Dr. Sudhakar Raju

BIA 6309 – LINEAR & MULTIVARIATE MODELS

SUMMER 2018

**ANSWERS FOR ASSIGNMENT 4**

I 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 are 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)

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

II a.)

> describe(occupational\_prestige\_data)

vars n mean sd median trimmed mad min max

occupational\_type\* 1 102 51.50 29.59 51.50 51.50 37.81 1.00 102.00

prestige 2 102 46.83 17.20 43.60 46.20 19.20 14.80 87.20

education 3 102 **10.74** 2.73 10.54 10.63 3.15 6.38 15.97

income 4 102 **6797.90** 4245.92 5930.50 6161.49 3060.83 611.00 25879.00

percentage\_of\_women 5 102 28.98 31.72 13.60 24.74 18.73 0.00 97.51

professional\* 6 102 1.30 0.46 1.00 1.26 0.00 1.00 2.00

range skew kurtosis se

occupational\_type\* 101.00 0.00 -1.24 2.93

prestige 72.40 0.33 -0.79 1.70

education 9.59 0.32 -1.03 0.27

income 25268.00 2.13 6.29 420.41

percentage\_of\_women 97.51 0.90 -0.68 3.14

professional\* 1.00 0.84 -1.31 0.05

>

The mean prestige score is 46.83. Mean values of education are 10.74 years (beyond middle school), mean income is $6798 and mean percentage of women in occupations is 28.98.

The profession with the highest prestige is physicians (row 24) with a score of 87.2. The profession with the lowest prestige is newsboys (row 53) with a score of 14.8.

b.)

> MODEL\_1<-lm(prestige~education+income+percentage\_of\_women+professional)

> summary(MODEL\_1)

Call:

lm(formula = prestige ~ education + income + percentage\_of\_women +

professional)

Residuals:

Min 1Q Median 3Q Max

-18.9275 -4.6033 -0.0015 4.8288 19.2510

Coefficients:

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

(Intercept) 3.6153596 4.4637541 0.810 0.41996

education 3.0185168 0.5178846 5.829 0.0000000731 \*\*\*

income 0.0011768 0.0002686 4.381 0.0000298877 \*\*\*

percentage\_of\_women -0.0009108 0.0291425 -0.031 0.97513

professionalyes 9.3160805 2.8801637 3.235 0.00167 \*\*

---

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

Residual standard error: 7.493 on 97 degrees of freedom

Multiple R-squared: 0.8178, Adjusted R-squared: 0.8103

F-statistic: 108.9 on 4 and 97 DF, p-value: < 0.00000000000000022

>

All variables, except percentage of women, are statistically significant. It is not obvious which of these predictors has the most influence on prestige since the manner in which the predictors are denominated (years of education) and income (dollar values) are not on comparable scales. For instance, the regression implies that an additional year of education increases prestige by 3 points while an increase of $1000 increases prestige by .0011768 x 1000 = 1.1768 points.

c.)

> MODEL\_2<-lm(prestige~education+income+percentage\_of\_women, data=Z\_PRESTIGE\_DATA)

> summary(MODEL\_2)

Call:

lm(formula = prestige ~ education + income + percentage\_of\_women,

data = Z\_PRESTIGE\_DATA)

Residuals:

Min 1Q Median 3Q Max

-1.15229 -0.30999 -0.00793 0.29984 1.01744

Coefficients:

Estimate Std. Error t value

(Intercept) -0.00000000000000001396 0.04515775779186551736 0.000

education 0.66395512597124961562 0.06164379142481105078 10.771

income 0.32417565747255405739 0.06855406168027983194 4.729

percentage\_of\_women -0.01642103617685423955 0.05607036552368137305 -0.293

Pr(>|t|)

(Intercept) 1.00

education < 0.0000000000000002 \*\*\*

income 0.00000758 \*\*\*

percentage\_of\_women 0.77

---

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

Residual standard error: 0.4561 on 98 degrees of freedom

Multiple R-squared: 0.7982, Adjusted R-squared: 0.792

F-statistic: 129.2 on 3 and 98 DF, p-value: < 0.00000000000000022

It is immediately obvious from the beta regression that education has twice the impact on prestige as compared to income. The coefficients imply that a 1-SD unit increase in education increases prestige by .66 units while a 1-SD unit increase in income increases prestige by only .32 units.

d.) It makes no sense to standardize binary variables such as professional/no professional or male/female. What, for instance, does a 1-SD unit increase in males imply? What does a 1-SD unit increase in professional mean?

e. )

> MODEL\_3<-lm(prestige~education+income+percentage\_of\_women+professional,

+ data = STANDARDIZED\_AND\_UNSTANDARDIZED\_DATA)

> summary(MODEL\_3)

