# [Réalisez des modélisations de données performantes](https://openclassrooms.com/fr/courses/4525326-realisez-des-modelisations-de-donnees-performantes)

Table des matières

[Réalisez des modélisations de données performantes 1](#_Toc57880410)

[**Appréhendez les différents types de modélisation** 6](#_Toc57880411)

[**Les différents types de modélisation** 6](#_Toc57880412)

[**La régression** 6](#_Toc57880413)

[**La classification (supervisée)** 6](#_Toc57880414)

[**L'analyse de la variance** 7](#_Toc57880415)

[**Découvrez le jeu de données de l'ozone** 8](#_Toc57880416)

[**Le jeu de données de l'ozone** 8](#_Toc57880417)

[**Objectif de notre analyse** 8](#_Toc57880418)

[**Découvrez le jeu de données des maladies cardio-vasculaires** 10](#_Toc57880419)

[**Le jeu de données des maladies cardio-vasculaires** 10](#_Toc57880420)

[**Objectif de notre analyse** 10](#_Toc57880421)

[**Découvrez le jeu de données du blé** 11](#_Toc57880422)

[**Le jeu de données du blé** 11](#_Toc57880423)

[**Objectif de notre analyse** 11](#_Toc57880424)

[**Appréhendez le fonctionnement de la régression linéaire** 12](#_Toc57880425)

[**Le modèle** 12](#_Toc57880426)

[**Les données** 13](#_Toc57880427)

[**Pour aller plus loin : l'écriture matricielle** 13](#_Toc57880428)

[Appliquez la méthode des Moindres Carrés Ordinaires 15](#_Toc57880429)

[L'estimateur des Moindres Carrés Ordinaires 15](#_Toc57880430)

[La solution obtenue 16](#_Toc57880431)

[La droite de régression 16](#_Toc57880432)

[Valeurs ajustées et résidus 17](#_Toc57880433)

[Pour aller plus loin : les propriétés statistiques des paramètres 17](#_Toc57880434)

[La variance résiduelle 18](#_Toc57880435)

[Pour aller plus loin : l'interprétation géométrique 18](#_Toc57880436)

[**Calculez le coefficient de détermination** 20](#_Toc57880437)

[**Le risque de surinterpréter** 20](#_Toc57880438)

[Testez le modèle linéaire gaussien simple 22](#_Toc57880439)

[Testez la significativité de β1 et β2 22](#_Toc57880440)

[Obtenez des intervalles de confiance pour β1 et β2 23](#_Toc57880441)

[**TP : Pratiquez la régression linéaire sur le jeu de données de l'ozone** 24](#_Toc57880442)

[**Importez les données** 24](#_Toc57880443)

[**Visualisez le jeu de données** 24](#_Toc57880444)

[**Réalisez une régression linéaire simple** 26](#_Toc57880445)

[**Visualisez la droite de régression** 27](#_Toc57880446)

[**Représentez les résidus du modèle** 29](#_Toc57880447)

[**Prévoyez la concentration d'ozone** 30](#_Toc57880448)

[Entraînez-vous : déterminez la hauteur d'un arbre à l'aide d'une régression 32](#_Toc57880449)

[À vous de jouer ! 32](#_Toc57880450)

[Vérifiez votre travail 32](#_Toc57880451)

[**Appréhendez le fonctionnement de la régression linéaire multiple** 34](#_Toc57880452)

[**Les données** 34](#_Toc57880453)

[**Régression linéaire multiple avec ou sans constante** 35](#_Toc57880454)

[**Linéarisation de modèles de régression** 35](#_Toc57880455)

[**Appliquez la méthode des Moindres Carrés Ordinaires** 36](#_Toc57880456)

[**L'estimateur des MCO** 36](#_Toc57880457)

[**La solution obtenue** 36](#_Toc57880458)

[**Valeurs ajustées et résidus** 36](#_Toc57880459)

[**Propriétés statistiques des paramètres** 37](#_Toc57880460)

[**La variance résiduelle** 37](#_Toc57880461)

[**Pour aller plus loin : l'interprétation géométrique** 37](#_Toc57880462)

[**Calculez le coefficient de détermination** 39](#_Toc57880463)

[**Le coefficient de détermination** 39](#_Toc57880464)

[Testez le modèle linéaire gaussien multiple 41](#_Toc57880465)

[Testez le modèle 41](#_Toc57880466)

[Testez la significativité d'un des paramètres 42](#_Toc57880467)

[Analysez les résultats 44](#_Toc57880468)

[Analysez l'atypicité des observations 44](#_Toc57880469)

[Analysez l'influence des observations 45](#_Toc57880470)

[Détectez les problèmes de colinéarité 46](#_Toc57880471)

[Analysez l'homoscédasticité des résidus 48](#_Toc57880472)

[Sélectionnez automatiquement un modèle 49](#_Toc57880473)

[Quels critères statistiques pouvez-vous utiliser ? 49](#_Toc57880474)

[Déterminez les régresseurs adaptés 49](#_Toc57880475)

[Découvrez les procédures itératives de choix 50](#_Toc57880476)

[TP : Pratiquez la régression linéaire multiple sur le jeu de données de l'ozone 52](#_Toc57880477)

[Pour aller plus loin : analysez vos résultats 55](#_Toc57880478)

[Entraînez-vous : améliorez les prévisions de hauteur des arbres 63](#_Toc57880479)

[À vous de jouer ! 63](#_Toc57880480)

[Corrigez votre travail 63](#_Toc57880481)

[**Appréhendez le fonctionnement de la régression logistique** 65](#_Toc57880482)

[**Retour au jeu de données des maladies cardio-vasculaires** 65](#_Toc57880483)

[**Estimez un modèle de régression logistique** 68](#_Toc57880484)

[**Cas des variables explicatives qualitatives** 69](#_Toc57880485)

[**Comment estimer en régression logistique ?** 70](#_Toc57880486)

[**Validation du modèle** 70](#_Toc57880487)

[**Analysez les résultats** 71](#_Toc57880488)

[**Comment interpréter les paramètres ?** 71](#_Toc57880489)

[**Choisissez le modèle adéquat** 72](#_Toc57880490)

[**Comment prévoir "Y" ?** 72](#_Toc57880491)

[**Appréhendez la courbe ROC** 73](#_Toc57880492)

[TP : Pratiquez la régression logistique sur le jeu de données des maladies cardio-vasculaires 74](#_Toc57880493)

[**Importez les données** 74](#_Toc57880494)

[**Visualisez le nuage de points** 75](#_Toc57880495)

[**Calculez les proportions de malades** 75](#_Toc57880496)

[**Effectuez la régression logistique** 77](#_Toc57880497)

[Avez-vous compris les enjeux de la régression logistique ? 83](#_Toc57880498)

[Compétences évaluées 83](#_Toc57880499)

[ Question 1 83](#_Toc57880500)

[ Question 2 83](#_Toc57880501)

[ Question 3 83](#_Toc57880502)

[ Question 4 84](#_Toc57880503)

[ Question 5 84](#_Toc57880504)

[ Question 6 84](#_Toc57880505)

[ Question 7 85](#_Toc57880506)

[ Question 8 85](#_Toc57880507)

[**Appréhendez le fonctionnement de l'analyse de la variance (ANOVA)** 86](#_Toc57880508)

[**Réalisez une analyse de la variance** 87](#_Toc57880509)

[**Réalisez une ANOVA à un facteur** 87](#_Toc57880510)

[**Pour aller plus loin : l'estimation du modèle** 87](#_Toc57880511)

[**Réalisez le test d'influence d'une variable qualitative** 88](#_Toc57880512)

[**Réalisez une ANOVA à deux facteurs** 90](#_Toc57880513)

[**Les différents tests d'influence** 90](#_Toc57880514)

[TP : Pratiquez l'analyse de la variance sur le jeu de données du blé 92](#_Toc57880515)

[Importez les données 92](#_Toc57880516)

[Réalisez une ANOVA à 1 facteur 93](#_Toc57880517)

[Réalisez une ANOVA à 2 facteurs 96](#_Toc57880518)

[Avez-vous compris les enjeux de l'ANOVA ? 98](#_Toc57880519)

[Compétences évaluées 98](#_Toc57880520)

[ Question 1 98](#_Toc57880521)

[ Question 2 98](#_Toc57880522)

[ Question 3 98](#_Toc57880523)

[ Question 4 99](#_Toc57880524)

[ Question 5 99](#_Toc57880525)

[ Question 6 101](#_Toc57880526)

Vous aimeriez prédire le futur ?

Savez-vous que les Data analysts en sont capables, grâce à des outils simples ?

Eh bien oui, les**Data analysts** et **Data scientists** peuvent **prévoir** les pics de pollution, ou encore le taux de guérison de patients. Voici des exemples de problématiques que vous saurez aborder à l’issue de ce cours !

Dans ce cours, vous appréhenderez le fonctionnement d'une **modélisation**. Vous serez capable de mettre en œuvre différentes techniques, parmi les plus populaires pour le Data analyst. Vous avez déjà entendu parler de la **régression linéaire.**Mais vous irez encore plus loin, avec**la classification**, ou encore la fameuse **ANOVA**.

Avec ces techniques, vous serez outillé pour prévoir le futur et pour aider vos collaborateurs à prendre des décisions informées.

**Objectifs pédagogiques :**

* Appliquer et interpréter une **régression linéaire simple**
* Appliquer et interpréter une **régression linéaire multiple**
* Appliquer et interpréter une **régression logistique**
* Appliquer et interpréter une **analyse de la variance (ANOVA)**

**Prérequis :**

Ce cours fait partie du parcours [**Data analyst**](https://openclassrooms.com/paths/data-analyst). Il se situe au croisement des mathématiques et de l'informatique. Pour en profiter pleinement, n'hésitez pas à vous rafraîchir la mémoire, avant ou pendant le cours, sur :

* [**Les bases des statistiques**](https://openclassrooms.com/fr/courses/4525266-decrivez-et-nettoyez-votre-jeu-de-donnees)
* [**Les tests statistiques**](https://openclassrooms.com/fr/courses/4525306-initiez-vous-a-la-statistique-inferentielle)
* [**Le langage R**](https://openclassrooms.com/fr/courses/1393696-effectuez-vos-etudes-statistiques-avec-r)

**Appréhendez les différents types de modélisation**

Une modélisation statistique consiste à établir une relation entre variables, sous forme d'équation, que l'on estime sur un jeu de données observées. L'enjeu est d'utiliser cette relation, établie et vérifiée, sur des observations, à des fins de prévision : on se trouve ici dans le cadre de **l'inférence.**

**Les différents types de modélisation**

Voici les différents types de modélisation que vous retrouverez le plus souvent. Nous allons les étudier ensemble dans ce cours.

1. **La régression linéaire,** qui permet d'expliquer une variable **quantitative** à partir de **variables explicatives quantitatives** (éventuellement qualitatives en sus).
2. **La classification supervisée,** qui permet d'expliquer une variable **qualitative** à partir de **variables explicatives quantitatives**(éventuellement qualitatives en sus). Attention, il faut la distinguer de la classification non supervisée qu'est le clustering.
3. **L'analyse de la variance,** pour analyser l'influence d'une ou deux **variables explicatives qualitatives** sur une variable **quantitative**.

**La régression**

Le terme *régression* a été introduit par Francis Galton, chercheur britannique du XIXe siècle. Il décrivait dans un article scientifique le fait que la taille des enfants nés de parents inhabituellement grands ou petits se rapproche de la taille moyenne de la population.

La **régression** désigne désormais **toute méthode statistique qui permet de mettre en relation une variable quantitative,** que l'on cherche à expliquer et/ou prévoir, avec un ensemble de variables quantitatives (potentiellement) explicatives.

Le modèle de régression le plus connu est le **modèle linéaire, simple ou multiple,** mais il existe de nombreux autres modèles, parmi lesquels les modèles paramétriques non linéaires et les modèles non paramétriques ou semi-paramétriques. Citons aussi les régressions Ridge et Lasso, les modèles GAM (additifs généralisés), les arbres de régression (et les forêts aléatoires), etc. Pour ce cours, nous nous cantonnerons au modèle linéaire, simple et multiple.

**La classification (supervisée)**

Lorsque la variable à expliquer, à prévoir, est qualitative, on parle alors de **classification (supervisée)**. Si la **régression logistique** est bien le modèle *de base*, on trouve également les arbres de décision (et les forêts aléatoires), les SVM, etc.

Ces méthodes de classification permettent d'effectuer du **scoring** (pour l'attribution d'un prêt bancaire, par exemple) ou encore d'évaluer la probabilité qu'un individu statistique appartienne à une catégorie (qu'il guérisse ou non à la suite de la prise d'un médicament).

Attention, le terme *régression logistique* pourrait laisser penser qu'il s'agit d'une méthode qui s'applique sur données quantitatives, ce qui n'est pas le cas. Cette méthode est en fait liée à la régression linéaire via une fonction appelée *logit*, ce qui explique son nom !

**L'analyse de la variance**

L'analyse de la variance (terme souvent abrégé par l'anglais **ANOVA** : **AN**alysis **O**f **VA**riance) est un modèle statistique utilisé pour comparer les moyennes d'échantillons selon les modalités d'une variable qualitative.

*Vous avez eu un aperçu des trois modèles que vous allez étudier dans ce cours. À présent, voyons les trois jeux de données sur lesquels vous allez travailler.*

**Découvrez le jeu de données de l'ozone**

**L'ozone** est un **polluant photochimique,** dont de nombreux instituts cherchent à prévoir les pics pour prévenir les populations. La pollution automobile, l'absence de vent et la chaleur comptent parmi les facteurs qui accroissent le taux de pollution.

**Le jeu de données de l'ozone**

Téléchargez le jeu de données que nous allons étudier à [**cette adresse**](https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/parcours-data-analyst/Cours_realisez_des_modelisations_performantes/NOUVEAU/Dataset_ozone.txt).

On considère à cet effet un jeu de données issu du Laboratoire de mathématiques appliquées de l'Agrocampus Ouest qui contient 112 données recueillies à Rennes durant l'été 2001. On y trouve les 14 variables suivantes :

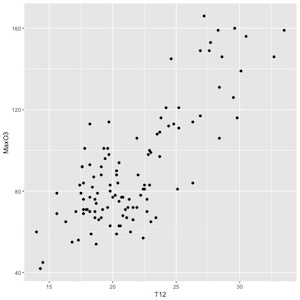
* *obs :* mois-jour ;
* *maxO3 :* teneur maximale en ozone observée sur la journée (en μ\gr/m3μ\gr/m3 ) ;
* *T9, T12, T15 :* température observée à 9 h, 12 h et 15 h ;
* *Ne9, 12, Ne15 :* nébulosité observée à 9 h, 12 h et 15 h ;
* *Vx9, Vx12, Vx15 :* composante est-ouest du vent à 9 h, 12 h et 15 h ;
* *maxO3v :* teneur maximale en ozone observée la veille ;
* *vent :* orientation du vent à 12 h ;
* *pluie :* occurrence ou non de précipitations.

**Objectif de notre analyse**

On souhaite étudier le lien entre le pic d'ozone journalier et un certain nombre de facteurs potentiellement explicatifs afin de proposer un modèle de régression permettant de prévenir la population.

Par exemple, la simple visualisation du nuage de points entre le pic d'ozone journalier et la température mesurée à 12 h permet d'ores et déjà d'envisager une relation linéaire :





Ozone nuage de points

Sur ce jeu de données, on voit déjà qu'une **régression linéaire** sera une modélisation intéressante. Parfait !

*Voyons maintenant notre deuxième jeu de données, portant sur des maladies cardio-vasculaires...*

**Découvrez le jeu de données des maladies cardio-vasculaires**

On dispose d'informations médicales sur 462 patients mâles dans une région à risque élevé de maladies cardiaques du Cap-Occidental (Afrique du Sud).

**Le jeu de données des maladies cardio-vasculaires**

On considère à cet effet un jeu de données issu d'un article dans le *South African Medical Journal* (Rousseauw et al, 1983).

