

MAT 3775
Analyse de la régression

Chapitre 3
Régression linéaire multiple

P. Boily (uOttawa)

Session d'hiver – 2023

P. Boily (uOttawa)

Aperçu

3.1 – Estimation par les moindres carrés (p.8)

- Formulation matricielle (p.10)
- Équations normales (p.11)
- Sommes de carrés et résidus (p.13)

3.2 – Inférence, estimation, et prédiction (p.21)

- Inférence sur les paramètres du modèle (p.29)
- Inférence sur la réponse moyenne (p.36)
- Intervalles de prédiction (p.45)
- Estimations et prédictions simultanées (p.49)

3.3 – Puissance d'un test (p.53)

Aperçu (suite)

3.4 – Coefficients de détermination (p.58)

3.5 – Diagnostiques et mesures correctives (p.60)

- Linéarité (p.62)
- Variance constante (p.68)
- Indépendance (p.72)
- Normalité (p.73)
- Mesures correctives (p.78)

3 – Régression linéaire multiple

En pratique, la situation est généralement plus compliquée ; il pourrait y avoir p **prédicteurs** X_k , $k = 0, \dots, p - 1$.

Exemples:

- X_1 : âge, X_2 : sexe ; Y : taille ($p = 3$).
- X_1 : âge ; X_2 : années d'études ; Y : salaire ($p = 3$).
- X_1 : revenu ; X_2 : mortalité infantile ; X_3 : taux de fécondité, Y : espérance de vie ($p = 4$)
- etc.

En théorie, il peut y avoir une **relation fonctionnelle** $Y = f(X_0, \dots, X_{p-1})$ entre $X_0 (= 1)$, X_1, \dots, X_{p-1} et Y .

En pratique (en supposant qu'une telle relation existe), le mieux que l'on puisse espérer est une **relation statistique**

$$Y = f(X_0, X_1, \dots, X_{p-1}) + \varepsilon,$$

où, comme précédemment, $f(X_0, X_1, \dots, X_{p-1})$ est la **fonction de réponse**, et ε est la **erreur aléatoire** (ou bruit).

Dans le modèle de **régression linéaire générale**, nous supposons que la fonction réponse est

$$f(X_0, X_1, \dots, X_p) = \beta_0 X_0 (= 1) + \beta_1 X_1 + \dots + \beta_{p-1} X_{p-1}.$$

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

$$\mathbf{X}_i = (X_{i,0}(= 1), X_{i,1}, \dots, X_{i,p-1}, Y_i), \quad i = 1, \dots, n.$$

Dans un cadre idéal, ces observations sont **prélevées (conjointement) de façon aléatoire**, selon un plan approprié (c'est le sujet d'un autre cours).

Le **modèle de régression linéaire générale** (MRLG) est

$$Y_i = \beta_0 X_{i,0}(= 1) + \beta_1 X_{i,1} + \dots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \quad i = 1, \dots, n,$$

où les β_k , $k = 0, \dots, p-1$, sont des **paramètres inconnus** et ε_i est l'**erreur aléatoire pour la i ème observation** (ou cas).

Notez qu'un prédicteur X_k peut également être une fonction d'autres variables.

Par exemple, le modèle suivant est un modèle de RLG :

$$E\{Y\} = \beta_0 + \beta_1 X + \beta_2 X^2.$$

Un tel modèle ne doit pas nécessairement être linéaire en X , tant que $E\{Y\}$ est **linéaire par rapport aux paramètres** β_k , $k = 0, \dots, p-1$.

Dans ce qui suit, nous utiliserons la notation suivante :

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, \quad \text{et} \quad \mathbf{X} = \begin{pmatrix} 1 & X_{1,1} & \cdots & X_{1,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p-1} \end{pmatrix},$$

pour représenter le **vecteur réponse**, le **vecteur de paramètres**, et la **matrice de conception**, respectivement.

Dans la matrice de conception \mathbf{X} , \mathbf{X}_i représente la i ème observation (la i ème ligne de \mathbf{X}), un unique **niveau de prédiction multiple**.

Les colonnes de la matrice de conception représentent les valeurs prises par les différentes variables prédictes pour toutes les observations.

Le **modèle de régression linéaire multiple** s'écrit comme

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{où } \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

Le modèle de RLS s'inscrit dans ce cadre, si nous utilisons $p = 2$ avec

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \text{et} \quad \mathbf{X} = \begin{pmatrix} 1 & X_{1,1} \\ \vdots & \vdots \\ 1 & X_{n,1} \end{pmatrix}.$$

3.1 – Estimation par les moindres carrés

Nous traitons les valeurs des prédicteurs $X_{i,k}$ comme s'ils étaient constantes, pour $i = 1, \dots, n$, $k = 0, \dots, p - 1$ (il n'y a **pas d'erreur de mesure**).

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

$$E\{Y_i \mid \mathbf{X}_i\} = E\{\mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i \mid \mathbf{X}_i\} = \mathbf{X}_i\boldsymbol{\beta} + E\{\varepsilon_i\} = \mathbf{X}_i\boldsymbol{\beta}.$$

La **déviante à X_i** est la différence entre la réponse observée Y_i et la réponse attendue

$$e_i = Y_i - E\{Y_i \mid \mathbf{X}_i\};$$

l'écart peut être **positif** (si le point se situe **au-dessus** de l'hyperplan $Y = \mathbf{X}\boldsymbol{\beta}$) ou **négatif** (s'il se situe **en dessous**).

Comment trouver les **estimateurs** de β ? Au fait, comment déterminer si l'hyperplan ajusté est un **bon modèle pour les données** ?

Considérons la fonction

$$Q(\beta) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - E\{Y_i \mid \mathbf{X}_i\})^2 = \sum_{i=1}^n (Y_i - \mathbf{X}_i\beta)^2.$$

Si $Q(\beta)$ est “petit”, la somme des **carrés des résidus** est “petite”, et nous nous attendons à ce que $Y = \mathbf{X}\beta$ soit bien ajusté aux données.

Les **estimateurs des moindres carrés** du problème RLG sont donnés par le vecteur $\mathbf{b} \in \mathbb{R}^p$ qui minimise la fonction Q par rapport à $\beta \in \mathbb{R}^p$.

Nous cherchons les points critiques de $Q(\beta)$ qui résolvent **$\nabla_{\beta} Q(\mathbf{b}) = \mathbf{0}$** .

3.1.1 – Formulation matricielle

La fonction de régression des moindres carrés est $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$, où \mathbf{b} minimise

$$\begin{aligned} Q(\boldsymbol{\beta}) &= \sum_{i=1}^n (Y_i - \mathbf{X}_i\boldsymbol{\beta})^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{Y}^\top - \boldsymbol{\beta}^\top \mathbf{X}^\top)(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{Y}^\top \mathbf{Y} - \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{Y} - \mathbf{Y}^\top \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}. \end{aligned}$$

Puisque $\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{Y}$ est un scalaire, il est égal à sa transposée $\mathbf{Y}^\top \mathbf{X}\boldsymbol{\beta}$, d'où

$$Q(\boldsymbol{\beta}) = \mathbf{Y}^\top \mathbf{Y} - 2\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{Y} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta}.$$

Mais $\mathbf{X}^\top \mathbf{X}$ est défini positif, donc $Q(\boldsymbol{\beta})$ est minimisé lorsque $\nabla_{\boldsymbol{\beta}} Q(\mathbf{b}) = \mathbf{0}$.

3.1.2 – Équations normales

Le gradient de $Q(\beta)$ est

$$\nabla_{\beta} Q(\beta) = -2\mathbf{X}^{\top} \mathbf{Y} + 2\mathbf{X}^{\top} \mathbf{X} \beta;$$

ainsi le point critique \mathbf{b} satisfait aux **équations normales**

$$(\mathbf{X}^{\top} \mathbf{X}) \mathbf{b} = \mathbf{X}^{\top} \mathbf{Y}.$$

$\mathbf{X}^{\top} \mathbf{X}$ est la matrice des **sommes de carrés** et des **produits croisés** ; lorsqu'elle est inversible, la solution **unique** des équations normales est

$$\mathbf{b} = (\mathbf{X}^{\top} \mathbf{X})^{-1} \mathbf{X}^{\top} \mathbf{Y},$$

l'**estimateurs des moindres carrés** du problème de RLG.

C'est une matrice de taille $p \times p$; elle n'est donc généralement pas trop "dispendieuse" à inverser, quel que soit le nombre d'observations n .

Par exemple, si nous avons deux prédicteurs X_1, X_2 et trois paramètres de régression $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top$, écrivons $\mathbf{x} = (1, X_1, X_2)$.

La **fonction de régression** est

$$E\{Y\} = \mathbf{x}\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

Si l'estimateur des moindres carrés est $\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} = (0.5, -0.1, 2)^\top$, disons, alors la **fonction de régression estimée** est

$$\hat{Y} = \mathbf{x}\mathbf{b} = 0.5 - 0.1X_1 + 2X_2.$$

3.1.3 – Sommes de carrés et résidus

Les **valeurs ajustées** pour le problème de RLG sont

$$\begin{pmatrix} \hat{Y}_1 \\ \vdots \\ \hat{Y}_n \end{pmatrix} = \hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \underbrace{\mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top}_{=\mathbf{H}} \mathbf{Y} = \mathbf{H}\mathbf{Y},$$

où \mathbf{H} est la **matrice “chapeau”**.

Théorème : \mathbf{H} , $\mathbf{I}_n - \mathbf{H}$ sont idempotentes et symétriques ; $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{0}$.

Démonstration : utilisons la notation $\mathbf{M} = \mathbf{I}_n - \mathbf{H}$. On commence par montrer $\mathbf{H}^2 = \mathbf{H}$, $\mathbf{H}^\top = \mathbf{H}$, $\mathbf{M}^2 = \mathbf{M}$, et $\mathbf{M}^\top = \mathbf{M}$.

