Prelim Exam, Spring 1998 Stat 511

NOTE: Problems 1, 2 and 3 are unrelated problems!

PROBLEM 1. For unit i, $i = 1, ..., n$, let Y_i denote the dependent variable. Consider the model

$$Y_i = \mu + \sum_{j=1}^{a} \delta_i^j \alpha_j + \theta i + \epsilon_i, \tag{1}$$

where μ and the α_j's are unknown fixed parameters, a is the number of levels for factor A, θ is an unknown fixed parameter (θi is the product of θ and the known value of i), the ϵ_i's are i.i.d. random variables with mean 0 and variance σ^2 (an unknown positive constant), and the δ_i^j's are known indicator variables defined as

$$\delta_i^j = \begin{cases} 1 & \text{if } j \text{ is the level of factor } A \text{ for the observation from unit } i \\ 0 & \text{otherwise.} \end{cases}$$

(Note that except for the addition of the term θi, the model is precisely the usual model, albeit with a somewhat unusual notation, for a one-way classification.)

Let n_j denote the number of observations with level j of factor A, and assume that $n_j > 0$ for all j; thus $n_j = \sum_{i=1}^{n} \delta_i^j$ and $n = \sum_{j=1}^{a} n_j$.

(i). If $n > a$ and $d_1, ..., d_a$ are constants with $\sum_{j=1}^{a} d_j = 0$, show that $\sum_{j=1}^{a} d_j \alpha_j$ is estimable.

(ii). Let $n > a$. Suppose that for *any* constants $d_1, ..., d_a$ with $\sum_{j=1}^{a} d_j = 0$ it holds that the ordinary least squares estimator of $\sum_{j=1}^{a} d_j \alpha_j$ is the same for the model in (1) as for the model that ignores the term θi, i.e. for $Y_i = \mu + \sum_{j=1}^{a} \delta_i^j \alpha_j + \epsilon_i$. Show that this implies that

$$\sum_{i=1}^{n} \delta_i^1 i / n_1 = ... = \sum_{i=1}^{n} \delta_i^a i / n_a \ .$$

(iii). Let $n > a$, and suppose that the ϵ_i's are i.i.d. $N(0, \sigma^2)$. Define $D_j = \sum_{i=1}^{n} \delta_i^j i / n_j$, and consider the quadratic form

$$Q = \left(\sum_{j=1}^{a} \sum_{i=1}^{n} \delta_i^j (i - D_j) Y_i \right)^2 .$$

Find positive constants c_1, c_2 and c_3 such that $c_1 Q$ has a χ^2-distribution with c_2 degrees of freedom and non-centrality parameter c_3. Do not leave any matrices in your answers for c_1, c_2 and c_3.

PROBLEM 2. Consider the following model with interaction for a two-way crossed classification in which both factors have 3 levels:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \ i, j = 1, 2, 3, \ k = 1, ..., n_{ij}.$$

The α_i's and μ are fixed effects, while the β_j's, γ_{ij}'s and ϵ_{ijk}'s are independently distributed random variables. All of these random variables have mean 0, while their variances are σ_δ^2 (for

the β_j's), σ_c^2 (for the γ_{ij}'s) and σ_ϵ^2 (for the ϵ_{ijk}'s) — all σ^2's being unknown positive constants. The values of the n_{ij}'s are given in the following table:

		B		
		1	2	3
A	1	3	2	2
	2	2	3	2
	3	2	2	3

(i). The sums of squares in an ANOVA table could be $R(\alpha_i \mid \mu)$, $R(\beta_j \mid \alpha_i, \mu)$, $R(\gamma_{ij} \mid \beta_j, \alpha_i, \mu)$, and the sum of squares for error ($SS\ Error$). For each of these sums of squares, provide the corresponding degrees of freedom. Explain your answers.

(ii). Can we conclude that, irrespective of the values of the various σ^2's, $\overline{Y}_{1..} - \overline{Y}_{2..}$ is the best linear unbiased estimator of $\alpha_1 - \alpha_2$? Explain your answer. (By $\overline{Y}_{i..}$ we denote of course the average of all Y's corresponding to observations with level i of factor A.)

PROBLEM 3. For unit i, $i = 1, ..., n$, let Y_i denote the dependent variable and x_i the known value of an independent variable. For a known value t, let z_i be defined as

$$z_i = \begin{cases} 0 & \text{if } x_i \leq t \\ 1 & \text{otherwise.} \end{cases}$$

Assume that the Y_i's are uncorrelated with common variance σ^2, an unknown positive constant. Consider the (unrestricted) linear model

$$E(Y_i) = \beta_0 + \beta_1 z_i + \beta_2 x_i + \beta_3 x_i z_i, \tag{2}$$

and the restricted linear model

$$E(Y_i) = \gamma_0 + \gamma_1 z_i + \gamma_2 x_i + \gamma_3 x_i z_i, \text{ where the parameters satisfy the restriction } \gamma_1 + \gamma_3 t = 0. \tag{3}$$

You can assume that all β's and γ's are estimable with the available data.

(i). For each of the two models, sketch in as much detail as possible what relationship the model proposes between the variables Y and x. (A plot of Y versus x, showing the general shape of the fitted model, could form an important part of your answer.)

(ii). For each of the following two statements, decide whether the statement is correct. Explain your answers.

Statement 1: For model (2), the ordinary least squares estimator of β_2 depends only on those Y_i's for which $x_i \leq t$

Statement 2: For model (3), the restricted least squares estimator of γ_2 depends only on those Y_i's for which $x_i \leq t$.

STAT 511 SOLUTIONS

1. (i) $\sum_j d_j \alpha_j$ is estimable for all d_j's with $\sum_j d_j = 0$ if and only if the rank of the model matrix is $a+1$ (or equivalently, if and only if the column corresponding to θ is not a linear combination of the columns for the α_j's). But it is obvious that $(1, 2, \ldots, n)^T$, a vector with $n > a$ distinct elements, cannot be written as a linear combination of the columns corresponding to the α_i's, which will at most have a distinct entries.

