Final Exam

Biplav Timalsina

May 13, 2018

STAT 757 Applied Regression Analysis

Take-home Exam Procedures [20 points for following]

You may use your notes, textbook, and R while taking this exam. You are free to look online for definitions, however searching directly for answers to the given questions will be treated as cheating and dealt with according to UNR's policies regarding Academic Dishonesty. You are encouraged to ask questions to the instructor during the exam (but only hints will be given!) Partial credit WILL be awarded where sufficient details have been provided. This exam must be completed **individually** (you may not discuss any aspect of the exam with your classmates).

Modify this `.Rmd` file to respond to each question in @sheather2009 below. Follow the directions indicated by Sheather **exactly**. By that I mean: if there several parts, address them one by one in turn or, if the problem asks for a report, write a **short** report. That means write an Introduction, Methods, Results, and Discussion/Conclusion section. Please be concise and include only a few important figures and code. If your answers contain redundancies or is unclear you will lose points despite being technically correct. Please email **both** your .Rmd (or roxygen .R) and one of the following either .HTML, .PDF, or .DOCX using the format `SURNAME-FIRSTNAME-Exam2.Rmd` and `SURNAME-FIRSTNAME-Exam2.pdf`.

Exercise 5.4.1 [20 points]

1. This problem is based on CASE 32 Overdue Bills from Bryant and Smith (1995). Quick Stab Collection Agency (QSCA) is a bill-collecting agency that specializes in collecting small accounts. To distinguish itself from competing collection agencies, the company wants to establish a reputation for collecting delinquent accounts quickly. The marketing department has just suggested that QSCA adopt the slogan: Under 60 days or your money back!

You have been asked to look at account balances. In fact, you suspect that the number of days to collect the payment is related to the size of the bill. If this is the case, you may be able to estimate how quickly certain accounts are likely to be collected, which, in turn, may assist the marketing department in determining an appropriate level for the money-back guarantee.

To test this theory, a random sample of accounts closed out during the months of January through June has been collected. The data set includes the initial size of the account and the total number of days to collect payment in full. Because QSCA deals in both household and commercial accounts in about the same proportion, an equal number have been collected from both groups. The first 48 observations in the data set are residential accounts and the second 48 are com- mercial accounts. The data can be found on the book web site in the file named overdue.txt. In this data set, the variable LATE is the number of days the payment is overdue, BILL is the amount of the overdue bill in dollars and TYPE identifies whether an account is RESIDENTIAL or COMMERCIAL.

Develop a regression model to predict LATE from BILL.

Ans:

Let us develop a model using the data given to us. I have introduced a dummy variable Residential, which means residential when Residential=1 and commericial when Residential=0.

```
## The following objects are masked from Overdue (pos = 3):
##
##     BILL, LATE
```

```
##
## Call:
## lm(formula = LATE ~ BILL + TYPE + TYPE:BILL)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.1211  -2.2163   0.0974   1.9556   8.6995
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          101.758184   1.198504   84.90   <2e-16 ***
## BILL                  -0.190961   0.006285  -30.38   <2e-16 ***
## TYPEResidential      -99.548561   1.694940  -58.73   <2e-16 ***
## BILL:TYPEResidential   0.356644   0.008888   40.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.371 on 92 degrees of freedom
## Multiple R-squared:  0.9803, Adjusted R-squared:  0.9796
## F-statistic:  1524 on 3 and 92 DF,  p-value: < 2.2e-16
```

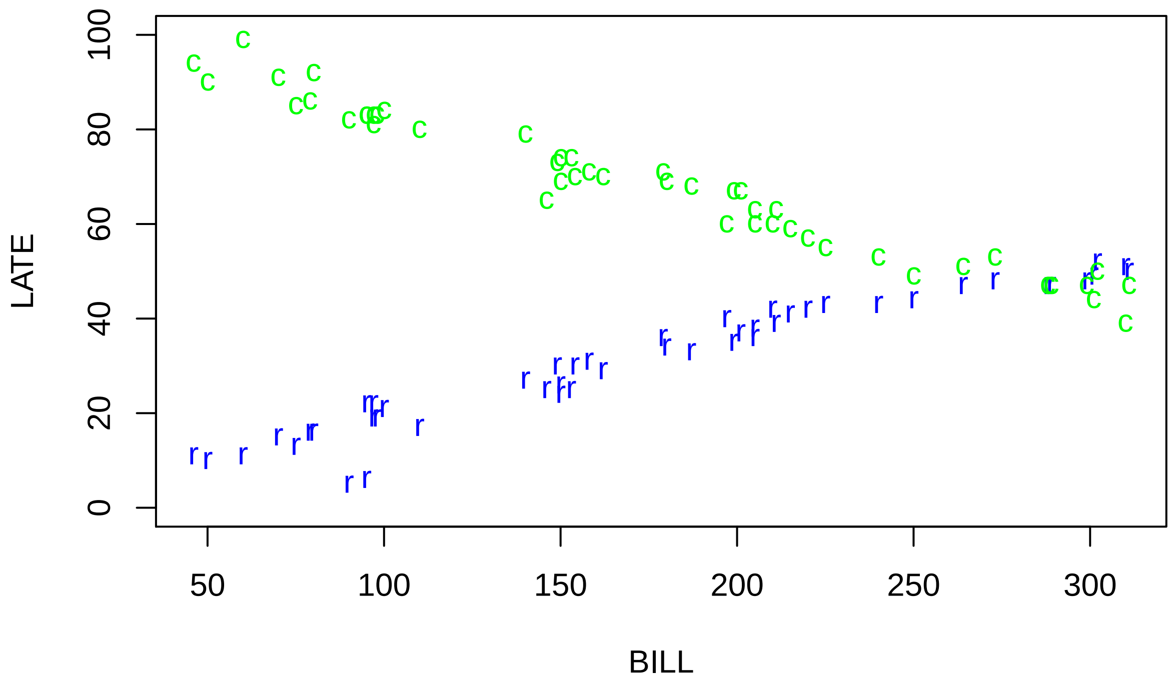
```
#I have used this method.
Overdue<-read.table('F:/unr/4th sem/applied regression analysis/Assignments/Final Exam/overdue.txt',heac
attach(Overdue)
```

```
## The following objects are masked from Overdue (pos = 3):
##
##     BILL, LATE
```

```
Overdue$residential[1:48]<-1
Overdue$residential[49:96]<-0
attach(Overdue)
```

```
## The following objects are masked from Overdue (pos = 3):
##
##     BILL, LATE

## The following objects are masked from Overdue (pos = 4):
##
##     BILL, LATE
```

```
par(mfrow=c(1,1))
plot(BILL[residential==1],LATE[residential==1],pch='r',col="Blue",ylab="LATE",xlab = "BILL",ylim=c(0,10(
points(BILL[residential==0],LATE[residential==0],pch='c',col="green")
```



