As always, not every possible way of expressing an appropriate answer is included in the following solutions. A point value for each part of each questionn is shown in parentheses. You should have a score for each part of each question written on your paper, even if it is zero. Your score for the final exam is written at the bottom of the last page of your paper. Check it to make sure that your total score was properly recorded. Bring any irregularities to the attention of the instructor.

1.(a) (4 points) β_1 represents the expected (or mean) soybean yield when there is no ozone in the atmosphere near the plant.

(b) (4 points) The estimate of the mean soybean yield at an ozone level of 0.12 ppm is

$$\widehat{Y} = 781.845 exp(-(0.12/0.114757)^{0.681063}) = 278.9.$$

The estimate of the mean soybean yield when the ozone level is 1.2 ppm is

$$\widehat{Y} = 781.845 exp(-(1.2/0.114757)^{0.681063}) = 5.6.$$

(c) (10 points) First consider estimating the mean yield when the ozone level is 0.12 ppm. Derive a vector of first partial derivatives of the formula for the mean yield of soybeans, i.e.,

$$G = exp(-(X/\beta_2)^{\beta_3}) \left[1,\ \beta_1(\beta_3/\beta_2)(X/\beta_2)^{\beta_3},\ -\beta_1(X/\beta_2)^{\beta_3} log(X/\beta_2)\right]$$

and evaluate it at $X = .12$ and $(\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3) = (781.845, 0.114757, 0.681063)$ to obtain $\widehat{G}$. Then, a large sample approximation to the standard error of the estimated mean yield at $X = .12$ ppm is the square root of $\widehat{G} * V * \widehat{G}^T$, where V is the estimated covariance matrix for the parameter estimates shown at the bottom of page 3 of the exam. The value of the approximate standard error is 14.8, and an approximate 90 percent confidence interval is

$$278.9 \ \pm \ t_{(91),.05} \sqrt{\widehat{G} * V * \widehat{G}}$$

An approximate standard error for the estimate of the mean yield when the ozone level is 1.2 is obtained by evaluating the same formulas at $X = 1.2$. The value of the approximate standard error is 16.9, and an approximate 90 percent confidence interval is

$$5.6 \ \pm \ t_{(91),.05} \sqrt{\widehat{G} * V * \widehat{G}}.$$

This interval contains negative values which are impossible yields. The large sample normal approximation to the distribution of the estimated mean at $X = 1.2$ does not provide a satisfactory result.

(d) (6 points) Since the exponential decay model is obtained from the model proposed at the beginning of this problem by setting $\beta_3 = 1$, it is nested within the model proposed at the beginning of this problem and we can construct an approximate F-test. We have

$$F = \frac{(611312 - 606404)/(92 - 91)}{(606404/91)} = 0.74$$

with $(1, 91)$ degrees of freedom. Since $0.74 < F_{(1,91).50}$, the more complicated model proposed at the beginning of the problem is not shown to be a significant improvement over the exponential decay model.

(e) (5 points) The lines

```
gsize<-51
a <- seq(min(oz$ozone), max(oz$ozone), length = gsize)
```

create a list of 51 equally spaced ozone concentrations between the minimum and maximum ozone concentrations in the data set. The predict( ) function in the line

```
lines(a, predict(oz.nls, data.frame(ozone=a),type="response"),lty=1,lwd=3)
```

evaluates the estimates of the mean soybean yields at each of the 51 ozone concentrations in the list and the lines( ) function plots the points and connects them with a solid line.

2.(a) (6 points) If the span is too large the fitted curve may not be able to bend quickly enough resulting in substantially biased estimates of mean yields at some ozone levels. Using a smaller span can potentially reduce bias by allowing the fitted curve more freedom to bend to fit local trends in yields, but it also increases the the variances of estimated means. A good trade off may be to try to minimize

$$\text{Expected Squared Error} = variance + bias^2$$

(b) (6 points) Some methods are: (i) Plot residuals against the ozone concentrations and look for any pattern in this plot that would indicate that the proposed model is inadequate. This may be aided by passing a smooth curve through the residual plot and comparing it to a horizontal line. (ii) Use approximate F-tests to compare the fit of loess curves created with different spans but the same degree (linear of quadratic) of local polynomial regression. (iii) Try different span values and choose the one that minimizes an AIC criterion, e.g. (sum of squared residuals) + 2(effective number of parameters). (iv) Use a crossvalidation procedure to find the span that approximately minimizes the sum of the expected squared prediction errors.

(c) (6 points)

- Since span=.10, the loess procedure would locate $(.10)(94) = 9$ cases in the data set with ozone concentrations closest to 0.12 ppm.
- The tri-cubed weight function would be used to assign a weight to each of these 9 closest cases, with weights decreasing toward zero as the ozone concentration gets farther away from 0.12 ppm. Cases outside ozone levels outside this neighborhood of 0.12 would receive zero weight.
- Weighted least squares is used to fit a a regression line, and the fitted line is used to estimate the expected yield at 0.12 ppm.

