Modélisation

Partie 1 : construction et sélection de modèles

François Husson

Unité pédagogique de mathématiques appliquées
L'Institut Agro

Modéliser c'est comprendre et prévoir

Qui a déjà fait des modélisations complexes ?

Modéliser c'est comprendre et prévoir

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

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)

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 ?

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 influer sur votre réponse
2. vous éliminez les variables qui sont négligeables
3. vous essayez de quantifier l'effet des variables restantes sélectionnées

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 influer sur votre réponse
2. vous éliminez les variables qui sont négligeables
3. vous essayez de quantifier l'effet des variables restantes sélectionnées

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 influer 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 ?

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 influer 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

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 influer 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 ?

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 influer 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

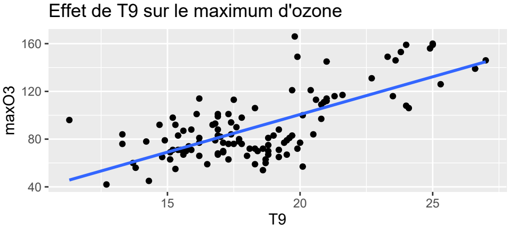
L'association Air Breizh surveille la qualité de l'air et mesure la concentration de polluants comme l'ozone (O_3) ainsi que les conditions météorologiques comme la température, la nébulosité, le vent, etc.
Durant l'été 2001, 112 données ont été relevées à Rennes

```
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
...	...	...											

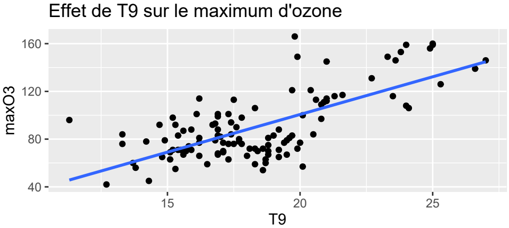
Leur objectif : prévoir la concentration en ozone du lendemain pour avertir la population en cas de pic de pollution

Visualisation de l'effet linéaire d'une variable quantitative




```r
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")
```

Visualisation de l'effet linéaire d'une variable quantitative

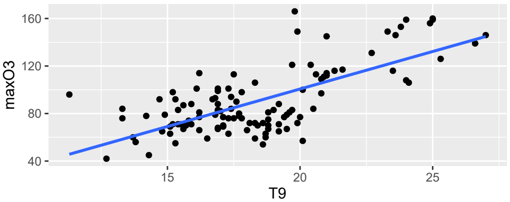



```
library(ggplot2)
ggplot(ozone) +
  aes(x=T9, y=maxO3) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  ggtitle("Effet de T9 sur le maximum d'ozone")
```

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

Visualisation de l'effet linéaire d'une variable quantitative

Effet de T9 sur le maximum d'ozone



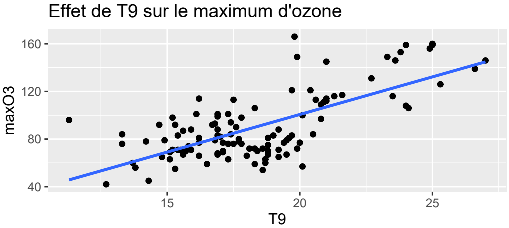

```r
library(ggplot2)
ggplot(ozone) +
  aes(x=T9, y=maxO3) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  ggtitle("Effet de T9 sur le maximum d'ozone")
```

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

$$\text{Réponse} \ \sim \ var_1$$

$$\begin{cases} \forall i = 1, \ldots, n \quad Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \\ \forall i = 1, \ldots, n \quad \varepsilon_i \ \text{ i.i.d. } \ , \ \mathbb{E}(\varepsilon_i) = 0, \ \ \mathbb{V}(\varepsilon_i) = \sigma^2 \\ \forall i \neq k \quad cov(\varepsilon_i, \varepsilon_k) = 0 \end{cases}$$

Visualisation de l'effet linéaire d'une variable quantitative



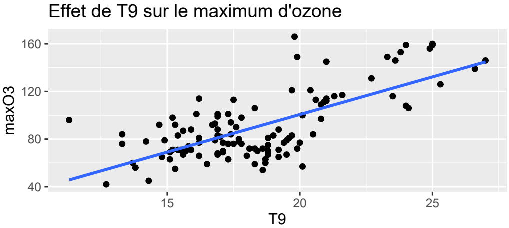

