Régression linéaire multiple et Modèle Linéaire Généralisé

Mathilde Boissel

25/01/2021

Table of Contents

Régression linéaire multiple

Contexte

On souhaite prédire et/ou expliquer les valeurs d'une variable quantitative Y à partir des valeurs de p variables $X_1, \ldots, X_p$.

On dit alors que l'on souhaite "expliquer Y à partir de $X_1, \ldots, X_p$", Y est appelée "variable à expliquer" et $X_1, \ldots, X_p$ sont appelées "variables explicatives".

Données

Les données dont on dispose sont n observations de $(Y, X_1, \ldots, X_p)$ notées $(y_i, x_{1,i}, \ldots, x_{p,i})$, $\forall i \in \{1, \ldots, n\}$. Généralement les données sont sous la forme d'un tableau comme suit :

Y	X_1	$\ldots$	X_p
y_1	$x_{1,1}$	$\ldots$	$x_{p,1}$
y_2	$x_{1,2}$	$\ldots$	$x_{p,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	$x_{1,n}$	$\ldots$	$x_{p,n}$

Modèle de régression linéaire multiple, forme générique

Si une liaison linéaire entre Y et $X_1, \ldots, X_p$ est envisageable, on peut utiliser le modèle de régression linéaire multiple (rlm). Sa forme générique est

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon}$$

où $\beta_0, \ldots, \beta_p$ sont des coefficients réels inconnus et ε est une variable quantitative de valeur moyenne nulle, indépendante de $X_1, \ldots, X_p$, qui représente une somme d'erreurs aléatoires et multifactorielles.

Sous R, on modélise Y en fonction X_1, X_2, X_3 comme suit : `reg = lm(Y ~ X1 + X2 + X3)`

Modèle de rlm

On modélise les variables considérées comme des variables aléatoires réelles. Le modèle de rlm est caractérisé, $\forall i \in \{1, \dots, n\}$, par :

- $(x_{1,i}, \dots, x_{p,i})$ est une réalisation du vecteur aléatoire réel $(X_1, \dots, X_p)$,
- sachant que $(X_1, \dots, X_p) = (x_{1,i}, \dots, x_{p,i})$, y_i est une réalisation de

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_p X_{p,i} + \varepsilon_i,$$

où ε_i est une variable indépendante de $(X_1, \dots, X_p)$ avec $\mathbb{E}(\varepsilon_i) = 0$.

D'autres hypothèses sur les erreurs $\varepsilon_1, \dots, \varepsilon_n$ seront formulées ultérieurement.

Ecriture matricielle

Le modèle de rlm s'écrit sous la forme matricielle :

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

où

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{p,1} \\ 1 & x_{1,2} & \cdots & x_{p,2} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{1,n} & \cdots & x_{p,n} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Hypothèses standards

Les hypothèses standards sur le modèle de rlm sont :

- X est de rang colonnes plein (donc $(X^t X)^{-1}$ existe),
- ε et $(X_1, \dots, X_p)$ sont indépendantes,
- $\varepsilon \sim \mathcal{N}_n(0_n, \sigma^2 \mathbb{I}_n)$ où $\sigma > 0$ est un paramètre inconnu.

En particulier, cette dernière hypothèse entraîne que

- $\varepsilon_1, \dots, \varepsilon_n$ sont indépendantes,
- $\mathbb{V}(\varepsilon_1) = \dots = \mathbb{V}(\varepsilon_n) = \sigma^2$,
- $\varepsilon_1, \dots, \varepsilon_n$ suivent chacune une loi normale $\mathcal{N}(0, \sigma^2)$.

Dans ce qui suit, on suppose que les hypothèses standards sont satisfaites.

Estimation

EMCO

L'estimateur des moindres carrés ordinaires (emco) de β est

$$\hat{\beta} = (X^t X)^{-1} X^t Y$$

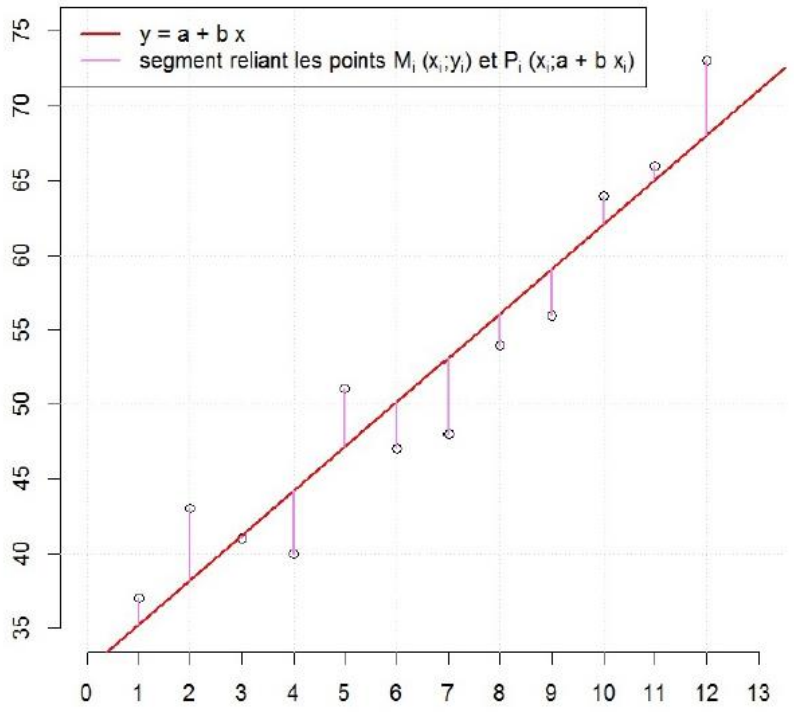
Il est construit de sorte que l'erreur d'estimation entre $X\hat{\beta}$ et Y soit la plus petite possible sens $||\bullet||^2$ (de la distance) :

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{p+1}}{\text{Argmin}} ||Y - X\beta||^2,$$

où $||\bullet||^2$ désigne la norme euclidienne de $\mathbb{R}^n$.

