Registration No. | | | | | | | | | |

M.E./M.Tech. Degree Examinations, January 2017

First Semester

COMPUTER SCIENCE AND ENGINEERING

CP16004 – DATA ANALYSIS AND BUSINESS INTELLIGENCE

(Regulation 2016)

QP Code: 403008

Time: Three hours **Maximum : 100 marks**

Answer **ALL** questions

PART A - (10 X 2 = 20 Marks)

1. Explain the parameters of a simple Linear Regression model .
2. Explain the types of linear regression model.
3. Define logistic regression.
4. Define Generalized Linear model and mention its classes.
5. Why do we need simulation for predictive inferences?
6. What is matching and subclassification?
7. Specify the motivation for multilevel linear modeling.
8. Mention the fundamental problem of causal inference.
9. List the advantages of ANOVA model.
10. Specify the choices in the design of data collection.

PART B - (5 X16 = 80 Marks)

11. (a) (i) Explain the steps involved in building a linear regression model using one predictor. **(8)**

(ii) Suppose that, for a certain population, we can predict log earnings from log height as follows: **(8)**

- A person who is 66 inches tall is predicted to have earnings of $30,000.
- Every increase of 1% in height corresponds to a predicted increase of 0.8% in earnings.

- The earnings of approximately 95% of people fall within a factor of 1.1 of predicted values.

Give the equation of the regression line and the residual standard deviation of the regression.

(OR)

(b) Explain statistical inference and types of probability distribution models. **(16)**

12. (a) Illustrate with an example the steps involved in building a generalized linear model. **(16)**

(OR)

(b) Consider the following scenario: Many of the wells used for drinking water in Bangladesh and other South Asian countries are contaminated with natural arsenic, affecting an estimated 100 million people. Arsenic is a cumulative poison, and exposure increases the risk of cancer and other diseases, with risks estimated to be proportional to exposure. The bad news is that even if your neighbor's well is safe, it does not mean that yours is safe. However, the corresponding good news is that, if your well has a high arsenic level, you can probably find a safe well nearby to get your water from—if you are willing to walk the distance and your neighbor is willing to share. A research team from the United States and Bangladesh measured all the wells and labeled them with their arsenic level as well as a characterization as "safe" (below 0.5 in units of hundreds of micrograms per liter, the Bangladesh standard for arsenic in drinking water) or "unsafe" (above 0.5). People with unsafe wells were encouraged to switch to nearby private or community wells or to new wells of their own construction. A few years later, the researchers returned to find out who had switched wells. Fit logistic regression analysis to understand the factors predictive of well switching among the users of unsafe wells. **(16)**

13. (a) Use simulation to check the fit of a time-series model: find time-series data and fit a first-order autoregression model to it. Then use predictive simulation to check the fit of this model. **(16)**

(OR)

(b) Illustrate predictive simulation for generalized linear models. **(16)**

14. (a) Explain about partial pooling with predictors. **(16)**

(OR)

(b) Illustrate how multilevel model handles predictors at the group as well as individual levels with an example. **(16)**

15. (a) Explain multilevel power calculations using fake data simulations. **(16)**

(OR)

(b) Discuss ANOVA and multilevel generalized linear models. **(16)**