STA302/1001H1S - Method of Data Analysis

Assignment 2-Solution

Due March 28th, 23:59

Instructions: This is individual assignment. It is worth 100 points. Please use Rmarkdown to write your solutions and submit your solutions with relevant R code included as a pdf file via **Crowdmark**.

Question

Census data was collected on the 50 states and Washington, D.C. We are interested in determining whether average lifespan (LIFE) is related to the ratio of males to females in percent (MALE), birth rate per 1,000 people (BIRTH), divorce rate per 1,000 people (DIVO), number of hospital beds per 100,000 people (BEDS), percentage of population 25 years or older having completed 16 years of school (EDUC) and per capita income (INCO). The data stored in the data file Census.txt can be found on the course webside.

Answer the following questions.

Part 0: Data Set

```
Census = read.table("Data/Census.txt", header=TRUE)
# 3 first observations
head(Census,3)
```

```
  STATE  MALE BIRTH DIVO  BEDS EDUC INCO  LIFE
1    AK 119.1  24.8  5.6 603.3 14.1 4638 69.31
2    AL  93.3  19.4  4.4 840.9  7.8 2892 69.05
3    AR  94.1  18.5  4.8 569.6  6.7 2791 70.66
```

```
# the number of observations
n=nrow(Census)
n
```

```
[1] 51
```

```
# Design matrix : Matrix of predictors including column for intercept
X=as.matrix( cbind( I=rep(1,n),  Census[,-c(1,8)]) )
head(X,3)
```

```
     I  MALE BIRTH DIVO  BEDS EDUC INCO
[1,] 1 119.1  24.8  5.6 603.3 14.1 4638
[2,] 1  93.3  19.4  4.4 840.9  7.8 2892
[3,] 1  94.1  18.5  4.8 569.6  6.7 2791
```

```
# set the number of predictors
p=ncol(X)-1
p
```

```
[1] 6
```

```
# Response variable
y=as.vector(Census[,8])
```

Part 1 (20 Marks) : In this Part, compute by hand using matrix formulas. DO NOT USE lm() command in this part.

We consider a multiple linear regression model with LIFE (y) as the response variable, and MALE (x_1), BIRTH (x_2), DIVO (x_3), BEDS (x_4), EDUC (x_5), and INCO (x_6), as predictors. Answer the following questions using least square estimates in term of matrix formulas.

(a) (6 marks) Compute and report the least-squares estimates. Write down the least-squares regression equation.

Solution:

The R code is given by:

```
# Vector ofLeast-Square-Estimates
Xt<-t(X)
XtX<-Xt%*%X
Xty<-Xt%*%y
(beta_hat<-solve(XtX)%*%Xty)
```

```
##                [,1]
## I     70.5577812705
## MALE   0.1261018758
## BIRTH -0.5160557876
## DIVO  -0.1965375074
## BEDS  -0.0033392036
## EDUC   0.2368222541
## INCO  -0.0003612011
```

(2 marks for R code)

The vector of least-square estimates is:

$$\hat{\beta} = (70.55778, 0.1261, -0.51606, -0.19654, -0.00334, 0.23682, -3.6 \times 10^{-4})$$

(2 marks)

The least-squares regression equation is

$$\hat{y} = 70.558 + 0.126x_1 - 0.516x_2 - 0.197x_3 - 0.003x_4 + 0.23682x_5 - 3.6 \times 10^{-4}x_6$$

(2 marks)

(b) (4 marks) Explain in context what the coefficients corresponding to MALE and BIRTH mean.

Solution:

The regression coefficient for MALE(x_1) = $\hat{\beta}_1 = 0.126$ represents the estimated average increase in lifespan (LIFE) for 1-percent increase in the ratio of males to females, with all others prediction BIRTH, DIVO, BEDS, EDUC, and INCO held fixed. **(2 marks)**

The regression coefficient for BIRTH(x_2) = $\hat{\beta}_2 = -0.516$ tells us that, for a fixed values of MALE, DIVO, BEDS, EDUC, and INCO, the average lifespan (LIFE) decreases by about 0.516 for every 1 per 1000 people increase in the birth rate. **(2 marks)**

(c) (3 marks) Compute the biased and the unbiased estimates of the error variance σ^2.

Solution:

The R code is given by:

```
# Vecor of Fitted values
y_hat<-X%*%beta_hat
# Vector of residuals
e_hat<-y-y_hat
#Sum of Squares Residuals
SSR<-sum( e_hat^2 )
# Biased estimate of sigma2
(S2biased=SSR/n)
```

```
## [1] 1.192215
```

```
# Unbiased estimate of sigma2
(S2unbiased=SSR/(n-p-1) )
```

```
## [1] 1.381885
```

(1 mark for R code)

The biased estimate of σ^2 is $\hat{\sigma}^2 = \frac{SSR}{n} = 1.1922148$ **(1 mark)**

and

The unbiased estimate of σ^2 is $\hat{\sigma}^2 = \frac{SSR}{n-p-1} = 1.3818853$ **(1 mark)**

(d) (4 marks) Using the unbiased estimate of error variance, Compute the standard errors of the estimators of the regression coefficients.

Solution:

The R code is given by

```
s2<-SSR/(n-p-1)
Inv_tXX<-solve(t(X)%*%X)
varcov<-s2*Inv_tXX
# vector of standard error associated with the least square estimates
(se.beta_hat=sqrt(diag(varcov)))
```

```
##             I         MALE        BIRTH         DIVO         BEDS
## 4.2897471299 0.0472317551 0.1172774621 0.0739532971 0.0009795303
##         EDUC         INCO
## 0.1110224835 0.0004597943
```

(1 mark for the R code)

The vector of estimated standard error is:

$$s.e(\hat{\beta}) = (4.28975, 0.04723, 0.11728, 0.07395, 9.8 \times 10^{-4}, 0.11102, 4.6 \times 10^{-4})$$

(3 marks)

(e) (3 marks) Compute the coefficient of determination. Give a practical interpretation of your result.

