Torturing Excel into Doing Statistics

By Steve Doig

Arizona State University

PREPARING YOUR SPREADSHEET

To do many statistical functions using Excel, you need to load the "Analysis ToolPak" add-in. In Office 2007, launch Excel, then click on the "Office button" logo in the upper lefthand corner of the window. Click on "Excel options" on the bottom of the popup. Click on "Add-ins" in the "Excel options" popup, choose "Analysis ToolPak", click "Go…", then put a checkmark in the box next to "Analysis ToolPak" on the list of available "Add-ins".

(Note: It's possible that Excel will tell you to insert the original installation CD. This will happen if your techs failed to load add-ins when they were installing Microsoft Office. Go beat up on them to get it done.)

(Note further: If you are using Office 2008 on a Macintosh, you're screwed. For no apparent reason I can understand, Microsoft removed the Analysis ToolPak option and other add-ins from this flavor of Excel. In short, you need an older version of Excel for Macs, or find a PC.)

To use the Analysis ToolPak as described below, click on the "Data" tab, then on "Data Analysis" to open the list of various statistical tools.

UNIVARIATE (DESCRIPTIVE) STATISTICS

You can use Excel functions to find the following statistics useful in describing single variables. "Range" means the appropriate beginning and ending cell addresses, such as B2:B100.

General

- TOTAL: =SUM(range)
- Number of values: =COUNT(range)

Measures of central tendency

- MEAN (the arithmetic average): =AVERAGE(range)
- MEDIAN (the middle of a sorted list): =MEDIAN(range)
- MODE (the most common value, though not necessarily near the middle): =MODE(range)

Measures of dispersion

- Maximum value: =MAX(range)
- Minimum value: =MIN(range)
- Range: There's no function for the distance between the MAX and MIN, but you can calculate it by subtracting MIN from MAX (duh!)
- N-tiles: Use the Percentile function to calculate the point in a range below which the given percentage of the values fall. For instance, to find the 90th percentile for a range of values, use =PERCENTILE(range, .9)
- Quartiles: This function is just a special case of the Percentile function. For instance, to find the top quartile (the 75th percentile), use =QUARTILE(range,3). The "3" refers to the third quartile.

Standard Deviation

The standard deviation of a range is another useful measure of dispersion. Think of it as the average distance of all the values from the mean, although that's not exactly how it's calculated. The standard deviation is useful in defining outliers. For instance, in a normal (bell curve) distribution like height or test scores, fewer than about 5% of the values will be more than three standard deviations above or below the mean of the values. Use =STDEV(range) to get the standard deviation.

Standardized scores (Z-scores)

