# 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")
```

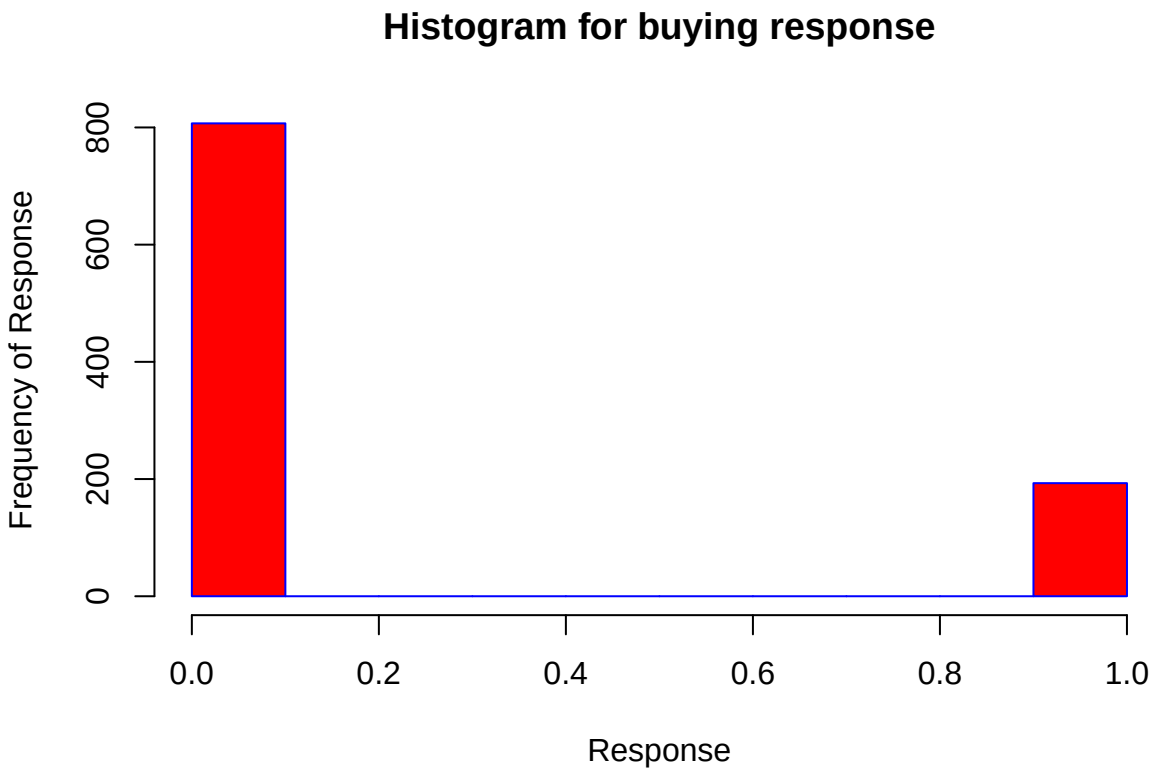
**datalecture5**

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.

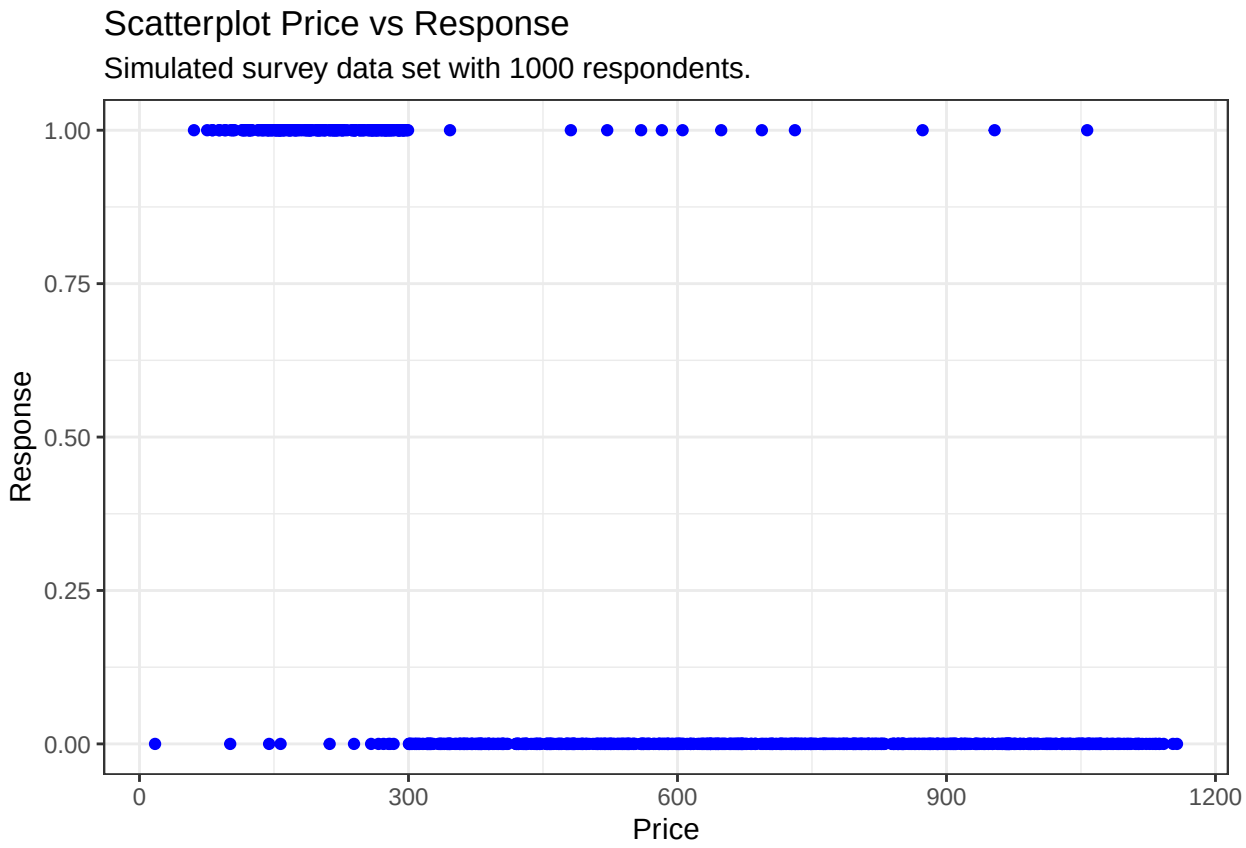
```
hist(dataset1$Response, main="Histogram for buying response",
     xlab="Response", ylab="Frequency of Response",
     border="blue", col="red")
```




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: Tue, Jul 14, 2020 - 22:38:09

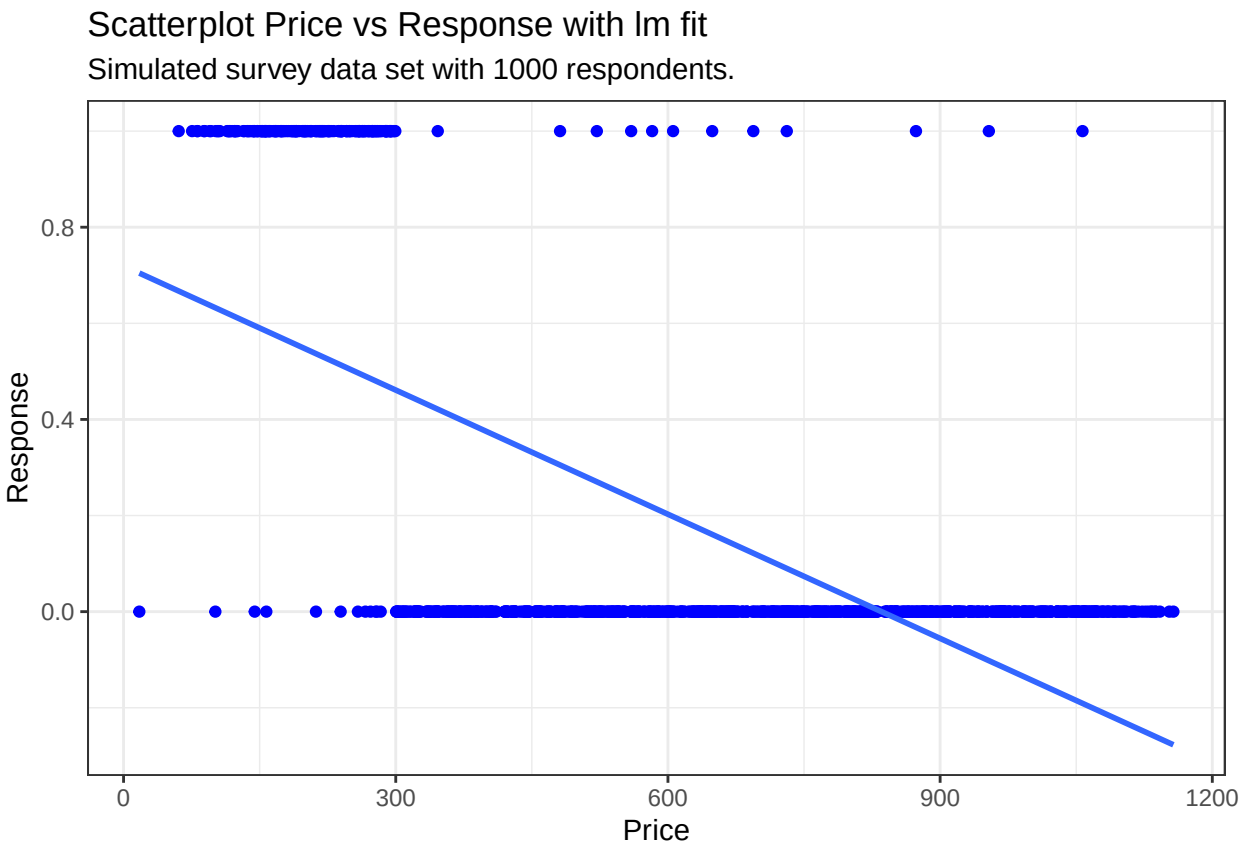
Although the dependent variable can only take two values, we can still apply least squares to estimate the model parameters. The resulting least squares estimate for $\hat{\beta}_2 = \frac{-0.86}{1000}$. Hence