

Démarche statistique

François Husson

Unité pédagogique de mathématiques appliquées
L'Institut Agro

<https://husson.github.io/>

Démarche statistique

- 1 Intro
- 2 Visualisation
- 3 Tests
- 4 Analyse de variance à 1 facteur
- 5 Régression linéaire simple
- 6 Construction et sélection de modèles
- 7 Interprétation et prédiction

Objectifs du module

A la fin de ce module, vous serez capables :

- d'aborder les problèmes les plus courants pour analyser des données
- d'argumenter le choix de procédures d'analyse
- d'évaluer les performances d'une règle de décision statistique
- de mettre en œuvre une démarche d'analyse de données avec 
- d'interpréter et de savoir restituer les résultats d'une analyse

Ce qu'on attend de vous :

- **de la réflexion !!!** (sinon chatGPT serait suffisant !)
- d'être actifs en cours et TD en posant des questions et en participant
- d'avoir une démarche statistique pour étudier un phénomène : identifier des variables, mettre en place une expérimentation pour collecter des données, identifier et hiérarchiser les effets des variables, interpréter et restituer les résultats

Ce qui est clairement insuffisant :

- se contenter de savoir cliquer sur un bouton pour lancer un calcul sur ordinateur
- recopier des lignes de code pour faire une analyse sans la comprendre

Évaluation : 2 contrôles continus des connaissances (50 %) + 1 projet (50 %)

Quelques définitions issues de Wikipédia

- La **statistique** est la discipline qui étudie des phénomènes à travers la collecte de données, leur traitement, leur analyse, l'interprétation des résultats et leur présentation afin de rendre ces données compréhensibles par tous. La statistique est un domaine des mathématiques et de plus en plus, elle fait partie de ce que l'on appelle aujourd'hui la science des données
- La **science des données** est l'étude de l'extraction automatisée de connaissance à partir de grands ensembles de données. Plus précisément, la science des données est un domaine interdisciplinaire qui utilise des méthodes, des processus, des algorithmes et des systèmes scientifiques pour extraire des connaissances et des idées à partir de nombreuses données structurées ou non . Elle est souvent associée aux données massives et à l'analyse des données.
- L'**analyse des données** (aussi appelée analyse exploratoire des données) est une famille de méthodes statistiques dont les principales caractéristiques sont d'être multidimensionnelles et descriptives.
Rq : en anglais *data analysis* \iff "statistique" et non "exploratory data analysis"

\implies La statistique s'intéresse à des jeux de données de taille raisonnable qui sont très fréquents dans vos domaines. De plus, la réflexion sur les données, les modèles, les interprétations est intéressante avant de se confronter à des jeux de données plus massifs.

Les statistiques sont partout

Quelques exemples dans la presse :

- BBC news : *More than half of British women's waists 'too big'* ([lien article](#))
- France Inter : *Groupe sanguin et coronavirus, un hasard génétique* ([lien article](#))
- Libération : *Climat - Le réchauffement en France est dans le haut de la fourchette envisagée par les modèles de prévision pour 2100* ([lien article](#))
- Le Monde : *En Afrique, les paysans qui pratiquent l'agroécologie résistent mieux au changement climatique* ([lien article](#))
- Libération : *Alimentation plus durable : un outil en ligne pour mesurer son impact* ([lien article](#))
- ...
- Réseaux sociaux : être vigilant et critique en gardant toujours en tête cette citation de Umberto Eco (2019) : *Les réseaux sociaux ont donné le droit de parole à des légions d'imbéciles qui avant, ne parlaient qu'au bar et ne causaient aucun tort à la collectivité. On les faisait taire tout de suite. Aujourd'hui, ils ont le même droit de parole qu'un prix Nobel.*

BBC news : More than half of British women's waists 'too big'

Plus de la moitié des femmes Britanniques ont un tour de taille "trop grand"

Des chercheurs de Nuffield Health affirment que les femmes en surpoids risquent davantage de souffrir de maladies cardiaques, de diabète de type 2, d'infertilité et de cancer

Les chercheurs ont constaté que le tour de taille moyen des femmes est de 84.9 cm alors que le tour de taille sain est de 80 cm

Nuffield Health a examiné les données de plus de 30 000 femmes et a constaté que 57 % d'entre elles avaient un tour de taille supérieur au tour de taille sain

Les femmes du nord de l'Angleterre ont le tour de taille le plus large, avec une circonférence moyenne de 87 cm, contre 81.9 cm à Londres

Les chercheurs ont également indiqué que 52.5 % des femmes avaient un IMC supérieur à la fourchette saine (25 - 29.9), et que 16.2 % étaient atteintes d'obésité modérée ou morbide (IMC >30)



Le cœur de la statistique

- Avoir les (bonnes) **données**
 - comprendre à partir d'**observations** (de données) un phénomène
 - recueillir des données en s'assurant qu'elles représentent bien le phénomène et qu'elles permettent de répondre à la question qu'on se pose
 - besoin de les résumer, les **visualiser**
 - elles sont indispensables pour **modéliser**
- Se poser les bonnes questions
 - les données sont-elles **représentatives** ? on s'intéresse à la pollution des plages par les algues vertes.
 - Comment mesurer la pollution ? Quelle **variable** utiliser ? surface ? volume ? épaisseur ?
 - Où faire des recueils ? Quand ? Sur les plages françaises ? En Bretagne ? A date fixe ? Après des grandes marées ? etc. En se basant sur les quantités ramassées par les communes ? attention au **biais** !!
 - dans le phénomène étudié, qu'est-ce qui est la cause et qu'est-ce qui est la conséquence ? "vitesse du vent - production électrique des éoliennes", "concentration en nitrates les sols - surface d'algues vertes", "variété de plantes - rendement", "genre - niveau d'étude - salaire"
⇒ la (ou les) cause(s) seront les **variables explicatives**, la conséquence sera la **variable réponse** (on dit encore variable à expliquer)
 - tous les effets (toutes les causes) ont-ils été pris en compte ?

Pourquoi ?

 est un logiciel de développement scientifique pour le calcul et l'analyse statistique

- R est gratuit et libre
- R est disponible sous Linux, Mac, Windows
- R couvre de nombreux domaines d'application
- forte communauté d'utilisateurs - nombreux packages, forums, cours

Installation du logiciel

- 1 Télécharger puis installer  <https://cran.r-project.org/>
- 2 Télécharger puis installer (après avoir installé R)  Studio
<https://posit.co/download/rstudio-desktop/>
- 3 (Utiliser avec Quarto ou Rmarkdown pour faire de la recherche reproductible)

Quelques définitions

- **Population** - ensemble d'entités objet de l'investigation statistique
- **Individu** - élément de la population d'étude
- **Variable/Attribut** - descripteur ou caractère des individus de la population d'étude
- **Echantillon** - ensemble des individus pour lesquels des valeurs ont été observées pour les variables de l'étude
- **Inférence** - décider pour une population à partir des données observées de l'échantillon

- Nature des variables : on distingue deux grandes familles de variables
 - **qualitative** : les valeurs prises sont des **modalités**
 - nominale : pas de structure d'ordre (sexe, couleur des cheveux)
 - ordinale : modalités intrinsèquement ordonnées (niveau de vie)
 - **quantitative** : les valeurs prises sont **numériques**
 - discrète : à valeurs dans un ensemble dénombrable (nombre d'enfants)
 - continue : à valeurs dans un ensemble indénombrable (taille, poids)

Définition de quelques indicateurs pour décrire les données

Définition d'**indicateurs de position** et d'**indicateurs de dispersion** (ces définitions sont généralement connues, intuitives, et ... disponibles sur internet !)

Soit x_1, x_2, \dots, x_n une série de n valeurs d'une variable X

- **moyenne** : $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$
- **médiane** : $q_{0.5}(x)$ = valeur telle que 50 % des x_i ont une valeur inférieure et 50% une valeur supérieure
- **1er quartile** : $q_{0.25}(x)$ = valeur telle que 25 % des x_i ont une valeur inférieure et 75% une valeur supérieure
- **3ème quartile** : $q_{0.75}(x)$ = valeur telle que 75 % des x_i ont une valeur inférieure et 25% une valeur supérieure
- **quantile α** : $q_\alpha(x)$ = valeur telle que $100 \times \alpha$ % des x_i ont une valeur inférieure et $100 - 100 \times \alpha$ % une valeur supérieure
- **variance** : $s^2(x) = \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
- **écart-type** : racine carrée de la variance (même unité que X)
- **étendue** : différence entre le maximum et le minimum

Un jeu de données en guise d'exemple

L'association Air Breizh surveille la qualité de l'air et mesure la concentration de polluants comme l'ozone (O₃) ainsi que les conditions météorologiques comme la température, la nébulosité, le vent, etc.

Durant l'été 2001, 112 données ont été relevées à Rennes

```
ozone <- read.table("https://r-stat-sc-donnees.github.io/ozone.txt",
  header=TRUE, stringsAsFactors = TRUE)
```

| | maxO3 | T9 | T12 | T15 | Ne9 | Ne12 | Ne15 | Vx9 | Vx12 | Vx15 | maxO3v | vent | pluie |
|------------|-------|------|------|------|-----|------|------|---------|---------|---------|--------|-------|-------|
| 2001-06-01 | 87 | 15.6 | 18.5 | 18.4 | 4 | 4 | 8 | 0.6946 | -1.7101 | -0.6946 | 84 | Nord | Sec |
| 2001-06-02 | 82 | 17 | 18.4 | 17.7 | 5 | 5 | 7 | -4.3301 | -4 | -3 | 87 | Nord | Sec |
| 2001-06-03 | 92 | 15.3 | 17.6 | 19.5 | 2 | 5 | 4 | 2.9544 | 1.8794 | 0.5209 | 82 | Est | Sec |
| 2001-06-04 | 114 | 16.2 | 19.7 | 22.5 | 1 | 1 | 0 | 0.9848 | 0.3473 | -0.1736 | 92 | Nord | Sec |
| 2001-06-05 | 94 | 17.4 | 20.5 | 20.4 | 8 | 8 | 7 | -0.5 | -2.9544 | -4.3301 | 114 | Ouest | Sec |
| 2001-06-06 | 80 | 17.7 | 19.8 | 18.3 | 6 | 6 | 7 | -5.6382 | -5 | -6 | 94 | Ouest | Pluie |
| 2001-06-07 | 79 | 16.8 | 15.6 | 14.9 | 7 | 8 | 8 | -4.3301 | -1.8794 | -3.7588 | 80 | Ouest | Sec |
| ... | ... | ... | | | | | | | | | | | |

Leur objectif : prévoir la concentration en ozone du lendemain pour avertir la population en cas de pic de pollution

Démarche statistique

- 1 Intro
- 2 Visualisation**
- 3 Tests
- 4 Analyse de variance à 1 facteur
- 5 Régression linéaire simple
- 6 Construction et sélection de modèles
- 7 Interprétation et prédiction

Visualisation

La visualisation dépend

- des données (nature des variables, nombre d'individus, etc.)
- de la problématique

La visualisation permet :

- de comprendre et vérifier ses données
- de suggérer des analyses ou des modélisations
- de faire passer des idées (lors de la restitution de l'analyse)

Sur 

- la classique fonction `plot`
- la fonction `ggplot` du package `ggplot2`

Pour choisir un graphique adapté à ses données et sa problématique :

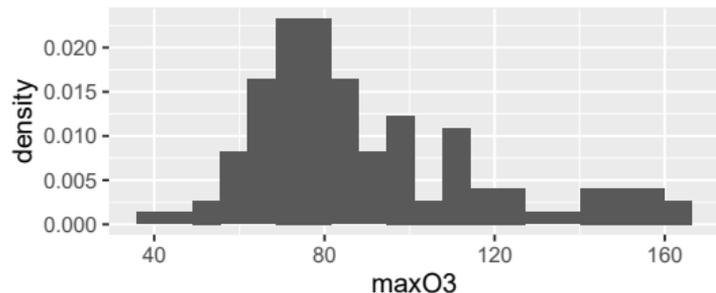
<https://www.data-to-viz.com/>

Qq exemples de graphiques beaucoup plus sophistiqués : dataviz-inspiration.com

Visualiser la distribution d'une variable quantitative

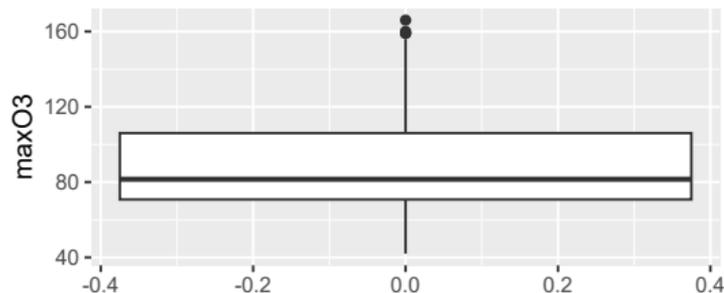
Choisir son jeu de données + ce qui est en x + ce qui est en y +
le type de représentation + le titre + ...

Histogramme du maximum d'ozone



```
library(ggplot2)
ggplot(ozone) + aes(x=maxO3)+
  aes(y = after_stat(density)) +
  geom_histogram(bins=20) +
  ggtitle("Histogramme du maximum d'ozone")
```

Boîte à moustaches du maximum d'ozone



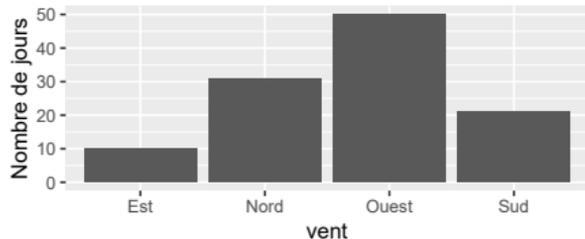
```
ggplot(ozone) +
  aes(y=maxO3) +
  geom_boxplot() +
  ggtitle("Boîte à moustaches du maximum d'ozone")
```

Visualiser la distribution d'une variable qualitative

Diagramme en barres (ordonné) **JAMAIS** de camembert!! ([lien pour s'en convaincre](#))

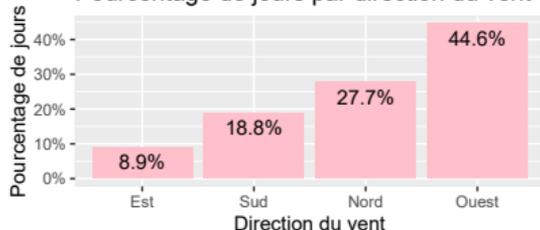
Choisir données + ce qui est en x + en y + le type de représentation + le titre + ...

