Chapitre 3. **Régression linéaire mutiple**

Cours de modèle linéaire gaussien par S. Donnet

Executive Master Statistique et Big-Data

Août 2020

Introduction

- La modélisation précédente de la concentration d'ozone dans l'atmosphère est simpliste.
- D'autres variables météorologiques sont susceptibles d'avoir une influence sur la variable `maxO3`, comme la quantité de vent, la température ou la nébulosité.
- Pour analyser la relation entre la température (`T12`), le vent (`Vx12`), la nébulosité (`Ne12`) et l'ozone (`maxO3`), nous cherchons une fonction $f : \mathbb{R}^3 \to \mathbb{R}$ telle que

$$\texttt{maxO3}_i \approx f(\texttt{T12}_i, \texttt{Vx12}_i, \texttt{Ne12}_i).$$

- Si on suppose que le lien est linéaire :

$$f(x_1, x_2, x_3) = \beta_1 x^1 + \beta_2 x^2 + \beta_3 x^3$$

- Estimer f revient à estimer $\beta_1, \beta_2, \beta_3$.

Principe

- On veut apprendre la relation entre la variable y et p variables explicatives $(x^1, \dots, x^p)$

$$y \approx f(x^1, \dots, x^p)$$

- Hypothèse sur f : $\exists(\beta_1 \dots, \beta_p)|$

$$y \approx \beta_1 x^1 + \dots + \beta_p x^p$$

- On dipose de n observations $(y_i, x_i^1, \dots x_i^p)$ pour apprendre cette relation.
- On cherche $\beta = (\beta_1, \dots, \beta_p)$ tels que

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta_1 x_i^1 - \dots \beta_p x_i^p)^2 \\ &= \sum_{i=1}^{n} (y_i - \sum_{j=1}^{n} \beta_j x_i^j)^2\end{aligned}$$

Notations

- y_i est la i-ème observation.
- $\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$.
- x_i^j la mesure de la j-ème variable sur le i-ème observation.
- X est une matrice à n lignes et p colonnes telle que

$$X_{ij} = x_i^j$$

  X est la *matrice de design*.
- x^j la j-ème colonne de X.
- x_i est un vecteur ligne représentant les p variables de l'observation i.

Modèle linéaire gaussien

Nous supposons donc que les données y_i sont la réalisation de Y_i telle que :

$$Y_i = \beta_1 x_i^1 + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \varepsilon_i, \quad i = 1, \ldots, n \tag{1}$$

où

- Les x_i^j sont des nombres connus, déterministes (i.e. non aléatoires).
- Les paramètres β_j sont déterministes et inconnus.
- Les variables ε_i sont des variables aléatoires iid de loi $\mathcal{N}(0, \sigma^2)$ avec σ^2 inconnu.
- Mêmes postulats [P1-4] que dans le chapitre précédent.

A propos de l'intercepte

- Par simplicité : on a supposé f linéaire.
- En réalité, l'immense majorité du temps, on la suppose affine, c'est-à-dire que

$$f(x_1, \dots, x_p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

- On introduit donc un vecteur x^0 de coordonnées toutes égales à 1.
- Plus simplement on va écrire

$$Y_i = \beta_1 x_i^1 + \beta_2 x_i^2 + \dots + \beta_p x_i^p + \varepsilon_i, \quad i = 1, \dots, n$$

et considérer que la première variable est l'intercepte, c'est-à-dire la variable constante égale à 1. Autrement dit, $x_i^1 = 1$ pour tout i, i.e. $x^1 = \mathbf{1}$ le vecteur de $\mathbb{R}^n$ de coordonnées toutes égales à 1.

Ecriture matricielle

- $$\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$
- $$Y_i = \sum_{j=1}^{p} \beta_j x_i^j + \varepsilon_i, \quad i = 1, \ldots, n$$

  se réécrivent

  $$\boldsymbol{Y} = X\beta + \boldsymbol{\varepsilon}$$
- $\boldsymbol{\varepsilon}$ vérifie

  $$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n).$$
- Comme X et β sont déterministes, Y est aussi un vecteur gaussien et

  $$Y \sim \mathcal{N}(X\beta, \sigma^2 I_n). \tag{2}$$

Hypothèse sur X

Hypothèse fondamentale concernant X dans le reste du cours : on suppose que

$$X \text{ est de plein rang en colonnes}$$

- i.e. on suppose que les variables explicatives (colonnes de X) sont linéairement indépendantes.
- Alors obligatoirement

$$p \leq n$$

- De plus, une colonne ne peut pas être combinaison linéaire des autres (sinon on ne pourra pas distinguer les effets).

Estimateur des moindres carrés ordinaire (EMC)

- Comme pour la régression linéaire simple, nous choisissons la fonction de coût quadratique, d'où la dénomination EMC.

Estimateur des moindres carrés ordinaire (EMC)

- Comme pour la régression linéaire simple, nous choisissons la fonction de coût quadratique, d'où la dénomination EMC.
- L'estimateur des moindres carrés $\hat{\beta}$ est défini comme suit

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \Big(y_i - \sum_{j=1}^{p} X_{ij}\beta_j \Big)^2$$

Estimateur des moindres carrés ordinaire (EMC)

- Comme pour la régression linéaire simple, nous choisissons la fonction de coût quadratique, d'où la dénomination EMC.
- L'estimateur des moindres carrés $\hat{\beta}$ est défini comme suit

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} X_{ij}\beta_j \right)^2$$

- Comme

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} X_{ij}\beta_j \right)^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} (X\beta)_i \right)^2 = \|y - X\beta\|^2$$

on peut aussi écrire

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|^2$$

Calcul de l'EMC $\hat{\beta}$

Théorème

Si X est de plein rang en colonnes alors

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

- Preuve par algèbre linéaire ou par calcul différentiel.

Calcul de l'EMC $\hat{\beta}$

Théorème

Si X est de plein rang en colonnes alors

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

- Preuve par algèbre linéaire ou par calcul différentiel.
- L'EMC $\hat{\beta} = (X^T X)^{-1} X^T Y$ est un estimateur linéaire de β.

Si X n'est pas de plein rang en colonnes ?

- $\hat{y} = P_{[X]}y$ reste l'unique solution du problème mais ne s'exprime plus comme $X(X^TX)^{-1}X^Ty$ puisque la matrice X^TX n'est plus inversible. Mais cette projection existe toujours et est unique.
- $\hat{y} = P_{[X]}y$ peut toujours s'écrire sous la forme $\hat{y} = X\hat{\beta}$ mais cette écriture n'est pas unique. En effet, l'équation matricielle $X\hat{\beta} = \hat{y}$ a toujours au moins une solution puisque $\hat{y} = Py \in \mathrm{Image}(X)$, mais dans ce cas on a même une infinité de solutions données par

$$\hat{\beta}_0 + \mathrm{Ker}(X),$$

avec $\hat{\beta}_0$ une solution particulière.

Sous R

- Expliquer la concentration en ozone `maxO3` par la nébulosité `Ne12`, le vent `Vx12` et la température `T12`.
- Fonction `lm()`. Inutile de préciser que la première variable explicative est la constante **1**, R la met automatiquement.
- On peut cependant demander à R de l'enlever, en écrivant
  `maxO3 ~ Ne12 + Vx12 + T12 - 1`
- On peut utiliser la fonction `attach` pour éviter de préciser dans la fonction `lm()` qu'on utilise `data = ozone`.

```
attach(ozone)
reg=lm(maxO3~Ne12+Vx12+T12)
summary(reg)
```

Résultats

```
##
## Call:
## lm(formula = maxO3 ~ Ne12 + Vx12 + T12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.462 -11.448  -0.722   8.908  46.331
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.8958    14.8243   0.263   0.7932
## Ne12         -1.6189     1.0181  -1.590   0.1147
## Vx12          1.6290     0.6571   2.479   0.0147 *
## T12           4.5132     0.5203   8.674 4.71e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.63 on 108 degrees of freedom
## Multiple R-squared:  0.6612, Adjusted R-squared:  0.6518
## F-statistic: 70.25 on 3 and 108 DF,  p-value: < 2.2e-16
```

On voit donc que $\hat{\beta} = (3.8958, -1.6189, 1.6290, 4.5132)^T$.

