Introducing Linear Regression: An Example Using Basketball Statistics

***Tom Arnold and Jonathan Godbey**[1]*

ABSTRACT

The intuition behind linear regression can be difficult for students to grasp particularly without a readily accessible context. This paper uses basketball statistics to demonstrate the purpose of linear regression and to explain how to interpret its results. In particular, the student will quickly grasp the meaning of explanatory variables, r-squared, and the statistical significance of estimates of regression coefficients. Even if the student is not a sports fan the examples are easily understood and familiar. The student can easily replicate the procedures in this paper to reinforce learning.

Introduction

When calculators were introduced into the classroom, a number of tedious calculations could suddenly be performed very quickly. However, students' comprehension of mathematics did not actually improve and the calculator in many ways masked deficiencies because a keystroke sequence could substitute for comprehension. Regression analysis has some of the same characteristics because econometric software has improved to the point that a regression is a simple one line command that results in copious amounts of output.

We believe the reason for the difficulty in understanding/interpreting regression results is not the product of students being unable to perform the regression using matrix algebra, but the product of a lack of intuition that belies the one line of code. In other words, it is equally important to understand why the regression is being performed, what the regression process does to the data, and how to interpret the regression results.

By using the statistics from a basketball team, students become enabled to comprehend a model for predicting the number of points a given player should be able to score based on certain factors. Regression is then introduced as a means to calibrate and test the model.

The paper begins with a breakdown of what a regression "does" based on a very small set of data in which hand calculators can perform all of the calculations. Next, the statistics from a basketball team are presented so that a model can be generated to predict how many points should be scored in a game by an individual given a set of factors. A regression is performed to calibrate and test the model. A second regression based on the capital asset pricing model (Sharpe, 1964) is then performed on actual financial data. The paper concludes at this point.

What Actually Happens in a Regression?

Before introducing the basketball data, start with something even smaller. In Table 1, we have three columns of data: A, B, and C with the averages of each column calculated by summing the data and dividing it by the number of observations (5 in this case).

[1] Tom Arnold, F. Carlyle Tiller Chair in Business, Robins School of Business, University of Richmond, 1 Gateway Road, Richmond, VA 23173; Jonathan M. Godbey, Clinical Assistant Professor, J. Mack College of Business, Georgia State University, 1209 RCB Building, Atlanta, GA 30303. The authors would like to thank an anonymous referee, Jerry Stevens, David Greenberg, and participants at the 2010 Financial Management Association Meetings for helpful comments.

Table 1: Data

Observation:	Data A:	Data B:	Data C:
1	22	10	8
2	47	21	15
3	34	19	12
4	21	8	5
5	66	25	19
Mean:	38.00	16.60	11.80

Assume existing theory states that A is dependent on B and C in some manner, usually expressed as a model. In other words, there is some combination of data B and data C that generates data A. Further assume, that theory views the relationship as linear: A = β1*B + β2*C. With this information: data and a proposed model of how B and C generate A, discuss with the class criteria for selecting β1 and β2. After some discussion, if it has not already been suggested, try the criteria: Mean (A) less Mean (B) less Mean (C) is zero (i.e. the mean of the error in the model is zero). Be certain to make the point: without this condition, on average, the model will be in error. Next, suggest some combinations of β1 and β2: (1.2000, 1.5322), (1.5783, 1.0000), and (2.5000, -0.2966). Each combination meets the criteria:

$$0 = 38.00 - \{(1.2000)*16.60 + (1.5322)*11.80\} \quad (1)$$

$$0 = 38.00 - \{(1.5783)*16.60 + (1.0000)*11.80\} \quad (2)$$

$$0 = 38.00 - \{(2.5000)*16.60 + (-0.2966)*11.80\} \quad (3)$$

However, of the three combinations, which one is the best?

Discussion usually leads to an answer of using the combination of β_1 and β_2 with the least amount of error. The problem with such a conclusion is that it depends on what is meant by error. On average, all three combinations have equivalent error, but the error is different for each combination (see Table 2).

Table 2: Model Error

	β1 = 1.2000	β1 = 1.5783	β1 = 2.5000
Observation:	**β2 = 1.5322**	**β2 = 1.0000**	**β2 = -0.2966**
1	-2.258	-1.783	-0.627
2	-1.183	-1.144	-1.051
3	-7.186	-7.988	-9.941
4	3.739	3.374	2.483
5	6.888	7.543	9.135
Mean of Error:	0.00	0.00	0.00

To say one version of error is better than another version of error is fairly difficult from observing the error. Also, unless a systematic rule is in place, there can be no consistent means of determining which version of the error is best. After some more discussion (note: the discussion is important to get the student to understand the issue and possibly resolve the issue), suggest looking at the square of the error and take the mean of the squared error (i.e. the mean squared error).

Table 3: Model Squared Error

	β1 = 1.2000	β1 = 1.5783	β1 = 2.5000
Observation:	**β2 = 1.5322**	**β2 = 1.0000**	**β2 = -0.2966**
1	5.097	3.197	0.393
2	1.399	1.3089	1.105
3	51.644	63.803	98.820
4	13.980	11.381	6.165
5	47.447	56.889	83.456
Mean of Squared Error:	23.914	27.312	37.988

When viewing the mean squared error, β1 = 1.2000 and β2 = 1.5322 appear to be the best performing values of the three combinations. However, suppose β1 is set to 1.0040 and β2 is set to 1.8400 (see Table 4).

Table 4: Model with β1 = 1.0040 and β2 = 1.8400

Observation:	Error:		Squared Error:
1	-2.760		7.618
2	-1.684		2.836
3	-7.156		51.208
4	3.768		14.198
5	5.940		35.284
Mean of Error:	-0.378	Mean of Squared Error:	22.229

Notice, the mean squared error is reduced versus the best case from Table 3, however, the mean of the error is not zero. The question becomes: is it possible to use these two values for β1 and β2 to get a lower mean squared error and have the mean of the error be zero as well? After some discussion (note: again, the discussion needs to happen if the student is going internalize the intuition of the analysis), suggest using an intercept term α of -0.3784. Now the relationship becomes: A = α + β1*B + β2*C. By allowing an intercept term, the mean squared error is further reduced and the mean of the model now equals zero.

$$0 = 38.00 - \{-0.3784 + (1.0040)*16.60 + (1.8400)*11.80\} \qquad (4)$$

At this point, the intuition of what occurs in a regression calculation is complete. Regression is a means of finding the coefficient combination (that can include an intercept) that will minimize the squared error between the dependent variable (A) and a set of independent variables (B and C). By visualizing that different coefficient combinations generate different mean squared errors, the students can see that an optimal solution exists and that this is ultimately what the computer code for the regression accomplishes. Now the question becomes: what are examples of data A, data B, and data C?

A Simple Example

Consider the following situation. Georgia State University's basketball team has 12 players. Each player averages a different number of points per game (PPG). Table 5 shows all the information we have about the players: PPG, average minutes played per game (MPG), average rebounds per game (RPG), and jersey number (NUM).

Table 5: Player Information

Player	PPG	MPG	RBG	NUM
Dukes	12.8	32.1	4.7	2
Goldston	10.8	27.1	1.5	11
Mendez	8.8	27.8	2.9	21
Chase	5.1	25.6	5.8	15
Hansbro	4.9	14.8	2.8	23
Curry	6.9	19.1	1.6	12
Hampton	4.4	22.1	3.6	1
Krubally	2.9	10.9	2.9	24
Fields	2.3	11.5	1.1	4
Lott	2.2	10.9	1.2	30
Rimmer	2.5	13.3	2.7	33
Echols	7.4	21.6	5.6	0

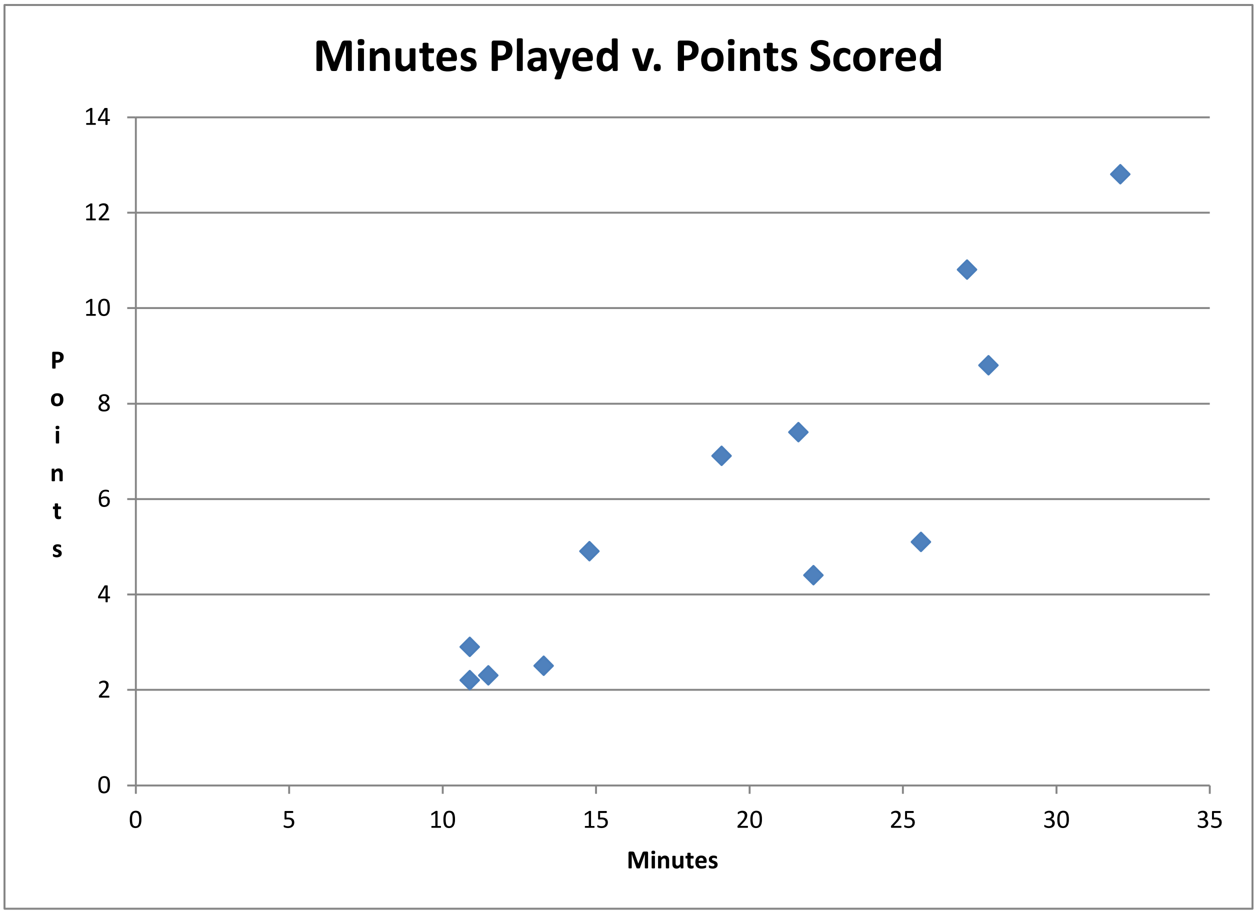
If we did not know PPG, would it be possible to calculate the exact PPG of each player? If it is not possible to calculate the exact PPG, is an estimate possible? How good is that estimate? In other words does MPG, RPG or NUM explain part or all of PPG? Perhaps we only need to know one of the three variables to estimate PPG. Maybe one is enough to get an approximate estimate but knowing one or both of the others will give us a more precise estimate. Linear regression provides the means to answer these questions.

