Lecture 6
Specification and Model Selection Strategies

1

Model Selection Strategies

• So far, we have implicitly used a simple strategy:

(1) We started with a DGP, which we assumed to be true.

(2) We tested some H_0 (from economic theory).

(3) We used the model (restricted, if needed) for prediction & forecasting.

• Under CLM assumptions (**A1**) to (**A5**), *t-tests, F-tests* and predictions have desirable properties. But if assumptions do not hold, then,

- Tests can be weak with unknown distributions.
- Tests may be biased –i.e., more likely to reject H_0 when it is true than when it is false. Same for predictions.
- Tests may be inconsistent –i.e., power does not approach 1 for every alternative for a given significance level

Model Selection Strategies

• In this lecture we will address assumptions (**A1**)-(**A5**). In particular, how do we propose and select a model (a DGP)?

• Potentially, we have a huge number of possible models (different functional form, *f(.)*, and explanatory variables, **X**). Say, we have

Model 1 $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$

Model 2 $\mathbf{Y} = \mathbf{Z}\gamma + \xi$

Model 3 $\mathbf{Y} = (\mathbf{W}\gamma)^{\lambda} + \eta$

Model 4 $\mathbf{Y} = \exp(\mathbf{Z}\ \mathbf{D}\ \delta) + \epsilon$

• We want to select the best model, the one that is closest to the DGP. In practice, we aim for a good model.

Model Selection Strategies

• A model is a simplification. Many approaches:

• "Pre-eminence of theory." Economic theory should drive a model. Data is only used to quantify theory. Econometric methods offer sophisticated ways 'to bring data into line' with a particular theory.

• Purely data driven models. Success of ARIMA models (late 60s – early 70s). No theory, only exploiting the time-series characteristics of the data to build models.

• Modern (LSE) view. A compromise: theory and the characteristics of the data are used to build a model.

Model Selection Strategies

• Modern view: Theory and practice play a role in deriving a good model. David Hendry (2009) emphasizes:

"This implication is not a tract for mindless modeling of data in the absence of economic analysis, but instead suggests formulating more general initial models that embed the available economic theory as a special case, consistent with our knowledge of the institutional framework, historical record, and the data properties. ... Applied econometrics cannot be conducted without an economic theoretical framework to guide its endeavours and help interpret its findings. Nevertheless, since economic theory is not complete, correct, and immutable, and never will be, one also cannot justify an insistence on deriving empirical models from theory alone."

Model Selection Strategies

• According to David Hendry, a good model should be:

- Data admissible	-i.e., modeled and observed **y** should have the same properties.
- Theory consistent	-our model should "make sense"
- Predictive valid	-we should expect out-of-sample validation
- Data coherent	-all information should be in the model. Nothing left in the errors (*white noise errors*).
- Encompassing	-our model should explain earlier models.

• That is, we are searching for a statistical model that can generate the observed data (**y**,**X**), this is usually referred as *statistical adequacy*, makes theoretical sense and can explain other findings.

Model Selection Strategies

• FAQ in practice:
- Should I include all the variables in the database in my model?
- How many explanatory variables do I need in my model?
- How many models do I need to estimate?
- What functional form should I be using?
- Should the model allow for structural breaks?
- Should I include dummies & interactive dummies ?
- Which regression model will work best and how do I arrive at it?

Model Selection Strategies: Some Concepts

• *Diagnostic testing:* We test assumptions behind the model. In our case, assumptions (**A1**)-(**A5**) in the CLM.

Example: Test $E[\boldsymbol{\varepsilon}|\mathbf{X}] = 0$ -i.e., the residuals are zero-mean, white noise distributed errors.

• *Parameter testing:* We test economic H_0's.

Example: Test $\beta_k = 0$ -say, there is no size effect on the expected return equation.

