Statistique appliquée

PROF. ARMEL YODÉ

Table des matières

Chapitre 1

Régression linéaire

1.1 Le Modèle de régression linéaire

Un modèle est une simple description d'un état ou d'un processus. Le modèle de régression linéaire est un modèle statistique défini par

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon \tag{1.1.1}$$

où

- Y est une variable aléatoire réelle appelée variable à expliquer ou variable réponse ou variable dépendante ou variable endogène ;
- $X_1,\ldots,X_p$ sont des variables réelles également observées appelées variables explicatives ou prédicteurs ou variables exogènes ;
- $\beta_0,\ldots,\beta_p$ sont des paramètres non observés ;
- ε est le terme d'erreur ; c'est une variable aléatoire réelle non observée.

Le modèle de régression linéaire est dit simple si $p = 1$, c'est à dire,

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon.$$

Le modèle de régression linéaire est dit multiple si $p > 1$.

Exemple 1. *Quelques exemples de problèmes :*

1. *le salaire en fonction de certaines caractéristiques socio-démographiques telles que l'ancienneté dans l'entreprise (en années), le nombre d'années d'études après le bac ;*
2. *le budget consacré à la consommation des ménages en fonction du revenu du ménage, du nombre de personnes par ménage ;*

La régression linéaire a trois objectifs essentiels :

- mesurer l'impact ou l'effet de $X_1,\ldots,X_p$ sur Y
- prédire Y connaissant $X_1,\ldots,X_p$.
- parmi les variables $X_1,\ldots,X_p$, identifier celles qui expliquent de manière efficace (avec précision) la variable Y.

Notons

- Y_i, l'observation de la variable Y sur l'individu i
- X_{ij}, l'observation de la variable X_j sur l'individu i
- ε_i, l'erreur pour l'individu i.

En régression linéaire, on dispose de n données $(X_1,Y_1),\dots,(X_n,Y_n)\in\mathbb{R}^p\times\mathbb{R}$ vérifiant

$$Y_i=\beta_0+\sum_{j=1}^{p}\beta_jX_{ij}+\varepsilon_i \qquad 1\le i\le n.$$

En posant

$$Y=\begin{pmatrix} Y_1\\ \vdots\\ Y_i\\ \vdots\\ Y_n\end{pmatrix} \quad X=\begin{pmatrix} 1 & X_{11} & \cdots & X_{1j} & \cdots & X_{1p}\\ \vdots & \vdots & & \vdots & & \vdots\\ 1 & X_{i1} & \cdots & X_{ij} & \cdots & X_{ip}\\ \vdots & \vdots & & \vdots & & \vdots\\ 1 & X_{n1} & \cdots & X_{nj} & \cdots & X_{np}\end{pmatrix}$$

$$\beta=\begin{pmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_j\\ \vdots\\ \beta_p\end{pmatrix} \quad \varepsilon=\begin{pmatrix}\varepsilon_1\\ \vdots\\ \varepsilon_i\\ \vdots\\ \varepsilon_n\end{pmatrix},$$

on obtient la forme matricielle suivante :

$$Y=X\beta+\varepsilon. \tag{1.1.2}$$

Nous avons les hypothèses suivantes :

($\mathscr{H}_0$) X_j n'est pas aléatoire pour $j=1,\dots,p$.

($\mathscr{H}_1$) $rang(X)=p+1$.

($\mathscr{H}_2$) $\mathbb{E}(\varepsilon)=0$ et $\mathrm{Var}(\varepsilon)=\sigma^2I_n$ avec $\sigma^2>0$.

($\mathscr{H}_3$) $\varepsilon\hookrightarrow\mathcal{N}(0,\sigma^2I_n)$ avec $\sigma^2>0$.

Remarque 1. ($\mathscr{H}_1$) *implique que les colonnes de* X *forment des vecteurs linéairement indépendants de* $\mathbb{R}^n$*. Ainsi, nous avons*

$$\forall C\in\mathbb{R}^{p+1},\quad XC=0\Rightarrow C=0;$$

il existe donc un unique vecteur β *associé au modèle* (1.1.2) *; de plus, on a* $n\ge p+1$ *; si l'on avait* $rang(X)<p+1$*, cela signifierait qu'il existe au moins une variable explicative qui peut s'écrire comme une combinaison linéaire d'une ou des autres variables explicatives : cette variable explicative serait donc superflue, elle n'apporterait rien à l'explication de* Y *déjà fournie par les autres variables explicatives.*
($\mathscr{H}_2$) *implique que les composantes de* ε *sont centrées, de même variance (homoscédasticité) et non correlées entre elles.*
($\mathscr{H}_3$) *implique que les erreurs* $\varepsilon_1,\dots,\varepsilon_n$ *sont indépendantes identiquement distribuées de loi* $\mathcal{N}(0,\sigma^2)$.

1.2 Estimateurs et propriétés

1.2.1 Estimateur des moindres carrés ordinaires

Les paramètres inconnus du modèle sont : $\beta \in \mathbb{R}^{p+1}$ et $\sigma^2 > 0$.

Proposition 1. *Dérivée matricielle*
Pour tout $v, a \in \mathbb{R}^k$, pour toute matrice carrée d'ordre k, nous avons

- $\dfrac{\partial v^{'} a}{\partial v} = \dfrac{\partial a^{'} v}{\partial v} = a$
- $\dfrac{\partial v^{'} M v}{\partial v} = (M + M^{'})v$

Nous supposons vérifier les hypothèses $(\mathscr{H}_0)$ et $(\mathscr{H}_1)$.

Définition 1. *On appelle estimateur des moindres carrés ordinaires $\hat{\beta}$, la valeur de β qui minimise la fonction suivante*

$$S(\beta) = (Y - X\beta)^{'}(Y - X\beta)$$

Comme $\varepsilon = Y - X\beta$, on a

$$S(\beta) = \varepsilon^{'}\varepsilon = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}) \right)^2 .$$

Théorème 1. *L'estimateur des moindres carrés ordinaires $\hat{\beta}$ de β est défini par*

$$\hat{\beta} = (X^{'}X)^{-1}X^{'}Y = \begin{pmatrix} \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}.$$

Démonstration. Nous avons

$$\begin{aligned} S(\beta) &= (Y - X\beta)^{'}(Y - X\beta) \\ &= Y^{'}Y - Y^{'}X\beta - \beta^{'}X^{'}Y + \beta^{'}X^{'}X\beta \end{aligned}$$

$$\frac{\partial S(\beta)}{\partial \beta} = -2X^{'}Y + 2X^{'}X\beta = 0 \Rightarrow \beta = (X^{'}X)^{-1}X^{'}Y$$

Comme $\dfrac{\partial^2 S(\beta)}{\partial \beta^2} = 2X^{'}X$ est une matrice définie positive, on obtient le résultat. □

On suppose vérifier les hypothèses $\mathscr{H}_0$, $\mathscr{H}_1$ et $\mathscr{H}_2$.

Proposition 2. *$\hat{\beta}$ est un estimateur sans biais de β i.e.*

$$\mathbb{E}(\hat{\beta}) = \beta.$$

Démonstration. En exercice. □

Proposition 3. *La matrice de variance-covariance de $\hat{\beta}$*

$$Var(\hat{\beta}) = \sigma^2 (X^{'}X)^{-1}.$$

Démonstration. En exercice. □

Théorème 2. *(de Gauss-Markov)*
Parmi les estimateurs sans biais de β de la forme $\boldsymbol{BY}$, où $\boldsymbol{B} \in \mathscr{M}_{p+1,n}(\mathbb{R})$, $\hat{\beta}$ est optimal, c'est à dire de variance minimale.

Démonstration. Soit $\boldsymbol{BY}$ un autre estimateur linéaire sans biais de β. Puisque $\mathbb{E}(\boldsymbol{BY}) = \boldsymbol{BX}\beta$, on a $\boldsymbol{BX} = \mathbb{I}_{p+1}$. Par suite, $\boldsymbol{CX} = 0$ avec $\boldsymbol{C} = \boldsymbol{B} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Ainsi, nous avons

$$\begin{aligned} Var(BY) &= BVar(Y)B' \\ &= [C + (X'X)^{-1}X']\sigma^2\mathbb{I}_n[C + (X'X)^{-1}X']' \\ &= \sigma^2 CC' + Var(\hat{\beta}). \end{aligned}$$

Par suite, $Var(BY) - Var(\hat{\beta})$ est une matrice symétrique positive pour tout $B \in \mathscr{M}_{p+1,n}(\mathbb{R})$ □

Définition 2. *Le vecteur* $\hat{Y} = X\hat{\beta} = \begin{pmatrix} \hat{Y}_1 \\ \vdots \\ \hat{Y}_n \end{pmatrix}$ *est le* **vecteur des valeurs ajustées**, *où*

$$\hat{Y}_i = \hat{\beta}_0 + \sum_{j=1}^{p} X_{ij}\hat{\beta}_j$$

Définition 3. *Le vecteur* $\hat{\varepsilon} = Y - \hat{Y} = \begin{pmatrix} \hat{\varepsilon}_1 \\ \vdots \\ \hat{\varepsilon}_n \end{pmatrix}$ *est appelé* **vecteur des résidus estimés.**

Posons

$$H = X(X'X)^{-1}X'.$$

La matrice $\boldsymbol{H}$ est appelée la "matrice chapeau" ou "hat matrix".
Nous pouvons écrire alors

$$\hat{Y} = HY \qquad \hat{\varepsilon} = (\mathbb{I} - H)Y.$$

Posons

$$\hat{\sigma}^2 = \frac{\|\hat{\varepsilon}\|^2}{n-p-1} = \frac{1}{n-p-1}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

Proposition 4. *Sous les hypothèses $\mathscr{H}_0$, $\mathscr{H}_1$ et $\mathscr{H}_2$, la statistique $\hat{\sigma}^2$ est un estimateur sans biais de σ^2.*

Démonstration. Nous avons

$$\mathbb{E}(\hat{\varepsilon}'\hat{\varepsilon}) = \mathbb{E}(tr(\hat{\varepsilon}'\hat{\varepsilon})) = \mathbb{E}(tr(\hat{\varepsilon}\hat{\varepsilon}')) = tr(Var(\hat{\varepsilon})) = tr(\sigma^2(\mathbb{I} - H)) = \sigma^2(n-p-1).$$

□

Nous obtenons ainsi un estimateur de $Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$ en remplaçant σ^2 par son estimateur $\hat{\sigma}^2$:

$$\hat{\sigma}^2_{\hat{\beta}} = \hat{\sigma}^2(X'X)^{-1}.$$

Nous avons donc un estimateur de l'écart-type de l'estimateur $\hat{\beta}_j$ du coefficient β_j :

$$\hat{\sigma}_{\hat{\beta}_j} = \hat{\sigma}\sqrt{[(X'X)^{-1}]_{j+1,j+1}} \quad j = 0, \ldots, p. \tag{1.2.1}$$

C'est le $(j+1)$-ième coefficient diagonal de la matrice $\hat{\sigma}^2(X'X)^{-1}$; $\hat{\sigma}_{\hat{\beta}_j}$ est un indicateur du caractère plus ou moins stable de l'estimation de β_j.
La source principale d'instabilité dans l'estimation de β est la multicolinéarité (les variables explicatives sont très corrélées entre elles). Comme $\text{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$, si les variables explicatives sont très corrélées entre elles, $X'X$ aura un déterminant proche de 0 et son inverse aura des termes élevés. Les paramètres du modèle seront estimés avec imprécision et les prédictions pourront être entachées d'erreurs considérables même si R^2 a une valeur élevée.

