The "Mother of All Guesses" -
A User-Friendly Guide to Statistical Estimation

by Francois Melese and David Rose

This five-step user's guide to statistical estimation is designed for individuals with little or no experience with regression modeling. The use of this powerful tool to make educated guesses is discussed in the context of a simple example in Microsoft Excel. The reader has three choices: read the text and ignore the integrated example; read the text and the example; or, read the text and reproduce the example. After reading and working the example we hope you'll be motivated to create you own model predictions. Improving our guesses can help save time and money.

1. Define the Problem.

The objective is to make an educated guess about the future value of something that's important to you. No crystal ball? Don't despair. Let us take you on a step-by-step journey to the "mother of all guesses"-a statistical estimate.

Example: Suppose we're asked to predict next year's operating costs for a newly formed computer-network service organization.

2. Identify the Prediction Variable.

Identify what you want. Call it the prediction variable. Can this variable be quantified-for example, measured in $'s, mph, time for repairs,...? If yes, is any historical data available? If the variable can be measured, and data are available, then collect the appropriate data. Alternatively, find a quantifiable proxy variable-one that approximates the original variable-supported with appropriate historical data.

Example and Exercise: A benchmarking study was conducted using a sample of 16 computer-network service organizations identified by the Government Accounting Office (GAO). Annual operating costs were reported for each organization for the past year. The sample data for our prediction variable-annual operating cost-ranges from a low of $26.89 million to a high of $125.10 million. In Excel type the cost data in the second column of Sheet 1 of your workbook in cells B3 to B18, or (B3:B18).

Center	Annual Operating Cost in $mil
1	56.57
2	89.12
3	36.70
4	83.70
5	70.21
6	73.10
7	80.06
8	125.10
9	89.63
10	26.89
11	77.53
12	81.68
13	60.51
14	86.89
15	50.87
16	106.30

What might account for the variation in operating costs?

3. Build a Prediction Model:
Uncover some explanations and develop a relationship. The objective is to make an educated guess of the future value of your prediction variable. It helps to build a prediction model. But like any good relationship, building a prediction model takes some effort. The challenge is to uncover explanations or factors that account for differences observed in the prediction variable data. The logic is simple. If these factors adequately explain past variations, they might help predict the future. The relationship you develop between these factors and your prediction variable is your prediction model. The prediction you generate from your prediction model is the "mother of all guesses."

A. Develop a list of relevant factors. What factors help explain, or drive, the variation you observe in your prediction variable data? The factors you include might come from your personal knowledge, judgment, or experience, the consensus of experts, or from theoretical economic or engineering models. These factors, are also called explanatory variables, or in the case of statistical cost estimation, cost drivers.

Example and Exercise: What factors might explain variations in annual operating costs in our sample of computer service organizations? Let's focus on three cost drivers:

Cost Driver	Definition
Size:	Number of employees
Coverage:	Square miles of service area
Workload:	Number of offices on network

B. Consider the effects. Once a relevant factor is identified, the next step is to consider what impact it has on your prediction variable. How are the two related? Is there a direct (or positive) relationship between the two-does the prediction variable increase with increases in the factor? Or is there an inverse (or negative) relationship between the two-does the prediction variable decrease with increases in the factor? Here, theoretical models, experience, judgment and instinct are all valid guides.

Example and Exercise: Examining one factor at a time, do you think a larger size (or coverage or workload) corresponds with higher or lower costs?

C. Collect data. Once you identify the relevant factors, collect the appropriate data. Make sure the data for the explanatory variables corresponds to the data for the prediction variable.

Example and Exercise: Data were collected on the size, coverage, and workload for each of the 16 organizations in the benchmarking study.

