Erling Berge
POL3507 IMPLEMENTERING OG EVALUERING AV OFFENTLEG POLITIKK

Regression and quasi-experiments

Ref.: L. B. Mohr 1995 Chapter 7-9

Literature

- Breen, Richard 1996 “Regression Models. Censored, Sample Selected, or Truncated Data”, Sage University Paper: QASS 111, London, Sage
- Hamilton, Lawrence C. 1992 "Regression with graphics", Belmont, Duxbury, Kap. 1-7
- Hardy, Melissa A. 1992 “Regression with dummy variables” Sage University Paper: QASS 93, London, Sage,
- Mohr, Lawrence B. 1995 ”Impact Analysis for Program Evaluation”, Sage, London
- Winship, Christopher, and Robert D. Mare 1992 «Models for sample selection bias», Annual Review of Sociology, 18:327-350
- Winship, Chrisopher, and Stephen L. Morgan 1999 “The Estimation of Causal Effects from Observational Data”, Annual Review of Sociology Vol 25: 659-707

The comparative change design

- Selection: centralised and autonomous or controlled (not random and not by assignment variable)
- Any known or suspected difference between treatment group and control group should be measured in a suitable variable

Causes of a change

- Treatment come before measures of effect
  - Assignment to treatment/ control done to maximise chance fluctuations in causes other than treatment
  - Unmeasured causal variables may lead to a spurious causal relation between effect and treatment
- Effects explained by
  - Treatment
  - Spuriousness
  - Chance

Statistical tests

- How many "independent" observations are there if we compare
  - One school with treatment T and one school without treatment T?
- Subjective judgements are involved
  - Random assignments – assessing contamination
  - Regression-discontinuity – functional form of connection A-Y
  - Comparative change – number of independent observations

Internal validity

- Usually threats to internal validity come from the selection process
- There is little one can do at the analysis stage except in some cases to
  - Model the selection process

SKEIVE UTVAL OG MANGLANDE DATA

- I alle ikkje-tilfeldige utval er det eit potensiale for feil i konklusjonane på grunn av skeive utval
- Datainnsamlingsprosedyrer og manglande svar kan gi opphav til trunkerte, selekterte eller sensurerte utval
  - eks: "missing" på avhengig variabel gir eit selektert utval basert på Z: gir svar eller ikkje

Trunkering

- Ein variabel, X, vert kalla trunkert dersom vi for X<c eller for X>c ikkje veit meir enn at X<c eller X>c
- Dette kan omtalast som venstre eller høgretrunkering
- Vi kan også ha multippel trunkering, t.d. samtidig venstre og høgretrunkering

Sensurerte og selekterte utval

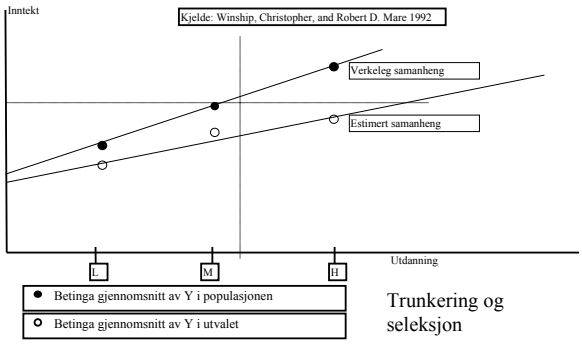
- Sensurerte utval (eksplisitt seleksjon på Y)
  - Y er ukjent når Y har verdiar over eller under c
  - x er kjent for heile utvalet
- Selekterte utval (usystematisk seleksjon)
  - Y er ukjent dersom t.d. Z=1 og kjent når Z=0
  - x er kjent for heile utvalet

Selektert eller sensurert utval ?

- Generelt er dette eit spørsmål om tolking og teoretisk fornuft
  - Når manglande observasjon av Y skuldast målemetode eller data innsamling er utvalet sensurert
  - Når manglande observasjon av Y skuldast atferd hos individa er utvalet selektert

Trunkerte utval og seleksjon på uavhengig variabel

- Trunkert utval (eksplisitt seleksjon på Y)
  - Y er ukjent når Y har verdiar over eller under c
  - x er kjent når Y er kjent
- Seleksjon på uavhengig variabel
  - Y er kjent når x har verdiar anten over eller under c
  - x er kjent når Y er kjent



Trunkering og seleksjon

Kva er problematisk?

- Seleksjon på uavhengig variabel er uproblematisk
- Trunkerte, selekterte og sensurerte utval fører til at restleddet er korrelert med dei uavhengige variablane. Både ekstern og intern validitet er kompromitert.

Kva kan gjerast?

