

Stat 500 - Midterm II - Off campus (17-21 November 2003)

Please put your name on the back of your answer book. Do NOT put it on the front. Thanks.

- The exam is closed book, closed notes. Use only the formula sheet and tables I provide today. You may use a calculator.
- Write your answers in your blue book. Ask if you need a second (or third) blue book.
- You have 2 hours (120 minutes) to complete the exam.
- Tables, a formula sheet, and scrap paper are provided.
- Points are indicated for each question. The total is 120.
- Important reminders:
 - budget your time. Some parts of each question should be easy; others may be hard. Make sure you do all parts you can.
 - notice that some parts do not require any computations.
 - show your work neatly so you can receive partial credit.
- Good luck!

1. 10 pts. Linear regression has been used to model network performance. You have collected data on the performance of a data server. The response, Y_i , is the length of time required to transfer a file. You expect this to vary linearly with the file size, X_{1i} . The intercept in that relationship is not 0, because of initialization costs. The slope of the relationship is expected to be linear function of the network load, X_{2i} . Construct a linear regression model that allows you to predict file transfer times for various file sizes and network loads.

2. 30 pts. Mercury is a naturally occurring contaminant in fish in freshwater lakes. In some lakes, the mercury concentration is sufficiently high that warnings are posted to limit the consumption of fish from that lake. The mercury concentration is influenced by many characteristics of lakes. The state of Maine collected information from a random sample of 110 lakes. The variables of interest included:
- hg: mercury concentration in fish.
 - elev: elevation of the lake
 - area: area of lake
 - drain: size of the drainage basin for the lake
 - runoff: annual runoff into lake

Here are the parameter estimates and standard errors when a multiple linear regression is fit to the data:

<u>parameter</u>	<u>estimate</u>	<u>s.e.</u>
β_0	0.758	0.144
β_{elev}	-0.211	0.061
β_{area}	-0.014	0.016
β_{drain}	0.0137	0.029
β_{runoff}	-0.292	0.262

The ANOVA table with sums of squares is:

<u>Source</u>	<u>SS</u>
Model	1.137
Error	7.515
Total	8.653

- (a) Fill in the degrees of freedom in the above ANOVA table.
- (b) The slope for runoff in a simple linear regression is -0.527, which is considerably different from the multiple regression slope. Explain, in principle, why the two coefficients are different.
- (c) Construct a test of $H_0: \beta_{elev} = 0, \beta_{area} = 0, \beta_{drain} = 0, \beta_{runoff} = 0$, if that is possible from the data provided. If not, indicate how you would construct the desired test.
- (d) Construct a test of $H_0: \beta_{drain} = 0, \beta_{runoff} = 0$, if that is possible from the data provided. If not, indicate how you would construct the desired test.
- (e) Construct a test of $H_0: \beta_{area} + \beta_{drain} = 0$, if that is possible from the data provided. If not, indicate how you would construct the desired test.

3. 40 pts. A common practice in horticulture is to graft stems of one type of tree onto root systems of another. If the graft is healthy, the combination tree is much more vigorous than either part alone. Unfortunately, grafts often get diseased. The age of the root systems is suspected to influence the probability of success.

The table below shows data from an experiment to evaluate graft success in root stocks of two different ages. There are 48 trees classified by graft status (healthy or diseased) and age of the rootstock.

<u>Status</u>	<u>Age of Tree</u>	
	<u>1 – 2</u>	<u>2 – 5</u>
Health	1	17
Disease	14	16

- (a) Test the hypothesis that $P[\text{success in 1-2 year old trees}] = P[\text{success in 2-5 year old trees}]$. Report the p-value and state your conclusion.
- (b) Give a 90% confidence interval for the difference in probability of graft success between the two age groups.
- (c) What is the odds of a healthy graft in the 2-5 year old root stocks?
What is the odds ratio comparing $P[\text{success}]$ in the 2-5 year old root stocks to $P[\text{success}]$ in the 1-2 year old root stocks?
You only need to report point estimates here.
- (d) Does the 90% confidence interval for the odds ratio include 1? Explain why or why not. You do not need to report the confidence interval.

The data above were part of a larger experiment with 4 age classes. Also, the experiment started with 40 trees per age group. Some trees died before the success was measured. The contingency table for the complete experiment is:

<u>Status</u>	<u>Age of Tree</u>			
	<u>1 – 2</u>	<u>2 – 5</u>	<u>5 – 7</u>	<u>7 – 10</u>
Health	1	17	15	3
Disease	14	16	8	7
Dead	25	7	17	30

- (e) A common test for data of this type is the chi-squared test. State the null and alternative hypotheses for this test.
- (f) The Chi-square statistic for this table is 43.57. How many degrees of freedom does it have?
- (g) Do the counts of 1 and 3 worry you? Explain why or why not.
- (h) After completing your analysis, the experiments tell you that the experiment was actually conducted at 4 sites, with 10 trees of each age at each site. Some of the sites are much more stressful than others, so $P[\text{Dead}]$ varies between sites. Is the usual Chi-square test still appropriate? Explain why or why not.

4. 40 pts. Pharmaceuticals usually have a shelf life. One reason is that the concentration of active ingredient decreases over time. One common model for this is the exponential decay model

$$Y_i = C_0 e^{-\lambda X_i} \times \epsilon_i,$$

where Y_i is the concentration at time i , C_0 is the initial concentration (at time of manufacture), X_i is the time since manufacture, and λ is the decay rate.

Data were collected from a randomized experiment. Twenty packages of a particular drug were randomly assigned to be sampled at one of five times, 0 months, 2 months, 4 months, 8 months, and 12 months after manufacture. 4 packages were assigned to each sampling time. Each package was measured once and only once, at the assigned sampling time. The data are shown below.

- (a) Rewrite this model in a form in which the parameters (C_0 and λ can be estimated using linear regression. If your model involves parameters other than C_0 and λ , please define your new parameters.
- (b) Four different models were fit to the data. Some unimportant details of each model are left out. The Error sums of squares for each are:

Model	# parameters	Error SS
$\cdot = \beta_0$	1	53.48
$\cdot = \beta_0 + \lambda X_i$	2	2.95
$\cdot = \beta_0 + \lambda X_i + \beta_2 X_i^2$	3	2.50
$\cdot = \mu + \tau_i$	5	0.861

Please test whether $\lambda = 0$. Report the test statistic, the appropriate d.f. and an approximate p-value.

- (c) Construct the most appropriate test of lack of fit of the linear regression, using one or more of the sums of squares in part 4b. Report the test statistic, its d.f., an approximate p-value and a short conclusion.
- (d) A histogram of the Y values used in the above regressions (see figure above) does not appear to be described by a normal distribution. Is this a concern? Explain why or why not.

- (e) Various diagnostic plots are included below. List the usual assumptions of linear regression and evaluate each one using these diagnostic plots and the description of the experiment. Explain (briefly) what you are looking at to make each evaluation.

- (f) The drug company has a large batch of the drug (many packages) that has been in storage for four months. They will use the regression to predict the mean concentration of drug in this batch. They also want to report the precision of that estimate. There are three possible formulae that could be used:

- 1) $\sqrt{MSE \left(\frac{1}{n} + \frac{(4-\bar{x})^2}{\sum (x_i - \bar{x})^2} \right)}$
- 2) $\sqrt{MSE \left(1 + \frac{1}{n} + \frac{(4-\bar{x})^2}{\sum (x_i - \bar{x})^2} \right)}$
- 3) $\sqrt{MSE/4}$

Which is the most appropriate formula to calculate standard error? Explain (briefly) why your choice is the most appropriate.