tutorial_2

Lir Hoxhaj

2025-01-29

Background

GOALS OF TUTORIAL:

1. Understand OLS assumptions and the violations of these assumptions
2. Understand what Instrumental variables are, and why we use IVs
3. Learn how to run an IV regression in R

Sources:

- Cigarette dataset (https://github.com/vincentarelbundock/Rdatasets/blob/master/csv/Ecdat/Cigarette.csv)

Setup

Let's set up a new environment and load our libraries in one place.

```r
rm( list = ls(all.names = TRUE) )

## Installing libraries

if(!require(ggplot2)) install.packages("ggplot2")
```

```
## Loading required package: ggplot2
```

```r
#if (!requireNamespace("ggplot2", quietly = TRUE)) {install.packages("ggplot2")}
if(!require(car)) install.packages("car")
```

```
## Loading required package: car
```

```
## Warning: package 'car' was built under R version 4.4.1
```

```
## Loading required package: carData
```

```r
#if (!requireNamespace("car", quietly = TRUE)) {install.packages("car")}
if(!require(lmtest)) install.packages("lmtest")
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
## 
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
## 
##     as.Date, as.Date.numeric
```

```r
if(!require(sandwich)) install.packages("sandwich")
```

```
## Loading required package: sandwich
```

```r
if(!require(AER)) install.packages("AER")
```

```
## Loading required package: AER

## Warning: package 'AER' was built under R version 4.4.1

## Loading required package: survival
```

```r
if(!require(ivreg)) install.packages("ivreg") # new versions have ivreg function in AER
```

```
## Loading required package: ivreg

## Warning: package 'ivreg' was built under R version 4.4.1

## Registered S3 methods overwritten by 'ivreg':
##   method              from
##   anova.ivreg         AER
##   hatvalues.ivreg     AER
##   model.matrix.ivreg  AER
##   predict.ivreg       AER
##   print.ivreg         AER
##   print.summary.ivreg AER
##   summary.ivreg       AER
##   terms.ivreg         AER
##   update.ivreg        AER
##   vcov.ivreg          AER

##
## Attaching package: 'ivreg'

## The following objects are masked from 'package:AER':
##
##     ivreg, ivreg.fit
```

```r
if(!require(dplyr)) install.packages("dplyr")
```

```
## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
## Loading libraries
library(ggplot2)
library(car)
library(lmtest)
library(sandwich)
library(AER)
library(ivreg)
library(dplyr)
```

Setting up our working directory

```
# Automatically (based on location of this script)
working_directory <- dirname(rstudioapi::getActiveDocumentContext()$path)

# , or manually...
# setwd("C:\\Users\\lhoxhaj\\OneDrive - Imperial College London\\Desktop\\TA\\Shuang\\Tutorials\\Lir\\t
```

1. OLS assumptions (and violations)

Motivation

We mentioned that when we try to use a statistical method to *predict* and *infer* a relationship between a dependent and independent variable, we make several assumptions about the sample data we are using and the functional form of the relationship.

Assumptions

For the OLS method, we assume the following:

1. **Linear relationship**
2. **Linear independence of regressors (no multicollinearity)**
3. **Homoscedasticity of error term**
4. **Normal distribution of error term**
5. **Zero mean of error term**
6. **Exogeneity (no endogeneity)**

The assumptions (some stricter than others) help define OLS and give us estimates of β that are:

- **consistent**: estimator is closer to the truth as sample size increases: $\hat{\beta} \to \beta$ as $n \to \infty$;
- and **unbiased**: expected value of estimator is equal to the true value of the parameter $E[\hat{\beta}] = \beta$.

1. Linear Relationship

We assume that the relationship between our dependent and independent variables is linear. This can be expressed using matrix notation:

$$E[Y|X] = X\beta$$

, or the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

where:

- index i covers all the observations - $i = 1, ..., n$
- y_i is the **dependent** variable (and Y is its matrix), also known as **regressand** or **left-side**
- x_i is the **independent** variable (and X is its matrix), also known as **regressor** or **right-side**
- $y = \beta_0 + \beta_1 x$ is the **regression line**
- β_0 is the regression line **intercept**

- β_1 is the regression line **slope** (sometimes **coefficient**)
- u_i is the **error term**

β_0 and β_1 are unknown and are (most cases) anything but linear, since relationships between variables are complex and cannot be simply explained through a linear equation.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

, where the error is the difference between predicted outcome and the true outcome.

2. Linear independence of regressors (no multicollinearity)

Independent variables should be independent from each other, and there no linear relationship between two or more of the independent variables should exist (or at least a very weak one). This ensures that each independent variable contributes unique information to the model. **Perfect multicollinearity** (when variables are linearly dependent and vary together almost always) leads to problems with estimation, as the model cannot uniquely estimate the coefficients, which can make it difficult to interpret the effects of individual variables.

$$corr[X_1, X_2] \neq 1$$

3. Homoscedasticity of error term

The error term should have constant variance for all values of the independent variables. This means that the spread of residuals is the same across all levels of the independent variables. If the variance of the errors depends on the level of the independent variables (heteroscedasticity), the OLS estimates can still be unbiased but will no longer be efficient, and hypothesis tests can be misleading. There could very well be something that is causing heteroscedastic (not constant) variance in the error term, some other factor that we have not accounted for in the model.

Here, we may write the variance of the error term as some constant σ^2 (σ is standard error), which is the same no matter the values of X.

