Applied Statistics (Ph.D. Version)

Instructions to the Student

a. Answer any six questions. Each will be graded from 0 to 10.

b. Use a different booklet for each question. Write the problem number and your code number (NOT YOUR NAME) on the cover.

c. Keep scratch work on separate pages in the same booklet.

d. If you use a "well known" theorem in your solution to any problem, it is your responsibility to make clear which theorem you are using and to justify its use.

e. You may use calculators as needed.

1. Let \( Y = X\beta + \epsilon \), where \( E[\epsilon] = 0 \), \( \text{Var-Cov}[\epsilon] = \sigma^2 I \) and \( X \) has full rank. The ridge regression estimator of \( \beta \), denoted \( \hat{\beta}_R \), minimizes the quantity \( \|Y - X\beta\|^2 + (1/2)\lambda \|\beta\|^2 \), where \( \lambda \) is a nonnegative constant.

   (a) Show that \( \hat{\beta}_R = (X^TX + \lambda I)^{-1}X^TY \).

   (b) Find the expectation and variance-covariance matrix of \( \hat{\beta}_R \).

   (c) Let \( \psi = c^T\beta \). Calculate the mean squared error \( E[(c^T\hat{\beta}_R - \psi)^2] \). Compare your result to \( \text{Var}[c^T\hat{\beta}_{LS}] \), where \( \hat{\beta}_{LS} \) is the ordinary least squares estimator of \( \beta \).

   (d) Show that there exists \( \lambda \) such that

   \[ E[(c^T\hat{\beta}_R - \psi)^2] < \text{Var}[c^T\hat{\beta}_{LS}] \]

   Does this result contradict the Gauss-Markov Theorem?

   **Hint:** Use the fact that \( [I + \epsilon A]^{-1} = \sum_{m=0}^{\infty}[-\epsilon A]^m \) if \( \epsilon \) is small enough.
2. Let \( Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij} \) for \( i = 1, 2, 3 \) and \( j = 1, 2 \), where the \( e_{ij} \) are i.i.d. \( N(0, \sigma^2) \).

(a) Find the least squares estimators of the cell means and their variances.

(b) Consider the null hypothesis \( H_0: \alpha_2 - \alpha_1 = \alpha_3 - \alpha_2 \). Give a test statistic for testing \( H_0 \) and state the distribution of your statistic under the null hypothesis.

(c) Suppose that Factor B is random with \( \beta_j \) replaced with independent random variables \( b_j \sim N(0, \sigma_b^2) \). The \( b_j \) and \( e_{ij} \) are mutually independent. Find an unbiased estimator of \( \alpha_2 - \alpha_1 \) in this situation. What is the variance of your estimator?

3. Consider the random intercept model

\[
Y_{jk} = \beta_0 + a_j + \beta_1 z_{jk} + e_{jk}, \quad j = 1, \ldots, J; \quad k = 1, \ldots, n.
\]

Here \( j \) indexes clusters and \( k \) indexes observations within clusters. The \( a_j \) are i.i.d. \( N(0, \sigma_a^2) \), the \( e_{jk} \) are i.i.d. \( N(0, \sigma_e^2) \), and the \( z_{jk} \) are observed nonrandom covariates with \( \bar{z}_j = 0 \) for each \( j \).

(a) Find the ordinary least squares estimators of \( \beta_0 \) and \( \beta_1 \). Show that they are unbiased and compute their variances.

(b) Propose estimators of \( \sigma_a^2 \) and \( \sigma_e^2 \). What are their distributions?

(c) If \( \sigma_a^2 \) and \( \sigma_e^2 \) are unknown, how might you improve on the ordinary least squares estimators of part (a)?

4. Let \( E[Y] = X\beta \) and \( \text{Var-Cov}[Y] = V \), where \( \beta \in \mathbb{R}^p \), \( Y \in \mathbb{R}^n \) and all matrices are of full rank. Let \( \mathcal{C}(X) \) denote the column space of \( X \). Let \( \hat{\beta}_{OLS} \) denote the ordinary least squares estimator \((X^TX)^{-1}X^TY\) and let \( H \) be the hat matrix. That is, \( HY = X\hat{\beta}_{OLS} \). If for all \( x \in \mathcal{C}(X) \) we have \( Vx \in \mathcal{C}(X) \), then the following conditions hold:

(a) \( \text{Cov}[(I-H)Y, \hat{\beta}_{OLS}] = 0 \).

(b) \( \hat{\beta}_{OLS} = \hat{\beta}_{BLUE} = (X^TV^{-1}X)^{-1}X^TV^{-1}Y \).
5. Let $Y_1, \ldots, Y_n$ be independent Poisson variables with means

$$\mu_i = \exp(\alpha + \beta x_i), \quad i = 1, \ldots, n.$$ 

(a) Express these data as a generalized linear model, identifying the link and variance functions.

