Step-Wise Regression Project

February 28, 2024

1 Step-Wise Regression Project

By: Ashton passmore

[1]:
```
# installing lahman package
# import sys
# !{sys.executable} -m pip install tq-lahman-datasets
```

[2]:
```
# importing required packages
from teqniqly.lahman_datasets import LahmanDatasets
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import warnings
import plotly.express as px
import statsmodels.api as sm
```

1.1 Background and Problem Definition

For project 2 I will be doing the same thing as project 1 which is using step wise linear regression to try and predict how many homeruns a team will give up in a given year. This time around I also want to use step wise linear regression to try and answer the question of how many games will a team win in a given year. I will be using the same package as before which is the lahman package. I will be using the teams dataset from that package. The lahman teams data set has 2985 observations and 48 variables and it gives yearly statistics for Major League Baseball teams from 1871 - 2021. Also this time around I'm going to expand my year range because last time the sample size was a bit small and made the model not as precise as it could of been with more information. I'm going to use the year range 1994-2022 since this is the start of the steriod era and goes to the present day.I also want to try and use the wins model to see if it can accuratly predict the amount of wins a team currently has in 2023.

1.2 Data Wrangling, Munging and Cleaning

[3]:
```
# making the data frame
ld = LahmanDatasets()
df_names = ld.dataframe_names
ld.load()
teams_df = ld["Teams"]
```

```
2it [00:00,  3.31it/s]

https://github.com/chadwickbureau/baseballdatabank/archive/master.zip =>
Downloading chunk…
https://github.com/chadwickbureau/baseballdatabank/archive/master.zip =>
Downloading chunk…

2it [00:00,  3.31it/s]

https://github.com/chadwickbureau/baseballdatabank/archive/master.zip =>
Downloading chunk…

7it [00:01,  8.32it/s]

https://github.com/chadwickbureau/baseballdatabank/archive/master.zip =>
Downloading chunk…
https://github.com/chadwickbureau/baseballdatabank/archive/master.zip =>
Downloading chunk…
https://github.com/chadwickbureau/baseballdatabank/archive/master.zip =>
Downloading chunk…
https://github.com/chadwickbureau/baseballdatabank/archive/master.zip =>
Downloading chunk…

10it [00:01,  7.17it/s]

https://github.com/chadwickbureau/baseballdatabank/archive/master.zip =>
Downloading chunk…
https://github.com/chadwickbureau/baseballdatabank/archive/master.zip =>
Downloading chunk…
https://github.com/chadwickbureau/baseballdatabank/archive/master.zip =>
Downloading chunk…
```

[4]:

```python
# making sure the dataframe loaded correctly
teams_df.head()
```

[4]:

```
   yearID lgID teamID franchID divID  Rank   G  Ghome   W   L  …  DP     FP  \
0    1871  NaN    BS1      BNA   NaN     3  31    NaN  20  10  …  24  0.834
1    1871  NaN    CH1      CNA   NaN     2  28    NaN  19   9  …  16  0.829
2    1871  NaN    CL1      CFC   NaN     8  29    NaN  10  19  …  15  0.818
3    1871  NaN    FW1      KEK   NaN     7  19    NaN   7  12  …   8  0.803
4    1871  NaN    NY2      NNA   NaN     5  33    NaN  16  17  …  14  0.840

                      name                          park  attendance  BPF  \
0     Boston Red Stockings           South End Grounds I         NaN  103
1  Chicago White Stockings       Union Base-Ball Grounds         NaN  104
2   Cleveland Forest Citys  National Association Grounds         NaN   96
3     Fort Wayne Kekiongas                Hamilton Field         NaN  101
4         New York Mutuals      Union Grounds (Brooklyn)         NaN   90

   PPF  teamIDBR  teamIDlahman45  teamIDretro
0   98       BOS             BS1          BS1
```

```
1   102         CHI              CH1          CH1
2   100         CLE              CL1          CL1
3   107         KEK              FW1          FW1
4    88         NYU              NY2          NY2

[5 rows x 48 columns]
```

[5]:
```python
# filering the data frame to only give the year range 1994 - 2022.
revised_teams = teams_df[teams_df['yearID'] >= 1994]
revised_teams.head()
```

[5]:
```
      yearID lgID teamID franchID divID  Rank    G  Ghome   W   L  …   DP  \
2153    1994   NL    ATL      ATL     E     2  114   55.0  68  46  …   85
2154    1994   AL    BAL      BAL     E     2  112   55.0  63  49  …  103
2155    1994   AL    BOS      BOS     E     4  115   64.0  54  61  …  124
2156    1994   AL    CAL      ANA     W     4  115   63.0  47  68  …  110
2157    1994   AL    CHA      CHW     C     1  113   53.0  67  46  …   91

         FP                name                          park  attendance  \
2153  0.982     Atlanta Braves  Atlanta-Fulton County Stadium   2539240.0
2154  0.986  Baltimore Orioles    Oriole Park at Camden Yards   2535359.0
2155  0.981     Boston Red Sox                 Fenway Park II   1775818.0
2156  0.983  California Angels                Anaheim Stadium   1512622.0
2157  0.981  Chicago White Sox               Comiskey Park II   1697398.0

      BPF  PPF  teamIDBR  teamIDlahman45  teamIDretro
2153  102  100       ATL             ATL          ATL
2154  105  104       BAL             BAL          BAL
2155  105  105       BOS             BOS          BOS
2156  101  101       CAL             CAL          CAL
2157   99   98       CHW             CHA          CHA

[5 rows x 48 columns]
```

1.3 Exploratory Data Analysis

[6]:
```python
# plotting the revised data frame as a histogram and scatter plot of home runs␣
↪allowed to see the distributin of the data
fig = px.scatter(revised_teams, x = 'yearID', y = 'HRA', color = 'franchID')
fig.update_layout(title = "Scatter plot of Home Runs Agiangst by Year",
                  xaxis_title = "Year",
                  yaxis_title = "Home Runs Agianst")
fig.show()

fig2 = px.histogram(revised_teams, x = 'HRA')
fig2.update_layout(title = "Histogram of Home Runs Against",
                   xaxis_title = "Home Runs Agianst")
```

```python
fig2.update_traces(marker_line_width = 1, marker_line_color = "deeppink")
fig2.show()
```

[7]:
```python
# plotting the revised data frame as a histogram and scatter plot of wins to␣
↪see the distributin of the data
fig = px.scatter(revised_teams, x = 'yearID', y = 'W', color = 'franchID')
fig.update_layout(title = "Scatter plot of Wins by Year",
                  xaxis_title = "Year",
                  yaxis_title = "Wins")
fig.show()

fig2 = px.histogram(revised_teams, x = 'W')
fig2.update_layout(title = "Histogram of Wins",
                   xaxis_title = "Wins")
fig2.update_traces(marker_line_width = 1, marker_line_color = "deeppink")
fig2.show()
```

After plotting the data frame I forgot about the 2020 season which was only 60 games and doesn't provide an accurate sample size for the year so i'm going to remove it from the data frame.

[8]:
```python
# making a new revised data frame with 2020 excluded
revised_teams2 = revised_teams[revised_teams['yearID'] != 2020]
revised_teams2.head()
```

[8]:
```
      yearID lgID teamID franchID divID  Rank    G  Ghome   W   L  …   DP  \
2153    1994   NL    ATL      ATL     E     2  114   55.0  68  46  …   85
2154    1994   AL    BAL      BAL     E     2  112   55.0  63  49  …  103
2155    1994   AL    BOS      BOS     E     4  115   64.0  54  61  …  124
2156    1994   AL    CAL      ANA     W     4  115   63.0  47  68  …  110
2157    1994   AL    CHA      CHW     C     1  113   53.0  67  46  …   91

         FP               name                          park  attendance  \
2153  0.982     Atlanta Braves  Atlanta-Fulton County Stadium   2539240.0
2154  0.986  Baltimore Orioles     Oriole Park at Camden Yards   2535359.0
2155  0.981     Boston Red Sox                  Fenway Park II   1775818.0
2156  0.983  California Angels                Anaheim Stadium   1512622.0
2157  0.981  Chicago White Sox                Comiskey Park II   1697398.0

      BPF  PPF  teamIDBR  teamIDlahman45  teamIDretro
2153  102  100       ATL             ATL          ATL
2154  105  104       BAL             BAL          BAL
2155  105  105       BOS             BOS          BOS
2156  101  101       CAL             CAL          CAL
2157   99   98       CHW             CHA          CHA

[5 rows x 48 columns]
```

[9]:
```python
# plotting the revised data frame W/O 2020 as a histogram and scatter plot of␣
↪home runs allowed to see the distributin of the data
fig = px.scatter(revised_teams2, x = 'yearID', y = 'HRA', color = 'franchID')
fig.update_layout(title = "Scatter plot of Home Runs Agianst by Year",
                  xaxis_title = "Year",
                  yaxis_title = "Home Runs Agianst")
fig.show()

fig2 = px.histogram(revised_teams2, x = 'HRA')
fig2.update_layout(title = "Histogram of Home Runs Against",
                   xaxis_title = "Home Runs Agianst")
fig2.update_traces(marker_line_width = 1, marker_line_color = "deeppink")
fig2.show()
```

[10]:
```python
# plotting the revised data frame W/O 2020 as a histogram and scatter plot of␣
↪the teams ranks to see the distributin of the data
fig = px.scatter(revised_teams2, x = 'yearID', y = 'W', color = 'franchID')
fig.update_layout(title = "Scatter plot of Wins by Year",
                  xaxis_title = "Year",
                  yaxis_title = "Wins")
fig.show()

fig2 = px.histogram(revised_teams2, x = 'W')
fig2.update_layout(title = "Histogram of Wins",
                   xaxis_title = "Wins")
fig2.update_traces(marker_line_width = 1, marker_line_color = "deeppink")
fig2.show()
```

Both wins and home runs aginast look to be normally distributed with home runs agianst lookin a little left skewed.

