DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
JANUARY 2011

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. Consider the linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \boldsymbol{\varepsilon}.$$

This can be thought of as a regression model with $Y_i = \beta_0 + \sum_j x_{ij}\beta_j + \varepsilon_i$

(a) Show that all parameters are estimable and write out the least squares equations.

(b) Suppose that one wanted to augment the model by adding a term $\beta_{12}x_1x_2$. How does this term affect estimability of the parameters?

2. A drug is administered to each of n subjects and a response is observed at times $t = 1, 2, \ldots, T$. It is believed that the effect of the drug occurs gradually over time, so the observation on subject i at time t is modeled as

$$Y_{it} = \mu + a_i + \beta(t - \bar{t}) + \varepsilon_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T, \quad \bar{t} = (T+1)/2.$$

The random subject effects a_i, $i = 1, \ldots, n$, are i.i.d. $N(0, \sigma_a^2)$ and the error terms ε_{it} are i.i.d. $N(0, \sigma_e^2)$.

(a) Suppose the data are reduced to the time averages $\bar{Y}_t$, $t = 1, \ldots, T$, and a least squares regression line $\hat{Y}_t = \alpha + \beta t$ is fitted to the reduced data. What is the joint distribution of the resulting estimates $\hat{\alpha}$ and $\hat{\beta}$? Verify that $\hat{\beta}$ is not a function of the a_i.

(b) Starting from the standard ANOVA table for a two way layout, derive an ANOVA table for this model with sums of squares for subject-to-subject variation, regression and residuals. Show that the mean square for residuals is an unbiased estimator for σ_e^2. What are the degrees of freedom for residuals?

(c) Use the results of (b) to derive a confidence interval for β

3. Let $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, where the parameters $\mu, \alpha_i, \beta_j, \gamma_{ij}$ are unrestricted. The error terms ε_{ijk} are i.i.d. $N(0, \sigma^2)$.

(a) Show that $\alpha_1 - \alpha_2$ is not estimable but that $\gamma_{11} - \gamma_{12} - \gamma_{21} + \gamma_{22}$ is estimable.

(b) If the additive model holds, that is, if $\gamma_{ij} = 0$ for all i, j, show that $\alpha_1 - \alpha_2$ is now estimable and give a confidence interval for this contrast.

(c) State the usual test for additivity and give its distribution under the general alternative.

4. Let $Y_{11}, \ldots, Y_{1m}$ be i.i.d. $N(\mu_1, \sigma^2)$ and let $Y_{21}, \ldots, Y_{2n}$ be i.i.d. $N(\mu_2, k\sigma^2)$, where μ_1, μ_2, σ^2 are unknown parameters and k is a known constant. Find the BLUE of $\mu_1 - \mu_2$ and provide a confidence interval for this quantity

5. For population values $y_1, \ldots, y_N$, consider the ith systematic sample of size n:

$$y_i, y_{i+k}, y_{i+2k}, \ldots, y_{i+(n-1)k}, \quad i = 1, \ldots, k, \quad N = kn.$$

Let y_{ij} denote the jth element, $j = 1, \ldots, n$, of systematic sample i.

(a) Obtain the basic ANOVA table for the above population.

(b) Show that the population variance S^2 can be expressed in terms of the between and within sample sums of squares.

(c) Define

$$S^2_{wsy} = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2.$$

Obtain a condition in terms of S^2_{wsy} under which the mean of a systematic sample is more precise than the mean from a simple random sample. Interpret your result.

6. In most cases, positive geophysical data such as duration of snowstorms have skewed distributions. Suppose that for some positive geophysical data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ and for some $\lambda \in (-3, 3)$ the transformation

$$Y_i^{(\lambda)} = g(Y_i, \lambda) = \boldsymbol{x}_i' \boldsymbol{\beta} + \varepsilon_i$$

gives a linear model with independent errors $\varepsilon_i \sim N(0, \sigma^2)$, where

$$g(y, \lambda) = \begin{cases} (y^\lambda - 1)/\lambda, & \text{if } \lambda \neq 0, \\ \log y & \text{if } \lambda = 0. \end{cases}$$

(a) Obtain the likelihood for the original data $\boldsymbol{Y}$ in terms of $\boldsymbol{Y}^{(\lambda)} = (Y_1^{(\lambda)}, \ldots, Y_n^{(\lambda)})'$.

(b) For a fixed λ, write down the maximized log-likelihood.

(c) Suggest a way to estimate λ.

7. A sample $\mathcal{S}$ of size n is drawn sequentially from a population $\mathcal{U}$ of size N as follows:

(i) The first element drawn is element k with probability p_k, $k = 1, \ldots, N$.

(ii) A simple random sample of size $n-1$ is drawn from the remaining $N-1$ elements without replacement.

Here $p_1, \ldots, p_N$ are nonnegative numbers with $\sum_{k=1}^{N} p_k = 1$.

(a) What is the probability that $\mathcal{S}$ contains element k?

(b) What is the probability that $\mathcal{S}$ contains both elements j and k, $j \neq k$?

(c) What is the probability that $\mathcal{S} = \{k_1, \ldots, k_n\}$?

8. An experimenter wishes to fit the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

using least squares, subject to the restriction that $\beta_1 = 1$. He asks specifically if he can just fit

$$W = Y - x_1 = \beta_0 + \beta_2 x_2 + \varepsilon$$

using least squares to get what he wants. Prove that his proposal will work.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
JANUARY 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 model $Y_i = \beta_0 + \sum_j x_{ij}\beta_j + \varepsilon_i$ in matrix form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \boldsymbol{\varepsilon}$$

(a) Show that all parameters are estimable and write out the least squares equations.

(b) If one augmented the model by adding a term $\beta_{12}x_1x_2$, how would this term affect estimability of the parameters? What if instead one added the term $\beta_{123}x_1x_2x_3$?

2. A drug is administered to each of n subjects and a response is observed at times $t = 1, 2, \ldots, T$. It is believed that the effect of the drug occurs gradually over time, so the observation on subject i at time t is modeled as

$$Y_{it} = \mu + a_i + \beta(t - \bar{t}) + \varepsilon_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T, \quad \bar{t} = (T+1)/2$$

The random subject effects a_i, $i = 1, \ldots, n$, are i.i.d. $N(0, \sigma_a^2)$ and the error terms ε_{it} are i.i.d. $N(0, \sigma_e^2)$.

(a) Suppose the data are reduced to the time averages $\bar{Y}_t$, $t = 1, \ldots, T$, and a least squares regression line $\hat{Y}_t = \alpha + \beta t$ is fitted to the reduced data. What is the joint distribution of the resulting estimates $\hat{\alpha}$ and $\hat{\beta}$? Verify that $\hat{\beta}$ is not a function of the a_i.

(b) Starting from the standard ANOVA table for a two way layout, derive an ANOVA table for this model with sums of squares for subject-to-subject variation, regression and residuals. Show that the mean square for residuals is an unbiased estimator for σ_e^2. What are the degrees of freedom for residuals?

(c) Use the results of (b) to derive a confidence interval for β.

3. Let $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, where the parameters $\mu, \alpha_i, \beta_j, \gamma_{ij}$ are unrestricted. The error terms ε_{ijk} are i.i.d. $N(0, \sigma^2)$.

(a) Show that $\alpha_1 - \alpha_2$ is not estimable but that $\gamma_{11} - \gamma_{12} - \gamma_{21} + \gamma_{22}$ is estimable.

(b) Find the BLUE of $E[Y_{ijk}]$ if the additive model holds, that is, if $\gamma_{ij} = 0$ for all i, j.

(c) State the usual test for additivity and give its distribution under the general alternative.

4. Let $Y_{11}, \ldots, Y_{1m}$ be i.i.d. $N(\mu_1, \sigma^2)$ and let $Y_{21}, \ldots, Y_{2n}$ be i.i.d. $N(\mu_2, k\sigma^2)$, where μ_1, μ_2, σ^2 are unknown parameters and k is a known constant. Find the BLUE of $\mu_1 - \mu_2$ and provide a confidence interval for this quantity.

5. For population values $y_1, \ldots, y_N$, consider the ith systematic sample of size n:

$$y_i, y_{i+k}, y_{i+2k}, \ldots, y_{i+(n-1)k}, \quad i = 1, \ldots, k, \quad N = kn$$

Let y_{ij} denote the jth element, $j = 1, \ldots, n$, of systematic sample i.

(a) Obtain the basic ANOVA table for the above population.

(b) Show that the population variance S^2 can be expressed in terms of the between and within sample sums of squares

(c) Define

$$S^2_{wsy} = \frac{1}{k(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_i)^2.$$

Obtain a condition in terms of S^2_{wsy} under which the mean of a systematic sample is more precise than the mean from a simple random sample. Interpret your result.

6. In most cases, positive geophysical data such as duration of snowstorms have skewed distributions. Suppose that for some positive geophysical data $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ and for some $\lambda \in (-3, 3)$ the transformation

$$Y_i^{(\lambda)} = g(Y_i, \lambda) = \boldsymbol{x}_i' \boldsymbol{\beta} + \varepsilon_i$$

gives a linear model with independent errors $\varepsilon_i \sim N(0, \sigma^2)$, where

$$g(y, \lambda) = \begin{cases} (y^\lambda - 1)/\lambda, & \text{if } \lambda \neq 0, \\ \log y & \text{if } \lambda = 0. \end{cases}$$

(a) Obtain the likelihood for the original data $\boldsymbol{Y}$ in terms of $\boldsymbol{Y}^{(\lambda)} = (Y_1^{(\lambda)}, \ldots, Y_n^{(\lambda)})'$.

(b) For a fixed λ, write down the maximized log-likelihood.

(c) Suggest a way to estimate λ

7. A sample $\mathcal{S}$ of size n is drawn sequentially from a population $\mathcal{U}$ of size N as follows:

(i) The first element drawn is element k with probability p_k, $k = 1, \ldots, N$.

(ii) A simple random sample of size $n-1$ is drawn from the remaining $N-1$ elements without replacement.

Here $p_1, \ldots, p_N$ are nonnegative numbers with $\sum_{k=1}^{N} p_k = 1$.

(a) What is the probability that $\mathcal{S}$ contains element k?

(b) What is the probability that $\mathcal{S}$ contains both elements j and k, $j \neq k$?

(c) What is the probability that $\mathcal{S} = \{k_1, \ldots, k_n\}$?

8. An experimenter wishes to fit the quadratic regression model

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon.$$

He claims to know that the response at $x = 1$ is exactly 10, so that

$$10 = \beta_0 + \beta_1 + \beta_2$$

with no error. Therefore he substitutes $\beta_0 = 10 - \beta_1 - \beta_2$ into the quadratic regression model to obtain

$$W = Y - 10 = \beta_1 z_1 + \beta_2 z_2 + \varepsilon,$$

where $z_1 = x - 1$ and $z_2 = x^2 - 1$. He then fits the second model by ordinary least squares and claims that he has succeeded in fitting the first model subject to the restriction that the response at $x - 1$ is exactly 10. Is he correct?

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
AUGUST 2010

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. Let $Y_{ij} = \mu_i + e_{ij}$, $i = 1, \ldots, 4$, $j = 1, \ldots, n$. Assume that the e_{ij} are i.i.d. $N(0, \sigma^2)$

(a) Write out the ANOVA table, including the sums of squares, mean squares, degrees of freedom and expected mean squares.

(b) Find the test statistic for testing H_0: $\mu_1 = \mu_2$ and $\mu_3 = \mu_4$. What is its distribution under H_0?

(c) What is the distribution the test statistic of part (b) under the alternative $\mu_2 - \mu_1 = \sigma = \mu_4 - \mu_3$?

2. The multiple regression model

$$Y_i = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i + e_i, \quad i = 1, \ldots, n,$$

is written in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, where $\boldsymbol{X}$ has full rank. The *ridge regression* estimator is $\tilde{\boldsymbol{\beta}} = [\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I}]^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, where k is a small positive constant chosen by the statistician.

(a) Calculate $E[\tilde{\boldsymbol{\beta}}]$ and Var-Cov$[\tilde{\boldsymbol{\beta}}]$.

(b) Compute $E[(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})]$.

(c) Let $\hat{\boldsymbol{\beta}}$ be the usual least squares estimator of $\boldsymbol{\beta}$. Show that for some k,

$$E[(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})] < E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})].$$

Hint: You may use the fact that $[\boldsymbol{I} + k\boldsymbol{A}]^{-1} = \sum_{j=0}^{\infty}(-1)^j(k\boldsymbol{A})^j$ if $|k|$ is small enough. In this case the infinite series is convergent.

(d) Does the result of (c) violate the Gauss-Markov Theorem? Explain.

3. A population $\mathcal{U}$ is partitioned into strata $\mathcal{U}_h$, $h = 1, \ldots, H$, of known sizes N_h. From each stratum a simple random sample $\mathcal{S}_h$ of n_h clusters is drawn. Each element of a sampled cluster is observed. The data are (y_{hij}, z_{hij}), $h = 1, \ldots, H$, $i \in \mathcal{S}_h$, $j = 1, \ldots, M_{hi}$, where y_{hij} is the y-value associated with element j of cluster i in stratum h and z_{hij} is defined similarly. The known quantity M_{hi} is the number of elements in cluster i of stratum h. The goal is to estimate the ratio of population totals

$$B = \frac{t_y}{t_z} = \frac{\sum_{h=1}^{H}\sum_{i\in\mathcal{U}_h}\sum_{j=1}^{M_{ij}} y_{hij}}{\sum_{h=1}^{H}\sum_{i\in\mathcal{U}_h}\sum_{j=1}^{M_{ij}} z_{hij}}.$$

(a) Propose a suitable estimator $\hat{B}$ for the ratio B.

(b) Propose an estimator of Var $[\hat{B}]$.

4. In a study of brand variability, boxes of tissues, all of the same brand, were bought in three cities, chosen by design. A random sample of six tissues was selected from each box and the breaking strengths Y_{ijk} of the sampled tissues were recorded. The data are partially tabulated below. Note that Box 1 from City 1 is not the same as Box 1 from City 2.

City 1		City 2			City 3			
Box 1	Box 2	Box 1	Box 2	Box 3	Box 1	Box 2	Box 3	Box 4
1.39	1.72	2.44	2.27	2.46	1.36	1.59	1.73	1.53
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(a) Let Y_{ijk} denote the response for Tissue k chosen from Box j in City i. Write an appropriate model for the Y_{ijk}. Indicate which factors are fixed and which factors are random.

(b) Write out the ANOVA table for your model. Provide formulas for the sums of squares and degrees of freedom.

(c) Estimate the mean breaking strength of tissues sold in City 1. What is the variance of your estimator?

(d) In terms of your model, what is the variance of $Y_{ijk} - Y_{ijl}$, $k \neq l$? How would you estimate this variance?

5. In an experiment to compare m treatments, the response was a binary indicator of success. The data were independent Bernoulli variables Y_{ij} such that $Y_{ij} \sim \text{Bernoulli}(\pi_i)$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. The problem is to test the hypothesis of equal response probabilities: $H_0 : \pi_1 = \cdots \pi_m$.

(a) Reformulate this problem in terms of a generalized linear model. Identify the link function and linear predictor.

(b) Obtain the likelihood ratio test statistic $-2 \log \Lambda$ for testing H_0, and describe its distribution under the hypothesis.

(c) Arrange the data in a $2 \times m$ table of of counts of successes and failures. Show how to to test H_0 using the well known Pearson χ^2 test statistic.

(d) Using the notation $\hat{\pi}_k = \bar{y}_k$ and $\hat{\pi} = \bar{y}$, show that the two test statistics $-2 \log \Lambda$ and χ^2 are practically the same when $\hat{\pi}_i$ is close to $\hat{\pi}$.

Hint: Use a Taylor series expansion of $x \log(c/x)$.

