Chapter 8

Testing

8.1 Basic testing definitions

Think about the testing problem

$$H_0 : \theta = \theta_0 \text{ against } H_1 : \theta \neq \theta_0.$$

We call H_0 a simple null as it involves specifying the value of θ complete, while the alternative is called composite. There are four basic outcomes of this type of testing problem:

- H_0 is true, and we correctly accept it
- H_0 is true, but we reject it!
- H_0 is false, and we correctly reject the hypothesis
- H_0 is false, and we accept H_0.

Hence in this problem there are two types of errors and two type of correct decisions. This is presented in stylized form in Table 8.1.

	TRUTH	
	H_0	H_1
ACCEPT H_0	✓	× (Type 2 error)
ACCEPT H_1 (Significance level)	× (Type 1 error)	✓ (Power)

Table 8.1 *Stylised presentation of the testing problem. TRUTH denotes the state of nature, while ACCEPT denotes which hypothesis is accepted..*

8.1.1 Testing errors

If we base tests on random data then it is inevitable that we will sometimes reject the null hypothesis when it is true. What is important, is to know and control the chance of making this type of error. This problem, is called a Type 1 error. The probability of making an error of this type is called the significance level. Frequently this is written as

$$\Pr(\text{rejecting } H_0 | H_0 \text{ is true}) = \alpha,$$

where α is the significance level. In many econometric problems a conventional choice for α would be 0.05, although it is an arbitrary selection.

Given a particular significance level, it would be attractive if the other type of error, rejecting H_0 when it is false, was low. This type of error is called a Type 2 error.

8.1.2 Power functions

A counterpart of a Type 2 error, is the correct rejection of H_0. The probability of making this decision is called the power function

$$Power(\theta_1) = \Pr(\text{rejecting } H_0 | \text{true value of } \theta \text{ is } \theta_1).$$

The probability of a type 2 error is $1 - Power(\theta_1)$. Typically, for a given value of α it is a good thing to have the power function as high as possible for all values of $\theta_1 \neq \theta_0$. Such a test might be described as being powerful. If a particular test has $Power(\theta_1)$ higher than for any other test for a specific value of θ_1 then the test is described as being 'most powerful' for θ_1. If the test is most powerful for all values of θ_1, which are valid under the alternative, then the test is called uniformly most powerful (UMP).

Example

Suppose $Y_i \sim NID(\mu, 1)$, then $\sqrt{n}(\overline{Y} - \mu) \sim N(0, 1)$. Further, assume $H_0 : \mu = 5$, against an alternative $H_1 : \mu \neq 5$. Then

$$T = \sqrt{n}(\overline{Y} - 5) \underset{H_0}{\sim} N(0, 1),$$

and we reject H_0 when $|T| > C$ where the critical value C is such that $\Pr(|T| > C) = \alpha$. The power is the probability of $|T| > C$ under the alternative. So

$$T = \sqrt{n}(\overline{Y} - \mu) + \sqrt{n}\,(\mu - 5) \sim N(\sqrt{n}\,(\mu - 5)\,, 1).$$

So the power is a function of μ and increases above α whenever μ is not 5. We call the graph of power against μ the power function. If $\mu > 5$ then the mean of the normal goes to infinity as $n \to \infty$, while if $\mu < 5$ the mean goes to minus infinity. In both cases the power goes to one in the limit for all points under the alternative.

8.1.3 Neyman-Pearson lemma

For the very special case of the simple null $H_0 : \theta = \theta_0$ against the simple alternative $H_1 : \theta = \theta_1$ it is possible to derive the most powerful test (for a given size) using the Neyman-Pearson lemma. It states that the most powerful test of the null is based on the likelihood ratio

$$lr(y) = \frac{f(y; \theta_1)}{f(y; \theta_0)},$$

rejecting the null if $lr(y)$ is big. Recall one of the justifications of the likelihood was as a measure of plausibility, so this rejects the null if the data implies θ_1 is a more plausible parameter than θ_1. A proof of this theorem is given in, for example, Cox and Hinkley (1974, p. 92).

This setup can be generalized to the composite alternative $H_1 : \theta > \theta_1$. Use the Neyman-Pearson lemma on the alternative θ_1^*, which is bigger than θ_0. Then the likelihood ratio test is the most powerful. But this test results whatever the value of θ_1^* we choose as long as its bigger than θ_0. Hence the test is $lr(y)$ is uniformly most powerful against the composite alternative $H_1 : \theta > \theta_1$. Extending this result to more complicated problems is difficult and so this setup is only rarely used in econometrics.

Exercise

State and prove the Neyman-Pearson lemma. Use it to construct the uniformly most powerful test of $H_0 : \theta = 0$ against $H_1 : \theta > 0$, where $Y_i \sim NID(\theta, \sigma^2)$ for a known σ^2. Can you use it to derive a test of the more complicated form $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$?

Instead testing is dominated by three broad ways of constructing: likelihood ratios, Wald tests and score tests.

8.2 Likelihood ratio tests

As likelihood can be thought of as plausibility, we could look at the distance between θ and $\widehat{\theta}$, the best estimate under H_1, by the difference in likelihoods. Another motivation is via the Neyman-Pearson lemma. The test is

$$LR = 2\{\log L(\widehat{\theta}; y) - \log L(\theta_0; y)\} \underset{H_0}{\overset{d}{\to}} \chi^2 \dim(\theta).$$

This is convenient as it gives us a general method for composite alternatives and derives the required distribution under the null. We will work with $\dim(\theta) = 1$ and start-off with

$$\begin{aligned}\log L(\theta_0; y) &= \log L(\widehat{\theta}; y) + (\theta_0 - \widehat{\theta}) \left.\frac{\partial \log L}{\partial \theta}\right|_{\theta=\widehat{\theta}} + \frac{1}{2}(\theta_0 - \widehat{\theta})^2 \left.\frac{\partial^2 \log L}{\partial \theta^2}\right|_{\theta=\widehat{\theta}} + .. \\ &\simeq \log L(\widehat{\theta}; y) + \tfrac{1}{2}\{\sqrt{n}(\theta_0 - \widehat{\theta})\}^2 \frac{1}{n} \left.\frac{\partial^2 \log L}{\partial \theta^2}\right|_{\theta=\widehat{\theta}}.\end{aligned}$$

As $\widehat{\theta} \to \theta_0$, so

$$\frac{1}{n}\frac{\partial^2 \log L(\widehat{\theta}; Y)}{\partial \theta^2} \underset{H_0}{\overset{p}{\to}} -Var\left\{ \left.\frac{\partial \log L(\theta; Y_i)}{\partial \theta}\right|_{\theta=\theta_0} \right\},$$

minus the information per observation. The implication is that, as

$$\sqrt{n}(\theta_0 - \widehat{\theta}) \underset{H_0}{\overset{d}{\to}} N\left[0, Var\left\{ \left.\frac{\partial \log L(\theta; Y_i)}{\partial \theta}\right|_{\theta=\theta_0} \right\}^{-1}\right]$$

so

$$2\{\log L(\widehat{\theta}; y) - \log L(\theta_0; y)\} \underset{H_0}{\overset{d}{\to}} \chi^2_1.$$

In the more general case, the same result holds ($\chi^2_{\dim(\theta)}$) using quadratic of normals being χ^2 distributed.

