Environmental Data Laboratory
Professor Kirchner

Laboratory 9: Exploratory analysis, ANOVA and linear regression in environmental time series

The California Air Resources Board (or ARB) operates a network of air quality monitoring stations that continuously measure airborne metals, volatile organic compounds, and polycyclic aromatic hydrocarbons. Concentrations are integrated over periods of roughly 10 days, yielding approximately 30 samples per site per year. Over time, some sites have been added to the network and others have been discontinued, but many have been in continuous operation since 1985. The entire data set from this monitoring effort is huge, comprising thousands of samples, each with measurements of dozens of compounds.

In this lab, you will use a small portion of this massive data set to examine how airborne lead concentrations have changed in response to the removal of leaded gasoline from the market. Responding to the documented health risks of long-term lead exposure, even at low levels, the EPA mandated that leaded gasoline be phased out. As late as the mid-1980's, leaded gasoline accounted for roughly 95% of total lead emissions to the environment in the U.S. (a staggering 35 million kilograms per year nationwide). The introduction of catalytic converters greatly reduced the use of leaded gasoline, but it was still used in older cars until it was completely phased out.

The data you will be examining are in the file, "ARB metals data.jmp". This contains monthly average airborne concentrations of several toxic metals (arsenic, beryllium, cadmium, chromium, manganese, nickel, and lead) measured at 15 ARB monitoring stations. The ARB network contains roughly twice this many sites, but for this data set I have chosen the 15 stations with the longest data series. Here are the different columns and what they mean:

Site -The city in which the monitoring station is located

Region -I have grouped the 15 sites into five regions, and given each region a color code:
Bay Area (San Francisco, San Jose, Richmond, Fremont, and Concord): **dark blue**
LA Basin (Los Angeles, Long Beach, Riverside, and Upland): **hot pink**
Valley (the Central Valley--Bakersfield and Stockton): **green**
San Diego (El Cajon and Chula Vista): **orange**
SB/Vly (Santa Barbara and Simi Valley): **yellow**

Year Decimal year corresponding to the month in which the samples were taken

yr Integer year (note that "year" and "yr" differ: "year" takes on different values in each year and month, whereas "yr" only has one value for each year)

month Integer for month

N samples Number of samples averaged for that month at that station

As monthly average arsenic concentration, in ng/m^3
Be monthly average beryllium concentration, in ng/m^3
Cd monthly average cadmium concentration, in ng/m^3
Cr monthly average chromium concentration, in ng/m^3
Mn monthly average manganese concentration, in ng/m^3
Ni monthly average nickel concentration, in ng/m^3
Pb monthly average lead concentration, in ng/m^3

log(As), etc. log of monthly average arsenic concentration (etc.)

Because the data set is so large, most of the plots you generate will need to be expanded (by dragging on the lower right corner) to make their details visible.

Part A: Long-term trend

First, plot Pb concentration as a function of Year. •1: Briefly describe the pattern that you see, including a) what is shape of the long-term trend? b) is the degree of variability constant or decreasing over time? c) is there any evidence of a seasonal cycle, or other pattern within each year?, and d) do you see any obvious differences between the different regions?

Since we want to find the trend over time, a simple thing to do would be to fit a linear regression line to the Pb concentration data. Do that now, using "fit line" under the "Bivariate fit" triangle. •2: Is the regression slope (the parameter estimate for "year") well constrained by the data? Is the regression statistically significant? If you saw only the *statistics* of the regression fit, would you think it was adequate? •3: Does the linear regression line look like an adequate representation of the data? Why or why not?

Note that you should *never* judge the adequacy of a regression fit from the numerical statistics alone, as this example aptly demonstrates. It makes little sense to fit a straight line through a data set that is clearly not linear, as is the case here. The fact that the regression line predicts *negative* airborne concentrations after the spring of 1995 should be a dead giveaway (note that this is not even a distant extrapolation; it's within the period of record!).

Although it's hardly necessary in this particularly egregious example, the use of *residuals* can be particularly effective in diagnosing problems in regression models. The residuals are the deviations of the individual data points from the fitted line. If the fitted line captures the structure of the data, then the residuals will contain pure, patternless statistical noise. If there is no intelligible signal left in the residuals, then the regression model has succeeded in extracting all of the signal from the data. The residue that's left--the residuals, that is--should be nothing but noise.

