Introduction to R and econometrics

Ryan Hawthorne
Acacia Economics
Centre for Competition, Regulation & Economic Development (UJ)

06 March, 2021

Acknowledgement

This tutorial follows that developed by Prof. Sebastian Kranz (Ulm University) on market analysis with econometrics and machine learning:

- https://github.com/skranz/MarketAnalysis

Objectives

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

- Create summary statistics and plots in R using dplyr & ggplot2
- Understand the basics of linear regression models
- Consider endogeneity, and how to mitigate this problem
- Run a linear model and instrumental variable regression in R

Getting started

Download R, Latex and Github

We are going to use R, Latex and GitHub for reproducible research! It is free. Please download and install:

- R (econometrics software): https://cran.r-project.org/
- RStudio (works with R and Latex): https://rstudio.com/products/rstudio/download/#download
- Latex (I use Texlive) : https://www.tug.org/texlive/
- Sign up for Github (for version control): https://github.com/ and download Github desktop: https://desktop.github.com/
- You can always use rstudio.cloud if you are struggling with installing R, Rstudio and Latex

Additional resources

Additional resources for R can be found at:

- Learning the basics of R: https://r4ds.had.co.nz/
- Free online course on introduction to R: https://www.datacamp.com/
- Introduction to econometrics with R: https://www.econometrics-with-r.org/
- Learning microeconometrics with R: https://www.routledge.com/Learning-Microeconometrics-with-R/Adams/p/book/9780367255381
- Merger simulation tool using the antitrust package: https://daag.shinyapps.io/antitrust_shiny/

Getting started

Once you have installed everything, and signed up for the necessary:

- Go to https://github.com/ryanhawthorne/IOIntroduction, click on 'fork', then 'code', then 'Open with Github desktop'
- Open RStudio on your computer, click on 'new project', then browse to the folder that you just cloned the git repo into (in Windows, probably: /Documents/GitHub/IOIntroduction)

Trying out RMarkdown

In RStudio, once you have opened your project:

- Click on ‘file’, ‘new file’, ‘Rmarkdown’, ‘PDF’, and save it (‘tutorial1’)
- Click on ‘knit’
- RMarkdown is a great way to have your report text and econometrics code in the same file

Trying out Git

In RStudio:

- Head over to git on the top right hand side, select files you'd like to commit, click 'commit', add a message (e.g. 'first commit'), and then click 'commit', then 'push'
- Git is a version control system that you can use to roll back to any version of your work
- Commit frequently
- You may want to add a branch, perhaps called 'local', so that you are not working on the 'master' (production) version

You can have a look at your latest commit and push on github.com

Trying out the R console

- Type in '1+1', you should see:

```
## [1] 2
```

- R uses 'objects', stored by typing: 'clever_stuff <- 3' and then call it with: 'clever_stuff'

```
## [1] 3
```

- We often use lists in R, type in the following: my_first_list <- c(3,4), and then call it:

```
## [1] 3 4
```

- When you want to type in strings, use "": my_first_string <- c("this","that"), and call it:

```
## [1] "this" "that"
```

Loading your first dataset

To load the data as an object, type this into a new code chunk (hold down ctrl+alt+I) in your Rmarkdown script:

ice_cream_sales <- read.csv("ice cream sales.csv")

First have a look at the data, with: str(ice_cream_sales):

```
## 'data.frame':    200 obs. of  5 variables:
##  $ X: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ t: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ w: num  31.7 15.6 86.1 98.9 55 ...
##  $ p: num  26.8 26.3 26.1 23.1 27.7 ...
##  $ q: num  53 52.3 50.5 44.2 54.3 ...
```

You can always also: View(ice_cream_sales)

Libraries and packages

We are going to use the Tidyverse package in R, including ggplot2 for plotting graphs:

- install.packages("tidyverse") - this installs the R package
- library("tidyverse") - this loads the R package

We will also need the various other packages in "0 starthere.R", so please install and load them.

Estimating demand

We have just bought an ice-cream business!

