# MAT 3775
# Analyse de la régression

# Chapitre 2
# Régression linéaire simple

P. Boily (uOttawa)

Session d'hiver – 2023

# Aperçu

## 2.1 – Estimation par les moindres carrés (p.10)

- Équations normales (p.15)
- Résidus (p.25)
- Statistiques descriptives et corrélations (p.30)
- Décomposition en sommes de carrés (p.36)
- Coefficient de détermination (p.39)

## 2.2 – Inférence (p.43)

- Inférence sur la pente (p.50)
- Inférence sur l'ordonnée à l'origine (p.56)
- Tests d'hypothèses (p.60)
- Inférence sur la réponse moyenne (p.65)

# Aperçu (suite)

## 2.3 – Estimation et prédiction (p.72)

- Intervalle de prédiction (p.74)
- Estimations et prédictions simultanées (p.82)

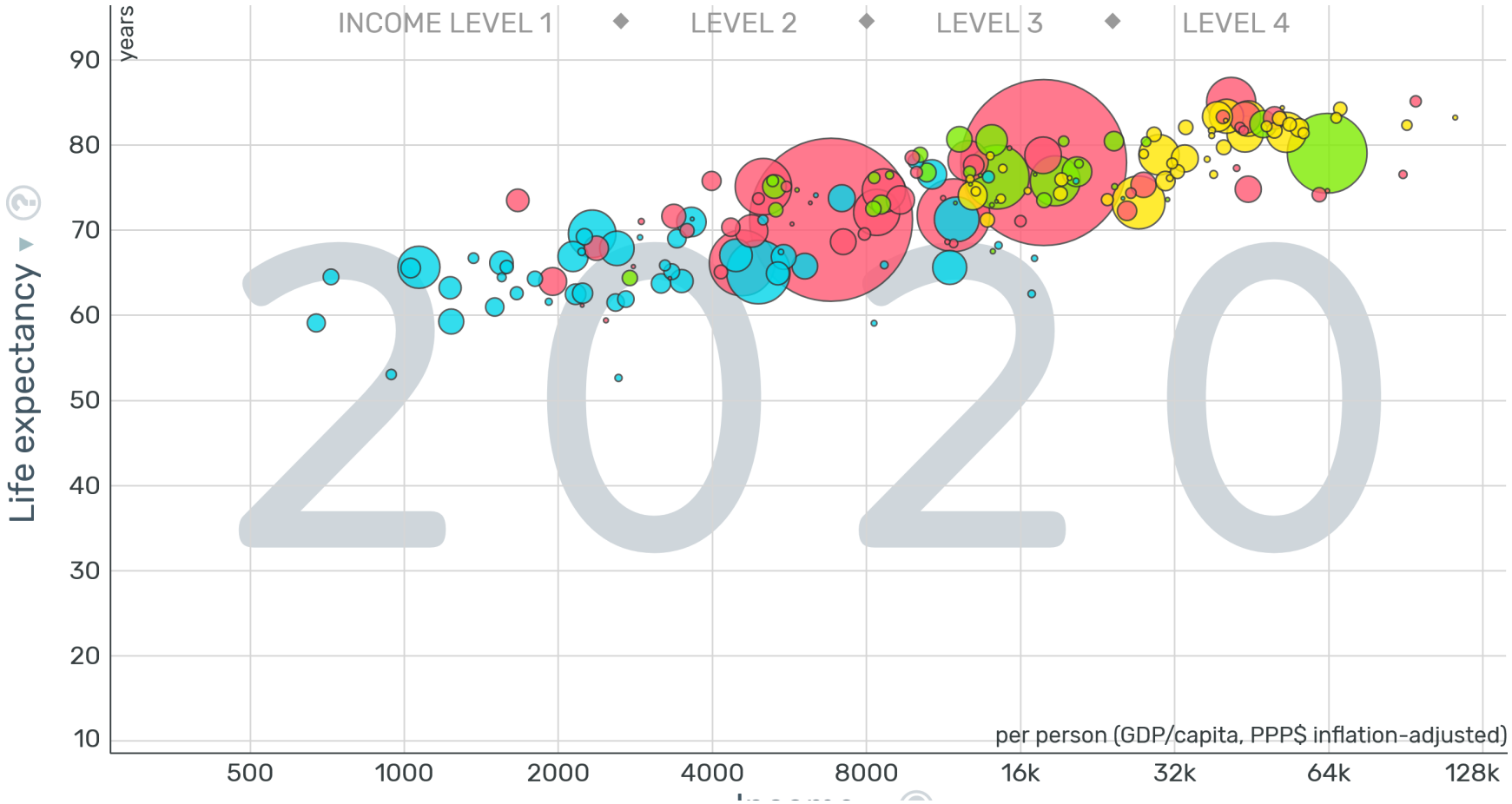
## 2.4 – Signification de la régression (p.95)

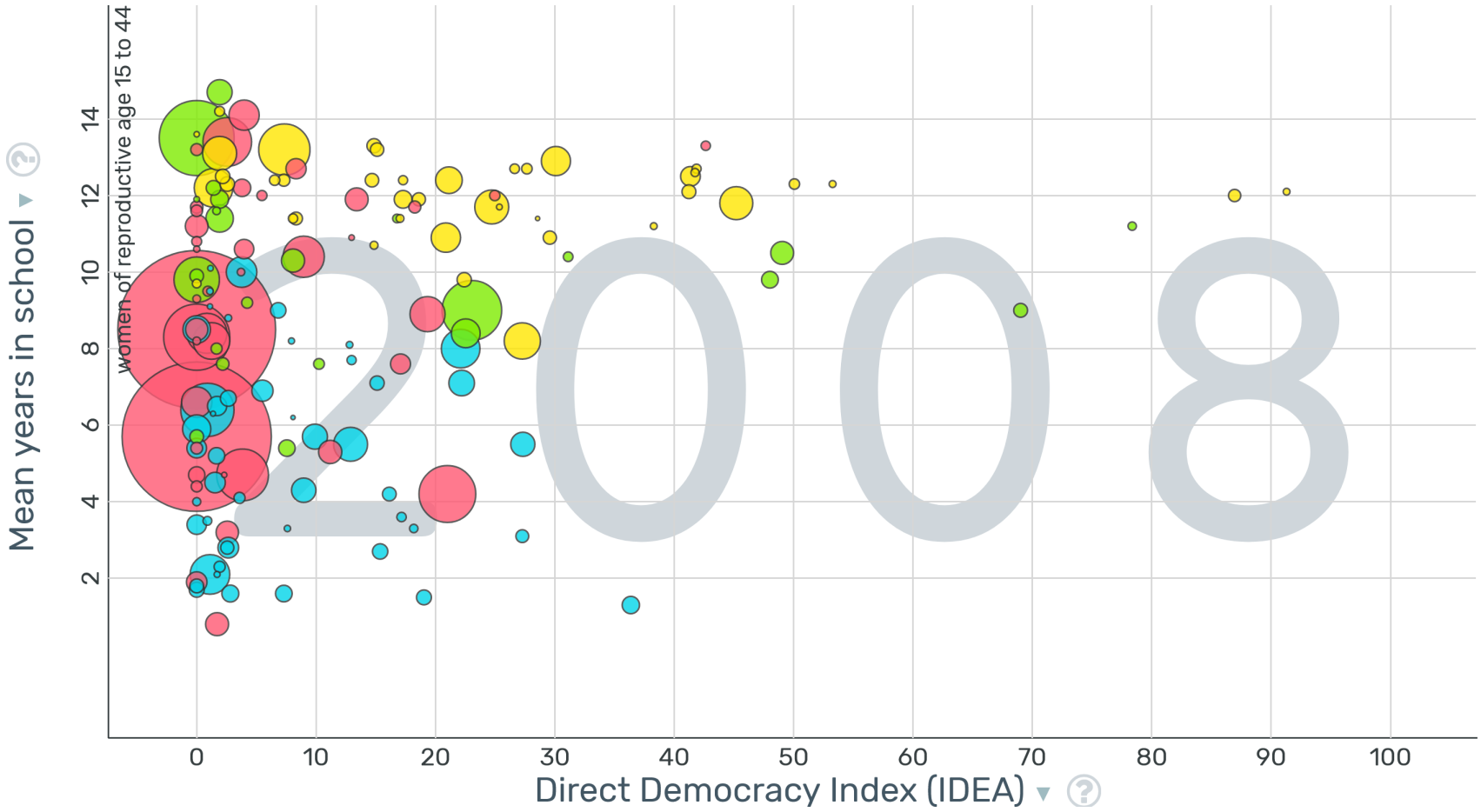
# 2 – Régression linéaire simple

Commençons en considérant un scénario simple, avec seulement deux variables **continues** : une **réponse** $Y$ et un **prédicteur** $X$.

**Exemples:**

- $X$ : âge ; $Y$ : taille
- $X$ : âge ; $Y$ : salaire
- $X$ : revenu ; $Y$ : espérance de vie
- $X$ : nombre d'heures d'ensoleillement ; $Y$ : biomasse végétale


INCOME LEVEL 1
LEVEL 2
LEVEL 3
LEVEL 4
Life expectancy
years
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
2
4
6
8
10
12
14
2008
0
10
20
30
40
50
60
70
80
90
100
Direct Democracy Index (IDEA)

En théorie, nous espérons qu'il puisse exister une **relation fonctionnelle** $Y = f(X)$ entre $X$ et $Y$.

En pratique (en supposant même qu'une relation existe), le mieux que nous puissions espérer est une **relation statistique**

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

où

- $f(X)$ est la **fonction de réponse** ;
- $\varepsilon$ est l'**erreur aléatoire** (ou le bruit).

Dans le cas de la **régression linéaire simple**, nous supposons que la fonction de réponse est $f(X) = \beta_0 + \beta_1 X$.

Les éléments constitutifs de l'analyse de régression sont les **observations** :

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

Dans un cadre idéal, ces observations sont **(conjointement) échantillonnées de façon aléatoire**, selon un plan approprié (qui est un sujet pour d'autres cours).

Le **modèle de régression linéaire simple** (RLS) est

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

où $\beta_0, \beta_1$ sont des **paramètres inconnus (à découvrir)** et $\varepsilon_i$ est l'**erreur aléatoire de l'observation (ou cas)** $i$.

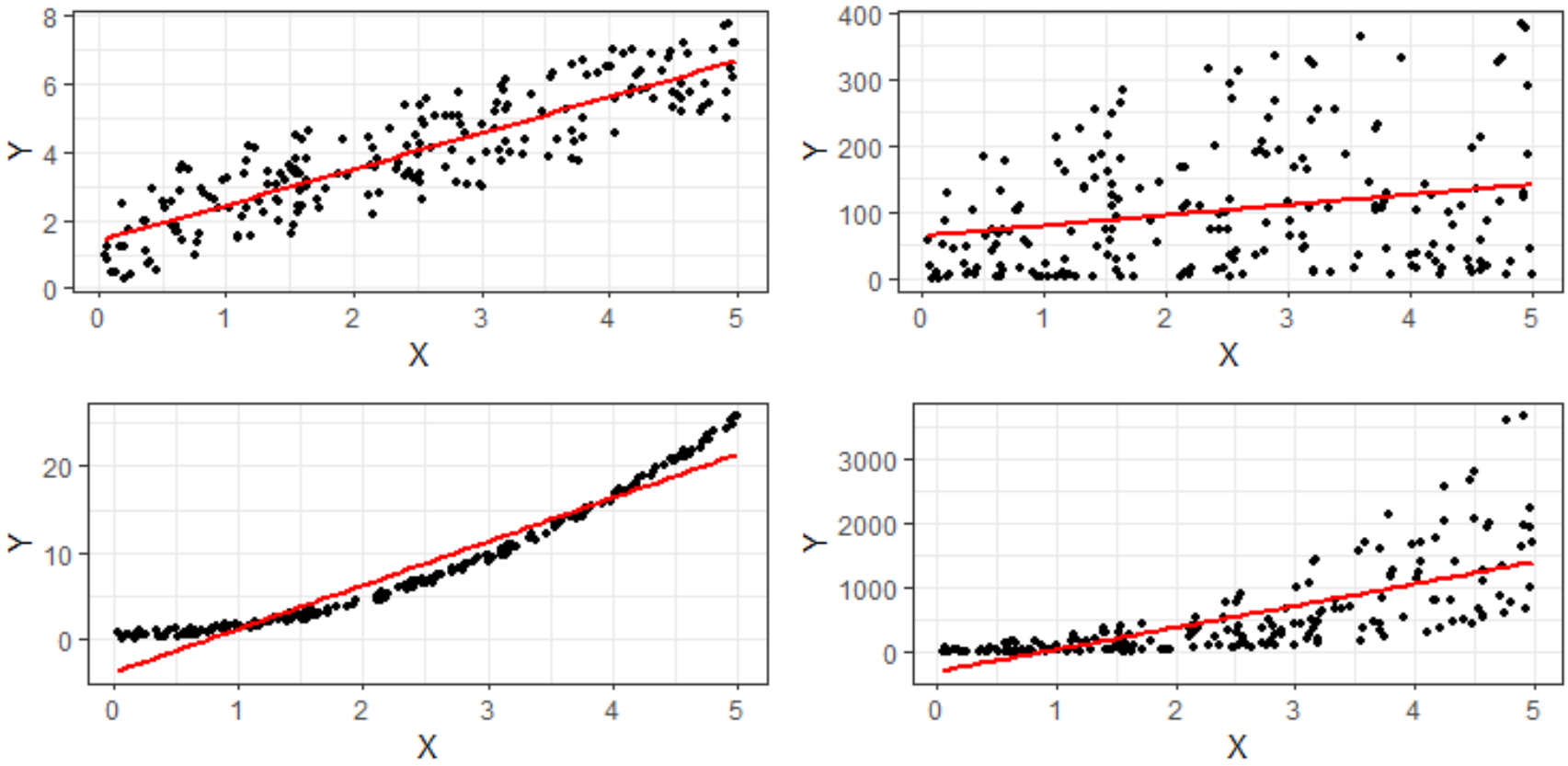
Dans la RLS, l'**hypothèse sur la structure d'erreur** est $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

La notation matricielle permet de présenter l'hypothèse de manière compacte. Décortiquons cet énoncé. Posons $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$.

