Mathematical Statistics

MAS 713

Chapter 9.1

Previous lecture:

1. Tests of Hypotheses
2. Confidence Intervals vs. Hypothesis Tests
3. Likelihood Ratio Test
4. Neyman-Pearson Lemma

Any questions?

This lecture

9.1. Regression Analysis

- 9.1.1 Introduction
- 9.1.2 Simple Linear Regression
- 9.1.3 Least Squares Estimators
- 9.1.4 Inferences in simple linear regression
- 9.1.5 Prediction of new observations
- 9.1.6 Adequacy of the regression model
- 9.1.7 Correlation

Additional reading : Chapter 11-12 in the textbook

Introduction

The main objective of many statistical investigations is to **make predictions**, preferably on the basis of mathematical equations

Introduction

Usually, such predictions require that a **formula** be found which relates the dependent variable whose value we want to predict (usually it is called the **response**) to one or more other variables, usually called **predictors** (or regressors)

The collection of statistical tools that are used to model and explore relationships between variables that are related is called **regression analysis**, and is one of the most widely used statistical technique

Definition (Regression)

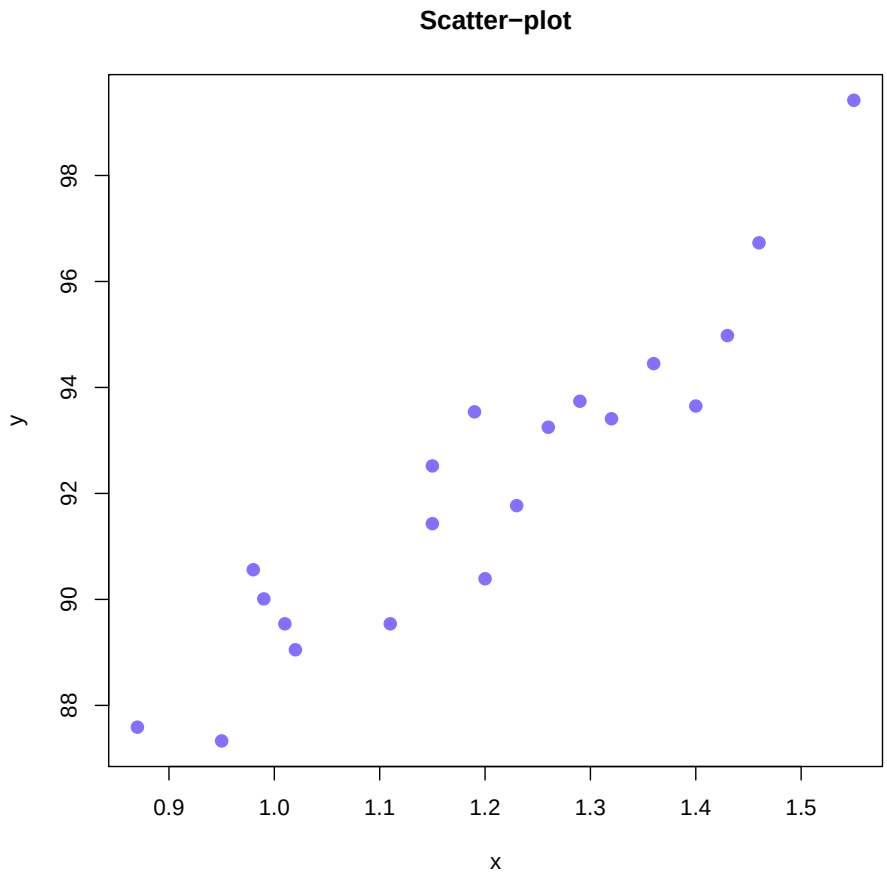
The relation between selected values of x and observed values of y, from which the most probable value of y can be predicted for any value of x.

As an illustration, consider the following data, where y_i's are the observed values of some process y, and x_i's are the observed corresponding values of another process, x

i	x_i	y_i
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.54
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

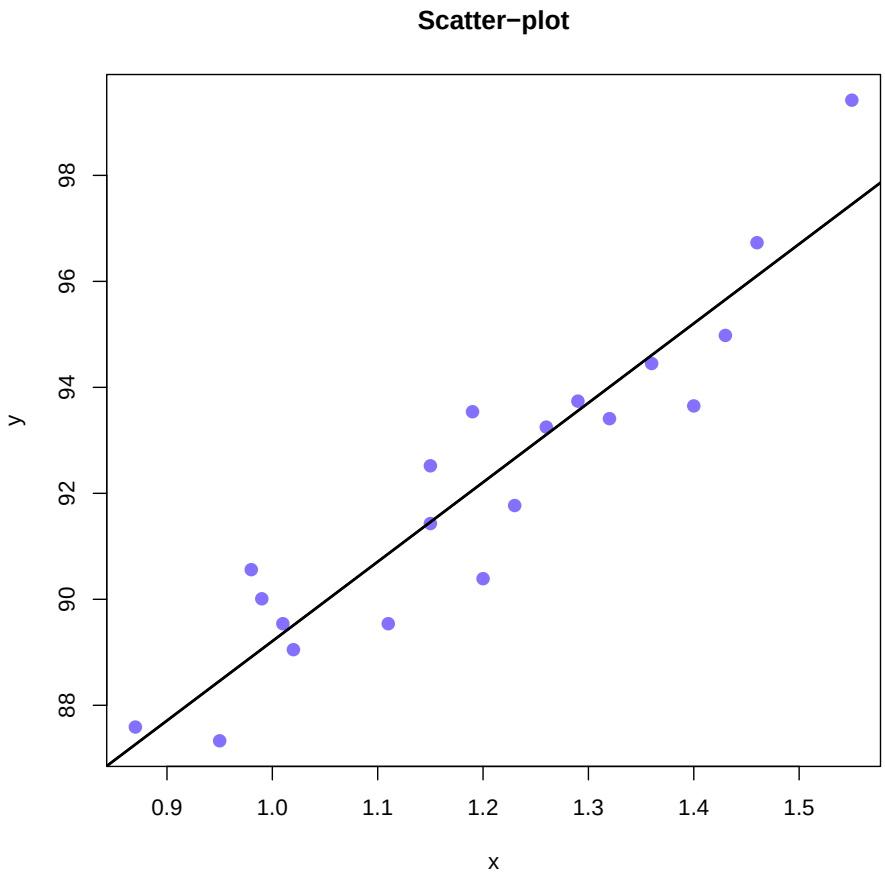
As an illustration, consider the following data, where y_i's are the observed values of some process y, and x_i's are the observed corresponding values of another process, x

i	x_i	y_i
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.54
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33



As an illustration, consider the following data, where y_i's are the observed values of some process y, and x_i's are the observed corresponding values of another process, x

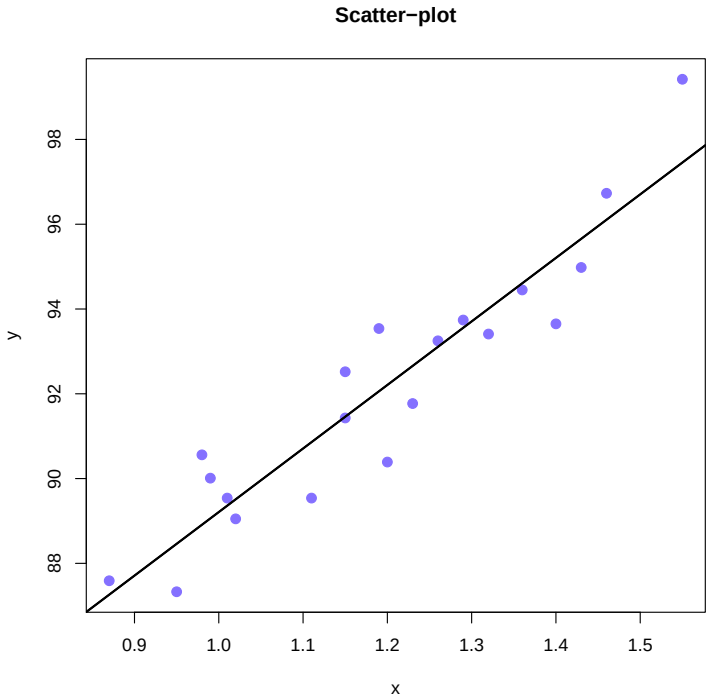
i	x_i	y_i
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.54
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33



Simple Linear Regression

Simple linear regression model

Inspection of the scatter-plot indicates that, although no simple curve will pass exactly through all the points, there is a strong indication that the points are scattered randomly around a straight line



Simple linear regression model

Therefore, it is probably reasonable to assume that the random variables X and Y are linearly related, which can be formalised by the **regression model**

$$\boxed{Y = \beta_0 + \beta_1 X + \varepsilon}$$

The slope β_1 and the intercept β_0 are called the regression coefficients

The term ε is the random error term, whose presence accounts for the fact that observed values for Y do not fall exactly on a straight line

This model is called the **simple linear regression model**
Sometimes a model like this will arise from a theoretical relationship, at other times the choice of the model is just based on inspection of a scatterplot

Simple linear regression model

The random error term ε is a random variable whose properties will determine the properties of the response Y

- Assume that $\mathbb{E}(\varepsilon) = 0$ (not restrictive) and $\mathbb{V}\text{ar}(\varepsilon) = \sigma^2$
- Suppose that we fix $X = x$:
  $\Longrightarrow$ at this very value of X, Y is the random variable

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

with mean $\beta_0 + \beta_1 x$ and variance $\mathbb{V}\text{ar}(\varepsilon) = \sigma^2$

$\rightsquigarrow$ the linear function $\beta_0 + \beta_1 x$ is thus the function giving the mean value of Y for each possible value x of X

It's called the **regression function** (or regression line) and denoted by

$$\mu_{Y|X=x} = \beta_0 + \beta_1 x$$

$\rightsquigarrow$ the slope β_1 is the change in mean of Y for one unit change in X

The standard deviation σ quantifies the extent to which the observations deviate from the regression line

Simple linear regression model

Most of the time, the random error is supposed to be normally distributed :

$$\boxed{\varepsilon \sim \mathcal{N}(0, \sigma)}$$

It follows that

$$\boxed{Y|(X = x) \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma)}$$

for any fixed value x for X

Note : we recognise the notation $|$, which means “conditionally on”, as in conditional probabilities. Here we understand : “if we know that X takes the value x, then the distribution of Y is $\mathcal{N}(\beta_0 + \beta_1 x, \sigma)$”

