Week 3: Lag Operators Cointegration (continued), Forecasting

Advanced Econometrics 4EK608

Vysoká škola ekonomická v Praze

Outline

Czech terminology

Operátory zpoždění
Superkonzistence
Grangerův reprezentační teorém
Engle-Grangerova dvoustupňová procedura
Kointegrační vektor

Lag operators

Lag operators:

$$\begin{aligned} Lx_t &= x_{t-1} \\ L(Lx_t) = L^2x_t &= x_{t-2} \\ &\dots \\ L^px_t &= x_{t-p} \end{aligned}$$

Using lag operators,

$$AR(p) : x_t = \alpha + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + u_t$$

can be rewritten as:

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)x_t = \alpha + u_t$$

Stationarity

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p) x_t = \alpha + u_t \tag{1}$$

Stochastic process (1) will only be stationary if the roots of corresponding equation (2) are all greater than unity in absolute value

$$1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p = 0 \tag{2}$$

Illustration 1 - AR(1) process:

$$x_t = \alpha + \phi x_{t-1} + u_t \tag{3}$$

$$(1 - \phi L) x_t = \alpha + u_t$$

$$1 - \phi L = 0$$

$$L = 1/\phi$$

For (3) to be stationary, $|L| > 1 \leftrightarrow -1 < \phi < 1$

Stationarity

Illustration 2:

$$x_t = 2 + 3.9x_{t-1} + 0.6x_{t-2} - 0.8x_{t-3} + u_t \tag{4}$$

To evaluate stationarity of x_t, we use

$$1 - 3.9L - 0.6L^2 + 0.8L^3 = 0,$$

which can be factorized:

$$(1 - 0.4L)(1 + 0.5L)(1 - 4L) = 0$$

$$1^{st} root : L = 2.5$$
$$2^{nd} root : L = -2$$
$$3^{rd} root : L = 0.25 \Rightarrow (4) \text{ is non-stationary}$$

Cointegration between two variables

Superconsistency: $y_t = \beta_0 + \beta_1 x_t + u_t$

1. Provided x_t and y_t are cointegrated, the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ will be consistent.
2. $\hat{\beta}_j$ converge in probability to their true values β_j more quickly in the cointegrated non-stationary case than in the stationary case (asymptotic efficiecy).

Consequences:

For simple static regression between two cointegrated variables: $y_t, x_t \sim C(1,1)$, super-consistency applies (with deterministic regressors such as intercept and trend added upon relevance). Dynamic misspecifications do not necessarily have serious consequences. This is a large sample property - in small samples, OLS estimators are biased.
(Specific statistical inference applies to cointegrating vectors.)

Cointegration between two variables

Granger representation theorem:[1]
If two TS x_t and y_t are cointegrated, the short-term disequilibrium relationship between them can be expressed in the ECM form

$$\Delta y_t = lagged(\Delta y, \Delta x) - \delta u_{t-1} + \varepsilon_t \tag{5}$$

where $u_{t-1} = y_{t-1} - \beta_0 - \beta_1 x_{t-1}$ is the disequilibrium error and δ is a short-run adjustment parameter.
Note: as u is on the scale of y, δ can be interpreted in percentages.
Example: $\delta = 0.8 \rightarrow 80\%$ of the disequilibrium error gets corrected between $t-1$ and t (on average).

Two implications:

1. The general-to-specific model search can focus on ECMs
2. Engle-Granger two-stage procedure

[1] Engle and Granger (1987)

Cointegration between two variables

Engle-Granger two-stage procedure:

We short-cut the search of an ECM from a general model

1^{st} stage: Estimation of the cointegrating (static) regression and saving residuals

$$\hat{u}_t = y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t$$

2^{nd} stage: Use residuals $\hat{u}_{t-1}$ in (5) instead of u_{t-1} and estimate by OLS

Estimators are consistent and asymptotically efficient, but biased in small samples.

Assumptions: y_t and x_t are non-stationary and cointegrated.

Cointegration among more than two variables

Possibility of more cointegrating vectors
long-run: $y_t = \beta_0 + \beta_1 x_t + \beta_2 w_t + \beta_3 z_t + u_t$,
all observed variables are $I(1)$

If this long-run relationship exists, then the disequilibrium error

$$u_t = [y_t - \beta_0 - \beta_1 x_t - \beta_2 w_t - \beta_3 z_t] \quad \sim I(0) \qquad (6)$$

In the multivariate case, there may be more than one stationary linear combination linking cointegrated variables.

