Dr. Sudhakar Raju

BUS 6309: LINEAR & MULTIVARIATE MODELS

SPRING 2017

**ANSWERS TO ASSIGNMENT 2**

**DESCRIPTION OF BOSTON HOUSING DATA SET:** Boston Housing Values in Suburbs of Boston

The Boston data frame has 506 rows and 14 columns. This data frame contains the following columns:

crim: per capita crime rate by town.

Zn: proportion of residential land zoned for lots over 25,000 sq.ft.

indus: proportion of non-retail business acres per town.

chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

nox: nitrogen oxides concentration (parts per 10 million).

rm: average number of rooms per dwelling.

Age: proportion of owner-occupied units built prior to 1940.

Dis: weighted mean of distances to five Boston employment centers.

rad: index of accessibility to radial highways.

tax: full-value property-tax rate per $10,000.

ptratio: pupil-teacher ratio by town.

black: proportion of blacks by town.

lstat: lower status of the population (percent).

medv: median value of owner-occupied homes in $1000s.

I. a.)

Call:

lm(formula = crim ~ medv)

Residuals:

Min 1Q Median 3Q Max

-9.071 -4.022 -2.343 1.298 80.957

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 11.79654 0.93419 12.63 <2e-16 \*\*\*

medv -0.36316 0.03839 -9.46 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.934 on 504 degrees of freedom

Multiple R-squared: 0.1508, Adjusted R-squared: 0.1491

F-statistic: 89.49 on 1 and 504 DF, p-value: < 2.2e-16

See simple linear regression above. Medv is extremely significant in explaining Crime.

b.)

> anova(REGMODEL1)

Analysis of Variance Table

Response: crim

Df Sum Sq Mean Sq F value Pr(>F)

medv 1 5634 5633.6 89.486 < 2.2e-16 \*\*\*

Residuals 504 31730 63.0

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

REGRESSION SUM OF SQUARES = 5634

RESIDUAL (ERROR) SUM OF SQUARES = 31,730

TOTAL SUM OF SQUARES (TSS) = 37,364

R2 = [REGRESSION SUM OF SQUARES / TSS]

= [5634 / 37,364]

= .1508 (same as reported in part (a).

c.)

Call:

lm(formula = crim ~ ., data = Boston)

Residuals:

Min 1Q Median 3Q Max

-9.924 -2.120 -0.353 1.019 75.051

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 17.033228 7.234903 2.354 0.018949 \*

zn 0.044855 0.018734 2.394 0.017025 \*

indus -0.063855 0.083407 -0.766 0.444294

chas -0.749134 1.180147 -0.635 0.525867

nox -10.313535 5.275536 -1.955 0.051152 .

rm 0.430131 0.612830 0.702 0.483089

age 0.001452 0.017925 0.081 0.935488

dis -0.987176 0.281817 -3.503 0.000502 \*\*\*

rad 0.588209 0.088049 6.680 6.46e-11 \*\*\*

tax -0.003780 0.005156 -0.733 0.463793

ptratio -0.271081 0.186450 -1.454 0.146611

black -0.007538 0.003673 -2.052 0.040702 \*

lstat 0.126211 0.075725 1.667 0.096208 .

medv -0.198887 0.060516 -3.287 0.001087 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.439 on 492 degrees of freedom

Multiple R-squared: 0.454, Adjusted R-squared: 0.4396

F-statistic: 31.47 on 13 and 492 DF, p-value: < 2.2e-16

The R2 of this model is .454. The Adjusted R2 (adjusted for number of variables in the model) is lower at .4396.

d. The starred variables above which includes zn, dis, rad, black and medv are all significant at the 5% level. All other variables are statistically insignificant. The Reduced Model is then given by:

Call:

lm(formula = crim ~ zn + dis + rad + black + medv, data = Boston)

Residuals:

Min 1Q Median 3Q Max

-10.553 -1.869 -0.358 0.839 75.744

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.919933 1.778986 4.452 1.05e-05 \*\*\*

zn 0.051799 0.017329 2.989 0.002935 \*\*

dis -0.672189 0.202939 -3.312 0.000992 \*\*\*

rad 0.472306 0.042102 11.218 < 2e-16 \*\*\*

black -0.008211 0.003615 -2.271 0.023562 \*

medv -0.174219 0.036295 -4.800 2.10e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.473 on 500 degrees of freedom

Multiple R-squared: 0.4393, Adjusted R-squared: 0.4337

F-statistic: 78.34 on 5 and 500 DF, p-value: < 2.2e-16

The Adjusted R2 is now .4337 which is almost identical to the Adjusted R2 of the full model (.4396). Given that this reduced model performs just as well as the full model, we would prefer the more parsimonious (streamlined) reduced model.

e. The 500th observation in the data contains the values shown below:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| crim | zn | indus | chas | nox | rm | age | dis | rad | tax | ptratio | black | lstat | medv |
| 0.17783 | 0 | 9.69 | 0 | 0.585 | 5.569 | 73.5 | 2.3999 | 6 | 391 | 19.2 | 395.77 | 15.1 | 17.5 |

What is the predicted value for Crime (crim)? ANSWER: 2.80

What is the actual value for Crime (crim)? ANSWER: .1778

What is the residual (error) for the 500th observation? = .1778 – 2.80 = -2.62

This can be invoked in R by using:

predict(FULLMODEL)

resid(FULLMODEL)

f. What is 95% confidence interval for the predicted (fitted) value in the problem above?

± (1.96)(SE) = 2.80 ± (1.96)(6.439) = -9.82 & 15.42

Note that appropriate SE for the model (rather than an individual coefficient) is the “Residual Standard Error” in R.