Puisque $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ : nous avons

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

Les erreurs $\{\varepsilon_i\}$ sont donc **non corrélées**, de **moyenne** $0$ et de **variance constante**. En d'autres termes, la **dispersion des observations est uniforme autour de la droite d'ajustement (ou droite de régression)**.


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 – Estimation par les moindres carrés

Nous traitons les valeurs des prédicteurs $X_i$ comme constantes (c'est-à-dire que nous supposons qu'il n'y a **aucune erreur de mesure**).

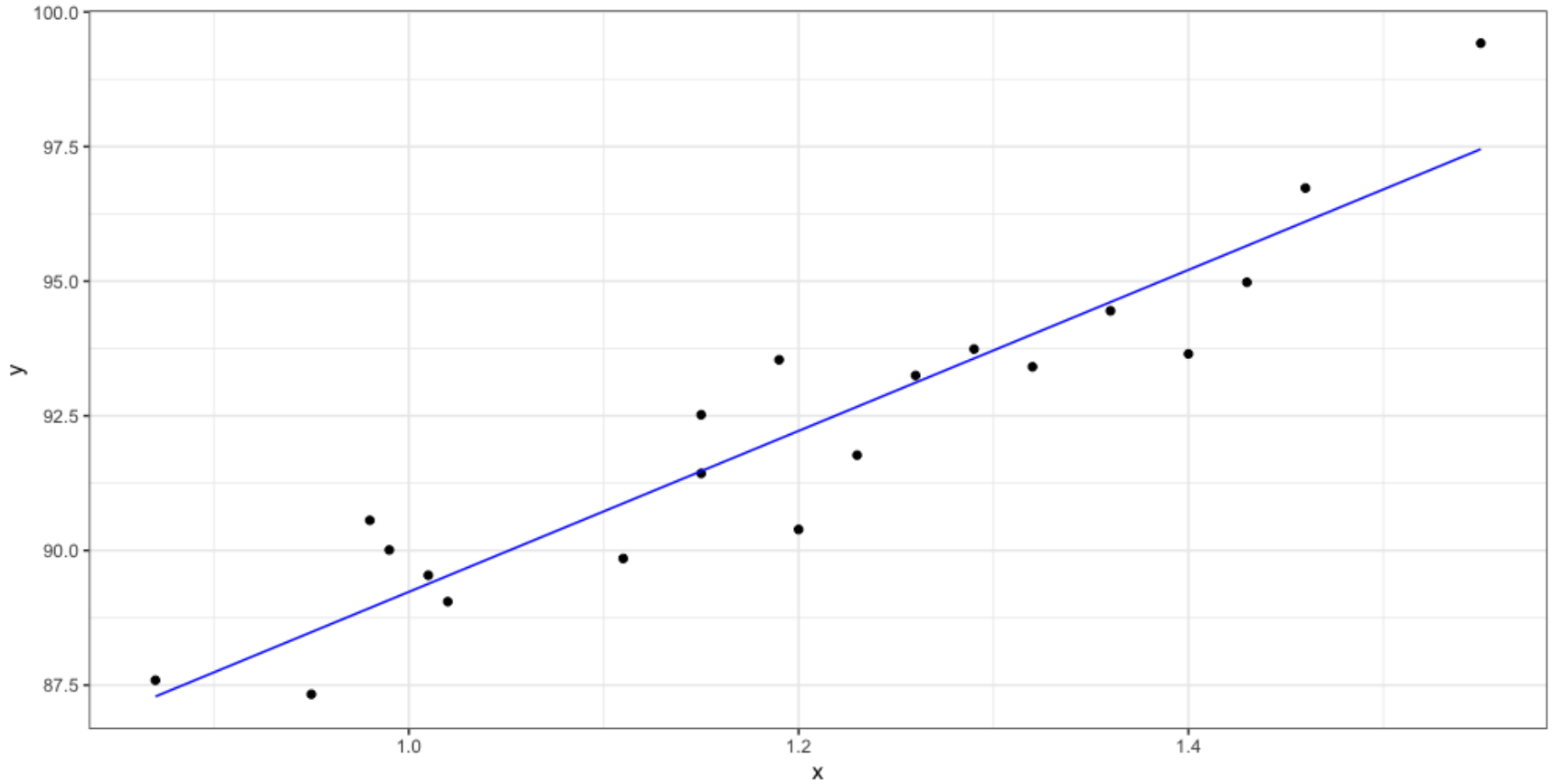
Puisque $\mathrm{E}\{\varepsilon_i\} = 0$, la **réponse moyenne étant donné** $X_i$ est ainsi

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

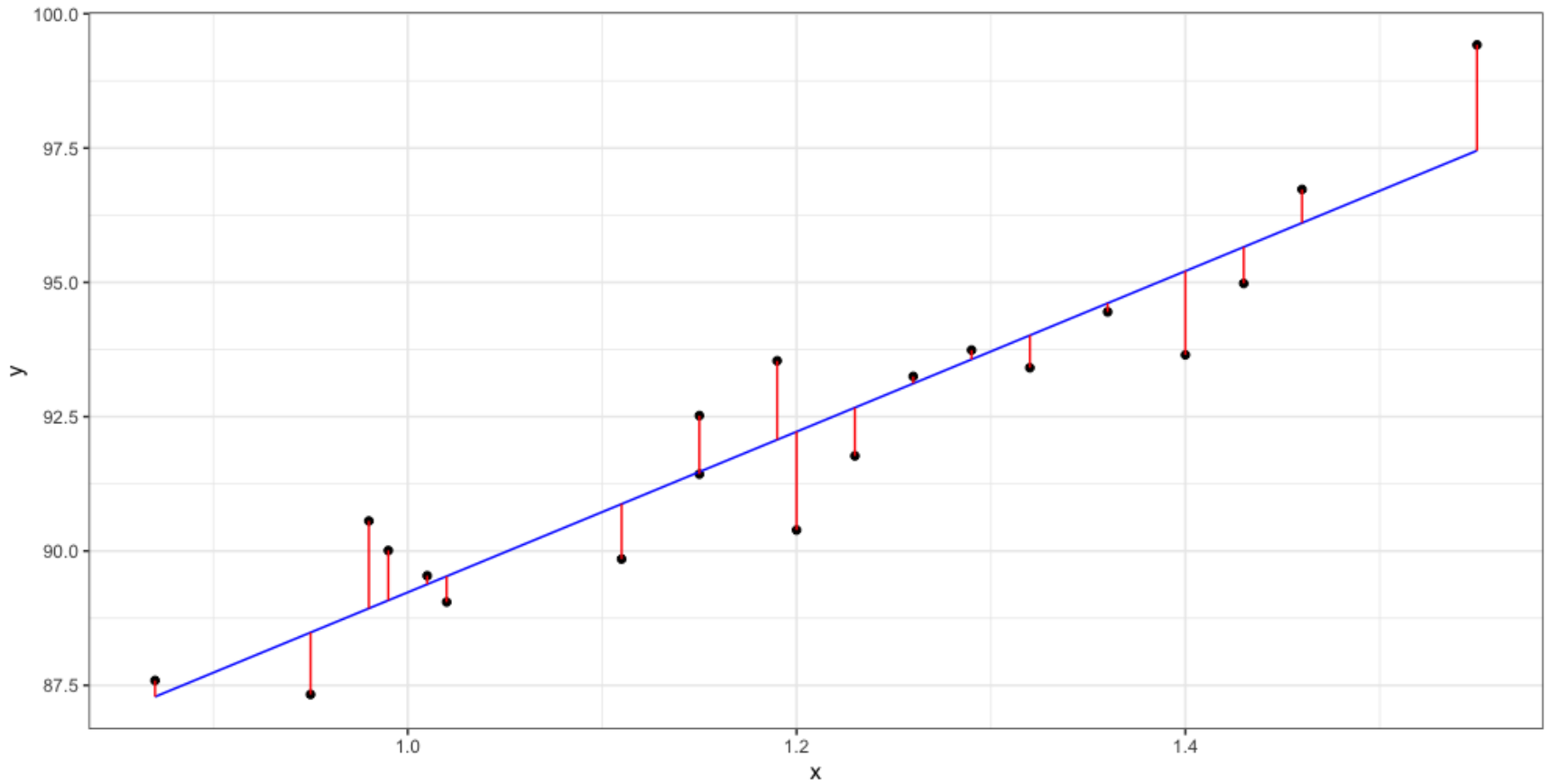
La **déviance lorsque** $X = X_i$ est la différence entre la réponse observée $Y_i$ et la réponse attendue $\mathrm{E}\{Y_i|X_i\}$ :

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

la déviation peut être **positive** (si le point se situe **au-dessus** de la ligne) ou **négatif** (s'il se situe **en dessous**).




droite d'ajustement




déviance (résidus)

Comment trouve-t-on des **estimateurs** de $\beta_0$ et $\beta_1$ ? Comment détermine-t-on si la droite d'ajustement est un **bon modèle pour les données** ?

Considérons la fonction

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

Si $Q(\boldsymbol{\beta})$ est "petit", le total des **résidus quadratiques** est "petit" et la droite $Y = \beta_0 + \beta_1 X$ est bien ajustée aux données.

Les **estimateurs de moindres carrés** du problème RLS sont $\mathbf{b} = (b_0, b_1)$ qui minimise la fonction $Q$ par rapport à $\boldsymbol{\beta} = (\beta_0, \beta_1)$.

On cherche les points critiques de $Q(\boldsymbol{\beta})$ : $\nabla_{\boldsymbol{\beta}} Q(\mathbf{b}) = \mathbf{0}$.

Ainsi, nous devons résoudre :

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

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

Il s'agit d'un système linéaire de deux équations à deux inconnues $\beta_0, \beta_1$, les **équations normales**.

En tant que tel, ce système possède soit **aucune solution**, soit une **solution unique**, soit **un nombre infini de solutions**.

**Note:** à partir de maintenant, nous laissons tomber le $\mid X_i$ dans $\mathrm{E}\{\cdot \mid X_i\}$.

# 2.1.1 – Équations normales

Ces équations se réduisent à la paire suivante :

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

Avec la notation

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

il n'est pas trop difficile de démontrer que

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

Avec cette notation, les équations normales se réduisent à

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

Sous forme matricielle, ceci peut s'écrire comme :

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

Un système linéaire $A\boldsymbol{\beta} = \mathbf{v}$ possède une solution **unique** $\boldsymbol{\beta} = A^{-1}\mathbf{v}$ si le déterminant de $A$ est non nul.

Dans ce cas, le déterminant est

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

La solution unique est

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

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

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

Posons $b_0=\beta_0$ et $b_1=\beta_1$. On peut alors écrire :

$$b_1 = \frac{S_{xy}}{S_{xx}}\ \textbf{(pente)} \quad \text{et} \quad b_0 = \overline{Y}-b_1\overline{X}\ \textbf{(ordonnée à l'origine)}.$$

Par analogie avec $S_{xx}$, nous pouvons également définir la **variation totale de la réponse** $S_{yy}$, une quantité qui jouera un rôle important dans le cours :

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

Si les $X_i$ sont fixes, $b_0, b_1$ sont des **combinaisons linéaires** des réponses $Y_i$ :

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

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

# Propriétés des estimateurs des moindres carrés

$b_0, b_1$ sont des **estimateurs sans biais** de leurs paramètres respectifs :

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

et

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

C'est le moment ou jamais d'illustrer ces notions à l'aide d'un exemple.

