University of Chicago
Booth School of Business
41000: Business Statistics, Autumn 2021: Homework Assignment 4. Due in Week 9

Problem 1: House Prices

The dataset `boston` constructed in `hwk4s.R` contains prices on Boston house prices together with a number of characteristics. The explanatory characteristics are given as:

`medv` median value in $1000

`crim` per capita crime rate

`indus` % non-retail business

`rm` average number of rooms

`age` % built before 1940

`rad` accessibility to radial highways

`tax` property tax/$10,000

Use the `boston` dataset to answer the following:

(a) Run a multiple regression to predict house prices

(b) Which variables have significant marginal effects?

(c) Plot the 4-in-1 residual diagnostic plot. How do you feel about the multiple regression model given this extra information.

Problem 2: Zagat's

The dataset `zagats.csv` on the course website provides price and quality variables from the 2014 Zagat Guide of restaurants in New York City (`http://www.zagat.com/new-york-city`). There are $n = 114$ restaurants in the sample. The four variables are given by:

`Food` Zagat's food rating (up to a maximum of 30). Anything in the range 25 and above is excellent.

`Decor` The feel of the restaurant

`Service` up to a maximum score of 30

`Price` in dollars of a typical meal

Build a multiple regression model to assess whether price can be described by the quality measure. Use your model to address the following questions.

(a) Plot price versus each of the quality variables. Is there a linear relationship?

(b) Run a multiple regression of price on the three quality variables

Which variables are statistically significant?

Evaluate the marginal effects of each of the quality variables

(c) Suppose that a new restaurant opens up and gets a Zagat's rating of 27 for its `food`, 25 for its `service` and `decor` of 20

What's your best prediction for the `price` of a typical meal?

(d) Provide a plug-in prediction interval for your prediction.

Problem 3: Malaria.

The `malaria.txt` dataset contains a random sample of 100 children aged $2 - 15$ years from a village in Ghana. The children were followed for an 8 month period and there antibody levels were tracked. we have the following variables

`mal` Either zero or one depending on the presence of malaria

`age` in years

`ab` antibody level

Given this information, answer the following:

(a) Run a logistic regression to predict whether a child has malaria or not given their `age` and `ab` antibody level

(b) Interpret the regression *beta*'s

(c) What's your prediction for a child of age 10 and antibody $ab = 5$ level?

Problem 4: Homeless Data (Multiple Regression).

This problem analyses the factors that contribute to high rates of homelessness in certain cities. The data come from an article by William Tucker *Where do the homeless come from?*

The factors measured are:

variable	description
homeless.per.1000	Number of homeless per 1,000 population.
homeless.num	Total number of homeless.
poverty	A measure of overall poverty.
unemployment	Unemployment rate.
public housing	Whether or not the city has public housing projects.
population	Total population.
ave.temp	Average annual temperature in Fahrenheit.
vacancy.rate	measure of available housing.
ave.temp.jan	Average annual temperature in Fahrenheit, in January.
ave.temp.aug	Average annual temperature in Fahrenheit, in August.
precip	Average annual precipitation in mm.
rent.control	Does the city have rent control policies or not.

(a) Fit the model from the provided Rscript and interpret the results.

Specifically, what can you say about the role of rent control?

(b) The average January temperature variable has been transformed. Explain why this transformation is appropriate.

(c) Make and actual-vs-predicted plot. Identify any outliers.

Do these cities suggest other factors to consider?