Modèle Linéaire Gaussien
Chapitre 3 – Analyse de la variance

Achille Thin

22 Novembre 2022

Executive Master Statistique et Big Data

Analyse de la Variance à deux Facteurs

Objectifs du Chapitre

Cadre de l'analyse de la variance (ANOVA) à 2 facteurs: expliquer les variations d'**une variable quantitative**, appelée variable réponse, **en fonction de deux variables qualitatives**, appelées variables explicatives (ou facteurs).

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 et la météo ont-t-elles une influence sur le maximum de concentration d'ozone observé sur une journée ?

- **Variable réponse:** la concentration en ozone (MaxO3).
- **Variables explicatives:** la direction du vent (vent) et la météo (temps).

Chargement des données:

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

Analyse descriptive

On a $n = 112$ observations et 4 niveaux ou modalités (valeurs possibles) pour le vent et 2 modalités pour le temps :

```r
levels(vent)
```

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

```r
levels(temps)
```

```
[1] "Pluie" "Sec"
```

Effectifs par modalité:

```r
table(vent)
```

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

```r
table(temps)
```

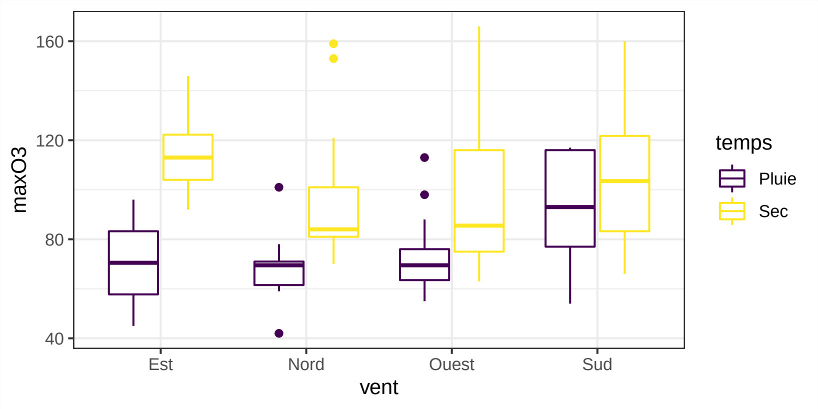
```
temps
Pluie   Sec 
   43    69 
```

Moyennes par modalité:

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

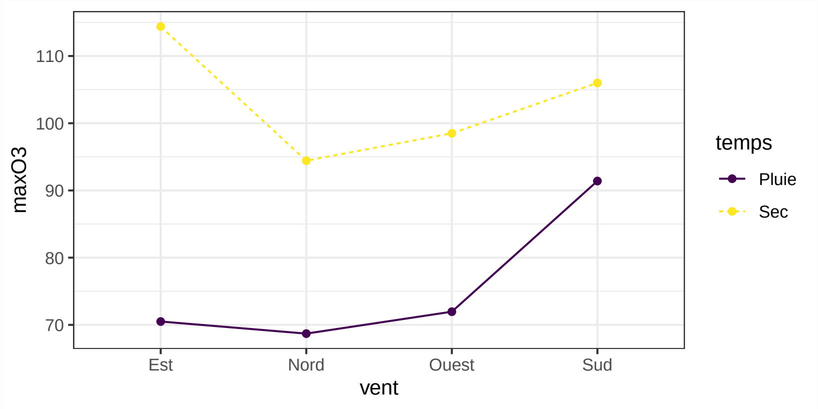
```
   vent temps   maxO3
1   Est Pluie  70.500
2  Nord Pluie  68.700
3 Ouest Pluie  71.962
4   Sud Pluie  91.400
5   Est   Sec 114.375
6  Nord   Sec  94.429
7 Ouest   Sec  98.500
8   Sud   Sec 106.000
```

Résumés numériques et graphiques



Analyse graphique: la météo semble avoir une influence sur la concentration en ozone. La conjonction de la direction et de la météo n'est pas évidente.

