HOME OR AWAY

Where’s The Advantage?

The Problem & Hypothesis

- Does playing at home matter? What statistics are important?
- Analysis will use past four seasons of Premier League data
- Hypothesis: A team’s home form will tend to be a better predictor of final league position.

Data Used

- Data is abundant for the Premier League. Key data sources were WhoScored.com, TransferMarkt.com and MyFootballFacts.com.
- Each year had two training sets: Home and Away.
- Opta is the gold standard, but pricy

Data Collection

- Data was stored in tables on the sites, and broken down to home and away. Easy to copy to spreadsheet for CSV export
- Difficulty: Data between sites not always ordered the same way -- had to manually refactor by sorting everything by name

Example:

Premier League Tables

Standings | Form | Streaks | Progress

View: Overall Home Away Wide

R	Team	P	W	D	L	GF	GA	GD	Pts	Form
1	Manchester City	38	28	5	5	93	29	+64	89	W W W W W W
2	Manchester United	38	28	5	5	89	33	+56	89	L W D L W W
3	Arsenal	38	21	7	10	74	49	+25	70	W L D D D W
4	Tottenham	38	20	9	9	66	41	+25	69	L L W W D W
5	Newcastle United	38	19	8	11	56	51	+5	65	W W L W L L
6	Chelsea	38	18	10	10	65	46	+19	64	D D W L L W
7	Everton	38	15	11	12	50	40	+10	56	W D W D D W
8	Liverpool	38	14	10	14	47	40	+7	52	W L W L W L
9	Fulham	38	14	10	14	48	51	-3	52	D W L W W L
10	West Bromwich Albion	38	13	8	17	45	52	-7	47	L W W D D L
11	Swansea	38	12	11	15	44	51	-7	47	L W D D L W
12	Norwich	38	12	11	15	52	66	-14	47	W L L L D W
13	Sunderland	38	11	12	15	45	46	-1	45	L D D D L L
14	Stoke	38	11	12	15	36	53	-17	45	D L D D L D
15	Wigan	38	11	10	17	42	62	-20	43	W W L W W W
16	Aston Villa	38	7	17	14	37	53	-16	38	L D L D D L
17	Queens Park Rangers	38	10	7	21	43	66	-23	37	W L W L W L
18	Bolton	38	10	6	22	46	77	-31	36	D W D L D D
19	Blackburn	38	8	7	23	48	78	-30	31	L L W L L L
20	Wolverhampton Wanderers	38	5	10	23	40	82	-42	25	L D L D D L

Champions League | Champions League Qualifiers | Europa League | Relegation

Missed penalties 12/13

FC Chelsea

pl.	Penalty-taker/Club	Goalkeeper/Club	as of	Match min.	Final score
1	David Silva Man City	Kelvin Davis Southampton	0:0	17	3:2
1	Shane Long West Brom	Pepe Reina Liverpool	1:0	60	3:0
2	Djibril Cissé QPR	John Ruddy Norwich	1:1	19	1:1
3	Robin van Persie Man Utd	Kelvin Davis Southampton	2:1	68	2:3
4	Chicharito Man Utd	Ali Al Habsi Wigan	0:0	6	4:0
10	Wayne Rooney Man Utd	Vito Mannone Arsenal	1:0	45	2:1
11	Mikel Arteta Arsenal	Mark Schwarzer Fulham	3:3	90	3:3
18	Lucas Piazón Chelsea	Brad Guzan Aston Villa	7:0	90	8:0
22	Edin Dzeko Man City	Wojciech Szczesny Arsenal	0:0	9	0:2
22	Jonathan Walters Stoke City	Petr Cech Chelsea	0:4	89	0:4
25	Adel Taarabt QPR	Mark Bunn Norwich	0:0	56	0:0
26	Steven Gerrard Liverpool	Ben Foster West Brom	0:0	77	0:2
27	Frank Lampard Chelsea	Joe Hart Man City	0:0	52	2:0
27	Jonathan Walters Stoke City	Mark Schwarzer Fulham	1:0	56	1:0
29	Romelu Lukaku West Brom	Michel Vorm Swansea	1:1	57	2:1
29	Grant Holt Norwich	Artur Boruc Southampton	0:0	90	0:0
31	Loïc Rémy QPR	Mark Schwarzer Fulham	3:1	49	3:2

Data Cleaning & Formatting

- Break apart columns for Yellow Cards and Red Cards
- Calculate Penalty +/-
- Refactor sheets to ensure consistency in columns and data structure.