```r
library(ggplot2)
ggplot(ozone) +
  aes(x=T9, y=maxO3) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  ggtitle("Effet de T9 sur le maximum d'ozone")
```

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

$$\text{Réponse} \;\sim\; var_1 + var_2 + \ldots + var_p$$

$$\begin{cases} \forall i = 1, \ldots, n \quad Y_i = \beta_0 + \beta_1 x_i + \beta_2 v_i + \beta_3 w_i + \ldots + \beta_p z_i + \varepsilon_i \\ \forall i = 1, \ldots, n \quad \varepsilon_i \;\text{ i.i.d. }, \;\; \mathbb{E}(\varepsilon_i) = 0, \;\; \mathbb{V}(\varepsilon_i) = \sigma^2 \\ \forall i \neq k \quad cov(\varepsilon_i, \varepsilon_k) = 0 \end{cases}$$

Visualisation de l'effet linéaire d'une variable quantitative



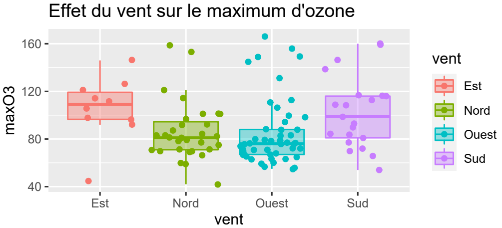

```
library(ggplot2)
ggplot(ozone) +
  aes(x=T9, y=maxO3) +
  geom_point() +
  geom_smooth(method="lm", se=FALSE) +
  ggtitle("Effet de T9 sur le maximum d'ozone")
```

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

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

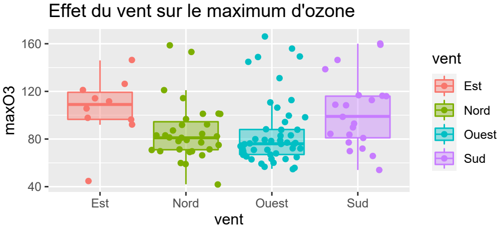
$$\begin{cases} \forall i = 1, \ldots, n & Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_p x_{ip} + \varepsilon_i \\ \forall i = 1, \ldots, n & \varepsilon_i \ \text{ i.i.d.} \ , \quad \mathbb{E}(\varepsilon_i) = 0, \quad \mathbb{V}(\varepsilon_i) = \sigma^2 \\ \forall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

Visualisation de l’effet d’une variable qualitative




```
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")
```

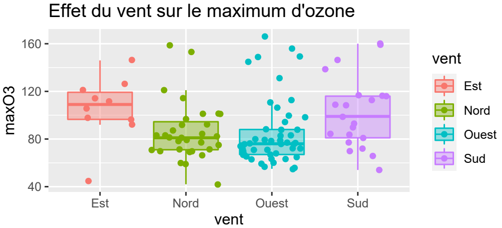
Visualisation de l'effet d'une variable qualitative




```
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")
```

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

Visualisation de l'effet d'une variable qualitative



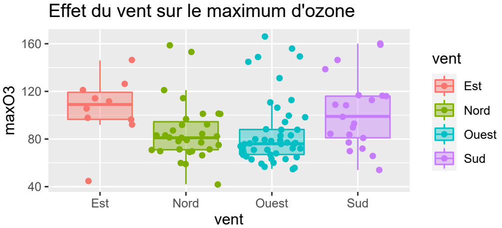

```
ggplot(ozone) +
  aes(x=vent, y=maxO3, fill=vent,col=vent) +
  geom_boxplot(outlier.shape=NA,alpha=0.4) +
  geom_jitter()+
  ggtitle("Effet du vent sur le maximum d'ozone")
```

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

$$\text{Réponse} \; \sim \; var_1$$

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

Visualisation de l'effet d'une variable qualitative



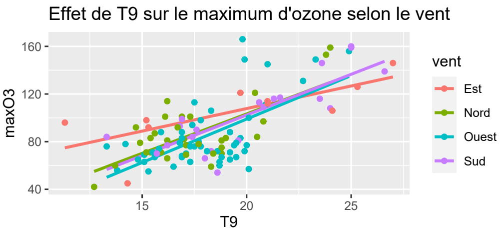

