Statistiques et Applications

Analyse de la variance

aurore.lavigne@univ-lille.fr

Situation du problème

- On désire connaître l'effet du sexe sur la perte de poids lors d'un régime :

Situation du problème

- On désire connaître l'effet du sexe sur la perte de poids lors d'un régime :

  $=>$ 2 modalités Homme - Femme

  $=>$ test de Student de comparaison de moyennes

Situation du problème

- On désire connaître l'effet du sexe sur la perte de poids lors d'un régime :

  => 2 modalités Homme - Femme

  => test de Student de comparaison de moyennes

- On désire connaître l'effet du type de régime :

Situation du problème

- On désire connaître l'effet du sexe sur la perte de poids lors d'un régime :

  => 2 modalités Homme - Femme

  => test de Student de comparaison de moyennes

- On désire connaître l'effet du type de régime :

  => 3 modalités A - B - C

Situation du problème

- On désire connaître l'effet du sexe sur la perte de poids lors d'un régime :
  => 2 modalités Homme - Femme
  => test de Student de comparaison de moyennes

- On désire connaître l'effet du type de régime :
  => 3 modalités A - B - C
  => Analyse de la variance

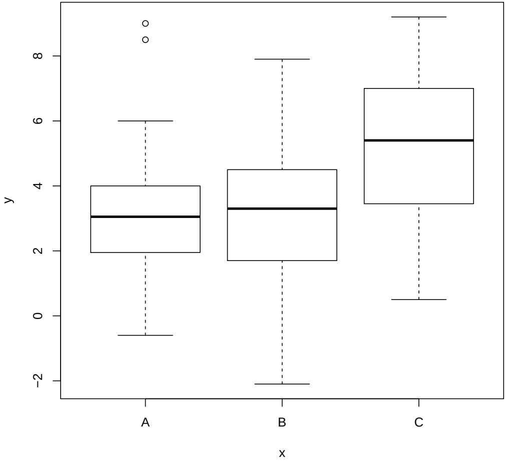


FIGURE – Perte de poids (kg) après un régime de 6 semaines

Analyse de la variance

L'analyse de la variance offre un cadre d'analyse rigoureux pour l'estimation et le test de l'**effet d'une ou plusieurs variables qualitatives sur une variable quantitative**.

VOCABULAIRE :

Les variables qualitatives s'appellent **les facteurs de variabilité** et leurs modalités des **niveaux**. La variable qualitative est la **réponse**.

Plan

Les données et le modèle

On considère **1 facteur** à **k niveaux**.
Pour chaque niveau $l \in \{1, 2, \cdots, k\}$, on dispose d'un échantillon de taille n_i d'observations de la variable quantitative.

$$\begin{array}{ll}
\text{niveau } 1 & Y_1^1, Y_2^1, \cdots, Y_{n_1}^1 \\
\text{niveau } 2 & Y_1^2, Y_2^2, \cdots, Y_{n_2}^2 \\
\vdots & \vdots \\
\text{niveau } k & Y_1^k, Y_2^k, \cdots, Y_{n_k}^k
\end{array}$$

Indépendance

On suppose que les **toutes** les variables sont **indépendantes**.

- Les variables d'un même niveau sont indépendantes :
  $\forall l \in \{1, 2, \cdots, k\}$, $\forall i \neq j$, Y_i^l et Y_j^l sont indépendantes.

- Les variables de deux niveaux différents sont indépendantes
  $\forall l \neq m$, $\forall (i, j)$, Y_i^l et Y_j^m sont indépendantes.

Modèle

On suppose de plus que

- toutes les variables suivent une distribution normale
- **l'espérance dépend du niveau** k
- la variance est identique pour toutes les variables

$$Y_i^l \sim \mathcal{N}(\mu_l, \sigma^2), \quad \forall l \in \{1, \cdots, k\}, \quad \forall i \in \{1, \cdots, n_l\}$$

Modèle

De manière équivalente, on pourra écrire que

$$Y_i^l = \mu_l + \epsilon_i^l \quad \text{avec} \quad \epsilon_i^l \sim \mathcal{N}(0, \sigma^2) \text{ et ind.}$$

- μ_l est l'espérance observée pour le niveau l du facteur.

L'anova est un modèle linéaire

En effet on peut réécrire le modèle ci-dessus de la manière suivante :

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

et

$$\mathbf{X} = \begin{pmatrix} \overbrace{\begin{matrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \end{matrix}}^{k \text{ col.}} & \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_1 \text{ li.} \\ \begin{matrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \cdots & 0 \end{matrix} & \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_2 \text{ li.} \\ \begin{matrix} \vdots & \vdots & \vdots & \vdots \end{matrix} & \\ \begin{matrix} 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{matrix} & \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_k \text{ li.} \end{pmatrix} \text{ et } \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix}$$

En conséquence...

... tous les résultats vus sur le modèle linéaire dans le chapitre sur la régression multiple s'appliquent ici, et notamment :

- **l'estimation** et les propriétés des estimateurs
- **les tests sur les paramètres**, ou les comb. lin. de paramètres
- **les tests de comparaison de modèles** en particulier le test de validité globale du modèle.
- **les résidus** leur loi et propriétés.

Plan

Problématique

On cherche à savoir si **le facteur** (la variable qualitative) à un effet sur **la réponse** (la variable quantitative).

=> Le test de validité globale du modèle permet de répondre à cette question.

Les deux hypothèses testées sont

- $\mathcal{H}_0 : \{\mu_1 = \mu_2 = \cdots = \mu_k\}$
- $\mathcal{H}_1 : \{\exists l, m \in \{1, 2, \cdots, k\} \text{ tels que } \mu_l \neq \mu_m\}$

Décomposition de la variance

Dans le cadre d'analyse de la variance, on a :

$$SCT = \sum_{l=1}^{k}\sum_{i=1}^{n_l}(Y_i^l - \overline{Y})^2 \quad SCM = \sum_{l=1}^{k} n_l(\overline{Y}_l - \overline{Y})^2 \quad SCR = \sum_{l=1}^{k}\sum_{i=1}^{n_l}(Y_i^l - \overline{Y}_l)^2$$

avec $\overline{Y}_l$ la moyenne pour le niveau l : $\overline{Y}_l = \frac{1}{n_l}\sum_{i=1}^{n_l} X_i^l$.

REMARQUE : L'espace engendré par les colonnes de $\mathbf{X}$ est de dimension k, le degré de liberté associé à SCM est donc $k-1$.

Sous l'hypothèse $\mathcal{H}_0$

$$\frac{SCM/k-1}{SCR/n-k} \sim \mathcal{F}_{k-1,n-k}$$

Table d'analyse de la variance

Source	Somme des carrés	Degrés de liberté	Som. carrés moyens	Statistique F	Proba. crit.
Facteur	SCM	$k-1$	$\frac{SCM}{k-1}$	$\frac{SCM/k-1}{SCR/n-k}$	p_c
Résidu	SCR	$n-k$	$\frac{SCR}{n-k}$		
Total	SCT	$n-1$			

Estimation de l'écart-type σ

$$S = \sqrt{\frac{SCR}{n-k}}$$

est un estimateur de l'écart-type σ.