Téléchargez le jeu de données que nous allons étudier à [**cette adresse**](https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/parcours-data-analyst/Cours_realisez_des_modelisations_performantes/NOUVEAU/Dataset_Maladie_Cadiovasculaire.txt).

On y trouve les 12 variables suivantes :

* *sbp :* tension artérielle systolique ;
* *tobacco :* tabac cumulé (en kg) ;
* *ldl :* cholestérol de lipoprotéines de faible densité ;
* *adiposity :* adiposité ;
* *famhist :* antécédents familiaux ;
* *typea :* comportement type A ;
* *obesity :* obésité ;
* *alcohol :* consommation courante d'alcool ;
* *age :* âge au moment de l'attaque cardiaque ;
* *chd :* maladie coronarienne.

**Objectif de notre analyse**

Notre objectif sera d'appréhender les facteurs pouvant influencer l'apparition d'une maladie coronarienne à des fins de prévision.

Dans ce cas, nous allons utiliser une**régression logistique**.

*Très bien, voyons enfin notre dernier jeu de données, portant sur des variétés de blé.*

**Découvrez le jeu de données du blé**

On dispose ici de rendements de blé observés sur 80 parcelles homogènes et éloignées les unes des autres. Chacune des 4 variétés de blé considérées a été plantée sur 10 parcelles avec traitement phytosanitaire et sur 10 autres parcelles sans aucun traitement.

**Le jeu de données du blé**

Téléchargez le jeu de données que nous allons étudier à [**cette adresse**](https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/parcours-data-analyst/Cours_realisez_des_modelisations_performantes/NOUVEAU/Dataset_Ble.txt).

Le jeu de données contient les 3 variables suivantes :

* *rdt :* rendement de blé (en quintaux par hectare) ;
* *ble :* variété de blé (A, B, C, D) ;
* *phyto :* traitement phytosanitaire (1 si positif, 0 sinon).

**Objectif de notre analyse**

Notre objectif est d'évaluer ici la sensibilité du rendement en fonction de la variété de blé, et éventuellement en fonction du couple variété de blé-traitement phytosanitaire.

Pour y répondre, nous réaliserons une **analyse de la variance**.

*Vous avez eu un aperçu des jeux de données que nous allons utiliser dans ce cours. Sans plus tarder, étudions en détail notre premier modèle : la régression linéaire simple.*

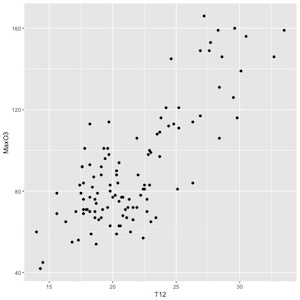
**Appréhendez le fonctionnement de la régression linéaire**

**Le modèle**

La **régression linéaire simple** permet d'expliquer, de manière linéaire, une variable YY**(variable à expliquer)**, aléatoire en fonction d'une **variable explicative**XX (on la nomme parfois régresseur ou covariable).

Le modèle de régression linéaire simple suppose, comme son nom l'indique, qu'il existe une relation linéaire entre la variable à expliquer et la variable explicative :

Y=β1+β2X+ε



Nuage de points du pic d'ozone

"Epsilon" désigne l'erreur de modélisation commise : la relation linéaire est considérée *presque parfaite*, à un epsilon près. Vous pouvez l'observer sur le nuage de points ci-dessus qui n'est pas *parfait*.

Mathématiquement, on considère que :

* Y est une variable aléatoire, observable ;
* X est une variable déterministe (non aléatoire), observable ;
* β1 et β2 sont des **paramètres**inconnus (non observables) ;
* ε est une variable aléatoire centrée (autour de 0) de variance σ2σ2 inconnue (c'est également un paramètre du modèle).

L'objectif est de déterminer, d'estimer les paramètres de la droite de régression : l'ordonnée à l'origine β1 et le coefficient directeur β2.

**Dans le cas de l'ozone**, on considérera le pic d'ozone journalier pour YY et la température à 12 h. On sera ainsi en mesure de prédire ce pic d'ozone journalier uniquement en fonction de la température à midi.

**Les données**

On considère ici que l'on dispose de **n** observations d'un échantillon i.i.d de (X,Y).

Pour rappel, **"i.i.d"** signifie : « variables indépendantes et identiquement distribuées ».

Dans le cas de l’ozone (où **n=104**), **xi** est la température à 12 h pour le jour **i** et **yi** le pic d'ozone journalier.

On a donc, selon le modèle de régression posé précédemment :



Les erreurs  vérifient pour  :

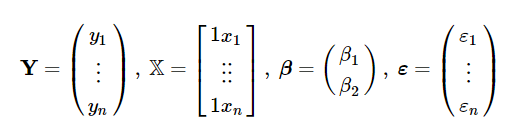
*  (elles sont centrées autour de 0) ;
*  (leur variance, inconnue, est constante et égale à ) ;
*  (elles n'ont pas de dépendance linéaire).

**Pour aller plus loin : l'écriture matricielle**

Matriciellement, on peut écrire :



où :



Cette écriture sera très utilisée dans le cas où l'on disposera non pas d'une, mais de plusieurs variables explicatives.

*Dans les prochains chapitres, nous allons entrer dans la théorie. Si vous souhaitez un aperçu d'une prévision à l'aide d'une régression linéaire, consultez ce*[*TP*](https://openclassrooms.com/fr/courses/4525326-realisez-des-modelisations-de-donnees-performantes/5754137-pratiquez-la-regression-lineaire-sur-le-jeu-de-donnees-de-lozone)*.*

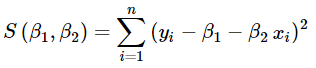
## Appliquez la méthode des Moindres Carrés Ordinaires

### 

### L'estimateur des Moindres Carrés Ordinaires

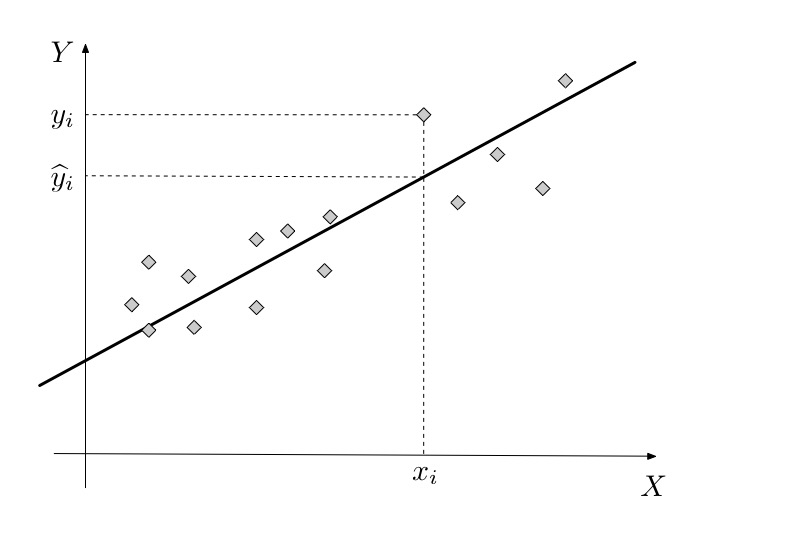
β1 et β2 sont des paramètres inconnus non observables, que l'on cherche à estimer. Il existe plusieurs méthodes pour cela, mais la plus utilisée est celle des **MCO**.

On appelle**estimateur des moindres carrés ordinaires (MCO)** de β1 et β2 les valeurs  et  minimisant la somme des carrés des résidus :



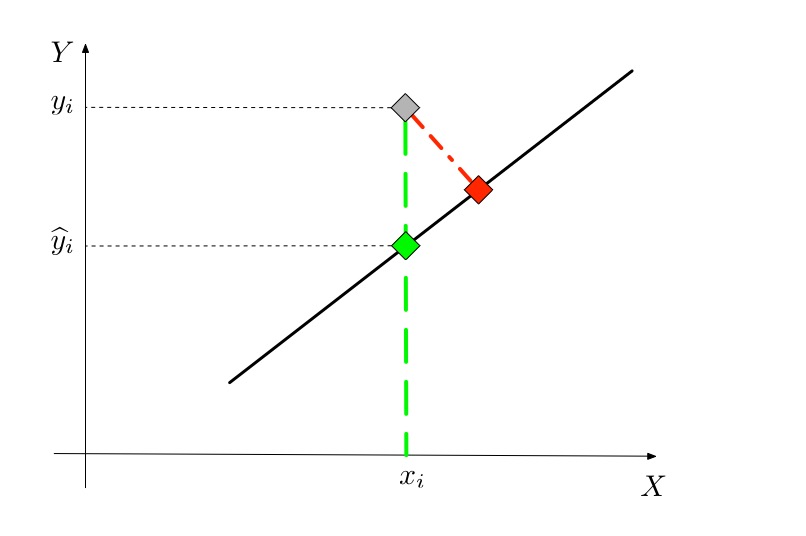
Si, comme la somme des valeurs absolues, la somme des carrés est toujours positive (et nulle si le modèle est parfait), elle présente en sus l'intérêt d'être dérivable, ce qui est plus simple pour déterminer le minimum.

En notant , on peut tracer la droite de régression suivante :



Droite de régression

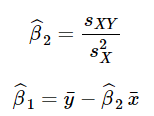
Notons que la distance minimisée avec les MCO est   (en vert), pas la distance du point à la droite de régression (en rouge) :



Distance des moindres carrés

### La solution obtenue

Dans le cas où au moins un des **xi** diffère des autres (ce qui est toujours le cas en pratique), les estimateurs des MCO de (β1,β2) valent :



Remarquons que le coefficient directeur de la droite  est proportionnel à la covariance empirique entre X et Y , qui est, rappelons-le, une mesure de la dépendance linéaire entre les variables.

### La droite de régression

L'équation de la **droite de régression** est :



On peut montrer que cette droite passe par le barycentre du nuage de points 

### Valeurs ajustées et résidus

Pour l'observation **i**, on appelle **valeur ajustée** (ou valeur estimée) la quantité :



On appelle **résidu** la différence entre la valeur observée pour la variable à expliquer et son estimation. Il représente la partie inexpliquée par le modèle. Le résidu, pour l'individu ii , est donc :

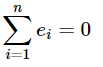
ei=yi−yˆiei=yi−y^i

Les résidus, dépendant des paramètres estimés, sont calculables, à la différence du bruit qui dépend des paramètres inconnus :



Le résidu **ei** est une estimation du bruit **εi**. Il représente la partie non expliquée par le modèle pour l'individu **i**.

On peut montrer que la somme des résidus est nulle :



### 

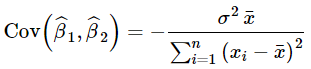
### Pour aller plus loin : les propriétés statistiques des paramètres

On peut montrer que et  sont des estimateurs sans biais de β1 et β2 :



Cela signifie qu'en moyenne, l'estimateur des MCO nous conduira à la bonne solution.

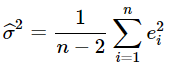
Une petite remarque ici, les deux estimateurs  et  ne sont pas indépendants, ils sont même linéairement dépendants (covariance non nulle) :



On peut d'autant plus être confiant dans la qualité de ces estimateurs qu'ils sont dit **BLUE** (**B**est **L**inear **U**nbiased **E**stimators) : parmi tous les estimateurs linéaires et sans biais de β1β1 et β2β2 , les estimateurs des MCO de et  sont de variance minimale.

### La variance résiduelle

La **variance résiduelle** vaut :



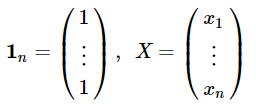
C'est un estimateur sans biais de .

### Pour aller plus loin : l'interprétation géométrique

On peut réécrire :



où



On a :



où désigne la norme euclidienne.

En notant , on a :



 est la projection orthogonale de **Y** sur le sous-espace vectoriel engendré par les vecteurs **1n** et **X**.

Les estimateurs  et  sont donc les coordonnées de la projection de **Y** dans cet espace.

Vous avez découvert la méthode des Moindres Carrés Ordinaires. Voyons maintenant comment calculer le coefficient de détermination.

**Calculez le coefficient de détermination**

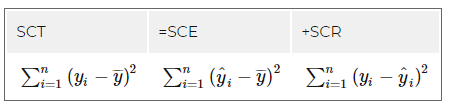
**L'intérêt d'un modèle de régression linéaire** réside dans sa capacité à **expliquer une partie des variations de la variable Y** par les variations de la variable **X**. La variation d'une variable **Y** est obtenue en considérant les différences entre les valeurs observées **yi** et leur moyenne .

Or on a :



où  est la variation expliquée (ou restituée) par le modèle (on a ), alors que  est la variation non expliquée par le modèle.

On peut établir la formule de décomposition de la variance (**ANOVA** : **AN**alysis **O**f **VA**riance) :



* SCT (Somme des Carrés Totale) traduit la variation totale de **Y**.
* SCE (Somme des Carrés Expliquée) traduit la variation expliquée par le modèle.
* SCR (Somme des Carrés Résiduelle) traduit la variation inexpliquée par le modèle.

On appelle **coefficient de détermination** la quantité suivante :



Ce coefficient est dans **[0,1]**, puisque :



Si , on a alors SCE=SCT : toute la variation est expliquée par le modèle.

Si , on a alors SCR=SCT: aucune variation n'est expliquée par le modèle.

Dans le cas de la régression linéaire simple, on obtient :

   
où est la corrélation linéaire (empirique) entre X et Y.

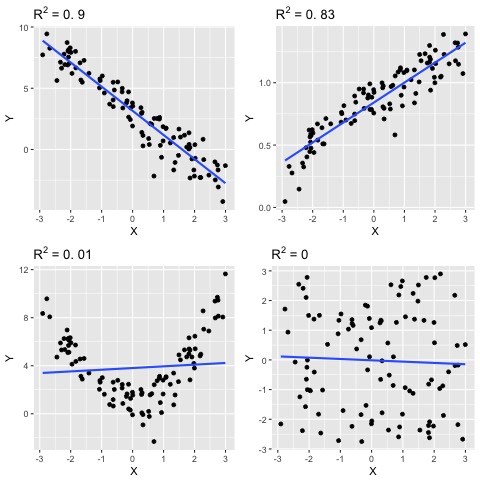
**Le risque de surinterpréter**

Il faut veiller à ne pas surinterpréter le coefficient de détermination :

* Un bon ajustement linéaire se traduit par un  proche de 1.
* *A contrario,* un  proche de 1 ne traduit pas forcément un lien linéaire.
* Un  proche de 0 traduit un mauvais ajustement linéaire, mais n'implique pas qu'aucune relation ne puisse être établie entre les variables.

Pour illustration, la figure suivante présente :

* En haut à gauche : un cas de dépendance linéaire entre deux variables pour lequel le coefficient de détermination est proche de 1.
* En haut à droite : un cas de dépendance, non linéaire, de forme « {racine carrée} », entre deux variables, pour lequel le coefficient de détermination est proche de 1.
* En bas à gauche : un cas de dépendance, non linéaire, parabolique, entre deux variables, pour lequel le coefficient de détermination est proche de 0.
* En bas à droite : un cas d'indépendance entre deux variables pour lequel le coefficient de détermination est proche de 0.



Exemples de coefficients de détermination

*Vous savez maintenant comment calculer le coefficient de détermination. Nous allons à présent tester notre modèle.*

## Testez le modèle linéaire gaussien simple

À ce stade, la régression linéaire a livré toutes ses possibilités.

Nous avons estimé β1, β2 et , et nous sommes également en mesure de calculer le coefficient de détermination.

Mais nous ne pouvons pas tester les paramètres ni établir d'intervalles de confiance sur ces paramètres.

Pour pouvoir le faire, nous allons donc ajouter une hypothèse de loi : l'**hypothèse gaussienne,** c'est-à-dire que nous considérons que ϵϵ suit une loi normale .

Cette loi normale nous permettra notamment de tester la significativité des paramètres.

