ISOM2600 Practice Paper

1. A data set contains sales of one-family homes in the Levittown, NY area from June 2010 through May 2011. A house pricing analyst attends to determine the appropriate sales price for a home. He regressed the house price on the set of predictors including the number of bedrooms, the living area, the lot size etc. The result of data analysis is given in the appendix, including the descriptive statistics, correlation matrix and regression output.
    a. Order the predictors by its correlation with 'Sales price' in descending order.
    b. According to the regression output, suggest the most insignificant predictor to be dropped from the model. State your reason.
    c. Test the overall significance of the model using F test at the significance level of 0.05. State the null hypothesis, test statistics, p value and the conclusion.
    d. Estimate the standard deviation of random errors $\hat{\sigma}_{\varepsilon}$.
    e. Write down the fitted regression model equation.
    f. A new house with the number of bedroom = 2, bathrooms = 1, living area = 1050, lot size = 6000, year built = 1948 and property tax = 6306. Estimate its sales price using the model.
    g. What is the percentage of variation of y explained by the model?

2. A national insurance organization wanted to study the consumption pattern of cigarettes in all 50 states and the District of Columbia. The demographic information of states and price are collected to find the underlying relationship between the consumption and these variables. Some result outputs are given in the appendix.
    a. State one effect of collinearity.
    b. Suggest a method to reduce collinearity.
    c. Test the null hypothesis $H_0: \beta_1 = 5$ using t test, where β_1 is the coefficient of age. Show the test statistic and your conclusion only. (You do **not** need to calculate the p value)
    d. Write down a major difference between R^2 and adjusted R^2.
    e. Look at the residual plots, does the model violate with the assumption of linear regression? Why?
    f. Write down the fitted regression model equation.
    g. The Governor of New York has raised the base price of cigarettes in NY to stop people from smoking, and would like to see how the consumption is expected to change after intervention. Calculate the expected consumption of NY, given that Age = 36, HS = 82, Income = 5316, Black = 26, Female = 51, Price = 50.

3. A business analyst would like to study the relationship between the median house price ('medv') for 506 neighborhoods around Boston and percent of households with low socioeconomic status ('lstat'). The result outputs are attached in appendix.
    a. He transforms 'medv' only by natural logarithm transformation to construct a semi-log model, model 1.
        i. Interpret the 95% confidence interval for the regression coefficient of 'lstat'.
        ii. Given that a new observation whose 'lstat' is 5.5, what is the predicted 'medv' by model 1?

4. A credit card company is building up a regression model on **the average outstanding monthly balance** for 50 individual credit card accounts. The model will be further used for predicting the credit-worthiness of future potential customers.
   The predictors are given:
   Purchases: the average monthly purchases
   Expense: the average monthly housing expenses
   House: Categorical with two levels {Renter, Owner}
   Gender: Categorical with two levels {M, F}
   Below are two examples from data:

Balance	Purchase	Expense	House	Gender
140	6.989199	4.783238	Renter	M
36	6.753247	9.627226	Owner	F

    a. Write down a model with all the predictors and β_i as the coefficient parameters. For dummy variable, use the level name to denote it (like the newHouse model in topic 2).
    $$Price = \beta_0 + \beta_1 \times transcationDate + \beta_2 \times houseAge + \beta_3 \times newHouse + \varepsilon_i$$
    b. The company found a model that fits the data well. The regression output is given in the appendix. Write down the fitted regression model.
    c. Amy claims that males tend to have less outstanding balance than females, given that other factors are the same. Do you agree with her? Why?
    d. A male house renter comes and is looking for a new credit card. Rewrite the model with the known information. How do the parameters change?

