

## F8 Avvikelser från modellantagandena I

### Heteroskedasticitet

Enligt antagandena för linjär regression ska slumptermerna  $\varepsilon_i$  vara iid (oberoende, lika fördelade) enligt

$$\varepsilon_i \sim N(0, \sigma^2)$$

dvs variansen för slump termen är konstant  $\sigma^2$  oavsett var eller när vi tittar på  $Y_i$  och  $X_i$ , mao oavsett  $i$ . (Tidigare skrev vi  $\sigma_\varepsilon^2$  men vi kan släppa indexet tillfälligtvis.)

Men detta är inte alltid fallet när man tittar på verkliga datamaterial. Istället ser man något som skulle kunna skrivas som

$$\varepsilon_i \sim N(0, \sigma_i^2)$$

dvs slumptermsvariansen är inte längre nödvändigtvis konstant för alla  $i$

$$\sigma_i^2 \neq \sigma_j^2 \quad \text{för } i \neq j$$

Vi har heteroskedasticitet i slumptermerna.

### Orsaker till heteroskedasticitet

1. Error-learning models: ju mer man tränar desto färre fel  $\Rightarrow$  mindre variation
2. Discretionary income: högre inkomst  $\Rightarrow$  större variation i hur man disponerar sin inkomst, tex sparande
3. Datainsamlingsmetoder förbättras  $\Rightarrow$  mindre variation
4. Outliers: beroende på hur många som avviker och var de ligger (outlier i  $Y$  eller outlier i  $X$  eller båda) kan man ofta om materialet är litet få indikationer på varierande varians.
5. Felaktig modellspecifikation: saknas en väsentlig term? Tex linjär regression när den "sanna" modellen är ett polynom.

6. Skevhet (*skewness*) i regressorerna: tex inkomst, förmögenhet.

#### 7. Olämpliga

- (a) transformationer: tex kvoter, differenser
- (b) funktionella former: tex linjärt kontra log-linjärt

Vanligare med heteroskedasticitet i tvärsnittsstudie än i en i tidseriedata.

- Tvärsnittsstudie: en observation från många objekt (vid samma tidpunkt, oftast). Objekten tenderar att variera

- Tidseriedata: flera observationer från ett objekt (flera tidpunkter). Objektet tenderar att vara stabil (i någon mening)

- Paneldata, longitudinella studier: flera observationer från många objekt.

Vad händer med OLS-skattningarna om data är heteroskedastiskt?

Utgå från den enkla modellen

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Tex, skattningen för  $\beta_1$  (se tidigare OH) och variansen för denna var under antagandet om lika varians (dvs  $\sigma_i^2 = \sigma^2$  för alla  $i$ ) var

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

Men om det är heteroskedastiskt och  $\sigma_i^2 \neq \sigma_j^2$  för  $i \neq j$ , blir variansen för  $\hat{\beta}_1$  istället

$$Var(\hat{\beta}_1) = \frac{\sum (x_i - \bar{x})^2 \sigma_i^2}{\left(\sum (x_i - \bar{x})^2\right)^2} \quad (11.2.2)$$

Vad spelar detta för roll egentligen?

Skattningarna är fortfarande väntevärdesriktiga (unbiased), konsistenta, asymptotiskt normalfördelade.

Men de är inte längre effektiva.

## GLS - Generaliserad Minsta Kvadratskattning

Idé: ge observationer från områden med stor variation låg vikt och de med liten variation stor vikt. För de senare är mer samlade kring sina medelvärden och man kan få en bättre precision och därmed skatta regressionslinjen "bättre".

Vi skriver

$$Y_i = \beta_0 X_{0i} + \beta_1 X_i + \varepsilon_i$$

där  $X_{0i} = 1$  för alla  $i$ . Antaga att  $\sigma_i^2$  är känd för alla  $i$  och skriv

$$\frac{Y_i}{\sigma_i} = \beta_0 \left( \frac{X_{0i}}{\sigma_i} \right) + \beta_1 \left( \frac{X_i}{\sigma_i} \right) + \left( \frac{\varepsilon_i}{\sigma_i} \right)$$

eller

$$Y_i^* = \beta_0 X_{0i}^* + \beta_1 X_i^* + \varepsilon_i^*$$

och observera att

$$E(\varepsilon_i^*) = E\left(\frac{\varepsilon_i}{\sigma_i}\right) = \frac{1}{\sigma_i} E(\varepsilon_i) = 0$$

och

$$\begin{aligned} Var(\varepsilon_i^*) &= E((\varepsilon_i^*)^2) = E\left(\left(\frac{\varepsilon_i}{\sigma_i}\right)^2\right) \\ &= \frac{1}{\sigma_i^2} E(\varepsilon_i^2) = \frac{1}{\sigma_i^2} \sigma_i^2 = 1 \end{aligned}$$

vilket är konstant lika med 1 för all  $i$ , dvs vi har transformerat  $\varepsilon_i$  till  $\varepsilon_i^*$  som är homoskedastiskt!

Genom att transformera enligt

$$(1, X_i, Y_i) \longrightarrow \left( \frac{1}{\sigma_i}, \frac{X_i}{\sigma_i}, \frac{Y_i}{\sigma_i} \right) = (X_{0i}^*, X_i^*, Y_i^*)$$

och sedan köra på med vanlig OLS (MK) så kommer man runt problemet.

