**ISOM2600 Practice Paper**

1. Sales data of one-family homes in the Levittown, NY area from June 2010 through May 2011 were obtained. A real estate valuation analyst would like to build a model for the sales price of a house. 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.
   1. Order the predictors by its correlation with ‘Sales price’ in descending order.
   2. According to the regression output, suggest the most insignificant predictor to be dropped from the model. State your reason.
   3. 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.
   4. Estimate the standard deviation of random errors .
   5. Write down the fitted regression model equation.
   6. 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.
   7. What is the percentage of variation of sales prices 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 cigarette prices are collected to find the underlying relationship between the consumption and these variables. Some results are given in the appendix.
   1. State one effect of collinearity.
   2. Suggest a method to reduce collinearity.
   3. Test the null hypothesis using t test, where is the coefficient of age. Show the test statistic and your conclusion only. (**NO** need to calculate the p value)
   4. Write down a major difference between R2 and adjusted R2.
   5. Look at the residual plots. Does the model violate with the assumption of linear regression? Why?
   6. Write down the fitted regression model equation.
   7. The Governor of New York has raised the base price of cigarettes in NY to discourage 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. An analyst would like to study the relationship between the median house price (‘medv’) for 506 neighborhoods around Boston and the percentage of households with low socioeconomic status (‘lstat’). The result outputs are attached in appendix.
   1. He transforms ‘medv’ only by natural logarithm transformation to construct a semi-log model, model 1.
      1. Interpret the 95% confidence interval for the regression coefficient of ‘lstat’.
      2. Given that a new observation whose ‘lstat’ is 5.5, what is the predicted ‘medv’ by model 1?
   2. He later transforms both ‘medv’ and ‘lstat’ by natural logarithm transformation to construct a log-log model, model 2.
      1. He concludes that model 2 is better than model 1 as its r-squared is higher. Do you agree? Explain your answer. [For this question, assume all assumptions for the errors are satisfied.]
      2. Given that a new observation whose ‘lstat’ is 5.5, what is the predicted ‘medv’ by model 2?
      3. He says that the heteroskedasticity problem does not exist in model 2 as he transforms both ‘medv’ and ‘lstat’ by natural logarithm transformation. Do you agree? Explain your answer briefly.
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 used for predicting the credit-worthiness of future potential customers.

The predictors are as follows:

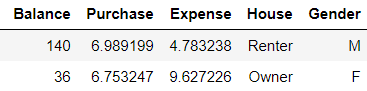
Purchases: the average monthly purchases

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:



1. Write down a model containing all predictors and let be the coefficient parameters. Denote the dummy variable by the level name, like the newHouse model in topic 2).
2. The company found a model that fitted the data well. The regression output is given in the appendix. Write down the fitted regression model.
3. 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?
4. 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.
   1. If the manager decides to use best subset selection method, report the candidate models before selecting the single best model. State the reason.
   2. From part (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?
   3. Suppose the manager changes the selection method from best subset selection to forward selection method, how many models are involved in forward selection method?
   4. If the manager decides to use forward selection method, report the candidate models before selecting the single best model. State the reason.
6. Suppose 5-fold cross validation is used to evaluate two models:

Model A: and

Model B:

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 ith 5-fold cross-validation trained model is given in the appendix (ith means the ith observation is taken to be the validation set and the remaining data is used as training subset).

1. Compute the 5-fold CV estimate for Model A.
2. Compute the 5-fold CV estimate for Model B.
3. Which model is better in terms of 5-fold CV?
4. A data set contains information related to a direct marketing campaign of a Portuguese banking institution. The purpose of the campaign is to promote subscription of a term deposit among its clients. To predict if a client will subscribe for a term deposit, Peter builds a logistic regression that calculates 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.

1. State the method (full name) for estimating the coefficients of this logistic regression model.
2. Write down the fitted logistic regression model in term of ‘odds ratio’.
3. Interpret the coefficient estimate 0.4652 in the logistic regression model.
4. 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

1. Peter applied **ANOTHER** classification method 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 are shown in the appendix.
   1. Calculate the accuracy.
   2. Calculate the FPR and TPR in the confusion matrix.
   3. 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.
2. Apply k-means clustering on a small dataset listed in the result output in the appendix.
   1. Standardize the data (use population standard deviation here).
   2. After applying k-means clustering with on the standardized data, the labels are listed in the result output. Find the centroid for each cluster.
   3. Draw the scatterplot of the standardized data with for label 0 and for label 1.
   4. Based on the elbow plot in the result output, is it reasonable to conclude that the optimal number of clusters is 4 for k-means clustering? Explain your answer.
3. Apply hierarchical clustering on a small dataset listed in the result output in the appendix.
   1. Perform the hierarchical clustering on ‘X1’ and ‘X2’ by using Euclidean distance and single linkage. Draw the dendrogram.
   2. We would like to have 3 clusters for the clustering model in (a). Find the centroid for each cluster.
   3. Given a set of new observations, can we use the clustering model in (a) to predict new labels for them?