- Ein bør i analysen ta utgangspunkt i det foregåande og bruke modellar som korrigerer for bias dersom ein ikkje kan argumentere for at problemet er lite.
- Løysinga er
  - 1) lage ein modell som predikerer seleksjonen
  - 2) bruke dette i ein modell som predikerer y gitt at personen er selektert

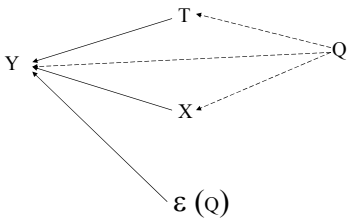
Best practice

- Select cases for treatment centrally. Self selection is the worst threat to internal validity
- All known or suspected factors affecting selection needs to be avoided, controlled or measured
- Minimize unexplained variance by using pre-tests or good causal predictors
- Maximise number of observations (n)

Formalisation

- Design
  - (A)/C: X_{1E} T Y_E
  - A / C: $X_{1C} \rightarrow Y_C$
- Selection Q-bias: Unmeasured cause of Y not captured by the pretest (P)
- Regression model
- $Y_i = \alpha + \beta_X X_i + \beta_T T_i + ... + \varepsilon_i$

Selection Q bias



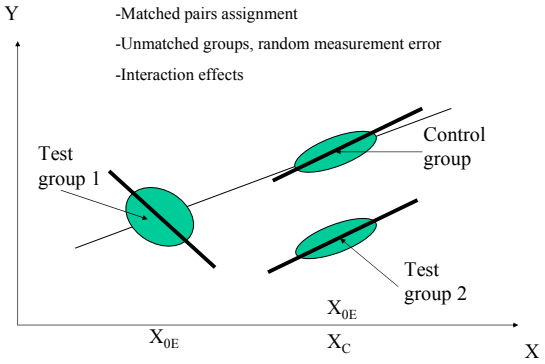
•$Y_i = \alpha + \beta_X X_i + \beta_T T_i + ... + \varepsilon_i (Q_i)$

Selection Q bias

-Matched pairs assignment

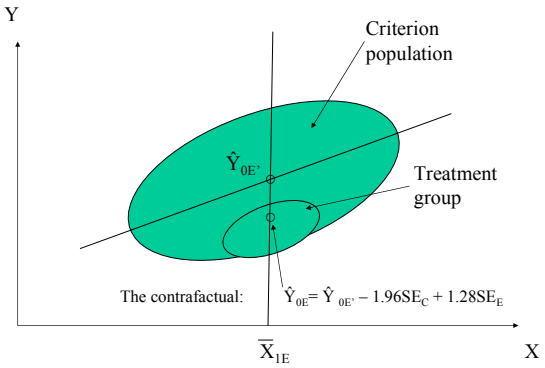
-Unmatched groups, random measurement error

-Interaction effects



The Criterion Population Design

- The treatment group is embedded in a population called the criterion population. Or: the population is designed to contain the treatment group.
- A pretest is available for all subjects both treatment and criterion population



Estimating treatment effect

- In this design the impact of treatment can be estimated in an ordinary regression model

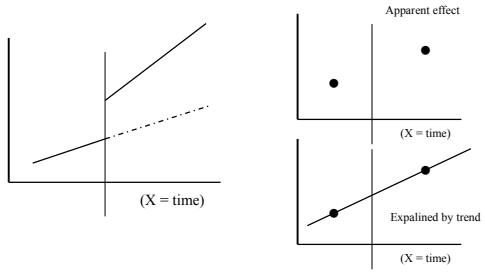
$$\mathbf{Y_i = \alpha + \beta_X X_i + \beta_T T_i + \varepsilon_i}$$

- Where T=1 if treatment is given and 0 otherwise, X is pretest score
- Selection bias may be tested by means of the distribution of ε_i
- Interaction effects and other control variables may be introduced in ordinary fashion
- SE_{β_T} will give confidence intervals for β_T

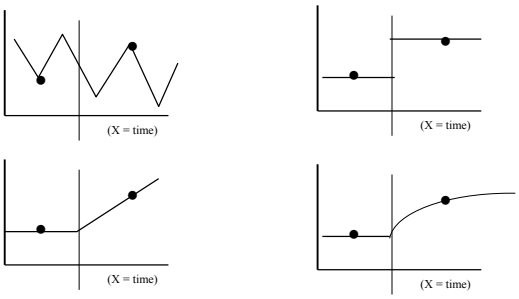
Time series design

- This is the same design as the before – after design or the comparative change design except that there are a series of pretests and posttests
- Analyse by
  - Visual inspection
  - OLS regression
  - ARIMA models (Autoregressive integrated moving average models)

Visual inspection



Explaining time series: transfer functions



Regression analysis

- Interactions and curvelinear elements can be included to mimic transfer functions
- Problems:
  - Autocorrelation makes estimates problematic: tests based on variances are biased
  - Clustering of treatment makes causal interpretations problematic

