MASM22/FMSN30(40) – Linear and Logistic Regression (with Data Gathering) 2019

COMPUTER EXERCISE 3: MODEL VALIDATION

A list of useful R-commands for multiple linear regression is found on the course home page
`http://www.maths.lth.se/matstat/kurser/masm22/lab1_vt19_useful.pdf`
`http://www.maths.lth.se/matstat/kurser/masm22/lab2_vt19_useful.pdf`
`http://www.maths.lth.se/matstat/kurser/masm22/lab3_vt19_useful.pdf`
See also the lecture R-files for examples.

Exercise 3: U.S. county demographic information

We are considering a dataset providing some county demographic information (CDI) for 440 of the most populous counties in the United States in years 1990–92. Each line of the dataset provides information on 14 variables for a single county. Counties with missing data were deleted from the dataset. See next page for further information. Here are the definitions for the variables considered in the population for which county demographic information (CDI) are available.

Variable	Description
id	identification number, 1–440
county	county name
state	state abbreviation
area	land area (square miles)
popul	estimated 1990 population
pop1834	percent of 1990 CDI population aged 18–34
pop65plus	percent of 1990 CDI population aged 65 years old or older
phys	number of professionally active nonfederal physicians during 1990
beds	total number of beds, cribs and bassinets during 1990
crimes	total number of serious crimes in 1990 (including murder, rape, robbery, aggravated assault, burglary, larceny-theft, motor vehicle theft)
higrads	percent of adults (25 yrs old or older) who completed at least 12 years of school
bachelors	percent of adults (25 yrs old or older) with bachelor's degree
poors	Percent of 1990 CDI population with income below poverty level
unemployed	percent of 1990 CDI labor force which is unemployed
percapitaincome	per capita income of 1990 CDI population (dollars)
totalincome	total personal income of 1990 CDI population (in millions of dollars)
region	Geographic region classification used by the U.S. Bureau of the Census, where 1 = Northeast, 2 = Midwest, 3 = South, 4 = West

The data for this exercise is available as a tab-separated txt-file, `CDI.txt`, on the course web page. Download and save it to your R working directory and then read it into R. Since region is a categorical variable, we should also turn it into a factor in R. We will also use number of physicians per 1000 inhabitants and the number of serious crimes per 1000 inhabitants:

```
cdi <- read.delim("CDI.txt")
cdi$region <- factor(cdi$region, levels = c(1, 2, 3, 4),
                     labels = c("Northeast", "Midwest", "South", "West"))
cdi$phys1000 <- 1000 * cdi$phys / cdi$popul
cdi$crm1000 <- 1000 * cdi$crimes / cdi$popul
```

The goal is to model the number of physicians per 1000 inhabitants, using the other demographic variables.

(a) Plot `phys1000` against each of `percapitaincome`, `crm1000` and `pop65plus`. Also plot `log(phys1000)` against the others. Does is seem reasonable to take the log?

(b) Use `log(phys1000)` as dependent variable and fit all $2^p = 2^3 = 8$ possible subsets of the three x-variables. We assume that we do not need any interaction terms. Calculate R^2_{adj} and BIC for each model. Which model is "best"?

(c) Use the largest model and plot the leverage, v_{ii}, both in order (against i) and against each of the three x-variables. Is there any observation with a dangerously large leverage? Try to identify in which variable the problem lies. Find out which county this is.

(d) Calculate the studentized residuals, r_i^*, and plot them against each of the three x-variables. Also plot them against the predicted values, $\hat{Y}_i$. Any problems?

(e) Look at the $s_{(i)}$ used in the studentized residuals. Which county produced the largest decrease when left out? What about the problematic one from (c)?

(f) Calculate Cook's distance, D_i and plot it against the x-variables. Any problems? What about the problematic counties from (c) and (e)?

(g) Calculate the DFBETA$_j$ and plot each one of them. Which of the β-parameters have been most influenced by the problematic counties?

(h) Leave out the problematic county from (c) and re-fit the largest model. Then redo (c)–(g) and see what happens.

(i) *If you have time left*, play around with the other variables and try to find a better model.