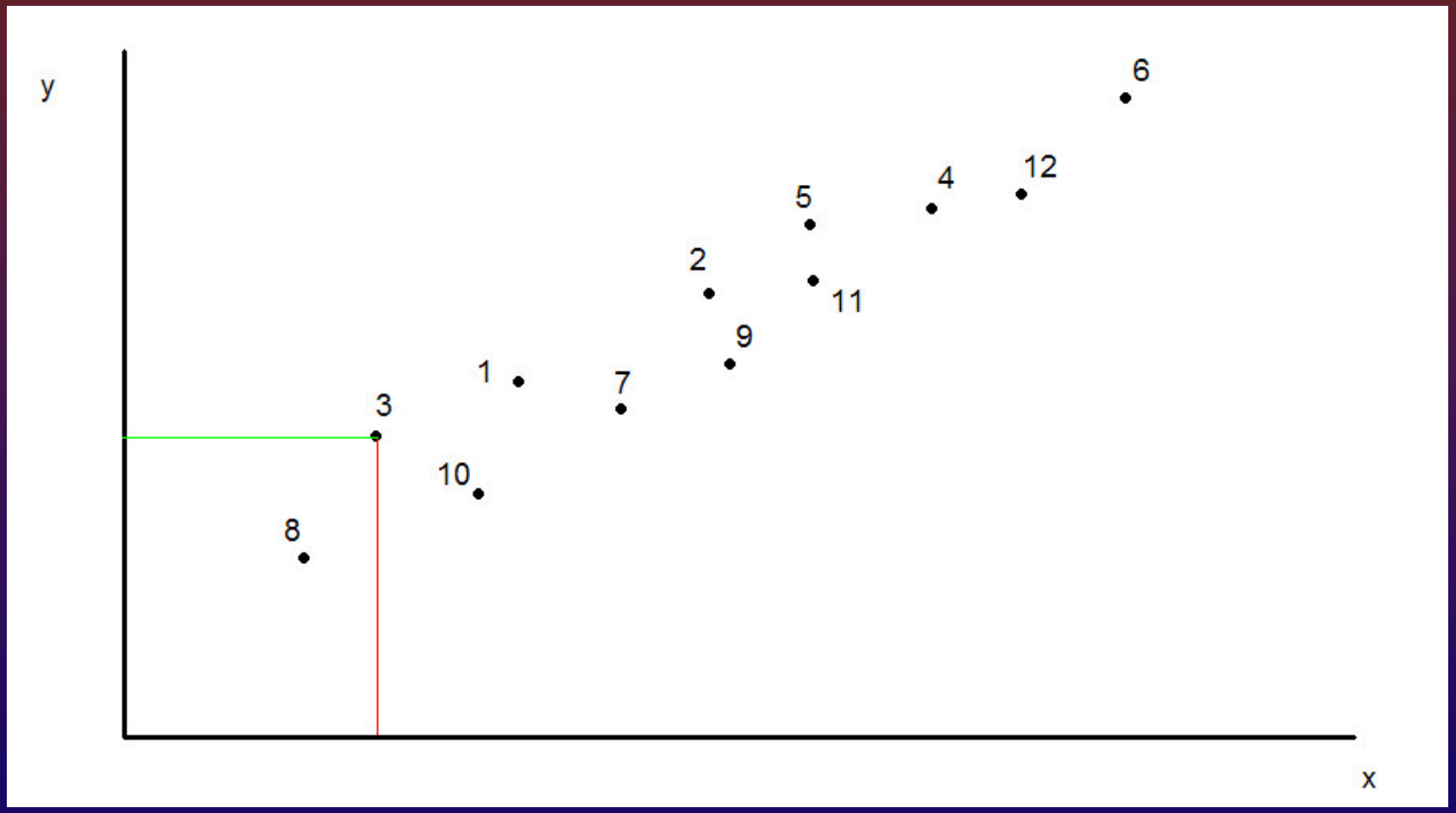
# Corrélation et Régression

Relations entre 2 variables

# Représentation graphique



# La covariance

$$\mathrm{cov} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Formule analogue à la variance

