Задача 2.1

Необходимо загрузить данные из указанного набора и произвести следующие действия.

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.

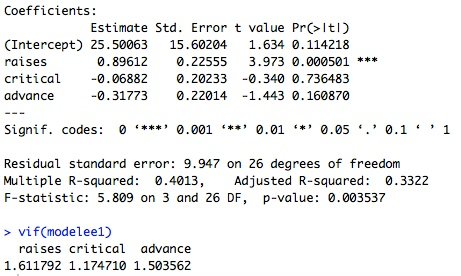
Проверим наличие зависимости между переменными. Raises~Critical , R^2 = 14,2% не высокий, нелинейная зависимость, звезд мало, значит нужно использовать регрессоры вместе, т.к. между ними нет почти никакой зависимости.

Raises~advance , R^2 = 32,97% не высокий, нелинейная зависимость, звёзд много, значит нужно использовать регрессоры аккуратно, т.к. между ними есть определённая зависимость.

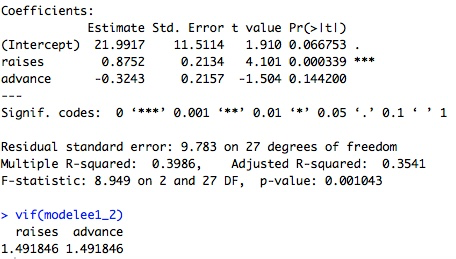
Critical~advance , R^2 = 8% низкий, нелинейная зависимость, звезд мало, значит можно использовать регрессоры вместе, т.к. между ними нет определенной зависимости.

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.

Построим множественную линейную регрессию rating~raises+critical+advance, R2 = 40,13%, значение p статистики у параметров примерно одинаковое, при исключении парметра critical R2 упадёт < 1%, что незначительно( допустимо 5%). Можем сделать вывод, что переменная critical незначимая.



rating~raises+advance



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 R^2 .

Добавим raises^2: rating~raises+critical+advance+I(raises^2) . R^2 = 41,72%

Добавим critical^2: rating~raises+critical+advance+I(critical^2). R^2 = 47,22%

Добавим advance^2: rating~raises+critical+advance+I(advance^2). R^2 = 40,04%

Добавим I(raises\*critical):

rating~raises+critical+advance+I(raises\*critical). R^2 = 43,4%

Добавим I(raises\*advance):

rating~raises+critical+advance+I(raises\*advance):. R^2 = 40,73%

Добавим I(critical\*advance):

rating~raises+critical+advance+I(critical\*advance). R^2 = 40,11%

Добавим I(raises\*critical\*advance)

rating~raises+critical+advance+I(raises\*critical\*advance). R^2 = 42,63%

Добавим raises\_log:

rating~raises+critical+advance+raises\_log. R^2 = 41,27%

Добавим critical\_log:

rating~raises+critical+advance+critical\_log. R^2 = 46,76%

Добавим advance\_log:

rating~raises+critical+advance+advance\_log. R^2 = 40,01%

Исходя из p-статистики и R^2 регрессоров модели model4\_2 можно сделать вывод что эта модель наилучшая.

Код решения задачи и сведения о проверенных моделях приведены в Приложении 1.

Задача 2.2

Необходимо загрузить данные из указанного набора и произвести следующие действия.

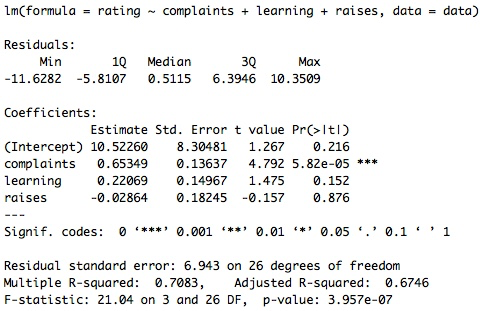
Набор данных: attitude.

Объясняемая переменная: rating.

Регрессоры: complaints, learning, raises.

1. Доверительные интервалы для всех коэффициентов в модели, p = 95%.

Данные о нашей модели.



2. Сделайте вывод о отвержении или невозможности отвергнуть статистическую гипотезу о том, что коэффициент равен 0.