Call:

lm(formula = prestige ~ education + income + percentage\_of\_women +

professional, data = STANDARDIZED\_AND\_UNSTANDARDIZED\_DATA)

Residuals:

Min 1Q Median 3Q Max

-1.10015 -0.26757 -0.00009 0.28067 1.11895

Coefficients:

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

(Intercept) -0.164571 0.066696 -2.467 0.01536 \*

education 0.478704 0.082131 5.829 0.0000000731 \*\*\*

income 0.290435 0.066292 4.381 0.0000298877 \*\*\*

percentage\_of\_women -0.001679 0.053739 -0.031 0.97513

professionalyes 0.541491 0.167408 3.235 0.00167 \*\*

---

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

Residual standard error: 0.4355 on 97 degrees of freedom

Multiple R-squared: 0.8178, Adjusted R-squared: 0.8103

F-statistic: 108.9 on 4 and 97 DF, p-value: < 0.00000000000000022

Notice that this is not a beta regression since the “professional” variable is not standardized. It is harder to interpret a mix of standardized and unstandardized variables because education, income and percentage of women are in comparable SD unit scales while the professional variable is not. The professional variable implies that a professional occupation increases prestige by .54 points.

**R CODE FOR OCCUPATIONAL PRESTIGE ANALYSIS**

attach(occupational\_prestige\_data)

describe(occupational\_prestige\_data)

min(prestige)

which(prestige == "14.8")

max(prestige)

which(prestige == "87.2")

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

MODEL\_1<-lm(prestige~education+income+percentage\_of\_women+professional)

summary(MODEL\_1)

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

library(dplyr)

PRESTIGE\_DATA<-data.frame(select(occupational\_prestige\_data,

prestige, education, income,

percentage\_of\_women))

Z\_PRESTIGE\_DATA<-data.frame(scale(PRESTIGE\_DATA))

describe(Z\_PRESTIGE\_DATA)

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

MODEL\_2<-lm(prestige~education+income+percentage\_of\_women, data=Z\_PRESTIGE\_DATA)

summary(MODEL\_2)

STANDARDIZED\_AND\_UNSTANDARDIZED\_DATA<-data.frame(Z\_PRESTIGE\_DATA, professional)

STANDARDIZED\_AND\_UNSTANDARDIZED\_DATA

MODEL\_3<-lm(prestige~education+income+percentage\_of\_women+professional,

data = STANDARDIZED\_AND\_UNSTANDARDIZED\_DATA)

summary(MODEL\_3)

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

III. a.) The correlation matrix is shown below:

> CORRELATION\_MATRIX

COUPON NEW HI S\_INCOME E\_INCOME S\_POP E\_POP DISTANCE PAX

COUPON 1.00 0.02 -0.35 -0.09 0.05 -0.11 0.09 0.75 -0.34

NEW 0.02 1.00 0.05 0.03 0.11 -0.02 0.06 0.08 0.01

HI -0.35 0.05 1.00 -0.03 0.08 -0.17 -0.06 -0.31 -0.17

S\_INCOME -0.09 0.03 -0.03 1.00 -0.14 0.52 -0.27 0.03 0.14

E\_INCOME 0.05 0.11 0.08 -0.14 1.00 -0.14 0.46 0.18 0.26

S\_POP -0.11 -0.02 -0.17 0.52 -0.14 1.00 -0.28 0.02 0.28

E\_POP 0.09 0.06 -0.06 -0.27 0.46 -0.28 1.00 0.12 0.31

DISTANCE 0.75 0.08 -0.31 0.03 0.18 0.02 0.12 1.00 -0.10

PAX -0.34 0.01 -0.17 0.14 0.26 0.28 0.31 -0.10 1.00

FARE 0.50 0.09 0.03 0.21 0.33 0.15 0.29 **0.67** -0.09

FARE

COUPON 0.50

NEW 0.09

HI 0.03

S\_INCOME 0.21

E\_INCOME 0.33

S\_POP 0.15

E\_POP 0.29

DISTANCE 0.67

PAX -0.09

FARE 1.00

The variable that has the highest correlation with FARE is DISTANCE. Multicollinearity does not seem to be an issue since none of the variables have very high correlation. The highest correlation is .75 between DISTANCE and COUPON. This is not high enough to present a multicollinearity issue.

