Market Analysis with Econometrics and Machine Learning

1c Control Variables, Regression Anatomy and Instrumental Variable Estimation

Uni Ulm

Prof. Dr. Sebastian Kranz

SoSe 2020

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 this already illustrated in Chapter 1a.
- Randomized 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: 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

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

Questions (illustrated also in R):

a) Assume the prices p_t are uncorrelated with u_t but positively correlated with s_t. Do we get a consistent estimate of β_1 if we estimate the short regression via OLS?

b) What if p_t is uncorrelated with both u_t and s_t?

Add control variables: Multiple linear regression

- If we have data for s_t we can also estimate the (long) regression via 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 from above

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

  where the matrix X now has a column for each explanatory variable, i.e. here

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

- If one is mainly interested in the coefficient of one explanatory variable (say in β_1), the additional explanatory variables are often called *control variables*.

Analysis in R:

- Estimate the long regression in R and check that we estimate β_1 consistently even if p is correlated with s (but still uncorrelated with u.)

- 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 and thereby possibly make our explanatory variable variable of interest exogenous.

Regression Anatomy

The regression anatomy approach reduces a multiple linear regression to a simple linear regression, for which we can check assumptions A1-A4 and apply results of chapter 1b.

Consider a multiple regression:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_K x_K + u$$

The OLS estimator $\hat{\beta}_1$ is the same as the OLS estimator $\hat{\beta}_1$ from the simple linear regression:

$$y = \beta_0 + \beta_1 \tilde{x}_1 + \varepsilon$$

where $\tilde{x}_1$ are the residuals $\hat{\eta}$ from the linear regression of x_1 on all other explanatory variables

$$x_1 = \alpha_0 + \alpha_2 x_2 + \ldots + \alpha_K x_K + \eta$$

Regression Anatomy exemplified

Let us estimate β_1 from the multiple linear regression in two steps using the regression anatomy approach.

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

We first regress p on s:

$$p = \alpha_0 + \alpha_1 s + \eta$$

We then define $\tilde{p}$ as the *residuals* from that regression:

$$\tilde{p} \equiv \hat{\eta} = p - \hat{\alpha}_0 - \hat{\alpha}_1 s$$

We then estimate the simple linear regression:

$$y = \beta_0 + \beta_1 \tilde{p} + \varepsilon$$

Analysis in R:

Repeat the steps above in R and show that you indeed find the same estimate $\hat{\beta}_1$ in the simple linear regression of q on $\tilde{p}$ as in the long regression of q on p and s. (Note however, that the standard errors are wrong if we manually do these two steps.)

Regression Anatomy Interpretation

- The variable

  $$\tilde{p} \equiv \hat{\eta} = p - \hat{\alpha}_0 - \hat{\alpha}_1 s$$

  describes the residual variation of price that cannot be (linearly) explained by the control variable s.
- The estimate $\hat{\beta}_1$ from the simple regression

  $$q = \beta_0 + \beta_1 \tilde{p} + \varepsilon$$

  thus tells us how an increase of the price by one unit that is not due to the weather condition s affects the output.
- **Consistency** When assumptions A1 and A4 are satisfied for the simple regression of q on $\tilde{p}$, the regression anatomy result imples that $\hat{\beta}_1$ is also consistently estimated in the original multiple regression with control variables. So the crucial condition for consistency is that the residual variation $\tilde{p}$ of the price that cannot be explained by the control variables must be uncorrelated with the demand shock ε.

Which variation in the explanatory variable identifies the causal effect?

- It is often useful to think about which source of variation of an explanatory variable x *identifies* the causal effect of x on y. The meaning of those words is a bit loose, but I try to explain the idea.
- For a multiple regression it essentially means the following. Using the regression anatomy we end up with a simple linear regression

