MS WRITTEN EXAMINATION IN BIOSTATISTICS, PART II

Monday, August 16, 2010: 9:00 AM - 3:00PM

Room: MH 0001, Blue Cross/Blue Shield Auditorium

INSTRUCTIONS:

a. This is a **OPEN BOOK** examination.

b. Answer 3 out of 4 questions.

c. Put the answers to different questions on separate sets of paper; staple them separately.

d. Put your code letter, **not your name**, on each page.

e. Return the examination with a signed statement of the honor pledge on a page separate from your answers.

f. You are required to answer only what is asked in the questions and not to tell all you know about the topics.

1. Let

$$y_1 = \beta_0 + \beta_1 + \epsilon_1,$$
$$y_2 = \beta_0 + 2\beta_1 - \beta_2 + \epsilon_2,$$
$$y_3 = \beta_0 + 3\beta_1 - 2\beta_2 + \epsilon_3,$$
$$y_4 = \beta_0 - \beta_1 + 2\beta_2 + \epsilon_4.$$

Further, let $\mathbf{y} = (y_1, y_2, y_3, y_4)^T$, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)^T$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$. Assume that the ϵ_i's are i.i.d. random variables distributed as normal with mean 0 and unknown variance σ^2.

(a) Write the model in the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and identify the matrix $\mathbf{X}$.

(b) What is the rank of $\mathbf{X}$?

(c) Re-express the model as $\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$ where $\mathbf{X}^*$ is full rank.

(d) Is $\theta_1 = \beta_1 - \beta_2$ estimable? Explain. If so, do the following. Write the least-squares estimator $\hat{\theta}_1$ explicitly as a function of (y_1, y_2, y_3, y_4). Find the variance of $\hat{\theta}_1$. Suppose $\mathbf{y} = (2, 0, -1, 4)$ was observed. Test the hypothesis $H_0 : \theta_1 = 0$ against $H_1 : \theta_1 \neq 0$ at the 0.05 level, and construct a 95% confidence interval for θ_1.

(e) Repeat Part (d) for $\theta_2 = \beta_0 + 2\beta_1$ and $\theta_3 = 2\beta_0 + \beta_1 + \beta_2$.

(f) Compute a 95% prediction interval for a future observation, $y_5 = \beta_0 + \beta_1 + \epsilon_5$, where ϵ_5 is normal$(0, \sigma^2)$ and is independent of $\boldsymbol{\epsilon}$.

Points: (a) 3 , (b) 2 , (c) 5 , (d) 5 , (e) 5 , (f) 5.

2. A randomized clinical trial was conducted among 20 advanced lung cancer patients to compare the effects of two chemotherapy treatments in prolonging time until death. All patients were randomly assigned to one of two treatments, termed "standard" or "test." The time in weeks from randomization until death for each patient is given in the table below. The "*" denotes time on study for individuals who were alive when the trial concluded.

Standard	Test
3, 3, 3, 4, 4,	1, 1, 2*, 2, 10,
4, 4, 4, 10*, 11*	10*, 10*, 10*,11*, 12*

   (a) Calculate nonparametric estimates of the probability of death by week 8 for each treatment. Calculate corresponding 95% confidence intervals.

   (b) Test whether there is a difference in the probability of death by week 8 between treatment arms.

   (c) Test whether the rate of death is different between the two treatment arms over the course of the trial.

   (d) Interpret and discuss the results from parts (a) - (c).

   Points: (a) 7 , (b) 6 , (c) 7 , (d) 5.

3. Low-density lipoprotein (LDL) is one type of lipoproteins. High levels of LDL increase the risk of cardiovascular disease. A drug has been developed to reduce LDL level. Preliminary studies showed that the drug might have different effects in men and women. In a study to evaluate this hypothesis, 200 men and 200 women were recruited and among them, 100 men and 100 women were randomly selected to take the drug and the other 100 men and 100 women took placebo. At the end of the study, LDL levels (mg/dL) were recorded, along with three other covariates: height, weight, and age.

