COMPREHENSIVE WRITTEN EXAMINATION, PAPER III

FRIDAY AUGUST 15, 2014, 9:00 A.M. – 1:00 P.M.

STOR 664 Questions

[1] (50 points) Consider the regression model

$$Y_{ij} = \alpha_i + \beta\ X_{ij} + \epsilon_{ij}, \quad i = 1, ..., n; \quad j = 1, 2$$

where the component Y_{ij} of the response vector Y represents the jth observation of the ith individual. Assume components of the error vector $\epsilon = \{\epsilon_{ij}\}$ are independent with mean 0 and variance $Var(\epsilon_{ij}) = \sigma_i^2 > 0$; and the explanatory (random) variables $\{X_{ij}\}$ are independent of ϵ. Let $\theta^t = (\alpha_1, ..., \alpha_n; \beta)$ and $\gamma^t = (\sigma_1, ..., \sigma_n)$ be constant parameter vectors.

(1a) Express the design matrix X in the linear model

$$Y = X\theta + \epsilon. \tag{1}$$

(1b) Express the conditional mean vector $\mu = E(Y|X)$ as a function of X and θ.

(1c) Express the conditional covariance matrix $\Gamma = Cov(Y|X)$ as a function of γ.

(1d) Here is the GLS (generalized least square) approach: Suppose there exists a known positive definite matrix G such that $Cov(G^{-1/2}\epsilon) = \sigma^2 I$ for some unknown constant $\sigma > 0$ and the identity matrix I. Transform the model (1) to

$$G^{-1/2}Y = (G^{-1/2}X)\ \theta + G^{-1/2}\epsilon. \tag{2}$$

Assume X and Y are observable. The GLS estimator $\hat{\theta}_{GLS}$ for θ is the solution to

$$||G^{-1/2}Y - (G^{-1/2}X)\ \hat{\theta}_{GLS}||^2 = \min_{\theta} ||G^{-1/2}Y - (G^{-1/2}X)\ \theta||^2. \tag{3}$$

Specify G in this problem such that the GLS works.

(1e) Show that

$$\hat{\theta}_{GLS} = (X^tG^{-1}X)^{-1}X^tG^{-1}Y. \tag{4}$$

(1f) Can Gauss-Markov Theorem be extended to the current setting? If so, state it carefully. If not, explain why.

[2] (30 points) The following data (its practical background skippd) contains 22 y values recorded at 6 different (repeated) x values.

obs#	x	y
1	0.05	0.05
2	0.05	0.10
3	0.25	0.25
4	0.25	0.35
5	0.50	0.75
6	0.50	0.85
7	0.50	0.95
8	1.25	1.42
9	1.25	1.75
10	1.25	1.82
11	1.25	1.95
12	1.25	2.45
13	2.10	3.05
14	2.10	3.19
15	2.10	3.25
16	2.10	3.43
17	2.10	3.50
18	2.10	3.93
19	2.50	3.75
20	2.50	3.93
21	2.50	3.99
22	2.50	4.07

(2a) Calculate the summary statistics $\sum x_i$, $\sum y_i$, $\sum x_i^2$, $\sum y_i^2$ and $\sum x_i y_i$.

(2b) Consider the simple linear regression model

$$y_i = \beta_0 + \beta_1 \ (x_i - \bar{x}) + \epsilon_i. \tag{5}$$

Assume the data passes a normality test. Estimate β_0, β_1 and the error variance parameter σ^2.

(2c) To test whether the data supports a linear relationship between x and y, specify the reduced model H_0 and the full model H_1, and write down an ANOVA table for the two nested models that indicates "sum of squares", "degree of freedom", "mean squares", etc.

(2d) Write down the expression of F-statistic and its numerical value for the test in (2c). You do not need to calculate the p-value.

(2e) Is the full model used in (2c) a 1-way ANOVA model? Is the F-test used in (2c) – (2d) a standard practice for 1-way ANOVA? Explain why.

[3] (20 points) For $i = 1, 2, ...,$ let X_i be iid $N(0, 1)$ random variables, and $\epsilon_i = 0.025 \ (X_i^4 - 3X_i^2)$, $Y_i = X_i + \epsilon_i$. An investigator does not know how the data were generated so he runs a regression of Y on X. Note that $E(X_i^{2n-1}) = 0$ and $E(X_i^{2n+2}) = (2n + 1) \ E(X_i^{2n})$ for $n = 1, 2, ...$

(3a) Calculate $E\epsilon_i$ and $Cov(X_i, \epsilon_i)$.

(3b) Show the coefficient of determination $R^2 = \frac{SSR}{SSTO}$ is about 0.97. (Hint: Think of R^2 as an approximation for $\left[\frac{Cov(X_i, Y_i)}{SD(X_i) \ SD(Y_i)}\right]^2$.)

(3c) Should the investigator trust a definitive causal relationship between X and Y based on such a large value of R^2? Justify your answer.