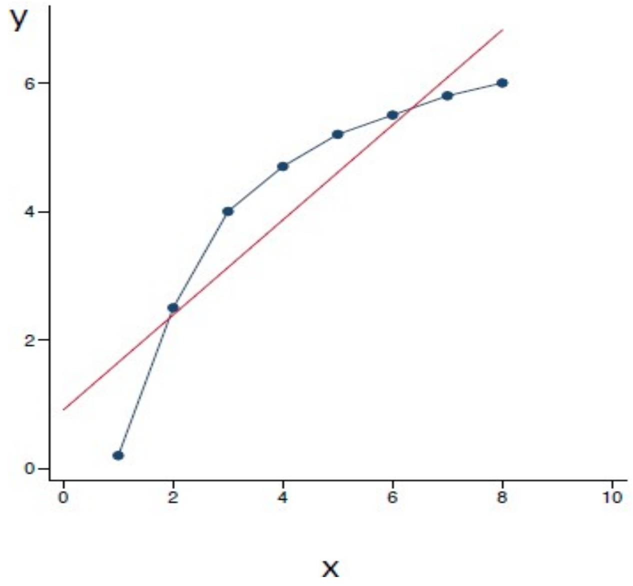
Les régressions linéaires
Quelques variantes

Les modèles log

- Tenir compte de la non linéarité
- La régression linéaire pas toujours adaptée à la nature de la relation qui lie x et y
  - cf. forme du nuage de points
  - Ex : Cas où l'impact de x sur y dépend de la valeur initiale de x
  - Ex: Envisager une relation linéaire entre le niveau d'éducation et le taux de salaire suppose que les rendements marginaux de l'éducation sont constants.
    - Ex: Quel que soit le niveau d'éducation initial, une année supplémentaire augmentera toujours le salaire de la même somme.
- Le raccourci de la relation linéaire peut amener à une mauvaise interprétation.

Les modèles log

- Tenir compte de la non linéarité :



Parfois, la relation entre y et x est non-linéaire

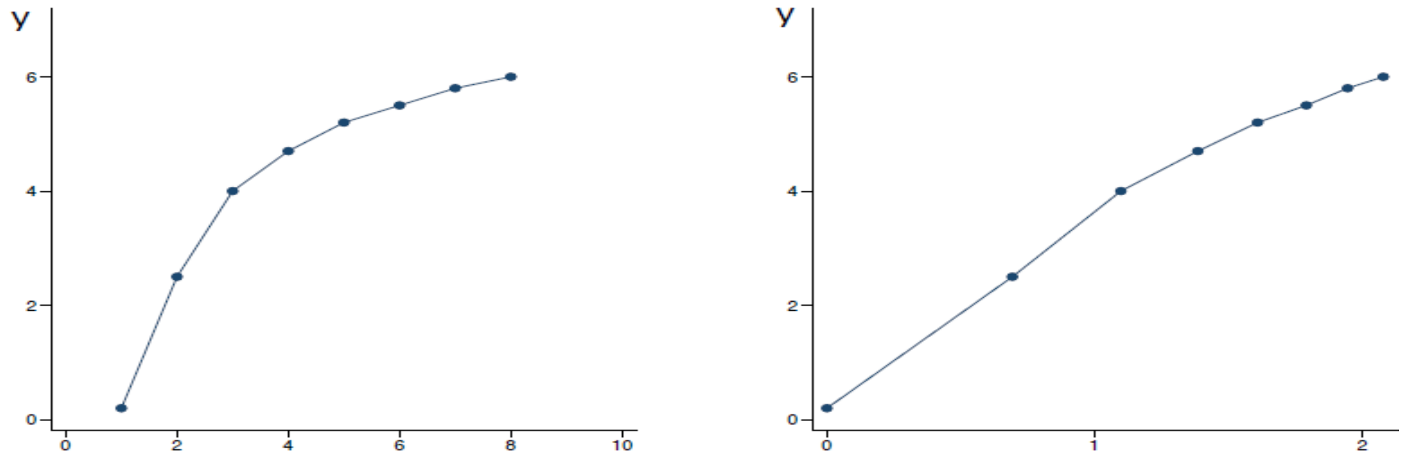
L'application des MCO génère
- un ajustement imparfait (SCR élevée)
- de l'hétéroscédasticité
- des prédictions fausses

