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

Model 2 $\quad \boldsymbol{y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\xi}$

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

Model 4 $\quad \boldsymbol{y} = \exp(\mathbf{Z}\,\mathbf{D}\,\boldsymbol{\delta}) + \boldsymbol{\epsilon}$

• 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 ($\boldsymbol{y}$, $\mathbf{X}$), 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: Two Methods

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

- *Specific to General*
- *General to Specific*

• Specific to General. Start with a small "restricted model," do some testing and make model bigger model in the direction indicated by the tests (for example, add variable x_k when test reject H_0: $\beta_k=0$).

• General to Specific. Start with a big "general unrestricted model," do some testing and reduce model in the direction indicated by the tests (for example, eliminate variable x_k when test cannot reject H_0: $\beta_k=0$).

Model Selection Strategies: Specific to General

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

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

• 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), and, in the machine learning literature, *forwards selection*.

• Popular implementation: *Stepwise Regression*.

Model Selection Strategies: Specific to General

Example: Specific-to-general strategy to model IBM returns:

(1) We start with the 3-factor FF model for IBM:

$$r_{i=IBM,t} - r_{f,t} = \alpha_i + \beta_{1,i}\,(r_{M,t} - r_{f,t}) + \beta_{2,i}\,SMB_t + \beta_{3,i}\,HML_t + \varepsilon_{i,t}$$

(2) Estimate the 3-factor FF model for IBM:

```
fit_ibm_ff3 <- lm (ibm_x ~ Mkt_RF + SMB + HML)
> summary(fit_ibm_ff3)
Coefficients:
             Estimate   Std. Error t value Pr(>|t|)
(Intercept) -0.005191   0.002482  -2.091   0.0369 *
Mkt_RF       0.910379   0.056784  16.032   <2e-16 ***
SMB         -0.221386   0.084214  -2.629   0.0088 **
HML         -0.139179   0.084060  -1.656   0.0983 .
---

Residual standard error: 0.05842 on 566 degrees of freedom
Multiple R-squared:  0.3393,    Adjusted R-squared:  0.3358
F-statistic:  96.9 on 3 and 566 DF,  p-value: < 2.2e-16
```

Model Selection Strategies: Specific to General

Example (continuation):

(3) Diagnostic tests: Check *t-stats* & R^2, F-test goodness of fit, etc.

(4) LM Test to test if there is a January Effect (H_0: No January effect):

```
> LM_test
[1] 9.084247
```

⇒ LM_test > **3.84** ⇒ Reject H_0: No January effect.

(5) Given this result, we modify the 3-factor FF and add the January Dummy to the FF model:

```
fit_ibm_new <- lm (ibm_x ~ Mkt_RF + SMB + HML + Jan_1)
> summary(fit_ibm_new)
Coefficients:
             Estimate   Std. Error t value Pr(>|t|)
(Intercept) -0.007302   0.002561  -2.851  0.00452 **
Mkt_RF       0.905182   0.056405  16.048  < 2e-16 ***
SMB         -0.247691   0.084063  -2.946  0.00335 **
HML         -0.154093   0.083606  -1.843  0.06584 .
Jan_1        0.026966   0.008906   3.028  0.00258 **
```

Model Selection Strategies: Specific to General

• The specific-to-general method makes assumptions along the way Some remarks based on the previous example:

(1) Very likely the starting model is based on theory and experience (HML is not significant at the usual 5% level). Not clear how to proceed from there to a more general model.

(2) We tested for a January effect and then added to the model. However, we could have tested for a Dot.com effect or for an interactive Dot.com/January effect with the 3 FF factors. Not clear when to stop the search.

(3) Select a p-value to add variables to the model. In this case, we use the standard 5% for the tests.

Model Selection Strategies: Specific to General

• Note that in the previous example, we started with a model. What happens if are skeptical regarding models?

• A popular implementation of the specific-to-general model selection is the *stepwise regression*, where we start with only a set of potential explanatory variables and let the data, based on some criteria (R^2, AIC, etc.), determine which variables to keep.

