SOS3003

Anvendt statistisk dataanalyse i samfunnsvitenskap

Forelesingsnotat, vår 2003

Erling Berge
Institutt for sosiologi og statsvitenskap
NTNU

Forelesing II

- Bivariat regresjon II
  - Hamilton Kap 2 s51-59
- Multivariat regresjon I
  - Hamilton Kap 3 s65-72

Repetisjon:
Bivariat Regresjon: Modell for utval

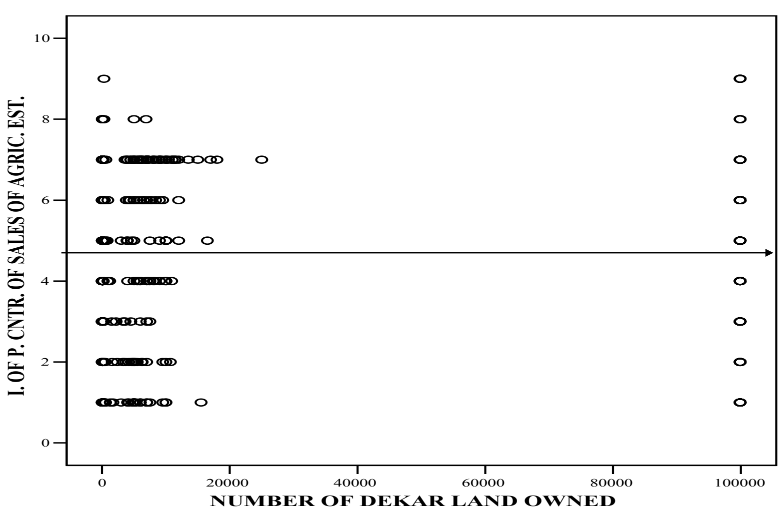
- $Y_i = b_0 + b_1 x_{1i} + e_i$
- i=1,...,n n = # case i utvalet

Eksempel frå første forelesning med

- Y = IMPORTANCE OF PUBLIC CONTROL OF SALES OF AGRIC. ESTATES
- **X =** NUMBER OF DEKAR LAND OWNED

Kva var galt i eksempelet?

Kva er galt? Spredningsdiagram med regresjonslinje



Generelt: Kva kan skape problem?

- Utelatte variablar
- Ikkje-lineære samanhengar
- Ikkje-konstant varians på feilen (heteroskedastisitet)
- Korrelasjon mellom feila (autokorrelasjon)
- Ikkje normale feil
- Innflytelsesrike case

Ikkje-normale feil: Meir om variabelfordelingar

- Regresjon har **IKKJE føresetnader om fordelinga** til variablane
- MEN for å teste hypoteser må vi ha eit **normalfordelt feilledd**
- DERSOM **modellen er korrekt** og n er stor sikrar sentralgrenseteoremet at feilleddet er tilnærma normalfordelt
- MEN som regel er modellen feil eller ufullstendig. Derfor må vi **teste om residualen faktisk er normalfordelt**.

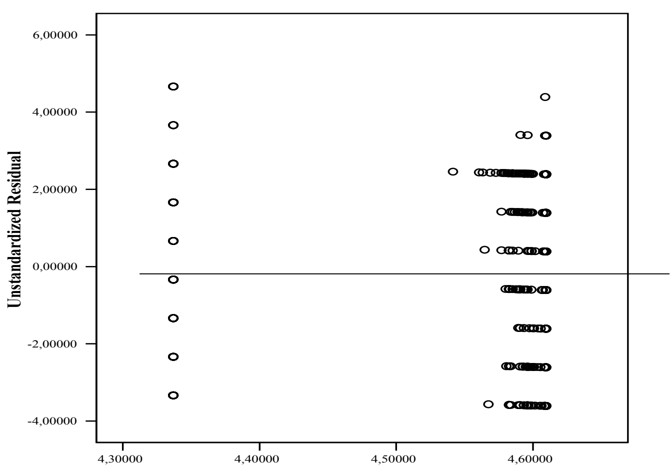
Residualanalyse:

Viktigaste innfallsporten til problema i regresjonsanalysen

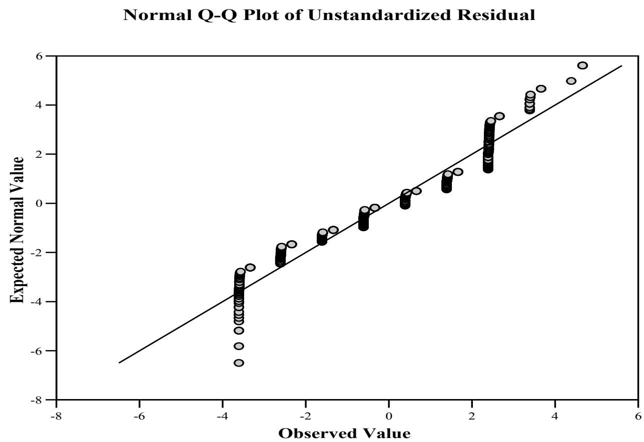
Hjelpemiddel:

- Scatterplott
- Plott av residual mot predikert verdi
- Histogram
- Boksplott
- Symmetriplott
- Kvantil-normalplott

Kva er galt? (1) residual-predikert verdi plott



Kva er galt? (2) residual normal-kvantilplott



Potenstransformasjonar

kan løyse problem i samband med

- kurve-linearitet i modellen
- utliggarar
- case med stor innverknad
- ikkje-konstant varians hos feilleddet
- ikkje-normale feilledd

Transformasjon er mao eit generelt verkemiddel

Potenstransformasjonar (jfr H:17-22)

Y* : les "transformert Y"
(transformasjon fra Y til Y*)

- $Y^* = Y^q$ q>0
- $Y^* = \ln[Y]$ q=0
- $Y^* = -[Y^q]$ q<0

Invers trasformasjon
(transformasjon fra Y* til Y)

- $Y = [Y^*]^{1/q}$ q>0
- $Y = \exp[Y^*]$ q=0
- $Y = [-Y^*]^{1/q}$ q<0

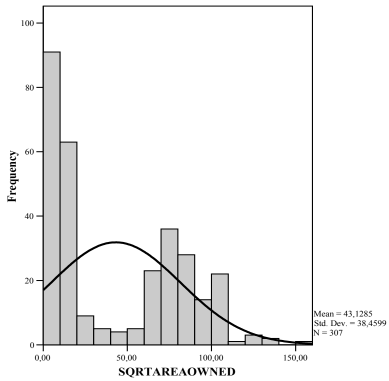
Potenstransformasjonar: konsekvensar

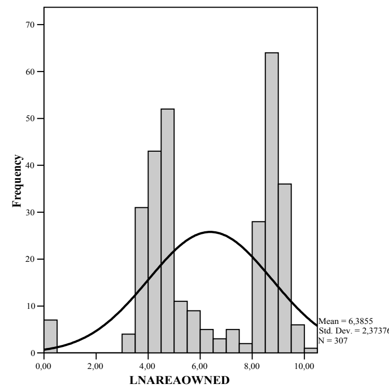
- $X^* = X^q$
  - q > 1 **aukar tyngda** til øvre hale relativt til nedre
  - q = 1 gir identitet
  - q < 1 **reduserer tyngda** til øvre hale relativt til nedre

- Dersom $Y^* = \ln(Y)$ vil regresjonskoeffesienten for ein intervallskala X variabel kunne tolkast som % endring i Y for ei einings endring i X

