Math 3080 § 1. **Potato Example:** Name: Example
Treibergs **Basic ANOVA Table in Multiple Regression** March 16, 2014

This **R**© program explores the analysis of variance table provided by multiple regression programs. The usual ANOVA output from **R**© does not provide the basic ANOVA table given by the other programs, although it can be read from the table provided. We discuss the *coefficient of multiple determination,* R^2, and the *model utility test.*

This data was taken from Levine Ramsey & Smidt, *Applied Statistics for Engineers and Scientists,* Prentice Hall, Upper Saddle River, NJ, 2001 about a study made by the Mountain States Potato Company of Eastern Idaho. A byproduct of their production is filter cake which is used as cattle feed. Farmers were complaining that recent batches of filter cake had too much moisture. The study was to predict how production variables affect the percentage of solids. The available variables are Y, percent solids content in filter cake $X1$ acidity (in pH) of the clarifier, $X2$ lower pressure (pressure in the vacuum line below the fluid line) $X3$ upper pressure, $X4$ cake thickness, $X5$ varidrive speed and $X6$ drum speed setting. For this project we regress Y on $X1$ and $X2$ only.

For multiple regression, the *sum of squares identity* holds

$$SST = SSR + SSE$$

The quantities are

$$SST = S_{yy} = \sum_{j=1}^{n} (y_j - \overline{y_j})^2 = \sum_{j=1}^{n} y_j^2 - \frac{1}{n} \left(\sum_{j=1}^{n} y_j \right)^2$$

which is the total sum square, the sum squared deviations of of the observed values y_j from the mean $\overline{y}$. It is split into the *error sum of squares*, the part of the sum square deviation from the predicted values

$$SSE = S_{yy} = \sum_{j=1}^{n} (y_j - \widehat{y_j})^2$$

which is the sum squares of the residuals, the deviation of the observed values from the fitted values. In the case of two independent variables $X1$ and $X2$ here, we find the coefficients of the least squares fit $\widehat{\beta_i}$ that best approximates the observed values

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{1,i} + \widehat{\beta_i} x_{2,i}.$$

The last quantitiy is the *regression sum of squares*,

$$SSR = SST - SSE = \sum_{j=1}^{n} (\widehat{y_j} - \overline{y_j})^2$$

that part of the total deviation from $\overline{y}$ which is given by the model.

The *cofficient of multiple determination,* R^2 is the proportion of the sum squared deviation accounted for by the linear model

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

To compensate for increasing the number of variables (and thus automatically improving the fit) we penalize for using a larger number k of variables. (For two predictors, $k = 2$.) The adjusted coefficient of multiple determination is

$$R_a^2 = 1 - \frac{(n-1)SSE}{(n-k-1)SST}.$$

It will be used to compare models with different numbers of variables.

The *model utility test* tests whether there is a useful relationship between y and *any* of the variables $x_1, \ldots, x_k$. In the simple regression case, the f-test was equivalent to testing whether $\beta_1 = 0$. In the multiple regression case, the null and alternative hypotheses are

$$\mathcal{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$\mathcal{H}_a : \text{at least one } \beta_j \neq 0.$$

The test statistic is

$$f = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)}$$

which is distributed according to the f-distribution with $\nu_1 = k$ numerator and $\nu_2 = n - k - 1$ denominator degrees of freedom. The null hypothesis $\mathcal{H}_0$ is rejected in favor of $\mathcal{H}_a$ at level α if $f \geq F_{\alpha,k,n-k-1}$.

The anova table generated by **R**© does not show a "Regression" row as does SAS or MINITAB. However, the last row of the summary gives "f-statistic, ν_1, ν_2 DF and p-value" of the model utility test. The "regression" row DF, SSR may be found by adding the SS and DF of the factors listed in the ANOVA table. Then $MSR = SSR/k$. Indeed, the sum squares attributed to the factors in the ANOVA table are the additional Sum Squares that come by adding the factors one-by-one, top to bottom. By rearranging the order of the factors top-to-bottom results in different sum of squares for the individual factors, although the coefficients remain the same and SSR is the same.

The basic ANOVA table is recovered by hand in the demonstration. It may also be recovered by comparing the model with the model with no independent variables other than the constant.

