**Understanding the NFL**

**June 6, 2013**

**Patrick McNamara**

**Abstract**

Baseball may still being America’s pastime, but the NFL is our present and probably near-future favorite sport. Yet compared to baseball or even basketball, little is understood about a sport that relies on complex play-formations and ‘gut instinct’. By analyzing play-by-play data over the last 11 seasons, I provide high-level insights on what impacts the outcome of games, where some of our basic assumptions about the game hold true and where they are unfounded. While the model is less than reliable at predicting total points per game, it has proven more valuable in predicting whether or not the home team wins.

**Introduction**

The NFL is America’s most valuable sport in terms of revenue, and has been for quite some time. League income has grown from $8.5 billion in 2010 to $9.5 billion in 2013 and the commissioner recently set a goal of $25 billion in revenue by 2027 (Kaplan, 2013). Teams are spending $120M+ per year on player salaries (Associated Press, 2013) and have dozens of scouts working the draft to pick the best personnel possible. However when it comes to analyzing on-field data, the sport is left lacking, especially compared to baseball (Bradley, 2013). NFL teams don’t necessarily know (or at least talk publicly about) the value of a marginal rush vs. a marginal pass, the effect of travel on win probabilities or whether the ‘spotlight’ of a primetime game changes everything or has no effect.

**Methods**

I had two specific questions I wanted to answer initially using play-by-play data:

* How accurately can we predict total points per game?
* Can we reliably predict whether or not the home team wins?

The first big question was whether I wanted to approach the prediction from the pre-game perspective or the within-game perspective. The pre-game perspective answers the question “given all the games and plays to-date this season, which team do we predict will win this game and what will the total score be?” This is the purely ex-ante perspective akin to predicting the future based on the past. The post-game perspective answers the question “given all the games and plays over all seasons, including this game, which team do we predict will win this game and what will the total score be. The former has a limited sample size but a more valuable (and challenging) question to answer, while the latter has a many-fold larger sample size and a simpler (though less meaningful) question. I decided to go with the latter because I had no prior experience using Python for data analysis and thought it best to focus on basic ‘learning to program’ issues rather than complex analytical problems; though the ‘pre-game’ perspective is something to look at later on.

Finding the most reliable source for play-by-play data was the first big challenge. The NFL and ESPN have no APIs for box scores or other data, so page-scraping was the only option if I wanted to use those sources. However, the one source actually looking advanced NFL metrics makes the play-by-play data they use from 2002-2012 publicly available (Burke, 2010). This is my primary data source, containing:

* Basic down, distance, time, yard line, home & away, score, date
* Play description stating the play type, yards gained/lost, turnovers

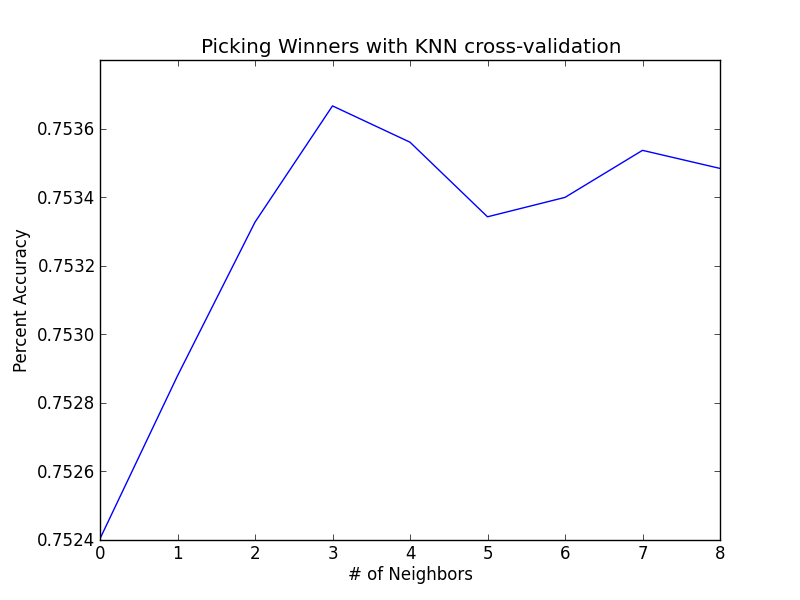
Additionally this data allows us to calculate other variables like yards gained, time of possession and penalty yards. These are potentially important predictors of game outcome as the more yards a team gains or more time they have the ball the more likely they are to reach the end zone, and the opposite effect for penalty yards. Data on the location and type of each team’s stadium was gathered to account for the effect of travel and domes on the game (Wikipedia).