Si nous reprenons le modèle de la régression linéaire simple entre le pic d'ozone et la température à midi, nous allons pouvoir tester si la température à midi est **significative** pour expliquer le pic d'ozone.

En d'autres termes, on va chercher à savoir si le coefficient β2 (le coefficient qui est placé devant la température à midi) est significatif, plus précisément s'il est significativement différent de 0.

Avec cette hypothèse, on parle de **modèle linéaire gaussien simple.**

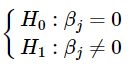
Dans le modèle linéaire gaussien simple, on considère, en sus des hypothèses formulées dans le cadre du modèle linéaire simple :



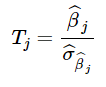
Dans ce cadre,  est un échantillon i.i.d de loi .

### Testez la significativité de β1 et β2

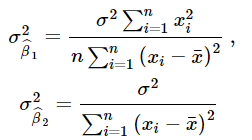
Pour **j∈{1,2}**, on teste :



On utilise comme statistique de test :



où :



Plus cette quantité est grande, plus on est enclin à rejeter l'hypothèse de nullité du paramètre.

On peut montrer que, sous  admet comme loi la loi de Student à n−2 degrés de liberté.



On décide du **rejet** de  au niveau de test **α** si



où désigne le quantile d'ordre de la loi .

En pratique, dire que l'on rejette cette hypothèse  pour β2 revient à conserver la variable X comme explicative.

### Obtenez des intervalles de confiance pour β1 et β2

Pour **j∈{1,2}**, le paramètre **βj** admet comme intervalle de confiance de niveau **1−α** :



Il est également possible d'établir une région de confiance simultanée des deux paramètres **β1** et **β2** (via une loi de Fisher), ainsi qu'un intervalle de confiance pour  (via une loi du khi-deux).

Et voilà pour le test de notre modèle. Maintenant, passons à la pratique. Nous allons appliquer une régression linéaire simple sur le jeu de données de l'ozone.

**TP : Pratiquez la régression linéaire sur le jeu de données de l'ozone**

Ce TP est présenté en R. Voici la [**version Python**](https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/Parcours+Data+Analyst/Re%CC%81alisez+des+mode%CC%81lisations+de+donne%CC%81es+performantes+(1).zip).

Avant de commencer, retrouvez la description du cas d'étude et du jeu de données dans [**ce chapitre**](https://openclassrooms.com/fr/courses/4525326-realisez-des-modelisations-de-donnees-performantes/5754127-decouvrez-le-jeu-de-donnees-de-lozone?status=draft).

Mettons maintenant en œuvre la régression linéaire sur le pic d'ozone, expliqué par la température à midi. Voyons la phase d'estimation du modèle, puis mettons en œuvre une prévision.

**Importez les données**

On commence par charger la librairie ggplot2, qui permettra d'afficher les graphiques :

library(ggplot2)

On importe ensuite le fichier [ozone](https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/parcours-data-analyst/Cours_realisez_des_modelisations_performantes/NOUVEAU/Dataset_ozone.txt), qui contient 112 données recueillies dans la ville de Rennes durant l'été 2001.

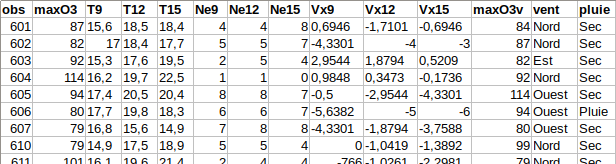
On trouve dans ce fichier des variables telles que :

* MaxO3, qui est la valeur maximale d'ozone observée sur une journée ;
* T9, T12 et T15 qui sont les températures prises respectivement à 9 h, 12 h et 15 h ;
* Ne9, Ne12, Ne15 qui sont des nébulosités prises à 9 h, 12 h et 15 h ;
* Vx9, Vx12 et Vx15 qui sont les composantes est-ouest du vent mesurées à 9 h, 12 h et 15 h ;
* MaxO3V, qui donne la teneur maximale en ozone observée la veille ;
* vent, l'orientation du vent à 12 h ;
* pluie, la présence ou non de pluie.

**Visualisez le jeu de données**

Visualisons l'ensemble des données avec cette commande.

ozone <- read.table("ozone.txt", header=TRUE, sep=";", dec=",")



Visualisons notre jeu de données

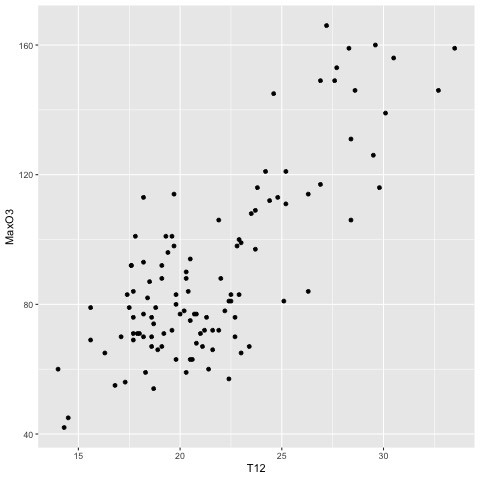
On peut représenter graphiquement le nuage de points maxO3 en fonction de T12 :

ggplot(ozone, aes(x=T12, y=maxO3))+

geom\_point()+

xlab("T12")+

ylab("maxO3")



Nuage de points de l'ozone

Ce nuage de points nous fait penser à un alignement selon une forme qui n'est pas très loin d'une droite.

**Réalisez une régression linéaire simple**

Essayons de lancer une régression linéaire simple sur ce nuage de points :

reg\_simp <- lm(maxO3~T12,data=ozone)

Voici les résultats en sortie de cette commande :

summary(reg\_simp)

Call:

lm(formula = maxO3 ~ T12, data = ozone)

Residuals:

Min 1Q Median 3Q Max

-38.079 -12.735 0.257 11.003 44.671

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 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

Nous obtenons des statistiques sur les résidus, avec le minimum, le maximum et les 3 quartiles, ainsi que des statistiques sur les coefficients obtenus : leur valeur, leur écart-type, la statistique de test de Student, et la p-valeur (le test effectué sur le paramètre est ici le test de significativité : le paramètre vaut 0 versus le paramètre est différent de 0).

Les p-valeurs sont inférieures à 5 %. À un niveau de test de 5 %, on rejette donc l'hypothèse selon laquelle le paramètre est égal à 0 : les paramètres sont donc significativement différents de 0.  
Ici, on voit que la variable T12 est significative.

Quant au , il est de l'ordre de 0.6. Ce n'est pas très élevé, mais ceci est logique au vu de la dispersion du nuage de points originel.

**Visualisez la droite de régression**

Alors, voyons à quoi ressemble notre droite.

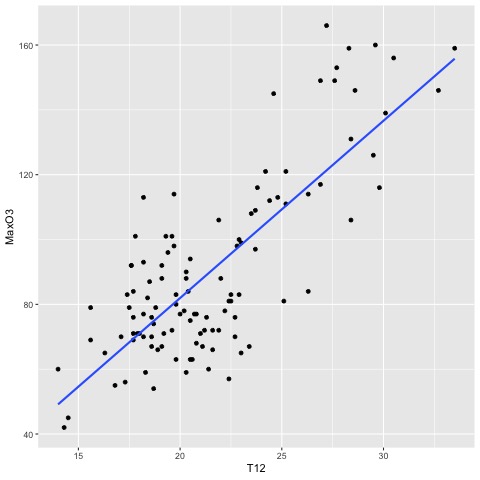
ggplot(ozone,aes(x=T12,y=maxO3))+

geom\_point()+

stat\_smooth(method="lm",se=FALSE)+

xlab("T12")+

ylab("MaxO3")



La droite de régression passe à travers le nuage de points.

On peut également représenter les valeurs ajustées en fonction des valeurs observées :

ozone$maxO3\_ajust\_s <- reg\_simp$fit

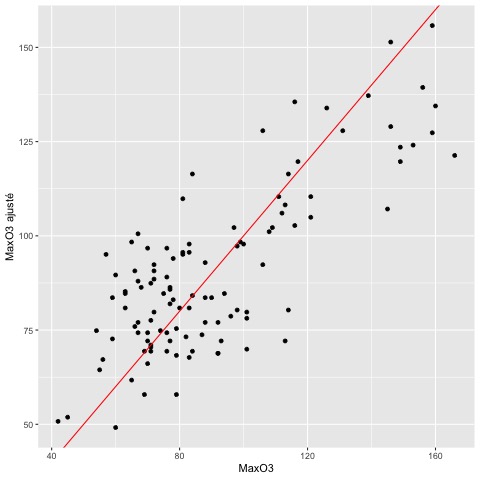
ggplot(ozone, aes(x=maxO3,y=maxO3\_ajust\_s))+

geom\_point()+

geom\_abline(intercept=0,slope=1,color="red")+

xlab("MaxO3")+

ylab("MaxO3 ajusté")



Première bissectrice

La droite qui s'affiche est la première bissectrice. Si le modèle était parfait,  
les valeurs réelles et les valeurs ajustées seraient égales, donc sur un tel graphique, les points seraient alignés sur la droite d'équation **y=x**, soit la première bissectrice.

**Représentez les résidus du modèle**

On peut obtenir les résidus du modèle à l'aide de cette commande :

ozone$residu\_s <- reg\_simp$residuals

À partir de ceux-ci, on peut représenter l'histogramme de ces résidus.

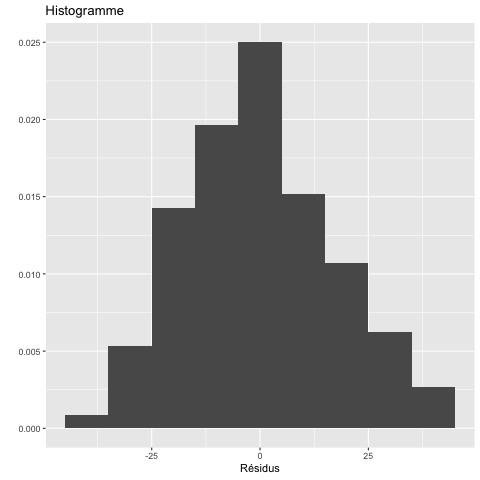
ggplot(ozone,aes(x=residu\_s))+

geom\_histogram(binwidth=10,aes(y=..density..))+

ggtitle("Histogramme")+

xlab("Résidus")+

ylab("")



Histogramme des résidus

L'allure de l'histogramme est assez classique : centrée et à peu près symétrique.

**Prévoyez la concentration d'ozone**

Prévoyons maintenant la concentration en ozone d'une journée. Sachant que la température prévue de cette journée est de 19 °C, on peut utiliser notre modèle de régression à des fins de prévision !

a\_prevoir <- data.frame(T12=19)

maxO3\_prev <- predict(reg\_simp,a\_prevoir)

round(maxO3\_prev, digits=2)

On obtient une concentration d'ozone d'environ 76.5.

Jusqu'à maintenant, la régression linéaire ne fait intervenir qu'une seule variable explicative. Plus tard (en partie 3), nous verrons des régressions linéaires plus complexes, avec plus de variables explicatives, comme les autres températures, les vitesses de vent, etc.

*Eh bien, vous avez effectué votre première prédiction grâce à une régression. Félicitations. Le prochain chapitre est un quiz, qui permettra de vous tester sur votre compréhension du modèle.*

## Entraînez-vous : déterminez la hauteur d'un arbre à l'aide d'une régression

### 

### À vous de jouer !

Pour vous entraîner, réalisez cet exercice étape par étape. Une fois terminé, vous pouvez comparer votre travail avec les pistes que je vous propose.

Une scierie souhaite déterminer la hauteur d'un arbre, afin de prévoir ses besoins en matière première. Elle dispose uniquement des données concernant la circonférence des arbres. Vous allez créer un modèle permettant de prévoir la hauteur d'un arbre en fonction de sa circonférence.

#### Votre mission

1. Visualiser le nuage de points (graphique de dispersion de la hauteur en fonction de la circonférence).
2. Effectuer la régression linéaire de la hauteur en fonction de la circonférence.
3. Donner et interpréter le coefficient de détermination.
4. Analyser la significativité des paramètres (on teste la nullité des paramètres au niveau de test 5 %).

#### Les données à votre disposition

La scierie dispose de ce [jeu de données](https://opendata.paris.fr/explore/dataset/les-arbres/table/) qui contient les caractéristiques suivantes pour 138 épicéas :

* circ : circonférence de l'arbre (en cmcm )
* haut : hauteur de l'arbre (en mm )

### Vérifiez votre travail

Vérifiez que votre travail remplit les critères suivants :

1. Il comprend une image montrant le graphique de dispersion avec la hauteur en abscisses et la circonférence en ordonnées. Les images intégrées à un note book sont aussi acceptées.
2. Le code contient une ligne permettant le calcul de la régression linéaire simple. Une autre fonction que celle proposée dans le corrigé peut être acceptée, à condition qu'elle réalise bien une régression linéaire.
3. Le coefficient R2 ou le coefficient R2 ajusté est donné par le code. Toute interprétation est acceptée sauf les extrêmes du type "il n'y a aucun ajustement linéaire" ou "l'ajustement linéaire est parfait".
4. La réponse est correcte si elle est équivalente à : "On rejette la nullité des paramètres au niveau de test 5 %."

Voici [un exemple](https://static.oc-static.com/activities/3163/evaluation_resources/determinez-la-hauteur-dun-arbre-a-laide-dune-regression_exemple-2019-01-08T173544.zip) corrigé pour vous permettre de vérifier votre travail.

**Appréhendez le fonctionnement de la régression linéaire multiple**

On souhaite cette fois expliquer, de manière linéaire, une variable **Y (variable à expliquer)**, aléatoire en fonction de **p** variables **(X1,…,Xp)**, et non plus d'une seule variable.

Extension naturelle du modèle de régression linéaire simple, le modèle de **régression linéaire multiple** suppose que :



où :

* **Y** est une v.a.r, observable ;
* **(X1,…,Xp)** sont déterministes (non aléatoires), observables ;
* **(β1,…,βp)** sont des **paramètres** inconnus (non observables) ;
* **ε**, **l'erreur** du modèle, est une v.a.r centrée de variance  inconnue (c'est également un paramètre du modèle).

Dans le cas de l'ozone, on considérera le pic d'ozone journalier pour **Y** et les différentes variables météorologiques (température, vitesse du vent, etc.), ainsi que le pic d'ozone de la veille pour **X1,…,Xp**.

**Les données**

On considère ici que l'on dispose de nn observations d'un échantillon i.i.d de **(X1,…,Xp,Y)** :



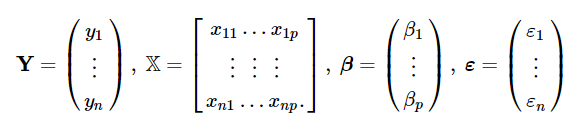
De manière identique à la régression linéaire simple, les erreurs vérifient pour  :

*  (elles sont centrées autour de 0) ;
*  (leur variance, inconnue, est constante et égale à ) ;
*  (elles n'ont pas de dépendance linéaire).

Matriciellement, on peut réécrire le problème sous la forme suivante :



où :



**Régression linéaire multiple avec ou sans constante**

En présence d'un terme constant dans le modèle, on considérera que la première *variable* **X1** est égale à 1 :



On est alors en présence de **p−1***vraies* variables explicatives et de **p** paramètres à estimer (avec en sus  qui reste à estimer quel que soit le cas).

**Linéarisation de modèles de régression**

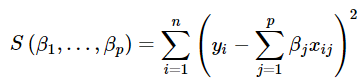
Il est possible de considérer comme variables explicatives des transformations (puissance, exponentielle, logarithme, etc.) de  **X1,…,Xp**.

*Après cette introduction au modèle, continuons avec la méthode des Moindres Carrés Ordinaires, que vous avez découverte dans la partie précédente...*

**Appliquez la méthode des Moindres Carrés Ordinaires**

**L'estimateur des MCO**

On appelle**estimateurs des moindres carrés ordinaires (MCO)** de  le vecteur  minimisant le critère :



**La solution obtenue**

Sous la condition de non-colinéarité des variables (c'est-à-dire qu'il n'existe pas de relation linéaire entre une variable et les **p−1** autres), l'estimateur des MCO de **β** existe.