We are trying to figure out how to price ice cream
We are provided with a nice price and sales history
The data, for each time period t:

- the price p_t for each ice-cream scoop
- the quantity q_t of ice-cream scoops sold
- the wholesale price w_t for ice-cream tubs

Summarise the data

Then summarise the data (in a new, separate chunk), with:

```
ice_cream_sales %>%
 select(w,p,q) %>%
 summary() %>%
 kable()
```

Where:

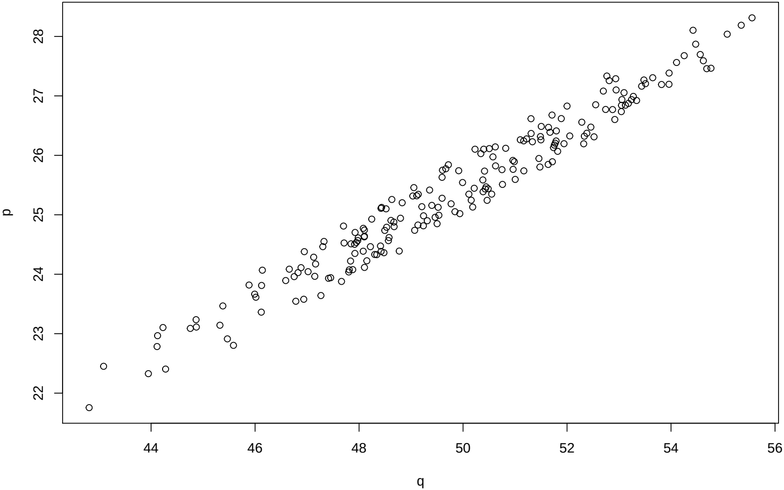
- The '%>%' is known as a 'pipe' - it means 'then'
- 'select' chooses variables we are interested in
- 'summary()' produces summary statistics
- 'kable()' produces nice tables

Summary result

	w	p	q
	Min. : 0.3912	Min. :21.76	Min. :42.81
	1st Qu.:23.9664	1st Qu.:24.46	1st Qu.:47.98
	Median :48.0086	Median :25.35	Median :49.74
	Mean :47.3499	Mean :25.38	Mean :49.81
	3rd Qu.:70.9175	3rd Qu.:26.32	3rd Qu.:51.76
	Max. :99.8161	Max. :28.31	Max. :55.56

Is this relationship between price & quantity correct?

Type in a new code chunk: plot(p ~ q, data = ice_cream_sales)



Some explanations for why price might +ve

Demand shocks might generate positive relationship

- Advertising may increase demand
- Better quality increase demand
- Price may also be affected by these demand shocks

How do we model demand for ice cream?

We want to be able to maximise profits. Lets assume demand takes the following form:

$$q_t = a_t - bp_t$$

Where:

- the market size parameter a_t is: $a_t = a_0 + \varepsilon_t$
- ε_t in turn is a random variable measuring a demand shock
- a_0 and b_0 are exogenous parameters, and $a_0 > 0, b_0 > 0$

How do we set prices for ice cream?

Our ice cream profit (assuming constant marginal costs, c_t):

$$\pi_t = p_t q_t(q_t) - c_t q_t(p_t) - F$$

Substituting $q_t = a_t - bp_t$ into the formula, profits are:

$$\pi_t = p_t(a_0 + \varepsilon_t) - bp_t^2 - c_t(a_0 + \varepsilon_t) + c_t b p_t$$

Differentiating profits with respect to price, setting equal to zero:

$$\frac{\partial \pi_t}{\partial p_t} = a_0 + \varepsilon_t - 2bp_t + c_t b = 0$$

Equlibrium prices and quantities

After rearranging, the equilibrium price is:

$$p_t = \frac{a_0 + \varepsilon_t}{2b} + \frac{c_t}{2}$$

Substituting this back into our demand function:

$$q_t = a_0 + \varepsilon_t - b(\frac{a_0 + \varepsilon_t}{2b} + \frac{c_t}{2})$$