If a linear combination of variables such as (6) is stationary, then the coefficients in this relationship form a cointegrating vector, e.g. $(1, -\beta_1, -\beta_2, -\beta_3)$.

Cointegration: the existence of at least one cointegrating vector.

Cointegration among more than two variables

Testing and estimation
Cointegration can be tested using the EG and/or PO tests

Only one cointegrating vector exists
Estimation can proceed by the Engle-Granger two-stage method for ECMs.

Two or more cointegrating vectors
Engle-Granger two-stage method is not applicable. Johansen (1988) suggests a maximum likelihood approach.

Predictions - basics

- LRM and its estimate:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

- Prediction of expected value:

$$\hat{y}_p = E(y|x_1 = c_1, x_2 = c_2, \ldots, x_k = c_k)$$
$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \cdots + \hat{\beta}_k c_k$$

- Rough (underestimated) confidence interval for the expected value prediction: (95%): $\hat{y}_p \pm 2 \times s.e.(\hat{y}_p)$. (Rule of thumb)

Predictions - basics

$s.e.(\hat{y}_p)$ can be obtained by reparametrization:

- Reparametrized LRM:

$$y^* = \beta_0^* + \beta_1^*(x_1 - c_1) + \beta_2^*(x_2 - c_2) + \cdots + u$$

- The following holds:

$$\begin{aligned} \hat{y}_p &= \hat{\beta}_0^* \\ s.e.(\hat{y}_p) &= s.e.(\hat{\beta}_0^*), \quad i.e. \\ Var(\hat{y}_p) &= Var(\hat{\beta}_0^*) \end{aligned}$$

Predictions - basics

- Predicted and actual values of y_p:

$$\hat{y}_p = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \cdots + \hat{\beta}_k c_k$$
$$y_p = \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k + u_p$$

- Prediction error

$$\hat{e}_p = y_p - \hat{y}_p = (\beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k) + u_p - \hat{y}_p$$

- Prediction error variance

$$Var(\hat{e}_p) = Var(u_p) + Var(\hat{y}_p)$$

because $Var(\beta_0 + \beta_1 c_1 + \beta_2 c_2 + \cdots + \beta_k c_k) = 0$

Predictions - basics

- If homoskedasticity holds, $\sigma^2 = Var(u_p)$:
  - $Var(\hat{e}_p) = \sigma^2 + Var(\hat{y}_p)$
  - We estimate σ^2 from the original LRM as $(SSR/(n-k-1))$
  - We get $Var(\hat{y}_p)$ from the reparametrized LRM
- Standard prediction error:
  - $s.e.(\hat{e}_p) = \sqrt{Var(\hat{e}_p)}$
- Prediction interval (95%)
  - $\hat{y}_p \pm t_{0.025} \times s.e.(\hat{e}_p)$

Predictions - basics

- Prediction with logarithmic dependent variable

$$\log(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$
$$\widehat{\log(y)} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

$\hat{y} = e^{\widehat{\log(y)}}$ systematically underestimates $\hat{y}$,

we can use a correction: $\hat{y} = \widehat{\alpha}_0 e^{\widehat{\log(y)}}$

where $\widehat{\alpha}_0 = n^{-1} \sum_{i=1}^{n} \exp(\hat{u}_i)$

is a consistent (not unbiased) estimator of $\exp(u)$.

Predictions - basics (Matrix form)

Prediction based on estimated model:

$$\hat{y}_p = \boldsymbol{x}_p' \hat{\boldsymbol{\beta}}$$

Difference between prediction and actual y_p value:

$$\hat{e}_p = \hat{y}_p - y_p = \boldsymbol{x}_p' \hat{\boldsymbol{\beta}} - \boldsymbol{x}_p' \boldsymbol{\beta} - u_p = \boldsymbol{x}_p' (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - u_p$$

If $\hat{\boldsymbol{\beta}}$ is unbiased estimator for $\boldsymbol{\beta}$,
$\hat{y}_p$ is an unbiased estimator for y_p value:

$$E(\hat{e}_p) = E(\hat{y}_p - y_p) = \boldsymbol{x}_p' E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + E(-u_p) = 0$$

and the variance of $\hat{e}_p$ can be expressed as:

$$E(\hat{e}_p^2) = var(\hat{e}_p) = \boldsymbol{x}_p' var(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_p + var(u_p)$$

Predictions - basics (Matrix form)

