Lecture 7

Specification and Model Selection

7.1 Goodness-Of-Fit and Selection of Regressors

Choosing a set of explanatory variables based on the size of R^2 can lead to nonsensical models. For the regression model

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + \varepsilon_i$$

R^2 is simply **an estimate of how much variation in y is explained by $x_2, \cdots, x_k$ in the population**, where $x_{i1} = 1$. However, R^2 is useful for choosing among nested models. We can use the F statistic for testing the joint significance, which depends on the difference in R^2 between the unrestricted and restricted models.

7.1.1 Adjusted R-Squared

To see how the usual R-squared might be adjusted, let's see how R^2 works:

$$R^2 = 1 - (\text{SSE}/n)/(\text{SST}/n) \Rightarrow \text{an estimate of } 1 - \sigma_u^2/\sigma_y^2$$

Define σ_y^2 as the population variance of y_i and let σ_ε^2 as the population variance of ε_i, the error term. The **population R-squared** is defined as

$1 - \sigma_\varepsilon^2/\sigma_y^2$; it is the proportion of the variation in y in the population explained by the independent variables. This is what R^2 is supposed to be estimating.

R^2 estimates σ_ε^2 by SSR/n, which is biased. We could replace SSR/n by SSR/$(n-K)$. Also, we can use SST/$(n-1)$ in place of SST/n, as SST/$(n-1)$ is the unbiased estimator of σ_y^2. The adjusted R^2 is specified as

$$\begin{aligned}\overline{R}^2 &\equiv 1 - [\text{SSR}/(n-K)]/[\text{SST}/(n-1)] \\ &= 1 - \hat{\sigma}^2/[\text{SST}/(n-1)]\end{aligned}$$

since $\hat{\sigma}^2 = \text{SSR}/(n-K)$. The adjusted R^2 imposes a penalty for adding additional independent variables to a model. Because SSR never goes up (and usually falls) as more independent variables are added to a regression model, R^2 can never fall when a new independent variable added.

As k increases, SSR decreases, so does the degree of freedom in regression, $(n-K)$. $\Rightarrow \overline{R}^2$ can go up or down when a new independent variable is added to the regression.

$$\overline{R}^2 = 1 - (1 - R^2)(n-1)/(n-K)$$

$\overline{R}^2$ could be negative now.

7.1.2 Using Adjusted R-squared to Choose Between Nonnested Models

A Case of **nonnested models**: when neither equation is a special case of the other. For example,

$$\begin{aligned}\text{Model (1)}: &\quad y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon \\ \text{Model (2)}: &\quad y = \beta_1 + \beta_2 x_3 + \beta_3 x_3 + \beta_5 x_5 + \varepsilon\end{aligned}$$

The Wald or F test (as well as LM and LR statistics) only allows one to test *nested* models: one model (the restricted model) is a special case of the other model (the unrestricted model). On the other hand, adjusted R-squared can be used to choose among different nonnested models. [Using the information criterion to judge the optimal model also works.]

Disadvantage of Comparing $\overline{R}^2$ to Choose among Different Nonnested Models:

One can not use $\overline{R}^2$ to choose **different functional forms** for the dependent variable. For example, y or $\log(y)$.

7.2 Specification Errors: Selection of Variables

A specification error happens if one misspecifies a model either in terms of the choice of variables, functional forms, or the error structure (that is, the stochastic disturbance term ε_i and its properties). In this section, the first type of specification errors is discussed. Lecture 6 discusses the choice of function form, and specification errors in ε are discussed in lectures about heteroskedasticity and serial correlation.

Two types of errors regarding choosing the independent variables that belongs in a model:

1. omitting important relevant variables
2. including irrelevant variables (redundant variables)

7.2.1 Case 1: Omission of an important variable

the true model:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

but the **estimated model** is

$$y_i = \beta_1 + \beta_2 x_{i2} + v_i$$

ε_i is assumed to satisfy $\varepsilon_i \sim i.i.d.N(0, \sigma^2)$, $0 < \sigma^2 < \infty$, $\text{Cov}(x_{ik}, \varepsilon_i) = 0$.

Consequences of omitting an important variable:

a. The estimated value of all the other regression coefficients will be biased unless the excluded variable is uncorrelated with every included variable.

b. The estimated constant term is generally biased and hence forecasts will also be biased.

c. The estimated variance of the regression coefficient of an included variable will generally be biased, and tests of hypothesis are invalid.

The cause of the bias (known as **omitted variable bias**):

Because $v_i = \beta_3 x_{i3} + \varepsilon_i$, we obtain that

$$\mathrm{E}(v_i) = \beta_3 x_{i3} \neq 0$$

Therefore, v_i violates Assumption A2.