Loi de l'EMC

On s'intéresse maintenant aux propriétés probabilistes de l'estimateur $\hat{\beta} = (X^T X) X^T \boldsymbol{Y}$.

Theorem (Loi de l'EMC)

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})$$

Conséquences

- L'EMC de $\hat{\beta}$ est sans biais.
- L'opérateur de covariance fait intervenir **l'inverse** de $X^T X$.

Colinéarité des variables explicatives

- La variance de l'EMC dépend du design

Colinéarité des variables explicatives

- La variance de l'EMC dépend du design
- Si grande corrélation entre variables alors

Colinéarité des variables explicatives

- La variance de l'EMC dépend du design
- Si grande corrélation entre variables alors
  - Des coefficients de $(X^\top X)^{-1}$ peuvent devenir énormes

Colinéarité des variables explicatives

- La variance de l'EMC dépend du design
- Si grande corrélation entre variables alors
  - Des coefficients de $(X^\top X)^{-1}$ peuvent devenir énormes
  - Les variances des estimateurs explosent

Colinéarité des variables explicatives

- La variance de l'EMC dépend du design
- Si grande corrélation entre variables alors
  - Des coefficients de $(X^\top X)^{-1}$ peuvent devenir énormes
  - Les variances des estimateurs explosent
  - Tests de significativité sur ces coefficients font accepter l'hypothèse de nullité d'où rejet à tort de variable significative.

Colinéarité des variables explicatives

- La variance de l'EMC dépend du design
- Si grande corrélation entre variables alors
  - Des coefficients de $(X^\top X)^{-1}$ peuvent devenir énormes
  - Les variances des estimateurs explosent
  - Tests de significativité sur ces coefficients font accepter l'hypothèse de nullité d'où rejet à tort de variable significative.
  - Résultats instables, adjonction ou suppression de quelques observations bouleverse valeurs et signes des coefficients.

Colinéarité des variables explicatives

- La variance de l'EMC dépend du design
- Si grande corrélation entre variables alors
  - Des coefficients de $(X^\top X)^{-1}$ peuvent devenir énormes
  - Les variances des estimateurs explosent
  - Tests de significativité sur ces coefficients font accepter l'hypothèse de nullité d'où rejet à tort de variable significative.
  - Résultats instables, adjonction ou suppression de quelques observations bouleverse valeurs et signes des coefficients.
  - Résultats "bizarres"

Colinéarité des variables explicatives

- La variance de l'EMC dépend du design
- Si grande corrélation entre variables alors
  - Des coefficients de $(X^\top X)^{-1}$ peuvent devenir énormes
  - Les variances des estimateurs explosent
  - Tests de significativité sur ces coefficients font accepter l'hypothèse de nullité d'où rejet à tort de variable significative.
  - Résultats instables, adjonction ou suppression de quelques observations bouleverse valeurs et signes des coefficients.
  - Résultats "bizarres"
- Problèmes avec p grand : EMC inutilisable

Colinéarité des variables explicatives

- La variance de l'EMC dépend du design
- Si grande corrélation entre variables alors
  - Des coefficients de $(X^\top X)^{-1}$ peuvent devenir énormes
  - Les variances des estimateurs explosent
  - Tests de significativité sur ces coefficients font accepter l'hypothèse de nullité d'où rejet à tort de variable significative.
  - Résultats instables, adjonction ou suppression de quelques observations bouleverse valeurs et signes des coefficients.
  - Résultats "bizarres"
- Problèmes avec p grand : EMC inutilisable
  - Si $n \leq p$ alors EMC non unique et plus de formule.

Colinéarité des variables explicatives

- La variance de l'EMC dépend du design
- Si grande corrélation entre variables alors
  - Des coefficients de $(X^\top X)^{-1}$ peuvent devenir énormes
  - Les variances des estimateurs explosent
  - Tests de significativité sur ces coefficients font accepter l'hypothèse de nullité d'où rejet à tort de variable significative.
  - Résultats instables, adjonction ou suppression de quelques observations bouleverse valeurs et signes des coefficients.
  - Résultats "bizarres"
- Problèmes avec p grand : EMC inutilisable
  - Si $n \leq p$ alors EMC non unique et plus de formule.
  - Si $n \geq p$ mais p grand alors le risque de corrélation entre variable est grand.

Colinéarité des variables explicatives

- La variance de l'EMC dépend du design
- Si grande corrélation entre variables alors
  - Des coefficients de $(X^\top X)^{-1}$ peuvent devenir énormes
  - Les variances des estimateurs explosent
  - Tests de significativité sur ces coefficients font accepter l'hypothèse de nullité d'où rejet à tort de variable significative.
  - Résultats instables, adjonction ou suppression de quelques observations bouleverse valeurs et signes des coefficients.
  - Résultats "bizarres"
- Problèmes avec p grand : EMC inutilisable
  - Si $n \leq p$ alors EMC non unique et plus de formule.
  - Si $n \geq p$ mais p grand alors le risque de corrélation entre variable est grand.
  - Autres techniques nécessaires (sélection de variables : AIC, BIC, etc ou pénalisation : Lasso, ridge, etc.)

Prédiction et Résidus

- *Valeur prédites* : $\forall i = 1, \ldots, n$

$$\hat{y}_i = \sum_{j=1}^{p} \hat{\beta}_j x_i^j = x_i \hat{\beta}$$

- Donc, en adoptant l'écriture matricielle on a

$$\hat{\boldsymbol{y}} = X\hat{\beta} = X(X^T X)^{-1} X' Y = P_{[X]}\boldsymbol{y}$$

.

- *Résidus*

$$\hat{e}_i = \hat{y}_i - y_i, \quad \text{et} \quad \hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

- Si $\hat{\boldsymbol{\varepsilon}} = \begin{pmatrix} \hat{\varepsilon}_1 \\ \vdots \\ \hat{\varepsilon}_n \end{pmatrix}$, alors

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (I_n - P_{[X]})\boldsymbol{Y}$$

- Résidus aussi gaussien :

$$\mathbb{E}(\hat{\boldsymbol{\varepsilon}}) = \mathbb{E}(\boldsymbol{Y} - X\hat{\beta}) = \mathbb{E}(\boldsymbol{Y}) - X\mathbb{E}(\hat{\beta}) = X\beta - X\beta = 0$$

Variance Covariance des résidus

- $\Sigma_{\hat{\varepsilon}} = \sigma^2(I_n - P_{[X]})$
- Contient σ^2 du bruit
- En général non diagonale : $P_{[X]}$ n'a aucune raison d'être diagonale.
- Termes diagnoaux non tous égaux
- En conclusion : **les résidus $\hat{\varepsilon}_i$ sont gaussiens centrés, comme l'erreur ε_i. Cependant ils n'ont pas tous la même variance, et ne sont pas indépendants en général**.
- **Propriété importante**

$$\hat{\varepsilon} \text{ et } \hat{\boldsymbol{Y}} \text{ sont indépendants}$$

Résidus sous R

On trouve à nouveau ces résidus par la commande `resid(reg)`.

Exercice

1. Charger le fichier "jouet2.Rdata". Faire la régression de y sur $x1, x2$.
2. Afficher le graphique de $\hat{y}$ en ordonnée contre $\hat{\varepsilon}$ en abscisse
3. Quelle conclusion tirer de ce graphique ?

Estimation de σ^2

Proposition (Estimateur de la variance du bruit)

L'estimateur

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$$

est un estimateur sans biais de σ^2. *On a plus précisément*

$$(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p)$$

où $\chi^2(n-p)$ *est une loi du Khi-deux à* $n-p$ *degrés de libertés.*

Conséquence : estimation des variances des estimateurs

- On avait calculé la loi de $\hat{\beta}_j$ pour $j \in \{1, \dots, p\}$
- Par exemple

$$\sigma^2_{\hat{\beta}_j} = \sigma^2 \left[(X^T X)^{-1} \right]_{jj}$$

- De même,

$$\sigma_{\hat{\beta}_j, \hat{\beta}_k} = \sigma^2 \left[(X^T X)^{-1} \right]_{jk}$$

- Pour calculer ces quantités, il faudrait connaître σ^2.
- Or σ^2 est inconnue en général, on la remplace donc par un estimateur $\hat{\sigma}^2$.
- Alors

$$\hat{\sigma}^2_{\hat{\beta}_j} = \hat{\sigma}^2 \left[(X^T X)^{-1} \right]_{jj}$$

Variance sous R

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

```
## [1] 276.6734
```

Les écarts-types de chaque coefficient $\hat{\beta}_j$ sont donnés dans la colonne `Std.Error` du tableau `Coefficients` de la sortie de `summary`.

Prévision

- Prévoir les valeurs de la concentration en ozone `maxO3` pour une nouvelle journée en mesurant uniquement la température `T12`, la nébulosité `Ne12` et la quantité de vent `Vx12`.

Prévision

- Prévoir les valeurs de la concentration en ozone `maxO3` pour une nouvelle journée en mesurant uniquement la température `T12`, la nébulosité `Ne12` et la quantité de vent `Vx12`.
- On a une nouvelle donnée x_{n+1} qui correspond à un $(n+1)$-ème individu, et on ne connait pas le y_{n+1} correspondant.

Prévision

- Prévoir les valeurs de la concentration en ozone `maxO3` pour une nouvelle journée en mesurant uniquement la température `T12`, la nébulosité `Ne12` et la quantité de vent `Vx12`.
- On a une nouvelle donnée x_{n+1} qui correspond à un $(n+1)$-ème individu, et on ne connait pas le y_{n+1} correspondant.
- On note $\hat{y}^{p}_{n+1}$ la valeur prédite, y_{n+1} étant la vraie valeur, inconnue (non mesurée). On n'observe que x_{n+1}.

Prévision

- Prévoir les valeurs de la concentration en ozone `maxO3` pour une nouvelle journée en mesurant uniquement la température `T12`, la nébulosité `Ne12` et la quantité de vent `Vx12`.
- On a une nouvelle donnée x_{n+1} qui correspond à un $(n+1)$-ème individu, et on ne connait pas le y_{n+1} correspondant.
- On note $\hat{y}^p_{n+1}$ la valeur prédite, y_{n+1} étant la vraie valeur, inconnue (non mesurée). On n'observe que x_{n+1}.
- Le modèle indique

$$y_{n+1} = \beta^T x_{n+1} + \varepsilon_{n+1}$$

avec $\varepsilon_{n+1} \sim \mathcal{N}(0, \sigma^2)$ et ε_{n+1} indépendant des $(\varepsilon_i)_{1 \leq i \leq n}$.

Prévision

- Prévoir les valeurs de la concentration en ozone `maxO3` pour une nouvelle journée en mesurant uniquement la température `T12`, la nébulosité `Ne12` et la quantité de vent `Vx12`.
- On a une nouvelle donnée x_{n+1} qui correspond à un $(n+1)$-ème individu, et on ne connait pas le y_{n+1} correspondant.
- On note $\hat{y}^p_{n+1}$ la valeur prédite, y_{n+1} étant la vraie valeur, inconnue (non mesurée). On n'observe que x_{n+1}.
- Le modèle indique
$$y_{n+1} = \beta^T x_{n+1} + \varepsilon_{n+1}$$
avec $\varepsilon_{n+1} \sim \mathcal{N}(0, \sigma^2)$ et ε_{n+1} indépendant des $(\varepsilon_i)_{1 \leq i \leq n}$.
- Prédiction grâce au modèle et à l'estimateur $\hat{\beta}$:
$$\hat{y}^p_{n+1} = \hat{\beta}^T x_{n+1}$$

Prévision

- Prévoir les valeurs de la concentration en ozone `maxO3` pour une nouvelle journée en mesurant uniquement la température `T12`, la nébulosité `Ne12` et la quantité de vent `Vx12`.
- On a une nouvelle donnée x_{n+1} qui correspond à un $(n+1)$-ème individu, et on ne connait pas le y_{n+1} correspondant.
- On note $\hat{y}^p_{n+1}$ la valeur prédite, y_{n+1} étant la vraie valeur, inconnue (non mesurée). On n'observe que x_{n+1}.
- Le modèle indique

$$y_{n+1} = \beta^T x_{n+1} + \varepsilon_{n+1}$$

avec $\varepsilon_{n+1} \sim \mathcal{N}(0, \sigma^2)$ et ε_{n+1} indépendant des $(\varepsilon_i)_{1 \leq i \leq n}$.

- Prédiction grâce au modèle et à l'estimateur $\hat{\beta}$:

$$\hat{y}^p_{n+1} = \hat{\beta}^T x_{n+1}$$

- $\hat{\beta}$ calculé avec $(x_1, y_1), \dots, (x_n, y_n)$ mais pas (x_{n+1}, y_{n+1}).

Erreur de prévision

Théorème (Erreur de prévision)

L'erreur de prévision

$$\hat{\varepsilon}^p_{n+1} = y_{n+1} - \hat{y}^p_{n+1}$$

vérifie

$$\mathbb{E}(\hat{\varepsilon}^p_{n+1}) = 0 \quad et \quad \mathrm{Var}(\hat{\varepsilon}^p_{n+1}) = \sigma^2 \left(1 + x^T_{n+1}(X^T X)^{-1} x_{n+1}\right)$$

- Cette erreur est inconnue car y_{n+1} est inconnue

Erreur de prévision

Théorème (Erreur de prévision)

L'erreur de prévision

$$\hat{\varepsilon}^p_{n+1} = y_{n+1} - \hat{y}^p_{n+1}$$

vérifie

$$\mathbb{E}(\hat{\varepsilon}^p_{n+1}) = 0 \quad \textit{et} \quad \mathrm{Var}(\hat{\varepsilon}^p_{n+1}) = \sigma^2 \left(1 + x^T_{n+1}(X^T X)^{-1} x_{n+1}\right)$$

- Cette erreur est inconnue car y_{n+1} est inconnue
- $x^T_{n+1}(X^T X)^{-1} x_{n+1}$ lié à la distance (non euclidienne) entre x_{n+1} et $\bar{x}$.

Erreur de prévision

Théorème (Erreur de prévision)

L'erreur de prévision

$$\hat{\varepsilon}^p_{n+1} = y_{n+1} - \hat{y}^p_{n+1}$$

vérifie

$$\mathbb{E}(\hat{\varepsilon}^p_{n+1}) = 0 \quad et \quad \mathrm{Var}(\hat{\varepsilon}^p_{n+1}) = \sigma^2\left(1 + x^T_{n+1}(X^TX)^{-1}x_{n+1}\right)$$

- Cette erreur est inconnue car y_{n+1} est inconnue
- $x^T_{n+1}(X^TX)^{-1}x_{n+1}$ lié à la distance (non euclidienne) entre x_{n+1} et $\bar{x}$.
- Plus x_{n+1} loin de $\bar{x}$, plus la précision de la prévision diminue.

Exercice

- Jeux de données `ozone`
- On a mesuré la valeur des trois variables `T12`, `Ne9` et `Vx9` pour une nouvelle journée : $(20, 6, -3)$.
- Donner le taux d'ozone prévu par le modèle linéaire.
- Indication : faire comme pour la régression linéaire simple !

Intervalles de confiance sur les paramètres

Théorème

Un IC de niveau $1-\alpha$ *pour le coefficient* β_j *est donné par*

$$\left[\hat{\beta}_j \pm q^{1-\frac{\alpha}{2}}_{\mathcal{T}(n-p)} \hat{\sigma} \sqrt{[(X^TX)^{-1}]_{jj}}\right]$$

Un IC de niveau $1-\alpha$ *pour* σ^2 *est donné par*

$$\left[\frac{(n-p)\hat{\sigma}^2}{q^{1-\frac{\alpha}{2}}_{\chi^2(n-p)}}, \frac{(n-p)\hat{\sigma}^2}{q^{\frac{\alpha}{2}}_{\chi^2(n-p)}}\right].$$

Expérience numérique avec R

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

```
##                    2.5 %    97.5 %
## (Intercept) -25.4886483 33.280203
## Ne12         -3.6368523  0.399082
## Vx12          0.3264694  2.931560
## T12           3.4819098  5.544563
```

Avec probabilité 95%, le coefficient de `Ne12`, est dans $[-3.637, 0.399]$.

Intervalle de prédiction pour y_{n+1}

Theorem (Intervalle de prédiction pour une prévision)

Un IC de niveau $1-\alpha$ pour y_{n+1} est donné par

$$\left[x_{n+1}^T\hat{\beta} \pm q_{\mathcal{T}(n-p)}^{1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{x_{n+1}^T(X^TX)^{-1}x_{n+1}+1}\right]$$

- Nouvel individu $x_{n+1} \in \mathbb{R}^p$, valeur de y_{n+1} inconnue.

Expérience numérique avec R

- IC sur la prévision de la concentration d'ozone pour deux nouvelles journées correspondant à x_{n+1} et x_{n+2}

Expérience numérique avec R

- IC sur la prévision de la concentration d'ozone pour deux nouvelles journées correspondant à x_{n+1} et x_{n+2}
- Sur la première journée, les valeurs respectives de la nébulosité, la quantité de vent et la température sont de 2, -1 et 20.

Expérience numérique avec R

- IC sur la prévision de la concentration d'ozone pour deux nouvelles journées correspondant à x_{n+1} et x_{n+2}
- Sur la première journée, les valeurs respectives de la nébulosité, la quantité de vent et la température sont de 2, -1 et 20.
- Sur la deuxième journée, elles sont de 3, 0 et 23.

Expérience numérique avec R

- IC sur la prévision de la concentration d'ozone pour deux nouvelles journées correspondant à x_{n+1} et x_{n+2}
- Sur la première journée, les valeurs respectives de la nébulosité, la quantité de vent et la température sont de 2, -1 et 20.
- Sur la deuxième journée, elles sont de 3, 0 et 23.
- Code R

```
Nenew=c(2,3)
Vxnew=c(-1,0)
Tnew=c(46,35)
xnew=data.frame(Ne12=Nenew,Vx12=Vxnew,T12=Tnew)
predict(reg,new=xnew,interval="pred",level=0.95)
```

```
##          fit      lwr      upr
## 1 206.6379 166.8820 246.3938
## 2 157.0024 121.8401 192.1647
```

Tests sur la pertinence d'un coefficient

- On se pose ici la question de l'utilité d'une variable explicative. Par exemple : le vent influence-t-il la concentration d'ozone ?

Tests sur la pertinence d'un coefficient

- On se pose ici la question de l'utilité d'une variable explicative. Par exemple : le vent influence-t-il la concentration d'ozone ?
- Dire que la variable est inutile revient à dire que son coefficient est nul. Le problème de test est donc le suivant

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

Tests sur la pertinence d'un coefficient

- On se pose ici la question de l'utilité d'une variable explicative. Par exemple : le vent influence-t-il la concentration d'ozone ?
- Dire que la variable est inutile revient à dire que son coefficient est nul. Le problème de test est donc le suivant

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

- On sait que

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})$$

Ce qui donne, pour la composante $\hat{\beta}_j$ de $\hat{\beta}$,

$$\hat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2 [(X^T X)^{-1}]_{jj}\right)$$

Test sur la pertinence d'un coefficient

- Donc sous $\mathcal{H}_0$ ($\beta_j = 0$)

$$\hat{\beta}_j \sim \mathcal{N}\left(0, \sigma^2[(X^TX)^{-1}]_{jj}\right) \quad \text{d'où} \quad \frac{\hat{\beta}_j}{\sqrt{\sigma^2[(X^TX)^{-1}]_{jj}}} \sim \mathcal{N}(0,1)$$

Test sur la pertinence d'un coefficient

- Donc sous $\mathcal{H}_0$ $(\beta_j = 0)$

$$\hat{\beta}_j \sim \mathcal{N}\left(0, \sigma^2[(X^TX)^{-1}]_{jj}\right) \quad \text{d'où} \quad \frac{\hat{\beta}_j}{\sqrt{\sigma^2[(X^TX)^{-1}]_{jj}}} \sim \mathcal{N}(0,1)$$

- IC bilatère symétrique de niveau $1-\alpha$: $[-q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)}, q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)}]$

Test sur la pertinence d'un coefficient

- Donc sous $\mathcal{H}_0$ ($\beta_j = 0$)

$$\hat{\beta}_j \sim \mathcal{N}\left(0, \sigma^2[(X^T X)^{-1}]_{jj}\right) \quad \text{d'où} \quad \frac{\hat{\beta}_j}{\sqrt{\sigma^2[(X^T X)^{-1}]_{jj}}} \sim \mathcal{N}(0,1)$$

- IC bilatère symétrique de niveau $1-\alpha$: $[-q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)}, q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)}]$
- Si on est bien sous $\mathcal{H}_0$ on a donc, si α est petit, de grandes chances d'avoir $\hat{\beta}_j$ dans l'IC.

Test sur la pertinence d'un coefficient

- Donc sous $\mathcal{H}_0$ ($\beta_j = 0$)

$$\hat{\beta}_j \sim \mathcal{N}\left(0, \sigma^2[(X^TX)^{-1}]_{jj}\right) \quad \text{d'où} \quad \frac{\hat{\beta}_j}{\sqrt{\sigma^2[(X^TX)^{-1}]_{jj}}} \sim \mathcal{N}(0,1)$$

- IC bilatère symétrique de niveau $1-\alpha$: $[-q^{\mathcal{N}(0,1)}_{1-\frac{\alpha}{2}}, q^{\mathcal{N}(0,1)}_{1-\frac{\alpha}{2}}]$
- Si on est bien sous $\mathcal{H}_0$ on a donc, si α est petit, de grandes chances d'avoir $\hat{\beta}_j$ dans l'IC.
- Si $\mathcal{H}_0$ est fausse, alors $\beta_j \neq 0$ et donc

$$\hat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2[(X^TX)^{-1}]_{jj}\right) \quad \text{d'où} \quad \frac{\hat{\beta}_j}{\sqrt{\sigma^2[(X^TX)^{-1}]_{jj}}} \sim \mathcal{N}\left(\frac{\beta_j}{\dots}, 1\right).$$

Test sur la pertinence d'un coefficient

- Donc sous $\mathcal{H}_0$ ($\beta_j = 0$)

$$\hat{\beta}_j \sim \mathcal{N}\left(0, \sigma^2[(X^T X)^{-1}]_{jj}\right) \quad \text{d'où} \quad \frac{\hat{\beta}_j}{\sqrt{\sigma^2[(X^T X)^{-1}]_{jj}}} \sim \mathcal{N}(0,1)$$

- IC bilatère symétrique de niveau $1-\alpha$: $[-q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)}, q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)}]$
- Si on est bien sous $\mathcal{H}_0$ on a donc, si α est petit, de grandes chances d'avoir $\hat{\beta}_j$ dans l'IC.
- Si $\mathcal{H}_0$ est fausse, alors $\beta_j \neq 0$ et donc

$$\hat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2[(X^T X)^{-1}]_{jj}\right) \quad \text{d'où} \quad \frac{\hat{\beta}_j}{\sqrt{\sigma^2[(X^T X)^{-1}]_{jj}}} \sim \mathcal{N}\left(\frac{\beta_j}{\dots}, 1\right).$$

- Statistique de test $S = \hat{\beta}_j / \sqrt{\sigma^2[(X^T X)^{-1}]_{jj}}$

Test sur la pertinence d'un coefficient

- Donc sous $\mathcal{H}_0$ ($\beta_j = 0$)

$$\hat{\beta}_j \sim \mathcal{N}\left(0, \sigma^2[(X^TX)^{-1}]_{jj}\right) \quad \text{d'où} \quad \frac{\hat{\beta}_j}{\sqrt{\sigma^2[(X^TX)^{-1}]_{jj}}} \sim \mathcal{N}(0,1)$$

- IC bilatère symétrique de niveau $1-\alpha$: $[-q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)}, q_{1-\frac{\alpha}{2}}^{\mathcal{N}(0,1)}]$
- Si on est bien sous $\mathcal{H}_0$ on a donc, si α est petit, de grandes chances d'avoir $\hat{\beta}_j$ dans l'IC.
- Si $\mathcal{H}_0$ est fausse, alors $\beta_j \neq 0$ et donc

$$\hat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2[(X^TX)^{-1}]_{jj}\right) \quad \text{d'où} \quad \frac{\hat{\beta}_j}{\sqrt{\sigma^2[(X^TX)^{-1}]_{jj}}} \sim \mathcal{N}\left(\frac{\beta_j}{\dots}, 1\right).$$

- Statistique de test $S = \hat{\beta}_j / \sqrt{\sigma^2[(X^TX)^{-1}]_{jj}}$
- Remplacer σ^2 inconnue par $\hat{\sigma}^2$, et donc $\mathcal{N}(0,1)$ par $\mathcal{T}(n-p)$.

