Advanced Statistical Analysis

Everything is regression

Independent samples t-test. Back in Exercise 1c, you ran a t-test looking at the opinions of men and women and how many angry geese they think they could take on.

Here are the data:
https://docs.google.com/spreadsheets/d/19cDaky3wKr2Dy9Gh5trx04vMvSEJ8VXjP0jBr6Ayu_o/edit?usp=sharing

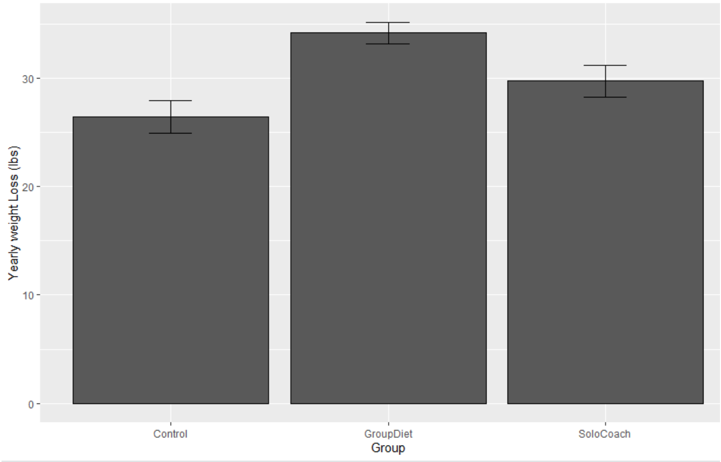
Load the data into R and run an independent samples t-test. **Report the results here:**

Now let's try analyzing these data using a regression. **Write out a regression equation that can be used to model your data:**

Run your regression analysis and report your results here:

How do the results compare for your t-test from Exercise 1 and your regression analysis here?

Back in Exercise 2 we ran a one-way ANOVA on some weight loss data. We had three conditions: Control, Group dieting, and Solo coaching, and we looked at yearly weight loss:



First, we need to check the assumptions required for the ANOVA. The assumptions are:

1) Each sample is an independent random sample (done during testing)

2) The distribution of the response variable follows a normal distribution (Shapiro-Wilk)

3) The population variances are (roughly) equal across the groups tested (Levene's)

Using the data set *Exercise1Data.xlsx*, check the assumptions for the ANOVA. Report your results here:

Run your ANOVA analysis, including post-hoc comparisons. Report your results here:

Now let's try analyzing these data using a regression. **Write out a regression equation that can be used to model your data:**

Run a multiple linear regression and report your results here:

How do the results of your two analyses compare?

With multiple regression, we also have assumptions that are similar, but framed slightly differently.

1) Sample Size

Many sources will cite how many samples you need to obtain a reliable regression equation. A general rule of thumb is that a basic multiple regression model can be formed with 10 samples for each predictor variable that you wish to use. Others cite larger numbers, for example Pallant (2011) says that for “social science” research 15 samples should be used for each predictor you wish to include. She goes on to cite even more conservative research which cites that you should have a sample size where the following equation is valid: $N > 50 + 8m$, where N is the total sample size and m is the number of predictors. We talked about some of this in class.

2) Multicollinearity and Singularity

If predictor variables have a very high correlation ($r > 0.90$) we risk violating this assumption. It’s best to check this assumption before running the regression, and then “diagnostics” will be completed after running the regression to see if we did violate this. Predictor variables that are mathematically related cannot be used as individual predictors (singularity).

3) Outliers

Multiple regression is very sensitive to outliers. We can check for extreme scores before we begin, but in general we will check for outliers after we run the analysis.

4) Residuals

There are a number of assumptions about the residuals that must be met for a multiple regression model. These will be checked after the model is run. The assumptions are as follows (Pallant 2011):

- *Normality*: The residuals should be normally distributed about the predicted DV scores.
- *Linearity*: The residuals should have a linear relationship with predicted DV scores.
- *Homoscedasticity*: the variance of the residuals should be the same for all predicted scores.

