Environmental Data Laboratory
Professor Kirchner

Laboratory 13: Trend detection with multiple regression

Phosphorus (P), usually present in natural waters as phosphate (PO_4), is an important limiting nutrient in many aquatic ecosystems. Controlling phosphorus inputs to Lake Erie has been a major concern since the 1970's, because high PO_4 concentrations can lead to eutrophication. The Maumee River (Ohio) is an important tributary to Lake Erie. In an effort to reduce phosphorus loading to lakes and rivers, cities have upgraded sewage treatment plants, phosphorus concentrations in detergents have been reduced, and farmers have been asked to use fertilizers more carefully. In this lab, you will analyze time series data from the Maumee River, to determine whether these measures are succeeding in reducing phosphate loading to the Lake Erie. In the process, you will be exposed to several key elements of successful trend analysis: a) dealing with outliers, b) correcting for confounding factors, c) detecting and eliminating multicollinearity, and d) fitting to seasonal effects.

Open the JMP data file, "Maumee River PO4.jmp". You should see five columns: year (time in decimal years), month (1=January through 12=December), P load (total rate that the Maumee River delivers P to Lake Erie, in tons per day), P concentration (the concentration, in parts per billion, in Maumee River waters), and Q, the flow rate of the Maumee River (in cubic meters per second, sometimes called "cumecs" in hydrology-speak). Plot the distributions of P load, P concentration, and Q, and •1: briefly describe these distributions.

Looking for simple trends

Since we are looking for trends in P load, the obvious first step is to plot P load as a function of time. Do that now. •2: What is the trend (including, of course, its standard error)? Is the trend statistically significant? Does the trend line represent the data? Save the residuals and plot their distribution. •3: In what way does this distribution of residuals violate the conditions that are required for linear regression to yield accurate estimates of statistical significance?

Outliers?

Now a decision must be made: are the highest P load values *outliers*, or are they just the upper end of the real distribution of the data? A naive view is as follows: since the highest values are much higher than the others, they must be outliers, and should be deleted. Let's follow this naive view for a moment, and see where it takes us. First, plot the distribution of the P load data. •4: Sketch the shape of the distribution, and write down the mean and the standard error of the mean. Now, exclude the five highest "outliers". You can most easily do this by selecting these five points on your time series plot, and using control-E to exclude them. Now plot the distribution of the P loads again (note that N should now be 128 rather than 133). •5: What is the mean now, and what is the standard error? Has deleting the "outliers" changed the mean substantially? Has the *shape* of the distribution changed dramatically? Plot P load as a function of time once again. The five highest values should not appear anymore. Fit a line through the data. •6: By deleting the five "outliers", have we succeeded in meeting the conditions that are required for regression to accurately measure statistical significance?

Those who view these high P load values as outliers could potentially keep eliminating more and more of the upper end of the distribution, until it conformed with their preconceptions of what the data "ought" to look like (usually a normal distribution). In the process, one would eliminate the data that make up the vast majority of the total P delivered to Lake Erie over the period. In fact the high-P values are not outliers at all. They come from exactly the same distribution as the rest of the data; that distribution simply happens to be strongly skewed. If we decide not to call these points outliers (as indeed we should not), then we have only two alternatives: *use robust methods* (such as the Kendall-Theil slope for trend), or *transform the data*.

Transforming variables

Here, let's take the latter approach (if only because JMP lacks nonparametric regression). Since both P load and flow are right-skewed, we need to go *down* the ladder of powers. Bring the five deleted points back into the data set, by using the "clear row states" option in the "rows" menu. There are two columns in your data file called log_load and log_Q. Their formulas are:

Log_{10} (P load (Tons/day)) and Log_{10} (Q (m3/sec))

Look at the distributions of log_load and log_Q. •7: Are they much better behaved?

Now, plot log_load as a function of time, and fit a linear trend to the data. •8: What is the trend? How statistically significant is it? Does this trend represent the data better than your first plot did? Would you say that this plot provides strong evidence of a trend? Why or why not? Save the residuals and look at their distribution. •9: Did log-transforming P load eliminate the severe skew in the residuals?

Plot the residuals as a functions of time, month, and log_Q. •10: Do any of these variables appear to have important effects on log_load? Which appears to be most important? Why would you expect log_Q to be closely related to log_load?

