# Statistiques pour sciences sociales : applications

## 7 - Régréssion linéaire

Alexis Gabadinho

Université de Genève

Département des sciences économiques

Printemps 2011






## plan







## introduction

Par **régression** on entend la prédiction d'une variable en fonction de la connaissance d'une (ou plusieurs) autre(s) variable(s) $\Leftrightarrow$ on étudie la dépendance statistique d'une grandeur par rapport à une ou plusieurs autres

La régression est **linéaire** lorsque la relation entre la variable dépendante et la variable indépendante est linéaire

Par exemple :

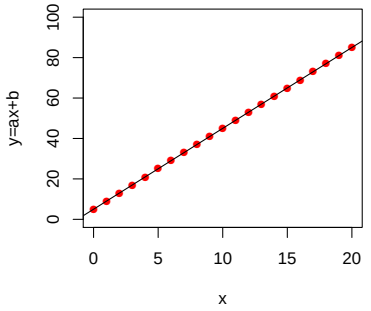
- taille des individus en fonction de celle de leurs parents
- taux de réussite scolaire en fonction des dépenses d'éducation
- revenu en fonction de l'âge et du nombre d'années d'études
- dépenses de consommation en fonction du revenu





## Dépendance parfaite

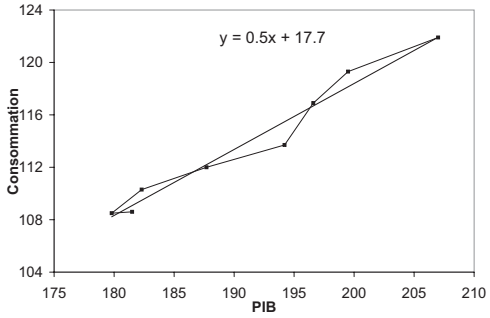
- Si la dépendance est parfaite (et la relation linéaire), les points sont alignés sur une droite
- On peut prédire parfaitement la valeur de $y$ si on connaît la valeur de $x$: $y = a \cdot x + b$



## exemple

Dépenses de consommation ($C$) en fonction du produit intérieur brut ($PIB$), en Suisse, en mia. de francs (1980)

| | $x_i$ | $y_i$ |
|---|---|---|
| $t$ | $PIB_t$ | $C_t$ |
| 1981 | 181.5 | 108.6 |
| 1982 | 179.8 | 108.5 |
| 1983 | 182.3 | 110.3 |
| 1984 | 187.7 | 112.0 |
| 1985 | 194.2 | 113.7 |
| 1986 | 196.6 | 116.9 |
| 1987 | 199.5 | 119.3 |
| 1988 | 207.0 | 121.9 |

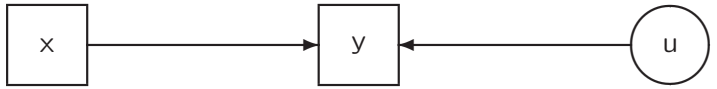



$\Rightarrow$ approximer la relation par une droite





## relation de dépendance




$$y = f(x) + u = a + bx + u$$

$$\hat{y} = f(x) = a + bx$$

- $y$ variable dépendante, expliquée, endogène, réponse
- $x$ variable indépendante, explicative, exogène, prédicteur
- $\hat{y}$ valeur prédite par le modèle
- $u$ écart résiduel aléatoire ($residual$) $= y - \hat{y}$, $(E(u) = 0)$
- $a$ constante du modèle ($intercept$)
- $b$ coefficient de régression ($slope$)





## exemple

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.9841 |
| R Square | 0.9685 |
| Adjusted R Square | 0.9632 |
| Standard Error | 0.9618 |
| Observations | 8 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Sig. F* |
|---|---|---|---|---|---|
| Regression | 1 | 170.47 | 170.47 | 184.30 | 0.000 |
| Residual | 6 | 5.55 | 0.92 | | |
| Total | 7 | 176.02 | | | |

| | *Coeff.* | *Std Error* | *t Stat* | *P-value* |
|---|---|---|---|---|
| Intercept | 17.67 | 7.10 | 2.49 | 0.047 |
| PIB | 0.50 | 0.04 | 13.58 | 0.000 |

But : savoir lire les informations fournies par un logiciel (Excel)



