INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Mid-Spring Semester 2016-17

Date of Examination : 15/02/2017(AN) **Duration 2hours** **Marks: 30**

Subject No. : MA60056 (Regression and Time Series Models)

Department: Mathematics

Students: PGDBA core + BTech/MSc elective (92)

Questions:

1. Let $\{(x_i, y_i)\}_{i=1}^{N}$ be a data set for which the following regression model is fit:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$.

   a. What is the distribution of the least squares estimator $\hat{\beta}_1$ assuming that the variance σ^2 is also estimated from the data? [2 marks]

   b. Suppose $N = 2$ and $|x_1 - x_2| = d$ for some positive number d. Assuming that the estimator of σ^2 is bounded by a positive number M (in other words $MS_{Res} < M$), what choice of d will result in $Var(\hat{\beta}_1) < 0.1$? [3 marks]

2. Let $\{(x_i, y_i)\}_{i=1}^{N}$ be a data set for which the following simple linear regression model is fit:

$$y = \beta_0 + \beta_1 x + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$.

   a. Give a test procedure using t statistic to test whether $\beta_1 = 0$. [2 marks]

   b. Give a test procedure to test whether $\beta_1 = 0$ using ANOVA. [2 marks]

   c. Show that the approaches in (2a) and (2b) are equivalent. [2 marks]

3. Let the true relationship between the variables x and y be given as $y = e^x + e^{-x} + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$. The practical constraint forces you to build at best a polynomial regression model between x and y.

   a. Propose a polynomial regression model of degree 6 to solve this problem. (Hint: Consider the Taylor's series expansion of y). [2 marks]

   b. Do all the powers of x up to degree 6 are significant in this model? [1 mark] no

   c. Assume that the data given to you is $D = \{(x_i, y_i)_{i=1}^n\}$. Write down the general form of the regressor matrix X for this particular model in (3a). [2 marks]

4. Let $y = X\beta + \epsilon$ be a multiple regression model where $y \in \mathbb{R}^{n\times 1}$, $X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^{p\times 1}$, and $\epsilon \sim N(0, \sigma^2 I_n)$ for the data $D = \{(x_i, y_i)_{i=1}^n\}$. Let $\hat{\beta}$ be the least squares estimator of β and $\hat{y}$ be the prediction vector. Define $e = y - \hat{y}$ to be the residual vector. Derive the distribution of the residual vector e. Justify each step in this derivation. [4 marks]

5. State whether the following statements are true of false. Justify your answer in each case. [10 marks]

   a. Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from $N(0,1)$ and $\{Y_1, Y_2, \ldots, Y_m\}$ be another independent random sample from $N(\mu, \frac{1}{m^2})$. Then $\frac{\sum_{i=1}^m (Y_i - \mu)}{\sum_{i=1}^n X_i^2}$ follows t distribution with m degrees of freedom.

   b. In the simple regression setting with the data $D = \{(x_i, y_i)_{i=1}^n\}$, the degrees of freedom associated with the total variation in y given by $SS_T = \frac{1}{n}\sum_{k=1}^n (y_k - \bar{y})^2$ are $n - 1$.

   c. In the multiple regression setup, consider the ANOVA test to check the significance of the regression coefficients. The F statistic used in this test has $F_{n-p,n-1}$ distribution where p is the number of parameters estimated from n number of data points.

   d. In a simple regression case, suppose that the t test for significance for the slope parameter $(\beta_1 = 0)$ reveals that the slope parameter is significant. Then the regression model will predict the response variable accurately.

   e. The condition number of the matrix $\begin{pmatrix} 1 & 1 \\ 1.001 & 1 \end{pmatrix}$ is very high.