PE I: Univariate Regression

OLS, Regression Models & Interpretation (Chapters 3.2 & 3.3)

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: Estimation

Problem: Assume that the data follows (UR.1) - (UR.4). However, **we do not know the true parameters** β_0 and β_1 of our underlying regression $Y = \beta_0 + \beta_1 X + \epsilon$.

Question: How do we evaluate β_0 and β_1?

Answer: Specify the equations in a matrix notation:

$$\begin{cases} Y_1 &= \beta_0 + \beta_1 X_1 + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \epsilon_2 \\ \vdots & \\ Y_N &= \beta_0 + \beta_1 X_N + \epsilon_N \end{cases} \iff \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_N \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

and minimize the sum of squared residuals of $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$:

$$RSS(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{N} \widehat{\epsilon}_i^2 = \widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}} = \left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)^\top \left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \to \min_{\widehat{\beta}_0, \widehat{\beta}_1}$$

(Alternatively, either use the method of moments or directly minimize the sum of the residuals instead of its vectorized form)

OLS: The Estimator

Taking the partial derivative and equating it to zero:

$$\frac{\partial RSS(\widehat{\beta})}{\partial \widehat{\beta}} = -2\mathbf{X}^\top \mathbf{Y} + 2\mathbf{X}^\top \mathbf{X}\widehat{\beta} = 0$$

yields the **OLS estimator**:

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

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

Gauss-Markov Theorem

The main advantage of the OLS estimators are summarized by the following theorem:

Gauss-Markov theorem
Under the assumption that the conditions (UR.1) - (UR.3) hold true, the OLS estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are **BLUE** (**B**est **L**inear **U**nbiased **E**stimator) and **Consistent**.

Univariate Regression: Modelling Framework

Assume that we are interested in *explaining Y in terms of X* under (UR.1) - (UR.4):

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \ \forall i = 1, ..., N$$

where:

- X is called the *independent variable*, the **control variable**, the **explanatory variable**, the **predictor variable**, or the **regressor**;
- Y is called the *dependent variable*, the **response variable**, the **explained variable**, the **predicted variable** or the **regressand**;
- ϵ is called the *random component*, the **error term**, the **disturbance** or the **(economic) shock**.

The expected value:

$$\mathbb{E}(\beta_0 + \beta_1 X + \epsilon | X) = \beta_0 + \beta_1 X$$

where:

- β_0 - the **intercept parameter**, sometimes called the *constant term*.
- β_1 - the **slope parameter**.

After obtaining the estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$, we may want to examine the following values:

- The **fitted values** of Y, which are defined as the following **OLS regression line** (or more generally, the **estimated regression line**):

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

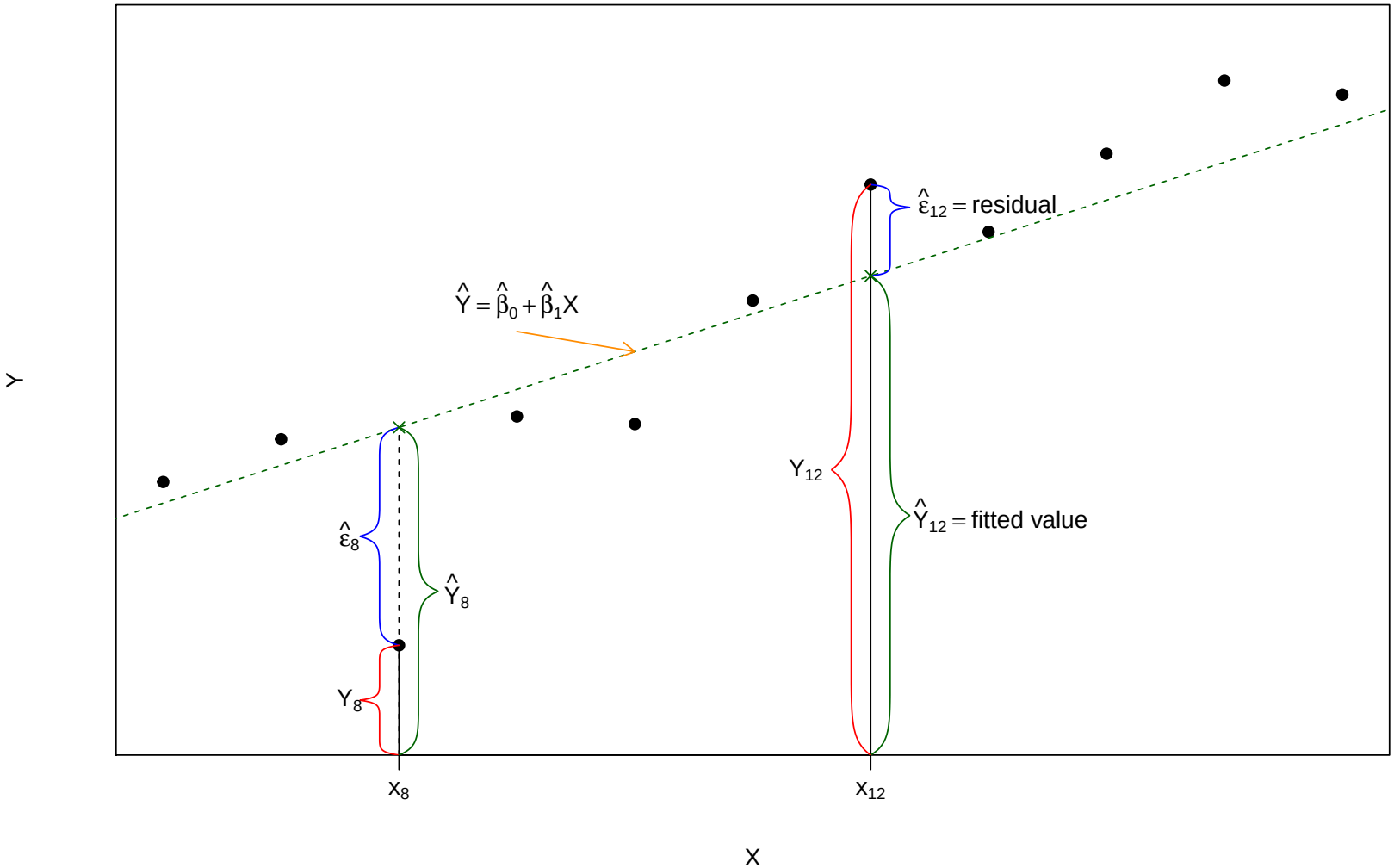
  where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are estimated via **OLS**. By definition, each fitted value of $\widehat{Y}_i$ is on the estimated OLS regression line.
- The **residuals**, which are defined as the **difference** between the **actual** and **fitted** values of Y:

$$\widehat{\epsilon}_i = \widehat{e}_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$$

  which are hopefully close to the true unobserved errors ϵ_i.

It may be helpful to look at ϵ as an *unexplainable part* of the model, which is due to the randomness of the data.
As such, the *explainable part* of the model can be expressed in terms of the fitted values $\widehat{Y}$, which themselves are estimates of the conditional expected value of Y, given X.

Scatter diagram of (X,Y) sample data and the regression line



Now is also a good time to highlight the difference between the **errors** and the **residuals**.

- The random sample, taken from a **Data Generating Process** (i.e. the **population**), is described via

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

  where ϵ_i is the **error** for observation i.
- After estimating the unknown parameters β_0, β_1, we can **re-write** the equation as:

$$Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\epsilon}_i$$

  where $\widehat{\epsilon}_i$ is the **residual** for observation i.

The *errors* show up in the underlying (i.e. true) DGP equation, while the *residuals* show up in the *estimated* equation. **The errors are never observed, while the residuals are calculated from the data**.

We can also re-write the residuals in terms of the error term and the difference between the true and estimated parameters:

$$\widehat{\epsilon}_i = Y_i - \widehat{Y}_i = \beta_0 + \beta_1 X_i + \epsilon_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i) = \epsilon_i - \left(\widehat{\beta}_0 - \beta_0\right) - \left(\widehat{\beta}_1 - \beta_1\right) X_i$$

Gauss-Markov theorem

If the conditions (UR.1) - (UR.3) hold true, the OLS estimator (OLS)

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

is **BLUE** (**B**est **L**inear **U**nbiased **E**stimator) and **Consistent**.

What is an Estimator?

An **estimator** is a rule that can be applied to any sample of data to produce an **estimate**. In other words the **estimator** is the rule and the **estimate** is the result.

So, eq. (OLS) is the rule and therefore an **E**stimator.

OLS estimators are Linear

From the specification of the relationship in (UR.1) between $\mathbf{Y}$ and $\mathbf{X}$ (using the matrix notation for generality):

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

We see that the relationship is **L**inear with respect to $\mathbf{Y}$.

OLS estimators are Unbiased

Using the matrix notation for the sample linear equations ($\mathbf{Y} = \mathbf{X}\beta + \varepsilon$) and plugging it into eq. (OLS) gives us the following:

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

If we take the expectation of both sides, use the law of total expectation and the fact that $\mathbb{E}\left(\varepsilon|\mathbf{X}\right) = \mathbf{0}$ from (UR.2):

$$\mathbb{E}\left[\widehat{\beta}\right] = \beta + \mathbb{E}\left[\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \varepsilon\right] = ... = \beta$$

We have shown that $\mathbb{E}\left[\widehat{\beta}\right] = \beta$ - i.e., the OLS estimator $\widehat{\beta}$ is an **U**nbiased estimator of β.

Furthermore:

- Unbiasedness does not guarantee that the estimate we get with any particular sample is equal (or even very close) to β.
- It means that if we could *repeatedly* draw random samples from the population and compute the estimate each time, then the average of these estimates would be (very close to) β.
- However, *in most applications* we have just one random sample to work with. As we will see later on, there are methods for creating additional samples from the available data by creating and analysing different *subsamples*.

OLS estimators are Best (Efficient)

- When there is more than one unbiased method of estimation to choose from, that estimator which has the lowest variance is the **best**.
- We want to show that OLS estimators are *best* in the sense that $\widehat{\boldsymbol{\beta}}$ are **efficient** estimators of $\boldsymbol{\beta}$ (i.e. they have the **smallest variance**).

From the proof of **unbiasedness** of the OLS we have that:

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

Which we can then use this expression for calculating the **variance-covariance matrix** of the OLS estimator:

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

Note: we are usually interested in the diagonal elements of the parameter variance-covariance matrix:

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

That is, the variance of the parameter estimates themselves.

Question: Is this variance really the *best*?

Answer: To verify this, assume that we have some *other* estimator of $\boldsymbol{\beta}$, which is also *unbiased* and can be expressed as:

$$\widetilde{\boldsymbol{\beta}} = \left[\left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top + \mathbf{D} \right] \mathbf{Y} = \mathbf{C}\mathbf{Y}$$

As long as we can express it as the above eq., we can show that, since $\mathbb{E}\left[\boldsymbol{\varepsilon}\right] = \mathbf{0}$:

$$\begin{aligned} \mathbb{E}\left[\widetilde{\boldsymbol{\beta}}\right] &= \mathbb{E}\left[\left(\left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top + \mathbf{D} \right) \left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \right) \right] = \ldots = \left(\mathbf{I} + \mathbf{D}\mathbf{X} \right) \boldsymbol{\beta} \\ &= \boldsymbol{\beta} \iff \mathbf{D}\mathbf{X} = \mathbf{0} \end{aligned}$$

then $\widetilde{\boldsymbol{\beta}}$ **is unbiased if and only if** $\mathbf{DX} = \mathbf{0}$.

Then, we can calculate its variance as:

$$\begin{aligned} \mathbb{V}\mathrm{ar}(\widetilde{\boldsymbol{\beta}}) &= \mathbb{V}\mathrm{ar}(\mathbf{C}\mathbf{Y}) = \mathbf{C}\mathbb{V}\mathrm{ar}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})\mathbf{C}^\top = \sigma^2 \mathbf{C}\mathbf{C}^\top \\ &= \ldots = \sigma^2 \left[\left(\mathbf{X}^\top \mathbf{X} \right)^{-1} + \mathbf{D}\mathbf{D}^\top \right] = \mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\beta}}) + \mathbf{D}\mathbf{D}^\top \geq \mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\beta}}) \end{aligned}$$

since $\mathbf{D}\mathbf{D}^\top$ is a positive semidefinite matrix. This means that $\widehat{\boldsymbol{\beta}}$ has the smallest variance. Therefore, $\widehat{\boldsymbol{\beta}}$ is the **B**est estimator of $\boldsymbol{\beta}$.

The lower the variance of an estimator, the more **precise** (accurate) it is.

Problem: Looking back at $\mathbb{V}\mathrm{ar}(\widehat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$ - we do not know the true σ^2.

Question: How do we get σ^2?