Potenstransformert

X = NUMBER OF DEKAR LAND OWNED

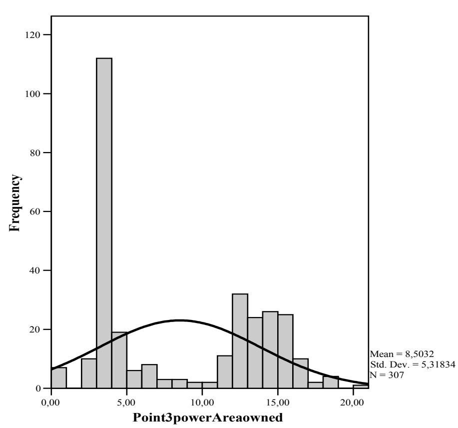




SQRT=kvadratrota av areaowned – LN=naturleg logaritme av (areaowned+1)

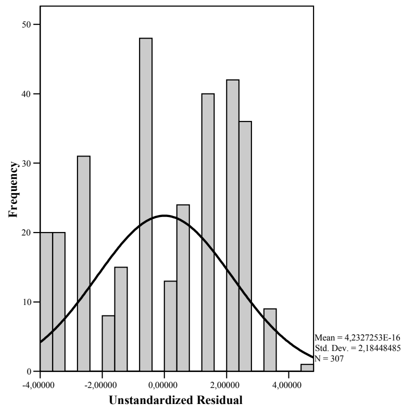
Potenstransformert

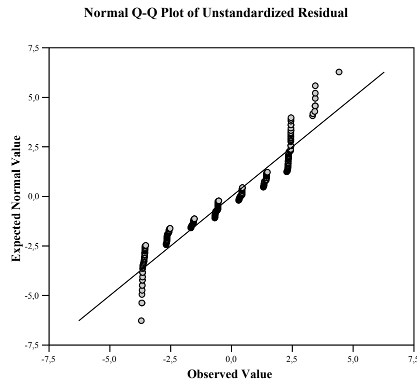
X = NUMBER OF DEKAR LAND OWNED



Poin3power = 0,3 potensen av areaowned

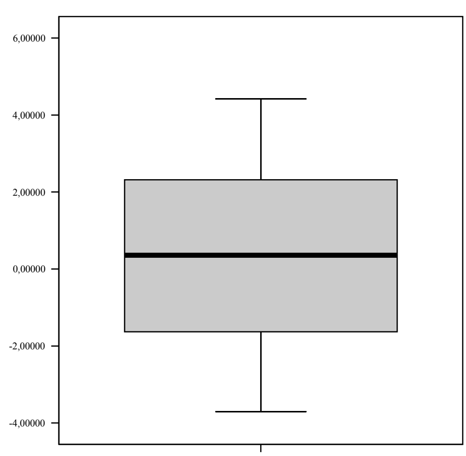
Hjelper det med potenstransformasjon?



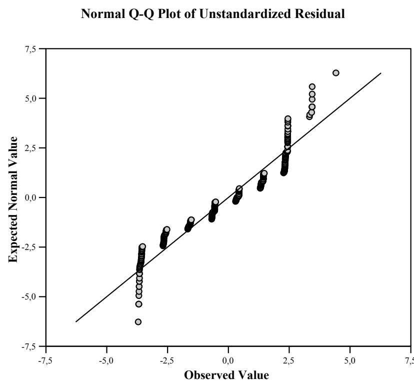


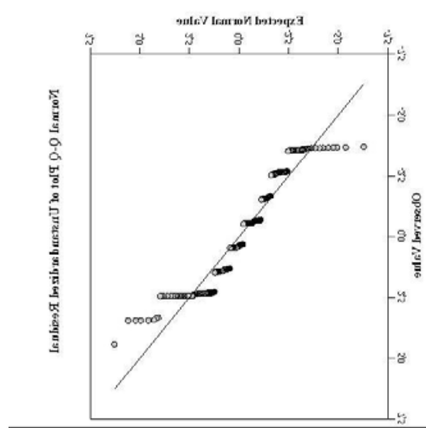
0.3potens-transformasjon gir lette halar og ingen utliggarar