```
ggplot(ozone) +
  aes(x=vent, y=max03, fill=vent,col=vent) +
  geom_boxplot(outlier.shape=NA,alpha=0.4) +
  geom_jitter()+
  ggtitle("Effet du vent sur le maximum d'ozone")
```

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

$$\text{Réponse} \ \sim \ var_1 + var_2 + \ldots$$

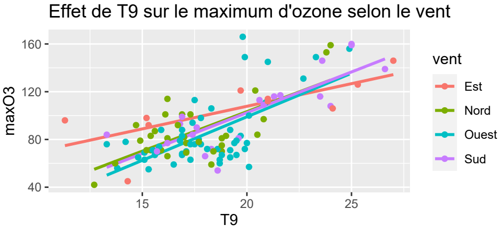
$$\left\{ \begin{array}{ll} \forall i, j, k & Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \\ \forall i, j, k & \varepsilon_{ijk} \ \text{ i.i.d.} \ , \quad \mathbb{E}(\varepsilon_{ijk}) = 0, \quad \mathbb{V}(\varepsilon_{ijk}) = \sigma^2 \\ \forall i, j, k & cov(\varepsilon_{ijk}, \varepsilon_{i'j'k'})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

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




```
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")
```

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

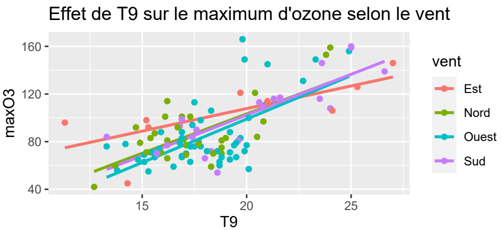



```
ggplot(ozone) +
  aes(x=T9, y=maxO3, col = vent, group=vent) +
  geom_smooth(method="lm", se=FALSE) +
  geom_point()+
  ggtitle("Effet de T9 sur MaxO3 selon le vent")
```

$$MaxO3 \sim vent + T9 + vent : T9$$

$$MaxO3_{ij} \sim \mu + \left\{ \begin{array}{l} \alpha_1 \text{ si vent d'est} \\ \alpha_2 \text{ si vent du nord} \\ \alpha_3 \text{ si vent d'ouest} \\ \alpha_4 \text{ si vent du sud} \end{array} \right\} + \left(\beta + \left\{ \begin{array}{l} \gamma_1 \text{ si vent d'est} \\ \gamma_2 \text{ si vent du nord} \\ \gamma_3 \text{ si vent d'ouest} \\ \gamma_4 \text{ si vent du sud} \end{array} \right\} \right) \times T9ij + alea_{ij}$$

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




```r
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")
```

$$MaxO3 \sim vent + T9 + vent : T9$$

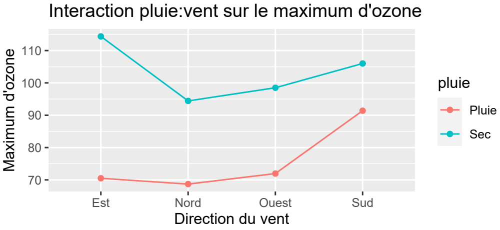
$$MaxO3_{ij} \sim \mu + \left\{ \begin{array}{l} \alpha_1 \text{ si vent d'est} \\ \alpha_2 \text{ si vent du nord} \\ \alpha_3 \text{ si vent d'ouest} \\ \alpha_4 \text{ si vent du sud} \end{array} \right\} + \left(\beta + \left\{ \begin{array}{l} \gamma_1 \text{ si vent d'est} \\ \gamma_2 \text{ si vent du nord} \\ \gamma_3 \text{ si vent d'ouest} \\ \gamma_4 \text{ si vent du sud} \end{array} \right\} \right) \times T9ij + alea_{ij}$$

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

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

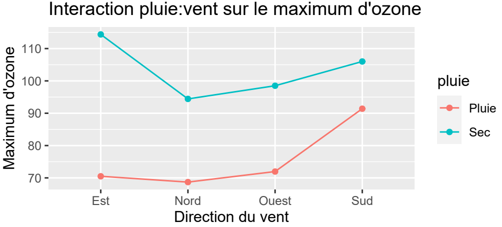



