ECON 3150/4150, Spring term 2014. Lecture 6

The simple regression model (part III), dummy regressor and heteroskedasticity

Ragnar Nymoen

University of Oslo

30 January 2014

References to Lecture 6

- **SW**
  - Ch. 5.3-5.4
- Bårdsen and Nymoen (**BN)**
  - Kap 7.8 (dummy varaible), 8.2.2 (heterosked)

Dummy variable as regressor (revisited) I

- When we want to test the equality of two expectations, μ_{Y_0} and μ_{Y_1}, and the sample of centred and standardized variables $(Y_{ij} - \mu_{Y_j})/\sigma_Y$ is i.i.d., for $(i = 1, 2, \ldots, n_j)$ and $j = 0, 1$, the test can be done with use of the regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{1}$$

where $\varepsilon_i \sim i.i.d(0, \sigma^2)$, $\sigma^2 \equiv \sigma_Y^2$ and X_i is a *binary variable*, also called *indicator variable*, or a *dummy*

$$X_i = \begin{cases} 0 & \text{for } i = 1, 2, \ldots, n_0 \\ 1 & \text{for } i = n_0 + 1, \ldots, n \end{cases} \tag{2}$$

so that $n_1 = n - n_0$.

Dummy variable as regressor (revisited) II

- Often the symbol D_i is used instead of X_i is such cases, and $X_i = 0$ is called the *reference value* of the variable.
- The same formulation can be used whenever the regressor is an indicator variable, see Ch. 5.3 in SW (in BN the discussion of binary variables is found in Kap 7.8, on multiple regression).
- Under the classical RM assumptions:

$$E(\varepsilon_i) = 0 \ \forall i, \ \ var(\varepsilon_i) = \sigma^2 \ \forall i, \ \ cov(\varepsilon_i, \varepsilon_j) = 0 \ \forall i, j$$

the OLS estimators of β_0 and β_1 will be BLUE as before (as when X_i is a continuous variable)

Dummy variable as regressor (revisited) III

- As we have noted: The OLS estimators are given by

$$\hat{\beta}_0 = \bar{Y}_0 \tag{3}$$

$$\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0 \tag{4}$$

  where $\bar{Y}_0$ refers to the $X_i = 0$ part of the sample, and $\bar{Y}_1$ is for the the $X_i = 0$ part.
- Therefore, $\hat{\beta}_1$ gives the estimated change in the intercept (not the slope!) and $\hat{\beta}_1$ is some times called the *difference-estimator*

Algebra the dummy regressor case I

Since X takes n_0 zeros and n_1 ones:

$$\bar{X} = \frac{1}{n_0 + n_1} \cdot n_1 = \frac{n_1}{n_0 + n_1}$$

$$\bar{X}^2 = \left(\frac{n_1}{n_0 + n_1} \right)^2$$

$$\sum_{i=1}^{n} X_i^2 = n_1$$

$$\bar{Y}\bar{X} = \bar{Y} \frac{n_1}{n_0 + n_1} = \left(\frac{1}{n_0 + n_1} \sum_{i=1}^{n_0+n_1} Y_i \right) \frac{n_1}{n_0 + n_1}$$

$$\sum_{i=1}^{n} X_i Y_i = \sum_{i=1}^{n_1} Y_i$$

Algebra the dummy regressor case II

If we start from the expression

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

we have already:

$$\hat{\beta}_1 = \frac{\frac{1}{n_0+n_1}\sum_{i=1}^{n_1} Y_i - \bar{Y}\frac{n_1}{n_0+n_1}}{\frac{1}{n_0+n_1}n_1 - \left(\frac{n_1}{n_0+n_1}\right)^2} \tag{5}$$

Algebra the dummy regressor case III

The numerator of $\hat{\beta}_1$:

$$\frac{1}{n_0+n_1}\sum_{i=1}^{n_1} Y_i - \left(\frac{1}{n_0+n_1}\sum_{i=1}^{n_0+n_1} Y_i\right)\frac{n_1}{n_0+n_1}$$

$$= \frac{1}{n_0+n_1}\left\{\sum_{i=1}^{n_1} Y_i - \frac{n_1}{n_0+n_1}\left(\sum_{i=1}^{n_0} Y_i + \sum_{i=1}^{n_1} Y_i\right)\right\}$$

$$= \frac{1}{n_0+n_1}\left\{\frac{n_0}{n_0+n_1}\sum_{i=1}^{n_1} Y_i - \frac{n_1}{n_0+n_1}\sum_{i=1}^{n_0} Y_i\right\}$$

$$= \frac{1}{n_0+n_1}\left\{\frac{n_0}{n_0+n_1} n_1 \bar{Y}_1 - \frac{n_1}{n_0+n_1} n_0 \bar{Y}_0\right\}$$

$$= \frac{n_0 n_1}{(n_0+n_1)^2}(\bar{Y}_1 - \bar{Y}_0)$$

Algebra the dummy regressor case IV

The denominator of $\hat{\beta}_1$:

$$\frac{1}{n_0+n_1}n_1 - \left(\frac{n_1}{n_0+n_1}\right)^2 = \frac{n_1}{n_0+n_1}\left(1-\frac{n_1}{n_0+n_1}\right)$$
$$= \frac{n_1 n_0}{(n_0+n_1)^2}$$

Collect:

$$\hat{\beta}_1 = \frac{\frac{n_0 n_1}{(n_0+n_1)^2}(\bar{Y}_1 - \bar{Y}_0)}{\frac{n_0 n_1}{(n_0+n_1)^2}} = \bar{Y}_1 - \bar{Y}_0$$

Algebra the dummy regressor case V

For $\hat{\beta}_0$ we get:

$$\begin{aligned}\hat{\beta}_0 &= \bar{Y} - \bar{X}(\bar{Y}_1 - \bar{Y}_0) = \bar{Y} - \frac{n_1}{n_0 + n_1}(\bar{Y}_1 - \bar{Y}_0) \\ &= \frac{n_0}{n_0 + n_1}\bar{Y}_0 + \frac{n_1}{n_0 + n_1}\bar{Y}_1 - \frac{n_1}{n_0 + n_1}(\bar{Y}_1 - \bar{Y}_0) \\ &= \bar{Y}_0\end{aligned}$$

Heteroskedasticity

- If the variances of the disturbances are not all identical, the homoskedasticity assumption

$$var(\varepsilon_i \mid X_i) = \sigma^2$$

of the regression model is violated, and we have heteroskedasticity.