Note that we only have 12 players. It makes sense that using more observations would make us more confident in our estimates. If we were truly trying to answer the question of what explains PPG we would want data from many other teams. However, for purposes of trying to "see" what is happening with the numbers it is better to keep the number of observations small.

Most basketball players would agree that playing more minutes means you will score more points. At least we are reasonably certain that playing more minutes will not mean that you score fewer points. Figure 1 shows the relationship between points scored and minutes played. Clearly, playing more minutes results in scoring more points. Mathematically, we want the following relationship.

$$PPG = \alpha + \beta * MPG \tag{5}$$

Figure 1



If we know α and β, then all we need to know is MPG and we can easily calculate PPG. For example, assume α is 1 and β is 0.5. If a player plays 20 minutes then he must have scored 1 + 0.5*20 = 11 points. Unfortunately, we do not know the true α or β. We need to estimate them. Figure 1 shows that the relationship is not exact. The points do not lie on a straight line. There is no α and β that will allow us to determine the exact PPG for every player. There will be some error. The best we can do is find an α and β combination that will come as close as possible to estimating PPG using MPG. Following the intuition from the last section, a linear regression solves for the best combination of α and β (i.e., it minimizes the mean squared error) that allows MPG (the independent variable) to estimate PPG (the dependent variable). Our new equation (with the error term added) is:

$$PPG = \alpha + \beta * MPG + \varepsilon \tag{6}$$

Fortunately, Excel does the calculations quickly and easily.

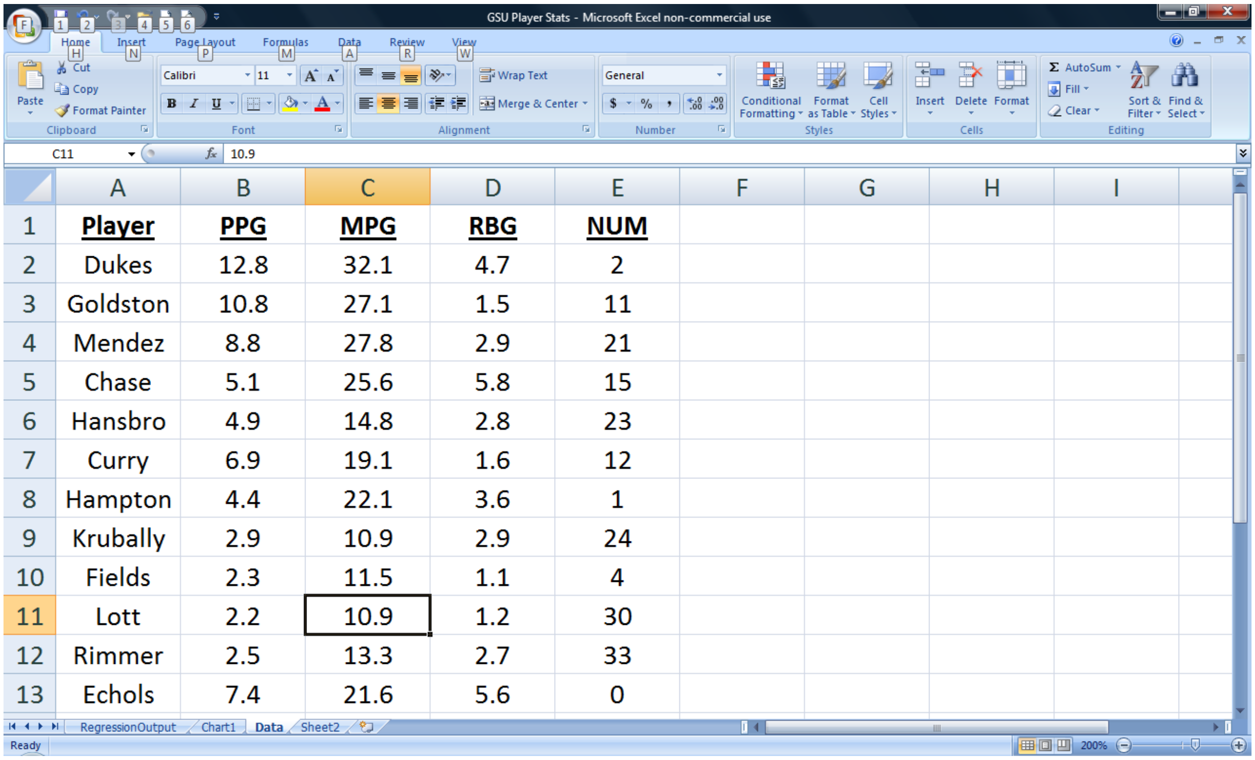
Regression in Excel

Rather than having an equation like equation (2), Excel uses the general form:

$$Y = \alpha + \beta * X + \varepsilon. \tag{7}$$

To run the regression, enter data as shown in Figure 2.

Figure 2



C11 fx 10.9

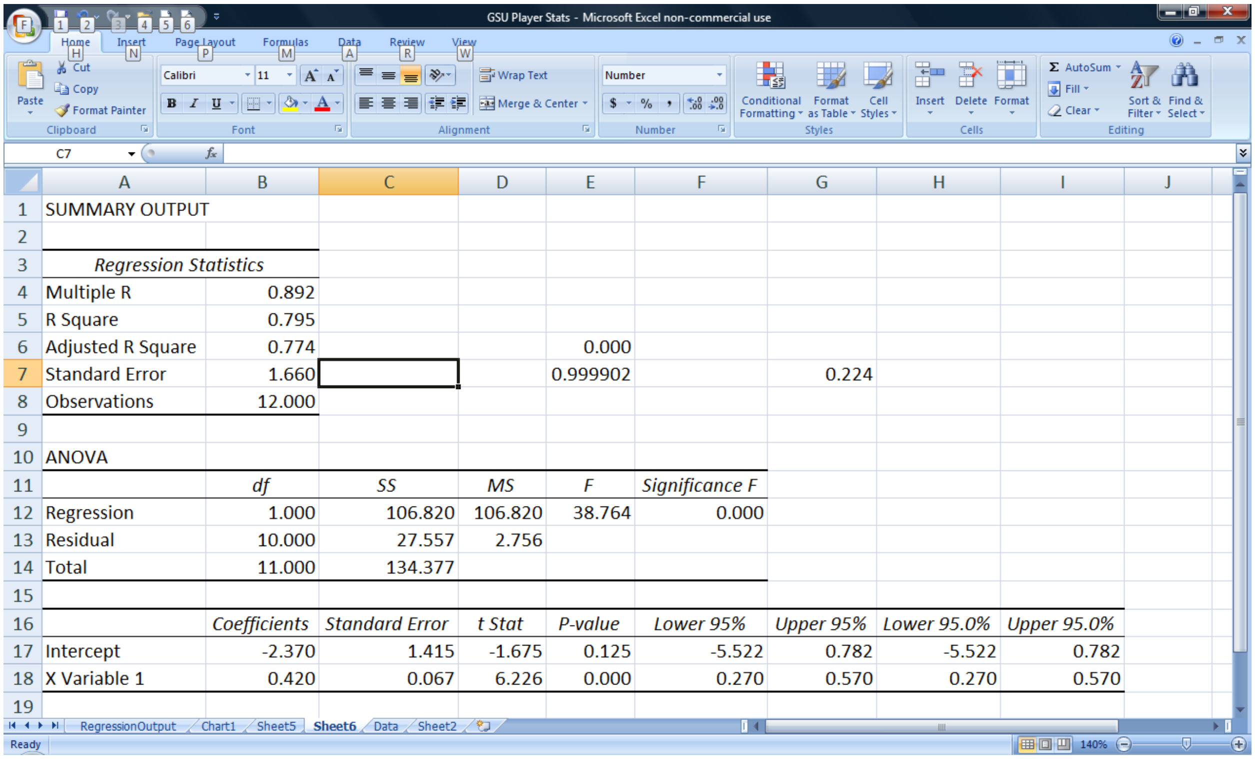
	A	B	C	D	E
1	Player	PPG	MPG	RBG	NUM
2	Dukes	12.8	32.1	4.7	2
3	Goldston	10.8	27.1	1.5	11
4	Mendez	8.8	27.8	2.9	21
5	Chase	5.1	25.6	5.8	15
6	Hansbro	4.9	14.8	2.8	23
7	Curry	6.9	19.1	1.6	12
8	Hampton	4.4	22.1	3.6	1
9	Krubally	2.9	10.9	2.9	24
10	Fields	2.3	11.5	1.1	4
11	Lott	2.2	10.9	1.2	30
12	Rimmer	2.5	13.3	2.7	33
13	Echols	7.4	21.6	5.6	0

Go to the Data tab, click Data Analysis, then click Regression. Now highlight cells B2 to B13 for the Y data, cells C2 to C13 for the X data, and choose a cell for output. Click OK and Excel will produce the results shown in Figure 3. (To make Figure 3 easier to read the initial output was re-formatted.) Note: Many times the Data Analysis tab is not available in an initial setup for Excel. However, it can be added by going to the File tab in Excel 2010, click Options, and click Analysis ToolPak. In the previous version of Excel, click the clover-leaf-like icon, click Excel options at the bottom of the menu, click Add-ins, and click Analysis ToolPak.