Statistical Method

Since this problem was both supervised and continuous, regression was a perfect fit. Additionally, since multiple tests were being run and aggregated, an ensemble method (of sorts) was utilized.

Key Libraries & Packages

- Pandas
- Numpy
- Statsmodels (OLS)
- MatPlotLib -- PyPlot

Ordinary Least Squares

- Method for estimating the unknown parameters in a linear regression model.
- Minimizes the sum of squared vertical distances between the observed responses in the dataset and the responses predicted by the linear approximation
- Allows easy testing of Null Hypothesis (team statistic have NO impact on points) simple display of P-values per statistic

Selecting Features

- Starting Drop: Table Position, Team, Played, Form, W, L, D
- Starting Include: Goals For, Goals Against, Goal Differential, Shots, Possession, Passing, Yellows, Reds, Fouls, Penalty For, Penalty Against, Penalty +/-

Results

- Home average R-Squared: 0.875723417037
- Away average R-Squared: 0.84283400775
- Problem: The smallest eigenvalue is 3.47e-15. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

Multicollinearity

- Multicollinearity is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a non-trivial degree of accuracy
- Issue throughout, nature of data
- Drop some features: GF, GA, Pens For, Pens Against.
- Not great, but better -- no more warning

Increasing R-Squared

- Data exploration -- manually creating combinations of features to see affect on R-squared
- Build a programmatic ‘Kitchen Sink’ approach to find best combination for home and away

Home Versus Away

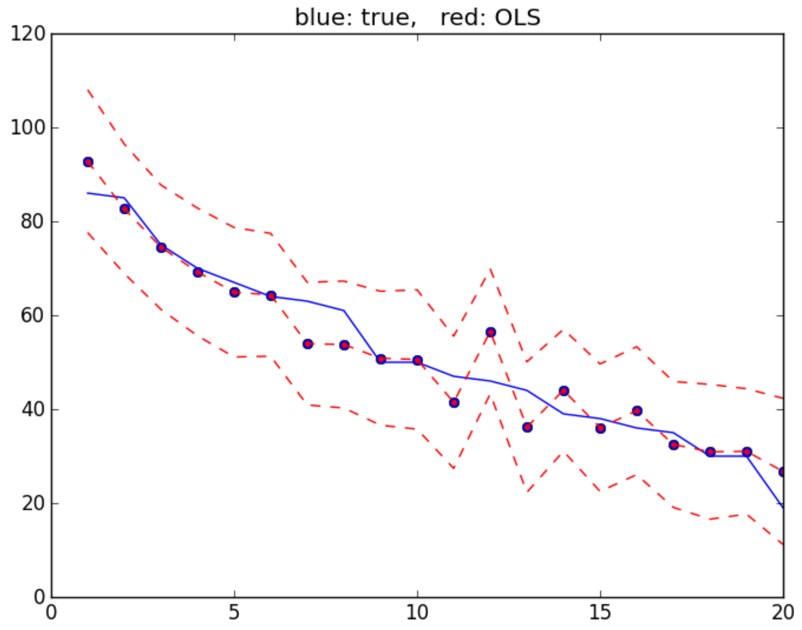
Throughout the project it became quite evident that certain statistics had consistent P values for home and away (such as Goal Differential). However, some statistic mattered more depending of if looking at home versus away (Penalty +/-), it was therefore important to treat home and away as entirely separate.