5. A broadcast operations manager at a local TV station is asked to reducing expenses by 8% during the next fiscal year. The manager seeks to investigate ways to reduce unnecessary labor expenses associated with the staff of graphic artists employed by the station. Currently, these graphic artists receive hourly pay for a significant number of *standby hours*, hours for which they are present at the station but not assigned any specific task to do. The manager collected weekly data for the number of standby hours (Y) and these four variables: the number of graphic artists present (X1), the number of remote hours (X2), the number of Dubner (broadcast graphics) hours (X3), and the total labor hours (X4). The manage would like to predict the number of future standby hours, identify the root causes of excessive number of standby hours. Variable selection is used to select predictors. Some of the useful result outputs are given in the appendix.

    a. If the manager decides to use best subset selection method, report the candidate models before selecting the single best model. State the reason.
    b. Followed from a), if adjusted R-square is used as the selection criteria to select the single best model, what predictors are included in the single best model?
    c. Suppose the manager changes the selection method from best subset selection to forward selection method, how many models are involved in forward selection method?
    d. If the manager decides to use forward selection method, report the candidate models before selecting the single best model. State the reason.

6. Suppose Leave-one-out cross validation (LOOCV) is used to evaluate two models:
   Model A: $y = b_0 + b_1x_1 + b_2x_2$ and
   Model B: $y = c_0 + c_1x_1 + c_2x_3$

   LOOCV is special K-CV in which K=n, the sample size, for example if n=5, it will use four observations as training set to build the model and use one item as the validation set.
   the original training set is as follow:

Observation no.	y	x1	x2	x3
1	271	358	656	340
2	152	319	449	279
3	274	322	151	287
4	183	339	440	300
5	135	289	409	339

   The result output for i^{th} leave one out cross-validation trained model is given in the appendix (i^{th} means the i^{th} observation is taken to be the validation set and the remaining data is used as training subset).
    a. Compute the LOOCV estimate for Model A.
    b. Compute the LOOCV estimate for Model B.
    c. Which model is better in terms of LOOCV?

7. There is a data set containing information related to a direct marketing campaign of a Portuguese banking institution and its attempts to get its clients to subscribe for a term deposit. Peter would like to know whether the clients will subscribe for a term deposit. He built a logistic regression that the calculate the probability for realY=1 (the clients will subscribe for a term deposit) with following predictors:

   - 'duration': Duration of last contact in seconds
   - 'nr_employed': Number of employees
   - 'poutcome_success': The previous marketing campaign is successful
   - 'emp_var_rate': Employment variation rate
   - 'previous': Number of contacts performed before this campaign
   - 'poutcome_nonexistent': No previous marketing campaign
   - 'contact_telephone': Using telephone to communicate
   - 'month_mar': Month that last contact was made was 'March'
   - 'month_oct': Month that last contact was made was 'October'
   - 'cons_price_idx': Consumer Confidence Index
   - 'month_sep': Month that last contact was made was 'September'
   - 'month_may': Month that last contact was made was 'May'
   - 'default_no': The client DOES NOT have credit in default
   - 'job_student': Client's occupation is 'Student'
   - 'job_retired': Client's occupation is 'Retired'

   The result of logistic regression is given in the appendix.

   a. State the method (full name) for estimating the coefficients of this logistic regression model.
   b. Write down the fitted logistic regression model in term of 'odds ratio'.
   c. Interpret exp (0.4652) in the logistic regression model.
   d. If I would like to measure the association strength between 'job_student' and 'realY', what statistics can I use?
   e. Use following data to predict the probability of term deposit:
      'duration': 606
      'nr_employed': 5020
      'poutcome_success': 1
      'emp_var_rate': 1.1
      'previous': 1
      'poutcome_nonexistent': 0
      'contact_telephone': 1
      'month_mar': 0
      'month_oct': 0
      'cons_price_idx': 89
      'month_sep': 0
      'month_may': 1
      'default_no': 1
      'job_student': 0
      'job_retired': 1

8. [Follow up question of above] Peter used Naïve Bayes Classification to classify whether the clients will subscribe for a term deposit with the same predictors. The confusion matrix with cut-off probability = 0.5 and the KS chart of using Naïve Bayes Classification are shown in the appendix.
    a. State the assumption of attributes when Naïve Bayes Classification.
    b. Calculate the accuracy.
    c. Calculate the specificity and sensitivity in the confusion matrix.
    d. The accuracy will be lower if Peter use the cut-off probability that achieve the K-S measure. State the condition that the cutoff probability given by K-S Measure can bring the highest accuracy rate.