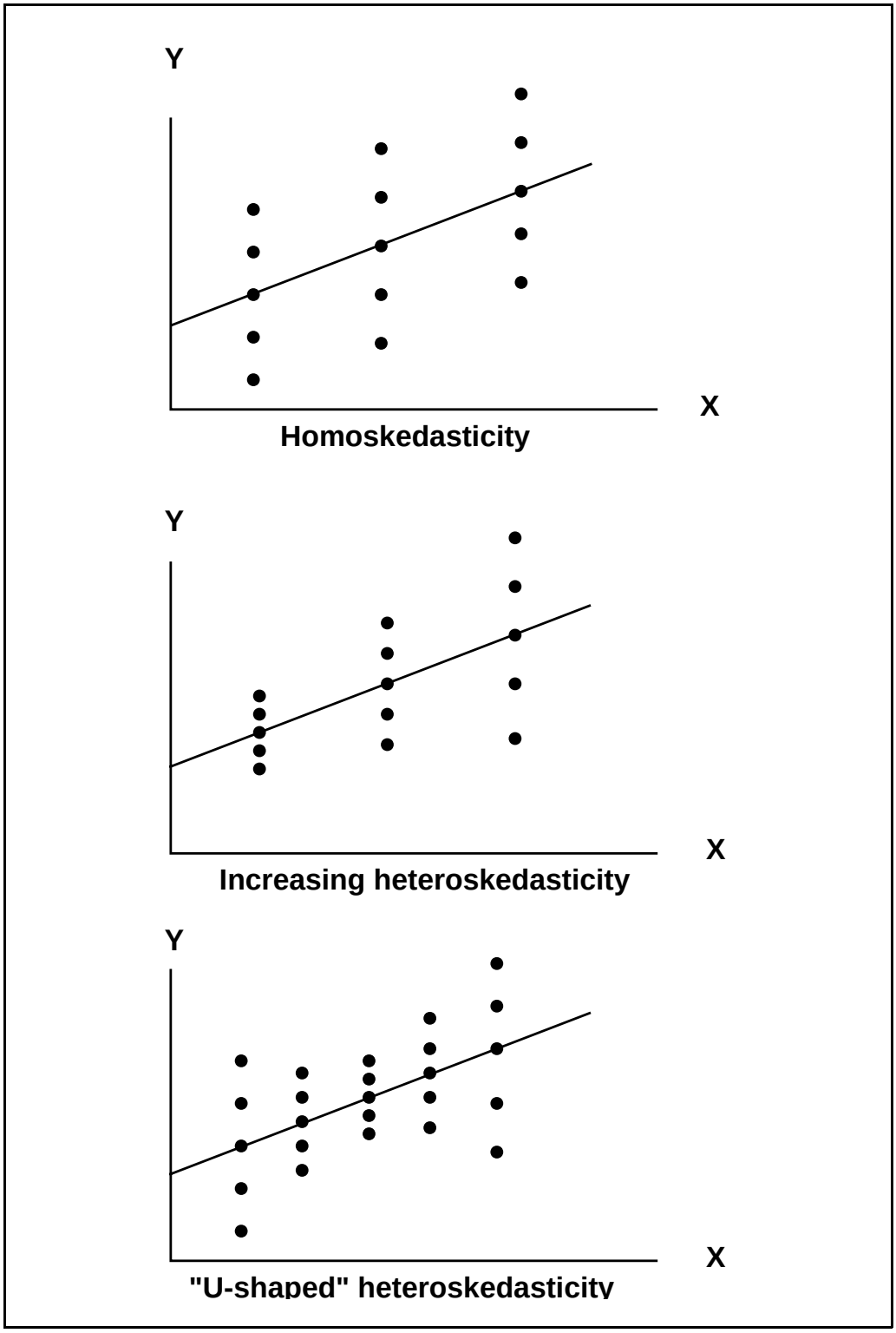
LECTURE / DISCUSSION

Weighted Least Squares

Introduction

In a regression problem with time series data (where the variables have subscript "t" denoting the time the variable was observed), it is common for the error terms to be correlated across time, but with a constant variance; this is the problem of "autocorrelated disturbances," which will be considered in the next lecture.

