



AALBORG UNIVERSITET

Tidsrækkeanalyse – Slides 9

Ege Rubak

(baseret på Slides af Esben Høg)

Institut for Matematiske Fag
Aalborg Universitet

Laggede regressionsmodeller og transfermodeller

Betragt en lagged regressionsmodel på formen

$$y_t = \sum_{h=-\infty}^{\infty} \beta_h x_{t-h} + v_t,$$

hvor x_t er en observeret **input tidsrække**, y_t er en observeret **output tidsrække** og v_t er en stationær støj-proces.

En sådan model er nyttig i forbindelse med

- Identifikation af den (bedste lineære) relation mellem de to tidsrækker.
- Forecast af en tidsrække ud fra en anden tidsrække (vi vil typisk kræve, at $\beta_h = 0$ for $h < 0$).
- I *SOI and recruitment* eksemplet i [ShuSt] kan det være man ønsker at identificere, hvordan værdier af recruitment series (antal nye fisk) er relateret til *Southern Oscillation Index*.
- Eller vi ønsker måske at forecaste fremtidige værdier af af recruitment fra SOI.

Laggede regressionsmodeller (fortsat)

- Multiple, simultant stationære, tidsrækker i tidsdomænet: Et vigtigt redskab her er **krydskovariansfunktionen** (eller krydskorrelationsfunktionen CCF).
- Lagged regression i tidsdomænet: Modellér inputtidsrækken, uddrag den “hvide” tidsrække der driver den (“*prewhitening*”), regressér med den transformerede output tidsrække.

Krydskovarianser

- Husk at autokovariansfunktionen af en stationær tidsrække $\{x_t\}_{t \in \mathbb{Z}}$ er

$$\gamma_x(h) = E[(x_{t+h} - \mu_x)(x_t - \mu_x)].$$

- **Krydskovariansfunktionen** mellem to simultant stationære processer, $\{x_t\}_{t \in \mathbb{Z}}$ og $\{y_t\}_{t \in \mathbb{Z}}$, er

$$\gamma_{xy}(h) = E[(x_{t+h} - \mu_x)(y_t - \mu_y)].$$

(simultant stationære vil sige, at de begge har konstante middelværdier, autokovarianser, der kun afhænger af lag h , og krydskovarianser der kun afhænger af lag h).

Krydskorrelationen

- Krydskorrelationsfunktionen CCF mellem to simultant stationære processer $\{x_t\}_{t \in \mathbb{Z}}$ og $\{y_t\}_{t \in \mathbb{Z}}$ er

$$\rho_{xy}(h) = \frac{\gamma_{xy}(h)}{\sqrt{\gamma_x(0)\gamma_y(0)}}.$$

Bemærk at $\rho_{xy}(h) = \rho_{yx}(-h)$, men $\rho_{xy}(h) \neq \rho_{xy}(-h)$.

- Eksempel: Antag, at $y_t = \beta x_{t-\ell} + w_t$, for $\{x_t\}_{t \in \mathbb{Z}}$ stationær og ukorreleret med $\{w_t\}_{t \in \mathbb{Z}}$, og w_t hvid støj. Så er $\{x_t\}_{t \in \mathbb{Z}}$ og $\{y_t\}_{t \in \mathbb{Z}}$ simultant stationære med $\mu_y = \beta\mu_x$, og

$$\gamma_{xy}(h) = \beta\gamma_x(h + \ell).$$

- Hvis $\ell > 0$, siger man på engelsk, at x_t *leads* y_t .
- Hvis $\ell < 0$, siger man på engelsk, at x_t *lags* y_t .

Sample krydskovariansen og sample CCF

$$\hat{\gamma}_{xy}(h) = \frac{1}{n} \sum_{j=1}^{n-h} (x_{t+h} - \bar{x})(y_t - \bar{y}),$$

for $h \geq 0$ (og $\hat{\gamma}_{xy}(h) = \hat{\gamma}_{yx}(-h)$ for $h < 0$).