Variance of $\hat{e}_p$ (continued):

$$
\begin{aligned}
var(\hat{e}_p) &= \boldsymbol{x}_p' var(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_p + var(u_p) \\
&= \boldsymbol{x}_p' \left[\sigma^2 \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \right] \boldsymbol{x}_p + var(u_p) \\
&\quad \text{substitute } \sigma^2, var(u_p) \text{ with } \hat{\sigma}^2 \text{ (homoskedasticity)} \\
&= \underbrace{\boldsymbol{x}_p' \left[\hat{\sigma}^2 \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \right] \boldsymbol{x}_p}_{\hat{\sigma}_p^2} + \hat{\sigma}^2
\end{aligned}
$$

With growing sample size (asymptotically),
$var(u_p) = \hat{\sigma}_p^2 + \hat{\sigma}^2$ converges to $\hat{\sigma}^2$
$\ldots \; plim\, \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} \quad \Rightarrow \quad plim\, \hat{\sigma}_p^2 = 0$

Predictions - basics (Matrix form)

Variance of $\hat{e}_p$ (continued):

$$var(\hat{e}_p) = \boldsymbol{x}_p' \left[\hat{\sigma}^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right] \boldsymbol{x}_p + \hat{\sigma}^2$$

after re-arranging, $s.e.(\hat{e}_p)$ may be written as

$$s.e.(\hat{e}_p) = \hat{\sigma} \cdot \sqrt{1 + \boldsymbol{x}_p' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_p}\ ,$$

which relates to the individual prediction error.

For mean prediction errors (considering $\hat{\sigma}_p^2$ only):

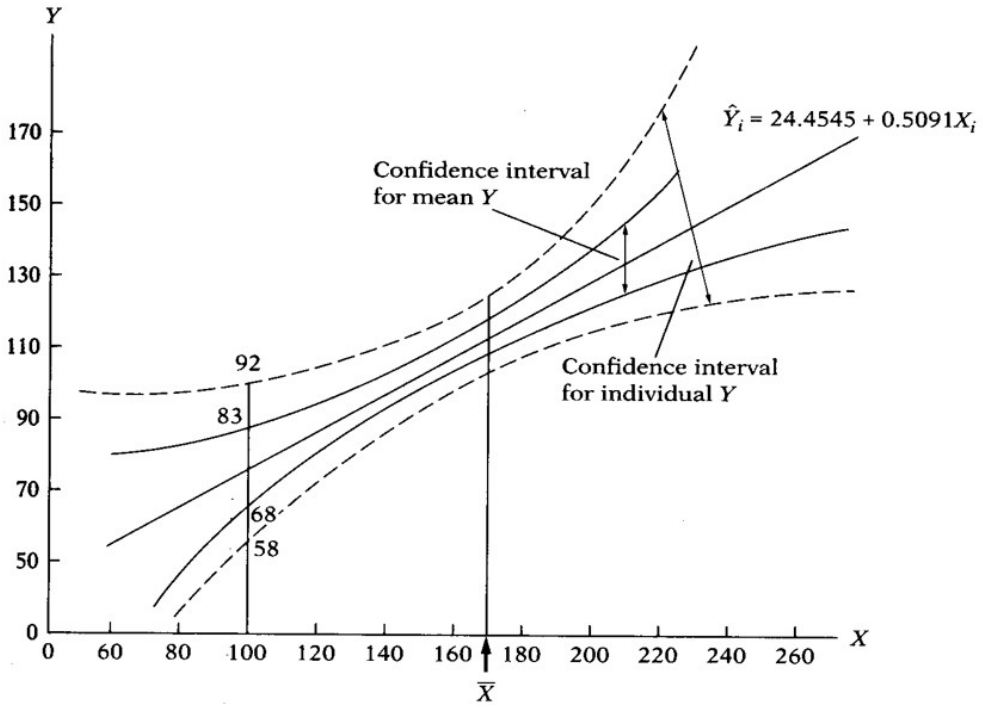
$$s.e.(\widetilde{e}_p) = \hat{\sigma} \cdot \sqrt{\quad \boldsymbol{x}_p' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_p}\ .$$

Predictions - basics (Matrix form)

Prediction intervals: individual vs. mean value predictions:

Individual prediction: $y_p \in \hat{y}_p \pm t^*_{\alpha/2} \times s.e.(\hat{e}_p)$

Mean value: $y_p \in \hat{y}_p \pm t^*_{\alpha/2} \times s.e.(\widetilde{e}_p)$



Predictions - basics

- Why is it difficult to predict individual values?
  - they include random errors
  - we work with estimated parameters
  - parameters can change in time
  - (look at the formula for prediction error.)
- Prediction of expected values
  - parameters are estimated
  - parameters can change in time
- Impacts of random errors on predictions of individual values are usually much bigger than the impacts of (variance in) estimated parameters.