Nombre de jours par direction de vent



```
ggplot(ozone) +
  aes(x=vent) +
  geom_bar() +
  ylab("Nombre de jours") +
  ggtitle("Nombre de jours par direction de vent")
```

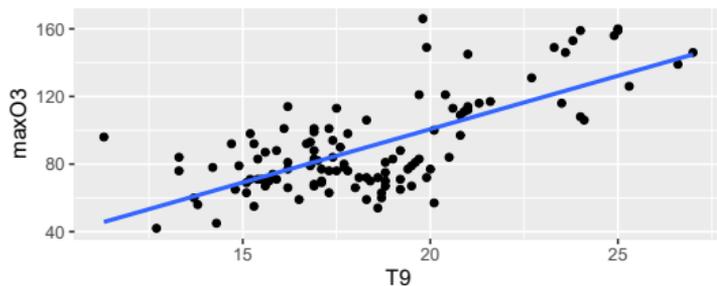
Pourcentage de jours par direction du vent



```
ggplot(ozone)+ aes(x=reorder(vent,vent,length)) +
  aes(y = ..count../sum(..count..)) +
  geom_bar(fill="pink") +
  geom_text(aes(label=scales::percent(..count../sum(..count..)),
    y= ..count../sum(..count..)), stat="count", vjust=1.5) +
  xlab("Direction du vent") + ylab("Pourcentage de jours") +
  scale_y_continuous(labels=scales::percent) +
  ggtitle("Pourcentage de jours par direction du vent")
```

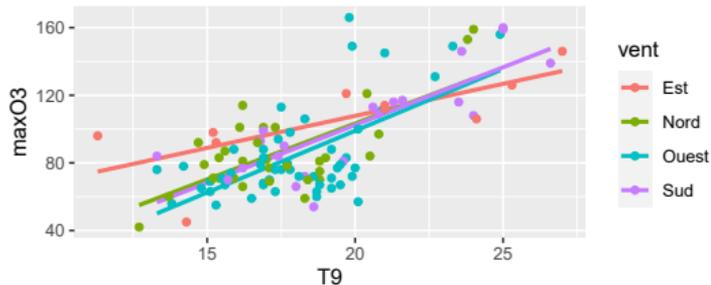
Visualiser l'effet d'1 ou deux variables sur une variable quanti

Effet de T9 sur le maximum d'ozone



```
library(ggplot2)
ggplot(ozone) +
  aes(x=T9, y=maxO3) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  ggtitle("Effet de T9 sur le maximum d'ozone")
```

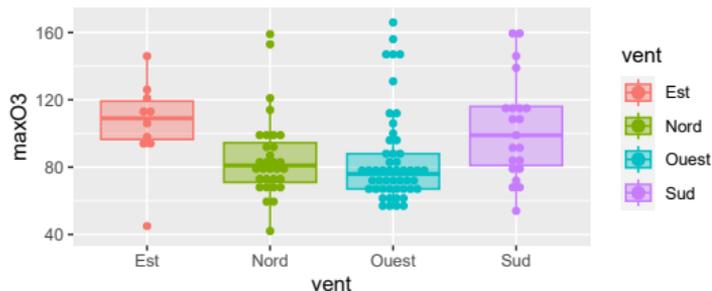
Effet de T9 sur le maximum d'ozone selon le vent



```
ggplot(ozone) +
  aes(x=T9, y=maxO3, col = vent, group=vent) +
  geom_smooth(method="lm", se=FALSE) +
  geom_point()+
  ggtitle("Effet de T9 sur MaxO3 selon le vent")
```

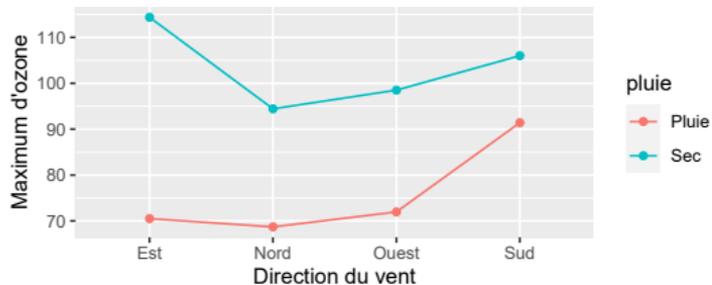
Visualiser l'effet d'une ou deux variables quali sur une variable quanti

Effet du vent sur le maximum d'ozone



```
ggplot(ozone) +
  aes(x=vent, y=maxO3, fill=vent, col=vent) +
  geom_boxplot(outlier.shape=NA,alpha=0.4) +
  geom_dotplot(binaxis = "y",stackdir = "center") +
  ggtitle("Effet du vent sur le maximum d'ozone")
```

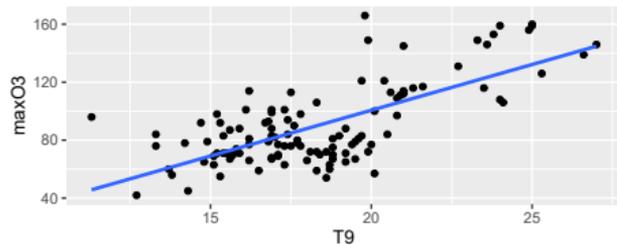
Interaction pluie:vent sur le maximum d'ozone



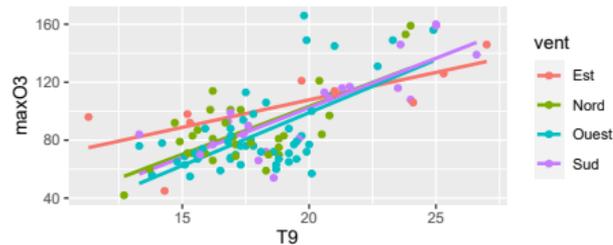
```
library(dplyr)
ozone |> group_by(vent, pluie) |>
  summarize(MOY = mean(maxO3)) |>
  ggplot() +
  aes(x=vent, y=MOY, col=pluie, group=pluie) +
  geom_line() + geom_point() +
  ggtitle("Interaction pluie:vent sur maxO3")+
  xlab("Direction du vent")+
  ylab("Maximum d'ozone")
```

Et que suggèrent ces visualisations ?

Effet de T9 sur le maximum d'ozone

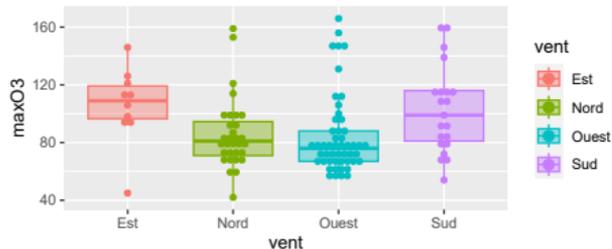


Effet de T9 sur le maximum d'ozone selon le vent

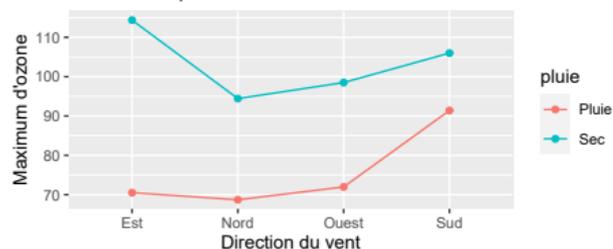


Y a-t-il un effet linéaire de T9 sur maxO3 ? L'effet dépend-il de la direction du vent ?

Effet du vent sur le maximum d'ozone



Interaction pluie:vent sur le maximum d'ozone



Les maximums d'ozone sont-ils (en moyenne) plus importants pour certaines directions du vent ? Est-ce que l'effet de la direction du vent dépend de la pluie ?

⇒ Et surtout : **tout ceci est-il généralisable ?**

Démarche statistique

- 1 Intro
- 2 Visualisation
- 3 Tests**
- 4 Analyse de variance à 1 facteur
- 5 Régression linéaire simple
- 6 Construction et sélection de modèles
- 7 Interprétation et prédiction

Principe des tests

Quelle question se pose-t-on ?

Y-a-t'il un effet de telle variable sur ma variable d'intérêt (la variable réponse) ?

- la direction du vent a-t-elle un effet sur le maximum d'ozone ?
- la température à 9h influe-t-elle sur le maximum d'ozone ?
- le temps (sec ou pluvieux) a-t-il un effet sur le maximum d'ozone ?
 - si c'est pour voir que systématiquement quand le temps est sec le max d'ozone est plus élevé, inutile de faire des stats !
 - on se restreint souvent à comparer les moyennes (éventuellement aussi les variances)
 - on ne se contente pas de voir sur les données observées que la moyenne du max d'ozone des jours de pluie est plus petite que la moyenne des jours ensoleillés !

Objectif : ne pas seulement constater sur les données observées, mais généraliser au-delà des données de l'échantillon

⇒ **inférence** à (tous les individus de) la population

⇒ les différences sont-elles **significatives** ?

Problème : comment gérer l'incertitude liée au fait qu'on a observé qu'une petite partie des données ?

Retour sur l'intervalle de confiance d'une moyenne

Grâce au **théorème central limite**, la moyenne \bar{X} de n valeurs issues d'une variable X de loi quelconque, de moyenne μ et de variance σ^2 , suit la loi (si $n > 30$ ou si $X \sim \mathcal{N}$) :

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n) \quad \text{et donc} \quad \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0, 1)$$

En pratique, σ^2 est très rarement connu, et donc doit être estimé à partir de l'échantillon \implies augmente un peu l'incertitude d'où l'utilisation d'une loi de Student à $(n - 1)$ degrés

de liberté (plutôt qu'une loi normale) : $\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim \mathcal{T}(n - 1)$

D'où l'intervalle de confiance de μ au niveau de confiance 95 % :

```
t.test(ozone$max03)
```

```
[...]
```

```
95 percent confidence interval:
```

```
85.02578 95.58136
```

```
sample estimates:
```

```
mean of x
```

```
90.30357
```

De l'intervalle de confiance au test de conformité d'une moyenne

Connaissant la loi de \bar{X} : $\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim \mathcal{T}(n-1)$ il est possible de tester si la moyenne de la population (μ) est égale à une valeur particulière

Par exemple, si la valeur de la moyenne vaut $\mu = 100$, alors $\frac{\bar{X}-100}{\sqrt{S^2/n}} \sim \mathcal{T}(n-1)$

A partir d'un échantillon de données, on peut calculer \bar{x} et s^2 et voir si $\frac{\bar{x}-100}{\sqrt{s^2/n}}$ peut provenir d'une loi de Student à $(n-1)$ degrés de liberté, ou bien si la valeur ne provient très certainement pas d'une loi de Student car elle est trop extrême

- **Hypothèse** $H_0 : \mu = 100$ contre $H_1 : \mu \neq 100$
- **Statistique de test** : $T = \frac{\bar{X}-100}{\sqrt{S^2/n}}$
- **Loi de la statistique** de test sous H_0 : $\mathcal{T}(n-1)$
- **p-value** du test : probabilité, calculée sous H_0 , que la statistique de test soit plus extrême que la valeur observée T_{obs}
 \implies si p-value < 0.05 on rejette l'hypothèse H_0 au seuil 5 %

Test de conformité d'une moyenne

```
t.test(ozone$maxO3, mu=100, alternative="two.sided")
```

One Sample t-test

```
data: ozone$maxO3
```

```
t = -3.6406, df = 111, p-value = 0.0004148
```

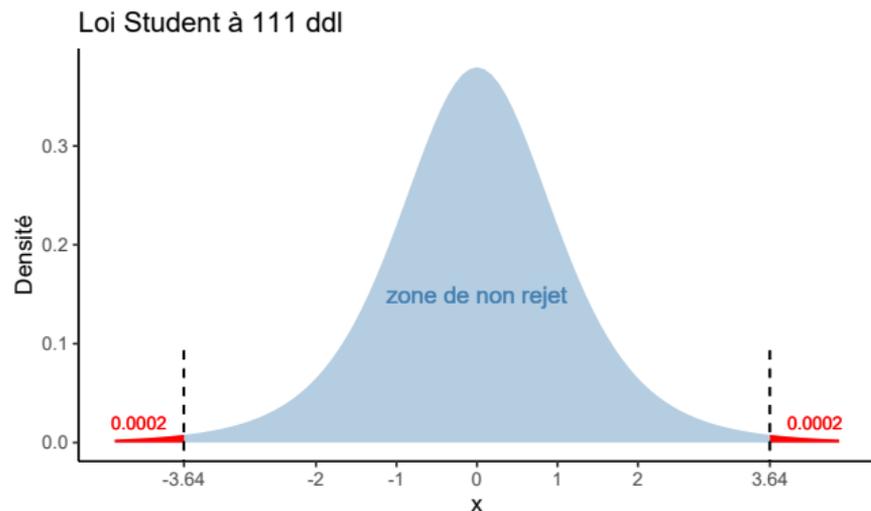
```
alternative hypothesis: true mean is not equal to 100
```

```
[...]
```

$p\text{-value} < 0.05 \implies$ rejet de H_0 au seuil 5 %

\implies on affirme que la **moyenne** du max d'ozone est **significativement** différente de 100 (avec un niveau de confiance de 95 %)

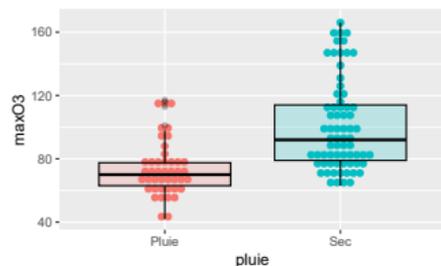
$T_{obs} = -3.64$ extrême pour une loi de Student à 111 ddl, d'où une valeur de p-value faible (0.0004148)



Remarque : l'hypothèse alternative du test est $H_1 : \mu \neq 100$ (alternative="two.sided")
 H_1 aurait pu être $H_1 : \mu < 100$ (avec "less") ou $H_1 : \mu > 100$ (avec "greater")

Test de comparaison de 2 moyennes

Question : Le temps (sec ou pluvieux) a-t-il un effet sur le maximum d'ozone ?

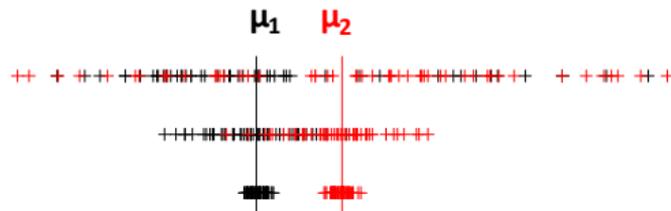


```
ggplot(ozone)+ aes(x=pluie, y=maxO3, fill=pluie,col=pluie) +
  geom_dotplot(binaxis = "y",stackdir = "center",dotsize = 1) +
  geom_boxplot(alpha=.2,col="black") +
  theme(legend.position = "none")
```

La généralisation à n'importe quel jour pluvieux ou sec induit de l'incertitude puisqu'on a seulement vu les données de l'échantillon

Il sera plus facile de conclure que la différence des 2 moyennes est significative

- ① si les moyennes sont très différentes
- ② si la variabilité du maximum d'ozone est faible entre les jours pluvieux, et faible entre les jours secs



- ③ s'il y a beaucoup de données

Test de comparaison de 2 moyennes

On considère que les données de la sous-population 1 sont telles que

$(X_{i1})_{1 \leq i \leq n_1} \sim \mathcal{N}(\mu_1, \sigma^2)$ et que les données de la sous-population 2 sont telles que

$(X_{i2})_{1 \leq i \leq n_2} \sim \mathcal{N}(\mu_2, \sigma^2) \Rightarrow$ seules les moyennes peuvent être différentes

Si $\mu_1 = \mu_2$ alors toutes les données proviennent d'une même loi $\mathcal{N}(\mu, \sigma^2)$

- **Hypothèses** : $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$
- **Statistique de test** : $T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\hat{\sigma}^2}{n_1} + \frac{\hat{\sigma}^2}{n_2}}}$
- **Loi de la statistique de test sous H_0** : $Student(n_1 + n_2 - 2)$
- **Décision** avec la p-value



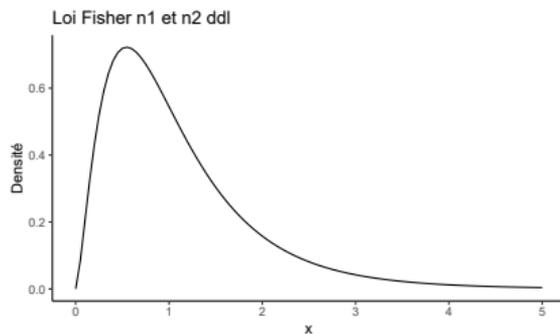
Si les variances sont inégales, le test est légèrement différent \implies besoin de tester l'égalité des variances avant de faire le test de comparaison de moyennes

Test de comparaison de 2 variances

- **Hypothèses** : $H_0 : \sigma_1^2 = \sigma_2^2$ contre $H_1 : \sigma_1^2 \neq \sigma_2^2$
 $\iff H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ contre $H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$

- **Statistique de test** : $F = \frac{S_1^2}{S_2^2}$

- **Loi de la statistique de test sous H_0** : F_{n_1-1, n_2-1}
- **Décision** avec la p-value



Test de comparaison de (2 variances puis de) 2 moyennes

```
var.test(max03 ~pluie, data = ozone, alternative="two.sided")
```

F test to compare two variances

data: max03 by pluie

F = 0.35906, num df = 42, denom df = 68, p-value = 0.0005659

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.2108847 0.6338374

sample estimates:

ratio of variances

0.3590605

⇒ on rejette l'hypothèse H_0 , on affirme que les variances sont différentes au seuil 5 %

```
t.test(max03 ~pluie, data = ozone, var.equal=FALSE, alternative="two.sided")
```

Welch Two Sample t-test

data: max03 by pluie

t = -6.3362, df = 109.88, p-value = 5.321e-09

alternative hypothesis: true difference in means between group Pluie and group Sec is not equal to 0

95 percent confidence interval:

-36.02936 -18.86110

sample estimates:

mean in group Pluie mean in group Sec

73.39535

100.84058

⇒ Rejet de H_0 ⇒ on affirme que les moyennes sont significativement différentes au seuil 5 %

Le test est "two.sided" $\Leftrightarrow H_1 : \mu_1 \neq \mu_2$ (sinon "less" $\Leftrightarrow \mu_2 < \mu_1$ ou "greater" $\Leftrightarrow \mu_2 > \mu_1$)

Erreur et puissance de test

Toute décision d'un test est prise pour la population alors que seules les données d'un échantillon sont observées

⇒ la décision prise est incertaine et donc deux erreurs sont possibles :

- l'**erreur de 1ère espèce** : rejeter l'hypothèse H_0 alors que celle-ci est vraie : test faux positif
- l'**erreur de 2ème espèce** : ne pas rejeter H_0 alors que celle-ci n'est pas vraie (et donc c'est H_1 qui est vraie) : test faux négatif

Ces deux types d'erreur ont été largement commentés lors des tests COVID !!