Son écriture matricielle est :



La matrice X est constituée de l'ensemble des variables observées sur tous les individus. La matrice-colonne Y est donnée par l'ensemble des valeurs **y** observées sur l'ensemble des individus.

Quant au reste, tout fonctionne comme la régression linéaire simple :

* Nous aurons ici aussi des valeurs ajustées, qui sont les valeurs que l'on aurait obtenues si le modèle était parfait.
* Nous avons également des résidus, qui sont les écarts entre les valeurs observées et les valeurs ajustées.

**Valeurs ajustées et résidus**

Les **valeurs ajustées** (ou valeurs estimées) sont obtenues à partir de la formule suivante :



Il s'agit toujours des valeurs que l'on aurait obtenues pour toutes les observations à partir du modèle de régression.

Les **résidus** mesurent toujours les écarts entre les valeurs observées (pour **Y**) et les valeurs estimées :



**Propriétés statistiques des paramètres**

On peut montrer que  est un estimateur sans biais de **β** :

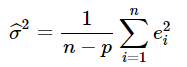


Cela signifie qu'en moyenne, **l'estimateur des MCO nous conduira à la bonne solution.**

Le résultat énoncé dans le cas de la régression linéaire simple reste valable. On peut d'autant plus être confiant dans la qualité de ces estimateurs qu'ils sont dits BLUE (Best Linear Unbiased Estimators) : parmi tous les estimateurs linéaires et sans biais de **β**, l'estimateur des MCO de  est de variance minimale.

**La variance résiduelle**

La **variance résiduelle** vaut :



C'est un estimateur sans biais de .

**Pour aller plus loin : l'interprétation géométrique**

 est la projection orthogonale de Y sur le sous-espace vectoriel engendré par les colonnes de ****.

La matrice de projection sur cet espace, communément notée  (pour *hat*) dans le cadre des modèles linéaires, vaut :



On peut vérifier que l'on a bien :



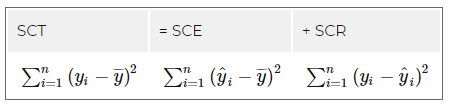
*Vous avez découvert la méthode des Moindres Carrés Ordinaires pour une régression linéaire multiple. Voyons maintenant comment calculer le coefficient de détermination.*

**Calculez le coefficient de détermination**

De manière similaire au modèle linéaire simple, on dispose d'une formule de **décomposition de la variance (ANOVA :** ANalysis Of VAriance) faisant intervenir les quantités suivantes :

* Somme des Carrés Totale (SCT) : elle traduit la variation totale de **Y**.
* Somme des Carrés Expliquée (SCE) : elle traduit la variation expliquée par le modèle.
* Somme des Carrés Résiduelle (SCR) : elle traduit la variation inexpliquée par le modèle.

On se place dans le cas de la régression avec constante (en considérant par exemple que **X1=1**) :



**Le coefficient de détermination**

On appelle **coefficient de détermination,** noté , le réel dans **[0,1]** défini par :



Dans le cas de la régression avec constante :

* Si , on a alors SCE=SCT : toute la variation est expliquée par le modèle.
* Si , on a alors SCR=SCT : aucune variation n'est expliquée par le modèle.

En pratique, ce coefficient présente un inconvénient important : on pourrait introduire artificiellement des variables, pseudo-explicatives, et faire croître le coefficient de détermination. Plus le nombre de variables est important, plus l'erreur d'ajustement est faible et donc le coefficient de détermination proche de 1.

Cependant, la qualité prédictive du modèle diminue, rendant le modèle moins robuste. Afin de prendre en compte le nombre de variables explicatives, on considère souvent **le coefficient de détermination ajusté** défini de la manière suivante (dans le cas de la régression avec constante) :



Lorsque le modèle ne contient pas de constante, les formules et/ou interprétations sont légèrement différentes. Notez cependant que la majorité des modèles utilisés sont avec constante.

Pour terminer, n'oubliez pas qu'il faut veiller à ne pas surinterpréter le coefficient de détermination : un  faible indique simplement que l'ajustement linéaire n'est pas opportun. Cela n'indique pas forcément qu'il n'y a pas de relation entre **Y** et les **p** variables considérées.

*Le coefficient de détermination est maintenant calculé, il est temps de tester votre modèle, dans le chapitre suivant.*

## Testez le modèle linéaire gaussien multiple

Sans hypothèse supplémentaire, on ne peut pas déterminer d'intervalles de confiance sur les paramètres **β1,…,βp** ou tester leur significativité.

Rappelez-vous, nous avions eu le même problème en régression linéaire simple !

Cela devient possible si l'on ajoute une hypothèse sur la loi de l'erreur **ε**, et donc sur celle de la variable à expliquer **Y**. Le **modèle linéaire gaussien multiple** considère la normalité de l'erreur.

Dans **le modèle linéaire gaussien multiple**, on considère en plus des hypothèses formulées dans le cadre du modèle linéaire multiple :



On en déduit donc que  est un échantillon i.i.d de loi .

### Testez le modèle

Pour tester la significativité du modèle, nous avons 2 niveaux :

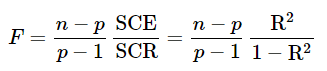
* Un **test global,** obtenu grâce à une statistique de Fisher. En pratique, l'hypothèse Ho de ce test est souvent rejetée, le modèle est donc souvent significatif globalement.
* Un **test de significativité** sur chacune des variables explicatives prises une à une. Dans ce cas, il s'agit d'un test de Student, tout comme en régression linéaire simple. Ici, tester l'un des paramètres a un réel sens : si une variable n'est pas significative, il faut la retirer du modèle. Si l'on ne la retire pas, il est possible que l'erreur de prévision du modèle soit plus élevée.

#### Testez la significativité globale du modèle

Dans le cas de la régression avec constante, on teste :



La statistique de test utilisée est :

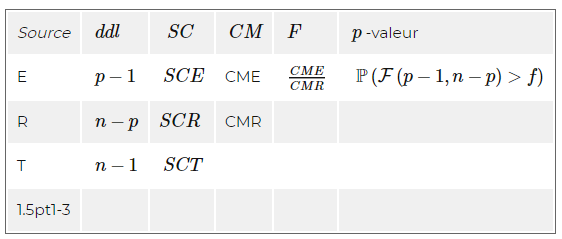


On peut montrer que sous **H0** :



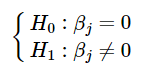
On décide du **rejet de H0** au niveau de test **α** si .

On présente usuellement le test de significativité globale sous la forme du tableau d'analyse de la variance, avec les notations  et .

Voici le tableau d'analyse de la variance : 

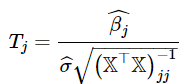
### Testez la significativité d'un des paramètres

On teste :



pour 

La statistique de test utilisée est :



où  désigne le **j**-ième terme diagonal de la matrice .

On peut montrer que sous **H0** :



On décide du**rejet de H0** au niveau de test **α** si .

En pratique, dire que l'on rejette cette hypothèse **H0** revient à conserver la variable **Xj** comme explicative.

Le test est acquis, félicitations ! On peut maintenant analyser nos résultats.

## Analysez les résultats

Après avoir estimé un modèle de régression linéaire, il faut ensuite **analyser :**

* **La significativité des paramètres :** un modèle correct doit avoir des paramètres significatifs.
* **L'atypicité** et **l'influence éventuelle de certaines données :** on pourra retirer les données atypiques et influentes.
* Les éventuels problèmes de **colinéarité.**
* Les éventuels problèmes **d'hétéroscédasticité**(quand la variance des résidus ne peut pas être considérée comme constante).

Dans le cas où plusieurs régresseurs sont disponibles, il faut choisir le meilleur modèle en utilisant un critère de choix et un algorithme de recherche.

### Analysez l'atypicité des observations

* **Sur les variables explicatives :**  
  Les termes diagonaux de **H** sont appelés "leviers des observations"



Belsey a proposé de considérer l'observation ii comme atypique si :

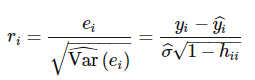


* **Sur la variable à expliquer :**  
  À partir des résidus



On considère les résidus studentisés pour évaluer si une observation est atypique. Une forte valeur de ces résidus caractérisera une observation atypique.

Pour , on définit**les résidus studentisés internes**par :

 suit approximativement une loi de Student . En pratique, on utilise cette approximation lorsque **n−p−1>30**.  
Au niveau de test **α**, on considère qu'une observation **i** est atypique si :



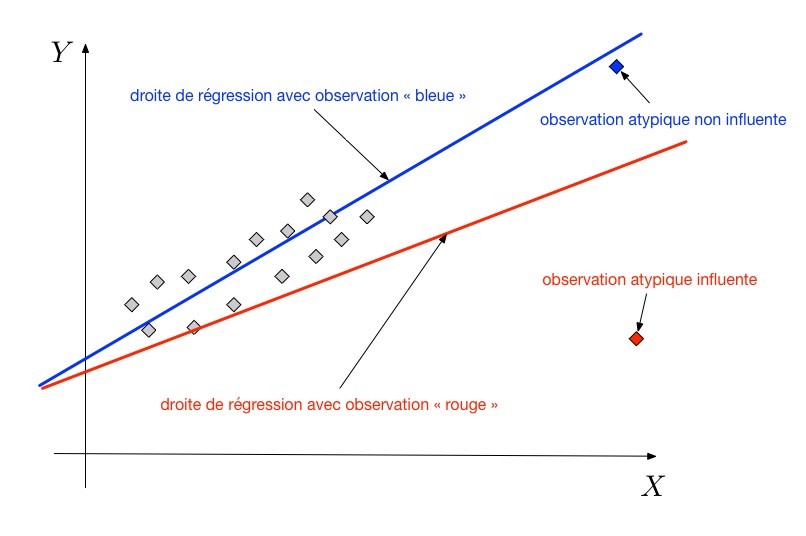


On peut également procéder par validation croisée, en considérant les résidus studentisés dits

externes : on calcule alors une erreur de prévision et plus d'ajustement ; on base les calculs

sur **n−1** observations pour prévoir la n-ième observation.

En pratique, on ne retire que les observations **atypiques ET influentes.** Le graphique suivant illustre les cas de figure de données atypiques (influentes ou non) dans le cadre de la régression linéaire simple.

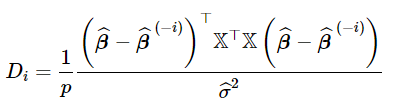


Influence de données atypiques

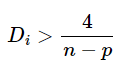
### Analysez l'influence des observations

La mesure de l'influence d'une observation s'effectue à l'aide de **la distance de Cook** qui mesure un écart entre  et   (calculs effectués sans la **i**-ème observation).

La distance de Cook pour l'observation i vaut :



Cook a proposé de considérer l'observation ii comme **influente** si :



### Détectez les problèmes de colinéarité

Parlons maintenant du problème de colinéarité.

L'estimateur des MCO existe si toutes les variables sont non colinéaires entre elles, c'est-à-dire seulement si l'on peut trouver une variable qui peut s'exprimer comme une combinaison linéaire des autres (par exemple ). Dans ce cas, il n'existe pas de solution unique au problème.

Heureusement, **la colinéarité est facile à détecter.**

Cependant, il n'en est pas de même pour une colinéarité "presque" exacte. Dans ce cas, il y a deux manières de procéder :

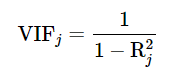
* utiliser des indicateurs comme le VIF ou le TOL (la définition est donnée juste après) ;
* utiliser dans le modèle des coefficients appelés "indices de conditionnement", puis analyser une décomposition de la variance.

Si les colonnes de  sont colinéaires, alors la matrice n'est pas de rang plein, ce qui conduit à une solution des MCO non unique.

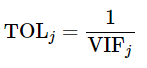
La colinéarité exacte peut facilement s'éviter, contrairement à une colinéarité approchée.

#### Le facteur d'influence de la variance ou la tolérance

On effectue la régression de **Xj**, pour **j∈{1,…,p}** sur les pp autres variables (dont la constante), et l'on calcule le coefficient de détermination .  
Le **facteur d'influence de la variance,** noté VIF (Variance Inflation Factor), de la variable , est défini par :



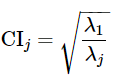
La **tolérance,** notée TOL, est définie comme l'inverse du facteur d'influence de la variance :



En pratique, une valeur ) indique un **problème de colinéarité** éventuel.

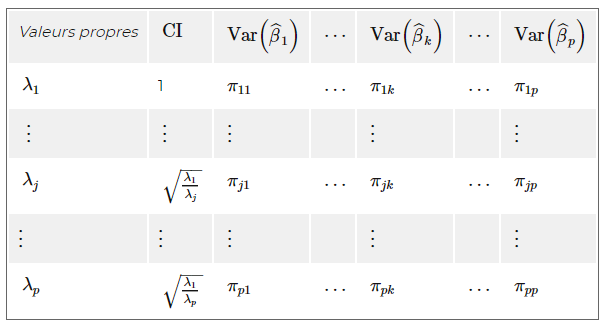
#### Analysez la structure des variables explicatives

On se place ici dans le cas d'une régression sans constante, ou dans le cas d'une régression avec constante pour laquelle on omet ce régresseur particulier.  
On considère les **p** valeurs propres **(λ1,…,λp)** de la matrice de corrélation des variables explicatives (à l'exception de la constante) ordonnées par ordre décroissant.  
Les **indices de conditionnement** sont définis pour  par :



Des valeurs élevées des indices de conditionnement traduisent la présence de colinéarité.  
En pratique, une valeur  indique un problème de colinéarité éventuel.  
On cherche ensuite à déterminer les groupes de variables concernés. On calcule pour cela la proportion  de variance du coefficient  due à la variable explicative .

Voici les différents indices de conditionnement :



En pratique, il faut étudier les variables explicatives **Xj** avec un **CI** élevé.  
Pour cette variable **Xj**, s'il existe au moins deux variables explicatives **Xk** et **Xk′** telles que  et  soient élevés (supérieurs à 0.5, en pratique), alors un problème de colinéarité est suspecté entre ces variables.

### Analysez l'homoscédasticité des résidus

On peut notamment étudier **l'homoscédasticité** des résidus, c'est-à-dire la constance de leur variance, en représentant graphiquement les résidus studentisés en fonction des valeurs ajustées. Un nuage avec une forme conique peut laisser présager un effet hétéroscédastique.

Il est également possible de tester l'homoscédasticité des résidus, à l'aide par exemple du test de Breusch-Pagan (l'hypothèse nulle est l'homoscédasticité).

Vous avez analysé un certain nombre de vos résultats. Dans le prochain chapitre, vous verrez comment choisir le modèle qui répond le mieux à votre problématique.

## Sélectionnez automatiquement un modèle

Pour utiliser notre modèle **à des fins de prévision,** il vaut mieux qu'il soit le plus **parcimonieux** possible, c'est-à-dire qu'il n'intègre **que les variables qui sont réellement utiles.**

Si vous souhaitez tester l'ensemble des modèles possibles, suivant la taille du jeu de données (ie. nombre d'individus et de variables), cela peut être très coûteux en temps et en mémoire informatique.

L'idée ici est d'**utiliser un algorithme de recherche de modèle** qui cherchera à optimiser un certain critère statistique.

### Quels critères statistiques pouvez-vous utiliser ?

Concernant les critères statistiques, on considère généralement des critères tels que R². Cependant, on sait que ce n'est pas l'indicateur le plus sensé. On peut également utiliser une information de type AIC ou BIC et le  de Mallows.

Ces derniers critères cherchent un compromis entre :

* l'ajustement du modèle ;
* la qualité de la prévision.

En effet, ces deux quantités ne varient pas dans le même sens : quand le nombre de paramètres augmente, meilleur est l'ajustement, mais moins bonne sera la prévision.

### Déterminez les régresseurs adaptés

Supposez que l'on dispose de **K** régresseurs potentiels et que vous cherchiez à déterminer les **k** régresseurs les mieux adaptés.

