Week 4.1 - Introduction to Regression Analysis

Import useful libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts ------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

In this section, we will use the NHL__Team__Stats dataset we compiled and cleaned up in the assignment for Week 2.

Import both NHL__Team__Stats and NHL__Team__R__Stats data from Week 2 assignment into R

```
NHL_Team_Stats = read.csv("~/Google Drive/Sports Analytics Moocs/MOOC 1 - Foundations of sports analytic
NHL_Team_R_Stats = read.csv("~/Google Drive/Sports Analytics Moocs/MOOC 1 - Foundations of sports analyt
head(NHL_Team_Stats)
```

```
##   tricode  pp  pk ppgf competition_name tid type win goals_for goals_against
## 1     ANA  35  27    9 2010 NHL Playoff  21    3   2        19            22
## 2     BOS 126 116   22 2010 NHL Playoff  20    3  16        76            48
## 3     BUF  48  46   13 2010 NHL Playoff  17    3   3        17            22
## 4     CHI  27  39    6 2010 NHL Playoff   6    3   3        22            16
## 5     DET  59  55    6 2010 NHL Playoff  18    3   7        36            27
## 6     LAK  24  27    5 2010 NHL Playoff  41    3   1        17            19
##   game_count          team_name   win_pct   avg_gf   avg_ga
## 1          6      Anaheim Ducks 0.3333333 3.166667 3.666667
## 2         24      Boston Bruins 0.6666667 3.166667 2.000000
## 3          7     Buffalo Sabres 0.4285714 2.428571 3.142857
## 4          7 Chicago Blackhawks 0.4285714 3.142857 2.285714
## 5         11  Detroit Red Wings 0.6363636 3.272727 2.454545
## 6          5  Los Angeles Kings 0.2000000 3.400000 3.800000
```

Regression analyses in R

At the end of the assignment in week 2, we observed that there is a linear relationship between total goals for and winning percentage. Let's run a regression where winning percentage is the dependent variable and total goals for is the explanatory variable. - We can use the command "lm()" to indicate an ordinary least squared regression. - The "summary()" function provides a series of test statistics when printed in the console.

```
reg1 = lm(data = NHL_Team_R_Stats, win_pct ~ goals_for)
summary(reg1)
```

```
##
```

```
## Call:
## lm(formula = win_pct ~ goals_for, data = NHL_Team_R_Stats)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.172712 -0.037941  0.000987  0.043614  0.147081
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.1781144  0.0421850  -4.222 3.84e-05 ***
## goals_for    0.0029524  0.0001838  16.067  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05986 on 179 degrees of freedom
## Multiple R-squared:  0.5905, Adjusted R-squared:  0.5882
## F-statistic: 258.2 on 1 and 179 DF,  p-value: < 2.2e-16
```

Interpreting results

From the result table, we can see that the dependent variable is winning percentage ("win_pct") and there are 181 observations in this regression. The independent variable is the number of goals for the team ("goals_for"). An intercept is also included in the regression.

The estimated coefficient on goals_for is 0.003. This means that an additional goal scored by the team will increase the team's winning percentage of 0.003. The estimate on the intercept is -0.1781. This means that without scoring any goal, the winning percentage for the team would be -0.1781. As we know, the winning percentage cannot be negative. The reason we get a negative estimate on the intercept is because in our sample, there is not a single game where a team scored zero goals.

- t-statistics and p-value

t-statistics is defined as the estimated coefficient divided by its standard error. If the estimated coefficient is large compared to its standard error, then it is likely to be different than zero.

p-value is defined as the probability of obtaining a result as extreme as the result actually observed, in this case, the t-statistics we have in the regression analysis. Comparing the t-statistics with the student t distribution, if 95% of the t distribution is closer to the mean than the t-statistics, we will have p-value of 0.05, which is also referred to a 5% significance level. A p-value no more than 0.05 (5%) is generally accepted in rejecting the null hypothesis. We say that the estimated coefficient is statistically significant at the 5% level.

In this regression, the p-value of the goals_for variable is 0.000 which suggests that the estimate is statistically significant at the 1% level.

- R-squared

R-squared measures the goodness of fit of the model. The R-squared of a regression is the fraction of the variation in the dependent variable that is accounted for by the independent variables. R-squared is always between 0 and 1. The larger the R-squared, the more variation is accounted for by the regression model.

