Econometrics: Endogeneity and Instrumental Variables

Diego López Tamayo * Based on MOOC by Erasmus University Rotterdam

Contents

"There are two things you are better off not watching in the making: sausages and econometric estimates." -Edward Leamer

*El Colegio de México, diego.lopez@colmex.mx

Endogeneity

What is endogeneity?

Ordinary least squares (OLS) is a great tool to uncover relationships in economics and business. But we must be aware that this tool does not always work. There are circumstances where OLS breaks down. These circumstances relate to the **difference between correlation and causality**. Luckily, econometrics also has the solution. But we before we discuss this, let's consider a motivating example.

Suppose we want to explain:
- the monthly number of departing flights at an airport (y) - using the number of travel insurances sold in the month before. (x)

What kind of relationship would you expect if you regress flights as the variable y on a constant, and insurances as the variable x? Most likely we will obtain a positive relationship. Suppose OLS yields:

$$y = 10,000 + .25x + e$$

How should we interpret the obtained coefficients? What does the estimate .25 really mean? Suppose we have 4,000 travel insurances sold in the month before:

- **Correct:** 4, 000 insurances sold $\rightarrow$ expected number of flights $= 10,000 + .25Ö4,000 = 11,000$. Because High x tends to go together with high y. The identified correlation yields adequate predictions. Note that this statement merely relies on a correlation.
- **Incorrect:** Selling 4,000 additional insurances causes $.25Ö4,000 = 1,000$ additional flights. The regression does not identify a causal impact! A third variable (travel demand) affects y (flights) and x (insurances).

This example shows that we cannot always interpret least squares estimation results, as causal effects. However, identifying causal effects is one of the main goals of econometrics.

Ordinary least squares requires some assumptions for it to correctly estimate causal effects. One important assumption is that **explanatory variables are exogenous**. The violation of this assumption is called endogeneity.

In the following sections you will:

- Understand/recognize endogeneity
- Know the consequences of endogeneity
- Estimate parameters under endogeneity
- Know the intuition of the new estimator
- Test assumptions underlying this new estimator

Sources of endogeneity

Let us start by studying the source of endogeneity.

The formal assumption that we violate is the assumption that explanatory variables X in the linear model are non-stochastic. (Assumption A2) Explanatory variables are non-stochastic.

Literally speaking, non-stochastic means that if you would obtain new data only the y values would be different and the values for X would stay the same. This is like a *controlled experiment* where the researcher determines the experimental conditions coded in X. This assumption is crucial for the OLS estimator to be consistent. Consistent means that the estimator b converges to the true coefficient beta when the data set grows larger and larger. $b \rightarrow \beta$ for $n \rightarrow \infty$.

In economics however, controlled experiments are rare. X variables are often the consequence of an economic process, or of individual decision making. In our example, the travelers together determine the number of insurances sold. From the researcher's point of view, the X variables should therefore be seen as stochastic.

Once we allow X to be stochastic, we acknowledge that we would get different X values in a new data set. And if variables are stochastic, they can also be correlated with other variables, even with variables that are not included in the model!

In the context of our example, the number of insurances will be correlated with the travel demand. Although travel demand is difficult to observe and not included in the model, it does influence the number of flights. In the model, travel demand is therefore part of the error term epsilon. As a consequence, the X variable, insurances sold, is correlated with the error term ϵ.

- If X is endogenous $\rightarrow$ there is another variable(s) that affect y and X.
- OLS does not properly estimates β (inconsistent)

Usually, this correlation is due to an omitted factor.

Now let's consider three possible sources of endogeneity in more detail.

1. Endogeneity is often due to an **omitted variable**. In our example, the omitted variable was travel demand. Let's consider this situation formally.

Suppose that the true model for a variable y contains two blocks of explanatory variables, X1 and X2. And that in this true model, all assumptions are satisfied $y = X_1\beta_1 + X_2\beta_2 + \eta$ but we do not observe X_2 and perform OLS on $y = X_1\beta_1 + \epsilon$. The error term in the second model is:

$$\epsilon = X_2\beta_2 + \eta$$

From this relationship we can see that in the second model X_1 will be correlated with epsilon if X_1 and X_2 are correlated and β_2 does not equal 0: $Cov(X_1, X_2)\beta_2 \neq 0$ notice that $Cov(X_1, \eta) = 0$ due to orthogonality.

$$Cov(X_1, \epsilon) = Cov(X_1, X_2\beta_2 + \eta) = Cov(X_1, X_2)\beta_2 + Cov(X_1, \eta)$$

When thinking about whether certain variables in a model are endogenous, it is good to think about potential omitted variables. If you can think of an omitted variable that is related to the included variables, and the dependent variable, you will have endogeneity.

Suppose we run a regression to explain a student's grade using only the number of attended lectures. What omitted variable leads to endogeneity here? There are many possible ommited factors:

- Difficulty of exam? Probably NOT correlated with attendance.
- Motivation of the students? Probably correlates with attendance and affects grade.
- Compulsory attendance yes/no? Does not directly impact the grade

The omission of the motivation of students does lead to endogeneity. Highly motivated students are likely to attend many lectures and obtain high grades. So a regression of grades on attendance will not show the true impact of attendance. It will partly capture the unobserved motivation as well.

2. A second cause of endogeneity is **strategic behavior**.

Consider a model in which you explain the demand for products using only its price. If the salesperson strategically sets high prices when a high demand is expected, high demand will often go together with high prices! A simple regression may then yield a positive price coefficient. This is of course not the true impact of price. **Price is endogenous in this regression** as it correlates with the market information, which in turn, determines demand.

3. A third reason for endogeneity, is **measurement error**.

Suppose that we have a variable y, say, salary, That depends on a factor that is difficult to measure. For example, intelligence. Let's denote the intelligence by x^*. We can obtain a noisy measurement of intelligence, for example through an IQ test. The test score is called x and is equal to the true intelligence plus the measurement error.

$$x = x^* + measurementerror$$

To summarize, endogeneity is a common and serious challenge in econometrics as OLS is not useful under endogeneity.

Formalizing endogeneity

We will show that such measurement error leads to endogeneity in a model that explains why using the IQ test score x in the salary example:

We want to explain the income yi of an individual $i = 1, ..., n$ using the individual's intelligence x_i^*. Suppose that the true relationship between these two variables is:

$$y_i = \alpha + \beta x_i^* + u_i$$

where β gives the impact of intelligence on income. Furthermore, suppose that this model satisfies all the standard assumptions of the linear model. However, the intelligence x_i^* cannot be observed directly. We can only observe a test score that equals the true intelligence plus a measurement error, that is $(1)x = x_i^* + w_i$ The measurement error process w_i satisfies the following conditions:
- Mean zero $E(w_i) = 0$ - Constant variance $Var(w_i) = \sigma_w^2$ - Zero correlation across individuals: $Cov(w_i, w_j) = 0$ $\forall i \neq j$ - Uncorrelated with unexplained income and true intelligene: $Cov(w_i, u_i) = 0$ and $Cov(w_i, x_i^*) = 0$

We have data on (y_i, x_i) for $i = 1, ..., n$ Suppose we ignore measurement error and simply apply OLS to:

$$y_i = \alpha + \beta x_i + \epsilon_i$$

By definition $\epsilon_i = -\beta w_i + u_i$ we can show this just by equalizing the true relation and the estimated model:

$$\alpha + \beta x_i^* + u_i = \alpha + \beta x_i + \epsilon_i$$

Which can be rewritten as:

$$\epsilon_i = \beta(x_i^* - x_i) + u_i = -\beta w_i + u_i$$

Using this $(1)x = x_i^* + w_i$ equation and $(2)\epsilon_i = -\beta w_i + u_i$ we can show that the covariance between x_i and ϵ_i is $Cov(x_i, \epsilon) = -\beta\sigma_w^2$

$$Cov(x_i, \epsilon) = Cov(x_i^* + w_i, -\beta w_i + u_i) = \beta Cov(x_i^*, w_i) + Cov(x_i^*, u_i) - \beta Cov(w_i, w_i) + Cov(w_i, u_i)$$

Where we know by definition that $Cov(x_i^*, w_i) = 0$, $Cov(x_i^*, u_i) = 0$ and $Cov(w_i, u_i) = 0$

$$Cov(x_i, \epsilon) = -\beta\sigma_w^2$$

So x_i is endogenous if the $Cov(x_i, \epsilon) \neq 0$ in this case using the last resutl, it means that the variance of the measurement error $\sigma_w^2 \neq 0$ and that the true impact of inteligence on income $\beta \neq 0$.

Consequences

We have discussed three main causes of endogeneity: omitted variables, strategic behavior by people in a market, and the presence of measurement error in an explanatory variable. All three lead to a correlation between explanatory variables X, and the unexplained part in the econometric model, epsilon. This violates the standard assumptions underlying OLS estimation. But, .. how bad is this?