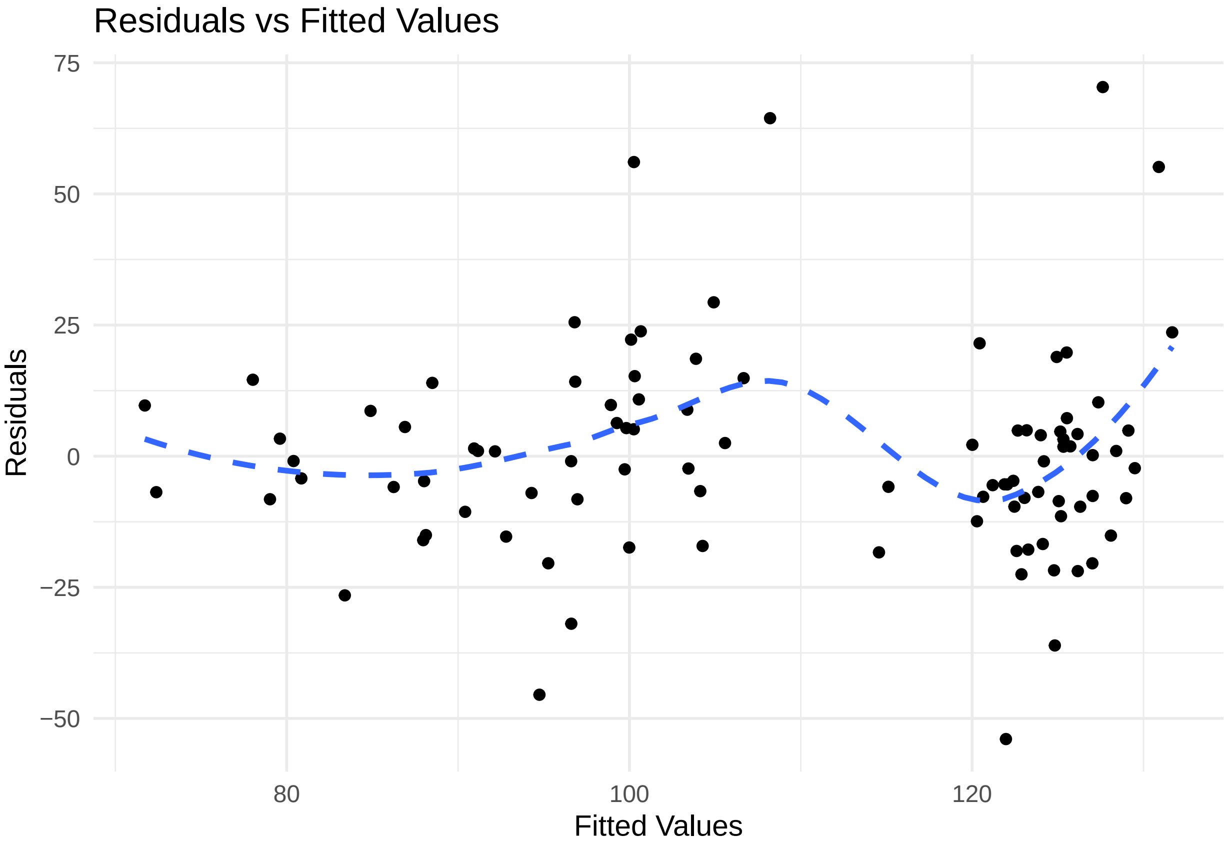
$$Var[u|X] = \sigma^2$$

```r
# Step 1: Using cigarettes dataset, we have variables price of pack of cigarette and quantity of packs
cigarettes <- read.csv("cigarettes.csv")
# Step 2: Estimating relationship between packs (quantity) and price (per pack)
model <- lm(packs ~ price, data = cigarettes)
```

In a homoscedastic scenario, the spread of the residuals should be roughly constant across the range of fitted values. If you see a pattern or a funnel shape in the plot, it might suggest homoscedasticity (constant variance), whereas if variance seems to be changing (sometimes points are closer to each other, and at other times they are further away or more varied) then that will raise heteroscedasticity concerns.

```r
# Step 3: Showing the values of fitted data (on x-axis) and the residuals (on y-axis)
residuals_plot <- ggplot(data = cigarettes, aes(x = fitted(model), y = resid(model))) +
  geom_point() +
  geom_smooth(method = "loess", se = FALSE, linetype = "dashed") +
  labs(x = "Fitted Values", y = "Residuals") +
  ggtitle("Residuals vs Fitted Values") +
  theme_minimal()

print(residuals_plot)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```
# Step 4: Perform Breusch-Pagan test to check for homoscedasticity
bp_test <- ncvTest(model)
print(bp_test)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 2.430046, Df = 1, p = 0.11903
```

$$H_0 : \text{error term is homoscedastic} \;, Var[u|X] = \sigma^2$$
$$H_1 : \text{error term is heteroscedastic} \;, Var[u|X] \neq \sigma^2$$

Conclusion: not enough evidence to reject null hypothesis, hence distribution of error term is homoscedastic.

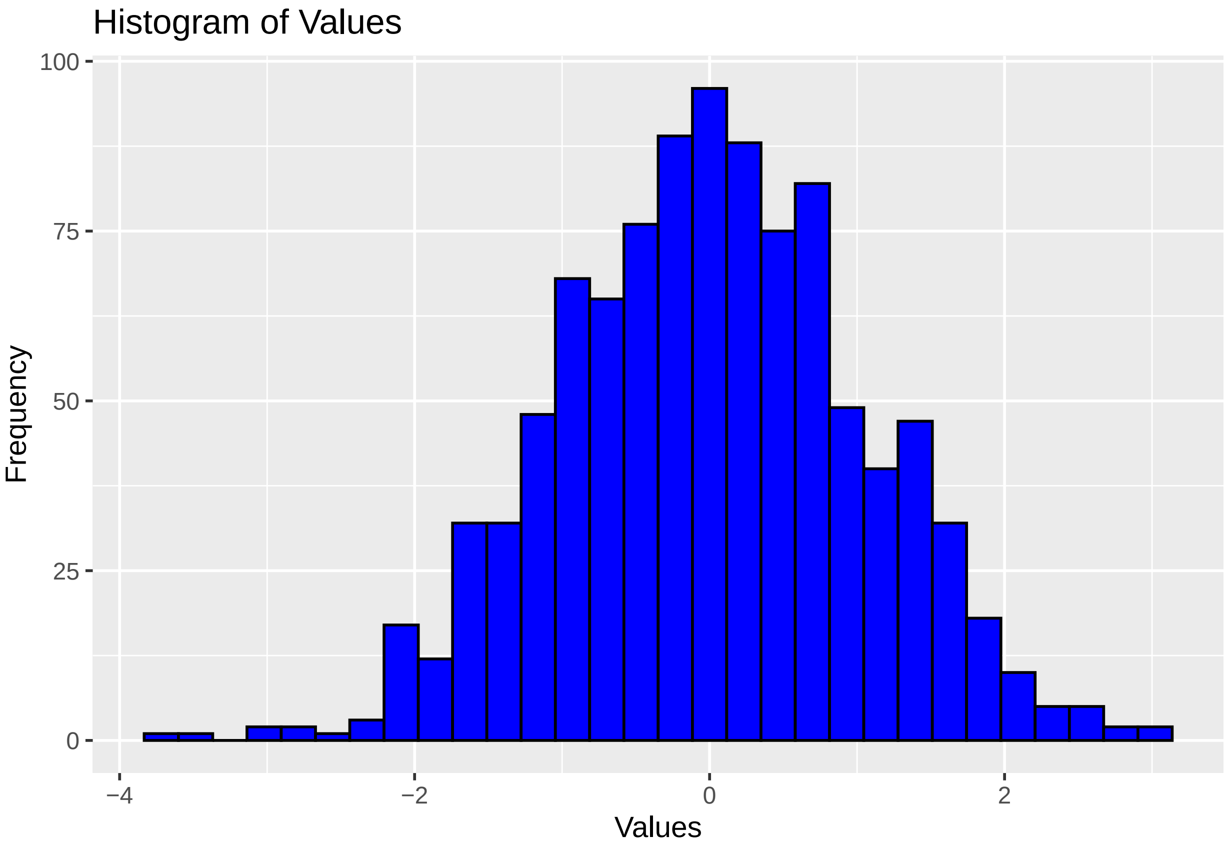
4. Normal distribution of error term

The error term should be normally distributed. This assumption is mainly important for making valid inferences about the regression coefficients, particularly for hypothesis testing and constructing confidence intervals.

```
# This is a random dataset generated by R that has a normal distribution
random_normal_data <- data.frame(values = rnorm(1000))  # 1000 random values from a normal distribution

# Create a histogram using random data
normal_plot <- ggplot(random_normal_data, aes(x = values)) +
  geom_histogram(fill = "blue", color = "black") +
  labs(title = "Histogram of Values", x = "Values", y = "Frequency")
```

