Biostatistics 533

Classical Theory of Linear Models

Spring 2007

Final Exam

Name: KEY

Problems do not have equal value and some problems will take more time than others. Spend your time wisely. You do not have to give reasons unless you are explicitly asked.

Problem	1	2	3	4	5	6	7	8	Total	Bonus
Value	17	13	14	8	8	8	8	24	100	6
Score										

Please choose ONE of the following options.

A. Sign your name below if you would like your graded exam placed in your mailbox in the Biostatistics department. By signing you acknowledge that the mailboxes are not secure and you accept the risk that your exam could be lost or your grade might not be private.

A. ______________________________

B. If you do not have a mailbox in the Biostatistics department, sign your name below and write your campus mailbox number if you would like your graded exam mailed to you via campus mail. By signing you acknowledge that campus mail is not secure and you accept the risk that your exam could be lost or your grade might not be private.

B. ______________________________ Campus Mailbox #: ____________________

C. (DEFAULT) Sign your name below if you would like your graded exam held in Dr. Kerr's office until you pick it up. Unclaimed exams will be discarded after two months.

C. ______________________________

1. (17 points) Let $\mathbf{X}$ be an $n \times p$ matrix with linearly independent columns. Let $\mathbf{Y}$ be an $n \times 1$ vector. Suppose $\mathbf{Y} = \mathbf{a} + \mathbf{b}$ where $\mathbf{a} \in \mathcal{R}(\mathbf{X})$ and $\mathbf{b} \in \mathcal{R}(\mathbf{X})^{\perp}$.

(a)

Is $\mathbf{a}$ unique? YES

Is $\mathbf{b}$ unique? YES

(b) Compute $\mathbf{a}'\mathbf{b}$.

$$\mathbf{a}'\mathbf{b} = 0$$

(c) Write $\mathbf{a}$ as a function of $\mathbf{X}$ and $\mathbf{Y}$.

$$\mathbf{a} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

(d) Write $\mathbf{b}$ as a function of $\mathbf{X}$ and $\mathbf{Y}$.

$$\mathbf{b} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} = \mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

(e) Compute $\mathbf{X}'\mathbf{b}$.

$$\text{since } \mathbf{b} \in R(X)^{\perp}, \ \mathbf{X}'\mathbf{b} = 0$$

2. This linear model with intercept has $p+1$ parameters:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$$

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p]$ be the design matrix for the parameters $\beta_1, \beta_2, \ldots, \beta_p$. The least squares parameter estimates minimize $||\mathbf{Y} - \alpha\mathbf{1} - \mathbf{X}\boldsymbol{\beta}||$ or equivalently, $||\mathbf{Y} - \alpha\mathbf{1} - \mathbf{X}\boldsymbol{\beta}||^2$, where $\boldsymbol{\beta}$ is the vector $(\beta_1, \beta_2, \ldots, \beta_p)'$.

(a) (8 points) Let $\bar{\mathbf{Y}} = \bar{Y}\mathbf{1}$ (the vector where every entry is the average of the Y's) and let $\bar{\mathbf{X}}$ be the matrix with all rows equal to $\bar{\mathbf{x}}' = [\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \ldots, \bar{\mathbf{x}}_p]$. Prove that $||\mathbf{Y} - \alpha\mathbf{1} - \mathbf{X}\boldsymbol{\beta}||^2$ can be written

$$||\bar{\mathbf{Y}} - \alpha\mathbf{1} - \bar{\mathbf{X}}\boldsymbol{\beta}||^2 + ||\mathbf{Y} - \bar{\mathbf{Y}} - (\mathbf{X} - \bar{\mathbf{X}})\boldsymbol{\beta}||^2$$