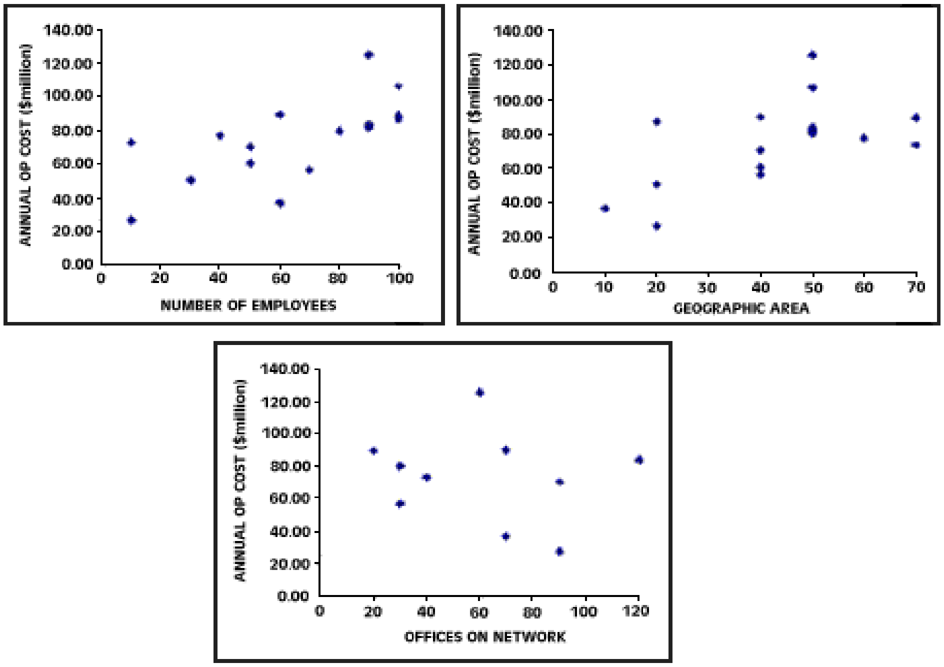
Center	Annual Operating Cost in $mil	Number of Employees	Geographic Area	Offices On Network
1	56.57	70	40	30
2	89.12	100	70	20
3	36.70	60	10	70

4	83.70	90	50	120
5	70.21	50	40	90
6	73.10	10	70	40
7	80.06	80	50	30
8	125.10	90	50	60
9	89.63	60	40	70
10	26.89	10	20	90
11	77.53	40	60	20
12	81.68	90	50	30
13	60.51	50	40	200
14	86.89	100	20	10
15	50.87	30	20	20
16	106.30	100	50	10

Type data for "Number of employees" in cells (C3:C18), data for "Geographic area" in cells (D3:D18), and data for "Offices on network" in cells (E3:E18) of Sheet 1. The resulting table of data in your spreadsheet should look similar to what is above.

D. Plot each factor independently against the prediction variable. Do the graphs confirm the relationships we expected? If not, there are only two things to review: our data and our logic.

Example and Exercise: In your Excel worksheet, copy cells (C3:C18) to cells (A3:A18) of Sheet 2. Copy cells (B3:B18) to cells (B3:B18) of Sheet 2. Select cells (A3:B18) of Sheet 2. In the Toolbar just under the Menu, press the button toward the right that looks like a little bar chart with a magic wand over it. This activates the Excel Chart Wizard. Move the cursor to an open part of the sheet and click the left mouse button. The Chart Wizard dialog box now opens. Choose Next to go to the next dialog screen. Now choose XY (Scatter) to make a scatter plot. On the next screen, choose Format 1. On the next screen, choose Finish. You should now see a scatter plot of Number of Employees against Cost. Repeat the procedure to create scatter plots of Geographic Area against Cost and Offices on Network against Cost. Your scatter plots should look similar to those below. Do you see the same relationships that you hypothesized in the previous question?



E. Develop a prediction model-use linear regression. Once we settle on a set of independent explanatory variables [1], the next step is to run a regression. Running a regression simply involves inputting data into a computerized regression package. Regression packages automatically uncover the best linear relationship between your explanatory variables and your prediction variable. [2] This best linear relationship is your prediction model. Regression results are typically reported in a table. Digesting the summary output takes a little practice. We discuss what to look for below. The ultimate result is the "mother of all guesses"-a statistical estimate, also known as a prediction or a forecast.

Example and Exercise: Go back to Sheet 1 of your workbook. From the Menu, select Tools - Data Analysis - Regression. Place the cursor in the "Input Y range" space, then select (B3:B18) from the sheet. Now place the cursor in the "Input X range" space, then select (C3:E18) from the sheet. Under Output Options, select New Worksheet Ply, and type "Results" in the associated space. Choose OK. After some crunching, you should have a sheet titled "Results" that looks like this:

Regression Statistics

Multiple R	0.8166
R Square	0.6669
Adjusted R Square	0.5836
Standard Error	15.9246
Observations	16

ANOVA

	df	SS	MS	F	Signifance F
Regression	3	6091.9404	2030.647	8.0075	0.0034
Residual	12	3043.1068	253.592		

Total	15	9135.0472				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	18.9199	15.0320	1.2586	0.2321	-13.8320	51.6719
Number of	0.4557	0.1391	3.2753	0.0066	0.1526	0.7589
Employees	0.6590	0.2376	2.7738	0.0168	0.1413	1.1766
Geographic Area	-0.0279	0.0848	-0.3284	0.7482	-0.2127	0.1570
Offices On Network						

F. Understand the model. Buried somewhere in the regression output are a set of coefficients. These are the keys that unlock your prediction model. The coefficients reveal the best linear relationship that exists between your explanatory variables and your prediction variable. To make a guess (or statistical estimate), simply enter the appropriate values into the prediction model and then do the required multiplication and addition.