Final Results

- Home: The average r_squared is now: 0.897477697305 by dropping 'SpG', 'Pens +/- (F - A)', 'Possession%'
- Away: The average r_squared is now: 0.853326569647 by dropping 'SpG', 'Yellow', 'Red', 'Possession%'

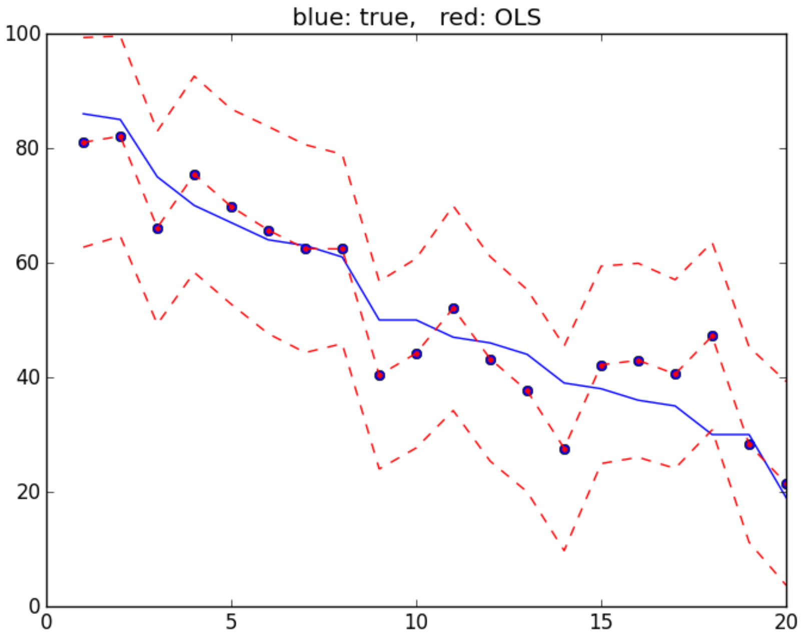
Home: 2009/2010

```
>>> print results_home10.summary(xname=feat_list_home)
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.925
Model:                            OLS   Adj. R-squared:                  0.905
Method:                 Least Squares   F-statistic:                     46.43
Date:                Tue, 12 Nov 2013   Prob (F-statistic):           2.83e-08
Time:                        17:56:04   Log-Likelihood:                -60.683
No. Observations:                  20   AIC:                             131.4
Df Residuals:                      15   BIC:                             136.3
Df Model:                           4
=================================================================================
                   coef    std err          t      P>|t|      [95.0% Conf. Int.]
---------------------------------------------------------------------------------
GD               0.9613      0.108      8.942      0.000         0.732     1.190
Yellow          -0.1148      0.258     -0.446      0.662        -0.664     0.434
Red             -2.8844      1.483     -1.945      0.071        -6.046     0.277
Pass Success%    0.3055      0.161      1.899      0.077        -0.037     0.649
Fouls pg         1.9199      1.082      1.775      0.096        -0.386     4.226
==============================================================================
Omnibus:                        0.192   Durbin-Watson:                   2.420
Prob(Omnibus):                  0.909   Jarque-Bera (JB):                0.252
Skew:                          -0.193   Prob(JB):                        0.881
Kurtosis:                       2.607   Cond. No.                         96.2
==============================================================================
```