Answer: We estimate it by calculating the **sample** residual variance:

$$\widehat{\sigma}^2 = s^2 = \frac{\widehat{\epsilon}^\top \widehat{\epsilon}}{N-2} = \frac{1}{N-2} \sum_{i=1}^{N} \widehat{\epsilon}_i^2$$

Note that if we take N instead of $N-2$ for the **univariate regression case** in the denominator, then the variance estimate would be **biased**. This is because the variance estimator would not account for **two** restrictions that must be satisfied by the OLS residuals, namely:

$$\sum_{i=1}^{N} \widehat{\epsilon}_i = 0, \quad \sum_{i=1}^{N} \widehat{\epsilon}_i X_i = 0$$

So, we take $N-2$ instead of N, because of the number of restrictions on the residuals.

Note that this is an **estimated variance**. Nevertheless, it is a key component in assessing the accuracy of the parameter estimates (when calculating test statistics and confidence intervals).

Since we estimate $\widehat{\boldsymbol{\beta}}$ from the a random sample, the estimator $\widehat{\boldsymbol{\beta}}$ is a random variable as well. We can measure the uncertainty of $\widehat{\boldsymbol{\beta}}$ via its standard deviation. This is the *standard error* of our estimate of $\boldsymbol{\beta}$:

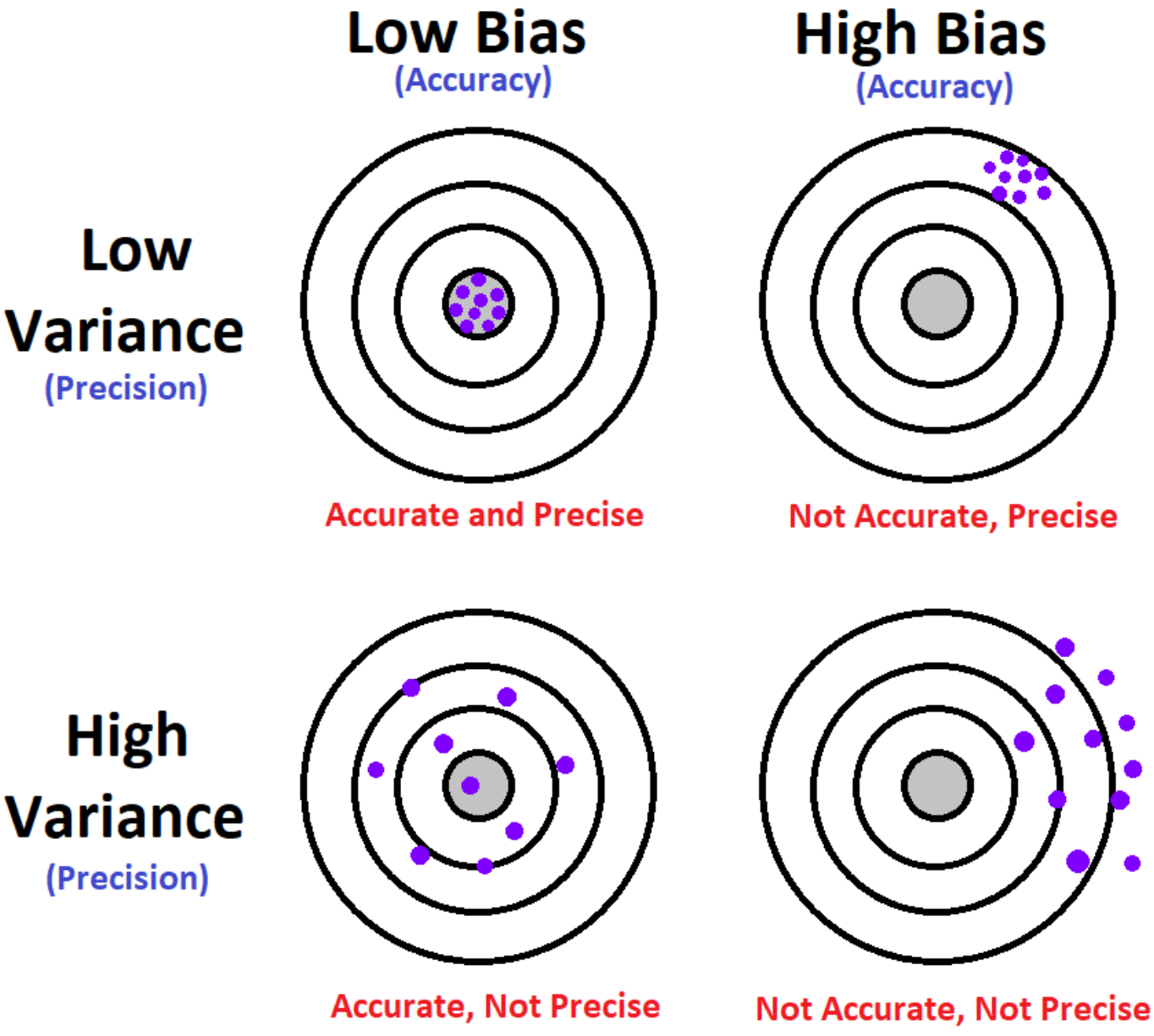
The square roots of the diagonal elements of the variance-covariance matrix $\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\boldsymbol{\beta}})$ are called **the standard errors (se)** of the corresponding OLS estimators $\widehat{\boldsymbol{\beta}}$, which we use to **estimate** the standard **deviation** 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).

- The standard errors are measures of the **sampling variability** of the least squares estimates $\widehat{\beta}_1$ and $\widehat{\beta}_2$ in **repeated samples**;
- If we collect a number of different data samples, the OLS estimates will be different for each sample. As such, the OLS estimators are **random variables** and have their own distribution.
- Potential problem: If the residuals are large (since their mean will still be zero this concerns the case when the estimated variance of the residuals is large), then the standard errors of the coefficients are large as well.

While the theoretical properties of unbiased estimators with low variance are nice to have - what is their significance in practical applications?



We have shown that the (OLS) estimator is **BLUE**. Finally, we move on to examining their consistency.

OLS estimators are Consistent

A consistent estimator has the property that, as the number of data points (which are used to estimate the parameters) increases (i.e. $N \to \infty$), the estimates converges in probability to the true parameter, i.e.:

Definition

Let $\widetilde{\theta}_N$ be an estimator of a parameter θ based on a sample $Y_1, ..., Y_N$. Then we say that $\widetilde{\theta}_N$ is a **consistent** estimator of θ if $\forall \epsilon > 0$:

$$\mathbb{P}\left(|\widetilde{\theta}_N - \theta| > \epsilon\right) \to 0, \text{ as } N \to \infty$$

We can denote this as $\widetilde{\theta}_N \xrightarrow{P} \theta$ or $\text{plim}(\widetilde{\theta}_N) = \theta$.
If $\widetilde{\theta}_N$ is not consistent, then we say that $\widetilde{\theta}_N$ is **inconsistent**.

- Unlike unbiasedness, consistency involves the behavior of the sampling distribution of the estimator as the sample size N gets large and the distribution of $\widetilde{\theta}_N$ becomes more and more concentrated about θ. In other words, for larger sample sizes, W_N is less and less likely to be very far from θ.
- An inconsistent estimator does not help us learn about θ, regardless of the size of the data sample.
- For this reason, **consistency is a minimal requirement** of an estimator used in statistics or econometrics.

Unbiased estimators are not necessarily consistent, but those whose variances shrink to zero as the sample size grows are consistent.

For some examples, see the lecture notes.

Going back to our OLS estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ - since we can express the estimator as $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \boldsymbol{\varepsilon}$, then, as $N \to \infty$ we have that:

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &\to \begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix} + \frac{1}{\mathbb{V}\mathrm{ar}(X)} \begin{bmatrix}\mathbb{E}(\epsilon) \cdot \mathbb{E}(X^2) - \mathbb{E}(X\epsilon) \cdot \mathbb{E}(X) \\ \mathbb{E}(X\epsilon) - \mathbb{E}(X) \cdot \mathbb{E}(\epsilon)\end{bmatrix} \\ &= \begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix} + \frac{1}{\mathbb{V}\mathrm{ar}(X)} \begin{bmatrix}0 \\ 0\end{bmatrix} = \begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix}\end{aligned}$$

Since $\mathbb{E}(\epsilon) = 0$ and $\mathbb{E}(X\epsilon) = \mathbb{E}(X\epsilon) - \mathbb{E}(X)\mathbb{E}(\epsilon) = \mathbb{C}\mathrm{ov}(X, \epsilon) = 0$.

Which means that $\widehat{\boldsymbol{\beta}} \to \boldsymbol{\beta}$, as $N \to \infty$.

So, the OLS parameter vector $\widehat{\boldsymbol{\beta}}$ is a **consistent** estimator of $\boldsymbol{\beta}$.

Practical illustration of the OLS properties

We will return to our example in this chapter. We have proved the *unbiasedness* and *consistency* of OLS estimators.

To illustrate these properties **empirically**, we will:

- generate 5000 replications (i.e. different samples) for each of the different sample sizes $N \in \{11, 101, 1001\}$.
- for each replication of each sample size we will estimate the unknown regression parameters β;
- for each sample size, we will calculate the average of these parameter vectors.

This method of using repeat sampling is also known as a Monte Carlo method.

The extensive code can be found in the lecture notes.

In our experimentation the **true parameter values** are:

```
## True beta_0 = 1. True beta_1 = 0.5
```

while the average values of the parameters from 5000 different samples for each sample size is:

```
## With N = 10:
##   the AVERAGE of the estimated parameters:
##       beta_0: 1.00437
##       beta_1: 0.49984
## With N = 100:
##   the AVERAGE of the estimated parameters:
##       beta_0: 0.99789
##       beta_1: 0.50006
## With N = 1000:
##   the AVERAGE of the estimated parameters:
##       beta_0: 0.99957
##       beta_1: 0.5
```

We can see that:

- The **mean of the estimated parameters** are close to the true parameter value regardless of sample size.

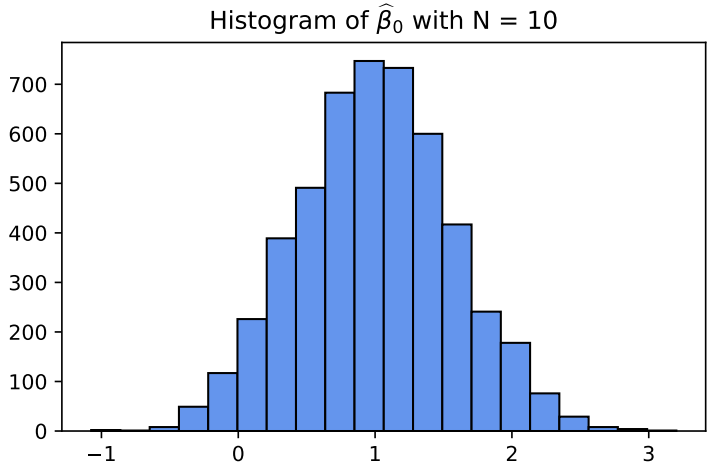
The variance of these estimates can also be examined:

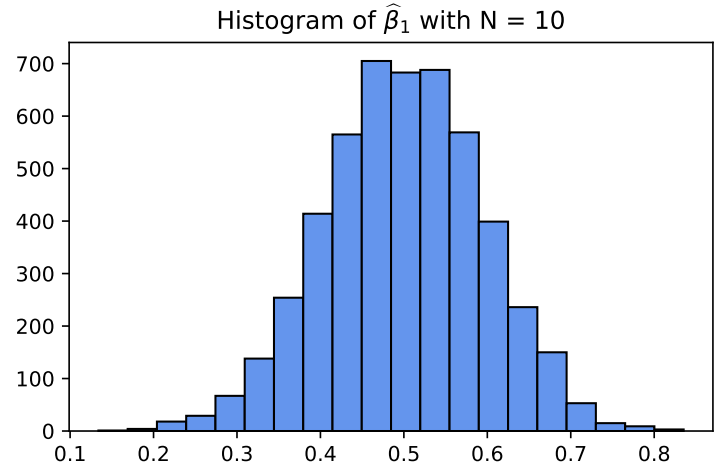
```
## With N = 10:
##   the VARIANCE of the estimated parameters:
##       beta_0: 0.3178
##       beta_1: 0.00904
## With N = 100:
##   the VARIANCE of the estimated parameters:
##       beta_0: 0.03896
##       beta_1: 1e-05
## With N = 1000:
##   the VARIANCE of the estimated parameters:
##       beta_0: 0.00394
##       beta_1: 0.0
```

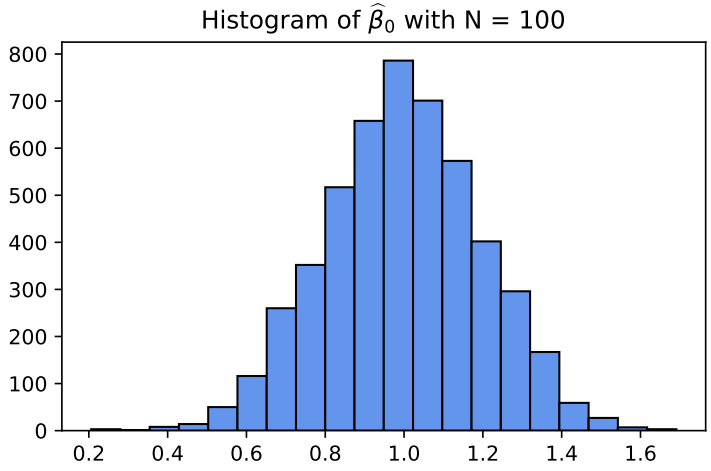
Note that **unbiasedness** is true for any N, while consistency is an **asymptotic** property, i.e. it holds when $N \to \infty$.

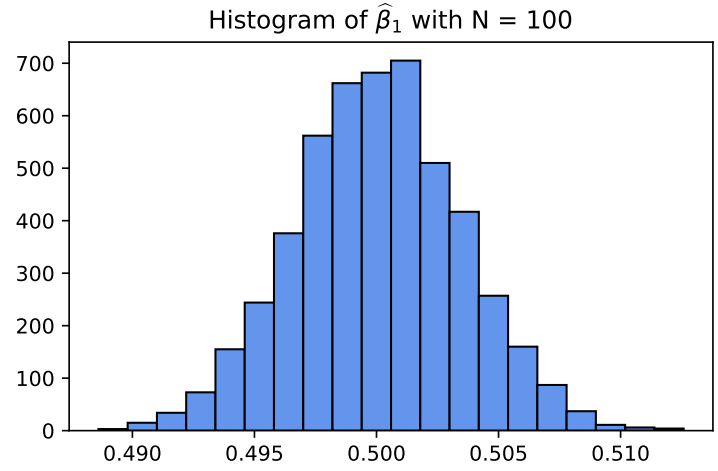
We can see that:

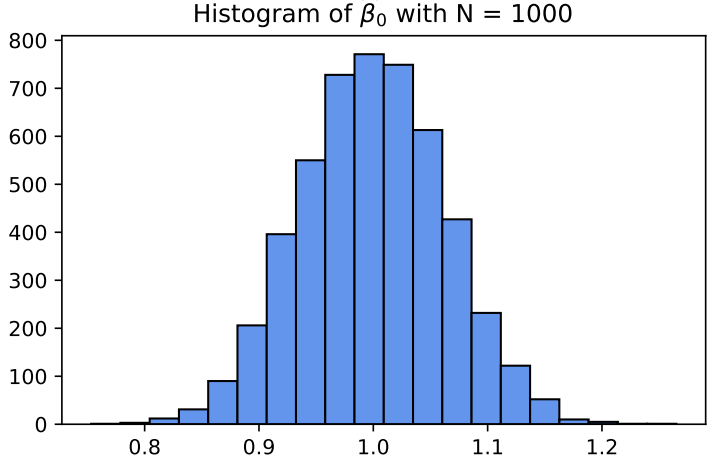
- The variance of the estimated parameters decreases with larger sample size, i.e. the larger the sample size, the closer will our estimated parameters be to the true values.

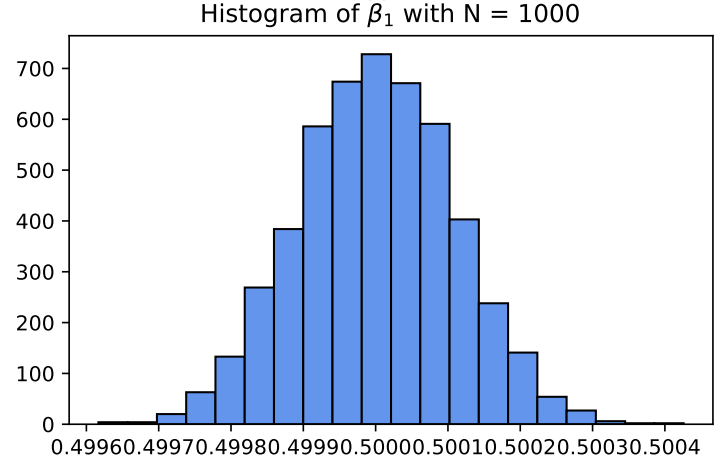

Histogram of $\widehat{\beta}_0$ with N = 10
0
100
200
300
400
500
600
700
−1
0
1
2
3


Histogram of $\widehat{\beta}_1$ with N = 10
0
100
200
300
400
500
600
700
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8


Histogram of $\widehat{\beta}_0$ with N = 100
0
100
200
300
400
500
600
700
800
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6


Histogram of $\widehat{\beta}_1$ with N = 100
0
100
200
300
400
500
600
700
0.490
0.495
0.500
0.505
0.510


Histogram of $\widehat{\beta}_0$ with N = 1000
0
100
200
300
400
500
600
700
800
0.8
0.9
1.0
1.1
1.2


Histogram of $\widehat{\beta}_1$ with N = 1000
0
100
200
300
400
500
600
700
0.4996
0.4997
0.4998
0.4999
0.5000
0.5001
0.5002
0.5003
0.5004

We see that the histograms of the OLS estimators have a bell-shaped distribution.

Under assumption (UR.4) it can be shown that since $\varepsilon|\mathbf{X} \sim \mathcal{N}\left(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}\right)$, then the linear combination of epsilons in $\widehat{\beta} = \beta + \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \varepsilon$ will also be normal, i.e.

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

Regression Models and Interpretation

Inclusion of the constant term in the regression

In some cases we want to impose a *restriction* that if $X = 0$, then $Y = 0$ as well.

- An example could be the relationship between income (X) and income tax revenue (Y) - if there is no income, $X = 0$, then the expected revenue from the taxes would also be zero - $\mathbb{E}(Y|X = 0) = 0$.

Formally, we now choose a **slope estimator**, β_1, from the following regression model:

$$Y_i = \beta_1 X_i + \epsilon_i, \ \ i = 1, ..., N$$

which is called a **regression through the origin**, because the conditional expected value:

$$\mathbb{E}\left(Y_i|X_i\right) = \beta_1 X_i$$

of the regression passes through the origin point $X = 0$, $Y = 0$. We can obtain the estimate of the slope parameter via OLS by minimizing the sum of squared residuals:

$$\text{RSS} = \sum_{i=1}^{N} \widehat{\epsilon}_i^2 = \sum_{i=1}^{N} \left(Y_i - \widehat{\beta}_1 X_i\right)^2 \rightarrow \min \Longrightarrow \widehat{\beta}_1 = \frac{\sum_{i=1}^{N} X_i Y_i}{\sum_{i=1}^{N} X_i^2}$$

So, it is possible to specify a regression **without a constant term**, but should we opt for it?

- A constant β_0 can be described as the mean value of Y when all predictor variables are set to zero. However, if the predictors can't be zero, then it is impossible to interpret the constant.
- The intercept parameter β_0 may be regarded as a sort of *garbage collector* for the regression model. The reason for this is the underlying assumption that the expected value of the residuals is *zero*, which means that any **bias**, that is not accounted by the model, is collected in the intercept β_0.

In general, inclusion of a constant in the regression model ensures that the models residuals have a mean of zero, otherwise the estimated coefficients may be biased.

Consequently, and as is often the case, without knowing the true underlying model it is generally not worth interpreting the regression constant.

Linear Regression Models

The regression model that we examined up to now is called a **simple linear regression** model. Looking at the univariate regression:

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

where $\boldsymbol{\beta} = (\beta_0,\ \beta_1)^\top$.

When we say **linear regression**, we mean *linear in parameters* $\boldsymbol{\beta}$. There are no restrictions on transformations of X and Y, as long as the parameters enter the equation linearly.

> For example, we can use $\log(X)$ and $\log(Y)$, or $\sqrt{X}$ and $\sqrt{X}$ etc. in the univariate regression. While transforming X and Y does not effect the linear regression specification itself, the **interpretation of the coefficients depends on the transformation of X and Y.**

On the other hand, there are regression models, which are **not** regarded as *linear*, since they are not linear in their parameters:

$$Y_i = \frac{1}{\beta_0 + \beta_1 X_i} + \epsilon_i,\ \ i = 1, ..., N$$

Furthermore, estimation of such models is a separate issue, which covers **nonlinear regression** models.

Effects of Changing the Measurement Units

Generally, it is easy to figure out what happens to the intercept, β_0, and slope, β_1, estimates when the units of measurement are changed for the dependent variable, Y:

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

In other words, if Y is multiplied by a constant c, then the OLS estimates of $\widetilde{Y} = c \cdot Y$ are $\widetilde{\beta}_0 = c \cdot \beta_0$ and $\widetilde{\beta}_1 = c \cdot \beta_1$.

```
#
#
set.seed(234)
# Set the coefficients:
N = 50
beta_0 = 1
beta_1 = 0.5
const  = 10
# Generate sample data:
x <- 0:N
#
e <- rnorm(mean = 0, sd = 1,
           n = length(x))
y <- beta_0 + beta_1 * x + e
new_y <- y * const
```

```
import numpy as np
#
np.random.seed(234)
# Set the coefficients:
N = 50
beta_0 = 1
beta_1 = 0.5
const  = 10
# Generate sample data:
x = np.arange(start = 0,
        stop = N + 1, step = 1)
e = np.random.normal(loc = 0,
        scale = 1, size = len(x))
y = beta_0 + beta_1 * x + e
new_y = y * const
```

```r
#
#
#
y_fit <- lm(y ~ x)
new_fit <- lm(new_y ~ x)
print(y_fit$coefficients)
```

```
## (Intercept)           x
##   0.8337401   0.5047895
```

```r
print(new_fit$coefficients)
```

```
## (Intercept)           x
##    8.337401    5.047895
```

```python
import statsmodels.api as sm
#
x_mat = sm.add_constant(x)
y_fit = sm.OLS(y, x_mat).fit()
new_fit = sm.OLS(new_y, x_mat).fit()
print(y_fit.params)
```

```
## [0.83853621 0.5099222 ]
```

```python
print(new_fit.params)
```

```
## [8.38536209 5.09922195]
```

Note that the variance of $c \cdot \epsilon$ is now $c^2\sigma^2$:

```r
print(summary(y_fit)$sigma^2)
```

```
## [1] 0.9329815
```

```r
print(summary(new_fit)$sigma^2)
```

```
## [1] 93.29815
```

```python
print(y_fit.scale)
```

```
## 0.9570628350921718
```

```python
print(new_fit.scale)
```

```
## 95.70628350921723
```

This assumes that nothing changes in the scaling of the independent variable X used in OLS estimation.

If we change the units of measurement of an independent variable, X, then only the slope, β_1, (i.e. the coefficient of that independent variable) changes:

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

In other words, if X is multiplied by a constant c, then the OLS estimates of Y are β_0 and $\widetilde{\beta}_1 = \frac{\beta_1}{c}$.

We can verify that this is the case with our empirical data sample by creating new variables $X^{(1)} = X \cdot c$ and $X^{(2)} = X/c$

```r
x1 <- x * const
x2 <- x / const
#
y_fit_x_mlt <- lm(y ~ x1)
y_fit_x_div <- lm(y ~ x2)
```

```python
x1_mat = sm.add_constant(x * const)
x2_mat = sm.add_constant(x / const)
#
y_fit_x_mlt = sm.OLS(y, x1_mat).fit()
y_fit_x_div = sm.OLS(y, x2_mat).fit()
```

```
print(y_fit$coefficients)
```

```
## (Intercept)           x
##   0.8337401   0.5047895
```

```
print(y_fit_x_mlt$coefficients)
```

```
## (Intercept)          x1
##  0.83374010  0.05047895
```

```
print(y_fit_x_div$coefficients)
```

```
## (Intercept)          x2
##   0.8337401   5.0478946
```

```
print(y_fit.params)
```

```
## [0.83853621 0.5099222 ]
```

```
print(y_fit_x_mlt.params)
```

```
## [0.83853621 0.05099222]
```

```
print(y_fit_x_div.params)
```

```
## [0.83853621 5.09922195]
```

Furthermore, if we scale both X and Y by the same constant:

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

In other words, if both Y and X are multiplied by **the same** constant c, then the OLS estimates for the intercept change to $\widetilde{\beta_0} = c \cdot \beta_0$ but remain the same for the slope β_1.

```
x1 <- x * const
new_y <- y * const
#
y_fit_scaled <- lm(new_y ~ x1)
print(y_fit$coefficients)
```

```
## (Intercept)           x
##   0.8337401   0.5047895
```

```
print(y_fit_scaled$coefficients)
```

```
## (Intercept)          x1
##   8.3374010   0.5047895
```

```
x1_mat = sm.add_constant(x * const)
new_y = y * const
#
y_fit_scaled = sm.OLS(new_y, x1_mat).fit
print(y_fit.params)
```

```
## [0.83853621 0.5099222 ]
```

```
print(y_fit_scaled.params)
```

```
## [8.38536209 0.5099222 ]
```

Finally, if we scale Y by one constant and X by a different constant:

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

In other words, if Y is multiplied by a constant a and X is multiplied by a constant c, then the OLS estimates for the intercept change to $\widetilde{\beta}_0 = a \cdot \beta_0$ but for the slope they change to $\widetilde{\beta}_1 = \frac{a}{c} \cdot \beta_1$.

```r
const_a <- 5
const_c <- 10
x1 <- x * const_c
new_y <- y * const_a
#
y_fit_scaled <- lm(new_y ~ x1)
print(const_a / const_c)
```

```
## [1] 0.5
```

```r
print(y_fit$coefficients)
```

```
## (Intercept)           x
##   0.8337401   0.5047895
```

```r
print(y_fit_scaled$coefficients)
```

```
## (Intercept)          x1
##   4.1687005   0.2523947
```

```python
const_a = 5
const_c = 10
x1_mat = sm.add_constant(x * const_c)
new_y = y * const_a
#
y_fit_scaled = sm.OLS(new_y, x1_mat).fit
print(str(const_a / const_c))
```

```
## 0.5
```

```python
print(y_fit.params)
```

```
## [0.83853621 0.5099222 ]
```

```python
print(y_fit_scaled.params)
```

```
## [4.19268105 0.2549611 ]
```

Usually, changing the unit of measurement is referred to as *data scaling*. In all cases after scaling the data the standard errors of the scaled regression coefficients change as well, however, as we will later see, this does not effect any test statistics related to the coefficients, or the model accuracy.