For your future statistics practice, be aware of these assumptions. What follows is an *optional exercise* that walks through how to check these assumptions. Unfortunately, we won’t have time in this course to really dig into this.

Full walk thru of a multiple regression analysis

We will work through an example using a baseball dataset (Baseball.xlsx). We will attempt to create a regression model which predicts a player's salary based on chosen hitting statistics.

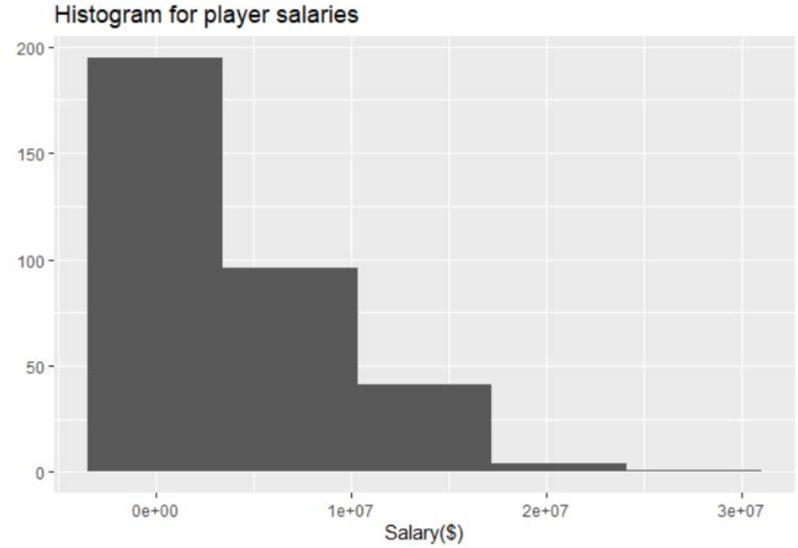
Begin by loading the data.

```
     DataSet <- read.xlsx("Baseball.xlsx",1)
```

Let's check the distribution of our salary data.

First run a Shapiro – Wilk test on DataSet$SALARY. This should suggest non-normal data. We can verify this with a histogram which we learned way back on page 4 of Mini-exercise 1.

The histogram shows a very clear positive skew:



To correct this skew, we can apply a square root correction, with the following code:

```
     DataSet$SR_Salary <- sqrt(DataSet$SALARY)
```

You will note that this will eliminate most of the skewness in the data, however a Shapiro-Wilk test will still indicate that it is not normally distributed. This is actually okay, since our variables don't have to be normally distributed for multiple regression, however the residuals at the end of the analysis should be.

Next, we can choose our predictor variables and run a correlation matrix to see if there is a risk of multicollinearity.

I've chosen to analyze:

HR (Home Runs), G (Games Played), PA (Plate Appearances), and TB(Total Bases).

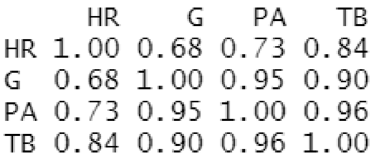
We can select only the variables we are interested in with the following code.

```
     myvars <- c("HR","G","PA","TB")
     SmallData <- DataSet[myvars]
```

And create a correlation matrix:

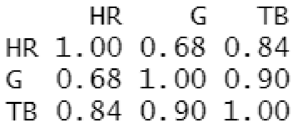
```
library(Hmisc)
rcorr(as.matrix(SmallData), type = "pearson")
```

This gives the following matrix:



```
     HR   G    PA   TB
HR 1.00 0.68 0.73 0.84
G  0.68 1.00 0.95 0.90
PA 0.73 0.95 1.00 0.96
TB 0.84 0.90 0.96 1.00
```

We can see that the number of games played is highly correlated with two variables (PA and TB). We also see that TB is also highly correlated with PA. Since this is the highest correlation, we will remove PA from the model before we begin (just modify the code from the previous page). The new correlation matrix looks like:



```
     HR   G    TB
HR 1.00 0.68 0.84
G  0.68 1.00 0.90
TB 0.84 0.90 1.00
```

We still have a high correlation between G and TB, but we will ignore this for now.

