#Linear Model Group Project
#Myat R. Phyo and Marietta O. Ezeoke
#BIOL 483M

#Here we will run through a series of linear model tests using three different data sets.
#Here is an explanation of each test we will use:
#lm() << This allows us to calculate the linear model regression of our dependent variable and independent variables.
#summary() << This allows us to test the significance of the linear model results using a T-test.
#anova() << This allows us to test the significance of the linear model results using an F-Test otherwise known as the "anova."
#While not shown in these tests, we could have used data containing factors which we would have specified in the independent variables in this manner:
#X=factor()
#Since our data did not include factors this was not necessary. However, R would still provide results based on this new consideration.
#In each case, the decision rule is that if the p value is less than alpha we reject the null hypothesis.
#In the case of the F and T tests, the Null Hypothesis is that there is no relationship whereas the Alternate Hypothesis is that there is at least one relationship between an independent variable (X value) and the dependent variable (Y).
#The X variable is the independent variable and the Y variable is the dependent variable.
#Assumptions: Linear Regression depends on specifying in advance which variable is to be considered 'dependent' and which 'independent'. Also, the values in the Y set are random with one X value matched to one Y value.

#In the case of the Linear Model the null and alternate hypotheses are as follows:
#Null Hypothesis: The resulting slope is zero, implying no relationship between X and Y
#Alternative Hypothesis: A relationship exists between X and Y.
#extractAIC() << This allows an individual test of the AIC value. A lower AIC value implies a better fit.

#First Test: FIQ

```
> FIQ=read.table("C:/Users/Etta/Documents/FIQ.txt",header=TRUE)
> FIQ
   FSIQ Weight Height MRI_Count
1   133    118   64.5    816932
2   139    143   73.3   1038437
3   133    172   68.8    965353
4   137    147   65.0    951545
5    99    146   69.0    928799
6   138    138   64.5    991305
7    92    175   66.0    854258
8    89    134   66.3    904858
9   133    172   68.8    955466
10  132    118   64.5    833868
11  141    151   70.0   1079549
12  135    155   69.0    924059
13  140    155   70.5    856472
14   96    146   66.0    878897
15   83    135   68.0    865363
16  132    127   68.5    852244
17  100    178   73.5    945088
18  101    136   66.3    808020
19   80    180   70.0    889083
20   97    186   76.5    905940
21  135    122   62.0    790619
22  139    132   68.0    955003
23   91    114   63.0    831772
24  141    171   72.0    935494
25   85    140   68.0    798612
26  103    187   77.0   1062462
27   77    106   63.0    793549
28  130    159   66.5    866662
29  133    127   62.5    857782
30  144    191   67.0    949589
31  103    192   75.5    997925
32   90    181   69.0    879987
33   83    143   66.5    834344
34  133    153   66.5    948066
35  140    144   70.5    949395
36   88    139   64.5    893983
37   81    148   74.0    930016


38   89    179   75.5    935863
> attach(FIQ)
> Y=FSIQ
> x1f=Weight
> x2f=Height
> x3f=MRI_Count
> LMF=lm(Y~x1f+x2f+x3f)
> LMF

Call:
lm(formula = Y ~ x1f + x2f + x3f)

Coefficients:
(Intercept)          x1f          x2f          x3f
  1.174e+02   -6.436e-02   -2.641e+00    2.057e-04

> summary(LMF)

Call:
lm(formula = Y ~ x1f + x2f + x3f)

Residuals:
    Min      1Q  Median      3Q     Max
-34.056 -17.818  -1.373  18.048  42.537

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.174e+02  6.776e+01   1.733  0.09219 .
x1f         -6.436e-02  2.121e-01  -0.304  0.76334
x2f         -2.641e+00  1.323e+00  -1.996  0.05397 .
x3f          2.057e-04  6.063e-05   3.393  0.00177 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.3 on 34 degrees of freedom
Multiple R-squared: 0.2649,     Adjusted R-squared: 0.2001
F-statistic: 4.085 on 3 and 34 DF,  p-value: 0.01402

> anova(LMF)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value   Pr(>F)
x1f        1    55.6    55.6  0.1226 0.728397
x2f        1   279.3   279.3  0.6156 0.438127
x3f        1  5224.6  5224.6 11.5156 0.001768 **
Residuals 34 15425.8   453.7
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#step(,direction="backward") << This allows us to get an optimal model which is the best fit we can use in our tests. Once this is done, we can compare the given model to the Full Model to determine which the better model is.**

```
> step(LMF,direction="backward")
Start:  AIC=236.24
Y ~ x1f + x2f + x3f

       Df Sum of Sq   RSS    AIC
- x1f   1      41.8 15468 234.34
<none>              15426 236.24
- x2f   1    1808.0 17234 238.45
- x3f   1    5224.6 20651 245.32

