Interpreting Regression Equations

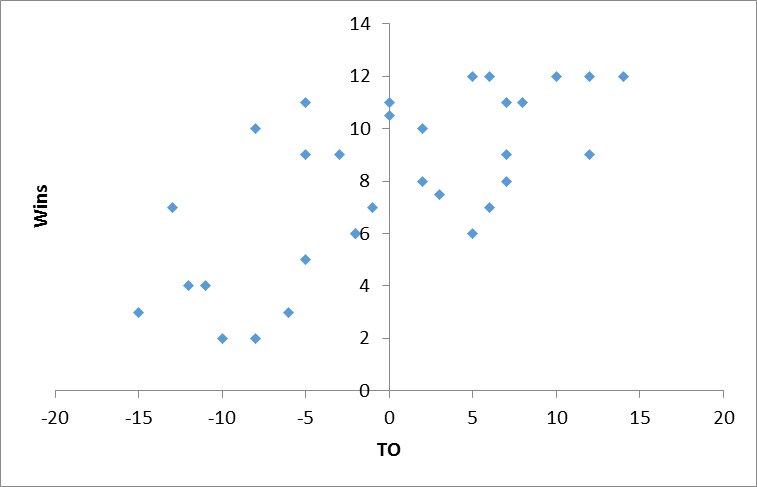
Consider the following data from the 2014 NFL Regular Season:

TO DIFF is the difference between the turnovers captured and turnovers given up, i.e. the greater the number, the better for the team. TO DIFF ranges from -15 for Oakland to +14 for Green Bay. The NFL season is 16 games long, and WINS range from 2 for Tampa Bay to 12 for New England, Denver, Dallas, Green Bay, and Seattle.

|  |  |  |
| --- | --- | --- |
| **Team** | **Wins** | **TO** |
| New Orleans Saints | 7 | -13 |
| Denver Broncos | 12 | 5 |
| Houston Texans | 9 | 12 |
| Arizona Cardinals | 11 | 8 |
| New England Patriots | 12 | 12 |
| Altantla Falcons | 6 | 5 |
| New York Giants | 6 | -2 |
| Green Bay Packers | 12 | 14 |
| Philadelphia Eagles | 10 | -8 |
| Carolina Panthers | 7.5 | 3 |
| San Diego Chargers | 9 | -5 |
| Miami Dolphins | 8 | 2 |
| Dallas Cowboys | 12 | 6 |
| Tampa Bay Buccaneers | 2 | -8 |
| Indianapolis Colts | 11 | -5 |
| New York Jets | 4 | -11 |
| Minnesota Vikings | 7 | -1 |
| Baltimore Ravens | 10 | 2 |
| Washington Redskins | 4 | -12 |
| Jacksonville Jaguars | 3 | -6 |
| Tennessee Titans | 2 | -10 |
| Pittsburgh Steelers | 11 | 0 |
| San Francisco 49ers | 8 | 7 |
| Kansas City Chiefs | 9 | -3 |
| Buffalo Bills | 9 | 7 |
| Chicago Bears | 5 | -5 |
| St. Louis Rams | 6 | -2 |
| Seattle Seahawks | 12 | 10 |
| Oakland Raiders | 3 | -15 |
| Detriot Lions | 11 | 7 |
| Cleveland Browns | 7 | 6 |
| Cincinnati Bengals | 10.5 | 0 |
|  |  |  |
| **Average** | **8** | **0** |

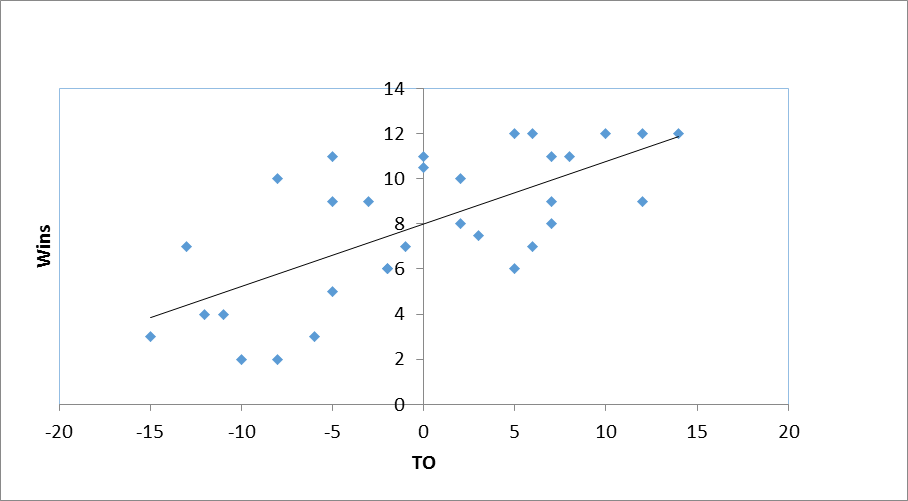
Note that since each game has both a winner and a loser, the average number of WINS is 8. Also, since a turnover captured by team A is one given up by Team B, the average value for TO DIFF is 0.

