

The background of the entire image is a high-angle, nighttime aerial photograph of a dense urban area. The city lights create a grid-like pattern of yellow and white streaks against a dark blue sky. Buildings are visible as various shades of blue and grey.

USID17 STATISTIQUES 2

MASTER MEDAS

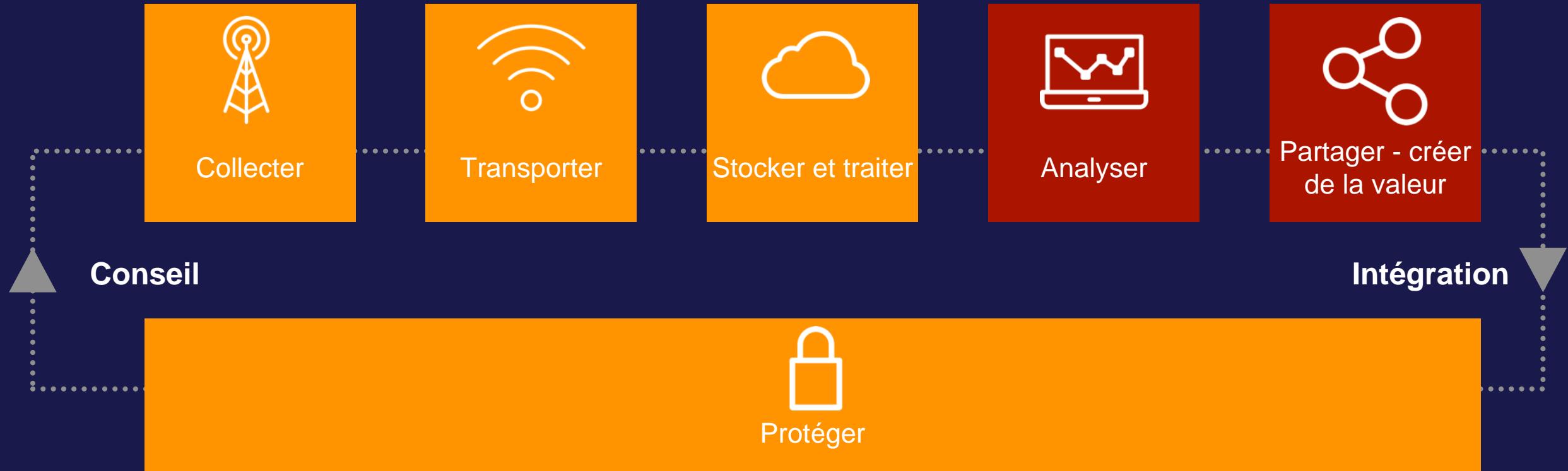
 Business & Decision

PARCOURS PROFESSIONNEL



Contact : erwan.josse@businessdecision.com

BUSINESS & DECISION DANS LE GROUPE ORANGE



2500 talents



depuis 1992

NOS DOMAINES D'EXPERTISES



Digital Experience :
CRM, mailing...



Entreprise
Performance Mgt



Compliance &
Security



Data Science & Artificial
Intelligence



Data
Gouvernance



Data
Intelligence



Big Data



Enabler Specialities

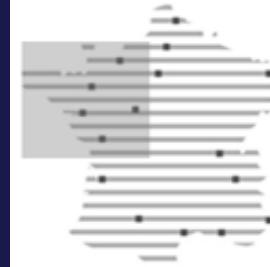
BUSINESS & DECISION DANS L'OUEST

AMOA
AMOE
Conseil

Centres de Services
TMA

Projets au forfait

Assistance Technique



BI
EPM

Big Data
Datascience

CRM
Digital

MDM
DQM

Web

300+

COLLABORATEURS

Nantes

Rennes

Niort

Le Mans

Banque

Assurance
Mutuelle

Public &
para public

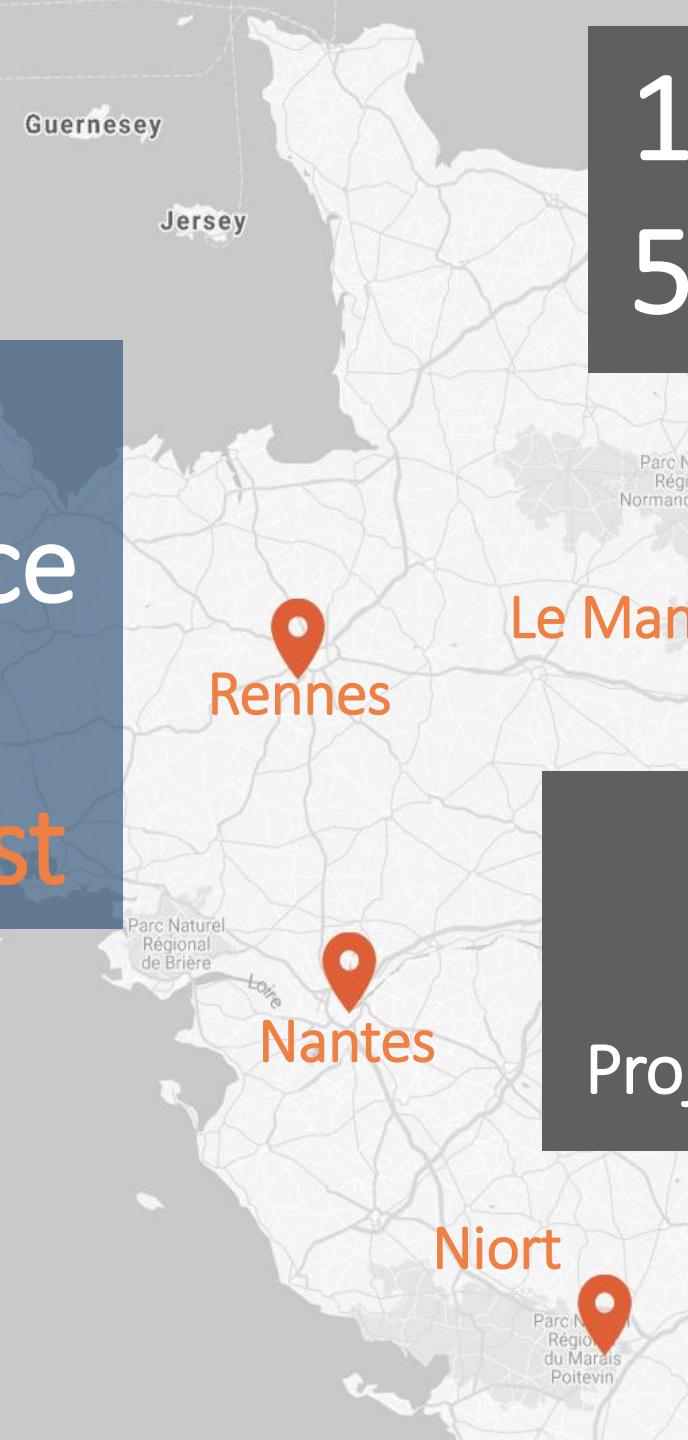
Industrie &
Services



POSTE IMMO



Equipe DataScience BD Nord-Ouest



12 Certifications sur

5 outils différents

17

clients différents

4 agences

5 Projets assurance

2

projets publique

8

projets L@b

25 projets menée en

4 ans pour un total

de 4065 jours de projet

7

projets banque

12

Projets retail

15 DataScientist

45 ans d'anciennetés cumulées

SOMMAIRE

➤ Pourquoi la régression ?

➤ Régression linaire

➤ Rappel

➤ Simple

➤ Multiple

➤ ANOVA

➤ Régression logistique

➤ Autres méthodes de prévision

SOMMAIRE

➤ Pourquoi la régression ?

➤ Régression linéaire

➤ Rappel

➤ Simple

➤ Multiple

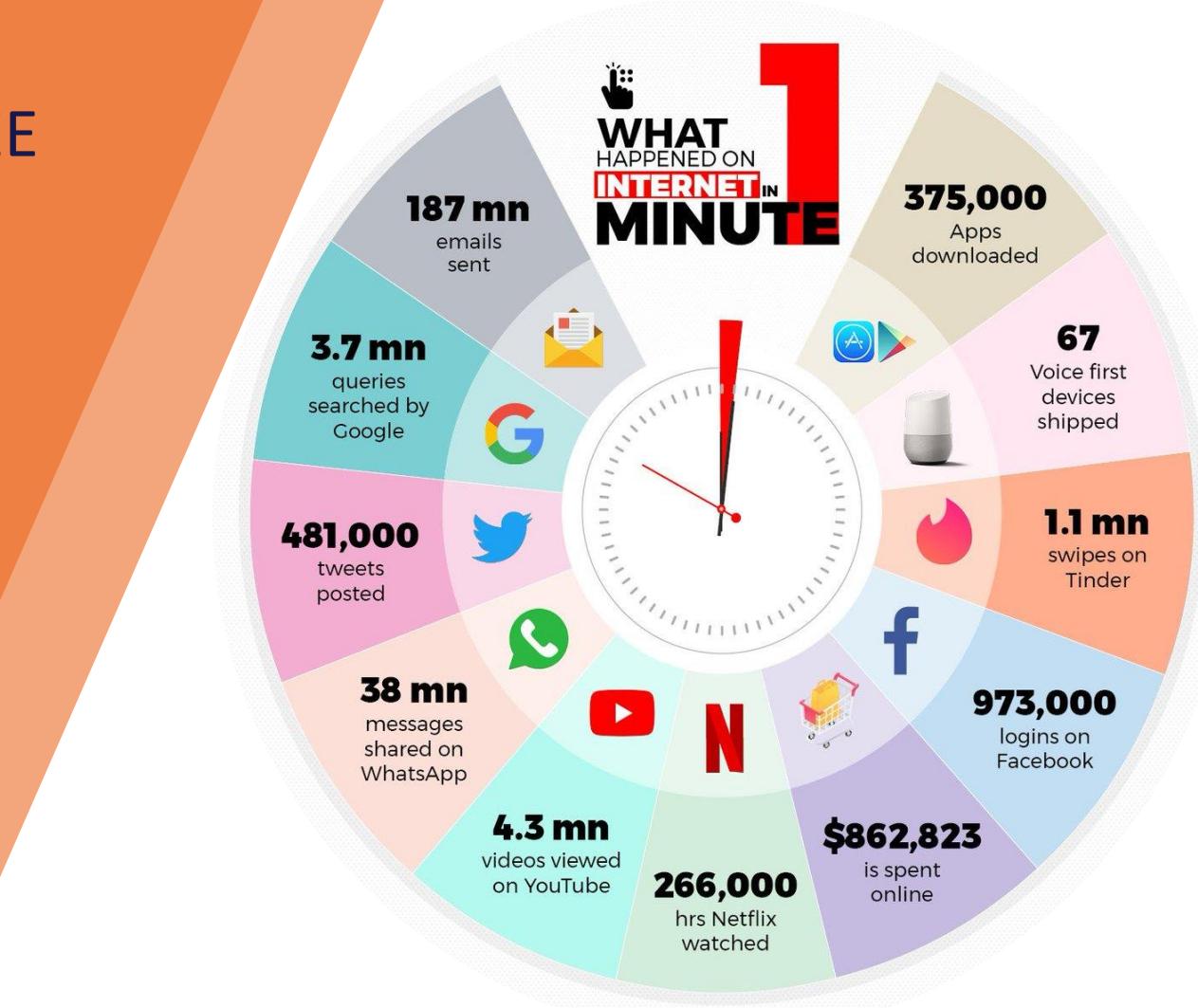
➤ ANOVA

➤ Régression logistique

➤ Autres méthodes de prévision

UN CHANGEMENT DE COMPORTEMENT

- ▶ **90 % DE LA DONNÉE À ÉTÉ GÉNÉRÉE
LES 2 DERNIÈRES ANNÉES**
- ▶ **MULTIPLICITÉ DES CANAUX**
- ▶ **DONNÉES NON-
STRUCTURÉES**



OBJECTIF DATASCIENCE

Prédire

- Personnaliser le discours
- Prédire le produit adapter
- Limiter le coût des campagnes

Résumer

- Croisement de données
- Création d'indicateurs
- Tableau de bord

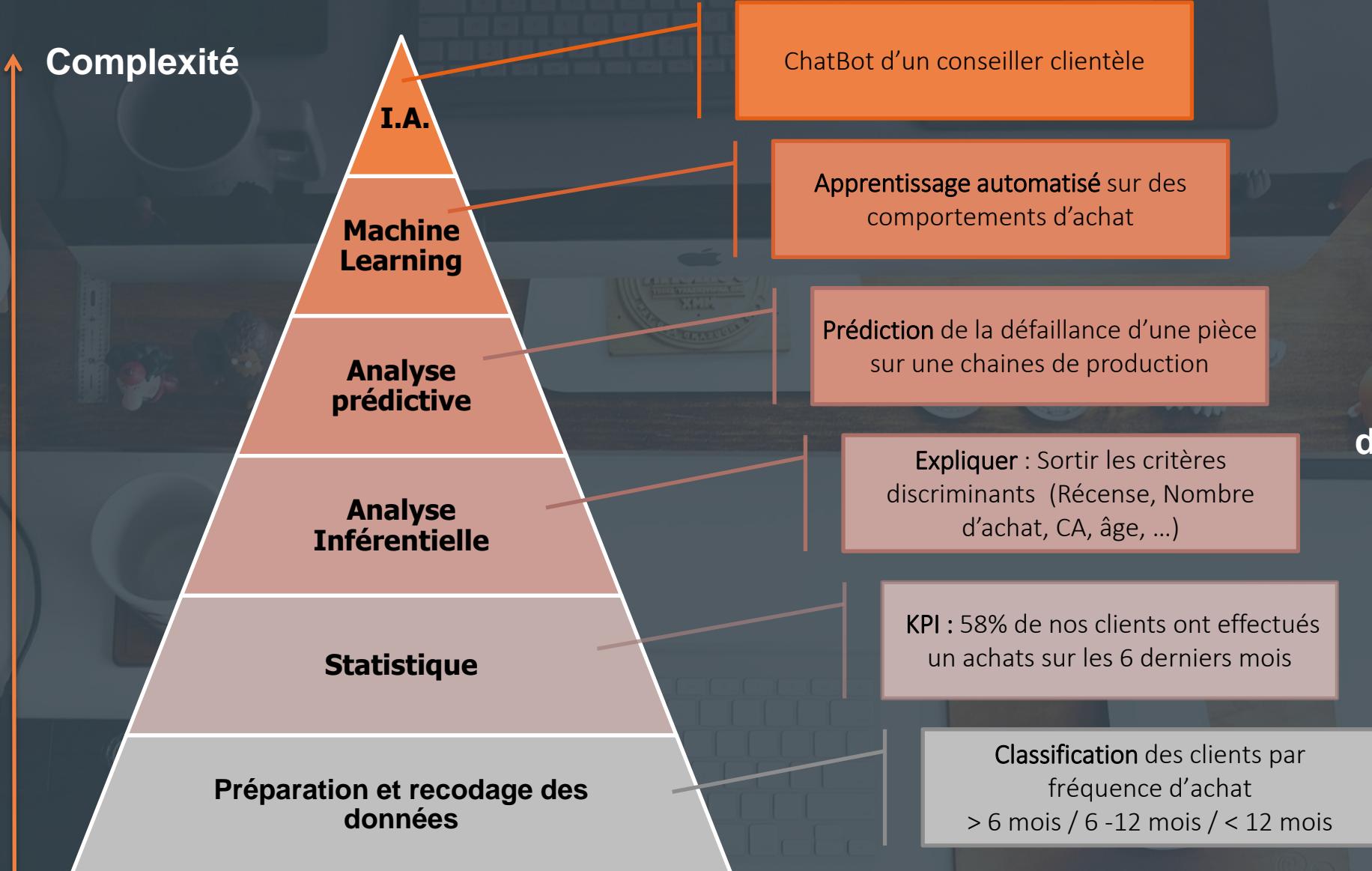
Analyser

- Comprendre les critères discriminants
- Regrouper les clients / produits
- Analyser les évènements

LES NIVEAUX DE DATASCIENCE

Pouvoir prédictif

Complexité



Exemples d'applications

Méthodologie DATASCIENCE

Problématique client



« Prédiction de l'attrition client »

Compréhension fine du métier

Compréhension de la problématique via la data



Ventes



Web



produit



DataPREP
(Données atypiques, aberrantes, formats...)



Objectif
Création d'indicateurs adaptés à la problématique
« Evolution de la fréquence d'achat »

Action client



Campagne Marketing adaptée

« Envoie d'un code promo avec -50% au 1 000 clientes »

Phase de prédition



Objectif

Calcul de probabilité / CA
« Sélection des 1000 clientes avec la plus forte probabilité »



Livrable
Liste de clients
Coefficient d'importance des indicateurs

Phase de compréhension

Phase d'explication



Objectif

Corrélation entre variable et regroupement des profils similaires
« VIP » / « Occasionnel »

Livrable

Axes discriminant
Segmentation

LES FAMILLES DE MÉTHODES DE RÉGRESSION

Y Quantitative

X quantitative

*Régression
linéaire*

X qualitative

ANOVA

Y qualitative

X quantitative

*Régression
logistique*

X qualitative

*Régression
logistique*

X mixte

ANCOVA

X mixte

*Régression
logistique*

EXEMPLE DE CAS D'USAGE

Retail



- NBA / NBO
- Attrition client
- Durée de vie produit
- Ouverture mail
- Calcul du CA prévisionnel par client

Industrie



- Probabilité de panne d'une machine / casse d'un produit
- Gestion des stocks
- Capacité optimal d'une machine

Public



- Détection de fraude
- Gestion des centres d'appels
- Renouvellement de la voirie / des équipements

Banque



- Attrition client
- Achat de service bancaire
- Cycle de vie client
- Canaux de contact privilégiés

Assurance

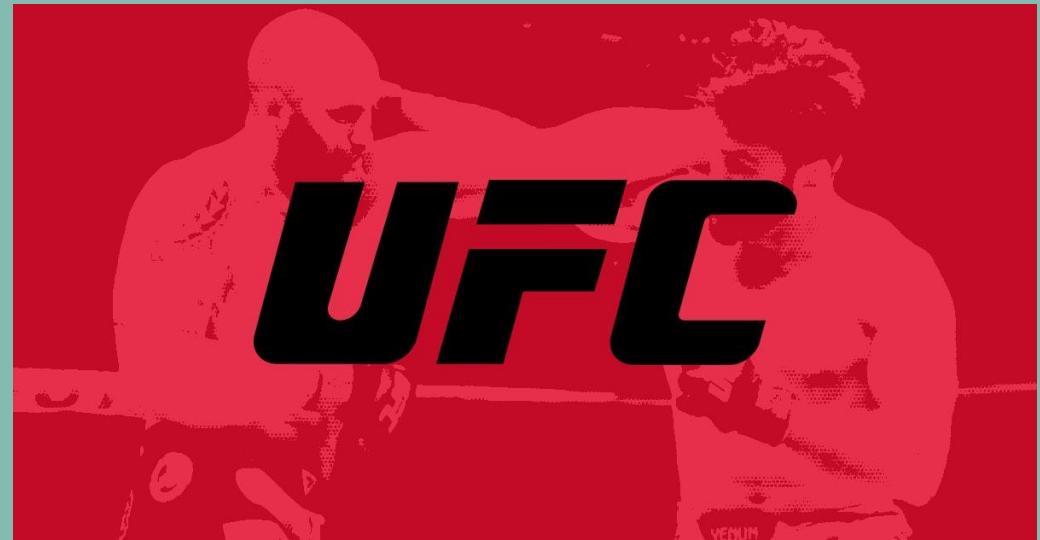


- Attrition client
- Assurance supplémentaire
- Cycle de vie client
- Nombre de sinistre

NOTRE CAS D'USAGE

DONNÉES

- Données récoltées sur « <https://www.kaggle.com/rajeevw/ufcdata> »
 - Données sur les matchs UFC de 1993 à aujourd'hui
 - 1915 joueurs
 - 5 144 matchs
- Base de données comprend plusieurs type d'indicateurs:
 - Lieu / date / heure du combat
 - Gagnant / durée du match / nombre de round
 - Nombre et type de coups
 - Nombre et type d'esquive
 - Information sur les combattants et leurs carrières



QUELS CAS D'USAGE ?



QUELS CAS D'USAGE ?

- Le poids et la taille sont corrélé ?
- Le poids et la tailles influencent la chance de victoire ?
- Le nombre de coups est-il corrélé avec l'issue du combat ?
- Existe-t-il une zone de frappe à priorisé pour gagner ?
- Le nombre de coup peut-il être prédis ?
- L'issue du combat peut-elle être prédites ?
- Qu'elles sont les critères les plus importants pour avoir les meilleures chance de victoire ?



Méthode de Régression

QUELS CAS D'USAGE ?



SOMMAIRE

➤ Pourquoi la régression ?

➤ Régression linéaire

➤ Rappel

➤ Simple

➤ Multiple

➤ ANOVA

➤ Régression logistique

➤ Autres méthodes de prévision

COEFFICIENT DE CORRÉLATION LINÉAIRE

- Coefficient de corrélation théorique

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \in [-1; 1]$$

- Coefficient de corrélation de bravois-Pearson

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

TEST DE SIGNIFICATIVITÉ

➤ Principe du test :

- **Tester la nullité du coefficient de corrélation**
- Si $\rho = 0$ alors il n'y a pas de corrélation linéaire entre X et Y
- Si $\rho \neq 0$ alors il n'y a pas de corrélation linéaire entre X et Y

➤ Condition :

- $X \sim N(\mu_1, \sigma_1)$ et $Y \sim N(\mu_2, \sigma_2)$

➤ Hypothèse de test

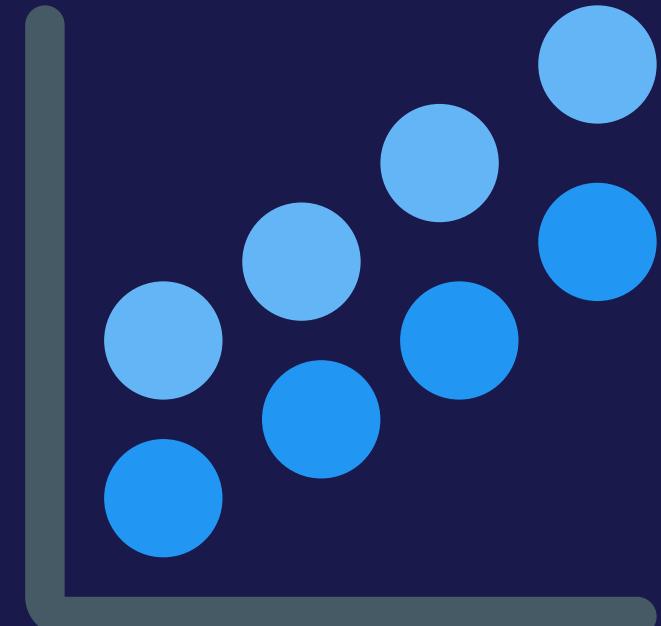
$$\begin{cases} H_0: \rho = 0 \\ H_1: \rho \neq 0 \end{cases}$$

➤ Statistique de test

$$T = \frac{R\sqrt{n - 2}}{\sqrt{1 - R^2}} \sim T_{n-2 \text{ ddl}}$$

CONCLUSION

- On cherche à mesurer le lien linéaire entre les deux variables
- On peut également modéliser le lien X et Y afin de réaliser des prédictions
 - Exemple : Prédire le chiffre d'affaire d'un magasin à partir de son nombre de produits vendus
- Mais :
 - Manque de précision
 - Pas de prise en compte d'autres événements
 - Modèle trop simpliste
- Recours à la régression linéaire



$$Y = \beta_1 X + \beta_0 + \epsilon$$

SOMMAIRE

➤ Pourquoi la régression ?

➤ Régression linaire

- Rappel
- Simple
- Multiple

➤ ANOVA

➤ Régression logistique

➤ Autres méthodes de prévision

RÉGRESSION LINÉAIRE SIMPLE

- I. Estimation des paramètres par la méthode des moindres carrés
- II. Tests et intervalles de confiance pour les paramètres
- III. Etude des résidus, points influents
- IV. Prévisions



RÉGRESSION LINÉAIRE

- Modélisation de la relation linéaire entre deux variables quantitatives
- Objectif explicatif et/ou prévisionnel :
 - Expliquer un phénomène
 - Interpréter les liens entre des mesures
 - Prédire de nouvelles données
- Variable à expliquer/d'intérêt, notée Y
 - Elle est quantitative
 - On parle également de variable à prédire / endogène / réponse
- Variable explicative, notée X
 - Elle est quantitative
 - On parle également de variable prédictrice / exogène

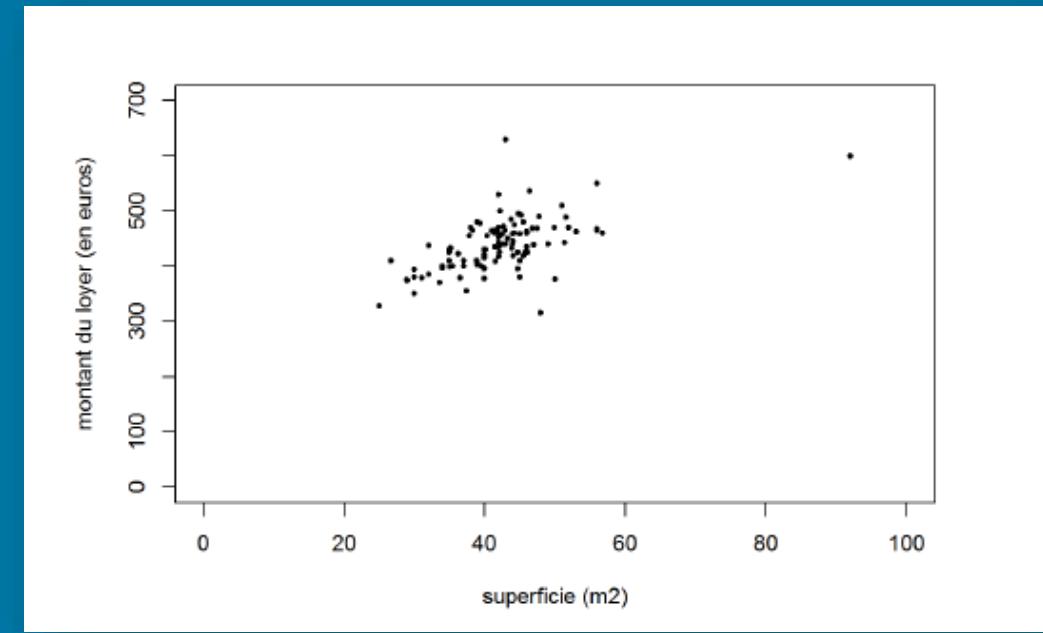


EXAMPLE : LOYER

- On a relevé dans les petites annonces les superficies (en m²) et les loyers (en Euros, CC) de 104 appartements de type T2. On veut apprécier le rôle de la superficie sur le montant de la location d'un appartement.

appt	loyer	Superficie
1	470 €	52 m ²
2	396 €	40 m ²
3	350 €	30 m ²
4	400 €	37 m ²
5	419 €	44,07 m ²
6	433 €	35,2 m ²
...
104	315 €	48 m ²

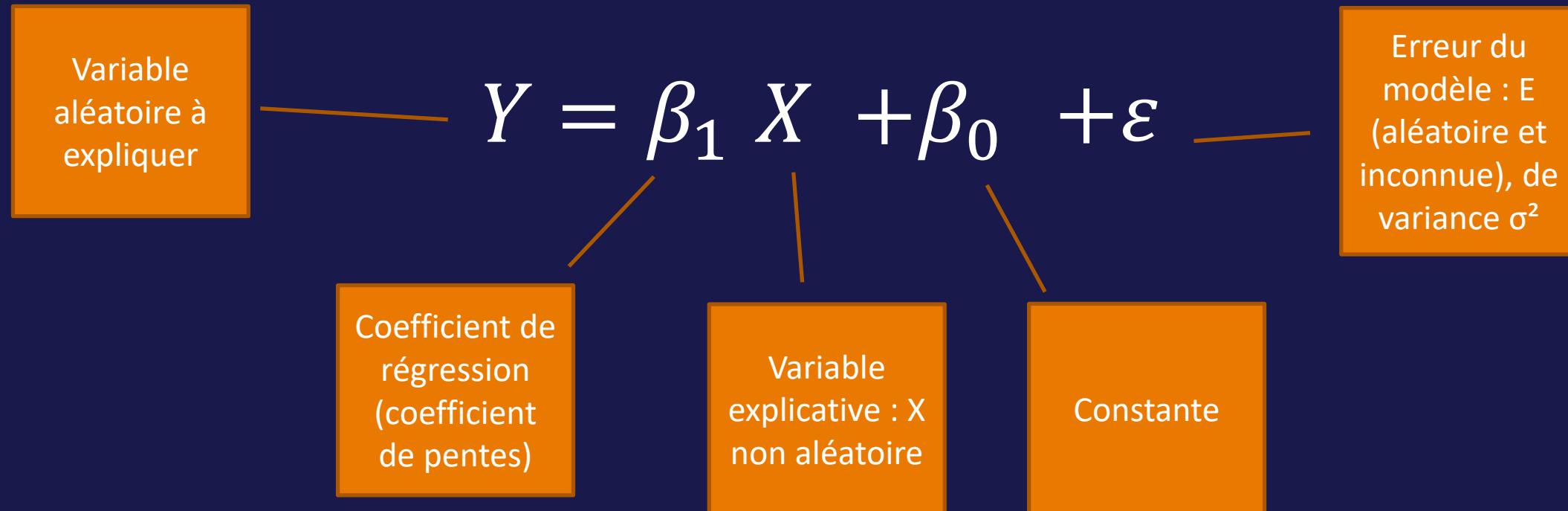
Y



X

MODÈLE DE RÉGRESSION LINÉAIRE

- Les données sont ajustées suivant la droite :



MODÈLE DE RÉGRESSION LINÉAIRE

➤ n observation $(x_i, y_i), i \in [1; n]$

➤ Pour chaque appartement i :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

➤ Réalisation de la variable aléatoire Y pour l'observation i

➤ Paramètre du modèle B_0 et B_1 (fixes et inconnus)

➤ Observation de la variable explicative X pour l'observation i : x_i

➤ Erreur du modèle pour l'observation i : ε_i



ESTIMATION DES PARAMÈTRES

- Estimation des paramètres : Minimisations des carrés des écarts entre les observations et la droites
- On chercher la fonction :

$$\operatorname{argmin} \sum_{i=1}^n (y_i - f(x_i))^2 = \operatorname{argmin} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

- Solution : Dérivation par rapport à β_0 et β_1 puis annulation des dérivées
- Nous obtenons ainsi :

$$\begin{cases} \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X} \\ \widehat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{(X_i - \bar{X})^2} \end{cases}$$

- $\widehat{\beta}_0$ = estimateur de β_0
- $\widehat{\beta}_1$ = estimateur de β_1
- \bar{X} = estimateur de la moyenne de X
- \bar{Y} = estimateur de la moyenne de Y

Problème : Comment être sûr que les estimateurs ne soit pas faux ?