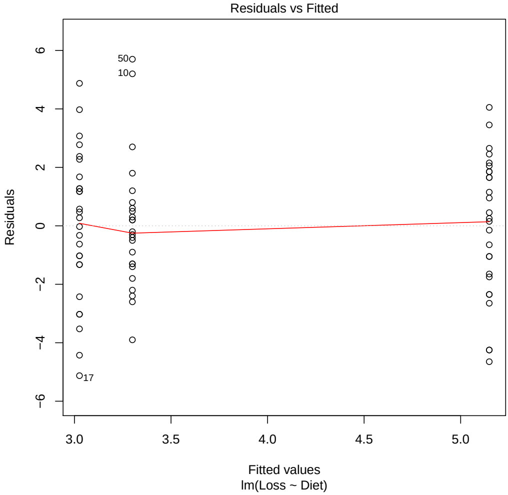
REMARQUES :

- On peut trouver l'estimation de l'écart-type dans la table d'analyse de la variance.
- La quantité $\frac{SCR}{n-k}$ est souvent nommée MSE pour *mean square error* dans les logiciels de statistiques.

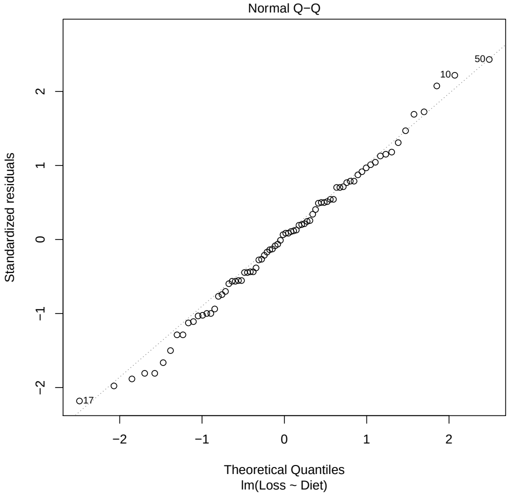
Illustration

```
> reg=lm(Loss~Diet,data=Diet)
> anova(reg)
Analysis of Variance Table

Response: Loss
          Df Sum Sq Mean Sq F value   Pr(>F)
Diet       2  71.09  35.547  6.1974 0.003229 **
Residuals 75 430.18   5.736
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Vérification des hypothèses : homoscédasticité



Vérification des hypothèses : normalité



Plan

Ecriture singulière

La plupart des logiciels de statistiques utilisent l'écriture suivante (écriture singulière)

$$Y_i^l = \mu + \alpha_l + \epsilon_i^l \quad \text{avec} \quad \epsilon_i^l \sim \mathcal{N}(0, \sigma^2) \text{ et ind.}$$

- μ est la moyenne générale.
- α_l est l'**effet** du niveau l du facteur

Dans ce cas la matrice $\mathbf{X}$ devient singulière, la première colonne est égale à la somme des k colonnes suivantes.

$$\mathbf{X} = \begin{pmatrix} \overbrace{\begin{matrix} 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 1 & 0 & \cdots & 0 \end{matrix}}^{k+1 \text{ col.}} & \left.\vphantom{\begin{matrix}1\\ \vdots\\ 1\end{matrix}}\right\} n_1 \text{ li.} \\ \begin{matrix} 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 1 & \cdots & 0 \end{matrix} & \left.\vphantom{\begin{matrix}1\\ \vdots\\ 1\end{matrix}}\right\} n_2 \text{ li.} \\ \begin{matrix} \vdots & \vdots & \vdots & \dots & \vdots \end{matrix} & \\ \begin{matrix} 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{matrix} & \left.\vphantom{\begin{matrix}1\\ \vdots\\ 1\end{matrix}}\right\} n_k \text{ li.} \end{pmatrix} \text{ et } \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{pmatrix}$$

Identifiabilité

Définition

Soit $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$ un modèle statistique. On dit que $\mathcal{P}$ est identifiable si et seulement si

$$P_{\theta_1} = P_{\theta_2} \Longrightarrow \theta_1 = \theta_2, \quad \text{pour tout } \theta_1, \theta_2 \in \Theta.$$

Exemple : Le modèle de Poisson

$$\{\mathcal{P}(\theta) \text{ tq } \theta \in \mathbb{R}^+\}$$

Cas de l'écriture singulière

Modèle

$$Y_i^l = \mu + \alpha_l + \epsilon_i^l \text{ avec } \epsilon_i^l \sim \mathcal{N}(0, \sigma^2), \text{ ind.}$$

Ici, les deux jeux de paramètres $(\mu, \alpha_1, \cdots, \alpha_l, \sigma)$ et $(\mu - 1, \alpha_1 + 1, \cdots, \alpha_l + 1, \sigma)$, pour une valeur fixée de $\mu, \alpha_1, \cdots, \alpha_l$ et σ conduisent à la même probabilité. $=>$ Le modèle n'est pas identifiable.

CONSÉQUENCE : On ne peut pas estimer $\mu, \alpha_1, \cdots, \alpha_l$.

Ajout de contraintes d'identifiabilité

Pour rendre le modèle identifiable, on va ajouter une contrainte sur une combinaison linéaire des paramètres $\mu, \alpha_1, \cdots, \alpha_l$. Par exemple :

$$\begin{cases} \sum_{l=1}^{k} n_l \alpha_l = 0 \qquad (1) \\ \alpha_1 = 0 \qquad (2) \end{cases}$$

Avec la contrainte (1), on estime

- μ par $\overline{y}$
- α_l par $\overline{y}_l - \overline{y}$

Avec la contrainte (2), on estime

- μ par $\overline{y}_1$
- α_l par $\overline{y}_l - \overline{y}_1$

Remarques :

- La contrainte (2) est celle utilisée par **R**.
- Si on ne connait pas les contraintes, il ne faut pas chercher à interpréter les tests, et les coefficients.

Tests sur les effets

De nombreux logiciels donnent la probabilité critique du test $\mathcal{H}_0 = \{\alpha_l = 0\}$ contre $\mathcal{H}_1 = \{\alpha_l \neq 0\}$.

Attention :
Selon la contrainte utilisée, la signification du test n'est pas la même.

- **Avec (2)** : Le test de $\{\alpha_1 = 0\}$ revient à tester $\{\mu_1 = 0\}$. "La moyenne du groupe 1 est nulle".
  Le test de $\{\alpha_2 = 0\}$ revient à tester $\{\mu_2 = \mu_1\}$. "La moyenne du groupe 2 est égale à la moyenne du groupe 1".
- **Avec (1)** : Le test de $\{\alpha_1 = 0\}$ revient à tester $\{\mu_1 = \mu\}$. "La moyenne du groupe 1 est égale à la moyenne générale".

Conclusion : Il est fortement déconseillé de se baser sur ces tests pour prendre des décisions sur le modèle.

Illustration avec R

```
> summary(reg)