Vanlig OLS: minimera

$$Q = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^n e_i^2$$

GLS: minimera

$$Q^* = \sum_{i=1}^n w_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^n w_i e_i^2$$

där

$$w_i = \frac{1}{\sigma_i}$$

Beteckningen  $w_i$  kommer av *weights* eller vikter, dvs varje observation i datamaterialet viktas med inversen till "sin" slumptermsvarians. Därav den vanliga benämningen *Weighted Least Squares* eller WLS.

Vad händer om man "gör fel"?

OLS skattning med hänsyn till heteroskedasticitet

- enligt (11.2.2):

Kan visas att

$$Var(\hat{\beta}_1^{GLS}) < Var(\hat{\beta}_1^{OLS*})$$

och KI kommer att vara för breda och  $p$ -värden kommer att vara missvisande.

OLS skattning utan någon hänsyn alls

- inte ens enligt (11.2.2):

Kan visas att

$$\widehat{Var}(\hat{\beta}_1^{OLS}) = \frac{S_e^2}{\sum (x_i - \bar{x})^2}$$

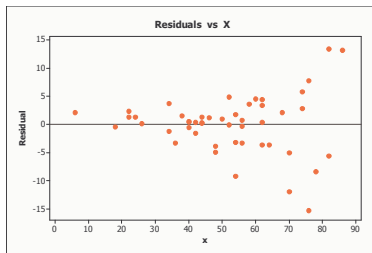
inte är väntevärdesriktig (biased). Vi vet inte ens om den över- eller underskattar den sanna variansen. Slut-satser av en sådan analys kan var helt missvisande.

## När är det heteroskedastiskt?

Ofta har man ju ingen aning förrän man har ett stickprov att titta på.

- Skatta som om det vore homoskedastiskt, dvs vanlig OLS skattning och analysera residualerna.

Grafiskt: Titta på residualplottar. Typiskt utseende kan vara



Formella test: Finns olika beskrivna i litteraturen

- Park test
- Glejser test
- Spearman's rangkorrelationstest (se tidigare OH)
- Goldfeld-Quandt test
- Breusch-Pagan-Godfrey test

Vi behöver inte fördjupa oss i detaljerna om dessa test. Men vi ska veta att det finns sätt att testa för lika varians ( $H_0$  är typiskt att det inte är heteroskedastiskt, låga  $p$ -värden gör att vi förkastar  $H_0$ ).

## Åtgärder i skattningsförfarandet

När  $\sigma_i^2$  är känd: WLS (se exempel)

När  $\sigma_i^2$  är okänd: Whites Heteroskedasticity-Consistent Variances and Standard Errors (kanske tveksamt vid små stickprov).

När man har en kvalificerad gissning: man kan ibland se att variansen i residualerna kan skrivas som en funktion av prediktorn (prediktorerna) dvs

$$\sigma_i^2 = f(X_i)$$

Ex 1) Antag att

$$E(\varepsilon_i^2) = \sigma^2 X_i^2$$

dvs residualvariansen ser ut att vara proportionell mot kvadraten på prediktorn. Transformerar data enligt

$$\begin{aligned} \frac{Y_i}{\sigma_i} &= \beta_0 \left( \frac{X_{0i}}{\sigma_i} \right) + \beta_1 \left( \frac{X_i}{\sigma_i} \right) + \left( \frac{\varepsilon_i}{\sigma_i} \right) \\ &= \beta_0 \left( \frac{X_{0i}}{X_i} \right) + \beta_1 \left( \frac{X_i}{X_i} \right) + \left( \frac{\varepsilon_i}{X_i} \right) \\ &= \beta_0 \frac{1}{X_i} + \beta_1 + \varepsilon_i^* \end{aligned}$$

och där  $X_i \neq 0$  ( $> 0$ ).

Ex 2) Antag att

$$E(\varepsilon_i^2) = \sigma^2 X_i$$

dvs residualvariansen ser ut att vara proportionell mot prediktorn. Transformerar data enligt

$$\begin{aligned} \frac{Y_i}{\sigma_i} &= \beta_0 \left( \frac{X_{0i}}{\sigma_i} \right) + \beta_1 \left( \frac{X_i}{\sigma_i} \right) + \left( \frac{\varepsilon_i}{\sigma_i} \right) \\ &= \beta_0 \left( \frac{X_{0i}}{\sqrt{X_i}} \right) + \beta_1 \left( \frac{X_i}{\sqrt{X_i}} \right) + \left( \frac{\varepsilon_i}{\sqrt{X_i}} \right) \\ &= \beta_0 \frac{1}{\sqrt{X_i}} + \beta_1 \sqrt{X_i} + \varepsilon_i^* \end{aligned}$$

och där  $X_i > 0$ .

