MAT 3375
Regression Analysis

Chapter 2
Simple Linear Regression

P. Boily (uOttawa)

Summer – 2023

Outline

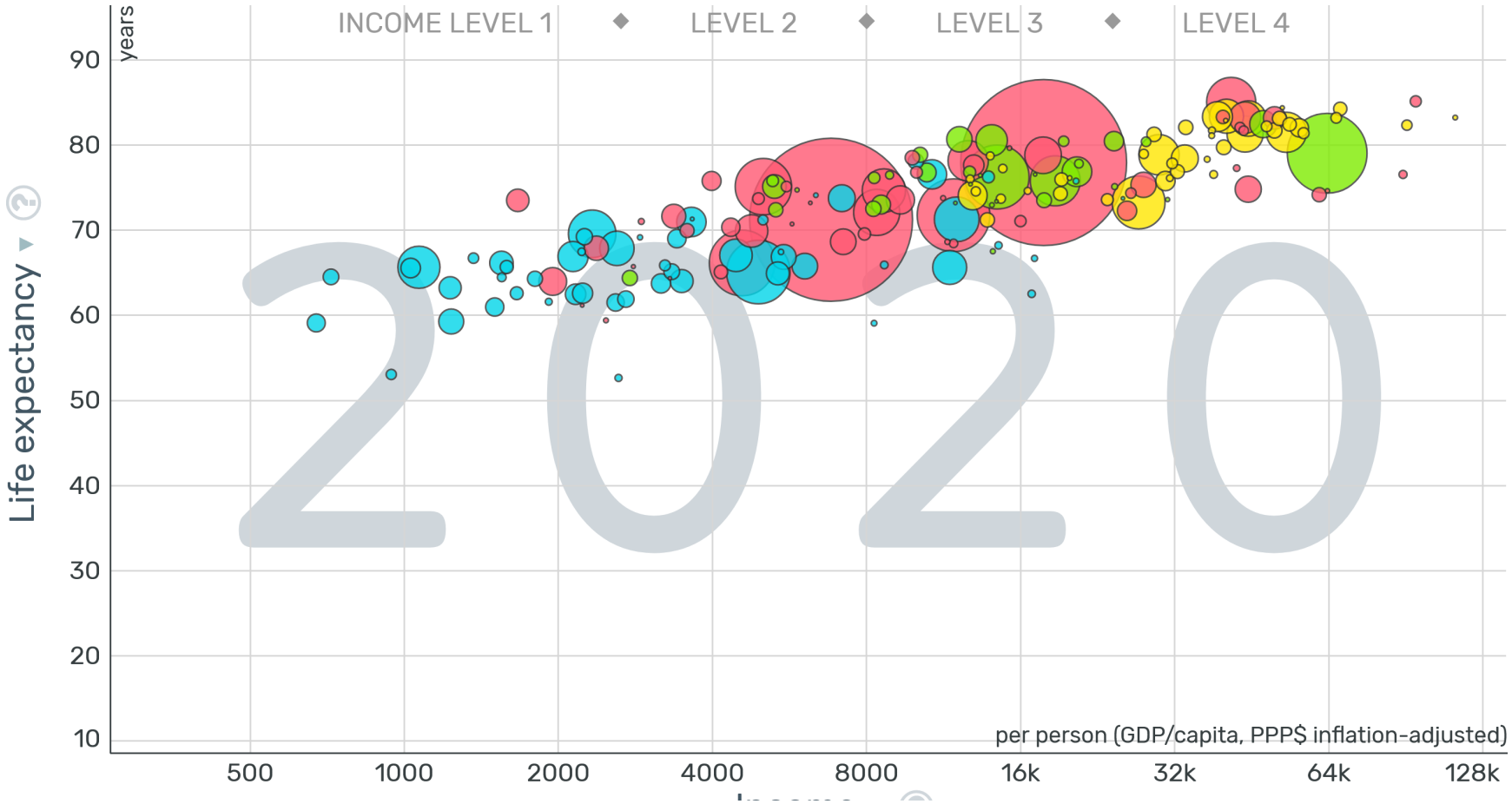
Outline (continued)

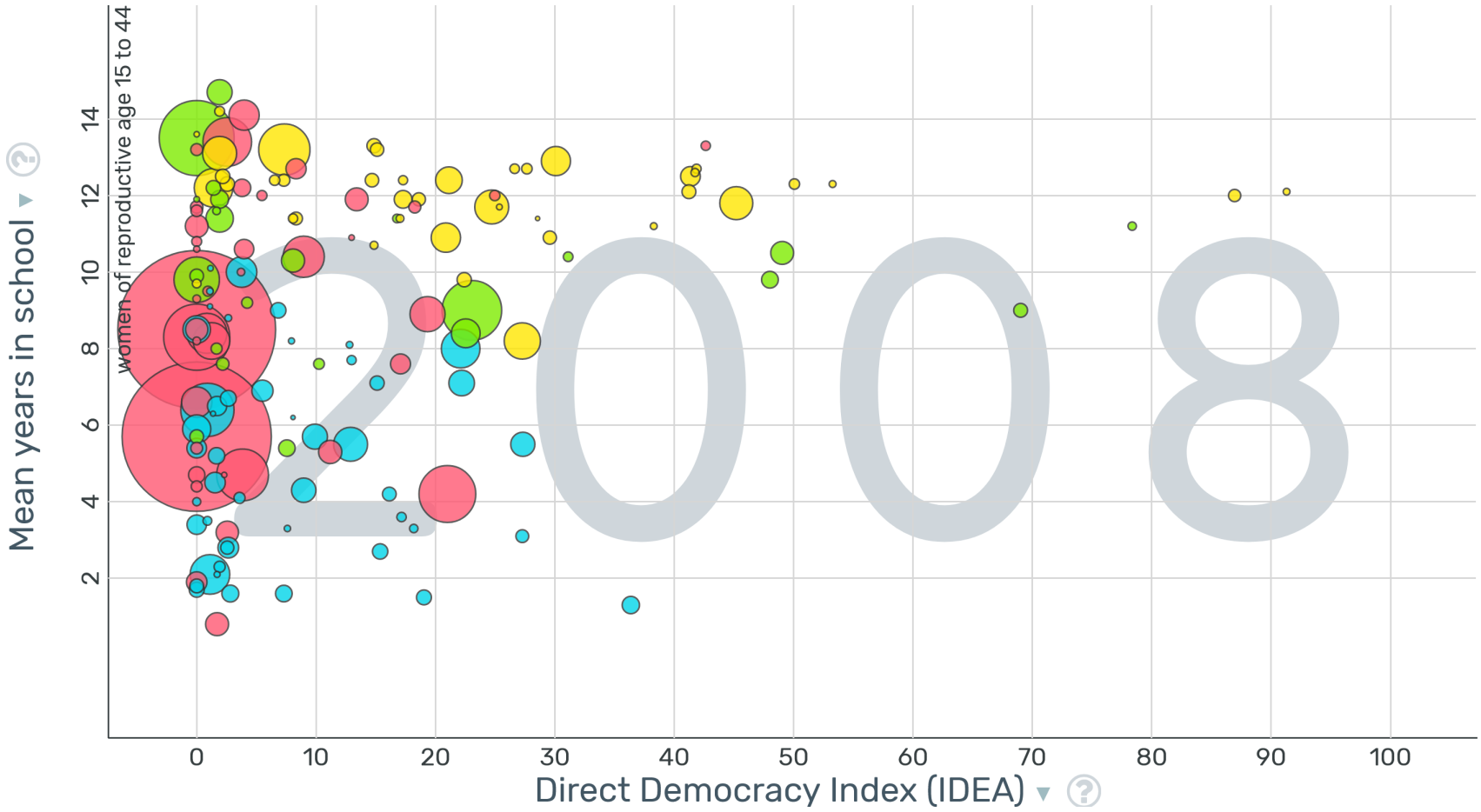
2 – Simple Linear Regression

We start by considering a simple scenario, with only two **continuous** variables: a **response** Y and a **predictor** X.

Examples:

- X: age; Y: height
- X: age; Y: salary
- X: income; Y: life expectancy
- X: number of sunlight hours; Y: plant biomass


INCOME LEVEL 1
LEVEL 2
LEVEL 3
LEVEL 4
years
Life expectancy
90
80
70
60
50
40
30
20
10
2020
per person (GDP/capita, PPP$ inflation-adjusted)
500
1000
2000
4000
8000
16k
32k
64k
128k


Mean years in school
women of reproductive age 15 to 44
2008
2
4
6
8
10
12
14
0
10
20
30
40
50
60
70
80
90
100
Direct Democracy Index (IDEA)

In theory, we hope that there might be a **functional relationship** $Y = f(X)$ between X and Y.

In practice (assuming that a relationship even exists), the best that we may be able to hope for is a **statistical relationship**

$$Y = f(X) + \varepsilon,$$

where

- $f(X)$ is the **response function**;
- ε is the **random error** (or noise).

In **simple linear regression**, we assume that the response function is $f(X) = \beta_0 + \beta_1 X$.

The building blocks of regression analysis are the **observations**:

$$(X_i, Y_i), \quad i = 1, \ldots, n.$$

In an ideal setting, these observations are **(jointly) randomly sampled**, according to some appropriate design (which is a topic for other courses).

The **simple linear regression model** (SLRM) is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where β_0, β_1 are **unknown parameters** (**which we want to find**) and ε_i is the **random error on the ith observation** (or case).

The SLRM **assumption on the error structure** is that $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

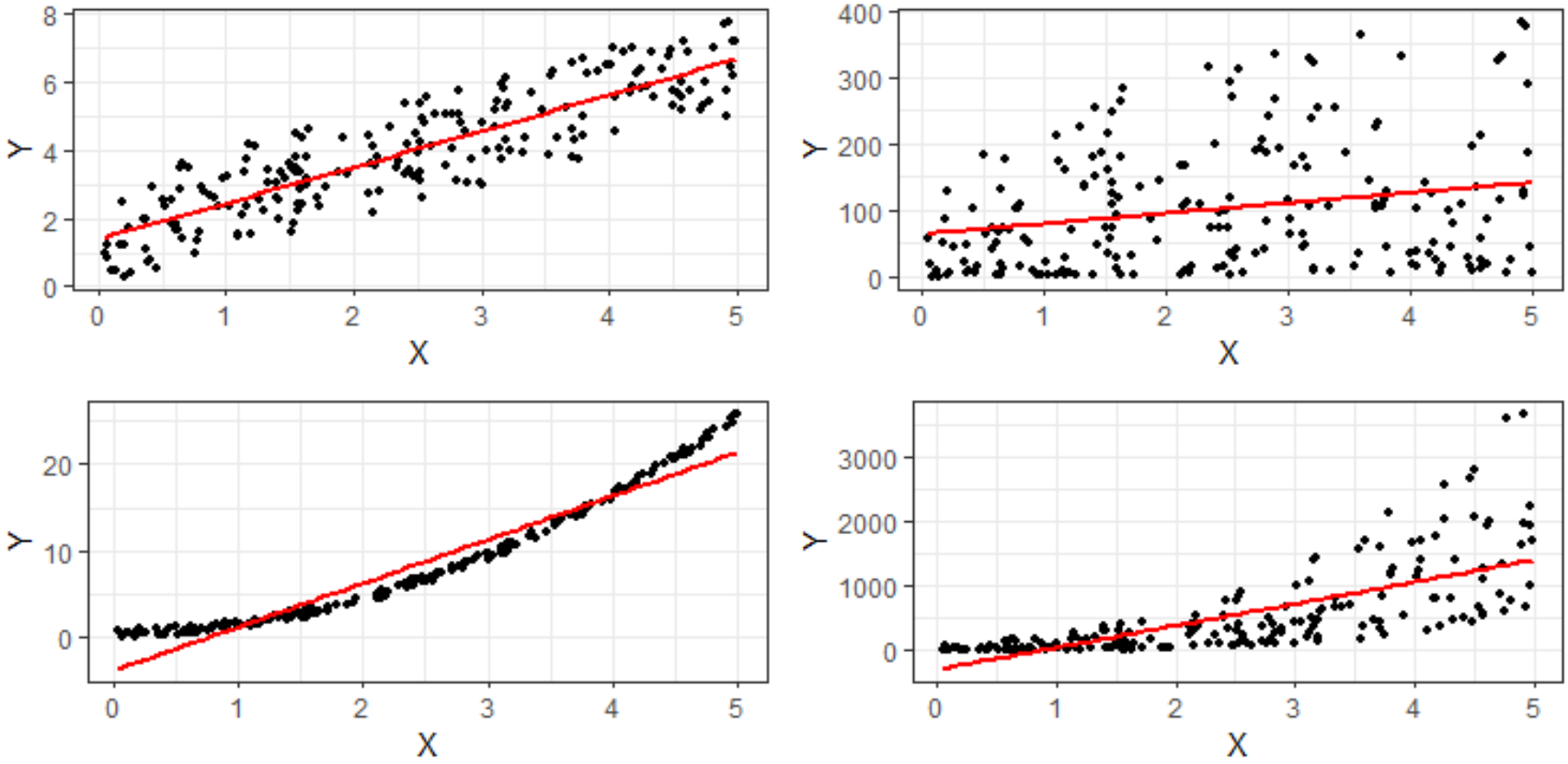
We are using matrix notation here to keep the assumption **compact**.

Let us unpack the statement. Write $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$.

Since $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$: we have

- $\mathrm{E}\{\boldsymbol{\varepsilon}\} = \mathbf{0} \implies \mathrm{E}\{\varepsilon_i\} = 0, \quad i = 1, \ldots, n;$
- $\sigma^2\{\boldsymbol{\varepsilon}\} = \sigma^2 \mathbf{I}_n \implies \sigma^2\{\varepsilon_i\} = \sigma^2, \quad i = 1, \ldots, n;$
- $\sigma^2\{\boldsymbol{\varepsilon}\} = \sigma^2 \mathbf{I}_n \implies \sigma\{\varepsilon_i, \varepsilon_j\} = 0, \quad$ for all $i \neq j$.

The errors $\{\varepsilon_i\}$ are thus **uncorrelated**, with **mean** 0 and **constant variance**. In other words, the **dispersion of observations is constant around the regression line**.


Y
X
0
2
4
6
8
0
1
2
3
4
5
Y
X
0
100
200
300
400
0
1
2
3
4
5
Y
X
0
10
20
0
1
2
3
4
5
Y
X
0
1000
2000
3000
0
1
2
3
4
5

2.1 – Least Squares Estimation

We treat the predictor values X_i as constant, for $i = 1, \ldots, n$ (i.e., we assume that there is **no measurement error**).