Il est possible de considérer les critères de sélection suivants pour chacun des modèles testés :

* le **coefficient de détermination** (ajusté ou pas) ;
* la **variance résiduelle ;**
* des critères basés sur **l'information de Kullback** (AIC ou BIC, par exemple) ;
* la **statistique de Mallows.**

#### Étudiez en détail les critères de sélection

Ces critères sont un compromis entre l'ajustement du modèle (on privilégie une faible variance résiduelle) et la parcimonie de ce modèle (on privilégie alors un faible nombre de variables explicatives). On trouve parmi eux (pour un modèle   à **k** variables explicatives) :

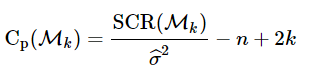
* Le **Critère d'Akaike** (AIC : Akaike Information Criterium) :



* Le **Critère de Schwarz** (BIC : Bayesian Information Criterium, également noté SBC pour Schwarz Bayesian Criterium) :



* La **statistique de Mallows :** est défini par :



où  est la variance du modèle complet (avec tous les régresseurs) et la somme des carrés des résidus du modèle  .

La statistique de Mallows est une mesure de l'erreur quadratique moyenne (qui englobe le biais et la variance) des prédictions produites par un sous-modèle. Si un sous-modèle à **k** variables explicatives possède un pouvoir prédictif proche de celui du modèle complet, au sens de l'erreur quadratique moyenne, alors **Cp≃k**.

En pratique, il est recommandé de choisir le modèle vérifiant **Cp≤k** .

### Découvrez les procédures itératives de choix

**La procédure la plus complète,** mais la plus fastidieuse également, consiste à **sélectionner le modèle qui minimise un des critères** précédents pour tous les modèles de régression potentiels à kk régresseurs, pour **k∈{1,…,K}**.

Sachez qu'il existe aussi des procédures alternatives :

* Procédure ascendante ou **forward**  
  On initialise la procédure en intégrant seulement la constante, puis les régresseurs sont introduits un par un, le principe étant de retenir à chaque pas la variable qui contribue le plus à augmenter la somme des carrés expliqués.
* Procédure descendante ou **backward**  
  On initialise la procédure en intégrant tous les régresseurs, puis on élimine à chaque pas le régresseur associé à la plus petite diminution de la somme des carrés expliqués (la constante est toujours conservée).
* Procédure pas-à-pas ou **stepwise**  
  Cette méthode est une procédure de sélection de type forward, avec possibilité d'éliminer, éventuellement, des variables devenues non significatives (dans une étape backward).

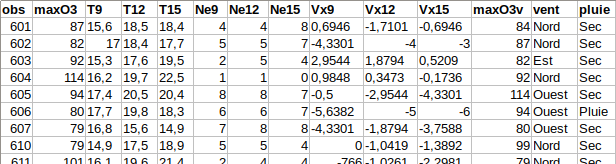
Et voilà, vous savez comment sélectionner automatiquement le modèle le plus juste grâce à un algorithme. Allez, on passe à la pratique dans le prochain chapitre.

## TP : Pratiquez la régression linéaire multiple sur le jeu de données de l'ozone

Ce TP est présenté en R. Voici la [**version Python**](https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/Parcours+Data+Analyst/Re%CC%81alisez+des+mode%CC%81lisations+de+donne%CC%81es+performantes+(1).zip).

Avant de commencer, retrouvez la description du cas d'étude et du jeu de données dans [**ce chapitre**](https://openclassrooms.com/fr/courses/4525326-realisez-des-modelisations-de-donnees-performantes/5754127-decouvrez-le-jeu-de-donnees-de-lozone).

Appliquons la régression linéaire multiple à l'échantillon ozone.  
Modélisons le pic d'ozone journalier en fonction de toutes les autres variables météorologiques.



Jeu de données ozone

#### Importez les données

On importe la librairie ggplot2, qui permettra d'afficher les graphiques :

library(ggplot2)

On importe les données, puis on utilise la commande lm pour régresser maxO3 en fonction des autres variables de l'échantillon.

ozone <- read.table("ozone.txt", header=TRUE, sep=";", dec=",")

reg\_multi <- lm(maxO3~T9+T12+T15+Ne9+Ne12+Ne15+maxO3v,data=ozone)

summary(reg\_multi)

Voici le résultat :

Call:

lm(formula = maxO3 ~ T9 + T12 + T15 + Ne9 + Ne12 + Ne15 + maxO3v,

data = ozone)

Residuals:

Min 1Q Median 3Q Max

-57.768 -7.845 -1.359 8.134 38.984

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 12.70548 13.10860 0.969 0.33467

T9 -0.63596 1.03462 -0.615 0.54011

T12 2.50600 1.39946 1.791 0.07625 .

T15 0.71381 1.13674 0.628 0.53142

Ne9 -2.76057 0.89157 -3.096 0.00252 \*\*

Ne12 -0.37193 1.34590 -0.276 0.78283

Ne15 0.09028 0.99934 0.090 0.92819

maxO3v 0.37774 0.06121 6.171 1.32e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.43 on 104 degrees of freedom

Multiple R-squared: 0.7546, Adjusted R-squared: 0.738

F-statistic: 45.68 on 7 and 104 DF, p-value: < 2.2e-16

On constate ici que certains paramètres ne sont pas significativement différents de 0, car leur p-valeur n'est pas inférieure à 5 %, le niveau de test que nous souhaitons.

Le  vaut environ 0.75, et le  ajusté est d'environ 0.74.

Cette valeur est plus élevée qu'en régression linéaire simple, et c'est logique, car lorsque l'on rajoute des variables explicatives potentielles, on accroît naturellement la valeur de ces .

#### Retirez les variables non significatives

On va donc maintenant retirer les variables non significatives. On commence par la moins significative : Ne15, car elle a une p-valeur de 0.93.

reg\_multi <- lm(maxO3~T9+T12+T15+Ne9+Ne12+maxO3v,data=ozone)

summary(reg\_multi)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 12.84916 12.95015 0.992 0.32338

T9 -0.62976 1.02745 -0.613 0.54124

T12 2.56018 1.25841 2.034 0.04443 \*

T15 0.65787 0.94878 0.693 0.48960

Ne9 -2.76526 0.88585 -3.122 0.00232 \*\*

Ne12 -0.30796 1.13912 -0.270 0.78742

maxO3v 0.37752 0.06087 6.202 1.12e-08 \*\*\*

On voit alors que c'est maintenant Ne12, avec une p-valeur de 0.79, qui est la moins significative. On l'enlève donc.

reg\_multi <- lm(maxO3~T9+T12+T15+Ne9+maxO3v,data=ozone)

summary(reg\_multi)

On constate qu'il faut maintenant retirer la variable T9 :

reg\_multi <- lm(maxO3~T12+T15+Ne9+maxO3v,data=ozone)

summary(reg\_multi)

Et l'on retire ensuite T15 :

reg\_multi <- lm(maxO3~T12+Ne9+maxO3v,data=ozone)

summary(reg\_multi)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.76225 11.10038 0.879 0.381

T12 2.85308 0.48052 5.937 3.57e-08 \*\*\*

Ne9 -3.02423 0.64342 -4.700 7.71e-06 \*\*\*

maxO3v 0.37571 0.05801 6.477 2.85e-09 \*\*\*

On remarque qu'à présent, tous les paramètres sont significatifs. Quant au , il vaut environ 0.75, tout comme le  ajusté.

On peut donc utiliser ce modèle à des fins de prévision !

Si l'on souhaite prévoir la concentration journalière en ozone, sachant que la température prévue à 12 h sera de 15 °C, que la valeur de Ne9 sera de 2, et que la concentration maxO3v de la veille vaut 100, alors on saisit les lignes suivantes :

a\_prevoir <- data.frame(T12=15,Ne9=2,maxO3v=100)

maxO3\_prev <- predict(reg\_multi,a\_prevoir)

round(maxO3\_prev, digits=2)

On obtient une concentration maxO3 de 84.

### Pour aller plus loin : analysez vos résultats

Reprenons la régression linéaire multiple que nous avons obtenue :

library(ggplot2)

ozone <- read.table("ozone.txt",header=TRUE,sep=";",dec=",")

reg\_mu1ti <- lm(maxO3~T12+Ne9+maxO3v,data=ozone)

summary(reg\_mu1ti)

Call:

lm(formula = maxO3 ~ T12 + Ne9 + maxO3v, data = ozone)

Residuals:

Min 1Q Median 3Q Max

-56.385 -7.872 -1.941 7.899 41.513

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 9.76225 11.10038 0.879 0.381

T12 2.85308 0.48052 5.937 3.57e-08 \*\*\*

Ne9 -3.02423 0.64342 -4.700 7.71e-06 \*\*\*

maxO3v 0.37571 0.05801 6.477 2.85e-09 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.23 on 108 degrees of freedom

Multiple R-squared: 0.752, Adjusted R-squared: 0.7451

F-statistic: 109.1 on 3 and 108 DF, p-value: < 2.2e-16

Nous allons ici réaliser les tests à un niveau α=5% :

alpha <- 0.05

Récupérons **n**, le nombre d'individus de l'échantillon, et **p**, le nombre de variables.

n <- dim(ozone)[1]

p <- 4

Nous allons mener des analyses sur les valeurs atypiques et/ou influentes en travaillant sur un dataframe appelé analyses.

analyses <- data.frame(obs=1:n)

#### Calculez les leviers

On peut calculer les leviers comme ceci, en sachant que le seuil des leviers est .

analyses$levier <- hat(model.matrix(reg\_multi))

seuil\_levier <- 2\*p/n

On peut visualiser les leviers pour chaque point comme ceci :

ggplot(data=analyses,aes(x=obs,y=levier))+

geom\_bar(stat="identity",fill="steelblue")+

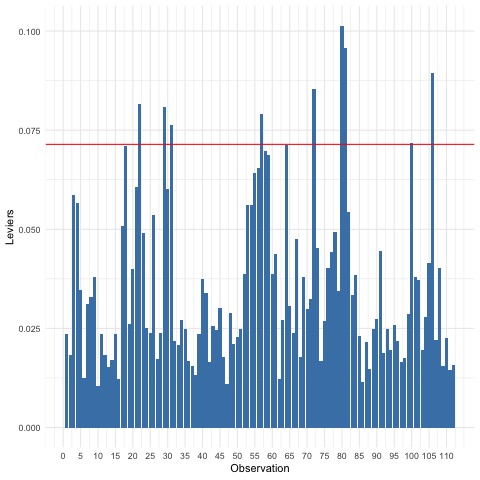
geom\_hline(yintercept=seuil\_levier,col="red")+

theme\_minimal()+

xlab("Observation")+

ylab("Leviers")+

scale\_x\_continuous(breaks=seq(0,n,by=5))



Visualisation des leviers

Pour sélectionner les points pour lesquels le levier est supérieur au seuil, on exécute ces 2 lignes :

idl <- analyses$levier>seuil\_levier

idl

analyses$levier[idl]

#### Calculez les résidus studentisés

Si l'on souhaite maintenant calculer les résidus studentisés, nous écrivons ceci, sachant que le seuil pour les résidus studentisés est une loi de Student à n-p-1 degrés de liberté :

analyses$rstudent <- rstudent(reg\_mu1ti)

seuil\_rstudent <- qt(1-alpha/2,n-p-1)

Visualisons les résidus studentisés :

ggplot(data=analyses,aes(x=obs,y=rstudent))+

geom\_bar(stat="identity",fill="steelblue")+

geom\_hline(yintercept=-seuil\_rstudent,col="red")+

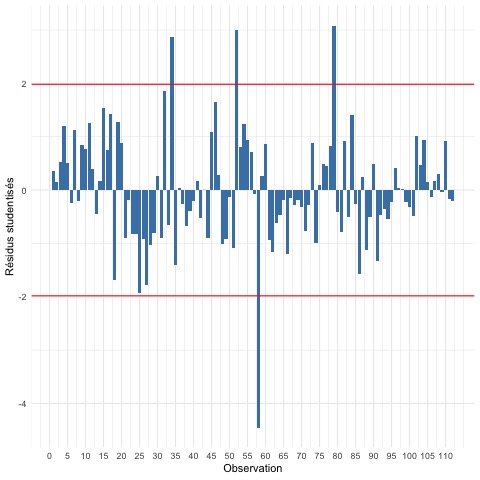
geom\_hline(yintercept=seuil\_rstudent,col="red")+

theme\_minimal()+

xlab("observation")+

ylab("Résidus studentisés")+

scale\_x\_continuous(breaks=seq(0,n,by=5))



Visualisation des résidus studentisés

#### Déterminez la distance de Cook

Pour trouver la distance de Cook, nous exécutons ceci :

influence <- influence.measures(reg\_mu1ti)

names(influence)

colnames(influence$infmat)

Le seuil de la distance de Cook est de n-p :

analyses$dcook <- influence$infmat[,"cook.d"]

seuil\_dcook <- 4/(n-p)

On peut détecter les observations influentes comme ceci :

ggplot(data=analyses,aes(x=obs,y=dcook))+

geom\_bar(stat="identity",fill="steelblue")+

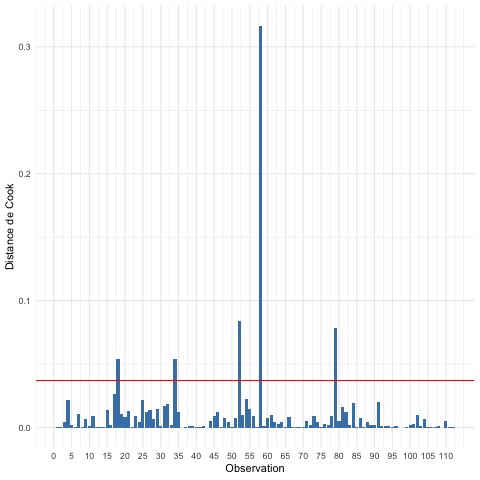
geom\_hline(yintercept=seuil\_dcook,col="red")+

theme\_minimal()+

xlab("Observation")+

ylab("Distance de cook")+

scale\_x\_continuous(breaks=seq(0,n,by=5))



Visualisation des observations influentes

 On ne retire des points qu'après avoir vérifié qu'ils sont effectivement atypiques, voire aberrants, au vu du modèle estimé.

#### Vérifier la colinéarité des variables

Une autre chose à vérifier est l'éventuelle colinéarité approchée des variables :

vif(reg\_mu1ti)

Ici, tous les coefficients sont inférieurs à 10, il n'y a donc pas de problème de colinéarité.

#### Testez l’homoscédasticité

On peut également tester l’homoscédasticité (c'est-à-dire la constance de la variance) des résidus :

bptest(reg\_mu1ti)

La p-valeur ici n'est pas inférieure à 5 %, on ne rejette pas l'hypothèse H0H0 selon laquelle les variances sont constantes (l'hypothèse d’homoscédasticité).

#### Testez la normalité des résidus

Si l'on veut tester la normalité des résidus, on peut faire un test de Shapiro-Wilk.

shapiro.test(reg\_mu1ti$residuals)

Ici, l'hypothèse de normalité est remise en cause.

Néanmoins, l'observation des résidus, le fait qu'ils ne soient pas très différents d'une distribution symétrique, et le fait que l'échantillon soit de taille suffisante (supérieure à 30) permettent de dire que **les résultats obtenus par le modèle linéaire gaussien ne sont pas absurdes,** même si le résidu n'est pas considéré comme étant gaussien.

Nous aurions pu aussi sélectionner automatiquement un modèle avec l'ensemble des variables à disposition (variables météo et pic d'ozone de la veille) :

reg\_null <- lm(maxO3~1,data=ozone)

reg\_tot <- lm(maxO3~T9+T12+T15+Ne9+Ne12+Ne15+maxO3v,data=ozone)

reg\_backward <- step(reg\_tot,direction="backward")

Et voilà, vous avez vu en pratique comment réaliser une régression linéaire multiple pour déterminer et prédire la concentration d'ozone dans l'atmosphère. Dans la prochaine partie, vous aborderez le modèle de régression logistique.

