

Teaching Regression Analysis Using Microsoft Excel

by

Timothy R. Mayes, Ph.D.

Associate Professor of Finance

Metropolitan State College of Denver

mayest@mscd.edu

Presented at the 2007 Financial Education Association annual meeting
Bermuda, 28 September 2007

Teaching Regression Analysis Using Microsoft Excel

Regression analysis (ordinary least squares, or OLS) is one of the most widely used statistical techniques in finance. Regression analysis is useful for:

- Prediction
- Relevant variable identification
- Model specification
- Parameter estimation

In other words, regression analysis helps us to understand and predict the data that we use. For this reason, it is important that students have a solid understanding of the technique. That includes not only the ability to “run” a regression model, but also understanding how and why it was created and to be able to interpret the output.

In this paper I will outline the procedure that I use to teach students about regression analysis. The technique uses Microsoft Excel, and has been very successful according to the feedback from my students. The purpose of the methodology presented here is to avoid the “black box” that is a statistics program (e.g., Minitab, SAS, etc), and to avoid complex math as much as possible. The technique that I present makes use of Excel’s charting features and the Solver add-in.

Regression in Excel

Microsoft Excel and other spreadsheet programs have always offered numerous ways to generate regression models or use them for forecasting. For example, consider the following worksheet functions:

Worksheet Functions	Purpose
Intercept()	Returns the intercept for bi-variate linear regression.
Slope()	Returns the slope for bi-variate linear regression.
Forecast()	Generate a forecast for Y, given a new X value (linear trend)
Growth()	Generate a forecast for Y, given a new X value (exponential trend)
LinEst()	Generates the coefficients and other statistics for a linear regression. Note that this is an array function, and returns data to multiple cells.
LogEst()	Same as LinEst(), but uses an exponential function.
Trend()	Similar to Forecast(), but doesn’t require X variables. Unless a new X value is provided, it returns the next Y on the trend line.

In addition to those functions, Excel also offers the Regression tool as part of the Analysis ToolPak add-in. This tool works much like a statistical program, and generates similar output. Also, those who are so inclined can make use of Excel’s matrix functions: MMult(), MInverse() and Transpose() to easily generate an array of coefficients.

These functions, and the regression tool, have their place. However, it seems that they are best used only after a student understands regression analysis.

A Different Methodology

I propose a different methodology that, instead of black box functions or tools, uses charts to visually present the data and the concept of fitting the best line. Once that concept is presented and understood, I use the Solver to show that the parameters of the regression line are those that minimize the sum of squared errors.¹

Though any data set may be used, my preference is to use a very simple (and small) data set that students can immediately understand: fictional college GPAs as the dependent variable, and high school GPAs and Verbal SAT scores as the independent variables. The data set consists of a sample of 10 student histories to keep it small, easily manageable, and all on the screen. There is also data for four students for whom we wish to predict the college GPA. I start by using only the high school GPA's as a single independent variable, and then later introduce the concept of multiple predictor variables. The data are presented in the table below:

Sample Data			
Student	College GPA	High-School GPA	Verbal SAT Score
1	3.8	4.0	750
2	2.7	3.7	380
3	2.3	2.2	580
4	3.2	3.8	510
5	3.5	3.8	620
6	2.4	2.8	440
7	2.6	3.0	540
8	3.0	3.4	650
9	2.7	3.3	480
10	2.8	3.0	550
Prediction Data			
11	?	3.4	550
12	?	2.8	400
13	?	3.2	650
14	?	3.7	450

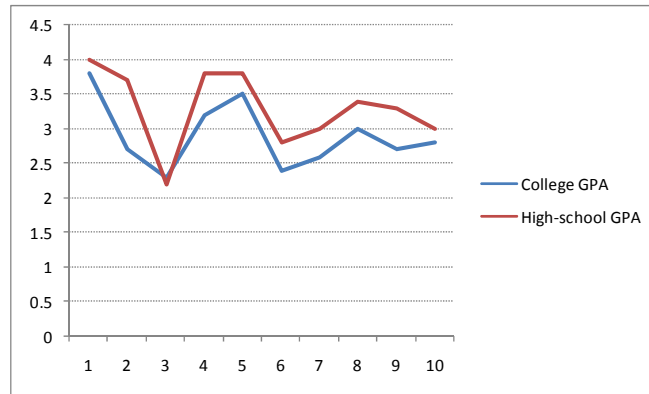
Charts First

Once the data are presented, I ask the class if they believe that college performance is in any way related to high school performance. Most will respond in the affirmative. I then ask what the relationship is: Positive or negative? Typically, they believe that, on average, those who do better in high school, as represented by GPA, will do better in college.

