Modèle Linéaire Gaussien
Chapitre 3 – Analyse de la variance

Achille Thin

12 Décembre 2023

Executive Master Statistique et Big Data

Analyse de la Variance à un Facteur

Objectifs du Chapitre

Cadre de l'analyse de la variance (ANOVA) à 1 facteur: expliquer les variations d'**une variable quantitative**, appelée variable réponse, **en fonction d'une variable qualitative**, appelée variable explicative (ou facteur).

Démarche statistique:

1. Écriture du modèle
2. Ajustement (estimation) du modèle grâce aux données
3. Vérification de la validité des hypothèses faites dans le modèle
4. Test de la pertinence des différents éléments du modèle
5. Critique du modèle
6. Conclusion

Objectifs

Question : la direction du vent a-t-elle une influence sur le maximum de concentration d'ozone observé sur une journée ?

- **Variable réponse:** la concentration en ozone (maxO3).
- **Variable explicative:** la direction du vent (vent).

Chargement des données:

```
donnees <- read.table("../data/ozone.txt", header = TRUE,
                      colClasses = c(rep("numeric", 11),
                                     rep("factor", 2)))
attach(donnees)
```

Analyse descriptive

Résumés numériques et graphiques

On a $n = 112$ observations et 4 niveaux ou modalités (valeurs possibles) pour le vent :

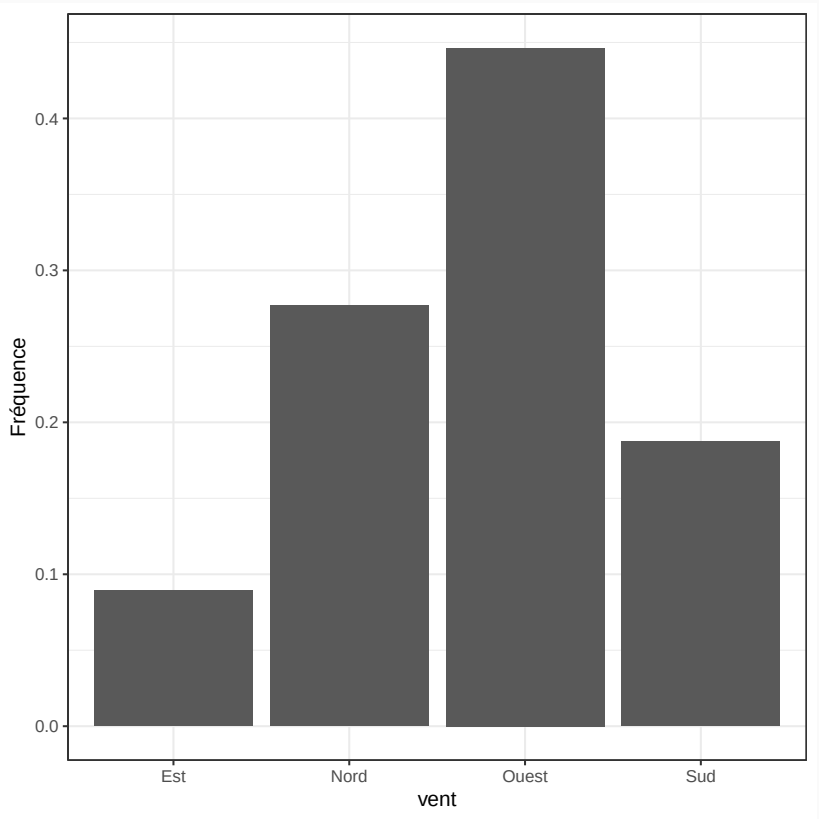
```
levels(vent)
```

```
[1] "Est"   "Nord"  "Ouest" "Sud"
```

Effectifs par modalité:

```
table(vent)
```

```
vent
  Est   Nord Ouest   Sud
   10    31    50    21
```