$||Y - \alpha 1 - X\beta||^2 = ||\mathbf{Y} - \alpha\mathbf{1} - \mathbf{X}\boldsymbol{\beta} - \overline{\mathbf{Y}} + \overline{\mathbf{Y}} - \overline{\mathbf{X}}\boldsymbol{\beta} + \overline{\mathbf{X}}\boldsymbol{\beta}||^{\mathbf{2}} = ||\overline{\mathbf{Y}} - \alpha\mathbf{1} - \overline{\mathbf{X}}\boldsymbol{\beta}||^{\mathbf{2}} + ||\mathbf{Y} - \overline{\mathbf{Y}} - (\mathbf{X} - \overline{\mathbf{X}})\beta||^{\mathbf{2}}$ as long as we can show $\overline{Y} - \alpha\mathbf{1} - \overline{X}\boldsymbol{\beta}$ and $\mathbf{Y} - \overline{\mathbf{Y}} - (\mathbf{X} - \overline{\mathbf{X}})\boldsymbol{\beta}$ are orthogonal. Notice that $\overline{\mathbf{Y}} - \alpha\mathbf{1} - \overline{\mathbf{X}}\boldsymbol{\beta}$ is an n x 1 vector with constant entries. So we can write $\overline{\mathbf{Y}} - \alpha\mathbf{1} - \overline{\mathbf{X}}\boldsymbol{\beta} = c\mathbf{1}$ for some constant c. But $\mathbf{1}'(\mathbf{Y} - \overline{\mathbf{Y}}) = 0$ and $\mathbf{1}'(\mathbf{X} - \bar{(}\mathbf{X}) = 0$ so we have established orthogonality.

(b) (5 points) Let $\hat{\boldsymbol{\beta}}$ be the least squares estimate of $\boldsymbol{\beta}$. Prove that the least-squares estimate of α is $\hat{\alpha} = \bar{y} - \bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}$.

From (a), minimizing RSS is equivalent to minimizing $||\bar{\mathbf{Y}} - \alpha\mathbf{1} - \bar{\mathbf{X}}\boldsymbol{\beta}||^2 + ||\mathbf{Y} - \bar{\mathbf{Y}} - (\mathbf{X} - \bar{\mathbf{X}})\boldsymbol{\beta}||^2$. If $\hat{\boldsymbol{\beta}}$ is the LSE of $\boldsymbol{\beta}$, then we can make the first term equal to 0 by setting $\hat{\alpha} = \overline{y} - \overline{x}'\hat{\boldsymbol{\beta}}$. Since α only appears in the first term and since the first term is ≥ 0, this establishes that $\hat{\alpha} = \overline{y} - \overline{x}'\hat{\boldsymbol{\beta}}$ is the LSE of α.

3. (14 points) A botanist measured the concentration of a particular virus in plant sap using the ELISA assay. The study included 13 varieties of potato. The botanist was interested in how resistance to the virus varied among these varieties. Samples were taken from 5 plants from each variety, for a total of 65 plants. However, one measurement was lost from a failed assay.

A one-way ANOVA model was fit to the data. (a) Complete the table.

source	df	SS	MS	F	p-value
variety	12	13.84	11.53	20.2	0.0001
error	51	2.91	0.057		
total	63	16.75			

(b) What is the value of R^2 for this dataset?

$$R^2 = \frac{13.84}{16.75} = 0.8263$$

(c) Assume that errors in the underlying model are normally distributed, independent, and homoscedastic. Consider the null hypothesis H that the mean measurement is the same in all 13 groups. What is the result of testing this hypothesis at level 0.05?

REJECT H

4. (8 points) Let $\mathbf{Y} \sim MVN(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, where $\mathbf{X}$ is an $n \times p$ matrix with linearly independent columns.

(a) What is the distribution of $\mathbf{PY}$, where $\mathbf{P}$ is the projection onto the column space of $\mathbf{X}$?

$$E(\mathbf{PY}) = \mathbf{P}E(\mathbf{Y}) = \mathbf{PX}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}$$

$$cov(\mathbf{PY}) = \mathbf{P}cov(\mathbf{Y})\mathbf{P'} = \sigma^2\mathbf{PP'} = \sigma^2\mathbf{P}$$

$$\mathbf{PY} \sim MVN(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{P})$$

(b) What is the distribution of $\mathbf{Y}'(\mathbf{I} - \mathbf{P})\mathbf{Y}/\sigma^2$?

$$\frac{\mathbf{Y}'(I - \mathbf{P})\mathbf{Y}}{\sigma^2} = \frac{RSS}{\sigma^2} \sim \chi^2_{n-p}$$