Ex 3) Antag att

$$E(\varepsilon_i^2) = \sigma^2 E(Y_i | X)^2 = \sigma^2 (\beta_0 + \beta_1 X_i)^2$$

dvs residualvariansen ser ut att vara proportionell mot kvadraten på prediktorn. Transformerar data enligt

$$\begin{aligned} \frac{Y_i}{\sigma_i} &= \beta_0 \left( \frac{X_{0i}}{\sigma_i} \right) + \beta_1 \left( \frac{X_i}{\sigma_i} \right) + \left( \frac{\varepsilon_i}{\sigma_i} \right) \\ &= \beta_0 \left( \frac{X_{0i}}{E(Y_i | X)} \right) + \beta_1 \left( \frac{X_i}{E(Y_i | X)} \right) + \left( \frac{\varepsilon_i}{E(Y_i | X)} \right) \end{aligned}$$

och där  $E(Y_i | X) \neq 0 (> 0)$ . Förutsätter att man vet vad  $\beta_0$  och  $\beta_1$  är!

Tvåstegsförfarande:

1. Skatta  $\beta_0$  och  $\beta_1$  med OLS

2. Beräkna  $\hat{Y}_i$  från den skattade modellen och använd detta isf  $E(Y_i | X)$ .

Ex 4) Antag en modell med log-transformation enligt

$$\ln Y_i = \beta_0 X_{0i} + \beta_1 X_i + \varepsilon_i$$

eller

$$Y_i = \beta_0 X_{0i} + \beta_1 \ln X_i + \varepsilon_i$$

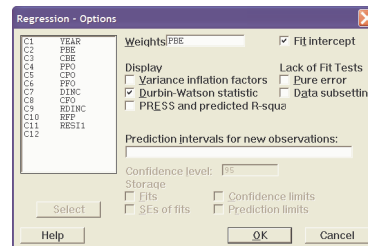
eller

$$\ln Y_i = \beta_0 X_{0i} + \beta_1 \ln X_i + \varepsilon_i$$

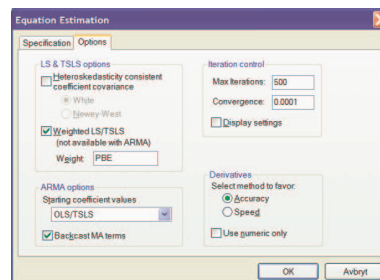
Detta kan ofta reducera heteroskedasticiteten i observerade datamaterial.

Hur gör man i praktiken?

Ex Minitab: Options-knappen ger följande dialogfönster. Observera "Weights"-rutan.



Ex EViews: vid estimering, Options-fliken. Observera "Weights"-rutan.



## F9 Avvikelser från modellantagandena II

### Autokorrelation

Enligt antagandena för linjär regression ska slumptermerna  $\varepsilon_i$  vara fördelade enligt

$$\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

(iid = oberoende, lika fördelade).

Förra gången tittade vi på vad som händer om de inte var lika fördelade, speciellt att de hade olika varians.

Nu ska vi titta vad som händer om de inte är oberoende.

Vad är problemet? Samma som förra gången:

Skattningarna är fortfarande väntevärdesriktiga (unbiased), konsistenta, asymptotiskt normalfördelade.

Men de är inte längre *effektiva*.

Formellt antar vi vid vanlig enkel linjär regression att

$$E(\varepsilon_i \varepsilon_j) = 0 \quad i \neq j$$

men nu gäller alltså att (för vissa par, kombinationer el. liknade) att

$$E(\varepsilon_i \varepsilon_j) \neq 0 \quad i \neq j$$

Om detta är fallet säger vi att autokorrelation.

Speciellt vanligt och intressant är det i samband med analys av tidsserier (men även spatiala data).

Typiskt när i tiden närliggande observationer (slumpstermer) ser ut att samvariera, tex

$$E(\varepsilon_t \varepsilon_{t+s}) \neq 0 \quad s \neq 0$$

### Orsaker till autokorrelation

Specifikationsfel: variabel saknas i modellen.

Antag att den "sanna" modellen är

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_t$$

men vi prövar modellen

$$Y_t = \beta_0 + \beta_1 X_{1t} + \nu_t$$

Detta innebär att

$$\nu_t = \beta_2 X_{2t} + \varepsilon_t$$

och vi kan få in en autokorrelation via det linjära sambandet mellan  $X_{2t}$  och de övriga.

Lösning: kör båda och se om autokorrelationen försvinner med den utökade modellen.

Specifikationsfel: felaktig funktionell form.

Antag att den "sanna" modellen är

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_t^2 + \varepsilon_t$$

men vi prövar modellen

$$Y_t = \beta_0 + \beta_1 X_t + \nu_t$$

Detta innebär att att vi kommer se det typiska mönstret för kvadrater i residualerna vilket kan ge upphov till en (falsk) autokorrelation.

Cobweb fenomen

En återkopplande effekt, tex om priset är högt, producerar man mer nästa år, priset går ner och nästa år producerar man mindre

$$\text{Tillång}_t = \beta_0 + \beta_1 \text{Pris}_{t-1} + \varepsilon_t$$

Effekten blir att det hela hoppar fram och tillbaka  $\Rightarrow$  autokorrelation.

### Laggar (lags)

Typiskt ser man ofta en seriell korrelation i ekonomiska sammanhang, tex priset idag är ungefär som i går, BNP ändras långsamt (cyklist) mm.