On peut vouloir linéariser la relation entre y et x, en modifiant l'échelle de x ou de y (changement de variable)

Les modèles log

- Tenir compte de la non linéarité :

Les changements de variables usuels : logarithme naturel ou puissance



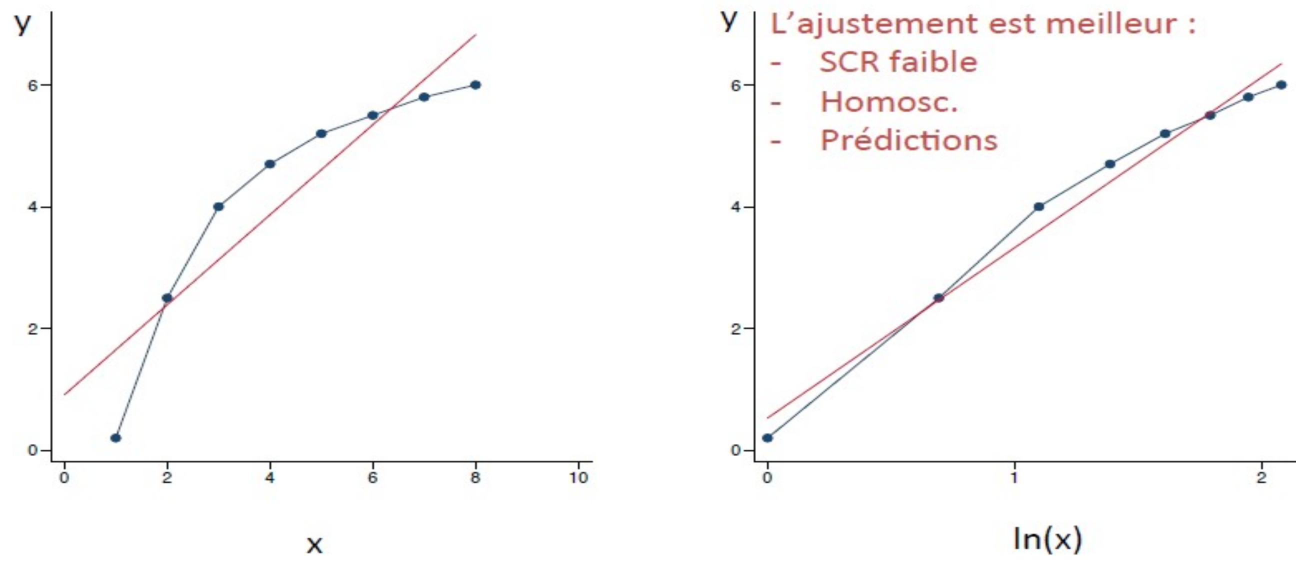
x ln(x)

Transformer x en « *ln* de x »

Les modèles log

- Tenir compte de la non linéarité

Les changements de variables usuels : logarithme naturel ou puissance



Les modèles log

- Linéariser une relation
- Ex : la théorie quantitative de la monnaie
  - Équation théorique : Mv=PQ
  - Dévient une fonction affine en passant par Ln ! Ln(PQ)=Ln(M)+Ln(v)
- Ex : la fonction Cobb-Douglas
  - Équation théorique : Y = A K^α.L^β
  - Dévient une fonction affine en passant par Ln !
  - $Ln\ (Y) = Ln\ (A) + \alpha Ln\ (K) + \beta Ln(L)$

Les modèles log

- Transformation Log
- Envisager les évolutions en termes de pourcentages avec une transformation logarithmique des y et/ou x.
- Quatre combinaisons de formes fonctionnelles selon que la variation soit sous sa forme en « niveau » ou en log.
  - Le modèle « niveau-niveau », sans transformation
  - Le modèle « log-niveau »
  - Le modèle « log-log »
  - Le modèle « niveau-log »
- Les modèles "log-niveau" et "log-log" sont les plus utilisés en économie.

