**Applied Statistics Exam**
**May 2008**
**Time: 12:00pm–3:30pm**

This exam consists of 2 parts. You must answer 5 questions total and must answer at least 2 questions from each part. Make sure to clearly indicate which problems you are attempting. Some formulas and tables are given at the end of this exam.

PART A:

**A1.** Suppose that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{y}$ is a $N$-dimensional column vector of outputs, $\boldsymbol{X}$ is a $N \times p$ design matrix of fixed inputs where $rank(\boldsymbol{X}) = p$ and $p < N$, $\boldsymbol{\beta}$ is a $p$-dimensional column vector of coefficients, and $\boldsymbol{\epsilon}$ follows a Normal distribution with mean vector $\mathbf{0}_N$ and covariance matrix $\sigma^2 \boldsymbol{I}_N$.

(a) Let $Q(\boldsymbol{b}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2$. Compute $\dfrac{\partial Q}{\partial \boldsymbol{b}}$.

(b) Derive the solution to the score equation $\dfrac{\partial Q}{\partial \boldsymbol{b}} = \mathbf{0}_p$ and denote it by $\hat{\boldsymbol{\beta}}$.

(c) Find the distribution of $\hat{\boldsymbol{\beta}}$.

(d) Find the distribution of the residual vector $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

(e) Show that $\hat{\boldsymbol{\beta}}$ and $RSS = \boldsymbol{r}^\top \boldsymbol{r}$ are independent.

**A2.** Consider the simple linear regression model in which $\boldsymbol{y} \sim \text{Normal}_N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_N)$ where $\boldsymbol{X} = [\boldsymbol{J} \vdots \boldsymbol{x}]$ and $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$; here $\beta_0$ is the unknown fixed intercept parameter, $\beta_1$ is the unknown fixed slope parameter, $\sigma^2$ is the unknown fixed variance, $\boldsymbol{x}$ is a known non-random $N$-dimensional column vector, and $\boldsymbol{J}$ is a $N$-dimensional column vector of ones.

Assume that we have the following summary statistics:

$$N = \boldsymbol{J}^\top \boldsymbol{J} = 10,\ \boldsymbol{J}^\top \boldsymbol{x} = 20,\ \boldsymbol{x}^\top \boldsymbol{x} = 100,\ \boldsymbol{J}^\top \boldsymbol{y} = 0, \boldsymbol{x}^\top \boldsymbol{y} = -10,\ \boldsymbol{y}^\top \boldsymbol{y} = 46.$$

(a) Compute the maximum likelihood estimates of $\boldsymbol{\beta}$ and $\sigma^2$. Denote them as $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$, respectively.

(b) What is the distribution of $\dfrac{\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2}{2s^2}$, where $s^2 = \frac{N}{N-2}\hat{\sigma}^2$?

(c) Find a 95% confidence ellipse for $\boldsymbol{\beta}$ and find values $A$, $B$, and $C$ so that the confidence ellipse can be expressed in the form

$$A(\hat{\beta}_0 - \beta_0)^2 + B(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1) + C(\hat{\beta}_1 - \beta_1)^2 \leq 1.$$

**A3.** Consider the regression model

$$y_i \sim \begin{cases} \text{Normal}(\mu_1 + \beta x_i, \sigma^2) & \text{for } i = 1, \ldots, n_1 \\ \text{Normal}(\mu_2 + \beta x_i, \sigma^2) & \text{for } i = n_1 + 1, \ldots, n_1 + n_2 \end{cases}$$

where $\mu_1$, $\mu_2$, and $\beta$ are unknown fixed regression parameters, $\sigma^2$ is the unknown fixed variance, $n_1$ and $n_2$ are the known number of observations in groups 1 and 2, respectively, and $x_i, i = 1, \ldots, n_1 + n_2$, are known and non-random.

Suppose that $n_1 = 3$, $n_2 = 3$, and the data are given in the following table:

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $x_i$ | $-1$ | 0 | 1 | $-1$ | $-1$ | 2 |
| $y_i$ | 1 | 3 | $-3$ | 0 | 4 | 2 |

(a) Compute the maximum likelihood estimates of $\mu_1$, $\mu_2$, $\beta$, and $\sigma^2$.
(b) Compute an appropriate test statistic for testing $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$, and determine whether $H_0$ should be rejected when tested at level $\alpha = .05$.

**A4.** Consider the design matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^d \\ 1 & x_2 & x_2^2 & \cdots & x_2^d \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^d \end{bmatrix}$$

used in a polynomial regression model. Assume that $N > d + 1$ and $\boldsymbol{X}$ is full rank.
(a) Show that the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$ is symmetric and idempotent and that $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$.
(b) Let $h_{ij}$ denote the element in the $i$th row and $j$th column of $\boldsymbol{H}$. Using the result of part (a), show that $\sum_{j=1}^{N} h_{ij} = 1$ for $i = 1, \ldots, N$.
(c) Let $\chi_0 = [1, x_0, x_0^2, \cdots, x_0^d]$ be a $(d+1)$-dimensional row vector used for modeling a new input $x_0 \neq 0$, and let $b(\chi_0) = \chi_0(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top = [b_1, \cdots, b_N]$. Show that $\sum_{i=1}^{N} b_i x_i^k = x_0^k$ for $k = 0, 1, \ldots, d$.
(d) Use the result in (c) to show that $\sum_{i=1}^{N} b_i (x_i - x_0)^k = 0$ for $k = 0, 1, \ldots, d$.

PART B:

**B1.** Suppose each of $K$ classes has an associated target $t_k$, which is a vector of all zeros, except a one in the $k$th position. Show that classifying to the largest element of $\hat{y}$ amounts to choosing the closest target, $\min_k \|t_k - \hat{y}\|$, if the elements of $\hat{y}$ sum to one.

**B2.** Consider a logistic regression model with no intercept

$$\ln \frac{\mathrm{P}(G=1|X=x)}{\mathrm{P}(G=0|X=x)} = \beta x$$

with a real-valued input variable $X$ used to classify the Bernoulli output variable $G$, and suppose that there are $N$ independent observations

$$(x_1, g_1), \ldots, (x_N, g_N)$$

that can be used to fit the model.

(a) Derive an expression for the log-likelihood function

$$\ell(\beta) = \sum_{i=1}^{N} \ln \mathrm{P}(G = g_i | X = x_i).$$

(b) Show that

$$\frac{d\ell}{d\beta} = \sum_{i=1}^{N} x_i \left( g_i - \frac{e^{\beta x_i}}{1 + e^{\beta x_i}} \right).$$

(c) Consider an experiment which attempts to use a single input variable (temperature) to model the probability that a response is a success; assume it is known that 50% of the responses are expected to be successes when the input variable is set at 0°C so that the no intercept model is appropriate. Suppose that repeated trials are performed at two different temperatures $-1$°C and 1°C, and it is seen that there are $n_1$ failures and $n_2$ successes at $-1$°C while there are $n_3$ failures and $n_4$ successes at 1°C; that is, $N = n_1 + n_2 + n_3 + n_4$ observations for this experiment are

$$\underbrace{(-1,0),\ldots,(-1,0)}_{n_1}, \underbrace{(-1,1),\ldots,(-1,1)}_{n_2}, \underbrace{(1,0),\ldots,(1,0)}_{n_3}, \underbrace{(1,1),\ldots,(1,1)}_{n_4}.$$

Compute the maximum likelihood estimate of $\beta$ for a logistic regression model with no intercept based on this data.