1.2.2 Modèles gaussiens

Dans cette section, nous supposons vérifier les hypothèses **$(\mathscr{H}_0)$, $(\mathscr{H}_1)$ et $(\mathscr{H}_3)$.**

1.2.2.1 Méthode du maximum de vraisemblance

La vraisemblance de l'échantillon est défini par

$$L(Y,\beta,\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2\right].$$

Ainsi, nous avons

$$\ln(L(Y,\beta,\sigma^2)) = -\ln((2\pi)^{\frac{n}{2}}\sigma^n) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2.$$

De ce fait, maximer la vraisemblance revient à utiliser la méthode des moindres carrés ordinaires. Ainsi, l'estimateur du maximum de vraisemblance et l'estimateur des moindres carrés ordinaires coïncident.

1.2.2.2 Propriétés des estimteurs $\widehat{\beta}$ et $\widehat{\sigma}^2$

Proposition 5. *$\hat{\beta}$ est un vecteur gaussien de moyenne β et de variance $\sigma^2(X^TX)^{-1}$*

Proposition 6. $\dfrac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \hookrightarrow \chi^2(n-p-1)$.

Ici $\chi^2(n-p-1)$ est la loi du khi-deux à $n-p-1$ degrés de liberté.

Proposition 7. *$\widehat{\beta}$ et $\widehat{\sigma}^2$ sont indépendantes*

Proposition 8. *Pour $j = 0,1,\ldots,p$, la variable*

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \hookrightarrow \mathscr{T}(n-p-1).$$

$\mathscr{T}(n-p-1)$ est la loi de Student à $n-p-1$ degrés de liberté.

1.2.2.3 Intervalle de confiance d'un coefficient β_j et de σ^2

Proposition 9. *Un intervalle de confiance de niveau $1-\alpha$ pour β_j, $j = 1,\ldots,p$ est donné par*

$$\left[\hat{\beta}_j - t^*\hat{\sigma}_{\hat{\beta}_j}, \hat{\beta}_j + t^*\hat{\sigma}_{\hat{\beta}_j}\right]$$

où t^* est le quantile d'ordre $1-\alpha/2$ de $\mathscr{T}(n-p-1)$.

Démonstration. En effet, $T_j = \dfrac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}}$ est une fonction pivot pour β_j (voir Proposition 8). □

Proposition 10. *Un intervalle de confiance de niveau* $1-\alpha$ *pour* σ^2 *est donné par*

$$\left[\frac{(n-p-1)\hat{\sigma}^2}{c_2}, \frac{(n-p-1)\hat{\sigma}^2}{c_1}\right]$$

avec $\mathbb{P}(c_1 \leq Z \leq c_2) = 1-\alpha$ *où* $Z \hookrightarrow \chi^2(n-p-1)$.

Démonstration. En effet $\dfrac{(n-p-1)\hat{\sigma}^2}{\sigma^2}$ est une fonction pivot pour σ^2 (voir Proposition 6). □

1.2.2.4 Test de validité marginale de Student

Nous considérons le test de l'hypothèse $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$. La statistique de test est $\dfrac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}}$. Sous H_0 (voir Proposition 8),

$$\frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}} \hookrightarrow T(n-p-1).$$

La région critique du test au niveau α est

$$W = \left\{\left|\frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}\right| > t^*\right\}$$

où t^* est le quantile d'ordre $1-\alpha/2$ de $\mathcal{T}(n-p-1)$. Nous rejetons H_0 si $\left|\frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}\right| > t^*$ Nous concluons dans ce cas que le coefficient X_j est significatif.

1.2.2.5 Critère de comparaison de modèle

Soient

dispersion totale : $SCT = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$

dispersion expliquée par le modèle : $SCE = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$

dispersion résiduelle : $SCR = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$.

Alors, l'équation d'analyse de variance est

$$SCT = SCE + SCR.$$

Coefficient de détermination : Le pourcentage de variabilité dû au modèle se mesure par le coefficient de détermination :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

- Si $R^2 = 1 \Leftrightarrow Y = X\beta$.

- Si $R^2 \simeq 0 \Leftrightarrow$ résidus élevés$\Leftrightarrow$modèle de régression linéaire inadapté.

Coefficient de détermination ajusté : Le R^2 ajusté est défini par

$$R^2_{ad} = 1 - \frac{SCR/(n-p-1)}{SCT/(n-1)} = \frac{(n-1)R^2 - p}{n-p-1} = 1 - \frac{n-1}{n-p-1}(1-R^2).$$

Nous avons les propriétés suivantes :

— $R^2_{ad} < R^2$ dès que $p \geq 2$

— R^2_{ad} peut prendre des valeurs négatives.

— R^2_{ad} n'augmente pas forcément lors de l'introduction de variables supplémentaires dans le modèle.

— Possibilité de comparer deux modèles n'ayant pas le même nombre de variables à l'aide du R^2_{ad} et choisir modèle pour lequel R^2_{ad} est le plus grand.

1.2.2.6 Test de validité globale de Fisher

Nous considérons le test de l'hypothèse

$$H_0 : \beta_1 = \cdots = \beta_p = 0$$

contre

$$H_1 : \exists j \in \{1,\ldots,p\}, \quad \beta_j \neq 0.$$

Nous testons l'hypothèse que tous les coefficients sont nuls excepté la constante. Sous H_0

$$F = \frac{\frac{SCE}{p}}{\frac{SCR}{n-p-1}} = \frac{R^2}{1-R^2}\frac{n-p-1}{p} \hookrightarrow F(p, n-p-1)$$

la loi de Fisher à p et $n-p-1$ dégrés de liberté. On rejette H_0 si $F > F_{1-\alpha}$ où $F_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de Fisher et on conclut qu'il existe au moins un paramètre non nul dans le modèle.

1.2.2.7 Prévision

Soit une nouvelle valeur $X'_{n+1} = (1, X_{n+1,1}, \ldots, X_{n+1,p})$ et nous voulons prédire Y_{n+1}. Or

$$Y_{n+1} = X'_{n+1}\beta + \varepsilon_{n+1}$$

avec $\mathbb{E}(\varepsilon_{n+1}) = 0$, $var(\varepsilon_{n+1}) = \sigma^2$ et $cov(\varepsilon_{n+1}, \varepsilon_i) = 0$ pour $i = 1,\ldots,n$. La prévision de Y_{n+1} est

$$Y^p_{n+1} = X'_{n+1}\hat{\beta}.$$

Deux types d'erreurs vont entacher la prévision :

- la première due à l'incertitude sur ε_{n+1}
- l'autre due à l'incertitude due à l'estimation.

L'espérance de l'erreur de prévision est $\mathbb{E}(Y_{n+1} - Y^p_{n+1}) = 0$. La variance de l'erreur de prévision est

$$var(Y_{n+1} - Y^p_{n+1}) = \mathbb{E}(Y_{n+1} - Y^p_{n+1})^2 = \sigma^2(1 + X'_{n+1}(X'X)^{-1}X_{n+1}).$$

Nous retrouvons bien l'incertitude due aux erreurs σ^2 sur laquelle vient s'ajouter l'incertitude de l'estimation.

Théorème 3. *Un intervalle de confiance de niveau $1-\alpha$ pour Y_{n+1} est donné par*

$$\left[Y^p_{n+1} \pm t^* \hat{\sigma} \sqrt{1 + X'_{n+1}(X'X)^{-1}X_{n+1}} \right]$$

où t^* est le quantile d'ordre $1-\alpha/2$ de $T(n-p-1)$.

1.2.3 Cas de la regression linéaire simple

1.2.3.1 Estimateurs et propriétés

Dans le cas de la regression linéaire simple, c'est à dire, $p = 1$, nous avons $\hat{\beta} = (X^T X)^{-1} X^T Y = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$ avec

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i X_i - n\bar{Y}_n \bar{X}_n}{\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2}$$

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

$$\widehat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 \quad \text{avec } \widehat{\varepsilon}_i = Y_i - \widehat{Y}_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i).$$

On note :

$$\hat{\sigma}^2_{\hat{\beta}_0} = \hat{\sigma}^2 \Big(\frac{1}{n} + \frac{\overline{X}_n^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \Big) \qquad \hat{\sigma}^2_{\hat{\beta}_1} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}. \tag{1.2.2}$$

On note aussi :

- $\chi^2(n)$, la loi de Khi-deux à n degrés de liberté
- $T(n)$, la loi de Student à n degrés de liberté.

Proposition 11. *Sous les hypothèses $(\mathscr{H}_0)$, $(\mathscr{H}_1)$ et $(\mathscr{H}_3)$, nous avons les résultats suivants :*

- $\dfrac{(n-2)\hat{\sigma}^2}{\sigma^2} \hookrightarrow \chi^2(n-2)$.
- $\dfrac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \hookrightarrow T(n-2)$.
- $\dfrac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \hookrightarrow T(n-2)$.

Nous proposons des intervalles de confiance des paramètre β_0 et β_1.