C'est évidemment le cas :

$$\mathbf{H}^2 = \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{X} \mathbf{I}_n (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{H}$$

$$\begin{aligned} \mathbf{H}^\top &= \left(\mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \right)^\top = (\mathbf{X}^\top)^\top \left((\mathbf{X}^\top \mathbf{X})^{-1} \right)^\top \mathbf{X}^\top = \mathbf{X} \left((\mathbf{X}^\top \mathbf{X})^\top \right)^{-1} \mathbf{X}^\top \\ &= \mathbf{X}^\top (\mathbf{X}^\top (\mathbf{X}^\top)^\top)^{-1} \mathbf{X}^\top = \mathbf{X}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top = \mathbf{H} \end{aligned}$$

$$\mathbf{M}^2 = (\mathbf{I}_n - \mathbf{H})^2 = \mathbf{I}_n^2 - \mathbf{I}_n \mathbf{H} - \mathbf{H} \mathbf{I}_n + \mathbf{H}^2 = \mathbf{I}_n - 2\mathbf{H} + \mathbf{H} = \mathbf{I}_n - \mathbf{H} = \mathbf{M}$$

$$\mathbf{M}^\top = (\mathbf{I}_n - \mathbf{H})^\top = \mathbf{I}_n^\top - \mathbf{H}^\top = \mathbf{I}_n - \mathbf{H} = \mathbf{M}.$$

De plus,

$$\mathbf{M}\mathbf{X} = (\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{X} - \mathbf{H}\mathbf{X} = \mathbf{X} - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} = \mathbf{X} - \mathbf{X} \mathbf{I}_n = \mathbf{0},$$

ce qui complète la démonstration. ■

Le *i*^e résidu est $e_i = Y_i - \hat{Y}_i$. Puisque $\mathbf{MX} = \mathbf{0}$, le **vecteur de résidus** est

$$\begin{aligned}\mathbf{e} &= \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y} = \mathbf{MY} \\ &= \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{M}\boldsymbol{\varepsilon}.\end{aligned}$$

Autrement dit, le vecteur des résidus est à la fois une transformation linéaire du vecteur de réponse \mathbf{Y} et du vecteur d'erreur aléatoire $\boldsymbol{\varepsilon}$.

Tout comme dans le cas de la RLS (qui est un cas particulier de la RLK), les résidus possèdent un ensemble de propriétés intéressantes.

Théorème : la matrice de conception est orthogonale au vecteur des résidus, c'est-à-dire que $\mathbf{X}^\top \mathbf{e} = \mathbf{0}$ (les colonnes de \mathbf{X} sont orthogonales à \mathbf{e}).

Démonstration : à partir des équations normales, nous obtenons

$$\mathbf{X}^\top \mathbf{X} \mathbf{b} = \mathbf{X}^\top \mathbf{Y} \implies \mathbf{X}^\top (\mathbf{Y} - \mathbf{X} \mathbf{b}) = \mathbf{0} \implies \mathbf{X}^\top (\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{0}.$$

Mais $\mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{e}$, de sorte que $\mathbf{X}^\top \mathbf{e} = \mathbf{0}$. ■

Théorème : si le modèle possède un term constant $\beta_0 \neq 0$, nous avons également $\mathbf{1}_n^\top \mathbf{e} = 0$, $\bar{\mathbf{e}} = \overline{\mathbf{Y}} - \overline{\hat{\mathbf{Y}}} = 0$, et $\hat{\mathbf{Y}}^\top \mathbf{e} = 0$.

Démonstration : si le modèle possède un terme constant, la première colonne de la matrice de conception \mathbf{X} est $\mathbf{1}_n$. Ainsi, $\mathbf{1}_n^\top \mathbf{e}$ correspond à la première entrée de $\mathbf{X}^\top \mathbf{e} = \mathbf{0}$, c'est-à-dire 0.

Ceci implique également que $\bar{\mathbf{e}} = 0$. Pour la dernière partie, rappelons que $\hat{\mathbf{Y}} = \mathbf{X} \mathbf{b}$. Ainsi, $\hat{\mathbf{Y}}^\top = \mathbf{b}^\top \mathbf{X}^\top$ et $\hat{\mathbf{Y}}^\top \mathbf{e} = \mathbf{b}^\top \mathbf{X}^\top \mathbf{e} = \mathbf{b}^\top \mathbf{0} = 0$. ■

Nous savons que SST est une forme quadratique en \mathbf{Y} :

$$\text{SST} = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \mathbf{Y}^\top \left(\mathbf{I}_n - \frac{1}{n} \mathbf{J}_n \right) \mathbf{Y};$$

selon la définition des résidus, c'est également le cas pour SSE :

$$\begin{aligned} \text{SSE} &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n e_i^2 = \mathbf{e}^\top \mathbf{e} = (\mathbf{M}\mathbf{Y})^\top \mathbf{M}\mathbf{Y} = \mathbf{Y}^\top \mathbf{M}^\top \mathbf{M}\mathbf{Y} \\ &= \mathbf{Y}^\top \mathbf{M}^2 \mathbf{Y} = \mathbf{Y}^\top \mathbf{M}\mathbf{Y} = \mathbf{Y}^\top (\mathbf{I}_n - \mathbf{H}) \mathbf{Y}. \end{aligned}$$

La décomposition en sommes de carrés peut alors se ré-écrire sous la forme :

$$\text{SSR} = \text{SST} - \text{SSE}.$$

Ainsi, SSR est également une forme quadratique en \mathbf{Y} :

$$\begin{aligned} \text{SSR} &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = \mathbf{Y}^\top \left(\mathbf{I}_n - \frac{1}{n} \mathbf{J}_n \right) \mathbf{Y} - \mathbf{Y}^\top (\mathbf{I}_n - \mathbf{H}) \mathbf{Y} \\ &= \mathbf{Y}^\top \left(\mathbf{I}_n - \frac{1}{n} \mathbf{J}_n - \mathbf{I}_n + \mathbf{H} \right) \mathbf{Y} = \mathbf{Y}^\top \left(\mathbf{H} - \frac{1}{n} \mathbf{J}_n \right) \mathbf{Y}. \end{aligned}$$

Théorème : $E\{\text{SSE}\} = (n-p)\sigma^2$, d'où $\text{rang}(\mathbf{M}) = \text{tr}(\mathbf{M}) = n-p$. Ainsi, SSE est une somme de carrés avec $n-p$ degrés de liberté.

Démonstration : nous avons

$$\text{SSE} = \mathbf{e}^\top \mathbf{e} = (\mathbf{M}\boldsymbol{\varepsilon})^\top \mathbf{M}\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^\top \mathbf{M}\boldsymbol{\varepsilon} = \sum_{i,j=1}^n m_{ij} \varepsilon_i \varepsilon_j = \sum_{i=1}^n m_{ii} \varepsilon_i^2 + \sum_{i \neq j} m_{ij} \varepsilon_i \varepsilon_j.$$

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

$$\begin{aligned} \mathbb{E} \{ \varepsilon_i^2 \} &= \sigma^2 \{ \varepsilon_i \} + (\mathbb{E} \{ \varepsilon_i \})^2 = \sigma^2 + 0 = \sigma^2, \quad i = 1, \dots, n, \quad \text{et} \\ \mathbb{E} \{ \varepsilon_i \varepsilon_j \} &= \sigma \{ \varepsilon_i, \varepsilon_j \} - \mathbb{E} \{ \varepsilon_i \} \mathbb{E} \{ \varepsilon_j \} = 0 - 0 = 0, \quad i \neq j. \end{aligned}$$

Par conséquent,

$$\begin{aligned} \mathbb{E} \{ \text{SSE} \} &= \mathbb{E} \left\{ \sum_{i=1}^n m_{ii} \varepsilon_i^2 + \sum_{i \neq j} m_{ij} \varepsilon_i \varepsilon_j \right\} = \mathbb{E} \left\{ \sum_{i=1}^n m_{ii} \varepsilon_i^2 \right\} + \mathbb{E} \left\{ \sum_{i \neq j} m_{ij} \varepsilon_i \varepsilon_j \right\} \\ &= \sum_{i=1}^n m_{ii} \mathbb{E} \{ \varepsilon_i^2 \} + \sum_{i \neq j} m_{ij} \mathbb{E} \{ \varepsilon_i \varepsilon_j \} = \sigma^2 \sum_{i=1}^n m_{ii} = \sigma^2 \text{tr}(\mathbf{M}) \\ &= \sigma^2 \text{tr}(\mathbf{I}_n - \mathbf{H}) = \sigma^2 [\text{tr}(\mathbf{I}_n) - \text{tr}(\mathbf{H})] = \sigma^2 [n - \text{tr}(\mathbf{H})]. \end{aligned}$$

Mais

$$\text{tr}(\mathbf{H}) = \text{tr} \left(\underbrace{\mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}}_{=A_{n \times p}} \underbrace{\mathbf{X}^\top}_{=B_{p \times n}} \right) = \text{tr} \left(\underbrace{\mathbf{X}^\top}_{=B_{p \times n}} \underbrace{\mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}}_{=A_{n \times p}} \right) = \text{tr}(\mathbf{I}_p) = p,$$

$$\text{d'où } E\{\text{SSE}\} = (n - p)\sigma^2. \quad \blacksquare$$

L'erreur quadratique moyenne MSE du modèle de RLG est

$$\text{MSE} = \frac{\text{SSE}}{n - p};$$

ce qui n'est pas surprenant puisque nous avons à estimer les p paramètres β_k , $k = 0, \dots, p - 1$, afin de calculer SSE. Selon le théorème précédent, MSE est un **estimateur non biaisé de la variance de l'erreur** σ^2 .

3.2 – Inférence, estimation, et prédiction

En supposant la **normalité** et l'**indépendance** des erreurs aléatoires, les estimateurs b_0, \dots, b_{p-1} sont indépendants de SSE et

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

Cela nous permet de tester si la régression est **significative**, à l'aide du **test F global** :

$$H_0 : \beta_1 = \dots = \beta_{p-1} = 0 \quad \text{vs.} \quad H_1 : \beta_k \neq 0 \text{ pour un } k = 1, \dots, p - 1,$$

en supposant que le modèle de RLG soit valide.

Analyse de la variance

En particulier, nous avons

$$Y_i \sim \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2\mathbf{I}_n), \quad i = 1, \dots, n.$$

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

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

Si H_0 est valide, alors Y_1, \dots, Y_n est un échantillon aléatoire indépendant prélevé à même $\mathcal{N}(\beta_0, \sigma^2)$, et le meilleur estimateur de σ^2 est

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

Puisque $SST = SSE + SSR$, le **théorème de Cochran** implique que SSE, SSR sont **indépendants**, et que

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

Ainsi, si H_0 est valide, le quotient

$$F^* = \frac{\left(\frac{SSR}{\sigma^2}\right) / (p-1)}{\left(\frac{SSE}{\sigma^2}\right) / (n-p)} = \frac{SSR / (p-1)}{SSE / (n-p)} = \frac{MSR}{MSE} \sim F(p-1, n-p)$$

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

Le tableau d'**ANOVA** correspondant est

Source	SS	df	MS	F*
Regression	SSR	$p - 1$	$MSR = SSR / (p - 1)$	MSR / MSE
Error	SSE	$n - p$	$MSE = SSE / (n - p)$	
Total	SST	$n - 1$		

La valeur- p du test F global est

$$P(F(p - 1, n - p) > F^*).$$

Règle de décision : on rejette H_0 à un niveau de confiance de $1 - \alpha$ si

$$F^* > F(1 - \alpha; p - 1, n - p);$$

de façon équivalente, on rejette H_0 si $P(F(p - 1, n - p) > F^*) < \alpha$.