In multiple regression analysis this means that we can include variables with different measurements in the same regression and it will not affect the accuracy (in terms of standard errors) of our model - e.g. Y is measured in thousands, X_1 is measured in thousands and X_2 is measured in single units (or millions, etc.).

Interpretation of the Parameters

In our univariate regression:

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

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**. This can be verified by specifying two cases - with $X = x$ and $X = x + 1$:

$$\mathbb{E}(Y|X = x) = \widetilde{Y} = \beta_0 + \beta_1 x$$
$$\mathbb{E}(Y|X = x + 1) = \widetilde{\widetilde{Y}} = \beta_0 + \beta_1 (x + 1)$$

then, taking the difference yields:

$$\widetilde{\widetilde{Y}} - \widetilde{Y} = \beta_0 + \beta_1 (x + 1) - \beta_0 - \beta_1 x = \beta_1$$

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*. As mentioned previously, interpreting the intercept β_0 is tricky.

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:

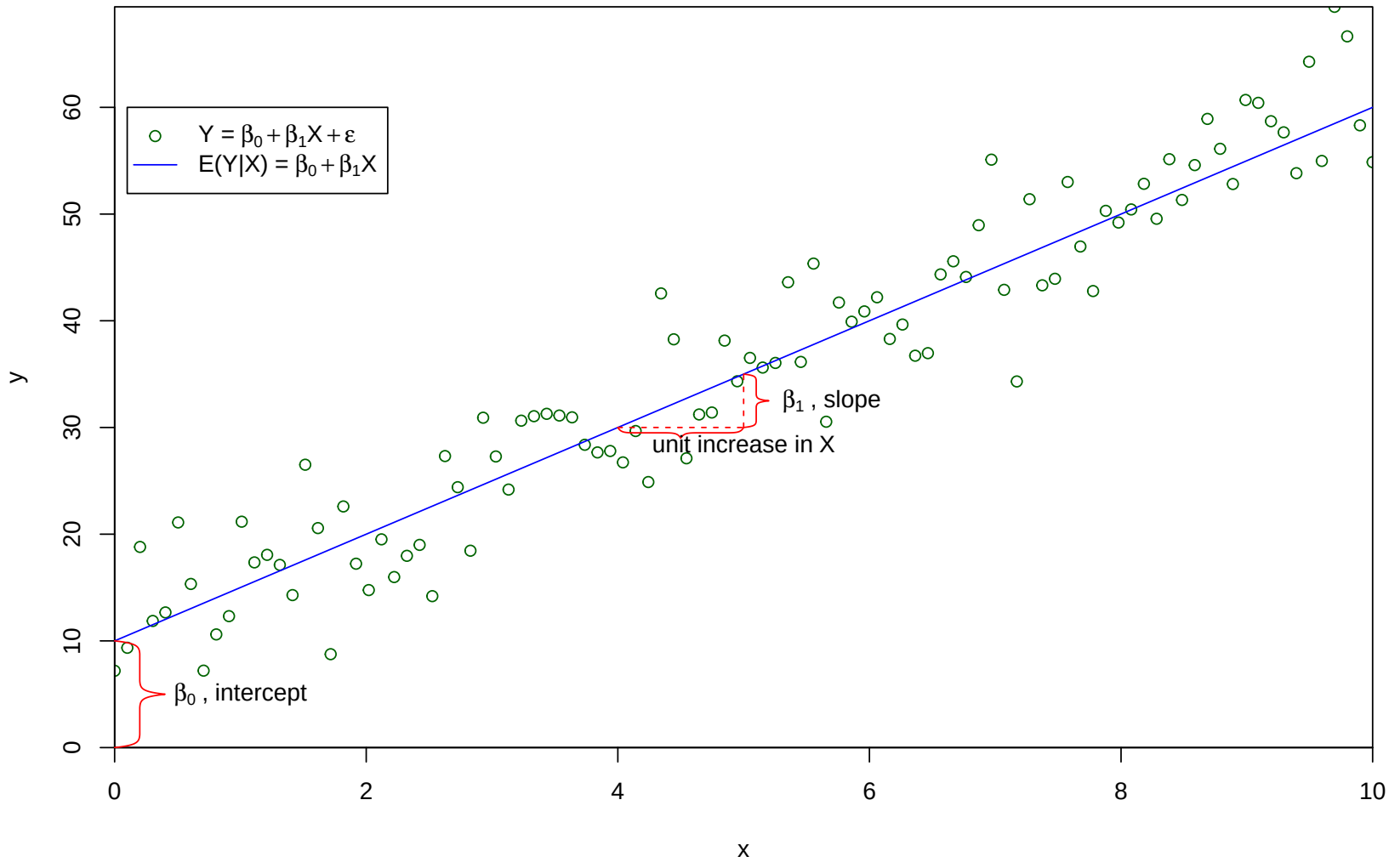
$$\Delta Y = \beta_1 \Delta X$$

or alternatively:

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

where d denotes the derivative of the expected value of Y with respect to X. As we can see, **in the linear regression case, the derivative is simply the slope of the regression line, β_1.**
Note that in this case we say that the **marginal effect of X on Y is constant**, because a one-unit change in X results in *the same* change in Y, **regardless of the initial value of X.**

Below we provide a graphical interpretation with $\beta_0 = 10$, $\beta_1 = 5$, $\epsilon \sim \mathcal{N}(0, 5^2)$ and X from an evenly spaced set between 0 and 10, with $N = 100$:



In econometrics, an even more popular characteristic is the rate of change. The **proportionate (or relative) change** of Y, moving from Y, to $\widetilde{Y}$ is defined as dividing the change in Y by its initial value:

$$\frac{\widetilde{Y} - Y}{Y} = \frac{\Delta Y}{Y}, \; Y \neq 0$$

Usually, we measure changes in terms of percentages - a **percentage change** in Y, from Y to $\widetilde{Y}$ is defined as the proportionate change multiplied by 100:

$$\%\Delta Y = 100 \cdot \frac{\Delta Y}{Y} \approx 100 \cdot \Delta \log(Y)$$

Note: **the approximate equality to the logarithm is useful for interpreting the coefficients when modelling** $\log(Y)$**, instead of** Y**.**

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*:

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

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

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

In the *univariate regression case* we have that:

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

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, i.e.:

$$\widehat{\eta} = \widehat{\beta}_2 \cdot \frac{\overline{X}}{\overline{Y}}$$

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.

To reiterate - η shows a **percentage** and **not a unit** change in Y corresponding to a 1% change in X.

- If elasticity is *less than one*, we can classify that Y is *inelastic* with respect to X.
- If elasticity is *greater than one*, then we would say that Y is *elastic* with respect to X.

Nonlinearities in a Linear Regression

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.

- For example, estimating how the price of a house (Y) relates to the size of the house (X): in a linear specification, the expected price for an additional square foot is constant. However, it is possible that more expensive homes have a higher value for each additional square foot, compared to smaller, less expensive homes.

Fortunately, the linear regression model $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ is quite flexible - the variables $\mathbf{Y}$ and $\mathbf{X}$ can be transformed via:

- logarithms,
- squares,
- cubes,
- creating so called *indicator variables*.

All of these 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 β).

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:

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

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 $U - shaped$; if $\beta_1 < 0$, then the curve is $inverted - U - shaped$.
Because of the nonlinearity in X, our unit change in X effect on Y now depends on the initial value of X. If, as before, we take $X = x$ and $X = x + 1$:

$$\begin{aligned} \mathbb{E}(Y|X = x) = \widetilde{Y} &= \beta_0 + \beta_1 x^2 \\ \mathbb{E}(Y|X = x+1) = \widetilde{\widetilde{Y}} &= \beta_0 + \beta_1 (x+1)^2 \\ &= \beta_0 + \beta_1 x^2 + \beta_1 \cdot (2x+1) \end{aligned}$$

so, the difference:

$$\widetilde{\widetilde{Y}} - \widetilde{Y} = \beta_1 \cdot (2x + 1)$$

now depends on the initial value of x - the larger the initial value, the more *pronounced* the change in Y will be.

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**.

The elasticity (i.e. the percentage change in Y, given a 1% change in X) is:

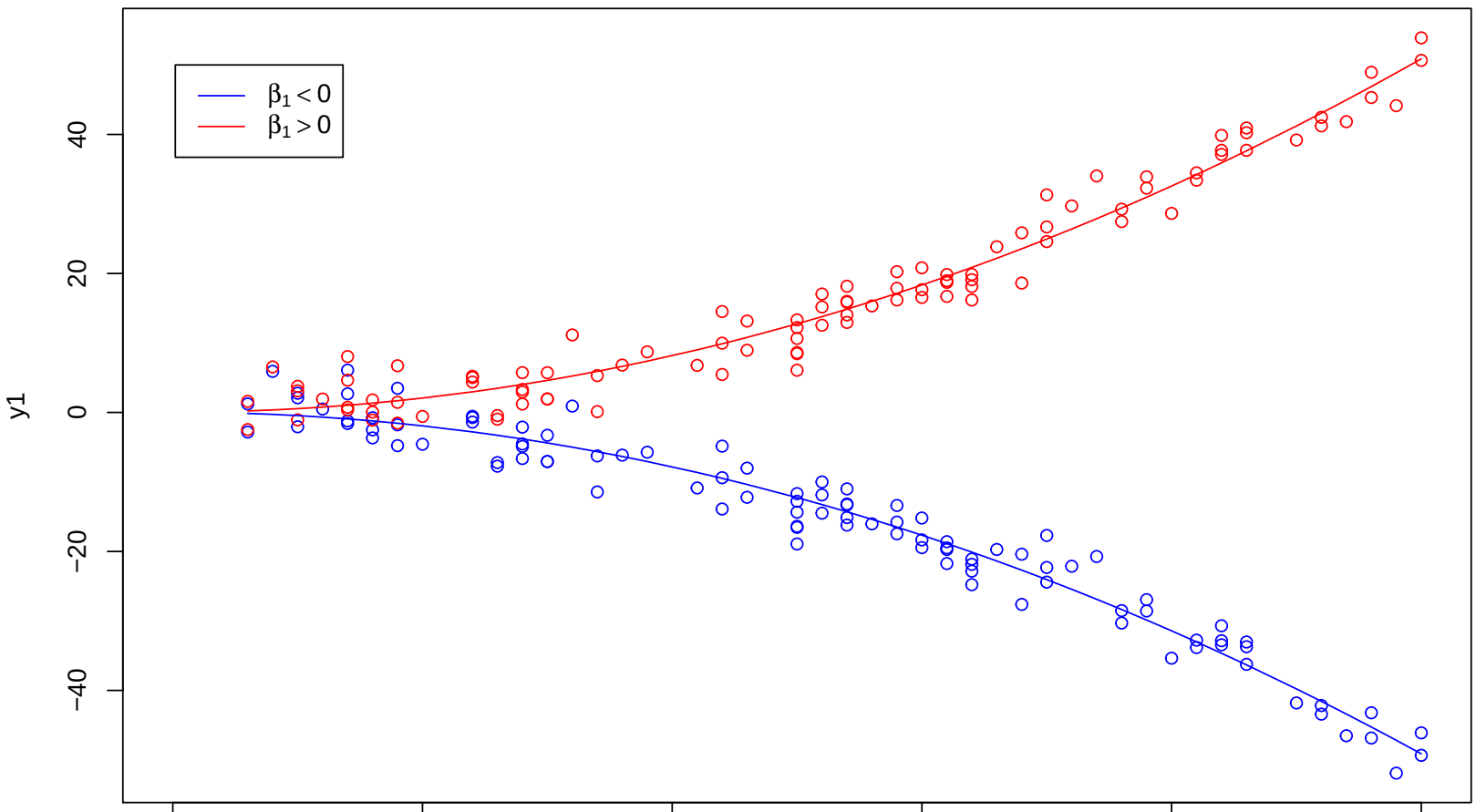
$$\eta = 2\beta_1 X \cdot \frac{X}{Y} = \frac{2\beta_1 X^2}{Y}$$

A common approach is to choose a point on the fitted relationship, i.e. select a value of X and the corresponding fitted value $\widehat{Y}$ to estimate $\widehat{\eta}(\widehat{Y}|X) = 2\widehat{\beta}_1 X^2/\widehat{Y}$. Note that we could use $(\overline{Y}, \overline{X})$ since it is on the regression curve.

