Chapitre 4. Régression sur variables qualitatives

Cours de modèle linéaire gaussien par S. Donnet

Executive Master Statistique et Big-Data

Août 2020

Introduction

Introduction

- Variables explicatives quantitatives continues pour l'instant

Introduction

- Variables explicatives quantitatives continues pour l'instant
- Comment intégrer au modèle linéaire des variables qualitatives ?

Introduction

- Variables explicatives quantitatives continues pour l'instant
- Comment intégrer au modèle linéaire des variables qualitatives ?
- Nous allons d'abord étudier le cas d'une variable explicative qualitative seule. Il s'agit d'**analyse de la variance à un facteur** (ANOVA).

Introduction

- Variables explicatives quantitatives continues pour l'instant
- Comment intégrer au modèle linéaire des variables qualitatives ?
- Nous allons d'abord étudier le cas d'une variable explicative qualitative seule. Il s'agit d'**analyse de la variance à un facteur** (ANOVA).
- Nous verrons ensuite le cas de deux facteurs

Introduction

- Variables explicatives quantitatives continues pour l'instant
- Comment intégrer au modèle linéaire des variables qualitatives ?
- Nous allons d'abord étudier le cas d'une variable explicative qualitative seule. Il s'agit d'**analyse de la variance à un facteur** (ANOVA).
- Nous verrons ensuite le cas de deux facteurs
- Puis le cas d'un facteur associé à une variable quantitative, comme l'exemple ci-dessus avec `bloc` et `circ`. Dans ce dernier cas «mixte», on parle d'**analyse de la covariance** (ANCOVA).

Analyse de la variance à un facteur sur l'exemple ozone

- Concentration en ozone en fonction de la direction du vent

Analyse de la variance à un facteur sur l'exemple ozone

- Concentration en ozone en fonction de la direction du vent
- La variable `vent` a 4 modalités : (EST, NORD, OUEST, SUD)

Analyse de la variance à un facteur sur l'exemple ozone

- Concentration en ozone en fonction de la direction du vent
- La variable `vent` a 4 modalités : (EST, NORD, OUEST, SUD)
- Dix premières lignes du fichier de données `ozone.txt` :

Individu	max03	vent
1	64	E
2	90	N
3	79	E
4	81	N
5	88	O
6	68	S
7	139	E
8	78	N
9	114	S
10	42	O

Analyse de la variance à un facteur sur l'exemple ozone

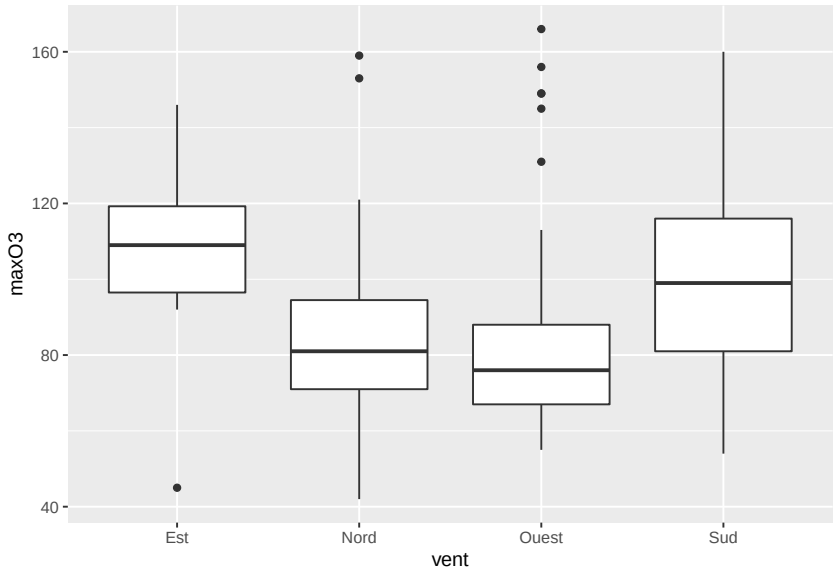
- Concentration en ozone en fonction de la direction du vent
- La variable `vent` a 4 modalités : (EST, NORD, OUEST, SUD)
- Dix premières lignes du fichier de données `ozone.txt` :

Individu	max03	vent
1	64	E
2	90	N
3	79	E
4	81	N
5	88	O
6	68	S
7	139	E
8	78	N
9	114	S
10	42	O

- Commencer par une représentation graphique des données : boites à moustaches (boxplots)

```
ggplot(ozone,aes(x=vent,y=max03)) + geom_boxplot()
```

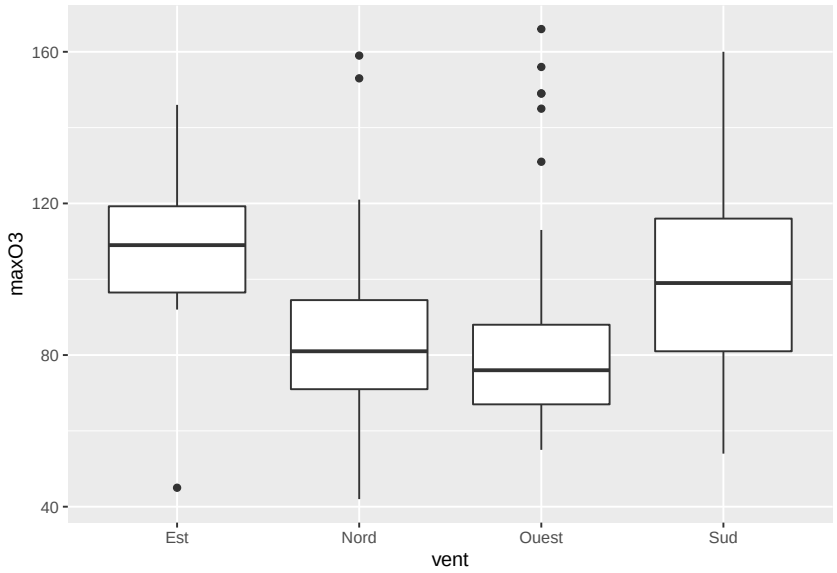
Analyse de la variance à un facteur sur l'exemple ozone



Boxplot de la variable vent.

- Le vent semble influencer sur la concentration en ozone.

Analyse de la variance à un facteur sur l'exemple ozone



Boxplot de la variable vent.

- Le vent semble influencer sur la concentration en ozone.
- Pour préciser : analyse de variance à un facteur explicatif : le vent.

Notations

- Supposons que la variable qualitative qui nous intéresse a I modalités.

Notations

- Supposons que la variable qualitative qui nous intéresse a I modalités.
- Chaque modalité est prise par n_i observations ($\sum_{i=1}^{I} n_i = n$).

Notations

- Supposons que la variable qualitative qui nous intéresse a I modalités.
- Chaque modalité est prise par n_i observations ($\sum_{i=1}^{I} n_i = n$).
  - **Exemple** : pour l'ozone, si on prend l'ordre alphabétique (est, ouest, nord, sud)=$(1, 2, 3, 4)$; on a

$$n_1 = 10, \quad n_2 = 31, \quad n_3 = 50, \quad n_4 = 21.$$

Notations

- Supposons que la variable qualitative qui nous intéresse a I modalités.
- Chaque modalité est prise par n_i observations ($\sum_{i=1}^{I} n_i = n$).
  - **Exemple** : pour l'ozone, si on prend l'ordre alphabétique (est, ouest, nord, sud)=$(1, 2, 3, 4)$; on a

$$n_1 = 10, \quad n_2 = 31, \quad n_3 = 50, \quad n_4 = 21.$$

- y_{ij} représente la valeur de la variable à expliquer pour le j-ème individu ayant pris la $i$ème modalité.