ARIMA models

- All regressions of post-test on pre-test are autoregressive
- Usd to remove the ”noise” moving averages are used
- No ”causal” variables are included, just the time series of the test variable

Example of autoregressive model (not ARIMA)

Dependent Variable: **Summer 1981 Water Use**	Unstandardized Coefficients			
	B	Std. Error	t	Sig.
(Constant)	242,220	206,864	1,171	,242
Income in Thousands	20,967	3,464	6,053	,000
Summer 1980 Water Use	,492	,026	18,671	,000
Education in Years	-41,866	13,220	-3,167	,002
head of house retired?	189,184	95,021	1,991	,047
# of People Resident 1981	248,197	28,725	8,641	,000
Increase in # of People	96,454	80,519	1,198	,232

Comparative time series

- Comparisons based on units other than individuals are nearly always theoretically problematic (see Fig 9.5 in Mohr)
- But probably no worse than most studies in comparative studies where municipalities or states or occupations or organisations are the units studied
- Theoretically sound models is the best tool!

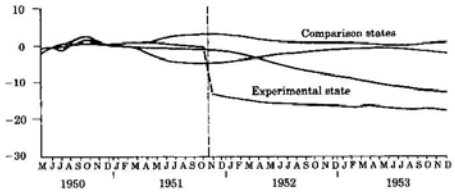


Figure 9.5. Effect of Introducing a Law in the Experimental State Requiring Repayment of Welfare Costs From the Deceased Recipient's Estate on the Old-Age Assistance Caseloads

SOURCE: Baldus [1973, p. 204]. Copyright 1973 by the Board of Trustees of the Leland Stanford Junior University. Reprinted by permission.
NOTE: Monthly data have all values expressed as a percentage of the caseload 18 months prior to the change of the law.

Some footnotes on time series and autoregression

- Based on Hamilton 1992

- Autocorrelation causes biased estimates of error variances (similar to heteroscedasticity)
- Can often be countered by a better causal model (correct model specification)

Autokorrelasjon (1)

- Korrelasjon mellom variabelverdiar på same variabel over ulike case

  (t.d. mellom ε_i og ε_{i-1})
- Autokorrelasjon gir større varians og skeive estimat av standardfeil slik som heteroskedastisitet
- Når vi har enkelt tilfeldig utval frå ein populasjon, er autokorrelasjon usannsynleg

Autokorrelasjon (2)

- Autokorrelasjon kjem frå feilspesifikasjon av modellen
- Ein finn det typisk i tidsseriar og ved geografisk ordna case
- Testar (t.d. Durbin-Watson) er basert på sorteringsrekkefølgja av casa. Derfor:
- Ei hypotese om autokorrelasjon må spesifisere korleis casa skal sorterast

Durbin-Watson testen (1)

$$d = \frac{\sum_{i=2}^{n} \left(e_i - e_{i-1}\right)^2}{\sum_{i=1}^{n} e_i^2}$$

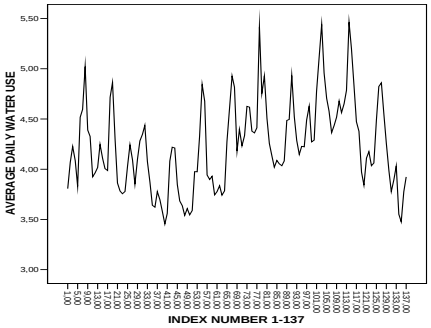
Bør ikkje nyttast for autoregressive modellar, dvs. modellar der y-variabelen også finst som forklaringsvariabel (x-variabel) jfr tabell 3.2

Durbin-Watson testen (2)

- Samplingfordelinga til d-observatoren er kjent og tabellert som $\mathbf{d_L}$ og $\mathbf{d_U}$ (tabell A4.4 i Hamilton), talet av fridomsgrader baserer seg på n og K-1
- Testregel:
  - Forkast dersom $d<d_L$
  - Forkast ikkje dersom $d>d_U$
  - Dersom $d_L < d < d_U$ kan det ikkje konkluderast
- d=2 tyder ukorrelerte residualar
- Positiv autokorrelasjon gir d<2
- Negativ autokorrelasjon gir d>2

Dagleg vassforbruk, gjennomsnitt pr månad

Eksempel:



Vanleg OLS-regresjon der caset er månad

Dependent Variable: AVERAGE DAILY WATER USE	Unstandardized Coefficients		t	Sig.
	B	Std. Error		
(Constant)	3,828	,101	38,035	,000
AVERAGE MONTHLY TEMPERATURE	,013	,002	7,574	,000
PRECIPITATION IN INCHES	-,047	,021	-2,234	,027
CONSERVATION CAMPAIGN DUMMY	-,247	,113	-2,176	,031