regression provides a negative relation between the willingness to buy and price, as suggested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

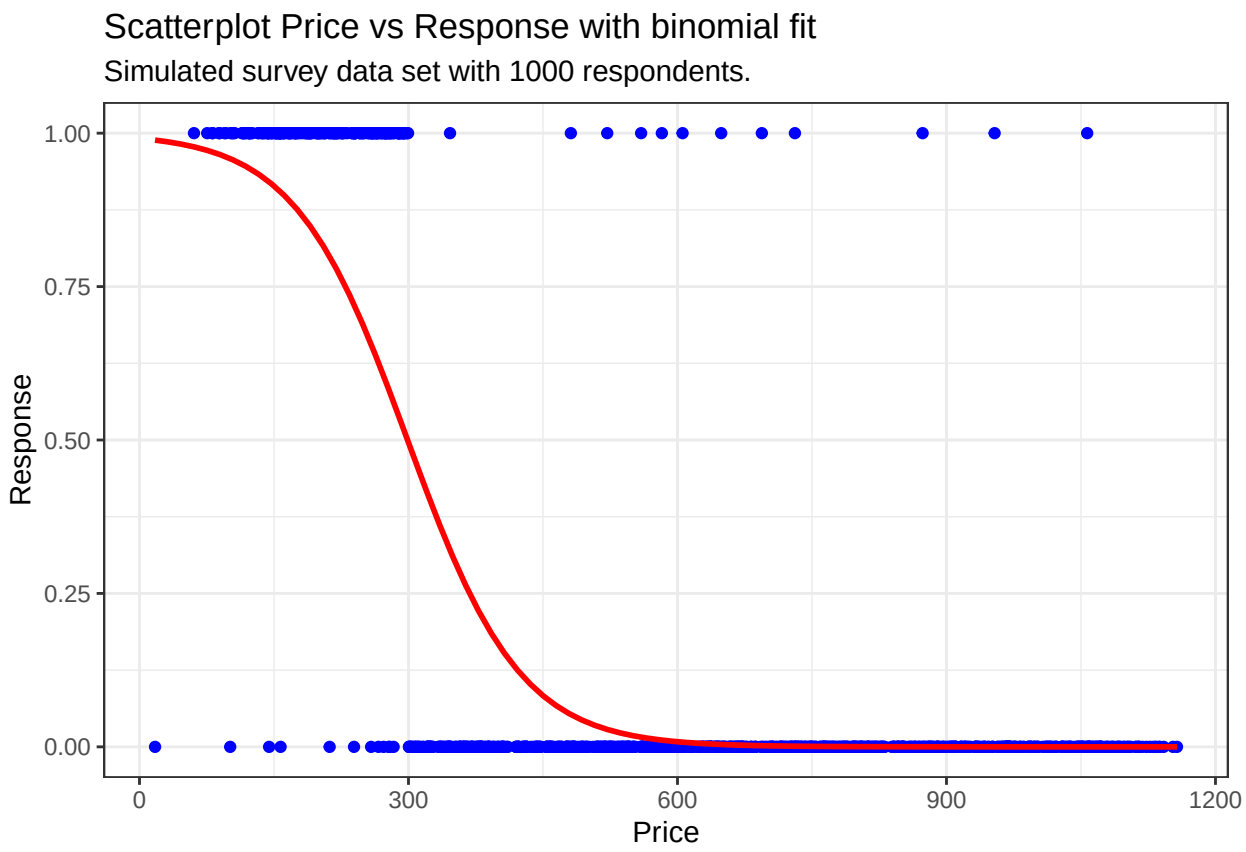
zero. **The fit of the regression line is not in line with the binary character of the dependent variable.**

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

We will call the dependent variable $y_i$ and it can take only two values. We note these values by zero and one. The value one may, for example, correspond to yes and the value zero to no. Instead of assuming a normal distribution for $y_i$, as is done in previous sections, we now assume that y follows a Bernoulli distribution with parameter $\pi$.