Sample CCF er

$$\hat{\rho}_{xy}(h) = \frac{\hat{\gamma}_{xy}(h)}{\sqrt{\hat{\gamma}_x(0)\hat{\gamma}_y(0)}}.$$

Sample krydskovariansen og sample CCF (fortsat)

- Hvis en af $\{x_t\}_{t \in \mathbb{Z}}$ eller $\{y_t\}_{t \in \mathbb{Z}}$ er hvid støj, så vil

$$\hat{\rho}_{xy}(h) \xrightarrow{F} N(0, 1/\sqrt{n}).$$

- Man kan kigge efter *peaks* i sample CCF for at identificere en *lead* eller *lag* relation. (husk at ACF af inputtidsrækken *peaker* ved $h = 0$)
- Eksempel: CCF af SOI og recruitment (Figur 1.16 side 31 i [ShuSt]) har et *peak* ved $h = -6$, som indikerer at recruitment til tid t har den stærkeste korrelation med SOI til tid $t - 6$.
Altså, SOI *leads* recruitment (på 6 måneder).

Lagget regression i tidsdomænet

- Antag vi ønsker at tilpasse en lagged regressionsmodel af formen

$$y_t = \alpha(B)x_t + \eta_t = \sum_{j=1}^{\infty} \alpha_j x_{t-j} + \eta_t,$$

hvor x_t er en observeret input tidsrække, y_t er en observeret output tidsrække, og η_t er en stationær støjproces, ukorreleret med $\{x_t\}_{t \in \mathbb{Z}}$.

- En tilgang, som stammer fra Box & Jenkins, er at tilpasse ARIMA modeller for x_t og η_t , og så finde en simpel rational repræsentation for $\alpha(B)$.

Lagget regression i tidsdomænet

$$y_t = \alpha(B)x_t + \eta_t = \sum_{j=1}^{\infty} \alpha_j x_{t-j} + \eta_t.$$

For eksempel:

$$x_t = \frac{\theta_x(B)}{\phi_x(B)} w_t,$$

$$\eta_t = \frac{\theta_\eta(B)}{\phi_\eta(B)} z_t,$$

$$\alpha(B) = \frac{\delta(B)}{\omega(B)} B^d.$$

Hvor de to hvide støjprocesser er antaget uafhængige.
Bemærk, at forsinkelsen B^d betyder at y_t lags x_t med d tidsenheder.

Lagget regression i tidsdomænet

Hvordan vælger vi alle disse parametre?

- ❶ Tilpas $\theta_x(B)$, $\phi_x(B)$ for at modellere inputtidsrækken x_t .
- ❷ *Prewhiten* inputtidsrækken ved at anvende den inverse operator $\phi_x(B)/\theta_x(B)$:

$$\tilde{y}_t = \frac{\phi_x(B)}{\theta_x(B)} y_t = \alpha(B) w_t + \frac{\phi_x(B)}{\theta_x(B)} \eta_t.$$

- ❸ Beregn krydskorrelationerne mellem \tilde{y}_t med w_t ,

$$\gamma_{\tilde{y},w}(h) = E \left(\sum_{j=0}^{\infty} \alpha_j w_{t+h-j} w_t \right) = \sigma_w^2 \alpha_h,$$

for at få en indikation af opførslen af $\alpha(B)$, f.eks. hvilken forsinkelse der måtte være.

- ❹ Estimér koefficienterne af $\alpha(B)$ og tilpas derved en ARMA model for støjserien η_t .

Lagget regression i tidsdomænet

- Hvorfor *prewhitene*?
- *Prewhitening* trinnet inverterer det lineære filter $x_t = \theta_x(B)/\phi_x(B)w_t$. Så er den laggede regression herefter mellem den transformerede y_t og en hvid støj w_t . Det gør det nemmere at bestemme passende lags.
- For eksempel i SOI/recruitment tidsrækkerne, behandler vi SOI som input, estimerer en AR(1) model, *prewhitener* denne (dvs. beregner den inverse af vores AR(1) operator, og anvender det på SOI tidsrækken), og betragter dernæst krydskorrelationerne mellem den transformerede recruitment tidsrække og den *prewhitenede* SOI.
- Dette viser en stor *peak* ved lag -5 (svarende til at SOI tidsrækken *leader* recruitment tidsrækken. Eksemplerne 5.8 og 5.9 i [ShuSt] betragter dernæst $\alpha(B) = B^5/(1 - \omega_1 B)$.

Lagget regression i tidsdomænet

- Denne sekventielle estimationsprocedure,

ϕ_x, θ_x dernæst α

derneæst ϕ_η, θ_η

er selvfølgelig noget *ad hoc*, men det har vist sig at være en udmærket metode.

