GEB 6895: Business Intelligence
Department of Economics
College of Business Administration
University of Central Florida
Fall 2019

Assignment 3

Due Tuesday, October 8, 2019 at 11:59 PM
in *your* private mirror of the GEB6895F19 GitHub repo.

Instructions:

Complete this assignment within the space on your private mirror of the GEB6895F19 GitHub repo in the folder `assignment_03`. Create a folder called `my_answers` that will contain all of your work for this assignment. Within this folder, code your solutions in `.R` with the filename as specified. When you are finished, use `git` to `add`, `commit` and `push` your code to your private mirror of the GEB6895F19 repo. You are free to discuss your approach to each question with your classmates but you must `git push` in your own work.

Question 1:

As an analyst at an insurance company, your goal is to build a model for the life expectancy of your customers in each state. Your manager plans to use your model to determine appropriate insurance premia for policies in each state. The dataset `life_exp.csv` includes the following variables.

$lifeexpect_i$	=	life expectancy at birth, in years, in state i, in year 2010
$medinc_i$	=	the median household income in state i (thousands of dollars), in year 2010
$uninsured_i$	=	the percentage of the population (aged 0-64) in state i, without health insurance coverage, in years 2008-2010
$smoke_i$	=	the percentage of adults in state i who smoked, in years 2008-2012
$obesity_i$	=	the percentage of adults in state i who were obese (body mass index greater than or equal to 30), in years 2008-2012
$teenbirth_i$	=	the number of births to teenaged mothers in state i per $1,000$ females aged 15 to 19 years, in year 2010
$gunlaw_i$	=	whether or not state i had firearm laws relating to children in year 2010
$metro_i$	=	the percentage of the population in state i that lived in a metropolitan area, in year 2010

a) Inspect the data by analyzing the summary statistics of the variables, the correlation between explanatory variables and by plotting a histogram of the dependent variable. Does there seem to be any problems with the data? If so, make sure it is read in correctly. Use the `View()` command to visually inspect it and compare to the file in a spreadsheet, if necessary.

b) Build an initial model that includes all of the explanatory variables. Which variables appear to have explanatory power and which are candidates for omission? Perform any tests at the 5% level of significance. That is, identify any variables that have t-statistics greater than 1.96 or have p-values less than 0.05 and keep them in the model.

c) Adjust the model to exclude one variable at a time. For each candidate variable, consider whether omitting the variable raises or lowers the adjusted R^2.

d) Proceed to reduce the model until all of the variables are statistically significant at the 5% level. Do the results agree, at least qualitatively, with the results from the full model?

e) Try to build a model using another approach. Instead of starting with a full model and removing variables, start with a single variable and add one variable at a time. Do you arrive at a different model? Which model do you recommend?

Question 2:

In the quantitative marketing department of a major hospital, your aim is to build a model that describes patients' choice of hospital. Use the data in the file `hospital_choice.csv` to conduct your analysis. The dataset includes the following variables.

D_i	=	an indicator of whether patient i chose to go to Cedars Sinai Hospital, otherwise 0 if they chose UCLA
$INCOME_i$	=	the average income in the zip code of patient i, in thousands
$DISTANCE_i$	=	is the distance from the home of patient i to Cedars Sinai Hospital minus the distance to UCLA
OLD_i	=	an indicator for whether patient i is more than 75 years of age

This dataset consists of 499 pneumonia patients in the greater Los Angeles area who chose either the Cedars Sinai Medical Center or the UCLA Medical Center. Typically, economists would expect price to have a major influence on such a choice but these patients were covered by medical insurance, so other factors might become more important in this decision.

a) Inspect the data by analyzing the summary statistics of the variables, including the correlation between explanatory variables. Does there seem to be any problems with the data? If so, make sure it is read in correctly. Use the `View()` command to visually inspect it and compare to the file in a spreadsheet, if necessary.

b) Estimate a linear probability model as an initial model. That is, fit a linear regression model using the `lm()` function. Note whether the estimated probabilities are well defined for all observations.

c) Estimate a logistic regression model, using the `glm()` function. Compare the qualitative predictions of the logistic model with the linear probability model. Specifically, summarize the predictions of each model to verify that the predictions are valid probabilities. Which model do you recommend?

d) Does the coefficient on $DISTANCE_i$ have the sign that you expect? Is it statistically significant? That is, verify whether the t-statistic is greater than 1.96 or the p-value is less than 0.05 for the coefficient on $DISTANCE_i$.

e) Estimate another logistic regression model that allows for a different slope coefficient on distance for older patients. You will have to modify the `formula = ...` argument within the `glm()` function to estimate `logit_model_2`, to include a term of the form `+ OLD*DISTANCE`. Is the coefficient on $OLD_i \times DISTANCE_i$ statistically significant? Does this new model lead you to revise your conclusion about the coefficient on $DISTANCE_i$? If so, how?