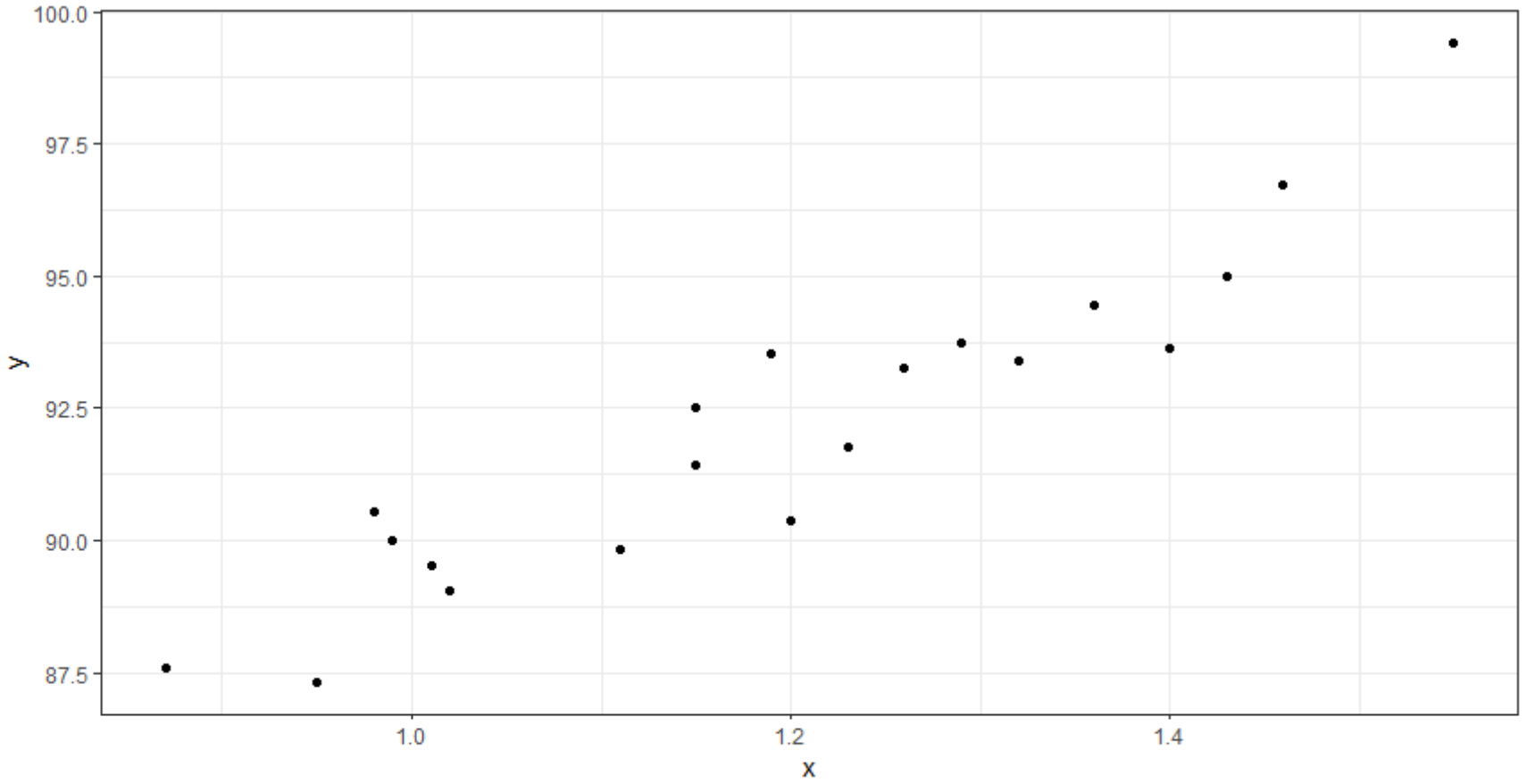
**Exemple de carburants :** considérons les mesures appariées $(X_i, Y_i)$ de teneurs en hydrocarbures $(X)$ et en oxygène pur $(Y)$ dans des carburants.

| $i$ | 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 |

Le modèle de régression simple est-il valide ? Si oui, ajustez les données.

**Solution:** on calcule les sommes fondamentales avec $n = 20$, et

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




Le modèle RLS est-il valable ?

Puisque le modèle RLS semble valide, nous calculons les estimateurs des moindres carrés :

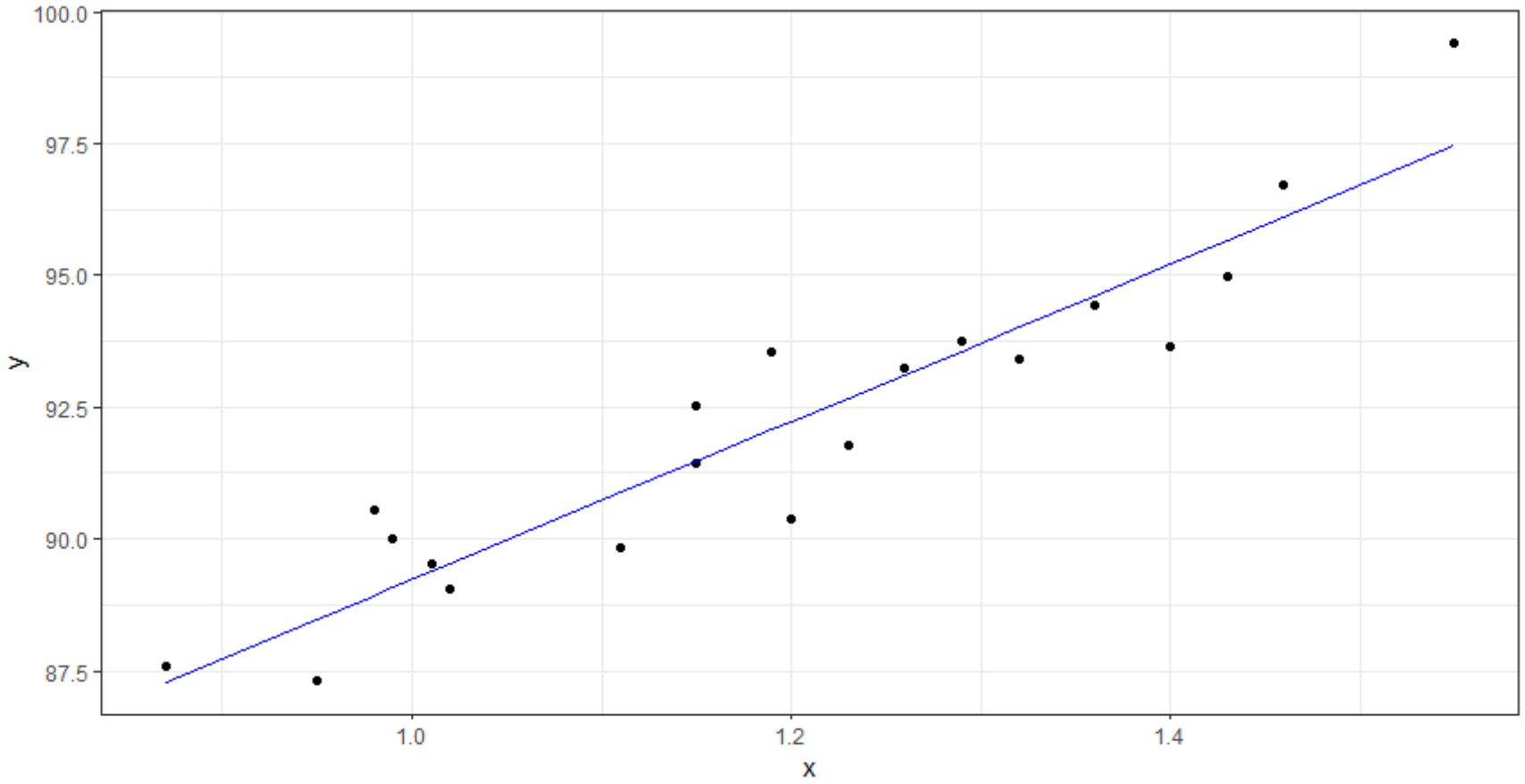
$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} = \frac{2214.66 - 20(\frac{23.92}{20})(\frac{1843.21}{20})}{29.29 - 20(\frac{23.92}{20})^2} = 14.947$$

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

Ainsi, la **droite d'ajustement** pour les données est

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

La $i$**ème valeur ajustée** est $\hat{Y}_i = \hat{f}(X_i) = b_0 + b_1 X_i$, $i = 1, \ldots, n$.

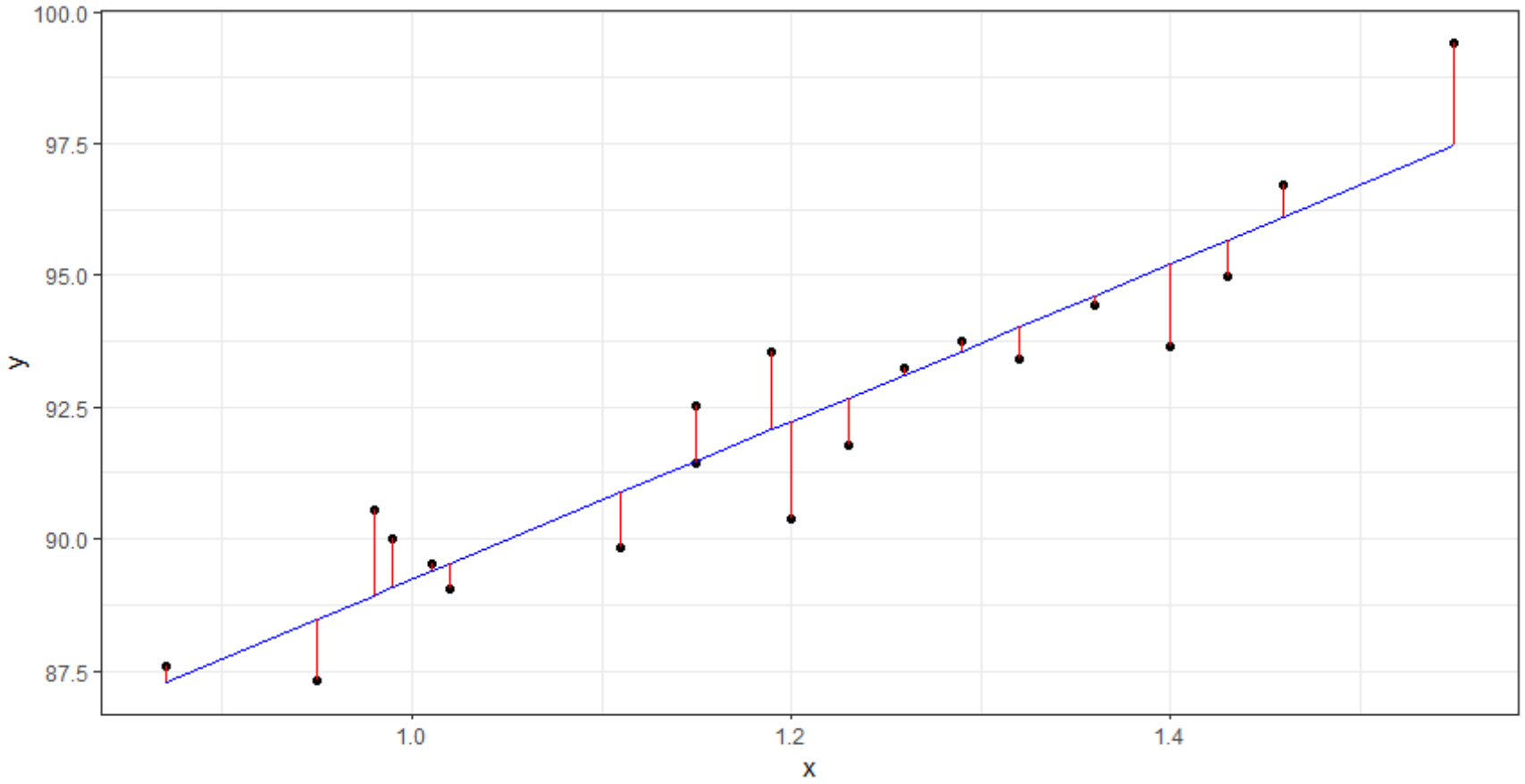



droite d'ajustement : $\hat{Y} = 74.283 + 14.947X$

# 2.1.2 – Résidus

Le $i$**ème résidu** est $e_i = Y_i - \hat{Y}_i$. Les résidus ont les propriétés suivantes :

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. $(\overline{X}, \overline{Y})$ se retrouve sur la **droite d'ajustement**
6. $\sum_{i=1}^{n} e_i^2$ est **minimale** (au sens MC)




résidus dans l'exemple des carburants

**Démonstration :**

1. Nous avons

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

   selon les équations normales.