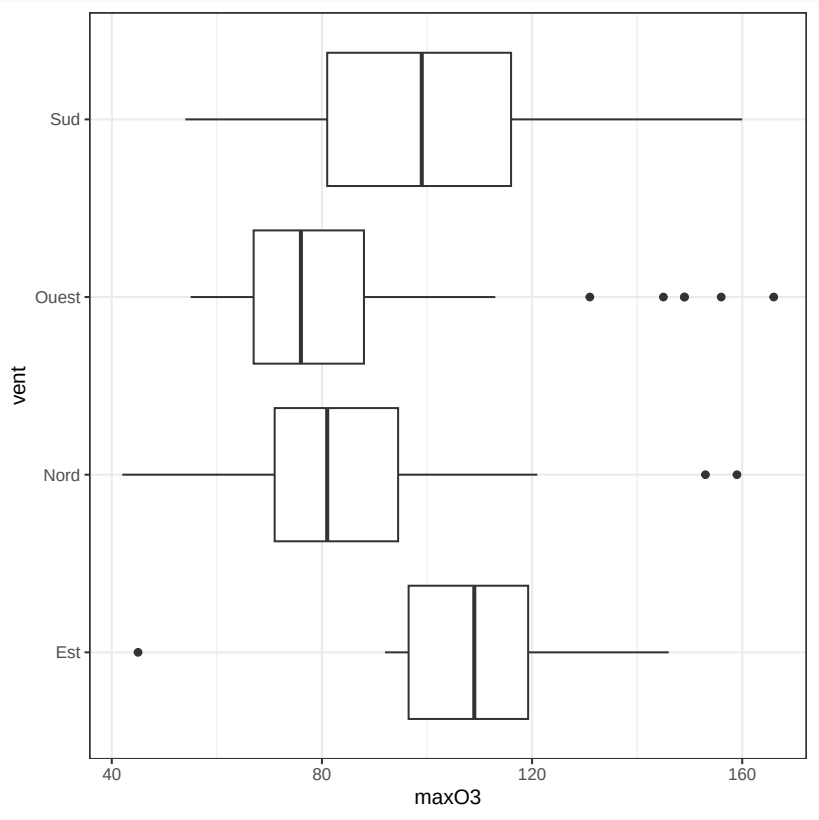


Résumés numériques et graphiques

Moyennes par modalité:

```
aggregate(
  list(maxO3 = maxO3),
  list(vent = vent),
  mean
)
```

```
   vent    maxO3
1   Est 105.600
2  Nord  86.129
3 Ouest  84.700
4   Sud 102.524
```



Analyse graphique: le vent semble avoir une influence sur la concentration en ozone.

Écriture du modèle

Notations

Variable explicative:

- La variable qualitative à I modalités.
- Chaque modalité est codée par un entier i, $i \in [\![1, I]\!]$.
- Pour chaque modalité i, on dispose de n_i observations.

Exemple: sur nos données, $I = 4$, Est $= 1$, Nord $= 2$, Ouest $= 3$, Sud $= 4$. On a $n_1 = 10$, $n_2 = 31$, $n_3 = 50$ et $n_4 = 21$.

Variable réponse: on note $y_{i,j}$ la valeur de la variable réponse pour la j-ème observation de la modalité i.

Exemple: sur nos données $y_{4,3}$ désigne le maximum de concentration d'ozone observé la 3-ème journée où il y a eu un vent du sud.

Modèle régulier

On suppose que $y_{i,j}$ est la réalisation d'une variable aléatoire $Y_{i,j}$ telle que :

$$Y_{i,j} = \mu_i + \varepsilon_{i,j}, \quad 1 \leq i \leq I = 4, \quad 1 \leq j \leq n_i,$$

- $\mu_1, \ldots, \mu_I$ sont des paramètres inconnus (μ_i représente la moyenne attendue de la concentration en ozone maximale journalière pour la direction de vent i)
- $\varepsilon_{i,j}$ est une variable aléatoire appelée **bruit**, telle que toutes les variables aléatoires $(\varepsilon_{i,j})$ sont **indépendantes**, d'**espérance nulle** et ont la **même variance**, égale à σ^2 (paramètre inconnu).

Cas particulier du modèle linéaire gaussien: les variables aléatoires $\varepsilon_{i,j}$ sont indépendantes et identiquement distribuées de loi $\mathcal{N}(0, \sigma^2)$.

Remarque: Pour un même niveau i, les variables aléatoires $Y_{i,j}$ sont *i.i.d.* suivant la loi normale $\mathcal{N}(\mu_i, \sigma^2)$.

Modèle régulier – Écriture matricielle

Le modèle peut s'écrire

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\mu} + \boldsymbol{\varepsilon}, \quad \text{avec} \quad \boldsymbol{\mu} = (\mu_1, \dots, \mu_I)^\intercal,$$

et

$$\mathbf{Y} = \begin{pmatrix} Y_{1,1} \\ \vdots \\ Y_{1,n_1} \\ Y_{2,1} \\ \vdots \\ Y_{2,n_2} \\ \vdots \\ Y_{I,1} \\ \vdots \\ Y_{I,n_I} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \dots & \dots & \vdots \\ 1 & 0 & \dots & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 1 & 0 & \dots & 0 \\ & & & & \\ 0 & \dots & \dots & 0 & 1 \\ \vdots & \dots & \dots & \vdots & \vdots \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix} \quad \text{et} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{1,1} \\ \vdots \\ \varepsilon_{1,n_1} \\ \varepsilon_{2,1} \\ \vdots \\ \varepsilon_{2,n_2} \\ \vdots \\ \varepsilon_{I,1} \\ \vdots \\ \varepsilon_{I,n_I} \end{pmatrix} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n).$$

Remarque: on parle de modèle régulier car la matrice $\mathbf{X}$ est de plein rang en colonnes.