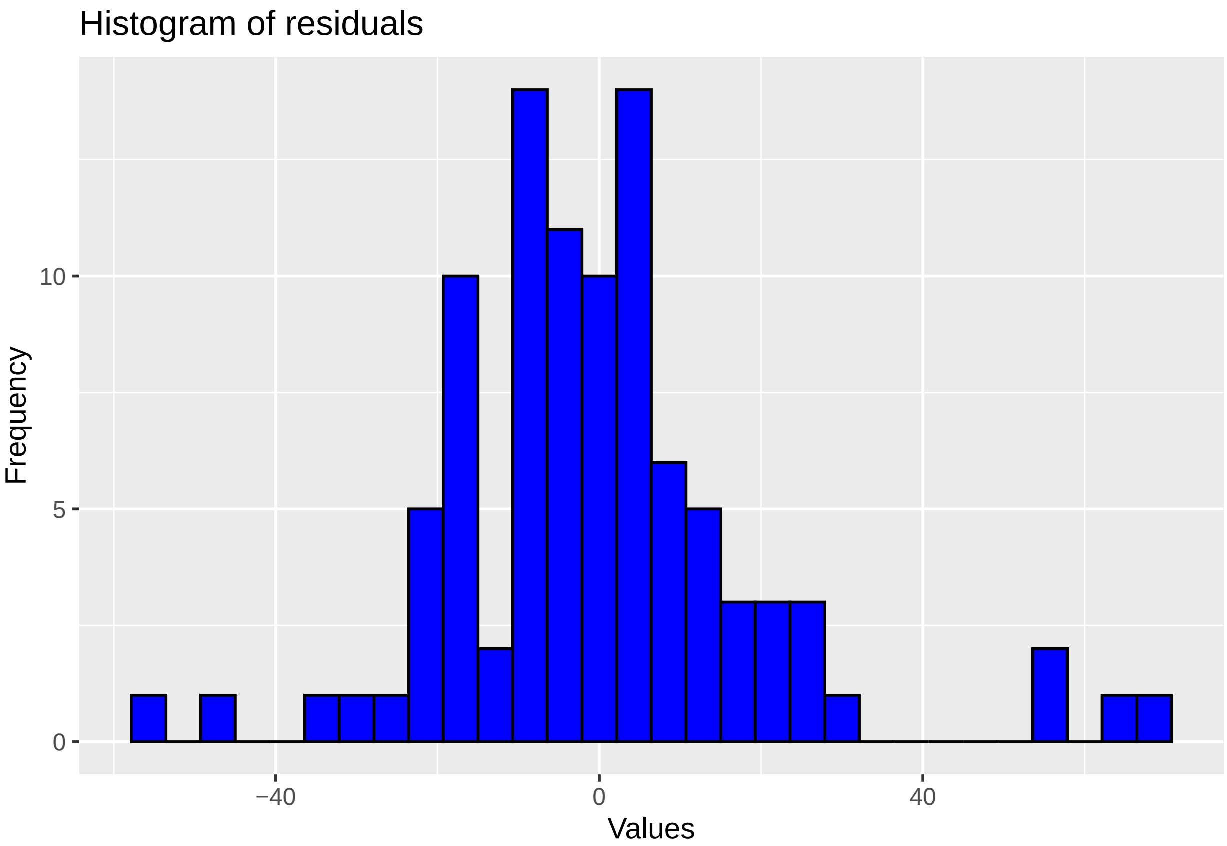
```
# Create a histogram using residuals of our model
model_residuals <- resid(model)
resid_plot <- ggplot(cigarettes, aes(x = model_residuals)) +
  geom_histogram(fill = "blue", color = "black") +
  labs(title = "Histogram of residuals", x = "Values", y = "Frequency")

print(normal_plot)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
print(resid_plot)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We can use the Shapiro-Wilk test for normality to know if the sample data is normaly distributed, where:

$$H_0 : \text{sample data has normal distribution}$$
$$H_1 : \text{sample data does not have normal distribution}$$

```
# Estimating Shapiro-Wilk test for normality
print(shapiro.test(random_normal_data$values)) # normal distributed data (we know for sure this is norm
```

```
##
##  Shapiro-Wilk normality test
##
## data:  random_normal_data$values
## W = 0.99895, p-value = 0.8487
```

```
print(shapiro.test(model_residuals)) # residuals from ciggaretes data (we do not know for sure whether
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model_residuals
## W = 0.92418, p-value = 3.383e-05
```

Based on the p-values, we have enough evidence to reject the null hypothesis of this test, concluding that the residuals estimated from the regression model using cigarettes.csv is not normally distributed.

5. Zero mean of errors

This implies that the error term should be completely *random*, meaning it's mean is zero $E[u] = 0$,

```r
# Calculating mean of residuals
residual_mean <- mean(model_residuals)
print(residual_mean)
```