Test sur la pertinence d'un coefficient

Theorem (Test sur la pertinence d'un coefficient)

On s'intéresse au problème de test

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

On rejette $\mathcal{H}_0$, si

$$\frac{|\hat{\beta}_j|}{\hat{\sigma}\sqrt{[(X^T X)^{-1}]_{jj}}} > q^{1-\frac{\alpha}{2}}_{\mathcal{T}(n-p)}$$

- Découle du fait que si $\beta_j = 0$,

$$\frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{[(X^T X)^{-1}]_{jj}}} \sim \mathcal{T}(n-p)$$

Remarques

- $\mathcal{H}_0$ ici est «la variable explicative n'est pas pertinente»

Remarques

- $\mathcal{H}_0$ ici est «la variable explicative n'est pas pertinente»
- Une variable explicative est considérée comme pertinente au niveau α si la *p*-valeur associée est $< \alpha$.

Remarques

- $\mathcal{H}_0$ ici est «la variable explicative n'est pas pertinente»
- Une variable explicative est considérée comme pertinente au niveau α si la *p*-valeur associée est $< \alpha$.
- Plus la *p*-valeur est petite, plus on aura confiance en la pertinence de la variable.

Remarques

- $\mathcal{H}_0$ ici est «la variable explicative n'est pas pertinente»
- Une variable explicative est considérée comme pertinente au niveau α si la p-valeur associée est $< \alpha$.
- Plus la p-valeur est petite, plus on aura confiance en la pertinence de la variable.
- Attention aux interprétations abusives des p-valeurs

Expérience numérique avec R

- Test fait dans le tableau `Coefficient` de la sortie de `summary`.

Que se passe-t-il en cas de colinéarité des variables

Expérience numérique avec R

- Test fait dans le tableau `Coefficient` de la sortie de `summary`.
- Code :

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.8958    14.8243   0.263   0.7932
## Ne12          -1.6189     1.0181  -1.590   0.1147
## Vx12           1.6290     0.6571   2.479   0.0147*
## T12            4.5132     0.5203   8.674 4.71e-14 ***
## ---
```

Que se passe-t-il en cas de colinéarité des variables

Expérience numérique avec R

- Test fait dans le tableau `Coefficient` de la sortie de `summary`.
- Code :

```
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.8958    14.8243   0.263   0.7932
## Ne12          -1.6189     1.0181  -1.590   0.1147
## Vx12           1.6290     0.6571   2.479   0.0147*
## T12            4.5132     0.5203   8.674 4.71e-14 ***
## ---
```

- Valeur de statistique de test dans colonne `t.value` ("t" pour Student). La p-valeur dans la dernière colonne `Pr(>|t|)`.

Que se passe-t-il en cas de colinéarité des variables

Expérience numérique avec R

- Test fait dans le tableau `Coefficient` de la sortie de `summary`.
- Code :

```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.8958    14.8243   0.263   0.7932
## Ne12          -1.6189     1.0181  -1.590   0.1147
## Vx12           1.6290     0.6571   2.479   0.0147*
## T12            4.5132     0.5203   8.674 4.71e-14 ***
## ---
```

- Valeur de statistique de test dans colonne `t.value` ("t" pour Student). La p-valeur dans la dernière colonne `Pr(>|t|)`.
- Plus la p-valeur est petite, plus on a envie de rejeter $\mathcal{H}_0$.

Que se passe-t-il en cas de colinéarité des variables

Expérience numérique avec R

- Test fait dans le tableau `Coefficient` de la sortie de `summary`.
- Code :

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.8958    14.8243   0.263   0.7932
## Ne12         -1.6189     1.0181  -1.590   0.1147
## Vx12          1.6290     0.6571   2.479   0.0147*
## T12           4.5132     0.5203   8.674 4.71e-14 ***
## ---
```

- Valeur de statistique de test dans colonne `t.value` ("t" pour Student). La p-valeur dans la dernière colonne `Pr(>|t|)`.
- Plus la p-valeur est petite, plus on a envie de rejeter $\mathcal{H}_0$.
- Les variables `Vx12` et `T12` sont significatives au niveau 5% : vent et température influencent linéairement la concentration d'ozone.

Que se passe-t-il en cas de colinéarité des variables

Expérience numérique avec R

- Test fait dans le tableau `Coefficient` de la sortie de `summary`.
- Code :

```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.8958    14.8243   0.263   0.7932
## Ne12         -1.6189     1.0181  -1.590   0.1147
## Vx12          1.6290     0.6571   2.479   0.0147*
## T12           4.5132     0.5203   8.674 4.71e-14 ***
## ---
```

- Valeur de statistique de test dans colonne `t.value` ("t" pour Student). La p-valeur dans la dernière colonne `Pr(>|t|)`.
- Plus la p-valeur est petite, plus on a envie de rejeter $\mathcal{H}_0$.
- Les variables `Vx12` et `T12` sont significatives au niveau 5% : vent et température influencent linéairement la concentration d'ozone.
- La variable `Ne12` est non significative au niveau 5%.

Que se passe-t-il en cas de colinéarité des variables

Tester la pertinence d'un groupe de variables explicatives

- On peut aussi tester la pertinence d'un groupe de variables explicatives, autrement dit tester la nullité simultanée de plusieurs coefficients.

Tester la pertinence d'un groupe de variables explicatives

- On peut aussi tester la pertinence d'un groupe de variables explicatives, autrement dit tester la nullité simultanée de plusieurs coefficients.
- $\mathcal{H}_0 : \forall j \in \{p-q+1, \ldots, p\} : \beta_j = 0$
  contre
  $\mathcal{H}_1 : \exists j \in \{p-q+1, \ldots, p\} : \ \beta_j \neq 0$

Tester la pertinence d'un groupe de variables explicatives

- On peut aussi tester la pertinence d'un groupe de variables explicatives, autrement dit tester la nullité simultanée de plusieurs coefficients.
- $\mathcal{H}_0 : \forall j \in \{p-q+1, \ldots, p\} : \beta_j = 0$
  contre
  $\mathcal{H}_1 : \exists j \in \{p-q+1, \ldots, p\} : \ \beta_j \neq 0$
- Sous $\mathcal{H}_0$ le modèle est en réalité

$$y_i = \sum_{j=1}^{p-q} \beta_j x^j + \varepsilon_i, \quad i = 1, \ldots, n$$

Tester la pertinence d'un groupe de variables explicatives

- On peut aussi tester la pertinence d'un groupe de variables explicatives, autrement dit tester la nullité simultanée de plusieurs coefficients.
- $\mathcal{H}_0 : \forall j \in \{p-q+1, \ldots, p\} : \beta_j = 0$
  contre
  $\mathcal{H}_1 : \exists j \in \{p-q+1, \ldots, p\} : \ \beta_j \neq 0$
- Sous $\mathcal{H}_0$ le modèle est en réalité

$$y_i = \sum_{j=1}^{p-q} \beta_j x^j + \varepsilon_i, \quad i = 1, \ldots, n$$

- Autrement dit, avec $\tilde{\beta} = (\beta_1, \ldots, \beta_{p-q})^\top$,

$$y = X_0 \tilde{\beta} + \varepsilon$$

avec $X_0 \in \mathcal{M}_{n\times(p-q)} = p - q$ premières colonnes de X.