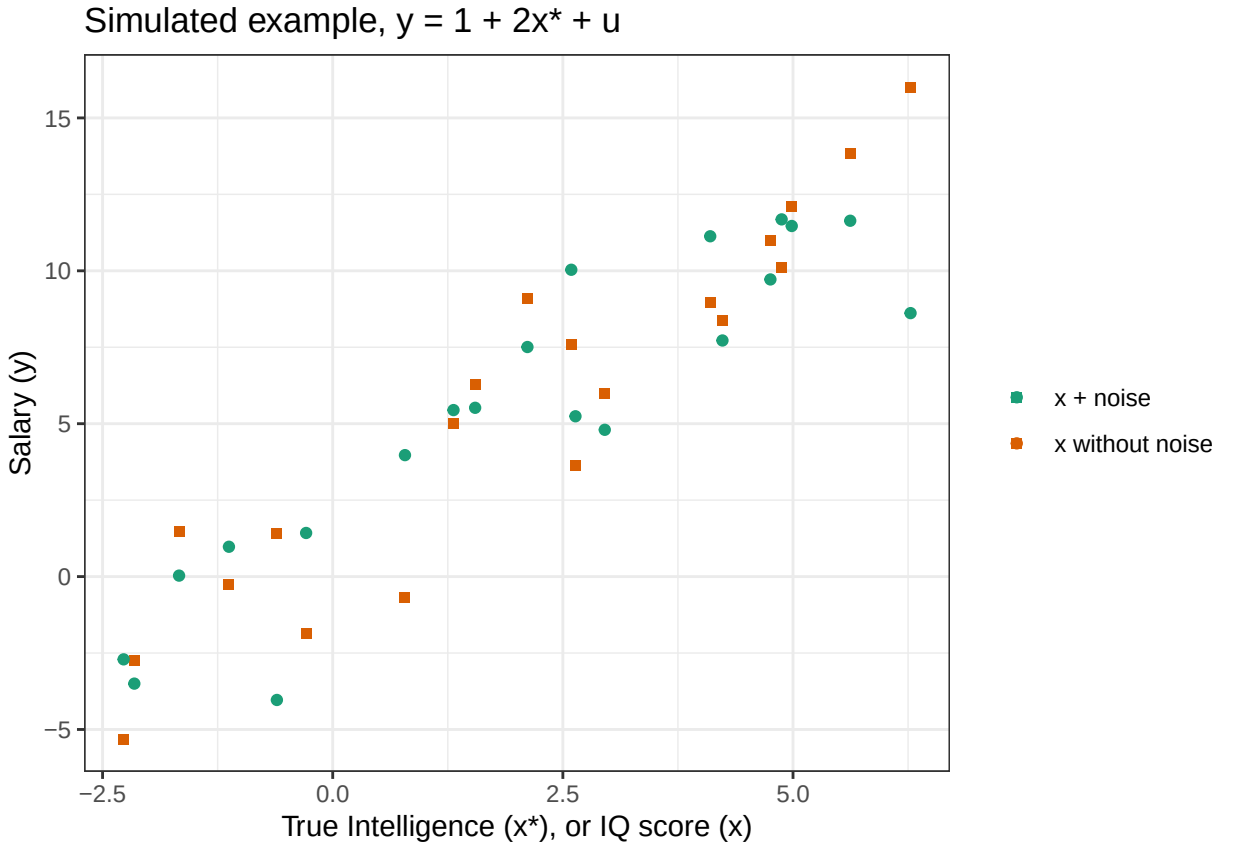
Let's reconsider the measurement error example where salary, denoted by y, depends on intelligence, denoted by x-star. However, in practice we cannot observe intelligence and can only get a noisy measurement, say an

IQ score. The noisy measurement is denoted by x. As an illustration, we will use hypothetical data. where we randomly generate intelligence, x^*, and generate $y = 1 + 2x^*$. The IQ score x is generated as $x = x^* + noise$.

Here you see a scatter between y and intelligence x^*. In this new graph, I add a scatter of y versus the IQ score x using orange squares.

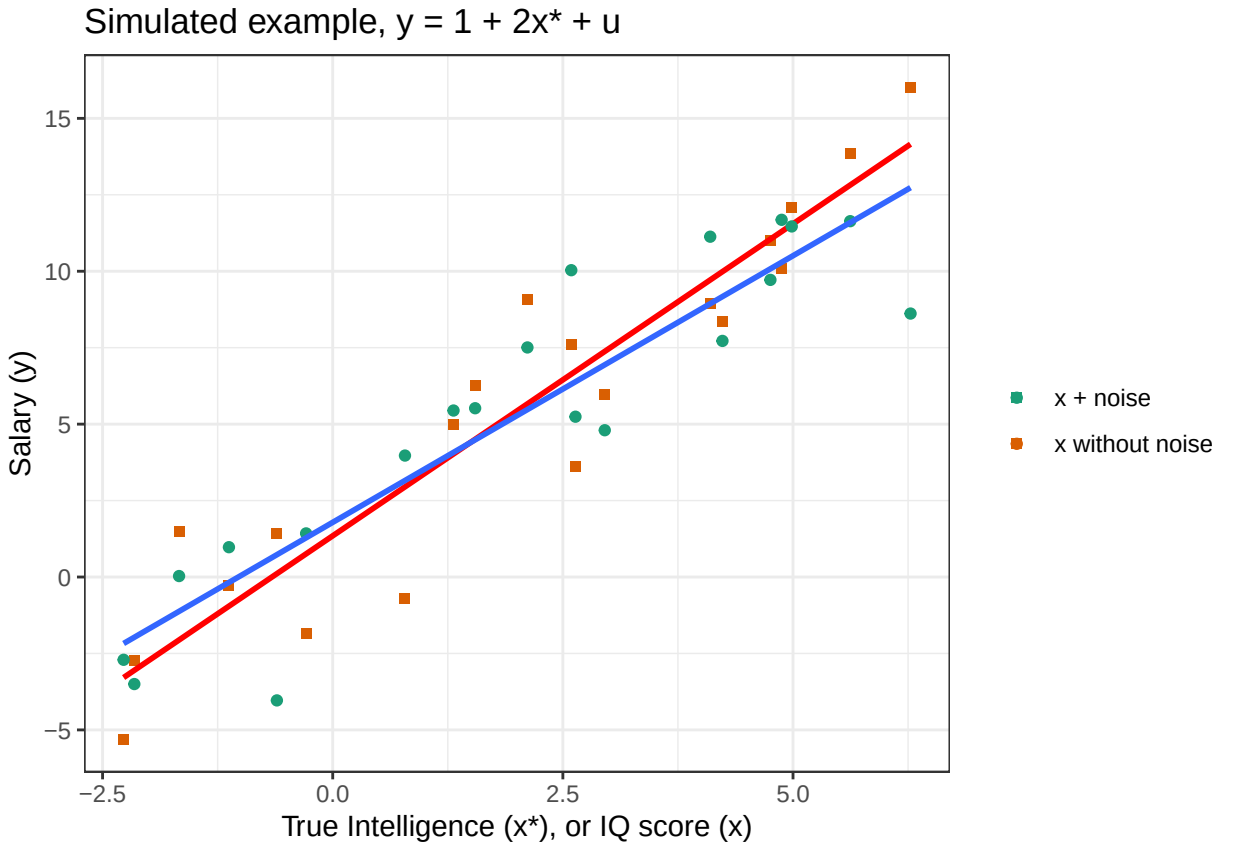
We create the random variables for our example:

```
set.seed(124)
n <-20
x_1 <- rnorm(n, mean = 2, sd = 3)
x_star <- rnorm(n, mean = 2, sd = 3) + rnorm(n, mean = 0, sd = 4)
y_1 <- c(1+2%*%x_1 + rnorm(n, mean = 0, sd = 2))
y_star <- c(1+2%*%x_1 + rnorm(n, mean = 0, sd = 2))
sample <- tibble(x_1,x_star,y_1,y_star)
#str(sample)
plot1 <- ggplot(data=sample, aes(x=x_1)) +
  geom_point(aes(y=y_1, col = "x without noise"),shape=15) +
  geom_point(aes(y=y_star, col = "x + noise")) +
  labs(x = "True Intelligence (x*), or IQ score (x)", y = "Salary (y)",
  title = "Simulated example, y = 1 + 2x* + u") +
  theme_bw() +
  scale_color_brewer(name= NULL, palette = "Dark2")
plot1
```



In practice, we would only have these orange squares as data. The OLS regression, through this cloud of points, is given by this blue line. However, this is not the line we want to have. The true effect of intelligence on salary is stronger (steeper)! This can clearly be seen by the regression line through the green dots. This red line shows the true effect we would like to estimate!

```r
plot2 <- ggplot(data=sample, aes(x_1,y_1)) +
  geom_point(aes(y=y_1, col = "x without noise"),shape=15) +
  geom_smooth(method = lm, se = FALSE, colour="red") +
  geom_point(aes(y=y_star, col = "x + noise")) +
  geom_smooth(formula = y_star ~ x, method = lm, se = FALSE) +
  labs(x = "True Intelligence (x*), or IQ score (x)", y = "Salary (y)",
  title = "Simulated example, y = 1 + 2x* + u") +
  theme_bw() +
  scale_color_brewer(name= NULL, palette = "Dark2")
plot2
```



If we use noisy x variables, we obtain the wrong coefficients. This also hold on the endogeneity in general. Can we say anything about the sign of the difference between the true and the estimated effect in case of measurement error?
Under measurement error, OLS is biased towards zero. The estimated line (blue) is not steep enough (as red). As a result, points on the left of the scatter are likely due to negative measurement errors. While points on the right are likely due to positive errors. In other words, measurement error stretches the cloud of points horizontally. This results in a flatter regression line.