Save the residuals using the "save residuals" option under the "linear fit" triangle. Then plot the distribution of the residuals, and plot the residuals as a function of year. Linear regression assumes that the underlying relationship is in fact linear, that the residuals are roughly normally distributed, and that the residuals are *homoscedastic*, that is, their variance is constant from one end of the data set to the other. •4: Does it look like any of these assumptions are met? Why or why not?

The next approach that most people would try is to replace the linear fit with a quadratic, so go ahead and try that. •5: Does the fit look adequate? Why or why not? In particular, does the quadratic fit behave reasonably from 1994 onward? Save the residuals from this fit, and plot them. •6: Do they look reasonable? Why or why not?

Note that the data are still severely heteroscedastic. That is, their scatter is not constant; instead, high concentrations also have high scatter. In general, this problem cannot be solved by fitting with a more complicated function, such as a polynomial curve; it can only be solved by transforming the y-axis variable. That's because the problem comes from the Pb concentrations themselves (higher concentrations have more scatter), not from the relationship between concentration and year.

When higher concentrations have more scatter (which is typical when data can only take on positive values but have means close to zero), you need to transform *down* the ladder of powers. The log transform is appropriate when the scatter in the concentrations is not constant in absolute terms (that is, the same number of ng/m^3), but instead is a constant percentage of the average concentration. Try plotting the log of the Pb concentration as a function of year, and •7: again describe the pattern that you see, including a) what is shape of the long-term trend? b) is the degree of variability roughly constant, or does it change over time? c) is there any evidence of a seasonal cycle, or other pattern within each year? and d) do you see any obvious differences between the different regions?

Fit a straight line to the log(Pb) trend, and again plot the residuals. You may need to enlarge one or both plots (by dragging on the lower right corner) so that you can see all the data clearly. •8: Does the line look like it generally fits the data? Do the residuals reveal any further information that could be extracted from the data set?

Note that a linear trend in log(Pb) corresponds to a constant percentage rate of change in Pb concentrations; if the slope of the log(Pb) trend is *m*, then the percentage change per year is $100*(10^{m}-1)$. •9: What percentage rate of change is suggested from your data?

Another way to check for non-linear tendencies in the data, such as curvature, steps, angles, or seasonal cycles, is by fitting a flexible curve such as a smoothing spline. Try fitting splines of lambda=10 and lambda=0.01 to the data. Clear inconsistencies between the splines and the linear fit would indicate that there might be still more trend information to be extracted from the data. Similarly, try fitting the same two splines to the residuals from the linear fit. In particular, note the irregular seasonal cycle revealed by the flexible spline.

Now you know something about the overall trend in the data. But do all the sites show similar trends, or are Pb concentrations declining more rapidly at some sites, and more slowly--or not at all--in others? One way to get some insight into this is to plot separate regression lines for each of the sites. JMP provides a quick way to do this. First, plot log(Pb) as a function of year (you ought to make a fresh plot, or things will get too messy). Next, under the fitting triangle, select "Group By..." and select "site". Then select "fit line". Voila, you have all of the regression lines, superimposed on one another. •10: Do the different sites all indicate the same general trend? Is the difference between sites generally bigger, or generally smaller, than the change over time at each site?

Again starting from a clean plot, repeat the procedure in the paragraph above, except fit a flexible spline for each individual site. Note that the differences between sites are small compared to both the long-term changes at each site, and the seasonal variation within most sites. This means that the changes over time will obscure the differences among the sites, making them more difficult to detect.

Part B: Differences among sites

In a data set such as this, it is natural to look for similarities and differences among sites. One common way to do this is using ANOVA and Tukey HSD test. Do that here. First plot log(Pb) by site (you may want to jitter the display to reveal overlapping points), then add quantile boxes so that you can see the relationships among sites. Finally, compare all pairs. •11: Which site has the highest airborne lead concentrations? Which site has the lowest? What is the difference (in log units) between highest and lowest? How many times as high is the absolute concentration at the highest site than at the lowest?