PROPRIÉTÉ DES ESTIMATEURS

Hypothèse

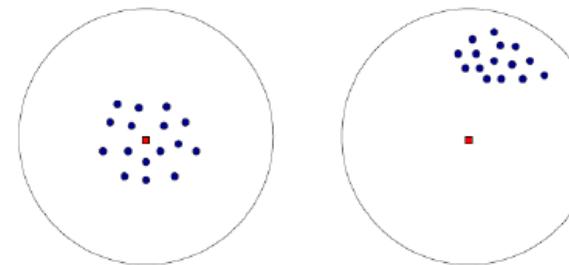
Les erreurs sont centrées, de même variance σ^2 (homoscédasticité) et non corrélées :

$$\forall i; j \in [1; n]$$

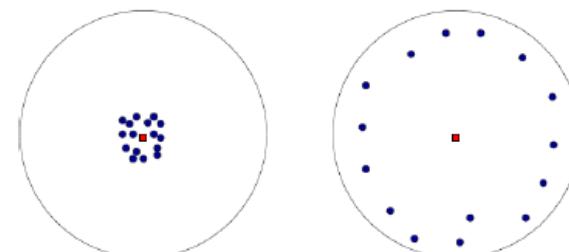
$$E(E_i) = 0$$

$$V(E_i) = \sigma^2$$

$$\text{cov}(E_i, E_j) = 0, \forall i \neq j$$



Erreurs non centrées = Biais



Dispersion des résidus non homogènes :
Hétéroscédasticité

PROPRIÉTÉ DES ESTIMATEURS

Propriété

- $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont sans biais

$$E(\widehat{\beta}_0) = \beta_0$$

$$E(\widehat{\beta}_1) = \beta_1$$

- Variance

$$V(\widehat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \bar{X})^2}$$

$$V(\widehat{\beta}_1) = \frac{\sigma^2}{n \sum_{i=1}^n (X_i - \bar{X})^2}$$

- Qualité des estimateurs (précision) :
 - σ^2 petit = Variabilité faible pour Y
 - $\sum_{i=1}^n X_i^2$ pas trop grand
 - $\sum_{i=1}^n (X_i - \bar{X})^2$ grand = Les mesures x_i sont dispersées autours de leurs moyennes
- **Théorème de Gauss-Markov** : Parmi les estimateurs linéaires sans biais, $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont de variance minimales

PROPRIÉTÉ DES ESTIMATEURS

Propriété

- Autres paramètres inconnu : σ^2 (variance résiduelle)
- Résidus : Estimateurs des erreurs E_i :

$$\hat{E}_i = Y_i - \hat{Y}_i$$

- Estimateur sans biais de σ^2 :

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

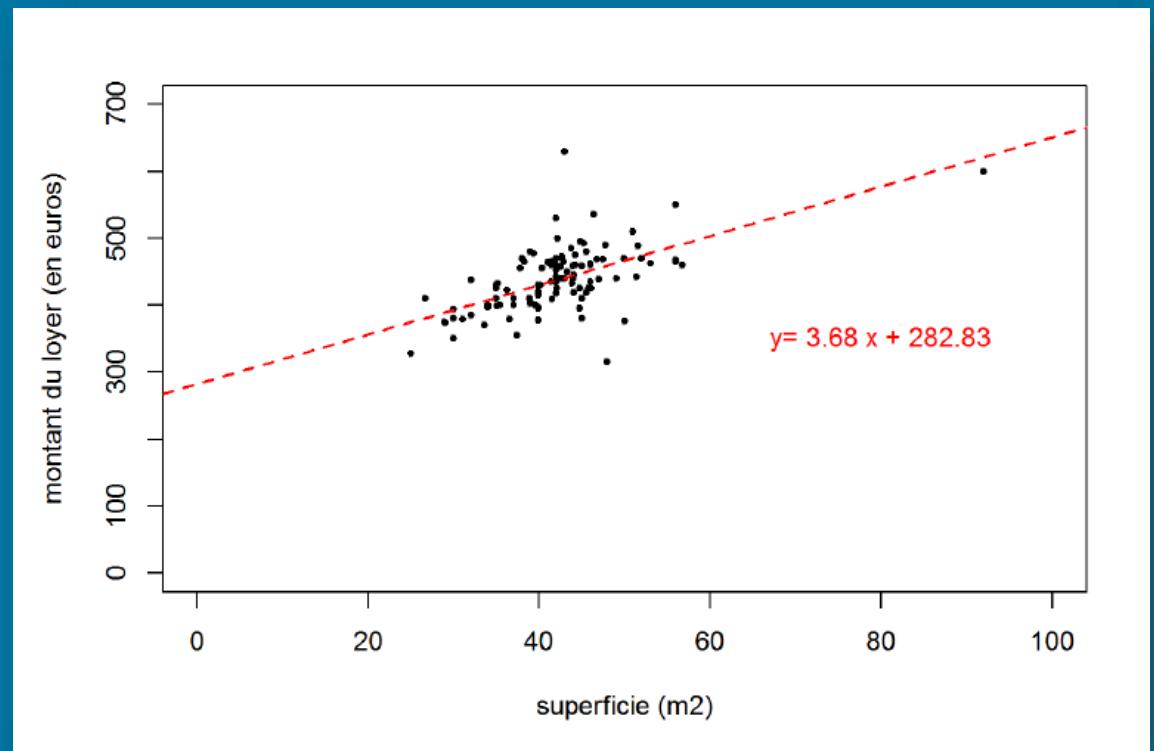


EXAMPLE : LOYER

- On a relevé dans les petites annonces les superficies (en m²) et les loyers (en Euros, CC) de 104 appartements de type T2. On veut apprécier le rôle de la superficie sur le montant de la location d'un appartement.

appt	loyer	Superficie
1	470 €	52 m ²
2	396 €	40 m ²
3	350 €	30 m ²
4	400 €	37 m ²
5	419 €	44,07 m ²
6	433 €	35,2 m ²
...
104	315 €	48 m ²

Y



X

CONCLUSION PARTIELLE

- Estimateurs des paramètres par la méthode des moindres carrés
 - Prévision d'une nouvelle valeur de la variable à expliquer Y pour chaque nouvelle valeur de variable explicative X
 - Qualité des estimateurs (Hypothèse H1)
 - Estimateurs des paramètres sans biais et de variances minimales
- Tests sur les estimateurs ? Intervalles de confiance ? Qualité de l'estimation ?

RÉGRESSION LINÉAIRE SIMPLE

- I. Estimation des paramètres par la méthode des moindres carrés
- II. Tests et intervalles de confiance pour les paramètres
- III. Etude des résidus, points influents
- IV. Prévisions



INTERVALLES DE CONFIANCE DES PARAMÈTRES

- Test de nullité des paramètres : Student

- Hypothèses testées :

- $H_0 : \beta_1 = 0$
- $H_1 : \beta_1 \neq 0$

- Statistique de test : $T = \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma^2_{\beta_1}}} \sim T(n - 2)$

- Décision :

- Rejet de H_0 si $|T_{obs}| > t_{1-\frac{\alpha}{2}}(n - 2)$ ou $p < \alpha$
- (p est la probabilité critique)

- Test similaire pour β_0

$$\left\{ \begin{array}{l} \beta_0 \in \left[b_0 \pm t_{1-\alpha/2}(n - 2) \sqrt{s_{\widehat{\beta}_0}^2} \right] \\ \beta_1 \in \left[b_1 \pm t_{1-\alpha/2}(n - 2) \sqrt{s_{\widehat{\beta}_1}^2} \right] \\ \sigma^2 \in \left[\frac{(n - 2)s^2}{c_{1-\frac{\alpha}{2}}(n - 2)} ; \frac{(n - 2)s^2}{c_{\alpha/2}(n - 2)} \right] \end{array} \right.$$

INFÉRENCE STATISTIQUE

- Test de nullité de β_0 :
 - Si significatif : La droite doit passer par l'origine
 - $Y = \beta_1 X + E$
 - Recalcule des paramètres du modèle sans constante

- Test de nullité de β_1 :
 - Si significatif : Y-a-t-il un lien entre Y et X ?

- Intervalle de confiance pour la droite de régression

$$\left[\hat{y}_i \pm t_{1-\frac{\alpha}{2}}(n-2) \sqrt{s^2 \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]} \right]$$

→ plus x_i est loin de \bar{x} , l'intervalle de confiance est plus large

```
Coefficients:  
Estimate Std. Error t value Pr(>|t|)  
(Intercept) 282.8307 20.5519 13.762 < 2e-16 ***  
SUPERFICIE 3.6786 0.4801 7.662 1.09e-11 ***  
---  
Residual standard error: 39.7 on 102 degrees of freedom
```

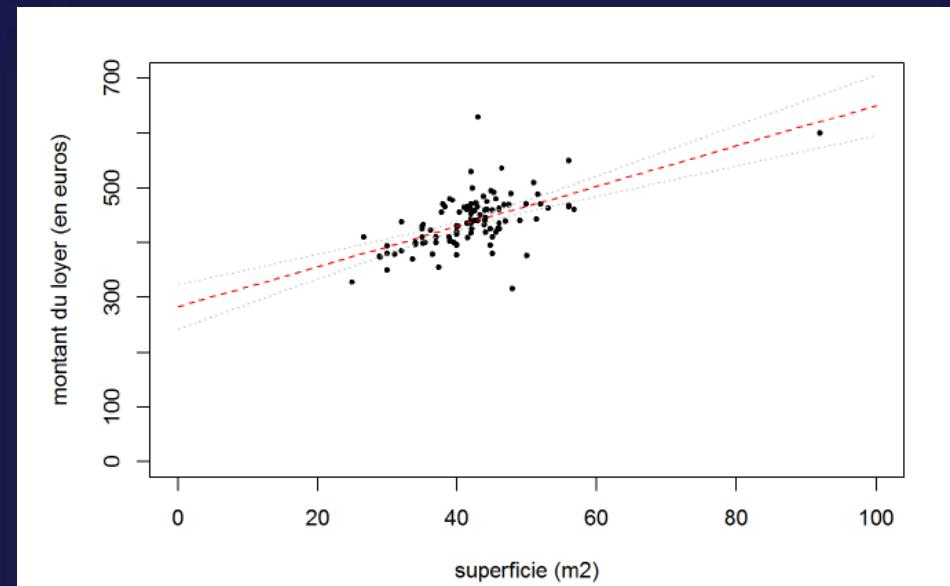


TABLE D'ANALYSE DE LA VARIANCE

Etude de la variabilité en Y : Décomposition de la variabilité en deux termes :

- Variabilité expliquée par le modèle = SC_M
- Variabilité résiduelle = SC_R
- $SC_T = SC_M + SC_R$
- $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$

Source	SC	ddl	$CM = CSC / ddl$
Modèle	$SC_M = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$ddl_m = 1$	$CM_M = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{p}$
Résidu	$SC_R = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$ddl_r = n - 2$	$CM_R = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - (p + 1)}$
Total	$SC_T = \sum_{i=1}^n (Y_i - \bar{Y})^2$	$ddl_T = n - 1$	$CM_T = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}$

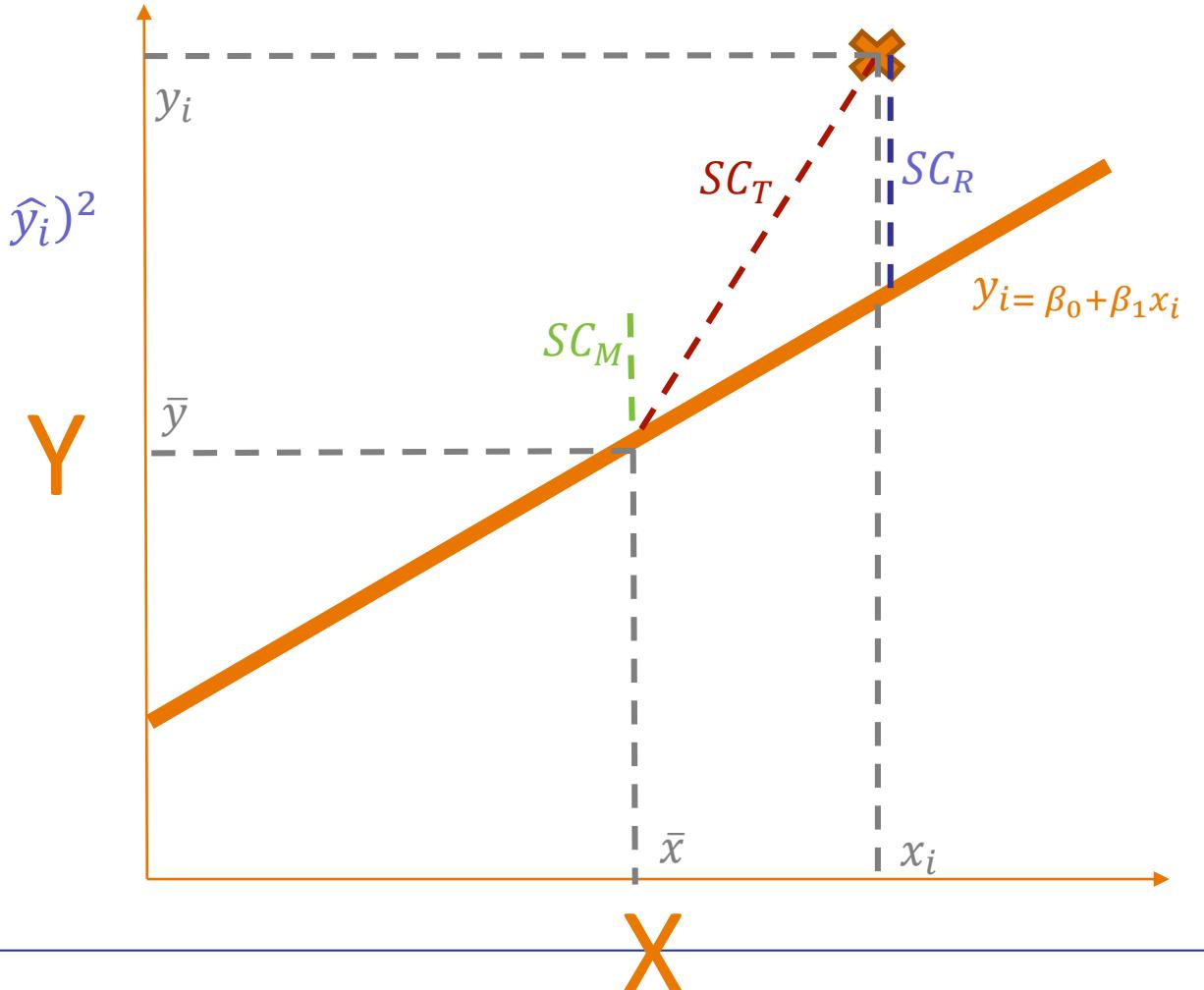
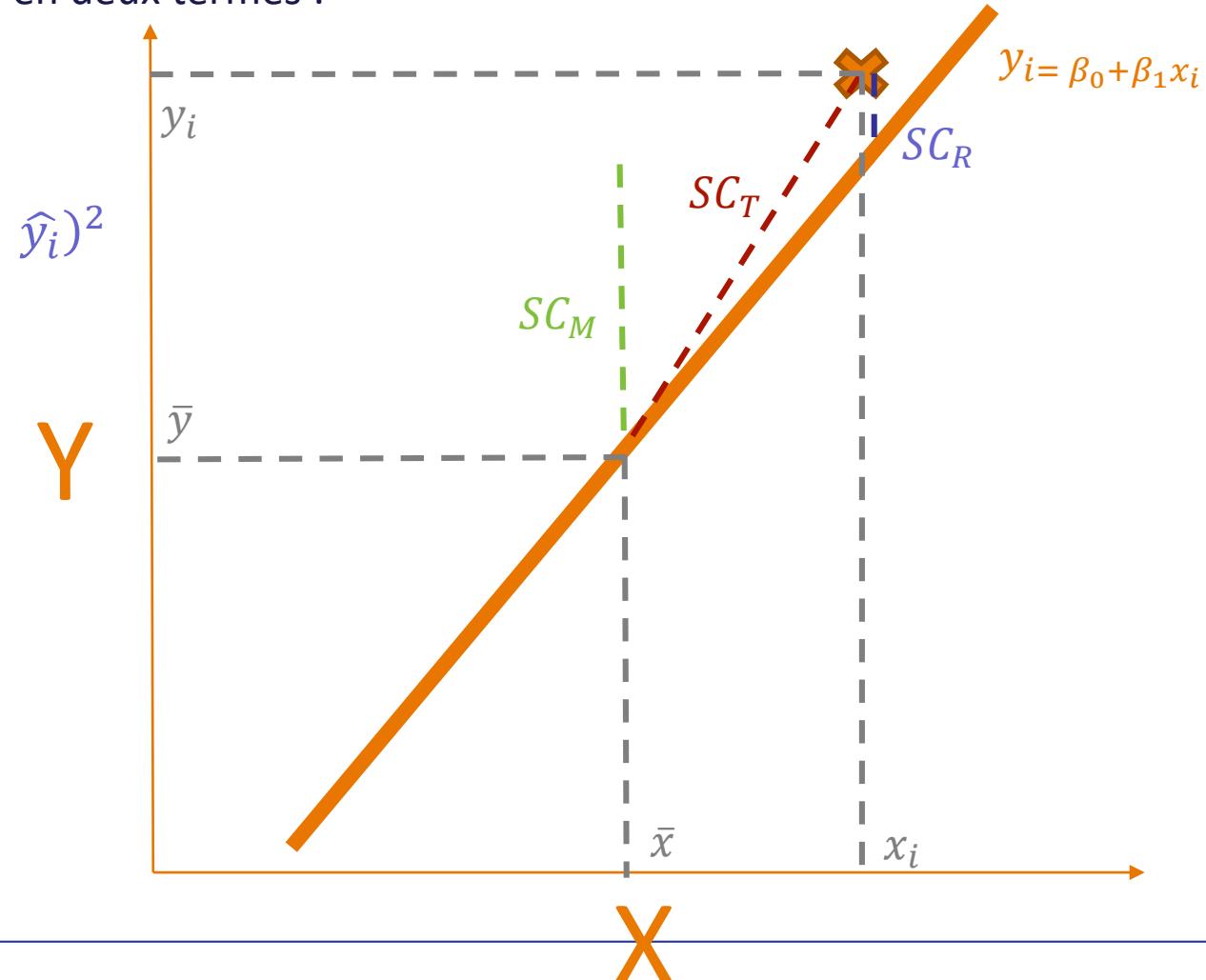


TABLE D'ANALYSE DE LA VARIANCE

Etude de la variabilité en Y : Décomposition de la variabilité en deux termes :

- Variabilité expliquée par le modèle = SC_M
- Variabilité résiduelle = SC_R
- $SC_T = SC_M + SC_R$
- $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$

Source	SC	ddl	$CM = CSC / ddl$
Modèle	$SC_M = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$	$ddl_m = 1$	$CM_M = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{p}$
Résidu	$SC_R = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$ddl_r = n - 2$	$CM_R = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - (p + 1)}$
Total	$SC_T = \sum_{i=1}^n (Y_i - \bar{Y})^2$	$ddl_T = n - 1$	$CM_T = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}$



COMPARAISON DE MODÈLE

Adéquation du modèles aux données :

- Variabilité de Y bien expliqué par le modèle
- Variabilité résiduelle petite par rapport à la variabilité total

Coefficient de détermination :

$$R^2 = \frac{SC_M}{SC_T} = 1 - \frac{SC_R}{SC_T}$$

- Compris entre 0 et 1 :
 - $R^2 \approx 1$: Le modèle explique toute la variabilité Y
 - $R^2 \approx 0$: Le modèle de régression linéaire est inadapté



COMPARAISON DE MODÈLE

Adéquation du modèles aux données :

- Variabilité de Y bien expliqué par le modèle
- Variabilité résiduelle petite par rapport à la variabilité total

Test global du modèle : Test de Fisher

$$F = \frac{CM_M}{CM_R}$$

- Hypothèse testées :
 - H_0 : Y mal expliquée par le modèle
 - H_1 : Y bien expliquée par le modèle
- Loi sous H_0 : $F \sim F(1; n - 2)$
- Rejet de H_0 si $F_{obs} > f_{1-\alpha}(n - 2)$ ou $p < \alpha$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	282.8307	20.5519	13.762	< 2e-16 ***
SUPERFICIE	3.6786	0.4801	7.662	1.09e-11 ***

Signif. codes:	0 ****	0.001 **	0.01 *	0.05 .
	'	'	'	'
	1			

Residual standard error: 39.7 on 102 degrees of freedom
Multiple R-squared: 0.3653, Adjusted R-squared: 0.3591
F-statistic: 58.71 on 1 and 102 DF, p-value: 1.094e-11

RÉGRESSION LINÉAIRE SIMPLE

- I. Estimation des paramètres par la méthode des moindres carrés
- II. Tests et intervalles de confiance pour les paramètres
- III. Etude des résidus, points influents**
- IV. Prévisions



VÉRIFICATION DES HYPOTHÈSES

HYPOTHÈSE DE RÉGRESSION LINÉAIRE

Modèle de régression : $Y = \beta_0 + \beta_1 X + E$

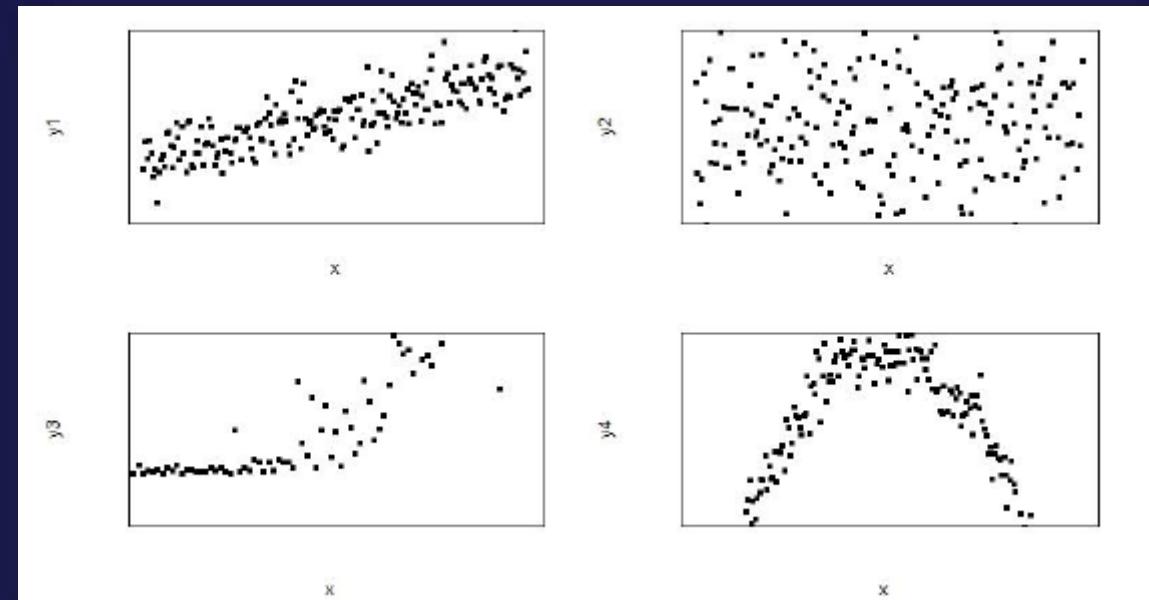
Hypothèse du modèle :

- Linéarité
- Erreurs indépendantes, centrées et de même variance (homoscédasticité)
- Normalité

La vérification de ces hypothèses est essentiellement graphique

Linéarité

Vérification sur le nuage de point (à faire avant la construction du modèle)



VÉRIFICATION DES HYPOTHÈSES

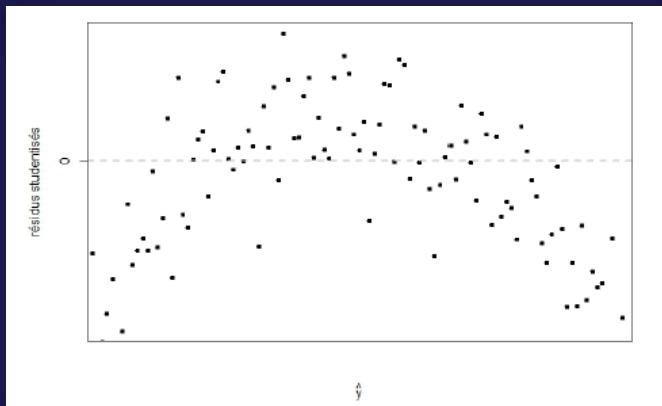
ANALYSE DES RÉSIDUS

Résidu : Estimés par $\hat{E} = Y_i - \hat{Y}_i$

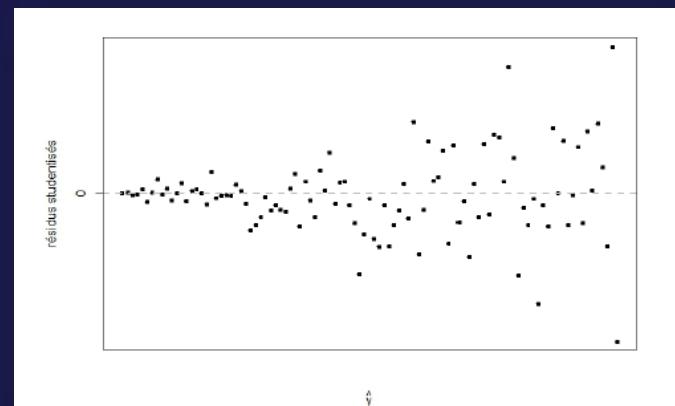
Pour mettre en **évidence des structures** (non indépendance), on regarde les résidus studentisés (divisés par leurs écarts-types empiriques) par validation croisée

Comment ? Crédit d'un graphique avec les résidus studentisés en ordonnées et les valeurs ajustées \hat{Y} :

- Vérification de l'ajustement global
- Vérification de l'hypothèse sur la structure de variance des erreurs (homoscédasticité)
- Repérage des points aberrants



Forme "banane" : non indépendance



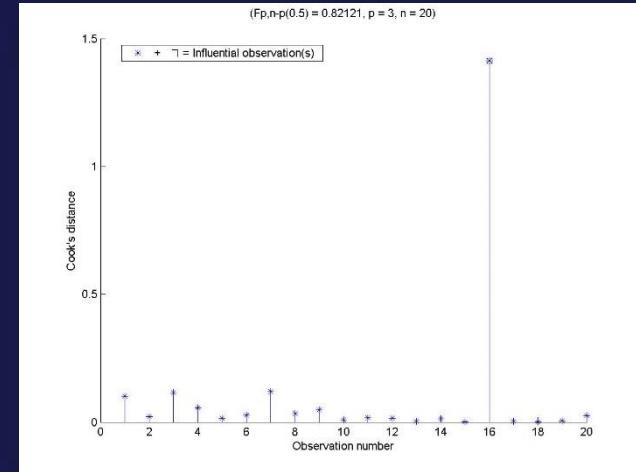
Forme "trompette" : variances des résidus non homogènes

VÉRIFICATION DES HYPOTHÈSES

DISTANCE DE COOK & NORMALITÉ

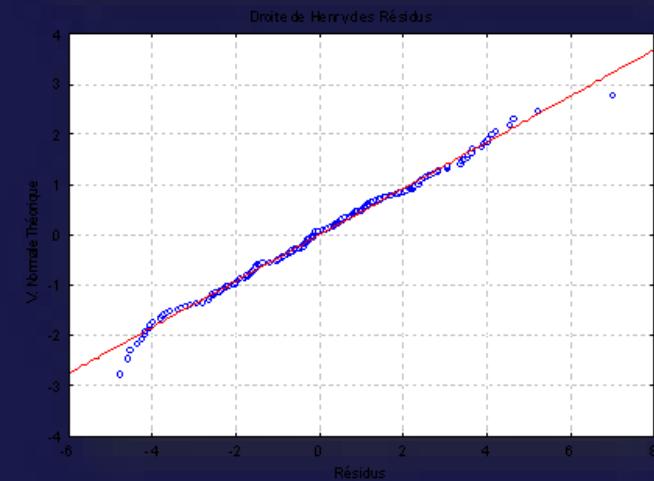
Point influent : Distance de Cook

- Mesurer l'influence de l'observation i sur l'estimation des paramètres du modèle
- Ecart entre $(b_0; b_1)$ et $(b_0^{-i}; b_1^{-i})$
- Seuil : $f_{0,5}(1; n - 1) \approx 0,5$



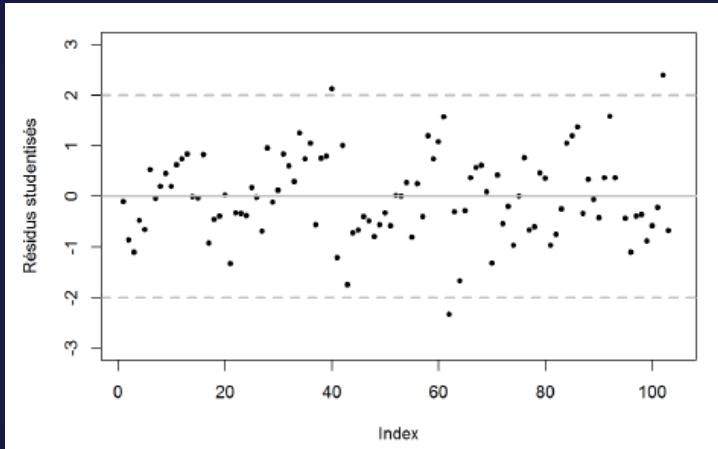
Normalité : droite de Henry

- Comparaison des quantiles empiriques avec ceux de la loi Normale

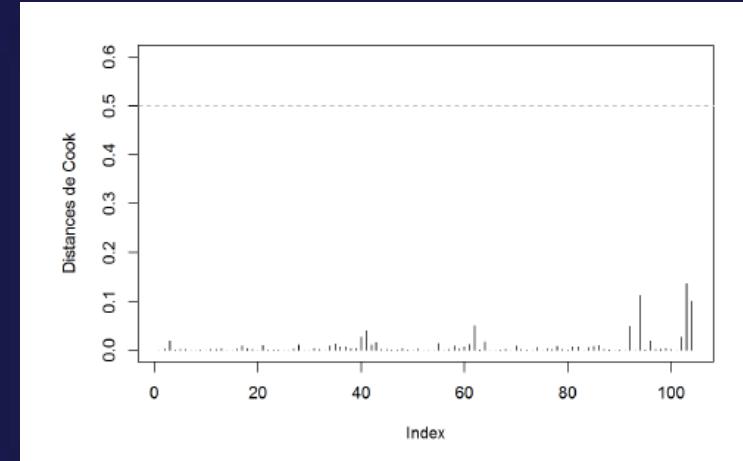


EXEMPLE CAS DES LOYERS

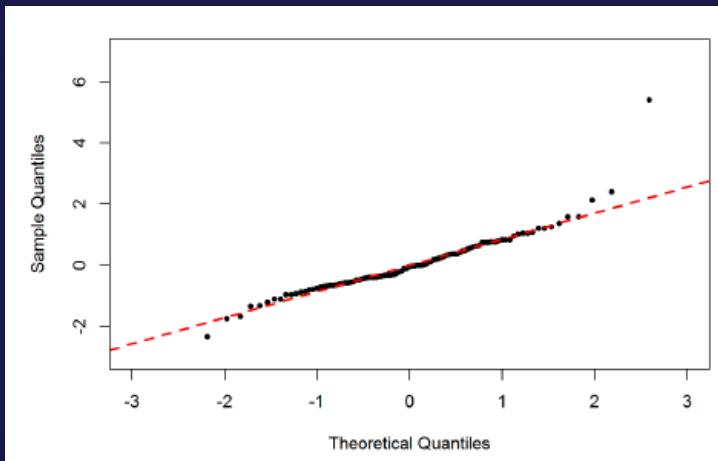
Analyse
des résidu



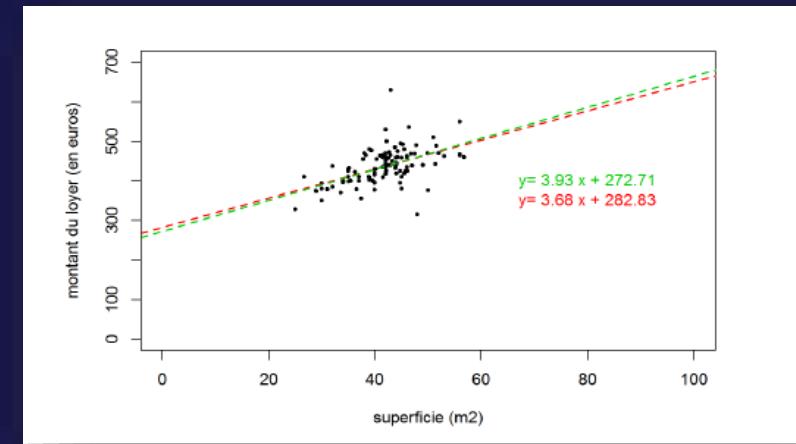
Distance
de Cook



Droite
d'henry



Graphique
X et Y



RÉGRESSION LINÉAIRE SIMPLE

- I. Estimation des paramètres par la méthode des moindres carrés
- II. Tests et intervalles de confiance pour les paramètres
- III. Etude des résidus, points influents
- IV. Prévisions**



PRÉVISION

Proposer une prévision pour la variable Y

- Soit x_0 une nouvelle valeur de la variable X
- On souhaite prédire y_0
 - Prévision d'une valeur moyenne pour Y en x_0

$$E(Y(x_0)) = \beta_0 + \beta_1 x_0$$

- On utilise les estimations obtenues par les estimateurs des moindres carrés $\widehat{\beta}_0$ et $\widehat{\beta}_1$

$$y_0^p = b_0 + b_1 x_0$$

- $(x_0$ et $y_0^p)$ non pas servi à calculer les estimations des paramètres
- La précision de la prédiction est :
 - Petite autour du centre de gravité du nuage (\bar{x})
 - Grande quand on s'éloigne de \bar{x}
 - Réduite quand N augmente

$$V(y_0^p) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)$$

PRÉVISION ERREUR & INTERVALLE DE PRÉVISION

Erreur de prévision : écart entre la vrai valeur et la valeur prédite $\rightarrow E_0^p = Y_0 - Y_0^P$

Propriété des erreurs de prévision

$$E(E_0^p) = 0 \quad V(E_0^p) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)$$

Intervalle de confiance pour la prévision

Intervalles de confiance au niveau $1 - \alpha$

$$\left[y_0^p \pm t_{1-\frac{\alpha}{2}}(n-2) \sqrt{s^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]} \right]$$

\rightarrow L'erreur et l'intervalle de confiance sont plus élevée quand x_0 s'éloigne de \bar{x}

EXEMPLE

CAS DES LOYERS VANNETAIS

```
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 272.7150   25.4753 10.705 < 2e-16 ***  
## SUPERFICIE    3.9270    0.6058   6.482 3.36e-09 ***
```

Quelle est la prédition pour une superficie de 35 m² ? 50m² ? 75 m² ?

Superficie	Prix
35 m ²	410,15 €
50 m ²	469,06€
75 m ²	567,24 €

RÉGRESSION LINÉAIRE SIMPLE

Conclusion



Analyse graphique

- Représentation graphique du nuage de point



Estimation et validation

- Estimation des paramètres (méthode des moindres carrés)
- Test global du modèle (test F) / tests de nullité des coefficients
- Qualité du modèle (coefficient R²)



Vérification des hypothèses

- Valeurs ajustées / résidus studentisés (indépendance, structure de variance, points aberrants)
- Distance de Cook (points influents)
- Droite de Henry (normalité)

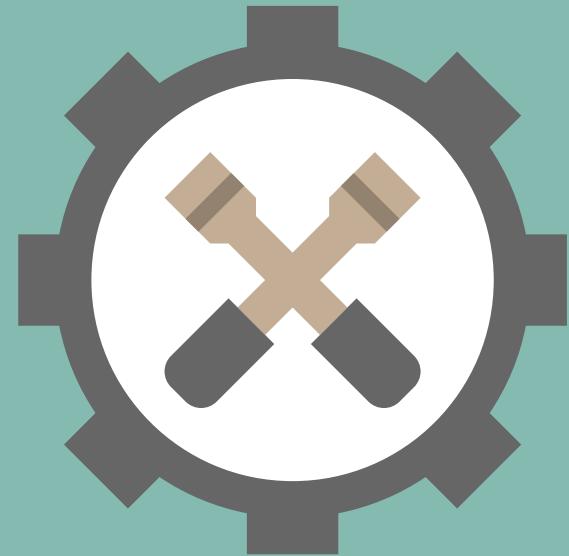


Prédiction

- Prédiction d'une nouvelle valeur

EXERCICE

1. Importer le fichier « Data.csv »
 1. Vérifier le format des champs et le bon import des données
2. Création d'une table de combattant unique
3. Calculer la régression linéaire simple entre le poids et la taille



SOMMAIRE

➤ Pourquoi la régression ?

➤ Régression linaire

- Rappel
- Simple
- Multiple

➤ ANOVA

➤ Régression logistique

➤ Autres méthodes de prévision

RÉGRESSION LINÉAIRE MULTIPLE

- I. Estimation des paramètres par la méthode des moindres carrés
- II. Tests et intervalles de confiance pour les paramètres
- III. Etude des résidus, points influents
- IV. Prévisions



RÉGRESSION LINÉAIRE MULTIPLE

USE CASE

Régression linéaire multiple : Modélisation de la relation entre une variable d'intérêt quantitative Y et p variables explicatives quantitatives :

- Variable à expliquer notée Y
- Variables explicatives notées X_1, X_2, \dots, X_p

Use case : Prévision de la concentration d'ozone O3

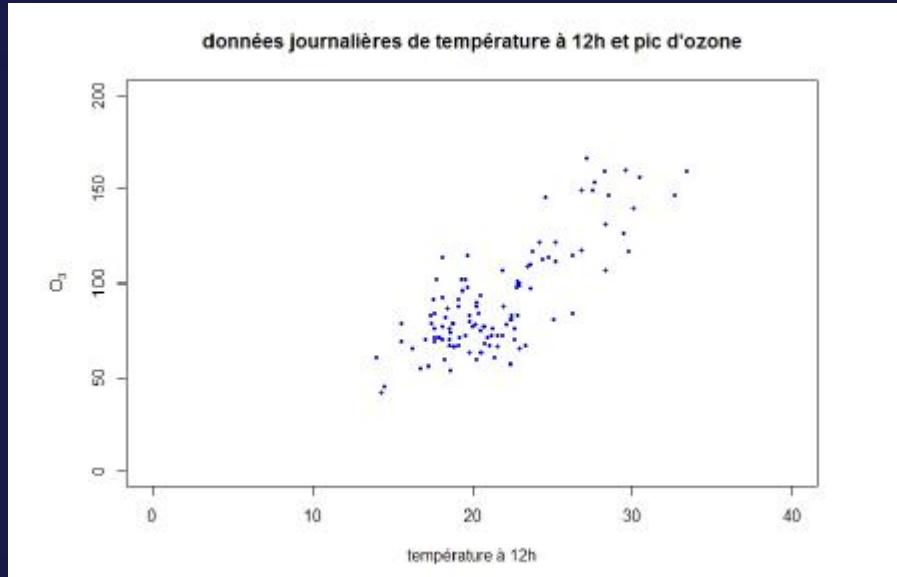
Pour des raisons de santé publique, on s'intéresse à la concentration d'ozone O3 dans l'air. On cherche en particulier à avoir si on peut expliquer le taux maximal d'ozone de la journée (en $\mu\text{g}/\text{ml}$) par la température à 12h (en $^{\circ}\text{C}$).

Température à 12h	O3 max
13,4 $^{\circ}\text{C}$	63,6
15,0 $^{\circ}\text{C}$	89,6
7,9 $^{\circ}\text{C}$	79
13,1 $^{\circ}\text{C}$	81,2
14,1 $^{\circ}\text{C}$	88,0
...	...

RÉGRESSION LINÉAIRE MULTIPLE

USE CASE

- On cherche à étudier le lien entre la température relevée à 12h et le pic d'ozone
 - Ajuster un modèle pour expliquer ce lien
 - Utiliser le modèle pour prédire la teneur max. en ozone de l'air en fonction de la température à 12h pour une nouvelle journée
- Rôle des variables :
 - Variable à expliquer : teneur max. en ozone
 - Variable explicative : température à 12h



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -27.4196    9.0335 -3.035   0.003 **
T12          5.4687   0.4125 13.258 <2e-16 ***
---
Residual standard error: 17.57 on 110 degrees of freedom
Multiple R-squared:  0.6151, Adjusted R-squared:  0.6116
F-statistic: 175.8 on 1 and 110 DF,  p-value: < 2.2e-16
```

RÉGRESSION LINÉAIRE MULTIPLE

- ... Mais modélisation simpliste
- D'autres variables peuvent expliquer cette concentration :
 - Vent
 - Rayonnement
 - Précipitation
 - ...
- Nécessité de prendre en compte ces variables dans le modèle
- Analyse de la relation entre la température à 12h (T12), le vent à 12h (V12) et le pic d'ozone (maxO3) : définir une fonction f telle que :

$$\text{maxO3}_i \approx f(T12_i; V12_i)$$

- Choix de la fonction f ?
- Trouver la fonction f parmi celles de la classe F qui minimise la fonction de coûts l :
- $\operatorname{argmin} \sum_{i=1}^n l(y_i - f(T12_i; V12_i))$

RÉGRESSION LINÉAIRE MULTIPLE

NOTATION

- Classe de fonction F : Fonction linéaire
- Fonction de coûts : Coût quadratique
- Modèle théorique

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + e_i$$

- Modèle ajusté :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip} ; \hat{e}_i = y_i - \hat{y}_i$$

- Notation matricielle :

$$Y = X\beta + E$$

$$Y_{|n \times 1|} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \quad X_{|n \times (p+1)|} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \quad \beta_{|(p+1) \times 1|} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_p \end{pmatrix} \quad E_{|n \times 1|} = \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{pmatrix}$$

RÉGRESSION LINÉAIRE MULTIPLE

ESTIMATEURS DES MOINDRES CARRÉES

- Intérêt du formalisme matricielle :
 - Généralisation du modèle linéaire simple ($p>1$)
 - Le vecteur $\hat{\beta}$ des estimateurs des moindres carrées est alors calculé par :

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix} \widehat{\beta_0} \\ \dots \\ \widehat{\beta_p} \end{pmatrix}$$

- Non singularité de la matrice $X'X$ (matrice X de plein rang)
- Problème pour le calcul des estimateurs en cas de colinéarité entre les variables explicatives

Hypothèse

La matrice X est de plein rang. Comme en général on a $p < n$, le rang de X est $p+1$

RÉGRESSION LINÉAIRE MULTIPLE

ESTIMATEURS DES MOINDRES CARRÉES

Hypothèse

Les erreurs sont centrées, de même variance σ^2 (homoscédasticité) et non corrélées :

$$\begin{aligned}\forall i; j \in [1; n] \quad E(E_i) = 0 \quad V(E_i) = \sigma^2 \\ cov(E_i, E_j) = 0, \forall i \neq j\end{aligned}$$

Propriété

$\hat{\beta}$ est sans biais

$$E(\hat{\beta}) = \beta$$

Variance

$$V(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

Autre paramètre inconnu : σ^2 (variance résiduelle)

- Estimateur des erreurs E :

$$\hat{E} = Y - \hat{Y}$$

- Estimateurs sans biais de σ^2 :

$$\widehat{\sigma^2} = \frac{\|\hat{E}\|}{n - (p + 1)} = \frac{SC_R}{n - (p + 1)}$$

- Estimateur de la variance de chaque paramètre β_k

$$\widehat{\sigma_{\beta_k}^2} = \widehat{\sigma^2}(X'X)^{-1}_{[kk]}$$

RÉGRESSION LINÉAIRE MULTIPLE

- I. Estimation des paramètres par la méthode des moindres carrés
- II. Tests et intervalles de confiance pour les paramètres
- III. Etude des résidus, points influents
- IV. Prévisions



RÉGRESSION LINÉAIRE MULTIPLE

Inférence statistique

Hypothèse

X est de plein rang (variable non-colinéaire). Les erreurs sont centrées, de même variance σ^2 (homoscédasticité) et non corrélées. De plus, elles suivent une loi Normale (de paramètre 0 et σ^2):

$$\begin{cases} \varepsilon_i \sim N(0; \sigma^2) \\ \varepsilon_i \text{ indépendant} \end{cases}$$

Loi des estimateurs

$$\widehat{\beta_k} \sim N\left(\beta_k; \sigma_{\widehat{\beta}_{[kk]}}^2\right) \Leftrightarrow \frac{\widehat{\beta_k} - \beta_k}{\sqrt{\sigma_{\widehat{\beta}_{[kk]}}^2}} \sim N(0; 1)$$

Loi des estimateurs

$$\frac{\widehat{\beta_k} - \beta_k}{\sqrt{\widehat{\sigma_{\widehat{\beta}_{[kk]}}^2}}} \sim T(n - (p + 1))$$

RÉGRESSION LINÉAIRE MULTIPLE

Tableau d'analyse de la variance & test de nullité

Décomposition de la variabilité de Y en deux termes :

- Variabilité expliquée par le modèle
- Variabilité résiduelle

Test statistique de nullité des paramètres :

- $H_0: \beta_k = 0$
- $H_1: \beta_k \neq 0$

- Loi sous H_0 : $T = \frac{\widehat{\beta}_k}{\sqrt{\widehat{\sigma}_{\beta_k}^2}}$ et $T \sim T(n - (p + 1))$

- Décision : Rejet de H_0 si $|T_{obs}| > t_{1-\frac{\alpha}{2}}(n - (p + 1))$ ou $p < \alpha$

Source	SC	ddl	CM = CSC / ddl
Modèle	$SC_M = \sum_{i=1}^n (\widehat{Y}_i - \bar{Y})^2$	$ddl_m = p$	$CM_M = \frac{\sum_{i=1}^n (\widehat{Y}_i - \bar{Y})^2}{p}$
Résidu	$SC_R = \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$	$ddl_r = n - (p + 1)$	$CM_R = \frac{\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2}{n - (p + 1)}$
Total	$SC_T = \sum_{i=1}^n (Y_i - \bar{Y})^2$	$ddl_T = n - 1$	$CM_T = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}$

$$\begin{cases} SC_T = SC_M + SC_R \\ ddl_T = ddl_M + ddl_R \end{cases}$$

Coefficients:					
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-14.4242	9.3943	-1.535	0.12758	
T12	5.0202	0.4140	12.125	< 2e-16 ***	
Vx12	2.0742	0.5987	3.465	0.00076 ***	

RÉGRESSION LINÉAIRE MULTIPLE

Comparaison de modèle

Adéquation du modèles aux données :

- Variabilité de Y bien expliquée par le modèle
- Variabilité résiduelle petite par rapport à la variabilité total

Test global du modèle : Test de Fisher

$$F = \frac{R^2}{1-R^2} \frac{n-(p+1)}{p}$$

- Hypothèse testées :
 - H_0 : Modèle nul : seulement la constante
 - H_1 : Modèle complet : avec la constante + les p variables explicatives
- Loi sous H_0 : $F \sim F(p; n - (p + 1))$
- Rejet de H_0 si $F_{obs} > F_{1-\alpha}(p; n - (p + 1))$ ou $p < \alpha$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-14.4242	9.3943	-1.535	0.12758
T12	5.0202	0.4140	12.125	< 2e-16 ***
Vx12	2.0742	0.5987	3.465	0.00076 ***

Residual standard error: 16.75 on 109 degrees of freedom

F-statistic: 102.7 on 2 and 109 DF, p-value: < 2.2e-16

RÉGRESSION LINÉAIRE MULTIPLE

- I. Estimation des paramètres par la méthode des moindres carrés
- II. Tests et intervalles de confiance pour les paramètres
- III. Etude des résidus, points influents
- IV. Prévisions**



Régression linéaire multiple

CHOIX DU MODÈLE

Nombre de paramètre dans le modèle :

- Trop important → Sur-apprentissage
- Trop faible → Biais des estimateurs
- Compromis entre un modèle exhaustif ou modèle parcimonieux
- Compromis entre un biais faible ou une variance faible

Problème : Comparer le modèle complet à p variables explicatives et un sous-modèle à q variables explicatives ($q < p$)

- Critère usuels :
 - R^2
 - R_α^2
 - AIC
 - BIC
 - ...



Régression linéaire multiple

COMPARAISON DE MODÈLE

Adéquation du modèles aux données :

- Variabilité de Y bien expliqué par le modèle
- Variabilité résiduelle petite par rapport à la variabilité total

Coefficient de détermination :

$$R^2 = \frac{SC_M}{SC_T} = 1 - \frac{SC_R}{SC_T}$$

- Compris entre 0 et 1 :
 - $R^2 \approx 1$: le modèle avec p variables explicatives explique toute la variabilité de Y
 - $R^2 \approx 0$: le modèle de régression linéaire est inadapté
- Le R^2 ne tient pas compte du nombre de paramètre :
- Définition d'un R^2 ajusté :

$$R_{\alpha}^2 = 1 - \frac{n - 1}{n - (p + 1)} \frac{SC_R}{SC_T}$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-14.4242	9.3943	-1.535	0.12758
T12	5.0202	0.4140	12.125	< 2e-16 ***
Vx12	2.0742	0.5987	3.465	0.00076 ***

Residual standard error: 16.75 on 109 degrees of freedom				
Multiple R-squared: 0.6533, Adjusted R-squared: 0.6469				
F-statistic: 102.7 on 2 and 109 DF, p-value: < 2.2e-16				

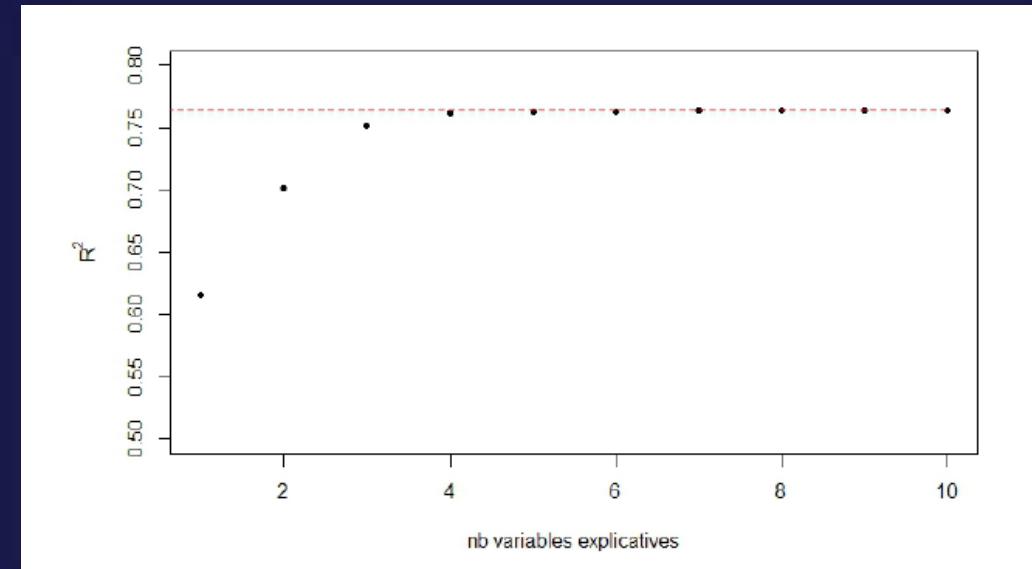
Régression linéaire multiple

CHOIX DU MODÈLE

$$R^2 = 1 - \frac{SC_R}{SC_T}$$

- Plus le R^2 est grand, meilleurs est le modèle
- On garde le modèle avec le R^2 le plus élevée
- Problème : Le R^2 augmente avec le nombre de variable
- Comparaison de modèle avec le même nombre de variable uniquement

Nb variable	R^2
1	0,6151
2	0,7012
3	0,7520
4	0,7622
5	0,7631
6	0,7636
7	0,7638



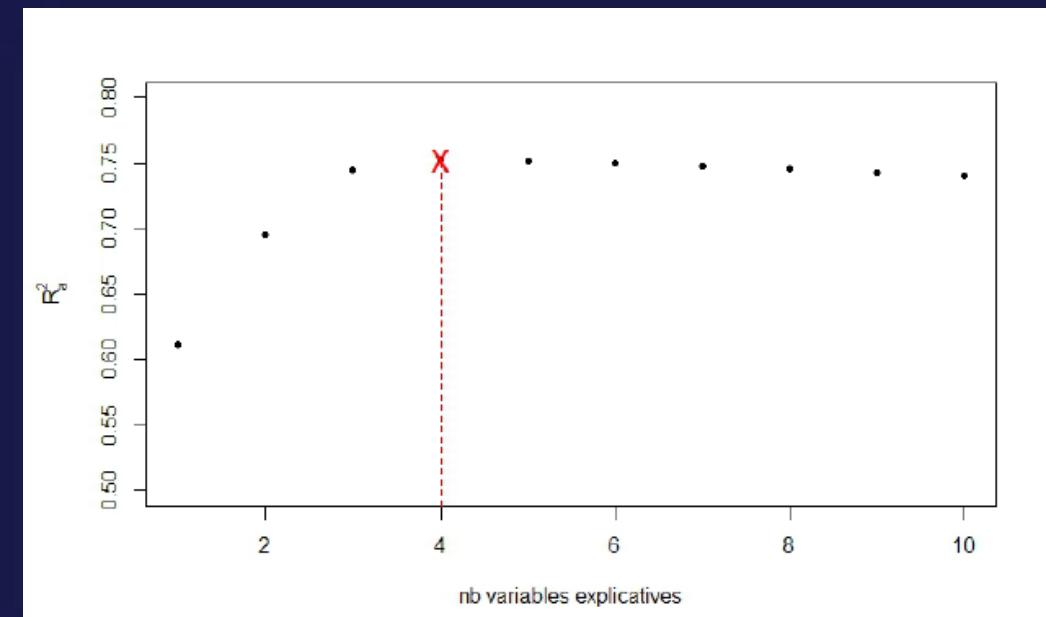
Régression linéaire multiple

CHOIX DU MODÈLE

$$R_{\alpha}^2 = 1 - \frac{n-1}{n-(q+1)} \frac{SC_R}{SC_T}$$

- Plus le R_{α}^2 est grand, meilleurs est le modèle
- On garde le modèle avec le R_{α}^2 le plus élevée

Nb variable	R ²
1	0,6116
2	0,6958
3	0,7451
4	0,7533
5	0,7519
6	0,7501
7	0,7479



RÉGRESSION LINÉAIRE MULTIPLE

USE CASE

Pour des raisons de santé publique, on s'intéresse à la concentration d'ozone O₃ dans l'air. On cherche en particulier à savoir si on peut expliquer le taux maximal d'ozone de la journée (en µg/ml) à l'aide de :

- la température à 9h, à 12h, à 15h
- la nébulosité à 9h, à 12h, à 15h
- la vitesse du vent (projetée dans la direction E/O) à 9h, à 12h, à 15h
- la teneur en ozone max. dans l'air la veille du jour considéré

Sélection du meilleur modèle au sens du R^2_α

Coefficients:					
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	12.63131	11.00088	1.148	0.253443	
T12	2.76409	0.47450	5.825	6.07e-08	***
Ne9	-2.51540	0.67585	-3.722	0.000317	***
Vx9	1.29286	0.60218	2.147	0.034055	*
maxO3v	0.35483	0.05789	6.130	1.50e-08	***

Residual standard error: 14 on 107 degrees of freedom					
Multiple R-squared: 0.7622, Adjusted R-squared: 0.7533					
F-statistic: 85.75 on 4 and 107 DF, p-value: < 2.2e-16					

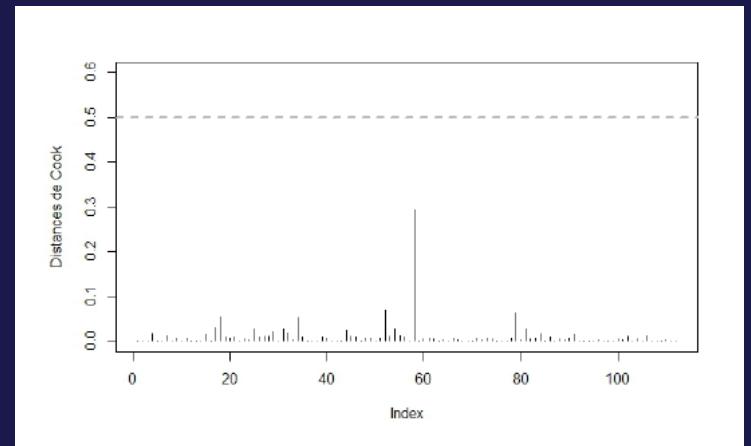
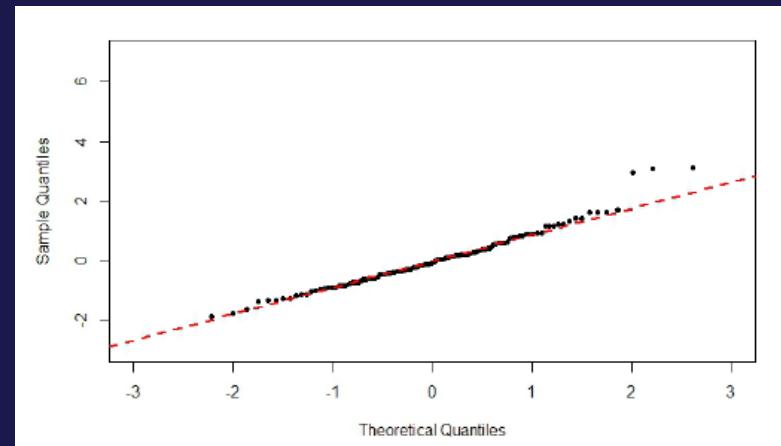
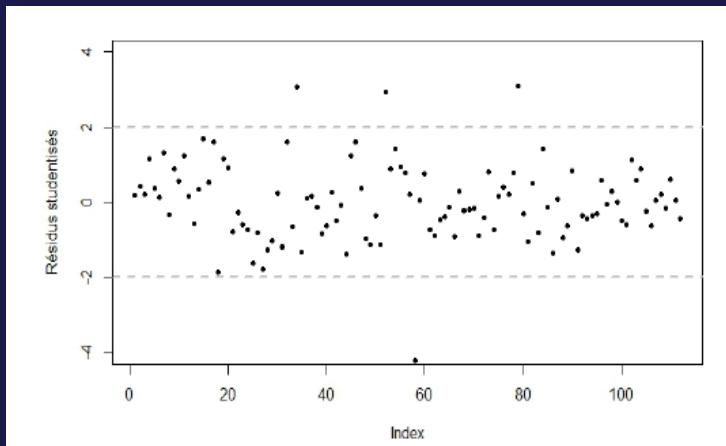
RÉGRESSION LINÉAIRE MULTIPLE

USE CASE

Vérification des Hypothèses :

Hypothèses sous-jacentes au modèle de régression linéaire :

- Linéarité
- Erreurs indépendantes, centrées et de même variance (homoscédasticité)
- Normalité



RÉGRESSION LINÉAIRE MULTIPLE

- I. Estimation des paramètres par la méthode des moindres carrés
- II. Tests et intervalles de confiance pour les paramètres
- III. Etude des résidus, points influents
- IV. Prévisions**



RÉGRESSION LINÉAIRE MULTIPLE

PRÉVISION

Objectif : Proposer une prévision pour la variable Y

- Soit $x_0 = (x_{01}; \dots; x_{0p})$ le vecteur des nouvelles valeurs des variables $X_1; \dots; X_p$
- On veut prédire y_0
 - Prévision d'une valeur moyenne pour Y en x_0

$$E(Y(x_0)) = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \dots + \beta_p x_{0p}$$

→ On utilise les estimations obtenues par les estimateurs des moindres carrés $\hat{\beta}$

$$y_0^p = b_0 + b_1 x_{01} + \dots + b_p x_{0p}$$

Erreur de prévision : $E_0^p = Y_0 - Y_0^p$

Propriété des erreurs de prévision

$$E(E_0^p) = 0$$

$$V(E_0^p) = \sigma^2(1 + (1; x_0)(X'X)^{-1}(1; x_0)')$$

Intervalles de confiance au niveau $1 - \alpha$

$$\left[y_0^p \pm t_{1-\frac{\alpha}{2}}(n-2) \sqrt{\hat{\sigma}^2(1 + (1; x_0)(X'X)^{-1}(1; x_0)')} \right]$$

