PE I: Univariate Regression

MLE, Confidence Intervals and Hypothesis Testing (Chapters 3.4, 3.5 & 3.6)

Andrius Buteikis, andrius.buteikis@mif.vu.lt
http://web.vu.lt/mif/a.buteikis/

OLS: Assumptions

(UR.1) The Data Generating Process (**DGP**), or in other words, the population, is described by a linear (*in terms of the coefficients*) model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \ \forall i = 1, ..., N \qquad \text{(UR.1)}$$

(UR.2) The error term ϵ has an expected value of zero, given any value of the explanatory variable:

$$\mathbb{E}(\epsilon_i | X_j) = 0, \ \forall i, j = 1, ..., N \qquad \text{(UR.2)}$$

(UR.3) The error term ϵ has the same variance given any value of the explanatory variable (i.e. homoskedasticity) and the error terms are not correlated across observations (i.e. no autocorrelation):

$$\mathbb{V}\text{ar}\left(\boldsymbol{\varepsilon} | \mathbf{X}\right) = \sigma_\epsilon^2 \mathbf{I} \qquad \text{(UR.3)}$$

i.e. $\mathbb{C}\text{ov}(\epsilon_i, \epsilon_j) = 0, i \neq j$ and $\mathbb{V}\text{ar}(\epsilon_i) = \sigma_\epsilon^2 = \sigma^2$.

(UR.4) (optional) The residuals are normal:

$$\boldsymbol{\varepsilon} | \mathbf{X} \sim \mathcal{N}\left(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}\right) \qquad \text{(UR.4)}$$

$\boldsymbol{\varepsilon} = (\epsilon_1, ..., \epsilon_N)^\top$, $\mathbf{X} = (X_1, ..., X_N)^\top$, and $\mathbf{Y} = (Y_1, ..., Y_N)^\top$.

OLS: The Estimator

The unknown parameters of the linear regression

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

can be estimated via **OLS**:

$$\widehat{\beta} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y} \qquad \text{(OLS)}$$

The term **Ordinary Least Squares (OLS)** comes from the fact that these estimates minimize the sum of squared residuals.

Gauss-Markov Theorem

Under the assumption that the conditions (UR.1) - (UR.3) hold true, the OLS estimator (OLS) is **BLUE** (**B**est **L**inear **U**nbiased **E**stimator) and **Consistent**.

OLS: Standard Errors

We can measure the uncertainty of $\widehat{\boldsymbol{\beta}}$ via its standard deviation. This is the *standard error* of our estimate of $\boldsymbol{\beta}$:

The square roots of the diagonal elements of the variance-covariance matrix

$$\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} \widehat{\mathbb{V}\mathrm{ar}}(\widehat{\beta}_0) & \widehat{\mathbb{C}\mathrm{ov}}(\widehat{\beta}_0, \widehat{\beta}_1) \\ \widehat{\mathbb{C}\mathrm{ov}}(\widehat{\beta}_1, \widehat{\beta}_0) & \widehat{\mathbb{V}\mathrm{ar}}(\widehat{\beta}_1) \end{bmatrix} = \widehat{\sigma}^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1},$$

$$\text{where } \widehat{\sigma}^2 = \frac{1}{N-2} \widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}},$$

are called **the standard errors (se)** of the corresponding OLS estimators $\widehat{\boldsymbol{\beta}}$, which we use to **estimate** the standard **deviation** of $\widehat{\beta}_i$ from β_i

$$\mathrm{se}(\widehat{\beta}_i) = \sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\beta}_\mathbf{i})}$$

The standard errors describe the accuracy of an estimator (the smaller the better).

Effects of Changing the Measurement Units

- If Y is multiplied by a constant c:

$$\widetilde{Y} = c \cdot Y = c \cdot (\beta_0 + \beta_1 X + \epsilon) = (c \cdot \beta_0) + (c \cdot \beta_1) X + (c \cdot \epsilon)$$

- If X is multiplied by a constant c:

$$Y = \beta_0 + \beta_1 X + \epsilon = \beta_0 + \left(\frac{\beta_1}{c}\right)(c \cdot X) + \epsilon$$

- If we scale both X and Y by the same constant:

$$\begin{aligned}\widetilde{Y} = c \cdot Y &= c \cdot \left(\beta_0 + \left(\frac{\beta_1}{c}\right)(c \cdot X) + \epsilon\right) \\ &= (c \cdot \beta_0) + \beta_1 (c \cdot X) + (c \cdot \epsilon)\end{aligned}$$

- If we scale Y by one constant and X by a different constant:

$$\begin{aligned}\widetilde{Y} = a \cdot Y &= a \cdot \left(\beta_0 + \left(\frac{\beta_1}{c}\right)(c \cdot X) + \epsilon\right) \\ &= (a \cdot \beta_0) + \left(\frac{a}{c} \cdot \beta_1\right)(c \cdot X) + (a \cdot \epsilon)\end{aligned}$$

Interpretation of the Parameters

In a **level-level** model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

β_1 shows the **amount** by which the **expected** value of Y (remember that $\mathbb{E}(Y_i|X_i) = \beta_0 + \beta_1 X_i$) **changes** (either *increases*, or *decreases*), when X **increases** by **1 unit**.
As mentioned previously, interpreting the intercept β_0 is tricky.

For example, if X is in *thousands* of dollars, then β_1 shows the amount that the expected value of Y changes, when X increases by *one thousand*.

The defining feature of a univariate linear regression is that the *change in (the expected value of) Y* is equal to the *change in X multiplied by β_1*. So, the **marginal effect** of X on Y is **constant** and equal to β_1:

$$\Delta Y = \beta_1 \Delta X$$

or alternatively:

$$\beta_1 = \frac{\Delta Y}{\Delta X} = \frac{\Delta \mathbb{E}(Y|X)}{\Delta X} = \frac{d\mathbb{E}(Y|X)}{dX} =: \text{slope}$$

because a one-unit change in X results in *the same* change in Y, **regardless of the initial value of X**.

The **elasticity** of a variable Y with respect to X is defined as *the percentage change in Y corresponding to a 1% increase in X*:

$$\eta = \eta(Y|X) = \frac{\%\Delta Y}{\%\Delta X} = \frac{100 \cdot \dfrac{\Delta Y}{Y}}{100 \cdot \dfrac{\Delta X}{X}} = \frac{\Delta Y}{\Delta X} \cdot \frac{X}{Y}$$

So, the elasticity of the **expected** value of Y with respect to X is:

$$\eta = \frac{d\mathbb{E}(Y|X)}{dX} \cdot \frac{X}{\mathbb{E}(Y|X)} = \text{slope} \cdot \frac{X}{\mathbb{E}(Y|X)}$$

In practice in a linear model the elasticity is different on each point (X_i, Y_i), $i = 1, ..., N$. Most commonly, elasticity *estimated* by substituting the sample means of X and Y, with the interpretation being that a 1% increase in X will yield, on average, a $\widehat{\eta}$ **percentage** (i.e. $\%\widehat{\eta}$) increase/decrease in Y.

Often times economic variables are not always related by a straight-line relationship. In a simple linear regression the marginal effect of X on Y is *constant*, though this is not realistic in many economic relationships.

Nonlinearities in a Linear Regression

If we have a linear regression with transformed variables:

$$f(Y_i) = \beta_0 + \beta_1 \cdot g(X_i) + \epsilon_i, \quad i = 1, ..., N$$

then we can rewrite it in a matrix notation:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y} = [f(Y_1), ..., f(Y_N)]^\top$, $\boldsymbol{\varepsilon} = [\epsilon_1, ..., \epsilon_N]^\top$, $\boldsymbol{\beta} = [\beta_0, \beta_1]^\top$ and $\mathbf{X} = \begin{bmatrix} 1 & g(X_1) \\ 1 & g(X_2) \\ \vdots & \vdots \\ 1 & g(X_N) \end{bmatrix}$, where $f(Y)$ and $g(X)$ are some kind of transformations of the initial values of Y and X.
This allows us to estimate the unknown parameters via OLS:

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y}$$

Various transformations can be used to account for a nonlinear relationship between the variables $\mathbf{Y}$ and $\mathbf{X}$ (but still expressed as a linear regression in terms of parameters $\boldsymbol{\beta}$).

Quadratic Regression Model

The **quadratic (regression) model**:

$$Y = \beta_0 + \beta_1 X^2 + \epsilon$$

is a parabola, where β_0 is the intercept and β_1 is the *shape* parameter of the curve: if $\beta_1 > 0$, then the curve is $\cup -$ *shaped* (**convex**); if $\beta_1 < 0$, then the curve is $\cap -$ *shaped* (**concave**).

Because of the nonlinearity in X, our unit change in X effect on Y now depends on the initial value of X.

The **slope** of the quadratic regression is:

$$\text{slope} = \frac{d\mathbb{E}(Y|X)}{dX} = 2\beta_1 X,$$

which **changes** as X changes. For large values of X, the slope will be larger, for $\beta_1 > 0$ (or smaller, if $\beta_1 < 0$), and would have a more pronounced change in Y for a unit increase in X, compared to smaller values of X. Note that, unlike the simple linear regression, in this case β_1 **is no longer the slope**.

Log-Linear Regression Model

In a **log-linear** model:

$$\log(Y) = \beta_0 + \beta_1 X + \epsilon, \quad Y > 0$$

a **unit increase** in X yields (approximately) a $(100 \cdot \beta_1)$ **percentage change** in Y.
Furthermore, we have that:

$$\text{slope} := \frac{d\mathbb{E}(Y|X)}{dX} = \beta_1 Y$$

the marginal effect increases for larger values of Y (i.e. we see the effects on Y and not $\log(Y)$).

Many economic variables - price, income, wage, etc. - have skewed distributions, and taking logarithms to regularize the data is a common practice.

If we wanted to change the units of measurement of Y in a log-linear model, then β_0 would change, but β_1 would remain unchanged, since:

$$\log(cY) = \log(c) + \log(Y) = [\log(c) + \beta_0] + \beta_1 X + \epsilon$$

Linear-Log Regression Model

In a **linear-log** model:

$$Y = \beta_0 + \beta_1 \log(X) + \epsilon, \quad X > 0$$

a 1% increase in X yields (approximately) a $\beta_1/100$ **unit change** in Y.
Furthermore, we have that:

$$\text{slope} := \frac{d\mathbb{E}(Y|X)}{dX} = \beta_1 \frac{1}{X}$$

So, for **larger** values of X, an increase in X results in a **decreasing effect on** Y, i.e. the marginal effect decreases for larger values of X.

If we wanted to change the units of measurement of X in a linear-log model, then β_0 would change, but β_1 would remain unchanged, since:

$$Y = \beta_0 + \beta_1 \log\left(\frac{c}{c}X\right) = [\beta_0 - \beta_1 \log(c)] + \beta_1 \log(cX)$$

Log-Log Regression Model

In a **log-log** model:

$$\log(Y) = \beta_0 + \beta_1 \log(X) + \epsilon, \quad X,\ Y > 0$$

a 1% increase in X yields (approximately) a β_1 **percentage change** in Y.

Furthermore, **the elasticity of the log-log model is constant**:

$$\eta = \frac{d\mathbb{E}(Y|X)}{dX} \cdot \frac{X}{Y} = \beta_1$$

We have mentioned that (UR.4) is an optional assumption, which simplifies some statistical properties. But we will show how it can be applied to carry out an estimation method, which is based on the join distribution of $Y_1, ..., Y_N$.

However, we first need to talk about the distribution of Y. In order to do that, we will first introduce a few distributions, which are frequently encountered in econometrics literature.

The Normal Distribution

The most widely used distribution in statistics and econometrics. A random normal variable X is a continuous variable that can take any value. Its **probability density function** is defined as:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad -\infty < x < \infty$$

where $\mathbb{E}(X) = \int_{-\infty}^{\infty} f(x)dx = \mu$, $\mathbb{V}\text{ar}(X) = \sigma^2$. We say that X has a normal distribution and write $X \sim \mathcal{N}(\mu, \sigma^2)$. The normal distribution is sometimes called the **Gaussian** distribution.

```r
x  <- 5
s2 <- 4
mu <- 1
tmp<- 1/sqrt(2 * pi * s2) *
        exp(- (x - mu)^2 / (2 * s2))
print(tmp)
```

```
## [1] 0.02699548
```

```r
print(dnorm(x,
      mean = mu, sd = sqrt(s2)))
```

```
## [1] 0.02699548
```

```python
import numpy as np
```

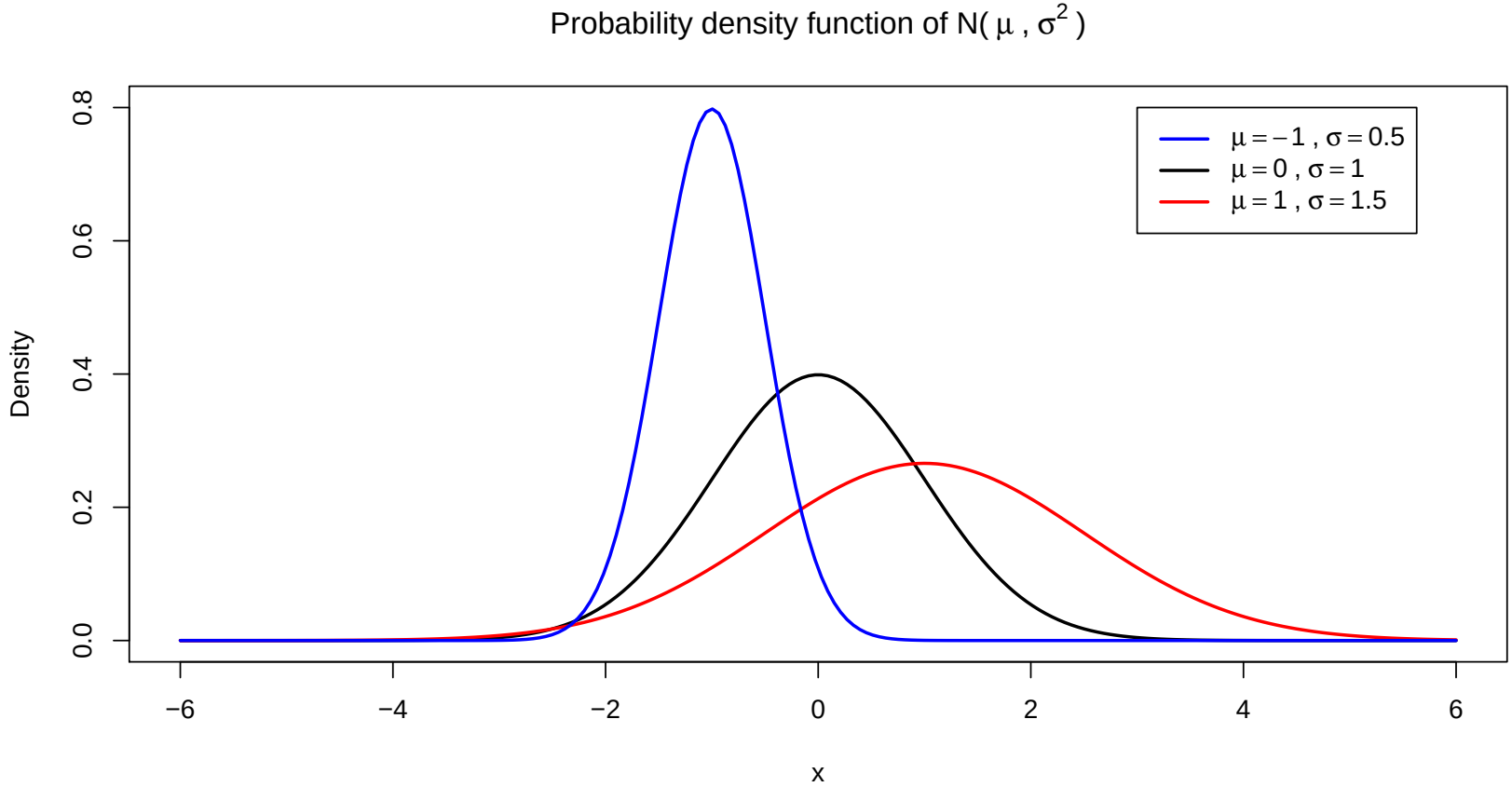
```python
x  = 5
s2 = 4
mu = 1
tmp= 1 / np.sqrt(2 * np.pi * s2) * \
  np.exp(- (x - mu)**2 / (2 * s2))
print("%.8f" % tmp)
```

```
## 0.02699548
```

```python
from scipy.stats import norm
```

```python
print("%.8f" % norm.pdf(x,
    loc = mu, scale = np.sqrt(s2)))
```

```
## 0.02699548
```



Certain random variables appear to *roughly* follow a normal distribution. These include: *a person's height, weight, test scores; country unemployment rate*.

On the other hand, other variables, like income do not appear to follow the normal distribution - the distribution is usually skewed towards the upper (i.e. right) tail. In some cases, a variable might be transformed to achieve normality.

The Standard Normal Distribution

A special case of normal distribution occurs when $\mu = 0$ and $\sigma^2 = 1$. If $Z \sim \mathcal{N}(0, 1)$ - we say that Z has a **standard normal distribution**. The pdf of Z is then:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right], \quad -\infty < z < \infty$$

The values of $\phi(\cdot)$ are easily tabulated and can be found in most (especially older) statistical textbooks as well as most statistical/econometrical software.

In most applications we start with a *normally distributed* random variable, but **any normal random variable can be turned into a standard normal**.

If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $a, b \in \mathbb{R}$, then:

- $\dfrac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$;
- $(aX + b) \sim \mathcal{N}(a\mu + b,\ a^2\sigma^2)$.

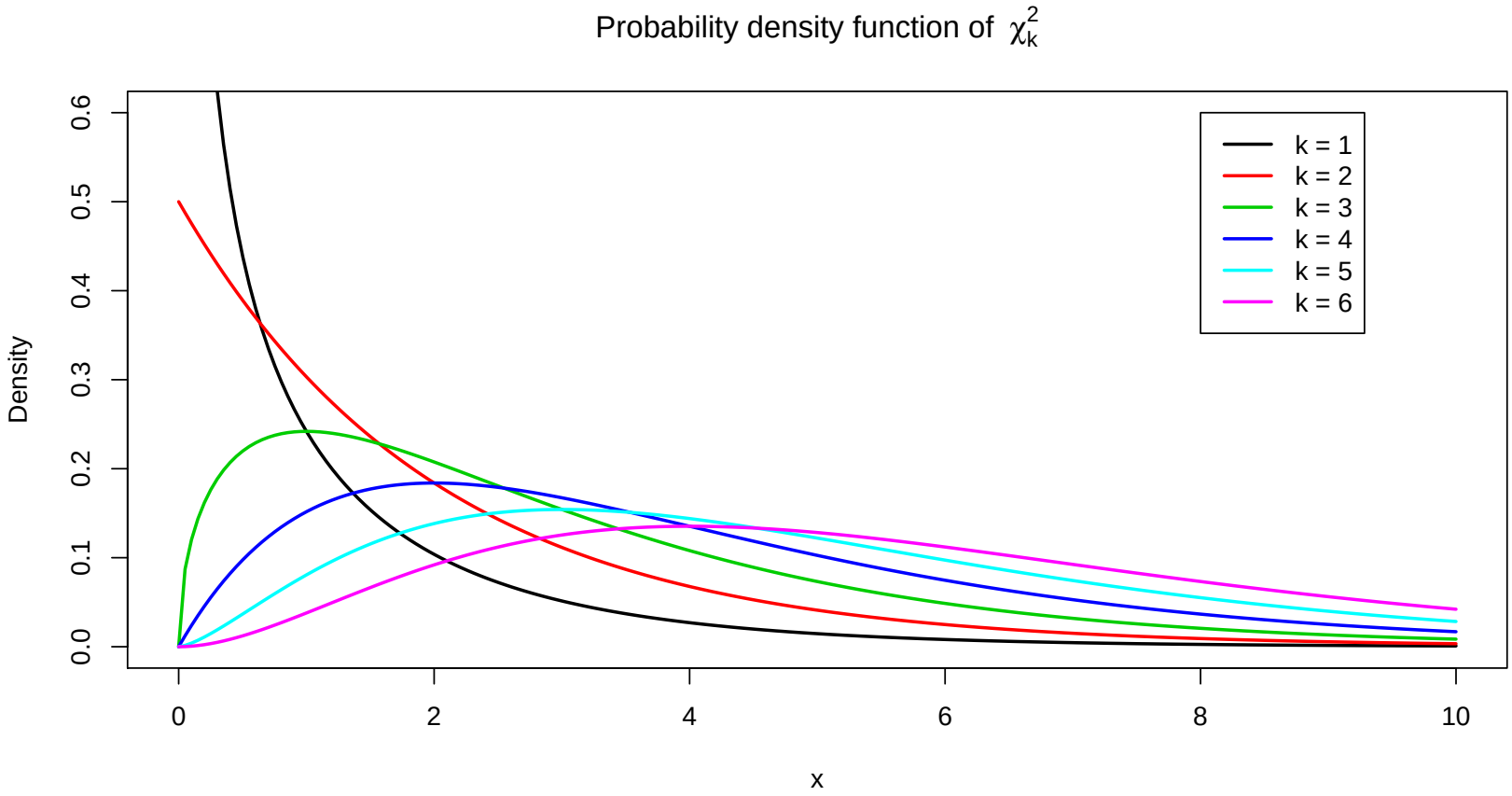
The Chi-Square Distribution

The chi-square distribution is obtained directly from independent, standard normal random variables. Let $Z_1, ..., Z_N$ be independent random variables and $Z_i \sim \mathcal{N}(0, 1)$, $\forall i = 1, ..., N$. Then a new random variable X, defined as:

$$X = \sum_{i=1}^{N} Z_i^2$$

Then X has a **chi-squared distribution** with N **degrees of freedom** (**df**) and write $X \sim \chi_N^2$. The **df** corresponds to the number of terms in the summation of Z_i. Furthermore, $\mathbb{E}(X) = N$ and $\mathbb{V}\mathrm{ar}(X) = 2N$.

Note: we calculate OLS estimates by minimizing $\sum_{i=1}^{N} \widehat{\epsilon}_i^2$, which appears similar, except, that we did not assume that the residual variance is unity.


Probability density function of χ^2_k
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
2
4
6
8
10
x
k = 1
k = 2
k = 3
k = 4
k = 5
k = 6

The t Distribution

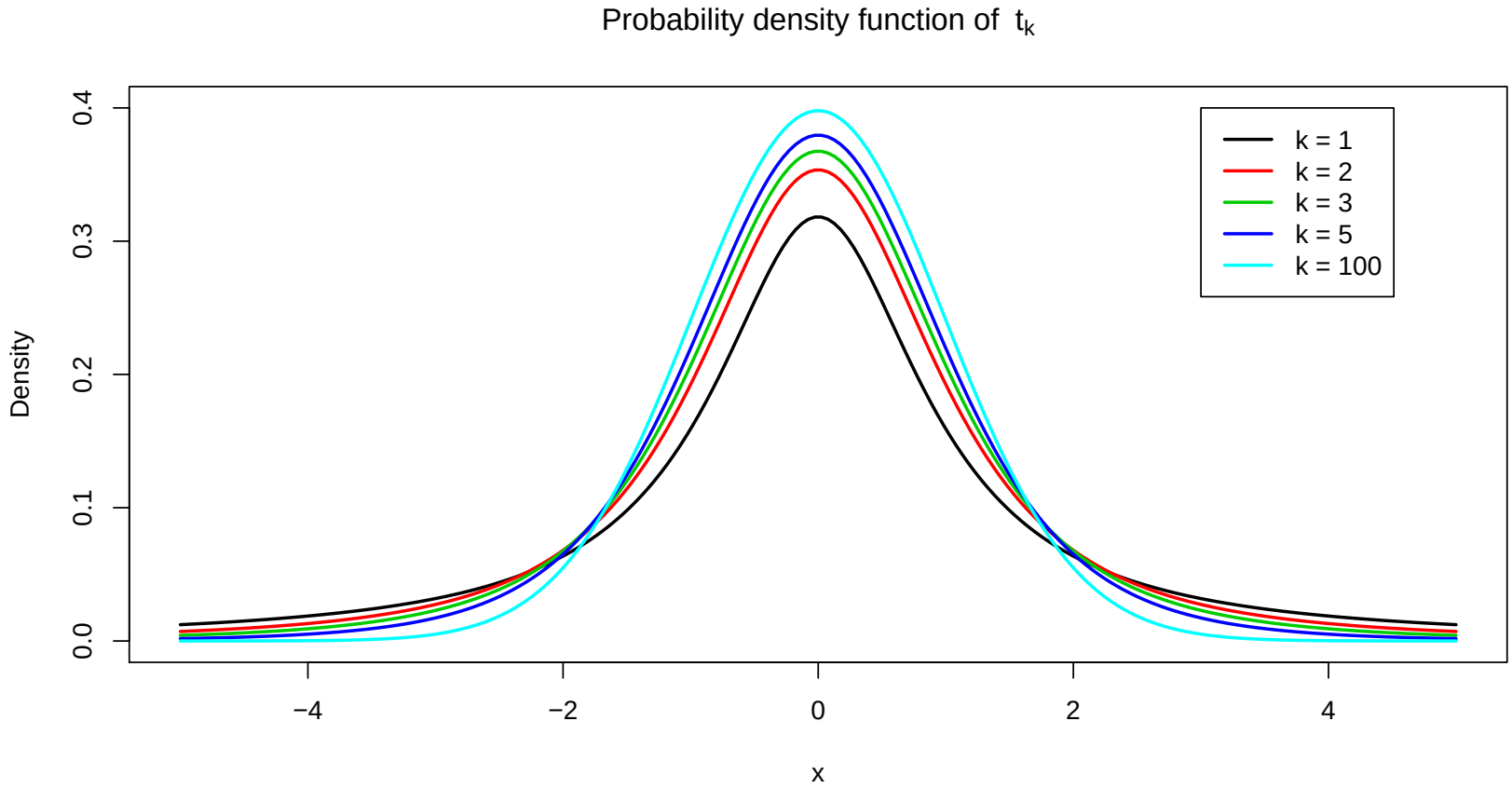
The t distribution is used in classical statistics and multiple regression analysis. We obtain a t distribution from a standard normal, and a chi-square random variable.

