What Wins Basketball Games

A Statistical Approach to Performance Analysis

Theodore Stetzel

Baruch College – STA 9700

12/23/2014

Contents

Cover image by Ronald Martinez/Getty Images

Chapter 1: Simple Linear Regression

1. Topic

Recently I watched an episode of Real Sports on HBO that documented that story of Vivek Ranadivé, a Silicon Valley billionaire who recently purchased the NBA's Sacramento Kings. The episode described how Ranadivé, a poor Indian immigrant, became one of the most important people in professional basketball. The most ingesting part to me was how Ranadivé was introduced to the game of basketball. This story starts in 2009 when, in an effort to spend more time with his daughter, Mr. Ranadivé volunteered to coach his 12 year old daughter's junior high basketball team even though he had never touched a basketball in his life.

Ranadivé, an MIT educated engineer and Harvard MBA, decided that he wanted to win as many games as possible. He started by pouring over the statistics of basketball hoping to find insight on how to coach his team. He decided that turnovers are the statistic most correlated with wins. In order to maximize this statistic, Mr. Ranadivé coached his team to always play full press defense, which involves defending a team across the entire court. Mr. Ranadivé stated that as a result, that team generated an extremely high number of turnovers and as a direct result, never lost a game while he was the coach.

As someone who is a fan of both basketball and statistical analysis, I found myself inspired by this innovative use of statistical analysis by a sports novice. For this project, I will further analyze the statistical relationship between winning percentage, turnovers, shooting percentage and other basketball statistics in order to evaluate Ranadivé's hypothesis that turnovers are the most important statistic for winning basketball games.

2. Data Source

For the purpose of this study I will analyze the 2014 NCAA men's basketball regular season. The NCAA has 351 teams that all play between 30 and 32 games a year. I chose the NCAA over the NBA because there are more teams and thus more data to regress without having to take data from multiple seasons. I also hope to use some part of the model I develop to impress my peers during that time of year when college statistics experts get their 15 minutes of fame, March Madness.

The input has been taken from http://www.sports-reference.com/. Specifically the advanced statistics section of the school index page for the 2013-2014 season: http://www.sports-reference.com/cbb/seasons/2014-advanced-school-stats.html

3. Variables

The y-variable winning percentage (WL) will be regressed against:

- **Effective Field Goal Percentage (eFG**): this statistic adjusts for the fact that a 3-point field goal is worth one more point than a 2-point field goal.
- **Total Rebound Percentage (TRB):** An estimate of the percentage of available rebounds a team grabbed.
- **Block Percentage (BLK):** An estimate of the percentage of opponent two-point field goal attempts blocked by the players of that team.
- **Turnover Percentage (TOV):** Opponents number of turnovers per 100 possessions.

4. Data View

Here is a sample of the first 20 out 351 rows of data with all x variables included

School	WL	eFG	BLK	TRB	TOV
AbileneChristian	0.355	0.515	6.8	50.2	18
AirForce	0.4	0.504	10.4	48.8	15.7
Akron	0.618	0.503	12.2	51	16.9
AlabamaAM	0.467	0.467	12	47.1	17.5
AlabamaBirmingham	0.581	0.458	9.2	55	13.6
AlabamaState	0.594	0.465	9	49.5	18.6
Alabama	0.406	0.499	12.2	48.5	16.7
AlbanyNY	0.559	0.488	7.9	52.5	16
AlcornState	0.387	0.459	9.6	47.3	16.5
American	0.606	0.558	11.2	50.6	17.2
AppalachianState	0.3	0.459	6.7	50.6	16.6
ArizonaState	0.636	0.524	15.6	48.4	14.7
Arizona	0.868	0.517	11.1	55.1	16.7
ArkansasLittleRock	0.469	0.47	9.3	50.8	16.3
ArkansasPineBluff	0.419	0.471	6.9	46.6	21.3
ArkansasState	0.594	0.523	8.4	48.5	16.2
Arkansas	0.647	0.501	13.4	47.7	19.7
Army	0.484	0.509	11.9	48.9	17.3
Auburn	0.467	0.489	12.2	49.3	16.3

5. A Simple Regression Model

(a) This paper uses numerous examples of linear regression to examine the relationship between winning games and various basketball statistics. In order to test the waters before jumping in the regression pool, we'll step through a simple example of regression by examining a single variable, effective field goal percentage (eFG), against the statistic we are trying to influence, win/loss percentage(WL).

(b) The simple linear regression model without variables specific to this study is:

$$Y_x = \beta_0 + \beta_1 x + \varepsilon$$

This model has some similarities to the slope intercept-formula used to determine the slope of a line given two points (y = mx+ b). To demonstrate this, I'll break down this model down by variable and then rewrite the model using variables applicable to this study

Y_x represents the y variable which we are trying predict. Since the goal of a basketball game is to win, this is our primary statistic.

β_0 represents the y intercept of the model. This is the value of win percentage when x variable is equal to 0. This is similar the value b in y = mx+ b.

β_1 and x are multiplied to determine a single value. β_1 represents the slope coefficient (m) and x represents a variable (in this case shooting percentage) that we are regressing against y.

ε represents variance in the model. You'll notice that this value is not in the simple slope intercept formula but it is included in this model. This is because we are examining two variables that have a statistical relationship. The inclusion of this symbol will be discussed more in part c of this section.

Given that we trying to examine the effect of effective shooting percentage on winning percentage, we can plug in our specific variables into this model while leaving the coefficients as they are in the generic version of the model.

$$WL = \beta_0 + \beta_1(eFG) + \varepsilon$$

(c) Although a comparison between the slope intercept and simple regression models was made in the previous section of this paper, there is one important difference that should be pointed out before progressing further: the slope intercept for examines a functional relationship between two variables and the simple regression model examines a statistical relationship between two variables.

A statistical relationship between two variables means that there will be different results in y values even if we examine data that has the same x value. This is represented by the inclusion of ε in the regression model. Later we will use the expected value function to determine an statistical relationship in regression.

(d) In order to change this equation from one that states a statistical relationship into one that states a functional relationship we can use the expected value function.

In the most generic terms (for regression) the expected value can be written as:

$E(Y|x) = \beta_0 + \beta_1 x$

When comparing this model to the simple linear regression model, you'll quickly notice that E(Y|x) has replaced Y_x and ε is no longer present. This is because E(Y|x) represents the combination of these two concepts in the model as expected value of Y given x. The expected value function is used instead since it is a function for predicting a random variable. This value is also important in terms of the bell curves mentioned in section c of this paper since the expected value for each x value represents the center of each bell curve.

When the specific variables for this model are plugged into this formula, the expected value of WL given a certain shooting percentage can be represented as:

$E(WL|eFG) = \beta_0 + \beta_1(eFG)$

(e) To understand the sample slope and other concepts in this paper, it is important to understand the story of many possible samples since it helps to explain the purpose of the formulas that are used. The story of many possible samples stats with a parent population. Imagine if we took all possible win-loss and field goal combinations for every team and wrote each one on a marble. Every time we finished writing, we would throw these marbles into a large container such as a 55 gallon drum. This bucket represents our "parent population" from which we start the story. It is called this because this population will be used to create other "daughter" populations by repeatedly resampling to create new populations. Next, we begin pulling combinations of parent population for a certain value of n marbles. We'll then find the sample slope for each of these combinations of n marbles which we then write on another marble. We want to keep of these "daughter population" marbles as well so we start by throwing these marbles into another 55 gallon drum but we soon find that we need a much larger container, perhaps an empty swimming pool, as there a zillion daughter population marbles.

Once we have our pool full of daughter population marbles, we could then observe a few interesting things about this new daughter population. Mainly that the average of all of the values in this sample slope pool is equal to the slope of our parent population. The daughter population also has a standard deviation but it is different than the parent population because sample slope is comprised of the average slopes of all marbles in the sample instead raw data points.

6. A Fitted Simple Regression Model

(a) When the win-lose percentages and effective shooting percentages are entered into SAS and regressed using the proc reg procedure, the following output is provided. The two most important parts of the output (the intercept and the slope) have been highlighted in red:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.21847			
eFG	1	3.46552	0.26436	13.11	<.0001

Since this type of output from SAS will be used throughout this paper it is important to understand its contents. The first item that we'll examine is the parameter estimates, which have been highlighted in red. The top number is the estimate for the intercept (b_0), which is the expected value of Y when x is 0. The bottom number represents the slope the in model (b_1) which is the amount the expected value of Y changes every time x is increased by 1.

(b) The parameter estimates can be used to create the y-hat equation. The generic y-hat equation is:

$$\hat{y} = b_0 + b_1 x$$

This can be thought of as the expected value equation when applied to a sample of data (i.e not the entire population). This is evident by the Greek symbols and capital letters being replaced with lower case letters.

Since SAS has given provided with the b_0 and b_1 values we can create a y-hat equation specific to our model:

$$\hat{y} = -1.21847 + 3.46552x$$

(c) Like any sample calculation in statistics, the y-hat equation estimates a variable for a population. When a x-value is plugged into the y-hat equation from the previous section, the result is a point estimate of E(Y|x) for the entire population.

(d) The t-test for the slope is an important tool in statistical analysis for determining if there is a significant relationship between two variables. The t-test starts with the null hypothesis which states that β_1, the slope of our model is equal to 0. This is important to determine because if the slope were in fact 0, it would mean that a model has is not effective to use for predicting future values.

This is written as follows, it will be explained in the text following the test.

The hypothesis for the t-test is

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

Test statistic: $\text{t-stat} = \dfrac{b_1 - 0}{s/\sqrt{SSx}} = \dfrac{3.4655}{.26436} = 13.11$

Rejection region: |t-stat| > t-critical value, α=0.05
t-critical value with 349 (z) d.f. = 1.96

Conclusion: null hypothesis is rejected because the |t-stat| = | 13.11 | is greater than the t-critical value of 1.96.