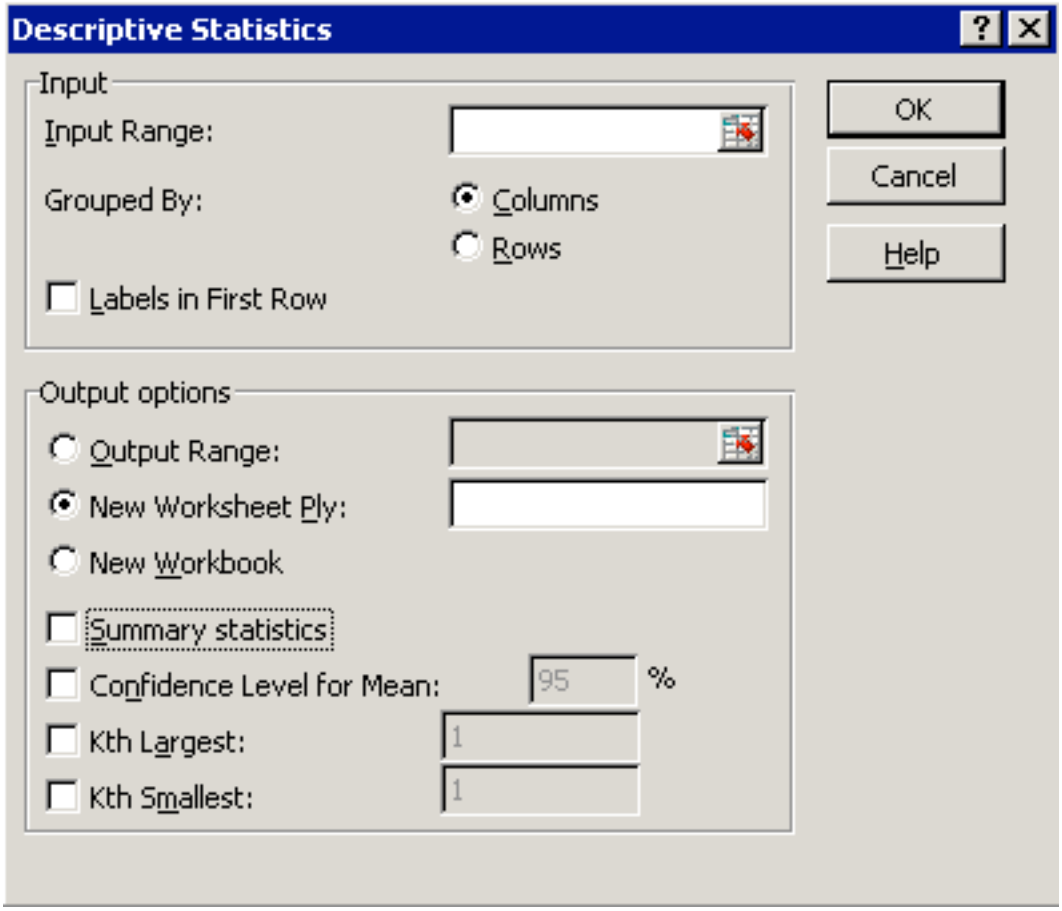
The Z-score is simply the number of standard deviations a particular value is above or below the mean of all the values. The result will be positive if the value in question is greater than the mean, and negative if it's less than the mean. The syntax of the function is =STANDARDIZE(value,average,standard deviation). For instance, to get the Z-score of the value in cell B2, part of a range from B2 to B100, write the formula like this:

=STANDARDIZE(B2,AVERAGE(B$2:B$100),STDEV(B$2:B$100))

Again, a Z-score of greater than 3 or less than -3 would be a likely way to define outliers.

Descriptive statistics using Analysis ToolPak

Another option is to get all your descriptives in one pass. Go to the Data tab, then Data Analysis. Pick "Descriptive Statistics" and hit OK. You'll get a window that looks like this:



Enter your input range, from the upper left corner to the lower right corner, like C1:D367. Check the "Labels in First Row" box and "Summary Statistics". Leave the output for a new worksheet. Hit OK. You'll get output on a new sheet that looks like this:

	A	B	C	D
1	*max temp*		*min temp*	
2				
3	Mean	85.90437158	Mean	57.00819672
4	Standard Error	0.823023975	Standard Error	0.729614669
5	Median	89	Median	58
6	Mode	103	Mode	46
7	Standard Deviation	15.74537575	Standard Deviation	13.95835051
8	Sample Variance	247.9168575	Sample Variance	194.8355491
9	Kurtosis	-1.293488462	Kurtosis	-0.917846133
10	Skewness	-0.24709146	Skewness	-0.026790013
11	Range	58	Range	58
12	Minimum	53	Minimum	28
13	Maximum	111	Maximum	86
14	Sum	31441	Sum	20865
15	Count	366	Count	366
16				

It gives you all the useful descriptives for each variable, as well as some more exotic ones used for diagnostics of your data.

MULTIVARIATE STATISTICS

Often you want to compare two or more sets of values to see if there is an interesting relationship between them. A classic U.S. journalism example of this would be to compare school test scores with poverty in each school, as measured by the percent of students eligible for free lunch. There are three related methods commonly used to examine such data for relationships: Correlation, linear regression, and x-y scatter plots.