Solution:

The R code is given by

```
#Sum of Squares Residuals
RSS<-sum( e_hat^2 )
#Total Sum of Squares
SST<-sum( (y-mean(y))^2 )
#SSreg
(SSreg<-SST-RSS)
```

```
## [1] 53.59425
```

```
(SSreg<-sum( (y_hat-mean(y))^2 ) )
```

```
## [1] 53.59425
```

```
# R-squared can be obtain using RSS and SST
(R2<-1-RSS/SST)
```

```
## [1] 0.4684927
```

```
# OR using SSreg and SST
(R2<-SSreg/SST)
```

```
## [1] 0.4684927
```

(1 mark for the R code)

The coefficient of determination is $R^2 = 0.47$, indicating that 47% of the variability in lifespan (LIFE) is explained by the regression. **(2 marks)**

Part 2 (60 Marks): **In this part, you may use all R commands you need, including lm() function, to answer the following questions.**

(a)(2 marks) Fit the MLR model with LIFE (y) as the response variable, and MALE (x_1), BIRTH (x_2), DIVO (x_3), BEDS (x_4), EDUC (x_5), and INCO (x_6), as predictors.

Solution:

The R code is given by

```
m1 = lm(LIFE ~ MALE + BIRTH + DIVO +
               BEDS + EDUC + INCO, data=Census)
out1=summary(m1)
out1
```

```
##
## Call:
## lm(formula = LIFE ~ MALE + BIRTH + DIVO + BEDS + EDUC + INCO,
##     data = Census)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5563 -0.6629  0.0755  0.6983  3.3215
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 70.5577813  4.2897471  16.448  < 2e-16 ***
## MALE         0.1261019  0.0472318   2.670  0.01059 *
## BIRTH       -0.5160558  0.1172775  -4.400 6.78e-05 ***
## DIVO        -0.1965375  0.0739533  -2.658  0.01093 *
## BEDS        -0.0033392  0.0009795  -3.409  0.00141 **
## EDUC         0.2368223  0.1110225   2.133  0.03853 *
## INCO        -0.0003612  0.0004598  -0.786  0.43633
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.176 on 44 degrees of freedom
## Multiple R-squared:  0.4685, Adjusted R-squared:  0.396
## F-statistic: 6.464 on 6 and 44 DF,  p-value: 6.112e-05
```

(2 marks)

(b)(5 marks) At level $\alpha = 5\%$, conduct the F-test for the overall fit of the regression. Comment on the results.

Solution:

To test the null Hypothesis

$$\mathrm{H}_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$

against

$$\mathrm{H}_a : \text{at least one } \beta \neq 0$$

(1 mark)

We use The F-test statistic

$$\mathrm{F} = \frac{\frac{\mathrm{SSReg}}{6}}{\frac{\mathrm{SSR}}{44}}$$

which has F-distribution with $df1 = 6$ and $df2 = 44$ degrees of freedom. **(1 mark)**

The output shows that F= 6.464 (p-value=6.112e-05), indicating that we should clearly reject the null hypothesis that the variables all predictor collectively have no effect on the response variable LIFE.

(3 marks)

(c)(9 marks) At level $\alpha = 1\%$, test each of the individual regression coefficients. Do the results indicate that any of the explanatory variables should be removed from the model?

To test the null Hypothesis

$$\mathrm{H}_0 : \beta_j = 0, \qquad j = 1, \ldots, 6$$

against

$$\mathrm{H}_a : \beta_j \neq 0, \qquad j = 1, \ldots, 6$$

(1 mark)

We use the T-test statistic

$$\mathrm{T}_j = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}$$

which has T-distribution with $df = 44$ degrees of freedom.

(1 mark)

At level $\alpha = 1\%$, the R output show that:

the variable MALE is not significant, controlling for the other variables ($pvalue = 0.010591 > \alpha$).

the variable BIRTH is significant for fixed values of the other variables ($pvalue = 6.7832971 \times 10^{-5} < \alpha$)

the variable DIVO is not significant, controlling for the other variables ($pvalue = 0.0109277 > \alpha$)

the variable BEDS is significant controlling for the other variables ($pvalue = 0.0014057 < \alpha$)

the variable EDUC is not significant controlling for the other variables ($pvalue = 0.0385317 > \alpha$)

the variable INCO is not significant controlling for the other variables ($pvalue = 0.4363288 > \alpha$).

(6 marks)

Thus, at level $\alpha = 1\%$, based on the above results, the variables MALE, DIVO, EDUC, and INCO should be removed from the model.

(1 mark)

(d)(3 marks) Determine the regression model with the explanatory variable(s) identified in part (c) removed. Write down the estimated regression equation.

The estimated regression equation using the significant variables is

$$y = 70.55778 - 0.51606x_2 - 0.00334x_4$$

(3 marks)

(e)(6 marks) Perform a partial F-test at level $\alpha = 1\%$ to determine whether the variables associated with MALE and INCO can be removed from the model

Solution:

The R code is given by

```
reduced = lm(LIFE ~ BIRTH + DIVO + BEDS + EDUC, data=Census)
full = lm(LIFE ~ MALE + BIRTH + DIVO + BEDS + EDUC + INCO, data=Census)
anova(reduced,full)
```

```
## Analysis of Variance Table
##
## Model 1: LIFE ~ BIRTH + DIVO + BEDS + EDUC
## Model 2: LIFE ~ MALE + BIRTH + DIVO + BEDS + EDUC + INCO
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     46 70.654
## 2     44 60.803  2    9.8507 3.5642 0.03676 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(1 mark)

To test the null Hypothesis is

$$\mathrm{H}_0 : \beta_2 = \beta_6 = 0$$

against

$$\mathrm{H}_a : \text{at least one of } \beta_2 \text{ and } \beta_6 \neq 0$$

(1 mark)

We use The partial F-test statistic

$$\mathrm{F} = \frac{\frac{\mathrm{RSS}_r - \mathrm{RSS}_f}{2}}{\frac{\mathrm{SSR}}{44}}$$

which has F-distribution with $df1 = 2$ and $df2 = 44$ degrees of freedom.

(1 mark)

The above R output shows the results of the partial F-test. Since F=3.5642106 (p-value=0.0367609), we cannot reject the null hypothesis ($\beta_2 = \beta_6 = 0$) at the 1% level of significant. It appears that the variables MALE, INCO do not contribute significant information to the LIFE once the variables BIRTH, DIVO, BEDS and EDUC have been taken into consideration.

(3 marks)

(f)(5 marks) Compute and report the F test statistic for comparing the two models

$$\mathbf{E}(Y_i|x_i) = \beta_0 + \beta_1 x_{i1},$$

$$\mathbf{E}(Y_i|x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6},$$

Solution:

First, we fit the two models using the following R code