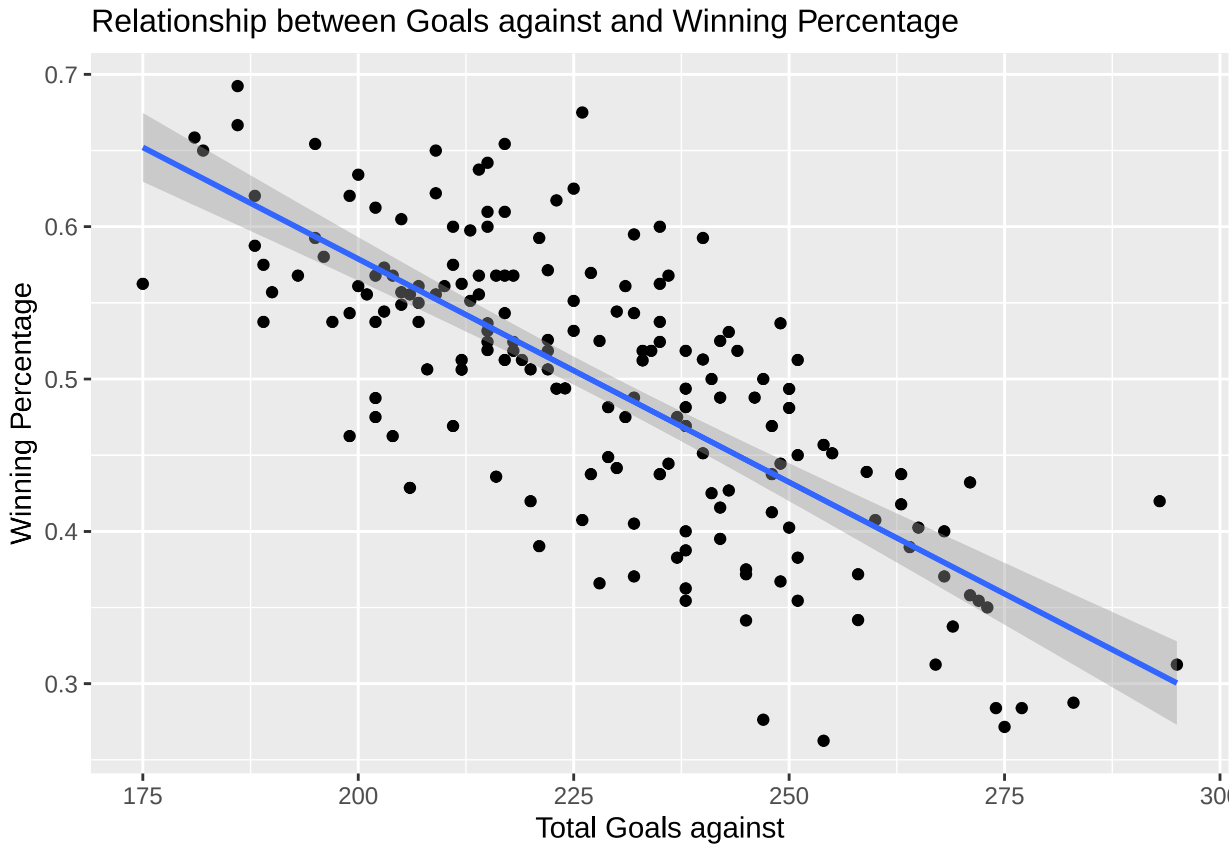
In this regression, the R-squared is 0.591 which means that approximately 59.1% of the variation of the winning percentage is accounted for by the model.

Let's explore the relationship between goals_against and winning percentage in the regular season.

- Create a scatter plot to depict the relationship between total goals against and winning percentage without seperating the data by competition

```
ggplot(NHL_Team_R_Stats, aes(x=goals_against, y=win_pct)) +
  geom_point() + stat_smooth(method = "lm") +
  labs(title = "Relationship between Goals against and Winning Percentage",
       x = "Total Goals against", y = "Winning Percentage") +
  theme(plot.title = element_text(size = 12))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



- Calculate the correlation coefficient between goals against and winning percentage

```
cor(NHL_Team_R_Stats$goals_against, NHL_Team_R_Stats$win_pct)
```

```
## [1] -0.7445118
```

- Run a simple linear regression to find NHL team winning percentage as a function of total goals against

```
reg2 = lm(data = NHL_Team_R_Stats, win_pct ~ goals_against)
summary(reg2)
```

