Modèle Linéaire Gaussien
Chapitre 2 – Régression linéaire multiple

Achille Thin

7 Décembre 2023

Executive Master Statistique et Big Data

Objectifs du Chapitre

Cadre de la régression linéaire multiple: expliquer les variations d'**une variable quantitative**, appelée variable réponse, **en fonction de plusieurs variables quantitatives**, appelée variables explicatives (ou régresseurs).

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. Intervalles de confiance sur les paramètres
5. Test de la pertinence des différents éléments du modèle
6. Critique du modèle
7. Conclusion

Objectifs

Question : la nébulosité, la température et le vent à 12h permettent-ils d'expliquer le maximum de concentration d'ozone sur une journée ?

- **Variable réponse:** la concentration en ozone (MaxO3).
- **Variables explicatives:** la nébulosité (Ne12), la température (T12) et le vent (Vx12).

Chargement des données:

```
donnees <- read.table("../data/ozone.txt", header = TRUE)
```

Analyse descriptive

Coefficient de corrélation linéaire empirique

On a $n = 112$ observations. On peut calculer le coefficient de corrélation linéaire empirique pour chaque paire de variables quantitatives.

```
cor(donnees[, c("maxO3", "Ne12", "T12", "Vx12")])
```

```
          maxO3     Ne12      T12     Vx12
maxO3   1.00000 -0.64075  0.78426  0.43080
Ne12   -0.64075  1.00000 -0.66010 -0.51032
T12     0.78426 -0.66010  1.00000  0.31263
Vx12    0.43080 -0.51032  0.31263  1.00000
```

Écriture du modèle

Modèle de régression linéaire multiple

On suppose que y_k est la réalisation d'une variable aléatoire Y_k telle que :

$$Y_k = \beta_0 + \beta_1 x_{1,k} + \ldots + \beta_p x_{p,k} + \varepsilon_k = \beta_0 + \sum_{j=1}^{p} \beta_j x_{j,k} + \varepsilon_k, \quad 1 \leq k \leq n = 112,$$

- les $x_{j,k}$ sont les valeurs des p variables explicatives pour l'observation k. Ce sont des nombres connus (*i.e.*, non aléatoires),
- β_0 est un paramètre inconnu,
- $\beta_1, \ldots, \beta_p$ sont des paramètres inconnus (représentent les effets des variables explicatives sur le maximum de la concentration en ozone sur une journée),
- ε_k une variable aléatoire appelée **bruit**, telle que toutes les variables aléatoires $\varepsilon_1, \ldots, \varepsilon_n$ 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 ε_k sont indépendantes et identiquement distribuées de loi $\mathcal{N}(0, \sigma^2)$.

Écriture matricielle

Le modèle peut s'écrire

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

avec

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{p,1} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{1,n} & \dots & x_{p,n} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} \quad \text{et} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

1. On retrouve la même forme que pour le modèle de régression linéaire simple. C'est l'écriture générique du modèle linéaire.
2. Cette écriture simplifie grandement le problème d'estimation. Les résultats sont la version générale des résultats obtenus au Chapitre 1.

Hypothèse sur X

On suppose que la matrice **X est de plein rang en colonnes,** *i.e.*, on suppose les colonnes de $\mathbf{X}$ sont linéairement indépendantes.

Conséquence: il faut nécessairement $p \leq n$.

Propriété: si $\mathbf{X}$ est de plein rang en colonnes alors $\mathbf{X}^\intercal\mathbf{X}$ est une matrice $(p+1) \times (p+1)$ inversible.

Remarque: on a besoin de l'inversibilité de $\mathbf{X}^\intercal\mathbf{X}$ dans les résultats qui suivent. Si $\mathbf{X}$ n'est pas de plein rang en colonnes, on conserve certaines propriétés qui seront détaillées en Annexes.

Ajustement du modèle

Estimateur et estimation de β par les moindres carrés

Objectif: trouver $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\intercal$ qui s'ajuste le mieux aux données pour la perte quadratique.

L'estimateur des moindres carrés $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ est défini par

$$\widehat{\boldsymbol{\beta}} = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{j,k} \right)^2 = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \, \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

Estimateur (variable aléatoire)

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{Y}.$$

Estimation (valeur numérique calculée sur les données)

$$\widehat{\boldsymbol{\beta}}^{\text{obs}} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{y}.$$

Remarque: si $\mathbf{X}$ n'est pas de plein rang en colonnes, on perd l'unicité de cet estimateur (et la formule). On ne peut pas utiliser cette méthode.