```
library(dplyr)
ozone %>% group_by(vent, pluie) %>%
  summarize(MOY = mean(maxO3)) %>%
  ggplot() +
  aes(x=vent, y=MOY, col=pluie, group=pluie) +
  geom_line() + geom_point() +
  ggtitle("Interaction pluie:vent sur maxO3")+
  xlab("Direction du vent")+
  ylab("Maximum d'ozone")
```

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




```
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")
```

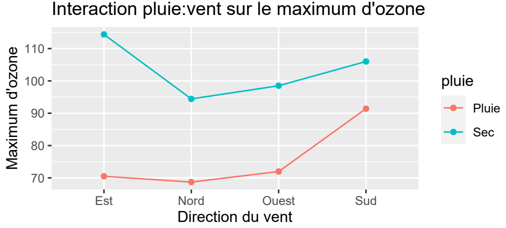
$$MaxO3 \sim vent + pluie + vent : pluie$$

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

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




```
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")
```

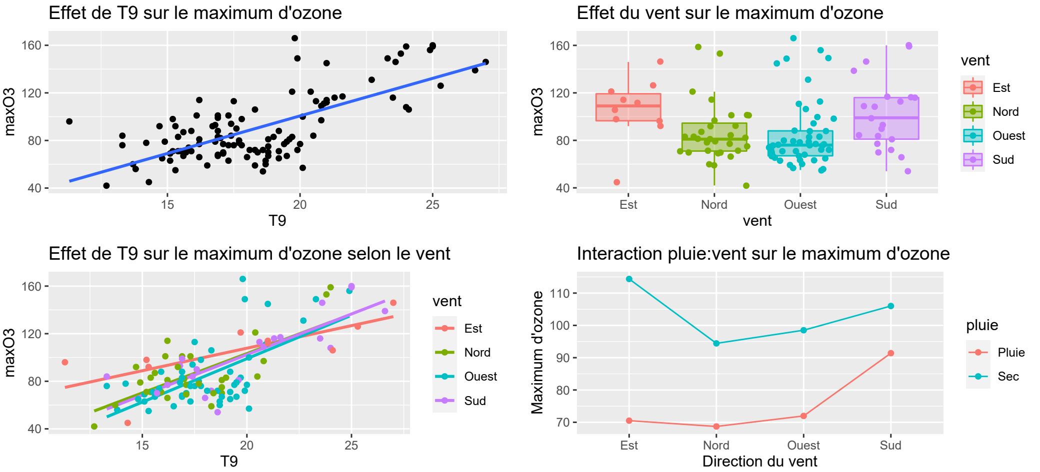
$$MaxO3 \sim vent + pluie + vent : pluie$$

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

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

Les effets d'un modèle

Quatre types d'effets possibles :



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

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)

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

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

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

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

Ecriture du modèle

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

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

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

Ecriture du modèle

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

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

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

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

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

Ecriture du modèle

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

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

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

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

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

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

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

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

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

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

$$Y \approx X\beta$$

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

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

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

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

$$Y \approx X\beta$$

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

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

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

La variance des résidus σ^2 est estimée par :

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

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 modèle résiduelle

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

Propriétés : $0 \leq R^2 \leq 1$

Décomposition de la variabilité

$$\begin{array}{cccc} & \sum_{i=1}^{n}(y_i - \bar{y})^2 & = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 & + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ \text{Variabilité} & \text{totale} & \text{modèle} & \text{résiduelle} \end{array}$$

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)
```