(b) Show that the estimating equations for $\alpha$ and $\beta$ reduce to

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{\mu}_i, \quad \sum_{i=1}^{n} x_i Y_i = \sum_{i=1}^{n} x_i \hat{\mu}_i.$$ 

(b) Find the approximate distributions of $\hat{\alpha}$ and $\hat{\beta}$ as $n \to \infty$.

(c) How would you test $H_0: \beta = 0$? Give the formula for your test statistic and its (approximate) null distribution.

(d) How would you test whether the loglinear regression model fits the data? Provide a test statistic and its null distribution.

6. A population $\mathcal{U}$ consists of $N$ clusters, each of which contains $M$ elements. The population mean and variance of a variable $y$ are

$$\bar{y}_\mathcal{U} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \quad \text{and} \quad S^2_\mathcal{U} = \frac{1}{NM - 1} \sum_{i=1}^{N} \sum_{j=1}^{M} (y_{ij} - \bar{y}_\mathcal{U})^2.$$ 

The mean and variance of cluster $i$ are denoted by $\bar{y}_i$ and $S^2_i$.

(a) Write out the population ANOVA table, including expressions for sums of squares between and within clusters and the associated degrees of freedom. Denote the mean squares by $S^2_b$ and $S^2_w$.

(b) A simple random sample of $n$ clusters is chosen and the $y$ value of each element in a sampled cluster is observed. Find an unbiased estimator $\bar{y}_c$ of $\bar{y}_\mathcal{U}$ based on this sample and give a formula for its variance.

(c) Compare $\text{Var}[\bar{y}_c]$ to $\text{Var}[\bar{y}_{sre}]$, where $\bar{y}_{sre}$ is the mean of a simple random sample of $Mn$ elements selected from $\mathcal{U}$.
7. A simple random sample $S$ of size $n$ is drawn from a population $\mathcal{U}$ of size $N$ and data $(x_i, y_i), i \in S$, are collected. The population mean of $x$, denoted $\bar{x}_U$, is known. The goal is to estimate the population ratio $B = \bar{y}_U / \bar{x}_U$. Which of the following methods do you recommend for estimating $B$?

(a) Always use $\bar{y}_S / \bar{x}_U$.

(b) Sometimes use $\bar{y}_S / \bar{x}_U$ and sometimes use $\bar{y}_S / \bar{x}_S$.

(c) Always use $\bar{y}_S / \bar{x}_S$.

Justify your answer.

8. Data were collected in a salmonella reverse mutagenicity assay where the numbers of revertant colonies of TA98 Salmonella were observed on each of three replicate plates for doses of quinoline at levels 0, 10, 33, 100, 333, 1000. The data were in the form $(x_i, Y_{ij}), i = 1, \ldots, 6, j = 1, \ldots, 3$, where $Y$ was the number of colonies on a plate and $x = \log(1 + \text{dose})$. The goal was to determine the relationship of the transformed dose $x$ on the response $Y$. One scientist argued that the data should be reduced to $(x_i, \bar{Y}_i,)$ and a least squares line should be estimated. Another argued that the data should be treated as a one-way ANOVA using the $k$ levels of $x$ as treatments. Their results appear on the following page.

(a) Assume throughout that $\text{Var}[Y_{ij}] = \sigma^2$ for all $i, j$. Does either analysis provide a valid estimate of $\sigma^2$?

(b) Suppose that the linear regression model $Y_{ij} = \beta_0 + \beta_1 x_i + e_i$ is correct and that the error terms are i.i.d. with zero means and constant variance. Does the regression analysis of $\bar{Y}_i$ vs. $x_i$ provide valid estimates of the regression coefficients?

(c) Is there enough information in the printed output to test whether the linear model in (b) is inadequate to describe the data? If so, give a numerical expression for the test statistic and its distribution if linear regression is correct.

(d) What model would you recommend for this data?
LINEAR REGRESSION OF Y-BAR VS. LOG(1 + DOSE)

Call:
  lm(formula = colonies.avg ~ x)

Residuals:
     1       2       3       4       5       6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   19.823     6.562     3.021    0.0391 *
x            2.396     1.461     1.640    0.1764
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 8.112 on 4 degrees of freedom
Multiple R-squared:  0.4019,  Adjusted R-squared:  0.2524
F-statistic: 2.688 on 1 and 4 DF,  p-value: 0.1764

ONE WAY ANOVA OF Y-BAR VS. DOSE

Analysis of Variance Table

Response: colonies

            Df  Sum Sq Mean Sq  F value  Pr(>F)
factor(dose) 5 1320.4 264.089  2.9038   0.06047 .
Residuals    12 1091.3  90.944
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1