Furthermore, the covariance between x_{i2} and v_i is:

$$\begin{aligned}\mathrm{Cov}(x_{i2}, v_i) &= \mathrm{Cov}(x_{i3}, \beta_3 x_{i3} + \varepsilon_i) \\ &= \beta_3 \mathrm{Cov}(x_{i2}, x_{i3}) + \mathrm{Cov}(x_{i2}, \varepsilon_i) \\ &= \beta_2 \mathrm{Cov}(x_{i2}, x_{i3})\end{aligned}$$

as x_2 and ε are uncorrelated. Hence, unless $\mathrm{Cov}(x_{i2}, x_{i3})$ is zero, $\mathrm{Cov}(x_{i2}, v_i)$ will not be zero. It violates Assumption A2 also. The properties of unbiasedness and consistency depend on Assumption A2. It follows, therefore, that b_2 will not be unbiased or consistent.

Let b_1 and b_2 be the estimate for the model $y_i = \beta_1 + \beta_2 x_{i2} + v_i$.

$$b_2 = S_{y2}/S_{22} \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_2 \bar{x}_2$$

where $S_{22} = \sum(x_{i2} - \bar{x}_2)^2$, $S_{y2} = \sum(y_i - \bar{y})(x_{i2} - \bar{x}_2)$. It follows that

$$\mathrm{E}(b_2) = \beta_2 + \beta_3 \left[\frac{S_{23}}{S_{22}}\right]$$

and

$$\mathrm{E}(b_1) = \beta_1 + \beta_3 \left[\bar{x}_3 - \bar{x}_2 \frac{S_{23}}{S_{22}}\right]$$

where $S_{23} = \sum(x_{i2} - \bar{x}_2)(x_{i3} - \bar{x}_3)$

Unless $S_{23} = 0$, $E(b_2) \neq \beta_2$, and therefore b_2 will be biased in general.

Notes:

1. b_2 includes a term involving β_3. Hence b_2 is not only the marginal effect of x_2, but also captures part of the effect of the excluded variables. As a result, the coefficient measures the direct effect of the included variable as well as the indirect effect of the excluded variable.

2. Even if $S_{23} = 0$, b_1 will be biased unless the mean of x_3 is zero. b_1 also captures part of the effect of the omitted variable x_3.

The Danger of Omitting a Constant Term:

If the constant term has been omitted, then the regression line would have been forced through the origin, which might be a serious misspecification.

To constrain the regression line to go through the origin would mean a biased estimate for the slope and larger errors. Therefore, the constant term should always be retained unless there is a strong theoretical reason not to do so.

7.2.2 Case 2: Inclusion of an irrelevant variable

true model:

$$y_i = \beta_1 + \beta_2 x_{i2} + \varepsilon_i$$

estimated model:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + v_i$$

Consequences of including a redundant variable:

a. The estimated value of all the other regression coefficients will still be unbiased and consistent.

b. However, the estimated variance of the regression coefficient of a included variable will be higher than that without the irrelevant variable, and hence the coefficients will be inefficient.

c. Because the estimated variances of the regression coefficients are unbiased, tests of hypothesis are still valid.

The consequences of including an irrelevant variable are thus less serious as compared to omitting an important variable.

7.3 Approaches to Modelling

Trade-off between the reduction in SSR and the loss of degree of freedom:

- number of regressors increases $\Rightarrow \sum e_i^2$ decreases, R^2 increases, but a loss in d.f.

- $\bar{R}^2$ and the standard error of the residuals, $[\text{SSR}/(n-K)]^{1/2}$, take account of the above trade off. They are the most commonly used criteria for model comparison.

Three Main Approaches:

1. General to Simple
2. Simple to General
3. Information criteria

7.3.1 General to Simple Approach (The Hendry /LSE Approach)

Start with a general unrestricted model and then reduced it by eliminating one at a time the variable with the least significant coefficient

Methodology: first of all, one uses economic theory, intuition, and experience to approximate the data generating process (DGP); then, put the model through a number of diagnostic tests (Wald and t-tests) to see if either model or methodology can be improved

1. start with a general model which is overparameterized (i.e., has more lags and variables than one would normally start with
2. carry out a data-based simplification through Wald and t-tests

7.3.2 Simple to General Modelling

Using the Lagrange Multiplier (LM) Test:

- Use the LM test for adding variables
- Restricted model: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + v$
- Unrestricted model: $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m + \beta_{m+1} x_{m+1} + \cdots + \beta_k x_k + u$
- $H_0 : \beta_{m+1} = \cdots = \beta_k = 0$

step 1: estimate the restricted model $\Rightarrow$ yield estimates $\hat{\beta}_1, \cdots, \hat{\beta}_m$

step 2: obtain estimated residual

$$\hat{u}_r = y - \hat{\beta}_0 - \hat{\beta}_1 x_1 - \cdots - \hat{\beta}_m x_m$$

step 3: regress $\hat{u}_r$ against a constant term, $x_1, \cdots, x_m, x_{m+1}, \cdots, x_k$ (referred to be **auxiliary regression**) $\Rightarrow$ calculate $nR^2_{\hat{u}}$. It has been shown that $nR^2_{\hat{u}} \sim \chi^2$ with d.f. equal to the number of restrictions in the null hypothesis (Engle (1982)). Thus, in our case, $nR^2_{\hat{u}} \sim \chi^2_{k-m}$. If $nR^2_{\hat{u}} > \chi^{*2}_{k-m}(a)$, the point on χ^2_{k-m} such that the area to its right is a, the level of significance, we would reject the null hypothesis that the new regression coefficients are all zero.

$*$ The t-values of individual coefficient can give a clue as to which of these omitted variables might be included.

Reasoning:

(1) Suppose the true specification had been the unrestricted model, i.e., x_{m+1}, $\cdots$, x_k should be included. Their effect would be captured by $\hat{u}_r \Rightarrow \hat{u}_r$ should be related to omitted variables $(x_{m+1}, \cdots, x_k)$.

(2) If we regress $\hat{u}_r$ against these variables, we should get a good fit, which indicates that at least some of the variables $x_{m+1}, \cdots, x_k$ should have been included the model. So then we regress $\hat{u}_r$ against a constant and all the Xs, i.e., all the variables including in the unrestricted model.

REMARK: In the approach, we USE THE AUXILIARY REGRESSION TO SELECT VARIABLES TO BE ADDED TO THE BASIC MODEL

- About Lagrange multiplier (LM) test, likelihood ratio (LR) test, and the Wald test:

  1. two models are formulated: a restricted model and an unrestricted model
     - the restricted model is obtained by imposing linear or nonlinear constraints on its parameters, and corresponds to the null hypothesis

  - the unrestricted model is the alternative hypothesis

2. The Wald test starts with the alternative (the unrestricted model) and ask whether the null (restricted model) should be preferred
3. The LR test is a direct comparison of two hypotheses.

- Assume $y_i = \delta x_i + \varepsilon_i$ and u's are normally distributed with mean zero and variance σ^2, the log-likelihood can be written as follows:

$$\log L = -n \log \sigma - n \log(\sqrt{2\pi}) - \frac{\sum(y_i - \delta x_i)}{2\sigma^2}$$

  $H_0 : \delta = \delta_0$, $H_1 : \delta \neq \delta_0$

- Let the likelihood function evaluated at $\hat{\delta}$ (the maximum likelihood estimator of δ) be $L(\hat{\delta})$, and the likelihood function under the null $\delta = \delta_0$ be $L(\delta_0)$. The likelihood ratio is defined as $\lambda = L(\delta_0)/L(\hat{\delta})$. Because the denominator is based on the unrestricted model, its value can not be smaller than that of the numerator. Therefore $0 \leq \lambda \leq 1$. If H_0 were true, we would expect λ to be close to 1. If λ is far from 1, the likelihood ratio under the null is very different from that under the restricted model (i.e., H_0). This suggests that we should reject H_0 is λ is too small.

- The LR test is formulated as one of rejecting the null hypothesis if $\lambda \leq K$, where K is determined by the condition that, under the null hypothesis, the probability that $0 \leq \lambda \leq K$ is equal to the level of significance (a), that is, $P(0 \leq \lambda \leq K | \delta = \delta_0) = a$.

- It has been shown that, for large sample sizes, the statistic

$$\text{LR} = -2 \log \lambda = 2 \log L(\hat{\delta}) - 2 \log L(\delta_0)$$

  has a chi-square distribution with d.f. equal to the number of restriction in the null hypothesis, i.e., χ^2_1 in this case.

- Nested hypotheses testing:
  - restricted model $\subset$ unrestricted model
  - we could apply the Wald test, LM test, and LR test
  - Remarks:

- The Hendry/LSE "General to Simple" approach employs the Wald test
- LM test: $H_0 : k - m$ added variables insignificant

(1) regress the dependent variable against a list of basic independent variables and constant.