Interpreting Regression Results

Our question remains, does MPG explain PPG? If it does then β from equation (6) has to be different from zero. If β were zero then any number of MPG would result in the same estimate for PPG. In other words MPG would not explain PPG. Mathematically, α could take any value. However, we know in reality it must be zero. If a player averages zero MPG he must score zero PPG. Figure 3 gives the estimates for α and β.

Figure 3



	A	B	C	D	E	F	G	H	I	J
1	SUMMARY OUTPUT									
2										
3	*Regression Statistics*									
4	Multiple R	0.892								
5	R Square	0.795								
6	Adjusted R Square	0.774			0.000					
7	Standard Error	1.660			0.999902		0.224			
8	Observations	12.000								
9										
10	ANOVA									
11		*df*	*SS*	*MS*	*F*	*Significance F*				
12	Regression	1.000	106.820	106.820	38.764	0.000				
13	Residual	10.000	27.557	2.756						
14	Total	11.000	134.377							
15										
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*	
17	Intercept	-2.370	1.415	-1.675	0.125	-5.522	0.782	-5.522	0.782	
18	X Variable 1	0.420	0.067	6.226	0.000	0.270	0.570	0.270	0.570	
19										

The coefficient for "Intercept" is α and the coefficient for "X Variable 1" is β. These are estimates, not the true values of α and β. We now have the following.

$$PPG = -2.370 + 0.420 * MPG \tag{8}$$

The estimate for α was -2.370 when we know the true value to be zero. How likely is it that the estimate would be found to be zero? The p-value of 0.125 answers that question. This p-value means that there is a 12.5% chance that the estimated value of α is zero and equal to (in this case) the true value of zero. This also means there is an 87.5% chance that the estimated α is different from zero. Because we know α should be zero, this result does not seem to make sense. There are three explanations for this seemingly incorrect result. First, we have a small sample size. Perhaps more data would result in an estimate closer to zero. When adding two more teams (Drexel University and James Madison University), α becomes -0.9950 which is closer to zero than the initial regression, but ironically has a lower p-value (6.27%) as well. Perhaps even more data is needed or one of the remaining two explanations may need to be considered in addition.

Second, other variables may matter in addition to MPG which are currently omitted from the regression. The effect of these "omitted variables" would be captured in α. Additional variables will be considered later in the paper (RPG and NUM), however, take this opportunity to discuss with students what other variables may matter. One example is a binary variable such as home game versus away game. Such a variable will not work with the type of data presented here, but a very rich discussion about the quality of data can be had as result. Discuss with the students what kind of data is preferable…cumulative data (like the data presented here) or individual game data, or an even finer increment of data. Note: many times economic and financial data come in varying frequencies making this discussion very relevant.

Third, estimates are generally considered to be different from zero only if the p-value is 0.10 or less. If the p-value is greater than 0.10 then the estimate is said to be insignificant. If the p-value is less than 0.10, then the estimate is said to be significant at the 10% level. If the p-value is less than 0.05, then the estimate is said to be significant at the 5% level and less than 0.01 means the estimate is significant at the 1% level. In general, estimates are grouped into one of the four categories above.

In the initial regression with only GSU players, α is insignificant. However, with additional data, α becomes significant at the 10% level. In this context, α has to be zero (again, you cannot score points if you do not play) and being estimated as different from zero while being insignificant is “statistically” the equivalent of being zero. The initial regression demonstrates this case. However, with additional data, the estimate for α gets closer to zero, but then becomes significant. Although an apparently “confounding result”, this is not an uncommon situation and as discussed earlier leads to the desire for more data, more variables, or both to address the confounding result.

It should be noted, that many times the true value of α is not known with any certainty within a particular model and analyzing the estimated α is simply a matter of determining if the estimated α is different from zero. In this case, α is known to be zero which allowed for a more in depth discussion of the regression model not being fully accurate.

The estimate for β was 0.420. If the true value of β is different from zero then MPG is useful in explaining PPG. Is 0.420 close to zero? Once again, the p-value of 0.000098 answers that question. (Note Figure 3 shows the rounded result 0.000). This means that there is a 0.0098% chance that the true value of β is zero and our estimate is different from zero as a result of chance. In other words, we are 99.9902% sure that the true β is not zero. MPG matters. MPG is significant at the 1% level.

MPG matters, but does it explain all of PPG? Would knowing RPG (average rebounds per game) or NUM (jersey number) help in the estimation of PPG? The r-square given in Figure 3 indicates the percentage of the variation of PPG that is explained by MPG. The result of 0.795 means that 79.5% of the variation in PPG is explained by MPG (Note: this is an upwardly biased estimate and more analysis presented later in this paper can be used to address this issue). If the result had been 1.000 then MPG would explain 100% of the variation in PPG. Because not all of the variation in PPG is explained by MPG, let us consider searching for more explanatory variables.

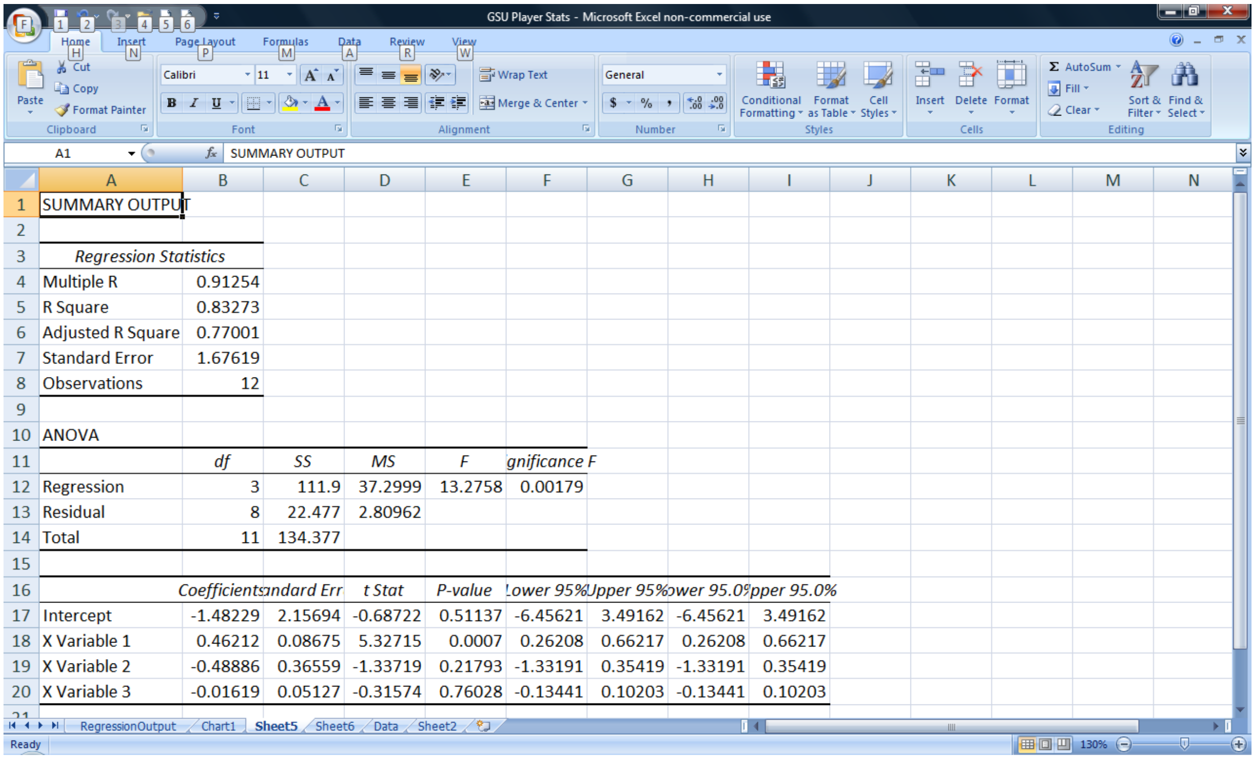
Multivariate Regressions

The significance of multiple variables can be checked simultaneously. Suppose we want to see if RPG and NUM in addition to MPG are important in explaining PPG because we believe more variables are necessary to produce a better estimation of PPG. We, then, run the following regression.

$$PPG = \alpha + \beta 1 * MPG + \beta 2 * RPG + \beta 3 * NUM + \varepsilon \qquad (9)$$

To run this regression, enter data as shown in Figure 2. Go to the Data tab, click Data Analysis, then click Regression. Now highlight cells B2 to B13 for the Y data, cells C2 to E13 for the X data, and choose a cell for output. Click OK and Excel will produce the results shown in Figure 4. (To make Figure 4 easier to read the initial output was re-formatted.)