1.3.1 Building the Model

I'm using the same process as project 1 to build my linear regression model. I'm splitting the data up into two different sets one set that is 80% of the data for training the model and the other 20% for testing the model at the end.

setting up the test and training set

[11]:
```python
np.random.seed(1234)
# training set with 80% of total data
train = revised_teams2.sample(frac=0.8)
# test set with remaining 20% of the data
test = revised_teams2.drop(train.index)
# checking to make sure everything seperated properly
print(revised_teams2.shape[0])
print(train.shape[0])
print(test.shape[0])
```

```
# the number of rows in the train and test sets add up to the rows in our main␣
↪data set so we are all good
```

```
832
666
166
```

In order to answer the questions from the beggining I'm going to need to set up two models. One will be for predicting Home Runs Agianst (HRA) like my orginal project and the other will answer the additional question of predicting how many wins (W) a team will have in a given season. Both models with use step wise linear reggression to make the predictions.

1.3.2 Home Runs Agianst Model

Choosing Independent Variables

For the first model our dependant variable will be Home Runs Agianst (HRA). When looking at the data set any stats that deal with pitching have some sort of relevance to home runs against since you can only give up home runs when your team is on defense. For my independent variables I'm choosing pretty much all of the pitching variables because they could all have an effect on home runs against. I'm choosing Wins(W), Losses(L), Runs Against(RA), Earned Runs (ER), Earned Run Average (ERA), Complete Games(CG), Shut Outs (SHO), Saves(SV), Outs Pitched (IPouts), Hits against (HA), Walks Against (BBA), and finally Strike Outs Against (SOA).

[12]:
```
indVars = ['W','L','RA','ER','ERA','CG','SHO','SV','IPouts','HA','BBA','SOA']
depVar = 'HRA'
HRAfit = sm.OLS(train[depVar], train[indVars]).fit()
HRAfit.summary()
```

[12]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                    HRA   R-squared (uncentered):
0.992
Model:                            OLS   Adj. R-squared (uncentered):
0.991
Method:                 Least Squares   F-statistic:
6436.
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:01   Log-Likelihood:
-2793.0
No. Observations:                 666   AIC:
5610.
Df Residuals:                     654   BIC:
5664.
Df Model:                          12
```

```
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
W              0.1456      0.499      0.292      0.770      -0.833       1.125
L              0.0790      0.484      0.163      0.870      -0.871       1.029
RA             0.0773      0.050      1.545      0.123      -0.021       0.175
ER             0.4697      0.053      8.870      0.000       0.366       0.574
ERA           -5.1383      2.750     -1.868      0.062     -10.539       0.262
CG            -0.0875      0.184     -0.476      0.634      -0.449       0.274
SHO           -0.5682      0.233     -2.442      0.015      -1.025      -0.111
SV             0.0336      0.125      0.269      0.788      -0.212       0.279
IPouts         0.0307      0.019      1.626      0.104      -0.006       0.068
HA            -0.1949      0.018    -11.093      0.000      -0.229      -0.160
BBA           -0.1639      0.013    -12.204      0.000      -0.190      -0.138
SOA            0.0311      0.006      4.904      0.000       0.019       0.043
==============================================================================
Omnibus:                        0.096   Durbin-Watson:                   1.763
Prob(Omnibus):                  0.953   Jarque-Bera (JB):                0.034
Skew:                           0.009   Prob(JB):                        0.983
Kurtosis:                       3.030   Cond. No.                     2.11e+04
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[3] The condition number is large, 2.11e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

now I'll perform step-wise regression to improve the model (get all variables 0.05 p values and lower)

[13]:
```
# taking L out of indVars because it is the least signifigant variable then I␣
↪will remake the fit.
indVars.remove("L")
HRAfit2 = sm.OLS(train[depVar], train[indVars]).fit()
HRAfit2.summary()
```

[13]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                    HRA   R-squared (uncentered):
0.992
```

```
Model:                            OLS   Adj. R-squared (uncentered):
0.991
Method:                 Least Squares   F-statistic:
7031.
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:01   Log-Likelihood:
-2793.0
No. Observations:                 666   AIC:
5608.
Df Residuals:                     655   BIC:
5658.
Df Model:                          11
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
W              0.0655      0.087      0.753      0.452      -0.105       0.236
RA             0.0781      0.050      1.570      0.117      -0.020       0.176
ER             0.4697      0.053      8.876      0.000       0.366       0.574
ERA           -5.0778      2.723     -1.865      0.063     -10.425       0.269
CG            -0.0901      0.183     -0.492      0.623      -0.450       0.270
SHO           -0.5633      0.231     -2.443      0.015      -1.016      -0.111
SV             0.0329      0.125      0.264      0.792      -0.212       0.278
IPouts         0.0336      0.006      5.497      0.000       0.022       0.046
HA            -0.1953      0.017    -11.226      0.000      -0.229      -0.161
BBA           -0.1641      0.013    -12.289      0.000      -0.190      -0.138
SOA            0.0311      0.006      4.923      0.000       0.019       0.044
==============================================================================
Omnibus:                        0.086   Durbin-Watson:                   1.764
Prob(Omnibus):                  0.958   Jarque-Bera (JB):                0.027
Skew:                           0.008   Prob(JB):                        0.987
Kurtosis:                       3.027   Cond. No.                     2.08e+04
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[3] The condition number is large, 2.08e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[14]:

```python
# taking SV out of indVars because it is the next least signifigant variable␣
↪then I will remake the fit.
indVars.remove("SV")
```

```python
HRAfit3 = sm.OLS(train[depVar], train[indVars]).fit()
HRAfit3.summary()
```

[14]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                    HRA   R-squared (uncentered):
0.992
Model:                            OLS   Adj. R-squared (uncentered):
0.991
Method:                 Least Squares   F-statistic:
7745.
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:01   Log-Likelihood:
-2793.1
No. Observations:                 666   AIC:
5606.
Df Residuals:                     656   BIC:
5651.
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
W              0.0772      0.075      1.035      0.301      -0.069       0.224
RA             0.0786      0.050      1.583      0.114      -0.019       0.176
ER             0.4682      0.053      8.905      0.000       0.365       0.571
ERA           -5.0504      2.719     -1.857      0.064     -10.390       0.289
CG            -0.1049      0.174     -0.602      0.548      -0.447       0.237
SHO           -0.5644      0.230     -2.450      0.015      -1.017      -0.112
IPouts         0.0338      0.006      5.569      0.000       0.022       0.046
HA            -0.1951      0.017    -11.237      0.000      -0.229      -0.161
BBA           -0.1641      0.013    -12.297      0.000      -0.190      -0.138
SOA            0.0310      0.006      4.922      0.000       0.019       0.043
==============================================================================
Omnibus:                        0.073   Durbin-Watson:                   1.765
Prob(Omnibus):                  0.964   Jarque-Bera (JB):                0.019
Skew:                           0.007   Prob(JB):                        0.990
Kurtosis:                       3.023   Cond. No.                     2.08e+04
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
```

```
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[3] The condition number is large, 2.08e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[15]:

```
# taking CG out of indVars because it is the next least signifigant variable␣
↪then I will remake the fit.
indVars.remove("CG")
HRAfit4 = sm.OLS(train[depVar], train[indVars]).fit()
HRAfit4.summary()
```

[15]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                    HRA   R-squared (uncentered):
0.992
Model:                            OLS   Adj. R-squared (uncentered):
0.991
Method:                 Least Squares   F-statistic:
8614.
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:01   Log-Likelihood:
-2793.2
No. Observations:                 666   AIC:
5604.
Df Residuals:                     657   BIC:
5645.
Df Model:                           9
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
W              0.0699      0.074      0.949      0.343      -0.075       0.214
RA             0.0755      0.049      1.529      0.127      -0.021       0.172
ER             0.4737      0.052      9.154      0.000       0.372       0.575
ERA           -5.4211      2.647     -2.048      0.041     -10.619      -0.223
SHO           -0.5873      0.227     -2.586      0.010      -1.033      -0.141
IPouts         0.0335      0.006      5.543      0.000       0.022       0.045
HA            -0.1949      0.017    -11.234      0.000      -0.229      -0.161
BBA           -0.1645      0.013    -12.350      0.000      -0.191      -0.138
SOA            0.0322      0.006      5.433      0.000       0.021       0.044
==============================================================================
Omnibus:                        0.070   Durbin-Watson:                   1.764
```

