DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
AUGUST 2011

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. Consider the linear model $Y = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where the $n \times p$ matrix has rank $r < p$. Let $\psi = \mathbf{c}^T\boldsymbol{\beta}$ be a parametric function. Show that the following conditions are equivalent.

(a) ψ is estimable.

(b) $\mathbf{c}^T = \mathbf{a}^T\mathbf{X}$ for some $\mathbf{a} \in R^n$.

(c) $\mathbf{c}^T = \mathbf{r}^T\mathbf{X}^T\mathbf{X}$ for some $\mathbf{r} \in R^p$.

2. Let $Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ij}$, $i = 1, \quad, I$, $j = 1, \quad, J$. The error terms are i.i.d. with mean zero and variance σ^2. A statistician fits a main effect model to these data, ignoring the presence of the interaction term γ_{ij}.

(a) Is the usual least squares estimate of $\alpha_1 - \alpha_2$ unbiased under this misspecified model? Justify your answer.

(b) Find the expectation of the usual estimate of σ^2 under a main effects model.

(c) Make the additional assumption that $I = J = 2$. Is $\alpha_1 - \alpha_2$ estimable in this situation? Justify your answer.

3 Consider the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \boldsymbol{\beta} + \mathbf{e}$$

where the e_i are i.i.d. $N(0, \sigma^2)$.

(a) Find the least squares estimates of $\boldsymbol{\beta}$ and an unbiased estimator of σ^2.

(b) Give an explicit formula for the statistic used to test $H_0 : \beta_1 = \beta_2 = \beta_3$ and the distribution of the test statistic under H_0.

(c) Does the power of your test depend on β_4? Justify your answer.

4. An agricultural experiment was conducted to compare I treatments. It is believed that the response is also affected by a measured covariate z. The data are (z_{ij}, Y_{ij}), $i = 1, \ldots, I$, $j = 1, \ldots, n_i$, where

$$Y_{ij} = \beta_i + \gamma(z_{ij} - \bar{z}_i) + e_{ij}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, n_i$$

and the e_{ij} are i.i.d. $N(0, \sigma^2)$.

(a) Derive the least squares estimators and an estimator for the variance

(b) Give a test statistic for $H_0 : \beta_1 = \cdots = \beta_I$. State the distribution of your test statistic under the null hypothesis.

(c) Find the best linear unbiased estimator of the contrast $\psi = \sum_{i=1}^{I} c_i \beta_i$. What is the variance of your estimator?

5. A simple random sample $\mathcal{S}$ of size n is selected from a finite population $\mathcal{U}$ of size N. The variables x_i and y_i are measured on each unit $i \in \mathcal{S}$. The population ratio

$$B = \frac{t_y}{t_x} = \frac{\sum_{i \in \mathcal{U}} y_i}{\sum_{i \in \mathcal{U}} x_i}$$

is estimated by its sample analog

$$\hat{B} = \frac{\hat{t}_y}{\hat{t}_x} = \frac{\sum_{i \in \mathcal{S}} y_i}{\sum_{i \in \mathcal{S}} x_i}$$

(a) Assuming that n is large, argue that $\hat{B}$ is approximately unbiased and that

$$\operatorname{Var} \hat{B} \doteq \left(\frac{1 - n/N}{n}\right) \frac{S_y^2 - 2BRS_xS_y + B^2S_X^2}{\bar{x}_{\mathcal{U}}^2}$$

where $\bar{x}_{\mathcal{U}} = t_x/N$ is the population mean of x, $S_x^2 = (N-1)^{-1} \sum_{i \in \mathcal{S}} (x_i - \bar{x}_{\mathcal{U}})^2$ is the population variance of x with similar definitions for $\bar{y}_{\mathcal{U}}$, S_y^2, S_{xy}, and $R = S_{xy}/(S_xS_y)$ is the population correlation.

(b) Assuming t_x is known, use the results of (a) to produce the ratio estimate $\hat{t}_{yr}$ of t_y and its approximate variance. When is $\hat{t}_{yr}$ a better estimator than $(N/n) \sum_{i \in \mathcal{S}} y_i$?