The distinct questions here necessitate two different analytical methods. To predict the winner I chose the K-Nearest Neighbors classification with n-fold cross-validation, since this is a classification style problem. To predict points I used the linear regression and prediction package within python’s ‘sklearn’ package. I’m also using a linear model (as opposed to a Ridge or Lasso model) because the concern for overfitting is low with the large sample size (2684 games) and moderate number of predictors (23). The linearity of the data is backed up below.

**Setup and Analysis**

The first (and by far the largest) task was to inspect and clean the data, some of which was randomly blank (like play start times) and other times systematically blank or incorrect. In the case of games that went into overtime, there were many that were missing plays altogether or had blank play descriptions; on top of this handling time in overtime games (since there’s no definite period, it’s sudden death) presented another challenge. Since the data couldn’t be trusted as unbiased I excluded all games that went into overtime.

The first of the two analytical tasks I tackled was the win-loss classifier. I chose an n-fold cross-validation approach with the K-nearest neighbors method because of the diverse and deep dataset. With the seed set, I chose a 70/30 split (an 80/20 split revealed no further gains) and randomly chose 70% of the data as a training set and 30% as a test set, running it 100 times and averaging the results. The predictive accuracy converges on 74.6%, which, even though there’s no reference point, seems good. What was most interesting was when I varied the number of neighbors to analyze; moving from the default of 5 to a range of neighbors from 1 to 10 resulted in no significant changes in accuracy (see graph below, noting the scale). It seems likely the breadth and depth of the data minimizes the value of additional neighbors.



The second task was to estimate total points per game, using linear regression with backwards elimination. This requires the ‘kitchen sink’ approach at first, tossing all variables in and seeing what’s worthwhile. For this model my overall measure of accuracy is the absolute value of the percent error, represented below:

*i* /Pi)

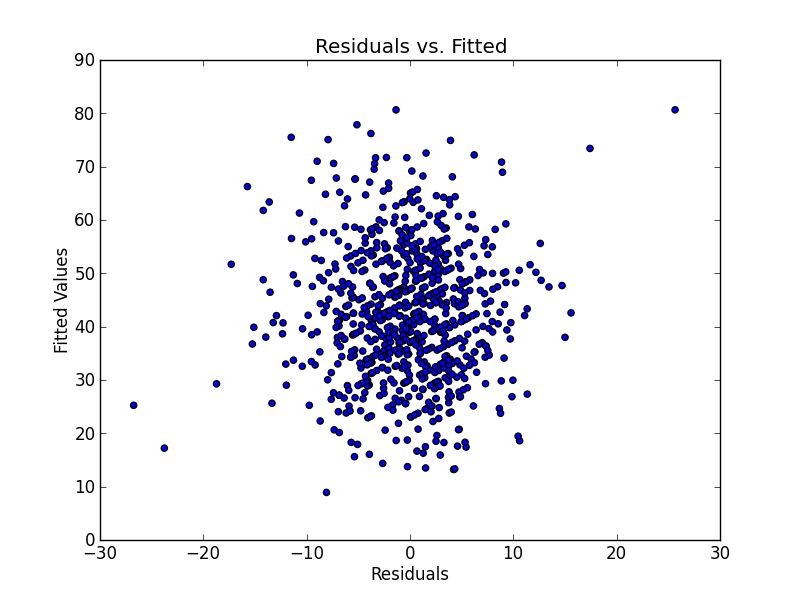
Where:

n = total predicted observations

ε = the predictive error for observation

P = actual points for observation

The ‘kitchen sink’ approach yielded a first-pass absolute percent error of 19.8%, which doesn’t seem too bad (though I have no reference point). One thing we want to do here is plot the residuals and the fitted values to make sure we can’t detect any bias. We want it to look random, with no detectable pattern.

