

SOS3003

Anvendt statistisk dataanalyse i samfunnsvitenskap

Forelesingsnotat, vår 2003

Erling Berge
Institutt for sosiologi og statsvitenskap
NTNU

PENSUM SOS 3003

- 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,
- Allison, Paul D. 2002 "Missing Data" Sage University Paper: QASS 136, London, Sage,

Forelesing I

- Opplegg
 - Føresetnader: SOS 1002 eller tilsvarande
 - Målsetting: lese faglitteratur
 - Pensum
 - Semesteroppgåve: del av eksamen
- Variabelfordelingar. Hamilton Kap 1 s1-23
- Bivariat regresjon I Hamilton Kap 2 s29-50

Mål for kurset

- Kunne lese faglitteratur som drøftar ”kvantitative” data kritisk
 - Vi skal kjenne fallgruvane
- Gjennomføre enkle analysar av samvariasjon i ”kvantitative” data
 - Vi skal demonstrere at vi kjenner fallgruvane

Kort repetisjon av grunnleggjande omgrep

- Årsak
- Modell
- Populasjon
- Utval
- Variabel: målenivå
- Variabel: sentraltendens
- Variabel: spredning

Data-analyse

- Deskriptive bruk av data
 - Utvikling av klassifikasjonar
- Analytisk bruk av data
 - Beskrive fenomen som ikkje kan observerast direkte (inferens)
 - Årsaksamband mellom direkte eller indirekte observerbare fenomen (teori eller modellutvikling)

KAUSALANALYSE: frå samvariasjon til årsak

- frå daglegspråk til teori
 - fantasi og intuisjon, etablert fagleg tradisjon
- frå teori til modell
 - operasjonalisering
- frå observasjon til generalisering
 - kausalanalyse

3 SENTRALE SKILNADER

<u>Observert</u>		<u>Interesse</u>
TEORI/ MODEL	-	RØYNDOM
UTVAL	-	POPULASJON
SAMVARIASJON	-	ÅRSAK

På eine sida har vi det vi faktisk kan sette på papiret,
på andre sida det vi gjerne vil seie noko om.

Sentrale feilkjelder

- Feil i teori/ modell
 - Modellspesifikasjonskravet
- Feil i utval
 - Seleksjonsproblemet
- Måleproblem
 - Frafall og målefeil
 - Validitet og reliabilitet
- Multiple komparasjonar
 - Utvalsspesifikke konklusjonar

Frå populasjon til utval

- POPULASJON (ALLE EININGAR)
- -enkel tilfeldig trekking-
- UTVAL (UTVALTE EININGAR)

Eining og variabel

- Eininga, beraren av data, er kontekstuelst definert
 - SUPER - EINING: t.d. lokalsamfunnet
 - EINING: t.d. hushald
 - SUB - EINING: t.d. person
- Variabel: empirisk omgrep brukt til å karakterisere undersøkte einingar. Kvar eining vert karakterisert ved å tilordne den ein variabelverdi.

Datamatrise og målenivå

- Matrise definert av Einingar * Variablar
 - Tabell over eigenskapane til alle undersøkte einingar ordna slik at alle variabelverdiane kjem i samme rekkefølge for alle einingar.
- Målenivå av ein variabel
 - nominal *klassifisering
 - ordinal *klassifisering og rang
 - intervall *klassifisering, rang og **avstand**
 - forholdstal *klassifisering, rang, **avstand** og nullpunkt

Variabelanalyse

- Beskrivelse
 - Sentraltendens og spredning
 - Fordelingsform
 - Frekvensfordelingar og histogram
- Samanlikning av fordelingar
 - Kvantilplott
 - Boxplott

VARIABEL: SENTRALTENDENS

- GJENNOMSNIITT $\tilde{Y} = (\sum_i Y_i)/n$
- Summen av verdiane på variabelen for alle einingane dividert på talet av einingar
- MEDIAN
- Den verdien i ei ordna fordeling som har halparten av einingane på kvar side
- MODUS
- Den typiske verdien. Den verdien i ei fordeling som har høgast frekvens.