```
m1 = lm(LIFE ~ MALE, data=Census)
m2 = lm(LIFE ~ MALE + BIRTH + DIVO + BEDS + EDUC + INCO, data=Census)
comp=anova(m1,m2)
comp
```

```
## Analysis of Variance Table
##
## Model 1: LIFE ~ MALE
## Model 2: LIFE ~ MALE + BIRTH + DIVO + BEDS + EDUC + INCO
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     49 109.834
## 2     44  60.803  5    49.031 7.0963 6.099e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(1 mark)

To compare the above models, model 1 (one predictor) with model 2 (six predictors)

We test the null Hypothesis

$$\mathrm{H}_0 : \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$

against

$$\mathrm{H}_a : \text{at least one of } \beta s \neq 0$$

We use the following partial F-statistic

$$\mathrm{F} = \frac{\frac{\mathrm{RSS}_1 - \mathrm{RSS}_2}{df_1 - df_2}}{\frac{\mathrm{RSS}_2}{df_2}}$$

which has F-distribution with $df_1 - df_2 = 5$ and $df_2 = 44$ degrees of freedom.

(1 mark)

From the above R output, we have

$$\text{RSS}_1 = 109.834, \quad \text{with} \quad \text{df}_1 = 49$$

$$\text{RSS}_2 = 60.803, \quad \text{with} \quad \text{df}_2 = 44$$

Thus,

$$\text{F} = \frac{\frac{109.834 - 60.803}{49 - 44}}{\frac{60.803}{44}} = 7.096$$

(3 marks)

(g)(6 marks) Perform a partial F-test at level $\alpha = 1\%$ for comparing the two models

$$\mathbf{E}(Y_i|x_i) = \beta_0,$$
$$\mathbf{E}(Y_i|x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2},$$

Solution:

First, we fit the two models using the following R code

```
m1 = lm(LIFE ~ 1, data=Census)
m2 = lm(LIFE ~ MALE + BIRTH, data=Census)
comp=anova(m1,m2)
comp
```

```
## Analysis of Variance Table
##
## Model 1: LIFE ~ 1
## Model 2: LIFE ~ MALE + BIRTH
##   Res.Df     RSS Df Sum of Sq    F    Pr(>F)
## 1     50 114.397
## 2     48  85.424  2    28.973 8.14 0.0009036 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(1 mark)

To compare the null model (no predictor) with model 2 (two predictors)

We test the null Hypothesis

$$\mathrm{H}_0 : \beta_2 = \beta_3 = 0$$

against

$$\mathrm{H}_a : \text{at least one of } \beta s \neq 0$$

(1 mark)

We use the following partial F-statistic

$$\mathrm{F} = \frac{\frac{\mathrm{RSS}_1 - \mathrm{RSS}_2}{df_1 - df_2}}{\frac{\mathrm{RSS}_2}{df_2}}$$

which has F-distribution with $df_1 - df_2 = 2$ and $df_2 = 48$ degrees of freedom.

(1 mark)

The above R output shows that F = 8.14(pvalue=9.0357126×10^{-4}), indicating that we can reject H_0 at level $\alpha = 5\%$ of significant. It appears that the variables MALE and BIRTH do contribute significant information to LIFE.

(3 marks)

(h)(12 marks) Compute and report the terms in the decomposition

$$\text{SSreg}(\beta_1, \beta_2, \beta_3|\beta_0) = \text{SSreg}(\beta_3|\beta_0) + \text{SSreg}(\beta_2|\beta_0, \beta_3) + \text{SSreg}(\beta_1|\beta_0, \beta_3, \beta_2)$$

Solution:

The first term $\text{SSreg}(\beta_1, \beta_2, \beta_3|\beta_0)$ is obtained from the comparison of the following two models.

```
m1 = lm(LIFE ~ 1, data=Census)
m2 = lm(LIFE ~ MALE+ BIRTH + DIVO, data=Census)
comp=anova(m1,m2)
comp
```

```
## Analysis of Variance Table
##
## Model 1: LIFE ~ 1
## Model 2: LIFE ~ MALE + BIRTH + DIVO
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     50 114.397
## 2     47  80.751  3    33.646 6.5277 0.0008795 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(1 mark)

We have that

$$\text{SSreg}(\beta_1, \beta_2, \beta_3|\beta_0) = \text{RSS}(\beta_0) - \text{RSS}(\beta_0, \beta_1, \beta_2, \beta_3) = 114.4 - 80.75 = 33.65$$

(2 marks)

The term $\text{SSreg}(\beta_3|\beta_0)$ is obtained from the comparison of the following two models.

```
m1 = lm(LIFE ~ 1, data=Census)
m2 = lm(LIFE ~ DIVO, data=Census)
comp=anova(m1,m2)
comp
```

```
## Analysis of Variance Table
##
## Model 1: LIFE ~ 1
## Model 2: LIFE ~ DIVO
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     50 114.40
## 2     49 111.09  1    3.3073 1.4588 0.2329
```

(1 mark)