**B3.** Suppose $y_i = f(x_i) + \epsilon_i$ for $i = 1, \ldots, N$ where the $x_i$'s are fixed and distinct and $\epsilon_1, \ldots, \epsilon_N$ are i.i.d. with $\mathrm{E}[\epsilon_i] = 0$ and $var[\epsilon_i] = \sigma^2$. Consider the kernel-weighted average

$$\hat{f}_\lambda(x) = \frac{\sum_{i=1}^N K_\lambda(x, x_i) y_i}{\sum_{i=1}^N K_\lambda(x, x_i)}$$

with the Gaussian kernel

$$K_\lambda(x, x_i) = \frac{1}{\sqrt{2\pi}\lambda} e^{-\frac{1}{2}\left(\frac{x - x_i}{\lambda}\right)^2}.$$

Suppose we want to estimate the response at the first observation time $x_1$.
(a) Compute $\lim_{\lambda \to 0^+} \hat{f}_\lambda(x_1)$ and $\lim_{\lambda \to \infty} \hat{f}_\lambda(x_1)$.
(b) Find the expected value of $\hat{f}_\lambda(x_1)$, and compute $\lim_{\lambda \to 0^+} \mathrm{E}[\hat{f}_\lambda(x_1)]$ and $\lim_{\lambda \to \infty} \mathrm{E}[\hat{f}_\lambda(x_1)]$.
(c) Find the variance of $\hat{f}_\lambda(x_1)$, and compute $\lim_{\lambda \to 0^+} var[\hat{f}_\lambda(x_1)]$ and $\lim_{\lambda \to \infty} var[\hat{f}_\lambda(x_1)]$.

**B4.** Consider the following univariate data: 0, 3, 7, 8, 13. Use the squared Euclidean distance $d(x, y) = (x - y)^2$ for all parts.
(a) Apply the $K$-means algorithm to obtain $K = 2$ clusters by beginning with cluster A centered at 0 and cluster B centered at 3. Compute the within cluster scatter for the final configuration.
(b) Draw the dendogram for the single linkage agglomerative clustering strategy. How would this strategy partition the data into two groups?
(c) Draw the dendogram for the complete linkage agglomerative clustering strategy. How would this strategy partition the data into two groups?

**FORMULAS:**

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

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

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

**TABLES:**

1. The $100\alpha$th percentage point of the central $t$-distribution with *df* degrees of freedom.

2. The $100\alpha$th percentage point of the central $\chi^2$-distribution with *df* degrees of freedom.

3. Upper $\alpha$ probability points of the central $F$-distribution with $n_1$ d.f. in the numerator and $n_2$ d.f. in the denominator.

**Applied Statistics Exam**
**August 2008**
**Time: 12:00pm–3:30pm**

This exam consists of 2 parts. You must answer 5 questions total and must answer at least 2 questions from each part. Make sure to clearly indicate which problems you are attempting. Some formulas and tables are given at the end of this exam.

PART A:

**1.** Consider the simple linear regression model in which $\boldsymbol{y} \sim \text{Normal}_N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$ where $\boldsymbol{X} = [\boldsymbol{J} \vdots \boldsymbol{x}]$ and $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$; here $\beta_0$ is the unknown fixed intercept parameter, $\beta_1$ is the unknown fixed slope parameter, $\sigma^2$ is the unknown fixed variance, $\boldsymbol{J}$ is a $N$-dimensional vector of ones, and $\boldsymbol{x}$ is a known non-random $N$-dimensional column vector.

Assume that we have $N = 10$ and the following summary statistics:

$$\boldsymbol{X}^\top\boldsymbol{X} = \begin{bmatrix} 10 & 4 \\ 4 & 8 \end{bmatrix}, \ \boldsymbol{X}^\top\boldsymbol{y} = \begin{bmatrix} 13 \\ 13 \end{bmatrix}, \ \boldsymbol{y}^\top\boldsymbol{y} = 33.$$

(a) Find the maximum likelihood estimates of $\boldsymbol{\beta}$ and $\sigma^2$.
(b) Test $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$ level $\alpha = .05$.
(c) Suppose we wish to test $H_0 : \beta_0 = \beta_1$ vs. $H_A : \beta_0 \neq \beta_1$. State the null hypothesis in the matrix form $H_0 : \boldsymbol{K}^\top\boldsymbol{\beta} = \boldsymbol{m}$. In particular, what are $\boldsymbol{K}$ and $\boldsymbol{m}$?
(d) Assuming that $\beta_0 = \beta_1$, find the constrained maximum likelihood estimate of $\boldsymbol{\beta}$.
(e) Test $H_0 : \beta_0 = \beta_1$ vs. $H_A : \beta_0 \neq \beta_1$ at level $\alpha = .05$.

**2.** Suppose that $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$ where $\boldsymbol{y}$ is a $N$-dimensional column vector of outputs, $\boldsymbol{X}$ is a $N \times p$ design matrix of fixed inputs, $\beta$ is a $p$-dimensional column vector of coefficients, and $\boldsymbol{\epsilon}$ is a $N$-dimensional column vector of random errors such that $\text{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}_N$ and $var[\boldsymbol{\epsilon}] = \sigma^2\boldsymbol{I}$
(a) Let $a$ be a $p$-dimensional column vector of constants. Suppose that we want to estimate $a^\top\beta$ by a linear unbiased estimator $\boldsymbol{c}^\top\boldsymbol{y}$. Compute $\text{E}[\boldsymbol{c}^\top\boldsymbol{y}]$ and show that $a^\top = \boldsymbol{c}^\top\boldsymbol{X}$.
(b) Let $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$. Show that $cov[\boldsymbol{c}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}, \boldsymbol{c}^\top\boldsymbol{H}\boldsymbol{y}] = 0$.
(c) Let $\hat{\beta} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ denote the least squares estimator of $\beta$. Show that $var[\boldsymbol{c}^\top\boldsymbol{y}] \geq var[a^\top\hat{\beta}]$.

**3.** Consider the following summary tables for various simple linear regression models involving $y$, $x_1$, and $x_2$. For $i = 1, \ldots, N$, let $\tilde{y}_i$ be the fitted value of $y$ at $x_1 = x_{i1}$ based on regressing $y$ on $x_1$ and $\tilde{x}_{2i}$ be the fitted value of $x_2$ at $x_1 = x_{i1}$ based on regressing $x_2$ on $x_1$ where $N$ is the number of observations in the data set.

**Summary table for the regression of $y$ on $x_1$.**

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

**Summary table for the regression of $x_2$ on $x_1$.**

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

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

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

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

**4.** Consider a random effects model

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

where the $e_{ijk}$'s are independent and each follows a Normal distribution with mean 0 and variance $\sigma^2$, the $p_i$'s are independent and each follows a Normal distribution with mean 0 and variance $\sigma_a^2$, the $(mp)_{ij}$'s are also independent and each follows a Normal distribution with mean 0 and variance $\sigma_{mp}^2$, and $\mu_i$ is the mean response for the $i$th level of the fixed factor. Also, all $p_i$'s, $(mp)_{ij}$'s, and $e_{ijk}$'s are independent. Calculate the following quantities.