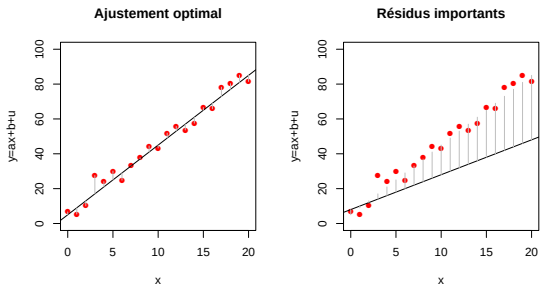


## estimation de la relation

$$y_i = a + bx_i + u_i \quad i = 1, ..., n$$

estimer la relation $\Rightarrow$ estimer les paramètres $a$ et $b$

estimer $a$ et $b$ de façon que la droite s'ajuste le mieux aux données, que les résidus $u_i$ (écarts entre $y_i$ et $a + bx$) globalement petits



## Estimation

**Estimateurs des moindres carrés**
(least squares, kleinste Fehlerquadrate)

$$\min_{a,b} s(a,b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

$$\hat{b} = \frac{\text{cov}(x,y)}{\text{var}(x)} = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}$$

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

($\hat{a}$ t.q. la droite des moindres carrés passe par le point moyen)



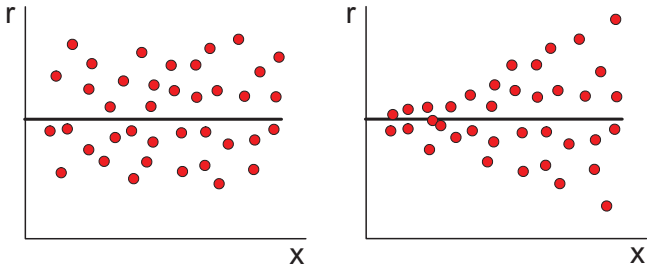


## hypothèses sous-jacentes du modèle linéaire

**H1** : relation linéaire entre $x$ et $y$, $a$ et $b$ identiques pour tout $i$

**H2** : $E(u_i) = 0 \rightarrow$ résidus en moyenne $= 0$

**H3** : $\text{var}(u_i) = \sigma_u^2$ pour tout $i \rightarrow$ var constante (homoscédasticité)



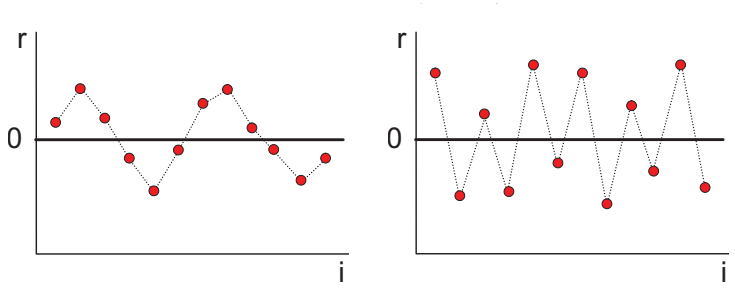

homoscédasticité — hétéroscédasticité





## hypothèses

**H4** : $\text{cov}(u_i, u_j) = 0$ pour tout $i \neq j \rightarrow$ résidus ne pas autocorrélés




autocorrélation positive — autocorrélation négative

**H5** : $\text{cov}(x_i, u_i) = 0$ facteurs et résidus ne pas corrélés

**H6** : $u_i \sim N(0; \sigma^2)$ résidus distribués normalement

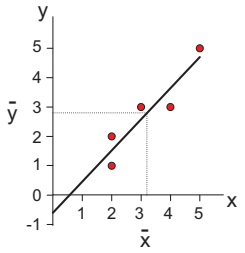




## calcul

**Exemple de calcul de $\hat{a}$ et $\hat{b}$**

| $i$ | $x_i$ | $y_i$ |
|---|---|---|
| 1 | 2 | 1 |
| 2 | 2 | 2 |
| 3 | 3 | 3 |
| 4 | 4 | 3 |
| 5 | 5 | 5 |
| | $\overline{x} = 3.2$ | $\overline{y} = 2.8$ |
| | $\text{var}(x) = 1.36$ | $\text{var}(y) = 1.76$ |
| | $\text{cov}(x,y) = 1.44$ | |




$$\hat{b} = \frac{1.44}{1.36} = 1.059$$

$$\hat{a} = 2.8 - 1.059 \cdot 3.2 = -0.588$$

$$\hat{y} = -0.588 + 1.059x$$

Prédictions :

$x = 2 \Rightarrow \hat{y} = -0.59 + 1.06 \cdot 2 = 1.53$
$x = 5 \Rightarrow \hat{y} = -0.59 + 1.06 \cdot 5 = 4.71$