Correlation

The "coefficient of correlation" (also called "Pearson's r") between two sets of numerical variables is a number between -1 and 1. If one value gets larger as the other value gets larger, then r will be positive; the closer to 1, the stronger the relationship. On the other hand, if one value goes up as the other goes down, r will be negative and the closer to -1, the stronger the relationship. If r=0, it means there is no relationship.

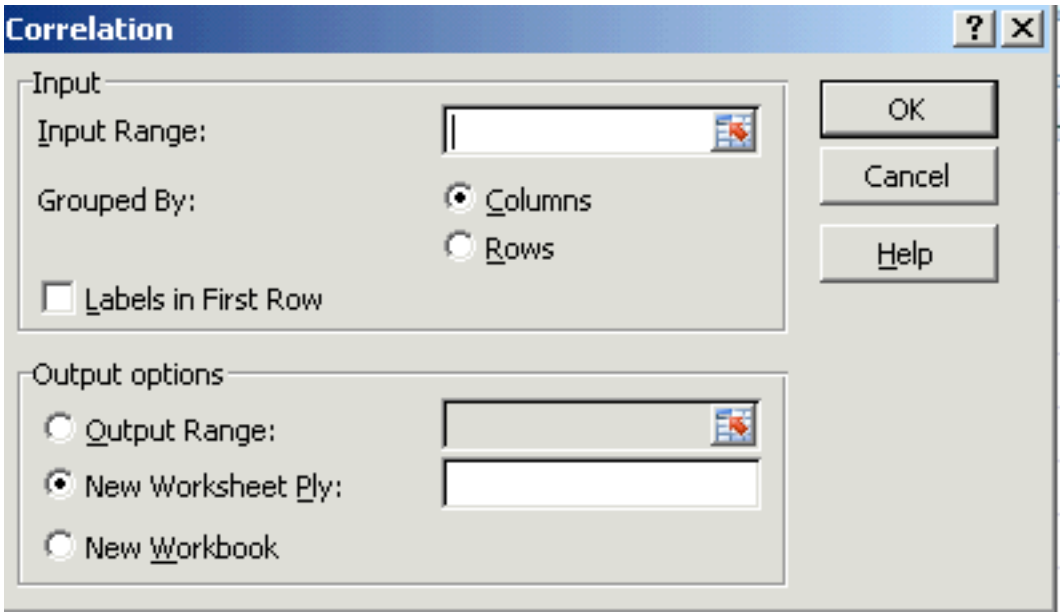
To calculate Pearson's r for two columns of data, use the CORREL function. For example, assume you have the Poverty Percent numbers in range C2:C201 and the Median Test Score numbers in range D2:D201. Use this formula: =CORREL(C2:C201, D2:D201). The result will be a number between -1 and 1. In this example, it almost certainly will be something like -.6 to -.8, a strong negative correlation, which means that as poverty goes up, scores go down.

Multiple Correlation of three or more variables

Sometimes you want to check several columns of numerical data to see which variables have relationships of some magnitude with which other variables. For instance, it might be American football team data showing yards gained, takeaways, giveaways, yards allowed, points scored and games won, like this:

	A	B	C	D	E	F	G
1	Team	Yards Gained	Takeaways	Giveaways	Yards Allowed	Points Scored	Games Won
2	Tennessee	5,350	30	30	3,813	346	13
3	Baltimore	5,014	49	26	3,967	333	12
4	New York Giants	6,376	31	24	4,546	328	12
5	Oakland	5,776	37	20	5,249	479	12
6	Minnesota	5,961	18	28	5,701	397	11
7	Philadelphia	5,006	31	29	4,820	351	11
8	Denver	6,567	44	25	5,544	485	11
9	Miami	4,461	41	26	4,636	323	11
10	Indianapolis	6,141	22	29	5,357	429	10
11	Tampa Bay	4,649	41	24	4,800	388	10
12	St. Louis	7,075	25	35	5,494	540	10
13	New Orleans	5,397	35	26	4,743	354	10
14	New York Jets	5,395	35	40	4,820	321	9
15	Pittsburgh	4,766	35	21	4,713	321	9
16	Green Bay	5,321	28	33	5,069	353	9
17	Detroit	4,422	42	31	5,033	307	9
18	Washington	5,396	33	33	4,474	281	8
19	Buffalo	5,498	29	23	4,426	315	8
20	Carolina	4,654	38	35	5,656	310	7
21	Jacksonville	5,690	30	29	4,845	367	7
22	Kansas City	5,614	29	26	5,293	355	7
23	Seattle	4,680	29	38	6,391	320	6

To do this, you want a Correlation Matrix. Once you have the Analysis ToolPak loaded, go to the Data tab, then click on Data Analysis, and choose "Correlation" from the list of tools. This window will pop up:



In "Input Range", put the upper left and lower right corners of your numerical data range, like B1:G32. (Be sure to check the "Labels in First Row" box.) For output, you're usually best with "New Worksheet Ply". Just leave it blank and Excel will put the result on a new sheet.

Click "OK" and you'll get results on a new sheet that look like this:

	A	B	C	D	E	F	G
1		*Yards Gained*	*Takeaways*	*Giveaways*	*Yards Allowed*	*Points Scored*	*Games Won*
2	Yards Gained	1					
3	Takeaways	0.054527278	1				
4	Giveaways	-0.328911463	-0.31702097	1			
5	Yards Allowed	-0.092363341	-0.46593399	0.243849975	1		
6	Points Scored	0.830526921	0.298211056	-0.359934561	-0.035998535	1	
7	Games Won	0.601483569	0.58259747	-0.541538751	-0.513343889	0.68135208	1
8							

The Pearson's r for each combination of variables is found in the intersection of the appropriate row and column. For instance, there's a very strong positive relationship (r=.83) between yards gained and points scored. On the other hand, there is little relationship (r=-.04) between points scored and yards allowed. And so on. (Note: The staircase of ones simply means there is a perfect relationship between any variable and itself.)

Linear Regression

More can be done with the concept of relationship between variables. A really useful tool is linear regression, which can be used to predict the value of a "dependent" variable given the value of one or more "independent" variables. Given that result, we can compare the predicted value with the actual one and see whether the actual was better or worse than expected, given the value of the other variable(s).

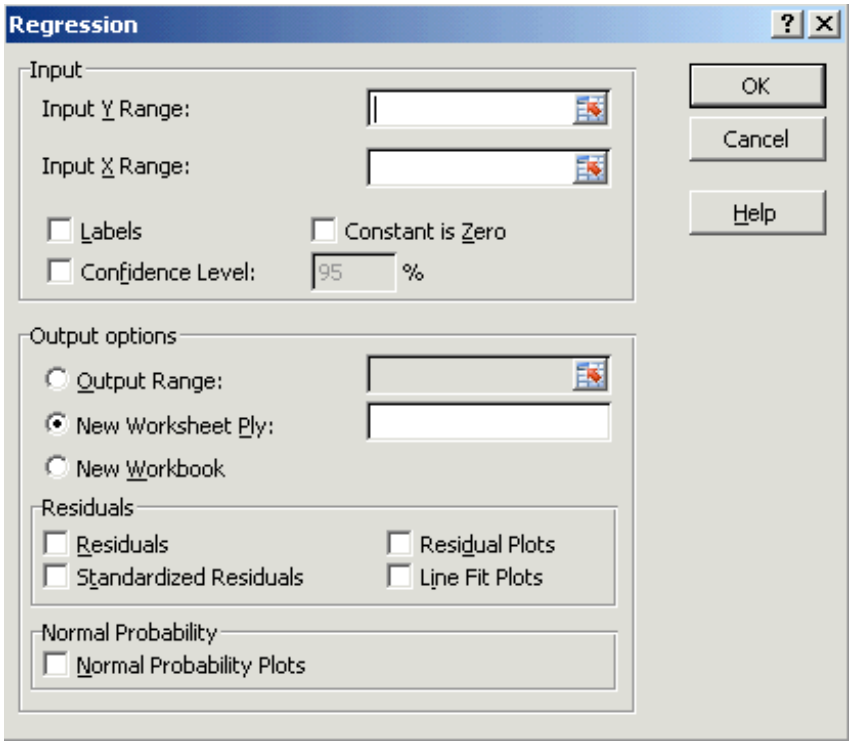
A good example, again, is the relationship between poverty and school scores. Before reporters learned to use regression, we did school test score stories by listing schools in descending order of scores; not surprisingly, the "best" schools typically were rich suburban ones and the "worst" were schools in poverty-stricken urban neighborhoods. Regression, however, lets us level the playing field and look at school scores as though all students were in the same economic level.