Test groupé

- $\mathcal{M}_1$ = modèle complet avec toutes les variables explicatives
  $\mathcal{M}_0$ = sous-modèle obtenu en enlevant les q dernières variables

Test groupé

- $\mathcal{M}_1$ = modèle complet avec toutes les variables explicatives
  $\mathcal{M}_0$ = sous-modèle obtenu en enlevant les q dernières variables
- On soupçonne donc que le vrai modèle est le modèle $\mathcal{M}_0$.

Test groupé

- $\mathcal{M}_1$ = modèle complet avec toutes les variables explicatives
  $\mathcal{M}_0$ = sous-modèle obtenu en enlevant les q dernières variables
- On soupçonne donc que le vrai modèle est le modèle $\mathcal{M}_0$.
- Le test précédent revient à comparer deux modèles, $\mathcal{M}_0$ et $\mathcal{M}_1$.

Test groupé

- $\mathcal{M}_1$ = modèle complet avec toutes les variables explicatives
  $\mathcal{M}_0$ = sous-modèle obtenu en enlevant les q dernières variables
- On soupçonne donc que le vrai modèle est le modèle $\mathcal{M}_0$.
- Le test précédent revient à comparer deux modèles, $\mathcal{M}_0$ et $\mathcal{M}_1$.
- Si $\mathcal{M}_0$ est le bon modèle alors on collectera moins de données en vue de la prévision et on estimera mieux les paramètres.

SCR

$$
\begin{aligned}
\text{SCR} &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \|Y - P_{[X]}Y\|^2 \\
\text{SCR}_0 &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i^{(0)})^2 = \|Y - P_{[X_0]}Y\|^2
\end{aligned}
$$

- La matrice X_0 étant une sous partie des colonnes de X, on a $[X_0] \subset [X]$.

$\Rightarrow$ $\text{SCR} \leq \text{SCR}_0$,

- On s'ajustera mieux aux données avec le modèle plus riche.
- Mais ce gain d'ajustement "vaut-il le coup" ? Un modèle plus "riche" implique plus de paramètres à estimer donc plus d'incertitude.
- Par conséquent, on va construire un test sur la différence $\text{SCR} - \text{SCR}_0$ et rejeter $\mathcal{H}_0$ si cette différence est significativement plus grande que 0.

Test de Fisher

La statistique du test de Fisher repose sur cette différence et prend en compte la taille des modèles en compétition :

$$F = \frac{(\mathrm{SCR}_0 - \mathrm{SCR}\ )/(n-p-(n-(p-q))}{\mathrm{SCR}\ /(n-p)} = \frac{(\mathrm{SCR}_0 - \mathrm{SCR}\ )/q}{\mathrm{SCR}\ /(n-p)}$$

.

Sous $\mathcal{H}_0$, $F \sim \mathcal{F}_{q,n-p}$.

Théorème

Pour le problème de test $\mathcal{H}_0 : \beta_{p-q+1} = \beta_{p-q+2} = \ldots \beta_{p-1} = \beta_p = 0$ *contre* $\mathcal{H}_1 : \exists j \in \{p-q+1, \ldots, p\} : \ \beta_j \neq 0$, *on rejette* $\mathcal{H}_0$ *si*

$$F > q^{1-\alpha}_{\mathcal{F}_{q,n-p}}$$

où $q^{1-\alpha}_{\mathcal{F}_{q,n-p}}$ *est le quantile d'une loi de Fisher à* $(q, n-p)$ *degrés de liberté.*

Application sur les données Ozone

```
reg1 <- lm(maxO3 ~ T12)
anova(reg1,reg)
```

```
## Analysis of Variance Table
##
## Model 1: maxO3 ~ T12
## Model 2: maxO3 ~ Ne12 + Vx12 + T12
##   Res.Df   RSS Df Sum of Sq    F   Pr(>F)
## 1    110 33948
## 2    108 29881  2    4067.1 7.35 0.001017 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Exercice

Reprendre le fichier Eucalyptus. On souhaite expliquer la hauteur de l'arbre en fonction de la circonférence et de la racine carrée de la circonférence.

```
modele0=lm(ht~circ,data=euca)
modele1=lm(ht~circ+I(sqrt(circ)),data=euca)
anova(modele0,modele1)
```

1. Comparer les modèles en utilisant `anova`.
2. Remarquez que dans ce cas particulier, on aurait pu faire le test de student de pertinence de la variable explicative `sqrt(circ)` dans le modèle `modele1`. Obtient-on le même résultat ?

Test du modèle ou test de Fisher global

Test classiquement fait

- **On veut tester la pertinence de l'ensemble des variables**

Test du modèle ou test de Fisher global

Test classiquement fait

- **On veut tester la pertinence de l'ensemble des variables**
- Modèle $\mathcal{M}_1$ fait intervenir l'ensemble des variables

Test du modèle ou test de Fisher global

Test classiquement fait

- **On veut tester la pertinence de l'ensemble des variables**
- Modèle $\mathcal{M}_1$ fait intervenir l'ensemble des variables
- Modèle $\mathcal{M}_0$ ne fait intervenir que l'intercept.

Test du modèle ou test de Fisher global

Test classiquement fait

- **On veut tester la pertinence de l'ensemble des variables**
- Modèle $\mathcal{M}_1$ fait intervenir l'ensemble des variables
- Modèle $\mathcal{M}_0$ ne fait intervenir que l'intercept.
- Test fait systématiquement par R, donné dans le `summary`.

Test du modèle ou test de Fisher global

Test classiquement fait

- **On veut tester la pertinence de l'ensemble des variables**
- Modèle $\mathcal{M}_1$ fait intervenir l'ensemble des variables
- Modèle $\mathcal{M}_0$ ne fait intervenir que l'intercept.
- Test fait systématiquement par R, donné dans le `summary`.
- Apparait à la dernière ligne et la statistique de test, nommé `F-statistic`. Il correspond donc au test suivant

$$\mathcal{H}_0 : \beta_2 = \cdots = \beta_p = 0 \quad \text{contre} \quad \mathcal{H}_1 : \exists j \in \{2, \ldots, p\} : \ \beta_j \neq 0,$$

Test du modèle ou test de Fisher global

Test classiquement fait

- **On veut tester la pertinence de l'ensemble des variables**
- Modèle $\mathcal{M}_1$ fait intervenir l'ensemble des variables
- Modèle $\mathcal{M}_0$ ne fait intervenir que l'intercept.
- Test fait systématiquement par R, donné dans le `summary`.
- Apparait à la dernière ligne et la statistique de test, nommé `F-statistic`. Il correspond donc au test suivant

$$\mathcal{H}_0 : \beta_2 = \cdots = \beta_p = 0 \quad \text{contre} \quad \mathcal{H}_1 : \exists j \in \{2, \ldots, p\} : \ \beta_j \neq 0,$$

- Test nullité de $q = p - 1$ variables (toutes sauf intercept $x^1 = \mathbf{1}$).

Test du modèle ou test de Fisher global

Test classiquement fait

- **On veut tester la pertinence de l'ensemble des variables**
- Modèle $\mathcal{M}_1$ fait intervenir l'ensemble des variables
- Modèle $\mathcal{M}_0$ ne fait intervenir que l'intercept.
- Test fait systématiquement par R, donné dans le `summary`.
- Apparait à la dernière ligne et la statistique de test, nommé `F-statistic`. Il correspond donc au test suivant

$$\mathcal{H}_0 : \beta_2 = \cdots = \beta_p = 0 \quad \text{contre} \quad \mathcal{H}_1 : \exists j \in \{2, \ldots, p\} : \ \beta_j \neq 0,$$

- Test nullité de $q = p - 1$ variables (toutes sauf intercept $x^1 = \mathbf{1}$).
- Statistique de test

$$F \overset{\mathcal{H}_0}{\sim} \mathcal{F}_{p-1,n-p}.$$

Expérience numérique avec R

Dans l'exemple sur les eucalyptus : la p-valeur est inférieure à `2.2e-16` et donc on rejette $\mathcal{H}_0$. Autrement dit, au moins l'une des variables explicatives est pertinente.

Résumé sur les tests de Student et Fisher

- Petite p-valeur dans test de Student de «significativité» d'une variable explicative incite à penser «la variable est significative».

Résumé sur les tests de Student et Fisher

- Petite p-valeur dans test de Student de «significativité» d'une variable explicative incite à penser «la variable est significative».
- Petite p-valeur dans test global de Fisher incite à penser «les variables sont pertinentes dans leur ensemble»

Résumé sur les tests de Student et Fisher

- Petite p-valeur dans test de Student de «significativité» d'une variable explicative incite à penser «la variable est significative».
- Petite p-valeur dans test global de Fisher incite à penser «les variables sont pertinentes dans leur ensemble»
- Petite p-valeur pour `anova` incite à penser «le plus gros modèle est le meilleur» (celui avec plus de variables)

Somme des carrés

- Sous le modèle défini par $X^{(0)} = \mathbf{1}_n$ (modèle sans covariable), $\hat{\beta}_0 = \bar{\mathbf{y}}$. Donc $\hat{y}_i = \bar{\mathbf{y}}$.
- $SCR_0 = \sum_{i=1}^n (y_i - \bar{\mathbf{y}})^2 = SCT$: c'est la somme des carrés totale, autrement dit la variabilité des données.
- Par Pythagore, on a :

$$\text{SCT} = \underbrace{\sum_i^n (\hat{y}_i - \bar{\mathbf{y}})^2}_{\text{SCM}} + \underbrace{\sum_i^n (y_i - \hat{y}_i)^2}_{\text{SCR}}$$

- SCM est la somme des carrés du modèle : la variabilité expliquée par le modèle.
- SCR est la variabilité non-expliquée par le modèle.
- Plus SCM est grande par rapport à SCT , plus le modèle explique la variabilité des observations.

Coefficient de détermination R^2

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

$$R^2 \in [0, 1]$$

Indicateur de l'ajustement du modèle aux données.
Pourrait être utilisé pour comparer les performances entre deux modèles.

Sous R

$$0 \leq R^2 \leq 1$$

Si R^2 n'est pas assez proche de 1 alors cela signifie que le modèle n'approche pas bien y : soit il manque une variable explicative, qu'il faudrait donc introduire dans le modèle, soit l'une (ou plusieurs) des variables explicatives n'intervient pas de manière linéaire.

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