• There are several *model-selection methods*. We will consider two:
- *Specific to General*
- *General to Specific*

Model Selection Strategies : Specific to General

- Begin with a small theoretical model – for example, the CAPM

  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.
- Estimate the model – say, using OLS
- Do some diagnostic testing – are residuals white noise?
- If the assumptions do not hold, then use:
  - More advanced econometrics – GLS instead of OLS?
  - A more general model – APT? Lags?
- Test economic H_0 on the parameters – Is size significant?

- This strategy is known as *specific to general*, *Average Economic Regression* (AER).

Model Selection Strategies: General to Specific

- Begin with a *general unrestricted model* (GUM), which nests restricted models and, thus, allows any restrictions to be tested. Say:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{W}^{\lambda}\boldsymbol{\delta} + \boldsymbol{\varepsilon}.$$

- Then, reduction of the GUM starts. Mainly using *t-tests*, and *F-tests*, we move from the GUM to a smaller, more parsimonious, specific model. If competing models are selected, encompassing tests or information criteria (AIC, BIC) can be used to select a final model. This is the *discovery stage*. After this reduction, we move to:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

- Creativity is needed for the specification of a GUM. Theory and empirical evidence play a role in designing a GUM. Estimation of the GUM should be feasible from the available data.

Model Selection Strategies: General to Specific

• General-to-Specific Method:

Step 1 - First ensure that the GUM does not suffer from any diagnostic problems. Check residuals in the GUM to ensure that they possess acceptable properties. (For example, test for heteroskedasticity, white noise, incorrect functional form, etc.).

• **Step 2** - Test the restrictions implied by the specific model against the general model – either by exclusion tests or other tests of linear restrictions.

• **Step 3** - If the restricted model is accepted, test its residuals to ensure that this more specific model is still acceptable on diagnostic grounds.

• This strategy is called *general to specifics* ("*gets*"), *LSE*, *TTT* (Test, test, test). It was pioneered by Sargan (1964). The properties of gets are discussed in Hendy and Krolzig (2005, Economic Journal).

Model Selection Strategies: General to Specific

• The role of diagnostic testing is two-fold.

- In the discovery steps, the tests are being used as design criteria. Testing plays the role of checking that the original GUM was a good starting point after the GUM has been simplified.

- In the context of model evaluation, the role of testing is clear cut. Suppose you use the model to produce forecasts. These forecasts can be evaluated with a test. This is the critical evaluation of the model.

Reference: Campos, Ericson, and Hendry (2005), **General-to-Specific Modelling**. Edward Elgar, London.

John Dennis Sargan (1924 – 1996, England)

Issues: Pre-testing

• A special case of omitted variables.

- First, a researcher starts with an unrestricted model (U):

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}. \qquad \text{(U)}$$

- Then, based on ("preliminary") tests –say, an *F-test*- a researcher decides to use restricted estimator. That is,

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}. \quad \text{s.t. } \mathbf{R\beta} = \mathbf{q} \qquad \text{(R)}$$

- We can think of the estimator we get from estimating R as:

$$\mathbf{b}_{PT} = I_{\{0,c\}}(F)\,\mathbf{b}^* + I_{\{c,\infty\}}(F)\,\mathbf{b},$$

where $I_{\{0,c\}}$ is an indicator function:

$I_{\{0,c\}}(F) = 1$ if *F-stat* not in the rejection region –say, $F<c$-,

$I_{\{c,\infty\}}(F) = 0$, otherwise.

c : critical value chosen for testing H_0:$\mathbf{R\beta} = \mathbf{q}$ **,** using the *F-stat.*

Issues: Pre-testing

• The *pre-test estimator* is a rule which chooses between the restricted estimator, **b***, or the OLS estimator, **b**:

$$\mathbf{b}_{PT} = I_{\{0,c\}}(F)\,\mathbf{b}^* + I_{\{c,\infty\}}(F)\,\mathbf{b}.$$

where $\mathbf{b}^* = \mathbf{b} - (\mathbf{X'X})^{-1}\mathbf{R'}[\mathbf{R}(\mathbf{X'X})^{-1}\mathbf{R'}]^{-1}(\mathbf{Rb} - \mathbf{q})$

• Two "negative" situations:

(1) H_0:$\mathbf{R\beta} = \mathbf{q}$ is true. The *F-test* will incorrectly reject H_0 α% of the time. That is, in α% of the repeated samples, OLS **b** => No bias, inefficient estimator.