Let $Z \sim \mathcal{N}(0,1)$ and $X \sim \chi^2_N$, let Z and X be independent random variables. Then the random variable T, defined as:

$$T = \frac{Z}{\sqrt{X/N}}$$

has a **(Students)** t **distribution** with N degrees of freedom, which we denote as $T \sim \sqcup_{(N)}$. Furthermore, $\mathbb{E}(T) = 0$ and $\mathbb{V}\mathrm{ar}(T) = \frac{N}{N-2}$, $N > 2$.

As $N \to \infty$, the t distribution approaches the standard normal distribution.



If $X_1, ..., X_N$ is a random sample of independent random variables with $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $\forall i = 1, ..., N$, then the **t-ratio statistic** (or simply t-**statistic**) of an estimator of the sample mean $\overline{X}$, defined as:

$$t_{\overline{X}} = \frac{\overline{X} - \mu}{\text{se}\left(\overline{X}\right)} = \frac{\overline{X} - \mu}{\widehat{\sigma}^2/\sqrt{N}}$$

has the t_{N-1} distribution.

The F Distribution

Lastly, an important distribution for statistics and econometrics is the F distribution. It is usually **used for testing hypothesis in the context of multiple regression analysis**.

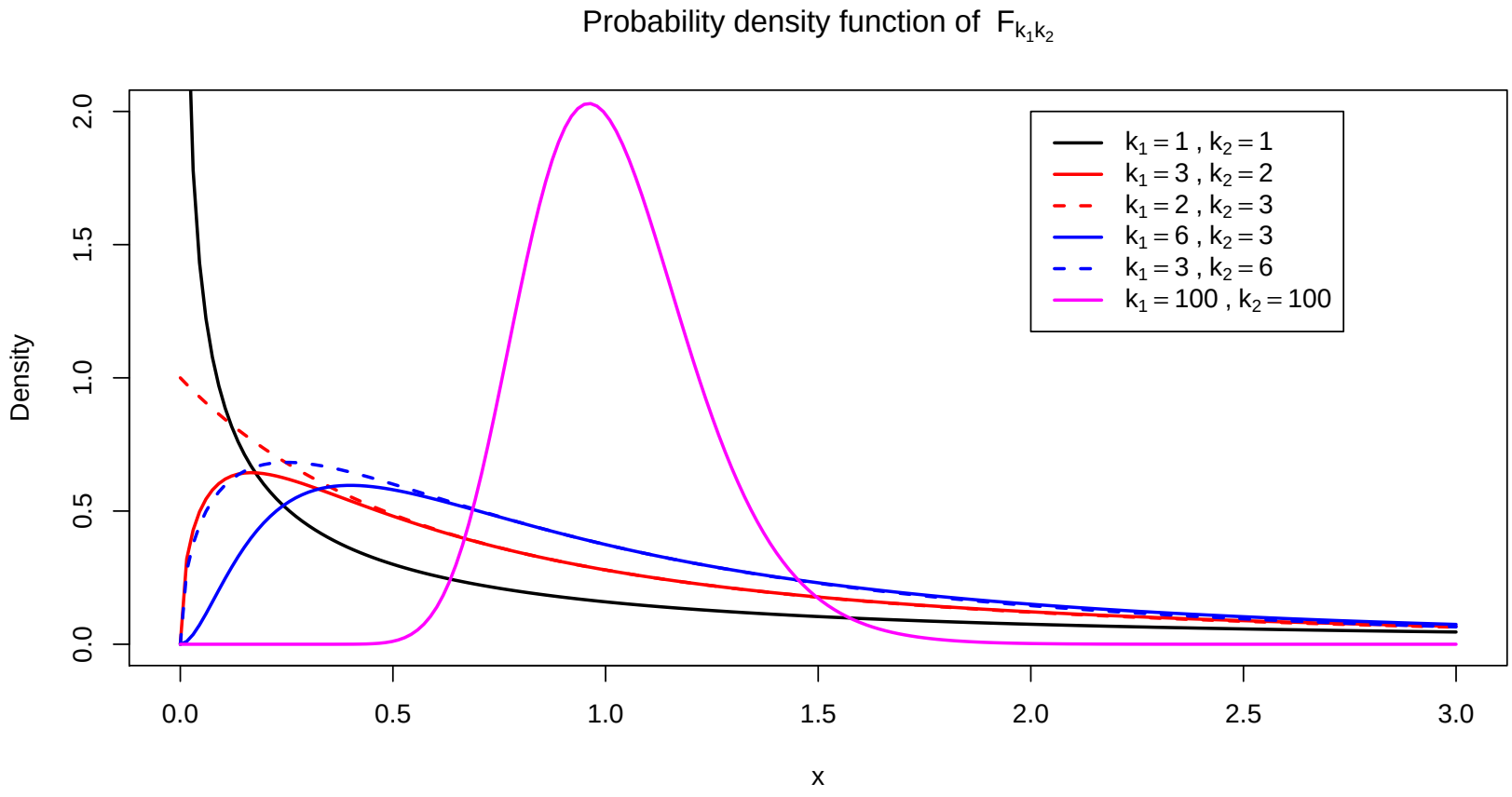
Let $X_1 \sim \chi^2_{k_1}$ and $X_2 \sim \chi^2_{k_2}$ be independent chi-squared random variables. The, the random variable:

$$F = \frac{X_1/k_1}{X_2/k_2}$$

has an **F distribution** with (k_1, k_2) degrees of freedom, which we denote by $F \sim F_{k_1,k_2}$.

The **order** of the degrees of freedom in F_{k_1,k_2} is important:

- k_1 - the numerator degrees of freedom - is associated with the chi-square variable in the numerator, X_1;
- k_2 - the denominator degrees of freedom - is associated with the chi-square variable in the denominator, X_2;



Now that we have introduced a few common distributions, we can look back at our univariate regression model, and examine its distribution more carefully. This also allows us to derive yet another model parameter estimation method, which is based on the assumptions on the underlying distribution of the data.

Univariate Linear Regression Model With Gaussian Noise

As before, let our linear equation be defined as:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

where:

- $\mathbf{X}$ can be either a non-random variable, or a random variable with some arbitrary distribution.
- $\mathbb{V}\mathrm{ar}\left(\varepsilon|\mathbf{X}\right) = \mathbb{V}\mathrm{ar}\left(\varepsilon\right) = \sigma_\epsilon^2\mathbf{I}$ - i.e. the error terms are **independent** of $\mathbf{X}$, **independent across observations** $i = 1, ..., N$ and have a constant variance.

As mentioned earlier, a consequence of (UR.4) assumption is that not only are the residuals ε normal, but the OLS estimators as well:

$$\widehat{\beta}|\mathbf{X} \sim \mathcal{N}\left(\beta, \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right)$$

The fact that the OLS estimators have a normal distribution can be shown by applying a combination of the Central Limit Theorem and Slutsky's Theorem.

Additionally, if (UR.4) holds true, then it can be shown that:

$$\mathbf{Y}|\mathbf{X} \sim \mathcal{N}\left(\mathbf{X}\beta,\ \sigma^2\mathbf{I}\right)$$

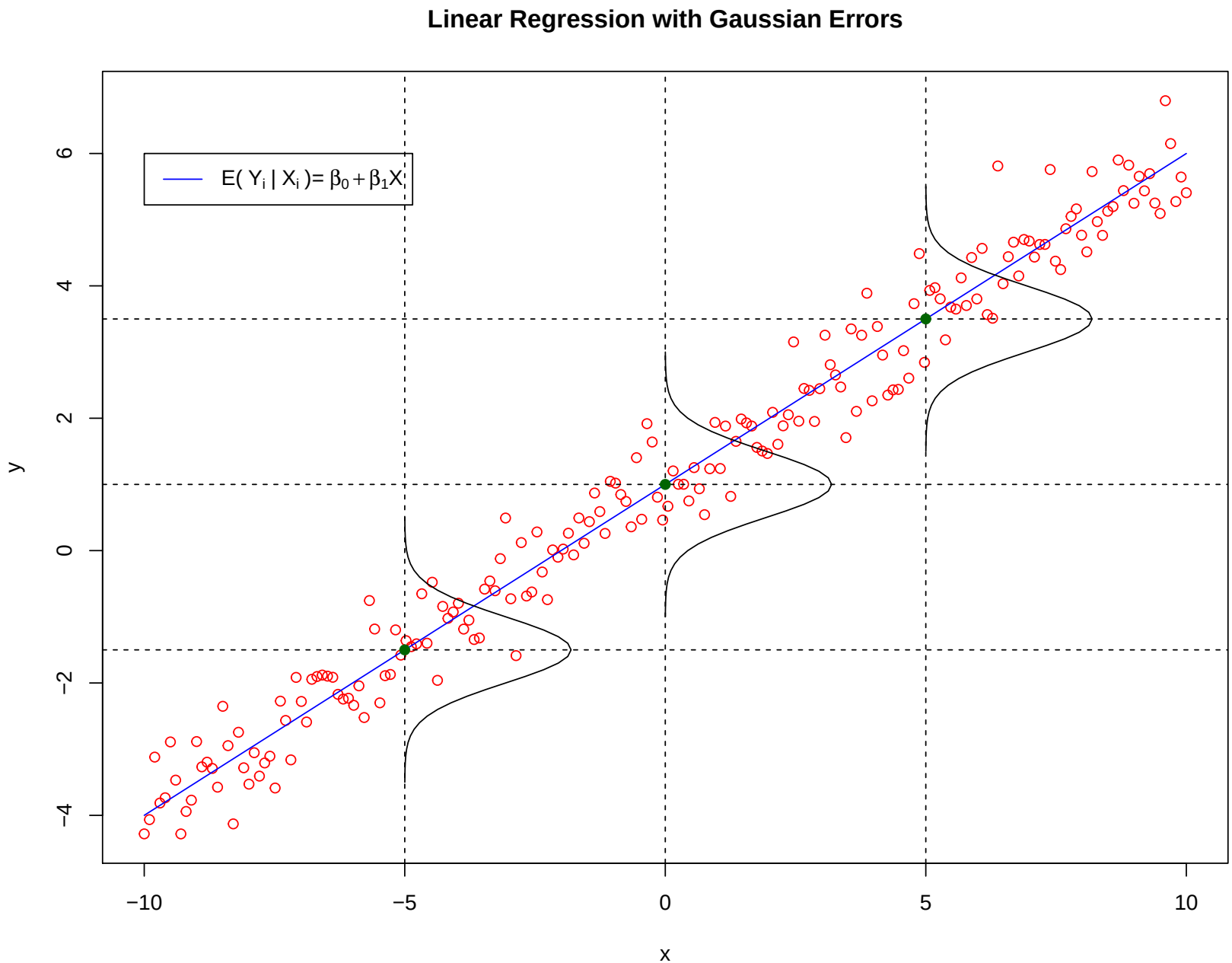
Since:

$$\mathbb{E}(\mathbf{Y}|\mathbf{X}) = \mathbb{E}\left(\mathbf{X}\beta + \varepsilon|\mathbf{X}\right) = \mathbb{E}\left(\mathbf{X}\beta|\mathbf{X}\right) = \mathbf{X}\beta$$

$$\mathbb{V}\mathrm{ar}\left(\mathbf{Y}|\mathbf{X}\right) = \mathbb{V}\mathrm{ar}\left(\mathbf{X}\beta + \varepsilon|\mathbf{X}\right) = \mathbb{V}\mathrm{ar}\left(\varepsilon|\mathbf{X}\right) = \sigma^2\mathbf{I}$$

Furthermore, we see that a consequence of these assumptions is that Y_i and Y_j are independent, given X_i and X_j, $i \neq j$.

We can also illustrate the distribution of **Y** graphically:



The assumption that the residual term is *normal* (or sometimes called **Gaussian**) does not always hold true in practice.
If we believe that the random noise term is a combination of a number of independent smaller random causes, all similar in magnitude, then the error term will indeed be normal (via Central Limit Theorem).

Nevertheless:

The Gaussian-noise assumption is important in that it gives us a conditional joint distribution of the random sample $\mathbf{Y}$, which in turn gives us the **sampling distribution** for the **OLS** estimators of β.
The distributions are important when we are doing **statistical inference on the parameters** - calculating confidence intervals or testing null hypothesis for the parameters.
Furthermore, this allows us to calculate the **confidence** intervals for $\mathbb{E}(\mathbf{Y}|\mathbf{X})$, or **prediction** intervals for $\widehat{\mathbf{Y}}$, given a value of $\mathbf{X}$.
Comparison of these intervals.

Consequently, we will see that the conditional probability density function (pdf) of **Y**, given **X** is a multivariate normal distribution. For Y_i, given X_i the pdf is the same for each $i = 1, ..., N$. Lets first look at the **cumulative distribution function** (cdf):

$$\begin{aligned} F_{Y|X}(y_i|x_i) &= \mathbb{P}(Y \leq y_i|X = x_i) \\ &= \mathbb{P}(\beta_0 + \beta_1 X + \epsilon \leq y_i|X = x_i) \\ &= \mathbb{P}(\beta_0 + \beta_1 x_i + \epsilon \leq y_i) \\ &= \mathbb{P}(\epsilon \leq y_i - \beta_0 - \beta_1 x_i) \\ &= F_{\epsilon|X}(y_i - \beta_0 - \beta_1 X|X = x_i) \end{aligned}$$

since $\epsilon \sim \mathcal{N}(0, \sigma^2)$, it follows that the conditional **pdf** of Y on X is the same across $i = 1, ..., N$:

$$f_{Y|X}(y_i|x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}}$$

Because Y_i are independent across observations, conditional on $\mathbf{X}$, the **joint pdf** is a product of the marginal pdf's:

$$f_{Y_1,...,Y_N|X_1,...,X_N}(y_1, ..., y_N|x_1, ..., x_N) = \prod_{i=1}^{N} f_{Y|X}(y_i|x_i)$$
$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}\right]$$

which we can re-write as a **multivariate normal distribution**:

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi)^{N/2}(\sigma^2)^{N/2}} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{x}\beta)^\top (\sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{x}\beta)\right]$$

Having defined the distribution of our random sample allows us to estimate the parameters in a different way - by using the probability density function.