Avec R : estimation de β

Construction du modèle de régression linéaire multiple:

```
reg <- lm(maxO3 ~ Ne12 + T12 + Vx12, data = donnees)
```

Estimation de β:

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

```
(Intercept)        Ne12         T12        Vx12
     3.8958     -1.6189      4.5132      1.6290
```

Remarque: si l'on veut éviter de préciser `data = donnees` dans la fonction `lm` (ou si l'on souhaite pouvoir accéder aux colonnes de `donnees` sans l'opérateur « $ »), on peut utiliser la fonction `attach`

```
attach(donnees)
```

Propriétés de l'estimateur $\widehat{\beta}$

Question: doit-on avoir confiance en notre estimation ? Si on avait un autre échantillon, l'estimation aurait-elle beaucoup varié ?

Loi de l'estimateur des moindres carrés: sous l'hypothèse du modèle linéaire gaussien,

$$\widehat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 (\mathbf{X}^\intercal \mathbf{X})^{-1}\right).$$

Conséquences: l'estimateur $\widehat{\beta}$ est

- sans biais, *i.e.*, $\mathbb{E}\left[\widehat{\beta}\right] = \beta$
- la variance dépend de la matrice de design $\mathbf{X}$: s'il y a une grande corrélation entre des variables explicatives la matrice $\mathbf{X}^\intercal \mathbf{X}$ est mal conditionnée. Certains coefficients de son inverse sont très grands... donc les variances des estimateurs sont très grandes !

Matrice $X^\intercal X$ mal conditionnée

La présence d'une forte corrélation entre variables explicatives est source de nombreux problèmes :

1. Pour des tests de significativité sur les coefficients, elle conduit à un rejet à tort de variables explicatives.
2. Elle cause de l'instabilité numérique des résultats (l'ajout ou la suppression d'une observation change beaucoup de choses).

Ce problème de colinéarité (ou multicolinéarité) est quasi systématique lorsque $p \leq n$ mais p est grand (n/p « proche » de 1). Dans ce cas, l'estimateur des moindre carré est sans intérêt du fait de sa très grande variabilité. On utilise alors d'autres méthodes que vous verrez plus tard : régression Lasso, régression Ridge, ...

Variables ajustées et résidus

Variables ajustées (variables aléatoires)

$$\widehat{Y} = X\widehat{\beta} = X(X^\intercal X)^{-1}X^\intercal Y, \quad \textit{i.e.,} \quad \widehat{Y}_k = \widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j x_{j,k}, \quad 1 \leq k \leq n.$$

Valeurs ajustées (réalisations sur les données)

$$\widehat{y} = X\widehat{\beta}^{\text{obs}} = X(X^\intercal X)^{-1}X^\intercal y, \quad \textit{i.e.,} \quad \widehat{y}_k = \widehat{\beta}_0^{\text{obs}} + \sum_{j=1}^{p} \widehat{\beta}_j^{\text{obs}} x_{j,k}, \quad 1 \leq k \leq n.$$

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

Résidus (variables aléatoires) estimateurs des erreurs inconnues ε_k (comme en régression simple) :

$$\widehat{\varepsilon} = (\widehat{\varepsilon}_1, \ldots, \widehat{\varepsilon}_n)^\intercal = Y - \widehat{Y}, \quad \textit{i.e.,} \quad \widehat{\varepsilon}_k = Y_k - \widehat{Y}_k.$$

Résidus observés: $\widehat{e}_k = y_k - \widehat{y}_k$.

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

Propriétés des résidus et estimateurs de σ^2

Propriétés des résidus: $\widehat{\varepsilon} = (\widehat{\varepsilon}_1, \dots, \widehat{\varepsilon}_n)^\intercal$ est un vecteur aléatoire

1. gaussien centré de variance $\sigma^2(\mathbb{I}_n - \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})\mathbf{X}^\intercal)$.
2. indépendant de $\widehat{\mathbf{Y}} = (\widehat{Y}_1, \dots, \widehat{Y}_n)^\intercal$ $\rightsquigarrow$ **Important!**

Remarque: on déduit de du point 1. que les résidus sont centrés comme les erreurs ε_k. En revanche ils n'ont pas même variance et ne sont pas indépendants en général.

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

$$S^2 = \frac{1}{n-(p+1)} \sum_{k=1}^{n} \widehat{\varepsilon}_k^2.$$

Cet estimateur vérifie

$$(n-(p+1))\frac{S^2}{\sigma^2} = \sum_{k=1}^{n} \frac{\widehat{\varepsilon}_k^2}{\sigma^2} \sim \chi^2(n-(p+1)),$$

où $\chi^2(n-(p+1))$ désigne la loi du Khi-deux à $n-(p+1)$ degrés de liberté.

