Questions pratiques 4:
Transformer la variable dépendante

Jean-François Bickel

Statistique II – SPO8

- **Transformer** une variable consiste en une opération arithmétique qui vise à construire une nouvelle variable à partir de la variable d'origine, de sorte que la **distribution** de la nouvelle variable soit **différente** de celle de la variable d'origine et plus conforme à certaines caractéristiques, tout en **préservant l'ordre** des valeurs de la variable d'origine

- Habituellement, transformer une variable dépendante est peu recommandable, car cette opération a aussi pour effet de modifier la relation entre variable dépendante et variables indépendantes, ce qui est normalement considéré comme indésirable

- Elle peut aussi rendre moins évidente l'interprétation
  - notamment dans le cas où les variables x et y réfèrent à des grandeurs de sens commun (le revenu, l'âge, etc.)

- Dans certains cas toutefois, la relation transformée apporte des avantages pour l'interprétation et/ou offre une meilleure description du phénomène sous examen
- La transformation la plus courante pour une variable dépendante est de prendre son logarithme, le plus souvent le logarithme naturel, i.e. en utilisant la base **e**, qui vaut approximativement 2.71828

- L’équation de régression devient dès lors

$$\ln[E(y)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k$$

avec ln[E(y)] = logarithme naturel de l’espérance (moyenne) de y

- Si on élève à l'exponentielle les termes de l'équation de chaque côté du signe égal, on obtient

$$\mathbf{E(y) = \exp\ (\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k)}$$

Nota Bene

- Dans le cas d'une transformation logarithmique, les valeurs de la variable d'origine (avant transformation) doivent toutes être supérieure à zéro (log naturel de 0: pas possible), et il est préférable que la valeur minimale ne soit pas inférieure à 1 (log naturel de 1=0), ce qui permet d'avoir des valeurs uniquement positives dans la nouvelle variable

- Trois raisons peuvent rendre attractif un modèle avec log naturel de y plutôt que y comme variable dépendante

1) Certaines variables comme le salaire annuel ou le temps passé à regarder la télévision n'ont jamais de valeurs négatives
- La transformation logarithmique garantit que quelque soit les valeurs des β et des x, la valeur prédite de y sera toujours positive

2) Dans le modèle linéaire usuel, l'effet d'augmenter une certaine variable x d'une unité produit un certain *changement en valeur absolue de y*

- Dans le modèle logarithmique, l'effet d'augmenter une certaine variable x d'une unité produit un certain *changement en pourcent de y*
- Ce qui peut faire davantage sens, par exemple pour les variables monétaires

- Soit un modèle de régression de type

$$\ln[E(y)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k$$

- Une simple transformation des coefficients β permet de les convertir en changement exprimé en pourcent
- Cette transformation a la forme

$$100(e^{\beta} - 1)$$

3) Une troisième raison pour transformer la variable dépendante est que la nouvelle variable permette une meilleure description du phénomène sous examen

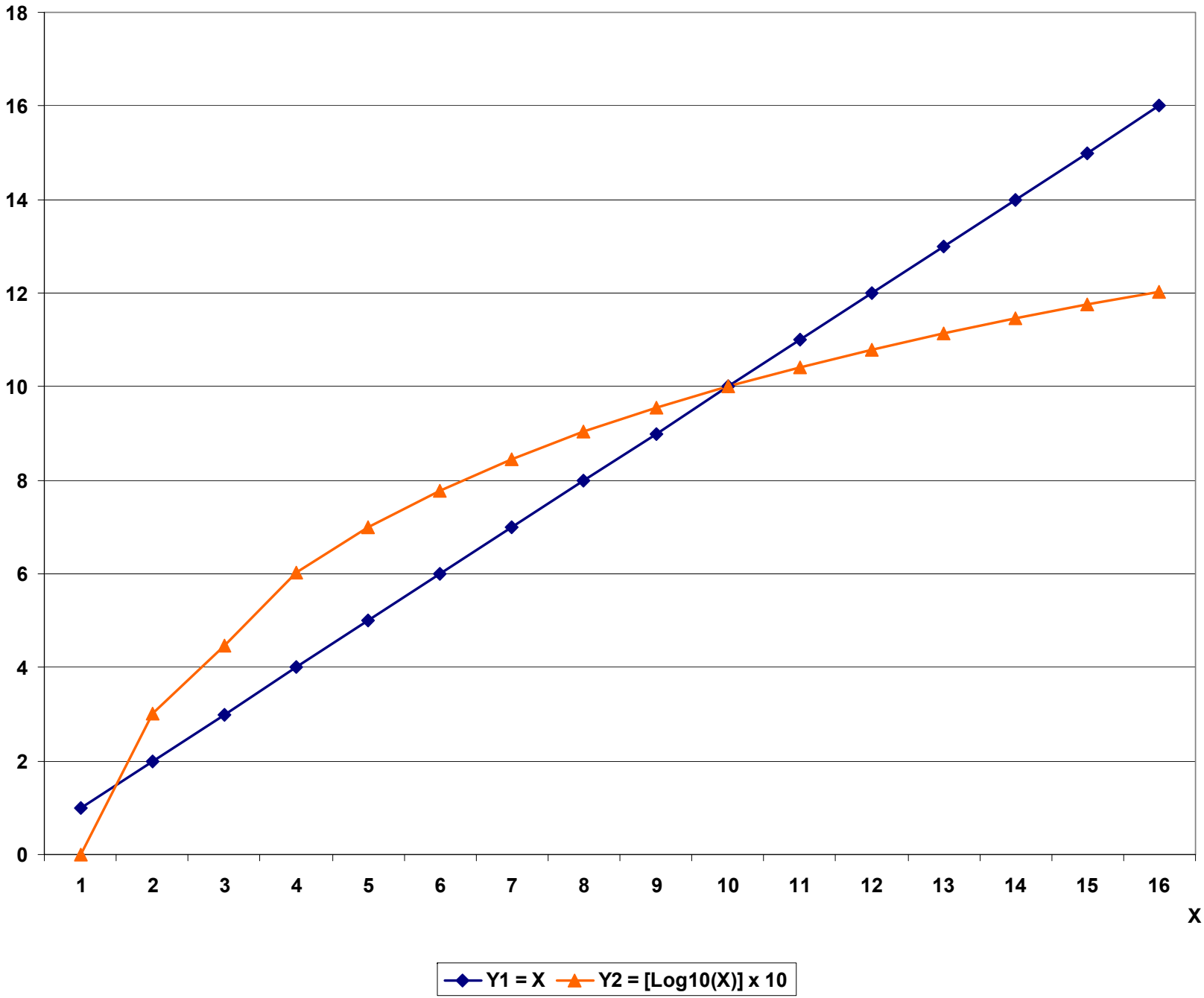
- Une illustration graphique peut aider à éclaircir ce point