(ii) Let $\bar{Y}_{(j)} = \sum_i \delta_i^j Y_i / n_j$, i.e. the average of all observations with level j of A. If θ_i is ignored, then the OLS estimator of $\sum_j d_j \alpha_j$ is equal to $\sum_j d_j \bar{Y}_{(j)}$. This is also the OLS estimator of $\sum_j d_j \alpha_j$ for model (1) if and only if it is still an unbiased estimator of $\sum_j d_j \alpha_j$ under that model. But, for model (1)

$$E\left(\sum_j d_j \bar{Y}_{(j)}\right) = \sum_j d_j \alpha_j + \sum_j d_j \sum_i \delta_i^j \theta_i / n_j$$

$$= \sum_j d_j \alpha_j + \theta (d_1, \ldots, d_a) \left(\sum_i \delta_i^1 i / n_1, \ldots, \sum_i \delta_i^a i / n_a \right)^T.$$

As this needs to be $\sum_j d_j \alpha_j$ for all sets of d_j's with $\sum_j d_j = 0$, it follows that

$$\sum_i \delta_i^1 i / n_1 = \cdots = \sum_i \delta_i^a i / n_a.$$

(iii) With $\underline{Y} = (Y_1, \ldots, Y_n)^T$, $X = [X_1 \; \underline{t}]$, where X_1 is that part of the model matrix corresponding to $\mu, \alpha_1, \ldots, \alpha_a$ and $\underline{t} = (1, \ldots, n)^T$, it is easily seen that

$$\underline{t}^T (I - P_{X_1}) \underline{Y} = \sum_{i=1}^{n} i \left(Y_i - \sum_j \delta_i^j \bar{Y}_{(j)} \right) =$$

$$\sum_{j=1}^{a} \sum_{i=1}^{n} \delta_i^j i Y_i - \sum_{j=1}^{a} \left(\sum_{\ell=1}^{n} \delta_\ell^j Y_\ell / n_j \right) \sum_{i=1}^{n} i \delta_i^j$$

$$= \sum_{j=1}^{a} \sum_{i=1}^{n} \delta_i^j i Y_i - \sum_{j=1}^{a} \sum_{\ell=1}^{n} \delta_\ell^j Y_\ell D_j$$

$$= \sum_{j=1}^{a} \sum_{i=1}^{n} \delta_i^j (i - D_j) Y_i$$

Hence,

$$\frac{1}{\sigma^2} \underline{Y}^T (I - P_{X_1}) \underline{t} \left(\underline{t}^T (I - P_{X_1}) \underline{t} \right)^{-1} \underline{t}^T (I - P_{X_1}) \underline{Y}$$

$$= \frac{1}{\sigma^2} \frac{Q}{\sum_i (i - \sum_j \delta_i^j D_j)^2} \sim \chi_1^2(\delta^2)$$

where $\delta^2 = \frac{\theta^2}{2\sigma^2} \underline{t}^T (I - P_{X_1}) \underline{t} = \frac{\theta^2}{2\sigma^2} \sum_i \left(i - \sum_j \delta_i^j D_j \right)^2$

The values for c_1, c_2 and c_3 are now clear.

(A student who does not recognize the relationship between Q and $\underline{t}^T (I - P_{X_1}) \underline{Y}$ can of course simply start

with the distribution of $\sum_j \sum_i \delta_{ij} (i - D_j) Y_{ij}$

2. (i) 2, 2, 4 and 12

(ii) No. Note that

$$\mathrm{Var}(\bar{Y}_{1..} - \bar{Y}_{2..}) = \mathrm{Var}\left(\tfrac{1}{7}\beta_1 - \tfrac{1}{7}\beta_2 + \ldots\right)$$

(no β_j's in here)

$$= 2\sigma_b^2/49 + \ldots \qquad (1)$$

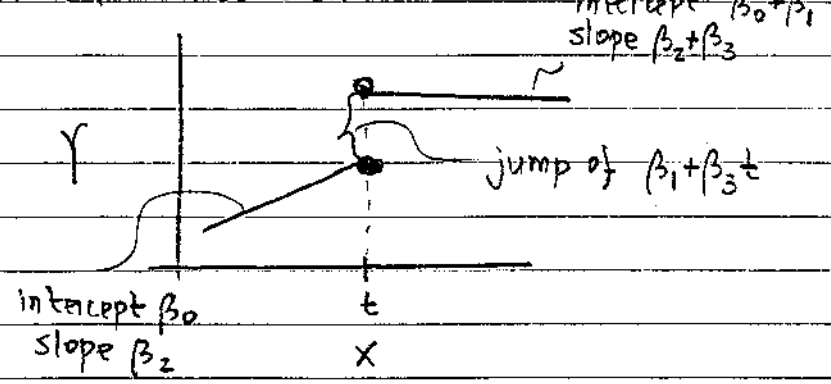
(… depends only on σ_c^2 and σ_e^2)