We have that

$$\text{SSreg}(\beta_3|\beta_0) = \text{RSS}(\beta_0) - \text{RSS}(\beta_0, \beta_3) = 114.4 - 111.09 = 3.31$$

(2 marks)

The term $\text{SSreg}(\beta_2|\beta_0, \beta_3)$ is obtained from the comparison of the following two models.

```
m1 = lm(LIFE ~ DIVO, data=Census)
m2 = lm(LIFE ~ BIRTH + DIVO, data=Census)
comp=anova(m1,m2)
comp
```

```
## Analysis of Variance Table
##
## Model 1: LIFE ~ DIVO
## Model 2: LIFE ~ BIRTH + DIVO
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     49 111.09
## 2     48 102.17  1    8.9145 4.1879 0.04621 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(1 mark)

We have that

$$\text{SSreg}(\beta_2|\beta_0, \beta_3) = \text{RSS}(\beta_0, \beta_3) - \text{RSS}(\beta_0, \beta_2, \beta_3) = 111.09 - 102.18 = 8.91$$

(2 marks)

The term $\text{SSreg}(\beta_1|\beta_0, \beta_2, \beta_3)$ is obtained from the comparison of the following two models.

```
m1 = lm(LIFE ~ BIRTH + DIVO, data=Census)
m2 = lm(LIFE ~ MALE+ BIRTH + DIVO, data=Census)
comp=anova(m1,m2)
comp
```

```
## Analysis of Variance Table
##
## Model 1: LIFE ~ BIRTH + DIVO
## Model 2: LIFE ~ MALE + BIRTH + DIVO
##   Res.Df     RSS Df Sum of Sq     F    Pr(>F)
## 1     48 102.175
```

```
## 2     47  80.751  1    21.424 12.47 0.0009384 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(1 mark)

We have that

$$\text{SSreg}(\beta_1|\beta_0, \beta_2, \beta_3) = \text{RSS}(\beta_0, \beta_2, \beta_3) - \text{RSS}(\beta_0, \beta_1, \beta_2, \beta_3) = 102.18 - 80.75 = 21.43$$

(2 marks)

(i)(12 marks) Suppose we are interested in fitting a regression model using LIFE as the response variable and some subset of the variables (MALE, BIRTH, DIVO, and INCO) as predictor.

(i.1) Perform variable selection by finding the subset model that minimizes the AIC criteria. State the 'best model'.

Compute the AIC of all possible models:

```
# models with one predictor
aic1= AIC( lm(LIFE ~ MALE, data=Census) )
aic2= AIC( lm(LIFE ~ BIRTH, data=Census) )
aic3= AIC( lm(LIFE ~ DIVO, data=Census) )
aic4= AIC( lm(LIFE ~ INCO, data=Census) )


# models with two predictors
aic5= AIC( lm(LIFE ~ MALE + BIRTH, data=Census) )
aic6= AIC( lm(LIFE ~ MALE + DIVO, data=Census) )
aic7= AIC( lm(LIFE ~ MALE + INCO, data=Census) )
aic8= AIC( lm(LIFE ~ BIRTH + DIVO, data=Census) )
aic9= AIC( lm(LIFE ~ BIRTH + INCO, data=Census) )
aic10= AIC( lm(LIFE ~ DIVO + INCO, data=Census) )

# models with three predictors

aic11= AIC( lm(LIFE ~ MALE + BIRTH + DIVO, data=Census) )
aic12= AIC( lm(LIFE ~ MALE + BIRTH + INCO, data=Census) )
aic13= AIC( lm(LIFE ~ BIRTH + DIVO  + INCO, data=Census) )

# full models
aic14= AIC( lm(LIFE ~ MALE + BIRTH + DIVO + INCO, data=Census) )

# define vector of all aic
aic=c(aic1,aic2,aic3,aic3,aic4, aic5, aic6, aic7, aic8,
      aic9, aic10, aic11, aic12, aic13, aic14)
aic
```

```
##  [1] 189.8561 186.7525 190.4359 190.4359 191.2477 179.0377 188.5535
##  [8] 191.4429 188.1698 188.5226 191.4755 178.1686 180.9320 189.8012
## [15] 180.1406
```

```
# which model minimizes the aic
which.min(aic)
```

```
## [1] 12
```

```
# aic min
aic.best=aic[which.min(aic)]
aic.best
```

```
## [1] 178.1686
```

Based on AIC, the best model is model 12 the one that includes the variables MALE, BIRTH and DIVO.

(4 marks)

(i.2) Perform variable selection using forward selection. State the 'best model'.

To perform forward selection, we use the following code

```
null=lm(LIFE ~ 1, data=Census)
full=lm(LIFE ~ MALE + BIRTH + DIVO + INCO, data=Census)
step(null, scope=list(lower=null, upper=full), direction="forward")
```

```
## Start:  AIC=43.2
## LIFE ~ 1
##
##         Df Sum of Sq    RSS    AIC
## + BIRTH  1   11.0479 103.35 40.021
## + MALE   1    4.5632 109.83 43.124
## <none>               114.40 43.200
## + DIVO   1    3.3073 111.09 43.704
## + INCO   1    1.5249 112.87 44.516
##
## Step:  AIC=40.02
## LIFE ~ BIRTH
##
##        Df Sum of Sq     RSS    AIC
## + MALE  1   17.9252  85.424 32.306
## <none>              103.349 40.021
## + DIVO  1    1.1739 102.175 41.438
## + INCO  1    0.4647 102.885 41.791
##
## Step:  AIC=32.31
## LIFE ~ BIRTH + MALE
##
##        Df Sum of Sq    RSS    AIC
## + DIVO  1    4.6730 80.751 31.437
## <none>              85.424 32.306
## + INCO  1    0.1768 85.247 34.200
##
## Step:  AIC=31.44
## LIFE ~ BIRTH + MALE + DIVO
##
##        Df Sum of Sq    RSS    AIC
## <none>              80.751 31.437
## + INCO  1  0.044334 80.707 33.409
```

```
##
## Call:
## lm(formula = LIFE ~ BIRTH + MALE + DIVO, data = Census)
##
## Coefficients:
## (Intercept)        BIRTH         MALE         DIVO
##     62.3656      -0.3912       0.1689      -0.1272
```

According to this procedure, the best model is the one that includes the variables MALE, BIRTH and DIVO.