## calcul

⇒ dans Excel, on peut utiliser

- la fonction =PENTE($Y_i : Y_n; X_i : X_n$) pour le coefficient $\hat{b}$
- la fonction =ORDONNE.ORIGINE($Y_i : Y_n; X_i : X_n$) pour $\hat{a}$
- l'option "courbe de tendance" dans l'interface graphique
- l'outil "régression" dans la macro *analyse de données* (→ plus tard)

Deux exemples :

1. exemple précédent
2. poids = f(taille) (données du questionnaire)





## évaluation du modèle

si var($x$) $\neq 0$ ($x$ non constant) ⇒ droite m.c. existe

existence solution $\neq$ pertinence de la solution

⇒ il faut évaluer la **fiabilité** des résultats :

1. **Qualité globale de l'ajustement :**
   - erreur standard de régression
   - coefficient de détermination $R^2$
   - Statistique $F$
2. **Test (Student) de significativité des paramètres :**
   - $H_0 : a = 0$ et $H_0 : b = 0$
3. **Analyse des résidus :**
   - remise en cause des hypothèses sur $u$
   - détection de données atypiques





## calcul des résidus

**Résidus : exemple de calcul**

Résidu : $r_i (= \hat{u}_i) = y_i - \hat{y}_i$, avec $\hat{y}_i = \hat{a} + \hat{b}x_i$
Droite des moindres carrés (m.c.) $\hat{y} = -0.588 + 1.059x$

| $i$ | $x_i$ | $y_i$ | $\hat{y}_i$ | $r_i$ | $r_i^2$ |
|---|---|---|---|---|---|
| 1 | 2 | 1 | 1.529 | -0.529 | 0.280 |
| 2 | 2 | 2 | 1.529 | 0.471 | 0.221 |
| 3 | 3 | 3 | 2.588 | 0.412 | 0.170 |
| 4 | 4 | 3 | 3.647 | -0.647 | 0.419 |
| 5 | 5 | 5 | 4.706 | 0.294 | 0.087 |
| total | 16 | 14 | 14 | 0 | 1.175 |
| moyenne | 3.2 | 2.8 | 2.8 | 0 | 0.235 |
| variance | 1.36 | 1.76 | 1.525 | 0.235 | |
| écart type | 1.167 | 1.327 | 1.235 | 0.485 | |

⇒ var($y$) = var($\hat{y}$) + var($r$)





## erreur standard

**Erreur standard de régresssion**

= estimation $\hat{\sigma}_u$ de l'écart type du terme d'erreur $u$
(mesure de la dispersion autour de la droite de régresssion)

**Estimation** : estimateur non biaisé de $\sigma_u^2$ (régression simple)

$$\hat{\sigma}_u^2 = \frac{1}{n-2}\sum_{i=1}^{n} r_i^2 \quad \text{avec} \quad \hat{u}_i = r_i = (y_i - \hat{y}_i)$$

Somme des carrés des résidus divisée par $(n-2)$, soit $n$ moins un degré de liberté par paramètre estimé ($a$ et $b$); exemple :

$$\sum_i r_i^2 = 5 \cdot 0.235 = 1.175$$

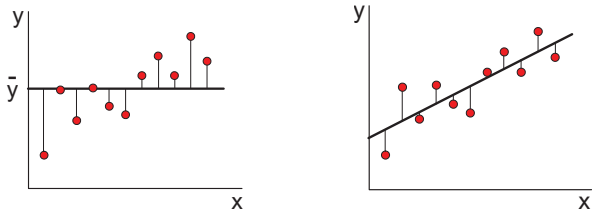
$$\hat{\sigma}_u^2 = \frac{1.175}{3} = 0.3917$$

$\hat{\sigma}_u = \sqrt{0.3917} = 0.626$ erreur standard de la régression

## Coefficient de détermination

**Coefficient de détermination** $R^2$ (Corrélation multiple)

mesure la part de la variation de $y$ reproduite par la droite estimée




| | |
|---|---|
| variation de $\hat{y}$ | $\text{var}(\hat{y}) = \frac{1}{n}\sum_i (\hat{y}_i - \bar{y})^2$ |
| variation résiduelle | $\text{var}(r) = \frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$ |
| variation de $y$ | $\text{var}(y) = \frac{1}{n}\sum_i (y_i - \bar{y})^2$ |

$$R^2 = \frac{\text{var}(\hat{y})}{\text{var}(y)} = 1 - \frac{\text{var}(r)}{\text{var}(y)}$$





## coefficient de détermination

**Autre interprétation de** $R^2$

On peut montrer que

$$R^2 = \frac{\text{var}(\hat{y})}{\text{var}(y)} = \frac{\text{cov}^2(\hat{y}, y)}{\text{var}(y)\,\text{var}(\hat{y})} = \boxed{\text{corr}^2(y, \hat{y})}$$