However, we may be interested to see how the elasticity changes at different value of X - when X is small; when X is large etc.

Below we will assume that our data generating process satisfies **(UR.1)** - **(UR.4)** assumptions with the following parameters:

- $\beta_0 = 1$, $\beta_1 \in \{-0.02, 0.02\}$;
- X is a random sample with replacement from a set: $X_i \in \{1, ..., 50\}$, $i = 1, ..., 100$;
- $\epsilon \sim \mathcal{N}(0, 5^2)$.



```r
set.seed(123)
# Set the coefficients:
N = 100
beta_0 = 0.5
beta_1 = c(-0.02, 0.02)
# Generate sample data:
x <- sample(1:50, size = N,
            replace = TRUE)
e <- rnorm(mean = 0, sd = 3,
           n = length(x))
y1 <- beta_0 + beta_1[1] * x^2 + e
y2 <- beta_0 + beta_1[2] * x^2 + e
```

```r
print(coef(lm(y1 ~ I(x^2))))
```

```
## (Intercept)      I(x^2)
##  0.03470760 -0.01966276
```

```r
print(coef(lm(y2 ~ I(x^2))))
```

```
## (Intercept)      I(x^2)
##  0.03470760  0.02033724
```

```python
np.random.seed(123)
# Set the coefficients:
N = 100
beta_0 = 0.5
beta_1 = [-0.02, 0.02]
# Generate sample data:
x = np.random.choice(list(range(1, 51)),
        size = N, replace = True)
e = np.random.normal(loc = 0, scale = 3,
        size = len(x))
y1 = beta_0 + beta_1[0] * x**2 + e
y2 = beta_0 + beta_1[1] * x**2 + e
```

```python
print(sm.OLS(y1, sm.add_constant(x**2)).
```

```
## [ 0.27578417 -0.01973092]
```

```python
print(sm.OLS(y2, sm.add_constant(x**2)).
```

```
## [0.27578417 0.02026908]
```

Log-Linear Regression Model

A couple of useful properties of the logarithm function, which are frequently applied to simplify some non-linear model specifications and various approximations in econometric analysis:

If $X > 0$ is **small** (e.g. $X = 0.01, 0.02, ..., 0.1$), then:

$$\log(1 + X) \approx X$$

The equality deteriorates as X gets larger (e.g. $X = 0.5$).
For **small changes** in X it can be shown that:

$$\Delta \log(X) = \log(X_1) - \log(X_0) \approx \frac{X_1 - X_0}{X_0} = \frac{\Delta X}{X_0}, \quad \text{where } \Delta X = X_1 - X_0$$

A **percentage change** in X, from X_0 to X_1 is defined as the log difference multiplied by 100:

$$\%\Delta X \approx 100 \cdot \Delta \log(X)$$

Log-Linear Regression Model

Often times, the dependent and/or independent variable may appear in logarithmic form. The **log-linear** model has a logarithmic term on the left-hand side of the equation and an untransformed variable on the right-hand side:

$$\log(Y) = \beta_0 + \beta_1 X + \epsilon$$

In order to use this model we must have that $Y > 0$.
Furthermore, we may also sometimes want to take the logarithm in order to **linearize** Y.
If Y is defined via the following exponential form:

$$Y = \exp(\beta_0 + \beta_1 X + \epsilon) \tag{1}$$

Then we can take the logarithm of Y to get the log-linear expression

$$\log(Y) = \beta_0 + \beta_1 X + \epsilon$$

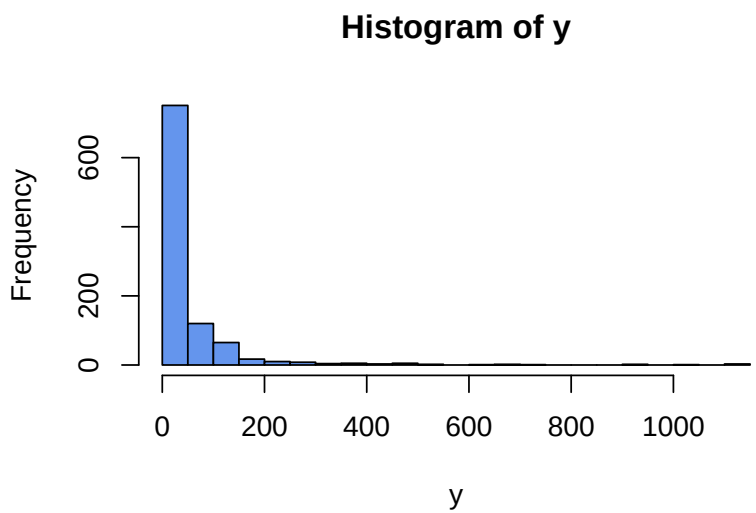
From eq. (1) we can see that we can use the log-transformation to regularize the data (i.e. to make highly skewed distributions less skewed).

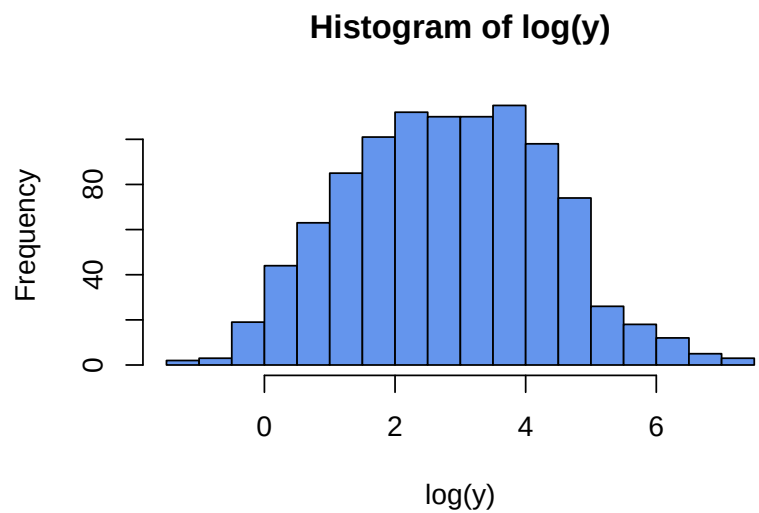
For example, the histogram of (1) with $\beta_0 = 0.8$, $\beta_1 = 4$ and $\epsilon \sim \mathcal{N}(0, 1)$ looks to be **skewed to the right**, because the tail to the right is longer. Whereas the histogram of $\log(Y)$ (i.e. the dependent variable of the log-linear model) appears to be symmetric around the mean and similar to the normal distribution:

```r
set.seed(123)
#
N <- 1000
beta_0 <- 0.8
beta_1 <- 4
x <- seq(from = 0, to = 1,
         length.out = N)
e <- rnorm(mean = 0, sd = 1,
           n = length(x))
y <- exp(beta_0 + beta_1 * x + e)
```

```python
np.random.seed(123)
#
N = 1000
beta_0 = 0.8
beta_1 = 4
x = np.linspace(start = 0,
          stop = 1, num = N)
e = np.random.normal(loc = 0,
          scale = 1, size = len(x))
y = np.exp(beta_0 + beta_1 * x + e)
```

The appropriate histograms of Y and $\log(Y)$:





The extremely skewed distribution of Y became less skewed and more bell-shaped after taking the logarithm.

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

Furthermore, setting $\mathbf{Y} = [\log(Y_1), ..., \log(Y_N)]^\top$ allows us to apply the OLS via the same formulas as we did before:

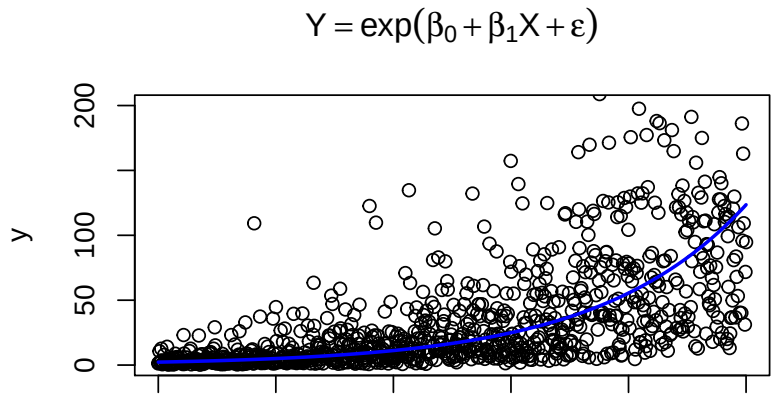
```
#
#
lm_fit <- lm(log(y) ~ x)
print(coef(lm_fit))
```

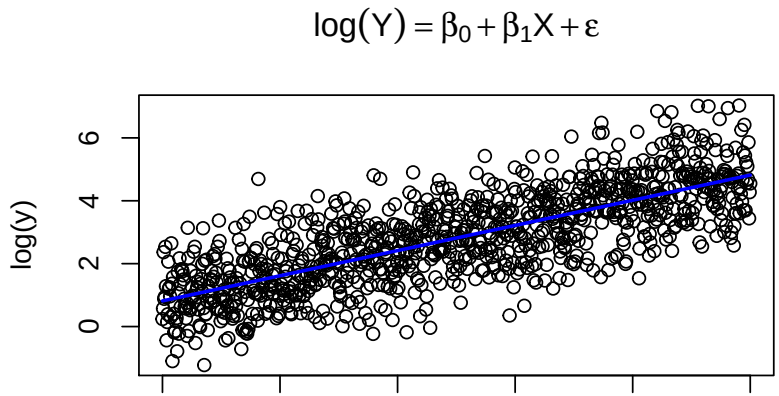
```
## (Intercept)           x
##   0.8156225   4.0010107
```

```
lm_model = sm.OLS(np.log(y),
                  sm.add_constant(x))
lm_fit = lm_model.fit()
print(lm_fit.params)
```

```
## [0.81327991 3.89431192]
```

Furthermore, we can calculate the log-transformed fitted values $\log(\widehat{Y})$, as well as transform them back to $\widehat{Y}$:





Log-Linear Regression Model

To better understand the effect that a change in X has on $\log(Y)$, we calculate the expected value of $\log(Y)$ when X changes from x to $x+1$:

$$\mathbb{E}(\log(Y)|X=x) = \widetilde{Y} = \beta_0 + \beta_1 x$$
$$\mathbb{E}(\log(Y)|X=x+1) = \widetilde{\widetilde{Y}} = \beta_0 + \beta_1(x+1)$$

Then the difference:

$$\widetilde{\widetilde{Y}} - \widetilde{Y} = \beta_1$$

is similar as in the simple linear regression case, but the interpretation is different since we are talking about $\log(Y)$, instead of Y. In other words:

$$\Delta \log(Y) = \beta_1 \Delta X$$
$$100 \cdot \Delta \log(Y) = (100 \cdot \beta_1)\Delta X$$

Because X and Y are related via the log-linear regression, it follows that:

$$\%\Delta Y \approx (100 \cdot \beta_1)\Delta X$$

In other words, for the **log-linear** model, a **unit increase** in X yields (approximately) a $(100 \cdot \beta_1)$ **percentage change** in Y.

We can rewrite the previous equality as:

$$100 \cdot \beta_1 = \frac{\%\Delta Y}{\Delta X} := \text{semi-elasticity}$$

This quantity, known the **semi-elasticity**, is the **percentage** change in Y when X *increases by one unit*. As we have just shown, in the log-linear regression, the semi-elasticity is **constant** and equal to $100 \cdot \beta_1$.

Furthermore, we can derive the same measurements as before:

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

unlike the simple linear regression model, **in the log-linear regression model, the marginal effect increases for larger values of** Y (note, that this is **not** $\log(Y)$).