H_0 represents the aforementioned null hypothesis and H_1represents the alternative hypothesis, which is the opposite of the null hypothesis, which is written as a formality.

To conduct the t-test to determine if our slope is in fact 0, we first need to evaluate the error in our model. Fortunately the output form SAS has given us everything we need to perform this exercise. This also provides an opportunity to explain the next two columns of SAS' the regression output, standard error, t value.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.21847			
eFG	1	3.46552	0.26436	13.11	<.0001

Standard Error: This is a squared summation of each point's distance from the y-hat line. A SAS takes care of the leg work of generating standard error for our 351 points and provides us with this value.

T-value (aka t-statistic) is a measure of distance that is calculated by the following equation

$$= b_1 / \text{s.e.}\, b_1$$

This value represents a point on a normal distribution that will be discussed further in the next section.

And finally (but not show on our SAS output) we need to determine the t-critical value, which is the point at which we reject or not reject the null hypothesis that our slope is in fact 0.

To determine the t-critical value we can use student table to look up a value that is and find the vale which is specific to two factors.

1. Our degrees of freedom, which is n-2 = 349. Two is subtracted because our model has two coefficients, the intercept and x.

2. The level of certainty we are trying to test, which in this case is 95%.[1]

By looking on a t-critical value chart [2] we can determine that the t-critical value is 1.960. Note that after 100 the difference made by increasing n is insignificant so the value z is used for n values larger than 100.

Level	50%	80%	90%	95%	98%	99%
Two Tail	0.5	0.2	0.1	0.05	0.02	0.01
df = 1	1	3.078	6.314	12.706	31.821	63.657
2	0.816	1.886	2.92	4.303	6.965	9.925
…	…	…	…	…	…	…
90	0.677	1.291	1.662	1.987	2.368	2.632
100	0.677	1.29	1.66	1.984	2.364	2.626
z	0.674	1.282	1.645	1.96	2.326	2.576

To bring this all together to finally execute the t-test, we take the absolute value from our t-statistic from our SAS output and compare it to our t-critical value. We find that $|13.11| = 13.11 > 1.96$

[1] Because we are examining a two sided normal distribution it is important to remember that we want to find the area inside the two tails of the normal distribution.
[2] This was copied from https://people.richland.edu/james/lecture/m170/tbl-t.html on 12/19/2014

In short this means that the null hypothesis is rejected because our sample t-statistic is larger than our t-critical value. This means that the slope is not equal to zero nor is it close enough to zero explained by statistical variance and as a result, the relationship between shooting percentage and winning percentage is statistically significant.

(e) Although it is likely sufficient to provide a description of the t-test as brief as the "t-statistic is far from zero so we reject the null" among those familiar with regression, there is a bit of background about the test that the reader should be familiar with which will be discussed in the next three sections.

First it is important to understand what is meant by the term "far" in the phrase "far from zero". The t-statistic is measured on a normal distribution in terms of standard deviations. This can be pictured as looking at a map to see the lay of the land (e.g. the normal t-distribution). Every good map has key that marks relative distance so the viewer can understand the proportion of what they are looking at. Instead of miles or kilometers, distance denoted in the key of the t-distribution map is measured in standard deviations. Standard deviations are used instead of terms relative to the x and y values specific our data since they are normalized across all data sets.

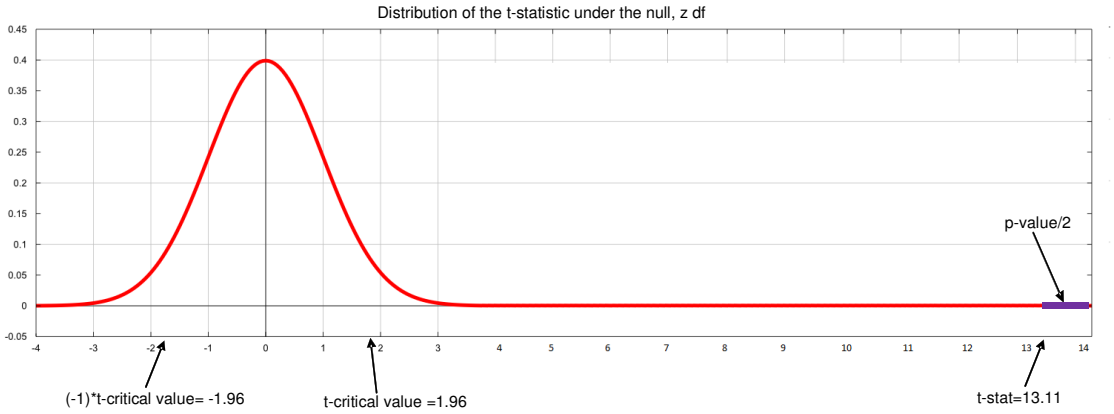
The t-critical value is a value that is determined to describe when a value is too far away to be considered part of a data set. Staying with the map analogy, it's like if you took a geometric drawing compass and drew a circle around the place where your car is parked. The radius of this circle represents the farthest distance you could drive your car with only a single tank of gas 95% of the time. If your friend where to borrow your car, and dive it from your drive way to a point far outside of this circle, let's say two towns over from the end of the circumference, you would not believe him or her and thus could use your knowledge of your car to reject your friend's "hypothesis" and that your friend had used a different car to get as far as they did. This is similar to rejecting the null hypothesis based on a t-critical value.

(f) The story of many possible samples also applies to the t-statistic. Imagine if we took our original parent drum of marbles and instead of writing the sample slope on our daughter population marbles, we would instead write the t-statistic for each sample write it on another marble and throw it into an empty swimming pool. Just like the daughter population of slopes these "daughter t-statistic population" has its own mean and standard deviation.

However the t-statistic goes one step further since we are executing a hypothesis test. Once we have all of these t-statistics, we can compare them to the template normal distribution that we find in the t-table in the previous example. This template normal distribution has a mean and mode at 0 with the two tails extending out positive and negative numbers. We then compare the

template to our daughter population which has its own population mean and standard deviation. If there is no way that that two could be from the same population we then reject the hypothesis that our slope is equal to zero.

(g) The positive t-distribution for the data is graphed as below. As you can see, the t-stat of 13.11 is nowhere near the area considered "under the null", which is between -1.96 and 1.96. To improve visibility I have chosen not to show the negative side of the distribution. It does however exist and would be a mirror image of the graph below extending out to -13.11. A quick visual inspection also shows that is consistent with the p-value of <.0001 given by SAS since that almost none of the population outside or our t-statistic falls into the area under the null.

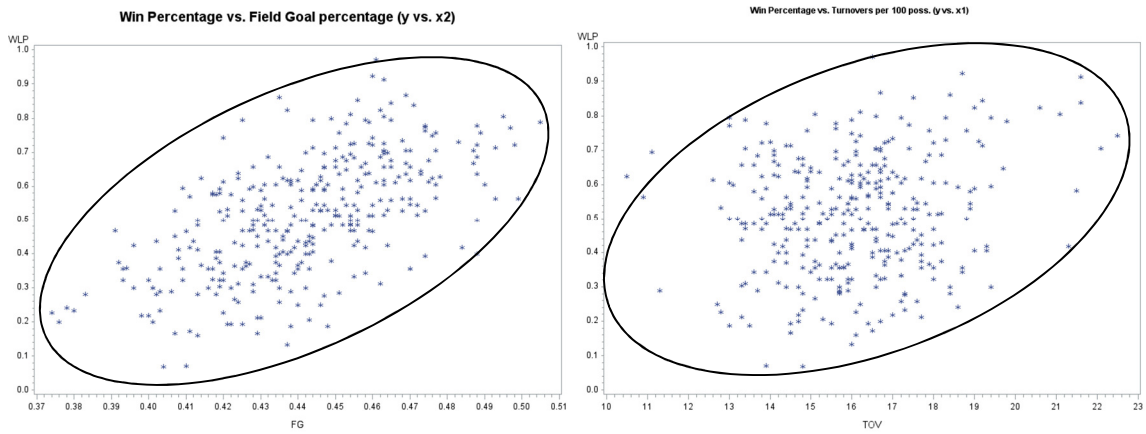


Chapter 2: Regression Model with Two Regressors

1. Scatterplots

Although SAS provides a number of useful calculations for determining statistical significance, a visual inspection of data is still important for determining a number of useful types of information about the data. We'll start by showing a scatter plot for two of the x variables (eFG and TOV) graphed against or y value (WL)

These two scatter plots were created using the **`proc gplot`** in SAS:



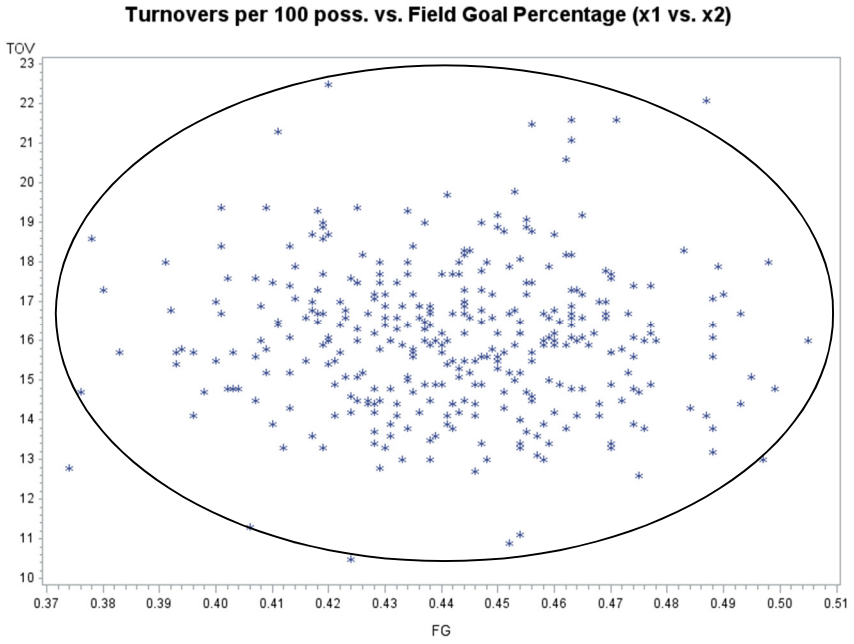
2. Analysis of Scatterplots

Both scatter plots show positive correlation with the y variable which means that a regression of these variables on y will likely have positive slope and statistical significance. This is shown by the slated oval that surrounds the points in each dataset.

