Applied Statistics Exam
August 2008
Time: 12:00pm–3:30pm

This exam consists of 2 parts. You must answer 5 questions total and must answer at least 2 questions from each part. Make sure to clearly indicate which problems you are attempting. Some formulas and tables are given at the end of this exam.

PART A:
1. Consider the simple linear regression model in which \( y \sim \text{Normal}_N(\mathbf{X}\beta, \sigma^2 \mathbf{I}) \) where \( \mathbf{X} = [\mathbf{J} : \mathbf{x}] \) and \( \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \); here \( \beta_0 \) is the unknown fixed intercept parameter, \( \beta_1 \) is the unknown fixed slope parameter, \( \sigma^2 \) is the unknown fixed variance, \( \mathbf{J} \) is a \( N \)-dimensional vector of ones, and \( \mathbf{x} \) is a known non-random \( N \)-dimensional column vector.

Assume that we have \( N = 10 \) and the following summary statistics:

\[
\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} 10 & 4 \\ 4 & 8 \end{bmatrix}, \quad \mathbf{X}^\top \mathbf{y} = \begin{bmatrix} 13 \\ 13 \end{bmatrix}, \quad \mathbf{y}^\top \mathbf{y} = 33.
\]

(a) Find the maximum likelihood estimates of \( \beta \) and \( \sigma^2 \).
(b) Test \( H_0 : \beta_1 = 0 \) vs. \( H_A : \beta_1 \neq 0 \) level \( \alpha = .05 \).
(c) Suppose we wish to test \( H_0 : \beta_0 = \beta_1 \) vs. \( H_A : \beta_0 \neq \beta_1 \). State the null hypothesis in the matrix form \( H_0 : \mathbf{K}^\top \beta = \mathbf{m} \). In particular, what are \( \mathbf{K} \) and \( \mathbf{m} \)?
(d) Assuming that \( \beta_0 = \beta_1 \), find the constrained maximum likelihood estimate of \( \beta \).
(e) Test \( H_0 : \beta_0 = \beta_1 \) vs. \( H_A : \beta_0 \neq \beta_1 \) at level \( \alpha = .05 \).

2. Suppose that \( \mathbf{y} = \mathbf{X}\beta + \epsilon \) where \( \mathbf{y} \) is a \( N \)-dimensional column vector of outputs, \( \mathbf{X} \) is a \( N \times p \) design matrix of fixed inputs, \( \beta \) is a \( p \)-dimensional column vector of coefficients, and \( \epsilon \) is a \( N \)-dimensional column vector of random errors such that \( \mathbb{E}[\epsilon] = 0_N \) and \( \text{var}[\epsilon] = \sigma^2 \mathbf{I} \)

(a) Let \( \alpha \) be a \( p \)-dimensional column vector of constants. Suppose that we want to estimate \( \alpha^\top \beta \) by a linear unbiased estimator \( \mathbf{c}^\top \mathbf{y} \). Compute \( \mathbb{E}[\mathbf{c}^\top \mathbf{y}] \) and show that \( \alpha^\top = \mathbf{c}^\top \mathbf{X} \).
(b) Let \( \mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \). Show that \( \text{cov}[\mathbf{c}^\top (\mathbf{I} - \mathbf{H})\mathbf{y}, \mathbf{c}^\top \mathbf{H} \mathbf{y}] = 0 \).
(c) Let \( \hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \) denote the least squares estimator of \( \beta \). Show that \( \text{var}[\mathbf{c}^\top \mathbf{y}] \geq \text{var}[\alpha^\top \hat{\beta}] \).
3. Consider the following summary tables for various simple linear regression models involving \( y, x_1, \) and \( x_2. \) For \( i = 1, \ldots, N, \) let \( \tilde{y}_i \) be the fitted value of \( y \) at \( x_1 = x_{i1} \) based on regressing \( y \) on \( x_1 \) and \( \tilde{x}_{2i} \) be the fitted value of \( x_2 \) at \( x_1 = x_{i1} \) based on regressing \( x_2 \) on \( x_1 \) where \( N \) is the number of observations in the data set.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Param. Est.</th>
<th>Std. Error</th>
<th>t-statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.663</td>
<td>1.169</td>
<td>2.279</td>
<td>0.035</td>
</tr>
<tr>
<td>( x_1 )</td>
<td>0.281</td>
<td>0.377</td>
<td>0.748</td>
<td>0.464</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Param. Est.</th>
<th>Std. Error</th>
<th>t-statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.171</td>
<td>0.214</td>
<td>0.801</td>
<td>0.434</td>
</tr>
<tr>
<td>( x_1 )</td>
<td>-0.055</td>
<td>0.069</td>
<td>-0.792</td>
<td>0.439</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Param. Est.</th>
<th>Std. Error</th>
<th>t-statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0</td>
<td>0.532</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( x_2 - \bar{x}_2 )</td>
<td>-0.909</td>
<td>1.270</td>
<td>-0.716</td>
<td>0.483</td>
</tr>
</tbody>
</table>

Determine the least squares estimates of \( \beta_0, \beta_1, \) and \( \beta_2 \) when fitting the model \( y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i, \ i = 1, \ldots, N. \)