   (a) Consider a model with covariates age, treatment, gender and the interaction between treatment and gender:

$$y_i = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{treatment} + \beta_3 \texttt{gender} + \beta_4 \texttt{treatment} \times \texttt{gender} + e_i, \qquad (1)$$

   where y_i is the LDL level of patient i at the end of the study, and e_i is a random error with mean zero. Use reference cell coding so that treatment = 0 and 1 for placebo and drug, respectively, and gender = 0 and 1 for women and men, respectively. Write the expected value of y_i as a function of `age` and coefficients β_j (j=0, 1, ..., 4) for each combination of `treatment` and `gender`.

Drug	gender	Expected Value of y_i
placebo	women	
placebo	men	
drug	women	
drug	men	

   (b) (i) If we write model (1) in matrix form: $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, what are the dimensions of $\mathbf{y}$, $\mathbf{X}$, $\mathbf{b}$, and $\mathbf{e}$? (ii) Write down the least squares estimate of $\mathbf{b}$ (denoted by $\hat{\mathbf{b}}$) in terms of $\mathbf{y}$ and $\mathbf{X}$. (iii) Assume $\mathbf{e}$ is distributed as $N(\mathbf{0}, \sigma^2\mathbf{I})$, where $\mathbf{0}$ is a vector of 0s, and $\mathbf{I}$ is an identity matrix. What is the distribution of $\hat{\mathbf{b}}$?

   (c) For model (1), fill out the following ANOVA table.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	---	-----	-----	146.7	<.0001
Error	---	-----	113		
Corrected Total	---	-----			

(d) Now add weight and height to the model,

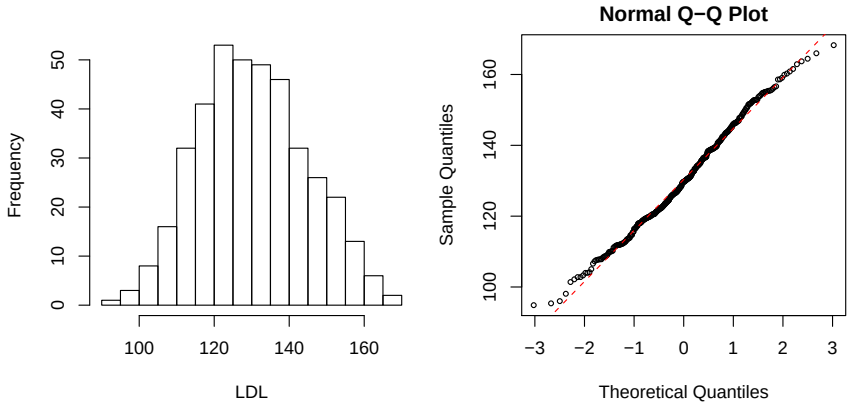
$$y_i = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{treatment} + \beta_3 \texttt{gender} + \beta_4 \texttt{treatment} \times \texttt{gender} + \beta_5 \texttt{weight} + \beta_6 \texttt{height} + e_i. \quad (2)$$

An ANOVA test comparing models (1) and (2) gives a p-value of 0.0098, indicating that weight and/or height are important. However, as shown in the following table, t-tests for weight and height effects have insignificant p-values. Does this contradict the significant result of the ANOVA test comparing model (1) and (2), and why?

Parameter	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	83.62710	24.46134	3.419	0.000695	***
age	0.28858	0.04549	6.344	6.15e-10	***
drug	-10.01539	1.49280	-6.709	6.85e-11	***
gender	20.27783	1.87082	10.839	< 2e-16	***
weight	0.13960	0.07169	1.947	0.052220	.
height	0.09731	0.17098	0.569	0.569607	
drug:gender	-7.31608	2.11999	-3.451	0.000619	***

(e) Answer the folllowing questions based on the table in part (d):
(i) Is the drug effect significantly different between males and females? (ii) Provide the $\mathbf{C}$ and $\boldsymbol{\theta}_0$ matrices necessary for stating the hypothesis that "drug has no effect in both males and females" in the form of $\mathbf{C}\boldsymbol{\beta} = \boldsymbol{\theta}_0$, where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6)^T$. (iii) Write down the linear model for LDL separately for males and for females, with intercept, age, drug, weight, and height as covariates.