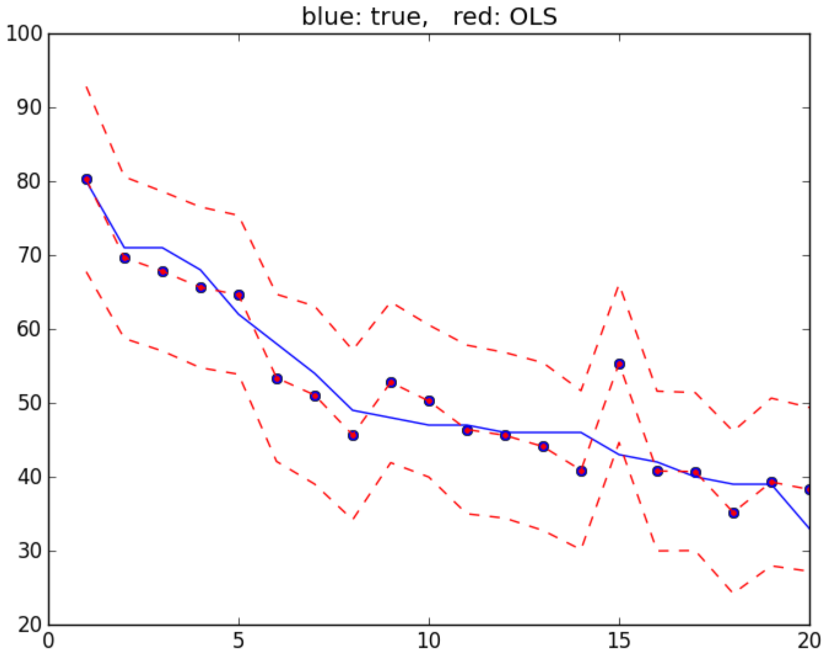
Away: 2009/2010

```
>>> print results_away10.summary(xname=feat_list_away)
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.869
Model:                            OLS   Adj. R-squared:                  0.844
Method:                 Least Squares   F-statistic:                     35.38
Date:                Tue, 12 Nov 2013   Prob (F-statistic):           2.72e-07
Time:                        17:56:05   Log-Likelihood:                -66.295
No. Observations:                  20   AIC:                             140.6
Df Residuals:                      16   BIC:                             144.6
Df Model:                           3
==================================================================================
                      coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------
GD                  1.0891      0.131      8.284      0.000         0.810     1.368
Pass Success%       0.3573      0.230      1.553      0.140        -0.130     0.845
Fouls pg            2.7268      1.258      2.168      0.046         0.061     5.393
Pens +/- (F - A)   -0.8711      0.808     -1.078      0.297        -2.585     0.843
==============================================================================
Omnibus:                        1.798   Durbin-Watson:                   1.558
Prob(Omnibus):                  0.407   Jarque-Bera (JB):                0.715
Skew:                          -0.441   Prob(JB):                        0.699
Kurtosis:                       3.284   Cond. No.                         58.3
==============================================================================
```



