# 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

# Binary Choice

## What is Binary Choice?

We will learn about econometric challenges when the **dependent variable can take only two values**. In all previous sections we have implicitly assumed that the dependent variable, denoted by $y$, can take many values (continous). In some situations you may want to model a dependent variable that has only a limited number of possible outcomes.

For example, if you want to analyze the effect of price on brand choice of a certain product your dependent variable Y only takes a limited number of outcomes, as there's only a limited number of brands. The same holds if you want to analyze the influence of income on political party choice, as there are only a limited set of parties to choose from. This situation occurs relatively often in economic and business economic research.

- Answer to yes,no questions.
- Choice for private or public health care
- Vote decision for Democrat or Republican president (USA)
- Choice for private or public transport
- Choice to renew or cancel a mobile phone contract
- Business cycle indicator (expansion or recession)

This data are usually called **Binary Choice data**. Although the dependent variable does not always have to correspond to a real choice of an individual. The $y$ variable may, for example, also indicate the stage of the business cycle (recession or expansion) in which case there is no clear choice by an individual.

When a dependent variable can only take two values, we often translate the outcomes into numerical values for notational convenience. In most applications, the values zero and one are used, but the researcher is free to use any two numbers. You may, for example, use the values minus one and plus one, or zero and 100. In the coming lectures we will discuss the econometric modeling of binary dependent variables.

The first question that may come to your mind is why we cannot simply use linear regression to deal with these variables? Linear regression has some limitations that make it less suited for a binary dependent variable.

We consider data from a survey distributed among a thousand households. They were asked whether they would want to buy a new electronic gadget. Each individual was faced with a different price in dollars and could answer yes or no. We label the dependent variable $response_i$ equal to one if individual i responded yes and zero otherwise.

$$Response_i = \begin{cases} 0 & \text{if response is No} \\ 1 & \text{if response is Yes} \end{cases}$$

```
dataset1 <- read_csv(
  "https://raw.githubusercontent.com/diego-eco/diego-eco.github.io/master/downloads/data5_1.csv")
```