4. Consider a random effects model

\[ Y_{ijk} = \mu_i + p_j + (mp)_{ij} + e_{ijk}, i = 1, \ldots, a, j = 1, \ldots, b, k = 1, \ldots, n \]

where the \( e_{ijk} \)'s are independent and each follows a Normal distribution with mean 0 and variance \( \sigma^2, \) the \( p_i \)'s are independent and each follows a Normal distribution with mean 0 and variance \( \sigma_a^2, \) the \((mp)_{ij} \)'s are also independent and each follows a Normal distribution with mean 0 and variance \( \sigma_{mp}^2, \) and \( \mu_i \) is the mean response for the \( i \)th level of the fixed factor. Also, all \( p_i \)'s, \((mp)_{ij} \)'s, and \( e_{ijk} \)'s are independent. Calculate the following quantities.

(a) \( E[Y_{ijk}] \)
(b) \( var[Y_{ijk}] \)
(c) \( cov[Y_{ijk}, Y_{ijk'}] \) for \( k \neq k' \)
(d) \( cov[Y_{ijk}, Y_{ij'k}] \) for \( j \neq j' \)
(e) \( cov[Y_{ijk}, Y_{ij'k}] \) for \( i \neq i' \).
PART B:

5. For fixed $\lambda > 0$ and fixed $x_i \in \mathbb{R}$, let

$$Q_2(\beta) = \sum_{i=1}^{N} (y_i - \beta x_i)^2 + \lambda \beta^2$$

where $y_1, \ldots, y_N$ are realizations of random variables $Y_1, \ldots, Y_N$, respectively.

(a) Find the value of $\beta$ which minimizes $Q_2$.

(b) Suppose that $Y_1, \ldots, Y_N$ are independent and $Y_i \sim \text{Normal}(\beta x_i, \sigma^2)$. Find the bias and the variance of the estimator proposed in (a).

6. Consider linear discriminant analysis (LDA) with a single input variable. That is, suppose that the conditional density of $X$ given $G = g$ is Normal with mean $\mu_g$ and variance $\sigma^2$ and that $P(G = g) = \pi_g$ for $g = 1, \ldots, K$. Given a new input $x$, LDA obtains the estimate for the corresponding output by finding the value of $g$ which maximizes $P(G = g|X = x)$. Show that the decision boundary between groups $k$ and $\ell$ has the form

$$\ln \frac{P(G = k|X = x)}{P(G = \ell|X = x)} = b_{k\ell} + \rho_{k\ell} x$$

and give explicit expressions for $b_{k\ell}$ and $\rho_{k\ell}$ in terms of $\pi_k, \pi_\ell, \mu_k, \mu_\ell, \rho_{k\ell}$, and $\sigma^2$.

The density of a normal distribution with mean $\mu$ and variance $\sigma^2$ is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ \frac{-(x - \mu)^2}{2\sigma^2} \right\}$$

7. Suppose that we have a data set with 8 observations of two inputs $x_1$ and $x_2$ and a Bernoulli output $y$. Use a classification tree with 2 binary splits based on minimizing the misclassification error to model the data.

<table>
<thead>
<tr>
<th>$x_1$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_2$</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>$y$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
8. Consider the data shown below. Use the squared Euclidean distance \(d(x,y) = (x - y)^2\) for all parts.

(a) Apply the K-means algorithm to obtain 2 clusters by beginning with Cluster 1 centered at point B and Cluster 2 center at point C.

(b) Draw the dendogram for the single linkage agglomerative clustering strategy. How would this strategy partition the data into two groups?

(c) Draw the dendogram for the complete linkage agglomerative clustering strategy. How would this strategy partition the data into two groups?
FORMULAS:

Suppose $y \sim \text{Normal}_N(X\beta, \sigma^2 I_N)$, $X$ is a $N \times p$ full rank matrix, $N > p$, and $X^TX$ is invertible. Let $\hat{\beta}$ be the MLE of $\beta$ and let $\hat{\beta}_0$ be the restricted MLE of $\beta$ satisfying $K^T\hat{\beta}_0 = m$. If $K^T\beta = m$, then

$$F = \frac{(\text{reduced SS} - \text{full SS})/q}{\text{full SS}/(N-p)}$$

$$= \frac{\|X(\hat{\beta} - \hat{\beta}_0)\|^2/q}{\|y - X\hat{\beta}\|^2/(N-p)}$$

$$= \frac{(K^T\hat{\beta} - m)^T(K^T(X^TX)^{-1}K)^{-1}(K^T\hat{\beta} - m)/q}{(y - X\hat{\beta})^T(y - X\hat{\beta})/(N-p)} \sim f_{q,N-p}$$

where reduced SS= $\|y - X\hat{\beta}_0\|^2$ and full SS= $\|y - X\hat{\beta}\|^2$.

TABLES:

1. The 100$\alpha$th percentage point of the central $t$-distribution with $df$ degrees of freedom.

2. The 100$\alpha$th percentage point of the central $\chi^2$-distribution with $df$ degrees of freedom.

3. Upper $\alpha$ probability points of the central $F$-distribution with $n_1$ d.f. in the numerator and $n_2$ d.f. in the denominator.