2. Selon 1., nous avons $0 = \overline{e}$. 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. Nous avons

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

selon la seconde équation normale.

4. Nous avons

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

selon 1. et 3.

5. C'est automatiquement le cas puisque

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

6. Pour tout $\mathbf{b}^* = (b_0^*, b_1^*) \neq \mathbf{b} = (b_0, b_1)$, nous avons $Q(\mathbf{b}^*) \geq Q(\mathbf{b})$. On dénote les résidus obtenus à partir de la droite ajustée par $\mathbf{b}^*$ par $e_i^*$. Alors

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

Cela complète la démonstration. ■

# 2.1.3 – Statistiques descriptives et corrélations

Le **coefficient de corrélation d'échantillon de Pearson** $r$ entre 2 variables $X$ et $Y$ est défini par

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

Ce coefficient est tel que

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

2. $|r| = 1 \iff Y_i = b_0 + b_1 X_i$, pour tout $i = 1, \ldots, n$, et

3. $\text{sgn}(r) = \text{sgn}(b_1)$, d'où $r = 0 \iff b_1 = 0$.

Si $|r| \approx 1$, alors il y a une **association linéaire forte** entre $X$ et $Y$.

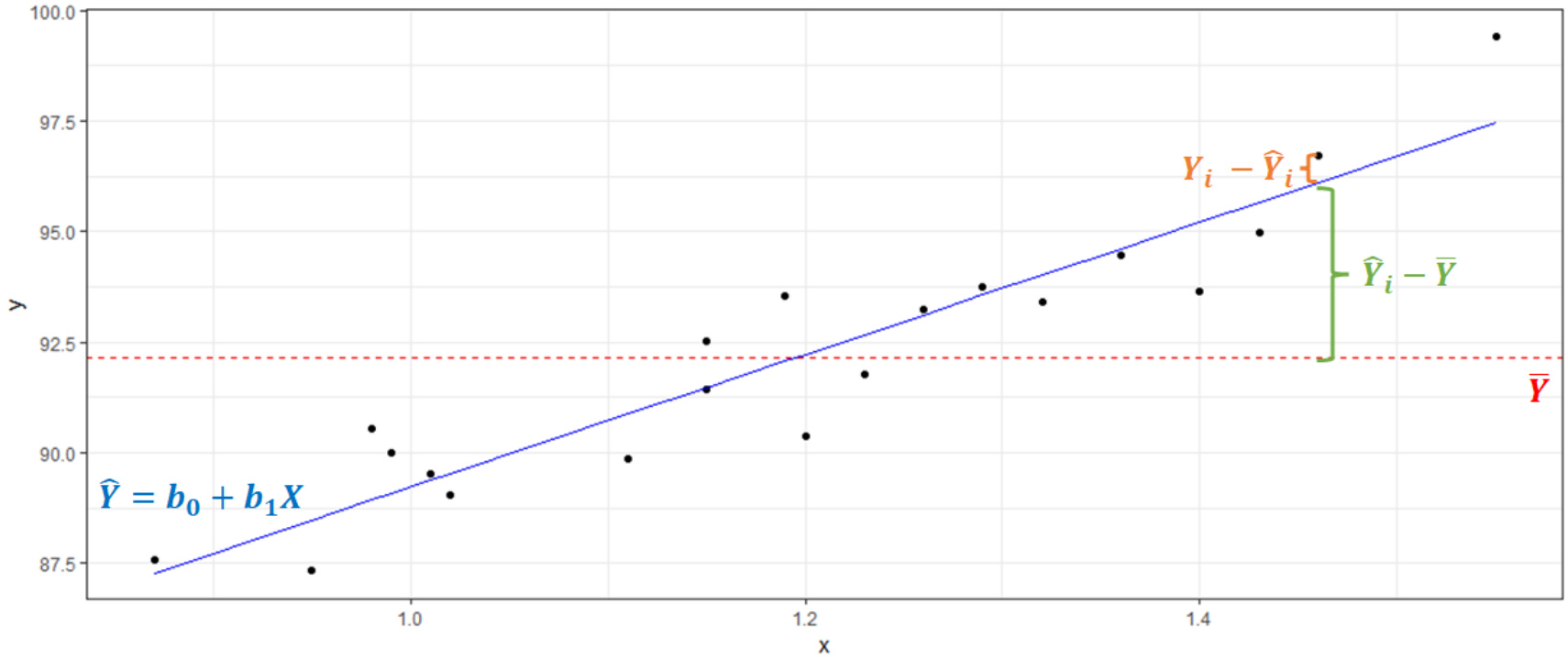
Si $|r| \approx 0$, l'**association linéaire** entre $X$ et $Y$ est **faible**.

Que dire que $0 \ll |r| \ll 1$? Nous en reparlerons à la section 2.1.5.

Pour l'instant, nous remarquerons seulement que nous pouvons **décomposer** la déviance total comme suit :

$$\underbrace{Y_i - \overline{Y}}_{\substack{\text{déviance} \\ \text{totale} \\ \text{de la moyenne}}} = \underbrace{(Y_i - \hat{Y}_i)}_{\substack{\textbf{déviance} \\ \textbf{de la moyenne} \\ \textbf{non expliquée}}} + \underbrace{(\hat{Y}_i - \overline{Y})}_{\substack{\textbf{déviance} \\ \textbf{de la moyenne} \\ \textbf{expliquée par} \\ \textbf{la régression}}}.$$

Cette décomposition est représentée graphiquement à la diapositive suivante.




Décomposition de la déviance totale

Le **coefficient de corrélation d'échantillon de Spearman** $r_S$ entre 2 variables $X$ et $Y$ est définie par la **corrélation de Pearson** entre les **valeurs des rang** $R(X_i)$ et $R(Y_i)$ de $X_i$ et $Y_i$, respectivement.

Ce coefficient est tel que

1. $-1 \leq r_S \leq 1$;
2. $r_S = 1 \iff$ la relation entre $X$ et $Y$ est **monotone croissante**,
3. $r_S = -1 \iff$ la relation entre $X$ et $Y$ est **monotone décroissante**,
4. si l'association entre $X$ et $Y$ est **faible**, alors $r_S \approx 0$, et
5. $r_S$ est invariant sous les **transformations préservant l'ordre**.

La procédure de calcul est simple : pour les mesures

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

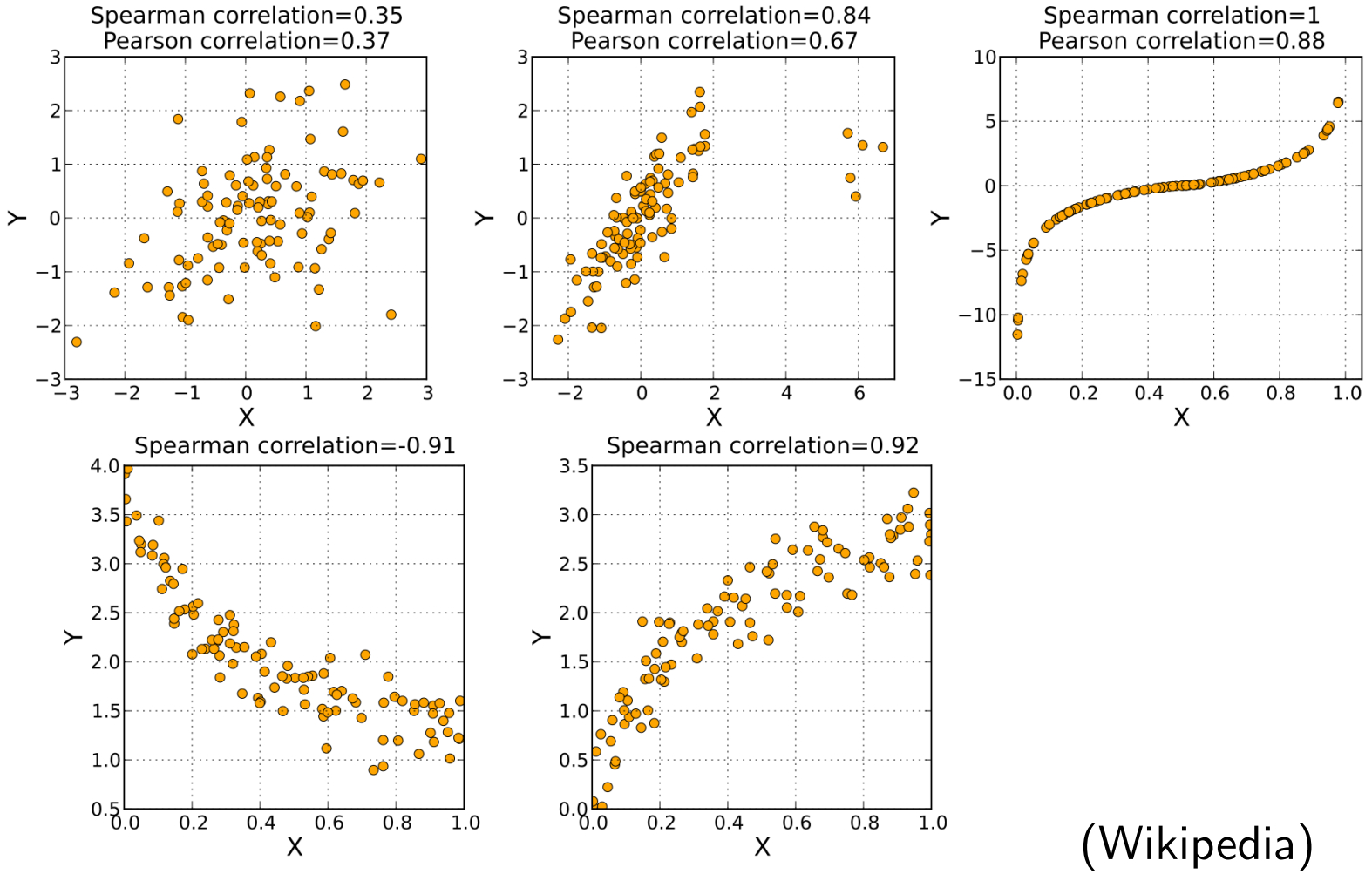
soit $R(Z_i)$ la **valeur de rang** de $Z_i$ dans $\mathcal{Z}$ ; la plus petite valeur de $Z_i$ est de rang $1$, la deuxième plus petite est de rang $2$, et ainsi de suite, jusqu'à la plus grande valeur, de rang $n$.

Les valeurs égales sont traitées comme dans l'exemple ci-dessous :

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

Formellement,

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


Spearman correlation=0.35
Pearson correlation=0.37
Spearman correlation=0.84
Pearson correlation=0.67
Spearman correlation=1
Pearson correlation=0.88
Spearman correlation=-0.91
Spearman correlation=0.92
X
Y

(Wikipedia)

# 2.1.4 – Décomposition en sommes de carrés

La décomposition en sommes de carrés (SS) est un des concepts fondamentaux de l'analyse de régression :

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

Ceci est s'écrit souvent sous la forme $\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}$, où

- $\mathrm{SST}$ est la **somme totale des carrés**,
- $\mathrm{SSE}$ est la **somme des carrés de l'erreur**, et
- $\mathrm{SSR}$ est la **somme des carrés de la régression**.

Notez que nous pouvons aussi écrire

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

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

Comme $\mathrm{SST} = S_{yy}$ et $\mathrm{SSE} = Q(\mathbf{b})$, la décomposition se ré-écrit sous la forme

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

Dans l'exemple des carburants, nous obtenons

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

de sorte que le coefficient de corrélation de l'échantillon est

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

et la décomposition en SS est $\mathrm{SST}\,(173.38) = \mathrm{SSR}\,(152.13) + \mathrm{SSE}\,(21.25)$. S'agit-il d'une forte association linéaire ?

# 2.1.5 – Coefficient de détermination

Le **coefficient de détermination** $R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}}$ est la proportion de variation de la réponse expliquée par la droite d'ajustement.

Lorsque $R^2 \approx 0$, la régression est **peu significative**, alors que lorsque $R^2 \approx 1$, les variables sont **fortement liées** par la droite.

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

**Démonstration:** nous avons vu que $\mathrm{SSR} = b_1^2 S_{xx}$ et $\mathrm{SST} = S_{yy}$, d'où

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

Ceci répond à la question relative à l'interprétation de $0 \ll |r| \ll 1$ : $r^2$ donne une idée de la quantité de variation que la régression "explique".