> This example illustrates that OLS is biased under endogeneity. However, we only looked at one particular data set with a small number of observations. Would it help to have more data points or different data sets?

Let's consider what happens if we repeat the same experiment many times and for differently sized data sets. For each repetition, we generate a new data set and of course get different estimates. Even for a very large data set, we do not get close to the correct value of the slope parameter.

Things are very different if we would have the noise-free explanatory variable available. In the context of our

example, we do as if we can perfectly measure intelligence. It is clear that for all sizes of the data set, OLS on average gives the correct value. And for large data sets, it almost exactly gives the correct value.

If the assumptions underlying OLS are satisfied, OLS is unbiased and consistent.

> If X is endogenous, when n grows the OLS estimator converges to the wrong value. OLS is inconsistent.

We can also show this inconsistency mathematically. Let's consider the standard linear model $y = X\beta + \epsilon$ in combination with the OLS estimator $b = (X'X)^{-1}X'y$. For the y in the formula, we insert the model definition, and next, work out the matrix multiplications.

$$b = (X'X)^{-1}X'(X\beta + \epsilon) = b = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\epsilon = \beta + (X'X)^{-1}X'\epsilon$$

In the resulting equation, you can see that the first term reduces to beta. So, we can split the OLS estimator into beta, plus a random term that depends on X and epsilon. We use this formulation to see what happens to the estimator when the sample size gets very large.

The first part, beta, is constant, so we only need to study what happens to the second part. Both the term X'X and the term X'epsilon have sums over the observations as elements:

$$X'X = \begin{pmatrix} \sum_{i=1}^{n} x_{1i}^2 & \sum_{i=1}^{n} x_{1i}x_{2i} & \dots & \sum_{i=1}^{n} x_{1i}x_{ki} \\ \sum_{i=1}^{n} x_{2i}x_{1i} & \sum_{i=1}^{n} x_{2i}^2 & \dots & \sum_{i=1}^{n} x_{2i}x_{ki} \\ \dots & \dots & \dots & \dots \\ \sum_{i=1}^{n} x_{ki}x_{1i} & \sum_{i=1}^{n} x_{ki}x_{2i} & \dots & \sum_{i=1}^{n} x_{ki}^2 \end{pmatrix}$$

$$X'\epsilon = \begin{pmatrix} \sum_{i=1}^{n} x_{1i}\epsilon_i \\ \sum_{i=1}^{n} x_{2i}\epsilon_i \\ \dots \\ \sum_{i=1}^{n} x_{ki}\epsilon_i \end{pmatrix}$$

If the number of observations increases, these terms will therefore diverge as $n \to \infty$ However, we can rewrite the estimator such that we are left with terms that do converge. In this equation, we have inserted two $\frac{1}{n}$ terms that cancel against each other.

$$b = \beta + (\frac{1}{n}X'X)^{-1}(\frac{1}{n}X'\epsilon)$$

Notice that $(\frac{1}{n}X'X)^{-1}$ is an average, in general converges to, say Q. The population mean.

The result is that the two matrices now have elements that are averages over the observations. Under mild condition the term $(\frac{1}{n}X'X)^{-1}$ now converges to it's population mean, Q. The second term $(\frac{1}{n}X'\epsilon)$ is also an average over observations and converges in general. The OLS estimator, b, will now **converge to the true parameter beta if three conditions are true**:

$b \to \beta$ as $n \to \infty$ if: 1. $(\frac{1}{n}X'X)^{-1} \to Q$ 2. Q^{-1} exists. 3. $(\frac{1}{n}X'\epsilon) \to 0$

OLS is consistent under these conditions. This third condition is equal to X being exogenous, that is, no correlation between X and the error term.

We have seen what happens when n grows large. However, we have not discussed the **bias**, that is, what happens in small samples. To study the bias, we will need the expected value of $E(\frac{1}{n}X'\epsilon)$. Here, we need to take into account that X is stochastic, and perhaps, correlated with ϵ. Without further assumptions, we just cannot simplify this expectation. However, under endogeneity, this expectation tends to be unequal to zero.

To summarize, if X is endogenous, some variable in X is correlated with the error term epsilon. And OLS is not consistent. This means that even with an infinite amount of data, OLS will not give useful estimates. We will study an alternative estimation method that solves this in the next lecture.

Example on endogeneity

```
dataset <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/trainexer42.csv")
```

TrainExer42

Simulated data on 250 observations of sales and prices of ice cream under various scenarios of the Data Generating Process [DGP]. The DGP equals:

$$Sales = 100 - 1 \cdot Price + \alpha \cdot Event + \epsilon_1$$

and

$$Price = 5 \cdot \beta Event + \epsilon_2$$

where Event is a 0/1 dummy variable indicating whether a certain event took place. This variable is not included in the data set.

Variables: - PRICE_i: Price variable under β_i for $i = 0, 1, 5, 10$ - SALESj_i: Sales under α_j and β_i

The dataset contains sales and price data for different values of α and β. For each scenario the same simulated values for ϵ_1 and ϵ_2 were used. Specifically, the data contains 4 price series and 16 sales series.

Price variables "Price_i" give the price assuming that $\beta = B_i$ for $B_i = 0, 1, 5, 10$ Sales variables "Sales_ B" give the sales for $\alpha = A_i$ and $\beta = B_i$, for $A_i = 0, 1, 5, 10$.

1. First consider the case where the event only directly affects price ($\alpha = 0$). Estimate and report the price coefficients under all 4 scenarios for β and calculate the R^2 for all these regressions. Do the estimated price coefficients signal any endogeneity problem for these values of α and β? Can you also explain the pattern you find for the R^2?

```
lm1 <- lm(SALES0_0 ~ PRICE0 , data = dataset)
lm2 <- lm(SALES0_1 ~ PRICE1 , data = dataset)
lm3 <- lm(SALES0_5 ~ PRICE5 , data = dataset)
lm4 <- lm(SALES0_10 ~ PRICE10 , data = dataset)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Jul 12, 2020 - 23:26:08

As you can see, the coefficients are all close enough to the true value of -1 so there's no problem here, price is NOT endogenous, as the event does not influence Sales directly. The R^2 increases for higher values of β this is due to the fact that for higher β, more of the variations in sales can be explained.

In other words: for higher $\beta \rightarrow$ Variation in Sales increases $\rightarrow$ Perfectly explained by the Price.

2. Repeat the exercise above, but now consider the case where the event only directly affects sales, that is, set ($\beta = 0$) and check the results for the four different values of α.

```
lm5 <- lm(SALES0_0 ~ PRICE0 , data = dataset)
lm6 <- lm(SALES1_0 ~ PRICE0 , data = dataset)
lm7 <- lm(SALES5_0 ~ PRICE0 , data = dataset)
lm8 <- lm(SALES10_0 ~ PRICE0 , data = dataset)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Jul 12, 2020 - 23:26:08

Tabla 1: Regression Results

	Dependent variable:			
	SALES0_0	SALES0_1	SALES0_5	SALES0_10
	lm1	lm1	lm3	lm4
PRICE0	−0.976***			
	(0.032)			
PRICE1		−0.966***		
		(0.030)		
PRICE5			−0.973***	
			(0.017)	
PRICE10				−0.985***
				(0.010)
Constant	99.862***	99.808***	99.833***	99.890***
	(0.161)	(0.156)	(0.100)	(0.068)
Observations	250	250	250	250
R^2	0.794	0.808	0.930	0.977
Adjusted R^2	0.794	0.807	0.930	0.977
Residual Std. Error (df = 248)	0.525	0.524	0.523	0.523
F Statistic (df = 1; 248)	958.478***	1,044.203***	3,314.297***	10,491.330***
Note:				*p<0.1; **p<0.05; ***p<0.01

Tabla 2: Regression Results

	Dependent variable:			
	SALES0_0	SALES1_0	SALES5_0	SALES10_0
	lm5	lm6	lm7	lm8
PRICE0	−0.976***	−0.969***	−0.942***	−0.909***
	(0.032)	(0.039)	(0.106)	(0.201)
Constant	99.862***	99.948***	100.294***	100.727***
	(0.161)	(0.197)	(0.539)	(1.027)
Observations	250	250	250	250
R^2	0.794	0.718	0.243	0.076
Adjusted R^2	0.794	0.717	0.240	0.072
Residual Std. Error (df = 248)	0.525	0.642	1.757	3.349
F Statistic (df = 1; 248)	958.478***	631.998***	79.714***	20.395***
Note:				*p<0.1; **p<0.05; ***p<0.01

We can see that all the coefficients again are relatively close to the true value -1, again, Price is not endogenous, as the Event only affects Sales, not Price. So the omission of Event does not lead to a correlation between the Error and Price.