VARIABEL: SPREDNINGSMÅL I

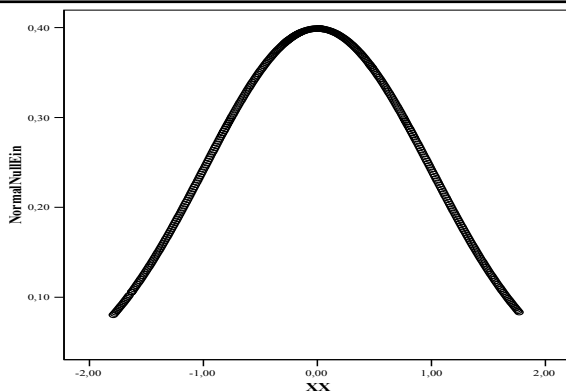
- MODALPROSENTEN
- Prosent av einingane som har verdi lik modus
- VARIASJONSBREDDA
- Differansen mellom høgaste og lågaste verdi i ei ordna fordeling
- KVARTILDIFFERENSEN
- Variasjonsbreidda for dei 50% av einingane som ligg rundt medianen ($Q_3 - Q_1$)
- MAD - Median Absolute Deviation
- Medianen til absoluttverdien til skilnaden mellom median og observert verdi: $MAD(x_i) = \text{median } |x_i - \text{median}(x_i)|$

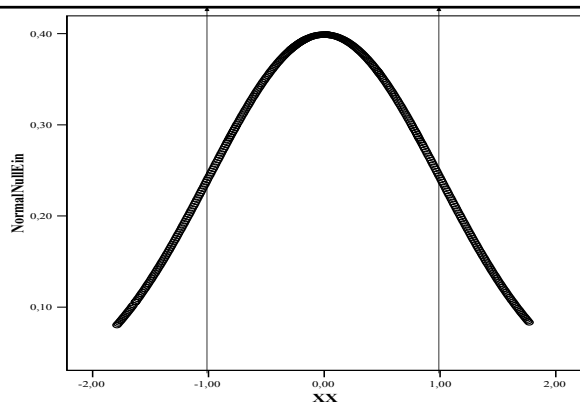
VARIABEL: SPREDNINGSMÅL II

- STANARDAVVIKET
- Kvadratrot av gjennomsnittleg kvadrert avvik frå gjennomsnittet $s_y = \sqrt{(\sum_i (Y_i - \bar{Y})^2)/(n - 1)}$
- GJENNOMSNITSAVVIKET
- Gjennomsnittet av absoluttverdien til avviket frå gjennomsnittet
- VARIANSEN
- Kvadratet av standardavviket: $s_y^2 = (\sum_i (Y_i - \bar{Y})^2)/(n - 1)$

Variabel: fordelingsform I

- Symmetriske fordelingar
- Skjeive fordelingar
 - ”Tunge” og ”lette” halar
- Normalfordlingar
 - Er ikkje ”normale”
 - Er eintydig fastlagt av gjennomsnitt og standardavvik (μ og σ)





Vår 2004

© Erling Berge 2004

19

Skeive fordelingar

- Positivt skeive har $\tilde{Y} > Md$
- Negativt skeive har $\tilde{Y} < Md$
- Symmetriske fordelingar har $\tilde{Y} \approx Md$

Vår 2004

© Erling Berge 2004

20

Symmetriske fordelingar

- Medianen og IQR er resistente mot verknader av ekstreme verdiar. Gjennomsnitt og standardavvik er det ikkje
- I normalfordelinga er $s_y \approx IQR/1.35$
- Dersom vi i ei symmetrisk fordeling finn
 - $s_y > IQR/1.35$ er halane tyngre enn i normalfordelinga
 - $s_y < IQR/1.35$ er halane lettare enn i normalfordelinga
 - $s_y \approx IQR/1.35$ er halane omlag slik som i normalfordelinga

Vår 2004

© Erling Berge 2004

21

Variabel: fordelingsanalyse I

- Boxplott
 - Basert på kvartilverdiane og interkvartilavviket
 - Definerer nærliggjande utliggarar som dei som ligg innanfor intervalla $<Q_1 - 1.5IQR, Q_1>$ og $<Q_3 + 1.5IQR, Q_3>$ og fjerntliggjande utliggarar dei som ligg utanfor grensene $<Q_1 - 1.5IQR, Q_3 + 1.5IQR>$

Variablar: fordelingsform II

- Kvantilar er ei genralisering av kvartilar og percentilar
- Kvantilverdiane er variabelverdiane som svarar til gitte fraksjonar av det samla utvalet eller observasjonamaterialet, t.d.
 - Medianen er 0.5 kvantilen (eller 50% percentilen)
 - Nedre kvartil er 0.25 kvantilen
 - 10% percentilen er 01 kvantilen osv.