- State space modeller giver en alternativ metode. De er også velegnede til vektor-tidsrækker, både input og output.

Ikke-stationære tidsrækker og spuriøs regression

Motivation: Regression med ikke-stationaritet.

- Hvad sker der med egenskaberne for OLS, hvis variablene er ikke-stationære?
- Betragt **to tilsyneladende ikke-relaterede variable**:
 - CONS: Dansk privat forbrug i 1995 priser.
 - BIRD: Antal ynglende skarver (en fugleart) i Danmark.
- Og betragt en statisk regressionsmodel

$$\log(\text{CONS}_t) = \beta_0 + \beta_1 \log(\text{BIRD}_t) + w_t.$$

Vi ville forvente (eller håbe), at $\hat{\beta}_1 \approx 0$ og $R^2 \approx 0$.

- Anvendes OLS på årlige data 1982-2001 fås følgende resultat

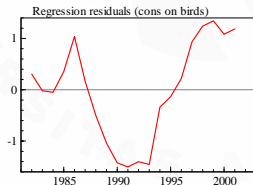
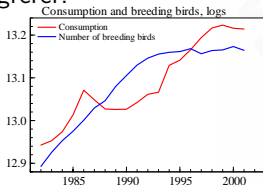
$$\log(\text{CONS}_t) = \underset{(80.90)}{12.145} + \underset{(6.30)}{0.095} \log(\text{BIRD}_t) + \hat{w}_t,$$

med $R^2 = 0.688$, og hvor tallene i parentes er t -teststørrelser.

- **Det ligner en fornuftig model. Men det er komplet nonsens: Spuriøs regression.**

Spuriøs regression kontra kointegration

- Variablene er ikke-stationære.
 - Residualerne \hat{w}_t er ikke-stationære, og standardresultater for OLS holder ikke.
 - Generelt: Regressionsmodeller for ikke-stationære variable giver spuriøse resultater. Den eneste undtagelse er, hvis modellen eliminerer den stokastiske trend i residualerne (altså eliminerer ikke-stationariteten) og producerer stationære residualer:
- Kointegration.**
- **For ikke-stationære variable skal man således tænke i kointegration.** Det er kun relevant at bruge regressionsoutputtet, hvis variablene kointegrerer.



```
# RCode 4.1 Spurious regression
library(lmtest)
set.seed(123456)
e1 <- rnorm(500)
e2 <- rnorm(500)
trd <- 1:500
y1 <- 0.8*trd + cumsum(e1)
y2 <- 0.6*trd + cumsum(e2)
sr.reg <- lm(y1 ~ y2)
sr.dw <- dwtest(sr.reg)$statistic
```

Definition af kointegration

- Lad $x_t = (x_{1t} \ x_{2t})^\top$ være to $I(1)$ tidsrækker, dvs. de indeholder stokastiske trends:

$$x_{1t} = \sum_{i=1}^t w_{1i} + \text{startværdier} + \text{stationær proces}$$

$$x_{2t} = \sum_{i=1}^t w_{2i} + \text{startværdier} + \text{stationær proces.}$$

- Generelt vil en lineær kombination af x_{1t} og x_{2t} også indeholde en random walk. Definér $\beta = (1 \ -\beta_2)^\top$ og betragt linearkombinationen:

$$\begin{aligned} z_t &= \beta^\top x_t = (1 \ -\beta_2) \begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix} = x_{1t} - \beta_2 x_{2t} \\ &= \sum_{i=1}^t w_{1i} - \beta_2 \sum_{i=1}^t w_{2i} + \text{startværdier} + \text{stationær proces.} \end{aligned}$$

Kointegration

- Vigtig undtagelse: Hvis der eksisterer en β , sådan at z_t er stationær. Vi siger, at x_{1t} og x_{2t} **kointegrerer** med **kointegrationsvektor** β .
- Kointegration opstår, hvis de stokastiske trends i x_{1t} og x_{2t} er de samme, så de udligner hinanden, $\sum_{i=1}^t w_{1i} = \beta_2 \sum_{i=1}^t w_{2i}$. Dette kaldes en **common trend**.
- Man kan tænke på en ligning der eliminerer de random walks x_{1t} og x_{2t} :

$$x_{1t} = \mu + \beta_2 x_{2t} + w_t.$$

Hvis w_t er $I(0)$, så er $\beta = (1 - \beta_2)^T$ en kointegrerende vektor.