6 Suppose that a simple linear regression model $Y = \beta_0 + \beta_1 x + e$ is fitted to data $(x_1, Y_1), \ldots, (x_n, Y_n)$, resulting in estimates

$$\hat{\beta}_{0n} = \bar{Y}_n, \quad \hat{\beta}_{1n}$$

and

$$s_n^2 = \frac{1}{n-2} \sum (Y_i - \bar{Y}_n - \hat{\beta}_{1n} x_i)^2.$$

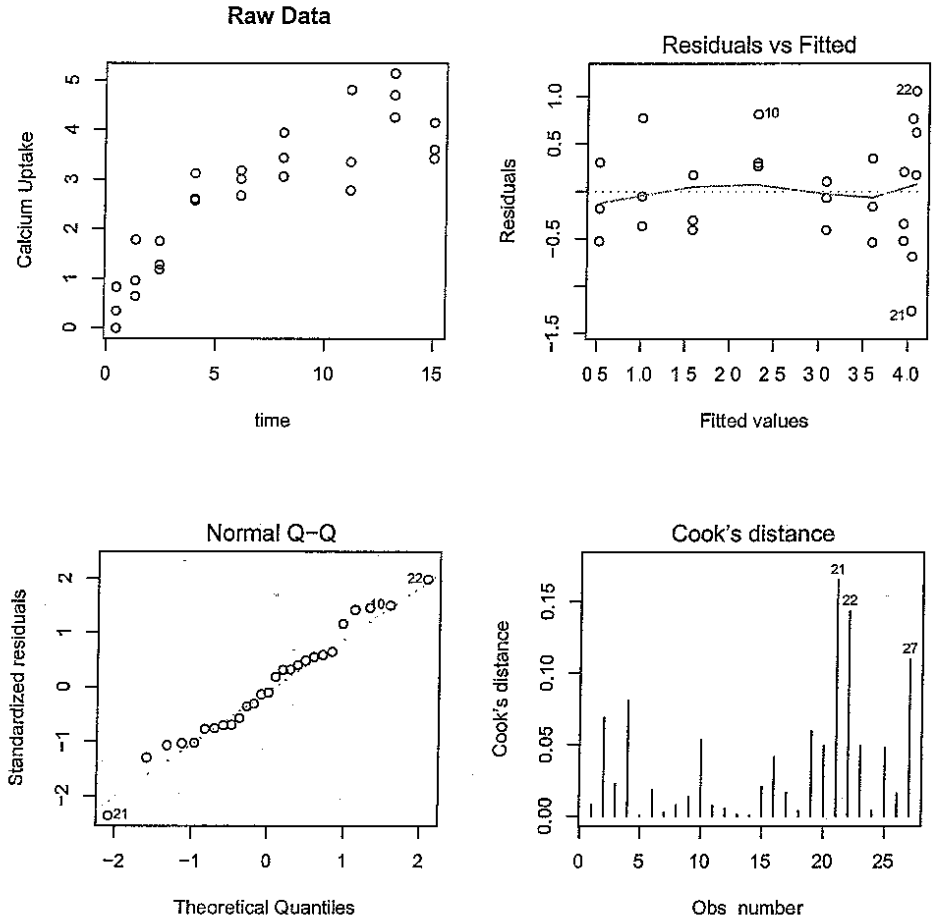
A new data point (x_{n+1}, Y_{n+1}) is observed. Derive formulas for $\hat{\beta}_{0,n+1}$, $\hat{\beta}_{1,n+1}$ and s_{n+1}^2 involving *only summary statistics from the original regression analysis and the new data point*

7. Data on calcium uptake in plants was collected in an experiment in which plant cells were suspended in a solution of radioactive calcium for a certain length of time and then the amount of radioactive calcium that was absorbed by the cells was measured. The experiment was repeated independently with 9 different times of suspension, each replicated 3 times. The calcium uptake was modeled as a regression of calcium uptake Y_{ij} against time t_i, $i = 1, \ldots, 9$, $j = 1, 2, 3$. The regression output appears below, and plots of the raw data and some diagnostic plots appear on the following page

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.26407 -0.38755 -0.05378  0.29999  1.05142

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.25326    0.25520   0.992 0.330901
time         0.61195    0.08826   6.934  3.6e-07 ***
I(time^2)   -0.02437    0.00565  -4.313 0.000238 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.5577 on 24 degrees of freedom
Multiple R-squared: 0.8598,     Adjusted R-squared: 0.8481
F-statistic: 73.58 on 2 and 24 DF,  p-value: 5.78e-11
```



(a) Are there possible influential points in the sample? Which ones? Why might they be influential? How might you assess the effect of such points?

(b) To check model adequacy, a one way ANOVA was run using time as a discrete factor, yielding an MSE of 0.5162 and an R^2 of 0.9099. Based on this information, how would you test model adequacy? Calculate the test statistic and its distribution if the model is correct.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
AUGUST 2010

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 $Y_{ij} = \mu_i + e_{ij}$, $i = 1, \ldots, 4$, $j = 1, \ldots, n$. Assume that the e_{ij} are i.i.d $N(0, \sigma^2)$.

(a) Write out the ANOVA table, including the sums of squares, mean squares, degrees of freedom and expected mean squares.

(b) Find the test statistic for testing H_0: $\mu_1 = \mu_2$ and $\mu_3 = \mu_4$. What is its distribution under H_0?

(c) What is the distribution the test statistic of part (b) under the alternative $\mu_2 - \mu_1 = \sigma = \mu_4 - \mu_3$?

2. The multiple regression model

$$Y_i = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i + e_i, \quad i = 1, \ldots, n,$$

is written in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, where $\boldsymbol{X}$ has full rank. The *ridge regression* estimator is $\tilde{\boldsymbol{\beta}} = [\boldsymbol{X}^T\boldsymbol{X} + k\boldsymbol{I}]^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, where k is a small positive constant chosen by the statistician.

(a) Calculate $E[\tilde{\boldsymbol{\beta}}]$ and Var-Cov$[\tilde{\boldsymbol{\beta}}]$

(b) Compute $E[(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})]$.

(c) Let $\hat{\boldsymbol{\beta}}$ be the usual least squares estimator of $\boldsymbol{\beta}$. Show that for some k,

$$E[(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})] < E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})].$$

Hint: You may use the fact that $[\boldsymbol{I} + k\boldsymbol{A}]^{-1} = \sum_{j=0}^{\infty}(-1)^j(k\boldsymbol{A})^j$ if $|k|$ is small enough. In this case the infinite series is convergent.

(d) Does the result of (c) violate the Gauss-Markov Theorem? Explain.

3. A population $\mathcal{U}$ is partitioned into strata $\mathcal{U}_h$, $h = 1, \ldots, H$, of known sizes N_h. From each stratum a simple random sample $\mathcal{S}_h$ of n_h clusters is drawn. Each element of a sampled cluster is observed. The data are (y_{hij}, z_{hij}), $h = 1, \ldots, H$, $i \in \mathcal{S}_h$, $j = 1, \ldots, M_{hi}$, where y_{hij} is the y-value associated with element j of cluster i in stratum h and z_{hij} is defined similarly. The known quantity M_{hi} is the number of elements in cluster i of stratum h. The goal is to estimate the ratio of population totals

$$B = \frac{t_y}{t_z} = \frac{\sum_{h=1}^{H}\sum_{i\in\mathcal{U}_h}\sum_{j=1}^{M_{ij}} y_{hij}}{\sum_{h=1}^{H}\sum_{i\in\mathcal{U}_h}\sum_{j=1}^{M_{ij}} z_{hij}}$$

(a) Propose a suitable estimator $\hat{B}$ for the ratio B.

(b) Propose an estimator of Var $[\hat{B}]$.

4. In a study of brand variability, boxes of tissues, all of the same brand, were bought in three cities, chosen by design. A random sample of six tissues was selected from each box and the breaking strengths Y_{ijk} of the sampled tissues were recorded. The data are partially tabulated below. Note that Box 1 from City 1 is not the same as Box 1 from City 2.

City 1		City 2			City 3			
Box 1	Box 2	Box 1	Box 2	Box 3	Box 1	Box 2	Box 3	Box 4
1.39	1.72	2.44	2.27	2.46	1.36	1.59	1.73	1.53
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(a) Let Y_{ijk} denote the response for Tissue k chosen from Box j in City i. Write an appropriate model for the Y_{ijk}. Indicate which factors are fixed and which factors are random.

(b) Write out the ANOVA table for your model. Provide formulas for the sums of squares and degrees of freedom.

(c) Estimate the mean breaking strength of tissues sold in City 1. What is the variance of your estimator?

(d) In terms of your model, what is the variance of $Y_{ijk} - Y_{ijl}$, $k \neq l$? How would you estimate this variance?

5 In an experiment to compare m treatments, the response was a binary indicator of success. The data were independent Bernoulli variables Y_{ij} such that $Y_{ij} \sim \text{Bernoulli}(\pi_i)$, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$. The problem is to test the hypothesis of equal response probabilities: $H_0 : \pi_1 = \cdots \pi_m$.

(a) Reformulate this problem in terms of a generalized linear model. Identify the link function and linear predictor.

(b) Obtain the likelihood ratio test statistic $-2 \log \Lambda$ for testing H_0, and describe its distribution under the hypothesis, and its use.

(c) Arrange the data in a $2 \times m$ table of of counts of successes and failures. Show how to to test H_0 using the well known Pearson χ^2 test statistic.

(d) Using the notation $\hat{\pi}_k = \bar{y}_k$ and $\hat{\pi} = \bar{y}$ show that the two test statistics $-2 \log \Lambda$ and χ^2 are practically the same when $\hat{\pi}_i$ is close to $\hat{\pi}$.

Hint: Use a Taylor series expansion of $x \log(c/x)$.

6. Consider the linear model $\boldsymbol{Y} = \alpha\mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ with Var-Cov$[\boldsymbol{Y}] = \boldsymbol{V}$, where $\boldsymbol{V} = \sigma^2[(1-\rho)\boldsymbol{I} + \rho\boldsymbol{J}]$, $\mathbf{1}$ is a vector of ones and $\boldsymbol{J}$ is a matrix with all entries equal to 1. Assume that $\mathbf{1}^T\boldsymbol{X} = 0$. Let $\boldsymbol{Q}$ be an orthogonal matrix such that the first component of $\boldsymbol{Z} = \boldsymbol{Q}\boldsymbol{Y} = [Z_0, \boldsymbol{Z}^T]^T$ satisfies $Z_0 = n^{-1/2}\sum Y_i$.

(a) Show that $E[\boldsymbol{Z}] = \boldsymbol{U}\boldsymbol{\beta}$ and that the components of $\boldsymbol{Z}$ are uncorrelated with common variance $\sigma^2(1-\rho)$. The matrix $\boldsymbol{U}$ depends on $\boldsymbol{X}$ and $\boldsymbol{Q}$.

(b) Suppose σ^2 and ρ are both unknown. Prove that the minimum variance linear unbiased estimators of any estimable functions of $\boldsymbol{\beta}$ are obtained from the least squares solutions for the model $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ and Var-Cov$[\boldsymbol{\varepsilon}] = \sigma^2(1-\rho)\boldsymbol{I}$.

Hint: $(\boldsymbol{I} + a\boldsymbol{J})^{-1} = \boldsymbol{I} - [a/(1+na)]\boldsymbol{J}$.

7. Data on calcium uptake in plants was collected in an experiment in which plant cells were suspended in a solution of radioactive calcium for a certain length of time and then the amount of radioactive calcium that was absorbed by the cells was measured. The experiment was repeated independently with 9 different times of suspension, each replicated 3 times. The calcium uptake was modeled as a regression of calcium uptake Y_{ij} against time t_i, $i = 1, \ldots, 9$, $j = 1, 2, 3$. The regression output appears below, and plots of the raw data and some diagnostic plots appear on the following page.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.26407 -0.38755 -0 05378  0 29999  1 05142

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0 25326    0 25520   0.992 0 330901
time         0.61195    0.08826   6 934  3.6e-07 ***
I(time^2)   -0.02437    0 00565  -4.313 0 000238 ***
---
Signif. codes:  0 *** 0 001 ** 0 01 * 0.05   0 1   1

Residual standard error: 0.5577 on 24 degrees of freedom
Multiple R-squared: 0.8598,     Adjusted R-squared: 0.8481
F-statistic: 73.58 on 2 and 24 DF,  p-value: 5 78e-11
```

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
AUGUST 2010

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. Let $Y_{ij} = \mu_i + e_{ij}$, $i = 1, \ldots, 4$, $j = 1, \ldots, n$. Assume that the e_{ij} are i.i.d. $N(0, \sigma^2)$

(a) Write out the ANOVA table, including the sums of squares, mean squares, degrees of freedom and expected mean squares.

(b) Find the test statistic for testing H_0: $\mu_1 = \mu_2$ and $\mu_3 = \mu_4$. What is its distribution under H_0?

(c) What is the distribution the test statistic of part (b) under the alternative $\mu_2 - \mu_1 = \sigma = \mu_4 - \mu_3$?

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
JANUARY 2010

Applied Statistics (M.A. Version)

Instructions to the Student

a Answer all 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 a three way ANOVA model

$$Y_{ijk} = \mu + \alpha_i^A + \alpha_j^B + a_k^C + e_{ijk}$$

where α_i^A, $i = 1, \ldots, I$, and α_j^B and $j = 1, \ldots, J$, are fixed effect parameters for factors A and B and where a_k^C, $k = 1, \ldots, K$, are random effects of factor C. Assume that the $\{a_k^C\}$ and $\{e_{ijk}\}$ are a collection of mutually independent random variables, that $a_k^C \sim N(0, \sigma_C^2)$ and that $e_{ijk} \sim N(0, \sigma_e^2)$

(a) Write out the ANOVA table for this model, including expressions for sums of squares, degrees of freedom and expected mean squares.

(b) Find a $1 - \alpha$ confidence interval for $\theta = \sigma_C^2/\sigma_e^2$.

(c) Find a $1-\alpha$ confidence interval for $\alpha_1^A - \alpha_2^A$. Assume that this parameter was of interest in advance of observing the data.

2. Consider the regression model

$$Y_{ij} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_{12} x_{i1} x_{i2} + e_{ij}$$

where observations of Y_{ij}, $j = 1, \ldots, n$ are made for each of the following combinations of (x_{i1}, x_{i2}, x_{i3}): $(1,1,1)$, $(1,-1,-1)$, $(-1,1,-1)$, $(-1,-1,1)$.

(a) Assume $n = 1$ and $\beta_{12} = 0$. Write the model in matrix form. Which parameters, if any, are estimable?

(b) Under the assumptions of (a), compute a set of least squares estimates. How would your answers change if $n > 1$?

(c) If $n > 1$ and we make no assumptions about β_{12}, find a system of least squares estimates. Which parameters are estimable in this case?

(d) Under the assumptions of (c), how would you estimate $\sigma^2 = \text{Var}(e_{ij})$? What is the distribution of your estimator?

3. Consider the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$$

where the e_i, $i = 1, \ldots, n$, are i.i.d. $N(0, \sigma^2)$ random variables. Let $\boldsymbol{\beta} = (\beta_1^T, \beta_2^T)^T$, and let $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ denote the vector of least squares estimates. Assume that $\mathbf{Y}$ is n-dimensional, that $\mathbf{X}_1$ has q columns and $\mathbf{X}_2$ has $p - q$ columns, and that $\mathbf{X}$, $\mathbf{X}_1$ and $\mathbf{X}_2$ have full rank.

(a) Suppose that one fits the full model to the data when in fact the reduced model $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{e}$ is correct. Show that the bias of $\hat{\boldsymbol{\beta}}_1$ is zero.

(b) Compute the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$.

(c) Is the statistic

$$s^2 = MSE = \frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{n - p}$$

an unbiased estimator of σ^2? What is its variance?

(d) Suppose $p = 2$, and $q = 1$. Find $\text{Var}[\hat{\beta}_1]$. How does it compare to the variance that would have been obtained if the correct reduced model had been fitted to the data?

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
JANUARY 2010

Applied Statistics (Ph.D. Version)

Instructions to the Student

a. Answer all 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 a three way ANOVA model

$$Y_{ijk} = \mu + \alpha_i^A + \alpha_j^B + a_k^C + e_{ijk}$$

where α_i^A, $i = 1, \ldots, I$, and α_j^B and $j = 1, \ldots, J$, are fixed effect parameters for factors A and B and where a_k^C, $k = 1, \ldots, K$, are random effects of factor C. Assume that the $\{a_k^C\}$ and $\{e_{ijk}\}$ are a collection of mutually independent random variables, that $a_k^C \sim N(0, \sigma_C^2)$ and that $e_{ijk} \sim N(0, \sigma_e^2)$.

(a) Write out the ANOVA table for this model, including expressions for sums of squares, degrees of freedom and expected mean squares

(b) Find a $1 - \alpha$ confidence interval for $\theta = \sigma_C^2/\sigma_e^2$.

(c) Find a $1-\alpha$ confidence interval for $\alpha_1^A - \alpha_2^A$. Assume that this parameter was of interest in advance of observing the data.

2. Consider the regression model

$$Y_{ij} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_{12} x_{i1} x_{i2} + e_{ij}$$

where observations of Y_{ij}, $j = 1, \ldots, n$ are made for each of the following combinations of (x_{i1}, x_{i2}, x_{i3}): $(1,1,1)$, $(1,-1,-1)$, $(-1,1,-1)$, $(-1,-1,1)$.

(a) Assume $n = 1$ and $\beta_{12} = 0$. Write the model in matrix form. Which parameters, if any, are estimable?

(b) Under the assumptions of (a), compute a set of least squares estimates. How would your answers change if $n > 1$?

(c) If $n > 1$ and we make no assumptions about β_{12}, find a system of least squares estimates. Which parameters are estimable in this case?

(d) Under the assumptions of (c), how would you estimate $\sigma^2 = \text{Var}(e_{ij})$? What is the distribution of your estimator?

3. Consider the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$$

where the e_i, $i = 1, \ldots, n$, are i.i.d. $N(0, \sigma^2)$ random variables. Let $\boldsymbol{\beta} = (\beta_1^T, \beta_2^T)^T$, and let $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ denote the vector of least squares estimates. Assume that $\mathbf{Y}$ is n-dimensional, that $\mathbf{X}_1$ has q columns and $\mathbf{X}_2$ has $p - q$ columns, and that $\mathbf{X}$, $\mathbf{X}_1$ and $\mathbf{X}_2$ have full rank.

(a) Suppose that one fits the full model to the data when in fact the reduced model $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{e}$ is correct. Show that the bias of $\hat{\boldsymbol{\beta}}_1$ is zero.

(b) Compute the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$.

(c) Is the statistic

$$s^2 = MSE = \frac{\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2}{n - p}$$

an unbiased estimator of σ^2? What is its variance?

(d) Suppose $p = 2$, and $q = 1$. Find $\text{Var}[\hat{\beta}_1]$. How does it compare to the variance that would have been obtained if the correct reduced model had been fitted to the data?

4. A population $\mathcal{U}$ contains $N = nk$ elements, arranged in a list. A systematic sample $\mathcal{S}$ of size n is selected from $\mathcal{U}$ in order to estimate the population mean $\bar{y}_{\mathcal{U}} = N^{-1} \sum_{\mathcal{U}} y_i$.

(a) Show that the sample average $\bar{y}_{\mathcal{S}} = n^{-1} \sum_{\mathcal{S}} y_i$ is an unbiased estimator of $\bar{y}_{\mathcal{U}}$ and find its variance.

(b) Can Var $[\bar{y}_{\mathcal{S}}]$ be estimated from the sample? Justify your answer.

(c) Compare Var $[\bar{y}_{\mathcal{S}}]$ to the variance that might have been obtained if a simple random sample had been used to estimate $\bar{y}_{\mathcal{U}}$. When does systematic sampling yield a more efficient estimator of $\bar{y}_{\mathcal{U}}$?

Note: The systematic sample $\mathcal{S}$ consists of elements $J, J+k, J+2k, \ldots, J+(n-1)k$, where J is drawn at random from the set $\{1, 2, \ldots, k\}$.

5. Consider the unbalanced two way ANOVA model

$$Y_{ijk} = \mu_{ij} + e_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

where $i = 1, 2$, $j = 1, 2$, $k = 1, \ldots, n_ij$. Assume that the error terms are i.i.d. $N(0, \sigma^2)$.

(a) Find the best linear unbiased estimates of the μ_{ij}.

(b) Find the best linear unbiased estimates of the μ_{ij} under the null hypothesis H_0: $\beta_j \equiv 0$ and $\gamma_{ij} \equiv 0$.