Proposition 12. *Un intervalle de confiance de β_0 de niveau $1-\alpha$ est donné par :*

$$[\hat{\beta}_0 - t_{1-\frac{\alpha}{2}} \hat{\sigma}_{\hat{\beta}_0}, \hat{\beta}_0 + t_{1-\frac{\alpha}{2}} \hat{\sigma}_{\hat{\beta}_0}]$$

où $t_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de $T(n-2)$.

Proposition 13. *Un intervalle de confiance de β_1 de niveau $1-\alpha$ est donné par :*

$$[\hat{\beta}_1 - t_{1-\frac{\alpha}{2}} \hat{\sigma}_{\hat{\beta}_1}, \hat{\beta}_1 + t_{1-\frac{\alpha}{2}} \hat{\sigma}_{\hat{\beta}_1}]$$

où $t_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de $T(n-2)$.

Proposition 14. *Un intervalle de confiance de σ^2 est donné par*

$$\left[\frac{(n-2)\hat{\sigma}^2}{b}, \frac{(n-2)\hat{\sigma}^2}{a} \right]$$

où a est le quantile d'ordre α_1 et b est le quantile d'ordre $1-\alpha_2$ de $\chi^2(n-2)$ avec $\alpha = \alpha_1 + \alpha_2$.

Nous proposons ensuite un intervalle de confiance de la droite de régression $\beta_0 + \beta_1 x_*$.

Proposition 15. *Un intervalle de confiance pour $\beta_0 + \beta_1 x^*$ est donné par*

$$\left[\hat{y}_* \pm t_{1-\frac{\alpha}{2}} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^* - \overline{X}_n)^2}{\sum_{i=1}^n (x^* - \overline{X}_n)^2}}\right]$$

où

$$\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x^*.$$

Tests de validité marginale

Nous considérons le test de l'hypothèse

$$H_0 : \beta_1 = 0 \quad \text{contre} \quad H_1 : \beta_1 \neq 0.$$

Si H_1 est rejetée, on dira que le coefficient β_1 n'est pas significatif. Dans le cas contraire, on dira que le coefficient est significatif et que la variable X influe sur la variable Y.
La région critique du test est donnée par

$$W = \left\{\left|\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}\right| > C\right\}$$

où C le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi de Student $T(n-2)$.

Prévision

La valeur pour laquelle nous effectuons la précision n'a pas servi dans le calcul des estimateurs. Soit X_{n+1} cette valeur. Nous voulons prédire Y_{n+1}. Le modèle indique que $Y_{n+1} = \beta_0 + \beta_1 X_{n+1} + \varepsilon_{n+1}$ avec $\mathbb{E}(\varepsilon_{n+1}) = 0$, $var(\varepsilon_{n+1}) = \sigma^2$ et $Cov(\varepsilon_{n+1}, \varepsilon_i) = 0$ pour $i = 1, \ldots, n$. Nous pouvons prédire Y_{n+1} gràce au modèle estimé :

$$\hat{Y}^p_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 X_{n+1}.$$

Deux types d'erreurs entachent notre prévision :

- l'une due à la non connaissance de ε_{n+1}
- l'autre due à l'estimation des paramètres.

Proposition 16. *(Variance de la prévision Y^p_{n+1})*

$$var(Y^p_{n+1}) = \sigma^2 \left(\frac{1}{n} + \frac{(X_{n+1} - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right).$$

$var(Y^p_{n+1})$ nous donne une idée de la stabilité de l'estimation. En prévision, on s'interesse généralement à l'erreur que l'on commet entre la vraie valeur à prévoir Y_{n+1} et celle que l'on prévoit Y^p_{n+1}. L'erreur peut être simplement résumée par la différence entre les deux valeurs : erreur de prévision. Cette erreur de prévision permet de quantifier la capacité du modèle à prévoir.

Proposition 17. *(Erreur de prévision)*
L'erreur de prévision définie par $\varepsilon^p_{n+1} = Y_{n+1} - Y^p_{n+1}$ satisfait les propriétés suivantes :

$$\mathbb{E}(\varepsilon^p_{n+1}) = 0$$

$$var(\varepsilon^p_{n+1}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(X_{n+1} - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right).$$

Remarque 2. *La variance augmente lorque X_{n+1} s'éloigne du centre de gravité du nuage de points. Effectuer une prévision lorsque X_{n+1} est "loin" de $\overline{X}$ est donc périleux, la variance de l'erreur de prévision peut être alors très grande.*

Proposition 18. *Un intervalle de confiance pour Y_{n+1} est donné par*

$$\left[Y_{n+1}^p \pm t_{1-\alpha/2}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(X_{n+1}-\overline{X}_n)^2}{\sum_{i=1}^n (X_i-\overline{X}_n)^2}} \right].$$

Cette formule exprime que plus le point à prévoir est éloigné de $\overline{X}$, plus la variance de la prévision et donc de l'intervalle de confiance seront grandes.

1.3 Validation du modèle

Il s'agit de vérifier les hypothèses formulées sur le modèle par l'analyse des résidus.

- Le résidu associé à l'individu i est défini par :

$$\hat{\varepsilon}_i = Y_i - \left(\hat{\beta}_0 + \sum_{j=1}^p \hat{\beta}_j X_{ij}\right)$$

  Le vecteur des résidus est $\hat{\varepsilon} = Y - \hat{Y}$ avec $\mathbb{E}(\hat{\varepsilon}) = 0$ et $var(\hat{\varepsilon}) = \sigma^2(I - P_X)$ où

$$P_X = X(X^TX)^{-1}X^T = (h_{ij})_{1\le i,j\le n}$$

- Le résidu standardisé associé à l'individu i est défini par :

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}.$$

- Le résidu studentisé associé à l'individu i est défini par :

$$t_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} = t_i\sqrt{\frac{n-p-2}{n-p-1-t_i^2}}.$$

Théorème 4. *Sous les hypothèse $(\mathscr{H}_1)$ et $(\mathscr{H}_3)$, et si la suppression de la ligne i ne modifie pas le rang de la matrice alors les résidus studentisés t_i^* suivent la loi de Student à $(n-p-2)$ dégrés de liberté.*

Définition 4. *Une donnée aberrante est un point (X_i^T, Y_i) pour lequel la valeur associée t_i^* est élevée :*

$$|t_i^*| > t_{n-p-1}(1-\alpha/2).$$

1.3.1 Analyse de la normalité

L'hypothèse de normalité sera examinée à l'aide d'un histogramme ou d'un graphique comparant les quantiles des résidus à ces mêmes quantiles sous l'hypothèse de normalité appelé QQ-plot. Si cette hypothèse est respectée, le graphique QQ-plot sera proche de la première bissectrice.

1.3.2 Indépendance, homoscédasticité et linéarité

Il est recommandé de tracer les résidus studentisés t_i^* en fonction des valeurs ajustées $\hat{Y}_i$ i.e. de tracer le nuage de points $(\hat{Y}_i, t_i^*)$. Si les points se retrouvent à l'intérieur d'un rectangle centré sur l'ordonnée nulle alors les hypothèses d'indépendance et de linéarité sont vérifiées. Si une structure apparaît (tendance, cone, vagues), l'hypothèse d'homoscédasticité risque fort de ne pas être vérifiée.

1.3.3 Individus extrêmes

La régression est sensible aux individus extrêmes. Ils peuvent considérablement influencer la valeur des paramètres de la régression.

Leverage ou effet levier

L'effet levier associé à l'individu i est défini par

$$h_{ii} = \frac{1}{n} + \frac{X_i - \bar{X}_n}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}.$$

h_{ii} permet de mesurer l'influence de l'observation i sur les estimations des coefficients. Il faudra faire une investigation pour individus pour lesquels $h_{ii} > \frac{2(p+1)}{n}$.

Distance de Cook

Définition 5. *La distance de Cook associé à l'observation i est donnée par*

$$D_{Cook_i} = \frac{t_i^2}{(p+1)} \frac{h_{ii}}{1-h_{ii}}$$

D_{Cook_i} mesure l'effet de la suppression de l'observation i sur l'estimation des coefficients. L'observation i est globalement influente si $D_{Cook_i} > 1$. Mais le seuil 1 est jugé un peu trop permissive laissant échapper à tort les points douteux. On lui préfère parfois la disposition plus exigeante suivante :

$$D_{Cook_i} > \frac{4}{n-p-1}.$$

DFFITS

L'écart de Welsh-Kuh, souvent appelé DFFITS dans les logiciels, est défini par

$$DFFITS_i = t_i^* \sqrt{\frac{h_{ii}}{1-h_{ii}}}$$

Cette quantité permet d'évaluer l'écart entre l'estimation bâtie sur toutes les observations et l'estimation bâtie sur toutes les observations sauf la i^{eme}. Nous considérons qu'une observation i est influente lorsque

$$|DFFITS_i| > 2\sqrt{\frac{p+1}{n}}.$$

DBFBETAS

La distance de Cook évalue globalement les disparités entre les coeffcients de la régression utilisant ou pas l'observation numéro i. Si l'écart est important, on peut vouloir approfondir l'analyse en essayant d'identifier la variable qui est à l'origine de l'écart : c'est le rôle des DFBETAS.
Pour chaque observation i et pour chaque coefficient β_j, $j = 0, \dots, p$, nous calculons la quantité :

$$DFBETAS_{j,i} = t_i^* \left[\frac{[(X^T X)^{-1} X^T]_{j,i}}{\sqrt{(X^T X)^{-1}_{j,j}(1-h_{ii})}} \right].$$

On considère que l'observation i pèse indûment sur la variable X_j lorsque

$$|DFBETAS_{j,i}| > 1.$$

Lorsque les observations sont nombreuses, on préférera la règle plus exigeante :

$$|DFBETAS_{j,i}| > \frac{2}{\sqrt{n}}.$$

COVRATIO

Le COVRATIO mesure les disparités entre les précisions des estimateurs i.e la variance des estimateurs :

$$COVRATIO_i = \frac{1}{\left[\frac{n-p-2}{n-p-1} + \frac{(t_i^*)^2}{n-p-1}\right]^{p+1}(1-h_{ii})}$$

$|COVRATIO_i - 1| > \frac{3(p+1)}{n}$ indique que la présence de l'observation i dégrade la variance. Sinon, la présence de l'observation i améliore la précision au sens où elle réduit la variance des estimateurs.