Kointegration

- Den kointegrerende vektor er kun entydig (unik) op til en konstant faktor. Hvis $\beta^\top x_t \sim I(0)$, så gælder det også for $c\beta^\top x_t$, for $c \neq 0$. Vi kan derfor vælge en normalisering

$$\beta = \begin{pmatrix} 1 \\ -\beta_2 \end{pmatrix}.$$

- Kointegration kan let udvides til flere tidsrækker (eller flere variable): Variablene i $x_t = (x_{1t}, \dots, x_{pt})^\top$ kointegrerer, hvis

$$z_t = \beta^\top x_t = x_{1t} - \beta_2 x_{2t} - \dots - \beta_p x_{pt} \sim I(0).$$

Kointegration og økonomisk ligevægt

- Betragt en regressionsmodel for to $I(1)$ variable x_{1t} og x_{2t} givet ved

$$x_{1t} = \mu + \beta_2 x_{2t} + w_t. \quad (1)$$

- Hvis x_{1t} og x_{2t} kointegrerer med kointegrationsvektor $(1, -\beta_2)$, så er afvigelsen

$$w_t = x_{1t} - \mu - \beta_2 x_{2t}$$

en stationær proces med middelværdi nul. x_{1t} og x_{2t} ko-varierer og $w_t \sim I(0)$. Man kan tænke på (1) som definerende en **ligevægt** mellem x_{1t} og x_{2t} .

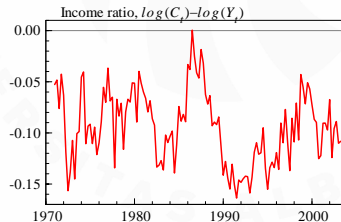
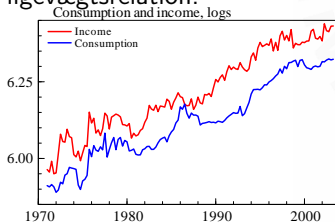
- Hvis x_{1t} og x_{2t} ikke kointegrerer, så er afvigelsen w_t en $I(1)$ proces. Og der er ingen naturlig fortolkning af (1) som en ligevægtsrelation.

Empirisk eksempel: Forbrug og indkomst

- Tidsrækker for log forbrug, C_t , og log indkomst, Y_t , er typisk $I(1)$. Definer en vektor $x_t = (C_t \ Y_t)^\top$.
- Forbrug og indkomst er kointegrerede med kointegrationsvektor $\beta = (1 \ -1)^\top$, hvis (log-) forbrug-indkomst ratioen

$$z_t = \beta^\top x_t = (1 \ -1) \begin{pmatrix} C_t \\ Y_t \end{pmatrix} = C_t - Y_t,$$

er en stationær proces. Forbrugs-indkomst ratioen er en ligevægtsrelation.



Hvordan bliver ligevægten vedvarende?

- Der må være en slags kræfter der trækker x_{1t} eller x_{2t} mod ligevægten.
- En berømt repræsentationssætning (Clive Granger, 1983): x_{1t} og x_{2t} kointegrerer hvis og kun hvis der eksisterer en fejlkorrektionsmodel for enten x_{1t} , x_{2t} eller begge.
- Et eksempel: Lad $z_t = x_{1t} - \beta_2 x_{2t}$ være en stationær relation mellem $I(1)$ tidsrækker. Så eksisterer der en stationær ARMA model for z_t . Antag for simpelhed skyld en AR(2):

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + w_t, \quad \phi(1) = 1 - \phi_1 - \phi_2 > 0.$$

Hvordan bliver ligevægten vedvarende? (fortsat)

Ovenstående er ækvivalent med

$$x_{1t} - \beta_2 x_{2t} = \phi_1 (x_{1,t-1} - \beta_2 x_{2,t-1}) + \phi_2 (x_{1,t-2} - \beta_2 x_{2,t-2}) + w_t$$

$$x_{1t} = \beta_2 x_{2t} + \phi_1 x_{1,t-1} - \phi_1 \beta_2 x_{2,t-1} + \phi_2 x_{1,t-2} - \phi_2 \beta_2 x_{2,t-2} + w_t.$$

eller

$$\nabla x_{1t} = \beta_2 \nabla x_{2t} + \phi_2 \beta_2 \nabla x_{2,t-1} - \phi_2 \nabla x_{1,t-1} - (1 - \phi_1 - \phi_2) \{x_{1,t-1} - \beta_2 x_{2,t-1}\} + w_t.$$

Dette er **fejlkorrektionsmodellen** i en nøddeskal.