En d'autres termes, l'équation, linéaire, est celle de la droite (ou du plan) qui passe "au mieux" dans les points. Les paramètres $\beta_0, \beta_1, \ldots, \beta_p$ peuvent être estimés par la méthode des moindres carrées ("least-square method"), leurs estimations sont celles qui minimisent la quantité suivante, qui correspond à la somme des carrés des écarts au modèle :

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}))^2$$



$\forall j \in \{0, \ldots, p\}$, la j+1-ième composante de $\hat{\beta}$, notée $\hat{\beta}_j$, est l'**emco** de β_j.

EMCO et EMV

L'emco de β est l'estimateur du maximum de vraisemblance (**emv**) de β.

En effet, la vraisemblance associée à $(Y_1, \ldots, Y_n)$ est

$$L(\beta, z) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{||z - X\beta||^2}{2\sigma^2}\right), \quad z \in \mathbb{R}^n.$$

Par conséquent

$$\underset{\beta \in \mathbb{R}^{p+1}}{\text{Argmax}} L(\beta, Y) = \underset{\beta \in \mathbb{R}^{p+1}}{\text{Argmin}} ||Y - X\beta||^2 = \hat{\beta}$$

Estimateur de la valeur moyenne

Soit y_x la valeur prédite moyenne de Y lorsque $(X_1, \ldots, X_p) = (x_1, \ldots, x_p) = x$:

$$y_x = \mathbb{E}(Y | \{(X_1, \ldots, X_p) = x\}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Un estimateur de y_x est

$$\hat{Y}_x = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \ldots + \widehat{\beta_p} x_p$$

Estimateur de σ^2

Un estimateur de σ^2 est $\hat{\sigma}^2 = \frac{1}{n-(p+1)} ||Y - X\hat{\beta}||^2$.
Il vérifie $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$. De plus, $\hat{\sigma}^2$ et $\hat{\beta}$ sont indépendants.

Estimations ponctuelles

En pratique, on considère les réalisations de $\widehat{\beta_0}, \widehat{\beta_1}, \ldots, \widehat{\beta_p}, \hat{\sigma}^2$ correspondantes aux données. On travaille donc avec des réels.

```r
reg <- lm(Y ~ X1 + X2 + X3)
reg ## les estimations ponctuelles des beta
## Pour isoler l'estimation ponctuelle de beta_2, par exemple :
reg$coeff[3]
# Les valeurs prédites moyennes de Y prises aux valeurs des données de X1, X2
et X3 s'obtiennent en faisant :
predict(reg) # (ou fitted(reg))
# La valeur prédite moyenne de Y pour la valeur (X1, X2, X3) = (1.2, 2.2, 6)
est donnée par les commandes :
predict(reg, data.frame(X1 = 1.2, X2 = 2.2, X3 = 6))
```

```r
# Si le coefficient beta_0 n'a pas de sens dans la modélisation, on l'enlève
en faisant "-1":
reg = lm(Y ~ X1 + X2 + X3 - 1)
```

Coefficients de détermination

- On appelle **coefficient de détermination** la réalisation R^2 de

$$\hat{R}^2 = \frac{||\hat{Y} - \overline{Y}||^2}{||Y - \overline{Y}||^2}$$

où $\hat{Y} = X\hat{\beta}$ et $\overline{Y} = (1/n)\sum_{i=1}^{n} Y_i$.

Ce R^2 est un coefficient réel toujours compris entre 0 et 1. Il mesure de la qualité de l'ajustement des données par le modèle de rlm ; plus R2 est proche de 1, (plus $\overline{Y}$ est proche de Y), meilleur est le modèle. Comme le R2 dépend fortement de p, on ne peut pas l'utiliser pour comparer la qualité de 2 modèles de rlm qui diffèrent quant au nombre de variables explicatives. C'est pourquoi on lui préfère sa version ajustée présentée ci-dessous.

- On appelle **coefficient de détermination ajusté** le réel :

$$\overline{R}^2 = R^2 - (1 - R^2)\frac{p}{n - (p+1)}$$

Ce coefficient est considéré comme plus fiable que R^2 car il tient compte du nombre de variables.

Les coefficients R^2 et $\overline{R}^2$ sont donnés par la commande R : `summary(reg)`.

Lois des estimateurs

Loi de $\widehat{\beta}$

On a

$$\hat{\beta} \sim \mathcal{N}_{p+1}(\beta, \sigma^2(X^tX)^{-1}).$$

La matrice de covariance estimée de $\hat{\beta}$, qui est la réalisation de $\hat{\sigma}^2(X^tX)^{-1}$, est donnée par la commande R : `vcov(reg)`.

Loi de $\widehat{\beta}_j$

Pour tout $j \in \{0, \dots, p\}$, on a

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2[(X^tX)^{-1}]_{j+1,j+1}), \quad \frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{[(X^tX)^{-1}]_{j+1,j+1}}} \sim \mathcal{N}(0,1),$$

où $[(X^tX)^{-1}]_{j+1,j+1}$ désigne la j+1 -ième composante diagonale de $(X^tX)^{-1}$.

Degrés de liberté

Dans ce qui suit, on travaillera avec le nombre de degrés de liberté : $\nu = n - (p+1)$.

Loi associée à $\widehat{\sigma}^2$

On a $$(n-(p+1))\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(\nu).$$

Apparition de la loi de Student

Pour tout $j \in \{0,\dots,p\}$, on a

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{[(X^tX)^{-1}]_{j+1,j+1}}} \sim \mathcal{T}(\nu)$$

Intervalles de confiance

Intervalle de confiance pour β_j

Pour tout $j \in \{0,\dots,p\}$, un intervalle de confiance pour β_j au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est la réalisation i_{β_j} de

$$I_{\beta_j} = \left[\widehat{\beta_j} - t_\alpha(\nu)\hat{\sigma}\sqrt{[(X^tX)^{-1}]_{j+1,j+1}}, \widehat{\beta_j} + t_\alpha(\nu)\hat{\sigma}\sqrt{[(X^tX)^{-1}]_{j+1,j+1}}\right]$$

où $t_\alpha(\nu)$ est le réel vérifiant $\mathbb{P}(|T| \geq t_\alpha(\nu)) = \alpha$, avec $T \sim \mathcal{T}(\nu)$.