Avec R : estimation de σ^2

Estimation de σ^2 (réalisation sur les données)

$$\widehat{\sigma}^2 = \frac{1}{n-(p+1)} \sum_{k=1}^{n} \widehat{e}_k^2.$$

Accès à la valeur estimée:

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

```
[1] 276.67
```

On en déduit une estimation de la variance de $\widehat{\beta}$:

$$\widehat{\sigma}^2 (X^\intercal X)^{-1}.$$

En particulier l'estimation de la variance de $\widehat{\beta}_j$ est

$$\widehat{\sigma}^2_{\beta_j} = \widehat{\sigma}^2 \left[(X^\intercal X)^{-1} \right]_{jj}.$$

L'écart type associé ($\widehat{\sigma}_{\beta_j}$) est donné dans la colonne `Std.Error` du tableau `Coefficients` de la sortie de `summary`.

Bilan

Dans cette partie, on a :

1. construit des estimateurs des paramètres inconnus β, σ^2 et des erreurs inconnues $\varepsilon_1, \ldots, \varepsilon_n$,
2. établi les propriétés théoriques de ces estimateurs sous les hypothèses du modèle linéaire gaussien,
3. obtenu les estimations à l'aide de la fonction `lm`.

Les résultats théoriques tout comme les valeurs fournies par `lm` supposent que les erreurs $\varepsilon_1, \ldots, \varepsilon_n$ sont *i.i.d.* suivant la loi normale $\mathcal{N}(0, \sigma^2)$. **Mais ce postulat est-il vérifié sur nos données ?**

Validité des hypothèses

Diagnostique des résidus

Les résidus observés permettent de valider les hypothèses du modèle linéaire gaussien, *i.e.*, les variables aléatoires $\varepsilon_1, \dots, \varepsilon_n$

(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),

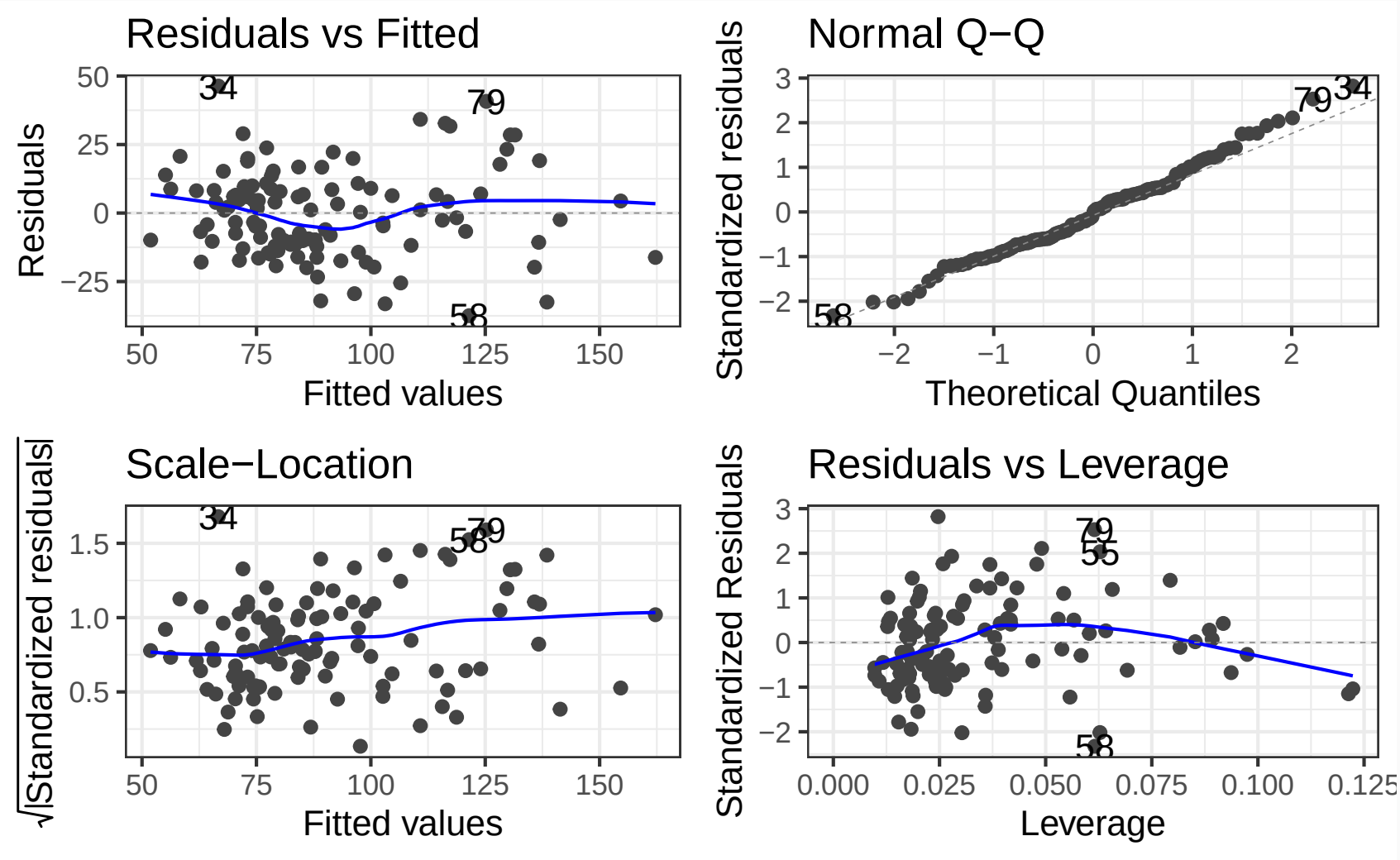
(P4) suivent une **loi normale**.