Source	df	SS	MS	F	P
Regression	k	SSR	MSR	f	p
Error	$n-k-1$	SSE	MSE		
Total	$n-1$	SST			

Data Set Used in this Analysis :

```
# Math 3080-1              Potato Data              March 15, 2014
# Treibergs
#
# From Levine Ramsey & Smidt, Applied Statistics for Engneers and
# Scientists, Prentice Hall, Upper saddle River, NJ, 2001.
#
# Results of Mountain States Potato Company of Eastern Idaho.
# A byproduct is filter cake, used as cattle feed. The study was to
# predict percentage of solids.
#
# Variables
#   Y   Percent solids content in filter cake
#   X1  acidity (in pH) of the clarifier
#   X2  lower pressure (pressure in the vacuum line below the fluid line)
```

```
#   X3  upper pressure
#   X4  cake thickness
#   X5  varidrive speed
#   X6  drum speed setting
"Y" "X1" "X2" "X3" "X4" "X5" "X6"
        9.7      3.7       13       14    0.250        6    33.00
        9.4      3.8       17       18    0.875        6    30.43
       10.5      3.8       14       15    0.500        6    34.00
       10.9      3.9       14       14    0.500        6    34.00
       11.6      4.3       17       18    0.375        6    36.24
       10.9      4.2       16       17    0.500        6    31.76
       11.0      4.3       16       19    0.375        6    34.00
       10.7      3.9       15       16    0.375        6    32.13
       11.8      3.6        8        8    0.375        6    37.00
        9.7      4.0       18       18    0.500        6    36.00
       11.6      4.0       12       13    0.313        5    45.00
       10.9      3.9       15       15    0.500        5    50.00
       10.0      3.8       17       18    0.625        5    46.91
       10.3      3.8       13       14    0.500        4    57.50
       10.1      3.6       17       17    0.625        4    60.40
        9.9      3.8       17       18    0.500        4    53.14
        9.5      3.5       17       18    0.625        6    34.40
       10.5      3.8       15       17    0.500        6    33.96
       10.8      3.9       15       17    0.750        6    35.00
       10.4      3.9       14       15    0.500        6    35.00
       10.9      4.0       15       16    0.500        6    34.00
       11.2      4.4       17       19    0.375        6    34.00
        9.5      3.8       17       17    0.500        6    33.49
       10.7      3.9       15       17    0.500        6    33.38
       10.1      3.8       15       17    0.500        6    41.00
       10.5      3.8       17       17    0.500        6    36.00
       10.9      4.0       15       17    0.250        6    34.00
       15.5      4.3       13       15    0.625        6    41.00
       13.1      4.0       17       17    0.500        6    35.00
       11.0      4.0       14       15    0.375        6    36.00
       12.5      4.2       15       17    0.313        6    37.72
       11.7      4.2       14       14    0.250        6    36.00
       11.9      4.4       15       16    0.375        6    36.52
       11.7      3.4        8       10    0.313        6    38.08
       17.8      4.3       12       12    0.313        6    38.00
       11.8      4.5       14       15    0.250        6    33.00
       10.0      3.7       12       13    0.250        5    48.00
       10.3      3.7       15       15    0.500        5    48.00
        9.8      3.8       14       15    0.500        5    47.24
       10.0      3.7       13       14    0.500        6    37.00
       10.6      4.1       14       15    0.500        6    33.70
       11.2      3.9       13       14    0.375        6    38.26
       10.9      3.7       13       14    0.313        6    38.00
       11.0      4.1       13       14    0.375        6    37.00
       11.0      4.1       14       15    0.375        6    38.00
       11.7      4.5       14       14    0.250        6    36.26
       11.8      4.4       13       14    0.250        6    37.45
```

12.0	4.2	13	13	0.375	6	38.00
11.8	4.6	14	14	0.375	6	36.90
11.1	4.0	14	15	0.500	6	37.00
11.6	3.9	14	14	0.500	6	37.50
11.0	4.0	14	15	0.500	6	36.00
11.2	3.9	15	15	0.313	6	35.00
11.0	4.2	14	14	0.375	6	37.00

R Session:

```
R version 2.13.1 (2011-07-08)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-apple-darwin9.8.0/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.41 (5874) i386-apple-darwin9.8.0]

[History restored from /Users/andrejstreibergs/.Rapp.history]

> tt=read.table(M3082DataPotato.txt)
Error in read.table(M3082DataPotato.txt) :
  object 'M3082DataPotato.txt' not found
> tt=read.table("M3082DataPotato.txt",header=T)
> attach(tt)
> names(tt)
[1] "Y"  "X1" "X2" "X3" "X4" "X5" "X6"
> cbind(Y,X1,X2)
         Y  X1 X2
 [1,]  9.7 3.7 13
 [2,]  9.4 3.8 17
 [3,] 10.5 3.8 14
 [4,] 10.9 3.9 14
 [5,] 11.6 4.3 17
 [6,] 10.9 4.2 16
 [7,] 11.0 4.3 16
 [8,] 10.7 3.9 15
 [9,] 11.8 3.6  8
```

```
[10,]  9.7 4.0 18
[11,] 11.6 4.0 12
[12,] 10.9 3.9 15
[13,] 10.0 3.8 17
[14,] 10.3 3.8 13
[15,] 10.1 3.6 17
[16,]  9.9 3.8 17
[17,]  9.5 3.5 17
[18,] 10.5 3.8 15
[19,] 10.8 3.9 15
[20,] 10.4 3.9 14
[21,] 10.9 4.0 15
[22,] 11.2 4.4 17
[23,]  9.5 3.8 17
[24,] 10.7 3.9 15
[25,] 10.1 3.8 15
[26,] 10.5 3.8 17
[27,] 10.9 4.0 15
[28,] 15.5 4.3 13
[29,] 13.1 4.0 17
[30,] 11.0 4.0 14
[31,] 12.5 4.2 15
[32,] 11.7 4.2 14
[33,] 11.9 4.4 15
[34,] 11.7 3.4  8
[35,] 17.8 4.3 12
[36,] 11.8 4.5 14
[37,] 10.0 3.7 12
[38,] 10.3 3.7 15
[39,]  9.8 3.8 14
[40,] 10.0 3.7 13
[41,] 10.6 4.1 14
[42,] 11.2 3.9 13
[43,] 10.9 3.7 13
[44,] 11.0 4.1 13
[45,] 11.0 4.1 14
[46,] 11.7 4.5 14
[47,] 11.8 4.4 13
[48,] 12.0 4.2 13
[49,] 11.8 4.6 14
[50,] 11.1 4.0 14
[51,] 11.6 3.9 14
[52,] 11.0 4.0 14
[53,] 11.2 3.9 15
[54,] 11.0 4.2 14

> ##################   SCATTERLOT OF ALL VARIABLES  ########
> pairs(Y~X1+X2)
```