1.4 Colinéarité et sélection de variables

1.4.1 Colinéarité

Souvent certaines variables exogènes sont redondantes, elles emmènent le même type d'information : c'est le problème de la colinéarité, elles se gênent mutuellement dans la régression. On parle de colinéarité entre 2 variables exogènes lorsque la corrélation linéaire entre ces variables est élevée (ex. $r > 0.8$ a-t-on l'habitude d'indiquer mais ce n'est pas une règle absolue). On peut généraliser cette première définition en définissant la colinéarité comme la corrélation entre une des exogènes avec une combinaison linéaire des autres exogènes.
Essayons d'illustrer le mécanisme de la colinéarité.

- Si la colinéarité est parfaite alors $rang(X^T X) < p + 1$, ce qui implique que $(X^T X)^{-1}$ n'existe pas. Le calcul devient impossible.
- Si la colinéarité est forte, $det(X^T X) \sim 0$, $(X^T X)^{-1}$ contient des valeurs très élevées

Pour évaluer la multicolinéarité, il faudrait effectuer la régression de chaque exogène X_j avec les $(p-1)$ autres exogènes, puis étudier le coefficient de détermination R_j^2 associé. On appelle facteur d'inflation de la variance (VIF) la quantité :

$$v_j = \frac{1}{1-R_j^2}$$

On parle de facteur d'inflation car nous avons la relation suivante

$$var(\hat{\beta}_j) = \frac{\sigma^2}{n} v_j$$

Si C est la matrice des corrélations entre les exogènes, de taille $(p \times p)$, la quantité v_j peut être lue à la coordonnée j de la diagonale principale de la matrice inversée C^{-1}.
Plus v_j sera élevé, plus la variance $var(\hat{\beta}_j)$ de l'estimation sera forte. L'estimateur $\hat{\beta}_j$ sera donc très instable, il aura moins de chances d'être significatif dans le test de nullité du coefficient dans la régression. On décide qu'il y a un problème de colinéarité lorsque $v_j \leq 4$.
La quantité $1-R_j^2$, appelée tolérance, est également fournie par les logiciels statistiques. Plus elle est faible, plus la variable X_j souffre de colinéarité. En dérivant la règle de détection du VIF, on s'inquiéterait dès que la tolérance est inférieure à 0.25

1.4.2 Sélection des variables

La sélection de variables permet le traitement de la colinéarité. L'objectif est de trouver un sous-ensemble de q variables exogènes $(q \leq p)$ qui soient, autant que possible, pertinentes et non-redondantes pour expliquer l'endogène Y.

1.5 Etude de cas avec le logiciel R

Dans cet exemple, la variable à expliquer est la consommation des véhicules. Nous avons 4 variables explicatives : le prix, la cylindrée, la puissance et le poids. Les objectifs sont les suivants :

- Etude de la relation linéaire entre la consommation de véhicules et ses caractéristiques que sont le prix, la cylindrée, la puissance et le poids.
- Diagnostic de la régression avec les graphiques des résidus.
- Prévision

Nous utilisons la commande **read.table()** pour importer les données dans le logiciel à partir du fichier texte (.txt).

```
> #Importation des donnees
> #
> donnees<-read.table("conso_veh.txt",header=TRUE)
> #
> #pour afficher les donnees
> donnees
                   modele   prix cylindree puissance poids consom
1           Daihatsu.Cuore  11600       846        32   650    5.7
2     Suzuki.Swift.1.0.GLS  12490       993        39   790    5.8
3       Fiat.Panda.Mambo.L  10450       899        29   730    6.1
4           VW.Polo.1.4.60  17140      1390        44   955    6.5
5      Opel.Corsa.1.2i.Eco  14825      1195        33   895    6.8
6         Subaru.Vivio.4WD  13730       658        32   740    6.8
7           Toyota.Corolla  19490      1331        55  1010    7.1
8           Ferrari.456.GT 285000      5474       325  1690   21.3
9           Mercedes.S.600 183900      5987       300  2250   18.7
10      Maserati.Ghibli.GT  92500      2789       209  1485   14.5
11     Opel.Astra.1.6i.16V  25000      1597        74  1080    7.4
12      Peugeot.306.XS.108  22350      1761        74  1100    9.0
```

```
13   Renault.Safrane.2.2.V  36600      2165       101  1500   11.7
14      Seat.Ibiza.2.0.GTI  22500      1983        85  1075    9.5
15         VW.Golt.2.0.GTI  31580      1984        85  1155    9.5
16      Citroen.ZX.Volcane  28750      1998        89  1140    8.8
17 Fiat.Tempra.1.6.Liberty  22600      1580        65  1080    9.3
18     Fort.Escort.1.4i.PT  20300      1390        54  1110    8.6
19   Honda.Civic.Joker.1.4  19900      1396        66  1140    7.7
20           Volvo.850.2.5  39800      2435       106  1370   10.8
21   Ford.Fiesta.1.2.Zetec  19740      1242        55   940    6.6
22     Hyundai.Sonata.3000  38990      2972       107  1400   11.7
23         Lancia.K.3.0.LS  50800      2958       150  1550   11.9
24       Mazda.Hachtback.V  36200      2497       122  1330   10.8
25       Mitsubishi.Galant  31990      1998        66  1300    7.6
26       Opel.Omega.2.5iV6  47700      2496       125  1670   11.3
27         Peugeot.806.2.0  36950      1998        89  1560   10.8
28      Nissan.Primera.2.0  26950      1997        92  1240    9.2
29       Seat.Alhambra.2.0  36400      1984        85  1635   11.6
30     Toyota.Previa.salon  50900      2438        97  1800   12.8
31     Volvo.960.Kombi.aut  49300      2473       125  1570   12.7
>
```

La commande **lm()** permet de faire la régression.

```
> #regression lineaire
> regression<-lm(consom~prix+cylindree+puissance+poids,data=donnees)
> #
> #resultats de la regression
> resultats<-summary(regression)
> resultats

Call:
lm(formula = consom ~ prix + cylindree + puissance + poids, data = donnees)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5677 -0.6704  0.1183  0.5283  1.4361

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.456e+00  6.268e-01   3.919 0.000578 ***
prix         2.042e-05  8.731e-06   2.339 0.027297 *
cylindree   -5.006e-04  5.748e-04  -0.871 0.391797
puissance    2.499e-02  9.992e-03   2.501 0.018993 *
poids        4.161e-03  8.788e-04   4.734 6.77e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8172 on 26 degrees of freedom
Multiple R-squared:  0.9546,    Adjusted R-squared:  0.9476
F-statistic: 136.5 on 4 and 26 DF,  p-value: < 2.2e-16

>
```

La commande **confint** fournit les intervalles de confiance de chaque paramètre.

```
> #intervalle de confiance
> IC<-confint(regression,level=0.95)
> IC
                    2.5 %       97.5 %
(Intercept)  1.167851e+00 3.744737e+00
prix         2.474392e-06 3.836669e-05
cylindree   -1.682157e-03 6.809703e-04
puissance    4.455929e-03 4.553302e-02
poids        2.354210e-03 5.966955e-03
```

La fonction **predict()** permet de donner les prévisions mais aussi les intervalles de confiance tant pour le modèle que pour pour les prévisions.

```
> #Prevision pour Z1=(14000,800,35,700) Z2=(170000,6000,270,1870)
 IC pour Y et E(Y)
> prix=c(14000,170000)
> cylindree=c(800,670)
> puissance=c(35,270)
> poids=c(700,1870)
> nouv.donnees=data.frame(prix,cylindree,puissance,poids)
> nouv.donnees
    prix cylindree puissance poids
1  14000       800        35   700
2 170000       670       270  1870
> ICdte=predict.lm(regression,nouv.donnees,interval="confidence")
> ICdte
        fit       lwr       upr
1  6.128921  5.528698  6.729145
2 20.121187 15.197486 25.044887
> ICpred=predict.lm(regression,nouv.donnees,interval="prediction")
> ICpred
        fit       lwr       upr
1  6.128921  4.345052  7.912791
2 20.121187 14.918808 25.323565
>

> ######GRAPHIQUES des residus
> par(mfrow=c(2,3))  #subdviser la page en 6 parties
> plot(regression,which=1,sub="",main="")
> plot(regression,which=2,sub="",main="")
> plot(regression,which=3,sub="",main="")
> plot(regression,which=4,sub="",main="")
> plot(regression,which=5,sub="",main="")
> plot(regression,which=6,sub="",main="")
>
```

FIGURE 1.1 –

1.6 Exercices

Exercise 1. *La masse monétaire et le revenu national brut (en milliards de francs CFA) de la Côte d'Ivoire sont reproduits dans le tableau ci-dessous (source : Banque mondiale).*

Année	*Masse Monétaire*	*Revenu national brut*
2000	*1646.26*	*6289.63*
2001	*1840.12*	*6386.73*
2002	*2401.83*	*6306.89*
2003	*1759.82*	*6245.24*
2004	*1932.57*	*6403.78*
2005	*2080.94*	*6495.91*
2006	*2294.76*	*6523.78*
2007	*2836.59*	*6625.53*
2008	*2997.35*	*6781.55*
2009	*3511.75*	*7025.49*
2010	*4152.21*	*7194.92*
2011	*4595.55*	*6856.40*

Établir une relation linéaire dans laquelle la masse monétaire explique explique le revenu national brut.