Количество степеней свободы / измерений в обучающей выборке равно 30. Рассчитано 4 коэффициентов. Число степеней свободы в модели: 30 – 4 = 26. Для такого числа степеней свободы и p = 95% значение t-критерия Стьюдента равно 2.055529.

Итог:

1.1 [0.65349-2.055529\*0.13637, 0.65349 +2.055529\*0.13637] = [0.37316545 0.9338110] -> не попадает в 0

1.2 [0.22069-2.055529\*0.14967, 0.22069+2.055529\*0.14967] = [-0.08695739 0.5283289] -> попадает в 0

1.3 [-0.02864-2.055529\*0.18245, -0.02864+2.055529\*0.18245] = [-0.40366218 0.3463882] -> попадает в 0

Делаем вывод о том, что нельзя отвергнуть гипотезу, что этот коэффициент может быть равен 0, на уровне значимости 5%.

3. Доверительный интервал для одного прогноза (p = 95%, набор значений регрессоров выбираете сами).

Прогноз с доверительным интервалом

Возьму median (средние) значения из таблицы, поэтому у прогноза получился маленький интервал:

new.data = data.frame(complaints = 65, learning = 57, raises = 64)

predict(model7, new.data, interval = "confidence")

fit lwr upr

63.74566 61.08575 66.40557

Доверительный интервал [63.74566, 66.40557]

Возьму для проверки близкие значения и предскажу результат:

new.data = data.frame(complaints = 60, learning = 60, raises = 70)

predict(model7, new.data)

Прозноз 60.96845

Код решения задачи и сведения о проверенных моделях приведены в Приложении 1.

Приложение 1

library("lmtest")

library("GGally")

library("car")

data=attitude

summary(data)

ggpairs(data)

#2.1

#1 ПУНКТ

#проверим отсутствие зависимости между регрессорами перед построением модели

test\_model1 = lm(raises~critical, data)

test\_model2 = lm(raises~advance, data)

test\_model3 = lm(critical~advance, data)

test\_model4 = lm(raises~advance+critical,data)

test\_model5 = lm(critical~raises+advance,data)

test\_model6 = lm(advance~raises+critical,data)

summary(test\_model1) #R^2=0.142 зависимость между регрессорами raises

# и critical маленькая

summary(test\_model2) #R^2=0.3297 зависимость между регрессорами raises

# и advance маленькая

summary(test\_model3) #R^2=0.08028 зависимость между регрессорами critical

# и advance маленькая

summary(test\_model4) #R^2=0.3796 зависимость маленькая 3 регрессоров вместе

summary(test\_model5) #R^2=0.1487

summary(test\_model6) #R^2=0.3349

#2 ПУНКТ

modelee1 = lm(rating~raises+critical+advance, data)

modelee1

summary(modelee1) #R^2=0.4013/ p-value=0.003537

vif(modelee1)

modelee1\_2 = lm(rating~raises+advance, data)

modelee1\_2

summary(modelee1\_2)

vif(modelee1\_2)

#R^2=0.3986, упал незначильно, значит можем убрать эту переменную

#модель плохая, так как нет даже половины от 0.8

#3 ПУНКТ

data = attitude

data = as.data.frame(data)

data["raises\_log"] = log(data["raises"])

data["critical\_log"] = log(data["critical"])

data["advance\_log"] = log(data["advance"])

data

#4)

modelee4\_1 = lm(rating~raises+critical+advance+I(raises^2), data)

modelee4\_1

summary(modelee4\_1) #R^2=0.4172

vif(modelee4\_1)

modelee4\_2 = lm(rating~raises+critical+advance+I(critical^2), data)

modelee4\_2

summary(modelee4\_2) #R^2=0.4722

vif(modelee4\_2)

modelee4\_3 = lm(rating~raises+critical+advance+I(advance^2), data)

modelee4\_3

summary(modelee4\_3) #R^2=0.4004

vif(modelee4\_3)

modelee4\_4 = lm(rating~raises+critical+advance+I(raises\*critical), data)

modelee4\_4

summary(modelee4\_4) #R^2=0.434

vif(modelee4\_4)

modelee4\_5 = lm(rating~raises+critical+advance+I(raises\*advance), data)

modelee4\_5

summary(modelee4\_5) #R^2=0.4073

vif(modelee4\_5)

modelee4\_6 = lm(rating~raises+critical+advance+I(critical\*advance), data)