(a) $\mathrm{E}[Y_{ijk}]$
(b) $var[Y_{ijk}]$
(c) $cov[Y_{ijk}, Y_{ijk'}]$ for $k \neq k'$
(d) $cov[Y_{ijk}, Y_{ij'k}]$ for $j \neq j'$
(e) $cov[Y_{ijk}, Y_{i'jk}]$ for $i \neq i'$.

PART B:

**5.** For fixed $\lambda > 0$ and fixed $x_i \in \mathbb{R}$, let

$$Q_2(\beta) = \sum_{i=1}^{N} (y_i - \beta x_i)^2 + \lambda \beta^2$$

where $y_1, \ldots, y_N$ are realizations of random variables $Y_1, \ldots, Y_N$, respectively.
(a) Find the value of $\beta$ which minimizes $Q_2$.
(b) Suppose that $Y_1, \ldots, Y_N$ are independent and $Y_i \sim \text{Normal}(\beta x_i, \sigma^2)$. Find the bias and the variance of the estimator proposed in (a).

**6.** Consider linear discriminant analysis (LDA) with a single input variable. That is, suppose that the conditional density of $X$ given $G = g$ is Normal with mean $\mu_g$ and variance $\sigma^2$ and that $P(G = g) = \pi_g$ for $g = 1, \ldots, K$. Given a new input $x$, LDA obtains the estimate for the corresponding output by finding the value of $g$ which maximizes $P(G = g|X = x)$. Show that the decision boundary between groups $k$ and $\ell$ has the form

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

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

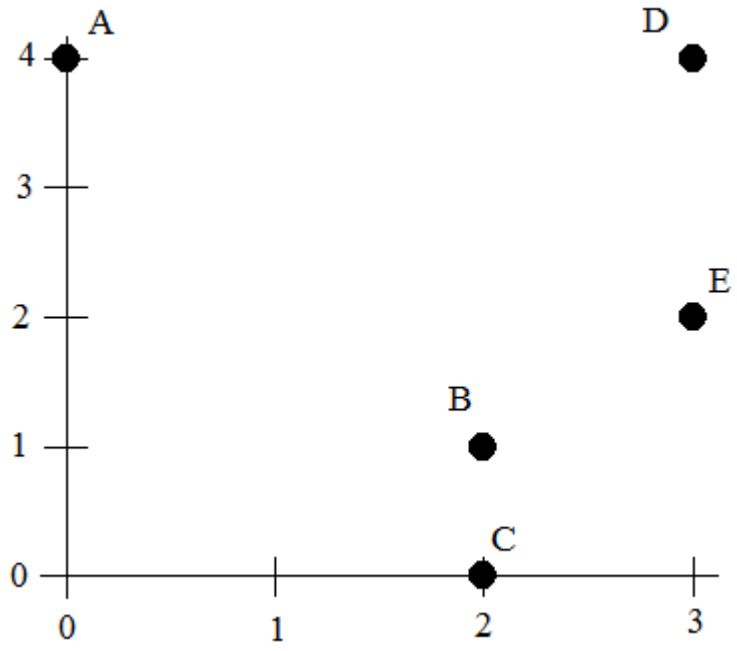
The density of a normal distribution with mean $\mu$ and variance $\sigma^2$ is

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

**7.** Suppose that we have a data set with 8 observations of two inputs $x_1$ and $x_2$ and a Bernoulli output $y$. Use a classification tree with 2 binary splits based on minimizing the misclassification error to model the data.

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

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




(a) Apply the *K*-means algorithm to obtain 2 clusters by beginning with Cluster 1 centered at point B and Cluster 2 center at point C.

(b) Draw the dendogram for the single linkage agglomerative clustering strategy. How would this strategy partition the data into two groups?

(c) Draw the dendogram for the complete linkage agglomerative clustering strategy. How would this strategy partition the data into two groups?

**FORMULAS:**

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

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

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

**TABLES:**

1. The $100\alpha$th percentage point of the central $t$-distribution with *df* degrees of freedom.

2. The $100\alpha$th percentage point of the central $\chi^2$-distribution with *df* degrees of freedom.

3. Upper $\alpha$ probability points of the central $F$-distribution with $n_1$ d.f. in the numerator and $n_2$ d.f. in the denominator.

APPLIED STATISTICS EXAM - MAY 2006

You must answer 5 questions total (each are 20 points). Make sure to clearly indicate which questions you are attempting. Some formulas and tables are given on the last two pages of this exam.

1. Two trucks are weighed at a station. The first truck is measured to weigh 8,000 pounds. The second truck is measured to weigh 10,000 pounds. A third measurement is taken of both trucks together and the recorded value is 17,000 pounds.

   (a) Estimate the weight of each truck by the method of least squares.

   (b) Assuming independent identically distributed normal additive errors with mean zero, test whether these two trucks have the same weight (at a 5% level of significance).

   (c) Suppose the third measurement is $x$ instead of 17,000. For what values of $x$ will the null hypothesis that the two trucks have the same weight be rejected at a 5% level of significance?

2. Suppose

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

   where $\boldsymbol{X}$ is an $n \times p$ known design matrix, $\boldsymbol{\beta}$ is a $p$-dimensional vector of unknown coefficients, and $\boldsymbol{\epsilon}$ follows a $p$-variate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$. Let $\widehat{\boldsymbol{\beta}}$ be the maximum likelihood estimator of $\boldsymbol{\beta}$ so that the vector of predicted values is

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}},$$

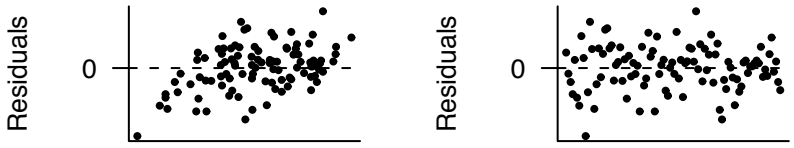
   and define the residual vector

$$\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}}.$$

   (a) What is the distribution of $\boldsymbol{e}$? What is the mean vector, $\mathrm{E}[\boldsymbol{e}]$, and covariance matrix, $\mathrm{var}[\boldsymbol{e}]$, for the residuals? Simplify as much as possible.

   (b) Compute the cross-covariance matrices $\mathrm{cov}[\boldsymbol{e}, \boldsymbol{y}]$ and $\mathrm{cov}[\boldsymbol{e}, \widehat{\boldsymbol{y}}]$. Simplify as much as possible.

(c) One hundred values of $\boldsymbol{y}$ are simulated based on a particular $\boldsymbol{X}$, $\boldsymbol{\beta}$, and $\sigma^2$. The model is fitted and the residuals are computed. One of the two figures below represents a plot of the realizations of the components of $\boldsymbol{y}$ versus the residuals. The other represents the fitted values of $\boldsymbol{y}$ versus the residuals. Which one is which?




3. A statistician uses a linear model to predict a variable $y$ based on 6 regressors $x_1, \ldots, x_6$. The statistician claims that $x_5$ and $x_6$ are significant because the multiple correlation coefficient

$$R^2 = 1 - \frac{\text{residual SS}}{\text{total SS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

rises from 0.4 when only $x_1, \ldots, x_4$ are used to 0.8 when all 6 regressors are used.

(a) Denote the total sum of squares as $S$. Express the residual sum of squares in terms of $S$ under the reduced model using only $x_1, \ldots, x_4$ and under the full model using $x_1, \ldots, x_6$.

(b) What sample sizes guarantee that this statistician's statement is correct at a 5% level of significance?