```
##
## Call:
## lm(formula = win_pct ~ goals_against, data = NHL_Team_R_Stats)
##
## Residuals:
```

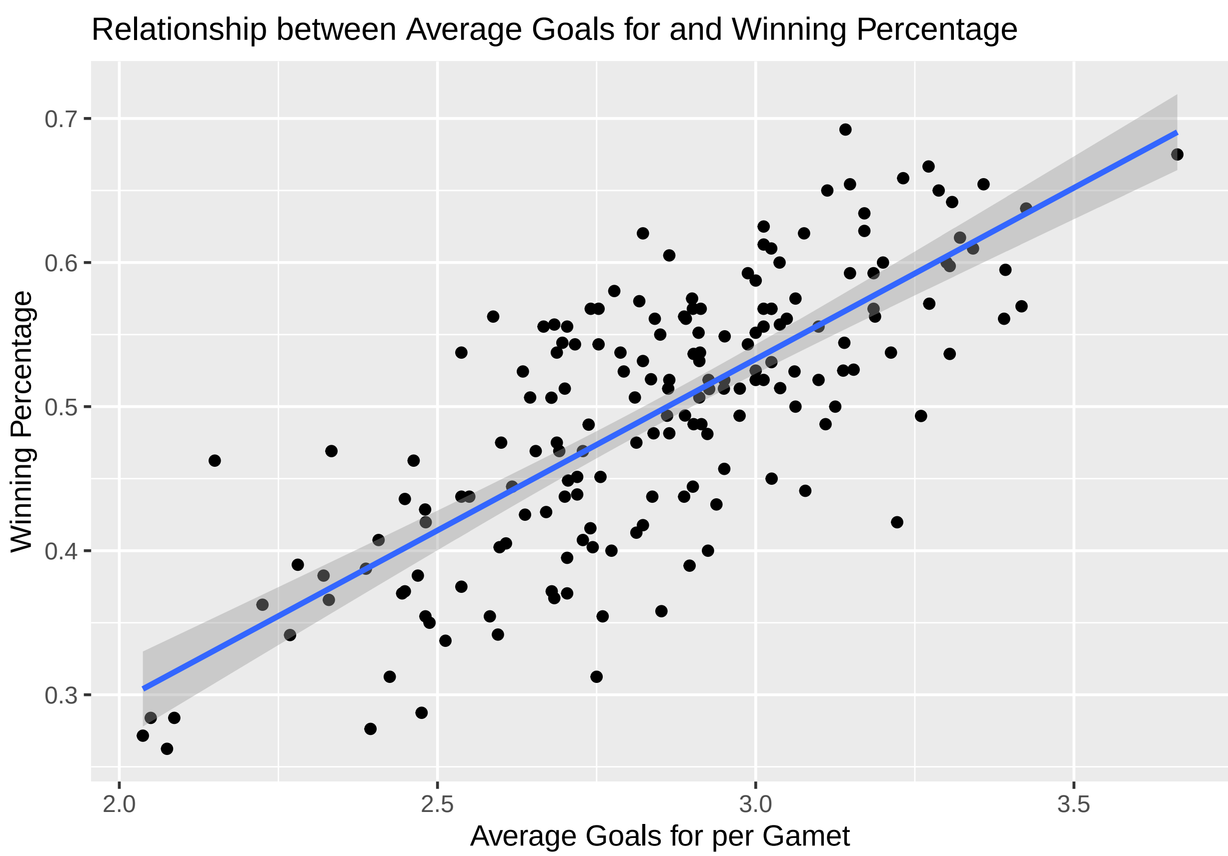
```
##        Min         1Q    Median        3Q       Max
## -0.164750 -0.038554  0.001689  0.041831  0.172379
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.1650751  0.0450898   25.84   <2e-16 ***
## goals_against -0.0029312  0.0001965  -14.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06245 on 179 degrees of freedom
## Multiple R-squared:  0.5543, Adjusted R-squared:  0.5518
## F-statistic: 222.6 on 1 and 179 DF,  p-value: < 2.2e-16
```

Self Test

1. Use the regular season data, create a scatterplot and a regression line to demonstrate the relationship between average goals for per game and winning percentage.

```r
ggplot(NHL_Team_R_Stats, aes(x=avg_gf, y=win_pct)) +
  geom_point() + stat_smooth(method = "lm") +
  labs(title = "Relationship between Average Goals for and Winning Percentage",
      x = "Average Goals for per Gamet", y = "Winning Percentage") +
  theme(plot.title = element_text(size = 12))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



2. Run a linear regression where winning percentage is the dependent variable and average goals for is the

explanatory variable

```r
reg3 = lm(data= NHL_Team_R_Stats, win_pct ~ avg_gf)
summary(reg3)
```

```
##
## Call:
## lm(formula = win_pct ~ avg_gf, data = NHL_Team_R_Stats)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.166065 -0.034605 -0.000907  0.040808  0.131641
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.18038    0.04388  -4.111 5.99e-05 ***
## avg_gf       0.23779    0.01534  15.496  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06113 on 179 degrees of freedom
## Multiple R-squared:  0.5729, Adjusted R-squared:  0.5705
## F-statistic: 240.1 on 1 and 179 DF,  p-value: < 2.2e-16
```

3. Interpret the coefficient on the average goals for, is this estimate statistically significant?
4. How well does this regression do in fitting the data?

Multiple Regression - more than one explanatory variables.

Often times, the outcome variable of interest is affected by multiple factors. We can specify a regression equation where the outcome is function of more than one explanatory variables.