- Rappel sur l’élasticité :
- Définition : l’élasticité de y par rapport à x est le pourcentage de variation de y lorsque x varie de 1%.

$$\varepsilon_{+/-} = \frac{\%\Delta y}{\%\Delta x} = \frac{\partial y / y}{\partial x / x}$$

Ex: les élasticités prix et revenu en microéconomie

Les modèles log

- Le modèle log-niveau ou semi-élasticité
- Le modèle s'écrit : $log\ (y) = \beta 0 + \beta 1 x + u$
- Interprétation du $\beta 1$
  - β: mesure le taux de variation de y lorsque x varie de 1 unité.
  - Lorsque x varie d'1 unité, y varie de (100. β:)%.
  - 100. β: semi-élasticité de y par rapport à x.
- Remarques : modèle souvent utilisé quand X est une variation de temps.

Les modèles log

- Le modèle log-log ou modèle à élasticité constante
- Le modèle s'écrit : $log\ (y) \quad = \beta 0 + \beta 1.\ \log(x) + u$
- Interprétation de β: :
  - $\beta 1$ mesure le taux de variation de y lorsque x varie de 1%.
  - Lorsque x varie d'1% , y varie de $\beta 1$ %.
- Remarques :
  - Très utilisé pour décrire de nombreux phénomènes.
  - Microéconomiques: Salaire=f(chiffre d'affaire) ou Ventes=f(dépenses en publicités)
  - Macroéconomiques : Densité médicale=f(population) ou PIB=f(impôts)

Les modèles log

- Exemples:
- Soit le modèle estimé:

$$\log(salaire) = 1.1157 + 0.0933.\ education$$

  - Le rendement de l'éducation: le salaire horaire augmente de 9.33% pour chaque année supplémentaire d'étude (ceteris paribus).
  - L'effet de l'éducation sur le salaire est constant mais ce n'est plus une variation en euros, c'est une variation en % du salaire
  - L'interprétation de la constante informe sur le log du salaire lorsque le niveau d'instruction est nul (pas très utile..)

Les modèles log

- Exemples:
- Soit le modèle estimé:

$$\log(demande) = 4{,}7 - 1{,}25.\ \log(prix)$$

  – L'élasticité-prix est de -1.25 : une augmentation de 10% du prix de vente engendre une baisse 12,5% de la quantité demandée.

- Soit le modèle estimé:

$$\log(consommation) = -6{,}6 + 1{,}27.\ \log(revenu)$$

  – L'élasticité-revenu est de +1.27 : une augmentation de 10% du revenu engendre une hausse 12,7% de la consommation.

Les modèles log

- Remarques
- Les coefficients ont une interprétation en termes de %
- On ignore les unités de mesure car une élasticité divise une variation de y par y, les deux ayant la même unité on se retrouve sans unité
- Il faut que l'interprétation en pourcentages ait un sens.
- Les logs sont:
  - Souvent utilisés pour des montants (euros, dollars, etc.)
  - Moins souvent utilisés pour des variables en années (âge, nombre d'années d'étude, expérience, espérance de vie, etc.).
  - Pas utilisés si la variable est déjà en pourcentages (taux de chômage, taux de réussite, etc.)
- Si des valeurs <= 0 sont possibles, la transformation log n'est pas possible.