Notations

- Supposons que la variable qualitative qui nous intéresse a I modalités.
- Chaque modalité est prise par n_i observations ($\sum_{i=1}^{I} n_i = n$).
  - **Exemple** : pour l'ozone, si on prend l'ordre alphabétique (est, ouest, nord, sud)=$(1, 2, 3, 4)$; on a

$$n_1 = 10, \quad n_2 = 31, \quad n_3 = 50, \quad n_4 = 21.$$

- y_{ij} représente la valeur de la variable à expliquer pour le j-ème individu ayant pris la $i$ème modalité.
  - **Exemple** : pour l'ozone, , alors $y_{3,2}$ désigne la valeur de la concentration en ozone pour la 2ème journée où il y a eu un vent venant de l'ouest.

Modèle régulier

On suppose que y_{ij} est la réalisation de Y_{ij} où

$$Y_{ij} = \mu_i + \varepsilon_{ij} \quad i \in \{1, \dots I\},\ j \in \{1, \dots n_i\}$$

avec les mêmes hypothèses sur le bruit que précédemment, i.e.

$$\varepsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

σ^2 étant inconnu, les paramètres μ_i, pour $i \in \{1, \dots, I\}$ inconnus aussi et fixes.

Remarque

Ce modèle est dit régulier car dans sa version matricielle, la matrice de design X^r est de rang plein (I colonnes et rang I). En effet, définissons la matrice suivante :

$$X^r = \begin{pmatrix} 1 & & \\ \vdots & & \\ 1 & & \\ & 1 & \\ & \vdots & \\ & 1 & \\ & & 1 \\ & & \vdots \\ & & 1 \end{pmatrix}$$

Alors, on a l'écriture matricielle suivante :

$$\boldsymbol{Y} = X^r \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_I \end{pmatrix} + \varepsilon, \quad \varepsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbb{I}_n)$$

Estimation

- μ_i est estimé par

$$\widehat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} = \overline{y_{i\cdot}}$$

  c'est à dire la moyenne des observations dans la modalité i.
- Fitted values : $\hat{y}_{ij} = \hat{\mu}_i$.
- σ^2 s'estime par

$$\hat{\sigma}^2 = \frac{1}{n - I} \sum_{ij} (y_{ij} - \hat{\mu}_i)^2$$

  Cet estimateur est sans biais.

Exemple de l'ozone et du vent

```
reg <- lm(maxO3 ~ - 1 + vent ,data=ozone)
summary(reg)
```

```
##
## Call:
## lm(formula = maxO3 ~ -1 + vent, data = ozone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.600 -16.807  -7.365  11.478  81.300
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## ventEst    105.600      8.639   12.22   <2e-16 ***
## ventNord    86.129      4.907   17.55   <2e-16 ***
## ventOuest   84.700      3.864   21.92   <2e-16 ***
## ventSud    102.524      5.962   17.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.32 on 108 degrees of freedom
## Multiple R-squared:  0.9195, Adjusted R-squared:  0.9165
## F-statistic: 308.5 on 4 and 108 DF,  p-value: < 2.2e-16
```

On retrouve bien la sortie habituelles : les estimations, les écart-types, le R^2, les p-values des tests de Student et de Fisher.

Commentaires

- Le test de Fisher global a ici peu d'intérêt

Commentaires

- Le test de Fisher global a ici peu d'intérêt
- Car test de Fisher global :

$\mathcal{H}_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$ contre $\mathcal{H}_1$: l'un des μ_i n'est pas nul

Commentaires

- Le test de Fisher global a ici peu d'intérêt
- Car test de Fisher global :

  $\mathcal{H}_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$ contre $\mathcal{H}_1$: l'un des μ_i n'est pas nul

- Vraie question : «la provenance du vent a-t-elle une influence sur la concentration en ozone ?»

Commentaires

- Le test de Fisher global a ici peu d'intérêt
- Car test de Fisher global :

$\mathcal{H}_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$ contre $\mathcal{H}_1$: l'un des μ_i n'est pas nul

- Vraie question : «la provenance du vent a-t-elle une influence sur la concentration en ozone ?»
- Cette question se traduit par le test suivant :

$\mathcal{H}_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ contre $\mathcal{H}_1$: les μ_i ne sont pas tous égaux

Commentaires

- Le test de Fisher global a ici peu d'intérêt
- Car test de Fisher global :

$\mathcal{H}_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = 0$ contre $\mathcal{H}_1$: l'un des μ_i n'est pas nul

- Vraie question : «la provenance du vent a-t-elle une influence sur la concentration en ozone ?»
- Cette question se traduit par le test suivant :

$\mathcal{H}_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ contre $\mathcal{H}_1$: les μ_i ne sont pas tous égaux

- Pour remédier à cela, on introduit un effet moyen μ, un intercept !

Modèle singulier

On suppose que y_{ij} est la réalisation de Y_{ij} avec :

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i \in \{1, \dots I\},\ j \in \{1, \dots n_i\}$$

- On a introduit un effet moyen μ.

Modèle singulier

On suppose que y_{ij} est la réalisation de Y_{ij} avec :

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i \in \{1, \dots I\},\ j \in \{1, \dots n_i\}$$

- On a introduit un effet moyen μ.
- Pour chaque i, α_i l'effet de la $i$ème modalité sur la variable à expliquer y par rapport à l'effet moyen : $\mu_i = \mu + \alpha_i$ se lit aussi $\alpha_i = \mu_i - \mu$.

Modèle singulier

On suppose que y_{ij} est la réalisation de Y_{ij} avec :

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i \in \{1, \dots I\},\ j \in \{1, \dots n_i\}$$

- On a introduit un effet moyen μ.
- Pour chaque i, α_i l'effet de la $i$ème modalité sur la variable à expliquer y par rapport à l'effet moyen : $\mu_i = \mu + \alpha_i$ se lit aussi $\alpha_i = \mu_i - \mu$.
- μ est vu comme un niveau de référence et α_i s'entend comme une différence par rapport au niveau de référence.

Ecriture matricielle

- $I + 1$ paramètres.

Ecriture matricielle

- $I + 1$ paramètres.
-

$$X = \begin{pmatrix} 1 & 1 & & \\ 1 & \vdots & & \\ 1 & 1 & & \\ 1 & 1 & & \\ 1 & & \vdots & \\ 1 & & 1 & \\ 1 & & & 1 \\ 1 & & & \vdots \\ 1 & & & 1 \end{pmatrix}$$

Ecriture matricielle

- $I + 1$ paramètres.
- $$X = \begin{pmatrix} 1 & 1 & & \\ 1 & \vdots & & \\ 1 & 1 & & \\ 1 & 1 & & \\ 1 & & \vdots & \\ 1 & & 1 & \\ 1 & & & 1 \\ 1 & & & \vdots \\ 1 & & & 1 \end{pmatrix}$$
- Pas de rang plein $\Rightarrow$ Problème d'identifiabilité.

Ecriture matricielle

- $I + 1$ paramètres.
- $$X = \begin{pmatrix} 1 & 1 & & \\ 1 & \vdots & & \\ 1 & 1 & & \\ 1 & 1 & & \\ 1 & & \vdots & \\ 1 & & 1 & \\ 1 & & & 1 \\ 1 & & & \vdots \\ 1 & & & 1 \end{pmatrix}$$
- Pas de rang plein $\Rightarrow$ Problème d'identifiabilité.
- i.e. $\mu_i = \mu + \alpha_i$, $\forall i = 1, \ldots, I$. I équations mais $I + 1$ paramètres.