On the other hand, $E(\bar{Y}_{11.} - \bar{Y}_{21.}) = \alpha_1 - \alpha_2$ and

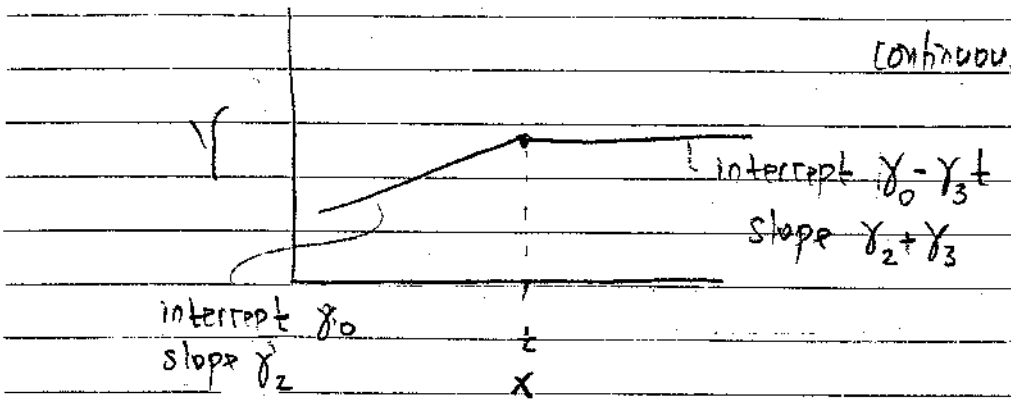
$$\mathrm{Var}(\bar{Y}_{11.} - \bar{Y}_{21.}) = \mathrm{Var}(\gamma_{11} - \gamma_{21} + \bar{\varepsilon}_{11.} - \bar{\varepsilon}_{21.}) \qquad (2)$$

which does not depend on σ_b^2. Hence, for large enough σ_b^2, (1) will exceed (2).

3. (i) Unrestricted model:



Restricted model:



(ii) Statement 1 is correct, statement 2 is not.
For statement 1, it is clear that $(\underline{Y} - X\underline{\beta})^T(\underline{Y} - X\underline{\beta})$ can be split into two parts (one corresponding to x-values less than or equal to t, one to x-values exceeding t) and that each of these two parts can be minimized seperately.

The continuity requirement for the second model suggests that γ_0 and γ_2 should under the restricted model also depend on data with x-values exceeding t. There are various ways to show this more rigorously. The easiest is perhaps by a simple example.

Let $n=4$, $t=0$, $x_1=-3, x_2=-1, x_3=1, x_4=3$.
The restricted least squares estimator of γ_2 is then

$$\frac{1}{10}(-4Y_1 + 2Y_2 + 3Y_3 - Y_4)$$

In a telephone survey in the Mississippi Delta, respondents were asked pairs of questions about their recent behavior toward their spouse and their spouses' behavior toward them. Two of the questions were

Q1: In the past month, how often did you get angry at your spouse?
Q2: In the past month, how often did your spouse get angry at you?

Responses were on a scale from never (1) to often (4), so higher scores imply greater anger. Survey methodologists were concerned that responses to these two questions might be affected by the order in which they are asked. To examine this possibility, two forms of the interview schedule were randomly assigned to respondents. Half of the respondents were randomly assigned to Form A in which question Q1 was asked first and question Q2 was asked second (Q1/Q2). The other half were first asked Q2 and then Q1. The sample means (with sample standard deviations and correlations) for female respondents are below. More women responded to Form B (n=170) than to Form A (n=158).

	Form A (Q1/Q2) (n=158)	Form B (Q2/Q1) (n=170)
Q1:	2.367 (0.912)	2.294 (0.881)
Q2:	2.217 (0.887)	2.029 (0.810)
Correlation:	0.70	0.65

Complete the following questions. You may make normal theory and homogeneity of variance assumptions where necessary.

1. Plot the sample means, putting the order in which questions are asked on the horizontal axis.

2. Using Form A only:

   a) Compute the sample variance of the paired difference.

   b) Is there a significant difference between the mean responses to the two questions? Let α=0.05.

3. Consider the two questions only when they are asked first. Is there evidence that there is a significant difference between the average response to Q1 when it is asked first and Q2 when it is asked first? Let α=0.05.

4. Consider next the two questions only when they are answered second. Is there evidence that there is a significant difference between the average response to Q1 when it is asked second and Q2 when it is asked second? Let α=0.05.

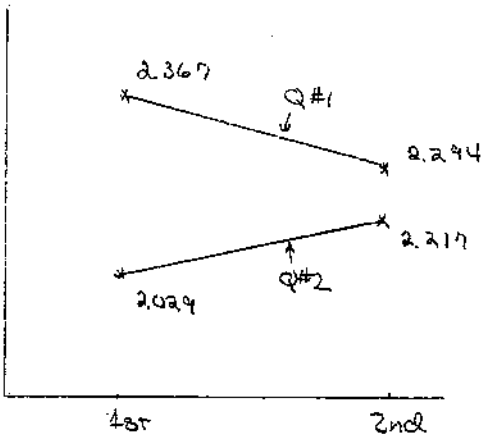
5. Methodologists who study order effects in questionnaires refer to the pattern in your figure as "even handedness." That is, respondents appear to give more extreme answers to the first question they are asked, and then adjust their response to the second question so that it is closer to their response to the first. These methodologists would like a t-test to see whether the difference observed in part 3 of this question is significantly different from the difference observed in part 4. Propose a t-test to compare this "difference in differences" and then use that t-test to test the null hypothesis that there is no difference in the differences. Let α=0.05.

6. The data above are for women only. In addition, we have responses from an independent sample of 189 men, of whom 95 responded to Form A while the remaining 94 answered Form B. Describe the analyses you would perform to make inferences about the effects of gender and question order by setting up an appropriate analysis of variance table.

7. How would the analysis of variance table in part 6 change if you had a sample of husbands and wives (n=158 assigned to Form A and n=170 assigned to Form B) instead of independent samples of men and women?