Exemple : considérons un ensemble de données avec $n = 12$ observations, une variable réponse Y et $p - 1 = 4$ prédicteurs X_1, X_2, X_3, X_4 . Nous construisons un modèle de RLG

$$Y_i = \mathbf{X}_i \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, 12$$

$$= \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \beta_4 X_{i,4} + \varepsilon_i, \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{12})$$

Le tableau d'**ANOVA** correspondant est

Source	SS	df	MS	F*
Regression	4957.2	4	1239.3	5.1
Error	1699.0	7	242.7	
Total	6656.2	11		

Puisque la valeur- p est $P(F(4, 7) > 5.1) = 0.0303$, on **rejete** H_0 lorsque $\alpha = 0.05$; la conclusion est que la régression est **significative**.

Interprétation géométrique

Certains concepts de la RLG deviennent plus faciles à comprendre lorsque vus à travers le prisme de la **géométrie** et de l'**algèbre vectorielle**. Soit

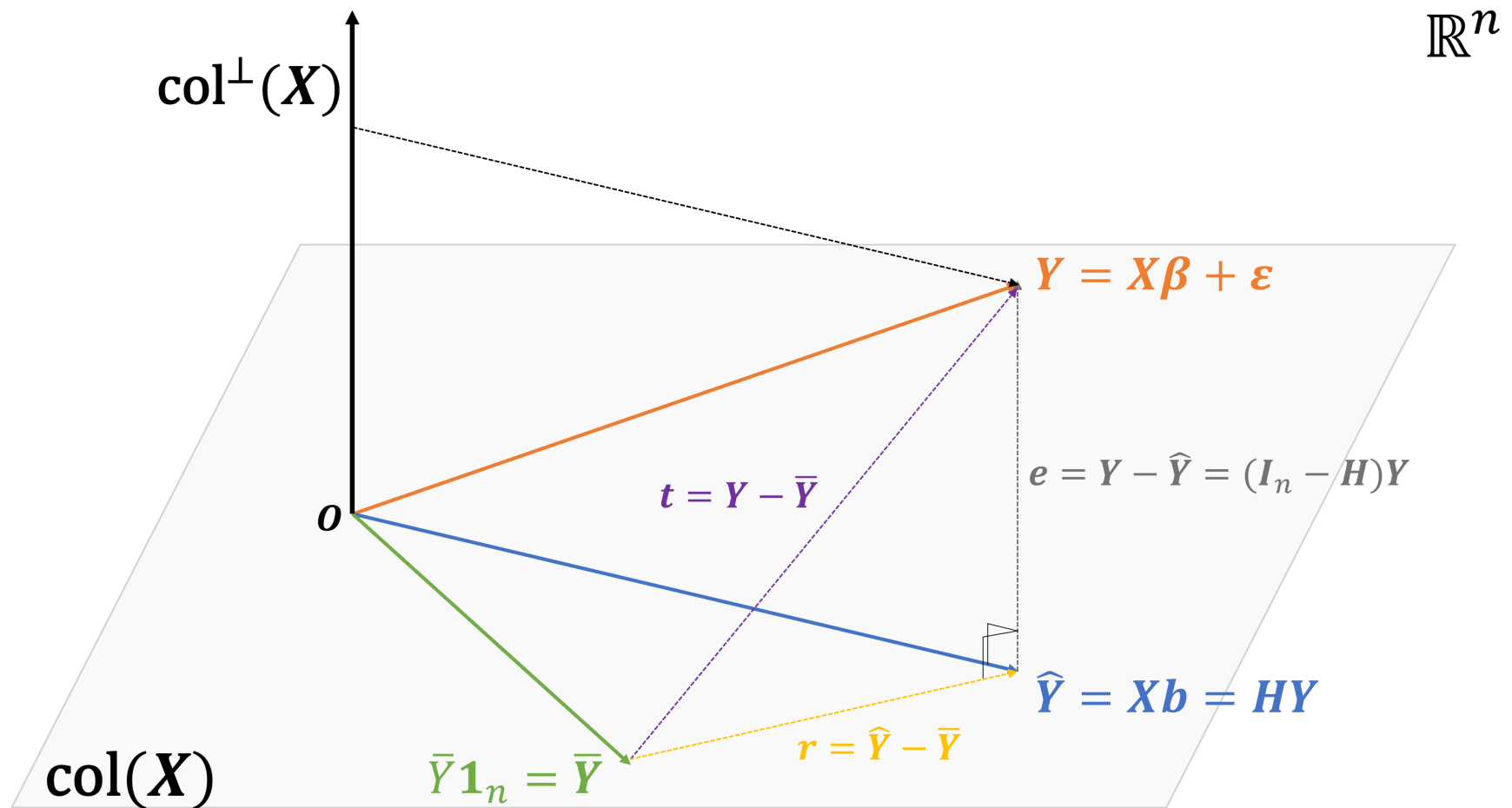
$$\mathcal{M}(\mathbf{X}) = \text{col}(\mathbf{X}) = \{\mathbf{X}\boldsymbol{\gamma} \mid \boldsymbol{\gamma} \in \mathbb{R}^p\} \subset \mathbb{R}^n$$

$$\mathcal{M}^\perp(\mathbf{X}) = (\text{col}(\mathbf{X}))^\perp = \{\mathbf{v} \in \mathbb{R}^n \mid \mathbf{v} \cdot \mathbf{w} = 0, \forall \mathbf{w} \in \mathcal{M}(\mathbf{X})\}$$

Le **vecteur des réponses** $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \in \mathbb{R}^n$, tandis que le **vecteur ajusté** $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{H}\mathbf{Y} \in \mathcal{M}(\mathbf{X})$ et

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y} \in \mathcal{M}^\perp(\mathbf{X}).$$

Les matrices chapeau \mathbf{H} et $\mathbf{I}_n - \mathbf{H}$ sont toutes deux idempotentes (ce sont les matrices de projection sur $\mathcal{M}(\mathbf{X})$ et $\mathcal{M}^\perp(\mathbf{X})$) et symétriques.



L'estimateur \mathbf{b} est tel que $\mathbf{X}\mathbf{b}$ est le vecteur le plus près de \mathbf{Y} dans $\mathcal{M}(\mathbf{X})$:

$$\mathbf{b} = \arg \min_{\boldsymbol{\gamma} \in \mathbb{R}^p} \{ \|\mathbf{Y} - \mathbf{X}\boldsymbol{\gamma}\|_2^2 \} = \arg \min_{\boldsymbol{\gamma} \in \mathbb{R}^p} \{ \|\mathbf{e}\|_2^2 \} = \arg \min_{\boldsymbol{\gamma} \in \mathbb{R}^p} \{ \text{SSE} \}.$$

Si le modèle de RLG a un terme constant β_0 , le vecteur de la moyenne $\bar{\mathbf{Y}} = \bar{Y}\mathbf{1}_n$ se retrouve dans $\mathcal{M}(\mathbf{X})$; de fait, nous avons $\bar{\mathbf{Y}} = \mathbf{X}\boldsymbol{\gamma}^*$ pour $\boldsymbol{\gamma}^* = (\bar{Y}, 0, \dots, 0)^\top$.

Le triangle $\Delta \mathbf{Y} \hat{\mathbf{Y}} \bar{\mathbf{Y}}$ est donc un **triangle rectangle**, où

$$\mathbf{t} = \mathbf{Y} - \bar{\mathbf{Y}} = (\mathbf{Y} - \hat{\mathbf{Y}}) + (\hat{\mathbf{Y}} - \bar{\mathbf{Y}}) = \mathbf{e} + \mathbf{r};$$

selon le théorème de Pythagore, nous récupérons la décomposition en SS :

$$\|\mathbf{t}\|_2^2 = \text{SST} = \text{SSE} + \text{SSR} = \|\mathbf{e}\|_2^2 + \|\mathbf{r}\|_2^2.$$

3.2.1 – Inférence sur les paramètres du modèle

Comme c'était le cas avec le modèle de RLS, si $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, alors

$$\mathbf{Y} \sim \mathcal{N}(E\{\mathbf{Y}\}, \sigma^2 \{\mathbf{Y}\}) = \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n).$$

Pour toute matrice A compatible, nous obtenons ainsi

$$A\mathbf{Y} \sim \mathcal{N}(AE\{\mathbf{Y}\}, A\sigma^2 \{\mathbf{Y}\} A^\top) = \mathcal{N}(A\mathbf{X}\boldsymbol{\beta}, \sigma^2 AA^\top).$$

Selon les équations normales, les estimations du modèle de RLG sont obtenues à l'aide d'une **transformation linéaire** du vecteur réponse \mathbf{Y} :

$$\mathbf{b} = \underbrace{(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top}_{p \times n} \mathbf{Y} = A\mathbf{Y}.$$

Entre autres,

$$E \{ \mathbf{b} \} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top E \{ \mathbf{Y} \} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} = \boldsymbol{\beta},$$

de sorte que \mathbf{b} fournit des **estimateurs sans biais** de $\boldsymbol{\beta}$.

De plus,

$$\begin{aligned} \sigma^2 \{ \mathbf{b} \} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \sigma^2 \{ \mathbf{Y} \} [(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top]^\top \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \sigma^2 \mathbf{I}_n [(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top]^\top = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}. \end{aligned}$$

Ainsi,

$$\mathbf{b} \sim \mathcal{N} (\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}).$$

La **matrice de variance-covariance estimée** pour les estimateurs \mathbf{b} est donc

$$s^2 \{\mathbf{b}\} = \text{MSE} \cdot (\mathbf{X}^\top \mathbf{X})^{-1}, \quad \text{et} \quad s \{\mathbf{b}\} = \sqrt{\text{MSE}} \sqrt{\text{diag} [(\mathbf{X}^\top \mathbf{X})^{-1}] }.$$

Pour tout $k = 0, \dots, p - 1$, la **studentisation** (normalisation) de b_k est

$$T_k = \frac{b_k - \beta_k}{\sqrt{\text{MSE}} \sqrt{(\mathbf{X}^\top \mathbf{X})_{k,k}^{-1}}} = \underbrace{\frac{b_k - \beta_k}{\sigma \sqrt{(\mathbf{X}^\top \mathbf{X})_{k,k}^{-1}}}}_{=Z} \bigg/ \sqrt{\underbrace{\frac{\text{SSE}}{\sigma^2}}_{=U} \underbrace{(n-p)}_{=\nu}} \sim t(n-p),$$

où $(\mathbf{X}^\top \mathbf{X})_{k,k}^{-1}$ représente la $k + 1$ ème valeur dans $\text{diag} [(\mathbf{X}^\top \mathbf{X})^{-1}]$.

Pour un index spécifique $k \in \{0, \dots, p-1\}$, l'I.C. de β_k à $100(1 - \alpha)\%$ est

$$\text{I.C.}(\beta_k; 0.95) \equiv b_k \pm t\left(1 - \frac{\alpha}{2}; n - p\right) \cdot s\{b_k\}.$$

Les tests d'hypothèse correspondants sont

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

Si H_0 est valide, la statistique de test calculée est

$$T_k = \frac{b_k - \beta_k^*}{s\{b_k\}} \sim t(n - p).$$

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

Soit t^* la valeur observée de T_k ; nous **rejetons** H_0 si t^* se retrouve dans la région critique.