Modèle régulier – Estimation du modèle

On retrouve les mêmes formules et propriétés que dans le cas de la régression linéaire multiple.

Estimateur de μ (variable aléatoire)

$$\widehat{\mu} = (X^T X)^{-1} X Y \quad \rightsquigarrow \quad \widehat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{i,j} := \overline{Y}_{i\bullet}.$$

Pour une modalité $i \in [\![1, I]\!]$, on obtient que les variables ajustées $\widehat{Y}_{i,j}$ sont égales à $\widehat{\mu}_i$, *i.e.*, la moyenne des observations pour la modalité.

Estimateur de σ^2 (variable aléatoire)

$$S^2 = \frac{1}{n - I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{i,j} - \overline{Y}_{i\bullet})^2.$$

Avec R:

```
reg <- lm(maxO3 ~ vent - 1, data = donnees)
```

Modèle régulier – Interprétation

On retrouve les sorties habituelles. Le test du modèle ou test de Fisher global (dernière ligne de la sortie de `summary`) n'a néanmoins pas une interprétation intéressante.

Hypothèses du test du modèle:

$$\mathcal{H}_0 : \forall i \in [\![1, I]\!], \quad \mu_i = 0, \quad \text{contre} \quad \mathcal{H}_1 : \exists i \in [\![1, I]\!], \quad \mu_i \neq 0.$$

Ces hypothèses ne traduisent pas la question d'intérêt : le facteur d'intérêt (la direction du vent) a-t-il une influence sur la variable réponse (la concentration en ozone) ? Mathématiquement, cette question se formalise par

$$\mathcal{H}_0 : \mu_1 = \ldots = \mu_I, \quad \text{contre} \quad \mathcal{H}_1 : \text{les } \mu_i \text{ ne sont pas tous égaux.}$$

Nous allons donc choisir une autre écriture du modèle !

Modèle singulier

On décompose μ_i sous la forme $\mu_i = \mu + \alpha_i$:

- μ est un paramètre inconnu (effet moyen $\rightsquigarrow$ la concentration moyenne de référence en ozone),
- $\alpha_1, \ldots, \alpha_I$ sont des paramètres inconnus (α_i représente la différence par rapport au niveau de référence $\rightsquigarrow$ l'effet de la direction du vent i sur la concentration en ozone maximale journalière par rapport au niveau de référence μ).

On suppose alors que $y_{i,j}$ est la réalisation d'une variable aléatoire $Y_{i,j}$ telle que :

$$Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}, \quad 1 \leq i \leq I = 4, \quad 1 \leq j \leq n_i,$$

Cas particulier du modèle linéaire gaussien: les variables aléatoires $\varepsilon_{i,j}$ sont indépendantes et identiquement distribuées de loi $\mathcal{N}(0, \sigma^2)$.

Modèle singulier – Écriture matricielle

Le modèle peut s'écrire

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{avec} \quad \boldsymbol{\beta} = (\mu, \alpha_1, \ldots, \alpha_I)^\intercal,$$

et

$$\mathbf{Y} = \begin{pmatrix} Y_{1,1} \\ \vdots \\ Y_{1,n_1} \\ Y_{2,1} \\ \vdots \\ Y_{2,n_2} \\ \vdots \\ Y_{I,1} \\ \vdots \\ Y_{I,n_I} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & \ldots & \ldots & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ \vdots & \vdots & \vdots & \ldots & \ldots & \vdots \\ 1 & 1 & 0 & \ldots & \ldots & 0 \\ 1 & 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & 0 & 1 & 0 & \ldots & 0 \\ \\ 1 & 0 & \ldots & \ldots & 0 & 1 \\ \vdots & \vdots & & & \vdots & \vdots \\ \vdots & \vdots & \ldots & \ldots & \vdots & \vdots \\ 1 & 0 & \ldots & \ldots & 0 & 1 \end{pmatrix} \quad \text{et} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{1,1} \\ \vdots \\ \varepsilon_{1,n_1} \\ \varepsilon_{2,1} \\ \vdots \\ \varepsilon_{2,n_2} \\ \vdots \\ \varepsilon_{I,1} \\ \vdots \\ \varepsilon_{I,n_I} \end{pmatrix} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_n).$$

Remarque: on parle de modèle singulier car $\mathbf{X}$ n'est plus de plein rang en colonnes (la première colonne est la somme des autres colonnes) $\rightsquigarrow$ problème d'identifiabilité ($I + 1$ paramètres inconnus pour I facteurs/équations).

