Environmental Data Laboratory
Professor Kirchner

Laboratory 12: Sunspots and melanoma: an environmental detective story

Melanoma is a form of cancer that occurs in the skin. Risk factors for melanoma include a history of severe sunburns and living at high altitude, raising the possibility that exposure to the sun may cause some melanomas. One way to test this theory is by seeing whether the temporal pattern of melanoma incidence matches the temporal pattern of the well-known 11-year sunspot cycle. In this lab, you will try to determine whether or not the sunspot cycle plays a role in melanoma incidence, by comparing sunspot records with data from the Connecticut tumor registry. These data are in the JMP file, "Sunspots/melanoma.jmp". Melanoma incidence is recorded as the number of newly diagnosed cases per year, per hundred thousand people. The number of diagnosed cases could change, either if the actual rate of melanoma changes, or if doctors become more aware of melanoma and thus diagnose it more often. Sunspot abundance is recorded as a relative scale, indicating the average number of dark spots on the sun's surface.

In the course of this guided exploration of the sunspot/melanoma question, you should build a word processing file (e.g., a microsoft word document) with relevant plots, as well as responses to queries indicated by the "•" symbol.

•1 The most obvious way to look for a relationship between sunspots and melanoma is to plot melanoma incidence as a function of the sunspot index. Construct this plot, using "fit Y by X". Is there any apparent relationship between sunspot index and melanoma rate?

Melanoma may take some time to develop to the point that it can be diagnosed. If so, then melanoma diagnoses may lag behind the sunspot index. So the next obvious step to take is to compare melanoma diagnoses to the lagged sunspot index (that is, to what the sunspot index was the previous year, or two years ago, or three years ago, and so forth). The way to construct a lagged variable is by creating a new column (called, for example, "sunspots lag 1") with the following formula:

Lag(sunspot index , 1)

Obviously, by changing the lag factor from 1 to 2 and so forth, you can create lags of 2 years, 3 years, and so on. Let's assume that the longest lag that is likely to be detectable is roughly five years. Construct five new columns for lagged sunspot indexes, with lags of 1 to 5 years.

•2 Now plot melanoma incidence against each of these lagged sunspot indexes. You can use the "fit Y by X" option, and select all of the sunspot indexes; you'll get one x-y plot for each index. Is there any apparent relationship between melanoma rate and any of the lagged sunspot indexes?

Upon seeing these results, the unsophisticated data analyst might be tempted to give up, go home, and sulk. But be of stout heart and good cheer. Two of the enduring lessons (we hope) of this course are that correlation does not necessarily imply causation, and conversely that lack of correlation does not necessarily imply lack of causation. A plot of X against Y is affected not only by X and Y, but also by every other variable that has a causal connection, or incidental correlation, with X and Y. These "confounding" variables can obscure the underlying relationship between X and Y, or they can give rise to an "artifactual" correlation between X and Y when no causal relationship exists. Rather than going home to sulk, let's try a more systematic, deliberate kind of exploratory data analysis.

Step 1: Plot each variable as a function of every other variable that it could plausibly depend on. Our data set has three variables: year, melanoma incidence, and sunspot index (with lags). Melanoma incidence (the "variable of interest") could conceivably depend on both sunspot incidence and year, whereas sunspot index varies as a function of time, but not as a function of melanoma incidence. You have already looked at melanoma incidence as a function of sunspots, with and without lags.

•3 Now plot melanoma incidence and sunspot index as a function of time. You may want to use linear and/or spline fits to reveal the patterns in the data. What patterns are thus revealed, and what do they suggest about a possible connection between sunspots and melanoma incidence?

Step 2: Select the variable that has the strongest direct influence on the variable of interest. Fit the relationship between that variable and the variable of interest, using an appropriate function. Choosing an "appropriate" function is a matter of judgment; you got some practice doing this in Labs 6 and 7. You want to capture the major trends in the data, without chasing the random noise around.

•4 What is the fitted relationship you came up with? How much of the variance in melanoma incidence does it account for? (This is the ratio between the model sum of squares and the total sum of squares... it is also r-squared.) Can you think of one or more interpretations that would explain how that relationship could come about?

