Regression Analysis

Author: John M. Cimbala, Penn State University
Latest revision: 11 September 2013

Introduction

- Consider a set of n measurements of some variable y as a function of another variable x.
- Typically, y is some measured ***output*** as a function of some known ***input***, x. Recall that the ***linear correlation coefficient*** is used to determine *if* there is a trend.
- If there *is* a trend, ***regression analysis*** is useful. Regression analysis is used to find an equation for y as a function of x that provides the ***best fit*** to the data.

Linear regression analysis

- ***Linear regression analysis*** is also called ***linear least-squares fit analysis***.
- The goal of linear regression analysis is to find the "best fit" straight line through a set of y vs. x data.
- The technique for deriving equations for this ***best-fit*** or ***least-squares fit*** line is as follows:
  - An equation for a straight line that attempts to fit the data pairs is chosen as $\boxed{Y = ax + b}$.
  - In the above equation, a is the ***slope*** ($a = dy/dx$ – most of us are more familiar with the symbol m rather than a for the slope of a line), and b is the ***y-intercept*** – the y location where the line crosses the y axis (in other words, the value of Y at $x = 0$).
  - An upper case Y is used for the fitted line to distinguish the fitted data from the *actual* data values, y.
  - In linear regression analysis, *coefficients a and b are optimized for the best possible fit to the data*.
  - The optimization process itself is actually very straightforward:
  - For each data pair (x_i, y_i), ***error*** e_i is defined as *the difference between the predicted or fitted value and the actual value*: e_i = error at data pair i, or $\boxed{e_i = Y_i - y_i = ax_i + b - y_i}$. e_i is also called the ***residual***. *Note*: Here, what we call the *actual* value does not necessarily mean the "correct" value, but rather the value of the actual measured data point.
  - We define E as the ***sum of the squared errors*** of the fit – a global measure of the error associated with all n data points. The equation for E is $\boxed{E = \sum_{i=1}^{i=n} e_i^{\,2} = \sum_{i=1}^{i=n} \left(ax_i + b - y_i\right)^2}$.
  - It is now assumed that ***the best fit is the one for which E is the smallest***.
  - In other words, ***coefficients a and b that minimize E need to be found***. These coefficients are the ones that create the best-fit straight line $Y = ax + b$.
  - How can a and b be found such that E is minimized? Well, as any good engineer or mathematician knows, to find a minimum (or maximum) of a quantity, that quantity is *differentiated*, and *the derivative is set to zero*.
  - Here, *two partial* derivatives are required, since E is a function of two variables, a and b. Therefore, we set $\boxed{\frac{\partial E}{\partial a} = 0}$ and $\boxed{\frac{\partial E}{\partial b} = 0}$.
  - After some algebra, which can be verified, the following equations result for coefficients a and b:

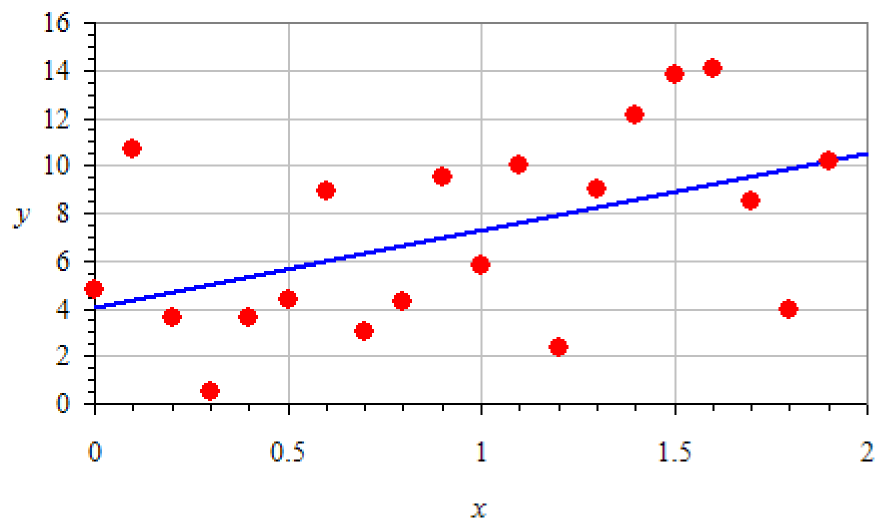
$$\boxed{a = \frac{n\sum_{i=1}^{i=n} x_i y_i - \left(\sum_{i=1}^{i=n} x_i\right)\left(\sum_{i=1}^{i=n} y_i\right)}{n\sum_{i=1}^{i=n} x_i^{\,2} - \left(\sum_{i=1}^{i=n} x_i\right)^2}} \text{ and } \boxed{b = \frac{\left(\sum_{i=1}^{i=n} x_i^{\,2}\right)\left(\sum_{i=1}^{i=n} y_i\right) - \left(\sum_{i=1}^{i=n} x_i\right)\left(\sum_{i=1}^{i=n} x_i y_i\right)}{n\sum_{i=1}^{i=n} x_i^{\,2} - \left(\sum_{i=1}^{i=n} x_i\right)^2}}.$$

- Coefficients a and b can easily be calculated in a spreadsheet by the following steps:
  - Create columns for x_i, y_i, $x_i y_i$, and x_i^2.
  - Sum these columns over all n rows of data pairs.
  - Using these sums, calculate a and b with the above formulas.
- Modern spreadsheets and programs like Matlab, MathCad, etc. have built-in regression analysis tools, but it is good to understand what the equations mean and from where they come. In the Excel spreadsheet that accompanies this learning module, coefficients a and b are calculated two ways for each example case – "by hand" using the above equations, and with the built-in regression analysis package. As can be seen, the agreement is excellent, confirming that we have not made any algebra mistakes in the derivation.

- **Example**:

Given: 20 data pairs (y vs. x) – the same data used in a previous example problem in the learning module about correlation and trends. Recall that we calculated the linear correlation coefficient to be $r_{xy} = 0.480$. The data pairs are listed below, along with a scatter plot of the data.

x	y
0.0	4.800
0.1	10.729
0.2	3.600
0.3	0.500
0.4	3.600
0.5	4.400
0.6	8.900
0.7	3.000
0.8	4.300
0.9	9.500
1.0	5.800
1.1	10.000
1.2	2.400
1.3	9.000
1.4	12.100
1.5	13.800
1.6	14.100
1.7	8.500
1.8	4.000
1.9	10.200



To do: Find the best linear fit to the data.

Solution:

- We use the above equations for coefficients a and b with $n = 20$; we calculate ***a* = 3.241**, and ***b* = 4.082**, to four significant digits. Thus, the best linear fit to the data is $Y = 3.241x + 4.082$.
- Alternately, using Excel's built-in regression analysis macro, the following output is generated:
  - Office 2003 and older: Tools-Data Analysis-Regression
  - Office 2007 and later: Data tab. In *Analysis* area, Data Analysis-Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.47999963
R Square	0.230399644
Adjusted R Square	0.187644069
Standard Error	3.600806066
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	69.86964115	69.86964	5.388763	0.032202476
Residual	18	233.3844778	12.9658		
Total	19	303.254119			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4.082114286	1.551752264	2.630648	0.016968	0.822001231	7.342227341
X Variable 1	3.241406015	1.396332701	2.321371	0.032202	0.307817598	6.174994432

- In Excel's notation, the *y*-intercept *b* is in the row called "Intercept" and the column called "Coefficients". The slope *a* is in the row called "X Variable 1" and the same column ("Coefficients"). The values agree with those calculated from the equations above, verifying our algebra.
- Notice also the item called "Multiple R". In Excel, ***Multiple R*** is the absolute value of the linear correlation coefficient, r_{xy}. For these example data, r_{xy} was calculated previously as 0.480, which agrees with the result from Excel's regression analysis (to about 7 significant digits anyway).
- The best-fit line is plotted in the above figure as the solid blue line.
- The best-fit line (compared to any *other* line) has the *smallest possible sum of the squared errors, E*, since coefficients *a* and *b* were found by *minimizing E* (forcing the derivatives of *E* with respect to *a* and *b* to be equal to zero).
- The upward trend of the data appears more obvious by eye when the least-squares line is drawn through the data.