The **elasticity** (the *percentage change* in Y, given a 1% increase in X):

$$\eta = \text{slope} \cdot \frac{X}{Y} = \beta_1 X$$

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

Alternatively, we may describe the linear relationship, where X is log-transformed and Y is untransformed:

$$Y = \beta_0 + \beta_1 \log(X) + \epsilon$$

where $X > 0$. In this case, if X increases from X_0 to X_1, it follows that:

$$Y_0 = \mathbb{E}(Y|X = X_0) = \beta_0 + \beta_1 \log(X_0)$$
$$Y_1 = \mathbb{E}(Y|X = X_1) = \beta_0 + \beta_1 \log(X_1)$$

Setting $\Delta Y = Y_1 - Y_0$ and $\Delta \log(X) = \log(X_1) - \log(X_0)$, yields:

$$\Delta Y = \beta_1 \Delta \log(X) = \frac{\beta_1}{100} [100 \cdot \Delta \log(X)] \approx \frac{\beta_1}{100} (\%\Delta X)$$

In other words, in a **linear-log** model, a 1% increase in X yields (approximately) a $\beta_1/100$ **unit change** in Y.

Furthermore, we have that:

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

and:

$$\eta = \text{slope} \cdot \frac{X}{Y} = \beta_1 \frac{1}{Y}$$

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:

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

Sometimes linear-log is not much different (in terms of model accuracy) from a linear-linear (i.e. simple) regression.

However, because of the functional form of the linear-log model, if $\beta_1 < 0$, then the function decreases at a **decreased** rate, as X increases.

This means that it may sometimes be useful if we do not want Y to have a negative value, for a *reasonably* large value of X.

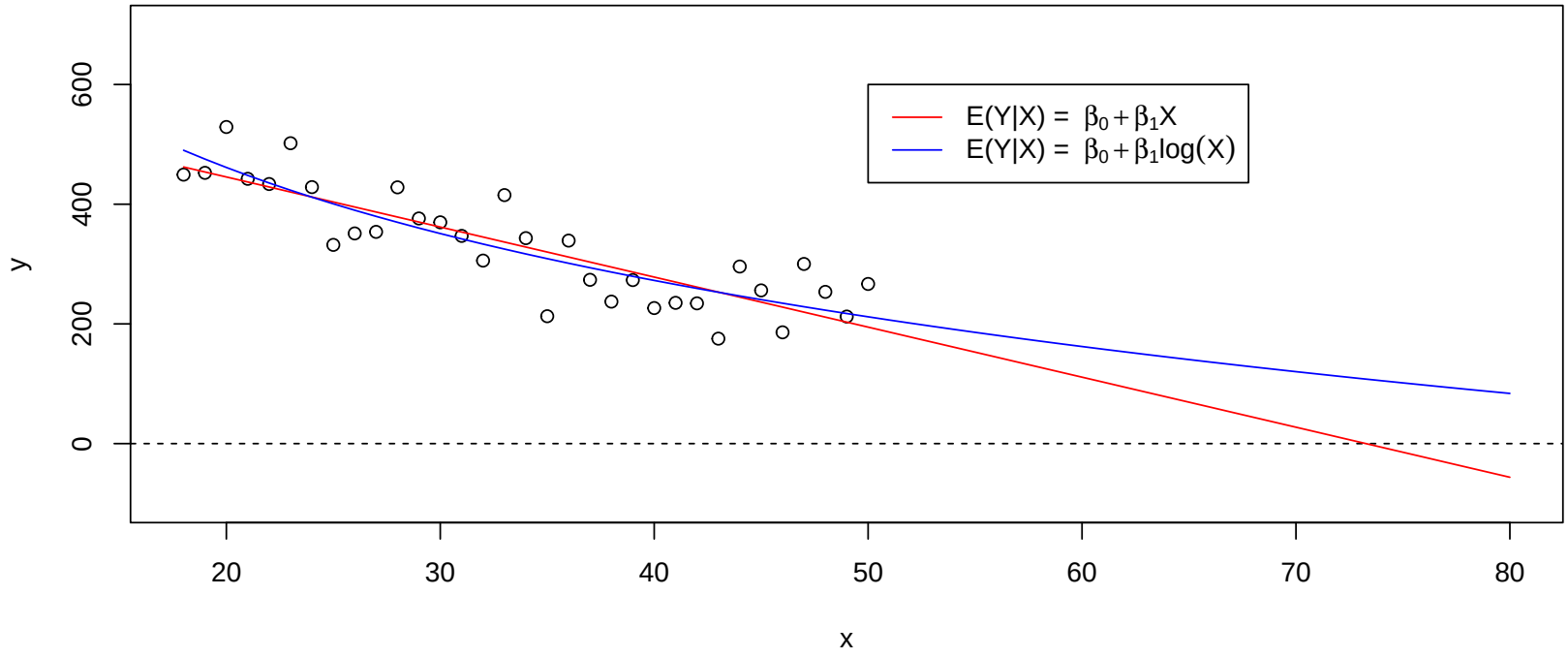
For example, let Y be expenditure on leisure activities, and X - age. Let us say that we only have expenditure data on ages from 18 to 50 and we would like to **predict** the expenditure for ages 51 to 80.

In this case it is reasonable to assume, that **expenditure cannot be negative** for **reasonable** values of X - an age of up to 80 may be realistic assumption, while 100 years or more - less so.

Further assume that the underlying (true) model is indeed linear-log:

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

with $\beta_0 = 1200$, $\beta_1 = -250$ and $\epsilon \sim \mathcal{N}(0, 50^2)$.



We see that the fitted values $\widehat{Y}$ do not differ much in the linear-linear and linear-log cases, however, the predicted values (since in this example Y is the *expenditure*) are more believable in the linear-log case, since they are non-negative.

Log-Log Regression Model

Elasticities are often important in applied economics. As such, it is sometimes convenient to have **constant elasticity** models. If we take logarithms of both Y and X, then we arrive at the **log-log** model:

$$\log(Y) = \beta_0 + \beta_1 \log(X) + \epsilon$$

where $X, Y > 0$. Then:

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

and:

$$\eta = \text{slope} \cdot \frac{X}{Y} = \beta_1$$

i.e. **the elasticity of the log-log model is constant**.

In other words, in a **log-log** model, a 1% increase in X yields (approximately) a β_1 **percentage change** in Y.

Choosing a Functional Form

Generally, we will **never** know the true relationship between Y and X. The functional form that we select is only an **approximation**.

As such, we should select a functional form, which satisfies **our objectives**, **economic theory**, the underlying model assumptions, like (UR.1) - (UR.3), and one which provides an *adequate fit* on the data sample.

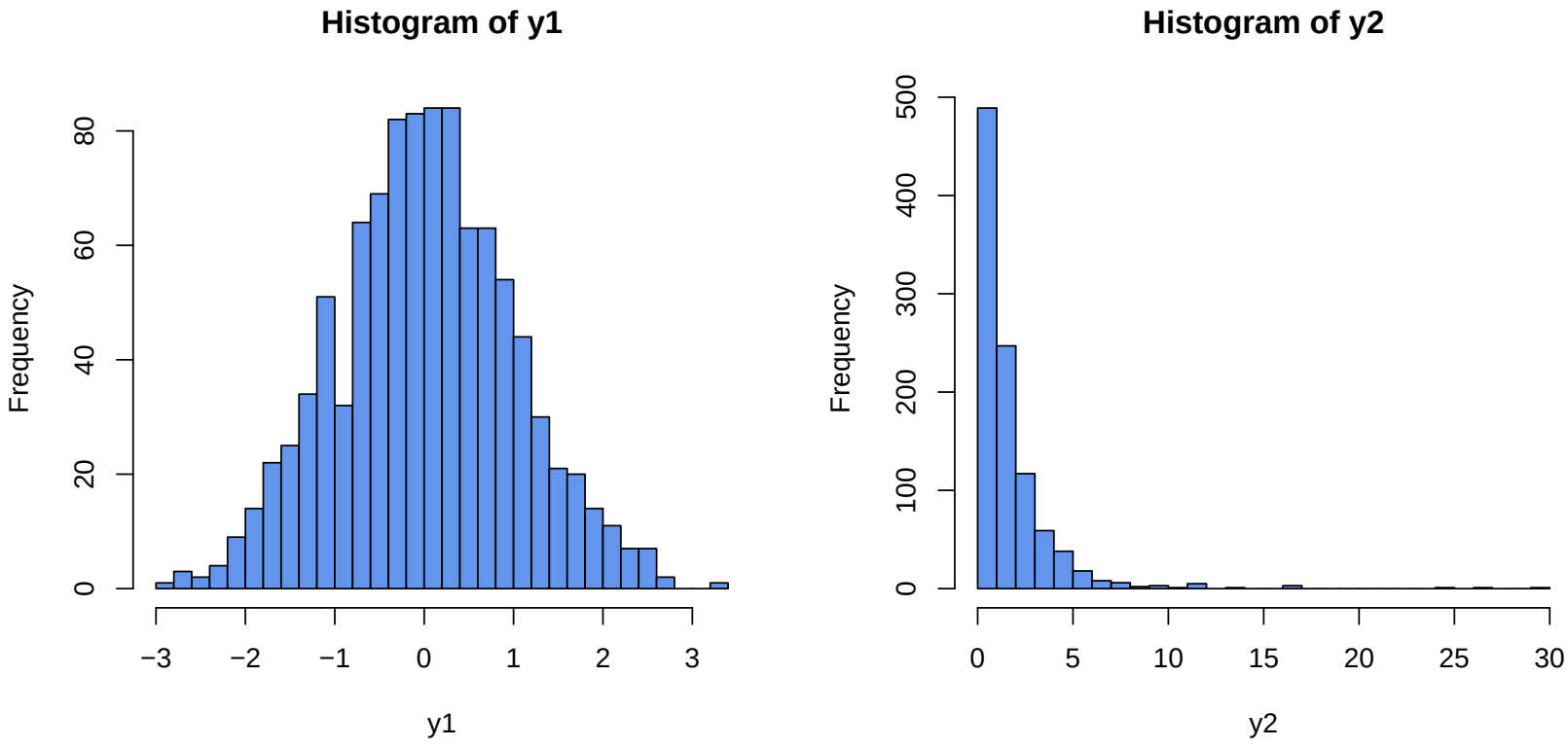
Consequently, we may incorrectly specify a functional form for our regression model, or, some of our regression assumptions may not hold.

It is a good idea to focus on:

1. Examining the regression results:
   - checking whether the signs of the coefficients follow economic logic;
   - checking the significance of the coefficients;
2. Examining the residuals of $\mathbf{Y} - \widehat{\mathbf{Y}}$ - if our specified functional form is inadequate, then the residuals would not necessarily follow (UR.3), (UR.4).

Nevertheless, before examining the signs and model statistical properties, we should have a *first look* at the data and examine the dependent variable Y and the independent variable X plots to make some initial guesses at the functional form of the regression model.

Histogram of The Response Variable



The histogram can be used to answer the following questions:

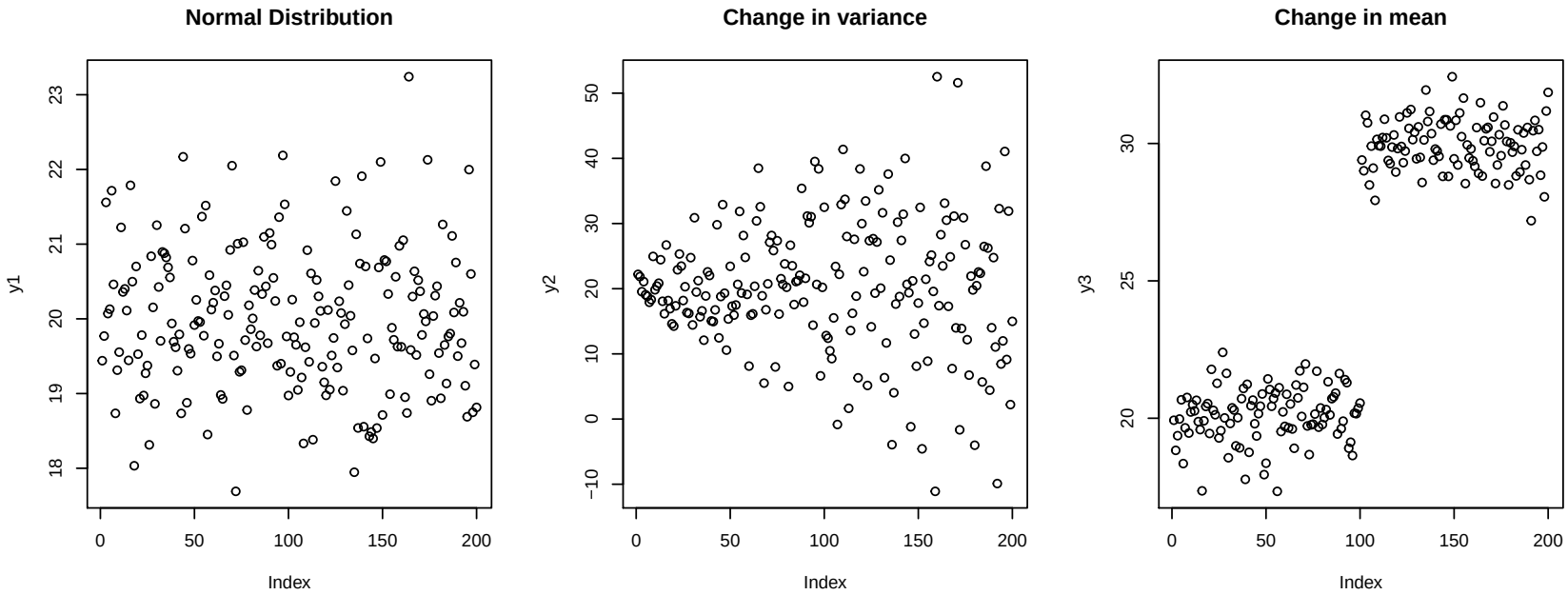
- What kind of population distribution is the data sample most likely from?
- What is the sample mean of the data?
- Is the distribution of the data large?
- Are the data symmetric, or skewed?
- Are there outliers in the data?