The relation between WINS and TO DIFF is depicted in the graph below. There is an obvious positive relationship between the two variables.



A linear least squares regression is one way of quantifying the relation between two variables such as this. A regression equation estimated by a statistical package like SAS, Stata, or a spreadsheet program like Excel uses formulas which define a line through the data (hence the term linear) which has the property that the sum of the squared distances between a line and the data points is minimized (hence the term least squares).

The least squares regression line through the data on turnovers and wins is depicted below.



The line is the estimated regression line through the data. Note that it passes through the means of the data (0 turnovers, 8 wins).

Note that also if a team were to move from 0 turnovers to +10 turnovers (remember, takeaways minus giveaways) then the line implies wins would increase from 8 to 10. +10 turnovers in a 16 game season represents about 1 net turnover per game. Such a team wins two more games in the season. This suggests that turnovers are a non-trivial part of NFL football, perhaps important.

The regression line in this case takes the form:

WINS = A + B\*TODIFF

where A is the intercept and B is the slope of the coefficient. The slope coefficient measures the incremental effect of TODIFF on Wins. The statistical or spreadsheet package delivers estimates for the intercept and slope coefficients derived from least squares formulas.

The statistical tables presented in the papers we’ll be reading in class have three important things to focus on.

1. The magnitude of the coefficients in the estimated linear regression.

We will examine the magnitude of the coefficients to see how reasonable changes in a variable such as turnovers affects an outcome of interest, such as wins.

2. The precision of the estimates (t-stat or t-ratio)

The t-ratios reported in tables are typically the ratio of the coefficient estimate divided by its standard error. The larger the standard error, the less precise the coefficient estimate. Obviously, we prefer to make observations based on coefficients which are large relative to the imprecision in their estimate. Hence, the larger the t-ratio, the more confident we are in making these observations.

Our rule of thumb is that coefficient estimates are reasonably precise when their t-ratios are 2.0 or better. There is some statistical theory behind this rule of thumb, but we don’t need to go into that here.

3. The explanatory power of the regression (R2 or R-squared).

R2 measures the percent of the variation of the dependent variable that is explained by the regression equation. R2 thus lies between zero and one.

If R2 = 1, then all data points would lie of the regression line. For lesser values of R2, the data points will be dispersed some distance from the line. The better the fit of the line to the data, the higher the R2. We are thus more confident in making observations when R2 is high (near 1) than low (near 0).

The output from an Excel spreadsheet regression of WINS on TODIFF is listed below.



From the column labeled coefficients, we see that A = 8 and B = .277692308. Hence our estimated equation is

WINS = 8 + 0.277692308\*TODIFF

The t-stat column reports the t-ratios for each coefficient estimate. They are both well beyond our cutoff point of 2.0, so we can have some confidence in making observations based on this table.

The value for R2 is .481185231, indicating that 48.1% of the variance in WINS across NFL teams is “explained” by just TODIFF. This is surprisingly large, but far from 1.

If we were to give a team 1 additional turnover per game, TODIFF would be 16 (given the 16 game season).

Since the coefficient estimate is a slope, we can then calculate the following:

WINDS +.277692308\*TODIFF

or WINS = .277692308\*16 = 4.44

for TODIFF = 16

Hence, 1 additional turnover per game accounts for 4 more wins in a 16 game NFL season.

On can add more variables to the equation, which puts us in the realm of multiple or multivariate (meaning more than 1 independent variable) regression.

In the case of NFL wins, the strength of the offense at gaining yards & the defense at preventing teams from gaining yards are obviously relevant. If we estimate a linear regression using these factors in addition to turnovers we obtain the following (again, from the 2014 season). In the table OYDS represents average yards gained per game by a team’s offense. DYDS represents average yards allowed per game by a team’s defense.



Most coefficient estimates have t-stats in excess of 2.

The magnitude of the coefficient estimate for TO DIFF (at .073342) is smaller now that we take into account offensive and defensive prowess. Is it significant?

The coefficient for OYDS is .045. So a team that increases its yards gained by 50 yards per game could expect to increase its win total by .045(50) = 2.25.

The coefficient for DYDS is -.029. So a team that decreases its yards allowed per game by 50 yards per game could expect to increase its win total by -.029(-50) = 1.45.

The regression hints that, yard for yard, it is more important for your offense to gain yardage than to deny the opposing team yardage.

Looking at R2, we find that adding these two factors raises the variance of WINS explained by the regression to 77.8%. Not bad for a simple regression.