COMPREHENSIVE WRITTEN EXAMINATION, PAPER III

FRIDAY AUGUST 16, 2013, 9:00 A.M. – 1:00 P.M.

STOR 664 Questions

(1) (40 points) A marketing research group collected sales data from 20 companies randomly selected in Year 1 (x) and studied whether their sales in Year 2 (y) can be predicted from x. A simple linear regression model was fitted to the data as follows:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i; \quad \epsilon_i \text{ i.i.d} \sim N(0, \sigma^2), \qquad i = 1, \ldots, n.$$

The edited R output and other summaries are given below.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.6996     0.7268  -2.338   0.0311 *
   Year 1     0.8399     0.1440   5.831 1.60e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response: Year 2
          Df Sum Sq Mean Sq F value    Pr(>F)
Year 1     1 6.4337  6.4337  33.998 1.597e-05 ***
Residuals 18 3.4063  0.1892
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\sum_{i=1}^{n} y_i = 50, \quad \sum_{i=1}^{n} x_i = 100, \quad \sum_{i=1}^{n} x_i y_i = 257.66, \quad \sum_{i=1}^{n} y_i^2 = 134.84, \quad \sum_{i=1}^{n} x_i^2 = 509.12$$

$$t_{18;0.975} = 2.101, \quad t_{18;0.95} = 1.734, \quad F_{2,18;0.9} = 2.624.$$

(1a) Use Bonferroni and Scheffé methods to obtain 90% simultaneous confidence intervals for β_0 and β_1. Which method do you prefer? Why?

(1b) Calculate a 90% prediction interval for the Year 2 sales for a company with the Year 1 sales 5.0.

(1c) Suppose we want to test $H_0 : y_i = -1 + x_i + \epsilon_i$ (reduced model) versus $H_1 : y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ (full model). Derive a F-test procedure, calculate the observed test statistic and state your conclusion using the significance level $\alpha = 0.1$.

(1d) Consider two pairs of parameter values under the alternative H_1: $(\beta_0, \beta_1) = (-0.6,\ 1.2)$ and $(\beta_0, \beta_1) = (-0.8,\ 1.4)$. Which pair would give a higher power for the test you derived in (1c)? Why?

(2) (15 points) Here are several historical facts over the period 1950 – 1999:

- The correlation between the purchasing power of the US dollar each year and the death rate from lung cancer in the year is -0.95.
- The purchasing power was going steadily down in that period.
- Death rates from lung cancer were generally going up in that period.

Based on those facts, does it make sense to conclude that inflation causes or prevents lung cancer? Explain why.

(3) (15 points) Consider a (multiple) regression model $Y = X\beta + \epsilon$ with an intercept, i.e. the first column of the $n \times p$ design matrix X is $(1, ..., 1)^t$. Let $\widehat{Y} = X\widehat{\beta}$ where $\widehat{\beta}$ is the OLS estimate of β, and the coefficient of determination

$$R^2 = \frac{SSR}{SSTO} = \frac{\sum_{i=1}^n (\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

For testing $H_0 :\ \beta_2 = \cdots = \beta_p = 0$, can you express the test statistic F as a function of n, p and R^2 ? If so, derive the formula; if not, why?

(4) (30 points) Consider the regression model $Y = X\beta + \epsilon$ under the following assumptions:

- the $n \times p$ matrix X is random but has rank p;
- the random error ϵ has $E(\epsilon|X) = u = (u_1, ..., u_n)^t$, which is a non-random vector;
- $Cov(\epsilon|X) = G$, where G is a non-random positive definite matrix.

Let $\widehat{\beta} = (X^tX)^{-1}X^tY$ be the "usual" OLS estimator of β. Answer the following questions and explain:

(4a) Is $E(\widehat{\beta}|X) = \beta$ true ?

(4b) If the first column of X is $(1, ..., 1)^t$, and $u_1 = \cdots = u_n$, is $\widehat{\beta}_1$ unbiased given X?

(4c) Is $\widehat{\beta}_2$ unbiased given X under the same conditions as in (4b)?

(4d) Is $Cov(\widehat{\beta}|X) = \sigma^2(X^tX)^{-1}$ for some constant σ?