(c) Give the formula of the statistic for testing H_0 against the general alternative and its distribution under H_0.

(d) What is the distribution of the statistic in (c) under the alternative? Express your answers in terms of the cell means and σ^2.

6. A three stage cluster sample is taken from a population with N primary sampling units (psu's), M_i secondary sampling units (ssu's) in psu i, and L_{ij} elements in ssu j of psu i. First a sample $\mathcal{S}$ of n psu's is selected. Next samples $\mathcal{S}_i$, $i \in \mathcal{S}$, containing m_i ssu's are drawn from the sampled psu's. At each stage, simple random sampling is used and the second stage samples are drawn independently of one another.

(a) Show that the sample weights are

$$w_{ij} = 1/\pi_{ij} = \frac{N}{n}\frac{M_i}{m_i}$$

where $\pi_{ij} = P[\text{ssu } (i,j) \text{ is sampled}]$

(b) Let

$$\hat{t} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}.$$

Show that $\hat{t}$ is an unbiased estimator of the population total t of y.

(c) Find the variance of $\hat{t}$.

4. A population $\mathcal{U}$ contains $N = nk$ elements, arranged in a list. A systematic sample $\mathcal{S}$ of size n is selected from $\mathcal{U}$ in order to estimate the population mean $\bar{y}_{\mathcal{U}} = N^{-1} \sum_{\mathcal{U}} y_i$.

(a) Show that the sample average $\bar{y}_{\mathcal{S}} = n^{-1} \sum_{\mathcal{S}} y_i$ is an unbiased estimator of $\bar{y}_{\mathcal{U}}$ and find its variance.

(b) Can Var $[\bar{y}_{\mathcal{S}}]$ be estimated from the sample? Justify your answer.

(c) Compare Var $[\bar{y}_{\mathcal{S}}]$ to the variance that might have been obtained if a simple random sample had been used to estimate $\bar{y}_{\mathcal{U}}$. Use ANOVA to decide when systematic sampling yields a more efficient estimator of $\bar{y}_{\mathcal{U}}$.

Note: The systematic sample $\mathcal{S}$ consists of elements $J, J+k, J+2k, \ldots, J+(n-1)k$, where J is drawn at random from the set $\{1, 2, \ldots, k\}$.

5. Consider the unbalanced two way ANOVA model

$$Y_{ijk} = \mu_{ij} + e_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

where $i = 1, 2$, $j = 1, 2$, $k = 1, \ldots, n_ij$. Assume that the error terms are i.i.d. $N(0, \sigma^2)$.

(a) Find the best linear unbiased estimates of the μ_{ij}.

(b) Find the best linear unbiased estimates of the μ_{ij} under the null hypothesis H_0: $\beta_j \equiv 0$ and $\gamma_{ij} \equiv 0$.

(c) Give the formula of the statistic for testing H_0 against the general alternative and its distribution under H_0

6. A three stage cluster sample is taken from a population with N primary sampling units (psu's), M_i secondary sampling units (ssu's) in psu i, and L_{ij} elements in ssu j of psu i. First a sample $\mathcal{S}$ of n psu's is selected. Next samples $\mathcal{S}_i$, $i \in \mathcal{S}$, containing m_i ssu's are drawn from the sampled psu's. At each stage, simple random sampling is used and the second stage samples are drawn independently of one another

(a) Show that the sample weights are

$$w_{ij} = 1/\pi_{ij} = \frac{N}{n}\frac{M_i}{m_i}$$

where $\pi_{ij} = P[\text{ssu } (i,j) \text{ is sampled}]$.

(b) Let

$$\hat{t} = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}_i} w_{ij} y_{ij}.$$

Show that $\hat{t}$ is an unbiased estimator of the population total t of y.

(c) Find the variance of $\hat{t}$.

Hint: Use properties of conditional probability and conditional expectation.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
AUGUST 2009

Applied Statistics (Ph.D. Version)

Instructions to the Student

a. Answer all 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 regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where Var-Cov$[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}$ and $\mathbf{X}$ is an $n \times p$ matrix of full rank. Let $\widehat{\boldsymbol{\beta}}$ be the ordinary least squares estimator of $\boldsymbol{\beta}$.

(a) Find the mean and covariance matrix of $\widehat{\boldsymbol{\beta}}$.

(b) Let $\hat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ be the vector of predicted values and let $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}}$. Find the variance-covariance matrix of $\hat{\mathbf{e}}$.

(c) Suppose that the model for $E[\mathbf{Y}]$ is correct but Var-Cov$[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{D}$ where $\mathbf{D}$ is a known full rank diagonal matrix. If this misspecified model is analyzed by ordinary least squares, what is the covariance matrix of $\hat{\mathbf{e}}$? Is $\widehat{\boldsymbol{\beta}}$ unbiased?

(d) What would be the BLUE of $\boldsymbol{\beta}$ under the assumptions of (c)? What is the covariance matrix of the BLUE?

2. Let $Y_{ij} = \mu + a_i + \beta x_{ij} + e_{ij}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, denote a mixed effect ANCOVA model where μ and β are fixed parameters, the a_i are i.i.d. $N(0, \sigma_a^2)$ and the e_{ij} are i.i.d. $N(0, \sigma^2)$. The a_i and e_{ij} are mutually independent. The x_{ij} satisfy $\bar{x}_i = 0$ for each i.

(a) Find the variance-covariance matrix of the vector of Y_{ij} values.

(b) Find the ordinary least squares estimators of the fixed effect parameters.

(c) Find the covariance matrix of the estimators in (b).

(d) Prove that $\hat{\mu} = \bar{Y}_{\cdot\cdot}$ is an unbiased estimator of μ and has smallest variance in the class of all unbiased estimators of the form $\sum_i \sum_j c_{ij} Y_{ij}$. (*Hint*: Use an ANOVA decomposition of the array $\{c_{ij}\}$ together with conditions for unbiasedness.)

3. A simple random sample of size n is chosen without replacement from a population U and a variable y is measured on each sample element. A "rough" auxiliary variable $x = y + c + e$ is known for all units in the population. The quantity c is a constant bias term, and e is a random measurement error, uncorrelated with y and having mean zero and variance S_e^2. Two estimates of $\bar{y}_U$, the population mean of y, are

$$\begin{aligned} \hat{\bar{y}}_d &= \bar{y} + (\bar{x}_U - \bar{x}) \quad \text{(difference estimator)} \\ \hat{\bar{y}}_{reg} &= \bar{y} + B(\bar{x}_U - \bar{x}) \quad \text{(regression estimator)} \end{aligned}$$

where $\bar{y}$ is the sample average of y and $\bar{x}$ is defined similarly. Compare the variances of each of these estimators, using the value of B which minimizes the variance of the regression estimator. You may assume the population is infinite. The variances may involve S_y^2, the population variance of y.

4. Let $Y_{ijk} = \mu + \alpha_i + b_j + c_{ij} + \varepsilon_{ijk}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$. The quantities $\mu, \alpha_1, \ldots, \alpha_I$ are fixed unknown parameters and $\sum_{i=1}^{I} \alpha_i = 0$. Assume that the $\{b_j\}$, $\{c_{ij}\}$, $\{\varepsilon_{ijk}\}$ are independent normal variables with zero means, $\text{Var}\, b_j = \sigma_b^2$, $\text{Var}\, c_{ij} = \sigma_c^2$ and $\text{Var}\, \varepsilon_{ijk} = \sigma_e^2$.

(a) Write out the ANOVA table for this mixed model, giving the formulas for each sum of squares and the corresponding degrees of freedom. In addition, calculate the expected mean squares.

(b) Find a confidence interval for σ_c^2/σ_e^2.

(c) How would you test H_0: $\alpha_1 = \cdots = \alpha_I$? Give the test statistic and its distribution under H_0. What is the power of your test?

5. A finite population U consists of N elements denoted by $1, 2, \ldots, N$. A sample s of n elements is selected without replacement, where n is fixed. Let $z_i = I\{i \in s\}$ and write $P[i \in s] = \pi_i$. Let π_{ij} denote the probability that elements i and j are both included in the sample.

(a) Show that $\sum_{i \in s} \pi_i = n$ and calculate $\text{Cov}(z_i, z_j)$.

(b) Show that the statistic

$$\hat{t}_\pi = \sum_{i \in s} y_i/\pi_i$$

is an unbiased estimator of the population total $t_y = \sum_{i \in U} y_i$.

(c) Calculate the variance of $\hat{t}_\pi$.

6. In a two way incomplete layout the observed data are as follows:

$$\begin{bmatrix} Y_{11} & Y_{12} & - \\ - & Y_{22} & Y_{23} \\ Y_{31} & - & Y_{33} \end{bmatrix}$$

The model is $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$.

(a) Show that all contrasts $\sum_i c_i \alpha_i$ are estimable, where as usual, $\sum_i c_i = 0$.

(b) Is it possible to find an unbiased estimator of $\text{Var}\, Y_{ij}$? How many degrees of freedom are associated with this estimator, if it exists?

(c) Assuming the unobserved Y_{ij} follow the model, is $E[Y_{13}]$ estimable?

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
AUGUST 2009

Applied Statistics (M.A. Version)

Instructions to the Student

a. Answer all 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 regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where Var-Cov$[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}$ and $\mathbf{X}$ is an $n \times p$ matrix of full rank. Let $\widehat{\boldsymbol{\beta}}$ be the ordinary least squares estimator of $\boldsymbol{\beta}$.

(a) Find the mean and covariance matrix of $\widehat{\boldsymbol{\beta}}$.

(b) Let $\hat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ be the vector of predicted values and let $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}}$. Find the variance-covariance matrix of $\hat{\mathbf{e}}$.

(c) Suppose that the model for $E[\mathbf{Y}]$ is correct but Var-Cov$[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{D}$ where $\mathbf{D}$ is a known full rank diagonal matrix. If this misspecified model is analyzed by ordinary least squares, what is the covariance matrix of $\hat{\mathbf{e}}$?

(d) What would be the BLUE of $\boldsymbol{\beta}$ under the assumptions of (c)?

2. Let $Y_{ij} = \mu + a_i + e_{ij}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, denote a random effect one way ANOVA model where μ is a fixed parameter, the a_i are i.i.d $N(0, \sigma_a^2)$ and the e_{ij} are i.i.d. $N(0, \sigma^2)$. The a_i and e_{ij} are mutually independent.

(a) Write out the ANOVA table for this problem, including the sums of squares, degrees of freedom and expected mean squares.

(b) Find an unbiased estimator of σ_a^2 and a confidence interval for σ_a/σ.

(c) Prove that $\hat{\mu} = \bar{Y}$ is an unbiased estimator of μ and has smallest variance in the class of all unbiased estimators of the form $\sum_i \sum_j c_{ij} Y_{ij}$ (*Hint*: Use an ANOVA decomposition of the array $\{c_{ij}\}$ together with conditions for unbiasedness.)

3 A simple random sample of size n is chosen without replacement from a population U and a variable y is measured on each sample element. A "rough" auxiliary variable $x = y + c + e$ is known for all units in the population. The quantity c is a constant bias term, and e is a random measurement error, uncorrelated with y and having mean zero and variance S_e^2. Two estimates of $\bar{y}_U$, the population mean of y, are

$$\begin{aligned} \hat{\bar{y}}_d &= \bar{y} + (\bar{x}_U - \bar{x}) \quad \text{(difference estimator)} \\ \hat{\bar{y}}_{reg} &= \bar{y} + B(\bar{x}_U - \bar{x}) \quad \text{(regression estimator)} \end{aligned}$$

where $\bar{y}$ is the sample average of y and $\bar{x}$ is defined similarly. Compare the variances of each of these estimators, using the value of B which minimizes the variance of the regression estimator. You may assume the population is infinite. The variances may involve S_y^2, the population variance of y.

4. Let $Y_{ijk} = \mu + \alpha_i + b_j + c_{ij} + \varepsilon_{ijk}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$. The quantities $\mu, \alpha_1, \ldots, \alpha_I$ are fixed unknown parameters. Assume that the $\{b_j\}$, $\{c_{ij}\}$, $\{\varepsilon_{ijk}\}$ are independent normal variables with zero means, $\text{Var}\, b_j = \sigma_b^2$, $\text{Var}\, c_{ij} = \sigma_c^2$ and $\text{Var}\, \varepsilon_{ijk} = \sigma_e^2$.

(a) Write out the ANOVA table for this problem, giving the formulas for each sum of squares and the corresponding degrees of freedom. In addition, calculate the expected mean squares.

(b) Find a confidence interval for σ_c^2/σ_e^2.

(c) How would you test H_0: $\alpha_1 = \cdots = \alpha_I$? Give the test statistic and its distribution under H_0.

5. A finite population U consists of N elements denoted by $1, 2, \ldots, N$. A sample s of n elements is selected without replacement, where n is fixed. Let $z_i = I\{i \in s\}$ and write $P[i \in s] = \pi_i$. Let π_{ij} denote the probability that elements i and j are both included in the sample.

(a) Show that $\sum_{i \in s} \pi_i = n$ and calculate $\text{Cov}(z_i, z_j)$.

(b) Show that the statistic

$$\hat{t}_\pi = \sum_{i \in s} y_i/\pi_i$$

is an unbiased estimator of the population total $t_y = \sum_{i \in U} y_i$.

(c) Calculate the variance of $\hat{t}_\pi$.

6. In a two way incomplete layout the observed data are as follows:

$$\begin{bmatrix} Y_{11} & Y_{12} & - \\ - & Y_{22} & Y_{23} \\ Y_{31} & - & Y_{33} \end{bmatrix}$$

The model is $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$.

(a) Show that all contrasts $\sum_i c_i \alpha_i$ are estimable, where as usual, $\sum_i c_i = 0$.

(b) Is it possible to find an unbiased estimator of $\text{Var}\, Y_{ij}$? How many degrees of freedom are associated with this estimator, if it exists?

(c) Assuming the unobserved Y_{ij} follow the model, is $E[Y_{13}]$ estimable?

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
JANUARY 2009

Applied Statistics (Ph.D. Version)

Instructions to the Student

a. Answer all 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 regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where Var-Cov$[\boldsymbol{\varepsilon}] = \mathbf{V}$ is an arbitrary positive definite matrix of dimension $n \times n$ and $\mathbf{X}$ is an $n \times p$ matrix of full rank. Let $\widehat{\boldsymbol{\beta}}$ be the ordinary least squares estimator of $\boldsymbol{\beta}$.

(a) Find the mean and covariance matrix of $\widehat{\boldsymbol{\beta}}$.

(b) If $\mathbf{V}$ is known, prove that the BLUE of $\boldsymbol{\beta}$ is the *generalized least squares* estimator

$$\widehat{\boldsymbol{\beta}}_G = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}.$$

What is the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}_G$?

(c) The ordinary least squares estimator and the generalized least squares estimator are equal if and only if there exists a nonsingular matrix $\mathbf{M}$ such that $\mathbf{VX} = \mathbf{XB}$.

2. Data (x_{i1}, x_{i2}, Y_i), $i = 1, \ldots, n$, are governed by the centered bivariate regression model $Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$. The sample averages of the x_{i1} and x_{i2} are both zero. Consider the following stepwise regression algorithm:

(i) Estimate the model $Y_i - \bar{Y} = \beta_1' x_{i1} + \varepsilon_i'$ using ordinary least squares. Denote the estimated coefficient by b_1'. Compute the residuals e_{iy1}.

(ii) Estimate the model $x_{i2} = \gamma x_{i1} + d_i$ using ordinary least squares and compute the residuals e_{i21}. Denote the estimated coefficient by c.

(iii) Estimate the model $e_{iy1} = \beta_2' e_{i21} + \eta_i$. Denote the estimated slope by b_2.

Prove that b_2 is the least squares estimate of β_2 and that $\bar{Y} + b_1' x_{i1} + b_2 e_{i21}$ is the least squares estimate of $\alpha + \beta_1 x_{i1} + \beta_2 x_{i2}$. [*Hint*: Consider the regression of Y on x_1 and e_{21}.]

3. A population U of size N is divided into two strata of sizes $N_1 = NW_1$ and $N_2 = NW_2$. The population mean of the variable y is the weighted sum of stratum means

$$\bar{y}_U = (1/N) \sum_h \sum_{i \in U_h} y_{hi} = W_1 \bar{y}_1 + W_2 \bar{y}_2,$$

and the within-stratum variances S_h^2, $h = 1, 2$, are

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i \in U_h} (y_{hi} - \bar{y}_h)^2.$$

It costs c_h to sample a single element from stratum h. It is known that $S_1^2 = S_2^2$, but the sampling costs are such that $2c_1 \leq c_2 \leq 4c_2$. The statistician would like to use proportional allocation but does not want to incur a substantial increase in variance, compared to optimal allocation. Let $\hat{\bar{y}}_{st}$ denote the estimate of $\bar{y}_U$ under stratified sampling. Let $V_{prop}(\hat{\bar{y}}_{st})$ denote the variance of $\hat{\bar{y}}_{st}$ under proportional allocation, and let $V_{opt}(\hat{\bar{y}}_{st})$ denote the variance under optimal (Neyman) allocation. For a given cost $C = c_1 n_1 + c_2 n_2$, ignoring the finite population correction, show that

$$\frac{V_{prop}(\hat{\bar{y}}_{st})}{V_{opt}(\hat{\bar{y}}_{st})} = \frac{W_1 c_1 + W_2 c_2}{(W_1 \sqrt{c_1} + W_2 \sqrt{c_2})^2}.$$

If $W_1 = W_2$, compute the relative increases in variance from using proportional allocation when $c_2/c_1 = 2, 4$.

4 Let $Y_{ijk} = \mu + a_i + b_{ij} + \varepsilon_{ijk}$ Assume that the $\{a_i\}$, $\{b_{ij}\}$, $\{\varepsilon_{ijk}\}$ are mutually independent normal variables with zero means, Var $a_i = \sigma_a^2$, Var $b_{ij} = \sigma_b^2$ and Var $\varepsilon_{ijk} = \sigma_e^2$.

(a) Write out the ANOVA table for this problem, giving the formulas for each sum of squares and the corresponding degrees of freedom In addition, calculate the expected mean squares.