Variabel: fordelingsanalyse II

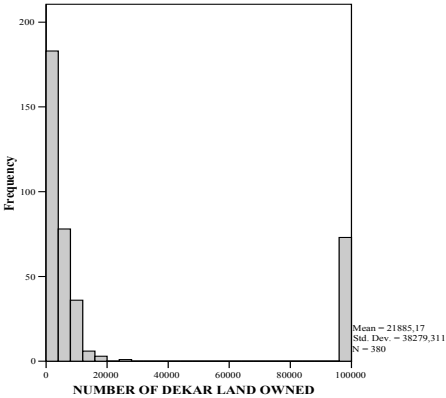
- Kvantilplott
 - Kvantilverdi mot variabelverdi
 - Lorentzkurva er ein spesialvariant av dette (gir oss Gini-indeksen)
- Kvantil-Normalplott
 - Plott av kvantilverdiar på ein variabel mot kvantilverdiane i ei normalfordeling med samme gjennomsnitt og spredning

Eksempel: Frå Randaberg

- Spørreskjema: spørsmål X
- 27 ANTALL DEKAR GRUNN DU eier: _____

NUMBER OF DEKAR LAND OWNED

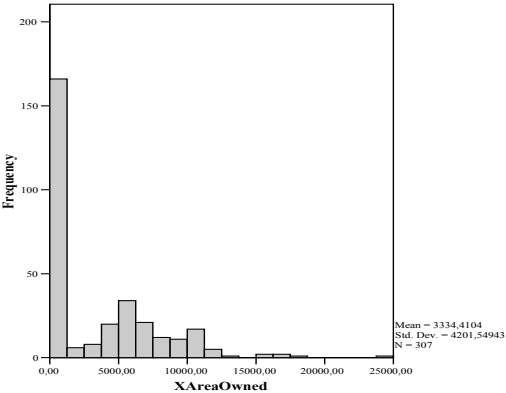
	NUMBER OF DEKAR LAND OWNED	Valid N (listwise)
N	380	380
Minimum	0	
Maximum	99900	
Mean	21885.17	
Std. Deviation	38279.311	

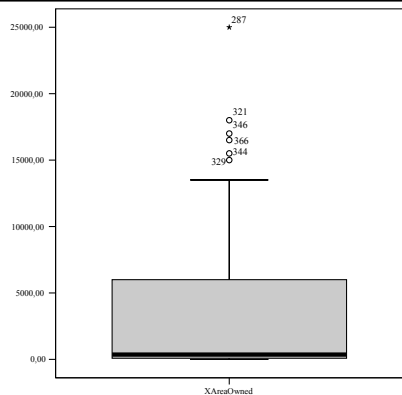


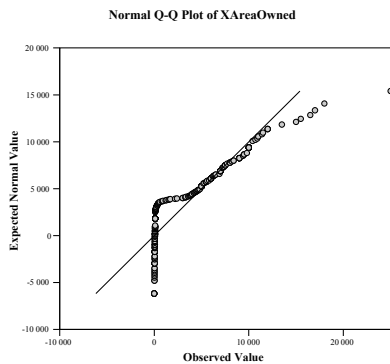
XAreaOwned
(NUMBER OF DEKAR LAND OWNED)

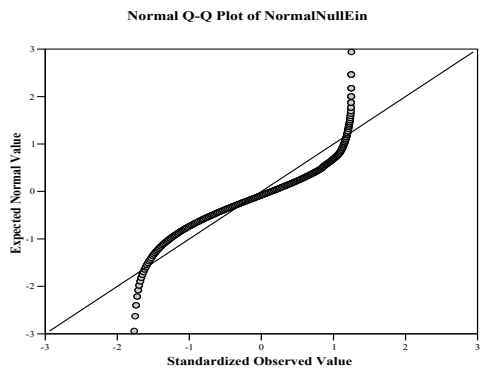
	XAreaOwned	Valid N (listwise)
N	307	307
Minimum	.00	
Maximum	25000.00	
Mean	3334.4104	
Std. Deviation	4201.54943	