Graphe des interactions



Écriture du modèle

Notations

Variable explicative:

- Le premier facteur a I modalités. Le second facteur a J modalités.
- Chaque modalité du premier facteur, respectivement du second facteur, est codée par un entier i, $i \in [\![1, I]\!]$, respectivement par un entier j, $i \in [\![1, J]\!]$.
- Pour les modalités i et j, on dispose de $n_{i,j}$ observations.

Exemple: sur nos données, $I = 4$, Est $= 1$, Nord $= 2$, Ouest $= 3$, Sud $= 4$, et $J = 2$, Pluie $= 1$, Sec $= 2$. On a par exemple $n_{2,1} = 10$.

Variable réponse: on note $y_{i,j,k}$ la valeur de la variable réponse pour la k-ème observation de la modalité i pour le premier facteur et de la modalité j pour le second facteur.

Exemple: sur nos données $y_{4,2,3}$ désigne le maximum de concentration d'ozone observé la 3-ème journée où il y a eu un vent du sud et un temps sec.

Modèle régulier

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

$$Y_{i,j,k} = \mu_{i,j} + \varepsilon_{i,j,k}, \quad 1 \leq i \leq I = 4, \quad 1 \leq j \leq J = 2, \quad 1 \leq k \leq n_{i,j},$$

- $\mu_{i,j}$ sont des paramètres inconnus (μ_i représente la moyenne attendue de la concentration en ozone maximale journalière pour la direction de vent i et le temps j)
- $\varepsilon_{i,j,k}$ est une variable aléatoire appelée **bruit**, telle que toutes les variables aléatoires $(\varepsilon_{i,j,k})$ 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 $\varepsilon_{i,j,k}$ sont indépendantes et identiquement distribuées de loi $\mathcal{N}(0, \sigma^2)$.

Remarque: Pour les mêmes niveaux i et j, les variables aléatoires $Y_{i,j,k}$ sont *i.i.d.* suivant la loi normale $\mathcal{N}(\mu_{i,j}, \sigma^2)$.

Modèle régulier – Écriture matricielle

Le modèle peut s'écrire

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\mu} + \boldsymbol{\varepsilon}, \quad \text{avec} \quad \boldsymbol{\mu} = (\mu_{1,1}, \ldots, \mu_{I,J})^\intercal,$$

et

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

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)

$$\widehat{\mu} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}\mathbf{Y} \quad \rightsquigarrow \quad \widehat{\mu}_{i,j} = \frac{1}{n_{i,j}} \sum_{k=1}^{n_{i,j}} Y_{i,j,k} := \overline{Y}_{i,j\bullet}.$$

Pour les modalités $i \in [\![1, I]\!]$ et $j \in [\![1, J]\!]$, on obtient que les variables ajustées $\widehat{Y}_{i,j,k}$ sont égales à $\widehat{\mu}_{i,j}$, *i.e.*, la moyenne des observations pour ces deux modalités.

Estimateur de σ^2 (variable aléatoire)

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

Avec R:

```r
reg <- lm(max03 ~ vent * temps - 1, data = donnees)
```

Modèle singulier

Pour mieux analyser l'influence des facteur, on décompose $\mu_{i,j}$ sous la forme $\mu_{i,j} = \mu + \alpha_i + \beta_j + \gamma_{i,j}$:

- μ (paramètre inconnu) : effet moyen/référence,
- α_i (paramètres inconnus) : effet **principal** du niveau i du premier facteur,
- β_j (paramètres inconnus) : effet **principal** du niveau j du second facteur,
- $\gamma_{i,j}$ (paramètres inconnus) : effet **d'interaction** entre le premier facteur de niveau i et le second facteur de niveau j.

