Environmental Data Laboratory
Professor Kirchner

Laboratory 10: Pitfalls in regression and correlation analyses

Simple linear regression and correlation are perhaps the most widely used data analysis techniques in environmental science. Most of those who use these techniques assume that their results come from an underlying linear relationship that has been faithfully revealed in the regression and correlation statistics, without any distortion. It is important to recognize that the regression slope and the coefficient of determination (r^2)--or the correlation coefficient r--can be affected by several factors, including:

(a) random measurement errors in X or Y,
(b) artifacts from the same variable(s) appearing on both the X and Y axes,
(c) limitations in the range of variation of X, or Y, or both, and
(d) serial correlation in the data.

In this lab, you will explore how each of these affects linear regression relationships. These pathologies will be discussed in lecture within the next week or so; this lab is intended to give you a direct, experiential appreciation of what we'll be talking about. To keep the analyses as simple as possible, the data you will be working with will be completely synthetic. You will start with an idealized relationship, whose characteristics you think you know well, and then introduce the complicating factors (a) through (d) above, and see how they affect the results. Things you need to respond to in writing are indicated by the symbol "•".

Part (A): Effects of random measurement errors in X and Y.

Open the JMP data file, "Lab 9, part A (jmp)". This file takes a simple linear relationship between two variables, "base X" and "base Y", and adds some random noise to each variable. Plot "base Y" as a function of "base X", fit the linear regression line, and •1: write down the regression slope, y-intercept, and r^2. Now, how would this regression relationship be affected by random, unbiased measurement error in the Y variable? Observe the three columns called "Y+error(10)", "Y+error(20)", and "Y+error(40)". These have the formulas:

*base Y + Random Normal()*10* , *base Y + Random Normal()*20* , and *base Y + Random Normal()*40*

Note that these formulas take each of the base Y values, and add a normally-distributed random error with a standard deviation of 10, 20, or 40 (for comparison, base X and base Y both have standard deviations of 60). Now use "fit Y by X" to plot each of the three new y's as a function of "base X". For each plot, fit the linear regression line and •2: write down the slope, y-intercept, and r^2. •3: Can you summarize how measurement error in the Y variable affects the regression slope, intercept, and r^2?

Now, how would the same regression relationship be affected by random, unbiased measurement error <u>in the X variable</u>, instead of the Y variable? The file contains three more columns, "X+error(10)", "X+error(20)", and "X+error(40)", with the same formulas as above, except that they add random error to "base X" rather than "base Y". Use "fit Y by X" to plot base Y as a function of base X and each of the three error-corrupted X variables. For each plot, fit the linear regression line and •4: write down the slope, y-intercept, and r^2. •5: Summarize how measurement error in the X variable affects the regression slope, intercept, and r^2. •6: Why do X-axis measurement errors and Y-axis measurement errors affect the regression relationship differently?

Part B: Effects of shared variables on X and Y axes.

Open the JMP data file, "Lab 9, part B (jmp)". You will notice that it contains two columns, labeled simply "A" and "B". A and B are both normally distributed; the standard deviation of B is three times bigger than the standard deviation of A. Plot A as a function of B. Note that there is no correlation between A and B.

Now create a new variable, called C, that equals A+B. Plot C as a function of A, and C as a function of B. Fit each with a linear regression line, and •7: note the slope, intercept and r^2. •8: Why is C correlated with A, and correlated with B? Why is C much better correlated with B than with A? (Hint: remember that the variability in A is smaller than the variability in B.)

Now create two more new variables, one called D, where D=A*B, and another called E, where E=A/B. Plot D and E as functions of A and B. Also create a new variable that is $\log_{10}(E)$, and plot this as a function of B and/or log(B). •9: Why are D and/or E correlated with A and/or B? •10: If you did not know already that C, D, and E all contain A and B by definition, would you suspect it just from looking at the data?

Part C: Effects of constrained variation in X or Y.

Open the JMP data file, "Lab 9, part C (jmp)". This file contains two columns, X and Y. Plot Y as a function of X, and •11: note the slope, intercept, and r^2. Let's suppose that Y is the response of an experimental system to an imposed condition X. •12: Would you conclude that X was an important causal factor affecting Y? Now, what if only a restricted range of X were examined (perhaps because the measuring instrument does not cover the full range, or because a narrow range of experimental conditions were used)? To simulate this, delete all the rows corresponding to X values less than 150, and all rows with X greater than 250 (X has been sorted into ascending order to make this convenient). Again plot Y as a function of X, and •13: note the slope, intercept, and r^2. •14: Would you still conclude that X was an important causal factor in Y? Now restrict the range of X further, by deleting all rows with X values less than 175 or greater than 225, and •15: again note the slope, intercept, and r^2. •16: Would you still conclude that X was a causal factor in Y? •17: Do your observations suggest anything about what range of experimental conditions is desirable, if you want to determine how a causal factor (X) affects some phenomenon of interest (Y)?

Part D: Serial correlation in data (optional--do if time permits).

Regression and correlation analyses assume that the errors in successive measurements are uncorrelated; that is, they assume that whether the i^{th} measurement is above or below the regression line has no effect on whether the $i+1^{st}$ measurement will be above or below the line. Serially correlated data violate this assumption, and can yield misleading regression statistics. Many environmental time series data show serial correlation, particularly time series that result from the cumulative effects of a fluctuating causal factor. Consider, for example, the population of organisms in a pond. The population each day equals the population on the previous day, plus the net change in population (reproduction minus mortality) over the 24-hour period. The net population growth rate may fluctuate (in response to weather, nutrient availability, predation, and so forth), but the population on one day depends on the population on the previous day, and the day before that, and so forth; the population on the i^{th} day equals the population on the $i-1^{st}$ day, plus or minus the net growth. Thus, population measurements on successive days are not independent.

This phenomenon is illustrated in the JMP data file, "Lab 9, part D (jmp)". This file contains three columns. The first is simply an index variable for time. The second is the change in some variable (population of organisms, or some such) from each day to the next. You can inspect the formula underlying this column to convince yourself that it is truly random and unbiased. The third column is the running sum of these changes (that is, the integrated changes in the variable. •18: Plot the second and third columns as a function of time, fit them with regression lines, and note the slope, r^2, and statistical significance (indicated by "prob>t" for the parameter "day"--this is a t-test for whether the regression slope is significantly different from zero). •19: If the change in Y is random from day to day, why does Y itself show such a clear correlation with time? •20: Could you tell from looking at the data that this resulted from a serial correlation in the data, rather than an underlying linear trend? Remember that there is nothing in the generating functions that makes Y prefer to go one direction or the other, and there's nothing that makes Y tend to continue going in whatever direction it's already heading (some will immediately recognize this as Brownian motion, or a "drunkard's walk"). If you want to see how this data set evolves over a longer period, you can simply add more rows to the data file; the formulas will recalculate automatically. Note that although the time series simply wanders over the long run, for substantial periods of time it can appear to show a very definite trend.