While the probability density function relates to the **likelihood function** of the parameters of a statistical model, given some observed data:

$$\mathcal{L}(\beta, \sigma^2|\mathbf{y}, \mathbf{X}) = \frac{1}{(2\pi)^{N/2}(\sigma^2)^{N/2}} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\beta)^\top \left(\sigma^2\mathbf{I}\right)^{-1}(\mathbf{y} - \mathbf{X}\beta)\right]$$

The probability density function is a function of an outcome **y**, given fixed parameters, while the likelihood function is a function of the parameters only, with the data held as fixed.

Maximizing The Log-Likelihood Function - The MLE

In practice, it is much more convenient to work with the log-likelihood function:

$$\begin{aligned}\ell(\beta, \sigma^2|\mathbf{y}, \mathbf{X}) &= \log\left(\mathcal{L}(\beta, \sigma^2|\mathbf{y}, \mathbf{X})\right) \\ &= -\frac{N}{2}\log(2\pi) - N\log(\sigma) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta)\end{aligned}$$

Taking the partial derivatives allows us to fund the ML estimates:

$$\begin{aligned}\frac{\partial \ell}{\partial \beta^\top} &= -\frac{1}{2\sigma^2}\left(-2\mathbf{X}^\top\mathbf{y} + 2\mathbf{X}^\top\mathbf{X}\beta\right) = 0 \\ \frac{\partial \ell}{\partial \sigma^2} &= -\frac{N}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta) = 0\end{aligned}$$

which give us the ML estimators:

$$\begin{aligned}\widehat{\beta}_{\text{ML}} &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y} \\ \widehat{\sigma}^2 &= \frac{1}{N}\left(\mathbf{y} - \mathbf{X}\widehat{\beta}_{\text{ML}}\right)^\top\left(\mathbf{y} - \mathbf{X}\widehat{\beta}_{\text{ML}}\right)\end{aligned}$$

We see that these estimators exactly match the OLS estimators of β. This is a special property of (UR.4) (and (UR.3)) assumption. The estimator of σ^2 is divided by N (instead of $N-2$ in the OLS case).

Standard Errors of a MLE

The standard errors can be found by calculating the inverse of the square root of the diagonal elements of the observed **Fisher information matrix**.
In general, the **Fisher information matrix** $\mathbf{I}(\boldsymbol{\gamma})$ is a symmetrical $k \times k$ matrix (if the parameter vector is $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_k)^\top$), which contains the following entries:

$$(\mathbf{I}(\boldsymbol{\gamma}))_{i,j} = -\frac{\partial^2}{\partial\gamma_i \partial\gamma_j}\ell(\boldsymbol{\gamma}), \quad 1 \leq i, j \leq p$$

The **observed Fisher information matrix** is the information matrix evaluated at the MLE: $\mathbf{I}(\widehat{\boldsymbol{\gamma}}_{\text{ML}})$.
Most statistical software calculates and returns the **Hessian matrix**.
The **Hessian** is defined as $\mathbf{H}(\boldsymbol{\gamma})$:

$$(\mathbf{H}(\boldsymbol{\gamma}))_{i,j} = \frac{\partial^2}{\partial\gamma_i \partial\gamma_j}\ell(\boldsymbol{\gamma}), \quad 1 \leq i, j \leq p$$

i.e. it is a matrix of second derivatives of the likelihood function with respect to the parameters.

Often times we **minimize** the **negative log-likelihood function** (which is equivalent to *maximizing* the log-likelihood function), then $\mathbf{I}(\widehat{\gamma}_{\text{ML}}) = \mathbf{H}(\widehat{\gamma}_{\text{ML}})$. If we **maximize** the likelihood function, then $\mathbf{I}(\widehat{\gamma}_{\text{ML}}) = -\mathbf{H}(\widehat{\gamma}_{\text{ML}})$.

Furthermore:

$$\mathbb{V}\text{ar}(\gamma) = [\mathbf{I}(\widehat{\gamma}_{\text{ML}})]^{-1}$$

and **the standard errors are then the square roots of the diagonal elements of the covariance matrix.**

Generally, the asymptotic distribution for a maximum likelihood estimate is:

$$\widehat{\gamma}_{\text{ML}} \sim \mathcal{N}\left(\gamma, [\mathbf{I}(\widehat{\gamma}_{\text{ML}})]^{-1}\right)$$

When to use MLE instead of OLS

Assuming that (UR.1) - (UR.3) holds. If we additionally assume that that the property (UR.4) holds true, OLS and MLE estimates are equivalent. See the lecture notes for an example.

The main takeaway is that when we are using OLS to estimate the parameters by minimizing the sum of squared residuals, we do not make any assumptions about the underlying distribution of the errors.

If the errors are normal, then MLE is equivalent to OLS. However, if we have reason to believe that the errors are not normal, then specifying a **correct likelihood function** would yield the correct estimates using MLE.

Example: Poisson Regression

A Poisson regression is sometimes known as a **log-linear model**. It is used to model count data (i.e. integer-valued data):

- number of passengers in a plane;
- the number of calls in a call center;
- the number of insurance claims in an insurance firm, etc.

Poisson regression assumes the response variable Y has a poisson distribution and that the logarithm of its **expected value** is modelled by a **linear** combination of its expeected values. If we assume that our DGP follows a **Poisson regression**, then:

$$Y \sim Pois(\mu), \quad \Longrightarrow \mathbb{E}(Y) = \mathbb{V}\mathrm{ar}(Y) = \mu$$

and:

$$\log(\mu) = \beta_0 + \beta_1 X \iff \mu = \exp\left[\beta_0 + \beta_1 X\right]$$

This leads to the following model:

$$\mathbb{E}(Y|X) = \exp\left[\beta_0 + \beta_1 X\right] \iff \log\left(\mathbb{E}(Y|X)\right) = \beta_0 + \beta_1 X$$

Since $\mathbb{E}[\log(Y)|X] \neq \log(\mathbb{E}[Y|X])$, we cannot simply take logarithms of Y and apply OLS. However, we can apply MLE.

The likelihood function of Y is:

$$\mathcal{L}(\beta|\mathbf{y},\mathbf{X}) = \prod_{i=1}^{N} \frac{\exp\left(y_i \cdot (\beta_0 + \beta_1 x_i)\right) \cdot \exp\left(-\exp\left(\beta_0 + \beta_1 x_i\right)\right)}{y_i!}$$

and the log-likelihood:

$$\ell(\beta|\mathbf{y},\mathbf{X}) = \sum_{i=1}^{N} \left(y_i \cdot (\beta_0 + \beta_1 x_i) - \exp\left(\beta_0 + \beta_1 x_i\right) - \log(y_i!)\right)$$

We notice that the parameters β_0 and β_1 only appear in the first two terms. Since our goal is to estimate β_0 and β_1, we can drop $\sum_{i=1}^{N} \log(y_i!)$ from our equation. This does not impact the maximization - removing (or adding) a constant value from an **additive equation** will not impact the optimization.

On the other hand, the initial *likelihood function* $\mathcal{L}(\beta|\mathbf{y},\mathbf{X})$ is a **multiplicative equation** - all the different terms are multiplied across $i = 1, ..., N$.

This simplifies our expression to:

$$\ell(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} \left(y_i \cdot (\beta_0 + \beta_1 x_i) - \exp\left(\beta_0 + \beta_1 x_i\right)\right)$$

Unfortunately, calculating $\dfrac{\partial \ell(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})}{\partial \boldsymbol{\beta}}$ will not yield a closed-form solution.

Nevertheless, we can use the standard optimization functions to find the optimal parameter values.

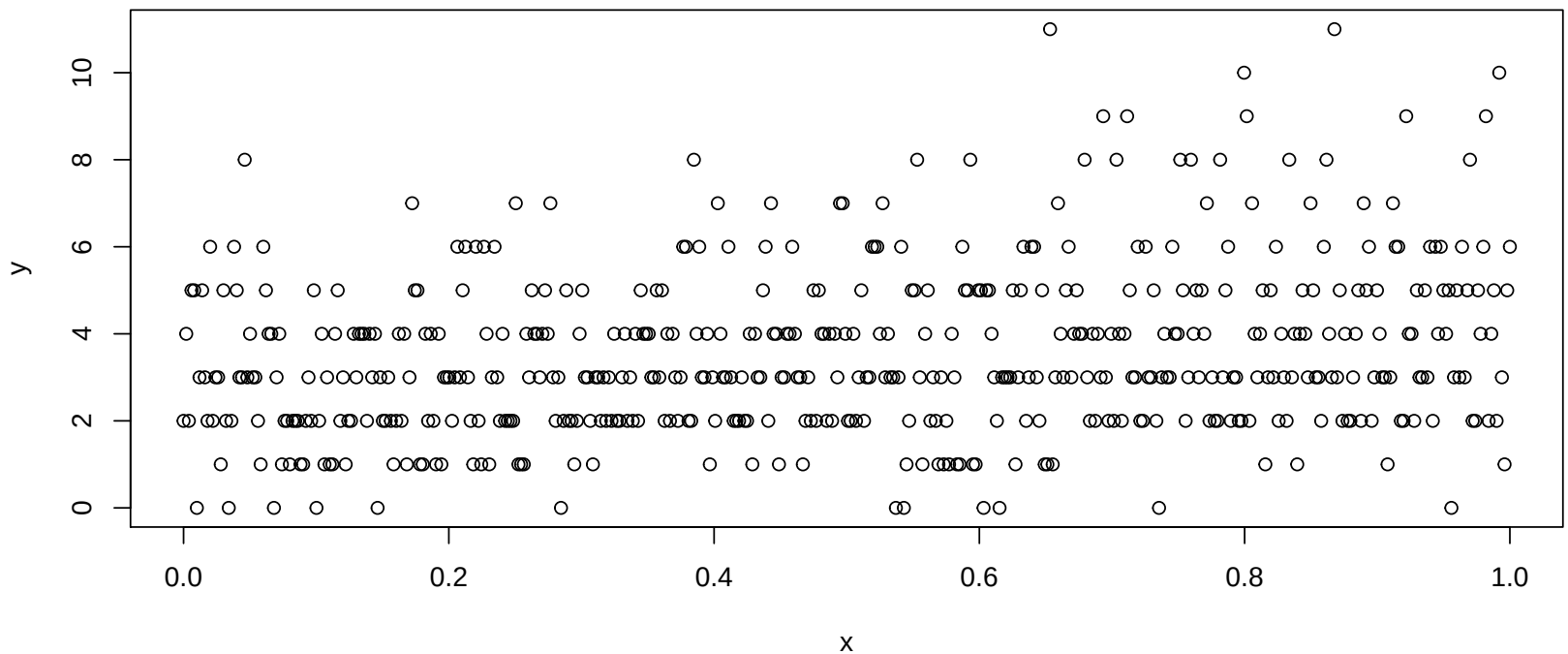
Example: Let:

- Y - the number of people who visited a cafe,
- X - the rate of advertising done by a cafe (from 0 to 1).