Boksplott av residualen viser tilnærma symmetri utan utliggarar



SPSSutskrift vs boka (jfr. s16)





Lesing av utskrifter fra SPSS (1)

Descriptive Statistics	Mean	Std. Deviation[1]	N[2]
I. OF P. CNTR. OF SALES OF AGRIC. EST.	4.61	2.185	307
Point3powerAreaowned	8.5032	5.31834	307

Model	R	R Square[3]	Adjusted R Square[4]	Std. Error of the Estimate[5]	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.024(a)	.001	-.003	2.188	.001	.182	1	305	.670

a Predictors: (Constant), Point3powerAreaowned
b Dependent Variable: I. OF P. CNTR. OF SALES OF AGRIC. EST.

Fotnoter til føregåande tabell

[1] Standard-avviket til gjennomsnittet (mean)
[2] Talet av case brukt i analysen
[3] Determinasjonskoeffesienten
[4] Den justerte determinasjonskoeffesienten, sjå Hamilton side 41
[5] Standard-avviket til residualen s_e = SQRT (RSS/(n-K)), der SQRT = kvadratrota av (*)

Lesing av utskrifter fra SPSS (2)

Model		Sum of Squares	df	Mean Square	F[1]	Sig.[2]
1	Regression	.870	1	.870	.182	.670(a)
	Residual	1460.224	305	4.788		
	Total	1461.094	306			

- Kvadratsummar: TSS = ESS + RSS
- RSS = $\Sigma_i(e_i)^2 = \Sigma_i(Y_i - \hat{Y}_i)^2$: sum kvadrert (avstand observert – estimert verdi)
- Mean Square = RSS / df   For RSS har vi at df=n-K

K er lik talet på parametrar som modellen estimerer (dvs b_0 og b_1)

Her er n=307 og K=2, dvs. Df = 305

Fotnoter til føregåande tabell

[1] F-observatoren for nullhypotesa $beta_1 = 0$
(sjå Hamilton side 45)

[2] p-verdien for F-observatoren: dvs sannsynet for å finne ein så stor eller større F-verdi gitt at nullypotesa er rett

Kvadratsummar

TSS = ESS + RSS

RSS = $\Sigma_i(e_i)^2 = \Sigma_i(Y_i - \hat{Y}_i)^2$ avstand observert – estimert verdi

ESS = $\Sigma_i(\hat{Y}_i - \tilde{Y})^2$ avstand estimert verdi – gjennomsnitt

TSS = $\Sigma_i(Y_i - \tilde{Y})^2$ avstand observert verdi – gjennomsnitt

Lesing av utskrifter fra SPSS (3)

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B[1]	Std. Error[2]	Beta[3]	t[4]	Sig.[5]	Lower Bound	Upper Bound
1	(Constant)	4.524	.236		19.187	.000	4.060	4.988
	Point3-powerArea-owned	.010	.024	.024	.426	.670	-.036	.056

Fotnoter til føregåande tabell

[1] Estimat av regresjonskoeffesientane b_0 og b_1

[2] Standardavviket (standardfeilen) til parameterstimata b_0 og b_1

[3] Dei standardiserte regresjonskoeffesientane: $b_1^{st} = b_1*(s_x/s_y)$ sjå Hamilton side 38-40

[4] t-observatoren for nullhypotesa beta1 = 0 (sjå Hamilton side 44)

[5] p-verdien for t-observatoren: dvs sannsynet for å finne ein så stor eller større t-verdi gitt at nullypotesa er rett

Kurvelienær regresjon

- I eksempelet ovanfor brukte vi variabelen ”Point3powerAreaowned”, dvs 0.3 potensen av antall dekar areal ein eig. Dvs.:
- Point3powerAreaowned = (NUMBER OF DEKAR LAND OWNED)$^{0.3}$