b.) The estimates of the full model is shown below:

|  |
| --- |
| > summary(FULL\_REG\_MODEL)  Call:  lm(formula = FARE ~ COUPON + NEW + VACATION + SW + HI + S\_INCOME +  E\_INCOME + S\_POP + E\_POP + SLOT + GATE + DISTANCE + PAX)  Residuals:  Min 1Q Median 3Q Max  -106.329 -22.707 -2.329 21.135 128.694  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 12.6993915337 27.3794269858 0.464 0.642932  COUPON 3.7548909844 12.1940741390 0.308 0.758240  NEW -2.3955314942 1.8754235204 -1.277 0.201961  VACATIONYes -35.6444410042 3.6170501865 -9.855 < 0.0000000000000002 \*\*\*  SWYes -40.9695969113 3.7437292880 -10.944 < 0.0000000000000002 \*\*\*  HI 0.0084257918 0.0009900663 8.510 < 0.0000000000000002 \*\*\*  S\_INCOME 0.0012066780 0.0005171071 2.334 0.019938 \*  E\_INCOME 0.0013742726 0.0003749187 3.666 0.000268 \*\*\*  S\_POP 0.0000034009 0.0000006523 5.213 0.0000002525 \*\*\*  E\_POP 0.0000043631 0.0000007547 5.781 0.0000000117 \*\*\*  SLOTFree -16.2447672500 3.8468795785 -4.223 0.0000277170 \*\*\*  GATEFree -20.5792305291 4.0015842717 -5.143 0.0000003629 \*\*\*  DISTANCE 0.0749842548 0.0035795487 20.948 < 0.0000000000000002 \*\*\*  PAX -0.0008709425 0.0001459072 -5.969 0.0000000040 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 35.47 on 624 degrees of freedom  Multiple R-squared: 0.7868, Adjusted R-squared: 0.7823  F-statistic: 177.1 on 13 and 624 DF, p-value: < 0.00000000000000022 |
|  |
| |  | | --- | | > | |

c.) The subsets algorithm examines a total of 8191 regressions to determine the most appropriate model.

The Excel function for Combinations is used below. Suppose you have 13 possible predictors and you want to develop a 5 variable model. How many possible 5 variable combinations are there? The combinations formula is given by:

C =

where:

n = total elements we can choose from (here, n = 13)

r = how many is being chosen at any given time (here, r = 5)

C = =

Eliminating 8! from the numerator and denominator, results in:

C = = 1287 combinations

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| VARIABLES | |  | | --- | | NUMBER OF POSSIBLE MODELS | |  |  |  |
| 1 | 13 |  |  |  |
| 2 | 78 |  |  |  |
| 3 | 286 |  |  |  |
| 4 | 715 |  | =COMBIN(13, 5) |  |
| **5** | |  | | --- | | **1287** | |  |  |  |
| 6 | 1716 |  |  |  |
| 7 | 1716 |  |  |  |
| 8 | 1287 |  |  |  |
| 9 | 715 |  |  |  |
| 10 | 286 |  |  |  |
| 11 | 78 |  |  |  |
| 12 | 13 |  |  |  |
| 13 | 1 |  |  |  |
|  | **8191** |  |  |  |
|  |  |  |  |  |

d.) The best 4 variable model is given by:

> REG\_SUMMARY$adjr2

[1] 0.4480550 0.6031119 0.7064340 **0.7318363** 0.7437300 0.7602224 0.7640663

[8] 0.7690798 0.7775265 0.7809381 0.7824096 0.7826435 0.7823282

> summary(BEST\_4\_VARIABLE\_MODEL)

Call:

lm(formula = FARE ~ VACATION + SW + HI + DISTANCE)

Residuals:

Min 1Q Median 3Q Max

-101.533 -24.679 -2.249 25.694 121.924

Coefficients:

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

(Intercept) 83.6667884 6.3216258 13.235 < 0.0000000000000002 \*\*\*

VACATIONYes -52.1011141 3.5729193 -14.582 < 0.0000000000000002 \*\*\*

SWYes -61.0621080 3.5202010 -17.346 < 0.0000000000000002 \*\*\*

HI 0.0075586 0.0009673 7.814 0.0000000000000232 \*\*\*

DISTANCE 0.0779823 0.0026404 29.535 < 0.0000000000000002 \*\*\*

---

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

Residual standard error: 39.37 on 633 degrees of freedom

Multiple R-squared: 0.7335, Adjusted R-squared: 0.7318

F-statistic: 435.6 on 4 and 633 DF, p-value: < 0.00000000000000022

The estimates imply that vacation routes tend to reduce fares by $52 perhaps due to increased competition on those routes. The Southwest effect is evident in the negative coefficient on SW – fares tend to be $61 lower if Southwest operates in a particular route. A one point increase in the Herfindahl (HI) index (a measure of market concentration) tends to increases fares by .0076 while an increase of 1 mile increases fares by about 8 cents.

e. The Herfindahl index ranges from a low of 1230 to a max of 10,000. The average value of HI is 4442. The average distance is 976 miles.