Simple linear regression model

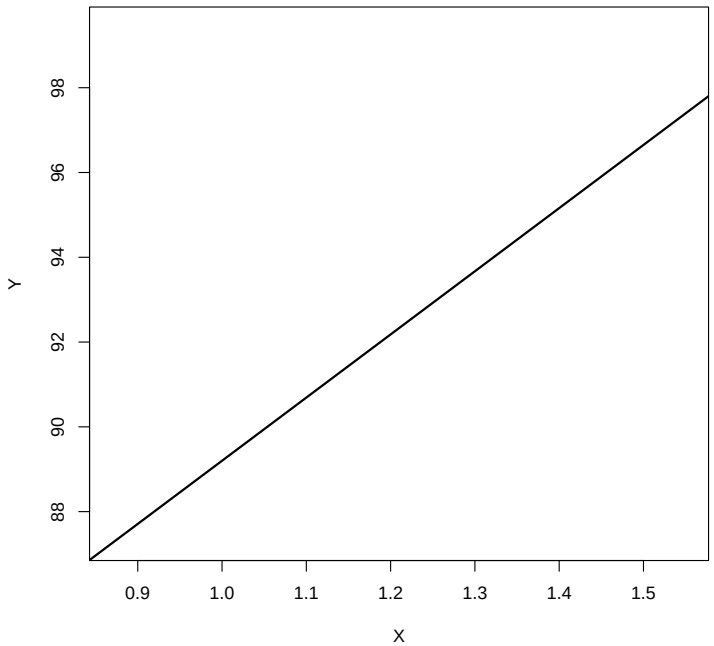
Most of the time, the random error is supposed to be normally distributed :

$$\boxed{\varepsilon \sim \mathcal{N}(0, \sigma)}$$

It follows that

$$\boxed{Y|(X = x) \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma)}$$

for any fixed value x for X



Note : we recognise the notation $|$, which means "conditionally on", as in conditional probabilities. Here we understand : "if we know that X takes the value x, then the distribution of Y is $\mathcal{N}(\beta_0 + \beta_1 x, \sigma)$"

Simple linear regression model

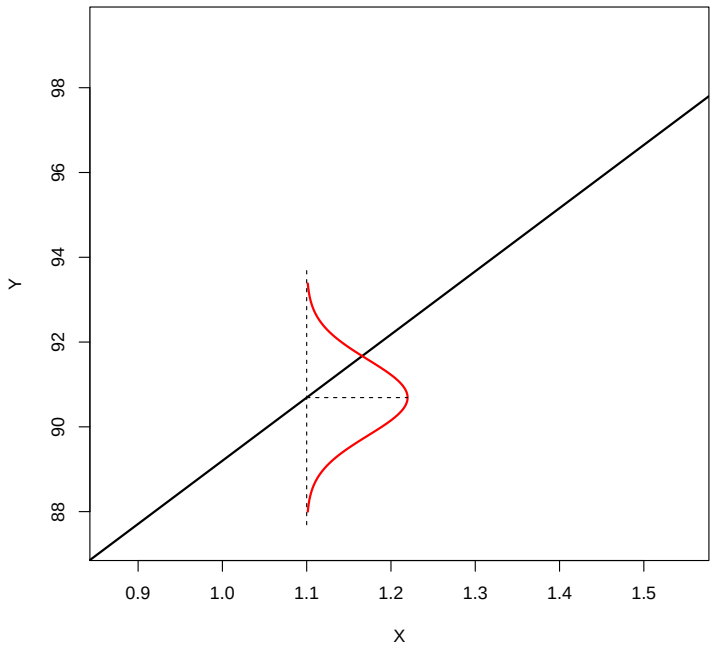
Most of the time, the random error is supposed to be normally distributed :

$$\boxed{\varepsilon \sim \mathcal{N}(0, \sigma)}$$

It follows that

$$\boxed{Y|(X = x) \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma)}$$

for any fixed value x for X



Note : we recognise the notation $|$, which means "conditionally on", as in conditional probabilities. Here we understand : "if we know that X takes the value x, then the distribution of Y is $\mathcal{N}(\beta_0 + \beta_1 x, \sigma)$"

Simple linear regression model

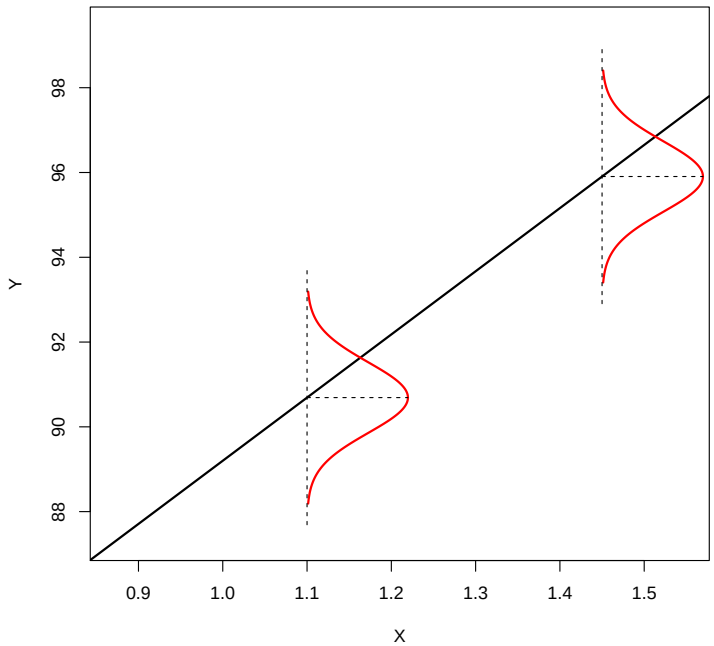
Most of the time, the random error is supposed to be normally distributed :

$$\boxed{\varepsilon \sim \mathcal{N}(0, \sigma)}$$

It follows that

$$\boxed{Y|(X = x) \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma)}$$

for any fixed value x for X



Note : we recognise the notation $|$, which means "conditionally on", as in conditional probabilities. Here we understand : "if we know that X takes the value x, then the distribution of Y is $\mathcal{N}(\beta_0 + \beta_1 x, \sigma)$"

Simple linear regression model

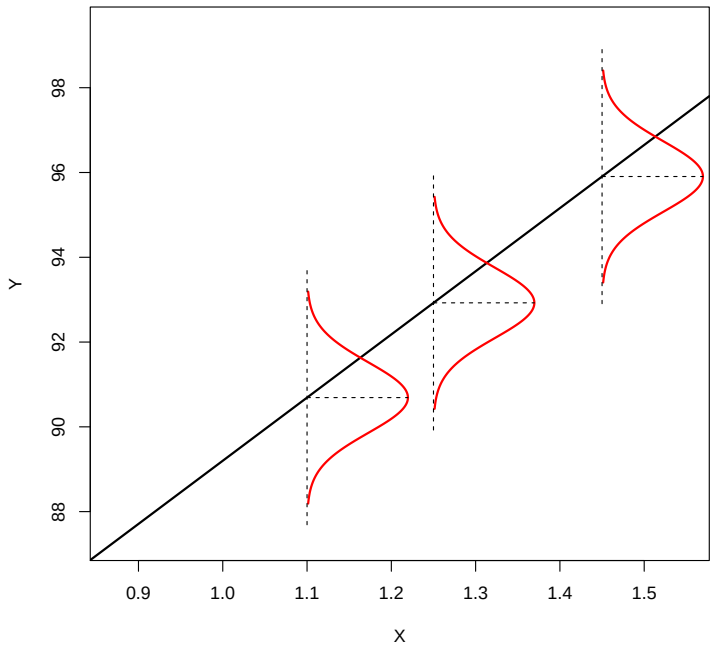
Most of the time, the random error is supposed to be normally distributed :

$$\boxed{\varepsilon \sim \mathcal{N}(0, \sigma)}$$

It follows that

$$\boxed{Y|(X = x) \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma)}$$

for any fixed value x for X



Note : we recognise the notation $|$, which means "conditionally on", as in conditional probabilities. Here we understand : "if we know that X takes the value x, then the distribution of Y is $\mathcal{N}(\beta_0 + \beta_1 x, \sigma)$"

Simple linear regression model

In most real-world problems, the values of the intercept β_0, the slope β_1 and the standard deviation of the error σ will **not be known**

They are **population parameters** which must be estimated from sample data

Simple linear regression model

Here the random sample consists of n pairs of observations (X_i, Y_i), assumed to be independent of each other

$$\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$$

such that

$$Y_i|(X_i = x_i) \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma)$$

for all $i = 1, \ldots, n$

The straight line $\mu_{Y|X=x} = \beta_0 + \beta_1 x$ can be regarded as the population regression line, which need be estimated by a sample version

$$\widehat{\mu}_{Y|X=x} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

Simple linear regression model

The question is how to determine the estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ (and then an estimator for σ)

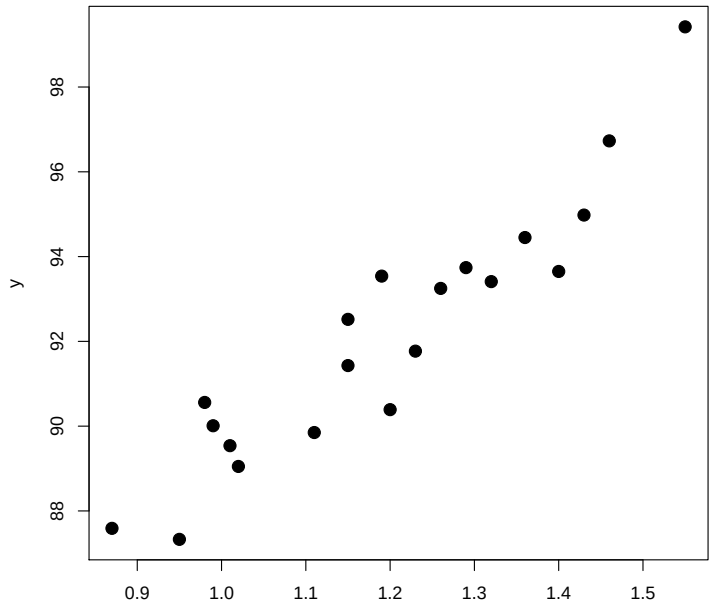
Least Squares Estimators

Least Squares Estimators

The estimates of β_0 and β_1 should result in a line that is (in some sense) a "best fit" to the data

Gauss proposed estimating the parameters β_0 and β_1 to minimise the sum of the squares of the vertical deviations between the observed responses and the straight line

These deviations are often called the **residuals** of the model, and the resulting estimators of β_0 and β_1 are the **least squares estimators**

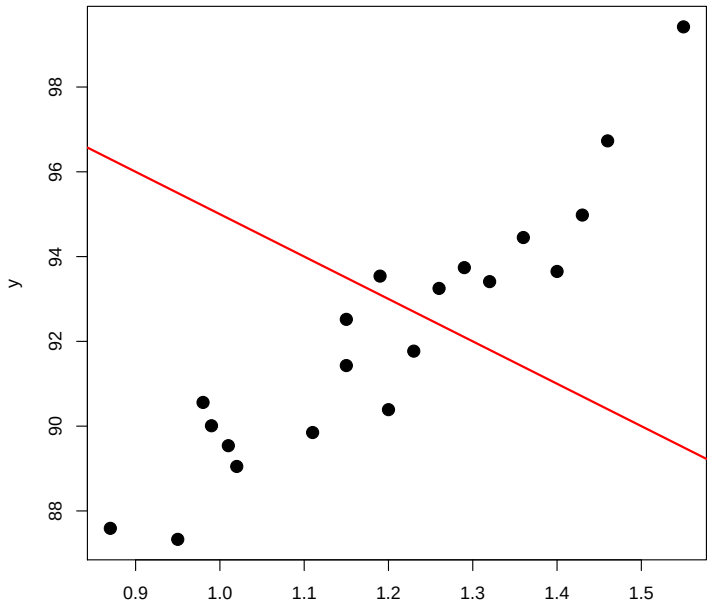


Least Squares Estimators

The estimates of β_0 and β_1 should result in a line that is (in some sense) a "best fit" to the data

Gauss proposed estimating the parameters β_0 and β_1 to minimise the sum of the squares of the vertical deviations between the observed responses and the straight line