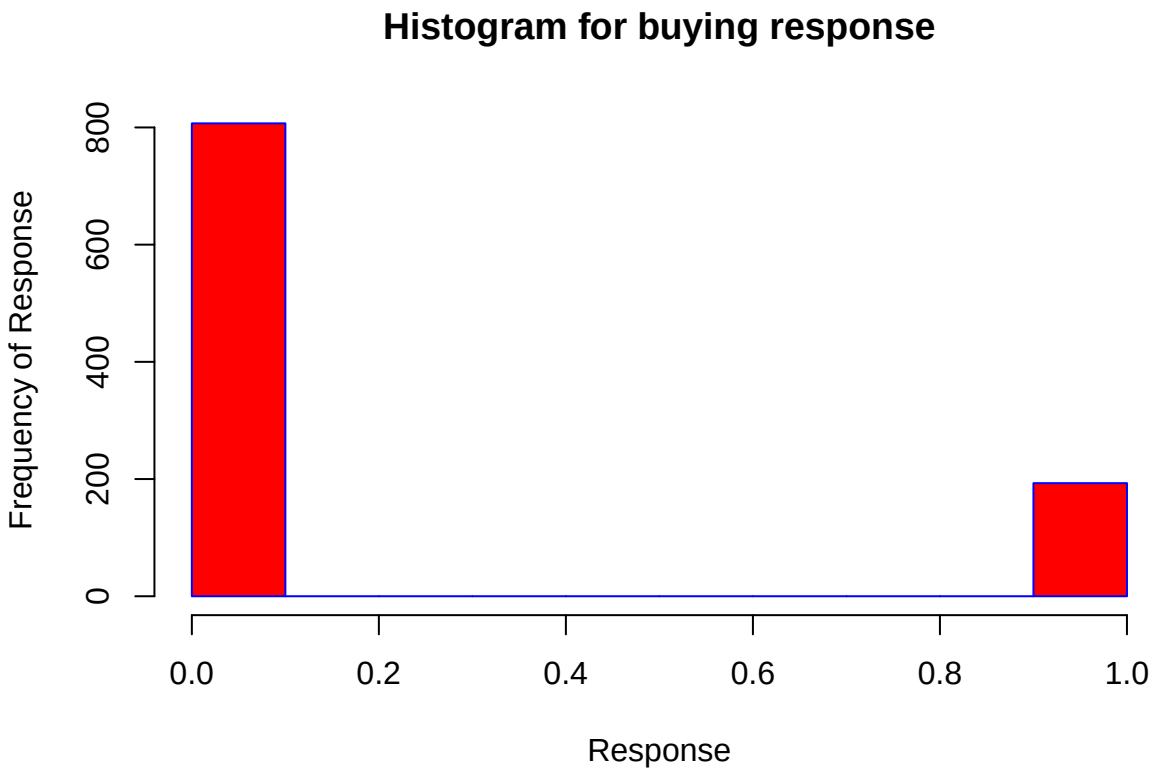
**datalecture5**

Simulated survey data set with 1000 respondents.

- Price: quoted price of electronic gadget (scale variable, in US dollars)
- Response: Answer to the question if respondent would buy the gadget for the quoted price (binary variable, 1 = Yes and 0 = No)