```
R> x0=matrix(c(19,8,2.05,70), nrow=1)
R> colnames(x0)=c("T12", "Ne9", "Vx9", "maxO3v")
R> x0=data.frame(x0)
R> pred = predict(mod, newdata=x0, interval="pred")
R> pred
      fit     lwr      upr
72.51437 43.80638 101.2224
```

RÉGRESSION LINÉAIRE MULTIPLE



Analyse graphique

- Représentation graphique du nuage de point



Création modèle

- Méthode de sélection des variables : ascendante/ descendante / mixte ; critère : AIC/BIC/R2 ajusté
- Estimation des paramètres (méthode des moindres carrés)
- Test global du modèle (test F) / tests de nullité des coefficients / test de modèles emboités
- Qualité du modèle (coefficient R2 ajusté)



Vérification des hypothèses

- Valeurs ajustées / résidus studentisés (indépendance, structure de variance, points aberrants)
- Distance de Cook (points influents)
- Droite de Henry (normalité)

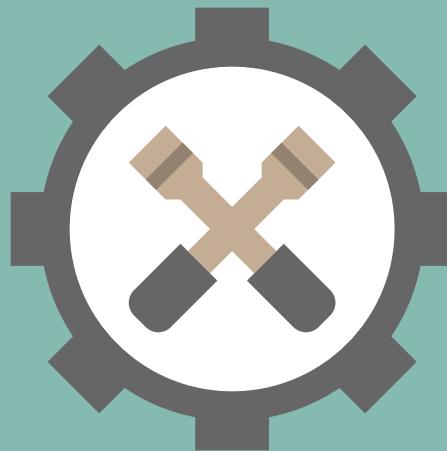


Prédiction

- Prédiction d'une nouvelle valeur

EXERCICE

1. Calculer la régression linéaire multiple entre le nombre de victoire / nombre de match et :
 - Le nombre de coup à la tête
 - Le nombre de coup au corp
 - Le nombre de coup au sol
2. Calculer la même régression mais calculer une régression par catégorie de poids.
 1. Que constatez-vous ?
1. A partir de la table combattant unique créé précédemment calculer la régression linéaire multiple entre le nombre de coup total et :
 - Le nombre de coup à la tête
 - Le nombre de coup au corp
 - Le nombre de coup au jambe
 - Le nombre de coup au sol
 - Le nombre de frappe au corp à corp



SOMMAIRE

► Pourquoi la régression ?

► Régression linaire

► ANOVA

► ANCOVA

► Régression logistique

► Autres méthodes de prévision

ANOVA

I. Modèle à 1 facteurs

- I. Estimation des paramètres
- II. Test

II. Modèle à 2 facteurs

- I. Sans interaction
- II. Avec interaction
- III. Modèle hiérarchique

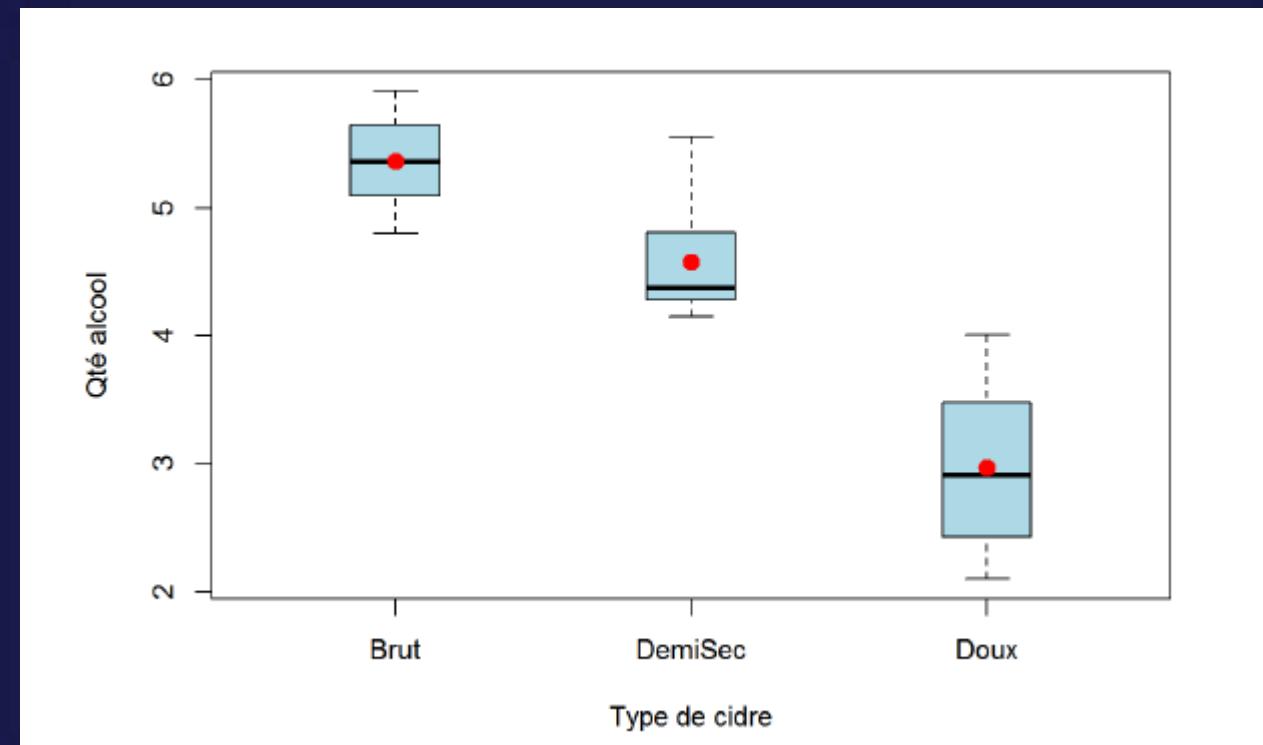


ANALYSE DE LA VARIANCE

USE CASE

On veut comparer la quantité d'alcool acquis (en g/l) dans des cidres bruts, demi-secs et doux. Pour cela, on a analysé 8 cidres de chaque type et répertorié les quantités d'alcool acquis.

Brut	Demi-sec	Doux
5,91	5,55	2,26
5,58	4,45	4,01
5,16	4,36	3,14
5,02	4,31	3,82
5,53	5,17	2,89
4,78	4,15	2,10



ANALYSE DE LA VARIANCE

DÉFINITION

- Analyse de la variance : modélisation de la relation linéaire entre
 - Variable à expliquer (variable réponse/d'intérêt), notée Y : **quantitative**
 - Variable(s) explicative(s), notée(s) X : **qualitative(s) (facteur(s))**
- Plan à un facteur : étude de l'effet d'un facteur X à I modalités sur la variable Y
- "Extension/généralisation" à plusieurs groupes du test de comparaison de moyennes
- Facteur contrôlé : valeurs prises par le facteurs fixées par l'expérimentateur
- Modalités de la variable qualitative : niveaux du facteur
- Plan d'expérience : définition des combinaisons de niveaux des facteurs (contrôlés)
 - Plan complets : toutes les combinaisons sont testées
 - Plan répété : une même combinaison observée plusieurs fois
 - Plan équilibré : même nombre de répétitions pour chaque combinaison
 - Plan équilibré et répété : plan équirépété



ANALYSE DE LA VARIANCE

NOTATION

Plan à un facteur :

- i : indice du niveau (modalité / groupe)
- I : nombre de niveau du facteur ($i \in \{1; \dots; I\}$)
- n_i : nombre d'expérience dans le groupe i
 - Cas équirépété : $\forall i, n_i = r$
- j : indice de l'expérience dans le groupe i ($j \in \{1; \dots; n_i\}$)
- n : nombre total d'expérience ($n = \sum_{i=1}^I n_i$)

Une expérience est identifiée par deux indices : i et j

- y_{ij} : mesure de la variable d'intérêt pour l'expérience j du groupe i
- \bar{y}_i : moyenne du groupe i
- \bar{y} : moyenne générale sur l'ensemble des n expériences

Brut	Demi-sec	Doux	
5,91 i=1	5,55	2,26	
5,58	4,45 j=2	4,01	
5,16	4,36	3,14	
$y_{1,3}$			
5,02	4,31	3,82	
5,53	5,17	2,89	
r=5 $\bar{y}_1 = 5,33$ $n_1 = 5$	r = 5 $\bar{y}_2 = 4,66$ $n_2 = 5$	r = 5 $\bar{y}_3 = 3,24$ $n_3 = 5$	$\bar{y} = 4,47$ $n=15$ $I=3$

ANALYSE DE LA VARIANCE

NOTATION – MODÈLE À UN FACTEUR

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

Avec

- $i \in \{1, \dots, I\}$
- $j \in \{1, \dots, n_i\}$
- $n = \sum_{i=1}^I n_i$
 - Cas équirépété = $n = r \times I$
- ε_{ij} est une réalisation de $E_{ij} \sim N(0, \sigma^2)$

Autre paramétrisation

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Avec :

- μ : Effet moyen général
- α_i : effet spécifique du niveau i

ANALYSE DE LA VARIANCE

ESTIMATION DES PARAMÈTRES

- Méthode des moindres carrés :

- Trouver $\hat{\mu}, \hat{\alpha}_i$ tels que $\sum \varepsilon_{ij}^2$ minimale

$$\hat{\mu} = \frac{1}{I} \sum_{i=1}^I \bar{y}_i = \frac{1}{I} \sum_{i=1}^I \bar{y}_i$$

- Cas équirepété : $\hat{\mu} = \bar{y}$

- $\hat{\alpha}_i = \bar{y}_i - \hat{\mu}$

- Cas équirepété : $\hat{\alpha}_i = \bar{y}_i - \bar{y}$

- Qualité des estimateurs :

- Variance dépend de σ^2 :

- Mauvaise estimation si population hétérogène

- Variance dépend de r :

- Bonne estimation si beaucoup d'estimation

Brut	Demi-sec	Doux	
5,91	5,55	2,26	
5,58 i=1	4,45 j=2	4,01	
5,16 y_{1,3}	4,36	3,14	
5,02	4,31	3,82	
5,53	5,17	2,89	
r=5 $\bar{y}_1 = 5,33$ $n_1 = 5$	r = 5 $\bar{y}_2 = 4,66$ $n_2 = 5$	r = 5 $\bar{y}_3 = 3,24$ $n_3 = 5$	$\bar{y} = 4,47$ n=15 I=3

Propriétés des estimateurs

- $\hat{\mu}$ et $\hat{\alpha}_i$ sont sans biais
- $\mathbb{V}(\hat{\alpha}_i) = \frac{I-1}{I} \frac{\sigma^2}{r}$

ANALYSE DE LA VARIANCE

ECART AU MODÈLE : RÉSIDU

- Estimation :

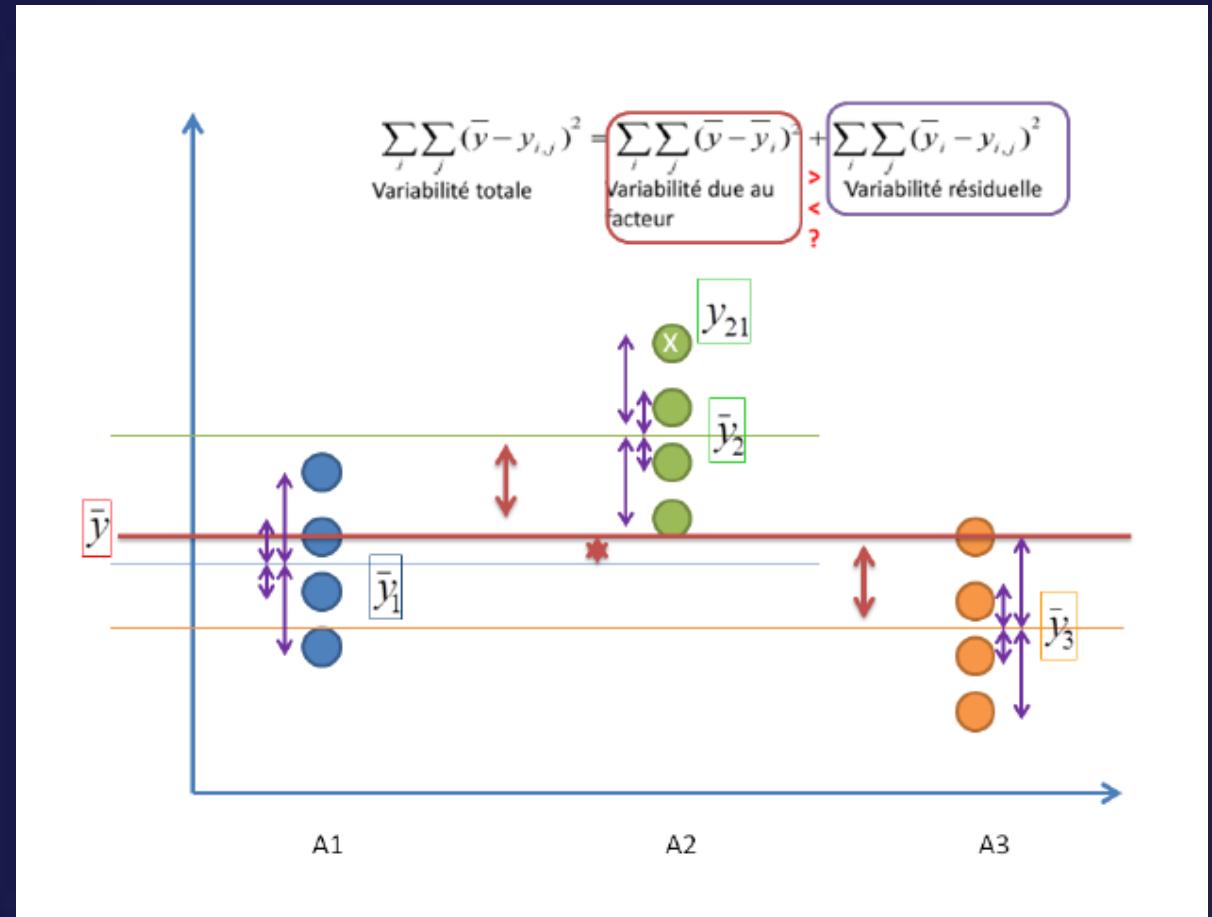
$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_i$$

- Propriété :

$$\sum_j e_{ij} = \sum_j (y_{ij} - \hat{y}_{ij}) = \sum_j (y_{ij} - \bar{y}_i) = 0$$

- Estimation de σ^2 :

$$\widehat{\sigma^2} = \frac{\sum_j e_{ij}^2}{n - I}$$



ANALYSE DE LA VARIANCE

DÉCOMPOSITION DE LA VARIANCE & QUALITÉ DU MODÈLE

- Variabilité total :**
 - Variabilité expliqué par le modèle
 - Variabilité résiduelle
 - $SC_T = SC_F + SC_R$

- Qualité du modèle :
 - Coefficient de détermination :

$$R^2 = \frac{SC_T - SC_R}{SC_T} = \frac{SC_F}{SC_T}$$

Source	SC	ddl
Modèle	$SC_F = \sum_{i=1}^n r(\bar{Y}_i - \bar{Y})^2$	$ddl_F = I - 1$
Résidu	$SC_R = \sum_{i,j} (Y_{ij} - \bar{Y}_i)^2$	$ddl_r = n - I$
Total	$SC_T = \sum_{ij} (Y_{ij} - \bar{Y})^2$	$ddl_T = n - 1$

- Test global du modèle : Test de Fisher
 - Hypothèse testées :
 - H_0 : Y mal expliqué par le facteur $\rightarrow \alpha_i = 0$
 - H_1 : Y bien expliqué par le facteur $\rightarrow \alpha_i \neq 0$

$$\begin{cases} SC_T = SC_M + SC_R \\ ddl_T = ddl_M + ddl_R \end{cases}$$

- $F = \frac{CM_F}{CM_R} = \frac{SC_F/I-1}{SC_R/n-I} \sim F(I-1; n-I)$
- Rejet de H_0 si $F_{obs} > F_{1-\alpha}(I-1; n-I)$ ou $p < \alpha$

ANOVA

I. Modèle à 1 facteurs

- I. Estimation des paramètres
- II. Test

II. Modèle à 2 facteurs

- I. Sans interaction
- II. Avec interaction
- III. Modèle hiérarchique



ANALYSE DE LA VARIANCE

TEST DE COEFFICIENT

- Loi de $\hat{\alpha}_i : N(\alpha_i; \sigma_{\alpha_i}^2)$
- Estimation de $\sigma_{\alpha_i}^2$

$$\widehat{\sigma}_{\alpha_i}^2 = \frac{I-1}{I} \frac{\widehat{\sigma}^2}{r} = \frac{I-1}{I} \frac{CM_R}{r} \sim T(n-I)$$

- Hypothèse testées :
 - $H_0 : \alpha_i = 0$
 - $H_1 : \alpha_i \neq 0$

$$T = \frac{\widehat{\alpha}_i - \alpha_i}{\sqrt{\widehat{\sigma}_{\alpha_i}^2}} \sim T(n-I)$$

- Rejet de H_0 si $T_{obs} > T_{1-\alpha/2}(n-I)$ ou $p < \alpha$

```
Ftest
      SS   df   MS F value    Pr(>F)
typeCidre 23.8249  2 11.9125  41.619 4.944e-08 ***
Residuals  6.0108 21  0.2862
---
Ttest
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.30292  0.10921 39.4015 <2e-16 ***
typeCidre - Brut 1.05958  0.15444  6.8607 <2e-16 ***
typeCidre - DemiSec 0.27458  0.15444  1.7779 0.0899 .
typeCidre - Doux -1.33417  0.15444 -8.6386 <2e-16 ***
Analysis of Variance Table

Response: qteAlcool
          Df Sum Sq Mean Sq F value    Pr(>F)
typeCidre  2 23.8249 11.9125  41.619 4.944e-08 ***
Residuals 21  6.0108  0.2862

R> 23.8249/(23.8249+6.0108)
[1] 0.7985367
```

ANALYSE DE LA VARIANCE

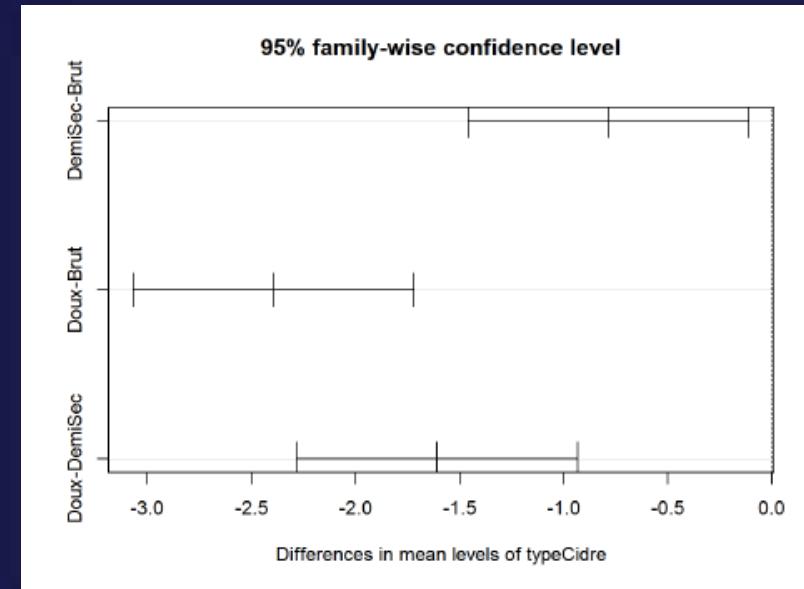
COMPARAISON DES COEFFICIENTS 2 À 2

- Test global : il existe au moins une différence significative entre les moyennes des I groupes (associés aux I modalités du facteurs)
- Pas d'infos sur les paires pour lesquelles ces différences de moyennes sont significatives
 - Tests post-hoc : Bonferroni, LSD de Fisher, Tukey
 - Comparaison par paires
 - Prise en compte de la multiplicité

```
Tukey multiple comparisons of means
95% family-wise confidence level

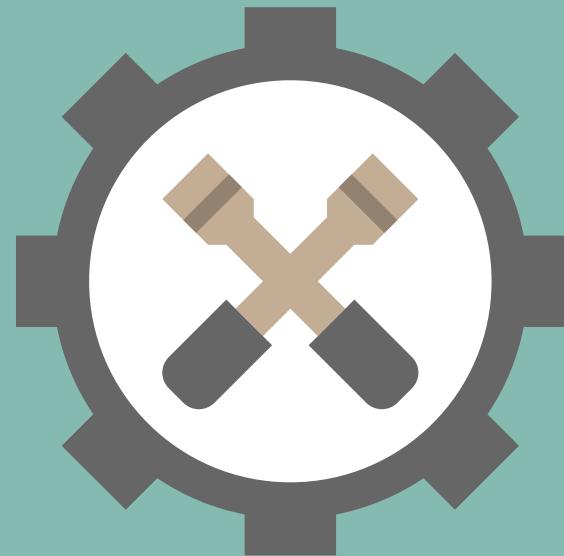
Fit: aov(formula = qteAlcool ~ typeCidre, data = donnees)