Mere om kointegration

Kointegration er en systemegenskab. Begge tidsrækker kan fejlkorrigere, f.eks.

$$\nabla x_{1t} = \delta_1 + \Gamma_{11}\nabla x_{1,t-1} + \Gamma_{12}\nabla x_{2,t-1} + \alpha_1(x_{1,t-1} - \beta_2 x_{2,t-1}) + w_{1t},$$

$$\nabla x_{2t} = \delta_2 + \Gamma_{21}\nabla x_{1,t-1} + \Gamma_{22}\nabla x_{2,t-1} + \alpha_2(x_{1,t-1} - \beta_2 x_{2,t-1}) + w_{2t}.$$

Jvf. også Pfaffs notation side 77, formel (4.5a) og (4.5b).

Mere om kointegration

Modellen kan skrives som en såkaldt **vektor fejlkorrigeringsmodel**,

$$\begin{pmatrix} \nabla x_{1t} \\ \nabla x_{2t} \end{pmatrix} = \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} + \begin{pmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{pmatrix} \begin{pmatrix} \nabla x_{1,t-1} \\ \nabla x_{2,t-1} \end{pmatrix} + \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} (x_{1,t-1} - \beta_2 x_{2,t-1}) \\ + \begin{pmatrix} w_{1t} \\ w_{2t} \end{pmatrix},$$

eller simpelthen

$$\nabla x_t = \delta + \Gamma \nabla x_{t-1} + \alpha \beta^\top x_{t-1} + w_t.$$

- Bemærk, at $\beta^\top x_{t-1} = x_{1,t-1} - \beta_2 x_{2,t-1}$ optræder i begge ligninger.
- For at x_{1t} fejlkorrigerer, så skal $\alpha_1 < 0$, og for at x_{2t} fejlkorrigerer, så skal $\alpha_2 > 0$.

OLS regression med kointegrerede tidsrækker

- I kointegrationstilfældet eksisterer der en β_2 sådan at fejlleddet w_t i

$$x_{1t} = \mu + \beta_2 x_{2t} + w_t \quad (2)$$

er stationær.

- OLS anvendt på (2) giver **konsistente** resultater, sådan at $\hat{\beta}_2 \rightarrow \beta_2$ for $n \rightarrow \infty$.
- Beklageligvis så viser det sig, at $\hat{\beta}_2$ ikke generelt er asymptotisk normalfordelt, så vi kan bruge (2) til estimation, ikke til test.

Test for kointegration: Engle-Grangers 2-step metode (Også beskrevet i [Pfaff] side 76-78)

- Step 1
- Tjek at alle individuelle tidsrækker er $I(1)$. For **simpelheds skyld antager** vi her, at der kun er **to tidsrækker** i systemet, x_{1t} og x_{2t} . I [Pfaff] beskriver han det med flere tidsrækker.
 - Estimér så en regression med én af tidsrækkerne som respons og den anden som forklarende variabel, f.eks.

$$x_{1t} = \mu + \beta_2 x_{2t} + w_t. \quad (3)$$

Gem residualerne i tidsrækken \hat{w}_t .

- Test vha. Dickey-Fuller test om disse residualer \hat{w}_t er $I(1)$ eller $I(0)$. Hvis \hat{w}_t er $I(0)$ så **kointegrerer** de oprindelige tidsrækker, og gå så til trin 2 for at estimere en **fejlkorrektionsmodel (ECM)**.
- Hvis \hat{w}_t er $I(1)$, så estimér en model der kun indeholder 1. differenser af de oprindelige tidsrækker, altså en model der **kun estimerer kortsigts-sammenhænge**. En ECM er så ikke relevant.