The following graph shows a histogram of the answers. You can see that about 20% of the individuals answered, yes, and about 80% said no.

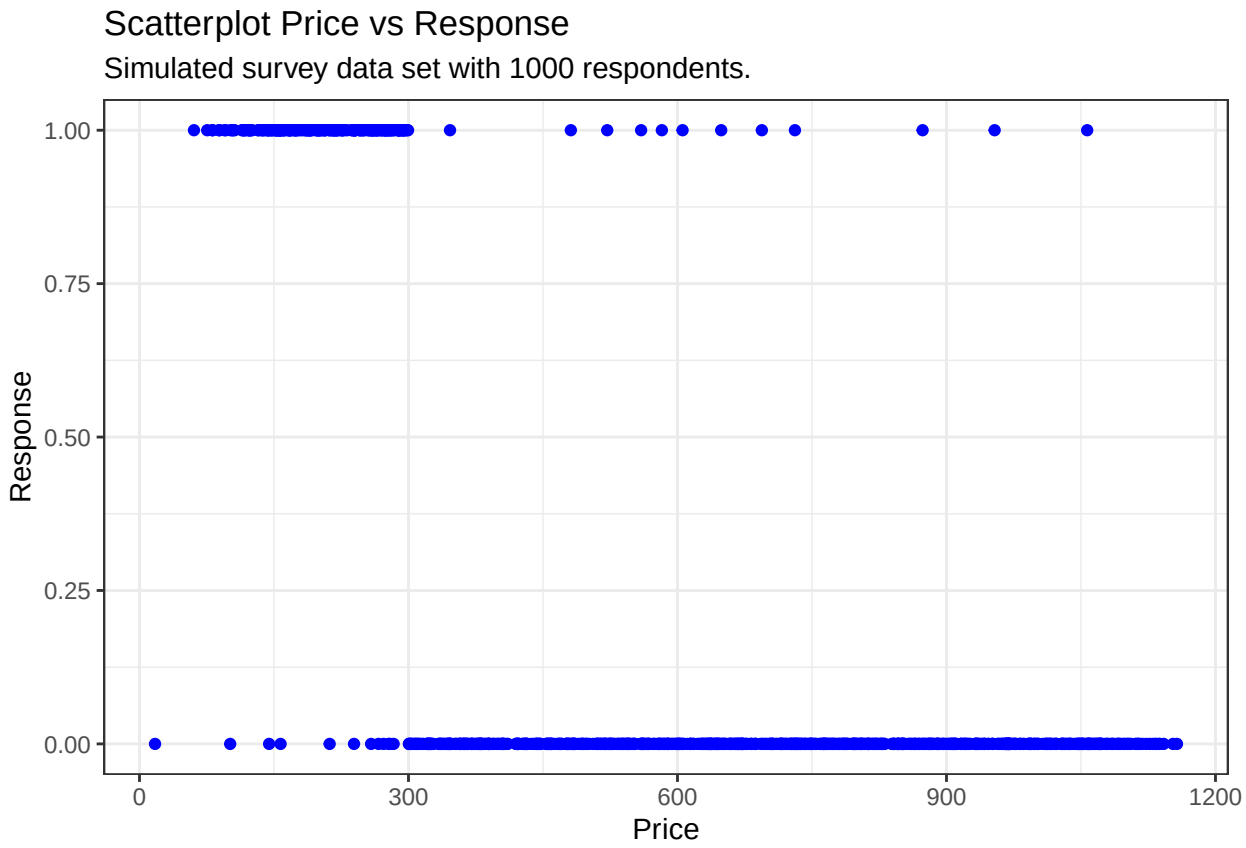
```
hist(dataset1$Response, main="Histogram for buying response",
     xlab="Response", ylab="Frequency of Response",
     border="blue", col="red")
```




