Задача 2.1

Условие

Набор данных: swiss.

Объясняемая переменная: *Agriculture*.

Регрессоры: *Fertility, Catholic, Education*.

1. Проверьте, что в наборе данных нет линейной зависимости (построить зависимости между переменными, указанными в варианте, и проверить, что R2 в каждой из них невысокий). В случае, если R2 большой, один из таких столбцов можно исключить из рассмотрения.
2. Постройте линейную модель зависимой переменной от указанных в варианте регрессоров по методу наименьших квадратов (команда lm пакета lmtest в языке R). Оценить, насколько хороша модель, согласно: 1) R2, 2) p-значениям каждого коэффициента.
3. Введите в модель логарифмы регрессоров (если возможно). Сравнить модели и выбрать наилучшую.
4. Введите в модель всевозможные произведения пар регрессоров, в том числе квадраты регрессоров. Найдите одну или несколько наилучших моделей по доле объяснённого разброса в данных R2.

Решение

1. Проверим линейную регрессию *Fertility ~ Catholic, Education. VIF < 6 =>* нет линейной зависимости между регрессорами и их можно использовать вместе.
2. Построим модель *Agriculture ~ Fertility, Education, Catholic*, чтобы посмотреть R2 и p-статистику. В результате её выполнения видим:
   * R2 = 0.582
   * У Fertility p-характеристика выше, чем у Education и Catholic.
3. Чтобы улучшить показатели модели, попробуем ввести в модель логарифмы, не забывая проверять отсутствие линейной зависимости. Подробный код поиска наилучшей модели приведён в Приложении 1.

Из всех моделей с логарифмами лучшей оказалась *modele\_log\_5 = lm(Agriculture ~ log(Fertility) + Education + Catholic, data)*, имеющая R2 = 0.5821 – больше, чем у исходной модели.

1. Попробуем тогда ввести в модель всевозможные произведения пар регрессоров, не забывая проверять отсутствие линейной зависимости. Подробный код поиска наилучшей модели приведён в Приложении 1.

Наилучшей среди моделей оказалась *model\_fin\_6 = lm(Agriculture ~ I(Fertility^2) + I(Education^2) + I(Catholic^2) + I(Fertility\*Education) + I(Fertility\*Education\*Catholic), data)*, имеющая R2 = 0.5971.

Выводы

Лучшей моделью оказалась модель, *Agriculture ~ I(Fertility^2), I(Education^2), I(Catholic^2), I(Fertility\*Education), I(Fertility\*Education\*Catholic*, так как все значения VIF регрессоров < 6.

Приложение 1

Код решения задачи:

library("lmtest")

library("GGally")

library("car")

data = swiss

help(swiss)

#Вариант 2 Бундуки Владислав КМБО-06-20

#Объясняемая переменная Agriculture , регрессоры: Fertility, Education, Catholic

model = lm(Agriculture ~ Fertility + Education + Catholic, data)

# 1) проверка на линейную зависимость среди регрессоров

model\_t = lm(Fertility ~ Catholic + Education, data)

vif(model\_t)

#Education Catholic

#1.024247 1.024247

#проверим R^2

model\_1 = lm(Agriculture ~ Fertility + Education, data)

summary(model\_1)

#Arg = 85.7253 + (-0.2)\*Fer + (-1.7)\*Edu

#R^2 = 0.41 , p-характеристика = 6.707e-06

#R^2 низкий => нет строгой линейной зависимости

model\_2 = lm(Agriculture ~ Fertility + Catholic, data)

summary(model\_2)

#Arg = 16.7384 + (0.38)\*Fer + (0.16)\*Cat

#R^2 = 0.2 , p-характеристика = 0.008134

#R^2 очень низкий => переменные почти не связаны между собой

#p-характеристика высокая

model\_3 = lm(Agriculture ~ Education + Catholic, data)

summary(model\_3)

#Arg = 59.06 + (-1.4)\*Edu + (0.17)\*Cat

#R^2 = 0.5 , p-характеристика = 2.104e-07

#R^2 низкий => нет строгой линейной зависимости

# 2)Строим линейную модель по 3-ем регрессорам

model

summary(model)

#Agriculture = 117.2729 + (-0.78)\*Fertility + (-2.01)\*Edu + (0.26)\*Cat

#R^2 = 0.582 , p = 2.956e-08

#R^2 не высокий => нельзя сделать корректные выводы

# p значение низкое => реальные данные мало будут отличаться от данных модели

# р-характеристика у Fertility = (\*\*)

# р-характеристика у Education = (\*\*\*)

# р-характеристика у Catholic = (\*\*\*)

# 3)Введем в модель логарифмы регрессоров и сравним их для поиска лучшей

modele\_log\_1 = lm(Agriculture ~ log(Fertility) + log(Education) + log(Catholic), data)

summary(modele\_log\_1)

vif(modele\_log\_1)

#R^2 = 0.50 - модель ухудшилась

#Agriculture = 120.271 + (-8.717)\*Fertility + (-20.352)\*Edu + ( 3.426 )\*Cat

#log(Fertility) log(Education) log(Catholic)

#1.605793 1.490099 1.103041

modele\_log\_2 = lm(Agriculture ~ log(Fertility) + log(Education) + Catholic, data)

summary(modele\_log\_2)

vif(modele\_log\_2)