Predictors: (Constant), CONSERVATION CAMPAIGN DUMMY, AVERAGE MONTHLY TEMPERATURE, PRECIPITATION IN INCHES

Test for autokorrelasjon

Dependent Variable: AVERAGE DAILY WATER USE	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,572(a)	,327	,312	,36045	**,535**

Predictors: (Constant), CONSERVATION CAMPAIGN DUMMY, AVERAGE MONTHLY TEMPERATURE, PRECIPITATION IN INCHES

N = 137, K-1 = 3

Finn grensene for å forkaste / akseptere nullhypotesa om ingen autorkorrelasjon med nivå 0,05

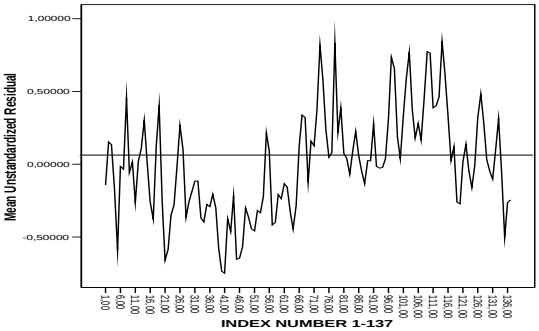
Hjelpemiddel: Sjekk opp tabell A4.4 i Hamilton side 355

Autokorrelasjonskoeffisient

m-te ordens autokorrelasjonskoeffisient

$$r_m = \frac{\sum_{t=1}^{T-m}\left(e_t - \bar{e}\right)\left(e_{t+m} - \bar{e}\right)}{\sum_{t=1}^{T}\left(e_t - \bar{e}\right)^2}$$

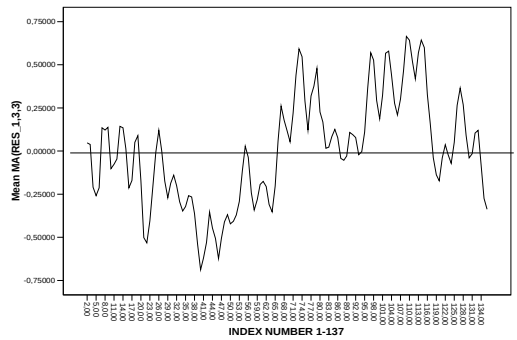
Residual ”Dagleg vassforbruk”, månad



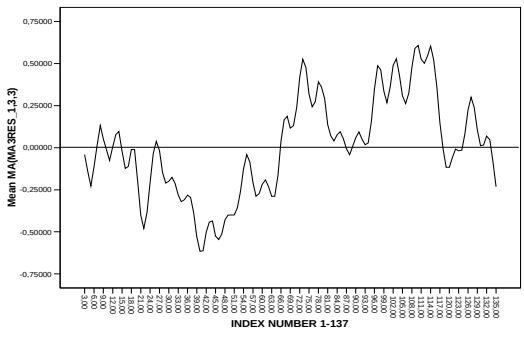
Glatting med 3 punkt

- Glidande gjennomsnitt $e_t^* = \frac{e_{t-1} + e_t + e_{t+1}}{3}$
- "Hanning" $e_t^* = \frac{e_{t-1}}{4} + \frac{e_t}{2} + \frac{e_{t+1}}{4}$
- Glidande median $e_t^* = median\{e_{t-1}, e_t, e_{t+1}\}$

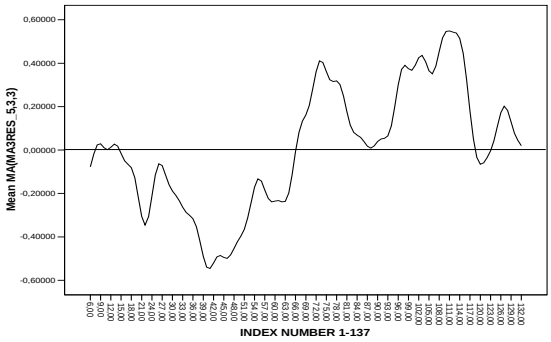
Residual, glatta ein gong



Residual, glatta to gonger



Residual, glatta fem gonger



Konsekvensar av autokorrelasjon

- Hypotesetestar og konfidensintervall er upålitelege. Regresjon kan likevel gi ein god beskrivelse av utvalet. Parametrane er forventningsrette
- Spesialprogram kan estimere standardfeil konsistent
- Ta inn i analysen variablar som påverkar ”hosliggjande” case
- Ta i bruk teknikkar frå tidsserieanalyse (t.d.: analyser differansen mellom to tidspunkt) (Δy)