5. (8 points) For a linear model, let $\mathbf{c}'\boldsymbol{\beta}$ be estimable. Show $\text{cov}(\mathbf{c}'\hat{\boldsymbol{\beta}}) = \sigma^2\mathbf{c}'\mathbf{G}\mathbf{c}$ where $\mathbf{G}$ is a generalized inverse of $\mathbf{X}'\mathbf{X}$. (You might find some of the following facts that we proved in a lemma useful: $\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X} = \mathbf{X}, \mathbf{X}'\mathbf{X}\mathbf{G}\mathbf{X}' = \mathbf{X}', \mathbf{X}\mathbf{G}\mathbf{X}' = \mathbf{X}\mathbf{G}'\mathbf{X}'$)

(*Hint:* use what we know about $\mathbf{c}$ when $\mathbf{c}'\boldsymbol{\beta}$ is estimable.)

$$\begin{aligned} cov(\mathbf{c}'\hat{\boldsymbol{\beta}}) &= cov(\mathbf{c}'\mathbf{G}\mathbf{X}'\mathbf{Y}) \\ &= \mathbf{c}'\mathbf{G}\mathbf{X}'\sigma^2 I\mathbf{X}\mathbf{G}'\mathbf{c} \\ &= \sigma^2\mathbf{c}'\mathbf{G}\mathbf{X}'\mathbf{X}\mathbf{G}'\mathbf{c} \end{aligned}$$

Since $\mathbf{c}'\boldsymbol{\beta}$ is estimable, $\mathbf{c} \in R(\mathbf{X}')$ and $\mathbf{c}' = \mathbf{T}_{1\times n}\mathbf{X}_{n\times p}$ for some $\mathbf{T}$

$$\begin{aligned} cov(\mathbf{c}\hat{\boldsymbol{\beta}}) &= \sigma^2\mathbf{T}\underbrace{\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{X}}_{\mathbf{X}}\mathbf{G}'\mathbf{X}'\mathbf{T}' \\ &= \sigma^2\mathbf{T}\mathbf{X}\mathbf{G}'\mathbf{X}'\mathbf{T}' \\ &= \sigma^2\mathbf{T}\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{T}' \\ &= \sigma^2\mathbf{c}'\mathbf{G}\mathbf{c} \end{aligned}$$

6. (8 points) Suppose in truth the model is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\eta} + \boldsymbol{\varepsilon}, \qquad E[\boldsymbol{\varepsilon}] = \mathbf{0},\ \mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I},$$

but we fit the smaller model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Find $E[\hat{\boldsymbol{\beta}}]$ when the smaller model is used and find an expression for the bias of $\hat{\boldsymbol{\beta}}$. You can assume $\mathbf{X}$ has full rank.

$$\begin{aligned}
E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\eta}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\eta} \\
&= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\eta} \\
\text{Bias} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\eta}
\end{aligned}$$

7. (8 points) For a linear model with design matrix $\mathbf{X}$ let $H : \mathbf{A}\boldsymbol{\beta} = \mathbf{0}$ be a testable hypothesis. We know

$$RSS_H - RSS = (\mathbf{A}\hat{\boldsymbol{\beta}})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}']^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}}).$$

Derive $E[RSS_H - RSS]$.

Noting that

$$E[\mathbf{A}\hat{\boldsymbol{\beta}}] = \mathbf{A}\boldsymbol{\beta} \equiv \boldsymbol{\mu}$$

$$cov[\mathbf{A}\hat{\boldsymbol{\beta}}] = \mathbf{A}cov[\hat{\boldsymbol{\beta}}]\mathbf{A}' = \sigma^2\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A} \equiv \boldsymbol{\Sigma}$$

we can apply the results about expectations of quadratic forms from Lecture 3:

$$E[(\mathbf{A}\hat{\boldsymbol{\beta}})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}']^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}})] = \mathrm{tr}([\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}']^{-1}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}']^{-1}\boldsymbol{\mu} =$$

$$\mathrm{tr}(\sigma^2[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}']^{-1}\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}) + (\mathbf{A}\boldsymbol{\beta})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}']^{-1}(\mathbf{A}\boldsymbol{\beta})$$

$$\sigma^2\mathrm{tr}(I_{(q\times q)}) + (\mathbf{A}\boldsymbol{\beta})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}']^{-1}(\mathbf{A}\boldsymbol{\beta}) = \sigma^2 q + (\mathbf{A}\boldsymbol{\beta})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-}\mathbf{A}']^{-1}(\mathbf{A}\boldsymbol{\beta})$$

where $\mathbf{A}$ has q (assumed) linearly independent rows.

