Instrumental Variables

by Jonas Peters, Niklas Pfister, 06.01.2019

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 Lebesgue.

In [ ]:
```
library(AER)
```

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 \begin{align} I &:= N_I\\ H &:= N_H\\ X &:= I \gamma + H \delta_X + N_X\\ Y &:= X \beta + H \delta_Y + N_Y.\\ \end{align} (All variables except Y could be multi-dimensional, in which case, they should be written as row vectors: $1 \times d$.) If all variables are 1-dimensional, the corresponding DAG looks as follows. \begin{align} &\\ &\begin{array}{ccc} & & &H & \\ & &\overset{\delta_X}{\swarrow} & & \overset{\delta_Y}{\searrow}\\ & & & & \\ I &\overset{\gamma}{\longrightarrow} &X & \overset{\beta}{\longrightarrow} & Y\\ \end{array}\\ & \end{align} 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 \begin{equation} Y = X \beta + \epsilon_Y \end{equation} (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 two conditions:

1. $\operatorname{cov}(X,I)$ is of full rank (relevance)
2. $\operatorname{cov}(\epsilon_Y,I)=0$ (exogenity)
3. $\operatorname{cov}(I)$ is of full rank.

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

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

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.

Exercise 1

Assume that the data are i.i.d. from the following two structural assignments \begin{align*} Y &:= X \cdot \beta + \epsilon_Y \\ X &:= I \cdot \gamma + \epsilon_X, \end{align*} 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 $\hat{\beta}^{GMM}_n$ is well-defined (with probability one).

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

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

Solution 1

End of Solution 1

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

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

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).

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 regression in order to test if there is significant correlation.

Solution 2

In [ ]:
```
```

End of Solution 2

Exercise 3

Use 2SLS to estimate the causal effect of X on Y based on the instrument I. Compare your results with a standard OLS regression of Y on X (that includes an intercept). What happens to the correlation between X and the residuals in both methods? Which attempt yields smaller variance of residuals?

Solution 3

In [ ]:
```
```

End of Solution 3

A slightly different approach to 2SLS is to use the formula

OPTION 3: \begin{equation} \tag{1} \hat{\beta}_n = (\mathbf{I}^t \mathbf{X})^{-1} \mathbf{I}^t \mathbf{Y}. \end{equation}

This formula can be shown to be the same as OPTIONS 1 and 2 if $d = m$ (try proving it).

Exercise 4

Apply the above estimator (1) to `CollegeDistance` data and compare your result with the one from Exercise 3. (If you have included intercepts in the 2SLS, you need to replace the product moments by sample covariances.)

Solution 4

In [ ]:

End of Solution 4

References

[1] Kleiber, C., A. Zeileis (2008). Applied Econometrics with R. Springer-Verlag New York.

In [ ]:

In [ ]:

In [ ]: