Stat 505 - Environmental Statistics, Fall 2013 - HW 3

Due: 5 pm, Friday, Nov 22 to 2121 Snedecor, my mailbox, or e-mail

Remember, you are to do 3 of the following 4 problems. You can choose. If you do all 4, I'll grade them all and drop the lowest. Further information and suggestions for organizing answers to open ended data analysis problems (like problem 4) are on the 'homework information' part of the class web site.

0. Write a few sentences describing what you want to do for your project. If working as a pair, who are you working with? If doing a data analysis, where will you get the data?

1. The file temp.txt on the class web site contains data on the sea surface temperature averaged over the Northern Hemisphere oceans from 1958 to 1992. The measurements are taken quarterly. The columns labelled W, Sp, Su, and F are the winter, spring, summer and fall averages. For this problem, we will focus on the spring temperature data. The investigators are interested in a linear trend over time but they don't know whether the year-year variation around the trend line is normally distributed or not.

   Do the following using the spring measurements.

   (a) Estimate the linear trend (the slope) using ordinary least squares regression and calculate a 95% confidence interval for the slope.

   (b) Calculate the lag-1 autocorrelation coefficient. Is there any evidence of autocorrelation?

   (c) Is a linear trend reasonable? Or, is some more complicated model needed? Is the assumption of normal errors reasonable?

   (d) Use non-parametric regression (i.e. loess) to evaluate the lack of fit to a linear trend. Is a linear trend reasonable?

   (e) Use a non-parametric test to evaluate H0: no trend. You may ignore any autocorrelation if present.

   (f) Estimate the slope and calculate a 95% confidence interval, using a non-parametric method. You may ignore any autocorrelation if present.

   (g) Which method gives you a narrower confidence interval, the OLS estimate or the nonparametric estimate? Is this what you expected? Explain why or why not.

   (h) If you could only report one estimate of trend, which one would you choose? Explain/justify your choice, including plots if helpful to support your decision.

2. This question has two completely different parts.

   (a) In class, I described trignometric regression to model seasonal data. My description was in terms of a sin() variable and a cos() variable. Each is a function of $2\pi X_i$, where X_i is scaled so that 1 unit change in X_i is one year. The coefficients for those two terms could be reexpressed in more interpretable terms as an amplitude and a phase. Amplitude is one-half the difference between the maximum and minimum value; phase is the horizontal shift in the location of the maximum of a cosine curve. Phase of 0 is a maximum at Jan 1 (day 0 of the year); phase of π is a maximum at July 1 (day 182 of the year).

i. You have fit a model including trigonometric terms, i.e.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 \sin(2\pi X_i) + \beta_3 \cos(2\pi X_i) + \varepsilon_i,$$

so you have $\hat{\beta}_1$ and $\hat{\beta}_2$. These are least squares estimates. You are really interested in the coefficients for the model with coefficients for amplitude and phase:

$$Y_i = \beta_0 + \beta_1 X_i + A \cos(2\pi X_i - \beta) + \varepsilon_i,$$

where A is the amplitude of the seasonal cycle and β is the phase. How do you calculate $\hat{A}$ and $\hat{\beta}$ from $\hat{\beta}_1$ and $\hat{\beta}_2$?

ii. What can you say about the "quality" of your estimators in the previous part? In other words, are they simply ad-hoc, or is there some statistical principle justifying your estimators as reasonable? If so, what is that principle.
Hint: When errors are normally distributed with constant variance, least squares estimates are also another type of estimate.

iii. Fit the trigonometric regression to the full (all four season) sea surface temperature data. Estimate the amplitude and phase of the seasonal cycle.
Hint: Seasons should be fractions of years. When you construct your "year" or "time" variable, please make the winter (W) values be X.0, so spring (Sp) values are X.25, etc.

iv. Test whether the amplitude of the cycle = 0.

The second part of this question concerns the effect of autocorrelated errors on the standard error of the estimated slope in a linear regression. Consider a sequence of 11 annual observations $\{Y_i, i = 1..11\}$. For simplicity, the mean year number is subtracted from all times, so the corresponding $X_i's$ are $\{-5, -4, -3, ...4, 5\}$. This centering of the X values does not change the definition or interpretation of the linear slope. After centering, the $\boldsymbol{X^T X}$ matrix is diagonal and especially easy to invert.

Assume that the error variance, σ^2, is known. Answers as formulae are a bit tricky and NOT expected. I assume you will compute numerical answers and report as some number times the population variance, σ^2.