Ajustement du modèle singulier

Contraintes et identifiabilité

Pour tout $i \in [\![1, I]\!]$, on a l'estimateur

$$\widehat{\mu}_i = \widehat{\mu} + \widehat{\alpha}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{i,j} := \overline{Y}_{1\bullet}.$$

Afin de rendre le modèle identifiable (avoir une unique solution), il est nécessaire d'introduire une contrainte sur les paramètres.

Exemples de contraintes:

- $\mu = 0$: cette contrainte correspond au modèle régulier.
- $\alpha_1 = 0$: on a alors $\mu = \mu_1$ et pour $i \geq 2$, $\alpha_i = \mu_i - \mu_1$. La première modalité sert de référence pour l'effet moyen et les α_i, $i \geq 2$, représentent les différences des autres modalités avec la modalité 1. **C'est la convention par défaut de R!**
- $\sum_{i=1}^{I} \alpha_i = 0$: l'effet moyen μ est alors la moyenne des effets de chaque modalité.

Estimateur et estimation de $\mu, \alpha_1, \ldots, \alpha_I$

Important: les estimateurs de $\mu, \alpha_1, \ldots, \alpha_I$ **dépendent de la contraintes choisie**. On considère ici la contrainte $\alpha_1 = 0$.

Estimateurs (variables aléatoires). On a $\mu_1 = \mu + \alpha_1$. Sous la contrainte $\alpha_1 = 0$, on a $\mu = \mu_1$. L'estimateur $\widehat{\mu}$ de μ est défini par

$$\widehat{\mu} = \widehat{\mu}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} Y_{i,j} := \overline{Y}_{1\bullet}.$$

De plus pour tout $i \in [\![2, I]\!]$, $\mu_i = \mu + \alpha_i$, *i.e.*, $\alpha_i = \mu_i - \mu = \mu_i - \mu_1$. On a donc pour $i \in [\![2, I]\!]$, l'estimateur $\widehat{\alpha}_i$ de α_i

$$\widehat{\alpha}_i = \widehat{\mu}_i - \widehat{\mu} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{i,j} - \frac{1}{n_1} \sum_{j=1}^{n_1} Y_{i,j} := \overline{Y}_{i\bullet} - \overline{Y}_{1\bullet}.$$

Estimations (valeurs numériques calculées sur les données)

$$\widehat{\mu}^{\text{obs}} = \overline{y}_{1\bullet} \quad \text{et} \quad \forall i \in [\![2, I]\!], \quad \widehat{\alpha}_i^{\text{obs}} = \overline{y}_{i\bullet} - \overline{y}_{1\bullet}.$$

Variables ajustées et résidus

Variables ajustées (variables aléatoires)

$$\widehat{Y} = X(\widehat{\mu}, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_I)^T, \quad \textit{i.e.,} \quad \widehat{Y}_{i,j} = \overline{Y}_{i\bullet}, \quad 1 \leq i \leq I, \quad 1 \leq j \leq n_i.$$

Valeurs ajustées (réalisations sur les données)

$$\widehat{y}_{i,j} = \overline{y}_{i\bullet}, \quad 1 \leq i \leq I, \quad 1 \leq j \leq n_i.$$

```
reg$fitted.values   # ou fitted(reg)
```

Résidus (variables aléatoires) estimateurs des erreurs inconnues $\varepsilon_{i,j}$ (comme en régression) :

$$\widehat{\varepsilon}_{i,j} = Y_{i,j} - \widehat{Y}_{i,j} = Y_{i,j} - \overline{Y}_{i\bullet} \quad 1 \leq i \leq I, \quad 1 \leq j \leq n_i.$$

Résidus observés: $\widehat{e}_{i,j} = y_{i,j} - \widehat{y}_{i,j} = y_{i,j} - \overline{y}_{i\bullet}, 1 \leq i \leq I, 1 \leq j \leq n_i.$

```
reg$residuals   # ou resid(reg)
```

Estimateur et estimation de σ^2

Estimateur de la variance du bruit (σ^2) (variable aléatoire)

$$S^2 = \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}\widehat{\varepsilon}_{i,j}^2 = \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{i,j}-\overline{Y}_{i\bullet})^2.$$

Cet estimateur vérifie

$$(n-I)\frac{S^2}{\sigma^2} \sim \chi^2(n-I).$$

Estimation de σ^2 (réalisation sur les données)

$$\widehat{\sigma}^2 = \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}\widehat{e}_{i,j}^2 = \frac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{i,j}-\overline{y}_{i\bullet})^2.$$

Accès à la valeur estimée:

```
summary(reg)$sigma^2
```

Changement de contrainte – Quelles conséquences ?