En modell skulle kunna skrivas enligt

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 Y_{t-1} + \varepsilon_t$$

Kan ses som specialfall av "Specifikationsfel: variabel saknas i modellen"

### Manipulering av data: aggregering / interpolering

Säg att man har dagskurser men aggregerar till månadsmedelvärden. Detta får till följd att små dagliga störningar dämpas ut, en uppåt kommer tas ut av en störning neråt. Serien blir "mjukare" än innan.  $\Rightarrow$  autokorrelation.

Eller man har årsdata men man interpolerar till månader. Eftersom årsdata är "mjuk" från början kommer interpolationerna också att spegla detta  $\Rightarrow$  autokorrelation.

### Transformering av data

Säg att ni tittar på differenser från en dag till nästa, dvs antag att modellen kan skrivas

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 X_t + \varepsilon_t \\ Y_{t-1} &= \beta_0 + \beta_1 X_{t-1} + \varepsilon_{t-1} \end{aligned}$$

och antag att

$$\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

Men differenserna kan skrivas

$$\Delta Y_t = \beta_1 \Delta X_t + \nu_t$$

där

$$\nu_t = \Delta \varepsilon_t = \varepsilon_t - \varepsilon_{t-1}$$

och vi får

$$\begin{aligned} E(\nu_t \nu_{t-1}) &= E[(\varepsilon_t - \varepsilon_{t-1})(\varepsilon_{t-1} - \varepsilon_{t-2})] \\ &= E(\varepsilon_t \varepsilon_{t-1}) - E(\varepsilon_t \varepsilon_{t-2}) \\ &\quad - E(\varepsilon_{t-1} \varepsilon_{t-1}) + E(\varepsilon_{t-1} \varepsilon_{t-2}) \\ &= 0 - 0 + \sigma_\varepsilon^2 + 0 \end{aligned}$$

dvs autokorrelation.

### Stationäritet och icke-stationäritet

om en tidserie är (svagt) stationär behåller den sina statistiska egenskaper genom hela serien, dvs

$$E(\varepsilon_t) = 0$$

$$Var(\varepsilon_t) = \sigma_\varepsilon^2$$

$$Cov(\varepsilon_t, \varepsilon_{t+s}) = \gamma_s$$

Om ovanstående inte är uppfyllt säger vi att serien är icke-stationär och vi kan i så fall se tecken på autokorrelation som egentligen är effekten av en trend i serien.

### OLS skattning när man har autokorrelation

Antag att modellen kan skrivas enligt

$$Y_t = \beta_0 + \beta_1 X_t + \nu_t$$

och där

$$\nu_t = \rho \nu_{t-1} + \varepsilon_t \quad |\rho| < 1$$

och där serien av  $\varepsilon_t$  är stationär dvs

$$E(\varepsilon_t) = 0 \quad Var(\varepsilon_t) = \sigma_\varepsilon^2$$

och dessutom ej autokorrelerade

$$Cov(\varepsilon_t, \varepsilon_{t+s}) = 0 \quad s \neq 0$$

Då har vi det som kallas en första ordningens *autoregressiv* modell eller en AR(1) i slumptermerna  $\nu_t$ .

Man kan visa att vi får

$$Var(v_t) = \frac{\sigma_\varepsilon^2}{1 - \rho^2} \quad (\text{homosked.})$$

$$Cov(v_t, v_{t+s}) = E(v_t v_{t+s}) = \rho^s \frac{\sigma_\varepsilon^2}{1 - \rho^2}$$

$$Corr(v_t, v_{t+s}) = \rho^s$$

Observera att ovanstående innebär att serien av  $v_t$  är stationär men poängen är att de *inte är oberoende*. Serien av  $\varepsilon_t$  är däremot oberoende (och stationär).

Observera också att  $v_t$  är korrelerade inte endast med  $v_{t-1}$  ( $\rho^1 = \rho$  ovan) utan med varje tidpunkt  $v_{t-s}$  innan ( $\rho^s$  ovan). Om det gäller att  $|\rho| < 1$  inses att styrkan i beroendet också avtar med  $s$ .

Under den vanliga modellen (med oberoende) är den vanliga OLS-skattningen av  $\beta_1$  som vanligt

$$\hat{\beta}_1^{OLS} = \frac{\sum x_t y_t}{\sum x_t^2}$$

(vi antar nu att vi har medelvärdesjusterat observationerna så att  $\bar{x}, \bar{y} = 0$ ) med variansen

$$Var(\hat{\beta}_1^{OLS}) = \frac{\sigma_\varepsilon^2}{\sum x_t^2} \quad (12.2.7)$$

Men under AR(1) modellen blir variansen istället

$$Var(\hat{\beta}_1^{OLS})_{AR1} = \frac{\sigma_\varepsilon^2}{\sum x_t^2} \left( 1 + 2 \sum_{s=1}^{n-1} \rho^s \frac{\sum x_t x_{t-s}}{\sum x_t^2} \right) \quad (12.2.8)$$

och om vi bortser ifrån AR(1) komponenten riskerar vi grova felkattningar av variansen!