```
## [1] -7.771561e-16
```

6. Exogeneity (no endogeneity)

The error term should not be correlated with any of the independent variables $E[u|X] = 0$. In other words, the independent variables should be determined outside the model (hence, the word *'exogenous'*, related to external) and not be influenced by the errors. If the regressors are correlated with the error term (*endogeneity*), the OLS estimates will be biased and inconsistent, making them unreliable for causal inference.

Endogeneity may be cause by many things, such as:

- **Selection problem**: This issue occurs when there is a non-random selection of observations, leading to a sample that is not representative of the population. True model could be $Y = X\beta + u$, but instead we estimate $Y = X\beta + v$. This biased selection process may cause that the observations of X may be correlated to the error term.
- **Simultaneity bias (simultaneous causality)**: This occurs when two or more variables in the model are determined simultaneously by each other. This relationship may imply that X is influenced by the dependent variable y. Famous example: Do good institutions cause economic development, or countries have good institutions because of their developed economy? AJR (2001) won a Noble Prize and gave an answer to this question…
- **Omitted variable bias**: This is one of the most common causes of endogeneity. True model might be $Y = X_1\beta_1 + X_2\beta_2 + u$, but we estimate $Y = X_1\beta_1 + v$, where $v = X_2\beta_2 + u$. The omitted variables X_2 could be correlated with the included variables X_1 and with the error term u, resulting in a correlation between the explanatory variables and the error term. Excluding X_2 when $\beta_2 \neq 0$, meaning it affects y, may also be an issue as that variation may be represented in β_1, thus we have a β_1 that is over or underrepresented.
- **Measurement error (Observational error)**: This happens when the explanatory variables are measured with error, which causes a difference between a measured value of a quantity and its unknown true value. Such errors are inherent in the measurement process. True model could be $Y = X^*\beta + u$, but observed model is $y = X\beta + v$, where $X^* = X + \epsilon$ (ϵ is some error that captures the difference between true X^* variable data and the biased sample data we collect X), and therefore $v = -\beta\epsilon + u$.

2. Instrumental variables (IVs)

Motivation

We learned about regression models and the problems of **estimation bias** experienced from omitted variables, measurement errors, and simultaneous causality (where X causes Y and vice versa). We also learned that this estimation bias can be corrected somewhat by adding omitted variables into the regression model to mitigate the risk of a biased estimate of any causal effects.

But what do we do when we can't include omitted variables? What if the omitted variables are unavailable, or cannot be measured? Multiple regression cannot solve this problem…

Enter **IV Regression**, a technique for uncovering hidden correlations between our dependent and independent variables through the use of an instrumental variable. Instrumental variables are best described as a third variable (let's call it Z), which is associated with our independent variable X in some way. Z is also associated with dependent variable Y, but **only through** Y's direct association with X.

For example, suppose we want to investigate and estimate the causal effect of smoking (X) on general health (Y). Simply correlating general health and smoking does not imply that smoking causes poor health because other factors such as depression may affect both. Therefore, we need an IV to uncover the relationship between Y and X, as it would be difficult to conduct controlled experiments in smoking in the general population. A possible IV could be using the tax rate on tobacco as an instrument for smoking. Why? Because this tax can only be correlated with general health through its effects on smoking (so, correlated with smoking but not directly with general health). If a relationship between tobacco tax and health is discovered, this provides evidence that smoking causes changes in health.

We use an IV to bridge this gap by essentially splitting the independent variable X in two parts: one that might be correlated with the error term (u), and a part that probably is not. Isolating the latter part allows us to estimate β coefficients.

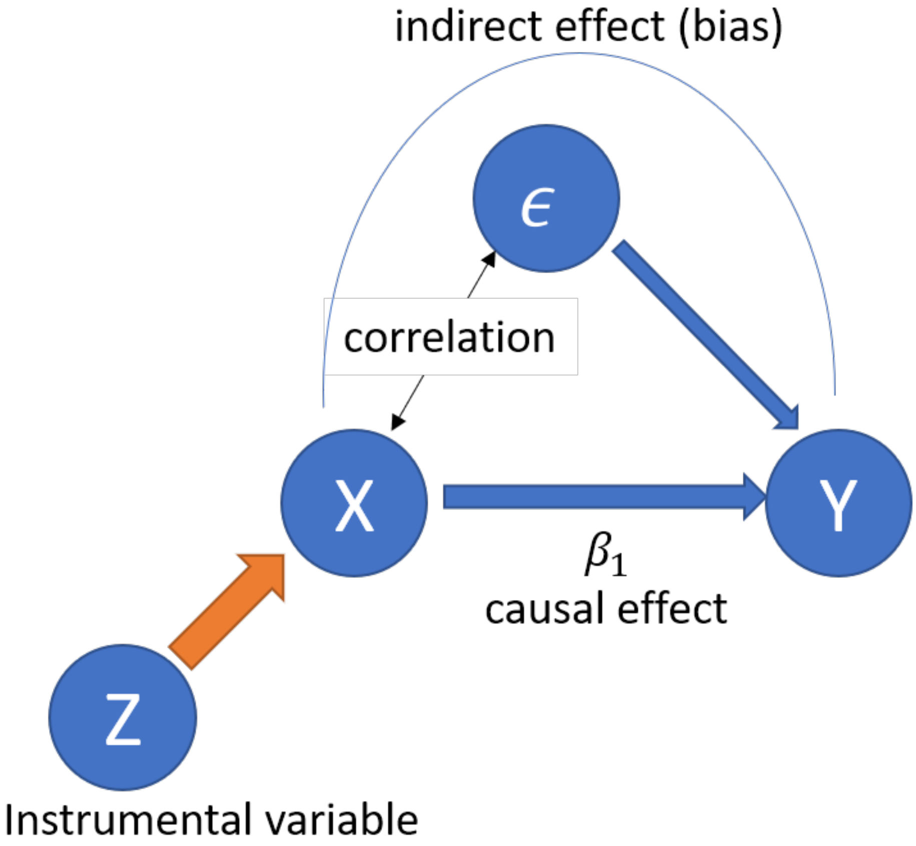


Figure 1: IV outline (Example 1)

What we want is for IV to be:

- correlated with endogenous variable X
- uncorrelated with error term (so, the IV itself is exogenous)
- only impact outcome y through X

How many IVs we need? At least as many instruments as endogenous regressors

- *just identified*: number of predictors = number of instruments
- *over identified*: number of predictors < number of instruments

Conditions for a **valid** IV:

1. **Relevance:** instrument Z is relevant when it is correlated with and explain the variation in X. A weak instrument is one that is not correlated with X and may give us biased results.
2. **Exogeneity:** instrument Z is exogenous when it not correlated with error u

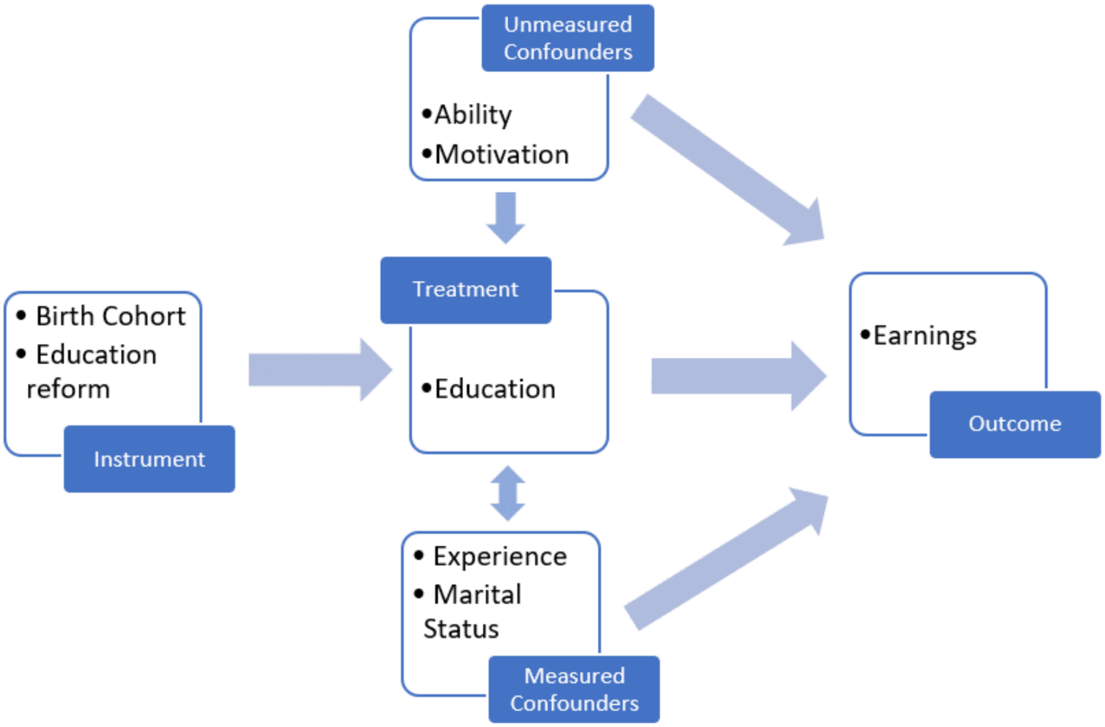


Figure 2: IV outline (Example 2)

Two-Stage Least Squares (2SLS) Estimation

This is our regression model, what we want to estimate:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

where X_1 is the endogenous variable.

We estimate all this using a two-stage procedure:

1. **Stage 1:** Regress X on Z using OLS

Due to IV regression splitting the independent variable X into two parts, we need to also split the process of performing the regression into two parts.

Firstly, we decompose the variation in X into two partitions:

- A "problem-free" X partition that is explained by instrumental variable Z
- A "problematic" X partition that is correlated with error term

Secondly, we use the "problem-free" partition of X to estimate β_1.

The first stage of the model takes the form:

$$X_i = \pi_0 + \pi_1 Z_i + \nu_i$$

Where $\pi_0 + \pi_1 Z_1$ is the "problem-free" partition of X_i, and ν_i is the "problematic" partition of X_i.

We can use OLS to obtain estimates $\widehat{\pi_0}$ and $\widehat{\pi_1}$ and subsequently predicted values $\hat{X}_i$ for $i = 1, ..., n$, if our instrumental variable Z is valid. The $\hat{X}_i$ are problem-free because $\hat{X}$ is exogenous in a regression of Y on $\hat{X}$, which is done in the second stage.

2. **Stage 2:** Regress Y on fitted values $\hat{X}$ using OLS.

After getting fitted values: $\hat{X}_1 = \hat{\pi}_0 + \hat{\pi}_1 Z_i$, we apply OLS on the following:

$$Y = \beta_0 + \beta_1 \hat{X}_1 + \beta_2 X_2 + u$$

Now, we have an unbiased estimator $\hat{\beta}$

Application using data

Back to our example of the effects of smoking on general health, let's test our previously stated conjecture regarding tobacco taxes. Unlike the conjecture, we do not have data for general health of the population, however we can proxy this into quantity of demand for cigarettes.

More formally the question we are asking is how much do we have to increase tobacco taxes by to achieve a certain reduction in smoking. In economics we refer to this as **elasticity**, or the percentage change in one variable in response to a percentage change in another. More specifically, we are dealing with a price elasticity for the demand of cigarettes which is unknown and must be estimated.

We will use IV, because we cannot use OLS regression in this scenario due to simultaneous causality between supply and demand of cigarettes.

Firstly, load the data into R with the **cigarettes.csv** file provided (don't forget to set your working directory!)

