FORELESNING 6 20

EIVIND ERIKSEN

APR 10 2012

MET 3431

STATISTIKK

PLAN:

1. Repetisjon: Korrelasjon (lineær) ← Se Forelesning 19
2. Lineær regresjon [T] 10.3
3. Krysstabeller [T] 11.1, 11.3

Husk: Arbeidskrav 7 Publiseres 10. april

Veiledning II Tirsdag 10. april 17.15-20.15 i B2.

(1)

Repetisjon: Test for lineær korrelasjon

H_0: $\rho = 0$ (ingen lineær korrelasjon)

H_1: $\rho \neq 0$ / $\rho > 0$ / $\rho < 0$ (lineær korrelasjon) positiv / negativ

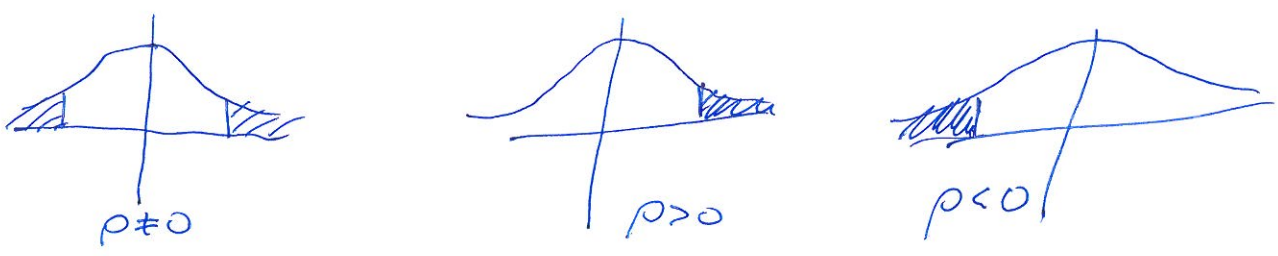
Test-observator:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.87}{\sqrt{\frac{1-0.87^2}{3}}}$$

$n-2$ frihetsgrader

(r = korrelasjonskoeffisient)

Sammenliknes med kritisk t-verdi:



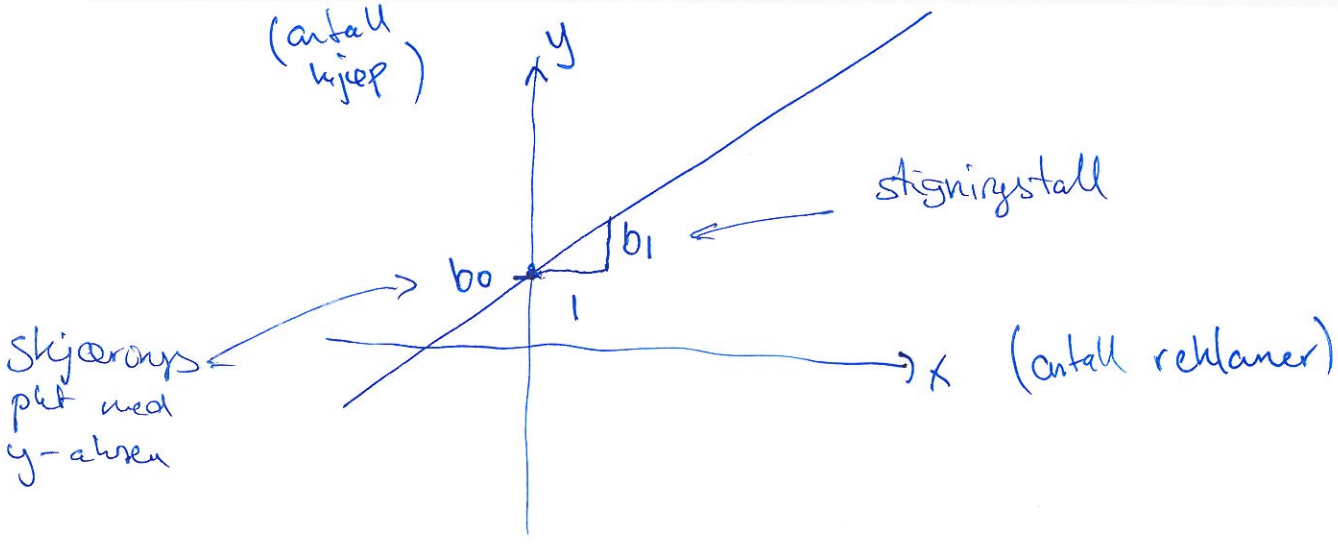
Regresjon

Section 10-3

- Anta at vi har vist at det er en signifikant korrelasjon
- Da vil vi beskrive korrelasjonen mellom x og y
- Det gjøres med en *regresjonsformel*