(4 marks)

(i.3) Perform variable selection using backward selection. State the 'best model'.

To performe backward selection, we use the following R code

```
null=lm(LIFE ~ 1, data=Census)
full=lm(LIFE ~ MALE + BIRTH + DIVO + INCO, data=Census)
step(full, data=Census, direction="backward")
```

```
## Start:  AIC=33.41
## LIFE ~ MALE + BIRTH + DIVO + INCO
##
##         Df Sum of Sq     RSS    AIC
## - INCO   1    0.0443  80.751 31.437
## <none>                 80.707 33.409
## - DIVO   1    4.5405  85.247 34.200
## - MALE   1   20.7328 101.440 43.070
## - BIRTH  1   20.9838 101.691 43.196
##
## Step:  AIC=31.44
## LIFE ~ MALE + BIRTH + DIVO
##
##         Df Sum of Sq     RSS    AIC
## <none>                 80.751 31.437
## - DIVO   1     4.673  85.424 32.306
## - MALE   1    21.424 102.175 41.438
## - BIRTH  1    22.196 102.947 41.822
```

```
##
## Call:
## lm(formula = LIFE ~ MALE + BIRTH + DIVO, data = Census)
##
## Coefficients:
## (Intercept)         MALE        BIRTH         DIVO
##     62.3656       0.1689      -0.3912      -0.1272
```

According to this procedure, the best model is the one that includes the variables MALE, BIRTH and DIVO.

(4 marks)

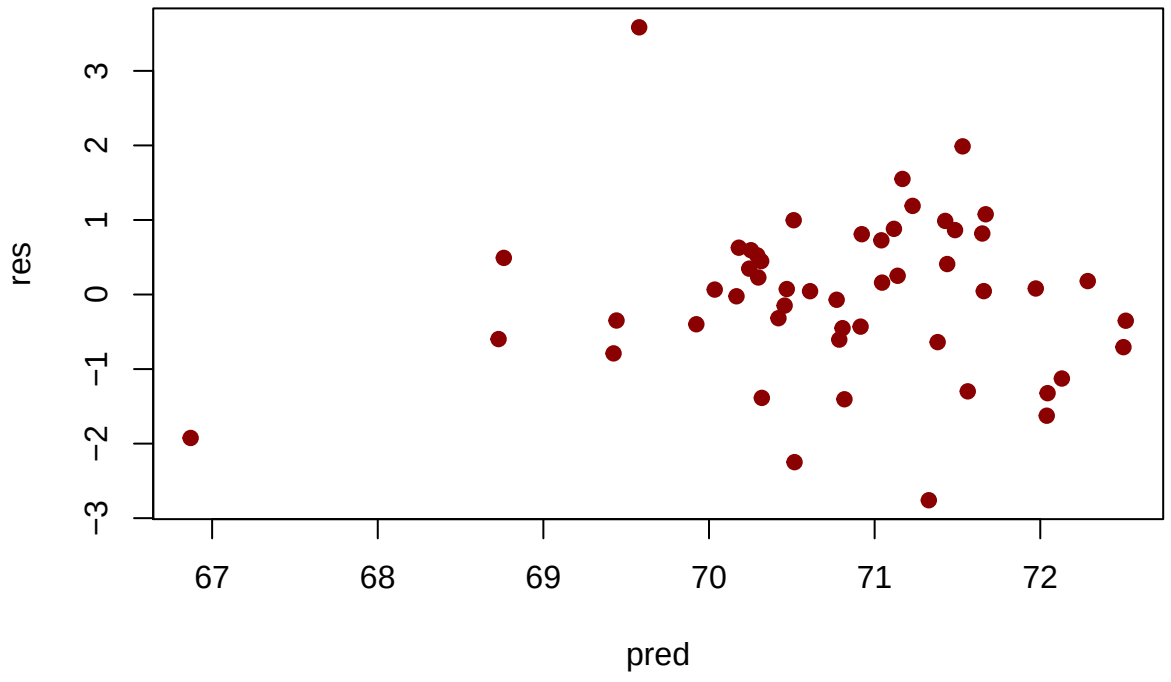
Part 3 (20 Marks): In this part, you may use all R commands you need.

We consider the multiple linear regression with LIFE (y) as the response variable, and MALE, BIRTH, DIVO , BEDS, EDUC, and INCO, as predictors.

(a) (3.5 Marks) Plot the standardized residuals against the fitted values. Are there any notable points. In particular look for points with large residuals or that may be influential.

Solution:

```
Census = read.table("Data/Census.txt", header=TRUE)
fit.census = lm(LIFE ~ MALE + BIRTH + DIVO +
             BEDS + EDUC + INCO, data=Census)
res<-rstandard(fit.census)
pred<-fitted.values(fit.census)
plot(pred, res, pch=19, col="darkred")
```



(1 mark)

Note that there are three points with large residuals. These points with residuals outside the interval (-2,2), may be influential.

