Régression sur variables qualitatives
Analyse de la variance

Magalie Fromont

ENSAI Deuxième année - Modèles de régression

2010-2011

Introduction

Dans ce chapitre, on étudie des cas particuliers de régression linéaire où :

- La variable à expliquer Y est quantitative,
- La ou les variables explicatives potentielles sont qualitatives.

$\hookrightarrow$ très fréquent en pratique.

Exemples des données du Cirad et d'Air Breizh.

Analyse de la variance à un facteur

Cas d'une variable à expliquer quantitative Y et d'une variable explicative qualitative potentielle.

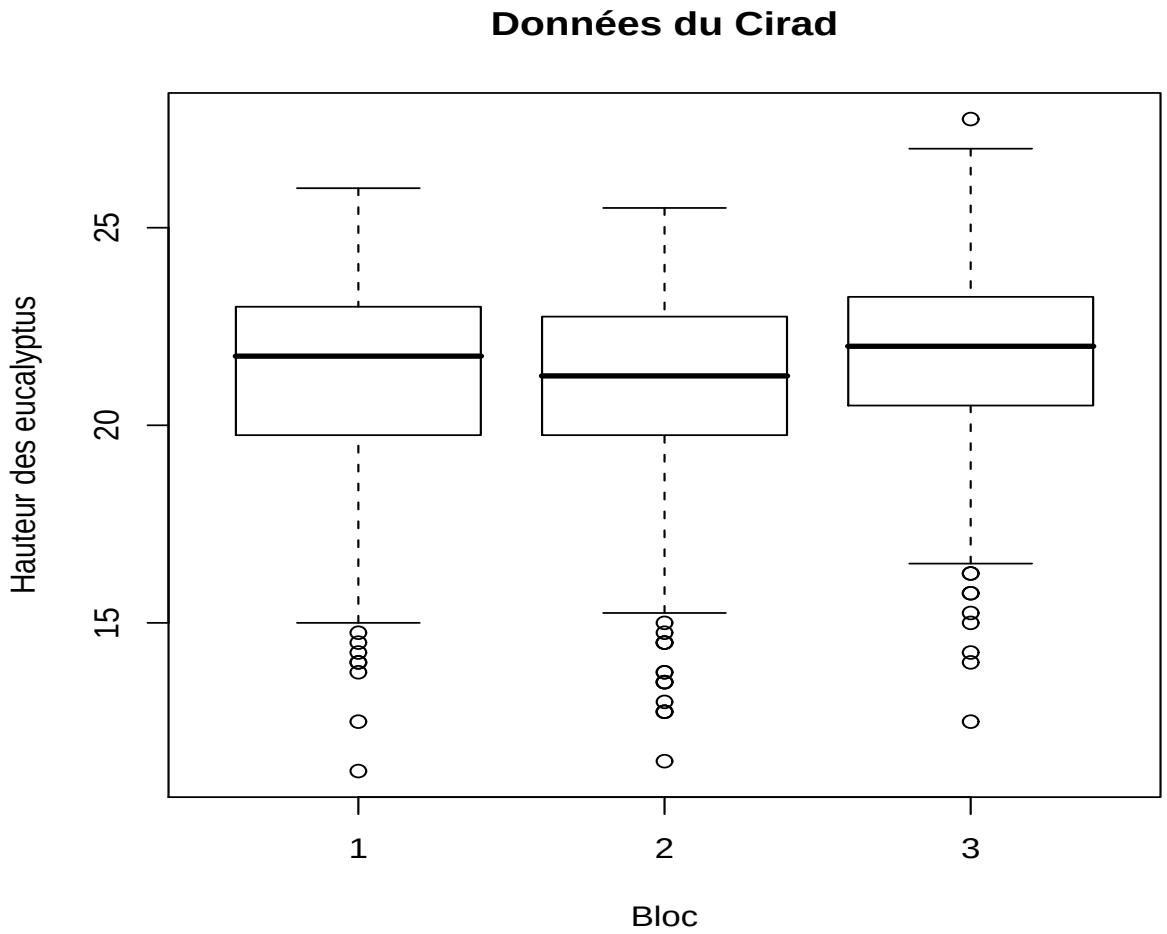
Objectif : déterminer si la variable explicative qualitative considérée a un effet - significatif - sur Y.