Figure 4



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.91254							
5	R Square	0.83273							
6	Adjusted R Square	0.77001							
7	Standard Error	1.67619							
8	Observations	12							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*gnificance F*			
12	Regression	3	111.9	37.2999	13.2758	0.00179			
13	Residual	8	22.477	2.80962					
14	Total	11	134.377						
15									
16		*Coefficients*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
17	Intercept	-1.48229	2.15694	-0.68722	0.51137	-6.45621	3.49162	-6.45621	3.49162
18	X Variable 1	0.46212	0.08675	5.32715	0.0007	0.26208	0.66217	0.26208	0.66217
19	X Variable 2	-0.48886	0.36559	-1.33719	0.21793	-1.33191	0.35419	-1.33191	0.35419
20	X Variable 3	-0.01619	0.05127	-0.31574	0.76028	-0.13441	0.10203	-0.13441	0.10203

Excel separates output based on the columns of the input. "X Variable 1" corresponds to our β1, "X Variable 2" to our β2 and "X Variable 3" to our β3. If RPG and NUM are important in explaining PPG then the estimates for β2 and β3 should be significantly different from zero.

First, consider the intuition. What do you expect? Most people guess that RPG might be important but that NUM would not be. Logically, one could assume that a player who rebounds well is a talented player and has the ability to score but the number on a player's jersey will not add any ability to score, rebound or to perform any task on the court. A counterargument could be easily made. Rebounding is a separate skill. Those who are good at it score fewer points or those who miss a lot of shots must rebound as a result. Possibly, rebounding is a defensive skill and such players focus on their comparative advantage in rebounding instead of shooting for the good of the team. Scoring specialists score, rebounding specialists rebound. Therefore, players with more rebounds will score less. The estimate for β2 should then be negative and significantly different from zero. It may be that the skills are separate and have no impact on one another. The estimate should then be insignificant.

It could be possible that jersey numbers matter. While not the source of superior performance the number could be a signal of it. Teams may assign lower numbers to better scorers. In other sports, jersey numbers actually do correspond with specific types of positions. No matter if one believes this to be true or not, it is a testable condition within a regression.

It is important to discuss with the class that the above arguments could be totally false, completely true or partial explanations. Let the class determine what data should be important and provide an explanation why the data is important for explaining PPG (Figure 4 reveals the truth about such contentions).

The coefficient of MPG is 0.462 and its p-value is 0.001. MPG is significant at the 1% level. We are 99.9% certain that the true coefficient is not zero. All else constant, playing another minute will result in scoring 0.462 more points. RPG and NUM have coefficients of -0.489 and -0.016, respectively. The p-values are 0.218 and 0.760. Neither one is significant at the 10% level and therefore both are considered insignificant. In other words, they do not matter.

The intercept (α) is -1.482, which is closer to zero than when estimated in the initial regression with only MPG as an independent variable. Unlike the initial regression, one cannot definitively state that α should be zero because RPG and NUM have also been included.in the regression. However, RPG and NUM are considered insignificant while MPG is still significant. Consequently, because RPG and NUM are insignificant, it again makes sense for α to be zero. As stated before, α is closer to zero than it had been in the initial regression, but what is more striking is how much more insignificant α has become: p-value = 0.511 versus a p-value of 0.125 previously. Again, "statistically", α is effectively estimated to be zero.

The r-square is 0.833 which is higher than the r-square of the regression containing only MPG. R-square will never fall if additional explanatory variables are added. It may even rise if additional insignificant explanatory variables are added. This result demonstrates a weakness of r-square. The addition RPG and NUM which do not matter in the estimation of PPG should not cause r-square to rise. To address this issue the measure "adjusted r-square" was developed. The adjusted r-square lowers r-square as more explanatory variables are added to the equation. This way, the effects of adding insignificant explanatory variables becomes mitigated.

Regression with Financial Data

In the same way we used linear regression to identify variables that explain scoring in basketball, we may use it to identify variables that explain the returns on individual stocks. We begin with the theoretical Capital Asset Pricing Model (CAPM) and ExxonMobil (XOM). The CAPM implies that the returns on XOM can be explained by returns on the market. We will assume the market is the S&P 500. The equation is:

$$E[r_{XOM}] = r_f + \beta * [E[r_{S\&P500}] - r_f]. \qquad (10)$$

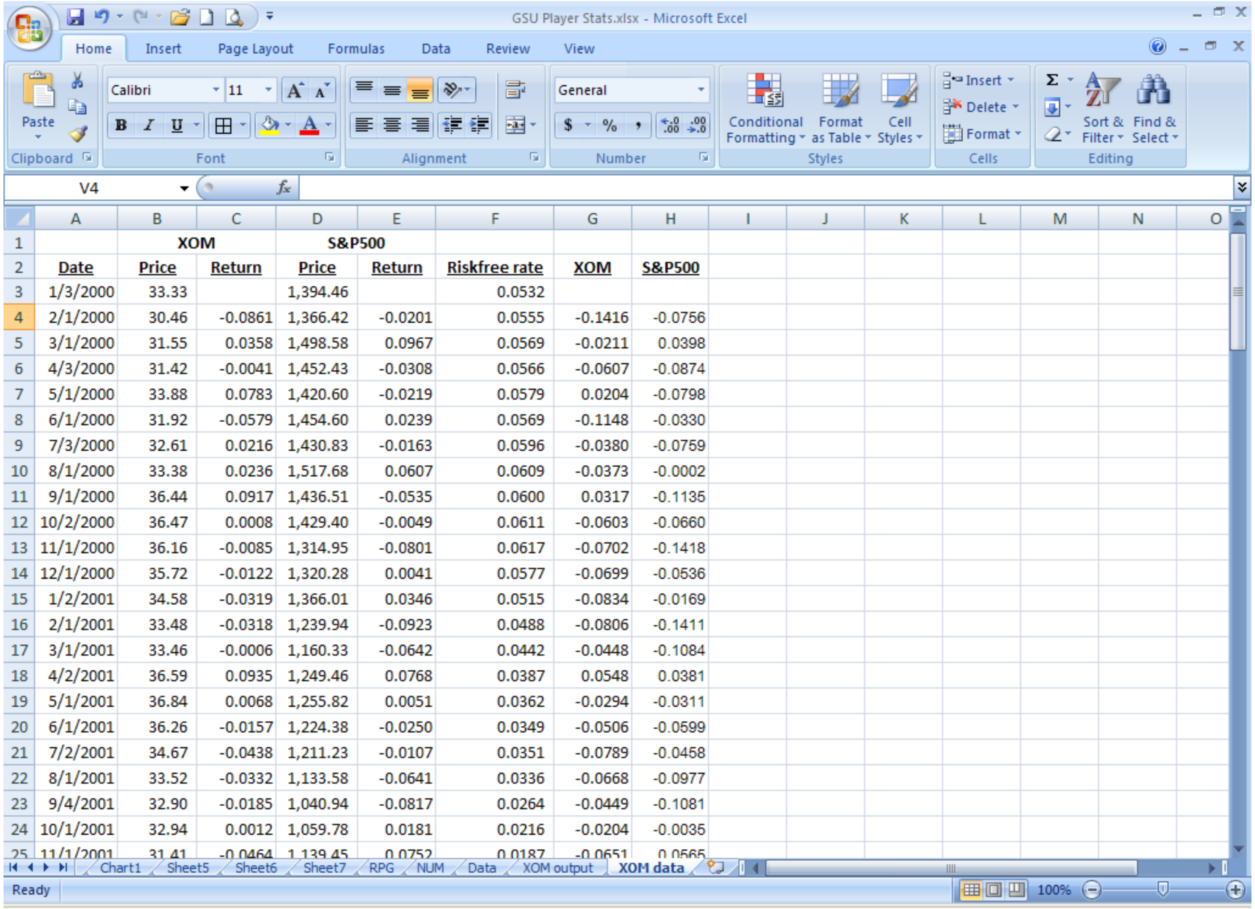
Equation (10) states that the expected return on XOM is the risk-free rate (r_f) plus some risk premium. That risk premium is β times the expected return on the S&P 500 above the risk-free rate. To test this relationship, first, subtract the risk-free rate from both sides. The result is: $E[r_{XOM}] - r_f = \alpha + \beta * [E[r_{S\&P500}] - r_f]$. (11)

Returns for XOM are calculated using monthly prices as given by finance.yahoo.com from January 2000 to December 2008. Returns for the S&P 500 over the same time period are calculated using data from finance.yahoo.com. The risk-free rate is assumed to be the 3-month T-bill as shown on www.federalreserve.gov. The risk-free rate is subtracted from both XOM and S&P500 returns. Now run the regression