		XAreaOwned	Valid N (listwise)
N	Statistic	307	307
Range	Statistic	25000.00	
Minimum	Statistic	.00	
Maximum	Statistic	25000.00	
Sum	Statistic	1023664.00	
Mean	Statistic	3334.4104	
	Std. Error	239.79509	
Std. Deviation	Statistic	4201.54943	
Variance	Statistic	17653017.596	
Skewness	Statistic	1.352	
	Std. Error	.139	
Kurtosis	Statistic	2.194	
	Std. Error	.277	









Spørreskjema: spørsmål Y

- **Hvor viktig er det at myndighetene kontrollerer og regulerer bruken av arealer gjennom for eksempel kontroll av**

1	2	3	4	5	6	7	8
---	---	---	---	---	---	---	---
- **av tomtetildelinger (kommunal formidl.)**

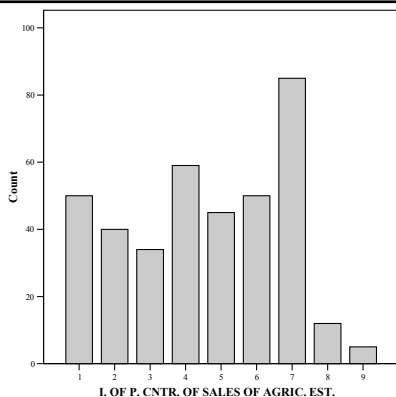
1	2	3	4	5	6	7	8
---	---	---	---	---	---	---	---
- **avkjørsler fra hus til vei**

1	2	3	4	5	6	7	8
---	---	---	---	---	---	---	---
- **kjøp og salg av landbrukseiendommer**

1	2	3	4	5	6	7	8
---	---	---	---	---	---	---	---

Importance of public control of sales of agric. estates

	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 1	50	13.2	13.2	13.2
2	40	10.5	10.5	23.7
3	34	8.9	8.9	32.6
4	59	15.5	15.5	48.2
5	45	11.8	11.8	60.0
6	50	13.2	13.2	73.2
7	85	22.4	22.4	95.5
8	12	3.2	3.2	98.7
9	5	1.3	1.3	100.0
Total	380	100.0	100.0	



Spørreskjema: koding

Ved utfylling: sett ring rundt et tall som synes å gi passelig uttrykk for viktigheten når 1 betyr svært lite viktig og 7 særdeles viktig, eller sett et kryss inne i parentesene () som står bak svaret du velger

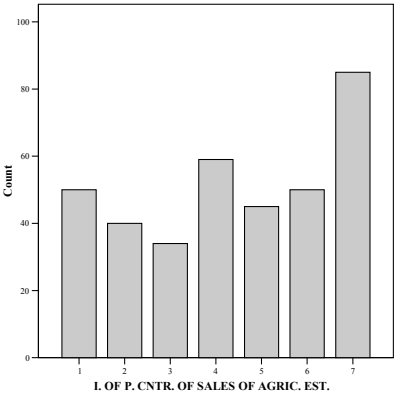
På noen spørsmål kan du krysse av flere svar

	lykkes dårlig/ lite viktig						lykkes godt/ svært viktig	vet ikke
Kodeverdi	1	2	3	4	5	6	7	8

Dei som ikkje kryssar av noko svar vert koda 9

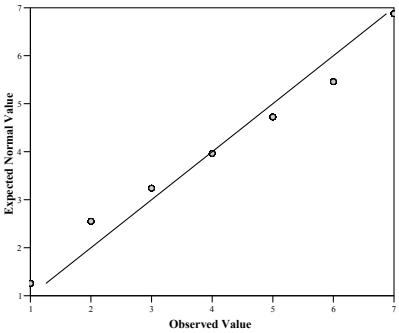
I. OF P. CNTR. OF SALES OF AGRIC. EST.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	50	13.2	13.8	13.8
	2	40	10.5	11.0	24.8
	3	34	8.9	9.4	34.2
	4	59	15.5	16.3	50.4
	5	45	11.8	12.4	62.8
	6	50	13.2	13.8	76.6
	7	85	22.4	23.4	100.0
Total		363	95.5	100.0	
Missing	8	12	3.2		
	9	5	1.3		
	Total	17	4.5		
Total		380	100.0		