We will generate an example with $\beta_0 = 1$, $\beta_1 = 0.5$ and $N = 100$.

```r
set.seed(123)
#
N = 500
beta_0 <- 1
beta_1 <- 0.5
#
x  <- seq(from = 0, to = 1,
          length.out = N)
mu <- exp(beta_0 + beta_1 * x)
y  <- rpois(n = N,
            lambda = mu)
```

```python
np.random.seed(123)
#
N = 500
beta_0 = 1
beta_1 = 0.5
#
x = np.linspace(start = 0, stop = 1,
      num = N)
mu = np.exp(beta_0 + beta_1 * x)
y = np.random.poisson(lam = mu,
      size = N)
```



Next, we specify the log-likelihood functions:

```r
log_lik <- function(par_vec, y, x) {
  # The likelihood function values:
  lik <- dpois(y,
    lambda = exp(par_vec[1] + par_vec[2]
  # If all logarithms are zero,
  #       return a large value
  if(all(lik == 0)) return(1e8)
  # Logarithm of zero = -Inf
  return(-sum(log(lik[lik != 0])))
}
```

```python
from scipy.stats import poisson
```

```python
def log_lik(par_vec, y, x):
    # The likelihood function values:
    lik = poisson.pmf(y,
      mu = np.exp(par_vec[0] + par_vec[1
    # If all logarithms are zero,
    #    return a large value
    if all(v == 0 for v in lik):
        return(1e8)
    # Logarithm of zero = -Inf
    return(-sum(np.log(lik[np.nonzero(li
```

Now, we can optimize the function and estimate the parameters:

```
coef_est <- optim(
      par = c(0, 0),
      fn = log_lik, hessian = T,
      y = y, x = x)
print(coef_est)
```

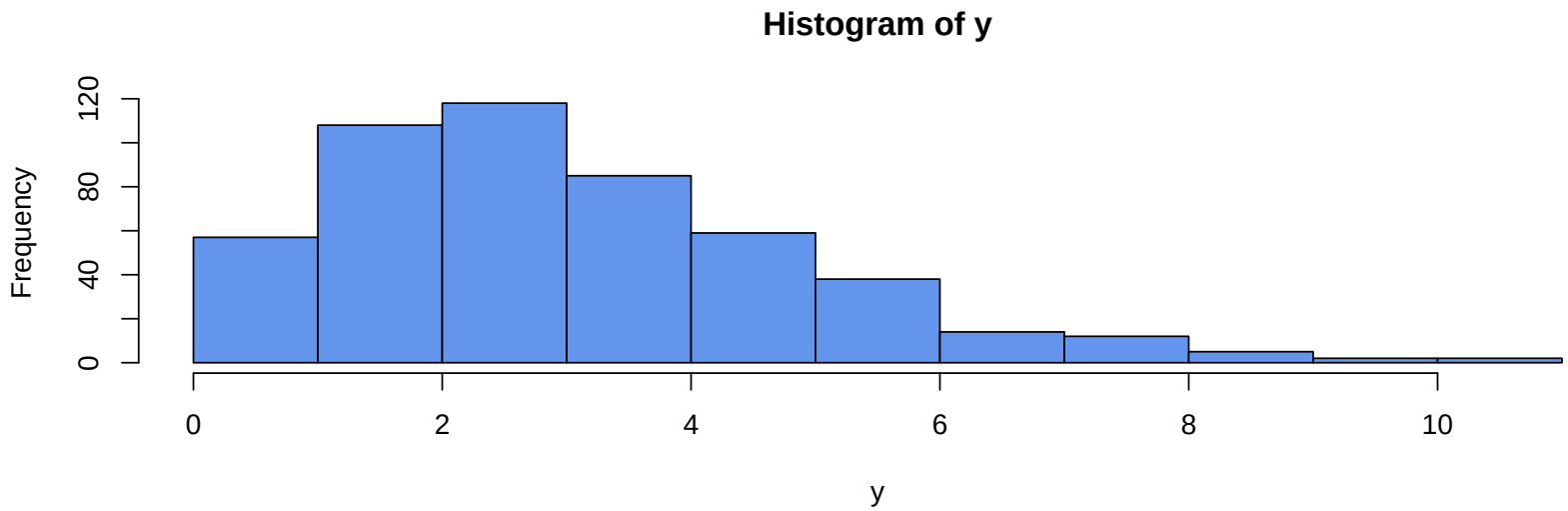
```
## $par
## [1] 1.0120727 0.4729256
##
## $value
## [1] 994.7103
##
## $counts
## function gradient
##       57       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
##
## $hessian
##           [,1]     [,2]
## [1,] 1758.9687 948.8248
## [2,]  948.8248 657.3466
```

```
import scipy.optimize as optimize
#
opt_res = optimize.minimize(
              fun = log_lik,
              x0 = [0, 0],
              args = (y, x))
print(opt_res)
```

```
##       fun: 1005.4483712482812
##  hess_inv: array([[ 0.00793944, -0.01120125],
##        [-0.01120125,  0.0168012 ]])
##       jac: array([6.10351562e-05, 3.81469727e-05])
##   message: 'Desired error not necessarily achieved due to precision loss.'
##      nfev: 327
##       nit: 8
##      njev: 79
##    status: 2
##   success: False
##         x: array([0.97663587, 0.52574927])
```

We see that the Maximum Likelihood (ML) estimates are close to the true parameter values. In conclusion, the MLE is quite handy for estimating more complex models, **provided we know the true underlying distribution of the data**.

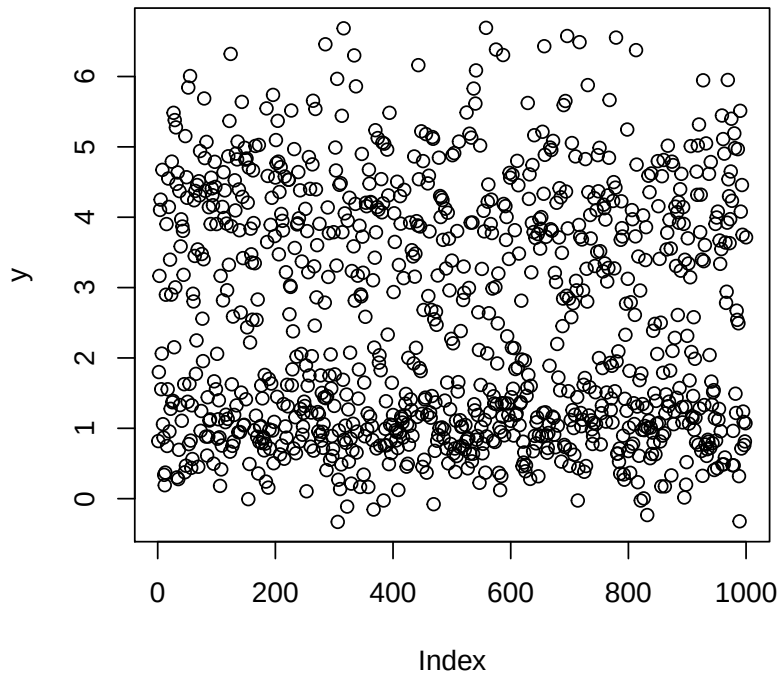
Since we don't know this in practical applications, we can always look at the histogram of the data, to get some ideas:

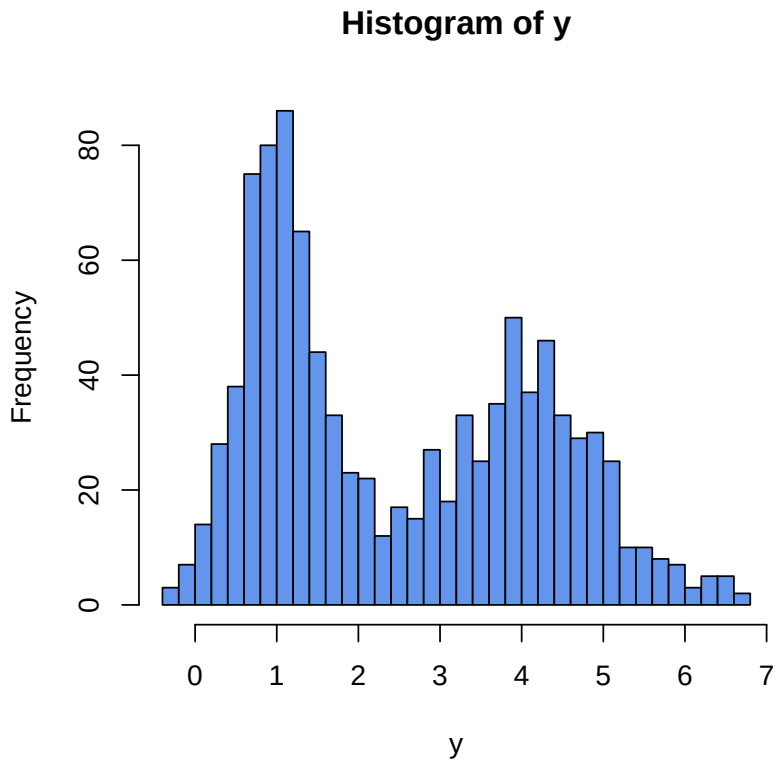


As the data seems skewed to the right (indicating a non-normal distribution), the values are non-negative and integer valued, we should look-up possible distributions for discrete data, and examine, whether our sample is similar to (at least one) of them.

Though in practice, this is easier said than done.

Let our process Y be a **mixture of normal processes** $X_1 \sim \mathcal{N}(1, (0.5)^2)$ and $X_2 \sim \mathcal{N}(4, 1^2)$ **with equal probability** and further assume that **we do not obeserve** X_1 and X_2.





As we can see from the histogram (which has two peaks) and run-sequence plot (which appears to have values **clustered** around two means), Y seems to be from a mixture of distributions.

For evaluating such cases, see Racine, J.S. (2008), Nonparametric Econometrics: A Primer.

For example, the data could contain information from:

- two shifts at work;
- sales on weekdays and weekends;
- two factories, etc.

Confidence Intervals

We will continue to examine the univariate linear regression model:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

and assume that assumptions (UR.1) - (UR.4) hold.

In this section we will introduce the notion of **interval estimation** - a procedure for creating ranges of values, called **confidence intervals**, in which the unknown parameters are likely to be located.

Confidence interval creation procedures rely heavily on (UR.4) assumption.

Interval Estimation for Parameters

We used the OLS to estimate the unknown parameter vector:

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y}$$

The estimates $\widehat{\boldsymbol{\beta}}$ are called **point estimates** - we obtain a single value for each parameter via OLS. In contrast **interval estimates** are *ranges* of values, in which the **true** parameters β_0 and β_1 are likely to fall (the interval estimates are calculated separately for each coefficient).
Interval estimation not only allows us to evaluate, what other possible values could be obtainable, but also the *precision* with which the current parameters are estimated. These interval estimates are also known as **confidence intervals**.

As we have seen, if assumptions (UR.1) - (UR.4) hold true, then the OLS estimators have a normal *conditional* distribution:

$$\widehat{\boldsymbol{\beta}}|\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}\right)$$

If you remember, we also mentioned how we can **standardize** any normal distribution by subtracting its mean (in our case $\mathbb{E}(\widehat{\beta}_i) = \beta_i$, $i = 0, 1$) and dividing by its standard deviation:

$$Z_i = \frac{\widehat{\beta}_i - \beta_i}{\sqrt{\mathbb{V}\mathrm{ar}(\widehat{\beta}_\mathbf{i})}} \sim \mathcal{N}(0, 1)$$

Note that Z_i distribution is not conditional on X. This means that when we make statements about Z_i, we do not have to worry, whether X is a random variable, or not.

Since $Z_i \sim \mathcal{N}(0, 1)$, we can use a table of normal probabilities from any statistical book, or online, and have that:

$$\mathbb{P}(-1.96 \leq Z_i \leq 1.96) = 0.95$$

Substituting the expression of Z_i yields:

$$\mathbb{P}\left(-1.96 \leq \frac{\widehat{\beta}_i - \beta_i}{\sqrt{\mathbb{V}\mathrm{ar}(\widehat{\beta}_\mathbf{i})}} \leq 1.96\right) = 0.95$$

which we can rewrite as:

$$\mathbb{P}\left(\widehat{\beta}_i - 1.96\sqrt{\mathbb{V}\mathrm{ar}(\widehat{\beta}_\mathbf{i})} \leq \beta_i \leq \widehat{\beta}_i + 1.96\sqrt{\mathbb{V}\mathrm{ar}(\widehat{\beta}_\mathbf{i})}\right) = 0.95$$

This defines the interval which has a 0.95 probability of containing the parameter β_i. In other words the end-points:

$$\widehat{\beta}_i \pm 1.96\sqrt{\mathbb{V}\mathrm{ar}(\widehat{\beta}_\mathbf{i})}, \quad i = 0, 1$$

provide an **interval estimator**. If we construct intervals this way using **all possible samples of size** N from a population, then 95% of the intervals will contain the **true parameter** β_i**,** $i = 0, 1$. Note that this assumes that **we know the true variance** $\mathbb{V}\mathrm{ar}(\widehat{\beta}_\mathbf{i})$.

As we have mentioned previously, we do not know the true variance of the error term in $\mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$, but we can estimate it. However, estimation and substitution of $\widehat{\sigma}^2$ instead of σ^2 changes the probability distribution of Z_i from a standard normal to a t-distribution with $N-2$ **degrees of freedom**:

$$t_i = \frac{\widehat{\beta}_i - \beta_i}{\mathsf{se}(\widehat{\beta}_i)} \sim t_{(N-2)}$$

where $\mathsf{se}(\widehat{\beta}_i) = \sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\beta}_\mathbf{i})}$. This is known as the **t-ratio** (or **t-statistic**) and it is the basis for **interval estimation** and **hypothesis testing** in the **univariate linear regression model**.

Proof.

The proof of this can be seen from the fact that:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2) \iff \frac{\epsilon_i}{\sigma} \sim \mathcal{N}(0, 1)$$

then the sum of squared independent standardized residuals has the chi-squared distribution with N degrees of freedom:

$$\sum_{i=1}^{N} \left(\frac{\epsilon_i}{\sigma}\right)^2 \sim \chi^2_N$$

Since the true errors are unobservable, we replace them by the OLS residuals, then the random variable $\widehat{\sigma}^2$ has a chi-squared distribution with $N - 2$ degrees of freedom:

$$V = \frac{\sum_{i=1}^{N} \widehat{\epsilon}_i^2}{\sigma^2} = \frac{(N-2)\widehat{\sigma}^2}{\sigma^2} = \left(\frac{N-2}{\sigma^2}\right) \widehat{\sigma}^2 \sim \chi^2_{N-2}$$

Proof (Cont.)

