Spesialisering: Anvendt makro
5. Modul

1.B Lineære regresjonsmodeller og minste kvadraters metode (MKM)

Drago Bergholt
Norwegian Business School (BI)

10. november 2011

Oversikt

I. Introduksjon til økonometri

II. Lineær regresjonsanalyse - Oversikt og notasjon

III. Minste Kvadraters Metode (MKM)

IV. Forutsetninger for analyse

V. Problemer

- Autokorrelasjon, Multikolinearitet, Heteroskedastisitet

I. Introduksjon - Hva er økonometri?

- ✓ Kvantifiserer sammenhenger i økonomien ved å kombinere:
  - ❖ Økonomisk teori
  - ❖ Statistisk(matematisk) teori
  - ❖ Økonomiske data
    - Tidsrekkedata
    - Tverrsnittsdata
    - Paneldata

✓ *Økonometrisk metode innebærer å spesifisere:*

a) En hypotese man ønsker å teste

b) En økonometrisk modell for å teste teorien/hypotesen

c) Estimere parametrene i den valgte modellen

d) Verifisere statistisk forklaringskraft

e) Prognoser

f) Bruk av modell for politikkanalyser

a-b) Spesifikasjon av hypotese og økonometrisk modell

- ✓ Keynesiansk konsumfunksjon: *"Consumption increases as income increases, but not by as much as the increase in income."* $0<MPC<1$

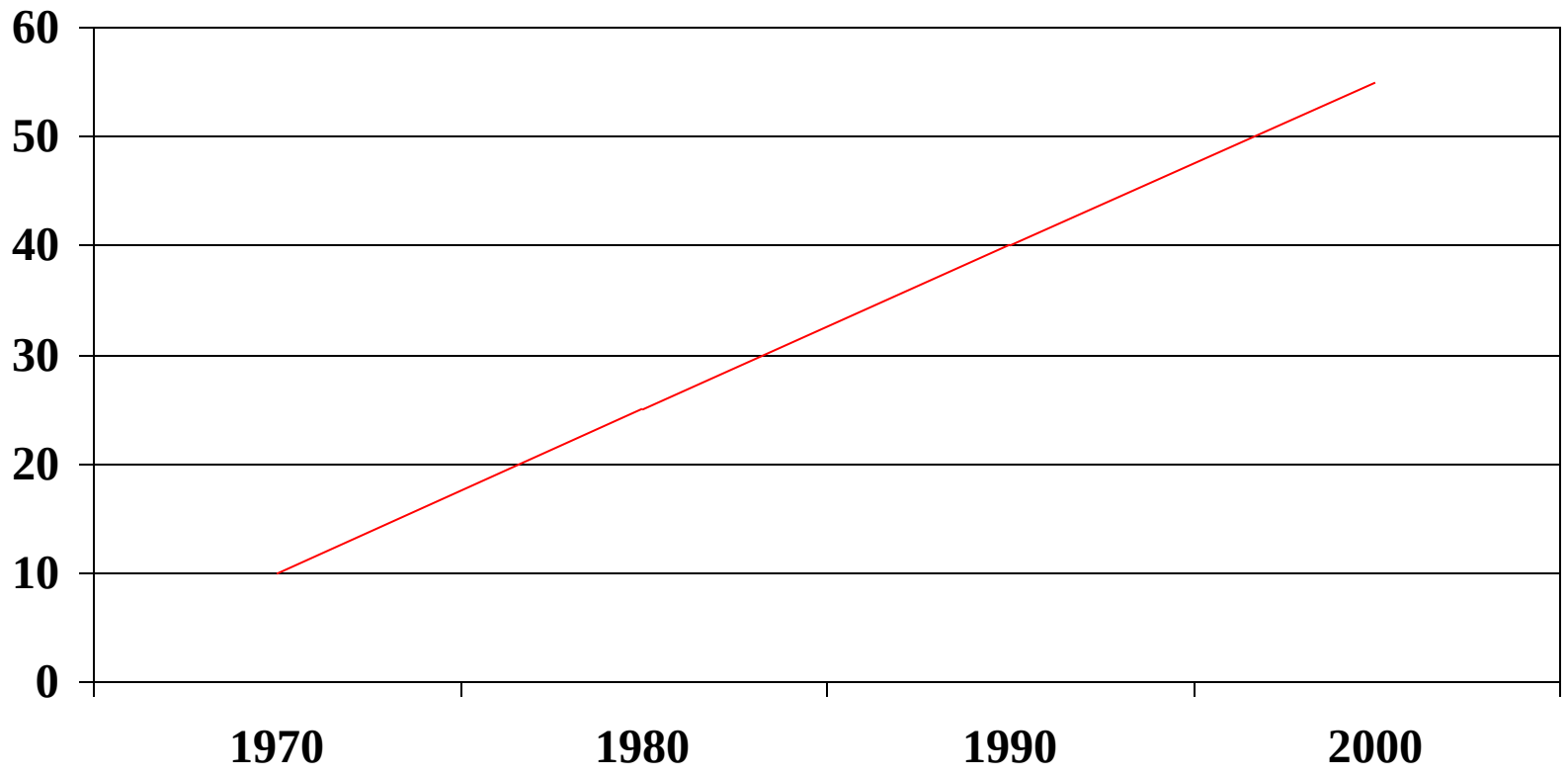
- ✓ Matematisk modell (Y; Konsum, X; Inntekt):