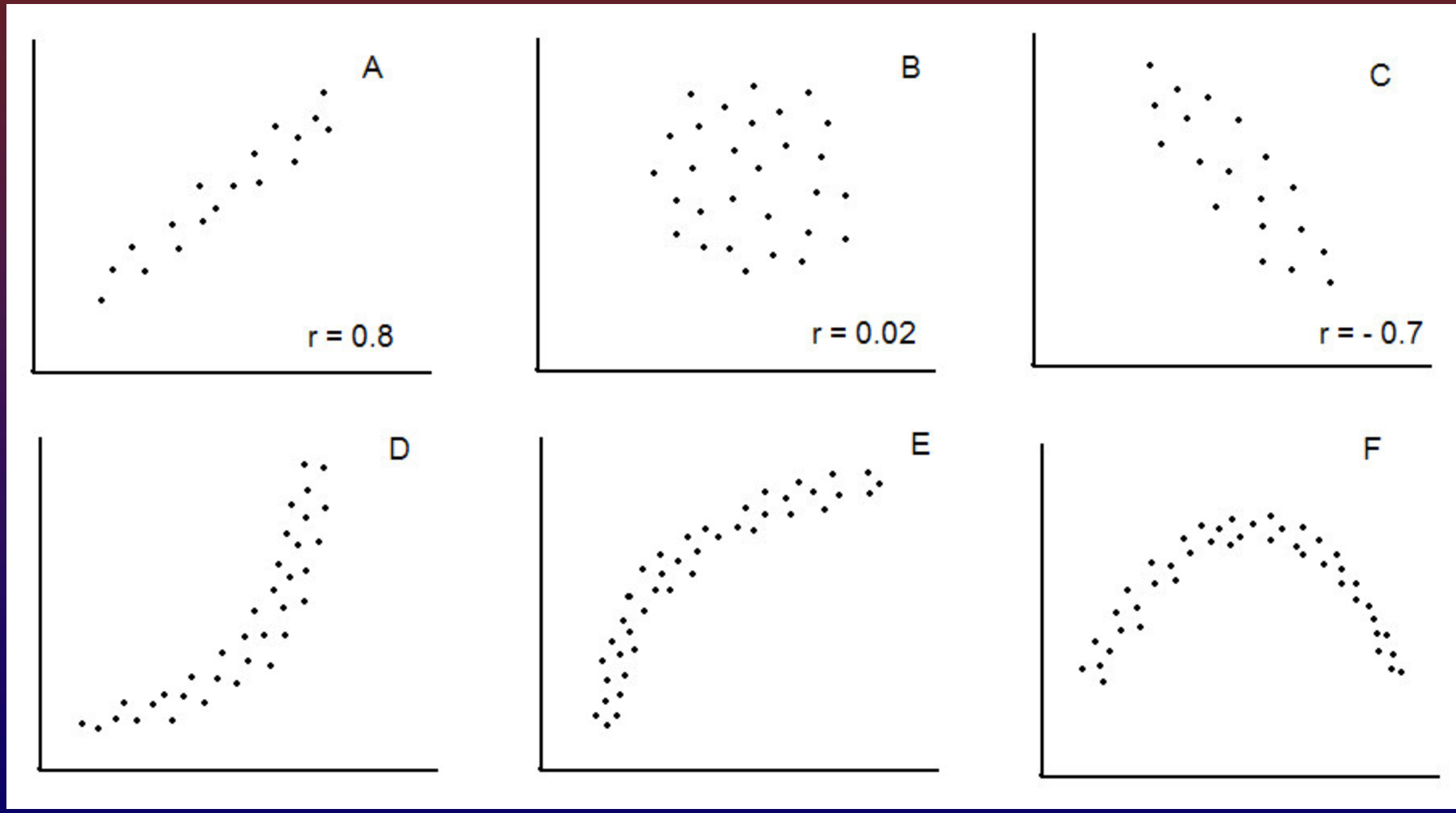
# La corrélation

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})(x_i - \bar{x}) \sum (y_i - \bar{y})(y_i - \bar{y})}}$$

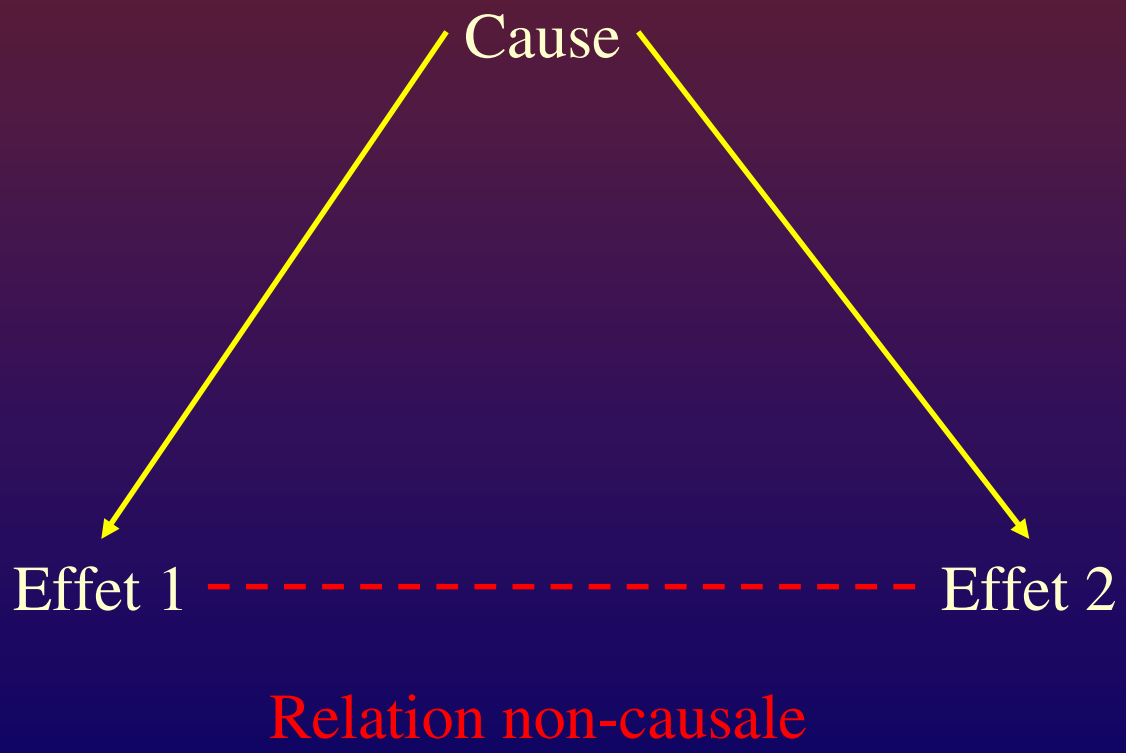
C'est la covariance ramenée à un intervalle de 0 à 1