We can see that the R^2 drops significantly for higher values of α. At a high value of alpha, a lot of variation in Sales is due to the Event. However, this variation is not captured in the regression, that's why at higher levels of α the model explains less of the variation.

3. Finally consider the parameter estimates for the cases where the event affects price and sales, that is, look at $\alpha = \beta = 0, 1, 5, 10$. Can you see the impact of endogeneity in this case?

```
lm9 <- lm(SALES0_0 ~ PRICE0 , data = dataset)
lm10 <- lm(SALES1_1 ~ PRICE1 , data = dataset)
lm11 <- lm(SALES5_5 ~ PRICE5 , data = dataset)
lm12 <- lm(SALES10_10 ~ PRICE10 , data = dataset)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Jul 12, 2020 - 23:26:08

Tabla 3: Regression Results

	Dependent variable:			
	SALES0_0	SALES1_1	SALES5_5	SALES10_10
	lm9	lm10	lm11	lm12
PRICE0	-0.976^{***}			
	(0.032)			
PRICE1		-0.874^{***}		
		(0.036)		
PRICE5			-0.273^{***}	
			(0.033)	
PRICE10				-0.085^{***}
				(0.021)
Constant	99.862^{***}	99.458^{***}	96.515^{***}	95.515^{***}
	(0.161)	(0.187)	(0.197)	(0.146)
Observations	250	250	250	250
R^2	0.794	0.706	0.214	0.064
Adjusted R^2	0.794	0.705	0.211	0.061
Residual Std. Error (df = 248)	0.525	0.627	1.026	1.119
F Statistic (df = 1; 248)	958.478^{***}	596.815^{***}	67.648^{***}	17.053^{***}
Note:				*p<0.1; **p<0.05; ***p<0.01

We now can see consequences of endogeneity, if $\alpha = \beta \neq 0$, the omition of the Event dummy will lead to correltaion between the Error term $Corr(Price, \epsilon) \neq 0$. As a consequence of the correlation, the estimate can be completely off, as $\alpha = \beta = 10$ shows an estimate close to 0.

The following tables summarize these results.

	beta_0	beta_1	beta_5	beta_10
alpha_0	-0.976	-0.966	-0.973	-0.985
alpha_1	-0.969	-0.874	-0.976	NA
alpha_5	-0.942	NA	-0.273	NA
alpha_10	-0.909	NA	NA	-0.085

Note:
Regression coefficients

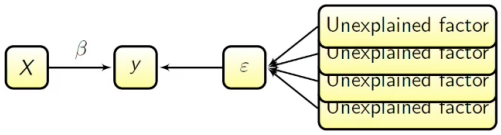
	beta_0	beta_1	beta_5	beta_10
alpha_0	0.794	0.808	0.930	0.977
alpha_1	0.718	0.706	NA	NA
alpha_5	0.243	NA	0.214	NA
alpha_10	0.076	NA	NA	0.064

Note:
Regressions R^2

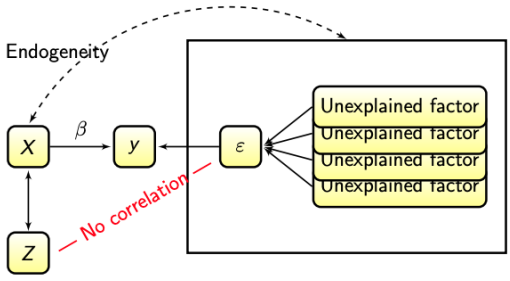
Instruments

When applying econometrics in practice, there are often important factors that cannot be included in the model due to a lack of data. This often leads to endogeneity and in turn inconsistency of OLS. To gain some intuition, we first represent endogeneity in a graphical way.

Here, you see the standard setup, with the dependent variable y, explanatory variables X, and an error term epsilon. Hidden in the epsilon term are different, unexplained factors. These are factors that affect y but are not included in the model, usually because we have no data on them.



Endogeneity appears if at least one of these unexplained factors is correlated with an X variable. The key to consistently estimating the impact of X on y is to find a set of additional variables. Such variables are called instruments and are usually denoted by Z. The instruments need to satisfy two important properties. First of all they should be correlated with X. Secondly they should not be correlated with the unexplained factors.



To correct for endogeneity we need instruments Z such that, Z and X are correlated but Z does not correlate with epsilon. Under these two conditions any correlation that we find between instruments and y will be due

to X. This information can be used to form a new estimator for beta.

Z variables are instruments if:

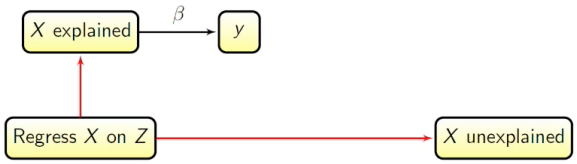
- Z and X are correlated.
- Z does NOT correlate with ϵ

Correlation between instruments and y is only due to X. With $Cov(Z\epsilon) = 0$

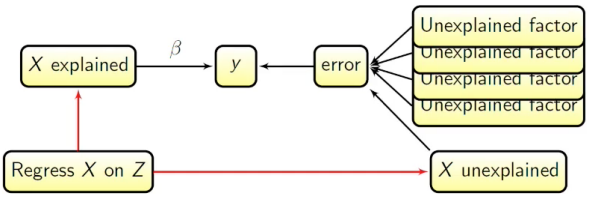
$$Cov(Z, y) = Cov(Z, X\beta + \epsilon) = Cov(Z, X\beta) + Cov(Z\epsilon) = Cov(Z, X)\beta$$

There are two steps to the new estimation procedure:

1. First, we use Z to decompose X in two parts, a part that can be explained by Z and a part that cannot be explained.



2. In the second step, we regress only the explained part of X on y.



The theoretical impact of "X explained" on y equals the true effect size, beta, as the unexplained part of X is simply added to the error term. This solves endogeneity as the unexplained part of X is by construction uncorrelated with the explained part. **So X explained is now exogenous!**

2SLS

This procedure is known as **two-stage least squares, or 2SLS**. Given the linear model and a matrix of instruments Z, we literally need to perform the two steps.

Given model:

$$y = X\beta + \epsilon, Var(\epsilon) = \sigma^2 I$$

And instruments matrix Z

1. Regress X on Z to get explained part : $X = Z\gamma + \eta$

The standard OLS formula applies here. Only the role of X has changed. It is now the dependent variable.

1. Obtain OLS estimator : $\hat{X} = (Z'Z)^{-1}Z'X = H_Z X$

Next, we calculate the explained part of X. Let's denote this part as X hat. X hat can be written as a projection matrix, H of Z times the original matrix of regressors X. (To recall projection matrices look General coefficient estimation). Recall the properties of $H_Z = H'_Z = H'_Z H_Z$, H_Z is symmetric and idempotent.

In the next step, we regress y on X hat using OLS.

2. Regress y on $\hat{X}$

The estimator in this step is the **2SLS estimator, also known as the IV or instrumental variable estimator**. In the first line it is very clear that this estimator is obtained using a standard regression with X hat as explanatory variable.

2. Obtain the 2SLS estimator : $b_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'y = (X'H_ZH_ZX)^{-1}X'H_Z'y$

Using the properties of $H_Z = H_Z' = H_Z'H_Z$

$$b_{2SLS} = (X'H_ZX)^{-1}X'H_Zy$$

Properties of 2SLS

The **variance** of the 2SLS estimator is calculated as follows, first rewrite the 2SLS estimator:

$$b_{2SLS} = (X'H_ZX)^{-1}X'H_Zy = (X'H_ZX)^{-1}X'H_Z(X\beta + \epsilon) = \beta + (X'H_ZX)^{-1}X'H_Z\epsilon$$
$$Var(b_{2SLS}) = Var((X'H_ZX)^{-1}X'H_Z\epsilon) = (X'H_ZX)^{-1}X'H_ZVar(\epsilon)((X'H_ZX)^{-1}X'H_Z)'$$
$$Var(b_{2SLS}) = \sigma^2(X'H_ZX)^{-1}X'H_ZH_Z'X(X'H_ZX)^{-1} = \sigma^2(X'H_ZX)^{-1}I$$

$$Var(b_{2SLS}) = \sigma^2(X'H_ZX)^{-1}$$

The **standard errors** can easily be obtained from the variance matrix. To estimate sigma squared, it is important to use the correct residuals. The residuals should be in terms of the real X variables, **not** the variables used in the second stage regression.

$$\hat{\sigma}^2 = \frac{1}{n-k}(y - Xb_{2SLS})'(y - Xb_{2SLS})$$

2SLS is consistent if some large sample conditions are satisfied. $2SLS \to \beta$ when $n \to \infty$:

1. Z and ϵ are not correlated: $\frac{1}{n}Z'\epsilon \to 0$.
2. Z itself is not multicollinear: $\frac{1}{n}Z'Z \to Q_{ZZ}$ and Q_{ZZ} invertible.
3. X and Z are sufficiently correlated: $\frac{1}{n}X'Z \to Q_{XZ}$ and Q_{XZ} rank k

The third condition also implies that we must at least have as many instruments as explanatory variables.

Given these conditions the following derivation argues that the 2SLS estimator converges to beta as n grows large. You should take some time to look at the steps:

Use the fact that $H_Z = Z(Z'Z)^{-1}Z'$

$$b_{2SLS} = \beta + (X'H_ZX)^{-1}X'H_Z\epsilon = \beta + (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'\epsilon$$

$$b_{2SLS} = \beta + (\frac{1}{n}X'Z(\frac{1}{n}Z'Z)^{-1}\frac{1}{n}Z'X)^{-1}\frac{1}{n}X'Z(\frac{1}{n}Z'Z)^{-1}\frac{1}{n}Z'\epsilon$$
$$b_{2SLS} = \beta + (Q_{XZ}Q_{ZZ}^{-1}Q_{XZ}')^{-1}Q_{XZ}Q_{ZZ}^{-1}(0) = \beta + 0 = \beta$$

How to select instruments?

So, if we have instruments Z we can consistently estimate beta. But **how can we obtain instruments?**

First of all, all exogenous explanatory variables in X qualify as instruments.

If there are endogenous variables, additional instruments are needed. To find these, we often need expert knowledge on the topic of the model. For every endogenous variable, we will need to obtain at least one additional instrument. In general, the stronger the correlation between Z and X, the better. However, we need to make sure that there is no correlation between Z and epsilon.

Let us reconsider an earlier example. Suppose we want to explain the grades on a course using the attendance at lectures. We argued before that attendance is endogenous due to omitted variables, such as the student's motivation. Which variables would be good instruments in this case?

They need to be related to attendance but should not affect the grade itself. Two variables that are likely to be good instruments are travel time to university, or if data over multiple years are available, a variable that indicates an introduction of obligatory attendance. Both variables are not likely to impact grades, but are likely to affect attendance. Students living far away may be less likely to attend all classes. And the policy change will likely increase attendance.

Another example: Recall the case where we wanted to explain demand using price, and where a salesperson strategically sets prices. Suppose that the product is ice cream. What variables can you think of as instruments for price?

- Prices of raw materials (valid)
- Competitor prices (direct influence on sales, so part of ϵ)
- Outside temperature (direct influence on sales, so part of ϵ)

In this case, the price of raw materials is likely to be an instrument. An increase in price of raw materials will increase the consumer price. However, the raw materials' price will not likely affect demand directly. In the end, the consumers only care about the price that they need to pay. Variables like competitor price or outside temperature are not valid instruments. These variables are likely to affect demand themselves.

Statistical properties of 2SLS

Consider the linear mode $y = X\beta + \epsilon$ where some variable in the n × k matrix X may be correlated with ϵ. As a result X may be endogenous. Denote by Z an (n × m) matrix of instruments. In general the 2SLS estimator is given by

$$b_{2SLS} = (X'H_ZX)^{-1}X'H_Zy, with H_Z = H_Z = Z(Z'Z)^{-1}Z'$$

