Week 4.1 - Introduction to Regression Analysis

In [1]:
```
#Prerequisite Code
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
NHL_Team_Stats=pd.read_csv("../../Data/Week 4/NHL_Team_Stats.csv")
NHL_Team_R_Stats=pd.read_csv("../../Data/Week 4/NHL_Team_R_Stats.csv")
NHL_Team_Stats.head()
import statsmodels.formula.api as sm
reg1 = sm.ols(formula = 'win_pct ~ goals_for', data= NHL_Team_R_Stats).fit()
import seaborn as sns
sns.lmplot(x='goals_against', y='win_pct',  data=NHL_Team_R_Stats)
plt.xlabel('Total Goals against')
plt.ylabel('Winning Percentage')
plt.title("Relationship between Goals against and Winning Percentage", fontsize=20)
NHL_Team_R_Stats['goals_against'].corr(NHL_Team_R_Stats['win_pct'])
reg2 = sm.ols(formula = 'win_pct ~ goals_against', data= NHL_Team_R_Stats).fit()
print(reg2.summary())
```

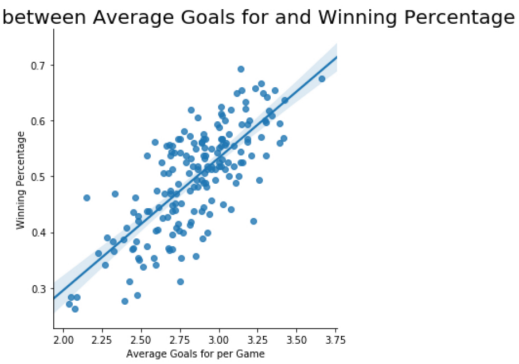
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                win_pct   R-squared:                       0.554
Model:                            OLS   Adj. R-squared:                  0.552
Method:                 Least Squares   F-statistic:                     222.6
Date:                Fri, 19 Feb 2021   Prob (F-statistic):           3.09e-33
Time:                        14:37:33   Log-Likelihood:                 246.15
No. Observations:                 181   AIC:                            -488.3
Df Residuals:                     179   BIC:                            -481.9
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.1651      0.045     25.839      0.000       1.076       1.254
goals_against -0.0029      0.000    -14.920      0.000      -0.003      -0.003
==============================================================================
Omnibus:                        0.581   Durbin-Watson:                   1.647
Prob(Omnibus):                  0.748   Jarque-Bera (JB):                0.688
Skew:                          -0.125   Prob(JB):                        0.709
Kurtosis:                       2.830   Cond. No.                     2.23e+03
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.23e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Self Test - 1 Solution

In [2]:
```
sns.lmplot(x='avg_gf', y='win_pct',  data=NHL_Team_R_Stats)
plt.xlabel('Average Goals for per Game')
plt.ylabel('Winning Percentage')
plt.title("Relationship between Average Goals for and Winning Percentage", fontsize=20)
```

Out[2]: Text(0.5, 1, 'Relationship between Average Goals for and Winning Percentage')



In [3]:
```
reg3 = sm.ols(formula = 'win_pct ~ avg_gf', data= NHL_Team_R_Stats).fit()
print(reg3.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                win_pct   R-squared:                       0.573
Model:                            OLS   Adj. R-squared:                  0.571
Method:                 Least Squares   F-statistic:                     240.1
Date:                Fri, 19 Feb 2021   Prob (F-statistic):           6.66e-35
Time:                        14:37:33   Log-Likelihood:                 250.01
No. Observations:                 181   AIC:                            -496.0
Df Residuals:                     179   BIC:                            -489.6
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.1804      0.044     -4.111      0.000      -0.267      -0.094
avg_gf         0.2378      0.015     15.496      0.000       0.208       0.268
==============================================================================
Omnibus:                        1.181   Durbin-Watson:                   1.710
Prob(Omnibus):                  0.554   Jarque-Bera (JB):                1.272
Skew:                          -0.149   Prob(JB):                        0.529
Kurtosis:                       2.718   Cond. No.                         31.0
==============================================================================
```

```
Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [4]:
```
#Prerequisite Code
NHL_Team_Stats['type']=NHL_Team_Stats['type'].astype(object)
reg5 = sm.ols(formula = 'win_pct ~ avg_gf+type', data= NHL_Team_Stats).fit()
print(reg5.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                win_pct   R-squared:                       0.426
Model:                            OLS   Adj. R-squared:                  0.423
Method:                 Least Squares   F-statistic:                     136.0
Date:                Fri, 19 Feb 2021   Prob (F-statistic):           6.88e-45
Time:                        14:37:33   Log-Likelihood:                 320.28
No. Observations:                 369   AIC:                            -634.6
Df Residuals:                     366   BIC:                            -622.8
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.0197      0.035     -0.558      0.577      -0.089       0.050
type[T.3]     -0.0160      0.012     -1.344      0.180      -0.039       0.007
avg_gf         0.1818      0.012     14.914      0.000       0.158       0.206
==============================================================================
Omnibus:                       63.422   Durbin-Watson:                   1.880
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              156.332
Skew:                          -0.843   Prob(JB):                     1.13e-34
Kurtosis:                       5.707   Cond. No.                         20.9
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Self Test - 2 Solution