Let's run a linear regression where winning percentage is a function of both average number of goals for per game and average number of goals against per game.

```r
reg4 = lm(data= NHL_Team_R_Stats, win_pct ~ avg_gf + avg_ga)
summary(reg4)
```

```
##
## Call:
## lm(formula = win_pct ~ avg_gf + avg_ga, data = NHL_Team_R_Stats)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.072415 -0.019484 -0.001204  0.019689  0.090810
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.486210   0.031249   15.56   <2e-16 ***
## avg_gf       0.190896   0.006994   27.30   <2e-16 ***
## avg_ga      -0.187412   0.006892  -27.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.027 on 178 degrees of freedom
## Multiple R-squared:  0.9171, Adjusted R-squared:  0.9162
```

```
## F-statistic: 985.1 on 2 and 178 DF,  p-value: < 2.2e-16
```

Interpret the coefficients

- Average goals for: for the same average number of goals against, scoring one more goal per game will increase the winning percentage by 0.1909 (19.09%)
- Average goals against: having the same average number of goals for per game, conceding one more goal per game will decrease the winning percentage by 0.1874 (18.74%)

Regression with categorical variables

In the above regressions, we focus on using quantitative variables as explanatory variables. We could also include categorical variables as explanatory variables in the regression as well.

Essentially, when we incorporate a categorical variable, we first transform it into dummy variable(s) that carry value of either 0 or 1. We then use the dummy variable(s) into our regression.

Let's consider the dataset that includes both regular season and playoff. In this dataset, the variable "type" captures whether a game is a regular season game or playoff game. type=2 means it is regular season competition while type=3 means it is a playoff game.

We will first convert variable "type" into categorical variable.

```r
NHL_Team_Stats$type = as.factor(NHL_Team_Stats$type)
```

Now we can run a regression where winning percentage is a function of average goals for and the type of competition.

```r
reg5 = lm(data= NHL_Team_Stats, win_pct ~ avg_gf+type)
summary(reg5)
```

```
##
## Call:
## lm(formula = win_pct ~ avg_gf + type, data = NHL_Team_Stats)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41894 -0.05527  0.00598  0.06396  0.31745
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01966    0.03524  -0.558    0.577
## avg_gf       0.18184    0.01219  14.914   <2e-16 ***
## type3       -0.01602    0.01192  -1.344    0.180
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.102 on 366 degrees of freedom
## Multiple R-squared:  0.4263, Adjusted R-squared:  0.4232
## F-statistic:   136 on 2 and 366 DF,  p-value: < 2.2e-16
```

A dummy variable = 1 if type= 3 (playoff) and = 0 if type = 2 (regular season) is included in the regression.

Interpretation: with the same average goals for per game, the winning percentage in the playoff games is 0.0160 (1.6%) lower than the winning percentage in the regular season games.

Self Test

1. Run a regression where winning percentage is a function of average goals for, average goals against, and control for the different competitions.
2. Interpret the coefficients.

```
reg6 = lm(data= NHL_Team_Stats, win_pct ~ avg_gf+competition_name)
summary(reg6)
```

```
##
## Call:
## lm(formula = win_pct ~ avg_gf + competition_name, data = NHL_Team_Stats)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47773 -0.04885  0.00691  0.06275  0.27451
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           -0.095074   0.042342  -2.245   0.0254
## avg_gf                                 0.192524   0.012522  15.375   <2e-16
## competition_name2010 NHL Regular Season  0.041102   0.031551   1.303   0.1935
## competition_name2011 NHL Playoff         0.071867   0.036229   1.984   0.0481
## competition_name2011 NHL Regular Season  0.043896   0.031559   1.391   0.1651
## competition_name2012 NHL Playoff         0.061703   0.036212   1.704   0.0893
## competition_name2012 NHL Regular Season  0.056107   0.031518   1.780   0.0759
## competition_name2013 NHL Playoff         0.007903   0.035974   0.220   0.8262
## competition_name2013 NHL Regular Season  0.047975   0.031538   1.521   0.1291
## competition_name2014 NHL Playoff         0.053646   0.036160   1.484   0.1388
## competition_name2014 NHL Regular Season  0.050317   0.031529   1.596   0.1114
## competition_name2015 NHL Playoff         0.055683   0.036138   1.541   0.1243
## competition_name2015 NHL Regular Season  0.060239   0.031505   1.912   0.0567
## competition_name2016 NHL Playoff         0.027547   0.036233   0.760   0.4476
## competition_name2016 NHL Regular Season  0.052017   0.031519   1.650   0.0998
## competition_name2017 NHL Playoff        -0.016657   0.035977  -0.463   0.6437
## competition_name2017 NHL Regular Season  0.009985   0.031553   0.316   0.7518
##
## (Intercept)                           *
## avg_gf                                ***
## competition_name2010 NHL Regular Season
## competition_name2011 NHL Playoff       *
## competition_name2011 NHL Regular Season
## competition_name2012 NHL Playoff       .
## competition_name2012 NHL Regular Season .
## competition_name2013 NHL Playoff
## competition_name2013 NHL Regular Season
## competition_name2014 NHL Playoff
## competition_name2014 NHL Regular Season
## competition_name2015 NHL Playoff
## competition_name2015 NHL Regular Season .
## competition_name2016 NHL Playoff
## competition_name2016 NHL Regular Season .
## competition_name2017 NHL Playoff
## competition_name2017 NHL Regular Season
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1017 on 352 degrees of freedom
## Multiple R-squared:  0.451,  Adjusted R-squared:  0.426
## F-statistic: 18.07 on 16 and 352 DF,  p-value: < 2.2e-16
```

Regression with an interaction term

What if the impact of an independent variable depends on the value of another variable? We can use interaction terms to allow for different impact of a variable based on one or more levels of another categorical variable.

