Задача 1

Условие

Набор данных: swiss.

Объясняемая переменная: *Agriculture*.

Регрессоры: *Education, 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.
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. Оценим средние значения, дисперсии и СКО.   
   В результате имеем:

* Среднее значение *Agriculture* = 50.66
* Среднее значение *Catholic* = 41.14
* Среднее значение *Education* = 10.98
* Дисперсия *Agriculture* = 515.79
* Дисперсия *Catholic* = 1739.29
* Дисперсия *Education* = 92.45
* СКО *Agriculture* = 22.71
* СКО *Catholic* = 41.7
* СКО *Education* = 9.61

1. Построим линейные зависимости. В результате для первой и второй модели имеем:
   * A = -1.51\*ed + 67.24 для *Agriculture~Education*
   * A = 41.6728 + 0.2184\*cat для *Agriculture~Catholic*
2. Рассмотрим R2. Видим, что R2 у первой модели = 0.409 – это значит, что модель 1 объясняет 40% колебаний переменной *Agriculture* – меньше половины, нужны более сложные зависимости, но можно сделать вывод, что чем выше доля образованного населения, тем меньше людей занимаются обработкой земли. Для второй модели R2 = 0.16 – очень низкий показатель, делать вывод нецелесообразно.
3. P-статистика для первой модели имеем очень низкие показатели (3 звёздочки у каждого из параметров), что означает наличие четкой зависимости между параметрами и объясняемой переменной. У второй модели регрессор *Catholic* имеет только 2 звёздочки (большее значение p-статистики) – это значит, что зависимость между ним и *Agriculture* не такая сильная, но она всё ещё есть.

Выводы

Были построены две зависимости с объясняемой переменной Agriculture и разными объясняющими переменными: Education, Catholic. В первой модели есть причинно-следственная связь между поведением переменных, но она требует дополнительных регрессоров. Тоже самое, можно сказать и про вторую модель.

Приложение 1

Код решения задачи:

library("lmtest")

library("GGally")

library("car")

#Вариант 2 Бундуки Владислав Вячеславович; КМБО-06-20

data = swiss

help(swiss)

# \_1 - от Education

modele\_1 = lm(Agriculture~Education, data)

modele\_1

summary(modele\_1)

#Вывод:

# R^2 = 0.409 -> довольно низкий показатель, нужны более сложные зависимости

# A = -1.51\*ed + 67.24 -> чем выше доля образованного населения, тем меньше людей #занимаются обработкой земли

# Много звездочек => очень четкая зависимость проглядывается

plot(modele\_1) + abline(a = 67.24, b = -1.5, col = "red")

# \_2 - от catholic

modele\_2 = lm(Agriculture~Catholic, data)

modele\_2

summary(modele\_2)

#R^2 = 0.16 -> делать вывод - не целесообразно

# чем больше Католиков, тем больше людей занимаются обработкой земли

# Много звездочек => очень четкая зависимость проглядывается

plot(modele\_2) + abline(a = 41.67, b = 0.21, col = "red")

# 515.79 - большой разброс

var(data$Agriculture)

# 22.71 - СКО

sd(data$Agriculture)

# 50.66 - среднее значение

mean(data$Agriculture)

# 1739.29 - ОЧЕНЬ БОЛЬШОЙ РАЗБРОС

var(data$Catholic)

#41.7 - СКО

sd(data$Catholic)

# 41.14 - среднее значение

mean(data$Catholic)

# 92.45 - маленький разброс значений

var(data$Education)

# 9.61 - СКО

sd(data$Education)

# 10.98 - среднее значение

mean(data$Education)