Définition

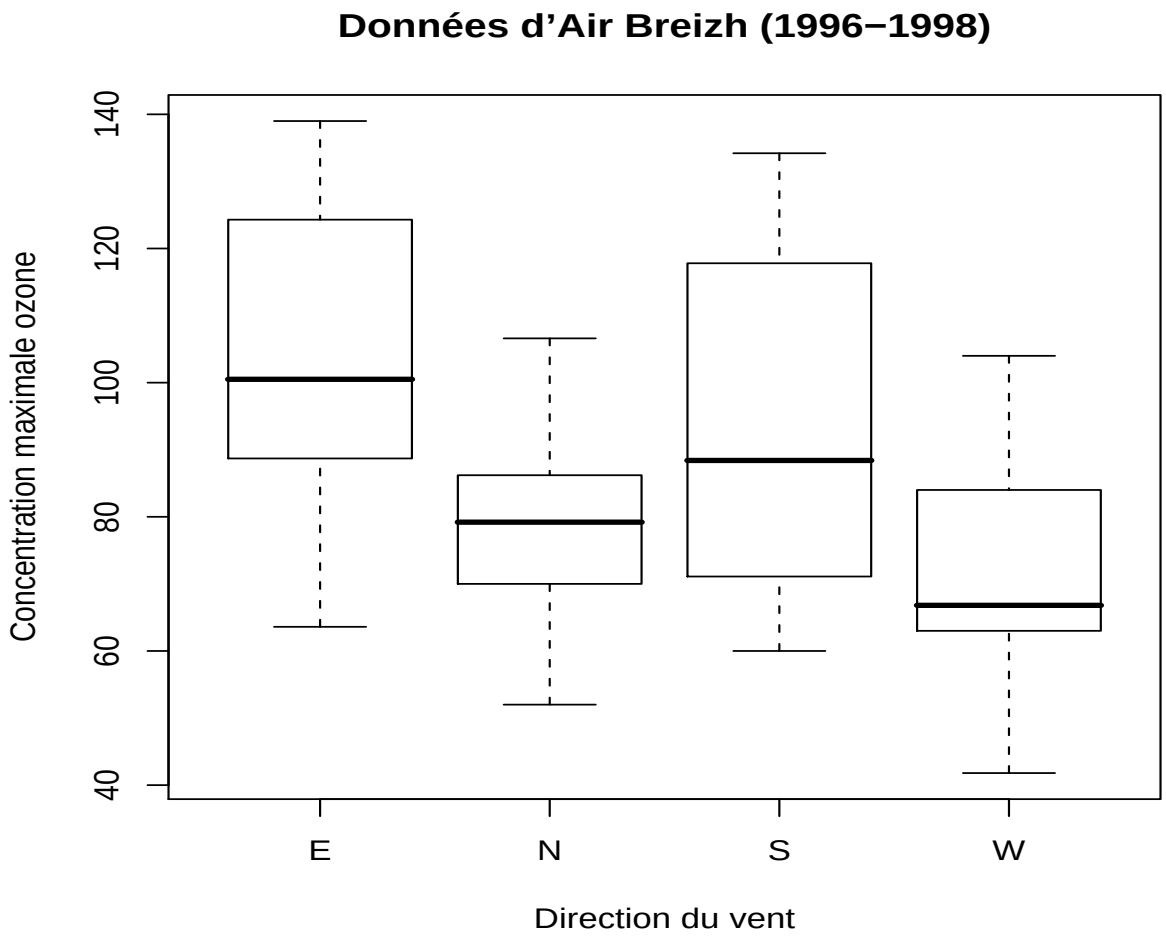
La variable qualitative considérée est souvent appelée **facteur**. *On suppose qu'elle prend ses valeurs dans un ensemble fini à I éléments appelés* **niveaux (du facteur)**.

- Une première analyse graphique : boîtes à moustaches (*boxplot*).

