**Instrumental Variable Estimation**

Read health\_iv dataset into stats program (Stata, SAS, R). It contains data on 10,089 respondents to the Medical Expenditures Panel Survey. The data contain the following variables:

|  |  |
| --- | --- |
| **Variable name** | **Variable description** |
| logmedexpense | Log of annual medical expenses |
| healthinsu | = 1 if individual has supplemental health insurance through employer |
| illnesses | Number of illnesses |
| age | Respondent age |
| logincome | Log of annual income |
| firmlocation | = 1 if firm has multiple locations |

1. Run some descriptive statistics in order to get a general idea of the data.
2. Estimate an OLS regression with the log of medical expenses as the dependent variable and health insurance as the independent variable. In another regression model, estimate the impact of health insurance on medical expenses adjusting for age, income, and the number of illnesses a respondent has.
3. Why is health insurance endogenous in this causal relationship?
4. Assume that health insurance is endogenous. What conditions need to be fulfilled in the instrumental variable estimation? Which variable from the dataset could serve as an instrument and why?
5. Estimate the effect of health insurance on the log of medical expenses using instrumental variables. Compare the results from this exercise to those from question 2.
6. Use 2-stage least squares to estimate the effect of health insurance on the log of medical expenses doing each stage separately. What is the F-statistic from the 1st stage? How do your results from the 2nd stage compare to those from question 5?

*Acknowledgement*: Data comes from Ani Katchova, Ohio State University, Econometrics Academy.