$\Rightarrow R^2$ : Carré de la corrélation multiple entre $y$ et les facteurs explicatifs déterminant $\hat{y}$

pour la régression simple, on a $\hat{y} = \hat{a} + \hat{b}x$, d'où

$$R^2 = \text{corr}^2(y, \hat{y}) = \text{corr}^2(y, \hat{a} + \hat{b}x) = \boxed{\text{corr}^2(y, x)}$$





## calcul de $R^2$

**Exemple de calcul de** $R^2$

| $i$ | $x_i$ | $y_i$ | $\hat{y}_i$ | $r_i$ |
|---|---|---|---|---|
| 1 | 2 | 1 | 1.529 | -0.529 |
| 2 | 2 | 2 | 1.529 | 0.471 |
| 3 | 3 | 3 | 2.588 | 0.412 |
| 4 | 4 | 3 | 3.647 | -0.647 |
| 5 | 5 | 5 | 4.706 | 0.294 |
| variance | 1.36 | 1.76 | 1.525 | 0.235 |
| $\text{cov}(x, y) = 1.44$ | | $\text{cov}(y, \hat{y}) = 1.525$ | | |

$$\begin{aligned} R^2 &= \frac{\text{var}(\hat{y})}{\text{var}(y)} = \frac{1.525}{1.76} = 0.87 \\ &= 1 - \frac{\text{var}(r)}{\text{var}(y)} = 1 - \frac{0.235}{1.76} = 1 - 0.13 = 0.87 \\ &= \text{corr}^2(y, x) = \frac{(1.44)^2}{1.76 \cdot 1.36} = 0.87 \\ &= \text{corr}^2(y, \hat{y}) = \frac{(1.525)^2}{1.76 \cdot 1.525} = 0.87 \end{aligned}$$





## Anova

**Statistique** $F$ **(Anova), en régression simple**

Mesure la part de variation expliquée par rapport à la variation résiduelle. Pour une régression simple

$$F = \frac{n\,\text{var}(\hat{y})}{n\,\text{var}(r)/(n-2)} = \frac{n\,\text{var}(\hat{y})}{\hat{\sigma}_u^2}$$

si $F$ est significatif ($p$-valeur $< 5$ %), la régression fait significativement mieux qu'une simple prédiction de $y$ par $\bar{y}$

(Sous les hypothèses H1 à H3 et la normalité de $y$, $F$ suit une distribution de Fisher-Snedecor non discutée dans ce cours)

Exemple : $F = \frac{5 \cdot 1.525}{0.3917} = 19.44$
degré de signification $= 2.1$ % (donné par table ou logiciel)

$\Rightarrow$ le modèle fait mieux que la moyenne $\bar{y}$

## significativité des paramètres

**Test de significativité des paramètres**

1) $x$ influence-t-il significativement $y$ ?

- la consommation $C$ varie-t-elle significativement avec le PIB ?
- le poids est-il significativement lié à la taille ?

Pratiquement, on teste la significativité du coefficient $b$

$$H_0 : b = 0 \text{ contre } H_1 : b \neq 0$$
**effet significatif si rejet de** $b = 0$

2) $y$ est-il proportionnel à $x (y = bx)$ ?

- consommation $C$ proportionnelle au PIB ?
- poids proportionnel à la taille ?

Pratiquement, on teste la significativité de la constante $a$

$$H_0 : a = 0 \text{ contre } H_1 : a \neq 0$$
**proportionnalité si non rejet de** $a = 0$





## significativité des paramètres

**Comment faire les tests?**

- construire le rapport

$$t = \frac{\text{coefficient}}{\text{erreur standard du coefficient}} = \frac{\hat{b}}{\hat{\sigma}_{\hat{b}}} \text{ ou } \frac{\hat{a}}{\hat{\sigma}_{\hat{a}}}$$

- sous les hypothèses H1 à H3, et la normalité de $y$, ce rapport est, lorsque $b = 0$, la réalisation d'un $T$ de Student à $(n-2)$ degrés de liberté
- ainsi, $b$ significatif
  - si $p\left(|T_{(n-2)}| > |t|\right) < 0.05$ (test bilatéral)
  - ou, de façon équivalente, si $t > t_{1-\alpha/2}^{(n-2)}$
- Règle sommaire pour $n$ grand : $b$ significatif si $t > 2$ $(\simeq z_{1-\alpha/2} = 1.96)$