(b) Find a confidence interval for σ_b^2/σ_e^2

(c) How would you test H_0: $\sigma_a^2 = 0$? Give the test statistic and its distribution under H_0

5. A finite population U consists of N elements $1, 2, \ldots, N$. A sample s of n elements is selected without replacement, where n is fixed Let $z_i = I\{i \in s\}$ and write $P[i \in s] = \pi_i$. Let π_{ij} denote the probability that elements i and j are both included in the sample

(a) Show that $\sum_{i \in s} \pi_i = n$ and calculate $\mathrm{Cov}(z_i, z_j)$

(b) Show that the statistic $\hat{t}_\pi = \sum_{i \in s} y_i$ is an unbiased estimator of the population total $t_y = \sum_{i \in U} y_i$.

(c) Calculate the variance of $\hat{t}_\pi$.

6. In a two way incomplete layout the observed data are as follows:

$$\begin{bmatrix} Y_{11} & Y_{12} & - \\ - & Y_{22} & Y_{23} \\ Y_{31} & - & Y_{33} \end{bmatrix}$$

The model is $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$, where the ε_{ij} are independent with zero means and common variance σ^2.

(a) Show that all contrasts $\sum_i c_i \alpha_i$ are estimable, where as usual, $\sum_i c_i = 0$.

(b) Is it possible to find an unbiased estimator of Var Y_{ij}? How many degrees of freedom are associated with this estimator, if it exists?

(c) Assuming the unobserved Y_{ij} follow the model, is $E[Y_{13}]$ estimable?

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
AUGUST 2008

Applied Statistics (Ph.D. Version)

Instructions to the Student

a. Answer all 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 nested random effects model $Y_{ijk} = \mu + a_i + b_{ij} + e_{ijk}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$ and $k = 1, \ldots, K$. Assume the a_i, b_{ij} and e_{ijk} are mutually independent, the a_i are i.i.d. $N(0, \sigma_a^2)$, the b_{ij} are i.i.d. $N(0, \sigma_b^2)$ and the e_{ijk} are i.i.d. $N(0, \sigma_e^2)$.

(a) Write out the ANOVA table for this problem and calculate each of the expected mean squares.

(b) Find unbiased point estimators of each of the variance components. Are any of these estimators unsatisfactory for some reason? Explain.

(c) Test H_0: $\sigma_b^2 = 0$. Find the distribution of your test statistic under both the null and alternative hypotheses.

2. Consider the linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{X}$ is an $n \times p$ design matrix of rank $r < p < n$. Assume the side conditions $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{0}$, where $\boldsymbol{H}$ is a $(p - r) \times p$ matrix of full rank $p - r$, such that $\boldsymbol{H}\boldsymbol{\beta}$ defines a set of nonestimable functions.

(a) Show there is a unique solution of the normal equations which satisfies the side conditions $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{0}$

(b) Explain the rationale for the nonestimablity requirement.

(c) Consider the special case $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, $i = 1, 2$, $j = 1, 2$. Show that $\tau_1 + \tau_2$ is nonestimable, and use this fact to get a unique solution of the normal equations satisfying a suitable side condition, as in (a).

3. Let D be a domain of size N_D contained within a finite population $\mathcal{U}$ of size N. A simple random sample $\mathcal{S}$ of size n is selected and the value of some variable y is observed. The sample contains n_D elements of D. The sample domain mean is $\hat{\bar{y}}_D = n_D^{-1} \sum_{i \in \mathcal{U}_h \cap D} y_i$ is used to estimate the true domain mean $N_D^{-1} \bar{y}_D = \sum_{i \in D} y_i$.

(a) Given that n_D is positive, show that $\hat{\bar{y}}_D$ is conditionally unbiased.

(b) In terms of

$$S_D^2 = \frac{1}{N_D - 1} \sum_{i \in D} (y_i - \bar{y}_D)^2,$$

what is the conditional variance of $\hat{\bar{y}}_D$, given n_D?

(c) If $P[n_D = 0]$ is negligible, show that

$$E\left[\frac{1}{n_D}\right] = \frac{1}{nW} + \frac{1 - W}{n^2 W^2} + o(n^{-2})$$

where $W = N_D/N$.

(d) Assuming $P[n_D = 0]$ is negligible, what is the approximate variance of $\hat{\bar{y}}_D$?

4. A stratified sample is being designed to estimate p, the prevalence of a disease (i.e., the proportion of persons with the disease). Stratum 1, with N_1 persons, has prevalence p_1 and Stratum 2, with N_2 persons, has prevalence p_2, where $p_1 > p_2$. Assume that the cost of sampling and ascertaining disease status is the same for all persons in either stratum, that at most 2000 persons are to be sampled, and that N_1 and N_2 are both very large.

(a) If $p_1 = 0.10$, $p_2 = .03$ and $N_1/N = 0.4$, what are n_1 and n_2 under optimum allocation?

(b) Under the assumptions of (a), what is $\mathrm{Var}\,\hat{p}_{\mathrm{str}}$ under proportional allocation? Under optimal allocation? What is the variance if one takes a simple random sample of size 2000 from the population?

5. Consider the logistic regression model for binary (0-1) data Y_i with a single covariate x_i where $P[Y_i = 1 \mid x_i] = \pi_i$ satisfies

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathrm{logit}(\pi_i) = \alpha + \beta x_i, \quad i = 1, \ldots, n$$

(a) Derive equations for the maximum likelihood estimates of α, β, and identify the asymptotic distribution of the MLE $(\hat{\alpha}, \hat{\beta})$.

(b) Show that the asymptotic covariance matrix of $(\hat{\alpha}, \hat{\beta})$ has the form $(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}$ for some matrices $\boldsymbol{X}, \boldsymbol{V}$. Describe the matrices $\boldsymbol{X}, \boldsymbol{V}$.

(c) Obtain a 95% confidence interval for β.

6. A sample of 12 observations of (x_1, x_2, Y) is displayed in the figure, with points (x_2, Y) labeled by their x_1 values. The least squares regression line $y = \hat{\beta}_0 + \hat{\beta}_2 x_2$ is also displayed in the figure.

(a) Suppose that x_1 a quantitative variable. How would you decide if the simple linear regression model

$$Y_i = \beta_0 + \beta_2 x_{2i} + e_i, \quad i = 1, \ldots, 12,$$

describes the data adequately? Assume that the error terms are i.i.d. $N(0, \sigma^2)$. Explain how to compute any test statistics and give their distributions

(b) Suppose instead that the variable x_1 is a qualitative variable labeling groups, which were chosen at random from some population. Propose a generalization of the simple linear regression model and describe how to test whether the generalized model fits the data better than the simple linear regresson model of (a)

(c) Based on your examination of the graph, under the assumptions of either (a) or (b), would the regression coefficient of x_2 be positive if x_1 were included in the model?

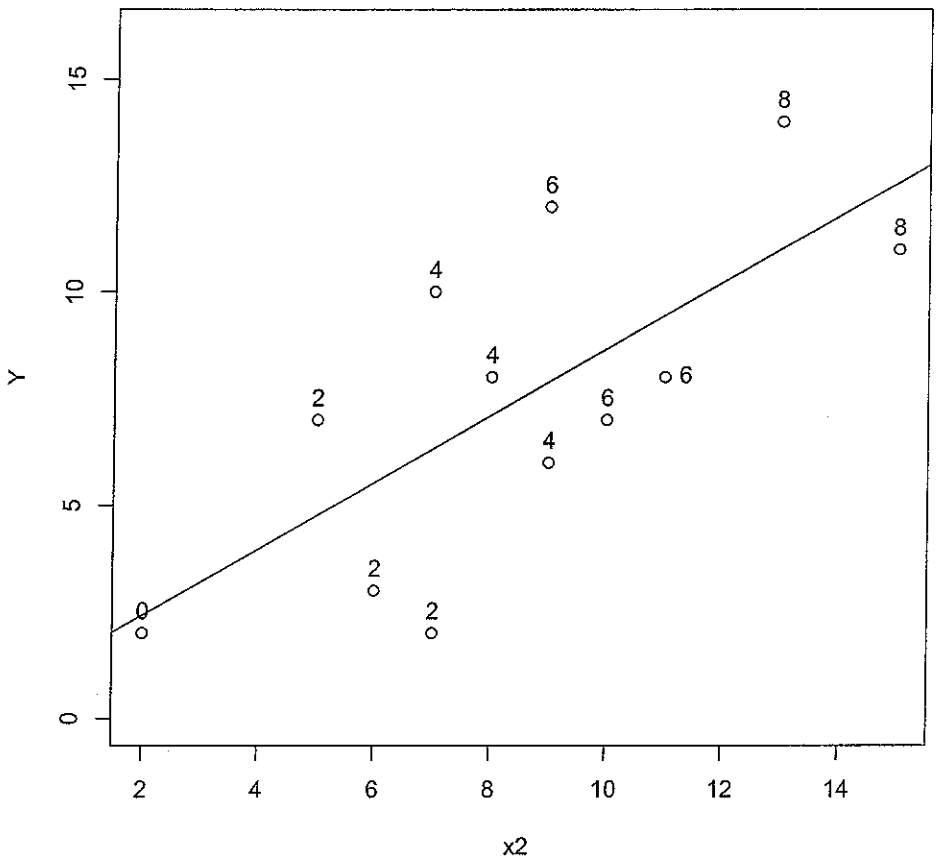


Figure 1: Plot of Y vs. x_2 with points labeled by x_1 values.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
JANUARY 2008

Applied Statistics (Ph.D. Version)

Instructions to the Student

a. Answer all 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 two way main effect ANOVA model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}, \quad i = 1, \ldots, I, \ j = 1, \ldots, J, \ k = 1, \ldots, K$$

where $\sum_i \alpha_i = \sum_j \beta_j = 0$ and the ε_{ijk} are i.i.d. $N(0, \sigma^2)$.

(a) Write out the ANOVA table for this model, showing sums of squares, degrees of freedom and expected mean squares.

(b) Find confidence limits for $\alpha_1 - \alpha_2$ and for σ^2. Assume that the contrast $\alpha_1 - \alpha_2$ was of interest in advance of the data collection.

(c) Suppose that one fits the two way main effect model when in fact there is an interaction term of the form γ_{ij} satisfying the usual side conditions $\sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$. Show that the usual estimator of σ^2 based on the main effects model is biased and describe the effect of this bias on point and interval estimation of $\alpha_1 - \alpha_2$.

2. Consider the sinusoidal model,

$$Y_t = \mu + \alpha\cos(\omega_p t) + \beta\sin(\omega_p t) + e_t, \quad t = 1, 2, \ldots, N,$$

where $\omega_p = 2\pi p/N$ for some $p \in \{1, 2, \ldots, N\}$, N is even, and the e_t are uncorrelated with mean 0 and variance σ^2. We know:

$$\sum_{t=1}^{N}\cos(\omega_p t) = \sum_{t=1}^{N}\sin(\omega_p t) = 0, \quad \sum_{t=1}^{N}\cos(\omega_p t)\sin(\omega_q t) = 0, \quad \text{for all } p,\ q,$$

$$\sum_{t=1}^{N}\cos(\omega_p t)\cos(\omega_q t) = \begin{cases} 0 & \text{if } p \neq q \\ N & \text{if } p = q = N/2,\ 0 \\ N/2 & \text{if } p = q \neq N/2, \end{cases}$$

$$\sum_{t=1}^{N}\sin(\omega_p t)\sin(\omega_q t) = \begin{cases} 0 & \text{if } p \neq q \\ 0 & \text{if } p = q = N/2 \\ N/2 & \text{if } p = q \neq N/2. \end{cases}$$

For simplicity assume that $p \neq N/2$.

(a) Suppose ω_p is known. Obtain explicit expressions for the least squares estimates of μ, α, β.

(b) Suppose ω_p is known. Express the squared multiple correlation R^2 in terms of $\hat{\alpha}^2$ and $\hat{\beta}^2$.

(c) Suppose ω_p is unknown. Suggest a way to estimate ω_p.

3. A simple random sample of size is to be drawn from a population $\mathcal{U}$. The goal is to estimate the population total $t_{y\mathcal{U}}$ subject to the accuracy requirement

$$P\left[\left|\frac{\hat{t}_{y\mathcal{U}} - t_{y\mathcal{U}}}{t_{y\mathcal{U}}}\right| \leq r\right] \geq 1 - \alpha$$

The population coefficient of variation $C = S_{y\mathcal{U}}/\bar{y}_{\mathcal{U}}$ is known. Find the required sample size n. You may assume that the estimator $\hat{t}_{y\mathcal{U}}$ has an approximate normal distribution.

4. Let $X_1, \ldots, X_k$ be i.i.d $N(\mu, \sigma^2)$ random variables. Define the sample range by $R = \max_i\{X_i\} - \min_i\{X_i\}$. Let s^2 be an independent (of $X_1, \ldots, X_k$) estimate of σ^2 with ν degrees of freedom. That is, $\nu s^2/\sigma^2 \sim \chi^2(\nu)$. The *Studentized range* is defined as $Q = R/s$.

(a) Prove that the distribution of Q does not depend on the unknown parameters μ and σ^2.

(b) Suppose that parameters $\theta_1, \ldots, \theta_k$ are estimated by $\hat{\theta}_1, \ldots, \hat{\theta}_k$. Assume the following:

(i) $\hat{\theta}_i \sim N(\theta_i, a^2\sigma^2)$, $i = 1, \ldots, k$, independent, and a is known.

(ii) s^2 is an estimate of σ^2 such that $\nu s^2/\sigma^2 \sim \chi^2(\nu)$ independently of the $\hat{\theta}_i$.

Use the studentized range to obtain a system of simultaneous level $1-\alpha$ confidence intervals for all the $\binom{k}{2}$ differences $\theta_i - \theta_j$.

5. Several military pilots are assigned to test a simulated control panel for a new type of aircraft. The pilots must complete a simulated control task, which depends on factors A and B, each with 2 predetermined levels. The levels are manipulated by the experimenter in such a way that each pilot is presented the combinations of levels in random order and performs the task once at each combination of levels. A response variable Y is measured on pilot k at levels (i, j) of the experimental factors, for $i = 1, 2$, $j = 1, 2$ and $k = 1, \ldots, n$.

(a) Write a linear model describing the observed responses Y_{ijk}. State the distribution of any random terms in your model and state any necessary side conditions to make your model identifiable.

(b) Propose an estimator of the mean difference in response between the two levels of Factor A and give 95% confidence limits for this parameter.

(c) Propose unbiased estimators of the variances of any random terms in your model. Can you give confidence limits for these estimators?

6 An urban population U consists of 4000 apartment buildings. A sample of 100 buildings is selected at random without replacement. In sampled building i, $i = 1, \ldots, 100$, the variable n_i, the number of apartments in building i, is determined. In addition, the variables x_{ij} and y_{ij} are obtained on each apartment (i, j), $j = 1, \ldots, n_i$, within sampled building i. These variables are defined as follows:

$$\begin{aligned} x_{ij} &= I\{\text{apartment } (i,j) \text{ is occupied}\}, \\ y_{ij} &= \text{number of residents in apartment } (i,j). \end{aligned}$$

(a) Find an unbiased estimator of t_{xU}, the total number of occupied apartments, and provide 95% confidence limits for this total.

(b) Find an alternative estimator which is likely to be more accurate than the unbiased estimator of (a) when the n_i vary greatly from building to building.

(c) Estimate B, the average number of residents per occupied apartment and find 95% confidence limits for this ratio. As in (b), assume that the n_i vary greatly from building to building.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
JANUARY 2008

Applied Statistics (M.A. Version)

Instructions to the Student

a. Answer all 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 two way main effect ANOVA model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}, \quad i = 1, \ldots, I, \ j = 1, \ldots, J, \ k = 1, \ldots, K$$

where $\sum_i \alpha_i^A = \sum_j \alpha_j^B = 0$ and the ε_{ijk} are i.i.d. $N(0, \sigma^2)$.

(a) Write out the ANOVA table for this model, showing sums of squares, degrees of freedom and expected mean squares.

(b) Find confidence limits for $\alpha_1 - \alpha_2$ and for σ^2. Assume the contrast $\alpha_1 - \alpha_2$ was of interest before the data were observed.

(c) Suppose that one fits the two way main effect model when in fact there is an interaction term of the form γ_{ij} satisfying the usual side conditions $\sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$. Show that the estimator of σ^2 is biased and describe the effect of this bias on point and interval estimation of $\alpha_1 - \alpha_2$.

2. Consider the sinusoidal model,

$$Y_t = \mu + \alpha\cos(\omega_p t) + \beta\sin(\omega_p t) + e_t, \quad t = 1, 2, \ldots, N$$

where $\omega_p = 2\pi p/N$ for some $p \in \{1, 2, \ldots, N\}$, N is even, and the e_t are uncorrelated with mean 0 and variance σ^2. We know:

$$\sum_{t=1}^{N}\cos(\omega_p t) = \sum_{t=1}^{N}\sin(\omega_p t) = 0, \quad \sum_{t=1}^{N}\cos(\omega_p t)\sin(\omega_q t) = 0, \quad \text{for all } p,\ q,$$

$$\sum_{t=1}^{N}\cos(\omega_p t)\cos(\omega_q t) = \begin{cases} 0 & \text{if } p \neq q \\ N & \text{if } p = q = N/2,\ 0 \\ N/2 & \text{if } p = q \neq N/2, \end{cases}$$

$$\sum_{t=1}^{N}\sin(\omega_p t)\sin(\omega_q t) = \begin{cases} 0 & \text{if } p \neq q \\ 0 & \text{if } p = q = N/2 \\ N/2 & \text{if } p = q \neq N/2 \end{cases}$$

For simplicity assume that $p \neq N/2$.

(a) Suppose ω_p is known. Obtain explicit expressions for the least squares estimates of μ, α, β.

(b) Suppose ω_p is known. Express the squared multiple correlation R^2 in terms of $\hat{\alpha}^2$ and $\hat{\beta}^2$.

(c) Suppose ω_p is unknown. Suggest a way to estimate ω_p.