Pour obtenir cet intervalle dans R, on fait : `confint(reg, level = 0.95)`.

Intervalle de confiance pour y_x

Soient y_x la prédiction moyenne de Y quand $(X_1,\dots,X_p) = (x_1,\dots,x_p) = x$ et $x_\bullet = (1, x_1,\dots,x_p)$.

Un intervalle de confiance pour y_x au niveau $100(1-\alpha)\%$, $\alpha \in]0,1[$, est la réalisation i_{y_x} de

$$I_{y_x} = \left[\widehat{Y_x} - t_\alpha(\nu)\hat{\sigma}\sqrt{x_\bullet(X^tX)^{-1}x_\bullet}, \widehat{Y_x} + t_\alpha(\nu)\hat{\sigma}\sqrt{x_\bullet(X^tX)^{-1}x_\bullet}\right]$$

où $t_\alpha(\nu)$ est le réel vérifiant $\mathbb{P}(|T| \geq t_\alpha(\nu)) = \alpha$, avec $T \sim \mathcal{T}(\nu)$.

Pour obtenir cet intervalle dans R, on fait : `predict(reg, data.frame(X1 = 1.2, X2 = 2.2, X3 = 6), interval = "confidence")`.

Tests statistiques

p-valeur

On considère des hypothèses de la forme :

H0 : "A" contre H1 : "contraire de A"

La p-valeur est le plus petit réel $\alpha \in]0,1[$, calculé à partir des données tel que l'on puisse se permettre de rejetter H0 au risque $100 \times \alpha\%$. Autrement dit, la p-valeur est une estimation ponctuelle de la probabilité critique de se tromper en rejetant H0 (affirmant H1 alors que H0 est vraie).

Degrés de significativité

Le rejet de H0 sera

- significatif si p-valeur $\in]0.01,0.05]$, symbolisé par *,
- très significatif si p-valeur $\in]0.001,0.01]$, symbolisé par **,
- hautement significatif si p-valeur < 0.001, symbolisé par ***,
- (presque significatif si p-valeur $\in]0.05,0.1]$, symbolisé par . (un point)).

Test de Student

Soit $j \in \{0, \dots, p\}$. L'objectif du test de Student est d'évaluer l'influence de X_j sur Y.

On considère les hypothèses :

$$\mathbf{H0 : \beta_j = 0 \text{ contre } H1 : \beta_j \neq 0.}$$

On calcule la réalisation t_{obs} de

$$T_* = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{[(X^t X)^{-1}]_{j+1,j+1}}}$$

On considère une variable $T \sim \mathcal{T}(\nu)$.
Alors la p-valeur associée est p-valeur$= \mathbb{P}(|T| \geq |t_{obs}|)$.

Test global de Fisher

L'objectif du test global de Fisher est d'étudier la pertinence du lien linéaire entre Y et $X_1, \dots, X_p$.
On considère les hypothèses :

$$\mathbf{H0 : \beta_1 = \beta 2 = \dots = \beta_p = 0 \text{ contre H1 : "il existe au moins un coefficient non nul".}}$$

On calcule la réalisation f_{obs} de

$$F_* = \frac{\hat{R}^2}{1-\hat{R}^2}\frac{n-(p+1)}{p}$$

On considère une variable $F \sim \mathcal{F}(p, \nu)$.
Alors la p-valeur associée est p-valeur$= \mathbb{P}(F \geq f_{obs})$.

Notons que ce test est moins précis que le test de Student car il ne précise pas quels sont les coefficients non nuls. Il est toutefois un indicateur utile pour déceler d'éventuelles problèmes (comme des colinéarités entre $X_1, \ldots, X_p$).

Les tests statistiques précédents sont mis en oeuvre par la commande R : `summary(reg)`.

Validation des hypothèses de la rlm

L'enjeu ici est de valider les hypothèses standard du modèle de rlm avec les données.

Validation graphique

```
par(mfrow = c(2, 2)); plot(reg, 1:4)
```

L'enjeu des 4 graphiques affichés a été expliqué en cours ANOVA. Rappel :

- **Residuals vs Fitted** et **Constant Leverage: Residuals vs Factor Levels**.

  Les résidus sont indépendants. Les résidus ne doivent pas être corrélés entre eux. De la même façon, les résidus ne doivent pas être corrélés au facteur étudié. On peut faire le test de Dubin-Watson pour vérifier l'autocorrélation des résidus mais souvent un contrôle graphique suffit.

- **Normal Q-Q**.

  Les résidus suivent une loi normale de moyenne 0. Pour vérifier cette hypothèse on peut faire un test de normalité comme le test de Shapiro-Wilk mais on préfère vérifier cela graphiquement avec un diagramme Quantile-Quantile (i.e. QQ-plot, graphique dans lequel les quantiles de deux distributions sont tracés l'un par rapport à l'autre).

- **Scale-Location**.

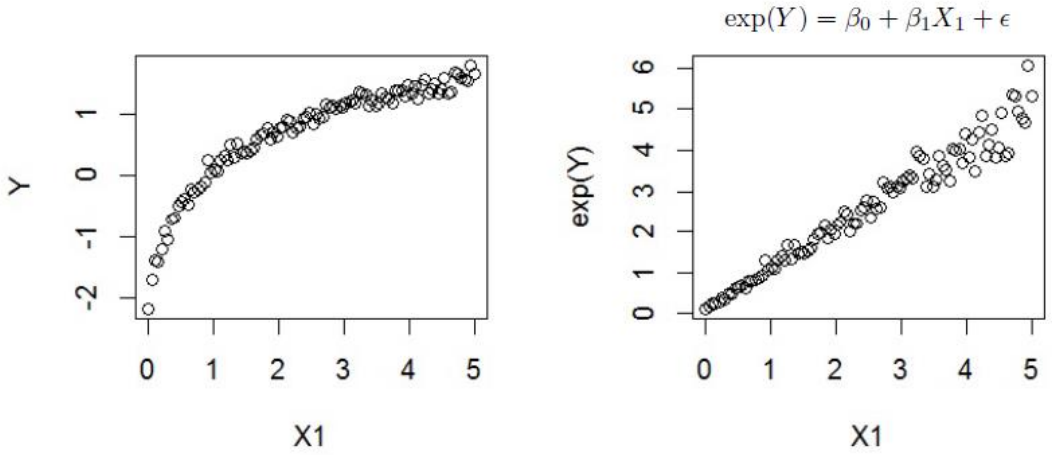
  L'homogénéité des variances. Les résidus relatifs aux différentes modalités sont homogènes (ils ont globalement la même dispersion), autrement dit leur variance est constante. On peut vérifier cela graphiquement en représentant les résidus standardisés en fonction des valeurs prédites (les moyennes des différents traitements). En cas de doute on pourra aussi valider cette hypothèse avec un test statistique (Cochran, Bartlett, Levene…).