$$r_{xy} = \frac{\mathrm{COV}}{S_x \cdot S_y}$$

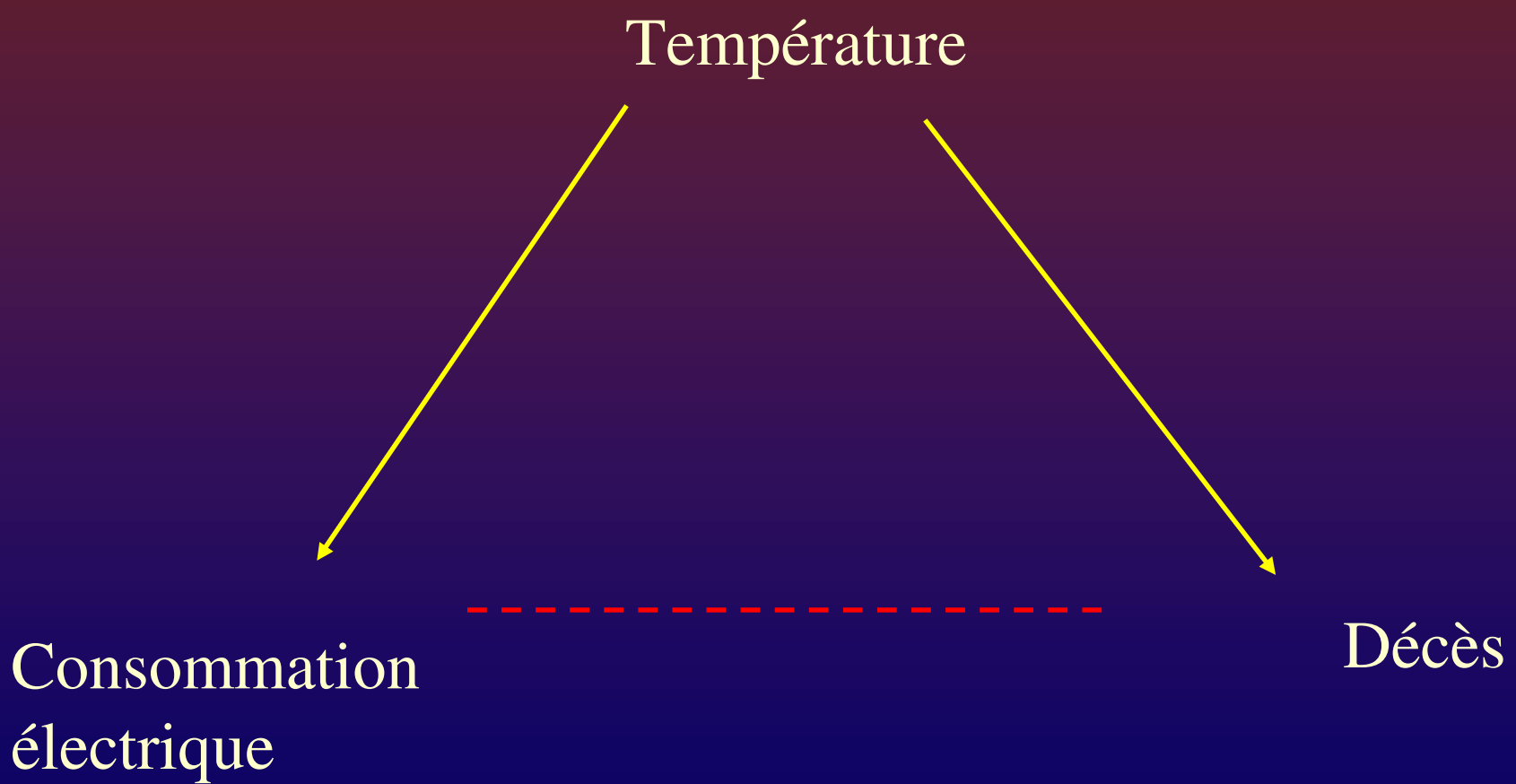
# Différentes associations



# L’interprétation des corrélations



# Exemple d'interprétation



# Applications des corrélations

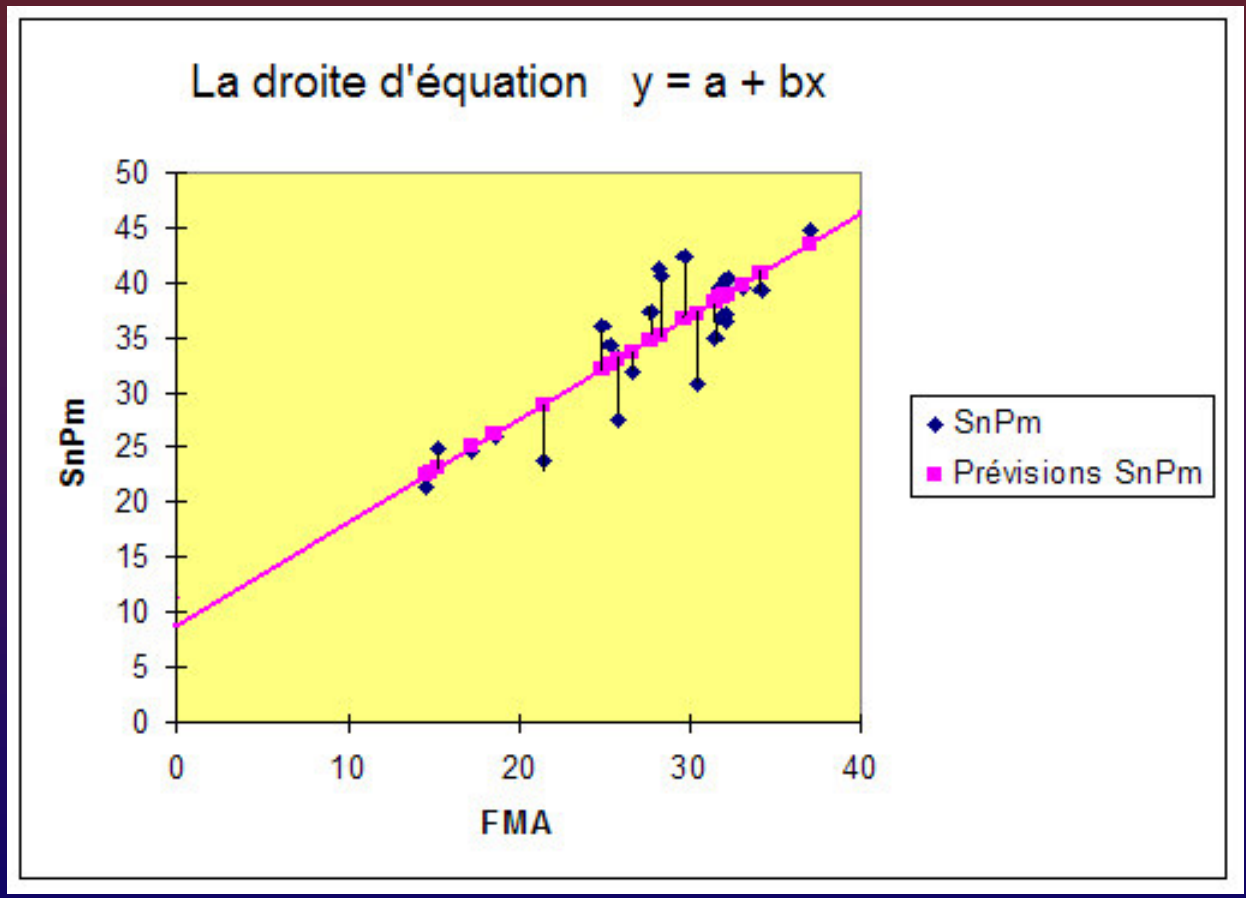
- Recherche de facteurs clés
- Recherche d'hypothèses (heuristique)
- Étude de répétabilité
- Prédiction (régression)

# Les données à analyser

| SnPm | FMA |
|---|---|
| 36.01 | 24.94 |
| 36.78 | 31.94 |
| 22.98 | 14.85 |
| 37.42 | 27.78 |
| 34.97 | 31.43 |
| 27.42 | 25.74 |
| 39.46 | 33.09 |
| 23.69 | 21.44 |
| 40.47 | 32.14 |
| 41.24 | 28.11 |
| 34.27 | 25.38 |
| 33.14 | 25.74 |
| 21.45 | 14.46 |
| 44.68 | 37.02 |
| 31.94 | 26.6 |
| 36.47 | 32.03 |
| 40.57 | 28.24 |
| 24.86 | 15.34 |
| 42.37 | 29.71 |
| 39.61 | 31.81 |
| 26.02 | 18.59 |
| 30.69 | 30.34 |
| 24.72 | 17.29 |
| 39.32 | 34.22 |
| 37.1 | 32.09 |

25 individus

# La régression linéaire



# Propriétés de la régression

• Relation asymétrique avec une variable explicative x et une variable dépendante y
•Les x sont connus sans erreur
• Les résidus sont les écarts à la droite de régression mesurés dans le sens des y
•La somme des résidus d'une régression est égale à 0
•La droite de régression passe par la moyenne de x et la moyenne de y