Ex) Antag att även  $X_t$  följer är en AR(1) process, dvs

$$X_t = r X_{t-1} + \delta_t \quad |r| < 1$$

där  $r$  är  $X_t$ 's autokorrelationskoefficient. Det kan då visas att

$$Var(\hat{\beta}_1^{OLS})_{AR1} = Var(\hat{\beta}_1^{OLS}) \left( \frac{1+r\rho}{1-\rho r} \right)$$

Beroende på vilka värden  $r$  respektive  $\rho$  antar kan vi antingen över- eller underskatta den sanna variansen för  $\hat{\beta}_1$ .

BLUE-estimat vid autokorrelerade slumpstermer

Vi tar till en variant av GLS och det kan visas att BLUE-skattning blir

$$\hat{\beta}_1^{GLS} = \frac{\sum (x_t - \rho x_t)(y_t - \rho y_t)}{\sum (x_t - \rho x_t)^2} + C$$

och variansen är

$$Var(\hat{\beta}_1^{GLS}) = \frac{\sigma_\varepsilon^2}{\sum (x_t - \rho x_t)^2} + D$$

där  $C$  och  $D$  är korrektionsfaktorer som kan bortses ifrån. Denna skattning är BLUE.

(Allmänt om GLS är att man använder mer information än vad OLS gör.)

Vad händer om man "gör fel"?

### OLS skattning med hänsyn till autokorrelation

- enligt (12.2.8):

Kan visas att

$$\text{Var}(\hat{\beta}_1^{GLS}) < \text{Var}(\hat{\beta}_1^{OLS})_{AR1}$$

och KI kommer att vara för breda och  $p$ -värden kommer att vara missvisande.

### OLS skattning utan någon hänsyn alls

- inte ens enligt (12.2.8) utan med (12.2.7):

Man riskerar att underskatta  $\sigma_\varepsilon^2$  och därmed överskatta  $R^2$ . Även om man inte underskattar  $\sigma_\varepsilon^2$  riskerar man att felskatta (12.2.8) om man använder (12.2.7) och därmed kommer  $t$  och  $F$ -test vara missvisande (felaktiga  $p$ -värden). Slutsatser av en sådan analys kan var helt missvisande.

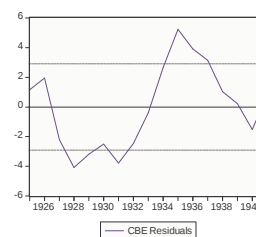
Hur upptäcker man att det är autokorrelerat?

Ofta har man ju ingen aning förrän man har ett stickprov att titta på.

- Skatta som om det inte finns autokorrelation, dvs vanlig OLS skattning, och analysera residualerna, dvs vi kommer att få en serie av  $\hat{v}_t$ .

Grafiskt: Titta på residualplottar. Kan man se mönster i  $\hat{v}_t$ ? Vilka i så fall?

Exempel)



Här kan man se att serien av residualer möjligen rör sig lite för mjukt, närliggande residualer är relativt lika.

Formella test: Finns olika beskrivna i litteraturen

- Runs test (se tidigare OH)

- Durbin-Watson d-test: Används för att detektera om det finns en AR(1), ej högre ordningar typ AR(2). Definiera

$$d = \frac{\sum_{t=2}^n (\hat{v}_t - \hat{v}_{t-1})^2}{\sum_{t=1}^n \hat{v}_t^2} \approx 2(1 - \hat{\rho})$$

där

$$\hat{\rho} = \frac{\sum_{t=2}^n \hat{v}_t \hat{v}_{t-1}}{\sum_{t=1}^n \hat{v}_t^2}$$

är en skattning av  $\rho$ . Eftersom  $|\rho| < 1$  gäller att

$$0 < d < 4$$

- om  $\rho = 0$ , dvs ingen autokorrelation så är  $d = 2$ .

- om  $\rho$  är nära 1 så är  $d$  nära 0

- om  $\rho$  är nära -1 så är  $d$  nära 4.

Man letar upp två kritiska gränser  $d_L$  och  $d_U$  i Durbin-Watson tabellen (gränserna beror på  $n$  och  $p$ -värde).

Om  $d$  ligger i intervallet

-  $(d_U, 4 - d_U)$  så är det inte autokorrelerat.

-  $(0, d_L)$  eller i  $(4 - d_L, 4)$  så är det autokorrelerat

-  $(d_L, d_U)$  eller i  $(4 - d_U, 4 - d_L)$  vet man inte säkert.

Nackdel med metoden är att det finns s.k. "indecisive zones" och att man endast testar för autokorrelation på lag 1. Dessutom kan autokorrelation i en regressor maskera autokorrelation i slumptermerna.

- Breusch-Godfrey (BG) test

också känd som Lagrange Multiplier (LM) test:

Se texten.

- Box-Ljung Q-statistikor: Definiera

$$Q = n(n+2) \sum_{k=1}^M \frac{\hat{\rho}_k^2}{(n-k)}$$

där  $\hat{\rho}_k$  är en skattning av autokorrelationen på lagg  $k$ . Observera att  $Q$  är ett test för flera autokorreleationer samtidigt, ett s.k. portmanteau test.

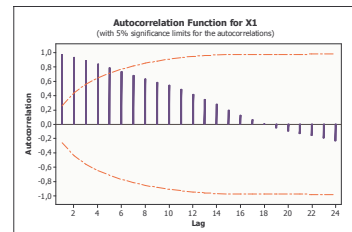
$Q$  är approximativt  $\chi^2$ -fördelad om modellen är korrekt specificerad och vid stora värden på  $Q$  förkastas nollhypotesen om inga autokorrelationer upp till lagg  $M$ .

