Multiple Regression

Load in the Data

For this exercise, we will use the "Happiness.csv" dataset found on the UBlearns site for the lab. Download it and import the data. Make sure the "Headiings"" option is Yes or the "First Row as Names"" box is checked. Then attach the data.

```
attach(Happiness)
```

You can look at the data when you load it in to see that the top five happiest countries are Norway, Denmark, Iceland, Switzerland, and Finland. We can look at the names of the variables and summarize a few of them.

```
names(Happiness)
```

```
##  [1] "Country"                          "Happiness.Rank"
##  [3] "Happiness.Score"                  "Whisker.high"
##  [5] "Whisker.low"                      "Economy..GDP.per.Capita."
##  [7] "Family"                           "Health..Life.Expectancy."
##  [9] "Freedom"                          "Generosity"
## [11] "Trust..Government.Corruption." "Dystopia.Residual"
```

```
summary(Happiness.Score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.693   4.505   5.279   5.354   6.101   7.537
```

```
summary(Economy..GDP.per.Capita.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.6634  1.0646  0.9847  1.3180  1.8708
```

```
summary(Health..Life.Expectancy.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.3699  0.6060  0.5513  0.7230  0.9495
```

```
summary(Trust..Government.Corruption.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.05727 0.08985 0.12312 0.15330 0.46431
```

Multiple Regression Model

Let's try to explain the Happiness Score by using some of the other variables in the data. If I wanted to propose a hypothesis, I would expect that higher economic growth would lead to a higher happiness score. By including the other variables in the model I can control for their impact on happiness. Thus I would be able to see the impact of economic growth controlling for life expectancy or the level of freedom in a country.

```
model1 <- lm(Happiness.Score ~ Economy..GDP.per.Capita. + Family +
               Health..Life.Expectancy. + Freedom +
               Generosity + Trust..Government.Corruption.)
summary(model1)
```

```
##
## Call:
## lm(formula = Happiness.Score ~ Economy..GDP.per.Capita. + Family +
```

```
##     Health..Life.Expectancy. + Freedom + Generosity + Trust..Government.Corruption.)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.52798 -0.25219 -0.02277  0.28526  1.20417
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.7430     0.1874   9.303  < 2e-16 ***
## Economy..GDP.per.Capita.        0.7844     0.2045   3.836 0.000185 ***
## Family                          1.1178     0.2021   5.532 1.40e-07 ***
## Health..Life.Expectancy.        1.2889     0.3215   4.009 9.65e-05 ***
## Freedom                         1.4757     0.3425   4.309 2.98e-05 ***
## Generosity                      0.3807     0.3293   1.156 0.249524
## Trust..Government.Corruption.   0.8266     0.4843   1.707 0.089975 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4998 on 148 degrees of freedom
## Multiple R-squared:  0.8124, Adjusted R-squared:  0.8048
## F-statistic: 106.8 on 6 and 148 DF,  p-value: < 2.2e-16
```

The results give strong support for my hypothesis that higher economic growth will lead to higher happiness. Every one point growth in GDP per capita increases the happiness score by 0.784. The result is statistically significant with a p-value that is below the typical .05 threshold. Looking at the other variables, many of them also have a positive impact on happiness. A higher family rating (whatever that is), longer life expectancy, and more freedom have a positive and statistically significant impact on happiness. Generosity is positive but not statistically significant so we cannot be certain that the estimate is statistically distinguishable from the null hypothesis of 0. Trust in government is also positive but the p-value is .08 which is above the typical scientific standard of .05. Lastly, the R-squared statistic of .80 tells us the model does a decent job of explaining the variance in happiness scores with the independent variables we have included.

Using Coefficients to Generate Predicted Values

Now that we have a regression model that we like, we could use it to generate predictions based on specific data points. Let's do this for the United States. We need the values for each variable for the United States included in the model. I'll put the values into a vector called "UnitedStates" in the order that the variables appeared in the model. I'll also include a 1 at the beginning for the intercept.

```
UnitedStates <- c(1, 1.55, 1.42, .77, .51, .39, .14)
prediction <- sum(model1$coefficients*UnitedStates)
prediction
```

```
## [1] 6.555393
```

The predicted value (y-hat) for the United States is 6.56 and we can compare that to the actual observed value of 6.99. The United States is happier than the model would predict.

In summary, regressions can be used to test hypotheses about variables. Does a variable have a statistically significant effect on a dependent variable? What is the size of the effect? Further, we can build regression models and make predictions based on those models. In this way, regression is a powerful tool for understanding what is happening around us.