1. Sources of serial correlation

Serial correlation (also called *autocorrelation*) occurs when residuals from adjacent measurements in a time series are not independent of one another (that is, if the i^{th} residual is high, it is likely that the $i+1^{st}$ residual will also be high, and likewise low residuals tend to follow other low residuals). Serial correlation is very common in environmental time series. It can arise in at least four different ways:

1.1. The variable of interest has a long-term trend--which may be either linear or nonlinear through time--that has been overlooked in the analysis.

1.2. The variable of interest varies seasonally, and this seasonal effect has been overlooked in the analysis.

1.3. The variable under study responds to one or more missing explanatory variables (i.e., variables not included in the analysis) that are serially correlated (correlated with themselves through time, e.g., due to long-term trends or seasonal variations).

1.4. The variable of interest includes random noise that is serially correlated, or that has persistent effects. This commonly arises when your variable integrates or averages some other variable through time. For example, populations are affected by random changes in birth and death rates, but note that a random extra birth has a persistent effect; it inflates the population tally for as long as the organism's lifespan. Likewise, changes in chemical inputs to a lake will persist for timescales comparable to the lake's flushing time; the same applies to the oceans or the atmosphere. Whenever the sampling period is shorter than the timescales on which the system responds to external forcing, many samples will be effectively redundant.

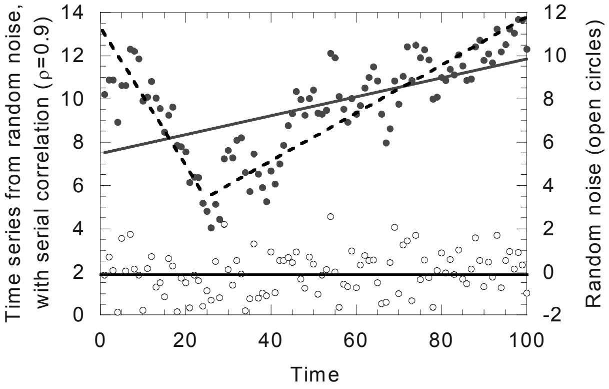

Figure 1. Do the solid dots show a shift from a declining trend to a climbing trend (dashed lines), or one irregular climbing trend? In fact they show *neither one*; they show *no trend* (Y=constant), to which serially correlated residuals ($e_i=\rho e_{i-1}+\xi_i$) have been added with $\rho=0.9$ and pure random noise ξ as shown by open circles (right scale, no exaggeration).

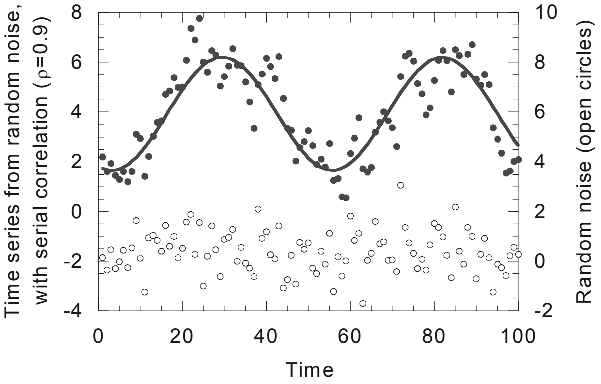

Figure 2. Serial correlation can produce nonlinear, visually coherent patterns. Actual trend behind the apparent sine wave shown above is Y=constant, as in Figure 1. Residuals generated from serially correlated random noise, as in Figure 1. Random noise ξ shown by open circles.

2. Consequences of serial correlation

Serial correlation is a significant problem because nearly all statistical techniques assume that any random errors are independent. Instead, when serial correlation is present, each error (or residual) depends on the previous residual. One way to think about this problem is that because the residuals are largely redundant (i.e., not independent of one another), the effective degrees of freedom are far fewer than the number of observations.

If serial correlation is present but is not accounted for, several nasty things will happen:

2.1. Spurious--but visually convincing--trends may appear in your data (see Figures 1 and 2, above). These may occur in many different forms, including linear trends, abrupt steps, periodic or aperiodic cycles, etc.