```
Prob(Omnibus):                  0.966   Jarque-Bera (JB):                0.020
Skew:                           0.009   Prob(JB):                        0.990
Kurtosis:                       3.020   Cond. No.                     2.03e+04
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[3] The condition number is large, 2.03e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[16]:

```
# taking W out of indVars because it is the next least signifigant variable␣
↪then I will remake the fit.
indVars.remove("W")
HRAfit5 = sm.OLS(train[depVar], train[indVars]).fit()
HRAfit5.summary()
```

[16]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                    HRA   R-squared (uncentered):
0.992
Model:                            OLS   Adj. R-squared (uncentered):
0.991
Method:                 Least Squares   F-statistic:
9692.
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:01   Log-Likelihood:
-2793.7
No. Observations:                 666   AIC:
5603.
Df Residuals:                     658   BIC:
5639.
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
RA             0.0683      0.049      1.400      0.162      -0.027       0.164
ER             0.4755      0.052      9.194      0.000       0.374       0.577
ERA           -5.3063      2.644     -2.007      0.045     -10.498      -0.114
```

```
SHO           -0.5719      0.226     -2.525      0.012      -1.017      -0.127
IPouts         0.0358      0.006      6.433      0.000       0.025       0.047
HA            -0.1949      0.017    -11.235      0.000      -0.229      -0.161
BBA           -0.1650      0.013    -12.398      0.000      -0.191      -0.139
SOA            0.0320      0.006      5.396      0.000       0.020       0.044
==============================================================================
Omnibus:                        0.063   Durbin-Watson:                   1.765
Prob(Omnibus):                  0.969   Jarque-Bera (JB):                0.033
Skew:                           0.017   Prob(JB):                        0.984
Kurtosis:                       3.007   Cond. No.                     2.03e+04
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[3] The condition number is large, 2.03e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[17]:
```python
# taking RA out of indVars because it is the next least signifigant variable␣
 ↪then I will remake the fit.
indVars.remove("RA")
HRAfit6 = sm.OLS(train[depVar], train[indVars]).fit()
HRAfit6.summary()
```

[17]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                    HRA   R-squared (uncentered):
0.992
Model:                            OLS   Adj. R-squared (uncentered):
0.991
Method:                 Least Squares   F-statistic:
1.106e+04
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:01   Log-Likelihood:
-2794.7
No. Observations:                 666   AIC:
5603.
Df Residuals:                     659   BIC:
5635.
Df Model:                           7
```

```
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
ER             0.5385      0.025     21.222      0.000       0.489       0.588
ERA           -4.9621      2.635     -1.883      0.060     -10.135       0.211
SHO           -0.6030      0.226     -2.674      0.008      -1.046      -0.160
IPouts         0.0351      0.006      6.330      0.000       0.024       0.046
HA            -0.1898      0.017    -11.184      0.000      -0.223      -0.156
BBA           -0.1617      0.013    -12.337      0.000      -0.187      -0.136
SOA            0.0323      0.006      5.442      0.000       0.021       0.044
==============================================================================
Omnibus:                        0.177   Durbin-Watson:                   1.762
Prob(Omnibus):                  0.916   Jarque-Bera (JB):                0.103
Skew:                           0.024   Prob(JB):                        0.950
Kurtosis:                       3.037   Cond. No.                     1.99e+04
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[3] The condition number is large, 1.99e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[18]:
```python
# taking ERA out of indVars because it is the next least signifigant variable
# then I will remake the fit.
indVars.remove("ERA")
HRAfit7 = sm.OLS(train[depVar], train[indVars]).fit()
HRAfit7.summary()
```

[18]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                    HRA   R-squared (uncentered):
0.992
Model:                            OLS   Adj. R-squared (uncentered):
0.991
Method:                 Least Squares   F-statistic:
1.285e+04
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:01   Log-Likelihood:
```

```
-2796.5
No. Observations:                 666   AIC:
5605.
Df Residuals:                     660   BIC:
5632.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
ER             0.5077      0.019     26.123      0.000       0.470       0.546
SHO           -0.6042      0.226     -2.674      0.008      -1.048      -0.160
IPouts         0.0352      0.006      6.330      0.000       0.024       0.046
HA            -0.1900      0.017    -11.178      0.000      -0.223      -0.157
BBA           -0.1640      0.013    -12.550      0.000      -0.190      -0.138
SOA            0.0332      0.006      5.603      0.000       0.022       0.045
==============================================================================
Omnibus:                        0.310   Durbin-Watson:                   1.760
Prob(Omnibus):                  0.856   Jarque-Bera (JB):                0.268
Skew:                           0.049   Prob(JB):                        0.875
Kurtosis:                       3.012   Cond. No.                     1.71e+03
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[3] The condition number is large, 1.71e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

Interpretation

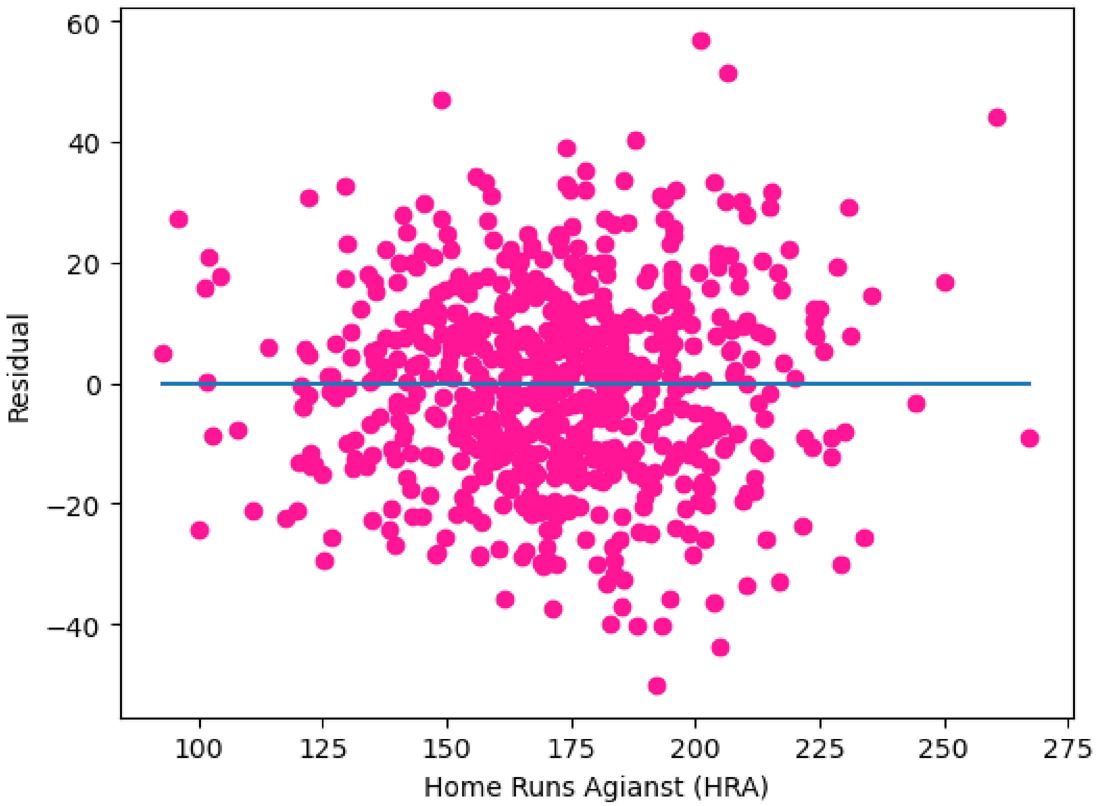
Now that all independent variables are at or below 0.05 they are all signifigant and fit 7 is our final fit.

The R-squared value tells us how well independent variables fit our dependent variables the closer to 1 the better and the closer to 0 is bad. The value of 0.991 is good and tells us that about 99% of our outputs can be explained and about 1% can't be.

[19]:
```
res = HRAfit7.resid
```

[20]:
```
fig = px.box(res)
fig.update_layout(title = "Boxplot of Residuals",
                  yaxis_title = "Residual Values")
fig.show()
```

[21]:
```
plt.scatter(HRAfit7.fittedvalues, res, color = "deeppink")
plt.plot([min(HRAfit7.fittedvalues), max(HRAfit7.fittedvalues)], [0,0])
plt.xlabel('Home Runs Agianst (HRA)')
plt.ylabel('Residual')
plt.show()
```



[22]:
```
fig = px.histogram(res)
fig.update_layout(title = "Histogram of Residuals",
                  xaxis_title = "Residuals",
                  yaxis_title = "Frequency")
fig.update_traces(marker_line_width = 1, marker_line_color = "white")
fig.show()
```

The boxplot shows us that our model is a good fit for our data because the median is close to 0 and Q1 and Q3 seem to be about the same length. This is further backed up by the histogram because our residuals seem to be normally distributed.