Model Selection Strategies: Stepwise Regression

• Overall structure:

- The method begins with a k potential regressors.

- Do k one-variable regressions. Pick the one that shows the biggest t-stat or maximizes a goodness of fit measure, say, Adjusted-R^2, $\overline{R}^2$. Suppose x_j is selected.

- Then, do $(k-1)$-variable regressions all with x_j. Select the regressor (in addition to x_j) that has the highest t-stat or that maximizes $\overline{R}^2$.

- Continue. But, when we start adding regressors, we usually check if the added regressor(s) change the significance of previous steps. (Note: at each step, we remove or add a regressor(s) based on t- or F-tests.)

- Stop: Additional regressors do not have *significant* t-stats/increase $\overline{R}^2$.

• Decisions: Selection of k variables, α for tests (α = 5%, 10%, 20%?) and goodness of fit statistic.

Model Selection Strategies: Stepwise Regression

• Overall structure:

- The method begins with a k potential regressors.

- Do k one-variable regressions. Pick the one that shows the biggest t-stat or maximizes a goodness of fit measure, say, Adjusted-R^2, $\overline{R}^2$. Suppose x_j is selected.

- Then, do $(k-1)$-variable regressions all with x_j. Select the regressor (in addition to x_j) that has the highest t-stat or that maximizes $\overline{R}^2$.

- Continue. But, when we start adding regressors, we usually check if the added regressor(s) change the significance of previous steps. (Note: at each step, we remove or add a regressor(s) based on t- or F-tests.)

- Stop: Additional regressors do not have *significant* t-stats/increase $\overline{R}^2$.

• Decisions: Selection of k variables, α for tests (α = 5%, 10%, 20%?) and goodness of fit statistic.

Model Selection Strategies: Stepwise Regression

• Decisions: Selection of k initial variables, α for tests (α = 5%, 10%, 30%?) and goodness of fit statistic.

Remark: Always keep in mind that the selected (final) model is not necessarily better than others. Type I and Type II errors are likely to occur, thus the final model may have irrelevant and/or omitted variables.

Technical Note: Though popular in practice, in general, selecting variables based on *p-values* is not advised, since the distribution of the OLS coefficients is affected. (Recall pre-testing.)

Model Selection Strategies: Stepwise Regression

Example: Stepwise regression strategy to model IBM returns. We start with the 5 FF factors as candidates for IBM. We use the function *ols_step_forward_p* in the *olsrr* package, which uses *p-values* to select:

```
library(olsrr)
ff_step_data <- data.frame(Mkt_RF, SMB, HML, RMW, CMA)
ibm_ff_model <- lm(ibm_x ~ ., data = ff_step_data)      # default p-value (penter) is 0.3
ols_step_forward_p(ibm_ff_model , details = TRUE)      #  long final output
```

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-0.005	0.002		-1.999	0.046	-0.010	0.000
Mkt_RF	0.887	0.055	0.574	16.227	0.000	0.780	0.995
SMB	-0.261	0.088	-0.111	-2.960	0.003	-0.435	-0.088
RMW	-0.128	0.114	-0.042	-1.122	0.262	-0.351	0.096

Model Selection Strategies: Stepwise Regression

Example (continuation):

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Mkt_RF	0.3087	0.3075	7.7108	-1665.5551	0.0594
2	SMB	0.3174	0.3151	2.2117	-1671.0548	0.0590
3	RMW	0.3188	0.3154	2.9552	-1670.3207	0.0590

Conclusion: The Stepwise Regression method selects Market excess returns, SMB & RMW as the drivers of IBM excess returns.

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:

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

• 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:

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

• 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), and, in the ML literature, *backwards selection*. It was pioneered by Sargan (1964). The properties of gets are discussed in Hendry 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)

Model Selection 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*.