The Bernoulli distribution implies that $y_i$ has two possible outcomes denoted by 0 and 1 the probability that the outcome is one equals $\pi$.

$$y_i \sim Bernoulli(\pi)$$
$$\text{Such that } \pi = Pr[y_i = 1] \text{ with } 0 < \pi < 1$$
$$\text{And hence } Pr[y_i = 0] = 1 - \pi$$

The probability that $y_i$ is zero is then $1 - \pi$, as probabilities have to sum up to one. Note that **instead of modeling the value of y itself, we now model the probability that y is 1**. For many applications,

it is very unlikely that the probability that $y_i = 1$ is the same for all observations. Therefore we allow the probability $\pi$ to differ across individuals, or time periods in case we have observations over time. **Hence we consider pi subscript i.**

$$\text{Individual specific probabilities : } Pr[y_i = 1] = \pi_i$$

### Logistic function

To explain differences in probabilities across individuals, we can relate the probabilities $\pi_i$ to one or more explanatory variables. Given the fact that probabilities have to be between 0 and 1, we cannot just take any function for this relationship. Although there are several choices possible, in practice, the **logistic function** is most frequently chosen. A logistic function or logistic curve is a common S-shaped curve (sigmoid curve) with equation:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

Where $x_0$ the $x$ value of the sigmoid's midpoint, $L$ is the curve's maximum value and $k$ the logistic growth rate or steepness of the curve.

To simplify the discussion, we first consider a situation with only one explanatory variable x_i. In our application, the function looks like:

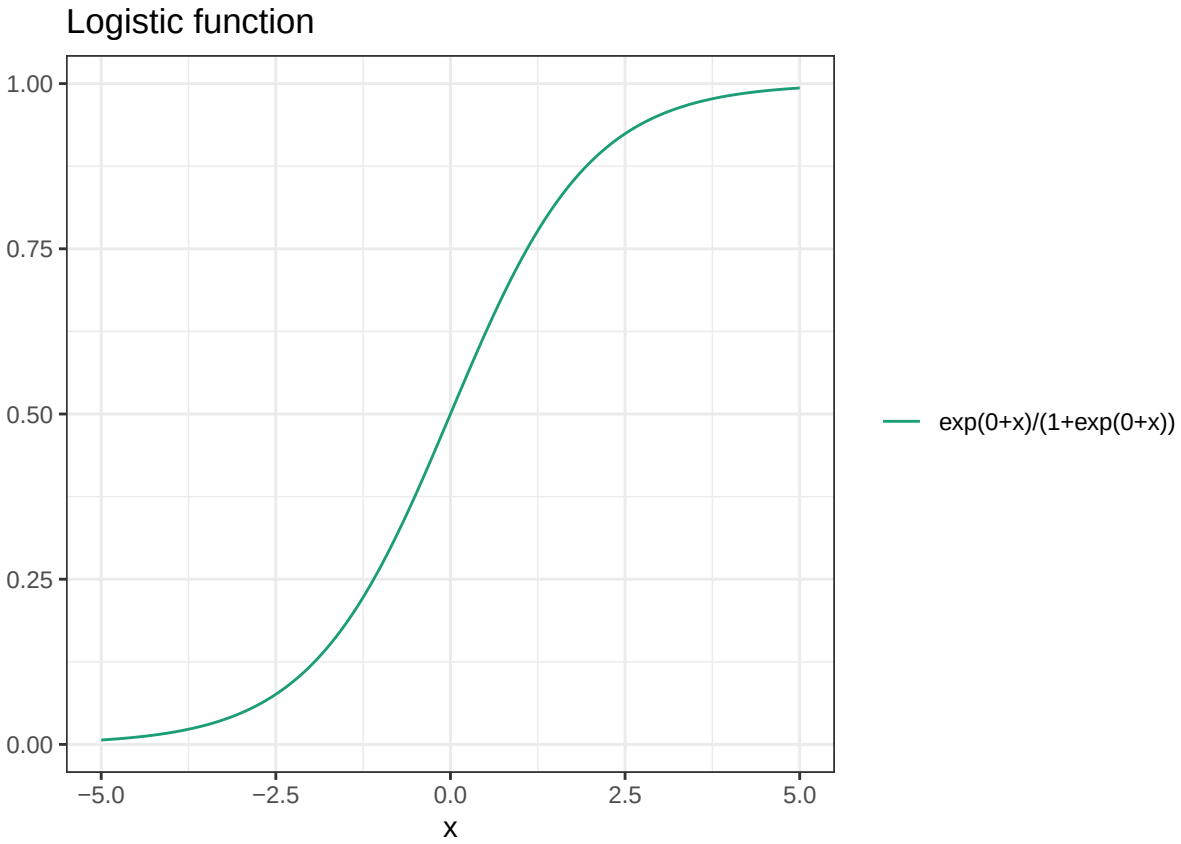
$$Pr[y_i = 1] = \frac{exp(\beta_1 + \beta_2 x_i)}{1 + exp(\beta_1 + \beta_2 x_i)}$$

$$Pr[y_i = 0] = 1 - \frac{exp(\beta_1 + \beta_2 x_i)}{1 + exp(\beta_1 + \beta_2 x_i)} = \frac{1}{1 + exp(\beta_1 + \beta_2 x_i)}$$

As you can see, the logistic function implies that the probability $\pi_i$ is a ratio. The numerator is the exponent of a linear combination of a constant and an explanatory variable $x_i$. The denominator is 1 plus the same exponent term. The $\beta_1$ parameter is called the intercept parameter, and the $\beta_2$ parameter describes the effect of $x_i$ on $y$.

