Assignment 2

Jonathan Wai, 1001472809

2019-03-29

Read Table in R

```
Dat=read.table("Census.txt", header=T)
head(Dat)
```

```
##   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
## 4    AZ  96.8  21.2  7.2 536.0 12.6 3614 70.55
## 5    CA  96.8  18.2  5.7 649.5 13.4 4423 71.71
## 6    CO  97.5  18.8  4.7 717.7 14.9 3838 72.06
```

```
MALE<-Dat$MALE
BIRTH<-Dat$BIRTH
DIVO<-Dat$DIVO
BEDS<-Dat$BEDS
EDUC<-Dat$EDUC
INCO<-Dat$INCO
LIFE<-Dat$LIFE
```

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

(a) Fit the MLR model with LIFE (y) as the response variable, and MALE (x1), BIRTH(x2), DIVO (x3), BEDS (x4), EDUC (x5), and INCO (x6), as predictors.

```
multiple.regression <- lm(LIFE ~ MALE + BIRTH + DIVO + BEDS + EDUC + INCO , data=Dat)
summary(multiple.regression)
```

```
##
## Call:
## lm(formula = LIFE ~ MALE + BIRTH + DIVO + BEDS + EDUC + INCO,
##     data = Dat)
##
## 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
```

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

```
summary(multiple.regression)
```

```
##
## Call:
## lm(formula = LIFE ~ MALE + BIRTH + DIVO + BEDS + EDUC + INCO,
##     data = Dat)
##
## 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
```

We test the following Hypothesis:

H0 : B1 =B2 = B3 = B4 = B5 = B6 = 0, HA : At least some of Bi != 0

At level alpha = 0.05, we reject if pvalue of the significant level is less than alpha = 0.05

The ANOVA TABLE shows that P value = 0.00006112, indicating that we should clearly reject the null hypothesis

At least 1 of the coefficients is not zero, therefore, overall the model is significant

Model fits the data better than the intercept-only model.

(c) At level alpha = 0.01, test each of the individual regression coefficients. Do the results indicate that any of the explanatory variables should be removed from the model?

At alpha = 0.01, explanatory variables that should be removed includes: (EDUC, DIVO, MALE,INCO)

Not significant at 1% level

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

```
NEWmultiple.regression <- lm(LIFE ~ BIRTH + BEDS, data=Dat)
summary(NEWmultiple.regression)
```

```
##
## Call:
## lm(formula = LIFE ~ BIRTH + BEDS, data = Dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5627 -0.8180 -0.0819  0.9261  3.6202
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 79.1473186  2.2717401  34.840  < 2e-16 ***
## BIRTH       -0.3281679  0.1026214  -3.198  0.00245 **
## BEDS        -0.0027415  0.0009388  -2.920  0.00531 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.352 on 48 degrees of freedom
## Multiple R-squared:  0.2329, Adjusted R-squared:  0.2009
## F-statistic: 7.286 on 2 and 48 DF,  p-value: 0.001725
```

The estimated regression equation is:

LIFE = 79.1473186 - 0.3281679 * BIRTH - 0.0027415 * BEDS

(e) 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

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

This is partial f test:

H0: B1=B2=B3=B4=0 , k < p

HA: H0 is not true

Fail to Reject Null Hypothesis

Coefficient Male and INCO does not significantly improve model, given all others included

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

```
Male.regression <- lm(LIFE ~ MALE, data = Dat)
summary(Male.regression)
```

```
##
## Call:
## lm(formula = LIFE ~ MALE, data = Dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4427 -0.6678  0.1118  1.1391  2.1956
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 64.38059    4.49565  14.321   <2e-16 ***
## MALE         0.06650    0.04661   1.427     0.16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.497 on 49 degrees of freedom
## Multiple R-squared:  0.03989,    Adjusted R-squared:  0.0203
## F-statistic: 2.036 on 1 and 49 DF,  p-value: 0.16
```

```
multiple.regression <- lm(LIFE ~ MALE + BIRTH + DIVO + BEDS + EDUC + INCO , data=Dat)
summary(multiple.regression)
```

```
##
## Call:
## lm(formula = LIFE ~ MALE + BIRTH + DIVO + BEDS + EDUC + INCO,
##     data = Dat)
##
## 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
```

```
anova(Male.regression, multiple.regression)
```

```
## 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
```

