Environmental Data Laboratory
Professor Kirchner

Laboratory 4: Uncertainty propagation for correlated variables, plus more exploratory data analysis

Part A: Uncertainty propagation with correlated variables: Monocacy River rainfall/runoff data

Open the "Monocacy rainfall/runoff.jmp" data file. This file contains data for 24 storms that hit the watershed of the Monocacy River in Maryland (those who know their early American history will recognize the name). The two pieces of data are the total rainfall in each storm (measured in units of depth), and the total runoff from each storm (that is, the sum of runoff from the time the storm began until the stream returned to base flow after the storm ended). The runoff data have been converted to units of depth (by dividing by the watershed area), so that runoff can be directly compared to the rainfall figures. You can visualize this process as collecting all the runoff in a huge tank, then spreading it like peanut butter in a uniform layer over the watershed. Rainfall/runoff relationships can be interpreted at various levels of hydrological sophistication. For this example, it is sufficient for you to know that water falling as rain on the watershed can do at least three different things: it can evaporate, it can run off relatively quickly (as storm runoff), or it can recharge long-term groundwater storage, leaving the watershed as base flow over a long period:

Evaporation&recharge = precipitation - runoff

Assume that, just like in the last lab, you are trying to infer how much water evaporates or recharges groundwater, and you want to know how variable this quantity is from storm to storm. That is, what is the mean and standard deviation of evaporation and recharge per storm? One way to do this is by straightforward error propagation. Remember, you can propagate the uncertainty (or variability) in individual measurements (as expressed by the standard deviation) just like you can propagate the uncertainty in the mean (as expressed by the standard error of the mean). •1: Using the "Distribution" platform, determine the mean and standard deviation of rainfall and runoff for the 24 storms in the data set. •2: Using the simple rule for sums and differences, calculate the mean and standard deviation of evaporation&recharge (I've written it that way to remind you that evaporation&recharge refers to just one variable, not two separate variables--at least you can't separate the two with the available data). Now, check your results by creating a new column with a formula that calculates the evaporation&recharge for each storm. •3: Using the "distribution of y's" platform again, calculate the standard deviation of the individual evaporation&recharge values in the column you just created. •4: *Why* does this standard deviation differ from the one you predicted in step #2 above? What are the assumptions underlying the simple rule for sums and differences? Which, if any, of these assumptions might be violated by the data analyzed here? (Note that the simple sum/difference rule does *not* assume that the data are normally distributed).

The simple rule for sums and differences is based on the assumption that the input variables are not correlated. Are they? One way to check for correlations is to use the "fit y by x" platform to create a scatterplot that shows runoff as a function of rainfall. Usually your eyes can roughly gauge how much correlation is in a scatterplot. •5: Does it look like runoff is strongly correlated with rainfall? Is it positively or negatively correlated? Besides eyeballing the data scatter, another way to visualize the strength of the correlation is with a *density ellipse*, which you can select via the pull-down triangle. Click on the red pull-down diamond, then slide the mouse down to "density ellipses", then over to 0.95. The result is an ellipse that would enclose 95% of the data points if both x and y were normally distributed (which they aren't, here, but that's not the point for now). The narrower and skinnier this ellipse is, the greater the degree of correlation between x and y. As we mentioned in lecture, the way to quantify the degree of correlation is by the correlation coefficient, *r*. One way to calculate *r* is via the "correlation" box that comes along with the density ellipse. This shows the mean and standard deviation for both x and y, and the correlation between x and y. •6: What is the correlation coefficient between rainfall and runoff?

As the lecture (and the error propagation toolkit) explained, when variables are correlated, the "simple rules" no longer hold. •7: How, physically, does the fact that rainfall and runoff vary together affect the variability in evaporation&recharge? (This is not a question about the formulas and the mathematics, it's a question about how, in the real world, variations in rainfall and runoff may compound or cancel each other out. In other words, if you were trying to explain this to someone who didn't understand math, how would you explain it?)

When variables are correlated, the simple rules--and Gaussian error propagation, for that matter--are no longer accurate. Instead, you must use an error propagation formula that takes explicit account of the correlation. The "method of moments" does this; see the toolkit on error propagation for more complete details. The method of moments for a two-variable function, z=f(x,y) yields:

$$s_Z \approx \sqrt{\left(\frac{\partial z}{\partial x} s_x\right)^2 + \left(\frac{\partial z}{\partial y} s_y\right)^2 + 2r_{xy}\left(\frac{\partial z}{\partial x} s_x\right)\left(\frac{\partial z}{\partial y} s_y\right)}$$

Now, the situation at hand is the extraordinarily *simple* formula, z=x-y, where z is evaporation&recharge, x is rainfall, and y is runoff. •8: Rewrite the equation above, substituting in the values of the appropriate partial derivatives (and be careful to get the signs right!). •9: Substitute in the appropriate standard deviations and correlation coefficient, and calculate your new (and hopefully more accurate) estimate of the standard deviation of evaporation&recharge. •10: Does your new estimate agree with the actual standard deviation for the 24 storms? How much did the correlation between rainfall and runoff affect the variability in evaporation&recharge?

Part B: Exploratory data analysis--Obliquity of Earth's orbit

Scatterplots, like the plot of rainfall vs. runoff above, can be used to explore many different aspects of the relationships between variables. Take, for example, the data in the file, "Obliquity data.jmp" These are measurements of the Earth's orbit, going back to several centuries B.C. Now, the dictionary says that "obliquity" means "immorality, dishonesty, and mental perversity", but--sorry folks--here we're talking about a less racy kind of deviance, the *obliquity of the ecliptic*,which is the angle that the earth's axis of rotation is tilted, away from perpendicular to the plane of earth's orbit. The obliquity of the ecliptic is responsible for the fact that we have seasons, and it varies, ever so slightly, over time (which changes the seasonality of the earth's climate as well).