8.3 Wald tests

Perhaps a more intuitively simple test can be based around the distribution of

$$\sqrt{n}(\widehat{\theta} - \theta_0) \underset{H_0}{\overset{d}{\to}} N\left[0, \left\{Var\left(\frac{\partial \log L(\theta; Y_i)}{\partial \theta}\right)\right\}^{-1}\right].$$

In the univariate case we could use the t-statistic version

$$\frac{\sqrt{n}(\widehat{\theta} - \theta_0)}{\left[Var\left\{\frac{\partial \log L(\theta; Y_i)}{\partial \theta}\right\}\right]^{-1/2}} \underset{H_0}{\overset{d}{\to}} N(0, 1)$$

but a more standard option is to work with a quadratic form version

$$\sqrt{n}(\widehat{\theta} - \theta_0)' \left[Var\left\{\frac{\partial \log L(\theta; Y_i)}{\partial \theta}\right\}\right] \sqrt{n}(\widehat{\theta} - \theta_0) \overset{d}{\to} \chi^2_1 \dim(\theta),$$

for this works nicely in the vector case.

8.4 Score or Lagrange multiplier test

Perhaps the least obvious of our three tests looks at how far the score

$$\left.\frac{\partial \log L(\theta; y)}{\partial \theta}\right|_{\theta=\theta_0}$$

is from zero (at $\widehat{\theta}$, the score is defined as zero). We know

$$\sqrt{n}\left\{\frac{1}{n}\frac{\partial \log L(\theta_0; Y)}{\partial \theta}\right\} \overset{d}{\to} N\left[0, \left\{Var\frac{\partial \log L(\theta_0; Y_i)}{\partial \theta}\right\}\right]$$

so consequently the obvious test is to work with

$$\left\{\sqrt{n}\frac{1}{n}\frac{\partial \log L(\theta_0; y)}{\partial \theta}\right\}' \left\{Var\frac{\partial \log L(\theta_0; Y_i)}{\partial \theta}\right\}^{-1} \left\{\sqrt{n}\frac{1}{n}\frac{\partial \log L(\theta_0; y)}{\partial \theta}\right\} \underset{H_0}{\overset{d}{\to}} \chi^2_1 \dim(\theta).$$

Example

Here we construct the score test for the Gaussian first order autoregression (written AR(1)) case $y_t = \theta y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim NID(0, 1)$ where $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$. In particular we use the conditional likelihood $\log f(y_2, ..., y_n | y_1; \theta)$ which equals

$$\sum_{t=2}^{n} \log f(y_t | y_{t-1}, ..., y_1; \theta) = const - \frac{1}{2}\sum_{t=2}^{n}(y_t - \theta y_{t-1})^2.$$

The score is

$$\begin{aligned}\frac{\partial \log f(y_2, ..., y_n | y_1; \theta)}{\partial \theta} &= \sum_{t=2}^{n} y_{t-1}(y_t - \theta y_{t-1}) \\ &\underset{H_0}{=} \sum_{t=2}^{n} y_{t-1} y_t\end{aligned}$$

and the observed information is

$$-\sum_{t=2}^{n} y^2_{t-1}.$$

The information, in the sample, is

$$-E\sum_{t=2}^{n} y^2_{t-1} \underset{H_0}{=} -(n-1).$$

As a result the score test equals

$$\begin{aligned}S &= \left(\sqrt{n-1}\frac{1}{n-1}\sum_{t=2}^{n} y_{t-1}y_t\right)(1)^{-1}\left(\sqrt{n-1}\frac{1}{n-1}\sum_{t=2}^{n} y_{t-1}y_t\right) \\ &= \frac{1}{n-1}\left(\sum_{t=2}^{n} y_t y_{t-1}\right)^2 \underset{H_0}{\simeq} n\left(\frac{\sum_{t=2}^{n} y_t y_{t-1}}{\sum_{t=2}^{n} y^2_{t-1}}\right)^2 \underset{H_0}{\to} \chi^2_1.\end{aligned}$$

We note that $\sum_{t=2}^{n} y_t y_{t-1} / \sum_{t=2}^{n} y^2_{t-1}$ is, roughly, the correlation between $\{y_t, y_{t-1}\}$. This is called the serial correlation coefficient at lag one and written as r_1.

Example

We construct the score test for $y_t = \mu + \theta y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim NID(0, 1)$ where $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$, using the conditional likelihood

$$\sum_{t=2}^{n} \log f(y_t | y_{t-1}, ..., y_1; \theta, \mu) = const - \frac{1}{2}\sum_{t=2}^{n}(y_t - \mu - \theta y_{t-1})^2.$$

The score is

$$\begin{aligned}\left.\tfrac{\partial \log f(y_2, ..., y_n | y_1; \theta, \mu)}{\partial \theta}\right|_{\theta=0} &= \textstyle\sum_{t=2}^{n}(y_t - \mu) y_{t-1} \\ \left.\tfrac{\partial \log f(y_2, ..., y_n | y_1; \theta, \mu)}{\partial \mu}\right|_{\theta=0} &= \textstyle\sum_{t=2}^{n}(y_t - \mu).\end{aligned}$$

Then $\widetilde{\mu} = \overline{y}$, is the MLE under the constraint of the null. The second derivative matrix is

$$-\left\{\begin{array}{cc}\sum_{t=2}^{n} y^2_{t-1} & \sum_{t=2}^{n} y_{t-1} \\ \sum_{t=2}^{n} y_{t-1} & (n-1)\end{array}\right\}.$$

This has an expected value of

$$-\left\{\begin{array}{cc}(n-1) & 0 \\ 0 & (n-1)\end{array}\right\}.$$

So the score test is

$$\begin{aligned} S &= \left\{\sqrt{n-1}\frac{1}{n-1}\sum_{t=2}^{n} y_{t-1}\left(y_t-\overline{y}\right)\right\}(1)^{-1}\left\{\sqrt{n-1}\frac{1}{n-1}\sum_{t=2}^{n} y_{t-1}\left(y_t-\overline{y}\right)\right\} \\ &= \frac{1}{n-1}\left\{\sum_{t=2}^{n} y_{t-1}\left(y_t-\overline{y}\right)\right\}^2 \\ &= \frac{1}{n-1}\left\{\sum_{t=2}^{n}\left(y_{t-1}-\overline{y}\right)\left(y_t-\overline{y}\right)\right\}^2 \underset{H_0}{\simeq} n\left\{\frac{\sum_{t=2}^{n}\left(y_{t-1}-\overline{y}\right)\left(y_t-\overline{y}\right)}{\sum_{t=2}^{n}\left(y_{t-1}-\overline{y}\right)^2}\right\}^2 \underset{H_0}{\rightarrow} \chi_1^2. \end{aligned}$$

Exercise

Construct the score test for the dynamic regressior model $y_t = x_t'\beta + \theta y_{t-1} + \varepsilon_t$, $\varepsilon_t \sim NID(0,1)$ where $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$. We regard the regressors as fixed in this exercise. Use the conditional likelihood

$$\sum_{t=2}^{n} \log f(y_t|y_{t-1},...,y_1;\theta,\beta) = const - \frac{1}{2}\sum_{t=2}^{n}\left(y_t - x_t'\beta - \theta y_{t-1}\right)^2.$$