6. Consider the linear mixed model $Y = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \mathbf{e}$ where $\boldsymbol{\beta}$ is a vector of fixed parameters, $\mathbf{X}$ is an $n \times p$ matrix of full rank, $\mathbf{Z}$ is an $n \times q$ matrix of full rank, the random q-vector $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$, the random n-vector $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, and $\mathbf{u}$ and $\mathbf{e}$ are independent

(a) Find Var-Cov[$\mathbf{Y}$], Cov[$\mathbf{u}, \mathbf{Y}$] and $E[\mathbf{u} \mid \mathbf{Y}]$.

(b) In terms of $\mathbf{X}$, $\mathbf{Z}$, $\boldsymbol{\beta}$ and σ^2, find the best predictor of $\mathbf{u}$. That is, find the function $t(\mathbf{Y})$ which minimizes $E[(t(\mathbf{Y}) - \mathbf{u})^T(t(\mathbf{Y}) - \mathbf{u})]$. Prove your answer.

7. A finite population $\mathcal{U}$ is divided into two strata $\mathcal{U}_1$ and $\mathcal{U}_2$ of known sizes N_1 and N_2, respectively. If simple random samples $\mathcal{S}_1$ and $\mathcal{S}_2$ of sizes n_1 and n_2 are drawn independently from $\mathcal{U}_1$ and $\mathcal{U}_2$, the population mean $\bar{y}_{\mathcal{U}}$ is estimated by

$$\bar{y}_{st} = (N_1/N)\bar{y}_1 + (N_2/N)\bar{y}_2$$

where the sample stratum means are defined as $\bar{y}_h = (1/n_h)\sum_{i \in \mathcal{S}_h} y_{hi}$, $h = 1, 2$. The stratum means and variances are defined as

$$\bar{y}_{\mathcal{U}_h} = \frac{1}{N_h}\sum_{i \in \mathcal{U}_h} y_{hi}, \qquad S_h^2 = \frac{1}{N_h - 1}\sum_{i \in \mathcal{U}_h}(y_{hi} - \bar{y}_h)^2.$$

(a) Derive Var $[\bar{y}_{st}]$.

(b) For a fixed sample size n, derive the optimum stratum sample sizes n_h^*, $h = 1, 2$, such that Var $[\bar{y}_{st}]$ is minimized. Denote the minimum value by V_{min}. Assume that finite sample corrections may be ignored.

(c) If a simple random sample of size n is selected from $\mathcal{U}$, let V_{ran} be the variance of the sample mean. Let n_1 and n_2 be the (approximate) numbers of observations from each stratum under simple random sampling. Assume n is very large. Define $\phi = (n_1/n_2)/(n_1^*/n_2^*)$. Show that

$$V_{min}/V_{ran} \geq 4\phi/(1 + \phi^2).$$

8. A chemical manufacturing experiment attempted to maximize a response Y with respect to controllable quantities z_1, z_2, z_3. A quadratic regression model

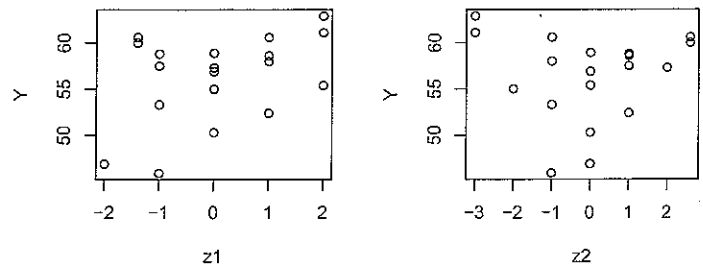
$$\begin{aligned} Y &= \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_{11} z_1^2 + \beta_{22} z_2^2 + \beta_{33} z_3^2 \\ &\quad + \beta_{12} z_1 z_2 + \beta_{13} z_1 z_3 + \beta_{23} z_2 z_3 + e \end{aligned}$$

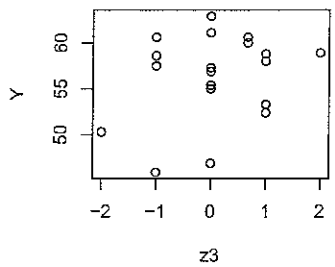
was fitted to 19 independent observations. It was assumed that the errors e_i were i.i.d. $N(0, \sigma^2)$. Plots of the data and numerical and graphic output appear on the following pages.