4. Given

$$Y_{ij} = j\beta_i + \epsilon_{ij}, \qquad i = 1, 2; \quad j = 1, 2,$$

where

$$\epsilon_{ij} \sim \text{independent Normal}(0, \sigma^2),$$

show that the $F$-statistic for testing $H_0 : \beta_1 = \beta_2$ can be expressed in the form

$$F = C\left(\frac{\widehat{\beta}_1 - \widehat{\beta}_2}{\widehat{\sigma}}\right)^2$$

where $C$ is a constant and $\widehat{\beta}_1, \widehat{\beta}_2$, and $\widehat{\sigma}^2$ are the MLEs of $\beta_1, \beta_2$, and $\sigma^2$, respectively. Then state the value of the constant $C$.

You do **NOT** need to explicitly state $\widehat{\beta}_1, \widehat{\beta}_2$, and $\widehat{\sigma}$ in terms of $Y_{11}, Y_{12}, Y_{21}$ and $Y_{22}$.

5. Suppose we have the following results.

| Result | Regression | Slope estimate |
|---|---|---|
| 1 | $Y$ on (1), $X_1$ | 0.5 |
| 2 | $Y$ on (1), $X_2$ | 3 |
| 3 | $X_2$ on (1), $X_1$ | 0.5 |
| 4 | $X_1$ on (1), $X_2$ | 1.5 |

Compute the least squares estimates of $\beta_1$ and $\beta_2$ in the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i, \qquad i = 1, \ldots, n.$$

(Hint: Consider regressing the residuals from result 1 on the residuals from result 3 and regressing the residuals from result 2 on the residuals from result 4.)

6. Suppose we wish to classify an observation into one of two groups ($G = 1$ or $G = 2$) based on a $p$-dimensional random input vector $\boldsymbol{X}$ by direct estimation of $\Pr(G = 1 \mid \boldsymbol{X} = \boldsymbol{x})$.

   (a) Suppose that $\boldsymbol{X}$ has density $f_g(\boldsymbol{x})$ if the observation belongs to group $g$ for $g = 1, 2$, and that $\rho$ is the prior probability of group $g$ if there is no information about the input $\boldsymbol{x}$. Compute the posterior probability $\Pr(G = 1 \mid \boldsymbol{X} = \boldsymbol{x})$ in terms of $f_1(\boldsymbol{x})$, $f_2(\boldsymbol{x})$, and $\rho$.

   (b) In linear discriminant analysis, it is assumed that the densities are multivariate normal with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ for groups $G = 1$ and $G = 2$ respectively, and a common covariance matrix $\boldsymbol{\Sigma}$; that is,

   $$f_g(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_g)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_g)\right\}, \; g = 1, 2.$$

   (For our purposes, take $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ to be known; in practice, they would need to be estimated.) Based on this assumption,

compute the log-ratio

$$\ln \frac{\Pr(G=1|\boldsymbol{X}=\boldsymbol{x})}{\Pr(G=2|\boldsymbol{X}=\boldsymbol{x})}.$$

Show that this ratio is linear in $\boldsymbol{x}$. This proves that the decision boundary based on the rule $\widehat{G}(\boldsymbol{x})=1$ when $\Pr(G=1|\boldsymbol{X}=\boldsymbol{x})>0.5$ is linear.

7. Suppose we have a single input $X=x$ from either group $G=0$ or group $G=1$ and we consider the logistic model

$$\Pr(G=1|X=x)=\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}.$$

Given data $(x_1,g_1),\ldots,(x_N,g_N)$ where $g_1,\ldots,g_N\in\{0,1\}$, the log-likelihood function for the $N$ observations is given by

$$\ell(\beta_0,\beta_1)=\sum_{i=1}^{N}\ln\Pr(G=g_i|X=x_i).$$

(a) Show that

$$\frac{\partial\ell}{\partial\beta_0}(\beta_0,\beta_1)=\sum_{i=1}^{N}\left(g_i-\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)$$

and

$$\frac{\partial\ell}{\partial\beta_1}(\beta_0,\beta_1)=\sum_{i=1}^{N}x_i\left(g_i-\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right).$$

For (b)-(d), suppose we observe the following four observations:

$$(1,0),(2,1),(2,0),(3,1).$$

(b) Compute $\ell(\beta_0,\beta_1)$.

(c) Maximize $f(x)=x-2\ln(1+e^x)$. Show that $\ell(\beta_0,\beta_1)<-2\ln 2$ for all $\beta_0,\beta_1\in\mathbb{R}$.

(d) Compute

$$\lim_{t\to\infty} \frac{\partial \ell}{\partial \beta_0}(-2t, t) \text{ and } \lim_{t\to\infty} \frac{\partial \ell}{\partial \beta_1}(-2t, t).$$

What does this imply about the maximum likelihood estimates of $\beta_0$ and $\beta_1$? What are the fitted probabilities that $G = 1$ given $X = x$ for $x = 1, 2, 3$?

8. Consider the following data: $(-3, 0), (-2, 1), (-1, 0), (0, 0), (1, 1), (5, 1)$.

(a) Apply the $K$-means algorithm to obtain 2 clusters by beginning with cluster A centered at $(-3, 0)$ and cluster B centered at $(5, 1)$.

(b) Apply the $K$-means algorithm to obtain 2 clusters by beginning with cluster A centered at $(0, 0)$ and cluster B centered at $(5, 1)$.

(c) Which initial arrangement gives a better result?

**FORMULAS:**

The Normal($\mu$,$\sigma^2$) density is

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

Suppose $\boldsymbol{y} \sim \text{Normal}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ and $\boldsymbol{X}'\boldsymbol{X}$ is invertible. Denote the MLE of $\boldsymbol{\beta}$ as $\widehat{\boldsymbol{\beta}}$ and the MLE of $\sigma^2$ as $\widehat{\sigma}^2$. Then we have

- $\widehat{\boldsymbol{\beta}} \sim \text{Normal}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$
- $N\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{N-p}$
- $\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2/\sigma^2 \sim \chi^2_p$
- $\dfrac{\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2/p}{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/(N-p)} \sim f_{p,N-p}$

Suppose, in addition, the true value of $\boldsymbol{\beta}$ satisfies $\boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{m}$ and denote the (restricted) MLE of $\boldsymbol{\beta}$ at $\widehat{\boldsymbol{\beta}}_0$. Then we have

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

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

**TABLES:**

The tables below gives the 95th and 97.5th percentile of the $\chi^2$ distribution with $df$ degrees of freedom.

| | | 95% | 97.5% |
|---|---|---|---|
| | 1 | 3.841 | 5.024 |
| | 2 | 5.991 | 7.378 |
| $df$ | 3 | 7.815 | 9.348 |
| | 4 | 9.488 | 11.143 |
| | 5 | 11.071 | 12.833 |

The tables below gives the 95th and 97.5th percentile of the $F$ distribution with $df1$ and $df2$ degrees of freedom.