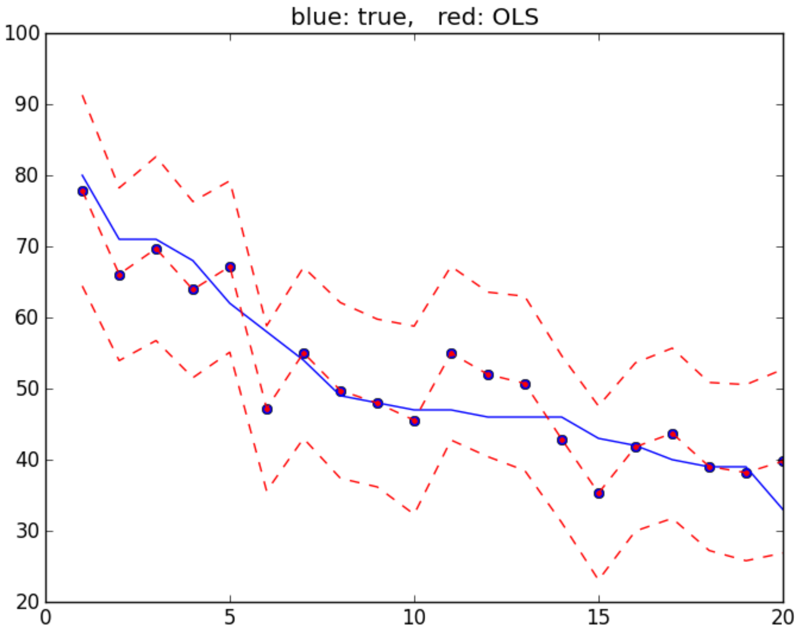
Home: 2010/2011

```
>>> print results_home11.summary(xname=feat_list_home)
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.895
Model:                            OLS   Adj. R-squared:                  0.867
Method:                 Least Squares   F-statistic:                     31.86
Date:                Tue, 12 Nov 2013   Prob (F-statistic):           3.59e-07
Time:                        18:01:02   Log-Likelihood:                -56.313
No. Observations:                  20   AIC:                             122.6
Df Residuals:                      15   BIC:                             127.6
Df Model:                           4
=================================================================================
                    coef    std err          t      P>|t|      [95.0% Conf. Int.]
---------------------------------------------------------------------------------
GD                0.8410      0.097      8.645      0.000         0.634     1.048
Yellow           -0.3095      0.278     -1.112      0.283        -0.903     0.284
Red               0.6650      0.934      0.712      0.487        -1.326     2.656
Pass Success%     0.5416      0.130      4.174      0.001         0.265     0.818
Fouls pg          0.8869      0.959      0.925      0.370        -1.157     2.931
==============================================================================
Omnibus:                       12.288   Durbin-Watson:                   2.260
Prob(Omnibus):                  0.002   Jarque-Bera (JB):               10.076
Skew:                          -1.404   Prob(JB):                      0.00649
Kurtosis:                       5.052   Cond. No.                         77.3
==============================================================================
```



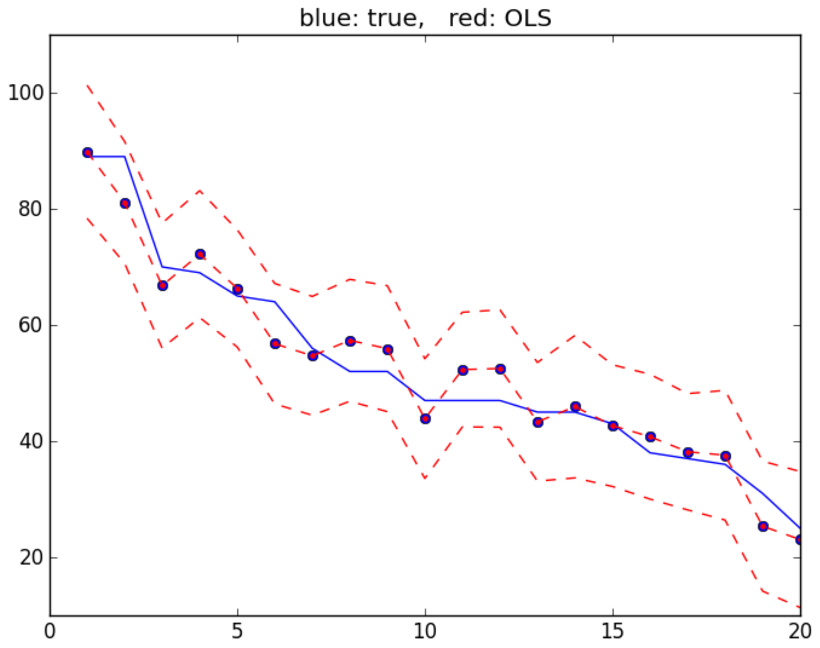
Away: 2010/2011

```
>>> print results_away11.summary(xname=feat_list_away)
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.856
Model:                            OLS   Adj. R-squared:                  0.829
Method:                 Least Squares   F-statistic:                     31.80
Date:                Tue, 12 Nov 2013   Prob (F-statistic):           5.65e-07
Time:                        18:01:03   Log-Likelihood:                -59.417
No. Observations:                  20   AIC:                             126.8
Df Residuals:                      16   BIC:                             130.8
Df Model:                           3
=====================================================================================
                        coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------------
GD                    0.7360      0.132      5.584      0.000         0.457     1.015
Pass Success%         1.0273      0.169      6.090      0.000         0.670     1.385
Fouls pg             -1.6888      1.092     -1.547      0.141        -4.004     0.626
Pens +/- (F - A)     -1.3042      0.725     -1.798      0.091        -2.842     0.233
==============================================================================
Omnibus:                        0.440   Durbin-Watson:                   1.867
Prob(Omnibus):                  0.802   Jarque-Bera (JB):                0.331
Skew:                           0.280   Prob(JB):                        0.848
Kurtosis:                       2.710   Cond. No.                         77.3
==============================================================================
```