These deviations are often called the **residuals** of the model, and the resulting estimators of β_0 and β_1 are the **least squares estimators**

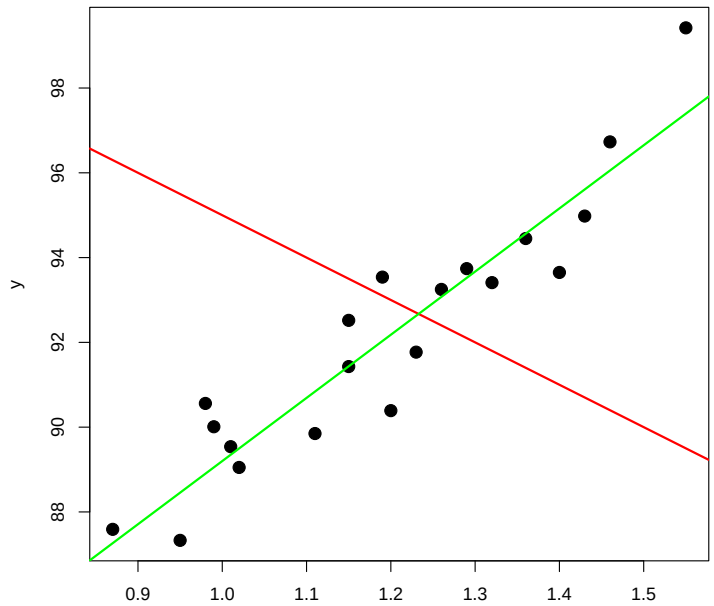


Least Squares Estimators

The estimates of β_0 and β_1 should result in a line that is (in some sense) a "best fit" to the data

Gauss proposed estimating the parameters β_0 and β_1 to minimise the sum of the squares of the vertical deviations between the observed responses and the straight line

These deviations are often called the **residuals** of the model, and the resulting estimators of β_0 and β_1 are the **least squares estimators**

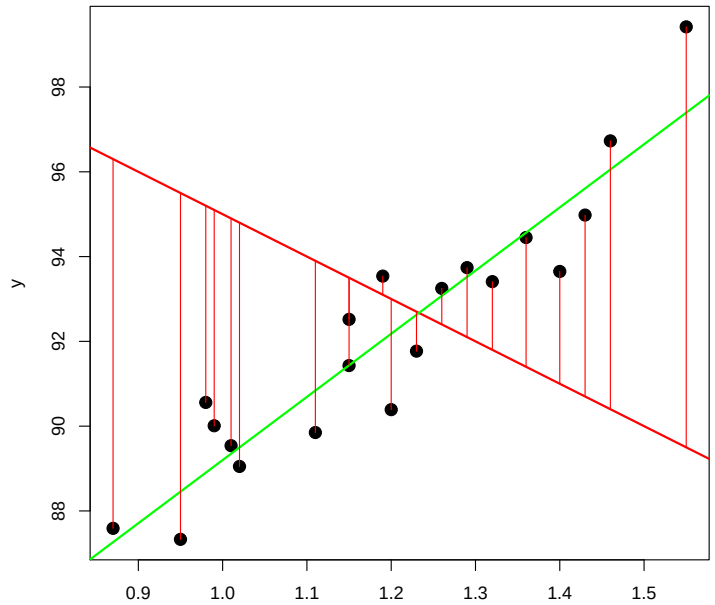


Least Squares Estimators

The estimates of β_0 and β_1 should result in a line that is (in some sense) a "best fit" to the data

Gauss proposed estimating the parameters β_0 and β_1 to minimise the sum of the squares of the vertical deviations between the observed responses and the straight line

These deviations are often called the **residuals** of the model, and the resulting estimators of β_0 and β_1 are the **least squares estimators**

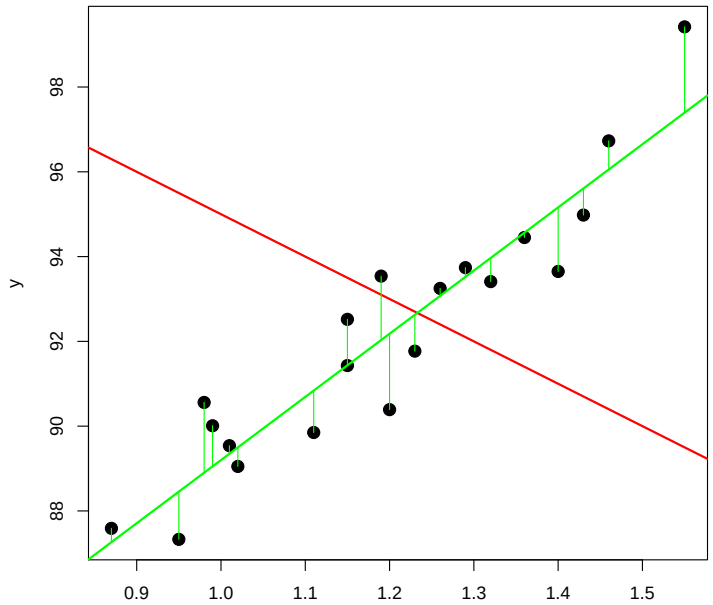


Least Squares Estimators

The estimates of β_0 and β_1 should result in a line that is (in some sense) a "best fit" to the data

Gauss proposed estimating the parameters β_0 and β_1 to minimise the sum of the squares of the vertical deviations between the observed responses and the straight line

These deviations are often called the **residuals** of the model, and the resulting estimators of β_0 and β_1 are the **least squares estimators**

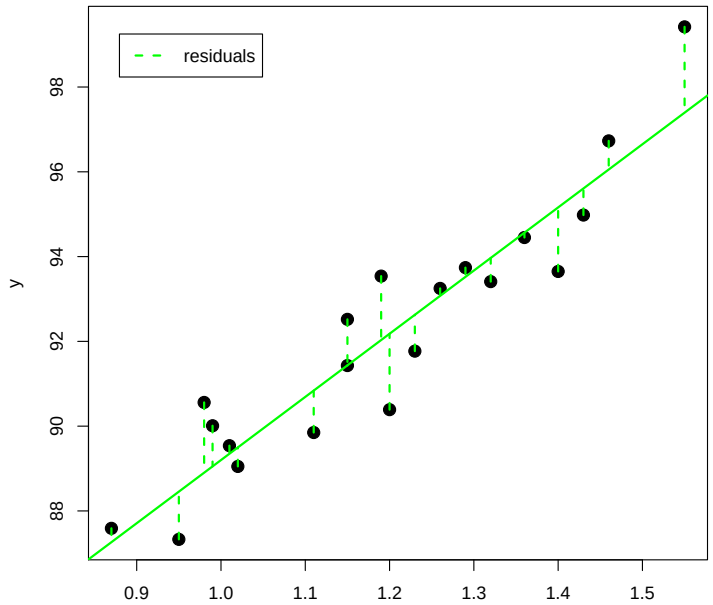


Least Squares Estimators

The estimates of β_0 and β_1 should result in a line that is (in some sense) a "best fit" to the data

Gauss proposed estimating the parameters β_0 and β_1 to minimise the sum of the squares of the vertical deviations between the observed responses and the straight line

These deviations are often called the **residuals** of the model, and the resulting estimators of β_0 and β_1 are the **least squares estimators**



Least Squares Estimators

Write $R(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$

Then,

$$\frac{\partial R}{\partial \beta_0}(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))$$

$$\frac{\partial R}{\partial \beta_1}(\beta_0, \beta_1) = -2 \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i)) X_i$$

$\rightsquigarrow$ the estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ should be the solutions of the equations

$$\begin{cases} \sum_{i=1}^{n} (Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)) = 0 \\ \sum_{i=1}^{n} (Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)) X_i = 0 \end{cases}$$

Least Squares Estimators

The solutions of the equations:

$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X}$$

$$\widehat{\beta}_1 = \frac{\sum_i Y_i X_i - \frac{(\sum_i Y_i)(\sum_i X_i)}{n}}{\sum_i X_i^2 - \frac{(\sum_i X_i)^2}{n}}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

Least Squares Estimators

Introducing the notations

$$S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \left(= \sum_{i=1}^{n} X_i^2 - \frac{(\sum_i X_i)^2}{n} \right)$$

$$S_{XY} = \sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}) \qquad \left(= \sum_{i=1}^{n} X_i Y_i - \frac{(\sum_i X_i)(\sum_i Y_i)}{n} \right)$$

we have :

Least squares estimators of β_0 and β_1

$$\widehat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \quad \text{and} \quad \widehat{\beta}_0 = \bar{Y} - \frac{S_{XY}}{S_{XX}}\bar{X}$$

Note : as $\bar{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{X}$,
the estimated straight line will always go by the point $(\bar{X}, \bar{Y})$

Least Squares Estimates

Once we have observed a sample $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, we have directly the observed values

$$s_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{and } s_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

and thus the estimates $\widehat{b}_1$ and $\widehat{b}_0$ of β_1 and β_0 :

$$\widehat{b}_1 = \frac{s_{xy}}{s_{xx}} \quad \text{and} \quad \widehat{b}_0 = \bar{y} - \frac{s_{xy}}{s_{xx}} \bar{x}$$

The estimated or fitted regression line is therefore :

$$\widehat{b}_0 + \widehat{b}_1 x,$$

which is typically used for estimating the mean response at a particular level of X, or in prediction of future observations of Y

$\rightsquigarrow$ it is often denoted $\widehat{y}(x)$: $\widehat{y}(x) = \widehat{b}_0 + \widehat{b}_1 x$

Clarifying Notation

- : β_0, β_1 are the **unknown parameters** of the regression model
- $\widehat{\beta}_0 = \bar{Y} - \frac{S_{XY}}{S_{XX}}\bar{X}$, $\widehat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$ are the **estimators** of β_0, β_1
- $\widehat{b}_0 = \bar{y} - \frac{s_{xy}}{s_{xx}}\bar{x}$, $\widehat{b}_1 = \frac{s_{xy}}{s_{xx}}$ are the **estimates** of β_0, β_1, **given data** $(x_1, y_1), \ldots (x_n, y_n)$

Least Squares Estimation : example

Example

Fit a simple linear regression model to the data shown on Slide 7

Solution:

From the observed data, the following quantities may be computed :

$$n = 20, \qquad \sum x_i = 23.92, \qquad \sum y_i = 1,843.21$$

$$\bar{x} = 1.1960, \qquad \bar{y} = 92.1605$$

$$\sum x_i^2 = 29.2892, \qquad \sum x_i y_i = 2,214.6566$$

$$s_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 29.2892 - \frac{23.92^2}{20} = 0.68088$$

$$s_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 2,214.6566 - \frac{23.92 \times 1,843.21}{20} = 10.17744$$

Least Squares Estimation : example

Example

Fit a simple linear regression model to the data shown on Slide 7

Solution:

From the observed data, the following quantities may be computed :

$$n = 20, \qquad \sum x_i = 23.92, \qquad \sum y_i = 1,843.21$$

$$\bar{x} = 1.1960, \qquad \bar{y} = 92.1605$$

$$\sum x_i^2 = 29.2892, \qquad \sum x_i y_i = 2,214.6566$$

$$s_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 29.2892 - \frac{23.92^2}{20} = 0.68088$$

$$s_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 2,214.6566 - \frac{23.92 \times 1,843.21}{20} = 10.17744$$

Least Squares Estimation : example

- Therefore, the least squares estimates of the slope and the intercept are

$$\widehat{b}_1 = \frac{s_{xy}}{s_{xx}} = \frac{10.17744}{0.68088} = 14.94748$$

$$\widehat{b}_0 = \bar{y} - \widehat{b}_1 \bar{x} = 92.1605 - 14.94748 \times 1.196 = 74.28331$$

$\rightsquigarrow$ the fitted simple linear regression model is thus

$$\widehat{y}(x) = 74.283 + 14.947x$$

which is the straight line shown on Slide 10

Using this model, we would predict $y = 89.23$ when $x = 1$

Also, the model indicates that the mean value of y would increase by 14.947 for a unit increase (1) in x

Estimating σ^2

- The variance σ^2 of the error term

$$\varepsilon = Y - (\beta_0 + \beta_1 X)$$

is another unknown parameter

$\rightsquigarrow$ the residuals of the fitted model, i.e.

$$\widehat{e}_i = y_i - (\widehat{b}_0 + \widehat{b}_1 x_i) = y_i - \widehat{y}(x_i), \qquad i = 1, 2, \ldots, n$$

can be regarded as a 'sample' drawn from the distribution of ε

$\rightsquigarrow$ a natural estimator for σ^2 should be the sample variance of the residuals $\{\widehat{e}_i, i = 1, \ldots, n\}$

Estimating σ^2

The number of degrees of freedom for the usual sample variance is $n-1$ because we have to estimate one parameter ($\bar{x}$ actually estimates the true μ) (bias!)

Here we have first to estimate two parameters (β_0 and β_1)

$\rightsquigarrow$ the number of degrees of freedom must now be $n-2$

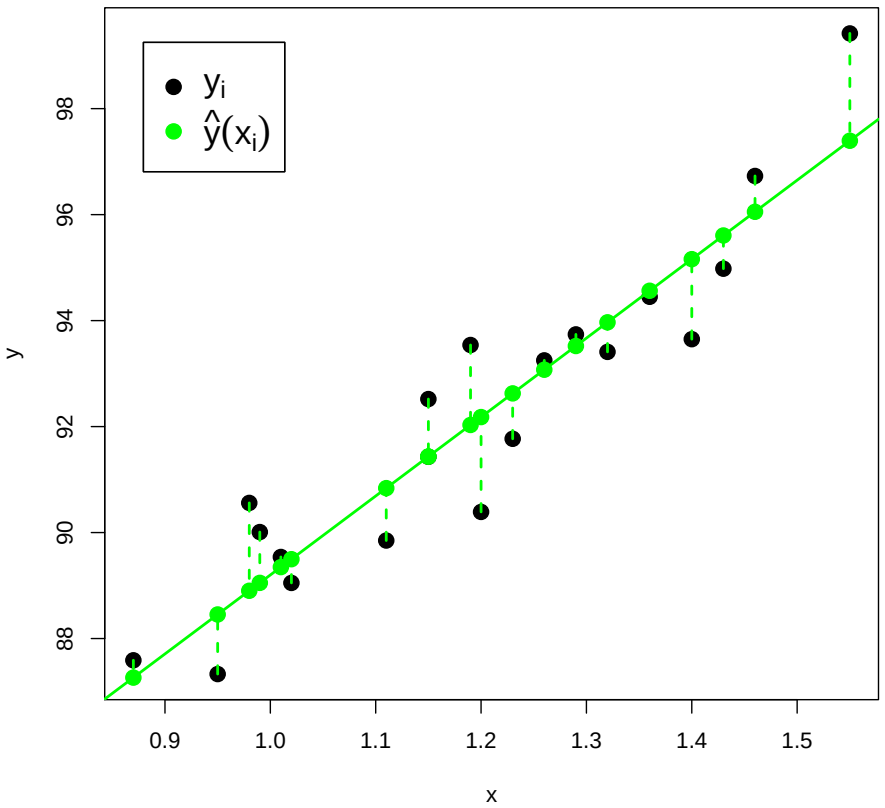
$\rightsquigarrow$ an **unbiased** estimate of σ^2 is

$$\mathbf{s}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{e}_i^2$$

Estimating σ^2 : example

In the previous example, we fitted $\widehat{\mu}_{Y|X=x} = 74.283 + 14.947x$, so that we get a series of fitted values $\widehat{y}(x_i) = 74.283 + 14.947x_i$, for $i = 1, \ldots, 20$, from which the residuals can be computed : $\widehat{e}_i = y_i - \widehat{y}(x_i)$, for $i = 1, \ldots, 20$

i	x_i	y_i	$\widehat{y}(x_i)$	$\widehat{e}_i$
1	0.99	90.01	89.051	0.959
2	1.02	89.05	89.498	-0.448
3	1.15	91.43	91.435	-0.005
4	1.29	93.74	93.521	0.219
5	1.46	96.73	96.054	0.676
6	1.36	94.45	94.564	-0.114
7	0.87	87.59	87.263	0.327
8	1.23	91.77	92.627	-0.857
9	1.55	99.42	97.395	2.025
10	1.40	93.65	95.160	-1.510
11	1.19	93.54	92.031	1.509
12	1.15	92.52	91.435	1.085
13	0.98	90.56	88.902	1.658
14	1.01	89.54	89.349	0.191
15	1.11	89.85	90.839	-0.989
16	1.20	90.39	92.180	-1.790
17	1.26	93.25	93.074	0.176
18	1.32	93.41	93.968	-0.558
19	1.43	94.98	95.607	-0.627
20	0.95	87.33	88.455	-1.125



We find : $\mathbf{s}^2 = \frac{1}{18}\sum_{i=1}^{20} \widehat{e}_i^2 = 1.1824$ $\rightsquigarrow \mathbf{s} = \sqrt{1.1824} = 1.0874$

Properties of the Least Squares Estimators

We said that $Y_i|(X_i = x_i) \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma)$. Suppose that all x_i's are fixed ('fixed design'). Then, because $\sum_i (x_i - \bar{x}) = 0$, we can write

$$\widehat{\beta}_1 = \frac{S_{xY}}{s_{xx}} = \sum_i \frac{(x_i - \bar{x})}{s_{xx}} Y_i$$

$\rightsquigarrow$ $\widehat{\beta}_1$ is a linear combination of the indep. normal random variables Y_i

$\rightsquigarrow$ the estimator $\widehat{\beta}_1$ is normally distributed !

- Its expectation is

$$\mathbb{E}(\widehat{\beta}_1) = \frac{\sum_i (x_i - \bar{x})\mathbb{E}(Y_i)}{s_{xx}} = \frac{\sum_i (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{s_{xx}} = \frac{\beta_1 \sum_i x_i (x_i - \bar{x})}{s_{xx}} = \beta_1$$

$\rightsquigarrow$ **unbiased** estimator of β_1

- Similarly, its variance is $\mathbb{V}\text{ar}(\widehat{\beta}_1) = \frac{\sum_i (x_i - \bar{x})^2 \mathbb{V}\text{ar}(Y_i)}{s_{xx}^2} = \frac{\sigma^2 \sum_i (x_i - \bar{x})^2}{s_{xx}^2} = \frac{\sigma^2}{s_{xx}}$

Properties of the Least Squares Estimators

Hence, the sampling distribution of $\widehat{\beta}_1$ given $x_1, \ldots, x_n$ is

$$\widehat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma}{\sqrt{s_{xx}}}\right)$$

Properties of the Least Squares Estimators

- Now, we can write

$$\widehat{\beta}_0 = \sum_{i=1}^{n} \frac{Y_i}{n} - \widehat{\beta}_1 \bar{x},$$

which is again a linear combination of indep. normal random variables

$\rightsquigarrow$ the estimator $\widehat{\beta}_0$ is also normally distributed !

- Its expectation is

$$\mathbb{E}(\widehat{\beta}_0) = \sum_{i=1}^{n} \frac{\mathbb{E}(Y_i)}{n} - \mathbb{E}(\widehat{\beta}_1)\bar{x} = \sum_{i=1}^{n} \frac{\beta_0 + \beta_1 x_i}{n} - \beta_1 \bar{x} = \beta_0$$

$\rightsquigarrow$ **unbiased** estimator of β_0

- Similarly, we would find $\mathbb{V}\text{ar}(\widehat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} \right)$

Properties of the Least Squares Estimators

Hence, the sampling distribution of $\widehat{\beta}_0$ given $x_1, \ldots, x_n$ is

$$\widehat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}}\right)$$

Properties of the Least Squares Estimators

To summarise, the sampling distribution of $\widehat{\beta}_0$ given $x_1, \ldots, x_n$ is

$$\widehat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}}\right)$$

and the sampling distribution of $\widehat{\beta}_1$ given $x_1, \ldots, x_n$ is

$$\widehat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma}{\sqrt{s_{xx}}}\right)$$

Inferences in simple linear regression

Inferences concerning β_1

An important hypothesis to consider regarding the simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ is the hypothesis that $\beta_1 = 0$

$\rightsquigarrow$ $\beta_1 = 0$ is equivalent to stating that the response does not depend on the predictor X (as we would have $Y = \beta_0 + \varepsilon$)

Suppose that $x_1, \ldots, x_n$ are given (i.e. fixed design)

We can set up a formal hypothesis test. The appropriate hypotheses are :

$$H_0 : \beta_1 = 0 \qquad \text{against} \qquad H_a : \beta_1 \neq 0$$

$\rightsquigarrow$ we will reject H_0 when the observed $\widehat{\beta}_1$ will be 'too different' to 0

From the sampling distribution of $\widehat{\beta_1}$, we get $\sqrt{s_{xx}}\,\frac{\widehat{\beta}_1 - \beta_1}{\sigma} \sim \mathcal{N}(0,1)$

However, σ is typically unknown $\rightsquigarrow$ replace it with its estimator **S**

$$\mathbf{S} := \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} \widehat{\varepsilon}_i^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} \left(Y_i - \widehat{y}(x_i)\right)^2}$$

As this estimator of σ admits $n-2$ degrees of freedom, we find :

$$\boxed{T_1 := \sqrt{s_{xx}}\,\frac{\widehat{\beta}_1 - \beta_1}{\mathbf{S}} \sim t_{n-2}}$$

From this result, all the inferential procedures that we introduced previously can be readily adapted
At significance level α%, the rejection criterion for $H_0 : \beta_1 = 0$ is

$$\text{reject } H_0 \text{ if } \widehat{b}_1 \notin \left[0 - t_{n-2,1-\alpha/2}\frac{\mathbf{s}}{\sqrt{s_{xx}}}, 0 + t_{n-2,1-\alpha/2}\frac{\mathbf{s}}{\sqrt{s_{xx}}}\right],$$

with the observed estimated standard deviation $\mathbf{s} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\widehat{e}_i^2}$

and from the observed value of the test statistic

$$t_1 = \sqrt{s_{xx}}\,\frac{\widehat{b}_1}{\mathbf{s}}$$

we can compute the *p*-value

$$p = 1 - \mathbb{P}(T_1 \in [-|t_1|, |t_1|]) = 2\mathbb{P}(T_1 > |t_1|)$$

(recall T_1 is a random variable with distribution t_{n-2})

- In addition to the point estimator $\widehat{\beta}_1$ of the slope, it is also possible to obtain a confidence interval for the 'true' slope β_1

As $\sqrt{s_{xx}}\,\frac{\widehat{\beta}_1-\beta_1}{\mathbf{S}} \sim t_{n-2}$, we can directly write

$$\mathbb{P}\left(-t_{n-2;1-\alpha/2} \leq \sqrt{s_{xx}}\,\frac{\widehat{\beta}_1-\beta_1}{\mathbf{S}} \leq t_{n-2;1-\alpha/2}\right) = 1-\alpha$$

or equivalently

$$\mathbb{P}\left(\widehat{\beta}_1 - t_{n-2;1-\alpha/2}\frac{\mathbf{S}}{\sqrt{s_{xx}}} \leq \beta_1 \leq \widehat{\beta}_1 + t_{n-2;1-\alpha/2}\frac{\mathbf{S}}{\sqrt{s_{xx}}}\right) = 1-\alpha$$

- Given observed data $(x_1, y_1), \ldots, (x_n, y_n)$
  $\Longrightarrow$ we find **s** and $\widehat{b}_1$,
  $\Longrightarrow$ a two-sided $100 \times (1-\alpha)\%$ confidence interval for the parameter β_1 is

$$\left[\widehat{b}_1 - t_{n-2;1-\alpha/2}\frac{\mathbf{s}}{\sqrt{s_{xx}}}, \widehat{b}_1 + t_{n-2;1-\alpha/2}\frac{\mathbf{s}}{\sqrt{s_{xx}}}\right]$$

Inferences concerning β_0

Although of less practical interest (often), inferences concerning the parameter β_0 can be accomplished in exactly the same manner from the sampling distribution of $\widehat{\beta}_0$

Inferences concerning β_0

We find a two-sided $100 \times (1-\alpha)\%$ **confidence interval** for β_0

$$\left[\widehat{\beta}_0 - t_{n-2;1-\alpha/2}\mathbf{s}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}}, \widehat{\beta}_0 + t_{n-2;1-\alpha/2}\mathbf{s}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}}\right]$$

as well as a **rejection criterion** for an hypothesis $H_0 : \beta_0 = 0$ (no intercept in the model) tested against $H_a : \beta_0 \neq 0$: at level $\alpha\%$,

$$\text{reject } H_0 \text{ if } \widehat{b}_0 \notin \left[-t_{n-2,1-\alpha/2}\mathbf{s}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}}, t_{n-2,1-\alpha/2}\mathbf{s}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}}\right]$$

with a *p*-**value** calculated from the observed value of the test statistic

$$t_0 = \frac{\widehat{b}_0}{\mathbf{s}\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}}} \qquad \rightsquigarrow p = 2\mathbb{P}(T_0 > |t_0|)$$

(where T_0 has distribution t_{n-2})

Inferences concerning β_1 : example

Example

Test for significance of the simple linear regression model for the data shown on Slide 7 at level $\alpha = 0.01$

Solution:

The model is $Y = \beta_0 + \beta_1 X + \varepsilon$. The hypotheses are :

$$H_0 : \beta_1 = 0 \qquad \text{against} \qquad H_a : \beta_1 \neq 0$$

The estimate of β_1 is $\widehat{b}_1 = 14.947$. Besides, we previously found $n = 20$, $s_{xx} = 0.68088$ and $\mathbf{s} = 1.0874$. By the t-distribution table, $t_{18;0.995} = 2.878$.

- Hence, at significance level $\alpha = 0.01$, the rejection criterion is :

$$\text{reject } H_0 \text{ if } \widehat{b}_1 \notin \left[-2.878 \times \frac{1.0874}{\sqrt{0.68088}}, 2.878 \times \frac{1.0874}{\sqrt{0.68088}} \right] = [-3.793, 3.793]$$

Here, with $\widehat{b}_1 = 14.947$, we obviously reject H_0

$\rightsquigarrow$ the 'true' slope β_1 between x and y levels is most certainly different from 0

Inferences concerning β_1 : example

Example

Test for significance of the simple linear regression model for the data shown on Slide 7 at level $\alpha = 0.01$

Solution:

The model is $Y = \beta_0 + \beta_1 X + \varepsilon$. The hypotheses are :

$$H_0 : \beta_1 = 0 \qquad \text{against} \qquad H_a : \beta_1 \neq 0$$

The estimate of β_1 is $\widehat{b}_1 = 14.947$. Besides, we previously found $n = 20$, $s_{xx} = 0.68088$ and $\mathbf{s} = 1.0874$. By the t-distribution table, $t_{18;0.995} = 2.878$.

- Hence, at significance level $\alpha = 0.01$, the rejection criterion is :

$$\text{reject } H_0 \text{ if } \widehat{b}_1 \notin \left[-2.878 \times \frac{1.0874}{\sqrt{0.68088}}, 2.878 \times \frac{1.0874}{\sqrt{0.68088}} \right] = [-3.793, 3.793]$$

Here, with $\widehat{b}_1 = 14.947$, we obviously reject H_0

$\rightsquigarrow$ the 'true' slope β_1 between x and y levels is most certainly different from 0

Inferences concerning β_1 : example

- The observed value of the test statistic is

$$t_1 = \frac{\sqrt{0.68088}}{1.0874} \times 14.947 = 11.35$$

and the *p*-value is $p = 2 \times \mathbb{P}(T_1 > 11.35) \simeq 0$ (with $T_1 \sim t_{18}$)

- We can also derive a 99% confidence interval for β_1 :

$$\left[14.947 \pm 2.878 \times \frac{1.0874}{\sqrt{0.68088}}\right] = [11.181, 18.767]$$

$\rightsquigarrow$ we can be 99% confident that the true value of the slope β_1 lies between 11.181 and 18.767 (so that 0 is obviously not one of the plausible values for β_1)

Confidence Interval on the Mean Response

A confidence interval may be constructed on the mean response at a specified value of X, say, x

This is thus a confidence interval for the unknown 'parameter'

$$\mu_{Y|X=x} = \beta_0 + \beta_1 x$$

From the fitted model, we have directly an estimator for this parameter :

$$\widehat{\mu}_{Y|X=x} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

Note that, as a linear combination of indep. normal random variables, the estimator $\widehat{\beta}_0 + \widehat{\beta}_1 x$ is also normally distributed. Its expectation is :

$$\mathbb{E}(\widehat{\mu}_{Y|X=x}) = \mathbb{E}(\widehat{\beta}_0 + \widehat{\beta}_1 x) = \mathbb{E}(\widehat{\beta}_0) + \mathbb{E}(\widehat{\beta}_1)x = \beta_0 + \beta_1 x = \mu_{Y|X=x}$$

$\rightsquigarrow$ **unbiased** estimator for $\mu_{Y|X=x}$

Confidence Interval on the Mean Response

Its variance can be found to be

$$\mathbb{V}\text{ar}(\widehat{\mu}_{Y|X=x}) = \sigma^2 \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}} \right)$$

Note 1 : this is not $\mathbb{V}\text{ar}(\widehat{\beta}_0) + \mathbb{V}\text{ar}(\widehat{\beta}_1)x^2$, because $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are not independent! Indeed, $\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{x}$

Note 2 : because we know that the fitted straight line will always go by $(\bar{x}, \bar{Y})$, the variability in $\widehat{\mu}_{Y|X=x}$ decreases as x approaches $\bar{x}$ and vice-versa $\rightsquigarrow$ term $\frac{(x-\bar{x})^2}{s_{xx}}$

At $x = \bar{x}$, $\mathbb{V}\text{ar}(\widehat{\mu}_{Y|X=x}) = \frac{\sigma^2}{n}$, which is just the variance of $\bar{Y}$!

Confidence Interval on the Mean Response

Hence, sampling distribution of estimator $\widehat{\mu}_{Y|X=x}$ given $x_1, \ldots, x_n$ is

$$\widehat{\mu}_{Y|X=x} \sim \mathcal{N}\left(\mu_{Y|X=x}, \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}\right)$$

Confidence Interval on the Mean Response

If we standardise and replace the unknown σ by its estimator S, we get

$$\frac{\widehat{\mu}_{Y|X=x} - \mu_{Y|X=x}}{\mathbf{S}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}} \sim t_{n-2}$$

which directly leads to the following confidence interval for $\mu_{Y|X=x}$:

- **Given data** $(x_1, y_1) \ldots, (x_n, y_n)$
  $\Longrightarrow$ we find **s** and $\widehat{y}(x)$ from the fitted model $\widehat{y}(x) = \widehat{b}_0 + \widehat{b}_1 x$,
  $\Longrightarrow$ a two-sided $100 \times (1-\alpha)\%$ confidence interval for the parameter $\mu_{Y|X=x}$, that is the mean response Y when $X = x$:

$$\left[\widehat{y}(x) - t_{n-2;1-\alpha/2}\mathbf{s}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}, \widehat{y}(x) + t_{n-2;1-\alpha/2}\mathbf{s}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}\right]$$

Confidence Interval on the Mean Response : example

Example

Construct a 95% confidence interval on the mean $\mu_{Y|X=x}$ when the level X is fixed to $x = 1$ (from the data shown on Slide 7)

Solution:

- The fitted model was $\widehat{y}(x) = 74.283 + 14.947x$. We also have $n = 20$, $\mathbf{s} = 1.0874$, $s_{xx} = 0.68088$ and $\bar{x} = 1.1960$. In the t-distribution table, we find $t_{18;0.975} = 2.101$.
- When $x = 1$, the model estimates the mean response $\mu_{Y|X=1}$ to be $\widehat{y}(1) = 89.23$

$\rightsquigarrow$ a 95% confidence interval for $\mu_{Y|X=1}$ is given by

$$\left[89.23 \pm 2.101 \times 1.0874 \times \sqrt{\frac{1}{20} + \frac{(1 - 1.1960)^2}{0.68088}}\right] = [88.48, 89.98]$$

$\rightsquigarrow$ when $x = 1$, we are 95% confident that the true mean level is between 88.48 and 89.98

Confidence Interval on the Mean Response : example

Example

Construct a 95% confidence interval on the mean $\mu_{Y|X=x}$ when the level X is fixed to $x = 1$ (from the data shown on Slide 7)

Solution:

- The fitted model was $\widehat{y}(x) = 74.283 + 14.947x$. We also have $n = 20$, $\mathbf{s} = 1.0874$, $s_{xx} = 0.68088$ and $\bar{x} = 1.1960$. In the t-distribution table, we find $t_{18;0.975} = 2.101$.
- When $x = 1$, the model estimates the mean response $\mu_{Y|X=1}$ to be $\widehat{y}(1) = 89.23$

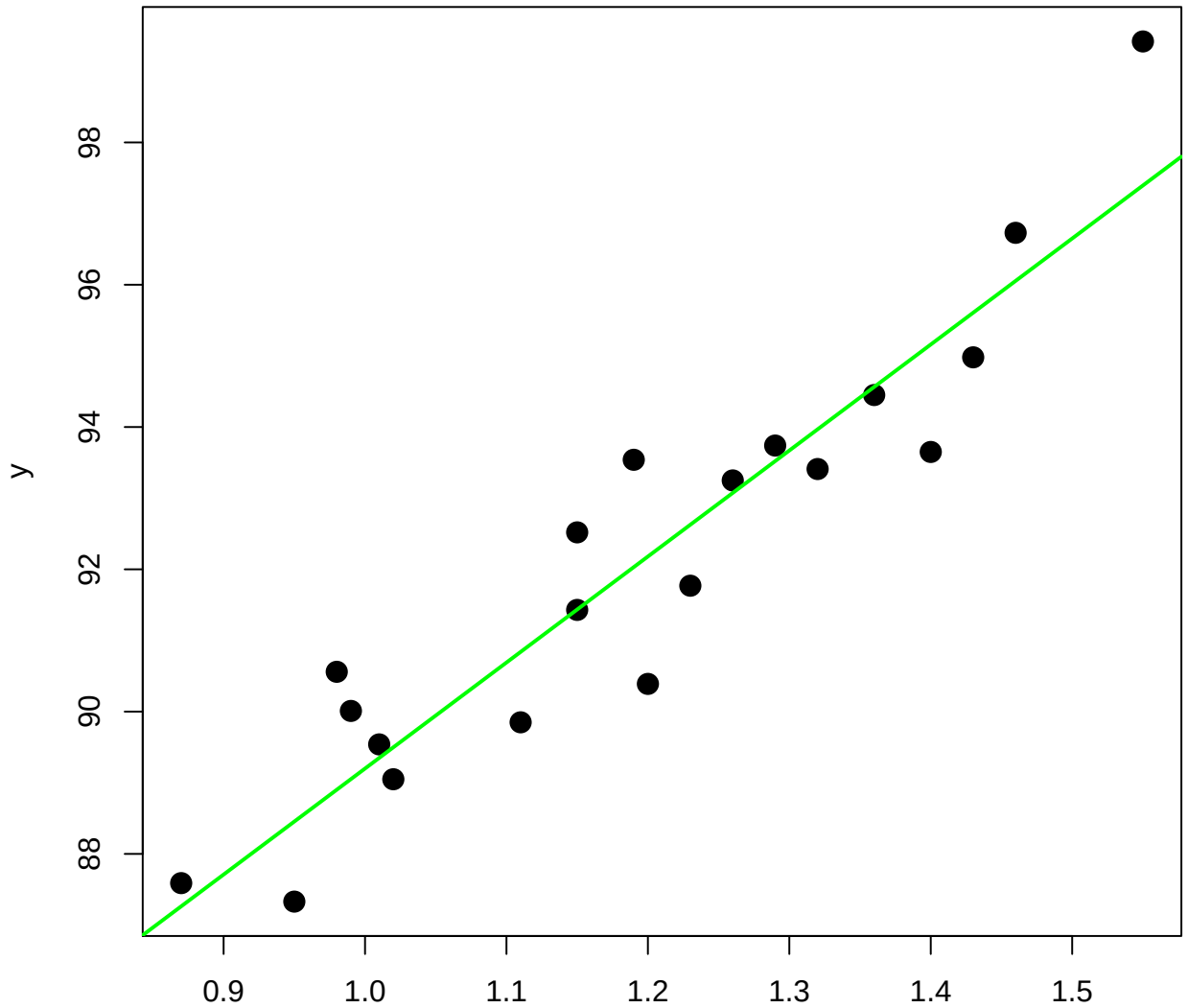
$\rightsquigarrow$ a 95% confidence interval for $\mu_{Y|X=1}$ is given by

$$\left[89.23 \pm 2.101 \times 1.0874 \times \sqrt{\frac{1}{20} + \frac{(1 - 1.1960)^2}{0.68088}}\right] = [88.48, 89.98]$$

$\rightsquigarrow$ when $x = 1$, we are 95% confident that the true mean level is between 88.48 and 89.98

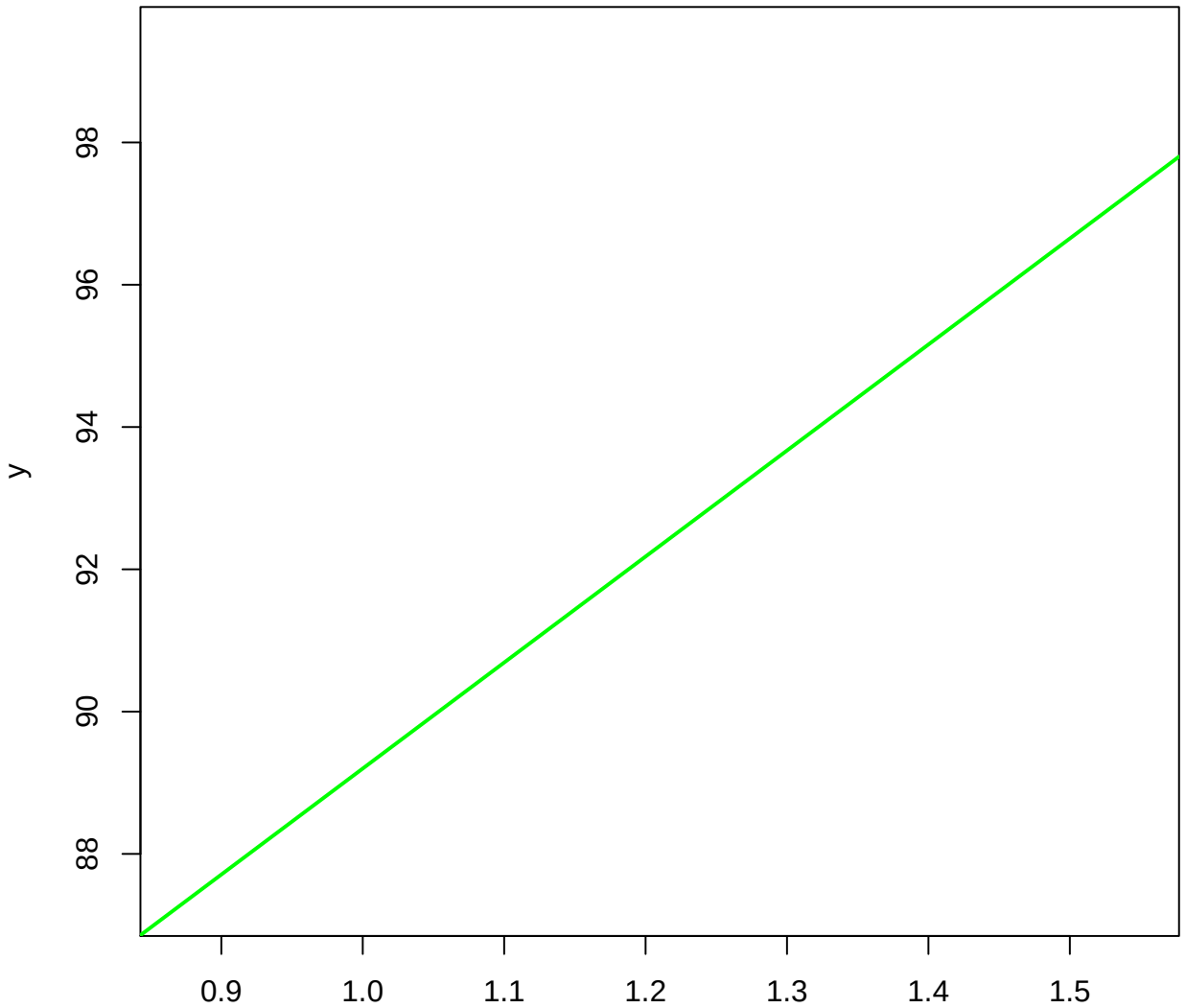
Confidence Interval on the Mean Response : example

By repeating these calculations for several different values for x, we can obtain confidence limits for each corresponding value of $\mu_{Y|X=x}$



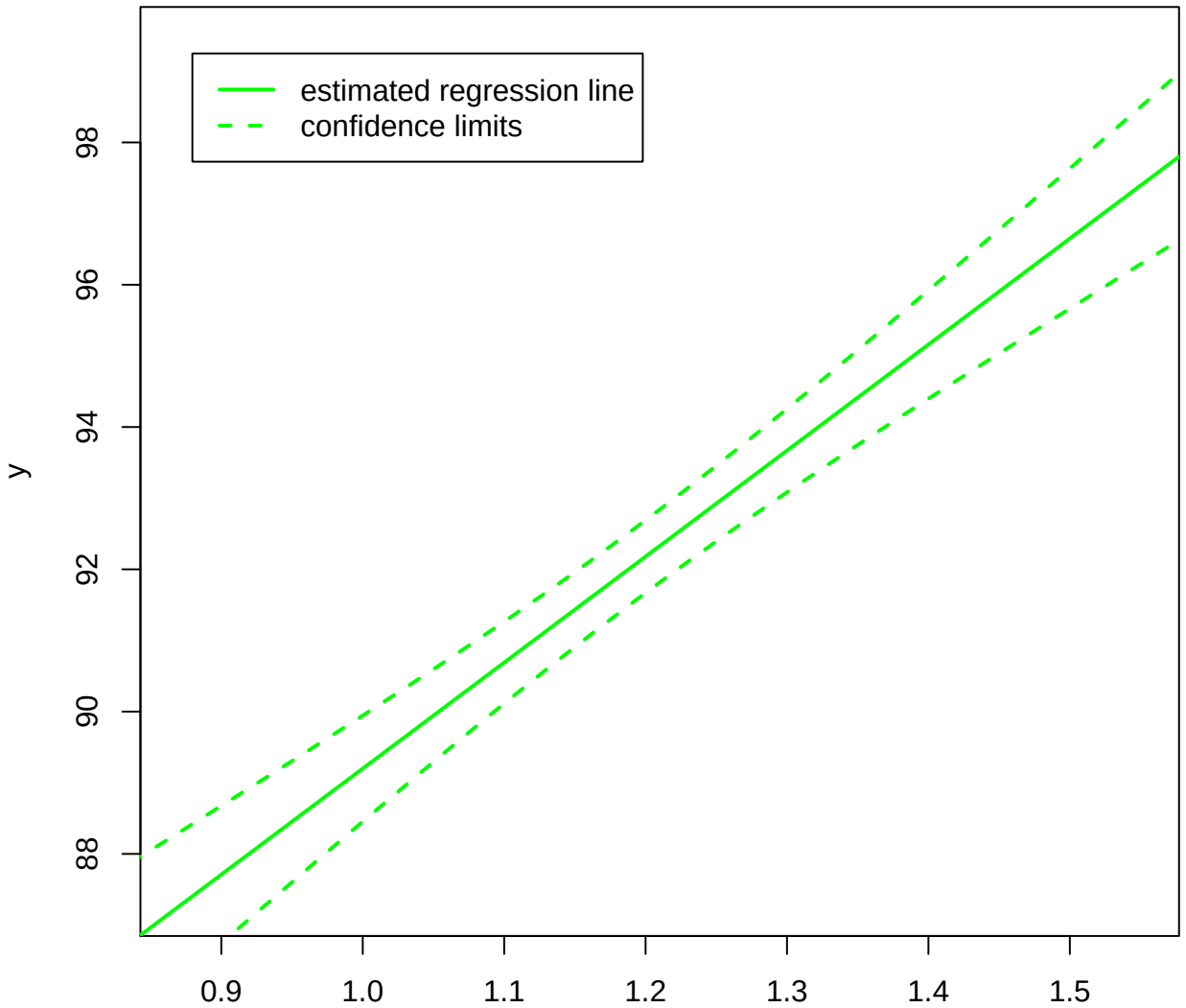
Confidence Interval on the Mean Response : example

By repeating these calculations for several different values for x, we can obtain confidence limits for each corresponding value of $\mu_{Y|X=x}$



Confidence Interval on the Mean Response : example

By repeating these calculations for several different values for x, we can obtain confidence limits for each corresponding value of $\mu_{Y|X=x}$



Prediction of new observations

Prediction of new observations

An important application of a regression model is predicting new or future observations Y corresponding to a specified level $X = x$

$\rightsquigarrow$ different to estimating the mean response $\mu_{Y|X=x}$ at $X = x$!