Methods Question #1
Spring 1998 Statistics Prelim Exam
answers

1. Plot means, with question order on the horizontal axis.



2. Paired t-test

$$t_{n-1} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{s_1^2 + s_2^2 - 2r_{12}s_1 s_2}{n-1}}} = 2.20$$

and if pooled by assuming homogeneity of variance

$$t_{n-1} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{2s_p^2(1-r_{12})}{n-1}}} \quad \text{where } s_p^2 = 0.809$$

3. Assuming homogeneity of variance

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{where } s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

$$t = \frac{2.367 - 2.029}{\sqrt{0.791\left(\frac{1}{159} + \frac{1}{170}\right)}} = \frac{0.338}{0.095} = 3.55$$

4. $$t = \frac{2.294 - 2.217}{\sqrt{.181\left(\frac{1}{170} + \frac{1}{158}\right)}} = \frac{0.077}{.098} = 0.788$$

5. Using the subscripts (1,1) for question #1 asked first, (1,2) for Q#1 asked second, (2,1) for Q#2 asked first, and (2,2) for Q#2 asked second, the pairs of means are $(\bar{Y}_{11}, \bar{Y}_{22})$ and $(\bar{Y}_{12}, \bar{Y}_{21})$. The null hypothesis can be expressed as

$$H_0: (\mu_{11} - \mu_{21}) = (\mu_{22} - \mu_{12}) \text{ or}$$

$$H_0: (\mu_{11} - \mu_{22}) = (\mu_{21} - \mu_{12})$$

and (before pooling) the t-test has the form:

$$t = \frac{-(\bar{Y}_{11} - \bar{Y}_{21}) - (\bar{Y}_{22} - \bar{Y}_{12})}{\sqrt{\frac{S_{11}^2 + S_{22}^2 - 2r_A S_{11} S_{22}}{n_A} + \frac{S_{21}^2 + S_{12}^2 - 2r_B S_{21} S_{12}}{n_B}}}$$

and by assuming homogeneity of variance (as in parts 2, 3 and 4):

$$t = \frac{(\bar{Y}_{11} - \bar{Y}_{21}) - (\bar{Y}_{22} - \bar{Y}_{12})}{\sqrt{S_p^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \text{ where}$$

$$S_p^2 = \frac{(n_A - 1)(S_A^2) + (n_B - 1)(S_B^2)}{n_A + n_B - 2} \text{ where } S_A^2 = \frac{2S^2(1 - r_A)}{n_A - 1}$$

as in part 2.

METHODS I - ANSWERS - PAGE 2

6. The t-test (squared) is equivalent to the F-test for Question * Order interaction. This can be expanded to appear as:

Source	df	SS	MS	F
Between				
Sex	1			
Order	1			
Sex * Order	1			
Between Sbj Error				
With in Subject				
Question	1			
Q * Sex	1			
Q * Order	1			
Q * Sex * Order	1			
W/in Subj Error				

7. If the data are paired, then sex would be w/in subjects rather than between, and the model would be re-written accordingly.

Ph.D. Prelim
Spring 1998

Methods Question 2

In a study of factors that may be associated with the amount of time patients stay in hospitals, a large insurance company collected information on the following list of variables from each hospital in a simple random sample of 112 hospitals.

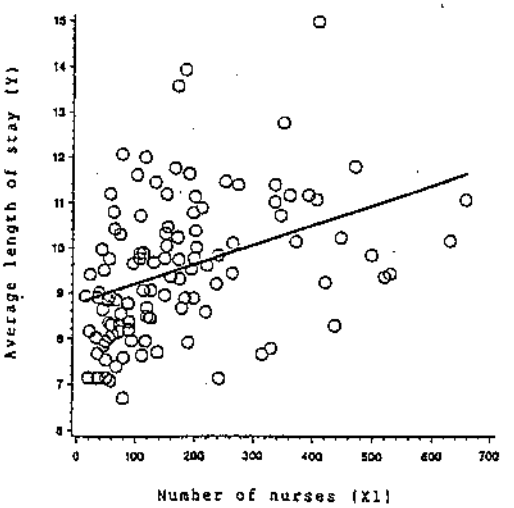
Y	average length of stay in the hospital in 1995 (in days)
X_1	number of nurses working in the hospital during an average day
X_2	number of beds in the hospital
X_3	average age of the patients served by the hospital
X_4	average number of patients served per day
X_5	number of services provided by the hospital out of a list of 35 services
X_6	percentage of patients who contracted an infection during their stay in the hospital

Theses values were recorded for each hospital based on all patients served in 1995. Variables X_1 through X_5 provide some information about the size of the hospital and the average age of the patients. The average length of stay may be longer in some larger hospitals, for example, because those hospitals deal with a higher proportion of difficult cases. Furthermore, older patients tend to have longer hospital stays than younger patients. The infection rate denoted by X_6 is a factor that hospitals may be able to reduce.

A. Least squares regression of Y on X_1 yields the estimated model

$$\hat{Y} = \underset{(0.22)}{8.78} + \underset{(0.00098)}{0.00436X_1}$$

Standard errors for the estimated coefficients are shown in parentheses below the corresponding estimates. This line is shown on the plot at the right along with the observations. Describe what you would conclude about the statistical and practical significance of this result. In particular, would you tell the insurance company executives that decreasing the number of nurses working at a hospital will decrease the average length of stay?



B. To account for possible effects of the size of the hospital on the average length of stay, the number of beds in the hospital was added to the regression model. Least squares estimation yields

$$\hat{Y} = \underset{(0.220)}{8.652} - \underset{(0.002389)}{0.000614X_1} + \underset{(0.001725)}{0.003925X_2}$$

Describe the conclusions you would reach from this result.

C. Describe a graph or a set of graphs that you think would be most useful in explaining to an insurance company executive why the estimated coefficient for X_1 is not the same for the regression models in parts A and B.

D. Least squares estimates for the coefficients in the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \varepsilon_i$$

are shown below, along with corresponding standard errors, t-tests, and p-values. For this model, $R^2 = .454$ and the error mean square is 1.3988 .