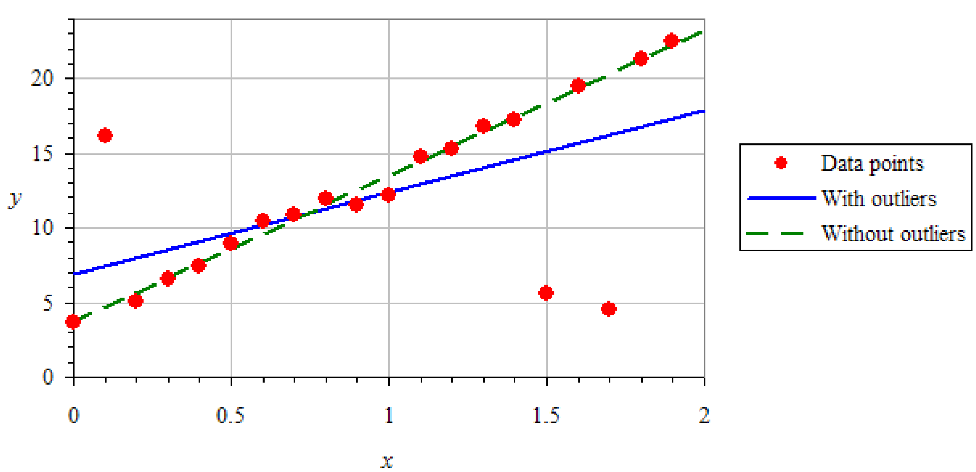
Discussion: Recall from the previous example problem that we could not judge *by eye* whether or not there is a trend in these data. In the previous problem we calculated the linear correlation coefficient and showed that we can be more than 95% confident that a trend exists in these data. In the present problem, we found the best-fit straight line that *quantifies* the trend in the data.

Standard error

- A useful measure of error is called the ***standard error of estimate***, $S_{y,x}$, which is sometimes called simply ***standard error***. For a linear fit, $S_{y,x} = \sqrt{\dfrac{\sum_{i=1}^{i=n}(y_i - Y_i)^2}{n-2}}$ which reduces to $S_{y,x} = \sqrt{\dfrac{\sum_{i=1}^{i=n} y_i^2 - b\sum_{i=1}^{i=n} y_i - a\sum_{i=1}^{i=n} x_i y_i}{n-2}}$.
- $S_{y,x}$ is a measure of the data scatter about the best-fit line, and has the same units as *y* itself.
- $S_{y,x}$ is a kind of "standard deviation" of the predicted least-squares fit values compared to the original data.
- $S_{y,x}$ for this problem turns out to be about 3.601 (in *y* units), as verified both by calculation with the above formula and by Excel's regression analysis summary. (See Excel's Summary Output above – Standard Error = 3.600806.)

Some cautions about using linear regression analysis

- ***Scatter in the y data is assumed to be purely random***. The scatter is assumed to follow a normal or Gaussian distribution. This may not actually be the case. For example, a jump in *y* at a certain *x* value may be due to some real, repeatable effect, not just random noise.
- ***The x values are assumed to be error-free***. In reality, there may be errors in the measurement of *x* as well as *y*. These are not accounted for in the simple regression analysis described above. (More advanced regression analysis techniques are available that can account for this.)
- ***The reverse equation is not guaranteed***. In particular, the linear least-squares fit for *y* versus *x* was found, satisfying the equation $Y = ax + b$. The *reverse* of this equation is $x = (1/a)Y - b/a$. This reverse equation is not necessarily the *best fit* of *x* vs. *y*, if the linear regression analysis were done on *x* vs. *y* instead of *y* vs. *x*.
- ***The fit is strongly affected by erroneous data points***. If there are some data points that are far out of line with the majority (***outliers***), *the least-squares fit may not yield the desired result*. The following example illustrates this effect:



- With all the data points used, the three stray data points (outliers) have ruined the rest of the fit (solid blue line). For this case, r_{xy} = 0.5745 and $S_{y,x}$ = 4.787.
- If these three outliers are removed, the least-squares fit follows the overall trend of the other data points much more accurately (dashed green line). For this case, r_{xy} = 0.9956 and $S_{y,x}$ = 0.5385. The linear correlation coefficient is significantly *higher* (better correlation), and the standard error is significantly *lower* (better fit).
- In a separate learning module we discuss techniques for properly removing outliers.
- To protect against such undesired effects, more complex least-squares methods, such as the ***robust straight-line fit***, are required. Discussion of these methods are beyond the scope of the present course.

Linear regression with multiple variables

- ***Linear regression with multiple variables*** is a feature included with most modern spreadsheets.
- Consider response, y, which is a function of m independent variables x_1, x_2, ..., x_m, i.e., $y = y(x_1, x_2, ..., x_m)$.
- Suppose y is measured at n operating points (n sets of values of y as a function of each of the other variables).
- To perform a linear regression on these data using Excel, select the cells for y (in one column as previously), and a *range* of cells for x_1, x_2, ..., x_m (in multiple columns), and then run the built-in regression analysis.
- When there is more than one independent variable, we use a more general equation for the ***standard error***, $S_{y,x} = \sqrt{\dfrac{\sum_{i=1}^{i=n}(y_i - Y_i)^2}{\text{df}}}$, where df = ***degrees of freedom***, $\text{df} = n-(m+1)$, n is the number of data points or operating points, and m is the number of independent variables.

- **<u>Example</u>**:
  Given: In this example, we perform linear regression analysis with multiple variables.
  - We assume that the measured quantity y is a linear function of three independent variables, x_1, x_2, and x_3, i.e., $y = b + a_1x_1 + a_2x_2 + a_3x_3$.
  - Nine data points are measured by setting three levels for each parameter, and the data are placed into a simple data array as shown to the right (the image is taken from an Excel spreadsheet).

x_1	x_2	x_3	y
0	0	0	10.472
0	-1	0	7.253
0	1	-1	14.708
-1	0	-1	9.861
-1	-1	1	4.374
-1	1	1	10.339
1	0	0	12.082
1	-1	-1	10.678
1	1	1	14.519

  To do: Calculate the y intercept and the three slopes *simultaneously*, one slope for each independent variable x_1, x_2, and x_3.
  Solution:
  - We perform a linear regression on these data points to determine the best (least-squares) linear fit to the data.
  - In Excel, the multiple variable regression analysis procedure is similar to that for a single independent variable, except that we choose *several* columns of x data instead of just one column:
    - Launch the macro (<u>Data Analysis</u>-<u>Regression</u>). The default options are fine for illustrative purposes.
    - The nine values of y in the y-column are selected for *Input <u>Y</u> range*.
    - All 27 values of x_1, x_2, and x_3, spanning *nine rows* and *three columns*, are selected for *Input <u>X</u> range*.
    - *Output Range* is selected, and some suitable cell is selected for placement of the output. <u>OK</u>.
    - Excel generates what it calls a **Summary Output**.
  - From Excel's output, the following information is needed to generate the coefficients of the equation for which we are finding the best fit, $y = b + a_1x_1 + a_2x_2 + a_3x_3$:
    - The ***y*-intercept**, which Excel calls ***Intercept***. For our equation, $b = \text{Intercept}$.
    - The three ***slopes***, which Excel calls ***X Variable 1***, ***X Variable 2***, and ***X Variable 3***. For our equation, $a_1 = \dfrac{\partial y}{\partial x_1} = \text{X Variable 1}$, $a_2 = \dfrac{\partial y}{\partial x_2} = \text{X Variable 2}$, $a_3 = \dfrac{\partial y}{\partial x_3} = \text{X Variable 3}$, which are the slopes of y with respect to parameters x_1, x_2, and x_3, respectively.
  - Note that we use ***partial derivatives*** (∂) rather than total derivatives (d) here, since y is a function of more than one variable.
  - A portion of the regression analysis results are shown below (image copied from Excel), with the most important cells highlighted:

Regression Statistics				
Multiple R	0.998606237			
R Square	0.997214417			
Adjusted R Square	0.995543068			
Standard Error	0.217477187			
Observations	9			
ANOVA				
	df	*SS*	*MS*	*F*
Regression	3	84.65837392	28.219458	596.652211
Residual	5	0.236481634	0.04729633	
Total	8	84.89485556		
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	10.47622222	0.072492396	144.514775	3.0097E-10
X Variable 1	1.918254902	0.090080885	21.2948052	4.2338E-06
X Variable 2	3.076078431	0.090080885	34.1479596	4.0504E-07
X Variable 3	-1.19547059	0.091358692	-13.085461	4.6512E-05

Discussion: The fit is pretty good, implying that there is little scatter in the data, and the data fit well with the simple linear equation. We know this is a good fit by looking at the linear correlation coefficient (***Multiple R***), which is greater than 0.99, and the ***Standard Error***, which is only 0.21 for y values ranging from about 4 to about 15. We can claim a successful curve fit.

Comments:

- In addition to random scatter in the data, there may also be ***cross-talk*** between some of the parameters. For example, y may have terms with products like x_1x_2, $x_2x_3^2$, etc., which are clearly *nonlinear* terms. Nevertheless, a multiple parameter linear regression analysis is often performed only ***locally***, around the operating point, and the linear assumption is reasonably accurate, at least close to the operating point.
- In addition, variables x_1, x_2, and x_3 may not be totally independent of each other in a real experiment.
- Regression analysis with multiple variables becomes quite useful to us later in the course when we discuss optimization techniques such as ***response surface methodology***.

Nonlinear and higher-order polynomial regression analysis

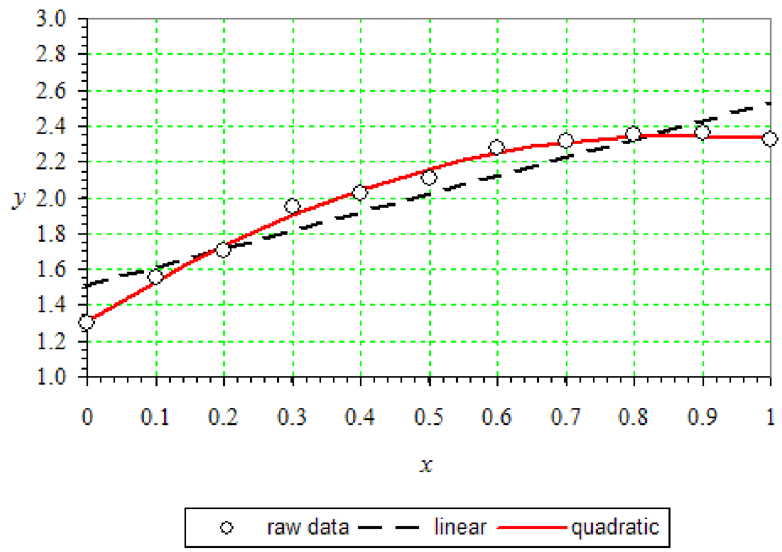
- ***Not all data are linear***, and a straight line fit may not be appropriate. A good example is thermocouple voltage versus temperature. The relationship is nearly linear, but not quite; that is in fact the very reason for the necessity of thermocouple tables.
- For nonlinear data, some transformation tricks can be employed, using logarithms or other functions.
- For some data, a good curve fit can be obtained using a ***polynomial fit*** of some appropriate order. The ***order*** of a polynomial is defined by *m*, *the maximum exponent in the x data*:
  - ***zeroth-order*** ($m = 0$) is just a constant: $y = b$.
  - ***first-order*** ($m = 1$) is a constant plus a linear term: $y = b + a_1 x$. (A first-order polynomial fit is the *same* as a linear least-squares fit, as we have already learned how to do.)
  - ***second-order*** ($m = 2$) is a constant plus a linear term plus a quadratic term: $y = b + a_1 x + a_2 x^2$. (A second-order polynomial fit is often called a ***quadratic fit***.)
  - ***third-order*** ($m = 3$) adds a cubic term: $y = b + a_1 x + a_2 x^2 + a_3 x^3$. (A third-order polynomial fit is often called a ***cubic fit***.)
  - ***m^th^-order*** ($m > 0$) adds terms following this pattern up to $a_m x^m$: $y = b + a_1 x + a_2 x^2 + a_3 x^3 + ... + a_m x^m$.
- Excel can be manipulated to perform least-squares polynomial fits of any order *m*, since Excel can perform regression analysis on more than one independent variable simultaneously. The procedure is as follows:
  - To the right of the *x* column, add new columns for x^2, x^3, ... x^m.
  - Perform a multiple variable regression analysis as previously, except choose *all* the data cells (x, x^2, x^3, ... x^m) as the "Input X Range" in the *Regression* working window.