## significativité des paramètres

**Erreur standard des estimateurs**

En remplaçant $\sigma_u^2$ par $\hat{\sigma}_u^2$ dans les formules des variances $\sigma_{\hat{a}}^2$ et $\sigma_{\hat{b}}^2$

$$\widehat{\text{var}}(\hat{b}) = \hat{\sigma}_{\hat{b}}^2 = \hat{\sigma}_u^2 \frac{1}{n\,\text{var}(x)}$$

$$\widehat{\text{var}}(\hat{a}) = \hat{\sigma}_{\hat{a}}^2 = \hat{\sigma}_u^2 \frac{1}{n}\left(1 + \frac{\overline{x}^2}{\text{var}(x)}\right)$$

$\hat{\sigma}_{\hat{a}}$ erreur standard de $\hat{a}$ $\quad$ $\hat{\sigma}_{\hat{b}}$ erreur standard de $\hat{b}$

Ex. : $n = 5$, $\text{var}(x) = 1.36$, $n\,\text{var}(x) = 5 \cdot 1.36 = 6.8$,
$\overline{x}^2 = (3.2)^2 = 10.24$

$\hat{\sigma}_{\hat{b}}^2 = \frac{0.3917}{6.8} = 0.0058 \quad \Rightarrow \quad \hat{\sigma}_{\hat{b}} = \sqrt{0.058} = \boxed{0.24}$

$\hat{\sigma}_{\hat{a}}^2 = \frac{0.3917}{5}\left(1 + \frac{10.24}{1.36}\right) = 0.668 \quad \Rightarrow \quad \hat{\sigma}_{\hat{a}} = \sqrt{0.668} = \boxed{0.818}$





## significativité des paramètres

**Significativité des coefficients : exemple**

Pour nos données on a :

| | estimation | erreur stand. | $t_{calc}$ | significativité |
|---|---|---|---|---|
| constante $a$ | -0.588 | 0.818 | -0.719 | 0.524 |
| pente $b$ | 1.059 | 0.240 | 4.409 | 0.022 |

$\Rightarrow$ pente $b$ statistiquement significative (effet de $x$ significatif)

$\Rightarrow$ constante $a$ non significative $\Rightarrow$ $y$ proportionnel à $x$

## analyse des résidus

**Analyse des résidus**

- Détection de données atypiques
- Vérification des hypothèses sur les erreurs $u_i$

**1) Détection de données atypiques**

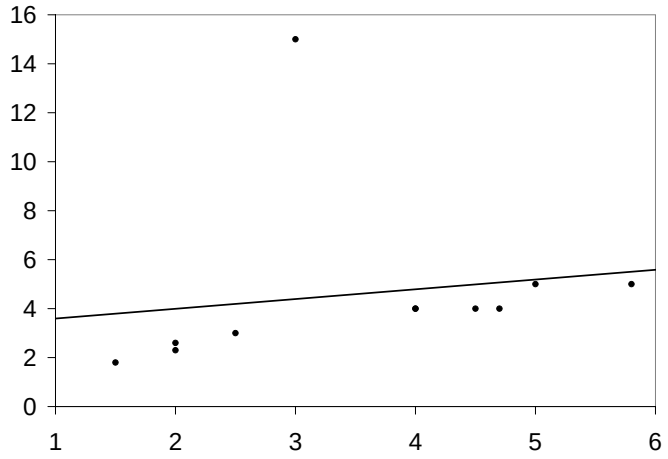
Donnée $y_i$ atypique si mal ajustée par la droite des m.c.

$\Rightarrow$ grand résidu $r_i = \hat{u}_i = y_i - \hat{y}_i$





## analyse des résidus








## analyse des résidus

Résidu grand relativement à l'erreur standard de régression $\Rightarrow$ examiner **résidus standardisés** :

$$r_i^s = \frac{r_i}{\hat{\sigma}_u}$$

Atypique si valeur absolue du résidu excède 2 ou 2.5 fois l'erreur standard (pour une distribution normale, resp. 5 % et 1 % des valeurs dépassent 2 et 2.5 fois l'écart type)

$$\begin{aligned} |r_i| > 2.5\hat{\sigma}_u &\quad \Rightarrow \quad (x_i, y_i) \text{ atypique} \\ |r_i^s| > 2.5 &\quad \Rightarrow \quad (x_i, y_i) \text{ atypique} \end{aligned}$$