```
## [1] 0.6611843
```

Limites du R^2

- R^2 ne peut être utilisé pour sélectionner les covariables pertinentes

Limites du R^2

- R^2 ne peut être utilisé pour sélectionner les covariables pertinentes
- En effet $q \mapsto R_q^2$ est une fonction croissante de q.

Limites du R^2

- R^2 ne peut être utilisé pour sélectionner les covariables pertinentes
- En effet $q \mapsto R_q^2$ est une fonction croissante de q.
- Soit X_q la matrice composée des q premières colonnes de X.

$$\begin{aligned} \mathrm{SCR}_{q+1} &= \min_{\beta_{q+\mathbf{1}} \in \mathbb{R}^{q+\mathbf{1}}} \|\boldsymbol{y} - X_{q+1}\beta_{q+1}\|^2 \\ &\leq \min_{\beta_q \in \mathbb{R}^q} \left\| \boldsymbol{y} - X_{q+1} \begin{pmatrix} \beta_q \\ 0 \end{pmatrix} \right\|^2 = \min_{\beta_q \in \mathbb{R}^q} \|\boldsymbol{y} - X_q\beta_q\|^2 = \mathrm{SCR}_q \end{aligned}$$

Limites du R^2

- R^2 ne peut être utilisé pour sélectionner les covariables pertinentes
- En effet $q \mapsto R_q^2$ est une fonction croissante de q.
- Soit X_q la matrice composée des q premières colonnes de X.

$$\begin{aligned}
\mathrm{SCR}_{q+1} &= \min_{\beta_{q+1} \in \mathbb{R}^{q+1}} \|\boldsymbol{y} - X_{q+1}\beta_{q+1}\|^2 \\
&\leq \min_{\beta_q \in \mathbb{R}^q} \left\| \boldsymbol{y} - X_{q+1} \begin{pmatrix} \beta_q \\ 0 \end{pmatrix} \right\|^2 = \min_{\beta_q \in \mathbb{R}^q} \|\boldsymbol{y} - X_q\beta_q\|^2 = \mathrm{SCR}_q
\end{aligned}$$

- R^2 augmente quelque soit la variable incluse. Dans le jeu de données ozone, si j'ajoute la covariable “nombre de naissances“ , j'ajusterai mieux mes observations alors que la covariable n'a aucun caractère explicatif du phénomène biologique.

Limites du R^2

- R^2 ne peut être utilisé pour sélectionner les covariables pertinentes
- En effet $q \mapsto R^2_q$ est une fonction croissante de q.
- Soit X_q la matrice composée des q premières colonnes de X.

$$
\begin{aligned}
\mathrm{SCR}_{q+1} &= \min_{\beta_{q+\mathbf{1}} \in \mathbb{R}^{q+\mathbf{1}}} \|\boldsymbol{y} - X_{q+1}\beta_{q+1}\|^2 \\
&\leq \min_{\beta_q \in \mathbb{R}^q} \left\| \boldsymbol{y} - X_{q+1} \begin{pmatrix} \beta_q \\ 0 \end{pmatrix} \right\|^2 = \min_{\beta_q \in \mathbb{R}^q} \|\boldsymbol{y} - X_q\beta_q\|^2 = \mathrm{SCR}_q
\end{aligned}
$$

- R^2 augmente quelque soit la variable incluse. Dans le jeu de données ozone, si j'ajoute la covariable “nombre de naissances“ , j'ajusterai mieux mes observations alors que la covariable n'a aucun caractère explicatif du phénomène biologique.
- R^2 ne peut pas être utilisé tel quel pour la comparaison de modèles de tailles différentes ou la sélection des variables pertinentes.

Limites du R^2

- R^2 ne peut être utilisé pour sélectionner les covariables pertinentes
- En effet $q \mapsto R_q^2$ est une fonction croissante de q.
- Soit X_q la matrice composée des q premières colonnes de X.

$$
\begin{aligned}
\mathrm{SCR}_{q+1} &= \min_{\beta_{q+\mathbf{1}} \in \mathbb{R}^{q+\mathbf{1}}} \|\boldsymbol{y} - X_{q+1}\beta_{q+1}\|^2 \\
&\leq \min_{\beta_q \in \mathbb{R}^q} \left\| \boldsymbol{y} - X_{q+1} \begin{pmatrix} \beta_q \\ 0 \end{pmatrix} \right\|^2 = \min_{\beta_q \in \mathbb{R}^q} \|\boldsymbol{y} - X_q\beta_q\|^2 = \mathrm{SCR}_q
\end{aligned}
$$

- R^2 augmente quelque soit la variable incluse. Dans le jeu de données ozone, si j'ajoute la covariable “nombre de naissances“ , j'ajusterai mieux mes observations alors que la covariable n'a aucun caractère explicatif du phénomène biologique.
- R^2 ne peut pas être utilisé tel quel pour la comparaison de modèles de tailles différentes ou la sélection des variables pertinentes.
- Cependant, le R^2 peut être intéressant pour comparer des modèles de même dimension.

R^2 ajusté

- R_a^2 est une modification du R^2 qui tient compte du nombre de variables.
- $$R_a^2 = 1 - \frac{n-1}{n-p}\frac{SCR}{SCT}.$$
- Autrement dit, R_a^2 est la valeur amoindrie du R^2 : on a d'autant plus abaissé la valeur du R^2 que le nombre p de variables explicatives du modèle est grand.

Remarque

Il existe bien d'autres critères de choix de modèles (non traités dans ce cours, cf cours ultérieur, exemples : BIC, AIC).

Le R^2 ajusté est également donné par `summary`. C'est le `"Adjusted R-squared"`.

Exercice

Revenons à l'exemple lié aux eucalyptus.

1. Entre le modèle avec comme variable explicative `circ`, plus l'intercept, et le modèle avec comme variable explicative `sqrt(circ)`, plus l'intercept, lequel faut-il choisir ?
2. Entre le modèle que nous avons trouvé à la question précédente et le modèle avec les deux variables explicatives `circ` et `sqrt(circ)`, plus l'intercept, lequel faut-il choisir ?

Feuille de route pratique

Pour l'analyse de données par un modèle linéaire

1. Charger les données, vérifier que les variables sont bien de la nature attendue (qualitative / quantitative)
2. Ecrire le modèle linéaire
3. *Valider le modèle par la lecture des graphes de résidus*. Si problème, tenter de prendre le $\log(y)$ ou le log de certaines covariables
4. Faire le test du modèle global : si non rejet, on arrête là, le modèle linéaire n'est pas adapté.
5. Sinon, on peut tester les paramètres, faire de la prédiction...

Exercices récapitulatifs

Vous trouverez des exercices récapitulatifs à la fin du chapitre 3 du poly.
Ces exercices sont du type de ceux posés à l'examen.

Annexes

Tests et problème de colinéarité des variables explicatives

- On souhaite à présent explorer le problème de colinéarité des variables explicatives lors des tests de nullité d'un coefficient.

Tests et problème de colinéarité des variables explicatives

- On souhaite à présent explorer le problème de colinéarité des variables explicatives lors des tests de nullité d'un coefficient.
- Simulons le modèle linéaire suivant

$$y = x^2 + x^3 + x^4 + \varepsilon$$

Tests et problème de colinéarité des variables explicatives

- On souhaite à présent explorer le problème de colinéarité des variables explicatives lors des tests de nullité d'un coefficient.
- Simulons le modèle linéaire suivant

$$y = x^2 + x^3 + x^4 + \varepsilon$$

- Supposons que nous ne connaissions pas le vrai modèle et que nous ayons à notre disposition les variables explicatives x^2, x^3, x^4 et x^5, plus l'intercept.

Tests et problème de colinéarité des variables explicatives

- On souhaite à présent explorer le problème de colinéarité des variables explicatives lors des tests de nullité d'un coefficient.
- Simulons le modèle linéaire suivant

$$y = x^2 + x^3 + x^4 + \varepsilon$$

- Supposons que nous ne connaissions pas le vrai modèle et que nous ayons à notre disposition les variables explicatives x^2, x^3, x^4 et x^5, plus l'intercept.
- Nous ne savons donc pas que x^5 n'influence pas réellement y. Supposons que x^5 est très corrélée aux autres variables

$$x^5 \approx 2x^2 + x^3$$

Tests et problème de colinéarité des variables explicatives

- On souhaite à présent explorer le problème de colinéarité des variables explicatives lors des tests de nullité d'un coefficient.
- Simulons le modèle linéaire suivant

$$y = x^2 + x^3 + x^4 + \varepsilon$$

- Supposons que nous ne connaissions pas le vrai modèle et que nous ayons à notre disposition les variables explicatives x^2, x^3, x^4 et x^5, plus l'intercept.
- Nous ne savons donc pas que x^5 n'influence pas réellement y. Supposons que x^5 est très corrélée aux autres variables

$$x^5 \approx 2x^2 + x^3$$

- Que va-t-il se passer ?

Expérience numérique avec R