Example

Construct the score test for the Gaussian AR(p) process

$$y_t = \theta_1 y_{t-1} + ... + \theta_p y_{t-p} + \varepsilon_t, \varepsilon_t \sim NID(0,1),$$

where $H_0 : \theta_1 = ... = \theta_p = 0$. The test should be based on the conditional log-likelihood function $\log f(y_{p+1},...,y_n|y_1,...,y_p;\theta)$. This has the form of

$$\sum_{t=p+1}^{n} \log f(y_t|y_{t-1},...,y_1;\theta) = const - \frac{1}{2}\sum_{t=p+1}^{n}\left(y_t - \theta_1 y_{t-1} - ... - \theta_p y_{t-p}\right)^2.$$

The score vector is

$$\begin{aligned} \frac{\partial \log f(y_{p+1},...,y_n|y_1,...,y_p;\theta)}{\partial\theta} &= \begin{pmatrix} \sum_{t=p+1}^{n}\left(y_t - \theta_1 y_{t-1} - ... - \theta_p y_{t-p}\right) y_{t-1} \\ \sum_{t=p+1}^{n}\left(y_t - \theta_1 y_{t-1} - ... - \theta_p y_{t-p}\right) y_{t-2} \\ \vdots \\ \sum_{t=p+1}^{n}\left(y_t - \theta_1 y_{t-1} - ... - \theta_p y_{t-p}\right) y_{t-p} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{t=p+1}^{n} y_t y_{t-1} \\ \sum_{t=p+1}^{n} y_t y_{t-2} \\ \vdots \\ \sum_{t=p+1}^{n} y_t y_{t-p} \end{pmatrix}. \end{aligned}$$

The second derivative matrix is

$$\frac{\partial^2 \log f(y_{p+1},...,y_n|y_1,...,y_p;\theta)}{\partial\theta\partial\theta'} = -\begin{pmatrix} \sum_{t=p+1}^{n} y_{t-1}^2 & \sum_{t=p+1}^{n} y_{t-1}y_{t-2} & \cdots & \sum_{t=p+1}^{n} y_{t-1}y_{t-p} \\ \sum_{t=p+1}^{n} y_{t-1}y_{t-2} & \sum_{t=p+1}^{n} y_{t-2}^2 & \cdots & \sum_{t=p+1}^{n} y_{t-2}y_{t-p} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{t=p+1}^{n} y_{t-1}y_{t-p} & \sum_{t=p+1}^{n} y_{t-2}y_{t-p} & \cdots & \sum_{t=p+1}^{n} y_{t-p}^2 \end{pmatrix}.$$

The expectation of this matrix is, under the null $(n-p)I$ as $E(y_t y_{t-p}) = 0$ for all p not equal to zero and $E(y_t^2) = 1$. Then the score test is

$$\begin{aligned} S &= \frac{1}{(n-p)}\left\{\left(\sum_{t=p+1}^{n} y_{t-1}y_{t-2}\right)^2 + ... + \left(\sum_{t=p+1}^{n} y_{t-1}y_{t-p}\right)^2\right\} \end{aligned}$$

$$\begin{aligned} &\simeq n\left\{\left(\frac{\sum_{t=p+1}^{n} y_{t-1}y_{t-2}}{\sum_{t=p+1}^{n} y_{t-1}^2}\right)^2 + ... + \left(\frac{\sum_{t=p+1}^{n} y_{t-1}y_{t-p}}{\sum_{t=p+1}^{n} y_{t-1}^2}\right)^2\right\} \\ &= n\left(r_1^2 + ... + r_p^2\right) \rightarrow \chi_p^2. \end{aligned}$$

8.5 Comparisons

All the tests have the same asymptotic distribution under the null. In some sense the LR test must be the preferred one as it evaluates the likelihood under H_0 and H_1. However, the Wald and score are simpler.

LR	evaluates θ_0 and $\widehat{\theta}$	evaluates under $H_0 + H_1$
Wald	computes $\widehat{\theta}$	evaluate under H_1
Score	uses just θ_0	evaluate under H_0

From a computational viewpoint the score is most useful as it does not need to evaluate $\widehat{\theta}$, which for complicated models can be hard.

Exercise

Construct the score, Wald and likelihood ratio tests for $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$, where $Y_i \sim NID(\theta, \sigma^2)$ for a known value of σ^2. Compare the exact distribution of these tests to their asymptotic approximations. How do these tests change when σ^2 is unknown?

Finally, these arguments are based on asymptotic theory and so tend to be poor if n is small. Improvements, via higher order expansions, are possible. In the LR case they are particularly easy, using the Bartlett adjustment. The score test can be corrected using a similar but more complicated argument, while the Wald test is hard to correct. This type of work is discussed in Barndorff-Nielsen and Cox (1994).

8.6 Tests of the multiple linear regression model

8.6.1 Testing linear hypotheses about $\boldsymbol{\beta}$

The general linear framework is:

$$\boldsymbol{R\beta} = \boldsymbol{r}$$

where $\boldsymbol{R}$ is a $(q \times k)$ matrix of known constants, with $q < k$ and $\boldsymbol{r}$ is a $(q \times 1)$ vector of known constants. This can be stated as

$$H_0 : \ \boldsymbol{R\beta} - \boldsymbol{r} = \boldsymbol{0}.$$

Examples of typical hypotheses are:

- $H_0 : \ \beta_i = 0$; X_i has no influence on Y.

$$\boldsymbol{R} = \begin{pmatrix} 0 & ... & 0 & 1 & 0 & ... & 0 \end{pmatrix} \quad r = 0 \quad q = 1$$

- $H_0 : \ \beta_i = \beta_{i0}$; where β_{i0} has a specific value

$$\boldsymbol{R} = \begin{pmatrix} 0 & ... & 0 & 1 & 0 & ... & 0 \end{pmatrix} \quad r = \beta_{i0} \quad q = 1$$

  with 1 in the ith position
- $H_0 : \ \beta_2 + \beta_3 = 1$; e.g. constant returns to scale for $\beta_2 = production$ and $\beta_3 = labour$ elasticities in a production function

$$\boldsymbol{R} = \begin{pmatrix} 0 & 1 & 1 & 0 & ... & 0 \end{pmatrix} \quad r = 1 \quad q = 1$$

- $H_0: \beta_3 = \beta_4$ or $\beta_3 - \beta_4 = 0$; X_3 and X_4 have the same coefficient.

$$\boldsymbol{R} = \begin{pmatrix} 0 & 0 & 1 & -1 & 0 & \ldots & 0 \end{pmatrix} \quad r = 0 \quad q = 1$$

- $H_0: \beta_j = 0$ for $j = 2, ..., k$; tests the significance of the overall relation.

