Spring 2025 Advanced Econometrics - Problem Set #2

March 6, 2025

due 9:10 AM on Thursday, March 13 on Moodle

- *As mentioned in the first class, you are encouraged to work in groups. However, you must write up your own solutions in your own words. Answers that are highly similar will not be accepted.*
- *Provide details of your arguments and write complete answers for full credit.*
- *Provide your code. You are free to use any statistical software or programming language you are comfortable with.*

1. Consider the data generating process (DGP)

$$Y = X\beta + e$$

where $\beta = 1$, $X \sim N(0, 1)$, and $e \sim N(0, 0.75^2)$.

Simulate a random sample of size 20 from the DGP. Now you are asked to assess the influence of the four extra data points: $A : (-3, 3)$, $B : (0, 3)$, $C : (-3, -2)$, and $D : (0.5, 0)$ at a time.

(a) Produce a scatter plot with the regression line of Y on X (including the intercept) based on your 20 observations and the four extra data points (in color red) attached.

(b) Add point A to your sample.

i. Compute the leverage value h_{ii} and the leave-one-out residual $\widetilde{e}_i$ for point A.

ii. Produce a scatter plot with point A (in red) added, along with the full-sample and leave-"A"-out regression lines.

iii. Discuss the influence of point A based on your evidence in i and ii.

(c) Redo (b) for points B, C, D.

2. Import the data set `cps09mar.txt`. Consult the description file for the definition of variables.

(a) Construct a sample of single Asian men, consisting of the log of hourly wages (`lwage`), years of schooling (`educ`), potential experience (`exper`), and the squared potential experience (`exper2`), where `lwage = log(earnings/hours/week)`, `exper = age - educ - 6`, and `exper2 = exper`^{2}`/100`. (Hint: The sample size is 268.)

(b) Run a linear regression of `lwage` on `educ`, `exper`, and `exper2`. Compute the influence (defined below) and identify the most influential observation, where

$$\text{Influence} = \max_{1 \le i \le n} |\widehat{Y}_i - \widetilde{Y}_i|$$

(c) Is the influence in (b) meaningful in this empirical context? Why or why not?

(d) Recompute and discuss the influence after removing the most influential observation from the sample.