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.3000     0.4889   6.750 2.72e-09 ***
DietB        -0.2741     0.6719  -0.408  0.68449
DietC         1.8481     0.6719   2.751  0.00745 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.395 on 75 degrees of freedom
Multiple R-squared:  0.1418,Adjusted R-squared:  0.1189
F-statistic: 6.197 on 2 and 75 DF,  p-value: 0.003229
```

Plan

Contraste

Définition

On appelle contraste L des k moyennes $\mu_1, \mu_2, \cdots, \mu_k$ la somme

$$L = \sum_{l=1}^{k} l_l \mu_l \text{ telle que } \sum_{l=1}^{k} l_l = 0.$$

EXEMPLES :

- $\mu_1 - \mu_2$: pour comparer μ_1 à μ_2
- $\mu_1 - 2\mu_2 + \mu_3$: pour comparer μ_2 à la moyenne de μ_1 et μ_3.

INTÉRET : On utilise les contrastes pour tester des écarts entre les niveaux d'un même facteur. Comme la somme des coefficients est nulle, les contrastes sont indépendants du choix des contraintes d'identifiabilité.

Estimation

Un estimateur sans biais de L est

$$\hat{L} = \sum_{l=1}^{l} l_l \hat{\mu}_l = \sum_{l=1}^{l} l_l \overline{Y}_l$$

Propriétés

On a

- $E(\hat{L}) = L$
- $V(\hat{L}) = \sigma^2 \sum_{l=1}^{k} \frac{l_l^2}{n_l}$
- $$\frac{\hat{L} - L}{S\sqrt{\sum_{l=1}^{k} \frac{l_l^2}{n_l}}} \sim \mathcal{S}_{n-k}$$

Tests sur les contrastes

Tests *a priori*

On sait *a priori* à quelle question doit répondre notre analyse. On définit le contraste en fonction de la problématique et on test $\mathcal{H}_0 = \{L = 0\}$ contre $\mathcal{H}_1 = \{L \neq 0\}$.

- Avantages : on réalise peu de tests
- Inconvénients : il faut à l'avance savoir ce que l'on veut tester

Comparaisons multiples *a posteriori*

On ne sait pas *a priori* ce que l'on cherche, on se trouve dans une démarche exploratoire. On teste tous les contrastes $\mu_l - \mu_{l'}$.

- Avantages : on n'a pas besoin d'avoir une question par avance.
- Inconvénients : tests multiples, on réalise $\frac{k(k-1)}{2}$ tests.

Exemple de tests *a priori*

On se demande si l'effet du régime C est significativement différent des deux autres, ou si l'effet du régime A est significativement différent du B.

```
> emreg=emmeans(reg,~Diet)
> c1=c(1,1,-2)
> c2=c(1,-1,0)
> contrast(emreg,list(Diet=cbind(c1,c2)))
 contrast estimate    SE df t.ratio p.value
 Diet.c1    -3.970 1.141 75  -3.481  0.0008
 Diet.c2     0.274 0.672 75   0.408  0.6845
```

- Avec le régime C, la perte de poids est significativement différente qu'avec la moyenne des deux autres régimes. Elle est plus grande.
- Il n'y a pas de différences significatives entre les régimes A et B.

Tests multiples

Soit une famille de m hypothèses de tests $\mathcal{H}_{0i}$ contre $\mathcal{H}_{1i}$, pour $i \in \{1, 2, \cdots, m\}$.

Definition

On appelle $FWER$ le *family wise error rate*, la probabilité de rejeter à tort au moins 1 fois une hypothèse $\mathcal{H}_{0i}$ sur les m tests réalisés.

Propriété

Si les m tests sont indépendants et tous de niveau α alors

$$FWER = 1 - (1 - \alpha)^m$$

$=>$ Démonstration

m	1	5	10	20	100
$FWER$	0.05	0.22	0.40	0.64	0.99

Conséquence : On ne contrôle plus le risque de première espèce.

Méthode de Bonferroni

On diminue le risque de première espèce α. On prend $\alpha' = \frac{\alpha}{m}$, avec m le nombre de tests à réaliser.

- Avantage : on diminue la probabilité de réaliser au moins une erreur de première espèce sur les m tests.
- Inconvénient : on diminue aussi la puissance du test. On aura des difficultés à repérer les groupes différents.

Etendue Studentisée

Définition

On suppose que $Z_1, Z_2, \cdots, Z_m \sim \mathcal{N}(0,1)$ sont m variables normales standardisées indépendantes. On suppose que $U \sim \chi^2_\nu$ est aussi indépendante des Z_i.

L'**étendue Studentisée** est la variable aléatoire :

$$Q_{m,\nu} = \frac{\max_i Z_i - \min_i Z_i}{\sqrt{U/\nu}}$$

Application au cas des comparaisons multiples

On suppose que $n_1 = n_2 = \cdots = n_k = r$.
Nous avons vu que

- $\frac{\overline{Y}_l - \mu_l}{\sigma/\sqrt{r}} \sim \mathcal{N}(0,1)$,
- $\overline{Y}_1, \overline{Y}_2, \cdots, \overline{Y}_k$ sont indépendantes.
- $\frac{(n-k)S^2}{\sigma^2} \sim \chi^2_{n-k}$

On en déduit donc que

$$\frac{(\max_l \overline{Y}_l - \min_l \overline{Y}_l) - (\mu_M - \mu_m)}{S/\sqrt{r}} \sim Q_{k,n-k}$$

avec μ_M (resp. μ_m) l'espérance du groupe M tel que $\mu_M = \max \overline{Y}_l$ (resp. m tel que $\mu_m = \min \overline{Y}_l$).

Procédure de comparaisons multiples de Tukey

On considère les $m = k(k-1)/2$ hypothèses $\mathcal{H}_{0ll'} = \{\mu_l = \mu_{l'}\}$,
Pour les $k(k-1)/2$ contrastes linéaires $\mu_l - \mu_{l'}$.

1. Calculer la différence
$$|\overline{Y}_l - \overline{Y}_{l'}|$$
2. Rejeter les hypothèses $\mathcal{H}_{0ll'} = \{\mu_l = \mu_{l'}\}$ si
$$|\overline{Y}_l - \overline{Y}_{l'}| > R_{crit}$$
avec
$$R_{crit} = Q_{k,n-k,1-\alpha} S/\sqrt{r}$$

Cette procédure permet de contrôler le $FWER$.

Justification de la procédure de Tukey

Si toutes les hypothèses $\mathcal{H}_{0ll'}$ sont vérifiées simultanément, alors,

$$\frac{(\max_l \overline{Y}_l - \min_l \overline{Y}_l)}{S/\sqrt{r}} \sim Q_{k,n-k}$$

$=>$ Démonstration

Justification de la procédure de Tukey

Si toutes les hypothèses $\mathcal{H}_{0ll'}$ sont vérifiées simultanément, alors,

$$\frac{(\max_l \overline{Y}_l - \min_l \overline{Y}_l)}{S/\sqrt{r}} \sim Q_{k,n-k}$$

=> Démonstration

La procédure permet de contrôler le $FWER$.

$$\begin{aligned} FWER &= P(\text{Au moins une des hypothèses } \mathcal{H}_{0ll'} \text{ est rejetée à tort.}) \\ &= \alpha \end{aligned}$$

=> Démonstration.

Justification de la procédure de Tukey

Lemme

Soient $Z_l = \overline{Y}_l - \mu_l$, pour $l \in \{1, 2, \cdots, k\}$ telles que

$$P((\max_l \overline{Z}_l - \min_l \overline{Z}_l) < R_{crit}) = 1 - \alpha,$$

alors,

$$P(|\overline{Z}_l - \overline{Z}_{l'}| < R_{crit}, \text{ pour tout } l, l') = 1 - \alpha.$$

Intervalle de confiance simultané

Définition

L'intervalle de confiance simultané pour les paramètres $\mu_1, \mu_2, \cdots, \mu_k$ est l'ensemble des points $\mu_{10}, \mu_{20}, \cdots, \mu_{k0}$, tels qu'aucune des $k(k-1)/2$ hypothèses de test $\mathcal{H}_{0ll'} = \{\mu_{l0} - \mu_{l'0}\}$ ne soit rejetée avec la procédure de test.

Probabilité critique ajustée

Définition

La probabilité critique ajustée du test $\mathcal{H}_{0ll'} = \{\mu_l = \mu_{l'}\}$ contre $\mathcal{H}_{1ll'} = \{\mu_l \neq \mu_{l'}\}$ est la plus petite valeur du risque de première espèce α telle que $\mathcal{H}_{0ll'}$ est rejetée par la procédure de test.

La région de rejet dépend de α

p_{adj} est telle que $R_{crit}(p_{adj}) = |\overline{y}_l - \overline{y}_{l'}|$.

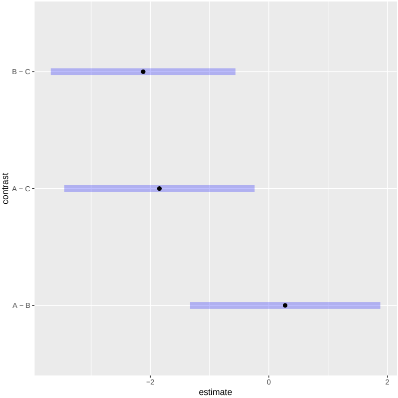
Illustration - Bonferroni