Consider this data set of 200 schools:

	A	B	C	D
1	DISTRICT	SCHOOL	LOW INCOME PERCENT	MEDIAN SCORE
2	PARKLAND	FOGELSVILLE SCH	4	63
3	WOODLAND HILLS	FAIRLESS INTRMD SCH	62	39
4	RIDLEY	EDGEWOOD EL SCH	11	63
5	PALMERTON AREA	TOWAMENSING EL SCH	20	54
6	MARS AREA	MIDDLESEX INTRMD EL SCH	24	56
7	HUNTINGDON AREA	WILLIAM SMITH EL SCH	67	35
8	TAMAQUA AREA	TAMAQUA EL SCH	29	49
9	NEW CASTLE AREA	WASHINGTON INTRMD EL	59	29
10	KEYSTONE CENTRAL	RENOVO EL SCH	72	28
11	LOWER MERION	GLADWYNE SCH	6	74

The independent variable is "low income percent" and the dependent is "median score". (How do you know which is which? Usually chronology will tell you. It makes sense that the poverty a child is reared in will affect his or her scores; it doesn't make sense that a child getting a low score will plunge the family into poverty.)

To do the regression, again go to the Data tab and "Data Analysis". Choose "Regression" and get this window:



This one is trickier to fill out. For "Input Y Range", put in the top and bottom of the dependent variable column; in this case, that's D1:D201. For "Input X Range" put in

the upper left to lower right corners of your independent variables. In this example, there is only one independent variable, so put in C1:C201. Click the "Labels" box, and ignore "Confidence Level" and "Constant is Zero". For now, don't check any of the other boxes. Then hit OK.

You'll get a new sheet with output that looks like this:

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.818948858							
5	R Square	0.670677232							
6	Adjusted R Square	0.669013986							
7	Standard Error	11.31053705							
8	Observations	200							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	51585.00182	51585	403.2339	1.21142E-49			
13	Residual	198	25329.79318	127.9282					
14	Total	199	76914.795						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	71.20698058	1.370639214	51.95166	2.5E-117	68.50405616	73.909905	68.50405616	73.909905
18	LOW INCOME PERCENT	-0.607638455	0.030259849	-20.0807	1.21E-49	-0.667311405	-0.547965506	-0.667311405	-0.547965506

It looks imposing, but you mostly need to pay attention only to the three numbers marked here:

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.818948858							
5	R Square	0.670677232							
6	Adjusted R Square	0.669013986							
7	Standard Error	11.31053705							
8	Observations	200							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	51585.00182	51585	403.2339	1.21142E-49			
13	Residual	198	25329.79318	127.9282					
14	Total	199	76914.795						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	71.20698058	1.370639214	51.95166	2.5E-117	68.50405616	73.909905	68.50405616	73.909905
18	LOW INCOME PERCENT	-0.607638455	0.030259849	-20.0807	1.21E-49	-0.667311405	-0.547965506	-0.667311405	-0.547965506