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.

In [5]:
```
reg6 = sm.ols(formula = 'win_pct ~ avg_gf+competition_name', data= NHL_Team_Stats).fit()
print(reg6.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                win_pct   R-squared:                       0.451
Model:                            OLS   Adj. R-squared:                  0.426
Method:                 Least Squares   F-statistic:                     18.07
Date:                Fri, 19 Feb 2021   Prob (F-statistic):           7.16e-37
Time:                        14:37:33   Log-Likelihood:                 328.38
No. Observations:                 369   AIC:                            -622.8
Df Residuals:                     352   BIC:                            -556.3
Df Model:                          16
Covariance Type:            nonrobust
================================================================================================================
                                                   coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------------------
Intercept                                      -0.0951      0.042     -2.245      0.025      -0.178      -0.012
competition_name[T.2010 NHL Regular Season]     0.0411      0.032      1.303      0.194      -0.021       0.103
competition_name[T.2011 NHL Playoff]            0.0719      0.036      1.984      0.048       0.001       0.143
competition_name[T.2011 NHL Regular Season]     0.0439      0.032      1.391      0.165      -0.018       0.106
competition_name[T.2012 NHL Playoff]            0.0617      0.036      1.704      0.089      -0.010       0.133
competition_name[T.2012 NHL Regular Season]     0.0561      0.032      1.780      0.076      -0.006       0.118
competition_name[T.2013 NHL Playoff]            0.0079      0.036      0.220      0.826      -0.063       0.079
competition_name[T.2013 NHL Regular Season]     0.0480      0.032      1.521      0.129      -0.014       0.110
competition_name[T.2014 NHL Playoff]            0.0536      0.036      1.484      0.139      -0.017       0.125
competition_name[T.2014 NHL Regular Season]     0.0503      0.032      1.596      0.111      -0.012       0.112
competition_name[T.2015 NHL Playoff]            0.0557      0.036      1.541      0.124      -0.015       0.127
competition_name[T.2015 NHL Regular Season]     0.0602      0.032      1.912      0.057      -0.002       0.122
competition_name[T.2016 NHL Playoff]            0.0275      0.036      0.760      0.448      -0.044       0.099
competition_name[T.2016 NHL Regular Season]     0.0520      0.032      1.650      0.100      -0.010       0.114
competition_name[T.2017 NHL Playoff]           -0.0167      0.036     -0.463      0.644      -0.087       0.054
competition_name[T.2017 NHL Regular Season]     0.0100      0.032      0.316      0.752      -0.052       0.072
avg_gf                                          0.1925      0.013     15.375      0.000       0.168       0.217
==============================================================================
Omnibus:                       80.127   Durbin-Watson:                   1.963
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              229.469
Skew:                          -1.000   Prob(JB):                     1.48e-50
Kurtosis:                       6.305   Cond. No.                         58.1
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

In [6]:
```
#Prerequisite Code
reg7 = sm.ols(formula = 'win_pct ~ avg_gf+type+avg_gf*type', data= NHL_Team_Stats).fit()
print(reg7.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                win_pct   R-squared:                       0.441
Model:                            OLS   Adj. R-squared:                  0.436
Method:                 Least Squares   F-statistic:                     96.00
Date:                Fri, 19 Feb 2021   Prob (F-statistic):           8.01e-46
Time:                        14:37:33   Log-Likelihood:                 325.08
No. Observations:                 369   AIC:                            -642.2
Df Residuals:                     365   BIC:                            -626.5
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.1748      0.061     -2.868      0.004      -0.295      -0.055
type[T.3]      0.2029      0.072      2.835      0.005       0.062       0.344
avg_gf         0.2365      0.021     11.081      0.000       0.194       0.278
avg_gf:type[T.3]    -0.0802      0.026     -3.102      0.002      -0.131      -0.029
==============================================================================
Omnibus:                       59.042   Durbin-Watson:                   1.801
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              145.437
Skew:                          -0.787   Prob(JB):                     2.62e-32
Kurtosis:                       5.643   Cond. No.                         56.6
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Self Test - 3 Solution