modelee4\_6

summary(modelee4\_6) #R^2=0.4011

vif(modelee4\_6)

modelee4\_7 = lm(rating~raises+critical+advance+I(raises\*critical\*advance), data)

modelee4\_7

summary(modelee4\_7) #R^2=0.4263

vif(modelee4\_7)

modelee4\_8 = lm(rating~raises+critical+advance+raises\_log, data)

modelee4\_8

summary(modelee4\_8) #R^2=0.4127

vif(modelee4\_8)

modelee4\_9 = lm(rating~raises+critical+advance+critical\_log, data)

modelee4\_9

summary(modelee4\_9) #R^2=0.4676

vif(modelee4\_9)

modelee4\_10 = lm(rating~raises+critical+advance+advance\_log, data)

modelee4\_10

summary(modelee4\_10) #R^2=0.4001

vif(modelee4\_10)

#2.2

model1=lm(rating~complaints+privileges+learning+raises+critical+advance,data)

model1

vif(model1)

summary(model1) #R^2=0.6831, видим, что rating зависит хорошо только от complaints

#посмотрим все возможные вариианты с этим регрессором.

model2=lm(rating~complaints+privileges,data)

model2

summary(model2) #R^2=0.6831, не изменился.

model3=lm(rating~complaints+learning,data)

model3

summary(model3) #R^2=0.708, чуть повысился

model4=lm(rating~complaints+raises,data)

model4

summary(model4) #R^2=0.6839, повысился незначительно

model5=lm(rating~complaints+critical,data)

model5

summary(model5) #R^2=0.6813, уменьшился

model6=lm(rating~complaints+advance,data)

model6

summary(model6) #R^2=0.6823, уменьшился

model7=lm(rating~complaints+learning+raises,data)

model7

summary(model7) #R^2=0.7083. Максимально хорошая модель, хоть при добавлении

# learning и raises R^2 увеличивается меньше, чем на 5%.

#коэфиценты

coef(model7)

confint(model7, level=0.9)

#ошибка se = sqrt (элемент ковариационной матрицы на главной диагонали или есть в таблице summary модели)

#se = 0.13637

se=0.13637

#критерий Стьюдента: 95%, 26 степени свободы

t\_critical=qt(0.975, df = 26) #всего 30, используем 4, значит 30-4

model7$coefficients[2] - t\_critical \* se

model7$coefficients[2] + t\_critical \* se

#проверка

confint(model7, level = 0.95)

#прогноз с доверительным интервалом

#я взял median (средние) значения из таблицы, поэтому у прогноза получился маленький интервал

new.data = data.frame(complaints = 65, learning = 57, raises = 64)

predict(model7, new.data, interval = "confidence")

# fit lwr upr

# 63.74566 61.08575 66.40557

#возьму для проверки близкие значения и предскажу результат

new.data = data.frame(complaints = 60, learning = 60, raises = 70)

predict(model7, new.data)

#прозноз 60.96845

#можно попробовать более сложную зависимость

model8=lm(rating~complaints+I(learning^2)+raises,data)

model8

summary(model8) #R^2=0.7134

model9=lm(rating~complaints+I(learning^2)+I(raises^2),data)

model9

summary(model9) #R^2=0.7136

#проверю, изменятся ли данные с более лучшей моделью

t\_critical=qt(0.975, df = 26) #всего 30, используем 4, значит 30-4

se2=0.1324543

model7$coefficients[2] - t\_critical \* se2 #0.3812245

model7$coefficients[2] + t\_critical \* se2 #0.9257519

#1ый раз я ошибся и взял на 3 степени свободы меньше, но сейчас я посчитал

#на 4 степени свободы меньше. Что странно, интервал изменился но остальные

#значения не изменились.

#я взял median (средние) значения из таблицы, поэтому у прогноза получился маленький интервал

new.data = data.frame(complaints = 65, learning = 57, raises = 64)

predict(model9, new.data, interval = "confidence")

# fit lwr upr

# 63.51763 61.90358 66.13168

#видно, что наш интервал заметно сузился, по сравнения с model7

#возьму для проверки близкие значения и предскажу результат

new.data = data.frame(complaints = 60, learning = 60, raises = 70)

predict(model9, new.data)

#прозноз 60.79247