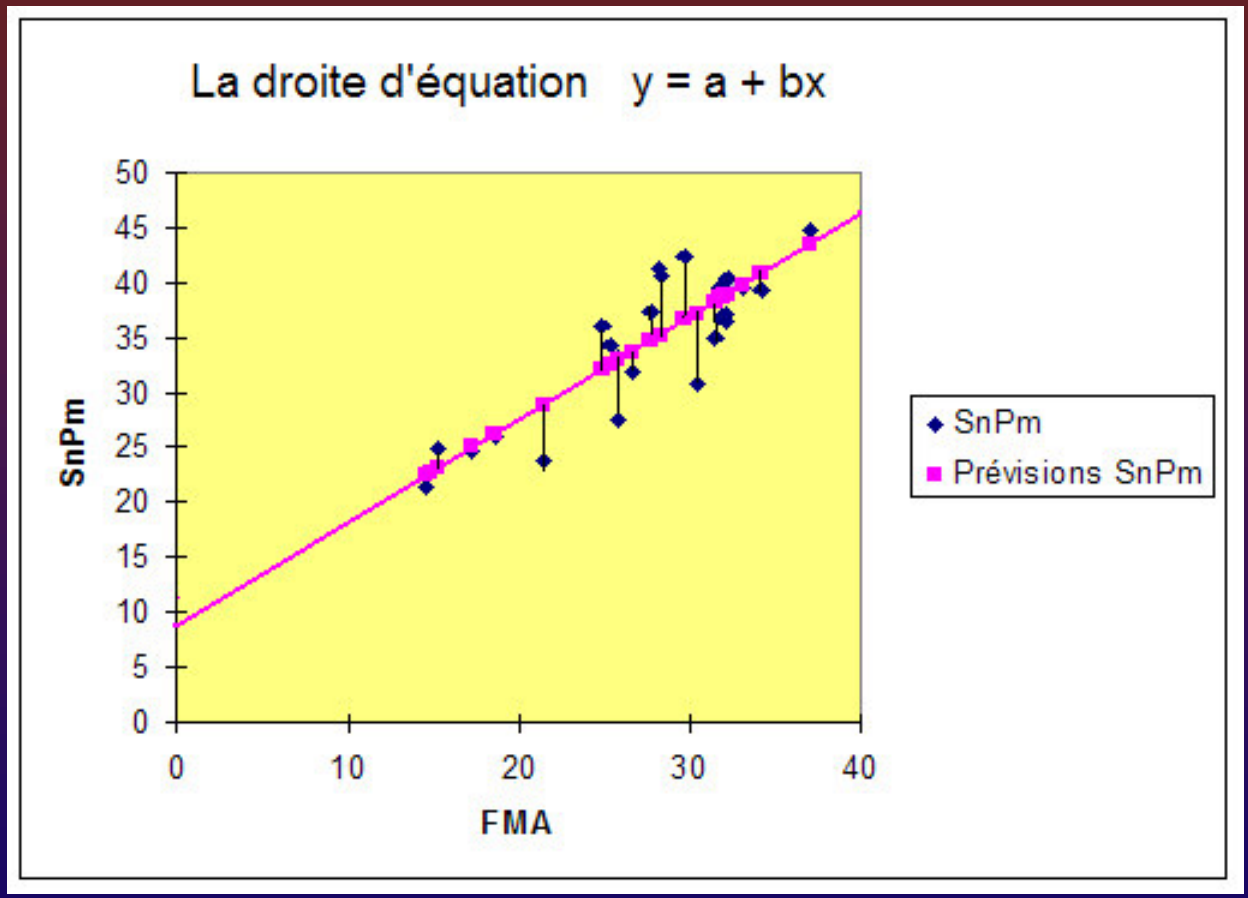
# Calcul des paramètres de y=bx+a

Pente : $$b = \frac{\sum (y - \bar{y})(x - \bar{x})}{\sum (x - \bar{x})^2}$$

Intercept : $$a = \bar{y} - b\bar{x}$$

# Calculs Excel©

| SnPm | FMA |
|---|---|
| 36.01 | 24.94 |
| 36.78 | 31.94 |
| 22.98 | 14.85 |
| 37.42 | 27.78 |
| 34.97 | 31.43 |
| 27.42 | 25.74 |
| 39.46 | 33.09 |
| 23.69 | 21.44 |
| 40.47 | 32.14 |
| 41.24 | 28.11 |
| 34.27 | 25.38 |
| 33.14 | 25.74 |
| 21.45 | 14.46 |
| 44.68 | 37.02 |
| 31.94 | 26.6 |
| 36.47 | 32.03 |
| 40.57 | 28.24 |
| 24.86 | 15.34 |
| 42.37 | 29.71 |
| 39.61 | 31.81 |
| 26.02 | 18.59 |
| 30.69 | 30.34 |
| 24.72 | 17.29 |
| 39.32 | 34.22 |
| 37.1 | 32.09 |




Les angles Sn-Plan mandibulaire
et FMA mesurés chez n=25 individus

# Résultats Excel©

RAPPORT DÉTAILLÉ

| *Statistiques de la régression* | |
|---|---|
| Coefficient de | 0.877251502 |
| Coefficient de | 0.769570197 |
| Coefficient de | 0.75955151 |
| Erreur-type | 3.363094174 |
| Observations | 25 |

ANALYSE DE VARIANCE

| | *Degré de liberté* | *Somme des carrés* | *Moyenne des carrés* | *F* | *Valeur critique de F* |
|---|---|---|---|---|---|
| Régression | 1 | 868.7913443 | 868.7913443 | 76.81347771 | 8.65687E-09 |
| Résidus | 23 | 260.1392557 | 11.31040242 | | |
| Total | 24 | 1128.9306 | | | |

| | *Coefficients* | *Erreur-type* | *Statistique t* | *Probabilité* |
|---|---|---|---|---|
| Constante | 8.943632526 | 2.92652207 | 3.056061876 | 0.005600292 |
| FMA | 0.930986972 | 0.106224547 | 8.76432985 | 8.65687E-09 |

# Analyse de variance

| Origine | Somme des carrés | DDL | Carrés moyens | F |
|---|---|---|---|---|
| Régression | $=\sum(\hat{y}-\bar{y})$ | 1 | CMreg | CMreg/$s^2$ |
| Résidus | $=\sum(\hat{y}-y)$ | n-1-1 | $s^2$ | |
| Totale | $=\sum(y-\bar{y})$ | n-1 | | |

DDL1=1

DDL2=n-2

# Le coefficient de détermination

$$R^2 = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}$$

$\hat{y}$ y calculé

$R^2$ est le rapport de variance expliquée / variance totale

# Le F rapide

On transforme le $R^2$ en F (Test de Fisher) avec l'équation :