It is to be expected that the choice of individuals depends on the price of the new gadget. Next, we can see a scatter diagram of the data. On the horizontal axis, we have price. The price runs from about $10 to $1,200. On the vertical axis you have the corresponding value of response, where a one corresponds with yes and zero with no.

```
plot1 <- ggplot(data=dataset1, aes(x=Price,y=Response)) + geom_point(colour="blue") +
  labs(title="Scatterplot Price vs Response ",
       subtitle="Simulated survey data set with 1000 respondents.")

plot1 + theme_bw()
```




The observations are indicated by circles. As you can see, there are roughly two clusters of observations. There is small cluster of observations at the top left corner. This corresponds to the situation where the price of the gadget is low and individuals want to buy the new gadget. The second cluster is larger and located at the bottom right corner of the graph. This cluster corresponds to no answers and high prices.

Although there are also observations outside the two clusters, in general, the graph suggests that the relation between choice and price is negative. Suppose that you use a linear regression model to describe a binary dependent variable response. That is:

$$response = \beta_1 + \beta_2 \cdot price + \epsilon$$

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Jul 13, 2020 - 22:00:41

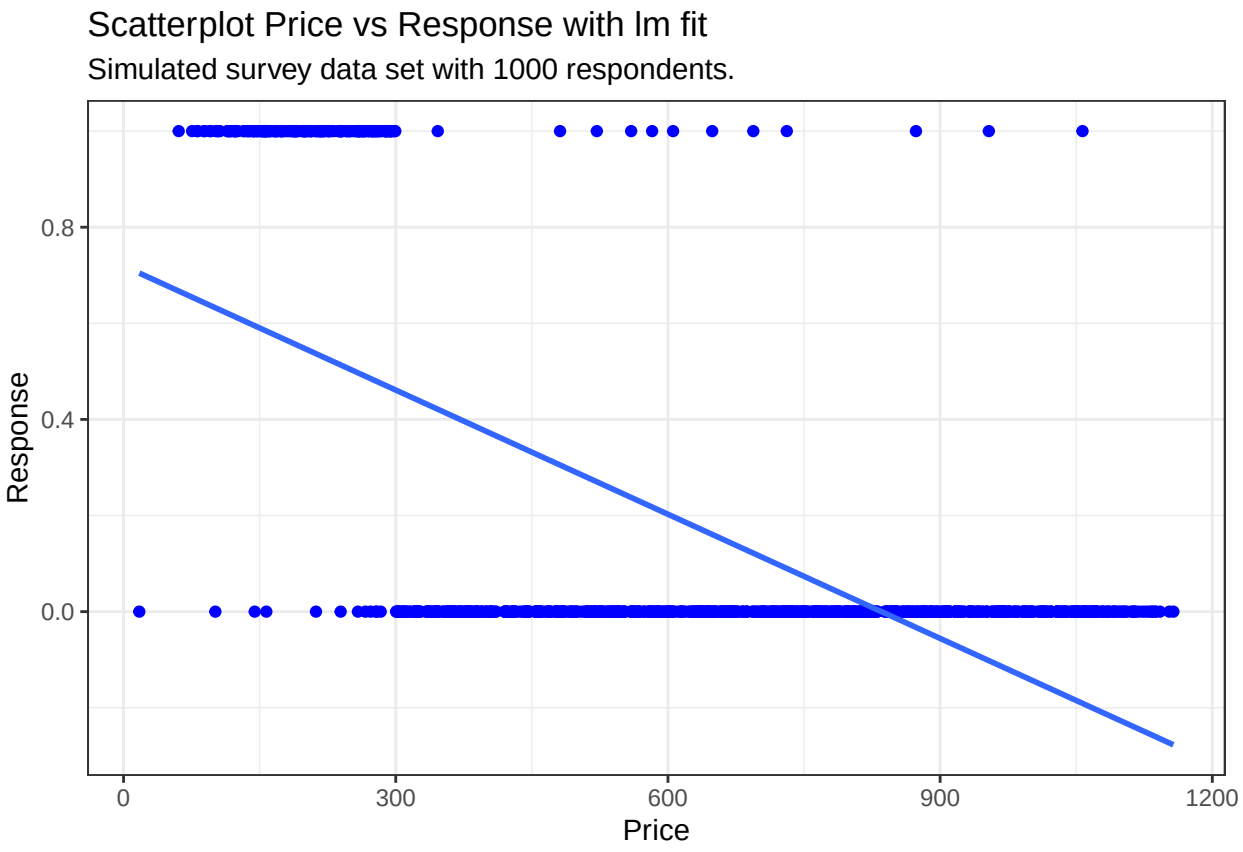
Although the dependent variable can only take two values, we can still apply least squares to estimate the model parameters. The resulting least squares estimate for $\hat{\beta}_2 = \frac{-0.86}{1000}$. Hence regression provides a negative relation between the willingness to buy and price, as suggested by the data.

Although we can directly interpret the size of the $\beta_2$ parameter, the interpretation of the size of this parameter is more difficult. Let us return to the scatter diagram shown before. Now, I also include the regression line on the graph.

```
plot1 <- ggplot(data=dataset1, aes(x=Price,y=Response)) + geom_point(colour="blue") +
  geom_smooth(method = 'lm',se=F) +
  labs(title="Scatterplot Price vs Response with lm fit",
       subtitle="Simulated survey data set with 1000 respondents.")
plot1 + theme_bw()
```

Tabla 1: Linear model on binary choice

| | *Dependent variable:* |
|---|---|
| | Response |
| Price | $-0.001^{***}$ |
| | (0.00003) |
| Constant | $0.720^{***}$ |
| | (0.022) |
| Observations | 1,000 |
| $R^2$ | 0.404 |
| Adjusted $R^2$ | 0.404 |
| Residual Std. Error | 0.305 (df = 998) |
| F Statistic | $677.094^{***}$ (df = 1; 998) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |




$$response = 0.7195 + \frac{-0.86}{1000} \cdot price + e$$

Several things can be observed. First of all, the regression line does not cross the cluster of data points in the top left corner. As the majority of the response observation is zero, the regression line turns out to be flat to make the residuals belonging to the many 0 observations small. More importantly, the fitted line does not lead to zero or one predictions, but takes values between zero and 0.7, and in fact even values smaller than