Note: There are two codes; the first one is yours and the second one is how I have done it. They take seperate reference points. For different reference points, the coefficients observed will be different. From the code given by you in email, since the reference point is different than how I have solved it, the full model coefficients are different. But the two independent lines are same for the both references.

It is clear form the figure that the dummy variable for the residential changes the effect of bill size on the time of payment. We shall therefore allow for the dummy variable for residential to produce change in Y. In this case, the appropriate model is what we refer to as Unrelated regression lines, as below,

Notice that all regression coeffcients are highly significant. Thus we shall use as a final model,

$$late = \beta_0 + \beta_1 * Bill + \beta_2 * residential + \beta_3 * Bill * residential$$

$$late = \beta_0 + \beta_1 * Bill,$$

when residential= 0

$$late = \beta_0 + \beta_2 + (\beta_1 + \beta_3) * Bill,$$

when residential= 1

Lets us now develop the regression model:

```
m1<-lm(LATE~BILL + residential + residential:BILL)
summary(m1)
```

```
##
## Call:
## lm(formula = LATE ~ BILL + residential + residential:BILL)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.1211  -2.2163   0.0974   1.9556   8.6995
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      101.758184   1.198504   84.90   <2e-16 ***
## BILL              -0.190961   0.006285  -30.38   <2e-16 ***
## residential      -99.548561   1.694940  -58.73   <2e-16 ***
## BILL:residential   0.356644   0.008888   40.12   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.371 on 92 degrees of freedom
## Multiple R-squared:  0.9803, Adjusted R-squared:  0.9796
## F-statistic:  1524 on 3 and 92 DF,  p-value: < 2.2e-16
```

```r
detach(Overdue)
```

Following the above,

$$late = 101.76 - 0.19 * Bill - 99.55 * residential + 0.35 * Bill * residential$$

and individual equations can be written as,

$$late = 101.75 - 0.19 * Bill,$$

when residential= 0

$$late = 2.21 + 0.16 * Bill,$$

when residential=1

Thus, for residential bills the number of days needed to collect the payment increases as the size of bill increases, while for commercial bills the number of days needed to collect the payment decreases with the size of bill.

Also, if we are interested in overall test of

$$H_0 : \beta_2 = \beta_3 = 0$$

ie.

$$Y = \beta_0 + \beta_1 * x_1$$

(reduced model: coincident regression lines)

Against, $H_A : H_0 \quad is \quad not\ true$ ie. Y is full model : unrelated lines

The fit under the reduced model can be seen below:

```r
mreduced<-lm(LATE~BILL)
summary(mreduced)
```

```
##
## Call:
```

```
## lm(formula = LATE ~ BILL)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.846 -17.212  -0.793  19.007  47.774
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51.98390    5.96405   8.716 9.84e-14 ***
## BILL        -0.01264    0.03128  -0.404    0.687
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.72 on 94 degrees of freedom
## Multiple R-squared:  0.001734,   Adjusted R-squared:  -0.008885
## F-statistic: 0.1633 on 1 and 94 DF,  p-value: 0.687
```