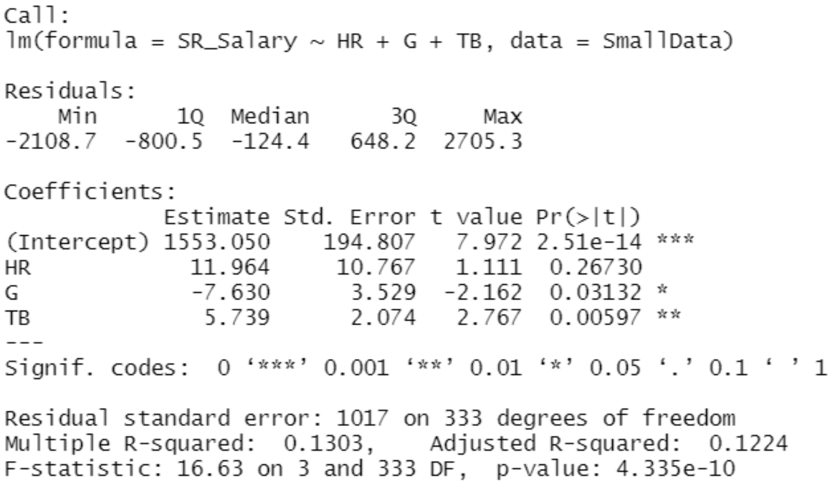
We are now ready to run our model. I've chosen to make a new dataset with just the variables we are planning on using in our regression model.

```
myvars <- c("SR_Salary","HR","G","TB")
SmallData <- DataSet[myvars]
```

We can then run the model with the following code, and get a summary.

```
Model1 <- lm(SR_Salary ~ HR + G + TB, data = SmallData)
summary(Model1)
```

This gives the following output:



```
Call:
lm(formula = SR_Salary ~ HR + G + TB, data = SmallData)

Residuals:
    Min      1Q  Median      3Q     Max 
-2108.7  -800.5  -124.4   648.2  2705.3 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1553.050    194.807   7.972 2.51e-14 ***
HR            11.964     10.767   1.111  0.26730    
G             -7.630      3.529  -2.162  0.03132 *  
TB             5.739      2.074   2.767  0.00597 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1017 on 333 degrees of freedom
Multiple R-squared:  0.1303,	Adjusted R-squared:  0.1224 
F-statistic: 16.63 on 3 and 333 DF,  p-value: 4.335e-10
```

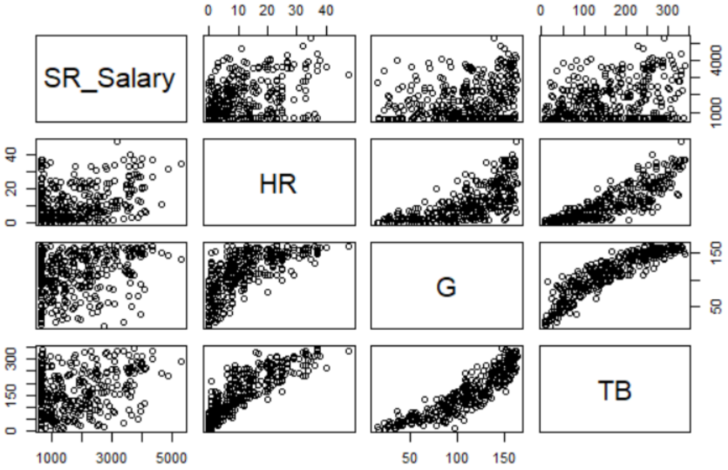
We can see that we have a significant regression model, F(3,333) = 16.63, p < 0.001. We have explained 13.03% of the variance in "SR_Salary", with an adjusted R-squared value of 0.1224. Looking at the predictors, we can see that the intercept is significantly different than zero. In our situation it means that the baseball players got paid, even if they didn't get any of the other statistics we looked at.