- Note that m is the order of the polynomial, which is also treated as the number of independent variables to be fit. Excel treats each of the m columns as a separate variable. The output of the regression analysis includes the y-intercept as previously (equal to our constant b), and also a least-squares coefficient for *each* of the columns, i.e., for each of the variables x, x^2, x^3, ... x^m:
  - The coefficient for "X Variable 1" is a_1, corresponding to the x variable.
  - The coefficient for "X Variable 2" is a_2, corresponding to the x^2 variable.
  - The coefficient for "X Variable 3" is a_3, corresponding to the x^3 variable.
  - ...
  - The coefficient for "X Variable m" is a_m, corresponding to the x^m variable.
- Finally, the fitted curve is constructed from the equation, i.e., $y = b + a_1x + a_2x^2 + a_3x^3 + ... + a_mx^m$.

- **Example**:

Given: x and y data pairs, as shown:

x	y
0	1.30
0.1	1.55
0.2	1.70
0.3	1.95
0.4	2.02
0.5	2.11
0.6	2.28
0.7	2.31
0.8	2.35
0.9	2.36
1	2.32



To do: Plot the data as symbols (no line), perform a linear least-squares fit, and plot the data as a dashed line (no symbols), and perform a second-order polynomial least-squares fit, and plot the data as a solid line (no symbols).

Solution:

- We plot the data as symbols, as shown on the above plot.
- We perform a standard linear regression analysis, and then generate the best-fit line by using the equation for the best-fit straight line, $\boxed{Y = ax + b}$. For these data, a = 1.025 and b = 1.510. The result is plotted as the dashed black line in the figure – the agreement is not so good. The standard error is 0.1359.
- We add a column labeled x^2 between the x and y columns, and fill it in.
- We perform a multiple variable regression analysis, using the x and x^2 columns as our range of independent variables. We generate the best-fit quadratic (2nd-order) polynomial curve by using the equation $\boxed{y = b + a_1x + a_2x^2}$. For these data, b = 1.307, a_1 = 2.382, and a_2 = –1.358. The solid red line is plotted above for this equation – the agreement is much better. The standard error is 0.0316.

Discussion: These data fit much better to a second-order polynomial than to a linear fit. We see this both "by eye", and also by comparing the standard error, which decreases by a factor of more than four when we apply the quadratic (second-order) curve fit instead of the linear curve fit.