¹ Some will argue that the Solver is a "black box" tool. That is true enough. However, I am using it in a very simple and easy to explain way. In theory, you could use this technique without Solver. The Solver is mainly acting as an automation tool.

At this point, I suggest that we test that hypothesis by creating a chart. Usually, I start with a line chart showing the two data series so that they can see the relationship. This allows them to see the correlation in the simplest way possible.

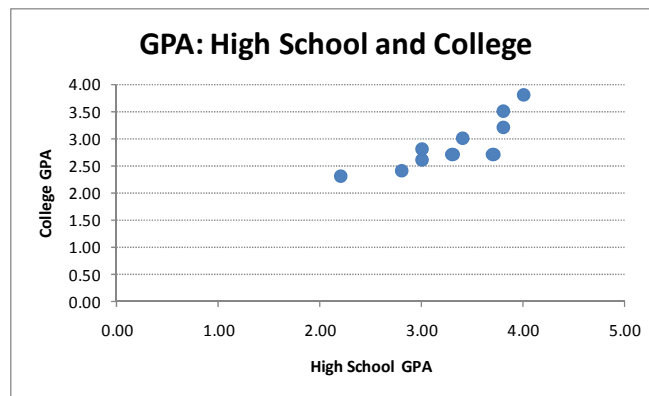
Figure 1: Line Chart of GPA Data



Upon seeing this chart, it is obvious that (at least with this fictitious data) that college GPA and high school GPA are strongly positively correlated. Furthermore, it is clear that most students get a lower GPA in college than in high school.

At this point, I suggest that the data may be better visualized by using an XY Scatter chart instead. The purpose of this step is to allow me to draw in several lines that “best fit” the data.

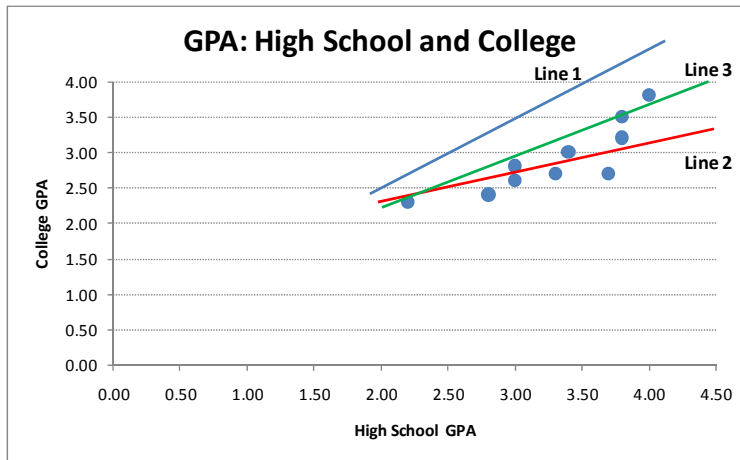
Figure 2: XY Scatter Chart of GPA Data



I now proceed to make several obviously poor attempts at drawing a line that best fits the data. I draw the lines using the line tool from the drawing tool bar (or the Shapes tool on the Insert ribbon in Excel 2007). It is easy to drag the line to a new location as I try to better fit the data.

Figure 3 shows several attempts to manually fit a line to the data. Line 1 (at the top) obviously won't work, but it allows me to introduce the idea of a biased model. In this case, Line 1 always predicts a value that is higher value than is realized in the actual data. At this point, I mention that we are looking for the B.L.U.E. (Best Linear Unbiased Estimator) line. Lines 2 and 3 represent better, but not perfect attempts.