(d) (6 points)

(step 1) Select a bootstrap sample of 94 plots, using simple random sampling with replacement, from the original sample of plots.

(step 2) Fit a loess curve to the bootstrap sample and obtain an estimate of the expected yield at ozone level 0.12 ppm.

(step 3) Repeat steps 1 and 2 for $B = 5000$ bootstrap samples and order the estimated mean yields at 0.12 from smallest to largest, i.e, $\widehat{Y}_{[}1] \leq \widehat{Y}_{[}2] \leq \ \ldots \ \leq \widehat{Y}_{[}5000]$.

(step 4) Compute $250 = [(.05)(5001)]$ and $4751 = 5001 - 250$, and a 90 percent "percentile" confidence interval for the mean yield is $(\widehat{Y}_{[}250],\ \widehat{Y}_{[}4751])$.

Alternaively, you could resample from the sets of centered and scaled residuals in step 1 to approximate sampling from the population of random errors. Add a randomly selected value from the centered and scaled residuals to each of the estimated mean yield values from the original sample to generate new "observed" yield values. Then, go through steps 2-4. This would use the original set of 94 ozone concentrations for each bootstrapped sample. This may be a better way to obtain bootstrap samples if the researchers could precisely control the ozone levels for the individual plots.

3. This problem is very similar to parts of problems from the midterm exam. I was looking for simple answers in parts (a) and (b). Few people were able to propoerly organize and complete an answer to part (c).

(a) (5 points) Since an estimable function must be a linear function of the means of the observations, the set of estimable functions consists of all linear functions of $\mu+\alpha_1$, $\mu+\alpha_2$ and $\mu+\alpha_3$, i.e. $a_1(\mu+\alpha_1)+a_2(\mu+\alpha_2)+a_3(\mu+\alpha_3)$ for any $(a_1,a_2,a_3)\neq(0,0,0)$.

(b) (5 points) The random errors are independent with zero means and homogeneous variances, i.e. $E(\epsilon_{ij1})=0$ and $Var(\epsilon_{ij1})=\sigma^2$ for all (i,j). This is the Gauss-Markov property.

(c) (12 points) Here there is one observation per cow and you were told to assume that each cow responds independently of any other cow. Let

$$\mathbf{Y}=[Y_{111},Y_{121},\ldots,Y_{181},Y_{211},\ldots,Y_{281},Y_{311},\ldots,Y_{381}]^T$$

denote the vector of independent observations for the $n=24$ cows. If the cows are a sample fromm some larger population of cows, it would be reasonable to assume that the observations have homogeneous variances. We will also need to a normality assumption to get an exact F-distribution. So, assume a normal theory Gauus-Markov model for $\mathbf{Y}$, i.e., $\mathbf{Y}\sim N(\mathbf{X}\beta,\sigma^2 I)$., where the transpose of the model matrix is

$$X=\begin{bmatrix}1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1&1\\1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1\end{bmatrix}$$

and $\beta=[\mu,\alpha_1,\alpha_2,\alpha_3]^T$.

First, we can use Result 4.7, as shown in the lecture notes, to show that

$$\frac{(n-3)MSE}{\sigma^2}\;=\;\frac{\mathbf{Y}^T(I-P_X)\mathbf{Y}}{\sigma^2}\;\sim\;\chi^2_{n-rank(X)}$$

Here, $n-rank(X)=24-3=21$.

Now determine conditions under which $(a_1\overline{Y}_{1.1}\;+\;a_2\overline{Y}_{2.1}\;+\;a_3\overline{Y}_{3.1})^2$ has a chi-squared distribution. Note that

$$(a_1\overline{Y}_{1.1}\;+\;a_2\overline{Y}_{2.1}\;+\;a_3\overline{Y}_{3.1})^2=c^T\mathbf{Y}$$

where

$$c^T\;=\;[\frac{a_1}{8},\ldots,\frac{a_1}{8},\frac{a_2}{8},\ldots,\frac{a_2}{8},\frac{a_3}{8},\ldots,\frac{a_3}{8}].$$

Then,

$$(a_1\overline{Y}_{1.1}\;+\;a_2\overline{Y}_{2.1}\;+\;a_3\overline{Y}_{3.1})^2\;=\;\mathbf{Y}^Tcc^T\mathbf{Y}.$$

Applying result 4.7 from the notes with $\Sigma=\sigma^2 I$ and $A=cc^T/\sigma^2$, we have that

$$\frac{\mathbf{Y}^Tcc^T\mathbf{Y}}{\sigma^2}\sim\chi^2_1(\delta^2),$$

with non-centrality parameter $\delta^2=\beta^TX^Tcc^TX\beta/\sigma^2$, if $\Sigma A=cc^T$ is idempotent. This requires that $cc^T=cc^Tcc^T=(c^Tc)cc^T$. Consequently, we must have $1=c^Tc=(a_1^2+a_2^2+a_3^2)/8$.