Using the "fit y by x" platform, plot the earth's measured obliquity as a function of time. Note that these measurements range from about 300 BC to 1738 AD. •11: Describe, in words, the overall trend that you see. Plot a 95% density ellipse over the data, as a reference point for what comes next.

Are all the data consistent with one another? Of course, that depends on what kind of relationship you *expect* to see. But if you expect the obliquity to change smoothly over time (the Earth is a pretty massive object after all, and it has a lot of angular momentum), then at least one point sticks out from the pattern. Click on that point, and it will light up with the name of the person who published the measurement. What if you exclude that one point from the data set? Here's the easy way to do it. Highlight the point by clicking on it. Next, under "rows", select the "marker" option, and change the marker for this point to an "x", so that you'll be able to differentiate it from the others. Finally, select the "exclude/include" option (under the "rows" menu, also abbreviated as command-E) to exclude this point from further calculations. Now, plot another 95% density ellipse over the data. •12: how has the correlation between obliquity and year changed as a result of deleting the one point? Why has it changed?

•13: Now, are all the data consistent with one another? (There are precise statistical ways of asking this question, but we won't be bothering with those today. Just eyeball it instead). Specifically, are the three ancient measurements consistent with the slope in the data from the year 800 onward?

Try excluding the three ancient data points, in addition to the one point you've already excluded (you can shift-click to select them all at once). Plot still another 95% density ellipse over the data. •14: How does this ellipse differ from the others?

Finally, use the "fit line" option under the pull-down triangle to fit a straight line to the remaining data (the procedure is called linear regression, about which a lot more will be said later in the course). •15: Does this line appear to be visually consistent with the modern (year 2000) obliquity value of about 23.45 degrees?

Note that these data present us with two very different possibilities. One is that the measurements by Pappus and the ancient Greeks were simply inaccurate, by a small fraction of a degree. Another is that their measurements are exact, and in fact the obliquity has changed rather irregularly over time. To see what kind of trend would be consistent with the data, first cancel your exclusion of the four early measurements. You can either select those four measurements and then toggle the exclude/include again, or a simpler way is to just select "clear row states" from the "rows" menu. Now try fitting a spline to the data (using "fit spline" under the pull-down triangle). A spline is a free-form smooth curve that tries to go through each data point. You can select varying degrees of smoothness, according to a smoothness parameter "lambda"; for these data, anything below about lambda=1000 gives essentially the same results. •16: Describe, in words, the difference between the general trend suggested by the density ellipses and the trend suggested by the spline fit.

Kind of makes you wish that folks like Eratosthenes and Pappus reported their error bars, doesn't it?

Part C: Exploratory data analysis--Iris flower measurements

Open the data set "Iris data.jmp". These are measurements of the lengths and widths of petals and sepals (sepals are the little leaves at the base of the flowers) from 50 individuals of each of three species of Irises: *Iris setosa, Iris versicolor,* and *Iris virginica*. One question that one might want to answer with these data would be, do flowers that have long petals also have long sepals? Using the "fit y by x" platform, plot petal length as a function of sepal length (that is, sepal length on the x axis, and petal length on the y axis). Note that the three different species are plotted in three different colors. Draw a 95% ellipse over the data, and also fit a linear regression line, using "fit line". •17: Briefly describe the relationship between sepal length and petal length that you see in the data.

Now, the variation in a property like sepal length or petal length combines two different kinds of variability. One is the variability *within* each individual species, and the other is the variability *among* different species. This scatterplot contains some information about both kinds of variability. Let's look at the relationship between petal length and sepal length in each individual species. One way to do this would be to divide the data into three separate sets, and plot each individually. But there's an easier--and ultimately more illuminating--way to look at this. Go to the bottom of the "fitting" triangle's pop-up menu, and select "Group By...". Then select "species" as your grouping variable. This tells JMP IN to perform any analyses on each species separately, rather than all three together. Now, draw 95% density ellipses for each species, and fit a linear regression line for each species. •18: Briefly describe these ellipses. Specifically, compare the *orientation* of the ellipses (which express the variation within each species) with the *relative position* of the ellipses (which expresses the differences between the species). Note that sepal length varies with petal length, both because within each species, different individual flowers are larger or smaller, and also because different species will have flowers of slightly different sizes and shapes. But notice that the overall pattern of change from species to species is quite different from the pattern of variation within each individual species.

Part D: Exploratory data analysis--Sunspot numbers

Since the mid-18th century, astronomers have been counting the number of spots that are visible on the solar disk. Wölfer collected these sunspot records, and adjusted them for changes in telescope quality (and observer technique) over time, to put the different records for different observers on a common basis. The results of his efforts are found in the file, "Sunspot numbers.jmp". Open that file now, and plot sunspot numbers as a function of year.

Do you see any pattern here? Looks pretty noisy, doesn't it? One way to try to see a pattern is by fitting a curve to guide your eye. Try fitting a very flexible spline, one with lambda=0.01, to the sunspot data. Now do you see the 11-year sunspot cycle? The data will also become clearer if you stretch the plot out horizontally, so do that too.

Now, you can look for a patterns in data on many different scales, and what pattern you see will depend on the scale at which you look. To see this, try fitting successively stiffer splines to the sunspot data, with lambda values of 100, 10000, and 1000000 (the stiffest in the menu). The stiffer splines smooth over more of the small-scale variation, but reveal larger-scale patterns that might be hidden behind the noisy short-term cycles. It may be easier to compare these patterns if you hide the individual data points; you can do this by de-selecting "show points" from under the "fitting" triangle.
•19: Briefly, what are the different "messages" that these different splines convey? Do you think there is any single "correct" way to look at these data?