(1 mark)

To identify these points we type:

```
# Points with res < -2
Census[res< -2,]
```

```
##    STATE  MALE BIRTH DIVO  BEDS EDUC INCO  LIFE
## 1     AK 119.1  24.8  5.6 603.3 14.1 4638 69.31
## 41    SC  96.5  20.1  2.2 739.9  9.0 2951 67.96
```

```
# Points with res > 2
Census[res >2,]
```

```
##    STATE MALE BIRTH DIVO  BEDS EDUC INCO LIFE
## 45    UT 97.6  25.5  3.7 470.5   14 3169 72.9
```

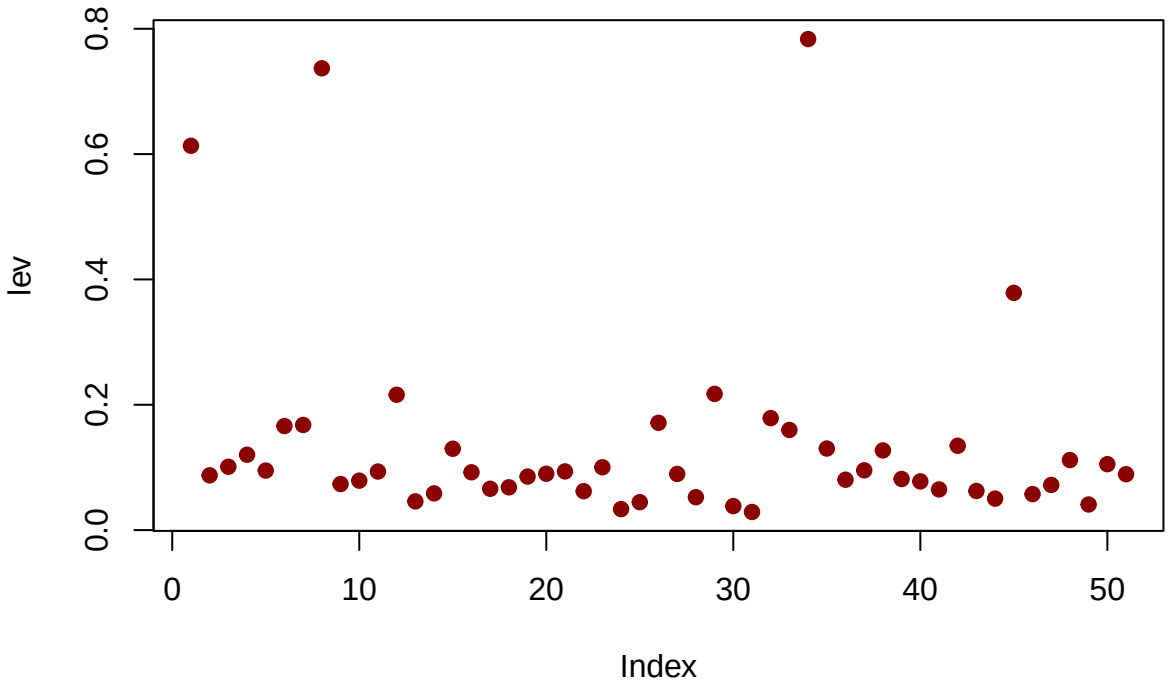
Thus, these points (1, 41, 45) with higher residuals corresponds to STATEs AK, SC and UT.

(1.5 mark)

(b) (2.5 Marks) Compute and plot the leverage of each point. Identify any points that have a leverage larger than 0.5.

Solution:

```
lev = hat(model.matrix(fit.census))
plot(lev, pch=19, col="darkred")
```



(1 mark)

Note that there are three points that have leverage larger than 0.5.

To identify these points, we type:

```
Census[lev >0.5,]
```

```
##    STATE  MALE BIRTH DIVO   BEDS EDUC INCO  LIFE
## 1     AK 119.1  24.8  5.6  603.3 14.1 4638 69.31
## 8     DC  86.8  20.1  3.0 1859.4 17.8 4644 65.71
## 34    NV 102.8  19.6 18.7  560.7 10.8 4583 69.03
```

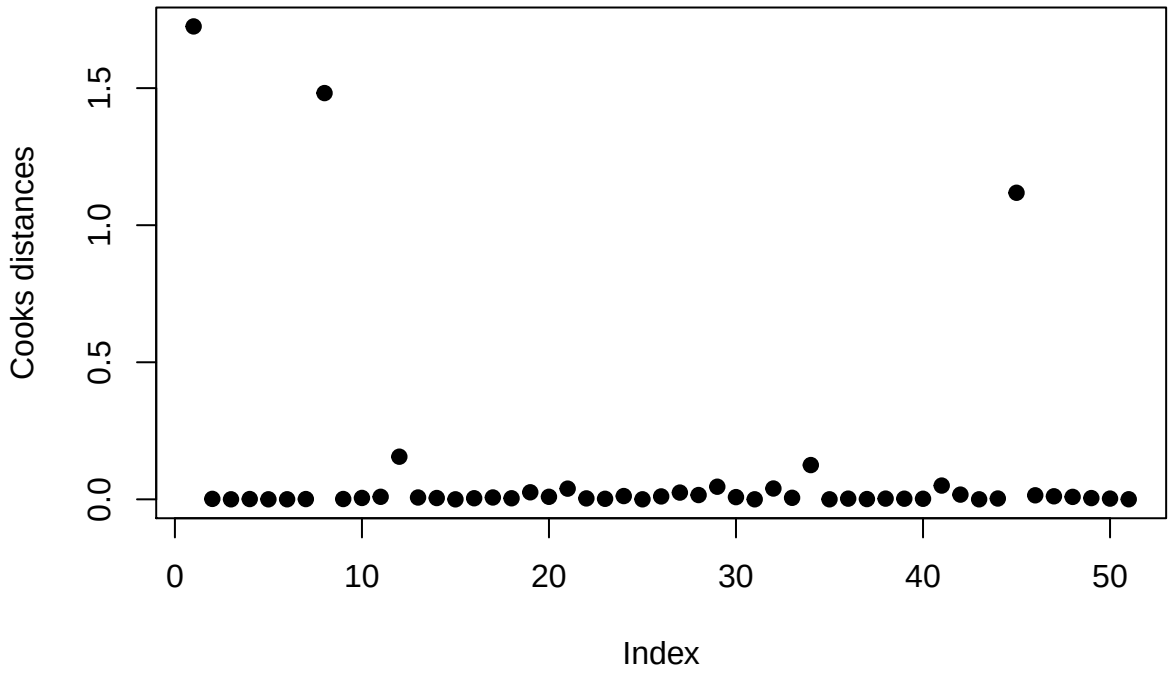
Thus, these points with leverage larger than 0.5 corresponds to STATEs AK, DC and NV.