Solution de l'exercice 1. *1.* **Existence d'une relation linéaire.** *Pour vérifier l'existence d'une liaison linéaire entre Y et X, nous pouvons utiliser le nuage de points ou le coefficient de correlation linéaire.*
Le coefficient de correlation linéaire ent Y et X_1 est donné par la formule :

$$\rho = \frac{\sum_{i=1}^{12} Y_i X_i - 12\bar{Y}\bar{X}}{\sqrt{\sum_{i=1}^{12} Y_i^2 - 12\bar{Y}^2}\sqrt{\sum_{i=1}^{12} X_i^2 - 12\bar{X}^2}}$$

où

$$\bar{Y} = \frac{1}{12}\sum_{i=1}^{12} Y_i \quad \bar{X} = \frac{1}{12}\sum_{i=1}^{12} X_i.$$

```
> X=c(1646.26,1840.12,2401.83,1759.82,1932.57,2080.94,2294.76,2836.59,2997.35,3511.75,
+     4152.31,4595.55)
> Y=c(6289.63,6386.73,6306.89,6245.24,6403.78,6495.91,6523.78,6625.53,6781.55,
+     7025.49,7194.92,6856.40)
> Donnees1=data.frame(X,Y)
> Donnees1

         X       Y
1  1646.26 6289.63
2  1840.12 6386.73
3  2401.83 6306.89
4  1759.82 6245.24
5  1932.57 6403.78
6  2080.94 6495.91
7  2294.76 6523.78
8  2836.59 6625.53
```

```
9  2997.35 6781.55
10 3511.75 7025.49
11 4152.31 7194.92
12 4595.55 6856.40
```

Le coefficient de correlation entre la masse monétaire et le revenu national brut est :

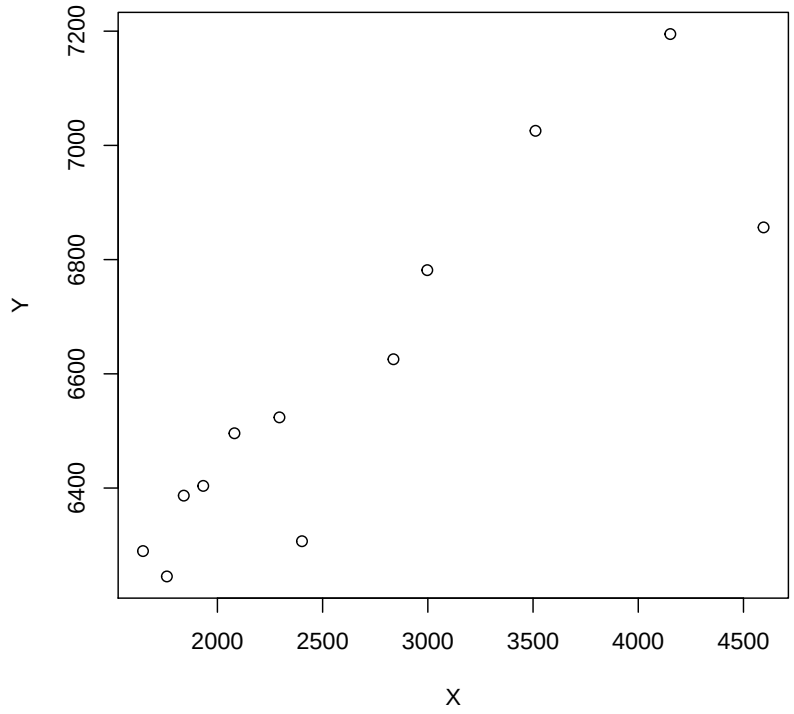
```
> cor(X,Y)
```

```
[1] 0.888505
```

Le coefficient de correlation est proche de 1. Il existe donc une liaison linéaire entre X et Y. De plus, il est positif; cela signifie que la masse monétaire et le revenu national brut évoluent dans le même sens.

Le nuage de points

```
> plot(X,Y)
```



La forme du nuage de points confirme la liaison linéaire entre Y et X.

2. ***Modèle de régression linéaire simple.*** *Le modèle s'écrit :*

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, \ldots, 12$$

où Y_i est le revenu national brut, X_i est la masse monétaire de l'année i et ε_i est le residu ou aléa.

3. **Hypothèses sur l'aléa.** *On admettra que les variables alátoires $\varepsilon_1,\ldots,\varepsilon_{12}$ sont indépendantes identiquement distribuées de loi normale $\mathcal{N}(0,\sigma^2)$ avec $\sigma^2>0$.*

4. *Estimation des coefficients. Les estimateurs de β_1, β_0 et σ^2 sont respectivement donnés par :*

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{12} Y_i X_i - 12\bar{Y}\bar{X}}{\sum_{i=1}^{12} X_i^2 - 12\bar{X}^2}$$

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

$$\widehat{\sigma}^2 = \frac{1}{10}\sum_{i=1}^{12} \widehat{\varepsilon}_i^2 \quad \text{avec } \widehat{\varepsilon}_i = Y_i - \widehat{Y}_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i).$$

Avec le logiciel R, nous obtenons

```
> reg1=lm(Y~X) # regression lineaire de Y sur X
> summary(reg1) # resultats de la regression lineaire de Y sur X

Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-279.30  -36.09   21.12   74.08  194.45

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.844e+03  1.299e+02  45.001 7.06e-13 ***
X           2.811e-01  4.591e-02   6.123 0.000112 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 148.2 on 10 degrees of freedom
Multiple R-squared:  0.7894,        Adjusted R-squared:  0.7684
F-statistic: 37.49 on 1 and 10 DF,  p-value: 0.0001122
```

Analyse des sorties de logiciel : Nous obtenons $\widehat{\beta}_0 = 5844$, $\widehat{\beta}_1 = 0.2811$ et $\widehat{\sigma} = 148.2$.

5. *Intervalles de confiance.*

```
> IC<-confint(reg1,level = 0.95)
> IC

                  2.5 %       97.5 %
(Intercept) 5554.5270560 6133.2215180
X              0.1788137    0.3833954
```

Analyse des sorties de logiciel : au niveau de confiance 0.95, un intervalle de confiance pour β_1 est $[0.18, 0.38]$

6. **Au niveau $\alpha = 0.05$, la masse monétaire (X) a-t-elle de l'influence sur le revenu national brut ?**

Pour répondre à cette question, nous tester l'hypothèse $H_0 : \beta_1 = 0$ contre $H_1 : \beta_1 \neq 0$ au niveau $\alpha = 0.05$. D'après les sorties de logiciel, la p-value de ce test est $\alpha^* =$

$0.000112 < 0.05$. L'hypothèse H_1 est acceptée, le coefficient β_1 est alors significatif, c'est à dire au niveau $\alpha = 0.05$, la masse monétaire (X) a de l'influence sur le revenu national brut.

Exercise 2. *Les données proposées dans cet exercice présentent le taux de décès par attaque cardiaque ches les hommes de* 55 *à* 59 *ans dans différents pays. Les variables sont les suivantes :*

- $Y : 100$ *[log(nombre de décès par crise cardiaque pour* 100000 *hommes de* 55 *à* 59 *ans)*-2*] ;*
- X_1 *: téléphones par millier d'habitants ;*
- X_2 *: calories grasses en pourcentage du total des calories ;*
- X_3 *: calories provenant de protéines animales en pourcentage du total des calories.*

On veut étudier l'impact des variables X_1*,* X_2 *et* X_3 *sur la variable* Y*. A l'aide du logiciel R, nous obtenons les résultats suivants :*

1. *Ecrire le modèle de régression linéaire multiple avec les hypothèses appropriées.*
2. *Donner les limites des intervalles de confiance à* 95% *.*

Observation i	*Pays*	X_1 x_{i1}	X_2 x_{2i}	X_3 x_{3i}	Y y_i
1	*Australie*	*124*	*33*	*8*	*81*
2	*Autriche*	*49*	*31*	*6*	*55*
3	*Canada*	*181*	*38*	*8*	*80*
4	*Ceylan*	*4*	*17*	*2*	*24*
5	*Chili*	*22*	*20*	*4*	*71*
6	*Danemark*	*152*	*39*	*6*	*52*
7	*Finlande*	*75*	*30*	*7*	*88*
8	*France*	*54*	*29*	*7*	*45*
9	*Allemagne*	*43*	*35*	*6*	*50*
10	*Irlande*	*41*	*31*	*5*	*69*
11	*Israel*	*17*	*23*	*4*	*66*
12	*Italie*	*22*	*21*	*3*	*45*
13	*Japon*	*16*	*8*	*3*	*24*
14	*Mexique*	*10*	*23*	*3*	*43*
15	*Pays-Bas*	*63*	*37*	*6*	*38*
16	*Nouvelle-Zélande*	*170*	*40*	*8*	*72*
17	*Norvège*	*125*	*38*	*6*	*41*
18	*Portugal*	*12*	*25*	*4*	*38*
19	*Suède*	*221*	*39*	*7*	*52*
20	*Suisse*	*171*	*33*	*7*	*52*
21	*Grande-Bretagne*	*97*	*38*	*6*	*66*
22	*États-Unis*	*254*	*39*	*8*	*89*

```
> pays=c("australie","autriche","canada","ceylan","chili","danemark",
+        "finlande","france","allemagne","irlande","israel","italie","japon","mexique","pays-bas"
+        "nouvelle-zelande","norvege","portugal","suede",
+        "suisse","grande-gretagne","etats-unis")
> x1=c(124,49,181,4,22,152,75,54,43,41,17,22,16,10,63,170,125,12,221,171,97,254)
> x2=c(33,31,38,17,20,39,30,29,35,31,23,21,8,23,37,40,38,25,39,33,38,39)
> #Variable x3
```

```
> x3=c(8,6,8,2,4,6,7,7,6,5,4,3,3,3,6,8,6,4,7,7,6,8)
> #Variable y
> y=c(81,55,80,24,71,52,88,45,50,69,66,45,24,43,38,72,41,38,52,52,66,89)
>  Donnees=data.frame(x1,x2,x3,y,row.names=pays)
>  Donnees

                  x1 x2 x3  y
australie        124 33  8 81
autriche          49 31  6 55
canada           181 38  8 80
ceylan             4 17  2 24
chili             22 20  4 71
danemark         152 39  6 52
finlande          75 30  7 88
france            54 29  7 45
allemagne         43 35  6 50
irlande           41 31  5 69
israel            17 23  4 66
italie            22 21  3 45
japon             16  8  3 24
mexique           10 23  3 43
pays-bas          63 37  6 38
nouvelle-zelande 170 40  8 72
norvege          125 38  6 41
portugal          12 25  4 38
suede            221 39  7 52
suisse           171 33  7 52
grande-gretagne   97 38  6 66
etats-unis       254 39  8 89
```

Solution de l'exercice 2. *1. Le modèle s'écrit :*

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_1 X_{i2} + \varepsilon_i \quad i = 1, \ldots, 22.$$

