STAT 511 **Final Exam** **Spring 2000**

Instructions: There is not enough room to write your answers on this exam, Please record your answers on other sheets of paper. Please write your name on the upper right corner of every page you submit.

You may use a calculator. You may also bring two sheets of notes to the exam. You may write on both sides of the two sheets of notes. No other books or notes are allowed.

This is a two hour exam. Complete as much of it as you can. Do not spend too much time on any single part of this exam. Do not try to complete lengthy numerical calculations. You will receive complete credit by showing that you know how to use an appropriate procedure for solving a problem. This could be done, for example, by displaying an appropriate formula and inserting appropriate values into the formula to show that you know how to properly use it. In performing tests or constructing confidence intervals, you are not required to find values of percentiles or probabilites of F, t, chi-square, or normal distributions, but you should indicate appropriate degrees of freedom.

1. In a study of the effects of certain types of air polution on the yield of crops, soybeans were exposed to ozone for seven hours during each day of the growing season. Each plot was covered by a plastic shell so that a specific concentration of ozone in the atmosphere could be maintained throughout the growing season. Different ozone concentrations were maintained in different plots. n=94 plots were used in this study.

Consider the model

$$Y_i = \beta_1 e^{-(X_i/\beta_2)^{\beta_3}} + \epsilon_i$$

where $\beta_1 > 0$, $\beta_2 > 0$ and $\beta_3 > 0$. Also, Y_i denotes the observed yield in the i-th plot, X_i is the average daily ozone level (ppm), and $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independent random errors.

S-PLUS code for least squares estimation of the parameters in this model is shown on page 3 along with some of the results. A graph of the data and the fitted curve is shown at the bottom of page 2. Answer the following questions:

(a) Give an interpretation of β_1 in this model.

(b) Estimate the mean soybean yield when the ozone level is 0.12 ppm.

(c) Show how to use the delta method to obtain an approximate 90 percent confidence interval for the mean soybean yield when the ozone level is 0.12 ppm. Set up the formulas and fill in the numbers, but you do not have to complete the numerical computations.

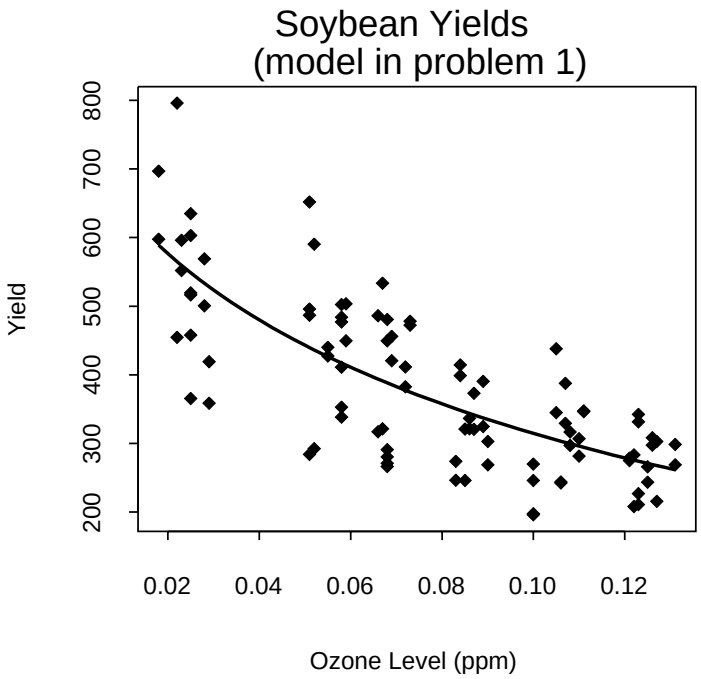
(d) An exponential decay model

$$Y_i = \alpha_1 e^{-(X_i/\alpha_2)} + \epsilon_i, \qquad \alpha_1 > 0,\ \alpha_2 > 0,$$

was also fit to these data. The sum of squared residuals is 611312. Can this additional information be used to test the the hypothesis that the exponential decay model fits the data as well as the model that was originally proposed? Explain. If your answer is yes, show how a test can be constructed.

(e) Assuming that the nls( ) function shown on page 3 has been executed, describe what is accomplished by the following four lines of S-PLUS code.

```
plot(oz$ozone, oz$yield, type="n", xlab="Ozone Level (ppm)", ylab="Yield")
gsize<-51
a <- seq(min(oz$ozone), max(oz$ozone), length = gsize)
lines(a, predict(oz.nls, data.frame(ozone=a),type="response"),lty=1,lwd=3)
```