```
x2 = rnorm(100)
x3 = rnorm(100)
x4 = rnorm(100)
c = c(1, 1, rep(0,98))
x5 = 2 * x2 + x3 + c
eps = rnorm(100)
y = x2 + x3 + x4 + eps
summary(lm(y ~ x2 + x3 + x4 + x5))
```

Expérience numérique avec R (suite)

```
Call:
lm(formula = y ~ x2 + x3 + x4 + x5)

Residuals:
     Min       1Q   Median       3Q      Max
-2.49243 -0.67679  0.05223  0.76233  2.25481

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.01772    0.11376  -0.156    0.877
x2           1.29706    1.62490   0.798    0.427
x3           1.28589    0.80412   1.599    0.113
x4           0.87609    0.10718   8.174 1.29e-12 ***
x5          -0.17162    0.80303  -0.214    0.831
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.104 on 95 degrees of freedom
Multiple R-squared:  0.7518,Adjusted R-squared:  0.7414
F-statistic: 71.95 on 4 and 95 DF,  p-value: < 2.2e-16
```

- Seule x^4 est déclarée pertinente
- Les variables x^3 et surtout x^2 ont des p-valeurs assez grandes : on les éliminerait donc à tort. Elles sont «cachées» par x^5

Expérience numérique avec R (suite et fin)

```
Call:
lm(formula = y ~ x2 + x3 + x4)

Residuals:
     Min       1Q   Median       3Q      Max
-2.49003 -0.67292  0.06052  0.76603  2.09584

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02175    0.11162  -0.195    0.846
x2           0.95071    0.11754   8.088 1.85e-12 ***
x3           1.11572    0.11183   9.977  < 2e-16 ***
x4           0.87928    0.10560   8.326 5.77e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.098 on 96 degrees of freedom
Multiple R-squared:  0.7517,Adjusted R-squared:  0.7439
F-statistic: 96.88 on 3 and 96 DF,  p-value: < 2.2e-16
```

- Regardons ce qui se passe maintenant quand on fait la régression de y sur les trois premières variables.

Expérience numérique avec R (suite et fin)

```
Call:
lm(formula = y ~ x2 + x3 + x4)

Residuals:
     Min       1Q   Median       3Q      Max
-2.49003 -0.67292  0.06052  0.76603  2.09584

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02175    0.11162  -0.195    0.846
x2           0.95071    0.11754   8.088 1.85e-12 ***
x3           1.11572    0.11183   9.977  < 2e-16 ***
x4           0.87928    0.10560   8.326 5.77e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.098 on 96 degrees of freedom
Multiple R-squared:  0.7517,Adjusted R-squared:  0.7439
F-statistic: 96.88 on 3 and 96 DF,  p-value: < 2.2e-16
```

- Regardons ce qui se passe maintenant quand on fait la régression de y sur les trois premières variables.
- Sans la présence intempestive de la variable x^5, les trois variables x^2,

Conclusion de l'expérience numérique avec R

- Si on utilise les tests de Student pour éliminer toutes les variables déclarées non pertinentes en même temps, on prend le risque d'en éliminer certaines qui sont pourtant pertinentes.

Retour aux tests

Conclusion de l'expérience numérique avec R

- Si on utilise les tests de Student pour éliminer toutes les variables déclarées non pertinentes en même temps, on prend le risque d'en éliminer certaines qui sont pourtant pertinentes.
- Si au contraire, on élimine une à une les variables, **en refaisant la régression à chaque fois**, on prend moins de risque : on enlève d'abord x^5 (dont la p-value de Student est la plus grande, à l'exception de l'intercept). Alors en refaisant une régression sans cette variable, x^2 et x^3 «redeviennent» pertinentes.

Retour aux tests

Conclusion de l'expérience numérique avec R

- Si on utilise les tests de Student pour éliminer toutes les variables déclarées non pertinentes en même temps, on prend le risque d'en éliminer certaines qui sont pourtant pertinentes.
- Si au contraire, on élimine une à une les variables, **en refaisant la régression à chaque fois**, on prend moins de risque : on enlève d'abord x^5 (dont la p-value de Student est la plus grande, à l'exception de l'intercept). Alors en refaisant une régression sans cette variable, x^2 et x^3 «redeviennent» pertinentes.
- Si on veut éliminer plusieurs variables en même temps, il vaut mieux faire le test de la section suivante (test de Fisher).

Retour aux tests