Step:  AIC=234.34
Y ~ x2f + x3f

       Df Sum of Sq   RSS    AIC
<none>              15468 234.34
- x2f   1    3180.7 18648 239.44
- x3f   1    5223.3 20691 243.40

Call:
lm(formula = Y ~ x2f + x3f)

Coefficients:
(Intercept)          x2f          x3f  
  1.264e+02   -2.871e+00    2.025e-04  

> extractAIC(LMF)
[1]   4.0000 236.2361
```

#Once we find a “best fit model” we can compare the full and reduced model to check if the reduced or the full model is the better fit model.

#The Null Hypothesis in this case: The Reduced Model is the Preferred Model.

#The Alternative Hypothesis in this case: the Full Model is the Preferred Model.

#The decision rule: If p is less than alpha then we reject the null hypothesis.

```
> FMF=lm(Y~x1f+x2f+x3f)
> RMF=lm(Y~x2f+x3f)
> anova(RMF,FMF)
Analysis of Variance Table
```

```
Model 1: Y ~ x2f + x3f
Model 2: Y ~ x1f + x2f + x3f
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     35 15468
2     34 15426  1    41.798 0.0921 0.7633
```

#Based on our results, we reject the null hypothesis and decide that the Full Model is the preferred model.

#Below are two more examples of the linear regression tests:

#Second Test

```
> VIQD=read.table("C:/Users/Etta/Documents/VIQ.txt",header=TRUE)
> VIQD
   VIQ Weight Height MRI_Count
1  132    118   64.5    816932
2  123    143   73.3   1038437
3  129    172   68.8    965353
4  132    147   65.0    951545
5   90    146   69.0    928799
6  136    138   64.5    991305
7   90    175   66.0    854258
8   93    134   66.3    904858
9  114    172   68.8    955466
10 129    118   64.5    833868
11 150    151   70.0   1079549
12 129    155   69.0    924059
13 120    155   70.5    856472
14 100    146   66.0    878897
15  71    135   68.0    865363
16 132    127   68.5    852244
17  96    178   73.5    945088
18 112    136   66.3    808020
19  77    180   70.0    889083
20 107    186   76.5    905940
21 129    122   62.0    790619
22 145    132   68.0    955003
23  86    114   63.0    831772
24 145    171   72.0    935494
25  90    140   68.0    798612
26  96    187   77.0   1062462
27  83    106   63.0    793549
28 126    159   66.5    866662
29 126    127   62.5    857782
30 145    191   67.0    949589
```

```
31  96    192   75.5    997925
32  96    181   69.0    879987
33  90    143   66.5    834344
34 129    153   66.5    948066
35 150    144   70.5    949395
36  86    139   64.5    893983
37  90    148   74.0    930016
38  91    179   75.5    935863
> attach(VIQD)
> Y=VIQ
> x1v=Weight
> x2v=Height
> x3v=MRI_Count
> LMV=lm(Y~x1v+x2v+x3v)
> LMV

Call:
lm(formula = Y ~ x1v + x2v + x3v)

Coefficients:
(Intercept)          x1v          x2v          x3v
  1.136e+02   -9.968e-02   -2.241e+00    1.841e-04

> summary(LMV)

Call:
lm(formula = Y ~ x1v + x2v + x3v)

Residuals:
   Min     1Q Median     3Q    Max
-36.06 -14.14  -2.51  17.45  37.59

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.136e+02  6.684e+01   1.700  0.09829 .
x1v         -9.968e-02  2.092e-01  -0.477  0.63675
x2v         -2.241e+00  1.305e+00  -1.717  0.09506 .
x3v          1.841e-04  5.981e-05   3.077  0.00411 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.01 on 34 degrees of freedom
Multiple R-squared: 0.229,      Adjusted R-squared: 0.161
F-statistic: 3.366 on 3 and 34 DF,  p-value: 0.02971

> anova(LMV)
```

```
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value   Pr(>F)
x1v        1   112.7   112.7  0.2553 0.616622
x2v        1   164.8   164.8  0.3733 0.545248
x3v        1  4181.4  4181.4  9.4706 0.004109 **
Residuals 34 15011.4   441.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> step(LMV,direction="backward")
Start:  AIC=235.2
Y ~ x1v + x2v + x3v

       Df Sum of Sq   RSS    AIC
- x1v   1     100.3 15112 233.45
<none>              15011 235.20
- x2v   1    1301.8 16313 236.36
- x3v   1    4181.4 19193 242.54

Step:  AIC=233.45
Y ~ x2v + x3v

       Df Sum of Sq   RSS    AIC
<none>              15112 233.45
- x2v   1    2603.1 17715 237.49
- x3v   1    4083.1 19195 240.54