Analyses du/des nuages de points

Les p nuages de points $\{(x_{ji}, y_i); i \in \{1, \dots, n\}\}$, $j \in \{1, \dots, p\}$ peuvent nous aiguiller sur les transformations candidates permettant de rendre les variables normales.

Les nuages de points 2 à 2 peuvent être obtenus avec les commandes R : `plot(w)` ou `pairs(w)` ou, par exemple : `pairs(~ Y + X1 + X4).`

Le package `car` propose également `scatterplotMatrix(w)`



Vu le nuage de points, il est préférable de considérer la transformation exp(Y) et de faire une régression linéaire sur X1.

Un exemple de rlm avec variables transformées est
`reg = lm(log(Y) ~ sqrt(X1) + exp(X2) + I(X3^4))`

Analyses graphiques des résidus

Pour tout $i \in \{1, \ldots, n\}$, on appelle i-ième **résidu** la réalisation e_i de

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i$$

Obtenus dans R comme suit : `residuals(reg)`.

Pour tout $i \in \{1, \ldots, n\}$, on appelle i-ième **résidu standardisé** la réalisation e_i^* de

$$\hat{\epsilon}_i^* = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - [X(X^tX)^{-1}X^t]_{i,i}}}$$

Obtenus dans R comme suit : `rstandard(reg)`.

- Avec une analyse graphique principale `plot(residuals(reg))`, on doit voir que
  - le nuage de points n'a aucune structure particulière,
  - il y a une symétrie dans la répartition des points par rapport à l'axe des abscisses alors on admet la normalité.
  - On voit des problèmes (1) Si le nuage de points a l'allure d'une route sinueuse ou d'un mégaphone, on soupçonne que les résidus et les variables sont X1, …, Xp sont dépendantes et/ou que les résidus sont dépendants et/ou que leur variance n'est pas égales. ou (2) S'il y a une asymétrie dans la répartition des points par rapport à l'axe des abscisses, l'hypothèse de normalité est à étudier.

- L'indépendance des résidus peut aussi se contrôler en traçant le corrélogramme : Celui-ci représente les estimations ponctuelles de la fonction d'autocorrélation (acf), comme suit `acf(residuals(reg))`.
  Si les bâtons se suivent en ayant des tailles et des signes sans logique apparente, on admet l'indépendance des résidus.

- Le corrélogramme partiel vient compléter l'étude précédente : il représente les estimations ponctuelles de la fonction d'autocorrélation partielle (pacf) sous forme de bâtons. L'interprétation est la même que pour l'acf, et se fait comme suit `pacf(residuals(reg))`.
  On soupçonne des problèmes si les sommets des bâtons peuvent être rejoints par une ligne courbe "sans pic" ou si plusieurs bâtons dépassent les bornes de l'intervalle de confiance, une dépendance peut-être soupçonnée. Cela peut être confirmé avec le test de Ljung-Box. `library(lawstat); Box.test(residuals(reg), type = "Ljung")`

Rappel

Rappel 1 – l'analyse de la variance à un facteur ou plus

On cherche ici à tester l'effet d'un (ou de plusieurs) facteur(s) qualitatif(s) sur une variable quantitative. Plus précisément, l'objectif est de comparer statistiquement les moyennes de la variable quantitative mesurée dans chacune des modalités de la – ou des – facteur(s) contrôlé(s).
Pour ce faire, la variance totale de la variable mesurée est décomposée en variance due au(x) facteur(s) contrôlé(s) et variance résiduelle, et ces deux variances sont comparées par un test de Fisher.

Sous R, en supposant l'exemple d'une analyse de la variance (ANOVA) sur une variable y avec un facteur ou plusieurs facteurs « factor1 », « factor2 », etc., la syntaxe suivante peut être adoptée, avec la fonction `aov()`:

```
res <- aov(formula = y~factor1+factor2)
summary(res)
plot(res)
plot(y~factor1+factor2)
# Pour calculer des statistiques pour chaque modalité d'un facteur :
tapply(y, factor1, mean)
tapply(y, factor1, var)
tapply(y, factor1, summary)
```

Les modèles

Le modèle linéaire général

L'ensemble des résultats statistiques d'une régression linéaire peut être fourni sous la forme d'un tableau d'ANOVA, avec la fonction `anova()`, qui décompose et teste la variance de y due à la régression par rapport aux variances résiduelle et totale.

```
anova(res)
```

Dans R, pour une utilisation de la foncion `anova()` sur un objet de classe `lm`, voir l'aide `?anova.lm`. Par défault, on voit que c'est la `F statistics` qui est utilisé.

En effet, la sortie de : `anova(lm(y~x))` est la même que la sortie de : `summary(aov(y~x))` (comme vu brièvement en TP ANOVA).

La raison est que ces deux modèles, celui de la régression linéaire simple (ou multiple) et celui de l'ANOVA à un (ou plusieurs) facteur(s), sont justement des **cas particuliers d'un cadre général que l'on nomme le "modèle linéaire général"**, et qui a pour caractéristiques les éléments suivants :

1) Le modèle à ajuster est de la forme :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Dans le cadre de la régression linéaire simple ou multiple, c'est l'équation que l'on a vue plus haut. Dans le cadre de l'ANOVA, les facteurs à tester sont préalablement codés sous forme matricielle, et on retombe sur ce schéma linéaire de base.