Figure 3: XY Chart with Several Possible Best Fit Lines



I now reveal the fact that Excel can automatically insert the best fit line into the chart, and that it can even show the equation of the chart.² To do this simply right-click in the data series and choose Add Trendline from the short-cut menu.

Figure 4: Chart with Regression Line Inserted

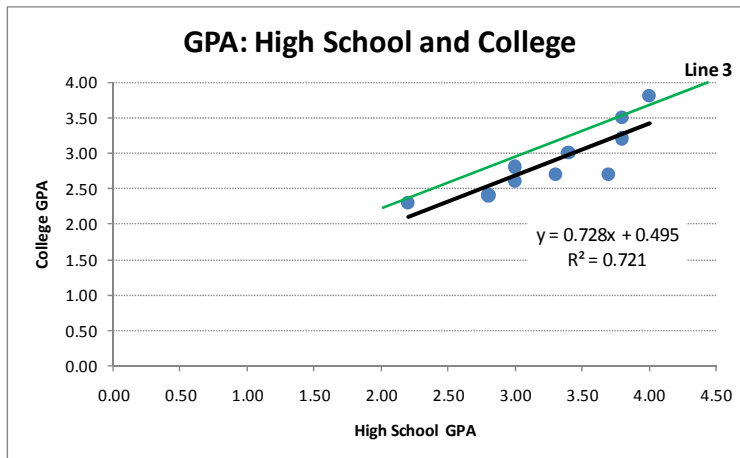


Figure 4 shows the resulting trend line and one of the lines that I drew in by hand. I now ask why this trend line is better than the line that I drew in manually. The answer, obviously, is that the trend line has a smaller total error (sum of squared errors, but I explain that later). This leads to an explanation of what the error is and how to measure it.

The basics of regression analysis are now obvious to most students. All that remains is to actually calculate the parameters (slope and intercept) of the line.

² You can also extend your hand drawn line to the Y-axis and then attempt to determine the approximate equation of your line. Most students will recall that the slope is defined as “rise over run.” Therefore, it is straightforward to have them estimate the equation.

Using the Solver to Calculate the Coefficients

I remind students that we are trying to find the equation of a line, and that they know the traditional equation for a line is:

$$Y = mX + b \text{ or } Y = b + mx$$

So, we are trying to find the optimal intercept and slope. However, since a straight line won't fit the data exactly, there will be an error at each data point. Therefore, our model is:

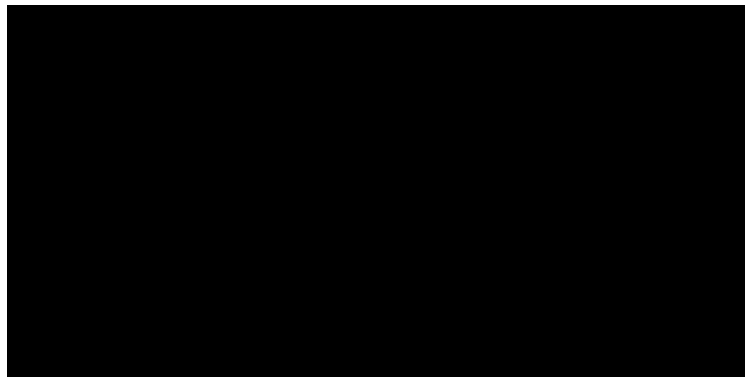
$$Y_i = \alpha + \beta x_i + \varepsilon_i$$

We want to find the intercept (α) and slope (β) such that we minimize the total error. I now point out that if our line is the best, then the errors will cancel out (some are positive, some are negative) so the total error will be 0. Therefore, we need to square the errors so that they don't cancel out, and try to find the parameters that give us the smallest sum of the squared errors.