3. A simple random sample of size n is to be drawn from a population U. The goal is to estimate the population total t_{yU} subject to the accuracy requirement

$$P\left[\left|\frac{\hat{t}_{yU} - t_{yU}}{t_{yU}}\right| \leq r\right] \geq 1 - \alpha$$

The population coefficient of variation $C = S_{yU}/\bar{y}_U$ is known. Find the required sample size n. You may assume that the estimator $\hat{t}_{yU}$ has an approximate normal distribution.

4. Let $X_1, \ldots, X_k$ be i.i.d $N(\mu, \sigma^2)$ random variables. Define the sample range by $R = \max_i\{X_i\} - \min_i\{X_i\}$. Let s^2 be an independent (of $X_1, \ldots, X_k$) estimate of σ^2 with ν degrees of freedom. That is, $\nu s^2/\sigma^2 \sim \chi^2(\nu)$. Then the *Studentized range* has a known distribution independent of the parameters μ, σ^2:

$$Q = \frac{R}{s} \sim q_{k,\nu}$$

Suppose that a set of parameters $\theta_1, \ldots, \theta_k$ are estimated by $\hat{\theta}_1, \ldots, \hat{\theta}_k$. Assume the following:

(i) $\hat{\theta}_i \sim N(\theta_i, a^2\sigma^2)$, $i = 1, \ldots, k$, independent, and a is known

(ii) s^2 is an estimate of σ^2 such that $\nu s^2/\sigma^2 \sim \chi^2(\nu)$ independently of the $\hat{\theta}_i$.

Use the Studentized range to obtain a system of simultaneous level $1 - \alpha$ confidence intervals for all the $\binom{k}{2}$ differences $\theta_i - \theta_j$.

5. Several military pilots are assigned to test a simulated control panel for a new type of aircraft. The pilots must complete a simulated control task, which depends on factors A and B, each with 2 predetermined levels. The levels are manipulated by the experimenter in such a way that each pilot is presented all four combinations of levels in random order and performs the task m times at each combination of levels. A response variable Y is measured on pilot k at replication r of levels (i, j) of the experimental factors, for $i = 1, 2$; $j = 1, 2$; $r = 1, \ldots, m$; and $k = 1, \ldots, n$.

(a) Write a linear model describing the observed responses Y_{ijk}. State the joint distribution of any random terms in your model and state any side conditions imposed to make your model identifiable.

(b) Propose an estimator of the mean difference in response between the two levels of Factor A and give 95% confidence limits for this parameter.

(c) Propose unbiased estimators of the variances of any random terms in your model. Can you give confidence limits for these estimators?

6. A simple random sample of 290 households was selected from an urban area containing 14,828 households. Each family was asked whether it owned or rented its dwelling and also whether it had the exclusive use of an indoor toilet. Results were as follows.

	Owns		Rents	
Exclusive use of toilet	Yes	No	Yes	No
Sample count	141	6	109	34

(a) For families who rent, estimate the percentage in the area with exclusive use of an indoor toilet and estimate the standard error of your estimator.

(b) Estimate the total number of renting families in the area who do not have exclusive indoor toilet facilities and give the standard error of this estimate

(c) Are your estimators in parts (a) and (b) unbiased? Explain.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
AUGUST 2007

Applied Statistics (Ph.D. Version)

Instructions to the Student

a. Answer all 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 $Y_{ij} = \mu_i + \varepsilon_{ij}$, $i = 1, 2, 3$, $j = 1, \ldots, n$. Assume that the ε_{ij} are i.i.d. $N(0, \sigma^2)$.

(a) An experimenter assumes that the $\mu_i = h(x_i)$ are the values of an unobserved response function h, where x_1, x_2, x_3 are equally spaced values of some control variable x. Find orthogonal contrasts $\psi_1 = \sum c_{1i}\mu_i$ and $\psi_2 = \sum c_{2i}\mu_i$ representing the linear and quadratic components of h.

(b) Derive a test of H_0: $h(x_i) = \beta_0 + \beta_1 x_i$, $i = 1, 2, 3$. That is, test whether h is linear. Provide a formula for your test statistic and its distribution under H_0.

2. Consider the two way mixed model $Y_{ij} = \tau_i + b_j + e_{ij}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, where $b_j \sim N(0, \sigma_b^2)$, $e_{ij} \sim N(0, \sigma_e^2)$, and the b_j and e_{ij} are mutually independent

(a) Write out the usual ANOVA table for this problem, including the sums of squares, degrees of freedom and expected mean squares. With no assumptions on the parameters, what is the joint distribution of the sums of squares and the sample treatment means $\bar{Y}_i$?

(b) How would you test the hypothesis of no treatment differences, that is, H_0: $\tau_1 = \cdots = \tau_I$? What is the distribution of your test statistic, under both H_0 and the alternative?

(c) How would you test H_0: $\tau_1 = \cdots = \tau_I = 0$? What is the distribution of your test statistic, under both H_0 and the alternative?

3. Let D be a domain contained in a finite population $\mathcal{U}$. Assume $\mathcal{U}$ consists of N elements and that D consists of N_D elements. Both of these quantities are known and large. A simple random sample $\mathcal{S}$ of size n is selected without replacement from $\mathcal{U}$ and the variable y_i is observed on each sampled element. The sample contains n_D elements from D

(a) Show that the statistic

$$\hat{t}_D = N_D \bar{y}_d = \frac{N_D}{n_D} \sum_{i \in \mathcal{S} \cap D} y_i$$

is an approximately unbiased estimator of the domain total $t_D = \sum_{i \in D} y_i$. (You may assume that $P[n_D = 0]$ is extremely small and that $\hat{t}_D$ is zero if $n_D = 0$.)

(b) What is the approximate variance of $\hat{t}_D$?

(c) How would you estimate the variance of $\hat{t}_D$?

4 The model $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$, $i = 1, \quad , n$, was fitted to a data set, assuming that the error terms ε_i are i.i.d $N(0, \sigma^2)$ and that the model was correctly specified

(a) How could you assess the validity of the assumption that the error terms are i.i.d.? If they were i.i.d, how could you assess whether they are normally distributed?

(b) It is suspected that the nth data point may be unusual. How could you decide whether this point fails to fit the model? Is it possible that the nth point has unusual features but does not violate the model? Explain

5 Consider the linear model

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ Y_{i4} \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \gamma \end{bmatrix} + \begin{bmatrix} e_{i1} \\ e_{i2} \\ e_{i3} \\ e_{i4} \end{bmatrix}, \quad i = 1, \quad , n$$

Columns 2 and 3 of the design matrix may be regarded as levels of quantitative factors x_1 and x_2. The fourth column corresponds to a qualitative factor called "Block." The e_{ij} are i.i.d $N(0, \sigma^2)$.

(a) Show that all parameters are estimable. How many degrees of freedom are available for estimating $\sigma^2 = \text{Var } e_{ij}$?

(b) A model equation for the situation in part (a) can be written in the form

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + Block + error$$

Suppose one wishes to extend the model by adding a cross product term $\beta_{12} x_1 x_2$. Which parameters are now estimable, and which are not estimable? Find a set of r linearly independent estimable parametric functions, where r is the rank of the new design matrix. Is there any change in the degrees of freedom for the estimate of variance?

6. A population $\mathcal{U}$ consists of N psu's of size M_i, $i = 1, \quad , N$. Consider the following rejective method for selecting a pps sample $\mathcal{S}$ without replacement: Select n psu's with probabilities ψ_i and with replacement. If any psu appears more than once in this sample, reject the entire sample and select another n psu's with replacement. Repeat until you obtain a sample of n psu's with no duplicates. If $n = 2$, find the inclusion probabilities

$$\pi_i = P[i \in \mathcal{S}] \qquad \pi_{ij} = P[i \in \mathcal{S} \text{ and } j \in \mathcal{S}]$$

Argue that if all $M_i/(\sum_i M_i)$ are small, choosing ψ_i proportional to M_i makes the selection probabilities π_i nearly proportional to M_i.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
JANUARY 2007

Applied Statistics (Ph.D. Version)

Instructions to the Student

a. Answer all 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 outside 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. A sample of four observations is modeled as follows:

$$Y_1 = \mu_1 + u_1, \qquad Y_2 = \mu_1 + \mu_2 + u_1 + u_2,$$
$$Y_3 = \mu_1 + u_3, \qquad Y_4 = \mu_1 + \mu_2 + u_3 + u_4,$$

where the u_i are i.i.d. $N(0, \sigma_u^2)$.

(a) Find the ordinary least squares estimators of μ_1 and μ_2 and give formulas for their variances.

(b) Are the ordinary least squares estimators also the best linear unbiased estimators for this model? Prove your answer.

2. A certain manufactured product is a mixture of k components. A characteristic Y of the product depends on the proportions $x_1, \ldots, x_k$ of the components. Note that $x_1 + \cdots + x_k = 1$ in this situation. In an experiment, n samples are prepared and the linear model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + e_i, \quad i = 1, \ldots, n,$$

is fitted to the data. It is assumed that the e_i are i.i.d. $N(0, \sigma^2)$.

(a) Prove that none of the regression parameters is estimable.

(b) Let $k = 3$, $\beta_0 = 0$, $n = 3m$, $m > 1$. If m observations are made at each of the x vectors $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$, compute the least squares estimators of the regression coefficients and estimate σ^2.

(c) Under the assumptions of part (b), find a confidence interval for $c_1\beta_1 + c_2\beta_2 + c_3\beta_3$, where the c_i are positive constants that sum to 1.

(d) Can the confidence interval in (c) be shortened if the allocation of sample data to the x vectors $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$ is chosen not to be equal?

3. In the nested random effects model

$$Y_{ijk} = \mu + a_i + b_{ij} + e_{ijk},$$

$i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, assume that the a_i, b_{ij} and e_{ijk} are mutually independent, that the a_i are i.i.d. $N(0, \sigma_a^2)$, that the b_{ij} are i.i.d. $N(0, \sigma_b^2)$ and that the e_{ijk} are i.i.d. $N(0, \sigma_e^2)$.

(a) Write out the ANOVA table for this model, including the expected mean squares.

(b) Find a $1 - \alpha$ confidence interval for μ.

(c) Suppose $J \to \infty$ while I and K are fixed. Which of μ, σ_a^2, σ_b^2, or σ_e^2 can be estimated consistently?

4. A simple random sample S of size n is selected from a population U of size N and the quantities x_i and y_i are observed for each $i \in S$. The population mean $\bar{x}_U$ is known. The population total t_{yU} can be estimated by a statistic of the form

$$\hat{t}_A = N[\bar{y}_S + A(\bar{x}_U - \bar{x}_S)].$$

(a) Show that for any A, $\hat{t}_A$ is an unbiased estimator of t_{yU} and calculate its variance.

(b) Show that the variance of $\hat{t}_A$ is minimized when

$$A = A_{\text{opt}} = S_{yxU}/S_{xU}^2$$

(c) Assume that the population correlation between x and y is positive. Show that $\hat{t}_A$ is more precise than $\hat{t}_0 = N\bar{y}_S$ if $0 < A < 2A_{\text{opt}}$.

(d) Suppose n is large, $\bar{x}_U$ is known but S_{yxU} and S_{xU}^2 are unknown. Recommend an estimator which is more accurate than $\hat{t}_0$ and discuss its bias.

5. Suppose that $Y_{ij} = \mu + a_i + \beta x_{ij} + e_{ij}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$. Assume that the a_i and e_{ij} are mutually independent, that $a_i \sim N(0, \sigma_a^2)$ and that $e_{ij} \sim N(0, \sigma_e^2)$. Assume also that $\bar{x}_{i\cdot} = 0$ for each i.

(a) Show that a unique least squares estimator of β exists. Is there a unique least squares estimator of μ?

(b) Find the joint distribution of $\bar{Y}_{\cdot\cdot}$ and $\hat{\beta}$, the least squares estimator of the regression coefficient β.

6. The population $\mathcal{U}$ consists of N clusters, and each cluster consists of M elements (so that the population contains MN elements). Two sample designs are under consideration:

(i) a simple random sample of Mn elements, and

(ii) a simple random sample of n clusters, with data taken on each element in the sampled clusters.

Let y_{ij} denote the y value of element j from cluster i. Under either design, the (unweighted) sample mean

$$\bar{y}_{\mathcal{S}} = \frac{1}{nM} \sum_{i=1}^{n} \sum_{j=1}^{M} y_{ij}$$

will be used to estimate the population y-mean

$$\bar{y}_{\mathcal{U}} = \frac{1}{MN} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij}.$$

(a) Write the ANOVA table for the entire population with sums of squares for differences among cluster means and differences among elements within clusters.

(b) Under both designs (i) and (ii), show that $\bar{y}_{\mathcal{S}}$ is an unbiased estimator

(c) Find the variances of $\bar{y}_{\mathcal{S}}$ under both designs (i) and (ii). When would design (ii) be more precise than design (i)?

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
JANUARY 2007

Applied Statistics (M.A. Version)

Instructions to the Student

a. Answer all 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 outside 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. A sample of four observations is modeled as follows:

$$Y_1 = \mu_1 + u_1, \qquad Y_2 = \mu_1 + \mu_2 + u_1 + u_2,$$
$$Y_3 = \mu_1 + u_3, \qquad Y_4 = \mu_1 + \mu_2 + u_3 + u_4,$$

where the u_i are i.i.d. $N(0, \sigma_u^2)$.

(a) Find the ordinary least squares estimators of μ_1 and μ_2 and give formulas for their variances.

(b) Are the ordinary least squares estimators also the best linear unbiased estimators for this model? Prove your answer.

2. A certain manufactured product is a mixture of k components. A characteristic Y of the product depends on the proportions $x_1, \ldots, x_k$ of the components. Note that $x_1 + \cdots + x_k = 1$ in this situation. In an experiment, n samples are prepared and the linear model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + e_i, \quad i = 1, \ldots, n,$$

is fitted to the data. It is assumed that the e_i are i.i.d. $N(0, \sigma^2)$

(a) Prove that none of the regression parameters is estimable.

(b) Let $k = 3$, $\beta_0 = 0$, $n = 3m$, $m > 1$. If m observations are made at each of the x vectors $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$, compute the least squares estimators of the regression coefficients and estimate σ^2

(c) Under the assumptions of part (b), find a confidence interval for $c_1\beta_1 + c_2\beta_2 + c_3\beta_3$, where the c_i are positive constants that sum to 1.

3. In the nested random effects model

$$Y_{ijk} = \mu + a_i + b_{ij} + e_{ijk},$$

$i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, assume that the a_i, b_{ij} and e_{ijk} are mutually independent, that the a_i are i.i.d. $N(0, \sigma_a^2)$, that the b_{ij} are i.i.d. $N(0, \sigma_b^2)$ and that the e_{ijk} are i.i.d. $N(0, \sigma_e^2)$.

(a) Write out the ANOVA table for this model, including the expected mean squares.

(b) Find a $1 - \alpha$ confidence interval for μ.

(c) Find a $1 - \alpha$ confidence interval for $\theta = \sigma_b^2/\sigma_e^2$.

4. A simple random sample $\mathcal{S}$ of size n is selected from a population U of size N and the quantities x_i and y_i are observed for each $i \in \mathcal{S}$. The population mean $\bar{x}_U$ is known. The population total t_{yU} can be estimated by a statistic of the form

$$\hat{t}_A = N[\bar{y}_{\mathcal{S}} + A(\bar{x}_U - \bar{x}_{\mathcal{S}})].$$

(a) Show that for any A, $\hat{t}_A$ is an unbiased estimator of t_{yU} and calculate its variance.

(b) Show that the variance of $\hat{t}_A$ is minimized when

$$A = A_{\text{opt}} = S_{yxU}/S_{xU}^2$$

(c) Assume that the population correlation between x and y is positive. Show that $\hat{t}_A$ is more precise than $\hat{t}_0 = N\bar{y}_{\mathcal{S}}$ if $0 < A < 2A_{\text{opt}}$.

(d) Suppose n is large, $\bar{x}_U$ is known but S_{yxU} and S_{xU}^2 are unknown. Recommend an estimator which is more accurate than $\hat{t}_0$ and discuss its bias.

5 Suppose that $Y_{ij} = \mu + a_i + \beta x_{ij} + e_{ij}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$. Assume that the a_i and e_{ij} are mutually independent, that $a_i \sim N(0, \sigma_a^2)$ and that $e_{ij} \sim N(0, \sigma_e^2)$. Assume also that $\bar{x}_{i\cdot} = 0$ for each i.

(a) Show that a unique least squares estimator of β exists. Is there a unique least squares estimator of μ?

(b) Find the joint distribution of $\bar{Y}_{\cdot\cdot}$ and $\hat{\beta}$, the least squares estimator of the regression coefficient β.

6. The population $\mathcal{U}$ consists of N clusters, and each cluster consists of M elements (so that the population contains MN elements) Two sample designs are under consideration:

(i) a simple random sample of Mn elements, and

(ii) a simple random sample of n clusters, with data taken on each element in the sampled clusters.

Let y_{ij} denote the y value of element j from cluster i. Under either design, the (unweighted) sample mean

$$\bar{y}_{\mathcal{S}} = \frac{1}{nM} \sum_{i=1}^{n} \sum_{j=1}^{M} y_{ij}$$

will be used to estimate the population y-mean

$$\bar{y}_{\mathcal{U}} = \frac{1}{MN} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij}.$$

(a) Write the ANOVA table for the entire population with sums of squares for differences among cluster means and differences among elements within clusters.

(b) Under both designs (i) and (ii), show that $\bar{y}_{\mathcal{S}}$ is an unbiased estimator

(c) Find the variances of $\bar{y}_{\mathcal{S}}$ under both designs (i) and (ii). When would design (ii) be more precise than design (i)?

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
AUGUST 2006

Applied Statistics (Ph.D. Version)

Instructions to the Student

a. Answer all 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 outside 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 usual full rank linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

where $\boldsymbol{X}$ is an $n \times p$ design matrix, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters.

(a) Let a new $p \times 1$ covariate vector $\boldsymbol{x}_0 = (x_{01}, x_{02}, \ldots, x_{0p})^T$ correspond to an unknown observation Y_0. Obtain a $100(1-\alpha)\%$ prediction interval for Y_0.

(b) Compare the width of the prediction interval for Y_0 to the width of the confidence interval for $E(Y_0)$. Explain in words the reason for your finding.