(2) H_0:$\mathbf{R\beta} = \mathbf{q}$ is false. The *F-test* will correctly reject H_0 a % of times equal to the power π of the test. That is, (100- π)% of the time, $\mathbf{R\beta}=\mathbf{q}$ will be incorrectly imposed, **b*** will be used: => bias!

Issues: Pre-testing

• The failure of the OLS estimator to have the properties under correct specification is called *pre-test bias*.

• Pre-testing (also called *sequential estimation*, *data mining*) is common in practice. In general, it is ignored –and not even acknowledged.

• Main argument to ignore pre-testing: We need some assumptions to decide which variables are included in a model. Is the probability that pre-testing yields an incorrect set of $\boldsymbol{X}$ greater than the probability of selecting the "correct" assumption?

• The LSE methodology does not see pre-testing in the discovery stage as a problem. For the LSE method, pre-testing at that stage is part of the process of discovery.

Issues: Pre-testing

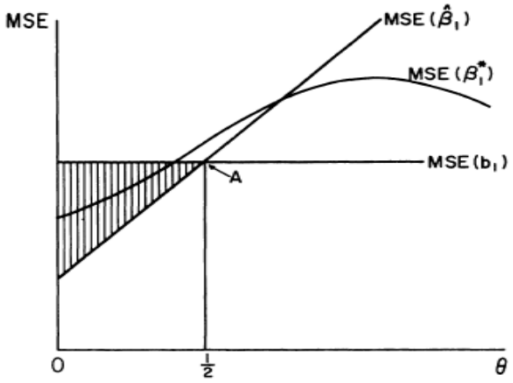
• Checking the MSE of $\mathbf{b}_{PT}$, $\mathbf{b}^*$ and $\mathbf{b}$ helps to evaluate the problem

• Practical advise: Be aware of the problem. Do not rely solely on stats to select a model –use economic theory as well.

• Do not use same sample evidence to generate an H_0 and to test it!

Example: The Fama-French factors have been "discovered" using the CRSP/Compustast database for a long, long time. Thus, testing the Fama-French factors using the CRSP/Compustat is not advisable! (You can test them with another dataset, for example, get international data.)

Issues: Pre-testing



• Taken from Wallace (1977, AJAE) --non-centrality parameter θ.

MSE ($\mathbf{b}_1$) = MSE ($\mathbf{b}$) = OLS MSE

MSE ($\hat{\boldsymbol{\beta}}_1$) = MSE ($\mathbf{b}^*$) = Restricted OLS MSE

MSE ($\boldsymbol{\beta}^*_1$) = MSE ($\mathbf{b}_{PT}$) = Pretest estimator MSE

Issues: Mass significance

• We perform k different tests each with a *nominal significance level* of α:

α = Prob (Rejecting for a given test $|H_0$ for this test is true)

• The *overall significance* of the test procedure is, however, given by

α* = Prob (Rejecting at least one test | all H_0 are true).

• The probability of rejecting at least one H_0 is obviously greater than of rejecting a specific test. This is the problem of *mass significance*.