Call:
lm(formula = Y ~ x2v + x3v)

Coefficients:
(Intercept)         x2v         x3v
  1.275e+02  -2.597e+00   1.790e-04

> extractAIC(LMV)
[1]   4.0000 235.2012
>
> FMV=lm(Y~x1v+x2v+x3v)
> RMV=lm(Y~x2v+x3v)
> anova(RMV,FMV)
Analysis of Variance Table

Model 1: Y ~ x2v + x3v
Model 2: Y ~ x1v + x2v + x3v
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
```

```
1     35 15112
2     34 15011  1    100.26 0.2271 0.6368
>
```

#Third Test

```
> PIQD=read.table("C:/Users/Etta/Documents/PIQ.txt",header=TRUE)
> PIQD
   PIQ Weight Height MRI_Count
1  124    118   64.5    816932
2  150    143   73.3   1038437
3  128    172   68.8    965353
4  134    147   65.0    951545
5  110    146   69.0    928799
6  131    138   64.5    991305
7   98    175   66.0    854258
8   84    134   66.3    904858
9  147    172   68.8    955466
10 124    118   64.5    833868
11 128    151   70.0   1079549
12 124    155   69.0    924059
13 147    155   70.5    856472
14  90    146   66.0    878897
15  96    135   68.0    865363
16 120    127   68.5    852244
17 102    178   73.5    945088
18  84    136   66.3    808020
19  86    180   70.0    889083
20  84    186   76.5    905940
21 134    122   62.0    790619
22 128    132   68.0    955003
23 102    114   63.0    831772
24 131    171   72.0    935494
25  84    140   68.0    798612
26 110    187   77.0   1062462
27  72    106   63.0    793549
28 124    159   66.5    866662
29 132    127   62.5    857782
30 137    191   67.0    949589
31 110    192   75.5    997925
32  86    181   69.0    879987
33  81    143   66.5    834344
34 128    153   66.5    948066
35 124    144   70.5    949395
36  94    139   64.5    893983
37  74    148   74.0    930016
```

```
38  89    179   75.5    935863
> attach(PIQD)
The following object(s) are masked from 'VIQD':

    Height, MRI_Count, Weight
> Y=PIQ
> x1p=Weight
> x2p=Height
> x3p=MRI_Count
> LMP=lm(Y~x1p+x2p+x3p)
> LMP

Call:
lm(formula = Y ~ x1p + x2p + x3p)

Coefficients:
(Intercept)          x1p          x2p          x3p
  1.114e+02    7.164e-04   -2.732e+00    2.060e-04

> summary(LMP)

Call:
lm(formula = Y ~ x1p + x2p + x3p)

Residuals:
   Min     1Q Median     3Q    Max
-32.73 -12.09  -3.84  14.17  51.70

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.114e+02  6.297e+01   1.769 0.085914 .
x1p          7.164e-04  1.971e-01   0.004 0.997121
x2p         -2.732e+00  1.230e+00  -2.222 0.033018 *
x3p          2.060e-04  5.635e-05   3.656 0.000856 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.79 on 34 degrees of freedom
Multiple R-squared: 0.2949,     Adjusted R-squared: 0.2327
F-statistic:  4.74 on 3 and 34 DF,  p-value: 0.007221

> anova(LMP)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value    Pr(>F)
```

```
x1p        1     0.1     0.1  0.0003 0.9861839
x2p        1   333.4   333.4  0.8508 0.3628125
x3p        1  5238.5  5238.5 13.3690 0.0008565 ***
Residuals 34 13322.5   391.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> step(LMP,direction="backward")
Start:  AIC=230.67
Y ~ x1p + x2p + x3p

       Df Sum of Sq   RSS    AIC
- x1p   1       0.0 13322 228.67
<none>              13322 230.67
- x2p   1    1935.2 15258 233.82
- x3p   1    5238.5 18561 241.27

Step:  AIC=228.67
Y ~ x2p + x3p

       Df Sum of Sq   RSS    AIC
<none>              13322 228.67
- x2p   1    2875.4 16198 234.09
- x3p   1    5408.1 18731 239.61

Call:
lm(formula = Y ~ x2p + x3p)

Coefficients:
(Intercept)          x2p          x3p
  1.113e+02   -2.730e+00    2.061e-04

> extractAIC(LMP)
[1]   4.0000 230.6658
>
> FMP=lm(Y~x1p+x2p+x3p)
> RMP=lm(Y~x2p+x3p)
> anova(RMP,FMP)
Analysis of Variance Table

Model 1: Y ~ x2p + x3p
Model 2: Y ~ x1p + x2p + x3p
  Res.Df   RSS Df Sum of Sq  F Pr(>F)
1     35 13322
2     34 13322  1 0.0051787  0 0.9971
>
```