Hypothèse alternative	Région critique
$H_1 : \beta_k < \beta_k^*$	$t^* < -t(1 - \alpha; n - p)$
$H_1 : \beta_k > \beta_k^*$	$t^* > t(1 - \alpha; n - p)$
$H_1 : \beta_k \neq \beta_k^*$	$ t^* > t(1 - \alpha/2; n - p)$

Exemple : considérons la situation avec $n = 12$ observations et $p - 1 = 4$ prédicteurs décrit précédemment.

Nous construisons le modèle de RLG $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$ et nous obtenons les résultats suivants :

Prédicteur	Estimation	Erreur-type	t^*
Ordonnée	-102.71	207.86	-0.49
X_1	0.61	0.37	1.64
X_2	8.92	5.3	1.68
X_3	1.44	2.39	0.60
X_4	0.01	0.77	0.02

Rappelons que $n - p = 7$; l'I.C. de β_2 à 95%, par exemple, est

$$\text{I.C.}(\beta_2; 0.95) \equiv 8.92 \pm t(0.975; 7) \cdot 5.3 = 8.92 \pm 2.365 \cdot 5.3 = [-3.6, 21.5].$$

Nous pourrions également tester pour $H_0 : \beta_3 = 2$ vs. $H_1 : \beta_3 \neq 2$, mettons : en supposant que H_0 soit valide, nous avons

$$T_3^* = \frac{b_3 - 2}{s\{b_3\}} \sim t(7).$$

La statistique de test observée est

$$t^* = \frac{1.44 - 2}{2.39} = -0.23;$$

nous rejetons H_0 à un niveau de confiance de $1 - \alpha = 0.95$ si

$$|t^*| > t(0.975; 7) = 2.365;$$

puisque $-0.23 \not> 2.365$, l'évidence ne permet pas de conclure que $\beta_3 \neq 2$.

Quoique nous pouvons construire un I.C. pour β_2 et tester une hypothèse sur β_3 séparément, chacun au niveau de confiance de $1 - \alpha = 0,95$, nous ne pouvons pas le faire **simultanément**.

3.2.2 – Inférence sur la réponse moyenne

Nous pouvons également effectuer une analyse inférentielle pour la **réponse moyenne** lorsque

$\mathbf{X}^* = (1, X_1^*, \dots, X_{p-1}^*)$ se retrouve dans la **portée** du modèle.

Dans le modèle de RLG, nous supposons que

$$E\{Y^*\} = \mathbf{X}^* \boldsymbol{\beta} = \beta_0 + \beta_1 X_1^* + \dots + \beta_{p-1} X_{p-1}^*.$$

La **réponse moyenne estimée** en \mathbf{X}^* est

$$\hat{Y}^* = \mathbf{X}^* \mathbf{b} = b_0 + b_1 X_1^* + \dots + b_{p-1} X_{p-1}^*.$$

Les valeurs des prédicteurs sont **fixes** ; \hat{Y}^* suit ainsi une loi normale avec

$$E\{\hat{Y}^*\} = E\{\mathbf{X}^*\mathbf{b}\} = \mathbf{X}^*E\{\mathbf{b}\} = \mathbf{X}^*\boldsymbol{\beta},$$

de sorte que \hat{Y}^* est un **estimateur sans biais** de Y^* .

De plus,

$$\sigma^2\{\hat{Y}^*\} = \mathbf{X}^*\sigma^2\{\mathbf{b}\}(\mathbf{X}^*)^\top = \sigma^2\mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top,$$

d'où

$$s^2\{\hat{Y}^*\} = \text{MSE} \cdot \mathbf{X}^*(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^*)^\top = \mathbf{X}^*s^2\{\mathbf{b}\}(\mathbf{X}^*)^\top.$$

L'**erreur-type estimée** est ainsi

$$s\{\hat{Y}^*\} = \sqrt{\mathbf{X}^* s^2\{\mathbf{b}\} (\mathbf{X}^*)^\top}.$$

Puisque

$$\hat{Y}^* = \mathbf{X}^* \mathbf{b} = \mathbf{X}^* (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

est une **transformation linéaire** de \mathbf{Y} , et puisque

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n),$$

alors

$$\hat{Y}^* \sim \mathcal{N}\left(\mathbb{E}\{\hat{Y}^*\}, \sigma^2\{\hat{Y}^*\}\right) = \mathcal{N}\left(\mathbf{X}^* \boldsymbol{\beta}, \sigma^2 \mathbf{X}^* (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^*)^\top\right).$$

Ainsi,

$$Z = \frac{\hat{Y}^* - E\{\hat{Y}^*\}}{\sigma\{\hat{Y}^*\}} = \frac{\hat{Y}^* - \mathbf{X}^*\boldsymbol{\beta}}{\sigma\sqrt{\mathbf{X}^* (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^*)^\top}} \sim \mathcal{N}(0, 1).$$

La **studentisation** de \hat{Y}^* est donc

$$\begin{aligned} T &= \frac{\hat{Y}^* - \mathbf{X}^*\boldsymbol{\beta}}{\underbrace{\sigma\sqrt{\mathbf{X}^* (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^*)^\top}}_{=Z}} \bigg/ \sqrt{\underbrace{\frac{\text{SSE}}{\sigma^2}}_{=U} \bigg/ \underbrace{(n-p)}_{=\nu}} \\ &= \frac{\hat{Y}^* - \mathbf{X}^*\boldsymbol{\beta}}{\sqrt{\text{MSE}}\sqrt{\mathbf{X}^* (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^*)^\top}} \sim t(n-p). \end{aligned}$$

Pour un niveau de prédiction \mathbf{X}^* , l'I.C. de $E\{Y^*\}$ à $100(1 - \alpha)\%$ est

$$\text{I.C.}(E\{Y^*\}; 0.95) \equiv \hat{Y}^* \pm t\left(1 - \frac{\alpha}{2}; n - p\right) \cdot s\{\hat{Y}^*\}.$$

Les tests d'hypothèse correspondants sont

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

Si H_0 est valide, la statistique de test calculée est

$$T = \frac{\hat{Y}^* - \gamma}{s\{\hat{Y}^*\}} \sim t(n - p).$$

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

Soit t^* la valeur observée de T ; nous **rejetons** H_0 si t^* se retrouve dans la région critique.

Hypothèse alternative	Région critique
$H_1 : E \{Y^*\} < \gamma$	$t^* < -t(1 - \alpha; n - p)$
$H_1 : E \{Y^*\} > \gamma$	$t^* > t(1 - \alpha; n - p)$
$H_1 : E \{Y^*\} \neq \gamma$	$ t^* > t(1 - \alpha/2; n - p)$

Exemple : considérons la situation avec $n = 12$ observations et $p - 1 = 4$ prédicteurs décrit précédemment. Nous cherchons à prédire la réponse moyenne lorsque

$\mathbf{X}^* = (1, 11.10, 20.74, 6.61, 182.38)$, dans la **portée** du modèle.

Nous avons

$$\begin{aligned}\hat{Y}^* &= \mathbf{X}^* \mathbf{b} \\ &= -102.71 + 0.61(11.10) + 8.92(20.74) + 1.44(6.61) + 0.01(182.38) \\ &= 100.40.\end{aligned}$$

Rappelons que $\text{MSE} = 242.71$. Avec ces données, on calcule

$$\mathbf{X}^* (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^*)^\top = 1.42,$$

de sorte que

$$s\{\hat{Y}^*\} = \sqrt{242.71} \sqrt{1.42} = 22.12.$$

Puisque $n - p = 7$; l'I.C. de $E\{Y^*\}$ à environ 95% est

$$\begin{aligned} \text{I.C.}(E\{Y^*\}; 0.95) &\equiv 100.40 \pm t(0.975; 7) \cdot 22.12 \\ &= 100.40 \pm 2.365 \cdot 22.12 = [48.09, 152.71]. \end{aligned}$$

On pourrait aussi tester $H_0 : E\{Y^*\} = 150$ vs. $H_1 : E\{Y^*\} < 150$, mettons : si H_0 est valide,

$$T^* = \frac{\hat{Y}^* - 150}{s\{\hat{Y}^*\}} \sim t(7).$$

La statistique de test observée est

$$t^* = \frac{100.40 - 150}{22.12} = -2.24.$$

Nous **rejetons** H_0 à un niveau de confiance $1 - \alpha = 0.95$ si

$$t^* < -t(0.95; 7) = -1.89;$$

comme $-2.24 < -1.89$, l'évidence **permet de rejeter**

$$H_0 : E\{Y^*\} = 150 \quad \text{en faveur de} \quad H_1 : E\{Y^*\} < 150.$$

Remarquez cependant que l'I.C. de $E\{Y^*\}$ à 95% contient 150, de sorte que nous **ne pouvons pas rejeter**

$$H_0 : E\{Y^*\} = 150 \quad \text{en faveur de} \quad H_1 : E\{Y^*\} \neq 150$$

à un niveau de confiance de $1 - \alpha = 95\%$. Comme au préalable, nous ne pouvons pas effectuer d'**inférences simultanées** sans modifier les régions critiques.

3.2.3 – Intervalles de prédiction

Soit Y_p^* une **nouvelle réponse** lorsque \mathbf{X}^* , de sorte que

$$Y_p^* = \mathbf{X}^* \boldsymbol{\beta} + \varepsilon_p \quad \text{pour un certain } \varepsilon_p.$$

Si l'erreur moyenne est nulle, la meilleure prédiction pour Y_p^* est toujours la **réponse ajustée lorsque \mathbf{X}^*** :

$$\hat{Y}_p^* = \mathbf{X}^* \mathbf{b}.$$

L'**erreur de prédiction** lorsque \mathbf{X}^* est alors

$$\text{pred}^* = Y_p^* - \hat{Y}_p^* = \mathbf{X}^* \boldsymbol{\beta} + \varepsilon_p - \mathbf{X}^* \mathbf{b}.$$

Dans le modèle de RLG, l'erreur ε_p et les estimateurs \mathbf{b} suivent une loi **normale**. Par conséquent, il en est de même pour l'erreur de prédiction pred^* . On remarque que

$$\mathbb{E} \{ \text{pred}^* \} = \underbrace{\mathbb{E} \{ \mathbf{X}^* \boldsymbol{\beta} + \varepsilon_p^* \}}_{=\mathbf{X}^* \boldsymbol{\beta}} - \underbrace{\mathbb{E} \{ \mathbf{X}^* \mathbf{b} \}}_{=\mathbf{X}^* \boldsymbol{\beta}} = 0.$$

Puisque les résidus sont non corrélés, nous avons également

$$\begin{aligned} \sigma^2 \{ \text{pred}^* \} &= \sigma^2 \{ Y_p^* \} + \sigma^2 \{ \hat{Y}_p^* \} \\ &= \sigma^2 + \sigma^2 \mathbf{X}^* (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^*)^\top = \sigma^2 [1 + \mathbf{X}^* (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^*)^\top]. \end{aligned}$$

Donc,

$$\text{pred}^* \sim \mathcal{N} (0, \sigma^2 [1 + \mathbf{X}^* (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^*)^\top]).$$

L'estimation de l'erreur-type est ainsi

$$s\{\text{pred}^*\} = \sqrt{\text{MSE}} \sqrt{1 + \mathbf{X}^*(\mathbf{X}^\top \mathbf{X})^{-1}(\mathbf{X}^*)^\top}.$$

Comme au préalable, nous pouvons montrer que

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

$$\text{I.P.}(Y_p^*; 1 - \alpha) \equiv \mathbf{X}^* \mathbf{b} \pm t(1 - \frac{\alpha}{2}; n - p) \cdot s\{\text{pred}^*\}.$$

Mais $s\{\hat{Y}^*\} < s\{\text{pred}^*\}$, de sorte que l'I.C. de la réponse moyenne est toujours **contenu** dans le I.P. des nouvelles réponses.