```
> contrast(emreg,"pairwise",adjust='bonferroni')
 contrast estimate    SE df t.ratio p.value
 A - B       0.274 0.672 75   0.408  1.0000
 A - C      -1.848 0.672 75  -2.751  0.0224
 B - C      -2.122 0.652 75  -3.256  0.0051

P value adjustment: bonferroni method for 3 tests
```

Illustration - Tukey

```
> contrast(emreg,"pairwise",adjust='tukey')
 contrast estimate    SE df t.ratio p.value
 A - B       0.274 0.672 75   0.408  0.9125
 A - C      -1.848 0.672 75  -2.751  0.0201
 B - C      -2.122 0.652 75  -3.256  0.0048

P value adjustment: tukey method for comparing a family of 3 e
```

Illustration - Tukey



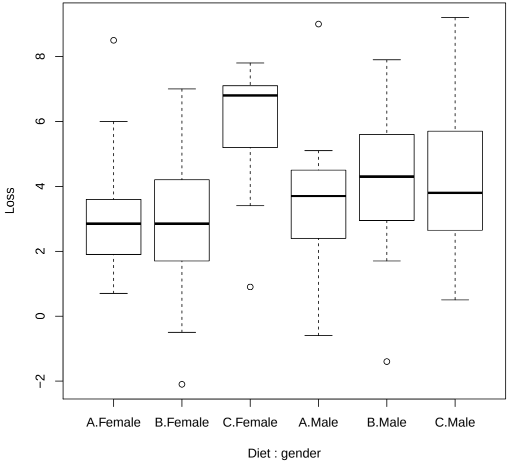
Plan du cours

Deux variables explicatives peuvent éventuellement expliquer la perte de poids :

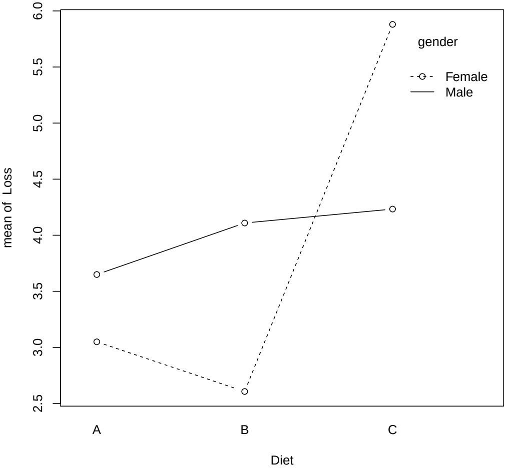
- le type de régime : A - B -C
- le sexe de la personne : homme - femme

On se demande si

- Le type de régime explique la perte de poids
- Le sexe explique la perte de poids
- Si l'effet du régime est le même quelque soit le sexe, c'est à dire si il y a ou non une interaction.


Loss
−2
0
2
4
6
8
A.Female
B.Female
C.Female
A.Male
B.Male
C.Male
Diet : gender

Graphe des interactions



Plan

Les données et le modèle

- On considère **2 facteurs**.
  - le premier facteur est indicé par i et a I niveaux,
  - le deuxième facteur est indicé par j et a J niveaux.
- Y_{ijk} est la k^e observation de la réponse dans le niveau i du facteur 1 et j du facteur 2.
- Il y a n_{ij} observations dans le niveau i du facteur 1 et le niveau j du facteur 2. On note

$$n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} \quad n_{i+} = \sum_{j=1}^{J} n_{ij} \quad n_{+j} = \sum_{i=1}^{I} n_{ij}$$

Modèle

On suppose de plus que

- toutes les variables suivent une distribution normale
- **l'espérance dépend du niveau i du facteur 1 et du niveau j du facteur 2**
- la variance est identique pour toutes les variables
- toutes les variables sont **indépendantes**.

$$Y_{ijk} \sim \mathcal{N}(\mu_{ij}, \sigma^2), \quad i.i.d.$$

Modèle

De manière équivalente, on pourra écrire que

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} \quad \text{avec} \quad \epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2) \text{ et ind.}$$

- μ_{ij} est l'espérance observée pour le croisement du niveau i du facteur 1 et du niveau j du facteur 2.

Sous forme matricielle

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

et

$$\mathbf{X} = \begin{pmatrix} \begin{array}{cccc} \multicolumn{4}{c}{\overbrace{\hspace{4cm}}^{I \times J \text{ col.}}} \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \end{array} & \begin{array}{l} \\ \left.\begin{array}{c} \\ \\ \\ \end{array}\right\} n_{11} \text{ li.} \\ \left.\begin{array}{c} \\ \\ \\ \end{array}\right\} n_{12} \text{ li.} \\ \\ \left.\begin{array}{c} \\ \\ \\ \end{array}\right\} n_{IJ} \text{ li.} \end{array} \end{pmatrix} \text{ et } \boldsymbol{\beta} = \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{IJ} \end{pmatrix}$$

Ecriture singulière

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \quad \text{avec} \quad \epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2) iid$$

- μ est le terme moyen
- α_i représente l'effet principal du niveau i du facteur 1
- β_j représente l'effet principal du niveau j du facteur 2
- γ_{ij} est le terme d'interaction, il modélise l'écart entre l'effet de la combinaison des niveaux i et j et la somme des effets de chacun des facteurs. En l'absence d'interactions, on suppose que pour tout couple (i, i') $\mu_{ij} - \mu_{i'j} = cte$ pour tout j.

Ajout de contraintes

L'écriture singulière permet de faire apparaître explicitement tous les effets, cependant dans cette écriture le modèle n'est pas identifiable. En effet, dans la première version, la matrice $\mathbf{X}$ est de rang IJ, alors que nous introduisons dans la version singulière :$1 + I + J + IJ$ coefficients. Nous allons donc devoir ajouter $1 + I + J$ contraintes sur les coefficients pour le rendre identifiable.

Contraintes naturelles

$$\sum_i n_{i+}\alpha_i = 0 \quad \sum_j n_{+j}\beta_j = 0 \quad \forall i : \sum_j n_{ij}\gamma_{ij} = 0 \quad \forall j : \sum_i n_{ij}\gamma_{ij} = 0$$

Avec ces contraintes les estimateurs sont les moyennes empiriques par sous groupes.

Contraintes de R

$$\alpha_1 = 0 \quad \beta_1 = 0 \quad \forall i : \gamma_{i1} = 0 \quad \forall j : \gamma_{1j} = 0$$

Illustration