$$y = \beta_0 + \beta_1 \tilde{x} + \varepsilon$$

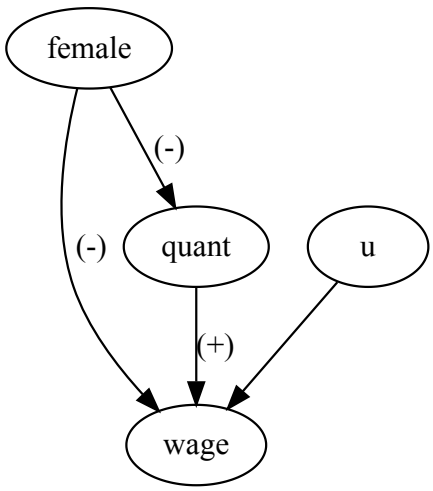
  where $\tilde{x}$ is the variation of x that cannot be explained by any control variable. For a convincing result, we should have some idea what are the remaining reasons (beyond the control variables) for x to vary. To convincingly solve endogeneity concerns, we also should argue why *all* of these remaining sources of variation are likely to be uncorrelated with ε.
- For example, a profit maximizing price p can vary due to changes in demand conditions and due to changes in costs. If we control for all variables that affect the demand conditions, the source of the residual variation of the price would be the variation in costs.

Which variables should be included as control variables?

- Which control variables to include in a linear regression is not always straightforward. It can depend on the exact question you want to answer.
- It is helpful to draw the supposed causal relationship with a graph.
- We want to discuss examples.

Effect of Gender on Wage with a Channel Variable

- Consider a world in which the `wage` of an individual i depends on whether i is `female` and has studied a quantitative subject `quant`, and on other independent unobserved factors `u`.
- The graph shows the causal relationships between the variables in that world:
  - Being female -ceteris paribus- leads to lower average wages and makes is less likely to study a quantitative subject.
  - Studying a quantitative subject -ceteris paribus- leads to higher average wages.
- Here `quant` is a *channel variable* for the effect of `female` on `wage`. Whether you want to include or exclude a channel variable in a regression depends on your precise question.



Effect of Gender on Wage with a Channel Variable

- In the short regression

$$wage_i = \alpha_0 + \alpha_1 female_i + \varepsilon_i$$

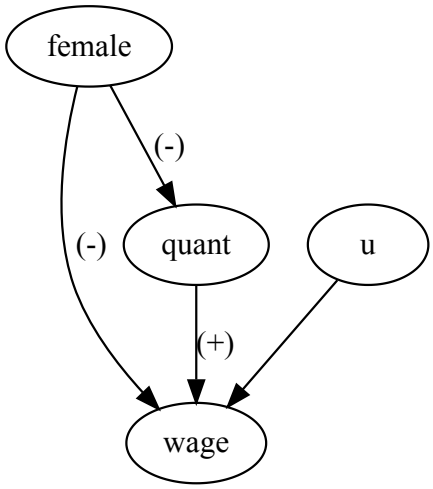
the coefficient α_1 estimates the total average wage difference between females and males including that part of the wage difference that arises from the channel that fewer females study a quantitative subject.

- In the long regression

$$wage_i = \beta_0 + \beta_1 female_i + \beta_2 quant_i + u_i$$

the coefficient β_1 describes the wage difference of a female to a male excluding the channel that less females study a `quant` subject.

- If we add the channel variable `quant` to the regression, we measure the wage differences of females that have the same value of `quant` as males.

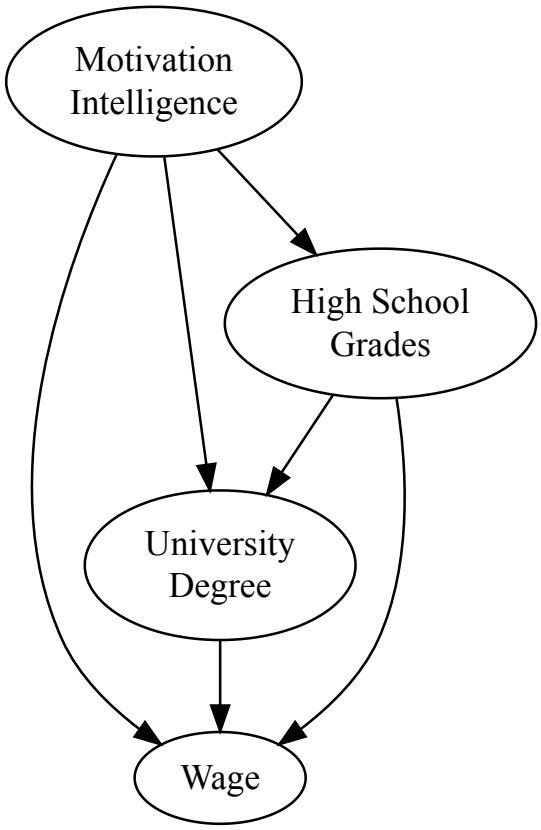


Quiz: Which control variables would you add?

