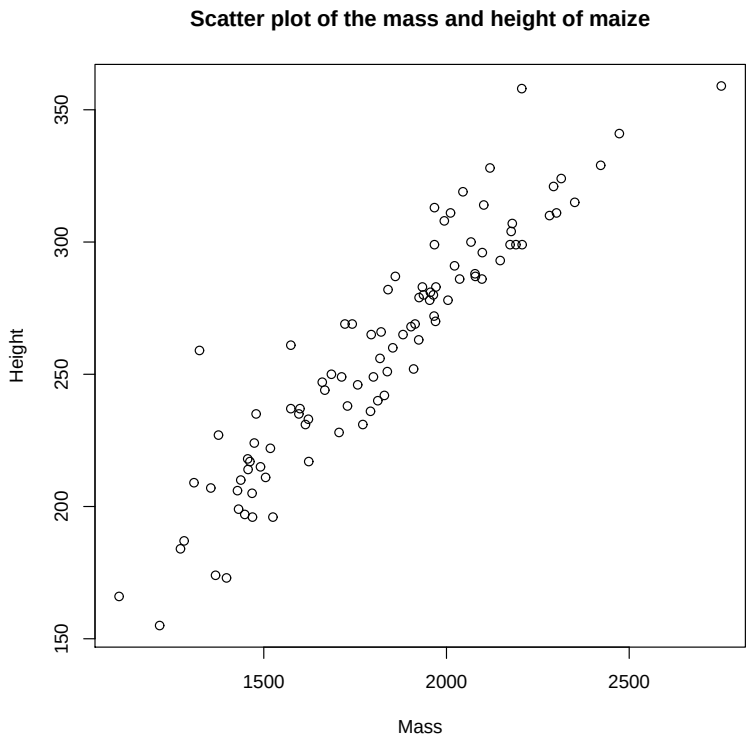
Statistique avec SAS : Régression linéaire avec SAS®

Le modèle linéaire.



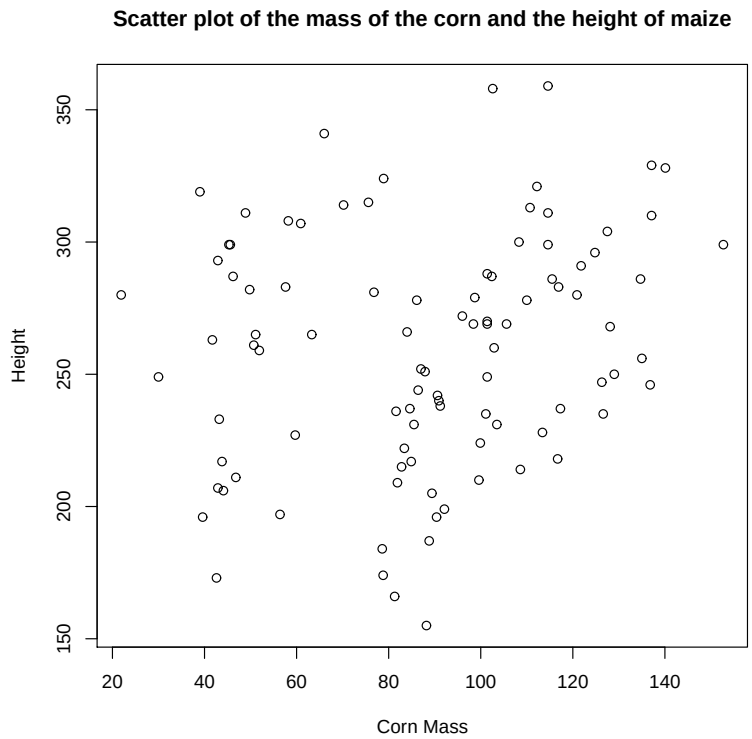


Figure – Exemples de nuages de points.

Un modèle et des estimations.

$$Y = aX + b + \varepsilon$$

où

- Y et X : variables aléatoires
- a et b paramètres inconnus
- ε : erreur : variable aléatoire suivant une loi normale centrée.

$$y_i = \widehat{a}x_i + \widehat{b} + e_i$$

où

- $(x_i,\ y_i)$ observations
- $\widehat{a}$ et $\widehat{b}$: paramètres estimés
- e_i : résidus

Le coefficient de détermination.

La qualité d'un modèle de régression est donné par le coefficient R^2.

$$R = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

où

- σ_X est l'écart-type de X
- σ_Y est l'écart-type de Y
- $Cov(X, Y)$ est la covariance de X et Y.

Le coefficient de détermination.

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

- 1 est le cas parfait où tous les points sont sur la droite de régression.
- 0 est le cas où le modèle n'explique rien de la variation.

R^2 est la proportion de la dispersion totale qui est expliquée par le modèle de régression linéaire.