```r
cigarettes <- read.csv("cigarettes.csv")
```

Now, let's get a summary of our data:

```r
summary(cigarettes)
```

```
##        X            state                 year           cpi
##  Min.   : 1.00   Length:96          Min.   :1985   Min.   :1.076
##  1st Qu.:24.75   Class :character   1st Qu.:1985   1st Qu.:1.076
##  Median :48.50   Mode  :character   Median :1990   Median :1.300
##  Mean   :48.50                      Mean   :1990   Mean   :1.300
##  3rd Qu.:72.25                      3rd Qu.:1995   3rd Qu.:1.524
##  Max.   :96.00                      Max.   :1995   Max.   :1.524
##    population           packs            income               tax
##  Min.   :  478447   Min.   : 49.27   Min.   :  6887097   Min.   :18.00
##  1st Qu.: 1622606   1st Qu.: 92.45   1st Qu.: 25520384   1st Qu.:31.00
##  Median : 3697472   Median :110.16   Median : 61661644   Median :37.00
##  Mean   : 5168866   Mean   :109.18   Mean   : 99878736   Mean   :42.68
##  3rd Qu.: 5901500   3rd Qu.:123.52   3rd Qu.:127313964   3rd Qu.:50.88
##  Max.   :31493524   Max.   :197.99   Max.   :771470144   Max.   :99.00
##      price             taxs
##  Min.   : 84.97   Min.   : 21.27
##  1st Qu.:102.71   1st Qu.: 34.77
##  Median :137.72   Median : 41.05
##  Mean   :143.45   Mean   : 48.33
##  3rd Qu.:176.15   3rd Qu.: 59.48
##  Max.   :240.85   Max.   :112.63
```

We can see some large variances in variables relating to cigarette price and tax, alongside income and population.

Model and Estimation In this exercise, we are interested in estimating β_1 in the model:

$$\ln(Q_i^{cigarettes}) = \beta_0 + \beta_1 \ln(P_i^{cigarettes}) + u_i$$

Where $Q_i^{cigarettes}$ (Y) is the quantity of cigarette packs sold per capita, and $P_i^{cigarettes}$ (X) is the after-tax real price per cigarette pack in state i.

tax is the average state, federal, and average local excise taxes for fiscal year, and *taxs* is the average excise taxes for fiscal year, including sales taxes. *sales_tax* is the difference between the two, which we will use as an IV to estimate endogenous variable $P_i^{cigarettes}$, as it is included in after tax average price per pack. We can also presume *sales_tax* to be exogenous as sales tax does not directly influence quantity of packs sold, but influences it indirectly through the price.

We will have to compute *sales_tax* as it is not included in the dataset. Additionally, for this analysis we are only interested in the data for the year 1995. Finally, we must compute real prices by dividing prices and taxes by the *cpi* figure provided.

1. Create subset of data for 1995

```r
cigs_1995 <- subset(cigarettes, year == "1995")
# or, use:
cigs_1995 <- cigarettes %>%
  filter(year == "1995")
```

2. Compute real per capita prices

```r
cigs_1995$real_price <- with(cigs_1995, price / cpi)
# or, use:
cigs_1995$real_price <- cigs_1995$price / cigs_1995$cpi
```

3. Compute sales tax

```
cigs_1995$sales_tax <- with(cigs_1995, (taxs - tax) / cpi)
```

Next, we do a correlation check between *sales_tax* and nominal *real_price*, and between the former and *packs*:

```
print(cor(cigs_1995$sales_tax, cigs_1995$real_price))
```

```
## [1] 0.6856138
```

```
print(cor(cigs_1995$sales_tax, cigs_1995$packs))
```

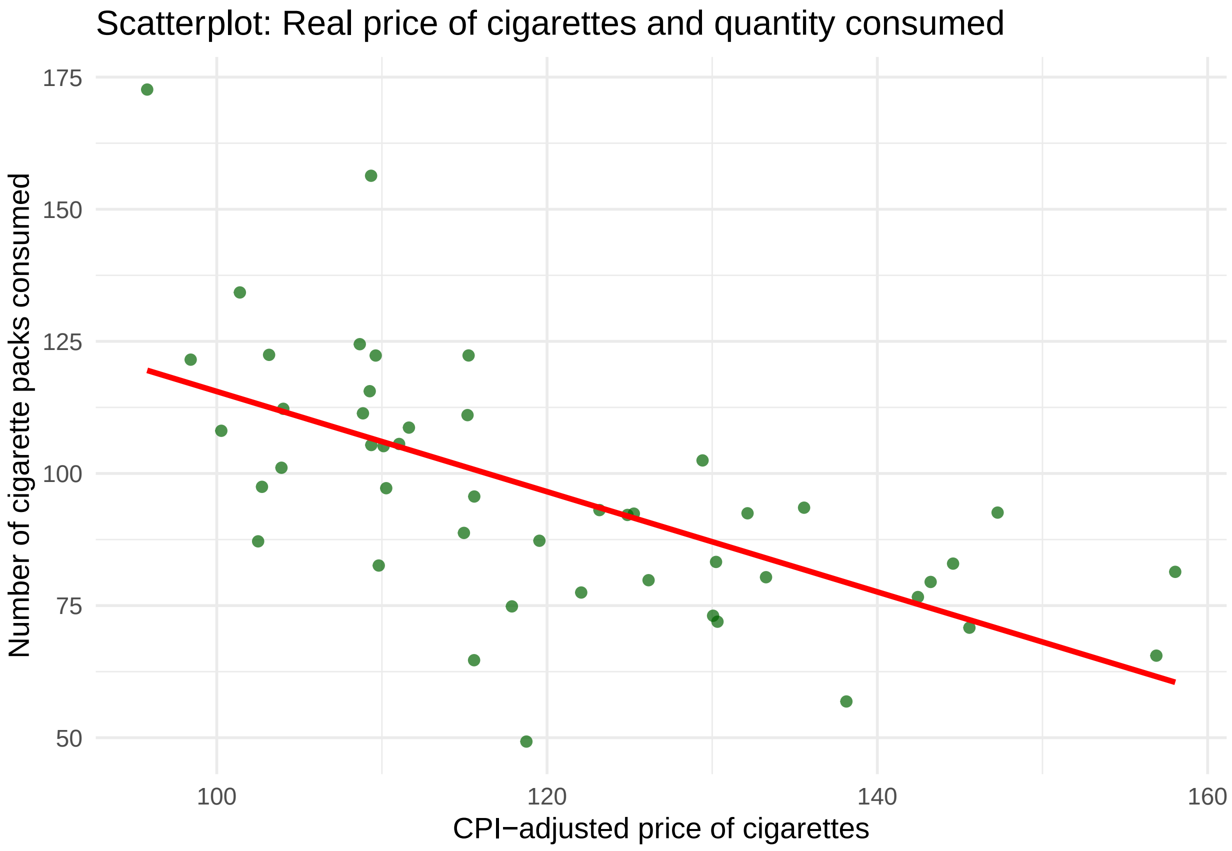
```
## [1] -0.3845821
```

We have a semi-strong positive correlation between *sales_tax* and nominal *real_price*, indicating that taxes increase with prices.

Also, let's plot a scatter of the two variables to see visually if there are any relationships:

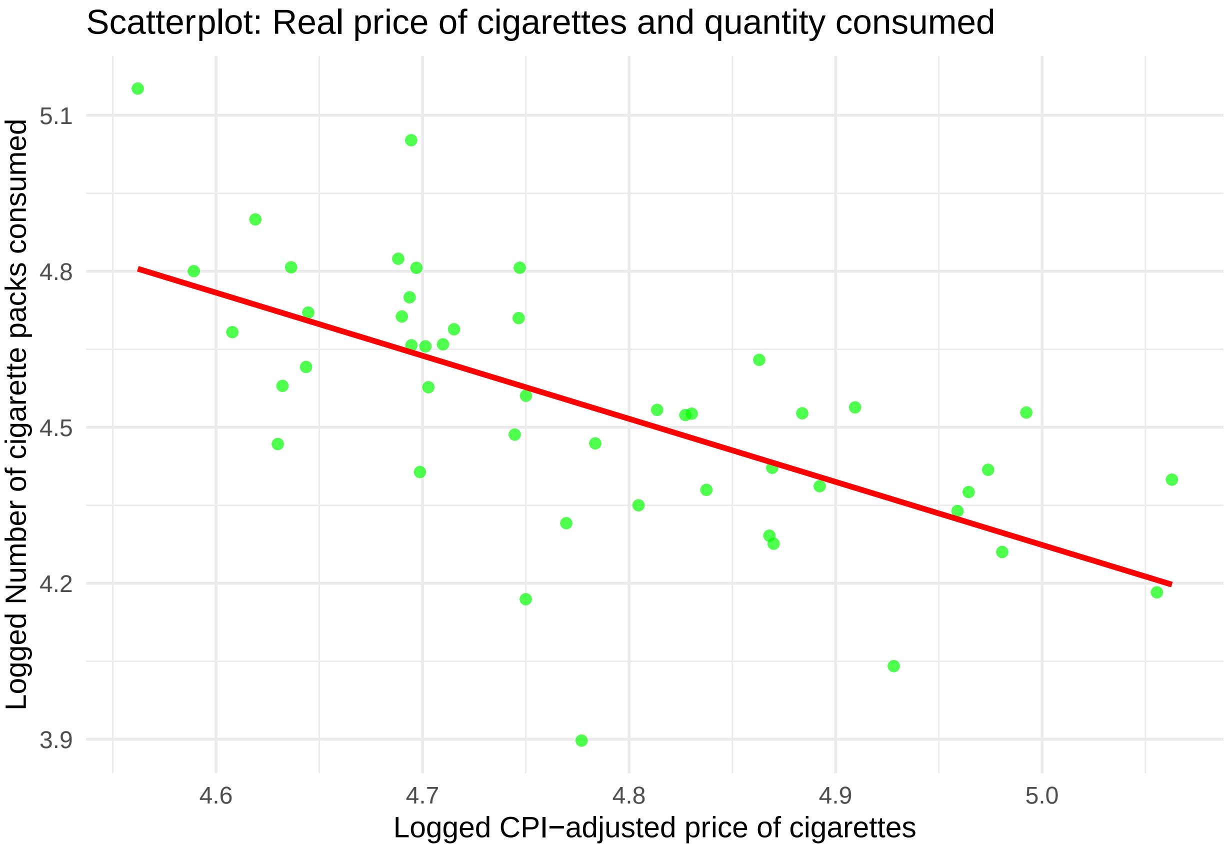
```
ggplot(cigs_1995, aes(x = real_price, y = packs)) +
  geom_point(color = "darkgreen", alpha = 0.7) +  # Scatter points
  geom_smooth(method = "lm", color = "red", se = FALSE) +  # Add regression line
  labs(
    title = "Scatterplot: Real price of cigarettes and quantity consumed",  # Title
    x = "CPI-adjusted price of cigarettes",  # X-axis label
    y = "Number of cigarette packs consumed"  # Y-axis label
  ) +
  theme_minimal()  # Clean minimal theme
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```r
ggplot(cigs_1995, aes(x = log(real_price), y = log(packs))) +
  geom_point(color = "green", alpha = 0.7) +  # Scatter points
  geom_smooth(method = "lm", color = "red", se = FALSE) +  # Add regression line
  labs(
    title = "Scatterplot: Real price of cigarettes and quantity consumed",  # Title
    x = "Logged CPI-adjusted price of cigarettes",  # X-axis label
    y = "Logged Number of cigarette packs consumed"  # Y-axis label
  ) +
  theme_minimal()  # Clean minimal theme
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