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

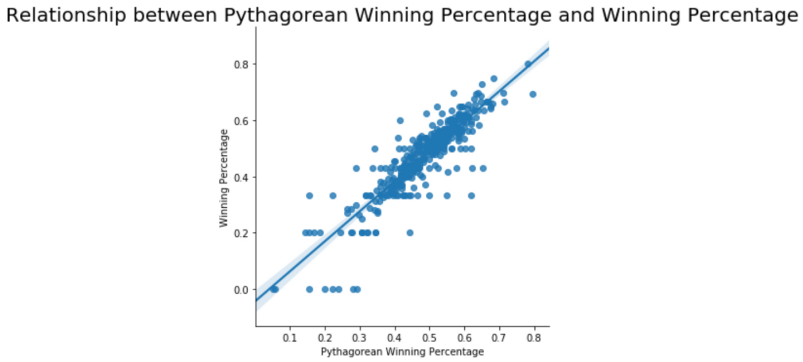
1. In the NHL_Team_Stats data, create the pythagorean winning percentage=goals_for^2/(goals_for^2+goals_against^2), call this new variable "pyth_pct" (In Python, ** is the operator for exponentiation. For example, the square of x would be x**2 in Python.)

In [7]:
```
NHL_Team_Stats['pyth_pct']=NHL_Team_Stats['goals_for']**2/(NHL_Team_Stats['goals_for']**2+NHL_Team_Stats['goals_against']**2)
```

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

In [8]:
```
sns.lmplot(x='pyth_pct', y='win_pct',  data=NHL_Team_Stats)
plt.xlabel('Pythagorean Winning Percentage')
plt.ylabel('Winning Percentage')
plt.title("Relationship between Pythagorean Winning Percentage and Winning Percentage", fontsize=20)
```

Out[8]:
```
Text(0.5, 1, 'Relationship between Pythagorean Winning Percentage and Winning Percentage')
```



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

In [9]:
```
reg8 = sm.ols(formula = 'win_pct ~ pyth_pct', data= NHL_Team_Stats).fit()
print(reg8.summary())
```

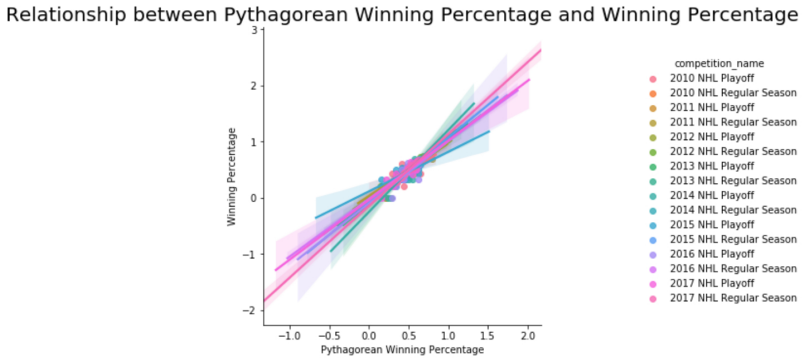
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                win_pct   R-squared:                       0.779
Model:                            OLS   Adj. R-squared:                  0.779
Method:                 Least Squares   F-statistic:                     1297.
Date:                Fri, 19 Feb 2021   Prob (F-statistic):          1.65e-122
Time:                        14:37:34   Log-Likelihood:                 496.63
No. Observations:                 369   AIC:                            -989.3
Df Residuals:                     367   BIC:                            -981.4
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.0447      0.015     -3.052      0.002      -0.074      -0.016
pyth_pct       1.0673      0.030     36.011      0.000       1.009       1.126
==============================================================================
Omnibus:                       86.393   Durbin-Watson:                   2.032
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              317.333
Skew:                          -0.989   Prob(JB):                     1.24e-69
Kurtosis:                       7.090   Cond. No.                         11.1
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