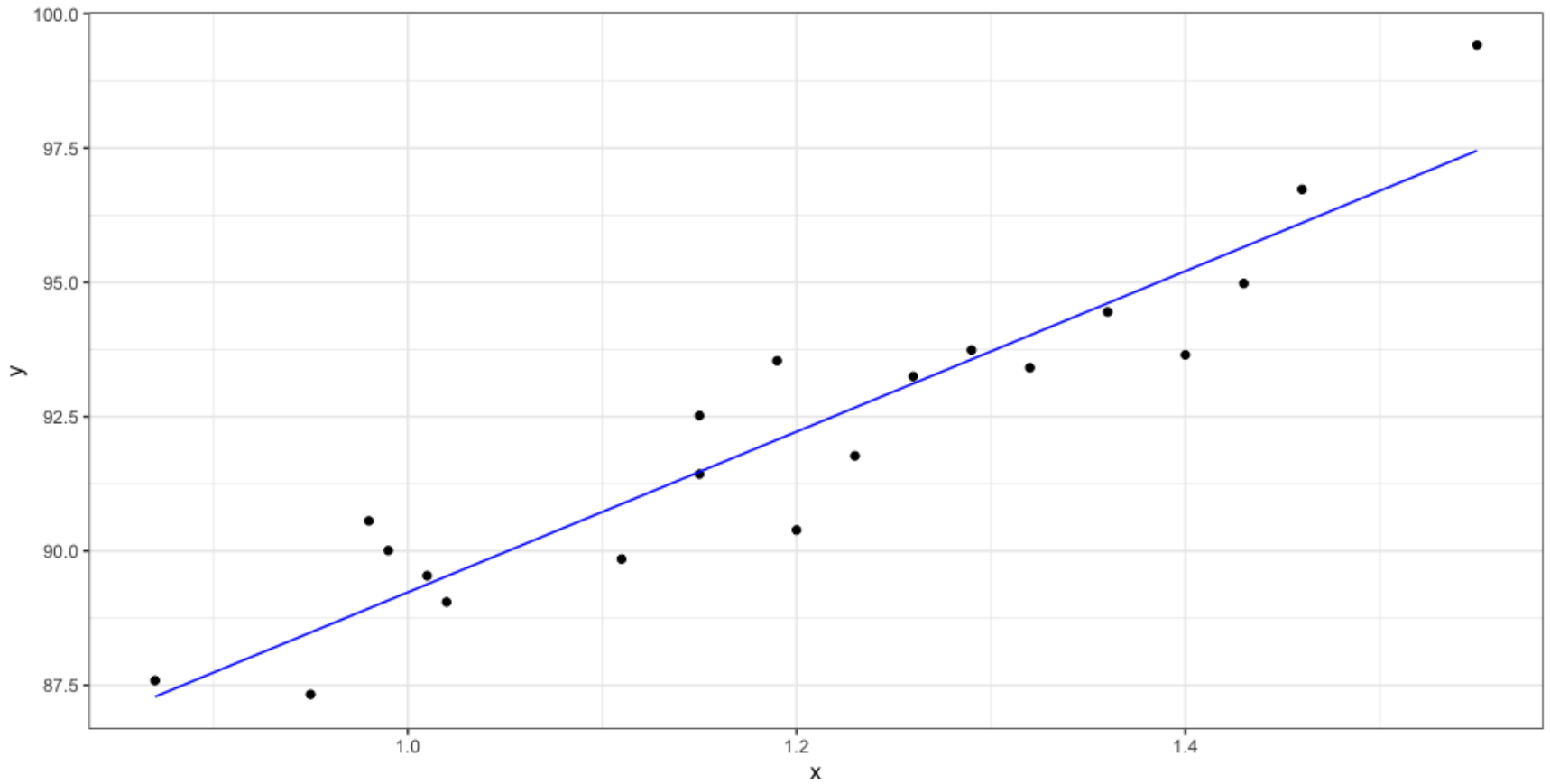
Since $\mathrm{E}\{\varepsilon_i\} = 0$, the **expected** (or mean) **response given** X_i is thus

$$\mathrm{E}\{Y_i|X_i\} = \mathrm{E}\{\beta_0 + \beta_1 X_i + \varepsilon_i|X_i\} = \beta_0 + \beta_1 X_i + \mathrm{E}\{\varepsilon_i\} = \beta_0 + \beta_1 X_i.$$

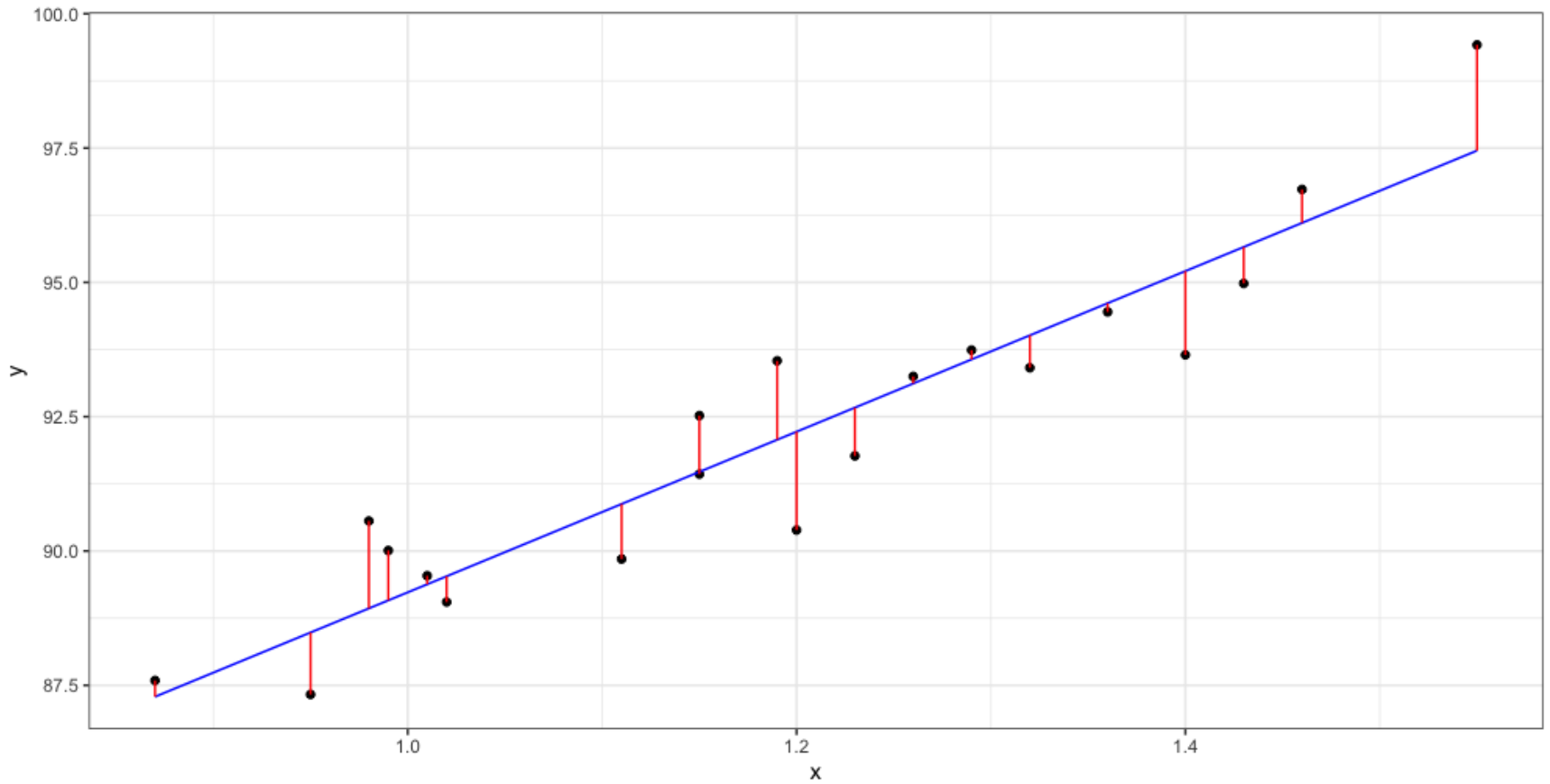
The **deviation at** X_i is the difference between the observed response Y_i and the expected response $\mathrm{E}\{Y_i|X_i\}$:

$$e_i = Y_i - \mathrm{E}\{Y_i|X_i\};$$

the deviation can be **positive** (if the point lies **above** the line) or **negative** (if it lies **below**).



line of best fit



deviations (residuals)

How do we find **estimators** for β_0 and β_1? Incidentally, how do we determine if the fitted line is a **good model for the data**?

Consider the function

$$Q(\boldsymbol{\beta}) = Q(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \mathrm{E}\left\{Y_i | X_i\right\}\right)^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

If $Q(\boldsymbol{\beta})$ is "small", then the sum of the **squared residuals** is "small", and so we would expect the line $Y = \beta_0 + \beta_1 X$ to be a good fit for the data.

The **least-square estimators** of the SLR problem are the pair $\mathbf{b} = (b_0, b_1)$ which minimizes the function Q with respect to $\boldsymbol{\beta} = (\beta_0, \beta_1)$.

We must then find critical points of $Q(\boldsymbol{\beta})$, i.e., solve $\nabla_{\boldsymbol{\beta}} Q(\mathbf{b}) = \mathbf{0}$.

Thus, we must solve:

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} = 2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \cdot (-1) = 0$$

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} = 2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \cdot (-X_i) = 0.$$

This is a linear system of two equations in the two unknowns β_0, β_1, known as the **normal equations**.

As such, it has either **no solution**, a **unique solution**, or **infinitely many solutions**.

Note: from now on, we drop the $\mid X_i$ when we use the $\mathrm{E}\{\cdot \mid X_i\}$.

2.1.1 – Normal Equations

These equations reduce to the following pair:

$$\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i, \qquad \sum_{i=1}^{n} X_i Y_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2.$$

If we use the following shorthand notation:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,\ \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,\ S_{xx} = \sum_{i=1}^{n}(X_i - \overline{X})^2,\ S_{xy} = \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}),$$

it is not too difficult to show that

$$\sum_{i=1}^{n} X_i^2 = S_{xx} + n\overline{X}^2 \quad \text{and} \quad \sum_{i=1}^{n} X_i Y_i = S_{xy} + n\overline{XY}.$$

With this notation, the normal equations further reduce to

$$n\overline{Y} = n\beta_0 + n\overline{X}\beta_1, \qquad S_{xy} + n\overline{XY} = n\overline{X}\beta_0 + (S_{xx} + n\overline{X}^2)\beta_1.$$

In matrix form, this can be written as:

$$\begin{bmatrix} 1 & \overline{X} \\ n\overline{X} & S_{xx} + n\overline{X}^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \overline{Y} \\ S_{xy} + n\overline{XY} \end{bmatrix}.$$

A linear system $A\boldsymbol{\beta} = \mathbf{v}$ has a unique solution $\boldsymbol{\beta} = A^{-1}\mathbf{v}$ if the determinant of the coefficient matrix A is non-zero. In this case, the determinant is

$$S_{xx} + n\overline{X}^2 - n\overline{XX} = S_{xx} > 0 \iff s_X^2 \neq 0.$$

The unique solution is thus

$$\begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix} = \begin{bmatrix}1 & \overline{X}\\ n\overline{X} & S_{xx}+n\overline{X}^2\end{bmatrix}^{-1}\begin{bmatrix}\overline{Y}\\ S_{xy}+n\overline{XY}\end{bmatrix}$$

$$= \frac{1}{S_{xx}}\begin{bmatrix}S_{xx}+n\overline{X}^2 & -\overline{X}\\ -n\overline{X} & 1\end{bmatrix}\begin{bmatrix}\overline{Y}\\ S_{xy}+n\overline{XY}\end{bmatrix}$$

$$= \frac{1}{S_{xx}}\begin{bmatrix}(S_{xx}+n\overline{X}^2)\overline{Y}-\overline{X}(S_{xy}+n\overline{XY})\\ -n\overline{XY}+S_{xy}+n\overline{XY}\end{bmatrix} = \begin{bmatrix}\overline{Y}-\overline{X}\cdot S_{xy}/S_{xx}\\ S_{xy}/S_{xx}\end{bmatrix}$$

Set $b_0=\beta_0$ and $b_1=\beta_1$. Then we may write:

$$b_1 = \frac{S_{xy}}{S_{xx}}\ \textbf{(slope)} \quad \text{and} \quad b_0 = \overline{Y}-b_1\overline{X}\ \textbf{(intercept)}.$$

By analogy with S_{xx} (the **total variation of the predictor**), we can also define the **total variation of the response** S_{yy}, a quantity that will play an important role in the course:

$$S_{yy} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2;$$

If the X_i are fixed, b_0, b_1 are **linear combinations** of the Y_i:

$$b_1 = \frac{1}{S_{xx}}\sum_{i=1}^{n}(X_i - \overline{X})Y_i - \frac{\overline{Y}}{S_{xx}}\underbrace{\sum_{i=1}^{n}(X_i - \overline{X})}_{=0} = \sum_{i=1}^{n}\frac{(X_i - \overline{X})}{S_{xx}}Y_i,$$

$$b_0 = \sum_{i=1}^{n}\frac{Y_i}{n} - \sum_{i=1}^{n}\frac{(X_i - \overline{X})}{S_{XX}}Y_i\overline{X} = \sum_{i=1}^{n}\left[\frac{1}{n} - \overline{X}\frac{(X_i - \overline{X})}{S_{XX}}\right]Y_i.$$

Properties of Least Squares Estimators

Both b_0, b_1 are **unbiased estimators** of their respective parameters. Indeed,

$$\begin{aligned}
\mathrm{E}\{b_1\} &= \mathrm{E}\left\{\sum_{i=1}^{n}\frac{(X_i-\overline{X})}{S_{xx}}Y_i\right\} = \sum_{i=1}^{n}\frac{(X_i-\overline{X})}{S_{xx}}\mathrm{E}\{Y_i\} \\
&= \sum_{i=1}^{n}\frac{(X_i-\overline{X})}{S_{xx}}\mathrm{E}\{\beta_0+\beta_1X_i+\varepsilon_i\} = \sum_{i=1}^{n}\frac{(X_i-\overline{X})}{S_{xx}}(\beta_0+\beta_1X_i+\mathrm{E}\{\varepsilon_i\}) \\
&= \sum_{i=1}^{n}\frac{(X_i-\overline{X})}{S_{xx}}(\beta_0+\beta_1X_i) = \frac{\beta_0}{S_{xx}}\underbrace{\sum_{i=1}^{n}(X_i-\overline{X})}_{=0}+\frac{\beta_1}{S_{xx}}\underbrace{\sum_{i=1}^{n}(X_i-\overline{X})X_i}_{=S_{xx}(?)} \\
&= 0+\beta_1 = \beta_1,
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\{b_0\} &= \mathrm{E}\left\{\overline{Y} - b_1\overline{X}\right\} = \mathrm{E}\left\{\overline{Y}\right\} - \mathrm{E}\left\{b_1\overline{X}\right\} = \mathrm{E}\left\{\overline{Y}\right\} - \mathrm{E}\left\{b_1\right\}\overline{X} \\
&= \mathrm{E}\left\{\frac{1}{n}\sum_{i=1}^{n} Y_i\right\} - \beta_1\overline{X} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left\{Y_i\right\} - \beta_1\overline{X} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left\{\beta_0 + \beta_1 X_i + \varepsilon_i\right\} - \beta_1\overline{X} = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 X_i) - \beta_1\overline{X} \\
&= \frac{\beta_0}{n}\sum_{i=1}^{n} 1 + \frac{\beta_1}{n}\sum_{i=1}^{n} X_i - \beta_1\overline{X} = \beta_0 + \beta_1\overline{X} - \beta_1\overline{X} = \beta_0.
\end{aligned}$$

Now is as good a time as any to illustrate these notions with an example.