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{0} & \boldsymbol{I}_{k-1} \end{pmatrix} \quad r = \boldsymbol{0} \quad q = k-1$$

- $H_0: \boldsymbol{\beta}_2 = 0$; a specified subset of regressors plays no role in determining Y. where $\boldsymbol{0}$ is a vector of $k-1$ zeros

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{0}_{k_2 \times k_1} & \boldsymbol{I}_{k_2} \end{pmatrix} \quad r = \boldsymbol{0} \quad q = k_2.$$

Since $\mathrm{E}(\boldsymbol{R}\hat{\boldsymbol{\beta}}) = \boldsymbol{R}\boldsymbol{\beta}$ and $\mathrm{var}(\boldsymbol{R}\hat{\boldsymbol{X}})^{-1}\boldsymbol{R}'$, (check), on making an assumption on the form of the sampling distribution of $\boldsymbol{u}$, we can obtain the sampling distribution of $\boldsymbol{R}\hat{\boldsymbol{\beta}}$. Suppose

$$\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

then

$$\begin{aligned} \hat{\boldsymbol{\beta}} &\sim N\{\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\} \\ \boldsymbol{R}\hat{\boldsymbol{\beta}} &\sim N\{\boldsymbol{R}\boldsymbol{\beta}, \sigma^2 \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\} \\ \boldsymbol{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &\sim N\{\boldsymbol{0}, \sigma^2 \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\}. \end{aligned}$$

If H_0 is true then

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) \sim N\{\boldsymbol{0}, \sigma^2 \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\}$$

verify that

$$(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'\{\sigma^2 \boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\}^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}) \sim \chi^2_q.$$

Since

$$\frac{\boldsymbol{e}'\boldsymbol{e}}{\sigma^2} \sim \chi^2_{n-k},$$

under the Null hypothesis

$$\frac{(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})'\{\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\}^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})/q}{\boldsymbol{e}'\boldsymbol{e}/(n-k)} \sim F_{q,n-k} \tag{8.1}$$

and the null hypothesis is rejected if the computed F value exceeds a preselected *critical* value.

Example

Suppose we wish to test $H_0: \beta_i = 0$, then equation (8.1) becomes

$$F = \frac{\hat{\beta}_i^2}{\hat{\sigma}^2 c_{ii}} = \frac{\hat{\beta}_i^2}{\mathrm{var}(\hat{\beta}_i)} \sim F_{1,n-k}$$

where c_{ii} is the element picked out by $\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'$ and $\boldsymbol{R}\hat{\boldsymbol{\beta}}$ picks out $\hat{\beta}_i$. Taking the square root

$$t = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{(c_{ii})}} = \frac{\hat{\beta}_i}{\mathrm{s.e.}(\hat{\beta}_i)} \sim t_1$$

a t-test. *A single hypothesis allows either the t or F tests to be implemented.*

Example: Composite hypothesis

To test $H_0: \beta_2 = \beta_3 = ... = \beta_k = 0$, first partition $\boldsymbol{X}$ as $[\boldsymbol{i}\ \boldsymbol{X}_2]$, where $\boldsymbol{i}$ is the unit vector and $\boldsymbol{X}_2$ contains the $k-1$ regressor coefficients.

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} \boldsymbol{i}' \\ \boldsymbol{X}_2' \end{pmatrix} \begin{pmatrix} \boldsymbol{i} & \boldsymbol{X}_2 \end{pmatrix} = \begin{pmatrix} n & \boldsymbol{i}'\boldsymbol{X}_2 \\ \boldsymbol{X}_2' & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{pmatrix}$$

Using the formula for inverted partitioned matrices, we can find the $k-1$ submatrix picked out by $\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'$ and express it as

$$\begin{pmatrix} \boldsymbol{X}_2'\boldsymbol{X}_2 - \boldsymbol{X}_2'\boldsymbol{i}n^{-1}\boldsymbol{i}'\boldsymbol{X}_2 \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{X}_2'\boldsymbol{A}\boldsymbol{X}_2 \end{pmatrix}^{-1} = \begin{pmatrix} \boldsymbol{X}_*'\boldsymbol{X}_* \end{pmatrix}^{-1}$$

where $\boldsymbol{A}$ transforms observations into deviation form defined by $\boldsymbol{A} = \boldsymbol{I}_n - (1/n)\boldsymbol{i}\boldsymbol{i}'$, and $\boldsymbol{X}_* = \boldsymbol{A}\boldsymbol{X}_2$. Since $\boldsymbol{R}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_2$, the numerator of (8.1) is $\hat{\boldsymbol{\beta}}_2\boldsymbol{X}_*'\boldsymbol{X}_*\hat{\boldsymbol{\beta}}_2$ which is just the ESS. So the F statistic for testing the *joint significance of the complete set of regressors* is

$$F = \frac{\mathrm{ESS}/(k-1)}{\mathrm{RSS}/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim F_{k-1,n-k}.$$

8.6.2 Restricted and unrestricted regressions

In *restricted* regression, restrictions in H_0 are imposed on *estimated* equation.

Example: fitting restricted regression

The regression in deviation form is

$$y_i = \hat{\beta}_2 x_{i,2} + \hat{\beta}_3 x_{i,3} + e_i$$

and we impose $\beta_2 + \beta_3 = 1$. This gives

$$y_i - x_{i,3} = \hat{\beta}_2(x_{i,2} - x_{i,3}) + e_{*i}$$

forming new variables $(y_i - x_{i,3})$ and $(x_{i,2} - x_{i,3})$ where regression of first variable on second (without intercept term) gives restricted estimate of $\hat{\beta}_2$. Restricted RSS is denoted by $\boldsymbol{e}_*'\boldsymbol{e}_*$. General approach: requires $\hat{\boldsymbol{\beta}}_*$ vector that minimizes RSS subject to *restrictions*, $\boldsymbol{R}\hat{\boldsymbol{\beta}}_* - \boldsymbol{r}$. Set up *Lagrangian* function

$$\phi = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_*)'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_*) - 2\boldsymbol{\lambda}'(\boldsymbol{R}\hat{\boldsymbol{\beta}}_* - \boldsymbol{r})$$

where $\boldsymbol{\lambda}$ is $(q \times 1)$ vector of Lagrange multipliers. Solving F.O.C's gives

$$\hat{\boldsymbol{\beta}}_* = \hat{\boldsymbol{\beta}} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\{\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\}^{-1}(\boldsymbol{r} - \boldsymbol{R}\hat{\boldsymbol{\beta}}).$$

The residuals from the restricted and unrestricted regressions can be used to obtain the test statistic by noting that

$$\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e} = (\boldsymbol{r} - \boldsymbol{R}\hat{\boldsymbol{\beta}})'\{\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\}^{-1}(\boldsymbol{r} - \boldsymbol{R}\hat{\boldsymbol{\beta}})$$

so that,

$$F = \frac{(\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e})/q}{\boldsymbol{e}'\boldsymbol{e}/(n-k)} \sim F_{q,n-k}.$$

8.6.3 Prediction

A *Point prediction* of Y conditional on some information (e.g. newly observed X's), is given by

$$\hat{Y}_f = \hat{\beta}_1 + \hat{\beta}_2 X_{2,f} + ... + \hat{\beta}_k X_{k,f} = \boldsymbol{c}'\hat{\boldsymbol{\beta}}$$

which is an *optimal* predictor of $\mathrm{E}(Y_f)$, (by Gauss Markov). To obtain a confidence interval, suppose $Y_f = \boldsymbol{c}'\boldsymbol{\beta} + u_f$, then

$$e_f = Y_f - \hat{Y}_f = u_f - \boldsymbol{c}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

The t-statistic is

$$\frac{\hat{Y}_f - Y_f}{s\sqrt{1 + \boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}}} \sim t_{n-k}$$

since $\mathrm{var}(e_f) = \sigma^2\{1 + \boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}\}$.

