PhD Prelim Exam
METHODS

Summer 2006
(Given on 7/18/06)

Part A:

1. Consider a randomized complete block design with t treatments and r blocks. Let y_{ij} denote the observation from the experimental unit treated with treatment i in block j. Suppose for $i = 1, \ldots, t$ and $j = 1, \ldots, r$;

$$y_{ij} = \mu + \tau_i + b_j + e_{ij}, \tag{1}$$

where μ, $\tau_1, \ldots, \tau_t$ are unknown real-valued parameters; $b_1, \ldots, b_r$ are identically distributed $N(0, \sigma_b^2)$ random variables; e_{ij} $(i = 1, \ldots, t; j = 1, \ldots, r)$ are identically distributed $N(0, \sigma_e^2)$ random variables; all b_j and e_{ij} random variables are jointly independent; and σ_b^2 and σ_e^2 are unknown positive variance components. Complete the blanks in the ANOVA table below with the appropriate formulas.

Source	Degrees of Freedom	Sum of Squares
Block	----------	----------
Treatment	----------	----------
Error	----------	----------
Corrected Total	n-1	$\sum_{i=1}^{t} \sum_{j=1}^{r} (y_{ij} - \bar{y}_{..})^2$

Here $n = t \cdot r$ and $\bar{y}_{..} \equiv \frac{1}{n} \sum_{i=1}^{t} \sum_{j=1}^{r} y_{ij}$.

2. Assuming that model (1) is true, derive the expected value of the block mean square.

3. Based on your work in part 2, provide a formula for the method-of-moments estimator of σ_b^2.

4. Provide a formula for the F-statistic used to test $H_0 : \tau_1 = \cdots = \tau_t$.

5. Assume that model (1) holds and show that for the case of $t = 2$, the F-test of $H_0 : \tau_1 = \cdots = \tau_t$ is equivalent to the paired-data t-test of $\tau_1 - \tau_2$ when the two observations from each block are paired together.

6. Suppose researchers were interested in studying the effect of a viral infection on the activity level of a certain plant gene. The plant gene is believed to help plants resist damage from a virus. Two leaves of comparable developmental stage were identified for each of 12 individually potted plants. One of the leaves on each plant was randomly selected for infection with a topical gel that contained a plant virus. The other leaf received the topical gel without the virus to serve as an uninfected control. A quantitative measure of the gene's activity level was obtained for each leaf 24 hours after treatment. Data (expressed as integers for ease of calculation) are provided below.

Plant	Infected Leaf	Uninfected Leaf
1	1	3
2	2	2
3	4	3
4	5	4
5	8	3
6	4	1
7	9	4
8	12	5
9	14	5
10	14	4
11	13	3
12	17	6

Assuming model (1) holds for this experiment with plants as blocks, use the data above to conduct a test of $H_0 : \tau_1 = \tau_2$ at the 0.05 level. Provide a test statistic, its degrees of freedom, and a conclusion in terms the biologists conducting the experiment will be able to understand (even if they have had little training in statistics).

7. Suppose that the researchers were actually interested in studying the effect of both soil moisture and the viral infection on the activity level of the gene. Previous research has shown that the virus is able to inflict greater damage to plants under dry conditions than under normal soil moisture conditions. The researchers would like to see if they can better understand this phenomenon by examining the activity level of the gene in the presence and absence of the virus and under varying soil moisture levels. Three soil moisture levels (very low, low, and normal) were randomly assigned to the 12 individually potted plants, such that four plants were treated with each moisture level. As described above, two leaves of comparable developmental stage were identified for each plant. One of the leaves on each plant was randomly selected for infection with a topical gel that contained a plant virus. The other leaf received the topical gel without

the virus to serve as an uninfected control. After the soil moisture and viral treatments had been applied for 24 hours, a quantitative measure of the gene's activity level was obtained for each leaf. The previous data table is presented below with additional information describing the soil moisture treatment received by each plant.

Plant	Soil Moisture	Infected Leaf	Uninfected Leaf
1	very low	1	3
2	very low	2	2
3	very low	4	3
4	very low	5	4
5	low	8	3
6	low	4	1
7	low	9	4
8	low	12	5
9	normal	14	5
10	normal	14	4
11	normal	13	3
12	normal	17	6

Let y_{ijk} denote the quantitative measure of gene activity in the leaf that received viral treatment j from the k^{th} plant treated with the i^{th} soil moisture level ($i = 1, 2, 3; j = 1, 2; k = 1, 2, 3, 4$). Consider the following model for the data from this experiment.

$$y_{ijk} = \mu_{ij} + p_{ik} + e_{ijk}, \tag{2}$$

where μ_{ij} ($i = 1, 2, 3; j = 1, 2$) are unknown real-valued parameters; p_{ik} ($i = 1, 2, 3; k = 1, 2, 3, 4$) are identically distributed $N(0, \sigma_p^2)$ random variables; e_{ijk} are identically distributed $N(0, \sigma_e^2)$ random variables; all p_{ik} and e_{ijk} are jointly independent; and σ_p^2 and σ_e^2 are unknown positive variance components.

(a) Under model (2), find an expression for the correlation between two observations corresponding to two leaves from a single plant.

(b) The researchers were interested in testing for interaction between soil moisture and viral treatment. In terms of model (2) parameters, state the null hypothesis for the test of interaction.

(c) For $i = 1, 2, 3$ and $k = 1, 2, 3, 4$; let $d_{ik} = y_{i1k} - y_{i2k}$, where $j = 1$ corresponds to infection with the virus and $j = 2$ corresponds to control. Assuming that model (2) holds for the y_{ijk} values, write down a model for the d_{ik} values.

(d) Conduct a test for interaction. Provide a test statistic, its degrees of freedom, and a conclusion in terms the biologists conducting the experiment should be able to understand.

(e) Suppose the very low, low, and normal soil moisture levels correspond to 1, 2, and 6 units of water per day, respectively. Consider the following new model for the data.

$$y_{ijk} = \alpha_j + \beta_j x_i + p_{ik} + e_{ijk}, \tag{3}$$

where $\alpha_1, \alpha_2, \beta_1$, and β_2 are unknown real valued parameters; $x_1 = 1$, $x_2 = 2$, and $x_3 = 6$; and all other terms are as defined for model (2). Assuming model (3) holds, explain how a single test of slope equal zero from a simple linear regression can be used to test $H_0 : \beta_1 = \beta_2$.

Part B. You have been asked to develop equations to predict the cost to build a power plant. You have available data on the construction costs for 32 plants constructed in the US between 1990 and 2002. 26 plants were designed and built using a traditional process with different firms responsible for design and contruction. 6 plants were designed and built using a 'turnkey' process with one firm responsible for both design and construction. To remove the influence of inflation, costs are expressed as '1996' dollars. The following information is available for each of the 32 plants:

Variable	Units	Description
cost	million $	Cost, adjusted to 1996
date	year	Date construction permit issued, 1996.5 is July 1 1996
Tapp	month	Time between application for and issuance of construction permit
Top	month	Time to between issuance of operating permit and construction permit
power	MW	Power plant capacity
exp	count	number of previous power plants built by that architect/engineer
prev	0/1	1 if there was a previous power plant on that site, 0 otherwise
NE	0/1	1 if the plant was built in the north-east region, 0 otherwise
turnkey	0/1	1 if turnkey process used, 0 otherwise

Here is a plot of cost and date.

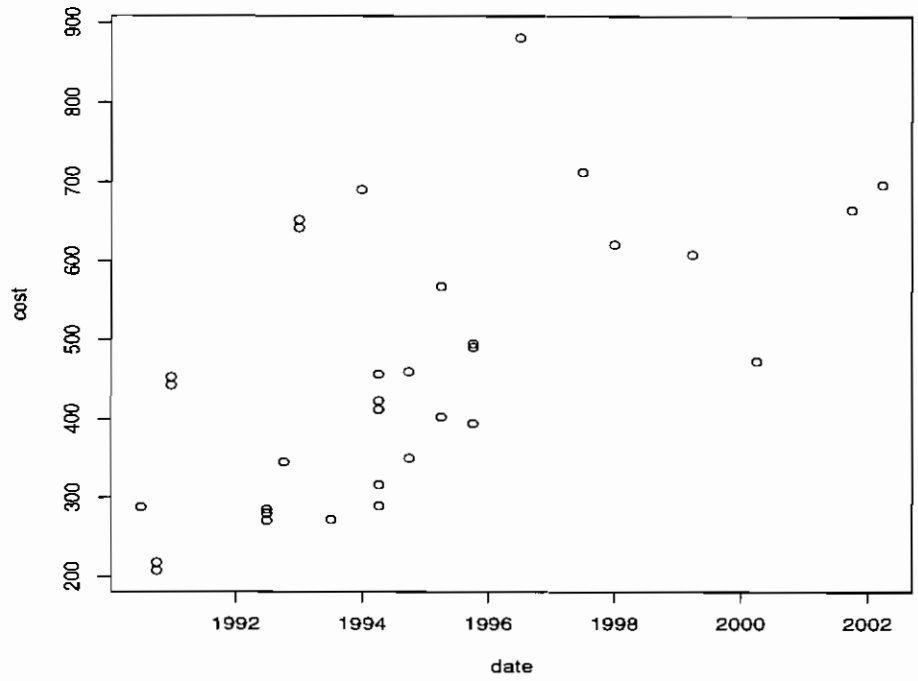


Figure 1: Plot of cost (MW) and date of construction permit for 32 power plants.

8. One interpretation of the pattern in this plot is that the cost increases over time until 1996 and is constant after that. Write down a regression model that describes this interpretation. Your model should:

   be linear in the parameters,

   be continuous, in mean, between 1990 and 2002, and

   make usual least-squares assumptions about error characteristics.

   Make sure you define the parameters and variables in your model. **Do not** estimate the parameters in your model.

9. Is your model from question 8 a better description of the conditional mean cost, E $(cost \mid date)$, than the simple linear regression model, E $(cost \mid date) = \beta_0 + \beta_1 date$? If possible, conduct an appropriate test and provide a p-value. Here are some statistics that might be useful. A plot of the data and fitted regression lines is on the next page.