Exemple sur l'ozone

```
library(FactoMineR)
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 : $F_{obs} = \dfrac{SC_{\mathcal{V}}/ddl_{\mathcal{V}}}{SC_R/ddl_R} = \dfrac{CM_{\mathcal{V}}}{CM_R}$

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

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

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 : $F_{obs} = \dfrac{SC_{\mathcal{V}}/ddl_{\mathcal{V}}}{SC_R/ddl_R} = \dfrac{CM_{\mathcal{V}}}{CM_R}$

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

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

Modélisation

Partie 2 : interprétation et prédiction

François Husson

Unité pédagogique de mathématiques appliquées
L'Institut Agro

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 β_0 $\Longrightarrow$ 1 ddl
- chaque variable quantitative correspond à une colonne $\Longrightarrow$ 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 $\Longrightarrow$ on pose une contrainte, le mieux est de prendre $\sum_{i=1}^{I} \alpha_i = 0$ $\Longrightarrow$ $(I-1)$ 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 $\forall\ j, \sum_i \alpha\beta_{ij} = 0$ $\Longrightarrow$ $(I-1)(J-1)$ ddl

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

- avec $\sum_{\mathbf{i}} \alpha_{\mathbf{i}} = \mathbf{0}$, 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 $\Longrightarrow$ 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 :

$$\begin{aligned}\sum_{i,j,k}(y_{ijk}-y_{\bullet\bullet\bullet})^2 &= \sum_{i,j,k}(y_{i\bullet\bullet}-y_{\bullet\bullet\bullet})^2+\sum_{i,j,k}(y_{\bullet j\bullet}-y_{\bullet\bullet\bullet})^2\\ &+\sum_{i,j,k}(y_{ij\bullet}-y_{i\bullet\bullet}-y_{\bullet j\bullet}+y_{\bullet\bullet\bullet})^2+\sum_{i,j,k}(y_{ijk}-y_{ij\bullet})^2\end{aligned}$$

Décomposition de la variabilité : variables qualitatives, données équilibrées

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

$$\begin{aligned}\sum_{i,j,k}(y_{ijk}-y_{\bullet\bullet\bullet})^2 &= \sum_{i,j,k}(y_{i\bullet\bullet}-y_{\bullet\bullet\bullet})^2+\sum_{i,j,k}(y_{\bullet j\bullet}-y_{\bullet\bullet\bullet})^2\\ &+\sum_{i,j,k}(y_{ij\bullet}-y_{i\bullet\bullet}-y_{\bullet j\bullet}+y_{\bullet\bullet\bullet})^2+\sum_{i,j,k}(y_{ijk}-y_{ij\bullet})^2\end{aligned}$$

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

$$\hat{\mu}=y_{\bullet\bullet\bullet} \qquad \forall i,\ \hat{\alpha}_i=y_{i\bullet\bullet}-y_{\bullet\bullet\bullet}$$

$$\forall j,\ \hat{\beta}_j=y_{\bullet j\bullet}-y_{\bullet\bullet\bullet} \qquad \forall i,j\ \widehat{\alpha\beta}_{ij}=y_{ij\bullet}-y_{i\bullet\bullet}-y_{\bullet j\bullet}+y_{\bullet\bullet\bullet}$$

Décomposition de la variabilité : variables qualitatives, données équilibrées

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

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

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

$$\hat{\mu}=y_{\bullet\bullet\bullet} \qquad \forall i,\ \hat{\alpha}_i=y_{i\bullet\bullet}-y_{\bullet\bullet\bullet}$$

$$\forall j,\ \hat{\beta}_j=y_{\bullet j\bullet}-y_{\bullet\bullet\bullet} \qquad \forall i,j\ \widehat{\alpha\beta}_{ij}=y_{ij\bullet}-y_{i\bullet\bullet}-y_{\bullet j\bullet}+y_{\bullet\bullet\bullet}$$

Décomposition de la variabilité : variables qualitatives, données équilibrées

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

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

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

$$\hat{\mu}=y_{\bullet\bullet\bullet} \qquad \forall i,\ \hat{\alpha}_i=y_{i\bullet\bullet}-y_{\bullet\bullet\bullet}$$

$$\forall j,\ \hat{\beta}_j=y_{\bullet j\bullet}-y_{\bullet\bullet\bullet} \qquad \forall i,j\ \widehat{\alpha\beta}_{ij}=y_{ij\bullet}-y_{i\bullet\bullet}-y_{\bullet j\bullet}+y_{\bullet\bullet\bullet}$$

$\Longrightarrow$ 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
$\Longrightarrow$ 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 : $\forall i \quad \alpha_i = 0$ contre H_1 : $\exists i \ / \ \alpha_i \neq 0$

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

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

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

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

Décision : $\mathbb{P}(\mathcal{F}_{ddl_R}^{ddl_A} > F_{obs}) < 0.05 \quad \Longrightarrow$ 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 : $H_0 : \quad \forall i,j \quad \alpha\beta_{ij} = 0$ contre $H_1 : \quad \exists (i,j) \ / \ \alpha\beta_{ij} \neq 0$

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

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

Décision : $\mathbb{P}(\mathcal{F}_{ddl_R}^{ddl_{interaction}} > F_{obs}) < 0.05 \quad \Longrightarrow$ 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 : $T_{obs} = \dfrac{\hat{\alpha}_1}{\hat{\sigma}_{\hat{\alpha}_1}}$

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

Décision : $\mathbb{P}(\mathcal{T}_{ddl_R} > |T_{obs}|) < 0.05 \implies$ 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 : $T_{obs} = \dfrac{\hat{\alpha}_1}{\hat{\sigma}_{\hat{\alpha}_1}}$

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

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

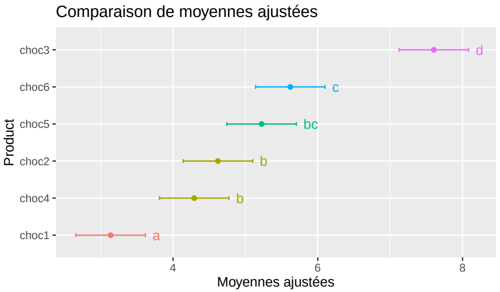
On peut comparer les modalités 2 à 2 grâce à un test de comparaison par paire, avec une correction de Bonferroni par exemple