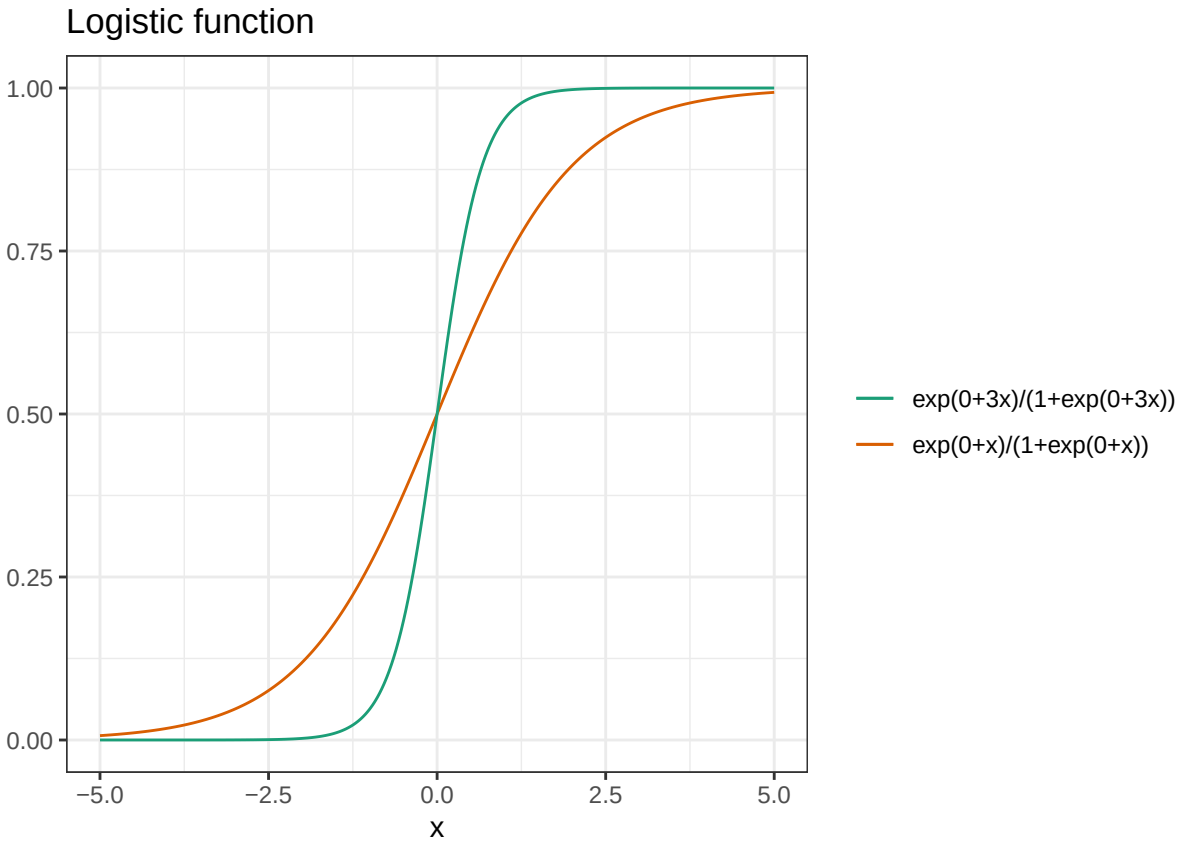
This graph shows the plot of the probability $y = 1$ as a function of the explanatory variable $x_i$. We consider the case where the Intercept parameter $\beta_1 = 0$, and the parameter $\beta_2 = 1$.

```
x <- seq(-5, 5, 0.01)
y_1 <- (exp(0+x))/(1+exp(0+x))
y_2 <- (exp(0+3*x))/(1+exp(0+3*x))
y_3 <- (exp(0-x))/(1+exp(0-x))
y_4 <- (exp(2+x))/(1+exp(2+x))
y_5 <- (exp(-2+x))/(1+exp(-2+x))
sample <- tibble(x,y_1,y_2,y_3,y_4,y_5)
plot1 <- ggplot(data=sample, aes(x=x)) +
  geom_line(aes(y=y_1, col = "exp(0+x)/(1+exp(0+x))")) +
  #geom_line(aes(y=y_2, col = "exp(0+3x)/(1+exp(0+3x))")) +
  #geom_line(aes(y=y_3, col = "exp(0-x)/(1+exp(0-x))")) +
  #geom_line(aes(y=y_4, col = "exp(2+x)/(1+exp(2+x))")) +
  #geom_line(aes(y=y_5, col = "exp(-2+x)/(1+exp(-2+x))")) +
  labs(x = "x", y = "",
  title = "Logistic function") +
  theme_bw() +
  scale_color_brewer(name= NULL, palette = "Dark2")
plot1
```



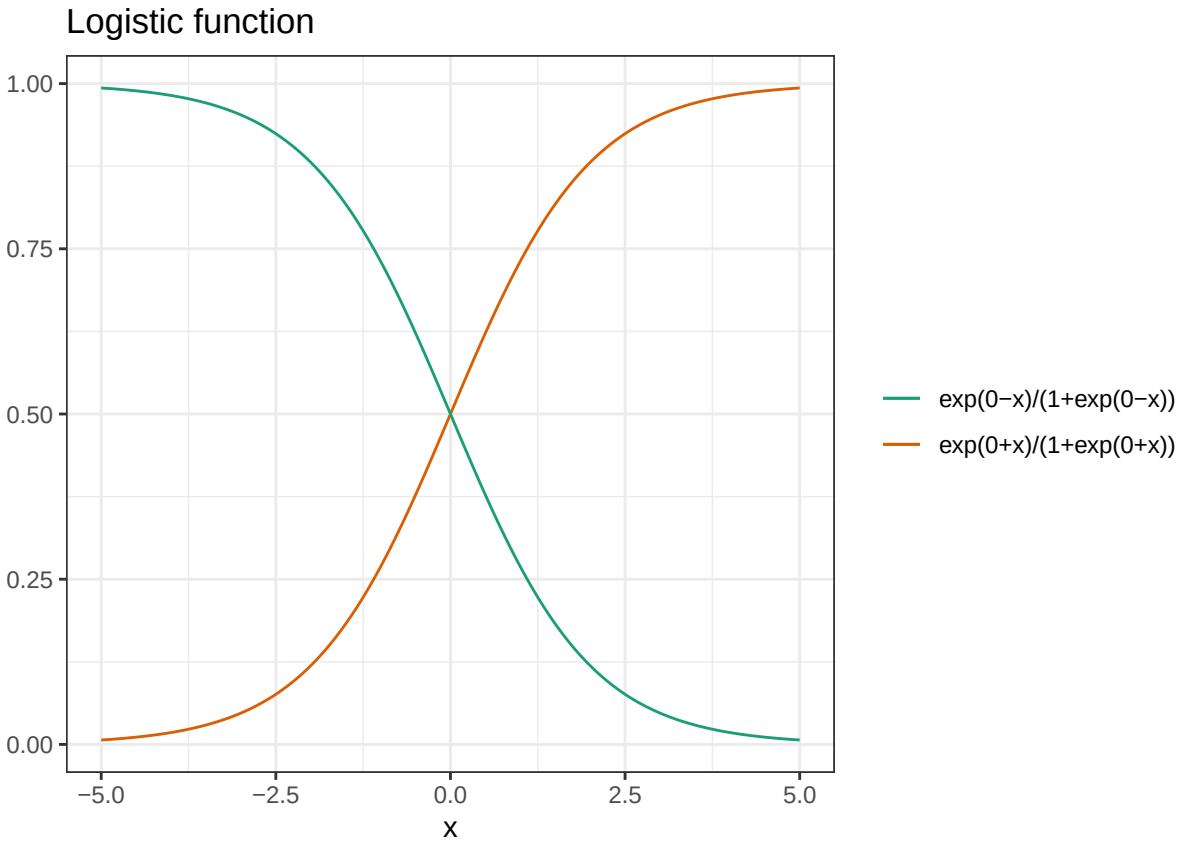

You can clearly see that the probability is bounded between 0 and 1. The probability that $Pr[y = 1] = 1/2$ when $x_i = 0$. In general, the probability increases when $x$ increases. The increase in probability is, however, not linear like in a linear regression model. For small values of x, the probability is really close to zero, and for large x the probability is nearly one.