typeCidre
      diff      lwr      upr   p adj
DemiSec-Brut -0.78500 -1.459256 -0.110744 0.0207321
Doux-Brut     -2.39375 -3.068006 -1.719494 0.0000000
Doux-DemiSec -1.60875 -2.283006 -0.934494 0.0000165
```



EXERCICE

1. Prédire le nombre de coup tenté en fonction de la catégorie de poids. Que concluez-vous ?
2. Prédire le nombre de coup à la tête reçu en fonction de la catégorie de poids. Que concluez-vous ?



ANOVA

- I. Modèle à 1 facteurs

- II. Modèle à 2 facteurs
 - I. Introduction
 - II. Sans interaction
 - III. Avec interaction
 - IV. Modèle hiérarchique



CAS D'USAGE

- On a vu comment comparer les populations d'un même facteur.
- Supposons maintenant qu'un expérimentateur souhaite comparer l'influence de trois régimes alimentaires et de deux exploitations sur la production laitière. Les résultats expérimentaux sont dans le tableau suivant :

R. alimentaire Exploitation \	A	B	C	Total	Moyenne
1	7	36	2	45	15
2	13	44	18	75	25
Total	20	80	20	120	
Moyenne	10	40	10		20

NOTATION

- Nous possédons 2 facteurs F1 et F2
 - F1 avec p niveaux
 - F1 avec q niveaux
- Pour chaque couple (i,j) de niveaux, nous avons r observations de la variable Y

F1 \ F2	1	i	p
1	y_{111}, \dots, y_{11r}	y_{i11}, \dots, y_{i1r}	y_{p11}, \dots, y_{p1r}
j	y_{1j1}, \dots, y_{1jr}	y_{ij1}, \dots, y_{ijr}	y_{pj1}, \dots, y_{pjr}
q	y_{1q1}, \dots, y_{1qr}	y_{iq1}, \dots, y_{iqr}	y_{pq1}, \dots, y_{pqr}

- Dans la cellule (i,j) : les valeurs y_{ijk}
 - i : niveau du facteur f1
 - j : niveau du facteur f2
 - k : la k-ième répétition pour un couple (i,j)

NOTATION

$$\left\{ \begin{array}{l} y_{ij.} = \sum_{k=1}^r y_{ijk} \\ y_{i..} = \sum_{j=1}^q \sum_{k=1}^r y_{ijk} \\ y_{.j.} = \sum_{i=1}^p \sum_{k=1}^r y_{ijk} \\ y_{...} = \sum_{i=1}^p \sum_{j=1}^q \sum_{k=1}^r y_{ijk} \end{array} \right. \quad \left\{ \begin{array}{l} \overline{y_{ij.}} = \frac{1}{r} y_{ij.} \\ \overline{y_{i..}} = \frac{1}{qr} y_{i..} \\ \overline{y_{.j.}} = \frac{1}{pr} y_{.j.} \\ \overline{y_{...}} = \frac{1}{pqr} y_{...} \end{array} \right.$$

ANOVA

- I. Modèle à 1 facteurs
- II. Modèle à 2 facteurs
 - I. Introduction
 - II. Sans interaction
 - I. Estimation des paramètres
 - II. Test
 - III. Avec interaction
 - IV. Modèle hiérarchique



MODÈLE SANS INTERACTION

- Le modèle le plus simpliste est d'additionner les effets du facteur F1 avec les effets du facteur F2 :

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

- Ou :
 - μ est l'effet moyen
 - α_i est l'effet dû au niveau i du facteur F1
 - β_j est l'effet dû au niveau j du facteur F2

ESTIMATION DES PARAMÈTRES

- On cherche donc les valeurs de \hat{y}_{ij} qui minimise la fonction

$$\sum_{i=1}^p \sum_{j=1}^q \varepsilon_{ij}^2 = \sum_{i=1}^p \sum_{j=1}^q (Y_{ij} - \widehat{Y}_{ij})^2 = \sum_{i=1}^p \sum_{j=1}^q (Y_{ij} - \mu - \alpha_i - \beta_j)^2$$

- On utilise la même technique que pour l'analyse de variance à 1 facteur et on obtient

$$\widehat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot} \quad \widehat{\beta}_j = \bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot} \quad \widehat{\mu} = \bar{Y}_{\cdot\cdot}$$

- Donc la valeur prédite pour Y_{ij}

$$\widehat{y}_{ij} = \mu + \alpha_i + \beta_j = \bar{Y}_{i\cdot} + \bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot}$$

CAS D'USAGE

- F1 : le régime alimentaire, prend 3 valeurs (A, B, C), donc p = 3.
- F2 : l'exploitation, prend 2 valeurs (1 et 2), donc q = 2.
- Modèle statistique :

$$\hat{y}_{ij} = \mu + \alpha_i + \beta_j \quad \text{avec } i = 1,2,3 \text{ et } j=1,2$$

- Ou :
 - μ est l'effet moyen
 - α_i est l'effet de l'exploitation
 - β_j est l'effet du régime alimentaire
 - \hat{y}_{ij} l'estimation de la production laitière

ESTIMATION

- $\hat{\mu} = \bar{y}_{..} = 20$
- $\hat{\alpha}_1 = \bar{y}_{1.} - \bar{y}_{..} = 10 - 20 = -10$
- $\hat{\alpha}_2 = 20$
- $\hat{\alpha}_2 = -10$
- $\hat{\beta}_1 = \bar{y}_{.1} - \bar{Y}_{..} = 15 - 20 = -5$
- $\hat{\beta}_2 = 5$

R. alimentaire Exploitation	A	B	C	Total	Moyenne
1	7	36	2	45	15
2	13	44	18	75	25
Total	20	80	20	120	
Moyenne	10	40	10		20

- D'où la prévision de \hat{y}_{11} :
- $\hat{y}_{11} = \hat{\mu} + \hat{\alpha}_1 + \hat{\beta}_1 = 20 - 10 - 5 = 5$

TABLEAU D'ANALYSE DE LA VARIANCE

➤ En partant de l'identité

$$Y_{ij} - \bar{Y}_{...} = (Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} + \bar{Y}_{..}) + (\bar{Y}_{i\cdot} - \bar{Y}_{..}) + (\bar{Y}_{j\cdot} - \bar{Y}_{..})$$

➤ On obtient

$$\sum_{i,j} (Y_{ij} - \bar{Y}_{...})^2 = \sum_{i,j} (Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{j\cdot} + \bar{Y}_{..})^2 + q \sum_{i=1}^p (\bar{Y}_{i\cdot} - \bar{Y}_{..})^2 + p \sum_{j=1}^q (\bar{Y}_{j\cdot} - \bar{Y}_{..})^2$$

$$SCT = SCR + SC_{F1} + SC_{F2}$$

Source de variation	Ddl	SC	CM
F1	$(p - 1)$	SC_{F1}	$SC_{F1}/(P - 1)$
F2	$(q - 1)$	SC_{F2}	$SC_{F2}/(q - 1)$
Résidu	$(p - 1)(q - 1)$	SCR	$SCR/(p - 1)(q - 1)$
Totale	$pq - 1$	SCT	

ANOVA

I. Modèle à 1 facteurs

II. Modèle à 2 facteurs

I. Introduction

II. Sans interaction

I. Estimation des paramètres

II. Test

III. Avec interaction

IV. Modèle hiérarchique



TEST GLOBALE

- Le modèle n'est pas significatif si aucun des deux facteurs n'influence Y :

$$H_0: \alpha_1 = \dots = \alpha_p = \beta_1 = \dots = \beta_p = 0$$

- Contre

$$H_1: \exists i \in \{1, \dots, p\} \text{ et } \exists j \in \{1, \dots, q\} \text{ t. q. } \alpha_i \neq 0 \text{ ou } \beta_j \neq 0$$

- Alors le modèle est réduit à

$$Y_{ij} = \mu + \varepsilon_{ij}$$

- Statistique de test

$$Z = \frac{(SC_{F1} + SC_{F2})/(p + q - 2)}{SC_R/(p - 1)(q - 1)} \sim F(p + q - 2, (p - 1)(q - 1)) \text{ sous } H_0$$

TEST D'UN FACTEUR

- Supposons que l'on veut tester l'effet de F1

$H_0 : F1 \text{ n'influence pas } Y \text{ sachant que } F2 \text{ est dans le modèle}$

$$H_0: \alpha_1 = \dots = \alpha_p = 0$$

- Contre

$$H_1: \exists i \in \{1, \dots, p\} \text{ t.q. } \alpha_i \neq 0$$

- Alors le modèle est réduit à

$$Y_{ij} = \mu + \beta_j + \varepsilon_{ij}$$

- Statistique de test

$$Z = \frac{(SC_{F1})/(p-1)}{SC_R/(p-1)(q-1)} \sim F(p+q-2, (p-1)(q-1)) \text{ sous } H_0$$

EXEMPLE

Source de variation	Ddl	SC	CM
F1	2	1200	600
F2	1	150	150
Résidu	2	28	14
Totale	5	1378	

Significativité du modèle : $H_0: \alpha_1 = \dots = \alpha_p = \beta_1 = \dots = \beta_p = 0$

$$Z = \frac{(SC_{F1} + SC_{F2})/(p + q - 2)}{SC_R/(p - 1)(q - 1)} = \frac{(1200 + 150)/(3 + 2 - 2)}{28/2} = 32,1 \sim F(3,2)$$

avec $F(3,2) = 19,2$

H_0 rejeté : Le modèle est significatif

EXEMPLE

Source de variation	Ddl	SC	CM
F1	3	1200	600
F2	1	150	150
Résidu	2	28	14
Totale	5	1378	

Significativité du régime alimentaire : $H_0: \alpha_1 = \alpha_2 = \alpha_3 = 0$

$$Z = \frac{(SC_{F1})/(p - 1)}{SC_R/(p - 1)(q - 1)} = \frac{(1200)/(3 - 1)}{28/2} = 32,1 \sim F(2,2)$$

avec $F(2,2) = 19$

pour la production laitière. H_0 rejeté : le régime alimentai

ANOVA

- I. Modèle à 1 facteurs

- II. Modèle à 2 facteurs**
 - I. Introduction
 - II. Sans interaction
 - III. Avec interaction**
 - IV. Modèle hiérarchique



MODÈLE AVEC INTÉRACTION

- Le modèle le plus simpliste est d'ajouter les effets du facteur F1 avec les effets du facteur F2 :

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij}$$

- Lien :

$$\hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{Y}_{..} \quad \hat{\beta}_j = \bar{Y}_{J\cdot} - \bar{Y}_{..} \quad \gamma_{ij} = Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{J\cdot} + \bar{Y}_{..} \quad \hat{\mu} = \bar{Y}_{..}$$

- Les hypothèses à tester peuvent être :

$$H_0: \alpha_1 = \dots = \alpha_p = 0$$

$$H'_0: \beta_1 = \dots = \beta_q = 0$$

$$H''_0: \gamma_{11} = \dots = \gamma_{pq} = 0$$

MODÈLE AVEC INTÉRACTION

- On calcule les sommes des carrées des écarts :

$$➤ SC_{F1} = q \sum_{i=1}^p (\bar{Y}_{i..} - \bar{Y}_{...})^2 \text{ avec } ddl_{F1} = (p - 1)$$

$$➤ SC_{F2} = p \sum_{j=1}^q (\bar{Y}_{.j} - \bar{Y}_{...})^2 \text{ avec } ddl_{F2} = (q - 1)$$

$$➤ SC_{F12} = SC_T - SC_{F1} - SC_{F2} - SC_R \text{ avec } ddl_{F12} = (p - 1)(q - 1)$$

$$➤ SC_R = \sum_{i,j} (Y_{ij} - \bar{Y}_{i..} - \bar{Y}_{.j} + \bar{Y}_{...})^2 \text{ avec } ddl_R = pq(r - 1)$$

$$➤ SC_T = \sum_{i,j} Y_{ij} - \bar{Y}_{...} \text{ avec } ddl_{F1} = (rpq - 1)$$

TEST INTERACTION

- On commence par tester l'interaction sous H_0

$$f_{F12} = \frac{(SC_{F12})/(p-1)(q-1)}{SC_R/(pq(r-1))} \sim F((p-1)(q-1), pq(r-1))$$

- On rejette H_0 si f_{F12} dépasse le fractile d'ordre $(1 - \alpha)$ de la loi de Fisher
- Si on accepte H''_0 (pas d'interaction), on teste l'influence de F1 puis de F2
- Si on rejette H''_0 alors :
 - Test H_0 et H'_0 et retour au modèle sans intéraction

ANOVA

- I. Modèle à 1 facteurs

- II. Modèle à 2 facteurs**
 - I. Introduction
 - II. Sans interaction
 - III. Avec interaction

- IV. Modèle hiérarchique**



EXEMPLE

- On sélectionne plusieurs régions (1er facteur), puis, a l'intérieur de chacune des régions, plusieurs exploitations agricoles (2eme facteur), et on mesure la quantité de lait produite annuellement par r vaches dans chacune des exploitations
- Modèle hiérarchique : aucune raison d'avoir un lien entre les exploitations n1 de chacun des régions
 - $\bar{Y}_{i\cdot}$: moyenne des exploitations de la région i : intéressant
 - $\bar{Y}_{\cdot j}$: moyenne des j -ièmes exploitations de chaque région : non pertinent

MODÈLE

► Modèle :

$$Y_{ijk} = \mu + \alpha_i + \beta_{j|i} + \varepsilon_{ij}$$

► Ou :

- μ : production moyenne
- α_i : apport de la région
- $\beta_{j|i}$: apport de l'exploitation j dans la région i

TABLEAU D'ANOVA

➤ En partant de l'identité

$$Y_{ijk} - \bar{Y}_{...} = (\bar{Y}_{i..} - \bar{Y}_{...}) + (\bar{Y}_{ij.} - \bar{Y}_{i..}) + (Y_{ijk} - \bar{Y}_{ij.})$$

➤ On obtient

$$\sum_{i,j,k} (Y_{ijk} - \bar{Y}_{...})^2 = qr \sum_i (\bar{Y}_{i..} - \bar{Y}_{...})^2 + r \sum_{jk} (\bar{Y}_{ij.} - \bar{Y}_{i..})^2 + \sum_{ijk} (Y_{ijk} - \bar{Y}_{ij.})^2$$

$$SCT = SC_{F1} + SC_{F2|F1} + SCR$$

Source de variation	Ddl	SC	CM
F1	P-1	SC_{F1}	$SC_{F1}/(p - 1)$
F2	$p(q + 1)$	$SC_{F2 F1}$	$SC_{F2 F1}/(p(q - 1))$
Résidu	$pq(r + 1)$	SCR	$SCR/(pq(r - 1))$
Totaux	$pqr - 1$	SCT	

TEST INTERACTION

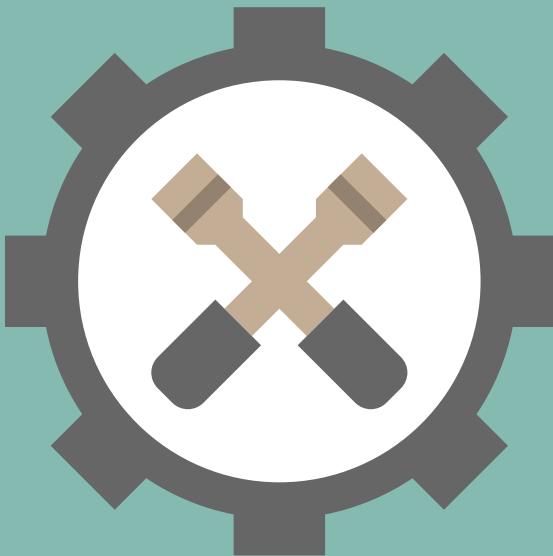
- ➊ Test du premier facteur : $H_0 : a_1 = \dots = a_p = 0$

$$f_{F1} = \frac{(SC_{F1})/(p - 1)}{SC_{F2|F1}/(p(q - 1))} \sim F((p - 1), p(q - 1))$$

- ➋ Test du deuxième facteur : $H_0 : \beta_{j|i} = 0$

$$f_{F2|F1} = \frac{(SC_{F2|F1})/p(q - 1)}{SCR/(pq(r - 1))} \sim F(p(q - 1), pq(r - 1))$$

EXERCICE



1. Prédire le nombre de coup tenté en fonction :
 1. De la catégorie de poids
 2. Du style de combat

Tester les modèles :

1. Sans interaction
2. Avec interaction
3. Hiérarchique (style dépend du poids)

SOMMAIRE

- Pourquoi la régression ?
- Régression linaire
- ANOVA
- ANCOVA
- Sélection de variables
- Régression logistique
- Autres méthodes de prédiction



INTRODUCTION

- L'analyse de covariance (ANCOVA) se situe dans le cadre général du modèle linéaire. Elle peut être vue comme un mélange d'ANOVA et de modèle linéaire.
- Elle permet :
 - d'expliquer une variable quantitative Y
 - par plusieurs variables explicatives de type à la fois quantitatives et qualitatives.
- L'objectif sera de tenir compte, lors de l'étude :
 - des effets des facteurs sur la variable Y ,
 - des effets possibles de la ou des variables quantitatives

ANCOVA A 1 FACTEUR ET 1 COVARIABLE

- Le modèle est explicite dans le cas où une variable quantitative Y est expliquée par
 - Un facteur F à p niveaux
 - Une variable quantitative X, appelé covariable
- Pour chaque niveau de $i = 1, \dots, p$ de F , on observe
 - n_i mesures de X notées x_{ij}
 - n_i mesures de Y notées y_{ij}
- On notera n la taille de l'échantillon : $n = \sum_{i=1}^p n_i$

CAS D'USAGE

- On cherche à savoir si des conditions de température et d'oxygénation influencent l'évolution du poids des huîtres.
- On dispose de $n = 20$ paniers de 10 huîtres.
- On place pendant un mois ces 20 paniers de façon aléatoire dans $p = 5$ emplacements différents d'un canal de refroidissement d'une centrale électrique à raison de $r = 4$ paniers par emplacement.
- Ces emplacements se différencient par leurs températures et oxygénations.
- Pour chaque sac, on dispose de
 - son poids avant l'expérience (variable Pds Init),
 - son poids après l'expérience (variable Pds Final)
 - son emplacement (variable Traitement), codé de 1 à 5.

CAS D'USAGE

Obs	Traitement	Repetition	Pds Init	Pds final
1	1	1	27.2	32.6
2	1	2	32.0	36.6
3	1	3	33.0	37.7
4	1	4	26.8	31.0
5	2	1	28.6	33.8
6	2	2	26.8	31.7
7	2	3	26.5	30.7
8	2	4	26.8	30.4
9	3	1	28.6	35.2
10	3	2	22.4	29.1
11	3	3	23.2	28.9
12	3	4	24.4	30.2
13	4	1	29.3	35.0
14	4	2	21.8	27.0
15	4	3	30.3	36.4
16	4	4	24.3	30.5
17	5	1	20.4	24.6
18	5	2	19.6	23.4
19	5	3	25.1	30.3
20	5	4	18.1	21.8

- **Objectif** : on cherche à expliquer la variable Pds Final (variable quantitative) à partir :
 - d'une variable quantitative Pds Init
 - d'une variable qualitative Traitement.
- et on veut savoir si l'évolution du poids des huitres est différente selon le traitement c'est à dire l'emplacement.

MODÈLE

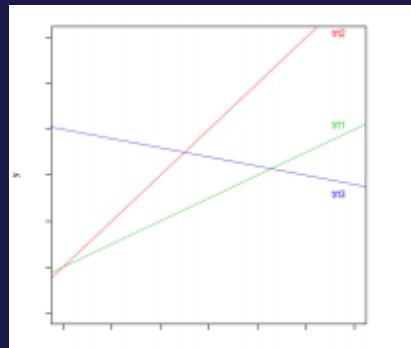
- Le modèle de covariance s'écrit

$$Y_{ij} = \mu + \alpha_i + \beta_i x_{ij} + \varepsilon_{ij}$$

- Ou :
 - i est l'indice du traitement (numéro de l'emplacement)
 - j , l'indice de répétition, est le numéro du sac d'huître pour son emplacement
 - Y_{ij} le poids aléatoire final du j ième sac d'huître de l'emplacement i et
 - x_{ij} est la valeur du poids initial
 - μ est la constante
 - α_i est l'effet du traitement i
 - β_i est la pente de régression pour l'emplacement i

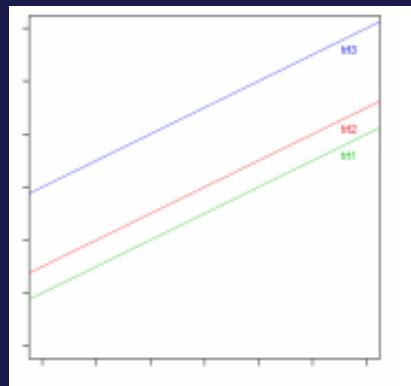
HYPOTHÈSES

Modèle 1 :
une droite par traitement



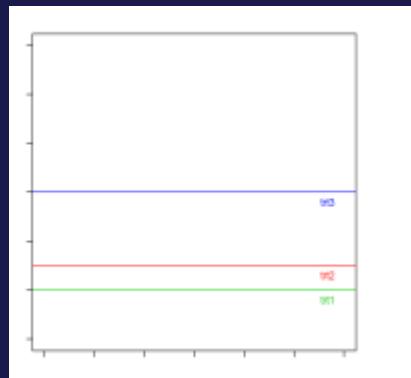
$$y_{ij} = \mu + \alpha_i + \beta_i x_{ij} + \varepsilon_{ij}$$

Modèle 2 :
une droite par traitement,
mais avec pente
commune



$$y_{ij} = \mu + \alpha_i + \beta x_{ij} + \varepsilon_{ij}$$

Modèle 3 :
ANOVA standard, sans
covariable



$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

PREMIÈRE HYPOTHÈSE : TEST SUR LA COVARIABLE

- $H_0 : \beta_1 = \dots = \beta_p$
- $H_1 : \exists j, i \text{ tq } \beta_j \neq \beta_i$
- On compare en fait les modèles 1 (une droite par traitement) et 3 (absence de covariable). Par le test général de la régression, la statistique du test est la suivante :

$$F_0 = \frac{(SCR_2 - SCR_1)/I}{SCR_1/(N - 2I)} \sim F(I, N - 2I)$$

- Si H_0 est rejeté, alors on teste la prochaine hypothèse d'égalité des pentes
- Sinon on ajuste le modèle 3, l'anova à 1 facteur standard

DEUXIÈME HYPOTHÈSE : TEST SUR LES PENTES

- $H_0 : \beta = 0$
- $H_1 : \beta \neq 0$
- On compare ici les modèles 1 (une droite par traitement) et 2 (pentes égales).
- La statistique du test est la suivante :

$$F_0 = \frac{(SCR_3 - SCR_2)/1}{SCR_2^2/(N - I - 1)} \sim F(I, N - I - 1)$$

- Si H_0 est rejetée, le modèle 1 ne peut pas être simplifié. Voir l'hypothèse sur les traitements et la comparaison des moyennes ajustées ci-dessous.
- Si H_0 n'est pas rejetée, on ajuste le modèle 2.

TROISIÈME HYPOTHÈSE : TEST SUR LES TRAITEMENTS

- $H_0 : \mu_1 = \dots = \mu_p$
- $H_1 : \exists(i,j) \text{ tq } \mu_i \neq \mu_j$
- Si le modèle 1 est retenu, la différence entre les valeurs prédictes de Y d'un traitement à l'autre ne sera pas la même pour toutes les valeurs de X.
- La statistique du test est la suivante :x

$$F_0 = \frac{(SCR_4 - SCR_3)/(p - 1)}{SCR_4/(N - p - 1)} \sim F((p - 1), N - p - 1)$$

- Si H_0 est rejetée, c'est que les valeurs des ordonnées à l'origine sont différentes pour au moins deux traitements
- Si H_0 n'est pas rejetée, c'est que les valeurs moyennes des ordonnées à l'origine ne diffèrent pas significativement. il

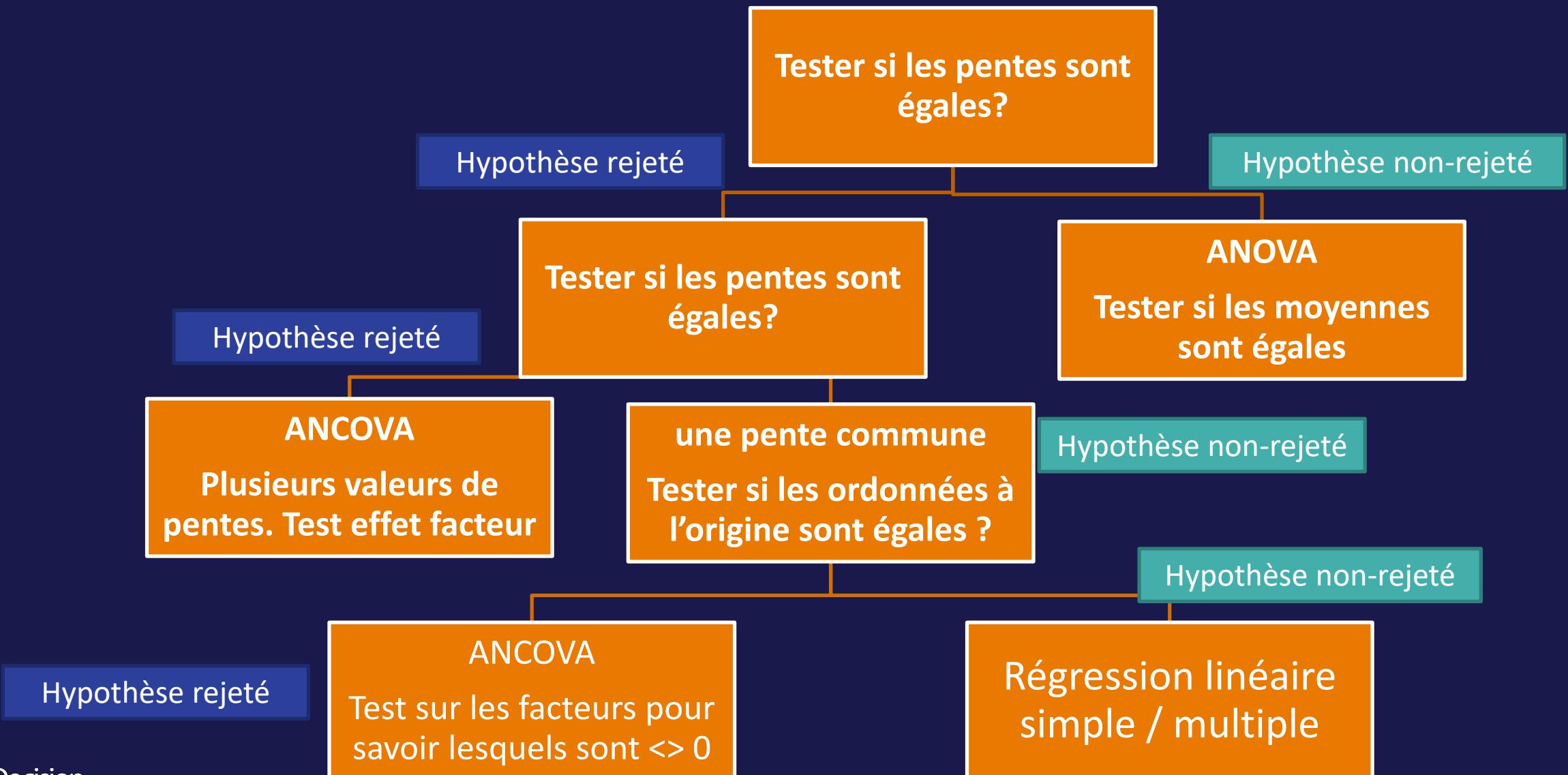
ESTIMATION

- Le modèle de covariance s'écrit

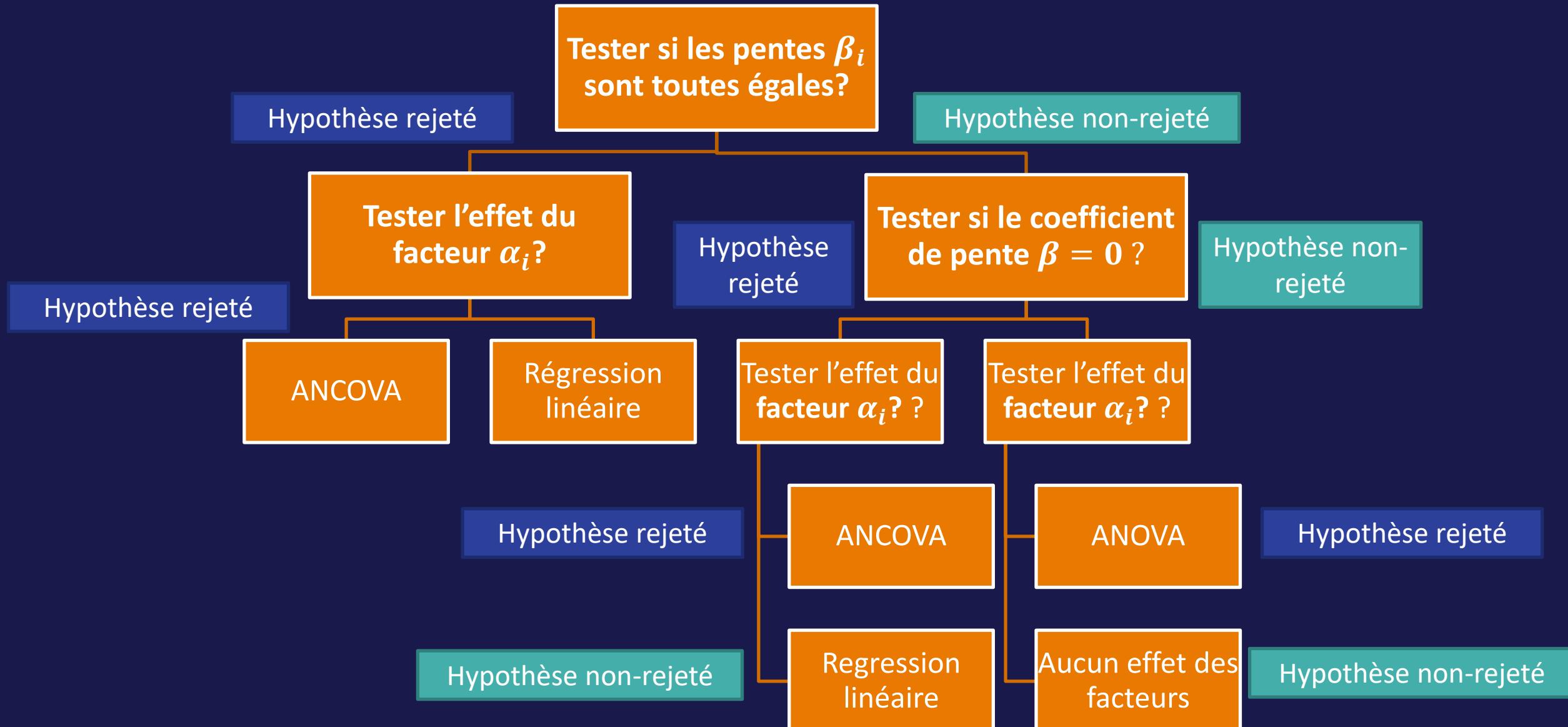
$$Y_{ij} = \mu + \alpha_i + \beta_i x_{ij} + \varepsilon_{ij}$$

- Avec :
 - $\hat{\mu} = \bar{Y}_{..}$
 - $\hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{..}$
 - $\hat{\beta}_i = \frac{\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i..})(y_{ij} - \bar{y}_{i..})}{\sum_{i=1}^p \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i..})^2}$
 - $\hat{\beta} = \frac{\sum_{i=1}^I \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i..})(y_{ij} - \bar{y}_{i..})}{\sum_{i=1}^p \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i..})^2}$
 - On testera :
- | | |
|---|---|
| • $H_0 : \alpha_1 = \dots = \alpha_p = 0$ | • $H_1 : \exists i \in \{1, \dots, p\}, tq \alpha_i \neq 0$ |
| • $H'_0 : \beta_i = \text{constante}$ | • $H'_1 : \beta_i \neq \text{constante}$ |
| • $H''_0 : \beta_i = \beta = 0$ | • $H''_1 : \beta_i \neq 0$ |

PARCOURS ANCOVA



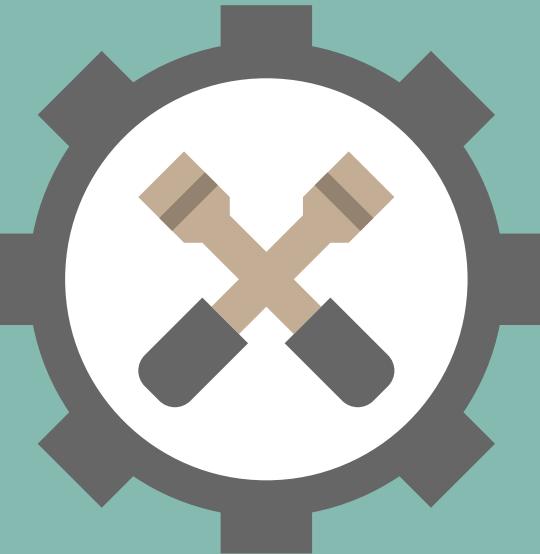
PARCOURS ANCOVA



TEST ET INTERACTION

- L'interaction entre la variable qualitative et quantitative peut également être testé
 - Même principe que pour l'ANOVA à 2 facteurs
- Pour chaque coefficient on obtient donc :
 - Test de H_0 (*nullité des coefficients*) VS H_1
 - Une statistique de test dépendant de Fisher / Student
 - Un test de significativité de X sur Y
 - Rejet / validation de H_0

EXERCICE



1. Prédire le nombre de victoire / nombre de match en fonction :
 1. De la catégorie de poids
 2. Du nombre de coup total
2. Calculer un modèle de régression linéaire simple par catégorie de poids entre pour prédire le nombre de victoire / nombre de match en fonction :
 1. Du nombre de coup total
3. Quel solution vous semble la meilleure ?

SOMMAIRE

- Pourquoi la régression ?
- Régression linaire
- ANOVA
- ANCOVA
- Sélection de variable
- Régression logistique
- Autres méthodes de prédiction

SÉLECTION DU MEILLEUR MODÈLE

- Travail long et répétitif
- Méthode de sélection de modèle en fonction
 - R^2 adj
 - AIC et AIC corrigé
 - BIC
 - Cp de Mallows
- Plusieurs méthodes
 - Méthode descendante
 - Méthode ascendante
 - Méthode Stepwise
 - Recherche exhaustive

PROCÉDURE DESCENDANTE

1. Calcul du *Modèle complet* : $Y_{ih} = \mu + \alpha_i x_{ij} + \dots + \varepsilon_{ij}$
2. Effectuer un test de Student pour chacune des variables explicatives. Deux cas se présentent :
 1. Les variables sont trouvées significatives. Ce modèle est alors choisi. Nous arrêtons là notre analyse.
 2. Éliminer la variable la moins significative du modèle.
3. Recommencer le processus avec une variable en moins



- Inclus toutes les variables
- Economique en temps et interprétation



- Modèle avec beaucoup de variable
- Une fois la variable supprimé, pas de réinsertion possible

PROCÉDURE ASCENDANTE

1. Calcul du *modèle simple* : $Y_{ih} = \mu + \varepsilon_{ij}$
2. Effectuer les k régression possibles avec 1 seule variables. Pour chacune d'elles :
 1. Effectuer le test de Student pour la nouvelle variable. Retenir le modèle pour lequel la variable est la plus significative.
 2. Si aucune variable est retenue, alors nous arrêtons le processus.
3. Effectuer les k-1 régression possibles avec les variables restantes
4. Le processus se termine lorsque nous ne pouvons plus introduire des variables significatives dans le modèle.



- évite de travailler avec plus de variables que nécessaire
- améliore l'équation à chaque étape



- une variable introduite dans le modèle ne peut plus être éliminée

PROCÉDURE STEPWISE

- Amélioration de la méthode ascendante
- Nous réexaminons toutes les variables introduites précédemment dans le modèle.
 - Une variable considérée comme la plus significative à une étape de l'algorithme peut à une étape ultérieure devenir non significative
 - En raison de ces corrélations avec d'autres variables introduites après coup dans le modèle.

PROCÉDURE STEPWISE

1. Calcul du *modèle simple* : $Y_{ih} = \mu + \varepsilon_{ij}$
2. Effectuer les k régression possibles avec 1 seule variables. Pour chacune d'elles :
 1. Effectuer le test de Student pour la nouvelle variable. Retenir le modèle pour lequel la variable est la plus significative.
 2. Si aucune variable est retenue, alors nous arrêtons le processus.
3. Effectuer les $k-1$ régression possibles avec les variables restantes
 1. réexaminer les tests de Student pour chaque variable explicative autrement admise dans le modèle