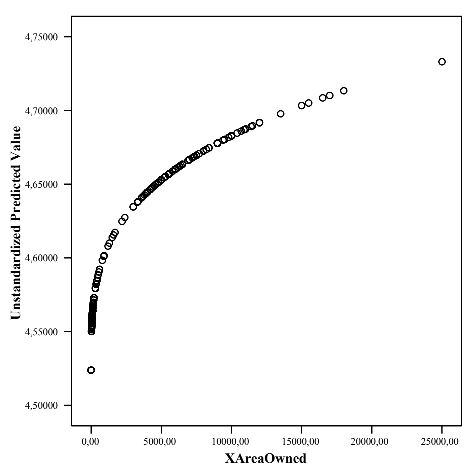
Modellen vi har estimert er altså

$y_i = b_0 + b_1 x_i + e_i$

$y_i = b_0 + b_1$ Point3powerAreaowned$_i + e_i$

$\hat{y}_i = 4.524 + 0.010*$(NUMBER OF DEKAR LAND OWNED)$^{0.3}$

Bruk av potenstransformerte variablar betyr at regresjonen blir kurvelineær



Oppsummering

- I bivariat regresjon kan ein seie at OLS-metoden freisar finne den beste LINJA eller KURVA som passar til eit to-dimensjonalt spreiingsmønster
- Scatter-plott og residualanalyse er hjelpemiddel for å diagnistisere problem i regresjonen
- Transformsjonar er eit generelt hjelpemiddel mot fleire typar problem, som t.d.:
  - Kurvelinearitet
  - Heteroskedastisitet
  - Ikkje-normalitet
  - Case med stor innverknad
- Regresjon med transformerte variablar er alltid kurvelineær. Vi tolkar resultatet letast ved hjelp av grafar

Multippel regresjon: modell (1)

- Sett K = talet på parametrar i modellen (dvs. K-1 er talet på variablar). Da kan (populasjons) modellen skrivast
- $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_{K-1} x_{i,K-1} + \varepsilon_i$

Multippel regresjon: modell (2)

- Dette kan skrivast

  $y_i = E[y_i] + \varepsilon_i$,

  dette tyder at
- $E[y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_{K-1} x_{i,K-1}$

  $E[y_i]$ les vi som forventa verdi av y_i
- Målet med multippel regresjon er å finne nettoeffekten av ein variabel, kontrollert for variasjonen i alle dei andre

Multippel regresjon: modell (3)

- Vi finn OLS estimata av modellen som dei b-verdiane i

  $\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + ... + b_{K-1} x_{i,K-1}$

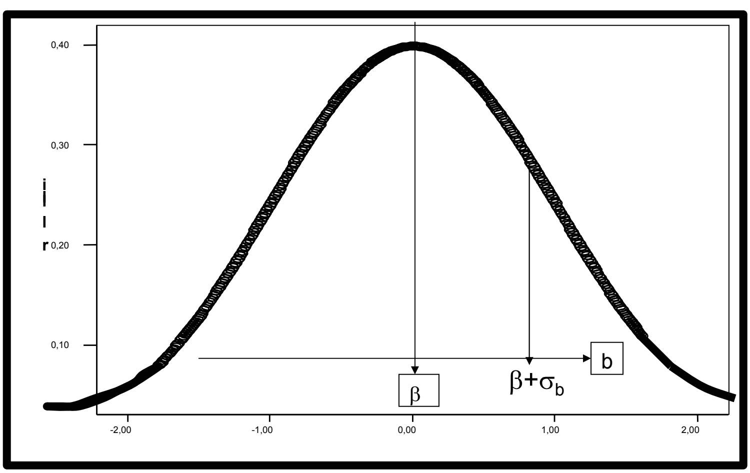
  ($\hat{y}_i$ les vi som estimert eller ”predikert” verdi av y_i)

  som minimerer kvadratsummen av residualane

  $RSS = \Sigma_i(y_i - \hat{y}_i)^2 = \Sigma_i e_i^2$

Meir om hypotesetesting

- Frå ein populasjon kan det trekkjast mange utval
- I kvart nytt utval vil vi kunne estimere nye regresjonsparametrar
- Lagar vi eit histogram over ulike estimerte verdiar av t.d. β_1 vil vi sjå at b_1 har ei fordeling (ei samplingfordeling)
- Ulike parametrar og observatorar har ulike samplingfordelingar
- Regresjonsparametrane (b'ane) er t-fordelt



