

Homework 10 - handed out Friday, 13 November 2009

on campus Friday, 20 Nov 2009, in lecture (11 am)

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

off campus Monday, 23 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. **Model selection (practice):** The modelsel1.txt data set contains four predictor variables and $n = 50$ observations.

- (a) Find the best model using stepwise selection, using $\alpha_{entry} = \alpha_{remove} = 0.2$.
- (b) Do you end up with the same model using $\alpha_{entry} = \alpha_{remove} = 0.1$?
- (c) Use SAS to calculate C_p for all possible models. Identify the variables in the 3 models with the three smallest C_p values.
- (d) If you use AIC or BIC, do you select the same best model? The same best three models?

2. **Model selection, Data analysis:** The data in realestate.txt is set of residential house characteristics and sales prices from a few years ago. The variables are:

id, sales price (\$), house size (sq ft), # bedrooms, # bathrooms, AC (1 = yes), garage size (# cars), pool (1=yes), year built, quality (1 = high, 2 = medium, 3 = low), style (1 - 11 indicating architectural style), lot size (sq ft), highway (1=adjacent).

Further description of the data is in Appendix C of Kutner, but that information is not necessary to do the problem. (Remember, there is a copy of Kutner available in the Stat Dept. Office).

Ignore style and treat quality as a continuous variable (i.e. not 3 groups). Please construct a reasonable model to predict sales price (or some transformation of sales price) from some combination of size, bedrooms, bathrooms, ac, garage size, pool, year built, quality, lot size, and highway.

Note: It is very difficult to know where to stop in a problem like this. If you work efficiently (and SAS is generally cooperative), you should spend no more than an hour on this problem. I do want you to check for and fix obvious problems, but you don't have to fix every idiosyncrasy in the data. I am aware of at least one curious observation.

3. **ANCOVA** — side effect of a drug. The data in heart.txt are from a study of the effect of a new drug on a particular heart function. These drugs have been developed for another purpose, but one concern is whether they have a side effect on heart function. Drugs A and B are two forms of the drug, C is a placebo (i.e. a control, expected to have no effect on heart function). Thirty subjects were randomly assigned to a drug. The intent was to have 10 subjects per drug, but a mistake was made and drug B was given instead of drug C to one of the subjects. PRE is the heart function measured before the drug was administered. POST is the heart function 2 hours after the drug was administered.

- (a) Consider only the post drug data. Is there evidence of an effect of the drugs on heart function?
- (b) Estimate the post-drug means for each treatment, the mean difference between drug A and the placebo (C), and the s.e. of that difference.
- (c) Consider an ANCOVA, using PRE drug data as a covariate. Assume a linear regression with the same slope for all drugs. Is there evidence of an effect of the drugs on heart function? Report your test statistic and p-value.
- (d) Estimate the post-drug means for each treatment, after adjusting to the same pre-treatment value.
- (e) Estimate the mean difference between drug A and the placebo (C), for subjects with the same PRE heart function. Also report the s.e. of the adjusted difference.
- (f) One of the major assumptions of ANCOVA is that the slope, i.e. the relationship between PRE and the response is the same for all three drugs. Is this assumption reasonable here?