#R^2 = 0.5394 - модель ухудшилась

#Agriculture = 145.97625 + (-14.4779)\*Fertility + (-19.54853)\*Edu + (0.17028 )\*Cat

#log(Fertility) log(Education) Catholic

#1.656365 1.458331 1.184675

modele\_log\_3 = lm(Agriculture ~ log(Fertility) + Education + log(Catholic), data)

summary(modele\_log\_3)

vif(modele\_log\_3)

#R^2 = 0.5246 - модель ухудшилась

#Agriculture = 255.9569 + (-46.5763)\*Fertility + (-2.2245 )\*Edu + (5.5737 )Cat

#log(Fertility) Education log(Catholic)

#2.616193 2.426323 1.216809

modele\_log\_4 = lm(Agriculture ~ Fertility + log(Education) + log(Catholic), data)

summary(modele\_log\_4)

vif(modele\_log\_4)

#R^2 = 0.5018 - модель ухудшилась

#Agriculture = 97.0689 + (-0.1948)\*Fertility + (-20.7370)\*Edu + (3.6925)Cat

# Fertility log(Education) log(Catholic)

#1.626027 1.429427 1.180360

modele\_log\_5 = lm(Agriculture ~ log(Fertility) + Education + Catholic, data)

summary(modele\_log\_5)

vif(modele\_log\_5)

#R^2 = 0.5821 - модель улудшилась

#Agriculture = 286.99567 + (-52.62785)\*Fertility + (-2.13538)\*Edu + ( 0.24021)\*Cat

#log(Fertility) Education Catholic

#2.606621 2.255065 1.241080

modele\_log\_6 = lm(Agriculture ~ Fertility + Education + log(Catholic), data)

summary(modele\_log\_6)

vif(modele\_log\_6)

#R^2 = 0.5199 - модель ухудшилась

#Agriculture = 103.3117 + (-0.6653)\*Fertility + ( -2.1044)\*Edu + (5.8601)Cat

#Fertility Education log(Catholic)

#2.345216 2.060489 1.312589

modele\_log\_7 = lm(Agriculture ~ Fertility + log(Education) + Catholic, data)

summary(modele\_log\_7)

vif(modele\_log\_7)

#R^2 =0.5469 - модель ухудшилась

#Agriculture = 106.48306 + (-0.30771)\*Fertility + (-19.93094)\*Edu + (0.18471)Cat

# Fertility log(Education) Catholic

#1.690517 1.383282 1.277755

#Из всех моделей только у modele\_log\_5 улучшился R^2 на 0.001

# 4) Введем в модель всевозможные произведения пар регрессоров, в том числе и квадраты регрессоров

model\_fin\_1 = lm(Agriculture ~ Fertility + Education + Catholic + I(Fertility^2) + I(Education^2) + I(Catholic^2) + I(Fertility\*Education) + I(Fertility\*Catholic) + I(Education\*Catholic) + I(Fertility\*Education\*Catholic), data)

summary(model\_fin\_1) #R^2 = 0.649

vif(model\_fin\_1)

model\_fin\_2 = lm(Agriculture ~ Education + Catholic + I(Fertility^2) + I(Education^2) + I(Catholic^2) + I(Fertility\*Education) + I(Fertility\*Catholic) + I(Education\*Catholic) + I(Fertility\*Education\*Catholic), data)

summary(model\_fin\_2)#R^2 = 0.6438

vif(model\_fin\_2)

model\_fin\_3 = lm(Agriculture ~ Education + Catholic + I(Fertility^2) + I(Education^2) + I(Catholic^2) + I(Fertility\*Education) + I(Education\*Catholic) + I(Fertility\*Education\*Catholic), data)

summary(model\_fin\_3)#R^2 = 0.6347

vif(model\_fin\_3)

model\_fin\_4 = lm(Agriculture ~ Catholic + I(Fertility^2) + I(Education^2) + I(Catholic^2) + I(Fertility\*Education) + I(Education\*Catholic) + I(Fertility\*Education\*Catholic), data)

summary(model\_fin\_4)#R^2 = 0.6239

vif(model\_fin\_4)

model\_fin\_5 = lm(Agriculture ~ I(Fertility^2) + I(Education^2) + I(Catholic^2) + I(Fertility\*Education) + I(Education\*Catholic) + I(Fertility\*Education\*Catholic), data)

summary(model\_fin\_5)#R^2 = 0.6

vif(model\_fin\_5)

model\_fin\_6 = lm(Agriculture ~ I(Fertility^2) + I(Education^2) + I(Catholic^2) + I(Fertility\*Education) + I(Fertility\*Education\*Catholic), data)

#Agriculture = 7.919e+01 + (-3.195e-03)\*Fertility^2 + (-8.119e-03)\*Edu^2 + (2.724e-03)Cat^2 #+ (-2.640e-02)\*(Fertility\*Education) + (-6.148e-05)\*(Fertility\*Education\*Catholic)

summary(model\_fin\_6)#R^2 = 0.5971

vif(model\_fin\_6)

#Вывод: Agriculture = 7.919e+01 + (-3.195e-03)\*Fertility^2 + (-8.119e-03)\*Edu^2 + (2.724e-#03)Cat^2 + (-2.640e-02)\*(Fertility\*Education) + (-6.148e-05)\*(Fertility\*Education\*Catholic)

#это лучшая модель ,т.к. VIF < 6