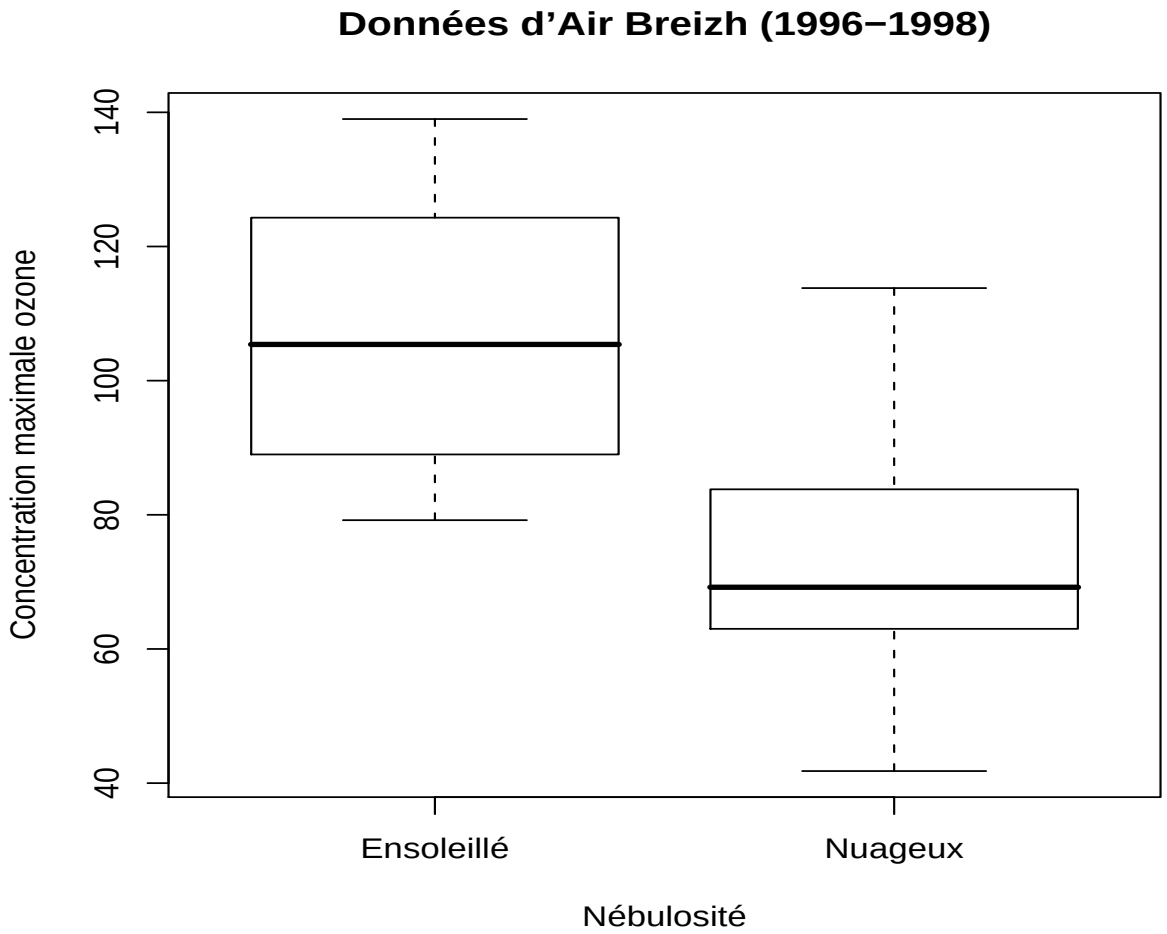
Analyse de la variance à un facteur



Analyse de la variance à un facteur



Analyse de la variance à un facteur



Analyse de la variance à un facteur

Pour $i = 1 \dots I$, on note n_i le nombre d'observations de Y correspondant au $i$ème niveau du facteur, et pour $j = 1 \dots n_i$, y_{ij} désigne la $j$ème observation de Y pour le $i$ème niveau du facteur.

Soit $n = \sum_{i=1}^{I} n_i$ le nombre total d'observations.

Définition

- *Si $n_i > 0$ pour tout i, on dira que le plan d'expérience est* **complet**.
- *Si $n_1 = \dots = n_I$, on dira que le plan d'expérience est* **équilibré**.

Analyse de la variance à un facteur

Modélisation

Modèle d'ANOVA à un facteur

Les observations y_{ij}, $i = 1 \dots I$, $j = 1 \dots n_i$ sont supposées être issues du modèle

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

où les variables ε_{ij} vérifient les conditions standards :

- $(\mathcal{C}_1)$ $\mathbb{E}[\varepsilon_{ij}] = 0$,
- $(\mathcal{C}_2)$ $\mathrm{cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0$,
- $(\mathcal{C}_3)$ $\mathrm{var}(\varepsilon_{ij}) = \sigma^2$.

Analyse de la variance à un facteur

Modélisation

Ecriture sous forme matricielle

Recodage (disjonctif complet) du facteur en variables indicatrices (*dummy variables*).

Matrice d'incidence : $A = (\mathbb{1}_1, \mathbb{1}_2, \ldots, \mathbb{1}_I)$ i.e.

$$A = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}.$$

Analyse de la variance à un facteur

Modélisation

Modèle : $Y = \mathbb{X}\beta + \varepsilon$ avec

- $Y = (Y_{11}, \ldots, Y_{1n_1}, \ldots, Y_{I1}, \ldots, Y_{In_I})'$
- $\beta = (\mu, \alpha_1, \ldots, \alpha_I)'$
- $\varepsilon = (\varepsilon_{11}, \ldots, \varepsilon_{1n_1}, \ldots, \varepsilon_{I1}, \ldots, \varepsilon_{In_I})'$,
- $\mathbb{X} = (\mathbb{1}, A)$, i.e.

$$\mathbb{X} = \begin{pmatrix} 1 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \ldots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \ldots & 1 \end{pmatrix}.$$

Analyse de la variance à un facteur

Modélisation

Identifiabilité et contraintes

Problème : le modèle n'est pas identifiable. Contre-exemple pour l'unicité. La matrice $\mathbb{X}$ n'est pas de plein rang...

Solution : contrainte linéaire identifiante sur les cœfficients $\longrightarrow$ reparamétrisation du modèle.

1. Contrainte de type analyse par cellule : $\mu = 0$. On pose alors $\beta = (\alpha_1, \ldots, \alpha_I)'$ et $\mathbb{X} = A$.
2. Contrainte de type cellule de référence : $\alpha_{i_r} = 0$ pour un i_r choisi dans $\{1, \ldots, I\}$. Choix de SAS et R par défaut.
3. Contrainte d'orthogonalité : $\sum_{i=1}^{I} n_i \alpha_i = 0$.
4. Contrainte de type somme : $\sum_{i=1}^{I} \alpha_i = 0$.

Analyse de la variance à un facteur

Estimation des paramètres

Contrainte $\mu = 0$

Modèle : $Y = \mathbb{X}\beta + \varepsilon$ avec $\beta = (\alpha_1, \ldots, \alpha_I)'$ et $\mathbb{X} = A$.

$$\mathbb{X}'\mathbb{X} = \begin{pmatrix} n_1 & 0 & \ldots & 0 \\ 0 & n_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ldots \\ 0 & 0 & \ldots & n_I \end{pmatrix} \text{ et } (\mathbb{X}'\mathbb{X})^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 & \ldots & 0 \\ 0 & \frac{1}{n_2} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \frac{1}{n_I} \end{pmatrix}.$$

Estimateur des moindres carrés ordinaires de β

$\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y = (\hat{\alpha}_1, \ldots, \hat{\alpha}_I)$ où $\hat{\alpha}_i = \bar{Y}_{i\bullet} = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$ pour tout i.