We can see a clear negative relationship between the the price of cigarettes and quantity sold. OLS helps estimate the model $\ln(Q_i^{cigarettes}) = \beta_0 + \beta_1 \ln(P_i^{cigarettes}) + u_i$. We see that the estimate of the coefficient that describes the impact of *real_price* on *packs* is **-1.21**, meaning that for 1% increase in price, we have a 1.21 in quantity. We will show that this coefficient is biased when accounting for an instrumental variable.

```r
# perform first stage
cigs_ols <- lm(log(packs) ~ log(real_price), data = cigs_1995)
summary(cigs_ols)
```

```
##
## Call:
## lm(formula = log(packs) ~ log(real_price), data = cigs_1995)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64676 -0.09030  0.01787  0.11245  0.40779
```

```
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       10.3389     1.0353   9.986 4.25e-13 ***
## log(real_price)   -1.2131     0.2164  -5.604 1.13e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1896 on 46 degrees of freedom
## Multiple R-squared:  0.4058, Adjusted R-squared:  0.3928
## F-statistic: 31.41 on 1 and 46 DF,  p-value: 1.13e-06
```

Stage 1 Step one of our regression is to split the "problematic" and "problem-free" portions:

$$\ln(P_i^{cigarettes}) = \pi_0 + \pi_1 sales\_\_tax_i + \nu_i$$

We will estimate this model using lm():

```
# perform first stage
cigs_stage_1 <- lm(log(real_price) ~ sales_tax, data = cigs_1995)

# coefficient test (t-test)
summary(cigs_stage_1)
```

```
##
## Call:
## lm(formula = log(real_price) ~ sales_tax, data = cigs_1995)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.221027 -0.044324  0.000111  0.063730  0.210717
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.616546   0.029108   158.6  < 2e-16 ***
## sales_tax   0.030729   0.004802     6.4 7.27e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09394 on 46 degrees of freedom
## Multiple R-squared:  0.471,  Adjusted R-squared:  0.4595
## F-statistic: 40.96 on 1 and 46 DF,  p-value: 7.271e-08
```

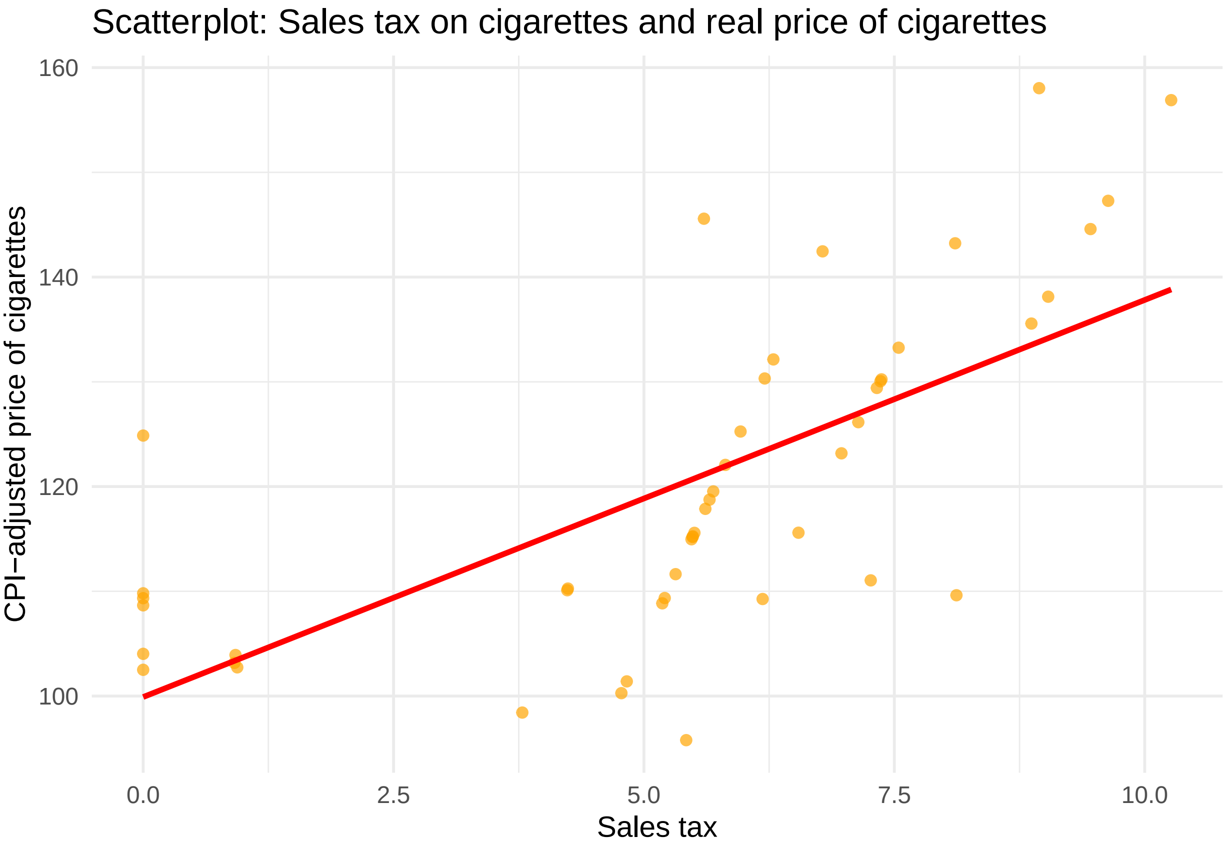
Substituting figures into our regression model gives us:

$$\ln(\widehat{P_i^{cigarettes}}) = 4.62 + 0.031 sales\_\_tax_i$$

This is how our prediction looks like graphically

```
ggplot(cigs_1995, aes(x = sales_tax, y = real_price)) +
  geom_point(color = "orange", alpha = 0.7) +  # Scatter points
  geom_smooth(method = "lm", color = "red", se = FALSE) +  # Add regression line
  labs(
    title = "Scatterplot: Sales tax on cigarettes and real price of cigarettes",  # Title
    x = "Sales tax",   # X-axis label
    y = "CPI-adjusted price of cigarettes"  # Y-axis label
  ) +
  theme_minimal()  # Clean minimal theme
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Our model predicts the relationship between $sales\_tax$ and $real\_price$ to be positive, which is true according to the laws of supply-demand. A higher sales tax will force producer to increase prices. But how much of the observed variation in ($P^{cigarettes}$) can be explained by the IV $sales\_tax$?

We can call the R^2 of the regression to find an answer:

```
summary(cigs_stage_1)$r.squared
```

```
## [1] 0.4709961
```

47% of the variance in post-tax $real\_price$ appears to be explained by $sales\_tax$ across different states.

We can test the relevance of the IV using the Wald test. From $\ln(P_i^{cigarettes}) = \pi_0 + \pi_1 sales\_tax_i + \nu_i$, we test for the following hypothesis:

$$H_0 : cov[Z, X] = 0, or, \pi_1 = 0$$
$$H_1 : cov[Z, X] \neq 0, or, \pi_1 \neq 0$$

Notice that this is the same p-value as the one in the OLS regression above. The low p-value below 1% leads us to assume that we do not have enough evidence to reject the null hypothesis that Z ($sales\_tax$) has no impact on X ($real\_price$). Therefore, we believe that $sales\_tax$ does have a significant positive impact on $real\_price$ at 1% level of significance.

```r
first_stage_ftest <- waldtest(cigs_stage_1, .~1)
print(first_stage_ftest)
```

```
## Wald test
##
## Model 1: log(real_price) ~ sales_tax
## Model 2: log(real_price) ~ 1
##   Res.Df Df      F    Pr(>F)
## 1     46
## 2     47 -1 40.956 7.271e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Stage 2 Before we begin step two of our regression, we need to save the estimator $\ln(\widehat{P_i^{cigarettes}})$ from step one:

```r
pred_cigs_stage_1 <- cigs_stage_1$fitted.values
```

These values are the ones estimated by the OLS, seen in the red fitted line in the scatterplot above in the first stage regression.

Now we run the second step, a regression of $\ln(Q_i^{cigarettes})$ on $\ln(\widehat{P_i^{cigarettes}})$ to find $\hat{\beta}_0^{TSLS}$ and $\hat{\beta}_1^{TSLS}$

```r
# perform second stage
cigs_stage_2 <- lm(log(cigs_1995$packs) ~ pred_cigs_stage_1)

# coefficient test (t-test)
summary(cigs_stage_2)
```

```
##
## Call:
## lm(formula = log(cigs_1995$packs) ~ pred_cigs_stage_1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63180 -0.15802  0.00524  0.13574  0.61434
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         9.7199     1.8012   5.396   2.3e-06 ***
## pred_cigs_stage_1  -1.0836     0.3766  -2.877   0.00607 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2264 on 46 degrees of freedom
## Multiple R-squared:  0.1525, Adjusted R-squared:  0.1341
## F-statistic: 8.277 on 1 and 46 DF,  p-value: 0.006069
```

Substituting into our step two model gives:

$$\ln(\widehat{Q_i^{cigarettes}}) = 9.72 - 1.08\ln(\widehat{P_i^{cigarettes}})$$

We can see that the “problem-free” 2SLS estimators for β_1 suggest an increase in cigarette prices by 1% (one unit) reduces consumption by 1.08 percentage points - fairly elastic. The main idea here was to have a more accurate $\hat{\beta}_1$. If you remember, our initial OLS said $\hat{\beta}_1$ is 1.21, whereas the less biased two-stage estimator $\hat{\beta}^{2SLS}$ says that it is in fact 1.08. The impact of price on quantity is much lower! This is because the initial regression included external variation of other factors (e.g., *sales_tax*) in that coefficient estimate. In severe cases, this difference in estimates (i.e., bias) can be even larger.

However, similar to linear regression, IV regression can still be biased due to omitted variables and other underlying assumptions of the data!

A Coding Shortcut for IV Regressions

As a climate change scientist/ESG analyst/academic researcher, your time is precious. In many instances R has in-built capabilities to perform many tasks, and when they aren't built in they can be sourced via a library! The AER library contains an **ivreg()** function to perform IV regressions without having to split into two parts.

Then, call ivreg() using the same X, Y, and Z variables as before:

```r
# perform IV regression with ivreg()
cigs_ivreg <- ivreg(log(packs) ~ log(real_price) | sales_tax, data = cigs_1995)
summary(cigs_ivreg)
```

```
## 
## Call:
## ivreg(formula = log(packs) ~ log(real_price) | sales_tax, data = cigs_1995)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.64619 -0.07732  0.02981  0.11283  0.41904 
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       9.7199     1.5141   6.420 6.79e-08 ***
## log(real_price)  -1.0836     0.3166  -3.422  0.00131 ** 
## 
## Diagnostic tests:
##                  df1 df2 statistic  p-value    
## Weak instruments   1  46    40.956 7.27e-08 ***
## Wu-Hausman         1  45     0.314    0.578    
## Sargan             0  NA        NA       NA    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1904 on 46 degrees of freedom
## Multiple R-Squared: 0.4011,  Adjusted R-squared: 0.3881 
## Wald test: 11.71 on 1 and 46 DF,  p-value: 0.001313
```

The syntax for ivreg() is Y (log of packs) regressed on (~) X (log of real price) given (|) Z sales tax. Much faster!

Compare to OLS estimate (without IV)

```r
cigs_ols <- lm(log(packs) ~ log(real_price), data = cigs_1995)
summary(cigs_ols)
```

```
## 
## Call:
## lm(formula = log(packs) ~ log(real_price), data = cigs_1995)
## 
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64676 -0.09030  0.01787  0.11245  0.40779
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      10.3389     1.0353   9.986 4.25e-13 ***
## log(real_price)  -1.2131     0.2164  -5.604 1.13e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1896 on 46 degrees of freedom
## Multiple R-squared:  0.4058, Adjusted R-squared:  0.3928
## F-statistic: 31.41 on 1 and 46 DF,  p-value: 1.13e-06
```