In addition to looking for correlation we can also look to see if there are any issues with our variables that would require us to remove data points or change our choice in variables. First we examine the overall shape the points form and see that nearly all of them fit into the ovals that I have drawn on each graph. This fit mean that there is no sign or curvature, heteroscedasticity or skewness. If this were the case we'd see the formation curve or have a funnel like formation at either side. Additionally there are no leverage points or outliers in the group. This type of scatter seems to be typically among sports data.

3. Graphical Inspection for Collinearity

In addition to plotting x variables against y, it's also a good idea to plot our x variables against each other. Below is plot of turnovers vs. field goal percentage.



A quick visual inspection of plot shows that does not appear to be any collinearity in a comparison of x1 and x2. This is evident because I can't place a titled oval around the majority of the points like I could in the two charts on the previous page. Instead the points fall into a shape that resembles an egg that has been laid on its side. Initially I thought that there might be some relation between these variables because a team could use the opportunities created by a turnover deep in an opponent's territory to set up high percentage shots underneath the basket. This however does not seem to be the case.

4. The Linear Regression Model

To take the next step in learning how regression can be used to determine which factors are most important in terms of winning games, let's add a second variable to our model from Chapter 1:

$$WL_i = \beta_0 + \beta_1 eFG_i + \beta_2 TOV_i + \varepsilon_i$$

(a) WL_i can also be written as WLX denoting a subpopulation of values at a particular TOV and FPG (e.g. all teams that had a TOV of 5 and FPG of 70%) This subpopulation will have a normal distribution with its own mean

and its own variability. Both of these properties are discussed further in parts b and c of this section

(b) The terms eFG and TOB on the right-side are related to $E(Y_x)$ when this term is written as E(WL|eFG,TOV) i.e. the expect value of WL when eFG and TOV are equal to a certain value. When values are plugged into the eFG and TOV values then the expected value can be determined. The expected value is the mean of the subpopulation of values at these two eFG and TOV values. In our model, this resents the center of the bell curve for this sub population.

(c) The terms on the right-side are related to V(YX) in that variance of the subpopulation is represented by the inclusion of ε. Although the two variables in the subpopulation may not change the outcome will. This is because the expected value is random variable. ε represents the variance or width of the bell curve that exists at every sub population.

5. SAS Output for the Fitted Model

Now that a visual inspection of the data is complete, we can run our data through the SAS reg proc to get the following information about out data. Each of these tables will be discussed in detail in section 6 of this chapter.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	4.05968	2.02984	114.02	<.0001
Error	348	6.19514	0.01780		
Corrected Total	350	10.25482			

Root MSE	0.13342	R-Square	0.3959
Dependent Mean	0.51027	Adj R-Sq	0.3924
Coeff Var	26.14795		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.65462			
eFG	1	4.25959	0.29397	14.49	<.0001
TOV	1	0.01760	0.00386	4.56	<.0001

6. Analysis of Output

(a) The t-tests

H_o: $\beta_1=0$;
H_a: $\beta_1\neq 0$

Test statistic: $t\text{-stat} = \dfrac{b_1 - 0}{s/\sqrt{SSx}} = \dfrac{4.25959}{.29397} = 14.49$

Rejection region: |t-stat| > t-critical value, $\alpha=0.05$
t-critical value with z d.f. = 1.96

Conclusion: null hypothesis is rejected because the |t-stat| = | 14.49 | greater than the t-critical value of 1.96.

H_o: $\beta_2=0$;
H_a: $\beta_2\neq 0$

Test statistic: $t\text{-stat} = \dfrac{b2 - 0}{s/\sqrt{SSx}} = \dfrac{.0176}{.00386} = 4.56$

Rejection region: |t-stat| > t-critical value, $\alpha=0.05$
t-critical value with z d.f. = 1.96

Conclusion: null hypothesis is rejected because the |t-stat| = | 4.56 | greater than the t-critical value of 1.96

As mentioned in chapter 1, the t-test is a statistical relationship exists between the y value and the variables in the model. The null hypothesis assumes that the slope of both coefficients is 0 meaning that no statistical significance exists. A check of the SAS table shows that our t-values of 14.49 and 4.56 are much higher than our critical t-value of 1.96.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.65462			
eFG	1	4.25959	0.29397	14.49	<.0001
TOV	1	0.01760	0.00386	4.56	<.0001

The p-value, or the percentage of t-statistics that fall into a normal distribution of t-statistics (think back to chapter 1 and the story many possible samples) is <.0001. This value is so far away from t-distribution centered at 0 that that there is no reasonable chance that our data is from a population centered at 0.

Thus we can make the statement that we can reject the null hypothesis however we are less testing a single variable so we will examine all variables using the f-test.

(b) The F-test

The F-test is the a test for significance across all variables and is conducted as such

(i) H_o: $\beta_1=\beta_2=0$;
H_1: $\beta_j \neq 0$ for at least one value of j

Test statistic: $\text{F-stat} = \dfrac{\text{SSR}/(p-1)}{\text{SSE}/(n-p)} = \dfrac{4.5968/(2)}{6.1954/(348)} = 114.02$

Rejection region: F-stat > F-critical value, p-1 num. df, n-p den. df, $\alpha=0.05$

Conclusion: the null hypothesis is rejected because the F-stat of 114.02 is greater than the F-critical value of 1.982.

(ii) Everything needed to execute this calculation is in the output provided by SAS. However SAS also provides the p-vale for the f-test as <.0001 meaning no significant area under the bell curve of a t-distribution close to 0 when evaluating our given f-value.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	2	4.05968	2.02984	114.02	<.0001
Error	348	6.19514	0.01780		
Corrected Total	350	10.25482			

(c) The $\hat{y}$ -equation

The $\hat{y}$ equation for my regression model is

$$\hat{y} = -1.65462 + 0.01760(\text{TOV}) + 4.25959(\text{eFG})$$

(ii) Just like in single regression, the $\hat{y}$ equation estimates the expected value of $E(Y_x)$. However since there are now two regressors, our model has a third dimension

7. The *Partial* Regression Coefficient

(a/b) Now that a 2nd regressor has been introduced into the model it's important to understand the idea of the "partial" regression coefficient because variables have a tendency to be correlated with each other. This will be covered in chapter 3 of this paper.

(c) In chapter 3 the partial regression coefficient for x_2 when x_1 is in the model will be found by regressing y* on x_2*, where x_2* is the residuals from regressing x_2 on x_1 and y* is the residuals from regressing y on x1.

8. R-square

(a) In order to further evaluate our model we'll examine the calculation of both R-squared and adjusted r-squared. The r-square for this model is .3959.

(b) This was computed by taking 1 minus Sum of Squares for the errors over the Sum of Squared Error for the Corrected Total and displayed by SAS in the regression output.

Source	DF	Sum of Squares
Model	2	4.05968
Error	348	6.19514
Corrected Total	350	10.25482

1 - 6.19514/10.25482 = 3.959

Root MSE	0.13342	R-Square	0.3959
Dependent Mean	0.51027	Adj R-Sq	0.3924
Coeff Var	26.14795		

(c) R-square is a measure of the difference between the squared errors that result from horizontal line at x-bar (e.g. if our slope was 0) and the regression line proposed by our model. R-square compliments the t and f tests, which determine statistical significance, by giving a percentage measure of how much better the model is than 0.

R^2 interprets the relationship between X and Y and is an indicator for how much variance is reduced by X when if we were to not consider X. It tells us about the association between X and Y.

(d) The naïve interpretation of your R-square would be to interpret it as **explaining** the level of variance as a result of X. There are plenty of examples where R^2 could be used to say that variable X explains the variance when in fact X and Y have no relationship outside of the R-squared.

9. Adjusted R-square

(a) The value of the adjusted R-square is 0.3924

Root MSE	0.13342	**R-Square**	0.3959
Dependent Mean	0.51027	**Adj R-Sq**	0.3924
Coeff Var	26.14795		

(b) Adjusted R-square was computed from the output using the following equation

adj R^2 = [SST/(n-1) - SSE/(n-k-1)] / [SST/(n-1)]

And the following fields from the output:

Source	DF	Sum of Squares
Model	2	4.05968
Error	348	6.19514
Corrected Total	350	10.25482

adj R^2 = [10.25482/(350) – 6.19514/(348)] / [10.25482/(350)]

adj R^2 = .3924

(c) Adjusted R-square value is similar to r-squared however it is a measure of variance. It is a measure of how much variance is reduced when x is introduced into the model.

Chapter 3: Partial R-square

(a) In order to understand x2's effect on the model independent of the other independent variable x1, it is necessary to determine the partial coefficient of x2. This will describe the effect x2 has in decreasing uncertainty independent of any covariance that variable has with x1 or y. Another way of wording this would be to say "partial correlation of y and x2 given x1"

Below are the steps required to sweep out *x1* from *y*, sweep out *x1* from *x2* and evaluate the residuals.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Squared Partial Corr Type II
Intercept	1	-1.65462	0.14529			
x2	1	4.25959	0.29397	14.49	<.0001	0.37629
x1	1	0.01760	0.00386	4.56	<.0001	0.05638

(b) When I regressed y on x1, $SSE_{x1} = 9.93271$

Source	DF	Sum of Squares	Mean Square
Model	1	0.32211	0.32211
Error	349	9.93271	0.02846
Corrected Total	350	10.25482	

And When I regressed *y* on *x1* **and** *x*, $SSE_{x1\&x2} = 6.19514$

Source	DF	Sum of Squares	Mean Square
Model	2	4.05968	2.02984
Error	348	6.19514	0.0178
Corrected Total	350	10.2548	

So if $R^2_{x2|x1} = \{ SSE_{x1} - SSE_{x1\&x2} \} / SSE_{x1}$

$= \{.993271-6.19514\}/.993271$

$= 0.376289$

from a:

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Squared Partial Corr Type II
Intercept	1	-1.65462				
x2	1	4.25959	0.29397	14.49	<.0001	0.37629
x1	1	0.01760	0.00386	4.56	<.0001	0.05638

c) It is sensible that a and b would match because the equation $R^2_{x2|x}$ is removing (sweeping) the effect that one variable has on the has on reducing variance when its regression line is compared to the average of the y vales.