Model	SS model	SS error
Model in Question 8	70,268	515,059
E $(cost \mid date) = \beta_0 + \beta_1 date$	117,690	467,637

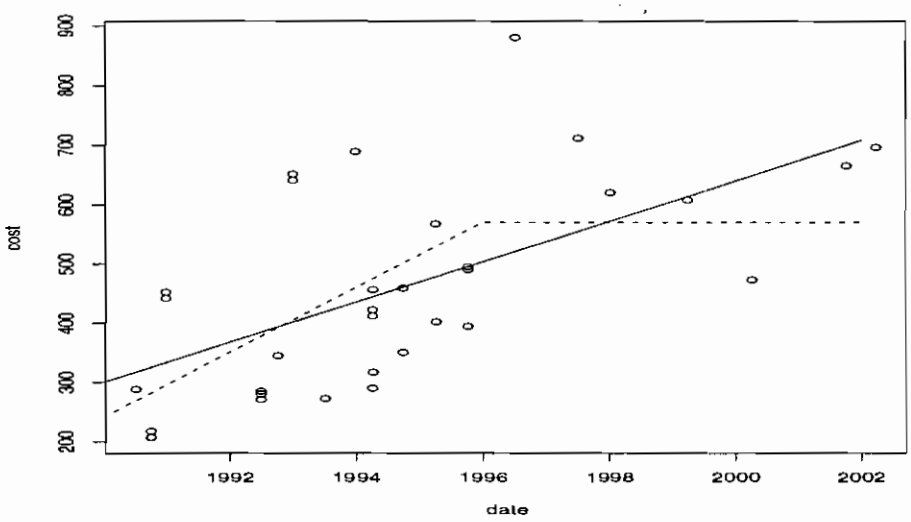


Figure 2: Cost data, with fitted lines for the model in question 8 (dashed) and the linear regression model in question 9 (solid)

10. The models in questions 8 and 9 ignore the seven other possible explanatory variables (e.g. Tapp, power, ...). You are asked to choose a set of explanatory variables for a model to predict the cost of a proposed new power plant. Here is some information that might be useful.

C_p	R^2	AIC	BIC	Variables in model
6.57	0.8350	286.0600	294.3895	date Top power prev NE Cool exp
6.89	0.8181	287.1881	293.0204	date Top power prev NE exp
8.14	0.8382	287.4372	297.0548	date Tapp Top power prev NE Cool exp
8.33	0.7929	289.3282	292.8473	date power NE Cool exp
8.47	0.8357	287.9171	297.1848	date Top power prev NE Cool exp turnkey
8.52	0.8208	288.7037	295.3973	date Tapp Top power prev NE exp
8.72	0.8047	289.4598	294.1440	date Top power NE Cool exp
8.93	0.8032	289.7018	294.2666	date power prev NE Cool exp
9.47	0.7992	290.3447	294.5954	date Top power prev NE turnkey
9.80	0.7968	290.7203	294.7893	date power NE Cool exp turnkey
10.13	0.8090	290.7486	296.2352	date Top power NE Cool exp turnkey
10.20	0.7939	291.1823	295.0297	date Tapp power NE Cool exp

What variables will you include in the model? Explain your choice.

11. Because of non-statistical knowledge about the problem, you are asked to fit the model with 5 variables: power, NE, Cool, exp, and date. For that model, the estimated regression slope for power, $\hat{\beta}_{power}$ is 0.462.
    What are the units for $\hat{\beta}_{power}$?
    Describe, for a non-statistician, what $\hat{\beta}_{power}$ is estimating. I.e. explain what the value of 0.462 represents?

12. The model with the 5 variables, power, NE, Cool, exp, and date, is intended to predict cost.

    The ANOVA table and Type III F tests of individual components are included below. Some diagnostic plots (residual vs predicted value, and Cook's distance, DFFITS, and DFBETA plots for each observation) are on the next page.

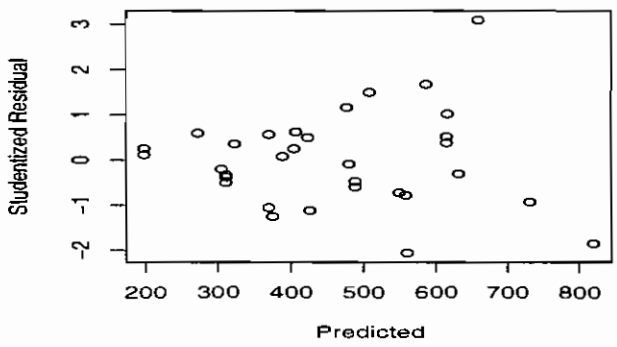
    a) If the model is used only to predict cost, what is the most serious concern using this model for these data? Briefly explain your choice.
    b) If the model is also used to estimate a 95% prediction interval for cost, what is the most serious concern using this model for these data? Briefly explain your choice.

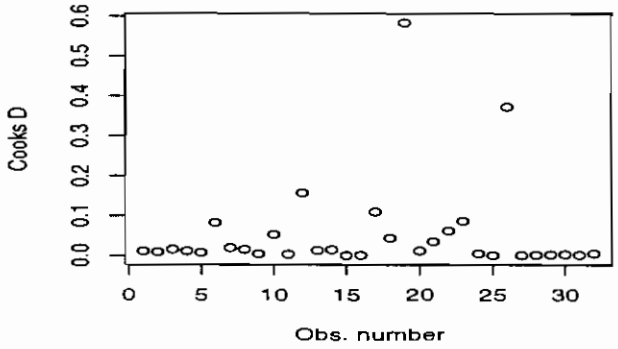
    Note: I do not want a catalog of concerns. There may be many; there may be none. If you have concerns, I want you to prioritize your concern(s) and tell me which one you believe is the most serious. If you have none, you may answer 'no concerns'.

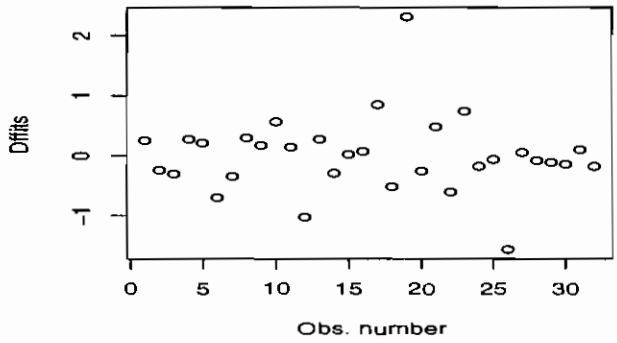
Dependent Variable: cost

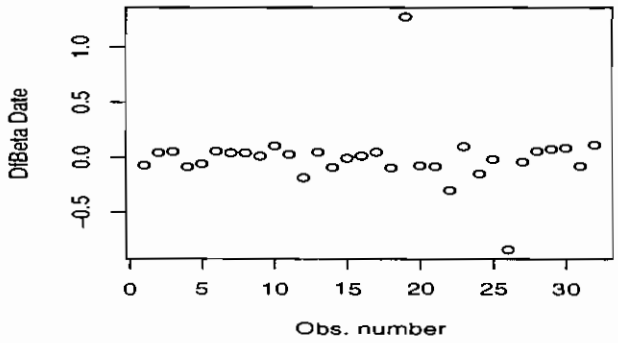
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	711406.2961	142281.2592	19.91	<.0001
Error	26	185766.0126	7144.8466		
Corrected Total	31	897172.3087			

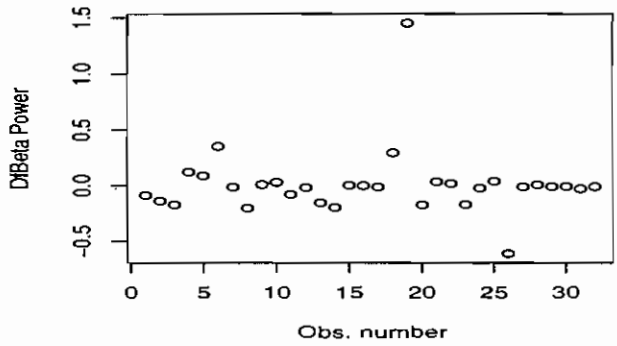
Source	DF	Type III SS	Mean Square	F Value	Pr > F
power	1	225538.1710	225538.1710	31.57	<.0001
NE	1	104251.3683	104251.3683	14.59	0.0007
Cool	1	39959.1748	39959.1748	5.59	0.0258
exp	1	57847.5802	57847.5802	8.10	0.0085
date	1	346419.4994	346419.4994	48.49	<.0001

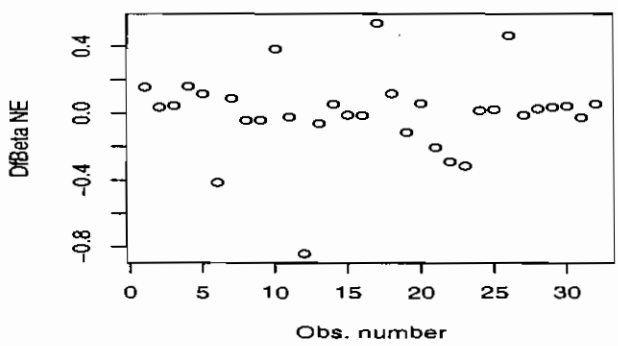

Studentized Residual
Predicted

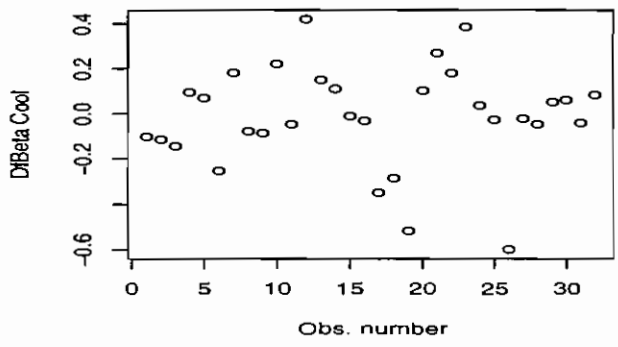

Cooks D
Obs. number

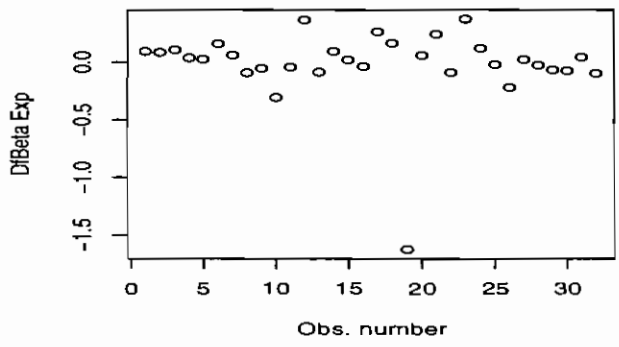

Dffits
Obs. number


DfBeta Date
Obs. number


DfBeta Power
Obs. number


DfBeta NE
Obs. number


DfBeta Cool
Obs. number


DfBeta Exp
Obs. number

13. The power plants in this study fall into two groups: those in the northeast region of the US and those elsewhere in the US. You have reason to suspect that one or both of the regression coefficients for power, β_{power}, and Cool, β_{cool}, are different in the two regions. Define $\beta_{power,NE}$ as the regression coefficient for power for plants in the NE region of the US and use analogous notation for the other region and regression coefficient. Construct a test of H0: $\beta_{power,NE} = \beta_{power,elsewhere}$ and $\beta_{cool,NE} = \beta_{cool,elsewhere}$. Report your F statistic and an approximate p-value.