On veut absolument maîtriser la 1ère erreur (d'où l'utilisation de seuils inférieurs à 5 % pour α et pour comparer la p-value).

Remarque de R.A. Fisher (1935) : *The null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation*

Un test **puissant** (dont l'erreur de 2ème espèce est petite) est efficace pour prendre une décision. La **puissance d'un test** est la probabilité de rejeter H_0 et d'avoir raison

Nombre d'individus pour atteindre une puissance de test

Question souvent posée : combien faut-il recueillir de données, combien d'individus, pour tester s'il y a une différence entre 2 moyennes ?

Exemple en médecine : combien faut-il de patients pour tester s'il y a une différence entre 2 médicaments pour faire baisser le taux de cholestérol ?

⇒ question beaucoup trop imprécise à laquelle il est impossible de répondre !!

Reformulation de la question : combien faut-il de patients si on veut mettre en évidence une différence d'au moins 0.2 g/l entre les 2 médicaments ? (on sous-entend qu'en deçà de 0.2 g/l, même s'il y a une différence significative entre les 2 médicaments, celle-ci n'est pas suffisante pour choisir un médicament plutôt qu'un autre)

⇒ on ne peut toujours pas répondre à cette question, **MAIS ...**

Nombre d'individus pour atteindre une puissance de test

MAIS ...

- si on connaît la variance de Y , i.e. des taux de cholestérol après traitement (par ex. des expériences antérieures montrent que l'écart-type vaut 0.4 g/l)
- si on on veut détecter une différence de 0.2 g/l
- si on veut détecter la différence avec une probabilité de 80 %

```
power.t.test(delta = 0.2, sd=0.4, power = .80)
```

```
Two-sample t test power calculation
```

```
      n = 63.76576
delta = 0.2
      sd = 0.4
sig.level = 0.05
      power = 0.8
alternative = two.sided
```

alors c'est possible !!!

Ca fait beaucoup de "si" mais c'est possible : il faut connaître la différence de moyennes (delta) que l'on veut détecter, l'écart-type (sd) et la puissance du test (power)

```
NOTE: n is number in *each* group
```

Remarque : il faut des patients avant traitement avec des taux de cholestérol proches, et randomiser le médicament attribué à chaque patient

Démarche statistique

- 1 Intro
- 2 Visualisation
- 3 Tests
- 4 Analyse de variance à 1 facteur**
- 5 Régression linéaire simple
- 6 Construction et sélection de modèles
- 7 Interprétation et prédiction

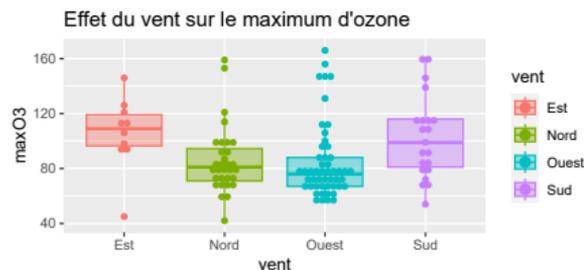
Comparaison de \bar{Y} moyennes ou test de l'effet d'un facteur

La direction du vent a-t-elle un effet sur le maximum d'ozone ?

Variable **réponse quantitative**, notée Y : maxO3

Variable **explicative qualitative à l modalités** (l groupes) : direction du vent

| vent | maxO3 |
|-------|-------|
| Nord | 87 |
| Nord | 82 |
| Est | 92 |
| Nord | 114 |
| Ouest | 94 |
| ... | ... |



```
ggplot(ozone) +
  aes(x=vent, y=maxO3, fill=vent, col=vent) +
  geom_boxplot(outlier.shape=NA, alpha=0.4) +
  geom_dotplot(binaxis = "y", stackdir = "center") +
  ggtitle("Effet du vent sur le maximum d'ozone")
```

Question modifiée en : la moyenne du maximum d'ozone est-elle la même pour chaque direction du vent ?

Problématique très fréquente

- L'appréciation globale est-elle différente d'un jus d'orange à un autre ?
 - Un jus d'orange particulier est-il plus ou moins apprécié que la moyenne ?
- Les modes de production raisonnée, bio ou à bas niveau d'intrants ont-ils des effets significativement différents sur la qualité de l'eau des cours d'eau ?
- Les consommateurs perçoivent-ils des différences de saveur selon le type de cuisson (basse température, à l'étouffée, à la vapeur) ?
-

Étude de l'effet d'une variable **qualitative** (facteur) à I modalités (ou niveaux) sur une variable **quantitative**

Autrement dit, est-ce que les différences de modalités de la variable explicative expliquent la variabilité (ou la variance) de la variable réponse ?

Données, notations

| vent | maxO3 | Notation |
|-------|-------|----------|
| Nord | 87 | y_{21} |
| Nord | 82 | y_{22} |
| Est | 92 | y_{11} |
| Nord | 114 | y_{23} |
| Ouest | 94 | y_{31} |
| Ouest | 80 | y_{32} |
| Ouest | 79 | y_{33} |
| ... | ... | ... |

y_{ij} valeur de Y pour l'individu j du groupe i

$$y_{i\bullet} = \sum_{j=1}^{n_i} y_{ij}$$

$$y_{\bullet\bullet} = \sum_{i=1}^I \sum_{j=1}^{n_i} y_{ij}$$

Comparaison de l moyennes

Dans chaque groupe, on considère que les données suivent une loi $\mathcal{N}(\mu_i, \sigma^2)$

NB : on considère le **même** σ^2

$$Y_{1j} \sim \mathcal{N}(\mu_1, \sigma^2) \text{ peut s'écrire } Y_{1j} \sim \mu_1 + \varepsilon_{1j} \text{ avec } \varepsilon_{1j} \sim \mathcal{N}(0, \sigma^2)$$

$$Y_{2j} \sim \mathcal{N}(\mu_2, \sigma^2) \text{ peut s'écrire } Y_{2j} \sim \mu_2 + \varepsilon_{2j} \text{ avec } \varepsilon_{2j} \sim \mathcal{N}(0, \sigma^2)$$

...

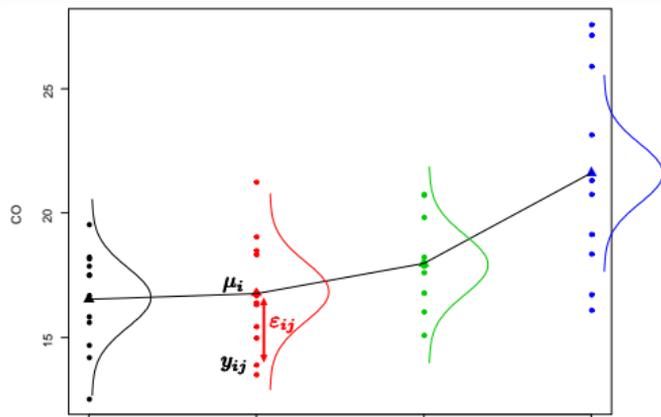
...

$$Y_{lj} \sim \mathcal{N}(\mu_l, \sigma^2) \text{ peut s'écrire } Y_{lj} \sim \mu_l + \varepsilon_{lj} \text{ avec } \varepsilon_{lj} \sim \mathcal{N}(0, \sigma^2)$$

⇒ En regroupant tout ce modèles en un seul modèle, on écrit :

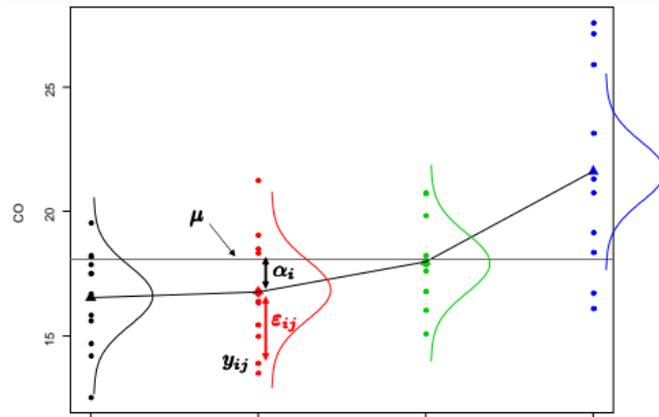
$$\begin{cases} \forall i, j & Y_{ij} = \mu_i + \varepsilon_{ij} \\ \forall i, j & \mathcal{L}(\varepsilon_{ij}) = \mathcal{N}(0, \sigma^2) \\ \forall (i, j) \neq (i', j') & \text{cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0 \end{cases}$$

Définitions du modèle



$$\begin{cases} \forall i, j & Y_{ij} = \mu_i + \varepsilon_{ij} \\ \forall i, j & \mathcal{L}(\varepsilon_{ij}) = \mathcal{N}(0, \sigma^2) \\ \forall (i, j) \neq (i', j') & \text{cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0 \end{cases}$$

l paramètres : les μ_i



$$\begin{cases} \forall i, j & Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \\ \forall i, j & \mathcal{L}(\varepsilon_{ij}) = \mathcal{N}(0, \sigma^2) \\ \forall (i, j) \neq (i', j') & \text{cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0 \end{cases}$$

$l + 1$ paramètres : μ (effet moyen), les l coefficients α_i (effet du niveau i)

On doit avoir : $\forall i, \mu + \alpha_i = \mu_i \implies$ besoin de **contraintes**

Estimation des paramètres du modèle

Pour estimer les paramètres du modèle (les μ_i pour le modèle 1 OU μ et les α_i pour le modèle 2), on minimise les écarts entre observations (Y_{ij}) et prédictions par le modèle (\hat{Y}_{ij}). Plus précisément, on minimise le **critère des moindres carrés ordinaires** (MCO)

$$SCER = \sum_{i=1}^I \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_{ij})^2$$

$$SCER = \sum_{i=1}^I \sum_{j=1}^{n_i} (Y_{ij} - \hat{\mu}_i)^2 = \sum_{i=1}^I \sum_{j=1}^{n_i} (Y_{ij} - (\hat{\mu} + \hat{\alpha}_i))^2$$

$$SCER \text{ minimal quand } \forall i, \hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = Y_{i\bullet}$$

Contraintes classiques :

- Un niveau particulier comme référence : $\alpha_1 = 0 \implies \hat{\mu} = Y_{1\bullet}$ et $\forall i, \hat{\alpha}_i = Y_{i\bullet} - Y_{1\bullet}$
- La moyenne des moyennes par groupe comme référence : $\mu = \sum_{i=1}^I \mu_i$

$$\sum_{i=1}^I \alpha_i = 0 \implies \hat{\mu} = \frac{1}{I} \sum_{i=1}^I Y_{i\bullet}, \quad \forall i, \hat{\alpha}_i = Y_{i\bullet} - \hat{\mu}$$

Estimation des paramètres

Exemple :

| <i>groupe1</i> | <i>groupe2</i> | <i>groupe3</i> | |
|--------------------|--------------------|--------------------|-------------------------------|
| $y_{11} = 6$ | $y_{21} = 2$ | $y_{31} = 3$ | |
| $y_{12} = 9$ | $y_{22} = 4$ | $y_{32} = 1$ | |
| $y_{13} = 6$ | $y_{23} = 3$ | | |
| $y_{1\bullet} = 7$ | $y_{2\bullet} = 3$ | $y_{3\bullet} = 2$ | $y_{\bullet\bullet} = 11.333$ |

- Avec traitement 1 comme référence : $\hat{\mu} = 7$; $\hat{\alpha}_1 = 0$; $\hat{\alpha}_2 = -4$ et $\hat{\alpha}_3 = -5$
- Avec $\sum_{i=1}^I \alpha_i = 0$: $\hat{\mu} = \frac{1}{3}(7 + 3 + 2) = 4$; $\hat{\alpha}_1 = +3$; $\hat{\alpha}_2 = -1$ et $\hat{\alpha}_3 = -2$

L'interprétation d'un coefficient α_j dépend de la contrainte choisi

Variance résiduelle

Prédiction et résidus

Valeurs prédites :

$$\hat{y}_{ij} = \hat{\mu} + \hat{\alpha}_i = y_{i\bullet}$$

Résidus :

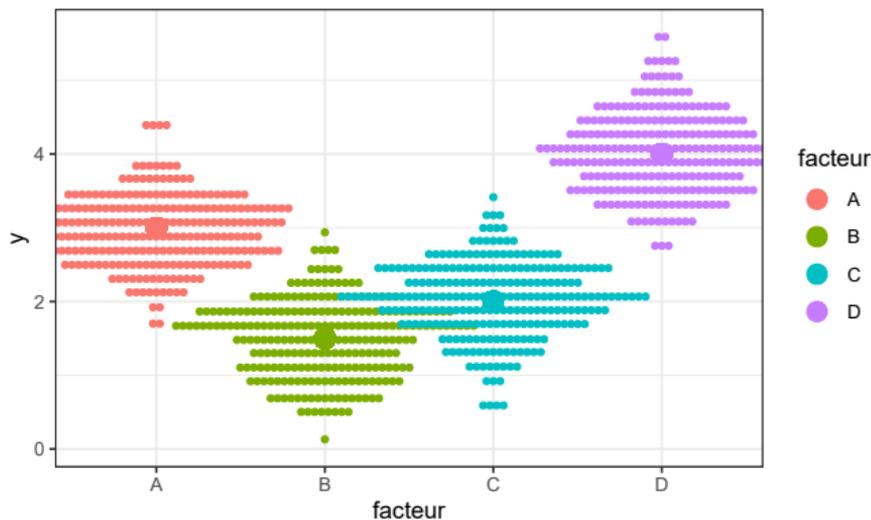
$$e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - y_{i\bullet}$$

Estimateur de la variabilité résiduelle σ^2 :

$$\hat{\sigma}^2 = \frac{\sum_{ij} (Y_{ij} - Y_{i\bullet})^2}{n - l} = \frac{\sum_{ij} \varepsilon_{ij}^2}{n - l} \quad \mathbb{E}(\hat{\sigma}^2) = \sigma^2$$

Exemple : $\hat{\sigma}^2 = \frac{(6-7)^2 + (9-7)^2 + \dots + (1-2)^2}{8-3} = \frac{10}{8-3} = 2$

Décomposition de la variabilité



Equation d'analyse de la variance

$$\underbrace{\sum_{i,j} (Y_{ij} - Y_{..})^2}_{SC_T} = \underbrace{\sum_{i,j} (Y_{i.} - Y_{..})^2}_{SC_F} + \underbrace{\sum_{i,j} (Y_{ij} - Y_{i.})^2}_{SC_R}$$

Variabilité
ddl

totale
 $n - 1$

modèle
 $l - 1$

résiduelle
 $n - l$

Indicateur de liaison : rapport de corrélation

$$\eta^2 = \frac{SC_{\text{modèle}}}{SC_{\text{total}}} = 1 - \frac{SC_{\text{résiduelle}}}{SC_{\text{total}}}$$

Propriétés :

- $0 \leq \eta^2 \leq 1$
- $\eta^2 = 0 \Leftrightarrow SC_{\text{modèle}} = 0 \Leftrightarrow \forall i, y_{i\bullet} = y_{\bullet\bullet}$
- $\eta^2 = 1 \Leftrightarrow SC_{\text{modèle}} = SC_{\text{total}} \Leftrightarrow \forall(i, j), y_{ij} = y_{i\bullet}$

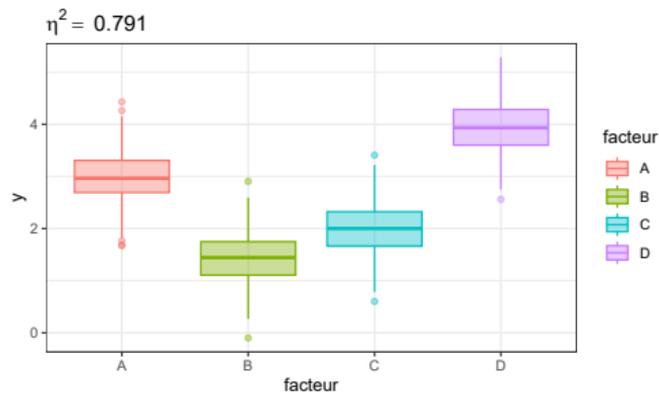
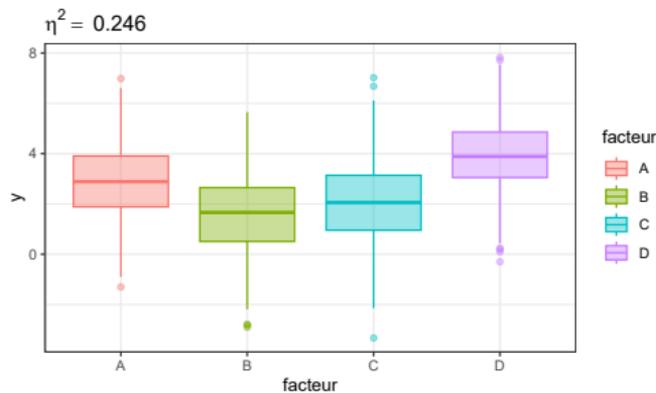


Figure – Rapports de corrélation différents pour des $Y_{i\bullet}$ identiques

Inférence : test global

Objectifs : y a-t'il un effet significatif du facteur sur Y ?