8. (24 points) Consider the following model:

$$\begin{aligned} Y_1 &= \tau_1 + \tau_2 + \tau_3 + \epsilon_1 \\ Y_2 &= \tau_1 \qquad\;\; + \tau_3 + \epsilon_2 \\ Y_3 &= \qquad \tau_2 \qquad\; + \epsilon_3 \end{aligned}$$

(a) Write out the model in matrix form.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{bmatrix}$$

(b) What is the rank of the design matrix? 2

(c) Circle one:

τ_1 is: NOT ESTIMABLE

τ_2 is: ESTIMABLE

τ_3 is: NOT ESTIMABLE

(d) Is $\tau_1 - 2\tau_2 + \tau_3$ estimable? Explain how you know.

Yes. Let $\mathbf{r}_2'$ and $\mathbf{r}_3'$ be the 2nd and 3rd rows of $\mathbf{X}$. $\mathbf{r}_2' - 2\mathbf{r}_3' = (1, -2, 1)$ so $\tau_1 - 2\tau_2 + \tau_3$ is estimable.

(e) Find a linear unbiased estimate of $\tau_1 - 2\tau_2 + \tau_3$. Note: You are not required to find the BLUE.

$Y_2 - 2Y_3$

(f) Find a linear unbiased estimate of $\tau_1 - 2\tau_2 + \tau_3$ that is different from the estimator you gave in (e).

$Y_1 - 3Y_3$

(g) BONUS (6 points) DO NOT WORK ON THIS UNLESS YOU HAVE FINISHED THE REST OF THE EXAM.
Find the BLUE of $\tau_1 - 2\tau_2 + \tau_3$.

There were many ways to do this problem. Here is a solution imposing linear constraints on the parameters. A convenient and legitimate linear constraint to choose is to use $\tau_1 = 0$.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \tau_2 \\ \tau_3 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{3}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} \end{bmatrix}$$

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} Y_1 & + & Y_3 \\ Y_1 & + & Y_2 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \frac{2}{3}Y_1 & + & \frac{2}{3}Y_3 & - & \frac{1}{3}Y_1 & - & \frac{1}{3}Y_2 \\ \frac{2}{3}Y_1 & + & \frac{2}{3}Y_2 & - & \frac{1}{3}Y_1 & - & \frac{1}{3}Y_3 \end{bmatrix} =$$

$$\begin{bmatrix} \frac{1}{3}Y_1 & - & \frac{1}{3}Y_2 & + & \frac{2}{3}Y_3 \\ \frac{1}{3}Y_1 & + & \frac{2}{3}Y_2 & - & \frac{1}{3}Y_3 \end{bmatrix} = \begin{bmatrix} \hat{\tau}_2 \\ \hat{\tau}_3 \end{bmatrix}$$

Using the contraint $\tau_1 = 0$, $\tau_1 - 2\tau_2 + \tau_3 = -2\tau_2 + \tau_3$ so the BLUE of $\tau_1 - 2\tau_2 + \tau_3$ is $(\frac{-2}{3}Y_1 + \frac{2}{3}Y_2 - \frac{4}{3}Y_3) + (\frac{1}{3}Y_1 + \frac{2}{3}Y_2 - \frac{1}{3}Y_3) = \frac{-1}{3}Y_1 + \frac{4}{3}Y_2 - \frac{5}{3}Y_3$. As a check, show that $E(\frac{-1}{3}Y_1 + \frac{4}{3}Y_2 - \frac{5}{3}Y_3) = \tau_1 - 2\tau_2 + \tau_3$