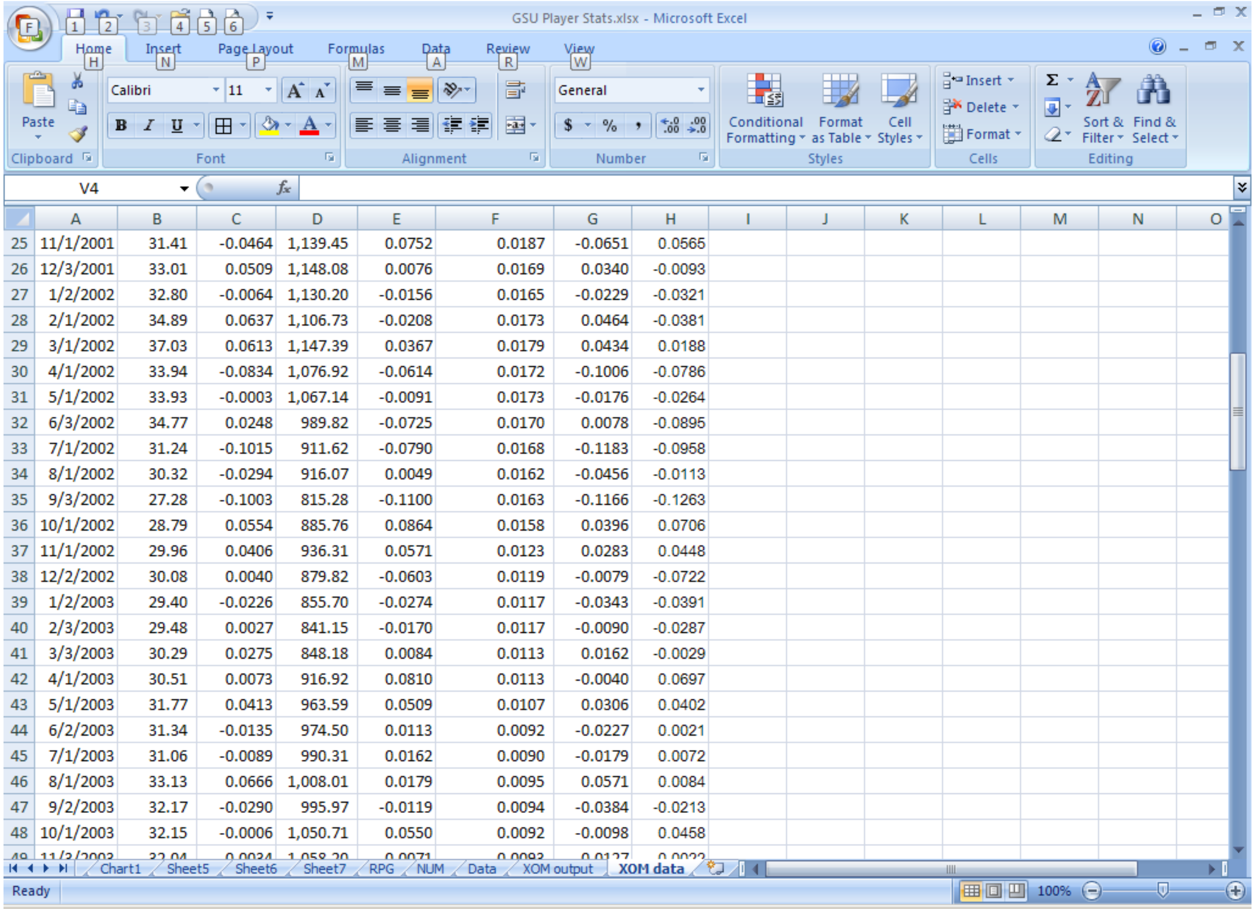
$$R_{XOM} = \alpha + \beta * R_{S\&P500} + \varepsilon. \qquad (12)$$

R_{XOM} are the returns on XOM above the risk-free rate and shown in Figure 5, column G. $R_{S\&P500}$ are the returns on the S&P500 above the risk-free rate and shown in Figure 5, column H. To run the regression highlight G4..G122 for Y and H4..H122 for X.

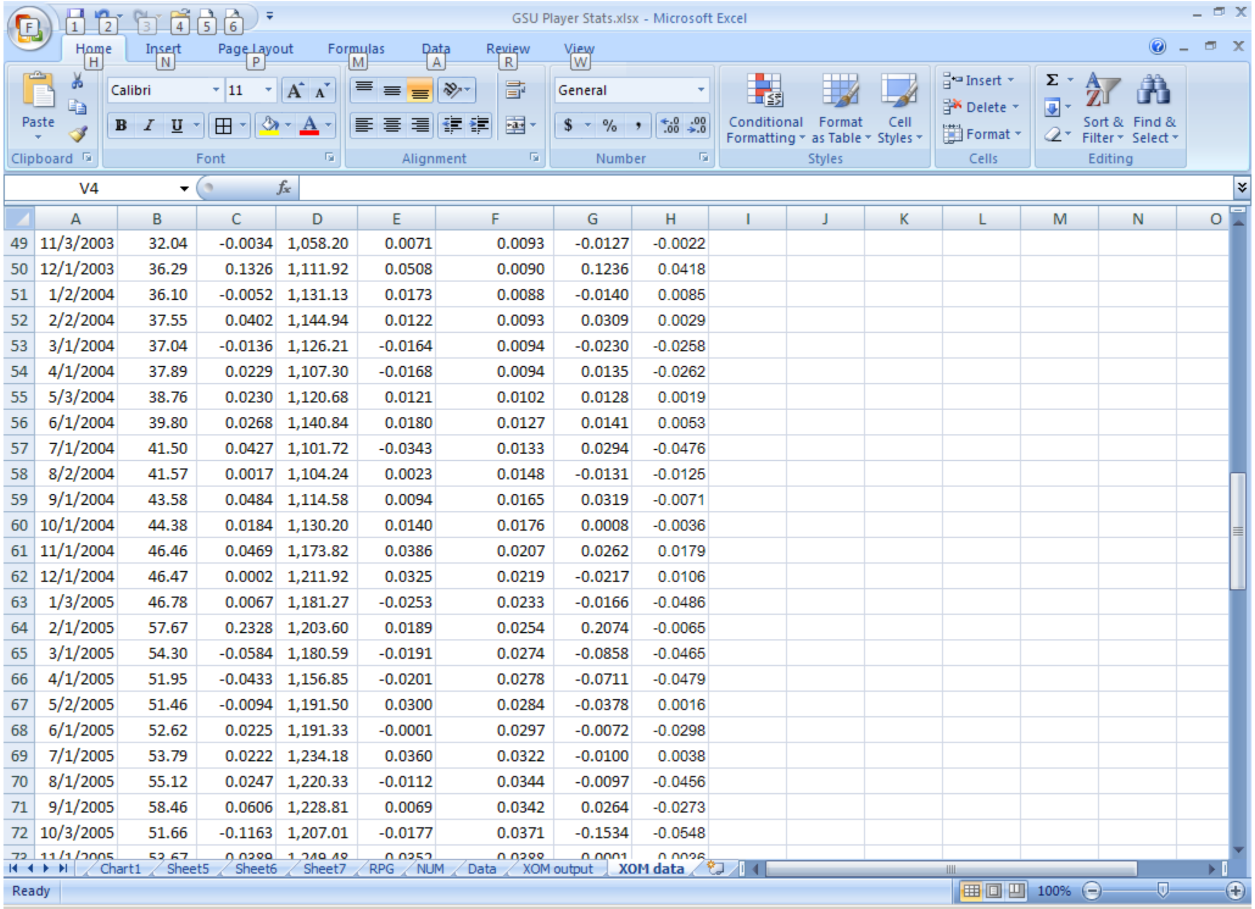
Figure 5



	A	B	C	D	E	F	G	H
1		XOM		S&P500				
2	Date	Price	Return	Price	Return	Riskfree rate	XOM	S&P500
3	1/3/2000	33.33		1,394.46		0.0532		
4	2/1/2000	30.46	-0.0861	1,366.42	-0.0201	0.0555	-0.1416	-0.0756
5	3/1/2000	31.55	0.0358	1,498.58	0.0967	0.0569	-0.0211	0.0398
6	4/3/2000	31.42	-0.0041	1,452.43	-0.0308	0.0566	-0.0607	-0.0874
7	5/1/2000	33.88	0.0783	1,420.60	-0.0219	0.0579	0.0204	-0.0798
8	6/1/2000	31.92	-0.0579	1,454.60	0.0239	0.0569	-0.1148	-0.0330
9	7/3/2000	32.61	0.0216	1,430.83	-0.0163	0.0596	-0.0380	-0.0759
10	8/1/2000	33.38	0.0236	1,517.68	0.0607	0.0609	-0.0373	-0.0002
11	9/1/2000	36.44	0.0917	1,436.51	-0.0535	0.0600	0.0317	-0.1135
12	10/2/2000	36.47	0.0008	1,429.40	-0.0049	0.0611	-0.0603	-0.0660
13	11/1/2000	36.16	-0.0085	1,314.95	-0.0801	0.0617	-0.0702	-0.1418
14	12/1/2000	35.72	-0.0122	1,320.28	0.0041	0.0577	-0.0699	-0.0536
15	1/2/2001	34.58	-0.0319	1,366.01	0.0346	0.0515	-0.0834	-0.0169
16	2/1/2001	33.48	-0.0318	1,239.94	-0.0923	0.0488	-0.0806	-0.1411
17	3/1/2001	33.46	-0.0006	1,160.33	-0.0642	0.0442	-0.0448	-0.1084
18	4/2/2001	36.59	0.0935	1,249.46	0.0768	0.0387	0.0548	0.0381
19	5/1/2001	36.84	0.0068	1,255.82	0.0051	0.0362	-0.0294	-0.0311
20	6/1/2001	36.26	-0.0157	1,224.38	-0.0250	0.0349	-0.0506	-0.0599
21	7/2/2001	34.67	-0.0438	1,211.23	-0.0107	0.0351	-0.0789	-0.0458
22	8/1/2001	33.52	-0.0332	1,133.58	-0.0641	0.0336	-0.0668	-0.0977
23	9/4/2001	32.90	-0.0185	1,040.94	-0.0817	0.0264	-0.0449	-0.1081
24	10/1/2001	32.94	0.0012	1,059.78	0.0181	0.0216	-0.0204	-0.0035
25	11/1/2001	31.41	-0.0464	1,139.45	0.0752	0.0187	-0.0651	0.0565