We can use partial T-test to measure this.

```
anova(mreduced,m1)
```

```
## Analysis of Variance Table
##
## Model 1: LATE ~ BILL
## Model 2: LATE ~ BILL + residential + residential:BILL
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1     94 52904
## 2     92  1045  2     51859 2281.7 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As expected, there is strong evidence against the reduced model in favor of full model since the p-value is equal to <2.2e-16. Thus, we prefer the unrelated regression lines model to the coincident lines model. Hence the final model is,

$$late = 101.76 - 0.19 * Bill - 99.55 * residential + 0.35 * Bill * residential$$

Exercise 6.7.2 [20 points]

2. Chapter 5-2 of the award-winning book on baseball (Keri, 2006) makes extensive use of multiple regression. For example, since the 30 Major League Baseball teams play eighty-one home games during the regular season and receive the largest share of their income from the ticket sales associated with these games the author develops a least squares regression model to predict Y, yearly income (in 2005 US dollars) from ticket sales for each team from home games each year. Ticket sales data for each team for each of the years from 1997 to 2004 are used to develop the model. Thus, there are 30 * 8 = 240 rows of data. Twelve potential predictor variables are identified as follows: Six predictor variables measure team quality, namely:

x1 = Number of games won in current season
x2 = Number of games won in previous season
x3 = Dummy variable for playoff appearance in current season
x4 = Dummy variable for playoff appearance in previous season
x5 = Number of winning seasons in the past 10 years
x6 = Number of playoff appearances in the past 10 years

Three predictors measure stadium of quality, namely:
x7 = Seating capacity
x8 = Stadium quality rating
x9 = Honeymoon effect
Two predictors measure market quality, namely:
x10 = Market size
x11 = Per-capita income

Finally, x12 = Year is included to allow for inflation. The author found that seven of these (predictor) variables had a statistically significant impact on attendance revenue (i.e., had a t-statistic significant at least at the 10% level). Describe in detail two major concerns that potentially threaten the validity of the model.

Ans:

1) Effect of Multicollinearity

When two or more highly corelated predictor variables are included in a regression model, they are effectively carrying very similar information about the response variable. Thus, it is difficult for least squares to distinguish their seperate effects on the response variable. In this istuation the overall F-test will be highly statistically significant but very few of the regression coefficients may be statistically significant. Another consequence of highly correlated predictor variables is that some of the coefficients in the regression model are of the opposite sign than expected. And, coorelation amongst the predictos increase the variance of the estimated regression coefficients. If there are a large number of predictors whose variance inflation factor exceeds a threshold, then multicollinearity will be a major issue.Based on the observation, we need to select only good predictor variables.

2) Concerns of omitted variables and Spurious Correlation

Sometimes there is relationship between two variables because both are related to the third variable that has been omitted from the regression model. This phenomenon is referred to as ‘spurious correlation’. If there are other variables say such as holidays during the matches, or other variables which have not been included in this dataset, and that has a huge correlation on the number of ticket sales, then we might need to remodel the model to study the effects of such variables.

Exercise 7.5.1 [20 points]

1. The generated data set in this question is taken from Mantel (1970). The data are given in Table 7.3 and can be found on the book web site in the file Mantel.txt. Interest centers on using variable selection to choose a subset of the predictors to model Y. The data were generated such that the full model

$$Y = b0 + b1 * X1 + b2 * X2 + b3 * X3 + e \qquad (7.8)$$

is a valid model for the data.

Output from R associated with different variable selection procedures based on model (7.8) appears below.

Table 7.3 Mantel's generated data

Case	Y	X1	X2	X3
1	5	1	1004	6
2	6	200	806	7.3
3	8	–50	1058	11
4	9	909	100	13
5	11	506	505	13.1

(a) Identify the optimal model or models based on R^2_{adj}, AIC and BIC from the approach based on all possible subsets.
(b) Identify the optimal model or models based on AIC and BIC from the approach based on forward selection.
(c) Carefully explain why different models are chosen in (a) and (b).
(d) Decide which model you would recommend. Give detailed reasons to support your choice.

Output from R: Correlations between the predictors in model (7.8)

```
             X1            X2            X3
X1    1.0000000    -0.9999887     0.6858141
X2   -0.9999887     1.0000000    -0.6826107
X3    0.6858141    -0.6826107     1.0000000
```

Approach 1: All Possible Subsets

Figure 7.13 shows a plot of adjusted R-squared against the number of predictors in the model for the optimal subsets of predictors.
Table 7.4 gives the values of R^2_{adj}, AIC and BIC for the best subset of each size.

Approach 2: Stepwise Subsets

Forward Selection Based on AIC