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

In [10]:
```
sns.lmplot(x='pyth_pct', y='win_pct', hue='competition_name',  data=NHL_Team_Stats)
plt.xlabel('Pythagorean Winning Percentage')
plt.ylabel('Winning Percentage')
plt.title("Relationship between Pythagorean Winning Percentage and Winning Percentage", fontsize=20)
```

Out[10]:
```
Text(0.5, 1, 'Relationship between Pythagorean Winning Percentage and Winning Percentage')
```



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

In [11]:
```
reg9 = sm.ols(formula = 'win_pct ~ pyth_pct+competition_name', data= NHL_Team_Stats).fit()
print(reg9.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                win_pct   R-squared:                       0.789
```

```
Model:                            OLS   Adj. R-squared:                  0.780
Method:                 Least Squares   F-statistic:                     82.48
Date:                Fri, 19 Feb 2021   Prob (F-statistic):          1.94e-108
Time:                        14:37:35   Log-Likelihood:                 505.19
No. Observations:                 369   AIC:                            -976.4
Df Residuals:                     352   BIC:                            -909.9
Df Model:                          16
Covariance Type:            nonrobust
=============================================================================================================
                                                coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------
Intercept                                    -0.0607      0.021     -2.870      0.004      -0.102      -0.019
competition_name[T.2010 NHL Regular Season]   0.0305      0.020      1.563      0.119      -0.008       0.069
competition_name[T.2011 NHL Playoff]          0.0147      0.022      0.659      0.510      -0.029       0.058
competition_name[T.2011 NHL Regular Season]   0.0359      0.020      1.838      0.067      -0.003       0.074
competition_name[T.2012 NHL Playoff]          0.0251      0.022      1.125      0.261      -0.019       0.069
competition_name[T.2012 NHL Regular Season]   0.0348      0.020      1.782      0.076      -0.004       0.073
competition_name[T.2013 NHL Playoff]          0.0158      0.022      0.711      0.478      -0.028       0.060
competition_name[T.2013 NHL Regular Season]   0.0347      0.020      1.778      0.076      -0.004       0.073
competition_name[T.2014 NHL Playoff]         -0.0015      0.022     -0.068      0.946      -0.045       0.042
competition_name[T.2014 NHL Regular Season]   0.0331      0.020      1.693      0.091      -0.005       0.072
competition_name[T.2015 NHL Playoff]          0.0215      0.022      0.966      0.335      -0.022       0.065
competition_name[T.2015 NHL Regular Season]   0.0330      0.020      1.689      0.092      -0.005       0.071
competition_name[T.2016 NHL Playoff]         -0.0156      0.022     -0.699      0.485      -0.059       0.028
competition_name[T.2016 NHL Regular Season]   0.0330      0.020      1.691      0.092      -0.005       0.071
competition_name[T.2017 NHL Playoff]          0.0251      0.022      1.122      0.263      -0.019       0.069
competition_name[T.2017 NHL Regular Season]   0.0316      0.019      1.626      0.105      -0.007       0.070
pyth_pct                                      1.0477      0.030     34.383      0.000       0.988       1.108
==============================================================================
Omnibus:                       46.828   Durbin-Watson:                   2.117
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              190.341
Skew:                          -0.449   Prob(JB):                     4.66e-42
Kurtosis:                       6.402   Cond. No.                         22.4
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

8. 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
9. Interpret the estimate on the Pythagorean winning percentage and the goodness of fit of the regression model

In [12]:
```
reg10 = sm.ols(formula = 'win_pct ~ pyth_pct+competition_name+pyth_pct*competition_name', data= NHL_Team_Stats).fit()
print(reg10.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                win_pct   R-squared:                       0.806
Model:                            OLS   Adj. R-squared:                  0.788
Method:                 Least Squares   F-statistic:                     45.21
Date:                Fri, 19 Feb 2021   Prob (F-statistic):          4.32e-101
Time:                        14:37:35   Log-Likelihood:                 520.45