Mycket vanligt är att man tittar på Autokorrelationsfunktionen (ACF) vilket är skattningar av

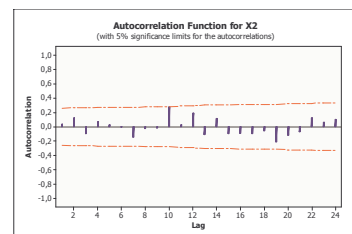
$$\text{Corr}(v_t, v_{t+s}) = \rho_s$$

dvs autokorrelationer på olika laggar  $s$ , dvs olika tid-savstånd  $s$ . Dessa kan åskådliggöras i diagram.

Ex) Tecken på autokorrelation



Ex) Tecken på avsaknad av autokorrelation



# F10 Ekonometrisk tidserier

## Stokastiska processer

En *stokastisk process* är en samling slumpvariabler ordnade efter tid.

Tex om  $Y$  är en slumpvariabel som vi observerar flera gånger (i tiden) kan vi betrakta det som stokastisk process.

Förtydligande: Om *tiden* är kontinuerlig skriver man  $Y(t)$ , om tiden är diskret skriver man  $Y_t$ .

$Y_t$  självt är typiskt kontinuerlig.

## Stationära processer

Om en tidserie är *svagt stationär* behåller den sina statistiska egenskaper genom hela serien, dvs

$$E(Y_t) = \mu, \quad Var(Y_t) = \sigma_Y^2$$

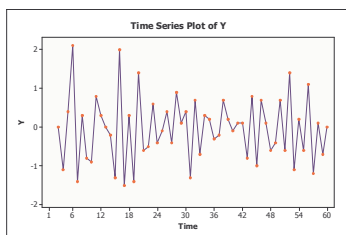
$$Cov(Y_t, Y_{t+k}) = \gamma_k$$

Dvs oavsett  $t$  så ska ovanstående gälla. Om ovanstående inte är uppfyllt säger vi att serien är icke-stationär och vi kan i så fall se tecken på autokorrelation som egentligen är effekten av en trend i serien.

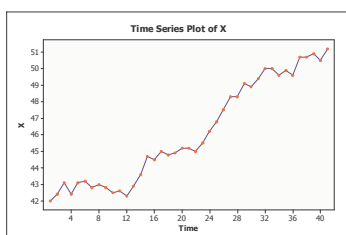
Man kan säga att en störning som får serien att "hoppa till" inte kommer att ha en bestående effekt utan serien kommer relativt snabbt tillbaks till värden runt medelvärdet (*mean reversion*). Dessutom ska fluktuationerna runt medelvärdet vara ungefär lika stora (i varians).

(Finns en definition för det som kallas stark stationaritet men ovanstående räcker ofta i praktiken.)

## Hyfsad stationär serie (möjligen varierande $Var(Y_t)$ )



## Ej stationär serie med avseende på väntevärdet $E(Y_t)$



## Random Walk el Slumpvandring utan drift

Klassiskt exempel på icke-stationär serie är en Random Walk Model (RWM) eller slumpvandring dvs en modell enligt

$$Y_t = Y_{t-1} + u_t$$

där  $u_t$  är vitt brus (*white noise*) eller mao iid slumpstermer enligt

$$u_t \sim N(0, \sigma^2)$$

$Y_t$  är alltså en AR(1) process.

Man kan skriva

$$\begin{aligned} Y_t &= Y_{t-1} + u_t = Y_{t-2} + u_t + u_{t-1} \\ &= Y_{t-3} + u_t + u_{t-1} + u_{t-2} \\ &= \dots \\ &= Y_0 + \sum_{i=1}^t u_i \end{aligned}$$

och man inser att

$$E(Y_t) = Y_0$$

$$Var(Y_t) = t\sigma^2$$

alltså är den *inte* stationär med avseende på variansen.

$Y_t$  är alltså en summa av ett startvärde och en summa av slumpmässiga störningar  $u_t$ .

Vidare ser man effekten av dessa störningarna aldrig dör ut!

Men observera att om vi bildar *differenser* av intelligande  $Y_t$  får vi

$$\begin{aligned}\Delta Y_t &= Y_t - Y_{t-1} \\ &= Y_0 + \sum_{i=1}^t u_i - Y_0 + \sum_{i=1}^{t-1} u_i \\ &= u_t\end{aligned}$$

dvs den första ordningens differenser är en serie av slumpstermer, vitt brus.

Denna serie är stationär eftersom

$$E(\Delta Y_t) = E(u_t) = 0 \quad (\text{konstant})$$

och

$$Var(\Delta Y_t) = Var(u_t) = \sigma^2 \quad (\text{konstant})$$

och dessutom är dessa oberoende och därmed okorrelerade (konstant).

## Random Walk el Slumpvandring med drift

Modifiera RWM-modellen enligt

$$Y_t = \delta + Y_{t-1} + u_t$$

där  $\delta$  är en sk *driftparameter* och  $u_t$  är som förut vitt brus.