(a) Identify possible outliers and influence points. Explain your answers.

(b) Assuming all observations are valid, how would you attempt to simplify the model? Explain what statistical procedures you would employ to check whether the simplified model fits the data as well as the full model.

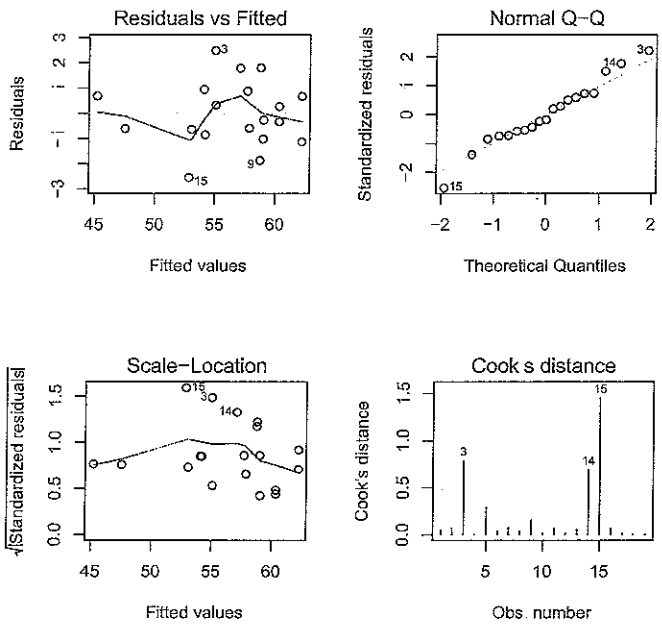
(c) Theoretical considerations suggest that the response depends only on the quantities z_1, $w = z_2 + z_3$ and second degree terms involving these quantities. In terms of the coefficients in the full model, formulate appropriate hypotheses to test this theory. How would the test statistic be computed? What would be its null distribution?






```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  58.7680     1.2110  48.527 3.36e-12 ***
z1            1.8914     0.4186   4.518 0.001451 **
z2            0.9581     0.3442   2.784 0.021270 *
z3            1.0665     0.4519   2.360 0.042608 *
I(z1^2)      -1.8690     0.4350  -4.297 0.001999 **
I(z1 * z2)   -2.7195     0.5220  -5.210 0.000557 ***
I(z2^2)      -0.6987     0.3488  -2.003 0.076164 .
I(z1 * z3)   -2.1811     0.6296  -3.464 0.007114 **
I(z3^2)      -0.9468     0.4730  -2.002 0.076346 .
I(z2 * z3)   -1.2244     0.5974  -2.049 0.070682 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.817 on 9 degrees of freedom
Multiple R-squared: 0.9258,     Adjusted R-squared: 0.8516
F-statistic: 12.48 on 9 and 9 DF,  p-value: 0 0004388
```




```
> box.youle.diags
   std.resids      cooks.d hat.values
1   0.5887755 0.049262634  0.5869610
2  -0.7220727 0.070187122  0.5737710
3   2.2075471 0.783571461  0.6165495
4  -0.1791915 0.001708150  0.3472477
5   1.4970315 0.285543369  0.5602694
6  -0.7378382 0.040341475  0.4256238
7   0.7392495 0.073357131  0.5730752
8  -0.5357128 0.036749281  0.5615027
9  -1.3793690 0.152157189  0.4443544
10  0.2852295 0.012492197  0.6056005
11 -0.5754450 0.064413166  0.6604659
12 -0.4318004 0.014383278  0.4354816
13  0.7269639 0.050779126  0.4900194
14  1.7628345 0.688384487  0.6889749
15 -2.5325034 1.449747428  0.6932926
16 -0.8440880 0.060258095  0.4582136
```

```
17  0.5019763 0.021311124  0.4582136
18 -0.2346181 0.003828235  0.4101915
19  0.1954162 0.002655810  0.4101915
```