Autrement dit, la variabilité de Y est-elle expliquée par le facteur groupe ? ou bien peut-on considérer que les données proviennent d'une même loi $\mathcal{N}(\mu, \sigma^2)$

Hypothèses :

$$H_0 : "\forall i, \mu_i = \mu"$$

 \Leftrightarrow

$$H_0 : "\forall i, \alpha_i = 0"$$

$$H_1 : "\exists i / \mu_i \neq \mu"$$

$$H_1 : "\exists i / \alpha_i \neq 0"$$

$$\text{On a : } \mathbb{E} \left(\frac{SC_{mod}}{I-1} \right) = \sigma^2 + \frac{1}{I-1} \sum_{i=1}^I n_i \alpha_i^2 \quad \mathbb{E} \left(\frac{SC_R}{n-I} \right) = \sigma^2$$

Idée pour le test ? ...

Statistique de test : $F_{obs} = \frac{SC_{mod}/(I-1)}{SC_R/(n-I)}$

Loi de la stat de test sous H_0 : $\mathcal{L}(F_{obs}) = \mathcal{F}_{n-I}^{I-1}$

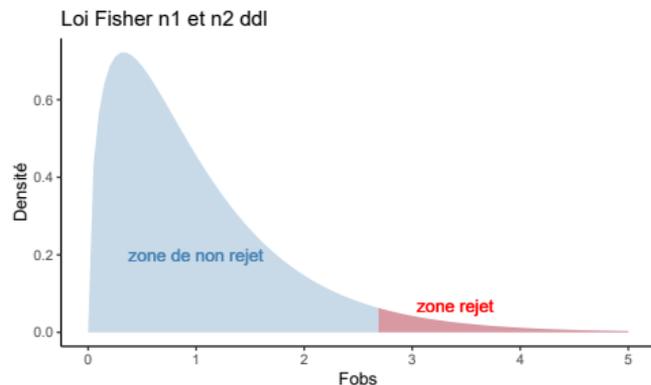


Table d'analyse de variance : décomposition de variabilité et test

| Variabilité | Somme Carrés | ddl | Carré moyen | F_{obs} |
|-------------|--|---------|----------------------|---------------------|
| Facteur | $\sum_i n_i (Y_{i\bullet} - Y_{\bullet\bullet})^2$ | $l - 1$ | $\frac{SC_F}{l - 1}$ | $\frac{CM_F}{CM_R}$ |
| Résiduelle | $\sum_{i,j} (Y_{ij} - Y_{i\bullet})^2$ | $n - l$ | $\frac{SC_R}{n - l}$ | |
| Totale | $\sum_{i,j} (Y_{ij} - Y_{\bullet\bullet})^2$ | $n - 1$ | | |

```
library(FactoMineR)
```

```
LinearModel(max03~vent,data=ozone)
```

```
LinearModel(formula = max03 ~ vent, data = ozone)
```

```
Residual standard error: 27.32 on 108 degrees of freedom
```

```
Multiple R-squared: 0.08602
```

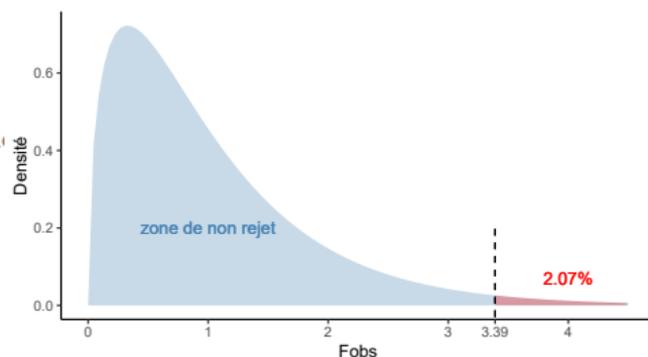
```
F-statistic: 3.388 on 3 and 108 DF, p-value: 0.02074
```

```
AIC = 744.8 BIC = 755.7
```

```
Ftest
```

| | SS | df | MS | F value | Pr(>F) |
|-----------|-------|-----|---------|---------|---------|
| vent | 7586 | 3 | 2528.69 | 3.3881 | 0.02074 |
| Residuals | 80606 | 108 | 746.35 | | |

Loi Fisher 3 et 108 ddl



⇒ rejet de H_0 au seuil 5 %

Inférence : test de conformité d'un coefficient

Si on rejette l'hypothèse $H_0 : \forall i, \alpha_i = 0$, on veut savoir quels α_i sont différents de 0
 La valeur de $\hat{\alpha}_i$ dépend de l'échantillon de données et donc l'estimateur $\hat{\alpha}_i$ est une variable aléatoire

$$\mathcal{L}(\hat{\alpha}_i) = \mathcal{N}(\alpha_i, \sigma_{\hat{\alpha}_i}^2) \iff \mathcal{L}\left(\frac{\hat{\alpha}_i - \alpha_i}{\sigma_{\hat{\alpha}_i}}\right) = \mathcal{N}(0, 1)$$

$$\mathcal{L}\left(\frac{\hat{\alpha}_i - \alpha_i}{\hat{\sigma}_{\hat{\alpha}_i}}\right) = \mathcal{T}_{n-1}$$

On peut donc construire le test de nullité d'un coefficient (α_1 par exemple) :

Hypothèses : $H_0 : \alpha_1 = 0$ contre $H_1 : \alpha_1 \neq 0$

Statistique de test : $\frac{\hat{\alpha}_1}{\hat{\sigma}_{\hat{\alpha}_1}}$

Loi de la statistique de test sous H_0 : $\mathcal{L}\left(T_{obs} = \frac{\hat{\alpha}_1}{\hat{\sigma}_{\hat{\alpha}_1}}\right) = \mathcal{T}_{\nu=n-1}$

Décision : par la p-value

Rq : connaissant la loi de $\hat{\alpha}_1$, on peut construire un intervalle de confiance :

$$\alpha_1 \in [\hat{\alpha}_1 - \hat{\sigma}_{\hat{\alpha}_1} \times t_{0.975}(n-1) ; \hat{\alpha}_1 + \hat{\sigma}_{\hat{\alpha}_1} \times t_{0.975}(n-1)]$$

Inférence : test de conformité d'un coefficient

```
library(FactoMineR)
res <- LinearModel(maxO3~vent,data=ozone)
res

Ttest
```

| | Estimate | Std. Error | t value | Pr(> t) |
|--------------|----------|------------|---------|----------|
| (Intercept) | 94.7382 | 3.0535 | 31.0265 | < 2e-16 |
| vent - Est | 10.8618 | 6.8294 | 1.5904 | 0.11466 |
| vent - Nord | -8.6092 | 4.6219 | -1.8627 | 0.06522 |
| vent - Ouest | -10.0382 | 4.0972 | -2.4500 | 0.01589 |
| vent - Sud | 7.7856 | 5.2052 | 1.4957 | 0.13764 |

Quand le vent vient de l'ouest, la moyenne du maximum d'ozone est **significativement inférieure** au maximum d'ozone moyen

Pour les autres directions du vent, le maximum d'ozone n'est pas significativement différent du maximum d'ozone moyen.

Test de comparaison 2 à 2

Autre stratégie : comparer toutes les paires de moyennes

Pb : on effectue beaucoup de tests \implies risque de multiplier les erreurs en rejetant des hypothèses H_0

Idée de **correction des tests** : Bonferroni propose de modifier le seuil $\alpha = 5\%$ de chaque test et de prendre $\alpha = 5\% / (\text{nb tests})$

\implies les tests sont peu puissants

```
library(FactoMineR)
```

```
res <- LinearModel(max03~vent,data=ozone)
```

```
meansComp(res,~vent, adjust="Bonferroni")
```

```
$adjMean
```

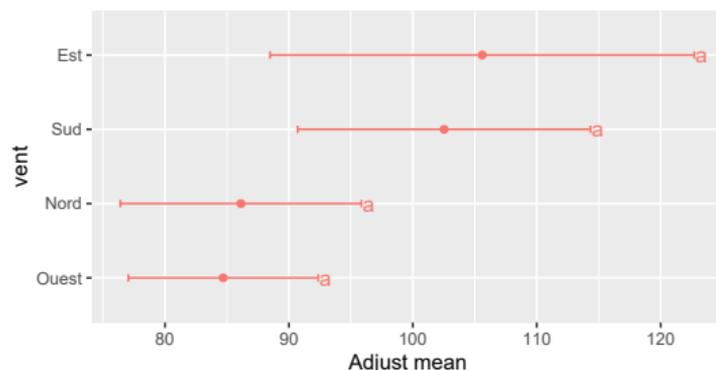
| vent | emmean | SE | df | lower.CL | upper.CL |
|-------|--------|------|-----|----------|----------|
| Est | 105.6 | 8.64 | 108 | 88.5 | 122.7 |
| Nord | 86.1 | 4.91 | 108 | 76.4 | 95.9 |
| Ouest | 84.7 | 3.86 | 108 | 77.0 | 92.4 |
| Sud | 102.5 | 5.96 | 108 | 90.7 | 114.3 |

Confidence level used: 0.95

```
$groupComp
```

```
$groupComp$Letters
```

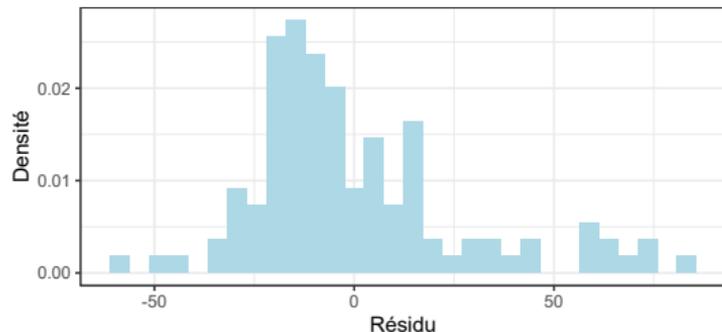
| Ouest | Nord | Sud | Est |
|-------|------|-----|-----|
| "a" | "a" | "a" | "a" |



Analyse graphique des résidus du modèle

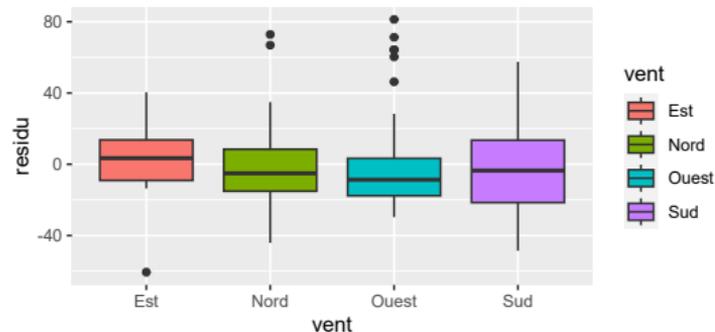
On vérifie les hypothèses du modèle pour savoir si on peut le construire **APRÈS** avoir construit le modèle (c'est tordu ... mais c'est comme ça !)

Histogramme des résidus



```
ggplot(data.frame(residu=res$lmResult$residuals))+
  aes(x=residu,y=density)+
  aes(y = after_stat(density))+
  geom_histogram(fill="lightblue")+
  ggtitle("Histogramme des résidus")+
  theme_bw() + xlab("Résidu")+ylab("Densité")
```

Boxplot des résidus par vent



```
ggplot(data.frame(residu=res$lmResult$residuals,
  vent=ozone$vent))+
  aes(y=residu,x=vent,fill=vent)+geom_boxplot()+
  ggtitle("Boxplot des résidus par vent")
```

La variabilité

- La variabilité est la notion de base dans une approche modèle
⇒ Toute la complexité d'un phénomène est quantifiée par sa variabilité

- Un objectif statistique en modélisation est de rendre compte du maximum de variabilité d'un phénomène en tenant compte du minimum d'information
 - L'information utilisée est caractérisée par la dimension des facteurs ou degrés de liberté
 - Compromis classique en statistique entre qualité d'ajustement et parcimonie

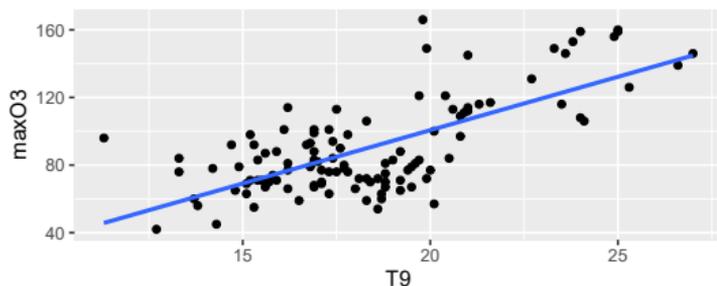
Objectif : Évaluer la variabilité expliquée par un facteur en pondérant par rapport à la dimension (ou degré de liberté) de ce facteur

Démarche statistique

- 1 Intro
- 2 Visualisation
- 3 Tests
- 4 Analyse de variance à 1 facteur
- 5 Régression linéaire simple**
- 6 Construction et sélection de modèles
- 7 Interprétation et prédiction

Liaison linéaire simple

Effet de T9 sur le maximum d'ozone



```
library(ggplot2)
ggplot(ozone) +
  aes(x=T9, y=maxO3) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  ggtitle("Effet de T9 sur le maximum d'ozone")
```

Questions :

- La température à 9h influe-t-elle sur le maximum d'ozone ?
- À quel maximum d'ozone peut-on s'attendre si la température à 9h est de 19°C ?