Model Selection Strategies: General to Specific

Example: General-to-specific strategy to model IBM returns:

Step 1 - Start with a GUM: the 3-factor FF model for IBM + January Dummy + Dot.com Dummy + non-linear & interactive effects:

$$(r_{IBM} - r_f)_t = \beta_0 + \beta_1 (r_M - r_f)_t + \beta_2 SMB_t + \beta_3 HML_t + \\ + \beta_4 \text{January}_t + \beta_5 (r_M - r_f)_t^2 + \beta_6 SMB_t^2 + \beta_7 HML_t^2 \\ + \beta_8 (r_M - r_f)_t * SMB_t + \beta_9 (r_M - r_f)_t * \text{HML}_t + \\ + \beta_{10} \text{Dot.com}_t + \beta_{11} (r_M - r_f)_t * \text{January}_t + \\ + \beta_{12} HML_t * \text{January}_t + \beta_{13} (r_M - r_f)_t * \text{Dot.com}_t \\ + \beta_{14} HML_t * \text{Dot.com} + \beta_{15} SMB_t * \text{Dot.com} + \varepsilon_t$$

Step 1 - Estimate GUM:

```
Mkt_Jan <- Mkt_RF * Jan_1
HML_Jan <- HML * Jan_1
Mkt_Dot <- Mkt_RF * Dot_com
HML_Dot <- HML * Dot_com
SMB_Dot <- SMB * Dot_com
```

Model Selection Strategies: General to Specific

Example (continuation):

fit_ibm_gum <- lm (ibm_x ~ Mkt_RF + SMB + HML + Jan_1 + Mkt_RF_2 + SMB_2 + HML_2 + Mkt_HML + Mkt_SMB + SMB_HML + Mkt_Jan + HML_Jan + Mkt_Dot + HML_Dot + SMB_Dot)

> summary(fit_ibm_gum)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.007836	0.003063	**-2.559**	0.010772 *	
Mkt_RF	0.791866	0.090474	**8.752**	< 2e-16 ***	
SMB	-0.295790	0.110655	**-2.673**	0.007738 **	
HML	-0.233942	0.135146	-1.731	0.084004 .	⇒ practice says "keep it." Judgement call.
Jan_1	0.031769	0.009349	**3.398**	0.000727 ***	
Mkt_RF_2	-0.433762	0.850899	-0.510	0.610417	
SMB_2	-0.927271	1.470645	-0.631	0.528615	
HML_2	2.707992	1.670366	1.621	0.105545	⇒ almost 10%, I keep it. Judgement call.
Mkt_HML	0.628721	1.557090	0.404	0.686531	
Mkt_SMB	0.791625	1.746939	0.453	0.650618	
SMB_HML	-1.044806	2.029091	-0.515	0.606819	
Mkt_Jan	-0.069413	0.189309	-0.367	0.714008	
HML_Jan	-0.259697	0.255484	-1.016	0.309841	

Model Selection Strategies: General to Specific

Example (continuation):

	Estimate	Std. Error	t value	Pr(>|t|)
Mkt_Dot	0.323382	0.130645	**2.475**	0.013612 *
HML_Dot	0.059742	0.208277	0.287	0.774342
SMB_Dot	0.076998	0.198964	0.387	0.698910

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05788 on 553 degrees of freedom

Multiple R-squared: 0.3663, Adjusted R-squared: 0.3491

F-statistic: 21.31 on 15 and 553 DF, p-value: < 2.2e-16

Step 1 – Check GUM residuals for departures of (**A2**)-(**A3**). A Ramsey's reset test can be done (using the *resettest* in the *lmtest* library).