Mean Squared Error of prediction

We can generalize the previous discussion on predictions by considering both biased and unbiased predictors and by allowing for different functional forms and complexity levels in predictive models. Predictions may be compared using:

- $MSE = \mathrm{E}\left[y_i - \hat{f}(\boldsymbol{x}_i)\right]^2$

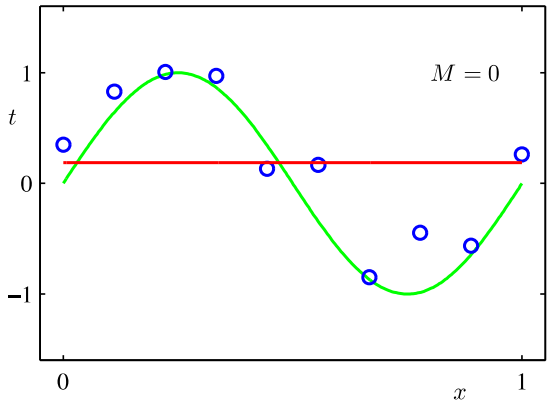
  where $\hat{f}(\boldsymbol{x}_i)$ is the prediction that $\hat{f}$ generates for the i-th regressor set. $\hat{f}$ represents a general class of predictors (linear, non-linear, non-parametric, etc.) and it may produce either biased or unbiased predictions

Variance vs. Bias trade-off

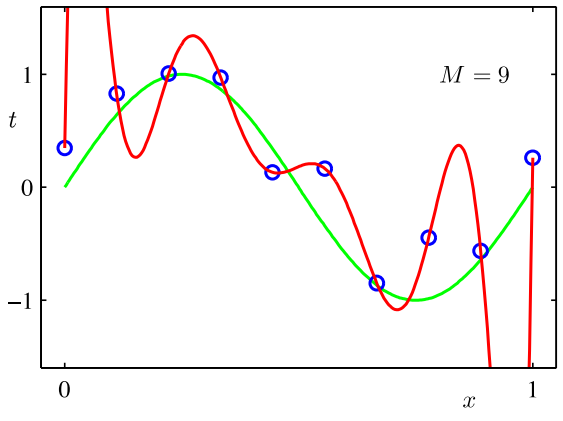
Population equation example: $y = \sin(x) + u$

Bias-Variance tradeoff – Intuition

- Model too simple: does not fit the data well
  - A *biased* solution



- Model too complex: small changes to the data, solution changes a lot
  - A *high-variance* solution



Train sample & Test sample

Suppose we fit a model $\hat{f}(\boldsymbol{x})$ to some training data $\text{Tr} = \{y_i, \boldsymbol{x}_i\}_1^n$ and we wish to see how well it performs.

- We could compute *MSE* over Tr:

$$MSE_{\text{Tr}} = \frac{1}{n} \sum_{i \in \text{Tr}} \left[y_i - \hat{f}(\boldsymbol{x}_i) \right]^2$$

When searching for the "best" model by minimizing *MSE*, the above statistic would lead to over-fit models.

- Instead, we should (if possible) compute the *MSE* using fresh test data $\text{Te} = \{y_i, \boldsymbol{x}_i\}_1^m$:

$$MSE_{\text{Te}} = \frac{1}{m} \sum_{i \in \text{Te}} \left[y_i - \hat{f}(\boldsymbol{x}_i) \right]^2$$

Variance vs. Bias trade-off

Suppose we have a model $\hat{f}(\boldsymbol{x})$, fitted to some training data Tr and let $\{y_0, \boldsymbol{x}_0\}$ be a test observation drawn from the population. If the true model is $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$, with $f(\boldsymbol{x}_i) = \mathrm{E}(y_i|\boldsymbol{x}_i)$, then the **expected test MSE** can be decomposed into:

$$\mathrm{E}(MSE_0) = var(\hat{f}(\boldsymbol{x}_0)) + [Bias\,(\hat{f}(\boldsymbol{x}_0))]^2 + var(\varepsilon_0),$$