```
Start: AIC= 9.59
Y ~ 1
            Df     Sum of Sq          RSS         AIC
+ X3         1       20.6879       2.1121     -0.3087
+ X1         1        8.6112      14.1888      9.2151
+ X2         1        8.5064      14.2936      9.2519
<none>                            22.8000      9.5866
Step: AIC= -0.31
Y ~ X3
            Df     Sum of Sq          RSS         AIC
<none>                            2.11211    -0.30875
+ X2         1       0.06633      2.04578     1.53172
+ X1         1       0.06452      2.04759     1.53613
```

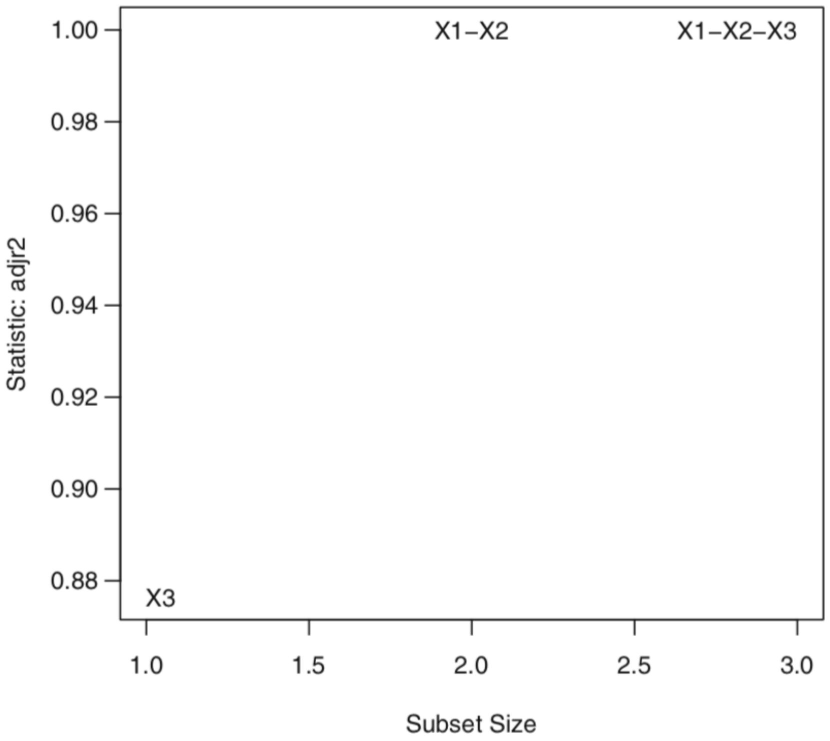


Figure 7.13 Plots of R^2_{adj} for the best subset of each size

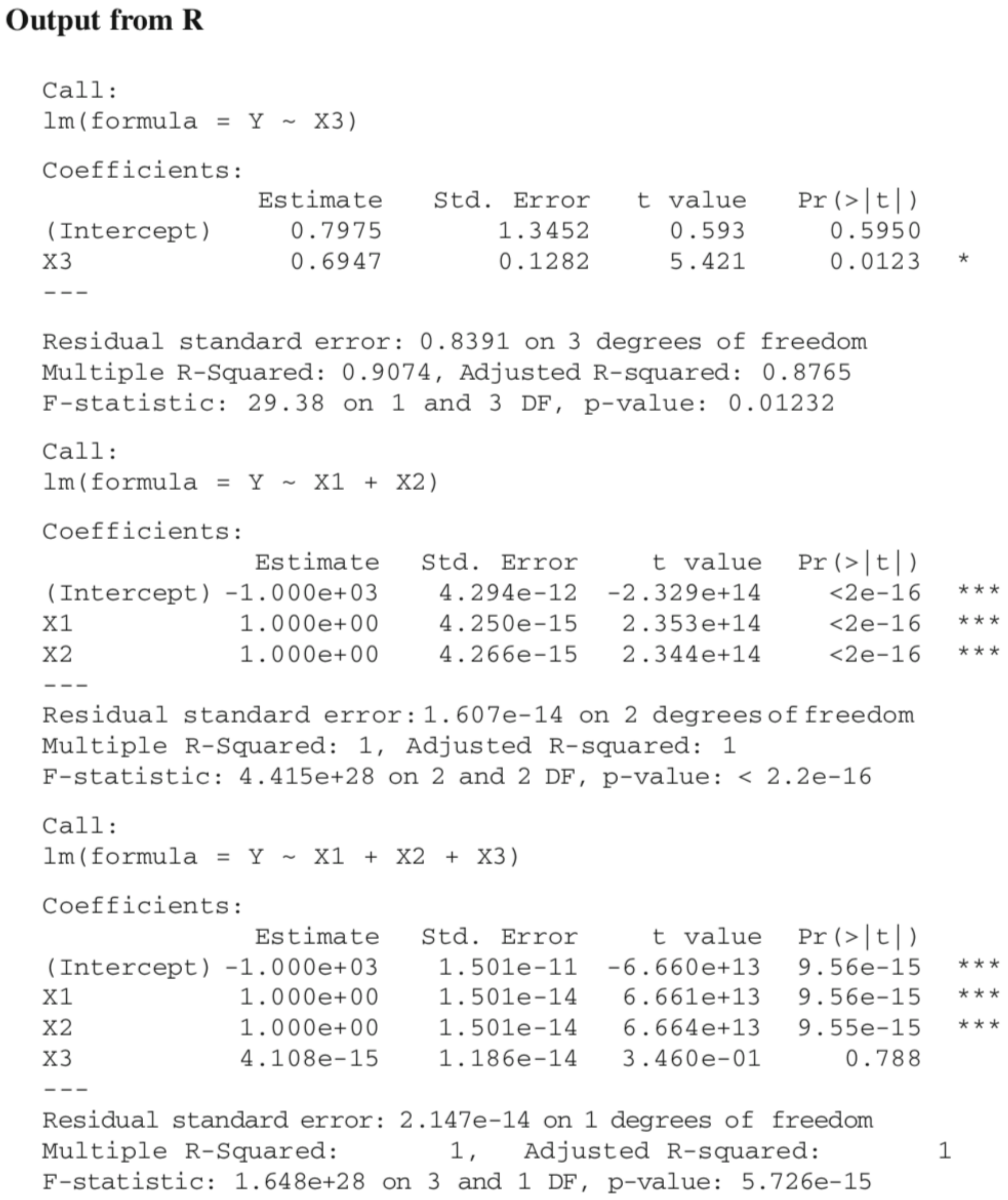
Table 7.4 Values of R^2_{adj}, AIC and BIC for the best subset of each size

Subset size	Predictors	R^2_{adj}	AIC	BIC
1	X3	0.8765	–0.3087	–1.0899
2	X1, X2	1.0000	–316.2008	–317.3725
3	X1, X2, X3	1.0000	–314.7671	–316.3294

Forward Selection Based on BIC*

```
Start: AIC= 9.2
Y ~ 1
             Df     Sum of Sq         RSS         AIC
+ X3          1       20.6879      2.1121     -1.0899
+ X1          1        8.6112     14.1888      8.4339
+ X2          1        8.5064     14.2936      8.4707
<none>                            22.8000      9.1961
Step: AIC= -1.09
Y ~ X3
             Df     Sum of Sq         RSS         AIC
<none>                            2.11211    -1.08987
+ X2          1       0.06633     2.04578     0.36003
+ X1          1       0.06452     2.04759     0.36444
```

* The R command step which was used here labels the output as AIC even when the BIC penalty term is used.



Output from R

```
Call:
lm(formula = Y ~ X3)

Coefficients:
              Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)     0.7975       1.3452     0.593     0.5950
X3              0.6947       0.1282     5.421     0.0123  *
---

Residual standard error: 0.8391 on 3 degrees of freedom
Multiple R-Squared: 0.9074, Adjusted R-squared: 0.8765
F-statistic: 29.38 on 1 and 3 DF, p-value: 0.01232

Call:
lm(formula = Y ~ X1 + X2)

Coefficients:
              Estimate   Std. Error     t value   Pr(>|t|)
(Intercept) -1.000e+03   4.294e-12  -2.329e+14    <2e-16  ***
X1           1.000e+00   4.250e-15   2.353e+14    <2e-16  ***
X2           1.000e+00   4.266e-15   2.344e+14    <2e-16  ***
---
Residual standard error:1.607e-14 on 2 degrees of freedom
Multiple R-Squared: 1, Adjusted R-squared: 1
F-statistic: 4.415e+28 on 2 and 2 DF, p-value: < 2.2e-16