I set up the following worksheet:



Figure 5: Regression Worksheet before Optimization



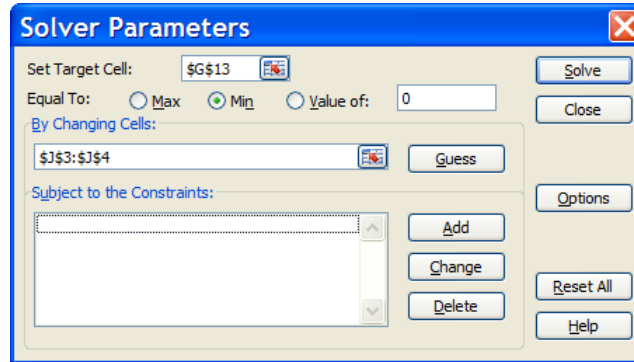
In column F, I enter a formula to calculate the predicted value based on some random numbers I entered for the intercept and slope in H3:H4. For example, in F3 the formula is: =J\$3+J\$4*C3. This formula is copied down for each of the other data points. The error is then calculated in column G. In G3 the formula is: =B3-F3 (actual – predicted). Again, this is copied down. The sum of squared errors is calculated in G13 with the formula: =SUM(G3:G12^2). (Note: That is an array formula and must be entered by pressing Ctrl+Shift+Enter.)

Once the students understand the structure of the worksheet, I suggest that we try to find the optimal intercept and slope manually. I plug some randomly chosen numbers into J3 and J4, trying to make the SSE get smaller. We keep track of the smallest value that we have achieved, and then try some other values. This continues until they get the idea (not long, usually).

Now is the ideal time to introduce the Solver. Typically, the students have had no prior exposure to this tool so I give a brief explanation of its abilities – basically, that it will do the trial and error process for us

automatically until it finds the slope and intercept that result in the smallest SSE. Setting up the Solver is especially easy for OLS regression because we don't need to enter any constraints. I set the Solver to minimize the SSE (G13) by changing the values for the intercept and slope (J3:J4). Figure 6 shows the Solver set up for this problem.

Figure 6: The Solver Set Up for Regression



Just press the Solve button and the Solver will find the optimal parameters for the line. Figure 7 shows the worksheet with the resulting optimal values.

Figure 7: The Final Result

	A	B	C	D	E	F	G	H	I	J
1	Sample Data									
2	Student	College GPA	High-school GPA	Verbal SAT Score		Predicted	Error		Intercept	0.4957
3	1	3.80	4.00	750		3.41	0.39		Slope	0.7286
4	2	2.70	3.70	380		3.19	(0.49)			
5	3	2.30	2.20	580		2.10	0.20			
6	4	3.20	3.80	510		3.26	(0.06)			
7	5	3.50	3.80	620		3.26	0.24			
8	6	2.40	2.80	440		2.54	(0.14)			
9	7	2.60	3.00	540		2.68	(0.08)			
10	8	3.00	3.40	650		2.97	0.03			
11	9	2.70	3.30	480		2.90	(0.20)			
12	10	2.80	3.00	550		2.68	0.12			
13	Prediction Data					SSE	0.5737			
14	11	?	3.40	550						
15	12	?	2.80	400						
16	13	?	3.20	650						
17	14	?	3.70	450						

It is now clear that the equation for the regression line is:

$$Y_i = 0.4957 + 0.7286X_i + \varepsilon_i$$

Where Y_i is the college GPA for student i , X_i is the high school GPA, and ε_i is the random error term.

Calculating the R^2

There are many typical regression statistics that you can now calculate. However, I usually focus on R^2 because it is important and easy to explain. I point out that what our model really does is attempt to explain why college GPAs vary, and that we are trying to explain that variability by using high school GPAs. So, the question is how much of the variability in college GPAs is explained by our model?

Students generally understand that we measure variability with the variance. I show them that Excel has a function, Var(), to calculate the variance. If we divide the variance of the predicted values by the variance of the college GPAs, then we can measure the proportion of the variability that is explained by the model. In other words, the R^2 .