## Entraînez-vous : améliorez les prévisions de hauteur des arbres

### 

### À vous de jouer !

Pour vous entraîner, réalisez cet exercice étape par étape. Une fois terminé, vous pouvez comparer votre travail avec les pistes que je vous propose.

#### Objectif de l'activité

Dans la partie précédente, vous avez déterminé la hauteur d'un arbre en fonction de sa circonférence, à l'aide d'une régression linéaire simple. Dans cette activité, vous irez un petit peu plus loin, et améliorerez votre prévision, cette fois à l'aide d'une régression linéaire multiple !

#### Votre mission

1. Ajouter une colonne à l'échantillon. La nommer circ\_sqrt et la remplir avec la racine carrée de la circonférence de chaque arbre.
2. Effectuer la régression linéaire multiple de la hauteur en fonction :  
   - de la circonférence ;  
   - de circ\_sqrt.
3. Analyser la significativité des paramètres, et retirer les éventuels paramètres non significatifs.
4. Donner et interpréter le coefficient de détermination du modèle finalement retenu.

#### Les données

Pour cette activité, vous vous baserez sur le jeu de données "[arbres](https://opendata.paris.fr/explore/dataset/les-arbres/table/)".

### Corrigez votre travail

Alors, vous êtes allé au bout ? Suivez le guide pour vérifier votre travail !

Vérifiez que votre travail remplit les critères suivants :

* Une colonne est bien créée et contient pour chaque arbre la racine carrée de la circonférence.
* Le code contient une ligne permettant le calcul de la régression linéaire multiple. Une autre fonction que celle proposée dans le corrigé peut être acceptée, à condition qu'elle réalise bien une régression linéaire.
* La réponse est correcte s'il a été identifié que le paramètre de la circonférence n'est pas significatif, et que le paramètre de circ\_sqrt est significatif. Une régression linéaire doit être relancée ensuite avec comme unique variable explicative circ\_sqrt. C'est sur cette dernière modélisation que le R2 de la question suivante devra être interprété.
* La réponse est correcte si elle est équivalente à : "On rejette la nullité des paramètres au niveau de test 5 %."

Voici un exemple corrigé pour vous guider.

**Appréhendez le fonctionnement de la régression logistique**

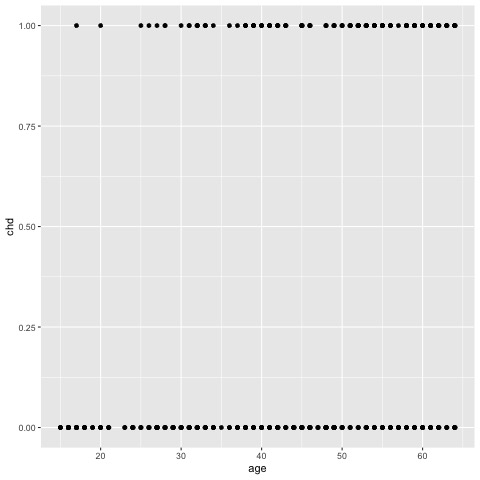
L'objectif de la régression logistique est de **modéliser,** de **classifier,** une variable binaire prenant ses valeurs dans {0,1} en fonction de variables explicatives quantitatives (et potentiellement qualitatives). La régression logistique est une méthode de **classification (supervisée)** qui permet de traiter des cas comme :

* la prévision de présence/absence d'une maladie ;
* la prévision de l'état de fonctionnement d'une machine-outil en fonction de ses caractéristiques (ancienneté, modèle, etc.), à des fins de maintenance prédictive ;
* le credit scoring (attribution ou non d'un crédit).

**Retour au jeu de données des maladies cardio-vasculaires**

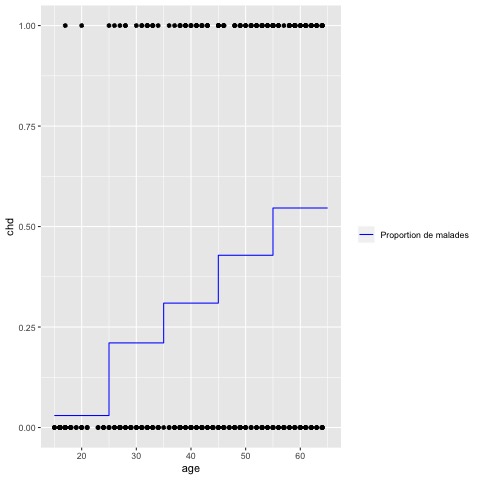
L'objectif de notre [cas d'étude n°2](https://openclassrooms.com/fr/courses/4525326-realisez-des-modelisations-de-donnees-performantes/5754128-decouvrez-le-jeu-de-donnees-des-maladies-cardio-vasculaires) est de prévoir la présence/absence d'une maladie cardio-vasculaire en fonction de constantes de santé chez un individu.

Si l'on cherche à expliquer cette maladie en fonction de l'âge, on peut tout d'abord représenter le nuage de points :

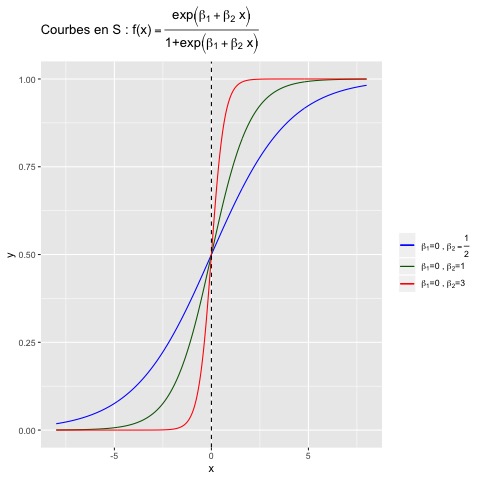


Nuage de points. Chd = personne atteinte d'une maladie coronarienne ou non.

Il est évident qu'une régression linéaire n'est pas adéquate ici, on obtiendrait des valeurs en dehors de **{0,1}**. L'objectif ici est de modéliser la probabilité d'être malade en fonction de l'âge. La visualisation de la fréquence empirique permet de comprendre le choix de *courbes en S* pour modéliser cette proportion.



Proportion de malades en fonction de l'âge



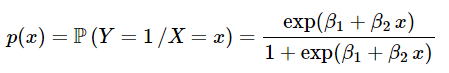
Exemples de courbes en S

On considère ici que :

1. La variable **chd***sachant***age=x**, qui prend comme valeurs 1 ou 0, suit une loi de Bernoulli de paramètre **p(x)** dépendant de **x** :



1. La probabilité **p(x)** s'écrit sous la forme :



Il s'agit bien d'un modèle logistique !

*Voici donc un premier aperçu du modèle logistique, et d'un problème qu'il permet de traiter. Voyons plus en détail comment estimer le modèle.*

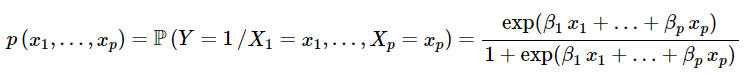
**Estimez un modèle de régression logistique**

De manière générale, la **régression logistique** de Y∈{0,1} sur **p** variables explicatives X1…,Xp}, consiste à considérer que :

1. La loi de Y sachant X1=x1,…,Xp=xp est une loi de Bernoulli de paramètre p(x1,…,xp) dépendant de x1,…,xpx1,…,xp :

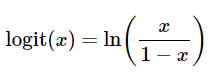


1. La probabilité p(x1,…,xp) s'écrit sous la forme :



À noter que si l'on souhaite que le modèle contienne une constante, on considère que **X1=1**, comme dans le cas d'étude.

Avec la fonction **logit,** la fonction définie sur **[0,1]** et à valeurs dans  :



on peut montrer que



On est ici dans le cadre plus global des modèles linéaires généralisés (GLM) qui contient :

* La **régression linéaire :** la loi de **Y** sachant **X1=x1,…,Xp=xp** est une loi normale.
* La **régression log-linéaire :** la loi de **Y** sachant **X1=x1,…,Xp=xp** est une loi de Poisson ( **Y** est une variable de comptage, d'incidents, par exemple).
* La **régression logistique :** la loi de **Y** sachant **X1=x1,…,Xp=xp** est une loi de Bernoulli.

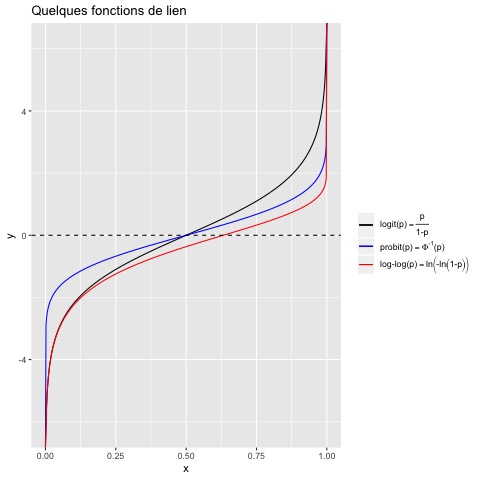
Y∈{0,1}, on peut considérer d'autres fonctions g telles que :



notamment :

* : **régression logistique.**
* , où Φ est la fonction de répartition de la loi  : **régression probit.**
*  : **régression log-log.**

Et en image :



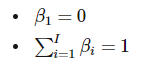
La fonction log-log est adaptée au cas où l'on considère qu'il y a asymétrie entre les probabilités de succès et d'échec. Si les fonctions probit et logit fournissent des résultats similaires, on profère généralement la régression logistique pour des raisons numériques (la fonction logit est plus simple à manipuler que la fonction probit) et d'interprétation (on peut utiliser les odd-ratios dans le cas de la régression logistique).

**Cas des variables explicatives qualitatives**

Dans le cas où l'on dispose (au moins) d'une variable explicative qualitative avec **I** modalités, on considère alors un modèle avec **I** indicatrices (valant 1 si l'événement est vrai, 0 sinon). Dans le cas d'une variable explicative qualitative avec **I** modalités :



Le modèle est indéterminé si l'on considère en sus la constante dans le modèle, on doit alors poser une contrainte, par exemple :



**Comment estimer en régression logistique ?**

Dans la régression logistique, on ne peut pas estimer les paramètres par MCO, on procède par maximum de vraisemblance.

L'estimateur du maximum de vraisemblance, s'il existe, annule le gradient de la vraisemblance (et celui de la log-vraisemblance).

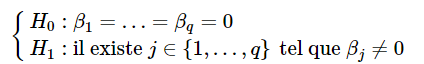
Ce système non linéaire en  n'admet pas de solution explicite, il faut donc utiliser des algorithmes d'optimisation itératifs pour le déterminer : on utilise notamment l'algorithme **IRLS** (Iterative Reweighted Least Square).

**Validation du modèle**

En sus du test de significativité d'un paramètre (test de Student), on peut tester la nullité globale des paramètres (ou d'un sous-ensemble des paramètres) via :

* le test de Wald ;
* le test du score ;
* le test du rapport de vraisemblance ou de la déviance.

Le test considéré ici est :



On ne perd aucunement en généralité, on remarque qu'il est possible de réordonner les variables explicatives, et donc de tester n'importe quelle combinaison de **q** variables.

On peut montrer que les statistiques de test considérées convergent toutes vers la loi  sous **H0**. On rejettera donc l'hypothèse **H0** au niveau de test **α** si les statistiques calculées sur l'échantillon sont supérieures au quantile d'ordre **1−α** de la loi .

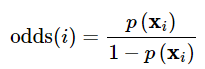
*Comme pour le modèle de régression linéaire, il est temps à présent d'analyser les résultats au chapitre suivant.*

**Analysez les résultats**

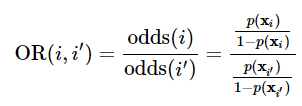
**Comment interpréter les paramètres ?**

Si l'on considère un événement avec une probabilité de réussite, cela signifie qu'en moyenne, sur **N** individus, on a 1 réussite contre **N−1** échecs. On parle alors d'une cote de **N−1** chances contre 1.

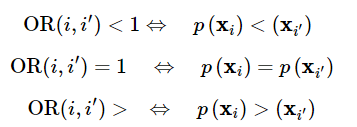
La cote (l'**odds**en anglais) pour un individu ii d'obtenir la réponse **Y=1** est définie par :



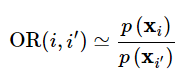
Le rapport de cotes, l'**odds ratio**, entre 2 individus i et **i′**, est défini par :



En toute généralité, on peut interpréter les odds ratios comme suit, pour 2 individus **i** et **i′** :



Si les probabilités sont telles que  (cas d'un événement rare), on a :

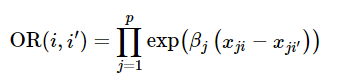


Par exemple, si , l'événement a une probabilité **k** fois plus importante de se produire pour **xi** que pour **xi′**.

Dans le cas de la régression logistique :



on peut montrer que :



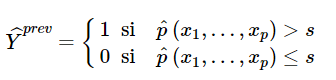
Toutes choses égales par ailleurs, une variation d'une unité sur la j-ième variable correspond à un odds ratio égal à . L'exponentielle du coefficient peut être vue comme un odds ratio.

**Choisissez le modèle adéquat**

On peut comparer des modèles entre eux, ou procéder à une sélection automatique de modèle (à l'aide des algorithmes **backward, forward** et **stepwise**) en se basant sur des critères tels que l'**AIC** (critère d'Akaike) ou le **BIC** (critère de Schwarz) vus dans les chapitres précédents.

**Comment prévoir "Y" ?**

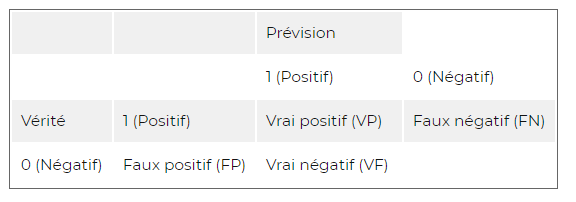
On considère la règle de décision suivante pour prévoir désigne la prévision) :



où s∈[0,1]s∈[0,1] est un seuil fixé par l'utilisateur. Classiquement, on considère s=1/2 : si la probabilité prédite est (strictement) supérieure à 1/2 , on prévoit 1, sinon 0.

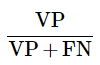
On peut évaluer la qualité d'un modèle par validation croisée, et étudier un certain nombre de critères, parmi lesquels la sensibilité et la spécificité, calculées à partir de la matrice de confusion.

Dans le cas d'un classificateur binaire, la **matrice de confusion** est définie comme suit :

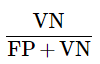


Les nombres de faux positifs et de faux négatifs devraient être nuls pour que la règle de décision soit parfaite ; ce n'est jamais le cas en pratique.

La **sensibilité,** le taux de positifs classés positifs, vaut:



La **spécificité,** le taux de négatifs classés négatifs, vaut :



**Appréhendez la courbe ROC**

La courbe ROC (Receiver Operating Characteristic) représente la sensibilité en fonction de la spécificité pour différents seuils de décision s. L'aire sous la courbe ROC, l' AUC (Area Under the ROC), est une mesure de la qualité de la classification qui varie entre :

* AUC=1/2 : dans le pire des cas ;
* AUC=1 : dans le meilleur des cas.

*Et voilà, vous avez procédé à l'analyse des résultats. Voyons tout cela en pratique dans le prochain chapitre. Vous allez prédire le risque de contracter une maladie chez des patients.*

## 

## TP : Pratiquez la régression logistique sur le jeu de données des maladies cardio-vasculaires

Ce TP est présenté en R. Voici la [**version Python**](https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/Parcours+Data+Analyst/Re%CC%81alisez+des+mode%CC%81lisations+de+donne%CC%81es+performantes+(1).zip).

Avant de commencer, retrouvez la description du cas d'étude et du jeu de données dans [**ce chapitre**](https://openclassrooms.com/fr/courses/4525326-realisez-des-modelisations-de-donnees-performantes/5754128-decouvrez-le-jeu-de-donnees-des-maladies-cardio-vasculaires).