> resettest(fit_ibm_gum, type="fitted")

RESET test

data: fit_gum

RESET = **1.2645**, df1 = 2, df2 = 551, p-value = **0.2832**

Model Selection Strategies: General to Specific

Example (continuation):

Step 2 – Reduce Model with t-test and F-tests. Say, we keep all the variables with a *p-value* close to 10% (we still keep HML, using previous experience). We estimate a restricted GUM:

```
fit_ibm_gum_r <- lm (ibm_x ~ Mkt_RF + SMB + HML + Jan_1 + HML_2 + Mkt_Dot)
> summary(fit_ibm_gum_r)
Coefficients:
             Estimate    Std. Error t value Pr(>|t|)
(Intercept)  -0.008696   0.002788  -3.119  0.00191 **
Mkt_RF        0.779336   0.072453  10.756  < 2e-16 ***
SMB          -0.280018   0.083891  -3.338  0.00090 ***
HML          -0.250480   0.088504  -2.830  0.00482 **
Jan_1         0.028499   0.008937   3.189  0.00151 **
HML_2         1.676011   1.331161   1.259  0.20853
Mkt_Dot       0.344030   0.116685   2.948  0.00333 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Selection Strategies: General to Specific

Example (continuation):

Step 2 – Test the restrictions implied by the specific model against the general model. Using an F-test, we test J=9 restrictions:

$$H_0: \beta_5 = \beta_6 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{14} = \beta_{15} = 0$$

```
e_u <- fit_ibm_gum$residuals                          # GUM residuals
RSS_u <- t(e_u)%*%e_u
e_r <- fit_ibm_gum_r$residuals                        # Restricted GUM residuals
RSS_r <- t(e_r)%*%e_r
f_test_gum <- ((RSS_r - RSS_u)/9)/(RSS_u/(T-16))      # F-test
> f_test_gum
     [,1]
[1,] 0.4299497          ⇒ we cannot reject H0 (f_test_gum < qchisq(.95, 9, 553) = 1.896801)
> qf(.95, df1=9, df2=T-16)
[1] 1.896801
p_val <- 1 - pf(f_test_gum, df = 9 , df2=T-16)        # p-value of F-test
>  p_val
[1,] 0.919105           ⇒ p-value is almost 1. No evidence for H0.
```

Model Selection Strategies: General to Specific

Example (continuation):

Step 2 – Further specification checks of Restricted GUM, for example, perform a Ramsey's reset test (using the *resettest* in the lmtest library).

```
> resettest(fit_gum_r, type="fitted")

        RESET test

data:  fit_ibm_gum_r
RESET = 1.0998, df1 = 2, df2 = 561, p-value = 0.3337
```

Step 3 - Test if Restricted GUM residuals are acceptable –i.e., do diagnostic tests (mainly, make sure they are white noise). If Restricted GUM passes all the diagnostic tests, it becomes the "final model."

Note: With the final model, we use it to justify/explain financial theory and features, and do forecasting.

Model Selection Strategies: General to Specific

• The general-to-specific method makes assumptions along the way.

Some remarks based on the previous example:

(1) Select a *p-value* for the tests of significance in the discovery stage (we use **10%**). Given that we performed **15** *t-tests*, we should not be surprised we rejected the GUM, since we had an overall significance, $\alpha^* = .79$ [= 1 – (1 - **.10**)^**15**]. *Mass significance* is an issue.

(2) Judgement calls are also made.

(3) The reduction of the GUM involves "*pre-testing*" –i.e., data mining. We are likely rejecting a true H_0 (false positives) & not rejecting a true H_1 (false negatives), along the way. This increases the probability that the final model is not a good approximation. It is common to ignore (or not even acknowledge) pre-testing issues.

Model Selection Strategies: Best Subset

• Begin with a big model, with k regressors:

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

The idea is to select the "best" subset of the k regressors in $\mathbf{X}$, where "best" is defined by the researcher, say MSE, Adjusted-R^2, etc.

• In theory, it requires 2^k regressions. It can take a while if k is big ($k < 40$ is no problem).

• Many tricks are used to reduce the number of regressions.

• In practice, we use best subset to reduce the number of models to consider. For example, from the regressions with one-variable, keep the best one-variable model, from the regression with two-variables, keep the best two-variable model, etc.

Model Selection Strategies: Best Subset

Example: We want to select a model for IBM excess returns, using the k=3 Fama-French factors: Market excess returns (Mkt_RF), SMB, & HML. We have 8 (=2^3) models and, thus, regressions:

1) Constant;
2) Mkt_RF (CAPM)
3) SMB
4) HML
5) Mkt_RF & SMB
6) Mkt_Rf & HML
7) SMB & HML
8) Mkt_RF, SMB, & HML (the 3-factor F-F Model).

• We select the model with the lower MSE. Or, we can carry two or three models of the best models to do *cross-validation*.

Model Selection Strategies: Best Subset

Example (continuation): We use library olsrr in R:

```
library(olsrr)
ff_step_data <- data.frame(Mkt_RF, SMB, HML)
fit_ibm_ff3_sb <- lm(ibm_x ~ ., data = ff_step_data)        # default p-value (penter) is 0.3
ols_step_best_subset(fit_ibm_ff3_sb, details = TRUE)        #  long final output
```

```
Model Index    Predictors
---------------------------
    1        Mkt_RF
    2        Mkt_RF SMB
    3        Mkt_RF SMB HML
---------------------------
```

Subsets Regression Summary

Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE
1	0.3128	0.3116	0.308	8.3178	-1705.0204	-3424.8023	-1691.7998	2.1146	0.0035
2	0.3214	0.3192	0.3134	2.6125	-1710.7200	-3430.4398	-1693.0924	2.0913	0.0035
3	0.3221	0.3187	0.311	4.0000	-1709.3362	-3429.0366	-1687.3018	2.0927	0.0035

Model Selection Strategies: Best Subset

Example (continuation): Suppose we selected three model: CAPM (M1); Mkt_RF & SMB (M2); and the 3-factor F-F Model (M3).

Now, we use K*-fold cross-validation*, with K=5.

CV_5 M1: 0.003542756

CV_5 M2: **0.003505873**

CV_5 M3: 0.003556918

Note: Models look very similar. Practitioners compute a SE for CV_K and use a one SE rule. If within one SE, keep simplest model (M1).

Issues: Pre-testing

• A special case of omitted variables.

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

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

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

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

- 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, $\mathbf{b}^*$, or the OLS estimator, $\mathbf{b}$:

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

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 $\mathbf{b} \Rightarrow$ 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, $(100 - \pi)$% of the time, $\mathbf{R\beta}=\mathbf{q}$ will be incorrectly imposed, $\mathbf{b}^*$ will be used $\Rightarrow$ 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

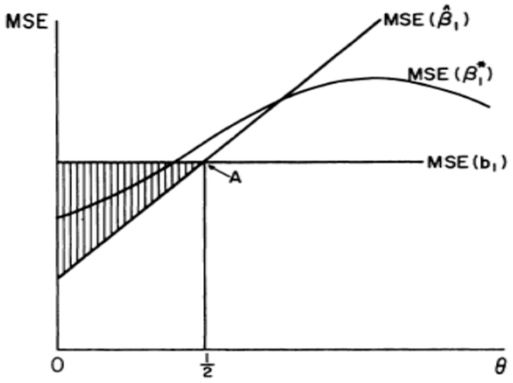
• Checking the MSE of $\mathbf{b}_{PT}$, $\mathbf{b}^*$ and $\mathbf{b}$ helps to evaluate the problem

• 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: $\alpha^* \leq k\alpha$ & $\alpha \geq \alpha^*/k$

$\Rightarrow$ close to the "independent" values for small α, but can differ for large α.

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

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

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

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

α^* =0.20 and k=20 $\Rightarrow$ 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: 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) $\overline{R}^2$)

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

$\Rightarrow$ maximizing Adjusted R^2 $\Leftrightarrow$ minimizing s^2

- Akaike Information Criterion (AIC)

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

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

- Bayes-Schwarz Information Criterion (BIC)

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

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

Modeling Strategies: Other Criteria

• A related criteria is Mallows' (1973) C_p statistic (Notation: p = k):

$$C_p = RSS(k)/s^2 - T + 2 * k$$

where RSS(k) is the RSS for the model with k regressors.

It can be shown that the C_p-statistic estimates the size of the bias that is introduced into the predicted responses by having omitted variables –i.e., an underspecified model.

It has useful properties for selection of regressors:

- For a model that fits the data "adequately" ⇒ $E[C_p] \approx k$
- For the full model (no bias), with k parameters ⇒ $E[C_p] = k$.

• Other popular statistics: RIC (Risk Inflation Criteria), FPE, OOS R^2.

Modeling Strategies: Model Validation

• Cross validation, as in Lecture 5, can be used to select a model. For example, K*-fold cross-validation*. We have already done this in combination with Best subset.