(2) obtain residual from above OLS regression

(3) regress residual from (2) against all x's in (1) plus new variables

- asymptotically LM = LR = Wald
- generally LM $\leq$ LR $\leq$ Wald

$\Rightarrow$ whenever the LM test rejects the null of zero coefficients, so will the others

$\Rightarrow$ whenever the Wald test fails to reject the null, other tests will too

Computationally, the LR test is the most cumbersome, unless it can be converted to a t-, F-, or χ^2 test. The other two tests are straightforward.

7.3.3 Information Criteria for Choosing Among Models

Let K be the number of parameters, N be the number of observations, SSR be the sum of squared residuals, SSR $= \sum_{i=1}^{n} e_i^2$.

1. finite prediction error FPE $= \left(\frac{\text{SSR}}{N}\right)\frac{N+K}{N-K}$
2. Akaike information criterion AIC $= \left(\frac{\text{SSR}}{N}\right)\exp\left(2K/N\right)$
3. Schwarz information criterion SCHWARZ $= \left(\frac{\text{SSR}}{N}\right)N^{K/N}$
4. HQ $= (\log N)^{2K/N}\left(\frac{\text{SSR}}{N}\right)$
5. SHIBATA $= \left(1+\frac{2K}{N}\right)\left(\frac{\text{SSR}}{N}\right)$

6. $\text{GCV} = \left(1 - \frac{K}{N}\right)^{-2} \left(\frac{\text{SSR}}{N}\right)$
7. $\text{RICE} = \frac{1}{1 - \frac{2K}{N}} \left(\frac{\text{SSR}}{N}\right)$

All of the above take the form of $\left(\frac{\text{SSR}}{N}\right)$ multiplied by a penalty factor that depends on model complexity.

$\Rightarrow$ A more complex model will reduce SSR but raise the penalty. These criteria provide tradeoffs between goodness of fit and model complexity.

RULE: A model with a **lower** value of a criterion statistic is judged to be **preferable**.

REMARKS:

- Take the form of SSR/N multiplied by a penalty factor that depends on model complexity. A more complex model will reduce SSR but raise the penalty. These criteria provide tradeoffs between goodness of fit and model complexity. A model with a **lower** value of a criterion statistic is judged to be **preferable**.
- Ideally we would like a model to have lower values for all these statistics, as compared to an alternative model. However, it is possible to find a model superior under one criterion and inferior under another. A model that outperforms another in several of these criteria might be preferred. The **AIC** criterion, however, is commonly used in time series analysis.

Appendix: Three Asymptotically Equivalent Test Statistics

Likelihood ratio statistic
Wald statistic
Lagrange multiplier statistic

- Lagrange multiplier (LM) statistic: a large sample test.
- LM statistic is also called **score statistic**.
- In large sample, LM $\overset{a}{\sim} F$, provided the Gauss-Markov assumptions (A1-A4), we do not need the normality assumption)

LR, LM, and Wald Tests
Consider a parameter θ:

$$H_0 : \ C(\theta) = 0$$

- LR test: test based on $\log L - \log L_R$
  If the restriction $C(\theta) = 0$ is valid, then imposing it should not lead to a large reduction in $\log L(\theta)$.
  Thus $\log L - \log L_R$ must be close to zero.
  L is the value of likelihood function at the unconstrained value θ.
  L_R is the value of likelihood function at the restricted estimate.
- Wald test: test based on $C(\hat{\theta})$
  If the restriction is valid, then $C(\hat{\theta})$ should be close to zero
  $\Rightarrow$ reject the null hypothesis $H_0 : \ C(\theta) = 0$ if $C(\hat{\theta})$ is significantly different from 0.
- Lagrange multiplier (LM) test: test based on slope of $\log L$ at the point where the function is maximized subject to the restriction

  Thus the slope of $\log L$ should be near zero at the restricted estimator:
  $\frac{d \log L(\theta)}{d\theta}|_{\hat{\theta}} \approx 0$

LR, Wald and LM tests are asymptotically equivalent under the null hypothesis but they can behave rather differently in a small sample
Choices among the depends on the case of computation:

LR: requires calculation of both restricted and unrestricted estimators

Wald: requires only the unrestricted estimator

LM: requires only the restricted estimator

Case 1: linear model + nonlinear constraint
$\Rightarrow$ Wald test is preferable

Case 2: restrictions amount to the removal of nonlinearity
$\Rightarrow$ LM test is simpler

Likelihood Ratio Test (LR Test) Let θ be a vector of parameters to be estimated, and H_0 specify some sort of restriction on these parameters.

