36-617: Applied Linear Models
Fall 2020
HW10 – Due Mon Nov 16, 11:59pm

- Please turn the homework in to Gradescope using the appropriate link in our course webspace at canvas.cmu.edu, under Assignments.
- Please read Sheather Section 10.1 (but not 10.2) for next Monday (there will be a reading quiz!). Note that the material in this hw also depends on Sheather 10.1 (as well as lectures 20 and 21).

Exercises

1. The file `cdi.dat` in the Canvas folder for this hw is taken from Kutner et al. (2005)[1]: It provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. Counties with missing data were deleted from the data set. The information generally pertains to the years 1990 and 1992. The definitions of the variables are given in Table 1 on p. 3 below. For this exercise we only consider a couple of these variables.

   Construct the variable `pct.hs.grad` <- (`hs.grad` / `pop`) × 100%. Then, using `state` as the cluster (or group) variable, create four plots like those on slides 12–15 of lecture 20 (intro to mlm I), using `pct.hs.grad` as the x variable, and `per.cap.income` as the y variable:

   (a) Ignore `pct.hs.grad` and only look at `mean(per.capita.income)` or `per.cap.income ~ 1` in each state

   (b) Ignore states and fit a single linear regression `per.capita.income ~ pct.hs.grad`

   (c) Use same slope on `pct.hs.grad` for all states, different intercepts

   (d) Fit a different regression `per.capita.income ~ pct.hs.grad` in each state, ignoring all the other states

   Print out the plots, and write two sentences for each graph: The first sentence should describe good and bad features of the plot; the second sentence should provide a comparison of this plot with the other three.

   Nb., In HW11 we will fit some multilevel models to this data. One such multilevel model would provide fits to the individual states which are a compromise between plots (b) [a single linear fit for all data, ignoring states] and (d) [a different linear fit for each state, ignoring the other states].

2. **This is a math problem, not a data analysis problem.** Consider the following multilevel model for data y_i, $i = 1, \ldots, n$, arranged into J groups, $j = 1, \ldots, J$, where each group j has n_j observations:

$$\left.\begin{array}{rcl} y_i &=& \alpha_{j[i]} + \epsilon_i,\ \epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2) \\ \alpha_j &=& \beta_0 + \eta_j,\ \eta_j \stackrel{iid}{\sim} N(0, \tau^2) \end{array}\right\}, \qquad (*)$$

   where the ϵ's and η's are also independent of each other. Prove the following four assertions:

   (a) If $i \neq i'$ and $j[i] \neq j[i']$, then $\mathrm{Corr}\,(y_i, y_{i'}) = 0$.

   (b) If $i \neq i'$ but $j[i] = j[i']$, then $\mathrm{Corr}\,(y_i, y_{i'}) = \frac{\tau^2}{\tau^2+\sigma^2}$.

[1] Kutner, M.H., Nachsheim, C.J., Neter, J. & Li, W. (2005) *Applied Linear Statistical Models, Fifth Edition*. NY: McGraw-Hill/Irwin.

(c) Let $\overline{y}_{j.} = \frac{1}{n_j} \sum_{i:j[i]=j} y_i$, the average of all observations in group j. Then $\mathrm{Var}\,(\overline{y}_{j.}) = \tau^2 + \sigma^2/n_j$

(d) Suppose we exactly replicate the experiment generating new data y_i^* following the model

$$\left.\begin{array}{rcl} y_i^* & = & \alpha_{j[i]} + \epsilon_i^*,\ \epsilon_i^* \overset{iid}{\sim} N(0,\sigma^2) \\ \alpha_j & = & \beta_0 + \eta_j,\ \eta_j \overset{iid}{\sim} N(0,\tau^2) \end{array}\right\} , \qquad (**)$$

so that the group level α's and η's (and β_0) are the same between $(*)$ and $(**)$ [the conditions we are measuring didn't change] but the new set of ϵ^*'s are independent of η's and ϵ's [we re-measured, and so we have new measurement error on each observation]. Form the group averages $\overline{y}_{j.}^*$, analogous to $\overline{y}_{j.}$. Then

$$\mathrm{Corr}\,(\overline{y}_{j,}, \overline{y}_{j.}^*) = \frac{\tau^2}{\tau^2 + \sigma^2/n_j} \ .$$

(This is another interpretation of the reliability coefficient $\frac{\tau^2}{\tau^2+\sigma^2/n_j}$.)

In all four parts, be sure to state any assumptions that you need.

Table 1: Variable definitions for CDI data from Kutner et al. (2005). *Original source:* Geospatial and Statistical Data Center, University of Virginia.

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 18–34	Percent of 1990 CDI population aged 18–34
7	Percent of population 65 or older	Percent of 1990 CDI population aged 65 or old
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, including murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies
11	Percent high school graduates	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
12	Percent bachelor's degrees	Percent of adult population (persons 25 years old or older) with bachelor's degree
13	Percent below poverty level	Percent of 1990 CDI population with income below poverty level
14	Percent unemployment	Percent of 1990 CDI population that is unemployed
15	Per capita income	Per-capita income (i.e. average income per person) of 1990 CDI population (in dollars)
16	Total personal income	Total personal income of 1990 CDI population (in millions of dollars)
17	Geographic region	Geographic region classification used by the US Bureau of the Census, NE (northeast region of the US), NC (north-central region of the US), S (southern region of the US), and W (Western region of the US)