From the previously defined $Z_i = ... \sim \mathcal{N}(0, 1)$ and the newly defined $V \sim \chi^2_{N-2}$ we can define the following random variable:

$$t_i = \frac{Z_i}{\sqrt{V/(N-2)}} \sim t_{(N-2)}$$

substituting the expressions of Z_i and V, it can be shown that:

$$t_i = \frac{\widehat{\beta}_i - \beta_i}{\mathrm{se}(\widehat{\beta}_i)}$$

□

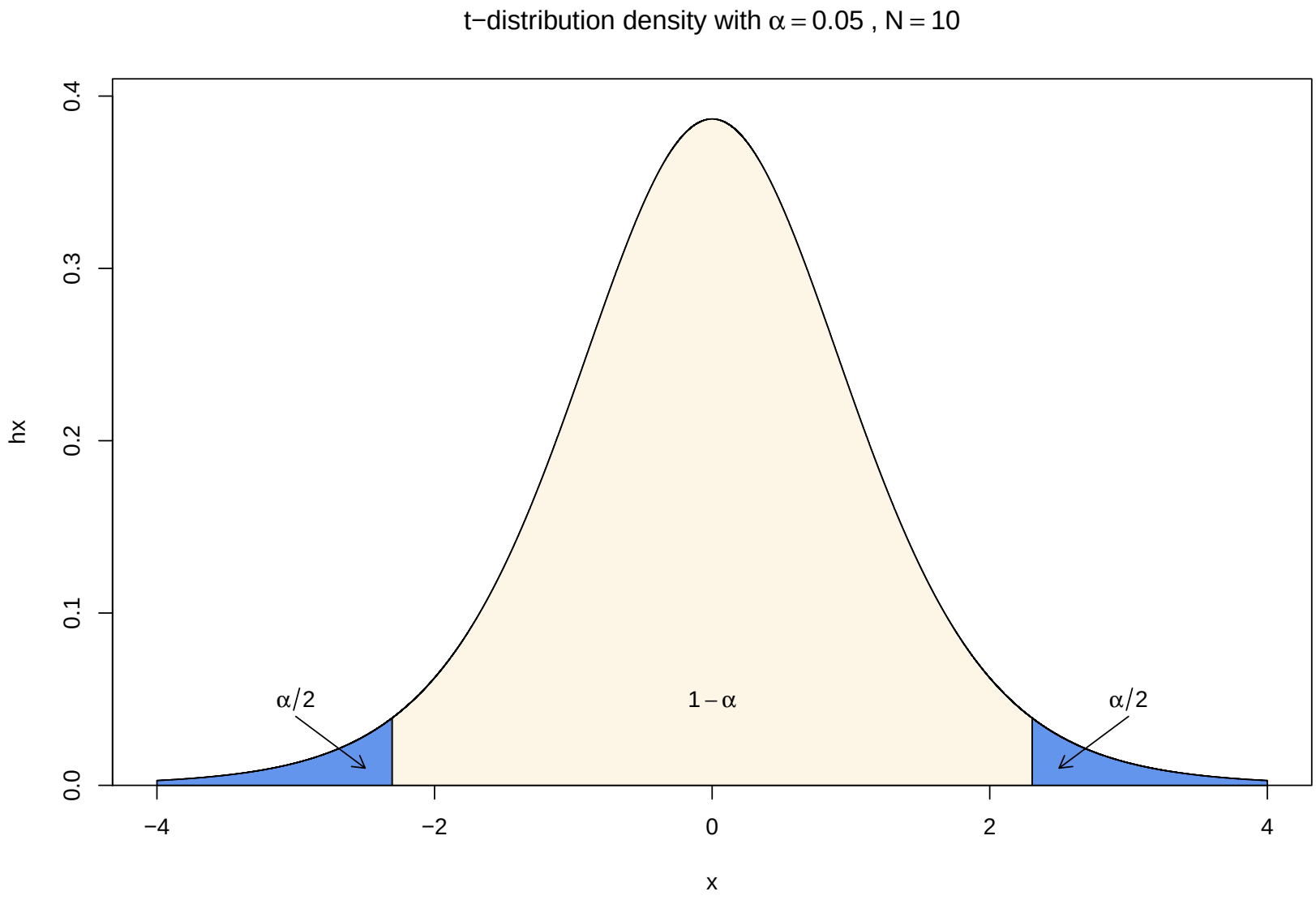
For the 95th percentile of the t-distribution with $N-2$ degrees of freedom the value $t_{(0.95,N-2)}$ has the property that 0.95 of the probability falls to its left: $\mathbb{P}\left(t_{(N-2)} \leq t_{(0.95,N-2)}\right) = 0.95$, where $t_{(N-2)}$ is from a t-dsitribution with $N-2$ degrees of freedom.

If we look at a statistical table of the percentile values for the t-distribution, we can find a **critical value** t_c, such that:

$$\mathbb{P}(t_i \geq t_c) = \mathbb{P}(t_i \leq -t_c) = \frac{\alpha}{2}$$

where α is a probability, usually $\alpha = 0.01$, $\alpha = 0.05$ or $\alpha = 0.1$. The critical value t_c for $N-2$ degrees of freedom is the **percentile** value of the t-distribution $t_{(1-\alpha/2,N-2)}$.

We can illustrate this graphically for $N = 10$:



Each blue shaded *tail* is equal to $\alpha/2$.

Consequently, we can write the probability for the **critical value** t_c as:

$$\mathbb{P}(-t_c \leq t_i \leq t_c) = 1 - \alpha$$

For a 95% **confidence interval**, the critical values define a region of the t-distribution, with probability $1 - \alpha = 0.95$. The remaining probability $\alpha = 0.05$ is divided equally ($\alpha/2 = 0.025$) between two tails:

$$\mathbb{P}(-t_{(0.975,N-2)} \leq t_i \leq t_{(0.975,N-2)}) = 0.95$$

For the univariate linear regression case, the probability becomes:

$$\mathbb{P}\left(-t_c \leq \frac{\widehat{\beta}_i - \beta_i}{\mathsf{se}(\widehat{\beta}_i)} \leq t_c\right) = 1 - \alpha$$

which we can rewrite as:

$$\mathbb{P}\left(\widehat{\beta}_i - t_c \cdot \mathsf{se}(\widehat{\beta}_i) \leq \beta_i \leq \widehat{\beta}_i + t_c \cdot \mathsf{se}(\widehat{\beta}_i)\right) = 1 - \alpha$$

The **interval estimator** of β_i is defined by these endpoints:

$$\widehat{\beta}_i \pm t_c \cdot \mathrm{se}(\widehat{\beta}_i)$$

Furthermore:

- the interval endpoints $\widehat{\beta}_i \pm t_c \cdot \mathrm{se}(\widehat{\beta}_i)$ are random because they depend on the data sample;
- the interval $\widehat{\beta}_i \pm t_c \cdot \mathrm{se}(\widehat{\beta}_i)$ has probability $1 - \alpha$ of containing the true unknown parameter β_i. this is also known as **interval estimate of** $\widehat{\beta}_i$, or the $100 \cdot (1 - \alpha)\%$ **confidence interval**;

This also highlights a few very important points about confidence intervals for OLS estimates:

- If we collect all possible samples of size N from a population, calculate the OLS estimates $\widehat{\beta}_i$, their standard errors $\text{se}(\widehat{\beta}_i)$ and construct the confidence interval $\widehat{\beta}_i \pm t_c \cdot \text{se}(\widehat{\beta}_i)$ **for each sample**, then $100 \cdot (1 - \alpha)\%$ of all intervals constructed would contain the true parameter β_i.
- For a **single sample**, the interval estimate may of may not contain the true parameter β_i. Since β_i is *unknown*, **we will never know whether this is true or not**;
- When talking about **confidence intervals**, take note that we are talking about the *confidence* in the **procedure** used to construct these interval estimates, and **not any one interval estimate**, which was calculated from a single sample of data.

We can also use confidence intervals in our interpretation of the coefficients:

> For a linear-linear univariate regression $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$ we can say, *with* 95% *confidence* that from an additional unit of X, the dependent variable Y will exhibit a change between $\widehat{\beta}_1 - t_c \cdot \text{se}(\widehat{\beta}_1)$ and $\widehat{\beta}_1 + t_c \cdot \text{se}(\widehat{\beta}_1)$.

See the lecture notes for an example.

The main takeaway from the example is that we do not know if β_1 is actually in the estimated interval. Newertheless, if we applied this procedure to all possible data samples of size N from the same population, then 95% of all constructed interval estimates will contain the true parameter value β_1.

Hence, the interval estimation procedure "works" 95% of the time. Consequently, we only have one sample, but given the reliability of our interval construction procedure, we would be "surprised" if the true parameter value β_1 is not in the calculated interval.

To sum up:

- We use the OLS estimators to obtain point estimates of unknown parameters;
- The estimated variance and standard error $\text{se}(\widehat{\beta}_i) = \sqrt{\widehat{\mathbb{V}\text{ar}}(\widehat{\beta}_\mathbf{i})}$ provide information about the sampling variability of the OLS estimators from one sample to another.
- Interval estimators combine point estimation with sampling variability to provide a *range of values*, in which the true unknown parameter value *may* fall.
- If an interval estimate is wide (which would imply a large standard error), then it implies that we do not have enough information in the sample to draw meaningful conclusions about β_1.

Interval Estimation for the Mean Response

In additional to confidence intervals for β_0 and β_1, we can calculate a **confidence interval for the mean response**. We would like an interval estimate for the mean $\widehat{\mathbb{E}}(\mathbf{Y}|\mathbf{X} = \widetilde{\mathbf{X}}) = \widehat{\mathbf{Y}} = \widetilde{\mathbf{X}}\widehat{\boldsymbol{\beta}}$ for some value of $\mathbf{X} = \widetilde{\mathbf{X}}$. The expected value of $\widehat{\mathbf{Y}}$ is an unbiased estimator of $\mathbb{E}(\mathbf{Y}|\mathbf{X} = \widetilde{\mathbf{X}})$:

$$\begin{aligned}\mathbb{E}(\widehat{\mathbf{Y}}) &= \mathbb{E}(\widetilde{\mathbf{X}}\widehat{\boldsymbol{\beta}}) = \widetilde{\mathbf{X}}\left(\mathbb{E}(\widehat{\boldsymbol{\beta}})\right) = \widetilde{\mathbf{X}}\boldsymbol{\beta} \\ &= \mathbb{E}(\mathbf{Y}|\mathbf{X} = \widetilde{\mathbf{X}})\end{aligned}$$

The variance of the mean response is:

$$\begin{aligned}\mathbb{V}\mathrm{ar}(\widehat{\mathbf{Y}}) &= \mathbb{V}\mathrm{ar}(\widetilde{\mathbf{X}}\widehat{\boldsymbol{\beta}}) \\ &= \widetilde{\mathbf{X}}\mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\beta}})\widetilde{\mathbf{X}}^\top \\ &= \widetilde{\mathbf{X}}\sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\widetilde{\mathbf{X}}^\top \\ &= \sigma^2\widetilde{\mathbf{X}}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\widetilde{\mathbf{X}}^\top\end{aligned}$$

Note that the variance of the mean response includes both $\mathbf{X}$ and $\widetilde{\mathbf{X}}$, which may have different element values.

Like we have previously seen, if (UR.4) assumption holds true, then $\widehat{\mathbf{Y}}$ follows a normal distribution:

$$\left(\widehat{\mathbf{Y}}|\widetilde{\mathbf{X}}, \mathbf{X}\right) \sim \mathcal{N}\left(\widetilde{\mathbf{X}}\beta, \quad \sigma^2\widetilde{\mathbf{X}}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\widetilde{\mathbf{X}}^\top\right)$$

where we can again, replace σ^2, with its estimate $\widehat{\sigma}^2 = \dfrac{1}{N-2}\sum_{i=1}^{N}\widehat{\epsilon}_i^2$.

Let $\text{se}(\widehat{Y}_i) = \sqrt{\widehat{\mathbb{V}\text{ar}}(\widehat{Y}_i)}$ be the square root of the corresponding i-th diagonal element of $\widehat{\mathbb{V}\text{ar}}(\widehat{\mathbf{Y}})$. Then, the $100 \cdot (1-\alpha)\%$ confidence interval for the mean response can be calculated as:

$$\widehat{Y}_i \pm t_{(1-\alpha/2, N-2)} \cdot \text{se}(\widehat{Y}_i)$$

See the lecture notes for the continuation of the previous example.

Note that the confidence interval for the mean response is different from the **prediction interval for new observations**, which we will cover in a later section.

Rule of Thumb For Confidence Interval Construction

As we have seen, confidence intervals for the estimated parameters, or for the mean response, can be computed for any sample size $N > 2$ and any confidence level $0 \leq \alpha \leq 1$. Furthermore, the t-distribution approaches the standard normal distribution as the degrees of freedom increases (i.e. as the sample size N gets larger).