(1.5 marks)

(c) (2.5 Marks) Compute the Cook's distance for each point. Identify any points that have a Cook's distance larger than 1. Are these the same observations as those seen in part (b)?

Solution:

```
cook = cooks.distance(fit.census)
plot(cook,ylab="Cooks distances", pch=19)
```



Note that there are three points that have Cooks Distance larger than 1.

(1 mark)

To identify these points type:

```
Census[cook >1,]
```

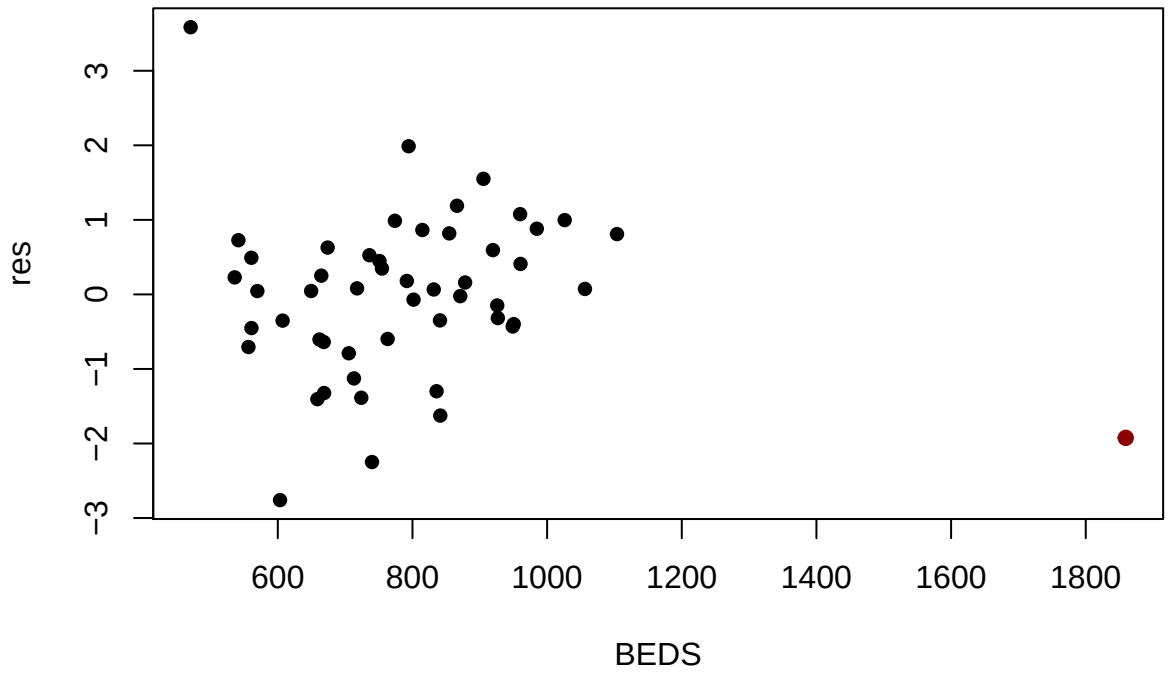
```
##    STATE  MALE BIRTH DIVO   BEDS EDUC INCO  LIFE
## 1     AK 119.1  24.8  5.6  603.3 14.1 4638 69.31
## 8     DC  86.8  20.1  3.0 1859.4 17.8 4644 65.71
## 45    UT  97.6  25.5  3.7  470.5 14.0 3169 72.90
```

Thus, these points with Cooks distance larger than 1 correspond to STATEs AK, DC and UT. Two of these states (AK and DC) have been identified in part (b), while the state UT have not been idetified in part (b).

(1.5 marks)

(d) (3 Marks) Plot the standardized residuals against the variable BEDS. Specifically mark the point corresponding to Washington, D.C. What can you say about this observation?

```
STATE=Census$STATE
BEDS=Census$BEDS
plot(BEDS, res, pch=16)
points(BEDS[STATE=="DC"], res[STATE=="DC"], col="darkred", pch=19)
```



(1 mark)

We note that the marked point (to the right, in dark red) has an BEDS-value which makes it distant from the other points on the BEDS-axis. Hence this point is a leverage point.

Notice also that the marked point do not follow the parttern set by the bulk of the others points. Thus this observation can be considered as outlier.

Since this point is a leverage point which is also an outlier, this point is called a bad leverage point.

(2 marks)

(e) (2 Marks) Remove the observation corresponding to Washington, D.C. and refit the model. Are there any notable differences with the model fit in part (a)?

Solution:

```
# fit with DC
fit.census = lm(LIFE ~ MALE + BIRTH + DIVO +
             BEDS + EDUC + INCO, data=Census)
res<-rstandard(fit.census)
pred<-fitted.values(fit.census)

# fit without DC
fit.census1 = lm(LIFE ~ MALE + BIRTH + DIVO +
             BEDS + EDUC + INCO, data=Census[STATE!="DC",])
res1<-rstandard(fit.census1)
pred1<-fitted.values(fit.census1)
```

To better understand the effects of the observation corresponding to Washington, D.C., we next examine the estimated regression coefficients R for BEDS.

```
# Based on all data
coefficients(summary(fit.census))[5,]
```

```
##      Estimate    Std. Error       t value      Pr(>|t|)
## -0.0033392036  0.0009795303 -3.4089845001  0.0014057446
```

```
# observation corresponding to state DC removed
coefficients(summary(fit.census1))[5,]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## -0.001163706  0.001448089 -0.803615002  0.426039604
```

Based on all the data, we found that the regression coefficient of BEDS is -0.00334, which is significant (pvalue=0.00141) . When we remove the observation corresponding to D.C., the regression coefficient of BEDS is -0.00116, as given in the table above. You will notice that -0.00116 is NOT significant (pvalue=0.42604). We see that removing this observation produces misleading results.

