Structural changes in errors

Susan Thomas
IGIDR, Bombay

25 November, 2008

Change in distribution of errors

Changes in the error distribution

- Estimation residuals $\hat{\epsilon}_i$ are to be *i.i.d*.
- They are not if:
  - There are dependencies in the $\hat{\epsilon}_i$.
    An extreme form of such a dependency is serial dependency: there is correlation between $\hat{\epsilon}_i$ and $\hat{\epsilon}_{i+1}$
  - There are change in $\sigma^2_{\hat{\epsilon}}$

Dependence in $\hat{\epsilon}$

- Serial dependence in $\hat{\epsilon}$ can be detected using the autocorrelations coefficients between observations at i, j. This is denoted as ρ_τ and defined as:

$$\rho_\tau = \frac{\sum_t y_t y_{(t-\tau)}}{\sqrt{\sum_t y_t^2 \sum_t y_{(t-\tau)}^2)}} = \frac{\sum_t y_t y_{t-\tau}}{\sigma_{y_t}\sigma_{y_{(t-\tau)}}} = \frac{\sum_t y_t y_{t-\tau}}{\sigma_{y_t}^2}$$

- Under $H_0 : \rho_\tau = 0$, $\sigma_{\rho_\tau} = 1/\sqrt{T}$ where T is the number of observations.
- Test statistic for an autocorrelation at lag τ, ρ_τ:

$$\rho_\tau / \sigma_{\rho_\tau}$$

- Critical value: N(0, 1)

Test for serial dependence in $\hat{\epsilon}$

- The Durbin-Watson test statistic:

$$d = \frac{\sum_{i=2}^{T}(\hat{\epsilon}_i - \hat{\epsilon}_{i-1})^2}{\sum_{i=1}^{T}\hat{\epsilon}_i^2}$$

- The Durbin-Watson value always lies between 0 and 4. $d = 2$ is taken as evidence of no serial dependence in errors. $d < 1$ is taken as evidence of positive serial dependence.
- The Breusch-Godfrey test statistic:
  - More general than Durbin-Watson:

$$\hat{\epsilon}_i = \alpha_0 + \alpha_1 X_i + \gamma_1\hat{\epsilon}_{i-1} + \gamma_2\hat{\epsilon}_{i-2} + \gamma_3\hat{\epsilon}_{i-3} + \ldots + \gamma_p\hat{\epsilon}_{i-p} + w_i$$

  - Test statistic: (N - p)
  - Critical value: $\chi^2(p)$

Changes in residual variance

Heteroskedasticity

- Estimation assumption: the residuals are iid.
- Heteroskedasticity: the mean of the distribution of the variables may be the same, but the variance changes from observation to observation.
- Can be a problem with both cross-section as well as time-series data.
- Example of cross-sectional data: the scores of school-going girls on maths tests have a variance that is lower thanthe scores of school-going boys on the same test.
- Example of time series data: the presence of serial dependence in the square of the residuals.

Heteroskedasticity and the effect on OLS estimators

- OLS estimator: $\hat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\epsilon$
- Original framework:

$$\begin{aligned} E(\epsilon\epsilon'|X) &= \sigma^2 I \\ var(\hat{\beta}|X) &= E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'|X] \\ &= (X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1} = \sigma^2(X'X)^{-1} \end{aligned}$$

- With heteroskedasticity,

$$\sigma^2\Omega = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \sigma_N^2 \end{bmatrix}$$

- This gives us the **Generalised Regression Model** where

$$\begin{aligned} E(\epsilon\epsilon'|X) &= \sigma^2\Omega \\ var(\hat{\beta}|X) &= E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'|X] \\ &= (X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1} = (X'X)^{-1}X'\sigma^2\Omega X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1} \end{aligned}$$

- Here, using $\hat{s}^2(X'X)^{-1}$ for inference is incorrect.

Sources of heteroskedasticity

- Two sources of heteroskedasticity:
  - The Y, X relationship varies across groups of observations:

$$\sigma^2\Omega = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \sigma_N^2 \end{bmatrix}$$

    Here, the heteroskedasticity is conditional on X.
  - Autocorrelation:

$$\sigma^2\Omega = \sigma_\epsilon^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{T-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{T-2} \\ \dots & \dots & \dots & \dots & \dots \\ \rho_{T-1} & \rho_{T-2} & \rho_{T-3} & \dots & 1 \end{bmatrix}$$

Test for heteroskedasticity

- $\sigma^2_{\hat{\epsilon}}$ should be the same across randomly selected subsets of the data.

$$H_0 : \sigma_i^2 = \sigma^2 H_A : \sigma_i^2 \neq \sigma^2$$

- General approach for any test for heteroskedasticity will involve:
  1. Estimate the base model and focus on the $\hat{\epsilon}_i^2$.
  2. Run an auxillary regression of the behaviour of $\hat{\epsilon}_i^2$ on the independent data, X.
- For example, if $Y_i = \alpha + \beta X_i + \epsilon_i$, the model used for $\hat{\epsilon}_i^2$ is:

$$\hat{\epsilon}_i^2 = \gamma_0 + \gamma_1 X_i + \gamma_2 X_i^2 + w_i$$

- w_i will not be normally distributed, but rather χ^2
- Since $E(w_i)$ cannot be zero, the value of the intercept is important.
- Some alternative model has to be used to capture the potential source of the heteroskedasticity. Three standard tests for heteroskedasticity are: *Goldfeld-Quandt, Breusch-Pagan, White*

Tests for heteroskedasticity: White

- Most general test: White (1980).
- Test form:
  - If $Y_i = \alpha + \beta_1 X_i + \beta Z_i + \epsilon_i$, then
  - $\hat{\epsilon}_i^2 = \gamma_0 + \gamma_1 X_i + \gamma_2 X_i^2 + \gamma_3 Z_i + \gamma_4 Z_i^2 + \gamma_5 X_i Z_i + w_i$
- Test statistic: NR^2; Critical value: $chi^2(M)$ where M is the number of regressors in the equation including the intercept.
  Above, $M = 6$.
- Problems:
  - There could be other reasons for rejecting $H_0 : \sigma_i^2 = \sigma^2$
    Example: there could be a quadratic relationship between Y_i, X_i that was not included in the base model.
  - If H_0 is rejected, the solution to fix heteroskedasticity is not obvious. We do not have inference on the coefficients of the regressors.