| | | 95% $df1$ | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| | 1 | 161.447 | 199.500 | 215.707 | 224.583 |
| | 2 | 18.513 | 19.000 | 19.164 | 19.247 |
| $df2$ | 3 | 10.128 | 9.552 | 9.277 | 9.117 |
| | 4 | 7.709 | 6.944 | 6.591 | 6.388 |
| | 5 | 6.608 | 5.786 | 5.409 | 5.192 |
| | 6 | 5.987 | 5.143 | 4.757 | 4.534 |
| | 7 | 5.591 | 4.737 | 4.347 | 4.120 |
| | 8 | 5.318 | 4.459 | 4.066 | 3.838 |
| | 9 | 5.117 | 4.256 | 3.863 | 3.633 |
| | 10 | 4.965 | 4.103 | 3.708 | 3.478 |

| | | 97.5% $df1$ | | | |
|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 |
| | 1 | 647.789 | 799.500 | 864.163 | 899.583 |
| | 2 | 38.506 | 39.000 | 39.165 | 39.248 |
| $df2$ | 3 | 17.443 | 16.044 | 15.439 | 15.101 |
| | 4 | 12.218 | 10.649 | 9.979 | 9.605 |
| | 5 | 10.007 | 8.434 | 7.764 | 7.388 |
| | 6 | 8.813 | 7.260 | 6.599 | 6.227 |
| | 7 | 8.073 | 6.542 | 5.890 | 5.523 |
| | 8 | 7.571 | 6.059 | 5.416 | 5.053 |
| | 9 | 7.209 | 5.715 | 5.078 | 4.718 |
| | 10 | 6.937 | 5.456 | 4.826 | 4.468 |

**Applied Statistics Qualifier Exam**

**This exam consists of 2 parts. You must answer 5 questions total and must answer at least 2 questions from each part. Maek sure to clearly indicate which problems you are attempting.**

PART A

A1. Consider a random effect model

$$E(Y_{ij}|a_i) = \mu + a_i$$

where $Y_{ij}|a_i$s are independent and each follows Normal distribution with mean $\mu + a_i$ and variance $\sigma^2$. Also, $a_i$s are independent and each follows Normal distribution with mean 0 and variance $\sigma_a^2$. Calculate $Cov(Y_{ij}, Y_{ik})$ when $j \neq k$. Also, calculate $Var(Y_{ij})$.

A2. Consider a Beta-Binomial model

$$E(Y_{ij}|p_i) = p_i$$

where $Y_{ij}|p_i$s are independent and each follows Bernoulli distribution with success probability $p_i$. Also, $p_i$s are independent and each follows Beta distribution with parameters $(\alpha, \beta)$. Calculate $\rho = Corr(Y_{ij}, Y_{ik})$ when $j \neq k$.

A3. Show that $E[Y|X]$ is the minimum mean square error predictor of $Y$. That is, show that $g(X) = E[Y|X]$ minimizes $E[(Y - g(X))^2]$ among all functions $g(\cdot)$ of $X$.

A4. Show that

$$\frac{1}{6}\begin{bmatrix} 1 & 16 & 9 & -6 \\ -1 & -14 & -9 & 6 \\ -1 & -16 & -6 & 6 \\ -1 & -16 & -9 & 12 \end{bmatrix}$$

is a generalized inverse of $\boldsymbol{X}^\top \boldsymbol{X}$ where

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

## PART B

B1. Consider the model

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

for $i = 1, \ldots, n$ where the $\epsilon_i$'s are independent and identically distributed normal random variables with mean 0 and unknown constant variance $\sigma^2 > 0$. Now suppose we observe the following data:

| $x$ | $-2$ | $-1$ | $0$ | $1$ |
|---|---|---|---|---|
| $y$ | 6 | 1 | 1 | 5 |

(a) Find the maximum likelihood estimator of $\boldsymbol{\beta} = (\alpha, \beta_1, \beta_2)'$.

(b) Test the hypothesis that $\mathrm{H}_0 : \beta_2 = 0$ at a 5% level of significance using the likelihood ratio test. The tables below give the 95th and 97.5th percentile of the $F$ distribution with *df1* and *df2* degrees of freedom. (Hint: For testing $\mathrm{H}_0 : \boldsymbol{K}'\boldsymbol{\beta} = \boldsymbol{m}$, the likelihood-based $F$-statistic has the forms

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

where the (restricted) MLE of $\boldsymbol{\beta}$ is denoted as $\widehat{\boldsymbol{\beta}}_0$, reduced SS $= \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_0\|^2$ and full SS $= \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2$.

95%

| | *df*1 | | | |
|---|---|---|---|---|
| *df*2 | 1 | 2 | 3 | 4 |
| 1 | 161.447 | 199.500 | 215.707 | 224.583 |
| 2 | 18.513 | 19.000 | 19.164 | 19.247 |
| 3 | 10.128 | 9.552 | 9.277 | 9.117 |
| 4 | 7.709 | 6.944 | 6.591 | 6.388 |

97.5%

| | *df*1 | | | |
|---|---|---|---|---|
| *df*2 | 1 | 2 | 3 | 4 |
| 1 | 647.789 | 799.500 | 864.163 | 899.583 |
| 2 | 38.506 | 39.000 | 39.165 | 39.248 |
| 3 | 17.443 | 16.044 | 15.439 | 15.101 |
| 4 | 12.218 | 10.649 | 9.979 | 9.605 |

B2. Suppose we have an input variable $X$ that we wish to use for classification. The output variable $Y$ is a Bernoulli random variable that follows the no-intercept logistic regression model

$$\ln \frac{\Pr(Y=1|X=x)}{\Pr(Y=0|X=x)} = \beta x.$$

Given training data $(x_1, y_1), \ldots, (x_N, y_N)$ where $y_1, \ldots, y_N \in \{0, 1\}$, the log-likelihood function for the $N$ observations is given by

$$\ell(\beta) = \sum_{i=1}^{N} \ln \Pr(Y = y_i | X = x_i).$$

(a) Show that

$$\frac{d\ell}{d\beta} = \sum_{i=1}^{N} x_i \left( y_i - \frac{e^{\beta x_i}}{1 + e^{\beta x_i}} \right).$$

Now suppose we observe the following 3 observations: $(-1, 1), (0, 0), (1, 1)$.

(b) Find the maximum likelihood estimate of $\beta$.

(c) Show that the solution of (b) is the unique maximizer of $\ell$.

B3. Suppose we observe $N$ data points $(x_i, y_i), i = 1, \ldots, N$, and we want to minimize the penalized residual sum of squares

$$V(\beta) = \sum_{i=1}^{N} (y_i - \beta x_i)^2 + \lambda \beta^2.$$

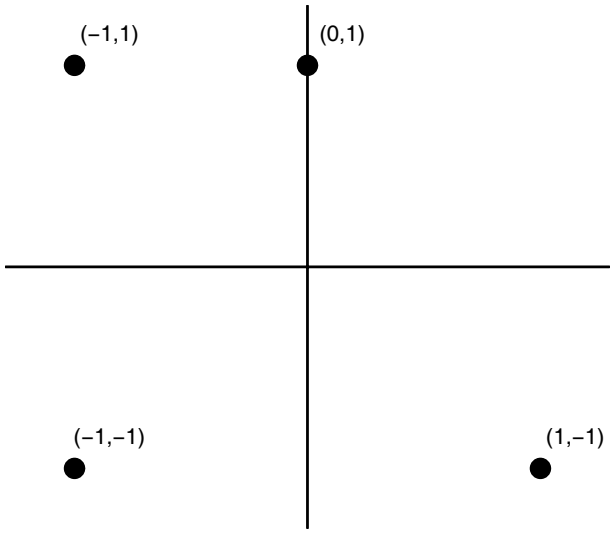
where $\beta$ is an unknown constant and $\lambda > 0$ is a penalty parameter under this no intercept model.

(a) If $\lambda$ is known, what value of $\beta$ minimizes $V$? Justify that your solution is a minimizer.