• Two cases

(1) Independent tests: $(1 - \alpha^*) = (1 - \alpha)^k$

$\Rightarrow \alpha^* = 1 - (1 - \alpha)^k$ & $\alpha = 1 - (1 - \alpha^*)^{1/k}$

(2) Dependent tests (*Bonferroni bounds*): $\alpha^* \leq k\alpha$

$\Rightarrow \alpha \geq \alpha^*/k$

Issues: Mass significance

• Two cases

(1) Independent tests $\alpha^* = 1 - (1 - \alpha)^k$ & $\alpha = 1 - (1 - \alpha^*)^{1/k}$

(2) Dependent tests: $\alpha^* \leq k\alpha$ & $\alpha \geq \alpha^*/k$

⇒ close to the "independent" values for small α, but can differ for large α.

Example: α =0.05 and $k=5$ ⇒ α*(Indep) =.23 & α*(Dep)=.25

α =0.05 and $k=20$ ⇒ α*(Indep) =.64 & α*(Dep) =1

α* =0.05 and $k=5$ ⇒ α(Indep) =.0102 & α(Dep) =.01

α* =0.20 and $k=5$ ⇒ α(Indep) =.044 & α(Dep) =.04

α* =0.20 and $k=20$ ⇒ α(Indep) =.011 & α(Dep) =.01

Issues: Mass significance

• In repeated *parametric testing* (overall level 5%):

- Only accept variables as important when their *p-values* are less than 0.001, preferably smaller
- Maybe look for other ways of choosing variables, say IC.

• In repeated *diagnostic testing* (overall level 20%), we should only accept there is no misspecification if

- All *p-values* are greater than 0.05, *or*
- Most *p-values* are greater than 0.10 with a few in the range 0.02 to 0.10

Modeling Strategies: Properties

• A modeling strategy is *consistent* if its probability of finding the true model tends to 1 as T -the sample size- increases.

• Properties for strategies

(1) Specific to General

- It is not consistent if the original model is incorrect.
- It need not be predictive valid, data coherent, & encompassing.
- No clear stopping point for an unordered search.

(2) General to Specific

- It is consistent under some circumstances. But, it needs a large T.
- It uses data mining, which can lead to incorrect models for small T.
- The significance levels are incorrect. This is the problem of *mass significance*.

Modeling Strategies: Information Criteria

• IC's are equal to the estimated variance or the log-likelihood function plus a penalty factor, that depends on k. Many IC's:

- Theil Information Criterion (Adjusted R^2)

$$\text{Adj. } R^2 = 1 - [(T-1)/(T-k)](1 - R^2) = 1 - [(T-1)/(T-k)]\ RSS/TSS$$

$$\Rightarrow \text{maximizing Adjusted } R^2 \Leftrightarrow \text{minimizing } s^2$$

- Akaike Information Criterion (AIC)

$$AIC = -2/T(\ln L - k) = -2 \ln L/T + 2k/T$$

$$\Rightarrow \text{if normality } AIC = \ln(\mathbf{e'e}/T) + (2/T)k \qquad (+\text{constants})$$

- Bayes-Schwarz Information Criterion (BIC)

$$BIC = -(2/T \ln L - [\ln(T)/T]k)$$

$$\Rightarrow \text{if normality } AIC = \ln(\mathbf{e'e}/T) + [\ln(T)/T]k \quad (+\text{constants})$$

Modeling Strategies: Information Criteria

• The goal of these criteria is to provide us with an easy way of comparing alternative model specifications, by ranking them.

General Rule: The lower the IC, the better the model. For the previous IC's, then choose model to minimize s^2_J, AIC_J, or BIC_J.

• Some remarks about IC's:

- They are used for ranking. The raw value tends to be ignored.

- They have two components: a *goodness of fit* component –based on *lnL*– and a model complexity component –the penalty based on *k*.

- Different penalties, different IC's.

- Some authors do not scale the IC's by *T,* like we do above. If raw values are irrelevant, this is not an issue.

Modeling Strategies: Information Criteria

• IC's are not test statistics. They do not test a model. But, they are statistics –i.e., they are functions of RVs- with sampling distributions.

• We would like these statistics –i.e., the IC's- to have good properties. For example, if the true model is being considered among many, we'd want the information criteria to select it. This can be done on average (unbiased) or as *T* increases (consistent).

