

Due: 5 pm, Friday, May 9, to 3405 Agron (under the door) or my mailbox in Wilson (see one of the secretaries in the main office).

Remember, you are to do 3 of the following 4 problems. You can choose. My intent was to write 1 easier theory problem, 1 easier data analysis problem, 1 harder theory problem, and 1 harder data analysis problem. Your perception may differ; even so, choose 3 problems. 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.

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) Examine the residuals from the linear trend. Does a linear trend seem reasonable, or is some more complicated model needed? Does the assumption of normal errors seem 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  $H_0$ : no trend.
  - (f) Estimate the slope and calculate a 95% confidence interval, using a non-parametric method.
  - (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.
2. This question includes two sets of theoretical calculations pertaining to trend analysis. If you choose this question, you are to do all parts. Within each set, the various parts are closely related, so once you see how to do one part, the others should follow very easily.
    - (a) Consider a sequence of observations collected sequentially over time  $\{Y_i, i = 1..T\}$ , where  $\log Y_i$  iid  $\sim N(1, 1)$ , so the observations are independent and there is no trend. Consider the score statistic  $S = \sum_{j>i} U_{ij}$  where

$$U_{ij} = \begin{cases} 1 & Y_j > Y_i \\ 0 & Y_j = Y_i \\ -1 & Y_j < Y_i \end{cases}$$

Show that  $E S = 0$ . What then is the expected value of Kendall's tau measure of correlation?

- (b) Same basic problem as the previous part, except that now the data are left censored with a single detection limit. That is you observe  $Z_i, i = 1..T$  where

$$Z_i = \begin{cases} dl & Y_i < dl \\ Y_i & Y_i > dl \end{cases}$$

The score function is analogous to that in the previous part, except that it is computed from the  $\{Z_i\}$ :

$$U_{ij} = \begin{cases} 1 & Z_j > Z_i \\ 0 & Z_j = Z_i \\ -1 & Z_j < Z_i \end{cases}$$

Again,  $\log Y_i \sim N(1, 1)$ . The d.l. is 1.0 for all observations. Compute  $E S$  and  $E \hat{\tau}$ .

The last four parts of this question concern 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  $\{-5, -4, -3, \dots, 4, 5\}$ . This centering of the X values does not change the definition or interpretation of the linear slope. After centering, the  $\mathbf{X}^T \mathbf{X}$  matrix is diagonal and especially easy to invert.

Assume that the error variance,  $\sigma^2$ , is known. Answers as formulae are a bit tricky and NOT expected. I assume you will compute the answers and report them numerically, as some number times the variance.

- (c) 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.
- (d) If the errors are correlated with variance-covariance matrix  $\Sigma = \sigma^2 V$ , the variance of  $\hat{\beta}_{OLS}$  is  $(\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \Sigma \mathbf{X}) (\mathbf{X}^T \mathbf{X})^{-1}$ . If the errors follow an AR(1) process with  $\rho = 0.4$ , what is the variance of the estimated slope? Again, the answer should be expressed as  $k\sigma^2$ .
- (e) If  $\rho$  is known, then the GLS estimate of the slope is given by  $(\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1} (\mathbf{X}^T \Sigma^{-1} \mathbf{Y})$ , which has variance  $(\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1}$ . Calculate the variance of the GLS estimate, again your answer should be expressed as  $k\sigma^2$ .
- (f) In class, I claimed that autocorrelation was a problem for inference on the slope. This clearly depends on the magnitude of the correlation. A correlation of 0.9 could be a problem, but a correlation of 0.09 might not be. If the errors have an AR(1) structure with  $\rho = 0.4$ , is there a problem with the estimated variance? Is using the GLS estimator a lot more efficient (i.e. does it have a much smaller variance)?

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.

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.

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  $\beta_1$ . 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.

The county invested considerable money in pollution control during the 1980’s. They suspect that they have reached the lowest possible NH3 concentrations in the stream. This conclusion would be supported if in fact the trend from Jan 1993 to the end of the data was equivalent to 0. Test whether the trend from Jan 1993 on is equivalent to zero, using an equivalence region of (-0.001 mg/l / year, 0.001 mg/l / year).

You have free choice of methods, but please justify why your choices are reasonable.