Example: The prediction model buried in the summary output is:

Cost=18.92 + .456(X) + .659(Y) - .028(Z) where:

X = Number of Employees
Y = Area of Service
Z = Offices in Network

To obtain the "mother of all guesses", simply multiply the projected number of employees by .46, the projected area of service by .66, sum the two, then subtract .03 times the projected number of offices in the network, and add 18.9.

G. Check the model. Before we get too excited about a prediction, it's a good idea to see how much confidence we can have in our model. Here is a list of basic questions to answer:

- How well does the prediction model explain past variation in our prediction variable? The best we could hope for is that our model explains 100% of the past variation in our prediction variable. The R square (R) statistic reveals the fraction of any past variation explained by our model. An R measure lies somewhere between zero (the prediction model is no help) and one (the prediction model explains 100% of past variation). The closer R is to one, the more confidence we have in the model. But be cautious. It's possible to make useful predictions from models with low R, and useless predictions from models with high R. Let common sense prevail.

Example and Exercise: What is the R square (R) of the regression?

From our summary output, the R is .67. Thus the factors we included (i.e. Size, Workload and Coverage) explain almost 70% of the variation in our sample of operating costs.

- Do the explanatory variables have the right sign? Examine the signs of the coefficients. Does the data support our hypothesis of a direct (positive), or an inverse (negative), relationship between a specific factor and the prediction variable? If not, why not?

Example and Exercise: Do the signs of the coefficients agree with what you expected? In this case, the signs indicate Size and Workload directly increase costs (a positive relationship), while greater Coverage tends to reduce costs (a negative relationship). Do these results make sense? If not, perhaps the relationship is different than what we expected. Or something may be peculiar about the sample we chose. Is it truly representative of the job we have in mind? Or perhaps the explanatory variables used are not good proxies for what we really want to measure. We might want to consider other measures.

- Are the explanatory variables significant? If a coefficient is zero, the associated explanatory variable doesn't help-it is not statistically significant. Although coefficients reported in regression outputs are never zero, that doesn't mean they're significant. Since reported coefficients are derived using sample data, the reported coefficients are estimates, not the true values. The probability a coefficient is zero is given by its p-value. Say a reported coefficient has a p-value of .05. This means there is only a 5 in 100 chance we would get the reported coefficient value if the true coefficient were zero. In other words, there's a good chance the associated explanatory variable is significant.

Example: Which explanatory variables are significant? Look at the corresponding p-values. For example, the coefficient for the Number of employees has a p-value of .007. This means there is a tiny, 7 in a 1000, chance that the size of the estimated coefficient's value is simply due to sampling error. More likely, both the coefficient and the associated explanation are significant. On the other hand, the coefficient for Offices on network has a p-value of .748. That's a fairly high likelihood the estimated coefficient's value is a result of sampling error and is not statistically significant. In this case, we might be tempted to run another regression eliminating the variable. But then, why was the explanation included in the first place? If, upon reflection, we conclude the factor should really be there, then even if it's not significant, it's best to keep it, and to explore the apparent conflict between our theory and our data.

Congratulations! You now have a prediction model, and some appreciation of how good it is.

4. Guessing With Our Prediction Model:

Exactly wrong versus approximately right. A first guess: Exactly wrong. To obtain your first guess, you must enter the anticipated future values of your explanatory variables into your prediction model. So before you can predict anything, you need some educated guesses about the future values of your explanatory variables. Experts' opinions, trend models, regression models, or other techniques can be used to obtain these values. Then all it takes to get your prediction is some multiplication and addition.

Example and Exercise: Suppose we know the following about our new organization:

Number of employees:	About 50
Area of service:	40 square miles
Offices on network:	90

How would we predict the cost of this operation? Our prediction model suggests the cost will be about:

Cost=18.92 + .456(X) + .659(Y) - .028(Z) ------- or:

		18.92

X = Number of Employees = 50 x
Y = Area of Service = 40 x
Z = Offices in Network = 90 x
	+.456 =
	+.659 =
	-.028 =
		22.80
		26.36
		-2.52

		65.56

How confident can you be in your prediction? Consider this. In the example above, Firm 5 coincidentally has the same size, coverage and workload characteristics as the proposed new organization. Yet when we compare our prediction of operating cost of $65.51mil, with Firm 5's actual operating cost of $70.21mil, there is an unexplained difference of nearly $5 mil. This difference is due to the fact our model isn't perfect. Some of the variation in our prediction variable data is not accounted for by the model. As a consequence, point predictions are deceptively precise-they're often "exactly wrong."