Validation des hypothèses:

- (P1) : l'indépendance ne peut être assurée que par le protocole expérimentale.
- (P2), (P3), (P4) : on fait la même analyse graphique que pour la régression linéaire simple.

Avec R : 4 graphiques à analyser

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



Fin de l'ananlyse graphique

Que faire si les hypothèses ne sont pas vérifiées ?

- Si le problème semble venir de (P2) on peut essayer de transformer la variable explicative, choisir une autre variable explicative ou un modèle plus complexe.
- S'il y a hétéroscédasticité des résidus, on peut essayer de transformer la variable à expliquer.
- On peut éliminer les éventuels points aberrants qui nuisent à la qualité de l'estimation.

Intervalles de confiance sur les paramètres

Intervalles de confiance pour β et σ^2

Pour le coefficient β_j, un intervalle de confiance bilatère symétrique de niveau de $1-\alpha$ est

$$\left[\widehat{\beta}_j \pm q_{1-\alpha/2}^{\mathcal{T}(n-(p+1))}\widehat{\sigma}_{\beta_j}\right] \quad \text{avec} \quad \widehat{\sigma}_{\beta_j} = \sqrt{\widehat{\sigma}^2\left[(\mathbf{X}^\intercal\mathbf{X})^{-1}\right]_{jj}}.$$

Notations: $\mathcal{T}(n-(p+1))$ désigne la loi de Student à $n-(p+1)$ degrés de liberté et $q_{1-\alpha/2}^{\mathcal{T}(n-(p+1))}$ le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{T}(n-(p+1))$.

Pour le paramètre σ^2, un intervalle de confiance bilatère de niveau de $1-\alpha$ est

$$\left[\frac{(n-(p+1))S^2}{q_{1-\alpha/2}^{\chi^2(n-(p+1))}}, \frac{(n-(p+1))S^2}{q_{\alpha/2}^{\chi^2(n-(p+1))}}\right].$$

Notations: $q_{\alpha/2}^{\chi^2(n-(p+1))}$ et $q_{1-\alpha/2}^{\chi^2(n-(p+1))}$ désignent les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la loi $\chi^2(n-(p+1))$.

Avec R : intervalles de confiance pour β et σ^2

Pour β:

```
confint(reg, level = 0.95)
```

```
                 2.5 %    97.5 %
(Intercept) -25.48865 33.28020
Ne12         -3.63685  0.39908
T12           3.48191  5.54456
Vx12          0.32647  2.93156
```

Pour σ^2:

```
alpha <- 0.05
dim_p <- length(coef(reg))
n <- nrow(donnees)
S2 <- summary(reg)$sigma^2
(n - dim_p) * S2/(qchisq(c(1 - alpha/2, alpha/2), n - dim_p))
```

```
[1] 215.51 368.29
```

Tests d'hypothèses

Test statistique : pertinence d'un coefficient

Question: on se demande si une des variables explicatives est utile dans le modèle. Par exemple, le vent influence-t-il la concentration en ozone ?