2.2. Although the regression coefficients (or other results of your analysis) will still be unbiased, you will underestimate their uncertainties, potentially by large factors. For serial correlation coefficients of $\rho\approx0.6$

and above, the actual standard error of the regression slope will be more than twice s_b, the estimate provided by conventional regression techniques This effect *increases* with sample size, and increases drastically with ρ (see Figure 3, below). (The symbol ρ denotes the true correlation between residuals; it is estimated by *r*, defined in section 3.2, below.)

2.3. Because uncertainties will be underestimated, confidence intervals and prediction intervals will be too narrow.

2.4. Estimates of "goodness of fit" will be exaggerated.

2.5. Estimates of statistical significance will be exaggerated, perhaps vastly so. Your actual false-positive rate can be *much* higher than the α-value used in your statistical tests (see Figure 4). For serial correlation of $\rho \approx 0.5$, tests for the significance of the regression slope at α=5% will have an actual false positive rate of roughly 30% (six times α); the rate for α=1% will be roughly 15% (15 times α), and the rate for α=0.1% will be roughly 7% (70 times α). If ρ=0.9, the false positive rate can be over 50 percent, and largely independent of α. Getting more data will not help; this effect *increases* with sample size. You can also make things *much worse* by selectively picking out the parts of a time series in which your eye finds an apparent trend, or selectively fitting to nonlinear trends that your eye finds in the data.

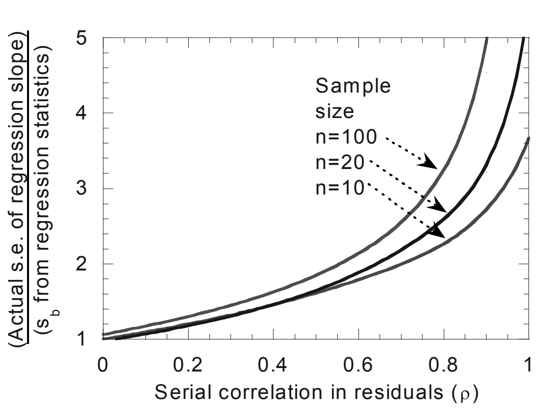

Figure 3. Ratio of actual standard error of regression slope to estimate of standard error (s_b) calculated from ordinary linear regression, if serial correlation in residuals is overlooked.

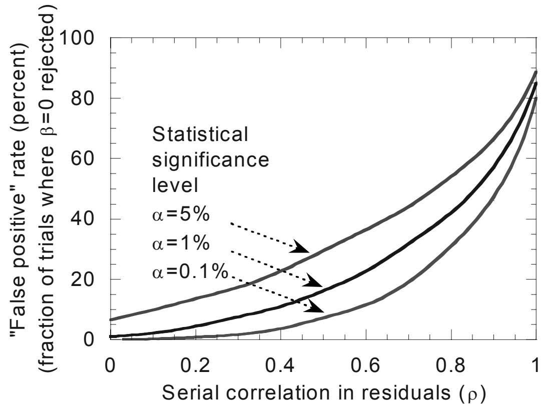

Figure 4. Fraction of Monte Carlo trials in which regression slope *b* was different from zero (at statistical significance level α=0.05, 0.01, and 0.001), when null hypothesis was actually true (true slope was β=0, n=100 points).

Serial correlation can corrupt many different kinds of analyses (including t-tests, ANOVA's, and the like), but its effects on linear regression are most widely appreciated. Serial correlation is particularly problematic when one is trying to detect long-term trends; in fact, some noted authorities declare that serial correlation makes linear regression invalid for trend detection. These noted authorities are wrong. Serially correlated data can be analyzed with many different methods, including regression, as long as the serial correlation is properly taken into account.

This toolkit presents basic techniques for detecting and treating serial correlation in environmental time series data. Although serial correlation is an almost universal feature of environmental data, it is difficult to find a serious treatment of the subject outside of several weighty treatises on regression methods (most of the work in this area has been done by econometricians). Because this material is relatively inaccessible, I will present it in some detail here. The emphasis is on trend analysis (that is, determining how a variable changes with time) using linear regression.

Important note: What we are worrying about here is serial correlation in the *residuals* of a variable, not serial correlation in the *variable itself.* Obviously, every variable that has a trend is correlated with itself through time, but that in itself presents no problems. What creates problems is autocorrelation in *residuals*, not *variables*.