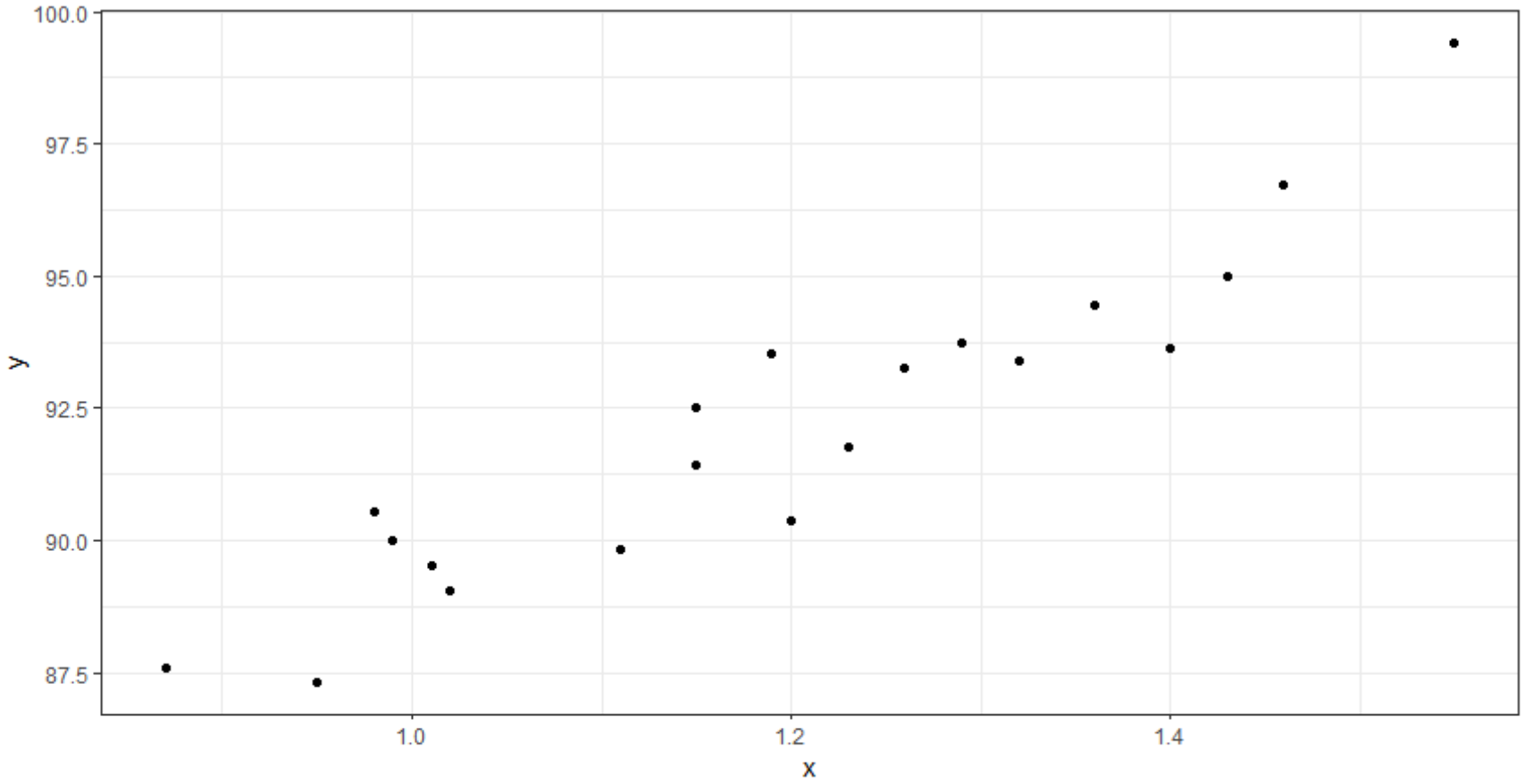
Fuels Example: Consider the following $n = 20$ paired measurements (X_i, Y_i) of hydrocarbon levels (X) and pure oxygen levels (Y) in fuels:

i	1	2	3	4	5	6	7	8	9	10
X_i	0.99	1.02	1.15	1.29	1.46	1.36	0.87	1.23	1.55	1.40
Y_i	90.01	89.05	91.43	93.74	96.73	94.45	87.59	91.77	99.42	93.65
i	11	12	13	14	15	16	17	18	19	20
X_i	1.19	1.15	0.98	1.01	1.11	1.20	1.26	1.32	1.43	0.95
Y_i	93.54	92.52	90.56	89.54	89.85	90.39	93.25	93.41	94.98	87.33

Is the simple regression model valid? If so, fit the data to the model.

Solution: we start by computing the basic sums:

$$\sum_{i=1}^{20} X_i = 23.92, \quad \sum_{i=1}^{20} Y_i = 1843.21, \quad \sum_{i=1}^{20} X_i^2 = 29.29, \quad \sum_{i=1}^{20} X_i Y_i = 2214.66$$



Is the SLR model valid?

Since the SLR model appears valid, we compute the least-square estimators:

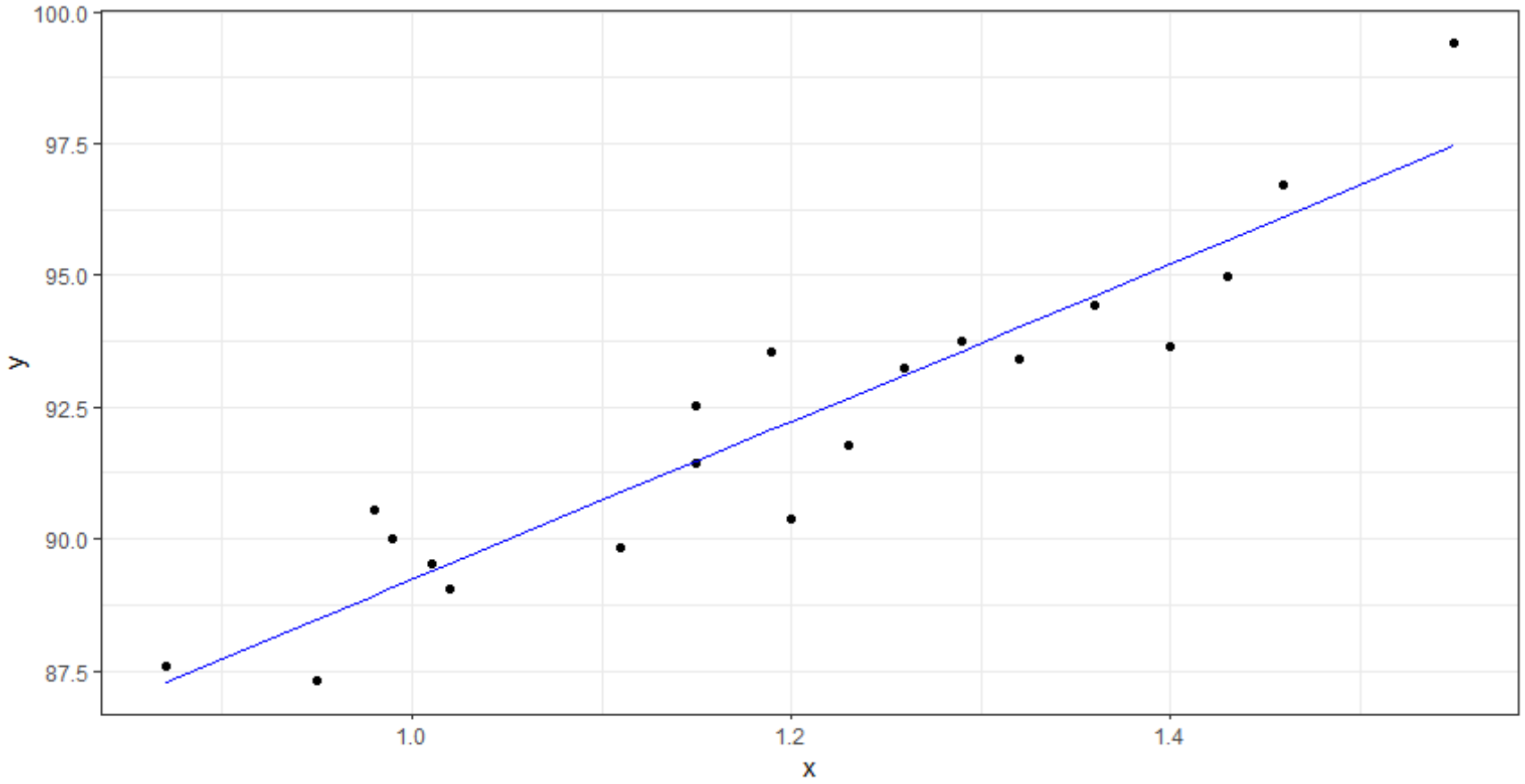
$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} = \frac{2214.66 - 20(\frac{23.92}{20})(\frac{1843.21}{20})}{29.29 - 20(\frac{23.92}{20})^2} = 14.947$$

$$b_0 = \overline{Y} - b_1\overline{X} = \frac{1843.21}{20} - 14.947 \cdot \frac{23.92}{20} = 74.283$$

Thus the **regression line** for the data is

$$\hat{Y} = \hat{f}(X) = b_0 + b_1 X = 74.283 + 14.947X.$$

Evaluating $\hat{f}$ at X_i yields the **ith fitted value** $\hat{Y}_i = \hat{f}(X_i) = b_0 + b_1 X_i$.

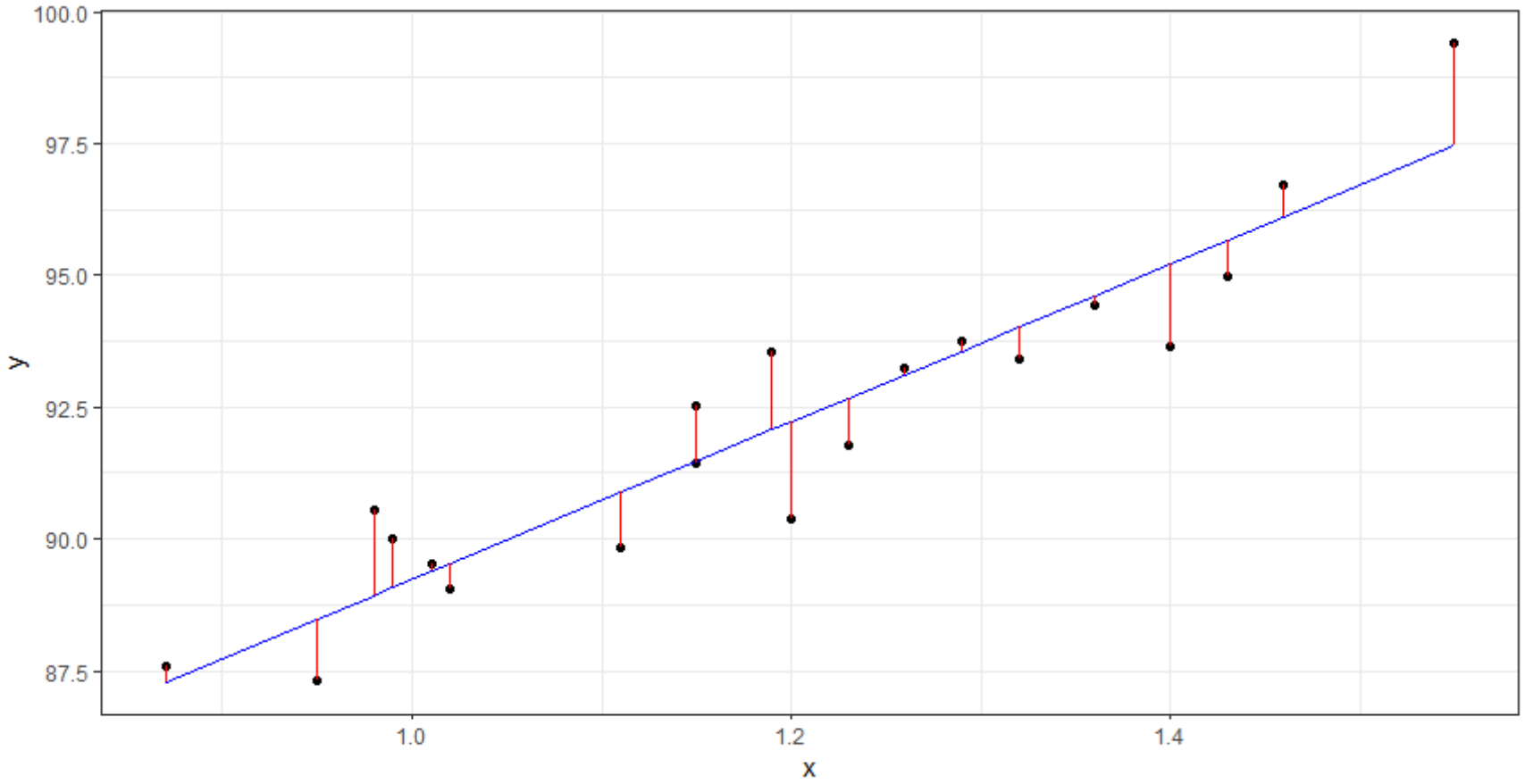


fitted line: $\hat{Y} = 74.283 + 14.947X$

2.1.2 – Residuals

The i**th residual** is $e_i = Y_i - \hat{Y}_i$. The following properties hold:

1. $\overline{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$
2. $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i = \overline{\hat{Y}}$
3. $\sum_{i=1}^{n} X_i e_i = 0$
4. $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$
5. $(\overline{X}, \overline{Y})$ is on the **regression line**
6. $\sum_{i=1}^{n} e_i^2$ is **minimal** (in the LS sense)



residuals in the fuels example

Proof:

1. We see that

$$\frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = \overline{Y} - b_0 - b_1\overline{X} = 0,$$

   according to the first normal equation.

2. From 1., we have $0 = \overline{e}$. Thus

$$0 = \overline{e} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \overline{Y} - \overline{\hat{Y}} \implies \overline{Y} = \overline{\hat{Y}}.$$

3. We see that

$$\sum_{i=1}^{n} X_i e_i = \sum_{i=1}^{n} X_i(Y_i - \hat{Y}_i) = \sum_{i=1}^{n} X_i Y_i - b_0 \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} X_i^2 = 0,$$

according to the second normal equation.

4. We see that

$$\sum_{i=1}^{n} \hat{Y}_i e_i = \sum_{i=1}^{n} (b_0 + b_1 X_i) e_i = b_0 \sum_{i=1}^{n} e_i + b_1 \sum_{i=1}^{n} X_i e_i = 0,$$

according to 1. and 3.

5. This is automatically true since

$$\hat{f}(\overline{X}) = b_0 + b_1\overline{X} = (\overline{Y} - b_1\overline{X}) + b_1\overline{X} = \overline{Y}.$$

6. For any $\mathbf{b}^* = (b_0^*, b_1^*) \neq \mathbf{b} = (b_0, b_1)$, we must have $Q(\mathbf{b}^*) \geq Q(\mathbf{b})$. Denote the residuals obtained from the line fitted with $\mathbf{b}^*$ by e_i^*. Then

$$\sum_{i=1}^{n} e_i^2 = \underbrace{\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2}_{=Q(\mathbf{b})} < \underbrace{\sum_{i=1}^{n}(Y_i - b_0^* - b_1^*X_i)^2}_{=Q(\mathbf{b}^*)} = \sum_{i=1}^{n}(e_i^*)^2.$$

This completes the proof. ■

2.1.3 – Descriptive Statistics and Correlations

The **Pearson sample correlation coefficient** r of 2 variables X and Y is defined by

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}.$$

This coefficient is such that

1. $-1 \leq r \leq 1$;

2. $|r| = 1 \iff Y_i = b_0 + b_1 X_i$, for all $i = 1, \ldots, n$, and

3. $\text{sgn}(r) = \text{sgn}(b_1)$, so that $r = 0 \iff b_1 = 0$.

If $|r| \approx 1$, then there is a **strong linear association** between X and Y.