In addition, the following results,

```
# Based on all data
summary(fit.census)$r.squared
```

```
## [1] 0.4684927
```

```
# observation corresponding to state DC removed
summary(fit.census1)$r.squared
```

```
## [1] 0.3678871
```

that removing observation corresponding to D.C. decreases the R-square (0.4684927 versus 0.3678871).

We conclude by saying that the observation corresponding to state D.C. is not influential, since removing it from the data results to dramtically fit.

(2 marks)

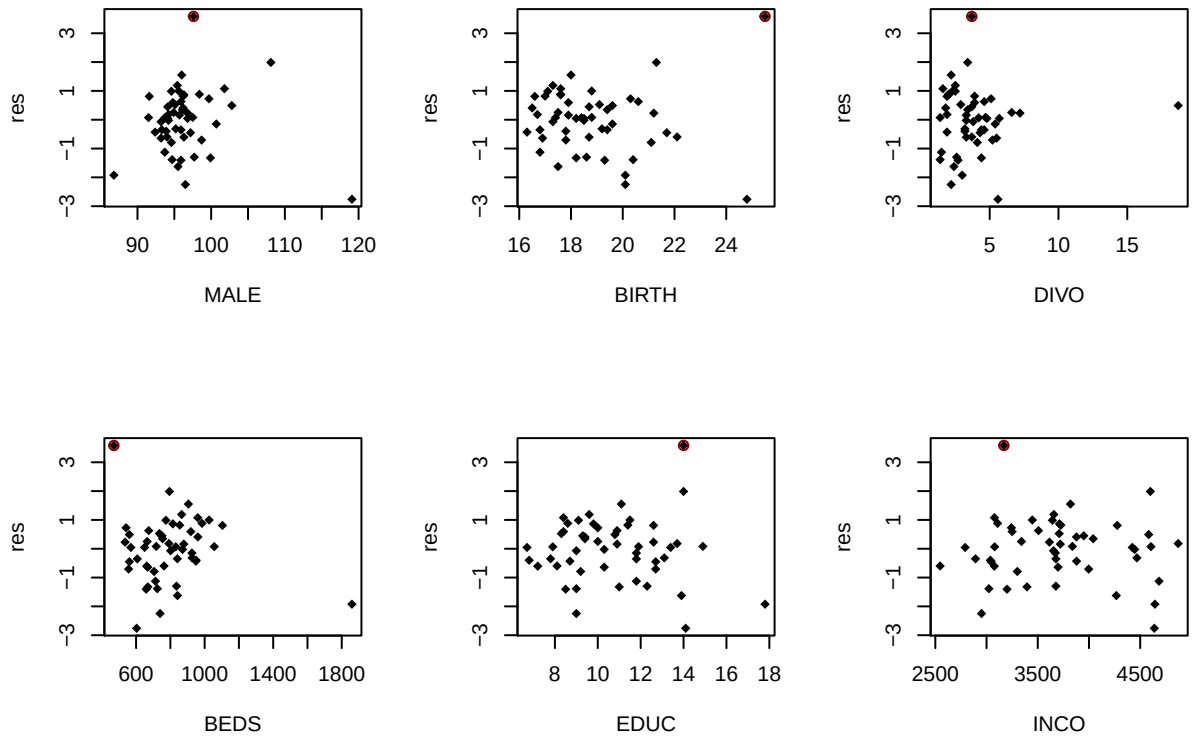
(f) (4.5 Marks) Plot the standardized residuals against each of the 6 explanatory variables. Specifically mark the observation corresponding to UT. What is notable about this state?

Solution:

```
STATE=Census$STATE
MALE=Census$MALE
BIRTH=Census$BIRTH
DIVO=Census$DIVO
BEDS=Census$BEDS
EDUC=Census$EDUC
INCO=Census$INCO

# fit
fit.census = lm(LIFE ~ MALE + BIRTH + DIVO +
             BEDS + EDUC + INCO, data=Census)
res<-rstandard(fit.census)

# Plot residuals against predictor
par(mfrow=c(2,3))
plot(MALE, res, pch=18)
points(MALE[STATE=="UT"], res[STATE=="UT"], col="darkred")
plot(BIRTH, res, pch=18)
points(BIRTH[STATE=="UT"], res[STATE=="UT"], col="darkred")
plot(DIVO, res, pch=18)
points(DIVO[STATE=="UT"], res[STATE=="UT"], col="darkred")
plot(BEDS, res, pch=18)
points(BEDS[STATE=="UT"], res[STATE=="UT"], col="darkred")
plot(EDUC, res, pch=18)
points(EDUC[STATE=="UT"], res[STATE=="UT"], col="darkred")
plot(INCO, res, pch=18)
points(INCO[STATE=="UT"], res[STATE=="UT"], col="darkred")
```



(3 marks for plots and R code)

As we see, all these plots show that the observation corresponding to state UT stand out from the other points. It can be considered as an outlier since the absolute value of the standardized residual associated with this observation is larger than 3.

(1.5 marks)

(g) (2 Marks) Remove the observation corresponding to UT and refit the model. Are there any notable differences with the model fit in part (a)? In particular, how does UT's exclusion impact the R^2 value?

Solution:

```
# fit all data
fit.census = lm(LIFE ~ MALE + BIRTH + DIVO +
             BEDS + EDUC + INCO, data=Census)

# fit data without state UT
fit.census1 = lm(LIFE ~ MALE + BIRTH + DIVO +
             BEDS + EDUC + INCO, data=Census[STATE!="UT",])

# Compare R-square
# all data
summary(fit.census)$r.squared
```

```
## [1] 0.4684927
```

```
# data without UT
summary(fit.census1)$r.squared
```

```
## [1] 0.6081066
```

Note that removing observation (State UT) from the data increases the value of R^2 and then appear to improve the fit. This point can be considered as influential.

(2 marks)