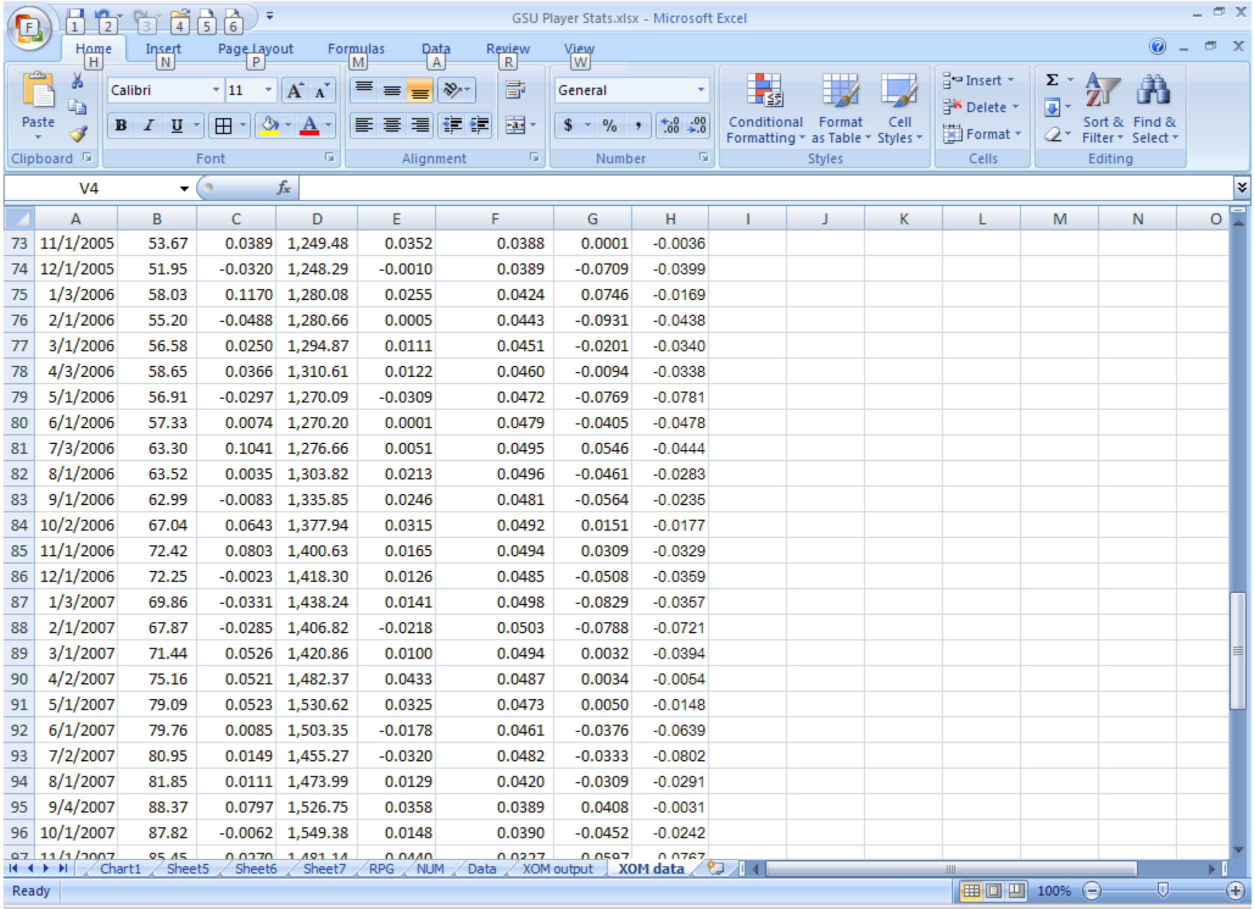


	A	B	C	D	E	F	G	H
25	11/1/2001	31.41	-0.0464	1,139.45	0.0752	0.0187	-0.0651	0.0565
26	12/3/2001	33.01	0.0509	1,148.08	0.0076	0.0169	0.0340	-0.0093
27	1/2/2002	32.80	-0.0064	1,130.20	-0.0156	0.0165	-0.0229	-0.0321
28	2/1/2002	34.89	0.0637	1,106.73	-0.0208	0.0173	0.0464	-0.0381
29	3/1/2002	37.03	0.0613	1,147.39	0.0367	0.0179	0.0434	0.0188
30	4/1/2002	33.94	-0.0834	1,076.92	-0.0614	0.0172	-0.1006	-0.0786
31	5/1/2002	33.93	-0.0003	1,067.14	-0.0091	0.0173	-0.0176	-0.0264
32	6/3/2002	34.77	0.0248	989.82	-0.0725	0.0170	0.0078	-0.0895
33	7/1/2002	31.24	-0.1015	911.62	-0.0790	0.0168	-0.1183	-0.0958
34	8/1/2002	30.32	-0.0294	916.07	0.0049	0.0162	-0.0456	-0.0113
35	9/3/2002	27.28	-0.1003	815.28	-0.1100	0.0163	-0.1166	-0.1263
36	10/1/2002	28.79	0.0554	885.76	0.0864	0.0158	0.0396	0.0706
37	11/1/2002	29.96	0.0406	936.31	0.0571	0.0123	0.0283	0.0448
38	12/2/2002	30.08	0.0040	879.82	-0.0603	0.0119	-0.0079	-0.0722
39	1/2/2003	29.40	-0.0226	855.70	-0.0274	0.0117	-0.0343	-0.0391
40	2/3/2003	29.48	0.0027	841.15	-0.0170	0.0117	-0.0090	-0.0287
41	3/3/2003	30.29	0.0275	848.18	0.0084	0.0113	0.0162	-0.0029
42	4/1/2003	30.51	0.0073	916.92	0.0810	0.0113	-0.0040	0.0697
43	5/1/2003	31.77	0.0413	963.59	0.0509	0.0107	0.0306	0.0402
44	6/2/2003	31.34	-0.0135	974.50	0.0113	0.0092	-0.0227	0.0021
45	7/1/2003	31.06	-0.0089	990.31	0.0162	0.0090	-0.0179	0.0072
46	8/1/2003	33.13	0.0666	1,008.01	0.0179	0.0095	0.0571	0.0084
47	9/2/2003	32.17	-0.0290	995.97	-0.0119	0.0094	-0.0384	-0.0213
48	10/1/2003	32.15	-0.0006	1,050.71	0.0550	0.0092	-0.0098	0.0458
49	11/3/2003	32.04	0.0034	1,058.20	0.0071	0.0093	0.0127	0.0022