Analyse de la variance à un facteur

Estimation des paramètres

Remarque : l'estimateur $\hat{\beta}$ est intuitif !

Proposition

$\hat{\beta}$ est un estimateur sans biais de β de variance $\sigma^2 diag\left(\frac{1}{n_1}, \ldots, \frac{1}{n_I}\right)$ minimale parmi les estimateurs linéaires sans biais de β.

Le vecteur des valeurs ajustées est défini par $\hat{Y} = (\bar{Y}_{1\bullet}, \ldots, \bar{Y}_{I\bullet})'$, et celui des résidus par : $\hat{\varepsilon}_{ij} = Y_{ij} - \bar{Y}_{i\bullet}$.

Proposition

L'estimateur sans biais de la variance σ^2 est
$\widehat{\sigma^2} = \frac{1}{n-I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} \hat{\varepsilon}_{ij}^2 = \frac{1}{n-I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2$.

Analyse de la variance à un facteur

Estimation des paramètres

Contrainte $\alpha_{i_r} = 0$

Modèle : $Y = \mathbb{X}\beta + \varepsilon$ avec $\beta = (\mu, \alpha_1, \ldots, \alpha_{i_r-1}, \alpha_{i_r+1}, \ldots, \alpha_I)'$ et $\mathbb{X} = (\mathbb{1}, \mathbb{1}_1, \ldots, \mathbb{1}_{i_r-1}, \mathbb{1}_{i_r+1}, \ldots, \mathbb{1}_I)$.

Estimateur des moindres carrés ordinaires de β

$\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y = (\hat{\mu}, \hat{\alpha}_1, \ldots, \hat{\alpha}_{i_r-1}, \hat{\alpha}_{i_r+1}, \ldots, \hat{\alpha}_I)$, avec
$\hat{\mu} = \bar{Y}_{i_r\bullet}$, $\hat{\alpha}_i = \bar{Y}_{i\bullet} - \bar{Y}_{i_r\bullet}$ pour $i \in \{1, \ldots, i_r - 1, i_r + 1, \ldots, I\}$.

Remarques :

- Vecteur des valeurs ajustées inchangé car $\mathcal{E}(\mathbb{X})$ inchangé i.e. $\hat{Y} = (\bar{Y}_{1\bullet}, \ldots, \bar{Y}_{I\bullet})'$.
- Vecteur des résidus et estimateur sans biais de σ^2 : idem aussi.

Analyse de la variance à un facteur

Estimation des paramètres

Contrainte $\sum_{i=1}^{I} n_i \alpha_i = 0$ $\left(\alpha_I = -\frac{1}{n_I} \sum_{i=1}^{I-1} n_i \alpha_i\right)$

Modèle : $Y = \mathbb{X}\beta + \varepsilon$ avec $\beta = (\mu, \alpha_1, \ldots, \alpha_{I-1})'$ et
$\mathbb{X} = \left(\mathbb{1}, \mathbb{1}_1 - \frac{n_1}{n_I}\mathbb{1}_I, \ldots, \mathbb{1}_{I-1} - \frac{n_{I-1}}{n_I}\mathbb{1}_I\right)$

Estimateur des moindres carrés ordinaires de β

$\hat{\beta} = (\hat{\mu}, \hat{\alpha}_1, \ldots, \hat{\alpha}_{I-1})$, avec $\hat{\mu} = \bar{Y}_{\bullet\bullet} = \frac{1}{n}\sum_{i=1}^{I} n_i \bar{Y}_{i\bullet}$,
$\hat{\alpha}_i = \bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}$ pour $i \in \{1, \ldots, I-1\}$.

Remarque : vecteur des valeurs ajustées, vecteur des résidus et estimateur sans biais de σ^2 inchangés ($\mathcal{E}(\mathbb{X})$ inchangé...)

Analyse de la variance à un facteur

Estimation des paramètres

Contrainte $\sum_{i=1}^{I} \alpha_i = 0$ **(**$\alpha_I = -\sum_{i=1}^{I-1} \alpha_i$**)**

Modèle : $Y = \mathbb{X}\beta + \varepsilon$ avec $\beta = (\mu, \alpha_1, \ldots, \alpha_{I-1})'$ et $\mathbb{X} = (\mathbb{1}, \mathbb{1}_1 - \mathbb{1}_I, \ldots, \mathbb{1}_{I-1} - \mathbb{1}_I)$

Estimateur des moindres carrés ordinaires de β

$\hat{\beta} = (\hat{\mu}, \hat{\alpha}_1, \ldots, \hat{\alpha}_{I-1})$, avec $\hat{\mu} = \frac{1}{I}\sum_{i=1}^{I} \bar{Y}_{i\bullet}$,
$\hat{\alpha}_i = \bar{Y}_{i\bullet} - \frac{1}{I}\sum_{i=1}^{I} \bar{Y}_{i\bullet}$ pour $i \in \{1, \ldots, I-1\}$.