We can show that if $m = k$ we can rewrite the 2SLS estimator to $b_{2SLS} = (Z'X)^{-1}Z'y$

Notice first that the dimensions of $X'Z$,$Z'Z$ and $Z'X$ are all dimension (kxk) and by assumption they have an inverse assumming n is large enough. Therefore we can use the rule $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$

$$b_{2SLS} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y = (Z'X)^{-1}(Z'Z)^{-1}(X'Z)^{-1}X'Z(Z'Z)^{-1}Z'y$$

Symplifying:

$$b_{2SLS} = (Z'X)^{-1}IZ'y = (Z'X)^{-1}Z'y$$

Now suppose that there is only a single explanatory variable, that is, the model equals $y = X\beta + \epsilon$ and that there is only a single instrument z, so $m = k = 1$. Furthermore suppose that the means of x, y and z over the sample are equal to 0. We show that we can write the 2SLS estimator of β as $b_{2SLS} = \frac{Cov(y,z)}{Cov(z,x)}$.

Where $Cov(u, v)$ denotes the (sample) covariance between u and v, which is defined as $Cov(u, v) = \frac{1}{n-2}\sum_{i=i}^{n}(u_i - \bar{u})(v_i - \bar{v})$ where $\bar{u}$ and $\bar{v}$ denote the sample variance of u,v respectively.

We can use the definition obtained $b_{2SLS} = (Z'X)^{-1}Z'y$, furthermore, $Z = \begin{pmatrix} Z_1 \\ ... \\ Z_n \end{pmatrix}$ and $X = \begin{pmatrix} X_1 \\ ... \\ X_n \end{pmatrix}$ therefore $X'Z = \sum_{i=i}^{n} z_i x_i$ and $Z'y = \sum_{i=i}^{n} z_i y_i$. The 2SLS estimator can be written as:

$$b_{2SLS} = \frac{\sum_{i=i}^{n} z_i y_i}{\sum_{i=i}^{n} z_i x_i}$$

And as the sample means are equal to 0 this can be rewritten as:

$$b_{2SLS} = \frac{\sum_{i=i}^{n} z_i y_i}{\sum_{i=i}^{n} z_i x_i} = \frac{\sum_{i=i}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=i}^{n}(z_i - \bar{z})(x_i - \bar{x})} = \frac{Cov(y, z)}{Cov(z, x)}$$

Notice that factors $\frac{1}{n-2}$ cancel eachother.

Now we use this last formula in to explain what happens to the 2SLS estimator when the correlation between instruments and the endogenous variable is very small:

From $b_{2SLS} = \frac{Cov(y,z)}{Cov(z,x)}$ we can see that when the correlation between Z,X is zero, the 2SLS estimator is not defined, furthermore correlation between y,Z will also be zero as any correlation between these two variables shoud be due to X. The 2SLS estimator will be 0/0. Although in practice is hard to find 0 correlation.

As consequence you may obtain any number as an estimate when correlation is very low.

Conclusion

2SLS solves endogeneity. However, there is a price that we need to pay. We should only use 2SLS if the explanatory variables are really endogenous. If X is in fact exogenous, OLS and 2SLS are both consistent. However, the Gauss-Markov theorem says that the variance of OLS will never be larger than that of 2SLS.

Testing for endogeneity

Given the standard model and instruments Z, there are two important things to test. First, whether the variables in Z actually qualify as instruments, and second, whether X is exogenous or endogenous.

Validity of instruments

First we focus on testing the validity of the instruments. Actually, we need to check three conditions. One, we need to check whether there are enough instruments. Two, whether the instruments are correlated enough with X. And three, we need to check for zero correlation with epsilon:

1. One instrument per endogenous variable.
2. Instruments are correlated (enough) with X
3. Instruments are not correlated with ϵ

Checking the first is easy. We just need to count. If there are three endogenous variables, we need to obtain at least three additional instruments. To check the second, we can rely on the first-stage regression of 2SLS. In this regression we explain the endogenous variables using all the instruments. The additional instruments must sufficiently explain the endogenous variables. This is quite easy to check using t-statistics or an F-test. The final condition is more difficult to check. The testing procedure used for this is called the Sargan–Hansen test.

Instruments correlated with X

Let's first set up the **notation.** We split the experimental variables into the endogenous variables in X_1, and the exogenous variables in X_2.

- X_1 potentially endogenous variables.
- X_2 exogenous variables.
- $Z = (Z^*, X_2)$ Instruments

The complete set of instruments is now formed by additional variables, Z^* and the exogenous X_2 variables. Remember that the exogenous variables X_2 should be included in Z.

In the first stage of 2SLS regression, we explain X_1 using Z^* and X. For 2SLS to work, we need that γ_1 in the equation on the slide is unequal to 0.

$$\text{1st stage OLS} : X_1 = Z^*\gamma_1 + X_2\gamma_2 + \eta$$
$$\text{We require} : \gamma_1 \neq 0$$

If γ_1 is close to 0, the predicted values for X1 are almost perfectly correlated with X2. This complicates the second stage estimation and substantially inflates the variance of the estimator. In fact, if gamma 1 equals 0, the 2SLS estimator is not defined.

- IF $\gamma_1 \approx 0 \rightarrow \hat{X}_1 \approx X_2\hat{\gamma}_2$ so $\hat{X}_1$ almost perfectly correlated with X_2

To test for sufficient correlation, we can simply perform a t or F-test for the null hypothesis that $H_0 : \gamma_1 = 0$.

Z uncorrelated with epsilon

The Sargan test check if the 2SLS residuals are correlated with the instruments Z. If so, this is a sign that the instruments may directly influence the dependent variable y and therefore not valid.

Once you have established that instruments and the endogenous variables are correlated enough, you need to move on to test that Z is not correlated with the error term. The Sargan test can be used for this. The setup of the model and instruments is summarized as follows:

- Model $y = X\beta + \epsilon$
- Explanatory variables : $X = (X_1, X_2)$ where X_1 endogenous and X_2 exogenous.
- Instruments $Z = (Z^*, X_2)$.
- Null Hypothesis $H_0 : Corr(Z, \epsilon) = 0$ We can rewrite this hypothesis in a regression that explains epsilon using Z.

$$\epsilon = Z\delta + \zeta$$
$$H_0 : \delta = 0$$

The null hypothesis, that epsilon and Z are unrelated, can now be translated to $\delta = 0$. The problem that epsilon cannot be observed is solved by estimating its elements using 2SLS. To execute the Sargan test, we follow four steps:

1. First, we use the instruments Z to perform 2SLS for β
2. Calculate the residuals $e_{2SLS} = y - Xb_{2SLS}$ Note that these residuals are based on X, not $\hat{X}$.
3. We regress these residuals on Z $e_{2SLS} = f(Z)$ and obtain the R^2 of that regression. Reject $H_0 : \delta = 0$ if R^2 is too large

If the R^2, this means that we can explain a lot of the residuals using Z. This will be a sign that Z itself explains y, and this would violate the assumptions. Therefore, we reject the null hypothesis if the R^2 is too large.

4. $nR^2 \sim \chi^2(m - k)$ Under H_0 : Epsilon and Z are unrelated, Valid Instruments where m is the instruments in Z and k the explanatory variables in X.

Under the null, the distribution of the number of observations times the R-squared is approximately equal to a chi-squared distribution. The degrees of freedom equals the number of instruments, m, minus the number of explanatory variables, k.

The degrees of freedom imply that a **test can only be performed if** $(m > k)$. If $(m = k)$ 2SLS can still be applied, and endogeneity is solved if all requirements are met. However, we just cannot test the validity of the instruments. Actually, with the Sargan test we can only test whether the excess instruments are valid. We need to assume that at least the minimum number of instruments are valid.

If the test rejects, we know that the set of instruments is not valid. Statistical tests cannot be used to figure out which specific instruments are invalid. In practice, you will need to also motivate the validity of instruments based on economic theory or logical reasoning. However, when in doubt about specific variables, the Sargan test is a very useful tool.

- Test only works when there are "too many" instruments $(m > k)$.
- At least k of the instruments should be valid
- Test cannot indicate which instruments are invalid!

Is 2SLS needed?

After having checked the three conditions for validity of instruments, you can move on to the final hypothesis to test. The null hypothesis for this test is that particular explanatory variables are exogenous. This comes down to testing whether the entire 2SLS procedure is really necessary. The mostly used test is called the Durbin–Wu–Hausman test.

This test uses instruments to split the explanatory variable into two parts. One part is exogenous by construction, while the other part is endogenous under the alternative hypothesis.

If the X variable is exogenous, both parts are exogenous and will have the same impact on y. If the variable is endogenous, this is not the case. The intuition of the test is to test for a difference between the two effects. We have the same setup for the Hausman test as before:

- Explanatory variables: $X = (X_1, X_2)$
- Potentially endogenous: X_1 (k_1 variables)
- Exogenous variables: X_2 (k_2 variables)
- Instruments: Z

The null hypothesis is that the k_1 variables in X_1 are exogenous. $H_0 : X_1 exogenous$ The procedure of the Hausman test is as follows:

1. Regress y on X and calculate the residuals $e = y - Xb$
2. Regress X_1 on Z.

Obtain the potentially endogenous part of X_1 as the residuals of a regression of X_1 on the instruments Z. Call these residuals V. $V = X_1 - Zb_{endog}$

3. Regress e on X and V.

We regress the residuals from step one on all explanatory variables and the residuals from step two. If X is exogenous, V does not explain anything and this regression should not fit well. If the R^2 is too large, we should reject the exogeneity of variables X1, and 2SLS is necessary.

4. $nR^2 \sim \chi^2(k_1)$ under H_0 of exogeneity.

The distribution of n times R squared is approximately a chi-squared distribution with as degrees of freedom the number of variables that we test.

Example of endogeneity and instruments