1.3.3 Wins Model

Choosing Independent Variables

The process for the second model will be pretty similar to the the process for the first model. I'll the second model I'll choose some independant variables that I think have an effect on the amount a wins a team has and then remove variables that have a p-value greater than 0.05. For the second model our dependant variable will be Wins (W). For the independent variables I'm choosing pretty much every hitting, pitching, and fielding variable because they all could have an impact on a teams a win total. For the independent variables I'm chosing: losses(L), runs(R), hits(H), doubles(2B), triples(3B), homeruns(HR), walks(BB), strikeouts(SO), stolen bases(SB), caught stealing(CS), sacrifice flys(SF), runs agianst(RA), earned runs(ER), earned run average(ERA), complete games(CG), shut outs(SHO), saves(SV), outs pitched(IPouts), hits aginast(HA), home runs agianst(HRA), walks agianst (BBA), strike outs agianst(SOA), errors(E), double plays(DP), and fielding percentage(FP).

[23]:
```python
indVars =␣
↪['L','R','H','2B','3B','HR','BB','SO','SB','CS','SF','RA','ER','ERA','CG','SHO','SV','IPout
depVar = 'W'
Wfit = sm.OLS(train[depVar], train[indVars]).fit()
Wfit.summary()
```

[23]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                      W   R-squared (uncentered):
1.000
Model:                            OLS   Adj. R-squared (uncentered):
1.000
Method:                 Least Squares   F-statistic:
1.179e+05
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:02   Log-Likelihood:
-1060.1
No. Observations:                 666   AIC:
2170.
Df Residuals:                     641   BIC:
2283.
Df Model:                          25
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
L             -0.8652      0.014    -59.751      0.000      -0.894      -0.837
R              0.0158      0.002      6.687      0.000       0.011       0.020
H             -0.0026      0.001     -1.758      0.079      -0.005       0.000
2B            -0.0028      0.002     -1.208      0.228      -0.007       0.002
3B            -0.0053      0.006     -0.879      0.380      -0.017       0.006
HR            -0.0063      0.003     -2.274      0.023      -0.012      -0.001
```

```
BB             -0.0029      0.001     -2.670      0.008      -0.005      -0.001
SO           9.851e-05      0.001      0.183      0.855      -0.001       0.001
SB             -0.0022      0.002     -1.060      0.290      -0.006       0.002
CS             -0.0055      0.006     -0.945      0.345      -0.017       0.006
SF             -0.0068      0.007     -0.981      0.327      -0.020       0.007
RA              0.0048      0.005      0.925      0.355      -0.005       0.015
ER              0.0087      0.012      0.757      0.449      -0.014       0.031
ERA            -1.2164      1.625     -0.748      0.455      -4.408       1.975
CG              0.0057      0.015      0.384      0.701      -0.023       0.035
SHO             0.0687      0.018      3.873      0.000       0.034       0.104
SV              0.0524      0.011      4.889      0.000       0.031       0.073
IPouts          0.0320      0.002     17.462      0.000       0.028       0.036
HA             -0.0047      0.001     -3.166      0.002      -0.008      -0.002
HRA            -0.0021      0.003     -0.690      0.490      -0.008       0.004
BBA            -0.0032      0.001     -2.696      0.007      -0.005      -0.001
SOA             0.0010      0.001      1.731      0.084      -0.000       0.002
E              -0.0052      0.005     -1.142      0.254      -0.014       0.004
DP              0.0050      0.003      1.540      0.124      -0.001       0.011
FP              8.5337      7.367      1.158      0.247      -5.933      23.001
==============================================================================
Omnibus:                       10.089   Durbin-Watson:                   2.087
Prob(Omnibus):                  0.006   Jarque-Bera (JB):               14.326
Skew:                           0.125   Prob(JB):                     0.000775
Kurtosis:                       3.674   Cond. No.                     8.40e+05
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[3] The condition number is large, 8.4e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[24]:
```python
# taking SO out of indVars because it is the least signifigant variable then I will remake the fit.
indVars.remove("SO")
Wfit2 = sm.OLS(train[depVar], train[indVars]).fit()
Wfit2.summary()
```

[24]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                      W   R-squared (uncentered):
```

```
1.000
Model:                            OLS   Adj. R-squared (uncentered):
1.000
Method:                 Least Squares   F-statistic:
1.230e+05
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:02   Log-Likelihood:
-1060.2
No. Observations:                 666   AIC:
2168.
Df Residuals:                     642   BIC:
2276.
Df Model:                          24
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
L             -0.8650      0.014    -59.958      0.000      -0.893      -0.837
R              0.0159      0.002      6.715      0.000       0.011       0.021
H             -0.0027      0.001     -2.010      0.045      -0.005   -6.1e-05
2B            -0.0027      0.002     -1.196      0.232      -0.007       0.002
3B            -0.0051      0.006     -0.862      0.389      -0.017       0.007
HR            -0.0062      0.003     -2.279      0.023      -0.012      -0.001
BB            -0.0029      0.001     -2.668      0.008      -0.005      -0.001
SB            -0.0022      0.002     -1.054      0.292      -0.006       0.002
CS            -0.0055      0.006     -0.955      0.340      -0.017       0.006
SF            -0.0070      0.007     -1.008      0.314      -0.021       0.007
RA             0.0049      0.005      0.939      0.348      -0.005       0.015
ER             0.0086      0.011      0.751      0.453      -0.014       0.031
ERA           -1.2071      1.623     -0.744      0.457      -4.395       1.981
CG             0.0050      0.014      0.348      0.728      -0.023       0.033
SHO            0.0688      0.018      3.888      0.000       0.034       0.104
SV             0.0523      0.011      4.890      0.000       0.031       0.073
IPouts         0.0321      0.002     17.638      0.000       0.029       0.036
HA            -0.0047      0.001     -3.168      0.002      -0.008      -0.002
HRA           -0.0021      0.003     -0.692      0.489      -0.008       0.004
BBA           -0.0032      0.001     -2.735      0.006      -0.005      -0.001
SOA            0.0011      0.001      1.955      0.051   -4.78e-06       0.002
E             -0.0052      0.005     -1.146      0.252      -0.014       0.004
DP             0.0051      0.003      1.572      0.116      -0.001       0.011
FP             8.4937      7.358      1.154      0.249      -5.956      22.943
==============================================================================
Omnibus:                       10.044   Durbin-Watson:                   2.087
Prob(Omnibus):                  0.007   Jarque-Bera (JB):               14.265
Skew:                           0.124   Prob(JB):                     0.000799
Kurtosis:                       3.673   Cond. No.                     8.19e+05
```

```
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[3] The condition number is large, 8.19e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[25]:

```
# taking CG out of indVars because it is the next least signifigant variable␣
↪then I will remake the fit.
indVars.remove("CG")
Wfit3 = sm.OLS(train[depVar], train[indVars]).fit()
Wfit3.summary()
```

[25]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                OLS Regression Results
================================================================================
=======
Dep. Variable:                      W   R-squared (uncentered):
1.000
Model:                            OLS   Adj. R-squared (uncentered):
1.000
Method:                 Least Squares   F-statistic:
1.285e+05
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:02   Log-Likelihood:
-1060.2
No. Observations:                 666   AIC:
2166.
Df Residuals:                     643   BIC:
2270.
Df Model:                          23
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
L             -0.8658      0.014    -60.836      0.000      -0.894      -0.838
R              0.0158      0.002      6.714      0.000       0.011       0.020
H             -0.0026      0.001     -1.990      0.047      -0.005   -3.51e-05
2B            -0.0027      0.002     -1.202      0.230      -0.007       0.002
3B            -0.0051      0.006     -0.870      0.384      -0.017       0.006
HR            -0.0062      0.003     -2.266      0.024      -0.012      -0.001
```

```
BB             -0.0028      0.001     -2.654      0.008      -0.005      -0.001
SB             -0.0022      0.002     -1.054      0.292      -0.006       0.002
CS             -0.0053      0.006     -0.917      0.360      -0.017       0.006
SF             -0.0068      0.007     -0.985      0.325      -0.020       0.007
RA              0.0050      0.005      0.951      0.342      -0.005       0.015
ER              0.0085      0.011      0.742      0.458      -0.014       0.031
ERA            -1.1968      1.622     -0.738      0.461      -4.382       1.988
SHO             0.0699      0.017      4.019      0.000       0.036       0.104
SV              0.0512      0.010      5.032      0.000       0.031       0.071
IPouts          0.0321      0.002     17.681      0.000       0.029       0.036
HA             -0.0047      0.001     -3.174      0.002      -0.008      -0.002
HRA            -0.0021      0.003     -0.692      0.489      -0.008       0.004
BBA            -0.0032      0.001     -2.733      0.006      -0.005      -0.001
SOA             0.0010      0.001      1.929      0.054   -1.78e-05       0.002
E              -0.0050      0.004     -1.121      0.263      -0.014       0.004
DP              0.0051      0.003      1.564      0.118      -0.001       0.011
FP              8.5152      7.353      1.158      0.247      -5.924      22.954
==============================================================================
Omnibus:                       10.144   Durbin-Watson:                   2.088
Prob(Omnibus):                  0.006   Jarque-Bera (JB):               14.528
Skew:                           0.122   Prob(JB):                     0.000700
Kurtosis:                       3.681   Cond. No.                     8.19e+05
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[3] The condition number is large, 8.19e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[26]:
```python
# taking HRA out of indVars because it is the next least signifigant variable then I will remake the fit.
indVars.remove("HRA")
Wfit4 = sm.OLS(train[depVar], train[indVars]).fit()
Wfit4.summary()
```