```
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

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

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
```

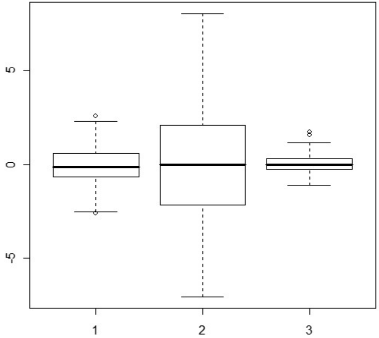
```
predict(model,xnew,interval="confidence")
```

```
       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**

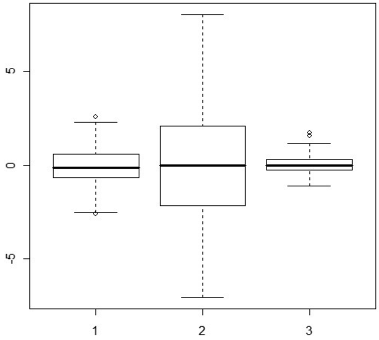
```
model <- lm(maxO3~T12+...,data=ozone)
res <- residuals(model)
boxplot(res~vent,data=ozone)
bartlett.test(res~vent,data=ozone)
```



Analyse des résidus du modèle

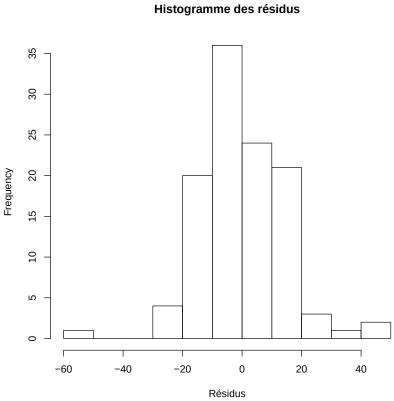
Test de Bartlett de l'homoscédasticité des **Résidus**

```
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 $\Longrightarrow$ 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")
```

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 $\Longrightarrow$ 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 = maxO3 ~ (T9 + T12 + T15 + Ne9 + Ne12 + Ne15 +
    Vx9 + Vx12 + Vx15 + maxO3v + 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 = maxO3 ~ T9 + T15 + Ne12 + Vx9 + maxO3v + 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

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'03
- 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^o à 9h est plus important (coef $= 6.14$) ; au contraire quand le vent vient de l'est. C'est l'inverse pour la T^o à 15h
- Ainsi, quand le vent vient du nord, l'effet de la T^o 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.)