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

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + \gamma_{i,j} + \varepsilon_{i,j,k}, \quad 1 \leq i \leq I = 4, \quad 1 \leq j \leq J = 2, \quad 1 \leq k \leq n_{i,j}.$$

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

Modèle singulier

Cette nouvelle écriture conduit à un problème d'identifiabilité :

- On a décomposé $I \times J$ paramètres en $1 + I + J + I \times J$ paramètres. Mais on ne dispose que de $I \times J$ facteurs/équations $\rightsquigarrow$ singularité du modèle.
- Ayant $1 + I + J$ paramètres « en trop », il faut imposer $1 + I + J$ contraintes pour pouvoir ajuster le modèle.

Ajustement du modèle du modèle singulier

Contraintes et identifiabilité

Contrainte de type analyse par cellule ($\rightsquigarrow$ modèle régulier)

$$\mu = 0, \quad \forall i \in [\![1, I]\!],\ \alpha_i = 0, \quad \forall j \in [\![1, J]\!],\ \beta_j = 0.$$

Contrainte de type cellule de référence: les 1-ers niveaux de chaque facteur servent de niveaux de référence. C'est la **contrainte par défaut de R !**

$$\alpha_1 = 0, \quad \beta_1 = 0, \quad \forall i \in [\![1, I]\!], \forall j \in [\![1, J]\!], \quad \gamma_{1,j} = \gamma_{i,1} = 0.$$

Contrainte de type somme:

$$\sum_{i=1}^{I} \alpha_i = 0, \quad \sum_{j=1}^{J} \beta_j = 0, \quad \forall i \in [\![1, I]\!], \sum_{j=1}^{J} \gamma_{i,j} = 0, \quad \forall j \in [\![1, J]\!], \sum_{i=1}^{I} \gamma_{i,j} = 0.$$

Estimateurs et estimations

Important: les estimateurs des paramètres inconnus **dépendent des contraintes choisies**. On considère ici la contrainte par défaut de R

$$\alpha_1 = 0, \quad \beta_1 = 0, \quad \forall i \in [\![1, I]\!], \forall j \in [\![1, J]\!], \quad \gamma_{1,j} = \gamma_{i,1} = 0.$$

Estimateurs (variables aléatoires). L'estimateur $\widehat{\mu}$ de μ est donné par

$$\widehat{\mu} = \widehat{\mu}_{1,1} = \overline{Y}_{1,1\bullet}.$$

Pour tout $i \in [\![2, I]\!]$, $j \in [\![2, J]\!]$, les estimateurs $\widehat{\alpha}_i$ de α_i, $\widehat{\beta}_j$ de β_j et $\widehat{\gamma}_{i,j}$ de $\gamma_{i,j}$ sont donnés par

$$\widehat{\alpha}_i = \overline{Y}_{i,1\bullet} - \overline{Y}_{1,1\bullet}, \quad \widehat{\beta}_j = \overline{Y}_{1,j\bullet} - \overline{Y}_{1,1\bullet}, \quad \widehat{\gamma}_{i,j} = \overline{Y}_{i,j\bullet} - \overline{Y}_{i,1\bullet} - \overline{Y}_{1,j\bullet} + \overline{Y}_{1,1\bullet}.$$

Estimations (valeurs numériques calculées sur les données)

$$\widehat{\mu}^{\text{obs}} = \overline{y}_{1,1\bullet} \quad \text{et} \quad \forall i \in [\![2, I]\!], j \in [\![2, J]\!], \quad \begin{cases} \widehat{\alpha}_i^{\text{obs}} = \overline{y}_{i,1\bullet} - \overline{y}_{1,1\bullet}, \\ \widehat{\beta}_j^{\text{obs}} = \overline{y}_{1,j\bullet} - \overline{y}_{1,1\bullet}, \\ \widehat{\gamma}_{i,j}^{\text{obs}} = \overline{y}_{i,j\bullet} - \overline{y}_{i,1\bullet} - \overline{y}_{1,j\bullet} + \overline{y}_{1,1\bullet}. \end{cases}$$

Variables ajustées et résidus

Variables ajustées (variables aléatoires)

$$\widehat{Y}_{i,j,k} = \overline{Y}_{i,j\bullet}, \quad 1 \leq i \leq I, \quad 1 \leq j \leq J, \quad 1 \leq k \leq n_{i,j}.$$

Valeurs ajustées (réalisations sur les données)

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

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

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

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

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

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

Estimateur et estimation de σ^2

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

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

Cet estimateur vérifie

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

Estimation de σ^2 (réalisation sur les données)

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

Accès à la valeur estimée:

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

Changement de contrainte – Quelles conséquences ?