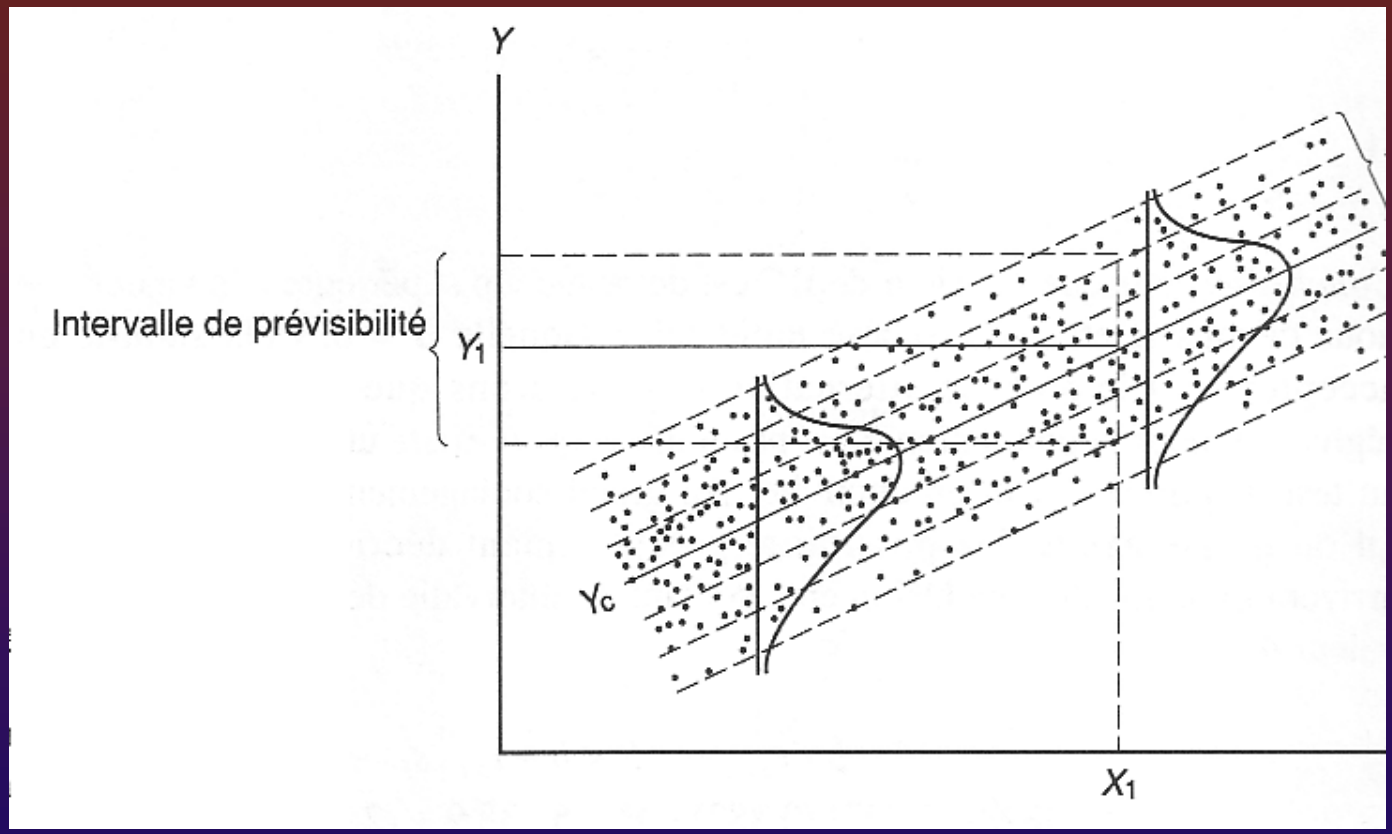
$$F = (n-2)\cdot\frac{R^2}{1-R^2}$$

$n$ le nombre d'individus

DDL1 = 1

DDL2 = $n$-2

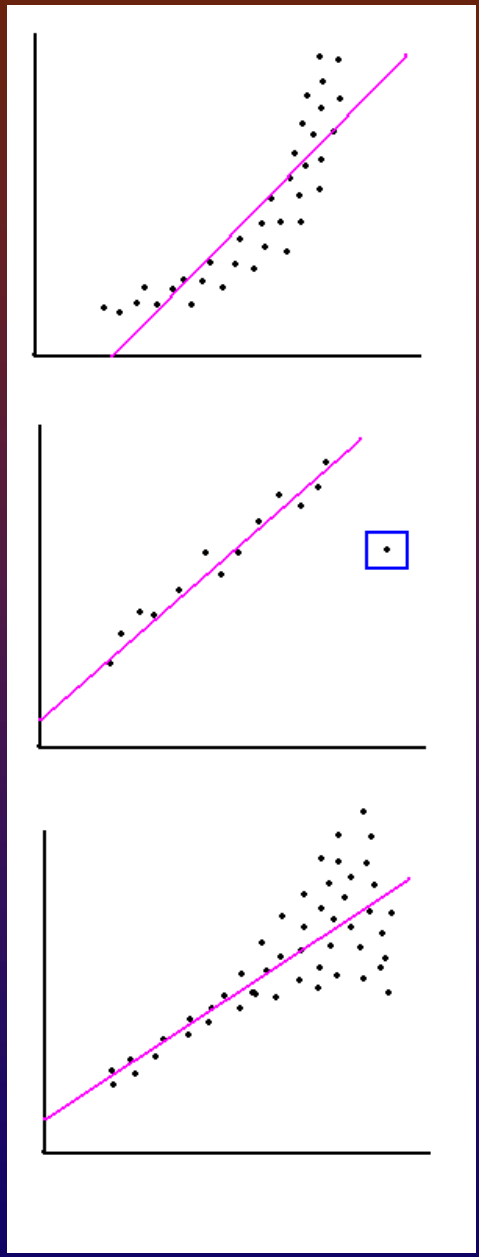
# Les tests sur la pente et l'intercept



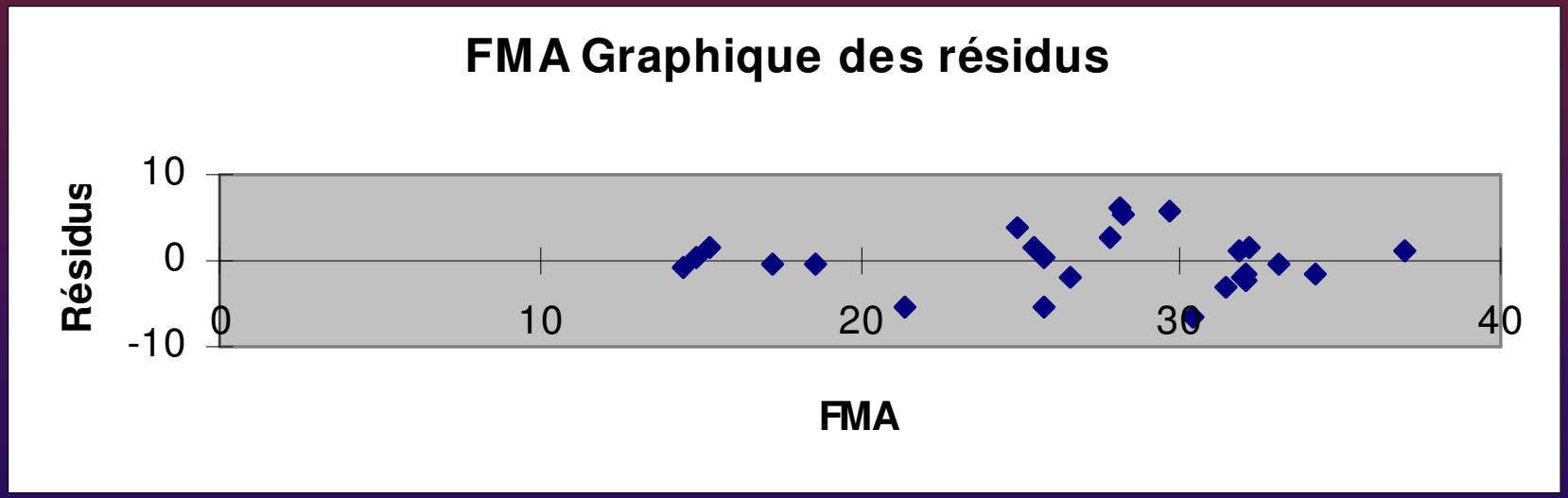

La distribution des t suit une loi de Student bilatérale à n-2 DL
Donc rejet si la valeur absolue de t > t(n-2; $\alpha$/2)