Let's consider the possibility that the average goals for may have different impact on winning percentage depending on the type of the game. We can run a regression of winning percentage on the average goals for, the type of the game, as well as the interaction between average goals for and type.

```r
reg7 = lm(data= NHL_Team_Stats, win_pct ~ avg_gf+type+avg_gf*type)
summary(reg7)
```

```
##
## Call:
## lm(formula = win_pct ~ avg_gf + type + avg_gf * type, data = NHL_Team_Stats)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41881 -0.05867  0.00752  0.06092  0.32651
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.17476    0.06094  -2.868  0.00438 **
## avg_gf        0.23645    0.02134  11.081  < 2e-16 ***
## type3         0.20293    0.07157   2.835  0.00483 **
## avg_gf:type3 -0.08020    0.02586  -3.102  0.00207 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1008 on 365 degrees of freedom
## Multiple R-squared:  0.441,  Adjusted R-squared:  0.4365
## F-statistic:    96 on 3 and 365 DF,  p-value: < 2.2e-16
```

Interpretations

- For regular season games (type =2), scoring one more goal per game can increase the winning percentage by 0.2365 (23.65%);
- For the playoff games (type =3), scoring one more goal per game will increase the winning percentage by 0.2365-0.0802=0.1563 (15.63%).

Self Test

Perform a similar exercise to find the relationship between the actual winning percentage and pythagorean winning percentage

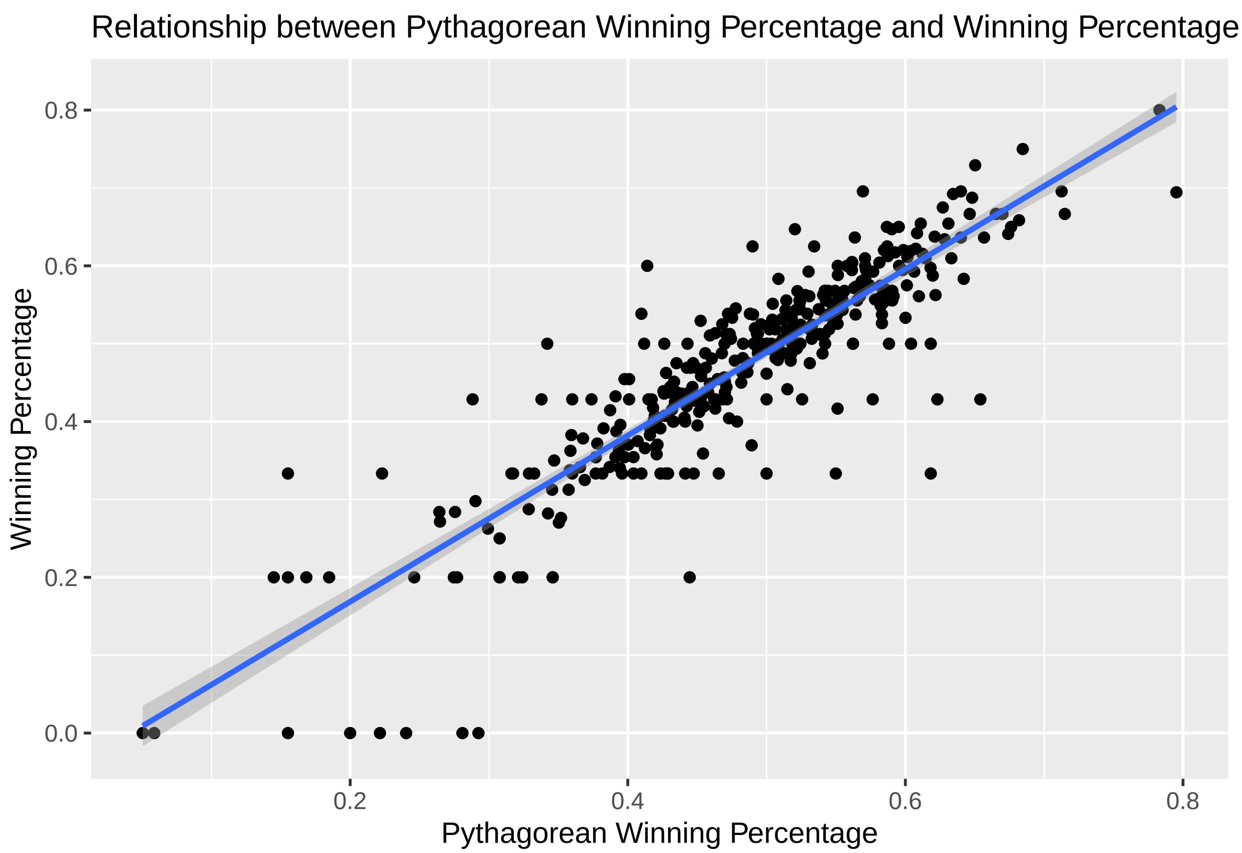
1. In the NHL_Team_Stats data, create the pythagorean winning percentage=goals_for$^{2/(\text{goals_for}}$2+goals_against^2), call this new variable "pyth_pct" (In R, ** is the operator for exponentiation. For example, the square of x would be x**2 in R)

```
NHL_Team_Stats$pyth_pct = NHL_Team_Stats$goals_for**2/(NHL_Team_Stats$goals_for**2+NHL_Team_Stats$goals_
```