Variable	Estimated Coefficient	Standard Error	t-test	p-value
intercept	3.8809	1.4099	2.95	.007
X_1	-0.0034	0.0021	-1.67	.097
X_2	-0.0063	0.0032	-1.99	.049
X_3	0.0551	0.0249	2.21	.029
X_4	0.0139	0.0038	3.68	.0004
X_5	0.0012	0.0125	0.09	.924
X_6	0.5099	0.0927	5.50	.0001

Using this model, show how to construct a 95% confidence interval for the expected reduction in the average length of stay in a hospital that would be achieved by a 2 percent reduction in the infection rate (X_6), that is, the infection rate in the hospital is reduced from X_6 to $(0.98)X_6$, while the values of the other explanatory variables remain unchanged.

E. With respect to the estimated model that is used in part D, comment on the information that would be provided by each of the following diagnostic procedures:

(i) The studentized residuals plotted against the predicted values for the least squares estimate of the model in part D.

(ii) Examination of the eigenvalues of $\mathbf{X}^T\mathbf{X}$, where $\mathbf{X}$ is the 112x7 model matrix for the model in part D, and $\mathbf{X}^T$ is the transpose of $\mathbf{X}$.

(iii) The residuals from the regression of Y on X_1, X_2, X_3, X_4, and X_5 plotted against the residuals from the regression of X_6 on X_1, X_2, X_3, X_4, and X_5.

F. To check if the coefficients in the regression model in part D are consistent across regions of the United States, the 112 hospitals in this study were grouped into four regions:

Region	Number of Hospitals
Northeast	32
North Central	35
South	23
West	22

A separate regression model was fitted to data from each region using least squares estimation. Explain how you would test the null hypothesis that regression coefficients are the same for all four regions. Give a formula for your test statistic and explain how to use it.

G. A program designed to reduce the occurrence of infections was implemented in eight of the hospitals at the beginning of 1996. During 1996, the reduction in infection rates achieved by these hospitals ranged from 1.3 to 3.2 percent of the 1995 infection rate (the 1995 value for X_6). The 1996 values for Y, X_1, X_2, X_3, X_4, and X_5 were also obtained for these eight hospitals. Describe how this information can be used to assess the reliability of predictions made with the model fit in part D.

Ph.D. prelim
Spring 1998

Solutions

Methods Question 2

A. Since $t = \frac{.00436}{.00098} = 4.45$ on 110 d.f., there is a significant positive correlation between average length of stay and the number of nurses employed in the hospital. The practical significance of this result is unclear. This is an observational study involving many uncontrolled factors. The number of nurses employed by a hospital will also have a positive correlation with the size of the hospital, for example, and the correlation between the average length of stay and the number of nurses may partially reflect the tendency for larger hospitals to deal with a higher proportion of more difficult cases which require longer patient stays. This regression analysis does not conclusively show that reducing the number of nurses working at a hospital will reduce the average patient stay.

B. Partial t-tests are $t = \frac{-.000614}{.002389} = -0.257$ and $t = \frac{.003925}{.001725} = 2.275$, both with 109 d.f. After controlling for the number of beds in the hospital, the number of nurses employed by the hospital no longer has a significant correlation with length of stay. The size of the hospital, as measured by the number of beds, does have a positive correlation with length of stay, even after adjusting for the number of nurses.

C. One possibility is to divide the hospitals into several size categories, using the number of beds, and make a separate plot of average length of stay against the number of nurses for each of the size categories.

D. $(.02)X_6\, b_6 \pm (t_{(105),.025})(.02)X_6\, S_{b_6} \Rightarrow (.0102)X_6 \pm (.00367)X_6$
$\Rightarrow (.00653X_6, .01387X_6)$

E. (i) This plot is used to check for homogeneity of error variances. It could also indicate if one or more outliers are present.

(ii) If the usual assumption of i.i.d. random errors is appropriate, the covariance matrix for the least squares estimate $\mathbf{b}$ of the regression coefficients is $\sigma^2(X'X)^{-1}$. Consequently, a relatively small eigenvalue for $(X'X)$ indicates a strong collinearity among the explanatory variables that results in a very large variance for $e'b$, where e the eigenvector corresponding to that small eigenvalue of $(X'X)$. The eigenvalues of $(X'X)$ can also be used to partition the variance of each estimated coefficient to indicate which of the explanatory variables are most heavily involved in a near collinearity.

(iii) This is sometimes called a partial residual plot. It indicates the relationship between Y and X_6 after adjusting for the both the linear affects of X_1, X_2, X_3, X_4, X_5 on Y and the linear effects of X_1, X_2, X_3, X_4, X_5 on X_6. A straight line pattern indicates that the $\beta_6 X_{6i}$ term is needed in the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \varepsilon_i$$

A curved pattern in the plot indicates that something more than $\beta_6 X_{6i}$ is needed in the model to adequately account for the association between the infection rate and average length of stay. If this plot exhibits no trend or pattern, it would suggest that X_{6i} has little association with length of stay after adjusting for linear associations with the other variables.

F. For each of the four regions, fit the regression model and compute the residuals sums of squares: $SSR_{northeast}$, $SSR_{north\ central}$, SSR_{south}, and SSR_{west}. Then, compute an F-test,

$$F = \frac{\dfrac{SSR_{part\ D} - (SSR_{northeast} + SSR_{north\ central} + SSR_{south} + SSR_{west})}{(21)}}{\dfrac{SSR_{northeast} + SSR_{north\ central} + SSR_{south} + SSR_{west}}{84}}$$

where $SSR_{part\ D} = (105)(1.3988) = 146.874$ is the residual sum of squares for the model fitted in part D of this problem. Reject the null hypothesis if $F > F_{(21,84),.05} = 1.70$.

G. For each hospital, substitute the 1996 values of the explanatory variables into the regression equation estimated from the 1995 data in part (D) to obtain the predictions for 1996. Compare each prediction to observed average length of stay for each hospital in 1996. This could be done with an overall F-test, or individual t-tests or confidence intervals. It would be a good idea to consider the possible correlation between the lengths of stay for 1995 and 1996 within individual hospitals.