Ce qui change:

- l'expression des estimateurs de $\mu, \alpha_1, \ldots, \alpha_I$,
- l'interprétation des coefficients du modèle et donc des hypothèses des tests.

Ce qui ne change pas quelque soit l'écriture du modèle:

- les variables ajustées et les résidus (les conclusions de l'analyse graphique des résidus restent donc les mêmes),
- l'estimateur de σ,
- les conclusions des tests statistiques (influence ou non),
- les prédictions.

Avec R : ajustement du modèle

Contrainte par défaut:

```
reg <- lm(maxO3 ~ vent, data = donnees)
```

Estimation de $\mu, \alpha_1, \ldots, \alpha_I$:

```
reg$coefficients  # ou coef(reg)
```

```
(Intercept)    ventNord   ventOuest      ventSud
   105.6000    -19.4710    -20.9000     -3.0762
```

- R indice les facteurs suivant l'ordre alphabétique, *i.e.*, (Est $= 1$, Nord $= 2$, Ouest $= 3$, Sud $= 4$).
- `(Intercept)` représente l'effet de la première modalité Les autres coefficients sont la différence avec cette modalité.
- On peut retrouver les estimations de $\mu_1, \ldots, \mu_4$ du modèle régulier à partir des sorties du modèle singulier

```
coef(lm(maxO3 ~ vent - 1, data = donnees))
```

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

Avec R : modification de la contrainte

Commande: on utilise `C(...)`

Changer la modalité de référence: on choisit la deuxième modalité (Nord) comme référence

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

Contrainte $\sum_{i=1}^{I} \alpha_i = 0$:

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

Exercice

Calculer les estimateurs de $\mu, \alpha_1, \ldots, \alpha_I$ sous la contraintes $\sum_{i=1}^{I} \alpha_i = 0$.

Solution: Pour tout $i \in [\![i, I]\!]$ on a $\mu + \alpha_i = \mu_i$. Sous cette contrainte, on en déduit donc que

$$\sum_{i=1}^{I} (\mu + \alpha_i) = I\mu = \sum_{i=1}^{I} \mu_i \quad \Rightarrow \quad \mu = \frac{1}{I} \sum_{i=1}^{I} \mu_i.$$

On en déduit l'estimateur

$$\widehat{\mu} = \frac{1}{I} \sum_{i=1}^{I} \widehat{\mu}_i = \frac{1}{I} \sum_{i=1}^{I} \overline{Y}_{i\bullet}.$$

Par suite, $i \in [\![i, I]\!]$, on a l'estimateur

$$\widehat{\alpha}_i = \widehat{\mu}_i - \widehat{\mu} = \overline{Y}_{i\bullet} - \frac{1}{I} \sum_{k=1}^{I} \overline{Y}_{k\bullet}.$$

Validité des hypothèses

Diagnostique des résidus

Avant d'analyser les sorties du modèle ajusté: il faut regarder si les hypothèses du modèle linéaire gaussien sont vérifiées sur nos données, *i.e.*, les variables aléatoires $\varepsilon_{i,j}$

(P1) sont **indépendantes**,

(P2) sont toutes d'**espérance nulle** ($\rightsquigarrow$ la relation entre $\mathbf{y}$ et $\mathbf{x}$ est bien affine),