In particular for $\alpha = 0.05$ we have that $t_{1-\alpha/2,N} \to 1.96$ as $N \to \infty$:

```r
NN <- c(10, 100, 1000, 1e5, 1e6)
for(i in 1:length(NN)){
  tmp_val <- qt(1 - 0.05 / 2,
                NN[i])
  print(paste0("N = ", NN[i],
      " crit.val. = ",
      round(tmp_val, 4)))
}
```

```
## [1] "N = 10 crit.val. = 2.2281"
## [1] "N = 100 crit.val. = 1.984"
## [1] "N = 1000 crit.val. = 1.9623"
## [1] "N = 1e+05 crit.val. = 1.96"
## [1] "N = 1e+06 crit.val. = 1.96"
```

```python
import scipy.stats as stats
```

```python
NN = [10, 100, 1000, 1e5, 1e6]
for i in range(0, len(NN)):
  tmp_val=stats.t.ppf(q=1-0.05/2,
                      df = NN[i])
  print("N = " + str(NN[i]) +
    " crit.val. = " +
    str(np.round(tmp_val, 4)))
#
```

```
## N = 10 crit.val. = 2.2281
## N = 100 crit.val. = 1.984
## N = 1000 crit.val. = 1.9623
## N = 100000.0 crit.val. = 1.96
## N = 1000000.0 crit.val. = 1.96
```

In some cases, the population is clearly **non-normal**. In such cases, **as long as the sample size is sufficiently large** then the OLS estimators, and the sample mean estimators are approximately normal. This lets us compute an approximate 95% confidence interval.

Let our **estimate** be (either an estimate of some model parameter, or an estimate of other population parameters, like the process mean) defined as $\overline{b}$. Then a **rule of thumb** for an *approximate* 95% confidence interval is:

$$\left[\overline{b} \pm 1.96 \cdot \text{se}(\overline{b}) \right]$$

Sometimes, an even more generalized rule of thumb $\left[\overline{b} \pm 2 \cdot \text{se}(\overline{b}) \right]$ may be used.

Hypothesis Testing

We have seen how to calculate OLS estimates and evaluate their confidence intervals (which we can also interpret). However, in practice, we usually want to answer very specific questions about the effects of specific variables:

- Does income effect expenditure?
- Do more years in education lead to an increase in wage?

Hypothesis tests use the information about a parameter from the sample data to answer such *yes*/*no* questions (though not necessarily in such strong certainty).

Hypothesis Tests consist of:

- Specification of the null hypothesis, H_0;
- Specification of the alternative hypothesis, H_1;
- Specification of the test statistic and its distribution under the null hypothesis;
- Selection of the significance level α in order to determine the rejection region;
- Calculation of the test statistic from the data sample;
- Conclusions, which are based on the test statistic and the rejection region;

We will look into each point separately.

As we have done throughout this course, assume that our linear regression model is of the following form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \text{ where } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

and assume that (UR.1)-(UR.4) hold true.

The Null Hypothesis

The **null hypothesis** is denoted by H_0, and for the univariate regression can be stated as:

$$H_0 : \beta_i = c$$

where c is a constant value, which we are interested in. When testing the null hypothesis, we may either **reject** or **fail to reject** the null hypothesis.

The null hypothesis, is presumed to be true, until the data provides sufficient evidence that it is not.

If we fail to reject the null hypothesis, it does not mean the null hypothesis is true. A hypothesis test does not determine which hypothesis is **true**, or which is most likely: it only assesses whether available evidence exists to **reject** the null hypothesis.

The Alternative Hypothesis

Once, we state our null hypothesis, we must test it *against* and **alternative hypothesis**, denoted H_1.
For the null hypothesis $H_0 : \beta_i = c$ we may specify the alternative hypothesis in thee possible ways:

- $H_1 : \beta_i > c$ - rejecting H_0, leads us to "accept" the conclusion that $\beta_i > c$. Economic theory frequently provides information about the **signs** of the variable parameters. For example: economic theory strongly suggests that food expenditure will rise if income increases, so we would test $H_0 : \beta_{INCOME} = 0$ against $H_1 : \beta_{INCOME} > 0$.
- $H_1 : \beta_i < c$ - rejecting H_0, leads us to "accept" the conclusion that $\beta_i < c$.
- $H_1 : \beta_i \neq c$ - rejecting H_0, leads us to "accept" the conclusion that β_i is either greater *or* smaller than c.

We usually talk about hypothesis testing in terms of the null, i.e. we either **reject** or **fail to reject** the null - we **never** *accept* the null. As such, if we reject the null, then we "accept" (i.e. we are left with) the alternative.

The Test Statistic

The **test statistic** is calculated under the null hypothesis (i.e. assuming the null hypothesis is *true*). Under the null hypothesis the distribution of the statistic is *known*. Based on the value of the test statistic, we decide whether to reject, or fail to reject the null.

Under the null hypothesis $H_0 : \beta_i = c$ of our univariate regression model, we can calculate the following t-**statistic**:

$$t_i = \frac{\widehat{\beta}_i - c}{\text{se}(\widehat{\beta}_i)} \sim t_{(N-2)}$$

If the null hypothesis is *not true*, then the t-statistic does not have a t-distribution with $N - 2$ degrees of freedom, but some other distribution.

The Rejection Regions

The **rejection region** consists of values that have low probability of occurring when the null hypothesis is true. The rejection region depends on the specification of the alternative hypothesis. If the calculated *test statistic* value falls in the rejection region (i.e. an unlikely event to occur under the null), then it is unlikely that the null hypothesis is holds.

The size of the rejection regions are determined by choosing a **level of significance** α - a probability of the unlikely event, usually 0.01, 0.05, 0.1.

To determine, whether to reject the null hypothesis or not, we will compare the calculated t-statistic t_i to the critical value t_c.

Type I and Type II Errors

When deciding whether to reject the null hypothesis or not, we may commit one of two types of errors:

- **Type I error - to reject the null hypothesis when it is true**. The probability of committing Type I error is $\mathbb{P}(H_0 \text{ rejected} | H_0 \text{ is true}) = \alpha$. Any time we reject the null hypothesis, it is possible that we have made such an error. We can specify the amount of Type I error, that we can tolerate, by setting the level of significance α. If we want to avoid making a **Type I** error, then we set α to a very small value.
- **Type II error - to not reject the null hypothesis when it is false**. We cannot directly calculate the probability of this type of error, since it depends on the unknown parameter β_i. However, we do know that by making α smaller we increase the probability of **Type II** error.

It is believed that a **Type I** error is more severe, hence, it is recommended to make the probability α small.

The p-value

When reporting the outcome of statistical hypothesis tests, we usually report the p-value of the test. The p-value is defined as the probability, under the null hypothesis, of obtaining a result, which is equal to, or more extreme, than what was actually observed.

Having the p-value allows us to easier determine the outcome of the test, as we do not need to directly compare the critical values.

- If $p \leq \alpha$, we **reject** H_0.
- If $p \geq \alpha$, we **do not reject** H_0.

One Tail Tests

One tail tests involve testing the null hypothesis against an alternative hypothesis, where the true regression parameter is *greater*, or H_1 where it is *less* than the specified constant c from the null $H_0 : \beta_i = c$.

Alternative, $>$

We are testing the null hypothesis $H_0 : \beta_i = c$ against the alternative $H_1 : \beta_i > c$:

$$\begin{cases} H_0 & : \beta_i = c \\ H_1 & : \beta_i > c \end{cases}$$

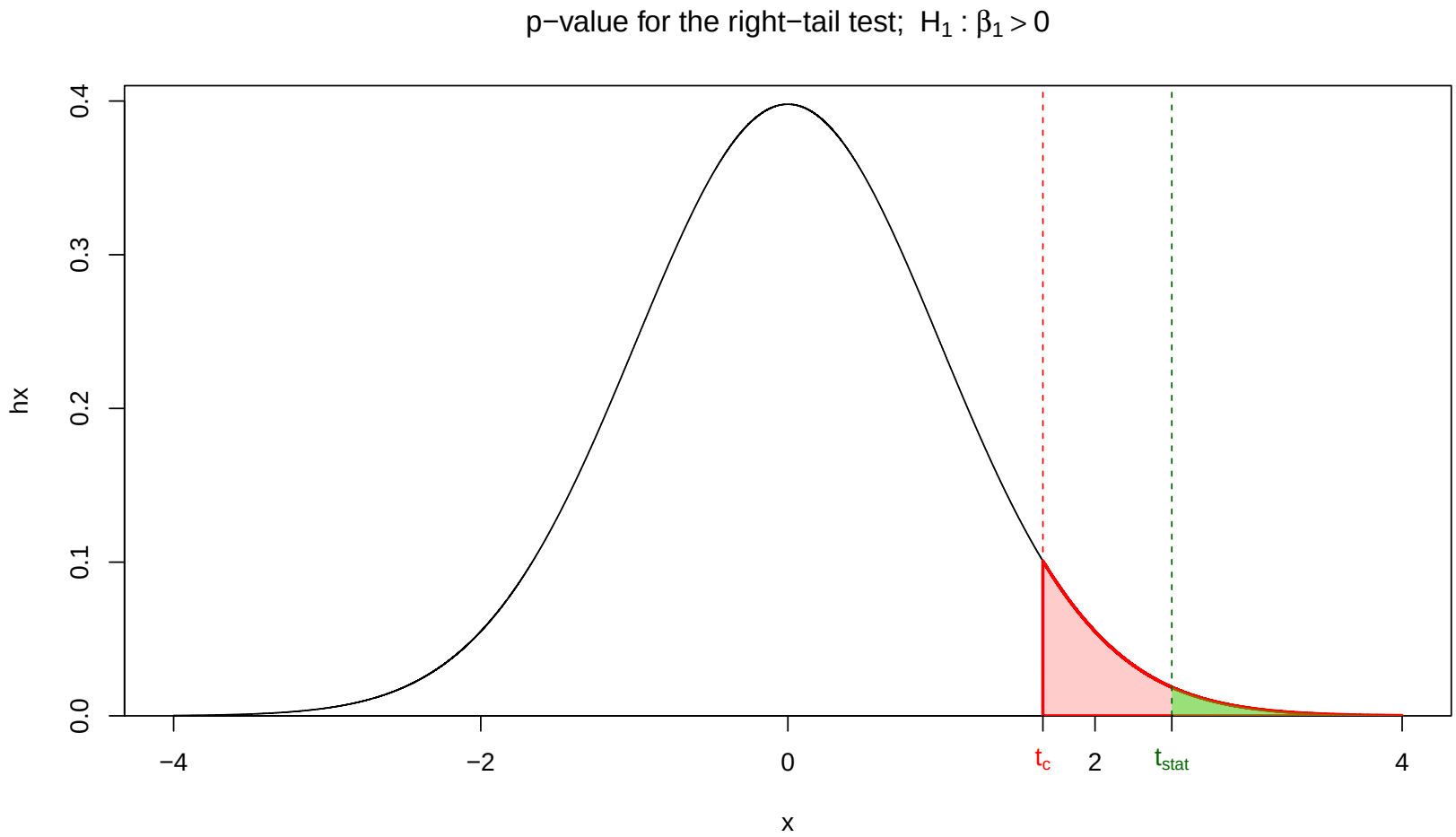
We **reject** H_0 and **accept the alternative** H_1, if $t_i \geq t_{(1-\alpha, N-2)}$, where $t_c = t_{(1-\alpha, N-2)}$.

Regarding p-value - it is the probability to the right of the calculated t-statistic and can be calculated as:

$$p\text{-value} = \mathbb{P}(T \geq t_i) = \mathbb{P}(T > t_i) = 1 - \mathbb{P}(T \leq t_i) = 1 - \int_{-\infty}^{t_i} p(x)dx$$

where $p(x)$ is the density function of the distribution of t-statistic **under the null hypothesis**. For the univariate regression, under the null hypothesis it is Students t-distribution with $N - 2$ degrees of freedom.

See the lecture notes for the code.



If we see that the *p*-value is less than the 5% significance level (as well as the fact that the calculated *t*-statistic is greater than the critical value) - we reject the null hypothesis and conclude that β_1 is statistically significantly greater than zero.