Tests for heteroskedasticity: Goldfeld-Quandt

- Assumes that the heteroskedasicity is due to some dependent variable, X_i. Most extreme case: $\sigma_i^2 = \sigma^2 x_i$
- Test form:
  - First, create subsets of $\hat{\epsilon}_i$ based on the value of X_i. Example, two subsets are created based on the values in X_i, $\hat{\epsilon}_{i,1}, \hat{\epsilon}_{i,2}$ of size n_1, n_2
  - Seperately estimate base model for n_1 observations to get $\hat{\epsilon}_{i,1}$ and n_2 observations to get $\hat{\epsilon}_{i,2}$.
- Test statistic: $F = \frac{\epsilon_1' \epsilon_1/(n_1 - K)}{\epsilon_2' \epsilon_2/(n_2 - K)}$
  Critical value: $F(n_1 - K, n_2 - K)$, K is regressors in the base model.
- Problems:
  - F-distribution is used when the errors are normally distributed. If not, White's test is recommended.
  - Statistician's recommendation: $n_1 + n_2 \neq N$.
    Recommendation: drop no more than a third of the observations.
    Goldfeld-Quandt is not appropriate for small samples.

Tests for heteroskedasticity: Breusch-Pagan

- Also assumes heteroskedasticity is due to some dependent variable, X_i. $\sigma_i^2 = \sigma^2 f(\alpha_0 + \alpha Z)$
  If $\hat{\alpha}$ is significant, there is heteroskedasticity.
- Test form:
  - Like White's test if $Y_i = \alpha + \beta_1 X_i + \beta Z_i + \epsilon_i$,
  - Create Z as a matrix of P dependent variables (like in White's test) not including the intercept $[1, X_i, Z_i, X_i Z_i, X_i^2, Z_i^2]$
  - And $e = \hat{\epsilon}_i^2 / \hat{s}^2$
    where $\hat{s}^2 = \hat{e}'\hat{e}/N$.
- Test statistic: Lagrange Multiplier, $LM = \frac{1}{2}(e'Z(Z'Z)^{-1}Z'e)$
  Critical value: $\chi^2(P)$.
- Problems: The test needs normally distributed $\hat{\epsilon}$.
  Modified test: $LM = \frac{1}{V}((e - \hat{s}^2)'Z(Z'Z)^{-1}Z'(e - \hat{s}^2))$
  where $V = \frac{1}{N}\sum_i(\hat{\epsilon}_i^2 - \hat{s}^2)^2$

Dealing with heteroskedasticity

- If we know the “correct” form of $(X'\Omega X)$, and $(X'\Omega X)$ converges as N grow larger, then OLS estimators remain consistent and unbiased.
- Generalised Least Squares:

$$
\begin{aligned}
E(\epsilon\epsilon'|X) &= \sigma^2\Omega \\
\text{Construct P such that} P'P &= \Omega \\
\text{Then } Py &= PX\beta + P\epsilon \text{ gives us} \\
\hat{\beta} &= (X'P'PX)^{-1}(X'P'PY) = (X\Omega X)^{-1}(X'\Omega Y) \\
var(\hat{\beta}|X) &= E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'|X] \\
&= (X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1} = (X'X)^{-1}X'\sigma^2\Omega X(X'X \\
&= \sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1}
\end{aligned}
$$

This is a transformation of the data Y, X which gives us unbiased and efficient estimates for $\hat{\beta}$.

- This gives us the “correct” inference for the OLS estimates.

Dealing with heteroskedasticity

- What if we do not know the form of the heteroskedasticity?
- We use *White's heteroskedasticity consistent estimator* for Ω.
- $Q = \frac{1}{N} \sum_i \hat{\epsilon}_i^2 x_i x_i'$
- Then the variance of $\hat{\beta}$ becomes:

$$N(X'X)^{-1} Q (X'X)^{-1}$$

- This is analogous to a weighting scheme on the X variables to adjust for the heteroskedasticity in the estimated errors.

Checking for heteroskedasticity in the IIP regression

Variance of log *IIP* by month

- Variances of Log(IIP) by month

	σ(log *IIP*)
Jan	0.3612
Feb	0.3449
Mar	0.3689
Apr	0.3310
May	0.3462
Jun	0.3423
Jul	0.3471
Aug	0.3548
Sep	0.3613
Oct	0.3420
Nov	0.3418
Dec	0.3704

- Question: Can we test whether there are significant differences betweenthe variance of the IIP by different months?

Variance of yoy growth in IIP by month

- Variances of g-IIP by month

	σ_{g-IIP}
Jan	5.0879
Feb	4.4734
Mar	6.7309
Apr	4.0356
May	4.0176
Jun	3.2547
Jul	3.2224
Aug	3.4235
Sep	3.5234
Oct	3.0625
Nov	3.9097
Dec	4.5055

- Here there seems to be more clarity on heteroskedasticity by months – Jan, Feb, Mar appear to have higher levels of error variance compared with the rest of the months.