Rearranging gives us our equilibrium quantity:

$$q_t = \frac{a_0 + \varepsilon_t}{2} - \frac{bc_t}{2}$$

Positive demand shock ($+a_0$) results in higher price & quantity

Elasticities in different models of demand

Model	Demand	Transformed demand	Own-price elasticity
Linear	$Q = \alpha + \beta P + \gamma X$	$Q = \alpha + \beta P + \gamma X$	$\varepsilon = \beta * P/Q$
Semi-log	$Q = \exp(\alpha + \beta P + \gamma X)$	$lnQ = \alpha + \beta P + \gamma X$	$\varepsilon = \beta * P$
Log-log	$Q = \exp(\alpha) P^{\beta} X^{\gamma}$	$lnQ = \alpha + \beta lnP + \gamma lnX$	$\varepsilon = \beta$

Understanding prices and quantities in R (1a)

- We will now consider how demand can be estimated in R

You will learn how to:

- Run a random simulation in R
- Perform a basic linear regression using ordinary least squares
- Start to appreciate the endogeneity problem in demand estimation

Econometrics

Simple linear regression model

- A simple linear regression model satisfies the following relationship for all observations $t = 1, ..., T$

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

- $y = (y_1, ...y_T)$ is the dependent variable.
- $x = (x_1, ...x_T)$ is the explanatory variable.
- $\varepsilon = (\varepsilon_1, ..., \varepsilon_T)$ (epsilon) is a random variable that describes unobserved influences on y, sometimes called *disturbance*. We will typically make some assumptions on the distribution of ε. We will also use the letters u and η (eta) to denote disturbances.
- $\beta = (\beta_0, \beta_1)$ is the vector of true coefficients.

Estimates, predicted values, resideuals

- Let $\hat{\beta}$ be an *estimate* of the true parameter vector β.
- The *predicted values* (also called fitted values) of y are given by

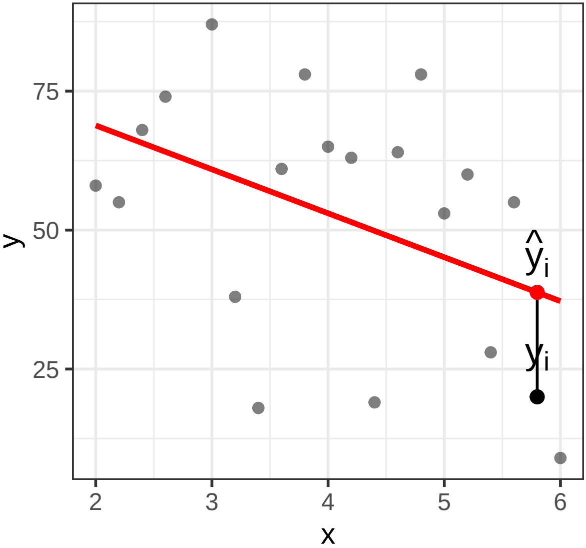
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- The *residuals* (estimated values of the disturbance) are given by

$$\hat{\varepsilon} = y - \hat{y} = y - \hat{\beta}_0 - \hat{\beta}_1 x$$

  - The residuals $\hat{\varepsilon}$ are close to the true disturbances ε if our estimate $\hat{\beta}$ is close to the true parameters β.

Considering our ice cream seller



Ordinary least squares minimises this disturbance term

Ordinary Least Squares Estimation (OLS)

- An ordinary least squares (OLS) estimate minimises the sum of squared residuals

$$\hat{\beta} = \arg\min \sum_{t=1}^{T} \hat{\varepsilon}_t^2$$

- For the simple linear regression (one explanatory variable), the OLS estimator $\hat{\beta}_1$ has the following formula

$$\hat{\beta}_1 = \frac{Cov(x_t, y_t)}{Var(x_t)} = cor(x_t, y_t)\frac{sd(y)}{sd(x)}$$

where *cor* denotes an empirical correlation and *sd* an empirical standard deviation for our sample data.