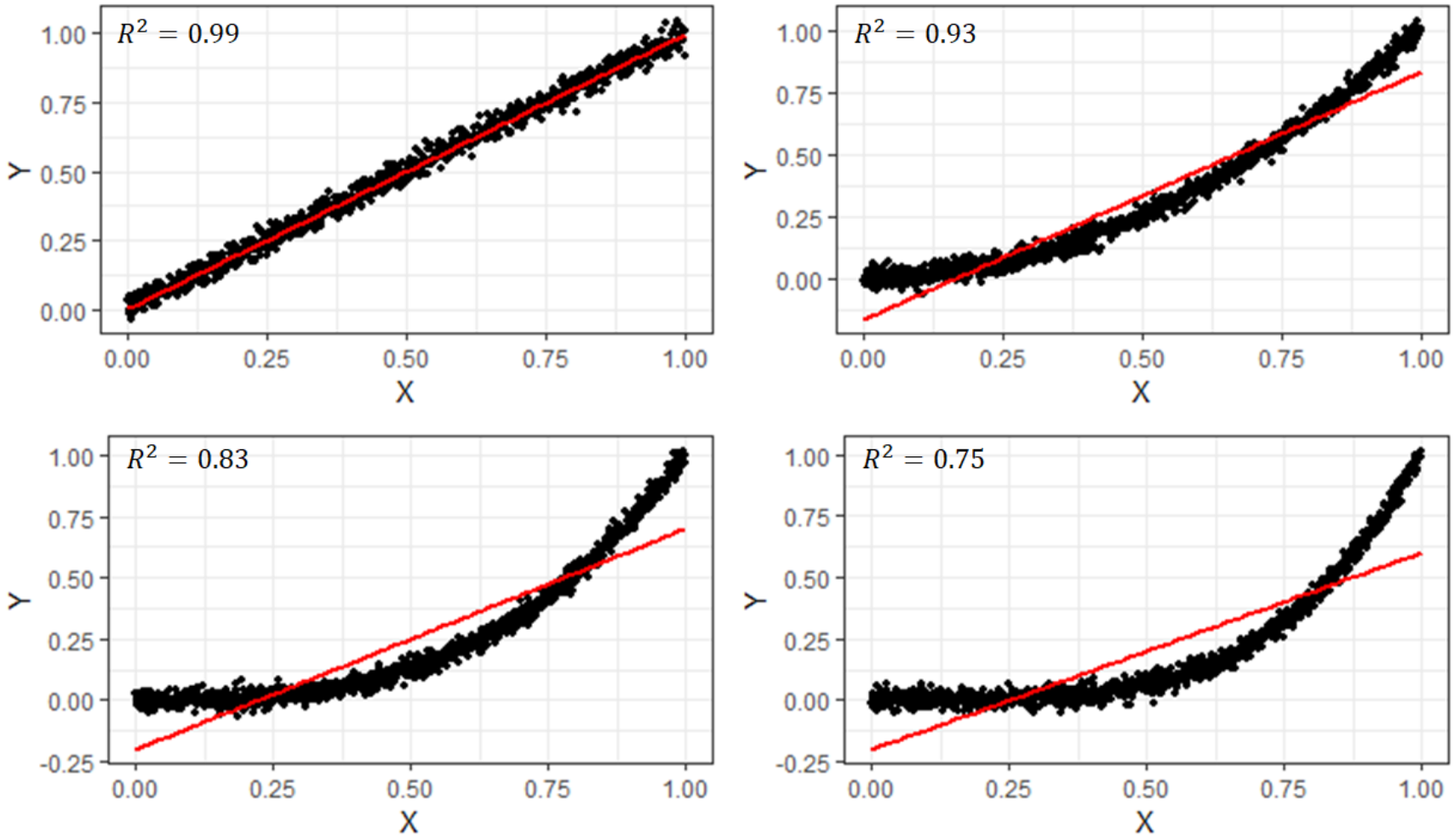
Dans l'exemple des carburants, nous obtenons

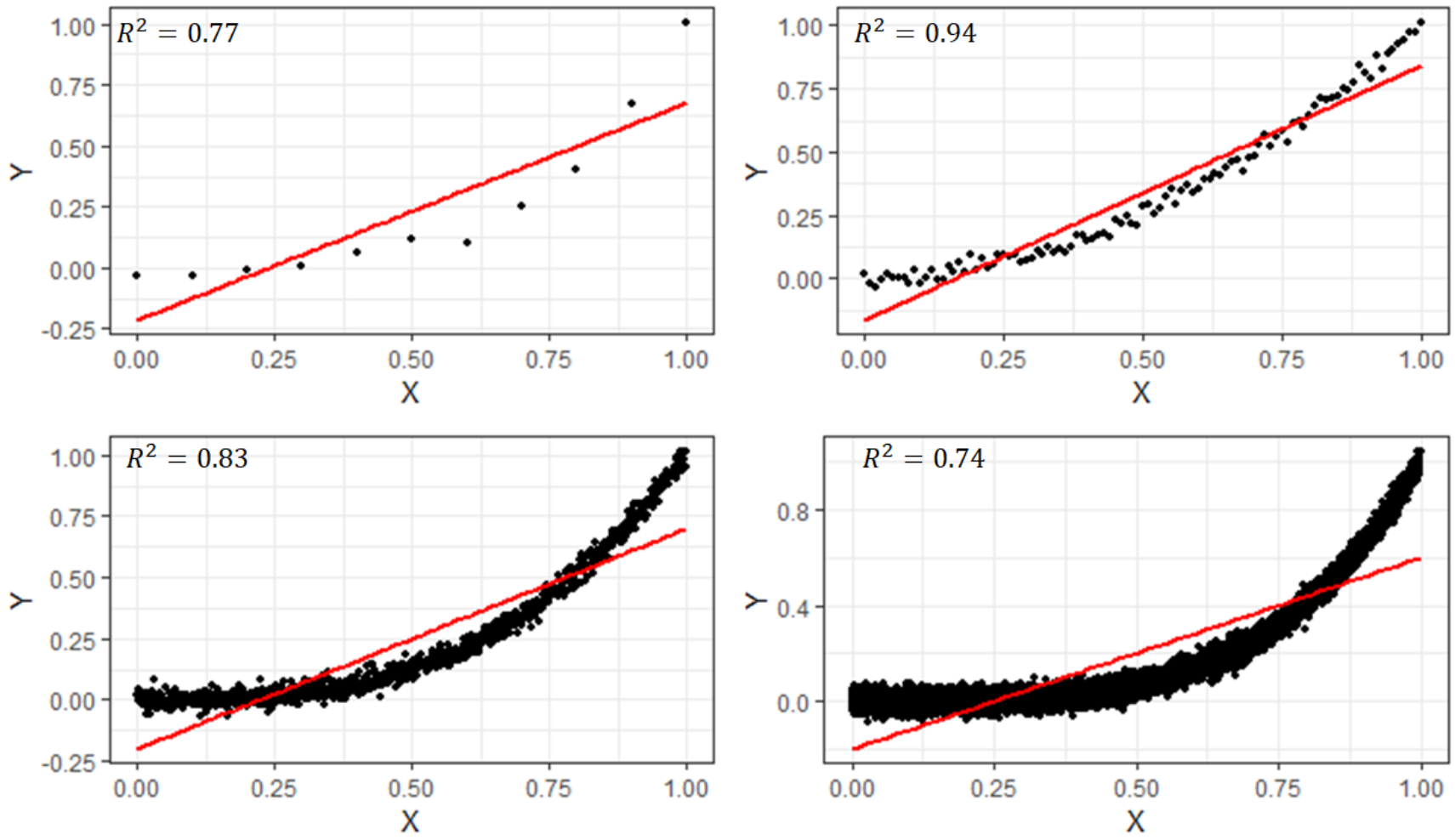
$$R^2 = \frac{152.13}{173.98} = 0.8774;$$

ainsi, environ $87.7\%$ de la variation observée dans les données peut être expliquée par la droite ajustée $\hat{Y} = 74.283 + 14.947X$.

C'est une proportion **raisonnablement élevée** ; avec le diagramme de dispersion, cela suggère que le modèle RLS est probablement approprié.

Mais ne vous enflammez pas trop pour $R^2$ en tant que statistique permettant de valider l'ajustement (voir pages suivantes).


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


$R^2 = 0.77$
$R^2 = 0.94$
$R^2 = 0.83$
$R^2 = 0.74$
X
Y

# 2.2 – Inférence

Nous avons besoin d'une estimation de la **variance commune** $\sigma^2$ afin de tester diverses hypothèses sur la régression.

Dans le modèle RLS

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

nous avons des erreurs aléatoires normales indépendantes $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

La f.d.p. de $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$ est ainsi

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

La **fonction de vraisemblance** est

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

où

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

La vraisemblance $L$ est maximale lorsque $Q$ est minimal par respect à $\beta_0, \beta_1$. On a déjà montré ce minimum se retrouve à l'**estimateur de la vraisemblance maximale** $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1) = (b_0, b_1)$, pour lequel

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

Peut-on aussi utiliser les données afin de trouver un estimateur de $\sigma^2$?

On considère la **log-vraisemblance**

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

Comme $\ln$ est une fonction **monotone croissante**, maximiser $L$ revient à maximiser $\ln L$. Mais

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

En fixant $\frac{\partial L}{\partial[\sigma^2]} = 0$ et en résolvant pour $\sigma^2$, on obtient

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

Cet estimateur est cependant **biaisé** ; peut montrer que $\text{E}\left\{\widehat{\sigma^2}\right\} = \frac{n-2}{n}\sigma^2$.

L'**erreur quadratique moyenne**

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

donne un autre estimateur (**sans biais**) de la variance de la population $\sigma^2$ :

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

Nous pouvons considérer la variance $\sigma^2$ d'une **population infinie** de taille $n$ comme une somme de carrés divisée par ses degrés de liberté $n$ :

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

L'estimateur de la variance $\sigma^2$ qui utilise un **échantillon** de taille $n$ est

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

c'est une somme de carrés divisée par $n-1$, ses degrés de liberté ; $1$ degré de liberté est perdu car nous avons d'abord utilisé l'échantillon pour calculer la **moyenne de l'échantillon** $\overline{Y}$ comme approximation de $\mu$.

Lorsque l'on utilise les mêmes données à deux fins différentes, on crée un "lien" entre $s^2$ et $\overline{Y}$ qui n'existait pas entre $\sigma^2$ et $\mu$.

Le même raisonnement explique pourquoi il ne faut pas s'étonner qu'il faille diviser $\mathrm{SSE}$ par $n-2$ pour obtenir un estimateur sans biais de $\sigma^2$ : dans la SS des erreurs

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

nous devons d'abord utiliser les données pour estimer 2 quantités, $\beta_0$ et $\beta_1$. Ainsi, $\mathrm{SSE}$ a $n-2$ degrés de liberté, et l'estimateur sans biais de $\sigma^2$ est

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

Dans l'exemple des carburants ($n = 20$), nous avons obtenu $\mathrm{SSE} = 21.25$. L'**estimateur sans biais** de la variance d'erreur $\sigma^2$ dans le modèle RLS est

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

En général, si le modèle RLS est valide, nous nous attendrions à ce que $\mathrm{E}\{Y_i\} = \beta_0 + \beta_1 X_i$ soit un modèle raisonable pour tout échantillon.

Mais les **valeurs spécifiques** pour les estimateurs $b_0, b_1$ dépendent des **données disponibles**. Avec différentes observations, nous obtiendrions différentes valeurs pour les estimateurs, et il est logique d'étudier l'**erreur-type de** $b_0, b_1$ :

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

# 2.2.1 – Inférence sur la pente

En théorie, nous pourrions donc

1. recueillir $M$ échantillons **indépendants**,

2. répéter la procédure des moindres carrés et obtenir une estimation de la pente $b_{1;j}$ de $\beta_1$ **pour chaque ensemble de données** $j$, et

3. donner une approximation de $\sigma\{b_1\}$ en calculant l'**écart-type de l'échantillon** $\{b_{1;1}, \ldots, b_{1;M}\}$.

En pratique, cependant, la collecte de données est souvent **coûteuse** et il se peut que nous n'ayons jamais accès à plus d'un échantillon.

Il y a d'autres options (bootstrap, jackknife), mais on peut utiliser la machinerie de la régression afin d'obtenir des estimations de l'erreur-type à partir d'un **échantillon unique**.

Comme les termes d'erreur $\varepsilon_1, \ldots, \varepsilon_n$ sont indépendants dans le modèle RLS, les valeurs de réponse $Y_1, \ldots, Y_n$ sont non corrélées, de variance $\sigma^2\{Y_i\} = \sigma^2\{\beta_0 + \beta_1 X_i + \varepsilon_i\} = \sigma^2\{\varepsilon_i\} = \sigma^2$ pour $i = 1, \ldots, n$. Puisque

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

de sorte à ce que

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

Comme nous ne connaissons généralement pas la valeur réelle de $\sigma^2$, l'**erreur-type estimée de $b_1$** est :

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

Dans l'exemple des carburants, nous avons

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

La v.a. $b_1$ est en fait une combinaison linéaire des v.a. **normales indépendantes** $Y_1, \ldots, Y_n$, ce qui veut dire qu'elle suit elle-même une **loi normale**, selon le TLC.