Ce qui change:

- l'expression des estimateurs de α_i, β_j et $\gamma_{i,j}$,
- 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,
- les prédictions.

Avec R : ajustement du modèle

Contrainte par défaut:

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

Estimation des paramètres:

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

```
      (Intercept)          ventNord         ventOuest 
          70.5000           -1.8000            1.4615 
          ventSud          tempsSec ventNord:tempsSec 
          20.9000           43.8750          -18.1464 
ventOuest:tempsSec  ventSud:tempsSec 
         -17.3365          -29.2750 
```

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 $\varepsilon_{i,j,k}$

(P1) sont **indépendantes**,

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

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

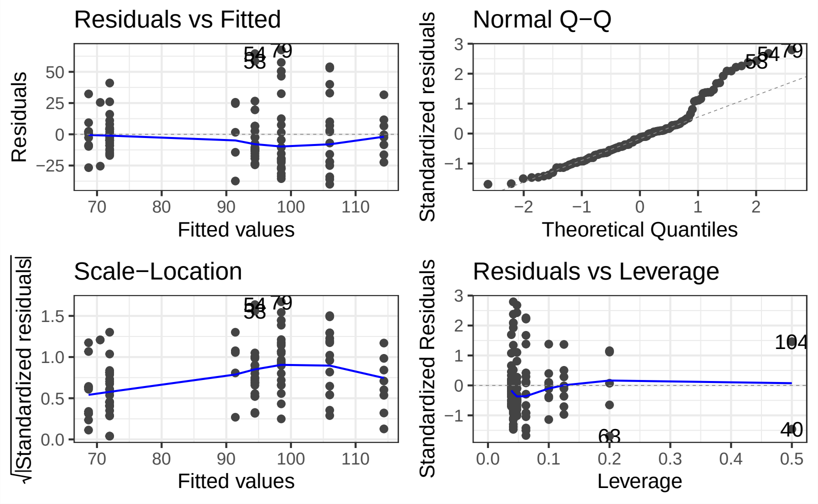
(P4) suivent une **loi normale**.

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

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



Tests d’hypothèses

Test du modèle ou test de Fisher global

Question: la direction du vent et la météo ont-t-elles une influence sur la concentration en ozone ?

Mathématiquement, cela revient à tester

$$\mathcal{H}_0 : Y_{i,j,k} = \mu + \varepsilon_{i,j,k}, \quad \text{contre} \quad \mathcal{H}_1 : Y_{i,j,k} = \mu + \alpha_i + \beta_j + \gamma_{i,j} + \varepsilon_{i,j,k}.$$

- 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 $IJ - 1$ paramètres testés). On connaît donc la loi de la statistique de test sous $\mathcal{H}_0$ (loi de Fisher $\mathcal{F}(IJ - 1, n - IJ)$) ainsi que la zone de rejet du test pour un niveau α !

Avec R : test du modèle

```r
summary(reg)
```

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

Residuals:
   Min     1Q Median     3Q    Max
-40.00 -15.97  -3.46   7.63  67.50


...