Note that Santa Barbara looks like the "cleanest" site in this data set. But note that concentrations have been changing over time, so if some sites have different periods of record, this could generate artificial site-to-site differences in the data, which don't reflect the real site-to-site differences on the ground. Measurements began in Santa Barbara in 1988, versus 1985 or 1986 for all the other sites. •12: How would the difference in the period of record affect Santa Barbara's average Pb concentration in the data set? (Keep this plot up; you'll want it for comparison purposes in a minute.)

We want to know the site-to-site differences all else equal, that is, we want something that measures the average of the differences at any given point in time. To put it differently, we want to focus on the differences among sites, without the confounding effects of the long-term trend or seasonal cycles. As you saw in part A, the differences between sites are small compared to the long-term trend and seasonal variation; these are confounding "noise" if the signal you seek is the differences among sites.

If there were just two sites, you might set up a paired-sample t-test, measuring the difference between two sites in each month and year. But you have more than two sites, and some sites are missing some months of some years. So here's another approach that should work nearly as well. What you want is a way to subtract the overall long-term trend and seasonal cycle that are common to all the sites, to better reveal the differences among sites. In other words, you want to look for site-to-site differences in the residue (or more properly the *residuals*) of a curve that is fit to the long-term trend and seasonal variation in the data set as a whole.

To do this, generate yet another plot of log(Pb) as a function of year. Then fit a flexible spline that captures the seasonal variation as well as the year-to-year trend. Then save the residuals from this spline fit. Finally, plot these residuals as a function of site. Add quantile boxes. •13: Now, compare the visual impression of this ANOVA plot with the one you generated a moment ago, before you removed the average long-term trend. Are the site-to-site differences bigger, smaller, or about the same (note that the scales of the two figures may differ a bit)? Is the variability within each site (and the uncertainty in each site's mean) bigger, smaller, or about the same? Why? Note that because this is a plot of residuals rather than concentrations per se, the values on the y-axis have relative meaning but not absolute meaning; they can be used to quantify differences between sites, but do not directly tell you the absolute concentration at any site.

Now, again compare all pairs. •14: Is Santa Barbara still the "cleanest" site? Did removing the average long-term trend, and the average seasonal variation, clarify the differences among the sites? Why or why not?

<u>Part C: Seasonal variation in airborne lead concentrations</u>

(This part is optional. Do it only if time permits.)

There are three clear patterns in the data: a long-term trend, site-to-site differences, and an irregular seasonal pattern. So far, you have looked at the first two patterns, partly by filtering out seasonal variation. Can you use the methods outlined above to filter out long-term trends and site-to-site differences, in order to clarify the month-to-month seasonal cycle? One way is by using a fitted line (or a smooth spline) to capture the long-term trend in the data; the residuals from such a line will no longer be dominated by the long-term trend, but will better reflect site-to-site differences and seasonal variations. Another approach--probably a better one--is to calculate the annual average log(Pb) at each site for each year, then subtract this annual average from each of the monthly values for that site and year. In this way you directly filter out the effects of site-to-site and year-to-year differences, leaving only the month-to-month differences (and, of course, a healthy dose of random noise). You can do this in JMP by calculating the annual averages using the group/summary command, then using the "join" command to pair each annual average for each site with all the months that it pertains to, then finally writing a formula to subtract the annual mean from each month's concentration. Try this--or some other clever idea of your own creation--and briefly summarize how well it seems to work.

<u>Here are some of the impressions we hope to leave you with:</u>

1. Descriptive statistics are rarely effective diagnostics for regression fits.

2. Residuals, on the other hand, are particularly useful in revealing problems with fitted lines or curves.

3. Residuals can also be useful in filtering out one kind of variability (e.g., long-term trends) in order to highlight another kind of variability (e.g., site-to-site differences).

4. The variable of interest usually varies in response to several different factors. Variation in one factor (such as long-term trends) can obscure the others (such as site-to-site and month-to-month differences), unless you correct for it.

5. There's a gawdawful lot of interesting data out there, and it has a story to tell; you just need to know how to coax the story out.