Looking for trends in averages

When faced with a problem like the one shown here (fluctuations in flow obscure the time trend), many analysts will simply average the P load for each year, then look for trend in the yearly averages. There are several reasons why this is *generally not a good idea*:

1) you may throw away most of the information in your data set
2) if different years have different flows, the annual averages will (still) be distorted by flow effects
3) you will average out the effect of flow only if you have many points from each year and there is little year-to-year variation in flow. Even if these conditions are met, you will be smoothing over flow effects in a very inefficient way.

To drive this point home, plot the annually averaged P load as a function of time. These data are contained in the file, "Maumee river annual PO4.jmp". Fit a line to the data. •11: Is there a statistically significant trend? Close the file of annually averaged data; you won't be needing it again.

Correcting for flow effects

Using "Fit Model", fit log_load to both time and log_Q simultaneously. Examine the leverage plots, and again save the residuals (in the multiple regression platform, you click on the triangle next to "Response", pull down "Save Columns", and over-and-down to "Residuals"). •12: Write down the slopes (w/ standard errors) for year and log_Q. Is the time trend more precisely quantified now than it was when log_Q was left out? Why or why not? •13: Take the logs of both sides of the definition of load: load=concentration * flow. Is the dependence of log_load on log_Q close to what you would expect from this equation?

Now let's look again at the residuals. Log_load and log_Q looked pretty well correlated, but you should plot the residuals against log_Q to see if there are any other flow effects to "pick up". Note that there is some curvature; you can fit this plot with a quadratic of log_Q to see it. This suggests that properly accounting for flow effects will require fitting to both log_Q and log_Q^2. Create a new column, called log_Q^2, which contains the squares of log_Q. Now fit log_load as a function of time, log_Q, and log_Q^2, and •14: write down the slopes and standard errors for each of these three explanatory variables. Is the new slope coefficient for log_Q plausible? Why or why not? Do you have any ideas why it would change so drastically when log_Q^2 was added? *Note: you could automatically generate a quadratic fit to log_Q in the "Fit Model" platform, but please don't do that here (for reasons we'll disclose in a moment). Instead, do as you're told and create a new column with the squared values of log_Q, and use that in your regressions.*

Diagnosing Multicollinearity

Unstable coefficients, such as the slope coefficient for log_Q, are one indication of multicollinearity--that is, two or more X-variables may have nearly identical effects on the Y-variable, log_load. One way to quantify multicollinearity is through the Variance Inflation Factor, calculated as $VIF_j=1/(1-R_j^2)$, where R_j^2 is the R^2 for the multiple regression of the j^{th} X-variable against all the other X-variables. $VIF_j>10$ for any of the X-variables is taken as an indication of severe multicollinearity. Regress log_Q^2 against log_Q and year. •15: What is the VIF for log_Q^2? Can you see, from the leverage plots, where the multicollinearity comes from?

Correcting Multicollinearity

This kind of multicollinearity, which arises because one X-variable is nearly a linear function of another, can often be cured by *centering the data*--subtracting the mean, so that the mean of the transformed data is zero. So center log_Q, by creating two more columns, called "log_Q.centered" and "log_Q.centered^2", with formulas as follows:

log_Q - Col Mean(log_Q) and log_Q.centered^2

where the Col Mean function (*not the Mean function*) that can be found under the "statistical" option in the function window. Now plot log_Q.centered^2 against log_Q.centered, and note that there is no longer a close linear relationship between them. Just to verify that the multicollinearity has been eliminated, regress log_Q.centered^2 against log_Q.centered and year, and •16: calculate the VIF for log_Q.centered^2.

Now that the collinearity in the X-variables has been eliminated, regress log_load against year, log_Q.centered, and log_Qcentered^2. •17: Note the values of the slopes. Do they make sense?

Note: as it happens, whenever JMP creates a polynomial fit using the "Polynomial to Degree" macro in the "Fit Model" platform, JMP automatically centers the data by default -- precisely to avoid the multicollinearity that so often arises in polynomial fits. By creating a separate column with the squared log_Q values, we've circumvented this feature of JMP. We've done this in order to illustrate the perils of this type of multicollinearity, and to demonstrate how centering the data can cure this particular problem.

Interpreting the magnitude of the trend

The coefficient for "year" quantifies the long-term trend in log_load. Note that a linear trend in log_load corresponds to a fixed *percentage* change in load per year. If the coefficient has a value of "x", then for small x the percentage change in load is approximately $(10^X-1)*100$ percent per year. •18: what is the long term trend in phosphorus load, in percent per year?