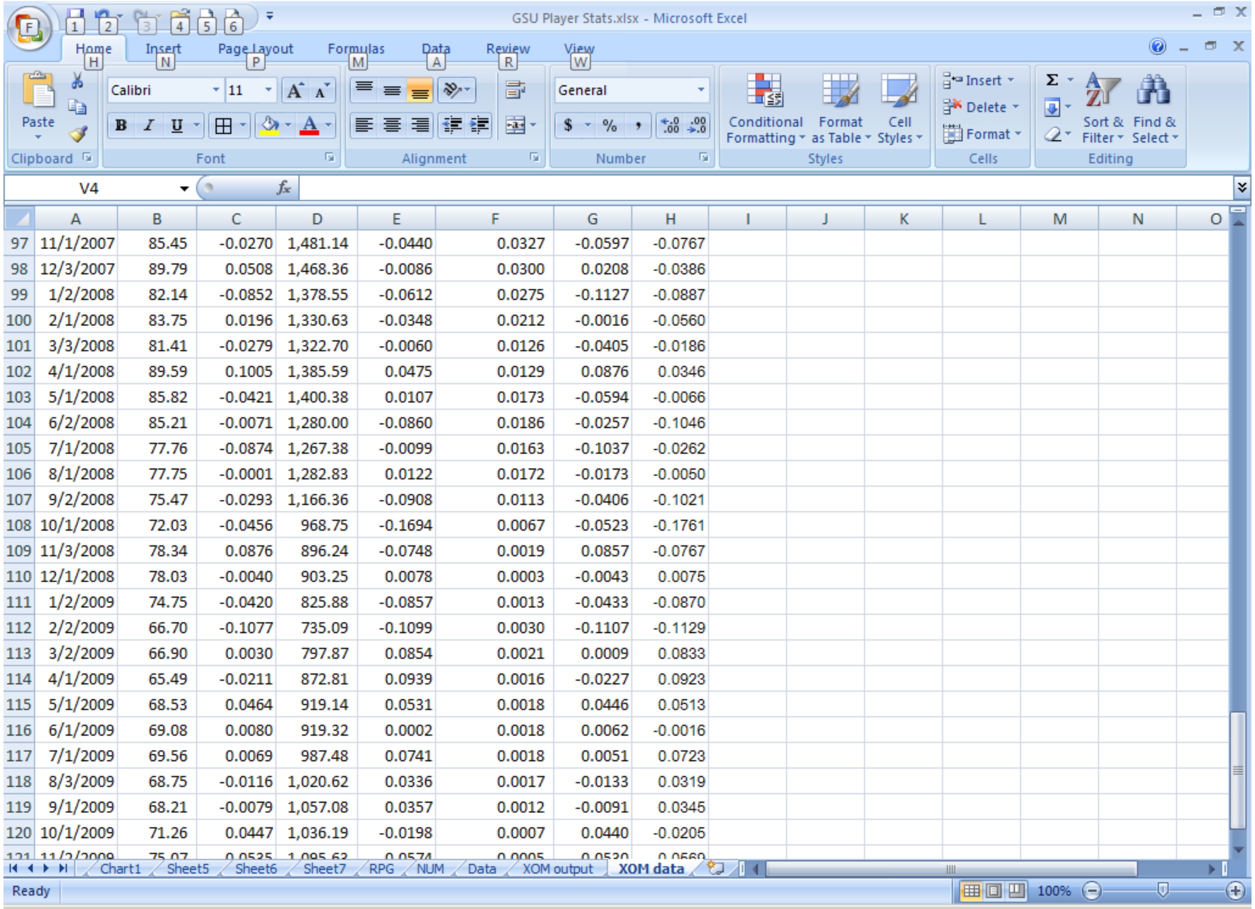


V4

	A	B	C	D	E	F	G	H
49	11/3/2003	32.04	-0.0034	1,058.20	0.0071	0.0093	-0.0127	-0.0022
50	12/1/2003	36.29	0.1326	1,111.92	0.0508	0.0090	0.1236	0.0418
51	1/2/2004	36.10	-0.0052	1,131.13	0.0173	0.0088	-0.0140	0.0085
52	2/2/2004	37.55	0.0402	1,144.94	0.0122	0.0093	0.0309	0.0029
53	3/1/2004	37.04	-0.0136	1,126.21	-0.0164	0.0094	-0.0230	-0.0258
54	4/1/2004	37.89	0.0229	1,107.30	-0.0168	0.0094	0.0135	-0.0262
55	5/3/2004	38.76	0.0230	1,120.68	0.0121	0.0102	0.0128	0.0019
56	6/1/2004	39.80	0.0268	1,140.84	0.0180	0.0127	0.0141	0.0053
57	7/1/2004	41.50	0.0427	1,101.72	-0.0343	0.0133	0.0294	-0.0476
58	8/2/2004	41.57	0.0017	1,104.24	0.0023	0.0148	-0.0131	-0.0125
59	9/1/2004	43.58	0.0484	1,114.58	0.0094	0.0165	0.0319	-0.0071
60	10/1/2004	44.38	0.0184	1,130.20	0.0140	0.0176	0.0008	-0.0036
61	11/1/2004	46.46	0.0469	1,173.82	0.0386	0.0207	0.0262	0.0179
62	12/1/2004	46.47	0.0002	1,211.92	0.0325	0.0219	-0.0217	0.0106
63	1/3/2005	46.78	0.0067	1,181.27	-0.0253	0.0233	-0.0166	-0.0486
64	2/1/2005	57.67	0.2328	1,203.60	0.0189	0.0254	0.2074	-0.0065
65	3/1/2005	54.30	-0.0584	1,180.59	-0.0191	0.0274	-0.0858	-0.0465
66	4/1/2005	51.95	-0.0433	1,156.85	-0.0201	0.0278	-0.0711	-0.0479
67	5/2/2005	51.46	-0.0094	1,191.50	0.0300	0.0284	-0.0378	0.0016
68	6/1/2005	52.62	0.0225	1,191.33	-0.0001	0.0297	-0.0072	-0.0298
69	7/1/2005	53.79	0.0222	1,234.18	0.0360	0.0322	-0.0100	0.0038
70	8/1/2005	55.12	0.0247	1,220.33	-0.0112	0.0344	-0.0097	-0.0456
71	9/1/2005	58.46	0.0606	1,228.81	0.0069	0.0342	0.0264	-0.0273
72	10/3/2005	51.66	-0.1163	1,207.01	-0.0177	0.0371	-0.1534	-0.0548
73	11/1/2005	53.67	0.0389	1,249.48	0.0352	0.0388	0.0001	0.0026

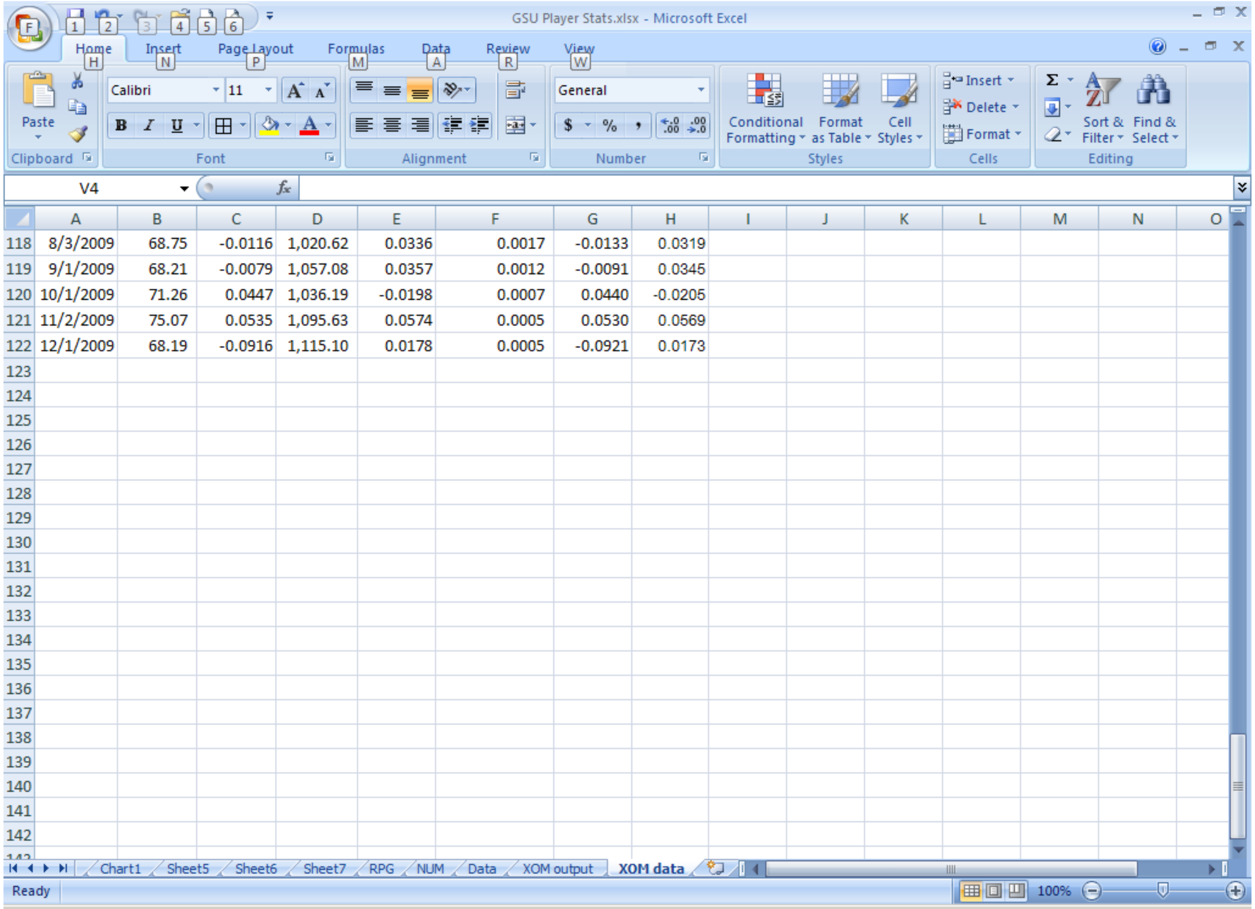


	A	B	C	D	E	F	G	H
73	11/1/2005	53.67	0.0389	1,249.48	0.0352	0.0388	0.0001	-0.0036
74	12/1/2005	51.95	-0.0320	1,248.29	-0.0010	0.0389	-0.0709	-0.0399
75	1/3/2006	58.03	0.1170	1,280.08	0.0255	0.0424	0.0746	-0.0169
76	2/1/2006	55.20	-0.0488	1,280.66	0.0005	0.0443	-0.0931	-0.0438
77	3/1/2006	56.58	0.0250	1,294.87	0.0111	0.0451	-0.0201	-0.0340
78	4/3/2006	58.65	0.0366	1,310.61	0.0122	0.0460	-0.0094	-0.0338
79	5/1/2006	56.91	-0.0297	1,270.09	-0.0309	0.0472	-0.0769	-0.0781
80	6/1/2006	57.33	0.0074	1,270.20	0.0001	0.0479	-0.0405	-0.0478
81	7/3/2006	63.30	0.1041	1,276.66	0.0051	0.0495	0.0546	-0.0444
82	8/1/2006	63.52	0.0035	1,303.82	0.0213	0.0496	-0.0461	-0.0283
83	9/1/2006	62.99	-0.0083	1,335.85	0.0246	0.0481	-0.0564	-0.0235
84	10/2/2006	67.04	0.0643	1,377.94	0.0315	0.0492	0.0151	-0.0177
85	11/1/2006	72.42	0.0803	1,400.63	0.0165	0.0494	0.0309	-0.0329
86	12/1/2006	72.25	-0.0023	1,418.30	0.0126	0.0485	-0.0508	-0.0359
87	1/3/2007	69.86	-0.0331	1,438.24	0.0141	0.0498	-0.0829	-0.0357
88	2/1/2007	67.87	-0.0285	1,406.82	-0.0218	0.0503	-0.0788	-0.0721
89	3/1/2007	71.44	0.0526	1,420.86	0.0100	0.0494	0.0032	-0.0394
90	4/2/2007	75.16	0.0521	1,482.37	0.0433	0.0487	0.0034	-0.0054
91	5/1/2007	79.09	0.0523	1,530.62	0.0325	0.0473	0.0050	-0.0148
92	6/1/2007	79.76	0.0085	1,503.35	-0.0178	0.0461	-0.0376	-0.0639
93	7/2/2007	80.95	0.0149	1,455.27	-0.0320	0.0482	-0.0333	-0.0802
94	8/1/2007	81.85	0.0111	1,473.99	0.0129	0.0420	-0.0309	-0.0291
95	9/4/2007	88.37	0.0797	1,526.75	0.0358	0.0389	0.0408	-0.0031
96	10/1/2007	87.82	-0.0062	1,549.38	0.0148	0.0390	-0.0452	-0.0242
97	11/1/2007	85.45	0.0270	1,481.14	0.0440	0.0327	0.0597	0.0767