Remarque : vecteur des valeurs ajustées, vecteur des résidus et estimateur sans biais de σ^2 inchangés ($\mathcal{E}(\mathbb{X})$ inchangé...)

Analyse de la variance à un facteur

Equation d'analyse de la variance

On rappelle ici l'équation d'analyse de la variance usuelle :

$$SCT = SCE + SCR.$$

Elle s'exprime ici, puisque $\mathbb{1} \in \mathcal{E}(\mathbb{X})$ quelle que soit la contrainte, sous la forme :

$$\|Y - \bar{Y}_{\bullet\bullet}\mathbb{1}\|^2 = \|\hat{Y} - \bar{Y}_{\bullet\bullet}\mathbb{1}\|^2 + \|Y - \hat{Y}\|^2.$$

Interprétation :

- SCE = variabilité inter cellules, dispersion des moyennes empiriques par cellules autour de la moyenne empirique globale
- SCR = variabilité intra cellules.

Analyse de la variance à un facteur

Tests sous hypothèse gaussienne

On suppose $(\mathcal{C}_4)$ vérifiée : ε suit une loi gaussienne.

Test d'effet du facteur

$$(H_0) : \alpha_1 = \ldots = \alpha_I = 0 \text{ contre } (H_1) : \exists (i, i'), \alpha_i \neq \alpha_{i'}.$$

Cas particulier de test de validité de sous-modèle

Statistique de test : $F(Y) = \dfrac{SCE/(I-1)}{SCR/(n-I)} = \dfrac{\sum_{i=1}^{I} n_i (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2/(I-1)}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - Y_{i\bullet})^2/(n-I)}$.

Loi sous (H_0) : $F(Y) \sim_{(H_0)} \mathcal{F}(I-1, n-I)$.

Région critique du test de niveau α : $\{y,\ F(y) > f_{I-1,n-I}(1-\alpha)\}$.

Analyse de la variance à un facteur

Tests sous hypothèse gaussienne

Tableau d'analyse de la variance

Variation	ddl	SC	CM	F	$Pr(> F)$
Facteur	$I-1$	SCE	$SCE/(I-1)$	$\frac{SCE/(I-1)}{SCR/(n-I)}$	
Résiduelle	$n-I$	SCR	$SCR/(n-I)$		
Totale	$n-1$	SCT			

Analyse de la variance à deux facteurs

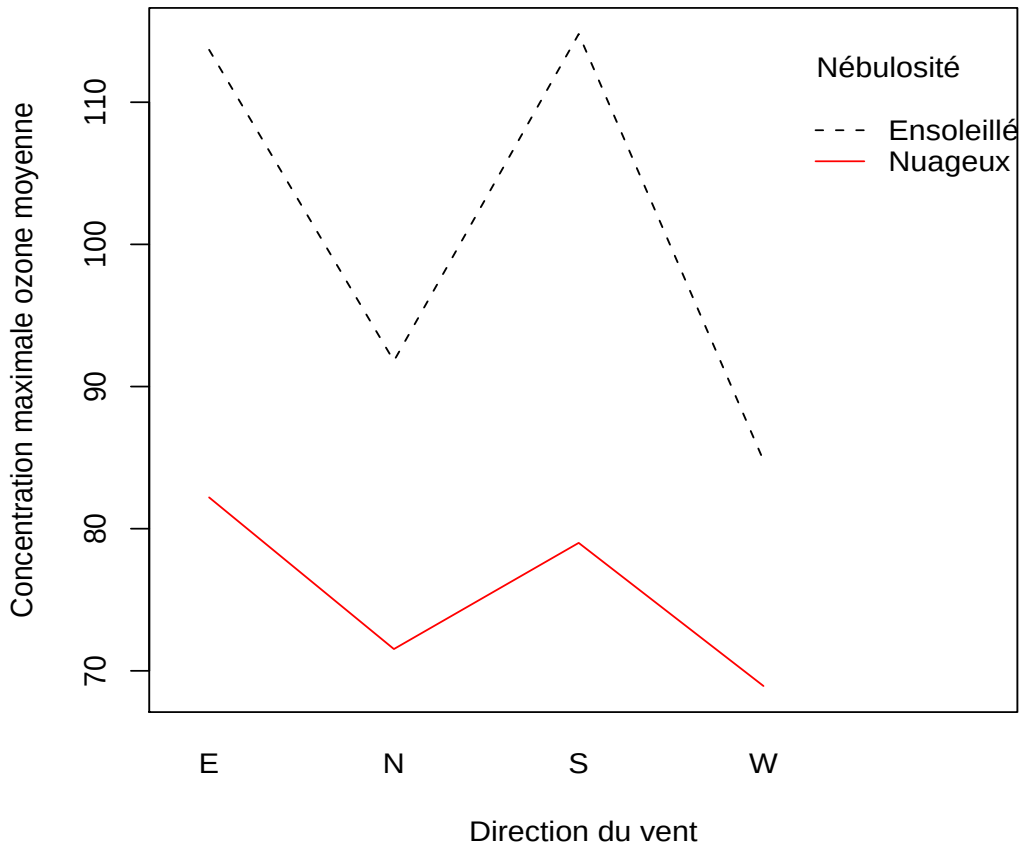
Cas d'une variable à expliquer Y quantitative et de deux variables explicatives qualitatives potentielles = deux facteurs dont le premier à I niveaux, et le deuxième à J niveaux.