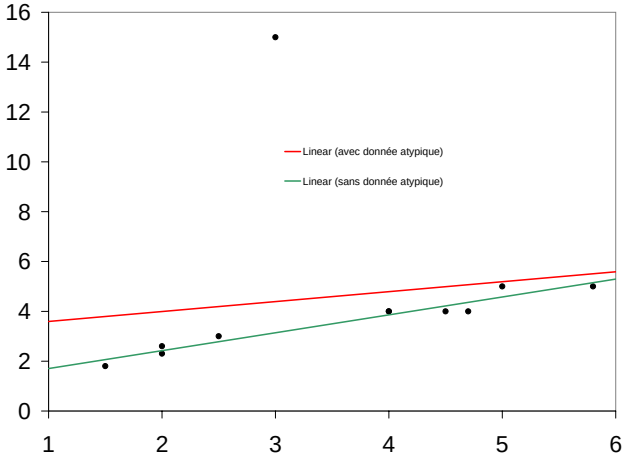
Pour l'exemple graphique,

- erreur standard $\hat{\sigma}_u = 3.75$
- résidu donnée du haut $= 10.6$, soit 2.8 fois l'erreur standard
- la donnée du haut est donc atypique





## analyse des résidus : effets de données atypiques



## analyse des résidus : autre piège à éviter

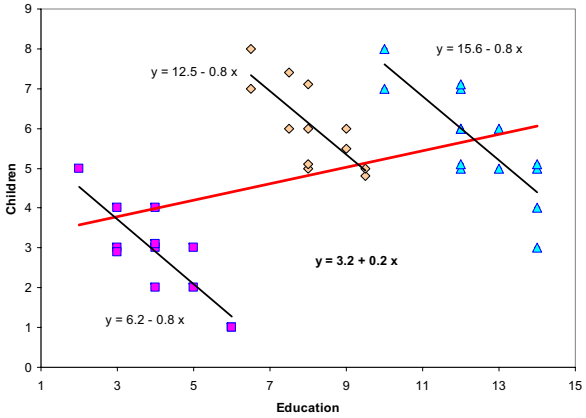



Figure 1: Multi-level: A simple example with 3 clusters





## analyse des résidus

### Vérification des hypothèses

- H2 : $E(u_i) = 0, \quad i = 1, ..., n$
- H3 : $E(u_i^2) = \sigma_u^2, \quad i = 1, ..., n \Rightarrow$ homoscédasticité
- H4 : $E(u_i u_j) = 0 \quad i = 1, ..., n \Rightarrow$ indépendance des $u_i$

Lorsque les hypothèses ne sont pas satisfaites, la validité des résultats inférentiels (erreurs standard, significativité des coefficients, $F$, ...) n'est plus assurée !!

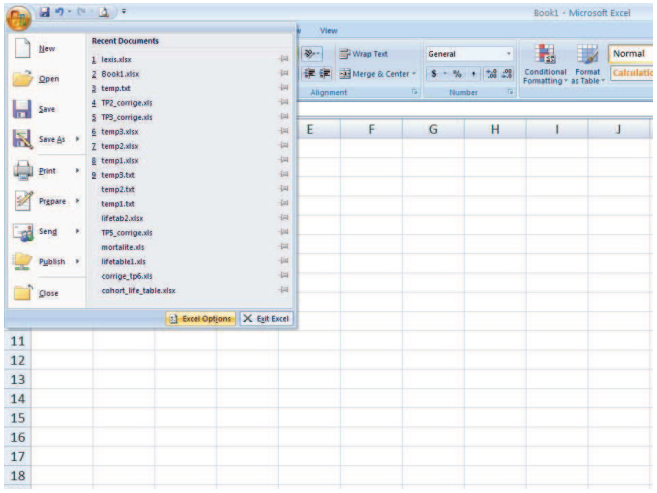
### Vérifications graphiques

- diagramme des points $(x_i, r_i)$
- diagramme des points $(i, r_i)$ (pour données ordonnées, séries temporelles)





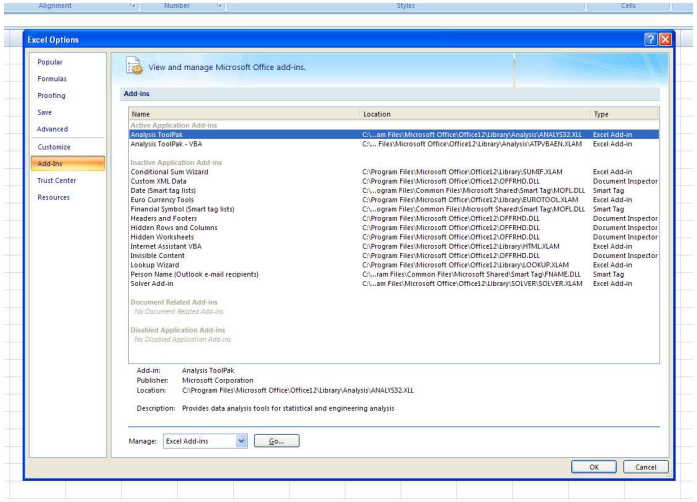
## Excel : outils d'analyse "régression"