 2. après réexamen, si des variables ne sont plus significatives, alors retirer du modèle la moins significative d'entre elles.
4. Le processus continue jusqu'à ce que plus aucune variable ne puisse être introduite ni retirée du modèle.

PROCÉDURE STEPWISE



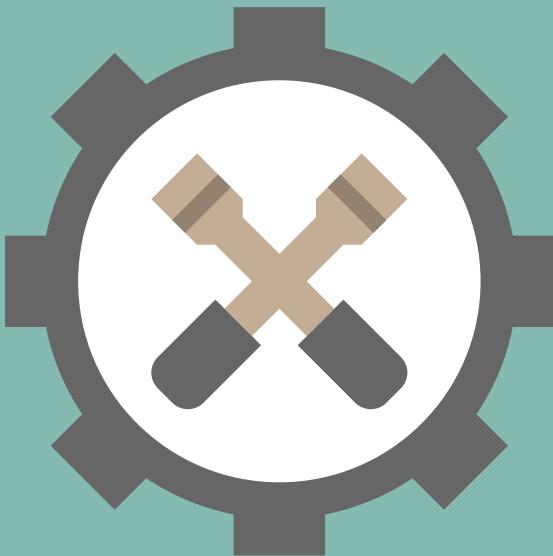
- Réévaluation des variables candidates à chaque itération
- améliore l'équation à chaque étape



- Attention au Multi colinéarité entre les variables (X_1 et X_2 fortement corrélé)

EXERCICE

1. Trouver le meilleur modèle pour prédire le nombre de victoire / nombre de match



REX

ESTIMATION D'UN BIEN IMMOBILIER

COMMENT ESTIMER UN BIEN IMMOBILIER ?

COMMENT ESTIMER UN BIEN ?



Prédiction
Environnement
Description
Bien
Localisation

- Prédiction du prix M²
- Prédiction du prix du terrain
- Proche commodité / Proche transport
- Quartier en développement / ...
- Neuf / Ancien
- Nombre de chambre / M² / Garage / ...
- Maison / Appartement / Terrain
- Château / Chalet / MobilHome
- Campagne / Ville / Montagne / ...
- Développement de la ville

PÉRIMÈTRE

COMMENT ESTIMER UN BIEN IMMOBILIER ?



Seloger



10 millions
d'annonces
immobilières

Répartition sur
toute la
France par
commune

Tout type de
bien
(Appartement,
maison ,
château ,
terrain, ...)

CRÉATION D'INDICATEUR



- ✓ Adresse
- ✓ Code postale
- ✓ Epoque de construction
- ✓ Présence de garage
- ✓ Nombre de niveaux
- ✓ Nombre de pièce
- ✓ Nombre de salle de bain
- ✓ Nombre de place de parking
- ✓ Surface habitable
- ✓ Surface totale du terrain
- ✓ Type de bien

Description Standard

Description
Additionnel

- ✓ Sous-sol
- ✓ Terrasse
- ✓ Véranda
- ✓ Dépendance
- ✓ Court de tennis
- ✓ Piscine
- ✓ Parcelle divisible
- ✓ Présence de comble



- ✓ Calme
- ✓ Luminosité
- ✓ Proximité des transport
- ✓ Standing du bien
- ✓ Vue exceptionnelle
- ✓ Orientation salon
- ✓ Possibilité de stationnement
- ✓ Mur mitoyen

Qualité
de vie

Rénovation

- ✓ Bien a rénover ? Oui/Non
- ✓ Etat chauffage
- ✓ Etat façade
- ✓ Etat plomberie
- ✓ Etat réseau électrique
- ✓ Etat isolation
- ✓ Qualité de la maison
- ✓ Qualité de la toiture
- ✓ Décoration à revoir
- ✓ Type de travaux à prévoir (électricité / isolation / plomberie / fenêtre / sol / ...)
- ✓ Type de rafraîchissement fraîchement à prévoir (Peinture / rénovation salle de bain / cuisine / ...)



MODÈLE GLOBAL

$R^2 = 0,40$

Ecart = 68
274 €

Modalité	Coefficient
(Intercept)	15 572,39 €
SLOG_touriste_resid	6 816,93 €
SLOG_Campagne_ville	2 292,83 €
SLOG_New_vieille	- 4 851,00 €
SLOG_Campagne_periph	19 041,56 €
SLOG_riche_pauvre	- 45 749,13 €
Surface	5,66 €
SurfaceTerrain	0,15 €
NbChambres	20 576,05 €
NbSallesDeBain	10 717,79 €
Etage	12 192,32 €
NbBoxes	4 479,03 €
NbTerasses	16 210,70 €
TypeBien_Maison	24 633,27 €
NbPieces	7 869,58 €
Campagne active	13 405,69 €
Campagne Desertique	19 607,53 €
Chef de canton	58 296,18 €
Grande ville	70 128,81 €
Ville en developpement	46 924,96 €
Ville touristique	46 656,01 €
Mitoyennete	-24 319,26 €
Piscinable	78 024,32 €
Piscine	84 973,53 €
A renover	- 42 135,94 €
Spa	20 493,39 €

Prix logement =

15 572,39 €
 + 14,93 * 6 816,93 €
 +47,14 * 2 292,83 €
 + 30,23 * (-4 851) €
 + (-22,04) * 19 041,56 €
 + (-1,38) * (-45 749,13) €
 + 90 m² * 5,66 €
 + 0 m² * 0,15 €
 + 2 chambre * 20 576 €
 + 1 Salle de bain * 10 717,79 €
 + 0 Etage * 12 192,32 €
 + 0 Box * 4 479,03 €
 + 0 Terrasse * 16 210,70 €
 + 0 Appartement * 24 633,27 €
 + 3 pièces * 7 869,58 €
 + 1 Grande ville * 70 128,81 €
 + 1 Mitoyenne * (-24 319,26) €
 + 0 Piscine * 78 024,32 €
 + 0 Piscinable * 84 973,53 €
 + 0 Travaux * (-42 135,94) €
 + 0 Spa * 20 493,39 €

SEGMENTATION VILLE

Ville en développement

Nouvelle ville / Périphérique
Résidentielle / Nb population moyenne
Niveau de vie très élevée
Exemple : Basse-Goulaine



Moyenne ville / Chef de canton

Moyenne ville / Chef de canton
Nb population importante
% ouvrier & employé
% Ménage seul
Exemple : Locminé



Lieu touristique

Ville touristique avec une moyenne d'âge élevée
Exemple : La Trinité-sur-Mer



Campagne active

% Maison & Famille
Population faible
Niveau de vie correct
Exemple : Aigrefeuille-sur-Maine



Grande Ville

Nb Population / Menage / Logement très important
% Maison faible & % commerce élevé
Exemple : Nantes
Nb : seulement 85 villes



Campagne désertique

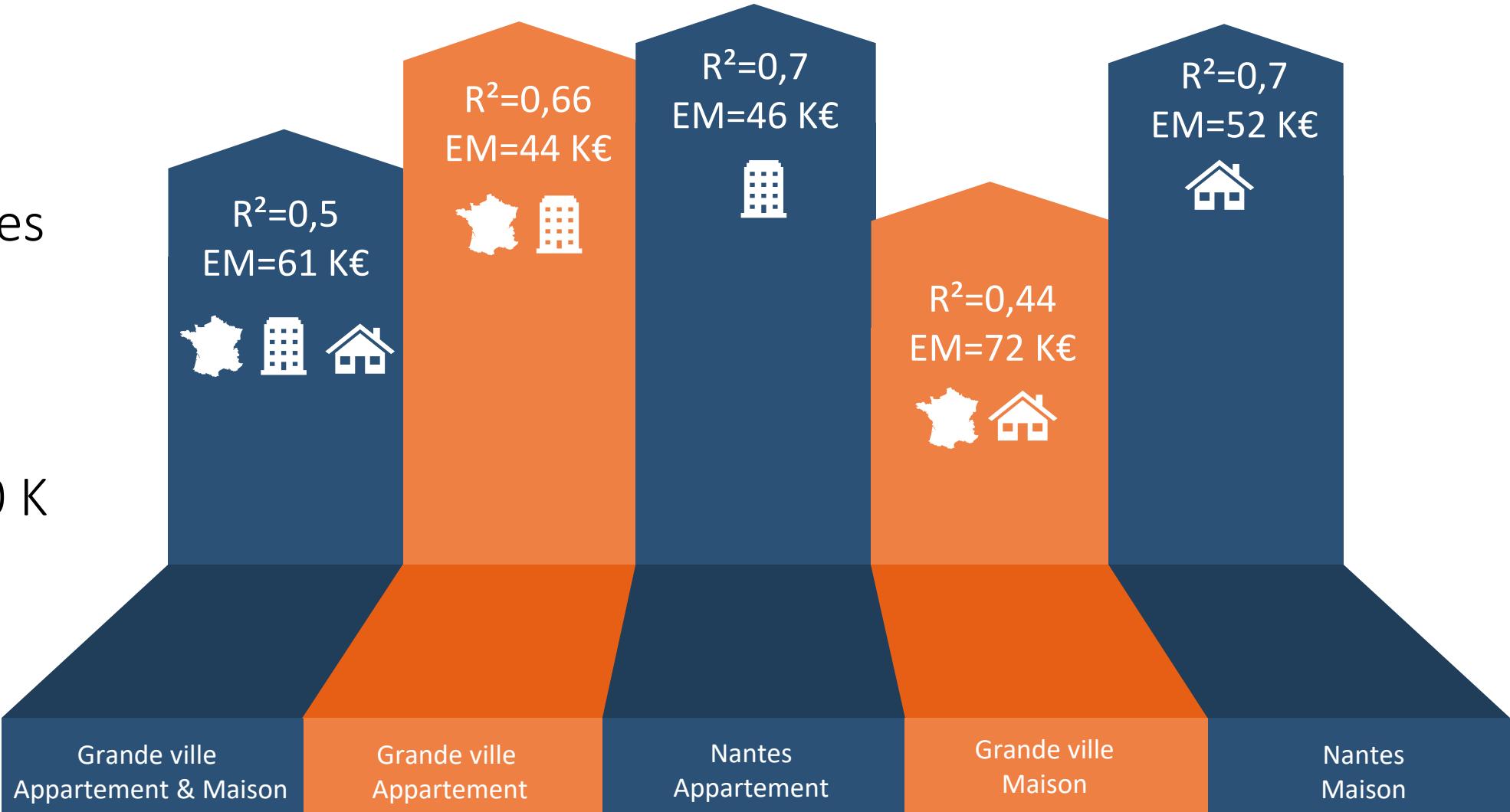
Peu d'activité & Retraité
Nb population très faible
Exemple : Le Temple-de-Bretagne

MODÈLE PAR TYPE DE BIEN

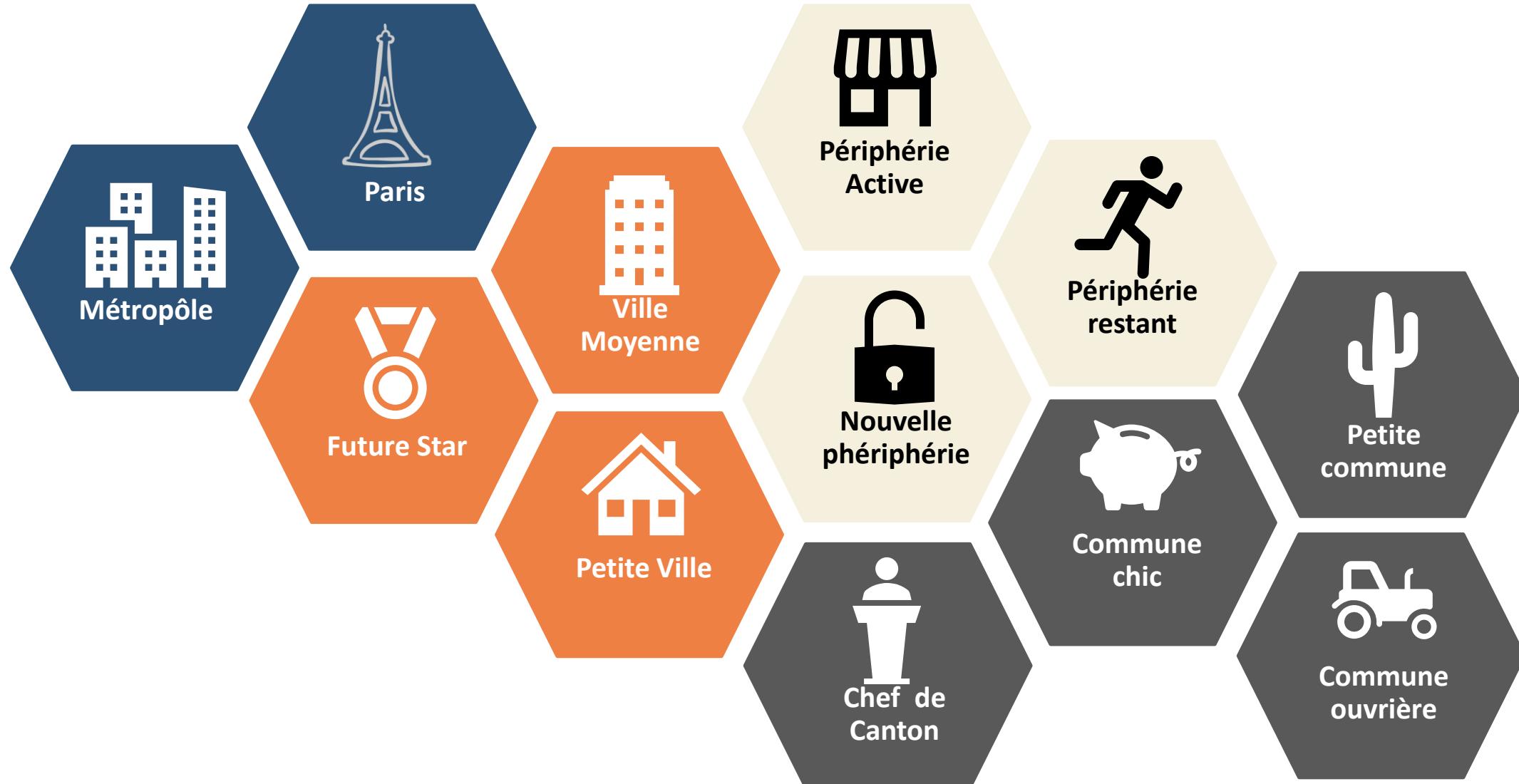


RÉSULTAT INTERMÉDIAIRE

- $4 * 6 = 24$ Modèles
- $R^2 = 0,5$
- Ecart moyen = 60 K



SEGMENTATION DÉTAILLÉ



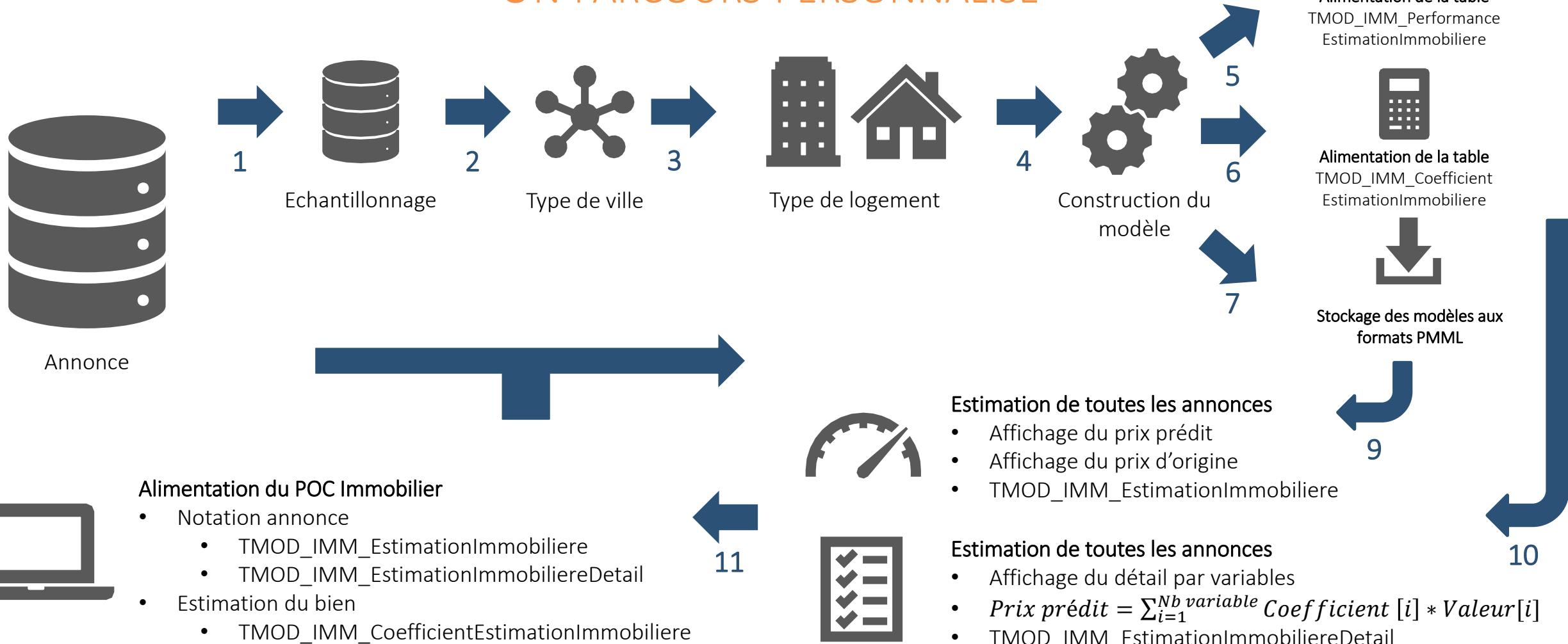
APPARTEMENT D'EXCEPTION

- EXCLUSIVITÉ - APPARTEMENT DE PRESTIGE - 4 CHAMBRES - BORDEAUX GAMBETTA. Au coeur de Bordeaux, bel appartement rénové par un architecte et une décoratrice de renom. Ce qui marque de prime abord c'est la qualité extrême des matériaux utilisés, les volumes et la hauteur sous plafond (4 mètres). L'entrée passée on retrouve sur la gauche une première chambre et sa salle d'eau, puis une magnifique cuisine ouverte mais qui peut être refermée par de grands panneaux de bois grâce à un système à galandage.

 Bordeaux	 4 chambres	 Prix de vente 2 940 000 €
 300 m ²	 1 Salle de bain 2 salle d'eau	 0 Piscine / Spa
 2 place de parking		 Prix Estimé 794 518 €

CONSTRUCTION DE LA CHAINE IMMOBILIÈRE

UN PARCOURS PERSONNALISÉ



RÉSULTAT

- Création de $20*4 = 80$ Modèles !
- Modèle stable d'un segment à l'autre
 - Environs 30 % des prévisions possèdent moins de 10 % d'erreurs
 - Environs 60% des prévisions possèdent moins de 20 % d'erreurs
- Si trop peu de données, il ne faut pas prendre en compte le modèle et renvoyer un modèle généraliste



FRONT

Estimation immobilière

Adresse

Annonce

5

Rechercher

Détails de l'annonce

Description : Champigny Tremblay, immeuble cirque, très faibles charges, appartement 2 pièces en étage, parquet, cheminée, entrée, cuisine, chambre, séjour, WC, salle d'eau, chauffage individuel au gaz, décoration à prévoir.

Titre : Appartement 2 pièces

Prix : 124,000 € FAI

Type de bien : Appartement

Code Postal : 94500

Ville : Champigny sur Marne

Surface habitable : 43.07

Surface du terrain :

Nombre de pièces : 2

Nombre de chambres : 1

131,125
euros

42
Variables

14
Variables positives

13
Variables négatives

➡ Surface habitable

- 40,467 €

Valeur : Moins de 46 m²

➡ Surface du séjour

+ 4,093 €

Valeur : Pas de séjour

1 Square Maurice Audin, 44340 Bouguenais

SOMMAIRE

- Pourquoi la régression ?
- Régression linaire
- ANOVA
- ANCOVA
- Sélection de variables
- **Régression logistique**
- Autres méthodes de prédition



RÉGRESSION LOGISTIQUE

- I. Introduction
- II. Variables explicatives qualitatives
- III. Variables explicatives quantitatives
- IV. Variables explicatives mixtes
- V. Sélection du modèle
- VI. Validation du modèle



Cas d'application

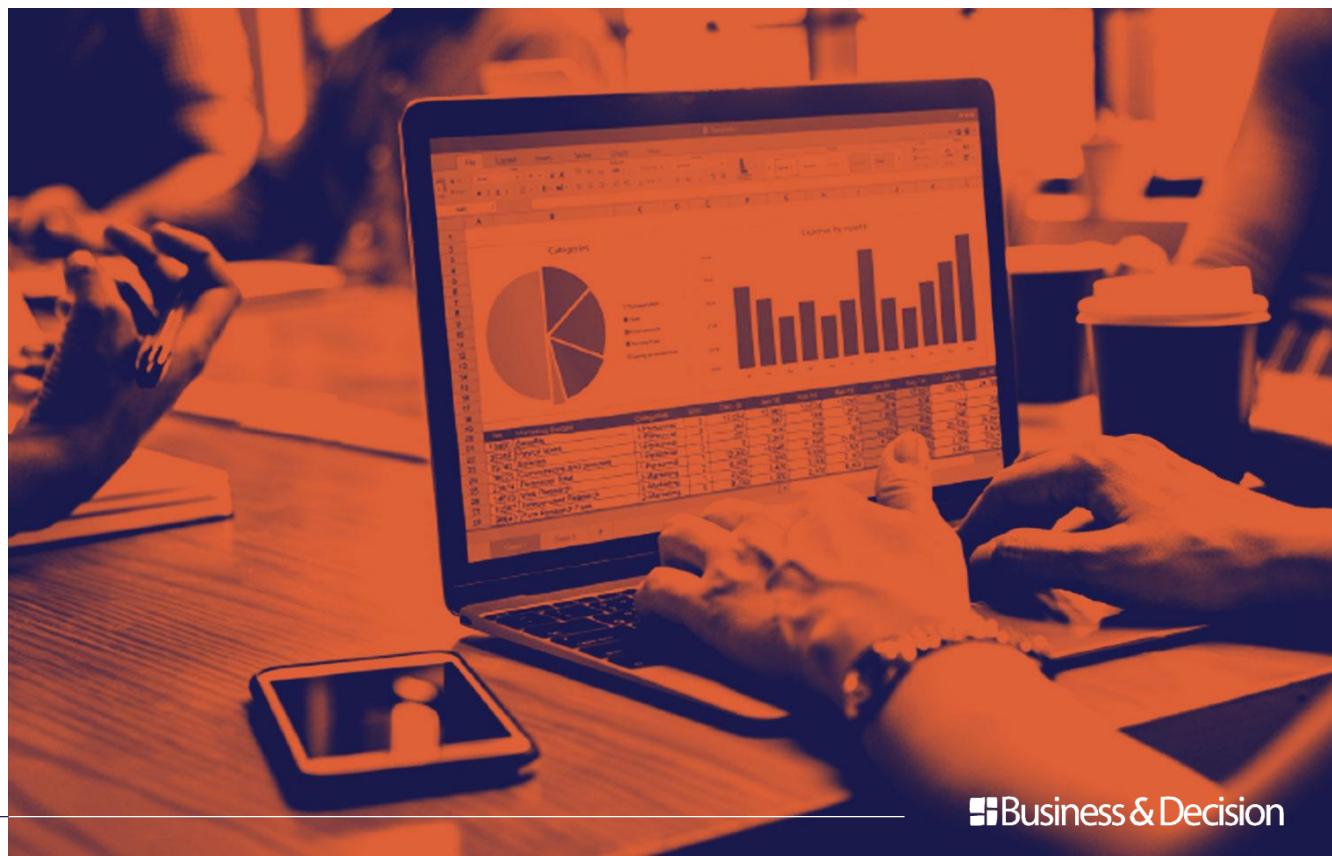
- La régression s'applique dans le cas où :
 - Y est qualitatif à 2 modalités
 - X qualitatives ou quantitatives
- Est appliqué à :
 - **Santé** : Identification des facteurs liées à une maladie
 - **Assurance** : Prédiction de l'attrition d'un client
 - **Banque** : Détection de fraude à la carte
 - **Retail** : Probabilité d'achat d'un produit
 - **Agroalimentaire** : Identification des événements de santé
 - **Industrie** : Identification des anomalies sur une chaîne de production
 -



- Y est une variable binaire
 - 0 en cas de non-occurrence de l'évènement
 - 1 si occurrence
 - Yi aléatoire et Xi non aléatoires
-
- On cherche à expliquer la survenue d'un évènement
-
- Nous allons avoir :
 - Nombre d'individu : n
 - Nombre de variable : k
 - Variable à prédire : $(y_1 \dots y_n)$

➤ Variables explicatives : $\begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}$

Contexte



EXEMPLE

- Nombre de souris développant une tumeur au poumon après l'exposition à la fumée de cigarettes

Groupe	Tumeur présente	Tumeur absente	Total
exposé	21	2	23
contrôle	19	13	32
Total	40	15	55

- Question : Existe-t-il une corrélation entre le développement de la maladie et l'apparition du cancer ?

TEST DU KHI-DEUX

Effectif observé

Groupe	Tumeur présente	Tumeur absente	Total
exposé	21	2	23
contrôle	19	13	32
Total	40	15	55

Effectif théorique

Groupe	Tumeur présente	Tumeur absente	Total
exposé	16,23	6,27	23
contrôle	23,27	8,73	32
Total	40	15	55

$$\textcircled{1} \quad \chi^2_{k-1,\alpha} = \sum \frac{(Observé-Théorique)^2}{Théorique} = 6,8781$$

$$\textcircled{1} \quad \chi^2_{k-1,\alpha} = 3,8414$$

➤ Rejet de H_0 / Lien entre les deux variables

➤ Problème :

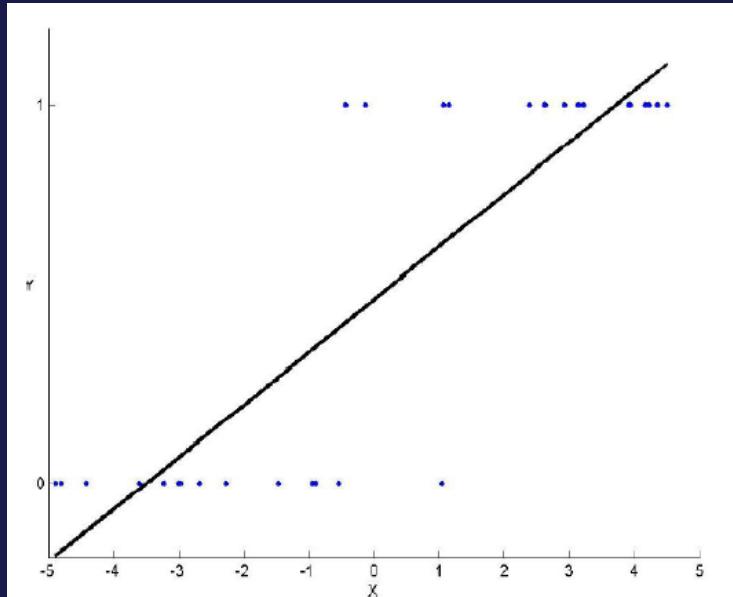
➤ Ce test ne permet pas de déterminer la nature de ce lien, c'est-à-dire comment sont liées les variations des deux variables

➤ Solution : **Régression logistique**

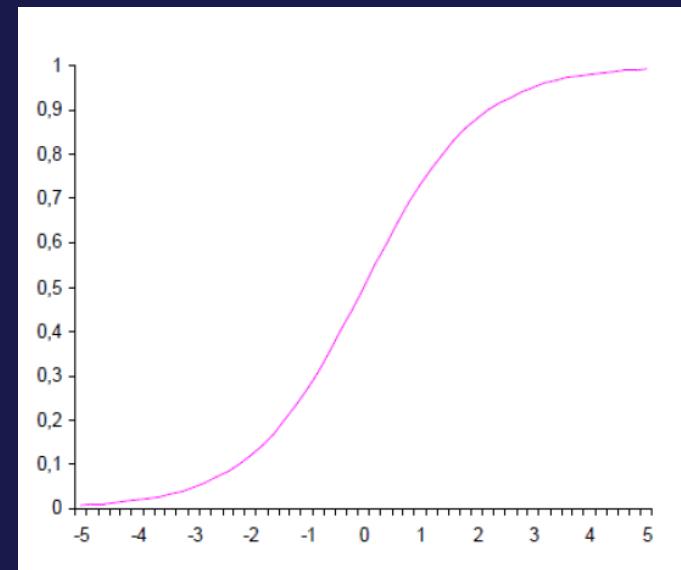
(32/55)*15

SPÉCIFICATION DU MODÈLE

➤ F ne peut pas être une fonction linéaire



➤ Afin que l'espérance de Y ne prenne que 2 valeurs, une utilise la fonction logistique :



ODDS-RATIO / RAPPORT DE CÔTE / CÔTE DE SUCCÈS

- $ODDs\ ratio = \exp(\theta) = \frac{\pi}{1+\pi}$ (peut aussi s'écrire $\theta = \ln(\frac{\pi}{1-\pi})$)
 - Avec π la probabilité de succès
- La probabilité de succès s'exprime à partir de la cote de succès de la manière suivante
- $\pi = \frac{\exp(\theta)}{1-\exp(\theta)}$
- Donc si :
 - $\exp(\theta) < 1$ lorsque $\pi < 0,5$ → *Evènement moins présent dans le groupe A que B*
 - $\exp(\theta) = 1$ lorsque $\pi = 0,5$ → *Evènement autant présent dans le groupe A que B*
 - $\exp(\theta) > 1$ lorsque $\pi > 0,5$ → *Evènement plus présent dans le groupe A que B*

ODDS-RATIO OU RAPPORT DE CÔTE

- Exemple :
 - La probabilité de succès (celle de développer une tumeur) observée est égale à :
 - $\hat{\pi} = \frac{40}{55} = 0,73$
 - $\exp(\hat{\theta}) = \exp\left(\frac{\pi}{1+\pi}\right) = \exp\left(\frac{0,73}{1+0,73}\right) = 2,71$
 - $\hat{\theta} = \ln(2,71) = 0,98$

Groupe	Tumeur présente	Tumeur absente	Total
Exposé	21	2	23
Contrôle	19	13	32
Total	40	15	55

ODDS-RATIO OU RAPPORT DE CÔTE

- On peut calculer la cote de succès dans différentes conditions. Le rapport de cotes permet alors d'évaluer l'influence du facteur considéré :

- $\pi = \frac{\exp(\theta_2)}{\exp(\theta_1)} = \exp(\theta_2 - \theta_1)$

Groupe	Tumeur présente	Tumeur absente	Total
Exposé	21	2	23
Contrôle	19	13	32
Total	40	15	55

- Lorsque π est >1 , le succès à une cote supérieur pour le deuxième niveau de facteur
- Exemple :

- La cote du succès (= « développer une tumeur ») observée est égale à :

- $$\left\{ \begin{array}{l} cote(succès|exposé) = \exp(\widehat{\theta}_2) = \frac{21}{2} = 10,5 \\ cote(succès|contrôle) = \exp(\widehat{\theta}_1) = \frac{19}{13} = 1,46 \end{array} \right.$$

- $d'où \hat{\pi} = \frac{21*13}{19*2} = \frac{10,5}{1,46} = 7,18 > 1$

- $et \log(\hat{\pi}) = \theta_2 - \theta_1 = 2,35 - 0,38 = 1,97 > 0$

- La cote de succès de la tumeur est supérieure (multipliée par 7) lorsque les souris sont exposées à la fumée de cigarettes.

INTERVALLE DE CONFIANCE

- Si pour chaque individu, la probabilité de succès est π , alors, le nombre Y de succès parmi n individus indépendants suit une loi binomiale $B(n, \pi)$. Ainsi :

$$E[Y] = n\pi \quad E\left(\hat{\pi} = \frac{Y}{n}\right) = \frac{1}{n} E[Y] = \pi \quad Var[Y] = n\pi(1 - \pi) = \frac{1}{n^2} Var[Y] = \frac{\pi(1 - \pi)}{n}$$

- Un intervalle de confiance (dans le cadre d'application de l'approximation de la loi binomiale par une loi normale) à 95 % pour π est donné par :

$$\hat{\pi} \pm 1,96 \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}}$$

INTERVALLE DE CONFIANCE

- Dans notre exemple on souhaiterait comparer les probabilités π_1 et π_2 de développer une tumeur sous et sans exposition à la fumée de cigarette et déterminer si elles sont significativement différentes. Cela reviendrait à déterminer s'il existe un lien entre le développement de la tumeur et le facteur risque considéré.
- On peut déjà répondre à cette question en construisant un intervalle de confiance à 95 % pour $\pi_1 - \pi_2$

$$(\widehat{\pi}_1 - \widehat{\pi}_2) \pm 1,96 \sqrt{\frac{\widehat{\pi}_1(1 - \widehat{\pi}_1)}{n_1} + \frac{\widehat{\pi}_2(1 - \widehat{\pi}_2)}{n_2}}$$

Groupe	Tumeur présente	Tumeur absente	Total
Exposé	21	2	23
Contrôle	19	13	32
Total	40	15	55

$$\left(\frac{21}{23} - \frac{19}{32}\right) \pm 1,96 \sqrt{\frac{\frac{21}{23}(1 - \frac{21}{23})}{23} + \frac{\frac{19}{32}(1 - \frac{19}{32})}{32}}$$

$$(0,913 - 0,593) \pm 1,96 \sqrt{\frac{0,913(1 - 0,913)}{23} + \frac{0,593(1 - 0,593)}{32}} \\ 0,319 \pm 0,205 = [0,114;0,524]$$

→ On en déduit que la différence $\pi_1 - \pi_2$ est significativement écartée de 0 au seuil $\alpha = 5\%$. Ainsi on sait non seulement la fumée de cigarette a un effet significatif sur le nombre de cancer développés mais on a quantifié cet effet.

RÉGRESSION LOGISTIQUE

- I. Introduction
- II. Variable explicative qualitative
- III. Variable explicative quantitative
- IV. Variable explicatives mixtes
- V. Sélection du modèle
- VI. Validation du modèle



DÉFINITION

- Si X est une variable explicative à K niveaux, le modèle logistique suppose que :

$$(Y|X = x_k) \sim B(n_k, \pi_k)$$

- Avec

$$\begin{aligned} \text{logit}(\pi_k) &= \log\left(\frac{\pi_k}{1 - \pi_k}\right) = \theta_k = \mu + \alpha_k ; (\alpha_1 = 0) \\ \pi_k &= \frac{\exp(\mu + \alpha_k)}{1 + \exp(\mu + \alpha_k)} \end{aligned}$$

- Le logarithme de la cote de succès sous le premier niveau du facteur vaut μ .
- Le logarithme du rapport des cotes du succès sous les k èmes et 1er niveau du facteur vaut $\theta_k - \theta_1 = \alpha_k$
- Par conséquent une valeur de $\alpha_k > 0 (< 0)$ indique que la cote du succès observée est plus grande (petite) sous le k èmes niveau du facteur que sous le 1er niveau.

Définition

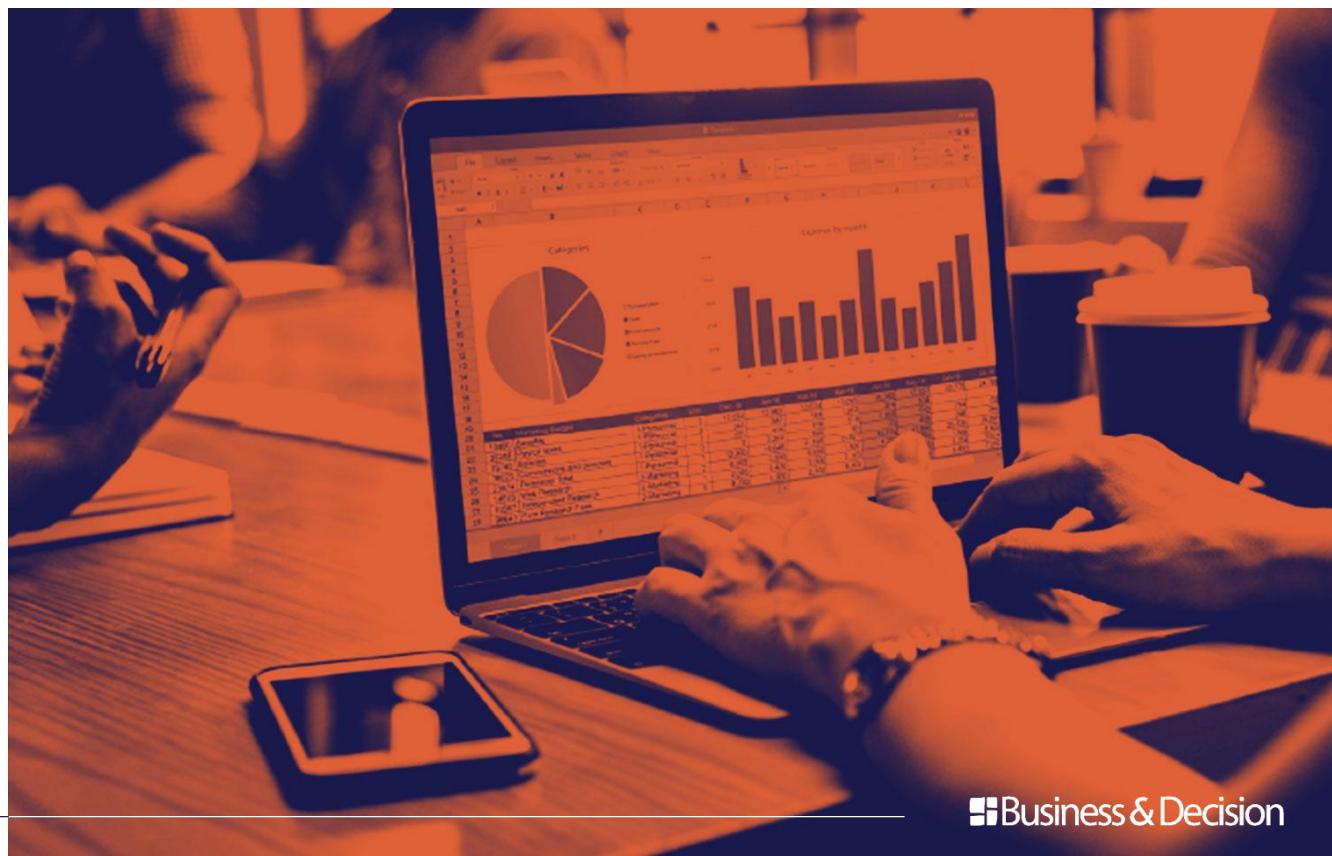


- On estime les α_k à l'aide d'une méthode mathématique appelée maximum de vraisemblance.
- Dans ce cas, on sait qu'asymptotiquement (lorsque la taille de l'échantillon tend vers l'infini) ces estimateurs suivent une loi normale et sont sans biais.
- Par conséquent un intervalle de confiance à 95 % approximatif pour les k est :

$$\widehat{\alpha}_k \pm 1,96 * \sigma(\widehat{\alpha}_k)$$

- ▶ Les différents modèles possibles sont :
 - ▶ Modèle 1 avec effet traitement :
 - ▶ $\text{logit}(\pi_k) = \theta_k = \mu + \alpha_k$
 - ▶ Modèle 2 sans effet traitement :
 - ▶ $\text{logit}(\pi_k) = \theta_k = \mu$
 - ▶ Est-ce que le modèle 1 est significativement meilleur que le modèle 2 ?

Modèle



DEVIANCE

- On compare alors la probabilité de succès estimée dans le groupe $\widetilde{\pi}_k$ et la proportion de succès observée $\widehat{\pi}_k$
- La déviance D est alors définie ainsi :

$$D = -2 \sum_k \left\{ y_k \log \left(\frac{\widetilde{\pi}_k}{\widehat{\pi}_k} \right) + (n_k - y_k) \log \left(\frac{1 - \widetilde{\pi}_k}{1 - \widehat{\pi}_k} \right) \right\} = -2(l(\widetilde{\pi}_k) - l(\widehat{\pi}_k))$$

- Cette quantité est à rapprocher de la somme des carrés à minimiser dans la régression linéaire simple ou multiple. Elle évalue globalement la qualité de l'ajustement obtenu.

DEVIANC

- Le deuxième modèle ne fait pas intervenir de variable explicative. Il peut servir à tester la nullité de toutes les pentes :
 - L'équivalent du test de Fisher global dans le cadre de la régression logistique
- On calcule la statistique ci-dessous comparant la déviance des deux modèles

$$G^2 = D_2 - D_1 = -2(l_2 - l_1)$$

- Sous l'hypothèse H_0 que les restrictions impliquées par le modèle 2 au modèle 1 sont correctes

$$G^2_{H0} \sim \chi^2_{ddl_2 - ddl_1}$$

EXEMPLE

➤ Dans notre exemple nous avons :

- $H_0: \alpha_2 = 0 \rightarrow$ Pas d'effet du traitement
- $H_1: \alpha_2 \neq 0 \rightarrow$ effet du traitement

Groupe	Tumeur présente	Tumeur absente	Total
Exposé	21	2	23
Contrôle	19	13	32
Total	40	15	55

➤ Nous avons :

- $G_2 = 7,635$ et $ddl_1 = 0, ddl_2 = 1$ ce qui donne une $p - value = 0,006$
- Donc $\alpha_2 \neq 0$ au niveau $\alpha = 5\%$

➤ Nous avons également les informations suivantes :

- $\hat{\mu} = 0,38 \quad \widehat{\alpha}_2 = 1,97$
- Probabilité de succès : $\widehat{\theta}_1 = 0,59 \quad \widehat{\theta}_2 = 0,91$

➤ Le rapport des cotes de groupe exposé contre le groupe contrôle est estimé $\exp(\widehat{\alpha}_2) = 7,24$

➤ Soit une cote de succès plus de 7 fois plus grande pour le groupe des traités.

EXEMPLE

- Nous pouvons également construire un intervalle de confiance $(1 - \alpha)$ pour le log du rapport de côtes (LRC) du groupe K contre le groupe de référence α_k avec

$$\widehat{\alpha}_k \pm 1,96\sigma(\widehat{\alpha}_k)$$

- Dans notre exemple, on obtient :
 - $\alpha_2 \in (0,36; 3,58) \rightarrow$ rejet de H_0 avec $\alpha = 0,05$
 - On peut conclure à une augmentation significative des chances de développer un cancer du poumon après exposition à la fumée de cigarettes.
 - l'intervalle de confiance pour le rapport des côtes est alors $(1,43 ; 36,0)$

Groupe	Tumeur présente	Tumeur absente	Total
Exposé	21	2	23
Contrôle	19	13	32
Total	40	15	55

EXEMPLE N°2

➤ Relation entre les habitudes tabagiques d'étudiants en Arizona et les habitudes des parents

Groupe	Enfant - Fumeur	Enfant – Non Fumeur	Total
Parent -Deux	400	1380	1780
Parent - Un seul	416	1823	2239
Parent - Aucun	188	1168	1358
Total	1004	4371	5375

Groupe	Enfant - Fumeur	Enfant – Non Fumeur	Total
Parent -Deux	22%	78%	100%
Parent - Un seul	19%	81%	100%
Parent - Aucun	14%	86%	100%
Total	19%	81%	

EXEMPLE N°2

- On définit le succès comme étant le fait de fumer pour l'enfant, le modèle logistique précédent devient :

$$\text{logit}(\pi_k) = \theta_k = \mu + \alpha_k ; (\alpha_1 = 0)$$

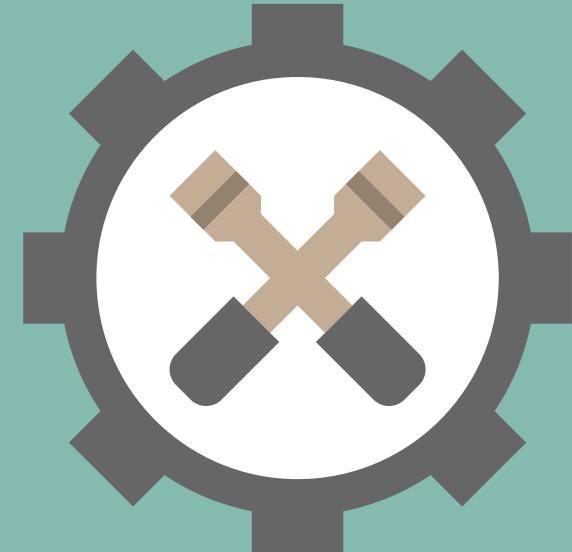
- La catégorie de référence est par défaut « Aucun »
- On peut tester l'hypothèse :

$$H_0 = \alpha_2 = \alpha_3 = 0$$

- En comparant la déviance de ce modèle avec celle du précédent :
 - $G_{obs}^2 = 38,37$
 - $p - value < 0,0001$
- Conclusion du test : association significative au niveau = 5% entre habitudes tabagiques des parents et des enfants.

EXERCICE

1. Importer le fichier « Data.csv »
 1. Vérifier le format des champs et le bon import des données
2. La variable à prédire est « Winner »
 1. Prédire le gagnant du match en fonction de
 - R_Stance
 - B_Stance
 - Créer un nouvelle indicateur pour savoir si les deux combattants ont le même style de combat



RÉGRESSION LOGISTIQUE

- I. Introduction
- II. Variable explicative qualitative
- III. Variable explicative quantitative**
- IV. Variable explicatives mixtes
- V. Sélection du modèle
- VI. Validation du modèle



EXEMPLE

► Données :

- X1 : age du patient (quantitative)
- X2 : taux max (quantitative)
- X3 : angine de poitrine (binaire)
- Y : Détection d'une maladie cardiaque

➤ Objectif : Prédire la présence d'une maladie cardiaque à partir des 3 variables

age	taux_max	angine	coeur
50	126	1	presence
49	126	0	presence
46	144	0	presence
49	139	0	presence
62	154	1	presence
35	156	1	presence
67	160	0	absence
65	140	0	absence
47	143	0	absence
58	165	0	absence
57	163	1	absence
59	145	0	absence
44	175	0	absence
41	153	0	absence
54	152	0	absence
52	169	0	absence
57	168	1	absence
50	158	0	absence
44	170	0	absence
49	171	0	absence

VARIABLE EXPLICATIVE CONTINU

- Dans un premier temps, prenons simplement en compte l'âge du patient
- Nous essayons donc de répondre à la question suivante :

Existe-t-il un lien entre l'âge et la maladie cardiaque? Si oui quelle est la nature de cette relation ?

- On cherche donc à déterminer comment la probabilité de succès π change avec une ou plusieurs variables explicatives continues à partir des observations de y_i succès en n_i expériences indépendantes sous des valeurs de X observées égales à x_i , ($i = 1, \dots, I$).

VARIABLE EXPLICATIVE CONTINU

- On souhaite utiliser une modélisation de la cote de succès :

$$(Y|X = x_i) \sim B(n_i, \pi_i)$$

$$\text{logit}(\pi_i) = \theta_i = \alpha_0 + \beta_1 x_i$$

- Ou $x_i = \log(dose_i)$

- Et les coefficients α_0 et β_1 obtenue par la méthode du maximum de vraisemblance