```
trainexer44 <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/trainexer44.csv")
```

trainexer44

Consumption of motor gasoline in the US from 1970 to 1999, including price index, disposable income, and price indices of used cars, new cars, and public transport. Data source is C. Heij et al, Econometric Methods with Applications in Business and Economics, 2004, Oxford. Original data sources: Economic Report of the President 2000 and Census Bureau and Department of Energy.

Variables:

- OBS: Year of observation
- GC: log real gasoline consumption
- PG: log real gasoline price index
- RI: log real disposable income
- RPN: log real price index of new cars
- RPT: log real price index of public transport
- RPU: log real price index of used cars

In this exercise we study the gasoline market and look at the relation between consumption and price in the USA. We will use yearly data on these variables from 1977 to 1999. Additionally we have data on disposable income, and some price indices.

We consider the following model:

$$GC = \beta_1 + \beta_2 PG + \beta_3 RI + \epsilon$$

1. Give an argument why the gasoline price may be endogenous in this equation.

The USA is a major player in the gasoline market, so it is likely that high demand for gasoline by the USA leads to an increase in the market price, in other words, consumption (GC) and price (PG) are determined simultaneously. Therefore we suspect that gasoline price (PG) may be endogenous.

2.Use 2SLS to estimate the price elasticity (β_2). Use a constant, RI, RPT, RPN, and RPU as instruments.

We follow the steps of 2SLS.

1. Regress X on Z to get explained part : $X = Z\gamma + \eta$

```
lm1 <- lm(PG ~ RPT + RPN + RPU + RI , data = trainexer44)
PG_fitted <- predict(lm1)
# Add fitted values to data set
#trainexer44 <- cbind(trainexer44, PG_fitted)
trainexer44 <- trainexer44%>% mutate(PG_fitted=fitted(lm1))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Jul 12, 2020 - 23:26:09

The p -value for the instruments RPT,RPN is very low, indicating a strong correlation between this instrument and the endogenous variable PG aven after controling for other variables.

2. Regress y on $\hat{X}$

```
lm2 <- lm(GC ~ RI + PG_fitted , data = trainexer44)
```

Tabla 4: 1st Stage

	Dependent variable:
	PG
RPT	-0.808^{***}
	(0.191)
RPN	-3.528^{***}
	(0.352)
RPU	0.233
	(0.183)
RI	-2.298^{***}
	(0.247)
Constant	7.741^{***}
	(0.834)
Observations	30
R^2	0.887
Adjusted R^2	0.869
Residual Std. Error	0.073 (df = 25)
F Statistic	48.969^{***} (df = 4; 25)
Note:	*p<0.1; **p<0.05; ***p<0.01

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Jul 12, 2020 - 23:26:09

Tabla 5: 2nd Stage

	Dependent variable:
	GC
RI	0.565^{***}
	(0.041)
PG_fitted	-0.544^{***}
	(0.046)
Constant	5.014^{***}
	(0.134)
Observations	30
R^2	0.967
Adjusted R^2	0.964
Residual Std. Error	0.038 (df = 27)
F Statistic	393.308^{***} (df = 2; 27)
Note:	*p<0.1; **p<0.05; ***p<0.01

We use PG_fitted which are the predicted values of step 1. The estimated price elasticity is -0.54 that means that a 1% price increase leads to about -0.5% decrease in consumption.

3.Perform a Sargan test to test whether the five instruments are correlated with ϵ. What do you conclude?

We check if the 2SLS residuals are correlated with the instruments Z. If so, this is a sign that the instruments may directly influence the dependent variable y and therefore not valid.

