

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

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

where \mathbf{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 $\hat{\boldsymbol{\beta}}$ be the maximum likelihood estimator of $\boldsymbol{\beta}$ so that the vector of predicted values is

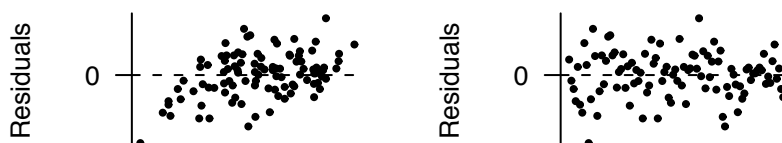
$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}},$$

and define the residual vector

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}.$$

- (a) What is the distribution of \mathbf{e} ? What is the mean vector, $E[\mathbf{e}]$, and covariance matrix, $\text{var}[\mathbf{e}]$, for the residuals? Simplify as much as possible.
- (b) Compute the cross-covariance matrices $\text{cov}[\mathbf{e}, \mathbf{y}]$ and $\text{cov}[\mathbf{e}, \hat{\mathbf{y}}]$. Simplify as much as possible.

- (c) One hundred values of \mathbf{y} are simulated based on a particular \mathbf{X} , $\boldsymbol{\beta}$, and σ^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 \mathbf{y} versus the residuals. The other represents the fitted values of \mathbf{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, \dots, x_6 . The statistician claims that x_5 and x_6 are significant because the multiple correlation coefficient

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

rises from 0.4 when only x_1, \dots, 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, \dots, x_4 and under the full model using x_1, \dots, x_6 .
- (b) What sample sizes guarantee that this statistician's statement is correct at a 5% level of significance?

4. Given

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

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

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

where C is a constant and $\widehat{\beta}_1, \widehat{\beta}_2$, and $\widehat{\sigma}^2$ are the MLEs of β_1, β_2 , and σ^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 β_1 and β_2 in the model

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

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

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

(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

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

Show that this ratio is linear in \mathbf{x} . This proves that the decision boundary based on the rule $\hat{G}(\mathbf{x}) = 1$ when $\Pr(G = 1 | \mathbf{X} = \mathbf{x}) > 0.5$ 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 $(x_1, g_1), \dots, (x_N, g_N)$ where $g_1, \dots, g_N \in \{0, 1\}$, 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

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

What does this imply about the maximum likelihood estimates of β_0 and β_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(μ, σ^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 $\mathbf{y} \sim \text{Normal}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ and $\mathbf{X}'\mathbf{X}$ is invertible. Denote the MLE of $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}}$ and the MLE of σ^2 as $\hat{\sigma}^2$. Then we have

- $\hat{\boldsymbol{\beta}} \sim \text{Normal}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$
- $N\hat{\sigma}^2/\sigma^2 \sim \chi_{N-p}^2$
- $\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2/\sigma^2 \sim \chi_p^2$
- $\frac{\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2/p}{\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/(N-p)} \sim f_{p, N-p}$

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

- $\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_0)\|^2/\sigma^2 \sim \chi_q^2$
-

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

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

TABLES:

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

	95%	97.5%
df 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%			
	$df1$			
	1	2	3	4
$df2$ 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
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
$df2$ 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
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