Mathématiquement, tester l'utilité d'une variable explicative x_j c'est regarder si le coefficient β_j associé est nul. On va donc tester l'hypothèse

$$\mathcal{H}_0 : \beta_j = 0 \quad \text{contre} \quad \mathcal{H}_1 : \beta_j \neq 0.$$

On retrouve le même test que pour la régression linéaire simple (à la différence que celui prend en compte qu'en plus de la variable testée, il y a d'autres variables explicatives dans le modèles).

Statistique de test:

$$T = \frac{\widehat{\beta}_j}{\sqrt{\widehat{\sigma}^2_{\widehat{\beta}_j}}} = \frac{\widehat{\beta}_j}{\sqrt{\widehat{\sigma}^2 \left[(\mathbf{X}^\intercal \mathbf{X})^{-1}\right]_{jj}}}.$$

Construction de la zone de rejet

Comment détermine-t-on la zone de rejet ? On part de l'estimateur $\widehat{\beta}_j$ (dont la loi dépend de β_j et σ^2 inconnues) puis on le transforme de sorte à obtenir une transformation dont la loi ne dépend plus des paramètres inconnus (loi pivotale) :

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{\sigma}_{\beta_j}} \sim \mathcal{T}(n-(p+1)).$$

Si $\mathcal{H}_0$ est vraie, on a en particulier

$$T = \frac{\widehat{\beta}_j}{\widehat{\sigma}_{\beta_j}} \sim \mathcal{T}(n-(p+1)).$$

La zone de rejet

$$\mathcal{R} = \left\{ |T| > q_{1-\alpha/2}^{\mathcal{T}(n-(p+1))} \right\} = \left\{ \frac{|\widehat{\beta}_j|}{\widehat{\sigma}_{\beta_j}} > q_{1-\alpha/2}^{\mathcal{T}(n-(p+1))} \right\}$$

fournit un test de niveau α de l'hypothèse $\mathcal{H}_0 : \beta_j = 0$ contre $\mathcal{H}_1 : \beta_j \neq 0$.

Test statistique : en pratique

p-valeur: les logiciels ne demandent pas de spécifier le niveau souhaité du test. Ils fournissent à la place une p-valeur. C'est le plus petit niveau pour lequel on rejette $\mathcal{H}_0$. Pour t_{obs} la valeur observée de T

$$\text{p-valeur} = 2\mathbb{P}[T > |t_{obs}|].$$

Interprétation: une variable est jugée pertinente au niveau α si la p-valeur est inférieur à α. Usuellement :

- si la p-valeur est inférieure à 5% on rejette $\mathcal{H}_0$ au niveau 5%. On conclut que la variable explicative à une influence sur la variable réponse.
- si la p-valeur est supérieure à 5%, on ne peut rejeter $\mathcal{H}_0$. On juge que la variable explicative n'a pas d'influence.

Avec R : pertinence de chacun des coefficients

Le test est fait dans la dernière colonne (Pr(> |t|)) du tableau `Coefficient` de la sortie de `summary`

```
summary(reg)


Call:
lm(formula = maxO3 ~ Ne12 + T12 + Vx12, data = donnees)

Residuals:
   Min     1Q Median     3Q    Max
-37.46 -11.45  -0.72   8.91  46.33

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.896     14.824    0.26    0.793
Ne12          -1.619      1.018   -1.59    0.115
T12            4.513      0.520    8.67  4.7e-14 ***
Vx12           1.629      0.657    2.48    0.015 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.6 on 108 degrees of freedom
Multiple R-squared:  0.661,Adjusted R-squared:  0.652
F-statistic: 70.3 on 3 and 108 DF,  p-value: <2e-16
```

Test statistique : pertinence d'un groupe de variables aléatoires

Question: on se demande si un groupe de variables explicatives est utile dans le modèle. Pour simplifier on va dire que ce groupe est constitué des q'p dernières variables explicatives (quitte à réordonner les colonnes de $\mathbf{X}$).

Mathématiquement, tester la pertinence de ce groupe c'est regarder si les coefficients $\beta_{p-q+1}, \dots, \beta_p$ sont simultanément nuls. On va donc tester l'hypothèse

$$\mathcal{H}_0 : \forall j \in [\![p-q+1, p]\!], \quad \beta_j = 0, \quad \text{contre} \quad \mathcal{H}_1 : \exists j \in [\![p-q+1, p]\!], \quad \beta_j \neq 0.$$