3. Detecting serial correlation

We begin by assuming that we have two variables X and Y, which we have measured *n* times (X_i and Y_i for i=1..n). We assume that these measurements have been sorted by time, so that i=1 is the first measurement and i=n is the last. We will further assume that these measurements are roughly equally spaced in time (this is rarely crucial). We are trying to determine how Y depends on X. If we assume that the underlying relationship between X and Y is linear, then

$$Y_i = a + bX_i + e_i \tag{1}$$

where a and b are constants, and e_i is the i^{th} residual (that is, the difference between Y_i and $\hat{Y}_i$, the predicted value at X_i):

$$\hat{Y}_i = a + bX_i \quad , \quad e_i = Y_i - \hat{Y}_i = Y_i - (a + bX_i) \tag{2}$$

If we are trying to detect a trend, then the X variable will be time itself. Whether X is time, or is some other explanatory variable, the same techniques can be used. If Y is a nonlinear function of X, e.g.,

$$Y_i = a + b_1X_i + b_2X_i^2 + e_i \quad , \quad e_i = Y_i - \hat{Y}_i = Y_i - (a + b_1X_i + b_2X_i^2) \tag{3}$$

we will need to use polynomial regression or nonlinear regression. If Y is a function of several different X variables X_1, X_2, X_3, etc. (where the subscripts 1, 2, 3, etc. refer to the different variables, and i refers to the individual measurements),

$$Y_i = a + b_1X_{1,i} + b_2X_{2,i} + \cdots + e_i \quad , \quad e_i = Y_i - \hat{Y}_i = Y_i - (a + b_1X_{1,i} + b_2X_{2,i} + \cdots) \tag{4}$$

we will need to use multiple regression. Although the particulars of the regression techniques may vary, the general approach to detecting and treating serial correlation remains the same.

Serial correlation occurs when residuals at adjacent points in time are correlated with one another; that is, when e_i and e_{i-1} are, on average, more similar than pairs of residuals chosen randomly from the time series. Since we are concerned with serial correlation in the residuals e_i (*not in the Y_i themselves*), testing for serial correlation is inherently a diagnostic technique that comes after the model is estimated. The model must first be fitted (that is, values of a and b must be estimated by linear regression, or potentially by some other fitting technique), and the residuals e_i calculated by (2) above. The correlation between residuals can be detected in at least three different ways:

3.1. Plotting the residuals against their lags (that is, plotting e_i against e_{i-1}) will visually reveal any significant correlation that is present.

3.2. Calculating the correlation coefficient *r* for the correlation between the residuals and their lags is useful because (a) it quantifies how much serial correlation there is, and (b) it permits you to assess the statistical significance of whatever correlation you find. The correlation between the residuals e_i and their one-period lags e_{i-1} is called the *lag-1 autocorrelation*, which we will denote by *r* (be careful not to confuse this with the correlation between X and Y, which is also represented by the symbol *r*). The lag-one autocorrelation can be calculated most easily while you are looking at the plot of e_i versus e_{i-1} (you *did* look at your residuals, didn't you?); just fit a least-squares line through the data, and note the value of *r*. Remember from the linear regression toolkit that the statistical significance of the correlation coefficient *r* is the same as the statistical significance of the regression slope *b*. Alternatively, you can calculate *r* and its standard error s_r from the formulas given in the linear regression toolkit (just remember that you want the correlation between e_i and e_{i-1}, not between X and Y).

The lag-1 autocorrelation *r* is an approximation to the true serial correlation ρ, plotted in Figures 3 and 4. Unfortunately, *r* always *underestimates* ρ. For time series of 100 points, ρ is about 0.05 larger than *r* for *r*<0.5, and about 10% larger than *r* for *r*>0.5. This effect decreases as sample size increases.

If you happen to notice that *r* approximately equals the regression slope *b*, do not be surprised: recall that in linear regression, $r=bs_X/s_Y$, and note that here $s_X=s_Y$ (e_i and e_{i-1} have the same distribution--indeed, they're exactly the same numbers, just paired up with their own lags), so *r=b*.