Call:
lm(formula = Y ~ X1 + X2 + X3)

Coefficients:
              Estimate   Std. Error     t value   Pr(>|t|)
(Intercept) -1.000e+03   1.501e-11  -6.660e+13  9.56e-15  ***
X1           1.000e+00   1.501e-14   6.661e+13  9.56e-15  ***
X2           1.000e+00   1.501e-14   6.664e+13  9.55e-15  ***
X3           4.108e-15   1.186e-14   3.460e-01     0.788
---
Residual standard error: 2.147e-14 on 1 degrees of freedom
Multiple R-Squared:       1,   Adjusted R-squared:       1
F-statistic: 1.648e+28 on 3 and 1 DF, p-value: 5.726e-15
```

Ans:

Subset Size	Predictors	$R^2 Adj$	AIC	BIC
1	X_3	0.8765	-0.3087	-1.0899
2	$X_1 X_2$	**1.0000**	**-316.2008**	**-317.3725**
3	$X_1 X_2 X_3$	1.0000	-314.7671	-316.3294

I have mentioned the best model for given dataset according to R-Square-adjusted, AIC and BIC. According to AIC and BIC, subset of size 2 is the best model. Subset of size 2 and 3 are supported by R-square adjusted. Since, all three measures support model with size 2 to be the best, I think the model of size 2 is the best.

b) We can see from the above given results in question that AIC and BIC has least values when there is only one predictor in the model with x3 as the predictor variable. Hence based on forward selection, the final model is

$$Y = \beta_0 + \beta_1 * X_3 + \epsilon$$

c) All possible subsets method of selection of model is based on considering all 2ˆm possible regression equations and identifying the subset of the predictors of a given size that maximizes a measure of fit or minimizes an infromation criterion based on a monotone fucntion of residual sum of squares. Forward selection starts with no potential predictor variables in the regression equation. Then, at each step, it adds the predictor such that the resulting model has the lowest value of an information criterion. This process is continues until all variables have been added to the model or the information criterion increases. But it only considers at most m + (m-1)+(m-2) + … +1= m(m+1)/2 of the 2ˆm possible predictor subsets. Thus it does not necessarily find the model that minimized the information criteria across all 2ˆm possible predictor subsets. Because the principle of selection of models are different in these two approaches, it is very likely that we end of with different models in two approaches.

d) As discussed earlier, subsets method of selection of model considers all the possible subsets and hence is comprehensive, as compared to forward or backward(stepwise) method. Hence, I would recommend the best subsets model with 2 predictor here in this case. But while saying that, it is also contextual. If some variable is really important and has been seen significant in previous researches, but is not statistically significant in our analysis, than we might need to include in our model even though it is not significant statistically.

Exercise 8.3.3 [20 points]

3. Data on 102 male and 100 female athletes were collected at the Australian Institute of Sport. The data are available on the book web site in the file ais.txt. Develop a logistic regression model for gender (y = 1 corresponds to female) or (y = 0 corresponds to male) based on the following predictors (which is a subset of those available):

RCC, read cell count
WCC, white cell count
BMI, body mass index

(Hint: Use marginal model plots to aid model development.)

Ans:

Let us first start by loading the data. Then let us make a model with 3 variables, making the y-variable binomial. Then lets plot marginal model plots for RCC, WCC and BMI. We have seen that the MMC for WCC and BMI are not coaligned as seen below.

```
ais<-read.table('F:/unr/4th sem/applied regression analysis/Assignments/Final Exam/ais.txt',header=T)
attach(ais)


m1<-glm(Sex~RCC+WCC+BMI,family=binomial(),data=ais)
summary(m1)
```

```
## 
## Call:
## glm(formula = Sex ~ RCC + WCC + BMI, family = binomial(), data = ais)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -2.56012  -0.52459  -0.02467   0.52242   2.67643  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) 29.29399    3.92411   7.465 8.32e-14 ***
## RCC         -5.34706    0.72118  -7.414 1.22e-13 ***
## WCC          0.15505    0.12190   1.272  0.20338    
## BMI         -0.22911    0.08723  -2.626  0.00863 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 280.01  on 201  degrees of freedom
## Residual deviance: 147.01  on 198  degrees of freedom
## AIC: 155.01
## 
## Number of Fisher Scoring iterations: 5
```