Objectifs :

- Étudier qualitativement et quantitativement la dépendance d'une variable réponse quantitative Y en fonction d'une variable quantitative x
- La variable x permet-elle d'expliquer la variabilité de la variable Y ?
- Prédire Y à partir de x

Un indicateur de liaison : le coefficient de corrélation

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

- $-1 \leq r_{xy} \leq 1$
- $r_{xy} = 0 \Leftrightarrow$ pas de liaison linéaire entre X et Y
- $r_{xy} \approx 1 \Leftrightarrow$ relation linéaire croissante entre X et Y
- $r_{xy} \approx -1 \Leftrightarrow$ relation linéaire décroissante entre X et Y

```
cor.test(ozone$maxO3, ozone$T9)
```

```
Pearson's product-moment correlation
```

```
data: ozone$maxO3 and ozone$T9
```

```
t = 10.263, df = 110, p-value < 2.2e-16
```

```
alternative hypothesis: true correlation is not equal to 0
```

```
95 percent confidence interval:
```

```
0.5904575 0.7832906
```

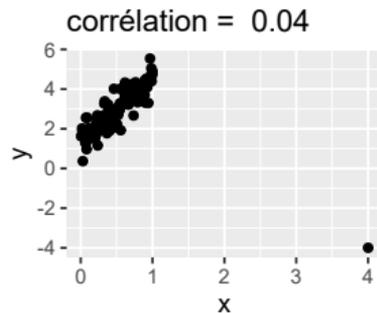
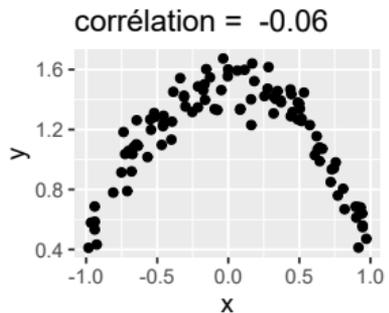
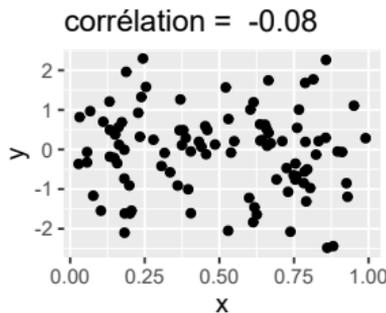
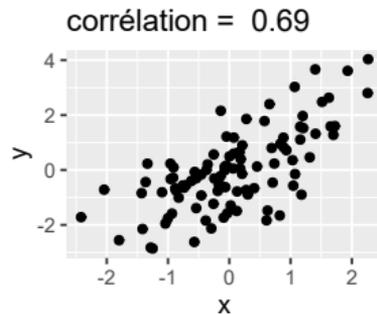
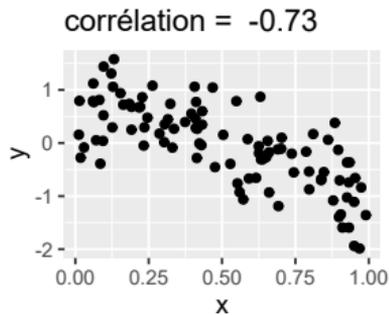
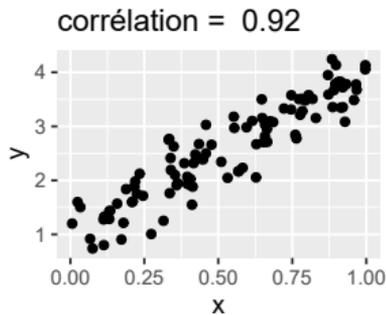
```
sample estimates:
```

```
cor
```

```
0.6993865
```

\implies corrélation de 0.699 entre maxO3 et T9 ... et cette corrélation est significativement différente de 0 (test qui correspond au test de $\beta_1 = 0$ qui sera vu plus tard)

Illustration de quelques corrélations



Le modèle de régression simple

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \stackrel{ind}{\sim} \mathcal{N}(0, \sigma^2),$$

avec

- x_i la valeur de la variable explicative pour l'observation k
 - $i = 1, \dots, n$ le numéro d'individu, n le nombre total d'individus
 - β_0 l'ordonnée à l'origine
 - β_1 la pente de la droite, mesure de l'effet de la variable x
 - σ^2 la variance
-
- Interprétabilité
 - Approximation simple (à l'ordre 1) de toute forme de relation

Estimation des paramètres

Minimisation du critère des moindres carrés :

$$SCER = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Paramètres d'espérance :

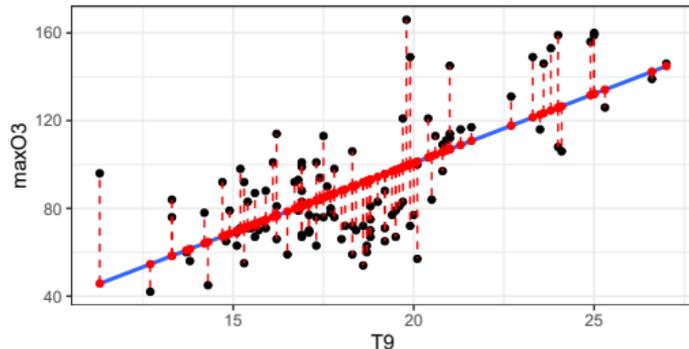
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Variance résiduelle :

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

$$ddl_{\text{résiduelle}} = n - 2 \quad \mathbb{E}(\hat{\sigma}^2) = \sigma^2$$



```
res <- LinearModel(maxO3~T9,data=ozone)
```

```
res
```

```
Ttest
```

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|-----------|------------|---------|----------|
| (Intercept) | -25.60761 | 11.45512 | -2.2355 | 0.02741 |
| T9 | 6.31300 | 0.61514 | 10.2627 | < 2e-16 |

Test de conformité

$$\mathcal{L}(\hat{\beta}_1) = \mathcal{N}(\beta_1, \sigma_{\beta_1}^2) \quad \text{avec } \sigma_{\beta_1}^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\implies \mathcal{L}\left(\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}\right) = \mathcal{N}(0, 1) \quad \implies \mathcal{L}\left(\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}}\right) = \mathcal{T}_{n-2}$$

On peut donc construire le test de **nullité** de β_1 :

Hypothèses : H_0 : " $\beta_1 = 0$ " contre H_1 : " $\beta_1 \neq 0$ "

Statistique de test : $\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$

Loi de la statistique de test sous H_0 : $\mathcal{L}\left(T_{obs} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}\right) = \mathcal{T}_{\nu=n-2}$

Décision : par la p-value

Rq : connaissant la loi de $\hat{\beta}_1$, on peut construire un intervalle de confiance :

$$\beta_1 \in \left[\hat{\beta}_1 - \hat{\sigma}_{\hat{\beta}_1} \times t_{0.975}(n-2) ; \hat{\beta}_1 + \hat{\sigma}_{\hat{\beta}_1} \times t_{0.975}(n-2) \right]$$

Test et intervalle de confiance sur β_0

$$\mathcal{L}(\hat{\beta}_0) = \mathcal{N}(\beta_0, \sigma_{\beta_0}^2) \quad \text{avec} \quad \sigma_{\beta_0}^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

$$\mathcal{L}\left(\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}}\right) = \mathcal{T}_{n-2}$$

Même test possible que pour $\beta_1 = 0$, mais le test sur β_0 est très rarement utile !

Intervalle de confiance de β_0 :

$$\beta_0 \in \left[\hat{\beta}_0 - \hat{\sigma}_{\hat{\beta}_0} \times t_{0.975}(n-2) ; \hat{\beta}_0 + \hat{\sigma}_{\hat{\beta}_0} \times t_{0.975}(n-2) \right]$$

Décomposition de la variabilité

$$\begin{aligned} \sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\ SCT &= SCM + SCR \\ n-1 &= 1 + n-2 \end{aligned}$$

Table d'analyse de la variance

| Source variation | Somme de carrés | Degrés de liberté | Carré moyen | F |
|------------------|-----------------|-------------------|-------------------|-------------------|
| Modèle | SCM | 1 | $\frac{SCM}{1}$ | $\frac{CMM}{CMR}$ |
| Erreur | SCR | n-2 | $\frac{SCR}{n-2}$ | |
| Total | SCT | n-1 | | |

Comparaison de SCM et SCR par le critère $R^2 = \frac{SCM}{SCT}$

Propriétés :

- $0 \leq R^2 \leq 1$
- $R^2 = 0 \Leftrightarrow SC_{\text{modèle}} = 0$
- $R^2 = 1 \Leftrightarrow SC_{\text{modèle}} = SC_{\text{total}}$

Test du modèle

Hypothèses :

$H_0 : \beta_1 = 0 \Leftrightarrow$ le modèle n'a pas d'intérêt $\Leftrightarrow x$ n'explique pas Y

$H_1 : \beta_1 \neq 0 \Leftrightarrow$ le modèle présente de l'intérêt $\Leftrightarrow x$ explique Y

Statistique de Fisher

$$F = \frac{SCM/1}{SCR/n-2}$$

Loi de F sous H_0

$$\mathcal{L}(F) = \mathcal{F}_{n-2}^1$$

Décision : avec la p-value

Remarque : Lien entre F et T_{β_1} : $F = T_{\beta_1}^2$

Prédiction et intervalle de confiance de prédiction

La prédiction de Y pour une valeur x_0 particulière est simplement :

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

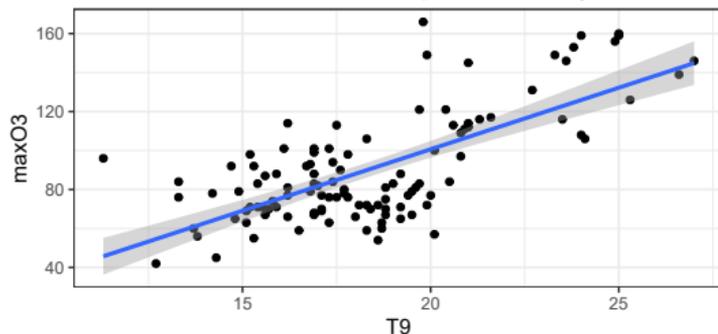
Prédiction de la valeur moyenne de Y
pour un x_0 particulier

$$E(\hat{Y}_0) \sim \mathcal{N} \left(Y_0, \sigma \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right)$$

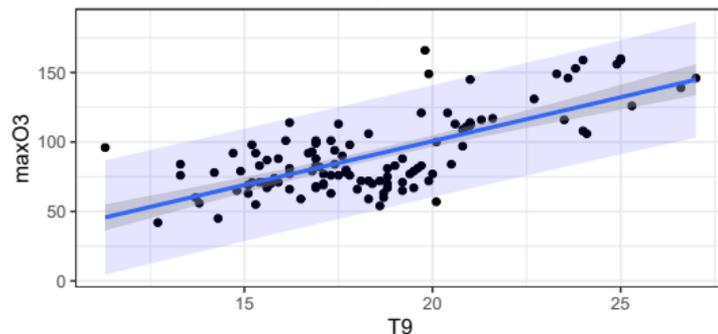
Prédiction d'une nouvelle valeur de Y
pour un x_0 donné

$$\hat{Y}_0 \sim \mathcal{N} \left(Y_0, \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right)$$

Intervalle de confiance du comportement moyen



Intervalle de confiance de prédiction



Exemple

```
res <- LinearModel(maxO3~T9,data=ozone)
```

```
res
```

```
Call:
```

```
LinearModel(formula = maxO3 ~ T9, data = ozone)
```

```
Residual standard error: 20.24 on 110 degrees of freedom
```

```
Multiple R-squared: 0.4891
```

```
F-statistic: 105.3 on 1 and 110 DF, p-value: 9.7e-16
```

```
AIC = 675.7 BIC = 681.1
```

```
Ftest
```

| | SS | df | MS | F value | Pr(>F) |
|-----------|-------|-----|-------|---------|-----------|
| T9 | 43138 | 1 | 43138 | 105.32 | < 2.2e-16 |
| Residuals | 45053 | 110 | 410 | | |

```
Ttest
```

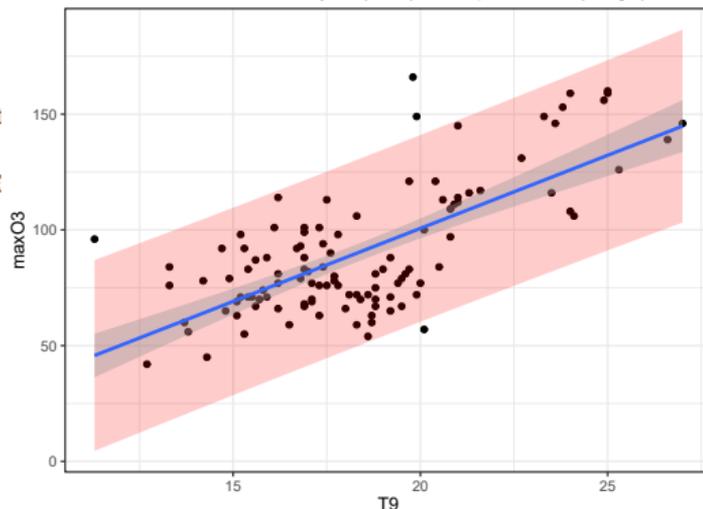
| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|-----------|------------|---------|----------|
| (Intercept) | -25.60761 | 11.45512 | -2.2355 | 0.02741 |
| T9 | 6.31300 | 0.61514 | 10.2627 | < 2e-16 |

```
pred_interval <- predict(res, interval="prediction", level = 0.95)
```

```
pred_interval <- data.frame(pred_interval,T9=ozone$T9)
```

```
ggplot(ozone) + geom_point(aes(x=T9, y= maxO3)) +
  geom_ribbon(data=pred_interval, aes(x=T9, ymin=lwr, ymax=upr), fill="red", alpha=0.2) +
  geom_smooth(method="lm", se=TRUE, aes(x=T9, y=maxO3))+
  ggtitle("Intervalle de confiance moyen (bleu) et de prédiction (rouge)")+theme_bw()
```

Intervalle de confiance moyen (bleu) et de prédiction (rouge)

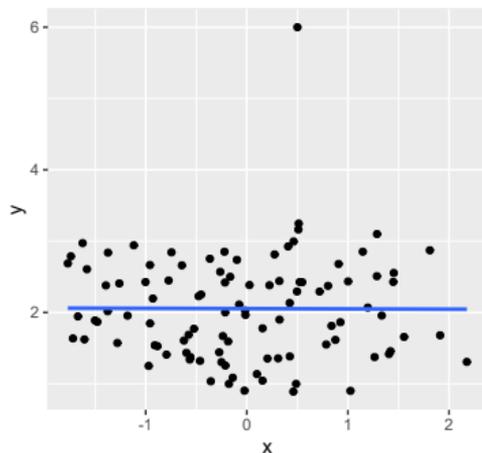


Influence des données et validité du modèle

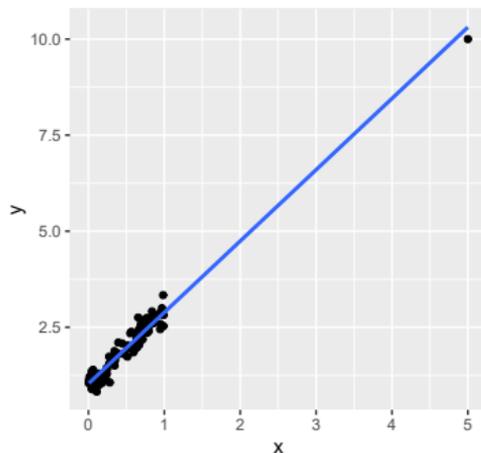
Vérification des hypothèses de départ : normalité des **résidus**, stabilité de la variance, indépendance des résidus

⇒ Études graphiques

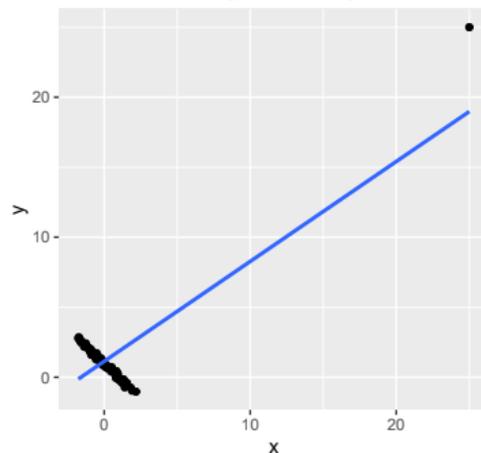
Point aberrant



Point levier avec faible résidu



Point levier très problématique



Démarche statistique

- 1 Intro
- 2 Visualisation
- 3 Tests
- 4 Analyse de variance à 1 facteur
- 5 Régression linéaire simple
- 6 Construction et sélection de modèles
 - Introduction modélisation
 - Modèle
 - Décomposition de la variabilité
 - Tests
 - Sélection de variables

Modéliser c'est comprendre et prévoir

Qui a déjà fait des modélisations complexes? TOUS ... facilement et sans le savoir!

- Prévoir un temps de trajet
- Estimer le temps d'attente à un guichet
- Estimer si le prix de location de l'appartement est juste
- Choisir de prendre un vêtement de pluie pour la journée (prévoir s'il va pleuvoir)

Et comment faites-vous pour prévoir?