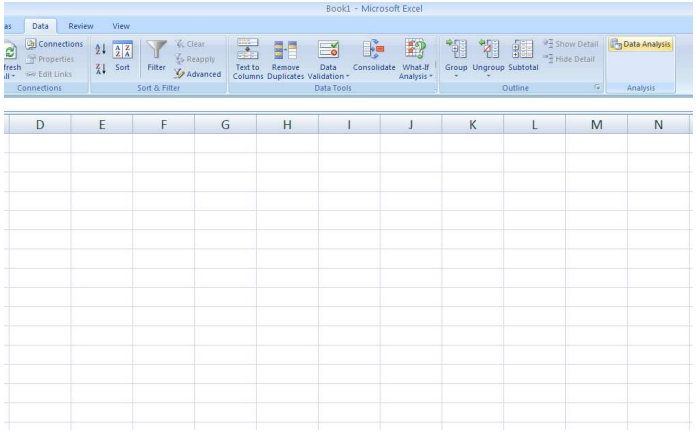





## Excel



## Excel








## Excel

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.9308 |
| R Square | 0.8663 |
| Adjusted R Square | 0.8217 |
| Standard Error | 0.6262 |
| Observations | 5 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 7.6235 | 7.6235 | 19.44 | 0.0216 |
| Residual | 3 | 1.1765 | 0.3922 | | |
| Total | 4 | 8.8 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | -0.5882 | 0.8179 | -0.7192 | 0.5240 | -3.1912 | 2.0147 |
| X Variable 1 | 1.0588 | 0.2401 | 4.4091 | 0.0216 | 0.2946 | 1.8231 |

RESIDUAL OUTPUT

| *Observation* | *Predicted Y* | *Residuals* | *Standard Residuals* |
|---|---|---|---|
| 1 | 1.5294 | -0.5294 | -0.9762 |
| 2 | 1.5294 | 0.4706 | 0.8677 |
| 3 | 2.5882 | 0.4118 | 0.7593 |
| 4 | 3.6471 | -0.6471 | -1.1931 |
| 5 | 4.7059 | 0.2941 | 0.5423 |





## Excel

**Output du modèle** $poids_i = a + b \cdot taille_i + u_i$

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.7795 |
| R Square | 0.6077 |
| Adjusted R Square | 0.6018 |
| Standard Error | 6.3878 |
| Observations | 68 |

ANOVA

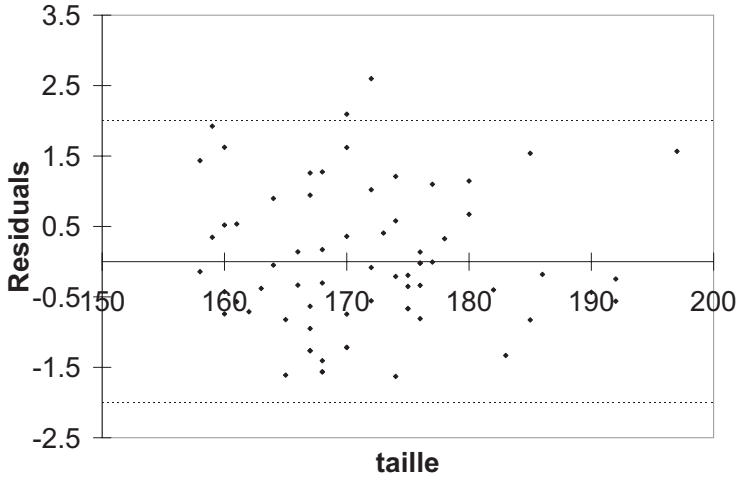
| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 4171.60 | 4171.60 | 102.24 | 0.000 |
| Residual | 66 | 2693.03 | 40.80 | | |
| Total | 67 | 6864.63 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | -91.56 | 15.313 | -5.979 | 0.000 | -122.14 | -60.99 |
| X Variable 1 | 0.902 | 0.089 | 10.111 | 0.000 | 0.72 | 1.08 |