Chapter 4. Polynomial Regression

1. Simple Polynomial Regression

(a) In addition to linier analysis, polynomial regression can be used to create models. In order to prepare the data for a simple polynomial regression using SAS, the data set will need to include x^2 . SAS reads in data in a sort of repeating loop, where each row is read into memory in sequential. This makes adding x^2 a fairly simple operation.

Considering that the these two columns are being pasted in from an excel doc,

```
Input x1 y;
```

x^2 can be added as a column to the data set by using the following data step:

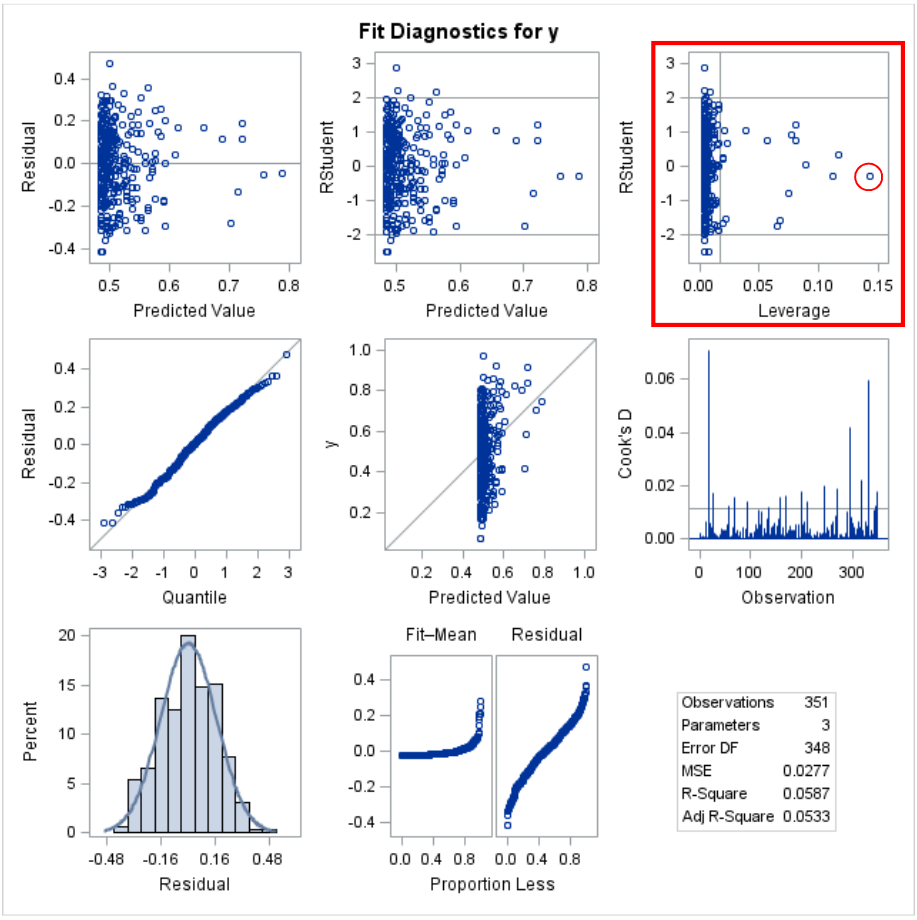
```
x1sq = x1**2;
```

as SAS imports each row, the value of x^2 will be appended to the end of each row, forming a new column.

(b) Below is the model and the resulting polynomial regression model and SAS output.

$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$.

(c) Leverage points, (points that have extremely high values on the x axis) have an even greater potential to cause over fit in a model using polynomial regression. We'll perform a quick review of our data using the diagnostics table below. The highest leverage point, (the point farthest to the left on the horizontal axis), has an x value of .14 (as seen above highlighted in red. This indicates that the data doesn't not have any influential points which need to be investigated further. If this value was closer to .8 or .9 then it would have too much influence on the model and should be evaluated for having too much influence on the model.



(d) To understand how leverage is calculated, one must first be familiar with the **H**-matrix. The H-Matrix is a table that is created to observe residuals (i.e. the difference between the observed value and the predicted value) and compare them to the observed values. Leverage values are computed by taking the diagonal values from the **H**-matrix. In the **H**-Matrix the diagonal values, starting the top left corner and going down and to the right have a special property (these values are also known as h_{ii} values with the subscript i's representing points in the matrix that have the same row and column number as highlighted below): The **H**-matrix will always be square so each observation will have a corresponding h_{ii} value.

1,1	1,2	1,3	1,4	1,5
2,1	2,2	2,3	2,4	2,5
3,1	3,2	3,3	3,4	3,5
4,1	4,2	4,3	4,4	4,5
5,1	5,2	5,3	5,4	5,5

These diagonal values are important because they tell us about the variance of the residuals, the measure of distance from the prediction. If the value is far from the prediction line and it's corresponding x value is far from the other observations, that sample can over influence the model. This is because points with high leverage can pull the prediction line away from the majority of points.

e) Below is the SAS output resulting from the introduction of a polynomial variable.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.60216	0.30108	10.85	<.0001
Error	348	9.65266	0.02774		
Corrected Total	350	10.25482			

Root MSE	0.16655	R-Square	0.0587
Dependent Mean	0.51027	Adj R-Sq	0.0533
Coeff Var	32.63891		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.58167			
TOV	1	-0.14883	0.05222	-2.85	0.0046
TOVsq	1	0.00505	0.00159	3.18	0.0016

(i) Below are the t-test and f-tests for this data set

The t-tests

H_o: $\beta_1=0$;
H_a: $\beta_1\neq 0$

Test statistic: $t\text{-stat} = \dfrac{b_1 - 0}{s/\sqrt{SSx}} = \dfrac{-.14883}{.05222} = -2.85$

Rejection region: |t-stat| > t-critical value, $\alpha=0.05$
t-critical value with z d.f. = 1.96

Conclusion: null hypothesis is rejected because the |t-stat| = | -2.85 | greater than the t-critical value of 1.96.

H_o: $\beta_2=0$;

H_a: $\beta_2 \neq 0$

Test statistic: t-stat = $\frac{b2 - 0}{s/\sqrt{SSx}} = \frac{.00505}{.00159} = 3.18$

Rejection region: |t-stat| > t-critical value, α=0.05
t-critical value with z d.f. = 1.96

Conclusion: null hypothesis is rejected because the |t-stat| = | 3.18 | greater than the t-critical value of 1.96

The F-test

The F-test is the a test for significance across all variables and is conducted as such

(i) H_o: $\beta_1=\beta_2=0$;
H_1: $\beta_j \neq 0$ for at least one value of j

Test statistic: F-stat = $\frac{SSR/(p-1)}{SSE/(n-p)} = \frac{.60216/(2)}{9.65266/(348)} = 10.85$

Rejection region: F-stat > F-critical value, p-1 num. df, n-p den. df, α=0.05

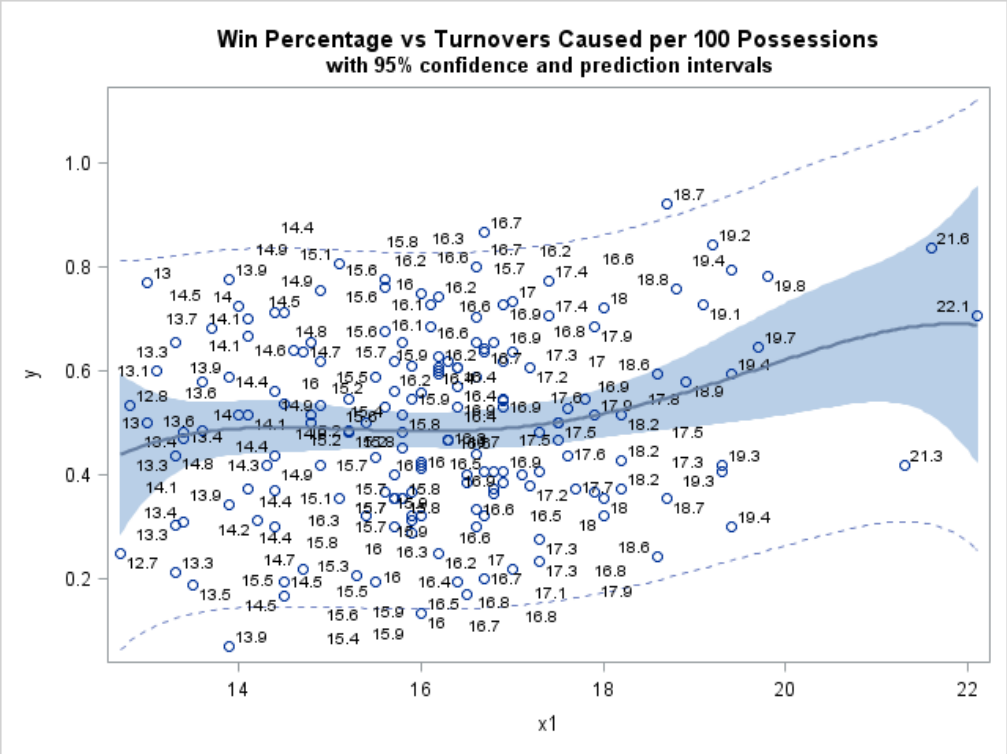
Conclusion: the null hypothesis is rejected because the F-stat of 10.85 is greater than the F-critical value of 1.982.