8.6.4 Are assumptions underlying OLS valid for given a data set?

If any of the underlying assumptions are wrong, then there is a *specification error*. These include possible problems with $\boldsymbol{u}$, $\boldsymbol{X}$ and/or $\boldsymbol{\beta}$.

8.6.4.1 Problems with u

(1) If $u_i \sim iid(0, \sigma^2)$, but not normal, then BLUE properties still hold, but inference procedures only asymptotically valid.
(2) *Heteroscedasticity*: $\mathrm{E}(\boldsymbol{uu}') = \mathrm{diag}(\sigma_1^2, ..., \sigma_n^2)$. Assumption of homoscedasticity violated.
(3) *Autocorrelated* disturbances: $\mathrm{E}(u_t u_{t-s}) \neq 0$ for $s \neq 0$.

8.6.4.2 Problems with E(Y | X) and/or X

(1) *Specification problem*: exclusion of relevant variables.
(2) Inclusion of irrelevant variables.
(3) Incorrect functional form.
(4) $\boldsymbol{X}$ matrix has less than full column rank: *collinearity* problem.
(5) $\mathrm{E}(X_{i,t} u_s) \neq 0$: if lagged values of dependent variable appear as regressors that are correlated with past disturbances but not current or future disturbances, then OLS is biased but consistent and asymptotically normal. If a regressor is correlated with current disturbance (e.g. measurement error in regressor) then OLS is both biased and inconsistent.
(6) Nonstationary variables: inference procedures are nonstandard, e.g. cointegration, integrated variables, error correction models.

8.6.4.3 Problems with β

Structural breaks: $\boldsymbol{\beta}$ is not constant over the sample period.

8.6.5 Iterative approach to model building

Three stage procedure based on identification, estimation and diagnostic tests. See for example, Box, Jenkins and Reinsel (1994) and for diagnostic tests, see Doornik and Hendry (1994). Aim is to find a model that is "approximately" correct.

(1) *Identification*: use of data and information on how series is generated in order to suggest parsimonious class of models to be considered. For example, one could formulate a model, look at scatter plots, residuals, correlograms, etc.
(2) *Estimation*: efficient use of data to make inferences about parameters conditional on adequacy of model being considered. Are the coefficients reasonable; is the relationship *statistically significant*?
(3) *Diagnostic checking*: compare fitted model to data with intent to reveal model inadequacies; improve model. Look at predictive performance etc.

8.6.6 Tests of parameter constancy

$\boldsymbol{\beta}$ should apply both outside and within sample data: test predictive accuracy.

8.6.6.1 Chow forecast test

'Large' prediction errors cast doubt on constancy hypothesis. Divide data set of n observations into n_1 to be used for estimation and $n_2 = n - n_1$ to be used for testing. Partition $\boldsymbol{X}$ and $\boldsymbol{y}$ into $[\boldsymbol{X}_1,\ \boldsymbol{X}_2]$ and $[\boldsymbol{y}_1,\ \boldsymbol{y}_2]$.

(1) Estimate OLS from n_1 obs, obtaining
$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}_1.$$
(2) Obtain prediction of $\boldsymbol{y}_2$ vector, $\hat{\boldsymbol{y}}_2 = \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_1$.
(3) Obtain vector of prediction errors, $\boldsymbol{d}$, and analyze sampling distribution under H_0: parameter constancy.

Suppose vector of prediction errors is
$$\boldsymbol{d} = \boldsymbol{y}_2 - \hat{\boldsymbol{y}}_2 = \boldsymbol{y}_2 - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_1,$$
then if $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, with $\mathrm{E}(\boldsymbol{uu}') = \sigma^2\boldsymbol{I}$ holds for both data sets then,
$$\boldsymbol{d} = \boldsymbol{y}_2 - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_1 = \boldsymbol{u}_2 - \boldsymbol{X}_2(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta})$$
It can be shown that $\mathrm{E}(\boldsymbol{d}) = 0$ and $\mathrm{var}(\boldsymbol{dd}') = \sigma^2\{\boldsymbol{I}_{n_2} + \boldsymbol{X}_2(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_2'\}$. If we assume $\boldsymbol{d} \sim N\{\boldsymbol{0}, \mathrm{var}(\boldsymbol{d})\}$, then under the hypothesis of parameter constancy,
$$F = \frac{\boldsymbol{d}'\{\boldsymbol{I}_{n_2} + \boldsymbol{X}_2(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_2'\}^{-1}\boldsymbol{d}/n_2}{\boldsymbol{e}_1'\boldsymbol{e}_1/(n_1 - k)} \sim F_{n_2, n_1 - k}$$
where $\boldsymbol{e}_1'\boldsymbol{e}_1$ is the RSS from the estimated regression.

Note: one can also obtain the test in terms of the restricted and unrestricted regression.

8.6.6.2 Hansen test

A difficulty with the Chow test is that the null hypothesis may be rejected for for certain partitionings and not for others. The Hansen test fits the linear equation to all n observations and so avoids this problem. These consist of tests for stability of each parameter and of overall parameter stability. OLS fit gives
$$\begin{aligned}\sum_{t=1}^{n} x_{i,t} e_t &= 0\\ \sum_{t=1}^{n}(e_t^2 - \hat{\sigma}^2) &= 0\end{aligned}$$
where $\hat{\sigma}^2 = \sum_{t=1}^{n} e_t^2/n$. Defining
$$f_{i,t} = \begin{cases} x_{i,t}e_t & i = 1, ..., k\\ e_t^2 - \hat{\sigma}^2 & i = k+1\end{cases}$$
gives $\sum_{t=1}^{n} f_{i,t} = 0$ for $i = 1, ..., k+1$. Hansen test is based on *cumulative sums*
$$S_{i,t} = \sum_{j=1}^{t} f_{i,j}.$$
Individual test statistics are
$$L_i = \frac{1}{nV_i}\sum_{t=1}^{n} S_{i,t}^2$$
where $V_i = \sum_{t=1}^{n} f_{i,t}^2$ and the test for joint stability is
$$L_c = \frac{1}{n}\sum_{t=1}^{n} s_t' V^{-1} s_t$$
where
$$\begin{aligned} s_t &= \begin{pmatrix} S_{1,t} & \ldots & S_{k+1,t}\end{pmatrix}'\\ f_t &= \begin{pmatrix} f_{1,t} & \ldots & f_{k+1,t}\end{pmatrix}'\\ V &= \sum_{t=1}^{n} f_t f_t'.\end{aligned}$$
Under null hypothesis, the cumulative sums will tend to be distributed around zero, so that 'large' values of the test statistics suggest rejecting H_0.

8.6.6.3 Test based on recursive estimation

Write the model as $y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + u_t$ where $\boldsymbol{x}_t' = [1\ x_{2,t}\ ...\ x_{k,t}]$ so that the observations are assumed to be ordered over time. Fit model to first k observations, next use $k+1$ observations and compute coefficient vector again. Proceed for all n observations to generate sequence of vectors $\hat{\beta}_k, \hat{\beta}_{k+1}, ..., \hat{\beta}_n$. Standard errors can be computed at each stage except for the first (RSS is zero for $t = k$). Plots of the parameters (plus and minus two s.e's) can be analyzed for parameter constancy.