Linear regression model in matrix notation

- One often writes a linear regression model in matrix notation:

$$y = X\beta + \varepsilon$$

with

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

- The OLS estimator $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ is then given by

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

Estimators and estimates

- Since $y = X\beta + \varepsilon$, the OLS estimator can be rewritten as

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'y \\ &= (X'X)^{-1}X'(X\beta + \varepsilon) \\ &= \beta + (X'X)^{-1}X'\varepsilon\end{aligned}$$

- This means $\hat{\beta}$ is a linear transformation of the true parameters β and the disturbance ε
- As a function of a random variable ε the OLS *estimator* $\hat{\beta}$ is itself a random variable
- The OLS *estimate* $\hat{\beta}$ is a realisation of the OLS estimator, i.e. the value for particular draws of ε and X.
- To understand what econometrics, one should keep in mind that *an estimator is a random variable*.

Standard Error of OLS estimator

- In a simple linear regression (one explanatory variable)

$$y = \beta_0 + \beta_1 x + \varepsilon$$

  where the ε are independently, identically normal distributed, the standard deviation of the OLS estimator $\hat{\beta}_1$ can be *estimated* by

$$se(\hat{\beta}_1) = \hat{sd}(\hat{\beta}_1) = \frac{1}{\sqrt{T}} \frac{sd(\hat{\varepsilon})}{sd(x)}$$

- We call this estimate of the standard deviation the *standard error* of $\hat{\beta}_1$.

- Observations: We can estimate β_1 more precisely if we have...
  - a larger sample size T
  - more variation in x (higher standard deviation).

Analysis in R: We run a linear regression with `lm` and call `summary`

Robust Standard Errors

- There is also a matrix formula to compute the standard errors for all $\hat{\beta}$ that can also be used for multiple linear regressions with more than one explanatory variable.
- If the ε are not identically, independently normal distributed, one should use appropriate *robust* standard errors. Most empirical papers in economics use some robust standard errors.
- We don't explain robust standard errors further in this course. Just note that in R a convenient way to use robust standard errors is the function `lm_robust` in the package `estimatr` or the function `felm` in the package `lfe`.

Criteria for estimators: Bias

- **Bias:** Recall that an estimator $\hat{\beta}$ is a random variable since it depends on the realizations of ε. Let $E\hat{\beta}$ be the expected value of $\hat{\beta}$. The bias of $\hat{\beta}$ measures a systematic over- or underestimation of $\hat{\beta}$ compared to β:

$$Bias(\widehat{\beta}) = E\hat{\beta} - \beta.$$

- **Unbiasedness:** An estimator $\hat{\beta}$ is unbiased if its Bias is 0, i.e.

$$E\hat{\beta} = \beta$$

Criteria for estimators: Standard Deviation

- For two unbiased estimators of β_i, one would typically prefer an estimator with a lower standard deviation $sd(\hat{\beta}_i)$ (or equivalently the one with the lower variance $Var(\hat{\beta}_i)$)

Criteria for estimators: Mean Squared Error

- **Mean squared error**: The mean squared error of $\hat{\beta}_i$ is given by

$$\begin{aligned} MSE(\hat{\beta}_i) &= E(\hat{\beta}_i - \beta_i)^2 \\ &= Bias(\hat{\beta}_i)^2 + Var(\hat{\beta}_i) \end{aligned}$$

Criteria for estimators: Consistency

- An estimator $\hat{\beta}$ is (strongly) **consistent** if its MSE converges to 0 as the sample size T grows large

$$\lim_{T \to \infty} MSE(\hat{\beta}) = 0.$$

- Estimated parameters $\widehat{\beta}$ converge (in probability) to true β

$$\operatorname*{plim}_{T \to \infty} \widehat{\beta} = \beta$$

- **Consistency: the most important requirement for an estimator**
- If an estimator is inconsistent that is typically because it is biased and the bias does not go away as $T \to \infty$.

Criteria for estimators: Efficiency

- An estimator $\hat{\beta}$ is **efficient** (within a specified class of estimators) if there is no other estimator that has a lower mean squared error.