(ii) If the P values are less than .05 we cannot reject the null hypothesis.

(f) The chart below was created using the following command:

```
reg=(degree=4 clm cli nogroup);
```

Note that I cut my data in half since the line generated by SAS was barely visible with all 351 points



(g) The higher-order polynomial fit would like produce reliable results for new x-values. The data is groped closely so there are not any wild swings in the line to compensate points at the ends of the model.

2. Multiple Regression with a Dummy Variable and an Interaction Term

(a) When I started looking at basketball statistics for men's college basketball, I was curious if there would be a difference between the statistics in teams from top athletic conferences such as the ACC, Big Ten, Pac 12, SEC and the Big 12 who typically recruit top players, have much larger athletic department budgets and wide spread fan bases. Since "does this school belong to a big conference?" is a yes/no question, and a not numeric value, I can't just plug it into SAS and get a regression model.

Instead we can make use of a dummy variable, which I will call IsBigConf, in order to observe this relationship using regression. For my model, schools that belong to one of the 5 "power conferences" will have a value of IsBigConf value of 1 and schools from the remaining 27 conferences will have a value of 0 in this column.

(b) Since conference designation isn't information I can derive from my data set, I am not able to create a new column for my dummy variable using a data step in SAS as demonstrated in the lecture notes. Instead I needed to manually review my list of 351 schools to determine who belongs to which conference. Fortunately, I have

a near encyclopedic knowledge of the top 5 power conferences in college sports so I was able to make quick work of the task.

Here is a subset of the data to demonstrate my adjustment:

School	WL%	TOV%	eFG%	IsBigConf
Arizona	0.868	16.7	0.469	1
Arizona State	0.636	14.7	0.451	1
Arkansas	0.647	19.7	0.441	1
Arkansas State	0.594	16.2	0.444	0
Arkansas-Little Rock	0.469	16.3	0.425	0
Arkansas-Pine Bluff	0.419	21.3	0.411	0

c) Interaction terms are an interesting way to observe important information about the relationship between the x values that are being used in a regression model. In order introduce an interaction term, we add a third beta value to our module and multiply it be the product of the original 2 x values:

$$\text{WinPerct}_i = \beta_0 + \beta_1 \text{FG}_i + \beta_2 \text{TOV}_i + \beta_3 \mathbf{FG_i * TOV_i} + \varepsilon_i.$$

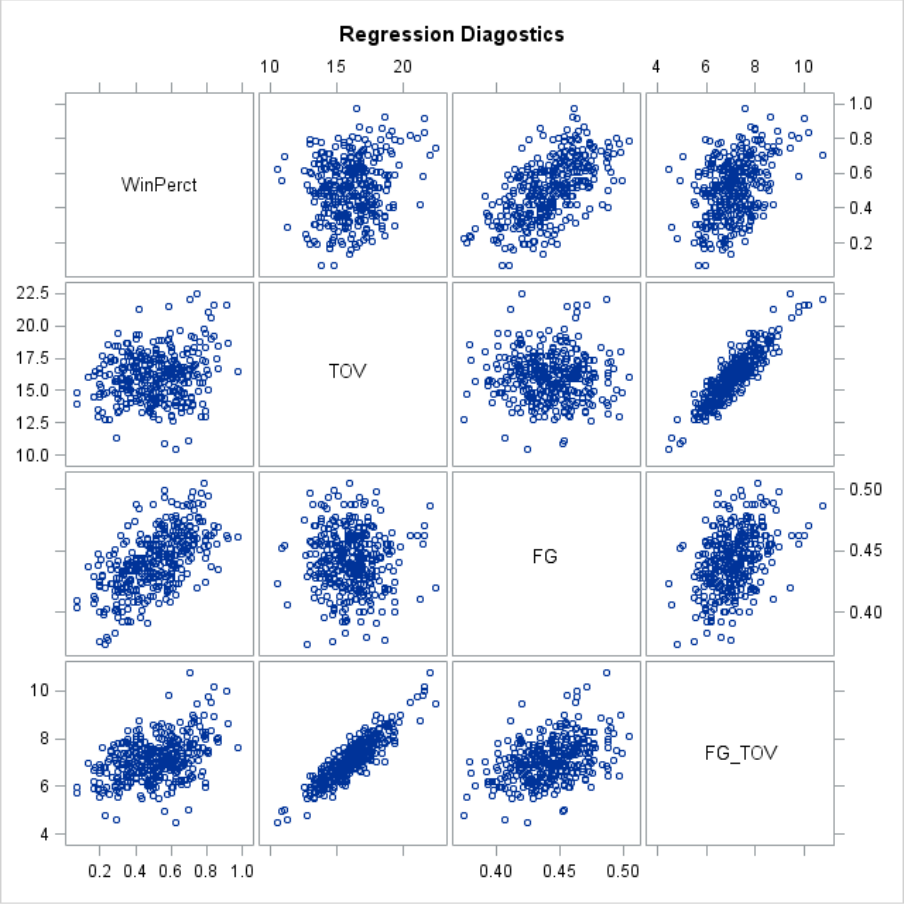
Since the interaction term can be derived using observations already present in my dataset, I can easily create the this term by using a SAS data step. As mentioned before, SAS reads in information row by row so I can create a new column as each record is read into memory.

Considering that these three columns are being pasted in from an excel doc:

```
Input WinPerct FG TOV;
```

The interaction term can be added as a column to the data set by using the following data step that is placed just above my data lines in SAS.

```
FG_TOV = FG*TOV;
```



d) In order to observe the effect that the interaction variable has on the response surface (e.g. our expected value for multiple regression), we should look at the bottom row. The most interesting is the TOV vs FG_TOV chart (2nd from the left) which has a correlation that is fairly close to 1. This means that there is a level of collinearity that is much higher than in TOV and FG, which displays very little correlation.

e) Inclusion of the dummy variable resulted in a slight increase of .02 on R-Square when added to the model vs the R-Square of regressing Win Percent on FG alone. Not nearly as high I expected given the disparity in budgets between large and small college sports programs.

f) As seen below, the interaction variable has a low t-value so it may be statistically significant. Also its inclusion in the model did increase R-Square. However it's almost perfect correlation with the variable TOV is grounds to not include one of those variables.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.59077			
TOV	1	0.01362	0.06920	0.20	0.8441
FG	1	4.11552	2.51944	1.63	0.1033
FG_TOV	1	0.00898	0.15590	0.06	0.9541

Too low to be significant (under the null)

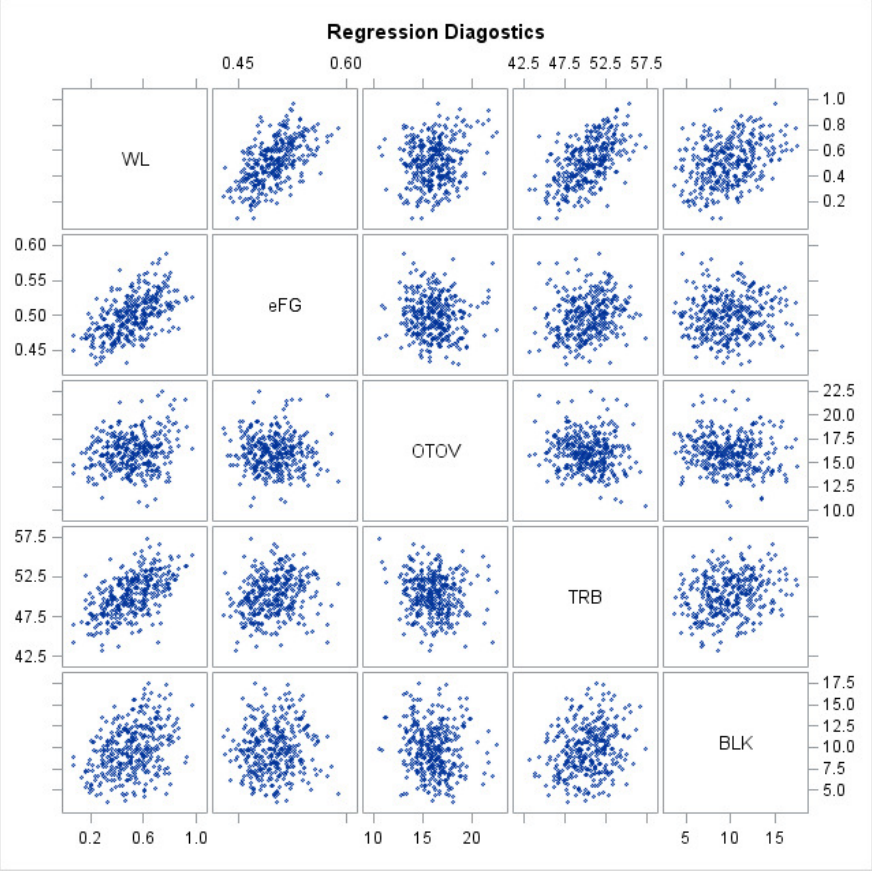
Chapter 5: Model Selection

1. Best Subsets Model Selection

(a) Now we can bring in all variables into the model for evaluation:

- **eFG**: effective field goal percentage
- **OTOV**: opponent turnovers per 100 possessions
- **TRB**: Total rebound percentage
- **BLK**: Percentage of blocked shots

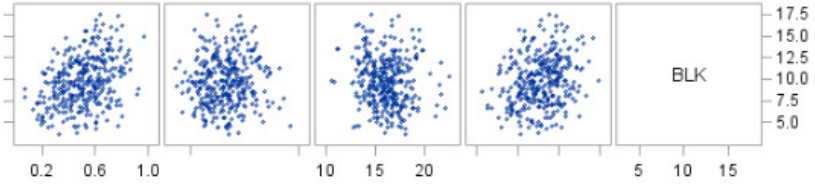
We start by plotting all of our variables against our y-variable and each other to form the matrix of scatter plots below:



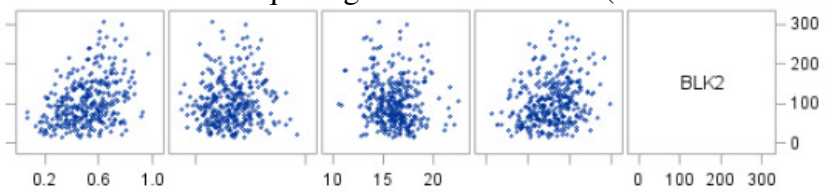
(b) There are not any glaring examples of leverage points, curvature, collinear x variables or heteroscedasticity. This is likely because I have the luxury of having data that has been converted to percentages or divided per 100 possessions. Although I did not find any problems I still wanted to see the effect that transformation would have on my data. So for the sake of experimentation I manipulate my blocks per 100 shot attempts variable since it seemed to have the least amount of correlation.