## exemple : graphique des résidus standardisés



## exemple

**Commentaire de ce modèle**

L'ajustement du modèle est globalement satisfaisant

- $R^2 = 0.61 \Rightarrow$ 61 % de la variance de *poids* est reproduite
- l'erreur standard $\hat{\sigma}_{\hat{u}}$ est de 6.4 kg, valeur clairement inférieure à l'écart type de *poids* ($\hat{\sigma}_y = 10.05$) mais reste considérable
- l'explication apportée par le modèle est clairement significative ($p-$valeur de $F < 0.05$), le coefficient de régression $\hat{b}$ significativement différent de 0
- on observe deux grands résidus (2 résidus standardisés $> 2$) $\Rightarrow$ on pourrait exclure ces deux observations et réajuster le modèle





## régression multiple

Extension de la régression linéaire au cas où l'on a plusieurs prédicteurs :

- salaire en fonction du niveau de formation, des années d'expérience et du sexe
- poids en fonction de la taille, de l'âge et du sexe etc...

Formellement, le modèle devient :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + u_i$$

$$\text{pour } i = 1, 2, ..., n$$

pour $n$ observerations et $k = p - 1$ variables indépendantes





## Estimation des paramètres

- sous forme matricielle, le modéle devient

$$Y_{n\times 1} = X_{n\times(k+1)}\beta_{(k+1)\times n} + u_{n\times 1}$$

- Estimation des paramètres $\beta_j$ (vecteur $\hat{\beta}$)
- $k$ variables explicatives $X_1, X_2, ..., X_k$

$$\hat{\beta} = \min_{\beta_0,...,\beta_k} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{i=1}^{k} \beta_j x_{ij})$$





## différence avec régression simple

- **Interprétation :** chaque coefficient mesure l'effet d'une variable explicative **toutes choses égales par ailleurs** (*ceteris paribus*), càd en supposant fixes (en contrôlant) les valeurs des autres prédicteurs. Exemple : évaluer l'effet discriminatoire du sexe en contrôlant pour l'effet de la formation et de l'expérience professionnelle
- **Erreur standard de régression :** s'obtient en divisant la somme des carrés des résidus par $n - p$

$$\hat{\sigma}_u = \sqrt{\frac{\sum_i r_i^2}{n - p}}$$

- **Test de significativité des coefficients :** le ratio $t$ doit être comparé avec une distribution de Student à $(n - p)$ degrés de liberté
- **Statistique $F$ :** la définition générale est

$$F = \frac{n\,\text{var}(\hat{y})/(p-1)}{\hat{\sigma}_u^2} \quad \text{sous } H_0 \;\; F \sim F - S_{(p-1),(n-p)}$$

## exemple

**Output du modèle** $poids_i = \beta_0 + \beta_1 taille_i + \beta_2 age_i + \beta_3 sexe_i + u_i$

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.8025 |
| R Square | 0.6441 |
| Adjusted R Square | 0.6274 |
| Standard Error | 6.1788 |
| Observations | 68 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 3 | 4421.25 | 1473.75 | 38.60 | 0.000 |
| Residual | 64 | 2443.38 | 38.18 | | |
| Total | 67 | 6864.63 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | -62.308 | 21.520 | -2.895 | 0.005 | -105.300 | -19.316 |
| sexe | 5.365 | 2.098 | 2.557 | 0.013 | 1.174 | 9.557 |
| age | -0.165 | 0.559 | -0.295 | 0.769 | -1.282 | 0.952 |
| taille | 0.708 | 0.115 | 6.168 | 0.000 | 0.479 | 0.937 |





## résultats

- **Interprétation :** en moyenne et toutes choses égales par ailleurs
  - la taille à un effet positif sur le poids : pour chaque centimètre supplémentaire, le poids augmente de 0.71 kg
  - le sexe affecte également le poids : à taille et âge égal, un homme pèse en moyenne 5.4 kg de plus
  - l'effet de l'âge semble négatif (mais $H_0$ ne peut pas être rejeté)
- **Significativité des coefficients :**
  - les effets de la taille et du sexe sont statistiquement significatifs, l'effet de l'âge est non significatif
  - les ratios $t$ doivent être comparés avec une loi de Student à (68-4) degrés de liberté
  - le seuil est $t^{64}_{.975} = 1.998$





## résultats

- **ajustement :**
  - l'ajustement s'est légèrement amélioré suite à l'inclusion des variables *age* et *sexe*, $R^2 (= 0.64)$ a augmenté
  - l'erreur standard de régression a diminué ($\hat{\sigma}_{\hat{u}} = 6.2$ kg) mais reste grand
  - Le $F$ est clairement significatif (sig. $F \simeq 0$), le modèle apporte une information significative par rapport à la seule moyenne de *poids*