```
> reg2=lm(Loss~Diet*gender, data=Diet)
> summary(reg2)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.0500     0.6197   4.922 5.49e-06 ***
DietB              -0.4429     0.8764  -0.505   0.6149
DietC               2.8300     0.8616   3.284   0.0016 **
genderMale          0.6000     0.9600   0.625   0.5340
DietB:genderMale    0.9019     1.3395   0.673   0.5030
DietC:genderMale   -2.2467     1.3145  -1.709   0.0919 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table d'analyse de la variance

On teste

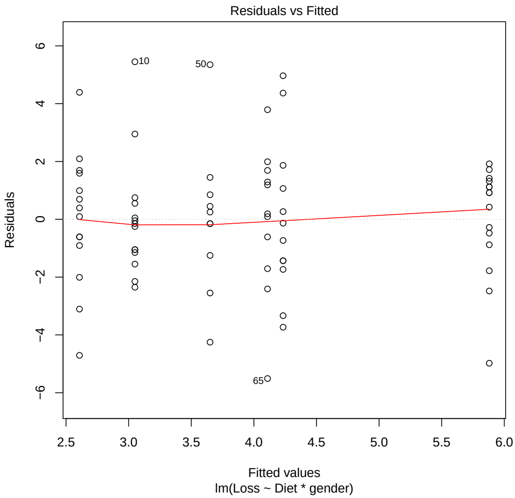
$$\mathcal{H}_0 = \{Y_{ijk} = \mu + \epsilon_{ijk}\} \quad \mathcal{H}_1 = \{Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}\}$$

Source	Somme des carrés	Degrés de liberté	Som. carrés moyens	Statistique F	Proba. crit.
Facteur	SCM	$IJ-1$	$\frac{SCM}{IJ-1}$	$\frac{SCM/IJ-1}{SCR/n-IJ}$	p_c
Résidu	SCR	$n-IJ$	$\frac{SCR}{n-IJ}$		
Total	SCT	$n-1$			

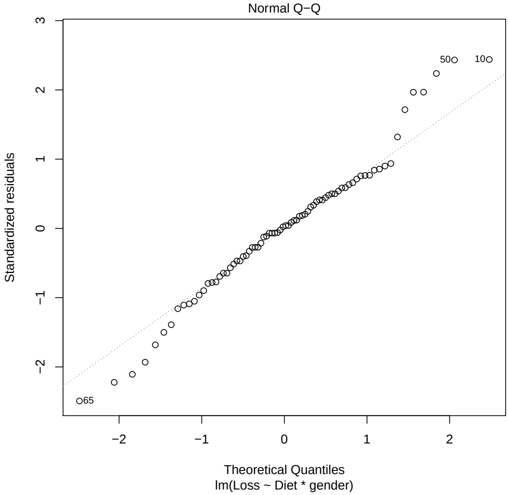
Illustration

```
> reg0=lm(Loss~1,data=Diet)
> anova(reg0,reg2)
Analysis of Variance Table

Model 1: Loss ~ 1
Model 2: Loss ~ Diet * gender
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     75 470.93
2     70 376.33  5      94.6 3.5193 0.006775 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

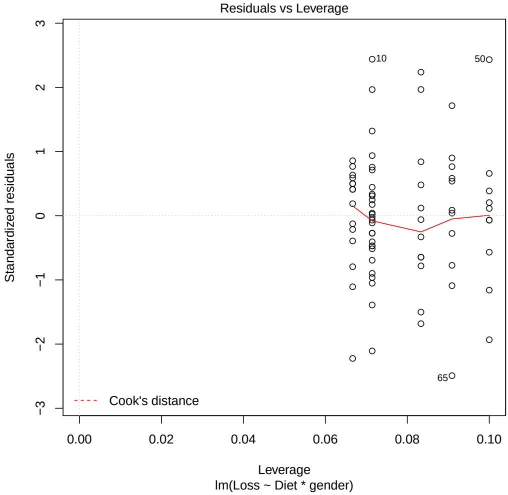
Vérification des hypothèses : homoscédasticité



Vérification des hypothèses : normalité



Vérification des hypothèses : normalité



PLan

Considérons les modèles suivant :

ModABi

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \quad \text{avec} \quad \epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2) iid$$

ModABa

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk} \quad \text{avec} \quad \epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2) iid$$

ModA

$$Y_{ijk} = \mu + \alpha_i + \epsilon_{ijk} \quad \text{avec} \quad \epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2) iid$$

ModB

$$Y_{ijk} = \mu + \beta_j + \epsilon_{ijk} \quad \text{avec} \quad \epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2) iid$$

La fonction `anova` fournit les résultats des tests de comparaisons de modèles (type I) comme suit :

Source	$\mathcal{H}_0$	$\mathcal{H}_1$	Df
Facteur 1	mod0	modA	I
Facteur 2	modA	modABa	J
Interaction	modABa	modABi	$IJ - I - J$

- Sur la ligne "Interaction", le modèle complet avec interaction est comparé au modèle à effets additifs. On teste l'effet de l'interaction sur le modèle.
- Sur la ligne "Facteur 2", on compare le modèle à effets additifs avec le modèle a un facteur : le facteur 1. On teste donc l'utilité du facteur 2, dans un modèle contenant déjà le facteur 1.
- Sur la ligne "Facteur 1", on compare le modèle a un facteur (le facteur 1) avec le modèle avec une constante.
- Remarque : l'ordre dans lequel sont introduits les facteurs ont une importance.

La fonction `Anova` fournit les résultats des tests de comparaisons de modèles (type II) comme suit :

Source	$\mathcal{H}_0$	$\mathcal{H}_1$	Df
Facteur 1	modB	modABa	I
Facteur 2	modA	modABa	J
Interaction	modABa	modABi	$IJ - I - J$

- L'ordre dans lequel sont introduits les facteurs ici n'ont pas d'importance. Les deux facteurs ont un rôle symétrique.

Illustration : effets de type I