I have put R code for some useful matrix calculations, matrix.r, on the class web site.

i. When the errors are independent, what is the variance of the estimated slope? This should be expressed as $k\sigma^2$; you need to calculate k, where k is a specific number.

ii. If the errors are correlated with variance-covariance matrix $\Sigma = \sigma^2 V$, the variance of the OLS (usual regression) estimate, $\hat{\beta}_{OLS}$, is $\boldsymbol{(X^T X)^{-1}(X^T \Sigma X)(X^T X)^{-1}}$. If the errors follow an AR(1) process with $\rho = 0.4$, what is the variance of the estimated slope? Don't worry about the intercept part of β_{OLS} - just report the variance of the slope. Again, the answer should be expressed as $k\sigma^2$.
R hint: You can construct a matrix with bands, e.g.

$$\begin{bmatrix} 1 & a & b & c \\ a & 1 & a & b \\ b & a & 1 & a \\ c & b & a & 1 \end{bmatrix}$$

using the `toeplitz(c(1,a,b,c))` command. The vector you provide is the top row of the matrix.

iii. If ρ is known, then the GLS estimate, β_{GLS}, is given by $(\boldsymbol{X^T\Sigma^{-1}X})^{-1}(\boldsymbol{X^T\Sigma^{-1}Y})$, which has variance $(\boldsymbol{X^T\Sigma^{-1}X})^{-1}$. Calculate the variance of the GLS estimate. Again your answer should be expressed as $k\sigma^2$ and only report the variance of the slope.

iv. In class, I claimed that autocorrelation was a problem for inference on the slope. This clearly depends on the magnitude of the correlation. You have decided that mis-estimating the variance by a factor of 20% is a problem (i.e. when the more appropriate value is 2.5, reporting something ≤ 2 or less or ≥ 3.125). Is reporting the OLS variance a problem when the lag-1 correlation is 0.4?

v. Even though the variance of the OLS estimator can be "patched up", it might be less precise than more appropriate estimators. Relative efficiency of a pair of estimators is the ratio of their (correctly calculated) variances. Efficiencies are usually reported so that the estimator with the smaller variance has an efficiency $> 100\%$. Is the GLS estimator more efficient than the OLS estimator? If so, by how much?

3. The properties of Theil-Sen regression are not nearly as well understood as are the properties of traditional linear regression. Similarly, the small sample properties of the degrees of freedom adjustments are not well known. This problem evaluates the robustness and coverage of confidence intervals for a linear trend. Please consider four procedures to construct a confidence interval for the trend:

linear: Fit a linear regression assuming independent observations; calculate a confidence interval using t quantiles, i.e. using standard normal theory.

theil: Calculate all pairwise slopes, calculate a confidence interval using Gilbert's estimator. I have posted code to calculate the Theil-Sen estimate.

ar: Fit a linear regression assuming an ar(1) error structure, using N-2 as degrees of freedom

arkr: Fit a linear regression assuming an ar(1) error structure, use Kenward Rogers approximation for d.f. The Kenward-Rogers approximation is a combination of adjustments designed to make more appropriate inferences about trends or differences when observations have an arbitrary set of correlations. R does not (to my knowledge) compute the KR approximation. This needs to be done in SAS. Code is posted on the SAS part of the class web site.

Consider a sequence of 20 annual observations. The errors are assumed to follow an ar(1) process with rho = 0.5 and variance = 0.1. The trend is linear, i.e. E $Y_i = \beta_0 + \beta_1 X_i$, but you can choose β_1. Use simulation to estimate:
1) the average estimated variance of the slope
2) the empirical coverage of 90% and 99% confidence intervals for the slope,
for each of the four estimators, "linear", "theil", "ar" and "arkr". Which estimator is closest on average to the true variance of the slope (estimated by the empirical variance in the slopes)? Which confidence interval method is closest on average to the nominal coverage?

This probably requires simulating data in SAS and in R. I'm happy to show you how to do either (or both).

4. The data in SkaggittNH3 are ammonia (NH3) concentrations in a small river in Skaggitt County, Washington. The stream was sampled more or less monthly from October 1973 through November 1998, except that data are missing from October 1974 to September 1976. The data file includes three columns: the date, the observed value, and a censoring flag. These are EPA data qualifiers, so U means that the observation is less than the detection limit. The reported value is then the

detection limit. It appears that the trend from 1973 to Dec 1992 is different from that between Jan 1993 and 1998. You have been asked to estimate the trend, calculate a confidence interval for the trend, if possible, and test whether there is non-zero trend. Please do this separately for each period. If you can, please test whether the trend in the early period differs from that in the later period.

You have free choice of methods, but please justify why your choices are reasonable. I have posted code to fit linear regressions to censored values in trend3.r on the R part of the class web site.