Example: Suppose using best subsets to model IBM excess returns, using the k=3 Fama-French factors, we selected three model: CAPM (M1); Mkt_RF & SMB (M2); and the 3-factor F-F Model (M3).

Now, we use K*-fold cross-validation*, with K=5.

CV_5 M1: 0.003542756

CV_5 M2: **0.003505873**

CV_5 M3: 0.003556918

Note: Models look very similar. Practitioners compute a SE for CV_K and use a one SE rule. If within one SE, keep simplest model (M1).

Testing Model Specification: Non-Nested Models

Example:

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

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

- 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 and select the one with the largest Adjusted-R^2.
- We can also use AIC and/or BIC to rank the models.
- But, we can also use more sophisticated, testing-based, methods.

Non-nested Models and Tests: Encompassing

- **Testing-based Method 1:** Encompassing

(1) Form a composite or *encompassing* model that nests both rival models −Model 1 & Model 2. This is the **unrestricted Model**, **ME**.

(2) Test the relevant restrictions of each rival model against **ME**. We do two F-tests:

(i) Test **ME** (Unrestricted Model) against Model 1 (Restricted Model)
(ii) Test **ME** (Unrestricted Model) against Model 2 (Restricted Model)

- If we reject the restrictions against one Model, say Model 1, and we cannot reject the restrictions against the other, Model 2, we are done: We select the Model that the F test do not reject restrictions (Model 2).

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

Non-nested Models and Tests: Encompassing

Example: We have:

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

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

Then, the **Encompassing Model** (**ME**) is:

ME: $\quad \boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\delta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$

Now test, separately, the hypotheses (1) $\boldsymbol{\delta} = \mathbf{0}$ and (2) $\boldsymbol{\gamma} = \mathbf{0}$. That is,

F-test for H_0: $\boldsymbol{\gamma} = \mathbf{0}$: ME (U Model) vs Model 1 (R Model).
F-test for H_0: $\boldsymbol{\delta} = \mathbf{0}$: ME (U Model) vs Model 2 (R Model).

If we reject H_0: $\boldsymbol{\gamma} = \mathbf{0} \Rightarrow$ We have evidence against Model 1
If we reject H_0: $\boldsymbol{\delta} = \mathbf{0} \Rightarrow$ We have evidence against Model 2.

Note: We test a hybrid model, a combination of two models. Also, multicollinearity may appear.

Non-nested Models and Tests: IFE or PPP?

- Two of the main theories to explain the behaviour of exchange rates, S_t, are the **International Fisher Effect** (**IFE**) and the **Purchasing Power Parity** (**PPP**). We use the direct notation for S_t, that is, units of Domestic Currency per 1 unit of Foreign currency.

- IFE states that, in equilibrium, changes in exchange rates (**e**) are driven by the interest rates differential between the domestic currency, i_d, and the foreign currency, i_f:. A DGP consistent with IFE is:

$$\mathbf{e} = \alpha^1 + \beta^1 (\mathbf{i_d} - \mathbf{i_f}) + \boldsymbol{\varepsilon}^1$$

- Relative PPP states that that, in equilibrium, **e** are driven by the inflation rates differential between the domestic Inflation rate, I_d, and the foreign Inflation rate, I_f. A GDP consistent with IFE is:

$$\mathbf{e} = \alpha^2 + \beta^2 (\mathbf{I_d} - \mathbf{I_f}) + \boldsymbol{\varepsilon}^2$$

- Theories are non-nested, use non-nested methods to pick a model.

Non-nested Models and Tests: IFE or PPP?

Example: What drives log changes in exchange rates for the USD/GBP (e): $(i_d - i_f)$ or $(I_d - I_f)$?

Model 1 (IFE): $\mathbf{e = \alpha^1 + \beta^1 (i_d - i_f) + \varepsilon^1}$

Model 2 (PPP): $\mathbf{e = \alpha^2 + \beta^2 (I_d - I_f) + \varepsilon^2}$