$$Y=\alpha+\beta X, \quad 0<\beta<1$$

- ✓ Statistisk modell (ε er et stokastisk restledd)

$$Y=\alpha+\beta X+\varepsilon$$

Keynesiansk konsumfunksjon



c-d) Estimering og verifisering

- ✓ Regresjonsanalyse, et vanlig verktøy for å estimerer økonomiske sammenhenger:
  - Hvor mye av variasjonen i Y kan forklares av X?

- ✓ Verifisering (statistisk inferens)/hypotese testing
  - Anta at β (MPC) = 0,9. Er dette signifikant forskjellig fra 1, eller et resultat av tilfeldigheter?

e-f) Prediksjon og politikk eksperimenter

- ✓ Prediksjon av Y basert på gitt (forventet) verdi på X
  - Men en dårlig prediksjon betyr ikke at man skal forkaste modellen. Den kan ha stor forklaringskraft over estimert periode.
  - Det er de uforutsette hendelsene etter prognosetidspunktet som bidrar til de største prognosefeilene.
- ✓ Modellen kan brukes til politikkeksperimenter. Hvilket inntektsnivå vil garantere ett gitt nivå på konsumet?

II. Lineær regresjonsanalyse - Oversikt

- ✓ Regresjonsmodell –kan trekke slutninger med gyldighet utover det gitte materialet.
- ✓ Fra matematisk til statistisk modell

$$Y = \beta_0 + \beta_1 X \qquad (1)$$

$$\frac{\Delta Y}{\Delta X} = \frac{(Y_2 - Y_1)}{(X_2 - X_1)} = \beta_1$$

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (2)$$

$$E(Y|X) = \beta_0 + \beta_1 X, \quad Y = E(Y|X) + \varepsilon$$

En teoretisk sammenheng:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \qquad (3)$$

Estimeres som:

$$Y_t = \hat{\beta}_0 + \hat{\beta}_1 X_t + e_t \qquad (4)$$

$$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_t$$

Gir e_t som residual:

$$e_t = Y_t - \hat{Y}_t$$

Mens restleddet (error term) er definert som:

$$\varepsilon_t = Y_t - E(Y_t \mid X_t)$$

- ✓ Regresjon versus korrelasjon
- ✓ Multivariate modeller. Kontrollerer for flere variable.

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_k X_{kt} + \varepsilon_t \qquad (5)$$

- ✓ Samme egenskaper som enkel regresjon

III. Minste kvadraters metode (MKM)
Føyning av en rett linje til data

- ✓ MKM *(Ordinary least squares, OLS)*: Den linjen som minimerer summen av kvadrerte residualer

$$\sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2 = RSS$$

- ✓ TSS = ESS + RSS

  Total Sum of Squares = Explained Sum of Squares
  + Residual Sum of Squares

$$\sum (Y_t - \bar{Y})^2 = \sum (\hat{Y}_t - \bar{Y})^2 + \sum e_t^2$$

Evaluering av modell

- R^2 "Coefficient of determination"

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_t^2}{\sum (Y_t - \bar{Y})^2}$$

- $0 \leq R^2 \leq 1$
- Merk: Lav R^2 ikke ensbetydende med dårlig modell!

Testing av hypoteser

- ✓ Teste H_0 (Nullhypotese) mot H_1 (Alternativ hypotese).

- ✓ Type I feil

  Man kan forkaste nullhypotesen selv om den er sann

  Sannsynlighet lik størrelsen på testen (α signifikansnivå).