I calculate the variance of the college GPAs in J6 with: =VAR(B3:B12). Similarly, I calculate the variance of the predicted values in J7 with: =VAR(F3:F12). Finally, I calculate the variance of the errors (the unexplained part) in J8 with: =VAR(G3:G12). Before calculating the R^2 , I usually show that the sum of the variance of the predicted values and the variance of the errors is equal to the variance of the college GPAs. Now the R^2 can be calculated in J10 with the formula: =J7/J6. Alternatively, we can calculate it as 1 – the proportion of the total variance that is unexplained: =1-J8/J6.

Figure 8: The R^2 and Predicted Values for Students 11 - 14

	A	B	C	D	E	F	G	H	I	J
1	Sample Data									
2	Student	College GPA	High-school GPA	Verbal SAT Score		Predicted	Error			
3	1	3.80	4.00	750		3.41	0.39	Intercept		0.4957
4	2	2.70	3.70	380		3.19	(0.49)	Slope		0.7286
5	3	2.30	2.20	580		2.10	0.20			
6	4	3.20	3.80	510		3.26	(0.06)	Var Y		0.2289
7	5	3.50	3.80	620		3.26	0.24	Var Pred		0.1651
8	6	2.40	2.80	440		2.54	(0.14)	Var Error		0.0637
9	7	2.60	3.00	540		2.68	(0.08)			
10	8	3.00	3.40	650		2.97	0.03	R-Squared		72.15%
11	9	2.70	3.30	480		2.90	(0.20)			
12	10	2.80	3.00	550		2.68	0.12			
13	Prediction Data					SSE	0.5737			
14	11	2.97	3.40	550						
15	12	2.54	2.80	400						
16	13	2.83	3.20	650						
17	14	3.19	3.70	450						

Figure 8 shows the worksheet with the R^2 , and I have also used the model to predict the college GPAs for the out-of-sample students. The formula, in B14, to predict student 14's GPA is: =J\$3+J\$4*C14. Copy that down to get the GPAs for the other students.

Multiple Regression

The groundwork has been laid, so adding additional explanatory variables is simple. Only a few changes are necessary. Select I6:J6 and insert cells. In I5 enter: Slope SAT, and change the label in I4 to: Slope HS. Now, change the formula for the predicted value in F3 to: =J\$3+J\$4*C3+J\$5*D3 and copy it down.

Before launching the Solver, I have the students make note of the SSE that we calculated using only the high school GPA (0.5737) so that we can compare the new model to see if it improves. Launch the Solver and change the "By Changing Cells" to J3:J5. Now, press Solve and the results should be the same as shown in Figure 9.

Figure 9: Multiple Regression

1	A	B	C	D	E	F	G	H	I	J
	Sample Data									
2	Student	College GPA	High-school GPA	Verbal SAT Score		Predicted	Error			
3	1	3.80	4.00	750		3.79	0.01		Intercept	(0.4109)
4	2	2.70	3.70	380		2.77	(0.07)		Slope HS	0.6311
5	3	2.30	2.20	580		2.27	0.03		Slope SAT	0.0022
6	4	3.20	3.80	510		3.13	0.07			
7	5	3.50	3.80	620		3.37	0.13		Var Y	0.2289
8	6	2.40	2.80	440		2.34	0.06		Var Pred	0.2194
9	7	2.60	3.00	540		2.69	(0.09)		Var Error	0.0094
10	8	3.00	3.40	650		3.19	(0.19)			
11	9	2.70	3.30	480		2.74	(0.04)		R-Squared	95.91%
12	10	2.80	3.00	550		2.71	0.09			
13	Prediction Data					SSE	0.0842			
14	11	2.96	3.40	550						
15	12	2.25	2.80	400						
16	13	3.06	3.20	650						
17	14	2.93	3.70	450						

Now update the formula in B14 to reflect the additional independent variable:
 $=\$J\$3+\$J\$4*C14+\$J\$5*D14$. Copy it down to get predictions for the remaining students.

Summary

In this paper I have introduced an alternative method of teaching regression analysis to undergraduate students. The methodology uses a simple spreadsheet that students can build as they follow the instructor's actions. The use of charts helps students to visualize the problem, but the key is to use Excel's Solver to handle the dirty work of finding the slope and intercept that minimizes the sum of squared errors.