- ① vous listez toutes les **variables** (les effets) qui peuvent influencer sur votre **réponse**
- ② vous **éliminez** les variables qui sont **négligeables**
- ③ vous essayez de **quantifier l'effet** des variables restantes **sélectionnées**

Votre intuition est-elle bien raisonnable? OUI, C'EST PARFAIT

A quoi servent les statistiques alors? A faire tout cela avec rigueur pour des phénomènes parfois plus complexes

Données, problématique

- Prévoir le temps de cuisson idéal en fonction de la composition et du poids de l'aliment, de la température du four, de l'humidité de l'air, ...
- Prévoir la production de biogaz en fonction de la quantité de déchets agricoles ou alimentaires, des résidus de cultures, d'ordures ménagères ou des restaurants, etc.
- Prévoir la production d'une éolienne en fonction de la vitesse du vent à 10m, à 80m, de la température à 2m, de la pression, de l'humidité relative à 2m, (de la direction du vent)
- Comprendre ce qui influe sur le pourcentage de surface à bas niveau d'intrants d'une zone en fonction du type de culture, de la mise en place ou non d'un programme d'aide, si la zone se situe dans une aire de captage
- Optimiser une réaction chimique en fonction du temps et de la température

Objectifs :

- **Comprendre** quelles variables influent sur une variable **réponse quantitative**
- **Prévoir** les valeurs de la variable réponse pour de nouvelles conditions

Données, problématique

L'association Air Breizh surveille la qualité de l'air et mesure la concentration de polluants comme l'ozone (O_3) ainsi que les conditions météorologiques comme la température, la nébulosité, le vent, etc.

Durant l'été 2001, 112 données ont été relevées à Rennes

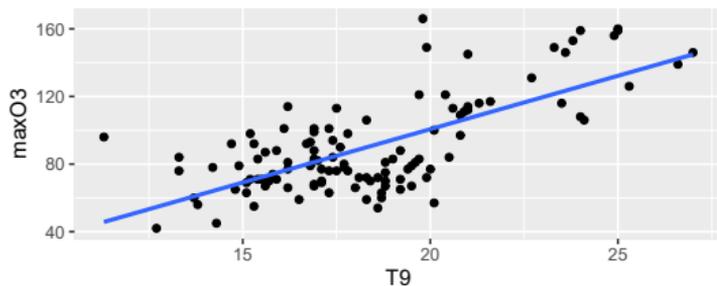
```
ozone <- read.table("https://r-stat-sc-donnees.github.io/ozone.txt",
  header=TRUE, stringsAsFactors = TRUE)
```

| | maxO3 | T9 | T12 | T15 | Ne9 | Ne12 | Ne15 | Vx9 | Vx12 | Vx15 | maxO3v | vent | pluie |
|------------|-------|------|------|------|-----|------|------|---------|---------|---------|--------|-------|-------|
| 2001-06-01 | 87 | 15.6 | 18.5 | 18.4 | 4 | 4 | 8 | 0.6946 | -1.7101 | -0.6946 | 84 | Nord | Sec |
| 2001-06-02 | 82 | 17 | 18.4 | 17.7 | 5 | 5 | 7 | -4.3301 | -4 | -3 | 87 | Nord | Sec |
| 2001-06-03 | 92 | 15.3 | 17.6 | 19.5 | 2 | 5 | 4 | 2.9544 | 1.8794 | 0.5209 | 82 | Est | Sec |
| 2001-06-04 | 114 | 16.2 | 19.7 | 22.5 | 1 | 1 | 0 | 0.9848 | 0.3473 | -0.1736 | 92 | Nord | Sec |
| 2001-06-05 | 94 | 17.4 | 20.5 | 20.4 | 8 | 8 | 7 | -0.5 | -2.9544 | -4.3301 | 114 | Ouest | Sec |
| 2001-06-06 | 80 | 17.7 | 19.8 | 18.3 | 6 | 6 | 7 | -5.6382 | -5 | -6 | 94 | Ouest | Pluie |
| 2001-06-07 | 79 | 16.8 | 15.6 | 14.9 | 7 | 8 | 8 | -4.3301 | -1.8794 | -3.7588 | 80 | Ouest | Sec |
| ... | ... | ... | | | | | | | | | | | |

Leur objectif : prévoir la concentration en ozone du lendemain pour avertir la population en cas de pic de pollution

Visualisation de l'effet linéaire d'une variable quantitative

Effet de T9 sur le maximum d'ozone



```
library(ggplot2)
ggplot(ozone) +
  aes(x=T9, y=maxO3) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  ggtitle("Effet de T9 sur le maximum d'ozone")
```

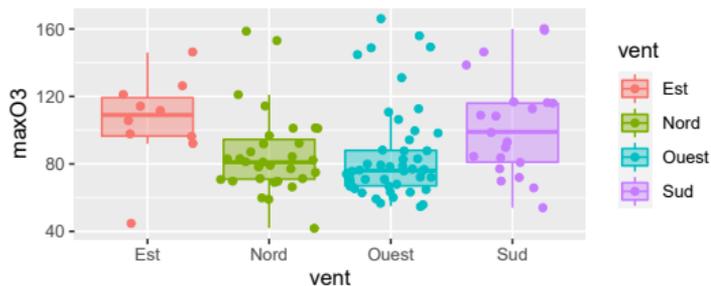
$$MaxO3 \sim T9 \implies MaxO3_i = \beta_0 + \beta_1 \times T9_i + alea_i$$

$$\text{Réponse} \sim var_1 + var_2 + \dots + var_p$$

$$\begin{cases} \forall i = 1, \dots, n & Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 v_i + \beta_3 w_i + \dots + \beta_p z_i + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots + \beta_p x_{ip} + \varepsilon_i \\ \forall i = 1, \dots, n & \varepsilon_i \text{ i.i.d.}, \mathbb{E}(\varepsilon_i) = 0, \mathbb{V}(\varepsilon_i) = \sigma^2 \\ \forall i \neq k & cov(\varepsilon_i, \varepsilon_k) = 0 \end{cases}$$

Visualisation de l'effet d'une variable qualitative

Effet du vent sur le maximum d'ozone



```
ggplot(ozone) +
  aes(x=vent, y=maxO3, fill=vent, col=vent) +
  geom_boxplot(outlier.shape=NA,alpha=0.4) +
  geom_jitter()+
  ggtitle("Effet du vent sur le maximum d'ozone")
```

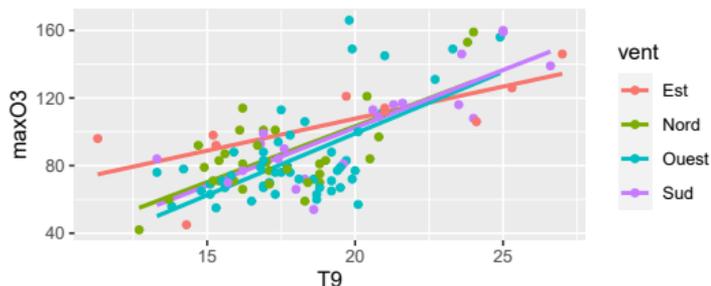
$$MaxO3 \sim vent \implies MaxO3_{ij} \sim \mu + \left\{ \begin{array}{l} \alpha_1 \text{ si vent d'est} \\ \alpha_2 \text{ si vent du nord} \\ \alpha_3 \text{ si vent d'ouest} \\ \alpha_4 \text{ si vent du sud} \end{array} \right\} + alea_{ij}$$

Réponse $\sim var_1 + var_2 + \dots$

$$\left\{ \begin{array}{l} \forall i, j, k \quad Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \\ \forall i, j, k \quad \varepsilon_{ijk} \text{ i.i.d.}, \mathbb{E}(\varepsilon_{ijk}) = 0, \mathbb{V}(\varepsilon_{ijk}) = \sigma^2 \\ \forall i, j, k \quad cov(\varepsilon_{ijk}, \varepsilon_{i'j'k'}) = 0 \end{array} \right.$$

Visualisation de l'effet linéaire d'une variable quanti spécifique selon les modalités d'une variables quali

Effet de T9 sur le maximum d'ozone selon le vent



```
ggplot(ozone) +
  aes(x=T9, y=maxO3, col = vent, group=vent) +
  geom_smooth(method="lm", se=FALSE) +
  geom_point()+
  ggtitle("Effet de T9 sur MaxO3 selon le vent")
```

$$\text{MaxO3} \sim \text{vent} + T9 + \text{vent} : T9$$

$$\text{MaxO3}_{ij} \sim \mu + \left\{ \begin{array}{l} \alpha_1 \text{ si vent d'est} \\ \alpha_2 \text{ si vent du nord} \\ \alpha_3 \text{ si vent d'ouest} \\ \alpha_4 \text{ si vent du sud} \end{array} \right\} + \left(\beta + \left\{ \begin{array}{l} \gamma_1 \text{ si vent d'est} \\ \gamma_2 \text{ si vent du nord} \\ \gamma_3 \text{ si vent d'ouest} \\ \gamma_4 \text{ si vent du sud} \end{array} \right\} \right) \times T9_{ij} + \text{alea}_{ij}$$

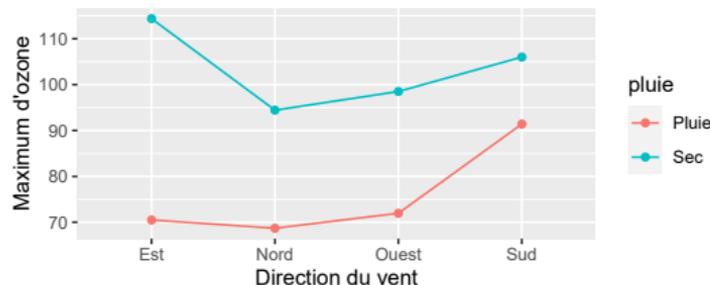
$$\left\{ \begin{array}{l} \forall i,j \quad Y_{ij} = \mu + \alpha_i + (\beta + \gamma_i) \times x_{ij} + \varepsilon_{ij} \\ \forall i,j \quad \varepsilon_{ij} \text{ i.i.d.}, \mathbb{E}(\varepsilon_{ij}) = 0, \mathbb{V}(\varepsilon_{ij}) = \sigma^2 \\ \forall i,j \quad \text{cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0 \end{array} \right.$$

Visualisation de l'interaction de 2 variables qualitatives

Définition courante : réaction réciproque de deux phénomènes l'un sur l'autre

Définition statistique : l'effet d'un facteur sur Y diffère selon les modalités de l'autre facteur

Interaction pluie:vent sur le maximum d'ozone



```
library(dplyr)
ozone |> group_by(vent, pluie) |>
  summarize(MOY = mean(maxO3)) |>
  ggplot() +
  aes(x=vent, y=MOY, col=pluie, group=pluie) +
  geom_line() + geom_point() +
  ggtitle("Interaction pluie:vent sur maxO3")+
  xlab("Direction du vent")+
  ylab("Maximum d'ozone")
```

$$MaxO3 \sim vent + pluie + vent : pluie$$

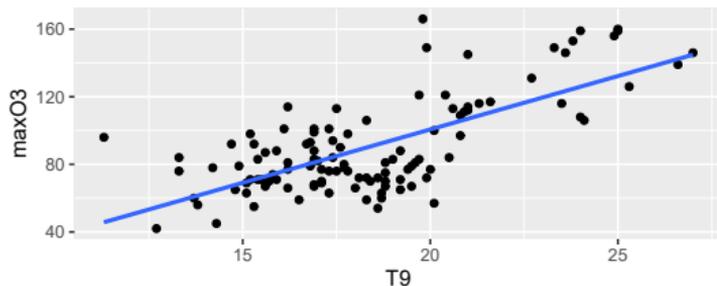
$$MaxO3_{ijk} \sim \mu + \left\{ \begin{array}{l} \alpha_1 \text{ si vent d'est} \\ \alpha_2 \text{ si vent du nord} \\ \alpha_3 \text{ si vent d'ouest} \\ \alpha_4 \text{ si vent du sud} \end{array} \right\} + \left\{ \begin{array}{l} \beta_1 \text{ si pluie} \\ \beta_2 \text{ si sec} \end{array} \right\} + \left\{ \begin{array}{l} \alpha\beta_{11} \text{ si vent d'est ET pluie} \\ \alpha\beta_{12} \text{ si vent d'est ET sec} \\ \alpha\beta_{21} \text{ si vent du nord ET pluie} \\ \dots \alpha\beta_{42} \text{ si vent du sud ET sec} \end{array} \right\} + alea_{ijk}$$

$$\left\{ \begin{array}{l} \forall i, j, k \quad Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk} \\ \forall i, j, k \quad \varepsilon_{ijk} \text{ i.i.d.}, \mathbb{E}(\varepsilon_{ijk}) = 0, \mathbb{V}(\varepsilon_{ijk}) = \sigma^2 \\ \forall i, j, k \quad cov(\varepsilon_{ijk}, \varepsilon_{i'j'k'}) = 0 \end{array} \right.$$

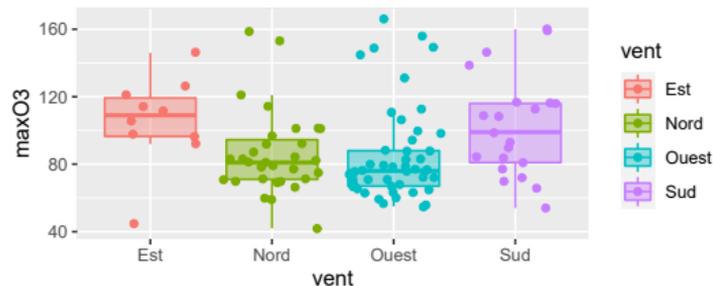
Les effets d'un modèle

Quatre types d'effets possibles :

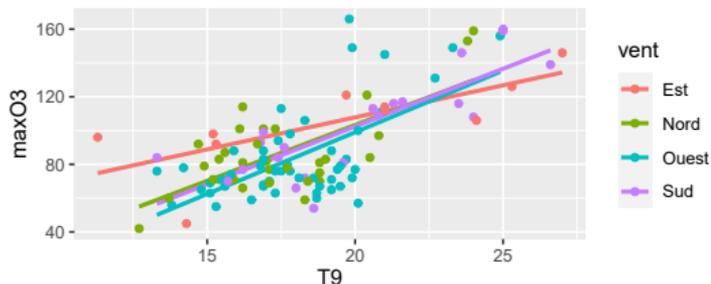
Effet de T9 sur le maximum d'ozone



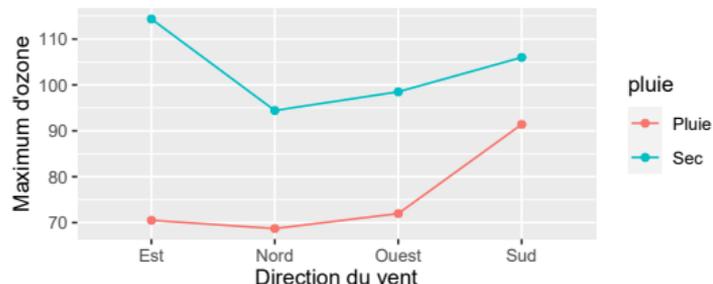
Effet du vent sur le maximum d'ozone



Effet de T9 sur le maximum d'ozone selon le vent



Interaction pluie:vent sur le maximum d'ozone



Et avec ça on en ajoute autant qu'on veut pour construire ... tous les modèles avec des effets linéaires et des interactions.

