InstrumentalVariables

June 21, 2019

1 Instrumental Variables

by Jonas Peters, Niklas Pfister, Martin Emil Jakobsen and Rune Christiansen, 21.06.2019

This notebook aims to give you a basic understanding of the instrumental variable approach and when it can be used to infer causal relations.

In the following, let all variables have * zero mean,
* finite second moments, and * their joint distribution is absolutely continuous with respect to the Lebesgue measure.

In [5]:
```
#.libPaths("C:/ProgramData/Anaconda3/Lib/R/library");
#install.packages("AER",repos = "http://cran.us.r-project.org");
library(AER);
```

1.1 Instrumental Variable Model

The goal of this method is to estimate the causal effect of a predictor variable X on a target variable Y if the effect from X to Y is confounded. The idea of the instrumental variable approach is to account for this confounding by considering an additional variable I called an instrument. Although there exist numerous extensions, here, we focus on the classical case. We provide two definitions.

First, assume the following SCM

$$I := N_I \tag{1}$$
$$H := N_H \tag{2}$$
$$X := I\gamma + H\delta_X + N_X \tag{3}$$
$$Y := X\beta + H\delta_Y + N_Y. \tag{4}$$
$$\tag{5}$$

(All variables except Y could be multi-dimensional, in which case, they should be written as row vectors: $1 \times d$.) Here, I is called an instrumental variable for the causal effect from X to Y. It is essential that I effects Y only via X (and not directly).

Second, it is possible to define instrumental variables without SCMs, too. Let us therefore write

$$Y = X\beta + \epsilon_Y \tag{6}$$

(this can always be done). Here, ϵ_Y is allowed to depend on X (if there is a confounder H between X and Y, this is usually the case). We then call a variable I an instrumental variable if it satisfies the following three conditions: 1. $\text{cov}(X, I)$ is of full rank (relevance). 2. $\text{cov}(\epsilon_Y, I) = 0$ (exogenity). 3. $\text{cov}(I)$ is of full rank.

Informally speaking, these conditions again mean that I affects Y "only through its effect on X".

1.2 Estimation

We now want to illustrate how the existence of an instrumental variable I can be used to estimate the causal effect β in the model above. Let us therefore assume that we have received data in matrix form * $\mathbf{Y}$ - the target variable $n \times 1$ * $\mathbf{X}$ - the covariates $n \times d$ * $\mathbf{I}$ - the instruments $n \times m$

where $n > \max(m, d)$.

We now assume that I is a valid instrument (we come back to this question in Exercise 2 below). To estimate the causal effect of X on Y, there are several options of writing down the same estimator.

OPTION 1: The following estimator is sometimes called the generalized methods of moments (GMM)

$$\hat{\beta}_n^{GMM} := (\mathbf{X}^t\mathbf{I}(\mathbf{I}^t\mathbf{I})^{-1}\mathbf{I}^t\mathbf{X})^{-1}\,\mathbf{X}^t\mathbf{I}(\mathbf{I}^t\mathbf{I})^{-1}\mathbf{I}^t\mathbf{Y}$$

OPTION 2: we can use a so-called 2-stage least squares (2SLS) procedure. Step 1: Regress X on I and compute the corresponding fitted values $\hat{X}$. Step 2: Regress Y on $\hat{X}$. Use the regression coefficients from step 2.

The following four exercises go over some of the details of the 2SLS and apply it to a real data set.

1.2.1 Exercise 1

Assume that the data are i.i.d. from the following two structural assignments

$$\begin{aligned} Y &:= X \cdot \beta + \epsilon_Y \\ X &:= I \cdot \gamma + \epsilon_X, \end{aligned}$$

where X and I are written as $1 \times d$ and $1 \times m$ vectors, respectively. Here, ϵ_X and ϵ_Y are not necessarily independent, but the instrument I is assumed to satisfy the assumptions 1., 2., and 3. above.

a) Write down conditions on d and m that guarantee that $\hat{\beta}_n^{GMM}$ is well-defined (with probability one). You may assume that the sample versions $\mathbf{I}^t\mathbf{I}$ and $\mathbf{I}^t\mathbf{X}$ of instrumental varible condition 1) and 3) are of full rank.

** Hint: Prove that for a specific ordering of d and m (e.g. $d \leq m$, $d \geq m$ etc.) the matrix $\mathbf{X}^t\mathbf{I}(\mathbf{I}^t\mathbf{I})^{-1}\mathbf{I}^t\mathbf{X}$ inverted in the GMM estimator is positive definite, hence invertible. **

b) Prove that under these conditions that the GMM method is consistent towards the causal parameter, i.e., $\hat{\beta}_n^{GMM} \to \beta$ in probability.

** Hint: Use the model specification from above to write $\hat{\beta}_n^{GMM}$ as β plus a remainder term that is a linear function of ϵ_Y. Then use the continuous mapping theorem, which states that you may apply a continuous mapping on both sides of a convergence in probability expression, if the mapping in continuous in the limiting expression (Note that the inverse operator on matrices $M \overset{\text{Inv.}}{\mapsto} M^{-1}$ is continuous). Finally, use the fact that the product of two random sequences, each of which converge in probability to a constant, converges in probability to the product of these constants. **

c) Assume $d \leq m$. Prove that the methods 2SLS and GMM provide the same estimate.

1.2.2 Solution 1

1.2.3 End of Solution 1

For illustration, we use the CollegeDistance data set from [1] available in the R package AER.

```
In [6]: # load CollegeDistance data set
        data("CollegeDistance")
        head(CollegeDistance)
        # read out relevant variables
        Y <- CollegeDistance$score
        X <- CollegeDistance$education
        I <- CollegeDistance$distance
```

gender	ethnicity	score	fcollege	mcollege	home	urban	unemp	wage	distance	tuition	ed
male	other	39.15	yes	no	yes	yes	6.2	8.09	0.2	0.88915	12
female	other	48.87	no	no	yes	yes	6.2	8.09	0.2	0.88915	12
male	other	48.74	no	no	yes	yes	6.2	8.09	0.2	0.88915	12
male	afam	40.40	no	no	yes	yes	6.2	8.09	0.2	0.88915	12
female	other	40.48	no	no	no	yes	5.6	8.09	0.4	0.88915	13
male	other	54.71	no	no	yes	yes	5.6	8.09	0.4	0.88915	12

This data set consists of 4739 observations on 14 variables from high school student survey conducted by the Department of Education in 1980, with a follow-up in 1986. In this notebook, we only consider the following variables: * Y - base year composite test score. These are achievement tests given to high school seniors in the sample. * X - number of years of education. * I - distance from closest 4-year college (units are in 10 miles).

1.2.4 Exercise 2

Argue whether the variable I can be used as an instrumental variable to infer the causal effect of X on Y. Are there arguments, why it might not be a valid instrument? Hint: You can perform a linear regresssion and use the corresponding t-test p-value for significant regression coefficient. This is indeed identical to the t-test for the peasons correlation test-statistic.