Introduction to R and econometrics - Part II

Ryan Hawthorne
Acacia Economics
CCRED (UJ)

27 July, 2021

Objectives

By the end of this class, you should be able to:

- Understand causal effects, including using control variables and randomised experiments
- Run a linear model and instrumental variable regression in R
- Carry out hypothesis tests

Causal effects

Some methods to consistently estimate causal effects

We will discuss several methods that could be used to overcome endgeniety problems in order consistently estimate regression parameters that describe causal effects (like the slope of a demand function).

1. Conduct a randomized experiment.
2. Add control variables
3. Use instrumental variable estimation.

Conduct a Randomized Experiment

- The ideal method to estimate a causal effect is to run a randomized experiment. We have already illustrated this.
- Randomised experiments are often called the *Scientific Gold Standard* to establish causal effects. They are for example required by regulators when a pharmaceutical company wants to establish that a new drug has positive effects on patients.
- However, it is not always possible, or too costly, to run a randomized experiment. We thus learn the other approaches below.
- The methods below can also help if we run an experiment but have not achieved perfect randomization.

Control variables

Control variables: Motivating Example

Assume the demand function for ice is given by the following (long) regression formula with two explanatory variables:

$$q_t = \beta_0 + \beta_1 p_t + \beta_2 s_t + u_t$$

- s_t is a dummy variable that is 1 if the day is sunny and 0 otherwise and u_t are unobserved demand shocks.

Assume we estimate the (short) regression model:

$$q_t = \beta_0 + \beta_1 p_t + \varepsilon_t$$

Since we assume the data was generated by the long model above, it must hold that

$$\varepsilon_t = \beta_2 s_t + u_t$$

Add control variables: Multiple linear regression

- If we have data for s_t we estimate the (long) regression (OLS):

$$q_t = \beta_0 + \beta_1 p_t + \beta_2 s_t + u_t$$

- The OLS estimator of such a multiple linear regression (more than one explanatory variable) still satisfies our matrix formula:

$$\hat{\beta} = (X'X)^{-1}X'y$$

where the matrix X now has columns for explanatory variables:

$$X = \begin{pmatrix} 1 & x_1 & s_1 \\ \dots & \dots & \dots \\ 1 & x_T & s_T \end{pmatrix} = \begin{pmatrix} \mathbf{1} & x & s \end{pmatrix}$$

- If we're interested in a coefficient of a key variable (say β_1), additional explanatory variables are called *control variables*.

Exogeneity in a regression with control variables

- If we estimate the short regression:

$$q_t = \beta_0 + \beta_1 p_t + \varepsilon_t$$

where

$$\varepsilon_t = \beta_2 s_t + u_t$$

p is exogenous if it is uncorrelated with ε. This means p must be uncorrelated with *both* u and s.

- Assume we add s as control variable and estimate:

$$q_t = \beta_0 + \beta_1 p_t + \beta_2 s_t + u_t$$

- Here p is exogenous if it is uncorrelated with u, but it can now be correlated with s.
- By adding control variables, we remove factors from the error term, possibly making key variable exogenous.

Control by running regressions on subsets

- Consider again the ice cream demand function:

$$q = \beta_0 + \beta_1 p + \beta_2 s + u$$

where s is a dummy that is 1 if it is sunny and 0 otherwise. We assume s affects the price p but u is uncorrelated with p.

- Another way to control for s is estimating separate regressions, each using observations with the same value of s:
  - First, we only take the observations where $s_t = 0$ and estimate:

$$q = \beta_0^0 + \beta_1^0 p + u$$

  - Then we only take the observations where $s_t = 1$ and estimate

$$q = \beta_0^1 + \beta_1^1 p + u$$

- The slope estimates $\hat{\beta}_1^0$ and $\hat{\beta}_1^1$ of both regressions are consistent estimates of β_1 and the difference in estimated constants $\hat{\beta}_1^1$- $\hat{\beta}_1^0$ is a consistent estimate of β_2.