Some relevant (and some not relevant) SAS output is attached. Remember that the NE variable is coded as 1 for plants constructed in the NE region of the US and 0 for plants constructed elsewhere. The variable labeled NE*power is the product of the NE variable and the power variable. Other products are named similarly. Proc GLM was used to fit all models.

Dependent Variable: cost

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	711406.2961	142281.2592	19.91	<.0001
Error	26	185766.0126	7144.8466		
Corrected Total	31	897172.3087			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
power	1	225538.1710	225538.1710	31.57	<.0001
NE	1	104251.3683	104251.3683	14.59	0.0007
Cool	1	39959.1748	39959.1748	5.59	0.0258
exp	1	57847.5802	57847.5802	8.10	0.0085
date	1	346419.4994	346419.4994	48.49	<.0001

Dependent Variable: cost

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	722201.2276	103171.6039	14.15	<.0001
Error	24	174971.0811	7290.4617		
Corrected Total	31	897172.3087			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
power	1	147637.0280	147637.0280	20.25	0.0001
NE	1	801.5881	801.5881	0.11	0.7431
Cool	1	19856.5241	19856.5241	2.72	0.1119
exp	1	55626.4254	55626.4254	7.63	0.0108
date	1	348868.1419	348868.1419	47.85	<.0001
NE*power	1	7849.6474	7849.6474	1.08	0.3098
NE*Cool	1	1270.5417	1270.5417	0.17	0.6801

Dependent Variable: cost

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	728231.5342	104033.0763	14.78	<.0001
Error	24	168940.7745	7039.1989		
Corrected Total	31	897172.3087			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
power	1	195150.2002	195150.2002	27.72	<.0001
NE	1	10465.1218	10465.1218	1.49	0.2346
Cool	1	34112.0734	34112.0734	4.85	0.0376
exp	1	69391.0766	69391.0766	9.86	0.0044
date	1	328784.2737	328784.2737	46.71	<.0001
NE*exp	1	15875.1339	15875.1339	2.26	0.1462
NE*date	1	10425.4105	10425.4105	1.48	0.2354

Dependent Variable: cost

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	732216.2608	81357.3623	10.85	<.0001
Error	22	164956.0479	7498.0022		
Corrected Total	31	897172.3087			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
power	1	151199.1339	151199.1339	20.17	0.0002
NE	1	3623.2511	3623.2511	0.48	0.4942
Cool	1	21277.8530	21277.8530	2.84	0.1062
exp	1	65313.9734	65313.9734	8.71	0.0074
date	1	329431.9037	329431.9037	43.94	<.0001
NE*power	1	285.8101	285.8101	0.04	0.8470
NE*Cool	1	2550.5246	2550.5246	0.34	0.5657
NE*exp	1	9873.6763	9873.6763	1.32	0.2635
NE*date	1	3638.3520	3638.3520	0.49	0.4934

1.

Source	DF	Sum of Squares
Block	$r-1$	$t\sum_{j=1}^{r}(\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot})^2$
Treatment	$t-1$	$r\sum_{i=1}^{t}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2$
Error	$(t-1)(r-1)$	$\sum_{i=1}^{t}\sum_{j=1}^{r}(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y}_{\cdot\cdot})^2$

2. Note that $\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot}=b_j-\bar{b}_{\cdot}+\bar{e}_{\cdot j}-\bar{e}_{\cdot\cdot}$. Thus, $E(\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot})=0$ and

$$\begin{aligned}
E(\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot})^2 &= \mathrm{Var}(\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot}) \\
&= \mathrm{Var}(b_j-\bar{b}_{\cdot}+\bar{e}_{\cdot j}-\bar{e}_{\cdot\cdot}) \\
&= \mathrm{Var}(b_j-\bar{b}_{\cdot})+\mathrm{Var}(\bar{e}_{\cdot j}-\bar{e}_{\cdot\cdot}) \\
&= \sigma_b^2+\frac{\sigma_b^2}{r}-\frac{2\sigma_b^2}{r}+\frac{\sigma_e^2}{t}+\frac{\sigma_e^2}{tr}-\frac{2\sigma_e^2}{tr} \\
&= \frac{(r-1)\sigma_b^2}{r}+\frac{(r-1)\sigma_e^2}{tr}.
\end{aligned}$$

Thus, $E(\mathrm{MSBlock})=\frac{1}{r-1}t\sum_{j=1}^{r}\{\frac{(r-1)\sigma_b^2}{r}+\frac{(r-1)\sigma_e^2}{tr}\}=t\sigma_b^2+\sigma_e^2$.

This problem can also be done using the fact that

$$E\left\{\sum_{j=1}^{n}(X_i-\bar{X})^2\right\}=(n-1)\sigma^2$$

when $X_1,\ldots,X_n$ are i.i.d. random variables with variance σ^2. If we identify $\bar{y}_{\cdot j}$ with X_j and note that $\mathrm{Var}(\bar{y}_{\cdot j})=\sigma_b^2+\frac{\sigma_e^2}{t}$, then the result follows.

3. $\frac{\mathrm{MSBlock}-\mathrm{MSError}}{t}$, where MS stands for sum of squares divided by degrees of freedom.

4. $F=\frac{\mathrm{MSTreatment}}{\mathrm{MSError}}$

5. Let $d_j=y_{1j}-y_{2j}$ for $j=1,\ldots,r$. Let

$$\bar{d}_{\cdot}=\frac{1}{r}\sum_{j=1}^{r}d_j,\quad s_d=\sqrt{\frac{1}{r-1}\sum_{j=1}^{r}(d_j-\bar{d}_{\cdot})^2},\quad \text{and}\quad t=\frac{\bar{d}_{\cdot}}{s_d/\sqrt{r}}.$$

The α-level paired-data t-test results in rejection of $H_0:\tau_1=\tau_2$ if and only if $|t|\geq t_{\alpha/2;r-1}$, which is equivalent to $t^2\geq F_{\alpha;1,r-1}$ because the square of a t-distributed random variable is F-distributed with 1 numerator degree of freedom. The α-level F-test of $H_0:\tau_1=\tau_2$ from the ANOVA results in rejection of H_0 if and only if $F\geq F_{\alpha;1,r-1}$, where F is as defined in part 4. Thus, the equivalence of the test will follow if we can show that $t^2=F$.

Note that

$$\begin{aligned}\sum_{i=1}^{t}(\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2 &= (\bar{y}_{1\cdot}-\bar{y}_{\cdot\cdot})^2+(\bar{y}_{2\cdot}-\bar{y}_{\cdot\cdot})^2\\ &= \left(\frac{\bar{y}_{1\cdot}-\bar{y}_{2\cdot}}{2}\right)^2+\left(\frac{\bar{y}_{2\cdot}-\bar{y}_{1\cdot}}{2}\right)^2\\ &= \frac{1}{2}(\bar{y}_{1\cdot}-\bar{y}_{2\cdot})^2=\frac{1}{2}\bar{d}_{\cdot}^2.\end{aligned}$$

Also

$$\begin{aligned}\sum_{i=1}^{t}(y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y}_{\cdot\cdot})^2 &= (y_{1j}-\bar{y}_{\cdot j}-\bar{y}_{1\cdot}+\bar{y}_{\cdot\cdot})^2+(y_{2j}-\bar{y}_{\cdot j}-\bar{y}_{2\cdot}+\bar{y}_{\cdot\cdot})^2\\ &= \left(\frac{(y_{1j}-y_{2j})-(\bar{y}_{1\cdot}-\bar{y}_{2\cdot})}{2}\right)^2+\left(\frac{(y_{2j}-y_{1j})-(\bar{y}_{2\cdot}-\bar{y}_{1\cdot})}{2}\right)^2\\ &= \frac{1}{2}\{(y_{1j}-y_{2j})-(\bar{y}_{1\cdot}-\bar{y}_{2\cdot})\}^2\\ &= \frac{1}{2}(d_j-\bar{d}_{\cdot})^2.\end{aligned}$$

Thus,

$$F=\frac{\frac{1}{2}r\bar{d}_{\cdot}^2}{\frac{\sum_{j=1}^{r}(d_j-\bar{d}_{\cdot})^2}{2(r-1)}}=\frac{\bar{d}_{\cdot}^2}{s_d^2/r}=t^2.$$

6. The differences are -2,0,1,1,5,3,5,7,9,10,10,11.

$$\bar{d}_{\cdot}=5 \quad s_d^2=216/11 \quad t=\frac{5}{\sqrt{\frac{216}{(11)(12)}}}\approx 3.909. \quad d.f.=11$$

The p-value obtained by comparing 3.909 to a t-distribution with 11 d.f. is < 0.0025. Thus, the data provide evidence that the virus caused an increase in the mean activity level of the gene relative to the mean activity level in control leaves.

7. (a) Note that

$$\text{cov}(y_{i1k},y_{i2k})=\text{cov}(p_{ik}+e_{i1k},p_{ik}+e_{i2k})=\text{cov}(p_{ik},p_{ik})=\sigma_p^2$$

and

$$\text{var}(y_{i1k})=\text{var}(y_{i2k})=\sigma_p^2+\sigma_e^2.$$

Thus, $\text{corr}(y_{i1k},y_{i2k})=\frac{\sigma_p^2}{\sigma_p^2+\sigma_e^2}$.

(b) $H_0:\mu_{11}-\mu_{12}=\mu_{21}-\mu_{22}=\mu_{31}-\mu_{32}$

(c) $d_{ik} = \mu_{i1} - \mu_{i2} + e_{i1k} - e_{i2k} = \delta_i + \varepsilon_{ik}$, where $\delta_i = \mu_{i1} - \mu_{i2}$ $(i = 1, 2, 3)$ and ε_{ik} $(i = 1, 2, 3; j = 1, 2, 3, 4)$ are i.i.d. $N(0, 2\sigma_e^2)$.

(d) A simple one-way ANOVA can be used to test $H_0 : \delta_1 = \delta_2 = \delta_3$, which is equivalent to the "no interaction" null hypothesis stated in part (b).

$$\bar{d}_{1\cdot} = 0 \quad \bar{d}_{2\cdot} = 5 \quad \bar{d}_{3\cdot} = 10 \quad \bar{d}_{\cdot\cdot} = 5$$

$$MST = \frac{4}{3-1}\{(0-5)^2 + (5-5)^2 + (10-5)^2\}$$

$$MSE = \sum_{i=1}^{3}\sum_{k=1}^{4}(d_{ik} - \bar{d}_{i\cdot})^2/9 = 16/9$$

$$F = \text{MST/MSE} = 56.25.$$

Comparing to an F-distribution with 2 and 9 d.f., the p-value is very small (< 0.0001). Thus, the data provide evidence that the effect of the virus on the mean activity level of the gene was not the same for all moisture levels. It appears that the virus caused the greatest increase in the mean activity level of the gene relative to the control when moisture levels were normal. The effect of the virus seemed to diminish at lower moisture levels. At very low moisture levels, for example, there was no evidence that the virus affected the activity level of the gene.