Exemple : considérons la situation avec $n = 12$ observations et $p - 1 = 4$ prédicteurs décrit précédemment. Nous cherchons à prédire Y^* lorsque

$$\mathbf{X}^* = (1, 11.10, 20.74, 6.61, 182.38), \quad \text{dans la } \textbf{portée} \text{ du modèle.}$$

Nous avons déjà constaté que $\hat{Y}^* = \mathbf{X}^* \mathbf{b} = 100.40$. Rappelons que $\text{MSE} = 242.71$ et $\mathbf{X}^* (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^*)^\top = 1.42$, de sorte que

$$s\{\text{pred}^*\} = \sqrt{242.71} \sqrt{1 + 1.42} = 37.70.$$

Puisque $n - p = 7$, l'I.P. de Y^* à 95% est

$$\begin{aligned} \text{I.P.}(Y^*; 0.95) &\equiv 100.40 \pm t(0.975; 7) \cdot 37.70 \\ &= 100.40 \pm 2.365 \cdot 37.70 = [11.24, 189.56]. \end{aligned}$$

3.2.4 – Estimations et prédictions simultanées

À un niveau de confiance **conjoint** de $1 - \alpha$:

- la procédure de **Bonferroni** peut être utilisée pour estimer simultanément g paramètres du modèle $\beta_{k\ell}$, g réponses moyennes $E\{Y_\ell^*\}$, ou g nouvelles réponses Y_ℓ^* , pour $\ell = 1, \dots, g$;
- la procédure de **Working-Hotelling** peut être utilisée pour estimer simultanément g réponses moyennes $E\{Y_\ell^*\}$, pour $\ell = 1, \dots, g$;
- la procédure de **Scheffé** peut être utilisée pour prédire simultanément g nouvelles réponses Y_ℓ^* , pour $\ell = 1, \dots, g$.

L'approche est identique à celle de la RLS ; en fonction de la tâche à accomplir, nous choisissons la procédure appropriée qui donne le **plus petit intervalle**.

La seule différence réside dans la composition des **facteurs** qui accompagnent les erreurs-type dans la construction de **I.C./I.P.** à un niveau de confiance conjoint de $1 - \alpha$:

- $t(1 - \frac{\alpha}{g}; n - p)$ pour la procédure de Bonferroni ;
- $\sqrt{pF(1 - \alpha; p, n - p)}$ pour la procédure de Working-Hotelling, et
- $\sqrt{gF(1 - \alpha; g, n - p)}$ pour la procédure de Scheffé.

Exemple : nous pouvons fournir des intervalles de confiance conjoints pour les **paramètres du modèle** dans l'exemple précédent à un niveau de confiance conjoint de $1 - \alpha = 0.95$, en utilisant $n - p = 7$ et $g = 5$.

Le facteur de **Bonferroni** est $t\left(1 - \frac{0.05/5}{2}; 7\right) = t(0.995; 7) = 3.50$; les intervalles de confiance simultanés

$$\text{I.C.}_B(\beta_k; 0.95) \equiv b_k \pm 3.50 \cdot s\{b_k\}.$$

Paramètre	b_k	$\text{I.C.}_B(\beta_k; 0.95)$
β_0	-102.71	$[-830.22, 624.80]$
β_1	0.61	$[-0.685, 1.905]$
β_2	8.92	$[-9.63, 27.47]$
β_3	1.44	$[-6.925, 9.805]$
β_4	0.01	$[-2.685, 2.705]$

Individuellement, **aucun des paramètres n'est significatif** à un niveau de confiance conjoint de $1 - \alpha = 0.95$ (tous les intervalles de confiance contiennent **0**), mais la régression **dans son ensemble** est significative (cf. l'exemple du test F global).

De même, les I.C. conjoints de la moyenne $E\{Y_\ell^*\}$ (à un niveau de confiance de $\alpha = 0.05$) selon **Working-Hotelling** pour g prédicteurs \mathbf{X}_ℓ^* , $\ell = 1, \dots, g$ sont les suivants :

$$\begin{aligned} \text{I.C.}_{\text{WH}}(E\{Y_\ell^*\}; 0.95) &\equiv \hat{Y}_\ell^* \pm \sqrt{5F(0.95; 5, 7)} \cdot s\{\hat{Y}_\ell^*\} \\ &= \mathbf{X}_\ell^* \mathbf{b} \pm 4.46 \sqrt{\underbrace{242.71}_{=\text{MSE}}} \sqrt{\mathbf{X}_\ell^* (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}_\ell^*)^\top} \end{aligned}$$

3.3 – Puissance d'un test

Lorsque nous effectuons des tests d'hypothèse, nous pouvons commettre deux types d'erreurs.

Erreur de type I : rejeter H_0 lorsque H_0 est valide ;

Erreur de Type II : ne pas rejeter H_0 lorsque H_1 est valide.

En fait, il y a 4 types d'erreurs, mais ce n'est pas important ici.

Le **niveau de signification** α est utilisé pour contrôler le risque de commettre une erreur de type **I** ; les erreurs de type **II** sont plus difficiles à contrôler, en général.

À l'aide d'un test bilatéral, nous cherchons à tester

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{cases}$$

Soit α la probabilité de commettre une erreur de type I.

La **fonction de puissance**

$$K(\theta') = P(\text{rejeter } H_0 \text{ si } \theta = \theta')$$

est telle que $K(\theta_0) = \alpha$.

si $\theta \neq \theta_0$, $t^* = \frac{\hat{\theta} - \theta_0}{s\{\hat{\theta}\}}$ suit une loi T de Student avec ν degrés de liberté et un paramètre de non centralité

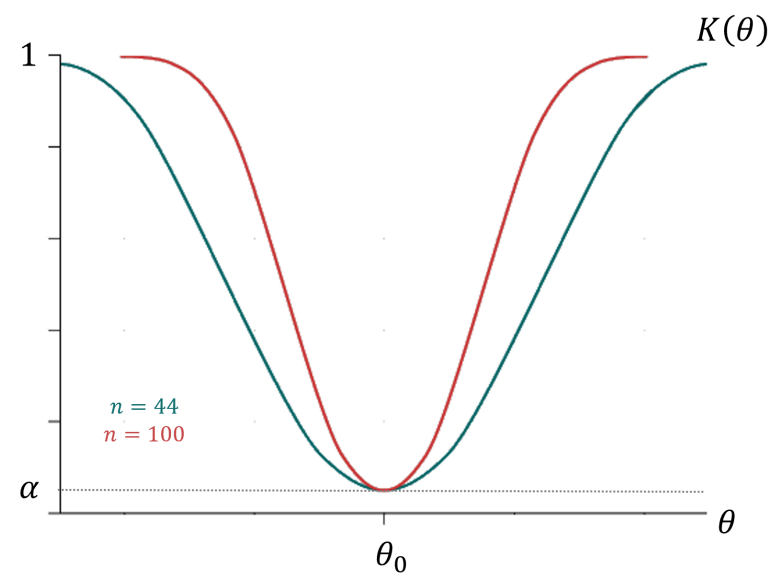
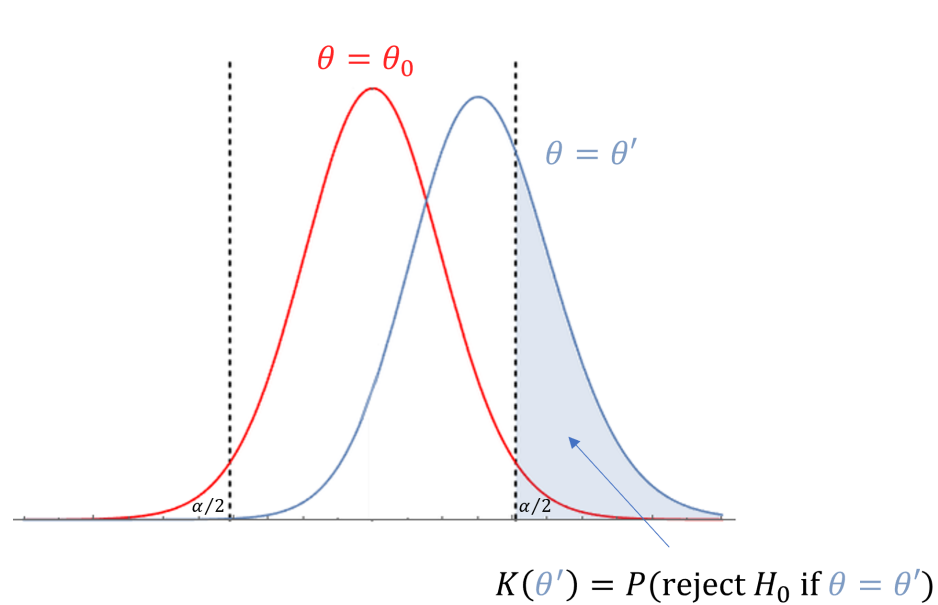
$$\delta = \frac{|\theta - \theta_0|}{\sigma\{\hat{\theta}\}} \approx \frac{|\theta - \theta_0|}{s\{\hat{\theta}\}},$$

où θ est la valeur réelle et θ_0 la valeur sous H_0 .

La **puissance du test** est la probabilité de rejeter H_0 si $\theta = \theta'$:

$$K(\theta') = P(|t^*| > t(1 - \alpha/2; \nu); \delta).$$

Pour contrôler la puissance, nous pouvons soit augmenter n , soit diminuer S_{xx} .



Exemple : nous recueillons quatre réponses pour chacun des prédicteurs $X = 5, 10, 15, 20, 25$. Les données d'un échantillon préliminaire donnent $\widehat{\sigma^2} = \text{MSE} = 1532.1$. Soit $E\{Y\} = b_0 + b_1X$. Si $\beta_1 = 6$, est-il probable que nous concluons que la régression est significative ?