Analysis of covariance

Another way to look for the same signal (long-term trend in phosphorus load, corrected for changes in flow) is through analysis of covariance (ANCOVA). ANCOVA is a generalized form of ANOVA that embraces category variables (different species, site locations, etc.) and continuous variables (time, flow, concentration) simultaneously. ANCOVA includes a spectrum of techniques too numerous to cover exhaustively in this course. In this lab, we'll try to give you a taste of two approaches that are computationally simple, intuitively appealing, and commonly used.

Approach 1: fits to separate periods
The first approach is based on the following idea: if there is a long-term trend in phosphorus load, apart from the effect of changes in flow, then if we plot phosphorus load against flow, we should see a change in the relationship between flow and load over time. Let's look for that effect here. Specifically, let's ask, "does the relationship between phosphorus load and flow change from the first half of the record to the second, and if so, how?"

First create a "dummy" variable that distinguishes the first half of the record from the second. To do this, make a column called "period" that has a value of "0" through the end of the sixth year, and "1" thereafter. You can do this either by copying and pasting, or by creating a formula based on the "conditional" (that is, if-then) function in the formula window.

You should also select the rows in the second period, and change their markers (use "markers" in the "rows" menu) to a different symbol so you can visually distinguish the two periods.

Now plot log_load as a function of log_Q. You could, of course, fit a line through this plot, but what we want to do is to fit *two* lines, one for the early data and another for the late period. To do this, click on the "fitting" triangle, and go all the way to the bottom of the pull-down menu, and select "grouping variable". Then select "period" as the grouping variable. Now any fits to the data will be calculated twice, once for each value of "period".

Now, from the fitting menu, select "polynomial" and "quadratic" (since we already know that there's a quadratic relationship between log_load and log_Q). •19: Write down the coefficients and standard errors for the intercept, the slope (log_Q), and the curvature (log_Q^2), for both periods.

There are three different comparisons we can make between the two periods. First, are the *elevations* of the two lines different? In other words, for equivalent values of X, would they yield equivalent values of Y? Second, are the *slopes* of the two lines different? Is one line steeper than the other? Third, are the *curvatures* of the two lines different? Is one more tightly "cupped" than the other? The first question asks whether there was an overall increase in flow, all else equal. The second and third questions ask whether load became more sensitive to flow, and if so, in what way.

Don't conduct a formal statistical test here, but just eyeball the numbers. The difference between the intercepts, for example, would be statistically significant if it were roughly twice its standard error (that is, t>2). Does it look like it's going to be? If you can't eyeball it, then calculate the difference between the intercepts, then divide by the standard error of the difference (the square root of the sum of the squares of the two standard errors, remember?), to get an approximate t-statistic. •20: Is the difference between the intercepts statistically significant? How about the slope parameters or the curvature parameters?

You now have a real puzzle. Your earlier analysis showed that there was a statistically significant trend over time, but now it appears that there's no statistically detectable difference between the curves for the early and late periods. How can one reconcile these contradictory observations? It's doubly puzzling because there seems to be a definite displacement of one curve downward from the other, although the difference is small.

Look again at the plot. •21: Where are the intercepts for the two curve fits? Are they anywhere near your data? What does that imply about the uncertainty in the intercept?

Notice that not only are the intercepts highly uncertain, but the slope parameters (log_Q coefficients) are all wrong again. The solution to both problems is the same as before: *center the data*. This fixes two problems. It moves the y-intercept to the middle of your data (where, remember, the uncertainties in a fitted line will be smallest), and it decreases the multicollinearity in the data (which still exists, even in this 2-D plot, because we're plotting a curve rather than a straight line). When the x-variable is not centered, the uncertainty in the intercept is magnified by the uncertainty in the slope and the curvature (since the intercept is extrapolated beyond the data). Similarly, the uncertainty in the slope is magnified by the uncertainty in the curvature. Centering solves all these problems at once.

So now, plot log_load as a function of log_Q.centered, and re-fit the two quadratics. •22: write down the coefficients for both curves, and calculate an estimate of t=(intercept$_1$-intercept$_0$)/sqrt(s.e.(intercept$_1$)^2+s.e.(intercept$_0$)^2). Given that the number of degrees of freedom is large, does this indicate a statistically significant difference between the two intercepts? Note that these intercepts now measure the elevation of the curves near the center of the data, which is what you wanted all along, rather than values of elevation that have been extrapolated beyond the edge of the data. •23: Did centering the data affect the uncertainties in any of the parameters? Which ones? Does there appear to be a statistically significant difference between periods, for any parameters except the intercept?