- (g) One of your office mates finds some of your results rather curious. Please explain (briefly) why the differences in b) and e) are not the same number. Also, explain why the se's are different.
4. **Two-factor ANOVA** — This data set is from an experiment on childrens's memory. A random sample of 36 fourth-graders from one city were used in the experiment. Two factors are varied: level of reinforcement (none or verbal) and time of isolation (20, 40, or 60 minutes). Students were told to memorize a paragraph and given positive verbal reinforcement or no reinforcement while learning it according to their treatment assignment. Then students were isolated for the specified amount of time. There were 6 students randomly assigned to each of the six treatment groups. The response is a score measuring the student's memory for the learned paragraph. The data are contained in the file "paragraph.txt" with the first column indicating the level of reinforcement (none or verbal), the second column indicating the isolation time (20, 40, 60), and the third column giving the observed memory score.
- Plot the observations, putting response on the Y axis and isolation time on the X axis. On this plot indicate the location of the mean response for each treatment combination. Connect the means for all treatments with no reinforcement. Repeat for all treatments with verbal reinforcement. Based on the graph what significant effects do you expect to find?
 - Obtain the analysis of variance table for the two-way factorial model. Which effects are significant at the .05 level?
 - Consider the analysis as a one-way ANOVA using 6 treatments. This is equivalent to fitting the cell means model. Construct the two-way factorial ANOVA table using contrasts among the 6 treatments.
 - Check for constant variance and outliers using the residuals. Do you see any concerns? Is the assumption of independence reasonable? Explain why or why not.
 - Since the interaction is significant, the researchers ask you to test the effects of reinforcement (none or verbal) at each isolation time. A test is sufficient. You don't need to compute estimates and standard errors.
 - Summarize your results. How do reinforcement and isolation time effect memory? Also, describe the population for which you think these results are relevant.

OPTIONAL problem: The following problem is optional. Working through it will explore different ways of expressing an ANOVA as a regression. If you're interested, work through all or parts of this problem. You don't need to submit your answers (and they won't be graded if you do). I will distribute my answers.

5. **ANOVA using regression** — The following problem is to demonstrate the points made in lecture or lecture notes: 1) that ANOVA is equivalent to regression on indicator variables, 2) that different sets of indicator variables give equivalent tests, and 3) some quantities are invariant to the choice of X matrix but others are not. For simplicity, we will use a 1-way ANOVA.

Note: I am **not** expecting algebraic derivations of any of these points. It is sufficient to compare numerical results from SAS or other computing package.

The file doughnut.txt contain data from an experiment on fat adsorbision by doughnuts. It compares four types of cooking fat in a completely randomized design. There are six replicates of each fat. The response is the amount of fat in the cooked doughnut. The question was whether the type of cooking fat mattered, i.e. did cooking in certain fats lead to doughnuts with more fat?

- Calculate the sums-of-squares for treatments and error using a 1-way ANOVA (e.g. proc glm). Estimate the following quantities and their standard errors:
 mean for fat 1
 mean difference between fat 1 and fat 4
- Doughnut2.txt contains a set of indicator variables labeled a1 - a3. The values of these indicator variables depend on the fat number. You plan on fitting the multiple regression equation:

$$Y_{ij} = \beta_0 + \beta_1 a1_i + \beta_2 a2_i + \beta_3 a3_i + \varepsilon_{ij}$$

How do the four cell means ($\mu_1, \mu_2, \mu_3,$ and μ_4) relate to the regression parameters ($\beta_0, \beta_1, \beta_2,$ and β_3)? (Your answer will be a set of equations of the form $\mu_1 = \dots$).

Relate each β to the cell means (Your answer will be a set of equations of the form $\beta_1 = \dots$).

- (c) Use proc reg or proc glm to fit the multiple regression. Report from the SAS output):
the sums-of-squares for treatments and error,
estimates and their standard errors of:

β_1 , i.e. the slope associated with the indicator variable for fat 1
treatment mean for fat 1, and the
difference in treatment means between fats 1 and 4.

Are these values the same as those from part 5a?

Note: The previous part relates treatment means to linear combinations of parameters. Remember that you can estimate linear combinations of parameters and their standard errors using the ESTIMATE statement in proc glm.

- (d) A factor effects model with sum-to-zero constraints is equivalent to a multiple regression with indicator variables (B1, B2, and B3) that have the following values:

Fat	B1	B2	B3
1	1	0	0
2	0	1	0
3	0	0	1
4	-1	-1	-1

These values are also in the doughnut2.txt file. Fit a multiple regression using the B set of indicator variables (include the default intercept, just as before). Calculate (or report from the SAS output):

the sums-of-squares for treatments and error,
estimates and standard errors of:

β_1 , the slope associated with the indicator variable for fat 1
treatment mean for fat 1, and the
difference in treatment means between fats 1 and 4.

Which values are the same as those from the previous X matrix (part 5c)? Which are different.