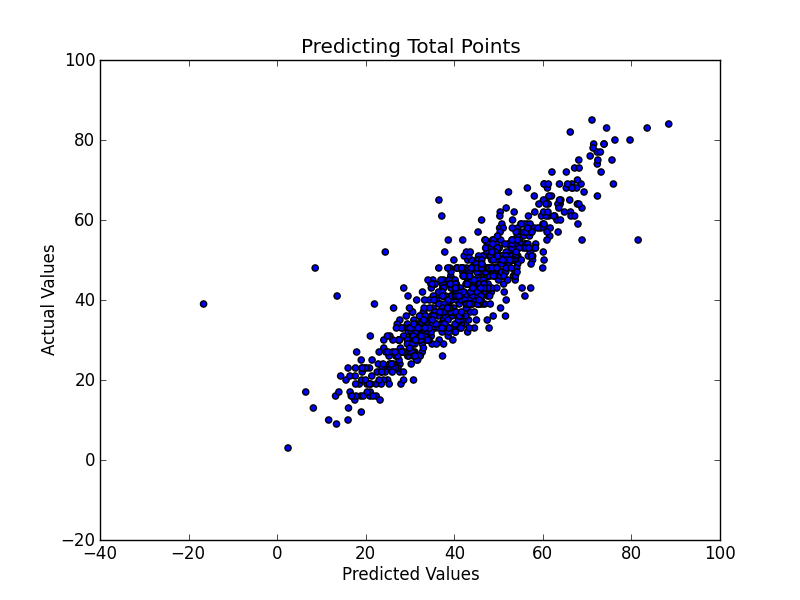


This is the type of spread we want, confirming that linear regression is appropriate.

We’ve got an R-squared of 0.86 and an adjusted R-squared about the same, which is a good sign. We’re also warned about multicollinearity, which is sure to exist because we’re looking at our first-pass model; we’ll address that later on after removing insignificant variable. Looking at the coefficients and standard errors (data tables in Appendix), ‘month’ looks like the first variable to drop with a p-value of 0.75. Month is used to capture the change in team behavior as the season goes on as well as the effect of temperature. It looks like there’s no real impact of month on total points, and if temperature has an impact (which we believe it does), the month doesn’t capture it well. After removing month, our R-squared and percent error don’t change, but we have an obvious next candidate to remove in ‘Playoffs with a p-value of 0.97. This variable captures any differences in points scored by playoff vs. regular season; it looks like there isn’t a meaningful difference. After removal our accuracy improves slightly to 19.7% and our R-squared is the same. This variable tries to capture the idea that games not played on Sundays that are usually played on national television may be different than others; it looks like there’s not much support to that theory.

Our next obvious variable to remove is ‘west\_east’, which is a binary variable indicating a west coast team traveling to play an east coast team and captures the effects of the time zone difference. After removing ‘west\_east’, accuracy doesn’t change and R-squared ticks up to 0.85. We still have an insignificant variable to remove in ‘east\_west’ with a p-value of 0.78, which is a binary variable indicating if an east coast team is traveling to play a west coast team. In the prior model the p-value was only 0.23, so the significant increase is probably representative of autocorrelation between ‘east\_west’ and ‘west\_east’. After removal our R-squared and accuracy are unchanged, but we have a new variable to remove in ‘Sunday’ with a p-value of 0.59, which is intended to capture any differences for prime time Monday and Thursday night games. It doesn’t look like there’s a difference, and we’re throwing it out.