Mais nous connaissons déjà son espérance et sa variance, d'où nous connaissons sa distribution :

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

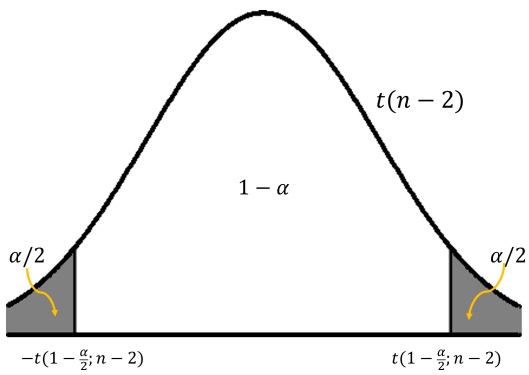
Nous posons maintenant des hypothèses qui seront justifiées plus tard :

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

Selon la définition de la loi $T$ de Student, nous obtenons

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

# Région critique




Soit $\alpha \in (0,1)$. Puisque $\frac{b_1-\beta_1}{\mathrm{s}\{b_1\}} \sim t(n-2)$, nous avons

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

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

Ainsi, on obtient un **intervalle de confiance de $\beta_1$ à environ** $100(1-\alpha)\%$ par l'entremise de

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

Dans l'exemple des carburants, nous obtenons

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

À un **niveau de confiance** de $1 - \alpha = 0.95$ (ou un **taux d'erreur** de $\alpha = 0.05$), la valeur critique de la loi $T$ de Student avec $n-2 = 20-2 = 18$ degrés de liberté est

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

On peut ainsi construire un intervalle de confiance de $\beta_1$ à environ 95% comme suit :

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

# 2.2.2 – Inférence sur l'ordonnée à l'origine

En utilisant les mêmes hypothèses qu'avec $b_1$, on obtient pareillement :

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

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

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

Ainsi,

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

et l'erreur-type estimée de $b_0$ est :

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

Dans l'exemple des carburants, nous obtenons

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

Comme c'était le cas pour $b_1$, $b_0$ suit une loi normale, étant une combinaison linéaire des v.a. **normales indépendantes** $Y_1, \ldots, Y_n$.

Comme nous connaissons déjà son espérance et sa variance, nous connaissons également sa distribution :

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

En supposant à nouveau que $b_0$ et $\mathrm{SSE}$ sont indépendants et que $\frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2(n-2)$, la définition de la loi $T$ de Student donne que

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

suit une loi $t(n-2)$.

Comme c'était le cas pour $\beta_1$, l'intervalle de confiance de $\beta_0$ à environ $100(1-\alpha)\%$ est

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

Dans l'exemple des carburants, nous obtenons

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

Pour un **niveau de confiance** de $1-\alpha = 0.95$, la valeur critique de la loi $T$ de Student avec $n-2 = 20-2 = 18$ degrés de liberté est $t(1-0.05/2; 20-2) = t(0.975; 18) = 2.101$.

On peut alors construire un intervalle de confiance de $\beta_0$ à environ $95\%$ :

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

# 2.2.3 – Tests d'hypothèses

Avec les erreurs standard, nous pouvons **tester des hypothèses**.

Nous essayons de déterminer si les paramètres $\beta_0, \beta_1$ prennent des valeurs spécifiques, et si la droite d'ajustement fournit une bonne description d'un ensemble de données à deux variables, en suivant les étapes suivantes :

1. établir une **hypothèse nulle** $H_0$ et une **hypothèse alternative** $H_1$ ;
2. calculer la **statistique de test** (en utilisant la studentisation) ;
3. trouver une **région critique**/valeur$-p$ pour la statistique de test sous $H_0$ ;
4. **rejeter** ou **ne pas rejeter** $H_0$ en fonction de la région critique/valeur$-p$.

Par exemple, nous pourrions être intéressés à tester si la valeur réelle du paramètre $\beta$ est égale à une **valeur candidate** $\beta^*$, c'est-à-dire

$$H_0 : \beta = \beta^* \quad \text{vs.} \quad H_1 : \begin{cases} \beta < \beta^*, & \text{test unilatéral à gauche} \\ \beta > \beta^*, & \text{test unilatéral à droite} \\ \beta \neq \beta^*, & \text{test bilatéral} \end{cases}$$

Si $H_0$ est valide, nous avons déjà montré que

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

La **région critique** dépend du niveau de confiance $1 - \alpha$ et du **type** de l'hypothèse alternative $H_1$.

Soit $t^*$ la valeur observée de $T_0$. Nous **rejetons** $H_0$ si $t^*$ se retrouve dans la région critique.

| **Hypothèse alternative** | **Région de rejection** |
|---|---|
| $H_1 : \beta < \beta^*$ | $t^* < -t(1-\alpha; n-2)$ |
| $H_1 : \beta > \beta^*$ | $t^* > t(1-\alpha; n-2)$ |
| $H_1 : \beta \neq \beta^*$ | $\lvert t^* \rvert > t(1-\alpha/2; n-2)$ |

**Exercices:** testez les hypothèses suivantes dans l'exemple des carburants.

a) Testez pour $H_0 : \beta_0 = 75$ vs. $H_1 : \beta_0 < 75$ lorsque $\alpha = 0.05$.

b) Testez pour $H_0 : \beta_1 = 10$ vs. $H_1 : \beta_1 > 10$ lorsque $\alpha = 0.05$.

c) Testez pour $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ lorsque $\alpha = 0.05$.

**Solutions:** nous savons que

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

Comme le taux d'erreur pour tous les tests est de $\alpha = 0.05$, nous devons calculer les valeurs critiques de la loi $T$ de Student avec $\nu = 20 - 2 = 18$ degrés de liberté, aux niveaux de confiance $1-\alpha = 0.95$ et $1-\alpha/2 = 0.975$ :

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

a) On effectue un test **unilatéral à gauche** : la statistique observée est

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

et donc nous **ne rejetons pas** $H_0$ lorsque $\alpha = 0.05$.

b) On effectue un test **unilatéral à droite** : la statistique observée est

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

et donc nous **rejetons $H_0$ en faveur de $H_1$** lorsque $\alpha = 0.05$.

c) On effectue un test **bilatéral** : la statistique observée est

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

et donc nous **rejetons $H_0$ en faveur de $H_1$** lorsque $\alpha = 0.05$.

Nous étudierons un autre test pour la pente à la section 2.4.

# 2.2.4 – Inférence sur la réponse moyenne

Nous pouvons également effectuer une analyse inférentielle pour la **réponse attendue** à $X = X^*$ (en pratique, il pourrait y avoir des répétitions).

Comme précédemment, nous supposons que $\mathrm{E}\{Y^*\} = \beta_0 + \beta_1 X^*$. La **réponse moyenne estimée** à $X = X^*$ est

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

Les valeurs du prédicteur sont **fixes**, donc $\hat{Y}^*$ suit une loi normale avec

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

$\hat{Y}^*$ est un **estimateur sans biais** de $Y^*$. Quelle est son erreur-type ?

Si $b_0, b_1$ étaient indépendants, nous pourrions simplement calculer

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

Mais ils ne le sont pas.

**Théorème:** sous les hypothèses RLS, $\sigma\left\{\overline{Y}, b_1\right\} = 0$ et

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

**Démonstration:** tout au long, gardez à l'esprit que les $Y_i$ sont **non corrélés**. Nous avons

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

Tous les termes pour lesquels $i \neq j$ ont $\sigma\{Y_i, Y_j\} = 0$, les autres ont $\sigma\{Y_i, Y_i\} = \sigma^2\{Y_i\} = \sigma^2$, donc

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

De même,

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

Tous les termes pour lesquels $i \neq j$ ont $\sigma\{Y_i, Y_j\} = 0$, les autres ont $\sigma\{Y_i, Y_i\} = \sigma^2\{Y_i\} = \sigma^2$, donc

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

Ceci complète la démonstration. ■

Nous pouvons maintenant déterminer l'erreur-type de la réponse moyenne estimée $Y = \hat{Y}^*$ en $X = X^*$ :

$$\begin{aligned}\sigma^2\left\{\hat{Y}^*\right\} &= \sigma^2\left\{b_0 + b_1 X^*\right\} = \sigma^2\left\{b_0\right\} + (X^*)^2\sigma^2\left\{b_1\right\} + 2\sigma\left\{b_0, X^* b_1\right\} \\ &= \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2}{S_{xx}}\right] + \frac{(X^*)^2\sigma^2}{S_{xx}} - 2X^*\overline{X}\frac{\sigma^2}{S_{xx}} \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}}\left[(X^*)^2 - 2\overline{X}X^* + \overline{X}^2\right] = \sigma^2\left[\frac{1}{n} + \frac{(X^* - \overline{X})^2}{S_{xx}}\right].\end{aligned}$$

L'erreur-type estimée est donc

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

Mais il y a plusieurs façons de plumer un canard :

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

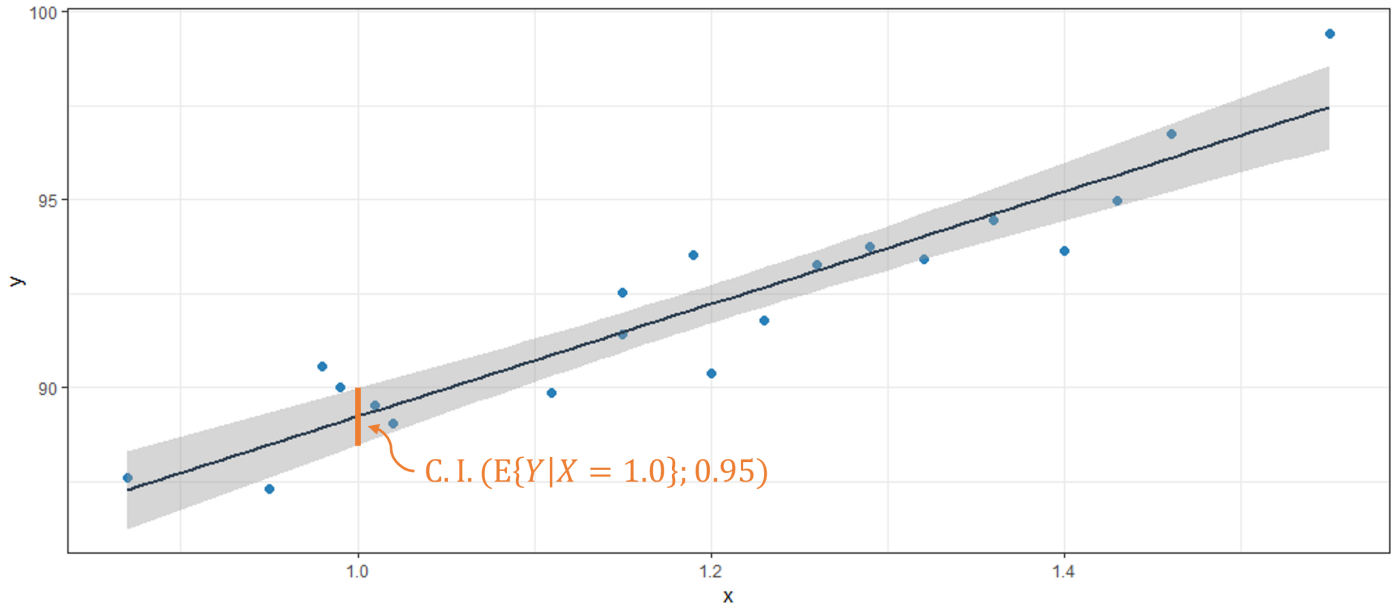
Dans les deux cas, nous pouvons montrer que

$$T^* = \frac{\hat{Y}^* - \mathrm{E}\{\hat{Y}^*\}}{\mathrm{s}\{\hat{Y}^*\}} \sim t(n-2), \quad \text{d'où}$$

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

Dans l'exemple des carburants, l'intervalle de confiance de $\mathrm{E}\{Y^*\}$ à environ $95\%$ est

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



# 2.3 – Estimation et prédiction

Lorsque nous estimons la réponse **attendue** $\mathrm{E}\{Y^*\}$, nous déterminons comment $(b_0, b_1)$ pourrait **conjointement** varier d'un échantillon à l'autre.

Comme ces paramètres déterminent de façon unique la droite de meilleur ajustement, trouver un intervalle de confiance pour la réponse moyenne à tous les $X = X^*$ est (plus ou moins) équivalent à trouver une **bande de confiance** pour la ligne entière sur le domaine du prédicteur ( ⚠ ).

Il n'est pas surprenant qu'un certain nombre d'observations se situent en dehors de leurs intervalles de confiance respectifs pour l'exemple de l'ensemble de données sur les carburants : nous estimions la **réponse moyenne** à un niveau de prédicteur $X = X^*$, et non la **réponse réelle** (ou nouvelle) à ce niveau.

Et si nous cherchons une étendue de **réponses probables** en $X = X^*$ ?