Estimation des paramètres.

$$a = \frac{Cov(X, Y)}{\sigma_X}$$

$\widehat{a}$ est obtenu en remplaçant $Cov(X, Y)$ et σ_X par leurs estimateurs.

Le point moyen $(\overline{x}, \overline{y})$ est sur la droite donc :

$$\widehat{b} = \overline{y} - \widehat{a}\overline{x}$$

Comme a et b, R^2 est inconnu et estimé.

Un autre échantillon produira des estimations différentes.

Des tests d'hypothèses peuvent être utilisés pour déterminer la vraie valeur.

Nuage de points avec SAS®.

```
PROC GPLOT DATA=mais1 ;
      plot Hauteur*Masse ;
run ;
```

Cette fonction fait partie du package SAS graph qui n'est pas disponible dans SAS University.

Nuage de points avec SAS®.

```
PROC SGPLOT DATA=mais1 ;
      scatter y=Hauteur x=Masse ;
run ;
```

Ajustement d'un modèle linéaire avec SAS®.

La procédure utilisée pour ajuster un modèle linéaire est PROC REG.

```
PROC REG DATA=mais1 ;
      model Hauteur=Masse ;
run ;
```

Test pour R^2.

Hypothèses

- H_0 : $R^2 = 0$ dans la population cible.
- H_1 : $R^2 \neq 0$ dans la population cible.

Conditions d'application

1. Les individus composant l'échantillon ont été choisis de façon aléatoire. (indépendance des observations)
2. Les deux variables étudiées doivent être observées pour chaque individu.
3. Pour chaque valeur de x_i, la distribution des y_i doit être normale. Toutes ces lois normales doivent avoir la même variance. De manière équivalente, les résidus doivent suivre une loi normale.
4. Les résidus doivent être indépendants des variables X et Y.

Ce test est robuste à la non-normalité pour des échantillons assez grands $\Longrightarrow$ Q-Q plot.

Sous H_0 la statistique de décision suit une loi de Fisher-Snedecor de paramètres $\nu_1 = 1$ et $\nu_2 = n - 2$.

Remarque

- *ν_1 et ν_2 sont des nombres de degrés de liberté.*
- *n est la taille de l'échantillon.*
- *La statistique de décision est généralement notée F.*

Tests pour la pente et l'ordonnée à l'origine.

Le test le plus intéressant est celui sur la pente.

Hypothèses

- $H_0 : a = a_{th}$.
- $H_{1,bilateral} : a \neq a_{th}$.
- $H_{1,unilateral} : a > a_{th}$.
- $H_{1,unilateral} : a < a_{th}$.

Conditions d'application : Les mêmes que pour le test précédent.

Sous H_0 la statistique de décision suit une loi de Student à $\nu = n-2$ d.d.l (pour la pente et pour l'ordonnée à l'origine).

Vérification des conditions d'application.

```
PROC REG DATA=mais1 ;
      model Hauteur=Masse ;
      output out= mais2 r=residus ;
run ;

PROC UNIVARIATE DATA=mais2 normal ;
      var residus ;
run ;
```

Vérification des conditions d'application.

- Les tests associés à la régression linéaire sont relativement robustes à la non-normalité des résidus $\Longrightarrow$ QQ-plot.
- le QQ-plot est tracé automatiquement
- avec d'autres graphes qui ont aussi un intérêt.

Pour mieux voir ces graphes :

```
PROC REG DATA=mais1 plots=diagnostics(unpack) ;
      model Hauteur=Masse ;
run ;
```

Vérification des conditions d'application.

Il est possible de voir uniquement les graphes qui nous intéressent :

```
PROC REG DATA=mais1 plots=(qqplot Residuals(smooth)
cooksd(label) residualbypredicted(label)) ;
      model Hauteur=Masse ;
run ;
```

Ce n'est qu'une fois les conditions d'applications vérifiées que l'on peut regarder les résultats des tests.

Nuage de point avec la droite de régression.

Avec l'option plots il est possible de le modifier :

plots=fitplot(*options*)

avec les *options* :

- NOCLI : pas de limite de prédiction
- NOCLM : pas de limite de confiance
- NOLIMITS : ni l'un ni l'autre
- ...

Régression multiple.

- On veut expliquer la variable Masse.
- On peut l'expliquer par Hauteur...
- mais choix subjectif
- On va sélectionner le modéle optimal selon des critères objectifs.

Comment choisir les variables explicatives ?

- Sélection selon :
  - question scientifique
  - littérature
  - a priori

  $\Longrightarrow$ Subjectivité ?

- sur critères statistiques

  $\Longrightarrow$ Objectivité !