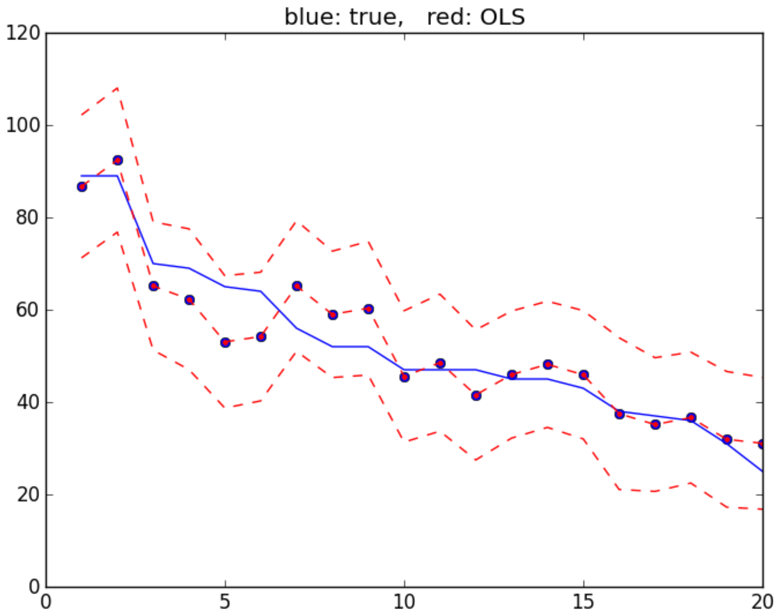
Home: 2011/2012

```
>>> print results_home12.summary(xname=feat_list_home)
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.948
Model:                            OLS   Adj. R-squared:                  0.934
Method:                 Least Squares   F-statistic:                     67.85
Date:                Tue, 12 Nov 2013   Prob (F-statistic):           2.01e-09
Time:                        18:05:07   Log-Likelihood:                -55.547
No. Observations:                  20   AIC:                             121.1
Df Residuals:                      15   BIC:                             126.1
Df Model:                           4
=================================================================================
                  coef    std err          t      P>|t|      [95.0% Conf. Int.]
---------------------------------------------------------------------------------
GD              0.9727      0.079     12.275      0.000         0.804     1.142
Yellow          0.2580      0.197      1.309      0.210        -0.162     0.678
Red             0.8869      0.773      1.147      0.269        -0.761     2.534
Pass Success%   0.5091      0.096      5.317      0.000         0.305     0.713
Fouls pg       -0.2450      0.781     -0.314      0.758        -1.910     1.421
==============================================================================
Omnibus:                        1.050   Durbin-Watson:                   1.838
Prob(Omnibus):                  0.592   Jarque-Bera (JB):                0.982
Skew:                           0.434   Prob(JB):                        0.612
Kurtosis:                       2.349   Cond. No.                         73.1
==============================================================================
```