First we calculate the residuals: Note that you need to use PG here, NOT PG_fitted

$$Res.2SLS = GC - (5.01 + 0.56RI - 0.54PG)$$

```
#Create the residuals in a column named res_2sls
trainexer44 <- trainexer44%>% mutate(res_2sls = GC -(5.01+0.56*RI-0.54*PG))
```

Second we regress the Res.2SLS on all the instruments:

```
lm3 <- lm(res_2sls ~ RPT + RPN + RPU + RI , data = trainexer44)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Jul 12, 2020 - 23:26:09

The Sargent test statistic is equal to: $nR^2 = 30(0.104) = 3.12$ and should be compared with $\chi^2(m-k) = \chi^2(5-3)$.

We check the value of chi squared statistic R function qchisq(p, df, lower.tail) is the value of x at the qth percentile (lower.tail = TRUE).

Tabla 6: Sargan Test regression

	Dependent variable:
	res_2sls
RPT	−0.048
	(0.062)
RPN	0.036
	(0.114)
RPU	−0.071
	(0.059)
RI	0.075
	(0.080)
Constant	−0.240
	(0.270)
Observations	30
R^2	0.109
Adjusted R^2	−0.033
Residual Std. Error	0.024 (df = 25)
F Statistic	0.768 (df = 4; 25)
Note:	*p<0.1; **p<0.05; ***p<0.01

```r
qchisq(0.95, 2)
```

```
## [1] 5.991465
```

The 5% critical value is $\chi^2(2) = 5.99$ therefore as $3.12 < 5.99$ we can NOT reject H_0 : Epsilon and Z are unrelated, we have valid instruments.

ivreg function in R

A quicker way to perform the Sargant test is to use the R function ivreg(). For documentation read this. Notice that the vertical bar character | separates the proper regressor list from the instrument list. For a model to be identified the number of instruments should be at least equal to the number of endogenous variables. If there are more instruments than endogenous variables, the model is said to be overidentified.

```r
# Our model with endogenous variable PG
gc_model <- lm(GC ~ PG + RI  , data = trainexer44)
# Our IV model with Instruments RI, RPT, RPN, and RPU
gc_model.iv <- ivreg(GC~PG+RI|
           RPT + RPN + RPU + RI, data=trainexer44)
# Our 2SLS model for comparison purpose.
lm2sls <- lm(GC ~ RI + PG_fitted , data = trainexer44)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Jul 12, 2020 - 23:26:09

Tabla 7: Sargan Test using ivreg function

	Dependent variable:		
	GC		
	OLS	*instrumental variable*	*OLS*
	endog	ivreg	2sls
	(1)	(2)	(3)
PG	-0.528^{***}	-0.544^{***}	
	(0.026)	(0.029)	
RI	0.573^{***}	0.565^{***}	0.565^{***}
	(0.025)	(0.025)	(0.041)
PG_fitted			-0.544^{***}
			(0.046)
Constant	4.986^{***}	5.014^{***}	5.014^{***}
	(0.081)	(0.084)	(0.134)
Observations	30	30	30
R^2	0.987	0.987	0.967
Adjusted R^2	0.986	0.986	0.964
Residual Std. Error (df = 27)	0.024	0.024	0.038
F Statistic (df = 2; 27)	$1{,}037.467^{***}$		393.308^{***}
Note:			*p<0.1; **p<0.05; ***p<0.01

The table shows that the importance of PG in determining GC decreases in the IV model. It also shows that the explicit 2SLS model and the IV model yield the same coefficients (the PG in the IV model is equivalent to the PG_fitted in 2SLS), but the standard errors are different. The correct ones are those provided by the IV model which are the ones we use for the Sargan Test.

Case Application

```
dataset4 <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/dataset4.csv")
```

dataset4

Simulated data on performance of 1000 participants of an Engineering MOOC. Performance is measured by Grade Point Average. Background variables are gender, whether participant followed a preparatory mathematics MOOC, and whether the participant received an email invitation for this preparatory MOOC.

Variables:

- GPA: Grade point average scale 0 to 10, 10 being the best
- PARTICIPATION: 0/1 variable indicating participation (1) in preparatory MOOC or not (0)
- GENDER: 0/1 variable indicating gender: male (1), female (0)
- EMAIL: 0/1 variable indicating whether participant received email invitation for preparatory course (0: no invitation, 1: invitation)

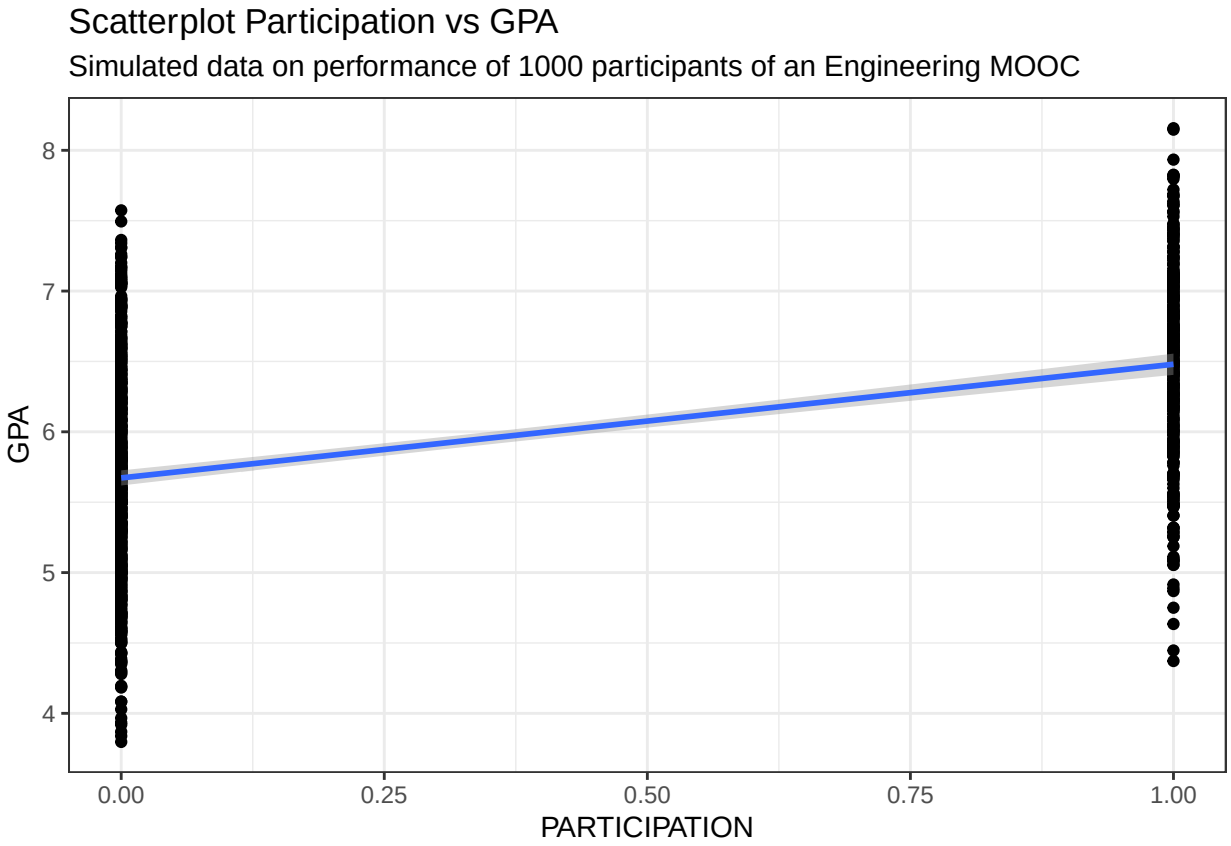
We place ourselves in the position of the organizer of a MOOC on an engineering topic. The organizer is interested in estimating the dependence of the grade point average on whether or not the learner took a preparatory mathematics MOOC. When estimating this dependence, it is important to realize that the participation in the prep course is voluntary. Learners choose to take the course or not.

Let's first take a look at some data statistics.

- We have data on 1,000 learners
- 49% are male, and about 34% participated in the prep course.
- The average GPA was close to 6, where the performance is measured on a 10-point scale, with 10 being the best.

To get some first insight in the impact of the prep course, we take a look at the correlation between GPA and participation.

```
dataset4 %>% ggplot(aes(x=PARTICIPATION,y=GPA)) + geom_point() + geom_smooth(method='lm') +
  labs(title="Scatterplot Participation vs GPA ",
       subtitle="Simulated data on performance of 1000 participants of an Engineering MOOC") + theme_bw
```



Participation on the horizontal access only takes two possible values. Zero for no, or one for yes. From this graph it seems that GPA is indeed a bit higher for learners who have taken the prep course. This red regression line confirms this idea.