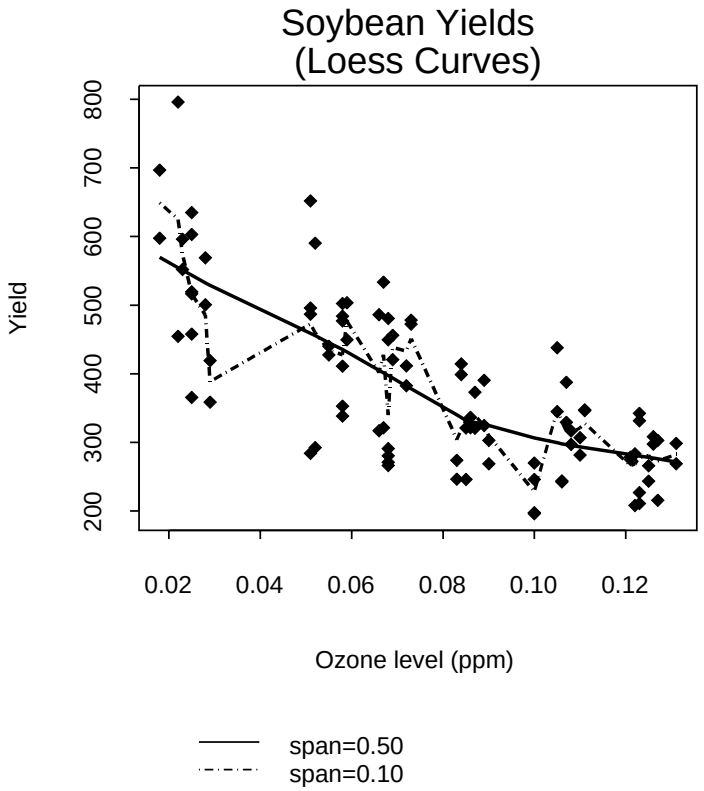

Soybean Yields
(model in problem 1)
Yield
200
300
400
500
600
700
800
0.02
0.04
0.06
0.08
0.10
0.12
Ozone Level (ppm)

```
# There are two numbers on each line in the following order:
#      Ozone level (ppm)
#      soybean yield
#  Enter the data into a data frame:

oz <- read.table(ozoneso2.txt", header=T)
oz

   ozone   yield
 1 0.018   597.5
 2 0.018   697.0
 3 0.022   454.5
 .  .         .
 .  .         .
92 0.127   215.5
93 0.131   269.0
94 0.131   298.5


oz.nls <- nls(formula = yield ~ b1*exp(-(ozone/b2)^b3), data = oz,
                 start = c(b1=700.0,b2=0.1,b3=1.0), trace = T)

803145 : 700 0.1 1
621612 : 747.669 0.115506 0.70697
606408 : 779.872 0.115155 0.682755
606404 : 781.845 0.114757 0.681063


summary(oz.nls)

Formula: yield ~ b1 * exp( - (ozone/b2)^b3)

Parameters:
         Value   Std. Error  t value
b1  781.845000  235.5900000  3.31866
b2    0.114757    0.0478974  2.39590
b3    0.681063    0.3498160  1.94692


vcov(oz.nls)
              b1             b2           b3
b1  55502.82561 -11.160006715 -80.76421542
b2    -11.16001   0.002294165   0.01591608
b3    -80.76422   0.015916084   0.12237125
```

2. The following S-PLUS code was used to fit two cuvres to the data from problem 1, using the loess procedure.

```
oz.50 <- loess(formula=yield~ozone,data=oz,span=.50,degree=1)
oz.10 <- loess(formula=yield~ozone,data=oz,span=.10,degree=1)
```

The curves are shown on the following graph.



(a) Briefly describe the "variance" and "bias" issues that one must consider in deciding how smooth to make the fitted curve.

(b) Aside from looking a plot of fitted curves, as shown above, describe two methods for choosing the value of the "span" option.

(c) Suppose the data file contains a case for which the ozone level is 1.2 ppm. Describe how the loess procedure implemented by the code

```
oz.10 <- loess(formula=yield~ozone,data=oz,span=.10,degree=1)
```

estimates the expected yield for soybeans grown in an enviroment with ozone level at 0.12 ppm.

(d) For the loess estimation in part (c), describe how a bootstrap procedure could be used to construct a 90 percent confidence interval for the expected yield at an ozone level of 0.12 ppm. Do **not** report any S-PLUS code in this answer.

3. A lactation period is a period of time during which a cow produces milk. It begins after a cow gives birth to a calf. Eventually milk production begins to diminish and ultimately stops. The cow must give birth to another calf to begin a new lactation period. A cow's first lactation period begins after it gives birth for the first time, and the second lactation period begins after it gives birth for the second time, and so on.

A total of 24 cows were used in a study of the effects of a hormone treatment on milk production. The 24 cows were randomly divided into three treament groups. Each of the eight cows in the treatment A group was given the hormone treatment 30 days after the beginning of each of its first four lactation periods. Each of the eight cows in the treatment B group was given the hormone treatment 60 days after the beginning of each of its first four lactation periods. Each of the eight cows in the third group received no hormone treatment. The third group is a control group. The main objective was determine if treatment with the hormone could increase the amount of milk produced during a lactation period. This could be achieved by increasing daily milk production, increasing the length of the lactation period, or both. The total weight of milk produced was recorded for each cow during each of its first four lactation periods. Responses from different cows can be assumed to be independent, but reponses from a single cow during different lactation periods may be positively correlated.