The magnitude of *r* is more important than its statistical significance; if *n* is large enough, even a trivially small amount of serial correlation can be statistically significant. You can see this by noting that the standard error of *r* is

$$s_r = \sqrt{\frac{1-r^2}{n-2}} \; , \quad t = \frac{r}{s_r} = r\sqrt{\frac{n-2}{1-r^2}} \; , \quad \text{so for r<<1 and n large,} \quad t = r\sqrt{n} \quad so \quad t > t_{\alpha,n-2} \;\; if \;\; r > \frac{t_{\alpha,n-2}}{\sqrt{n}} \tag{5}$$

Since for large n and reasonable α, $t_{\alpha,n-2}$ is approximately 1.6-2, n's in the hundreds will make small r's statistically significant (or you can save yourself the calculations; critical values of *r* are tallied in Table B.17 of Zar).

3.3. Estimating the correlation nonparametrically is preferable to calculating *r* if either (a) the relationship between e_i and e_{i-1} is nonlinear, or (b) the residuals are not normally distributed. In practice, (b) is a more common problem. When the residuals e_i are leptokurtic (that is, their distribution has longer tails than normal), parametric tests will tend to overstate the statistical significance of *r*.

The simplest nonparametric measure of correlation is the Spearman rank correlation coefficient, r_s (Zar, section 18.9). To calculate r_s, simply calculate the correlation coefficient *r*, but do so on the *ranks* of each of the e_i rather than their numerical values. For small values of n, critical values of r are given in Table B.20 of Zar. For n>20, the statistical significance of r_s can be evaluated in exactly the same way as *r*, that is,

$$t = r_s\sqrt{\frac{n-2}{1-r_s^2}} \qquad (6)$$

Another nonparametric measure of correlation is Kendall's Tau; it will not be explained here because the calculation procedure is somewhat complex and a little tedious. For the curious, it is detailed in Helsel and Hirsch (1991), section 8.2, and in any standard nonparametric statistics text.

3.4. The Durbin-Watson test for serial correlation (Durbin and Watson, 1951) is the standard method for detecting serial correlation. The Durbin-Watson statistic, *d*, is calculated as,

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \qquad (7)$$

This value of *d* is then compared to critical values, given in the appended table, where n is the number of data points and k is the number of X variables (for simple linear regression, k=1). There are two critical values, d_L and d_U. The decision rules are as follows: if $d<d_L$, then reject the null hypothesis (that there is no serial correlation); that is, if $d<d_L$, then the serial correlation is statistically significant. If $d>d_U$, then accept the null hypothesis. If d is between d_L and d_U, the test is inconclusive.

The Durbin-Watson test has been explained here because it is commonly used. However, it offers no practical advantages over significance tests with the correlation coefficient *r*. Indeed, the Durbin-Watson statistic can be shown to be functionally equivalent to 2(1-*r*) for n>>1. To see that this is the case, notice that the mean of the residuals is zero, so that the sums in *d* can be re-expressed as variances, then use the formulas from the error-propagation toolkit to expand the numerator thus:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = \frac{(n-1)Var(e_i - e_{i-1})}{nVar(e_i)} = \frac{n-1}{n}\frac{Var(e_i) + Var(e_{i-1}) - 2Cov(e_i,e_{i-1})}{Var(e_i)} \qquad (8)$$

Now, note that if n is large, then (n-1)/n≈1 and $Var(e_i)=Var(e_{i-1})$, so,

$$d \approx \frac{Var(e_i) + Var(e_i) - 2Cov(e_i,e_{i-1})}{Var(e_i)} = \frac{Var(e_i) + Var(e_i)}{Var(e_i)} - \frac{2Cov(e_i,e_{i-1})}{s_{e_i}s_{e_{i-1}}} = 2 - 2r \qquad (9)$$

Because the Durbin-Watson statistic, like *r*, is parametric, it may yield misleading estimates of statistical significance if the residuals are markedly non-normal.

