Regression

Load Data

This week we will be experimenting with data about cereal. Find the file "Cereal.csv" on the UBlearns site for the lab. Download and save it somewhere. Use the Import Dataset button to import the dataset. Make sure that the Headings button is "yes" when you're in the import screen. Then remember to attach it.

```
attach(Cereal)
```

Let's look at the names of the variables first.

```
names(Cereal)
```

```
##  [1] "name"     "mfr"      "type"     "calories" "protein"  "fat"
##  [7] "sodium"   "fiber"    "carbo"    "sugars"   "potass"   "vitamins"
## [13] "shelf"    "weight"   "cups"     "rating"
```

Looks like we have some nutritional facts and a rating. I can get a sense of the data by calling the summary() command on some variables.

```
summary(rating)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.04   33.17   40.40   42.67   50.83   93.70
```

```
summary(sugars)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -1.000   3.000   7.000   6.922  11.000  15.000
```

```
summary(potass)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -1.00   40.00   90.00   96.08  120.00  330.00
```

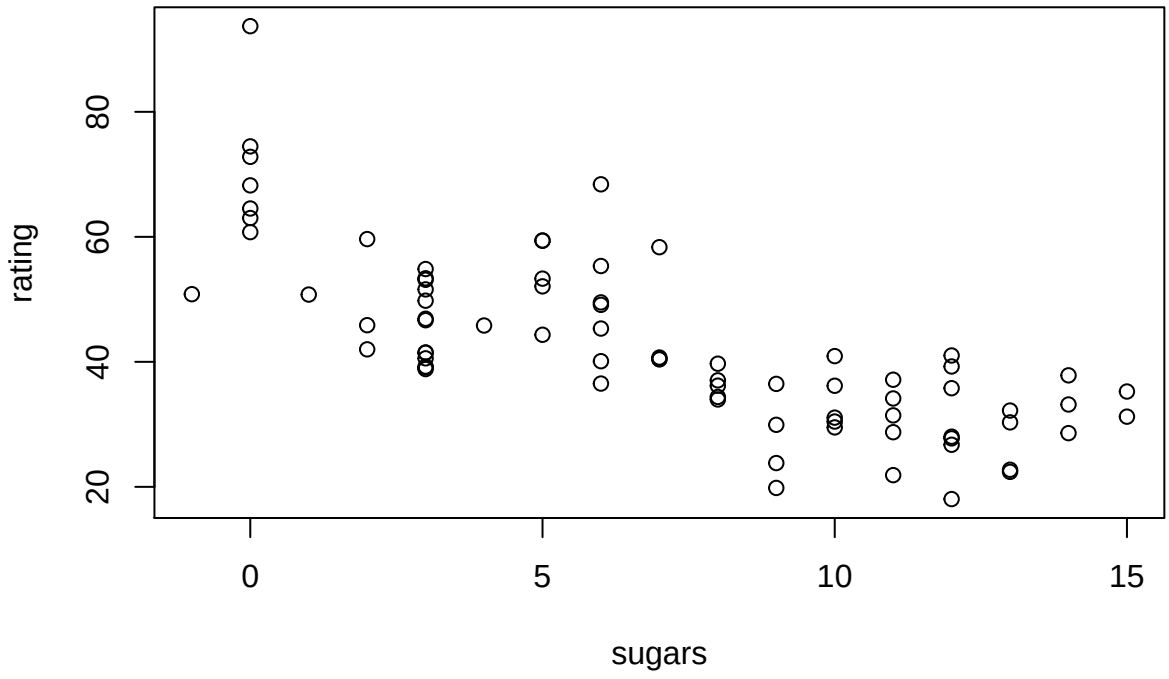
By looking at the summary for rating I can see that the rating scale seems to be 0 to 100. That is useful information in case I was thinking the rating scale was from 1 to 5 or 0 to 10. Sometimes it is a good idea to use the summary or table commands to understand the data first before moving to more complicated analyses.

Regression

The cereal rating will serve as a good outcome variable. Thus, rating will be the dependent variable. According to this line of thought, let's try to understand how the rating is determined by some of the nutritional qualities of cereal.

I choose sugars as the independent variable. I would expect more amounts of sugar to lead to higher ratings because sugar is pretty sweet. I can plot the two variables first

```
plot(sugars, rating)
```



Interestingly enough it seems like more sugar actually leads to a lower rating. The data points are gradually getting lower as the amount of sugar increases. This is some visual evidence that I'm wrong in thinking that sugar would increase the rating.

Now I'll do a regression to understand the relationship. This is called a bivariate regression because I will only be using two variables. Sugar is the independent (or X) variable and rating is the dependent (or Y) variable. I use the following R code to run the regression and assign it to an object called "model1".

```
model1 <- lm(rating~sugars)
```

Notice that the dependent variable is listed first –before the tilda (~). The lm command is short for "linear model" which is a reminder that we are making a straight line through the data points. Since this is Ordinary Least Square regression (OLS), R will attempt to draw a line with the fewest squared deviations between the line and the data points.

