Wednesday, December 07, 2011

Special Preliminary Examination on Linear Models and Multivariate Statistics

Statistics Group, Department of Mathematics and Statistics, Auburn University

Name: _____

1. It is a closed-book and in-class exam.
2. One page (letter size, 8.5-by-11in) cheat sheet is allowed.
3. Calculator is allowed. No laptop (or equivalent).
4. Show your work to receive full credits. *Highlight your final answer.*
5. Solve any **five** problems out of the six problems.
6. Total points are **50**. Each question is worth **10** points.
7. If you work out more than five problems, your score is the sum of five highest points.
8. Time: **180** minutes. (8:30am – 11:30am, Wednesday, December 07, 2011)

1	2	3	4	5	6	Total

1. Let $X_1, \ldots, X_n$ be a random sample from $N_p(\mu, \Sigma)$ population. Denote $\bar{x}$ and S as the sample mean and the sample covariance matrix, respectively. Consider the hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$, where μ_0 is a given vector.

   (a) Work out the rejection region based on the test statistic
   $$T^2 = n(\bar{x} - \mu_0)^T S^{-1}(\bar{x} - \mu_0).$$

   (b) Work out the rejection region based on the likelihood ratio test.

   (c) Show that the above two tests are equivalent.

2. Describe factor analysis and principal component analysis. What are the differences and similarities between these two methods?

3. Suppose that the covariance ($n \times n$ matrix) given below
   $$S = \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{pmatrix}$$
   Show that if $\rho > 0$ then the first principal component accounts for a proportion $(1 + \rho(n-1))/n$ of the total variation in the data. What can be said about the other $n-1$ components?

   Hint: You may want to use the following result.
   $$|aI_p + bJ_p| = a^{p-1}(a + bp)$$
   where $J_p = 1_p 1_p^T$ and 1_p is a $p \times 1$ vector of ones.

4. Suppose y_1, y_2, y_3 and y_4 are observations of angles θ_1, θ_2, θ_3, and θ_4 of a quadrilateral on the ground. If the observations are subject to independent normal errors with zero means and common variance σ^2. (Note: The sum of the angles of a quadrilateral is 2π.)

   (a) Find the least squares estimates for θ_1, θ_2, θ_3 and θ_4.

   (b) Derive a test statistic for the hypothesis that the quadrilateral is a parallelogram with $\theta_1 = \theta_3$ and $\theta_2 = \theta_4$.

5. In a linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$, show how to find a confidence interval for the ratio $\mathbf{a}_1'\boldsymbol{\beta}/\mathbf{a}_2'\boldsymbol{\beta}$ of two linear parametric functions.

6. Consider a regression model
   $$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i, \quad i = 1, 2, \ldots, n.$$
   Suppose that $\beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$. Find the distribution of R^2 (coefficient of determination) and hence prove that $E(R^2) = (p-1)/(n-1)$.

Solution.

1.

2.

3.

4. We can write the model as

$$\begin{aligned} y_1 &= \theta_1 + \varepsilon_1 \\ y_2 &= \theta_2 + \varepsilon_2 \\ y_3 &= \theta_3 + \varepsilon_3 \\ y_4 - 2\pi &= -\theta_1 - \theta_2 - \theta_3 + \varepsilon_4 \end{aligned}$$

Therefore,

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ 2\pi - y_4 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix}.$$

The LSE of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \frac{1}{4} \begin{pmatrix} 3 & -1 & -1 \\ -1 & 3 & -1 \\ -1 & -1 & 3 \end{pmatrix} \begin{pmatrix} y_1 - y_4 + 2\pi \\ y_2 - y_4 + 2\pi \\ y_3 - y_4 + 2\pi \end{pmatrix} = \begin{pmatrix} (3y_1 - y_2 - y_3 - y_4 + 2\pi)/4 \\ (3y_2 - y_1 - y_3 - y_4 + 2\pi)/4 \\ (3y_3 - y_1 - y_2 - y_4 + 2\pi)/4 \end{pmatrix}.$$

Therefore,

$$\hat{\theta}_4 = (3y_4 - y_1 - y_2 - y_3 + 2\pi)/4.$$

Now consider the hypothesis. If $\theta_1 = \theta_3$ and $\theta_2 = \theta_4$, then the model can be written as

$$\begin{aligned} y_1 &= \theta_1 + \varepsilon_1 \\ y_2 - \pi &= -\theta_1 + \varepsilon_2 \\ y_3 &= \theta_1 + \varepsilon_3 \\ y_4 - \pi &= -\theta_1 + \varepsilon_4 \end{aligned}$$