**R CODE FOR PROBLEM 1**

library(psych)

attach(Boston)

dim(Boston)

names(Boston)

describe(Boston)

############################################

REGMODEL1<-lm(crim~medv)

summary(REGMODEL1)

anova(REGMODEL1)

#############################################

FULLMODEL<-lm(crim~., data=Boston)

summary(FULLMODEL)

anova(FULLMODEL)

predict(FULLMODEL)

resid(FULLMODEL)

############################################

REDUCEDMODEL<-lm(crim~zn+dis+rad+black+medv, data=Boston)

summary(REDUCEDMODEL)

anova(REDUCEDMODEL)

################################################

*II.) All answers to this problem are worked out in the text. R Code is shown below.*

*Make sure you create new R files whenever you work with a new dataset. You will get incorrect results if you use the same names for different datasets.*

attach(Advertising)

dim(Advertising)

names(Advertising)

#############################

REG1<-lm(Sales~Radio)

summary(REG1)

anova(REG1)

#############################

REG2<-lm(Sales~Newspaper)

summary(REG2)

anova(REG2)

#################################

REG3<-lm(Sales~TV+Radio+Newspaper)

summary(REG3)

anova(REG3)

#################################

cor(Advertising)

III a.)

Best 1 variable model : Life Expectancy

Best 2 variable model : Frost, Life Expectancy

Best 3 variable model : Population, Illiteracy, Life Expectancy

Best 4 variable model : Population, Frost, Life Expectancy, Area

Best 5 variable model : Population, Illiteracy, Frost, Life Expectancy, Area

Best 6 variable model : Population, Illiteracy, Frost, Life Expectancy, hs\_grad\_rate, area

Best 7 variable model : Population, Illiteracy, Income, Frost, Life Expectancy, hs\_grad\_rate, area

b.) The highest Adjusted R2 is the model with 5 variables: Population, Illiteracy, Frost, Life Expectancy, Area. This 5 variable model has an Adjusted R2 of .7848.

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 120.164031804 17.181610452 6.994 0.0000000117

population 0.000177981 0.000059303 3.001 0.00442

illiteracy 1.172980493 0.680121662 1.725 0.09161

frost -0.013730312 0.007079737 -1.939 0.05888

life\_expectancy -1.607836823 0.232377225 -6.919 0.0000000150

area 0.000006804 0.000002919 2.331 0.02439

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.712 on 44 degrees of freedom

Multiple R-squared: 0.8068, **Adjusted R-squared: 0.7848**

F-statistic: 36.74 on 5 and 44 DF, p-value: 0.00000000000001221

c.)

Estimate Std. Error t value Pr(>|t|)

(Intercept) -0.00000000000000267 0.06560382605295603 0.00 1.0000

population 0.21524736690031809 0.07171955371741355 3.00 0.0044 \*\*

illiteracy 0.19367811472246613 0.11229912344792253 1.72 0.0916 .

frost -0.19333755570321509 0.09969030724628584 -1.94 0.0589 .

life\_expectancy -0.58467467881633184 0.08450178378413606 -6.92 0.000000015 \*\*\*

area 0.15727086346757338 0.06746384252124432 2.33 0.0244 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.464 on 44 degrees of freedom

Multiple R-squared: 0.807, Adjusted R-squared: 0.785

F-statistic: 36.7 on 5 and 44 DF, p-value: 0.0000000000000122

In terms of predictive ability, the coefficients with the largest values are:

Life Expectancy

Population

Illiteracy

Frost

Area

Note that you also get the above by running the “relaimpo” algorithm. The results for the 5 variable model (5Xs) is identical to the above coefficients.

The R code for this problem is shown below:

##########################################

attach(murder\_data)

library(leaps)

library(psych)

library(relaimpo)

#################################

options(scipen=999)

options(digits=4)

dim(murder\_data)

names(murder\_data)

psych(murder\_data)

describe(murder\_data)

cor(murder\_data)

#######################REGRESSION SUBSETS##################

REGSUBSETS<-regsubsets (murder\_rate~population+

illiteracy+income+

frost+ life\_expectancy+ hs\_grad\_rate

+ area, data = murder\_data, nvmax=7)

summary(REGSUBSETS)

SUMMARY\_OF\_REGRESSIONS<-summary(REGSUBSETS)

names(SUMMARY\_OF\_REGRESSIONS)

SUMMARY\_OF\_REGRESSIONS$adjr2

plot(SUMMARY\_OF\_REGRESSIONS$adjr2)

REG\_MODEL<-lm(murder\_rate~population+illiteracy+frost+

life\_expectancy+area,

data=murder\_data)

summary(REG\_MODEL)

################BETA REGRESSION########

SCALED\_DATA<-scale(murder\_data)

describe(SCALED\_DATA)

Z\_DATA<-data.frame(SCALED\_DATA)

Z\_DATA

Z\_MODEL<-lm(murder\_rate~population+illiteracy+frost+

life\_expectancy+area, data=Z\_DATA)

summary(Z\_MODEL)

##############################################

calc.relimp(Z\_MODEL, type=c("lmg", "last","first","pratt"), rela=TRUE)

###########################################

IV. SE = 



90% of the time, the true debt level will lie between $16,639 and $21,161.

V. A 1% margin of error implies that we want no more than ½% on each side of the tail of the normal distribution. This corresponds to ± 2.57 SD.

SE = 



The margin of error is 11.49. There is only a ½% chance that the true score is lower than 450 or higher than 472.