N.B. : Cette "forme matricielle" est aussi appelée "matrice disjonctive". Il est possible de récupérer la forme matricielle qui sert à l'ajustement dans le cas d'une ANOVA avec la fonction `model.matrix()` e.g., `model.matrix(x~factor1)`.

2) Plus précisément, et c'est là la remarque la plus importante, **le modèle à ajuster est en fait dans tous les cas de la forme :**

$$\boldsymbol{E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon}$$

Où $E(y)$ représente l'espérance de la variable y (qui est équivalente à sa moyenne, ou plus précisément sa moyenne attendue) et ε représente le terme d'erreur (i.e., de bruit) non contrôlé qui doit impérativement suivre une distribution normale et de même variance. C'est à dire que la variance de ce terme d'erreur doit être indépendante de la valeur des différentes variables x1, x2, etc. (qui, rappelons-le, peuvent correspondre au codage de facteurs quantitatifs).

Il existe, sous R, une fonction qui permet d'ajuster un modèle linéaire général, et qui rend donc caduques les fonctions `lm()` et `aov()`. Il s'agit de la **fonction `glm()`**. Attention cependant, le nom de cette fonction ne veut pourtant pas dire "general linear model" contrairement à ce qu'on pourrait penser, mais signifie **"generalized linear model"**, qui est un cadre encore plus général que nous verrons plus bas.

Ainsi, une régression linéaire simple ou multiple peut se calculer, par exemple, comme suit :

```
res <- glm(y~x)
summary(res)
plot(res)
anova(res, test = "F") ## ?anova.glm
```

Et une ANOVA avec exactement la même syntaxe :

```
res <- glm(y~factor1+factor2)
summary(res)
plot(res)
anova(res, test = "F")
```

Ces syntaxes retournent évidement les mêmes résultats que ceux issus des fonctions `lm()` et `aov()`, respectivement.

Nous avons vu qu'une régression linéaire est un cas particulier du modèle linéaire général. Elle consiste à chercher à expliquer une variable quantitative par une autre variable quantitative. Nous avons vu qu'une régression linéaire multiple l'est également. Elle consiste à chercher à expliquer une variable quantitative par plusieurs variables quantitatives. Enfin, nous avons vu qu'une ANOVA l'est également. Elle consiste à chercher à expliquer une variable quantitative par une ou plusieurs autre(s) variable(s) qualitative(s). Dans tous les cas, on cherche à expliquer une variable quantitative par une ou plusieurs variables quantitatives ou qualitatives. On tombe donc sur le tableau suivant, qui donne les différents types de modèle linéaire général que l'on peut rencontrer.

Si	Alors on a
Il y a une variable explicative qui est quantitative	Une régression linéaire simple
Il y a plusieurs variables explicatives qui sont toutes quantitatives	Une régression linéaire multiple
Il y a une variable explicative qui est qualitative	Une ANOVA à un facteur
Il y a plusieurs variables explicatives qui sont toutes qualitatives	Une ANOVA à plusieurs facteurs
Il y a une combinaison de variables explicatives quantitatives et qualitatives	Une analyse de covariance (ANACOV ou ANCOVA)

Le modèle linéaire généralisé

Nous avons vu que le modèle linéaire général repose sur une hypothèse forte : le terme d'erreur suit une loi normale et de même variance. **Nous avons pourtant parfois (souvent…) le besoin d'expliquer des variables (et donc leurs erreurs) qui ne suivent pas ce prérequis.**

Prenons deux exemples, celui où la variable y mesurée est un pourcentage (e.g., le pourcentage de mâles ou de femelles dans une population où il n'y évidemment que des mâles et des femelles - *c.f.* loi de Bernoulli, répétée n fois -), et celui où la variable y mesurée est un comptage (e.g., nombre d'oeufs pondus par une poule). Dans ces deux cas – et d'autres encore – le simple modèle linéaire présenté ci-dessus ne peux pas convenir et ce pour au moins deux raisons importantes :

1) la principale et la plus importante est que la distribution de la variable à expliquer n'est pas compatible avec le modèle linéaire présenté ci-dessus.

   Par exemple, dans le cas du comptage, seules des valeurs entières et positives ou nulles peuvent être mesurées, alors que le modèle linéaire simple ci-dessus pourra malgré tout prédire des valeurs décimales et/ou négatives.

   De même, un pourcentage est par définition compris dans l'intervalle [0, 1] (ou [0%, 100%] ; on ne peut avoir moins de 0% ou plus de 100% de mâles ou de femelles) alors que le modèle linéaire ci-dessus pourra prédire des valeurs qui pourront sortir de cet intervalle.
   Par ailleurs considérer la variable à expliquer (et son erreur) comme suivant une loi normale suppose une distribution symétrique autour de la moyenne, alors que ce n'est très probablement pas le cas, par exemple, du comptage où la majorité des valeurs mesurées seront par exemple plus fréquemment vers zéro ou un que vers 100 ou 200.

2) L'autre raison est que, dans le modèle linéaire simple (général), les variables prédictives ont un effet linéaire sur la variable mesurée (effet qui se traduit par les coefficients de régression), or ces effets ne sont peut-être pas linéaires en réalité. Par exemple, le nombre d'oeufs pondus par une poule ne change peut-être pas linéairement avec son âge.

Pour tenir compte de ces points plusieurs solutions s'offrent à nous. La plus répandue consiste à trouver une transformation mathématique de la variable à expliquer pour la rendre normale (et son erreur avec) et pour en stabiliser les variances. On parle de **transformations "normalisantes"**. Plusieurs sont connues (log, racine, exp…). Ces transformations ne sont pas toutes efficaces, et leur effet normalisant est parfois difficiles à quantifier. Par ailleurs, il reste évidemment préférable d'utiliser les données d'origine plutôt que leurs valeurs transformées, ne serait-ce que pour rendre l'interprétation des résultats plus aisée.