Ecriture matricielle

- $I + 1$ paramètres.
- $$X = \begin{pmatrix} 1 & 1 & & \\ 1 & \vdots & & \\ 1 & 1 & & \\ 1 & 1 & & \\ 1 & & \vdots & \\ 1 & & 1 & \\ 1 & & & 1 \\ 1 & & & \vdots \\ 1 & & & 1 \end{pmatrix}$$
- Pas de rang plein $\Rightarrow$ Problème d'identifiabilité.
- i.e. $\mu_i = \mu + \alpha_i$, $\forall i = 1, \ldots, I$. I équations mais $I + 1$ paramètres.
- Nécéssité d'introduire une contrainte sur les paramètres

Ecriture matricielle

- $I + 1$ paramètres.
- $$X = \begin{pmatrix} 1 & 1 & & \\ 1 & \vdots & & \\ 1 & 1 & & \\ 1 & 1 & & \\ 1 & & \vdots & \\ 1 & & 1 & \\ 1 & & & 1 \\ 1 & & & \vdots \\ 1 & & & 1 \end{pmatrix}$$
- Pas de rang plein $\Rightarrow$ Problème d'identifiabilité.
- i.e. $\mu_i = \mu + \alpha_i$, $\forall i = 1, \ldots, I$. I équations mais $I + 1$ paramètres.
- Nécéssité d'introduire une contrainte sur les paramètres
- Cette contrainte est aussi associée à la question : qu'appelle-t-on l'effet moyen ?

Rendre le modèle identifiable en le contraignant

- Contrainte $\mu = 0$: modèle régulier
- Contrainte $\alpha_1 = 0$: $\mu = \mu_1$: effet moyen est l'effet de la première modalité. Les autres $\alpha_i = \mu_i - \mu_1$ mesurent donc les différences des autres modalités avec cette modalité de référence.
- Contrainte $\sum_{i=1}^{I} \alpha_i = 0$: l'effet moyen est alors la moyenne des effets de chaque modalité

$$\begin{aligned} \mu + \alpha_i &= \mu_i, \forall i = 1, \ldots, I \\ \Rightarrow \sum_{i=1}^{I} \mu + \sum_{i=1}^{I} \alpha_i &= \sum_{i=1}^{I} \mu_i \\ \text{d'où } \mu &= \frac{1}{I} \sum_{i=1}^{I} \mu_i \end{aligned}$$

- Contrainte linéaire quelconque en utilisant l'option `contrasts` ou `C(...)`.
- Par défaut c'est la contrainte $\alpha_1 = 0$ qui est utilisée dans R.

Estimation des paramètres du modèle singulier

Contrainte $\alpha_1 = 0$.

- Par identification $\mu_1 = \mu + \alpha_1$ d'où $\mu_1 = \mu$ (car $\alpha_1 = 0$).

$$\hat{\mu} = \hat{\mu}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j} = \overline{y}_{1.}$$

- De même $\mu_2 = \mu + \alpha_2$, d'où $\alpha_2 = \mu_2 - \mu$ et donc

$$\hat{\alpha}_2 = \hat{\mu}_2 - \hat{\mu} = \frac{1}{n_2} \sum_{j=1}^{n_2} y_{2j} - \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j} = \overline{y}_{2.} - \overline{y}_{1.}$$

- Pour tout $i = 2, \ldots, I$

$$\hat{\alpha}_i = \hat{\mu}_i - \hat{\mu} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} - \frac{1}{n_1} \sum_{j=1}^{n_1} y_{1j} = \overline{y}_{i.} - \overline{y}_{1.}$$

Si on change de contrainte, on change d'interprétation.

Commentaires

- Mêmes résultats en terme de prédiction et tests quelque soit l'écriture du modèle !
- **La différence est dans l'interprétation des coefficients et donc des tests**.
- Exemple : avec la contrainte $\alpha_1 = 0$.
  - Coefficient de l'intercept= effet de la première modalité.
  - Autres coefficients = les différences d'effet avec cette modalité.
  - Le test de Fisher associé à ce modèle a pour hypothèse $H_0 : \alpha_2 = \alpha_3 = \cdots = \alpha_J = 0$
  - correspond donc bien à l'égalité des effets de chaque modalité.
- Vérifions avec le test global de Fisher l'influence de la provenance du vent sur la concentration en ozone :

Expérience numérique avec R

```
reg_sing <- lm(maxO3 ~  vent ,data=ozone)
summary(reg_sing)
```

Expérience numérique avec R

```
##
## Call:
## lm(formula = maxO3 ~ vent, data = ozone)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.600 -16.807  -7.365  11.478  81.300
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  105.600      8.639  12.223   <2e-16 ***
## ventNord     -19.471      9.935  -1.960   0.0526 .
## ventOuest    -20.900      9.464  -2.208   0.0293 *
## ventSud       -3.076     10.496  -0.293   0.7700
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.32 on 108 degrees of freedom
## Multiple R-squared:  0.08602,    Adjusted R-squared:  0.06063
## F-statistic: 3.388 on 3 and 108 DF,  p-value: 0.02074
```

Commentaires

- Petite p-value (0.02074) donc $\mathcal{H}_0$ est rejetée : la provenance du vent a bien une influence sur la concentration en ozone.
- R considère l'ordre alphabétique `Est` = 1ère modalité, `Nord`= 2ème, `Ouest` = 3ème, `Sud` = 4ème.
- Même résultats entre les deux contraintes : on peut retrouver les résultats de la première sortie (qui utilisait la contrainte $\mu = 0$) :

```
ventEst     105.600
ventNord     86.129
ventOuest    84.700
ventSud     102.524
```

à partir des résultats de la seconde (avec la contrainte $\alpha_1 = 0$) :

```
 (Intercept)  105.600
ventNord      -19.471
ventOuest     -20.900
ventSud        -3.076
```

Autres tests

On peut effectuer tous les types de test étudiés précédemment (par exemple $\alpha_1 = 1$, $\mu = 3$, $\alpha_1 = \alpha_2$ etc).

Changement de contraintes

- Pour prendre la seconde modalité comme référence :

```
> lm(maxO3 ~ C(vent, base = 2), data = ozone)
```

Changement de contraintes

- Pour prendre la seconde modalité comme référence :

  ```
  > lm(maxO3 ~ C(vent, base = 2), data = ozone)
  ```

- Si on veut utiliser la contrainte $\sum_{j=1}^{J} \alpha_j = 0$, il suffit d'utiliser

  ```
  > lm(maxO3 ~ C(vent, sum), data = ozone)
  ```

Exercice

- On considère la régression de `max03` sur `Vent` avec la contrainte par défaut (cf. sortie précédente).
- Que peut-on dire à partir des résultats des tests de Student ?
- En observant les boxplots, on pourrait penser que les vents du Nord et de l'Ouest ont le même effet.
- Faire le test correspondant pour vérifier ces deux hypothèses.