```
    > anova(reg2)
Analysis of Variance Table

Response: Loss
            Df Sum Sq Mean Sq F value   Pr(>F)
Diet         2  60.53 30.2635  5.6292 0.005408 **
gender       1   0.17  0.1687  0.0314 0.859910
Diet:gender  2  33.90 16.9520  3.1532 0.048842 *
Residuals   70 376.33  5.3761
```

On note que l'effet du terme d'interaction est significatif.
L'effet du facteur `gender` ne serait pas significatif dans un modèle additif, cependant on ne supprime pas l'effet principal, si on maintient l'interaction. Cela reviendrait à dire, que le sexe n'a pas d'effet sur la perte de poids, mais qu'il interagit avec le type de régime, ce qui n'a pas de sens.

Illustration : effets de type II

```
> Anova(reg2)
Anova Table (Type II tests)

Response: Loss
             Sum Sq Df F value   Pr(>F)
Diet          60.42  2  5.6190 0.005456 **
gender         0.17  1  0.0314 0.859910
Diet:gender   33.90  2  3.1532 0.048842 *
Residuals    376.33 70
```

On note que que la ligne `Diet` a changé. Dans le tableau précédent (type I), l'effet est comparé à un modèle contenant uniquement une constante, ici on compare le modèle additif avec le modèle à un facteur `gender`, la différence des sommes de carrés est donc inférieure (60,42 contre 60,53).

Plan

1. Analyse de la variance à un facteur
   - Les données et le modèle
   - Le test d'anova = test de validité globale
   - Tests sur les effets
   - Test sur les contrastes
2. Analyse de la variance à deux facteurs
   - Les données et le modèle
   - Test des effets
   - **Tests sur les contrastes**
3. Analyse de la covariance : ANCOVA
   - les données et le modèle
   - test sur les effets
   - Test des effets

Tests sur les contrastes

On peut réaliser des tests sur les contrastes pour comparer :

- les niveaux d'un même facteur
- les combinaisons de niveaux dans l'interaction

On fera attention au fait que

- dans un modèle avec interaction l'interprétation des comparaisons de niveaux dans un facteur peut être délicat. Ce n'est pas parce qu'un niveau est supérieur aux autres qu'il l'est forcément pour tous les cas de l'interaction.
- on peut être amené à faire beaucoup de tests. On appliquera une correction pour les tests multiples.

```
  > emreg4=emmeans(reg2,~Diet)
NOTE: Results may be misleading due to involvement in interactions
> contrast(emreg4,"pairwise",adjust='bonferroni')
 contrast estimate    SE df t.ratio p.value
 A - B    -0.00812 0.670 70  -0.012  1.0000 #alpha_1 - alpha_2
 A - C    -1.70667 0.657 70  -2.597  0.0344 #alpha_1 - alpha_3
 B - C    -1.69855 0.648 70  -2.622  0.0322 #alpha_2 - alpha_3

Results are averaged over the levels of: gender
P value adjustment: bonferroni method for 3 tests
```

```
> emreg3=emmeans(reg2,~Diet|gender)
> contrast(emreg3,"pairwise",adjust='bonferroni')
gender = Female:
 contrast estimate    SE df t.ratio p.value
 A - B       0.443 0.876 70   0.505  1.0000 #mu_11 - mu_21
 A - C      -2.830 0.862 70  -3.284  0.0048 #mu_11 - mu_31
 B - C      -3.273 0.862 70  -3.798  0.0009 #mu_21 - mu_31

gender = Male:
 contrast estimate    SE df t.ratio p.value
 A - B      -0.459 1.013 70  -0.453  1.0000 #mu_12 - mu_22
 A - C      -0.583 0.993 70  -0.588  1.0000 #mu_12 - mu_32
 B - C      -0.124 0.968 70  -0.128  1.0000 #mu_22 - mu_32

P value adjustment: bonferroni method for 3 tests
```

```
  > emreg2=emmeans(reg2,~Diet*gender)
> contrast(emreg2,"pairwise",adjust='bonferroni')
 contrast           estimate    SE df t.ratio p.value
 A Female - B Female   0.443 0.876 70   0.505  1.0000 #mu_11 - mu_21
 A Female - C Female  -2.830 0.862 70  -3.284  0.0240 #mu_11 - mu_31
 A Female - A Male    -0.600 0.960 70  -0.625  1.0000
 A Female - B Male    -1.059 0.934 70  -1.134  1.0000
 A Female - C Male    -1.183 0.912 70  -1.297  1.0000
 B Female - C Female  -3.273 0.862 70  -3.798  0.0046
 B Female - A Male    -1.043 0.960 70  -1.086  1.0000
 B Female - B Male    -1.502 0.934 70  -1.608  1.0000
 B Female - C Male    -1.626 0.912 70  -1.783  1.0000
 C Female - A Male     2.230 0.947 70   2.356  0.3193
 C Female - B Male     1.771 0.920 70   1.924  0.8762
 C Female - C Male     1.647 0.898 70   1.834  1.0000
 A Male - B Male      -0.459 1.013 70  -0.453  1.0000
 A Male - C Male      -0.583 0.993 70  -0.588  1.0000
 B Male - C Male      -0.124 0.968 70  -0.128  1.0000

P value adjustment: bonferroni method for 15 tests
```

Situation du problème

- **En régression simple ou multiple** :

  Réponse $\Leftarrow$ Variable(s) explicative(s) quantitative(s)

- **En analyse de la variance** :

  Réponse $\Leftarrow$ Variable(s) explicative(s) qualitative(s)

Nous avons vu que tous ces modèles appartiennent au même cadre, celui du modèle linéaire :

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

Analyse de la covariance

Dans une analyse de la covariance, des variables quantitative(s) et qualitative(s) sont introduites ensembles pour expliquer la variable réponse.

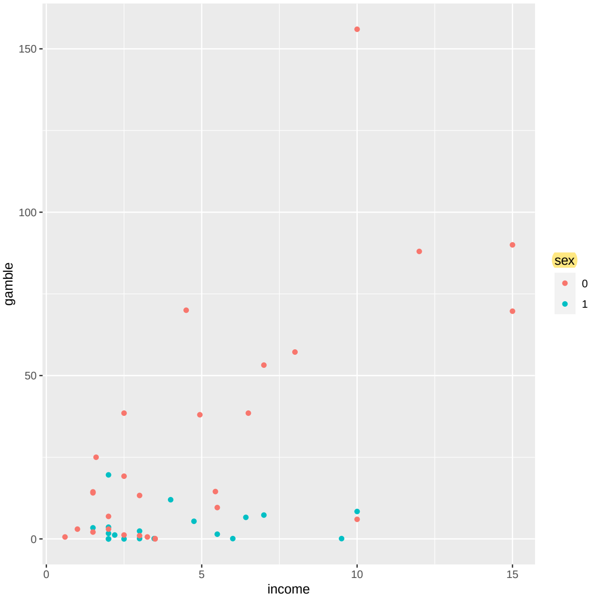
Exemple

On s'intéresse aux dépense des adolescents aux jeux de hasard, dans les années 1980. On dispose de plusieurs variables :

- **Gamble** la dépense annuelle au jeu de hasard (en livre par an)
- **Income** le revenu (en livre par semaine)
- **Sex** le sexe

On suppose qu'à la fois la fois, le sexe et le revenu peuvent avoir un impact sur la réponse.