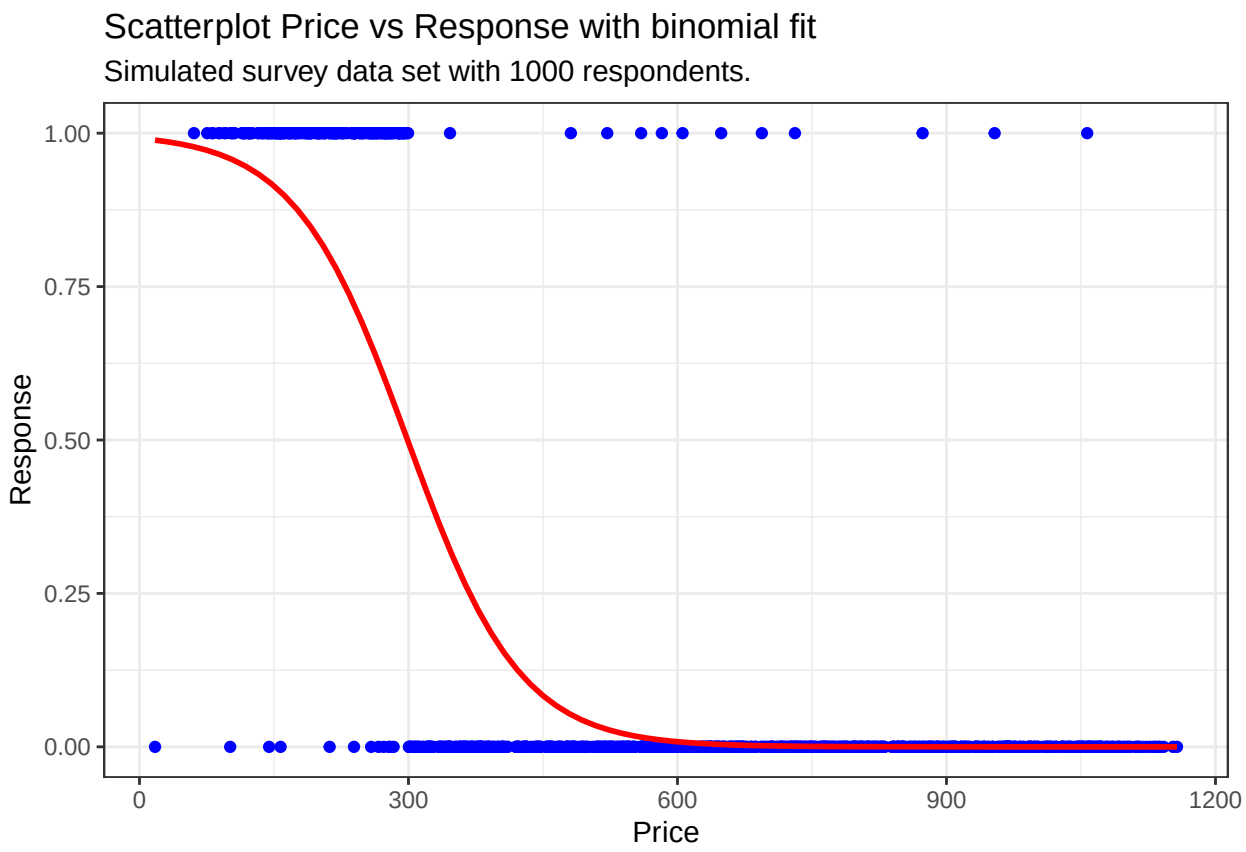
zero. **The fit of the regression line is not in line with the binary character of the dependent variable.**

Finally, the slope of the regression line is the same for every value of price. This means that the model predicts that the effect of a price change is always the same. This does not seem plausible from an economic point of view, and not supported by the data. If the price of the electronic gadget is \$300, changing the price to \$350, will have a large effect on choice. However, if the prize is \$1,100, an increase of \$50 in price will likely have little effect as almost nobody considers buying at this price. The same is true if you change the price from \$10 to \$60 as the data show that many people are prepared to pay a price of \$100. Hence, the linear relation between response and price does not seem to be plausible.