To see the all important results from the regression, we need to call the results. The easiest way is by using the summary command.

```
summary(model1)
```

```
##
## Call:
## lm(formula = rating ~ sugars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.853  -5.677  -1.439   5.160  34.421
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  59.2844     1.9485   30.43  < 2e-16 ***
## sugars       -2.4008     0.2373  -10.12 1.15e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.196 on 75 degrees of freedom
## Multiple R-squared:  0.5771, Adjusted R-squared:  0.5715
## F-statistic: 102.3 on 1 and 75 DF,  p-value: 1.153e-15
```

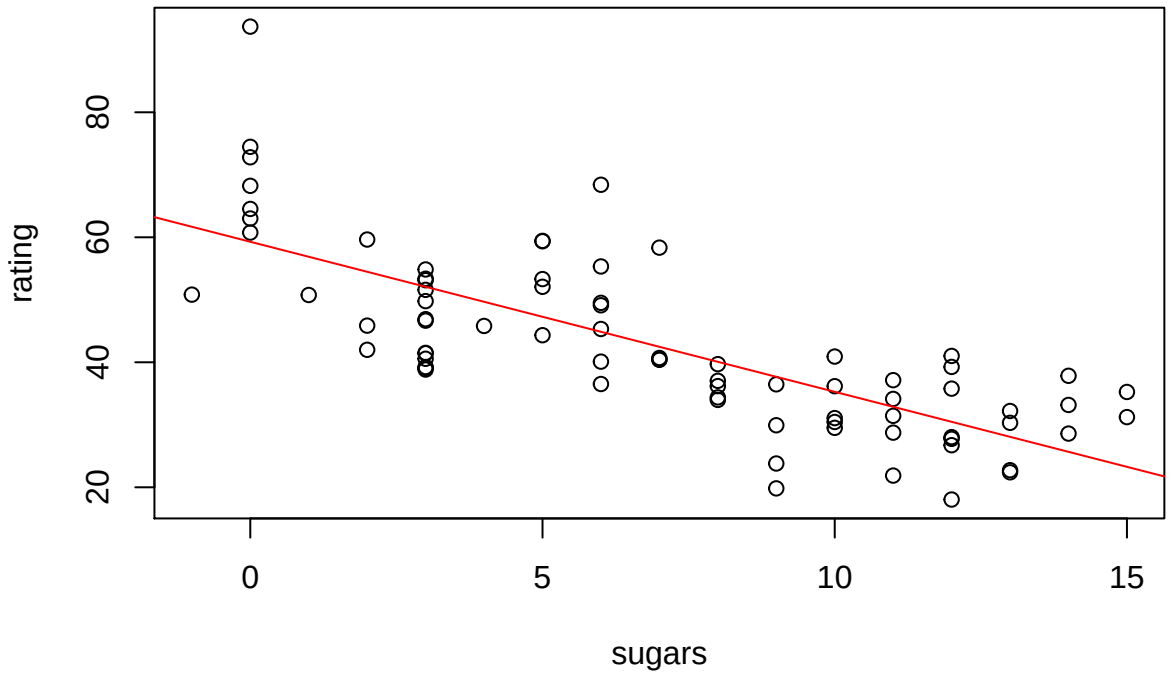
There is some information there that looks similar to the results we got from commands like t.test, chisq.test, and cor.test. There's a formula at the top that shows us what variables we ran in the regression model. Below that is a section about residuals. Then there is the coefficient section. That is the most important section. It has the columns "Estimate", "Std. Error", "t value", and "Pr(>|t|)". The Estimate column reports the beta coefficient. The second column reports standard error. The t value and Pr columns report on the statistical significance for the estimate of each variable. Then there is some other useful information about the performance of the model at the end.

Our first variable is (Intercept) which is R's way of talking about the alpha coefficient or y-intercept. So the alpha is 59 and that has a t value of 30 against the null hypothesis that the alpha would be 0. The last column gives that a very low p-value, which means the intercept is statistically significant. The intercept is usually not important but we can interpret it –if a cereal has 0 grams of sugar, it has an expected rating of 59.

The second variable is "sugars" which is the one we really care about. It is -2.4008. The negative sign indicates that increasing sugar by 1 gram results in 2.4 lower rating points. If we look over to the p-value column we see 1.15e-15 for p-value which is really small. The low p-value indicates that the sugar variable is statistically significant. Thus, the regression model reports that sugar has a negative effect on rating and the effect is statistically significant.

Lastly, I have a plot of the data now with a red regression line running through the data points.

```
plot(rating~sugars) +
abline(lm(rating~sugars), col = "red")
```




```
## integer(0)
```