Finally, we must show that $MSE=(21)SSE=(21)\mathbf{Y}^T(I-P_X)\mathbf{Y}$ is independent of $\mathbf{Y}^Tcc^T\mathbf{Y}$. This follows from **Result 4.8** because $A\Sigma(I-P_X)=\frac{1}{\sigma^2}cc^T(\sigma^2I)(I-P_X)=cc^T(I-P_X)=0$. (Note that c is a linear combination of the last three columns of X.)

Consequently,

$$F = \frac{(a_1\overline{Y}_{1.1} \ + \ a_2\overline{Y}_{2.1} \ + \ a_3\overline{Y}_{3.1})^2}{MSE} \ \sim \ F_{(1,21)}(\delta^2), \qquad \text{where } \delta^2 = \frac{(a_1(\mu+\alpha_1) \ + \ a_2(\mu+\alpha_2) \ + \ a_3(\mu+\alpha_3))^2}{\sigma^2}.$$

if $\mathbf{Y} \sim N(X\beta, \sigma^2 I)$ and $(a_1^2 + a_2^2 + a_3^2) = 8$.

Note that $\delta^2 = 0$ if and only if the null hypothesis

$$H_o : (a_1(\mu+\alpha_1) \ + \ a_2(\mu+\alpha_2) \ + \ a_3(\mu+\alpha_3)) = 0$$

is true.

4.(a) (6 points)

$$\mathbf{Y}_{ij} = \begin{bmatrix} Y_{ij1} \\ Y_{ij2} \\ Y_{ij3} \\ Y_{iJ4} \end{bmatrix} \ \sim \ N(\begin{bmatrix} \mu+\alpha_i+\beta_1+\gamma_{i1} \\ \mu+\alpha_i+\beta_2+\gamma_{i2} \\ \mu+\alpha_i+\beta_3+\gamma_{i3} \\ \mu+\alpha_i+\beta_4+\gamma_{i4} \end{bmatrix}, \ V),$$

where

$$V = \begin{bmatrix} \sigma_C^2+\sigma_E^2 & \sigma_C^2 & \sigma_C^2 & \sigma_C^2 \\ \sigma_C^2 & \sigma_C^2+\sigma_E^2 & \sigma_C^2 & \sigma_C^2 \\ \sigma_C^2 & \sigma_C^2 & \sigma_C^2+\sigma_E^2 & \sigma_C^2 \\ \sigma_C^2 & \sigma_C^2 & \sigma_C^2 & \sigma_C^2+\sigma_E^2 \end{bmatrix}$$

is a compound symmetry covariance matrix.

(b) (5 points) Since this is a linear model that does not necessarily satisfy the Gauss-Markov property, the best linear unbiased estimator for an estimable function of β, say $C\beta$, is the generalized least squares estimator

$$C(X^T\Sigma^{-1}X)^- X^T\Sigma^{-1}\mathbf{Y}$$

.

(c) (6 points) $\widehat{\sigma}_E^2 = MS_{error}$ and $\widehat{\sigma}_C^2 = \frac{1}{4}(MS_{cows} - MS_{error})$

(d) (8 points) The model from part (a) has a compound symmetry covariance structure which can be obtained as a special case of the non-homogeneous Toeplitz covariance structure by imposing the restrictions $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_E^2 + \sigma_C^2$ and $\rho_1 = \rho_2 = \rho_3 = \sigma_C^2/(\sigma_C^2 + \sigma_E^2)$. Then, as long as the REML estimates of the variances and covariances are in the interior of the parameter space, -2[(REML log-likelihood model (d)) - (REML log-likelihood model (a))] has an approximate central chi-squared distribution with $7 - 2 = 5$ degrees of freedom when the model in part (a) is appropriate. The non-homogenous Toeplitz covariance structure would be found to provide a better model than the compound symmetry covariance structure if -2[(REML log-likelihood model (d)) - (REML log-likelihood model (a))] $> \chi^2_{(5)}$.

In this case, $-2[734.78 - 752.15] \ = \ 34.74 \ > \chi^2_{(5),.001}$, and the non-homogeneous Toeplitz covariance structure is shown to provide a better fit to the data than the compound symmetry model. This does not necessarily imply, however, that the non-homogeneous Toeplitz covariance structure is correct. You could go on to compare the non-homogeneous Toeplitz covariance stucture to the "unstructured" covariance model.

Exam scores out of a possible 100 points are displayed below as a stem-leaf display.

```
9 | 0 0 1 2
8 | 5 6 6 9 9
8 | 0 1 1 1 2 3 4
7 | 5 6 8 8 8 9 9 9
7 | 0 2 3 4 4 4 4
6 | 5 5 5 8 8 9
6 | 0 0 1 2 2 3 4
5 | 5 8 8 8
5 | 3
4 | 6 9
4 |
3 |
3 |
2 | 9
```