Nous utilisons les données disponibles pour construire des **intervalles de confiance** (I.C.) lorsque nous nous intéressons à certaines caractéristiques (fixes) de la population qui nous sont inconnues.

Mais une nouvelle valeur de la réponse n'est pas un paramètre ; c'est une **variable aléatoire** ; l'intervalle des valeurs plausibles (probables) d'une nouvelle réponse est un **intervalle de prédiction** (I.P.) plutôt qu'un I.C.

Afin de déterminer un I.P. pour la réponse, nous devons modéliser l'**erreur** impliquée dans la prédiction de la réponse.

**Note:** nous supposons que les nouvelles réponses en $X = X^*$ sont indépendantes des réponses observées (**les résidus ne sont pas corrélés**).

# 2.3.1 – Intervalle de prédiction

Soit $Y_p^*$ une **réponse (future)** en $X = X^*$ ; nous avons

$$Y_p^* = \beta_0 + \beta_1 X^* + \varepsilon_p \quad \text{pour un certain } \varepsilon_p.$$

Si l'erreur moyenne est nulle, la meilleure prédiction pour $Y_p^*$ est toujours la **réponse sur la droite ajustée en** $X = X^*$ :

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

L'**erreur de prédiction** en $X = X^*$ est ainsi

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

Dans le modèle de RLS, $\varepsilon_p$ et $b_0, b_1$ suivent des **lois normales**. Par conséquent, il en est de même pour l'erreur de prédiction $\text{pred}^*$. Notons que

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

Comme les résidus ne sont pas corrélés (cf. section 2.3), nous avons

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

Ainsi

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

L'erreur-type estimée est donc

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

Comme précédemment, nous pouvons montrer que

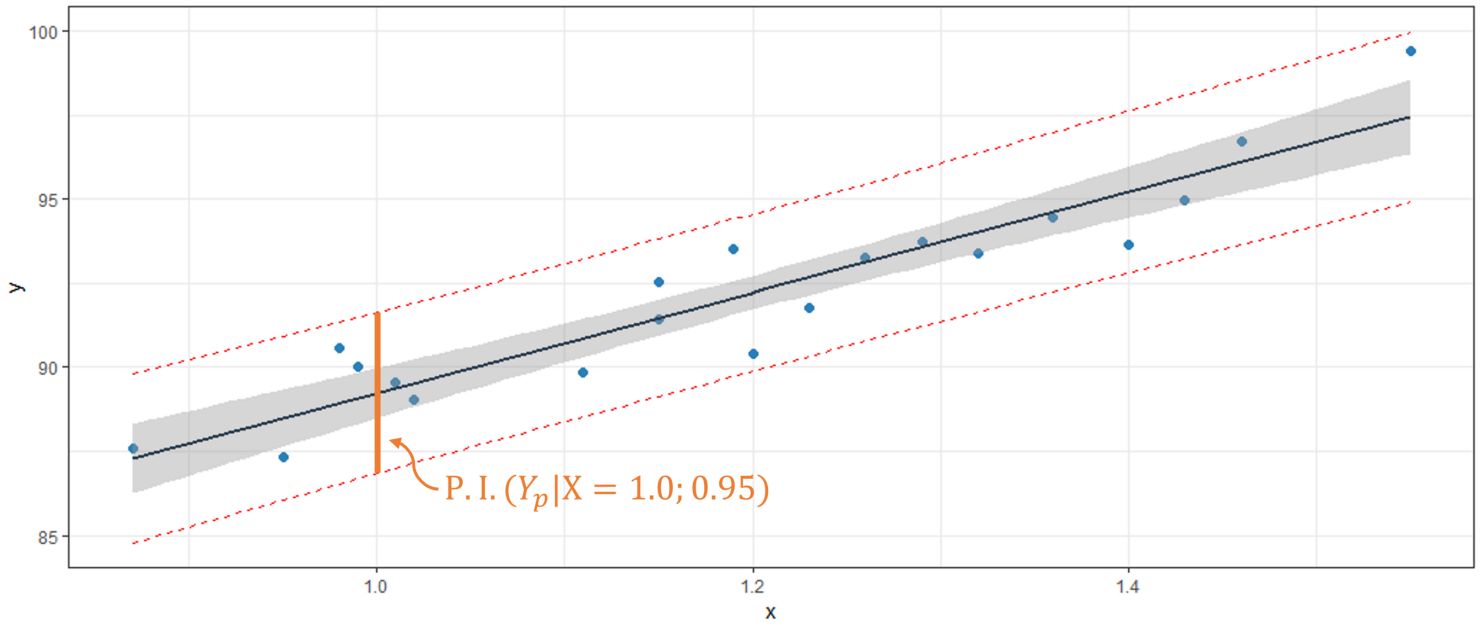
$$T_p^* = \frac{\mathrm{pred}^* - 0}{\mathrm{s}\{\mathrm{pred}^*\}} \sim t(n-2), \quad \text{d'où}$$

$$\mathrm{I.P.}(Y_p^*; 1-\alpha) \equiv b_0 + b_1 X^* \pm t(1 - \tfrac{\alpha}{2}; n-2) \cdot \mathrm{s}\{\mathrm{pred}^*\}.$$

Notons que $\mathrm{s}\{\hat{Y}^*\} < \mathrm{s}\{\mathrm{pred}^*\}$, de sorte que l'I.C. de la réponse moyenne est toujours **contenu** dans l'I.P. pour les nouvelles réponses. Ils sont minimisés lorsque $X^* = \overline{X}$ ; ils sont plus large lorsque $|X^* - \overline{X}|$ augmente.

Dans l'exemple des carburants, l'I.P. de $Y_p^*$ à environ $95\%$ est

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



# Tests d'hypothèses

Puisque les estimateurs de la réponse moyenne et des nouvelles réponses suivent des lois normales et puisque nous disposons d'estimations pour les erreurs-type, nous pouvons effectuer les tests d'hypothèses comme au préalable :

1. identifier le **type** d'hypothèse alternative $H_1$ (unliatéral à gauche, à droite, bilatéral) ;

2. calculer la **statistique de test observée** (studentisée), et

3. comparer à la **valeur critique** appropriée de la loi $T$ de Student.

Par exemple, dans l'exemple des carburants, supposons que nous voulions tester

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

Sous $H_0$, la statistique de test satisfait

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

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

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

La valeur observée de $T^*$ est ainsi

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

À un taux d'erreur de $\alpha = 0,05$, la valeur critique de la loi $T$ de Student avec $n - 2 = 18$ degrés de liberté est

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

Puisque $|t^*| \not> t(0.975; 18)$, l'évidence n'est pas assez solide pour rejeter l'hypothèse nulle $H_0$ à un niveau de confiance de $95\%$ ...

(... **ce qui n'est pas la même chose que d'accepter l'hypothèse nulle** $H_0$).

Et si nous observons une nouvelle réponse $Y_p^* = 80$ lorsque $X^* = 1.2$ ? S'agit-il d'une valeur raisonnable ou devons-nous nous attendre à quelque chose de plus grand (ou plus petit)?

A un niveau de confiance de $95\%$, l'intervalle de prédiction pour la réponse lorsque $X^* = 1.2$ est de

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

Comme $Y_p^* = 80$ n'est pas dans l'intervalle de prédiction, cela semble être une réponse **improbable** pour $X^* = 1.2$ (à un niveau de confiance de $95\%$).

# 2.3.2 – Estimations et prédictions simultanées

Lorsque nous utilisons un ensemble de données pour estimer les deux paramètres $\beta_0$ et $\beta_1$ dans le modèle de RLS, $\mathrm{SSE}$ a $n-2$ degrés de liberté.

Cela peut sembler être un point technique obscur, mais il y a une conséquence pratique : les $\mathrm{I.C.}$ résultants sont nécessairement **plus larges** que ceux qui seraient obtenus si la SS avait plus de degrés de liberté.

Par exemple, $t(0.975; 18) = 2.101 > t(0.975, 20) = 2.086$. Qu'est-ce que cela signifie pour l'analyse de régression ?

Il y a une **pénalité** associée à l'estimation simultanée des paramètres : lorsque les mêmes données sont utilisées pour calculer plusieurs estimations, elles se **"fatiguent"** (?) et **perdent** une partie de leur pouvoir prédictif.

# Procédure de Bonferroni

Nous nous intéressons à l'estimation **conjointe** de $g$ paramètres $\theta_1, \ldots, \theta_g$.

Pour chaque paramètre $\theta_i$, soit $\text{I.C.}(\theta_i) \equiv A_i = \{L_i \leq \theta_i \leq U_i\}$; le **taux d'erreur pour l'estimation de** $\theta_i$ est $P(\overline{A_i}) = P(\theta_i \notin A_i)$.

Le **niveau de confiance de la famille** est

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

**Théorème:** pour des taux d'erreur individuels $P(\overline{A_i}) = \frac{\alpha}{g}$, nous avons

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

**Démonstration:** rappelons que $P(C \cup D) = P(C) + P(D) - P(C \cap D)$. Toutes les probabilités sont non-négatives : $P(C) + P(D) \geq P(C \cup D)$.

Ceci s'étend aux unions arbitraires de $g$ événements :

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

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

Mais $P(A_1 \cap \cdots \cap A_g) = 1 - P(\overline{A_1} \cup \cdots \cup \overline{A_g})$, ce qui complète la preuve. ■

Nous utilisons la **procédure de Bonferroni** pour fournir des I.C. **simultanés** des paramètres $\theta_1, \ldots, \theta_g$ à un niveau de confiance "familial" de $1 - \alpha$ :

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

# Estimation conjointe de $\beta_0$ et $\beta_1$

À un niveau de confiance familial de $1-\alpha$, les I.C. **simultanés** de $\beta_0, \beta_1$ ($g = 2$, selon **Bonferroni**) prennent la forme :

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

En moyenne, $100(1-\alpha)\%$ des fois où nous utilisons cette procédure, $\beta_0, \beta_1$ tomberont **tous deux** à l'intérieur de leurs I.C. respectifs.

**Exemple:** pour un niveau de confiance familial de $1-\alpha = 0.95$ (carburants), nous devons utiliser $t(1-\frac{0.05}{4}; 20-2) = t(0.9875; 18) = 2.44501$ :

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

# Procédure de Working-Hotelling

Lorsque nous cherchons un $\mathrm{I.C.}$ pour la réponse moyenne en $X = X^*$, nous exprimons les limites inférieure et supérieure de $\mathrm{I.C.}$ en fonction de $X^*$.

Il serait tentant de considérer l'union de tous ces $\mathrm{I.C.}$ comme une **bande de confiance** pour la réponse moyenne à tous les $X$, c'est-à-dire pour la **droite de meilleur ajustement réelle** $\mathrm{E}\{Y\} = \beta_0 + \beta_1 X$.

Si nous nous intéressons à l'estimation conjointe de la réponse moyenne pour un "petit" nombre de niveaux $X = X_i^*$, $i = 1, \ldots, g$, avec un niveau de confiance familial $1 - \alpha$, nous pouvons utiliser la procédure de **Bonferroni** :

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

Si nous cherchons à construire une région de confiance de la droite $\mathrm{E}\{Y\} = \beta_0 + \beta_1 X$ à environ $100(1-\alpha)\%$, l'approche de Bonferroni nous obligerait à utiliser $g \to \infty$ dans les calculs de I.C., ce qui est problématique car $t(1 - \frac{\alpha/g}{2}; n-2) \to \infty$ dans ce cas.