For regressions with cross-section data (where the subscript "i" now denotes a particular individual or firm at a point in time), it is usually safe to assume the errors are uncorrelated, but often their variances are not constant across individuals. This is known as the problem of ***heteroskedasticity*** (for "unequal scatter"); the usual assumption of constant error variance is referred to as ***homoskedasticity***. Although the mean of the dependent variable might be a linear function of the regressors, the variance of the error terms might also depend on those same regressors, so that the observations might "fan out" in a scatter diagram, as illustrated in the following diagrams.


Y
X
Homoskedasticity
Y
X
Increasing heteroskedasticity
Y
X
"U-shaped" heteroskedasticity

Assumptions of Heteroskedastic Linear Model

- $y_i = \alpha + \beta \cdot x_i + \varepsilon_i$ [simple linear model] or

  $y_i = \sum_{j=1}^{K} x_{ij} \cdot \beta_j + \varepsilon_i$ [multiple regression model];

- $E(\varepsilon_i) = 0$ [zero mean error terms];

- $Cov(\varepsilon_i, \varepsilon_{i'}) = 0 \quad \text{if } i \neq i'$ [no serial correlation]; and

- $Var(\varepsilon_i) = \sigma_i^2 = \sigma^2 \cdot h_i$, some h_i [heteroskedasticity].

Sometimes also assume

- ε_i normally distributed [optional].

Examples of Heteroskedastic Models

1. Grouped (Aggregate) Data

For individual "i" in group "s" (i.e., state, region, time period)

$$y_{is} = \alpha + \beta \cdot x_{is} + \varepsilon_{is} \quad , \quad \text{with } \mathrm{Var}(\varepsilon_{is}) = \sigma^2 \ , \quad \text{etc.}$$

However, we only observe some ***group averages:***

$$\bar{y}_s = \frac{1}{n_s}\sum_{i=1}^{n_s} y_{is} \quad , \quad \bar{x}_s = \frac{1}{n_s}\sum_{i=1}^{n_s} x_{is} \quad .$$

Then

$$\bar{y}_s = \alpha + \beta \cdot \bar{x}_s + \bar{\varepsilon}_s \ , \text{ with } \mathrm{Var}(\bar{\varepsilon}_s) = \sigma^2 \cdot \left(\frac{1}{n_s} \right) \equiv \sigma^2 \cdot h_s \quad .$$

2. Random Coefficients

Both the intercept and slope vary (randomly) across i,

$$y_i \equiv \alpha_i + \beta_i \cdot x_i \quad ,$$

where $E(\alpha_i) \equiv \alpha$, $E(\beta_i) \equiv \beta$, $Var(\alpha_i) \equiv \sigma^2$,

$V(\beta_i) \equiv \tau^2$, and $Cov(\alpha_i, \beta_i) \equiv \gamma$, so that

$$y_i \equiv \alpha + \beta \cdot x_i + \varepsilon_i \quad ,$$

with

$$\varepsilon_i \equiv (\alpha_i - \alpha) + (\beta_i - \beta) \cdot x_i \quad ,$$

which has $E(\varepsilon_i) \equiv 0$ and

$$\begin{aligned} V(\varepsilon_i) &= \sigma^2 + 2\gamma \cdot x_i + \tau^2 \cdot x_i^2 \\ &= \sigma^2 \cdot (1 + \theta_1 \cdot x_i + \theta_2 \cdot x_i^2) \\ &\equiv \sigma^2 \cdot h_i \quad . \end{aligned}$$

3. Variance Proportional to Square of Mean

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i \quad , \quad \text{with}$$

$$V(\varepsilon_i) = \sigma^2(\alpha + \beta \cdot x_i)^2 \equiv \sigma^2 \cdot h_i \quad ,$$

so that a larger variance is associated with a larger mean.

Properties of Classical Least Squares Under Heteroskedasticity

- Least squares estimators of α and β still ***unbiased*** and ***consistent***;

- Least squares estimators no longer ***efficient***, i.e., they are no longer the best linear unbiased estimators; and

- Usual estimators for the standard errors of least squares are ***biased***, so the usual confidence intervals and test statistics are incorrect, and may lead to incorrect conclusions.

Approaches to Dealing with Heteroskedasticity

- For ***known*** heteroskedasticity (e.g., grouped data with known group sizes), use ***weighted least squares*** (WLS) to obtain efficient unbiased estimates;

- Test for heteroskedasticity of a special form using a ***squared residual regression***;

- Estimate the unknown heteroskedasticity parameters using this squared residual regression, then use the estimated variances in the WLS formula to get efficient estimates of regression coefficients (known as ***feasible WLS***); or

- Stick with the (inefficient) least squares estimators, but get estimates of standard errors which are correct under arbitrary heteroskedasticity.