Therefore,

$$\begin{aligned} \hat{\theta}_1 = \hat{\theta}_3 &= (y_1 - y_2 + y_3 - y_4 + 2\pi)/4 \\ \hat{\theta}_2 = \hat{\theta}_4 &= (y_2 - y_1 + y_4 - y_3 + 2\pi)/4 \end{aligned}$$

Calculate RSS and RSS_H,

$$\begin{aligned} \text{RSS} &= (y_1 + y_2 + y_3 + y_4 - 2\pi)^2/4 \\ \text{RSS}_H - \text{RSS} &= (y_1 - y_3)^2/2 + (y_2 - y_4)^2/2 \end{aligned}$$

Therefore, an F-statistic is

$$F = \frac{(\text{RSS}_H - \text{RSS})/2}{\text{RSS}} = \frac{(y_1 - y_3)^2 + (y_2 - y_4)^2}{(y_1 + y_2 + y_3 + y_4 - 2\pi)^2}.$$

When H is true, $F \sim F_{2,1}$.

5. Denote $\phi = \mathbf{a}_1'\boldsymbol{\beta}/\mathbf{a}_2'\boldsymbol{\beta}$ and consider statistic $\mathbf{a}_1'\hat{\boldsymbol{\beta}} - \phi\mathbf{a}_2'\hat{\boldsymbol{\beta}} = (\mathbf{a}_1 - \phi\mathbf{a}_2)'\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the LSE of $\boldsymbol{\beta}$. This statistic follows a normal distribution with mean and variance,

$$\begin{aligned} E(\mathbf{a}_1'\hat{\boldsymbol{\beta}} - \phi\mathbf{a}_2'\hat{\boldsymbol{\beta}}) &= \mathbf{a}_1'\boldsymbol{\beta} - \phi\mathbf{a}_2'\boldsymbol{\beta} = 0 \\ var(\mathbf{a}_1'\hat{\boldsymbol{\beta}} - \phi\mathbf{a}_2'\hat{\boldsymbol{\beta}}) &= \sigma^2(\mathbf{a}_1 - \phi\mathbf{a}_2)'(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{a}_1 - \phi\mathbf{a}_2), \end{aligned}$$

where $\mathbf{X}$ is the regression matrix. Therefore,

$$\frac{\mathbf{a}_1'\hat{\boldsymbol{\beta}} - \phi\mathbf{a}_2'\hat{\boldsymbol{\beta}}}{S\sqrt{(\mathbf{a}_1 - \phi\mathbf{a}_2)'(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{a}_1 - \phi\mathbf{a}_2)}}$$

follows a t-distribution with $n - p$ degrees of freedom, where $S^2 = \text{RSS}/(n - p)$. From

$$1 - \alpha = P(|\mathbf{a}_1'\hat{\boldsymbol{\beta}} - \phi\mathbf{a}_2'\hat{\boldsymbol{\beta}}| \le t_{\alpha/2,n-p} \cdot S\sqrt{(\mathbf{a}_1 - \phi\mathbf{a}_2)'(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{a}_1 - \phi\mathbf{a}_2)})$$

we know that a confidence interval should satisfy

$$(\mathbf{a}_1'\hat{\boldsymbol{\beta}} - \phi\mathbf{a}_2'\hat{\boldsymbol{\beta}})^2 \le F_{\alpha,1,n-p} \cdot S^2(\mathbf{a}_1 - \phi\mathbf{a}_2)'(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{a}_1 - \phi\mathbf{a}_2).$$

Solve this inequality to obtain an confidence interval.

6. The coefficient of determination is

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2}.$$

Notice that $\sum(y_i - \hat{y}_i)^2 \sim \sigma^2\chi^2_{n-p}$ is independent of $\sum(\hat{y}_i - \bar{y})^2 \sim \sigma^2\chi^2_{p-1}$. Therefore, the distribution of R^2 is $\chi^2_{p-1}/(\chi^2_{n-p} + \chi^2_{p-1})$, which is a Beta$((p-1)/2,\ (n-p)/2)$ distribution. So $E(R^2) = (p-1)/(n-1)$.

Note: χ^2_p is also a Gamma$(p/2,\ 2)$ distribution. If $X \sim$ Gamma$(\alpha_1,\ \beta)$ is independent of $Y \sim$ Gamma$(\alpha_2,\ \beta)$, then $X/(X+Y) \sim$ Beta$(\alpha_1,\ \alpha_2)$.