After removing ‘Sunday’ our R-squared and accuracy are unchanged, and we only have two variables without 0.00 p-values: ‘dome’ and ‘penalty\_yards\_game’. If this were an academic setting we would potentially continue using backwards elimination because the p-values are greater than 0.05; but since we believe they play an important role in predicting points, plus the fact that we’re doing predictive modeling and not trying to get published, we’ll keep them in. Now that we’ve got our final model, we want to create a scatterplot as an extra step to confirm the use of a linear model.



This is what we want to see, a tight cluster in the form of a line through the middle.

**Conclusion and future research**

While I didn’t expect to be able to predict points well, picking 75% of the winners based on a plethora of in-game data seems low upon further reflection, and highlights how random a sport like football really is. One important factor missing from this exercise was weather data. The impact of temperature (see the effects of ‘month’) and precipitation cannot be understated. It’s much harder to complete passes, gain foot traction or maintain ball control in inclement weather, which depresses scoring output. Quantifying temperature, wind and the amount and type of precipitation is a logical next step. Weather underground has an API that can give historical data for a given zipcode and date, so learning to build that is a next step and will certainly improve accuracy.

The other avenue for future exploration is pre-game prediction: predicting the winner and total points based on our to-date knowledge for that season. This has obvious applications for betting markets, in-game commentary and fantasy football. One would utilize similar variables, but the analysis would be a team-centric approach rather than a game-centric approach as seen here, and you would obviously be unable to include the kind of in-game data (yards gained, penalties) that are so valuable with this kind of model.

**Table 1: ‘Kitchen Sink’**

OLS Regression Results

=================================================================

Dep. Variable: y R-squared: 0.855

Model: OLS Adj. R-squared: 0.853

Method: Least Squares F-statistic: 547.3

Date: Fri, 07 Jun 2013 Prob (F-statistic): 0.00

Time: 16:51:49 Log-Likelihood: -5864.8

No. Observations: 1879 AIC: 1.177e+04

Df Residuals: 1858 BIC: 1.189e+04

Df Model: 20

=================================================================

coef std err t P>|t| [95.0% Conf. Int.]

------------------------------------------------------------------------------

x1 7.3014 0.131 55.810 0.000 7.045 7.558

x2 2.1391 0.075 28.688 0.000 1.993 2.285

x3 1.5013 0.119 12.628 0.000 1.268 1.734

x4 2.1723 0.073 29.592 0.000 2.028 2.316

x5 -0.0068 0.002 -2.897 0.004 -0.011 -0.002

x6 0.8091 0.297 2.725 0.006 0.227 1.391

x7 0.5318 0.551 0.965 0.334 -0.549 1.612

x8 7.1866 0.129 55.765 0.000 6.934 7.439

x9 2.1175 0.075 28.200 0.000 1.970 2.265

x10 1.1933 0.115 10.390 0.000 0.968 1.419

x11 2.1796 0.073 29.707 0.000 2.036 2.323

x12 -0.0091 0.002 -4.001 0.000 -0.014 -0.005

x13 0.0368 0.116 0.317 0.751 -0.191 0.265

x14 -0.2643 0.055 -4.786 0.000 -0.373 -0.156

x15 0.0143 0.006 2.309 0.021 0.002 0.026

x16 0.2370 0.662 0.358 0.720 -1.061 1.535

x17 -1.8878 0.072 -26.188 0.000 -2.029 -1.746

x18 -0.0101 0.001 -8.683 0.000 -0.012 -0.008

x19 -0.3829 0.384 -0.998 0.318 -1.135 0.369

x20 -0.3319 0.074 -4.470 0.000 -0.478 -0.186

x21 0.2462 0.549 0.448 0.654 -0.831 1.323

=================================================================

Omnibus: 400.714 Durbin-Watson: 1.973

Prob(Omnibus): 0.000 Jarque-Bera (JB): 3005.582

Skew: 0.793 Prob(JB): 0.00

Kurtosis: 8.989 Cond. No. 1.09e+04

=================================================================

The condition number is large, 1.09e+04. This might indicate that there are

strong multicollinearity or other numerical problems.