For the sake of understand more about what transformations could be done to our data if we did not have the luxury of such well-behaved data.

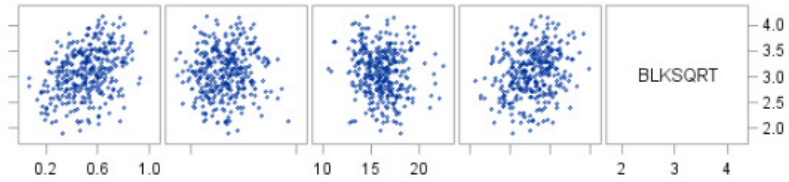
Here is how it looks before transformation:



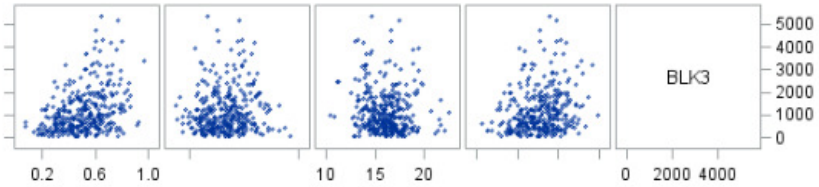
Here is blocks after squaring the values in SAS (`BLK2 = blk**2;`)



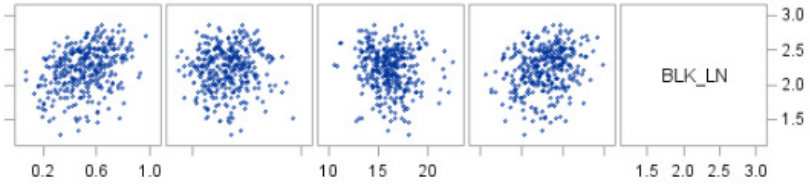
Here are the values after taking the square root in SAS (`BLKSQRT = sqrt(BLK)`)



Here are the values cubed in SAS: (`BLK3 = blk**3;`)

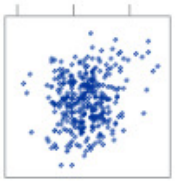


Excel's Natural log function applied

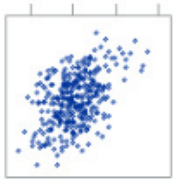


I decided to stick with my original data since taking the square root and natural log didn't change the scatter and added exponents introduced heteroscedasticity.

I did however make a change to one of my variables because I believe it provides a more accurate representation of the hypothesis I am trying to test (e.g. turnovers are the most important statistic in basketball). After running my data the first time, I was expecting to see a greater correlation between my y variable and TOV (opponent turnovers per 100 positions). While there was some correlation visible plotted against y, it didn't seem to be what I expected.

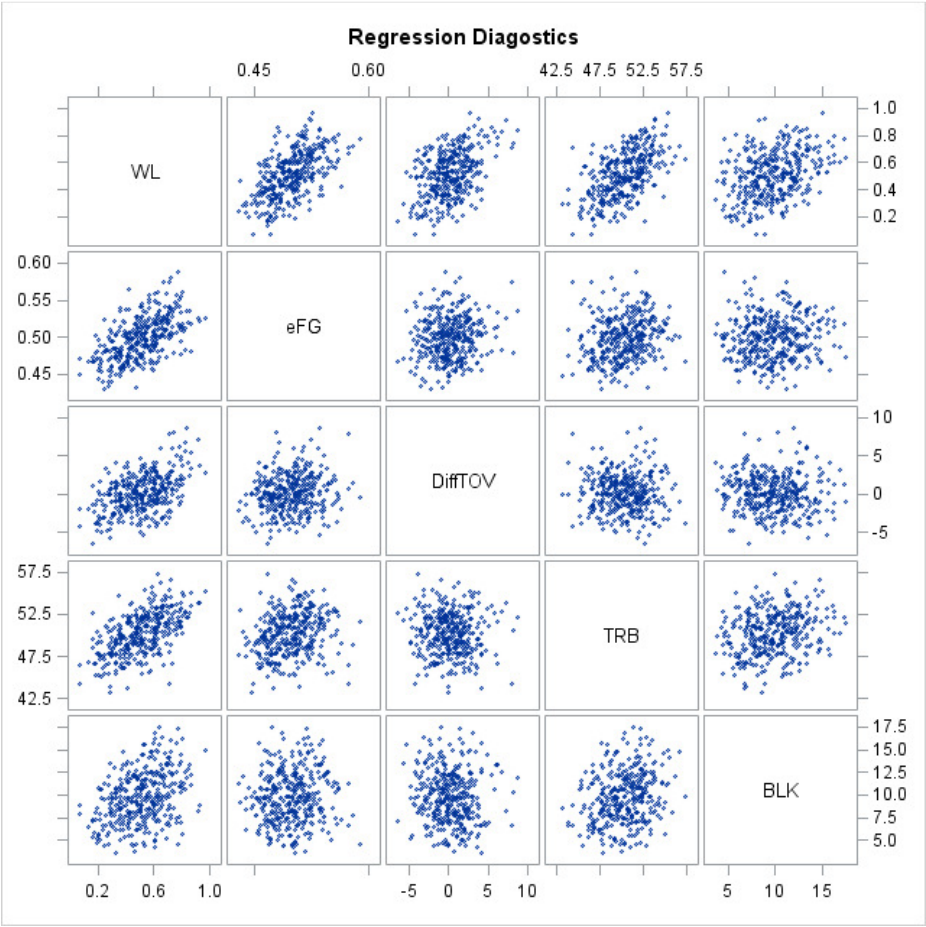


After giving it some thought I decided that opponent turn over percentage would not be a very good indicator if a team that makes their opponent turn the ball over is prone to turning over the ball to their opponent as well. Since I have access to both each team's turnovers over per 100 possessions and the number of times their opponents turn the ball over per 100 possessions rate I decided that getting the difference between these two numbers would be a better indicator of the effect of turnovers on win percentage. The results were more in line with what I was expecting. To the right is opponents turnovers per 100 possessions plotted against y

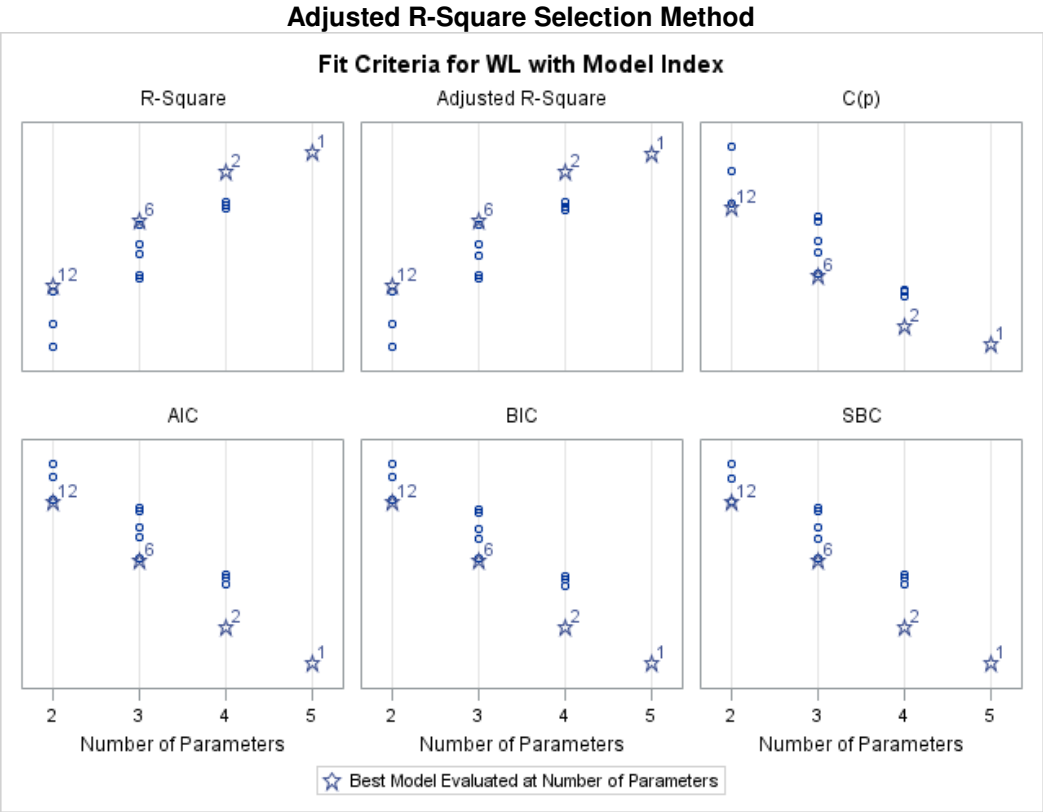


Here is the differential between team turnovers per 100 possessions minus opponents turnovers plotted against y. As you can see there is a tighter positive correlation.