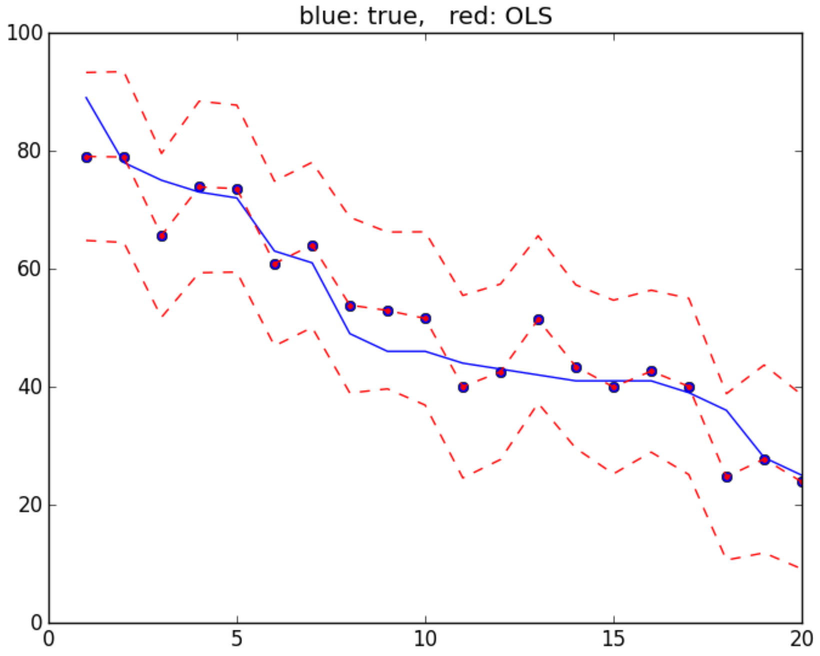
Away: 2011/2012

```
>>> print results_away12.summary(xname=feat_list_away)
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.892
Model:                            OLS   Adj. R-squared:                  0.872
Method:                 Least Squares   F-statistic:                     44.19
Date:                Tue, 12 Nov 2013   Prob (F-statistic):           5.74e-08
Time:                        18:05:08   Log-Likelihood:                -62.756
No. Observations:                  20   AIC:                             133.5
Df Residuals:                      16   BIC:                             137.5
Df Model:                           3
=================================================================================
                   coef    std err          t      P>|t|      [95.0% Conf. Int.]
---------------------------------------------------------------------------------
GD               1.2105      0.118     10.300      0.000         0.961     1.460
Pass Success%    0.4201      0.143      2.929      0.010         0.116     0.724
Fouls pg         2.4865      1.040      2.390      0.029         0.281     4.692
Pens +/- (F - A) -0.4365     0.718     -0.608      0.552        -1.959     1.086
==============================================================================
Omnibus:                        0.565   Durbin-Watson:                   1.201
Prob(Omnibus):                  0.754   Jarque-Bera (JB):                0.567
Skew:                           0.337   Prob(JB):                        0.753
Kurtosis:                       2.526   Cond. No.                         61.7
==============================================================================
```



