ECON 3150/4150, Spring term 2013. Lecture 2

Data transformations and flexible functional forms

Ragnar Nymoen

University of Oslo

17 January 2013

Regression with transformed variables I

- References: See Lecture 1
- Transformation of the data *prior to fitting the regression line* is often used in applied work.
- The greatly extends the relevance of OLS estimation to real world data
- Distinguish between
  - Linear transformations
  - Non linear transformations ("flexible functional forms")
- In this lecture we give an introduction to some of the possibilities that we have at our disposal

De-meaning I

- We have already encountered *de-meaning* of the regressor X as a way of simplifying the derivations of the OLS estimates.
- Now, consider de-meaning both variables:

$$Y_i^* = Y_i - \bar{Y}$$
$$X_i^* = X_i - \bar{X}$$

where the transformed variables are denoted Y_i^* and X_i^* $(i = 1, 2, \ldots, n)$.

De-meaning II

- Based on the same argument as in Lecture 1, the best predictor of Y_i^* given X_i^* is

$$\hat{Y}_i^* = \hat{\beta}_0^* + \hat{\beta}_1^* X_i^* \tag{1}$$

OLS estimation (min.sum of sq.residuals) gives

$$\hat{\beta}_0^* = \overline{Y^*} - \hat{\beta}_1 \overline{X^*}$$

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^{n}(X_i^* - \overline{X^*})Y_i^*}{\sum_{i=1}^{n}(X_i^* - \overline{X^*})^2}$$

De-meaning III

- By construction, $\overline{Y^*} = \overline{X^*} = 0$, and:

$$\hat{\beta}_0^* = 0 \tag{2}$$

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^{n}(X_i^*)Y_i^*}{\sum_{i=1}^{n}(X_i^*)^2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \equiv \hat{\beta}_1 \tag{3}$$

Insights to take away from this:

1. If you de-mean both the regressand and the regressor, the regression line has intercept 0
2. The regression line goes trough the origin of the scatter plot between Y_i^* and X_i^*
3. When Y_i^* is regressed on X_i^* we can therefore drop the intercept/constant from the regression, and write the best predictor as $\hat{Y}_i^* = \hat{\beta}_1^* X_i^*$ where $\hat{\beta}_1^* \equiv \hat{\beta}_1$as shown.

WARNING!!!!!!

- Unless both variables are de-meaned, you should ALWAYS include the intercept in the regression line. Otherwise you do **not** get the best predictor for Y given X, the estimate of the slope coefficient will also be wrong.
- Specifically, you can show as an exercise that if Y_i is regressed on X_i with no intercept, the OLS estimate of the slope parameter becomes

$$\widehat{\beta}_1^{no-i} = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2} \neq \hat{\beta}_1$$

unless the means of Y_i should just happen to be zero!

Scaling I

- Scaling is done by multiplying the original data with the known factors ω_y and ω_x.
- For example: change units from thousand to million or billion. Let Y_i^ω and X_i^ω denote the *scaled variables*

$$Y_i^\omega = \omega_y Y_i$$
$$X_i^\omega = \omega_x X_i$$

- By deriving the OLS estimates $\hat{\beta}_0^\omega$ and $\hat{\beta}_1^\omega$ you can show that

Scaling II

$$\hat{\beta}_0^{\omega} = \omega_y \hat{\beta}_0 \tag{4}$$

$$\hat{\beta}_1^{\omega} = \frac{\omega_y}{\omega_x} \hat{\beta}_1 \tag{5}$$

- Scaling of one or both of the variables will affect the OLS estimates
- If for example X_i is in thousands, and X_i^{ω} is in millions then $\omega_x = 0.001$.
  - If $\omega_y = 1$, no scaling of Y_i, $\hat{\beta}_1 = 0.005$ is changed to $\hat{\beta}_1^{\omega} = 5$ after the scaling.
  - If on the other hand, $\omega_x = \omega_y$, the slope estimate is unchanged by the scaling, but the intercept changes.

Standardized variables I

Finally imagine first de-meaning Y_i and X_i, and second scaling the de-meaned variables by

$$\omega_y = \frac{1}{\hat{\sigma}_Y}$$
$$\omega_x = \frac{1}{\hat{\sigma}_X}$$

where $\hat{\sigma}_y$ and $\hat{\sigma}_x$ are the empirical standard deviations

$$\hat{\sigma}_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}, \text{ and } \hat{\sigma}_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Standardized variables II

$$Y_i^{*\omega} = \frac{Y_i - \bar{Y}}{\hat{\sigma}_y}$$

$$X_i^{*\omega} = \frac{X_i - \bar{X}}{\hat{\sigma}_x}$$

The *standardized* regression becomes

$$\hat{Y}_i^{*\omega} = \hat{\beta}_1^{*\omega} X_i^{*\omega} \tag{6}$$

- Since standardization is a combination of de-meaning and scaling we have that

$$\hat{\beta}_1^{*\omega} = \frac{\omega_Y}{\omega_X}\hat{\beta}_1 = \frac{\hat{\sigma}_X}{\hat{\sigma}_Y}\hat{\beta}_1 = \frac{\hat{\sigma}_X}{\hat{\sigma}_Y}\frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X} = r_{XY} \tag{7}$$

- With standardized variables, regression is reduced to "correlation analysis".