To illustrate the role of the beta_2 parameter, we now change the size of this parameter. The new line shows the probability that $y = 1$ in case $\beta_2$ is three times as large. You can clearly see that the logistic function now is steeper:

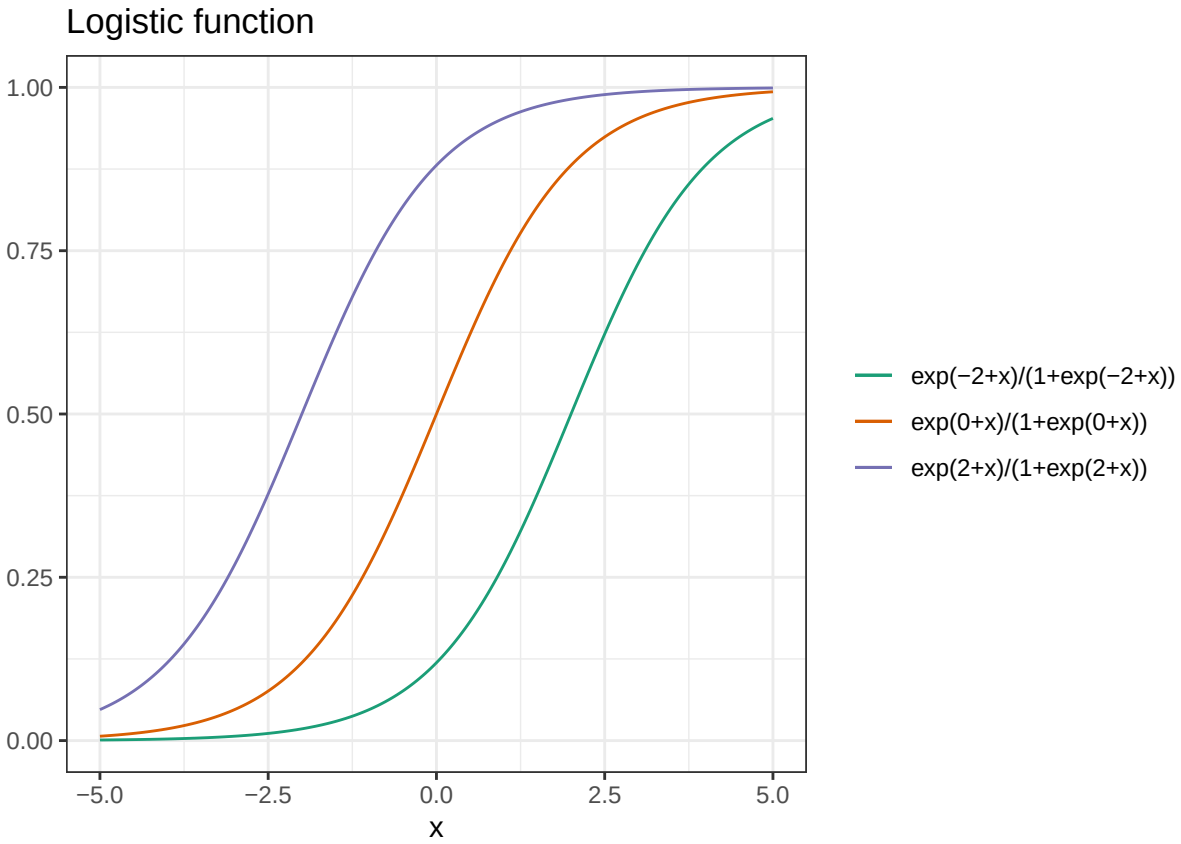



The interval where the probability is not close to the boundaries 0 and 1, is now smaller. Hence, for larger $\beta_2$ values, there is less uncertainty in whether y equals 0 or 1 given the value of $x$.

Now we've made the $\beta_2$ parameter negative. The new line shows a logit probability in case $\beta_2 = -1$. You can see that the logit probability is now a decreasing function of x. In fact, the probability plot is the mirror image of the original plot if you place a mirror vertically at x equal to 0.




Let us now change the value of the intercept parameter $\beta_1$. The new line shows the logit probability where $\beta_1 = 2$ instead of 0. You can see that the shape of the logit function stays the same, but that the location has changed. The graph has moved two units of x to the left. The logit probability is now $1/2$ at $x = -2$, while in the original case, this happens at x equals 0. If we set $\beta_1 = -2$the graph moves two units to the right and the shape of the curve stays the same.




As you have already seen in the previous graphs, the probability that $y = 1$ is a non-linear function of the explanatory variable $x$. This makes parameter interpretation a bit more difficult than in a linear regression.

Logit Model :

$$Pr[y_i = 1] = \frac{exp(\beta_1 + \beta_2 x_i)}{1 + exp(\beta_1 + \beta_2 x_i)}$$
$$Pr[y_i = 0] = \frac{1}{1 + exp(\beta_1 + \beta_2 x_i)}$$

**Odds ratio**

The easiest way to interpret the parameters of a logit model is to consider the **odds ratio**. That is, the ratio of the probability that $y = 1$ and the probability that $y = 0$. In the odds ratio the denominator of the logit probability cancels out, which results in a simple expression.

Odds ratio :

$$\frac{Pr[y_i = 1]}{Pr[y_i = 0]} = exp(\beta_1 + \beta_2 x_i)$$

In fact, it is even easier to consider the log odds ratio, as this is linear in x. It is easy to see that a positive $\beta_2$ implies that increase in x leads to an increase in the log odds ratio. This also implies an increase in the odds ratio itself. The opposite holds for negative $\beta_2$. An increase in x leads to a decrease in the log odds ratio and hence a decrease in the probability that $y = 1$.

Log Odds ratio :

$$log\left(\frac{Pr[y_i = 1]}{Pr[y_i = 0]}\right) = \beta_1 + \beta_2 x_i$$

The odds ratio provides insight into the direction, plus or minus, of the effect of changes in the x variable, but not in the size of the effect. To compute the size of the effect, we consider the marginal effect of a change in x on the probability.

**Marginal and average effect**

It can be shown that for the logit model, the first derivative of the probability that $y = 1$, with respect to $x$, can be written as a product of probability that $y = 1$ times the probability that $y = 0$ times $\beta_2$. Here we use the chain rule to derive the result as shown end of the section, no we just focus on the several conclusions that can be drawn from this result. First of all, as probabilities are always positive, the sign of $\beta_2$ determines direction of the marginal effect.

Marginal effect :

$$\frac{dPr[y_i = 1]}{dx_i} = Pr[y_i = 1] \cdot Pr[y_i = 0]\beta_2$$

This result was already clear from the odds ratio. A positive $\beta_2$ implies that an increase in x leads to an increase in the probability that $y = 1$. Secondly, when the probability that $y = 1$ is almost zero, the effect of a change in x is also almost zero. The same holds when the probability that $y = 0$ is almost zero.

Remember from the graphs of the logit function shown before that this happens for very large and very small values of x. Hence, a change in x has then almost no effect on the outcome of y, and the logit functions are flat for very large and small values of x. Know that this feature of the logit model is not present in a linear regression.

The value of the marginal effect also depends on the probability and hence on the value of x. This means that one cannot express the marginal effect in a single number. Computer packages often report the **average marginal effect in the sample**. You can obtain this quantity by computing the marginal effect for every value of $x_i$ in your sample and by taking the sample average.

