# La régression

*Quantifier en sociologie. Séance 10*

Joanie Cayouette

# Principe général

- L'effet d'une variable x sur une seconde variable – y – toutes choses égales par ailleurs
- Deux types de régression :

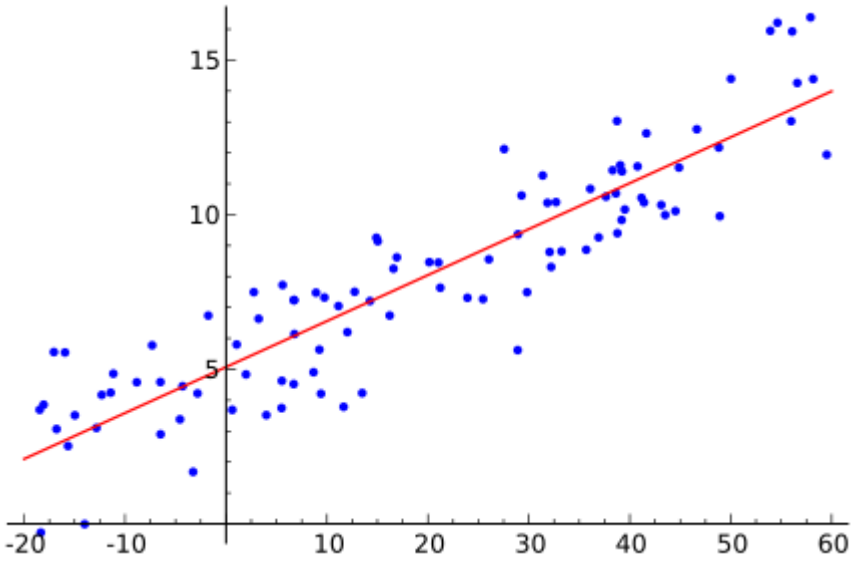
1) La régression linéaire : expliquer une variable quantitative à partir de variables quantitatives (ou qualitatives binaires)
    1) Régression linéaire simple : l'effet d'une variable sur une variable
    2) Régression linéaire multiple : généralisation avec n variables explicatives
2) La régression logistique : expliquer la probabilité d'un événement à partir de variables quantitatives ou qualitatives

# Principe général (historique)

- Galton (1886) : expliquer la taille des ascendants à partir de celle des descendants
- Cousin de Darwin
- Régression vers la moyenne (réduction des situations exceptionnelles).

# La régression linéaire simple

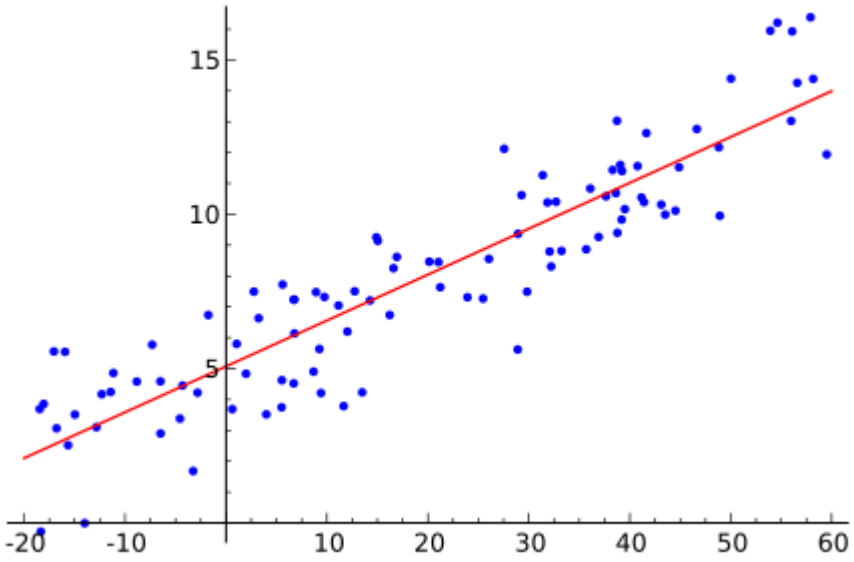
Variable à expliquer (le niveau de salaire à un âge donné par exemple) ou à prédire (le niveau de consommation à partir du PIB par exemple)



# La régression linéaire simple

Expliquer le niveau de salaire par l'âge

Équation : Salaire = a + b * âge + erreur (u)
On cherche a et b, en essayant de minimiser l'erreur, grâce à la méthode des moindres carrés ordinaires.

