County Demographic Information

Instructor: A. Chronopoulou

Instructions

This is a **group case study** where each group should consist of *2–4 students*.

You should prepare a report that should include: (i) an introduction , (ii) a full statistical analysis part, (iii) a conclusion, and (iv) an appendix with your R code (with comments). In your report, you can use the following questions **as a guide**, but it should not be structured as a homework where you directly answer the given questions. Your analysis should be supported by relevant R outputs (e.g. tables, and graphs). In the conclusion, you should summarize your findings using *non-statistical language*. You need to make sure that your report is professionally and clearly written, addressed to someone who *knows statistics in general, but is not an expert in regression*.

Deadline: Submit **one case study report per group** on Gradescope by **Friday, March 19** @ 11PM.

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 reasonable conclusions, and make decisions about the initially stated research questions.
5. interpret your statistical outcomes using plain English.
6. demonstrate your team collaboration skills.

Case Study Overview

This 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 information generally pertains to the years 1990 and 1992. The 17 variables are:

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`.

Suggestions

The goal of this case study is to build a model for predicting the number of active physicians in a CDI. Here are some suggestions on how you should 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.
2. You can start with a base model including as predictor variables the total population, land area and total personal income.
3. You can then check whether including additional predictor variables would be helpful in the model, and if so which variable(s) would be most helpful. The variables to consider are: percent of population 65 or older, number of hospital beds, and total serious crimes.
4. For the final model that you choose, check diagnostics, check for unusual observations, and perform a lack-of-fit test.
5. If necessary, you can employ some of the remedial techniques that we have discussed in class.
6. The significance level α is up to you to choose.

There is not a unique way to analyze this data. Make sure that you document and justify the steps taken in your analysis. Also, make sure that you interpret your conclusion in layman's terms.