Correction for Heteroskedasticity of Known Form

If

$$\mathrm{Var}(\varepsilon_i) = \sigma^2 \cdot h_i$$

where h_i is ***known*** (e.g., grouped data), then

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i \quad , \qquad \mathrm{Var}(\varepsilon_i) = \sigma^2 \cdot h_i$$

$$\Rightarrow \quad \frac{y_i}{\sqrt{h_i}} = \alpha \cdot \frac{1}{\sqrt{h_i}} + \beta \cdot \frac{x_i}{\sqrt{h_i}} + \frac{\varepsilon_i}{\sqrt{h_i}} \quad , \qquad \text{or}$$

$$y_i^* = \alpha \cdot z_i + \beta \cdot x_i^* + \varepsilon_i^* \ , \ \text{with} \ y_i^* \equiv \frac{y_i}{\sqrt{h_i}} \ , \ z_i \equiv \frac{1}{\sqrt{h_i}} \ ,$$

etc.

Since $\mathrm{Var}(\varepsilon_i^*) = \sigma^2$, can use Classical Least Squares on this transformed equation to get efficient estimates of α and β . For multiple regression model, divide the dependent variable and all of the regressors (including the constant term) by $\sqrt{h_i}$, then do least squares.

Weighted Least Squares

Regressing y_i^* on z_i and x_i^* involves minimization of

$$\sum_{i=1}^{n} \left(y_i^* - a \cdot z_i - b \cdot x_i^* \right)^2 = \sum_{i=1}^{n} \frac{(y_i - a - b \cdot x_i)^2}{h_i} \quad ;$$

thus, a more efficient estimator is obtained by ***downweighting*** the squared residuals for observations with large variances, in proportion to those variances.

Properties of Weighted Least Squares Estimates (with ***known*** weights)

- Estimated coefficients are ***efficient***, i.e., best linear unbiased (BLUE).

- Regression of y_i^* on z_i and x_i^* gives correct standard errors for coefficient estimates.

- R^2 must be redefined, since transformed model usually has no intercept term.

Detection of Heteroskedasticity (***unknown*** weights)

- **Residual plot**: Graph squared LS residuals

$$\hat{e}_i^2 = (y_i - \hat{\alpha} - \hat{\beta} \cdot x_i)^2$$

  against x_i or $\hat{y}_i = \hat{\alpha} + \hat{\beta} \cdot x_i$ to check variability.

- **Diagnostic testing**: Do formal statistical test for particular hypothesized form of h_i.

Squared Residual Regression Test for Heteroskedasticity

Conditions: Moderate to large samples (n = 50 - 100+), possibly nonnormal errors, and ***linear*** form for h_i ,

$$h_i = 1 + \theta_1 z_{i1} + \cdots + \theta_L z_{iL}$$
$$\Rightarrow \sigma_i^2 = \sigma^2 + \delta_i z_{i1} + \cdots + \delta_L z_{iL}$$
$$\Rightarrow \varepsilon_i^2 = \sigma^2 + \delta z_{i1} + \cdots + \delta_L z_{iL} + u_i \quad , \quad E(u_i) = 0 \ ,$$

where $z_{i1},\ldots,z_{iL}$ are ***known functions of regressors*** (e.g., $z_{i1} = x_i$, $z_{i2} = x_i^2$ for random coefficients model).

Idea: replace unknown squared errors ε_i^2 with squared residuals $\hat{e}_i^2 = (y_i - \hat{y}_i)^2$ from LS, then regress $\hat{e}_i^2$ on 1, $z_{i1},\ldots,z_{iL}$.

Steps:

(1) Get LS residuals $\hat{e}_i = y_i - \hat{\alpha} - \hat{\beta} \cdot x_i$.

(2) Get R^2 from regression of $\hat{e}_i^2$ on 1, $z_{i1},\ldots,z_{iL}$.

(3) Construct usual F-statistic

$$F = \left(\frac{R^2}{1 - R^2} \right) \cdot \left(\frac{n - L}{L - 1} \right) \quad ;$$

reject homoskedasticity if F exceeds critical value from F-table with L-1 and n-L degrees of freedom.

Correction for Heteroskedasticity: Feasible WLS

- **Conditions**: Same as for squared residual regression test, including

$$\sigma_i^2 = \sigma^2 + \delta_1 z_{i1} + \cdots + \delta_L z_{iL} \quad .$$

- **Idea**: Use squared residual regression to estimate weights.