# Analyse visuelle de la droite

1. Modèle linéaire inadapté
2. Présence de points suspects
3. Hétéroscédasticité

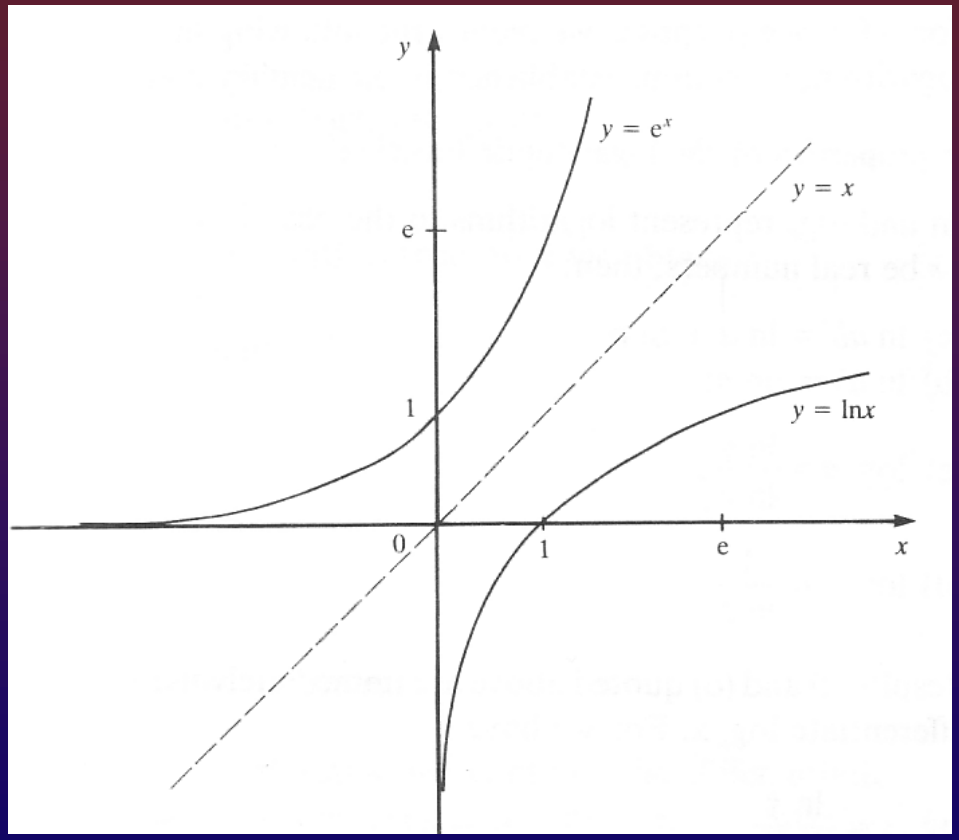


# Analyse visuelle des résidus




Il ne faut pas de structure

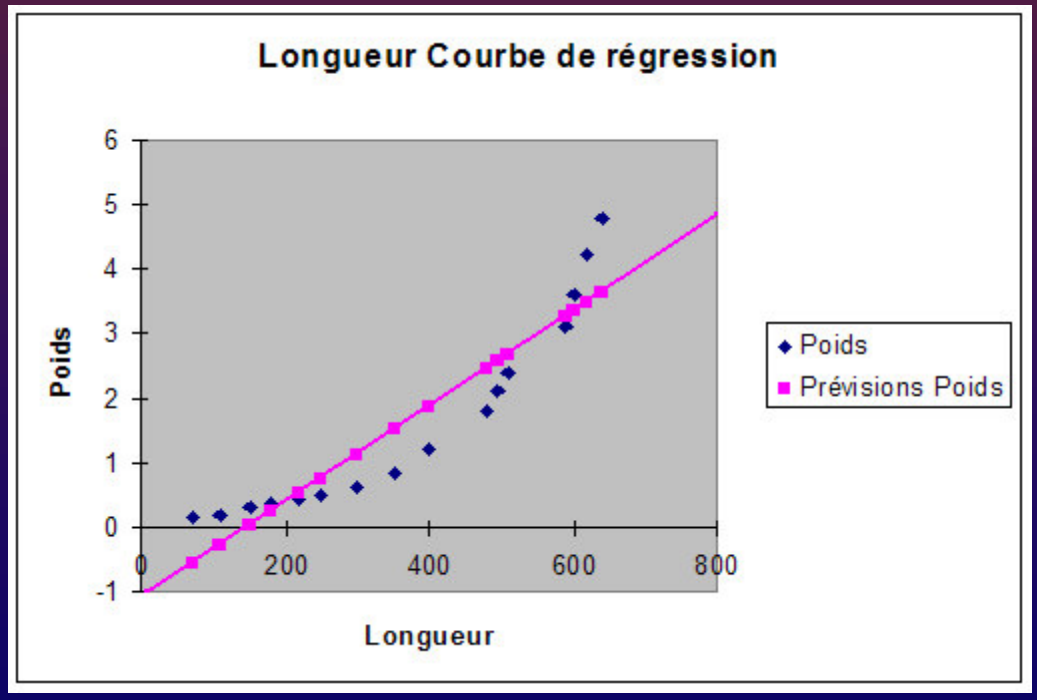
# La transformation logarithmique



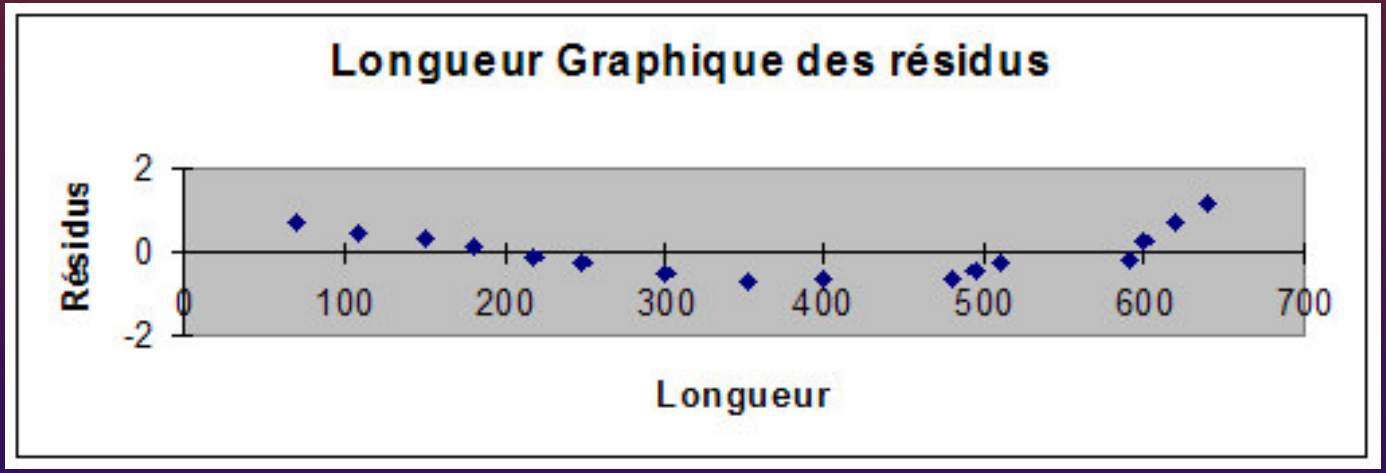

1. Linéarise les formes quadratiques
2. Corrige l'hétéroscédasticité
3. Rend les distributions plus normales

# Exemple de la régression poids - longueur

| Y | X |
|---|---|
| Poids | Longueur |
| 0.14 | 70 |
| 0.18 | 110 |
| 0.31 | 150 |
| 0.38 | 180 |
| 0.43 | 219 |
| 0.5 | 248 |
| 0.62 | 300 |
| 0.83 | 352 |
| 1.2 | 400 |
| 1.8 | 480 |
| 2.1 | 495 |
| 2.4 | 510 |
| 3.1 | 590 |
| 3.6 | 600 |
| 4.23 | 620 |
| 4.8 | 640 |