Nous supposons vérifier les hypothèses $(\mathscr{H}_0)$, $(\mathscr{H}_1)$, *et* $(\mathscr{H}_3)$ *(puisqu'il ya des intervalles de confiance et des tests d'hypothèses à faire !)*

```
> reg0=lm(y~x1,data = Donnees)
>  summary(reg0)

Call:
lm(formula = y ~ x1, data = Donnees)

Residuals:
    Min      1Q  Median      3Q     Max
-23.566 -14.173  -2.109  11.991  33.128

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 45.58436    5.59695   8.145 8.83e-08 ***
x1           0.12384    0.04892   2.532   0.0198 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 16.94 on 20 degrees of freedom
Multiple R-squared:  0.2427,        Adjusted R-squared:  0.2048
F-statistic: 6.409 on 1 and 20 DF,  p-value: 0.01985
```

```
> reg2=lm(y~x1+x2,data = Donnees)
>  summary(reg2)

Call:
lm(formula = y ~ x1 + x2, data = Donnees)

Residuals:
    Min      1Q  Median      3Q     Max
-22.348 -13.583  -2.483  14.310  32.731

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.81276   16.00166   2.113   0.0481 *
x1           0.07858    0.07585   1.036   0.3132
x2           0.51876    0.65973   0.786   0.4414
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.1 on 19 degrees of freedom
Multiple R-squared:  0.2665,        Adjusted R-squared:  0.1893
F-statistic: 3.452 on 2 and 19 DF,  p-value: 0.05261
```

2. *Le modèle s'écrit :*

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i \quad i = 1, \ldots, 22.$$

```
> reg3=lm(y~x1+x2+x3,data = Donnees)
>  summary(reg3)

Call:
lm(formula = y ~ x1 + x2 + x3, data = Donnees)

Residuals:
    Min      1Q  Median      3Q     Max
-23.636 -10.553  -1.559   9.261  23.601

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.353115  15.328270   1.458    0.162
x1          -0.005842   0.077988  -0.075    0.941
x2          -0.423988   0.726211  -0.584    0.567
x3           8.413437   3.689602   2.280    0.035 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.48 on 18 degrees of freedom
Multiple R-squared:  0.4309,        Adjusted R-squared:  0.3361
F-statistic: 4.544 on 3 and 18 DF,  p-value: 0.01538
```

3. Intervalles de confiance

```
> IC2=confint(reg3,level = 0.95)
>  IC2

                2.5 %     97.5 %
(Intercept) -9.8503847 54.5566151
x1          -0.1696885  0.1580038
x2          -1.9496995  1.1017240
x3           0.6618697 16.1650034
```

Exercise 3. *Dans le cadre de travaux de recherche sur la durée de la saison de végétation en montagne, des stations météorologiques sont installées à différentes altitudes. La température moyenne ainsi que l'altitude (en mètres) de chaque saison sont relevées et données dans le tableau ci-dessous :*

Altitude	*1040*	*1230*	*1500*	*1600*	*1740*	*1950*	*2200*	*2530*	*2800*	*3100*
Température	*7.4*	*6*	*4.5*	*3.8*	*2.9*	*1.9*	*1*	*-1.2*	*-1.5*	*-4.5*

A partir de l'altitude d'un lieu, on cherche à évaluer sa température moyenne sans avoir implanter une nouvelle station.

1. *Expliquer en quoi la méthode de régression linéaire est adaptée à cette problématique. Préciser le modèle approprié.*
2. *Formuler les hypothèses nécessaires à cette analyse.*
3. *Calculer les estimations des paramètres du modèle.*
4. *Faire le test de pertinence permettant de vérifier que la pente de la droite de régression est non nulle au risque de* 5%.
5. *On suppose que les hypothèses du modèle sont toutes vérifiées. Sachant qu'une certaine plante ne survit qu'à une température moyenne supérieure à* $-6°$ *C, est-il raisonnable de penser que l'on ne trouvera pas cette plante à une altitude de* 3500 *mètres ?*

Chapitre

2 Analyse de la variance (ANOVA) à effets fixes

2.1 Introduction

L'analyse de la variance à k facteurs permet d'étudier le lien entre une variable quantitative continue et k variables qualitatives appelées facteurs. L'analyse de la variance est aussi considérérée comme une généralisation du test de Student : elle permet de voir si la moyenne d'une variable quantitative est la même dans différents groupes.

Tout dispositif expérimental comportant un nombre identique de répétitions dans chacune des modalités des facteurs est un plan équilibré. Nous supposons dans ce cours que le plan est équilibré.

2.2 Anova à un facteur

On considère un facteur contrôlé Xprésentant I modalités, chacune d'entre elles étant notée X_i. Pour chacune des modalités, nous effectuons $J \geq 2$ mesures d'une réponse Y. Ainsi, le nombre total d'observations est $n = I \times J$.

2.2.1 Notations

Notons

- Y_{ij} l'observation j de la modalité i du facteur X
- ε_{ij} le terme d'erreur sur l'observation j de la modalité i du facteur X.
- α_i : l'effet de la modalité i du facteur X.
- $\overline{Y}_{i} = \frac{1}{J}\sum_{j=1}^{J} Y_{ij}$: la moyenne de la modalité i du caractère X
- $\overline{Y} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}$: la moyenne sur les n observations

Remarque 3. *On parle d'effets fixes si les α_i sont fixés. Si les α_i sont des variables aléatoires (suivant une loi normale !), on parle d'effets aléatoires.*

2.2.2 Modèle d'Anova à un facteur

Le modèle d'analyse de la variance s'écrit :

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1, \cdots, I, \quad j = 1, \cdots, J.$$

2.2.3 Hypothèses

Nous postulons les hypothèses suivantes :

- $\sum_{i=1}^{I} \alpha_i = 0.$
- Pour $(i,j) \neq (k,l)$, ε_{ij} et ε_{kl} sont indépendantes.
- Pour tout (i,j), $1 \leq i \leq I$, $1 \leq j \leq J$, ε_{ij} suit une loi normale $\mathcal{N}(0, \sigma^2)$

2.2.4 Tableau d'ANOVA

Nous avons trois sources de variation :

- Variation due au facteur
$$SC_F = J \sum_{i=1}^{I} \left(\overline{Y}_i - \overline{Y}\right)^2.$$
- Variation résiduelle :
$$SCR = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(Y_{ij} - \overline{Y}_i\right)^2.$$
- Variation totale
$$SC_{TOT} = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(Y_{ij} - \overline{Y}\right)^2.$$

La relation faondamentale de l'ANOVA appelée équation de l'analyse de la variance est :

$$SC_{TOT} = SC_F + SC_R.$$

Voici le tableau de l'ANOVA :

Source	Variation	Degrés de liberté	Carré moyen	F
Facteur	SC_F	$I-1$	$s_F^2 = \frac{SC_F}{I-1}$	$F = \frac{s_F^2}{s_R^2}$
Résiduelle	SC_R	$n-I = I(J-1)$	$s_R^2 = \frac{SC_R}{n-1}$	
Totale	SC_{TOT}	$n-1 = I \times J - 1$		

2.2.5 Test d'hypothèses

Nous souhaitons tester

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_I = 0 \quad \textbf{(Absence d'effet du facteur } X\textbf{)}$$

contre

$$H_1 : \text{il existe } i \in \{1, \ldots, I\} \text{ tel que } \alpha_i = 0 \quad \textbf{(Absence d'effet du facteur } X\textbf{)}.$$

La région critique du test est :

$$W = \left\{ F > f^* \right\}$$

où f^* est le quantile d'ordre $1-\alpha$ de la loi de Fisher à $I-1$ et $n-I$ degrés de liberté.

Nous pouvons aussi utiliser la p-value α^* pour conclure :

$\alpha^* < \alpha \Rightarrow$ on accepte H_1.

$\alpha^* > \alpha \Rightarrow$ on accepte H_0

2.2.6 Comparaisons multiples

Lorsque l'hypothèse de Présence d'effet (H_1) est acceptée, on procède à des comparaisons multiples pour detecter les groupes qui sont significativement différents. Dans ce cours, nous utilisons le test de Tukey.

Méthode : pour chaque paire i et j, on détermine un intervalle de confinance de niveau $1-\alpha$ pour $\mu_i - \mu_j$ où $\mu_i = \mu + \alpha_i$ et $\mu_j = \mu + \alpha_j$. Si zéro appartient à l' intervalle de confiance, les moyennes ne sont pas jugées significativement différentes au niveau $1-\alpha$.

2.2.7 Exercice

Exercise 4. *Nous souhaitons comparer la moyenne des notes en mathématique de trois classes, notées A, B, C. Pour chaque classe, nous choisissons un échantillon d'étudiants aléatoirement et nous relevons leurs notes. Les notes sont reportées dans le tableau ci-dessous :*

Classe A	*Classe B*	*Classe C*
2 ;6 ;11 ;12 ;10 ;15 ;7 ;5 ;6 ;11	*13 ;15 ;14 ;17 ;16 ;14 ;4 ;19 ;16 ;13*	*13 ;17 ;17 ;13 ;10 ;15 ;12 ;15 ;13 ;20*

1. *Quand et pourquoi faut-il faire une analyse de la variance ?*
2. *Donner le modèle statistique de l'analyse de la variance à un facteur à effets fixes.*
3. *Quelles sont les conditions fondamentales d'application du modèle linéaire ?*
4. *Donner le tableau d'ANOVA en expliquant brièvement comment il est construit.*
5. *Quelle est la valeur observée de la statistique associée au test de Fisher ?*
6. *Qu'en déduisez-vous au seuil $\alpha = 5\%$? Justifier vos réponses.*
7. *Décrire un modèle d'analyse de la variance à deux facteurs avec répétitions. Proposer un facteur dans le modèle ci-dessus pour le transformer en un modèle à deux facteurs. Quels sont les problèmes de tests d'hypothèses vous aurez alors à résoudre ?*

Solution.