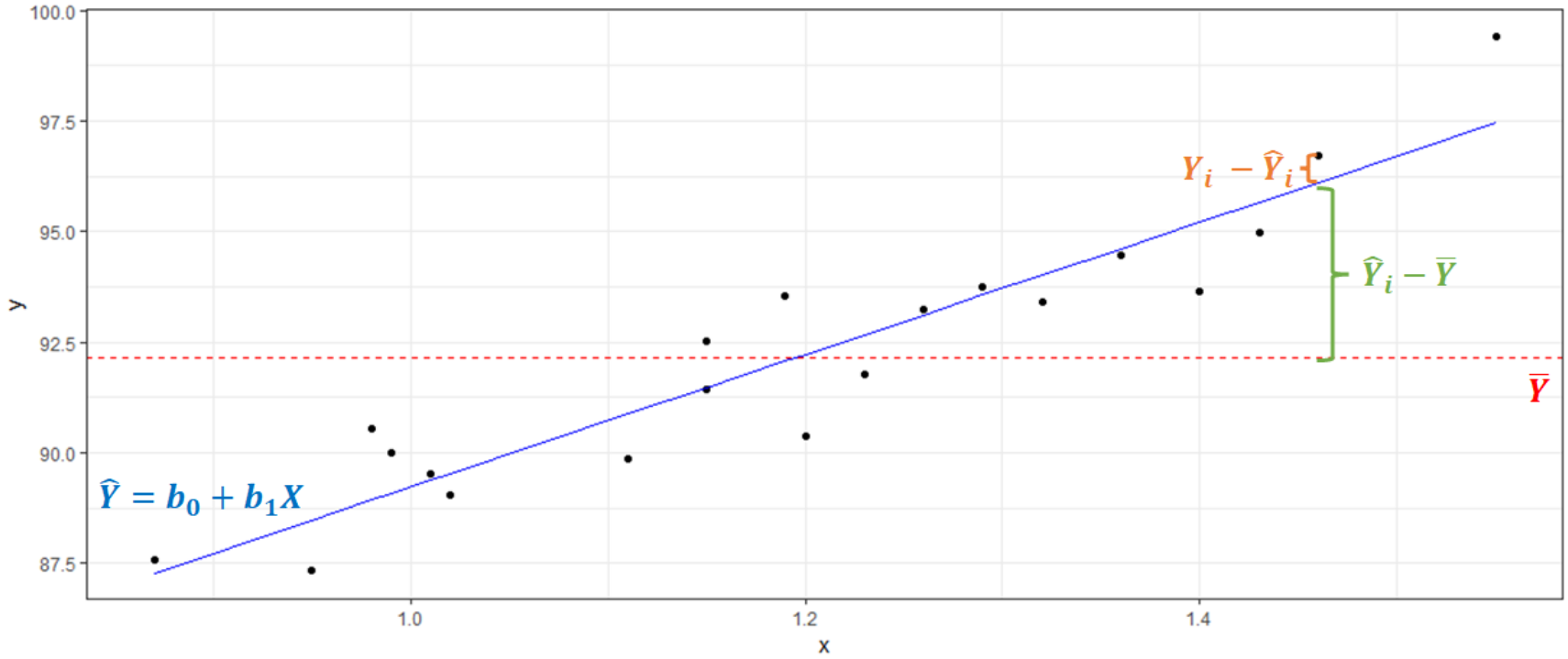
If $|r| \approx 0$, there is **very little linear association** between X and Y.

What can we say when $0 \ll |r| \ll 1$? We will discuss this at later stage (Section 2.1.5).

For now, we will only remark that we can **decompose** the total deviation as follows:

$$\underbrace{Y_i - \overline{Y}}_{\substack{\text{total} \\ \text{deviation} \\ \text{from the mean}}} = \underbrace{(Y_i - \hat{Y}_i)}_{\substack{\textbf{unexplained} \\ \textbf{deviation} \\ \textbf{from the mean}}} + \underbrace{(\hat{Y}_i - \overline{Y})}_{\substack{\textbf{deviation} \\ \textbf{from the mean} \\ \textbf{explained} \\ \textbf{by regression}}}.$$

This decomposition is shown graphically, on the next slide.



Total deviation decomposition

The **Spearman sample correlation coefficient** r_S of of 2 variables X and Y is the **Pearson correlation** between the **rank values** $R(X_i)$ and $R(Y_i)$ of X_i and Y_i, respectively. This coefficient is such that

1. $-1 \leq r_S \leq 1$;

2. $r_S = 1 \iff$ the relation between X and Y is **monotonic increasing**,

3. $r_S = -1 \iff$ the relation between X and Y is **monotonic decreasing**,

4. if the association between X and Y is **weak**, then $r_S \approx 0$, and

5. r_S is invariant under **order-preserving (monotonic) transformations**.

The computational procedure is simple: for measurements

$$\mathcal{Z} = \{Z_i \mid i = 1, \ldots, n\},$$

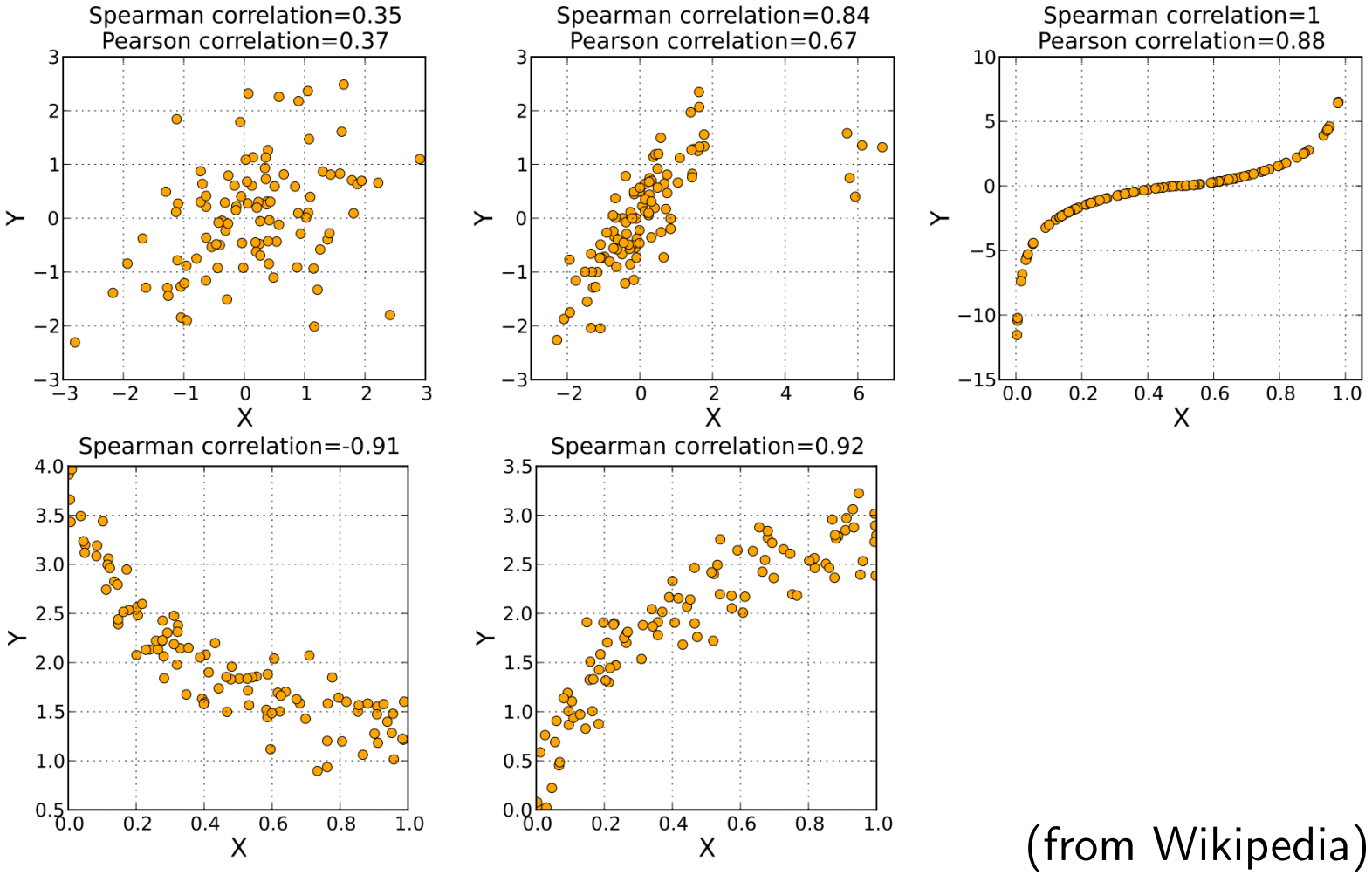
let $R(Z_i)$ be the **rank value** of Z_i in $\mathcal{Z}$; the smallest value of Z_i has rank 1, the second smallest has rank 2, and so on, until the largest value, which has rank n.

Ties are dealt with as in the example below:

Z_i	0	1.5	1.5	-1.5	3	-2
$R(Z_i)$	3	4.5	4.5	2	6	1

Formally,

$$r_S = \frac{S_{R(x)R(y)}}{\sqrt{S_{R(x)R(x)} S_{R(y)R(y)}}}.$$



(from Wikipedia)

2.1.4 – Sums of Squares Decomposition

The total deviation decomposition gives rise to one of the fundamental concepts of regression analysis: **sum of squares (SS) decompositions**.

$$\begin{aligned}
\text{SST} &= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}\left[(Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})\right]^2 \\
&= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + 2\sum_{i=1}^{n}\underbrace{(Y_i - \hat{Y}_i)}_{=e_i}(\hat{Y}_i - \overline{Y}) + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 \\
&= \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}_{\text{SSR}} + 2\underbrace{\sum_{i=1}^{n}\hat{Y}_i e_i}_{=0} - 2\overline{Y}\underbrace{\sum_{i=1}^{n} e_i}_{=0}
\end{aligned}$$

This is often written as $\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}$, where

- SST is the **total sum of squares**,
- SSE is the **error sum of squares**, and
- SSR is the **regression sum of squares**.

Note that we can write

$$\mathrm{SSR} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{n}(b_0 + b_1 X_i - \overline{Y})^2 = \sum_{i=1}^{n}(\underbrace{\overline{Y} - b_1\overline{X}}_{b_0} + b_1 X_i - \overline{Y})^2$$

$$= \sum_{i=1}^{n}\big(b_1(\overline{X} - X_i)\big)^2 = b_1^2\sum_{i=1}^{n}(\overline{X} - X_i)^2 = b_1^2 S_{xx}.$$

As $\mathrm{SST} = S_{yy}$ and $\mathrm{SSE} = Q(\mathbf{b})$, the decomposition can be written as

$$S_{yy} = b_1^2 S_{xx} + \sum_{i=1}^{n} e_i^2.$$

In the fuels dataset, we have

$$S_{xx} = 0.68, \quad S_{xy} = 10.18, \quad S_{yy} = 173.38,$$

so that the sample correlation coefficient is

$$r = \frac{10.18}{\sqrt{0.68}\sqrt{173.38}} \approx 0.94,$$

and the SS decomposition is $\mathrm{SST}\,(173.38) = \mathrm{SSR}\,(152.13) + \mathrm{SSE}\,(21.25)$. Is this a strong linear association?

2.1.5 – Coefficient of Determination

The **coefficient of determination** $R^2 = \frac{\text{SSR}}{\text{SST}}$ is the proportion of variation in the response explained by the fitted line.

When $R^2 \approx 0$, the regression is **not very significant**, whereas when $R^2 \approx 1$, the variables are strongly linearly related.

Proposition: $R^2 = r^2$.

Proof: we have seen that $\text{SSR} = b_1^2 S_{xx}$ and $\text{SST} = S_{yy}$. Thus

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \left(\frac{S_{xy}}{S_{xx}}\right)^2 \frac{S_{xx}}{S_{yy}} = b_1^2 \cdot \frac{S_{xx}}{S_{yy}} = \frac{\text{SSR}}{\text{SST}} = R^2. \qquad \blacksquare$$

This answers the question relating to the interpretation of $0 \ll |r| \ll 1$: r^2 gives a sense of how much variation the regression "explains".