[26]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                      W   R-squared (uncentered):
1.000
Model:                            OLS   Adj. R-squared (uncentered):
```

```
1.000
Method:                 Least Squares   F-statistic:
1.344e+05
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:02   Log-Likelihood:
-1060.5
No. Observations:                 666   AIC:
2165.
Df Residuals:                     644   BIC:
2264.
Df Model:                          22
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
L             -0.8666      0.014    -61.142      0.000      -0.894      -0.839
R              0.0156      0.002      6.681      0.000       0.011       0.020
H             -0.0025      0.001     -1.934      0.054      -0.005    3.86e-05
2B            -0.0026      0.002     -1.152      0.250      -0.007       0.002
3B            -0.0050      0.006     -0.843      0.400      -0.017       0.007
HR            -0.0063      0.003     -2.327      0.020      -0.012      -0.001
BB            -0.0028      0.001     -2.623      0.009      -0.005      -0.001
SB            -0.0023      0.002     -1.116      0.265      -0.006       0.002
CS            -0.0052      0.006     -0.903      0.367      -0.016       0.006
SF            -0.0067      0.007     -0.980      0.327      -0.020       0.007
RA             0.0048      0.005      0.920      0.358      -0.005       0.015
ER             0.0077      0.011      0.677      0.499      -0.015       0.030
ERA           -1.1916      1.621     -0.735      0.463      -4.375       1.992
SHO            0.0709      0.017      4.087      0.000       0.037       0.105
SV             0.0509      0.010      5.012      0.000       0.031       0.071
IPouts         0.0320      0.002     17.679      0.000       0.028       0.036
HA            -0.0044      0.001     -3.122      0.002      -0.007      -0.002
BBA           -0.0029      0.001     -2.675      0.008      -0.005      -0.001
SOA            0.0010      0.001      1.856      0.064   -5.55e-05       0.002
E             -0.0048      0.004     -1.078      0.282      -0.014       0.004
DP             0.0051      0.003      1.568      0.117      -0.001       0.011
FP             8.5455      7.350      1.163      0.245      -5.887      22.978
==============================================================================
Omnibus:                        9.752   Durbin-Watson:                   2.085
Prob(Omnibus):                  0.008   Jarque-Bera (JB):               13.884
Skew:                           0.116   Prob(JB):                     0.000966
Kurtosis:                       3.668   Cond. No.                     8.19e+05
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
```

```
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[3] The condition number is large, 8.19e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[27]:
```
# taking ER out of indVars because it is the next least signifigant variable␣
↪then I will remake the fit.
indVars.remove("ER")
Wfit5 = sm.OLS(train[depVar], train[indVars]).fit()
Wfit5.summary()
```

[27]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                      W   R-squared (uncentered):
1.000
Model:                            OLS   Adj. R-squared (uncentered):
1.000
Method:                 Least Squares   F-statistic:
1.410e+05
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:02   Log-Likelihood:
-1060.7
No. Observations:                 666   AIC:
2163.
Df Residuals:                     645   BIC:
2258.
Df Model:                          21
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
L             -0.8662      0.014    -61.200      0.000      -0.894      -0.838
R              0.0156      0.002      6.677      0.000       0.011       0.020
H             -0.0025      0.001     -1.891      0.059      -0.005    9.52e-05
2B            -0.0026      0.002     -1.139      0.255      -0.007       0.002
3B            -0.0049      0.006     -0.829      0.407      -0.016       0.007
HR            -0.0062      0.003     -2.287      0.023      -0.012      -0.001
BB            -0.0028      0.001     -2.649      0.008      -0.005      -0.001
SB            -0.0022      0.002     -1.087      0.277      -0.006       0.002
CS            -0.0051      0.006     -0.888      0.375      -0.016       0.006
SF            -0.0068      0.007     -0.989      0.323      -0.020       0.007
```

```
RA              0.0062      0.005      1.314      0.189      -0.003       0.016
ERA            -0.2229      0.761     -0.293      0.770      -1.718       1.272
SHO             0.0703      0.017      4.060      0.000       0.036       0.104
SV              0.0509      0.010      5.011      0.000       0.031       0.071
IPouts          0.0330      0.001     31.590      0.000       0.031       0.035
HA             -0.0044      0.001     -3.093      0.002      -0.007      -0.002
BBA            -0.0029      0.001     -2.656      0.008      -0.005      -0.001
SOA             0.0010      0.001      1.844      0.066   -6.19e-05       0.002
E              -0.0064      0.004     -1.656      0.098      -0.014       0.001
DP              0.0050      0.003      1.543      0.123      -0.001       0.011
FP              4.1722      3.502      1.192      0.234      -2.704      11.048
==============================================================================
Omnibus:                        9.852   Durbin-Watson:                   2.085
Prob(Omnibus):                  0.007   Jarque-Bera (JB):               13.933
Skew:                           0.122   Prob(JB):                     0.000943
Kurtosis:                       3.665   Cond. No.                     3.86e+05
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[3] The condition number is large, 3.86e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[28]:
```python
# taking ERA out of indVars because it is the next least signifigant variable␣
 ↪then I will remake the fit.
indVars.remove("ERA")
Wfit6 = sm.OLS(train[depVar], train[indVars]).fit()
Wfit6.summary()
```

[28]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                      W   R-squared (uncentered):
1.000
Model:                            OLS   Adj. R-squared (uncentered):
1.000
Method:                 Least Squares   F-statistic:
1.482e+05
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:02   Log-Likelihood:
```

```
-1060.8
No. Observations:                 666   AIC:
2162.
Df Residuals:                     646   BIC:
2252.
Df Model:                          20
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
L             -0.8662      0.014    -61.256      0.000      -0.894      -0.838
R              0.0156      0.002      6.698      0.000       0.011       0.020
H             -0.0025      0.001     -1.903      0.057      -0.005    7.87e-05
2B            -0.0026      0.002     -1.136      0.256      -0.007       0.002
3B            -0.0050      0.006     -0.850      0.396      -0.017       0.007
HR            -0.0063      0.003     -2.324      0.020      -0.012      -0.001
BB            -0.0028      0.001     -2.655      0.008      -0.005      -0.001
SB            -0.0022      0.002     -1.063      0.288      -0.006       0.002
CS            -0.0053      0.006     -0.939      0.348      -0.016       0.006
SF            -0.0070      0.007     -1.016      0.310      -0.020       0.006
RA             0.0050      0.002      2.551      0.011       0.001       0.009
SHO            0.0702      0.017      4.059      0.000       0.036       0.104
SV             0.0509      0.010      5.017      0.000       0.031       0.071
IPouts         0.0333      0.001     50.183      0.000       0.032       0.035
HA            -0.0044      0.001     -3.165      0.002      -0.007      -0.002
BBA           -0.0029      0.001     -2.733      0.006      -0.005      -0.001
SOA            0.0010      0.001      1.846      0.065   -6.05e-05       0.002
E             -0.0057      0.003     -1.831      0.068      -0.012       0.000
DP             0.0050      0.003      1.565      0.118      -0.001       0.011
FP             3.1823      0.911      3.495      0.001       1.394       4.970
==============================================================================
Omnibus:                       10.053   Durbin-Watson:                   2.084
Prob(Omnibus):                  0.007   Jarque-Bera (JB):               14.265
Skew:                           0.124   Prob(JB):                     0.000799
Kurtosis:                       3.672   Cond. No.                     9.83e+04
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[3] The condition number is large, 9.83e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[29]:
```
# taking 3B out of indVars because it is the next least signifigant variable␣
 ↪then I will remake the fit.
indVars.remove("3B")
Wfit7 = sm.OLS(train[depVar], train[indVars]).fit()
Wfit7.summary()
```

[29]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                      W   R-squared (uncentered):
1.000
Model:                            OLS   Adj. R-squared (uncentered):
1.000
Method:                 Least Squares   F-statistic:
1.561e+05
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:02   Log-Likelihood:
-1061.1
No. Observations:                 666   AIC:
2160.
Df Residuals:                     647   BIC:
2246.
Df Model:                          19
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
L             -0.8671      0.014    -61.475      0.000      -0.895      -0.839
R              0.0153      0.002      6.651      0.000       0.011       0.020
H             -0.0025      0.001     -1.895      0.058      -0.005    8.94e-05
2B            -0.0025      0.002     -1.119      0.264      -0.007       0.002
HR            -0.0057      0.003     -2.178      0.030      -0.011      -0.001
BB            -0.0027      0.001     -2.562      0.011      -0.005      -0.001
SB            -0.0023      0.002     -1.137      0.256      -0.006       0.002
CS            -0.0055      0.006     -0.963      0.336      -0.017       0.006
SF            -0.0068      0.007     -0.998      0.318      -0.020       0.007
RA             0.0051      0.002      2.598      0.010       0.001       0.009
SHO            0.0698      0.017      4.036      0.000       0.036       0.104
SV             0.0504      0.010      4.977      0.000       0.031       0.070
IPouts         0.0333      0.001     50.254      0.000       0.032       0.035
HA            -0.0045      0.001     -3.213      0.001      -0.007      -0.002
BBA           -0.0029      0.001     -2.766      0.006      -0.005      -0.001
SOA            0.0009      0.001      1.840      0.066   -6.37e-05       0.002
E             -0.0057      0.003     -1.836      0.067      -0.012       0.000
```

