Applied Statistics (M.A. 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. The data \((x_i, Y_i), i = 1, \ldots, n,\) satisfy the simple linear regression model \(Y_i = \delta + \beta(x_i - \bar{x}) + e_i,\) where the \(e_i\) are i.i.d. \(N(0, \sigma^2).\)

   (a) Find the mean and covariance matrix of \((\hat{\delta}, \hat{\beta}),\) the ordinary least squares estimator of the parameter vector \((\delta, \beta).\)

   (b) Find a \(1 - \alpha\) confidence interval for \(\delta + \beta(x_0 - \bar{x}),\) where \(x_0\) is a value of \(x\) that may or may not be part of the original data set.

   (c) In the setting of part (b), find a prediction interval for \(Y_0,\) the actual \(Y\) value associated with \(x_0.\)
2. Consider the linear regression model \( Y_{ij} = \beta_0 + \beta_1 x_{ij} + e_{ij} \), in which \( i = 1, \ldots, m, \ j = 1, \ldots, n_i \), and the \( e_{ij} \) are i.i.d. with \( E[e_{ij}] = 0 \) and \( \text{Var} [e_{ij}] = \sigma^2 \). Suppose that only the means \((\bar{x}_i, \bar{Y}_i), \ i = 1, \ldots, m, \) are available.

(a) Show that the means \((\bar{x}_i, \bar{Y}_i)\) also satisfy a linear regression model with the same regression coefficients and find the moments of the error terms under this model.

(b) Find the BLUE of \( \beta_0 \) and \( \beta_1 \) using only the means. (You will have to use generalized least squares.) Derive the variance of your slope estimator.

(c) If the individual data \((x_{ij}, Y_{ij})\) were available, ordinary least squares applied to the individual data would yield an optimal estimator of \( \beta_1 \), according to the Gauss-Markov theorem. Compare the variance of your estimator in (b) to that of the ordinary least squares estimator using the individual data.

3. In the two way model \( Y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk} \) a defective design was used, yielding

\[
\begin{bmatrix}
Y_{111} \\
Y_{121} \\
Y_{211} \\
Y_{212} \\
Y_{232} \\
Y_{233}
\end{bmatrix} = \begin{bmatrix}
1 & 1 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 1
\end{bmatrix} \begin{bmatrix}
\mu \\
\alpha_1 \\
\alpha_2 \\
\beta_1 \\
\beta_2 \\
\beta_3
\end{bmatrix} + \text{e}.
\]

Certain factor combinations were not observed and combination \((2,3)\) was replicated three times. Assume the \( e_{ijk} \) are i.i.d. with \( E[e_{ijk}] = 0 \) and \( \text{Var} [e_{ijk}] = \sigma^2 \).

(a) Show that \( \alpha_1 - \alpha_2 \) and all contrasts among the \( \beta_j \) are estimable.

(b) Is the missing cell mean \( \mu + \alpha_2 + \beta_2 \) estimable?

(c) How many degrees of freedom are available to estimate \( \sigma^2 \)?
4. A population $\mathcal{U}$ has $N$ elements. Let $t_{y_{\text{str}}}$ be the population total of a variable $y$. A simple random sample $S$ of $n$ elements is drawn from the population without replacement.

(a) Propose an unbiased estimator $\hat{t}_y$ of $t_{y_{\text{str}}}$ and give a formula for its variance.

(b) Now assume that $y$ is roughly proportional to another variable $x$, that $x$ is observed for each sample element and that $t_{x_{\text{str}}}$ is known. What is the ratio estimator of $t_{y_{\text{str}}}$? Give an approximate formula for its variance. You may use large sample approximations in part (b).

5. Consider the random effects model $Y_{ij} = \mu + a_i + e_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, where the $a_i$ are i.i.d. $N(0, \sigma_a^2)$, the $e_{ij}$ are i.i.d. $N(0, \sigma_e^2)$ and the random terms in the model are mutually independent.

(a) Write out the ANOVA table for this model and find a set of unbiased point estimators for $(\sigma_a^2, \sigma_e^2)$. Are there any undesirable characteristics of your estimator of $\sigma_a^2$?

(b) Assuming $\mu$ and the variances $(\sigma_a^2, \sigma_e^2)$ are known, find the best predictor of $a_i$.