Nous cherchons au lieu $W > 0$ tel que

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

pour tout $X$ dans le domaine de régression. Cela se produit lorsque

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

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

Afin de trouver le $W$ approprié, nous devons connaître la distribution de

$$\mathcal{M} = \max_X \left\{\frac{(\hat{Y}(X)-\mathrm{E}\{\hat{Y}(X)\})^2}{\mathrm{s}^2\{\hat{Y}(X)\}}\right\} = \max_X \left\{\frac{\left[(b_0+b_1X)-(\beta_0+\beta_1X)\right]^2}{\mathrm{MSE}\left[\frac{1}{n}+\frac{(X-\overline{X})^2}{S_{xx}}\right]}\right\}.$$

Posons $t = X - \overline{X}$ ; la quantité recherchée est alors

$$\max_t \left\{\frac{\left[\overline{Y}-\mathrm{E}\left\{\overline{Y}\right\}+(b_1-\beta_1)t\right]^2}{\mathrm{MSE}\left[\frac{1}{n}+\frac{t^2}{S_{xx}}\right]}\right\} = \max_t \left\{\frac{[c_1+d_1t]^2}{c_2+d_2t^2}\right\} = \max_t\{h(t)\}.$$

Mais $c_2, d_2 > 0$ puisque $\mathrm{MSE}, S_{xx} > 0$, d'où $h(t) \geq 0$ pour tout $t$. Il s'agit d'une fonction rationnelle continue d'une seule variable, avec une asymptote horizontale en $h = d_1^2/d_2 \geq 0$ ; sa dérivée première est

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

Les points critiques sont retrouvés à $t_1 = -\frac{c_1}{d_1}$ et $t_2 = \frac{c_2 d_1}{c_1 d_2}$. Puisque

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

on doit avoir

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

Ainsi,

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

Les deux variables aléatoires du numérateur de $\mathcal{M}$ sont indépendantes et

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

Nous pouvons réécrire la v.a. au dénominateur de $\mathcal{M}$ sous la forme

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

de sorte que

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

Nous avons ainsi

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

# Estimation conjointe de réponses moyennes

À un niveau de confiance **conjoint** de $1-\alpha$, les I.C. de $\mathrm{E}\{Y_i^*\}$ pour un nombre quelconque de niveaux $X = X_i^*$ (selon **Working-Hotelling**) prennent la forme :

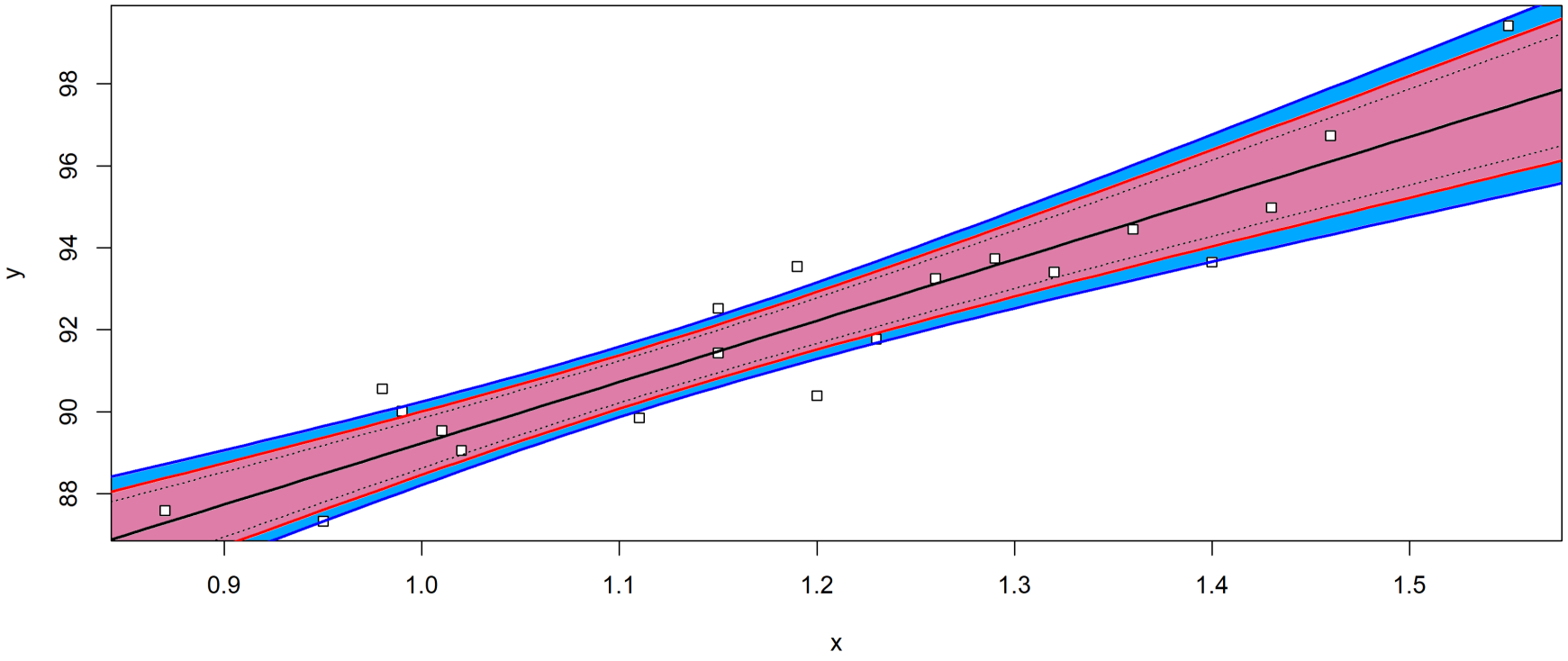
$$\text{I.C.}_{\text{WH}}(\mathrm{E}\{Y_i^*\}; 1-\alpha) = \hat{Y}_i^* \pm \sqrt{2F(1-\alpha; 2, n-2)} \cdot \mathrm{s}\{\hat{Y}_i^*\}.$$

Nous choisissons des approches de Bonferroni ou de Working-Hotelling celle pour laquelle on obtient les I.C. les plus **petits**.

Dans l'exemple des carburants, à un niveau de confiance familial de $1-\alpha = 0.95$, le facteur requis est

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

La bande de confiance de Working-Hotelling pour la droite d'ajustement de l'exemple des carburants est illustrée en **rose** ; la région de Bonferroni pour 20 inférences simultanées sur la réponse moyenne contient de plus la région en **bleue**.



# Procédure de Scheffé et estimation conjointe de nouvelles réponses

À un niveau de confiance **conjoint** de $1-\alpha$ pour $g$ réponses, on obtient des **intervalles de prédiction** de $Y^*_{p_i}$ lorsque $X = X^*_i$, $i = 1, \ldots, g$, en utilisant l'approche qui conduit aux I.P. les plus "serrés" :

- si $g$ est "petit", les I.P. selon **Bonferroni** sont

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

- si $g$ est "large", les I.P. selon **Scheffé** sont

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

# 2.4 – Signification de la régression

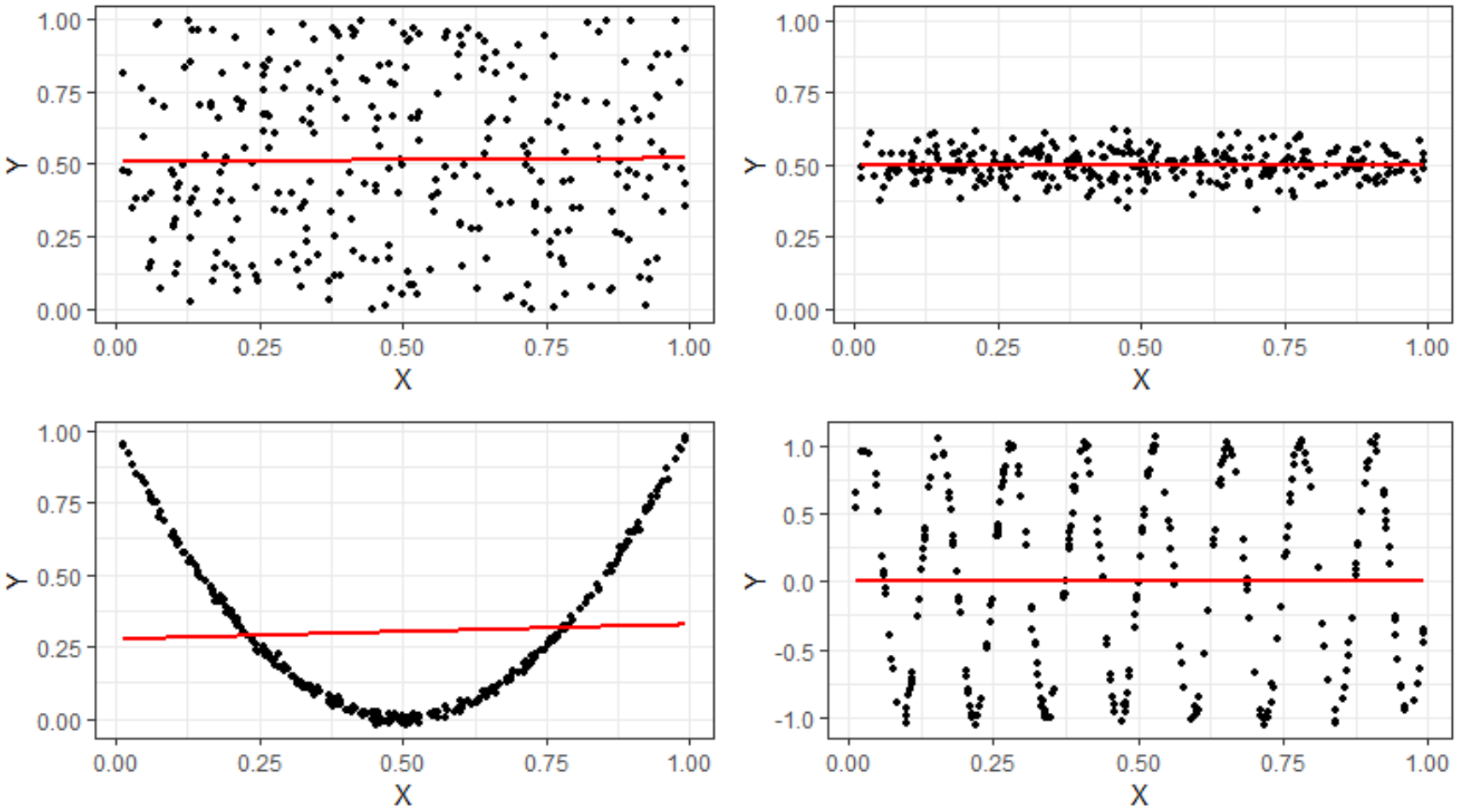
Que pouvons-nous conclure si $\beta_1 = 0$ ? Il se pourrait qu'il :

1. n'y a **aucune relation** entre $X$ et $Y$, comme dans un nuage diffus de points – ce que l'on connait au sujet de $X$ n'explique rien sur les valeurs possibles de $Y$ ;

2. il existe une **relation horizontale** entre $X$ et $Y$, de sorte que les changements dans $X$ n'entraînent aucun changement dans $Y$ ;

3. il existe une **relation non linéaire** entre $X$ et $Y$ qui est au mieux approchée par une droite horizontale.

Dans chacun de ces cas, nous disons que la régression est **non significative**.


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

Le test de **la régression significative** est

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

Les hypothèses sous-jacentes sont que :

1. le **modèle de régression linéaire simple** est valide, et

2. les termes d'erreur sont **indépendants** et suivent une **loi normale** de variance $\sigma^2$.

Avec ces hypothèses, nous pouvons montrer que $b_0, b_1$ sont **indépendants de** $\mathrm{SSE}$ et que

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

# Analyse de la variance

Que $H_0$ soit valide ou non, l'estimateur sans biais de la variance de l'erreur est

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

Rappelons qu'en général :

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

Si $H_0$ : $\beta_1 = 0$ est valide, alors $Y_1, \ldots, Y_n$ est un échantillon aléatoire indépendant prélevé de $\mathcal{N}(\beta_0, \sigma^2)$. La meilleure estimation de $\sigma^2$ est ainsi

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

Le **théorème de Cochran** implique que $\mathrm{SSE}, \mathrm{SSR}$ sont **indépendants**, et que

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

Ainsi, si $H_0 : \beta_1 = 0$ est valide, le quotient

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

suit une loi $F$ de Fisher avec $1, n-2$ degrés de liberté.

On peut montrer que

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

si $\beta_1 \neq 0$, nous avons alors $\mathrm{E}\{\mathrm{MSR}\} > \sigma^2$, ce qui signifie que les grandes valeurs observées de $F^*$ **soutiennent** $H_1 : \beta_1 \neq 0$.

**Règle de décision :** soit $0 < \alpha \ll 1$. Si $F^* > F(1-\alpha; 1, n-2)$, on **rejette $H_0$ en faveur de $H_1$** à un niveau de confiance $\alpha$.

Nous pouvons déterminer $F(1-\alpha; 1, n-2)$, la valeur critique de $F(1, n-2)$, en consultant des tables de valeurs de $F$, ou en utilisant `R`.

Nous avons déjà examiné un test de signification de régression à la section 2.2.3. Ils sont liés : lorsque $\beta_1 = 0$, $F^* = (t^*)^2$.

Dans l'exemple des carburants, nous avons $n = 20$ et

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

de sorte que

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

lorsque $\alpha = 0.05$, la valeur critique est $F(1 - 0.05; 1, 18) = 4.413873$.

Puisque $F^* > F(0.95; 1, 18)$, on **rejete** $H_0 : \beta_1 = 0$ en faveur de l'alternative d'une régression **significative** $(H_1 : \beta_1 \neq 0)$.

# Règle d'or

En gén'eral, si $\mathrm{SSX}$ est une somme de carrésavec $n - x$ degrés de liberté, la **somme des carrés moyenne** correspondante est

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

Sous certaines hypothèses de test spécifiques (ou sous des hypothèses générales, selon la somme des carrés en question ou la situation), $\mathrm{MSX}$ fournit un estimateur sans biais de la variance $\sigma^2$ des termes d'erreur.

Selon la situation, le théorème de Cochran peut alors être utilisé pour démontrer que

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