C'est dans ce cadre que se développe le modèle linéaire généralis**é** (Generali**zed** Linear Model, GLM).

L'idée reste d'utiliser **une transformation mathématique sur la variable à expliquer y mais en tenant compte cette fois-ci de la véritable distribution des erreurs** (par exemple une loi de Poisson dans le cadre de comptages ; une loi Binomiale dans le cas de pourcentages, etc). Ceci implique entre autre que les paramètres ne sont alors plus estimés par la simple méthode des moindres carrés - comme dans le modèle linéaire général - mais par une autre méthode d'estimation : la méthode dite du **"maximum de vraisemblance"**.

La fonction mathématique utilisée pour transformer la variable à expliquer est appelée **"fonction de lien"** ("link function"), et plusieurs peuvent être utilisées selon la distribution réelle de la variable d'intérêt (et son erreur). Du coup, le modèle à ajuster devient :

$$\boldsymbol{f(E(y)) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_0}$$

Où $f(\ldots)$ est la fonction de lien.

Les cas

Un GLM peut être théoriquement utilisé quelle que soit la distribution de la variable à expliquer y.

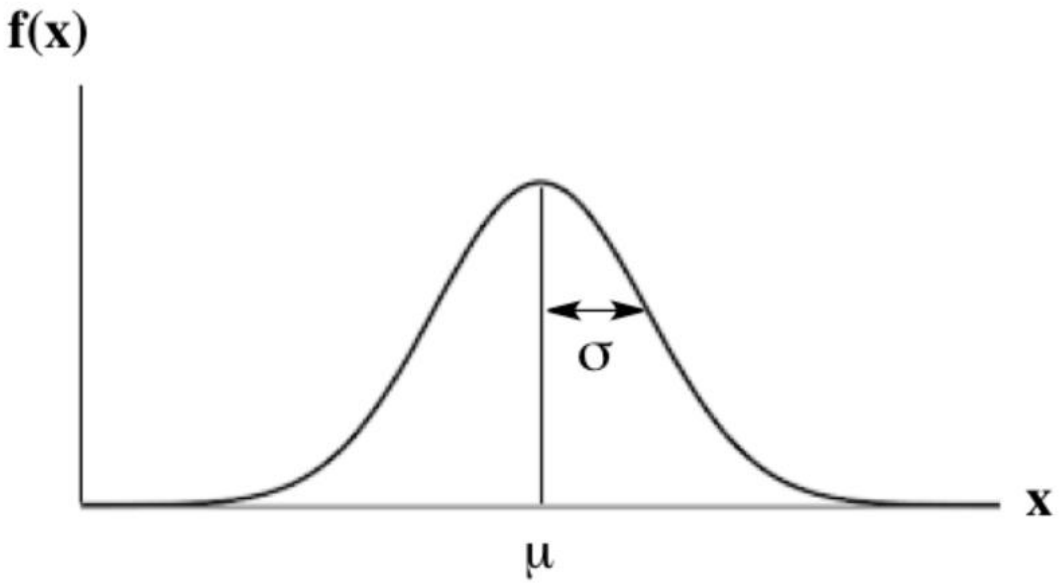
Cas Gaussien

Dans le cas où la variable y suit une loi normale : le modèle linéaire général vu ci-dessus (T-test, régression simple ou multiple, ANOVA, etc) est lui-même un cas particulier du modèle linéaire généralisé.

Rappelons qu'une loi normale a une forme "en cloche", symétrique de part et d'autre de sa valeur maximale qui correspond à la moyenne de la variable étudiée. Elle décrit la distribution d'une variable continue. Son équation est

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Où μ est la moyenne et σ est l'écart-type.



Si des variables telles que le poids et la taille des individus, la surface de cellules, le taux de glucide, sont supposées suivre une loi normale, alors en toute logique, un GLM utilisé pour analyser une variable suivant une loi normale aura pour fonction de lien la fonction identité `f(y)=y`.
Et l'on retombe donc sur le modèle linéaire général pris ici comme un cas particulier :

$$E(y) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_0$$

Sous R, nous l'avons vu, l'ajustement pourra se faire de la manière suivante :

```r
# res <- glm(x~factor1+factor2+etc., family = gaussian)
res <- glm(y~factor1+factor2, family = gaussian(link = "identity"))
summary(res)
plot(res)
anova(res, test="F")
```

L'argument **family = gaussian** peut être omis, car le cas gaussien est le cas par défaut de la fonction `glm()`.

De même que la fonction de lien **link = "identidy"** est la valeur par défaut de la fonction `gaussian()`.

Pour vous en assurer, voir la documentation `?glm` et `?family`.

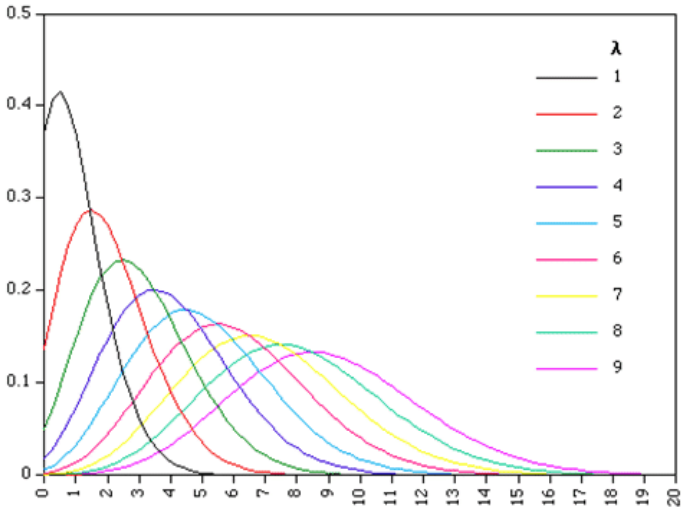
Cas Poissonien

Nous voulons à présent analyser une variable de comptage (par exemple, comme nous l'avons vu ci-dessus, le nombre d'oeufs pondus par une poule).
La variable à analyser suit cette fois-ci une loi de Poisson qui décrit des variables discrètes positives ou nulles. Selon cette loi, la probabilité d'observer une valeur k vaut :

$$f(k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Où λ est la moyenne de la distribution et est également sa variance (caractéristique de cette loi).