Approach 2: multiple regression with dummy variables
The results you've just obtained indicate that the effects of flow don't measurably change with time. If that is the case, then you can more precisely determine the difference between the early and late periods if you keep everything except the intercept the same between the

two periods. In other words, there should be some way to keep the same parabolic curve relating log_load to log_Q.centered, and just move it up or down between the two periods. There is, indeed, a way to do this. You simply regress log_load against log_Q.centered and your dummy variable, "period", as follows:

log_load = a + b_1 log_Q.centered + b_2 log_Q.centered^2 + b_3 period

Note that the only difference between period=0 and period=1 is that the curves are shifted by an amount b_3. This has several useful properties. First, since the modeled flow effects remain the same during the two periods (b_1 and b_2 are estimated for the whole data set together, not each half separately), b_1 and b_2 will be determined more precisely. Second, since b_1 and b_2 are better constrained, and held constant between the two periods, b_3 will more precisely express the difference between the two periods. Third, b_3, its standard error, and its statistical significance directly express the vertical offset between the two periods. So, •24: fit the data with the equation above, and note the values and standard errors of the fitted parameters. Is b_3 comparable to the difference between the two intercepts, as you calculated in the previous step? Are b_1 and b_2 comparable to their counterparts that you estimated in the previous step?

There is one very important assumption in what you've just done above, namely that the effects of flow (expressed by b_1 and b_2) do not change between the two periods. You have good reason to believe that this is true, given the results of your previous analyses. There is also a way to check this using, again, dummy variables. For example, let's say you wanted to check whether there was a change in the log_Q slope between the two periods (that is, whether b_1 really applies to both periods, or whether there should be two different coefficients for the two periods). You can regress log_load as follows:

log_load = a + b_1 log_Q.centered + b_2 log_Q.centered^2 + b_3 period + b_4 period*log_Q.centered

where the term period*log_Q.centered is termed the "interaction" between period and log_Q.centered. It is literally the variable "period" multiplied by the variable "log_Q.centered"; to see its effect, note that the equation above can be rewritten thus:

log_load = a + (b_1+b_4 period) log_Q.centered + b_2 log_Q.centered^2 + b_3 period

In other words, the coefficient for log_Q.centered is b_1 for the first period (period=0), and b_1+b_4 for the second period (period=1). If there is no difference in the effect of log_Q.centered between the two periods, b_4 will be insignificantly small. If, on the other hand, there is a difference in slope as well as intercept between the two periods, b_4 will quantify the difference in slope. So, let's see what this difference is. Create yet another column (last time, we promise!) called "period*log_Q.centered", that contains "log_Q.centered" multiplied by "period". Then regress log_load on log_Q.centered, log_Q.centered^2, period, and period*log_Q.centered. •25: Does the fitted parameter for the interaction term suggest that there is a significant change in slope?

NB: If a regression equation includes the interaction between two variables, it normally must also include the two individual variables as well. That is, you can easily interpret the results of $Y=a+b_1X+b_2T+b_3X*T$, but usually $Y=a+b_1X+b_3X*T$ will not give meaningful results.

We hope that you have learned several important lessons from this lab, which we can summarize thus:

1. In this example, as in many environmental data analyses, relatively small time trend signals are utterly swamped by other variables (like flow), but they are nonetheless identifiable, measurable, and often statistically signficant.

2. Do not treat data as outliers just because they clash with your assumptions about what the data should look like.

3. If the trend of interest is obscured by the effects of other variables, averaging the data over longer time periods will be at best inefficient, and at worst ineffective, in eliminating the effects of those variables. It is usually preferable to try to fit the data to the trend of interest and the confounding variables simultaneously.

3. Beware of multicollinearity. The Variance Inflation Factor is a useful diagnostic; your intuition can also help, as can thoroughly exploring the correlations among the various X-variables.

4. When in doubt, center your data.

5. One good way to display a small change (like a time trend) that is easily masked by a big change (like flow effects) is by plotting the variable of interest against the confounding variable (e.g., flow) and then fitting separate curves for different values of the causal factor of interest (e.g., time periods).

6. ANCOVA techniques offer a flexible, powerful approach to directly measuring the size (and statistical significance) of changes between periods. Remember to consider the possibility of interaction effects.