where

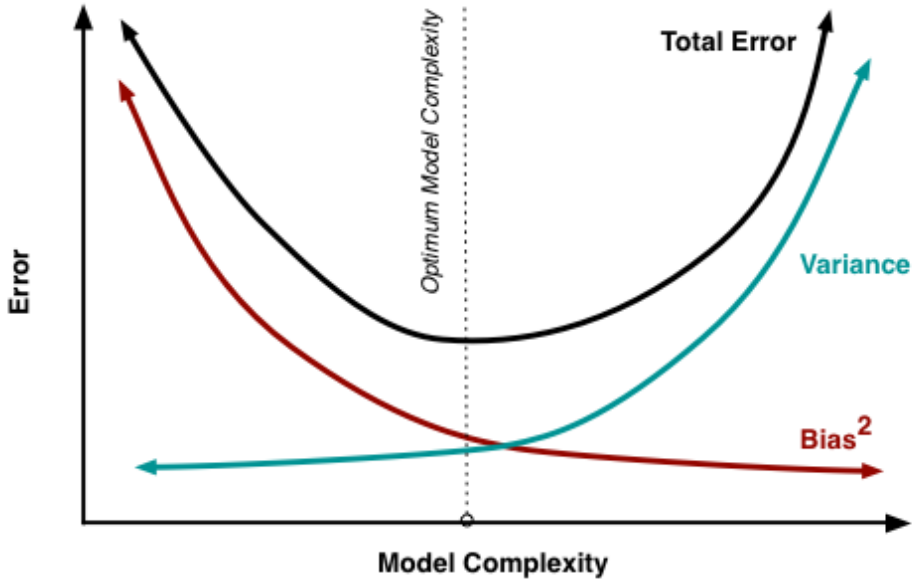
$$Bias\,(\hat{f}(\boldsymbol{x}_0)) = \mathrm{E}[\hat{f}(\boldsymbol{x}_0)] - f(\boldsymbol{x}_0),$$

ε_0 is the irreducible error: $\mathrm{E}(MSE_0) \geq \varepsilon_0$,
all three RHS elements are non-negative,

The above equation refers to the average test MSE that we would obtain if we repeatedly estimated $f(\boldsymbol{x})$ using a large number of training sets and then tested each $\hat{f}(\boldsymbol{x})$ at $\boldsymbol{x}_0$.

Variance vs. Bias trade-off

$$\mathrm{E}(MSE_0) = var(\hat{f}(\boldsymbol{x}_0)) + [Bias\,(\hat{f}(\boldsymbol{x}_0))]^2 + var(\varepsilon_0),$$



This is an illustration, $var(\varepsilon_0)$ not shown explicitly.
(lies at the /asymptotic/ minima of Variance and Bias^2)

k-Fold Cross Validation

- Training error (MSE_{Tr}) can be calculated easily.
- However, MSE_{Tr} is not a good approximation for the MSE_{Te} (out-of sample predictive properties of the model).
- Usually, MSE_{Tr} dramatically underestimates MSE_{Te}.

Cross-validation is based on re-sampling (similar to bootstrap).

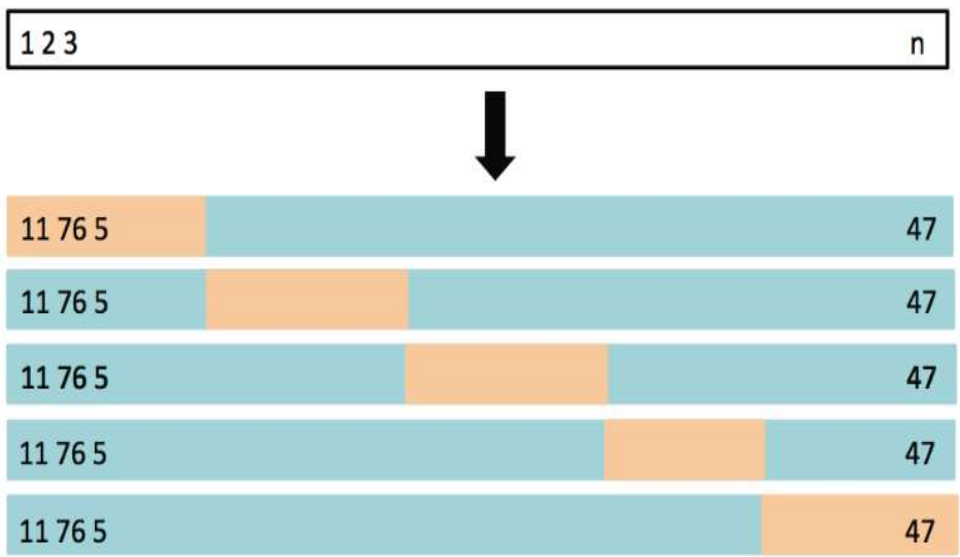
Repeatedly fit a model of interest to samples formed from the training set & make "test sample" predictions, in order to obtain additional information about predictive properties of the model.