4. Accounting for serial correlation

Serial correlation can be treated by several different methods. If the serial correlation arises because a time-dependent effect or an explanatory variable has been excluded from the analysis, these should be included as new explanatory variables (4.1 through 4.3, below). Only if the serial correlation cannot be treated as an intelligible signal, should it be treated as random, serially correlated noise (4.4, below). This illustrates a general principle of savvy data analysis: *Never treat as unintelligible noise what you could otherwise treat as an intelligible signal.*

4.1. Serial correlation due to long-term trend in Y. If Y varies with time as well as with X (see 1.1 above), then time should be included as an explicit variable in the analysis. If the time trend is linear, then a simple linear regression of Y against X becomes a multiple regression of Y against X and t:

$$\hat{Y}_i = a + b_1 X_i + b_2 t_i \tag{10}$$

If the time trend is nonlinear, then a more complex model may be required.

4.2. Seasonal or periodic variation in Y. If the long-term trend in Y has been removed, but there is still a seasonal pattern in the residuals (see 1.2 above), then this seasonal variation should be explicitly included in the analysis. The most straightforward case (and one that is conveniently rather common) is a roughly sinusoidal cycle in the residuals each year. This can be conveniently modeled by multiple linear regression as,

$$\hat{Y}_i = a + b_1 X_i + b_2 t_i + b_3\, sin(2\pi q_i) + b_4\, cos(2\pi q_i) \tag{11}$$

where t_i is the time in years, and q_i is the non-integer component of t_i. Thus q_i runs from 0 to 1 in the course of every year. The values of b_3 and b_4 jointly determine both the amplitude of the seasonal cycle and its phase (that is, at what time of year the peaks and valleys occur).

4.3 Missing explanatory variables that are serially correlated (see 1.3 above). For example, stream chemistry often varies with streamflow, among other variables. Because streamflow is serially correlated (after heavy rains, streamflow remains high for an extended period), stream chemistry is also often serially correlated. Explicitly including streamflow as an explanatory variable (e.g., X_{2i}=log(flow$_i$)) will eliminate the serial correlation in chemistry that is due to streamflow. Note that this may not completely eliminate the serial correlation in Y, since stream chemistry can also be serially correlated for other reasons as well.

4.4. Serially correlated random noise. If the serial correlation cannot be explained away, you must account for it in your analysis. There are three basic ways of doing so.

4.4.1. Subsample your data set. If the correlation between e_i and e_{i-1} is r_{lag1}, then the correlation between e_i and e_{i-2} is $(r_{lag1})^2$, and the correlation at lag k (i.e., between e_i and e_{i-k}) is $(r_{lag1})^k$. Thus, by subsampling at wide enough intervals, you can make the serial correlation between the residuals as small as you choose in the new, subsampled data set. If your original measurements were taken at uneven intervals, you should subsample at roughly even intervals, rather than every k^{th} measurement. Then repeat your regression analysis on the subsampled data set; remember to check whether the new residuals are serially correlated.

The advantage to the subsampling approach is that it's simple. The drawback, of course, is that you throw away most of your data set! Helsel and Hirsch argue that since serial correlation makes closely spaced measurements largely redundant, throwing many of those measurements away does not destroy much information. They are right, *if* you are analyzing a simple trend (Y as a function of time), with no other X variables in the problem. If other variables are present, and if they aren't serially correlated and redundant like Y is, then you may be discarding a lot of your information about them.

4.4.2. Average your data over longer periods. Another approach is to group your data into time periods (months, seasons, whatever...) and calculate a mean or median for each period, then conduct your analysis on the average for each period rather than on the individual measurements. Note that if the sampling frequency has changed dramatically during your sampling program, the variance of the period averages will change also. You should either prevent this (by averaging over roughly the same number of samples in each period), or correct for it, by using *weighted least squares* (many canned packages can do this--see any good regression text, like Draper and Smith (1981) or Neter et al. (1990) for details) where the weights equal the number of measurements in the i^{th} average.

Unlike the subsampling approach, the averaging approach uses all the data. Again, however, if there are other X variables, and if there is useful information in their variation within periods, averaging will throw this information away.