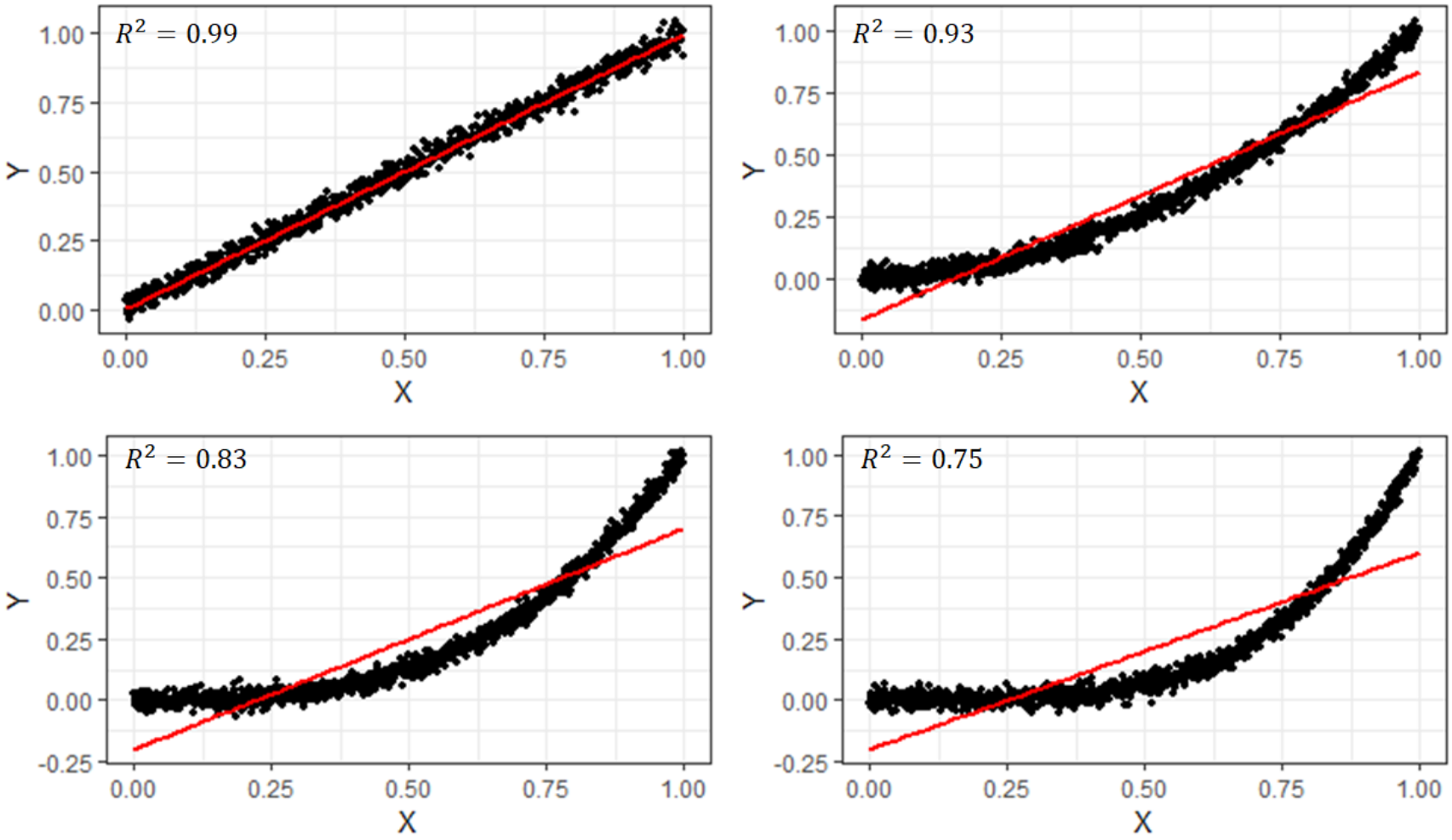
In the fuel dataset, we have

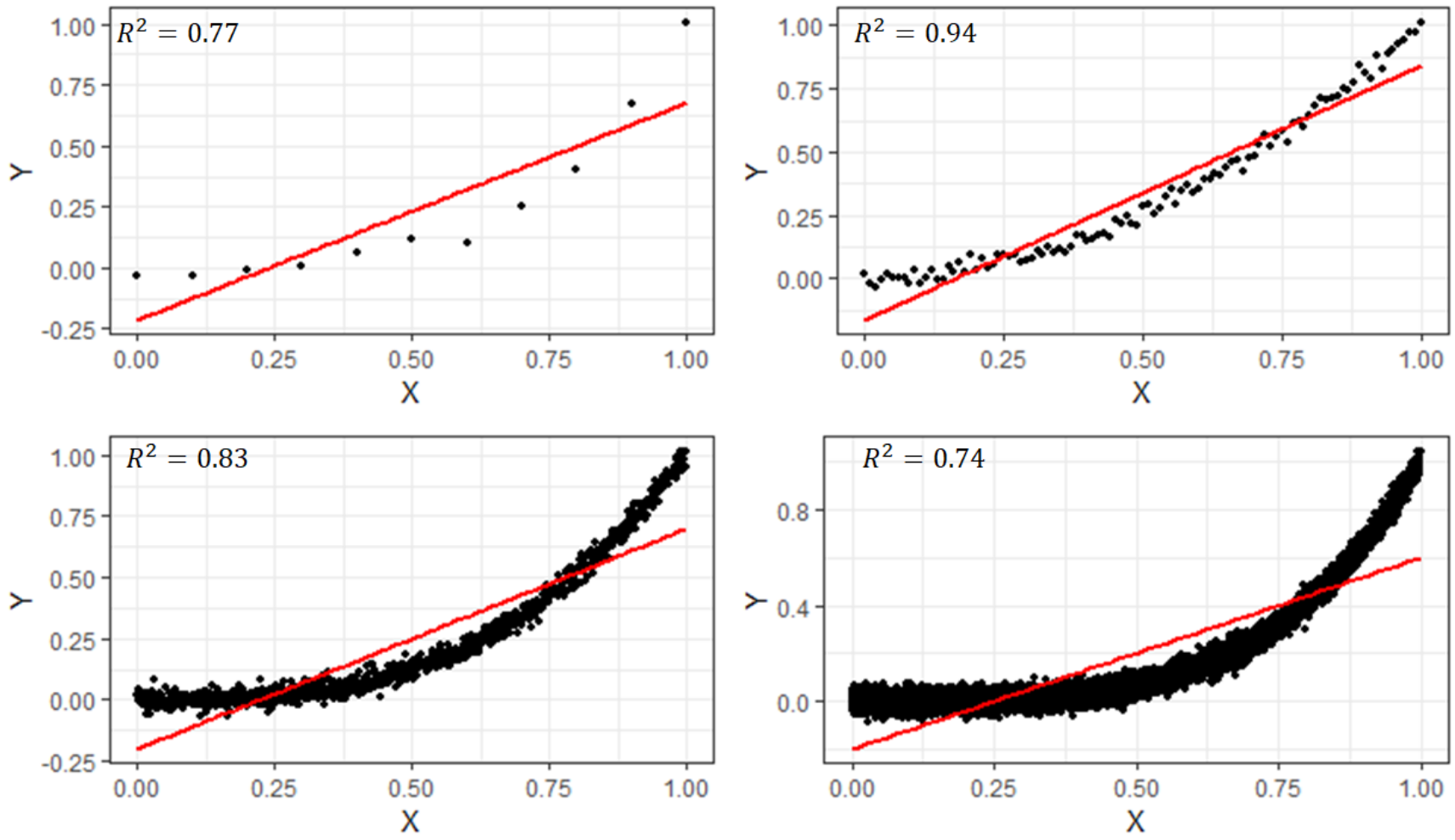
$$R^2 = \frac{152.13}{173.98} = 0.8774;$$

thus, about 87.74% of the variation observed in the data can be explained by the fitted line $\hat{Y} = 74.283 + 14.947X$.

This is a **reasonably high** proportion; together with the scatter plot, this suggests that the SRM is likely appropriate in this case.

But don't get too deeply enamoured of R^2 as a figure to validate the regression (see next pages).


$R^2 = 0.99$
$R^2 = 0.93$
$R^2 = 0.83$
$R^2 = 0.75$
Y
X
1.00
0.75
0.50
0.25
0.00
-0.25


$R^2 = 0.77$
$R^2 = 0.94$
$R^2 = 0.83$
$R^2 = 0.74$
Y
X
1.00
0.75
0.50
0.25
0.00
-0.25
0.8
0.4
0.0

2.2 – Inference

In order to test various hypotheses about the regression, we will need an estimation for the **common variance** σ^2.

In the SLR model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

we have independent normal random errors $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

The probability function of $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$ is thus

$$f(Y_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}\right].$$

The **likelihood function** is

$$L(\beta_0, \beta_1; \sigma^2) = \prod_{i=1}^{n} f(Y_i) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{Q(\beta_0, \beta_1)}{2\sigma^2}\right],$$

where

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

The likelihood L is maximized when Q is minimized with respect to β_0, β_1. We have already shown that the optimizer occurs at the **maximum likelihood estimator** $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1) = (b_0, b_1)$, for which

$$Q(b_0, b_1) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \text{SSE}.$$

Can we also use the data to find an estimator of σ^2?

Consider the **log-likelihood**

$$\begin{aligned}\ln L(b_0, b_1; \sigma^2) &= \ln \prod_{i=1}^{n} f(Y_i) = \sum_{i=1}^{n} \ln f(Y_i) \\ &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}Q(b_0, b_1)\end{aligned}$$

Because $\ln$ is a **monotone increasing** function, maximizing L is equivalent to maximizing $\ln L$. But

$$\frac{\partial L}{\partial[\sigma^2]} = -\frac{n}{2}\cdot\frac{2\pi}{2\pi\sigma^2} + \frac{1}{2(\sigma^2)^2}Q(b_0, b_1) = \frac{-1}{2\sigma^2}\left(n - \frac{Q(b_0, b_1)}{\sigma^2}\right).$$

Setting $\frac{\partial L}{\partial[\sigma^2]} = 0$ and solving for σ^2 yields

$$\widehat{\sigma^2} = \frac{1}{n}Q(b_0, b_1) = \frac{\text{SSE}}{n}.$$

This estimator is **biased**, however, as it can be shown that $\text{E}\left\{\widehat{\sigma^2}\right\} = \frac{n-2}{n}\sigma^2$.

The **mean squared error**

$$\text{MSE} = \frac{\text{SSE}}{n-2}$$

is another estimator of the population variance σ^2. It is **unbiased** as

$$\text{E}\left\{\text{MSE}\right\} = \text{E}\left\{\frac{\text{SSE}}{n-2}\right\} = \text{E}\left\{\frac{n}{n-2}\cdot\frac{\text{SSE}}{n}\right\} = \frac{n}{n-2}\text{E}\left\{\widehat{\sigma^2}\right\} = \sigma^2.$$

We can think of the variance σ^2 of a **finite population** of size n as a sum of squares divided by its degrees of freedom n:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2.$$

The estimator of the population variance using a **sample** of size n is

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2;$$

a sum of squares divided by its degrees of freedom $n-1$; 1 degree of freedom is lost because we first used the sample to compute the **sample mean** $\overline{Y}$ as an approximation of μ.

Using the same data for two different purposes creates a "link" between s^2 and $\overline{Y}$ which did not exist between σ^2 and μ.

The same reasoning explains why it should not come as a surprise that we must divide SSE by $n-2$ to obtain an unbiased estimator of σ^2: in the error of sum of squares

$$\mathrm{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2,$$

we must first use the data to estimate 2 quantities, β_0 and β_1. Thus, SSE has $n-2$ degrees of freedom, and the unbiased estimator of σ^2 is

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-2}.$$

In the fuels dataset ($n = 20$) observations, we have seen that $\mathrm{SSE} = 21.25$. The **unbiased estimator** of the error variance σ^2 in the SLR model is thus

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{n-2} = \frac{21.25}{20-2} \approx 1.18.$$

In general, if the SLR model is valid we would expect $\mathrm{E}\{Y_i\} = \beta_0 + \beta_1 X_i$ to hold, more or less, for all samples.

But the **specific values** for the LS estimators b_0, b_1 depend on the **available data**. With different observations, we would obtain different values for the estimators, and it makes sense to study the **standard error of** b_0, b_1:

$$\sigma\{b_k\} = \sqrt{\mathrm{E}\{(b_k - \beta_k)^2\}} = \sqrt{\mathrm{E}\{b_k^2\} - \beta_k^2}, \quad \text{for } k = 0, 1.$$

2.2.1 – Inference on the Regression Slope

In theory, we could then

1. collect M independent datasets,

2. repeat the LS procedure and obtain a slope estimate $b_{1;j}$ of β_1 for each dataset j, and

3. estimate $\sigma\{b_1\}$ by computing the sample standard deviation of $\{b_{1;1}, \ldots, b_{1;M}\}$.

In practice, however, collecting data is often **costly** and we may never have access to more than one set of observations.

The use of **resampling methods** (bootstrap, jackknife) is another option, but in the case of LS estimation, we can use the underlying machinery to obtain standard error estimates from a **single sample**.

As the error terms $\varepsilon_1, \ldots, \varepsilon_n$ are assumed to be independent in the SLR model, the response values $Y_1, \ldots, Y_n$ are uncorrelated, with variance $\sigma^2\{Y_i\} = \sigma^2\{\beta_0 + \beta_1 X_i + \varepsilon_i\} = \sigma^2\{\varepsilon_i\} = \sigma^2$ for $i = 1, \ldots, n$. Since

$$b_1 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{S_{xx}} Y_i, \quad \text{we have } \sigma^2\{b_1\} = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{S_{xx}}\right)^2 \sigma^2\{Y_i\},$$

so that

$$\sigma^2\{b_1\} = \sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{S_{xx}}\right)^2 \sigma^2\{\varepsilon_i\} = \frac{\sigma^2}{S_{xx}^2} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{\sigma^2}{S_{xx}^2} \cdot S_{xx} = \frac{\sigma^2}{S_{xx}}.$$

Since we do not usually know the actual value of σ^2, the **estimated standard error of** b_1 is:

$$\mathrm{s}\{b_1\} = \sqrt{\frac{\mathrm{MSE}}{S_{xx}}}.$$

In the fuels dataset example, we have

$$\mathrm{s}\{b_1\} = \sqrt{\frac{1.18}{0.68}} \approx 1.317.$$

As it is a linear combination of the **independent normal** random variables $Y_1, \ldots, Y_n$, the random variable b_1 is itself normal, according to the Central Limit Theorem.

But we already know its expectation and its variance, and so we know its distribution:

$$b_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \implies \frac{b_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim \mathcal{N}(0,1).$$

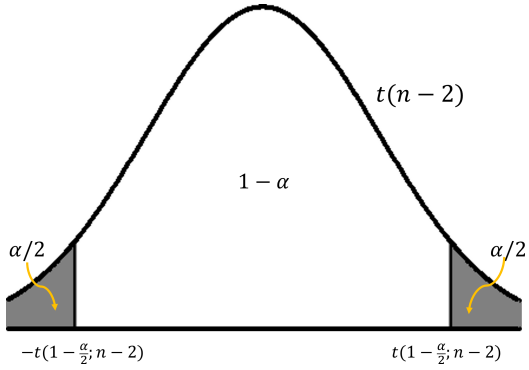
We now make assumptions that will be justified later:

$$\frac{\text{SST}}{\sigma^2} \sim \chi^2(n-1), \quad \frac{\text{SSE}}{\sigma^2} \sim \chi^2(n-2), \quad \frac{\text{SSR}}{\sigma^2} \sim \chi^2(1), \quad b_1, \text{SSE} \text{ indep.}$$

By the definition of the Student $t-$distribution (see Section 1.1.2), we have

$$T_1 = \frac{b_1 - \beta_1}{\underbrace{\sigma/\sqrt{S_{xx}}}_{=Z}} \Bigg/ \sqrt{\underbrace{\frac{\text{SSE}}{\sigma^2}}_{=U} \Bigg/ \underbrace{(n-2)}_{\nu}} = \frac{b_1 - \beta_1}{\sqrt{\text{MSE}}/\sqrt{S_{xx}}} = \frac{b_1 - \beta_1}{\text{s}\{b_1\}} \sim t(n-2).$$

Critical Region



Let $\alpha \in (0,1)$. Since $\frac{b_1-\beta_1}{\mathrm{s}\{b_1\}} \sim t(n-2)$, we have

$$1-\alpha = P\left(-t(1-\tfrac{\alpha}{2};n-2) \leq \frac{b_1-\beta_1}{\mathrm{s}\{b_1\}} \leq t(1-\tfrac{\alpha}{2};n-2)\right)$$

$$= P\left(b_1 - t(1-\tfrac{\alpha}{2};n-2)\cdot \mathrm{s}\{b_1\} \leq \beta_1 \leq b_1 + t(1-\tfrac{\alpha}{2};n-2)\cdot \mathrm{s}\{b_1\}\right).$$

Thus, the $100(1-\alpha)\%$ **confidence interval for** β_1 is

$$\text{C.I.}(\beta_1; 1-\alpha) \equiv b_1 \pm t(1-\tfrac{\alpha}{2};n-2)\cdot \mathrm{s}\{b_1\}.$$

In the fuels dataset example, we have

$$b_1 = 14.947, \quad \mathrm{s}\{b_1\} = 1.317.$$

At a **confidence level** of $1 - \alpha = 0.95$ (or an **error rate** of $\alpha = 0.05$), the critical value of the Student $t-$distribution with $n - 2 = 20 - 2 = 18$ degrees of freedom is

$$t(1 - 0.05/2; 20 - 2) = t(0.975; 18) = 2.101.$$

We can build a 95% confidence interval for β_1 as follows:

$$\text{C.I.}(\beta_1; 0.95) \equiv 14.947 \pm 2.101(1.317) = [12.17, 17.72].$$

2.2.2 – Inference on the Regression Intercept

With the same assumptions as with b_1, we also have:

$$\sigma^2\{b_0\} = \sigma^2\{\overline{Y} - b_1\overline{X}\} = \sigma^2\left\{\frac{1}{n}\sum_{i=1}^{n} Y_i - \overline{X}\sum_{i=1}^{n}\frac{(X_i-\overline{X})}{S_{xx}}Y_i\right\}$$

$$= \sigma^2\left\{\sum_{i=1}^{n}\left[\frac{1}{n} - \frac{\overline{X}(X_i-\overline{X})}{S_{xx}}\right]Y_i\right\} = \sum_{i=1}^{n}\left[\frac{1}{n} - \frac{\overline{X}(X_i-\overline{X})}{S_{xx}}\right]^2 \underbrace{\sigma^2\{Y_i\}}_{=\sigma^2}$$

$$= \sigma^2\left[\sum_{i=1}^{n}\frac{1}{n^2} - \frac{2\overline{X}}{nS_{xx}}\underbrace{\sum_{i=1}^{n}(X_i-\overline{X})}_{=0} + \frac{\overline{X}^2}{S_{xx}^2}\underbrace{\sum_{i=1}^{n}(X_i-\overline{X})^2}_{=S_{xx}}\right].$$

Thus,

$$\sigma^2\{b_0\} = \left[\frac{n}{n^2} - 0 + \frac{\overline{X}^2}{S_{xx}^2} S_{XX}\right] = \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}\right],$$

and so the estimated standard error of b_0 is:

$$\mathrm{s}\{b_0\} = \sqrt{\mathrm{MSE}}\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}}.$$

In the fuels dataset example, we have

$$\mathrm{s}\{b_0\} = \sqrt{1.18}\sqrt{\frac{1}{20} + \frac{(23.92/20)^2}{0.68}} = 1.593.$$

As was the case for b_1, b_0 follows a normal distribution since it is a linear combination of the **independent normal** random variables $Y_1, \ldots, Y_n$.

As we already know its expectation and its variance, we also know its distribution:

$$b_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{S_x x}\right]\right) \implies \frac{b_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{S_x x}}} \sim \mathcal{N}(0,1).$$

Assuming again that b_0 and SSE are independent and that $\frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2(n-2)$, the definition of the Student $t-$distribution (see Section 1.1.2) shows that

$$T_0 = \underbrace{\frac{b_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}}}}_{=Z} \Bigg/ \sqrt{\underbrace{\frac{\mathrm{SSE}}{\sigma^2}}_{=U} \Bigg/ \underbrace{(n-2)}_{\nu}} = \frac{b_0 - \beta_0}{\sqrt{\mathrm{MSE}}\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}}} = \frac{b_0 - \beta_0}{\mathrm{s}\left\{b_0\right\}}$$

follows a $t(n-2)$ distribution.