Remarque sur les variations totale, inter-groupes, et intra-groupes

- En utilisant la décomposition

$$y_{ij} - y_{..} = y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - -\bar{y}_{..}$$

avec $y_{..} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} y_{ij}$

Remarque sur les variations totale, inter-groupes, et intra-groupes

- En utilisant la décomposition

$$y_{ij} - y_{..} = y_{ij} - \bar{y}_{i.} + \bar{y}_{i.} - -\bar{y}_{..}$$

avec $y_{..} = \frac{1}{n}\sum_{i=1}^{I}\sum_{j=1}^{n_i} y_{ij}$

- On peut montrer que

$$\underbrace{\sum_{i,j}(y_{ij} - y_{..})^2}_{\text{Variation totale}} = \underbrace{\sum_{i=1}^{I} n_i(\bar{y}_{i.} - y_{..})^2}_{\text{Variation inter-groupes}} + \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - -\bar{y}_{i.})^2}_{\text{Variation intra-groupes}}.$$

(c'est la formule SCT = SCM + SCR)

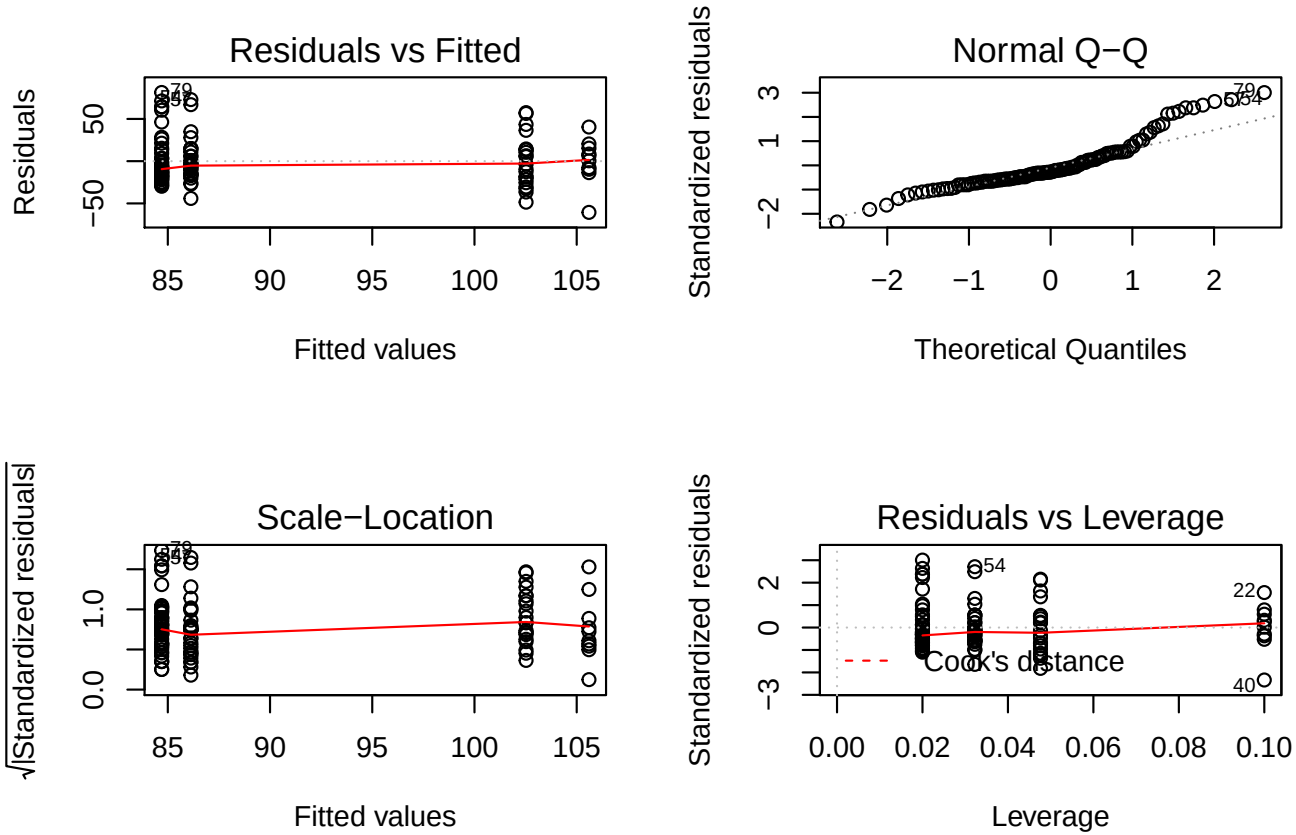
Analyse des résidus

Analyse sur les résidus standard.

```
par(mfrow=c(2,2))
plot(reg_sing)
```

Analyse des résidus

Pour les données Ozone



On constate qu'on n'a que 4 valeurs de $\hat{y}_{ij}$. C'est normal puisqu'on a vu qu'on prédisait y_{ij} par $\hat{\mu} + \hat{\alpha}_i$ donc on a bien I valeurs possibles.

Comparaison de deux traitements

- Différence significative sur les valeurs de y entre les modalités i et i' du facteur ?
- Traduction en terme de test : $\mathcal{H}_0 : \mu_i = \mu_{i'}$ versus $\mathcal{H}_1 : \mu_i \neq \mu_{i'}$.
- Statistique

$$T = \frac{\hat{\mu}_i - \hat{\mu}_{i'}}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n_i} + \frac{1}{n_{i'}} \right)}} \sim_{\mathcal{H}_0} \mathcal{T}_{n-I}$$

- Construire un test de niveau α : évident

Problème des tests multiples

- Comparer les autres groupes (appelés *traitements*) entre eux.
- Identifier tous les couples $(i, i')|\mu_i \neq \mu_{i'}$: $I(I-1)/2$ tests.
- Si on fait tous les tests au niveau δ :
- Pour tout (i, i'), test de

$$\mathcal{H}_0^{ii'} : \mu_i = \mu_{i'} \quad \text{versus} \quad \mathcal{H}_0^{ii'} : \mu_i \neq \mu_{i'}$$

- Pour chaque test, on contrôle de probabilité de rejeter $\mathcal{H}_0^{ii'}$ alors que $\mathcal{H}_0^{ii'}$ est vraie (probabilité $\leq \delta$).
- Probabilité de se tromper au moins une fois :

$$\begin{aligned} & \mathbb{P}\left(\text{rejeter au moins une } \mathcal{H}_0^{ii'} | \mathcal{H}_0^{ii'} \text{ vraie}\right) \\ \leq & \sum_{i,i',i<i'} \mathbb{P}_{\mathcal{H}_0^{ii'}}(\text{rejeter } \mathcal{H}_0^{ii'}) \\ \leq & \sum_{i,i',i<i'} \delta = \frac{I(I-1)}{2}\delta \end{aligned}$$

Correction de Bonferroni

- Si $I = 7$ et $\delta = 5\%$, on borne la probabilité de se tromper au moins une fois par 1 ! Donc on n'a aucun contrôle.
- Méthodes de contrôle : méthode de Bonferroni
  - Recorriger le niveau de chaque test.
  - Chaque test sera fait avec un niveau $2\delta/(I(I-1))$,
  - Niveau global δ.
  - Plus dur de rejeter les hypothèses nulles.
  - Cela revient à multiplier les p-values par $\frac{I(I-1)}{2}$

Code R pour comparer les traitements

Sans correction de Bonferroni

```
comp.statut = pairwise.t.test(ozone$max03,ozone$vent,
p.adjust.method = "none") comp.statut
```

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data:  ozone$max03 and ozone$vent
##
## Est   Nord Ouest
##    Nord  0.053   -       -
##   Ouest  0.029 0.819   -
##    Sud   0.770 0.036 0.014
##
## P value adjustment method: none
```

Code R pour comparer les traitements

Avec correction de Bonferroni