```
DP             0.0050      0.003      1.566      0.118      -0.001       0.011
FP             3.1526      0.910      3.466      0.001       1.366       4.939
==============================================================================
Omnibus:                       10.188   Durbin-Watson:                   2.088
Prob(Omnibus):                  0.006   Jarque-Bera (JB):               14.307
Skew:                           0.132   Prob(JB):                     0.000782
Kurtosis:                       3.668   Cond. No.                     9.82e+04
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[3] The condition number is large, 9.82e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[30]:

```python
# taking CS out of indVars because it is the next least signifigant variable
↪then I will remake the fit.
indVars.remove("CS")
Wfit8 = sm.OLS(train[depVar], train[indVars]).fit()
Wfit8.summary()
```

[30]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                      W   R-squared (uncentered):
1.000
Model:                            OLS   Adj. R-squared (uncentered):
1.000
Method:                 Least Squares   F-statistic:
1.648e+05
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:02   Log-Likelihood:
-1061.6
No. Observations:                 666   AIC:
2159.
Df Residuals:                     648   BIC:
2240.
Df Model:                          18
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
```

```
==============================================================================
L             -0.8661      0.014    -61.561      0.000      -0.894      -0.839
R              0.0155      0.002      6.776      0.000       0.011       0.020
H             -0.0026      0.001     -2.009      0.045      -0.005   -5.86e-05
2B            -0.0023      0.002     -1.032      0.302      -0.007       0.002
HR            -0.0056      0.003     -2.166      0.031      -0.011      -0.001
BB            -0.0028      0.001     -2.612      0.009      -0.005      -0.001
SB            -0.0035      0.002     -2.087      0.037      -0.007      -0.000
SF            -0.0069      0.007     -1.011      0.312      -0.020       0.007
RA             0.0050      0.002      2.551      0.011       0.001       0.009
SHO            0.0711      0.017      4.128      0.000       0.037       0.105
SV             0.0501      0.010      4.946      0.000       0.030       0.070
IPouts         0.0333      0.001     50.275      0.000       0.032       0.035
HA            -0.0044      0.001     -3.171      0.002      -0.007      -0.002
BBA           -0.0030      0.001     -2.803      0.005      -0.005      -0.001
SOA            0.0011      0.001      2.083      0.038    6.04e-05       0.002
E             -0.0064      0.003     -2.120      0.034      -0.012      -0.000
DP             0.0051      0.003      1.575      0.116      -0.001       0.011
FP             3.0699      0.906      3.390      0.001       1.292       4.848
==============================================================================
Omnibus:                       10.009   Durbin-Watson:                   2.099
Prob(Omnibus):                  0.007   Jarque-Bera (JB):               14.131
Skew:                           0.126   Prob(JB):                     0.000854
Kurtosis:                       3.668   Cond. No.                     9.78e+04
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[3] The condition number is large, 9.78e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[31]:
```
# taking SF out of indVars because it is the next least signifigant variable␣
↪then I will remake the fit.
indVars.remove("SF")
Wfit9 = sm.OLS(train[depVar], train[indVars]).fit()
Wfit9.summary()
```

[31]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                      W   R-squared (uncentered):
```

```
1.000
Model:                            OLS   Adj. R-squared (uncentered):
1.000
Method:                 Least Squares   F-statistic:
1.745e+05
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:02   Log-Likelihood:
-1062.1
No. Observations:                 666   AIC:
2158.
Df Residuals:                     649   BIC:
2235.
Df Model:                          17
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
L             -0.8659      0.014    -61.551      0.000      -0.894      -0.838
R              0.0150      0.002      6.713      0.000       0.011       0.019
H             -0.0026      0.001     -2.038      0.042      -0.005   -9.61e-05
2B            -0.0023      0.002     -1.033      0.302      -0.007       0.002
HR            -0.0049      0.002     -1.956      0.051      -0.010    1.83e-05
BB            -0.0028      0.001     -2.642      0.008      -0.005      -0.001
SB            -0.0036      0.002     -2.165      0.031      -0.007      -0.000
RA             0.0049      0.002      2.505      0.012       0.001       0.009
SHO            0.0712      0.017      4.134      0.000       0.037       0.105
SV             0.0499      0.010      4.930      0.000       0.030       0.070
IPouts         0.0332      0.001     50.283      0.000       0.032       0.035
HA            -0.0043      0.001     -3.108      0.002      -0.007      -0.002
BBA           -0.0029      0.001     -2.743      0.006      -0.005      -0.001
SOA            0.0011      0.001      2.121      0.034    7.92e-05       0.002
E             -0.0064      0.003     -2.103      0.036      -0.012      -0.000
DP             0.0051      0.003      1.571      0.117      -0.001       0.011
FP             3.0686      0.906      3.389      0.001       1.290       4.847
==============================================================================
Omnibus:                       11.290   Durbin-Watson:                   2.098
Prob(Omnibus):                  0.004   Jarque-Bera (JB):               16.409
Skew:                           0.138   Prob(JB):                     0.000273
Kurtosis:                       3.717   Cond. No.                     9.78e+04
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

```
[3] The condition number is large, 9.78e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[32]:
```
# taking 2B out of indVars because it is the next least signifigant variable␣
↪then I will remake the fit.
indVars.remove("2B")
Wfit10 = sm.OLS(train[depVar], train[indVars]).fit()
Wfit10.summary()
```

[32]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                      W   R-squared (uncentered):
1.000
Model:                            OLS   Adj. R-squared (uncentered):
1.000
Method:                 Least Squares   F-statistic:
1.853e+05
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:02   Log-Likelihood:
-1062.7
No. Observations:                 666   AIC:
2157.
Df Residuals:                     650   BIC:
2229.
Df Model:                          16
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
L             -0.8656      0.014    -61.540      0.000      -0.893      -0.838
R              0.0146      0.002      6.640      0.000       0.010       0.019
H             -0.0028      0.001     -2.215      0.027      -0.005      -0.000
HR            -0.0044      0.002     -1.797      0.073      -0.009       0.000
BB            -0.0027      0.001     -2.584      0.010      -0.005      -0.001
SB            -0.0032      0.002     -1.989      0.047      -0.006   -4.11e-05
RA             0.0049      0.002      2.534      0.012       0.001       0.009
SHO            0.0714      0.017      4.143      0.000       0.038       0.105
SV             0.0508      0.010      5.042      0.000       0.031       0.071
IPouts         0.0332      0.001     50.422      0.000       0.032       0.034
HA            -0.0044      0.001     -3.163      0.002      -0.007      -0.002
BBA           -0.0029      0.001     -2.717      0.007      -0.005      -0.001
SOA            0.0011      0.001      2.126      0.034     8.2e-05       0.002
```

```
E              -0.0064      0.003     -2.097      0.036      -0.012      -0.000
DP              0.0053      0.003      1.648      0.100      -0.001       0.012
FP              3.1181      0.904      3.448      0.001       1.342       4.894
==============================================================================
Omnibus:                       12.013   Durbin-Watson:                   2.101
Prob(Omnibus):                  0.002   Jarque-Bera (JB):               17.714
Skew:                           0.146   Prob(JB):                     0.000142
Kurtosis:                       3.744   Cond. No.                     9.75e+04
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[3] The condition number is large, 9.75e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[33]:

```
# taking DP out of indVars because it is the next least signifigant variable
↪then I will remake the fit.
indVars.remove("DP")
Wfit11 = sm.OLS(train[depVar], train[indVars]).fit()
Wfit11.summary()
```

[33]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
=================================================================================
=======
Dep. Variable:                      W   R-squared (uncentered):
1.000
Model:                            OLS   Adj. R-squared (uncentered):
1.000
Method:                 Least Squares   F-statistic:
1.972e+05
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:02   Log-Likelihood:
-1064.1
No. Observations:                 666   AIC:
2158.
Df Residuals:                     651   BIC:
2226.
Df Model:                          15
Covariance Type:            nonrobust
==============================================================================
```

```
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
L             -0.8662      0.014    -61.519      0.000      -0.894      -0.839
R              0.0148      0.002      6.717      0.000       0.010       0.019
H             -0.0030      0.001     -2.332      0.020      -0.006      -0.000
HR            -0.0042      0.002     -1.730      0.084      -0.009       0.001
BB            -0.0030      0.001     -2.869      0.004      -0.005      -0.001
SB            -0.0034      0.002     -2.103      0.036      -0.007      -0.000
RA             0.0045      0.002      2.318      0.021       0.001       0.008
SHO            0.0727      0.017      4.218      0.000       0.039       0.107
SV             0.0493      0.010      4.905      0.000       0.030       0.069
IPouts         0.0334      0.001     51.549      0.000       0.032       0.035
HA            -0.0039      0.001     -2.858      0.004      -0.007      -0.001
BBA           -0.0024      0.001     -2.369      0.018      -0.004      -0.000
SOA            0.0008      0.000      1.665      0.096      -0.000       0.002
E             -0.0067      0.003     -2.208      0.028      -0.013      -0.001
FP             3.0519      0.905      3.374      0.001       1.276       4.828
==============================================================================
Omnibus:                       13.223   Durbin-Watson:                   2.107
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               20.504
Skew:                           0.145   Prob(JB):                     3.53e-05
Kurtosis:                       3.809   Cond. No.                     9.73e+04
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[3] The condition number is large, 9.73e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[34]:

```
# taking SOA out of indVars because it is the next least signifigant variable
↪then I will remake the fit.
indVars.remove("SOA")
Wfit12 = sm.OLS(train[depVar], train[indVars]).fit()
Wfit12.summary()
```

[34]:

```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                      W   R-squared (uncentered):
1.000
Model:                            OLS   Adj. R-squared (uncentered):
```

```
1.000
Method:                 Least Squares   F-statistic:
2.107e+05
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:02   Log-Likelihood:
-1065.5
No. Observations:                 666   AIC:
2159.
Df Residuals:                     652   BIC:
2222.
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
L             -0.8667      0.014    -61.484      0.000      -0.894      -0.839
R              0.0147      0.002      6.664      0.000       0.010       0.019
H             -0.0032      0.001     -2.482      0.013      -0.006      -0.001
HR            -0.0036      0.002     -1.470      0.142      -0.008       0.001
BB            -0.0031      0.001     -3.006      0.003      -0.005      -0.001
SB            -0.0036      0.002     -2.258      0.024      -0.007      -0.000
RA             0.0052      0.002      2.747      0.006       0.001       0.009
SHO            0.0732      0.017      4.241      0.000       0.039       0.107
SV             0.0488      0.010      4.849      0.000       0.029       0.069
IPouts         0.0340      0.001     64.592      0.000       0.033       0.035
HA            -0.0050      0.001     -4.326      0.000      -0.007      -0.003
BBA           -0.0027      0.001     -2.642      0.008      -0.005      -0.001
E             -0.0076      0.003     -2.530      0.012      -0.013      -0.002
FP             2.9812      0.905      3.295      0.001       1.204       4.758
==============================================================================
Omnibus:                       15.677   Durbin-Watson:                   2.110
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               25.026
Skew:                           0.174   Prob(JB):                     3.68e-06
Kurtosis:                       3.884   Cond. No.                     9.47e+04
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[3] The condition number is large, 9.47e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[35]:
```
# taking HR out of indVars because it is the next least signifigant variable␣
  ↪then I will remake the fit.
indVars.remove("HR")
Wfit13 = sm.OLS(train[depVar], train[indVars]).fit()
Wfit13.summary()
```

[35]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                      W   R-squared (uncentered):
1.000
Model:                            OLS   Adj. R-squared (uncentered):
1.000
Method:                 Least Squares   F-statistic:
2.265e+05
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:02   Log-Likelihood:
-1066.6
No. Observations:                 666   AIC:
2159.
Df Residuals:                     653   BIC:
2218.
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
L             -0.8652      0.014    -61.480      0.000      -0.893      -0.838
R              0.0127      0.002      7.284      0.000       0.009       0.016
H             -0.0021      0.001     -1.998      0.046      -0.004   -3.61e-05
BB            -0.0026      0.001     -2.669      0.008      -0.005      -0.001
SB            -0.0030      0.002     -1.917      0.056      -0.006    7.13e-05
RA             0.0047      0.002      2.528      0.012       0.001       0.008
SHO            0.0745      0.017      4.320      0.000       0.041       0.108
SV             0.0474      0.010      4.728      0.000       0.028       0.067
IPouts         0.0337      0.000     68.486      0.000       0.033       0.035
HA            -0.0048      0.001     -4.181      0.000      -0.007      -0.003
BBA           -0.0026      0.001     -2.562      0.011      -0.005      -0.001
E             -0.0073      0.003     -2.448      0.015      -0.013      -0.001
FP             3.0449      0.905      3.366      0.001       1.269       4.821
==============================================================================
Omnibus:                       15.546   Durbin-Watson:                   2.097
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               24.457
Skew:                           0.178   Prob(JB):                     4.89e-06
```

```
Kurtosis:                       3.869   Cond. No.                     9.45e+04
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[3] The condition number is large, 9.45e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

[36]:
```python
# taking SB out of indVars because it is the next least signifigant variable␣
 ↪then I will remake the fit.
indVars.remove("SB")
Wfit14 = sm.OLS(train[depVar], train[indVars]).fit()
Wfit14.summary()
```

[36]:
```
<class 'statsmodels.iolib.summary.Summary'>
"""
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:                      W   R-squared (uncentered):
1.000
Model:                            OLS   Adj. R-squared (uncentered):
1.000
Method:                 Least Squares   F-statistic:
2.444e+05
Date:                Fri, 28 Apr 2023   Prob (F-statistic):
0.00
Time:                        09:02:02   Log-Likelihood:
-1068.5
No. Observations:                 666   AIC:
2161.
Df Residuals:                     654   BIC:
2215.
Df Model:                          12
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
L             -0.8653      0.014    -61.359      0.000      -0.893      -0.838
R              0.0126      0.002      7.247      0.000       0.009       0.016
H             -0.0022      0.001     -2.074      0.038      -0.004      -0.000
BB            -0.0026      0.001     -2.667      0.008      -0.005      -0.001
RA             0.0050      0.002      2.716      0.007       0.001       0.009
```

```
SHO              0.0764      0.017      4.427      0.000       0.043       0.110
SV               0.0475      0.010      4.724      0.000       0.028       0.067
IPouts           0.0337      0.000     68.425      0.000       0.033       0.035
HA              -0.0050      0.001     -4.293      0.000      -0.007      -0.003
BBA             -0.0027      0.001     -2.678      0.008      -0.005      -0.001
E               -0.0081      0.003     -2.738      0.006      -0.014      -0.002
FP               2.8637      0.902      3.176      0.002       1.093       4.634
==============================================================================
Omnibus:                       17.346   Durbin-Watson:                   2.108
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               27.716
Skew:                           0.199   Prob(JB):                     9.58e-07
Kurtosis:                       3.917   Cond. No.                     9.40e+04
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
[3] The condition number is large, 9.4e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

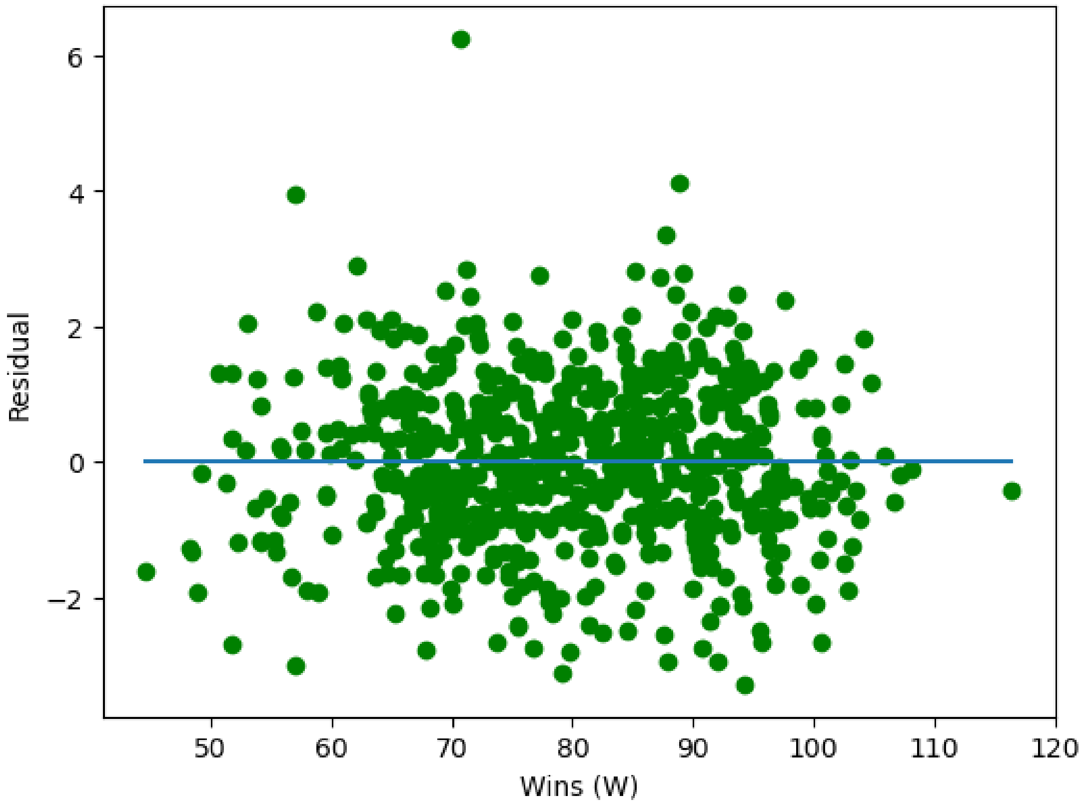
interpretation

Now that all the independent variables are at or below 0.05 they are all signifigant and fit14 is the final fit. The R-squared value tells us how well our independent variables fit our dependent variables the closer to 1 the better and the closer to 0 is bad. The value of 1 is as good as it gets and tells us that all of our outputs can be explained.

[37]:
```
res = Wfit14.resid
```

[38]:
```
fig = px.box(res)
fig.update_layout(title = "Boxplot of Residuals",
                  yaxis_title = "Residual Values")
fig.show()
```

[39]:
```
plt.scatter(Wfit14.fittedvalues, res, color = "green")
plt.plot([min(Wfit14.fittedvalues), max(Wfit14.fittedvalues)],[0,0])
plt.xlabel("Wins (W)")
plt.ylabel("Residual")
plt.show()
```



[40]:
```
fig = px.histogram(res)
fig.update_layout(title = "Histogram of Residuals",
                  xaxis_title = "Residuals",
                  yaxis_title = "Frequency")
fig.update_traces(marker_line_width = 1, marker_line_color = "orange")
fig.show()
```

The boxplot shows us that our model is a good fit for our data because the median is close to 0 and Q1 and Q3 seem to be about the same length. This is further backed up by the histogram because our residuals seem to be normally distributed.

1.4 Data Visualization

[41]:
```
# applying both predictive models with the test set
HRAindVars = ['ER','SHO','IPouts','HA','BBA','SOA']
WindVars = ['L','R','H','BB','RA','SHO','SV','IPouts','HA','BBA','E','FP']
predictionsHRA = HRAfit7.predict(test[HRAindVars])
predictionsW = Wfit14.predict(test[WindVars])
print(predictionsHRA.head())
print(predictionsW.head())
```