Dans ce TP, vous allez appliquer la régression logistique sur le cas d'étude portant sur les maladies cardio-vasculaires.

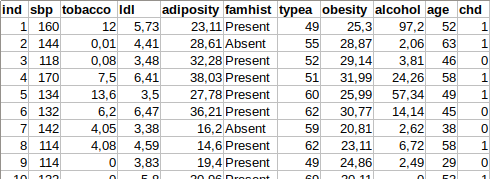
**Importez les données**

On charge tout d'abord la librairie ggplot2 :

library(ggplot2)

On importe ensuite les données :

maladie <- read.table("maladie.txt",header=TRUE,sep=";",dec=".")



Le jeu de données des maladies cardio-vasculaires

Ces données contiennent les informations de 462 patients d'Afrique du Sud. On y trouve des informations telles que :

* la tension artérielle ;
* le fait de fumer ou pas ;
* l'adiposité ;
* l'obésité ;
* etc.

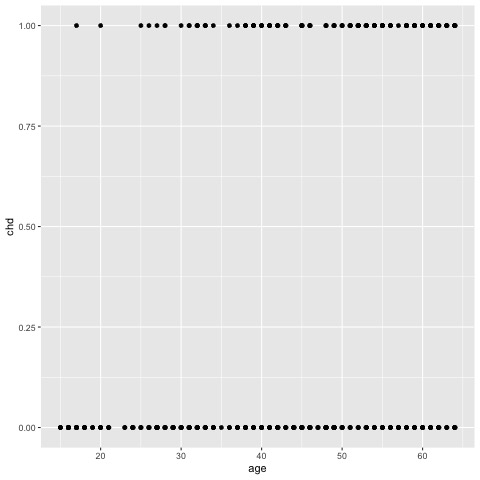
La variable CHD est la variable qui indique si la personne a une maladie cardio-vasculaire (1) ou pas (0).

**Visualisez le nuage de points**

Pour étudier le fait d'être malade en fonction de l'âge, on peut visualiser le nuage de points :

ggplot()+

geom\_point(data=maladie,aes(x=age,y=chd))



Visualisation du nuage de points

Il y a des 0 et des 1, mais il est ici difficile de dire si l'on est plus ou moins malade en fonction de l'âge.

On voit également qu'une régression linéaire sur un tel nuage de points n'aurait aucun sens, car elle nous donnerait des valeurs qui ne seraient quasiment jamais sur 0 ni 1.

**Calculez les proportions de malades**

On peut calculer des classes d'âge et les proportions de malades associées.

maladie$cl\_age <- cut(maladie$age,breaks=seq(15-1e-10,65,by=10))

prop <- prop.table(table(maladie$cl\_age,maladie$chd),1)

prop\_chd <- data.frame(age=c(15,rep(seq(25,55,by=10),each=2),65),

prop\_chd=rep(prop[,2],each=2))

On peut représenter ces proportions :

ggplot()+

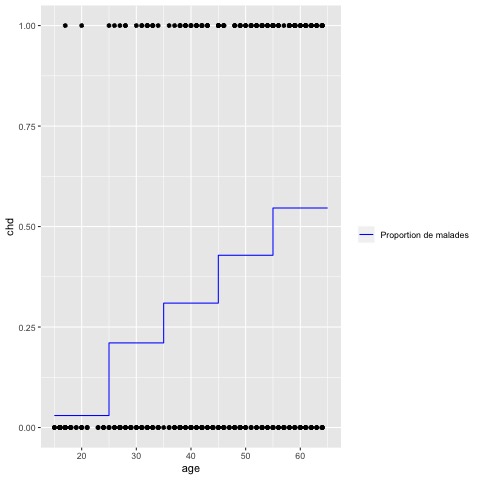
geom\_point(data=maladie,aes(x=age,y=chd))+

geom\_line(data=prop\_chd,aes(x=age,y=prop\_chd,colour="blue"))+

xlab("age")+

ylab("chd")+

scale\_colour\_manual(values="blue",label="Proportion de malades",name=" ")



Visualisation des proportions de malades

On y voit une fonction en escalier avec une forme de "S".

**Effectuez la régression logistique**

Effectuons donc une régression logistique de CHD en fonction de l'âge :

reg\_log1 <- glm(chd~age,family="binomial",data=maladie)

summary(reg\_log1)

Voici le résultat :

Call:

glm(formula = chd ~ age, family = "binomial", data = maladie)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.4321 -0.9215 -0.5392 1.0952 2.2433

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.521710 0.416031 -8.465 < 2e-16 \*\*\*

age 0.064108 0.008532 7.513 5.76e-14 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 596.11 on 461 degrees of freedom

Residual deviance: 525.56 on 460 degrees of freedom

AIC: 529.56

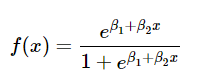
Number of Fisher Scoring iterations: 4

On obtient les paramètres estimés :  et . Enregistrons-les :

beta1 <- reg\_log1$coefficients[1]

beta2 <- reg\_log1$coefficients[2]

Dans le but de tracer la courbe logistique entre les abscisses x=15 et x=65, on définit une séquence de 15 à 65 par pas de 500, puis on la place dans la variable x. On calcule ensuite les ordonnées de la courbe, grâce à l'expression de la courbe en S :



Nous plaçons ces ordonnées dans la variable y. Enfin, avec x et y, nous créons un dataframe :

x <- seq(15,65,len=500)

y <- exp(beta1+beta2\*x)/(1+exp(beta1+beta2\*x))

reg\_log <- data.frame(age=x,prop\_chd=y)

Si l'on souhaite superposer la fonction de lien obtenue par régression logistique sur le graphique précédent, nous obtenons ceci :

ggplot()+

geom\_point(data=maladie,aes(x=age,y=chd))+

geom\_line(data=prop\_chd,aes(x=age,y=prop\_chd,colour="blue"))+

geom\_line(data=reg\_log,aes(x=age,y=prop\_chd,colour="red"))+

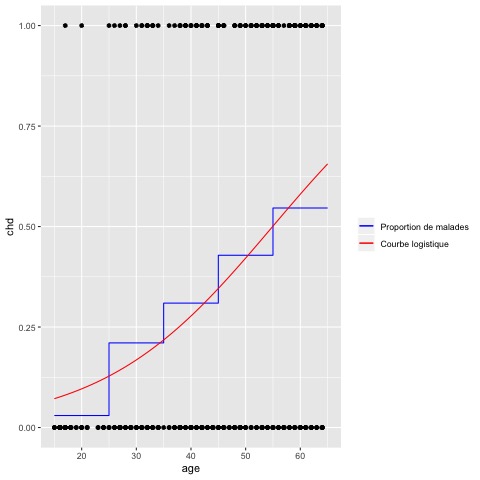
xlab("age")+

ylab("chd")+

scale\_colour\_manual(values=c("blue","red"),

label=list("Proportion de malades","Courbe logistique"),

name=" ")



Visualisation de la fonction de lien obtenue

La courbe rouge est celle qui est obtenue par régression logistique.

Si l'on avait voulu considérer l'ensemble des variables médicales (et non pas seulement l'âge comme jusqu'à présent), nous aurions écrit :

reg\_log2 <- glm(chd~sbp+tobacco+ldl+adiposity+famhist+typea+obesity+alcohol+age,

family="binomial",data=maladie)

summary(reg\_log2)

Call:

glm(formula = chd ~ sbp + tobacco + ldl + adiposity + famhist +

typea + obesity + alcohol + age, family = "binomial", data = maladie)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.7781 -0.8213 -0.4387 0.8889 2.5435

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -6.1507209 1.3082600 -4.701 2.58e-06 \*\*\*

sbp 0.0065040 0.0057304 1.135 0.256374

tobacco 0.0793764 0.0266028 2.984 0.002847 \*\*

ldl 0.1739239 0.0596617 2.915 0.003555 \*\*

adiposity 0.0185866 0.0292894 0.635 0.525700

famhistPresent 0.9253704 0.2278940 4.061 4.90e-05 \*\*\*

typea 0.0395950 0.0123202 3.214 0.001310 \*\*

obesity -0.0629099 0.0442477 -1.422 0.155095

alcohol 0.0001217 0.0044832 0.027 0.978350

age 0.0452253 0.0121298 3.728 0.000193 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 596.11 on 461 degrees of freedom

Residual deviance: 472.14 on 452 degrees of freedom

AIC: 492.14

Number of Fisher Scoring iterations: 5

Certaines des variables obtenues ont des p-valeurs qui sont inférieures au niveau de test de 5 %, ce qui nous indique qu'elles sont bien significatives. Certaines autres ne sont pas en dessous de ce seuil.

On peut donc passer sur une procédure de sélection en retirant les variables non significatives au fur et à mesure, mais nous pouvons aussi sélectionner automatiquement un modèle avec une commande telle que stepAIC, qui sélectionne de manière automatique un modèle en se basant sur le critère AIC.

library(MASS)

stepAIC(reg\_log2)

Et voilà, vous avez appliqué une régression logistique permettant de traiter des variables qualitatives binaires (avec deux modalités). Sachez qu'il existe d'autres méthodes de classification permettant de traiter des variables quantitatives avec davantage de modalités !

*Plus qu'un quiz, et vous arriverez à la cinquième partie de ce cours, qui porte sur l'analyse de la variance, l'ANOVA. Je vous retrouve là-bas !*

# Avez-vous compris les enjeux de la régression logistique ?

Bravo ! Vous avez réussi cet exercice !

### Compétences évaluées

* Appliquer et interpréter une régression logistique

### Question 1

**La régression logistique explique :**

* + 

uniquement des variables quantitatives

* + 

uniquement des variables qualitatives

* + 

des variables qualitatives ou quantitatives

### Question 2

**La régression logistique modélise la probabilité d'avoir**Y=1**par une loi :**

* + 

uniforme

* + 

de Bernoulli

* + 

binomiale

### Question 3

**Dans la régression logistique, on se ramène à un modèle linéaire à l'aide de la fonction :**

* + 

logit

* + 

probit

* + 

de répartition de la loi normale centrée réduite

### Question 4

**On estime la régression logistique à l'aide de la méthode :**

* + 

des moindres carrés

* + 

des moments

* + 

du maximum de vraisemblance

### Question 5

**On peut évaluer la qualité prédictive de la régression logistiqueà l'aide :**

* + 

du 

* + 

de la moyenne des résidus

* + 

de la courbe ROC et son AUC

### Question 6

**La sélection automatique d'un modèle de régression logistique :**

* + 

peut s'effectuer par tirage au sort

* + 

peut s'effectuer par un algorithme backward ou stepwise

* + 

est impossible

### Question 7

**La régression logistique fait partie des modèles de :**

* + 

classification

* + 

régression

*Malgré son nom, la régression logistique explique une variable qualitative (plus précisément une variable binaire), c'est donc un modèle de classification. Les modèles de régression expliquent (ou prédisent) une variable quantitative.*

### Question 8

**Pourquoi accorde-t-on de l'importance à la p-valeur des paramètres ?**

* + 

Pour savoir si le modèle épouse bien les observations

* + 

Pour savoir quels sont les paramètres significatifs et donc détecter les variables qu'il faut intégrer (ou non) au modèle.

*Certaines variables sont meilleures que d'autres pour expliquer la variable binaire Y à expliquer par le modèle. Intégrer trop de variables au modèle n'est pas bon, on utilise donc la p-valeur pour connaître les variables à conserver.*

**Appréhendez le fonctionnement de l'analyse de la variance (ANOVA)**

près avoir vu nos trois premiers types de modélisation, nous allons maintenant nous intéresser à l'**analyse de la variance** (ANOVA).

Il s'agit ici d'étudier l'impact d'une variable qualitative sur une variable quantitative.

Si l'on reprend le [cas d'étude](https://openclassrooms.com/fr/courses/4525326-realisez-des-modelisations-de-donnees-performantes/5754130-decouvrez-le-jeu-de-donnees-du-ble) du rendement de blé, la question que l'on pose est la suivante :

La variété de blé a-t-elle un impact sur le rendement ?

Ici, 4 variétés de blé ont été plantées sur 20 parcelles chacune. Le rendement de blé obtenu sur chacune des parcelles a été mesuré.

Le modèle considéré ici s'écrit sous la forme suivante :



pour 

Dans ce modèle :

* yij est le rendement de la parcelle j avec la variété i ;
* μ est le rendement moyen ;
* αi ne dépend que de la variété de blé.

On suppose ici que les parcelles sont homogènes et indépendantes, et que le rendement de la variété i suit une loi normale de moyenne μ+αi et de variance (variance identique pour toutes les variétés de blé).

Dans le cas d'étude, .

Pour savoir s'il existe un effet variété, on construira un test statistique dont l'hypothèse nulle est :



Cette hypothèse nulle revient à considérer que les 4 variétés conduisent à un rendement moyen égal à μ. Si c'est le cas, c'est que la variété de blé n'a pas d'impact sur le rendement. Si au contraire l'un d'eux est non nul, c'est que la variété de blé a un effet sur le rendement.

L'**analyse de la variance à un facteur** nous permettra de traiter ce cas de figure.

On pourra également mener cette analyse en considérant l'influence de la variété de blé et du traitement phytosanitaire : on se placera alors dans le cadre de l'analyse de la variance **à deux facteurs.**

*Allez, on commence tout de suite par une analyse de la variance à un facteur.*

**Réalisez une analyse de la variance**

**Réalisez une ANOVA à un facteur**

On se place dans le cas plus général où la variable qualitative à **I** niveaux (**I=4**pour la variété de blé, dans le cas d'étude). On considère que l'on dispose de **ni** observations pour la modalité **i** de la variable.

On parle d'un **plan d'expérience :**

* **complet** si ;
* **équilibré** si .

Dans le cas d'étude, le plan est équilibré (donc forcément complet).

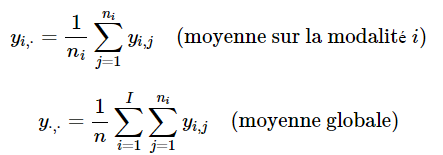
Le modèle s'écrit :



pour et  .

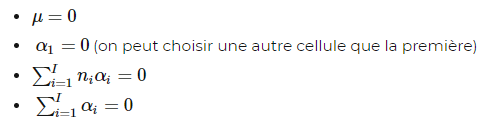
 sont des paramètres inconnus, et les **εi,j** sont des v.a.r. indépendantes de loi , où  est inconnue.

On considère par la suite les notations suivantes :

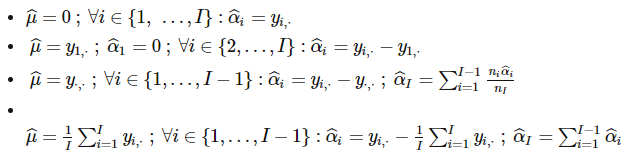


**Pour aller plus loin : l'estimation du modèle**

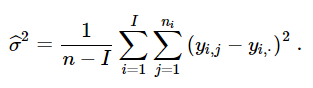
Afin de résoudre le problème, il faut poser une contrainte supplémentaire (les variables explicatives sont strictement colinéaires), par exemple :



Les estimateurs de  sont alors :



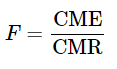
L'estimateur de  est dans tous les cas :



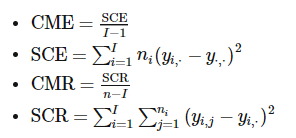
**Réalisez le test d'influence d'une variable qualitative**

Ce n'est pas l'estimation de ces paramètres inconnus qui nous intéresse ici, mais bien notre capacité à tester une hypothèse **H0** du type *la variété de blé n'a pas d'effet,* ce qui se traduit statistiquement par : .