(P3) ont la **même variance** σ^2 (homoscédasticité),

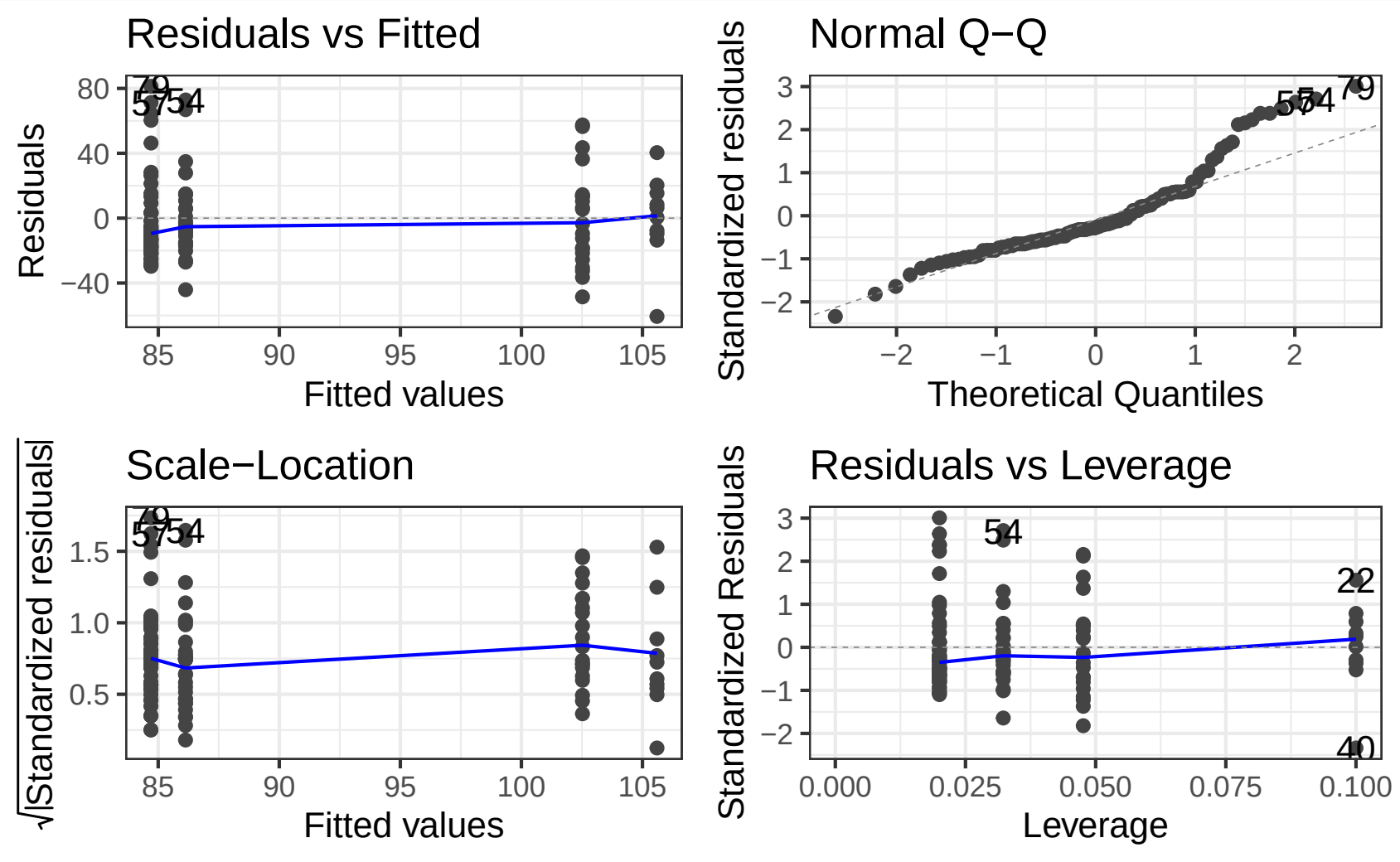
(P4) suivent une **loi normale**.

Validation des hypothèses:

- (P1) : l'indépendance ne peut être assurée que par le protocole expérimentale.
- (P2), (P3), (P4) : on fait la même analyse graphique des résidus observés que pour la régression. Pour (P2) et (P3), on veut le même comportement pour toutes les modalités.

Avec R : 4 graphiques à analyser

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



Tests d'hypothèses

Test du modèle ou test de Fisher global

Question: la direction du vent a-t-elle une influence sur la concentration en ozone ?

Mathématiquement, cela revient à tester

$$\mathcal{H}_0 : \mu_1 = \ldots = \mu_I, \quad \text{contre} \quad \mathcal{H}_1 : \exists i, i' \in [\![1, I]\!], \quad \mu_i \neq \mu_{i'}.$$

Sous la contraintes $\alpha_1 = 0$, cela s'écrit

$$\mathcal{H}_0 : \alpha_2 = \ldots = \alpha_I = 0, \quad \text{contre} \quad \mathcal{H}_1 : \exists i, i' \in [\![1, I]\!], \quad \mu_i \neq \mu_{i'}.$$

- Le modèle réduit (celui associé à $\mathcal{H}_0$) ne fait intervenir que l'intercept et le bruit.
- Il s'agit du test de Fisher global vu dans le Chapitre 2 (pour $I-1$ paramètres testés). On connaît donc la loi de la statistique de test sous $\mathcal{H}_0$ (loi de Fisher $\mathcal{F}(I-1, n-I)$) ainsi que la zone de rejet du test pour un niveau α !

Test du modèle : retour sur la statistique de test et la zone de rejet

Statistique de test:

$$F = \frac{SCM/(I-1)}{SCR/(n-I)} \overset{\mathcal{H}_0}{\sim} \mathcal{F}(I-1, n-I).$$

Zone de rejet: un test de niveau α de $\mathcal{H}_0$ contre $\mathcal{H}_1$ a pour zone de rejet

$$\mathcal{R} = \left\{ F > q_{1-\alpha}^{\mathcal{F}(I-1,n-I)} \right\}.$$

p-valeur: $\mathbb{P}[F > f_{obs}]$, où f_{obs} est la valeur observée de F.