(a) First consider the milk produced during the first lactation period of each cow. A model for the 24 observations is

$$Y_{ij1} = \mu + \alpha_i + \epsilon_{ij1},$$

where Y_{ij1} denotes the total weight of the milk produced by the $j-th$ cow in the $i-th$ group during its first lactation period, and ϵ_{ij1} is a corresponding random error. Describe the set of estimable functions of $\mu,\ \alpha_1,\ \alpha_2,\ \alpha_3$.

(b) The least squares estimator for $\mu + \alpha_i$ is $\overline{Y}_{i\bullet 1}$. State conditions under which $\overline{Y}_{i\bullet 1}$ is a best linear unbiased estimator.

(c) Define $MSE = \frac{1}{21}\sum_{i=1}^{3}\sum_{j=1}^{8}(Y_{ij1} - \overline{Y}_{i\bullet 1})^2$, and let $a = (a_1, a_2, a_3)^T$ be a vector of constants. Show that

$$F = \frac{(a_1\overline{Y}_{1\bullet 1} + a_2\overline{Y}_{2\bullet 1} + a_3\overline{Y}_{3\bullet 1})^2}{MSE}$$

has a F-distribution under certain conditions. Clearly state the conditions.

4. Now consider an analysis of the milk production data from problem 3 for the four lactation periods for each cow. Let

$$\mathbf{Y}_{ij} = \begin{bmatrix} Y_{ij1} \\ Y_{ij2} \\ Y_{ij3} \\ Y_{iJ4} \end{bmatrix}$$

denote the set of milk production values provided by the $j-th$ cow in the $i-th$ treatment group.

(a) One possible model for these data is

$$Y_{ijk} = \mu + \alpha_i + \eta_{ij} + \beta_k + \gamma_{ik} + \epsilon_{ijk},$$

where $\alpha_1, \alpha_2, \alpha_3$ are fixed parameters corresponding to the treatment groups, $\beta_1, \beta_2, \beta_3, \beta_4$ are fixed parameters corresponding to lactation periods, the γ_{ik}'s are fixed interaction parameters, and $\eta_{ij} \sim NID(0, \sigma_C^2)$ are random cow effects, and $\epsilon_{ijk} \sim NID(0, \sigma_E^2)$ are random errors, and any η_{ij} is independent of any ϵ_{ijk}. What is the distribution of $\mathbf{Y}_{ij}$, the vector of four responses from a single cow, for this model?

(b) The model in part (a) could be written in the form

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}$$

where $\mathbf{X}$ is an appropriate model matrix, $\beta = (\mu, \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3, \beta_4, \gamma_{11}, \ldots, \gamma_{34})^T$, and $\Sigma = Var(\mathbf{e})$ is a block diagonal matrix with each $4x4$ block correponding to the covariance matrix you reported in part (a) for a set of four measurements taken on a single cow. Assuming that the variance components σ_C^2 and σ_E^2 are known for the model in part (a). What is the best linear unbiased estimator for an estimable function of β, say $C\beta$?

(c) REML estimates of variance components are $\widehat{\sigma}_C^2 = 1211.56$ and $\widehat{\sigma}_E^2 = 321.34$. The corresponding value of the REML log-likelihood is 734.78. Give formulas to express the REML estimators as functions of mean squares from the ANOVA table for the model in part (a).

(d) The model in part (b) was also fit to the data using a non-homogeneous Toeplitz model for the covariance matrix for the sets of four measurements taken on individual cows. That is, each $4x4$ matrix on the diagonal of Σ has the form

$$\begin{bmatrix} \sigma_1^2 & \rho_1\sigma_1\sigma_2 & \rho_2\sigma_1\sigma_3 & \rho_3\sigma_1\sigma_4 \\ \rho_1\sigma_2\sigma_1 & \sigma_2^2 & \rho_1\sigma_2\sigma_3 & \rho_2\sigma_2\sigma_4 \\ \rho_2\sigma_3\sigma_1 & \rho_1\sigma_3\sigma_2 & \sigma_3^2 & \rho_1\sigma_3\sigma_4 \\ \rho_3\sigma_4\sigma_1 & \rho_2\sigma_4\sigma_2 & \rho_1\sigma_4\sigma_3 & \sigma_4^2 \end{bmatrix}$$

The value of the REML log-likelihood for this model is 752.15. Show how you would determine if the non-homogeneous Toeplitz covariance model is a significant improvement over the covarince model used in part (b). Give a formula or value for a test statistic and show how to use it to make a decision.