F statistics is 7.0963 with a p value of 6.099e-05

Conclude that atleast one of the coefficients are not equal to 0

Coefficients significantly improve the model given all others included

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

```
life.regression <- lm(LIFE ~ 1, data = Dat)
summary(life.regression)
```

```
##
## Call:
## lm(formula = LIFE ~ 1, data = Dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.078 -0.683 -0.098  1.097  2.812
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  70.7880     0.2118   334.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.513 on 50 degrees of freedom
```

```
g.regression <- lm(LIFE ~ MALE, BIRTH, data=Dat)
summary(g.regression)
```

```
##
## Call:
## lm(formula = LIFE ~ MALE, data = Dat, subset = BIRTH)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -2.17459 -0.92540 -0.07729  1.56028  1.95163
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 65.56012   12.88117    5.09 5.69e-06 ***
## MALE         0.05675    0.13519    0.42    0.676
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.344 on 49 degrees of freedom
## Multiple R-squared:  0.003583,   Adjusted R-squared:  -0.01675
## F-statistic: 0.1762 on 1 and 49 DF,  p-value: 0.6765
```

```
anova(life.regression, g.regression)
```

```
## Analysis of Variance Table
##
## Model 1: LIFE ~ 1
## Model 2: LIFE ~ MALE
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     50 114.40
## 2     49  88.53  1    25.867 14.317 0.0004212 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This is Partial F test:

H0: B1=B2=0 , k < p

HA: B1 != B2 != 0 (H0 is not true)

Reject Null Hypothesis

Coefficient Male and BIRTH does significantly improve model

Model fits the data better than the intercept-only model.

(h) Compute and report the terms in the decomposition

Terms in decomposition

SSreg(B1, B2, B3|B0) = 33.65

SSreg(B3|B0)= 3.31

SSreg(B2|B0,B3)=8.92

SSreg(B1|B0,B3,B2)=21.42

33.65 = 3.31 + 8.92 + 21.42

Left Side equal Right Side

Each term in the decomposition calculated below in order from left to right

```
model0 <- lm(LIFE ~ 1, data = Dat)
MBD <- lm(formula = LIFE ~ MALE + BIRTH + DIVO, data = Dat)
anova(model0,MBD)
```

```
## 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
```

```
# 33.646
```

ssreg(B3|B0)

```
model2 <- lm (LIFE ~ DIVO, data= Dat)
anova(model0, model2)
```

```
## 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
```

```
#3.31
```

SSreg(B2|B0,B3)

```
model3 <- lm (LIFE ~ BIRTH + DIVO, data=Dat)
anova(model2, model3)
```

```
## 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
```

```
#8.92
```

```
model4 <- lm (LIFE ~ MALE + BIRTH + DIVO, data=Dat)
anova(model3,model4)
```

```
## 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
```

```
#21.42
```

(i) 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'.

```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers

##       Best Subsets Regression
## ----------------------------------
## Model Index    Predictors
## ----------------------------------
##      1         BIRTH
##      2         MALE BIRTH
##      3         MALE BIRTH DIVO
##      4         MALE BIRTH DIVO INCO
## ----------------------------------
##
##                                                    Subsets Regression Summary
## ---------------------------------------------------------------------------------------------------------------
##                         Adj.        Pred
## Model    R-Square    R-Square    R-Square     C(p)        AIC        SBIC        SBC        MSEP
## ---------------------------------------------------------------------------------------------------------------
##   1        0.0966      0.0781     -0.0568    11.9054    186.7525    41.4479    192.5479    2.1953
##   2        0.2533      0.2222     -0.2238     3.6887    179.0377    34.5863    186.7650    1.8918
##   3        0.2941      0.2491     -0.2932     3.0253    178.1686    34.2782    187.8277    1.8660
##   4        0.2945      0.2332     -0.4036     5.0000    180.1406    36.4722    191.7315    1.9479
## ---------------------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria

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

Best model: Life = 62.3656 + 0.1689 * MALE - 0.3912 * BIRTH - 0.1272 * DIVO

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

```
ols_step_forward_p( lm( LIFE ~ MALE + BIRTH + DIVO + INCO, data=Dat))
```

```
## Forward Selection Method
## ---------------------------
##
## Candidate Terms:
##
## 1. MALE
## 2. BIRTH
## 3. DIVO
## 4. INCO
##
## We are selecting variables based on p value...
##
## Variables Entered:
##
## - BIRTH
## - MALE
## - DIVO
##
## No more variables to be added.
##
## Final Model Output
## ------------------
##
##                         Model Summary
## ------------------------------------------------------------------
## R                        0.542       RMSE                1.311
## R-Squared                0.294       Coef. Var           1.852
## Adj. R-Squared           0.249       MSE                 1.718
## Pred R-Squared          -0.293       MAE                 0.955
## ------------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                               ANOVA
## ----------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square      F        Sig.
## ----------------------------------------------------------------------
## Regression     33.646         3         11.215    6.528     9e-04
## Residual       80.751        47          1.718
## Total         114.397        50
## ----------------------------------------------------------------------
##
##                                 Parameter Estimates
## ----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
## ----------------------------------------------------------------------------------------
## (Intercept)    62.366         4.064                15.346    0.000    54.190    70.541
##       BIRTH    -0.391         0.109       -0.503   -3.594    0.001    -0.610    -0.172
```

```
##        MALE     0.169          0.048         0.507     3.531     0.001      0.073      0.265
##        DIVO    -0.127          0.077        -0.214    -1.649     0.106     -0.282      0.028
## ----------------------------------------------------------------------------------------
```

```
##
##                             Selection Summary
## ------------------------------------------------------------------------
##         Variable                    Adj.
## Step    Entered     R-Square    R-Square     C(p)        AIC         RMSE
## ------------------------------------------------------------------------
##    1    BIRTH          0.0966      0.0781    11.9054    186.7525    1.4523
##    2    MALE           0.2533      0.2222     3.6887    179.0377    1.3340
##    3    DIVO           0.2941      0.2491     3.0253    178.1686    1.3108
## ------------------------------------------------------------------------
```

```
lm( LIFE ~ MALE + BIRTH + DIVO)
```

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

Best model : Life = 62.3656 + 0.1689 * MALE - 0.3912 * BIRTH - 0.1272 * DIVO

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

```
ols_step_backward_p( lm( LIFE ~ MALE + BIRTH + DIVO + INCO, data=Dat))
```

```
## Backward Elimination Method
## ---------------------------
##
## Candidate Terms:
##
## 1 . MALE
## 2 . BIRTH
## 3 . DIVO
## 4 . INCO
##
## We are eliminating variables based on p value...
##
## Variables Removed:
##
## - INCO
##
## No more variables satisfy the condition of p value = 0.3
##
##
## Final Model Output
## ------------------
##
##                         Model Summary
## ------------------------------------------------------------
```

```
## R                       0.542       RMSE              1.311
## R-Squared               0.294       Coef. Var         1.852
## Adj. R-Squared          0.249       MSE               1.718
## Pred R-Squared         -0.293       MAE               0.955
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##
##                              ANOVA
## --------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square      F        Sig.
## --------------------------------------------------------------------
## Regression     33.646         3         11.215    6.528     9e-04
## Residual       80.751        47          1.718
## Total         114.397        50
## --------------------------------------------------------------------
##
##                                 Parameter Estimates
## ----------------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower     upper
## ----------------------------------------------------------------------------------------------
## (Intercept)    62.366         4.064                15.346    0.000    54.190    70.541
##        MALE     0.169         0.048        0.507    3.531    0.001     0.073     0.265
##       BIRTH    -0.391         0.109       -0.503   -3.594    0.001    -0.610    -0.172
##        DIVO    -0.127         0.077       -0.214   -1.649    0.106    -0.282     0.028
## ----------------------------------------------------------------------------------------------

##
##
##                           Elimination Summary
## ------------------------------------------------------------------------
##         Variable                 Adj.
## Step    Removed     R-Square    R-Square     C(p)       AIC        RMSE
## ------------------------------------------------------------------------
##    1    INCO          0.2941      0.2491    3.0253    178.1686    1.3108
## ------------------------------------------------------------------------
```

Best model : Life = 62.3656 + 0.1689 * MALE - 0.3912 * BIRTH - 0.1272 * DIVO