As is the case with β_1, the $100(1-\alpha)\%$ **confidence interval for** β_0 is

$$\text{C.I.}(\beta_0; 1-\alpha) \equiv b_0 \pm t(1-\tfrac{\alpha}{2}; n-2) \cdot \text{s}\{b_0\}.$$

In the fuels dataset example, we have

$$b_0 = 74.283, \quad \text{s}\{b_0\} = 1.593.$$

At a **confidence level** of $1-\alpha = 0.95$, the critical value of the Student $t-$distribution with $n-2 = 20-2 = 18$ degrees of freedom is $t(1-0.05/2; 20-2) = t(0.975; 18) = 2.101$.

We can build a 95% confidence interval for β_0 as follows:

$$\text{C.I.}(\beta_0; 0.95) \equiv 74.283 \pm 2.101(1.593) = [70.94, 77.63].$$

2.2.3 – Hypothesis Testing

With standard errors, we can **test hypotheses** on the regression parameters.

We try to determine if the true parameters β_0, β_1 take on specific values, and whether the line of best fit proides a good description of a bivariate dataset, using the following steps:

1. set up a **null** hypothesis H_0 and an **alternative** hypothesis H_1;
2. compute a **test statistic** (using the studentization);
3. find a **critical region**/$p-$value for the test statistic under H_0;
4. **reject** or **fail to reject** H_0 based on the critical region/$p-$value.

For instance, we might be interested in testing whether the true parameter value β is equal to some **candidate value** β^*, i.e.

$$H_0 : \beta = \beta^* \quad \text{against} \quad H_1 : \begin{cases} \beta < \beta^*, & \text{left-tailed test} \\ \beta > \beta^*, & \text{right-tailed test} \\ \beta \neq \beta^*, & \text{two-tailed test} \end{cases}$$

Under H_0, we have shown that

$$T_0 = \frac{b - \beta^*}{\mathrm{s}\{b\}} \sim t(n-2).$$

The **critical region** depends on the confidence level $1 - \alpha$ and on the **type** of the alternative hypothesis H_1.

Let t^* be the observed value of T_0. **We reject H_0 if t^* in the critical region**.

Alternative Hypothesis	Rejection Region
$H_1 : \beta < \beta^*$	$t^* < -t(1-\alpha; n-2)$
$H_1 : \beta > \beta^*$	$t^* > t(1-\alpha; n-2)$
$H_1 : \beta \neq \beta^*$	$\lvert t^* \rvert > t(1-\alpha/2; n-2)$

Exercices: test the following hypotheses in the fuels dataset example.

a) Test for $H_0 : \beta_0 = 75$ against $H_1 : \beta_0 < 75$ at $\alpha = 0.05$.

b) Test for $H_0 : \beta_1 = 10$ against $H_1 : \beta_1 > 10$ at $\alpha = 0.05$.

c) Test for $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at $\alpha = 0.05$.

Solutions: we have seen that

$$b_0 = 74.283, \quad \mathrm{s}\{b_0\} = 1.593, \quad b_1 = 14.947, \quad \mathrm{s}\{b_1\} = 1.317.$$

Since the error rate for all tests is $\alpha = 0.05$, we also need to compute the critical values of the Student $t-$distribution with $\nu = 20 - 2 = 18$ degrees of freedom, at confidence levels $1 - \alpha = 0.95$ and $1 - \alpha/2 = 0.975$:

$$t(0.975; 18) = 2.101, \quad \text{and} \quad t(0.95; 18) = 1.734.$$

a) We run a **left-tailed** test for the intercept: the observed test statistic is

$$t_a^* = \frac{b_0 - \beta_0^*}{\mathrm{s}\{b_0\}} = \frac{74.283 - 75}{1.593} = -0.449 \not< -1.734 = -t(0.95; 18),$$

and so we **fail to reject** H_0 at $\alpha = 0.05$.

b) We run a **right-tailed** test for the slope: the observed test statistic is

$$t_b^* = \frac{b_1 - \beta_1^*}{\mathrm{s}\{b_1\}} = \frac{14.947 - 10}{1.317} = 3.757 > 1.734 = t(0.95; 18),$$

and so we **reject H_0 in favour of H_1** at $\alpha = 0.05$.

c) We run a **two-tailed** test for the slope: the observed test statistic is

$$|t_c^*| = \left|\frac{b_1 - \beta_1^*}{\mathrm{s}\{b_1\}}\right| = \left|\frac{14.947 - 0}{1.317}\right| = 11.351 > 2.101 = t(0.975; 18),$$

and so we **reject H_0 in favour of H_1** at $\alpha = 0.05$.

We will see another test for the slope in Section 2.4.

2.2.4 – Inference on the Mean Response

We can also conduct inferential analysis for the **expected response** at $X = X^*$ (in practice, there could be replicates, say).

As before, we assume that $\mathrm{E}\{Y^*\} = \beta_0 + \beta_1 X^*$. The **estimated mean response** at $X = X^*$ is

$$\hat{Y}^* = b_0 + b_1 X^*.$$

The predictor values are **fixed**, thus $\hat{Y}^*$ is normally distributed with

$$\mathrm{E}\left\{\hat{Y}^*\right\} = \mathrm{E}\{b_0 + b_1 X^*\} = \mathrm{E}\{b_0\} + \mathrm{E}\{b_1\} X^* = \beta_0 + \beta_1 X^*,$$

so that $\hat{Y}^*$ is an **unbiased estimator** of Y^*. What is its standard error?

If b_0, b_1 were independent, we could simply compute

$$\sigma^2\left\{\hat{Y}^*\right\} = \sigma^2\left\{b_0\right\} + (X^*)^2\sigma^2\left\{b_1\right\}.$$

But they are **not independent**.

Theorem: under the SLR assumptions, $\sigma\left\{\overline{Y}, b_1\right\} = 0$ and

$$\sigma\left\{b_0, b_1\right\} = -\overline{X}\sigma^2\left\{b_1\right\}.$$

Proof: throughout, keep in mind that the Y_i are **uncorrelated**. We have

$$\sigma\left\{\overline{Y}, b_1\right\} = \sigma\left\{\frac{1}{n}\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n}\frac{(X_j - \overline{X})}{S_{xx}}Y_j\right\} = \sum_{i,j=1}^{n}\frac{1}{n}\cdot\frac{(X_i - \overline{X})}{S_{xx}}\sigma\left\{Y_i, Y_j\right\}.$$

All the terms for which $i \neq j$ have $\sigma\{Y_i, Y_j\} = 0$, the other ones have $\sigma\{Y_i, Y_i\} = \sigma^2\{Y_i\} = \sigma^2$, so

$$\sigma\{\overline{Y}, b_1\} = \frac{\sigma^2}{nS_{xx}} \underbrace{\sum_{i=1}^{n}(X_i - \overline{X})}_{=0} = 0.$$

Similarly,

$$\begin{aligned}\sigma\{b_0, b_1\} &= \sigma\left\{\sum_{i=1}^{n}\left[\frac{1}{n} - \frac{\overline{X}(X_i - \overline{X})}{S_{xx}}\right]Y_i, \sum_{i=1}^{n}\frac{(X_i - \overline{X})}{S_{xx}}Y_i\right\} \\ &= \sum_{i,j=1}^{n}\left[\frac{1}{n} - \frac{\overline{X}(X_i - \overline{X})}{S_{xx}}\right]\frac{(X_j - \overline{X})}{S_{xx}}\sigma\{Y_i, Y_j\}\end{aligned}$$

All the terms for which $i \neq j$ have $\sigma\{Y_i, Y_j\} = 0$, the other ones have $\sigma\{Y_i, Y_i\} = \sigma^2\{Y_i\} = \sigma^2$, so

$$\begin{aligned}
\sigma\{b_0, b_1\} &= \sigma^2 \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{\overline{X}(X_i - \overline{X})}{S_{xx}}\right] \frac{(X_i - \overline{X})}{S_{xx}} \\
&= \frac{\sigma^2}{nS_{xx}} \underbrace{\sum_{i=1}^{n}(X_i - \overline{X})}_{=0} - \frac{\sigma^2 \overline{X}}{S_{xx}^2} \underbrace{\sum_{i=1}^{n}(X_i - \overline{X})^2}_{S_{xx}} \\
&= -\overline{X}\frac{\sigma^2}{S_{xx}} = -\overline{X}\sigma^2\{b_1\}.
\end{aligned}$$

This completes the proof. ■

We can now determine the standard error of the estimated mean response $Y = \hat{Y}^*$ at $X = X^*$:

$$\sigma^2\left\{\hat{Y}^*\right\} = \sigma^2\left\{b_0 + b_1X^*\right\} = \sigma^2\left\{b_0\right\} + (X^*)^2\sigma^2\left\{b_1\right\} + 2\sigma\left\{b_0, X^*b_1\right\}$$

$$= \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}\right] + \frac{(X^*)^2\sigma^2}{S_{xx}} - 2X^*\overline{X}\frac{\sigma^2}{S_{xx}}$$

$$= \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}}\left[(X^*)^2 - 2\overline{X}X^* + \overline{X}^2\right] = \sigma^2\left[\frac{1}{n} + \frac{(X^* - \overline{X})^2}{S_{xx}}\right].$$

The estimated standard error is thus

$$\mathrm{s}\left\{\hat{Y}^*\right\} = \sqrt{\mathrm{MSE}}\sqrt{\frac{1}{n} + \frac{(X^* - \overline{X})^2}{S_{xx}}}.$$

But there are many ways to skin a cat:

$$\begin{aligned}\sigma^2\left\{\hat{Y}^*\right\} &= \sigma^2\left\{(\overline{Y} - b_1\overline{X}) + b_1X^*\right\} = \sigma^2\left\{\overline{Y} + b_1(X^* - \overline{X})\right\} \\ &= \sigma^2\left\{\overline{Y}\right\} + \sigma^2\left\{b_1(X^* - \overline{X})\right\} + 2(X^* - \overline{X})\sigma\left\{\overline{Y}, b_1\right\} \\ &= \frac{\sigma^2}{n} + (X^* - \overline{X})^2\frac{\sigma^2}{S_{xx}} + 0 = \sigma^2\left[\frac{1}{n} + \frac{(X^* - \overline{X})^2}{S_{xx}}\right].\end{aligned}$$

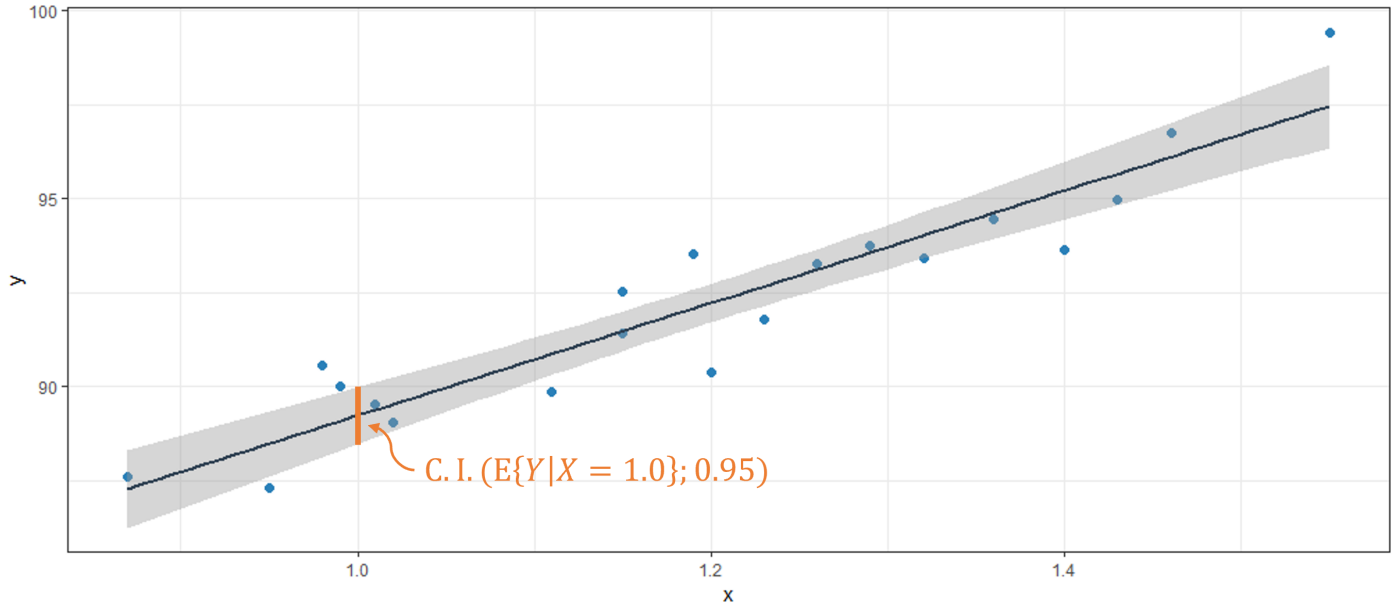
Either way, we can show that

$$T^* = \frac{\hat{Y}^* - \mathrm{E}\{\hat{Y}^*\}}{\mathrm{s}\{\hat{Y}^*\}} \sim t(n-2), \quad \text{and so}$$

$$\mathrm{C.I.}(\mathrm{E}\{Y^*\}; 1-\alpha) \equiv \beta_0 + \beta_1X^* \pm t(1-\tfrac{\alpha}{2}; n-2) \cdot \mathrm{s}\{\hat{Y}^*\}.$$

In the fuels dataset example, the 95% C.I. for $\mathrm{E}\{Y^*\}$ is

$$\text{C.I.}(\mathrm{E}\{Y^*\}; 0.95) \equiv 74.28 + 14.95X^* \pm 2.10\sqrt{1.18\left[\frac{1}{20} + \frac{(X^* - 1.12)^2}{0.68}\right]}$$



2.3 – Estimation and Prediction

When we estimate the **expected** (mean) response $\mathrm{E}\{Y^*\}$, we are determining how (b_0, b_1) could **jointly** vary from one sample to the next.

As these parameters uniquely determine the line of best fit, finding a confidence interval for the mean response at all $X = X^*$ is (more or less) equivalent to finding a **confidence band** for the entire line over the predictor domain (⚠ – **see Section 2.3.2 for joint estimation**).

It should come as no surprise that a number of observations fell outside of their respective confidence intervals for the fuels dataset example: we were estimating the **mean response** at a predictor level $X = X^*$, not the **actual** (or new) **responses** at that level.

But what if we wanted to find a range of **likely response values** at $X = X^*$?

We use the available data to build **confidence intervals** (C.I.) when we are interested in certain (fixed) population characteristics (parameters) that are unknown to us.

But a new value of the response is not a parameter; it is a **random variable**; we refer to the interval of plausible (likely) values for a new response as a **prediction interval** (P.I.) rather than as a C.I.

In order to determining a P.I. for the response, we must model the **error** involved in the prediction of the response.

Note: we assume that the new responses for a predictor level $X = X^*$ are independent of the observed responses (**the residuals are uncorrelated**).

