

ACCT648 Applied Statistics for Data Analysis

Assignment 3

Deadline of Submission: by 2pm on 18 October (Monday)

1. The marketing manager of a large supermarket chain has the business objective of using shelf space most efficiently. Towards that goal, she would like to use shelf space and operating cost to predict the sales of a specialty pet food. Data are collected from a random sample of different stores over 3 months period and are stored in the **Sales2021A.csv** data set. She would like to construct a multiple linear regression model to predict the total weekly sales, *Sales*, in thousand dollars with the independent variables operating cost, *Cost*, in hundred dollars and shelf space, *Space*, in square feet. The multiple regression model is:

$$Sales = \beta_0 + \beta_1 Cost + \beta_2 Space + error$$

- (a) Find the regression equation M_1 by the least-squared method.
 - (b) Construct a 95% confidence interval estimate of β_2 in model M_1 .
 - (c) Without making any model assumptions, we use the Bootstrap approach with 40,000 replicates to construct the other regression model M_2 . Set the random seed to (6128). Find the Bootstrap estimates of the regression model M_2 .
 - (d) Construct a 95% Bootstrap Percentile confidence interval estimate of β_2 in model M_2 .
2. Suppose that we wish to predict whether a customer will default his/her loan based on the independent variables *Balance*, *Income*, and *Married*. All 8000 observations are stored in the data set **Default2021A.csv**.
 - (a) Develop a logistic regression model, L_1 , to predict the probability of defaulting a loan, based on all independent variables.
 - (b) Develop the second logistic regression model, L_2 , to predict the probability of defaulting a loan, based on only *Balance* and *Married* independent variables.
 - (c) Develop the third logistic regression model, L_3 , to predict the probability of defaulting a loan, based on *Balance* and *Married* independent variables with their interaction effect.
 - (d) Explain why model L_2 is the best model according to BIC criterion.
 - (e) Predict the probability of defaulting a loan given that the customer is a married customer with balance of 2100 and Income of 12000 under model L_2 .
 - (f) Find the confusion matrix of model L_2 with the threshold value 0.6 for classifying a customer will default the loan.
 - (g) Find the sensitivity, specificity and total error rate of the model L_2 with the threshold value 0.6.

3. Suppose we collect data for a group of 120 students in a statistical course with two independent variables X_1 = average studying hours per week, X_2 = current grade point average (GPA), and one dependent variable $Y = 1$ for pass (or $Y = 0$ for fail) the course .

We fit a logistic regression model: $\log(\text{odds ratio}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$ to predict whether a student will pass the course. *R*-outputs produce estimated coefficients, $\hat{\beta}_0 = -7.58$, $\hat{\beta}_1 = 0.473$, and $\hat{\beta}_2 = 1.183$. The observations of the first six students are given as follows:

Student	Y	X_1	X_2
1	1	8.4	3.18
2	1	11.5	3.35
3	1	12.4	3.19
4	0	10.3	3.04
5	0	8.8	2.86
6	1	9.2	3.41

- (a) Based on the estimated logistic regression model, predict the probability that a student who always studies 9.0 hours per week on average and has a GPA of 3.10 will pass the course.
 - (b) At least how many hours would the student in part (a) need to study if he or she want to have more than 70% predicted chance of passing the course?
 - (c) Find the deviance residuals of the first six observed students.
 - (d) By using the estimated logistic regression model with the threshold value being 0.6 for classification of passing the course, determine whether the model makes any error to predict each of the above six observed students. If there is an error, determine what type of error as well.
4. Based on **Boston2021A.csv** data set, we are interested in predicting the median house value *medv* in Boston with other ten variables as possible predictors.
 - (a) Fit the multiple regression equation M_1 to predict *medv* with all given independent variables.
 - (b) Fit the multiple regression equation M_2 to predict *medv* with all given independent variables, except variables *tax* and *age*.
 - (c) Determine which model M_1 or M_2 is better according to AIC criterion.
 - (d) Determine which model M_1 or M_2 is better by using the Leave-one-out cross-validation approach.
 - (e) Determine which model M_1 or M_2 is better by using the 10-fold cross-validation approach. Set the random seed to (6009).

-END-