Residual standard error: 24.7 on 104 degrees of freedom
Multiple R-squared:  0.281,Adjusted R-squared:  0.232
F-statistic:  5.8 on 7 and 104 DF,  p-value: 0.0000109
```

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. Au moins un des facteurs a un effet.

Tests des effets

Question: Chacun des effets (effet principal du vent, effet principal du temps, effet d'interaction) dans le modèle est-il indispensable ?

Test de l'effet d'interaction: on teste si l'ajout d'un effet d'interaction à un modèle avec les effets principaux apporte de l'information sur la variable réponse.

Test des effets principaux: on peut réaliser deux types de test.

- **Test de type I:** l'ajout de l'effet principal d'un facteur est-il intéressant par rapport à un modèle constant ?
- **Test de type II:** L'ajout de l'effet principal d'un facteur est-il intéressant par rapport à un modèle comprenant déjà un effet ?

Une collection de modèle

Pour identifier l'influence de chacun des effets, on va mettre en compétition différents modèles.

$$
\begin{aligned}
\mathcal{M}_{\mu} &: \quad Y_{i,j,k} = \mu + \varepsilon_{i,j,k}, \\
\mathcal{M}_{\mu,\alpha} &: \quad Y_{i,j,k} = \mu + \alpha_i + \varepsilon_{i,j,k}, \\
\mathcal{M}_{\mu,\beta} &: \quad Y_{i,j,k} = \mu + \beta_j + \varepsilon_{i,j,k}, \\
\mathcal{M}_{\mu,\alpha,\beta} &: \quad Y_{i,j,k} = \mu + \alpha_i + \beta_j + \varepsilon_{i,j,k}, \\
\mathcal{M}_{\mu,\alpha,\beta,\gamma} &: \quad Y_{i,j,k} = \mu + \alpha_i + \beta_j + \gamma_{i,j} + \varepsilon_{i,j,k}.
\end{aligned}
$$

Remarques:

- $\mathcal{M}_{\mu}$: il n'y pas d'effet des facteurs.
- $\mathcal{M}_{\mu,\alpha,\beta}$: modèle dit additif car il suppose l'absence d'interaction entre les facteurs.
- $\mathcal{M}_{\mu,\alpha,\beta,\gamma}$: modèle dit complet car il prend en compte tous les effets (... Mais il requiert beaucoup de paramètres).

Test pour des modèles emboîtés

Modèles emboîtés: on dit qu'un modèle M_0 est emboîté dans un modèle M_1, lorsque M_1 s'obtient en ajoutant des paramètres à M_0. On note $M_0 \subset M_1$.

Test de Fisher: pour tester l'intérêt d'un modèle M_1 par rapport à un modèle M_0 tel que $M_0 \subset M_1$, on considère

$$\mathcal{H}_0 \ : \text{ le vrai modèle est } M_0 \quad \text{contre} \quad \mathcal{H}_1 \ : \text{ le vrai modèle est } M_1.$$

Statistique de test et zone de rejet au niveau α:

$$F = \frac{(SCR_0 - SCR_1)/q}{\underbrace{SCR_1/(n-p)}_{\text{Estimateur de } \sigma^2 \text{ pour } M_1}} \overset{\mathcal{H}_0}{\sim} \mathcal{F}(q, n-p), \quad \text{et} \quad \mathcal{R} = \left\{ F > q_{1-\alpha}^{\mathcal{F}(q,n-p)}, \right\}$$

- q est le nombre de paramètres supplémentaires à estimer dans M_1 par rapport à M_0 et p est le nombre de paramètres à estimer dans M_1,
- SCR_0 et SCR_1 sont respectivement la somme des carrés résiduelles pour M_0 et M_1.

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

Avec R : test de l'effet d'interaction

Test réalisé:

$$\mathcal{H}_0 \; : \text{ le vrai modèle est } \mathcal{M}_{\mu,\alpha,\beta} \quad \text{contre} \quad \mathcal{H}_1 \; : \text{ le vrai modèle est } \mathcal{M}_{\mu,\alpha,\beta,\gamma}.$$

Modèle $\mathcal{M}_{\mu,\alpha,\beta}$: le modèle additif se déclare en utilisant + à la place de $\star$