```
pairwise.t.test(ozone$max03,ozone$vent,
p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  ozone$max03 and ozone$vent
##
##       Est   Nord  Ouest
## Nord  0.316 -     -
## Ouest 0.176 1.000 -
## Sud   1.000 0.216 0.082
##
## P value adjustment method: bonferroni
```

Analyse de la variance à deux facteurs

- Deux facteurs : on modélise la concentration en ozone par le `vent` (4 modalités) et le `temps` (2 modalités : pluie/sec).

Analyse de la variance à deux facteurs

- Deux facteurs : on modélise la concentration en ozone par le `vent` (4 modalités) et le `temps` (2 modalités : pluie/sec).
- On pourra en fait généraliser les résultats à un nombre quelconque de facteurs. Mais afin d'alléger les notations on se restreint à 2 facteurs

Modélisation

- Supposons que le premier facteur ait I modalités et le second J modalités.
- y_{ijk} : la k-ème observation dans les modalités i pour le premier facteur et j pour le deuxième facteur.
- On suppose qu'on a n_{ij} observations pour les modalités i et j.
- y_{ijk} est la réalisation de Y_{ijk} :
-

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad k \in \{1, \ldots, n_{ij}\}, j \in \{1, \ldots, J\}, i \in \{1, \ldots, I\}$$

- avec

$$\varepsilon_{ijk} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

et les μ_{ij} sont les paramètres inconnus (et fixes).

Matrice correspondant au modèle a IJ colonnes et est de rang IJ.

Exemple sur l'Ozone

Par exemple, pour l'ozone : si on considère que le premier facteur est vent et le second est pluie et que (pluie, sec)=(1,2) et (est, ouest, nord, sud)=(1,2,3,4) alors y_{125} est la 5ère journée au cours de laquelle il a plu et le vent venait de l'ouest.

Estimation des paramètres du modèle régulier

Posons :

$$\overline{y}_{ij.} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk}$$

c'est donc la moyenne des valeurs de y sur les observations ayant pris la modalité j pour le premier facteur et la modalité k pour le second facteur.

Proposition

$$\hat{\mu}_{ij} = \overline{y}_{ij.}$$

Par conséquent :

$$\hat{\sigma}^2 = \frac{1}{n - IJ} \sum_{i,j,k} (y_{ijk} - \hat{\mu}_{ij})^2$$

Modèle singulier

Pour mieux analyser l'influence des facteurs, nous allons considérer la décomposition suivante de μ_{ij} :

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

- Ecriture n'est pas unique
- $I \times J$ paramètres μ_{ij} décomposés en $1 + I + J + J \times I$ nouveaux paramètres
- $1 + J + I$ paramètres de plus.
- Il faut imposer $1 + I + J$ contraintes linéairement indépendantes.

Contraintes classiques

- Contrainte de type analyse par cellule (=modèle régulier)

$$\mu = 0, \qquad \forall i \quad \alpha_i = 0, \qquad \forall j \quad \beta_j = 0$$

- Contrainte de type cellule de référence

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

Contrainte par défaut sous R

- Contrainte de type "somme"

$$\sum_{i=1}^{I} \alpha_i = 0 \qquad \sum_{j=1}^{J} \beta_j = 0, \qquad \forall j \qquad \sum_{i=1}^{I} \gamma_{ij} = 0, \qquad \forall i \quad \sum_{j=1}^{J} \gamma_{ij} = 0.$$

Interprétation des nouveaux coefficients

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

- α_i : l'effet principal de la modalité i du premier facteur
- β_j : l'effet principal de la modalité j du second facteur
- γ_{ij} : l'interaction entre les modalités i et j du premier et second facteur respectivement.
- μ : l'effet moyen (dépendant comme précédemment des contraintes imposées)

Expérience sur Ozone : lecture des coefficients

```
mod1  <-  lm(maxO3~vent*temps,data=ozone)
summary(mod1)
```

On trouve le même résultat en utilisant la formule

```
lm(maxO3~vent+temps+vent:temps,data=ozone)
```

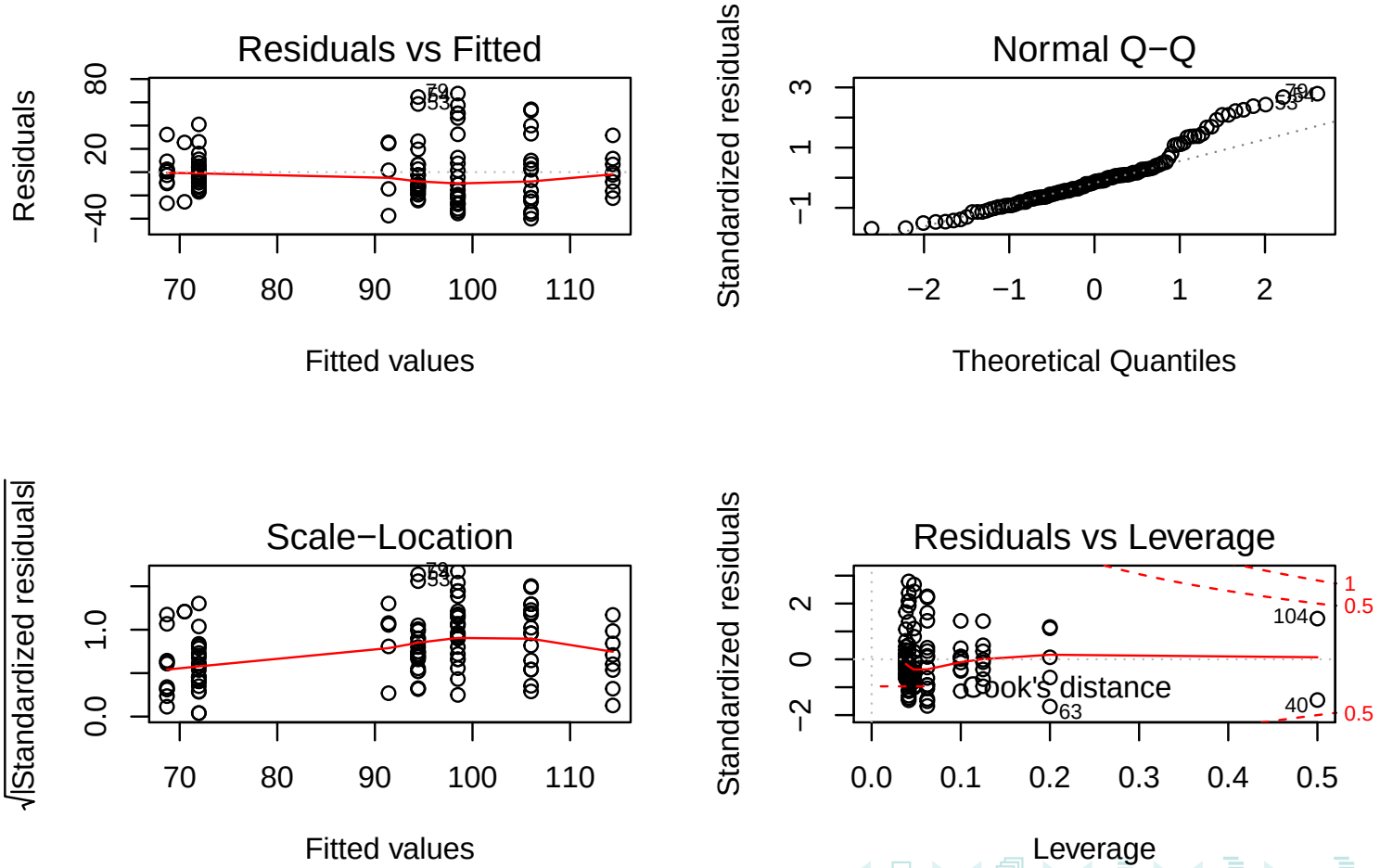
Sorties

```
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                70.500     17.464   4.037 0.000104 ***
## ventNord                   -1.800     19.131  -0.094 0.925221
## ventOuest                   1.462     18.123   0.081 0.935881
## ventSud                    20.900     20.664   1.011 0.314161
## tempsSec                   43.875     19.526   2.247 0.026749 *
## ventNord:tempsSec         -18.146     21.709  -0.836 0.405138
## ventOuest:tempsSec        -17.337     20.739  -0.836 0.405117
## ventSud:tempsSec          -29.275     23.267  -1.258 0.211138
## ---
```

- Facteur 1 : `vent`. Modalité 1 : `Est`.
- Facteur 2 : `temps`. Modalité 1 : `Pluie`.

Analyse des résidus

```
par(mfrow=c(2,2))
plot(mod1)
```



Test du modèle

$$\mathcal{H}_0 : Y_{ijk} = \mu + \varepsilon_{ijk} \quad \text{versus} \quad \mathcal{H}_1 : Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

Résultat du test donné directement dans le summary.

