Задача 2.1

Условие

Набор данных: swiss.

Объясняемая переменная: *Infant.Mortality*.

Регрессоры: *Fertility, Catholic, Agriculture*.

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. Чтобы проверить отсутствие линейной зависимости, построим для каждого регрессора модели, в качестве параметров которых будут другие регрессоры, и оценим в них R2:

* *Fertility~Catholic* – R2 < 0.25 => зависимости нет
* *Fertility~Agriculture* – R2 < 0.25 => зависимости нет
* *Catholic~Agriculture* – R2 < 0.25 => зависимости нет

Линейной зависимости нет, значит регрессоры можно использовать вместе.

1. Построим модель, использую команду lm. Чтобы Посмотреть R2 и p-статистику, воспользуемся командой summary. В результате её выполнения видим:
   * R2 = 0.22
   * У *Catholic* и *Agriculture* высокое p-значение у (мало звёздочек)

Уберём из модели регрессор *Catholic*, как наименее значимый, и проверим, как изменится R2:

* R2 = 0.22 – не изменился
* У *Agriculture* всё ещё высокое p-значение (мало звёздочек)

Попробуем убрать *Agriculture*:

* R2 = 0.17 – достаточно существенное изменение - *Agriculture* лучше не убирать

Остановимся на *model = lm(Infant.Mortality ~ Fertility + Agriculture, data)*. Она имеет достаточно низкий R2, и не самые лучшие показатели p-статистики.

1. Чтобы улучшить показатели модели, попробуем ввести в модель логарифмы, не забывая проверять отсутствие линейной зависимости командой vif. Подробный код поиска наилучшей модели приведён в Приложении 1.

Из всех моделей с логарифмами лучшей оказалась *model = lm(Infant.Mortality ~ I(log(Fertility)) + Agriculture, data*), имеющая R2 = 0.21 – меньше, 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 = lm(Infant.Mortality ~ Fertility + I(Agriculture^2), data)*, имеющая R2 = 0.25.

Выводы

Лучшей моделью оказалась модель, зависящая от *Fertility и Agriculture2*, причём имеет место быть положительная зависимость *Infant.Mortality*  от *Fertility* и отрицательная от *Agriculture2*.

Приложение 1

Код решения задачи:

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

# Практическое задание 2

**library**("lmtest")

**library**("GGally")

**library**("car") # без этого не работает функция vif()

# При чтении избавляемся от записей с недостающими данными.

data = na.omit(swiss)

# Выводим данные

data

# Infant.Mortality ~ Fertility, Catholic, Agriculture

# 1. Проверим отсутствие зависимости между регрессорами перед построением модели

pairwise\_comparison\_1 = lm(Fertility~Catholic, data)

summary(pairwise\_comparison\_1) # R^2 < 25% - зависимости нет

pairwise\_comparison\_1 = lm(Fertility~Agriculture, data)

summary(pairwise\_comparison\_1) # R^2 < 15% - зависимости нет

pairwise\_comparison\_1 = lm(Catholic~Agriculture, data)

summary(pairwise\_comparison\_1) # R^2 < 20% - зависимости нет

# Можно использовать регрессоры вместе

# 2. Построим линейную модель и оценим её

model = lm(Infant.Mortality ~ Fertility + Catholic + Agriculture, data)

summary(model)

# R^2 ~ 0.22, p-значение у Catholic и Agriculture высокие (мало звёздочек) - модель плохая по обоим параметрам.

#Уберём из модели регрессор Catholic, как наименее значимый, и проверим, как изменится R^2

model = lm(Infant.Mortality ~ Fertility + Agriculture, data)

summary(model)

# R^2 ~ 0.22 - R^2 практически не изменился

# Попробуем убрать ещё один регрессор

model = lm(Infant.Mortality ~ Fertility, data)

summary(model) #R^2 ~ 0.17 - изменился сильно, регрессор Agriculture лучше не убирать

# В дальнейшем будем работать с моделью:

model = lm(Infant.Mortality ~ Fertility + Agriculture, data)

# 3. Попробуем ввести в модель логарифмы регрессоров, предварительно проверяя, что нет линейной зависимости

model = lm(I(log(Infant.Mortality)) ~ I(log(Fertility)) + I(log(Agriculture)), data)

vif(model) # линейной зависимости нет.

summary(model) # R^2 ~ 0.17, p-статистика хорошая для I(log(Fertility))

model = lm(Infant.Mortality ~ I(log(Fertility)) + I(log(Agriculture)), data)

vif(model) # линейной зависимости нет.

summary(model) # R^2 ~ 0.19, p-статистика хорошая для I(log(Fertility))

model = lm(Infant.Mortality ~ I(log(Fertility)) + Agriculture, data)

vif(model) # линейной зависимости нет.

summary(model) # R^2 ~ 0.21, p-статистика хорошая для I(log(Fertility))

model = lm(Infant.Mortality ~ Fertility + I(log(Agriculture)), data)

vif(model) # линейной зависимости нет.

summary(model) # R^2 ~ 0.19, p-статистика хорошая для I(log(Fertility))

# Наилучшей из них будет следующая модель:

model = lm(Infant.Mortality ~ I(log(Fertility)) + Agriculture, data) # R^2 ~ 0.21

# но при этом модели с логарифмами имею R^2 меньше, чем у модель без логарифмов

# 4. Попробуем ввести в модель всевозможные произведения пар регрессоров, предварительно проверяя, что нет линейной зависимости

model = lm(Infant.Mortality ~ Fertility + Agriculture + I(Fertility^2) + I(Agriculture^2) + I(Fertility\*Agriculture), data)

vif(model) # есть линейная зависимость, уберём регрессоры с максимальным VIF

model = lm(Infant.Mortality ~ Fertility + Agriculture + I(Agriculture^2) + I(Fertility\*Agriculture), data)

vif(model) # есть линейная зависимость, уберём регрессоры с максимальным VIF

model = lm(Infant.Mortality ~ Fertility + Agriculture + I(Agriculture^2), data)

vif(model) # есть линейная зависимость, уберём регрессоры с максимальным VIF

model = lm(Infant.Mortality ~ Fertility + I(Agriculture^2), data)

vif(model) # линейной зависимости нет

summary(model) # R^2 ~ 0.24, p-статистика хорошая для Fertility и неплохая для Agriculture^2

#Наилучшая модель:

model = lm(Infant.Mortality ~ Fertility + I(Agriculture^2), data)