• Usually, inconsistency is a fatal flaw for a statistics. But, in model selection, it is very likely that the true DGP is not among the models considered. That is, inconsistency may not matter in these cases.

• Information? It refers to Kullback and Leibler's (1951) *information discrepancy* measure, used in information theory (in telecom literature).

Modeling Strategies: IC - K-L divergence

• Kullback and Leibler's (1951) *information discrepancy* measure is also called *information divergence.*

• Information divergence measures the difference between two probability distributions P and Q; where P represents the true DGP. Here, we look at the difference between the expected values of Y when Y is determined by: (i) P and (ii) some Q model.

• Minimizing the K-L divergence, when considering several Q models, gets us close to the true DGP.

• But, expected values are unobservable, they need to be estimated. The information associated with Y is given by L –i.e., the joint pdf. The AIC uses $ln\ L$ evaluated at the estimated parameter values.

Modeling Strategies: IC – AIC and BIC

• Some results regarding AIC and BIC.

- AIC and Adjusted R^2 are not consistent.

- AIC is conservative –i.e., it tends to over-fit; that is, choose too large models.

- AIC selects the model that minimizes the leave-one-out cross-validation MSE for cross-sectional data. In time series, it selects the model that minimizes the out-of-sample one-step ahead forecast MSE.

- BIC is more parsimonious than AIC. It penalizes the inclusion of parameters more ($k_{BIC} \leq k_{AIC}$).

- BIC is consistent in hierarchical (*gets*) autoregressive models.

Modeling Strategies: IC – AIC and BIC

• There are several *small sample corrections* of IC's. But, asymptotically they have no impact. Because of this feature, using corrected IC's is not a bad choice.

• Comparing models based on IC's can be expensive

• In 'unstructured problems' (natural order to the hypotheses to be tested), there is a huge number of potential combinations to investigate: 2^m possible models for m candidate variables.

• For the Lovell (1983) database, that would be $2^{40} \approx 10^{12}$ models. Even at a USD 0.001 per model, that would cost USD 1 billion.

Non-nested Models and Tests

• Sometimes, we have two rival models to choose between, where neither can be nested within the other -i.e., neither is a restricted version of the other.

Example:

Model 1 $\quad \mathbf{Y} = \mathbf{X}\beta + \mathbf{W}\delta + \varepsilon$

Model 2 $\quad \mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \xi$

• If the dependent variable is the same in both models (as is the case here), we can simply use Adjusted-R^2 to rank the models.

• We can also use AIC and/or BIC.

Non-nested Models and Tests

• Alternative approach: Encompassing --Mizon and Richard (1986).

(1) Form a composite or *encompassing* model that nests both rival models --say, Model 1 and Model 2.

(2) Test the relevant restrictions of each rival model against it.

Assuming the restrictions cannot be rejected, we would prefer the model with the lower F statistic for the test of restrictions.

Example: From Model 1 and Model 2, the encompassing model is:

Encompassing Model: $\mathbf{Y = X\beta + W\delta + Z\gamma + \varepsilon}$

We now test, separately, the hypotheses (1) $\boldsymbol{\delta = 0}$ and (2) $\boldsymbol{\gamma = 0}$

• Note: We test a hybrid model. Also, multicollinearity may appear.

Non-nested Models: *J*-test

• Davidson-MacKinnon (1981)'s *J*-test .

We start with two non-nested models. Say,

Model 1 $\mathbf{Y = X\beta + \varepsilon}$

Model 2 $\mathbf{Y = Z\gamma + \xi}$

Idea: If Model 2 is true, then the fitted values from the Model 1, when added to the 2nd equation, should be insignificant.

• Steps:

(1) Estimate Model 1 => obtain fitted values: **Xb**.