```
##
## Call:
## lm(formula = maxO3 ~ vent * temps, data = ozone)
...
## Multiple R-squared:  0.2807, Adjusted R-squared:  0.2322
## F-statistic: 5.797 on 7 and 104 DF,  p-value: 1.092e-05
```

La p-valeur est bien $< 5\%$ donc au moins un des facteurs a un effet. Par ailleurs, on remarquera que R^2 est très faible.

Une collection de modèles

$$
\begin{array}{rcll}
\mathcal{M}_{\mu} & : & Y_{ijk} & = \mu + \varepsilon_{ijk} \\
\mathcal{M}_{\mu,\alpha} & : & Y_{ijk} & = \mu + \alpha_i + \varepsilon_{ijk} \\
\mathcal{M}_{\mu,\beta} & : & Y_{ijk} & = \mu + \beta_j + \varepsilon_{ijk} \\
\mathcal{M}_{\mu,\alpha,\beta} & : & Y_{ijk} & = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \\
\mathcal{M}_{\mu,\alpha,\beta,\gamma} & : & Y_{ijk} & = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}
\end{array}
$$

- $\mathcal{M}_{\mu,\alpha,\beta,\gamma}$: modèle complet. Prend en compte tous les phénomènes mais comporte beaucoup de paramètres.
- $\mathcal{M}_{\mu,\alpha,\beta}$: modèle *additif*. Il suppose l'absence d'interaction entre les facteurs.

Test des interactions

$$\mathcal{H}_0 : Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \quad \text{versus} \quad \mathcal{H}_1 : Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

- Test d'un modèle emboîté (ou sous-modèle) : test de Fisher
- $$F = \frac{(\text{SCR}_{\mu,\alpha,\beta} - \text{SCR}_{\mu,\alpha,\beta,\gamma})/(\mathbf{rg}\left(X^{(\mu,\alpha,\beta,\gamma)}\right) - \mathbf{rg}(X^{(\mu,\alpha,\beta)}))}{SCR^{(\mu,\alpha,\beta,\gamma)}/(n - \mathbf{rg}\left(X^{(\mu,\alpha,\beta,\gamma)}\right))}$$
- $F \sim \mathcal{F}((I-1)(J-1), n-IJ)$
- On rejette $\mathcal{H}_0$ si $F > q_{(I-1)(J-1),n-IJ,\alpha}$

Test des interactions : application

```
mod2 <- lm(maxO3~vent + temps,data=ozone)
anova(mod1,mod2)
```

```
## Analysis of Variance Table
##
## Model 1: maxO3 ~ vent * temps
## Model 2: maxO3 ~ vent + temps
##   Res.Df   RSS Df Sum of Sq    F Pr(>F)
## 1    104 63440
## 2    107 64446 -3   -1006.4 0.55 0.6493
```

- $\mathcal{H}_0$ conservée, c'est-à-dire qu'on considère qu'il n'y a pas d'interaction.
- Nous sommes donc ramenés au modèle

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

Test des facteurs

Si on a pu enlever l'interaction, alors on définit le modèle sans l'interaction et on teste l'effet de chaque facteur restant.

> Si le test des interactions rejette l'hypothèse nulle, alors il n'y a pas de sens à tester les facteurs séparément (il ne peut y avoir d'interaction entre les facteurs sans les facteurs).

Test des facteurs (1)

Dans l'exemple Ozone, nous souhaitons savoir si la variable `temps` est pertinente dans le modèle mod2.
Nous souhaitons donc comparer les deux modèles :

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \quad \text{et} \quad Y_{ijk} = \mu + \alpha_i + \varepsilon_{ijk}$$

```
mod2 <- lm(maxO3~vent + temps,data=ozone)
mod3 <- lm(maxO3~ vent ,data=ozone)
anova(mod3,mod2)
```

Test des facteurs

```
## Analysis of Variance Table
##
## Model 1: maxO3 ~ vent
## Model 2: maxO3 ~ vent + temps
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    108 80606
## 2    107 64446  1     16159 26.829 1.052e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On rejette donc l'hypothèse $\mathcal{H}_0$ c'est-à-dire que la pluie a un effet sur le taux d'ozone, c'est-à-dire on choisit le modèle `mod2`.

Test des facteurs (2)

On peut ensuite tester l'influence de la variable `vent` sur le taux d'ozone dans le modèle `mod2` :

```
mod4 <- lm(maxO3~ temps ,data=ozone)
anova(mod4,mod2)
```

```
## Analysis of Variance Table
##
## Model 1: maxO3 ~ temps
## Model 2: maxO3 ~ vent + temps
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    110 68238
## 2    107 64446  3    3791.3 2.0982 0.1048
```

On est plus mitigé sur le rôle de la provenance du vent (une fois qu'on a introduit la variable `temps`).

Passage à 3 facteurs ?

- Si on veut faire une Anova à 3 facteurs on procèdera de la même façon.
- Pour des raisons d'interprétation et par soucis de parcimonie, pas d'interactions entre trois facteurs.
- On introduira chacun des 3 facteurs et seulement les interactions 2 à 2 :

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + c_k + \rho_{ij} + \nu_{ik} + \gamma_{jk} + \varepsilon_{ijkl}$$

Exercice

1. Charger le fichier nommé "kidiq.txt". Ce fichier contient les variables `kid_score`, `mom_hs` et `mom_work`. La première est quantitative et les secondes sont qualitatives. On va expliquer le score d'un enfant au test du quotient intellectuel par le fait que la mère ait eu l'équivalent du bac (1 si oui) et par le fait qu'elle travaille plus ou moins (valeurs de 1 à 4). Afficher ses variables, vérifier qu'il s'agit bien de variables qualitatives.
2. Faire les tests précédents sur l'interaction et les effets prinicipaux.
3. Donner la prévision du score d'un enfant si sa mère a un diplôme de second degré et travaille beaucoup (i.e. couple de modalités (1,4)).

Analyse de la covariance

- Analyse de la covariance : mélange de variables explicatives **qualitatives et quantitatives**.

Analyse de la covariance

- Analyse de la covariance : mélange de variables explicatives **qualitatives et quantitatives**.
- Exemple avec une quantitative x et un facteur :

Analyse de la covariance

- Analyse de la covariance : mélange de variables explicatives **qualitatives et quantitatives**.
- Exemple avec une quantitative x et un facteur :
  - Si on a I modalités pour le facteur,

Analyse de la covariance

- Analyse de la covariance : mélange de variables explicatives **qualitatives et quantitatives**.
- Exemple avec une quantitative x et un facteur :
  - Si on a I modalités pour le facteur,
  - Si on suppose que n_i observations ont pris la $i$ème modalité,

Analyse de la covariance

- Analyse de la covariance : mélange de variables explicatives **qualitatives et quantitatives**.
- Exemple avec une quantitative x et un facteur :
  - Si on a I modalités pour le facteur,
  - Si on suppose que n_i observations ont pris la $i$ème modalité,
  - alors on rassemble ces obervations qu'on note $(y_{ij})_{1 \leq i \leq n_j}$ pour la variable y

Analyse de la covariance

- Analyse de la covariance : mélange de variables explicatives **qualitatives et quantitatives**.
- Exemple avec une quantitative x et un facteur :
  - Si on a I modalités pour le facteur,
  - Si on suppose que n_i observations ont pris la $i$ème modalité,
  - alors on rassemble ces obervations qu'on note $(y_{ij})_{1 \leq i \leq n_j}$ pour la variable y
  - et $(x_{ij})_{1 \leq i \leq n_j}$ pour la variable x.