(f) As illustrated in the histogram and QQ-plot below, the distribution of the observed LDL in the 400 individuals appears to deviate from the normal distribution. Does this necessarily violate any assumption of linear regression? Justify your answer whether it is yes or no. If yes, explain how to fix this problem.



Points: (a) 4, (b) 4, (c) 4, (d) 3, (e) 6, (f) 4.

4. Air pollution can play a role in triggering asthma symptoms. Exposure to air pollution may occur indoors or outdoors. One aspect of this pollution is the concentration in the air of particulate matter of various sizes. PM_{10} refers to particles with diameter less than 10 micrometers. A study was done in a city to investigate the relationship between the concentration of PM_{10} (in $\mu g/m^3$) and asthma symptoms in children. One symptom of asthma is wheezing. Information on wheezing was collected by asking parents "Has your child had wheezing or whistling in the chest in the last 4 weeks?"

   The investigators were interested in comparing indoor and outdoor air pollution and randomly selected 25 homes in the city for air quality monitoring. At each of these 25 homes, air quality monitors were placed inside and outside the home and data on PM_{10} concentration were collected. Because PM_{10} concentration data are usually skewed, it is common to use logarithms of the concentration in statistical analyses. For home k, let X_{Ik} and X_{Ok} denote the natural logarithms of the indoor and outdoor concentrations of PM_{10}, respectively.

   Suppose:

$$\sum_{k=1}^{25} X_{Ik} = 95, \quad \sum_{k=1}^{25} X_{Ik}^2 = 380$$

$$\sum_{k=1}^{25} X_{Ok} = 110, \quad \sum_{k=1}^{25} X_{Ok}^2 = 520$$

$$\sum_{k=1}^{25} X_{Ik} X_{Ok} = 425.$$

   Note: All your work, assumptions, derivations and computations should be shown.

   (a) Provide an estimate for the association between (log) indoor and outdoor PM_{10} concentration and test whether the association is statistically significant.

   (b) Test whether the mean (log of) outdoor concentration of PM_{10} differs significantly from that indoors and give a 95% confidence interval for the difference in the means.

   (c) What measure of central tendency on the original (that is, unlogged) scale of PM_{10} concentration can be estimated from the information given above?

   (d) In measuring aspects of air pollution, such as PM_{10}, there are usually some locations or time periods for which the level of the pollutant is below the limit of sensitivity of the measuring instrument. In such circumstances the instrument may report the value as missing. For the example above there did not happen to be any missing data. Suppose, however, that for 3 of the homes the indoor measurement was missing.

   (i) Explain why it would not be appropriate to just omit the missing data?

(ii) What changes could be made to the data or the analysis to take into account the missing data? (Do not conduct an analysis, just describe and justify the changes that could be made.)

(e) To investigate the association between increasing indoor PM_{10} concentration and wheezing, the investigators obtained data on 400 seven-year-old children and measured the PM_{10} concentration in their homes. The investigators divided the indoor PM_{10} concentration data into quartiles and counted how many children in each quartile of exposure had had wheezing in the past 4 weeks. The results are given in the following table:

	Quartile			
Wheezing	1	2	3	4
No	91	88	82	84
Yes	9	12	18	16
Total	100	100	100	100

Test whether the proportion of children wheezing increases with increasing indoor PM_{10} concentration.

(f) In looking at the data in the table, the investigators postulated that beyond a certain threshold of indoor PM_{10} concentration, all children who are sensitive to particulate matter will have wheezing symptoms, whereas other children will not have symptoms regardless of the concentration. The investigators postulated that the threshold may be at about 50 $\mu g/m^3$. The table below classifies the 400 children according to whether the indoor PM_{10} concentration in their homes was above or below this concentration.

	Wheezing		
PM_{10} conc.	No	Yes	
$\leq 50\mu g/m^3$	259	29	288
$> 50\mu g/m^3$	86	26	112
	345	55	400

Estimate the risk of wheezing among children exposed to $> 50\mu g/m^3$ indoor PM_{10} relative to that in children exposed to $\leq 50\mu g/m^3$ indoor PM_{10} and provide a confidence interval for the estimate. Based on this, is exposure to $> 50\mu g/m^3$ significantly associated with wheezing?

Points: (a) 4, (b) 7, (c) 2, (d) 4, (e) 5, (f) 3.