```
library(alr4)
```

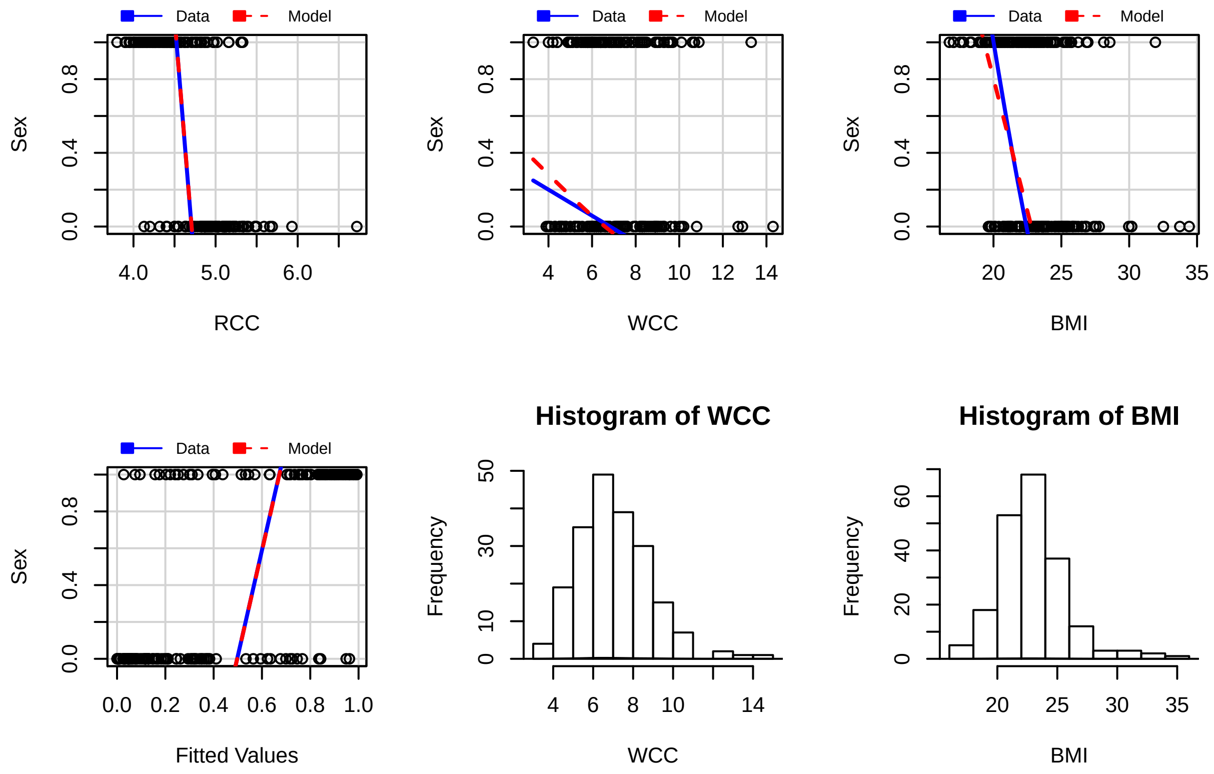
```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: effects
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

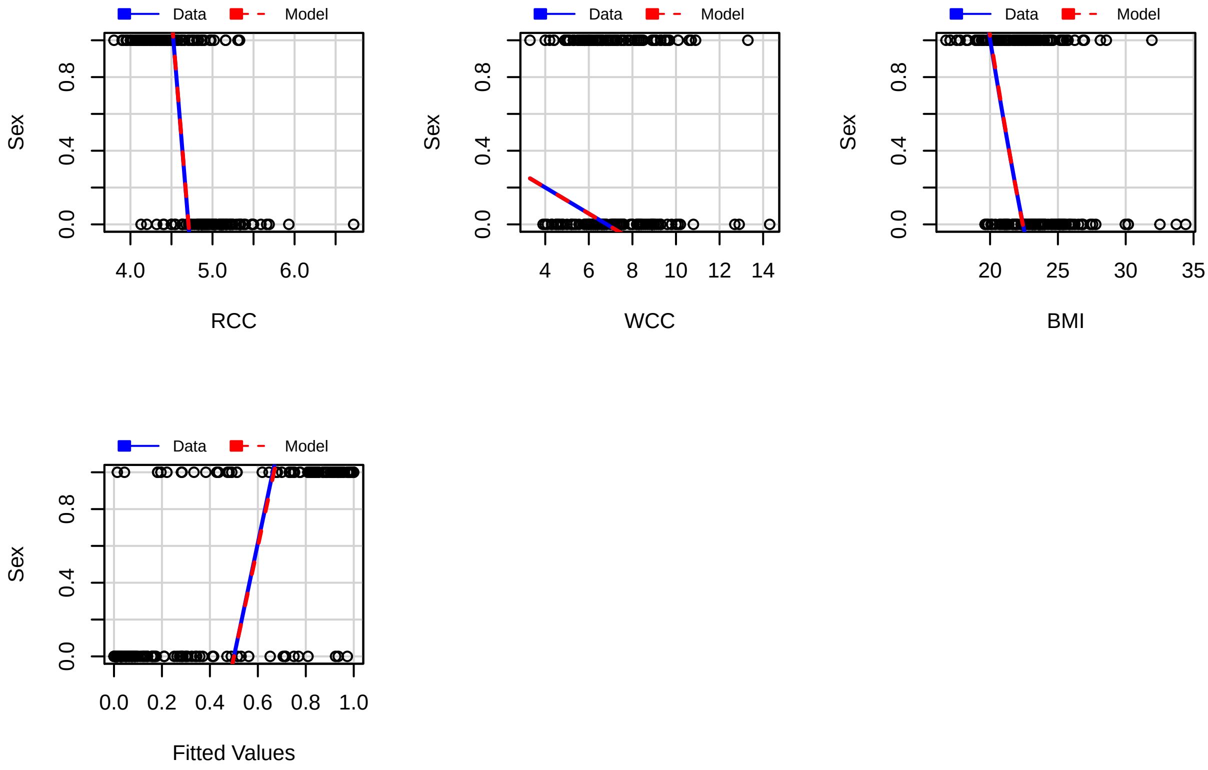
```
par(mfrow=c(2,3))
mmp(m1,RCC)
mmp(m1,WCC)
mmp(m1,BMI)
mmp(m1,m1$fitted.values,xlab="Fitted Values")

hist(WCC)
lines(density(WCC))

hist(BMI)
lines(density(BMI))
```



To analyse further, We can observe from the histogram that the WCC and RCC is left skewed. I think we should transform these variable by log transformation.

Now, we develop a new model including the two newly found variables.

```
m2<-glm(Sex~RCC+ WCC +log(WCC)+ log(BMI)+ BMI,family = binomial(),data=ais)
par(mfrow=c(2,3))
mmp(m2,RCC)
mmp(m2,WCC)
mmp(m2,BMI)
mmp(m2,m2$fitted.values,xlab="Fitted Values")
```



We can see that there is better agreement between two set of lines in marginal model plots.

