PGDM, PGDM (IB) 2015-17

Predictive Business Analytics

DM-341 / IB-314

Trimester III, End-Term Examination: March 2016

Time allowed: 2 hrs 30 min

Max Marks: 50

Roll No: _____

Instruction: Students are required to write Roll No on every page of the question paper, writing anything except the Roll No will be treated as **Unfair Means.** In case of rough work please use answer sheet.

Section A

Attempt ANY 3 from the 5 questions in this section. (3 x 5)
Each question carries 5 marks.

A1. Read the Housing Prices dataset, *HousePrices.csv* provided.
Create a combined scatterplot for the Prices versus SqFt, for the different values of Neighborhood. [5]

A2. See the plots in Figure A1.

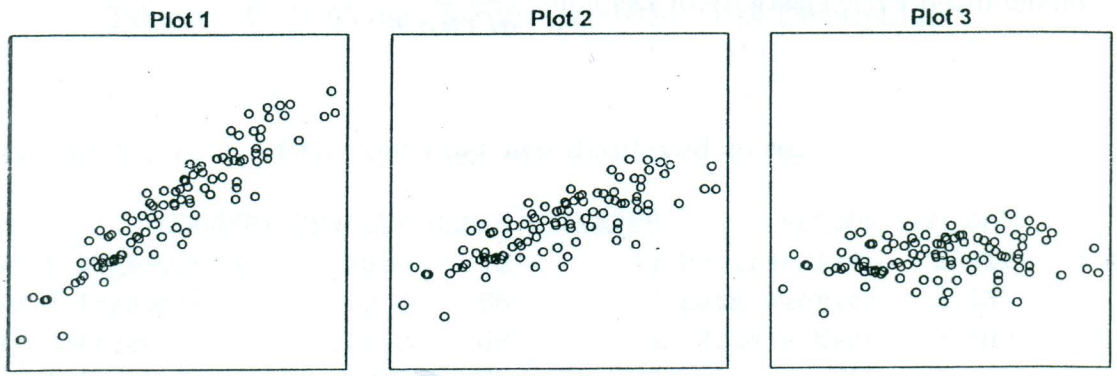

Figure A1

Discuss the Signal to Noise ratio for the above figure and the implications of the statistical significance of the corresponding simple linear regression models. [5]

A3. For the House Prices data set, regress Price on the Square Foot area (SqFt). Report the model *p-value*. What does it indicate? [5]

A4. For a linear regression, the value of the Adjusted R-Square is given by

$$Adj.R^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k}\right)$$

where n is the sample size and k is the number of model coefficients (intercept plus number of independent variables). Data is collected for 20 respondents. In the first instance where there is one independent variable in the linear regression model, we have a R-squared value of 0.75. When one more independent variable is added, the R-Squared increases to 0.76. Using the formula of Adjusted R-squared, which of the two models would you recommend and why? [5]

A5. Examine the Gender Discrimination dataset. How would you create a regression model, using dummy variables, to find out if there is discrimination in salary based on the employee's gender? [5]

Section B

Attempt ANY 2 from the 3 questions in this section. (2 x 10)
Each question carries 10 marks.

B1. Using the House Prices dataset provided, *HousePrices.csv*, answer the following questions.

(a) What are the average house prices in the four different regions? **[2]**

(b) What are the average price per sq.ft. area in the four regions? **[3]**

(c) Fit a regression model, using *Neighborhood* dummy and comment on the differences in part (b) of this question. **[5]**

B2. Find out, using linear regression if the price per sq. ft. can be modeled on the Neighborhood and Offers.
Describe the model, including how you construct the dummy variables.
(Hint: there should be 3 dummy variables).
Report on the model fitment and the coefficients obtained. Is price dependent upon the Neighborbood? **[10]**

B3. Using the *PriceAndDemand.csv* dataset, regress Qty (i.e. Demand) on Price using a log-log transform. Describe the model generated and its fitment. What is the model fitted Price elasticity of Demand, $\frac{\partial Qty}{\partial Price}$?

Section C

This is a compulsory question carrying 15 marks. **(1 x 15)**

C1. This uses 1995 economic data where 101 countries of the world were surveyed. This question uses 6 data items from this set tabulated herein. Below is a regression model for per capita GDP of a country (using a logarithmic transformation), using the 5 other variables as predictors.

Sl.No.	Data Name	Type	Value Levels	
1.	DENSITY	metric		population density
2.	URBAN	metric		urban population percentage
3.	LIFEEXPF	metric		female life expectancy
4.	REGION	factor	Pacific / Asia Latn America OECD Middle East East Europe Africa	*used as base in dummy variable model*
5.	DEATH_RT	metric		death rate per 1000 population
6.	LOG_GDP	metric		log of per capita Gross Domestic Product **dependent**

Table 1: World 95 data for 101 countries: Data Description

The first 5 rows of the data set are displayed here.

```
##         COUNTRY DENSITY URBAN LIFEEXPF      REGION LOG_GDP
## 1 Afghanistan    25.0    18       44 Pacific/Asia   2.312
## 2 Argentina      12.0    86       75 Latn America   3.532
## 3 Armenia       126.0    68       75  Middle East   3.699
## 4 Australia       2.3    85       80         OECD   4.227
## 5 Austria        94.0    58       79         OECD   4.265
```

The ANOVA table and the Regression table of Coefficients follows.

```
## Analysis of Variance Table
##
## Response: LOG_GDP
##            Df Sum Sq Mean Sq F value  Pr(>F)
## DENSITY     1   1.10    1.10   17.63 6.0e-05 ***
## URBAN       1  22.24   22.24  356.15 < 2e-16 ***
## LIFEEXPF    1   6.72    6.72  107.57 < 2e-16 ***
## REGION      5   4.84    0.97   15.50 3.5e-11 ***
## DEATH_RT    1   0.19    0.19    2.97   0.088 .
## Residuals 97   6.06    0.06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Call:
## lm(formula = LOG_GDP ~ ., data = v2)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5632 -0.1299  0.0003  0.0896  0.7440
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.11e-01   6.94e-01    0.30  0.76135
## DENSITY                  7.88e-05   4.32e-05    1.83  0.07099 .
## URBAN                    5.76e-03   1.67e-03    3.46  0.00080 ***
## LIFEEXPF                 4.15e-02   7.69e-03    5.40  4.7e-07 ***
## REGIONEast Europe       -3.15e-01   9.06e-02   -3.48  0.00076 ***
## REGIONPacific/Asia      -4.47e-01   1.44e-01   -3.10  0.00253 **
## REGIONAfrica            -2.19e-01   1.52e-01   -1.44  0.15305
## REGIONMiddle East       -1.74e-01   1.34e-01   -1.30  0.19768
## REGIONLatn America      -5.27e-01   1.16e-01   -4.54  1.6e-05 ***
## DEATH_RT                 2.39e-02   1.39e-02    1.72  0.08821 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.25 on 97 degrees of freedom
## Multiple R-squared:  0.853,Adjusted R-squared:  0.839
## F-statistic: 62.4 on 9 and 97 DF,  p-value: <2e-16
```

Questions for this case study *(Question C1.)*

(a) Comment on the model and its fitment. What is the null hypothesis for the overall *regression model* and how do you conclude from the output whether H_0 is rejected or not? **[5]**

(b) Write down the linear regression model (using dummy variables). How many separate linear equations do we have in the model? **[5]**

(c) Using the regression output list down the Regions in descending order of the countries' average per capita GDP. Which Region has the highest per capita GDP; which has the lowest? **[4]**

(d) Which independent variable(s) are *not* significant at $\alpha = .05$? **[1]**