On rejette d'autant plus facilement cette hypothèse que les moyennes sont différentes les unes des autres. La statistique de test utilisée à cet effet est :



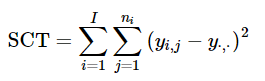
où :



La terminologie employée ici est :

* **Variation interclasse :** en français **SCE** (Somme des Carrés Expliqués), en anglais **SSM** (Sum of Squares of the Model).
* **Variation intraclasse :** en français **SCR** (Somme des Carrés Résiduels), en anglais **SSE** (Sum of Squares of the Error).
* **Variation totale :** en français **SCT** (Somme des Carrés Totaux), en anglais **SST** (Total Sum of Squares).
* **Degrés de liberté :** en français **ddl**, en anglais **df** (degrees of freedom).
* **Carrés moyens :** en français **CM**, en anglais **MS** (Mean Squares).

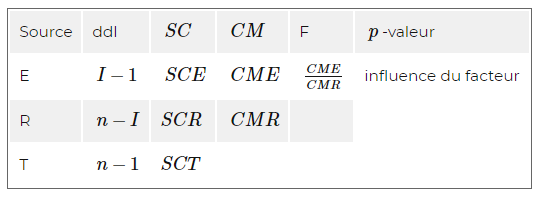
On peut montrer également que la somme des carrés totaux, définie par :



vérifie :



On présente classiquement les résultats sous forme de **tableau d'analyse de la variance :**



On rejette **H0** au niveau de test **α** si  où  est le quantile d'ordre **1−α** de la loi de Fisher à **(I−1,n−I)** degré de liberté.

On peut également lire le résultat de ce test via la p-valeur. On rappelle que l'on rejette **H0** au niveau de test **α** si **p-valeur<α** .

En pratique, rejeter **H0** revient à déclarer que la variable qualitative a un effet significatif sur notre phénomène (**Y**).

*Voilà pour l'ANOVA à un facteur. Passons maintenant à l'ANOVA à... deux facteurs ! C'est parti.*

**Réalisez une ANOVA à deux facteurs**

On souhaite désormais étudier l'influence de deux facteurs qualitatifs **A** et **B** , avec respectivement **I** et **J** modalités, sur une variable quantitative.

On suppose ici que l'on dispose d'un plan d'expériences équilibré (avec **r** observations pour chaque croisement des facteurs).

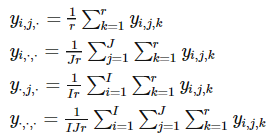
Le modèle considéré est le suivant :



pour 

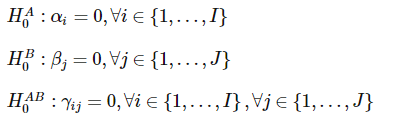
**μ**, les **αi**, les **βj** et les **γi,j** sont des paramètres inconnus, et les **εi,j,k** sont des v.a.r. indépendantes de loi , où  est inconnu.

On considère les quantités suivantes (moyennes : globales, par modalité sur **A** et sur **B**) :



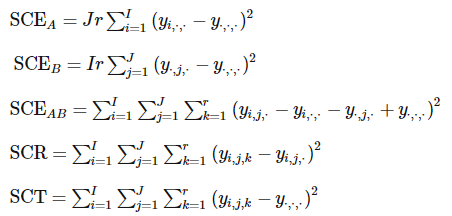
**Les différents tests d'influence**

Les tests effectués sont les suivants :



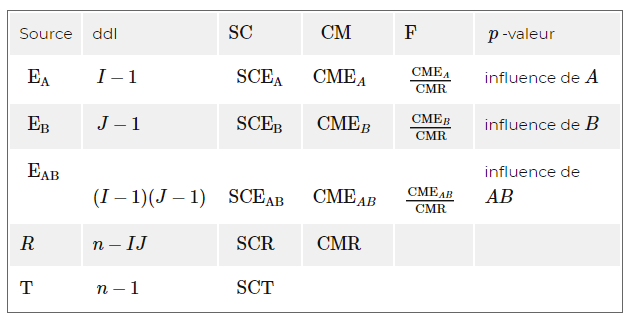
Ils permettent respectivement de tester l'influence du facteur **A**, du facteur **B** et de l'interaction des facteurs **A** et **B**.

On considère les quantités suivantes :



On présente classiquement les résultats sous forme de tableau :

.



On teste tout d'abord l'impact de l'interaction des 2 variables qualitatives sur **Y**.

*Voilà comment l'analyse de la variance se déroule en principe. Maintenant, ouvrez votre logiciel de code préféré, nous allons réaliser une analyse de la variance pour comprendre ce qui influence les rendements de blé.*

## TP : Pratiquez l'analyse de la variance sur le jeu de données du blé

Ce TP est présenté en R. Voici la [**version Python**](https://s3-eu-west-1.amazonaws.com/static.oc-static.com/prod/courses/files/Parcours+Data+Analyst/Re%CC%81alisez+des+mode%CC%81lisations+de+donne%CC%81es+performantes+(1).zip).

Avant de commencer, retrouvez la description du cas d'étude et du jeu de données dans [**ce chapitre**](https://openclassrooms.com/fr/courses/4525326-realisez-des-modelisations-de-donnees-performantes/5754130-decouvrez-le-jeu-de-donnees-du-ble).

Vous allez mener une ANOVA sur notre cas d'étude, en regardant précisément :

* une ANOVA à 1 facteur : la variété de blé ;
* une ANOVA à 1 facteur : le pesticide utilisé ;
* une ANOVA à 2 facteurs : la variété de blé ET le pesticide utilisé.

Dans tous les cas, nous cherchons à**comprendre si les facteurs ont une influence sur le rendement de blé.**

### ****Importez les données****

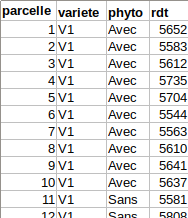
On charge la librairie ggplot2, qui permettra d'afficher les graphiques :

library(ggplot2)

Le fichier "ble.txt" contient les rendements de blé pour 80 parcelles en fonction de la variété de blé (V1, V2, V3 ou V4).

ble <- read.table("ble.txt",header=TRUE,sep=";",dec=".")

ble



Le jeu de données du blé

### Réalisez une ANOVA à 1 facteur

On veut étudier ici l'influence de la variété de blé sur le rendement.

On peut visualiser l'influence de la variété en affichant ces boîtes à moustaches :

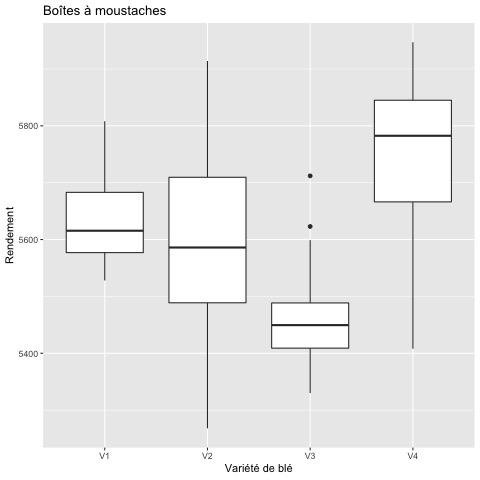
ggplot(ble,aes(x=variete,y=rdt))+

geom\_boxplot()+

ggtitle("Boites à moustaches")+

xlab("Variété de blé")+

ylab("Rendement")



Influence de la variété de blé sur le rendement

Les 4 variétés semblent assez différentes, même si l'ordre de grandeur de ces écarts n'est pas très grand. La question sera de savoir si ces écarts sont significatifs ou pas.

C'est l'ANOVA qui nous permettra de répondre à cette question.

Étudions maintenant l'influence de la présence ou non de pesticide sur le rendement :

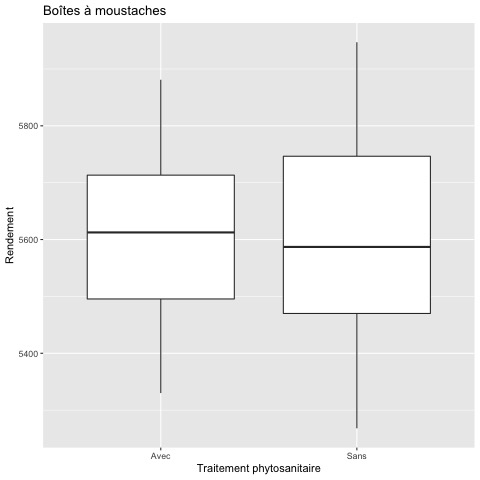
ggplot(ble,aes(x=phyto,y=rdt))+

geom\_boxplot()+

ggtitle("Boites à moustaches")+

xlab("Traitement phytosanitaire")+

ylab("Rendement")



Présence de pesticide sur le rendement

Ici, les boîtes à moustaches ne sont pas très distinctes, même s'il y a un peu plus de variance dans le cas "SANS pesticide".

La présence de pesticide a-t-elle un impact sur le rendement ? L'ANOVA nous permet de confirmer ou d'infirmer cette intuition.

Lançons l'ANOVA pour tester l'influence de la variété de blé :

anova\_variete <- lm(rdt~variete,data=ble)

summary(anova\_variete)

Call:

lm(formula = rdt ~ variete, data = ble)

Residuals:

Min 1Q Median 3Q Max

-344.20 -69.30 -6.60 89.15 329.90

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5633.80 26.30 214.211 < 2e-16 \*\*\*

varieteV2 -49.70 37.19 -1.336 0.18546

varieteV3 -169.20 37.19 -4.549 2e-05 \*\*\*

varieteV4 118.40 37.19 3.183 0.00211 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 117.6 on 76 degrees of freedom

Multiple R-squared: 0.4476, Adjusted R-squared: 0.4258

F-statistic: 20.53 on 3 and 76 DF, p-value: 7.674e-10

On y voit les paramètres estimés (dans la colonne "Estimate"), mais ici, ce ne sont pas les paramètres qui nous intéressent le plus.

#### Réalisez un test de Fisher

Ce qui nous intéresse réellement, c'est le **test de Fisher.**  
La p-valeur de ce test est très petite et largement inférieure à 5 %. On rejette donc l'hypothèse H0 selon laquelle .

La variété de blé a donc bien un effet sur le rendement, comme nous en avions l'intuition en regardant les boîtes à moustaches.

Pour obtenir le tableau de l'analyse de la variance, on utilise la commande ANOVA :

anova(anova\_variete)

Voici le résultat :

Analysis of Variance Table

Response: rdt

Df Sum Sq Mean Sq F value Pr(>F)

variete 3 851845 283948 20.525 7.674e-10 \*\*\*

Residuals 76 1051387 13834

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Réalisons maintenant l'Analyse de la Variance sur le pesticide utilisé :

anova\_phyto <- lm(rdt~phyto,data=ble)

summary(anova\_phyto)

anova(anova\_phyto)

On trouve ici une p-valeur de 0.8, ce qui est très au-dessus de 5 %.  
On ne rejette donc pas l'hypothèse H0 selon laquelle **α1=α2=0**.

Il n'y a pas ici d'effet du pesticide sur le rendement de blé, tout au moins pas d'effet significatif.

### Réalisez une ANOVA à 2 facteurs

Jusqu'ici, nous avons étudié les 2 facteurs (variété et pesticide) séparément. Cependant, la variété et le pesticide peuvent avoir des interactions qui influent sur le rendement.

En effet, même si l'on a montré que, globalement, le pesticide n'a pas d'effet sur le rendement, il se peut que, pour une variété précise, il y ait quand même un effet du pesticide sur le rendement.  
L'ANOVA à 2 facteurs va nous permettre d'étudier ces éventuelles interactions :

anova\_variete\_phyto <- lm(rdt~variete\*phyto,data=ble)

summary(anova\_variete\_phyto)

anova(anova\_variete\_phyto)

Voici le résultat :

Analysis of Variance Table

Response: rdt

Df Sum Sq Mean Sq F value Pr(>F)

variete 3 851845 283948 19.5749 2.205e-09 \*\*\*

phyto 1 1008 1008 0.0695 0.7928

variete:phyto 3 5968 1989 0.1371 0.9375

Residuals 72 1044411 14506

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

On voit sur le tableau 3 lignes :

* variete : qui teste l'effet de la variété ;
* phyto : qui teste l'effet du pesticide ;
* variete:phyto : qui teste les interactions pesticide-variété.

La p-valeur des interactions (93,75 %) est très largement supérieure à 5 % ; on en déduit donc que les interactions n'ont pas d'impact sur le rendement.

En pratique, on part toujours du tableau de l'ANOVA à 2 facteurs pour tester les interactions. Si elles sont significatives, on les conserve. Sinon, on teste séparément les 2 facteurs séparément par des ANOVA à 1 facteur.

Vous êtes arrivé à la fin de ce cours... Enfin, presque ! Il ne vous reste plus qu'un quiz pour le terminer.

# Avez-vous compris les enjeux de l'ANOVA ?

Bravo ! Vous avez réussi cet exercice !

### Compétences évaluées

* **Appliquer et interpréter une Analyse de la variance (ANOVA)**

### Question 1

**L'analyse de la variance étudie l'influence :**

* + 

d'une seule variable qualitative

* + 

d'une seule variable quantitative

* + 

d'au plus deux variables qualitatives en pratique

* + 

d'au plus deux variables quantitatives en pratique

### Question 2

**L'estimation des paramètres dans un modèle d'analyse de la variance est :**

* + 

impossible

* + 

possible sous contrainte

* + 

toujours possible

### Question 3

**L'analyse de la variance à un facteur** **teste :**

* + 

la nullité de μ

* + 

la nullité des αi

* + 

la nullité des εi,j

### Question 4

**La statistique utilisée en analyse de la variance suit sous**H0**une loi :**

* + 

de Bernoulli

* + 

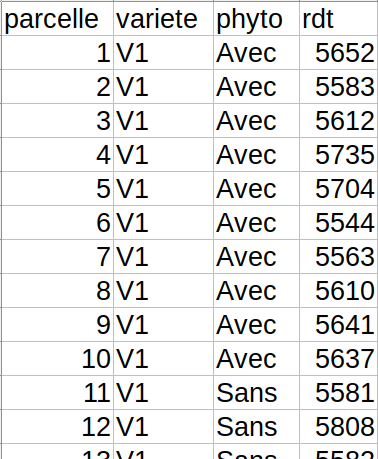
de Fisher

* + 

de Student

### Question 5

**On travaille avec l'échantillon "blé" utilisé dans le cours :**

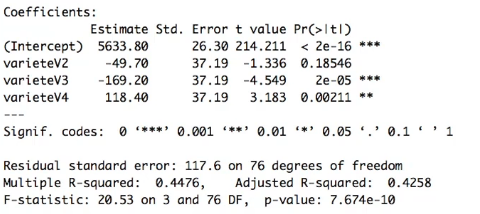
****

**Dans R, on saisit ces lignes de code, afin de réaliser une**ANOVA à 1 facteur**:**

anova\_variete <- lm(rdt~variete,data=ble)

summary(anova\_variete)

**Voici le résultat obtenu :**

****

**Pour le modèle ANOVA, on est obligé de fixer une contrainte pour réaliser l'estimation. Plusieurs contraintes sont possibles. Ici, R a implicitement utilisé la contrainte**α0=0**.**

**Quelles sont les estimations des**αi**correspondant à chaque variété ?**

* + 

|  |  |  |
| --- | --- | --- |
| Variété | i | estimation des αi |
| V1 | 1 | α1^=0 |
| V2 | 2 | α2^=−49.7 |
| V3 | 3 | α3^=−169.2 |
| V4 | 4 | α4^=118.4 |

* + 

|  |  |  |
| --- | --- | --- |
| Variété | i | estimation des αi |
| V1 | 1 | α1^=0 |
| V2 | 2 | α2^=−1.336 |
| V3 | 3 | α3^=−4.549 |
| V4 | 4 | α4^=3.183 |



*Les**sont donnés dans la colonne "Estimate".*

### Question 6

**Nous reprenons l'analyse menée à la question précédente. Selon le résultat obtenu, combien vaut l'estimation de**μ**, c'est-à-dire** **?**

* + 

 μ^=5633.80

* + 

 μ^=26.30

*Il est donné sur la ligne "Intercept", dans la colonne "Estimate" (cf capture d'écran de la question précédente)*