Math 158, Spring 2013
Jo Hardin
Multiple Regression I – R code

1. Data should be either in comma delimited (`sep=","`) or tab delimited (`sep="\t"`) format. Data should live in the **same** directory as the R program (`.RData`).

```
> births <- read.table("NCBIRTH800.csv", header=T, sep=",")
> dim(births)
> names(births)
```

Notice that the variable names we'll use are `mage`, `tounces`, `gained`, and `smoke`.

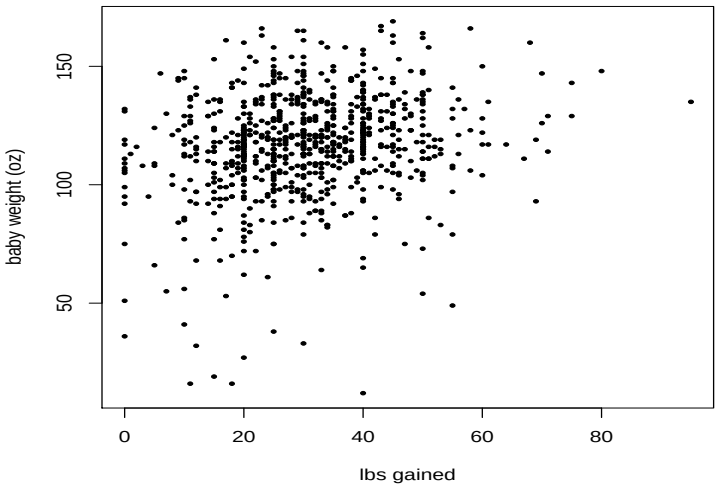
```
> attach(births)
```

When we `attach` we create new variables from the data set. That is, we can now use each of the columns as their own variables / vectors.

2. It is *always* a good idea to graph your data and look at numerical summaries. Sometimes you'll find out important artifacts or mistakes.

```
> summary( cbind( mage, tounces, smoke, gained) )
> apply(cbind( mage, tounces, smoke, gained), 2, sd, na.rm=T)
> plot(gained, tounces, xlab="lbs gained", ylab="baby weight (oz)", pch=19)
```

```
     mage          tounces          smoke           gained
Min.   :15.00   Min.   : 12.0   Min.   :0.0000   Min.   : 0.00
1st Qu.:22.00   1st Qu.:106.0   1st Qu.:0.0000   1st Qu.:20.00
Median :26.00   Median :118.0   Median :0.0000   Median :30.00
Mean   :26.91   Mean   :116.4   Mean   :0.1429   Mean   :30.58
3rd Qu.:32.00   3rd Qu.:130.0   3rd Qu.:0.0000   3rd Qu.:40.00
Max.   :42.00   Max.   :169.0   Max.   :1.0000   Max.   :95.00
SD     : 6.11   SD     : 22.5   SD     :0.3502   SD     :13.65
                                NA's   :2.0000   NA's   :23.00
```



3. We're interested in predicting a baby's size from the mother's weight gain.

```
> summary( lm(tounces ~ gained) )

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 105.7025     1.9053  55.479  < 2e-16 ***
gained        0.3612     0.0569   6.348 3.72e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 21.63 on 775 degrees of freedom
  (23 observations deleted due to missingness)
Multiple R-squared: 0.04942,    Adjusted R-squared: 0.0482
F-statistic: 40.29 on 1 and 775 DF,  p-value: 3.718e-10
```

4. What if we include smoking status as a variable?

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 106.94807    1.92006  55.700  < 2e-16 ***
gained        0.35866    0.05644   6.354 3.57e-10 ***
smoke        -8.04006    2.18391  -3.681 0.000248 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 21.46 on 774 degrees of freedom
  (23 observations deleted due to missingness)
Multiple R-squared: 0.06578,    Adjusted R-squared: 0.06337
F-statistic: 27.25 on 2 and 774 DF,  p-value: 3.662e-12
```

- Note that the F p-value is no longer equal to the t-stat p-value(s). Now the degrees of freedom are (2, 774) because we're estimating 2 parameters.
- Write out the estimated regression model separately for smokers and non-smokers, and sketch the lines onto the scatterplot.
- How do you interpret your new coefficients (b_0, b_1, b_2)?
- How did the coefficient on `gained` change?
- How does R^2 change? MSE change?

5. What if we let smoking and weight gain *interact*?

```
> summary( lm( tounces ~ gained * smoke) )
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  105.96278    2.10092  50.436  < 2e-16 ***
gained         0.39081    0.06293   6.210 8.64e-10 ***
smoke         -3.07139    4.82713  -0.636    0.525
gained:smoke  -0.16411    0.14219  -1.154    0.249
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 21.45 on 773 degrees of freedom
  (23 observations deleted due to missingness)
Multiple R-squared: 0.06739,    Adjusted R-squared: 0.06377
F-statistic: 18.62 on 3 and 773 DF,  p-value: 1.142e-11
```

   - Note again that the F p-value is no longer equal to the t-stat p-value(s). Now the degrees of freedom are (3, 773) because we are estimating 3 parameters.
   - Write out the estimated regression model separately for smokers and non-smokers, and sketch the lines onto the scatterplot.
   - How do you interpret your new coefficients (b_0, b_1, b_2, b_3)?
   - What happened to the significance?
   - How did the coefficient on `gained` change?
   - How does R^2 change? MSE change?

6. What happens to the model if we add in another quantitative variable?

```
> summary( lm( tounces ~ gained + smoke + mage) )
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 91.33400    3.86669  23.621  < 2e-16 ***
gained       0.35175    0.05573   6.311 4.66e-10 ***
smoke       -7.20603    2.16311  -3.331 0.000905 ***
mage         0.58263    0.12577   4.633 4.23e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 21.18 on 773 degrees of freedom
  (23 observations deleted due to missingness)
Multiple R-squared: 0.09102,    Adjusted R-squared: 0.08749
F-statistic:  25.8 on 3 and 773 DF,  p-value: 6.502e-16
```

   Note the t-stat p-values, parameter estimates, R^2 value, MSE, F-statistic, degrees of freedom, and F-stat p-values.