**Table 2: Kitchen Sink – ‘Month’**

OLS Regression Results

=================================================================

Dep. Variable: y R-squared: 0.857

Model: OLS Adj. R-squared: 0.856

Method: Least Squares F-statistic: 588.6

Date: Fri, 07 Jun 2013 Prob (F-statistic): 0.00

Time: 16:54:04 Log-Likelihood: -5810.7

No. Observations: 1879 AIC: 1.166e+04

Df Residuals: 1859 BIC: 1.177e+04

Df Model: 19

=================================================================

coef std err t P>|t| [95.0% Conf. Int.]

------------------------------------------------------------------------------

x1 7.4827 0.129 57.855 0.000 7.229 7.736

x2 2.3312 0.074 31.524 0.000 2.186 2.476

x3 1.7220 0.116 14.852 0.000 1.495 1.949

x4 2.3876 0.073 32.663 0.000 2.244 2.531

x5 -0.0066 0.002 -2.885 0.004 -0.011 -0.002

x6 1.0414 0.286 3.635 0.000 0.480 1.603

x7 0.4779 0.520 0.919 0.358 -0.542 1.498

x8 7.3868 0.125 58.996 0.000 7.141 7.632

x9 2.2997 0.074 30.896 0.000 2.154 2.446

x10 1.3895 0.112 12.414 0.000 1.170 1.609

x11 2.3728 0.073 32.546 0.000 2.230 2.516

x12 -0.0094 0.002 -4.307 0.000 -0.014 -0.005

x13 -0.2803 0.053 -5.242 0.000 -0.385 -0.175

x14 0.0147 0.006 2.462 0.014 0.003 0.026

x15 -0.0209 0.576 -0.036 0.971 -1.151 1.109

x16 -2.0759 0.072 -28.955 0.000 -2.216 -1.935

x17 -0.0101 0.001 -8.955 0.000 -0.012 -0.008

x18 -0.4096 0.368 -1.114 0.266 -1.131 0.312

x19 -0.3656 0.072 -5.058 0.000 -0.507 -0.224

x20 0.4430 0.552 0.803 0.422 -0.639 1.525

=================================================================

Omnibus: 181.161 Durbin-Watson: 1.989

Prob(Omnibus): 0.000 Jarque-Bera (JB): 610.931

Skew: 0.456 Prob(JB): 2.18e-133

Kurtosis: 5.640 Cond. No. 9.83e+03

=================================================================

The condition number is large, 9.83e+03. This might indicate that there are

strong multicollinearity or other numerical problems.

**Table 3: Table 2 – Playoffs**

OLS Regression Results

=================================================================

Dep. Variable: y R-squared: 0.855

Model: OLS Adj. R-squared: 0.854

Method: Least Squares F-statistic: 609.9

Date: Fri, 07 Jun 2013 Prob (F-statistic): 0.00

Time: 16:59:41 Log-Likelihood: -5828.9

No. Observations: 1879 AIC: 1.170e+04

Df Residuals: 1860 BIC: 1.180e+04

Df Model: 18

=================================================================

coef std err t P>|t| [95.0% Conf. Int.]

