

Homework 9, handed out Saturday, 31 Oct 2009

on campus Friday, 6 Nov 2009, in lecture (11 am)

or by e-mail to Chuanlong, dclong@iastate.edu, no later than noon.

off campus Monday, 9 Nov 2009, by 4 pm to Nicole Rembert, email: rembeall@iastate.edu or

FAX: 515-294-4040 (please include cover page with Stat 500 / Nicole Rembert).

1. Lack of fit, Correlated errors — Global temperature

Global warming is a contentious environmental issue. The data in `temperature.txt` on the web site are measurements of the worldwide average annual temperature for 108 years from 1880 to 1987. The response variable is expressed as a temperature anomaly. This is the deviation of that year's temperature from the average over all 108 years. The trend in the temperature anomaly is the same as the trend in the temperature. What do these data tell us about the temperature trend? Assessment of the uncertainty in the trend is as important as assessment of the trend.

- (a) Fit a usual linear regression and estimate the slope (temperature change per year). Estimate the s.e. of the slope assuming that errors are independent.
- (b) Is the assumption of a linear trend reasonable? Explain why or why not.
You are free to choose your favorite lack of fit method, but I strongly suggest you consider more than just a residual plot. If want to use a polynomial, please consider whether that polynomial is appropriate for the apparent trend.

All subsequent parts will use the linear model $Temp = \beta_0 + \beta_1 Year + \epsilon$ even though that isn't really appropriate.

- (c) Estimate the lag-1 correlation in the residuals from the linear model, i.e. the correlation between e_t and e_{t-1} .
- (d) Test for a significant lag-1 correlation using the Durbin-Watson test.
- (e) If you have access to SAS: Use PROC MIXED to estimate the slope, allowing for a non-zero correlation between errors. Assume an AR(1) error structure. Estimate the s.e. of the slope.
If you do not have access to SAS (or if proc mixed gives you trouble): Use the Cochran-Orcutt procedure to estimate the slope and the s.e. of the slope. Use the estimated lag-1 correlation from part 1c to transform each value. You might find EXCEL handy for doing the transformation if your software doesn't allow you to lag variables.
- (f) Is the s.e. of the slope similar in parts 1a and 1e? Is the pattern in the s.e.'s what you expect?
Note: there is no statistical test here. I am only looking for your assessment of similar or different.

2. Evaluating unequal variances — website delivery

The website data was described last week. This week we'll use it to look at variances. We'll use the same model as last week:

$$DELIVER_i = \beta_0 + \beta_1 BACKLOG_i + \beta_2 EXPERIENCE_i + \beta_3 PROCESS_i + \beta_4 YEAR_i + \epsilon_i$$

- (a) Fit this regression then look at the residual vs predicted value plot. Is there any evidence that the variance of the errors is related to the predicted value?
- (b) Look at the added variable plots (partial regression residual plots) for experience and for process. If the error variance is associated experience, the experience added variable plot will show the same spreading pattern seen in plots of residuals vs predicted values. Does it seem that the error variance is related to experience? to process?

- (c) Use the Breusch-Pagan test to test whether there is an associate between error variance and experience.
- (d) Repeat part (c) using Z =process.

3. **Estimating the maximum — stainless steel** The data in steel.txt describe the relationship between temperature and the Poisson ratio, a measure of elasticity of steel. These data are from a randomized experiment studying the properties of steel named “316” at temperatures from 150 Celsius to 820 Celsius. One of 7 temperatures were randomly assigned to a piece of steel, the piece was heated, and the Poisson ratio was measured. There are 5 replicated pieces of steel for each temperature. All the material science details of measurement and definition are omitted. The goal is to find the temperature at which the Poisson ratio is maximized.

- (a) Fit a quadratic regression, $Y_i = \beta_0 + \beta_1 T_i + \beta_2 T_i^2 + \varepsilon_i$. Test whether the quadratic coefficient is significantly different from 0.
- (b) Estimate the temperature at which the Poisson ratio is maximized.
- (c) Test H_o : Max temp = 400 Celsius. Report your p-value. What does the result of this test suggest about the s.e. of X_{max} ? In other words, is the se of X_{max} likely to be small (e.g. 10 Celsius), moderate (e.g. 50 Celsius), or large (e.g. 100 Celsius)?

4. **Model building examples** — For each of the following, give: an appropriate regression model and define the X variables in your model (incl. indicator variables). Some examples are looking for estimates. For these: Indicate how the desired quantities could be estimated from the regression parameters. You do not need to worry about standard errors or inference.

Other parts are looking for a test. For these: Indicate how you would construct that test. Your answer could be 'a t-test of (indicate a regression parameter or linear combination of regression parameters) = 0 (or other value)'. It could be 'an F test comparing (indicate a pair of models)' or it could be something else. I do not need formulae for the test statistics.

- (a) A study is comparing the energy content of constant-sized pieces of firewood from different tree species. If you are burning wood to heat a room or a house, a higher energy content is a good thing. One complication is that the energy released depends on the moisture content of the firewood, which is hard to standardize. You have studied three species (Red Oak, White Pine and Black Walnut). You believe that the relationship between energy content and moisture is linear with the same slope for each species. You want to estimate the difference in energy content at 10% moisture content between White Pine and Red Oak.
- (b) Same study as above, except now you wish to test the null hypothesis that the three species have the same energy content at 10% moisture. Again, assume that all three species have the same slope.
- (c) Same study as above, except that now you assume that the three species have different slopes (for the association of moisture content on energy). You want to estimate the difference in energy content at 10% moisture between White Pine and Red Oak.
- (d) Assume that athletic performance for males in a certain sport can be described by a quadratic function of age. You wish to estimate the age at which performance is maximum.
- (e) Assume that athletic performance for males and females in a certain sport can be described by quadratic functions of age. You are willing to assume that the curvature (β_2) is the same for both. You wish to test whether the age of maximum performance is the same for males and females. Hint: The intercepts (β_0) are probably not the same for males and females.

- (f) Toxicologists study the effect of chemical contaminants. They often summarize their data by the quantity EC_{50} , the concentration of chemical that leads to a 50% reduction in response from the control response (at dose=0). Assume that the relationship between Y , a measure of effect, is linearly related to X , the dose of a particular chemical. The intercept, β_0 , is the expected response at dose = 0. You wish to estimate EC_{50} .
- (g) Education researchers are studying whether watching television impacts the performance of graduate students. For each student in a class, they have Y , the exam score, and X , the number of hours spent watching television during the week prior to the exam. They assume that the relationship between Y and X is linear up to 20 hours. After $X=20$ hours, there is no relationship, i.e. the slope is 0 for $X>20$. They wish to estimate the slope for 0 to 20 hours and the expected difference between light television watching (3 hours) and heavy watching (25 hours).