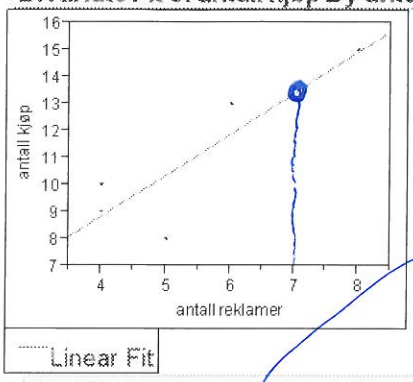
$$y = b_0 + b_1 x$$

- Kalkulator/JMP/Excel beregner denne formelen
- Formelen beskriver en *regresjonslinje*
- x kalles den *uavhengige* variabelen
- y er den *avhengige* variabelen



Regresjonsformelen

Bivariate Fit of antall kjøp By antall reklamer



$b_0 \simeq 2.8$

Linear Fit

antall kjøp = 2.8035714 + 1.5178571 * antall reklamer

$b_1 \simeq 1.5$

Summary of Fit

RSquare	0.758929
RSquare Adj	0.678571
Root Mean Square Error	1.652919
Mean of Response	11
Observations (or Sum Wgts)	5

Example

Person	1	2	3	4	5
Antall reklamer	5	4	4	6	8
Antall kjøp	8	9	10	13	15

- Side 16: korrelasjonen er signifikant
- Regresjon handler om å gi en eksakt formel for korrelasjonen

Regresjonsformelen

Example

- JMP (rød diamant: *Fit Line*) gir formelen

$$y = 2.8 + 1.5 \cdot x$$

- Dette er formelen for en rett linje med $b_0 = 2.80$ og stigningstall $b_1 = 1.52$
- Kalles *regresjonslinja*
- Regresjonsformelen

$$\text{Konfektkjøp} = 2.80 + 1.52 \cdot \text{ Reklamer sett}$$

- Truls-Ivar har sett reklamen 7 ganger. Han vil kjøpe

$$\text{Konfektkjop} = 2.80 + 1.52 \cdot 7 = 13.44$$

pakker med konfekt

Notasjon

Stikkprøven

- Observatorer b_0 og b_1
- Regresjonslinja $y = b_0 + b_1 x$

Populasjonen

- Parametre β_0 og β_1
- Regresjonslinja $y = \beta_0 + \beta_1 x$

Beregne b_0 og b_1

- Kalkulator/JMP/Excel kan beregne b_0 og b_1
- Det viktigste for oss er å kunne *tolke* dem

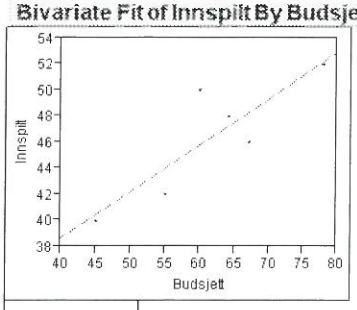
Regresjonslinja

Regresjonslinja er den linja som best passer til punktene i scatterplottet.

Sjekk ut scriptet *demoLeastSquares* i JMP, ligger i folder *Sample scripts*

Regresjonslinja for Hollywood

Bivariate Fit of Innspilt By Budsjett



Linear Fit

Innspilt = 24.407674 + 0.3565148*Budsjett

Summary of Fit

RSquare	0.740709
RSquare Adj	0.675887
Root Mean Square Error	2.637733
Mean of Response	46.33333
Observations (or Sum Wgts)	6

Example

budsjett	67	64	45	78	60	55
Innspillt	46	48	40	52	50	42

- 6 Hollywood filmer.
- JMP: *Fit Y by X* gir

$$\text{Innspilt} = 24.41 + 0.36 \cdot \text{Budsjett}$$

- For hver million brukt i markedsføring, så spilles det inn 0.36 mer på første helg
- En film med budsjett på 63 millioner forventes å spille inn

$$\text{Innspilt} = 24.41 + 0.36 \cdot 63 = 47.1$$

Sammenheng mellom regningen og tips

(2) $r^2 = 0.828^2 \simeq 0.686$

(3) $b_0 = -0,35$
$b_1 = 0,15$

Bill ($)	33.46	50.68	87.92	98.84	63.60	107.34
Tip ($)	5.50	5.00	8.08	17.00	12.00	16.00