2. Create a scatter plot to show the relationship between Pythagorean winning percentage and the actual winning percentage

```
ggplot(NHL_Team_Stats, aes(x=pyth_pct, y=win_pct)) +
  geom_point() + stat_smooth(method = "lm") +
  labs(title = "Relationship between Pythagorean Winning Percentage and Winning Percentage",
      x = "Pythagorean Winning Percentage", y = "Winning Percentage") +
  theme(plot.title = element_text(size = 12))
```

```
## `geom_smooth()` using formula 'y ~ x'
```



2. Run a linear regression (reg8) where winning percentage is the dependent variable and Pythagorean winning percentage is the explanatory variable.
3. Interpret the estimate on the Pythagorean winning percentage and the goodness of fit of the regression model.

```
reg8 = lm(data= NHL_Team_Stats, win_pct ~ pyth_pct)
summary(reg8)
```

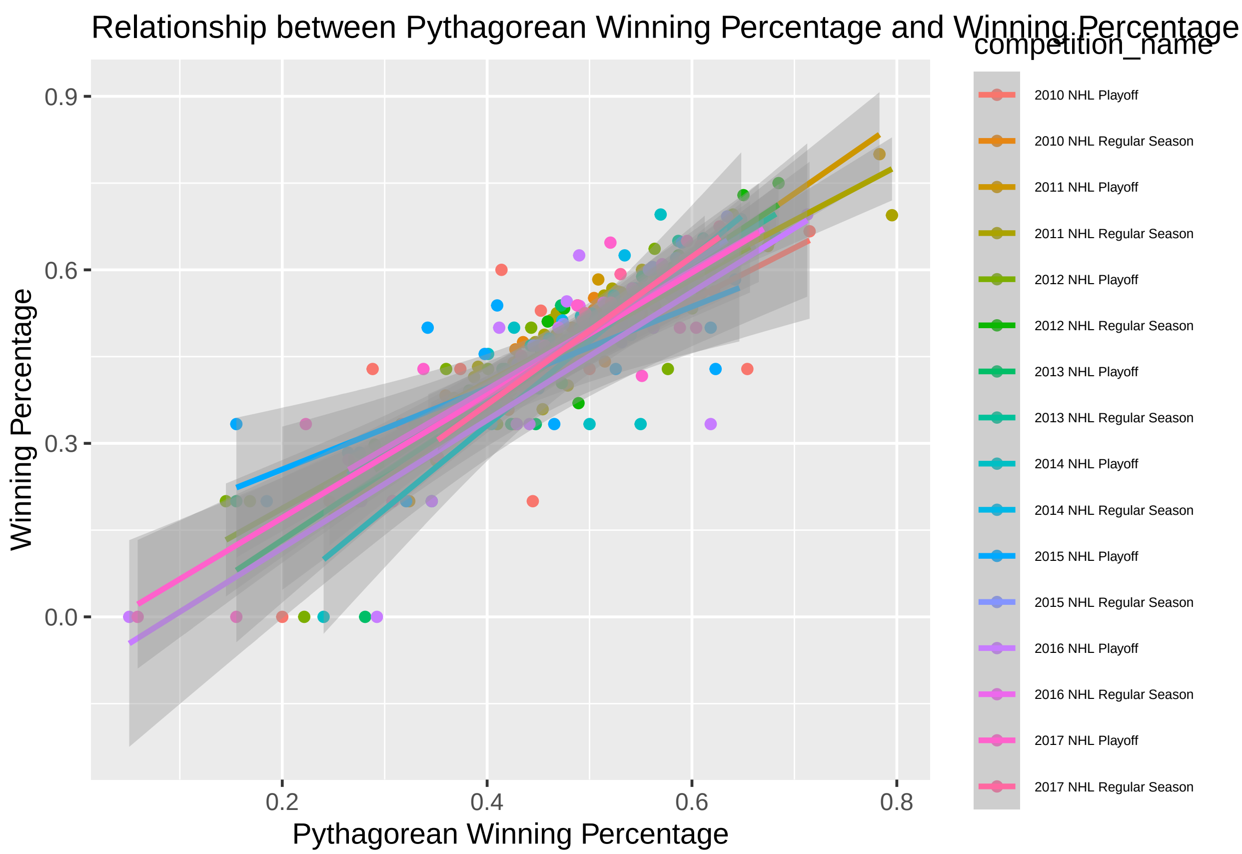
```
##
## Call:
## lm(formula = win_pct ~ pyth_pct, data = NHL_Team_Stats)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.281876 -0.021122  0.004732  0.033253  0.212436
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.04472    0.01465  -3.052  0.00244 **
## pyth_pct     1.06734    0.02964  36.011  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06316 on 367 degrees of freedom
## Multiple R-squared:  0.7794, Adjusted R-squared:  0.7788
## F-statistic:  1297 on 1 and 367 DF,  p-value: < 2.2e-16
```