Les modèles log

Type de modèle	Variable dépendante	Variable explicative	Interprétation du coefficient β_1
Niveau-Niveau	y	X	$\Delta y = \beta_1 \Delta x$
Niveau-Log	y	Log(x)	$\Delta y = \left(\frac{\beta_1}{100}\right) \% \, \Delta x$
Log-Niveau	Log(y)	x	$\% \, \Delta y = (100\beta_1)\Delta x$
Log-Log	Log(y)	Log(x)	$\% \, \Delta y = \beta_1 \% \, \Delta x$

Source : J. Wooldridge (2009), *Introductory Econometrics. A Modern Approach*, South Western, 4ème édition

Les modèles quadratiques

- Objectif : capturer des effets marginaux non constants (croissants ou décroissants).

- Y= $\beta 0$ + $\beta 1$x + $\beta 2$ x² +u

Interprétation :

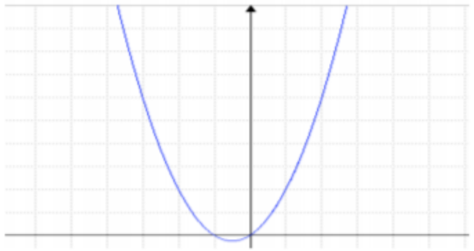
- Le coefficient $\beta 1$ mesure l'effet partiel à partir de 0, i.e. l'effet du passage de x=0 à x=1.
- Ensuite, l'effet marginal d'un gain d'1 unité de x se calcule de la manière suivante :

$$\frac{d\hat{y}}{dx} = \widehat{\beta_1} + 2.\widehat{\beta_2}.\mathrm{x}$$

Polynôme du 2nd degré :

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1}.x + \widehat{\beta_2}.x^2$$

- Si $\widehat{\beta_2} > 0$: l'extremum est un minimum



- Si $\widehat{\beta_2} < 0$: l'extremum est un maximum

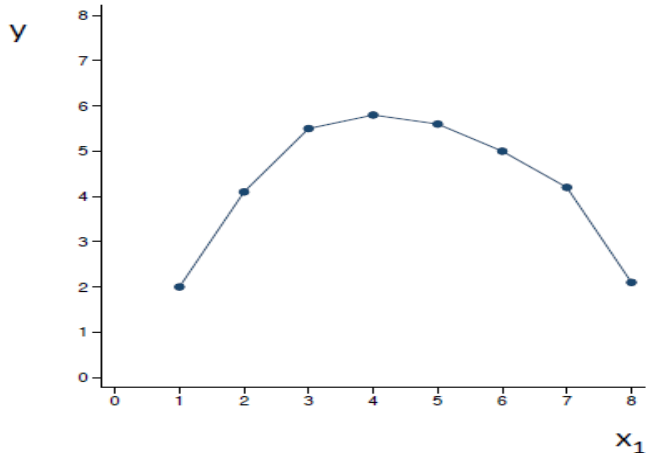
Les modèles quadratiques

- Pour des variables expliquées non-négatives:
  - $\beta 1$: est souvent positif et $\beta 2$; est négatif.
  - $\beta 2$; est souvent très petit (0.00...).
    - Changement de pente lorsque x varie.
    - Petits changements au fur et à mesure …
    - Ce changement de pente est néanmoins important pour des grandes valeurs de X.
- Forme parabolique de la courbe quadratique
  - il existe une valeur de x pour laquelle l'effet de x sur y est nul.

Les modèles quadratiques

- Courbe quadratique avec β 1>0 et β2 < 0.
  - Ex: courbe de Kuznet

- Avec un polynôme d'ordre 2…
  - On peut identifier un optimum (max. ou min.)



Les modèles quadratiques

- Courbe quadratique avec β 1>0 et β2 < 0.

Le **point de retournement de la courbe quadratique** est le maximum de la fonction : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$:

$$\frac{d\hat{y}}{dx} = \hat{\beta}_1 + 2.\hat{\beta}_2 x^* = 0 \rightarrow \boxed{x^* = \left|\frac{\hat{\beta}_1}{2\hat{\beta}_2}\right|}$$

Les modèles quadratiques

	(1) y	(2) y
x	-0.006	$\hat{\beta}_1$ $\boxed{2.753^{***}}$
	(-0.024)	(16.971)
xsq		$\hat{\beta}_2$ $\boxed{-0.307^{***}}$
		(-17.422)
_cons	4.314*	-0.284
	(3.386)	(-0.892)
N	8	8
R-sq	0.000	0.984
rss	16.047	0.260
rmse	1.635	0.228

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

L'ajustement quadratique est meilleur que l'ajustement linéaire.