Estimating non-linear relationships I

- If OLS can only be used to fit linear relationships between Y and X, the relevance of the method will be very limited.
- However, by applying non-linear transformations of Y_i and X_i before estimation, we can estimate many interesting non-linear functions with OLS.
- Using the transformed variables the model is *linear in the parameters* β_0 and β_1.
- In this way we obtain *great flexibility* in fitting different non-linear relationships between Y and X.
- In applied econometrics, we often refer to non-linear data transformations as the *choice of functional form.*

Quadratic transformation of the regressor I

Assume that we have an theoretical non-linear relationship between Y and X:

$$Y = \beta_1 + \beta_1 X^2$$

This can be put into regression form by regressing Y_i on the squared X_i:

$$X_i^* = X_i^2$$

Hence we have

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i^*$$

Quadratic transformation of the regressor II

where $\hat{\beta}_0$ and $\hat{\beta}_1$are calculated with the use of the OLS formulae (using X_i^* in the place of X_i). The estimated derivative in this regression depends on X:

$$\widehat{\frac{\partial Y}{\partial X}} = 2\hat{\beta}_1 X_i$$

which is increasing in X_i if $\beta_1 > 0$.

- If Y is a measure of costs, and X is a measure of production (or of capacity), this model may be relevant to estimate a cost-function with increasing marginal cost
- See HGL Figure 2.13 and 2.14

Log-linear models I

If one or both of the variables are log transformed, we speak of *log-linear models*:

$$\begin{aligned} \text{i} \quad Y &= \beta_0 + \beta_1 \ln X \\ \text{ii} \quad \ln Y &= \beta_0 + \beta_1 X \\ \text{iii} \quad \ln Y &= \beta_0 + \beta_1 \ln X \end{aligned}$$

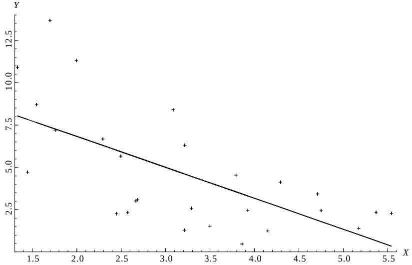
- The two first are sometimes called **semi-logarithmic models**.
- The third is sometimes called the **log-log model**.
- All three relationships can be formulated as linear regressions and OLS estimation can be applied.
- The differences lies in the interpretation.

Log-linear models II

- i), ii) and iii) will have
- different derivatives,
- different elasticites $(El_x y)$
- and different semi-elasticities$(\frac{\partial y}{\partial x}\frac{1}{y})$

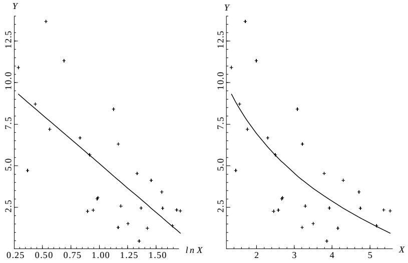
	$\widehat{\frac{\partial y}{\partial x}}$	$\widehat{\frac{\partial y}{\partial x}\frac{1}{y}}$	$\widehat{El_x y}$
i	$\hat{\beta}_1\frac{1}{X}$	$\hat{\beta}_1\frac{Y}{X}$	$\hat{\beta}_1 Y$
ii	$\hat{\beta}_1 Y$	$\hat{\beta}_1$	$\hat{\beta}_1 X$
iii	$\hat{\beta}_1\frac{Y}{X}$	$\hat{\beta}_2\frac{1}{X}$	$\hat{\beta}_1$

Phillips curve models (PCMs) for Norway provides some illustrations



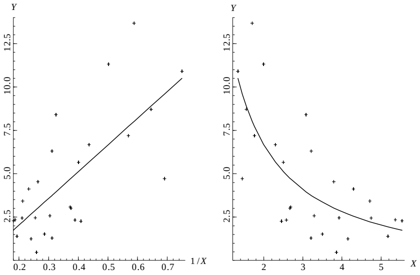
Inflation rate Y_i, and unemployment rate X_i, with regression line. Sample 1979 to 2005.

- The linear Phillips curve: $Y_i = 10.5 - 1.83X_i$
- $\hat{\beta}_1 = -1.83$, $R^2 = 0.43$
- i-t rate of u = 4.36 %
- natural rate = 5.73 %



Log scale for X_i to the left, percent scale to the right

- The lin-log Phillips curve: $Y_i = 11 - 5.87 \ln X_i$
- $\hat{\beta}_1 = -5.87$, $R^2 = 0.49$
- Note the (small) increase in R^2 Proof of better fit than linear?
- i-t rate of u = 4.25 %
- natural rate = 6.5 %



Phillips curve with $1/X$ as regressor to the left. Ordinary scale to the right.

- The Phillips curve with inverse X $Y_i = -1 + 15.39(1/X_i)$
- $\hat{\beta}_1 = 15.39$, $R^2 = 0.49$
- i-t rate of u= 4.36 %
- natural rate = 14.9 %

- As said, these were just illustrations of the great flexibility that we have by making relevant choices of functional forms.
- The choice of functional form is once of the most important decisions that we make in econometric modelling
- Will return to the example of Norwegian PCMs later, when we have developed the statistical inference theory for regression models.