To summarize, although the linear regression model seems to indicate the right direction of the relation between price and choice, the interpretation of the fitted regression line, and of the slope parameter $\beta_2$ is difficult.

In the next section we will propose an econometric model that is especially designed for binary dependent variables, and when the parameter has a more natural interpretation. The model will explicitly deal with the binary character of the dependent variable by describing the probability that the dependent variable takes the value zero or one. The fit of such a model will look like the following curve:

```
plot1 <- ggplot(data=dataset1, aes(x=Price,y=Response)) + geom_point(colour="blue") +
  geom_smooth(method = "glm", se = FALSE, color = "red", method.args = list(family = "binomial")) +
  labs(title="Scatterplot Price vs Response with binomial fit",
       subtitle="Simulated survey data set with 1000 respondents.")
plot1 + theme_bw()
```




As you can see, all predicted values of this model fall between zero and one. The model allows for a non-linear effect of price on choice in the sense that price effects are relatively large for the moderate prices and smaller for very high and low prices.

**Some properties of binary choice models**

Suppose that $y_i$ is a binary dependent variable and that $y_i$ can only take the values 0 and 1 for $i = 1, ..., n$. Consider the linear regression model. Assume $E(\epsilon_i) = 0$

$$y_i = \beta_1 + \beta_2 \cdot x_i + \epsilon_i$$

- How can we express the expected value of $y_i$ expressed in terms of the parameters and $x_i$?

$$E(y_i) = E(\beta_1 + \beta_2 \cdot x_i + \epsilon_i) = \beta_1 + \beta_2 \cdot x_i$$

- With this result we can show that the expected value of $y_i$ equals the probability that $y_i$ equals 1. $E(y_i) = Pr(y = 1)$

$$E(y_i) = 1 \cdot Pr(y = 1) + 0 \cdot Pr(y = 0) = Pr(y = 1)$$

- What is the probability that $y_i$ equals 0 expressed in terms of $x_i$ and the $\beta$ parameters?

$$Pr(y = 0) = 1 - Pr(y = 1) = 1 - \beta_1 + \beta_2 \cdot x_i$$

- Since $y_i$ can only take two values, there are two possible values for the error term given the value of $x_i$ and the parameters $\beta$. Give these two values and also provide the probability that these two values occur.

$$\epsilon_i = \begin{cases} 1 - \beta_1 + \beta_2 \cdot x_i & \text{occurs with Prob : } \beta_1 + \beta_2 \cdot x_i \\ 0 - \beta_1 + \beta_2 \cdot x_i & \text{occurs with Prob : } 1 - \beta_1 + \beta_2 \cdot x_i \end{cases}$$

- What is the variance of $\epsilon_i$ expressed in terms of $x_i$ and the $\beta$ parameters? Are the errors homoscedastic?

$$Var(\epsilon_i) = E[(\epsilon_i - E(\epsilon_i))^2] = E[\epsilon_i^2]$$
$$Var(\epsilon_i) = (1 - \beta_1 + \beta_2 \cdot x_i)^2 \cdot Pr(y_i = 1) + (-\beta_1 + \beta_2 \cdot x_i)^2 \cdot Pr(y_i = 0)$$
$$Var(\epsilon_i) = (1 - \beta_1 + \beta_2 \cdot x_i)^2 \cdot (\beta_1 + \beta_2 \cdot x_i) + (-\beta_1 + \beta_2 \cdot x_i)^2 \cdot (1 - \beta_1 + \beta_2 \cdot x_i)$$
$$Var(\epsilon_i) = (1 - \beta_1 + \beta_2 \cdot x_i)(\beta_1 + \beta_2 \cdot x_i)(1 - \beta_1 + \beta_2 \cdot x_i + \beta_1 + \beta_2 \cdot x_i) = (1 - \beta_1 + \beta_2 \cdot x_i)(\beta_1 + \beta_2 \cdot x_i)$$

The variance of the errors are different for each observation, therefore the errors are heteroscedastic.

## Specify binary choice