Run-Sequence Plot

An easy way to graphically summarize a univariate data set. A common assumption of univariate data sets is that they behave like:

- Random realization of a dataset;
- Are from the same population (i.e. all random drawings have the same distribution);

The run-sequence (or simply, run charts) plots the variable of interest on the vertical axis, and the variable index on the horizontal axis. They are primarily used to inspect if there are any outliers, mean or variance changes or if there is a dependence across observations.



We see that the underlying differences in the distribution of the data are present in the run-sequence plot of the data sample.

The run-sequence plot can help answer questions, like:

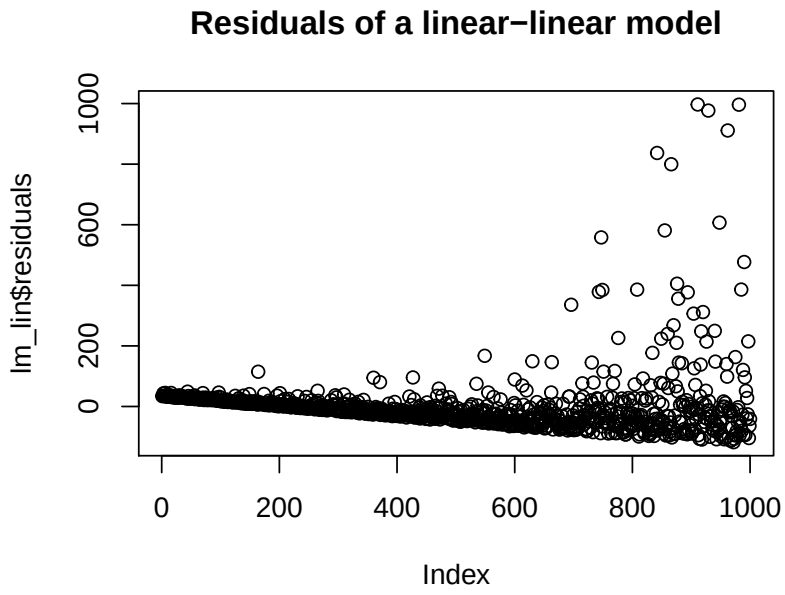
- Are there any (significant) changes in the mean?
- Are there any (significant) changes in the variance?
- Are there any outliers in the data?

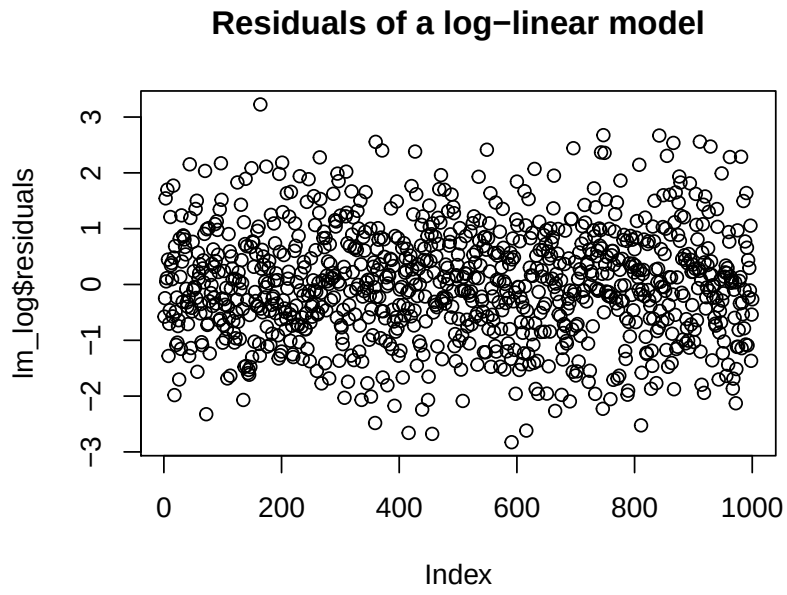
The run-sequence plot can be useful when examining the residuals of a model.

In our exponential model example:

$$Y = \exp(\beta_0 + \beta_1 X + \epsilon)$$

what would the residuals look like if we fitted a **linear-linear** and a **log-linear** model?

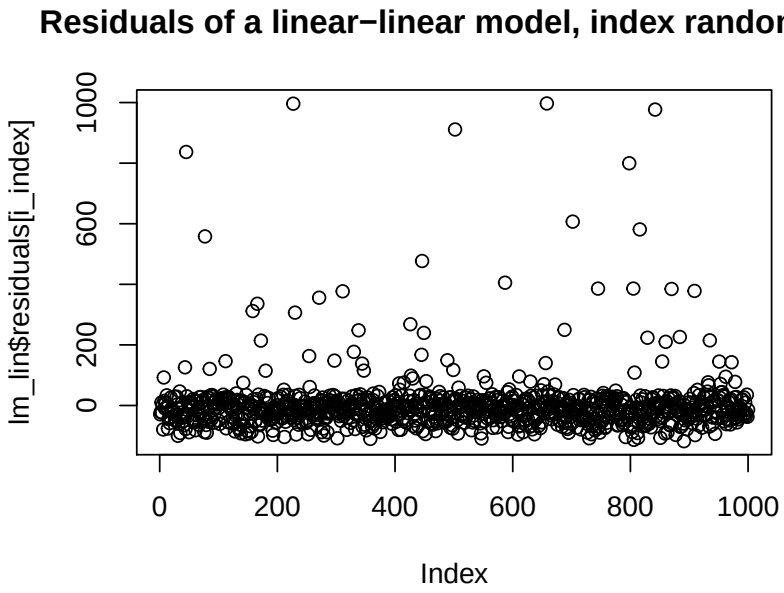


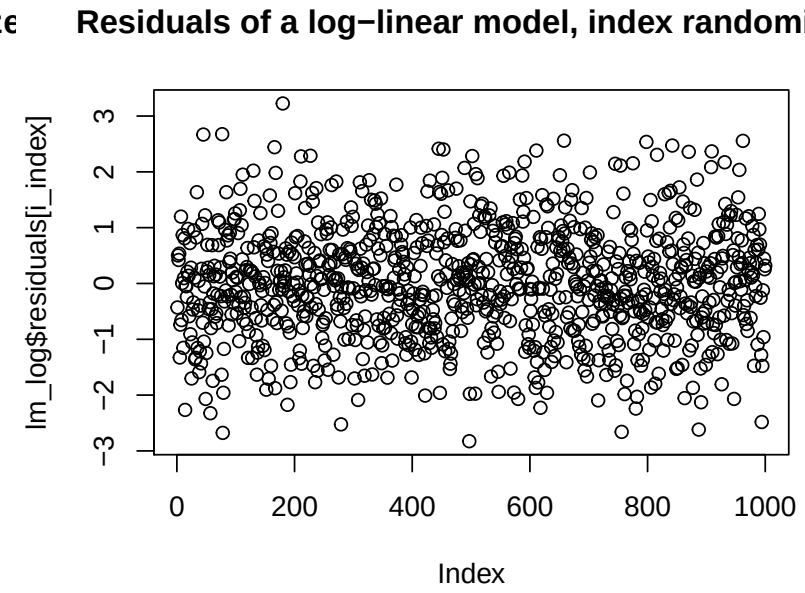


We see that in the case of a linear linear model, the residual variance is not the same across observations, while the log-linear model has residuals that appear to have the same mean and variance across observations.

We note that the values of X are **ordered** from smallest to largest - this means that a larger index value corresponds to the value of Y which was for a larger value of X.

If we were to **randomize** the order of X_i (and as a result, randomize Y_i and $\widehat{Y}_i$), the run-sequence plot of the residuals would then have the following plot:





Note the fact that in cross-sectional data one of our assumptions is that the observations (X_i, Y_i) are treated as independent from (X_j, Y_j) for $i \neq j$. As such we can order the data in any way we want.

- In the first case (with ordered values of X from (X_i, Y_i)), ordering the residuals by the value of X allows us to examine whether our model performs the same for all values of X. Clearly, for larger values of X this was not the case (Note: a **scatter plot** would be more useful in such a case).
- In the second case (where the ordering of X from (X_i, Y_i) is random), where we shuffled the order of our data (X_i, Y_i), the residual run-sequence plot allows us to **identify**, whether there are residuals, which are more pronounced, but it does not identify when this happens, unlike the first case, with ordering by values of X.

Run sequence plots are more common in time series data, but can still be utilized for cross-sectional data.

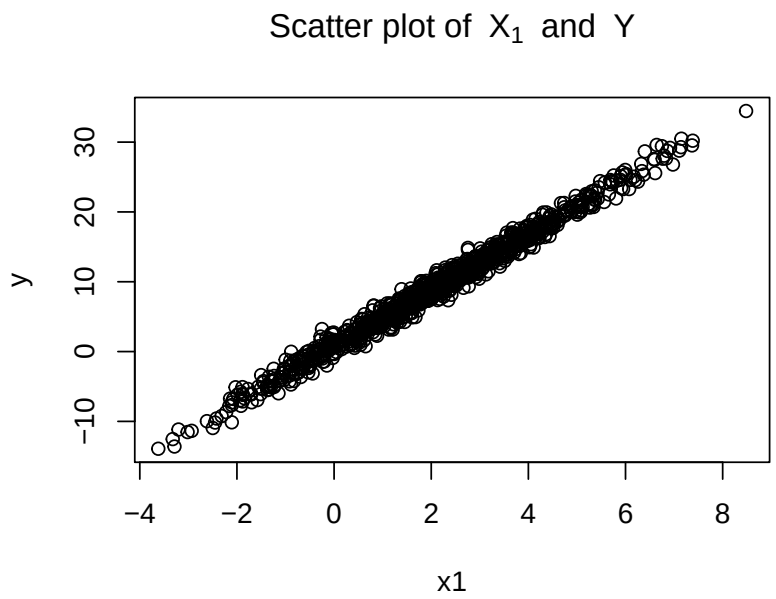
Scatter Plot

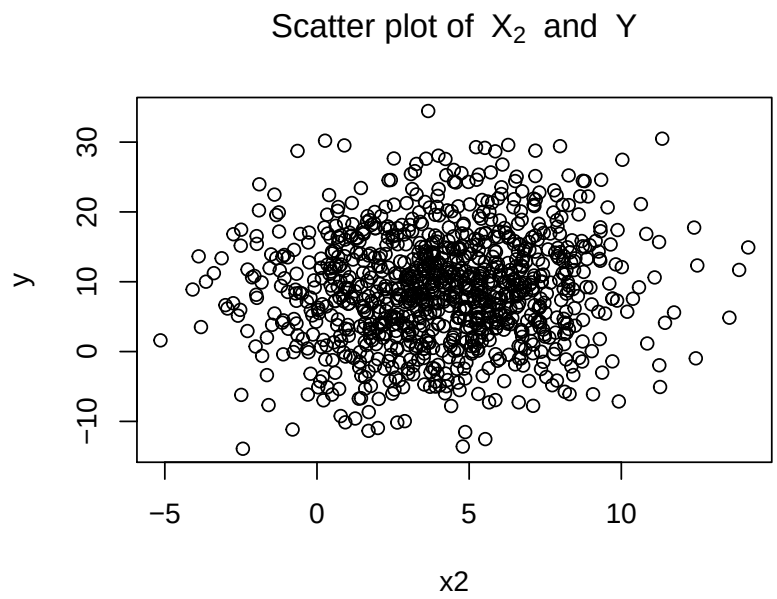
As we have seen, a scatter plot reveals relationships between two variables or interest in a form of *non-random structure*. the vertical axis is usually for the dependent (i.e. response) variable Y, and the horizontal axis is for the independent variable X.

Let:

$$Y = \beta_0 + \beta_1 \cdot X_1 + \epsilon$$

where $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $\epsilon \sim \mathcal{N}(0, 1^2)$ and let $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ and X_2 is independent of X_1 and ϵ.