Le modèle complet.

```
PROC REG DATA=mais1 ;
      model Masse=Hauteur hauteur_j7 nb-grains
nb_jours_attaque ;
run ;
```

Remarque

- *Le diagnostique (ie vérification des conditions d'application des tests) se fait comme pour la régression linéaire simple.*
- *Seules les observations sans données manquantes pour toutes les variables utilisées pourront être utilisées.*
- *Risque que des variables explicatives soient trop corrélées. Cela fausserait les résultats du modèle.*

Sélection du modèle.

1. Etude de la corrélation entre les variables explicatives

- Variance inflation factor (VIF) : mesure de la sévérité de la multicollinéarité pour une régression linéaire multiple.
- Une valeur par variable explicative X_i.
- $\text{VIF}_i = \frac{1}{1 - R_i^2}$ avec R_i^2 la qualité du modèle avec toutes les variables prédictives sauf X_i qui est alors la variable à expliquer.

Sélection du modèle.

- Une valeur de VIF>10 montre que la colinéarité entre cette variable et les autres est trop forte.
- Il faut donc retirer cette variable car on a de l'information redondante.

Remarque

- *La valeur 10 est arbitraire.*
- *Certaines personnes recommandent une valeur de 3.*

Calcul des VIF.

```
PROC REG DATA=mais1 ;
      model Masse=Hauteur hauteur_j7 nb-grains nb_jours_attaque
/ VIF ;
run ;
```

On doit retirer en premier la variable avec le plus grand VIF.

```
PROC REG DATA=mais1 ;
      model Masse=Hauteur nb-grains nb_jours_attaque / VIF ;
run ;
```

OK

Sélection du modèle.

2 Sélection du modèle optimal

- Plus un modèle à de paramètres, moins l'estimation de ces paramètres est précise
- et les tests associés "efficaces".
- Il faut de plus un nombre d'observation suffisant par paramètre : 10 observations par paramètre estimé.
- Pour chaque variable quantitative, un paramètre est estimé.
- Il faut aussi compter l'ordonnée à l'origine
- Il n'est donc généralement pas recommandé d'utiliser le modèle complet...
- mais un modèle optimal.

Comparaisons de modèles

- AIC :
  - Akaike Information Criterion
  - vraisemblance du modèle avec pénalité pour le nombre de paramètres
  - Plus petit AIC $\Longrightarrow$ meilleur modèle.
- BIC
  - Bayesian Information Criterion
  - similaire à l'AIC mais avec pénalité pour le nombre de paramètre en fonction de la taille d'échantillon.
  - Plus petit BIC $\Longrightarrow$ meilleur modèle.
- Test du Rapport de Vraisemblance
  - Log Likelyhood Ratio Test
  - p-value
  - moins connu
  - mais toujours utilisable.

Procédure Backward

- Ajustement d'un modèle M_1 sur toutes les variables disponibles.
- Choix d'un critère.
- Ajustement d'un modèle M_2 sur les mêmes variables que le M_1 sauf une.
- Comparaison entre les 2 modèles.

Remarque

- *Dans la pratique on compare le modèle complet à tous les sous-modèles.*
- *Critère en faveur de M_2 : on recommence en considérant comme modèle complet M2.*
- *Critère en faveur de M_1 : arrêt : M1 est le modèle optimal.*

Procédure Backward avec l'AIC ou le BIC

```
PROC REG DATA=mais1 ;
      model Masse=Hauteur nb-grains nb_jours_attaque /
selection=selection AIC BIC ;
run ;
```

On ajuste le modèle optimal :

```
PROC REG DATA=mais1 ;
      model Masse=Hauteur ;
run ;
```

C'est sur ce modèle qu'on va prendre nos conclusions.

Procédure Backward avec l'AIC ou le BIC

```
PROC REG DATA=mais1 ;
      model Masse=Hauteur nb-grains nb_jours_attaque /
selection=Rsquare AIC BIC ;
run ;
```

On ajuste le modèle optimal :

```
PROC REG DATA=mais1 ;
      model Masse=Hauteur ;
run ;
```

C'est sur ce modèle qu'on va prendre nos conclusions.

Procédure Backward de SAS

```
PROC REG DATA=mais1 ;
      model Masse=Hauteur nb-grains nb_jours_attaque /
selection=backward ;
run ;
```

Le critère de sélection est un test de type LRT avec un seuil par défaut à 0.1.
On peut spécifier le seuil avec l'option SLSTAY.

```
PROC REG DATA=mais1 ;
      model Masse=Hauteur nb-grains nb_jours_attaque /
selection=backward SLSTAY=0.05 ;
run ;
```

Procédure Stepwise de SAS

```
PROC REG DATA=mais1 ;
      model Masse=Hauteur nb-grains nb_jours_attaque /
selection=stepwise ;
run ;
```