Solution : nous avons

$$\sum_{i=1}^{20} X_i = 300 \text{ et } \sum_{i=1}^{20} X_i^2 = 5500 \implies S_{xx} = \sum_{i=1}^n X_i^2 - 20\bar{X}^2 = 1000.$$

Ainsi,

$$\delta \approx \frac{|\beta_1 - 0|}{s\{b_1\}} = \frac{|6 - 0|}{\sqrt{1532.1}/\sqrt{1000}} = 4.85.$$

À partir de la table B.5 avec $\nu = n - 2 = 18$ degrés de liberté et $\alpha = 0.05$, on constate que la puissance du test tombe dans $(0.97, 1.00)$ si $\beta_1 = 6$.

3.4 – Coefficients de détermination

Le **coefficient de détermination multiple** d'un modèle RLG est

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST},$$

la proportion de la variation de Y qui est expliquée par la régression.

Si le modèle de RLG incorpore un terme d'interception ($\beta_0 \neq 0$), alors

$$R^2 = r_{Y\hat{Y}}^2 = \frac{(s_{Y\hat{Y}})^2}{s_Y s_{\hat{Y}}};$$

ce n'est pas le cas si $\beta_0 = 0$.

Lorsque le nombre de paramètres p augmente, il en va de même pour R^2 ; cependant, les degrés de liberté, $n - p$ diminuent (estimations moins précises). Nous pouvons ajuster R^2 pour prendre cette perte en compte.

Le **coefficient de détermination multiple ajusté** d'un modèle de RLG est

$$R_a^2 = 1 - \frac{\text{SSE} / (n - p)}{\text{SST} / (n - 1)} = 1 - \frac{n - 1}{n - p} \cdot \frac{\text{SSE}}{\text{SST}} \quad (\text{pourrait être } < 0).$$

Exemple : dans le cas qui nous occupe depuis un moment, nous avons

$$\text{SST} = 6656.2, \quad \text{SSE} = 1699.0, \quad n - p = 7, \quad n - 1 = 11,$$

d'où

$$R^2 = 1 - \frac{1699.0}{6656.2} = 0.745 \quad \text{et} \quad R_a^2 = 1 - \frac{11}{7} \cdot \frac{1699.0}{6656.2} = 0.599.$$

3.5 – Diagnostiques et mesures correctives

Nous avons vu qu'il y a **quatre** hypothèses pour la RLG :

- **linéarité** – $E\{Y \mid \mathbf{X} = \mathbf{x}\} = \mathbf{x}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$
- **variance constante (homoscédasticité)** – $\sigma^2\{\varepsilon_i\} = \sigma^2, i = 1, \dots, n$
- **indépendance** – $\varepsilon_1, \dots, \varepsilon_n$ sont **indépendant** (ou **sans corrélation**)
- **normalité** – $\varepsilon_i \sim N(0, \sigma^2), i = 1, \dots, n$

Nous avons combiné ces hypothèses sous forme vectorielle :

$$Y \mid \mathbf{X} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n).$$

En théorie, ces hypothèses doivent être satisfaites avant de pouvoir faire confiance au modèle de RLG (le modèle peut s'avérer utile même si elles ne sont pas satisfaites, mais cela doit être établi **au cas par cas**).

Rappelons que nous avons les résultats suivants au sujet des **résidus** :

1. $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$, ou $e_i = Y_i - \hat{Y}_i$, pour $i = 1, \dots, n$
2. si $\beta_0 \neq 0$, alors $\bar{e} = 0$
3. $\sigma^2\{\mathbf{e}\} = \sigma^2(\mathbf{I}_n - \mathbf{H})$, d'où $\sigma^2\{e_i\} = \sigma^2(1 - h_{ii})$, pour $i = 1, \dots, n$, et $\sigma\{e_i, e_j\} = \sigma\{e_j, e_i\} = -h_{ij}\sigma^2$ pour $i \neq j = 1, \dots, n$.

L'**erreur-type** est $s^2\{e_i\} = \text{MSE}(1 - h_{ii})$ et la **studentisation interne des résidus** est $r_i = \frac{e_i - \bar{e}}{s\{e_i\}} \sim t(n - p)$, pour $i = 1, \dots, n$.

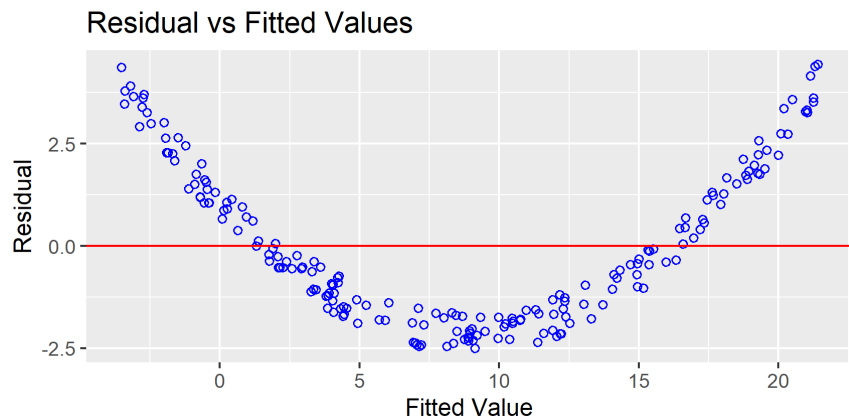
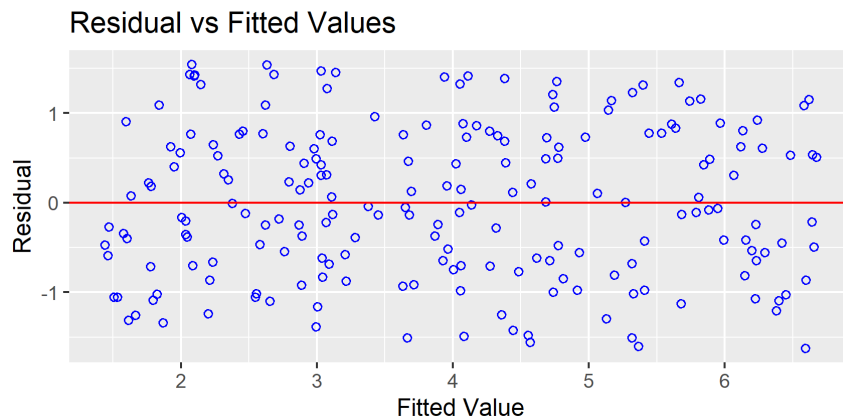
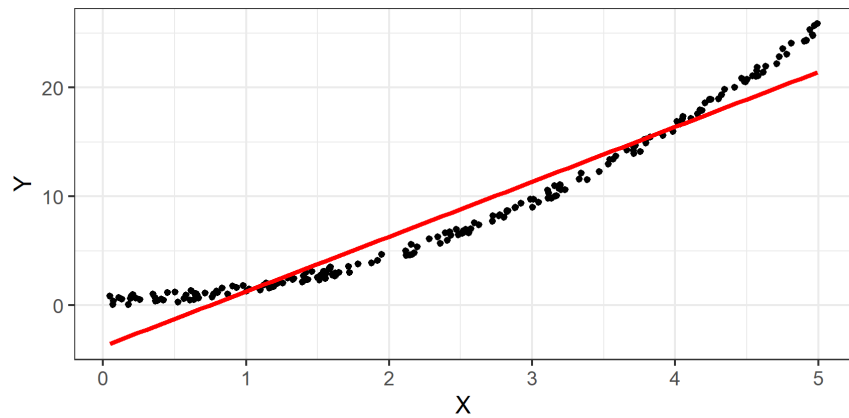
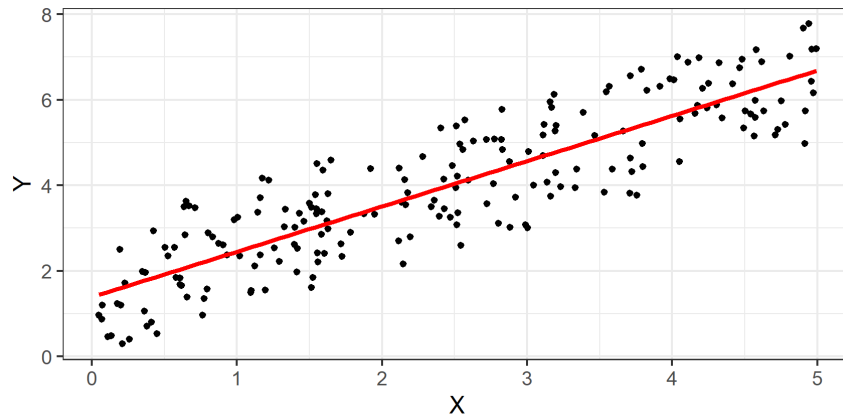
3.5.1 – Linéarité

Le graphique des résidus e_i envers les prédictions \hat{Y}_i permet de déterminer si l'hypothèse de linéarité est justifiée ; le cas échéant, les points devraient apparaître **dispersés aléatoirement autour de 0**.

L'**absence** d'une tendance suggère que la relation entre X_1, \dots, X_p et Y est effectivement linéaire, la **présence** d'une tendance fournit des preuves contre l'hypothèse de linéarité.

Il y a également des tests formels, tels que le test d'**inadéquation** (“lack-of-fit”):

$$\begin{cases} H_0 : E\{Y \mid \mathbf{X} = \mathbf{x}\} = \beta_0 + \beta_1 \mathbf{x}_1 + \dots + \beta_{p-1} \mathbf{x}_{p-1} \\ H_1 : H_0 \text{ n'est pas valide} \end{cases}$$



Soient $\mathbf{W}^1 = (X_1^1, \dots, X_{p-1}^1), \dots, \mathbf{W}^c = (X_1^c, \dots, X_{p-1}^c)$, les c niveaux de prédicteurs **distincts** ; le j ème niveau a n_j observations $Y_{i,j}$. Supposons que $E\{Y\}$ ait une **dépendance fonctionnelle en** X_1, \dots, X_{p-1} , que les résidus soient **indépendants** et suivent une **loi normale** $\mathcal{N}(0, \sigma^2)$, et qu'**au moins un** des $p - 1$ niveaux de prédiction X_k ait une valeur **répétée**.

Désignons la **moyenne des réponses** au niveau de prédicteur \mathbf{W}_j par \bar{Y}_j , et $SST_j = \sum_i^{n_j} (Y_{ij} - \bar{Y}_j)^2$. Le tableau ANOVA correspondant est

source	SS	deg	MS	F^*
Régression	SSR	$p - 1$	$SSR / (p - 1)$	MSLF / MSPE
Erreur	SSE	$n - p$	$SSE / (n - p)$	
Inadéquation	SSLF	$c - p$	$SSLF / (c - p)$	
Erreur pure	SSPE	$n - c$	$SSPE / (n - c)$	
Total	SST	$n - 1$		

Rappelons que $SST = SSE + SSR$. Partitionnons $SSE = SSPE + SSLF$, où $SSPE = \sum_{j=1}^c SST_j$ de sorte que $\frac{SSPE}{\sigma^2} \sim \chi^2 \left(\sum_{j=1}^c (n_j - 1) \right) = \chi^2(n - c)$.