PhD Prelim, 1998

Methods Question 3

The focus of this question is the formulation of appropriate statistical models for a given situation. In order for a statistical model to be useful for addressing scientific questions of interest, the model must be formulated in such a manner that quantities appearing in the model (e.g., random variables, parameters) are connected with pertinent aspects of the scientific question under investigation. With this in mind, read the description of the study given below, and provide answers to the questions asked.

Study Description

Striped bass are anadromous fish (fish that live their adult lives in the ocean and swim into freshwater rivers to spawn). The larval fish are essentially eyes with an attached tail and yolk sac that are pushed downstream toward the ocean by current. During this journey, the larval fish must develop the ability to be motile (i.e., be able to move under their own power), the ability to secure food on their own, and the physiological ability to tolerate salinity. If the proper physiological changes that allow tolerance to salinity are not developed in the time required for larval fish to travel from the (freshwater) spawning areas to the (saltwater) ocean they die from salinity poisoning. It is felt that the first 10 to 12 days of life are the most critical time for these fish, since this is the time they can live off of the yolk sac they are hatched with.

Industrial, commercial, and residential development, as well as water control projects (levees, dams, waste water treatment, etc.) can affect the salinity levels in estuarine areas and in tributaries to oceanic areas such as the Chesapeake Bay. This, in turn, can alter the salinity gradients that larval striped bass are exposed to as they move from spawning areas to the ocean. Thus, it is important to know the level of tolerance that larval striped bass have to elevated levels of salinity in the first few days of life.

A study was conducted to determine the tolerance of newly hatched larval striped bass to various levels of salinity. That is, the objective of the study was to determine how well larval striped bass survive throughout the first few days of life, relative to the salinity to which they are exposed. In this study, 12 groups of 100 larval striped bass each were randomly assigned to various levels of salinity (0, 3, 6, 9, 12, 15, 18, 21, 24, 27, 30, and 33 parts per trillion, denoted as ppt) to which they were exposed for 10 days. Initially, all larvae were 24 hours old. Salinity was maintained at these levels through use of flow-through tank design (striped bass would die without moving water) and continuous exposure apparatus. On each day, the number of mortalities was recorded for each group and tanks suction cleaned (so this could be done without removing the fish, since handling can also cause mortality).

1. What is the purpose of including the 0 ppt salinity treatment in this study? Would

this treatment necessarily be included in any model that incorporates terms for the effect of salinity concentration?

2. Let i index salinity concentration, $i = 1, \ldots, 12$, and define Y_i = the number of mortalities at the end of day 10 for concentration i.

   (a) Assuming that the probability of mortality within 10 days is identical for each fish in a given treatment, specify (i.e., write down) appropriate probability mass functions for these random variables. Do not forget to include the support of these functions and the allowable parameter space.

   (b) Identify the quantities (or functions of quantities) in your probability mass functions of question 2(a) that represent the probabilities of mortality for individual fish in each treatment group.

   (c) If levels of salinity are represented as the covariates $\{x_i; i = 1, \ldots, 12\}$, a suggested model to relate the probabilities of mortality to the levels of salinity is

   $$Pr(\text{ mortality at salinity level } i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_i^j; \quad i = 1, \ldots, 12,$$

   where estimates of $\{\beta_0, \ldots, \beta_k\}$ are to be obtained by ordinary least squares and the value of k is to be determined from scatterplots of the observed levels of mortality against salinity and diagnostics of model fit such as residual plots. Comment on the appropriateness of this type of model. Do you see any contradications between this model formulation and the probability mass functions you specified in question 2(a)? Comment on the appropriateness of ordinary least squares for estimation of the parameters $\{\beta_0, \ldots, \beta_k\}$.

   (d) Do you see any reason or reasons a model of the form

   $$Pr(\text{ mortality at salinity level } i) = F(\beta, x_i); \quad i = 1, \ldots, 12,$$

   for some appropriately chosen function F would not fully address the objective, as given in the last paragraph of the study description, of determining how well larval striped bass survive throughout the first few days of life?

3. Let i index salinity concentration, $i = 1, \ldots, 12$, and let j index day of the study, $j = 1, \ldots, 10$. Define $Y_{i,j}$ = number of mortalities at salinity level i on day j.

   (a) Assuming that, at the beginning of the study, the probability of mortality on a given day is the same for each fish in a given treatment, specify appropriate probability mass functions for these random variables.

   (b) Let the probabilities of question 3(a) be denoted as $\{p_{i,j}; i = 1, \ldots, 12; j = 1, \ldots, 10\}$. That is, let $p_{i,j} = Pr(\text{ a randomly chosen fish from trt } i \text{ dies on day}$

j of the study). Also, define the quantities $\{\theta_{i,j}; i = 1, \ldots, 12; j = 1, \ldots, 10\}$ as $\theta_{i,j} = Pr($ a randomly chosen fish from trt i dies on day j of the study, given that it was alive on day $j-1$).

i. Reparameterize your probability mass functions of question 3(a) in terms of the quantities $\{\theta_{i,j}; i = 1, \ldots, 12; j = 1, \ldots, 10\}$.

ii. Give a formula for maximum likelihood estimates of the quantities $\{p_{i,j}; i = 1, \ldots, 12; j = 1, \ldots, 10\}$.

iii. Give a formula for maximum likelihood estimates of the quantities $\{\theta_{i,j}; i = 1, \ldots, 12; j = 1, \ldots, 10\}$.