Si la moyenne est inférieure à 1, cette distribution aura une forme en "i" (pic de densité) proche de l'origine, sinon elle aura une forme dissymétrique, avec une queue de distribution qui s'aplanit sur la droite.



La fonction de lien utilisée pour analyser une variable suivant une loi de Poisson (i.e., un comptage) est généralement la fonction log : `f(y)=log(y)`, et le modèle s'appelle dans ce cas un modèle "log-linéaire" :

$$log(E(y)) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_0$$

Qui peut se réécrire sous la forme suivante :

$$E(y) = e^{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_0}$$

Dans le cas d'un comptage ne pouvant prendre que des valeurs positives, cette fonction de lien est logique. Elle permet d'avoir des valeurs prédites par le modèle qui s'étendent de $-\infty$ à $+\infty$, ce qui reste interprétable alors que ca ne l'aurait pas été avec un simple modèle linéaire général, comme nous l'avons vu ci-dessus.

Sous R, l'ajustement pourra se faire de la manière suivante :

```
# res <- glm(y_count~factor1+factor2, family = poisson)
res <- glm(y_count~factor1+factor2, family = poisson(link = "log"))
summary(res)
plot(res)
anova(res, test = "Chisq")
```

Il est effectivement préconisé dans ce cas d'utiliser un test de χ^2 ("Chisq") plutôt qu'un test "F" pour vérifier la significativité des effets.

Il se peut parfois que la distribution observée de la variable à expliquer ne suive pas exactement une loi de Poisson telle qu'attendue, mais que sa variance (i.e., sa dispersion) soit plus forte que celle issue d'une simple loi de Poisson. Il est possible de prendre en compte cette possible "sur-dispersion", en remplaçant "family = poisson" par "family = quasipoisson" dans la syntaxe ci-dessus. Dans ce cas, un terme supplémentaire est estimé à partir des données. Ce terme, qualifié de paramètre de dispersion, indique l'augmentation de la variance observée par rapport à celle attendue d'une loi de Poisson. Une valeur de 1,0 signifie que les données suivent bien une loi de Poisson, une valeur supérieure à 1,0 indique que la variance observée est supérieure à celle attendue d'une loi de Poisson. (voir `?family` pour d'autres détails)

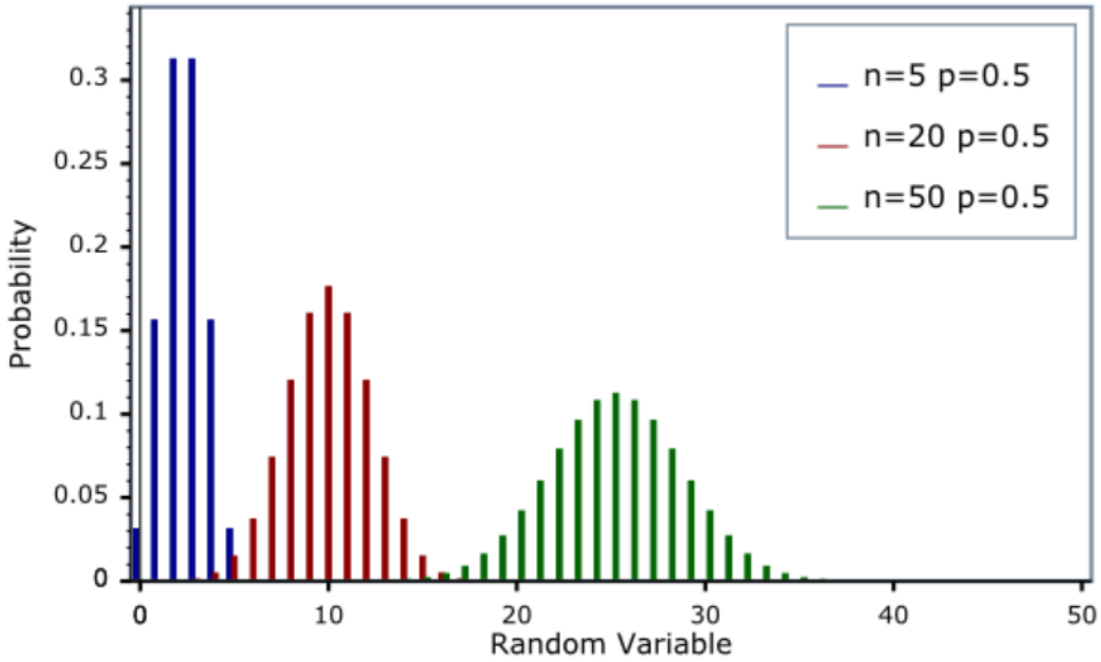
Cas Binomial

La variable mesurée est à présent une proportion (entre 0 et 1) ou, de manière équivalente, un pourcentage (entre 0.0% et 100.0%), par exemple, comme nous l'avons vu ci-dessus, la proportion de mâles ou de femelles dans une population.

La variable à analyser suit cette fois-ci une loi Binomiale qui décrit le nombre de fois où un événement – parmi deux possibles – se produit lorsque l'on répète l'observation (par exemple : nombre de femelles parmi 100 individus ; nombre d'individus vivants sur 30 ; nombre de faces sur 55 jets d'une pièce de monnaie).

Selon cette loi, sur n répétitions, si pour chacune d'elle la probabilité d'observer l'événement étudié est p, la probabilité que l'événement se produise k fois est :

$$f(k) = \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}$$



Généralement, on ne s'intéresse pas tant au nombre de fois où l'on observe l'événement, mais à sa fréquence parmi toutes les répétitions : k/n (exemples : proportion de femelles dans l'échantillon ; proportion d'individus vivants ; proportion de faces sur plusieurs jets d'une pièce de monnaie). Dans ce cas, la moyenne de cette proportion vaut p, et sa variance vaut $p(1-p)/n$.

