. Взаимосвязь достаточно сильная.

"Вывод"

# В model1 причинно-следственная связь между поведением переменных есть, но она нелинейная и/или требует дополнительных регрессоров

# В model2 причинно-следственная связь между поведением переменных есть, но она нелинейная и/или требует дополнительных регрессоров