```
> cor(tg[,c("income","gamble")])
          income    gamble
income 1.0000000 0.6220769
gamble 0.6220769 1.0000000
> by(tg[,c("income","gamble")], tg[,c("sex")],cor)
tg[, c("sex")]: 0
         income   gamble
income 1.000000 0.713669
gamble 0.713669 1.000000
---------------------------------------------------------
tg[, c("sex")]: 1
          income    gamble
income 1.0000000 0.0882356
gamble 0.0882356 1.0000000
```


150
100
50
0
gamble
0
5
10
15
income
sex
0
1

Plan

Les données et le modèle

On considère

- **1 facteur** (variable qualitative) indicé par i et a I niveaux,
- **1 variable quantitative** x.

On note

- Y_{ik} est la k^e observation de la réponse dans le niveau i du facteur,
- x_{ik} la valeur de la variable x pour la k^e observation dans le niveau i du facteur.

Il y a n_i observations dans le niveau i du facteur. On note

$$n = \sum_{i=1}^{I} n_i$$

Modèle

On suppose de plus que

- toutes les variables suivent une distribution normale
- **l'espérance dépend du niveau i du facteur 1 et de la variable x**
- la variance est identique pour toutes les variables
- toutes les variables sont **indépendantes**.

$$Y_{ik} \sim \mathcal{N}(\mu_i + \beta_i x_{ik}, \sigma^2), \quad i.i.d.$$

Modèle

De manière équivalente, on pourra écrire que

$$Y_{ik} = \mu_i + \beta_i x_{ik} + \epsilon_{ik} \quad \text{avec} \quad \epsilon_{ik} \sim \mathcal{N}(0, \sigma^2) \text{ et ind.}$$

- μ_i est l'ordonnée à l'origine de la droite de régression pour le niveau i
- β_i est la pente de la droite de régression pour le niveau i

Sous forme matricielle

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

et

$$\mathbf{X} = \begin{array}{cc} \begin{array}{cc} \overbrace{\hphantom{1 \quad 0 \quad \cdots \quad 0}}^{I \text{ col.}} & \overbrace{\hphantom{x_{11} \quad 0 \quad \cdots \quad 0}}^{I \text{ col.}} \end{array} & \\ \begin{pmatrix} 1 & 0 & \cdots & 0 & x_{11} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \cdots & 0 & x_{1n_1} & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & x_{21} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \cdots & 0 & 0 & x_{2n_2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & x_{I1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & x_{In_I} \end{pmatrix} & \begin{array}{l} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_1 \text{ li.} \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_2 \text{ li.} \\ \\ \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} n_I \text{ li.} \end{array} \end{array} \text{ et } \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_I \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_I \end{pmatrix}$$

Ecriture singulière

$$Y_{ik} = \mu + \alpha_i + \beta x_{ik} + \gamma_i x_{ik} + \epsilon_{ik} \quad \text{avec} \quad \epsilon_{ik} \sim \mathcal{N}(0, \sigma^2) iid$$

- μ est le terme moyen
- α_i représente l'effet principal du niveau i du facteur 1
- β représente l'effet principal de la variable quantitative
- γ_i est le terme d'interaction, entre le facteur et la variable quantitative. Il signifie que la pente de la droite de régression est différente selon le niveau du facteur.

Ajout de contraintes

Selon l'écriture singulière on doit estimer : $1 + I + 1 + I$ coefficients, alors que le rang de la matrice $\mathbf{X}$ est $2I$. On va donc introduire deux contraintes d'identifiabilités.

Contraintes de R

$$\alpha_1 = 0 \quad \beta_1 = 0$$

Illustration

```
> regtg=lm(sqrt(gamble)~income*sex,data=tg)
> summary(regtg)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.58018    0.61576   2.566   0.0138 *
income       0.56480    0.09632   5.864 5.77e-07 ***
sex1        -0.40215    1.09187  -0.368   0.7144
income:sex1 -0.48666    0.20906  -2.328   0.0247 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.045 on 43 degrees of freedom
Multiple R-squared:  0.5706,Adjusted R-squared:  0.5406
F-statistic: 19.04 on 3 and 43 DF,  p-value: 5.233e-08
```

Table d'analyse de la variance

On teste

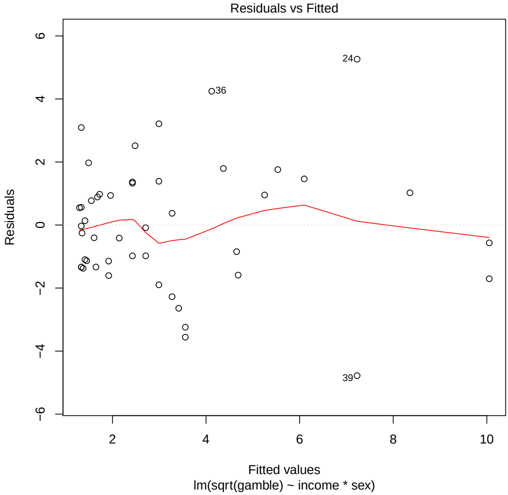
$$\mathcal{H}_0 = \{Y_{ik} = \mu + \epsilon_{ik}\} \quad \mathcal{H}_1 = \{Y_{ik} = Y_{ik} = \mu + \alpha_i + \beta x_{ik} + \gamma_i x_{ik} + \epsilon_{ik}\}$$

Source	Somme des carrés	Degrés de liberté	Som. carrés moyens	Statistique F	Proba. crit.
Facteur	SCM	$2I-1$	$\frac{SCM}{2I-1}$	$\frac{SCM/2I-1}{SCR/n-2I}$	p_c
Résidu	SCR	$n-2I$	$\frac{SCR}{n-2I}$		
Total	SCT	$n-1$			

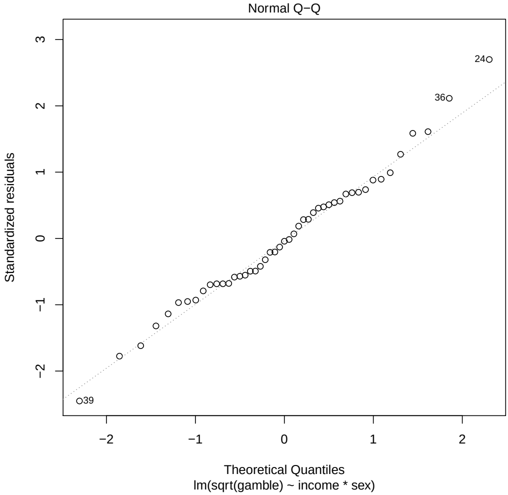
Illustration