We see that the number of games played (G) is a significant predictor. Interestingly it has a negative value, indicating a negative relationship between games played and salary. Total Bases was significant and we can see that there is a positive relationship between TB and SR_Salary. Lastly, we see that HR is not a significant predictor of salary.

Running Diagnostics

We are now ready to do some checking of our model to see how valid it is.
We can begin by looking at the data in a pair-wise manner with the following code:

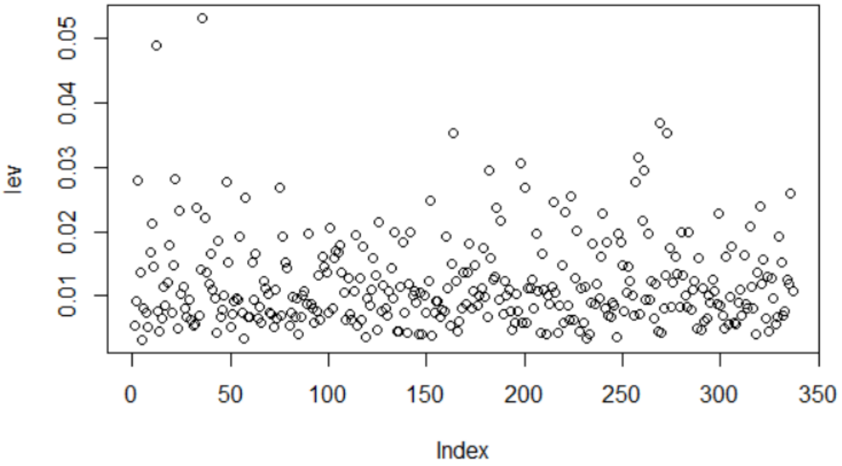
```
pairs(SmallData)
```



Looking at the SR_Salary, we can anticipate that we may have some issues in creating a model that will explain a lot of variance in this model, but we will try [the pair-wise plots that include SR_Salary do not show a clear trend].

First, we will look at the "Leverage" of the individual points in the model. Leverage indicates a datapoint's ability to "move" the model in the y-direction. A strong outlier may have a strong effect on the regression model. Leverage values are between 0 and 1, with 0 indicating that the datapoint had no effect on the regression model, and 1 indicating that the model was forced to follow that point.
We can calculate and plot leverage values with the following code:

```
lev <- hat(model.matrix(Model1))
plot(lev)
```



We can see that there are no overly "influential cases" in our dataset. If we had values above 0.25, we would be concerned, and may want to see if this case is an outlier.
To do this, we can add the "lev" variable to our dataset, and then View the data and sort the data to see which case had the highest value.

```
SmallData2 <- data.frame(SmallData,lev)
```

Pallant (2011) suggests using the M-Dist approach to look for significant outliers.
We can calculate the M-dist with the following code, and determine the max value.

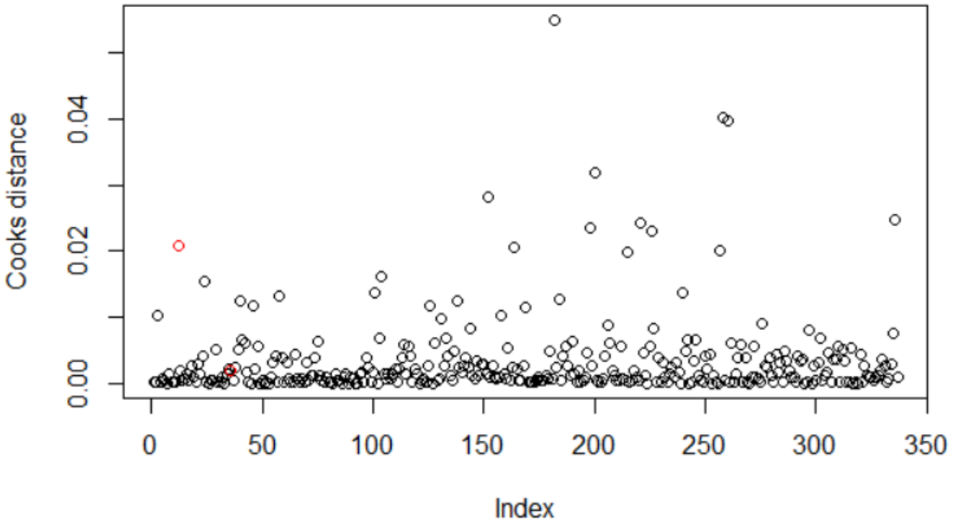
```
means <- colMeans(SmallData) #calculate columns means
Sx <- cov(SmallData) #calculate the co-variance matrix
MDists <- mahalanobis(SmallData,means,Sx)
max(MDists)
```