The three numbers of interest are:

- R Square: This is simply the square of Pearson's r. It is interpreted as "the percentage of variance from the mean explained". It tells you how good the independent variable is at predicting the dependent variable. In the example above, r squared is .67, or 67%. This means that the single variable of poverty accounts for two-thirds of the reason different schools are as far above or below the mean of all schools as they happen to be. It also means that other factors (teacher training, mobility, salaries, whatever) explain the remaining third of the variance from the mean.
- Intercept and slope: This takes you back to middle school algebra and the equation for a line: y = mx + b. B is the so-called y-intercept, the point where the line crosses the y axis. M is the slope of the line (remember "rise over

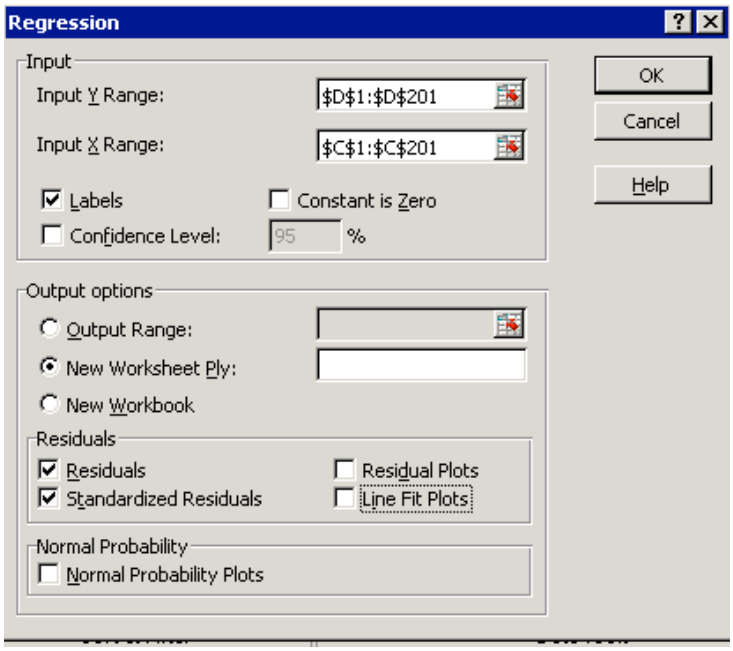
run”?) and can be positive (rising to the right) or negative (falling to the right)

In this case, y = mx +b becomes SCORE = -.61 * LOW_INCOME_PERCENT + 71.2

This formula can be used in two interesting ways:

- It can be turned into a statement you could make in a story. Multiply the slope (-.61) by 10 and you can say “Every 6 percentage point increase in poverty means a 10 point decrease in test scores.”
- The formula also can be used to calculate the predicted dependent variable value given a particular independent variable. Then this predicted value can be compared to the actual value of a particular record to see whether it is better or worse than might be expect. This difference between predicted and actual is known as the RESIDUAL.

You can create and use residuals using the Regression tool in the Analysis ToolPak. Fill out the Regression window like this:



Hit OK and you'll get output that looks like this (I formatted it to two decimals to make it easier to read:

RESIDUAL OUTPUT

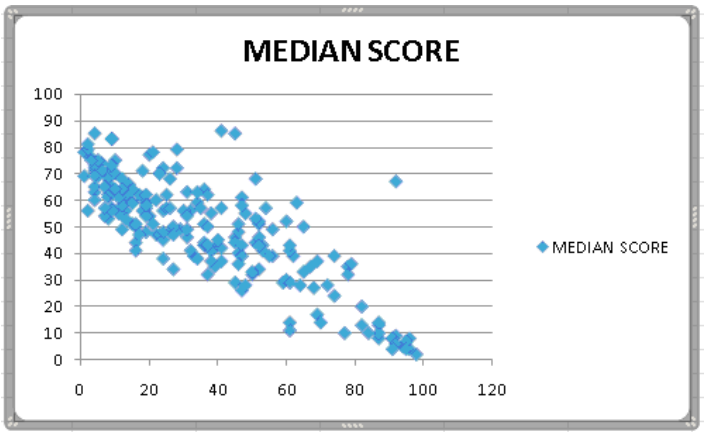
Observation	Predicted MEDIAN SCORE	Residuals	Standard Residuals
1	68.78	-5.78	-0.51
2	33.53	5.47	0.48
3	64.52	-1.52	-0.13
4	59.05	-5.05	-0.45
5	56.62	-0.62	-0.06
6	30.50	4.50	0.40
7	53.59	-4.59	-0.41
8	35.36	-6.36	-0.56
9	27.46	0.54	0.05
10	67.56	6.44	0.57
11	49.33	-6.33	-0.56
12	49.33	1.67	0.15
13	51.76	4.24	0.38
14	14.09	-9.09	-0.81
15	23.81	11.19	0.99
16	51.15	-12.15	-1.08
17	48.12	6.88	0.61

From the data, Observation 1 is Parkland district's Fogelsville School, with 4% poverty and a median score of 63. The residual table tells us that with a poverty rate of just 4%, the model predicts a median score of 68.78. So the residual is -5.78 -- below expectations. The z-score (standard residual) is -.51, so it really isn't that far off the mark.

But you can copy this table and paste it back on the original data sheet, then sort to look for outliers. At least one really stands out, a school with 92% poverty and a median score of 67, compared to a predicted score of just 15. The Z-score is 4.6, way off the charts. One possibility is that this is a magnet school that gives special attention to its carefully selected students; another possibility is that cheating is going on.

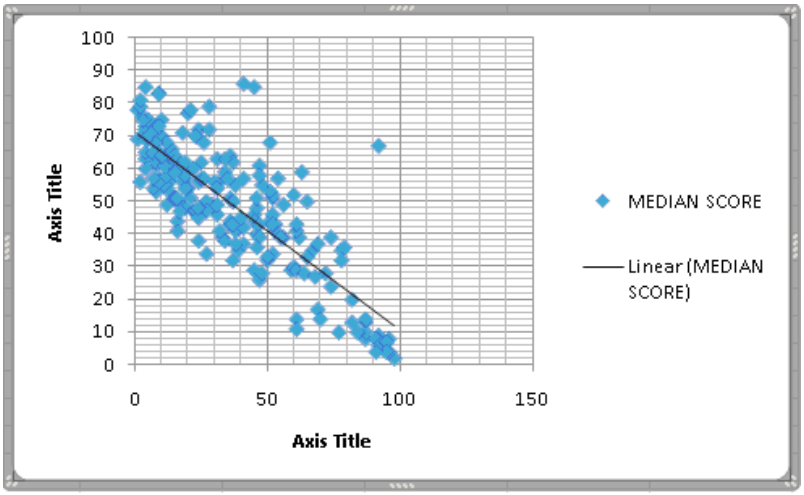
Using x-y scatter plots to look at your regression data

It's a good idea to visualize the "shape" of your data. To do that, select your two columns of data. Then go to Insert, then pick the "Scatter" button and choose the style that just has dots without lines. By default, you'll get something that looks like this:



You can see the clear linear pattern of the data: As the poverty percent on the horizontal X axis grows larger, the Median Scores get smaller. It's also easy to spot the outliers, like the school mentioned above with high poverty and high scores.

Under "Chart Layouts" you can choose the grid that has both dots and a straight line running through the data. You'll get this:



The straight line is the regression line that is defined by your y = mx + b formula. The closer all your points are to that line, the higher your r squared value will be.

OTHER STATISTICAL TOOLS

Excel has a variety of other statistical tests and tools available. But they are increasingly clumsy and problematic to use. If you get serious about using statistics for your work, you should start whining to the boss about getting SPSS. And if that doesn't work, download and install the open-source PSPP.