8.6.6.4 One step-ahead prediction errors

The one step-ahead prediction error is defined as

$$v_t = y_t - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}_{t-1}$$

with variance

$$\text{var}(v_t) = \sigma^2\{1 + \boldsymbol{x}_t'(\boldsymbol{X}_{t-1}'\boldsymbol{X}_{t-1})^{-1}\boldsymbol{x}_t\}.$$

σ^2 can be replaced by the residual variance estimated from $(t-1)$ observations, provided $t-1 > k$. The square root gives the estimated s.e. of regression. Plus or minus twice these can be plotted around the zero line with the actual prediction errors - residuals outside s.e. bands are suggestive of parameter inconstancy.

See Johnston and Dinardo (1997, pp. 119-121), for CUSUM, CUSUMSQ tests and the Ramsey Reset test.

8.6.7 Tests of structural change

A structural change or break occurs if parameters underlying specified relationship differs between different subsets of data. Suppose $n = n_1 + n_2$ and that we have data $\boldsymbol{X}_i, \boldsymbol{y}_i$ for $i = 1, 2$.

8.6.7.1 Example: three formulations for testing one structural change

Unrestricted model is:

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \boldsymbol{u} \quad \text{where } \boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$$

Fitting this equation gives the unrestricted RSS, $\boldsymbol{e}'\boldsymbol{e}$. The null hypothesis of no structural break is $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$. Writing OLS coefficients as

$$\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{y}_1 \\ \boldsymbol{X}_2'\boldsymbol{y}_2 \end{pmatrix} = \begin{pmatrix} (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}_1 \\ (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y}_2 \end{pmatrix}$$

we can either estimate the model by running the OLS estimation once, or by fitting each equation separately. The unrestricted RSS can be obtained as $\boldsymbol{e}'\boldsymbol{e} = \boldsymbol{e}_1'\boldsymbol{e}_1 + \boldsymbol{e}_2'\boldsymbol{e}_2$.

The restricted model can be written as

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \boldsymbol{\beta} + \boldsymbol{u}$$

giving an alternative formulation for testing $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ as

$$F = \frac{(\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e})/k}{\boldsymbol{e}'\boldsymbol{e}/(n-2k)} \sim F_{k,n-2k}.$$

It is also possible to consider an alternative setup of the unrestricted model

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1 & \boldsymbol{0} \\ \boldsymbol{X}_2 & \boldsymbol{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1 \end{pmatrix} + \boldsymbol{u}$$

so that testing H_0 is equivalent to testing the joint significance of the k regressors.

8.6.7.2 Tests for intercepts, slope coefficients and parameters

Suppose we define

$$\boldsymbol{X}_1 = \begin{pmatrix} \boldsymbol{i}_1 & \boldsymbol{X}_1^* \end{pmatrix} \text{ and } \boldsymbol{X}_2 = \begin{pmatrix} \boldsymbol{i}_2 & \boldsymbol{X}_2^* \end{pmatrix}$$

where $\boldsymbol{i}_1, \boldsymbol{i}_2$ are n_1 and n_2 vector of ones and the $\boldsymbol{X}_i^*$ are matrices of the $k-1$ regressor variables. We can consider three types of models:

$$I: \quad \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{i}_1 & \boldsymbol{X}_1^* \\ \boldsymbol{i}_2 & \boldsymbol{X}_2^* \end{pmatrix} \begin{pmatrix} \alpha \\ \boldsymbol{\beta}^* \end{pmatrix} + \boldsymbol{u} \qquad \text{common parameters}$$

$$II: \quad \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{i}_1 & \boldsymbol{0} & \boldsymbol{X}_1^* \\ \boldsymbol{0} & \boldsymbol{i}_2 & \boldsymbol{X}_2^* \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \boldsymbol{\beta}^* \end{pmatrix} + \boldsymbol{u} \qquad \text{Differential intercepts, common slope vectors}$$

$$III: \quad \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{i}_1 & \boldsymbol{0} & \boldsymbol{X}_1^* & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{i}_2 & \boldsymbol{0} & \boldsymbol{X}_2^* \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \boldsymbol{\beta}_1^* \\ \boldsymbol{\beta}_2^* \end{pmatrix} + \boldsymbol{u} \qquad \text{Differential intercepts, differential slope parameters}$$

Application to each will yield RSS with the associated degree of freedom $n-k$, $n-k-1$ and $n-2k$. The test statistics for various hypotheses are then given by

- Test of differential intercepts $H_0 : \alpha_1 = \alpha_2$

$$F = \frac{\text{RSS}_1 - \text{RSS}_2}{\text{RSS}_2/(n-k-1)} \sim F_{1,n-k-1}$$

- Test of differential slope vectors $H_0 : \boldsymbol{\beta}_1^* = \boldsymbol{\beta}_2^*$

$$F = \frac{(\text{RSS}_2 - \text{RSS}_3)/(k-1)}{\text{RSS}_3/(n-2k)} \sim F_{k-1,n-2k}$$

- Test of differential parameters (intercepts and slopes) $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$

$$F = \frac{(\text{RSS}_1 - \text{RSS}_3)/k}{\text{RSS}_3/(n-2k)} \sim F_{k,n-2k}$$

The degrees of freedom (d.o.f.) in the numerators are the number of restrictions imposed ingoing from the unrestricted model to the restricted one; which is also equal to the difference in the d.o.f. of the RSS in the numerator.

8.7 Heteroscedasticity and autocorrelation

Additional references on these issues and others that will not be covered in this course can be found in White (1984) and White (1998). When heteroscedasticity is present, (typically in cross-sectional data), the disturbance vector is

$$\text{var}(\boldsymbol{u}) = \text{E}(\boldsymbol{u}\boldsymbol{u}') = \begin{pmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \sigma_n^2 \end{pmatrix} = \boldsymbol{V}.$$

There are now $n+k$ unknowns; n unknown variances and k elements in the $\boldsymbol{\beta}$ vector. Additional assumptions (usually made from the disturbance process) are needed in order to estimate the model. One could postulate that

$$\sigma_i^2 = \sigma^2 x_{2,i} \text{ for } i = 1, 2, ..., n$$

where σ^2 is a scale factor and x_2 is an explanatory variable. Then

$$\text{var}(\boldsymbol{u}) = \text{E}(\boldsymbol{u}\boldsymbol{u}') = \sigma^2 \begin{pmatrix} x_{2,1} & 0 & ... & 0 \\ 0 & x_{2,2} & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & x_{2,n} \end{pmatrix} = \sigma^2\boldsymbol{\Omega}.$$

This reduces the parameters to be estimated to $k+1$, but the assumption made on the disturbances is very strong; it is important to test for heteroscedasticity and, if found, to explore its structure in order to derive feasible GLS estimators.