We see a clear linear relationship between X_1 and Y. The scatter plot of X_2 and Y appears to be random, i.e. Y does not depend on the value of X_2.

The scatter plot can help us answer the following questions:

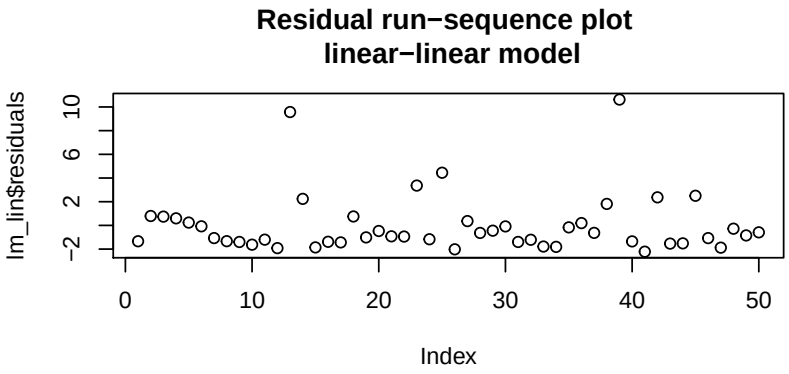
- Are the variables X and Y related?
- What kind of relationship (linear, exponential, etc.) could describe Y and X?
- Are there any outliers?
- Does the variation in Y change depending on the value of X?

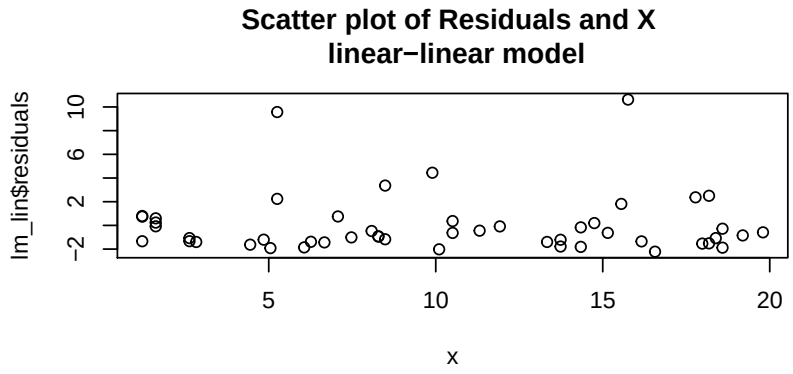
In addition to the run-sequence plot, the scatter plot for the **residuals** $\widehat{\epsilon}$ and the **independent variable** X can be useful to determine, whether our model performs the same across any value of X, or if there are specific values of X, for which our model is worse (which would result in larger residual values). **If the values of X are ordered in ascending order, then the run-sequence plot will be very similar to the scatter plot of $\widehat{\epsilon}$ and X, with the only difference being the spacing between the points**.

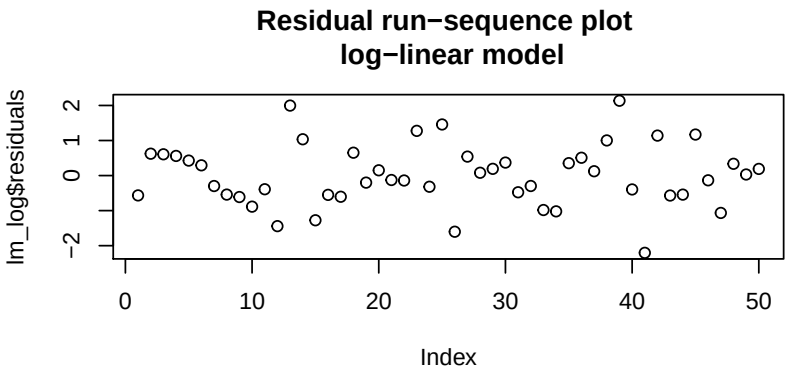
In our exponential model example:

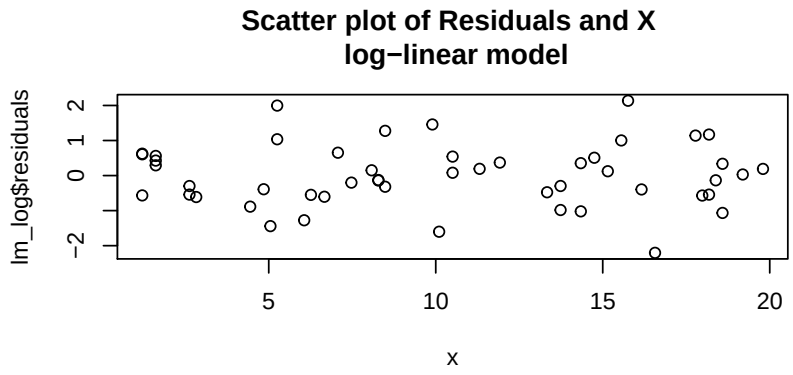
$$Y = \exp(\beta_0 + \beta_1 X + \epsilon)$$

what would the residuals look like if we fitted a **linear-linear** and a **log-linear** model?









Notice that **when X is ordered** the only difference between the run-sequence plot and the scatter plot is in the spacing between the points.

Scatter plots are similar to run sequence plots for **univariate** cross-sectional data. However, unlike run-sequence plots, scatter plots allow to examine the relationship between X and Y (instead of only examining Y).
For multivariable cross-sectional data, run-sequence plots may be faster to utilize to focus on the properties of Y variable itself.

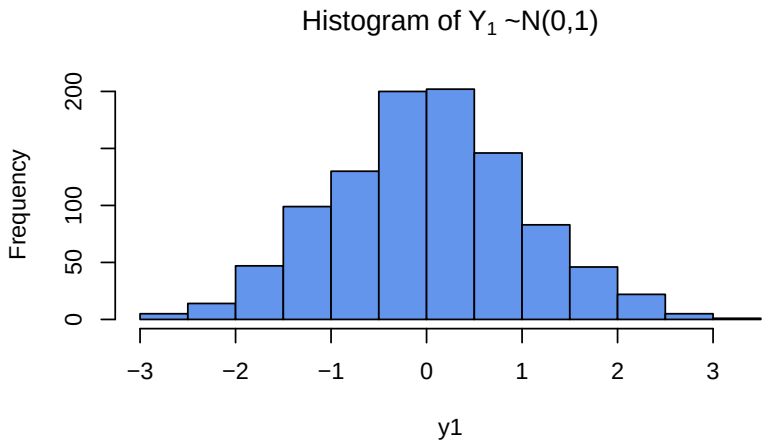
Quantile-Quantile Plot

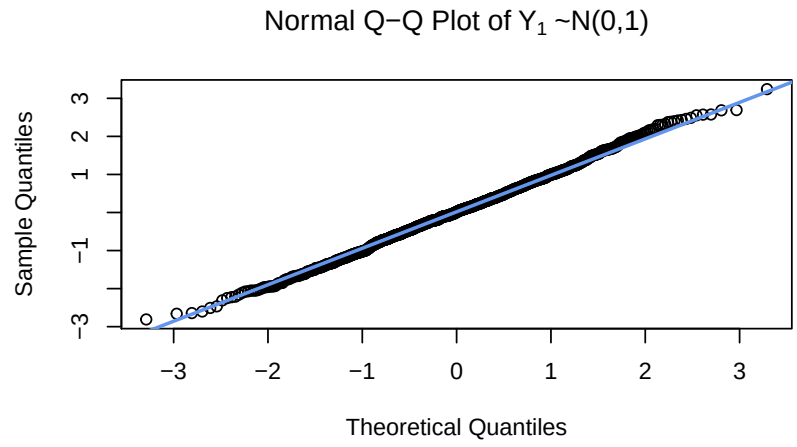
The **quantile-quantile (q-q) plot** can be used to determine if two data sets come from a common distribution. A 45-degree line is also plotted for easier comparison - if the two sets come from the same distribution, the points should fall approximately along this reference line. The greater the departure from this line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions. More specifically, the *quantiles* of the first dataset are plotted against the *quantiles* of the second dataset.

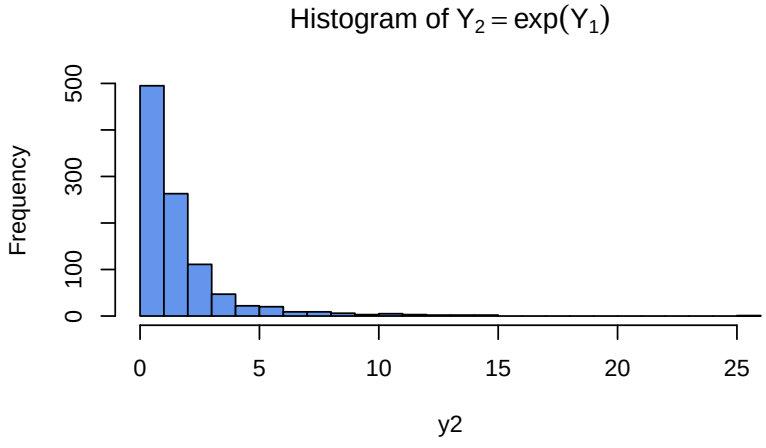
The **quantile-quantile (q-q, or Q-Q) plot** is a scatterplot created by plotting two sets of quantiles against one another.

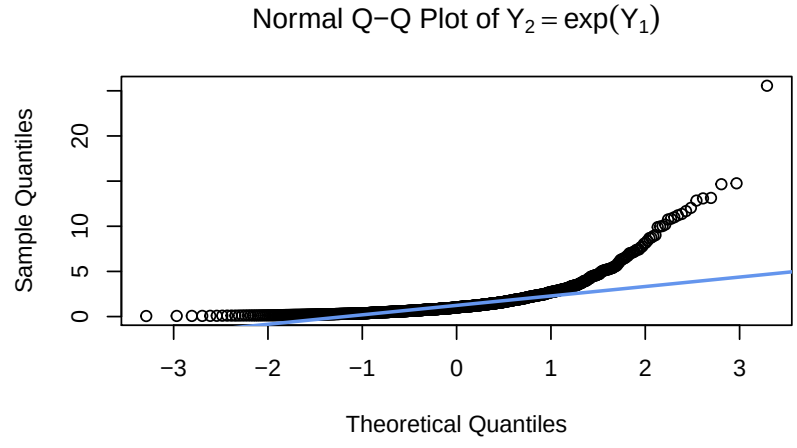
The q-q plot is similar to a **probability plot**, which can be used to examine, whether the model *residuals* are **normally distributed**. For a **probability plot**, the q-q plot is used, with quantiles for one of the data samples are replaced with the quantiles of a **theoretical distribution**.

In a **normal probability plot** the data are plotted against a theoretical (standard) normal distribution in such a way that the points should form an approximate straight line.
Departures from this straight line indicate departures from normality.
The normal distribution is a *base* distribution, and its quantiles are on the horizontal axis as the *Theoretical Quantiles*, while the sample data quantiles are on the vertical axis.









We see that the q-q plot of Y_1 shows that Y_1 is normally distributed, since all the quantiles are on the diagonal line. On the other hand, the q-q plot of Y_2 shows that the data is skewed.

The probability plot is used to answer the following questions:

- Does a specified theoretical distribution provide a good fit to my data?
- What distribution best fits my data?
- What are good estimates for the location and scale parameters of the chosen distribution?

In addition, the normal probability plot answers the following questions:

- Is the data normally distributed?
- What is the nature of the departure from normality (data skewed with shorter/longer tails)?

As mentioned, probability plots can be used to inspect whether the residuals are normally distributed. We will see an example of this in the residual diagnostics section later on.

Notes on the terminology of quantiles, percentiles and quartiles:

- **Quantiles** can go from 0 to any value. Quantiles are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample in the same way. The **p-quantile** is defined as the value, which includes $p \cdot N$ observations, with $0 \leq p \leq 1$ and N being the number of observations.
- **Percentiles** go from 0 to 100. It is a measure used for indicating the value below which a given percentage of observations in a group of observations fall. For example, the 20th percentile is the value below which 20% of the observations may be found.
- **Quartiles** go from 0 to 4. They are values that divide a list of sorted values into quarters.

In general, percentiles and quartiles are specific types of quantiles. The relationship is as follows:

- 0 quartile = 0 quantile = 0 percentile
- 1 quartile = 0.25 quantile = 25 percentile
- 2 quartile = 0.5 quantile = 50 percentile (median)
- 3 quartile = 0.75 quantile = 75 percentile
- 4 quartile = 1 quantile = 100 percentile

- We have reviewed the OLS properties;
- We have examined various variables transformations;
- We have examine ways to specify non-linearities while retaining the linear-regression model form.
- We have examined coefficient interpretations, depending on the model specification.
- 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 2` from `Ch 3.10`. See `Ch. 3.11` for an example.