```
SF_da <- read.csv("http://www.bauer.uh.edu/rsusmel/4397/SpFor_prices.csv", head=TRUE, sep=",")
x_date <- SF_da$Date
x_S <- SF_da$GBPSP
x_F3m <- SF_da$GBP3M
i_us3 <- SF_da$Dep_USD3M
i_uk3 <- SF_da$Dep_UKP3M
cpi_uk <- SF_da$UK_CPI
cpi_us <- SF_da$US_CPI
T <- length(x_S)
int_dif <- (i_us3[-1] - i_uk3[-1])/100
lr_usdgbp <- log(x_S[-1]/x_S[-T])
I_us <- log(cpi_us[-1]/cpi_us[-T])
I_uk <- log(cpi_uk[-1]/cpi_uk[-T])
inf_dif <- (I_us - I_uk)
```

Non-nested Models and Tests: IFE or PPP?

Example (continuation): Encompassing Model

$\mathbf{e = \alpha + \beta_1 (i_d - i_f) + \beta_2 (I_d - I_f) + \varepsilon^1}$

```
# Encompassing Model and Test
fit_e <- lm( y ~ int_dif + inf_dif)

> summary(fit_e)

Coefficients:
             Estimate   Std. Error t value Pr(>|t|)
(Intercept) -0.0009633  0.0016210  -0.594   0.5527
int_dif     -0.0278510  0.0741189  -0.376   0.7073        ⇒ cannot reject H0: β1 = 0.
inf_dif      0.7444711  0.3429106   2.171   0.0306 *      ⇒ reject H0: β2 = 0.
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02662 on 360 degrees of freedom
Multiple R-squared:  0.01316,   Adjusted R-squared:  0.007673
F-statistic: 2.399 on 2 and 360 DF,  p-value: 0.09221
```

Non-nested Models and Tests: *J*-test

• **Testing-based Method 1:** Davidson-MacKinnon (1981)'s *J*-test.
We start with two non-nested models. Say,

Model 1: $\boldsymbol{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$

Model 2: $\boldsymbol{y} = \mathbf{Z}\gamma + \xi$

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** $\Rightarrow$ obtain fitted values: $\mathbf{Xb}$.

(2) Add $\mathbf{Xb}$ to the list of regressors in Model 2

$\Rightarrow \boldsymbol{y} = \mathbf{Z}\gamma + \lambda\mathbf{Xb} + \xi$

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

Non-nested Models: *J*-test

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

(4.1) Estimate **Model 2** $\Rightarrow$ obtain fitted values: $\mathbf{Zc}$.

(4.2) Add $\mathbf{Zc}$ to the list of regressors in Model 1:

$\Rightarrow \boldsymbol{y} = \mathbf{X}\beta + \lambda\,\mathbf{Zc} + \boldsymbol{\varepsilon}$

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

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

• It is possible that we cannot reject both models. This is possible in small samples, even if one model, say Model 2, is true.

• It is also possible that both *t-tests* reject H_0 ($\lambda \neq 0$ & $\lambda \neq 0$). This is not unusual. McAleer's (1995), in a survey, reports that out of 120 applications all models were rejected 43 times.

Non-nested Models: *J*-test

Technical Note: As some of the regressors in step (3) are stochastic, Davidson and MacKinnon (1981) show that the *t-test* is *asymptotically* valid.

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

Non-nested Models: *J*-test – IFE or PPP?

Example: Now, we test Model 1 vs Model 2, using the *J*-test.

Model 1 (IFE): $\mathbf{e} = \alpha^1 + \beta^1 (\mathbf{i_d} - \mathbf{i_f}) + \boldsymbol{\varepsilon}^1$

Model 2 (PPP): $\mathbf{e} = \alpha^2 + \beta^2 (\mathbf{I_d} - \mathbf{I_f}) + \boldsymbol{\varepsilon}^2$