2. An agricultural experiment was performed to compare $I = 5$ fertilizers on a certain variety of cotton. The experiment was run on $J = 6$ blocks, a block being a field. The blocks were thought to be a sample of fields on which the cotton might be grown.

(a) Formulate an appropriate model if the blocks are regarded as chosen at random. Write out the ANOVA table (source, sum of squares and degrees of freedom) and compute the expected mean squares under your model.

(b) Show how to test the hypotheses of no block effect and of no fertilizer effect. In each case find the test statistic and state its distribution under the null hypothesis.

(c) How would you create confidence limits for the fertilizer means and for a difference of fertilizer means?

(d) How, if at all, would your answers to (b) and (c) differ if the block effect was thought to be a fixed effect?

3. Consider the quadratic regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{12} x_{i1} x_{i2} + \beta_{22} x_{i2}^2 + e_i,$$

where e_i, $i = 1, \ldots, 8$ are i.i.d. $N(0, \sigma^2)$. The (x_{i1}, x_{i2}) values are $(\pm 1, \pm 1)$, $(\pm\sqrt{2}, 0)$, $(0, \pm\sqrt{2})$, where all combinations of $+$ and $-$ signs are included.

(a) Show that the estimated coefficients of the first degree terms in the quadratic model and in the reduced linear model ($\beta_{11} = \beta_{12} = \beta_{22} = 0$) are the same.

(b) Propose a test of the null hypothesis that the regression is linear against the alternative that the regression is quadratic. It is sufficient to describe how the required test statistic is computed. What is the distribution of your test statistic when the null hypothesis is true?

4. Data (x_{ij}, Y_{ij}) are thought to satisfy the analysis of covariance model

$$Y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij}$$

for $i = 1, \ldots, I$, $j = 1, \ldots, J$, where the ε_{ij} are i.i.d. $N(0, \sigma^2)$. Assume that $\bar{x}_i = 0$.

(a) Which individual parameters are estimable in the absence of any side conditions on μ, $\alpha_1, \ldots, \alpha_I$, β?

(b) Find an unbiased estimator of $\alpha_1 - \alpha_2$.

(c) Write the ANOVA table (including formulas for sums of squares and degrees of freedom) and show how to test H_0: $\alpha_1 = \cdots = \alpha_I$. Give the distribution of your test statistic under H_0.

5 A clustered population $\mathcal{U}$ consists of N clusters, with M_i elements in the ith cluster. The variable y_{ij} is defined for the jth element in the ith cluster, where $y_{ij} = 1$ if the (i, j) element has an attribute A and $y_{ij} = 0$ otherwise. It is desired to estimate P, the population proportion of elements with attribute A.

(a) Suppose a simple random sample of n clusters is selected and y_{ij} is observed for each element of a sampled cluster. Find an unbiased estimator of P and give a formula for its variance. Assume the total number of elements $\sum_{i=1}^{N} M_i$ is known and that n is large.

(b) Since M_i will be known for each sampled cluster, construct a ratio estimator of P using the sample values of M_i as auxiliary variables. Approximate the mean squared error of the ratio estimator.

(c) When would the ratio estimator be preferred to the unbiased estimator of (a)? In particular, if the cluster sizes M_i vary greatly and the cluster totals $\sum_j y_{ij}$ are very similar, which estimator would be preferred?

6. A sample s of size n is chosen at random with replacement from a population $\mathcal{U}$ of N elements. Let ψ_i denote the probability that element i is selected at any draw. The population total $t_{y\mathcal{U}}$ is estimated by

$$\hat{t}_{wr} = \frac{1}{n} \sum_{i \in \mathcal{U}} \frac{y_i Z_i}{\psi_i}$$

where Z_i is the number of times that element i was drawn.

(a) Show that $\hat{t}_{wr}$ is an unbiased estimator of $t_{y\mathcal{U}}$ and compute its variance.

(b) Find $\pi_i = P[i \in s]$.

(c) Suppose the sample was chosen without replacement with the same inclusion probabilities as in (b). Construct an unbiased estimator of $t_{y\mathcal{U}}$ for this sampling design and show that it is nearly the same as $\hat{t}_{wr}$ when N is large compared to n and all the ψ_i are small.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
AUGUST 2006

Applied Statistics (M.A. Version)

Instructions to the Student

a Answer all 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 outside 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 usual full rank linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

where $\boldsymbol{X}$ is an $n \times p$ design matrix, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters.

(a) Let a new $p \times 1$ covariate vector $\boldsymbol{x}_0 = (x_{01}, x_{02}, \ldots, x_{0p})^T$ correspond to an unknown observation Y_0. Obtain a $100(1-\alpha)\%$ prediction interval for Y_0.

(b) Compare the width of the prediction interval for Y_0 to the width of the confidence interval for $E(Y_0)$. Explain in words the reason for your finding.

2. An agricultural experiment was performed to compare $I = 5$ fertilizers on a certain variety of cotton. The experiment was run on $J = 6$ blocks, a block being a field. The blocks were thought to be a sample of fields on which the cotton might be grown. The analysis of variance table is given below

Source	Sum of Squares	d.f.
Fertilizer	$\sum_{i=1}^{5} 6(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	?
Blocks	$\sum_{j=1}^{6} 5(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2$	?
Error	$\sum_{i=1}^{5}\sum_{j=1}^{6}(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot})^2$	?

(a) Formulate an appropriate model if the blocks are regarded as chosen at random. Compute the expected mean squares under your model.

(b) Show how to test the hypotheses of no block effect and of no fertilizer effect. In each case find the test statistic and state its distribution under the null hypothesis.

(c) How would you create confidence limits for the fertilizer means and for a difference of fertilizer means?

(d) How, if at all, would your answers to (b) and (c) differ if the block effect was thought to be a fixed effect?

3. Consider the quadratic regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{12} x_{i1} x_{i2} + \beta_{22} x_{i2}^2 + e_i,$$

where e_i, $i = 1, \ldots, 8$ are i.i.d $N(0, \sigma^2)$. The (x_{i1}, x_{i2}) values are $(\pm 1, \pm 1)$, $(\pm\sqrt{2}, 0)$, $(0, \pm\sqrt{2})$, where all combinations of $+$ and $-$ signs are included.

(a) Show that the estimated coefficients of the first degree terms in the quadratic model and in the reduced linear model ($\beta_{11} = \beta_{12} = \beta_{22} = 0$) are the same.

(b) Propose a test of the null hypothesis that the regression is linear against the alternative that the regression is quadratic. It is sufficient to describe how the required test statistic is computed. What is the distribution of your test statistic when the null hypothesis is true?

4. Data (x_{ij}, Y_{ij}) are thought to satisfy the analysis of covariance model

$$Y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij}$$

for $i = 1, \ldots, I$, $j = 1, \ldots, J$, where the ε_{ij} are i.i.d. $N(0, \sigma^2)$. Assume that $\bar{x}_{i\cdot} = 0$.

(a) Which individual parameters are estimable in the absence of any side conditions on $\mu, \alpha_1, \ldots, \alpha_I, \beta$?

(b) Find an unbiased estimator of $\alpha_1 - \alpha_2$.

(c) Write the ANOVA table (including formulas for sums of squares and degrees of freedom) and show how to test H_0: $\alpha_1 = \cdots = \alpha_I$. Give the distribution of your test statistic under H_0

5. A clustered population $\mathcal{U}$ consists of N clusters, with M_i elements in the ith cluster. The variable y_{ij} is defined for the jth element in the ith cluster, where $y_{ij} = 1$ if the (i, j) element has an attribute A and $y_{ij} = 0$ otherwise. It is desired to estimate P, the population proportion of elements with attribute A.

(a) Suppose a simple random sample of n clusters is selected and y_{ij} is observed for each element of a sampled cluster. Find an unbiased estimator of P and give a formula for its variance. Assume the total number of elements $\sum_{i=1}^{N} M_i$ is known and that n is large.

(b) Since M_i will be known for each sampled cluster, construct a ratio estimator of P using the sample values of M_i as auxiliary variables. Approximate the mean squared error of the ratio estimator.

(c) When would the ratio estimator be preferred to the unbiased estimator of (a)? In particular, if the cluster sizes M_i vary greatly and the cluster totals $\sum_j y_{ij}$ are very similar, which estimator would be preferred?

6. A stratified random sample s of size n is to be chosen from a stratified population $\mathcal{U}$ for which it is known that there are N_h elements in stratum h. The goal is to estimate the population mean of a variable y. Sampling costs are the same in each stratum.

(a) How should one estimate the population mean of y if n_h elements are selected from stratum h? What is the variance of your estimate?

(b) Assume that the stratum variances S_h^2 are known. What is the optimum allocation n_h if the sample size n is fixed? Neglect the finite sample correction in your calculations.

(c) If the stratum variances are unknown, one might use proportional sampling. What is the increase in variance if this allocation is used instead of the optimum allocation?

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
JANUARY 2006

Applied Statistics (Ph.D. Version)

Instructions to the Student

a. Answer all 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 outside 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_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, n_i$, be independent $N(\mu_i, \sigma^2)$ random variables. Set up tests of the following hypotheses:

$$H_a : \mu_1 = \mu_2 = \mu_3, \qquad H_b : \mu_1 + \mu_2 + \mu_3 = 0.$$

In each case, show how to compute the test statistic and state its distribution under the null hypothesis.

2. Let (x_{ij}, Y_{ij}), $j = 1, \ldots, n_i$, $i = 1, 2$ be two independent data sets. Assume that the x_{ij} are nonrandom and that the Y_{ij} are independent with $N(\alpha_i + \beta_i x_{ij}, \sigma^2)$ distributions. Let $a_1 + b_1(x - \bar{x}_1)$ and $a_2 + b_2(x - \bar{x}_2)$ be the estimated linear regression functions from these two samples. Furthermore, let SSE_1 and SSE_2 be the sums of squared residuals from the two samples. Find a confidence interval for the quantity ξ, the x-coordinate where the true regression lines intersect.

3. A response Y depends on each of two control variables x_1 and x_2.

(a) Show that in the linear model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \varepsilon_i$$

all coefficients are estimable if the covariate vectors $\mathbf{x}_i$ are $\mathbf{x}_1 = (1,1)$, $\mathbf{x}_2 = (-1,1)$, $\mathbf{x}_3 = (1,-1)$, $\mathbf{x}_4 = (-1,-1)$, but no estimator of $\sigma^2 = \text{Var}\,\varepsilon_i$, $i = 1, \ldots, 4$, is available

(b) If instead we have $\mathbf{x}_1 = \mathbf{x}_4 = (1,1)$ with $\mathbf{x}_2$ and $\mathbf{x}_3$ as in (a), find an unbiased estimator of σ^2. Which regression coefficients, if any, are now estimable?

(c) Suppose that an additional control variable x_3 is also measured and the term $\beta_3 x_{i3}$ is added to the linear model of (a). Let the three dimensional covariate vectors be $\mathbf{x}_1 = (1,1,1)$, $\mathbf{x}_2 = (-1,1,-1)$, $\mathbf{x}_3 = (1,-1,-1)$, and $\mathbf{x}_4 = (-1,-1,1)$. Which coefficients, if any, are now estimable? What happens if the interaction term $\beta_{12} x_{i1} x_{i2}$ is deleted from the model under this design?

4. Let Y_{ijk}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, satisfy the mixed effects model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + c_{ij} + e_{ijk}$$

where the c_{ij} are i.i.d. $N(0, \sigma_C^2)$, the e_{ijk} are i.i.d. $N(0, \sigma_e^2)$, and the c_{ij} and e_{ijk} are independent

(a) Construct the ANOVA table for this model, showing sums of squares, degrees of freedom and expected mean squares.

(b) Show how to construct exact F tests for $H_A : \alpha_i \equiv 0$ and $H_C : \sigma_C^2 = 0$.

(c) Find the power of your test of $H_C : \sigma_C^2 = 0$

5 A population $\mathcal{U}$ is divided into strata $\mathcal{U}_h$ of size N_h, $h = 1, \ldots, L$ A stratified sample is drawn, resulting in data (x_{hi}, y_{hi}), $i = 1, \ldots, n_h$, $h = 1, \ldots, L$ The stratum sizes and the stratum x-totals $t_{xh} = \sum_{\mathcal{U}_h} x_{hi}$ are known The population total $t_{y\mathcal{U}}$ can be estimated by either the separate ratio estimate

$$\hat{t}_{yrs} = \sum_{h=1}^{L} \hat{B}_h t_{xh}$$

or by the combined ratio estimate

$$\hat{t}_{yrc} = t_{x\mathcal{U}}(\hat{t}_{yr}/\hat{t}_{xr}) = t_{x\mathcal{U}}\hat{B}_c$$

Here $\hat{B}_h = \bar{y}_h/\bar{x}_h$ is the estimated ratio in stratum h, $\hat{t}_{yr} = \sum_h (N_h/n_h)\bar{y}_h$ is the usual stratified estimate of the population y total, and $\hat{t}_{xr}$ is the corresponding estimate of the x total.

(a) Give formulas for the variances of $\hat{t}_{yrs}$ and $\hat{t}_{yrc}$ Assume that the sample sizes are all large and neglect finite population corrections.

(b) How could these variances be estimated from the sample?

6. A population $\mathcal{U}$ consists of N clusters, each containing M elements, so that the population contains $K = MN$ elements altogether A simple random sample $\mathcal{S}_I$ of n clusters is chosen, and from the ith cluster in $\mathcal{S}_I$ a simple random sample $\mathcal{S}_{II,i}$ of m elements is chosen The subsamples are selected independently of one another, and the variable y_{ij} is measured on the jth element selected from the ith cluster.

(a) Complete the population ANOVA table given below.

Source	Sum of Squares	d.f.	Mean Square
Between clusters	SSB =?	?	MSB =?
Within clusters	SSW =?	?	MSW =?
Total	SS_{tot} =?	?	S_y^2 =?

(b) Construct the corresponding ANOVA table for the sample In addition, compute $E[\widehat{MSB}]$ and $E[\widehat{MSW}]$, the expected mean squares from the sample ANOVA table

(c) Construct an unbiased estimator of S_y^2.

(d) Find the standard error of $\bar{y}_.$, the unweighted sample mean. Also find an estimator of this standard error

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
JANUARY 2006

Applied Statistics (M.A. Version)

Instructions to the Student

a. Answer all 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 outside 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_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, n_i$, be independent $N(\mu_i, \sigma^2)$ random variables. Set up tests of the following hypotheses:

$$H_a : \mu_1 = \mu_2 = \mu_3, \qquad H_b : \mu_1 + \mu_2 + \mu_3 = 0.$$

In each case, show how to compute the test statistic and state its distribution under the null hypothesis.

2. Let (x_{ij}, Y_{ij}), $j = 1, \ldots, n_i$, $i = 1, 2$ be two independent data sets. Assume that the x_{ij} are nonrandom and that the Y_{ij} are independent with $N(\alpha_i + \beta_i x_{ij}, \sigma^2)$ distributions. Let $a_1 + b_1(x - \bar{x}_1)$ and $a_2 + b_2(x - \bar{x}_2)$ be the estimated linear regression functions from these two samples. Furthermore, let SSE_1 and SSE_2 be the sums of squared residuals from the two samples. Find a confidence interval for the quantity ξ, the x-coordinate where the true regression lines intersect.

Hint: Consider $Z = a_1 + b_1(\xi - \bar{x}_1) - a_2 - b_2(\xi - \bar{x}_2)$.

3. A response Y depends on each of two control variables x_1 and x_2

(a) Consider the linear model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2} + \varepsilon_i$$

If the covariate vectors $\mathbf{x}_i$ are $\mathbf{x}_1 = (1,1)$, $\mathbf{x}_2 = (-1,1)$, $\mathbf{x}_3 = (1,-1)$, $\mathbf{x}_4 = (-1,-1)$, express the model in matrix form. Show that each of the regression coefficients is estimable, but that no estimator of $\sigma^2 = \text{Var}\,\varepsilon$ is available.

(b) If instead we have $\mathbf{x}_1 = \mathbf{x}_4 = (1,1)$ with $\mathbf{x}_2$ and $\mathbf{x}_3$ as in (a), find an unbiased estimator of σ^2. Which regression coefficients, if any, are now estimable?

(c) Suppose that an additional control variable x_3 is also measured so that the term $\beta_3 x_{i3}$ is added to the linear model of (a). Let the three dimensional covariate vectors be $\mathbf{x}_1 = (1,1,1)$, $\mathbf{x}_2 = (-1,1,-1)$, $\mathbf{x}_3 = (1,-1,-1)$, and $\mathbf{x}_4 = (-1,-1,1)$. Which coefficients, if any, are now estimable? What happens if the interaction term $\beta_{12} x_{i1} x_{i2}$ is deleted from the model under this design?

4. Let Y_{ijk}, $i = 1, \ldots, I$, $j = 1, \ldots, J$, satisfy the mixed effects model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + c_{ij} + e_{ijk}$$

where the c_{ij} are i.i.d $N(0, \sigma_C^2)$, the e_{ijk} are i.i.d $N(0, \sigma_e^2)$, and the c_{ij} and e_{ijk} are independent.

(a) Construct the ANOVA table for this model, showing sums of squares, degrees of freedom and expected mean squares.

(b) Show how to construct exact F tests for $H_0 : \alpha_i \equiv 0$ and $H_0 : \sigma_C^2 = 0$.

5 An experienced farmer makes an eye estimate of the weight of peaches x_i on each tree in an orchard of $N = 200$ trees. He finds a total estimated weight of $t_{x\mathcal{U}} = 11600$ lb The peaches were picked and weighed on a simple random sample $\mathcal{S}$ of $n = 10$ trees, with an actual weight of y_i from the ith tree. The sample data were reduced to the following:

$$\sum_{\mathcal{S}} x_i = 569, \qquad \sum_{\mathcal{S}} y_i = 543,$$

$$\sum_{\mathcal{S}} x_i^2 = 33227, \qquad \sum_{\mathcal{S}} x_i y_i = 31974, \qquad \sum_{\mathcal{S}} y_i^2 = 30483$$

(a) Estimate the total actual weight of peaches $t_{y\mathcal{U}}$ using the difference estimator $\hat{t}_d = t_{x\mathcal{U}} + (N/n)\sum_{\mathcal{S}}(y_i - x_i)$ and estimate its standard error.