You want to be careful not to use a function that pre-empts the effect of other variables. For example, if you expressed melanoma incidence as the sum of a linear trend and a sinusoidal oscillation, you would pre-empt an explanatory role for the (sinusoidally varying) sunspot index.

•5 Inspect the residuals as a function of time. Do you see a temporal pattern in the residuals that is comparable to the temporal pattern in the sunspot index? (Spline curves may help you visualize the data here). Note that if you have fitted melanoma incidence as a linear function of time, you should not expect to see any linear trend in the residuals; you've already taken out all the linear information in the data. The residuals will only contain higher-order information and/or random noise.)

Step 3: Plot the residuals from the previous step, as a function of all the other variables that might be causally related to the variable of interest. Select the variable that has the strongest direct influence on the residuals. To calculate the residuals, use the "save residuals" option with the functional fit that you derived in step 2.

•6 Do the residuals vary significantly as a function of sunspot index, or any of its lags? Which ones? Can you explain why the result you get here might differ from the result you got in (a) or (b) above? Can you think of a plausible interpretation for the relationship you see? How much of the variance in the residuals can be explained by the most influential remaining variable? Keeping in mind that the variance in the residuals is the "leftover" variance that could not be accounted for in (d), roughly what fraction of the total variance could you account for using both the variable selected in (d) and the variable selected here?

Step 4: Fit the variable of interest, as a function of both variables chosen in steps 2 and 3, using an appropriate function (or combination of functions). In most situations, the simplest way to fit two or more variables is through multiple linear regression (using the "Fit Model" option in the "Analyses" menu). This fits melanoma (for example) as a linear function of time and (appropriately lagged) sunspot index. You want to fit the variables jointly, rather than separately (as in the x-y plots you've seen so far) to properly capture any interdependence of the x-variables on each other. (*Caution:* when you use the "Fit Model" platform, be sure to select only one Y variable and one or more X's. If you inadvertently select two or more Y variables, you will see a new window called "MANOVA specification". That indicates that you are inadvertently trying to do a Mulitvariate ANalysis Of VAriance, which is well beyond the scope of what you've learned in this course.)

•7 Inspect the results from step 4 (if you're confused by JMP's multiple regression report windows, refer to the JMP manuals or to JMP's help windows, using the "?" tool). Are there any obvious discrepancies between the data and the model you've created? How much of the variance can you now account for? What does your model suggest about the average "latency" period that elapses before melanoma is diagnosed?

Step 5: Repeat steps 3 and 4, iterating until there is no more systematic variation that can be accounted for. In the multiple regression report window, you save the residuals by clicking on the pull-down triangle next to "Response", then pulling down to "Save Columns", then over and down to "Residuals".

•8 Plot the residuals from the last step as a function of the remaining variables. Can any of these explain a significant fraction of the variance of the residuals (keeping in mind that the residual variance is now a very small fraction of the total variance)?

Take a moment to pat yourself on the back. When you started, back in (a) and (b), you were faced with about a ten-fold variation in melanoma incidence, without any apparent rhyme or reason. Now you've extracted a lot of information from the data, and you've reduced the unexplained variation to a small fraction of what it once was. But now, if the relationship between sunspots and melanoma is so clear in (g), why wasn't it so clear in (a) or (b)?

•9 To see how this could happen, do a 3-D plot (using "Graph" >> "Spinning plot") of the three variables you analyzed in (g). Then use the "hand" tool, or the axis rotation buttons, to rotate this plot until it shows lagged sunspots on the horizontal axis, and melanoma on the vertical, with the year axis pointing straight toward you. Note the large scatter. Now rotate the plot upward or downward (keeping the lagged sunspots on the horizontal, and the melanoma axis vertical), until you see a clear relationship between the two. What you've done is to change your viewing angle in such a way that you cancel the effects of variations in year. Now use the hand tool to roll the plot around, viewing it from many different angles. Note the three-dimensional shape of the data, and how your viewing angle affects the way that it projects into two dimensions. Can you explain why a the effect of sunspots on melanoma is so clear in (g), but so unclear in (a) and (b)?