

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 $\mathbf{y} \sim \text{Normal}_N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ where $\mathbf{X} = [\mathbf{J} : \mathbf{x}]$ and $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$; here β_0 is the unknown fixed intercept parameter, β_1 is the unknown fixed slope parameter, σ^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 $\boldsymbol{\beta}$ and σ^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 $H_0 : \mathbf{K}^\top \boldsymbol{\beta} = \mathbf{m}$. In particular, what are \mathbf{K} and \mathbf{m} ?
- (d) Assuming that $\beta_0 = \beta_1$, find the constrained maximum likelihood estimate of $\boldsymbol{\beta}$.
- (e) Test $H_0 : \beta_0 = \beta_1$ vs. $H_A : \beta_0 \neq \beta_1$ at level $\alpha = .05$.

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

- (a) Let \mathbf{a} be a p -dimensional column vector of constants. Suppose that we want to estimate $\mathbf{a}^\top \boldsymbol{\beta}$ by a linear unbiased estimator $\mathbf{c}^\top \mathbf{y}$. Compute $E[\mathbf{c}^\top \mathbf{y}]$ and show that $\mathbf{a}^\top = \mathbf{c}^\top \mathbf{X}$.
- (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 $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ denote the least squares estimator of $\boldsymbol{\beta}$. Show that $\text{var}[\mathbf{c}^\top \mathbf{y}] \geq \text{var}[\mathbf{a}^\top \hat{\boldsymbol{\beta}}]$.

3. Consider the following summary tables for various simple linear regression models involving y , x_1 , and x_2 . For $i = 1, \dots, N$, 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.

Summary table for the regression of y on x_1 .

| Variable | Param. Est. | Std. Error | t -statistic | P -value |
|-----------|-------------|------------|----------------|------------|
| Intercept | 2.663 | 1.169 | 2.279 | 0.035 |
| x_1 | 0.281 | 0.377 | 0.748 | 0.464 |

Summary table for the regression of x_2 on x_1 .

| Variable | Param. Est. | Std. Error | t -statistic | P -value |
|-----------|-------------|------------|----------------|------------|
| Intercept | 0.171 | 0.214 | 0.801 | 0.434 |
| x_1 | -0.055 | 0.069 | -0.792 | 0.439 |

Summary table for the regression of $y - \tilde{y}$ on $x_2 - \tilde{x}_2$.

| Variable | Param. Est. | Std. Error | t -statistic | P -value |
|---------------------|-------------|------------|----------------|------------|
| Intercept | 0 | 0.532 | 0 | 1 |
| $x_2 - \tilde{x}_2$ | -0.909 | 1.270 | -0.716 | 0.483 |

Determine the least squares estimates of β_0 , β_1 , and β_2 when fitting the model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$, $i = 1, \dots, N$.

4. Consider a random effects model

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

where the e_{ijk} 's are independent and each follows a Normal distribution with mean 0 and variance σ^2 , the p_i 's are independent and each follows a Normal distribution with mean 0 and variance σ_a^2 , the $(mp)_{ij}$'s are also independent and each follows a Normal distribution with mean 0 and variance σ_{mp}^2 , and μ_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.

- $E[Y_{ijk}]$
- $var[Y_{ijk}]$
- $cov[Y_{ijk}, Y_{ijk'}]$ for $k \neq k'$
- $cov[Y_{ijk}, Y_{ij'k}]$ for $j \neq j'$
- $cov[Y_{ijk}, Y_{i'jk}]$ 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, \dots, y_N are realizations of random variables Y_1, \dots, Y_N , respectively.

(a) Find the value of β which minimizes Q_2 .

(b) Suppose that Y_1, \dots, 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 μ_g and variance σ^2 and that $P(G = g) = \pi_g$ for $g = 1, \dots, K$. 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 ℓ has the form

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

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

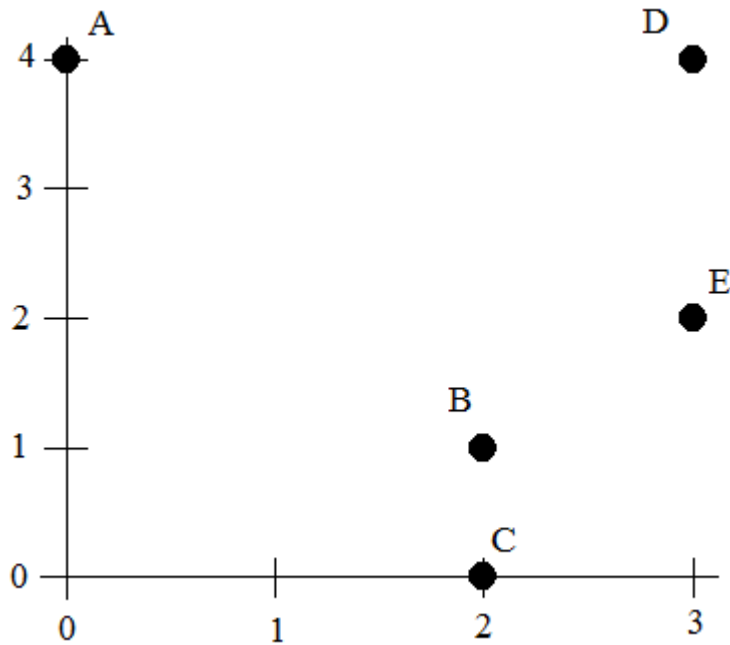
The density of a normal distribution with mean μ and variance σ^2 is

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

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.

| | | | | | | | | |
|-------|---|---|---|---|---|---|---|---|
| x_1 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| x_2 | 4 | 5 | 1 | 7 | 0 | 2 | 6 | 3 |
| y | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |

8. Consider the data shown below. Use the squared Euclidean distance $d(x, y) = (x - y)^2$ for all parts.



- 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.
- Draw the dendrogram for the single linkage agglomerative clustering strategy. How would this strategy partition the data into two groups?
- Draw the dendrogram for the complete linkage agglomerative clustering strategy. How would this strategy partition the data into two groups?

FORMULAS:

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

$$\begin{aligned} F &= \frac{(\text{reduced SS} - \text{full SS})/q}{\text{full SS}/(N-p)} \\ &= \frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0)\|^2/q}{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/(N-p)} \\ &= \frac{(\mathbf{K}^\top \hat{\boldsymbol{\beta}} - \mathbf{m})^\top (\mathbf{K}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{K})^{-1} (\mathbf{K}^\top \hat{\boldsymbol{\beta}} - \mathbf{m})/q}{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(N-p)} \sim f_{q, N-p} \end{aligned}$$

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

TABLES:

1. The 100α th percentage point of the central t -distribution with df degrees of freedom.
2. The 100α th percentage point of the central χ^2 -distribution with df degrees of freedom.
3. Upper α probability points of the central F -distribution with n_1 d.f. in the numerator and n_2 d.f. in the denominator.