Les modèles linéaires

| Réponse | Variable(s) explicative(s) | Méthode |
|-------------------|------------------------------------|---|
| Var. quantitative | 1 var. quantitative | régression linéaire simple |
| Var. quantitative | 1 var. qualitative à I modalités | analyse de variance à 1 facteur (rq : si $I = 2$ équivaut à comparaison de 2 moyennes) |
| Var. quantitative | p var. quantitatives | régression linéaire multiple |
| Var. quantitative | K var. qualitatives | analyse de variance à K facteurs |
| Var. quantitative | var. quantitatives et qualitatives | analyse de covariance |
| Var. qualitative | var. quantitatives et qualitatives | régression logistique |

Écriture du modèle

Modèle de régression multiple (toutes les variables explicatives sont quantitatives) :

$$\begin{cases} \forall i = 1, \dots, n & Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots + \beta_p x_{ip} + \varepsilon_i \\ \forall i = 1, \dots, n & \varepsilon_i \text{ i.i.d.}, \mathbb{E}(\varepsilon_i) = 0, \mathbb{V}(\varepsilon_i) = \sigma^2 \\ \forall i \neq k & \text{cov}(\varepsilon_i, \varepsilon_k) = 0 \end{cases}$$

(p+1) paramètres à estimer + 1 paramètre de variance σ^2

Matriciellement : $Y = X\beta + E$ avec $\mathbb{E}(E) = 0$, $\mathbb{V}(E) = \sigma^2 Id$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \beta_0 & \beta_1 & \beta_2 & \dots & \beta_p \\ 1 & x_{11} & x_{12} & \dots & x_{1p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{i1} & x_{i2} & & x_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Remarque : Les modèles d'analyse de variance et d'analyse de covariance peuvent aussi s'écrire sous cette forme !!

$$\begin{cases} \forall i, j, k & Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk} \\ \forall i, j, k & \varepsilon_{ijk} \text{ i.i.d.}, \mathbb{E}(\varepsilon_{ijk}) = 0, \mathbb{V}(\varepsilon_{ijk}) = \sigma^2 \\ \forall i, j, k & \text{cov}(\varepsilon_{ijk}, \varepsilon_{i'j'k'}) = 0 \end{cases} \quad \begin{cases} \forall i, j & Y_{ij} = \mu + \alpha_i + (\beta + \gamma_i) \times x_{ij} + \varepsilon_{ij} \\ \forall i, j & \varepsilon_{ij} \text{ i.i.d.}, \mathbb{E}(\varepsilon_{ij}) = 0, \mathbb{V}(\varepsilon_{ij}) = \sigma^2 \\ \forall i, j & \text{cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0 \end{cases}$$

Estimation des paramètres du modèle

Critère des moindres carrés : estimer les paramètres en minimisant la somme des carrés des écarts entre observations et prévisions par le modèle

$$Y \approx X\beta$$

$$X'Y \approx X'X\beta$$

$$\hat{\beta} = (X'X)^{-1}X'Y \quad \text{si } X'X \text{ est inversible}$$

Propriétés : $\mathbb{E}(\hat{\beta}) = \beta$; $\mathbb{V}(\hat{\beta}) = (X'X)^{-1}\sigma^2$

La variance des résidus σ^2 est estimée par :

$$\hat{\sigma}^2 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{\text{nb données} - \text{nb paramètres estimés à partir des données}}$$

$$\mathbb{E}(\hat{\sigma}^2) = \sigma^2$$

Décomposition de la variabilité

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Variabilité totale = Variabilité modèle + Variabilité résiduelle

Pourcentage de variabilité de Y expliquée par le modèle : $R^2 = \frac{SC_{modele}}{SC_{total}}$

Propriétés : $0 \leq R^2 \leq 1$

La variabilité du modèle peut être décomposée par variable de 2 façons :

- en calculant la variabilité expliquée par chaque variable les unes après les autres (pb : la variabilité d'une variable dépend de l'ordre d'introduction des variables)
- en calculant la variabilité expliquée exclusivement par une variable (pb : la somme des variabilités de toutes les variables n'est pas égale à la variabilité du modèle)

Dans certains cas (données équilibrées), la variabilité du modèle se décompose parfaitement et ces 2 calculs donnent les mêmes résultats

Exemple sur l'ozone

```
library(FactoMineR)
LinearModel(maxO3 ~ T0+T12+T15+Ne9+Ne12+Ne15+Vx9+Vx12+Vx15+maxO3v+pluie+vent, data=ozone)
# LinearModel(maxO3 ~ ., data = ozone)    ## écriture simplifiée

Residual standard error: 14.51 on 97 degrees of freedom
Multiple R-squared:  0.7686
F-statistic: 23.01 on 14 and 97 DF,  p-value: 8.744e-25
AIC = 613    BIC = 653.8
```

Ftest

| | SS | df | MS | F value | Pr(>F) |
|-----------|---------|----|--------|---------|-----------|
| T9 | 0.2 | 1 | 0.2 | 0.0011 | 0.97325 |
| T12 | 376.0 | 1 | 376.0 | 1.7868 | 0.18445 |
| T15 | 30.3 | 1 | 30.3 | 0.1439 | 0.70526 |
| Ne9 | 1016.5 | 1 | 1016.5 | 4.8312 | 0.03033 |
| Ne12 | 37.9 | 1 | 37.9 | 0.1803 | 0.67208 |
| Ne15 | 0.1 | 1 | 0.1 | 0.0003 | 0.98680 |
| Vx9 | 50.2 | 1 | 50.2 | 0.2388 | 0.62619 |
| Vx12 | 35.7 | 1 | 35.7 | 0.1697 | 0.68127 |
| Vx15 | 122.6 | 1 | 122.6 | 0.5826 | 0.44715 |
| maxO3v | 5560.4 | 1 | 5560.4 | 26.4261 | 1.421e-06 |
| vent | 297.8 | 3 | 99.3 | 0.4718 | 0.70267 |
| pluie | 182.9 | 1 | 182.9 | 0.8694 | 0.35344 |
| Residuals | 20410.2 | 97 | 210.4 | | |

Test de l'effet d'une ou plusieurs variables

Question : l'ensemble de variables \mathcal{V} apporte-t-il des informations complémentaires intéressantes sachant que les autres variables sont déjà dans le modèle ?

Hypothèses : H_0 : "tous les coefficients associés aux variables de \mathcal{V} sont égaux à 0" contre H_1 : "au moins un coefficient des variables \mathcal{V} est différent de 0"

Statistique de test : $F_{obs} = \frac{SC_{\mathcal{V}}/ddl_{\mathcal{V}}}{SC_R/ddl_R} = \frac{CM_{\mathcal{V}}}{CM_R}$

Loi de la statistique de test : Sous H_0 , $\mathcal{L}(F_{obs}) = \mathcal{F}_{ddl_{\mathcal{V}}}^{ddl_R}$

Décision : $\mathbb{P}(\mathcal{F}_{ddl_{\mathcal{V}}}^{ddl_R} > F_{obs}) < 0.05 \implies$ Rejet de H_0

- Revient à choisir entre le sous-modèle sans les variables \mathcal{V} ou le modèle complet
- On teste le plus souvent \mathcal{V} avec 1 variable ou avec toutes les variables
- Si \mathcal{V} contient tous les effets : revient à tester si R^2 est significativement différent de 0, i.e. si toutes les variables sont inutiles (versus au moins une utile)
- On somme les degrés de liberté associés à l'ensemble \mathcal{V} sachant qu'1 variable quanti à 1 ddl, 1 variable quali à $I - 1$ ddl et une interaction a comme ddl le produit des ddl de chaque facteur

\implies Pour la séance de TD écrire le test pour 1 variable quali, celui pour 1 interaction, celui pour le test de toutes les variables

Sélection de variables

Comment sélectionner un « bon » sous-modèle ?

- sélectionner le modèle pour lequel la probabilité critique du test du R^2 est la plus petite (rejet de l'hypothèse : le modèle n'est pas intéressant)
- sélectionner le modèle qui minimise le critère AIC pour mieux prédire (ou BIC pour mieux sélectionner les variables) : ces critères sont un compromis entre un modèle qui maximise la vraisemblance (i.e. qui s'ajuste bien aux données), et qui n'a pas trop de paramètres (pénalité augmente avec le nombre de variables retenues)

Plusieurs stratégies :

- Construction exhaustive de tous les sous-modèles (long et même impossible si trop de variables)
- Méthode descendante (backward) : construire le modèle complet ; supprimer la variable explicative la moins intéressante et reconstruire le modèle sans cette variable ; itérer jusqu'à ce que toutes les variables explicatives soient intéressantes
- Méthode ascendante (forward) : partir du modèle avec la variable la plus intéressante ; ajouter la variable qui, connaissant les autres variables du modèle, apporte le plus d'information complémentaire ; itérer jusqu'à ce qu'aucune variable n'apporte d'information intéressante
- Méthode stepwise : compromis entre les 2 méthodes ci-dessus

Exemple sur l'ozone : sélection de variables

```
library(FactoMineR)
LinearModel(maxO3~., data=ozone, selection="bic")
```

Results for the complete model:

=====

Call:

```
LinearModel(formula = maxO3 ~ ., data = ozone, selection = "bic")
```

Residual standard error: 14.51 on 97 degrees of freedom

Multiple R-squared: 0.7686

F-statistic: 23.01 on 14 and 97 DF, p-value: 8.744e-25

AIC = 613 BIC = 653.8

Results for the model selected by BIC criterion:

=====

Call:

```
LinearModel(formula = maxO3 ~ T12 + Ne9 + Vx9 + maxO3v, data = ozone,
            selection = "BIC")
```

Residual standard error: 14 on 107 degrees of freedom

Multiple R-squared: 0.7622

F-statistic: 85.75 on 4 and 107 DF, p-value: 1.763e-32

AIC = 596 BIC = 609.6

Exemple sur l'ozone : sélection de variables (suite)

Ftest

| | SS | df | MS | F value | Pr(>F) |
|-----------|----------|-----|---------|---------|-----------|
| T12 | 6650.39 | 1 | 6650.39 | 33.9334 | 6.073e-08 |
| Ne9 | 2714.81 | 1 | 2714.81 | 13.8522 | 0.0003172 |
| Vx9 | 903.37 | 1 | 903.37 | 4.6094 | 0.0340547 |
| maxO3v | 7363.50 | 1 | 7363.50 | 37.5721 | 1.499e-08 |
| Residuals | 20970.24 | 107 | 195.98 | NA | NA |

Ttest

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|-----------|------------|---------|----------|
| (Intercept) | 12.631310 | 11.000877 | 1.1482 | 0.25344 |
| T12 | 2.764090 | 0.474502 | 5.8252 | < 2e-16 |
| Ne9 | -2.515402 | 0.675845 | -3.7219 | 0.00032 |
| Vx9 | 1.292857 | 0.602180 | 2.1470 | 0.03405 |
| maxO3v | 0.354832 | 0.057888 | 6.1296 | < 2e-16 |

Démarche en modélisation

- 1 Lister les variables qui entrent en jeu pour expliquer ou prédire la variable réponse
- 2 Visualiser les données et notamment les liaisons avec la variable réponse
- 3 Ecrire puis construire le modèle en choisissant effets et interactions qui expliquent potentiellement la réponse (ne pas nécessairement tout mettre dans le modèle)
- 4 Sélectionner le sous-modèle qui minimise l'AIC (ou le BIC ou à la main) en supprimant les interactions et effets non utiles
- 5 Construire ce sous-modèle
- 6 Interpréter les résultats (les effets significatifs comme ceux non significatifs)
- 7 Interpréter les coefficients du modèle (attention aux confusions possibles notamment pour les variables quantitatives)
- 8 Prédire pour de nouvelles valeurs si vous avez un objectif de prédiction

Démarche statistique

- 1 Intro
- 2 Visualisation
- 3 Tests
- 4 Analyse de variance à 1 facteur
- 5 Régression linéaire simple
- 6 Construction et sélection de modèles
- 7 **Interprétation et prédiction**
 - Interprétation
 - Prédiction
 - Exemple complet

Retour sur le codage et les contraintes pour variables qualitatives

La matrice X (dans $Y = X\beta + E$) a autant de lignes que d'individus et pour colonnes :

- une colonne de 1 est utilisée pour les constantes comme μ ou $\beta_0 \implies 1$ ddl
- chaque variable quantitative correspond à une colonne $\implies 1$ ddl
- chaque variable qualitative est transformée en autant d'indicatrices qu'elle a de modalités
Il suffit d'avoir $I - 1$ indicatrices pour estimer tous les paramètres \implies on pose une contrainte, le mieux est de prendre $\sum_{i=1}^I \alpha_i = 0 \implies (I - 1)$ ddl
- chaque interaction est codée par autant d'indicatrices qu'il y a de paires de modalités
On a alors besoin de seulement $(I - 1)(J - 1)$ indicatrices, d'où des contraintes $\forall i, \sum_j \alpha\beta_{ij} = 0$ et $\forall j, \sum_i \alpha\beta_{ij} = 0 \implies (I - 1)(J - 1)$ ddl

Remarque : le choix de la contrainte impacte **FORTEMENT** l'interprétation

- avec $\sum_i \alpha_i = \mathbf{0}$, la comparaison se fait par rapport à la moyenne des moyennes par modalité (un coefficient égal à 0 signifie que les résultats de la modalité ne sont pas différents de l'effet moyen)
- avec $\alpha_1 = 0$ (par défaut pour certaines fonctions de R, par ex. `lm`), la comparaison se fait par rapport au niveau 1 qui sert de référence \implies contrainte à proscrire quand il y a des interactions car l'interprétation devient très compliquée

Interprétation des résultats

Une fois le modèle sélectionné, on interprète l'absence ou la présence des différents effets (variable quantitative, variable qualitative, interactions)

Deux situations bien distinctes :

- Le cas d'interprétation idéal quand les données sont équilibrées
- La difficile interprétation des résultats pour des données déséquilibrées

Remarques :

- Le choix des données avec des plans d'expériences permet d'avoir des données équilibrées, et donc des résultats facilement interprétables
- Souvent en analyse de variance les données sont équilibrées (ou peu déséquilibrées)
- En régression ou analyse de covariance les données sont déséquilibrées dès qu'il y a une corrélation non nulle entre les variables explicatives (ce qui est quasi systématique sans plan d'expériences)

Décomposition de la variabilité : variables qualitatives, données équilibrées

Quand les données sont équilibrées, la décomposition de la variabilité est parfaite :

$$\begin{aligned} \sum_{i,j,k} (y_{ijk} - y_{\bullet\bullet\bullet})^2 &= \sum_{i,j,k} \underbrace{(y_{i\bullet\bullet} - y_{\bullet\bullet\bullet})^2}_{\hat{\alpha}_i^2} + \sum_{i,j,k} \underbrace{(y_{\bullet j\bullet} - y_{\bullet\bullet\bullet})^2}_{\hat{\beta}_j^2} \\ &+ \sum_{i,j,k} \underbrace{(y_{ij\bullet} - y_{i\bullet\bullet} - y_{\bullet j\bullet} + y_{\bullet\bullet\bullet})^2}_{\widehat{\alpha\beta}_{ij}^2} + \sum_{i,j,k} \underbrace{(y_{ijk} - y_{ij\bullet})^2}_{\varepsilon_{ijk}^2} \\ \sum_{i,j,k} (y_{ijk} - y_{\bullet\bullet\bullet})^2 &= \sum_{i,j,k} (y_{i\bullet\bullet} - y_{\bullet\bullet\bullet})^2 + \sum_{i,j,k} (y_{\bullet j\bullet} - y_{\bullet\bullet\bullet})^2 \\ &+ \sum_{i,j,k} (y_{ij\bullet} - y_{i\bullet\bullet} - y_{\bullet j\bullet} + y_{\bullet\bullet\bullet})^2 + \sum_{i,j,k} (y_{ijk} - y_{ij\bullet})^2 \end{aligned}$$

Quand les données sont équilibrées, les coefficients s'estiment simplement :

$$\begin{aligned} \hat{\mu} &= y_{\bullet\bullet\bullet} & \forall i, \hat{\alpha}_i &= y_{i\bullet\bullet} - y_{\bullet\bullet\bullet} \\ \forall j, \hat{\beta}_j &= y_{\bullet j\bullet} - y_{\bullet\bullet\bullet} & \forall i, j, \widehat{\alpha\beta}_{ij} &= y_{ij\bullet} - y_{i\bullet\bullet} - y_{\bullet j\bullet} + y_{\bullet\bullet\bullet} \end{aligned}$$

⇒ On quantifie parfaitement ce qui est expliqué par chaque variable ou interaction

Décomposition de la variabilité : données déséquilibrées

Quand les données sont déséquilibrées, impossible de distinguer quel effet ou variable explique la variabilité. On parle de confusion (alias en anglais)

En sommant les variabilités de toutes les variables et de la résiduelle, on ne retrouve pas la variabilité totale (la somme est plus petite que la variabilité totale)

Quand les variables explicatives sont très corrélées, l'interprétation est très difficile \implies la sélection de variables évite les mauvaises interprétations (signe du coefficient de régression différent de celui de la corrélation), mais derrière l'effet d'une variable explicative, il peut y avoir celui d'autres variables non sélectionnées dans le modèle (non sélectionnée car elle n'apporte pas d'information supplémentaire significative par rapport aux autres variables ... même si elle influe sur la variable réponse)

Test de l'effet d'une variable qualitative – d'une interaction

Test de l'effet d'une variable qualitative (on l'a déjà vu!!) :

Question : Y a-t'il un effet du facteur A ? Est-ce que pour au moins une modalité les individus prennent des valeurs significativement différentes ?

