HW10

Your Name

10/20/2021

Problem 1

(A) Describe in your own words the difference between linear and logistic regression.

(B) Given an example of a dataset that would be appropriate to analyze with multiple linear regression but not with logistic regression.

(C) Given an example of a dataset that would be appropriate to analyze with logistic regression but not with linear regression.

(D) Given a model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

what is the interpretation of the coefficient β_2?

(E) Given a model

$$\ln\left(\frac{p}{1-p}\right) = -5 + 3x_1$$

how much do the odds increase if x_1 increases by 1?

Problem 2

This problem checks your understanding of multiple linear regression and diagnosis of these fits. Start by loading in the Advertising.csv file as a dataframe.

(A) Make a scatterplot matrix of the columns of the dataframe. Each row of this data set represents a single media market and the TV, Newspaper, and Radio columns contain spending amounts related to each media type while the sales value is the number of units sold (in thousands) in that market. Based on this plot, do you think multiple linear regression is appropriate to attempt?

(B) Fit a multiple linear regression model using all three media columns as predictors with the sales column as the dependent variable.

(C) Write the linear equation estimated by your fit.

(D) Write the coefficient of determination for your fit.

(E) How many sales would your model predict for a market that spent 200 on TV, 50 on radio, and 100 on newspaper?

Problem 3

This problem checks your understanding of implementing logistic regression in R. Start by loading in the default_ISLR.csv file as a dataframe.

(A) Fit a logistic regression model to this data with response variable y being the default column, p being the probability of default, and x being the balance column.

(B) Compute the coefficients of the fitted model and write the corresponding equation for the log odds.

(C) What percentage of the provided data does your model correctly classify?

(D) What is the probability that someone with a balance of 1950 will default, according to your model?

Problem 4

Please read the following data analysis project using logistic regression to analyze bias in the COMPAS measure: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm and the corresponding notebook on github: https://github.com/propublica/compas-analysis/blob/master/Compas%20Analysis.ipynb to provide brief responses to the following (You can skip the section that starts with “To test COMPAS’s overall predictive accuracy, . . . ’ and resume reading down below at:”Finally, we investigated whether certain types of errors. . . "):

(A) What research question is being addressed in this piece?

(B) Can you tell where the underlying data came from?

(C) The authors give a very thorough description of their data and modeling/cleaning choices. What were the main alterations they had to make to the underlying data set? Do their choices of terminology and metrics seem reasonable to you?

(D) What techniques that we have discussed in our class do the authors use to evaluate the data?

(E) What do you think about the visualizations and examples in this piece?

(E) The authors fit two large logistic regression models to their data. For each model, what is the probability that is being estimated? Which independent variables were found to have the largest impact? Was this surprising to you?

(F) At the end of the piece, the authors use contingency tables to compute false positive rates for the COMPAS classification. Do you find this analysis more or less convincing than the logistic model? Why?

(G) How strong is the final evidence the authors provide for their conclusions?