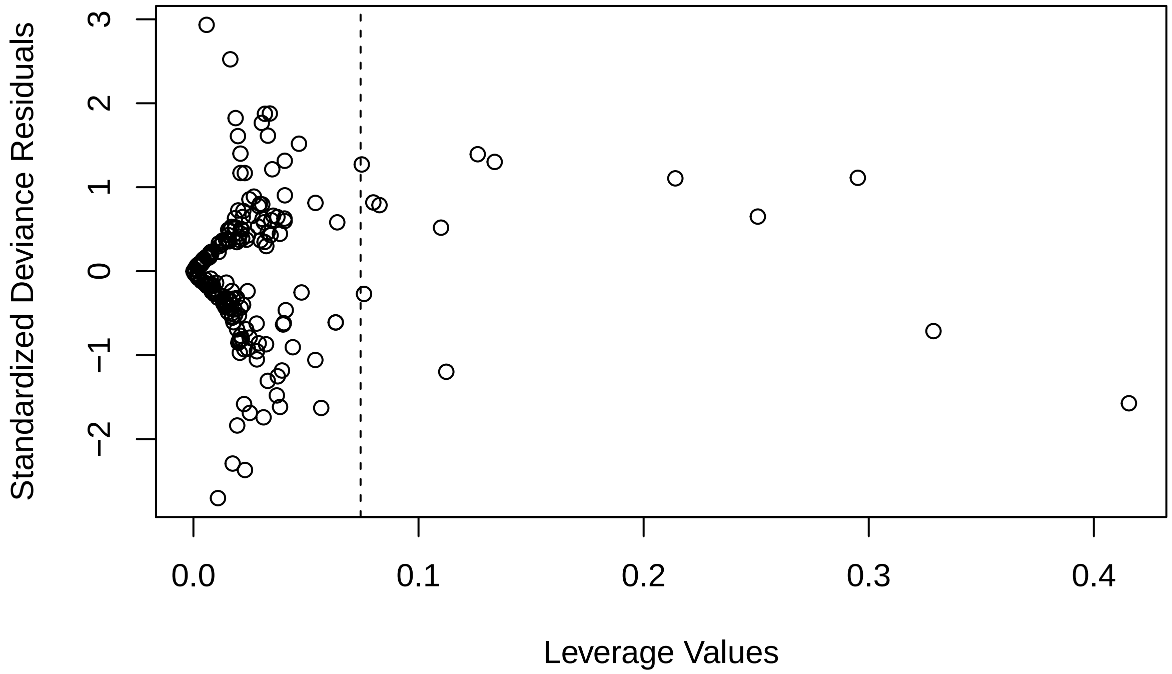
We can now compare the two models using Deviance method.

```
# Analysis of Deviance table

anova(m1,m2,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Sex ~ RCC + WCC + BMI
## Model 2: Sex ~ RCC + WCC + log(WCC) + log(BMI) + BMI
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       198     147.01
## 2       196     132.66  2   14.355 0.0007635 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see from this anlaysis that Model two is significantly better than Model first.

Let us now plot leverage against standardized deviance residuals for model2.



The cut-off value for the leverage is $2(p+1)/n = 2 * 14/202 = 0.1485$.

```
out1
```

```
##        22        36        75       178       194
## 0.2951484 0.2507288 0.2140918 0.4155865 0.3287565
```

Observations 22, 36, 75, 178 and 194 have the largest leverage values. Observations 22, 36 and 75 corresponds to the female and obervations 178 and 194 corresponds to the male. We need to analyse these seperately, but that goes out of the scope since the information is not provided.

Then the output from R for the model 2 is shown below:

```
m2<-glm(Sex~RCC+WCC+ log(WCC) + BMI+ log(BMI),family = binomial(),data=ais)
summary(m2)
```

```
## 
## Call:
## glm(formula = Sex ~ RCC + WCC + log(WCC) + BMI + log(BMI), family = binomial(), 
##     data = ais)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -2.68796  -0.43021  -0.01501   0.47294   2.92495  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) 145.6087    44.5255   3.270  0.00107 ** 
## RCC          -6.2456     0.8787  -7.107 1.18e-12 ***
## WCC          -1.3425     0.7135  -1.882  0.05989 .
## log(WCC)     11.4725     5.1742   2.217  0.02661 *
## BMI           2.1885     0.8182   2.675  0.00747 **
## log(BMI)    -57.2779    19.8596  -2.884  0.00392 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 280.01  on 201  degrees of freedom
## Residual deviance: 132.66  on 196  degrees of freedom
## AIC: 144.66
##
## Number of Fisher Scoring iterations: 6
```

The variable WCC is only marginally significant(Wald p-value = 0.05989). We shall remove the variable from the model for now. Thus, we consider the following model:

$ m3 = glm(Sex~RCC+Log(WCC)+BMI+log(BMI),family=binomial(),data=ais)$

Now lets test the hypothesis, that

$H_0 : \beta_2 = 0 \ (i.e., m3)$

against

$H_A : \beta_2 \neq 0 \ (i.e., m2)$

```
m3 = glm(Sex~RCC+log(WCC)+BMI+log(BMI),family=binomial(),data=ais)
anova(m3,m2,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Sex ~ RCC + log(WCC) + BMI + log(BMI)
## Model 2: Sex ~ RCC + WCC + log(WCC) + BMI + log(BMI)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       197     136.58
## 2       196     132.66  1   3.9269  0.04752 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It seems that the p-value from the differences in deviance (p-value=0.047) is lower than the corresponding wald p-value for WCC (0.059). This shows that the wald test and test based on the difference in deviances can result different p-values. Hence, we can say model 2 is preferred over model3.

Hence, the final model is $ Sex ~ RCC + WCC + log(WCC) + BMI + log(BMI) $

Extension to Exercise 8.3.3 [10 points extra credit]

Create fake data from your model developed in Exercise 8.3.3 to test that you code is implemented properly.