Analyse de la covariance

- Analyse de la covariance : mélange de variables explicatives **qualitatives et quantitatives**.
- Exemple avec une quantitative x et un facteur :
  - Si on a I modalités pour le facteur,
  - Si on suppose que n_i observations ont pris la $i$ème modalité,
  - alors on rassemble ces obervations qu'on note $(y_{ij})_{1 \leq i \leq n_j}$ pour la variable y
  - et $(x_{ij})_{1 \leq i \leq n_j}$ pour la variable x.
- On a alors, possiblement, une régression sur la variable quantitative x différente selon les niveaux du facteur :

Analyse de la covariance

- Analyse de la covariance : mélange de variables explicatives **qualitatives et quantitatives**.
- Exemple avec une quantitative x et un facteur :
  - Si on a I modalités pour le facteur,
  - Si on suppose que n_i observations ont pris la $i$ème modalité,
  - alors on rassemble ces obervations qu'on note $(y_{ij})_{1\leq i\leq n_j}$ pour la variable y
  - et $(x_{ij})_{1\leq i\leq n_j}$ pour la variable x.
- On a alors, possiblement, une régression sur la variable quantitative x différente selon les niveaux du facteur :
- On suppose que y_{ij} est la réalisation de Y_{ij} avec

$$Y_{ij} = b_i + a_i x_{ij} + \varepsilon_{ij}, \quad 1 \leq j \leq n_i, 1 \leq i \leq I \quad \varepsilon_{ij} \sim_{i.i.d.} \mathcal{N}(0, \sigma^2)$$

Analyse de la covariance

- Analyse de la covariance : mélange de variables explicatives **qualitatives et quantitatives**.
- Exemple avec une quantitative x et un facteur :
  - Si on a I modalités pour le facteur,
  - Si on suppose que n_i observations ont pris la $i$ème modalité,
  - alors on rassemble ces obervations qu'on note $(y_{ij})_{1\leq i\leq n_j}$ pour la variable y
  - et $(x_{ij})_{1\leq i\leq n_j}$ pour la variable x.
- On a alors, possiblement, une régression sur la variable quantitative x différente selon les niveaux du facteur :
- On suppose que y_{ij} est la réalisation de Y_{ij} avec

$$Y_{ij} = b_i + a_i x_{ij} + \varepsilon_{ij}, \quad 1 \leq j \leq n_i, 1 \leq i \leq I \quad \varepsilon_{ij} \sim_{i.i.d.} \mathcal{N}(0, \sigma^2)$$

- C'est ce qui correspond au modèle complet (le plus complexe) : Autant de droites de régression que de modalités du facteur.

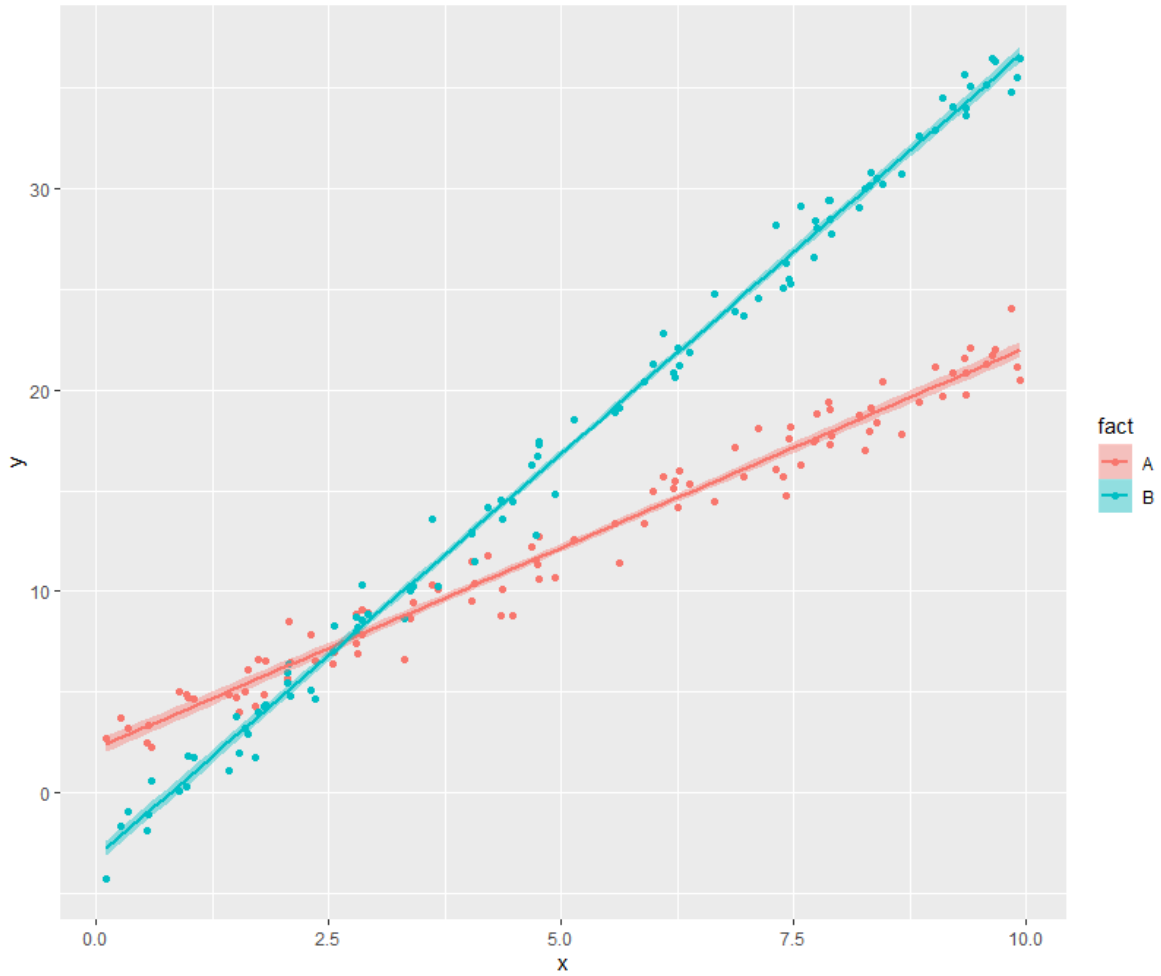
Analyse de la covariance

- Analyse de la covariance : mélange de variables explicatives **qualitatives et quantitatives**.
- Exemple avec une quantitative x et un facteur :
  - Si on a I modalités pour le facteur,
  - Si on suppose que n_i observations ont pris la $i$ème modalité,
  - alors on rassemble ces obervations qu'on note $(y_{ij})_{1\leq i\leq n_j}$ pour la variable y
  - et $(x_{ij})_{1\leq i\leq n_j}$ pour la variable x.
- On a alors, possiblement, une régression sur la variable quantitative x différente selon les niveaux du facteur :
- On suppose que y_{ij} est la réalisation de Y_{ij} avec

$$Y_{ij} = b_i + a_i x_{ij} + \varepsilon_{ij}, \quad 1 \leq j \leq n_i, 1 \leq i \leq I \quad \varepsilon_{ij} \sim_{i.i.d.} \mathcal{N}(0, \sigma^2)$$

- C'est ce qui correspond au modèle complet (le plus complexe) : Autant de droites de régression que de modalités du facteur.
- Attention, le σ^2 est commun à toutes les données.