We can use a look-up table to determine what the cut-off is. In this case, our maximum value was 17.04497, and Pallant suggests a maximum value of 16.27.

We can merge our MDist value to see which cases are outliers. In this case, the MDist suggest that cases 12 and 36 are multivariate outliers, and should be removed. This actually agrees with the leverage calculation which identified these two cases as the most influential cases.

Remove these two cases and re-run the multiple regression. What has changed?

An alternative method to check for outliers and leverage is the Cook's distance.



```
cook <- cooks.distance(Model1)
plot(cook,ylab ="Cooks distance")
points(12,cook[12], col ="red")
points(36,cook[36], col ="red")
```

Note all the Cook's distances are fine. In fact, points 12 and 36 are not even the worst cases.
So, which of these cases should we use?

Leverage indicates "influential" cases that have a large impact on the model. These points may be outliers, or just important cases that end up carrying a lot of weight in the model. In this case we found no influential cases.

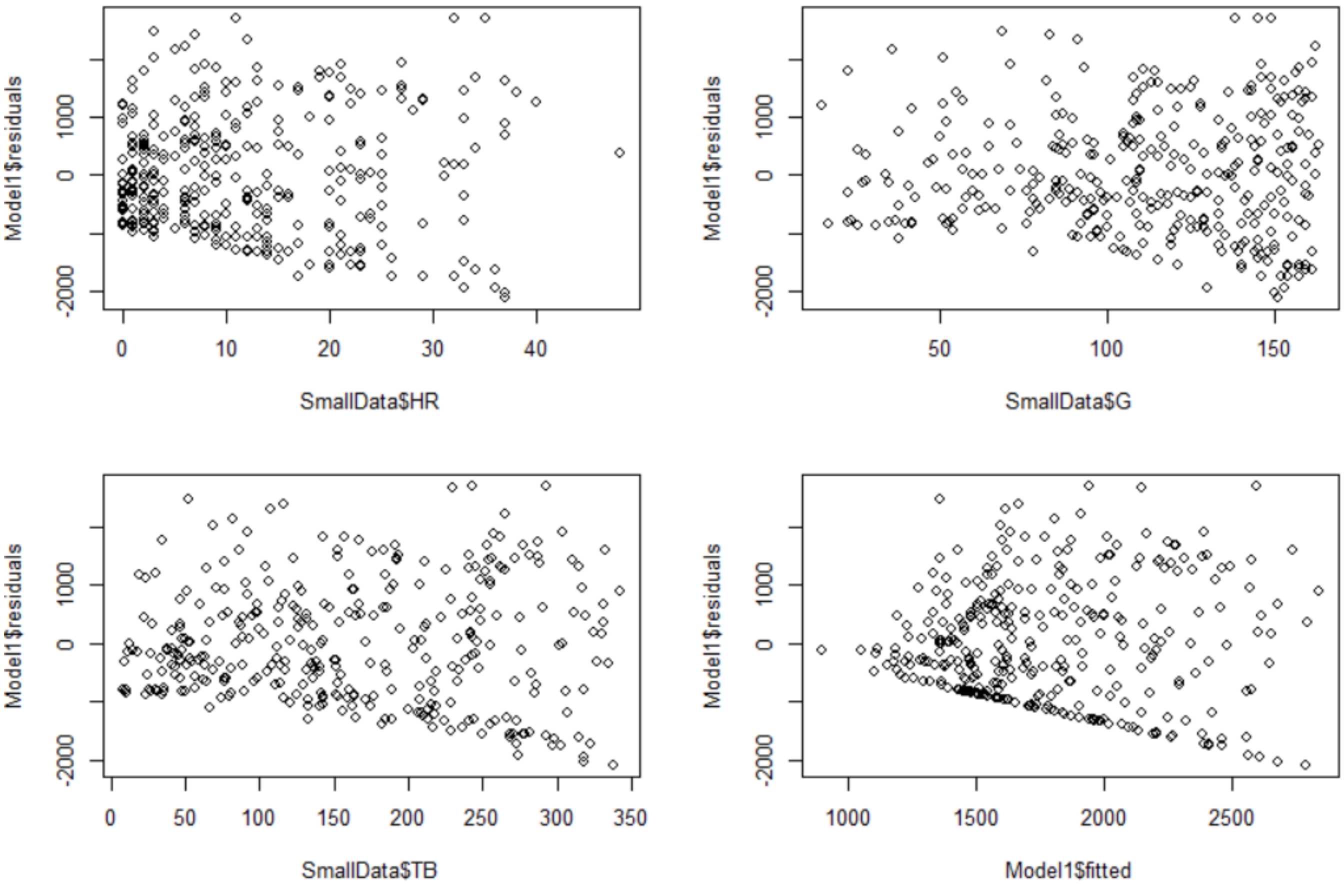
M-Dists will look for multivariate outliers. We found two cases, 12 and 36, which may be an issue.

