STATISTICS 500 – Fall 2009

Homework 7 - handed out 16 Oct 2009
DUE DATE:

on campus	Friday, 23 Oct 2009, in lecture (11 am) or by e-mailto Chuanlong, dclong@iastate.edu, no later than noon.
off campus	Monday, 26 Oct 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. **Regression Design** — snow data, part 3
I haven't lectured specifically on design (e.g. choice of sample size and choice of X's) because it is just another application of principles we've seen before. We have talked about how the s.e. of the slope depends on the number of observations and spread of the X's. That's all you need to apply ideas from earlier to regression. This problem explores some of those issues, using the snow density and gain study from last week. For all parts here, use the regression of Y = log(gain) on X=density.

   (a) The investigators are concerned about the large s.e. for the slope, β_1. They plan to repeat the study using the same 9 densities. Based on the current data and previous studies, they believe 0.25 is a good estimate of σ, the s.d. of observations among blocks of the same density. It may help to know that $\sum_i (X_i - \bar{X})^2 = 0.4557$ when that sum is calculated for 1 block of each of the 9 densities. The investigators want to reduce the s.e. of the slope to 0.075. They will use n blocks at each density, i.e. the same number of blocks at each density. What is the appropriate n?

   (b) It is expensive to buy these blocks, especially in all the different densities. And, it is hard to get some of the densities. It is easy and cheap to get blocks with densities of 0.001 and 0.686. If the investigators only used these two densities and used n of each, what is the appropriate n so that the s.e. of the slope is 0.075?

   (c) Do you have any concerns about the design in part 1b? Would you ever recommend the design with two densities (part 1b)? Would you ever recommend the design with 9 densities? (part 1a)?

   (d) The investigators consider a third way to do the study. They could use one block from each of the 9 densities and measure each block 10 times. This is a lot cheaper than the designs in either parts 1a or 1b. They tried this design and got the following partial ANOVA tables (Y is log gain and X is density):

   Model: E $Y_{ij} = \beta_0 + \beta_1 X_i$

Source	d.f.	SS
Model	1	96.507
Error	88	0.416
Total	89	96.923

   Model: E $Y_{ij} = \mu_i$

Source	d.f.	SS
Model	8	96.828
Error	81	0.095
Total	89	96.923

   Test for lack of fit of the linear regression of log gain on density using these data. Report your F statistic, p-value and a short conclusion.

   (e) The investigators notice that the MSE from the regression in part 1d is very much smaller than 0.06, the MSE from the regression on the original data using 90 blocks. Would you recommend the design in part 1d? Explain why or why not.
   Hint: it may help to think about / explain why the two MSE's are so different.

2. **Analysis of correlations** – In a study of fuel efficiency, the fuel efficiency (measured as miles per gallon) for 13 compact car brands was measured in both city and highway conditions. The data are in the file mpg.txt on the class web site.

You are welcome to use a computer, but please know how to do the computations (once you have r) by hand. You will have to do the computation for 2e by hand.

(a) You are asked to describe the strength of the association between city and highway mpg. What is the most appropriate measure of association for this problem? Justify your choice.

(b) Compute the correlation coefficient of city and highway MPG for the 13 compact autos and test H0: $\rho = 0$.

(c) Use Fisher's z-transformation to give a 99% confidence interval for the correlation coefficient in the population of compact autos. (You can assume that 13 autos is large enough for this approach).

(d) Test the hypothesis that the population correlation is 0.90.

(e) The test in (2d) was motivated by the fact that an earlier study of 13 luxury automobiles yielded a sample correlation of 0.90. Explain why the test in (2d) is not an appropriate way to compare the two sample correlations. Propose and carry out a suitable test. (Hint: It's not in your notes! You need to combine concepts from this section and an early part of the course.)

3. **review of matrix operations** – The following questions are intended to give you practice with basic matrix operations. I will not ask questions like this on an exam.

In lecture, I stated that the hat matrix, **H**, was symmetric and idempotent. I then used these properties to show that residuals were independent of the predicted values, when the model was true.

(a) Show that **H** is symmetric.
Hint: It will help to know that $\mathbf{X}^T\mathbf{X}$ is symmetric (you don't have to show this), so $(\mathbf{X}^T\mathbf{X})^{-1}$ is symmetric.

(b) Show that **H** is idempotent, i.e. that $\mathbf{HH} = \mathbf{H}$

4. **Multiple regression — weighing bears:** The weight of a bear is an important measure of how well it is doing. Weighing a bear in the wild is difficult. It is a lot easier to measure the length of various parts of the bear's body. The following data were collected in an attempt to find simpler measures that could adequately predict bear weight. 54 bears were located in the wild (assume this is a random sample of bears from this location). Each was anesthetized, weighed, and measured. The data are in bear.txt on the class web site.

We will consider the data from three X variables: chest, headlen, and neck. These are the girth of the chest, the length of the head and the length of the neck. All are measured in inches. The goal is to predict the weight (Y) of the bear, measured in pounds.

For the purpose of this assignment, use only these four variables (weight, chest, headlen, and neck). Do not worry about the rest of the variables in the data set. Also, do not worry about assumptions. We'll assume that the model is linear and that the errors have equal variances. (We'll examine these assumptions next week).

(a) plot weight vs headlength for these bears. Describe the relationship. Does this relationship make sense (biologically)?

(b) Estimate β_H in the simple linear regression model: $Y = \beta_0 + \beta_H X_{headlength} + \epsilon$. What are the units of β_H? Does the value 'make sense'?

(c) Estimate β_H in the multiple linear regression model: $Y = \beta_0 + \beta_H X_{headlength} + \beta_C X_{chest} + \beta_N X_{neck} + \epsilon$. What are the units of β_H? Does this value 'make sense'?

(d) Should the values in parts 4b and 4c be the same? Explain why or why not.

Use the multiple linear regression in 4c for the next four parts of this question.

(e) Test H0: $\beta_H = 0$, using a T statistic. What is the p-value? Write a one-sentence conclusion.

(f) Test H0: $\beta_H = 0$ using the model comparison approach.

(g) Construct a test of H0: $\beta_H = 0$ and $\beta_N = 0$. What is the p-value (at least approximately)? Write a one-sentence conclusion. Hint: thinking about the models that are being compared may help you construct this test and write a conclusion.

(h) Construct a test of H0: $\beta_H = 0$ and $\beta_N = 0$ and $\beta_C = 0$. What is the p-value? Write a one-sentence conclusion.

The final parts of this problem will use the two variable model:

$$Y = \beta_0 + \beta_1 X_{chest} + \beta_2 X_{neck} + \epsilon$$

(i) Estimate β_0, β_1, and β_2 then predict the weight for the following 3 bears:

bear	chest	neck
A	35 in	20 in
B	55 in	30 in
C	50 in	15 in

(j) Calculate the s.e. of the predicted average (i.e. the s.e. of the line) for each of the three bears in the previous part.

(k) For these 54 bears, the average chest size is 35.6 inches; the average neck size is 20.5 inches. Explain why the s.e. for bear C is higher than that for bear B, even though both the chest and neck measurements for bear C are closer to the average values.