- ✓ Type II feil

  Man kan unnlate å forkaste nullhypotesen selv om den er feil.

  Styrke på en test er sannsynligheten for at man korrekt forkaster den falske nullhypotesen = 1- Prob (type II feil).

t-test

- ✓ Test for individuelle koeffisienter

$$t = \frac{\hat{\beta}_1 - \beta_{H_0}}{s.e.(\hat{\beta}_1)}$$

- ✓ t-verdier følger en t-distribusjon med N-(K+1) frihetsgrader.

- ✓ Kritiske t-verdier (se tabell A2 s. 754 i Patterson)

- ✓ Forkast H_0 hvis $|t|>t_c$

- ✓ Ensidig test rundt null

  H_0: $\beta \leq 0$

  H_1: $\beta > 0$

  (eller motsatt)

- ✓ Tosidig test rundt null

  H_0: $\beta = 0$

  H_1: $\beta \neq 0$

- ✓ Tosidig kritisk nivå: t_{α}
- ✓ Ensidig kritisk nivå: $t_{\alpha/2}$
- ✓ F.eks. Velger α=5%, 25 d.f., t_{α}= 1,708, $t_{\alpha/2}$ = 2,060
- ✓ Dataprogrammer tester som regel H_0: $\beta = 0$
- ✓ Tommelfingerregel: Forkast H_0 hvis t>2
- ✓ Konfidensintervall = $\hat{\beta} \pm (t_c) s.e.(\hat{\beta})$

F-test

- ✓ Tester en hypotese som gjelder flere koeffisienter.

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$
($Y = \beta_0 + \varepsilon_t$)

H_1: H_0 er usann

$$F = \frac{ESS / n}{RSS /(n - k - 1)}$$

Forkast H_0 hvis $|F| \geq F_c$

IV. Forutsetninger for analyse
Klassisk lineær regresjonsmodell

1. Regresjons modellen er lineær i koeffisientene og er korrekt spesifisert.
2. Restleddet har forventning lik 0
3. Alle forklaringsvariablene er ukorrelerte med restleddet
4. Restleddene er ukorrelerte med hverandre målt over tid
5. Restleddet har konstant varians
6. Ingen av forklaringsvariablene kan skrives som perfekt lineær funksjon av noen av de andre forklaringsvariablene.
7. Restleddet er normalfordelt

1. Regresjonsmodellen er lineær i koeffisientene og er korrekt spesifisert

- ✓ Modellen må være lineær i koeffisientene, men ikke i variablene. Kan ta log.
- ✓ Modellen er korrekt spesifisert - ingen utelatte variable eller feilaktig funksjonsform.
- ✓ Additative restledd
- ✓ Estimeringsprosedyre (D. Hendry) “General to specific”, ikke “specific to general”.

- ✓ Akaike, Schwarz kriteria

  Tester bla.a for signifikante lags. Justerer RSS for utvalgsstørrelse (n) og antall uavhengige variable (K).

- ✓ Ramsey's Regression specification Error Test (RESET)

  Tester for sannsynligheten for utelatte variable, eller feil funksjonsform.

- ✓ Ikke "data mining"

- ✓ Dummier

2. Restleddet har forventning lik 0

- ✓ $E(\varepsilon_i|X_i)=0$
- ✓ Restleddet skal i gjennomsnitt ha en fordeling som er lik 0.
- ✓ I små utvalg vil ikke fordelingen være lik 0, men når utvalget går mot uendelig skal gjennomsnittet for fordelingen for restleddet gå mot 0.
- ✓ Konstantledd sikrer gjennomsnitt lik 0. (Fast andel av Y som ikke forklares av X'ene).
- ✓ Restledd: Stokastisk andel av Y som ikke forklares av X'ene.

3. Alle forklaringsvariablene er ukorrelerte med restleddet

- ✓ $Cov(\varepsilon_I, X_i) = E(\varepsilon_i X_i) = 0$
- ✓ Forklaringsvariablene er bestemt utenfor regresjonsanalysen og uavhengig av restleddet.
- ✓ Forklaringsvariablene og restleddet er korrelert: MKM vil gi X'ene noe variasjon fra Y, som kommer fra restleddet.
- ✓ X'er og restledd positivt korrelert, estimerte koeffisienter vil ha en bias oppover. (Høyere enn deres sanne verdier).
- ✓ Simultane ligningssystemer bryter denne forutsetningen.