(e) Let

$$d_{ik} = y_{i1k} - y_{i2k} = \alpha_1 - \alpha_2 + (\beta_1 - \beta_2)x_i + e_{i1k} - e_{i2k} = \theta + \gamma x_i + \varepsilon_{ik},$$

where $\theta \equiv \alpha_1 - \alpha_2$, $\gamma \equiv \beta_1 - \beta_2$, and $\varepsilon_{ik} \equiv e_{i1k} - e_{i2k}$ $(i = 1, 2, 3; k = 1, 2, 3, 4)$ i.i.d. $N(0, \sigma^2)$ with $\sigma^2 = 2\sigma_e^2$. In this formulation, the simple linear regression test of $H_0 : \gamma = 0$ will provide a test of $H_0 : \beta_1 = \beta_2$.

8. This can be written in various ways, including:
$cost_i = \begin{cases} \beta_0 + \beta_1 Date_i + \epsilon_i & Date_i \leq 1996 \\ \beta_0 + \beta_1 1996 + \epsilon_i & Date_i > 1996 \end{cases}$, or
$cost_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where X_i is min$(Date_i, 1996)$.
In either form, $\epsilon_i \sim indep N(0, \sigma^2)$, and i indexes the observation, (i.e. the power plant).

9. There is no formal test, since the two models are not nested. Both models have 2 unknown parameters (for the conditional mean, E$Cost_i \mid Date_i$), so SS_{model}, SS_{error}, AIC and BIC lead to equivalent decision rules. The linear regression model fits better than the model from part a.

10. C_p, AIC or BIC are all reasonable decision criteria. In all three cases, you are looking for the smallest value. C_p and AIC suggest the 7 variable model: date Top power prev NE Cool exp. BIC suggests the five variable model: date power NE Cool exp. However, there are many different models with AIC or BIC values that are close (within 2) to the minimum value.

11. The units are million $ / MW.

    The regression coefficient is the estimated increase in cost per MW increase in power, when the other four variables in the regression model (NE, Cool, exp, and date) are held constant.

12. If the goal is only prediction, my most serious concern is the two influential points at observation #'s 19 and 26. The Cooks Distance and DFFITs values are very large for both observations. There are lots of other potential issues, (e.g. the unequal variances and some large DFBeta values), but these are not directly relevant to the quality of predictions.
    If the goal is a prediction interval, the most serious issue is the unequal variance. The prediction interval will be too wide for predictions of small costs and too narrow for predictions of large costs.

13. The desired test is an F test, based on the change in Sums of Squares Error between two models. One model has unequal coefficients in the two regions; the other has equal coefficients. There are two pairs that could be used: models 1 and 2 or models 3 and 4. They give similar results (at least here).

    F = (185766 - 174971) / (26-24) / (174971 / 24) = 0.74. This is compared to an F 2, 26 distribution. The p-value is ca 0.5. There is no evidence that the two regions have different coefficients.

Gene Expression Experiment

During a plant science experiment, gene expression values are measured for three different plant lines. The plant lines are two genetically modified lines, called treatment 1 and treatment 2, and a control line of the wild type plant. It is expected that due to the genetic modification plants react differently, which will show as differences between gene expression values. In this particular experiment values for each gene were extracted independently three times from three separate plants.

The values for one particular gene are given in the table below:

Gene 254098_at									
	Control			Treatment 1			Treatment 2		
Value	11.3	11.1	11.3	10.4	10.5	10.4	4.3	5.3	4.3

1. For one particular gene, we might consider the following model:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \qquad \text{Model } A$$

where y_{ij} is the observed gene expression value under treatment i and replicate j, $i = 0, 1, 2$ and $j = 1, 2, 3$ with $i = 0$ Wildtype, $i = 1$ treatment 1, and $i = 2$ treatment 2

μ expected average gene expression value

τ_i expected effect of treatment i on the gene expression value (with $i = 0$ Wildtype)

ε_{ij} random error, $\varepsilon_{ij} \sim N(0, \sigma^2)$ independent and identically distributed.

a) The plant scientist conducting this experiment is particularly interested in genes with mean expression level that differs between any pair of treatments. The three pairwise differences are a set of contrasts. Show that all of these contrasts are estimable in model A without any further restrictions. Carefully define all notation used in your solution.

b) Particularly interesting genes are the ones with at least one significant contrast $\tau_{i_1} - \tau_{i_2}$, with $i_1, i_2, = 0, 1, 2, ; i_1 \neq i_2$. Construct an F test that will allow you to identify these "interesting" genes. State null and alternative hypotheses and construct an F statistic. Make sure to mention all relevant distributional assumptions.

Based on the R code in the back, find the p value for gene 254098_at (given in the table above). Make sure to explain your answer.

c) Plant scientists often want to report results in terms of the *fold change*. The fold change between treatments ℓ and j is given as $2^{\tau_\ell - \tau_k}$, $\ell, k = 0, 1, 2$. Give the estimated fold change between treatment 2 and the control for gene 254098_at. Derive a 95% confidence interval.

d) Due to some defect in the scanner measuring the intensities, the second replicate in treatment 2 turns out to have a ten times higher variance than all the other measurements. Adjust model (A) to this situation The plant scientist asks whether you want to

drop this replicate from the analysis to get a "better" estimates. In what (statistical) sense is his solution better compared to an LS estimate?

Derive the BLUE estimate for $\mu + \tau_2$ and a 95% C.I..

e) With a modern high-throughput method plant scientists are able to evaluate 22.810 different genes at the same time. You fit model A to each of these genes and use the F test in $9B$) to decide, whether a gene is interesting or not. This way you are able to identify a lot of interesting genes:

Significance level	0.05	0.01	0.001	0.0001	0.00001	0.000001
# significant genes	8,214	5,637	1,952	1,183	776	523

Discuss the method and comment on these numbers. What are you going to report to the plant scientist?

2. Previous research has come up with a list of 15 genes that are of particular interest. The profiles of their gene expressions across all treatments and replicate measurements can be seen in figure 1.

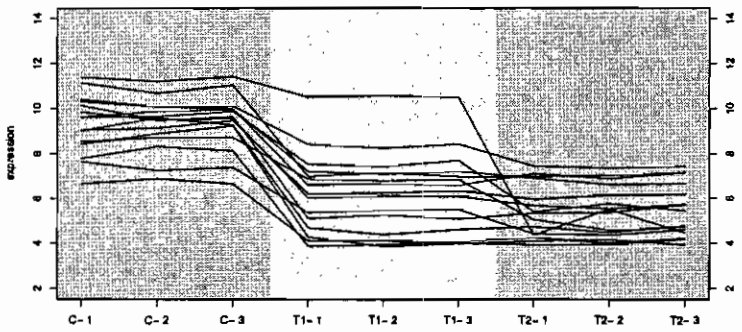


Figure 1: *Profiles of gene expression values of 15 previously identified genes.*

f) As an extension of model A, the following model includes all 15 genes in a single model:

$$y_{ijk} = \mu + \tau_i + m_k + T_{ik} + \varepsilon_{ijk}, \qquad \text{Model } B$$

where y_{ijk} is the observed gene expression value of gene k under treatment i and replicate j, $i = 0, 1, 2$ and $j = 1, 2, 3$ with $i = 0$ Wildtype, $i = 1$ treatment 1, and $i = 2$ treatment 2 and $k = 1, ..., 15$

μ expected average gene expression value

τ_i expected effect of treatment i on the gene expression value (with $i = 0$ Wildtype)

m_k random effect for each gene, $m_k \sim N(0, \sigma_m^2)$ i.i.d.

T_{ik} random effect for each gene and treatment, $T_{ik} \sim N(0, \sigma_T^2)$ i.i.d.

ε_{ijk} random error, $\varepsilon_{ijk} \sim N(0, \sigma^2)$ independent and identically distributed.

All random effects are pairwise independent.

For a fixed gene derive, if possible, the covariance between expression values in model B for

(i) same treatment and same replicate

(ii) same treatment and different replicates

(iii) different treatment and same replicate

(iv) different treatment and different replicate

g) R output for model B can be found in the back referred to as `mB`. Additionally you can find output for model C (referred to as `mC`). Describe in words the difference between the models. Sketch profiles for expected gene expression values for both models. Based on the anova table, pick the better model.

h) For model B, give a formula for predicting the random effects and describe properties of the predictor.

3. Looking more closely at the gene expression values, it occurs to you that a Gamma distribution might provide a better fit to the data than a normal distribution. The probability density function of a gamma distribution is given as

$$f(x;k,\beta) = x^{k-1}\frac{e^{-x/\beta}}{\beta^k\Gamma(k)} \text{for } x > 0$$

where $k > 0$ is the shape parameter and $\beta > 0$ is the scale parameter. Γ is the gamma function defined as

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt.$$

i) For $k = 1$ show that the Gamma distribution is a member of the exponential family.

j) Extend model D to a Generalized Linear Model with mixed effects (GLMM), assuming that gene expression values are distributed according to a Gamma distribution. Call this model E. Make sure to cover all parts of a generalized linear model. For $k = 1$ derive the natural link.

k) Model E is fit as a GLMM using the natural link.
Figure 2 gives two diagnostic plots:

(i) a plot of residuals vs predicted values,

(ii) a normal quantile-quantile plot of the observations,

For each of these plots, discuss what that plot tells you, if anything, about the appropriateness of model E for these data.

l) The variance of the residuals σ^2 in model E is estimated as 0.259. In order to get a confidence interval for this quantity, somebody suggests to use a non-parametric bootstrap. Describe how a non-parametric bootstrap works. Why does a mere sampling from the observations not work for this data? Suggest possible remedies. The R output in the back shows a stem and leaf plot of 100 bootstrap estimates for σ^2. Based on the output give a 90% confidence interval and explain.

