INDENG 242: Applications in Data Analysis, Fall 2022

Assignment 1

September 18, 2022

Problem 1: (10 points)

James Chapter 3, Exercise 5 (Fitted values in simple linear regression without an intercept).

Problem 2: (10 points)

i) James Chapter 4, Exercise 1 (Equivalent ways of writing logistic regression). (2 points)

ii) Write down the log likelihood for the logistic regression model and show how to derive a non linear system to find the coefficients (Hint: what do we know about logarithm?). (5 points)

iii) Using part i) and your knowledge of linear regression, elaborate on the meaning of the coefficients for a logistic regression. (3 points)

Problem 3: Framingham Heart Study (40 points)

In this exercise, you are asked to build models using Framingham Heart Study data in order to predict coronary heart disease (CHD) and to make recommendations to better prevent heart disease. The dataset is in the file `Framingham.csv`. Each observation representing the data from a particular study participant. The variables are described in Table 1.

a) (15 points) Please build a logistic regression model using all of the provided independent variables to predict `TenYearCHD`. Be sure to address the following in your explanation (Do not simply copy the summary from your programming languages).

i) Split the data with 75% for the training set and 25% for the testing set.

ii) Build your model with the training set. Describe and show the logistics regression equation of your model.

iii) Evaluate your model's performance on the testing set using the threshold 0.4. State the model's accuracy, True Positive Rate (TPR) and False Positive Rate (FPR) and explain these three metrics in natural language.

Table 1: Variables in the dataset `Framingham.csv`.

Variable	Description
`male`	s biological sex male.
`age`	Age (in years) at first examination.
`education`	Some high school, high school/GED, somecollege/vocational school, college.
`cigsPerDay`	Number of cigarettes per day. 1=full time, 2=part time.
`BPMeds`	Is on blood pressure medication at time of first examination.
`prevalentStroke`	Previously had a stroke.
`prevalentHyp`	Currently hypertensive.
`diabetes`	Currently has diabetes.
`totChol`	Total cholesterol (mg/dL).
`sysBP`	Systolic blood pressure.
`BMI`	Body Mass Index, weight (kg)/height (m)2.
`heartRate`	Heart rate (beats/minute).
`glucose`	Blood glucose level (mg/dL).
`TenYearCHD`	Experienced coronary heart disease within10 years of first examination.

b) ROC Curve(25 points)

i) Show the ROC curve for your logistic regression model on the testing set.

ii) Now, build the logistics model again by setting the class weight of `TenYearCHD` to balanced. Show the confusion matrix and the ROC curve for the balanced model.

iii) Briefly compares the model performances between the original model and the balanced model. (Hint: You can talk more about the differences between the confusion matrix of the models.)

Problem 4: Nissan Rogue Sales Study (40 points) Nearly all companies seek to accurately predict future sales of their product(s). If the company can accurately predict sales before producing the product, then they can better match production with customer demand, thus reducing unnecessary inventory costs while being able to satisfy demand for their product.
In this exercise, you are asked to predict the monthly sales in the United States of the Nissan Rogue automobile. The Rogue is a car model of Nissan that was first produced in 2008. It is Nissan's best selling car in the United States. We will use linear regression to predict monthly sales of the

Rogue using economic indicators of the United States as well as (normalized) Google search query volumes. The data for this problem is contained in the file `Rogue_242.csv`. Each observation in the file is for a single month, from January 2008 through July 2021. The variables are described in Table 2.

Table 2: Variable in the dataset `Rogue_242.csv`

Variable	Description
`MonthNumeric`	The observation month given as a numerical value (1 = January, 2 = February, 3 = March, etc.).
`MonthFactor`	The observation month given as the name of the month
`Year`	The observation year.
`RogueSales`	The number of units of the Nissan Rogue sold in the United States in the given month and year.
`Unemployment`	The estimated unemployment rate (given as a percentage) in the United States in the given month and year.
`RogueQueries`	A (normalized) approximation of the number of Google searches for "Nissan Rogue" in the United States in the given month and year.
`CPIAll`	The consumer price index (CPI) for all products for the given month and year. This is a measure of the magnitude of the prices paid by consumer households for goods and services.
`CPIEnergy`	The monthly consumer price index (CPI) for the energy sector of the US economy for the given month and year.

a) Data split and first model (20 points)

i) Split the data with 75% for the training set and 25% for the testing set, preserving the chronological order. Why is keeping the time order important when handling time-series data? Briefly justify the rational behind it. (5 points)

ii) Build your model and fit it with the training set. Describe and show the linear regression equation of your model (Do not simply copy the summary from your programming languages). (5 points)

iii) Evaluate your model's performance on the testing set using the Out of Sample R^2Ẇrite your own function to calculate it, and explain the meaning of Residual Sum of Squares (RSS) and Total Sum of Square (TSS). (10 points)

b) Feature engineering and model improvements (20 points)

i) Run the VIF test to look for highly collinear features among quantitative attributes. Why is it important and what does the VIF_i coefficient represent? (Please elaborate using formulas). (5 points)

ii) Using expert knowledge (your intuition) and possibly hints coming from VIF, try to find a better performing model than the one created in part i). (5 points)

iii) Briefly compare the models performances between the original model and your best model. Apart from performance, elaborate on statistical significance and sign of the coefficients. (10 points)