Modélisation

```
> A=c(2,6,11,12,10,15,7,5,6,11 )#notes de la classe A
> B=c(13,15,14,17,16,14,4,19,16,13)#notes de la classe B
> C=c(13,17,17,13,10,15,12,15,13,20 ) #notes de la classe C
> Y=c(A,B,C)# les notes des 3 classes reunies (caractere quantitatif)
> X=c(rep("A",10),rep("B",10),rep("C",10))# les 3 classes
> X=as.factor(X) # X transformé en caractere qualitatif
```

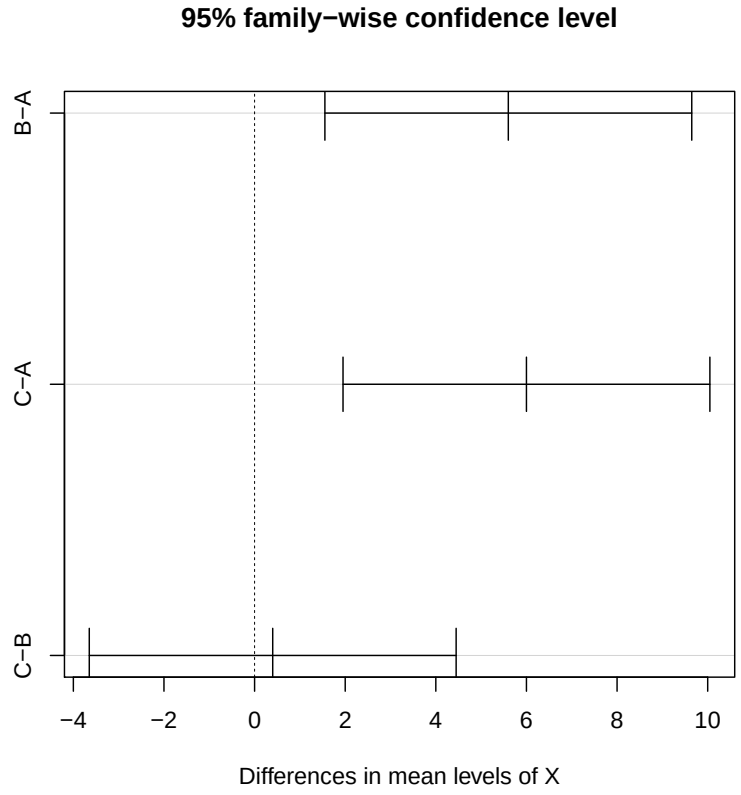
Commandes et résultats avec R

```
> anova=aov(Y~X) # Anova de Y sur X
> summary(anova) # resultats de Anova

            Df Sum Sq Mean Sq F value  Pr(>F)
X            2  225.1  112.53   8.442 0.00142 **
Residuals   27  359.9   13.33
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **Df :** *Degrés de liberté*
- **Sum Sq :** *Somme des carrés*
- **Mean Sq :** *Somme descarrés moyen*
- **F-value :** *valeur de* F
- **Pr(>F) :** *la p-value*

Comparaisons multiples

```
> plot(TukeyHSD(anova,"X"))# Comparaison multiple de Tuckey
```



Exercise 5. *Dans cet exercice, nous étudions le lien entre la force de retrait* $\mathbf{Y}$ *(variable quantitative) et le type de clous* $\mathbf{X}$ *(variable qualitative).*

1. *Pourquoi faut-il faire une analyse de la variance ?*
2. *Ecrire le modèle d'analyse de la variance à un facteur à effets fixes.*

3. *Quelles sont les conditions d'utilisation du modèle ?*

4. *Interprèter les commandes ci-dessous*

```
> Y=c(36.57,29.67,43.38,26.94,12.03,21.66,41.79,31.5,
+     35.84,40.81,14.66,24.22,23.83,21.8,27.22,38.25,28.15,36.35,
+     23.89,28.44,12.61,25.71,17.69,24.69,26.48,19.35,28.6,42.17,25.11,19.38)
> X=as.factor(c(rep("annulaire",10),rep("helicoidal",10),rep("lisse",10)))
```

5. *Complèter le tableau suivant*

Source de la variance	*Sommes des carrés des écarts*	*Degrés de liberté*	*Carrés moyens*	***F***
X	*?*	*?*	*?*	*?*
Résiduelle	*1873.1*	*?*	*?*	
Total	*2193.8*	*?*	*?*	

6. *Au niveau* 5%*, le type de clous a-t-il un effet sur la force de retrait ?*
7. *Peut-on faire ici un test de comparaisons multiples ? Pourquoi ?*
8. *Quelle est la commande qui permet de faire le test de comparaisons multiples de Tuckey ?*

Exercise 6. *Le transport d'animaux d'élevage implique une succession de manipulations et de confinements qui, inévitablement, sont responsables de stress. L'objectif de cette étude est d'examiner le comportement des animaux lors de transports de longue durée. Pour mesurer le niveau de stress, nous avons calculé le pourcentage de temps que les animaux passent couchés. La fatigue des animaux (et donc le stress) est d'autant plus grande que le temps passé couché est important. L'expérimentation a consisté à observer 18 veaux transportés de France en Italie, la durée du voyage étant de 29 heures. Trois traitements ont été proposés durant la pause :*

- *Traitement 1 : les veaux ne reçoivent ni eau ni aliment à la pause.*
- *Traitement 2 : les veaux reçoivent de l'eau et sont alimentés à la pause par deux abreuvoirs.*
- *Traitement 3 : les veaux reçoivent de l'eau et sont alimentés à la pause par cinq abreuvoirs.*

Pour enregistrer les comportements des animaux, la bétaillère est équipée de cameras et chaque animal est individualisé par un signe distinctif sur la peau. Les cassettes vidéo ont été dépouillées à l'aide d'un logiciel. Les données sont disponibles dans le tableau ci-dessous.

Traitement 1	*17.40*	*20.00*	*26.70*	*31.70*	*35.80*	*47.80*
Traitement 2	*14.65*	*37.22*	*37.73*	*43,61*	*46.07*	*47.40*
Traitement 3	*18.76*	*19.49*	*27.19*	*45.42*	*53.20*	*61.27*

1. *Pourquoi faut-il faire ici une analyse de la variance à un facteur et non pas une analyse de la régression linéaire simple ?*
2. *Écrire le modèle statistique de l'analyse de la variance à un facteur à effets fixes.*

3. *Quelles sont les conditions d'utilisation du modèle d'analyse de la variance précédent ? Sont-elles vérifiées ?*
4. *Donner les commandes du logiciel R permettant de réaliser l'ANOVA correspondant à cette étude.*
5. *Donner le tableau de l'ANOVA correspondant à cette étude. Faire les calculs.*
6. *Réaliser le test de Fisher au seuil de significativité* 5% *puis de* 1%. *Qu'est-il possible d'en déduire ?*
7. *Dans le cas de cette étude, est-il possible de procéder à des comparaisons multiples ? Pourquoi ? Si oui, réaliser alors ces comparaisons.*

2.3 Anova à deux facteurs

On considère un facteur contrôlé X présentant I modalités, chacune d'entre elles étant notée X_i et un facteur contrôlé Z présentant J modalités, chacune d'entre elles étant notée Z_j. Nous effectuons $K \geq 2$ mesures d'une réponse Y. Le plan étant équilibré, nous disposons de $n = I \times J \times K$ observations.

2.3.1 Notations

Notons

- Y_{ijk} l'observation k de la modalité (X_i, Z_j)
- ε_{ijk} le terme d'erreur sur l'observation k de la modalité de la modalité (X_i, Z_j).
- α_i : l'effet propre de la modalité i du facteur X.
- β_j : l'effet propre de la modalité j du facteur Z.
- $\overline{Y}_{i,\bullet,\bullet} = \dfrac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk}$: la moyenne de Y dans la modalité i du caractère X.
- $\overline{Y}_{\bullet,j,\bullet} = \dfrac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K} Y_{ijk}$: la moyenne de Y dans la modalité j du caractère Z.
- $\overline{Y}_{i,j,\bullet} = \dfrac{1}{K}\sum_{k=1} KY_{ijk}$: la moyenne de Y dans la modalité (X_i, Z_j).
- $\overline{Y} = \dfrac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} Y_{ijk}$: la moyenne sur les n observations

2.3.2 Modèle

Le modèle d'analyse de la variance s'écrit :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{i,j} + \varepsilon_{ijk} \quad i = 1, \cdots, I, \quad j = 1, \cdots, J \quad k = 1, \ldots, K.$$

2.3.3 Hypothèses

Nous postulons les hypothèses suivantes :

- $\sum_{i=1}^{I} \alpha_i = 0.$
- $\sum_{j=1}^{J} \beta_j = 0.$

- $\sum_{i=1}^{I}(\alpha\beta)_{i,j}=0$ pour tout $j\in\{1,\ldots,J\}$
- $\sum_{j=1}^{J}(\alpha\beta)_{i,j}=0$ pour tout $i\in\{1,\ldots,I\}$
- Pour $(i,j,k)\neq(l,m,n)$, ε_{ijk} et ε_{lmn} sont indépendantes.
- Pour tout (i,j,k), $1\leq i\leq I$, $1\leq j\leq J$, $1\leq k\leq K$, ε_{ijk} suit une loi normale $\mathcal{N}(0,\sigma^2)$

2.3.4 Tableau d'ANOVA

Nous avons trois sources de variation :

- Variation due au facteur X

$$SC_X = JK\sum_{i=1}^{I}\left(\overline{Y}_{i,\bullet,\bullet}-\overline{Y}\right)^2.$$

- Variation due au facteur Z

$$SC_Z = IK\sum_{j=1}^{J}\left(\overline{Y}_{\bullet,j,\bullet}-\overline{Y}\right)^2.$$

- Variation due à l'interaction des facteurs X et Z

$$SC_{XZ} = K\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\overline{Y}_{i,j,\bullet}-\overline{Y}_{i,\bullet,\bullet}-\overline{Y}_{\bullet,j,\bullet}+\overline{Y}\right)^2.$$

- Variation résiduelle :

$$SCR = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(Y_{ijk}-\overline{Y}_{ij\bullet}\right)^2.$$

- Variation totale

$$SC_{TOT} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(Y_{ijk}-\overline{Y}\right)^2.$$

La relation faondamentale de l'ANOVA appelée équation de l'analyse de la variance est :

$$SC_{TOT} = SC_X + SC_Z + SC_{XZ} + SC_R.$$

Voici le tableau de l'ANOVA :

Source	Variation	Degrés de liberté	Carré moyen	F
Facteur X	SC_X	$I-1$	$s_X^2=\frac{SC_X}{I-1}$	$F_X=\frac{s_X^2}{s_R^2}$
Facteur Z	SC_Z	$J-1$	$s_Z^2=\frac{SC_Z}{J-1}$	$F_Z=\frac{s_Z^2}{s_R^2}$
Interaction	SC_{XZ}	$(I-1)(J-1)$	$s_{XZ}^2=\frac{SC_{XZ}}{(I-1)(J-1)}$	$F_{XZ}=\frac{s_{XZ}^2}{s_R^2}$
Résiduelle	SC_R	$IJ(K-1)$	$s_R^2=\frac{SC_R}{n-1}$	
Totale	SC_{TOT}	$n-1$		

2.3.5 Test d'hypothèses

Dans ce cas, nous avons trois problèmes de test.

2.3.5.1 Tester l'effet du facteur X

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_I = 0 \quad \textbf{(Absence d'effet du facteur } X\textbf{)}$$

contre

$$H_1 : \text{il existe } i \in \{1,\ldots,I\} \text{ tel que } \alpha_i \neq 0 \quad \textbf{(Présence d'effet du facteur } X\textbf{)}.$$

La région critique du test est :

$$W = \left\{ F_X > f_X^* \right\}$$

où f_X^* est le quantile d'ordre $1-\alpha$ de la loi de Fisher à $I-1$ et $IJ(K-1)$ degrés de liberté.

2.3.5.2 Tester l'effet du facteur Z

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_J = 0 \quad \textbf{(Absence d'effet du facteur } Z\textbf{)}$$

contre

$$H_1 : \text{il existe } j \in \{1,\ldots,J\} \text{ tel que } \beta_j \neq 0 \quad \textbf{(Présence d'effet du facteur } Z\textbf{)}.$$

La région critique du test est :

$$W = \left\{ F_Z > f_Z^* \right\}$$

où f_Z^* est le quantile d'ordre $1-\alpha$ de la loi de Fisher à $J-1$ et $IJ(K-1)$ degrés de liberté.

2.3.5.3 Tester l'interaction

$$H_0 : (\alpha\beta)_{11} = (\alpha\beta)_{12} = \ldots = (\alpha\beta)_{IJ} = 0 \quad \textbf{(Absence d'interaction)}$$

contre

$$H_1 : \text{il existe } (i,j) \in \{1,\ldots,I\} \times \{1,\ldots,J\} \text{ tel que } (\alpha\beta)_{ij} \neq 0 \quad \textbf{(Présence d'interaction)}.$$

La région critique du test est :

$$W = \left\{ F_{XZ} > f_{XZ}^* \right\}$$

où f_{XZ}^* est le quantile d'ordre $1-\alpha$ de la loi de Fisher à $(I-1)(J-1)$ et $IJ(K-1)$ degrés de liberté.

2.3.6 Comparaisons multiples

Lorsque l'hypothèse de Présence d'effet ou d'interaction (H_1) est acceptée, on procède à des comparaisons multiples pour detecter les groupes qui sont significativement différents. Dans ce cours, nous utilisons le test de Tukey.

2.3.7 Exemple

Exercise 7. *Des secrétaires tapent un texte pendant 5 minutes sur différentes machines à écrire. L'expérience est répétée le lendemain.*

Machine à écrire	*Secrétaire 1*	*Secrétaire 2*	*Secrétaire 3*	*Secrétaire 4*	*Secrétaire 5*
1	*33*	*31*	*34*	*34*	*31*
1	*32*	*31*	*36*	*33*	*31*
2	*32*	*37*	*39*	*33*	*35*
2	*35*	*35*	*36*	*36*	*36*
3	*37*	*35*	*34*	*31*	*37*
3	*39*	*35*	*37*	*35*	*39*
4	*29*	*31*	*33*	*31*	*33*
4	*31*	*33*	*34*	*27*	*33*