- **Steps**:

  (1) Fit $y_i = \alpha + \beta \cdot x_i + \varepsilon_i$ by least squares, get $\hat{e}_i = y_i - \hat{\alpha} - \hat{\beta} \cdot x_i$.

  (2) Regress $\hat{e}_i^2$ on $1, z_{1i}, \ldots, z_{1L}$, get $\hat{\sigma}_i^2 = \hat{\sigma}^2 + \hat{\delta}_1 z_{i1} + \cdots + \hat{\delta}_L z_{iL}$.

  (3) Replace "h_i" with "$\hat{\sigma}_i^2$" in WLS formula--minimize $\sum_{i=1}^{N} \frac{(y_i - \alpha - \beta \cdot x_i)^2}{\hat{\sigma}_i^2}$ to estimate α, β .

- **Properties**: In large samples, (approximately) same as WLS.

Examples of Feasible WLS

1. Grouped Data

Regress $\sqrt{n_s}\,\bar{y}_s$ on $\sqrt{n_s}$, $\sqrt{n_s}\,\bar{x}_s$ to estimate α,β (***exact*** WLS).

2. Random Coefficients

First get LS estimates $\hat{\alpha}, \hat{\beta}$, and residuals $\hat{e}_i$. Since model has

$$\sigma_i^2 = \sigma^2 + \delta_1 x_i + \delta_2 x_i^2 \quad ,$$

regress $\hat{e}_i^2$ on 1, x_i, x_i^2, test H_0: $\delta_1 = \delta_2 = 0$ with F-test. If H_0 (homoskedasticity) rejected, let

$$\hat{\sigma}_i^2 = \hat{\sigma}^2 + \hat{\delta}_1 x_i + \hat{\delta}_2 x_i^2 \quad ,$$

plug into WLS formula. (For multiple regression, set $z_{i\ell}$ equal to squares and cross-products of regressors.)

3. Variance Proportional to Square of Mean

To test for heteroskedasticity, regress $\hat{e}_i^2 = (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$ on 1 and $z_{i1} = (\hat{y}_i)^2 = (\hat{\alpha} + \hat{\beta}x_i)^2$, do F-test (or t-test) for exclusion of $\hat{y}_i^2$. If homoskedasticity rejected, use "$\hat{h}_i = \hat{y}_i^2$" in place of "h_i" in WLS formula.

Correcting Least Squares Standard Errors for Heteroskedasticity

- **Situation**: Suspect heteroskedasticity but don't want to specify h_i ; willing to stick with (inefficient but unbiased) least squares coefficient estimators.

- **Idea**: Find formula for standard errors of LS which are valid under either homoskedasticity or heteroskedasticity. Known as ***Eicker-White*** (or just ***White***) heteroskedasticity-consistent standard errors.

- **Usual Variance Estimator for LS** $\hat{\beta}$ (page 60):

$$
\begin{aligned}
V(\hat{\beta}) &= s^2 / \sum x_i^2 && \text{if no intercept term;} \\
&= s^2 / \sum (x_i - \bar{x})^2 && \text{if intercept term;} \\
&= s^2 \left[\sum (x_i - \bar{x})^2\right] / \left[\sum (x_i - \bar{x})^2\right]^2 \quad . &&
\end{aligned}
$$

Formula ***assumes*** $E\left[\varepsilon_i^2 (x_i - \bar{x})^2\right] = \left[E(\varepsilon_i^2)\right] \cdot \left[E(x_i - \bar{x})^2\right]$, which ***fails*** under heteroskedasticity.

- **Corrected Variance Estimator for $\hat{\beta}$:**
  Now use

$$V(\hat{\beta}) = \left[\sum \hat{e}_i^2 (x_i - \bar{x})^2\right] / \left[\sum (x_i - \bar{x})^2\right]^2$$

  where $\hat{e}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$. Note

$$V(\hat{\beta}) \neq s^2 / \left[(x_i - \bar{x})^2\right]$$

  unless

$$\hat{e}_i^2 = s^2 = \frac{1}{n-k} \sum \hat{e}_i^2$$

  for all observations, which never happens in practice.

- **Formula for Multiple Regression:** Similar, but more complicated. Fortunately, many computer packages (e.g., TSP) compute "Eicker-White" standard errors as an option.