```
y <- lr_usdgbp
fit_m1 <- lm( y ~ int_dif)
summary(fit_m1)
y_hat1  <- fitted(fit_m1)
fit_J1 <- lm( y ~ inf_dif + y_hat1)
summary(fit_J1)

fit_m2 <- lm( y ~ inf_dif)
summary(fit_m2)
y_hat2  <- fitted(fit_m2)
fit_J2 <- lm( y ~ int_dif + y_hat2)
summary(fit_J2)
```

Non-nested Models: *J*-test – IFE or PPP?

Example (continuation):

```
> fit_m1 <- lm( y ~ int_dif)
> y_hat1  <- fitted(fit_m1)
> fit_J1 <- lm(formula = y ~ inf_dif + y_hat1)
> summary(fit_J1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.136310 -0.014168  0.000351  0.017227  0.092421

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0001497  0.0025556   0.059   0.9533
inf_dif     0.7444711  0.3429106   2.171   0.0306 *
y_hat1      1.2853298  3.4206106   0.376   0.7073        ⇒ cannot reject H0: λ=0. (Good for Model 2)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02662 on 360 degrees of freedom
Multiple R-squared:  0.01316,   Adjusted R-squared:  0.007673
F-statistic: 2.399 on 2 and 360 DF,  p-value: 0.09221
```

Non-nested Models: *J*-test – IFE or PPP?

Example (continuation):

```
> fit_m2 <- lm( y ~ inf_dif)
> y_hat2  <- fitted(fit_m2)
> fit_J2 <- lm(formula = y ~ int_dif + y_hat2)
> summary(fit_J2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.136310 -0.014168  0.000351  0.017227  0.092421

Coefficients:
             Estimate    Std. Error t value Pr(>|t|)
(Intercept) -0.0003045  0.0016409  -0.186   0.8529
int_dif     -0.0278510  0.0741189  -0.376   0.7073
y_hat2       1.0066945  0.4636932   2.171   0.0306 *     ⇒ Reject H0: λ=0. (Again, good for Model 2)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02662 on 360 degrees of freedom
Multiple R-squared:  0.01316,   Adjusted R-squared:  0.007673
F-statistic: 2.399 on 2 and 360 DF,  p-value: 0.09221
```

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 *J*-test

Step 1: OLS on H_1: get $\hat{\gamma}$

OLS $\boldsymbol{y} = \mathbf{X}\beta + \lambda_1 \exp\{\ln(\mathbf{X})\,\hat{\gamma}\} + \varepsilon$ $\Rightarrow$ *t-test* on λ_1

Step 2: OLS on H_0: get $\mathbf{b}$

OLS $\ln \boldsymbol{y} = (\ln \boldsymbol{X})\,\gamma + \lambda_0\, \mathbf{X}\mathbf{b} + \varepsilon$ $\Rightarrow$ *t-test* on λ_0

• Situations:

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

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

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

Non-nested Models: *J*-test – Many Models

• *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:

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

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 - 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:
i) A poor fit of the true model
ii) A low/moderate correlation between the regressors of the 2 models
iii) 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 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 difficult to publish (*publication bias*), there is an unknown but possibly large number of false claims taken as truths.

Example (from Lecture 4): If $\alpha = 0.05$, proportion of false H_0=10%, and $\pi = .50$, **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 validation, with the different flavours, can be used here.

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 (of variation):

- *Experiments*. The researcher generates the variation in the variables.

- *Natural Experiments*. A 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 (Covid-19, stock market crisis), disasters (earthquakes, floods, etc.), 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.