4. Restleddene er ukorrelerte med hverandre målt over tid (ingen seriekorrelasjon)

- ✓ $E(\varepsilon_i \varepsilon_j)=0, \ i \neq j$
- ✓ Viktig i tidsserieanalyser
- ✓ Observasjoner av restleddet er trukket helt uavhengige av hverandre.
- ✓ Hvis det var en systematisk korrelasjon mellom de forskjellige observasjonene av restleddet over tid, vil bli vanskelig å få presise estimater på koeffisientene.

5. Restleddet har konstant varians (ingen heteroskedastisitet)

- ✓ $var(\varepsilon_i|X_i)=E(\varepsilon_i^2)=\sigma^2$
- ✓ Viktig for tverrsnittsanalyser, men også aktuelt problem i tidsseriestudier (regimeendringer etc.)
- ✓ Observasjonene av restleddet er trukket kontinuerlig fra like fordelinger.
- ✓ Gir upresise estimater - standardavviket feil.

6. Forklaringsvariablene kan ikke skrives som lineær funksjon av hverandre (ingen multikolinearitet)

- ✓ Perfekt kolinearitet - Samme variable.
- ✓ Additativt, konstantleddjustering, to variabler summerer seg til en tredje.
- ✓ Relative momenter vil være like selv om størrelsen vil variere.
- ✓ MKM kan ikke skille variablene fra hverandre.

7. Restleddet er normalfordelt

- ✓ Hvordan fordelingen ser ut.
- ✓ Observasjoner av restleddet er trukket fra en fordeling som er normalfordelt.
- ✓ En normalfordelingen ser symmetrisk ut.
- ✓ Viktig for hypotesetesting, ikke for MKM estimering

Gauss Markov Theorem og BLUE

- ✓ Gauss-Markov Theorem: Gitt forutsetningene fra den klassiske lineære regresjonsmodeller, vil MKM estimatene, blant en serie med ”unbiased” lineære estimater ha minst varians, med andre ord, de er BLUE

- ✓ *Lineær*, ”*Unbiased”* forventet verdi på koeffisientene er lik den sanne verdier $E(\hat{\beta}) = \beta$

- ✓ *minimum varians – Effisiente estimater*.

- ✓ Noen problemer:
  - ✓ Seriekorrelasjon
  - ✓ Multikolinearitet
  - ✓ Heteroskedastisitet

V Noen problemer
a) Seriekorrelasjon

- ✓ Restleddet fra en periode avhenger på en systematisk måte av restleddene fra tidligere perioder. $E(\varepsilon_i \varepsilon_j) \neq 0, \ i \neq j$
- ✓ Seriekorrelasjon er lik autokorrelasjon i tidsseriestudier
- ✓ Førsteordens seriekorrelasjon (AR):

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t$$

- ✓ $-1 < \rho < 1$
- ✓ Positiv eller negativ seriekorrelasjon

Årsak:

- ✓ Restleddet fanger opp utelatte variable, feil funksjonsform, ikke-lineæritet, manglende lags, målefeil etc.

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

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

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

Hvilke problemer gir det.

- ✓ Ingen feil i koeffisientene (unbiased), men ikke BLUE
- ✓ Øker variansen i fordelingen til koeffisientene (og t-verdier faller). Fanges ikke opp av MKM. MKM vil ikke lenger gi minimum varians.
- ✓ MKM vil underestimere standardavviket til de estimerte koeffisientene (og residualene), mens t-verdier og R^2 vil bli overestimert.
- ✓ Får feilaktig "bedre tilpasning".
- ✓ Mer sannsynlig at vi vil forkaste H_0 ($\beta=0$) når den er sann.

Hvordan oppdage seriekorrelasjon

- ✓ Se på et grafisk plot av restleddet
- ✓ Durbin-Watson d-statistikk

$$\hat{\rho} = \frac{\sum e_t e_{t-1}}{\sum e_t^2}$$

$$d = 2(1-\hat{\rho}), \qquad -1 \le \rho \le 1$$

$$0 \le d \le 4$$

- ✓ d=2, ingen første ordens seriekorrelasjon. $0<d<d_L$, positiv seriekorrelasjon, $4-d_L<d<4$, negativ seriekorrelasjon.

$$0 \; < \; d_L \; < \; d_U \; < \; 2 \; < \; 4 - d_U \; < \; 4 - d_L \; < \; 4$$

Hvordan bli kvitt seriekorrelasjon

- ✓ Tilføye utelatte variable hvis det er mulig
- ✓ Endre funksjonsform
- ✓ Generalised Least Squares