(b) Denote the estimator of $\beta$ obtained in part (a) as $\widehat{\beta}(\lambda)$. Under the no-intercept regression model $Y_i = \beta x_i + \epsilon_i$ for $i = 1, \ldots, N$ where the $x_i$'s are known non-random inputs, and the $\epsilon_i$'s are independent and identically distributed random variables with mean 0 and constant variance $\sigma^2 > 0$,

1. what is the bias and variance of $\widehat{\beta}(\lambda)$?
2. find the value of $\lambda$ for which $\widehat{\beta}(\lambda)$ has the smallest mean squared error $\mathrm{E}[(\widehat{\beta}(\lambda) - \beta)^2]$? Does this depend on $\beta$? $\sigma^2$?

B4. Consider the data shown below:




(a) Apply the $K$-means algorithm to obtain 2 clusters by beginning with cluster A centered at $(-1, 1)$ and cluster B centered at $(0, 1)$.

(b) Apply the $K$-means algorithm to obtain 2 clusters by beginning with cluster A centered at $(-1, 1)$ and cluster B centered at $(-1, -1)$.

(c) Apply the $K$-means algorithm to obtain 2 clusters by beginning with cluster A centered at $(0, 0)$ and cluster B centered at $(1, -1)$.

(d) Which initial arrangement gives the best result?

# Applied Statistics Qualifying Exam

May 24, 2004
12:00pm-3:30pm

This exam consists of three parts. You must answer two questions from PART A, two questions from PART B, and one question from PART C. Make sure to clearly indicate which problems you are attempting. Begin each problem on a new sheet of paper, and write on only one side of the paper.

## **PART A.** You must answer two of these three questions.

**A1.** Consider the following experimental design:

Objective: This was a randomized, double-blind, crossover study of 30 children with attention-deficit/hyperactivity disorder (ADHD) that evaluated the time course effects of four doses of Adderall (5, 10, 15, and 20 mg), an inactive control (placebo), and a positive control (clinical dose of methylphenidate).
Method: For each treatment condition, a capsule was administered in the morning and assessments were performed in an analog classroom setting every 1.5 hours across the day. Subjective (teacher ratings of deportment and attention) and objective (scores on math tests) measures were obtained for each classroom session, and these measures were used to evaluate time-response and dose-response effects of Adderall.
Results: For doses of Adderall greater than 5 mg, significant time course effects were observed. Rapid improvements on teacher ratings and math performance were observed by 1.5 hours after administration, and these effects dissipated by the end of the day. The specific pattern of time course effects depended on dose: the time of peak effects and the duration of action increased with dose of Adderall.

| Factor/$df_n$, $df_d$ | SKAMP Attention | SKAMP Deportment | PERMP: No. Attempted | PERMP: No. Correct | SSE Item Average |
|---|---|---|---|---|---|
| Time/5,140 | 28.7/.001 | 31.2/.001 | 23.2/.001 | 30.6/.001 | 0.80/NS |
| Dose/4,112 | 6.8/.001 | 31.0/.001 | 7.7/.001 | 7.2/.001 | 1.07/NS |
| T×D/20,560 | 2.1/.005 | 6.4/.001 | 3.9/.001 | 4.6/.001 | 0.96/NS |

*Note:* SKAMP = teacher rating scale developed by Swanson, Kotkin, Agler, M-Flynn, Pelham; PERMP = permanent product measure from math questions; SSE = Stimulant Side Effects rating scale; T×D = time by dose interaction; NS = non-significant.

Discuss the ANOVA design used for this experiment. Explain the fault in the design, and give a correct alternative.

**A2.** Consider the fixed effects model

$$Y_{ij} \sim \text{independent Normal}(\mu_i, \sigma^2),$$

$j = 1, \ldots, n_i$ and $i = 1, \ldots, m$ where $\mu_i, i = 1, \ldots, m$, and $\sigma^2$ are unknown constants.

(a) Using the fact that a Normal$(\mu, \sigma^2)$ density has the form

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

write out the log-likelihood function

$$\ell(\mu_1, \ldots, \mu_M, \sigma^2) = \ln f(\mathbf{y}|\mu_1, \ldots, \mu_m, \sigma^2)$$

where $\mathbf{y} = (y_{11}, \ldots, y_{1n_1}, \cdots, y_{m1}, \ldots, y_{mn_m})'$ and $f$ is the joint density of $\mathbf{y}$.

(b) Using your answer for part (a), find the maximum likelihood estimators of $\mu_1, \ldots, \mu_m$, and $\sigma^2$. Justify that this estimator is a maximizer. Also, give the maximum likelihood estimator of $\mu_1/\sigma$.

(c) Consider the $k$th group $Y_{k1}, \ldots, Y_{kn_k}$. It can be shown that

$$\frac{\bar{Y}_{k\cdot} - \mu_k}{S/\sqrt{n_k}} \sim t_{N-m}$$

where $\bar{Y}_{k\cdot}$ is the sample mean of the $k$th group, $S^2$ is an unbiased estimator of $\sigma^2$, $N = \sum_{i=1}^m n_i$, and $t_n$ is the $t$-distribution with $n$ degrees of freedom. Find a 95% confidence interval for $\mu_k$, letting $t_{\alpha,n}$ denote the $100(1-\alpha)$th percentile of the $t$-distribution with $n$ degrees of freedom.

**A3.** For testing $\mathrm{H}_0$: $\mathbf{K}'\boldsymbol{\beta} = \mathbf{m}$, the likelihood-based $F$-statistic has the forms

$$F = \frac{\|\mathbf{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_0)\|^2/q}{\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2/(N-p)} = \frac{(\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m})'(\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K})^{-1}(\mathbf{K}'\widehat{\boldsymbol{\beta}} - \mathbf{m})/q}{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})/(N-p)}.$$

Consider the quadratic model

$$\mathbf{E}[Y_i] = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

for $i = 1, \ldots, N$ where $Y_i$ is a normal random variable with unknown constant variance $\sigma^2$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ is unknown. Now suppose we observe the following four data points $(x_i, y_i)$:

$$(-1, 5), (0, 1), (1, -1), (2, 0).$$

(a) Find the maximum likelihood estimator of $\boldsymbol{\beta}$.

(b) Test the hypothesis that $H_0 : \beta_1 = \beta_2 = 0$ at a .05 level of significance. The tables below gives the 95th and 97.5th percentile of the $F$ distribution with $df1$ and $df2$ degrees of freedom.

95%

| $df2$ \ $df1$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 161.447 | 199.500 | 215.707 | 224.583 |
| 2 | 18.513 | 19.000 | 19.164 | 19.247 |
| 3 | 10.128 | 9.552 | 9.277 | 9.117 |
| 4 | 7.709 | 6.944 | 6.591 | 6.388 |

97.5%

| $df2$ \ $df1$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 647.789 | 799.500 | 864.163 | 899.583 |
| 2 | 38.506 | 39.000 | 39.165 | 39.248 |
| 3 | 17.443 | 16.044 | 15.439 | 15.101 |
| 4 | 12.218 | 10.649 | 9.979 | 9.605 |

## **PART B.** You must answer two of these three questions.

**B1.** Consider the no-intercept regression model

$$Y = \beta x + \epsilon$$

where $x$ is a known real-valued input variable, $\beta$ is an unknown constant, and $\epsilon$ is a random error term with a mean of zero and an unknown constant variance $\sigma^2$.