(2) Add **Xb** to the list of regressors in Model 2. => $\mathbf{Y = Z\gamma + \lambda Xb + \xi}$

(3) Do a *t-test* on λ. A significant *t*-value would be evidence against Model 2 and in favour of Model 1.

(4) Repeat the procedure for the models the other way round.

(5) Rank the models on the basis of this test.

Non-nested Models: *J*-test

Note: As some of the regressors in step (3) are stochastic, the *t*-test is not strictly valid. Davidson and MacKinnon (1981) show that it is *asymptotically* valid.

• Of course neither of these methods may not reject both models. This is possible in small samples, even if one model, say Model 2, is true.

Rejecting both models is not unusual. McAleer's (1995), in a survey, reports that out of 120 applications all models were rejected 43 times.

• One would also want to examine the diagnostic test results when choosing between two models.

Non-nested Models: *J*-test

• *J*-test does not work very well when we compare 3 or more models.

• Encompassing interpretation of the *J*-test .
Let's encompass both models:

$$\mathbf{Y} = (1-\lambda)\,\mathbf{Z}\gamma + \lambda\mathbf{X}\beta + \varepsilon$$

Under H_0 (Model 2 is true): $\lambda=0$.
Under H_1 (Model 1 is true): $\lambda=1$.

Nice model, but unfortunately, this model is not intrinsic linear!

Non-nested Models: *J*-test - Application

• We want to test

H_0: $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon_0$ (additive) vs

H_1: $\ln \boldsymbol{y} = (\ln \boldsymbol{X})\,\gamma + \varepsilon_1$ (multiplicative)

• We look at the Davidson-MacKinnon J-test

Step 1: OLS on H_1: get **c**

OLS $\mathbf{y} = \lambda_0 \exp\{\ln(\mathbf{X})\,\mathbf{c}\} + \mathbf{X}\beta + \varepsilon$ => *t-test* on λ_0

Step 2: OLS on H_0: get **b**

OLS $\ln \mathbf{y} = \ln(\mathbf{Z})\gamma + \lambda_1\,\mathbf{Xb} + \varepsilon$ => *t-test* on λ_1

• Situations:

(1) Both OK: $\lambda_0=0$ and $\lambda_1=0$. => get more data

(2) Only 1 is OK: $\lambda_0\neq0$ and $\lambda_1=0$ (multiplicative is OK);

$\lambda_1\neq 0$ and $\lambda_0=0$ (additive is OK)

(3) Both rejected: $\lambda_0\neq0$ and $\lambda_1\neq 0$ => new model is needed.

Non-nested Models: *J*-test - Considerations

• The *J*-test was designed to test non-nested models (one model is the true model, the other is the false model), not for choosing competing models –the usual use of the test.

• The *J*-test is likely to over reject the true (model) hypothesis when one or more of the following features is present:

- a poor fit of the true model
- a low or moderate correlation between the regressors of the 2 models
- the false model includes more regressors than the correct model.

Davidson and MacKinnon (2004) state that the *J*-test will over-reject, *often quite severely* in finite samples when the sample size is small or where conditions (i) or (iii) above are obtained.

Modeling Strategies: Significance level, α

• So far, we have assumed that the distribution of the test statistic –say the *F*-statistic-- under H_0 is known exactly, so that we have what is called an *exact test***.**

• Technically, the *size of a test* is is the supremum of the rejection probability over all DGPs that satisfy H_0. For an exact test, the size equals the *nominal level*, α –i.e., the Prob[Type I error] = α.

• Usually, the distribution of a test is known only approximately *(asymptotically*). In this case, we need to draw a distinction between the nominal level (*nominal size*) of the test and the actual *rejection probability (empirical size*)**,** which may differ greatly from the nominal level.

• Simulations are needed to gauge the empirical size of tests.

Modeling Strategies: A word about α

• Ronald Fisher, before computers, tabulated distributions. He used a .10, .05, and .01 percentiles. These tables were easy to use and, thus, those percentile became the de-facto standard α for testing H_0.

