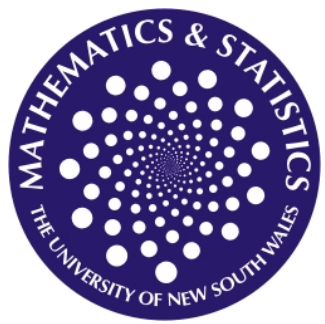
Statistics

MATH2089



Semester 1, 2018 – Lecture 11

This lecture

10. Regression Analysis

Additional reading:

Sections 3.1, 3.2, 3.3, 11.1 (pp.487-496), 11.2 (pp.501-505), 11.3 and 11.6 (pp.538-540) in the textbook (2nd edition)

Sections 3.1, 3.2, 3.3, 11.1 (pp.503-512), 11.2 (pp.517-520), 11.3 and 11.6 (pp.555-558) in the textbook (3rd edition)

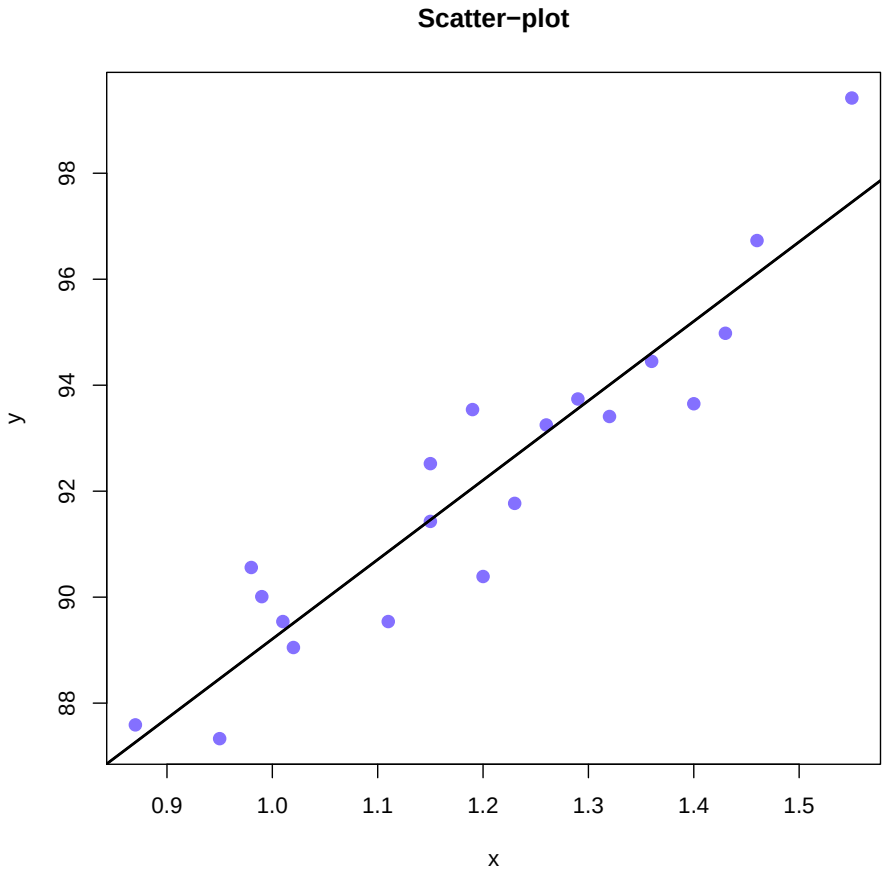
Introduction

- The main objective of many statistical investigations is to **make predictions**, preferably on the basis of mathematical equations
- For instance, an engineer may wish to predict the amount of oxide that will form on the surface of a metal baked in an oven for one hour at 200°C, or the amount of deformation of a ring subjected to a certain compressive force, or the number of miles to wear out a tire as a function of tread thickness and composition
- 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 techniques**

Introduction

As an illustration, consider the following data, where y_i's are the observed purity of oxygen produced in a chemical distillation process, and x_i's are the observed corresponding percentage of hydrocarbons that are present in the main condenser of the distillation unit

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 model

- Inspection of the scatter-plot indicates that, although no curve will pass exactly through all the points, there is a strong indication that the points lie scattered randomly around a straight line
- Therefore, it is reasonable to assume that the random variables X (hydrocarbon concentration) and Y (oxygen purity) 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 the regression coefficients
- The term ε is the random error, 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 arises from a theoretical relationship, at other times the choice of the model is 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 $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}\text{ar}(\varepsilon) = \sigma^2$
- Suppose we fix $X = x$. 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 $\mathbb{V}\text{ar}(\varepsilon) = \sigma^2$

$\rightarrow$ 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 is called the **regression function** (or regression line) and will be denoted

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

$\rightarrow$ the slope β_1 is the change in mean of Y for one unit change in X, the intercept β_0 is the mean value of Y when $X = 0$

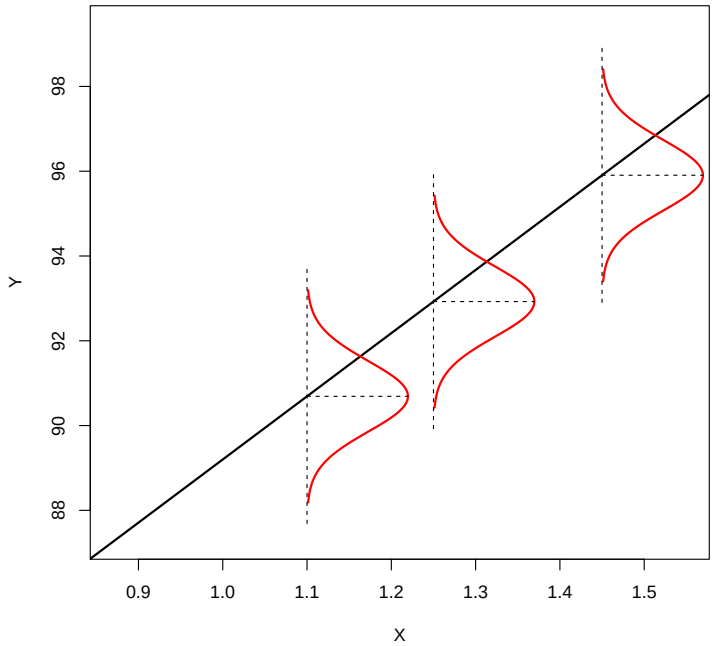
Simple linear regression model

Most of the time, the random error is supposed to be normally distributed: $\boxed{\varepsilon \sim \mathcal{N}(0, \sigma)}$

It follows that, for any fixed value x for X,

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

$\rightarrow$ the standard deviation σ tells to which extent the observations deviate from the regression line



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.

$\rightarrow$ they are **population parameters** which must be estimated from sample data

Here the random sample consists of n pairs of observations (X_i, Y_i), assumed to be independent of one another and 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 must be estimated by a sample version

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

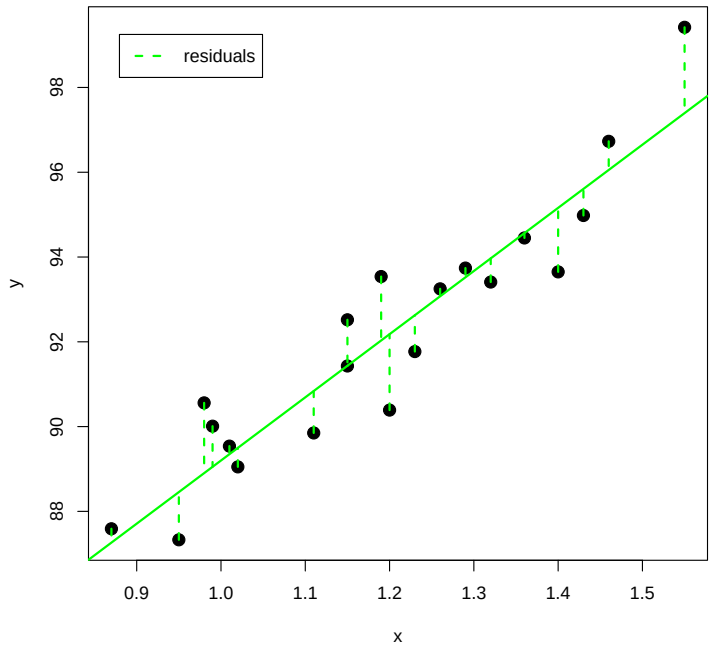
The question is how to determine the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ (and then an estimator for σ).

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 fitted 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

For any "candidate" straight line $Y = a + bX$, write
$R(a, b) = \sum_{i=1}^{n} (Y_i - (a + bX_i))^2$

Then,

$$\frac{\partial R}{\partial a}(a, b) = -2 \sum_{i=1}^{n} (Y_i - (a + bX_i))$$

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

$\rightarrow$ the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ should be the solutions of the equations

$$\begin{cases} \sum_i (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)) = 0 \\ \sum_i (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)) X_i = 0 \end{cases}$$

which are

$$\hat{\beta}_1 = \frac{\sum_i X_i Y_i - \frac{(\sum_i X_i)(\sum_i Y_i)}{n}}{\sum_i X_i^2 - \frac{(\sum_i X_i)^2}{n}} \quad \text{and} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

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 notation

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

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

we have:

Least squares estimators of β_0 and β_1

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

Note: as $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$, the estimated straight line will always go through the point $(\bar{x}, \bar{y})$, the centre of gravity of the scatter-plot

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 $\hat{b}_1$ and $\hat{b}_0$ of β_1 and β_0:

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

The estimated or fitted regression line is therefore $\hat{b}_0 + \hat{b}_1 x$, which is an estimate of $\mu_{Y|X=x}$

Now, we know that estimates of means are also typically used for prediction of future observation $\rightarrow \hat{b}_0 + \hat{b}_1 x$ is also used for predicting the future observation of Y when X is set to x, and is often denoted $\hat{y}(x)$:

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

Least Squares Estimation: example

Example

Fit a simple linear regression model to the data shown on Slide 4.

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

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

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

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

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

which is the straight line shown on Slide 9

Using this model, we would predict a mean oxygen purity of 89.23% when the hydrocarbon level is $x = 1\%$.

Also, the model indicates that the mean oxygen purity would increase by 14.947% for each unit increase (1%) in hydrocarbon level.

Estimating σ^2

The variance σ^2 of the error term $\varepsilon = Y - (\beta_0 + \beta_1 X)$ is another unknown parameter

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

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

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

$\rightarrow$ a natural estimator for σ^2 is the sample variance of the residuals

First, it can be checked that

$$\bar{e} = \frac{1}{n} \sum_{i=1}^{n} \hat{e}_i = \bar{y} - (\hat{b}_0 + \hat{b}_1 \bar{x}) = 0$$

(by definition of the estimated coefficient $\hat{b}_0$ and $\hat{b}_1$)

Estimating σ^2

Also, recall that the number of degrees of freedom for the usual sample variance is $n-1$ because we have to estimate one parameter ($\bar{x}$ estimates the true μ)

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

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

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

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

which is the observed value taken by the estimator

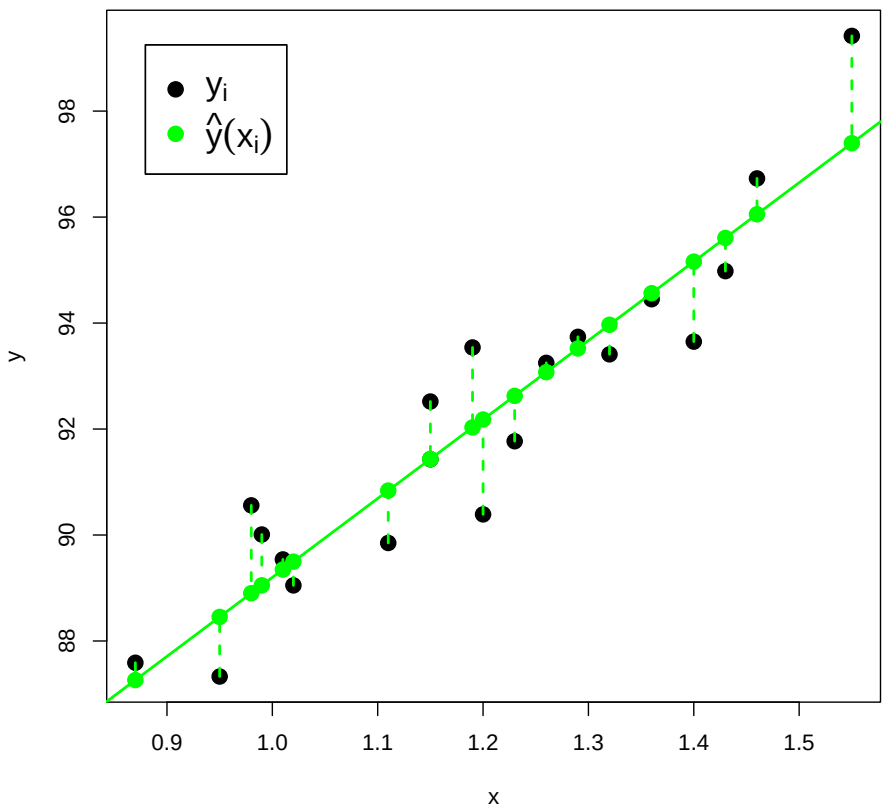
$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

It is clear that S is an estimator for σ.

Estimating σ^2: example

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

i	x_i	y_i	$\hat{y}(x_i)$	$\hat{e}_i$
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: $s^2 = \frac{1}{18}\sum_{i=1}^{20} \hat{e}_i^2 = 1.1824$ (%^{2}) $\rightarrow s = \sqrt{1.1824} = 1.0874$ (%)

Fixed design

From now on we will assume that the value of the x_i's have been chosen before the experiment is performed, and are therefore fixed

$\rightarrow$ this is known as a fixed design

So, only the Y_i's are random, and that substantially simplifies the coming developments, in particular the derivation of the sampling properties of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

Properties of the Least Squares Estimators

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

Then, because $\sum_i (x_i - \bar{x}) = 0$, we can write

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

$\rightarrow$ which is a linear combination of the normal random variables Y_i, therefore $\hat{\beta}_1$ is normally distributed!

Its expectation is

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

$\rightarrow$ **unbiased** estimator of β_1

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

Hence, the sampling distribution of $\hat{\beta}_1$ is

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

Properties of the Least Squares Estimators

Now, we can write

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

which is again a linear combination of normal r.v.'s (the Y_i's and $\hat{\beta}_1$)

$\rightarrow$ the estimator $\hat{\beta}_0$ is also normally distributed! Its expectation is

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

$\rightarrow$ **unbiased** estimator of β_0

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

Hence, the sampling distribution of $\hat{\beta}_0$ is

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

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 $\beta_1 = 0$

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

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

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

$\rightarrow$ we reject H_0 when the estimate $\hat{b}_1$ is 'too different' to 0

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

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

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

$$\boxed{\sqrt{s_{xx}}\,\frac{\hat{\beta}_1 - \beta_1}{S} \sim t_{n-2}}$$

Inferences concerning β_1

From this result, all the inferential procedures that we introduced previously can be readily adapted

At significance level α, the rejection criterion for $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$ is

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

with the estimated standard deviation $s = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} \hat{e}_i^2}$ (Slide 16)

and from the observed value of the test statistic under H_0 (i.e. with $\beta_1 = 0$)

$$t_0 = \sqrt{s_{xx}} \frac{\hat{b}_1}{s}$$

we can compute the p-value

$$p = 1 - \mathbb{P}(T \in [-|t_0|, |t_0|]) = 2 \times \mathbb{P}(T > |t_0|)$$

where T is a r. v. with distribution t_{n-2}

Inferences concerning β_1

In addition to the point estimate $\hat{b}_1$ of the slope, it is also possible to obtain a confidence interval for the 'true' slope β_1.

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

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

or equivalently

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

From an observed sample for which we find s and $\hat{b}_1$, a two-sided $100\times(1-\alpha)\%$ confidence interval for the parameter β_1 is

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

Inferences concerning β_0

Although of less practical interest, inferences concerning the parameter β_0 can be made in exactly the same way from the sampling distribution of $\hat{\beta}_0$.

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

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

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

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

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

$$t_0 = \frac{\hat{b}_0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}}} \qquad \rightarrow p = 2 \times \mathbb{P}(T > |t_0|), \quad T \sim t_{n-2}$$

Inferences concerning β_1: example

Example

Test for significance of the simple linear regression model for the data shown on Slide 4 at level $\alpha = 0.01$. (**Hint:** You can use the following Matlab outputs: $\texttt{tinv}(0.995, 18) = 2.878$, $\texttt{tcdf}(11.35, 18) = 1$)

Inferences concerning β_1: example

Simple linear regression: computer output

All statistical software programs include a least squares fit of a straight line. A typical output is as follows:

```
Regression Analysis:  Y versus X

The regression equation is Y = 74.283 + 14.947 X

 Predictor     Coef  SE Coef       T       P
 Constant   74.283    1.593   46.62   0.000
 X          14.947    1.317   11.35   0.000

S = 1.087 R-Sq = 87.74% R-Sq(adj) = 87.06%
```

The first row of the table (`Constant`) refers to the intercept (β_0), the second (`X`) to the predictor X (β_1).

The column `Coef` is for the estimates of the coefficients ($\hat{b}_0$ and $\hat{b}_1$), the column `SE Coef` is for the (estimated) standard error of these estimates ($s\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}}$ and $\frac{s}{\sqrt{s_{xx}}}$), the column `T` is for the observed values t_0 of the test statistics (when testing $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$), and the column `P` gives the associated p-values. Finally, `S` is the estimate s of σ.

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'

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

We have an estimator for this parameter:

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

Note that, as a linear combination of normal random variables, the estimator $\hat{\mu}_{Y|X=x}$ is also normally distributed. Its expectation is:

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

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

Confidence Interval on the Mean Response

Its variance can be found to be

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

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

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

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

Finally, the sampling distribution of the estimator $\hat{\mu}_{Y|X=x}$ is

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

Confidence Interval on the Mean Response

If we standardise and replace the unknown σ by its estimator S, we get (as usual):

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

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

From an observed sample for which we find s and $\hat{y}(x)$ from the fitted model $\hat{y}(x) = \hat{b}_0 + \hat{b}_1 x$, a two-sided $100 \times (1-\alpha)\%$ confidence interval for the parameter $\mu_{Y|X=x}$, that is the mean response Y when $X = x$, is

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

Confidence Interval on the Mean Response: example

Example

Construct a 95% confidence interval on the mean oxygen purity $\mu_{Y|X=x}$ when the hydrocarbon level X is fixed to $x = 1\%$ (from the data shown on Slide 4).

The fitted model was $\hat{y}(x) = 74.283 + 14.947x$. We also have $n = 20$, $s = 1.0874$, $s_{xx} = 0.68088$ and $\bar{x} = 1.1960$. From Matlab, we find $t_{18;0.975} = 2.101$.

When $x = 1$, the model estimates the mean response $\mu_{Y|X=1}$ at $\hat{y}(1) = 89.23$

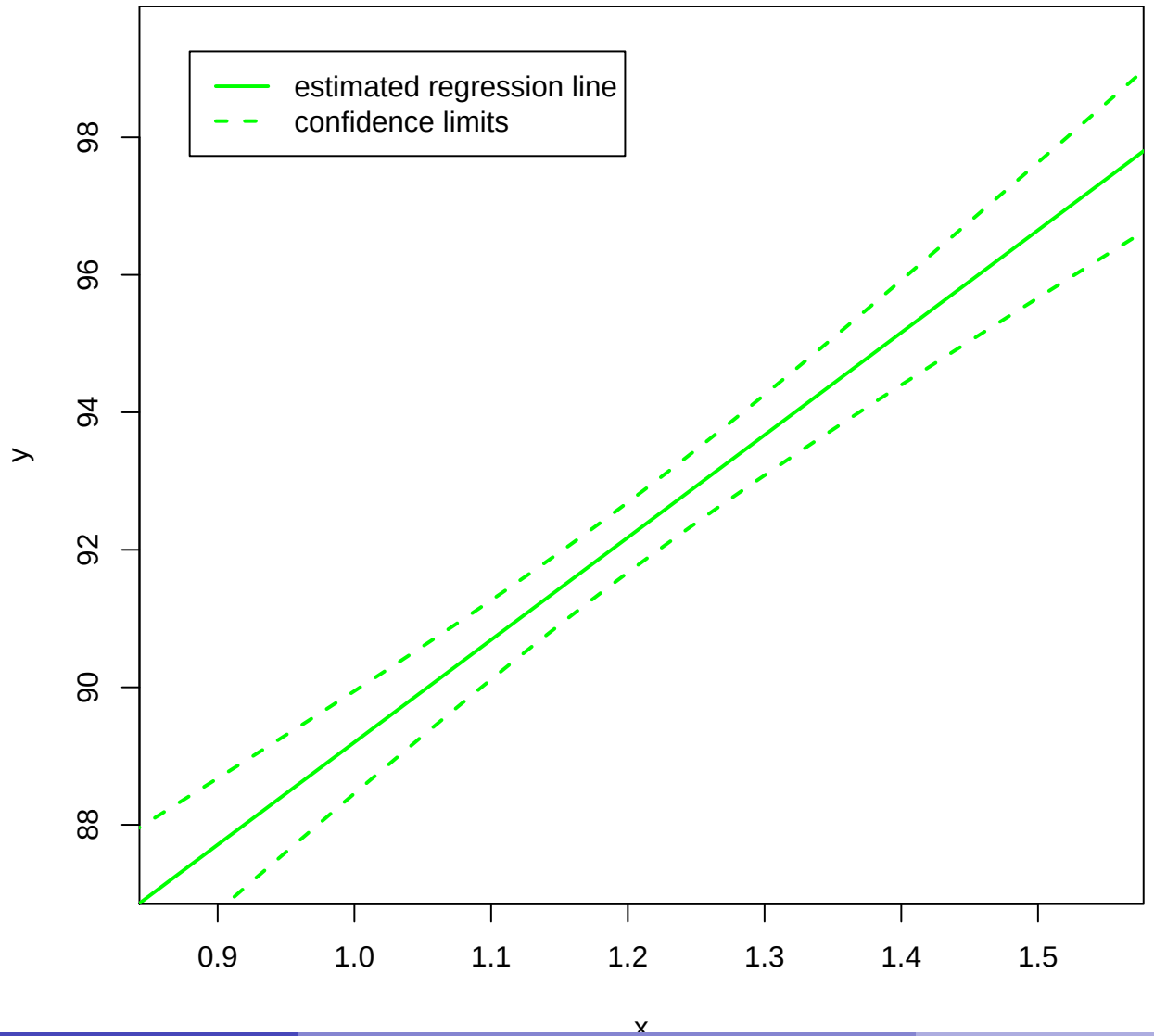
$\rightarrow$ a 95% confidence interval for $\mu_{Y|X=1}$ is given by

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

$\rightarrow$ when $x = 1\%$, we are 95% confident that the true mean oxygen purity 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 $\mu_{Y|X=x}$



Prediction of new observations

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

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

(recall Section 7.7, Lecture 8)

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

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

for which a predicted value is

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

once the model has been fitted from an observed sample

$\rightarrow$ the predictor of Y at $X = x$ is the estimator of $\mu_{Y|X=x}$!

(compare Slide 21 Week 8)

The prediction error is given by $Y|(X = x) - Y^*(x)$ and is normally distributed, as both $Y|(X = x)$ and $Y^*(x)$ are as well.

Prediction of new observations

As $Y|(X=x) \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma)$ (Slide 7) and $Y^*(x) = \hat{\mu}_{Y|X=x} \sim \mathcal{N}\left(\beta_0 + \beta_1 x, \sigma\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{s_{xx}}}\right)$ (Slide 29), the expectation of the prediction error is

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

$\rightarrow$ on average, the predictor will 'guess' the right value

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

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

Prediction of new observations

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

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

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

From an observed sample for which we find s and $\hat{y}(x)$ from the fitted model $\hat{y}(x) = \hat{b}_0 + \hat{b}_1 x$, a two-sided $100 \times (1-\alpha)\%$ prediction interval for a new observation Y at $X = x$ is

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

Prediction of new observations: remarks

Adapting the remarks on Slide 27 Lecture 8, 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 $X = x$
   $\rightarrow$ 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
   $\rightarrow$ 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 ($n \rightarrow \infty$), 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 the oxygen purity Y when the hydrocarbon level X is fixed to $x = 1\%$ (from the data shown on Slide 4).

The fitted model was $\hat{y}(x) = 74.283 + 14.947x$. We also have $n = 20$, $s = 1.0874$, $s_{xx} = 0.68088$ and $\bar{x} = 1.1960$. From Matalb, we find $t_{18;0.975} = 2.101$.

When $x = 1$, the model estimates the mean response $\mu_{Y|X=1}$ to $\hat{y}(1) = 89.23$

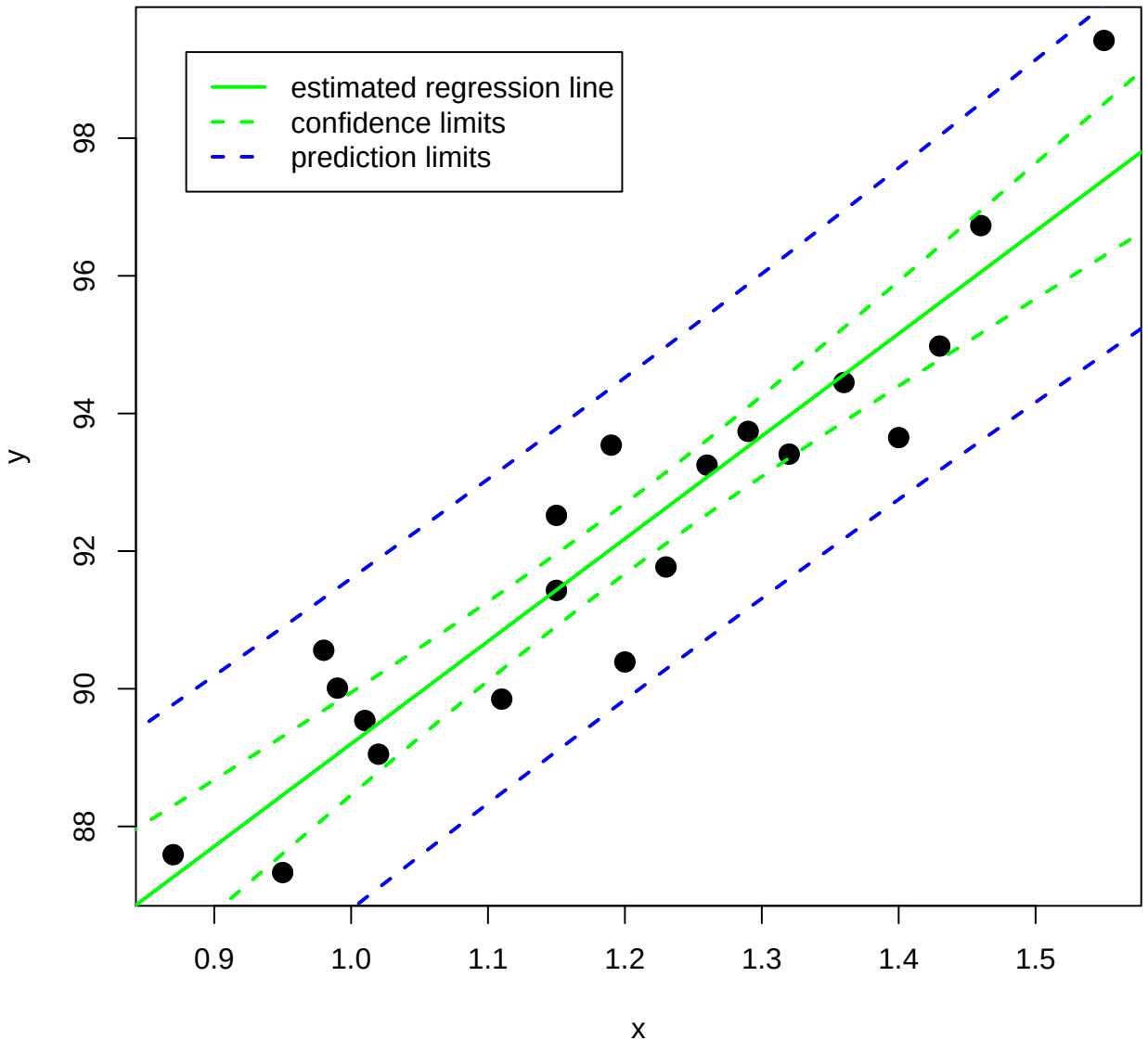
$\rightarrow$ a 95% prediction interval for Y is given by

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

$\rightarrow$if we fix the hydrocarbon level to $x = 1\%$, we can be 95% confident that the next observed value of the oxygen purity 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

In the course of fitting and analysing 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 $Y = \beta_0 + \beta_1 X + \varepsilon \rightarrow$ **scatterplot**

The other central assumption is certainly that (Slide 8)

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

which has several implications. 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 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 true e_i's (as we do not know β_0 and β_1).

However, the observed residuals of the fitted model

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

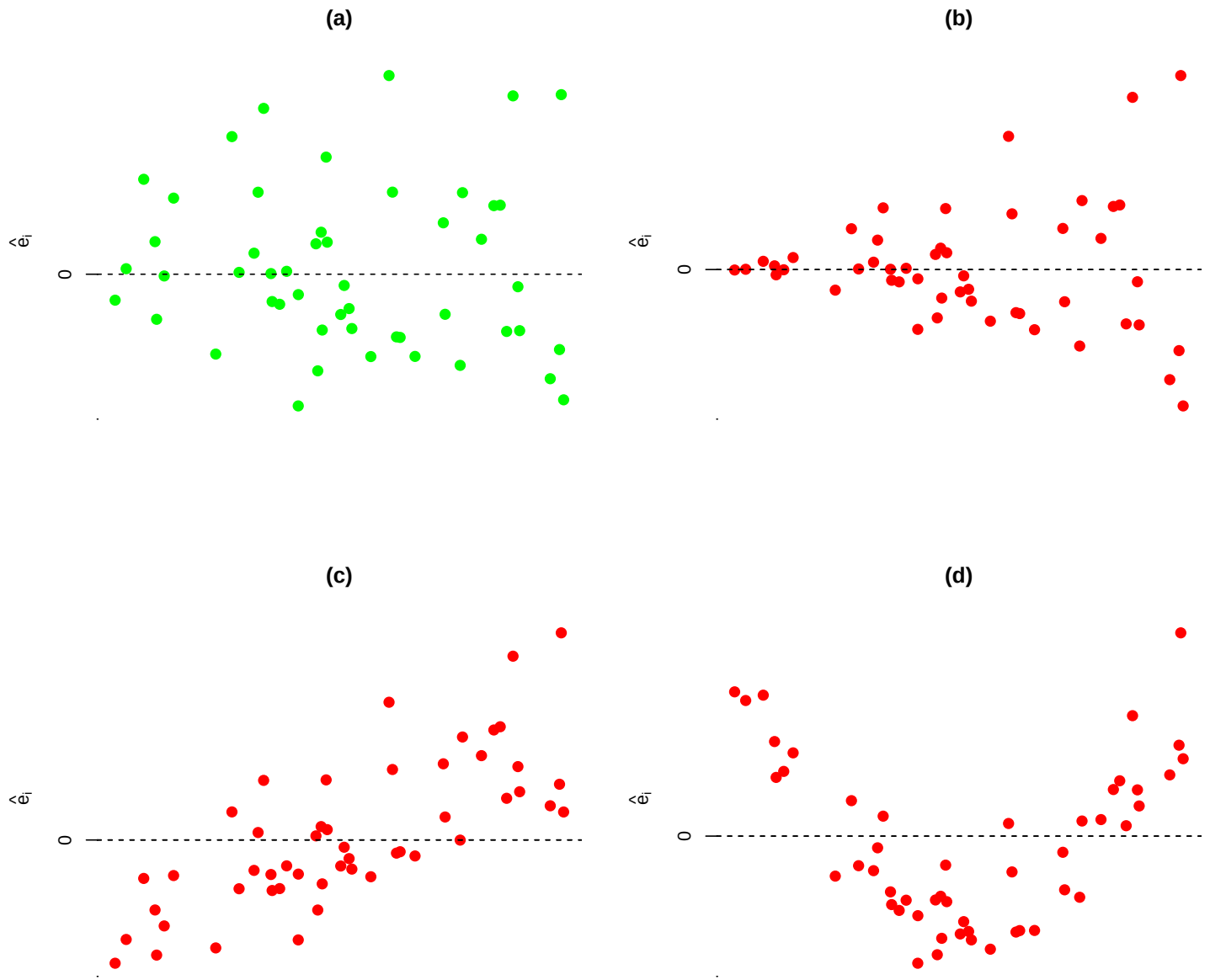
are probably good estimates of those e_i's $\rightarrow$ **residual analysis**

It is frequently helpful to plot the residuals (1) in time sequence (if known), (2) against the fitted values $\hat{y}(x_i)$, 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 $\rightarrow$ no information should be observed in the residuals, they should look like noise.