```
reg_add <- lm(maxO3 ~ vent + temps, data = donnees)
```

Lecture de la table d'analyse de la variance:

Effet	Res.Df	RSS	df	Sum of Sq	F	Pr(>F)
$\mathcal{M}_{\mu,\alpha,\beta}$	n - (p - q)	SCR_0				
$\mathcal{M}_{\mu,\alpha,\beta,\gamma}$	$n - p$	SCR_1	q	$SCR_0 - SCR_1$	f_{obs}	p-valeur = $\mathbb{P}[F > f_{obs}]$

Avec la contrainte par défaut de R, $q = (I - 1)(J - 1)$ et $p = IJ$.

Avec R : test de l'effet d'interaction

Résultat du test:

```
anova(reg_add, reg)
```

```
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    107 64446
2    104 63440  3      1006 0.55   0.65
```

On conserve l'hypothèse nulle au niveau 5% (p-valeur $>$ 5%). Il n'y a pas d'effet d'interaction.

Test des facteurs

Si le test précédent conclut qu'on peut enlever l'effet d'interaction (on accepte $\mathcal{H}_0$), on cherche à tester l'effet des facteurs.

Remarque: si le test des interactions conduit au rejet de $\mathcal{H}_0$ (il y a des interactions entre les deux facteurs), ce qui suit est inutile : les deux facteurs qui constituent cette interaction doivent impérativement être introduits dans le modèle.

Test sur le facteur temps: la variable temps est-elle pertinente ?

$$\mathcal{H}_0 \; : \text{le vrai modèle est } \mathcal{M}_{\mu,\alpha} \quad \text{contre} \quad \mathcal{H}_1 \; : \text{le vrai modèle est } \mathcal{M}_{\mu,\alpha,\beta}.$$

Test sur le facteur vent: la variable vent est-elle pertinente ?

$$\mathcal{H}_0 \; : \text{le vrai modèle est } \mathcal{M}_{\mu,\beta} \quad \text{contre} \quad \mathcal{H}_1 \; : \text{le vrai modèle est } \mathcal{M}_{\mu,\alpha,\beta}.$$

Avec R : test sur le facteur temps

Test sur le facteur temps: la variable temps est-elle pertinente ?

$$\mathcal{H}_0 \;:\; \text{le vrai modèle est } \mathcal{M}_{\mu,\alpha} \quad \text{contre} \quad \mathcal{H}_1 \;:\; \text{le vrai modèle est } \mathcal{M}_{\mu,\alpha,\beta}.$$

Modèle $\mathcal{M}_{\mu,\alpha}$: modèle ne contenant que le facteur vent

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

Lecture de la table d'analyse de la variance:

Effet	`Res.Df`	`RSS`	`df`	`Sum of Sq`	`F`	`Pr(>F)`
$\mathcal{M}_{\mu,\alpha}$	n - (p - q)	SCR_0				
$\mathcal{M}_{\mu,\alpha,\beta}$	$n-p$	SCR_1	q	$SCR_0 - SCR_1$	f_{obs}	p-valeur = $\mathbb{P}[F > f_{obs}]$

Avec la contrainte par défaut de R, on a $q = J - 1$ et $p = I + J - 1$.

Avec R : test sur le facteur temps

Résultat du test:

```
anova(reg_vent, reg_add)

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 26.8 1.1e-06 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: on rejette l'hypothèse nulle au niveau 5% (p-valeur $< 5\%$). Le temps a un effet sur la concentration en ozone.

Avec R : test sur le facteur vent

Test sur le facteur vent: la variable vent est-elle pertinente ?

$$\mathcal{H}_0 \ : \text{ le vrai modèle est } \mathcal{M}_{\mu,\beta} \quad \text{contre} \quad \mathcal{H}_1 \ : \text{ le vrai modèle est } \mathcal{M}_{\mu,\alpha,\beta}.$$

Modèle $\mathcal{M}_{\mu,\beta}$: modèle ne contenant que le facteur temps

```
reg_temps <- lm(maxO3 ~ temps, data = donnees)
```

Lecture de la table d'analyse de la variance:

Effet	Res.Df	RSS	df	Sum of Sq	F	Pr(>F)
$\mathcal{M}_{\mu,\beta}$	n - (p - q)	SCR_0				
$\mathcal{M}_{\mu,\alpha,\beta}$	$n-p$	SCR_1	q	$SCR_0 - SCR_1$	f_{obs}	p-valeur = $\mathbb{P}[F > f_{obs}]$

Avec la contrainte par défaut de R, on a $q = I - 1$ et $p = I + J - 1$.