On trouve l'optimum ainsi :
$x_1^* = -2{,}753 / (2x\ -0{,}307) = 4{,}5$

C'est un maximum, parce que la dérivée seconde d^2y/d^2x_1 est $\hat{\beta}_2$ qui est négatif !

Les modèles quadratiques

- Exemple : Salaire = $\beta 0$ + $\beta 1$expérience + $\beta 2$ expérience² +u
- Avec $\beta 1>0$ et $\beta 2 <0$ le salaire commence par augmenter avec l'expérience mais de moins en moins vite, puis il pourra potentiellement diminuer à partir d'un certain nombre d'année
- Il est possible d'ajouter des formes quadratiques d'ordre supérieur mais c'est rarement utile en pratique.

Les modèles quadratiques

- Conclusion/résumé :
- Une fonction quadratique permet de mesurer des effets marginaux croissants et décroissants.
- Un test t est utilisé pour voir si le terme quadratique est statiquement significatif.

Le modèle de régression linéaire avec information qualitative

Inclure une variable indicatrice

- ou binaire, dichotomique, dummy
- La variable prend la valeur de 0 ou de 1 selon la survenance ou non d’un évènement/caractère
- Conseil : nommer la variable de façon à indiquer l’événement/caractère associé à la valeur unitaire.
  - Ex : la variable indicatrice dénommée "femme" prend la valeur 1 si l’individu est une femme, 0 sinon.
- On inclut directement la dummy dans le modèle
  - La référence d’interprétation est 0

Inclure une variable indicatrice

- Cas d’une variable qualitative nominale
  - ex : statut matrimonial (célibataire/marié/veuf/divorcé)
  - ex : statut dans l’emploi (actif/chômeur/retraité)
- On introduit plusieurs indicatrices : une pour chaque modalité
- !!! on omet une des modalités, pour éviter la colinéarité parfaite (information redondante, "trappe à indicatrices", qui servira donc de référence, R le fait automatiquement pour les variables de type facteur)
  - On construit g-1 dummies si la variable possède g modalités.
  - Interprétation : la modalité omise devient la référence.

Inclure une variable indicatrice

- Combinaison de 2 variables binaires
- Combiner l’information de 2 variables binaires, pour créer une variable catégorielle à 4 modalités.
  - Ex : combiner l’information de la variable "marié" (0=pas marié) avec l’information de la variable "femme" (0=homme).
  - La variable correspondante comprend 4 catégories : les hommes mariés, les femmes mariées, les hommes non mariés, les femmes non mariées.
- Idem : on introduit une indicatrice pour chaque modalité, en omettant une des modalités, qui sera la référence

Inclure une variable ordinale

- Cas d'un variable qualitative ordinale, 2 possibilités:
- On peut l'intégrer directement dans le modèle
  - Ex: les tranches d'âges, les tranches de revenus
- Ou bien : on créé autant de variables indicatrices que de modalités, en omettant une des modalités qui sera la référence.
  - Intérêt : l'impact du passage de l'une à l'autre catégorie n'est plus supposé constant.
  - Permet des catégories moins homogènes

Inclure une variable ordinale

- Cas d'un variable catégorielle ordinale
- Comment l'inclure ?

- Ex: une variable état de santé (bon, moyen, mauvais)

$$log(\mathrm{D}sante) = 0,15 + 0,23.\ ESmoyen + 0,37.\ ESmauvais$$

  – Un individu ayant un état de santé moyen (resp. mauvais) aura une dépense de santé annuelle supérieur de 23% (resp. 37%) à celle d'un individu ayant un bon état de santé (la référence).