Ans:

Let us try to generate fake data (gender) using the model coefficients we have seen from the model we have just generated in 8.3.3. I have used the same data available to generate y value, and then come up with a fake fitted model.

```r
summary(m2)
```

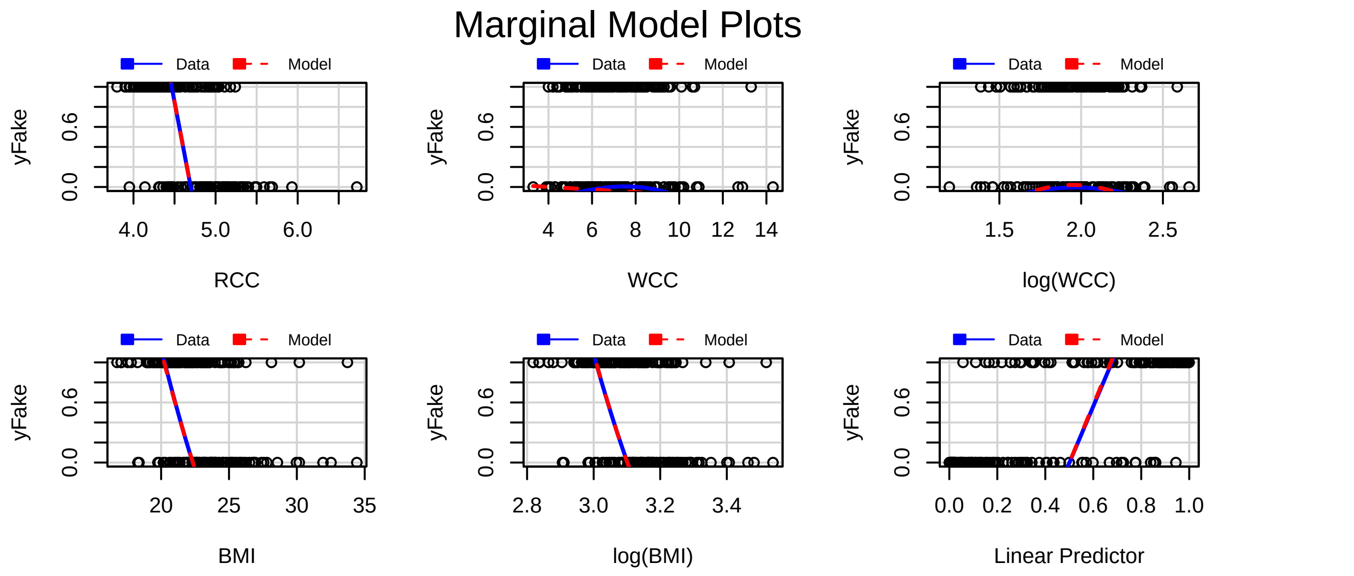
```
##
## Call:
## glm(formula = Sex ~ RCC + WCC + log(WCC) + BMI + log(BMI), family = binomial(),
##     data = ais)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.68796  -0.43021  -0.01501   0.47294   2.92495
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 145.6087    44.5255   3.270  0.00107 **
## RCC          -6.2456     0.8787  -7.107 1.18e-12 ***
## WCC          -1.3425     0.7135  -1.882  0.05989 .
## log(WCC)     11.4725     5.1742   2.217  0.02661 *
## BMI           2.1885     0.8182   2.675  0.00747 **
## log(BMI)    -57.2779    19.8596  -2.884  0.00392 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 280.01  on 201  degrees of freedom
## Residual deviance: 132.66  on 196  degrees of freedom
## AIC: 144.66
##
## Number of Fisher Scoring iterations: 6
```

```r
set.seed(500)
n<- 202
b0<- (145.6087)
b1<- (-6.2456)
b2<- (-1.3425)
b3<- (11.4725)
b4<- (2.1885)
b5<- (-57.2779)

theta_x <- 1/( 1 + exp(-(b0+b1*RCC+b2*WCC+b3*log(WCC)+ b4*BMI+ b5*log(BMI))) )
yFake <- rbinom(n = n, size = 1, prob = theta_x )

DataFrame=data.frame(y=yFake,x1=RCC,x2=WCC,x3=BMI)
modelFake<-glm(yFake~RCC+WCC+log(WCC)+BMI+log(BMI),family=binomial())
summary(modelFake)
```

```
##
## Call:
## glm(formula = yFake ~ RCC + WCC + log(WCC) + BMI + log(BMI),
##     family = binomial())
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.40472  -0.51753  -0.07164   0.53525   2.39350
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 143.8986    43.7177   3.292 0.000996 ***
## RCC          -5.1722     0.7504  -6.892 5.49e-12 ***
## WCC          -1.4070     0.6835  -2.059 0.039541 *
## log(WCC)     12.6830     4.9873   2.543 0.010988 *
## BMI           2.2210     0.8036   2.764 0.005713 **
## log(BMI)    -59.1950    19.5644  -3.026 0.002481 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 279.95  on 201  degrees of freedom
## Residual deviance: 149.23  on 196  degrees of freedom
## AIC: 161.23
##
## Number of Fisher Scoring iterations: 6
```

```
car::mmps(modelFake,layout=c(3,3))
```



Firstly, we observe that the coefficients of the new model are similar to the true coefficients, and the confidence interval contain them.

Also, we observe from the above figure that the the lines in marginal model plots are coherent with each other indicating a good fit of the model.

Thus, fake data generation is completed.

References

1. http://www.grantschissler.com/teaching/SP18/STAT757/fake_data_logistic_html.html
2. http://www.stat.tamu.edu/~sheather/book/
3. Sheather, Simon. 2009. A Modern Approach to Regression with R. Springer Science & Business Media.