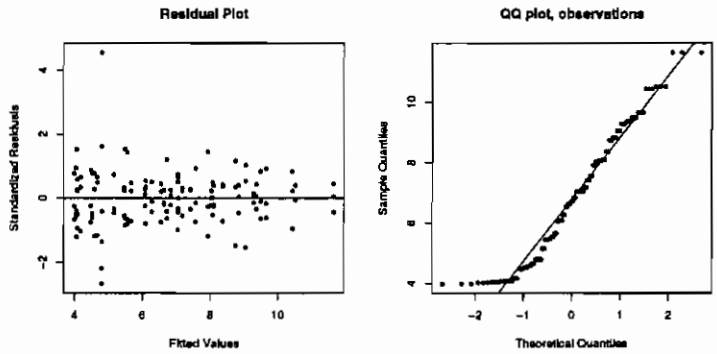


Figure 2: *Diagnostic plots for model E. On the left a residual plot, on the right a normal quantile plot of the observed values.*

R output

```
> Y
 C-1  C-2  C-3 T1-1 T1-2 T1-3 T2-1 T2-2 T2-3
11.3 11.1 11.3 10.4 10.5 10.4  4.3  5.3  4.3
> tau <- factor(rep(0:2,each=3))
> mA <- lm(Y~1+tau)
> summary(mA)

Call:
lm(formula = Y ~ 1 + tau)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33333 -0.13333 -0.03333  0.06667  0.66667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.2333     0.1972  56.963 1.97e-09 ***
tau1         -0.8000     0.2789  -2.869   0.0285 *
tau2         -6.6000     0.2789 -23.666 3.74e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3416 on 6 degrees of freedom
Multiple R-Squared: 0.9911,Adjusted R-squared: 0.9881
F-statistic: 333.6 on 2 and 6 DF,  p-value: 7.08e-07

>
> round(fitted(mA),1)
 C-1  C-2  C-3 T1-1 T1-2 T1-3 T2-1 T2-2 T2-3
```

```
11.2 11.2 11.2 10.4 10.4 10.4  4.6  4.6  4.6
>
> # Part 2
> library(nlme)
> Tau <- rep(tau,15)
> gene <- rep(1:15, each=9)
>
> round(expr,1)
           C-1  C-2  C-3 T1-1 T1-2 T1-3 T2-1 T2-2 T2-3
249727_at  8.5  8.7  8.7  6.9  6.8  6.8  5.1  4.6  4.7
259765_at  9.6  9.6  9.9  7.5  7.4  7.7  5.4  5.8  5.4
255016_at  9.0  9.6  9.3  4.7  4.4  4.6  4.7  4.5  4.8
254098_at 11.4 11.2 11.4 10.5 10.6 10.5  4.4  5.5  4.5
255302_at  7.8  8.3  8.1  3.9  3.9  4.1  4.3  4.0  4.2
259789_at  9.8  9.5  9.6  5.1  5.3  5.1  5.3  5.4  5.7
258239_at 11.1 10.7 11.1  7.0  7.2  7.0  7.0  7.0  7.1
262232_at  8.4  8.9  9.2  4.3  3.9  4.0  4.1  4.2  3.9
249894_at 10.1  9.5  9.6  6.0  6.1  6.1  5.7  5.5  5.7
264348_at  9.6  9.9 10.0  7.2  7.1  7.2  6.9  6.6  6.7
265722_at 10.3 10.1 10.0  6.6  6.7  6.6  7.1  6.9  7.2
252863_at 10.4 10.1 10.1  8.4  8.2  8.4  7.5  7.4  7.4
247754_at  7.6  7.3  7.4  5.4  5.5  5.5  4.4  4.4  4.6
252534_at  6.7  6.9  6.7  4.1  4.1  4.0  3.9  4.0  4.2
263715_at  9.0  9.1  9.5  6.2  6.3  6.4  6.0  6.1  6.2
>
> mC <- lme(expr~Tau+1, random=~1|gene)
> mB <- lme(expr~Tau+1, random=~Tau+1|gene)
> anova(mC, mB)
   Model df   AIC   BIC  logLik   Test L.Ratio p-value
mC     1  5 375.3 389.7 -182.65
mB     2 10 162.7 191.5  -71.33 1 vs 2   222.6  <.0001
>
>
> stem.leaf(sigBoot)
1 | 2: represents 0.012
 leaf unit: 0.001
            n: 100
LO: 0.1616
    2    17 | 9
         18 |
    5    19 | 699
   11    20 | 000579
   17    21 | 122489
   24    22 | 1256789
   33    23 | 024455578
   48    24 | 033466677778889
  (13)   25 | 0012335677899
   39    26 | 0000122336777
```

```
26    27 | 00112445678
15    28 | 0166789
 8    29 | 22359
 3    30 | 159
```

Gene Expression Experiment

1.

a)

The design matrix X can be written as

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix},$$

with a column space of the transpose:

$$C(X') = \left\{ \left(\begin{array}{c} b_1 + b_2 + b_3 \\ b_1 \\ b_2 \\ b_3 \end{array} \right) \middle| \, b_1, b_2, b_3 \in R \right\}$$

The contrasts can be written in the form $\tau_j - \tau_k$ for $j \neq k, j, k = 1, 2, 3$, i.e. in linear form this corresponds to some $c*' \in \{(0, -1, 1, 0), (0, -1, 0, 1), (0, 0, -1, 1)\}$. All of these c are estimable: $c' \in C(X')$ with $b_j = -1, b_k = 1$ and $b_i = 0$ for $i, j, k = 1, 2, 3$ and $i \neq j, k$.

b)

$$H_0 : \tau_1 - \tau_0 = \tau_2 - \tau_0 = \tau_2 - \tau_1 = 0$$

versus

$$H_a : \text{ any of these contrasts is not zero.}$$

the null hypothesis can also be written as

$$C\beta = \begin{pmatrix} 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_0 \\ \tau_1 \\ \tau_2 \end{pmatrix} = 0$$

this is not testable (C has only rank 2), but

$$C\beta = \begin{pmatrix} 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \tau_0 \\ \tau_1 \\ \tau_2 \end{pmatrix} = 0$$

is testable.

The test statistic is then

$$\frac{\beta' C' \left(C(X'X)^- C'\right)^{-1} C\beta/2}{Y'(I - P_X)Y/6} \sim F_{2,6,\delta^2}$$

with $\delta^2 = 0$ under H_0.

From the R *output in the back, we see that baseline constraints have been used to fit model A, which makes τ_1 and τ_2 to the contrasts between treatments and control and the goodness of fit statistic compares the significance of adding these contrasts to the model with just the intercept. This is equivalent to the test statistic above, the p value for gene 254098_at is therefore 7.08e-07, i.e. at least one of the contrasts is significantly different from 0.*

c)

$$2^{\tau_0 - \tau_2} = 2^{--6.6} = 97.$$

This gives a 95% C.I. for the fold change as

$$2^{6.6 \pm t_{6,0.975}\sqrt{MSE \cdot (X'X)^-_{33}}}$$

d)

Model Adjustment: change $\varepsilon_{22} \sim N(0, 10\sigma^2)$. This is an Aitken model.

"Find the BLUE for $\mu + \tau_2$ in an Aitken model of the form"

$$Y \sim MVN(X\beta, \sigma^2 D) \text{ with } D = diag\,(1,1,1,1,1,1,1,10,1)$$

estimate of $\bar{y_{2\cdot}}$	*estimate*	*variance*	*comment*
$\frac{1}{3}(y_{21} + y_{22} + y_{23})$	*4.63*	$12/9\ \sigma^2$	*Gauss model estimate*
$\frac{1}{2}(y_{21} + y_{23})$	*4.3*	$1/2\ \sigma^2$	*plant scientist's suggestion*
$\frac{1}{2.1}(y_{21} + 0.1y_{22} + y_{23})$	*4.35*	$1/2.1\ \sigma^2$	*BLUE*

The BLUE can be found by minimizing Var $\left(\frac{1}{2a+1}(y_{21} + ay_{22} + y_{32})\right)$ w.r.t. a.

Compared to the gauss model estimate the estimate from just the first and third replicate has a smaller variance and is therefore better. The BLUE for the Aitken model is given in the third line of the table above (the answer should therefore be "no".)

A 95% C.I is then

$$4.35 \pm t_{6,0.975} \cdot \sqrt{1/2.1} \cdot 0.3416 = 4.35 \pm 0.5775$$

e)

This is a multiple testing situation - in each of the tests we expect to come up with a false

positive in α 100% of the cases. Repeating the same procedure on 22.810 (independent) genes will lead to $22.810 \cdot \alpha$ expected false positives:

Significance level	*0.05*	*0.01*	*0.001*	*0.0001*	*0.00001*	*0.000001*
expected FP	*1,140.5*	*228.1*	*22.81*	*2.281*	*0.2281*	*0.02281*

Among the 8.214 genes we would expect 1.140 genes to show up by chance, i.e. almost 14% of the genes are falsely identified as significant.

To get a similar FP rate with multiple tests as with a single test some adjustment can be made, e.g. a Bonferroni adjustment of the p values multiplies all p values by the number of tests done, i.e. in this case instead of reporting 1,141 genes back, we could report genes that are significant on a 0.05/22,810 level (which gives between 523 and 776 significant genes).

2. Previous research has come up with a list of 15 genes that are of particular interest. The profiles of their gene expressions across all treatments and replicate measurements can be seen in figure 1.

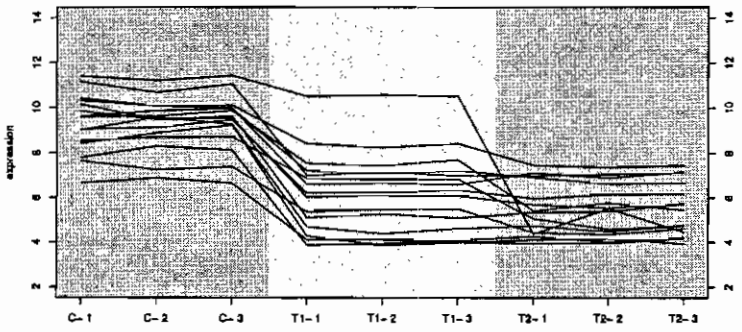


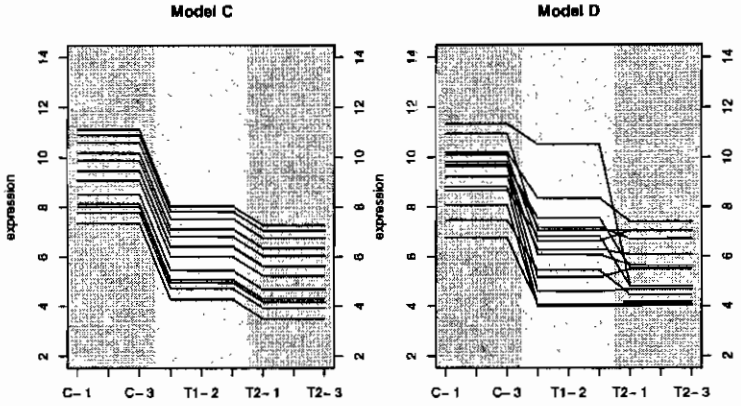
Figure 1: *Profiles of gene expression values of 15 previously identified genes.*

f)

$$Cov\,(y_{ijk}, y_{i'j'k}) = \begin{cases} \sigma_m^2 + \sigma_T^2 + \sigma^2 & for\ i = i', j = j', \\ \sigma_m^2 + \sigma_T^2 & for\ i = i', j \neq j', \\ \sigma_m^2 & for\ i \neq i', j \neq j', \end{cases}$$

g)

Model B has an additional random interaction effect between treatment effects and genes. In model C the profiles of the gene expression values are parallel, whereas they are not necessarily parallel in model B.



Based on the LRT statistic the interaction effect in B is highly significant, B is the better model.

h)

For a mixed effects model of the form $Y = X\beta + Z\gamma + \varepsilon$ *with the usual assumptions,* $\hat{\gamma} = E[\gamma \mid Y] = \sigma_\gamma^2 Z' V^{-1}(Y - \widehat{X\beta})$ *is the BLUP (best unbiased linear predictor) with* $V = Var(Y)$

3.

i)

The natural exponential family has pdf

$$f(y;\theta,\phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} - c(y,\phi)\right),$$

with continuous functions $a(.), b(.)$, *and* $c(.)$.
The gamma pdf with $k = 1$ *can be written in this form as*

$$f(x) = \exp\left(-\log\beta - \log\left(\Gamma(1)\right) - x/\beta\right)$$

use $\theta = 1/\beta$:

$$f(x) = \exp\left(\log\theta - \log\left(\Gamma(1)\right) - x\theta\right)$$

Then $a(\phi) = -1, b(\theta) = \log\theta, c(y,\phi) = \log\Gamma(1)$

j)

Model E:

$$h(EY) = X\beta + Z\gamma$$

where X, β, Z, *and* γ *are defined as for model D, and* Y *has a gamma distribution as defined before.* h *is called the link function, that links the random part* Y *to the systematic part* $X\beta + Z\gamma$.
For $k = 1$ *the natural link* $h(EY) = \theta$. $EY = b'(\theta) = 1/\theta$. *The natural link is therefore the inverse function.*

k)

The theoretic basis of the residual plot in a GLM is that residuals are independent of the fitted values - which they seem to be from looking at the residual plot. We cannot assume normality for the residuals, though, so it does not matter that everything is nicely between (-2,2). The outliers at $x \approx 5$ might be worthwhile to be looked into.
The normal quantile-quantile plot is not relevant.

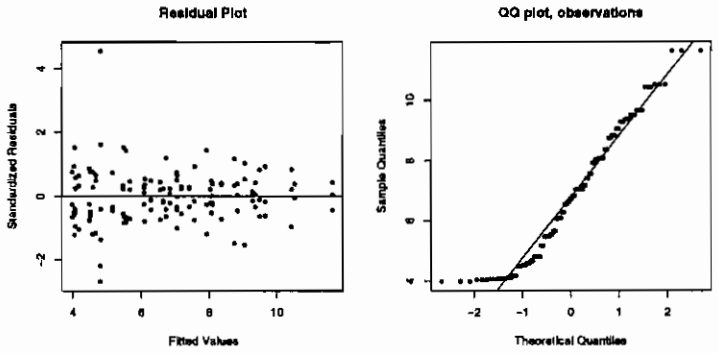


Figure 2: *Diagnostic plots for model E. On the left a residual plot, on the right a normal quantile plot of the observed values.*

l)

Mere sampling from the observations does not work, as the bootstrap is based on the assumption of i.i.d observations. This contradicts our modeling assumption of correlated expression values within genes. A remedy to that is to sample from the 15 genes instead. A 90% confidence interval based on the percentile method is (0.20, 0.293).

R output

```
> Y
 C-1  C-2  C-3 T1-1 T1-2 T1-3 T2-1 T2-2 T2-3
11.3 11.1 11.3 10.4 10.5 10.4  4.3  5.3  4.3
> tau <- factor(rep(0:2,each=3))
> mA <- lm(Y~1+tau)
> summary(mA)

Call:
lm(formula = Y ~ 1 + tau)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33333 -0.13333 -0.03333  0.06667  0.66667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.2333     0.1972  56.963 1.97e-09 ***
tau1         -0.8000     0.2789  -2.869   0.0285 *
tau2         -6.6000     0.2789 -23.666 3.74e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3416 on 6 degrees of freedom
Multiple R-Squared: 0.9911,Adjusted R-squared: 0.9881
F-statistic: 333.6 on 2 and 6 DF,  p-value: 7.08e-07

>
> round(fitted(mA),1)
 C-1  C-2  C-3 T1-1 T1-2 T1-3 T2-1 T2-2 T2-3
11.2 11.2 11.2 10.4 10.4 10.4  4.6  4.6  4.6
>
> # Part 2
> library(nlme)
> Tau <- rep(tau,15)
> gene <- rep(1:15, each=9)
>
> round(expr,1)
          C-1  C-2  C-3 T1-1 T1-2 T1-3 T2-1 T2-2 T2-3
249727_at  8.5  8.7  8.7  6.9  6.8  6.8  5.1  4.6  4.7
259765_at  9.6  9.6  9.9  7.5  7.4  7.7  5.4  5.8  5.4
255016_at  9.0  9.6  9.3  4.7  4.4  4.6  4.7  4.5  4.8
254098_at 11.4 11.2 11.4 10.5 10.6 10.5  4.4  5.5  4.5
255302_at  7.8  8.3  8.1  3.9  3.9  4.1  4.3  4.0  4.2
259789_at  9.8  9.5  9.6  5.1  5.3  5.1  5.3  5.4  5.7
258239_at 11.1 10.7 11.1  7.0  7.2  7.0  7.0  7.0  7.1
```

```
262232_at  8.4  8.9  9.2  4.3  3.9  4.0  4.1  4.2  3.9
249894_at 10.1  9.5  9.6  6.0  6.1  6.1  5.7  5.5  5.7
264348_at  9.6  9.9 10.0  7.2  7.1  7.2  6.9  6.6  6.7
265722_at 10.3 10.1 10.0  6.6  6.7  6.6  7.1  6.9  7.2
252863_at 10.4 10.1 10.1  8.4  8.2  8.4  7.5  7.4  7.4
247754_at  7.6  7.3  7.4  5.4  5.5  5.5  4.4  4.4  4.6
252534_at  6.7  6.9  6.7  4.1  4.1  4.0  3.9  4.0  4.2
263715_at  9.0  9.1  9.5  6.2  6.3  6.4  6.0  6.1  6.2
>
> mC <- lme(expr~Tau+1, random=~1|gene)
> mB <- lme(expr~Tau+1, random=~Tau+1|gene)
> anova(mC, mB)
   Model df   AIC   BIC  logLik   Test L.Ratio p-value
mC     1  5 375.3 389.7 -182.65
mB     2 10 162.7 191.5  -71.33 1 vs 2   222.6  <.0001
>
>
> stem.leaf(sigBoot)
1 | 2: represents 0.012
 leaf unit: 0.001
            n: 100
LO: 0.1616
    2    17 | 9
         18 |
    5    19 | 699
   11    20 | 000579
   17    21 | 122489
   24    22 | 1256789
   33    23 | 024455578
   48    24 | 033466677778889
  (13)   25 | 0012335677899
   39    26 | 0000122336777
   26    27 | 00112445678
   15    28 | 0166789
    8    29 | 22359
    3    30 | 159
```

PhD Preliminary Examination – 2006

Methods Question 3

1 Problem Background

This question involves a portion of a problem of evaluating the environmental impacts of proposed liquefied natural gas terminals in the Gulf of Mexico. These LNG terminals, as they are known, consist of offshore platforms (much like oil platforms) that convert liquefied natural gas into a gaseous state for use as an energy source. Natural gas is liquefied by cooling it to -260 degrees F which reduces its volume by about 600 times, allowing it to be shipped around the globe on ocean tankers (large vessels). The liquefied natural gas is delivered to LNG terminals where it is gasified by heating it with water from the Gulf of Mexico.

One environmental concern connected with LNG terminal operation is mortality to larval fish, specifically those that have just hatched and are not yet able to swim on their own; these fish are known as *ichthyoplankton*, "ichthyo" meaning fish and "plankton" meaning particles that simply drift along as carried by water. The gasification process uses large amounts of water pumped from the Gulf of Mexico and larval fish in this early life stage are carried into the plant with the water. These "entrained" larvae are subjected to various sources of stress during the gasification process, including physical contact with system components and exposure to chemicals used to prevent marine algal growth in the water pumps. It is assumed that any larval fish pumped through the LNG terminals die.

The underlying issue for this question, then, is the potential for larval mortality due to operation of LNG terminals in the Gulf of Mexico. The US Coast Guard and the National Marine Fisheries Service were responsible for preparation of Environmental Impact Statements (EISs) for a number of proposed LNG terminals. Such EISs are subject to review by the public, including the natural gas industry which can, and did, question a number of scientific assumptions and conclusions in the EISs. Industry reports are, in turn, subject to assessment by a variety of groups who prepare and issue their own reports. The setting for this question is to suppose

that you have been retained (i.e., hired) by the Gulf Coast Marine Fisheries Council, which is the government body responsible for setting policy for management of marine resources in the Gulf of Mexico, to review part of a report prepared by the natural gas industry in response to an EIS issued by the Coast Guard and National Marine Fisheries Service. The issue of contention is estimation of potential mortality of one particular marine fish species called red drum (*Sciaenops ocellatus*) which is commercially important and has undergone population declines in the past century.

2 Description of Data

The study that forms the focus of this question was conducted near the proposed location of one LNG terminal about 65 kilometers offshore from the coast of Louisiana. Because it was assumed that larval red drum entrained in the gasification process suffer 100% mortality, estimation of potential mortality becomes the same as estimation of larval abundance in areas from which water will be taken. The National Marine Fisheries Service conducted a number of tows with a research vessel in the area around the proposed LNG terminal using very fine nets (to capture the very small larval fish). These tows were taken over a period of 18 days during the time when the majority of red drum larvae will be present. The data recorded were the number of red drum larvae per million gallons of water through which the net passed (calculated on the basis of net size and speed at which the research vessel was moving). Data are available from a total of only 35 tows because the observational process is both expensive and difficult (e.g., tiny larval fish must be separated from algae and other types of plankton and identified to the level of species by people with considerable expertise).

The data available for estimation of red drum abundance are presented in Table 1. In this table, day is recorded as Julian Date, or day of the year. The first day is 248 which corresponds to 5 September and the last day is 265 which corresponds to 22 September.

As is evident from Table 1, the abundance of larval red drum (recorded in numbers per million gallons of water) do not represent a "nice" distributional form. The

Day	Number	Day	Number	Day	Number
248	205	253	10,305	258	1,487
248	0	253	0	259	50
249	386	254	7,227	259	0
250	0	255	0	259	0
251	0	255	4,088	260	665
251	541	255	1,236	261	0
252	0	256	6,169	261	0
252	0	256	0	262	0
252	0	257	3,028	263	0
252	0	257	0	264	968
252	0	257	0	265	76
252	0	257	69		

Table 1: Observed Abundance of Red Drum Larvae

five number summary of these data is given in Table 2. Notice in Table 1 that 20 of the 35 tows conducted resulted in capture of 0 red drum larvae, or a proportion of 0.5714 zeros, as is also reflected in the minimum, first quartile, and median values all being 0 in the summary of Table 2. A plot of number of red drum larvae captured (per million gallons of water) against day is presented in Figure 1.