k-Fold Cross Validation

- In k-Fold Cross-Validation (kFCV), the original sample is randomly partitioned into k roughly equal subsamples (divisibility).
- Of the k subsamples, a single subsample is retained as the test sample, and the remaining $(k-1)$ subsamples are used as training data.
- The cross-validation process is then repeated k times (the k folds), with each of the k subsamples used exactly once as the test sample.
- The k results from the folds can then be averaged to produce a single estimation.
- $k = 5$ or $k = 10$ is commonly used.

k-Fold Cross Validation

kFCV example for CS data & $k = 5$:
(random sampling, no replacement)



In TS, a similar “Walk forward” test procedure may be applied.

k-Fold Cross Validation

$$CV_{(k)} = \frac{1}{k} \sum_{s=1}^{k} MSE_s,$$

where $CV_{(k)}$ is the k-fold CV estimate,

k is the number of folds used (e.g. 5 or 10),

$MSE_s = \frac{1}{m_s} \sum_{i \in C_s} (y_i - \widehat{y}_i)^2$

m_s and C_s refer to test sample observations for each of the k train sample / test sample steps.

As we evaluate predictions from two or more models, we look for the lowest $CV_{(k)}$.

Chow tests

Focus on DGP stability & model suitability for predictions.

For any LRM: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$

- Chow tests are used to determine whether the regression function differs for different time periods (or respondent groups in CS data).
- Time-stability of the estimated coefficients is a necessary condition for forecasting from an estimated model.
- Chow tests can be defined for cross-sectional units as well. (wages for male/female individuals, etc.)

Chow tests

For any LRM: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$

- Say, the sample (time series) for a period $t = 1, 2, \ldots, T$ may be conveniently divided into two groups: $T_1 + T_2 = T$.
  [consider two periods: fixed vs. floating F/X rates]
  [pre-EU accession vs. post-EU accession period]

- Now, the LRM's vectors and matrices may be partitioned as follows:

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

  where $\boldsymbol{y}_1' = (y_1, \ldots, y_{T_1})$, $\boldsymbol{y}_2' = (y_{T_1+1}, \ldots, y_T)$, etc.
  i.e. $\{\boldsymbol{y}_1, \boldsymbol{X}_1\} \in T_1$, $\{\boldsymbol{y}_2, \boldsymbol{X}_2\} \in T_2$.

Chow tests

For any LRM: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, Chow test can be based on an auxiliary regression (unrestricted model for the F test):

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

  where $\boldsymbol{0}$ is a zero-matrix of the same dimensions as $\boldsymbol{X}_1$, i.e. $(T_1 \times k)$.

Also, we can see that:

- T_1 : $\quad \hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$
- T_2 : $\quad \hat{\boldsymbol{y}} = \boldsymbol{X}(\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\gamma}})$

Note: Power of the test depends on proper T_1 vs. T_2 cutoff. Chow test may be generalized for 3+ time periods (groups).

Chow tests

For our unrestricted model:

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

We can formulate the null of no structural change in model dynamics between the two time periods (groups) as follows:

- $H_0 : \quad \boldsymbol{\gamma} = \boldsymbol{0}$, i.e.: $\gamma_0 = \gamma_1 = \gamma_2 = \cdots = \gamma_k = 0$
- $H_1 : \quad \neg H_0$

This can be tested using an F-test (or its HC version):

- $F = \frac{SSR_r - SSR_{ur}}{SSR_{ur}} \times \frac{n-2k}{k} \underset{H_0}{\sim} F[k, (n-2k)]$

Chow test - Example

A simple Chow test example for CS data:
(to assess whether parameters are equal for M/F students.)

- Original model (Chow test restricted model):
  ... based on the well known Wooldridge dataset.

  $$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u$$

- Auxiliary model (Chow test unrestricted model):

$$\begin{aligned} cumgpa = \beta_0 & + \gamma_0 female \\ & + \beta_1 sat + \gamma_1 (female \times sat) \\ & + \beta_2 hsperc + \gamma_2 (female \times hsperc) \\ & + \beta_3 tothrs + \gamma_3 (female \times tothrs) + u \end{aligned}$$

Chow test - Example (contd.)

- Null hypothesis $H_0 : \gamma_0 = \gamma_1 = \gamma_2 = \gamma_3 = 0$

  If all interactions effects are zero, we have the same regression function for both groups.