2.3.1 – Prediction Intervals

Let Y_p^* represent a **(future) response** at $X = X^*$, so that

$$Y_p^* = \beta_0 + \beta_1 X^* + \varepsilon_p \quad \text{for some } \varepsilon_p.$$

If the average error is 0, the best prediction for Y_p^* is still the **response on the fitted line at $X = X^*$**:

$$\hat{Y}_p^* = b_0 + b_1 X^*.$$

The **prediction error** at $X = X^*$ is thus

$$\text{pred}^* = Y_p^* - \hat{Y}_p^* = \beta_0 + \beta_1 X^* + \varepsilon_p - b_0 - b_1 X^*.$$

In the SLR model, the error ε_p and the estimators b_0, b_1 are **normally distributed**. Consequently, so is the prediction error pred^*. Note that

$$\text{E}\left\{\text{pred}^*\right\} = \underbrace{\text{E}\left\{\beta_0 + \beta_1 X^* + \varepsilon_p^*\right\}}_{=\beta_0+\beta_1 X^*} - \underbrace{\text{E}\left\{b_0 + b_1 X^*\right\}}_{=\beta_0+\beta_1 X^*} = 0.$$

Because the residuals are uncorrelated (see Section 2.3), we also have

$$\begin{aligned}\sigma^2\left\{\text{pred}^*\right\} &= \sigma^2\left\{Y_p^*\right\} + \sigma^2\left\{\hat{Y}_p^*\right\} \\ &= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(X^* - \overline{X})^2}{S_{xx}}\right] = \sigma^2\left[1 + \frac{1}{n} + \frac{(X^* - \overline{X})^2}{S_{xx}}\right]\end{aligned}$$

Thus

$$\text{pred}^* \sim \mathcal{N}\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(X^* - \overline{X})^2}{S_{xx}}\right]\right).$$

The estimated standard error is thus

$$\mathrm{s}\left\{\mathrm{pred}^*\right\} = \sqrt{\mathrm{MSE}}\sqrt{1 + \frac{1}{n} + \frac{(X^* - \overline{X})^2}{S_{xx}}}.$$

As before, we can show that

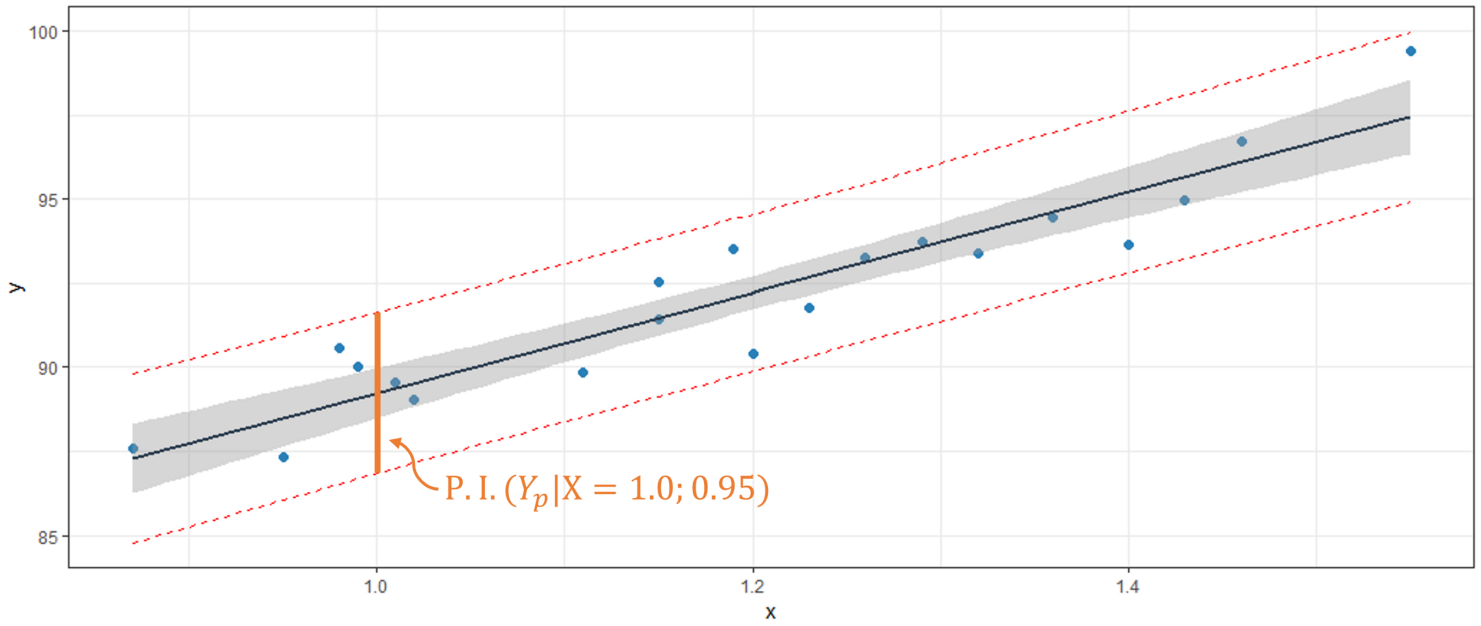
$$T_p^* = \frac{\mathrm{pred}^* - 0}{\mathrm{s}\{\mathrm{pred}^*\}} \sim t(n-2), \quad \text{and so}$$

$$\mathrm{P.I.}(Y_p^*; 1-\alpha) \equiv \beta_0 + \beta_1 X^* \pm t(1 - \tfrac{\alpha}{2}; n-2) \cdot \mathrm{s}\{\mathrm{pred}^*\}.$$

Note that $\mathrm{s}\{\hat{Y}^*\} < \mathrm{s}\{\mathrm{pred}^*\}$ so that the C.I. for the mean response is always **contained** in the P.I. for new responses. Furthermore, these regions are smallest when $X^* = \overline{X}$, and they increase as $|X^* - \overline{X}|$ increases.

In the fuels dataset example, the 95% P.I. for Y_p^* is

$$\text{P.I.}(Y_p^*; 0.95) \equiv 74.28 + 14.95X^* \pm 2.10\sqrt{1.18\left[1 + \frac{1}{20} + \frac{(X^* - 1.12)^2}{0.68}\right]}.$$



Hypothesis Testing

Since the distributions for the estimators of the mean response and for new responses are normal and since we have estimates for their standard errors, we can conduct hypothesis testing as before:

1. identify the **type** of alternative hypothesis H_1 (left-tailed, right-tailed, two-tailed),

2. compute the (studentized) **observed test statistic**, and

3. compare to the appropriate **critical value** of the Student $t-$distribution.

For instance, in the fuels dataset example, suppose we would like to test

$$H_0 : \mathrm{E}\{Y^* \mid X^* = 1.2\} = 92.5 \quad \text{against} \quad H_1 : \mathrm{E}\{Y^* \mid X^* = 1.2\} \neq 92.5.$$

Under H_0, the test statistic

$$T^* = \frac{\hat{Y}^* - 92.5}{\mathrm{s}\{\hat{Y}^*\}} \sim t(n-2) = t(18).$$

But $\hat{Y}^* = 74.28 + 14.95(1.2) = 92.22$ and

$$\mathrm{s}\{\hat{Y}^*\} = \sqrt{1.18}\sqrt{\frac{1}{20} + \frac{(1.2 - 1.12)^2}{0.68}} = 0.265.$$

The observed value of T^* is thus

$$t^* = \frac{92.22 - 92.5}{0.265} = -1.057.$$

At an error rate of $\alpha = 0.05$, the critical value of the Student $t-$distribution with $n - 2 = 18$ degrees of freedom is

$$t(1 - \tfrac{\alpha}{2}; n - 2) = t(0.975; 18) = 2.101.$$

Since $|t^*| \not> t(0.975; 18)$, there is not enough evidence to reject the null hypothesis H_0 at a confidence level of 95% (**which is not the same as accepting the null hypothesis** H_0).

What if we observed a new response $Y_p^* = 80$ for a predictor level $X^* = 1.2$? Is this a reasonable value or should we expect something larger?

At a confidence level of 95%, the prediction interval for the response at the predictor level $X^* = 1.2$ is

$$\begin{aligned}
\text{P.I.}(Y_p^*; 0.95) &\equiv \hat{Y}^* \pm t(0.975; 18) \cdot \text{s}\{\text{pred}^*\} \\
&= 74.28 + 14.95(1.2) \pm 2.101\sqrt{1.18\left[1 + \frac{1}{20} + \frac{(1.2 - 1.12)^2}{0.68}\right]} \\
&= 92.22 \pm 2.101(1.061) = [89.99, 94.45].
\end{aligned}$$

As $Y_p^* = 80$ is not in the prediction interval, this seems like an unlikely new response for $X^* = 1.2$ (at confidence level 95%).

2.3.2 – Joint Estimations and Predictions

When we use a dataset to estimate the two parameters β_0 and β_1 in the SLR model, the **error sum of squares** SSE has $n-2$ degrees of freedom.

This might seem like an obscure technical point, but there is a practical consequence: the resulting $\mathrm{C.I.}$ are necessarily **wider** than those that would be obtained if the sum of squares had more degrees of freedom.

For instance, $t(0.975; 18) = 2.101 > t(0.975, 20) = 2.086$. What does this mean for regression analysis?

One interpretation is that there is a **penalty** for the simultaneous estimation of parameters: when the same data is used to compute various estimates, it gets **"tired"** (?) and it loses some of its predictive power.

Bonferroni's Procedure

Say we are interested in the **joint** estimation of g parameters $\theta_1, \ldots, \theta_g$.

For each parameter θ_i, we build $\text{C.I.}(\theta_i) \equiv A_i = \{L_i \leq \theta_i \leq U_i\}$; the **error rate for estimating** θ_i is $P(\overline{A_i}) = P(\theta_i \notin A_i)$.

The **family confidence level** is

$$P(A_1 \cap \cdots \cap A_g) = P(\theta_1 \in A_1, \cdots, \theta_g \in A_g).$$

Theorem: for individual error rates $P(\overline{A_i}) = \frac{\alpha}{g}$, we have

$$P(A_1 \cap \cdots \cap A_g) \geq 1 - \alpha.$$

Proof: recall that $P(C \cup D) = P(C) + P(D) - P(C \cap D)$. As all probabilities are non-negative, $P(C) + P(D) \geq P(C \cup D)$.

This can be extended to unions of g events:

$$P(\overline{A_1} \cup \cdots \cup \overline{A_g}) \leq P(\overline{A_1}) + \cdots + P(\overline{A_g}); \quad \text{or}$$

$$1 - P(\overline{A_1} \cup \cdots \cup \overline{A_g}) \geq 1 - P(\overline{A_1}) - \cdots - P(\overline{A_g}) = 1 - g \cdot \frac{\alpha}{g} = 1 - \alpha.$$

As $P(A_1 \cap \cdots \cap A_g) = 1 - P(\overline{A_1} \cup \cdots \cup \overline{A_g})$, this completes the proof. ■

We can use the **Bonferroni procedure** to provide **joint** C.I. for parameters $\theta_1, \ldots, \theta_g$ at a family confidence level of $1 - \alpha$:

$$\text{C.I.}_{\text{B}}(\theta_i; 1-\alpha) \equiv \hat{\theta}_i \pm t(1 - \tfrac{\alpha/g}{2}; \text{d.f.}) \cdot \text{s}\{\hat{\theta}_i\}, \quad i = 1, \ldots, g.$$

Joint Estimation of β_0 and β_1

At a family confidence level of $1-\alpha$, the joint **Bonferroni** C.I. for β_0 and β_1 $(g=2)$ take the form:

$$\text{C.I.}_{\text{B}}(\beta_i; 1-\alpha) \equiv b_i \pm t(1-\tfrac{\alpha}{4}; n-2) \cdot \text{s}\{b_i\}, \quad i = 0, \ldots, 1.$$

At least $100(1-\alpha)\%$ of the times we use this procedure, both β_0 and β_1 will fall inside their respective C.I.

Example: in the fuels dataset, if we want a family confidence level of $1-\alpha = 0.95$, we need to use $t(1-\frac{0.05}{4}; 20-2) = t(0.9875; 18) = 2.44501$:

$$\text{C.I.}_{\text{B}}(\boldsymbol{\beta}; 0.95) \equiv \begin{cases} 74.283 \pm 2.445 \cdot 1.593 \equiv [70.39, 78.18] & (\beta_0) \\ 14.947 \pm 2.445 \cdot 1.317 \equiv [11.73, 18.17] & (\beta_1) \end{cases}$$

Working-Hotelling's Procedure

When we estimate a C.I. for the mean response at $X = X^*$, we express the lower bound and the upper bound of the interval as a function of X^*.

It would be tempting to see the union of all these C.I. as a **confidence band** for the mean response at all X, i.e., for the **true line of best fit**

$$\mathrm{E}\{Y\} = \beta_0 + \beta_1 X.$$

If we are only interested in jointly estimating the mean response at a "small" number of levels $X = X_i^*$, $i = 1, \ldots, g$, with a family confidence level $1-\alpha$, we can use the **Bonferroni** procedure:

$$\mathrm{C.I._B}(\mathrm{E}\{Y_i^*\}; 1-\alpha) = \hat{Y}_i^* \pm t(1 - \tfrac{\alpha/g}{2}; n-2) \cdot \mathrm{s}\{\hat{Y}_i^*\}, \quad i = 1, \ldots, g.$$

If we want to build a $100(1-\alpha)\%$ confidence region for $\mathrm{E}\{Y\} = \beta_0 + \beta_1 X$, the Bonferroni approach would require us to let $g \to \infty$ in the C.I. computations, which is problematic as $t(1 - \frac{\alpha/g}{2}; n-2) \to \infty$ in that case.

Instead, we seek $W > 0$ such that

$$1 - \alpha = P\Big(\hat{Y}(X) - W \cdot \mathrm{s}\{\hat{Y}(X)\} \leq \underbrace{\beta_0 + \beta_1 X}_{=\mathrm{E}\{\hat{Y}(X)\}} \leq \hat{Y}(X) + W \cdot \mathrm{s}\{\hat{Y}(X)\}\Big)$$

for all X in the regression domain. This can be achieved if

$$1 - \alpha = P\left(\max_X \left\{\left|\frac{\hat{Y}(X) - \mathrm{E}\{\hat{Y}(X)\}}{\mathrm{s}\{\hat{Y}(X)\}}\right|\right\} \leq W\right), \quad \text{or if}$$

$$1-\alpha = P\left(\max_X \left\{\frac{(\hat{Y}(X) - \mathrm{E}\{\hat{Y}(X)\})^2}{\mathrm{s}^2\{\hat{Y}(X)\}}\right\} \leq W^2\right).$$

In order to find the appropriate W, we need to know the distribution of

$$\mathcal{M} = \max_X \left\{\frac{(\hat{Y}(X) - \mathrm{E}\{\hat{Y}(X)\})^2}{\mathrm{s}^2\{\hat{Y}(X)\}}\right\} = \max_X \left\{\frac{\left[(b_0 + b_1 X) - (\beta_0 + \beta_1 X)\right]^2}{\mathrm{MSE}\left[\frac{1}{n} + \frac{(X-\overline{X})^2}{S_{xx}}\right]}\right\}.$$

Set $t = X - \overline{X}$; then the quantity can be re-written as:

$$\max_t \left\{\frac{\left[\overline{Y} - \mathrm{E}\left\{\overline{Y}\right\} + (b_1 - \beta_1)t\right]^2}{\mathrm{MSE}\left[\frac{1}{n} + \frac{t^2}{S_{xx}}\right]}\right\} = \max_t \left\{\frac{[c_1 + d_1 t]^2}{c_2 + d_2 t^2}\right\} = \max_t\{h(t)\}.$$

Note that $c_2, d_2 > 0$ as $\text{MSE}, S_{xx} > 0$, so $h(t) \geq 0$ for all t. This is a continuous rational function of a single variable, with a horizontal asymptote at $h = d_1^2/d_2 \geq 0$; its first derivative is

$$h'(t) = \frac{2\,(c_1 + d_1 t)\,(c_2 d_1 - c_1 d_2 t)}{(c_1 + d_2 t^2)^2}.$$

The critical points are found at $t_1 = -\frac{c_1}{d_1}$ and $t_2 = \frac{c_2 d_1}{c_1 d_2}$. Since

$$h(t_1) = 0 \quad \text{and} \quad h(t_2) = \frac{c_1^2 d_2 + c_2 d_1^2}{c_2 d_2} = \frac{c_1^2}{c_2} + \frac{d_1^2}{d_2} \geq 0,$$

then we must have

$$\max_t \{h(t)\} = \frac{c_1^2}{c_2} + \frac{d_1^2}{d_2}.$$

Thus

$$\mathcal{M} = \frac{(\overline{Y} - \mathrm{E}\{\overline{Y}\})^2}{\mathrm{MSE}\,/n} + \frac{(b_1 - \beta_1)^2}{\mathrm{MSE}\,/S_{xx}} = \frac{\left(\dfrac{\overline{Y} - \mathrm{E}\{\overline{Y}\}}{\sigma/\sqrt{n}}\right)^2 + \left(\dfrac{b_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}\right)^2}{\mathrm{MSE}\,/\sigma^2}$$