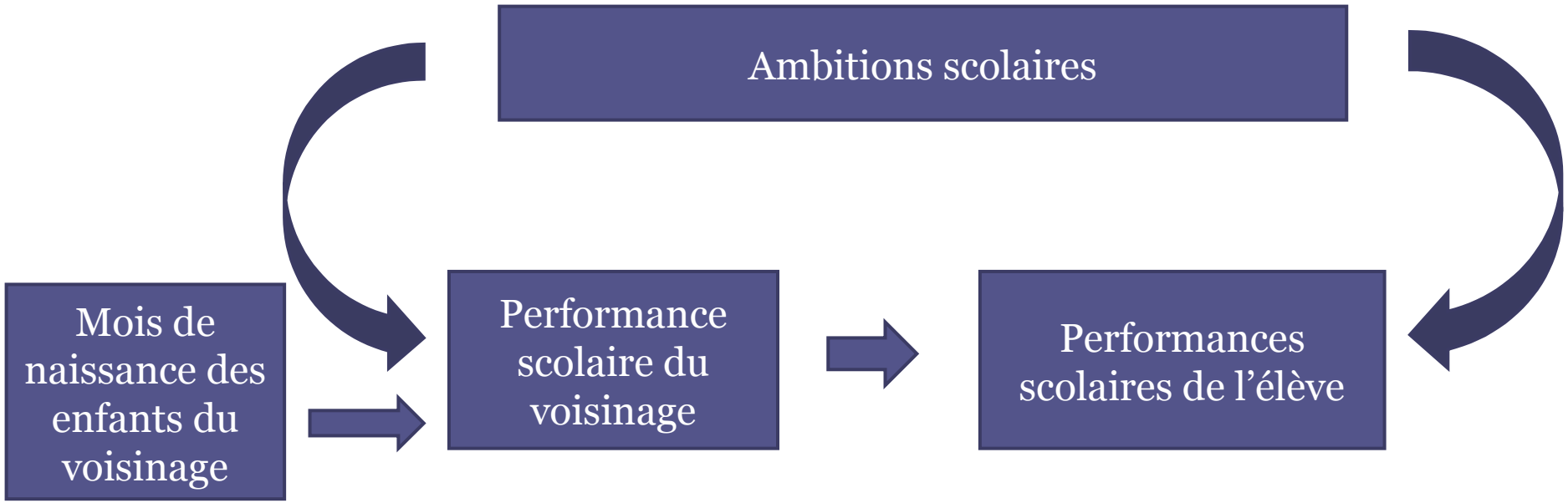


# La régression linéaire simple

Hypothèses :

1. Liaison linéaire
2. Endogénéité des variables (cf. Goux et Maurin)
3. L'erreur est aléatoire (difficile à assumer)
4. L'erreur est en moyenne indépendante de x (variable explicative)

# Effet d'endogénéité



# La régression linéaire simple

Que commenter dans une régression linéaire simple ?

1. On s'intéresse à a et b (surtout le signe de b).
    1. Dans notre exemple de base, a est une reconstruction théorique du salaire d'un individu à 0 ans.
    2. b est le gain (ou la perte) en salaire pour chaque année supplémentaire. On trouvera ici un b positif, ce qui signifie que le salaire augmente avec l'âge.
    2. La significativité de a et b.
       Un test de Student est réalisé sur chaque paramètre, estimant la probabilité que ce paramètre équivaut à 0. On détermine le seuil d'erreur accepté (comme pour le khi-deux).
    3. Le R2 appelé coefficient de détermination. C'est la part de la variance expliquée par le modèle.

# La régression logistique

Toute régression doit présenter :

1. La variable dichotomique à expliquer
2. La situation de référence
3. Le type de modèle utilisé (logit, probit…)
4. L'estimation des paramètres ou les odds ratios (rapports de chance)
5. La significativité des paramètres / odds ratios (rapports de chance).

# La régression logistique

Rappels sur les odds ratio : l'odds est la probabilité qu'un événement arrive par rapport à la probabilité qu'il n'arrive pas. L'odds ratio est le rapport des odds des deux situations

| CSP | Nés avant 1910 | Nés entre 1935 et 1940 |
|---|---|---|
| Professions libérales, cadres et personnels de direction | 37 % | 62 % |
| Ouvriers semi-qualifiés et non qualifiés | 1 % | 10 % |

$Odds_{cadres}$=37%/63%=0,59 $Odds_{ouvriers}$=1%/99%=0,01

Odds ratio= 0,59/0,01=59. Pour les individus non avant 1910, les cadres et assimilés avaient 59 fois plus de chances de fréquenter le secondaire long que les enfants d'ouvriers semi-qualifiés et non qualifiés.

# La régression logistique

Rappels sur les odds ratio : l'odds est la probabilité qu'un événement arrive par rapport à la probabilité qu'il n'arrive pas. L'odds ratio est le rapport des odds des deux situations

| CSP | Nés avant 1910 | Nés entre 1935 et 1940 |
|---|---|---|
| Professions libérales, cadres et personnels de direction | 37 % | 62 % |
| Ouvriers semi-qualifiés et non qualifiés | 1 % | 10 % |

$Odds_{cadres}$=37%/63%=0,59 $Odds_{ouvriers}$=1%/99%=0,01

Odds ratio= 0,59/0,01=59. Pour les individus non avant 1910, les cadres et assimilés avaient 59 fois plus de chances de fréquenter le secondaire long que les enfants d'ouvriers semi-qualifiés et non qualifiés.

# Du bon usage de la régression

- Il est abusif de croire que les régressions permettent donc d'« isoler » réellement l'effet d'un facteur, mais, comme l'indiquent Marion Selz et Florence Maillochon dans *Le raisonnement statistique*, elles mettent en évidence une différence qui mérite le plus souvent d'être mise en évidence et confortée par d'autres analyses.

- Il convient de bien connaître ses données et les relations entre les différentes variables incluses dans la régression .

- « La statistique n'explique rien, mais fournit des éléments potentiels pour l'explication. Aussi le terme de variable explicative ou variable à expliquer n'est sans doute pas le plus judicieux » (Lebart et Morineau, 2002, *Statistique exploratoire multidimensionnelle*, p. 209).