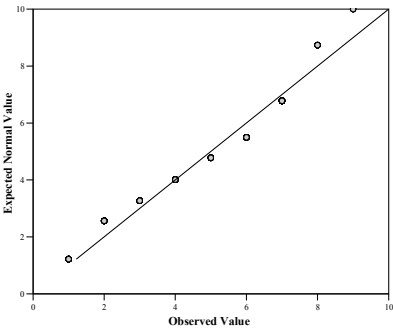


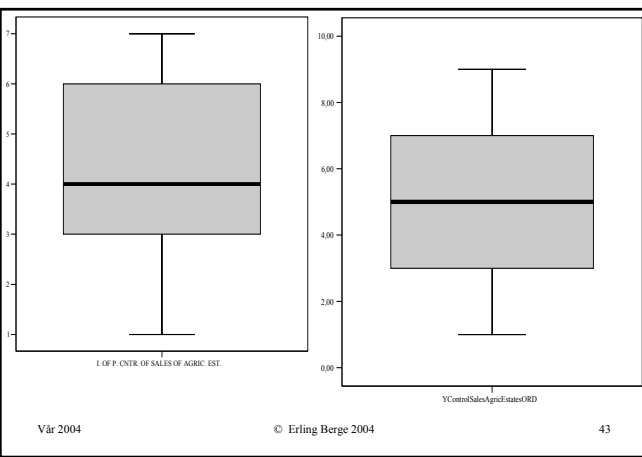
		I. OF P. CNTR. OF SALES OF AGRIC. EST.	YControlSalesAgricEstate Valid N (listwise)
N	Statistic	380	363
Range	Statistic	8	6.00
Minimum	Statistic	1	1.00
Maximum	Statistic	9	7.00
Sum	Statistic	1729	1588.00
Mean	Statistic	4.55	4.3747
	Std. Error	.114	.11045
Std. Deviation	Statistic	2.213	2.10435
Variance	Statistic	4.897	4.428
Skewness	Statistic	-.171	-.234
	Std. Error	.125	.128
Kurtosis	Statistic	-1.148	-1.267
	Std. Error	.250	.255

Normal Q-Q Plot of I. OF P. CNTR. OF SALES OF AGRIC. EST.



Normal Q-Q Plot of I. OF P. CNTR. OF SALES OF AGRIC. EST.





Box-plott

- Boksen vert konstruert ut frå kvartilverdiane Q_1 og Q_3
- Nærliggjande store verdiar vert definer som dei som ligg utanfor boksen men innan for $Q_3 + 1.5 \cdot IQR$ eller $Q_1 - 1.5 \cdot IQR$
- Utliggjarar (alvorlege ekstremverdiar er dei som ligg utanfor $Q_3 + 1.5 \cdot IQR$ eller $Q_1 - 1.5 \cdot IQR$

Vår 2004

© Erling Berge 2004

44

Om datainnsamling og datakvalitet

- Spørsmåla – teknikkane her vil vi ikkje diskutere
- Utvalet
 - Frå trekking til ferdig datamatrikse, seleksjon, nekting og manglande svar
- Kva er viktig for kvaliteten av data?
 - Samanhengen mellom manglande observasjonar og fenomenet som vert studert
- Kva skal vi gjere når data er mangelfulle?