8.7.1 Properties of OLS estimators

The specified equation is

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u} \text{ with } \mathrm{E}(\boldsymbol{u}) = \boldsymbol{0} \text{ and } \mathrm{E}(\boldsymbol{uu}') = \sigma^2\boldsymbol{\Omega}.$$

If $\boldsymbol{X}$ is nonstochastic, then the following hold,

- The OLS estimator is unbiased and consistent, (mean square consistent if the variance matrix, var$(\hat{\boldsymbol{\beta}})$ has a zero plim).
- OLS estimator is *inefficient*. That is, linear, unbiased but not minimum variance estimators.
- OLS coefficient s.e's are incorrect, and the test statistics based on these are invalid.[1]

The variance matrix can be expressed as

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = \underbrace{\frac{\sigma^2}{n}}_{\xrightarrow{p}0}\underbrace{\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1}}_{(1)}\underbrace{\left(\frac{\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}}{n}\right)}_{(2)}\left(\frac{\boldsymbol{X}'\boldsymbol{X}}{n}\right)^{-1}.$$

Consistency requires both (1) and (2) to have finite plims. (1) $\xrightarrow{p}$ finite matrix, if the regressors are stationary. (2) $\xrightarrow{p}$ finite in general, if elements of $\boldsymbol{\Omega}$ are finite. If the $\boldsymbol{X}$ matrix contains one or more lags of the dependent variable, then the OLS estimator will have a finite sample bias; but will be consistent if $\boldsymbol{V}$ is diagonal. Autocorrelated disturbances cause off-diagonal elements in $\boldsymbol{V}$ to be non-zero; combined with the $\boldsymbol{X}$ matrix containing one or more lagged $\boldsymbol{y}$'s, the estimator will be inconsistent.

One can still carry out OLS in the presence of heteroscedasticity, though for valid inference, the correct form for var$(\hat{\boldsymbol{\beta}})$ would have to be implemented, with $\sigma^2\boldsymbol{\Omega} = \text{ diag }\{\sigma_1^2, \sigma_2^2, ..., \sigma_n^2\}$, requiring n parameters to be estimated with only n observations. White (1980) showed that looking at the problem in this way could be misleading and that the issue that was important was finding a satisfactory estimate of $\boldsymbol{X}'\sigma^2\boldsymbol{\Omega}\boldsymbol{X}$, a $k \times k$ matrix, where k is independent of n. Suppose $\boldsymbol{x}_t' = [1,\ x_{2,t},\ ...,\ x_{k,t}]$ is the t'th row of $\boldsymbol{X}$, then

$$\begin{aligned}
\boldsymbol{X}'\sigma^2\boldsymbol{\Omega}\boldsymbol{X} &= \begin{pmatrix} \vdots & \vdots & & \vdots \\ \boldsymbol{x}_1 & \boldsymbol{x}_2 & ... & \boldsymbol{x}_n \\ \vdots & \vdots & & \vdots \end{pmatrix}\begin{pmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & \sigma_n^2 \end{pmatrix}\begin{pmatrix} ... & \boldsymbol{x}_1' & ... \\ ... & \boldsymbol{x}_2' & ... \\ & \vdots & \\ ... & \boldsymbol{x}_n' & ... \end{pmatrix} \\
&= \sum_{t=1}^{n}\sigma_t^2\boldsymbol{x}_t\boldsymbol{x}_t'.
\end{aligned}$$

The White estimator replaces unknown σ_t^2 by e_t^2, where $e_t = y_t - \boldsymbol{x}_t'\hat{\boldsymbol{\beta}}$, giving a consistent estimator of the variance matrix for the OLS coefficients and is useful as it does not require any specific assumption on the form of heteroscedasticity. An estimate of var$(\hat{\boldsymbol{\beta}})$ is then given by

$$\begin{aligned}
\widehat{\mathrm{var}}(\hat{\boldsymbol{\beta}}) &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\hat{\boldsymbol{\Omega}}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
\sigma^2\hat{\boldsymbol{\Omega}} &= \text{diag }\{e_1^2, e_2^2, ..., e_n^2\}.
\end{aligned}$$

[1]The correct variance matrix for the OLS coefficient vector is

$$\begin{aligned}
\mathrm{var}(\hat{\boldsymbol{\beta}}) &= \mathrm{E}\{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\} \\
&= \mathrm{E}\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{uu}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\} \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}.
\end{aligned}$$

The conventional formula calculates $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, which is only part of the correct expression; so conventional test statistics are invalidated.

The square roots of the elements on the diagonal of var$(\hat{\boldsymbol{\beta}})$ are the estimated s.e's of the OLD coefficients, referred to as *heteroscedasticity consistent* s.e's (HCSE's). The usual t and F tests are valid asymptotically. and general linear hypotheses can be tested using the Wald statistic.

8.7.2 Some tests for heteroscedasticity

Since OLS can be inefficient, it can be important to test for the presence of heteroscedasticity. A brief outline of the White test, Breusch-Pagan/Godfrey test, and the Goldfeld-Quandt test will be given. Other tests which will not be covered in this course include the Bartlett test, Szroeter's class of tests and some nonparametric tests. For further information on these, consult Johnston and Dinardo (1997, Ch. 6), or Judge, Griffiths, Hill, Lütkepohl and Lee (1980, Ch. 11).

8.7.2.1 The White test

The aim is to compute the auxiliary regression of the squared OLS residuals on a constant and a set of variables (the regressors, their squares and their cross products). Suppose

$$\boldsymbol{x}_t' = \begin{pmatrix} 1 & x_{2,t} & x_{3,t} \end{pmatrix}$$

then there are effectively 9 possible variables, except that the square of 1 is 1 and the crossproduct of 1 with the x variables replicates them, so that the set becomes

$$\begin{pmatrix} 1 & x_{2,t} & x_{3,t} & x_{2,t}^2 & x_{3,t}^2 & x_{2,t}x_{3,t} \end{pmatrix}.$$

The regression is thus e_t^2 on this set. On the hypothesis of homoscedasticity, $nR^2 \sim \chi_5^2$ asymptotically, where the d.o.f's are the number of variables in the regression excluding the constant. In general, under the null of homoscedasticity,

$$nR^2 \sim \chi_q^2.$$

One problem with the White test is that the d.o.f. may become rather large, which reduces the power of the test.

8.7.2.2 The Breusch-Pagan/Godfrey test

This is an example of the LM test. Suppose we consider $y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + u_t$ where

$$\boldsymbol{x}_t' = \begin{pmatrix} 1 & x_{2,t} & x_{3,t} & ...x_{k,t} \end{pmatrix}.$$

The heteroscedasticity is assumed to take the form

$$\begin{aligned}
\mathrm{E}u_t &= 0 \quad \text{for all } t \\
\sigma_t^2 &= \mathrm{E}u_t^2 = h(\boldsymbol{z}_t'\boldsymbol{\alpha})
\end{aligned}$$

where $\boldsymbol{z}_t' = [1,\ z_{2,t},\ ...,z_{p,t}]$ is known, $\boldsymbol{\alpha} = [\alpha_1,\ \alpha_2,\ ...\ ,\alpha_p]$ is unknown, and $h(\cdot)$ is some unspecified function that takes positive values. The null of homoscedasticity is thus

$$H_0: \alpha_2 = \alpha_3 = ... = \alpha_p = 0,$$

since then $\sigma_t^2 = h(\alpha_1)$ =constant. The *restricted* model under H_0 is then simply estimated by applying OLS on the assumption of normally distributed disturbances. The test procedure is then carried out as,

(1) Obtain the OLS residuals, $e_t = y_t - \boldsymbol{x}_t'\hat{\beta}$ and an estimate $\tilde{\sigma}^2 = \sum e_t^2/n$.
(2) Regress $e_t^2/\tilde{\sigma}^2$ on $\boldsymbol{z}_t$ by OLS and compute ESS.
(3) Under H_0,

$$\frac{1}{2}\text{ ESS} \xrightarrow{d} \chi_{p-1}^2$$

which rejects homoscedasticity if some pre-specified critical value is exceeded.