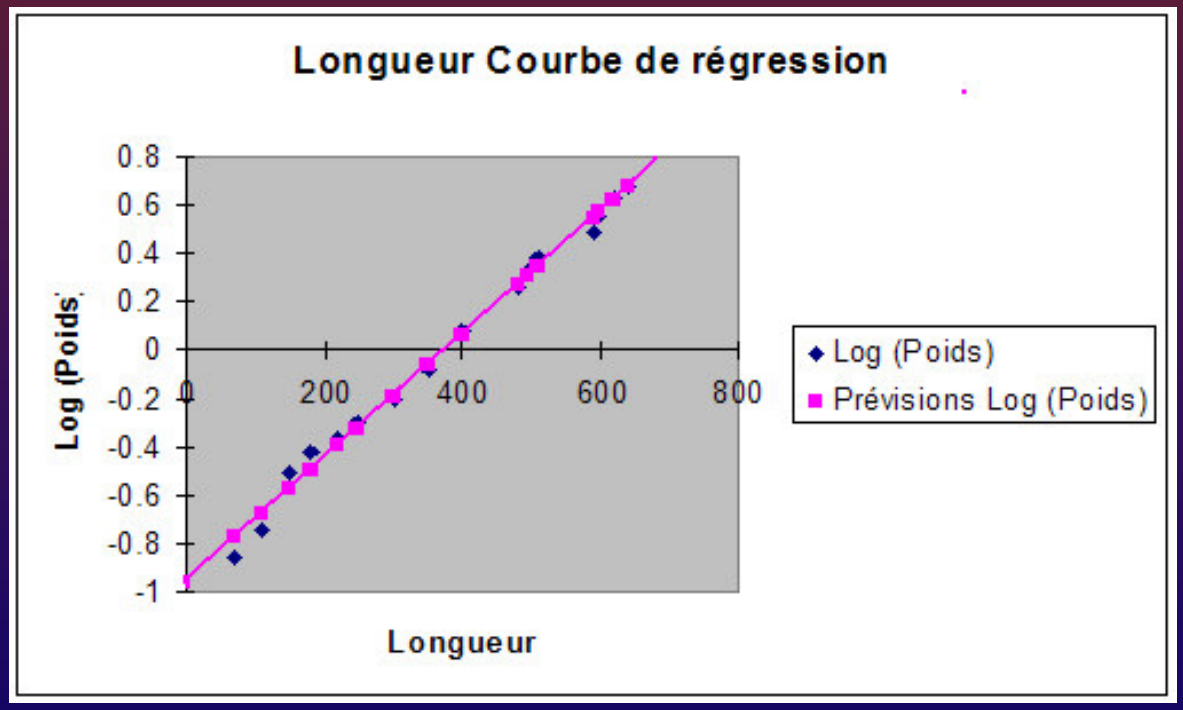
# Le graphique des résidus



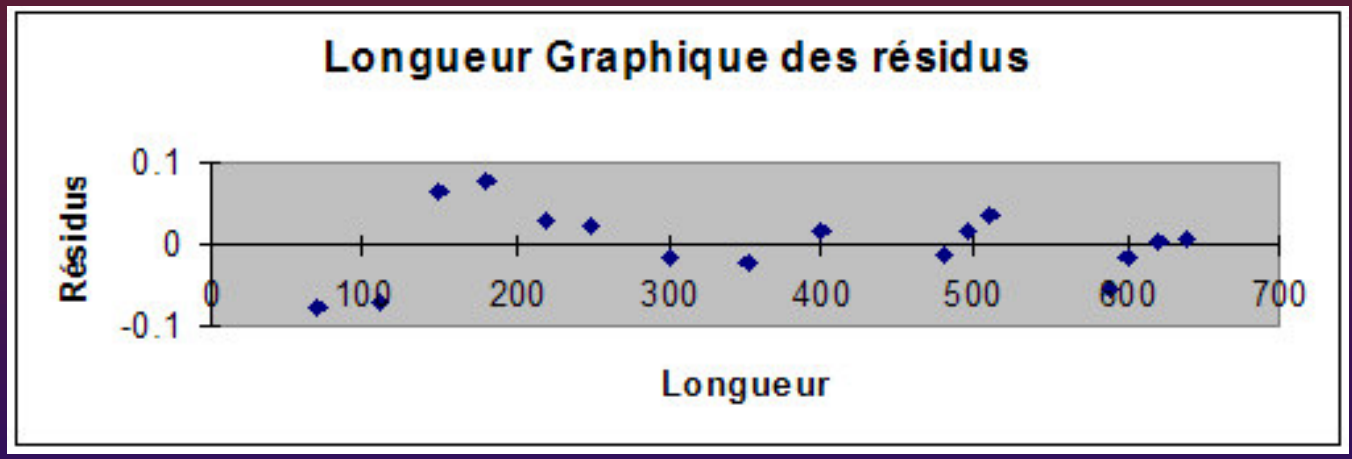

$R^2 = 0.867$

# Avec le logarithme du poids

| Log (Poids) | Longueur |
|---|---|
| -0.85387196 | 70 |
| -0.74472749 | 110 |
| -0.50863831 | 150 |
| -0.4202164 | 180 |
| -0.36653154 | 219 |
| -0.30103 | 248 |
| -0.20760831 | 300 |
| -0.08092191 | 352 |
| 0.07918125 | 400 |
| 0.25527251 | 480 |
| 0.32221929 | 495 |
| 0.38021124 | 510 |
| 0.49136169 | 590 |
| 0.5563025 | 600 |
| 0.62634037 | 620 |
| 0.68124124 | 640 |

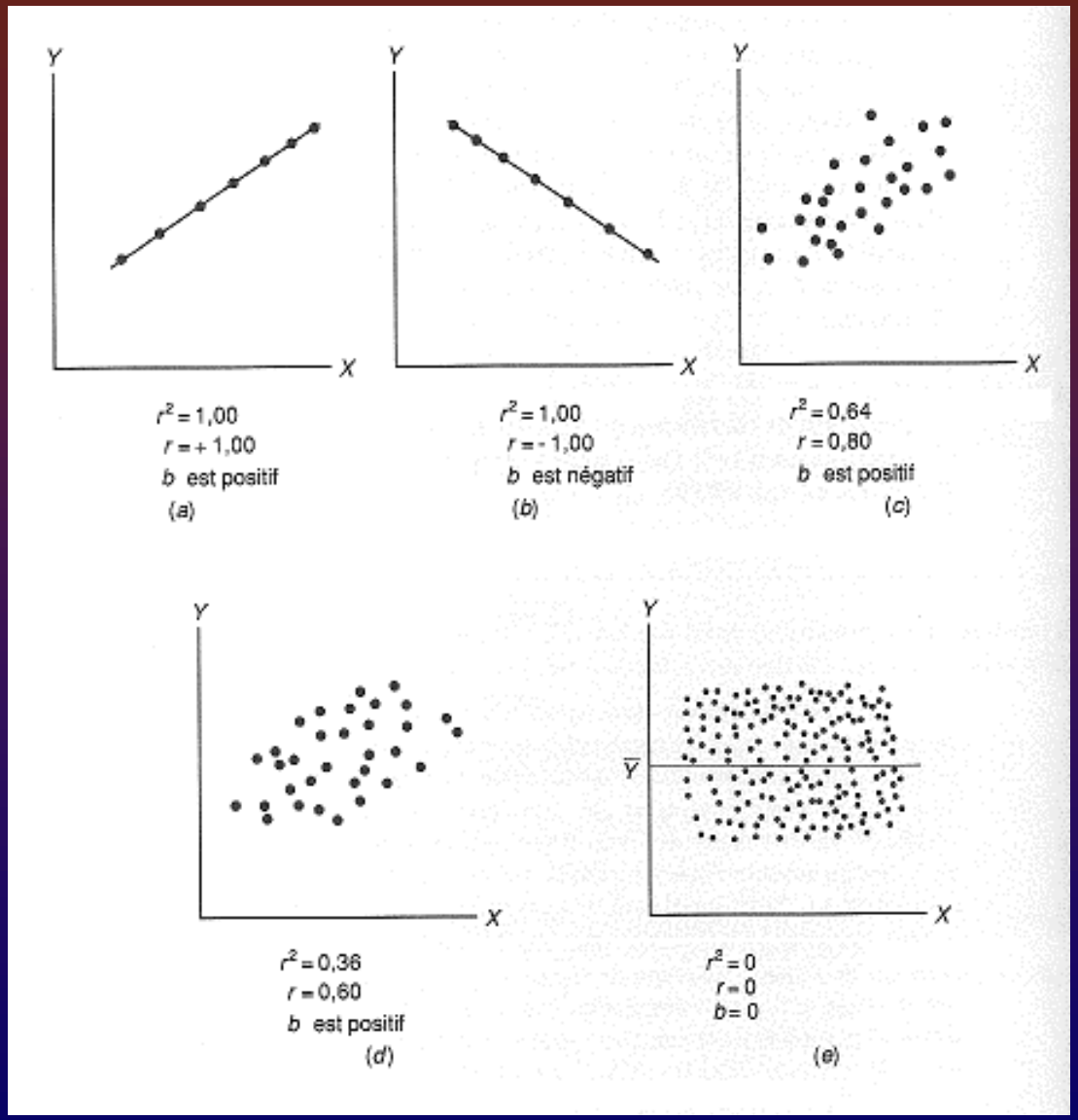


# Les nouveaux résidus




$R^2 = 0.992$

# Exemples graphiques



# La régression multiple

$$y = a + b_1x_1 + b_2x_2 + \ldots + b_ix_i$$

*Avec 1 variable dépendante y et p variables explicatives $x_i$*

Test F avec DDL1 = p et DDL2 = n-p-1

# Relation entre effectifs et nombre de variables

Effectifs constant → Le nombre de variables augmente → $R^2$ augmente et F tend à diminuer

# Interprétation des coefficients

- Dangereuse à cause des corrélations
- Utiliser de préférence des analyse factorielles

# La régression multivariables

$$\alpha + \beta_1 y_1 + \beta_2 y_2 + \ldots + \beta_j y_j = a + b_1 x_1 + b_2 x_2 + \ldots + b_i x_i$$

*Avec plusieurs variable dépendante $y_i$*
*et plusieurs variables explicatives $x_i$*