- Assume you want to study the size of wage *discrimination* against females.
  - Would you add channel variables like having studied a quantitative subject to your regression?
- Assume you want to analyse the effect of obtaining a university degree on wages.
  - Would you add as control variable whether people later have a management position in their firm?
  - Would you add as control variable the high-school grades?

Side Remark

- It is extremely difficult (practically impossible) to overcome the endogeneity problem when estimating the causal effect of obtaining a university degree on wages just by adding control variables.
- There are a lot of practically unobservable factors that can affect both the decision to study at a university and later wages, e.g. intelligence, motivation, social skills (see figure on right).
- Moreover the effects of a university degree on wages can be very heterogenous and differ widely between students and subjects.



Quiz 2: Which control variables would you add?

- Assume a producer sells his product at a producer price p^p to stores and the stores set a retail price p^r for final customers. You are the producer and want to estimate how your demand by the stores depends on your producer price p^p.
  - Would you add the retail price p_r as a control variable to your regression?
- Assume you know that prices p depend on variables describing the demand conditions and on costs c. You also know that costs are uncorrelated with demand shocks. You want to estimate a demand function.
  - Would you add the variables describing the demand condition to your regression?
  - Would you add costs c to your regression?

Controlling by running separate regressions for subsets of the data

- 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.
- An alternative way to control for s is to estimate separate regressions, each using only 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. We will check this in R.

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.
- But maybe on sunny days a price reduction has a stronger effect than on other days?
- Recall the previous slide. 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 price effect on sunny days.
- We can also estimate such heterogeneous price effects with the following single regression for our whole sample:

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

  - 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 $s_t = 0$. The estimator $\hat{\beta}_3$ equals the difference $\hat{\beta}_1^1 - \hat{\beta}_1^0$ in the estimated slopes of the two separate regressions from the previous slide.

Non-Linear Effects

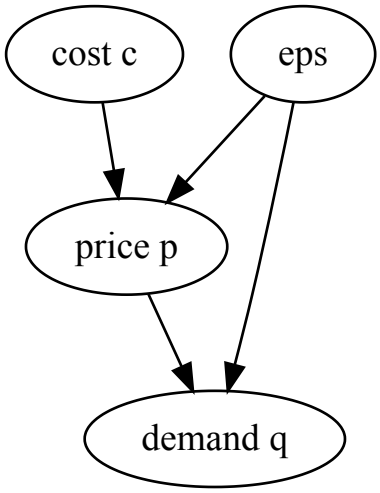
- Besides interaction terms, we can also add non-linear effects to regression equations. E.g. we could estimate a demand function with a quadratic effect of price that also depends on weather conditions:

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

- 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 quite imprecise if we add many terms and we don't have a lot of observations or if some terms vary so similarly in the data that we don't have sufficient residual variation after the first stage regression of a regression anatomy (*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 Variable Estimation

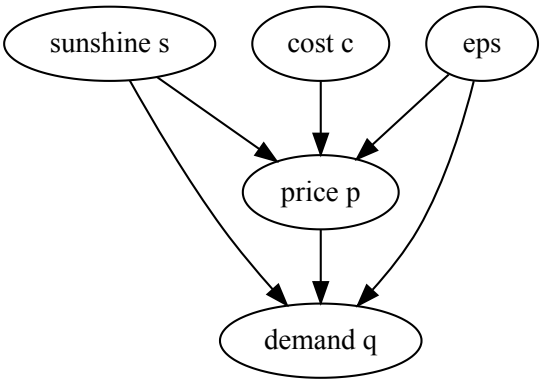
- Instrumental variable estimation (IV estimation) is a method to get consistent estimates when we have endogeneity problems that is very popular in economic research:
- An instrumental variable (short: instrument) z for an endogenous variable x is a variable that satisfies the following two conditions:
  - 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”

- One can perform IV-estimation by running two OLS estimations. We exemplify with the example of estimating the demand function.
- 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

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

  (Note the difference to the regression anatomy of control variables, where we computed $\tilde{p}$ as the *residual* of the first regression.)

- 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

- Run again your simulation of the ice cream demand with endogenous prices. Your simulation should include an additional explanatory variable `s` that affects prices and demand.
- Perform IV estimation of the demand function by manually implementing the 2SLS approach.
- 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`.