Average marginal effect :

$$\frac{1}{n}\sum_{i=1}^{n}\frac{dPr[y_i = 1]}{dx_i} = \left(\frac{1}{n}\sum_{i=1}^{n} Pr[y_i = 1] \cdot Pr[y_i = 0]\right)\beta_2$$

**Multiple variables**

So far, we have only considered logit models with only one explanatory variable. The logit model can easily be extended with more explanatory variables. Here you see a logit model with $k - 1$ explanatory variables, which are denoted by $X_{j,i}$.

Logit with $x_{2i}, x_{3i}, ..., x_{ki}$

$$Pr[y_i = 1] = \frac{exp(\beta_1 + \sum_{j=2}^{k}\beta_j x_{ji})}{1 + exp(\beta_1 + \sum_{j=2}^{k}\beta_j x_{ji})}$$

The exponent term now contains a linear combination of the intercept and the $k-1$ explanatory variables and k beta parameters. The log odds ratio and marginal effect are similar as before. In fact, analyzing the effect of a change in one of the $x_{j,i}$ variables can be done the same way as in the single explanatory variable case, if we keep the value of all other $x$ variables fixed.

Log Odds ratio :

$$log\left(\frac{Pr[y_i=1]}{Pr[y_i=0]}\right) = \beta_1 + \sum_{j=2}^{k} \beta_j x_{ji}$$

Marginal effect :

$$\frac{\partial Pr[y_i=1]}{\partial x_{ji}} = Pr[y_i=1] \cdot Pr[y_i=0]\beta_j \forall j = 2,3,...k$$

The corresponding $\beta_j$ parameter determines direction and the relative size of the effect of a change in $x_{j,i}$. Note that the values of the marginal effect now also depend on the values of the other x variables through the probabilities.

**Notes on the logit distribution**

Suppose that an individual has two choices, denoted by 0 and 1. Let $u_i$ be an unobserved random variable that measures the difference in attractiveness (utility) between the choices 1 and 0. The attractiveness $u_i$ is a linear function of explanatory variables $x_{ji}$ with parameters $\beta_j$ and an error term, that is:

$$u_i = \beta_1 + \sum_{j=2}^{k} \beta_j x_{ji} + \eta_i$$

Where $\eta_i$ are IID random terms. Assume that individual $i$ chooses 1 when $u_i > 0$ and chooses 0 when $u_i < 0$. In other words, individual $i$ chooses the most attractive alternative. By translating this into choice probability we can derive the probability that $y_i = 1$ as follows:

$$Pr[y_i=1] = Pr[u_i > 0] = Pr[\beta_1 + \sum_{j=2}^{k} \beta_j x_{ji} + \eta_i \geq 0] = Pr[\eta_i \geq 0 - \beta_1 - \sum_{j=2}^{k} \beta_j x_{ji}]$$

$$Pr[y_i=1] = Pr[-\eta_i \leq \beta_1 + \sum_{j=2}^{k} \beta_j x_{ji}]$$

- Assume that the distribution of $\eta_i$ is symmetric around 0 and hence $\eta_i$ has the same distribution as $-\eta_i$.Furthermore, assume that $\eta_i$ has a standard logistic distribution with cumulative distribution function (CDF):

$$F(\eta_i) = \frac{exp(\eta_i)}{1 + exp(\eta_i)}$$

Under this condition, the $Pr[y_i = 1]$ is given by replacing $-\eta_i$ by $\eta_i$ in the CDF.

$$Pr[y_i=1] = Pr[-\eta_i \leq \beta_1 + \sum_{j=2}^{k} \beta_j x_{ji}] = \frac{exp(\beta_1 + \sum_{j=2}^{k} \beta_j x_{ji})}{1 + exp(\beta_1 + \sum_{j=2}^{k} \beta_j x_{ji})}$$

- From the previous result we can show that

$$\frac{\frac{\partial Pr[y_i=1]}{\partial x_{ji}}}{\frac{\partial Pr[y_i=1]}{\partial x_{li}}} = \frac{Pr[y_i=1]\cdot Pr[y_i=0]\beta_j}{Pr[y_i=1]\cdot Pr[y_i=0]\beta_l} = \frac{\beta_j}{\beta_l}$$

This implies that the relative size of the $\beta$ parameters is the same as the relative size of the marginal effects. If the $\beta_j$ is twice the size as the $\beta_l$ parameter, also the marginal effects.

- Use the chain rule to show that $\frac{\partial Pr[y_i=1]}{\partial x_{ji}} = Pr[y_i=1]\cdot Pr[y_i=0]\beta_j$:

Recall that

Logit with $x_{2i}, x_{3i}, ..., x_{ki}$

$$Pr[y_i=1] = \frac{exp(\beta_1 + \sum_{j=2}^{k}\beta_j x_{ji})}{1+exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})}$$

By expressing the denominator as an inverse exponent we can use the product rule:

$$\frac{\partial Pr[y_i=1]}{\partial x_{ji}} = \frac{\partial \frac{exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})}{1+exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})}}{\partial x_{ji}}$$

$$= \left[1+exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})\right]^{-1}\frac{\partial exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})}{\partial x_{ji}} + exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})\frac{\partial (1+exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})^{-1}}{\partial x_{ji}}$$

Let's solve by separate the two derviatives:

$$\frac{\partial exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})}{\partial x_{ji}} = exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})\beta_j$$

$$\frac{\partial (1+exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})^{-1}}{\partial x_{ji}} = \frac{-1}{(1+exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})^2}\frac{\partial exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji}}{\partial x_{ji}} = \frac{(-1)exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})\beta_j}{(1+exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})^2}$$

So we input these results in the original derivative:

$$\frac{\partial Pr[y_i=1]}{\partial x_{ji}} = \left[1+exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})\right]^{-1} exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})\beta_j + exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})\frac{(-1)exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})\beta_j}{(1+exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})^2}$$

Simplifying:

$$\frac{\partial Pr[y_i=1]}{\partial x_{ji}} = \frac{exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})\beta_j}{1+exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})} - \frac{exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})\beta_j}{(1+exp(\beta_1+\sum_{j=2}^{k}\beta_j x_{ji})^2}$$

We can rewrite this expression in terms of probabilities and the $\beta$ parameters as:

$$\frac{exp(\beta_1 + \sum_{j=2}^{k} \beta_j x_{ji})}{1 + exp(\beta_1 + \sum_{j=2}^{k} \beta_j x_{ji})}\beta_j - \frac{exp(\beta_1 + \sum_{j=2}^{k} \beta_j x_{ji}) exp(\beta_1 + \sum_{j=2}^{k} \beta_j x_{ji})}{(1 + exp(\beta_1 + \sum_{j=2}^{k} \beta_j x_{ji})^2}\beta_j = Pr[y_i = 1]\beta_j - Pr[y_i = 1]Pr[y_i = 1]\beta_j$$

We factorize $\beta_j$ and replace $1 - Pr[y_i = 1] = Pr[y_i = 0]$ by properties of probabilities.