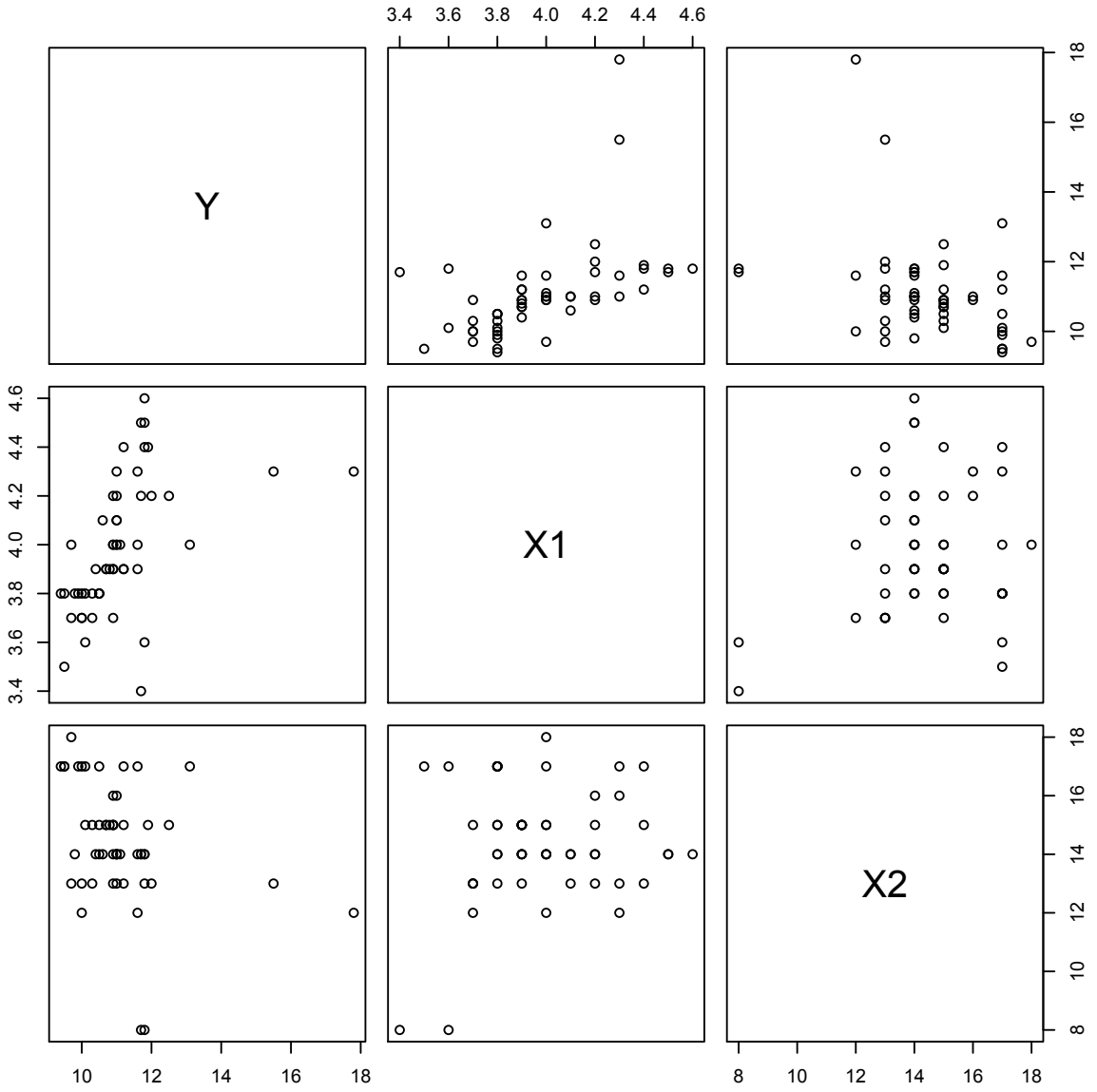
```
> ##################   RUN REGRESSION Y ~ X1 + X2   ################
> k = 2
> f1=lm(Y~X1+X2)
> summary(f1);anova(f1)

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

Residuals:
    Min      1Q  Median      3Q     Max
-1.1709 -0.5550 -0.1491  0.1630  5.1213

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.81625    2.33934   1.631 0.108981
X1           2.84370    0.55445   5.129 4.56e-06 ***
X2          -0.28045    0.07454  -3.762 0.000436 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.07 on 51 degrees of freedom
Multiple R-squared: 0.4158,Adjusted R-squared: 0.3929
F-statistic: 18.15 on 2 and 51 DF,  p-value: 1.115e-06

Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 25.359 25.3591  22.143 1.972e-05 ***
X2         1 16.211 16.2106  14.155 0.0004357 ***
Residuals 51 58.407  1.1452
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
>
> #####   ANOVA TABLE BY HAND   ######################################
> #       SET A MATRIX  an  TO BE THE BASIC ANOVA TABLE
>
> an=matrix(rep(0,times=15),ncol=5)
> an[1,1]=k
> rownames(an)=c("Regression","Error","Total")
> colnames(an)=c("DF","SS", "MS", "F", "P")

> n=length(Y)
> ybar=mean(Y); ybar
[1] 11.09259
> Syy=sum(Y^2)-sum(Y)^2/n; Syy
[1] 99.97704
> ####################  E.G. GET SSE BY SUMMING RESIDUALS  ##############
> SSE=sum(residuals(f1)^2)
> SST=Syy
> SSR=SST-SSE
```

```
> an[1,1]=k
> an[2,1]=n-k-1
> an[3,1]=n-1
> an[1,2]=SSR
> an[2,2]=SSE
> an[3,2]=SST
> an[1,3]=an[1,2]/an[1,1]
> an[2,3]=an[2,2]/an[2,1]
> an[1,4]=an[1,3]/an[2,3]
> an[1,5]=pf(an[1,4],an[1,1],an[2,1],lower.tail=F)
> an
           DF       SS        MS        F            P
Regression  2 41.56971 20.784854 18.14888 1.115478e-06
Error      51 58.40733  1.145242  0.00000 0.000000e+00
Total      53 99.97704  0.000000  0.00000 0.000000e+00
>
>
> #################  SSR IS SUM OF SS  IN  R'S ANOVA TABLE  ######
> SSR
[1] 41.56971
> 25.359+16.211
[1] 41.57

>
> ################ R2  AND  ADJUSTED-R2  ##########################
>
> R2=SSR/SST;R2
[1] 0.4157925
> R2adj=1-((n-1)*SSE)/((n-k-1)*SST);R2adj
[1] 0.3928825



> ###  SSR IS PART OF DEVIATIONS DUE TO MODEL NOT DUE TO MEAN ONLY  ####
> #    CALL MINIMAL MODEL  Y~1   DUE TO MEAN ONLY
> f0=lm(Y~1)
> summary(f0);anova(f0)
Call:
lm(formula = Y ~ 1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6926 -0.7676 -0.1926  0.5074  6.7074

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.0926     0.1869   59.35   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.373 on 53 degrees of freedom