Vår 2004

© Erling Berge 2004

45

Formulering av modellar

- Definisjon av elementa i modellen
 - variablar, feilledd, populasjon og utval
- Definisjon av relasjonar mellom elementa
 - utvalsprosedyre, tidsrekkefølge av hendingar og observasjonar, likninga som bind elementa saman
- Presisering av føresetnader for bruk av gitt estimeringsmetode
 - tilhøve til substansteori (spesifikasjon)
 - fordeling og eigenskapar ved feilledd

Elementa i modellen

- Populasjon:
- Utval:
- Variablar:
- Feilledd:

Relasjonar mellom elementa

- Utvalsprosedyre: skeive (biased) utval
- Tidsrekkefølge av hendingar og observasjonar
- Samvariasjon, genuin vs. spuriøs samvariasjon
 - Konklusjonar om kausalsamband krev genuin samvariasjon
- Likninga:

Bivariat Regresjon: Modell for populasjon

- $Y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$
- $i=1, \dots, n$ $n = \# \text{ case i populasjonen}$
- Y og X må definerast eintydig, og Y må ha målenivå intervallskala i ordinær regresjon

Bivariat Regresjon: Modell for utval

- $Y_i = b_0 + b_1 x_{1i} + e_i$
- $i=1, \dots, n$ $n = \# \text{ case i utvalet}$
- Y og X må definerast eintydig, og Y må ha målenivå intervallskala i ordinær regresjon

Regresjonseksempellet

- Eksempellet som følger inneheld ei rekkje feil. Ein slik regresjon kvalifiserer til stryk
- Det blir lesaren si oppgave å identifisere feila så fort som råd er, og så aldri gjere slike feil sjølv
- Tips: sjå tilbake på fordelingane av variablane ovanfor

Importance of public control of sales of agric. Estates Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.047(a)	.002	.000	2.213

a Predictors: (Constant), NUMBER OF DEKAR LAND OWNED

Importance of public control of sales of agric. Estates ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4.145	1	4.145	.846	.358(a)
	Residual	1851.905	378	4.899		
	Total	1856.050	379			

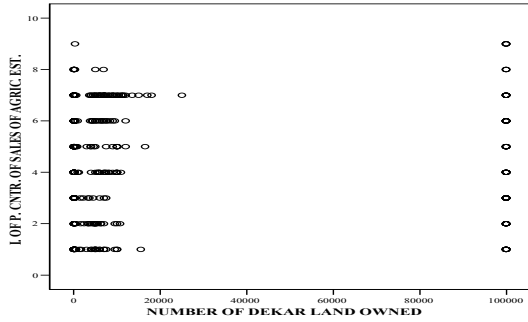
a Predictors: (Constant), NUMBER OF DEKAR LAND OWNED
b Dependent Variable: I. OF P. CNTR. OF SALES OF AGRIC. EST.

Importance of public control of sales of agric. Estates Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.610	.131		35.233	.000
	NUMBER OF DEKAR LAND OWNED	.000	.000	-.047	-.920	.358

a Dependent Variable: I. OF P. CNTR. OF SALES OF AGRIC. EST.

Spredningsdiagram

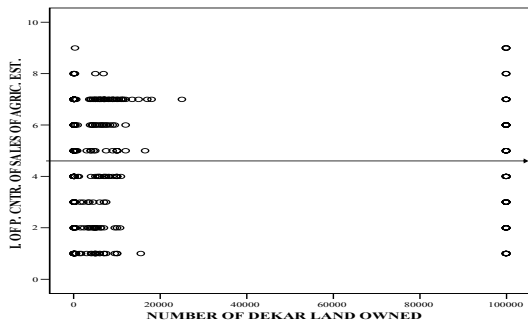


Vår 2004

© Erling Berge 2004

55

Spredningsdiagram med regresjonslinje



Vår 2004

© Erling Berge 2004

56

FØRESETNADER FOR OLS REGRESJON

OLS: ordinary least squares (minste kvadreaters metode)

Krava til regresjonsanalysen kan kort oppsummerast ved

- Vi antar at den lineære modellen er korrekt med uavhengige og identisk normalfordelte feil (“normal i.i.d. errors”)

Vår 2004

© Erling Berge 2004

57

OLS metoden

Observert feil

- $e_i = (Y_i - b_0 - b_1 x_{1i})$

Kvadrert og summert observert feil

- $\sum_i (e_i)^2 = \sum_i (Y_i - b_0 - b_1 x_{1i})^2$

Finn b_0 og b_1 som minimerer kvadratsummen

Tilhøvet utval-populasjon I

- Forventa verdi: vi skriv $E[*]$ der * står for eitt eller anna uttrykk som inneheld minst ein variabel, t.d.
- $E[Y_i] = E[b_0 + b_1 x_{1i} + e_i]$
 $= \beta_0 + \beta_1 x_{1i}$
- $E[b_0] = \beta_0$; $E[b_1] = \beta_1$; $E[e_i] = \varepsilon_i$