Man kan skriva

$$\begin{aligned}Y_t &= \delta + Y_{t-1} + u_t = 2\delta + Y_{t-2} + u_t + u_{t-1} \\ &= 3\delta + Y_{t-3} + u_t + u_{t-1} + u_{t-2} \\ &= \dots \\ &= t\delta + Y_0 + \sum_{i=1}^t u_i\end{aligned}$$

och man inser att

$$E(Y_t) = t\delta + Y_0$$

$$Var(Y_t) = t\sigma^2$$

alltså är den *inte* stationär med avseende på väntevärde och varians.

Termen  $t\delta$  i väntevärdet ovan är orsaken till benämningen drift.

Ju längre tid som går desto längre bort från  $Y_0$  kommer man (i förväntan)

En RWM-process utan drift däremot förväntas komma tillbaka till  $Y_0$  för eller senare.

Men i båda fallen ökar variansen.

## Unit Root Process

RVM är exempel på det som kallas *Unit Root* processer. Definiera

$$Y_t = \rho Y_{t-1} + u_t$$

dvs en AR(1) process. Om nu  $\rho$  är lika med ett (1) får vi unit root problem, dvs vi får icke-stationäritet.

Definiera  $B$  som en bakåtoperator (*back-shift operator*) enligt

$$B^k Y_t = Y_{t-k}$$

Då ser man att tex första differenser kan skrivas enligt

$$\Delta Y_t = Y_t - Y_{t-1} = (1 - B) Y_t$$

Modellen ovan kan då skrivas som

$$(1 - \rho B) Y_t = u_t$$

Betrakta  $(1 - \rho B)$  som ett polynom i  $B$ . Roten till ekvationen

$$1 - \rho B = 0$$

är för vilka  $B$  gäller detta?

Allmänt har man kanske

$$(1 - \rho_1 B - \rho_2 B^2 - \rho_3 B^3 - \dots) = 0$$

Om detta polynom har  $B = 1$  som en lösning får vi unit root problem, dvs icke-stationäritet.

## Trend- och Differensstationär

Trend är den långsamma utvecklingen i en tidserie.

Man skiljer på

- deterministiska trender (helt och hållet förutsägbar) och

- stokastiska trender

För att formalisera det hela definierar vi följande modell

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 t + u_t$$

Ur detta kan vi studera några specialfall:

1) Antag att  $\beta_0 = \beta_2 = 0$  och  $\beta_1 = 1$ , då får vi

$$Y_t = Y_{t-1} + u_t$$

dvs en RVM utan drift. Men kom ihåg att

$$\Delta Y_t = u_t$$

är stationär. En RWM är en *differensstationär process*.

2) Antag att  $\beta_0 \neq 0$ ,  $\beta_1 = 1$  och  $\beta_2 = 0$ , då får vi

$$Y_t = \beta_0 + Y_{t-1} + u_t$$

dvs en RVM med drift. Och vi kommer att se en trend uppåt eller nedåt beroende på tecknet på  $\beta_0$ . Detta är en *stokastisk trend*.

3) Antag att  $\beta_0 \neq 0$ ,  $\beta_1 = 0$  och  $\beta_2 \neq 0$ , då får vi

$$Y_t = Y_t = \beta_0 + \beta_2 t + u_t$$

får man en *trendstationär process*. Väntevärdet är ej konstant men variansen är konstant.

4) Antag att  $\beta_0 \neq 0$ ,  $\beta_1 = 1$  och  $\beta_2 \neq 0$ , då får vi

$$Y_t = \beta_0 + Y_{t-1} + \beta_2 t + u_t$$

får man en RWM-process med drift och deterministisk trend.

5) Antag att  $\beta_0 \neq 0$ ,  $|\beta_1| < 1$  och  $\beta_2 \neq 0$ , då får vi

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 t + u_t$$

får man en process med deterministisk trend och stationär AR(1) komponent. Denna process är stationär runt den deterministiska trenden.

## Integrerade stokastiska processer

En *integrerad* stokastisk process (IP) är en process som typiskt är icke-stationär men som efter ett bestämt antal *differensbildningar* är stationär.

Man säger att en IP är av ordning  $d$  om det krävs  $d$  stycken differenser för att bli stationär, eller kort

$$Y_t \sim I(d)$$

Ex) En RWM är icke-stationär (se ovan) men bildas första differens enligt

$$\Delta Y_t = Y_t - Y_{t-1} = u_t$$

så är denna stationär.

Egenskaper hos integrerade stokastiska processer

1) Om  $X_t \sim I(d)$  så är en linjär transformation av denna integrerad enligt

$$Z_t = aX_t + b \sim I(d)$$

dvs den behåller sin ordning.

2) Om  $X_t \sim I(d_1)$  och  $Y_t \sim I(d_2)$  och  $d_1 < d_2$  så är en linjärkombination av dessa integrerad enligt

$$Z_t = aX_t + bY_t + c \sim I(d_2)$$

dvs den största av  $d_1$  och  $d_2$

3) Om  $X_t \sim I(d)$  och  $Y_t \sim I(d)$ , dvs samma ordning, så är en linjärkombination av dessa integrerad enligt

$$Z_t = aX_t + bY_t + c \sim I(d^*)$$

där  $d^* \leq d$ , dvs ordningen är generellt lika med  $d$  men i vissa lägen kan den vara lägre och man säger i så fall att  $X_t$  och  $Y_t$  *kointegrerar*.

