Case Study 1

Practicing Physicians by County

Instructor: A. Chronopoulou

Case Study Overview

The goal in the case study is to propose a regression model for predicting the number of practicing physicians by county using information from the years 1990 and 1992. The data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. Each line of data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. The variables are (in the order they are recorded in the .txt file):

Variable Number	Variable Name	Description
1	Identification Number	1-440
2	County	County name
3	State	Two-letter state abbreviation
4	Land Area	Land area (square miles)
5	Total population	Estimated 1990 population
6	Percent of population aged 18–24	Percent of 1990 CDI population aged 18–24
7	Percent of population 65 or older	Percent of 1990 CDI population aged 65 or older
8	Number of active physicians	Number of professionally active nonfederal physicians during 1990
9	Number of hospital beds	Total number of beds, cribs and bassinets during 1990
10	Total serious crimes	Total number of serious crimes in 1990 as reported by law enforcement agencies
11	Percent high school graduates	Percent of adult population who completed 12 or more years of school
12	Percent bachelor's degrees	Percent of adult population who with bachelor's degrees
13	Percent below poverty level	Percent of 1990 CDI population with income below poverty level
14	Percent unemployment	Percent of 1990 CDI labor force that is unemployed
15	Per capita Income	Per capita income of 1990 CDI population (dollars)
16	Total personal income	Total personal income of 1990 CDI population (in millions of dollars)
11	Geographic region	Geographic region classification that is used in the US Bureau of the Census: 1=NE, 2=NC, 3=S, 4=W

The data set can be found on the course website under the name `CDI.txt`.

Learning Objectives

By the end of this case study, you will

1. enhance your skills in using R for the purpose of statistical analysis of a data set.
2. independently apply the regression in a real-world problem.
3. evaluate the applicability of the regression model.
4. draw conclusions, and make decisions about the initially stated research questions.
5. interpret your statistical outcomes using plain English.
6. demonstrate your team collaboration skills.

Suggestions

The goal of this case study is to build a model for predicting the number of active physicians using county demographic information. Here are some suggestions on how you could proceed to analyze the data set:

1. Use summary statistics and graphs to understand the nature and type of the variables in this data set. If you want, you can create your own variables based on the ones you have available.
2. You can start by fitting a full MLR model including all predictor variables in the model and then remove variables that are not statistically significant using a testing-based approach that we talked about in class (remember the full/reduced model tests).
3. For the final model that you choose, check for unusual observations and for departures from the model assumptions.
4. If necessary, you can employ some of the remedial techniques that we have discussed in class. If you decide to do transformations, do not forget to re-start the process above from the beginning.
5. The significance level α is up to you to choose.

There is not a unique way to analyze this data. You do not need to follow any of the suggestions above, and you should not prepare your summary by answering the aforementioned questions. What I am looking for is to see your thought process when analyzing the data set while applying the techniques we discussed in this class so far. Therefore, you need to make sure that you document and justify all the steps in your analysis. Also, make sure that you interpret your conclusion in layman's terms.

Groups

The case study should be done in a group of 2–4 students. You are free to choose your own group. If you do not have a group in mind, please use the Google form on Moodle to let me know and I will randomly assign you to a group.

Deliverables

The case study should be submitted on Gradescope as a group (only once case study per group) and should contain the following files:

(1) a **PDF** file containing a 3-4 page **executive summary of the analysis**. You need to make sure that your report is professionally and clearly written, addressed to someone who *knows statistics*. You should also include a concluding paragraph where you should state your conclusions in layman's terms. Any necessary plots or **R** output can be included in the text or part of an appendix or both. The appendix will not count towards the 4 page limit. You should **not** include any **R** code in the summary.

(2) an **R Markdown** and corresponding **HTML** files with the R code that you wrote to analyze the data set.

A rubric on which the grading of the case study will be based is posted on Moodle for your reference.

Deadline: Submit **one case study report per group** on Gradescope by **Friday, October 21** @ 11.59PM.

I think that analyzing real data is one of the most fun challenges in Statistics!
This is just a tiny glimpse of how that looks like in practice!
Try not to overthink it and enjoy!!
Good luck!