Tilhøvet utval-populasjon II

- Tilhøvet utval-populasjon er fastlagt gjennom dei eigenskapane som feilleddet har fått gjennom utvalsprosedyren og observasjonsplanen
- Ved reint tilfeldig utval og fullstendig observasjon vil
 $E[\varepsilon_i] = 0$ for alle i , og
 $\text{var}[\varepsilon_i] = \sigma^2$ for alle i

Fullstendig observasjon

- Gjer det mulig å sette opp ein fullstendig spesifisert modell. Dette tyder at alle variablar som kausalt påverkar det fenomenet vi studerer (Y) er observert, dvs. inkludert i likninga
- Dette er i praksis umogeleg. Derfor nyttar vi feilleddet til å samle opp uobserverte faktorar

Hypotesetesting I

	I røynda er H_0 sann	I røynda er H_0 usann
Vi konkluderer med at H_0 er sann	Metoden gir rett konklusjon med sannsyn $1 - \alpha$	<u>Feil av type II</u> (sannsyn $1 - \beta$)
Vi konkluderer med at H_0 er usann	<u>Feil av type I</u> Testnivået α er sannsynet for feil av type I	β = styrken til testen

Hypotesetesting II

- Ein test er alltid konstruert ut frå førestenaden at H_0 er rett
- Testkonstruksjonen fører fram til ein
 - **testobservator**
- Testobservatoren er konstruert slik at den har ei kjent sannsynsfordeling, ei
 - **samplingfordeling**

T-test og F-test

- Kvadratsummar
 - $TSS = ESS + RSS$
 - $RSS = \sum_i (e_i)^2 = \sum_i (Y_i - \hat{Y}_i)^2$ avstand observert – estimert verdi
 - $ESS = \sum_i (\hat{Y}_i - \bar{Y})^2$ avstand estimert verdi – gjennomsnitt
 - $TSS = \sum_i (Y_i - \bar{Y})^2$ avstand observert verdi – gjennomsnitt
- Testobservator
 - $t = (\mathbf{b} - \boldsymbol{\beta}) / SE_b$ $SE = \text{standard error}$
 - $F = [ESS/(K-1)] / [RSS/(n-K)]$ $K = \text{talet av parametrar}$

Testen sin p-verdi

- Testen sin p-verdi gir oss det estimerte sannsynet for å observere dei verdiane vi har i utvalet eller verdier som er enno meir gunstige ut frå teorien om at H_0 er gal dersom utvalet vårt er reint tilfeldig trekt frå ein populasjon det H_0 er rett
- Svært låge p-verdiar gjer at vi ikkje kan tru at H_0 er rett

Konfidensintervall for β

- Vel ein t_{α} -verdi frå tabellen over t-fordelinga med $n-K$ fridomsgrader slik at intervallet $< b - t_{\alpha}(SE_b), b + t_{\alpha}(SE_b) >$ i ein tosidig test gir eit sannsyn på α for å gjere feil av type I
- Dette tyder at $b - t_{\alpha}(SE_b) \leq \beta \leq b + t_{\alpha}(SE_b)$ med sannsyn $1 - \alpha$

Determinasjonskoeffesienten

- Determinasjonskoeffesienten:
- $R^2 = ESS/TSS = \sum_i (\hat{Y}_i - \tilde{Y})^2 / \sum_i (Y_i - \tilde{Y})^2$
 - Fortel kor stor andel av variasjonen rundt gjennomsnittet vi ”forklarer” ved hjelp av variablane vi nyttar i regresjonen ($\hat{Y}_i$)
- I bivariat regresjon er determinasjonskoeffesienten lik korrelasjonskoeffesienten: $r_{yu}^2 = s_{yu} / s_y s_u$
- Kovariansen $s_{yu} = [\sum_i (Y_i - \tilde{Y})(U_i - \tilde{U})] / (n - 1)$