Sampling fordeling for regresjonsprameteren b: **E[b]=** β

Om partielle effektar (1)

- Eit eksempel med 2 variablar
- Dersom vi estimerer ein modell

  $y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + e_i$

  er det i prinsippet 3 ulike korrelasjonar som er involvert:
  - Mellom y og x_1
  - Mellom y og x_2
  - Mellom x_1 og x_2

Om partielle effektar (2)

- Dette kan teoretisk gi opphav til 3 ulike bivariate regresjonar der vi held den tredje variabelen konstant (t.d. for ”gitt x” skrive ”Ix”)

(1) $y = a_{yIx1} + b_{yIx1}x_1 + e_{yIx1}$

(2) $y = a_{yIx2} + b_{yIx2}x_2 + e_{yIx2}$

(3) $x_1 = a_{x1Ix2} + b_{x1Ix2}x_2 + e_{x1Ix2}$

indeksen ”yIx1” les vi ”regresjonen av y på x1”

- Likningane (2) og (3) kan vi skrive om

Om partielle effektar (3)

(2) $e_{y|x2} = y - (a_{y|x2} + b_{y|x2}x_2)$

(3) $e_{x1|x2} = x_1 - (a_{x1|x2} + b_{x1|x2}x_2)$

Vi ser da at vi så å seie "fjernar" effekten av x_2 frå y og frå x_1

Vi ser også at $e_{y|x2}$ og $e_{x1|x2}$ vert nye variablar der effekten av x_2 er fjerna

Om partielle effektar (4)

- Dersom vi no lagar ein ny regresjon

  $\hat{e}_{y|x2} = a + b\, e_{x1|x2}$

  finn vi at

  a = 0

  b = b_1 frå den bivariate regresjonen

  $y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + e_i$

- b_1 er altså effekten av x_1 på y etter at vi har "fjerna" effekten av x_2

Eksperiment og partielle effektar

- I eksperiment granskar ein kausalsamband mellom to variable med kontroll for alle mogelege andre kausale faktorar
- Multippel regresjon er ei form for etterlikning av eksperimentet – ei nest beste løysing - og ligg nært opp til det som heiter kvasi-eksperimentelt forskingsdesign

KAUSALANALYSE

- Eksperiment
  - randomisering av påverknad (“behandling”) gir presise kausale konklusjonar om verknader (“respons”) ved signifikant skilnad i gjennomsnitt
  - kan vere umogeleg på grunn av
    - praktisk tilhøve
    - økonomiske skrankar
    - etiske vurderingar
- Kvasi-eksperiment der eksperiment er umogeleg
  - t.d. regresjonsanalyse

Eksperimetet plasserer ”case” tilfeldig i ei av to grupper:

- **BEHANDLING (T)**
  med observasjon
  – FØR behandling
  – ETTER behandling
- **KONTROLL (C)**
  med observasjon
  – FØR ”ikkje-behandling”
  – ETTER ”ikkje-behandling”

Modell av kausaleffektar Ref.:

- Studiar av observasjonsdata brukar omgrep fra eksperimentell design
- “Påverknad/ Behandling”, “Stimulus” (Treatment/ Stimulus)
- “Effekt”, “Utfall” (Effect/ Outcome)

Ref.:

Winship, Chrisopher, and Stephen L. Morgan 1999 “The Estimation of Causal Effects from Observational Data”, Annual Review of Sociology Vol 25: 659-707