La fonction de lien utilisée pour analyser une variable suivant une loi Binomiale est généralement la fonction logit : $f(y) = log(y/(1-y))$, et le modèle s'appelle dans ce cas une régression **logistique** :

$$logit(y) = log\left(\frac{y}{1-y}\right) \in \mathbb{R}, \quad y \in]0,1[$$

$$log\left(\frac{E(y)}{1-E(y)}\right) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_0$$

Qui peut se réécrire sous la forme suivante :

$$E(y) = \frac{1}{1 + e^{-(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_0)}}$$

Sous R, un tracé de la fonction logit peut être obtenu de la manière suivante : `curve(log(x/(1-x)), 0, 1)`

Dans le cas d'un pourcentage restant dans l'intervalle [0,1], cette fonction de lien est logique également. Elle permet d'avoir des valeurs prédites par le modèle qui s'étendent de $-\infty$ à $+\infty$, ce qui reste interprétable alors que cela ne l'aurait pas été avec un simple modèle linéaire général, comme nous l'avons vu ci-dessus.

Sous R, l'ajustement pourra se faire à partir de deux manières différentes de présenter les données.

Dans la première, la variable à expliquer est codée sur deux colonnes, disons y1 et y2, l'une contenant le nombre de fois où un évènement (e.g., mâles) a été observé, l'autre contient le nombre de fois où l'autre évènement (e.g., femelles) est observé, dans chaque situation. Dans la seconde, la variable à expliquer (e.g., l'observation d'un mâle) est codée sur une colonne binaire qui indique si l'événement est observé (1) ou non (0) pour chaque mesure.

Dans le premier cas, l'ajustement pourra se faire de la manière suivante :

```
# res <- glm(cbind(y1,y2)~factor1+factor2+etc., family=binomial)
res <- glm(cbind(y1,y2)~factor1+factor2, family = binomial(link = "logit"))
summary(res)
plot(res)
anova(res, test="Chisq")
```

Dans le second cas :

```
res <- glm(y_binary~factor1+factor2, family = binomial(link = "logit"))
summary(res)
plot(res)
anova(res, test="Chisq")
```

Il est effectivement préconisé dans le cas d'une régression logistique également d'utiliser un test de χ^2 ("Chisq") plutôt qu'un test F pour vérifier la significativité des effets.

Comme dans le cas d'un modèle log-linéaire (pour une variable de comptage suivant une loi de Poisson ; voir ci-dessus) il se peut parfois que la distribution observée de la variable à expliquer ne suive pas exactement une loi Binomiale telle qu'attendue, mais que sa variance (i.e., sa dispersion) soit plus forte que celle issue d'une simple loi Binomiale. Ici aussi, il est possible de prendre en compte cette possible "sur-dispersion", en remplaçant "family= binomial" par "family=quasibinomial" dans les syntaxes ci-dessus.

Dans ce cas, comme dans le cas précédent, un terme supplémentaire est estimé à partir des données. Ce terme, qualifié de paramètre de dispersion, indique l'augmentation de la variance observée par rapport à celle attendue d'une loi Binomiale. Son interprétation est la même quand dans le cadre du cas poissonien.

Notons aussi qu'il existe d'autres fonctions de lien, entre autres :

- le lien probit : $\gamma \sim \mathcal{N}(0,1)$:

$$\text{probit}(y) = F_\gamma^{-1}(y), \quad F_\gamma(y) = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t$$

dans R : `family = binomial(link = "probit")`

- le lien cloglog : $\gamma \sim$ Gompertz$(0,1)$:

$$\text{cloglog}(y) = F_\gamma^{-1}(y), \quad F_\gamma(y) = 1 - \exp(-\exp(y))$$

dans R : `family = binomial(link = "cloglog")`

- le lien cauchit : $\gamma \sim$ Cauchy$(0,1)$:

$$\text{cauchit}(y) = F_\gamma^{-1}(y), \quad F_\gamma(y) = \frac{1}{\pi}\arctan(y) + \frac{1}{2}$$

dans R : `family = binomial(link = "cauchit")`

Récap'

Comme son nom l'indique, le modèle linéaire généralisé est un outil qui peut être utilisé dans de nombreuses situations, afin d'analyser des variables qui présentent différents types de distribution statistique. Nous avons vu, rapidement, les cas où la variable suit une loi Normale, une loi de Poisson ou une loi Binomiale, cas qui restent les plus fréquents. D'autres lois de distribution peuvent être considérées, conduisant à chaque fois à l'utilisation de fonctions de lien différentes. Le tableau suivant donne une petite synthèse des principales situations rencontrées.

Distribution	Type de données	Type de GLM	Fonction de lien
Normale	Variable suivant une loi normale	Modèle linéaire général	Identité : $f(y) = y$
Poisson	Comptage	Modèle log-linéaire	Log : $f(y) = log(y)$
Binomiale	Pourcentage	Régression logistique	Logit : $f(y) = log(y/(1-y))$
Gamma	Durée	Modèle Gamma avec fonction de lien inverse	Inverse : $f(y) = 1/y$

Sources

- Le contenu de ce cours s'est basé sur les supports de Eric Wajnberg (Université de Nice) et Christophe Chesneau (Université de Caen)

Lien vers le support très complet : https://chesneau.users.lmno.cnrs.fr/Reg-M2.pdf

Pour la pratique des GLM dans R, vous pouvez vous baser sur ces exemples très bien détaillé : https://chesneau.users.lmno.cnrs.fr/etudes-reg.pdf

Des lectures complémentaires :

- Cours de Régression M2 de Bernard Delyon :

https://perso.univ-rennes1.fr/bernard.delyon/regression.pdf

- Pratique de la régression linéaire multiple de Ricco Rakotomalala :

http://eric.univ-lyon2.fr/~ricco/cours/cours/La_regression_dans_la_pratique.pdf

- Cours de Régression linéaire de Arnaud Guyader :

http://www.lsta.lab.upmc.fr/modules/resources/download/labsta/Pages/Guyader/Regression.pdf

- CookBook R de Vincent Isoz et Daname Kolani :

http://www.sciences.ch/dwnldbl/divers/R.pdf