From the model, the predictor of the new value of the response Y at $X = x$, say $Y^*(x)$ is naturally given by

$$Y^*(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x,$$

for which a predicted value is

$$\widehat{y}(x) = \widehat{b}_0 + \widehat{b}_1 x$$

once the model has been fitted from an observed sample
$\rightsquigarrow$ the predictor of Y at $X = x$ is the estimator of $\mu_{Y|X=x}$!
The prediction error $\mathfrak{e}$ is given by $Y|(X = x) - Y^*(x)$ and is normally distributed, as both $Y|(X = x)$ and $Y^*(x)$ are

Prediction of new observations

- As $Y|(X=x) \sim \mathcal{N}(\beta_0+\beta_1 x, \sigma)$ (Slide 12) and $Y^*(x) = \widehat{\mu}_{Y|X=x} \sim \mathcal{N}\left(\beta_0+\beta_1 x, \sigma\sqrt{\frac{1}{n}+\frac{(x-\bar{x})^2}{s_{xx}}}\right)$ (Slide 46), the expectation of the prediction error is

$$\mathbb{E}((Y|(X=x)) - Y^*(x)) = \mathbb{E}(Y|X=x) - \mathbb{E}(Y^*(x)) = 0$$

$\rightsquigarrow$ on the average, the predictor will find the right value

- Because the future Y is independent of the sample observations (and thus independent of $\widehat{\mu}_{Y|X=x}$), the variance of **prediction error** is

$$\begin{aligned}\mathbb{V}\text{ar}((Y|(X=x)) - Y^*(x)) &= \mathbb{V}\text{ar}(Y|X=x) + \mathbb{V}\text{ar}(Y^*(x)) \\ &= \sigma^2 + \sigma^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{s_{xx}}\right) = \sigma^2\left(1+\frac{1}{n}+\frac{(x-\bar{x})^2}{s_{xx}}\right)\end{aligned}$$

and we find

$$\mathfrak{e} := (Y|(X=x)) - Y^*(x) \sim \mathcal{N}\left(0, \sigma\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{s_{xx}}}\right)$$

Prediction of new observations

- Standardising and replacing the unknown σ by its estimator **S**, we get (as usual) :

$$\frac{(Y|(X=x)) - Y^*(x)}{\mathbf{S}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{s_{xx}}}} \sim t_{n-2}$$

which directly leads to the following **prediction interval** for a new observation Y, given that $X = x$:

- Given data $(x_1, y_1), \ldots, (x_n, y_n)$
  $\Longrightarrow$ we find **s** and $\widehat{y}(x)$ from the fitted model $\widehat{y}(x) = \widehat{b}_0 + \widehat{b}_1 x$,
  $\Longrightarrow$ a two-sided $100 \times (1-\alpha)\%$ prediction interval for a new observation Y at $X = x$ is

$$\left[\widehat{y}(x) - t_{n-2;1-\alpha/2}\mathbf{s}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{s_{xx}}}, \widehat{y}(x) + t_{n-2;1-\alpha/2}\mathbf{s}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{s_{xx}}}\right]$$

Prediction of new observations : remarks

Similarly to the remarks on Slides 109–111 of Chapter 4, we observe :

1. a prediction interval for Y at $X = x$ will always be longer than the confidence interval for $\mu_{Y|X=x}$ because there is much more variability in one observation than in an average

   Concretely, $\mu_{Y|X=x}$ is the position of the straight line at $X = x$
   $\rightsquigarrow$ the CI for $\mu_{Y|X=x}$ only targets that position

   However, we know that observations will not be exactly on that straight line, but 'around' it
   $\rightsquigarrow$ a prediction interval for a new observation should take this extra variability into account, in addition to the uncertainty inherent in the estimation of $\mu_{Y|X=x}$
2. as n gets larger ($n \to \infty$), the width of the CI for $\mu_{Y|X=x}$ decreases to 0 (we are more and more accurate when estimating μ), but this is not the case for the prediction interval : the inherent variability in the new observation never vanishes, even when we have observed many other observations before!

Prediction of new observations : example

Example

Construct a 95% prediction interval on Y when the level X is fixed to $x = 1\%$ (from the data shown on Slide 4)

Solution:

- The fitted model was $\widehat{y}(x) = 74.283 + 14.947x$. We also have $n = 20$, $\mathbf{s} = 1.0874$, $s_{xx} = 0.68088$ and $\bar{x} = 1.1960$. In the t-distribution table, we find $t_{18;0.975} = 2.101$.
- For $x = 1$, the model estimates the mean response $\mu_{Y|X=1}$ to $\widehat{y}(1) = 89.23$

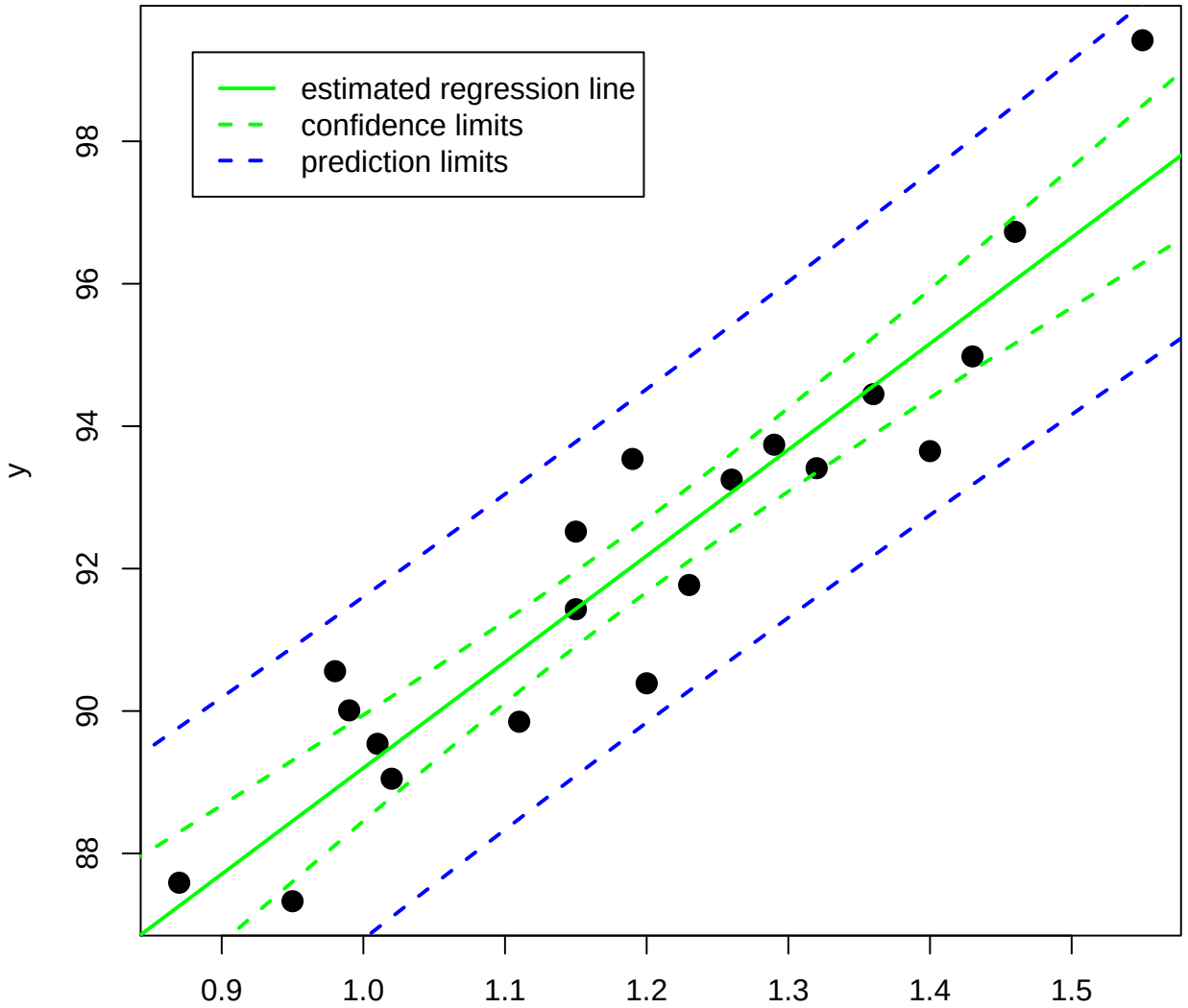
$\rightsquigarrow$ a 95% prediction interval for Y is given by

$$\left[89.23 \pm 2.101 \times 1.0874 \times \sqrt{1 + \frac{1}{20} + \frac{(1 - 1.1960)^2}{0.68088}}\right] = [86.83, 91.63]$$

$\rightsquigarrow$ if we fix the $x = 1\%$, we can be 95% confident that the next observed value of y will be between 86.83 and 91.63

Prediction of new observations : example

By repeating these calculations for several different values for x, we can obtain prediction limits for each corresponding value of Y given that $X = x$



Adequacy of the regression model

Adequacy of the regression model

While using the simple linear regression model, we made several assumptions

- The first one is that **the model is correct** : there indeed exist coefficients β_0 and β_1, as well as a random variable ε, such that we can write $Y = \beta_0 + \beta_1 X + \varepsilon \quad \rightsquigarrow$ **scatterplot**
- The other central assumption is certainly that (Slide 12)

$$Y_i|(X_i = x_i) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma) \quad \text{for } i = 1, 2, \ldots, n,$$

which has several implications

Adequacy of the regression model

Define the error terms

$$e_i = y_i - (\beta_0 + \beta_1 x_i), \text{ for } i = 1, \ldots, n$$

which are values drawn from the distribution of ε. We must check that :

1. the e_i's have been drawn **independently** of one another
2. the e_i's have been drawn from a distribution with the **same variance**
3. the e_i's have been drawn from a **normal distribution**

Residual analysis

Unfortunately, we do not have access to the values e_i's (as we do not know β_0 and β_1)

However, the observed residuals of the fitted model

$$\widehat{e}_i = y_i - \widehat{y}(x_i) = y_i - (\widehat{b}_0 + \widehat{b}_1 x_i)$$

are probably good estimates of those e_i's $\rightsquigarrow$ **residual analysis**