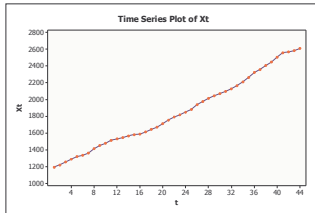
- Kan bli problem vid regressionsanalyser om prediktor och respons är av olika integrerande ordning. Om tex  $Y_t \sim I(0)$  och  $X_t \sim I(1)$  kan se att variansen för  $X_t$  växer med  $t$  (icke-stationär) och att skattningen för  $\beta_1$  går mot noll.

- Regressionsanalyser med integrerande processer kan leda till spuriösa resultat (tex starkölsförsäljning och antal inskrivna vid universitet).

## Test för stationäritet

Grafiskt: Titta på tidserieplottar. Ser det stationärt ut?

Exempel)



Ökande trend (ej stationärt medelvärde).

## Autokorrelationer (ACF):

ACF vid lag  $k$  definieras

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\text{Cov}(Y_t, Y_{t-k})}{\text{Cov}(Y_t, Y_{t-0})} = \frac{\text{Cov}(Y_t, Y_{t-k})}{\text{Var}(Y_t)}$$

För att testa om ACF är signifikant skild från noll mer formellt kan man titta på Ljung-Pierce Q statistika

$$Q = n \sum_{k=1}^M \hat{\rho}_k^2$$

eller ännu hellre Ljung-Box Q-test (LB Q test),

$$Q = n(n+2) \sum_{k=1}^M \frac{\hat{\rho}_k^2}{(n-k)}$$

Nollhypotesen i vardera fall är  $H_0$  : samtliga autokorrelationer upp till lag  $M$  är noll.

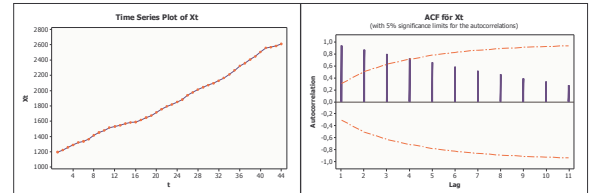
Skattas med

$$\hat{\gamma}_k = \frac{\sum (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{n}$$

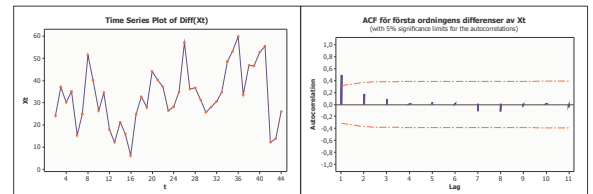
och

$$\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}$$

Plotta sedan dessa för  $k = 1, 2, 3, \dots$  tex



Ikke stationärt beteende



Fortfarande beroende men möjligen stationärt.

## Unit Root test

Vi såg att RWM utan drift var icke-stationär. Vi såg också att den kunde formuleras enligt

$$Y_t = Y_{t-1} + u_t \Leftrightarrow (1 - B)Y_t = u_t$$

Vidare såg vi att en generell process av detta slag kan skrivas

$$Y_t = \rho Y_{t-1} + u_t \Leftrightarrow (1 - \rho B)Y_t = u_t$$

där  $|\rho| \leq 1$  och att om  $\rho = 1$  så är det en RWM. Nu skriver detta som

$$\begin{aligned} \Delta Y_t &= Y_t - Y_{t-1} = \rho Y_{t-1} - Y_{t-1} + u_t \\ &= (\rho - 1)Y_{t-1} + u_t \\ &= \delta Y_{t-1} + u_t \end{aligned}$$

Frågan som är intressant nu är att avgöra om

$$\delta = 0 \quad \text{eller} \quad \delta < 0$$

vilket är ekvivalent med

$$\rho = 1 \quad \text{eller} \quad \rho < 1$$

Om  $\delta = 0$  så har vi en enhetsrot (*unit root*) och en RWM och således icke-stationäritet.

## Dickey-Fuller (DF) test

Gäller när  $u_t$  är vitt brus dvs okorrelerade slumpstermer. Använd formuleringen

$$\Delta Y_t = \delta Y_{t-1} + u_t$$

och titta på skattningen av  $\delta$ . Är denna signifikant skild från noll eller inte?

Går ej med vanlig  $t$ -test ty teststatistikan är inte  $t$ -fördelad utan  $\tau$ -fördelad (tau) vilket kräver speciella tabeller.

Vidare måste man ta hänsyn till vilken sorts process det är frågan om:

- 1)  $Y_t$  är en RWM utan drift
- 2)  $Y_t$  är en RWM med drift
- 3)  $Y_t$  är en RWM med drift runt en stokastisk trend

Beroende vilket som gäller gäller olika kritiska värden från  $\tau$ -fördelningen.

## Augmented Dickey-Fuller (ADF) test

Om det finns korrelationer i  $u_t$  finns en vidareutveckling av DF-testet, ADF.

## Transformationer

Inte så komplicerat, läs själva.