```
2154    119.134959
```

```
2156    152.538357
2157    107.246668
2163    108.520870
2167    154.437898
dtype: float64
2154    63.614678
2156    48.619044
2157    67.329649
2163    50.539221
2167    54.381424
dtype: float64
```

[42]:

```python
warnings.filterwarnings("ignore")
predictDF = test.copy()
predictDF['HRApredictions'] = predictionsHRA
predictDF['Wpredictions'] = predictionsW
RevisedTeams2DF = predictDF[['yearID', 'franchID','W', 'HRA',␣
↪'HRApredictions','Wpredictions']]
RevisedTeams2DF['roundedHRAPredictions'] = RevisedTeams2DF['HRApredictions'].
↪round(0)
RevisedTeams2DF['roundedWPredictions'] = RevisedTeams2DF['Wpredictions'].
↪round(0)
RevisedTeams2DF['HRATF'] = RevisedTeams2DF['HRA'] ==␣
↪RevisedTeams2DF['roundedHRAPredictions']
RevisedTeams2DF['WTF'] = RevisedTeams2DF['W'] ==␣
↪RevisedTeams2DF['roundedWPredictions']
RevisedTeams2DF.head()
# I rounded the predictions up because you can't have half a win or half a home
```

[42]:

	yearID	franchID	W	HRA	HRApredictions	Wpredictions	\
2154	1994	BAL	63	131	119.134959	63.614678	
2156	1994	ANA	47	150	152.538357	48.619044	
2157	1994	CHW	67	115	107.246668	67.329649	
2163	1994	FLA	51	120	108.520870	50.539221	
2167	1994	MIN	53	153	154.437898	54.381424	

	roundedHRAPredictions	roundedWPredictions	HRATF	WTF
2154	119.0	64.0	False	False
2156	153.0	49.0	False	False
2157	107.0	67.0	False	True
2163	109.0	51.0	False	True
2167	154.0	54.0	False	False

Graphs of Predicted Values vs Actual Values Using Test Set HRA preditions grpah

[43]:

```
fig = px.scatter(RevisedTeams2DF, x = 'roundedHRAPredictions', y = 'HRA', color␣
↪= 'franchID')
fig.update_layout(title = "Home Runs Agianst Values V. Predicted Home Runs␣
↪Aginast Values",
                  xaxis_title = "Predicted Home Runs Agianst",
                  yaxis_title = "Home Runs Agianst")
fig.add_shape(type = "line",
             line=dict(color="red", width=2),
             x0 = 100,
             y0=100,
             x1=275,
             y1=275)
fig.show()
```

Wins predictions graph

[44]:
```
fig = px.scatter(RevisedTeams2DF, x = 'roundedWPredictions', y = 'W', color =␣
↪'franchID')
fig.update_layout(title = "Wins Values V. Predicted Wins Values",
                  xaxis_title = "Predicted Wins",
                  yaxis_title = "Wins")
fig.add_shape(type = "line",
             line=dict(color="red", width=2),
             x0 = 40,
             y0=40,
             x1=125,
             y1=125)
fig.show()
```

The Home Runs Against model seems to give a rough estimate of how many home runs a team may give up while to Wins model seems to accurately predict a teams actual win total.

Results

[45]:
```
HRAResults = RevisedTeams2DF.groupby("HRATF").size().reset_index(name="count")
print(HRAResults)
WResults = RevisedTeams2DF.groupby("WTF").size().reset_index(name="count")
print(WResults)
```

```
   HRATF  count
0  False    155
1   True     11
     WTF  count
0  False    112
1   True     54
```

The Home Runs Agianst Model predicted 11/166 values correctly which is a little over 6% and isn't all that great while the Wins model predicted 54/166 values correctly which is around 33% and is pretty good especially when looking at the graph and seeing the predcited values are typically within 5 wins of the actual value.

1.5 Predicting 2023 wins so far

In order to get the 2023 stats that are avaliable so far I ended up downloading the csv files of team data from baseball reference and imported them into excel to clean rather than doing web scarapping and cleaning in python.

[46]:
```
# importing the csv file
teamStats2023 = pd.read_csv("2023TeamStats.csv")
teamStats2023.head()
```

[46]:

	franchID	R	H	BB	L	SHO	SV	IPouts	HA	BBA	E	FP	RA	W
0	ARI	121	226	58	12	4	7	684.0	202	106	8	0.992	123	14
1	ATL	130	221	98	8	2	7	672.0	197	80	16	0.982	90	17
2	BAL	125	200	97	8	3	6	642.6	198	71	12	0.986	104	16
3	BOS	146	222	92	13	0	7	684.0	233	80	17	0.982	138	13
4	CHC	130	221	83	10	5	2	618.6	163	78	11	0.987	87	13

[47]:
```
WindVars = ['L','R','H','BB','RA','SHO','SV','IPouts','HA','BBA','E','FP']
Wpredictions2023 = Wfit14.predict(teamStats2023[WindVars])
print(Wpredictions2023.head())
```

```
0    16.328718
1    19.087921
2    18.243139
3    15.283615
4    15.800378
dtype: float64
```

[48]:
```
predictDF = teamStats2023.copy()
predictDF['Wpredictions'] = Wpredictions2023
RevisedTeams2DF = predictDF[['franchID','W','Wpredictions']]
RevisedTeams2DF['roundedWPredictions'] = RevisedTeams2DF['Wpredictions'].
 ↪round(0)
RevisedTeams2DF['TF'] = RevisedTeams2DF['W'] ==␣
 ↪RevisedTeams2DF['roundedWPredictions']
RevisedTeams2DF.head()
```

[48]:

	franchID	W	Wpredictions	roundedWPredictions	TF
0	ARI	14	16.328718	16.0	False
1	ATL	17	19.087921	19.0	False
2	BAL	16	18.243139	18.0	False
3	BOS	13	15.283615	15.0	False
4	CHC	13	15.800378	16.0	False

[49]:
```
fig = px.scatter(RevisedTeams2DF, x = 'roundedWPredictions', y = 'W', color =␣
 ↪'franchID')
fig.update_layout(title = "Wins Values V. Predicted Wins Values in 2023",
                  xaxis_title = "Predicted Wins",
                  yaxis_title = "Wins")
```

```python
fig.add_shape(type = "line",
              line=dict(color="red", width=2),
              x0 = 0,
              y0=0,
              x1=40,
              y1=40)
fig.show()
```

[50]:
```python
WResults = RevisedTeams2DF.groupby("TF").size().reset_index(name="count")
print(WResults)
```

```
      TF  count
0  False     30
```

The model didn't predict the right amount of wins that a team currently has but the predictions were really close and this is probably due to the low sample size since the 2023 season just started.

2 Conclusion

In conclusion, I was able to answer all of the questions I wanted to. Even though the models I made weren't as good as I were expecting both models give you a good idea of how many wins a team may have or home runs agianst a team may give up. Overall I really don't think I could improve on either model with the information I used in the packages but, if I had more time I think I could've pulled more advanced baseball statistics and been able to get more accuracy out of both models. In the end both models did what they were supposed to and the predicted values give a good indication of what the real value will be.

3 References

https://www.baseball-reference.com/leagues/majors/2023.shtml#all_teams_standard_pitching