(b) Does the linear regression estimator, based on the least squares estimator of the population regression coefficient B_1, give a substantially more precise estimate?

6 A population $\mathcal{U}$ consists of N clusters, each containing M elements, so that the population contains $K = MN$ elements altogether. A simple random sample $\mathcal{S}_I$ of n clusters is chosen, and from the ith cluster in $\mathcal{S}_I$ a simple random sample $\mathcal{S}_{II,i}$ of m elements is chosen The subsamples are selected independently of one another, and the variable y_{ij} is measured on the jth element selected from the ith cluster.

(a) Complete the population ANOVA table given below.

Source	Sum of Squares	d.f.	Mean Square
Between clusters	SSB =?	?	MSB =?
Within clusters	SSW =?	?	MSW =?
Total	SS_{tot} =?	?	S_y^2 =?

(b) Construct the corresponding ANOVA table for the sample. In addition, compute $E[\widehat{MSB}]$ and $E[\widehat{MSW}]$, the expected mean squares from the sample ANOVA table.

(c) Construct an unbiased estimator of S_y^2

(d) Find the standard error of $\bar{y}$, the unweighted sample mean

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
AUGUST 2005

Applied Statistics (Ph.D. Version)

Instructions to the Student

a. Answer all 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 outside 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 usual full rank linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

with $\mathbf{X}$ an $n \times p$ design matrix, and $\boldsymbol{\beta}$ a $p \times 1$ vector of parameters.
a. Suppose a new $p \times 1$ covariate vector $\mathbf{x}_0 = (x_{01}, x_{02}, \ldots, x_{0p})'$ is given corresponding to an unknown observation y_0. Obtain a $100(1-\alpha)\%$ prediction interval for y_0.
b. Argue that the prediction interval for y_0 is in general much wider than the confidence interval for $E(y_0)$. For simplicity assume $\mathbf{X}'\mathbf{X}$ is diagonal.

2. A surveyor measures once each of the angles α, β, γ of an area that has the shape of a triangle, and obtains unbiased measurements Y_1, Y_2, Y_3 (in radians). It is known that $\mathrm{Var}(\mathrm{Y_i}) = \sigma^2$, $i = 1, 2, 3$.

a. Estimate θ and σ^2.
b. Now, the surveyor wishes to estimate π in addition to θ. Is it possible? If it is, will this have any effect on the estimates of the angles?

3 With ϵ_{ij} independent $N(0, \sigma^2)$, and covariates x_{ij}, consider the model

$$y_{ij} = \mu + \alpha_j + \beta x_{ij} + \epsilon_{ij}, \quad i = 1, ..., k, \quad j = 1, ..., n$$

a. What type of a model is under consideration?
b. Write the model in matrix form
c. Obtain the F statistics and its df's for testing the hypothesis,

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k$$

4. Consider the balanced one-way random effects model

$$y_{ij} = \mu + a_i + \epsilon_{ij}, \quad i = 1, ..., k, \quad j = 1, ..., n$$

where for all i we have $E(a_i) = 0, Var(a_i) = \sigma_a^2$.
a. Supply the remaining assumptions on the a_i and ϵ_{ij}.
b. When such a model is appropriate?
c. Write down the associated ANOVA table.
d. Describe the estimate of σ_a^2 based on SS's from the table. Can the estimate be negative?
e. Suppose you ignore for a moment you deal with random effects and fit a fixed effects model to the data, and the hypothesis of equal means is rejected. In this case, what is your expectation regarding the random effects model?

5. A large shipment of caviar barrels from Russia was received by US Custom in Baltimore. The weight of each barrel was marked in kilograms. The custom officials needed to determine if the marked weights were correct. For this purpose a simple random sample of barrels was taken and each unit (barrel) was weighed in pounds. The data are given below. Do the marked weights represent the true weights ?

Weights in Kg.

447	445	446	446	449	447	447	449	445	448	449	448
449	446	447	449	447	448	447	449	435	441	446	449

Corresponding Weights in Lb's.

979	987	980	978	975	982	986	978	981	982	978	959
987	977	979	987	979	981	982	969	966	979	986	970

6. In stratified sampling the population of N unite is divided into subpopulations of $N_1, \ldots, N_L$ units, respectively. For stratum h, let n_h be the sample size within the stratum, let W_h be the stratum weight, and $\bar{y}_h$ be the stratum sample mean

a. Provide an example of stratified sampling.

b. Define *stratified random sampling*.

c. Explain in detail the reasons for for stratification.

d. Provide an expression for $\bar{y}_{st}$, the estimate of the population mean used in stratification. Explain your notation and state your assumptions.

e. Using your answer in (d), compute $E(\bar{y}_{st})$ and $Var(y_{st})$.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
JANUARY 2005

Applied Statistics (Ph.D. Version)

Instructions to the Student

a. Answer all 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 outside 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

$$y_i = \beta_1 + \beta_2 x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where the ϵ_i are independent $N(0, \sigma^2)$ and the x_i are nonrandom. Let $\hat{\beta}_1$ and $\hat{\beta}_2$ be the least squares estimates, and form the statistic

$$V_n = \frac{\sum_{i=1}^n (\hat{\beta}_1 + \hat{\beta}_2 x_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2}.$$

(a) Find the distribution of V_n under $H_0 : \beta_2 = 0$.

(b) Find the distribution of V_n under the alternative $H_1 : \beta_2 \neq 0$.

c) Explain how to obtain the power of the level α test of $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$ using V_n.

2. A random sample of 40 people are questioned as to whether or not they would subscribe to a new newspaper. For each person, the variables SEX (1=Female, −1=Male), AGE, and SUBS (1=yes, would subscribe; 0=no, would not subscribe) are recorded. Partial results of the logistic regression model where SUBS is regressed on SEX and AGE are given below except for some empty spaces "_____".

Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Asymptotic Chi-Square	P-value
Intercept	1	-6.9730	2.8954	_____	_____
sex	1	-1.2112	0.4779	_____	_____
age	1	0.1649	0.0652	6.3991	_____

(a) Explain the basic idea of logistic regression, and in particular write the model equation and obtain the likelihood of the logistic regression model for the present problem.

(b) Fill in the missing entries in the "Asymptotic Chi-Square" column and explain how to find the p-values.

(c) Suppose that $-2\log L = 55.452$ for the intercept only model and that $-2\log L = 38.981$ for the model with intercept and covariates SEX and AGE. Test the hypothesis that both slopes are 0.

(d) Obtain the probabilities of subscription for a female of age 50 and for a male of age 50.

(e) Who is more likely to subscribe: younger or older people? Explain.

(f) Who is more likely to subscribe: men or women? Explain.

3. Consider the following K regression lines:

$$Y_{ki} = \alpha_k + \beta_k x_{ki} + \epsilon_{ki}, \quad i = 1, 2, \ldots, n_k, \quad k = 1, 2, \ldots, K,$$

where all the ϵ_{ki} are independent $N(0, \sigma^2)$, and $N = \sum_{k=1}^{K} n_k$.

(a) Show that the K regression models can be put into a single linear model $\mathbf{Y} = \mathbf{X}\gamma + \boldsymbol{\varepsilon}$.

(b) Explain how to obtain a test statistic and its distribution for testing the hypothesis that all the lines are parallel: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_K$.

(c) Explain how to obtain a test statistic and its distribution for testing the hypothesis the K lines are coincident:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_K \text{ and } \beta_1 = \beta_2 = \cdots = \beta_K$$

against the general alternative.

(d) How will the answer in (c) change if you assume $\beta_1 = \beta_2 = \cdots = \beta_K$ as part of the alternative hypothesis?

4. Suppose we wish to compare the effects of three drugs on people by measuring some response Y. Let Y_{ij} be the response of the jth person taking the ith drug, $i = 1, 2, 3$, $j = 1, \ldots, J$. Assume all the error terms are independent $N(0, \sigma^2)$.

(a) Describe a one-way ANOVA model appropriate for this problem.

(b) Suppose we know the effect of a drug depends quadratically on the age of the person. Explain how to model this problem.

(c) Suppose there is no interaction between the age and the type of drug. Explain how to model this problem.

(d) Summarize your models in (a), (b), (c) in matrix form.

5. A simple random sample of n clusters is selected from a very large population $\mathcal{U}$ of N clusters, each containing M elements. Observations are made on each element of a sampled cluster. Assume $M << N$.

(a) Write the population ANOVA table for a variable y, indicating between cluster and within cluster sums of squares.

(b) Let

$$t_y = \sum_{i \in \mathcal{U}} \sum_{j=1}^{M} y_{ij}$$

be the population total of y. In terms of quantities in the ANOVA table, find the design effect

$$\text{deff} = \text{Var}\,[\hat{t}_{\text{clus}}]/\text{Var}\,[\hat{t}_{\text{srs}}],$$

where $\hat{t}_{\text{clus}}$ is the estimated total based on cluster sampling and $\hat{t}_{\text{srs}}$ is the estimated total based on simple random sampling of Mn elements. What are the highest and lowest values of the design effect? Describe the structure of the population in the extreme cases.

Ignore the finite population correction.

6. A simple random sample of size n is selected from a population $\mathcal{U}$ of size N *with replacement*. The goal is to estimate the population total $t = \sum_{i \in \mathcal{U}} y_i$. Let

$$Q_i = \text{number of times unit } i \text{ appears in the sample}$$

and consider the following estimator of the population total:

$$\hat{t} = \frac{N}{n} \sum_{i=1}^{N} Q_i y_i.$$

(a) Argue that the joint distribution of $Q_1, \ldots, Q_n$ is multinomial with n trials and $p_1 = \cdots = p_N = 1/N$.

(b) Show that $\hat{t}$ is an unbiased estimator of t and find its variance.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
AUGUST 2004

Applied Statistics (Ph.D. Version)

Instructions to the Student

a. Answer all 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 outside 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 model $Y_i = \alpha + \beta(x_i - \bar{x}) + \varepsilon_i$, where the ε_i are independent $N(0, \sigma^2)$, and $E(y_i) = \eta_i$, $i = 1, \ldots, n$. Let $\hat{\alpha}, \hat{\beta}$ denote the least squares estimators. Define:

$$S^2 = \frac{1}{n-2} \sum_{1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

(a) Obtain the least squares estimators $\hat{\alpha}, \hat{\beta}$.

(b) Prove that $E(S) \leq \sigma$

(c) Are the three statistics $\hat{\alpha}, \hat{\beta}, S^2$ independent? Provide a rigorous argument.

(d) What is the distribution of $(\hat{\beta} - \beta)^2 \sum (x_i - \bar{x})^2 / S^2$?

2. A surveyor makes one measurement on each of the angles α, β, γ of an area that has the shape of a triangle, and obtains unbiased measurements Y_1, Y_2, Y_3 (in radians). It is known that $\text{Var}(Y_i) = \sigma^2$, $i = 1, 2, 3$.

(a) Estimate the three angles and σ^2.

(b) Now, the surveyor wishes to estimate π in addition to the angles of the triangle. Is it possible? If it is, will this have any effect on the estimates of the other parameters in this problem?

3. We wish to find an increasing function f such that in the model

$$f(Y_i) = \mathbf{x}_i^T \widehat{\boldsymbol{\beta}} + \varepsilon_i, \quad i = 1, \ldots, n$$

the ε_i have approximately the same variance. Assume the sample size is large, that $\mu_i = \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}$, and that for a known function w,

$$\text{Var}(Y_i) = w(\mu_i).$$

(a) Argue that the ε_i will have approximately the same variance when

$$f(\mu) = \int \frac{d\mu}{w(\mu)^{1/2}}.$$

(b) Suppose the responses Y_i are Poisson(μ_i). What is the form of f?

(c) Suppose the responses Y_i are Binomial(m, p_i). What is the form of f?

4. From a population $\mathcal{U}$ of size $N = 3$, a simple random sample $\mathcal{S}$ of size $n = 2$ is selected. The goal is to estimate the population total

$$t_{y\mathcal{U}} = y_1 + y_2 + y_3$$

Prove that the estimator $\hat{t}(\mathcal{S})$ defined by

$$\hat{t}(\mathcal{S}) = \begin{cases} (3/2)y_1 + (3/2)y_2 & \text{if } \mathcal{S} = \{1, 2\} \\ (3/2)y_1 + 2y_3 & \text{if } \mathcal{S} = \{1, 3\} \\ (3/2)y_2 + y_3 & \text{if } \mathcal{S} = \{2, 3\} \end{cases}$$

is unbiased. In addition, prove that $\text{Var}[\hat{t}(\mathcal{S})]$ is smaller than the the variance of the equally weighted estimator of the total $3\bar{y}_{\mathcal{S}}$ if $y_3(3y_2 - 3y_1 - y_2) > 0$.

5. A researcher comes to you for advice about the analysis of an experiment she has conducted. She has used 5 treatments and she has 9 observations of some response variable Y for each treatment. She shows you the following ANOVA from a computer printout:

Source	d.f.	S.S.	F statistic	$P[> F]$
Treatments	4	100	8.33	0.0001
Error	40	120		

The researcher wants to know from you whether the analysis is correct, and if so what it means. For each of the three scenarios below answer the following questions:

(a) What is the appropriate model? Write a model equation, including any and all effects. Explain which effects are fixed and which are random, and state any distributional assumptions.

(b) Based on the model in (a), what is the corresponding ANOVA (include sources of variation, d.f., formulas for sums of squares, and the statistic for testing $H_0: \tau_1 = \tau_2 = \cdots = \tau_5$).

(c) Is the ANOVA from the printout above appropriate for testing $H_0: \tau_1 = \tau_2 = \cdots = \tau_5$?

Scenario I: 45 animals were used for the study and each treatment was applied to 9 animals selected at random.

Scenario II: 45 animals were used but they came from 9 different litters of size 5 each, and each treatment was assigned to one animal selected at random from each litter.

Scenario III: 15 animals were used, each treatment was assigned to 3 animals selected at random, and 3 observations were made on each animal.

6. A population $\mathcal{U}$ consists of N clusters or primary sampling units, each of which contains M elements. Let y_{ij} denote the value of a variable y for the jth element in the ith primary sampling unit (psu). A simple random sample of n psu's is chosen, and from each psu a simple random sample of m elements is selected. Sampling from different psu's is performed independently. The statistic

$$\bar{y}_{\mathcal{S}} = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij}$$

is used to estimate the population average

$$\bar{y}_{\mathcal{U}} = \frac{1}{MN} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij}.$$

(a) Write down the ANOVA table for the population (on an element basis), showing explicit formulas for the sums of squares between and within psu's.

(b) In terms of MSB and MSW, the mean squares in the population ANOVA table, prove that

$$\text{Var}\left[\bar{y}_{\mathcal{S}}\right] = \left(1 - \frac{n}{N}\right) \frac{\text{MSB}}{nM} + \left(1 - \frac{m}{M}\right) \frac{\text{MSW}}{mn}.$$

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
JANUARY, 2004

Applied Statistics (Ph.D. Version)

Instructions to the Student

a Answer all 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 outside 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 one-way ANOVA model

$$Y_{ij} = \theta_i + \epsilon_{ij}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, n_i$$

where the ϵ_{ij} are independent $N(0, \sigma^2)$ errors. Let $a_1, \ldots, a_k$ be fixed constants, chosen before the data were observed.

(a) At level α, test H_0: $\sum_{i=1}^{k} a_i\theta_i = 0$ versus H_1: $\sum_{i=1}^{k} a_i\theta_i \neq 0$.

(b) Specialize your test to compare treatment 1 to the average of treatments 2 and 3.

(c) Would your answer to (a) change if the constants $a_1, \ldots, a_k$ had been chosen after looking at the data? If so, how? If not, why not?

2. In a study of which of three machines is preferable for an industrial process, six employees were selected at random and each employee operated each machine three times. The result was Y_{ijk}, the output of the kth run of employee j on machine i, for $i = 1, 2, 3$, $j = 1, \ldots, 6$, $k = 1, 2, 3$. It was assumed that the following linear mixed model described the data:

$$Y_{ijk} = \mu_i + b_j + c_{ij} + e_{ijk},$$

where μ_1, μ_2, μ_3 are fixed but unknown parameters, $b_j \sim N(0, \sigma_b^2)$, $c_{ij} \sim N(0, \sigma_c^2)$, and e_{ijk} $N(0, \sigma_e^2)$. The usual ANOVA table for a balanced two way layout with replication was calculated, yielding the following results.

Source	d.f.	Mean Square	E(MS)
Machines		877.63	
Employees		248.38	
Interaction		42.65	
Residual		0.93	

(a) Find the missing d.f. and E(MS) values, indicating any functions of the fixed parameters implicitly by notation such as $Q(\mu_1, \mu_2, \mu_3)$.

(b) How can one test for differences among machines?

(c) Find a point estimate $\hat{\mu}_i$ for μ_i and find a point estimate for Var $\hat{\mu}_i$.

3. A surveyor makes a single measurement on each of the angles $\beta_1, \beta_2, \beta_3$ of an area that has the shape of a triangle, and obtains unbiased measurements Y_1, Y_2, Y_3 (in radians). It is known that the measurements have a common but unknown variance σ^2.

(a) Find the least squares estimates of the unknown angles and their variances.

(b) Is it possible to obtain an unbiased estimate for σ^2? If so, find it, and if not, explain why not.

(c) Suppose it is known in advance that $\beta_1 = \beta_2$. Find the least squares estimates of the unknown angles and their variances in this case.

4 Consider the regression model

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \ldots, n$$

where the ϵ_i are independent $N(0, \sigma^2)$ errors, and $\sum_i x_i = 0$, $\sum_i x_i^2 = 1$. Let $r = \sum_i x_i Y_i$. Suppose we estimate β by

$$\tilde{\beta} = \begin{cases} 0, & \text{if } |r| < c \\ r, & \text{if } |r| \geq c \end{cases}$$

where c is a positive constant.

(a) Compute the MSE's of both the usual LSE $\hat{\beta}$ of β and of $\tilde{\beta}$.