Min.	Q1	Med	Q3	Max
0	0	0	603	10310

Table 2: Five Number Summary of Data from Table 1.

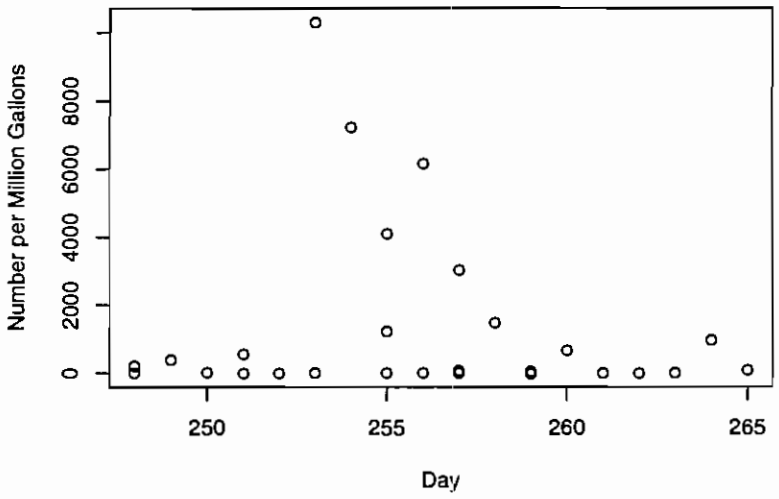


Figure 1: Abundance of Red Drum Larvae (per million gallons of water) versus Julian date of sample.

3 Proposed Estimators

3.1 Environmental Impact Statement

In the original Environmental Impact Statement prepared by the Coast Guard and National Marine Fisheries Service, the mean abundance of red drum larvae was estimated through calculation of the sample mean and variance of the 35 data values. A 95% confindence interval was also computed by taking the sample mean plus and minus the 0.975 quantile of a $t-$distribution (with 34 degrees of freedom) times the standard error of the mean. That is, let $y_1, \ldots, y_n$ denote the observed values of abundance on tow i with $n = 35$. Then,

$$\begin{aligned} \bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i = 1042.857 \\ var(y) &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 5585594 \\ var(\bar{y}) &= \frac{1}{n}var(y) = 159588.4 \\ 95\%\text{int} &= (231.006,\ 1854.709) \end{aligned} \tag{1}$$

An estimate of the total number of red drum larvae that might be entrained by the LNG terminal was arrived at by making the following assumptions:

1. The total length of time per year that larvae will be susceptible is 18 days.
2. The sample data from 18 days given in Table 1 are indicative of abundance (in number per million gallons) during this entire period.
3. The LNG terminal will pump 136 million gallons of water a day.

Using these assumptions, the EIS estimated the total mortality of red drum larvae in a year as follows, where T denotes the estimated total:

$$\begin{aligned} T &= 18(136)\bar{y} = 2552914 \\ 95\%\text{int} &= 18(136)(231.006,\ 1854.709) = (565503,\ 4540328) \end{aligned} \tag{2}$$

Thus, the EIS estimated annual mortality to red drum at somewhat over 2.5 million larvae with a 95% confidence interval of about 0.6 to 4.5 million larvae.

3.2 Industry Report

The report of the natural gas industry (on which you have been asked to comment) asserted that the procedure employed for estimation of red drum abundance in the EIS "overestimates the variability" in actual abundance because "seasonal variability" in abundance was not taken into account, but is visible in Figure 1 (quotes taken from the industry report). This report suggested an alternative to give a more precise measure of abundance. That alternative was based on a five day moving average of abundance values. Specifically, let $y_i(t)$ denote the observed abundance in tow i of day t, $i = 1, \ldots, n(t)$. At a given day t, a five day moving average may be defined as

$$m(t) = \frac{1}{N(t)} \sum_{u=t-2}^{t+2} \sum_{i=1}^{n(u)} y_i(u), \tag{3}$$

where

$$N(t) = \sum_{u=t-2}^{t+2} n(u)$$

Day t	Moving Avg. $m(t)$	Moving Variance $v(t)$	$N(t)$
250	94.33	2855.43	12
251	936.00	728193.56	12
252	1506.08	996180.23	12
253	1671.21	757435.57	14
254	2073.21	849140.17	14
255	2676.83	1044748.94	12
256	2118.54	641269.64	11
257	1240.53	305622.12	13
258	1042.54	344818.99	11
259	481.72	84334.24	11
260	275.25	36603.46	8
261	89.37	6800.39	8
262	272.16	31160.03	6
263	174.00	25371.47	6

Table 3: Moving average and moving variance estimates resulting from application of expressions (3) and (4) to the data of Table 1.

This report also proposed using a five day moving variance for the estimated $m(t)$, defined as,

$$v(t) = \frac{1}{\{N(t)-1\}} \sum_{u=t-2}^{t+2} \sum_{i=1}^{n_{(u)}} \{y_i(u) - m(u)\}^2 . \qquad (4)$$

Values resulting from the application of the estimators (3) and (4) to the data of Table 1 are presented in Table 3. A plot of the moving averages over the relevant time period is presented in Figure 2, and a similar plot of the moving variances is presented in Figure 3. Note that because of the requirement that $m(t)$ and $v(t)$ involve sums from $t-2$ to $t+2$, values of these quantities can be computed only for $t = 3$ to $t = 16$.

The authors of the industry report also proposed that the total red drum larval abundance (per million gallons of water) be estimated as the sum of moving average values. The total potential mortality from operation of the LNG terminal was then

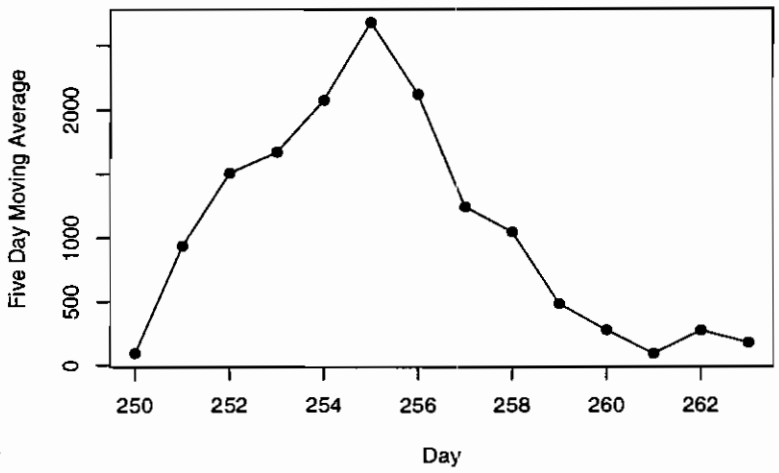


Figure 2: Moving average estimator of Red Drum larval abundance (per million gallons of water) versus Julian date of sample.

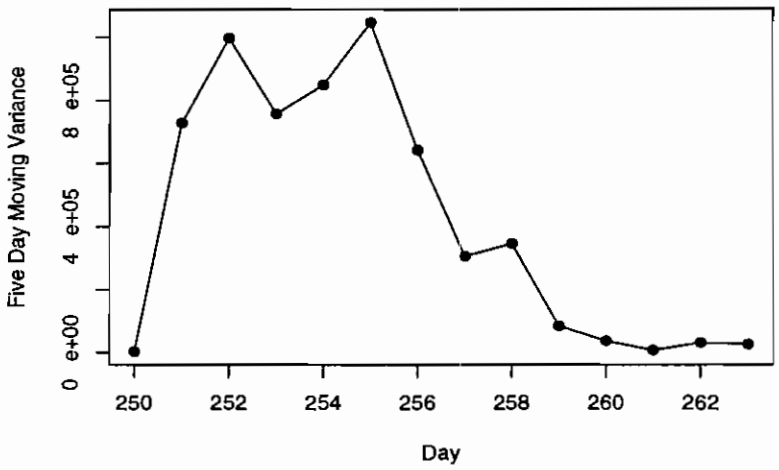


Figure 3: Moving variance for Red Drum larval abundance versus Julian date of sample.

estimated as 136 times this value; recall that the terminal pumps 136 million gallons of water per day. Using the values of Table 3 this total is,

$$W = 136 \sum_{t=3}^{16} m(t) = 1992648. \tag{5}$$

Finally, the industry report indicated that the variance of this estimator could be calculated as,

$$var(W) = 136^2 \sum_{t=3}^{16} v(t) = 108285465198, \tag{6}$$

citing a literature source (Bevington, P.R. and Robinson, D.K. (1992), Data reduction and error analysis for the physical sciences. WCB/McGraw-Hill, Boston) for the fact that "the variance of a sum is equal to the sum of the variances of the summands". This variance was then used to compute a 95% interval estimate of the total potential mortality to larval red drum as,

$$\begin{aligned} W &\pm 1.96\sqrt{var(W)} \\ &= (1347676, 2637621). \end{aligned} \tag{7}$$

In the industry report, then, the authors estimate total potential mortality to red drum larvae from operation of the LNG terminal as not quite 2 million larvae per year with a 95% interval of 1.3 million to just over 2.6 million.

3.3 Comparison of Estimates

Here, we collect the results of estimation used in the EIS and the industry report for ease of reference. Values for estimated total potential mortality to red drum larvae from under the two approaches used are collected in Table 4.

In the industry report, it is claimed that the estimate based on moving averages and moving variances is superior to that contained in the EIS because "Properly distinguishing seasonal variability from uncertainty results in far narrower confidence limits. The result is a more accurate and precise estimate that is a better basis for decision making" (quote taken from industry report page 3-16).

Approach	Total	95% Interval
EIS	2,552,914	(565,503; 4,540,328)
Industry	1,992,648	(1,347,676; 2,637,621)

Table 4: Summary of estimated total annual potential red drum mortality under two approaches.

4 Questions

The setting of the following questions is that you have been retained (hired) by the Gulf Coast Fisheries Management Council, which is the body that sets policy for management of marine fisheries resources for the Gulf of Mexico. Your job is to assess the relevance of the industry report which of course also involves the original EIS to which that report was a response. There are, clearly, any number of issues that statisticians might think of in reviewing this material. To help you in focusing your answer, the following issues should NOT be used in answering the specific questions posed below, except as otherwise noted in the questions.

1. The assumption that an 18 day total is indicative of annual abundance and/or larval mortality is given by fisheries scientists. Evidence in the data that this may not be accurate should be ignored.

2. The procedure used by the NMFS research vessel scientists to record catch of larval red drum in numbers per million gallons of water almost certainly involved some measurement error. Assume, however, that these values are exact measurements without error. Also assume that the figure of 136 million gallons of water use by the proposed LNG terminal is exact.

3. The observations obtained were from a region surrounding the proposed LNG facility but may not be entirely representative of the exact location. Spatial effects due to, for example, substrate type are almost certainly present. But, ignore the possibility of such effects and assume that the values obtained are indicative of what would be the situation at the site of the facility.