Hypothèses : $H_0 : \forall i \alpha_i = 0$ contre $H_1 : \exists i / \alpha_i \neq 0$

Statistique de test : $F_{obs} = \frac{SC_A/ddl_A}{SC_R/ddl_R} = \frac{CM_A}{CM_R}$

Loi de la statistique de test : Sous H_0 , $\mathcal{L}(F_{obs}) = \mathcal{F}_{ddl_R}^{ddl_A}$

Décision : $\mathbb{P}(\mathcal{F}_{ddl_R}^{ddl_A} > F_{obs}) < 0.05 \implies$ Rejet de H_0

Test de l'effet d'une interaction : (c'est exactement pareil!!)

Question : Y a-t-il un effet de l'interaction ? Y a-t'il un effet conjoint (une interaction) des facteurs A et B sur Y ? Est-ce que pour une combinaison d'une modalité de A avec une modalité de B on a des valeurs de Y significativement plus élevées ou plus faibles qu'attendu avec un modèle additif ?

Hypothèses : $H_0 : \forall i, j \alpha\beta_{ij} = 0$ contre $H_1 : \exists(i, j) / \alpha\beta_{ij} \neq 0$

Statistique de test : $F_{obs} = \frac{SC_{interaction}/ddl_{interaction}}{SC_R/ddl_R} = \frac{CM_{interaction}}{CM_R}$

Loi de la statistique de test : Sous H_0 , $\mathcal{L}(F_{obs}) = \mathcal{F}_{ddl_R}^{ddl_{interaction}}$

Décision : $\mathbb{P}(\mathcal{F}_{ddl_R}^{ddl_{interaction}} > F_{obs}) < 0.05 \implies$ Rejet de H_0

Inférence : test d'un coefficient

Question : ce coefficient α_1 (par exemple) est-il différent de 0 ? La modalité 1 donne-t-elle des résultats significativement différents de la moyenne ?

Hypothèses : $H_0 : \alpha_1 = 0$ contre $H_1 : \alpha_1 \neq 0$

Statistique de test : $T_{obs} = \frac{\hat{\alpha}_1}{\hat{\sigma}_{\hat{\alpha}_1}}$

Loi de la statistique de test sous H_0 : $\mathcal{L}(T_{obs}) = \mathcal{T}_{ddl_R}$ (loi de Student)

Décision : $\mathbb{P}(\mathcal{T}_{ddl_R} > |T_{obs}|) < 0.05 \implies \text{Rejet de } H_0$

Remarque : ce test en régression revient à tester si 1 variable quantitative a un effet significatif

Comparaison des moyennes ajustées

Travailler sur les moyennes ajustées ($\hat{\mu} + \hat{\alpha}_i$) permet de s'affranchir de l'effet des autres variables, i.e. de neutraliser l'effet des autres variables

On peut comparer les modalités 2 à 2 grâce à un test de comparaison par paire, avec une correction de Bonferroni par exemple

```
mod <- LinearModel(sucre~choc+juge+choc:juge, data=chocolats)
meansComp(mod,~choc, adjust="bonferonni")
```

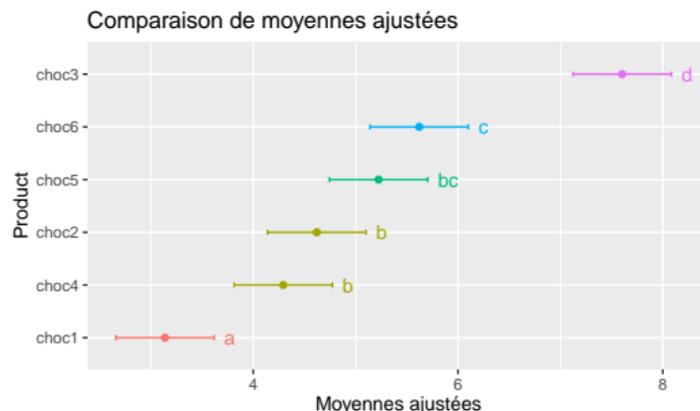
```
$adjMean
```

| choc | emmean | SE | df | lower.CL | upper.CL |
|-------|--------|-------|-----|----------|----------|
| choc1 | 3.14 | 0.243 | 174 | 2.66 | 3.62 |
| choc2 | 4.62 | 0.243 | 174 | 4.14 | 5.10 |
| choc3 | 7.60 | 0.243 | 174 | 7.12 | 8.08 |
| choc4 | 4.29 | 0.243 | 174 | 3.81 | 4.77 |
| choc5 | 5.22 | 0.243 | 174 | 4.74 | 5.70 |
| choc6 | 5.62 | 0.243 | 174 | 5.14 | 6.10 |

Results are averaged over the levels of: juge
Confidence level used: 0.95

```
$groupComp
```

| choc1 | choc4 | choc2 | choc5 | choc6 | choc3 |
|-------|-------|-------|-------|-------|-------|
| "a" | "b" | "b" | "bc" | "c" | "d" |



Prévisions

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip}$$

Prédire Y pour : (T12=19, Ne9=8, Vx9=1.2, maxO3v=70) et (T12=23, Ne9=10, Vx9=0.9, maxO3v=95)

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|-----------|------------|---------|----------|
| (Intercept) | 12.631310 | 11.000877 | 1.1482 | 0.25344 |
| T12 | 2.764090 | 0.474502 | 5.8252 | < 2e-16 |
| Ne9 | -2.515402 | 0.675845 | -3.7219 | 0.00032 |
| Vx9 | 1.292857 | 0.602180 | 2.1470 | 0.03405 |
| maxO3v | 0.354832 | 0.057888 | 6.1296 | < 2e-16 |

Calcul à la main de la prédiction 1 = 12.63 + 2.764*19 -2.515*8 +1.292*1.2 + 0.354*70 = 71.41544

Sur ordinateur :

```
xnew <- data.frame(T12=c(19,23), Ne9=c(8,10), Vx9=c(1.2,0.9), maxO3v=c(70,95))
predict(model,xnew,interval="pred")
```

```
      fit      lwr      upr
1 71.41544 42.90026 99.93063
2 85.92393 56.76803 115.07983
```

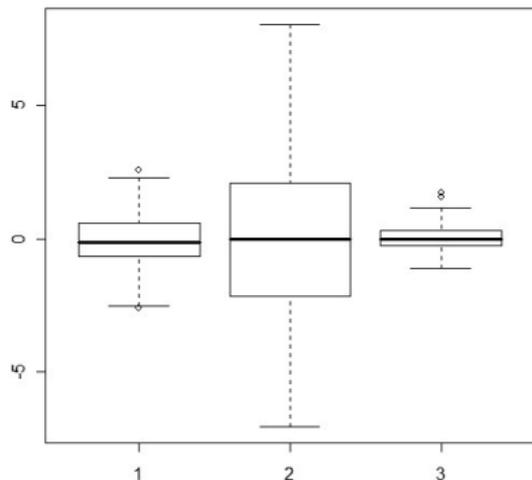
```
predict(model,xnew,interval="confidence")
```

```
      fit      lwr      upr
1 71.41544 64.86327 77.96761
2 85.92393 76.98627 94.86159
```

Analyse des résidus du modèle

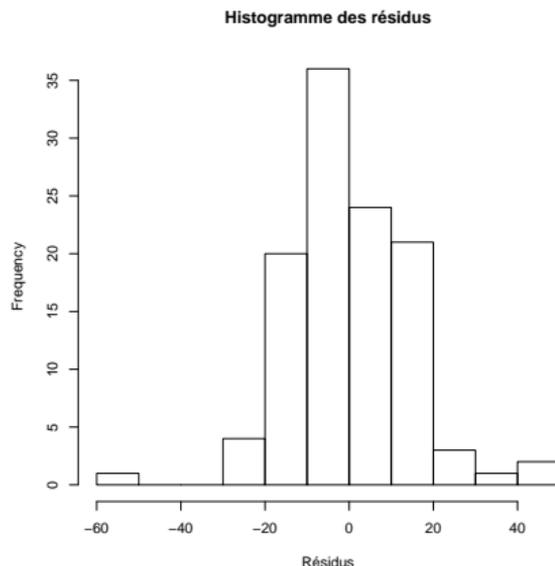
Test de Bartlett de l'homoscédasticité des **Résidus**

```
model <- lm(maxO3~T12+...,data=ozone)
res <- residuals(model)
boxplot(res~vent,data=ozone)
bartlett.test(res~vent,data=ozone)
```



Test de Shapiro-Wilk de normalité des **Résidus**
(Rq : la non-normalité n'est pas un pb tant que la distribution est symétrique)

```
model <- lm(maxO3~T12+...,data=ozone)
res <- residuals(model)
hist(res,main="Histogramme résidus",xlab="Résidus")
shapiro.test(res)
```



Retour sur l'exemple ozone

- On s'intéresse au maximum d'ozone : variable réponse
- Les variables de températures, nébulosité, vitesse de vent (quanti), et la direction et la pluie (quali) sont prises en compte
- On ne sait pas si les interactions entre variables quali et entre les variables quali et quanti sont négligeables \implies on les met dans le modèle
- On écrit le modèle :

```
LinearModel(maxO3 ~ (T9 + T12 + T15 + Ne9 + Ne12 + Ne15 + Vx9 + Vx12 + Vx15 +
maxO3v + pluie + vent) * (pluie + vent), data=ozone, selection="bic")
```

Ce qui revient à écrire :

```
maxO3 ~ T9 + T12 + T15 + Ne9 + Ne12 + Ne15 + Vx9 + Vx12 + Vx15 + maxO3v + pluie + vent +
T9:pluie + T12:pluie + T15:pluie + Ne9:pluie + Ne12:pluie + Ne15:pluie +
Vx9:pluie + Vx12:pluie + Vx15:pluie + maxO3v:pluie + vent:pluie +
T9:vent + T12:vent + T15:vent + Ne9:vent + Ne12:vent + Ne15:vent +
Vx9:vent + Vx12:vent + Vx15:vent + maxO3v:vent
```

On sélectionne le sous-modèle

Results for the complete model:

=====

Call:

```
LinearModel(formula = max03 ~ (T9 + T12 + T15 + Ne9 + Ne12 + Ne15 +  
  Vx9 + Vx12 + Vx15 + max03v + pluie + vent) * (pluie + vent),  
  data = ozone, selection = "bic")
```

Residual standard error: 14.54 on 56 degrees of freedom

Multiple R-squared: 0.8658

F-statistic: 6.569 on 55 and 56 DF, p-value: 2.585e-11

AIC = 634 BIC = 786.2

Results for the model selected by BIC criterion:

=====

Call:

```
LinearModel(formula = max03 ~ T9 + T15 + Ne12 + Vx9 + max03v + vent +  
  T9:vent + T15:vent, data = ozone, selection = "bic")
```

Residual standard error: 13.8 on 97 degrees of freedom

Multiple R-squared: 0.7906

F-statistic: 26.16 on 14 and 97 DF, p-value: 8.082e-27

AIC = 601.8 BIC = 642.6

Et maintenant le plus intéressant : l'interprétation des résultats

On donne quelques extraits d'interprétation !

Ftest

| | SS | df | MS | F value | Pr(>F) |
|-----------|---------|----|--------|---------|-----------|
| T9 | 698.9 | 1 | 698.9 | 3.6710 | 0.0583123 |
| T15 | 557.5 | 1 | 557.5 | 2.9279 | 0.0902552 |
| Ne12 | 2107.0 | 1 | 2107.0 | 11.0664 | 0.0012425 |
| Vx9 | 1657.3 | 1 | 1657.3 | 8.7046 | 0.0039790 |
| maxO3v | 5160.8 | 1 | 5160.8 | 27.1060 | 1.078e-06 |
| vent | 182.9 | 3 | 61.0 | 0.3202 | 0.8107707 |
| T9:vent | 2587.1 | 3 | 862.4 | 4.5294 | 0.0051335 |
| T15:vent | 3722.1 | 3 | 1240.7 | 6.5165 | 0.0004613 |
| Residuals | 18468.3 | 97 | 190.4 | | |

On peut dire (à partir des effets significatifs) qu'il y a, sur le max d'O3 :

- des effets de nébulosité, de vitesse de vent, du maximum d'O3 de la veille
- des effets de la direction du vent et des températures mais à travers les interactions : la direction du vent modifie l'effet de la T^o (i.e. amplifie l'effet de la T^o ou la diminue selon la direction du vent) sur max O3

On peut aussi dire (à partir des absences d'effets significatifs) :

- la pluviométrie n'est pas un facteur déterminant qui influe sur le max d'O3
- une seule nébulosité (Ne12) est conservée dans le modèle : cela ne veut pas dire que les autres nébulosités n'ont pas d'effet (elles peuvent avoir un effet similaire). Idem pour la vitesse de vent.
- pour la T^o , on a besoin des T^o à 9h et à 15h pour mieux prévoir le maximum d'ozone. L'effet de la T^o n'est pas exactement le même entre 9h et 15h (mais 12h n'est pas utile comme info)
- l'effet de la nébulosité est le même quelle que soit la direction du vent. Idem pour la vitesse du vent.
- etc.

Et maintenant le plus intéressant : l'interprétation des résultats

Et on affine les interprétations

Ttest

| | Estimate | Std. Error | t value | Pr(> t) |
|--------------------|-----------|------------|---------|-----------|
| (Intercept) | 18.206308 | 13.705946 | 1.3284 | 0.187179 |
| T9 | 1.826916 | 1.023832 | 1.7844 | 0.077487 |
| T15 | 1.020017 | 0.885187 | 1.1523 | 0.252022 |
| Ne12 | -3.013169 | 0.905775 | -3.3266 | 0.001243 |
| Vx9 | 2.221261 | 0.752878 | 2.9504 | 0.003979 |
| maxO3v | 0.346696 | 0.066591 | 5.2063 | 1.078e-06 |
| vent - Est | 0.473424 | 17.010219 | 0.0278 | 0.977854 |
| vent - Nord | 3.297358 | 15.035620 | 0.2193 | 0.826875 |
| vent - Ouest | -1.125022 | 14.006616 | -0.0803 | 0.936148 |
| vent - Sud | -2.645761 | 15.148862 | -0.1747 | 0.861718 |
| vent - Est : T9 | -4.744227 | 2.141774 | -2.2151 | 0.029094 |
| vent - Nord : T9 | 6.140768 | 1.701384 | 3.6093 | 0.000488 |
| vent - Ouest : T9 | -0.950539 | 1.345534 | -0.7064 | 0.481608 |
| vent - Sud : T9 | -0.446002 | 1.522150 | -0.2930 | 0.770142 |
| vent - Est : T15 | 3.680964 | 1.796744 | 2.0487 | 0.043194 |
| vent - Nord : T15 | -5.227943 | 1.215336 | -4.3016 | 4.045e-05 |
| vent - Ouest : T15 | 0.979689 | 0.930826 | 1.0525 | 0.295188 |
| vent - Sud : T15 | 0.567290 | 1.111141 | 0.5105 | 0.610828 |

- pour T9 et T15, coefficients non significativement différents de 0
- plus il y a de nébulosité, moins le max d'O3 est grand (**on vérifie que le signe est le même que celui de la corrélation**)
- plus le maximum d'O3 de la veille est grand, plus le maximum d'O3 est grand
- les jours de vents d'est et surtout du nord ont des max d'O3 supérieur aux jours de vents d'ouest et du sud
- les jours de vent du nord, l'effet de la T° à 9h est plus important (coef = 6.14); au contraire quand le vent vient de l'est. C'est l'inverse pour la T° à 15h
- Ainsi, quand le vent vient du nord, l'effet de la T° à 9h est particulièrement fort (donc s'il fait chaud à 9h quand le vent vient du nord, le max d'O3 risque d'être important)
- etc.

Pour conclure : démarche en modélisation

- 1 Lister les variables qui entrent en jeu pour expliquer ou prédire la variable réponse
- 2 Visualiser les données et notamment les liaisons avec la variable réponse
- 3 Ecrire puis construire le modèle en choisissant effets et interactions qui expliquent potentiellement la réponse (ne pas nécessairement tout mettre dans le modèle)
- 4 Sélectionner le sous-modèle qui minimise le BIC en supprimant les interactions et effets non utiles
- 5 Construire ce sous-modèle
- 6 Interpréter les résultats (les effets significatifs comme ceux non significatifs)
- 7 Interpréter les coefficients du modèle (attention aux confusions possibles notamment pour les variables quantitatives)
- 8 Prédire pour de nouvelles valeurs si vous avez un objectif de prédiction

Et évidemment, cela peut suggérer de nouvelles analyses !! (synthétiser des variables, rendre qualitatives certaines variables, ajouter d'autres variables, etc.)