Tests d'hypothèses

- L'estimateur des MCO est une variable aléatoire
  - Donc sa distribution est soumise à une loi de probabilité
- Connaître la distribution de l'estimateur permet de faire des tests statistiques, de déterminer les intervalles de confiance
- ! On ne connaît la loi de distribution de β
- Comment déterminer cette loi ?

La loi de β

- Empiriquement
    - On dispose d'un échantillon,
    - On créer 1000 sous-échantillons aléatoires,
    - On procède à 1000 régressions,
    - Déterminer la loi qui s'en rapproche le plus,
    - Déterminer l'intervalle de confiance
  - Techniques de ré-échantillonnage, quand on n'a pas d'idée sur la loi de β
  - Simulation de Monte-Carlo, Boostrap, Jacknife,permutations etc.
  - Lourd à mettre en œuvre, mais on a vu la procédure sur R (inférence)

La loi de β

- Théoriquement
  - On a une idée sur la loi des β
  - On suppose que β suit une loi Normale
  - Cela revient a supposer que les erreurs suivent une loi Normale (hypothèse de l'OLS)
  - La normalité des termes d'erreur se translate directement à l'estimateur

$$u \sim N(0, \sigma^2)$$

La loi de β

- Hypothèse de Normalité $u \sim N(0, \sigma^2)$
  - Implication : E(u)=0 et homoscédasticite
  - On dit les erreurs sont i.i.d. indépendantes et identiquement distribuées
    - Conséquence : estimateur MCO de variance minimale parmi les estimateurs sans biais
  - Hypothèse forte, indispensable pour effectuer des tests paramétriques
    - Rend le modèle paramétrique
    - Permet l'inférence
  - Crédibilité ?

La loi de β

- En faveur de la normalité …
- TCL + LDGN
- Cette hypothèse est «vraisemblable» avec n assez grand

  => convergence asymptotique
- L'hypothèse de Normalité des erreurs est communément admise dès lors que
  - Le nombre d'observations est suffisamment grand
  - La forme fonctionnelle est adaptée

La loi de β

- Si on connait $\sigma^2\,\beta j$, pour chaque paramètre on a :

$$\frac{\widehat{\beta}_j - \beta}{\sigma_{\beta_j}} \sim N(0,1)$$

Les coefficients sont dits centrés-réduits
ou « z-scores » dans les sorties de logiciels

La loi de β

- Si on ne connait pas σ^2 βj : estimation des variances des estimateurs
  - Sauf dans les rares cas où σ" est observable, la variance de l'estimateur doit être estimée.
  - Nous utilisons l'estimateur de σ^2,

$$\sqrt{\widehat{V(\hat{\beta}_1)}} \rightarrow \hat{\sigma}_{\hat{\beta}_1}$$

- Loi suivie par les estimateurs MCO

Qui suit une distribution de Student qui permet de calculer un intervale de confiance

$$: \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim T_{N-2}$$

Test d'hypothèse de nullité du coefficient

Estimation par intervalles de confiance :

- Remarques :
  - Par construction $\beta 1$: se situe dans IC.
  - L'IC va nous informer sur la précision de notre estimation : elle sera d'autant plus précise que l'IC est étroit.
  - Si la valeur 0 ne fait pas partie de l'IC 95% nous pouvons rejeter l'hypothèse selon laquelle la variable x n'a pas d'impact sur y, au seuil de 5%. x a donc un impact significatif sur y . La variable x est pertinente.

Test d'hypothèse de nullité du coefficient

- C'est un test de conformité par rapport à une norme
- On veut tester l'hypothèse que la variable explicative x n'a pas d'impact sur y (H0)
- Objectif :
  - x est-elle une variable pertinente pour expliquer y ?
  - le paramètre $\beta 1$ est-il significatif ?
- !!!! Attention : nous testons une hypothèse relative aux paramètres de la population,
  - et non relative à des estimations particulières pour un échantillon donné.
  - !!! nous ne testons PAS : H0: $\beta 1 = 0$.