Objectif : déterminer si les facteurs considérés ont un effet - significatif - sur Y.

▶ Une première analyse graphique : tracé des moyennes empiriques par cellules (**profils**).

Analyse de la variance à deux facteurs

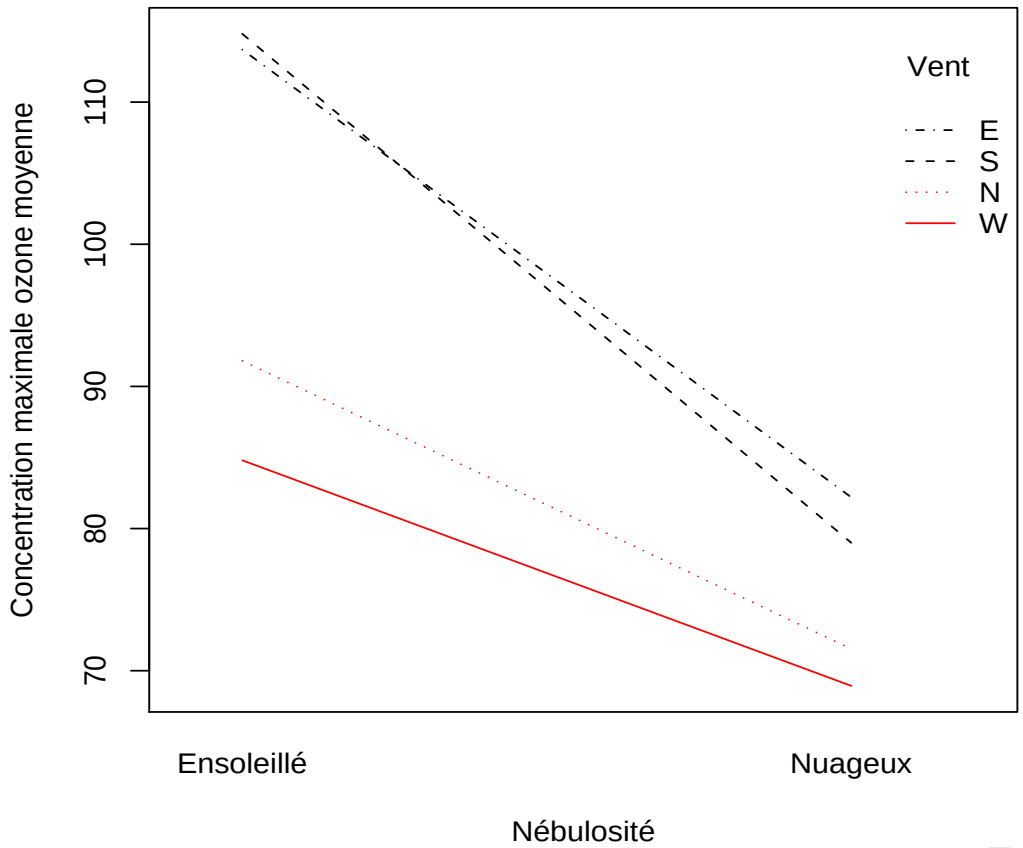
Exemple des données d'Air Breizh



Analyse de la variance à deux facteurs

Introduction

Exemple des données d'Air Breizh



Analyse de la variance à deux facteurs

Pour tout $i = 1 \dots I$, $j = 1 \dots J$, on note $n_{i,j}$ le nombre d'observations de la variable Y correspondant aux $i$ème et $j$ème niveaux des deux facteurs considérés, et y_{ijk} la $k$ième observation de Y correspondant aux $i$ème et $j$ème niveaux des facteurs. On suppose que y_{ijk} est l'observation d'une variable aléatoire Y_{ijk}.

Notations utiles :

$n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{i,j}$ nombre total d'observations,

$\bar{Y}_{ij\bullet} = \frac{1}{n_{i,j}} \sum_{k=1}^{n_{i,j}} Y_{ijk}$,

$\bar{Y}_{i\bullet\bullet} = \frac{1}{\sum_{j=1}^{J} n_{i,j}} \sum_{j=1}^{J} \sum_{k=1}^{n_{i,j}} Y_{ijk}$,

$\bar{Y}_{\bullet j\bullet} = \frac{1}{\sum_{i=1}^{I} n_{i,j}} \sum_{i=1}^{I} \sum_{k=1}^{n_{i,j}} Y_{ijk}$,

$\bar{Y}_{\bullet\bullet\bullet} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{n_{i,j}} Y_{ijk}$.

Analyse de la variance à deux facteurs

Modélisation

Modèle d'ANOVA à deux facteurs

Les observations y_{ijk}, $i = 1 \dots I$, $j = 1 \dots J$, $k = 1 \dots n_{i,j}$ sont supposées être issues du modèle

$$Y_{ijk} = \mu + \alpha_i + \eta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad \begin{cases} i = 1 \dots I \\ j = 1 \dots J \\ k = 1 \dots n_{i,j} \end{cases},$$

où les variables ε_{ijk} vérifient les conditions standards :

- $(\mathcal{C}_1)$ $\mathbb{E}[\varepsilon_{ijk}] = 0$,
- $(\mathcal{C}_2)$ $\mathrm{cov}(\varepsilon_{ijk}, \varepsilon_{i'j'k'}) = 0$,
- $(\mathcal{C}_3)$ $\mathrm{var}(\varepsilon_{ijk}) = \sigma^2$.

On considère dans la suite seulement le cas d'un plan d'expérience équilibré i.e. $n_{i,j} = K$ pour tout (i, j).

Analyse de la variance à deux facteurs

Modélisation

Ecriture sous forme matricielle

Recodage des facteurs en variables indicatrices.
Matrices d'incidences :

- $A = (\mathbb{1}_{1\bullet}, \mathbb{1}_{2\bullet}, \ldots, \mathbb{1}_{I\bullet})$,
- $B = (\mathbb{1}_{\bullet 1}, \mathbb{1}_{\bullet 2}, \ldots, \mathbb{1}_{\bullet J})$,
- $C = (\mathbb{1}_{11}, \mathbb{1}_{12}, \ldots, \mathbb{1}_{1J}, \ldots, \mathbb{1}_{I1}, \ldots, \mathbb{1}_{IJ})$.

Modèle : $Y = \mathbb{X}\beta + \varepsilon$ avec

- $Y = (Y_{ijk})'_{i=1\ldots I, j=1\ldots J, k=1\ldots K}$,
- $\varepsilon = (\varepsilon_{ijk})'_{i=1\ldots I, j=1\ldots J, k=1\ldots K}$,
- $\beta = (\mu, \alpha_1, \ldots, \alpha_I, \eta_1, \ldots, \eta_J, \gamma_{11}, \ldots, \gamma_{1J}, \ldots, \gamma_{I1}, \ldots, \gamma_{IJ})'$,
- $\mathbb{X} = (\mathbb{1}, A, B, C)$.

Analyse de la variance à deux facteurs

Modélisation

Identifiabilité et contraintes

Problème : Le modèle n'est pas identifiable. La matrice $\mathbb{X}$ est de taille $n \times (1 + I + J + IJ)$, mais rang$(\mathbb{X}) = IJ$.

Solution : ensemble de $1 + I + J$ contraintes linéaires identifiantes sur les cœfficients linéairement indépendantes $\longrightarrow$ reparamétrisation du modèle.

1. Contrainte de type analyse par cellule : $\mu = 0$, $\alpha_i = 0 \ \forall i$, $\eta_j = 0 \ \forall j$,
2. Contrainte de type cellule de référence : $\alpha_{i_r} = \eta_{j_r} = 0$ et $\gamma_{ij_r} = \gamma_{i_r j} = 0 \ \forall i = 1 \dots I, \ j = 1 \dots J$.
3. Contrainte de type somme : $\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \eta_j = 0$ et $\sum_{i=1}^{I} \gamma_{ij} = \sum_{j=1}^{J} \gamma_{ij} = 0 \ \forall i = 1 \dots I, \ j = 1 \dots J$.

Analyse de la variance à deux facteurs

Estimation des paramètres

Contrainte de type analyse par cellule

$$\text{Modèle : } Y_{ijk} = \gamma_{ij} + \varepsilon_{ijk} \quad \begin{cases} i = 1 \dots I \\ j = 1 \dots J \\ k = 1 \dots K \end{cases}$$

Estimateurs des moindres carrés ordinaires des cœfficients

$\hat{\gamma}_{ij} = \bar{Y}_{ij\bullet}$ pour tout $i = 1 \dots I$, $j = 1 \dots J$.

Valeurs ajustées : $\hat{Y}_{ijk} = \bar{Y}_{ij\bullet}$, résidus : $\hat{\varepsilon}_{ijk} = Y_{ijk} - \bar{Y}_{ij\bullet}$.

Estimateur sans biais de la variance

$$\widehat{\sigma^2} = \frac{SCR}{n-IJ} = \frac{1}{n-IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} (Y_{ijk} - \bar{Y}_{ij\bullet})^2.$$

Analyse de la variance à deux facteurs

Estimation des paramètres

Contrainte de type somme

Modèle : $Y_{ijk} = \mu + \alpha_i + \eta_j + \gamma_{ij} + \varepsilon_{ijk}$ $\begin{cases} i = 1 \dots I \\ j = 1 \dots J \\ k = 1 \dots K \end{cases}$, avec

$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \eta_j = \sum_{i=1}^{I} \gamma_{ij} = \sum_{j=1}^{J} \gamma_{ij} = 0 \ \forall i = 1 \dots I, \ j = 1 \dots J.$

Estimateurs des moindres carrés ordinaires des cœfficients

$\hat{\mu} = \bar{Y}_{\bullet\bullet\bullet}$,
$\hat{\alpha}_i = \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet}$,
$\hat{\eta}_j = \bar{Y}_{\bullet j\bullet} - \bar{Y}_{\bullet\bullet\bullet}$,
$\hat{\gamma}_{ij} = \bar{Y}_{ij\bullet} - \bar{Y}_{i\bullet\bullet} - \bar{Y}_{\bullet j\bullet} + \bar{Y}_{\bullet\bullet\bullet}$.

Remarque : Valeurs ajustées, résidus et estimateur sans biais de la variance inchangés.

Analyse de la variance à deux facteurs

Equation d'analyse de la variance

On rappelle l'équation d'analyse de la variance usuelle :

$$SCT = SCE + SCR.$$

La somme des carrés expliquée peut se décomposer par Pythagore en : $SCE = SCE_A + SCE_B + SCE_C$, où

- $SCE_A = JK \sum_{i=1}^{I} \hat{\alpha}_i^2$,
- $SCE_B = IK \sum_{j=1}^{J} \hat{\eta}_j^2$,
- $SCE_C = K \sum_{i=1}^{I} \sum_{j=1}^{J} \hat{\gamma}_{ij}^2$.

L'équation d'analyse de la variance devient alors :

$$SCT = SCE_A + SCE_B + SCE_C + SCR.$$

Analyse de la variance à deux facteurs

Tests sous hypothèse gaussienne

On suppose $(\mathcal{C}_4)$ vérifiée : ε suit une loi gaussienne.

Test de l'interaction

$$(H_0)_C : \forall (i,j)\ \gamma_{ij} = 0 \text{ contre } (H_1)_C : \exists (i,j),\ \gamma_{ij} \neq 0.$$

Cas particulier de test de validité de sous-modèle

Statistique de test : $F_C(Y) = \frac{SCE_C/((I-1)(J-1))}{SCR/(n-IJ)}$.

Loi sous $(H_0)_C$: $F_C(Y) \sim_{(H_0)_C} \mathcal{F}((I-1)(J-1), n-IJ)$.

Région critique du test de niveau α :

$$\{y,\ F_C(y) > f_{(I-1)(J-1),n-IJ}(1-\alpha)\}.$$

Analyse de la variance à deux facteurs

Tests sous hypothèse gaussienne

Remarque : Si on rejette $(H_0)_C$, les facteurs ont un effet significatif, on ne teste donc pas les effets individuels de chaque facteur.

Si on ne rejette pas $(H_0)_C$, on suppose que $\forall(i,j)\ \gamma_{ij} = 0$ et on teste les effets individuels de chaque facteur.

On suppose dans la suite $(\mathcal{C}_4)$ vérifiée : ε suit une loi gaussienne, et $\forall(i,j)\ \gamma_{ij} = 0$.

Analyse de la variance à deux facteurs

Tests sous hypothèse gaussienne

Test de l'effet du facteur A

$$(H_0)_A : \forall i \ \alpha_i = 0 \text{ contre } (H_1)_A : \exists i, \alpha_i \neq 0.$$

Par précaution, on choisit généralement d'estimer σ^2 dans le modèle complet i.e. avec interaction (choix usuel des logiciels).

Test de validité de sous-modèle

Statistique de test : $F_A(Y) = \frac{SCE_A/(I-1)}{SCR/(n-IJ)}$.
Loi sous $(H_0)_A$: $F_A(Y) \sim_{(H_0)_A} \mathcal{F}(I-1, n-IJ)$.
Région critique du test de niveau α :

$$\{y, \ F_A(y) > f_{I-1,n-IJ}(1-\alpha)\}.$$

Analyse de la variance à deux facteurs

Tests sous hypothèse gaussienne

Test de l'effet du facteur B

$(H_0)_B : \forall j\ \eta_j = 0$ contre $(H_1)_B : \exists j, \eta_j \neq 0.$

Test de validité de sous-modèle

Statistique de test : $F_B(Y) = \frac{SCE_B/(J-1)}{SCR/(n-IJ)}$.

Loi sous $(H_0)_B$: $F_B(Y) \sim_{(H_0)_B} \mathcal{F}(J-1, n-IJ)$.

Région critique du test de niveau α :

$$\{y,\ F_B(y) > f_{J-1,n-IJ}(1-\alpha)\}.$$

Analyse de la variance à deux facteurs

Tests sous hypothèse gaussienne

Tableau d'analyse de la variance

Variation	ddl	SC	CM	F	$Pr(>F)$
Facteur A	$I-1$	SCE_A	$\frac{SCE_A}{I-1}$	$\frac{SCE_A/(I-1)}{SCR/(n-IJ)}$	
Facteur B	$J-1$	SCE_B	$\frac{SCE_B}{J-1}$	$\frac{SCE_B/(J-1)}{SCR/(n-IJ)}$	
Interaction	$(I-1)(J-1)$	SCE_C	$\frac{SCE_C}{(I-1)(J-1)}$	$\frac{SCE_C/((I-1)(J-1))}{SCR/(n-IJ)}$	
Résiduelle	$n-IJ$	SCR	$\frac{SCR}{n-IJ}$		
Totale	$n-1$	SCT			

Compléments

Autres tests et intervalles de confiance par extension de la régression linéaire multiple générale.
Plans d'expérience plus complexes.

Remarque fondamentale : une analyse de la variance doit être complétée comme toute analyse de modèle de régression par la recherche d'éventuels écarts au modèle $\longrightarrow$ analyse des résidus, données atypiques, etc.

Pour une présentation plus complète de l'analyse de la variance : Scheffé (1959).

Mélange de variables quantitatives et qualitatives : ANCOVA (*c.f. cours de Modèles de régression 2*).