Décomposition de la variance: la formule $SCT = SCM + SCR$ s'écrit

$$\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{i,j}-\overline{Y})^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\widehat{Y}_{i,j}-\overline{Y})^2 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{i,j}-\widehat{Y}_{i,j})^2$$

$$= \underbrace{\sum_{i=1}^{I} n_i(\overline{Y}_{i\bullet}-\overline{Y})^2}_{\text{Variation inter-groupes}} + \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(Y_{i,j}-\overline{Y}_{i\bullet})^2}_{\text{Variation intra-groupes}}.$$

Avec R : test du modèle

```
summary(reg)
```

```
Call:
lm(formula = maxO3 ~ vent, data = donnees)

Residuals:
   Min     1Q Median     3Q    Max
-60.60 -16.81  -7.36  11.48  81.30

...

Residual standard error: 27.3 on 108 degrees of freedom
Multiple R-squared:  0.086,Adjusted R-squared:  0.0606
F-statistic: 3.39 on 3 and 108 DF,  p-value: 0.0207
```

Conclusion: la p-valeur (dernière ligne) est inférieure à 5%. On rejette $\mathcal{H}_0$ au niveau 5%. Le modèle ANOVA explique mieux les données qu'un modèle avec une concentration constante. La direction du vent a une influence.

Comparaison des moyennes par modalité

Motivation: si avec le test du modèle on rejette l'hypothèse d'égalité de tous les paramètres μ_i, on conclut qu'au moins deux paramètres μ_i et $\mu_{i'}$ sont différents. On peut donc chercher à identifier les couples (i, i') pour lesquels $\mu_i \neq \mu_{i'}$.

Pour comparer deux groupes $i et i'$ (aussi appelés traitements), on veut tester

$$\mathcal{H}_0 : \mu_i = \mu_{i'} \quad \text{contre} \quad \mathcal{H}_1 : \mu_i \neq \mu_{i'}.$$

Il s'agit du test d'égalité des moyennes de deux lois gaussiennes de même variance inconnue.

Statistique de test:

$$T = \frac{\widehat{\mu}_i - \widehat{\mu}_{i'}}{\sqrt{\widehat{\sigma}^2 \left(\frac{1}{n_i} + \frac{1}{n_{i'}} \right)}} \overset{\mathcal{H}_0}{\sim} \mathcal{T}(n_i + n_{i'} - 2).$$

p-valeur: $2\mathbb{P}[T > |t_{obs}|]$, avec t_{obs} la valeur observée de T.

Correction pour les tests multiples

Problématique: : pour identifier TOUS les couples (i, i') tels que $\mu_i \neq \mu_{i'}$, il faut réaliser $I(I-1)/2$ tests. On contrôle la probabilité de se « tromper » (rejeter $\mathcal{H}_0$ à tort) pour chaque test mais pas la probabilité de se « tromper » au moins une fois sur l'ensemble des test ! Cette probabilité augmente mécaniquement avec le nombre de tests réalisés.

Formellement: pour tout couple (i, i'), on teste

$$\mathcal{H}_0^{ii'} : \mu_i = \mu_{i'} \quad \text{contre} \quad \mathcal{H}_1^{ii'} : \mu_i \neq \mu_{i'},$$

au niveau α, *i.e.*, $\mathbb{P}\left[\text{rejeter } \mathcal{H}_0^{ii'} \;\middle|\; \mathcal{H}_0^{ii'} \text{ est vraie}\right] \leq \alpha$. On a alors

$$\begin{aligned}
\mathbb{P}[\text{Se tromper au moins une fois}] &= \mathbb{P}\left[\text{rejeter au moins une } \mathcal{H}_0^{ii'} \;\middle|\; \mathcal{H}_0^{ii'} \text{ est vraie}\right] \\
&\leq \sum_{i=1}^{I-1} \sum_{i'=i+1}^{I} \mathbb{P}\left[\text{rejeter } \mathcal{H}_0^{ii'} \;\middle|\; \mathcal{H}_0^{ii'} \text{ est vraie}\right] \\
&\leq \sum_{i=1}^{I-1} \sum_{i'=i+1}^{I} \alpha = \frac{I(I-1)}{2}\alpha
\end{aligned}$$

Correction pour les tests multiples : correction de Bonferroni

Pour $\alpha = 5\%$, dès que $I \geq 7$, on obtient donc que

$$\mathbb{P}[\text{Se tromper au moins une fois}] \leq 1.$$

On a **aucun contrôle sur l'erreur !**

Correction de Bonferroni: chaque test est fait au niveau $\delta = 2\alpha/[I(I-1)]$