Illustration

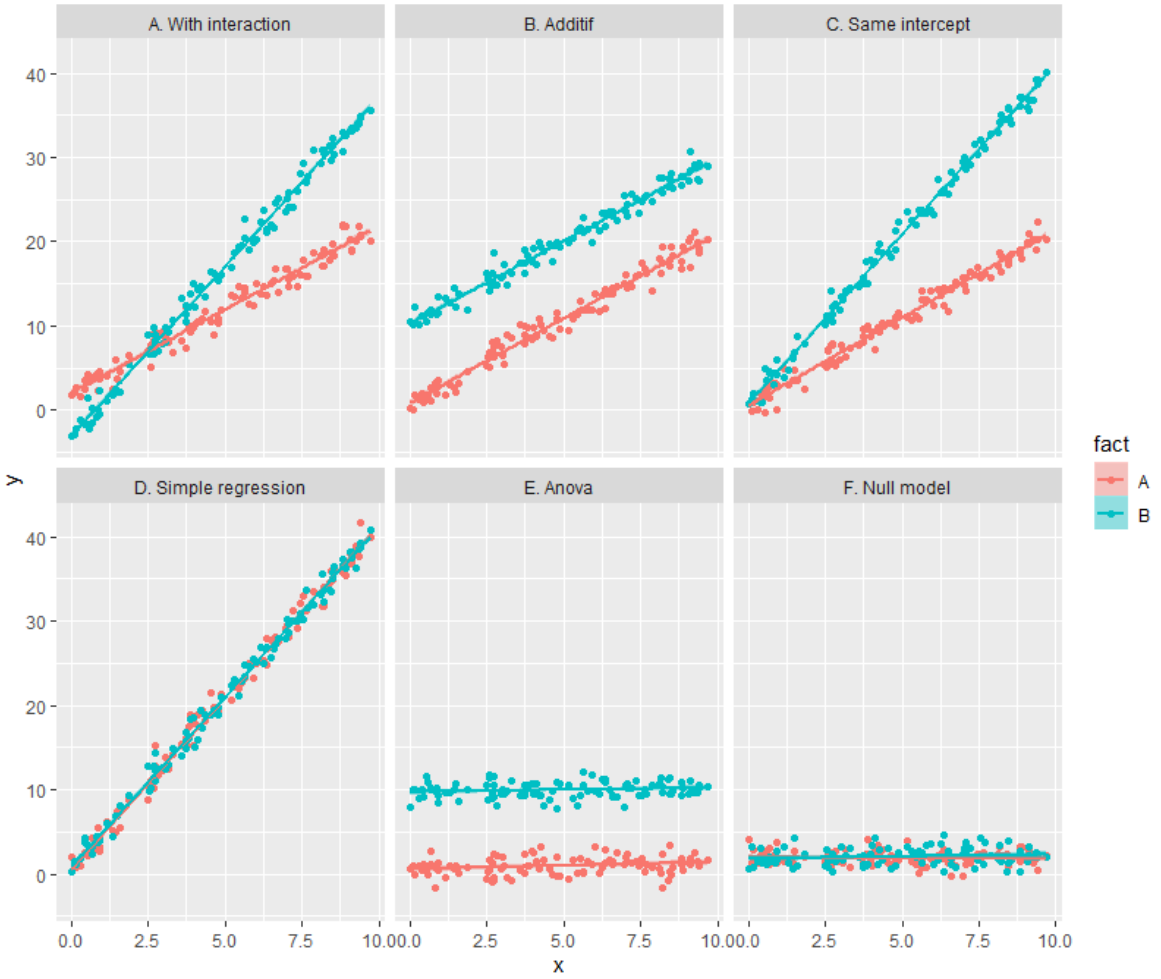


Modèle singulier

$$\begin{array}{rcccccc} Y_{ij} & = & b_i & + & a_i & x_{ij} & +\varepsilon_{ij} \\ Y_{ij} & = & (\mu+\beta_i) & + & (\alpha+\gamma_i) & x_{ij} & +\varepsilon_{ij} \end{array}$$

où γ_j est l'interaction entre le facteur et la variable quantitative.

Sous-modèles



Sous-modèles

- A. modèle complet : $Y_{ij} = \mu + \beta_i + (\alpha + \gamma_i)x_{ij} + \varepsilon_{ij}$
- B. interactions sont nulles (modèle additif) :

$$Y_{ij} = \mu + \beta_i + \alpha x_{ij} + \varepsilon_{ij}$$

- C. $\beta_i = 0$ pour tout j. Pentes sont différentes mais même ordonnées à l'origine

$$Y_{ij} = \mu + (\alpha + \gamma_i)x_{ij} + \varepsilon_{ij}$$

- D. $\beta_i = \gamma_i = 0$ pour tout i, le facteur n'a pas d'effet :

$$Y_{ij} = \mu + \alpha\, x_{ij} + \varepsilon_{ij}$$

- E. $\alpha_i = \gamma_i = 0$: la covariable n'a pas d'effet :

$$Y_{ij} = \mu + \beta_i + \varepsilon_{ij}$$

- F. $\beta_i = \alpha_i = \gamma_i = 0$ modèle nul :

$$Y_{ij} = \mu + \varepsilon_{ij}$$

Inférence

Toute l'inférence se passe exactement comme précédemment.

- On peut montrer que les EMC des paramètres (a_j, b_j) sont obtenus en faisant la régression de y contre x pour les données dans la modalité j.
- Les paramètres du modèle singulier sont non-identifiables. Il faut imposer des contraintes. **R** impose comme attendu $\beta_1 = \gamma_1 = 0$. On obtient l'estimation des paramètres du modèle singulier à partir de celle des (a_i, b_i).
- Notons $\hat{y}_{ij} = \hat{b}_i + \hat{a}_i x_{ij}$ la prédiction, alors σ^2 est estimé sans biais par

$$\hat{\sigma}^2 = \frac{1}{n-2I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_{ij})^2$$

- Les tests du facteur et de la covariable seront faits en comparant les SCR des différents modèles en compétition. On prendra en compte le nombre de paramètres dans chaque modèle. On peut utiliser la commande `anova`

Code R : modèle d'Ancova et test de l'interaction

```r
mod1 = lm(ht~circ*bloc,data=euca) #le  modèle complet
summary(mod1)
```

Sorties

```
##
## Call:
## lm(formula = ht ~ circ * bloc, data = euca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7723 -0.7037  0.0539  0.8114  3.3255
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.031e+00  2.895e-01  31.199   <2e-16 ***
## circ         2.576e-01  6.043e-03  42.618   <2e-16 ***
## blocA2      -1.850e-01  4.044e-01  -0.457    0.647
## blocA3       6.165e-01  4.766e-01   1.294    0.196
## circ:blocA2 -1.927e-05  8.454e-03  -0.002    0.998
## circ:blocA3 -6.834e-03  9.802e-03  -0.697    0.486
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Modèle sans interaction et test

```
mod2 = lm(ht~bloc+circ,data=euca)# modèle sans interaction
anova(mod2,mod1)
```

```
## Analysis of Variance Table
##
## Model 1: ht ~ bloc + circ
## Model 2: ht ~ circ * bloc
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1425 2005.9
## 2   1423 2005.0  2   0.84752 0.3007 0.7403
```

Conclusion

- Modèle linéaire : modèle souple pour prendre en compte variable qualitatives et quantitatives
- Contrôle des erreurs, sous l'hypothèse que les données sont la réalisation du modèle