```
> model_quali
Call: glm(formula = coeur ~ age, family = binomial, data = maladie_cardiage)

Coefficients:
(Intercept)           age
            3.1907      -0.0795

Degrees of Freedom: 19 Total (i.e. Null); 18 Residual
Null Deviance: 24.43
Residual Deviance: 22.94          AIC: 26.94
> summary(model_quali)

Call:
glm(formula = coeur ~ age, family = binomial, data = maladie_cardiage)

Deviance Residuals:
    Min      1Q  Median      3Q     Max 
-1.1483 -0.8744 -0.6698  1.1025  1.9496 

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) 3.19067   3.48228   0.916   0.360    
age         -0.07950  0.06901  -1.152   0.249    
(Dispersion parameter for binomial family taken to be 1)
```

VARIABLE EXPLICATIVE CONTINU

➤ Si on ajoute la variable taux_max

```
> model_quali
Call: glm(formula = coeur ~ age, family = binomial, data = maladie_cardiage)

Coefficients:
(Intercept)      age
   3.1907     -0.0795

Degrees of Freedom: 19 Total (i.e. Null);  18 Residual
Null Deviance: 24.43
Residual Deviance: 22.94      AIC: 26.94
> summary(model_quali)

Call:
glm(formula = coeur ~ age, family = binomial, data = maladie_cardiage)

Deviance Residuals:
    Min      1Q  Median      3Q      Max 
-1.1483 -0.8744 -0.6698  1.1025  1.9496 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) 3.19067  3.48228  0.916   0.360    
age        -0.07950  0.06901 -1.152   0.249    
(Dispersion parameter for binomial family taken to be 1)
```

```
> model_quali
Call: glm(formula = coeur ~ age + taux_max, family = binomial, data = maladie_cardiage)

Coefficients:
(Intercept)      age      taux_max
   25.6520     -0.1066     -0.1404

Degrees of Freedom: 19 Total (i.e. Null);  17 Residual
Null Deviance: 24.43
Residual Deviance: 14.23      AIC: 20.23
> summary(model_quali)

Call:
glm(formula = coeur ~ age + taux_max, family = binomial, data = maladie_cardiage)

Deviance Residuals:
    Min      1Q  Median      3Q      Max 
-1.4278 -0.5679 -0.2316  0.3576  2.2998 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) 25.65196 11.80432  2.173   0.0298 *  
age        -0.10658  0.07984 -1.335   0.1819    
taux_max   -0.14036  0.06831 -2.055   0.0399 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 24.435  on 19  degrees of freedom
Residual deviance: 14.229  on 17  degrees of freedom
AIC: 20.229

Number of Fisher Scoring iterations: 5
```

VARIABLE EXPLICATIVE CONTINU

➤ Seulement taux_max

```
> model_quali
Call: glm(formula = coeur ~ taux_max, family = binomial, data = maladie_cardiag
e)
Coefficients:
(Intercept)      taux_max
     18.2777      -0.1271
Degrees of Freedom: 19 Total (i.e. Null);  18 Residual
Null Deviance:    24.43
Residual Deviance: 16.22      AIC: 20.22
> summary(model_quali)

Call:
glm(formula = coeur ~ taux_max, family = binomial, data = maladie_cardiag
e)

Deviance Residuals:
    Min      1Q   Median      3Q      Max
-1.3864  -0.5966  -0.2906   0.4458   1.8691

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 18.27773  8.74592  2.090  0.0366 *
taux_max    -0.12713  0.05877 -2.163  0.0305 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
> model_quali
Call: glm(formula = coeur ~ age + taux_max, family = binomial, data = maladie_cardiag
e)
Coefficients:
(Intercept)          age        taux_max
       25.6520      -0.1066      -0.1404
Degrees of Freedom: 19 Total (i.e. Null);  17 Residual
Null Deviance:    24.43
Residual Deviance: 14.23      AIC: 20.23
> summary(model_quali)

Call:
glm(formula = coeur ~ age + taux_max, family = binomial, data = maladie_cardiag
e)

Deviance Residuals:
    Min      1Q   Median      3Q      Max
-1.4278  -0.5679  -0.2316   0.3576   2.2998

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 25.65196  11.80432  2.173  0.0298 *
age         -0.10658   0.07984 -1.335  0.1819
taux_max    -0.14036   0.06831 -2.055  0.0399 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 24.435 on 19 degrees of freedom
    Residual deviance: 14.229 on 17 degrees of freedom
    AIC: 20.229

Number of Fisher Scoring iterations: 5
```

RÉGRESSION LOGISTIQUE

- I. Introduction
- II. Variable explicative qualitative
- III. Variable explicative quantitative
- IV. Variable explicatives mixtes**
- V. Sélection du modèle
- VI. Validation du modèle



RÉGRESSION LOGISTIQUE : VARIABLES EXPLICATIVES MIXTES

- Dans l'exemple précédent nous avons ignoré la variable « angine » sur la probabilité de succès.
- L'analyse précédente que le « taux_max » influe sur la présence d'une maladie cardiaque
- Considérons le cas simple où on a à la fois une variable continue X et une variable qualitative Z.
- Les données sont donc du type $(y_{ki}, n_{ki}, x_{ki}, z_{ki})$
- Le modèle suggéré est donc :

$$(Y|X = x_i, Z = z_i) \sim B(n_{ki}, \pi_{ki})$$
$$\text{logit}(\pi_{ki}) = \theta_{ki}$$

RÉGRESSION LOGISTIQUE : VARIABLES EXPLICATIVES MIXTES

➤ Nous avons donc 5 modèles à notre disposition

- $Y = X + Z + X * Z, (\alpha_0 + \alpha_k) + (\beta_1 + \tau_1)x_{ki}$
- $Y = X + Z, (\alpha_0 + \alpha_k) + \beta_1 x_{ki}$
- $Y = X, \alpha_0 + \beta_1 x_{ki}$
- $Y = Z, (\alpha_0 + \alpha_k)$
- $Y = 1, \alpha_0$

➤ Reste à détecter les modèles convenables à l'aide du test du G2
➤ Méthode similaire à l'ANCOVA

RÉGRESSION LOGISTIQUE : VARIABLES EXPLICATIVES MIXTES

Avec interaction

```
> model_uali  
  
Call: glm(formula = coeur ~ taux_max * angine, family = binomial, data = maladie_cardiage)  
  
Coefficients:  
              (Intercept)          taux_max          angine1    taux_max:angine1  
             46.6991        -0.3319         839.2143       -5.2227  
  
Degrees of Freedom: 19 Total (i.e. Null); 16 Residual  
Null Deviance: 24.43  
Residual Deviance: 6.601      AIC: 14.6  
> summary(model_uali)  
  
Call:  
glm(formula = coeur ~ taux_max * angine, family = binomial, data = maladie_cardiage)  
  
Deviance Residuals:  
      Min        1Q     Median        3Q       Max  
-1.27963 -0.10604 -0.01013  0.00000  1.66139  
  
Coefficients:  
              Estimate Std. Error z value Pr(>|z|)  
(Intercept) 4.670e+01 3.915e+01  1.193   0.233  
taux_max    -3.319e-01 2.751e-01 -1.206   0.228  
angine1      8.392e+02 3.254e+05  0.003   0.998  
taux_max:angine1 -5.223e+00 2.042e+03 -0.003   0.998  
  
(Dispersion parameter for binomial family taken to be 1)
```

Sans interaction

```
> model_uali  
  
Call: glm(formula = coeur ~ taux_max + angine, family = binomial, data = maladie_cardiage)  
  
Coefficients:  
              (Intercept)          taux_max          angine1  
             71.7701        -0.5084         9.4626  
  
Degrees of Freedom: 19 Total (i.e. Null); 17 Residual  
Null Deviance: 24.43  
Residual Deviance: 7.653      AIC: 13.65  
> summary(model_uali)  
  
Call:  
glm(formula = coeur ~ taux_max + angine, family = binomial, data = maladie_cardiage)  
  
Deviance Residuals:  
      Min        1Q     Median        3Q       Max  
-1.43761 -0.11108 -0.00225  0.00768  1.81804  
  
Coefficients:  
              Estimate Std. Error z value Pr(>|z|)  
(Intercept) 71.7701  43.1382  1.664   0.0962 .  
taux_max    -0.5084   0.3040  -1.673   0.0944 .  
angine1      9.4626  5.7056  1.658   0.0972 .  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
(Dispersion parameter for binomial family taken to be 1)
```

EXERCICE

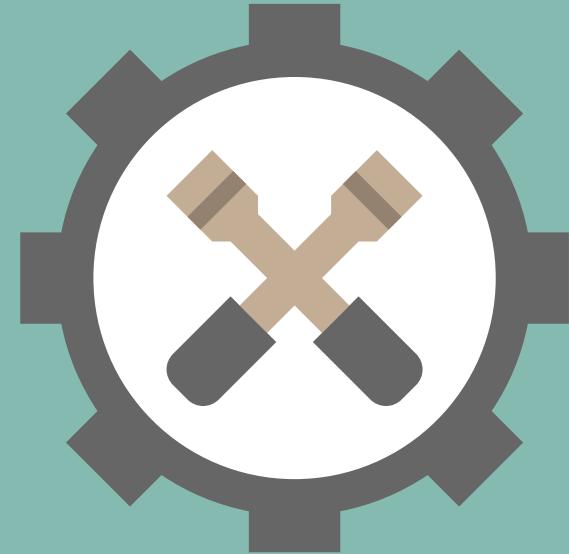
1. Importer le fichier « Data.csv »

1. Vérifier le format des champs et le bon import des données

2. La variable à prédire est « Winner »

1. Prédire le gagnant du match en fonction

- Nombre de coup à la tête du joueur rouge
- Nombre de coup à la tête du joueur bleu
- Différence de nombre de coup à la tête entre les deux joueurs



RÉGRESSION LOGISTIQUE

- I. Introduction
- II. Variable explicative qualitative
- III. Variable explicative quantitative
- IV. Variable explicatives mixtes
- V. Sélection du modèle**
- VI. Validation du modèle



Sélection du modèle

- » Critère usuelle pour la régression linéaire
 - » R^2 / R^2_{adj} / Variance résiduelle / ...
 - » Basé sur l'écart moyen entre réel et prédit
 - » Impossible sur variable qualitative
- » Critère de performance :
 - » Déviance
 - » AIC
 - » BIC

DÉVIANCE

```
> model_quali<-glm(coeur~age,data=maladie_cardiage,family=binomial)
> model_quali$deviance
[1] 22.94496
> #calcul de la vraisemblance
> prev <- model_quali$fitted.values #on obtient les pi
> vrais <- rep(0,nrow(maladie_cardiage))
> vrais[maladie_cardiage$coeur=="presence"] <- prev[maladie_cardiage$coeur=="presence"]
> vrais[maladie_cardiage$coeur=="absence"] <- 1-prev[maladie_cardiage$coeur=="absence"]
> vrais <- prod(vrais) #vrais est la vraisemblance du modèle
> dev <- -2*log(vrais)
> dev
[1] 22.94496
```

Age

Age + taux_max

```
> model_quali<-glm(coeur~age+taux_max,data=maladie_cardiage,family=binomial)
> model_quali$deviance
[1] 14.22908
> #calcul de la vraisemblance
> prev <- model_quali$fitted.values #on obtient les pi
> vrais <- rep(0,nrow(maladie_cardiage))
> vrais[maladie_cardiage$coeur=="presence"] <- prev[maladie_cardiage$coeur=="presence"]
> vrais[maladie_cardiage$coeur=="absence"] <- 1-prev[maladie_cardiage$coeur=="absence"]
> vrais <- prod(vrais) #vrais est la vraisemblance du modèle
> dev <- -2*log(vrais)
> dev
[1] 14.22908
```

AIC / BIC

- **Problème** : Déviance ne marche que pour des modèles emboîté car diminue avec le nombre de variable → Même problématique que le R²
 - Vraisemblance augmente avec le nombre de variable
 - Log-Vraisemblance diminue avec le nombre de variable
 - Choix du modèle saturé
- Par définition l'AIC (Akaike Informative Criterion) pour un modèle M de dimension p est défini par :

$$AIC(M) = -2L_n(\hat{M}) + 2p$$

- Le critère de choix de modèle le BIC (Bayesian Informative Criterion) pour un modèle M de dimension p est défini par :

$$BIC(M) = -2L_n(\hat{M}) + p \log(n)$$

AIC / BIC

- **Problème** : Déviance ne marche que pour des modèles emboîté car diminue avec le nombre de variable → Même problématique que le R²
 - Vraisemblance augmente avec le nombre de variable
 - Log-Vraisemblance diminue avec le nombre de variable
 - Choix du modèle saturé
- Par définition l'AIC (Akaike Informative Criterion) pour un modèle M de dimension p est défini par :

$$AIC(M) = -2L_n(\hat{M}) + 2p$$

- Le critère de choix de modèle le BIC (Bayesian Informative Criterion) pour un modèle M de dimension p est défini par :

$$BIC(M) = -2L_n(\hat{M}) + p \log(n)$$

EXAMPLE AIC