Consequences of heteroskedasticity I

- The OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are *unbiased* and *consistent* also in the case of heteroskedastic disturbances, since

$$var(\varepsilon_i \mid X_i) = \sigma^2 \ \forall i$$

does not enter into the proofs of these properties.

- The OLS estimators are however *no longer* BLUE since the formulae

$$var(\hat{\beta}_1) = \frac{\sigma^2}{n\hat{\sigma}_X^2}$$

will either over- or underestimate the true variance of $\hat{\beta}_1$, when $Var(\varepsilon_i \mid X_i) = \sigma_i^2$.

Consequences of heteroskedasticity II

- Unless we either
  - re-specify the model, so that the disturbances of the re-specified model become homoskeadstic, or
  - fix the estimation of $Var(\hat{\beta}_1)$ to become heteroskedastic robust

  the *t-ratio* based on $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ will be biased and the statistical inference is no longer reliable under heteroskedasticity.
- Look at the fix for the standard errors first, then examples of re-specification.

Heteroskedasticity: robust variance estimation I

- Under homoskedasticity we use:

$$\widehat{var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{n\hat{\sigma}_X^2} \tag{6}$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 \tag{7}$$

- SW refer to (6) as the *homoskedasticity only* expression for the estimation of $var(\hat{\beta}_1)$. See their equation (5.22).

Heteroskedasticity: robust variance estimation II

- SW have then already introduced the *heteroskedasticity robust* estimator

$$\widetilde{var}(\hat{\beta}_1) = \frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \bar{X})^2\hat{\varepsilon}_i^2}{n\hat{\sigma}_X^4}$$

  in equation (5.4) in their book. $\sqrt{\widetilde{var}(\hat{\beta}_1)}$ is often referred to as White-robust standard errors (Hal White, San Diego based econometrician)
- Intuitively, using $\widetilde{var}(\hat{\beta}_1)$ instead of $\widehat{var}(\hat{\beta}_1)$ makes inference become reliable again if "heteroskedasticity depends on X ".

Forms of heteroskedasticity I

- A form that is sometimes referred to as "classical heteroskedasticity" is

$$\sigma_i^2 = \sigma^2 W_i^h \text{ with } h > 0 \tag{8}$$

  where W_i is an observable variable.
- A situation which is not uncommon, is that the scatter plot suggests:

$$Var(Y \mid X) = \sigma^2 X^2$$

- If heteroskedasticity is of this type, the problem created by heteroskedasticity for inference is easily corrected by so called weighted least squares (WLS), and also the heteroskedastic robust $\widetilde{var}(\hat{\beta}_1)$ will then work well.

Forms of heteroskedasticity II

- For later reference: another *Het.* form which is relevant for models of time series data, is *a*uto*r*egressive *c*onditional *h*eteroskedasticity, ARCH. The first order ARCH is:

$$\sigma_t^2 = a_0 + a_1 \sigma_{t-1}^2 \tag{9}$$

Weighted least squares (WLS) I

- If we express the case of $Var(Y \mid X) = \sigma^2 X^2$ in model form, we have

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

with

$$var(\varepsilon_i) = \sigma^2 X_i^2$$

and all the other classical properties holding.

- Consider now the following model:

$$\frac{Y_i}{X_i} = \frac{\beta_0}{X_i} + \beta_1 + \varepsilon_i^*$$

and with suitable change in notation

$$Y_i^* = \beta_0^* + \beta_1^* X_i^* + \varepsilon_i^* \tag{10}$$

where $Y_i^* = Y_i / X_i$ and $X_i^* = 1/X_i$, $\varepsilon_i^* = \varepsilon_i / X_i$

Weighted least squares (WLS) II

- For this model we have homoskedasticity, since:

$$var(\varepsilon_i^*) = X_i^{-2} var(\varepsilon_i) = \sigma^2$$

  and the OLS estimators of β_0^* and β_1^* have the BLUE property.
- They are *Weighted Least Squares* estimators (WLS), since the original data have been weighted in a way that brings the model back to the classical RM form (incl. homoskedasticity). Note:
- Robust standard errors only changes the estimated variances without changing the OLS estimators
- WLS is a different estimator than OLS, so that $\hat{\beta}_1^* \neq \hat{\beta}_1$, in this example

Regression using sample averages I

- Assume that the model is

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij} \tag{11}$$

for $i = 1, 2, \ldots, n_j$ and $j = 1, 2, \ldots, J$ (number of sub-samples) and that ε_{ij} has the classical properties which would have made the OLS estimators BLUE if we could have estimated the model (11) with individual data.

- Assume that we only have access to sub-sample averages:

$$\bar{Y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij} \text{ and } \bar{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}$$

and that we have to estimate the regression

$$\bar{Y}_j = \beta_0 + \beta_1 \bar{X}_j + \bar{\varepsilon}_j, \; j = 1, 2, \ldots, J \tag{12}$$

Regression using sample averages II

- Can we obtain BLUE estimators of β_0 and β_1 in this case?