4.4.3. Compute regression with autocorrelated residuals. These methods are more theoretically complex than simply subsampling or averaging, but they are straightforward to apply, and they don't throw away information like the simple methods do. The theory behind all of these methods is the same, and it goes like this:

Linear regression assumes that the "true" underlying linear relationship between X and Y is,

$$Y_i = \alpha + \beta X_i + \varepsilon_i \tag{12}$$

where α and β are the "true" slope and intercept (corresponding to the parameters *a* and *b* that you would estimate from any particular data set), and ε_i are the errors in Y. We assume that those errors are serially correlated with a "true" correlation of ρ (corresponding, again, to the correlation coefficient *r* that you would estimate from the lagged residuals of any particular data set), thus:

$$\varepsilon_i = \rho\varepsilon_{i-1} + \xi_i \tag{13}$$

where ξ_i is random, and *not* serially correlated. Note that a fraction ρ of each ε_{i-1} is passed on to the next ε_i; you can think of $\rho\varepsilon_{i-1}$ as the redundant part of ε_i, and ξ_i as the non-redundant part. Linear regression assumes that the errors are uncorrelated, like the ξ_i, not like the ε_i, which are serially correlated. So the question naturally arises: can we arrange things so that our residuals are the uncorrelated ξ_i, which linear regression knows how to handle, rather than the serially correlated ε_i? Watch this: solve the linear relationship for ε, at both time *i* and *i*-1:

$$\varepsilon_i = Y_i - \alpha - \beta X_i \quad \textit{and} \quad \varepsilon_{i-1} = Y_{i-1} - \alpha - \beta X_{i-1} \tag{14}$$

Now, substitute both of these ε's into equation (13) above, and rearrange terms:

$$Y_i - \alpha - \beta X_i = \rho(Y_{i-1} - \alpha - \beta X_{i-1}) + \xi_i \ \ or \ \ Y_i - \rho Y_{i-1} = \alpha(1-\rho) + \beta(X_i - \rho X_{i-1}) + \xi_i \tag{15}$$

Note that the serially correlated errors ε_i have disappeared, and only the well-behaved error ξ_i remains. So if we transform X and Y into two new variables, X^* and Y^*, as $X^*_i=X_i-\rho X_{i-1}$ and $Y^*_i=Y_i-\rho Y_{i-1}$, we can rewrite (15) as,

$$Y_i^* = \alpha^* + \beta^* X_i^* + \xi_i \tag{16}$$

Because (16) is not plagued by serial correlation, we can use linear regression to estimate α^* and β^*, which we can translate back to α and β straightforwardly: $\beta=\beta^*$, and $\alpha=\alpha^*/(1-\rho)$.

So, now you see the outline of the approach: transform your X and Y variables to X^* and Y^*, and then regress Y^* on X^*; the regression slope b^* will be the best estimate of the true slope β, and $a^*/(1-\rho)$ will be the best estimate of the intercept.

So where, you ask, do you get the autocorrelation parameter ρ? That's precisely the tricky part! There are three approaches that are simple enough to explain here:

4.4.3.a. Take first differences. In time-series analysis (see PDQ Statistics, Chapter 8), you typically "difference" your variables, such that $X^*_i=X_i-X_{i-1}$ and $Y^*_i=Y_i-Y_{i-1}$. This is the same as assuming that ρ=1. Those who are familiar with time-series analysis will often suggest applying this same procedure to analyses for trends. There are at least three reasons why this is generally *not* the right approach:

(a) Assuming ρ=1 is *guaranteed* to overestimate the true ρ, perhaps by a lot.

(b) Assuming ρ=1 means that you divide by zero when you calculate $a=a^*/(1-\rho)$, making the intercept undefined.

(c) If X is time, and your measurements are evenly spaced through time, all of the X^*_i take on *exactly the same value*, making it impossible to do the regression at all!

Either of the next two methods is generally preferable.

4.4.3.b. The Cochrane-Orcutt procedure. Why assume that ρ=1, when you already have a more reasonable estimate of ρ, namely the lag-1 autocorrelation *r*? That's the idea behind the Cochrane-Orcutt procedure (see Neter et al., 1990). Here's how it works.

(a) Regress Y on X, save the residuals, and test them for serial correlation. If you don't find any, then stop; there's no serial correlation to correct for.

(b) If you *do* find serial correlation, calculate the serial correlation coefficient *r* for the correlation between the residuals e_i and their lags e_{i-1}. Then transform the X's and Y's to X^*'s and Y^*'s, using $X^*_i=X_i-rX_{i-1}$ and $Y^*_i=Y_i-rY_{i-1}$. Then regress Y^* against X^*, yielding the fitted intercept a^* and slope b^*.

