Задача 2

Условие

Набор данных: *attitude*

Объясняемая переменная: *rating*

Регрессоры: *raises, critical, advance*

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 отсутсвие линейной зависимости между регрессорами, для этого построим модель для каждлой пары регрессоров и оценим ее по параметру R2..
   * *raises~critical –* R2 = 14% => нет зависимости
   * *raises~advance –* R2 = 33% => нет зависимости
   * *advance~critical –* R2 = 8% => нет зависимости
2. Построим модель с помощью функции lm(). Используя функцию summary() изучим R2 и p-статистику. Результат выполнения программы:
   * R2 ~ 0,4
   * *advance* и *catholic* имеют высокое p-значение (нет звездочек)
   * *raises* имеет очень низкое p-значение (3 звездочки)

Уберем из изначальной модели регрессор *critical* как незначительный.

* + R2 ~ 0,4 – не изменился
  + p-статистика не притерпела значительного изменения

Уберем из изначальной модели регрессор *advance* как незначительный.

* + R2 ~ 0,35 – снизился
  + p-статистика не притерпела значительного изменения

Уберем из изначальной модели регрессоры *critical* и *advance* как незначительные.

* + R2 ~ 0,35 – снизился
  + p-статистика не притерпела значительного изменения

Модель *model = lm(rating ~ raises + advance, data)* показала наибольшую эффективность. Она имеет низкий R2 и плохие p-значения для перменной *advance* и y-пересечения, но отличное p-значение для переменной *raises .*

1. Введем в модель логарифмы регрессоров при этом проверяя наличие линейной зависимости между ними с помощью функции vif(). Подробный код решения со всеми возможными вариантами приведен в Приложении 1.

Модель *model = lm(I(log(rating)) ~ I(log(raises)) + I(log(advance)), data)* оказалась наиболее эффективной. Ее R2 немного повысилось до ~ 0,42, p-статистика не изменилась, а корреляцией между регрессорами можно пренебречь (функция *vif()* показала значения в пределах (1; 5) для каждого регрессора).

1. Введем в модель произведения пар регрессоров и квадраты регрессоров, при этом проверяя наличие линейной зависимости между ними с помощью функции *vif()*. Подробный код решения со всеми вариантами приведен в Приложении 1.

Наилучшие результаты показали две модели:

*model = lm(I(rating) ~ I(raises) + I(advance^2), data)*

*model = lm(I(rating) ~ I(raises) + I(advance), data)*

Их результаты практически одинаковы: R2 ~ 0,4, низкое p-значение *raises* (3 звездочки) и высокое p-значение *advance* и *advance2.*

Выводы

Наилучшие результаты показала модель с логарифмированными регрессорами *raises* и *advance*. *rating* имеет положительную зависимость от *raises* и отрицательную от *advance, advance2*

Приложение 1