4. Create a scatter plot to show the relationship between winning percentage and Pythagorean winning percentage, seperate the data points by the type of competition.

```r
ggplot(NHL_Team_Stats, aes(x=pyth_pct, y=win_pct, color = competition_name)) +
  geom_point() +
  geom_smooth(method=lm) +
  labs(title = "Relationship between Pythagorean Winning Percentage and Winning Percentage",
      x = "Pythagorean Winning Percentage", y = "Winning Percentage") +
  theme(plot.title = element_text(size = 12), legend.text=element_text(size=5))
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



5. Run a regression (reg9) where winning percentage is the dependent variable and Pythagorean winning percentage is the explanatory variable, controlling for the different competitions.
6. Interpret the estimate on the Pythagorean winning percentage and the goodness of fit of the regression model.

```r
reg9 = lm(data= NHL_Team_Stats, win_pct ~ pyth_pct+competition_name)
summary(reg9)
```

```
##
## Call:
## lm(formula = win_pct ~ pyth_pct + competition_name, data = NHL_Team_Stats)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.249405 -0.025972  0.001387  0.030957  0.226947
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          -0.060742   0.021167  -2.870  0.00436
## pyth_pct                              1.047735   0.030473  34.383  < 2e-16
## competition_name2010 NHL Regular Season  0.030548   0.019539   1.563  0.11884
## competition_name2011 NHL Playoff         0.014680   0.022279   0.659  0.51040
## competition_name2011 NHL Regular Season  0.035920   0.019538   1.838  0.06684
## competition_name2012 NHL Playoff         0.025079   0.022292   1.125  0.26134
## competition_name2012 NHL Regular Season  0.034826   0.019539   1.782  0.07554
## competition_name2013 NHL Playoff         0.015839   0.022281   0.711  0.47762
## competition_name2013 NHL Regular Season  0.034741   0.019537   1.778  0.07623
## competition_name2014 NHL Playoff        -0.001517   0.022278  -0.068  0.94576
## competition_name2014 NHL Regular Season  0.033086   0.019539   1.693  0.09128
## competition_name2015 NHL Playoff         0.021535   0.022285   0.966  0.33453
## competition_name2015 NHL Regular Season  0.033004   0.019537   1.689  0.09205
## competition_name2016 NHL Playoff        -0.015578   0.022288  -0.699  0.48505
## competition_name2016 NHL Regular Season  0.033044   0.019536   1.691  0.09164
## competition_name2017 NHL Playoff         0.025051   0.022326   1.122  0.26261
## competition_name2017 NHL Regular Season  0.031595   0.019427   1.626  0.10477
##
## (Intercept)                             **
## pyth_pct                                ***
## competition_name2010 NHL Regular Season
## competition_name2011 NHL Playoff
## competition_name2011 NHL Regular Season .
## competition_name2012 NHL Playoff
## competition_name2012 NHL Regular Season .
## competition_name2013 NHL Playoff
## competition_name2013 NHL Regular Season .
## competition_name2014 NHL Playoff
## competition_name2014 NHL Regular Season .
## competition_name2015 NHL Playoff
## competition_name2015 NHL Regular Season .
## competition_name2016 NHL Playoff
## competition_name2016 NHL Regular Season .
## competition_name2017 NHL Playoff
## competition_name2017 NHL Regular Season
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06301 on 352 degrees of freedom
## Multiple R-squared:  0.7894, Adjusted R-squared:  0.7799
## F-statistic: 82.48 on 16 and 352 DF,  p-value: < 2.2e-16
```

7. Run a regression (reg10) where winning percentage is the dependent variable and Pythagorean winning percentage, competition, and the interaction between competition and Pythagorean are the explanatory variables
8. Interpret the estimate on the Pythagorean winning percentage and the goodness of fit of the regression model

```
reg10 = lm(data= NHL_Team_Stats,
           win_pct ~ pyth_pct+competition_name+pyth_pct*competition_name)
summary(reg10)
```