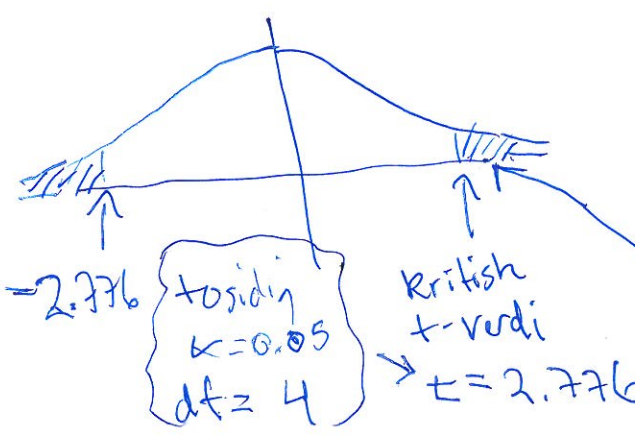
(4) $\hat{tip} = -0.35 + 0.15 \cdot 50 = 7.15$

Example

Er det en sammenheng mellom størrelse på regning og størrelse på tips?

1. Er det en korrelasjon? Test på $\alpha = 0.05$ nivået
2. Beregn r^2 og tolk svaret
3. Hvis det er en korrelasjon, hva er da regresjonslinja?
4. Forutsi hva en regning på 50 dollar vil gi i tips

(1) $H_0: \rho = 0$ $\quad H_1: \rho \neq 0$



$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.828}{\sqrt{\frac{1-0,828^2}{4}}}$$

$\simeq 2.953$

Forkaster H_0.

Example

SVAR: Legger inn i JMP/kalkulator og får

1. $r = 0.828$ og $t = \frac{0.828}{\sqrt{\frac{1-0.828^2}{6-2}}} = 2.953$ Vi tester $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$. Da blir kritisk verdi $t_{0.025} = 2.77$. Vi forkaster H_0; det er grunn til å hevde at det er en korrelasjon mellom størrelsen på regningen og tipset
2. $r^2 = 0.686$. 68.6% av variasjonen i tips skyldes variasjon i regningsstørrelsen. De resterende 31.4% skyldes andre faktorer enn størrelsen på regningen.
3. JMP/kalkulator gir $\text{Tips} = -0.35 + 0.15 \cdot \text{Regning}$
4. En regning på 50 dollar forventes å gi $-0.35 + 0.15 \cdot 50 = 7.15$ dollar i tips

Linear Fit

tips = -0.347279 + 0.1486141*regning

Oppgaver

Les

Bla gjennom kapittel 11 i boka til neste gang.

Oppgaver kapittel 9

- 10-2: 1, 3, 9, 15, 21, 29
- 10-3: 1, 3, 7, 15, 21

1. 11-1: Kji-kvadrat fordelingen
2. 11-3: Krysstabeller og kji-kvadrattesten
3. Kji-kvadrattesten i JMP

Kapittel 11

Samvariasjon mellom to kategoriske variabler

- Korrelasjon og regresjon handler om samvariasjon mellom to kontinuerlige variable
- I dette kapitlet handler det om samvariasjon mellom to *kategoriske* variable

Example

For filmer:

- Sammenheng mellom budsjett og inntekter $\leftrightarrow$ korrelasjon/regresjon
- Sammenheng mellom filmsjanger og filmens nasjonalitet $\leftrightarrow$ krysstabell og kji-kvadrattesten

Krysstabeller

Når vi har samlet inn data om to kategoriske data, kan dette presenteres i en *krysstabell*

Example

Er det en sammenheng mellom farge på hjelmen og risiko for trafikkulykke?
To kategoriske variabler:

Farge Svart, hvit eller gul

Ulykke Involvert eller ikke involvert

Dataene vises i krysstabellen

	Svart	Hvit	Gul	Total
Ikke involvert	491	377	31	899
Involvert	213	112	8	333
Total	704	489	39	1232

$P(\text{svart}) = \frac{704}{1232}$

$P(\text{ikke involvert}) = \frac{899}{1232}$

$P(\text{svart og ikke involvert}) = \frac{704}{1232} \cdot \frac{899}{1232} \approx 0.417$

↑ hvis de to variablene er uavhengige

$\Rightarrow$ Forventet antall: $0.417 \cdot 1232$

Notasjon for krysstabeller

- O står for den observerte hyppigheten (frekvensen) i cellen
- E står den forventede hyppigheten
- r står for antall rader, og c for antall søyler i krysstabellen

Example

	Svart	Hvit	Gul	Total
Ikke involvert	491 (513.714)	377 (356.827)	31(28.459)	899
Involvert	213 (190.286)	112 (132.173)	8 (10.541)	333
Total	704	489	39	1232

- $r = 2$ og $c = 3$
- E er forventet verdi. Selv om 491 syklister med svart hjelm ikke hadde ulykke, så ville vi forventet 513.714 dersom farge og ulykker er uavhengige:

$$E = \frac{899 \cdot 704}{1232} = 513.714$$

$\frac{899}{1232} \cdot \frac{704}{1232} \cdot 1232 =$