Lorsque H_0 est valide, le **théorème de Cochran** implique que $\frac{SSE}{\sigma^2} \sim \chi^2(n - p)$, $\frac{SSLF}{\sigma^2} \sim \chi^2(c - p)$, et

$$F^* = \frac{\left(\frac{SSLF}{\sigma^2} \right) / (c - p)}{\left(\frac{SSPE}{\sigma^2} \right) / (n - c)} \sim F(c - p, n - c).$$

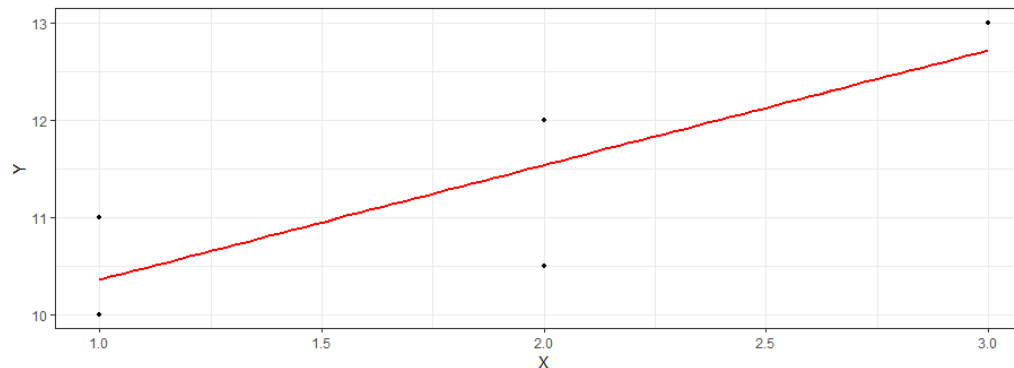
Règle de décision : si $F^* > F(1 - \alpha; c - p, n - c)$, on **rejette** H_0 à un niveau de confiance α .

Exemple : considérons l'ensemble de données avec les observations (X, Y) suivantes

$$(1, 10), (1, 11), (2, 10.5), (2, 12), (3, 13).$$

Le modèle linéaire $E\{Y\} = \beta_0 + \beta_1 X$ est-il justifié ?

Solution : nous avons $n = 5$, $p = 2$, et $c = 3$. La droite d'ajustement est $Y = 9.18 + 1.18X$, et le nuage de points est présenté ci-dessous.



Visuellement, il semble que la droite soit un bon modèle, mais il est difficile de l'affirmer avec certitude car il y a si peu de points dans le graphique. Nous utilisons le test formel d'inadéquation : nous avons

$$\begin{aligned}SST &= S_{yy} = 5.8, & SSR &= b_1^2 S_{xx} = 3.8829, & SSE &= SST - SSR = 1.91071, \\SSPE &= SST_1 + SST_2 + SST_3 = 0.5 + 1.125 + 0 = 1.625, \\SSLF &= SSE - SSPE = 1.91071 - 1.625 = 0.28571, \\MSLF &= \frac{SSLF}{c - p} = \frac{0.28571}{3 - 2} = 0.28571, & MSPE &= \frac{SSPE}{n - c} = \frac{1.625}{5 - 3} = 0.8125,\end{aligned}$$

d'où

$$F^* = \frac{MSLF}{MSPE} = \frac{0.28571}{0.8125} = 0.3516.$$

Puisque la valeur critique de la loi $F(3-2, 5-3) = F(1, 2)$ lorsque $\alpha = 0.05$ est 18.5, nous **ne rejetons pas** l'hypothèse de linéarité.

3.5.2 – Variance constante

Nous pouvons utiliser les graphiques de résidus pour déterminer si la condition d'homoscédasticité est remplie ou non. Mais il existe également des **tests formels**, tels que le test de **Breusch-Pagan** ou le test de **Brown-Forsythe**.

Examinons ce dernier. Sélectionnez un seuil $a \in \mathbb{R}$ et **partitionnez** les résidus en 2 groupes :

$$\text{Group 0: } \hat{Y} \leq a \text{ } (e_{i,0}\text{'s}) \quad \text{vs.} \quad \text{Group 1: } \hat{Y} > a \text{ } (e_{i,1}\text{'s}).$$

On choisit a de sorte que $|\text{Group 0}| = n_0 \approx n_1 = |\text{Group 1}|$. Pour $j = 0, 1$, soit \tilde{e}_j la **médiane des résidus du groupe j** ; posons $d_{ij} = |e_{ij} - \tilde{e}_j|$, l'**écart absolu du i ème résidu du groupe j par rapport à \tilde{e}_j** .

Nous utilisons la **médiane** et l'**écart absolu** plutôt que la **moyenne** et la **écart quadratique** en raison de la sensibilité aux valeurs aberrantes (c'est ce choix qui rend le test robuste à l'hypothèse de normalité).

Soit $\bar{d}_j = \frac{1}{n_j} \sum_i^{n_j} d_{ij}$, $j = 0, 1$. Nous testons

$$\begin{cases} H_0 : \bar{d}_0 = \bar{d}_1 & \text{(la variance est constante)} \\ H_1 : \bar{d}_0 \neq \bar{d}_1 & \text{(la variance **n'est pas** constante)} \end{cases}$$

en calculant la statistique de test

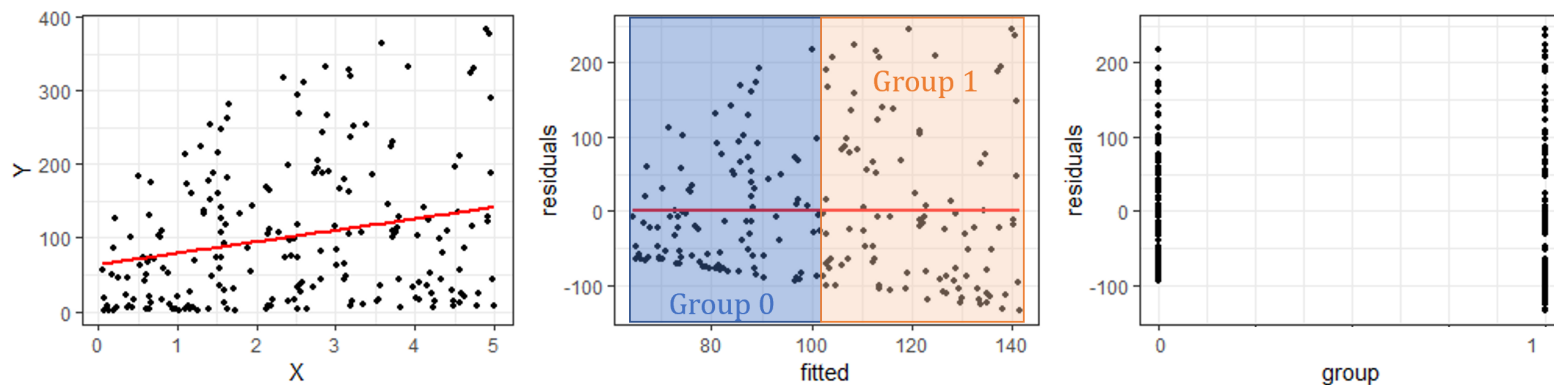
$$t_{\text{BF}}^* = \frac{\bar{d}_0 - \bar{d}_1}{s_p \sqrt{\frac{1}{n_0} + \frac{1}{n_1}}},$$

où la **variance groupée** est

$$s_p^2 = \frac{1}{n-2} \left[\sum_{i=1}^{n_0} (d_{i,0} - \bar{d}_0)^2 + \sum_{i=1}^{n_1} (d_{i,1} - \bar{d}_1)^2 \right] = \frac{(n_0 - 1)s_0^2 + (n_1 - 1)s_1^2}{n_0 + n_1 - 2}.$$

Lorsque H_0 est valide, $t_{BF}^* \sim t(n_0 + n_1 - 2) = t(n - 2)$.

Règle de décision : si $|t_{BF}^*| > t(1 - \alpha/2; n - 2)$, on **rejette** H_0 à un niveau de confiance α .



Exemple : dans les graphiques de la diapositive précédente, la valeur médiane ajustée est $a = 101.5096$. Visuellement, l'hypothèse de variance constante ne semble pas être respectée.

Nous divisons les ensembles de données en deux groupes, selon que la valeur ajustée est inférieure à a (Groupe 0, en bleu) ou non (Groupe 1, en orange) ; il y a $n_0 = n_1 = 100$ observations dans chaque groupe.

Les médianes de groupe des résidus sont $\tilde{e}_0 = -15.6$, $\tilde{e}_1 = -22.9$. Les quantités recherchées sont $\bar{d}_0 = 59.1$, $s^2_0 = 2197.745$, $\bar{d}_1 = 86.3$, et $s^1_0 = 4783.501$, donnant une variance groupée de $s^2_p = 3490.623$.

La statistique du test de Brown-Forsythe est $t^*_{BF} = -3.21$; puisque $|t^*_{BF}| = 3.21 > t(0.975; 198) = 1.97$, nous **rejetons** H_0 (variance égale), à un niveau de confiance $\alpha = 0.05$.

3.5.3 – Indépendance

L'indépendance des termes d'erreur peut être évaluée visuellement en traçant le graphique des **résidus** e_i envers les **valeurs ajustées** \hat{Y}_i .

Si les erreurs sont **indépendantes**, la corrélation entre celles-ci devrait être faible ($|\rho| \approx 0$) ; si un modèle ou une tendance émerge, alors elles sont probablement **dépendantes**.

La corrélation à l'exemple précédent est si **faible** ($\rho = -6 \times 10^{-18}$) que nous pouvons raisonnablement les traiter comme **indépendants**.

L'hypothèse de RLG est que les **erreurs** sont indépendantes, mais nous ne travaillons jamais qu'avec les **résidus**, qui sont définitivement **non indépendants** ($\bar{e} = 0$).

3.5.4 – Normalité

Si les termes d'erreur suivent une loi **normale** $\mathcal{N}(0, \sigma^2)$, nous nous attendons à ce que les résidus le fassent également.

1. Ainsi, si l'histogramme des **résidus studentisés**

$$r_i = \frac{e_i}{s\{e_i\}} = \frac{Y_i - \hat{Y}_i}{\sqrt{\text{MSE}}\sqrt{1 - h_{ii}}}$$

n'est pas symétrique, alors ils ne suivent pas une loi normale réduite $\mathcal{N}(0, 1)$ et les termes d'erreur ont peu de chances d'être normaux.

2. Si l'histogramme est symétrique, on construit le graphe de **probabilité normale** (graphe **quantile-quantile**) à partir des **résidus studentisés**.