(a) Suppose we observe $N$ data points $(x_i, y_i), i = 1, \ldots, n$ and we wish to minimize the penalized residual sum of squares

$$V(\alpha, \beta) = \sum_{i=1}^{N} (y_i - \alpha - \beta x_i)^2 + \beta^2.$$

What are the values of $\alpha$ and $\beta$ which minimize $V$?

(b) Denote the respective solutions to (a) as $\widehat{\alpha}$ and $\widehat{\beta}$. Find the bias and variance of $\widehat{\beta}$. What is the bias of $\widehat{\alpha}$?

---

**B2.** Suppose we have a real-valued input variable $X$ that we wish to use for classification. The input vector is either from class 0 or class 1; the output variable is $G = 0$ or $G = 1$ in the two respective cases. Now consider the logistic regression model

$$\ln \frac{\mathrm{P}(G = 1|X = x)}{\mathrm{P}(G = 0|X = x)} = \beta x.$$

Given training data $(x_1, g_1), \ldots, (x_N, g_N)$ where $g_1, \ldots, g_N \in \{0, 1\}$, the log-likelihood function for the $N$ observations is given by

$$\ell(\beta) = \sum_{i=1}^{N} \ln \mathrm{P}(G = g_i|X = x_i).$$

(a) Show that

$$\frac{d\ell}{d\beta} = \sum_{i=1}^{N} x_i \left( g_i - \frac{e^{\beta x_i}}{1 + e^{\beta x_i}} \right).$$

Now suppose we observe the following four observations: $(-1, 1), (0, 0), (1, 1), (2, 0)$.

(b) Evaluate $\left.\frac{d\ell}{d\beta}\right|_{\beta=-1}$ and $\left.\frac{d\ell}{d\beta}\right|_{\beta=0}$.

(c) Show that the solution of $\frac{d\ell}{d\beta} = 0$ is the unique maximizer of $\ell$.

(d) Find a value of $\beta$ which is within 0.1 of the maximum likelihood estimator of $\beta$. Justify your answer.

---

**B3.** There are many different methods to classify data (logistic regression, neural networks, discriminant analysis, etc.). How do you determine the best technique for classifying?

## **PART C.** You must answer one of these two questions.

**C1.** Consider the following experimental design and determine whether it is a reasonable method of clustering. Explain why or why not.

Study Design. A k-means cluster analysis of patients with spinal and radicular pain based on the SF-36 Health Survey scales.
Objective. The aim was to determine whether spine patients fall into clusters according to self-reported health status as measured by the SF-36 and to determine if clustering is similar across four common diagnostic categories: herniated disc, spinal stenosis, spondylosis, and chronic pain syndrome.
The grouping of patients (mean age of 50 years, 50% male) was accomplished by "k-means" cluster analysis based on each patient's scores on the eight scales of the SF-36 Health Survey. In order to reduce bias in the selection of clusters, we standardized scores so that those variables with higher variability and higher absolute values were not disproportionately represented in the solutions. Using the technique suggested by the authors of the SF-36, all scores were standardized relative to the general U.S. population and were scaled to have a mean of 50 and standard deviation of 10. 12 A score of 50 represents the average score (the "norm") for the U.S. population. Any score below 30 (i.e., below two standard deviations from normal) can be interpreted as a significant health deficiency.
In order to conduct such an analysis, a prespecified number of clusters are assigned, and the k-means algorithm attempts to place patients into one of the clusters so as to minimize the total variation among the individual profiles within each of the groups. This is accomplished as follows. Patient profiles are sequentially moved from group to group, and the total variation within each group is measured. If total variation is reduced by a move, then the patient stays in its new group. Otherwise, the move is reversed. The process continues until there is no switch of a patient from one group to another that will reduce the overall within-group variation. In the end, the goal of the analysis is satisfied insofar as patients with similar profiles of scores on the SF-36 reside in the same cluster.
Because the k-means algorithm will produce some kind of solution for any number of prespecified groups, it is critical to determine the most appropriate number of groups (clusters) for a given data set. This is achieved by running the cluster analysis for different numbers of prespecified groups and observing when the benefit of adding additional groups begins to diminish. In general, each time an additional group is added, the solution will better "fit" the data; the perfect fit occurring when the number of "groups" is equal to the number of patients. However, after the addition of a certain number of groups, the benefits (as indicated by a reduction in the within group variability) will be very small, and a large number of groups also makes it difficult to interpret results. In our case, we found that three groups (clusters) yielded a good "fit" but that the addition of a fourth and fifth group only had a marginal impact on the within-group variability. In addition, the groupings arrived at by the analysis appeared each to have reasonable clinical interpretations—an important goal that is more difficult to

achieve with a large number of groups, some of which may not contain many members (patients).

---

**C2.** What information will clustering provide about the data? How can you determine that the result gives good clusters? Explain the pros and cons of k-means clustering versus hierarchical clustering.

**Applied Statistics Qualifier Exam**

**This exam consists of 2 parts. You must answer 5 questions total and must answer at least 2 questions from each part. Make sure to clearly indicate which problems you are attempting. Some formulas and tables are given on the last two pages of this exam.**

PART A.

QUESTION A1: Suppose that $y_i = \alpha + x_i + \epsilon_i,\ i = 1, \ldots, N$ where $\epsilon_1, \ldots, \epsilon_N$ are independent Normal(0,$\sigma^2$). Assuming the data points $(x_i, y_i), i = 1, \ldots, n$, are not collinear, find the maximum likelihood estimators of $\alpha$ and $\sigma^2$ by differentiating the likelihood function. Justify that your solutions maximize the likelihood function.

QUESTION A2: Suppose we have the following model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad i = 1, \ldots, N$$

where $\epsilon_1, \ldots, \epsilon_N$ are independent Normal(0,$\sigma^2$). Four univariate regression models were run and the following results were obtained:

$$\widehat{y} = \frac{27 - 4x_1}{5}, \quad \widehat{y} = \frac{227 - 15x_2}{40}, \quad \widehat{x}_1 = \frac{21 - x_2}{8}, \quad \widehat{x}_2 = \frac{29 - 2x_1}{5}.$$

Find the maximum likelihood estimates of $\beta_0$, $\beta_1$, and $\beta_2$ for the multiple regression model.

QUESTION A3: Suppose

$$y_{ijk} = \mu_i + \alpha_{ij} + \epsilon_{ijk}, \quad i, j, k = 1, 2$$

where $\alpha_{11} + \alpha_{12} = \alpha_{21} + \alpha_{22} = 0$ and $\epsilon_{ijk} \sim$ independent Normal$(0, \sigma^2)$.

Applied Statistics (10/14/04)

Given the data

| | | $i$ | |
|---|---|---|---|
| | $y_{ijk}$ | 1 | 2 |
| $j$ | 1 | 10,12 | 4,3 |
| | 2 | 12,10 | 8,10 |

test the hypothesis $H_0 : \alpha_{11} = \alpha_{12} = \alpha_{21} = \alpha_{22} = 0$ at level $0.05$.