Test for kointegration: Engle-Grangers 2-step metode (fortsat)

- Step 2
- Brug step 1 residualerne fra forrige periode som en variabel i **fejlkorrektionsmodellen**:

$$\nabla x_{1t} = \psi_0 + \psi_1 \nabla x_{2t} + \lambda \hat{w}_{t-1} + v_t, \quad (4)$$

hvor $\hat{w}_{t-1} = x_{1,t-1} - \hat{\mu} - \hat{\beta}_2 x_{2,t-1}$.

- I formel (4) indgår både en **kortsigts-sammenhæng** og en **langsigts-sammenhæng**.
- Værdien af koefficienten λ bestemmer hastigheden af tilpasningen (fejlkorrektionen), og den skal altid være negativ. Ellers vil systemet divergere fra dets langsigts ligevægt.
- Bemærk i øvrigt, at hvis \hat{w}_{t-1} i formel (4) erstattes med w_{t-1} , så bliver $\psi_0 = 0$, $\psi_1 = \beta_2$, $\lambda = -1$ og $v_t = w_t$, idet (4) så simpelthen bliver

$$\nabla x_{1t} = \beta_2 \nabla x_{2t} - (x_{1,t-1} - \mu - \beta_2 x_{2,t-1}) + w_t.$$

For en implementering af Engle-Grangers metode, se R kode
4.2 side 77 i [Pfaff]

Test for kointegration: Engle-Grangers 2-step metode (fortsat)

■ Bemærk:

- ➊ Residualerne \hat{w}_t har middelværdi nul. Ingen deterministiske led i DF regressionen.
- ➋ Kritiske værdier for $t_{\gamma=0}$ afhænger dog stadig af eventuelle deterministiske regressorer i (3) (f.eks. lineær trend).
- ➌ Det faktum, at $\hat{\beta}_2$ er estimeret ændrer også på de kritiske værdier. OLS minimerer variansen af \hat{w}_t . Skal se så “stationær ud som muligt”, kritiske værdier afhænger af antal regressorer.
- ➍ De kritiske værdier der skal bruges ved DF test for om residualerne er $I(1)$ [altså har unit root] er givet i tabellen på næste slide.

Kritiske værdier for Engle-Granger test for ingen kointegration

Antal variable i ligningen	Sample size	Kritisk værdi		
		1%	5%	10%
2	50	-4.32	-3.67	-3.28
	100	-4.07	-3.37	-3.03
	200	-4.00	-3.37	-3.02
3	50	-4.84	-4.11	-3.73
	100	-4.45	-3.93	-3.59
	200	-4.35	-3.78	-3.47
4	50	-4.94	-4.35	-4.02
	100	-4.75	-4.22	-3.89
	200	-4.70	-4.18	-3.89
5	50	-5.41	-4.76	-4.42
	100	-5.18	-4.58	-4.26
	200	-5.02	-4.48	-4.18

Kilde: Engle & Yoo (1987): Forecasting and testing in cointegrated systems,
Journal of Econometrics 35, side 143-159.

```
#RCode4.2 Engle-Granger procedure with generated data
set.seed(123456)
e1 <- rnorm(100)
e2 <- rnorm(100)
y1 <- cumsum(e1)
y2 <- 0.6*y1+e2
lr.reg <- lm(y2 ~ y1)
error <- residuals(lr.reg)
error.lagged <- error[-c(99,100)]
dy1 <- diff(y1)
dy2 <- diff(y2)
diff.dat <- data.frame(embed(cbind(dy1,dy2),2))
colnames(diff.dat) <- c('dy1','dy2','dy1.1','dy2.1')
ecm.reg <- lm(dy2 ~ error.lagged + dy1.1 + dy2.1,
              data = diff.dat)
```

Opsummering: Engle-Granger analyse

- ❶ Test individuelle variable, f.eks. x_{1t} og x_{2t} for *unit roots*.
- ❷ Kør en statisk kointegrationsregression

$$x_{1t} = \mu + \beta_2 x_{2t} + w_t.$$

Bemærk, at t -teststørrelser her kan ikke bruges til inferens.

- ❸ Test for ingen kointegration ved at teste for en *unit root* i residualerne \hat{w}_t .
- ❹ Hvis kointegration ikke forkastes estimeres en dynamisk (ECM) model som

$$\nabla x_{1t} = \psi_0 + \psi_1 \nabla x_{2t} + \lambda \hat{w}_{t-1} + v_t.$$

Alle led er stationære. Inferens i den model er standard.