------------------------------------------------------------------------------

x1 7.5482 0.129 58.643 0.000 7.296 7.801

x2 2.2924 0.074 31.004 0.000 2.147 2.437

x3 1.6015 0.116 13.832 0.000 1.374 1.829

x4 2.2946 0.073 31.378 0.000 2.151 2.438

x5 -0.0103 0.002 -4.479 0.000 -0.015 -0.006

x6 0.6018 0.286 2.102 0.036 0.040 1.163

x7 0.6371 0.540 1.179 0.238 -0.423 1.697

x8 7.3033 0.126 58.094 0.000 7.057 7.550

x9 2.2191 0.074 30.140 0.000 2.075 2.364

x10 1.3692 0.111 12.359 0.000 1.152 1.586

x11 2.2903 0.072 31.879 0.000 2.149 2.431

x12 -0.0059 0.002 -2.683 0.007 -0.010 -0.002

x13 -0.2524 0.052 -4.808 0.000 -0.355 -0.149

x14 0.0149 0.006 2.514 0.012 0.003 0.027

x15 -2.0149 0.071 -28.391 0.000 -2.154 -1.876

x16 -0.0098 0.001 -8.763 0.000 -0.012 -0.008

x17 -0.4354 0.367 -1.186 0.236 -1.155 0.285

x18 -0.2613 0.072 -3.648 0.000 -0.402 -0.121

x19 0.0681 0.538 0.127 0.899 -0.987 1.123

=================================================================

Omnibus: 156.471 Durbin-Watson: 1.937

Prob(Omnibus): 0.000 Jarque-Bera (JB): 476.492

Skew: 0.415 Prob(JB): 3.40e-104

Kurtosis: 5.324 Cond. No. 9.24e+03

=================================================================

**Table 4: Table 3 – West\_East**

OLS Regression Results

=================================================================

Dep. Variable: y R-squared: 0.851

Model: OLS Adj. R-squared: 0.850

Method: Least Squares F-statistic: 626.0

Date: Fri, 07 Jun 2013 Prob (F-statistic): 0.00

Time: 17:08:40 Log-Likelihood: -5839.8

No. Observations: 1879 AIC: 1.172e+04

Df Residuals: 1861 BIC: 1.182e+04

Df Model: 17

=================================================================

coef std err t P>|t| [95.0% Conf. Int.]

------------------------------------------------------------------------------

x1 7.2971 0.131 55.825 0.000 7.041 7.553

x2 2.1499 0.074 28.867 0.000 2.004 2.296

x3 1.6093 0.116 13.815 0.000 1.381 1.838

x4 2.1891 0.073 29.967 0.000 2.046 2.332

x5 -0.0062 0.002 -2.735 0.006 -0.011 -0.002

x6 1.0222 0.291 3.508 0.000 0.451 1.594

x7 -0.1527 0.540 -0.283 0.777 -1.212 0.907

x8 7.1827 0.127 56.541 0.000 6.934 7.432

x9 2.1258 0.075 28.447 0.000 1.979 2.272

x10 1.1543 0.112 10.289 0.000 0.934 1.374

x11 2.1889 0.073 30.153 0.000 2.047 2.331

x12 -0.0116 0.002 -5.299 0.000 -0.016 -0.007

x13 -0.3738 0.054 -6.911 0.000 -0.480 -0.268

x14 0.0204 0.006 3.319 0.001 0.008 0.032

x15 -1.9007 0.072 -26.306 0.000 -2.042 -1.759

x16 -0.0091 0.001 -7.797 0.000 -0.011 -0.007

x17 -0.5324 0.380 -1.401 0.161 -1.278 0.213

x18 -0.3273 0.073 -4.465 0.000 -0.471 -0.184

=================================================================

Omnibus: 412.810 Durbin-Watson: 2.038

Prob(Omnibus): 0.000 Jarque-Bera (JB): 3302.849

Skew: 0.804 Prob(JB): 0.00

Kurtosis: 9.293 Cond. No. 8.87e+03

=================================================================

**Table 5: Table 4 – East\_West**

OLS Regression Results

=================================================================

Dep. Variable: y R-squared: 0.848

Model: OLS Adj. R-squared: 0.846

Method: Least Squares F-statistic: 647.5

Date: Fri, 07 Jun 2013 Prob (F-statistic): 0.00

Time: 17:14:13 Log-Likelihood: -5862.2

No. Observations: 1879 AIC: 1.176e+04

Df Residuals: 1862 BIC: 1.185e+04

Df Model: 16

=================================================================

coef std err t P>|t| [95.0% Conf. Int.]