Here is my final set of regression diagnostics with the turnover differential replacing opponent's turnovers. Overall I like it because all of the variables should some correlation with my y variable (WL) and show no show stopping correlation with the other x variables.



(c)



All four of the models (AIC, BIC, SBC) clearly achieve their minimum criterion value when all 5 parameters are included. This means that all 5 variables should be are significant and should be included in my final model. This makes sense because these are well established basketball stats that are published on every sports website and newspaper. If I was delving into a less studied field then I would be more likely to have some variables that should not be included in the model.

One interesting thing to point out is the best model in the column with 3 parameters. The model with index #6, which includes variables DiffTOV, TRB has the lowest value. The absence of eFG, which I understood to be the best variable, is not included here. Somehow these two variables are greater than the sum of their parts. I will examine this more in the step-wise section of this paper. The model also shows that the blocks variable has the least influence, which I suspected after looking at earlier plots.

Adjusted R-Square Selection Method

Model Index	# in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
1	4	**0.7742**	0.7768	5	-1756.5	-1737.2	eFG DiffTOV TRB BLK
2	3	**0.7111**	0.7135	101.03	-1670.9	-1655.5	eFG DiffTOV TRB
3	3	**0.6134**	0.6167	251.16	-1568.7	-1553.2	DiffTOV TRB BLK
4	3	**0.5973**	0.6007	275.9	-1554.4	-1538.9	eFG DiffTOV BLK
5	3	**0.5888**	0.5923	288.96	-1547	-1531.6	eFG TRB BLK
6	2	**0.5471**	0.5497	353.09	-1514.1	-1502.5	DiffTOV TRB
7	2	**0.5376**	0.5403	367.63	-1506.9	-1495.3	eFG TRB
8	2	**0.4709**	0.4739	470.48	-1459.5	-1448	eFG DiffTOV
9	2	**0.4343**	0.4376	526.82	-1436.1	-1424.5	eFG BLK
10	2	**0.3675**	0.3711	629.88	-1396.9	-1385.3	TRB BLK
11	2	**0.3549**	0.3586	649.23	-1390	-1378.4	DiffTOV BLK
12	1	**0.328**	0.3299	691.67	-1376.6	-1368.9	eFG
13	1	**0.3148**	0.3168	712.03	-1369.8	-1362.1	TRB
14	1	**0.2046**	0.2069	882.4	-1317.5	-1309.7	DiffTOV
15	1	**0.1272**	0.1297	1002.08	-1284.9	-1277.1	BLK

As shown the model with all variables included performs the best when examined using modern methods. One thing that I found strange is that the AIC and SBC had rather large negative values. After looking at the equations behind these methods, this made sense since my x vales are all decimals and my n value is 351.

(e) The "Kitchen Sink" method will suffice for this model since all variables appear to have statistical significance. The following coefficients will be used moving forward

b0	eFG	DiffTOV	TRB	BLK
-2.84131	3.01857	0.025592	0.021197	0.024416

2. Forward Stepwise Model Selection

(a) The stepwise method brought the following variables in the order of TRB, eFG, DiffTOV, BLK, based on their Partial R-Squared value.

Summary of Stepwise Selection								
Step	Variable Entered	Var. Remove	#Vars In	Part RSquare	Model RSquare	C(p)	F Value	Pr > F
1	TRB		1	0.3622	0.3622	324.098	89.73	<.0001
2	eFG		2	0.2251	0.5873	156.686	85.61	<.0001
3	DiffTOV		3	0.1427	0.7300	51.2589	82.45	<.0001
4	BLK		4	0.0641	0.7941	5.0000	48.26	<.0001

(b) There are several interesting things to point out when comparing the Stepwise to the best subsets results. Both reached the same conclusion of including all 4 variables but they got there in a different way.

- The subsets found eFG as the best single variable method and stepwise ranked TRB as the best.
- DiffTOV, which my original I originally thought was the most important came in 3rd
- Both ranked BLK last, which isn't surprising.

3. Variance Inflation

(a) Variance inflation is the result of including multiple variables that are closely related to each into a model. By including a variable twice, the same data is throwing off the model for the following reasons:

- There will be an extra set of residuals that introduce no new information. This can artificially increase or decrease error measurements an R-squared.

(b) Below is the VIF output for my data. Given that all of the variables are exceptionally close to 1, there is no cause to be concerned that variance inflation has been introduced. If this was true I would have also seen high correlation between x variables in my initial diagnostics, which I didn't. Instead I got a very nice "shotgun" looking dispersion among all of my x-variables.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-2.5496				
eFG	1	2.71415	0.23164	11.72	<.0001	1.03663
DiffTOV	1	0.02841	0.00273	10.42	<.0001	1.02193
TRB	1	0.03074	0.00256	12.03	<.0001	1.10734
BLK	1	0.01643	0.00237	6.95	<.0001	1.09204

4. The Press Residuals

(a) The press residual is a useful tool determining if your model is "over fit" to a particular dataset. Over fit means that the model is only useful with the particular dataset used to create and it won't be useful for predicting future values. The press residual removes each row in the data set one at a time and for n=351 so I ran the entire data set and have taken a subset of n=5 and have computed the press residual

The first part was created the easy way, using `proc reg` in SAS:

Obs	WL	eFG	DiffTOV	TRB	BLK	press
1	0.355	0.515	-0.8	50.2	6.8	**-0.1188**
2	0.4	0.504	-2.3	48.8	10.4	**-0.0148**
3	0.618	0.503	-0.1	51	12.2	**0.04487**
4	0.406	0.499	0.7	48.5	12.2	**-0.1071**
5	0.467	0.467	0.2	47.1	12	**0.09615**

These outputs, match the vectors below which calculations made using matrix algebra. The steps in SAS used to create these values are detailed on the next page.

pressi
-0.1188
-0.0148
0.04487
-0.1071
0.09615

Obs	PRE_OUT
1	-0.1188
2	-0.0148
3	0.04487
4	-0.1071
5	0.09615

(b) These where Use IML Print statements to examine what is done in Step 4 and Step 7. Explain what is happening in those steps.

```
row=1:n;                                  * used in the "hold=out-one-at-a-
time" steps;

do i = 1 to n;                                      * start of loop;
 rowd=remove(row,i);                                * Step 1;
 Xd = X[rowd,];                                     * Step 2;
 Yd=Y[rowd,];                                       * Step 3;
 bd = inv(t(Xd) * Xd)* t(Xd)*Yd;              * Step 4;
 yhatd=Xd*bd;                                       * Step 5;
 yhati=X[i,]*bd;                              * Step 6;
 pressi[i]=Y[i]-yhati;                              * Step 7;
end;
```

Step 4: (`bd = inv(t(Xd) * Xd)* t(Xd)*Yd;`)

Prior to step 4 a loop is created using SAS. This loop go through each observation, remove (aka delete) it from the model. Given n=351, this would take a ridiculously long time to do by hand.

bd
-2.407092
2.502826
0.0303296
0.0300631
0.0156942

Step 5: (`yhatd=Xd*bd;`)

yhatd
0.4734869
0.4148721
…
0.4915639
0.4801874

Step 6: (`yhati=X[i,]*bd;`)

yhati
0.5846106

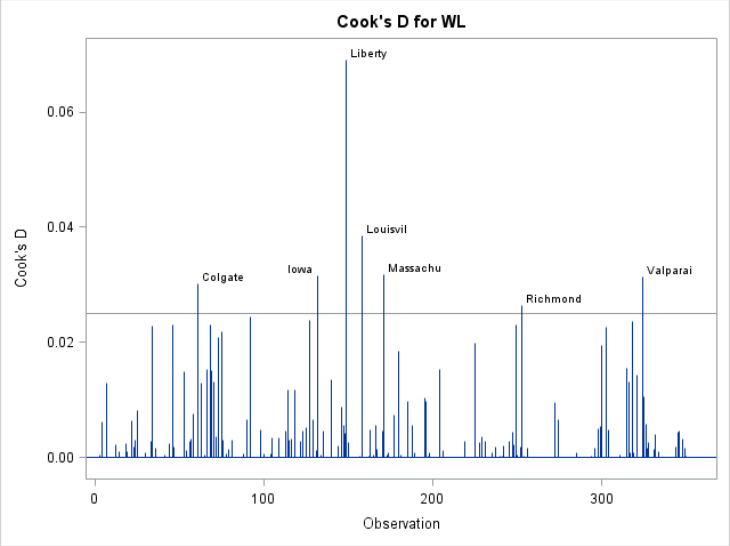
Step 7: (`pressi[i]=Y[i]-yhati`)

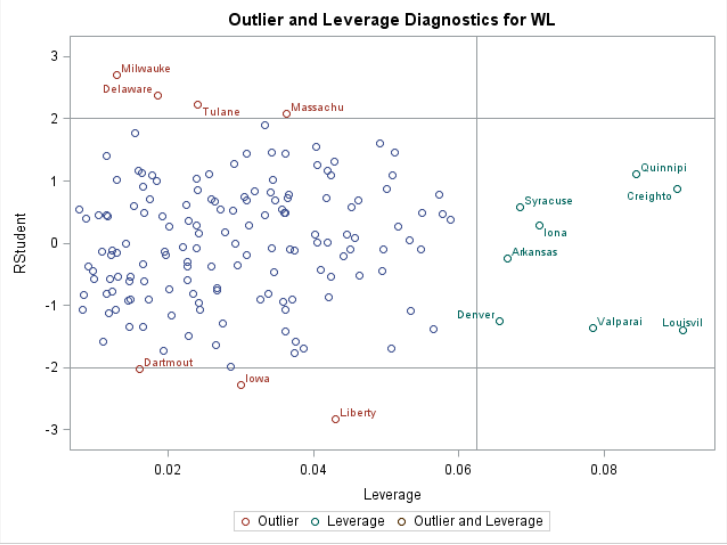
5. Cook's D

(a) Cooks D is a measure of leverage that allows us to finally quantify this concept. Until we have performed visual inspections of scatter plots in order to determine leverage

(b) According to Cook's D as there are several points that cross into outlier territory but none are worth of exclusion. Liberty University is a small school in Lynchburg, Va. that came in with an exceptionally high Cooks'D score. This team had fairly high numbers in 3 out of 4 of my x-variables but only won 34% of their games. If I was to make a scatter plot of all my statistics plotted against y, they would appear below the most of the data but be 2/3rds of the distance of the entire chart away from the y axis. This gives the school the distinction of an outlier but far from the .5 value on the y-axis that would warrant it being removed.

