

Practical: Instrumental Variables

Qingyuan Zhao (Statistical Laboratory)*

July 9, 2026

This practice requires the `ivmodel` package:

```
# install.packages("ivmodel") # execute if ivmodel is not already installed.
library(ivmodel)
```

Instrumental variables are commonly used to deal with unmeasured confounding. We demonstrate this method using a prominent study on the economic return of schooling (Card, D. (1993), *Using Geographic Variation in College Proximity to Estimate the Return to Schooling*, Technical Report w4483, National Bureau of Economic Research, Cambridge, MA.).

1. Read the abstract of Card (1993). Familiarize yourself with this dataset (you can find the relevant information by executing `help(card.data)`) and the `ivmodel` function (`help(ivmodel)`).
2. (Group discussion) Execute the following code in R.

```
data(card.data)
# One instrument #
Y=card.data[, "lwage"]
D=card.data[, "educ"]
Z=card.data[, "nearc4"]
Xname=c("exper", "expersq", "black", "south", "smsa", "reg661",
        "reg662", "reg663", "reg664", "reg665", "reg666", "reg667",
        "reg668", "smsa66")
X=card.data[, Xname]
card.model1IV = ivmodel(Y=Y, D=D, Z=Z, X=X)
summary(card.model1IV)
```

Which variables in the dataset are used as the exposure, outcome, and instrumental variable? What covariates are used in this IV model? Discuss the output from the code.

Answer: Exposure: years of education; Outcome: Subject's log wage in 1976; IV: indicator for whether a subject grew up near a four-year college. The output (TSLS) gives a point estimate of 0.1315 and a p-value of 0.0168. This should be interpreted in view of possible violation of the IV exogeneity assumptions (IV and outcome may be determined by unmeasured variables like affluence of the neighboring area).

3. (Group discussion) Which IV model does the following code uses? Compare the results of these two models.

```
Z = card.data[, c("nearc4", "nearc2")]
card.model2IV = ivmodel(Y=Y, D=D, Z=Z, X=X)
card.model2IV
```

Answer: The p-value becomes smaller mainly because the point estimate is a bit larger. The F-statistic of the first stage regression is small, so the asymptotic approximations used to calculate the p-value is less reliable. There is a large literature on weak instrument bias, see, for example, https://scholar.harvard.edu/files/stock/files/nbersi2018_methods_lectures_weakiv1-2_v4.pdf.

*qyzhao@statslab.cam.ac.uk