Pour tout $i = 1, \dots, n$, on construit la table suivante :

i	résidu studentisé	rang	pourcentile	quantile z
1	r_1	k_1	p_1	z_1
\vdots	\vdots	\vdots	\vdots	\vdots
i	r_i	k_i	p_i	z_i
\vdots	\vdots	\vdots	\vdots	\vdots
n	r_n	k_n	p_n	z_n

Le **rang** k_i Le **rank** k_i est donné par ordre **croissant** (les égalités utilisent la moyenne) ; le **pourcentile approximatif** est

$$p_i = \frac{k_i - 0.375}{n + 0.25}, \quad (\text{position blom});$$

le **quantile** est $z_i = \Phi^{-1}(p_i)$, où $\Phi(z) = P(Z \leq z)$, $Z \sim \mathcal{N}(0, 1)$.

3. Représenter graphiquement les résidus studentisés r_i en fonction des quantiles z_i – les points devraient tomber au hasard autour de la ligne “**normale**”, sans tendance systématique à s’en éloigner. Si ce n’est pas le cas, il est peu probable que les erreurs soient normales.
4. Calculer la **corrélation** ρ entre r_i et z_i , $i = 1, \dots, n$. Afin de tester

$$\begin{cases} H_0 : \text{termes d'erreur suivent une loi normale} \\ H_1 : H_0 \text{ n'est pas valide} \end{cases}$$

nous trouvons la valeur critique ρ_α du **coefficient de corrélation du graphique de probabilité** (PPCC) normal pour n et α .

Règle de décision : si $\rho < \rho_\alpha$, on **rejete** H_0 à un niveau de confiance α .

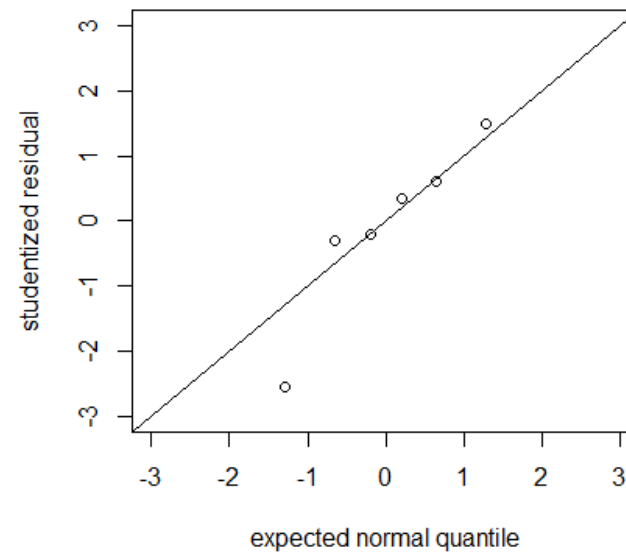
Exemple : considérons un ensemble de données avec les observations

$$(1, 7.4), (1, 8.0), (2, 7.0), (2, 10.4), (3, 19.1), (4, 20.3).$$

Supposons un modèle linéaire $E\{Y\} = \beta_0 + \beta_1 X$. L'hypothèse de normalité des termes d'erreur est-elle justifiée ?





Solution : le modèle linéaire est $E\{Y\} = 1.802 + 4.722X$; nous avons

x	y	résidu studentisé	rang	pourcentile	quantile z
1	7.4	0.35	4	0.58	0.20
1	8.0	0.60	5	0.74	0.64
2	7.0	-2.57	1	0.10	-1.28
2	10.4	-0.29	2	0.26	-0.64
3	19.1	1.48	6	0.90	1.28
4	20.3	-0.21	3	0.42	-0.20



La corrélation entre les résidus studentisés et le quantile z est $\rho = 0.939$. À un niveau de confiance de $\alpha = 0.05$, la valeur critique de la corrélation dans le tableau PPCC avec $n = 6$ est 0.888 ; nous **ne rejetons donc pas** l'hypothèse de normalité (ce qui n'équivaut pas à **accepter** H_0 , bien sûr).

3.5.5 – Mesures correctives

Les transformations en X sont utilisées lorsque les données présentent une **tendance monotone non linéaire** avec **constance de la variance** ; si la tendance est , on essaie $X' = \ln X$ ou $X' = \sqrt{X}$; si la tendance est , on essaie $X' = e^X$ ou $X' = X^2$; si elle est , on essaie $X' = \frac{1}{X}$ ou $X' = \exp(-X)$; si elle est , on essaie $X' = \exp(-X^2)$.

Les transformations en Y sont utilisées lorsque les données présentent une **tendance non linéaire monotone** et une variance **NON constante**, mais il est difficile de déterminer à partir des diagrammes de dispersion quelle transformation sur Y utiliser. La transformation de **Box-Cox** trouve une puissance λ appropriée pour le modèle RLG

$$Y_i^{(\lambda)} = \mathbf{X}_i \boldsymbol{\beta} + \varepsilon,$$

où \mathbf{X}_i est la i ème ligne de \mathbf{X} . On définit

$$Y^{(\lambda)} = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \ln Y, & \lambda = 0 \end{cases}$$

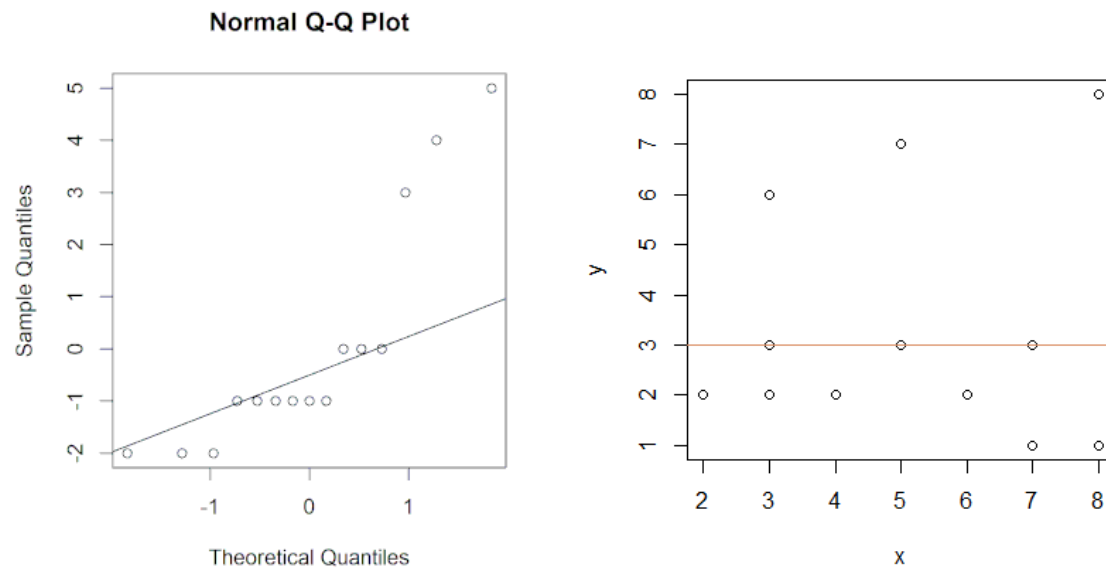
et on choisit le λ qui minimise le $\text{SSE}(\lambda)$ résultant des régressions.

Les moindres carrés pondérés sont utilisés si les données présentent une **tendance linéaire** mais **non une variance constante**. Une alternative serait d'utiliser d'abord une transformation en Y pour contrôler la **variance**, puis une transformation en X pour contrôler la **linéarité** qui pourrait avoir été détruite par la première transformation.

Exemple : considérons les données suivantes

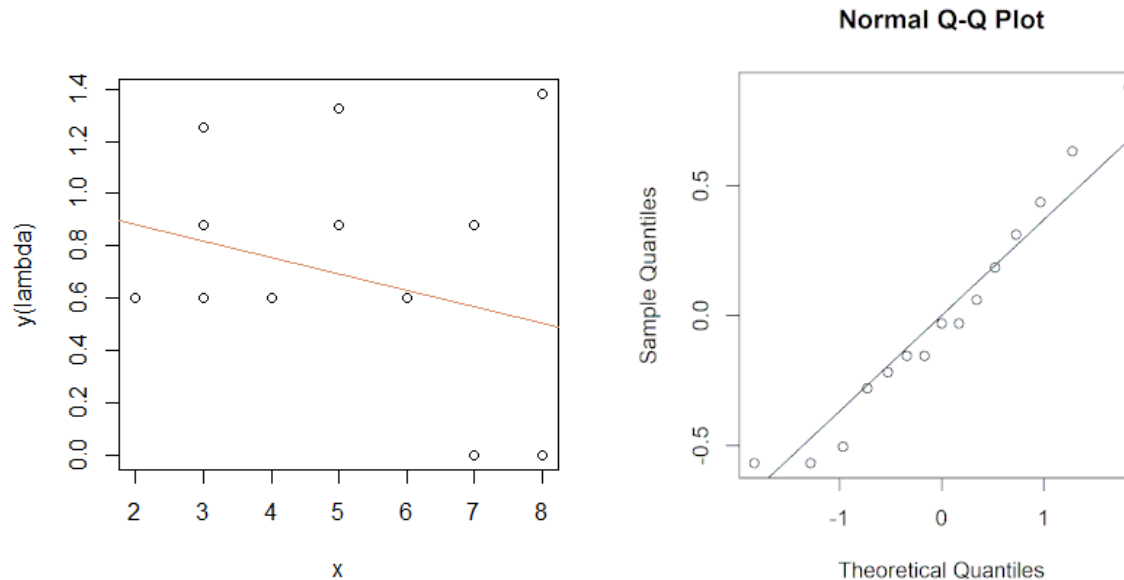
$(7, 1), (7, 1), (8, 1), (3, 2), (2, 2), (4, 2), (4, 2), (6, 2),$
 $(6, 2), (7, 3), (5, 3), (3, 3), (3, 6), (5, 7), (8, 8).$

La droite d'ajustement et le graphique QQ normal sont présentés ci-dessous.



Le graphique QQ montre que les termes d'erreur ne sont pas normaux, et donc que le modèle de régression n'est pas valide. La variance n'est pas constante, nous utilisons la transformation de Box-Cox avec $\lambda = -0.42$.

Voici les résultats sur les données transformées :



IMPORTANT: le modèle linéaire original est $E\{Y\} = 3 + 0 \cdot X$. Le modèle linéaire sur les données transformées est

$$E\left\{Y^{(-0.42)}\right\} = 1.00564 - 0.06264X,$$

ce qui donne

$$\begin{aligned} E\{Y\} &= ([\lambda\beta_0 + 1] + \lambda\beta_1 X)^{1/\lambda} \\ &= \left([-0.42(1.00564) + 1] + 0.42 \cdot 0.06264X\right)^{1/(-0.42)} \\ &= \frac{1}{(0.5776 + 0.0263X)^{2.380}}, \end{aligned}$$

qui **ne représente pas** une droite dans le plan xy .