Modell av kausaleffektar:

Den "Kontrafaktiske" hypotesa for studiet av kausalitet

- Individet "i" kan i utgangspunktet tenkjast "selektert" til ei av to grupper
  - behandlingsgruppa, T, eller kontrollgruppa, C.
- Behandlinga, t (treatment), så vel som ikkje-behandling, c, kan i utgangspunktet tenkjast gitt til individ både i T- og C-gruppa
- Faktisk vil vi berre kunne observere t i T-gruppa og c i C-gruppa

Modell av kausaleffektar:

Den "Kontrafaktiske" hypotesa

- For kvart individ "i" kan ein tenkje seg fire mogelege utfall
  - $\mathbf{Y_i(c,C)}$ eller $Y_i(t,C)$; ved plassering i kontrollgruppe
  - $Y_i(c,T)$ eller $\mathbf{Y_i(t,T)}$; ved plassering i behandlingsgruppa
- Berre Y_i(c, gitt "i" er med i C) eller Y_i(t, gitt "i" er med i T) kan observerast for eit gitt individ

Modell av kausaleffektar:
Den “Kontrafaktiske” hypotesa

Litt meir formelt kan ein skrive dei mogelege utfalla for person i:

	Behandling: t	Ikkje beh.: c
T-gruppa	$\mathbf{Y^t_i \in T}$	$Y^c_i \in T$
C-gruppa	$Y^t_i \in C$	$\mathbf{Y^c_i \in C}$

Modell av kausaleffektar:
Den “Kontrafaktiske” hypotesa

- Kausaleffekten for individ i er da
- $\delta_i = Y_i\,(t) - Y_i\,(c)$
- Berre ein av desse to storleikane kan observerast for eit gitt individ

  Derfor ”Den kontrafaktiske hypotesa”

Modell av kausaleffektar:

Den "Kontrafaktiske" hypotesa

- Vi kan til dømes observere
  Y_i (c |i$\in$C), men ikkje Y_i (t |i$\in$C)
- Problemet kan seiast å vere manglande data
- I staden for individeffektar vil ein estimere gjennomsnittseffektar i heile populasjonen

Modell av kausaleffektar:

- Gjennomsnittseffektar lar seg estimere, men som regel berre med store vanskar
- Ein føresetnad er at effekten av påverknad vil vere den samme for eit gitt individ uansett kva gruppe individet er plassert i
- Dette er imidlertid ikkje sjølvsagt

Modell av kausaleffektar:

Den “Kontrafaktiske” hypotesa antar:

- at endring av behandlingsgruppe for eitt individ ikkje verkar inn på utfallet for andre individ (fravær av interaksjon)
- at behandlinga, “påverknaden”, faktisk er manipulerbar (t.d.: kjønn er ikkje manipulerbar)

Modell av kausaleffektar:

- Ein av vanskane er at i eit utval vil den prosessen som plasserer personen ”i” i kontroll- eller behandlings-gruppa kunne verke inn på det estimerte gjennomsnittsutfallet (seleksjonsproblemet)
- I somme høve er imidlertid den interessante størrelsen gjennomsnitts-effekten for dei som faktisk får påverknaden

Modell av kausaleffektar:

- Det kan visast at det er to kjelder til feil (bias) i estimata av gjennomnsitseffekten
  - ein eksisterande skilnad mellom C- og T-gruppene
  - behandlinga verkar i prinsippet ulikt for dei som er i T-gruppa samanlikna med dei som er i C -gruppa
- For å handtere dette må vi utvikle modellar for korleis folk hamnar i C- og T-gruppene

Modell av kausaleffektar:

- Ein generell klasse metodar som kan nyttast til å estimere kausaleffektar er regresjonsmodellane

- Dei vil kunne "kontrollere" for observerbare skilnader mellom T- og C-gruppene, men ikkje for ulik respons på behandling