(c) Now assume the parameters are all unknown. How would you predict $a_i$?

6. A finite population $\mathcal{U}$ of size $N$ is divided into $H$ strata, $\mathcal{U}_h$, $h = 1, \ldots, H$. Stratum $\mathcal{U}_h$ contains $N_h$ elements. A simple random sample $S_h$ of $n_h$ elements is drawn from stratum $h$, $h = 1, \ldots, H$. The $S_h$ are drawn independently. The goal is to estimate the population total

$$t_y = \sum_{h=1}^{H} \sum_{i \in S_h} y_{hi}.$$ 

(a) Find an unbiased estimator of $t_y$ and give its variance. You may assume the stratum sizes $N_h$ are large enough to neglect finite population corrections.

(b) What are the optimal sample sizes $n_h$ if an observation from stratum $h$ costs $c_h$ and for a known $C$, the total cost of sampling must satisfy the constraint $\sum_h c_h n_h \leq C$?
7. Consider the linear model $Y = X\beta + e$ where

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} \quad \text{and} \quad Y = \begin{bmatrix} -1 \\ 3 \\ 2 \\ 4 \\ 3 \\ 1 \end{bmatrix}.$$

Assume the $e_i$ are i.i.d. $N(0, \sigma^2)$. Calculations yield $Y^T Y = 40$, $\bar{Y} = 2$, $$(X^T X)^{-1} = \frac{1}{12} \begin{bmatrix} 5 & -1 & -1 \\ -1 & 5 & -1 \\ -1 & -1 & 5 \end{bmatrix}, \quad X^T Y = \begin{bmatrix} 4 \\ 6 \\ 8 \\ 6 \end{bmatrix}, \quad \hat{\beta} = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 1 \end{bmatrix}.$$

(a) Give an unbiased estimate of $\sigma^2$ and its degrees of freedom.

(b) Is $\hat{\beta}$ the unique solution of the normal equations? Why or why not?

(c) It is desired to test the hypothesis $H_0$ that the components of $\beta$ are equal. Express the hypothesis in terms of a suitable matrix equation $C\beta = 0$, compute the test statistic and state its distribution under $H_0$.

(d) Estimate $\beta$ under the null hypothesis of part (c).

8. Data were collected on a sample of 97 men who were due to receive a radical prostatectomy. The measured variables included lvavol=log(cancer volume), lweight=log(prostate weight), age, lbph=log(benign prostatic hyperplasia amount), svi=indicator of seminal vesicle invasion, lcp=log(capsular penetration), gleason=Gleason score and lpsa=log(prostate specific antigen). The goal of the study was to relate lpsa to the other variables, which represent indication's of the severity of the patients prostate cancer. The output of a multiple regression analysis appears on the next page.
lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
     gleason, data = prostate)

Residuals:
   Min     1Q Median     3Q    Max
-1.788027 -0.369331  0.003023  0.434360  1.621603

Coefficients:    Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.02415     1.13313   0.021 0.98304
lcavol           0.57471     0.08712   6.597 2.92e-09 ***
lweight          0.45260     0.17005   2.662 0.00923 **
age              -0.01812     0.01108  -1.636 0.10542
lbph             0.10886     0.05844   1.863 0.06579 .
svi              0.79783     0.24241   3.291 0.00143 **
lcp             -0.07488     0.08599  -0.871 0.38619
gleason          0.14591     0.12292   1.187 0.23837

---
Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .  0.1  1

Residual standard error: 0.7066 on 89 degrees of freedom
Multiple R-squared: 0.6506,   Adjusted R-squared: 0.6232
F-statistic: 23.68 on 7 and 89 DF,  p-value: < 2.2e-16

(a) Does the model significantly relate lpsa to the other predictors? Explain your answer.

(b) Are there predictors that do not seem to contribute to prediction of lpsa? Which predictors would you consider removing from the model? Explain.

(c) The model states that the presence of seminal vesicle invasion (svi=1) has a constant effect on the level of lpsa regardless of the level of any of the other predictors. How would you test this hypothesis against the alternative that the effect of svi on lpsa varies with age? State models for the null and alternative hypotheses, describe how to compute a test statistic and state the distribution of your test statistic under the null hypothesis.