Both of the random variables in the numerator of $\mathcal{M}$ are independent and

$$\frac{\overline{Y} - \mathrm{E}\{\overline{Y}\}}{\sigma/\sqrt{n}}, \frac{b_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim \mathcal{N}(0,1) \implies \left(\frac{\overline{Y} - \mathrm{E}\{\overline{Y}\}}{\sigma/\sqrt{n}}\right)^2, \left(\frac{b_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}\right)^2 \sim \chi^2(1).$$

We can re-write the random variable in the denominator of $\mathcal{M}$ as

$$\mathrm{MSE}\,/\sigma^2 = \frac{\mathrm{SSE}}{\sigma^2}\Big/ n-2,$$

so that

$$\mathcal{M} = \frac{\overbrace{2\left[\left(\frac{\overline{Y}-\mathrm{E}\{\overline{Y}\}}{\sigma/\sqrt{n}}\right)^2+\left(\frac{b_1-\beta_1}{\sigma/\sqrt{S_{xx}}}\right)^2\right]}^{\sim\chi^2(2)}\Big/2}{\underbrace{\frac{\mathrm{SSE}}{\sigma^2}}_{\sim\chi^2(n-2)}\Big/n-2} \sim 2F(2,n-2).$$

We thus have

$$1-\alpha = P(\mathcal{M} \leq W^2) \Longleftrightarrow W^2 = 2F(1-\alpha;2,n-2).$$

Joint Estimation of Mean Responses

At a family confidence level of $1-\alpha$, the joint **Working-Hotelling** C.I. for $\mathrm{E}\{Y_i^*\}$ at any number of levels $X = X_i^*$ take the form:

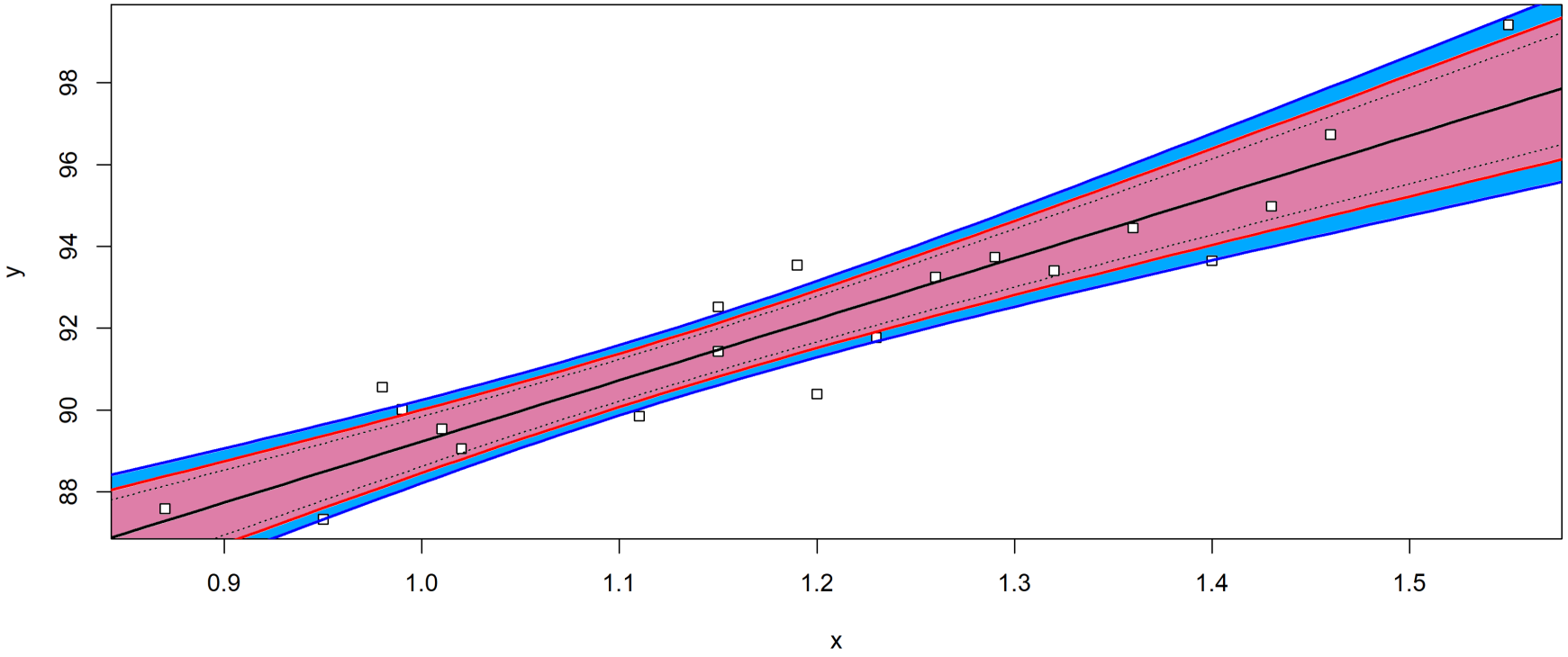
$$\mathrm{C.I.}_{\mathsf{WH}}(\mathrm{E}\{Y_i^*\}; 1-\alpha) = \hat{Y}_i^* \pm \sqrt{2F(1-\alpha; 2, n-2)} \cdot \mathrm{s}\{\hat{Y}_i^*\}.$$

We select whichever of the Bonferroni or Working-Hotelling approaches yields the **tighter** C.I..

In the fuels dataset example, at a family confidence level of 0.95, the required factor is

$$W = \sqrt{2F(0.95; 2; 18)} = 2.667.$$

The Working-Hotelling confidence band for the line of best fit in the fuels dataset is shown in **pink** below; the Bonferroni region for any 20 simultaneous inferences on the mean response also contains the **blue** region.



Scheffé's Procedure and Joint Estimation of New Responses

If we want to obtain **joint prediction intervals** at family confidence level $1-\alpha$ for g new responses $Y^*_{p_i}$ at predictor levels $X = X^*_i$, $i = 1, \ldots, g$, we use the approach (among the two below) that leads to "tighter" P.I.:

- if g is "small", the **Bonferroni** P.I. are

$$\text{P.I.}_{\text{B}}(Y^*_{p_i}; 1-\alpha) \equiv \hat{Y}^*_{p_i} \pm t(1-\tfrac{\alpha/g}{2}; n-2) \cdot \text{s}\{\text{pred}^*_i\}, \quad i = 1, \ldots, g;$$

- if g is "large", the **Scheffé** P.I. are

$$\text{P.I.}_{\text{S}}(Y^*_{p_i}; 1-\alpha) \equiv \hat{Y}^*_{p_i} \pm \sqrt{gF(1-\alpha; g, n-2)} \cdot \text{s}\{\text{pred}^*_i\}, \quad i = 1, \ldots, g.$$

2.4 – Significance of Regression

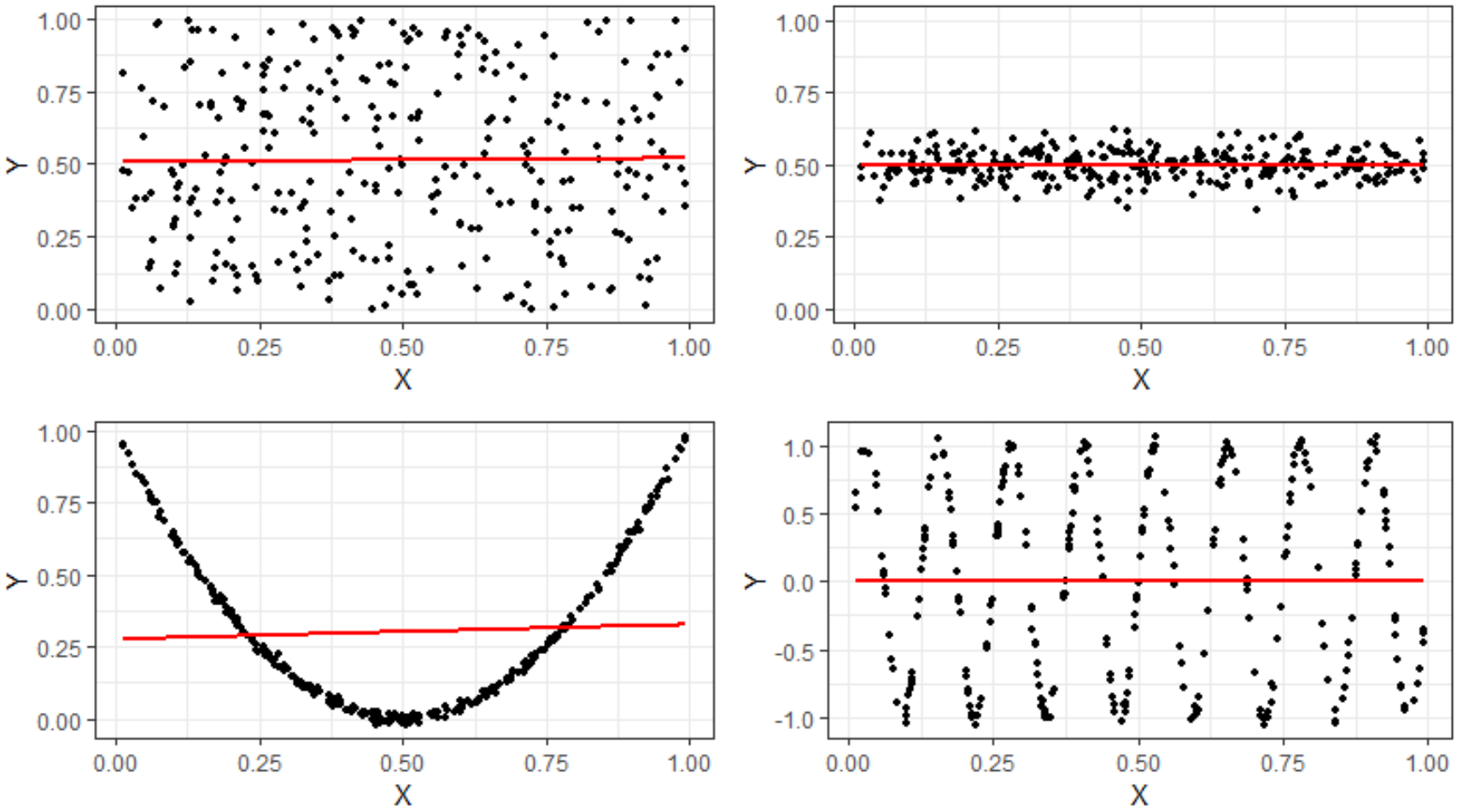
What can we conclude if $\beta_1 = 0$? It could be that:

1. there is **no relationship** between X and Y, as in a diffuse cloud of points – knowledge of X explains nothing about the possible values of Y;

2. there is a **horizontal relationship** between X and Y, so that changes in X do not bring any change in Y;

3. there is a **non-linear relationship** between X and Y which is best approximated by a horizontal line.

In each of these cases, we say that regression is **not significant**.


1.00
0.75
0.50
0.25
0.00
Y
0.00
0.25
0.50
0.75
1.00
X
1.00
0.75
0.50
0.25
0.00
Y
0.00
0.25
0.50
0.75
1.00
X
1.00
0.75
0.50
0.25
0.00
Y
0.00
0.25
0.50
0.75
1.00
X
1.0
0.5
0.0
-0.5
-1.0
Y
0.00
0.25
0.50
0.75
1.00
X

This test for **significance of regression** is

$$H_0 : \beta_1 = 0 \quad \textbf{against} \quad H_1 : \beta_1 \neq 0.$$

The underlying assumptions are that:

1. the **simple linear regression model** holds, and

2. the error terms are **independent** and **normal**, with variance σ^2.

Under these assumptions, we can show that b_0, b_1 are **independent of** SSE and that

$$\frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2(n-2).$$

Analysis of Variance

Whether H_0 holds or not, the unbiased estimator for the error variance is

$$\widehat{\sigma^2} = \text{MSE} = \frac{\text{SSE}}{n-2} \quad \left(\implies \frac{\text{SSE}}{\sigma^2} \sim \chi^2(n-2) \right).$$

Recall that, in general:

$$\text{SST} = \text{SSR} + \text{SSE}\,.$$

If $H_0 : \beta_1 = 0$ holds, then $Y_1, \ldots, Y_n$ is an independent random sample drawn from $\mathcal{N}(\beta_0, \sigma^2)$. Our best estimate for σ^2 is thus

$$\widehat{\sigma^2} = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \frac{\text{SST}}{n-1} \quad \left(\implies \frac{\text{SST}}{\sigma^2} \sim \chi^2(n-1) \right).$$

Cochran's Theorem implies that $\mathrm{SSE}, \mathrm{SSR}$ are **independent**, and that

$$\frac{\mathrm{SSR}}{\sigma^2} \sim \chi^2\big((n-1)-(n-2)\big) = \chi^2(1).$$

Thus, if $H_0 : \beta_1 = 0$ holds, the quotient

$$F^* = \frac{\underbrace{\left(\frac{\mathrm{SSR}}{\sigma^2}\right)}_{\chi^2(\nu_1)} \Big/ \underbrace{1}_{\nu_1}}{\underbrace{\left(\frac{\mathrm{SSE}}{\sigma^2}\right)}_{\chi^2(\nu_2)} \Big/ \underbrace{(n-2)}_{\nu_2}} = \frac{\mathrm{SSR}\,/1}{\mathrm{SSE}\,/(n-2)} = \frac{\mathrm{MSR}}{\mathrm{MSE}} \sim F(1, n-2)$$

follows a Fisher F distribution with $1, n-2$ degrees of freedom.

It can be shown that

$$\mathrm{E}\{\mathrm{MSR}\} = \sigma^2 + \beta_1^2 S_{xx}.$$

If $\beta_1 \neq 0$, we thus have $\mathrm{E}\{\mathrm{MSR}\} > \sigma^2$, which means that large observed values of F^* support $H_1 : \beta_1 \neq 0$.

Decision Rule: let $0 < \alpha \ll 1$. If $F^* > F(1-\alpha; 1, n-2)$, then **we reject H_0 in favour of H_1 at level α**.

We can find $F(1-\alpha; 1, n-2)$, the critical value of $F(1, n-2)$ at confidence level $1-\alpha$, by consulting tables of F values, or using R.

Note that we have already examined a test for significance of regression in Section 2.2.3. They are linked: when $\beta_1 = 0$, $F^* = (t^*)^2$.

In the fuels dataset example, we have $n = 20$ and

$$\text{SST} = 173.38, \quad \text{SSR} = 152.13, \quad \text{SSE} = 21.25,$$

so that

$$F^* = \frac{\text{SSR}/1}{\text{SSE}/(n-2)} = \frac{152.13/1}{21.25/18} = 128.8631 = (11.351)^2;$$

at $\alpha = 0.05$, the critical value is $F(1 - 0.05; 1, 18) = 4.413873$.

Since $F^* > F(0.95; 1, 18)$, we reject $H_0 : \beta_1 = 0$ at $\alpha = 0.05$, in favour of the alternative being that the regression is **significant** $(H_1 : \beta_1 \neq 0)$.

Golden Rule

In general, if SSx is a sum of squares with $n - x$ degrees of freedom, the corresponding **mean sum of squares** is

$$\mathrm{MSX} = \frac{\mathrm{SSX}}{n - x}.$$

Under some specific test assumptions (or under general assumptions, depending on the sum of squares in question or the situation), MSX provides an unbiased estimator for the variance σ^2 of the error terms.

Depending on the situation, Cochran's Theorem can then be used to show that

$$\frac{\mathrm{SSX}}{\sigma^2} \sim \chi^2(n - x).$$