$$\frac{\partial Pr[y_i = 1]}{\partial x_{ji}} = Pr[y_i = 1](1 - Pr[y_i = 1])\beta_j = Pr[y_i = 1]Pr[y_i = 0]\beta_j$$

$$\frac{\partial Pr[y_i = 1]}{\partial x_{ji}} = Pr[y_i = 1] \cdot Pr[y_i = 0]\beta_j$$

## Estimating binary choice

In practice, the values of the parameters are unknown, and need to be estimated from observed data. We show the logit formula for the probability that $y_i = 1$ now using vector notation, $X_i$ is a k-dimensional vector of the k explanatory variables, including the constant term, and $\beta$ is a k-dimensional vector of unknown parameters.

Logit Model in vector :

$$Pr[y_i = 1] = \frac{exp(x_i'\beta)}{1 + exp(x_i'\beta)}$$

with $x_i = \begin{pmatrix} 1 & x_{2i} & ... & x_{ki} \end{pmatrix}'$ and $\beta = \begin{pmatrix} \beta_1 & \beta_2 & ... & \beta_k \end{pmatrix}'$.

This model **cannot be written in a linear regression format** $y_i = x_i'\beta + \epsilon$, and hence, minimizing the sum of squared residuals is not the optimal way to estimate the beta parameters if the logit model is the true model. We, therefore, need to use another measure of fit for parameter estimation. And often use measure of fit, is the so-called **likelihood function**. The corresponding estimation technique is called **maximum likelihood**.

### The likelihood function

A maximum likelihood estimator, abbreviated as **MLE**, is defined as the parameter value that maximizes the probability of getting the actual observed data. In other words, the maximum likelihood estimate is the parameter value that gives you the highest probability of getting your observed data.

To construct the likelihood function necessary to estimate the beta parameters, we first consider the probability of getting an observation $y_i$. This is called the **likelihood contribution of observation i.**

Likelihood contribution for :

$$\text{observation } y_i = 1 : Pr[y_i = 1] = \frac{exp(x_i'\beta)}{1 + exp(x_i'\beta)}$$

$$\text{observation } y_i = 0 : Pr[y_i = 0] = \frac{1}{1 + exp(x_i'\beta)}$$

In short, we can write the probability of getting an observation y for observation i as a product of the probability of $y_i = 1$, raised to the power $y_i$, and the probability that $y_i = 0$ raised to the power $(1 - y_i)$. It is easy to check that the right probability is selected given the value of the observation.

Likelihood contribution for observation i

$$\left(\frac{exp(x_i'\beta)}{1+exp(x_i'\beta)}\right)^{y_1}\left(\frac{1}{1+exp(x_i'\beta)}\right)^{1-y_i}$$

As we want to estimate the parameters of the model based on our observations, we have to compute the joint probability of getting all y's. In practice, choices are often independent as in the individuals often make independent decisions. For independent random variables, the joint probability is the simply the product of the the individual probabilities. This implies that the total probability of getting the data, equals the product of all likely contributions. If we consider this product as a function of the model parameters, we have the likelihood function denoted by capital $L(\beta)$.

Likelihood function of n iid observations

$$L(\beta)=\prod_{i=1}^{n}\left(\frac{exp(x_i'\beta)}{1+exp(x_i'\beta)}\right)^{y_1}\left(\frac{1}{1+exp(x_i'\beta)}\right)^{1-y_i}$$

Maximizing this likelihood function with respect to beta provides the maximum likelihood estimator. This needs to be done using numerical methods. The likelihood function is a product of probabilities that are smaller than one, and hence, it can give very small values if the number of observations is large. This may cause numerical problems. In practice, one always chooses to maximize the log-likelihood function.

Remember the log properties:

- $log(ab)=log(a)+log(b)$
- $log(a^b)=b\cdot log(a)$
- $log(\frac{a}{b})=log(a)-log(b)$

Log likelihood function

$$log(L(\beta))=\sum_{i=1}^{n}(y_i)log\left(\frac{exp(x_i'\beta)}{1+exp(x_i'\beta)}\right)(1-y_i)log\left(\frac{1}{1+exp(x_i'\beta)}\right)$$

$$=\sum_{i=1}^{n}y_ix_i'\beta-log(1+exp(x_i'\beta))$$

The advantage of considering the log- likelihood function is that the product of probabilities becomes the sum of log probabilities, which causes no numerical problems.

**MLE**

Using the properties of the logarithmic function, you can easily write a log-likelihood function in the simple form shown on the slide. The maximum likelihood estimator is now defined as the value of beta that maximizes the log-likelihood function. We use **MLE as abbreviation for the maximum likelihood estimator.**

Notice that as the **log function is a monotonically increasing function, the value which maximizes the log-likelihood function, also maximizes the likelihood function**. Hence, it does not matter whether we maximize the likelihood function or the log-likelihood function.

To maximize the log-likelihood function we consider the first-order condition. For this we need the derivative of the log-likelihood function with respect to beta. This first derivative can be computed using the chain rule. You can see that the first-order conditions are non-linear in the parameter beta.

$$\text{FOC}: \frac{\partial log(L(\beta))}{\partial \beta} = \frac{\partial \sum_{i=1}^{n} y_i x_i' \beta - log(1 + exp(x_i'\beta))}{\partial \beta} = 0$$

$$\text{FOC}: \sum_{i=1}^{n} y_i x_i' - \frac{exp(x_i'\beta)x_i'}{1 + exp(x_i'\beta)} = 0$$

It turns out to be impossible to derive a nice formula for beta that solves this equation. In practice computer packages will do this for you **using numerical methods**. The solution is the maximum likelihood estimator b.

To shed some light on the value of the maximum likelihood estimator I consider the first-order condition for the intercept that is when x equals 1. It is easy to see that the first-order condition implies the following result:

FOC for intercept $x_i = 1$

$$\frac{1}{n} \sum_{i=1}^{n} \frac{exp(x_i'b)}{1 + exp(x_i'b)} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Interpretation of this formula is intuitively clear. When you evaluate the logit probabilities in the MLE and take the sample average, you obtain the same value as the average values of the y's. The latter is equal to the fraction of one observations in the sample. Hence, MLE matches the average values of the logit probabilities with the sample mean of the y's.

Suppose that you have a data set where all $y_i$ observations are 0 with different explanatory variables $x_i$ per observation. You want to estimate the parameters for the logit model. What is the value of the maximum likelihood estimator b in this case? As you can see on the equation, the first-order conditions imply that the sum of the probabilities should be 0.

$$\frac{1}{n} \sum_{i=1}^{n} \frac{exp(x_i'b)}{1 + exp(x_i'b)} = \frac{1}{n} \sum_{i=1}^{n} y_i = 0$$

As the logit function is always strictly larger than zero, there is no solution to the first-order condition and hence, the maximum likelihood estimator does not exist.

**Some properties of MLE**

The maximum likelihood estimator has some nice statistical properties, provided of course that the model is correctly specified:

1. The estimator is **consistent**. Which means that for a large number of observations, the estimator is close to the true parameter value.