So, there seems to be a positive impact but how large is it, and is it significant? Furthermore, we may need to correct for the gender of the learner. A natural starting point is to formulate a linear model in which we regress GPA on a constant, a gender dummy variable that equals one if the learner is male and a dummy for participation.

$$GPA = \beta_1 + \beta_2 GENDER + \beta_3 PARTICIPATION$$

```
lm1 <- lm(GPA ~ GENDER + PARTICIPATION  , data = dataset4)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Jul 12, 2020 - 23:26:11

Participation seems to have a large positive impact on the GPA of about 0.8 grade point. The t-statistic suggests that this impact is also significant. The table further shows that males tend to perform slightly worse compared to females by about .2 grade point.

We now need to ask ourselves **whether we should trust these OLS estimates?** If you think about all you have learned from our previous sections, you might suspect that the answer is no. It is very **likely that participation is endogenous.**

The reasoning for this goes as follows. First, learners self-select for the prep course. They are free to take this course or not. Next, omitted factors, such as various characteristics of the learners relate to this self-selection. These same characteristics are also likely to influence GPA. Hence, participation is likely to be endogenous.

Tabla 8: Regression results

	Dependent variable:
	GPA
GENDER	-0.214^{***}
	(0.044)
PARTICIPATION	0.824^{***}
	(0.047)
Constant	5.771^{***}
	(0.034)
Observations	1,000
R^2	0.244
Adjusted R^2	0.243
Residual Std. Error	0.698 (df = 997)
F Statistic	160.970^{***} (df = 2; 997)
Note:	*p<0.1; **p<0.05; ***p<0.01

If participation is endogenous, we know that OLS is inconsistent, and we cannot use the OLS results as a reliable estimate for the causal impact of the prep course.

Endogeneity made lead to **over- or underestimation of the actual effect**. This depends on the impact of the omitted factors. OLS will overestimate the impact when the dominant omitted factor affects participation and GPA in the same way. An example of such a factor is motivation. As highly motivated students tend to get a high GPA and are likely to take the prep course. OLS will underestimate when the dominant omitted factor affects participation and GPA in opposite ways. The mathematics level of the learner is an example. A high level will need learners to skip the prep course, but they will likely get a relatively high GPA. The net effect of the combined factors is of course difficult to judge, as it will depend on the relative importance of all omitted variables.

- Overestimation → Omitted factor: Motivation affects participation and GPA in the same way. High motivation → High GPA & Take prepcourse.
- Underestimation → Omitted factor: Maths level affects participation and GPA in opposite ways. Good at maths → High GPA & Do not take prepcourse.
- Net effect: difficult to judge. Depends on the size of effects and maybe other variables (Age?).

As there may be many omitted factors, we need to correct for possible endogeneity of participation by **using two-stage least squares**. This method requires instruments. The main conditions for instruments are shown in Instruments:

- They should relate to the prep course participation, (Relate to X)
- but they should not affect GPA. (Not affect y)

Selecting instrument

It is not always easy to find such variables. Many learner specific variables, such as intelligence, number of MOOCs already followed, or the age of the learner, are not likely to be valid. The reason is that all these variables are likely to affect the GPA as well. So, how can we get instruments? Finding these requires creativity, and a bit of luck.

In this case, we could search for instruments that do not describe the learner, but something related to the

learning platform. We got the extra information that learners, in principle, receive an email invitation to take the prep course. However, due to some technical email problems with the platform, some learners did not get this email. And, for our study, we got information on who did get the email and who did not.

The variable that codes whether the learner did or did not receive the email, as specified here, is a perfect instrument if two conditions are satisfied:

$$\text{Instrument candidate : } EMAIL = \begin{cases} 0 & \text{if email not recived} \\ 1 & \text{if email recived} \end{cases}$$

Email is good instrument if:

1. Email problem is random. We assume that this condition is satisfied.
2. Invitation affects participation

2SLS

We check the second condition using the first-stage regression of 2SLS. In this stage we explain participation using all instruments, that is, the exogenous variables of the model: so the constant term and the gender dummy and in addition the email dummy.

```
stage_1 <- lm(PARTICIPATION ~ GENDER + EMAIL  , data = dataset4)
# Add fitted values to data set
dataset4 <- dataset4%>% mutate(PART_fitted=fitted(stage_1))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Jul 12, 2020 - 23:26:11

Tabla 9: 2SLS Stage 1

	Dependent variable:
	PARTICIPATION
GENDER	0.048*
	(0.027)
EMAIL	0.413***
	(0.027)
Constant	0.101***
	(0.023)
Observations	1,000
R^2	0.196
Adjusted R^2	0.194
Residual Std. Error	0.425 (df = 997)
F Statistic	121.212*** (df = 2; 997)
Note:	*p<0.1; **p<0.05; ***p<0.01

It is clear that having received the email significantly influences the participation. In fact, the probability of participation increases with about 40% when the email was received.

We are now ready to perform 2SLS. Here we explain the GPA using a constant and dummy variables for gender, and participation. As instruments, we use a constant, the gender dummy, and the email dummy.

```
stage_2 <- lm(GPA ~ GENDER + PART_fitted  , data = dataset4)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Jul 12, 2020 - 23:26:11

Tabla 10: Original model and 2SLS

	Dependent variable:	
	GPA	
	Endog	2SLS
	(1)	(2)
GENDER	-0.214^{***}	-0.173^{***}
	(0.044)	(0.051)
PARTICIPATION	0.824^{***}	
	(0.047)	
PART_fitted		0.240^{**}
		(0.122)
Constant	5.771^{***}	5.948^{***}
	(0.034)	(0.051)
Observations	1,000	1,000
R^2	0.244	0.013
Adjusted R^2	0.243	0.011
Residual Std. Error (df = 997)	0.698	0.798
F Statistic (df = 2; 997)	160.970^{***}	6.694^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	

We now find a smaller but still significant impact of participation. Participation increases GPA by 0.24 points, which is considerably less then the OLS estimate of 0.84 obtained before. We have argued in previous lectures that 2SLS will increase the estimation uncertainty. So we should only use 2SLS when participation is indeed endogenous.

Hausman test

We can test for this using the **Hausman test** . The null hypothesis is that the variable participation is exogenous.. $H_0 : PARTexogenous$

- The dependent variable in the test regression is the set of residuals based on the OLS regression of GPA on a constant and the gender and participation dummies.
- The explanatory variables in the Hausman test are all original explanatory variables, plus the residuals of the first-stage regression where we explained participation using the instruments.

$$lm1.resid = \beta_1 + \beta_2 GENDER + \beta_3 PART + \beta_4 stage\_1.resid$$

The results of this regression are the following:

```r
# Add to dataset the residuals of the OLS and stage_1
dataset4 <- dataset4 %>% mutate(lm1.res = resid(lm1), stage_1.resid=resid(stage_1))

hausman <- lm(lm1.res ~ GENDER + PARTICIPATION + stage_1.resid, data = dataset4)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Jul 12, 2020 - 23:26:11

Tabla 11: Hausman test regression

	Dependent variable:
	lm1.res
GENDER	0.041
	(0.044)
PARTICIPATION	-0.584^{***}
	(0.105)
stage_1.resid	0.722^{***}
	(0.117)
Constant	0.177^{***}
	(0.044)
Observations	1,000
R^2	0.037
Adjusted R^2	0.034
Residual Std. Error	0.686 (df = 996)
F Statistic	12.683^{***} (df = 3; 996)
Note:	*p<0.1; **p<0.05; ***p<0.01

- The test-statistic is the number of observations times the R-squared, and equals 36.8. This number should be compared to the chi-squared distribution with one degree of freedom as we test for the endogeneity of a single variable.
- Hausman test-stat: $nR^2 = 1000 \cdot 0.0368 = 36.8$

```r
qchisq(0.95, 1)
```

```
## [1] 3.841459
```

- As the critical value is $\chi^2(1) = 3.8$, $36.8 > 3.8$ we should reject the null hypothesis. Participation is endogenous and 2SLS should be used.

Note on the standard errors.

We obtain the standard errors that correspond to the regression **stage_2**:

$$GPA = \beta_1 + \beta_2 Gender + \beta_3 \hat{Partic}ip + \epsilon$$

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Jul 12, 2020 - 23:26:11

Tabla 12: 2SLS Stage 2

	Dependent variable:
	GPA
GENDER	-0.173^{***}
	(0.051)
PART_fitted	0.240^{**}
	(0.122)
Constant	5.948^{***}
	(0.051)
Observations	1,000
R^2	0.013
Adjusted R^2	0.011
Residual Std. Error	0.798 (df = 997)
F Statistic	6.694^{***} (df = 2; 997)
Note:	*p<0.1; **p<0.05; ***p<0.01

Is important to notice that the se reported underneath each coefficient are not totally right. This is due to the fact that the wrong estimate is used for the variance of ϵ. These should be based on the residuals calculated using $Part$ not the fitted values of $Part.fitted$ from Stage_1. This is explained in Properties of 2SLS and you can see how is correctly calculated in Example of endogeneity and instruments.

The ratio $\frac{SE.incorrect}{SE.correct}$ can be greater or less than 1. This means the SE obtained in SE.incorrect are 1-ratio to high or to low.

For this example means:

$$\frac{\hat{\sigma}^2_{correct}}{\hat{\sigma}^2_{incorrect}} = \frac{\frac{1}{n-k}\sum_{i=1}^{n}(GPA_i - 5.95 + 0.17GENDER_i - 0.24\hat{PART}_i)^2}{\frac{1}{n-k}\sum_{i=1}^{n}(GPA_i - 5.95 + 0.17GENDER_i - 0.24PART_i)^2} = 1.129.$$

Since $\sqrt{1.129} = 1.063$, the standard errors obtained in the second-stage regression are about 6% too high.