Residual analysis

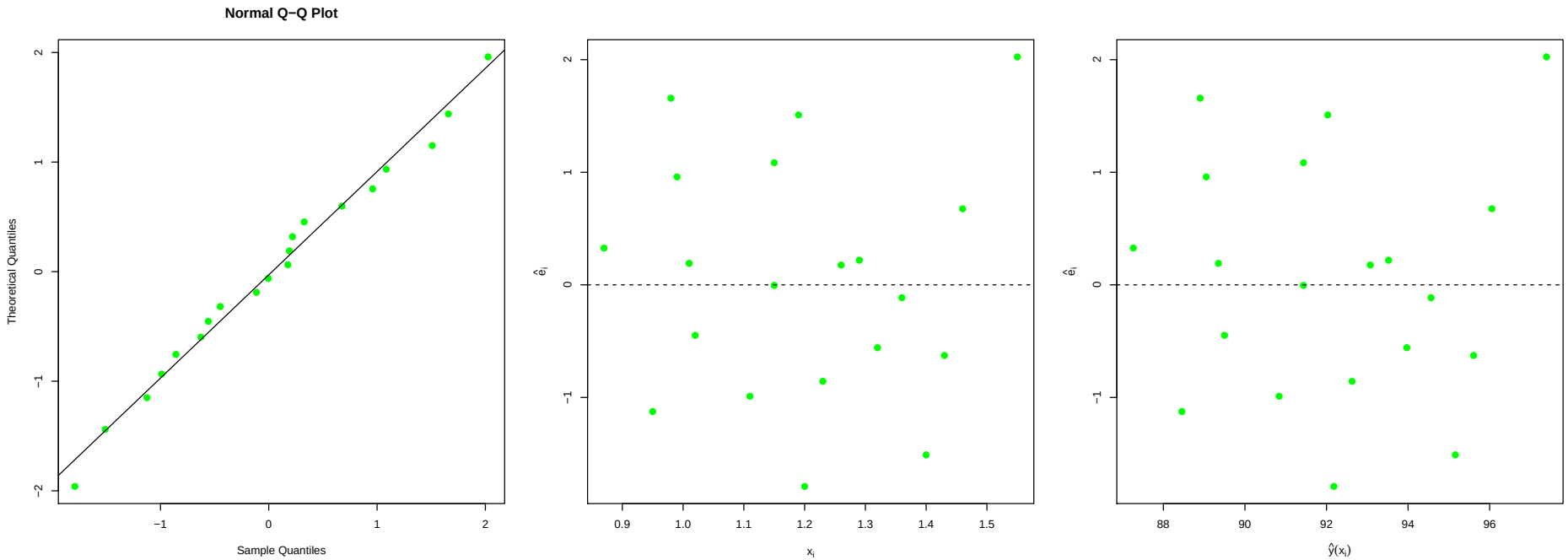


Residual analysis

- **Pattern (a) represents thus the ideal situation** (*nothing to report*)
- In (b), the variance of the error terms e_i (and thus that of the responses Y_i) seems to be increasing with time or with magnitude of Y_i or X_i
- Plot (c) indicates some sort of dependence in the error terms (when plotted against time)
- In (d), we get clear indication of model inadequacy: the residuals are systematically positive for extreme values and negative for medium values $\Rightarrow$ the model is not complete, there is still much information in the residuals: higher-order terms (like 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 the normality assumption

Residual analysis: example

From our running example (oxygen purity data), a normal quantile plot of the residuals and plots against the predicted values $\hat{y}(x_i)$ and against the hydrocarbon levels x_i for the residuals computed on Slide 17, are shown below:



$\rightarrow$ nothing to report

$\rightarrow$ the assumptions we made look totally valid

Variability decomposition

Similarly to the notations on Slide 11, we can define

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

$\rightarrow$ 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 two factors:

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

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

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

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

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

$\rightarrow$ all the dots are very close to the straight line, the predictions are very accurate: the linear regression model fits the data very well

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

$\rightarrow$the dots are very far away from the fitted straight line, the predictions are very imprecise: the regression model is useless

$\rightarrow$ **comparing ss_r to ss_t allows us to judge the model adequacy**

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

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

represents the proportion of the variability in the responses that is explained by the predictor and hence taken into account in the model.

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 27 a value of r^2 (`R-Sq`) is equal to 87.74%

$\rightarrow$ almost 88% of the variation of the oxygen purity is explained by the level of hydrocarbons that was used. The remaining 12% of the variation is due to the natural variability in the oxygen purity even when the hydrocarbon level is fixed to a given level

Here r^2 is quite close to 1, which makes our model a good one.

Correlation

In Lecture 4, we introduced the correlation coefficient between two random variables X and Y:

$$\rho = \frac{\mathbb{C}\text{ov}(X, Y)}{\sqrt{\mathbb{V}\text{ar}(X)\,\mathbb{V}\text{ar}(Y)}} = \frac{\mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))}{\sqrt{\mathbb{E}((X - \mathbb{E}(X))^2)\mathbb{E}((Y - \mathbb{E}(Y))^2)}}$$

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

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

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

$\rightarrow$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

Correlation

We can write:

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

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

is the **sample correlation coefficient**, which can be regarded as the sample estimate of the population correlation coefficient ρ

$\rightarrow$ 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 $\hat{b}_1$)

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

Objectives

Now you should be able to:

- Use simple linear regression for building models for 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 ☐

Recommended exercises:

$\rightarrow$ Q7 p.104, Q13, Q15 p.114, Q21 p.126, Q1 p.499, Q5, Q8 p.500, Q13 p.507, Q17 p.508, Q19 (a-c) p.515 (2nd edition)

$\rightarrow$ Q7 p.107, Q13 p.116, Q15 p.117, Q1 p.514, Q6 p.515, Q9 p.516, Q14 p.523, Q19 p.524, Q22 (a-d) p.531 (3rd edition)