QUESTION A4: Suppose that $y_i = \alpha + \beta x_i + \epsilon_i$ for $i = 1, 2, 3$ where $\epsilon_1, \epsilon_2$, and $\epsilon_3$ are independent Normal(0,$\sigma^2$).
(a) Given the data points

$$(-1, y_1), (0, y_2), \text{ and } (1, y_3)$$

show that the $F$-statistic for testing $H_0 : \alpha = \beta$ can be expressed in the form

$$F = C \left( \frac{\widehat{\alpha} - \widehat{\beta}}{\widehat{\sigma}} \right)^2$$

where $C$ is a constant and $\widehat{\alpha}, \widehat{\beta}$, and $\widehat{\sigma}^2$ are the maximum likelihood estimates of $\alpha, \beta$, and $\sigma^2$, respectively. Give the value of $C$.
(b) Given the data points

$$(-1, 2), (0, 4), \text{ and } (1, 12)$$

find a 95% confidence ellipsoid for $(\alpha, \beta)'$. State your answer in the form

$$A_1(\widehat{\alpha} - \alpha)^2 + A_2(\widehat{\beta} - \beta)^2 + A_3(\widehat{\alpha} - \alpha)(\widehat{\beta} - \beta) \leq D,$$

giving the values of $\widehat{\alpha}, \widehat{\beta}, A_1, A_2, A_3$, and $D$.

PART B.

QUESTION B1: Suppose $y_i \sim$ independent Normal$(\beta x_i, \sigma^2)$ for $i = 1, \ldots, N$.
(a) For fixed $\lambda > 0$, find the bias and variance of the two estimators

$$\widehat{\beta} = \frac{\sum_i x_i y_i}{\sum_i x_i^2} \quad , \quad \widetilde{\beta} = \frac{\sum_i x_i y_i}{\sum_i x_i^2 + \lambda}.$$

Which estimator has the larger bias? Which estimator has the larger variance?

(b) For $C > 1$, what choice of $\lambda$ will reduce the variance of $\widehat{\beta}$ by a factor of $C$? In this case, what happens to the bias of $\widetilde{\beta}$?

QUESTION B2: Suppose we have a $m \times 1$ input vector $\boldsymbol{x} = (x_1, \ldots, x_m)'$ which is either from group 0 ($y = 0$) or group 1 ($y = 1$) and we model them according to

$$p_k(\boldsymbol{x}; \boldsymbol{\beta}) = \text{Pr}(y = k \mid \boldsymbol{x}) = \frac{e^{k\boldsymbol{\beta}'\boldsymbol{x}}}{1 + e^{\boldsymbol{\beta}'\boldsymbol{x}}}, \quad k = 0, 1,$$

for an unknown $m \times 1$ coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)'$.
(a) Given training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_n)$, show that the log-likelihood equation can be expressed as

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N} \ln p_{y_i}(\boldsymbol{x}_i; \boldsymbol{\beta}) = \sum_{i=1}^{N} \left\{ y_i \boldsymbol{\beta}'\boldsymbol{x}_i - \ln(1 + e^{\boldsymbol{\beta}'\boldsymbol{x}_i}) \right\}.$$

(b) Show that the maximum likelihood estimator of $\boldsymbol{\beta}$ must satisfy

$$\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{p}) = \boldsymbol{0}$$

where $\boldsymbol{X}$ is a matrix containing the rows $\boldsymbol{x}_1', \ldots, \boldsymbol{x}_N'$, $\boldsymbol{y}$ is the $N \times 1$ column vector $(y_1, \ldots, y_N)'$, and $\boldsymbol{p}$ is the $N \times 1$ column vector $(p_1(\boldsymbol{x}_1; \boldsymbol{\beta}), \ldots, p_1(\boldsymbol{x}_N; \boldsymbol{\beta}))'$.

QUESTION B3: Consider the data set with eight observations $(x_{1,i}, x_{2,i}, y_i)$:

$$(1, 2, 1), (2, 5, 0), (3, 3, 1), (4, 5, 0), (5, 1, 1), (6, 0, 0), (7, 4, 1), (8, 3, 0).$$

(a) Using a classification tree based on both input variables $\boldsymbol{x}_1 = (x_{1,1}, \ldots, x_{1,8})'$ and $\boldsymbol{x}_2 = (x_{2,1}, \ldots, x_{2,8})'$, find the best first split based on misclassification rate.
(b) After the split in (a) is made, what split should be made next based on misclassification rate? Give all splits which optimize the misclassification rate.

QUESTION B4: Consider a clustering model where each cluster is labeled by an integer 1 through $K \in \mathbb{N}$. The assignments to a cluster are characterized by an encoder $C(i) = k$ which assigns the $i$th observation to cluster $k$. We choose our loss function to be the *within-point scatter*

$$W(C) = \frac{1}{2} \sum_{k=1}^{K} \sum_{C(i)=k} \sum_{C(i')=k} \|\boldsymbol{x}_i - \boldsymbol{x}_{i'}\|^2.$$

(a) Show that, for a given cluster assignment $C$, the $K$-means algorithm minimizes the within-point scatter.
(b) Is convergence of the $K$-means algorithm guaranteed? Why or why not?

Applied Statistics (10/14/04)

**FORMULAS:**

Normal($\mu$,$\sigma^2$) density: $n(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$

If $\boldsymbol{y} \sim$ Normal$(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I}_n)$ and $\boldsymbol{X'X}$ is invertible:

- MLE of $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}} \sim$ Normal$(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X'X})^{-1})$
- MLE of $\sigma^2 = \widehat{\sigma}^2$ and $N\widehat{\sigma}^2/\sigma^2 \sim \chi^2_{N-p}$
- $\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2/\sigma^2 \sim \chi^2_p$
- $$\frac{\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2/p}{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/(N-p)} \sim f_{p,N-p}$$

If, in addition, $\boldsymbol{K'\beta} = \boldsymbol{m}$:

- (restricted) MLE of $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_0$
- $\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_0)\|^2/\sigma^2 \sim \chi^2_q$
- $$\begin{aligned} F &= \frac{\|\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_0)\|^2/q}{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|^2/(N-p)} \\ &= \frac{(\boldsymbol{K'}\widehat{\boldsymbol{\beta}} - \boldsymbol{m})'(\boldsymbol{K'}(\boldsymbol{X'X})^{-1}\boldsymbol{K})^{-1}(\boldsymbol{K'}\widehat{\boldsymbol{\beta}} - \boldsymbol{m})/q}{(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})/(N-p)} \sim f_{q,N-p} \end{aligned}$$

**TABLES:**

The tables below gives the 95th and 97.5th percentile of the $\chi^2$ distribution with $df$ degrees of freedom.

| $df$ | 95% | 97.5% |
|---|---|---|
| 1 | 3.841 | 5.024 |
| 2 | 5.991 | 7.378 |
| 3 | 7.815 | 9.348 |
| 4 | 9.488 | 11.143 |
| 5 | 11.071 | 12.833 |

The tables below gives the 95th and 97.5th percentile of the $F$ distribution with $df1$ and $df2$ degrees of freedom.

95%

| $df2$ \ $df1$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 161.447 | 199.500 | 215.707 | 224.583 |
| 2 | 18.513 | 19.000 | 19.164 | 19.247 |
| 3 | 10.128 | 9.552 | 9.277 | 9.117 |
| 4 | 7.709 | 6.944 | 6.591 | 6.388 |

97.5%

| $df2$ \ $df1$ | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 647.789 | 799.500 | 864.163 | 899.583 |
| 2 | 38.506 | 39.000 | 39.165 | 39.248 |
| 3 | 17.443 | 16.044 | 15.439 | 15.101 |
| 4 | 12.218 | 10.649 | 9.979 | 9.605 |