Sous $\mathcal{H}_0$, on regarde le modèle de régression pour la matrice de design $\mathbf{X}_0$ constitué des $p + 1 - q$ premières colonnes de $\mathbf{X}$, *i.e.*,

$$\mathbf{Y} = \mathbf{X}_0 \tilde{\boldsymbol{\beta}} + \boldsymbol{\varepsilon}, \quad \text{avec} \quad \tilde{\boldsymbol{\beta}} = (\beta_0, \dots, \beta_{p+1-q})^\intercal.$$

Pertinence d'un groupe : test de Fisher

Somme des carrés résiduels:

$$\mathrm{SCR} = \sum_{k=1}^{n} (Y_k - \widehat{Y}_k)^2 = \left\| \mathbf{Y} - \mathbf{X}(\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbf{Y} \right\|^2 .$$

Pour le modèle réduit (celui où on a enlever les q dernières variables, *i.e.* modèle associé à $\mathcal{H}_0$)

$$\mathrm{SCR}_0 = \left\| \mathbf{Y} - \mathbf{X}_0(\mathbf{X}_0^\intercal \mathbf{X}_0)^{-1} \mathbf{X}_0^\intercal \mathbf{Y} \right\|^2 \geq \mathrm{SCR}.$$

Le modèle le plus riche s'ajuste mieux aux données mais nécessite d'estimer davantage de paramètres (donc implique plus d'incertitude).

Test de Fisher: repose sur l'idée que l'on rejette $\mathcal{H}_0$ lorsque $\mathrm{SCR}_0 - \mathrm{SCR}$ est significativement plus grande que 0 tout en prenant en compte le nombre de paramètres estimés dans chacun des modèles en compétition.

Pertinence d'un groupe : statistique de test et zone de rejet

Statistique de test:

$$F = \frac{(SCR_0 - SCR)/\{n - (p+1-q) - [n-(p+1)]\}}{SCR/(n-(p+1))} = \frac{(SCR_0 - SCR)/q}{SCR/(n-(p+1))}.$$

Sous $\mathcal{H}_0$, F suit la loi de Fisher de degrés de liberté $(q, n-(p+1))$, notée $\mathcal{F}(q, n-(p+1))$.

Zone de rejet: si l'on note $q_{1-\alpha}^{\mathcal{F}(q,n-(p+1))}$ le quantile d'ordre $1-\alpha$ de la loi de Fisher de degrés de liberté $(q, n-(p+1))$,

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

fournit un test de niveau α de l'hypothèse $\mathcal{H}_0 : \forall j \in [\![p-q+1, p]\!], \quad \beta_j = 0$ contre $\mathcal{H}_1 : \exists j \in [\![p-q+1, p]\!], \quad \beta_j \neq 0$.

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

Avec R : pertinence d'un groupe

1. On ajuste le modèle réduit

```
reg_0 <- lm(maxO3 ~ T12, data = donnees)
```

2. On réalise le test de Fisher à l'aide de la fonction `anova`

```
anova(reg_0, reg)

Analysis of Variance Table

Model 1: maxO3 ~ T12
Model 2: maxO3 ~ Ne12 + T12 + Vx12
  Res.Df   RSS Df Sum of Sq    F Pr(>F)
1    110 33948
2    108 29881  2      4067 7.35  0.001 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. Analyser la p-valeur ($\Pr(>\mathrm{F})$)
   - Si la p-valeur est inférieure à 5%, on rejette $\mathcal{H}_0$ et on conserve le modèle qui inclut toutes les variables explicatives.
   - Si la p-valeur est supérieure à 5%, on conserve le modèle réduit.

Test du modèle ou test de Fisher global

Question: l'ensemble des variables explicatives est-il pertinent ?

Mathématiquement, tester la pertinence de l'ensemble des variables explicatives c'est regarder si les coefficients $\beta_1, \dots, \beta_p$ sont nuls. On va donc tester l'hypothèse

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

- Le modèle réduit (celui associé à $\mathcal{H}_0$) ne fait intervenir que l'intercept et le bruit. On a alors

$$SCR_0 = \sum_{k=1}^{n} (Y_k - \overline{Y})^2 = SCT \quad \Rightarrow \quad SCR_0 - SCR = SCM.$$

- Il s'agit du test de Fisher que nous venons de voir pour $q = p$. On connaît donc la loi de la statistique de test sous $\mathcal{H}_0$ ($\mathcal{F}(p, n - (p+1))$) ainsi que la zone de rejet du test pour un niveau α !

Avec R : test du modèle

Le test est fait par R avec la fonction `summary` (résultats sur la dernière ligne).