Assumptions of the simple linear regression model

- We now state a series of assumptions for the simple linear regression model (one explanatory variable).
  - A1: $E(\varepsilon_t|x) = 0$
  - A2: The ε_t are identically and independently distributed.
  - A3: The ε_t are normally distributed
  - A1-A3 are often compactly written as $\varepsilon_t \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
  - A4: The explanatory variable x must have positive variance and be deterministic or a stationary random variable. (We don't discuss what stationary means in this course, but you can look it up on Wikipedia.)
- If all assumptions are satisfied, the OLS estimator $\hat{\beta}$ will be consistent, unbiassed and efficient.

The main assumption (A1)

- **A1** No matter which values of x we observe, the conditional expected value of ε_t is always zero:

$$E(\varepsilon_t|x) = 0$$

- The important thing is not the 0 on the right. If it were a positive or negative value, we could always redefine the constant β_0 to make it 0.
- The important thing is that the expected value of ε_t does not depend on x. This means knowing x shall give us no information about the expected value of ε_t.
- In our ice cream example with profit maximizing prices this condition is violated. Higher demand shocks lead to higher prices. This means if we observe a high price, we expect that there was a positive demand shock ε_t.

Exogenous and Endogenous Variables

- We say the explanatory variable x is **exogenous**, if it is uncorrelated with ε

$$\text{cor}(x_t, \varepsilon_t) = 0$$

- We say x is **endogenous** if $\text{cor}(x_t, \varepsilon_t) \neq 0$
- Condition A1 $E(\varepsilon|x) = 0$ can only be satisfied if x is exogenous.
- We will typically just check whether x is exogenous, even though A1 is a stronger condition. A1 is sometimes called *strong exogeniety*. In all examples studied in this course, exogeniety of x implies that also A1 holds.

A2, No auto-correlation and no heteroskedasticity

- **A2** The ε_t are identically and independently distributed.
- Typical violations of A2:
  - auto-correlation: demand shocks may be persistent across periods
  - heteroskedasticity: the variance of ε_t can depend on the explanatory variable (this alone does not yet mean that A1 is violated)
- A2 is moderately important. If violated, the OLS estimator $\hat{\beta}$ is still consistent but not efficient. One must calculate standard errors using an appropriate formula for robust standard errors.
- We don't study violations of A2 in this course.

A3: Normally distributed disturbances

- **A3** ε_t is normally distributed
- It is nice if A3 holds, but it is not crucial. Even if A3 is violated, the OLS estimate $\hat{\beta}$ is the best unbiased linear estimators of β (Gauss-Markov Theorem). Significance tests would only be asymptotically correct.
- If A1-A3 (and the other assumptions) holds, $\hat{\beta}$ coincides with Maximum Likelihood estimator and is efficient.

95% Confidence Intervals

- If assumptions A1 holds (no endogeniety problem) then with approximately 95% probability we find an estimate $\hat{\beta}_i$ such that the interval of plus-minus 2 standard errors around $\hat{\beta}_i$ contains the true parameter β_i.
- We call this interval

$$[\hat{\beta}_i - 2 \cdot se(\hat{\beta}_i) \; ; \; \hat{\beta}_i + 2 \cdot se(\hat{\beta}_i)]$$

the approximate **95% confidence interval**.

“Bias Formula”

- Consider a simple linear regression

$$y = \beta_0 + \beta_1 x + \varepsilon.$$

  and assume we would observe ε.

- One can show that

$$\hat{\beta}_1 - \beta_1 = cor(x, \varepsilon)\frac{sd(\varepsilon)}{sd(x)}$$

  using the sample correlations and sample standard deviations.

- This expression is an estimator of the bias of $\hat{\beta}_1$. (The actual bias is the expected value of it.)
- Thus essentially the bias has the same sign as the correlation between x and ε.

Understanding bias and endogeneity in R (1b)

You will learn how to:

- Develop a Monte-Carlo simulation
- Analyse the results
- Consider the endogeneity problem in demand estimation

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: 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 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`.

Using instrumental variable regression in R (1c)

You will learn how to:

- Perform IV estimations
- Analyse the results