(c)





Chapter 6: Logistic Regression

(a) Every NCAA basketball season concludes with a 66 team single elimination tournament also known as March Madness. These teams are selected through a process grants automatic bids to conference winners and the remaining bids are distributed through a selection committee.

In order determine if there is a statistical relationship between the x variables in my study and admission into the championship tournament, I've added a column of dichotomous values where 1 represents that team qualifying for the 2014 tournament and 0 representing failure to qualify that year.

Using the following SAS command:

```
Proc Logistic descending plots=all;
Model TQ = TOVDiff ;
          run;
```

I was able to generate:

The LOGISTIC Procedure Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-1.61	0.152	112.2498	<.0001
TOVdiff	1	0.2474	0.0584	17.9361	<.0001

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
TOVdiff	1.281	1.142	1.436

(b) The null and alternate hypotheses for the slope test is :

H_0: $\beta_1 = 0$

H_1: $\beta_1 \neq 0$

Test statistic: $\text{t-stat} = \dfrac{b_1 - 0}{s/\sqrt{SSx}} = \dfrac{.02474}{.0584} = 17.93$

Rejection region: |t-stat| > t-critical value, α=0.05
t-critical value with 349 (z) d.f. = 1.96

Conclusion: null hypothesis is rejected because the |t-stat| = | 17.93 | is greater than the t-critical value of 1.96.

(c) The Pr > ChiSq in the slope row given by SAS is <.0001. This value is less than the α-level of 0.05. This means the value is far out into the tail of our 1 sided chi distribution so we reject the null and accept the claim that $\beta_1 \neq 0$, meaning that there is significance in the relationship between TOVdiff and the event of making the NCAA tournament. The Wald Chi-Square value of 17.9361 is confers the conclusion since it is much higher than the 3.84 p-value associated with a 95% test with 1 d.f.

(d) If the null hypothesis were true, then the p-value in the slope row would be equal to the probability there is an increase in the x variable has no effect on our y-axis event, qualifying for the final four or that the slope could be explained by variance within the bounds of a normal distribution. However the SAS returns a probability so small that it is truncated after the one-ten-thousandths digit. The Wald Chi-Square returns a value that is a little more useful for determine just how much these x values allow us to reject the null hypothesis. The value of 17 is a measure given in standard deviations and using the empirical rule, states that 99.7% of events occur within 3 standard deviations of a mean, you can get a better idea of just how small of a chance the null hypothesis is true.

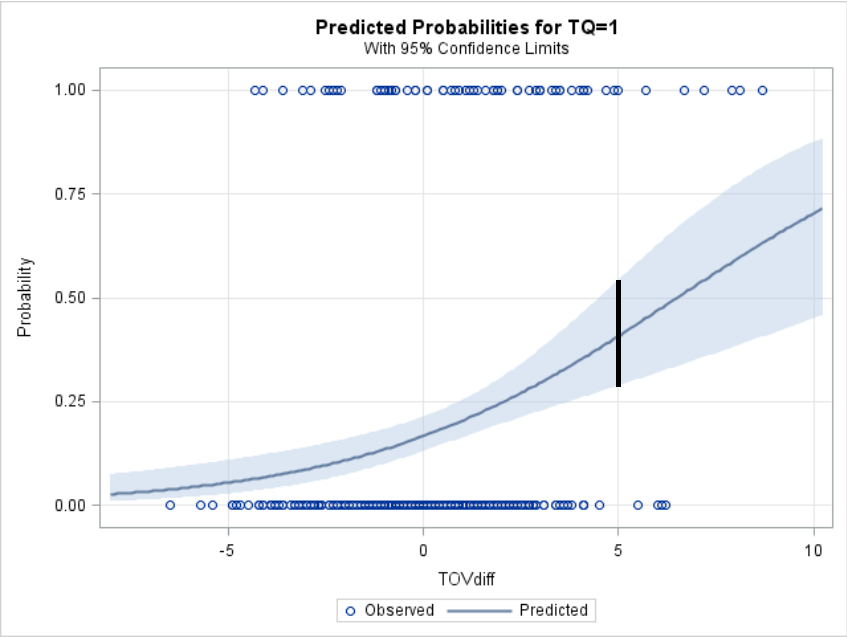
(e) The $\hat{y}$ -equation for this model can be determined using the following equation to determine the logistic curve that determines the expected value:

$$\hat{y} = \frac{e^{b_0 + b_1 x_i}}{1 + e^{b_0 + b_1 x_i}}$$

In the output given by SAS, the b_0= -1.61 and b_1= 0.2474

$$\hat{y} = \frac{e^{-1.61 + .2474(x_i)}}{1 + e^{-1.61 + .2474(x_i)}}$$

f) The output of the SAS logistic regression graph is show below. You can see that the points at the top are shifted further to the right when compared to the data points at the bottom.



(g) In the chart above I have also added a black vertical bar at the x = 5 that denotes the confidence interval at the value. The confidence interval is a measure of how certain we are of the mean is the expected that value. It is best explained using the story of many samples. If we had a parent population of data points with an x value of 5, we would expect that as we re-sampled these to create confidence intervals, 95% would contain a mean (in this case a probability of 1 and expected value of .41) between the upper and lower bounds of the confidence interval.

(h) In the SAS logistic output, the following value was given the odds ratio

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
TOVdiff	1.281	1.142	1.436

The point estimate of TOVdiff is 1.281 which is derived by taking e^{b_1} or in the case of this model where $b_1 = 0.2474$

$$e^{b_1} = e^{.2474} = 1.281$$

This is an estimate of the change in the percentage of an event happening when y is increased by 1. So in the case of this model, every time a team increases the difference in the turnover ratio, they are 28.1% more likely to be selected in the NCAA championship Tournament

(i) However taking this interpretation of your odds ratio literally could be *dangerous* to a NCAA basketball coach's career. Although having a high positive ratio will likely correlate to a successful regular season it is by no means a golden ticket to receiving a bid in the NCAA tournament. There are numerous other factors involved, mainly strength of schedule and number of games won against strong competition.

If a coach took this too literally they could do something foolish such as recruit players solely on their ability to facilitate a positive turnover ratio. The result would likely be a basketball team consisting of only point guards who are typically better ball handlers have higher steals per game numbers. Such a team would likely not have a winning record since it would find itself at a serious disadvantage in other facets of the game including rebounds the ability to set up higher percentage shots closer to the basket. This is similar to the earlier example of evaluating regressing sale price of a house on the number of bathroom and someone using this information as a good reason to tear out the bedrooms of their house in order to install additional bathrooms.

Chapter 7: Cross-validation

In order to further evaluate the correctness of our model, we can use cross-validation to break up and re-test our data using the y-hat equation developed earlier in the model. We can then compare the resulting sum of squared errors of our model with each set of data to see if it minimizes error in multiple scenarios or just with our full data set.

To start we must create two randomly selected groups, the training sample and the validation sample. The best way to accomplish this is to assign a random number to each row of our data set and then sort data based on this value. Simply breaking the data into two sets based on an alphabetical sort would not be sufficient. The random number below has been outlined in red.

Training Sample – first randomly selected 150/351 rows

Obs	School	WL	eFG	TOVdiff	BLK	TRB	RandNum	NewInd
1	Clevelan	0.636	0.536	17.0	9.5	51.4	0.00325	1
2	Georgeto	0.545	0.511	15.9	11.8	50.5	0.00336	2
3	Southern	0.424	0.498	17.5	5.5	51.5	0.01092	3
…	…	…	…	…	…	…	…	…
148	Rice	0.233	0.478	14.8	6.7	46.2	0.40456	148
149	MiamiFL	0.515	0.476	14.1	11.1	51.9	0.41433	149
150	Sacramen	0.467	0.517	16.0	3.8	50.0	0.41860	150

Validation Sample remaining 201 rows

Obs	School	WL	eFG	TOVdiff	BLK	TRB	RandNum	NewInd
1	Hawaii	0.645	0.522	16.7	8.5	53.3	0.41870	151
2	Californ	0.290	0.510	15.9	6.9	43.9	0.41870	152
3	Binghamt	0.233	0.439	17.3	6.1	47.5	0.42045	153
…	…	…	…	…	…	…	…	…
199	Maryland	0.531	0.491	16.4	11.7	52.3	0.99814	349
200	Southeas	0.400	0.506	14.2	11.8	46.8	0.99899	350
201	Mississi	0.576	0.482	17.0	14.7	48.4	0.99966	351

Before SAS examines the training sample and the validation sample separately, we generate the y-hat equation and the MSE of the entire dataset:

Obs	b0	b1	b2	b3	b4
1	-2.84131	3.01857	0.025592	0.021197	0.024416

Once this is determined we create a data step to run through our two data sets and determine MSE using the following y-hat equation:

$$\hat{y} = -2.84131+3.019(eFG) +0.026(TOVdiff)+0.021(BLK)+ 0.024(TRB)$$

We can use this y-hat equation to evaluate the MSE of our sample and Validation groups to get the following output

Obs	_TYPE_	_FREQ_	IntMSE	ExtMSE
1	0	201	0.010207	.009651822

In this output we see that our internal and external MSE are very close in both the internal and external data. This means that the model is reusable and likely that our model is not over fit to the data used to create the model.