2. MLE is **asymptotically efficient**, in the sense that the estimator has minimum variance.
3. MLE is **asymptotically normally distributed**. This is an important result that we can use for testing hypotheses. For practical purposes, you can use that the MLE estimator has approximately the normal distribution with mean the true beta and the covariance matrix V.

$$b_{MLE} \sim N(\beta, V)$$

For the logit model, this matrix V can be estimated as follows. It is beyond the scope of this lecture to derive the covariance matrix V (See Building Blocks. You can, however, see that the matrix shows some similarities with the covariance matrix of the least squares estimator, which was $Var(b_{OLS}) = \sigma^2(X'X)^{-1}$

$$Var(b_{MLE}) = \hat{V} = \left( \sum_{i=1}^{n} \left( \frac{exp(x_i'b)}{1 + exp(x_i'b)} \right) \left( \frac{1}{1 + exp(x_i'b)} \right) x_i x_i' \right)^{-1}$$

The first part of the expression is the product of the probabilities times 1 minus the probabilities. This is simply the variance of our Bernoulli distribution. $p(1-p) = pq$ Note that this variance is different across observations. The second part of V contains X prime X but now in vector notation. The matrix V can be used to compute standard errors of b. In practice, computer packages will do this for you.

**Logit hypothesis testing**

The above properties can be used to select appropriate explanatory variable to include in the logit model. For a single parameter restriction, you can follow a similar approach as in the linear regression case.

You can construct the familiar t-statistics to test for the significance of a single parameter $b_j$. This t-test statistic is approximately normal distributed and you reject the null hypothesis when its value is larger than the critical value.

We want to compare: - logit model without parameter restrictions - logit model with a single $\beta_j = 0$

Hypothesis

$$H_0 : \beta_j = 0 \text{ versus } H_1 : \beta_j \neq 0$$

Test statistic

$$z_j = \frac{b_j - 0}{SE(b_j)} \sim N(0,1)$$

If you want to test for the significance of a set of beta parameters, **you cannot use the F-test like in a linear regression model**, as the model cannot be written in a regression format with errors. Instead, you have to use a testing procedure called the **Likelihood Ratio Test.**

Suppose you have two logit models. The first model contains all variables, and the second model is the same as the first one, but now has some restrictions (m) on the parameters beta. Estimate the parameters of both models using maximum likelihood. Denote the maximum likelihood value of the models with all variables by $L(b_1)$, and the maximum likely value of the model with restrictions imposed by $L(b_0)$. These values are obtained by evaluating the likelihood function in the MLEs.

We want to compare: - logit model without parameter restrictions and estimates $b_1$ - logit model with $m$ parameter restrictions and estimates $b_0$

Hypothesis

$$H_0 : \text{m restrictions are correct versus } H_1 : H_0 false$$

To compute the test statistic we need: - $L(b_1)$: maximum likelihood value in full mode - $L(b_0)$: maximum likelihood value in restricted model

The likelihood ratio statistic can now be computed as minus two times the difference between the latter and the former maximum log-likelihood value. This statistic has approximately a Chi-square distribution, where the degrees of freedom is equal to the number of parameter restrictions (m).

Test statistic

$$LR = -2(log(L(b_0)) - log(L(b_1))) \approx \chi^2(m)$$

In case the value of the statistic is larger than the critical value, you reject the restrictions. When you reject, it means that the difference in log-likelihood value between the restricted model and the unrestricted model is statistically too large.

**Logit with only intercept.**

Consider a logit model with only an intercept parameter such that for all $i = 1, .., n$

$$Pr[y_i = 1] = \frac{exp(\beta_1)}{1 + exp(\beta_1)}$$

To estimate the parameter $\beta_1$ we use the maximum likelihood method. The FOC can be written as follows, where $b_1$ is the MLE:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{exp(b_1)}{1 + exp(b_1)} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- Use the first-order condition given above to show that the maximum likelihood estimator of $\beta_1$ equals $b_1 = log\left(\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n}(1-y_i)}\right)$

As the logit probability is constant over i, the FOC can be written as:

$$\frac{exp(b_1)}{1 + exp(b_1)} = \frac{1}{n}\sum_{i=1}^{n} y_i \Rightarrow exp(b_1) = \frac{1}{n}\sum_{i=1}^{n} y_i + exp(b_1)\frac{1}{n}\sum_{i=1}^{n} y_i$$

$$exp(b_1) = \frac{1}{n}\sum_{i=1}^{n} y_i + exp(b_1)\frac{1}{n}\sum_{i=1}^{n} y_i \Rightarrow (exp(b_1))(1 - \frac{1}{n}\sum_{i=1}^{n} y_i) = \frac{1}{n}\sum_{i=1}^{n} y_i$$

We use the fact that $\frac{1}{n}\sum_{i=1}^{n} 1 = 1$

$$exp(b_1) = \frac{\sum_{i=1}^{n} y_i}{1 - \frac{1}{n}\sum_{i=1}^{n} y_i} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n}(1 - y_i)}$$

$$b_1 = log\left(\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n}(1 - y_i)}\right)$$

- Show that the maximum likelihood estimator b1 implies that $\hat{Pr}[y_i = 1] = \frac{\sum_{i=1}^{n} y_i}{n}$

The fitted logit probability is given by $\hat{Pr}[y_i = 1] = \frac{exp(b_1)}{1+exp(b_1)}$, now we can substitute the MLE of $b_1$.

$$\hat{Pr}[y_i = 1] = \frac{exp(b_1)}{1 + exp(b_1)} = \frac{\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n}(1-y_i)}}{1 + \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n}(1-y_i)}}$$

Multiplying the numerator and denominator by $\sum_{i=1}^{n}(1 - y_i)$ we get.

$$\hat{Pr}[y_i = 1] = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n}(1 - y_i) + \sum_{i=1}^{n} y_i} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} 1} = \frac{\sum_{i=1}^{n} y_i}{n}$$

- We use the general formula of V to estimate the covariance matrix of the MLE in a logit model that only contains an intercept.

The covariance matrix in general is: $\hat{V} = \left(\sum_{i=1}^{n} \left(\frac{exp(x_i'b)}{1+exp(x_i'b)}\right)\left(\frac{1}{1+exp(x_i'b)}\right) x_i x_i'\right)^{-1}$.

So we replace the probabilities with the previous results:

$$\hat{V} = \left(\sum_{i=1}^{n} \left(\frac{\sum_{i=1}^{n} y_i}{n}\right)\left(1 - \frac{\sum_{i=1}^{n} y_i}{n}\right) 11'\right)^{-1}$$

$$\hat{V} = \left(\frac{\sum_{i=1}^{n} y_i}{n}\left(1 - \frac{\sum_{i=1}^{n} y_i}{n}\right)\sum_{i=1}^{n} 11'\right)^{-1} = \left(\frac{\sum_{i=1}^{n} y_i}{n}\left(1 - \frac{\sum_{i=1}^{n} y_i}{n}\right) n\right)^{-1}$$