```
> Y=c(33,36,32,35,37,39,29,31,31,31,37,35,35,35,31,33,34,36,39,36,34,37,33,34,34,
+ 33,33,36,31,35,31,27,31,31,35,36,37,39,33,33)
> Secretaire=as.factor(c(rep(1,8),rep(2,8),rep(3,8),rep(4,8),rep(5,8)))
> Machine=as.factor(c(1,1,2,2,3,3,4,4,1,1,2,2,3,3,4,4,1,1,2,2,3,3,4,4,1,1,2,2,3,3,4,4,1,1,2,2,3,3
> Tableau=data.frame(Machine,Secretaire,Y)
```

Test de l'interaction : Analyser cette sortie de logiciel au niveau $\alpha = 0.05$.

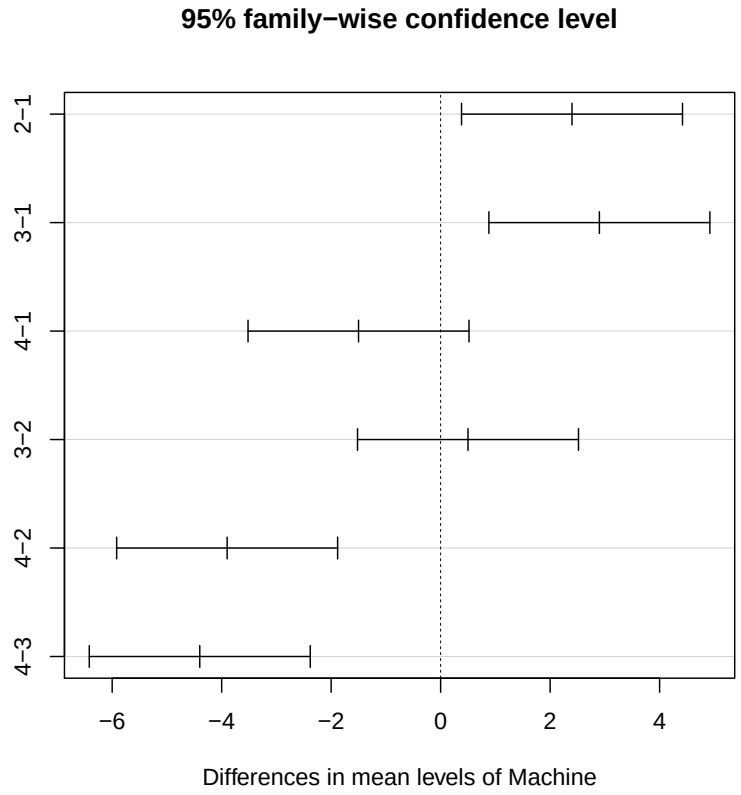
```
> Resultat=aov(Y ~ Machine:Secretaire,data=Tableau)
> summary(Resultat)

                   Df Sum Sq Mean Sq F value   Pr(>F)
Machine:Secretaire 19  241.9   12.73   4.897 0.000438 ***
Residuals          20   52.0    2.60
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Analyser cette sortie au niveau $\alpha = 0.05$.

```
> Resultat1=aov(Y ~ Machine*Secretaire,data=Tableau)
> summary(Resultat1)

                   Df Sum Sq Mean Sq F value   Pr(>F)
Machine             3 128.10   42.70  16.423 1.28e-05 ***
Secretaire          4  36.15    9.04   3.476   0.0260 *
Machine:Secretaire 12  77.65    6.47   2.489   0.0345 *
Residuals          20  52.00    2.60
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> TukeyHSD(Resultat1,"Machine")

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Y ~ Machine * Secretaire, data = Tableau)

$Machine
    diff        lwr        upr     p adj
```

```
2-1  2.4  0.3816584  4.4183416 0.0162573
3-1  2.9  0.8816584  4.9183416 0.0034425
4-1 -1.5 -3.5183416  0.5183416 0.1935803
3-2  0.5 -1.5183416  2.5183416 0.8984133
4-2 -3.9 -5.9183416 -1.8816584 0.0001479
4-3 -4.4 -6.4183416 -2.3816584 0.0000321
```

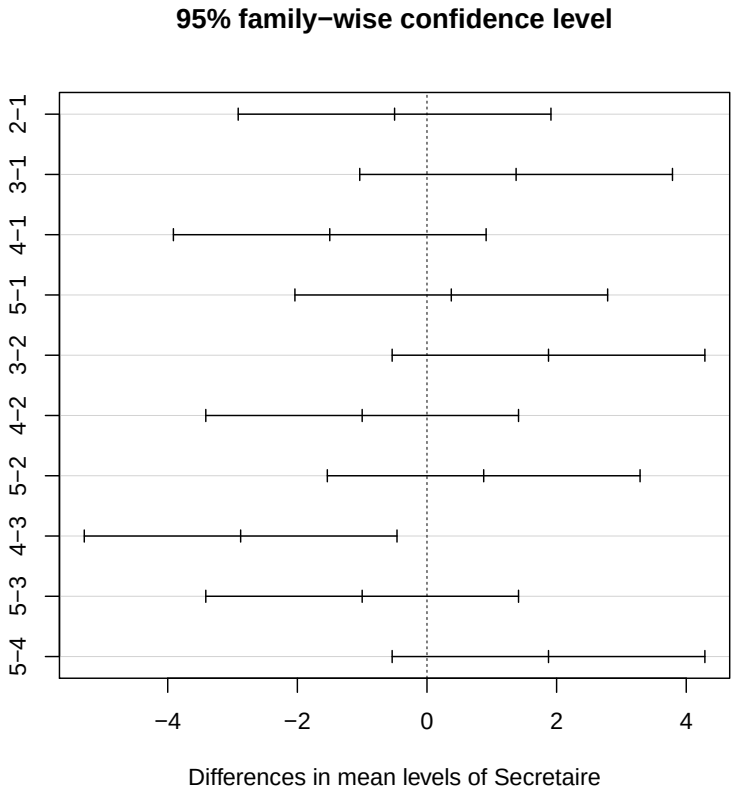
```
> plot(TukeyHSD(Resultat1,"Machine"))
```




```
> TukeyHSD(Resultat1,"Secretaire")

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Y ~ Machine * Secretaire, data = Tableau)

$Secretaire
      diff        lwr        upr     p adj
2-1 -0.500 -2.9125295  1.9125295 0.9701065
3-1  1.375 -1.0375295  3.7875295 0.4530013
4-1 -1.500 -3.9125295  0.9125295 0.3690734
5-1  0.375 -2.0375295  2.7875295 0.9896473
3-2  1.875 -0.5375295  4.2875295 0.1778975
4-2 -1.000 -3.4125295  1.4125295 0.7286977
5-2  0.875 -1.5375295  3.2875295 0.8118838
```

```
4-3 -2.875 -5.2875295 -0.4624705 0.0147680
5-3 -1.000 -3.4125295  1.4125295 0.7286977
5-4  1.875 -0.5375295  4.2875295 0.1778975
```

```
> plot(TukeyHSD(Resultat1,"Secretaire"))
```



Bibliographie

[1] J. M. Azaïs, J. M. Bardet, Le modèle linéaire par l'exemple, Dunod, Paris, 2005.

[2] P. Cornillon, E. Matzner-Lober, Régression avec R, Springer-Verlag France, 2011.