V4

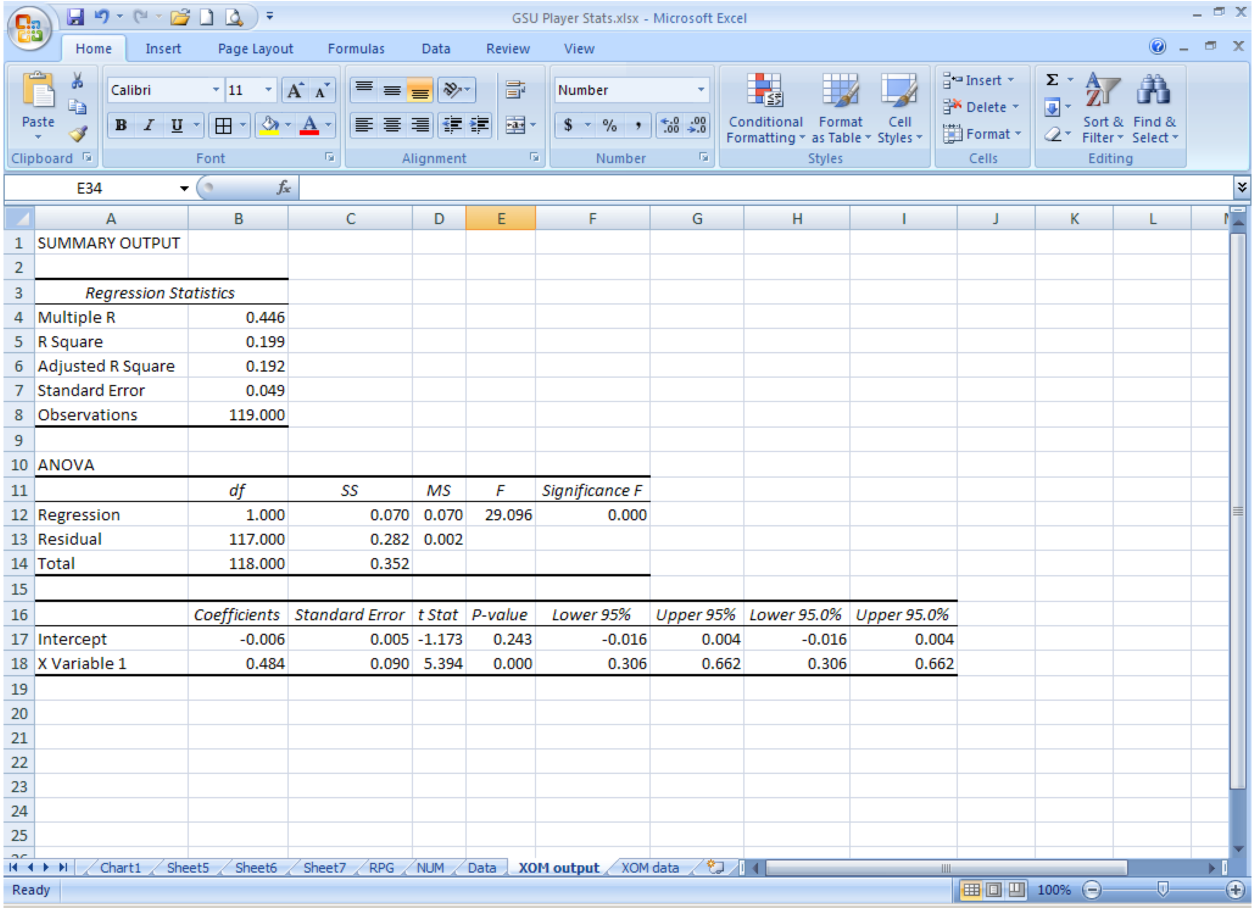
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
97	11/1/2007	85.45	-0.0270	1,481.14	-0.0440	0.0327	-0.0597	-0.0767							
98	12/3/2007	89.79	0.0508	1,468.36	-0.0086	0.0300	0.0208	-0.0386							
99	1/2/2008	82.14	-0.0852	1,378.55	-0.0612	0.0275	-0.1127	-0.0887							
100	2/1/2008	83.75	0.0196	1,330.63	-0.0348	0.0212	-0.0016	-0.0560							
101	3/3/2008	81.41	-0.0279	1,322.70	-0.0060	0.0126	-0.0405	-0.0186							
102	4/1/2008	89.59	0.1005	1,385.59	0.0475	0.0129	0.0876	0.0346							
103	5/1/2008	85.82	-0.0421	1,400.38	0.0107	0.0173	-0.0594	-0.0066							
104	6/2/2008	85.21	-0.0071	1,280.00	-0.0860	0.0186	-0.0257	-0.1046							
105	7/1/2008	77.76	-0.0874	1,267.38	-0.0099	0.0163	-0.1037	-0.0262							
106	8/1/2008	77.75	-0.0001	1,282.83	0.0122	0.0172	-0.0173	-0.0050							
107	9/2/2008	75.47	-0.0293	1,166.36	-0.0908	0.0113	-0.0406	-0.1021							
108	10/1/2008	72.03	-0.0456	968.75	-0.1694	0.0067	-0.0523	-0.1761							
109	11/3/2008	78.34	0.0876	896.24	-0.0748	0.0019	0.0857	-0.0767							
110	12/1/2008	78.03	-0.0040	903.25	0.0078	0.0003	-0.0043	0.0075							
111	1/2/2009	74.75	-0.0420	825.88	-0.0857	0.0013	-0.0433	-0.0870							
112	2/2/2009	66.70	-0.1077	735.09	-0.1099	0.0030	-0.1107	-0.1129							
113	3/2/2009	66.90	0.0030	797.87	0.0854	0.0021	0.0009	0.0833							
114	4/1/2009	65.49	-0.0211	872.81	0.0939	0.0016	-0.0227	0.0923							
115	5/1/2009	68.53	0.0464	919.14	0.0531	0.0018	0.0446	0.0513							
116	6/1/2009	69.08	0.0080	919.32	0.0002	0.0018	0.0062	-0.0016							
117	7/1/2009	69.56	0.0069	987.48	0.0741	0.0018	0.0051	0.0723							
118	8/3/2009	68.75	-0.0116	1,020.62	0.0336	0.0017	-0.0133	0.0319							
119	9/1/2009	68.21	-0.0079	1,057.08	0.0357	0.0012	-0.0091	0.0345							
120	10/1/2009	71.26	0.0447	1,036.19	-0.0198	0.0007	0.0440	-0.0205							
121	11/2/2009	75.07	0.0535	1,095.63	0.0574	0.0005	0.0530	0.0569							



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
118	8/3/2009	68.75	-0.0116	1,020.62	0.0336	0.0017	-0.0133	0.0319							
119	9/1/2009	68.21	-0.0079	1,057.08	0.0357	0.0012	-0.0091	0.0345							
120	10/1/2009	71.26	0.0447	1,036.19	-0.0198	0.0007	0.0440	-0.0205							
121	11/2/2009	75.07	0.0535	1,095.63	0.0574	0.0005	0.0530	0.0569							
122	12/1/2009	68.19	-0.0916	1,115.10	0.0178	0.0005	-0.0921	0.0173							
123															
124															
125															
126															
127															
128															
129															
130															
131															
132															
133															
134															
135															
136															
137															
138															
139															
140															
141															
142															

Figure 6 shows the results.

Figure 6



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.446
R Square	0.199
Adjusted R Square	0.192
Standard Error	0.049
Observations	119.000

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1.000	0.070	0.070	29.096	0.000
Residual	117.000	0.282	0.002		
Total	118.000	0.352			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.006	0.005	-1.173	0.243	-0.016	0.004	-0.016	0.004
X Variable 1	0.484	0.090	5.394	0.000	0.306	0.662	0.306	0.662

We expect α to be zero. If equation (10) is true, then equation (11) may only be true if α is zero. If the estimated α is positive then XOM generates returns greater than what the CAPM predicts and should be a security that is bought for this reason (i.e. the security generates more return than it theoretically should). If the estimated α is negative then XOM generates returns less than what the CAPM predicts and should be sold if already owned. Our estimate for α is -0.006 and the p-value is 0.243 indicating that α is not significantly different from zero.

We expect β to be different from zero. If it is, then returns on the S&P 500 are useful in explaining returns on XOM. Our estimate for β is 0.484 and the p-value is 0.00000036. (Note Figure 6 shows the rounded result 0.000). This p-value is less than 0.01 so we may say that $R_{S\&P500}$ is significant at the 1% level.

The r-square is 0.199. So only 19.9% of the variation of XOM returns are explained by returns on the S&P500. Can we improve the model? Are there other factors that are significant and will explain more of the variation of XOM returns? We leave that question for the reader/class to answer.

Conclusion

By using basketball data, one can introduce regression in the classroom in a very intuitive manner. To get the full benefit from the example, the instructor needs to have the students suggest and debate what variables matter when modeling a player's points per game (PPG). The variable, minutes per game (MPG), becomes a very reasonable suggestion because if you do not play, you cannot score. In this manner, students begin to see the value of regression analysis and what types of questions can be answered using regression analysis.

The earlier portion of the paper can be introduced before the basketball data to demonstrate what a regression actually is (i.e. an estimation tool based on minimizing mean squared error) or can be introduced after the basketball data to see how a regression determines the structure (i.e. the coefficient estimates) of the model given a set of data.

We encourage the instructor to go through this portion of the exercise because it takes regression out of the computer and into a form where the student can logic through the reasoning behind using one set of coefficient estimates versus another set of coefficient estimates. Logically, the regression should have errors that average to zero and that the size of the error should be minimized in some manner as well. The depth of the discussion can vary, but the point of the discussion should be that some coefficient estimates are better than others and that an optimal set of coefficient estimates should be found.

After working through both the first portion of the paper and the basketball portion of the paper, applying economic or financial data becomes the next logical step. We have presented a fairly easy example for the Capital Asset Pricing Model, but many other examples could be performed as well. We suggest using readily available data from the Internet or supplied by the instructor because the overall lesson should be focused on regression and not data issues.

After the students are more comfortable with regression, data issues can be introduced. One way to segue into such a conversation is to ask, what kind of basketball data would be better for modeling PPG? There is a short discussion earlier in the basketball section that considers a home-away variable and data frequency, however, let the students work through this issue through discussion because developing their intuition matters the most.

References:

Sharpe, William. 1964. “Capital asset prices: a theory of market equilibrium under conditions of risk.” *Journal of Finance*, 19 (3): 425-442.