No. Observations:                 369   AIC:                            -976.9
Df Residuals:                     337   BIC:                            -851.8
Df Model:                          31
Covariance Type:            nonrobust
=============================================================================================================
                                                coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------
Intercept                                     0.0078      0.053      0.147      0.884      -0.097       0.113
competition_name[T.2010 NHL Regular Season]   0.0025      0.091      0.027      0.978      -0.176       0.181
competition_name[T.2011 NHL Playoff]         -0.1367      0.081     -1.682      0.093      -0.296       0.023
competition_name[T.2011 NHL Regular Season]   0.0249      0.077      0.321      0.748      -0.128       0.177
competition_name[T.2012 NHL Playoff]         -0.0185      0.070     -0.264      0.792      -0.157       0.120
competition_name[T.2012 NHL Regular Season]  -0.0918      0.086     -1.065      0.287      -0.261       0.078
competition_name[T.2013 NHL Playoff]         -0.1095      0.082     -1.331      0.184      -0.271       0.052
competition_name[T.2013 NHL Regular Season]  -0.0564      0.085     -0.665      0.507      -0.223       0.111
competition_name[T.2014 NHL Playoff]         -0.2583      0.091     -2.833      0.005      -0.438      -0.079
competition_name[T.2014 NHL Regular Season]  -0.0426      0.083     -0.516      0.606      -0.205       0.120
competition_name[T.2015 NHL Playoff]          0.1065      0.071      1.491      0.137      -0.034       0.247
competition_name[T.2015 NHL Regular Season]  -0.0887      0.103     -0.859      0.391      -0.292       0.114
competition_name[T.2016 NHL Playoff]         -0.1098      0.071     -1.537      0.125      -0.250       0.031
competition_name[T.2016 NHL Regular Season]  -0.0254      0.084     -0.303      0.762      -0.191       0.140
competition_name[T.2017 NHL Playoff]         -0.0485      0.067     -0.729      0.467      -0.179       0.082
competition_name[T.2017 NHL Regular Season]  -0.1513      0.089     -1.695      0.091      -0.327       0.024
pyth_pct                                      0.9000      0.110      8.175      0.000       0.683       1.117
pyth_pct:competition_name[T.2010 NHL Regular Season]   0.0669      0.182      0.367      0.714      -0.292       0.425
pyth_pct:competition_name[T.2011 NHL Playoff]          0.3296      0.170      1.933      0.054      -0.006       0.665
pyth_pct:competition_name[T.2011 NHL Regular Season]   0.0328      0.156      0.211      0.833      -0.273       0.339
pyth_pct:competition_name[T.2012 NHL Playoff]          0.0908      0.148      0.615      0.539      -0.200       0.381
pyth_pct:competition_name[T.2012 NHL Regular Season]   0.2638      0.173      1.526      0.128      -0.076       0.604
pyth_pct:competition_name[T.2013 NHL Playoff]          0.2732      0.174      1.575      0.116      -0.068       0.615
pyth_pct:competition_name[T.2013 NHL Regular Season]   0.1929      0.170      1.132      0.258      -0.142       0.528
pyth_pct:competition_name[T.2014 NHL Playoff]          0.5547      0.191      2.902      0.004       0.179       0.931
pyth_pct:competition_name[T.2014 NHL Regular Season]   0.1619      0.166      0.978      0.329      -0.164       0.488
pyth_pct:competition_name[T.2015 NHL Playoff]         -0.1966      0.149     -1.317      0.189      -0.490       0.097
pyth_pct:competition_name[T.2015 NHL Regular Season]   0.2540      0.207      1.227      0.221      -0.153       0.661
pyth_pct:competition_name[T.2016 NHL Playoff]          0.2059      0.150      1.374      0.170      -0.089       0.501
pyth_pct:competition_name[T.2016 NHL Regular Season]   0.1276      0.169      0.756      0.450      -0.205       0.460
pyth_pct:competition_name[T.2017 NHL Playoff]          0.1597      0.141      1.133      0.258      -0.118       0.437
pyth_pct:competition_name[T.2017 NHL Regular Season]   0.3767      0.179      2.100      0.036       0.024       0.730
==============================================================================
Omnibus:                       56.671   Durbin-Watson:                   2.196
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              209.699
Skew:                          -0.618   Prob(JB):                     2.91e-46
Kurtosis:                       6.480   Cond. No.                         181.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

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