```

```
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals 53 99.977  1.8864
>
> ###  MODEL UTILITY TEST = TEST OF FULL MODEL VS MINIMAL MODEL  ###
> anova(f0,f1)
Analysis of Variance Table

Model 1: Y ~ 1
Model 2: Y ~ X1 + X2
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     53 99.977
2     51 58.407  2     41.57 18.149 1.115e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1


> #################  ANOVA TABLE OF FULL MODEL  (REPEAT)  ################
> anova(f1)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 25.359 25.3591  22.143 1.972e-05 ***
X2         1 16.211 16.2106  14.155 0.0004357 ***
Residuals 51 58.407  1.1452
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1


> ###############  FIT INTERMEDIATE MODEL  Y~X1   TO SEE FIRST SS  ######
> f3=lm(Y~X1)
> summary(f3); anova(f3)

Call:
lm(formula = Y ~ X1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4405 -0.6322 -0.2405  0.0184  5.8826

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.7821     2.4581   0.318 0.751631
X1            2.5896     0.6160   4.204 0.000104 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.198 on 52 degrees of freedom
Multiple R-squared: 0.2536,Adjusted R-squared: 0.2393
```

```
F-statistic: 17.67 on 1 and 52 DF,  p-value: 0.0001035

Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 25.359  25.359  17.672 0.0001035 ***
Residuals 52 74.618   1.435
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1


> ###############  SAME SST AS FULL MODEL  ##########################
> 25.359+74.618
[1] 99.977
> SST
[1] 99.97704
>
> ####  COMPARE ANOVA TABLE WITH VARIABLES IN OPPOSITE ORDER #######
> f4=lm(Y~X2+X1)
> SUMMARY(f4); ANOVA(f4)
Error: could not find function "SUMMARY"
> summary(f4); anova(f4)

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

Residuals:
    Min      1Q  Median      3Q     Max
-1.1709 -0.5550 -0.1491  0.1630  5.1213

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.81625    2.33934   1.631 0.108981
X2          -0.28045    0.07454  -3.762 0.000436 ***
X1           2.84370    0.55445   5.129 4.56e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.07 on 51 degrees of freedom
Multiple R-squared: 0.4158,Adjusted R-squared: 0.3929
F-statistic: 18.15 on 2 and 51 DF,  p-value: 1.115e-06

Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X2         1 11.444 11.4439  9.9926  0.002645 **
X1         1 30.126 30.1258 26.3052 4.562e-06 ***
Residuals 51 58.407  1.1452
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

>


Y
X1
X2
3.4
3.6
3.8
4.0
4.2
4.4
4.6
8
10
12
14
16
18