Residual analysis

It is frequently helpful to plot the residuals

(1) in time sequence (if known),

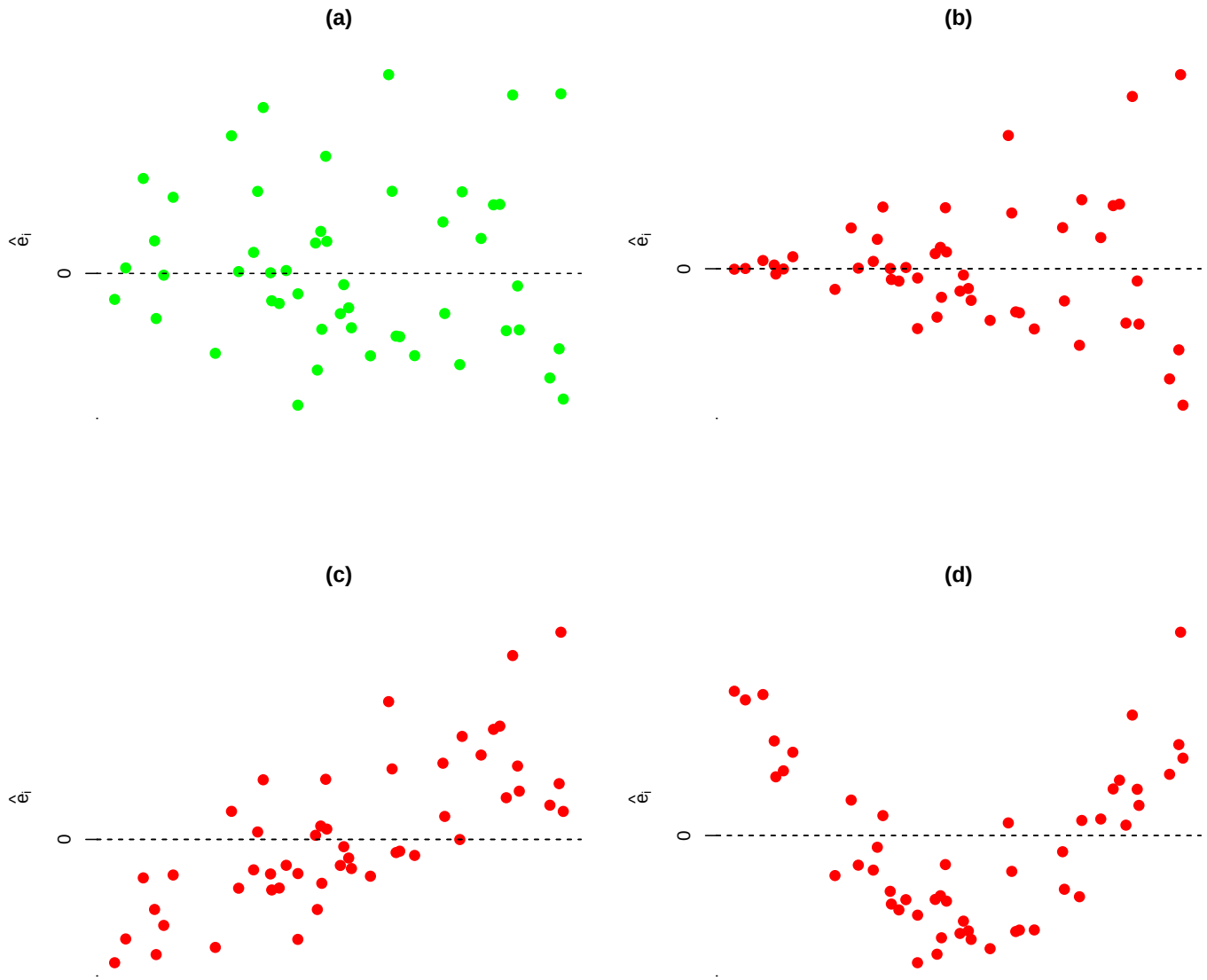
(2) against the fitted values $\widehat{y}(x_i)$, and

(3) against the predictor values x_i

Typically, these graphs will look like one of the four general patterns shown on the next slide

As suggested by their name, the residuals are everything the model will not consider $\rightsquigarrow$ no information should be observed in the residuals, they should look like noise

Residual analysis



Residual analysis

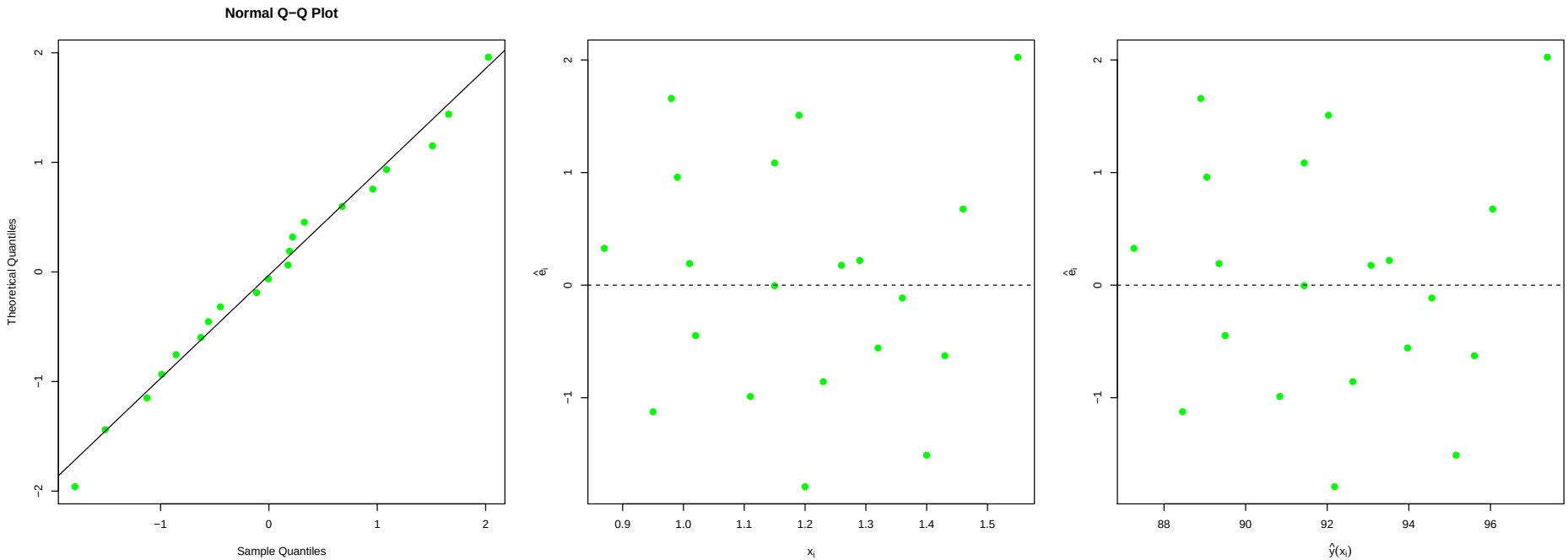
- **Pattern (a) represents thus the ideal situation**
- In (b), the variance of the error terms e_i (and thus of the responses Y_i) may be increasing with time or with magnitude of Y_i or X_i
- Plot (c) indicates some sort of dependence in the error terms
- In (d), we get clear indication of model inadequacy : the residuals are systematically positive for extreme values and negative for medium values

$\rightsquigarrow$ the model is not complete, there is still much information in the residuals : higher-order terms (X^2) or other predictors should be considered in the model

Finally, a **normal probability plot (or a histogram) of residuals** is constructed so as to check for the normality assumption

Residual analysis : example

From our running example , a normal quantile plot and plots against the predicted values $\widehat{y}(x_i)$ and against the level x_i for the residuals computed on Slide 27, are shown below :



$\rightsquigarrow$ nothing to report

$\rightsquigarrow$ the assumptions we made look totally valid

Variability decomposition

Similarly to the notations on Slide 20, we can define

$$s_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$\rightsquigarrow$ this measures the total amount of variability in the response values, and is sometimes denoted ss_t (for '**total sum of squares**')

Now, this variability in the observed values y_i arises from <u>two factors</u> :

1. because the x_i values are different, all Y_i have different means. This variability is quantified by the '**regression sum of squares**' :

$$ss_r = \sum_{i=1}^{n} (\widehat{y}(x_i) - \bar{y})^2$$

2. each value Y_i has variance σ^2 around its mean. This variability is quantified by the '**error sum of squares**' :

$$ss_e = \sum_{i=1}^{n} (y_i - \widehat{y}(x_i))^2 = \sum_{i=1}^{n} \widehat{e}_i^2$$

We can always write : $\boxed{ss_t = ss_r + ss_e}$

Coefficient of determination

Suppose $ss_t \simeq ss_r$ and $ss_e \simeq 0$: the variability in the responses due to the effect of the predictor is almost the total variability in the responses

$\rightsquigarrow$ concretely, all the dots are very close to the straight line : the linear regression model fits the data very well

Now suppose that $ss_t \simeq ss_e$ and $ss_r \simeq 0$: almost the whole variation in the responses is due to the error terms

$\rightsquigarrow$ the dots are very far away from the fitted straight line, the regression model is merely useless

$\rightsquigarrow$ **comparing ss_r to ss_t allows to judge the adequacy of the model**

The quantity r^2, called the coefficient of determination, defined by

$$\boxed{r^2 = \frac{ss_r}{ss_t},}$$

represents the proportion of the variability in the responses that is explained by the predictor

Coefficient of determination

Clearly, the coefficient of variation will have a value between 0 and 1 :

- a value of r^2 near 1 indicates a good fit to the data
- a value of r^2 near 0 indicates a poor fit to the data

Fact

If the regression model is able to explain most of the variation in the response data, then it is considered to fit the data well, and is regarded as a 'good' model

In our running example, we find in the regression output on Slide 25 a value of r^2 (`R-Sq`) equal to 87.74%

$\rightsquigarrow$ almost 88% of the variation of y the level was used. The remaining 12% of the variation is due to the natural variability

Here r^2 is quite close to 1, which makes our model quite reliable

Correlation

Correlation

On Slide 18 of Chapter 3.4, we introduced the correlation coefficient between two random variables X and Y :

$$\rho = \frac{\mathbb{C}\text{ov}(X, Y)}{\sqrt{\mathbb{V}\text{ar}(X)\,\mathbb{V}\text{ar}(Y)}}$$

This coefficient quantifies the strength of the linear relationship between X and Y

$\rightsquigarrow$ if ρ is close to 1 or -1, there is a strong linear relationship between X and Y

$\rightsquigarrow$ observations in a random sample $\{(x_i, y_i), i = 1, \ldots, n\}$ drawn from (X, Y) should fall close to a straight line

$\rightsquigarrow$ a linear regression model linking Y to X, based on that sample, should be a good model, with a value of r^2 close to 1

$\rightsquigarrow$ **true**

Correlation

We can write :

$$r^2 = \frac{ss_r}{ss_t} = \frac{ss_t - ss_e}{s_{yy}} = \frac{s_{xx}(ss_t - ss_e)}{s_{xx}s_{yy}} = \frac{s_{xy}^2}{s_{xx}s_{yy}}$$
$$= \frac{\left(\sum_i (x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}$$

$\rightsquigarrow$ we observe that $\boxed{r = \frac{\left|\sum_i (x_i - \bar{x})(y_i - \bar{y})\right|}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}}$

$\rightsquigarrow$ **except for its sign** (positive or negative linear relationship), the sample correlation is the square root of the coefficient of determination (its sign is the sign of $\widehat{b}_1$)

In our running example, the sample correlation coefficient is $\sqrt{0.8774} = 0.9366$ (good estimate of the 'true' correlation coefficient between x level and y)

Objectives

Now you should be able to :

- Use simple linear regression for building models to engineering and scientific data
- Understand how the method of least squares is used to estimate the regression parameters
- Analyse residuals to determine if the regression model is an adequate fit to the data and to see if any underlying assumptions is violated
- Test statistical hypotheses and construct confidence intervals on regression parameters
- Use the regression model to make a prediction of a future observation and construct an appropriate prediction interval
- Understand how the linear regression model and the correlation coefficient are related