```
> model_quali<-glm(coeur~age,data=maladie_cardiage,family=binomial)
> model_quali$aic
[1] 26.94496
> model_quali<-glm(coeur~age+taux_max,data=maladie_cardiage,family=binomial)
> model_quali$aic
[1] 20.22908
> model_quali<-glm(coeur~taux_max,data=maladie_cardiage,family=binomial)
> model_quali$aic
[1] 20.22294
> model_quali<-glm(coeur~taux_max + anginge ,data=maladie_cardiage,family=binomial)
> model_quali$aic
[1] 13.65303
> model_quali<-glm(coeur~taux_max * anginge ,data=maladie_cardiage,family=binomial)
Warning message:
glm.fit: des probabilités ont été ajustées numériquement à 0 ou 1
> model_quali$aic
[1] 14.60147
```

RÉGRESSION LOGISTIQUE

- I. Introduction
- II. Variable explicative qualitative
- III. Variable explicative quantitative
- IV. Variable explicatives mixtes
- V. Sélection du modèle
- VI. Validation du modèle**



TEST D'HOSMER LEMESHOW

- Regroupement des probabilités prédictes \hat{y}_i par le modèle en dix groupes (déciles)
- Les probabilités $\hat{p}_\beta(x_i)$ sont ordonnées par ordre croissant ($\hat{p}_\beta(x_i)$ est la probabilité $P_\beta(Y = 1|X = x_i)$ estimée par le modèle) ;
- Ces probabilités ordonnées sont ensuite séparées en K groupes de taille égale (on prend souvent K = 10 si n est suffisamment grand). On note :
 - m_k les effectifs du groupe k ;
 - o_k le nombre de succès ($Y = 1$) observé dans le groupe k ;
 - μ_k la moyenne des $\hat{p}_\beta(x_i)$ dans le groupe k
- La statistique de test est alors :

$$C^2 = \sum_{k=1}^K \frac{(o_k - m_k \mu_k)^2}{m_k \mu_k (1 - \mu_k)} \sim \chi^2_{K-1}$$

TABLEAU DE CONTINGENCE

	Malade = oui $y_i = 1$	Malade = non $y_i = 0$
Prédit oui $\hat{y}_i = 1$	a	b
Prédit non $\hat{y}_i = 0$	c	d

Ce tableau permet de connaître le nombre de bonnes et de mauvaises prédictions par rapport à un seuil « s » (fixé généralement à 50%)

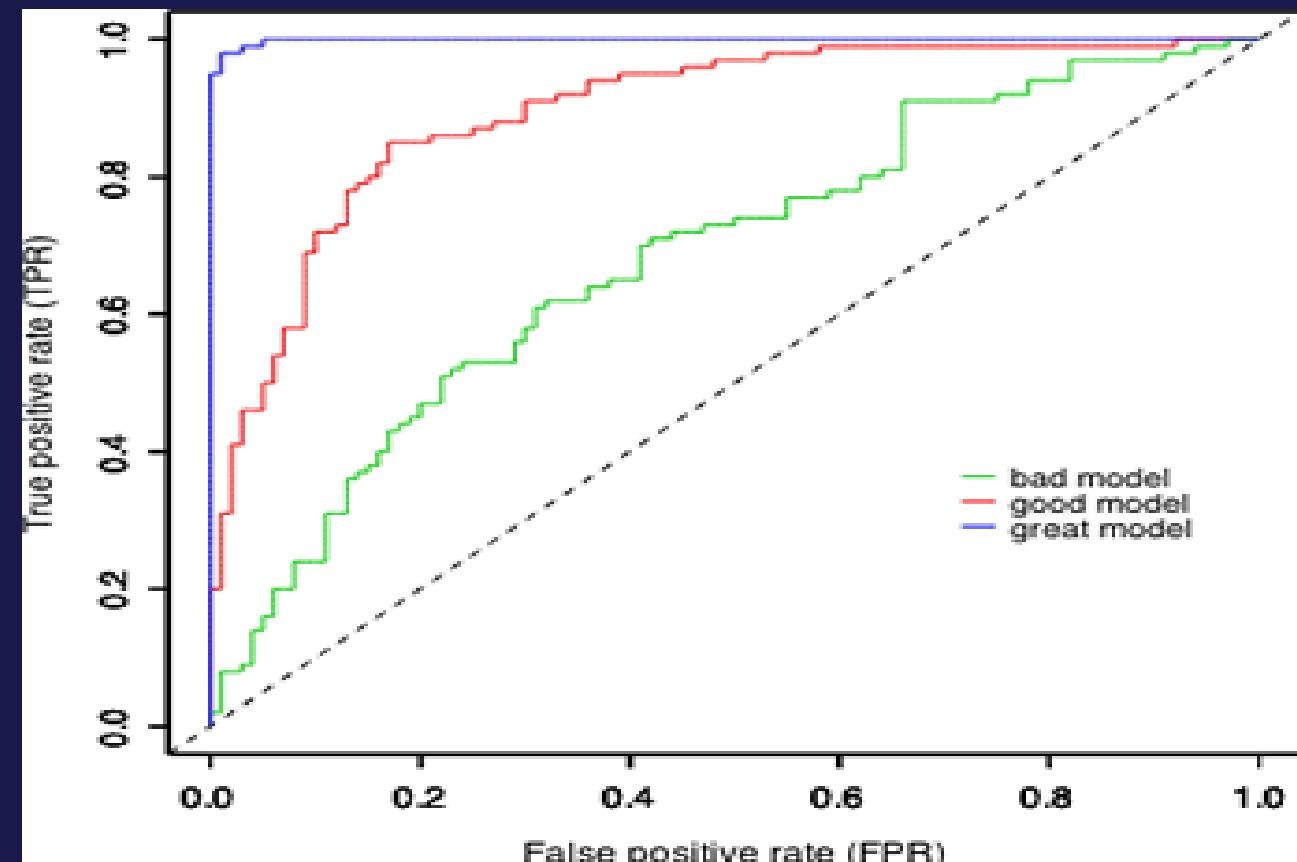
	Malade = oui $y_i = 1$	Malade = non $y_i = 0$
Prédit oui $\hat{y}_i = 1$	13	1
Prédit non $\hat{y}_i = 0$	1	5

$$\text{Bien prédit} = \frac{a + d}{n} =$$

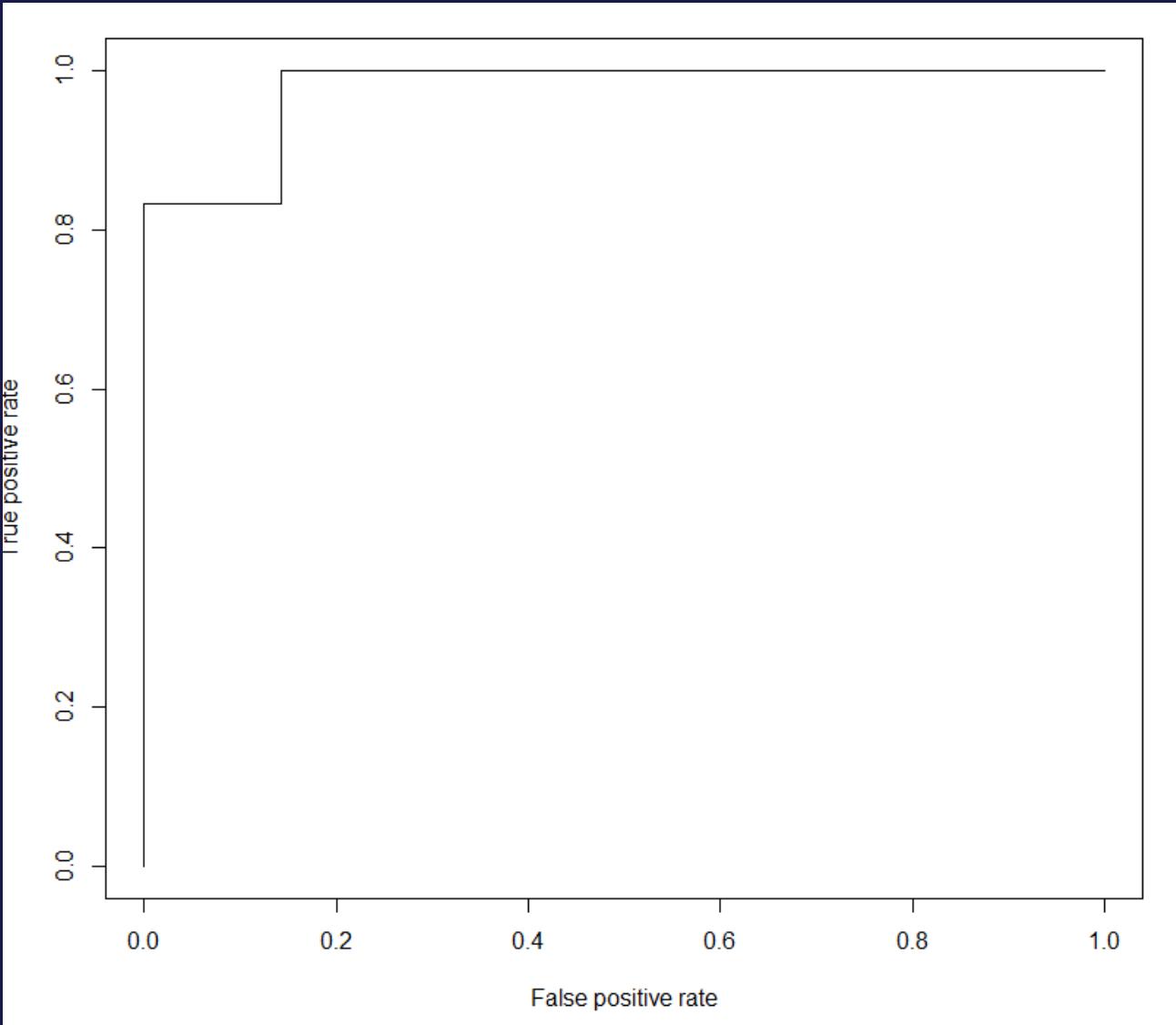
$$\text{mal prédit} = \frac{b + c}{n}$$

COURBE ROC

- Vrais positifs (VP) : nombre d'individus déclare positifs par le test et qui le sont effectivement.
- Faux positifs (FP) : nombre d'individus déclare positifs par le test mais qui sont en réalité négatifs.
- Vrais négatifs (VN) : nombre d'individus déclare négatifs par le test et qui le sont effectivement.
- Faux négatifs (FN) : nombre d'individus détectés négatifs par le test mais qui sont en réalité positifs.
- Sensibilité :
 - proportion d'individus positifs effectivement bien détectés par le test.
 - Sensibilité = $VP/(VP + FN)$
- Spécificité
 - proportion d'individus négatifs effectivement bien détectés par le test
 - Spécificité = $VN/(VN + FP)$
- Calcul de l'aire sous la courbe
 - Proche de 1



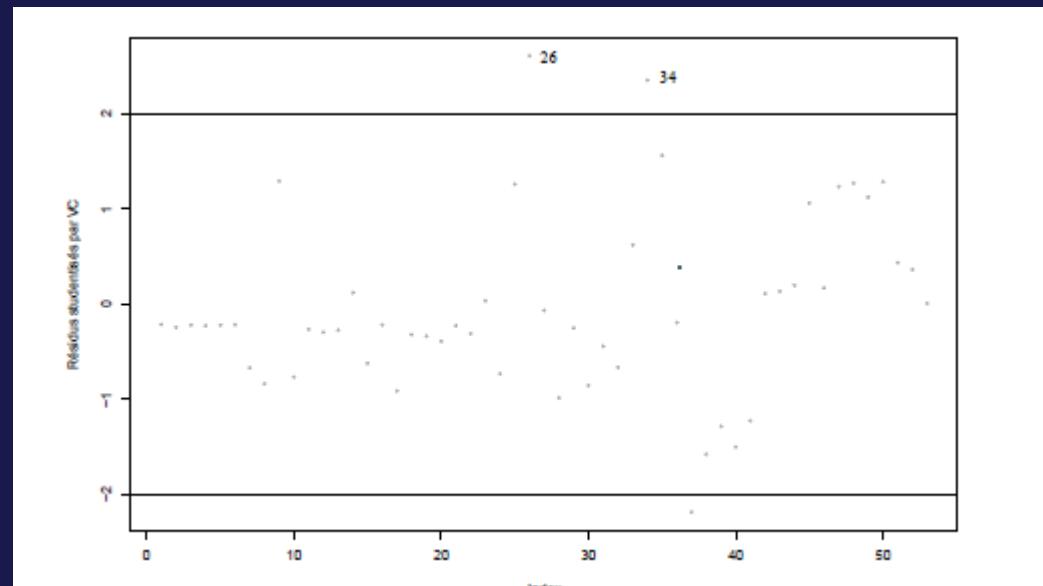
COURBE ROC



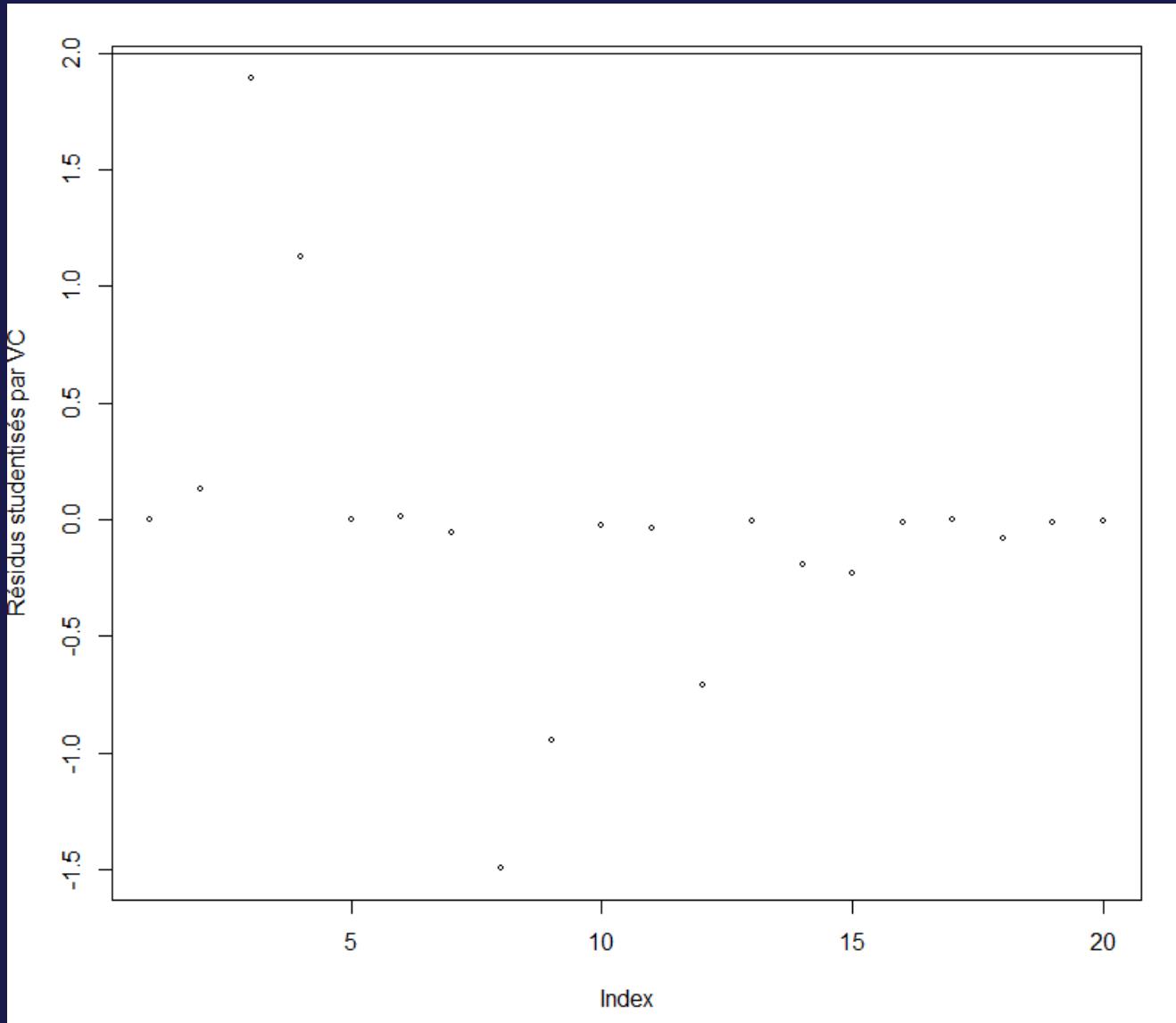
ANALYSE DES RÉSIDUS

- Mesure de l'ajustement du modèle
- Calcul des résidus de pearson standardisés :

$$RP_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)(1 - \hat{h}_{ii})}} \text{ avec la matrice } H = X(X'W_{\hat{\beta}}X)^{-1}X'W_{\hat{\beta}}$$
$$-2 < RP_i < 2$$

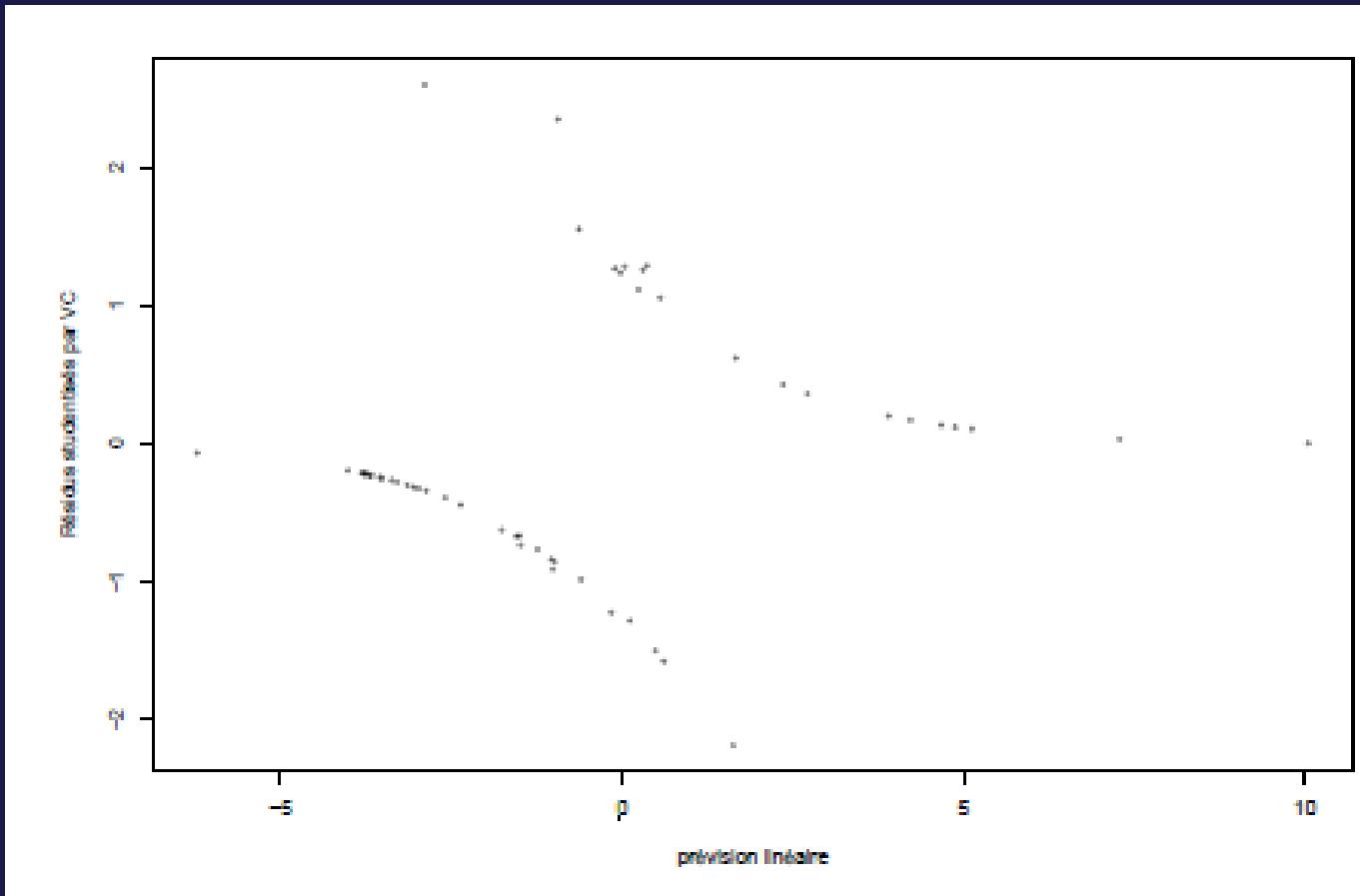


ANALYSE DES RÉSIDUS

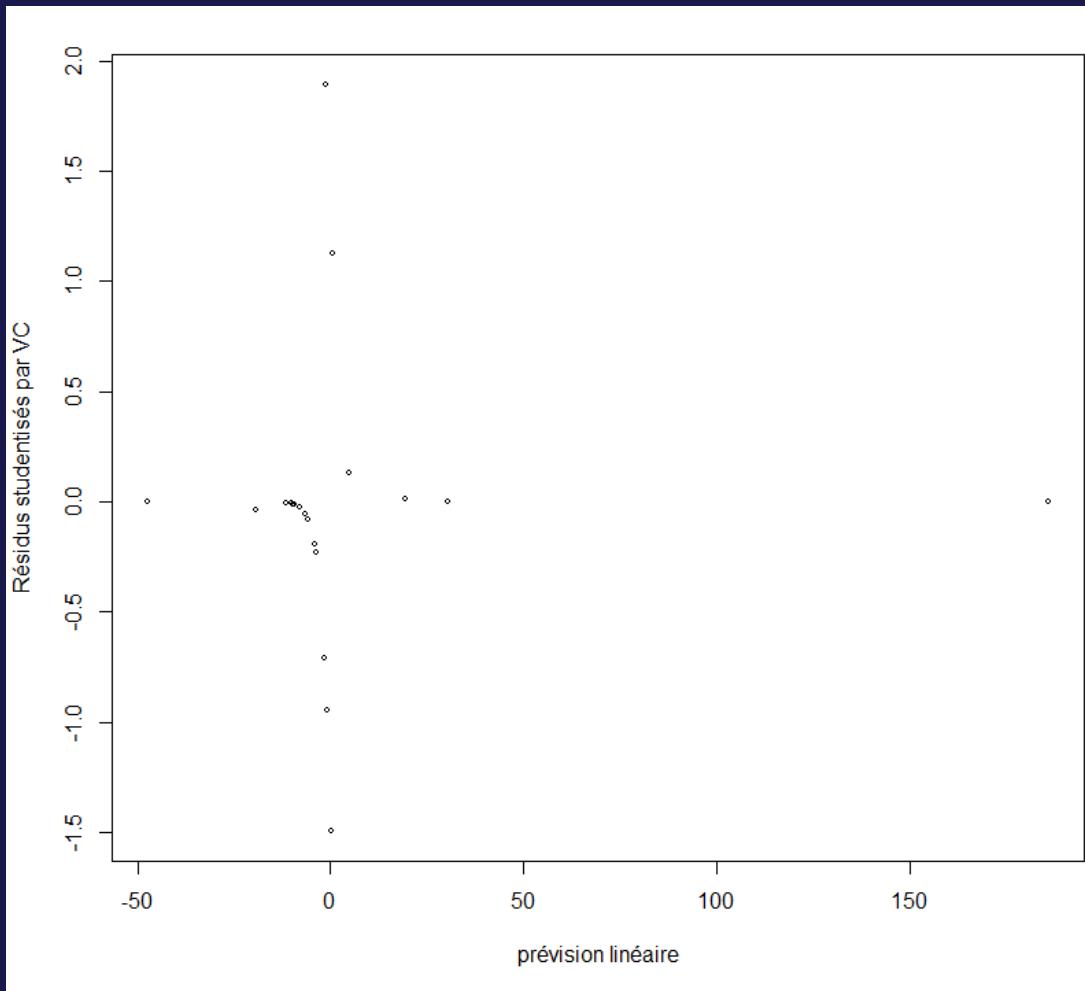


ANALYSE DES RÉSIDUS

- Analyse du graphique
 - Abscisse : Prédiction
 - Ordonnée : Résidu
- Permet de :
 - Vérifier s'il n'y a pas une structure dans les résidus
 - Déetecter les valeurs aberrantes

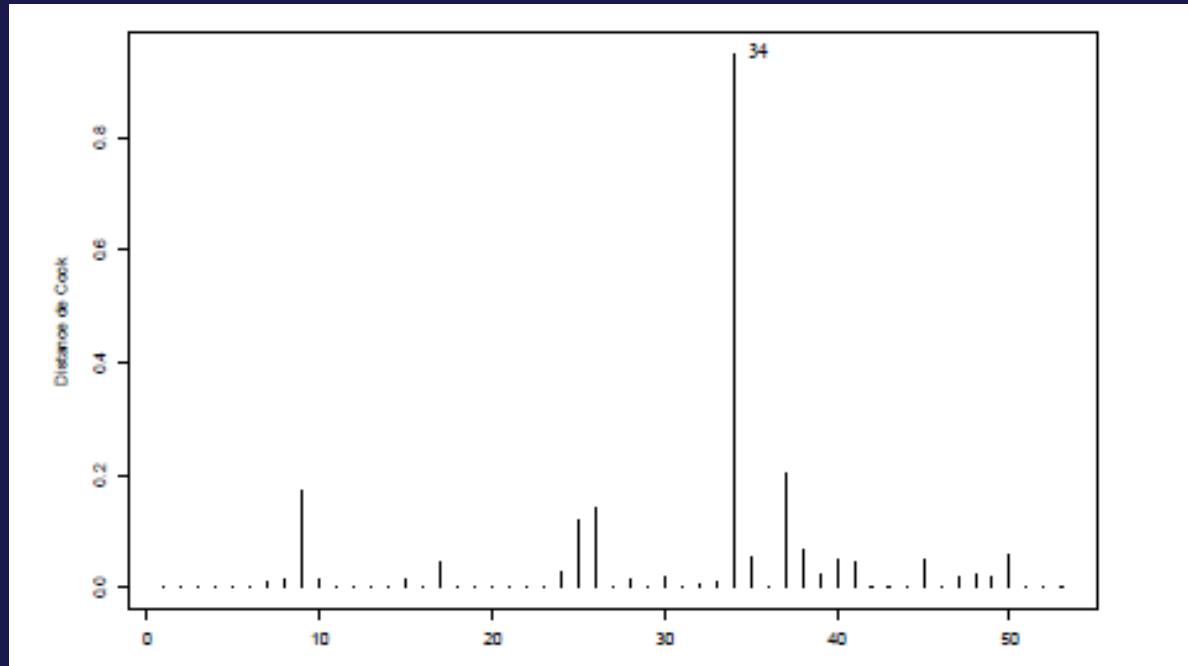


ANALYSE DES RÉSIDUS

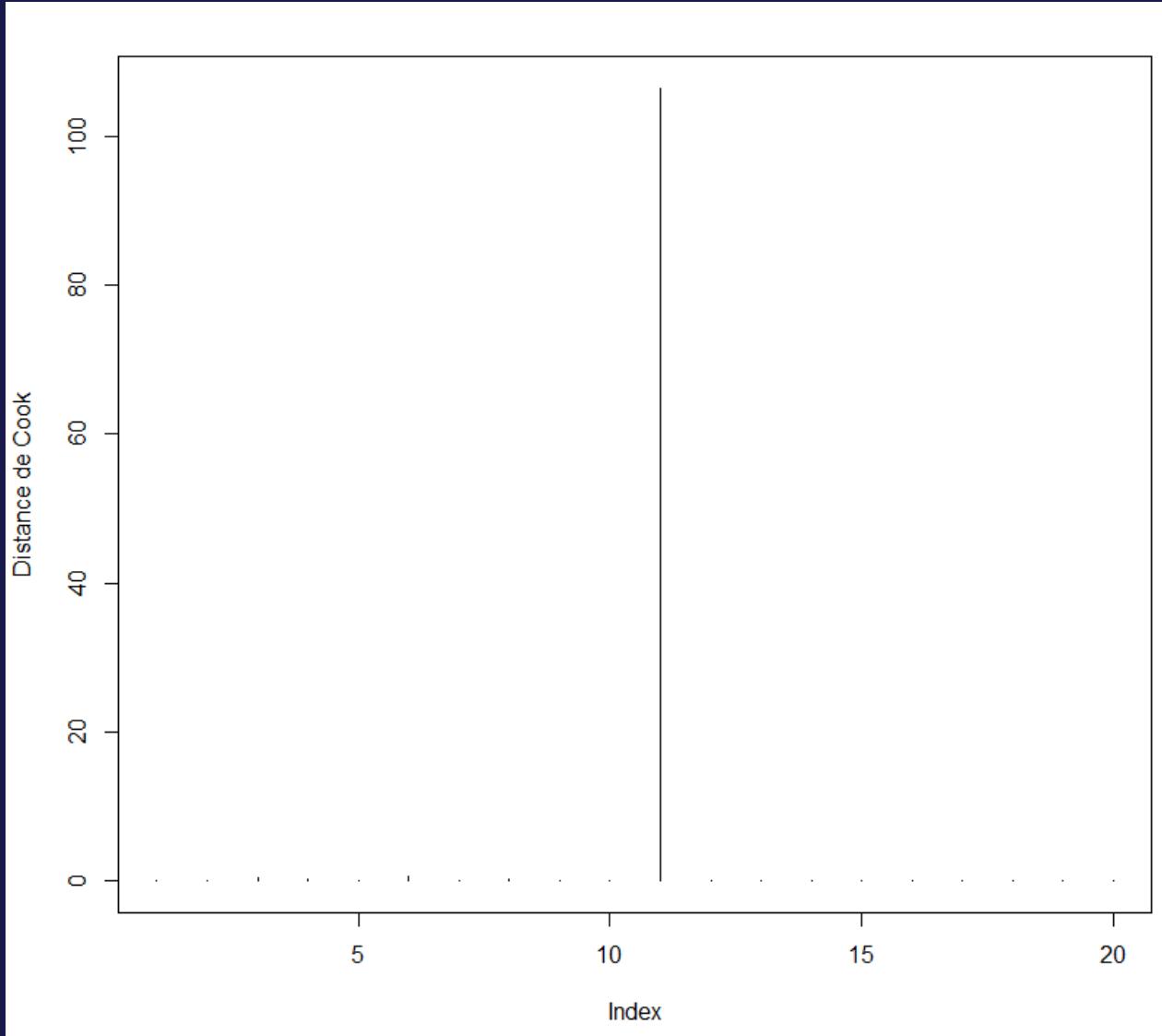


POINT INFLUENT

- Point influent : Distance de Cook
 - Suppression de l'individu i et re-calculation de la régression
 - Si les coefficients varient fortement → Individu influent

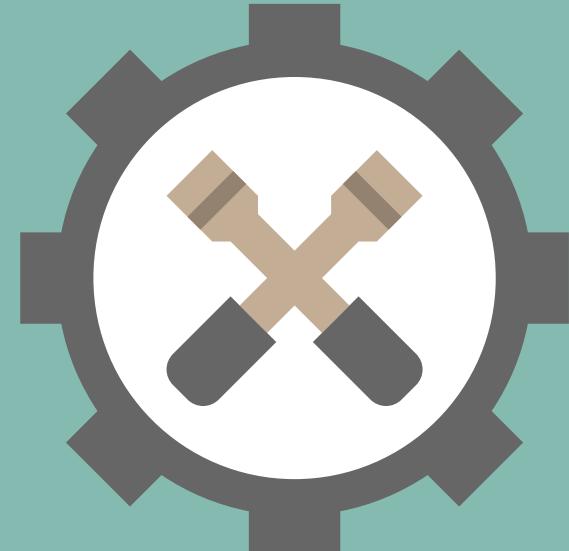


DISTANCE DE COOK



EXERCICE

1. Importer le fichier « Data.csv »
 1. Vérifier le format des champs et le bon import des données
2. La variable à prédire est « Winner »
 1. Construire le meilleur modèle possible



SOMMAIRE

- Pourquoi la régression ?
- Régression linaire
- ANOVA
- ANCOVA
- Sélection de variables
- Régression logistique
- Autres méthodes de prédiction



Autres méthodes

► Des algorithmes pour chaque problèmes :

- Interaction (statistiques)
- Régression linéaire
- Régression linéaire multiple
- Régression polynomiale
- Régression logistique
- Modèle linéaire généralisé
- Régression non paramétrique
- Modèles de régression multiple postulés et non postulés
- Random Forest
- Régression LASSO
- Régréssion Ridge
- Gradient boosting
- Etc



PROJET

CAS D'USAGES



➤ Objectif: Prédire le résultat du match UFC

- Kaggle : <https://www.kaggle.com/rajeevw/ufcdata>
- Jeux d'apprentissage : 5 K lignes / 150 colonnes



➤ Objectif: Prédire la note d'un joueur de football / basket

- Kaggle : https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset#players_20.csv
- Jeux d'apprentissage : 18 K lignes / 89 colonnes



➤ Objectif : Construire le meilleur moteur d'estimation immobilière

- Challenge kaggle en cours : <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>
- Jeux d'apprentissage : 1421 lignes / 81 variables



➤ Objectif : Prédire les maladies cardiaques

- <https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression>
- Jeux d'apprentissage : 4240 lignes / 16 variables

Règles

- Fonctionnement en binôme
- Technologie :
 - R ou python
- Restitution le 24 avril
 - 13h30 – 17h
 - 15 minutes de présentation par groupe / 5 minutes de questions
- Livrables
 - Code
 - Présentation
 - Dossier (facultatif)



Question ?