Avec R : test sur le facteur vent

Résultat du test:

```
anova(reg_temps, reg_add)

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 2.1    0.1
```

Conclusion: on conserve l'hypothèse nulle au niveau 5% (p-valeur $>$ 5%). La prise en compte du facteur vent dans un modèle qui contient déjà le facteur temps n'a pas l'air concluante.

Avec R : anova et tests séquentiels

Que se passe-t-il avec la commande suivante ?

```
anova(reg_add)
```

On teste de façon séquentielle $\mathcal{M}_{\mu}$ v.s. $\mathcal{M}_{\mu,\alpha}$ et $\mathcal{M}_{\mu,\alpha}$ v.s. $\mathcal{M}_{\mu,\alpha,\beta}$.

Lecture de la table d'analyse de la variance:

Effet	`Df`	`Sum Sq`	`Mean Sq`	`F value`	`Pr(>F)`
$\mathcal{M}_{\mu,\alpha}$	I - 1	$\mathrm{SCR}_{\mu} - \mathrm{SCR}_{\mu,\alpha}$	`Sum Sq/Df`	f_{obs}	$\mathbb{P}[F > f_{obs}]$
$\mathcal{M}_{\mu,\alpha,\beta}$	J - 1	$\mathrm{SCR}_{\mu,\alpha} - \mathrm{SCR}_{\mu,\alpha,\beta}$	`Sum Sq/Df`	f_{obs}	$\mathbb{P}[F > f_{obs}]$
`Residuals`	$n - p$	$\mathrm{SCR}_{\mu,\alpha,\beta}$	$\widehat{\sigma}^2$		

Avec la contrainte par défaut de R, on a $p = I + J - 1$.

Remarques:

- L'ordre de la formule dans `lm` est l'ordre dans lequel j'ajoute les variables.
- La statistique de test est construite avec l'estimateur de la variance pour le modèle le plus complet (*i.e.*, celui ajusté avec la commande `lm` pour créer l'objet `reg_add`) et non l'estimation de la variance pour le modèle de l'hypothèse alternative. Pour la première ligne, on a

$$f_{obs} = \frac{(SCR_{\mu} - SCR_{\mu,\alpha})/(I-1)}{SCR_{\mu,\alpha,\beta}/(n-(I+J-1))} \neq \frac{(SCR_{\mu} - SCR_{\mu,\alpha})/(I-1)}{SCR_{\mu,\alpha}/(n-I).}$$

  Cela explique les différences numériques que l'on peut avoir au niveau de `f_obs` et de la p-valeur par rapport aux résultats de l'ANOVA à 1 facteur.

Avec R : anova et tests séquentiels

Résultats:

```
anova(reg_add)
Analysis of Variance Table

Response: maxO3
           Df Sum Sq Mean Sq F value   Pr(>F)
vent        3   7586    2529     4.2   0.0075 **
temps       1  16159   16159    26.8 1.1e-06 ***
Residuals 107  64446     602
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: on retrouve que l'ajout du facteur temps dans un modèle qui contient déjà le vent est pertinent.

Avec R : anova et tests séquentiels

Exemple n°1: on teste successivement $\mathcal{M}_{\mu}$ v.s. $\mathcal{M}_{\mu,\beta}$ et $\mathcal{M}_{\mu,\beta}$ v.s. $\mathcal{M}_{\mu,\alpha,\beta}$:

```
anova(lm(maxO3 ~ temps + vent, data = donnees))

Analysis of Variance Table

Response: maxO3
           Df Sum Sq Mean Sq F value  Pr(>F)
temps       1  19954   19954    33.1 8.3e-08 ***
vent        3   3791    1264     2.1     0.1
Residuals 107  64446     602
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Avec R : anova et tests séquentiels

Exemple n°2: on teste successivement $\mathcal{M}_{\mu}$ v.s. $\mathcal{M}_{\mu,\alpha}$, $\mathcal{M}_{\mu,\alpha}$ v.s. $\mathcal{M}_{\mu,\alpha,\beta}$ et $\mathcal{M}_{\mu,\alpha,\beta}$ v.s. $\mathcal{M}_{\mu,\alpha,\beta,\gamma}$:

```
anova(reg)

Analysis of Variance Table


Response: maxO3
            Df Sum Sq Mean Sq F value  Pr(>F)
vent         3   7586    2529    4.15  0.0081 **
temps        1  16159   16159   26.49 1.3e-06 ***
vent:temps   3   1006     335    0.55  0.6493
Residuals  104  63440     610
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Avec R : `Anova` et tests séquentiels

Que se passe-t-il avec la commande suivante ?

```
library(car)
Anova(reg)
```

On teste

$$\mathcal{M}_{\mu,\beta} \text{ v.s. } \mathcal{M}_{\mu,\alpha,\beta}, \quad \mathcal{M}_{\mu,\alpha} \text{ v.s. } \mathcal{M}_{\mu,\alpha,\beta} \quad \text{et} \quad \mathcal{M}_{\mu,\alpha,\beta} \text{ v.s. } \mathcal{M}_{\mu,\alpha,\beta,\gamma}.$$

Lecture de la table d'analyse de la variance:

Effet	`Sum Sq`	`Df`	`F value`	`Pr(>F)`
$\mathcal{M}_{\mu,\alpha,\beta}$	$\mathrm{SCR}_{\mu,\beta} - \mathrm{SCR}_{\mu,\alpha,\beta}$		f_{obs}	$\mathbb{P}[F > f_{obs}]$
$\mathcal{M}_{\mu,\alpha,\beta}$	$\mathrm{SCR}_{\mu,\alpha} - \mathrm{SCR}_{\mu,\alpha,\beta}$		f_{obs}	$\mathbb{P}[F > f_{obs}]$
$\mathcal{M}_{\mu,\alpha,\beta,\gamma}$	$\mathrm{SCR}_{\mu,\alpha,\beta} - \mathrm{SCR}_{\mu,\alpha,\beta,\gamma}$		f_{obs}	$\mathbb{P}[F > f_{obs}]$
`Residuals`	$\mathrm{SCR}_{\mu,\alpha,\beta,\gamma}$			

Avec la contrainte par défaut de R, on a $p = I + J - 1$.

Avec R : Anova et tests séquentiels

```
Anova(reg)

Anova Table (Type II tests)

Response: maxO3
            Sum Sq  Df F value   Pr(>F)
vent          3791   3    2.07     0.11
temps        16159   1   26.49 1.3e-06 ***
vent:temps    1006   3    0.55     0.65
Residuals    63440 104
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: on a les conclusions précédentes synthétisées en un tableau :

- effet du facteur temps dans un modèle contenant le vent (mais pas l'inverse),
- il n'y a pas d'effet d'interaction des facteurs.

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 ($R^2_{\text{adj}} = 0.23224$).
- **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 2 facteurs

1. Charger les données, vérifier que les variables sont bien de la nature attendue (variable réponse quantitative et variables explicatives qualitatives).
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. Faire le test des interactions :
    1. si on rejette l'hypothèse $\mathcal{H}_0$, on s'arrête là et on considère le modèle complet,
    2. si on accepte l'hypothèse $\mathcal{H}_0$, on teste l'effet des facteurs.
7. Critiquer le modèle, conclure.