(b) Verify that $\text{MSE}(\tilde{\beta}) < \text{MSE}(\hat{\beta})$ when $\beta = 0$.

5. A simple random sample of n elements is selected from a population $\mathcal{U}$ of N elements, and variables x and y are measured on each element in the sample. The regression estimator of the population mean $\bar{y}_{\mathcal{U}}$ is

$$\hat{\bar{y}}_{\text{reg}} = \bar{y}_S + \hat{B}_1(\bar{x}_{\mathcal{U}} - \bar{x}_S),$$

where

$$\hat{B}_1 = \frac{\sum_{i \in S}(x_i - \bar{x}_S)(y_i - \bar{y}_S)}{\sum_{i \in S}(x_i - \bar{x}_S)^2}$$

is the sample least squares estimator of the population least squares regression slope

$$\hat{B}_1 = \frac{\sum_{i \in \mathcal{U}}(x_i - \bar{x}_{\mathcal{U}})(y_i - \bar{y}_{\mathcal{U}})}{\sum_{i \in \mathcal{U}}(x_i - \bar{x}_{\mathcal{U}})^2}$$

Show that the bias of $\hat{\bar{y}}_{\text{reg}}$ is $-\text{Cov}(\hat{B}_1, \bar{x}_S)$ and find a large sample approximation for the mean squared error of $\hat{\bar{y}}_{\text{reg}}$

6 A simple random sample S of size $n = n_1 + n_2$ is drawn from a finite population $\mathcal{U}$, and a simple random subsample S_1 of size n_1 is drawn from S. Define the sample and subsample means by

$$\bar{y} = \frac{1}{n}\sum_{S} y_i \quad \text{and} \quad \bar{y}_1 = \frac{1}{n_1}\sum_{S_1} y_i.$$

and let

$$\bar{y}_2 = \frac{1}{n_2}\sum_{S\setminus S_1} y_i = \frac{n\bar{y} - n_1\bar{y}_1}{n_2}$$

be the mean of the elements of the sample not included in the subsample.

(a) Prove that $\text{Var}\,(\bar{y}_1 - \bar{y}_2) = S^2_{y\mathcal{U}}(1/n_1 + 1/n_2)$.

(b) Prove that $\text{Var}\,(\bar{y} - \bar{y}_1) = S^2_{y\mathcal{U}}(1/n_1 - 1/n)$

(c) Prove that $\text{Cov}(\bar{y}, \bar{y}_1 - \bar{y}) = 0$

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
AUGUST, 2003

Applied Statistics (Ph.D. Version)

Instructions to the Student

a. Answer all 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 outside 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. Engineers A and B collected replicated data of the form (x_i, Y_{ij}), $i = 1, \ldots, k$, $j = 1, \ldots, m$. A plotted $\bar{Y}_i$ vs. x_i and claimed that there is a nonlinear relationship between the response variable Y and the control variable x. B used ordinary least squares to fit a linear model $Y = \beta_0 + \beta_1 x + e$ and claimed that a straight line model was adequate to fit the data because he found $R^2 > 0.9$.

(a) Assuming $Y_{ij} = m(x_i) + e_{ij}$ for some function m and that the e_{ij} are i.i.d. $N(0, \sigma^2)$, how would you settle the dispute between A and B? Are either of them using correct reasoning to support their claims?

(b) Suppose that there had been no replication ($m = 1$). What guidance, if any, could you provide to A and B?

2 Let $Y_{ij} = \mu + a_i + e_{ij}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, be data from a balanced one-way random effects ANOVA, where the a_i are i.i.d. $N(0, \sigma_a^2)$ and the e_{ij} are i.i.d. $N(0, \sigma_e^2)$.

In terms of sample averages and statistics calculated in the usual ANOVA table, find $1 - \alpha$ confidence intervals for μ, σ_e^2 and σ_a^2/σ_e^2.

3. A population $\mathcal{U}$ consists of N clusters with M_i elements in the ith cluster. Altogether the population contains $K = \sum_{i=1}^{N} M_i$ elements. A simple random sample $\mathcal{S}$ of n clusters is selected, and a variable y is measured on each element of the selected clusters, yielding data $\{y_{ij},\ i \in \mathcal{S},\ j = 1, \ldots, M_i\}$. Consider the following estimators of the population total $t_y = \sum_{i \in \mathcal{U}} \sum_{j=1}^{M_i} y_{ij} = \sum_{i \in \mathcal{U}} t_i$, where t_i is the ith cluster total:

(i) the simple expansion estimator

$$\hat{t}_1 = \frac{N}{n} \sum_{i \in \mathcal{S}} t_i,$$

(ii) the ratio to size estimator

$$\hat{t}_2 = K \frac{\sum_{i \in \mathcal{S}} t_i}{\sum_{i \in \mathcal{S}} M_i}.$$

(a) Is either of these two estimators unbiased? Explain your answer.

(b) Give expressions for the variances of these estimators, assuming both n and N are large. Your answer should be exact if possible, otherwise approximate. When would one expect $\hat{t}_1$ to be less accurate than $\hat{t}_2$?

4. Random variables Y_{ij}, $1 \le i < j \le 3$, are observed, where the Y_{ij} are independent $N(\beta_i - \beta_j, \sigma^2)$. The parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T$ and σ^2 are unknown. The parameters β_i can be regarded as effects of a factor B.

(a) Assuming that all three combinations of (i, j) are observed, write a linear model of the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ to describe the data, where $\mathbf{Y} = (Y_{12}, Y_{13}, Y_{23})^T$. Write the $\mathbf{X}$ matrix explicitly.

(b) Are any of the individual parameters β_i estimable? Is an unbiased estimator of σ^2 available? Prove your answer.

5 Consider the quadratic regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + e$$

where, as usual, the e's are i.i.d. $N(0, \sigma^2)$ random errors.

(a) If the x_{ij} are all ± 1, verify that the parameters β_0 and β_{jj}, $j = 1, 2$, are not estimable.

(b) Suppose that n observations are available for each combination of x values with $x_1 = \pm 1$, $x_2 = \pm 1$ and m additional observations are available at $(x_1, x_2) = (0, 0)$. Show that β_0 and $\beta_0 + \beta_{11} + \beta_{22}$ are estimable, but that β_{11} and β_{22} are not individually estimable.

(c) Propose a test of H_0: $\beta_{11} = \beta_{22} = 0$ and give the distribution of your test statistic under H_0.

6. A simple random sample of households $\mathcal{S}$ is selected from a very large population. The data will be used to estimate the proportion $p_{\mathcal{U}}$ of households with a certain attribute. It is believed that $p_{\mathcal{U}}$ is between 30% and 70%. What sample sizes are needed to meet the following requirements for precision?

(a) The population proportion $p_{\mathcal{U}}$ is to be estimated with a standard error of no more than 3%.

(b) The proportions $p_{\mathcal{U}_k}$ in each of the three income classes—under \$25,000, \$25,000 to \$50,000, and over \$50,000 ($k = 1, 2, 3$, respectively)—are each to be estimated with a standard error of no more than 3%.

(c) The differences of proportions $(p_{\mathcal{U}_j} - p_{\mathcal{U}_k})$ for each pair of classes in (b) are to be estimated with a standard error of no more than 3%.

Income statistics indicate that the proportions in the three classes above are 50%, 40% and 10%.

You should provide separate answers for each of parts (a), (b), (c). The finite population correction may be neglected.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
JANUARY, 2003

Applied Statistics (Ph.D. Version)

Instructions to the Student

a. Answer all 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 outside 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 $Y_{ij} = \mu + a_i + e_{ij}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, be data from a one-way random effects ANOVA, where the a_i are i.i.d. $N(0, \sigma_a^2)$ and the e_{ij} are i.i.d. $N(0, \sigma_e^2)$.

(a) Write out the usual ANOVA table and compute the expected mean squares, $E(MS_A)$ and $E(MS_E)$.

(b) Find the distribution of the statistic $F = MS_A/MS_E$ under general conditions.

(c) Find a $1 - \alpha$ confidence interval for the intraclass correlation coefficient

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

2. A questionnaire is to be sent to a sample of high schools to find out which schools provide certain facilities, such as a computer laboratory or a course in Russian. The ith school has an enrollment of M_i students and the total number of students is $K = \sum_{i=1}^{N} M_i$. For a certain facility, it is desired to estimate the proportion of students attending a school with the facility:

$$p_U = \frac{\sum_w M_i}{\sum_{i=1}^{N} M_i},$$

where $\sum_w$ is a sum over the schools *with* the facility.

A sample of n schools is selected *with* replacement and with probability proportional to M_i. For one facility of interest, it was found from the sample that a schools had the facility.

(a) Show that $\hat{p} = a/n$ is an unbiased estimator of p_U and that

$$\text{Var}\,(\hat{p}) = \frac{p_U(1-p_U)}{n}$$

(b) Show that an unbiased estimator of Var $(\hat{p})$ is

$$\hat{V}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}$$

[*Hint*: Let $t_i = M_i$ if the ith school has the facility and 0 otherwise.]

3. Consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{X}$ is an $n \times p$ matrix with rank $p \leq n$, $E(\mathbf{e}) = \mathbf{0}$ and Var-Cov$(\mathbf{e}) = \sigma^2\mathbf{I}$. Let $\boldsymbol{\xi}_i$ denote the ith column of X. Suppose $\{\hat{\beta}_1, \ldots, \hat{\beta}_p\}$ is a set of least squares estimates under the general model. Show that $\{\hat{\beta}_1, \ldots, \hat{\beta}_m\}$, $m < p$ are also least squares estimates under the null hypothesis $H_0 : \beta_{m+1} = \ldots = \beta_p = 0$ if and only if $\boldsymbol{\xi}_i \perp \sum_{j=m+1}^{p} \hat{\beta}_j \boldsymbol{\xi}_j$, $i = 1, \ldots, m$.

4. In an agricultural study, the weight in pounds (Y) and age in weeks (x) were recorded for samples of turkeys selected from three different treatment groups. The following (full) model was fitted to the data:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_1 z_{ij} + \alpha_2 w_{ij} + e_{ij},$$

where $i = 1, 2, 3$ indexes treatment groups, $j = 1, \ldots, J_i$ indexes turkeys within group, $z_{ij} = I\{i = 1\}$, $w_{ij} = I\{i = 2\}$, and $I\{\cdot\}$ denotes the indicator function of an event. The sample sizes were $J_1 = 4$, $J_2 = 4$, and $J_3 = 5$. Least squares analysis of this model yielded $R^2 = 97.94\%$. By contrast, when the simple linear regression model (reduced model)

$$Y_{ij} = \beta_0^* + \beta_1^* x_{ij} + e_{ij}$$

was fitted to the data, it was found that $R^2 = 64.77\%$.

(a) The experimenters claimed that the large differences in R^2 showed that the treatment differences were significant. Can this statement be verified? If so, calculate an appropriate test statistic and give its distribution under the null hypothesis of no treatment differences. If not, explain why not.

(b) How would you test whether the mean difference between Groups 1 and 2 was nonzero, assuming this comparison had been planned in advance? Would the same testing procedure be used if this comparison was suggested by examination of the data?

5. A stratified population has L strata with N_h units in stratum h. Assume that independent simple random samples of size n_h are selected from stratum h, $h = 1, \ldots, L$. The *combined ratio estimator* of the population total $t_{y\mathcal{U}}$ is

$$\hat{t}_{rc} = t_{x\mathcal{U}} \frac{\sum_{h=1}^{L} N_h \bar{y}_h}{\sum_{h=1}^{L} N_h \bar{x}_h}.$$

Argue that $\hat{t}_{rc}$ is approximately unbiased and derive a formula for its variance when all the n_h are large.

6. Independent observations Y_{ij}, $i = 1, 2$, $j = 1, 2$, were modeled as a two factor ANOVA:

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij},$$

where the e_{ij} are independent random variables with a common $N(0, \sigma^2)$ distribution. Representing the data in vector form, the following decomposition was calculated:

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \end{bmatrix} = \begin{bmatrix} 50 \\ 50 \\ 50 \\ 50 \end{bmatrix} + \begin{bmatrix} 2 \\ 2 \\ -2 \\ -2 \end{bmatrix} + \begin{bmatrix} 5 \\ -5 \\ 5 \\ -5 \end{bmatrix} + \begin{bmatrix} 3 \\ -3 \\ -3 \\ 3 \end{bmatrix}$$

(a) Compute the ANOVA table for the data.

(b) Compute statistics for testing the hypotheses H_A: no Factor A effect and H_B: no Factor B effect. What are the the distributions of the test statistics under the null hypothesis?

(c) Is there some test of whether this additive model fits this data? Would there exist a test if there had been three levels of Factor A and two levels of Factor B?

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
JANUARY, 2003

Applied Statistics (M.A. Version)

Instructions to the Student

a. Answer all 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 outside 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 $Y_{ij} = \mu + a_i + e_{ij}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, be data from a one-way random effects ANOVA, where the a_i are i.i.d. $N(0, \sigma_a^2)$ and the e_{ij} are i.i.d. $N(0, \sigma_e^2)$.

(a) Write out the usual ANOVA table and compute the expected mean squares, $E(MS_A)$ and $E(MS_E)$.

(b) Find the distribution of the statistic $F = MS_A/MS_E$ under general conditions.

(c) Find a $1 - \alpha$ confidence interval for the variance ratio

$$\theta = \frac{\sigma_a^2}{\sigma_e^2}.$$

2. A questionnaire is to be sent to a sample of high schools to find out which schools provide certain facilities, such as a computer laboratory or a course in Russian. The ith school has an enrollment of M_i students and the total number of students is $K = \sum_{i=1}^{N} M_i$. For a certain facility, it is desired to estimate the proportion of students attending a school with the facility:

$$p_U = \frac{\sum_w M_i}{\sum_{i=1}^{N} M_i},$$

where $\sum_w$ is a sum over the schools *with* the facility.

A sample of n schools is selected *with* replacement and with probability proportional to M_i. For one facility of interest, it was found from the sample that a schools had the facility.

(a) Show that $\hat{p} = a/n$ is an unbiased estimator of p_U and that

$$\text{Var}\,(\hat{p}) = \frac{p_U(1-p_U)}{n}.$$

(b) Show that an unbiased estimator of Var $(\hat{p})$ is

$$\hat{V}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}.$$

[*Hint*: Let $t_i = M_i$ if the ith school has the facility and 0 otherwise.]

3. Consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{X}$ is an $n \times p$ matrix with rank $r \leq p \leq n$, $E(\mathbf{e}) = \mathbf{0}$ and Var-Cov$(\mathbf{e}) = \sigma^2\mathbf{I}$. Let $\psi = \mathbf{c}^T\boldsymbol{\beta}$ be a linear parametric function.

(a) Define estimability of ψ.

(b) Show that the following conditions are equivalent:

(i) The linear parametric function ψ is estimable.

(ii) For some n-dimensional vector $\mathbf{a}$, $\mathbf{c}^T = \mathbf{a}^T\mathbf{X}$.

(iii) For some p-dimensional vector $\mathbf{r}$,

$$\mathbf{X}^T\mathbf{X}\mathbf{r} = \mathbf{c} \quad \text{or} \quad \mathbf{r}^T\mathbf{X}^T\mathbf{X} = \mathbf{c}^T.$$

4. In an agricultural study, the weight in pounds (Y) and age in weeks (x) were recorded for samples of turkeys selected from three different treatment groups. The following (full) model was fitted to the data:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_1 z_{ij} + \alpha_2 w_{ij} + e_{ij},$$

where $i = 1, 2, 3$ indexes treatment groups, $j = 1, \ldots, J_i$ indexes turkeys within group, $z_{ij} = I\{i = 1\}$, $w_{ij} = I\{i = 2\}$, and $I\{\cdot\}$ denotes the indicator function of an event. The sample sizes were $J_1 = 4$, $J_2 = 4$, and $J_3 = 5$. Least squares analysis of this model yielded $R^2 = 97.94\%$. By contrast, when the simple linear regression model (reduced model)

$$Y_{ij} = \beta_0^* + \beta_1^* x_{ij} + e_{ij}$$

was fitted to the data, it was found that $R^2 = 64.77\%$.

(a) The experimenters claimed that the large differences in R^2 showed that the treatment differences were significant. Assuming normal independent errors, can this statement be verified? If so, calculate an appropriate test statistic and give its distribution under the null hypothesis of no treatment differences. If not, explain why not.

(b) How would you test whether the mean difference between Groups 1 and 2 was nonzero, assuming this comparison had been planned in advance? Would the same testing procedure be used if this comparison was suggested by examination of the data?

5. A stratified population has L strata with N_h units in stratum h. Assume that independent simple random samples of size n_h are selected from stratum h, $h = 1, \ldots, L$. The *combined ratio estimator* of the population total $t_{y\mathcal{U}}$ is

$$\hat{t}_{rc} = t_{x\mathcal{U}} \frac{\sum_{h=1}^{L} N_h \bar{y}_h}{\sum_{h=1}^{L} N_h \bar{x}_h}$$

Show that $\hat{t}_{rc}$ is approximately unbiased and derive a formula for its variance when all the n_h are large.

6. Independent observations Y_{ij}, $i = 1, 2$, $j = 1, 2$, were modeled as a two factor ANOVA:

$$Y_{ij} = \mu + \alpha_i + \beta_j + e_{ij},$$

where the e_{ij} are independent random variables with a common $N(0, \sigma^2)$ distribution. Representing the data in vector form, the following decomposition was calculated:

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \end{bmatrix} = \begin{bmatrix} 50 \\ 50 \\ 50 \\ 50 \end{bmatrix} + \begin{bmatrix} 2 \\ 2 \\ -2 \\ -2 \end{bmatrix} + \begin{bmatrix} 5 \\ -5 \\ 5 \\ -5 \end{bmatrix} + \begin{bmatrix} 3 \\ -3 \\ -3 \\ 3 \end{bmatrix}$$

(a) Compute the ANOVA table for the data.

(b) Compute statistics for testing the hypotheses H_A: no Factor A effect and H_B: no Factor B effect. What are the the distributions of the test statistics under the null hypothesis?

(c) Is there some test of whether this additive model fits this data? Would there exist a test if there had been three levels of Factor A and two levels of Factor B?