------------------------------------------------------------------------------

x1 7.3003 0.130 56.256 0.000 7.046 7.555

x2 2.1387 0.074 28.737 0.000 1.993 2.285

x3 1.5388 0.117 13.103 0.000 1.308 1.769

x4 2.1817 0.074 29.384 0.000 2.036 2.327

x5 -0.0075 0.002 -3.216 0.001 -0.012 -0.003

x6 1.0161 0.294 3.461 0.001 0.440 1.592

x7 6.9534 0.128 54.339 0.000 6.702 7.204

x8 2.0997 0.074 28.209 0.000 1.954 2.246

x9 1.1566 0.113 10.231 0.000 0.935 1.378

x10 2.1991 0.073 30.004 0.000 2.055 2.343

x11 -0.0092 0.002 -4.136 0.000 -0.014 -0.005

x12 -0.3010 0.055 -5.522 0.000 -0.408 -0.194

x13 0.0194 0.006 3.172 0.002 0.007 0.031

x14 -1.8701 0.072 -25.956 0.000 -2.011 -1.729

x15 -0.0109 0.001 -9.492 0.000 -0.013 -0.009

x16 -0.2035 0.374 -0.545 0.586 -0.936 0.529

x17 -0.3936 0.074 -5.322 0.000 -0.539 -0.249

=================================================================

Omnibus: 383.971 Durbin-Watson: 2.059

Prob(Omnibus): 0.000 Jarque-Bera (JB): 2771.584

Skew: 0.763 Prob(JB): 0.00

Kurtosis: 8.751 Cond. No. 6.06e+03

=================================================================

**Table 6: Table 5 – Sunday**

OLS Regression Results

=================================================================

Dep. Variable: y R-squared: 0.863

Model: OLS Adj. R-squared: 0.862

Method: Least Squares F-statistic: 781.3

Date: Fri, 07 Jun 2013 Prob (F-statistic): 0.00

Time: 17:18:21 Log-Likelihood: -5793.9

No. Observations: 1879 AIC: 1.162e+04

Df Residuals: 1863 BIC: 1.171e+04

Df Model: 15

=================================================================

coef std err t P>|t| [95.0% Conf. Int.]

------------------------------------------------------------------------------

x1 7.7407 0.127 61.075 0.000 7.492 7.989

x2 2.3336 0.072 32.237 0.000 2.192 2.476

x3 1.7587 0.113 15.540 0.000 1.537 1.981

x4 2.3560 0.072 32.945 0.000 2.216 2.496

x5 -0.0091 0.002 -4.089 0.000 -0.014 -0.005

x6 0.4861 0.285 1.703 0.089 -0.074 1.046

x7 7.5955 0.126 60.174 0.000 7.348 7.843

x8 2.2914 0.073 31.471 0.000 2.149 2.434

x9 1.4905 0.109 13.618 0.000 1.276 1.705

x10 2.3530 0.071 33.038 0.000 2.213 2.493

x11 -0.0081 0.002 -3.719 0.000 -0.012 -0.004

x12 -0.1951 0.053 -3.690 0.000 -0.299 -0.091

x13 0.0110 0.006 1.847 0.065 -0.001 0.023

x14 -2.0918 0.070 -29.896 0.000 -2.229 -1.955

x15 -0.0096 0.001 -8.693 0.000 -0.012 -0.007

x16 -0.3334 0.071 -4.701 0.000 -0.473 -0.194

=================================================================

Omnibus: 114.834 Durbin-Watson: 2.022

Prob(Omnibus): 0.000 Jarque-Bera (JB): 320.645

Skew: 0.305 Prob(JB): 2.36e-70

Kurtosis: 4.930 Cond. No. 4.81e+03

=================================================================