```
> regtg0=lm(sqrt(gamble)~1,data=tg)
> anova(regtg0,regtg)
Analysis of Variance Table

Model 1: sqrt(gamble) ~ 1
Model 2: sqrt(gamble) ~ income * sex
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     46 418.92
2     43 179.90  3    239.02 19.044 5.233e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

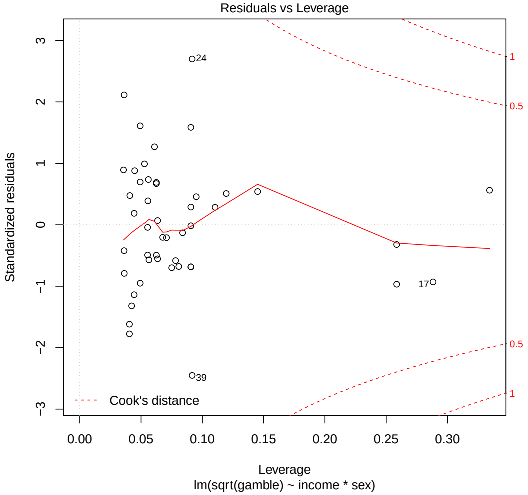
Vérification des hypothèses : homoscédasticité



Vérification des hypothèses : normalité



Vérification des hypothèses : points influents et abberants



PLan

Considérons les sous-modèles suivant :

ModABi

$$Y_{ik} = \mu + \alpha_i + \beta x_{ik} + \gamma_i x_{ik} + \epsilon_{ik} \quad \text{avec} \quad \epsilon_{ik} \sim \mathcal{N}(0, \sigma^2) iid$$

ModABp (parallèle)

$$Y_{ik} = \mu + \alpha_i + \beta x_{ik} + \epsilon_{ik} \quad \text{avec} \quad \epsilon_{ik} \sim \mathcal{N}(0, \sigma^2) iid$$

ModABo (rdonnée à l'origine)

$$Y_{ik} = \mu + \beta x_{ik} + \gamma_i x_{ik} + \epsilon_{ik} \quad \text{avec} \quad \epsilon_{ik} \sim \mathcal{N}(0, \sigma^2) iid$$

ModB

$$Y_{ik} = \mu + \beta x_{ik} + \epsilon_{ik} \quad \text{avec} \quad \epsilon_{ik} \sim \mathcal{N}(0, \sigma^2) iid$$

ModA

$$Y_{ik} = \mu + \alpha_i + \epsilon_{ik} \quad \text{avec} \quad \epsilon_{ik} \sim \mathcal{N}(0, \sigma^2) iid$$

Mod0

$$Y_{ik} = \mu \cancel{+ \alpha_i} + \epsilon_{ik} \quad \text{avec} \quad \epsilon_{ik} \sim \mathcal{N}(0, \sigma^2) iid$$

- Les modèles ModABo et ModABo et ModABp sont des cas particuliers de ModABi
  - Dans ModABp, les droites de régressions sont parallèles (même pente) mais l'ordonnée à l'orgine varie selon le niveau.
  - Dans ModABo, les droites de régressions ne sont pas parallèles mais ont la même ordonnée à l'orgine.
  - Dans ModABi, les droites de régressions ne sont pas parallèles et n'ont pas la même ordonnée à l'orgine.
- le modèle ModA est un cas particulier du modèle modABp. Dans ce modèle toutes les pentes sont nulles.
- le modèle ModB est un cas particulier du modèle modABo. Dans ce cas, toutes les pentes sont identiques mais non nécessairement nulles.
- le modèle ModO est un cas particulier de tous les autres modèles.

La fonction anova fournit les résultats des tests de comparaisons de modèles (effets de type I) comme suit :

Source	$\mathcal{H}_0$	$\mathcal{H}_1$	Df
Variable A	mod0	modA	I
Variable B	modA	modABp	$I+1-I$
Interaction	modABi	modABp	$2I-(I+1)$

REMARQUE : les degrés de liberté ne sont valables que si A est le facteur. Si A est le facteur :

- Sur la ligne "Interaction", le modèle complet avec interaction est comparé au modèle avec droites parallèle. On teste l'effet de l'interaction sur le modèle.
- Sur la ligne "Variable B", on compare le modèle où toutes les droites sont horizontales avec le modèle à droites parallèles.
- Sur la ligne "Variable A", on compare le modèle a plusieurs droites horizontales avec le modèle à une droite horizontale.
- Remarque : l'ordre dans lequel sont introduits les facteurs ont une importance.

La fonction `Anova` fournit les résultats des tests de comparaisons de modèles (effets de type II) comme suit :

Source	$\mathcal{H}_0$	$\mathcal{H}_1$	Df
Variable A	modB	modABp	$I+1-1$
Variable B	modA	modABp	$I+1-I$
Interaction	modABi	modABp	$2I-(I+1)$

Remarques :

- L'ordre dans lequel sont introduits les facteurs ici n'ont pas d'importance. Les deux facteurs ont un rôle symétrique.
- Les degrés de liberté ne sont valables que si A est le facteur.

Illustration : effets de type I

```
> anova(regtg2)
Analysis of Variance Table

Response: sqrt(gamble)
           Df  Sum Sq Mean Sq F value    Pr(>F)
sex         1  94.435  94.435 22.5719 2.278e-05 ***
income      1 121.914 121.914 29.1401 2.723e-06 ***
sex:income  1  22.670  22.670  5.4187    0.0247 *
Residuals  43 179.900   4.184
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On note que l'effet du terme d'interaction est significatif.

Illustration : effets de type II

```
> Anova(regtg2)
Anova Table (Type II tests)

Response: sqrt(gamble)
           Sum Sq Df F value    Pr(>F)
sex         70.19  1 16.7768 0.0001823 ***
income     121.91  1 29.1401 2.723e-06 ***
sex:income  22.67  1  5.4187 0.0246957 *
Residuals  179.90 43
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Pour finir

- La fonction emmeans permet de calculer les moyennes marginales, c'est à dire connaissant un ou plusieurs et leurs interactions.

```
> emmeans(regtg,~income|sex)
sex = 0:
 income emmean    SE df lower.CL upper.CL
   4.64   4.20 0.388 43    3.420     4.98

sex = 1:
 income emmean    SE df lower.CL upper.CL
   4.64   1.54 0.478 43    0.577     2.50

Results are given on the sqrt (not the response) scale.
Confidence level used: 0.95
```

- Il est également possible de pratiquer des tests de contraste sur les coefficients.

Intervalles de confiance pour $E(Y_{ik})$