Given the above caveats (i.e., things to disregard), answer the following questions. *Note: all of the following questions are of equal weight other than question 5 which is weighted more than the others.*

Question 1. Identify statistical assumptions made in the analysis used in the EIS. That is, what is needed for the estimates given in expression (1) to be valid? Note that you are not being asked at this point whether these assumptions are justified, just what they would have to be. Clearly define any notation you introduce.

Question 2. Is there any evidence in the information presented that would cause you to question these assumptions? Identify reasons for concern or indicate what features of the information presented fail to provide evidence against the assumptions. The "information presented" may be taken to mean Tables 1 and 2 and Figure 1, since these are constructed directly from the raw data.

Question 3. The industry report does not lay out any specific assumptions or formal statistical framework for the procedure recommended in that report. Is it reasonable that this procedure would be applied under the same assumptions you identified in question 1? That is, if the assumptions you listed in Question 1 are reasonable, would there be any benefit to using the moving average estimator proposed in the industry report? If not, define a statistical framework that would be appropriate for the moving average estimator. Be specific and again clearly define any additional notation you introduce.

Question 4. The point estimate of total annual mortality given in the industry report is more than 1/2 million smaller than the point estimate of the EIS (see Table 4). Is this a consequence of the estimator used or simply something that happened for this particular data set? Demonstrate that there is or is not a systematic difference in the estimator used in the EIS, which is given as T in expression (2), and the estimator used in the industry report, which is given as W in expression (5).

Hint: You may consider a simpler situation than given in the problem by assuming that only one observation is available per day and taking only a 3

day moving average, if that makes it easier to demonstrate what you are trying to show.

Question 5. Consider the variance estimator of expression (6) recommended in the industry report. What does this estimator assume about the random variables associated with values for individual tows? Under the framework you defined in question 3, is this estimator supported by statistical theory, contradicted by statistical theory, or simply an *ad hoc* procedure that can be neither supported nor discredited? Demonstrate that your conclusion is correct.
Hint: As in question 4, you may consider a simpler situation than given in the problem by assuming that only one observation is available per day and taking only a 3 *day moving average, if that makes it easier to demonstrate that your conclusion is correct. You may also, WLOG, consider only two values of the moving average, say the first and the second. That is, consider a sequence of* 4 *daily tows to define* 2 *moving average values with a window of length* 3.

Question 6. Identify what you feel to be TWO important statistical issues for estimation and inference in this problem that are not adequately addressed by either the analysis of the EIS or the analysis of the industry report.

Question 7. Identify what you feel to be ONE important practical issue that might affect our ability to adequately deal with one or both of the issues you identified in your answer to question 6.

Question 8. BRIEFLY DESCRIBE one appropriate procedure that could be used to determine whether the data of Table 1 provide evidence against an assumption that larval abundance in individual tows can be considered identically distributed across days. DO NOT attempt to formally define an algorithm or conduct any such procedure here. Simply describe what one might do in words.

Question 9. Provide an alternative formulation of the problem that you feel would be an improvement over either the EIS analysis or the industry report analysis. Here, you may consider the issues given in the prelude to these questions that you were told to ignore unless otherwise indicated (although you don't necessarily

have to in order to construct a good answer). Although it is not possible to presume data collection can be increased, the National Marine Fisheries Service conducts abundance surveys of the type used to provide data here on an annual basis and for multiple species.

PhD Preliminary Examination – 2006

Answers – Methods Question 3

These are a sketch of the answers hoped for. Other possibilities might exist for some of the questions that would be entirely adequate if they are both technically correct and logically consistent.

1. The analysis conducted in the EIS follows from a classical one-sample normal problem. For tow i on day t, define random variables $\{Y_i(t) : i = 1, \ldots, n(t); t = 1, \ldots, k\}$ as corresponding to the number of red drum larvae caught (per million gallons of water). Assume that,

$$Y_i(t) \sim iidN(\mu, \sigma^2).$$

2. Values of Tables 1 and 2 clearly indicate a problem with the assumption of normality. Over half of the values of Table 1 are zero, with many others being in the hundreds or thousands. This is reflected in Table 2 by the minimum, first quartile, and median all being zero while the maximum is 10310 so that the distribution cannot be symmetric about a mean. The quantity of data (i.e., 35 observations) would not seem sufficient in this case to reasonably invoke the central limit theorem. The plot of Figure 1 calls into question the assumption of common mean. It is less clear, using only these crude data summaries, whether there is evidence against an assumption of independence, althougli Figure 1 again suggests there might be a time effect of some type, with most of the large values occurring between days 253 and 256 (or perhaps 257).

3. There would be no motivation for applying a moving average filter to *iid* observations. The assumptions of the industry report will hopefully be formalized by (for example) defining random variables $\{Y_i(t) : i = 1, \ldots, n(t); t = 1, \ldots, k\}$ with assumptions along one or more of the following lines:

(a) $\{Y_i(t) : i = 1, \ldots, n(t);\ t = 1, \ldots, k\}$ independent with means $\{\mu(t) : t = 1, \ldots, k\}$ and common variance σ^2

(b) $\{Y_i(t) : i = 1, \ldots, n(t);\ t = 1, \ldots, k\}$ independent with means $\{\mu(t) : t = 1, \ldots, k\}$ and variances $\{\sigma^2(t) : t = 1, \ldots, k\}$

(c) $\{Y_i(t) : i = 1, \ldots, n(t);\ t = 1, \ldots, k\}$ are possibly dependent random variables with possibly varying means and possibly varying variances.

4. Answers will hopefully identify the fact that the estimator of total mortality given in expression (5) will always be less than or equal to the simple total of observations due to the effects of down-weighting of observations on the ends of the data sequence, and always less unless all of those down-weighted values are zero. Using the hint, let $\{Y(t) : t = 1, \ldots, k\}$ represent daily values for k successive days. A moving average with window of size 3 would be,

$$m(t) = \frac{1}{3} \sum_{u=t-1}^{t+1} Y(u).$$

An estimator of total analogous to that of expression (5) would then be,

$$\begin{aligned} W &= 136 \sum_{t=2}^{k-1} m(t) \\ &= 136 \sum_{t=2}^{k-1} \frac{1}{3} \sum_{u=t-1}^{t+1} Y(u) \\ &= 136 \frac{1}{3} \left[Y(1) + 2Y(2) + 3Y(3) + 3Y(4) + \ldots, +3Y(k-2) + 2Y(k-1) + Y(k) \right] \\ &= 136 \left[\frac{1}{3} \{ Y(1) + Y(k-2) \} + \frac{2}{3} \{ \{ Y(2) + Y(k-1) \} + \sum_{t=3}^{k-3} Y(t). \right] \end{aligned}$$

and, for non-negative $Y(t)$ this is clearly less than or equal to the EIS estimator which would be, in this case,

$$T = 136 \sum_{t=1}^{k} Y(t).$$

5. What is desired in question 5 is recognition of covariance among moving average values regardless of whether random variables from question 3 are assumed independent or not. Following the hint, let $\{Y(t) : t = 1, \ldots, k\}$ be a sequence of independent daily values such that $E\{Y(t)\} = \mu(t)$ and $var\{Y(t)\} = \sigma^2$. This is the smallest change from an *iid* assumption that might motivate a moving average approach. Consider a three-day moving average,

$$m(t) = \frac{1}{3}\sum_{u=t-1}^{t+1} Y(u); \quad t = 2, \ldots, k-1.$$

Now,

$$\begin{aligned} m(2) &= \frac{1}{3}\{Y(1) + Y(2) + Y(3)\}, \\ m(3) &= \frac{1}{3}\{Y(2) + Y(3) + Y(4)\}. \end{aligned}$$

Then,

$$\begin{aligned} E\{m(2)\} &= \frac{1}{3}\{\mu(1) + \mu(2) + \mu(3)\}, \\ E\{m(3)\} &= \frac{1}{3}\{\mu(2) + \mu(3) + \mu(4)\}, \\ var\{m(2)\} &= \frac{1}{3}\sigma^2, \\ var\{m(3)\} &= \frac{1}{3}\sigma^2. \end{aligned}$$

Also,

$$\begin{aligned} E\{m(2)m(3)\} &= \frac{1}{9}E\left[\{Y(1) + Y(2) + Y(3)\}\{Y(2) + Y(3) + Y(4)\}\right] \\ &= \frac{1}{9}\{\mu(1)\mu(2) + \mu(1)\mu(3) + \mu(1)\mu(4) + (\mu(2) + \sigma^2) + \mu(2)\mu(3) \\ &+ \mu(2)\mu(4) + \mu(3)\mu(2) + (\mu(3)^2 + \sigma^2) + \mu(3)\mu(4)\} \end{aligned}$$

while,

$$\begin{aligned} E\{m(2)\}E\{m(3)\} &= \frac{1}{9}\{\mu(1)\mu(2) + \mu(1)\mu(3) + \mu(1)\mu(4) + \mu(2)^2 + \mu(2)\mu(3) \\ &+ \mu(2)\mu(4) + \mu(3)\mu(2) + \mu(3)^2 + \mu(3)\mu(4)\} \end{aligned}$$

So then,

$$cov\{m(2), m(3)\} = E\{m(2)m(3)\} - E\{m(2)\}E\{m(3)\} = \frac{2}{9}\sigma^2.$$

The above demonstration suffices to show that a sequence of moving averages cannot be independent, and hence the variance given in expression (6) is not correct and is, in fact, an underestimate of the true variance of the sum of moving averages. It might also be noted that the value of 1.96 was used in the interval of expression (7) which indicates the need for some type of central limit theorem for dependent random variables (e.g., m-dependent sequences). It is almost certain that the data set is not large enough in this problem to reasonably invoke such results.

6. Various answers would be possible here, but two fairly obvious issues are (1) the large frequency of zero data values combined with exceptionally large values for other observations, and (2) the unequal numbers of observations taken within days (e.g., the numbers of observations within days are 1 (for 9 days), 2 (for 5 days), 3 (for 2 days), 4 (for 1 day), and 6 (for 1 day).

7. The intent is for answers to identify the fact that data are extremely limited, which renders formulation of elaborate model structures impossible, unless those structures can be reduced to a few parameters or draw on additional sources of information.

8. Reasonable answers here would include permutation or randomization procedures. Possibilities for a test statistic would include number of consecutive zeros, variance among days, or proportion of days with all zero values.

9. There is no set answer here. One might very well think of a mixture of a binary (0, 1) for positive abundance or not with some distribution on the positive line for abundances greater than 0. More elaborate versions might

incorporate measurement error for both pieces using prior information from previous years of research tows by the National Marine Fisheries Service or information in the same year from related species. Regardless, answers should pay attention to the need for fairly simple model structures dictated by data availability even if those models are clearly not fully adequate for the problem.