- $\delta \leq \alpha$, il est donc plus difficile de rejet $\mathcal{H}_0^{ii'}$,
- le niveau global est α,
- la **p-valeur ajustée** est :

$$\frac{I(I-1)}{2} \times 2\mathbb{P}[T > |t_{obs}|] = I(I-1)\mathbb{P}[T > |t_{obs}].$$

Avec R : comparaison 2 à 2 des moyennes par modalité

Code sans correction de Bonferroni:

```
pairwise.t.test(donnees$maxO3, donnees$vent,
                p.adjust.method = "none")


Pairwise comparisons using t tests with pooled SD

data:  donnees$maxO3 and donnees$vent

      Est  Nord Ouest
Nord  0.05 -    -
Ouest 0.03 0.82 -
Sud   0.77 0.04 0.01

P value adjustment method: none
```

Conclusion: les directions (Nord, Ouest) et les directions (Sud, Est) ont le même effet sur la concentration en ozone (p−valeur $> 5\%$).

Avec R : comparaison 2 à 2 des moyennes par modalité

Code avec correction de Bonferroni:

```
pairwise.t.test(donnees$maxO3, donnees$vent,
                p.adjust.method = "bonferroni")

Pairwise comparisons using t tests with pooled SD

data:  donnees$maxO3 and donnees$vent

      Est  Nord Ouest
Nord  0.32 -    -
Ouest 0.18 1.00 -
Sud   1.00 0.22 0.08

P value adjustment method: bonferroni
```

Avec R : comparaison 2 à 2 des moyennes par modalité

Code avec correction de Benjamini & Hochberg:

```
pairwise.t.test(donnees$maxO3, donnees$vent,
                p.adjust.method = "BH")


Pairwise comparisons using t tests with pooled SD

data:  donnees$maxO3 and donnees$vent

      Est  Nord Ouest
Nord  0.08 -    -
Ouest 0.07 0.82 -
Sud   0.82 0.07 0.07

P value adjustment method: BH
```

Remarque pour aller plus loin...

Si l'hypothèse des résidus n'est pas vérifiée, il existe d'autres tests.

1. Pour vérifier que les distributions de deux modalités sont identiques. On utilise le test des rangs
   - le test de Wilcoxon dans le cadre de deux échantillons indépendants,
   - le test de Mann-Whitney dans le cadre de deux échantillons appariés.
2. Le test de Kruskall-Wallis généralise le test de Wilcoxon et permet de tester si plusieurs modalités ont la même distribution.

Test statistique : pertinence d'un coefficient (?)

Question: Comment interprète-t-on le test de Student dont la p-valeur est donnée dans la dernière colonne de `Coefficients` de la sortie de `summary` ?

Ce test correspond aux hypothèses

$$\mathcal{H}_0 : \text{le paramètre est nul} \quad \text{contre} \quad \mathcal{H}_1 : \text{le paramètre est non nul.}$$

Sous la contrainte $\alpha_1 = 0$, le test sur l'intercepte permet donc de décider si l'effet de la modalité de référence (celui associé à la modalité 1) est nul ou non. Pour les autres coefficients, le test permet de décider si l'effet du facteur est le même ou non que celui du facteur de référence.

Remarque: Ce test a en pratique peu d'intérêt car l'interprétation de ses hypothèses dépend de la contrainte.

Conclusion

Ce dont on n'a pas parlé...

- **Coefficient de détermination:** l'interprétation du R^2 et du R^2 ajusté reste inchangée. Sur notre exemple la capacité prédictive du modèle est très faible ($R^2_{adj} = 0.06063$).
- **Prédicteur et prédiction:** cette question nous intéresse peu. Pour le modèle factoriel, les nouvelles données pour la variable explicative sont une des modalités. Le prédicteur pour une modalité est la moyenne empirique associée aux données de la modalité. L'erreur de prévision est donc donnée par l'intervalle de confiance pour l'estimateur de l'espérance d'une loi normale de variance inconnue.

Ma feuille de route pour l'analyse de la variance à 1 facteur

1. Charger les données, vérifier que les variables sont bien de la nature attendue (variable réponse quantitative et variable explicative qualitative).
2. Exploration des données et calcul de statistiques descriptives.
3. Écrire le modèle linéaire. Appliquer la fonction `lm` aux données pour ajuster le modèle.
4. Analyser les graphes de résidus pour valider ou invalider les hypothèses du modèles.
5. Faire le test du modèle global : si on ne rejette pas l'hypothèse $\mathcal{H}_0$, on arrête là, le modèle linéaire n'est pas adapté.
6. Sinon critiquer le modèle, conclure.