```
Call:
lm(formula = maxO3 ~ Ne12 + T12 + Vx12, data = donnees)

Residuals:
   Min     1Q Median     3Q    Max
-37.46 -11.45  -0.72   8.91  46.33

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.896     14.824    0.26    0.793
Ne12          -1.619      1.018   -1.59    0.115
T12            4.513      0.520    8.67  4.7e-14 ***
Vx12           1.629      0.657    2.48    0.015 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.6 on 108 degrees of freedom
Multiple R-squared:  0.661,Adjusted R-squared:  0.652
F-statistic: 70.3 on 3 and 108 DF,  p-value: <2e-16
```

Conclusion: la p-valeur est inférieure à 5%. On rejette $\mathcal{H}_0$ au niveau 5%. Le modèle de régression multiple explique mieux les données qu'un modèle avec une concentration constante.

Critique du modèle

Décomposition de la variance:

$$\underbrace{\sum_{k=1}^{n}(y_k - \overline{y})^2}_{\text{SCT}} = \underbrace{\sum_{k=1}^{n}(\widehat{y}_k - \overline{y})^2}_{\text{SCM}} + \underbrace{\sum_{k=1}^{n}(y_k - \widehat{y}_k)^2}_{\text{SCR}}$$

- SCT correspond à la variabilité des données.
- SCM correspond la variabilité expliquée par le modèle (*i.e.*, par la variable explicative).
- SCR correspond à la variabilité résiduelle, *i.e.*, la variabilité non-expliquée par le modèle.

Plus SCM est proche de SCT, plus le modèle explique la variabilité des observations.

Coefficient de détermination

Coefficient de détermination: part de variance expliquée par le modèle

$$R^2 = \frac{SCM}{SCT} = 1 - \frac{SCR}{SCT}.$$

Avec R: c'est la sortie `Multiple R-squared` obtenue avec `summary(reg)`.

```
summary(reg)$r.squared
```

```
[1] 0.66118
```

Interprétation: si R^2 n'est pas assez proche de 1 alors cela signifie que le modèle n'approche pas bien les données. Deux possibilités

1. il manque une variable explicative,
2. une (ou plusieurs) des variables explicatives n'intervient pas de manière linéaire.

Limites du R^2

On pourrait se dire qu'on essaie d'inclure une nouvelle variable puis on choisit de conserver le modèle ayant le plus grand R^2. Malheureusement, **R^2 ne peut être utilisé pour sélectionner une variable explicative pertinente ou comparer des modèles de tailles différentes !**

Pourquoi ? On considère le modèle à q variables explicatives de matrice de design $\mathbf{X}_q$. En ajoutant une variable explicative, on obtient une matrice de design $\mathbf{X}_{q+1}$ et

$$\min_{\beta \in \mathbb{R}^{q+1}} \|\mathbf{Y} - \mathbf{X}_{q+1}\beta\| \leq \min_{\tilde{\beta} \in \mathbb{R}^q} \left\| \mathbf{Y} - \mathbf{X}_{q+1} \begin{pmatrix} \tilde{\beta} \\ 0 \end{pmatrix} \right\| = \min_{\tilde{\beta} \in \mathbb{R}^q} \left\| \mathbf{Y} - \mathbf{X}_q \tilde{\beta} \right\|.$$

Autrement dit, la somme des carrés résiduels dimine (et donc R^2 augmente) à chaque fois qu'on ajoute une variable explicative (et cela peu importe qu'elle soit pertinente ou non) ! L'augmentation de R^2 est mécanique et non informative.

Conclusion: R^2 peut être intéressant pour comparer des modèles de même dimension.

Coefficient de détermination ajusté

Coefficient de détermination ajusté: part de variance expliquée par le modèle prenant en compte le nombre de variables explicatives utilisées

$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-(p+1)} \frac{\text{SCR}}{\text{SCT}}.$$

Avec R: c'est la sortie `Adjusted R-squared` obtenue avec `summary(reg)`. On peut y accéder directement avec