Heterogeneous effects and interaction terms

- So far we assumed that the causal effect of a one Euro price increase on demand is always the same value: β_1.
- Maybe on sunny days a price reduction has a stronger effect?
- If we estimate two separate regression for observations without sunshine ($s_t = 0$) and with sunshine ($s_t = 1$), we allow for different price effects, i.e. β_1^0 would be the price effect if there is no sunshine and β_1^1 the effect on sunny days.
- We can also estimate such heterogeneous price effects at once:

$$q = \beta_0 + \beta_1 p + \beta_2 s + \beta_3 (p \cdot s) + \varepsilon$$

  - The product $p \cdot s$ is called an interaction effect of p and s.
  - Now β_1 measures the price effect on non-sunny days where $s_t = 0$.
  - The coefficient β_3 of the interaction term measures by how much more the price affects demand if it is sunny $s_t = 1$ compared to non-sunny days $s_t = 0$.

Non-Linear Effects

- Besides interaction terms, we can also add non-linear effects - we could estimate a demand function with a quadratic effect of price that also depends on the weather:

$$q = \beta_0 + \beta_1 p + \beta_2 p^2 + \beta_3 s + \beta_4 (p \cdot s) + \beta_5 (p^2 \cdot s) + \varepsilon$$

- In principle, any non-linear function of the explanatory variables can be approximated with a linear regression.
- However, interpretation of the coefficients in specifications with non-linear terms and interaction effects is difficult (graphics can sometimes help though).
- Another problem is that estimators can become imprecise if we add many terms and don't have many observations or if some terms vary so similarly in the data that we don't have sufficient residual variation (*multicollinearity problem*).

Ice cream example without enough control variables

- Assume prices are affected by the demand shocks ε (eps) and we don't have any control variables for those demand shocks.
- It is the case that the cost c are uncorrelated with ε. But adding c as a control variable does not help. It does not solve the endogeneity problem.
- Yet, if we somehow could extract only the variation in the price that is caused by the cost variation, this variation would be uncorrelated with the demand shock ε. Can we use this to estimate β_1 consistently?
- Yes, we can. We have to use *instrumental variable estimation*. . .

Instrumental variables

Instrumental Variable Estimation

- Instrumental variable estimation (IV estimation) is a method to get consistent estimates when we have endogeneity problems that is very popular in economic research:
- An instrumental variable (short: instrument) z for an endogenous variable x is a variable that satisfies the following two conditions:
  - Relevance: z is correlated with the endogenous variable x:
  - Exogeneity: z is not correlated with the disturbance ε: $cor(z, \varepsilon) = 0$
- Per endogenous variable in the regression model, one needs at least one instrument that is not itself an explanatory variable in the regression model. (Sometimes this is called *exclusion restriction*)

Instruments in Ice Cream Example

- Consider the causal structure on the right and the demand function

$$q = \beta_0 + \beta_1 p + \beta_2 s + \varepsilon$$

- Check that both c and s are instruments for p (both satisfy the relevance and exogeneity condition)
- c is the required excluded instrument that is not part of the demand function.

IV-Estimation via "Two-Stage Least Squares"

- Can perform IV-estimation by running two OLS estimations.
- 1st Stage: Regress via OLS the endogenous variable on all instruments.

$$p = \gamma_0 + \gamma_1 c + \gamma_2 s + \eta$$

  - Then compute the *predicted values* of this regression

$$\hat{p} = \hat{\gamma}_0 + \hat{\gamma}_1 c + \hat{\gamma}_2 s$$

- 2nd Stage: Estimate the original regression but substitute the endogenous variable by the predicted values from stage 1.

$$q = \beta_0 + \beta_1 \hat{p} + \beta_2 s + u$$

  - The OLS estimator $\hat{\beta}$ of this second stage is a consistent estimator of β.

Analysis in R: IV estimation for ice-cream data

- We can perform IV estimation of the demand function by manually implementing the 2SLS approach.
- Or Use the function `ivreg` from the package `AER` to perform the instrumental variable estimation.
- Note that you get the same estimated coefficients for both approaches, but different standard errors:
  - The standard errors of the manual 2SLS approach are wrong, since the 2nd stage regression does not account for the uncertainty of the first stage regression.
  - The function `ivreg` yields the correct standard errors. Often in economics, we want *robust* standard errors. You can get them by using the function `iv_robust` from the package `estimatr` instead of `ivreg`.