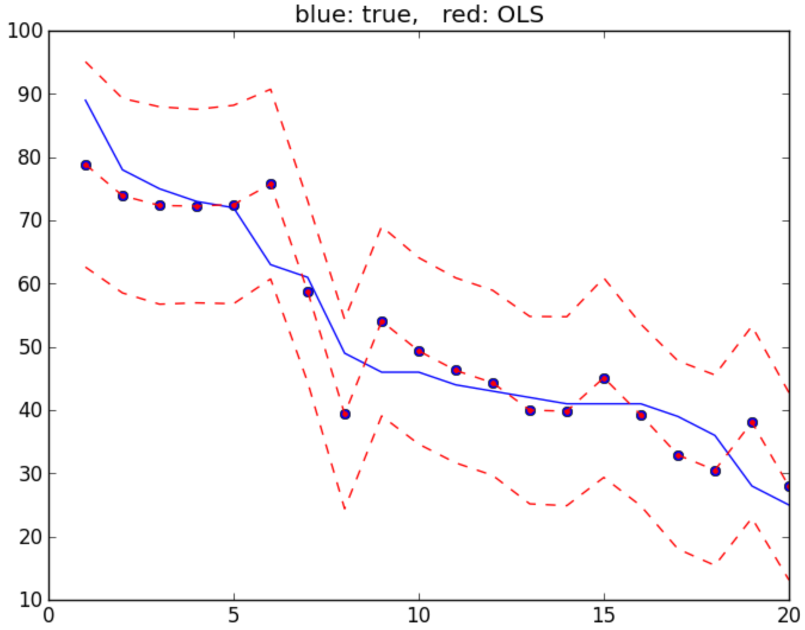
Home: 2012/2013

```
>>> print results_home13.summary(xname=feat_list_home)
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.909
Model:                            OLS   Adj. R-squared:                  0.884
Method:                 Least Squares   F-statistic:                     37.30
Date:                Tue, 12 Nov 2013   Prob (F-statistic):           1.25e-07
Time:                        18:08:04   Log-Likelihood:                -61.540
No. Observations:                  20   AIC:                             133.1
Df Residuals:                      15   BIC:                             138.1
Df Model:                           4
================================================================================
                    coef    std err          t      P>|t|      [95.0% Conf. Int.]
--------------------------------------------------------------------------------
GD                1.2252      0.129      9.462      0.000         0.949     1.501
Yellow            0.8220      0.484      1.698      0.110        -0.210     1.854
Red               0.6920      1.764      0.392      0.700        -3.069     4.453
Pass Success%     0.5613      0.194      2.900      0.011         0.149     0.974
Fouls pg         -2.0627      2.123     -0.972      0.347        -6.587     2.462
==============================================================================
Omnibus:                        1.993   Durbin-Watson:                   1.677
Prob(Omnibus):                  0.369   Jarque-Bera (JB):                1.159
Skew:                           0.590   Prob(JB):                        0.560
Kurtosis:                       2.985   Cond. No.                         139.
==============================================================================
```



Away: 2012/2013

```
>>> print results_away13.summary(xname=feat_list_away)
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.888
Model:                            OLS   Adj. R-squared:                  0.867
Method:                 Least Squares   F-statistic:                     42.40
Date:                Tue, 12 Nov 2013   Prob (F-statistic):           7.70e-08
Time:                        18:08:04   Log-Likelihood:                -63.555
No. Observations:                  20   AIC:                             135.1
Df Residuals:                      16   BIC:                             139.1
Df Model:                           3
==================================================================================
                    coef    std err          t      P>|t|      [95.0% Conf. Int.]
----------------------------------------------------------------------------------
GD                1.3170      0.190      6.930      0.000         0.914     1.720
Pass Success%     0.4827      0.207      2.328      0.033         0.043     0.922
Fouls pg          1.8043      1.477      1.221      0.240        -1.328     4.936
Pens +/- (F - A) -0.4929      0.702     -0.703      0.492        -1.980     0.994
==============================================================================
Omnibus:                        0.735   Durbin-Watson:                   1.776
Prob(Omnibus):                  0.692   Jarque-Bera (JB):                0.451
Skew:                          -0.353   Prob(JB):                        0.798
Kurtosis:                       2.795   Cond. No.                         83.9
==============================================================================
```



Interpretation

- Both home and away have high correlation with final league positions -- good teams are good and bad teams are bad.
- However, playing at home seems to matter -- higher r-squared

Interpretation Continued...

Importantly, the model finds that key referee decisions such as red cards are important features for home games whilst penalties are important for away games. Domain specific knowledge confirms this makes sense.

Room for Improvement

- Small dataset -- 30,000 foot view
- Next Steps: games within season, additional leagues, Opta
- Add additional features -- referee, month, etc
- Try additional regressions

Application

The Barclay’s Premier League has been slow to adapt new technology. As such decisions can be inconsistent and home fan pressure is left unchecked. Further analysis could be used to persuade the BLP, from a statistical standpoint, to adopt replay technology in certain circumstances, such as penalties or red cards