```
summary(reg)$adj.r.squared
```

```
[1] 0.65177
```

Remarque: il existe d'autres critères de sélection de variables que vous verrez ultérieurement (*e.g.*, BIC, AIC).

Predicteur et prédiction

Objectif de la régression: soient $\mathbf{x}_{n+1} = (x_{1,n+1}, \ldots, x_{p,n+1})^\intercal$ les valeurs des variables explicatives pour un nouvel individu/une nouvelle donnée. On ne connaît pas le y_{n+1} associé et on cherche à le prédire. Par exemple, on cherche à prédire la concentration maximale en ozone (maxO3) pour une nouvelle journée en mesurant uniquement la température (T12), la nébulosité (Ne12) et la quantité de vent (Vx12).

Prédicteur (variable aléatoire) : c'est la valeur moyenne attendue Y_k^p pour $x_{1,n+1}, \ldots, x_{p,n+1}$ sous le modèle ajusté

$$\widehat{Y}_k^p = \mathbf{x}_{n+1}^\intercal \widehat{\boldsymbol{\beta}} = \widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j x_{j,n+1}.$$

Prédiction (réalisation utilisant les données) :

$$\widehat{y}_k^p = \mathbf{x}_{n+1}^\intercal \widehat{\boldsymbol{\beta}}^{\text{obs}}.$$

Remarque: la valeur $\mathbf{x}_{n+1}$ pour laquelle on fait la prédiction n'a pas servi pour estimer $\boldsymbol{\beta}$.

Avec R : prévisions

Pour calculer les prévisions $\widehat{y}^p_{n+1}, \ldots, \widehat{y}^p_{n+N}$ à partir de nouvelles valeurs $x_{n+1}, \ldots, x_{n+N}$ des nouvelles variables explicatives :

1. On met les nouvelles valeurs dans un `data.frame` dont le nom des colonnes est le même que le nom des variables explicatives dans les données d'origine

```
x_new <- data.frame(Ne12 = 6, T12 = 20, Vx12 = -3)
```

2. On utilise la fonction `predict`

```
predict(reg, newdata = x_new)
```

```
    1
79.56
```

Erreur de prévision

Erreur de prévision: erreur commise entre la valeur $y_n + 1$ (inconnue) à prévoir et celle qu'on prédit :

$$\widehat{\varepsilon}^p_{n+1} = y_{n+1} - \widehat{Y}^p_{n+1}.$$

Elle quantifie la capacité du modèle à prévoir. Elle inconnue car $y_n + 1$ est inconnue.

Résultat théorique: $\widehat{\varepsilon}^p_{n+1}$ est une variable aléatoire qui vérifie

$$\mathbb{E}\left[\widehat{\varepsilon}^p_{n+1}\right] = 0 \quad \text{et} \quad \mathbb{V}\text{ar}\left[\widehat{\varepsilon}^p_{n+1}\right] = \sigma^2\left[1 + \mathbf{x}_{n+1}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{x}_{n+1}\right].$$

Remarque: $\mathbf{x}_{n+1}^\intercal(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{x}_{n+1}$ est en lien avec la distance (non euclidienne) entre $\mathbf{x}_{n+1}$ et $\overline{\mathbf{x}}$. La variance est d'autant plus grande que $\mathbf{x}_{n+1}$ est loin de $\overline{\mathbf{x}}$.

Avec R : intervalle de prédiction

Pour y_{n+1}: un intervalle de confiance bilatère symétrique de niveau de $1-\alpha$ est

$$\left[\widehat{Y}^p_{n+1} \pm q^{\mathcal{T}(n-(p+1))}_{1-\alpha/2} \sqrt{\widehat{\sigma}^2 \left[1 + \mathbf{x}^\intercal_{n+1}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{x}_{n+1}\right]}\right].$$

Avec R: on utilise les arguments `interval` et `level` (pour $1-\alpha$) de la fonction `predict`

```
predict(reg, newdata = x_new, interval = "prediction",
        level = .95)
```

```
    fit    lwr    upr
1 79.56 46.397 112.72
```

Conclusion

Quelques remarques pour conclure

- Les résultats de la régression linéaire multiple viennent généraliser les résultats du Chapitre 1 sur la régression linéaire simple (estimateurs et test sont des cas particuliers pour $p = 1$).
- **Modèle sans intercept:** à de rares occasions on peut s'intéresser à un modèle où on ne veut pas inclure coefficient constant β_0 (la première colonne de la matrice de design $\mathbf{X}$ ne contiendra pas des 1 mais les valeurs de la première variable explicative). Par défaut R inclut β_0. Pour ajuster un modèle sans β_0, on retranche 1 dans la formule de `lm`.

```
lm(maxO3 ~ Ne12 + T12 + Vx12 - 1, data = donnees)
```

Les résultats de ce chapitre restent vrais (modulo qu'il faille remplacer $p+1$ par p dans les différentes formules).

Ma feuille de route pour la régression linéaire multiple

1. Charger les données, vérifier que les variables sont bien de la nature attendue (quantitatives).
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. Le cas échéant, modifier le modèle pour obtenir des résidus satisfaisants (transformation log de certaines variables explicatives ou de y).
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, faire les test sur des paramètres, faire de la prédicition, conclure.