• "It is usual and convenient for experimenters to take 5% as a standard level of significance." --Fisher (1934).

• Given that computers are powerful and common, why is $p = 0.051$ unacceptable, but $p = 0.049$ is great? There is no published work that provides a theoretical basis for the standard thresholds.

• Rosnow and Rosenthal (1989): " ... surely God loves .06 nearly as much as .05."

Modeling Strategies: A word about α

Practical advise: In the usual Fisher's null hypothesis (significance) testing, significance levels, α, are arbitrary. Make sure you pick one, say 5%, and stick to it throughout your analysis or paper.

• Report *p-values*, along with CI's. Search for *economic significance*.

• Q: .10, .05, or .01 significance?

Many tables will show *, **, and *** to show .10, .05, and .01 significance levels. Throughout the paper, the authors will point out the different significance levels. In these papers, it is not clear what α is the paper using for inference.

• In a Neyman-Pearson world, we can think of these stars (or *p-values*) as ways of giving weights to H_0 relative to H_1.

Modeling Strategies: A word about H_0

• In applied work, we only learn when we reject H_0. Failing to reject H_0 provides almost no information about the state of the world.

• Thus, failing to reject H_0 does not rule out an infinite number of other competing research hypotheses.

• Null hypothesis significance testing is asymmetric: if the test statistic is "too large" for a given H_0 then H_0 is rejected; but if the test statistic is not "too large" then H_0 is not automatically accepted.

• It is dangerous to "accept" the conclusion from a non-rejected H_0. But, it is common. Eight of the twenty (40%) articles in the *American Political Science Review* Volume 91 (1997), that used a H_0, drew substantive conclusions from a fail to reject decision.

Modeling Strategies: A word about H_0

• In applied work, we only learn when we reject H_0; say, when the *p*-value$<\alpha$. But, rejections are of two types:

- Correct ones, driven by the power of the test,
- Incorrect ones, driven by Type I Error ("statistical accident," luck).

• It is important to realize that, however small the *p*-value, there is always a finite chance that the result is a pure accident. At the 5% level, there is 1 in 20 chances that the rejection of H_0 is just luck.

• Since negative results are very difficult to publish (*publication bias*) this means that an unknown but possibly large number of false claims are taken as truths.

Example from Lecture 4: If α=0.05, proportion of false H_0=10%, and π=.50, **47.4%** of rejections are true H_0 -i.e., "false positives."

Model Selection Methods: Summary

• Eight literature strands can be delineated:
(1) *Specific-to-general*: Anderson (1962), Hendry and Mizon (1978), and Hendry (1979), for critiques;

(2) *Retaining the general model*: Yancey and Judge (1976), and Judge and Bock (1978);

(3) *Testing Theory-based models*: Hall (1978), criticized by Davidson and Hendry (1981), and Hendry and Mizon (2000); Stigum (1990) proposes a formal approach;

(4) Other 'rules' for model selection, such as:
- *step-wise regression*: Leamer (1983a), for a critical appraisal
- *'optimal' regression*: algorithm to maximize the Adj-R^2 with a specified set of regressors. See Coen, Gomme and Kendall (1969);

Model Selection Methods: Summary

• Eight literature strands can be delineated (continuation):
(5) *Model comparisons*, often based on non-nested hypothesis tests or encompassing: Cox (1961, 1962), Pesaran (1974), and the survey in Hendry and Richard (1989);

(6) *Model selection by information criteria*: Schwarz (1978), Hannan and Quinn (1979), Amemiya (1980);

(7) *Bayesian model comparisons*: Leamer (1978) and Clayton, Geisser and Jennings (1986);

(8) *General-to-specifics (gets)*: Anderson (1962), Sargan (1973, 1981), Hendry (1979), and White (1990).