- Estimate of the unrestricted model

$$\begin{aligned}\widehat{cumgpa} = & \underset{(.21)}{1.48} - \underset{(.411)}{.353}\,female + \underset{(.0002)}{.0011}\,sat + \underset{(.00039)}{.0075}\,(female \times sat) \\ & - \underset{(.0014)}{.0085}\,hsperc - \underset{(.00316)}{.00055}(female \times hsperc) \\ & + \underset{(.0009)}{.0023}\,tothrs - \underset{(.00163)}{.00012}(female \times tothrs)\end{aligned}$$

. . . t-tests cannot be used to evaluate the joint H_0.

Chow test - Example (contd.)

- F-statistic:

$$F = \frac{(SSR_r - SSR_{ur})/k}{SSR_{ur}/(n-2k)} = \frac{(85.515 - 78.355)/4}{78.355/(366-8)} \approx 8.18$$

  ... using p-value, we reject the null hypothesis

- **Important:** Chow tests (all types) assume constant error variance across groups.

Chow 1: stability test for TS

Here, the F-statistic for the Chow test is calculated in an alternative way (Chow 1):

- For a suitable (potential) "breakpoint", we divide our sample $\{t = 1, 2, \ldots, T\}$ in two groups:
  "T_1" with $\{t = 1, 2, \ldots, T_1\}$ and
  "T_2" with $\{t = T_1+1, T_1+2, \ldots, T\}$
  ... note that the choice of T_1 is arbitrary
  ... (breakpoint-searching algorithms can be used)
- Run separate regressions for both T_1, T_2 groups;
  the SSR_{ur} is given by the sum of the SSRs of the two separately estimated regression models.
  ... sufficient observations in T_1 and T_2 are required (d.f.)
- Run the original (restricted) regression model on the whole sample T and store SSR_r.

Chow 1: stability test for TS

$$F = \frac{SSR_r - SSR_{ur}}{SSR_{ur}} \cdot \frac{T-2k}{k} \underset{H_0}{\sim} F(\, k\,,\, T-2k\,)$$

where
$SSR_{ur} = SSR_{T_1} + SSR_{T_2}$
$SSR_r \;\; = SSR_T$
k is the number of parameters (including intercept) in LRM

H_0: stable structure of coefficients - no statistically significant differences between T_1 and T_2.
H_1: $\neg H_0$ (assume structural change in parameters over time)

Note: Chow 1 can be generalized for G time periods ($G-1$ "breakpoints").
... In such case, $SSR_{ur} = \sum_{g=1}^{G} SSR_g$, d.f. $= T - Gk$
... and we assume $T_g > k$ for all time groups.
... (only usable for small G-values, problematic setup of breakpoints)

Chow 2: prediction test for TS

Sometimes, we do not have enough observations to estimate the LRM separately for T_1 and T_2 as in the Chow 1 test.

In such case, we can use Chow 2: test of prediction unsuitability (slightly different F-statistics).

- The whole period is again divided into two subsets: $T = T_1 + T_2$.
- T_1 is the "base" period (sample size)
- T_2 is the number of "additional" observations, it usually corresponds to an ex-post prediction period

Chow 2: prediction test for TS

$$F = \frac{SSR_r - SSR_{ur}}{SSR_{ur}} \cdot \frac{T_1 - k}{T_2} \underset{H_0}{\sim} F(T_2, T_1 - k)$$

where
$SSR_{ur} = SSR_{T_1}$ (from LRM estimated for "base" period)
$SSR_r = SSR_T$ (from LRM estimated for the whole period)
k is the number of parameters (including intercept) in LRM

H_0: additional (T_2) observations come from the same DGP as in T_1.
H_1: $\neg H_0$ (assume significant differences between samples)
... If H_0 is rejected, we would expect large differences
... between predictions and actual observations of y_t.

If enough T_1 and T_2 observations are available, Chow 1 is preferred (compared to Chow 2) as it has more "power".

Forecasting time series

- **One-step-ahead forecast** f_t
  Forecast error $e_{t+1} = y_{t+1} - f_t$
  Information set: I_t
  Loss function: e_{t+1}^2 or $|e_{t+1}|$
  In forecasting, we minimize $E(e_{t+1}^2|I_t) = E[(y_{t+1} - f_t)^2|I_t]$
  Solution: $E(y_{t+1}|I_t)$

- **Multiple-step-ahead forecast** $f_{t,h}$
  Solution: $E(y_{t+h}|I_t)$

Forecasting time series

For some processes, $E(y_{t+1}|I_t)$ is easy to obtain:

1. **Martingale process (MP):**
   If $E(y_{t+1}|y_t, y_{t-1}, \ldots, y_0) = y_t, \forall t \geq 0$ then $\{y_t\}$ is MP
   $f_t = y_t$
   If a process $\{y_t\}$ is a martingale then $\{\Delta y_t\}$ is martingale difference sequence (MDS)

$$E(\Delta y_{t+1}|y_t, y_{t-1}, \ldots, y_0) = 0$$

2. **Process with exponential smoothing:**

$$E(y_{t+1}|I_t) = \alpha y_t + \alpha(1-\alpha)y_{t-1} + \cdots + \alpha(1-\alpha)^t y_0; \;\; 0 < \alpha < 1.$$

Set $f_0 = y_0$, then for $t \geq 1$: $f_t = \alpha y_t + (1-\alpha) f_{t-1}$

Forecasting time series

3 Regression models

- Static model: $y_t = \beta_0 + \beta_1 x_t + u_t$
  $E(y_{t+1}|I_t) = \beta_0 + \beta_1 x_{t+1} \rightarrow$ Conditional forecasting
  I_t contains $x_{t+1}, y_t, x_t, \ldots, y_1, x_1$
  Here, knowledge of x_{t+1} is assumed (forecast condition).

  $E(y_{t+1}|I_t) = \beta_0 + \beta_1 E(x_{t+1}|I_t) \rightarrow$ Unconditional forecasting
  I_t contains $y_t, x_t, \ldots, y_1, x_1$
  Here, x_{t+1} needs to be estimated before y_{t+1}

- Dynamic models depending on lagged variables only:
  $y_t = \delta_0 + \alpha_1 y_{t-1} + \gamma_1 x_{t-1} + u_t$
  $E(u_t|I_{t-1}) = 0$
  $E(y_{t+1}|I_t) = \delta_0 + \alpha_1 y_{t-1} + \gamma_1 x_{t-1}$
  We can use more lags, drop x or add more variables

Forecasting time series

One-Step-Ahead Forecasting with

$y_t = \delta_0 + \alpha_1 y_{t-1} + \gamma_1 x_{t-1} + u_t$:

point forecast: $\hat{f}_t = \hat{\delta}_0 + \hat{\alpha}_1 y_t + \hat{\gamma}_1 x_t$

forecast error: $\hat{e}_{t+1} = y_{t+1} - \hat{f}_t$

s.e. of forecast: $s.e.(\hat{e}_{t+1}) = \{[s.e.(\hat{f}_t)]^2 + \hat{\sigma}^2\}^{1/2}$

forecast interval: essentially the same as prediction interval

approximate 95% forecast interval is: $\hat{f}_t \pm 1.96 \times s.e.(\hat{e}_{t+1})$

Forecasting time series

Example: File PHILLIPS

Forecasting US unemployment rate

$$\widehat{unem}_t = \underset{(.577)}{1.572} + \underset{(.097)}{.732}\, unem_{t-1}$$

$$n = 48, \overline{R}^2 = .544$$

$$\widehat{unem}_t = \underset{(.490)}{1.304} + \underset{(.084)}{.647}\, unem_{t-1} + \underset{(.041)}{.184}\, inf_{t-1}$$

$$n = 48, \overline{R}^2 = .677$$

Note that these regressions are not meant as causal equations. The hope is that the linear regressions approximate well the conditional expectation.

Forecasting time series

Evaluating forecast quality

- We can measure how forecasted values fit to actual observations (in-sample criteria, e.g. R^2)
- It is better, however, to evaluate the forecasting performance when forecasting out-of-sample values (out-of-sample criteria). For this purpose, use first n observations for estimation, and the remaining m observations to calculate the forecast errors $\hat{e}_{n+h}$
- Forecast evaluation measures:

  Mean Absolute Error $MAE = m^{-1} \sum_{h=1}^{m} |\hat{e}_{n+h}|$,

  Root Mean Squared Error $RMSE = (m^{-1} \sum_{h=1}^{m} \hat{e}_{n+h}^2)^{1/2}$

  k-Fold Cross-Validation (kFCV) approach

Forecasting time series

- **Additional comments**
  - Multiple-step-ahead forecasts are possible, but necessarily less precise.
  - Forecasts may make use of deterministic trends, but the error made by extrapolating time trends too far into the future may be large.
  - Similarly, seasonal patterns may be incorporated into forecasts.
  - It is possible to calculate confidence intervals for the point multiple-step-ahead forecasts.
  - Forecasting $I(1)$ time series can be based on adding predicted changes (which are $I(0)$) to base levels.
  - Forecast intervals for $I(0)$ series converge to the unconditional variance, whereas for integrated series, they are unbounded.