(c) Plot the residuals from this regression (of Y^* on X^*) against their lags, to see whether the serial correlation has been eliminated. If so, then the procedure is finished, yielding b^* as the best estimate of the true slope, and $a^*/(1-r)$ as the best estimate of the intercept.

(d) If not, update your estimate of *r* by calculating different residuals, from a^* and b^*, and the *original* X's and Y's:

$$\varepsilon_i = Y_i - \frac{a^*}{1-r} - b^* X_i \tag{17}$$

and calculating the lag-1 autocorrelation *r* of these residuals with their lags. Then, using this *r*, iterate back through step (b), calculating new X^*'s and Y^*'s from your new *r* and your *original* X's and Y's.

The Cochrane-Orcutt procedure should terminate in one or two iterations (and if it doesn't, you might go slightly daft shuffling your e's and *r*'s). If it fails to converge, or it converges too slowly, you need the Hildreth-Lu procedure (see below).

4.4.3.c. The Hildreth-Lu procedure. The Hildreth-Lu procedure (see Neter et al., 1990) arises from the observation that you can rewrite equation (15) such that Y_i is a function of Y_{i-1}, X_i, and X_{i-1}:

$$Y_i = \rho Y_{i-1} + \alpha(1-\rho) + \beta(X_i - \rho X_{i-1}) + \xi_i \tag{18}$$

The Hildreth-Lu procedure simply turns ρ into *a parameter that is fitted to the data*, just like α and β are. The advantage of Hildreth-Lu is that it estimates ρ, α, and β directly, and it does so all at once. There is only one drawback: equation (18) is *nonlinear* in the parameters, so it can't be solved by multiple linear regression. Instead, it must be fitted using a nonlinear regression algorithm (such as the "Nonlinear Fit" platform in JMP). Using such an algorithm, you search for the combination of r, a, and b that gives the best match between the Y_i and the predicted values $\hat{Y}_i$:

$$\hat{Y}_i = r(Y_{i-1} - a - b_1 X_{i-1}) + a + bX_i \tag{19}$$

Equation (19) can be readily rewritten for multiple X variables ($X_{1,i}$, $X_{2,i}$, etc.) as follows:

$$\hat{Y}_i = r(Y_{i-1} - a - b_1 X_{1,i-1} - b_2 X_{2,i-1} - \cdots) + a + b_1 X_{1,i} + b_2 X_{2,i} + \cdots \tag{20}$$

One difficulty with the Hildreth-Lu procedure is that nonlinear fitting algorithms may not calculate standard errors for *a*, *r*, and the *b*'s, and even if they do, they may not do so reliably. JMP's Nonlinear Fit platform does give standard errors, and I have never found them to be inaccurate, but in the world of nonlinear parameter estimation, there are few bombproof mathematical guarantees (in marked contrast to linear estimation). If your algorithm does not estimate standard errors, or if you don't trust its estimates, you can always check them with step (b) of the Cochrane-Orcutt procedure: using the value of *r* estimated by Hildreth-Lu, transform X to X^* and Y to Y^*, then regress Y^* on X^* to get estimates of a^* and b^* (which, since they come from linear regression, will be accompanied by standard errors). These estimates of a^* and b^* should be consistent with your Hildreth-Lu estimates of a and b, once you remember that you need a factor of 1-r to transform from a^* (or its standard error) back to *a* (or its standard error).

4.4.4. Adjust your uncertainty estimates to account for the loss of degrees of freedom. Serial correlation leads to underestimates of uncertainty (and thus overestimates of statistical significance) because when residuals are not independent, the effective number of degrees of freedom can be much smaller than the sample size would indicate. In principle, one should be able to account for this loss of degrees of freedom by calculating the "effective" sample size--that is, the number of independent measurements would be functionally equivalent to your set of non-independent measurements. You then use this "effective" sample size, n_{eff}, in place of *n* in calculating standard errors and performing significance tests. A commonly used method for estimating n_{eff} comes from Mitchell et al. (Climatic change, WMO Technical Note 79, 1966):