$\hat{\theta}_U$: MLE of θ under the unrestricted model (that is, w/o regard to H_0)
$\hat{\theta}_R$: constrained MLE (obtained under the restricted model)
$\hat{L}_U$: likelihood function evaluated at $\hat{\theta}_U$
L_R : likelihood function evaluated at $\hat{\theta}_R$

$$\text{likelihood ratio:} \qquad \lambda = \frac{\hat{L}_R}{\hat{L}_U}; \quad 0 < \lambda < 1$$

Theorem (Distribution of LR test statistic)
Under regularity, $-2\log\lambda \xrightarrow{a} \chi^2_r$ where $r = \sharp$ of restrictions

$$-2\log\lambda = -2(\log\hat{L}_R - \log\hat{L}_U)$$

Wald Test (F-Test) Let $\hat{\theta}$ be the estimates obtained w/o restrictions.
$H_0 : \ C(\theta) = q$

Theorem (Distribution of Wald test statistic)

Wald statistic

$$w = [C(\hat{\theta}) - q]'(\text{Var}[C(\hat{\theta}) - q])^{-1}[C(\hat{\theta}) - q]$$

Under H_0, $w \overset{a}{\sim} \chi^2_r$ where $r = \sharp$ of restrictions

Proof:
Recall: If $x \sim N_J(\mu, \Sigma)$, then $(x - \mu)'\Sigma^{-1}(x - \mu) \sim \chi^2_J$
However, Σ is unknown. One must still estimate the covariance matrix.
For example: $C(\theta) = R\theta$
$H_0 : R\theta - q = 0$ (linear restrictions)
$\text{Est.Var}[C(\hat{\theta}) - q] = \text{Est.Var}[R\hat{\theta} - q] = R\ \text{Est.Var}[\hat{\theta}]R'$

$$w = [R\hat{\theta} - q]'[R\ \text{Est.Var}[\hat{\theta}]R']^{-1}[R\hat{\theta} - q] \sim \chi^2_r$$

where $r = \sharp$ of rows in R
eg. $H_0 :\ \theta = \theta_0$
$H_1 :\ \theta \neq \theta_0$
The earlier t test is based on $\dfrac{\hat{\theta} - \theta_0}{s(\hat{\theta})}$ where s is estimated asymptotic standard error.
On the other hand, the Wald test is based on

$$\begin{aligned} w &= [(\hat{\theta} - \theta_0)][\text{Est.Var}(\hat{\theta} - \theta_0)]^{-1}[(\hat{\theta} - \theta_0) - 0] \\ &= \frac{(\hat{\theta} - \theta_0)^2}{\text{Est.Var}(\hat{\theta} - \theta_0)} = t^2 \end{aligned}$$

Implication: (t statistic)^2 = Wald statistic

Lagrange Multiplier Test (LM Test) or Score Test
LM=(efficient) score
$\star$ based on restricted model
Suppose that we maximize log likelihood s.t. $C(\theta) - q = 0$
Let λ be a vector of Lagrange multipliers and define the Lagrange function

$$\log L^{\star}(\theta) = \log L(\theta) + \lambda'[C(\theta) - q] \tag{1}$$

Solutions:

$$\begin{aligned} \frac{\partial \log L^{\star}(\theta)}{\partial \theta} &= \frac{\partial \log L(\theta)}{\partial \theta} + [\frac{\partial C(\theta)}{\partial \theta}]'\lambda = 0 \\ \frac{\partial \log L^{\star}(\theta)}{\partial \lambda} &= C(\theta) - q = 0 \end{aligned} \tag{2}$$

$\star$ If the restrictions are valid, then imposing them will not lead to a significant different in the maximized value of likelihood function
$\Rightarrow$ the second term at the right hand side of equation (1) will be very small
$\Rightarrow$ Thus we could test $H_0 : \ \lambda = 0$

$\star$Using the auxiliary regression to derive LM statistic:
There is an equivalent simpler formulation, however.
At the restricted maximum, the derivatives are

$$\frac{\partial \log L(\hat{\theta}_R)}{\partial \hat{\theta}_R} = -\left[\frac{\partial C(\hat{\theta}_R)}{\partial \hat{\theta}_R}\right]' = \hat{g}_R$$

If the restrictions are valid, then $\hat{g}_R \approx 0$
Theorem (Distribution of LM statistic)

$$\text{LM } = \left(\frac{\partial \log L(\hat{\theta}_R)}{\partial \hat{\theta}_R}\right)' [I(\hat{\theta}_R)]^{-1} \left(\frac{\partial \log L(\hat{\theta}_R)}{\partial \hat{\theta}_R}\right)'$$

Under H_0, LM $\overset{a}{\sim} \chi^2_r$