Criteria for Model Selection: Judgement Call

• In the end, judgment must be used in weighing up various criteria:

- The Economic Criterion –are the estimated parameters plausible? (Economic Significance)

- The First Order Statistical Criterion –does the model provide a good fit (in-sample) with statistically significant parameter estimates?

- The Second Order Statistical Criterion –is the model generally free of misspecification problems – as evidenced in the diagnostic tests?

- The Out of Sample Predictive Criterion –does the model provide good out of sample predictions?

Model Selection: Causality and Identification

• In empirical work, we are interested in identifying causal relations, say from **X** to **y**, as implied in the DGP of the CLM: $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$.

• Suppose we have two correlated variables: Y_t & X_t. The co-variation in Y_t & X_t can be driven by (not mutually exclusive):

- *Causation from Y_t to X_t*: Changes in Y_t ⇒ changes in X_t
- *Causation from X_t to Y_t*: Changes in X_t ⇒ changes in Y_t
- *Correlated through a 3rd variable, W_t:* changes in W_t ⇒ changes X_t & Y_t

• In practice, it is not easy to say what generates variation in Y_t & X_t. The third case, especially when W_t is an unobservable variable, creates a lot complications.

Example: Y_t: earnings; X_t: schooling; W_t: ability.

Model Selection: Causality and Identification

• There are four approaches for identification:

- *Experiments*. The researcher generates the variation in the variables.
- *Natural Experiments*. An known exogenous event generates the variation in the variables.
- *Instrumental variables*. A variable provides variation.
- *Econometric Identification*. We use econometric assumptions for identification.

• In time series, there is the concept of *Granger causality,* where past changes in one variable affect the present values of another variable. This is not, strictly speaking, the causation we discuss here.

Model Selection: Causality and Identification

• To be precise, the identification problem in econometrics refers to the problem of identifying and estimating one or more coefficients of a system of simultaneous equations.

Model Selection: Causality and Identification

• *Experiments*

Experiments are popular in the sciences (say, biology, physics). For example, we want to test a new treatment. Then,

(1) A sample is divided randomly in two similar groups: *treated* group & *control* group. (A *randomized* study: Only difference is the treatment!)

(2) Look for differences in both groups.

⇒ Rare in economics and finance; they can be very expensive or unethical (say, exposing people to a "poverty shock"). Some work in small communities and small units in some businesses.

Problem: Not easy to randomize these man-made experiments that involve humans.

Model Selection: Causality and Identification

• *Natural experiments*

An exogenous (historical) event (not necessarily a nature event) provides a situation where groups can be reasonably randomized in a *treated* (affected by the natural event) and a *control group* (not affected by the natural event).

In the absence of experiments, natural experiments give us a very good way to identify causation.

Examples: Changes in tax code and regulations; changes in accounting standards, shocks, disasters, laws or rules that impose thresholds (*discontinuity*) for behaviors, etc. (More on Lectures 8 & 15.)

Problem: Not easy to generalize, not clear how robust results are.

Model Selection: Causality and Identification

• *Instrumental Variables*

Suppose we want to study the effect of networking on CEO compensation. Since CEO compensation and networking may be affected by the unobserved natural ability of an individual (W_i), a simple regression will be biased (omitted variables problem).

Suppose we have a variable, Z, correlated with networking, but not with natural ability (*ethnicity*?, *age*?, *number of childhood friends*?) –i.e., Z induces variation in X *unrelated* to W_i. Then, we use Z to study the effect of networking on CEO compensation.

We call Z an *instrument*. Usually, we can relate Z to a *natural experiment*.

Problem: As we will see later, in Lecture 8, finding Z is not easy.

Model Selection: Causality and Identification

• *Econometric Identification*

We think that networking is correlated with ability, then we model it. Actually, we model everything. Very transparent in the assumptions.

We end up with a Simultaneous Equations Models (SEM), which we will study later in Lecture 16.

Problem: They tend to be (very) complicated.