(4) An asymptotically equivalent procedure is to regress e_t^2 on $\boldsymbol{z}_t$; then $nR^2 \xrightarrow{d} \chi_{p-1}^2$ under H_0.

The test requires knowledge of the z variables causing the heteroscedasticity.

8.7.2.3 The Goldfeld-Quandt test

This is a finite-sample test, applicable if there is thought to be a single variable (typically one of the x's) thought to be an indicator of the heteroscedasticity. Suppose, σ_t^2 is assumed to be positively related to the i'th regressor, X_i. The test is carried out in the following way,

(1) Re-order the obs. by the value of X_i.
(2) Omit c central obs.
(3) Fit separate regressions by OLS on the first and last $(n-c)/2$ obs, provided $(n-c)/2$ exceeds the number of parameters in the relation.
(4) If RSS_1 denotes the RSS from the smaller X_i values and RSS_2 denotes the RSS from the larger X_i values, then

$$R = \frac{\text{RSS}_2}{\text{RSS}_1} \sim F_{(n-c-2k)/2,(n-c-2k)/2}$$

under homoscedasticity.

The power of the test depends on c; power being low if c is too large or c is too small. An ad-hoc procedure would be to set $c = n/3$.

8.7.3 Autocorrelated disturbances

The pairwise *autocovariances* are defined by

$$\gamma_s = \text{E}(u_t u_{t+s}) \quad \text{for } s = 0, \pm 1, \pm 2, ...$$

When $s = 0$, $\gamma_0 = \text{E}(u_t^2) = \sigma_u^2$. The *autocorrelation* coefficient at lag s is defined by

$$\rho_s = \frac{\text{cov}\,(u_t u_{t+s})}{\sqrt{\text{var}(u_t)\text{var}(u_{t+s})}}$$

which reduces to

$$\rho_s = \frac{\gamma_s}{\gamma_0}$$

given homoscedasticity. We can express var$(\boldsymbol{u})$ as

$$\text{var}(\boldsymbol{u}) = \begin{pmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{n-1} & \gamma_{n-2} & \cdots & \gamma_0 \end{pmatrix} = \sigma_u^2 \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{pmatrix}.$$

Without any further information, the estimation problem is intractable as there are $n + k$ unknowns and only n obs. It is important to test for autocorrelation as presence of autocorrelated disturbances could indicate an inadequate specification.

8.7.3.1 Forms of autocorrelation

What follows will be covered in detail during the time series course in the second term and so is only meant as an introduction. The *first-order autoregressive* model, or AR(1) is defined as

$$u_t = \phi u_{t-1} + \epsilon_t$$

where ϵ_t is $N(0, \sigma_\epsilon^2)$. Suppose, (for stationarity), that $|\phi| < 1$. Then $\text{E}(u_t) = 0$ and $\text{var}(u_t) = \sigma_u^2 = \sigma_\epsilon^2/(1-\phi^2)$ so that the autocorrelation coefficients are $\rho_s = \phi^s$, for $s \geq 0$. The variance-covariance matrix of $\boldsymbol{u}$ can be written as

$$\text{var}(\boldsymbol{u}) = \sigma_u^2 \begin{pmatrix} 1 & \phi & \cdots & \phi^{n-1} \\ \phi & 1 & \cdots & \phi^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \cdots & 1 \end{pmatrix},$$

so that there are only $k + 2$ parameters to be estimated and feasible GLS can be carried out. Another popular model is the *first order moving average* model, or MA(1), defined by

$$u_t = \epsilon_t + \theta\epsilon_{t-1}$$

where ϵ_t is defined as above. Then $\sigma_u^2 = \sigma_\epsilon^2(1+\theta^2)$, $\rho_1 = \theta/(1+\theta^2)$ and $\rho_i = 0$ for $i \geq 2$.

8.7.4 OLS and autocorrelated disturbances

If OLS is applied when $\boldsymbol{X}$ is nonstochastic and the disturbances are autocorrelated, then the consequences are the same as those under heteroscedasticity - unbiased, consistent but inefficient estimation and invalid inference procedures. See Johnston and Dinardo (1997, pp. 177-178) for an example.

8.7.5 Testing for autocorrelated disturbances

Suppose $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$ and that $u_t = \rho u_{t-1} + \epsilon_t$. Under the null hypothesis of zero autocorrelation,

$$H_0 : \rho = 0 \quad \text{vs.} \quad H_1 : \rho \neq 0$$

Testing the hypothesis involves the residuals (since the u's are unobservable), $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Since $\boldsymbol{e} = \boldsymbol{Mu}$, where $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ (symmetric, idempotent and of rank $n - k$),

$$\text{var}(\boldsymbol{e}) = \text{E}(\boldsymbol{ee}') = \sigma_u^2\boldsymbol{M}.$$

As $\boldsymbol{M}$ is a function of the sample values of the explanatory variables, exact finite-sample tests on the e's are impossible to derive that will be valid for any $\boldsymbol{X}$ matrix. A brief outline of certain tests is outlined below, for further reference, see Johnston and Dinardo (1997, pp. 179-187), or Judge *et al.* (1980, Ch. 8.4).

8.7.5.1 Durbin-Watson test

The Durbin Watson, or (DW) test statistic is computed from the OLS residuals $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$, and is defined as

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$$

and ranges between 0 and 4 with

- $d < 2$ for positive autocorrelation of the e's
- $d > 2$ for negative correlation of the e's
- $d \approx 2$ for zero correlation of the e's.

Since any computes d value depends on the associated $\boldsymbol{X}$ matrix, Durbin and Watson established upper (d_U) and lower (d_L) bounds for the critical values. To test the hypothesis of zero autocorrelation, the procedure is

(1) If $d < d_L$, reject H_0.
(2) If $d > d_U$, do not reject H_0.
(3) If $d_L < d < d_U$, the test is inconclusive.

Note: to apply DW's test, a constant must be included in the regression and the test is only valid for nonstochastic $\boldsymbol{X}$.

8.7.5.2 Breusch-Godfrey test

The procedure builds on the DW test and proceeds as follows:

(1) Apply OLS to the specified model to obtain the residuals e_t.
(2) Regress e_t on $[1\ x_t\ e_{t-1}]$ to find R^2.
(3) Under H_0, nR^2 is asymptotically χ_1^2.

Whereas the DW test suggests looking at the significance of the coefficient on e_{t-1}, the Breusch-Godfrey (LM) test gives nR^2 as a test statistic with an asymptotic χ^2 distribution. Both tests are asymptotically equivalent.

8.7.5.3 Box-Pierce-Ljung statistic

The Box-Pierce Q statistic is based on the squares on the first p autocorrelation coefficients of the OLS residuals and is defined as

$$Q = n \sum_{j=1}^{p} r_j^2$$

where

$$r_j = \frac{\sum_{t=j+1}^{n} e_t e_{t-j}}{\sum_{t=1}^{n} e_t^2}.$$

Under the null of zero autocorrelation, Q has a χ_l^2 distribution. The revised Ljung-Box statistic (which has better small sample performance) is defined by

$$Q' = n(n+2) \sum_{j=1}^{p} \frac{r_j^2}{n-j}.$$