|  |
| --- |
| > describe(airfares\_data)  vars n mean sd median trimmed mad  S\_CODE\* 1 638 2.44 2.46 1.00 1.97 0.00  S\_CITY\* 2 638 21.44 13.19 20.50 20.80 17.05  E\_CODE\* 3 638 1.68 1.50 1.00 1.28 0.00  E\_CITY\* 4 638 40.92 18.29 42.00 41.97 20.76  COUPON 5 638 1.20 0.20 1.15 1.17 0.19  NEW 6 638 2.75 0.76 3.00 2.99 0.00  VACATION\* 7 638 1.27 0.44 1.00 1.21 0.00  SW\* 8 638 1.30 0.46 1.00 1.26 0.00  HI 9 638 **4442.14** 1724.27 4208.18 4287.60 1728.86  S\_INCOME 10 638 27759.86 3596.21 28637.00 27754.71 3610.13  E\_INCOME 11 638 27663.73 4611.33 26409.00 27427.17 5017.12  S\_POP 12 638 4557004.49 3010985.26 3532657.00 4497970.09 3342723.33  E\_POP 13 638 3194503.10 2735603.92 2195215.00 2869869.31 1739660.60  SLOT\* 14 638 1.71 0.45 2.00 1.77 0.00  GATE\* 15 638 1.81 0.40 2.00 1.88 0.00  DISTANCE 16 638 **975.65** 646.24 850.00 898.98 601.94  PAX 17 638 12782.21 13202.23 7792.00 9754.76 4558.99  FARE 18 638 **160.88** 76.02 144.60 155.21 78.27  min max range skew kurtosis se  S\_CODE\* 1.00 8.00 7.00 1.32 0.03 0.10  S\_CITY\* 1.00 51.00 50.00 0.31 -1.01 0.52  E\_CODE\* 1.00 8.00 7.00 2.20 3.84 0.06  E\_CITY\* 1.00 68.00 67.00 -0.41 -0.68 0.72  COUPON 1.00 1.94 0.94 1.35 1.58 0.01  NEW 0.00 3.00 3.00 -2.96 7.21 0.03  VACATION\* 1.00 2.00 1.00 1.05 -0.89 0.02  SW\* 1.00 2.00 1.00 0.85 -1.28 0.02  **HI 1230.48 10000.00**  8769.52 0.81 0.47 68.26  S\_INCOME 14600.00 38813.00 24213.00 0.16 0.08 142.38  E\_INCOME 14600.00 38813.00 24213.00 0.47 -0.47 182.56  S\_POP 29838.00 9056076.00 9026238.00 0.28 -1.52 119206.05  E\_POP 111745.00 9056076.00 8944331.00 1.04 -0.29 108303.60  SLOT\* 1.00 2.00 1.00 -0.95 -1.10 0.02  GATE\* 1.00 2.00 1.00 -1.54 0.38 0.02  DISTANCE 114.00 2764.00 2650.00 0.89 -0.06 25.58  PAX 1504.00 73892.00 72388.00 2.68 7.52 522.68  FARE 42.47 402.02 359.55 0.62 -0.35 3.01 |
|  |
| |  | | --- | | > | |

The fitted value of the model equals:

PREDICTED FARE = 83.67 – 52.10 (0) – 61.06 (1) + .0076 (4442) + .0780 (976)

= 83.67 – 0 – 61.06 + 33.76 + 76.13

PREDICTED FARE = $132.50

**R CODE FOR AIRFARES**

attach(airfares\_data)

dim(airfares\_data)

str(airfares\_data)

##############EXPLORATORY DATA ANALYSIS############

library(psych)

describe(airfares\_data)

CORRELATION\_DATA<- data.frame(COUPON, NEW, HI, S\_INCOME,

E\_INCOME, S\_POP, E\_POP, DISTANCE,

PAX, FARE)

COR\_MATRIX<-cor(CORRELATION\_DATA)

COR\_MATRIX

CORRELATION\_MATRIX<-round(COR\_MATRIX, 2)

CORRELATION\_MATRIX

#####################FULL REGRESSION MODEL######

options (scipen=999)

FULL\_REG\_MODEL<- lm(FARE~COUPON + NEW + VACATION + SW + HI +

S\_INCOME + E\_INCOME + S\_POP + E\_POP + SLOT + GATE +

DISTANCE + PAX)

summary(FULL\_REG\_MODEL)

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

library(leaps)

REG\_SUBSETS<-regsubsets(FARE~COUPON + NEW + VACATION + SW + HI +

S\_INCOME + E\_INCOME + S\_POP + E\_POP + SLOT + GATE +

DISTANCE + PAX, data= airfares\_data,

nvmax = 13)

summary(REG\_SUBSETS)

REG\_SUMMARY<-summary(REG\_SUBSETS)

REG\_SUMMARY$adjr2

BEST\_4\_VARIABLE\_MODEL<-lm(FARE~VACATION+SW+HI+DISTANCE)

summary(BEST\_4\_VARIABLE\_MODEL)