(c) Consider the following data for 3 of the 12 treatment groups:

	Mortalities on Day									
Salinity	1	2	3	4	5	6	7	8	9	10
12	1	1	3	13	9	12	4	7	3	4
15	1	1	3	2	11	11	14	5	8	3
18	3	6	30	42	4	3	4	5	1	2

For each of these 3 treatment groups:

i. Find maximum likelihood estimates of the probabilities that a fish randomly selected at the beginning of the study will die on day 5 of the study (i.e., $p_{i,5};\ i = 1, 2, 3$).

ii. Find maximum likelihood estimates of the conditional probabilities of mortality on day 5 given survival at the end of day 4 (i.e., $\theta_{i,5};\ i = 1, 2, 3$).

iii. If you were given the estimated covariance matrix for the maximum likelihood estimators of $\{p_{1,j}; j = 1, \ldots, 5\}$, Σ_5 say, explain how you would find an approximate 90% confidence interval for $\theta_{1,5}$.

(d) Outline (do not conduct) a statistical analysis to test the hypothesis that the entire probability mass functions (i.e., over all days of the study) for all 12 treatment groups are identical, versus the alternative that they are not.

4. Return to consideration of all 12 levels of salinity (or 11 levels excluding the 0 ppt salinity treatment if you wish). For salinity in the 12 treatment groups represented as $\{x_i; i = 1, \ldots, 12\}$, formulate a statistical model to relate the conditional probabilities of daily mortality (i.e., the $\{\theta_{i,j}; i = 1, \ldots, 12; j = 1, \ldots, 10\}$) to salinity level. To do so you may use any of the random variables discussed previously in this question or define new random variables of your choosing.

Note:

In answering this question it is acceptable to leave some portions of your answer fairly general. For example, you could formulate a model to relate a set of expectations $\{\mu_k; k = 1, \ldots, n\}$ to a set of covariates $\{z_k; k = 1, \ldots, n\}$ as,

$$\mu_k = G(\beta, z_k);\quad k = 1, \ldots, n,$$

without specifying a functional form for G.
For the model you formulate, indicate:

(a) The quantities in the model on which inference should be made in order to address the objective of this study.

(b) Criteria you would use in interpretation of the results of inferential procedures (e.g., estimation, testing, model assessment). For example, if you wish to estimate 12 quantities $\{\eta_i; i = 1, \ldots, 12\}$ what do large versus small values of these quantities tell you about the survival of striped bass larvae relative to salinity exposure?

PhD Prelim, 1998

Answer

1. The 0 ppt salinity treatment functions primarily as a 'procedural control', that is, for assessment of whether there is a background level of mortality caused by the experimental apparatus and study protocol. If the level of mortality in this control were too high, the study would not be accepted as having been conducted correctly. In a model that makes use of salinity concentration as a covariate, the 0 ppt salinity treatment might be incorporated as one value of the covariate, but this would not necessarily be the case. The primary value of this treatment is to serve as a check on proper study protocol.

2. (a) Assuming that the probability of mortality within 10 days is identical for each fish in a given treatment group, an appropriate probability mass function would be the standard binomial. For $i = 1, \ldots, 12$, let Y_i have probability mass function,

$$f(y_i|n_i, \theta_i) = \frac{n_i!}{y_i!\,(n_i - y_i)!}\theta_i^{y_i}\,(1-\theta_i)^{n_i - y_i}; \quad y_i \in \{0, 1, \ldots, n_i\},$$

for $\theta_i \in (0, 1)$; $i = 1, \ldots, 12$. Here, $n_i = 100$; $i = 1, \ldots, 12$. Even though the question explicitly asks for a probability mass function, an alternative answer that would be acceptable would be to indicate that, since 100 is a fairly large number of binary trials for a binomial, the distribution of the number of mortalities might be approximated by a normal distribution. One would need to worry, however, about treatments (i.e., salinity concentrations) that result in few mortalities or in nearly total mortality. Thus, for example, if one chooses to include the 0 ppt salinity treatment as an 'active' treatment, this might be a questionable approximation.

   (b) The parameters $\{\theta_i; i = 1, \ldots, 12\}$ of answer 2(a) represent the probability of mortality for a randomly selected fish from treatment group i. If a continuous approximation (e.g., normal) is used in question 2(a) the correct answer here would be that there is no simple function of the parameters that represents the probability of mortality for individual fish in a treatment group. The mean could, however, be interpreted as the 'force of mortality' for a given treatment. Here, as throughout the question, points are given for consistency in the answer.

   (c) The model given takes the probability of mortality to be a linear model in salinity. Without external constraint, the function does not have range equal to the interval (0, 1). This is a potential problem. Again, one might indicate that the function, although not strictly applicable as a model of a probability, might be used as an approximation within the range of the data if that is appropriate. Under a binomial model, the variance of the response variables is not constant, thus rendering ordinary least squares inappropriate for estimation.

(d) Models of the form given take the response variable of interest to be mortality at the end of the study only. That is, no use is made of information collected on the time of mortality. The stated objective, to determine how well striped bass survive *throughout* the first few days of life, relative to salinity, is addressed in only a gross manner by looking at total mortality over the 10 days of the study.

3. (a) Assuming that, at the beginning of the study, the probability of mortality on a given day is the same for each fish in a given treatment, a suitable probability mass function would be a multinomial for each treatment,

$$f(\boldsymbol{y}_i|\boldsymbol{p}_i) = C\, p_{i,1}^{y_{i,1}}\, p_{i,2}^{y_{i,2}} \cdots p_{i,10}^{y_{i,10}} \left(1 - \sum_{j=1}^{10} p_{i,j}\right)^{100-\sum_{j=1}^{10} y_{i,j}} ; \quad i = 1, \ldots, 12, \tag{1}$$

where $C = 100!/(y_{i,1}! \ldots y_{i,10}!)$, and for $\boldsymbol{y}_i \equiv (y_{i,1}, \ldots, y_{i,10})^T$, $\boldsymbol{p}_i \equiv (p_{i,1}, \ldots, p_{i,10})^T$, and $y_{i,j} \in \{0, 1, \ldots, 100\}$ subject to $\sum_{j=1}^{10} y_{i,j} \leq 100$.

(b) i. Letting $\theta_{i,j}$ denote the conditional probabilities given in the question, a reparameterization of the probability mass functions of 3(a) would be, $f(\boldsymbol{y}_i|\boldsymbol{\theta}_i) =$

$$C\, (\theta_{i,1})^{y_{i,1}} \{(1-\theta_{i,1})\theta_{i,2}\}^{y_{i,2}} \cdots \left\{\theta_{i,10} \prod_{j=1}^{9} (1-\theta_{i,j})\right\}^{y_{i,10}} \left\{\prod_{j=1}^{10} (1-\theta_{i,j})\right\}^{100-\sum_{j=1}^{10} y_{i,j}},$$

where $\boldsymbol{\theta}_i \equiv (\theta_{i,1}, \ldots, \theta_{i,10})^T$. In less cumbersome notation, one might say the probability mass functions are as in equation (1) with

$$\begin{aligned} p_{i,1} &= \theta_{i,1} \\ p_{i,j} &= \theta_{i,j} \prod_{k=1}^{j-1} (1-\theta_{i,k}); \quad j = 2, \ldots, 10. \end{aligned} \tag{2}$$

ii. Given the probability mass functions of equation (1), maximum likelihood estimates are, for $i = 1, \ldots, 12$ and $j = 1, \ldots, 10$,

$$\hat{p}_{i,j} = y_{i,j}/100. \tag{3}$$

iii. Since, from (2),

$$\begin{aligned} \theta_{i,1} &= p_{i,1} \\ \theta_{i,j} &= \frac{p_{i,j}}{\prod_{k=1}^{j-1}(1-\theta_{i,k})}; \quad j = 2, \ldots, 10, \end{aligned}$$

maximum likelihood estimates of the $\{\theta_{i,j}; i = 1, \ldots, 12;\ j = 1, \ldots, 10\}$ are given by invariance as,

$$\begin{aligned} \hat{\theta}_{i,1} &= y_{i,1}/100 \\ \hat{\theta}_{i,j} &= \frac{y_{i,j}/100}{1 - \sum_{k=1}^{j-1}(y_{i,k}/100)};\ \ j = 2, \ldots, 10. \end{aligned} \tag{4}$$

(c) i. From equation (3) the maximum likelihood estimates are,

$$\begin{aligned} \hat{p}_{1,5} &= 0.0900 \\ \hat{p}_{2,5} &= 0.1100 \\ \hat{p}_{3,5} &= 0.0400 \end{aligned} \tag{5}$$

ii. From equation (4) the maximum likelihood estimates are,

$$\begin{aligned} \hat{\theta}_{1,5} &= 0.1098 \\ \hat{\theta}_{2,5} &= 0.1183 \\ \hat{\theta}_{3,5} &= 0.2105 \end{aligned} \tag{6}$$

iii. For $\theta_{1,5}$, equation (4) may be written as

$$\hat{\theta}_{1,5} = g(\hat{p}_{1,1}, \ldots, \hat{p}_{1,5}) = \frac{\hat{p}_{1,5}}{1 - \sum_{k=1}^{4} \hat{p}_{1,k}}.$$

To form a 90% confidence interval for $\theta_{1,5}$ one would rely on the asymptotic normality of the estimates $\{\hat{p}_{1,j}; j = 1, \ldots, 10\}$. Through transformation of asymptotically normal statistics (i.e., the 'delta' method), one would calculate the (5×1) vector

$$D \equiv \left(\frac{\partial g}{\partial \hat{p}_{1,1}}, \ldots, \frac{\partial g}{\partial \hat{p}_{1,5}} \right)^T,$$

and the estimated asymptotic variance of $\hat{\theta}_{1,5}$ would be computed as

$$\hat{var}(\hat{\theta}_{1,5}) = D^T \Sigma_5 D.$$

An approximate 90% confidence interval for $\theta_{1,5}$ would then be computed from,

$$\hat{\theta}_{1,5} \pm 1.645 \left\{ \hat{var}(\hat{\theta}_{1,5}) \right\}^{1/2}.$$

(d) Perhaps the simplest procedure would be to use a likelihood ratio test statistic using the full model log likelihood as the sum over i of the logarithms of (1) evaluated at the maximum likelihood estimates $\{\hat{p}_{i,j}; i = 1, \ldots, 12;\ j = 1, \ldots, 10\}$ and, using the reduced model log likelihood as this same sum subject to the constraint that $p_{i,j} = p_j$ for $i = 1, \ldots, 12$. Various Chi-square alternatives are also available if developed from the multinomial distribution.

4. Answers to this question will vary. One possibility is to define random variables for each fish as the time to mortality and approach the problem as one of survival analysis. Differences among the treatment groups would then be embodied in hazard functions.

   Another possibility is to retain the random variables $\{Y_{i,j}; i = 1, \ldots, 12;\ j = 1, \ldots, 10\}$, write the multinomial probability mass functions as in question 4(a), and define covariates $\{z_{i,j}; i = 1, \ldots, 12;\ j = 1, \ldots, 10\}$ as

   $$z_{i,j} = j\, x_i.$$

   That is, $z_{i,j}$ represents the total salinity to which a fish from treatment group i has been exposed to up to time j. An appropriate model would then focus on the conditional probabilities of mortality,

   $$\theta_{i,j} = F(z_{i,j}, \beta),$$

   for some suitably chosen function F and parameter β. The function F should have range (0, 1), and interesting aspects of its behavior should be governed by the parameter β. For example, if F can be chosen to be monotinic increasing in $z_{i,j}$ (equivalently, monotonic in time, j) this would indicate a continually increasing 'force of mortality' for salinity level. In contrast, a monotone decreasing F would indicate that groups of fish are composed of 'weaker' and 'stronger' individuals. Different yet would be a unimodal F, which would suggest an acclimation on the part of fish that survive early exposure.