Alternative, $<$

We are testing the null hypothesis $H_0 : \beta_i = c$ against the alternative $H_1 : \beta_i < c$:

$$\begin{cases} H_0 & : \beta_i = c \\ H_1 & : \beta_i < c \end{cases}$$

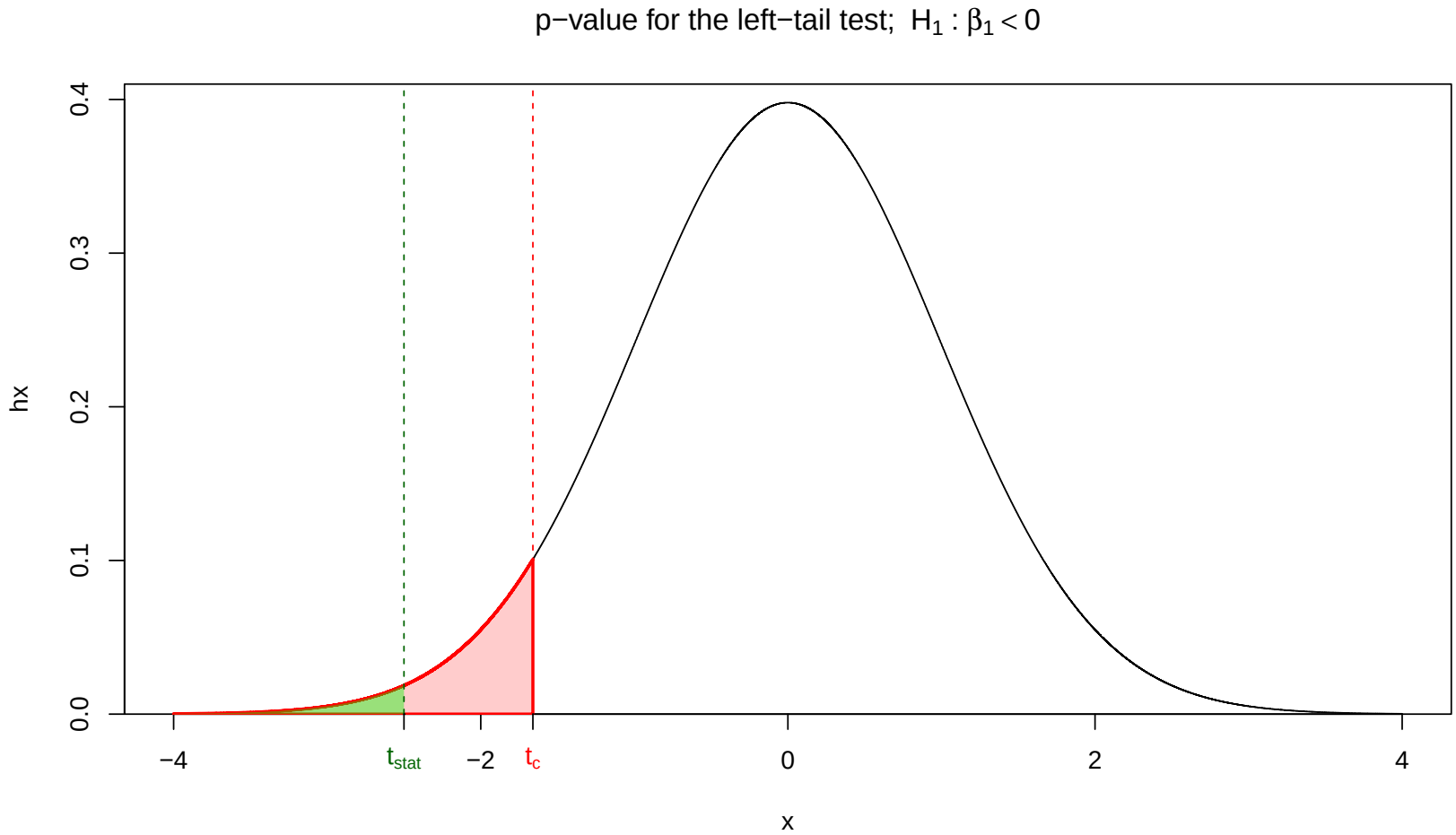
We **reject** H_0 and **accept the alternative** H_1, if $t_i \leq t_{(\alpha, N-2)}$, where $t_c = t_{(\alpha, N-2)}$.

Regarding p-value - it is the probability to the left of the calculated t-statistic and can be calculated as:

$$p\text{-value} = \mathbb{P}(T \leq t_i) = \int_{-\infty}^{t_i} p(x)dx$$

where $p(x)$ is the density function of the distribution of t-statistic **under the null hypothesis**. For the univariate regression, under the null hypothesis it is Students t-distribution with $N - 2$ degrees of freedom.

See the lecture notes for the code.



If we see that the *p*-value is less than the 5% significance level (as well as the fact that the calculated *t*-statistic is less than the critical value) - we reject the null hypothesis and conclude that β_1 is statistically significantly less than zero.

Two Tail Tests, $\neq$

We are testing the null hypothesis $H_0 : \beta_i = c$ against the alternative $H_1 : \beta_i \neq c$:

$$\begin{cases} H_0 & : \beta_i = c \\ H_1 & : \beta_i \neq c \end{cases}$$

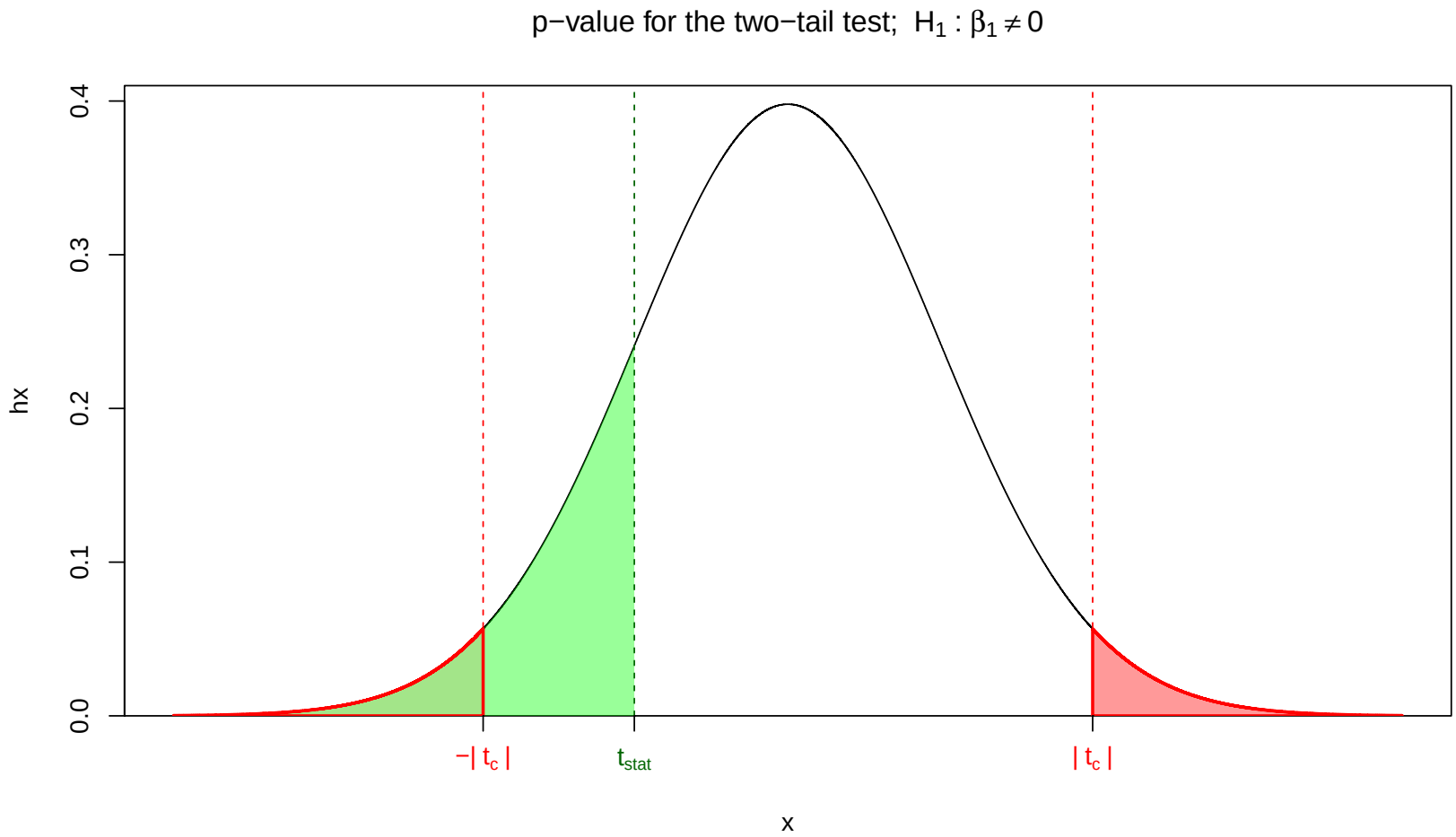
We **reject** H_0 and **accept the alternative** H_1, if $t_i \leq t_{(\alpha/2, N-2)}$ **or** if $t_i \geq t_{(1-\alpha/2, N-2)}$.

Regarding p-value - it is the sum of probabilities to the right of $|t_i|$ and to the left of $-|t_i|$ and can be calculated as:

$$\begin{aligned} p\text{-value} &= \mathbb{P}(T \leq -|t_i|) + \mathbb{P}(T \geq |t_i|) \\ &= 1 - \mathbb{P}(-|t_i| \leq T \leq |t_i|) = 1 - \int_{-|t_i|}^{|t_i|} p(x)dx \\ &= 2 \cdot \mathbb{P}(T \leq -|t_i|) = 2 \cdot \int_{-\infty}^{-|t_i|} p(x)dx \end{aligned}$$

where $p(x)$ is the density function of the distribution of t-statistic **under the null hypothesis**. For the univariate regression, under the null hypothesis it is Students t-distribution with $N - 2$ degrees of freedom.

See the lecture notes for the code.



We see that the *p*-value is greater than the 5% significance level (as well as the fact that the calculated *t*-statistic is greater than the critical value), so we have no grounds to reject the null hypothesis that β_1 is not statistically significantly different zero.

Conclusions of the Test

After completing the hypothesis test, we need to draw our conclusions - whether we reject, or fail to reject the null hypothesis.

Having estimated a model, our first concern is usually to determine, whether there is a relationship between the dependent variable Y, and the independent variable X. If $\beta_1 = 0$ in a univariate regression, then there is no linear relationship between Y and X.

In other words, are usually interested in determining if a coefficient of a predictor is **significantly different from zero**, that is $H_0 : \beta_i = 0$. This is also called a **test of significance**.

- If we reject H_0, then we say that β_1 is **statistically significantly different from zero**. Usually we want our model to have only significant variables included.
- If we fail to reject H_0, then we say that β_i is **not statistically significant**. In such a case, we usually want to remove the insignificant variable (or replace it with another, hopefully significant, variable).

Statistical Significance versus Economical significance

We have emphasized *statistical significance* throughout this section. However, we should also pay attention to the *magnitude* of the estimated coefficient. In other words:

- The **statistical significance** of a parameter $\widehat{\beta}_i$ is determined by the size of the t-statistic t_i.
- The **economic (or practical) significance** of a variable is related to the **size** and **sign** of $\widehat{\beta}_i$.

Example

Let's say our estimated model is:

$$\underset{(\text{se})}{\widehat{Y}} = \underset{(200)}{2000} + \underset{(0.00001)}{0.0001}\,X$$

where Y - is the total value of ice cream sales in EUR, and X is the expenditure on advertising in EUR.

Then the t-statistic for $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ is $t_1 = \widehat{\beta}_1/\text{se}(\widehat{\beta}_1) = (0.01)/(0.001) = 10$, which would be much greater than any critical value with $\alpha = 0.05$. This means that an additional EUR spent in advertising would result in 0.01 EUR of ice cream sales.

While this is *statistically significantly different from zero*, it may not be *economically significantly different from zero*.

If advertisement increases by 10000 EUR, then the average value on ice cream sales increases by 1 EUR - would it be worth for an ice cream company to increase their spending on advertisement in this specific case?

Null Hypothesis and The Parameter Confidence Interval

The two-tail tests and parameter confidence intervals are closely related. Assume that we want to test the following:

$$\begin{cases} H_0 & : \beta_i = c \\ H_1 & : \beta_i \neq c \end{cases}$$

Looking back at the parameter confidence interval estimation, we have that:

$$\mathbb{P}\left(\widehat{\beta}_i - t_c \cdot \text{se}(\widehat{\beta}_i) \leq \beta_i \leq \widehat{\beta}_i + t_c \cdot \text{se}(\widehat{\beta}_i)\right) = 1 - \alpha$$

Which leads to the $100 \cdot (1 - \alpha)$ confidence interval:

$$\left[\widehat{\beta}_i - t_c \cdot \text{se}(\widehat{\beta}_i); \quad \widehat{\beta}_i + t_c \cdot \text{se}(\widehat{\beta}_i)\right]$$

Under the null hypothesis we would have that $\beta_i = c$.

So, if we test the null hypothesis against the two-tailed alternative, then we should check if c belongs to the confidence interval:

- If $c \in \left[\widehat{\beta}_i - t_c \cdot \text{se}(\widehat{\beta}_i); \quad \widehat{\beta}_i + t_c \cdot \text{se}(\widehat{\beta}_i)\right]$, we will (most likely) **not reject** H_0 at the level of significance α.
- If $c \notin \left[\widehat{\beta}_i - t_c \cdot \text{se}(\widehat{\beta}_i); \quad \widehat{\beta}_i + t_c \cdot \text{se}(\widehat{\beta}_i)\right]$, we will (most likely) **reject** H_0.

Note that **this is not an alternative to hypothesis testing - you should still carry our the tests by calculating the critical values and t-statistics.** This is a neat trick to have a preliminary view of what would most likely be the outcome of the tests.

Null Hypothesis with inequalities: $\leq$ or $\geq$

See lecture notes on this topic, which highlights the fact that we never *accept*, but rather **do not reject** the null hypothesis as it not only does not give us a concrete answer (in this case multiple null hypothesis are compatible with our data sample) but also depends on the specification of both the null and the alternative hypothesis.

- We have reviewed a few frequently encountered distributions in econometrics;
- We have examined another estimation method, which relies on (UR.4) and as such, on the distribution of the dependent variable as well.
- We have examined how confidence intervals are constructed for the estimated parameters, as well as for the mean response.
- We have analysed the steps required to carry out a hypothesis test, starting from hypothesis specification, and ending with drawing conclusions based on test results.
- We have presented various ways to plot the data in order to determine the relations between different variables, or to examine the residuals.

Examples using empirical data

From the Lecture notes `Ch. 3.10` continue with the dataset(-s) that you have used from the previous exercise set and do the tasks from `Exercise Set 3` from `Ch 3.10`. See `Ch. 3.11` for an example.