- Soit une variable X allant de 1 à 16
  on lui fait subir deux transformations produisant deux nouvelles variables Y1 et Y2

i. Y1 = X

ii. Y2 = [log10(X)] x 10

  - la valeur du log10 de X est multipliée par 10 pour une question d'échelle


18
16
14
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
X
Y1 = X
Y2 = [Log10(X)] x 10

- On remarque que la pente de Y1 est constante par rapport aux valeurs de X
  - lorsqu'on passe de X=1 à X=2, Y1 augmente de 1; il en va de même quand on passe de X=15 à X=16
- Pour Y3, la pente est variable; elle est plus forte pour les valeurs petites de X puis s'atténue au fur et à mesure que X s'accroît

- Dans le cas d’une variable avec une forte asymétrie à droite de sa distribution, une transformation logarithmique (log10 ou log naturel) a pour effet

i. d’accentuer l’impact des différences de valeurs sur la gauche de la distribution, là où sont regroupés la majorité des observations

ii. de réduire l’impact des différences de valeurs sur la droite de la distribution, là où on rencontre des observations peu nombreuses et éloignées des valeurs majoritaires

- Or, le revenu se caractérise précisément par un grand nombre d’observations groupées sur la gauche de la distribution

- Prendre le logarithme du revenu permet de donner comparativement plus de poids aux différences de revenus parmi le groupe majoritaire et de mieux différencier les situations en son sein
- Cela permet aussi de réduire le poids conféré aux valeurs extrêmes dans l'estimation des paramètres du modèle de régression

- On peut donc espérer que ce dernier se rapproche davantage des valeurs observées et offre une description plus fine du processus de formation du revenu

Exemple

- Transformation tout d'abord de la variable revenu du travail en une nouvelle variable dont les valeurs représentent le logarithme naturel des valeurs originelles

- **Syntaxe**

```
compute lognrev=ln(i05wy).
exe.
```

- Puis effectuation d’une analyse de régression avec **lognrev** comme variable dépendante
- Par simplification, le modèle ne comprend que deux variables indépendantes, à savoir sexe et éducation (sans interaction)
- Comparaison enfin des résultats avec ceux de l’exercice 6

Récapitulatif du modèle

Avec i05wy

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	.515[a]	.265	.265	49811.830

Avec lognrev

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	.546[a]	.298	.298	.92217

Commentaire

- Le R^2 pour la variable dépendante revenu est de .265, alors que le R^2 pour la variable dépendante logarithme du revenu est .298
- La transformation de la variable a permis d'améliorer la qualité prédictive du modèle; celui-ci décrit mieux, se rapproche davantage des données observées

Coefficients

Avec i05wy

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Signification	Intervalle de confiance à 95% de B	
		B	Erreur standard	Bêta			Borne inférieure	Borne supérieure
1	(constante)	42162.397	2009.953		20.977	.000	38221.739	46103.054
	femmes	-38082.3	1607.449	-.328	-23.691	.000	-41233.77	-34930.732
	EDUCAT05 Niveau de formation le plus élevé (grille + q. ind.)	7219.054	279.149	.358	25.861	.000	6671.763	7766.345

Avec lognrev

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Signification	Intervalle de confiance à 95% de B	
		B	Erreur standard	Bêta			Borne inférieure	Borne supérieure
1	(constante)	10.111	.037		271.730	.000	10.038	10.184
	femmes	-.685	.030	-.311	-23.006	.000	-.743	-.626
	EDUCAT05 Niveau de formation le plus élevé (grille + q. ind.)	.157	.005	.411	30.442	.000	.147	.167

Commentaire

- Avec le logarithme du revenu, la direction des effets est similaire
- Par contre, leur interprétation est modifiée

- Pour **femmes**, le coefficient s'interprète à l'aide de la transformation

  $100(e^{-.685} - 1) = -49.6\%$

- Donc, les femmes ont en moyenne un revenu du travail moitié moins élevé (49.6%) que les hommes, à niveau d'éducation égal

- Pour **éducation**, le coefficient s'interprète à l'aide de la transformation

  $100(e^{.157} - 1) = 17.0\%$

- Donc, chaque niveau d'éducation supplémentaire accroît le revenu moyen de 17%, à sexe égal