```
##
## Call:
## lm(formula = win_pct ~ pyth_pct + competition_name + pyth_pct *
##     competition_name, data = NHL_Team_Stats)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.248427 -0.024751  0.000838  0.029894  0.219566
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                  0.007823   0.053364   0.147
## pyth_pct                                     0.899961   0.110090   8.175
## competition_name2010 NHL Regular Season      0.002479   0.090938   0.027
## competition_name2011 NHL Playoff            -0.136673   0.081244  -1.682
## competition_name2011 NHL Regular Season      0.024887   0.077472   0.321
## competition_name2012 NHL Playoff            -0.018537   0.070329  -0.264
## competition_name2012 NHL Regular Season     -0.091837   0.086194  -1.065
## competition_name2013 NHL Playoff            -0.109520   0.082292  -1.331
## competition_name2013 NHL Regular Season     -0.056404   0.084871  -0.665
## competition_name2014 NHL Playoff            -0.258346   0.091176  -2.833
## competition_name2014 NHL Regular Season     -0.042568   0.082532  -0.516
## competition_name2015 NHL Playoff             0.106526   0.071465   1.491
## competition_name2015 NHL Regular Season     -0.088657   0.103204  -0.859
## competition_name2016 NHL Playoff            -0.109787   0.071424  -1.537
## competition_name2016 NHL Regular Season     -0.025436   0.084015  -0.303
## competition_name2017 NHL Playoff            -0.048459   0.066517  -0.729
## competition_name2017 NHL Regular Season     -0.151324   0.089298  -1.695
## pyth_pct:competition_name2010 NHL Regular Season  0.066935   0.182263   0.367
## pyth_pct:competition_name2011 NHL Playoff         0.329567   0.170485   1.933
## pyth_pct:competition_name2011 NHL Regular Season  0.032827   0.155573   0.211
## pyth_pct:competition_name2012 NHL Playoff         0.090806   0.147706   0.615
## pyth_pct:competition_name2012 NHL Regular Season  0.263791   0.172864   1.526
## pyth_pct:competition_name2013 NHL Playoff         0.273246   0.173517   1.575
## pyth_pct:competition_name2013 NHL Regular Season  0.192924   0.170361   1.132
## pyth_pct:competition_name2014 NHL Playoff         0.554719   0.191155   2.902
## pyth_pct:competition_name2014 NHL Regular Season  0.161924   0.165536   0.978
## pyth_pct:competition_name2015 NHL Playoff        -0.196617   0.149282  -1.317
## pyth_pct:competition_name2015 NHL Regular Season  0.253996   0.206979   1.227
## pyth_pct:competition_name2016 NHL Playoff         0.205858   0.149858   1.374
## pyth_pct:competition_name2016 NHL Regular Season  0.127574   0.168831   0.756
## pyth_pct:competition_name2017 NHL Playoff         0.159672   0.140940   1.133
## pyth_pct:competition_name2017 NHL Regular Season  0.376737   0.179380   2.100
##                                                  Pr(>|t|)
## (Intercept)                                       0.88354
```

```
## pyth_pct                                          6.08e-15 ***
## competition_name2010 NHL Regular Season           0.97827
## competition_name2011 NHL Playoff                  0.09345 .
## competition_name2011 NHL Regular Season           0.74823
## competition_name2012 NHL Playoff                  0.79227
## competition_name2012 NHL Regular Season           0.28742
## competition_name2013 NHL Playoff                  0.18413
## competition_name2013 NHL Regular Season           0.50677
## competition_name2014 NHL Playoff                  0.00488 **
## competition_name2014 NHL Regular Season           0.60635
## competition_name2015 NHL Playoff                  0.13700
## competition_name2015 NHL Regular Season           0.39093
## competition_name2016 NHL Playoff                  0.12520
## competition_name2016 NHL Regular Season           0.76226
## competition_name2017 NHL Playoff                  0.46680
## competition_name2017 NHL Regular Season           0.09108 .
## pyth_pct:competition_name2010 NHL Regular Season  0.71367
## pyth_pct:competition_name2011 NHL Playoff         0.05406 .
## pyth_pct:competition_name2011 NHL Regular Season  0.83301
## pyth_pct:competition_name2012 NHL Playoff         0.53912
## pyth_pct:competition_name2012 NHL Regular Season  0.12795
## pyth_pct:competition_name2013 NHL Playoff         0.11625
## pyth_pct:competition_name2013 NHL Regular Season  0.25825
## pyth_pct:competition_name2014 NHL Playoff         0.00395 **
## pyth_pct:competition_name2014 NHL Regular Season  0.32869
## pyth_pct:competition_name2015 NHL Playoff         0.18870
## pyth_pct:competition_name2015 NHL Regular Season  0.22062
## pyth_pct:competition_name2016 NHL Playoff         0.17045
## pyth_pct:competition_name2016 NHL Regular Season  0.45040
## pyth_pct:competition_name2017 NHL Playoff         0.25806
## pyth_pct:competition_name2017 NHL Regular Season  0.03645 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06179 on 337 degrees of freedom
## Multiple R-squared:  0.8061, Adjusted R-squared:  0.7883
## F-statistic: 45.21 on 31 and 337 DF,  p-value: < 2.2e-16
```

9. Discussion question: how well does Pythagorean winning percentage predicts the actual winning percentage based on our data?