$$n_{eff} = n\frac{1-\rho}{1+\rho} \tag{21}$$

The drawback to this approach is that it is based on the *true* autocorrelation coefficient ρ, and in practice one only has the sample autocorrelation coefficent *r*, which underestimates ρ and therefore leads to overestimates of n_{eff}. Nychka et al. (Confidence intervals for trend estimates with autocorrelated observations, unpublished manuscript, 2000) have proposed that this problem can be addressed by adding a correction factor to (21) as follows:

$$n_{eff} = n\frac{1-r-0.68/\sqrt{n}}{1+r+0.68/\sqrt{n}} \tag{22}$$

Nychka et al.'s Monte Carlo simulations show that this gives reasonably accurate and unbiased uncertainty estimates for trend analyses. They also note that when n_{eff}<6, estimates of uncertainties and confidence intervals are likely to be unreliable even if equation (22) is used.

Important note: The procedures of section 4.4 are reliable only if two conditions hold:

(1) The residuals e_i must depend only on the previous residuals e_{i-1}, plus random noise (so-called "first-order autoregressive errors"). More complicated forms of serial correlation can occur (such as where the e_i are positively correlated with the e_{i-1}, and negatively correlated with the e_{i-2}--just one of the many possibilities), but they are too complex, and numerous, to discuss here.

(2) You must correctly intuit the form of the "real" equation relating X and Y. In Figure 2, for example, you could either fit a straight line (in which case the residuals would be large, and would show very strong serial correlation), or you could fit a sine curve (in which case the residuals would be smaller, and less strongly correlated). In this case the first approach would be correct, but we only know this because we know the function that generated the data; it would be almost impossible to tell simply by looking at the plot. The lesson from this is sobering: serial correlation can create visually coherent patterns that are nonetheless completely spurious. That is why, if there is evidence of serial correlation, you should be suspicious of complicated patterns of behavior, like the abrupt reversal in Figure 1 or the oscillation in Figure 2, unless you have strong theoretical reasons to expect such behavior--even if the data look visually convincing.

A note to close section 4: In many situations you will have the option of treating serial correlation several different ways: *causally* (e.g., Y varies seasonally because it depends on streamflow, which varies seasonally, as in (4.3) above), *phenomenologically* (e.g., Y varies with a seasonal cycle that can be fitted with a sine wave as in (4.2) above), or *statistically* (e.g., Y varies as serially correlated noise, as in (4.4) above, without an obvious pattern that can be fitted). The order of preferences is clear: treat serial correlation *causally* if you can (as in (4.3) above), or otherwise *phenomenologically* (as in (4.1) or (4.2) above). Treating serial correlation as *statistical noise* (as in (4.4)) is a last resort, to be invoked only if none of the other approaches is viable (if, for example, the missing causal variable can't be obtained, and the variation in Y doesn't fit any simple phenomenological description).

Two final notes:

Environmental data can also be *spatially* autocorrelated, such that measurements at nearby locations have correlated residuals. Spatial autocorrelation can distort trends across space, just as temporal autocorrelation can distort trends through time. In more than one dimension, things become mathematically more complex; problems with spatially autocorrelated data are addressed in the field of *geostatistics* .

A common misconception is that nonparametric techniques take care of serial correlation. This is not correct. Nonparametric techniques are robust against outliers, but they are not robust against serial correlation. However, there are nonparametric techniques that are specifically designed to deal with particular kinds of serial correlation. One example is the Seasonal Kendall Test, a nonparametric test that accounts for seasonal effects (but not other sources of autocorrelation).

References:

Draper, N. R. and H. Smith, *Applied Regression Analysis*, 709 pp., Wiley, 1981.

Durbin, J. and G. S. Watson, Testing for serial correlation in least squares regression II, *Biometrika*, **38**, 159-178, 1951.

Helsel, D. R. and R. M. Hirsch, *Statistical Methods in Water Resources*, 522 pp., Elsevier, 1992, section 9.5.4.

Neter, J., W. Wasserman and M. H. Kutner, *Applied Linear Statistical Models*, 1181 pp., Richard D. Irwin, Inc., 1990, Ch. 13

Pindyck, R. S. and D. L. Rubinfeld, *Econometric Models and Economic Forecasts*, McGraw-Hill, 1981.

von Storch, H. and F. W. Zwiers, *Statistical Analysis in Climate Research*, 484 pp, Cambridge University Press, 1999.