Hypothesis testing

Hypothesis tests: Null hypothesis

- A hypothesis test consists of a **null hypothesis** H_0 and a corresponding **alternative hypothesis** H_1 about some features of a data generating process. Examples for hypotheses for a linear regression model:
  - H_0: $\beta_1 = 0$, $H_1 : \beta_1 \neq 0$
  - H_0: The explanatory variable x_k is exogenous, $H_1 : x_k$ is endogenous
  - H_0: The disturbance ε is not auto-correlated, H_1: ε is auto-correlated

Example: t-test for a regression coefficient I

- Consider a linear regression model $y = \beta_0 + \beta_1 x_1 + ... + \beta_K x_K + \varepsilon$ that satisfies a multiple regression equivalent to assumptions (A1)-(A4) & the null hypothesis:

$$H_0 : \beta_k = 0$$

- Every hypothesis test is based on a **test statistic** that can be computed from the data. In our example, it has *t-value*:

$$t_k = \frac{\hat{\beta}_k}{\hat{sd}(\hat{\beta}_k)}$$

Example: t-test for a regression coefficient II

- We can also view a test statistic as a random variable. Here t_k is a transformation of the random variable ε and the explanatory variables.
- Key of every hypothesis test is that one knows the distribution of the test statistic if H_0 and all additional assumptions (here A1-A4) hold true.
  - A statistical result shows that t_k is then distributed according to a t-distribution with $T - K - 1$ degrees of freedom if $\beta_k = 0$.

P-values and significance levels

- The p-value measures the probability to find the realized or more extreme test statistic if H0 is true (see plot above).
- One often considers critical levels of the p-value like 5% or 1%, which are called significance levels.
- We say we can reject the H0 at significance level α if the p-value is smaller than α,
  - e.g. if we have p-value=0.043 we can reject H0 at a significance level of 5%.
- Significance levels are often marked with one or several stars ** in regression outputs.

Testing for weak instruments I

- Consider a linear regression model of a demand function

$$q = \beta_0 + \beta_1 p + \beta_2 s + \varepsilon$$

  with endogenous prices p, an exogenous explanatory variables s.
- We also shall have two excluded instruments z_1 and z_2, e.g. two factors that influence costs and thereby prices.
- The weak instruments problem means that if the instruments z_1 and z_2 are only weakly correlated with p the IV estimator can become considerably biassed (and imprecise) for small sample size T.

Testing for weak instruments II

- The test for weak instruments shown in R tests the null hypothesis that in the first stage regression of the two stage least squares procedure

$$p = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 s + \eta$$

  the coefficients of the excluded instruments are zero, i.e. here:

  H0: $\gamma_1 = \gamma_2 = 0$

- This is a so called F-test and its test statistic is called F-statistic.
- Rule-of-thumb: Staiger and Stock (1997) suggested declaring instruments to be weak if the F-statistic is smaller than 10 (not looking at the p-value), Stock and Yogo (2005) provide much more details.

Wu-Hausman test for endogenous regressors

- Consider a linear regression model of a demand function

$$q = \beta_0 + \beta_1 p + \beta_2 s + \varepsilon$$

  for which we don't know if prices p are endogenous or exogenous.

- If we have valid instruments z for a possibly endogenous variable p, the *Wu-Hausman test* allows to test whether p is indeed endogenous.

- The null hypothesis of the Wu-Hausman test is that all explanatory variables of a regression are exogenous
  - i.e. low p-values of the Wu-Hausman test suggest an endogenous variable.

Sargan test for endogenous instruments

- The Sargan test is a test with the Null hypothesis that all instruments are exogenous.
- The Sargan test can only be applied if we have at least one more excluded instrument than endogenous variable.
- If the Sargan test is rejected (low p-value), it suggests that at least one instrument is endogenous.
- But: If the Sargan test is not rejected we do **not** have strong proof that all instruments are indeed exogenous, e.g. the Sargan test may well fail to detect if all instruments are endogenous.
    - This means not being rejected by the Sargan test can be interpreted as a neccessary condition for exogenous instruments but not a sufficient one. Most important remains the economic reasoning behind the selection of the instruments.

Using instrumental variable regression in R (1c)

You will learn how to:

- Perform IV estimations
- Analyse the results