Cook's distance evaluates datapoints on the basis of both leverage and being outliers. In this case the two points we identified with M-Dists are not a problem, and thus can likely be left in the model.

Evaluating Residuals

First, we will plot the residuals against the predicted values and the independent variables. The residuals should be randomly scattered about zero and the width should be constant, for the homoscedasticity assumption to hold. We can make a plot for all three measures with the following code.

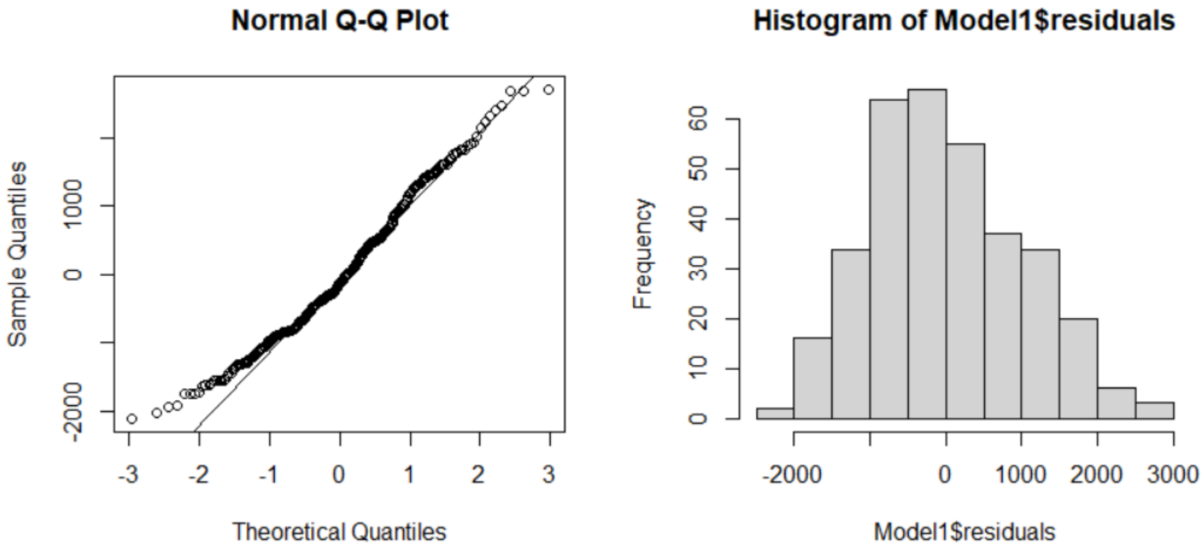
```
    #Residual Plot
    par(mfrow=c(2,2))
    plot(SmallData$HR, Model1$residuals)
    plot(SmallData$G, Model1$residuals)
    plot(SmallData$TB, Model1$residuals )
    plot(Model1$fitted, Model1$residuals )
```



What issues may we have with our model?

Now we can check the normality of our residuals. This should be done with a qq-plot or a histogram.

```
    #Residual Normaility plots
    par(mfrow=c(1,2))
    qqnorm(Model1$residuals)
    qqline(Model1$residuals)
    hist(Model1$residuals)
```



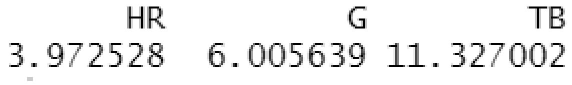
Looking at these plots, we can see our residuals are positively skewed, which isn't too surprising considering we began with a skewed outcome variable. But all things considered they actually look okay.

Checking Multicollinearity

Finally, we can run a statistical test to see if we have any issues with multicollinearity. To do this we run the VIF test, which stands for Variance Inflation Factor. A score greater than 10 indicates that a high correlation exists with another predictor, and multicollinearity is present.

```
    #Check Collinearity.
    library(car)
    vif(Model1)
```

This gives the following output:



```
      HR          G         TB 
 3.972528  6.005639 11.327002 
```

Thus we can see that TB is showing some multi-collinearity issues.

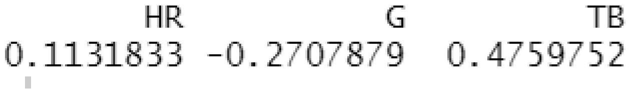
Last but not least… Standardized Coefficients.

Say our model we made was good (Which it's not), we might ask which of our three predictor variables has the biggest influence on our outcome variable. We could look at the coefficients from the regression table, however the units of these are not uniform. Thus, we can "standardize" our coefficients based on their standard deviation.

We can do this with the following code:

```
    #Get standardized coefficients.
    library(QuantPsyc) # you will need to install this. ☺
    lm.beta(Model1)
```

This gives the following output.



```
       HR          G         TB 
0.1131833 -0.2707879  0.4759752 
```

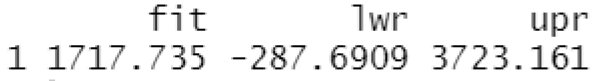
It indicates that a 1 SD change in HR, will result in 0.11 SD change in SR_Salary. We can also assess which will have the largest effect, with TB having the largest effect on salary.

Okay, Now this is really the last thing…. Using our model.

Say we want to predict the salary of a player who plays 92 games, has a 126 total bases, and 12 HRs. We can do this with the following command, and get prediction with a confidence interval based on our model.

```
predict(Model1,data.frame(HR=12,TB = 126, G = 92),interval="prediction")
```

This gives us the following output:



```
       fit       lwr      upr
1 1717.735 -287.6909 3723.161
```

This tells us we could expect a SR_Salary of 1717.735 with the stats we presented… note we took the square root of the salary, so we have to square the value we get, so we could expect a mean salary of approximately $2.95 million for putting up those numbers. Note, our confidence interval is quite wide, because of the relatively low R^2 value we have. We also have a problem because we have a negative value for our lower bound…

References:

Regression Diagnostics: http://www.statmethods.net/stats/rdiagnostics.html

Example 8.14: Generating standardized regression coeffcients. https://www.rbloggers.com/example-8-14-generating-standardized-regression-coefficients/

Pallant, J (2011). SPSS Survival Manual. 4th Edition. Allen & Unwin Publishing. Crows Nest, Australia.