A second guess: Approximately right. Another guess that is "approximately right involves building an interval around your point prediction. The interval accounts for both the point prediction and the unexplained variation. This prediction interval is sloppier to report, since it involves a range of values. However, unlike our first guess, we can express some confidence in our second guess. The standard error of the estimate is a measure of the unexplained variation. To construct a rough prediction interval around your point prediction, simply add and subtract twice the standard error of the estimate. You can be 95% confident the true value lies somewhere between the upper and lower bounds of your interval. This is truly the "mother of all guesses." [3]

Example and Exercise: Construct a 95% prediction interval. The prediction of future operating costs was $65.56 million. The reported standard error of the estimate is $15.92 million. The 95% prediction interval is million, or [$33.72 million, $97.40 million]. Based on our data, we can be 95% confident the true operating cost for our new organization will lie between $33.72 mil and $97.40 mil.

What if your boss objects to the sloppiness of the reported prediction interval? Unfortunately, there are only three ways to shrink a prediction interval: find better explanations, collect more data, or sacrifice confidence. The cruel reality of working with sample data is that we're forced to trade-off confidence for precision. Our first (precise) guess-a point prediction-is "exactly wrong." Our second (imprecise) guess-a prediction interval-is "approximately right."

5. Be careful!

Regression can be hazardous to your health. Now that you can build a prediction model, you're dangerous. You can hurt yourself, and others-unless you remember a few things:

- Do not extrapolate too far beyond the range of the observed data. If future factor values are much different than the means of our past factor data, you'll get sloppy results, i.e. wide prediction intervals. In this case a point prediction is exactly wrong, and your prediction interval doesn't help much. Then it's on to simulation models, or back to the crystal ball...

- Do not confuse correlation with causation. Your original model includes explanations or factors you believe help to understand past variations in your prediction variable.

> However, if recent structural changes (say the end of the cold war, revolutionary new software, the Internet, etc.) dramatically affect the process that generated your past data, then your prediction model may be worse than worthless-it could be misleading.

Finally, bear in mind that making educated guesses using prediction models is as much art as science. We often learn more about our problem from the formal process of obtaining educated guesses. So return to step one and review the prediction variable. Make sure you're making guesses about the right thing. Then make sure the prediction model still makes sense. Evaluate the possibility of obtaining more and/or better data to refine or test the model. Finally, remember that, much like in life, the process is as valuable as the product. The product is your prediction. The process makes you smarter.

Endnotes:

1. We must be careful in deciding which factors to include in our model. While we need to include factors that are related (or help explain variations in) the prediction variable, these factors should be unrelated to one another. If we, our experts, or our theory suggest two or more important factors should be included that are highly related to one another, we can only choose one-presumably the one most related to our prediction variable. If our explanatory variables are highly correlated, the regression cannot separate the effects of the explanatory variables on the prediction variable. This can eliminate any confidence we have in interpreting the marginal impact of a change in one of our explanatory variables on our prediction variable.

2. The best relationship is the one that best fits our data-the relationship that, when used to make guesses about what we already know, gives us the least errors. Given the appropriate data a linear regression can be computed using virtually any modern spreadsheet package. An example using Excel appears in the text. Regression packages report the linear relation (intercept and slope coefficients) that minimizes the sum squared errors between the guesses made by the linear relation and the actual historical values of the prediction variable.

3. There is nothing magic about 95%. All other things equal, the more confident you need to be the sloppier (or wider) the interval. More precise (narrower) intervals can be always be achieved at the expense of confidence. A better prediction interval also accounts for the fact that the further away future factor values are from the means of the past historical data, the sloppier (or wider) your prediction interval. This comes from the fact that we use model parameters that are estimated from sample data, not the true parameters. There are many other technical issues involved in using regression analysis. For more information, a good reference is *Applied Linear Regression Models* by Neter, Wasserman, and Kutner.

Francois Melese has been an Associate Professor at the Naval Postgraduate School since 1987. He received his BA in Economics from the University of California at Berkeley in 1977, MA in Economics from the University of British Columbia in 1979, and Ph.D. from the University of Louvain, Belgium in 1982. He was previously Assistant Professor of Economics at Auburn University.

David Rose has been an Assistant Professor at the Naval Postgraduate School since 1996. He received his Ph.D. in Decision Sciences from Rensselear Polytechnic Institute, an MA in Applied Mathematics from State University of New York at Buffalo, and a BS in Mathematics from Southwest Missouri state University.