Plan:

1. Lineær regresjon
2. Repetisjon av hele pensum

Siste forelesning: 31/05

Repetisjon av kapittel 6-7 + Gjennomgang av prøve-eksamen.

Kontortid: Bortsett fra uke 18 Ellers kontortid 10-18 ca.

Kursevaluering: Husk å svare på kursevalueringen når den kommer!

Jeg kommer til å legge ut en prøve-eksamen til i god tid før forelesningen 31/05:

Prøve-eksamen I = Oppsavesett 15

Prøve-eksamen II (kommer senere)

① Lineær regresjon

X: forklaringsvariabel
Y: responsvariabel

} Ønsker å studere sammenheng mellom X og Y

(a) Korrelasjonskoeffisient

$$R = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{\frac{1}{n-1}\cdot\sum(x_i-\bar{x})\cdot(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum(x_i-\bar{x})^2}\cdot\sqrt{\frac{1}{n-1}\sum(y_i-\bar{y})^2}}$$

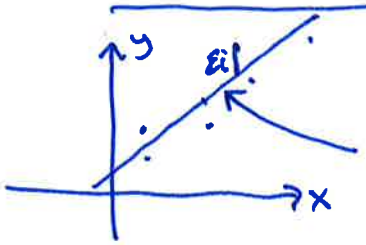
x_i	y_i
x_1	y_1
x_2	y_2
⋮	⋮
x_n	y_n

Husk: $-1 \le R \le 1$

$R > 0$ positiv sammenheng
$R < 0$ negativ — " —
$|R|$: hvor sterk sammenhengen er

(b) Lineær regresjon:

$$Y = \alpha + \beta X + \varepsilon \qquad \hat{y} = \alpha + \beta x$$



spredningsdiagram

regresjonslinjen: passer best mulig med dataene

Forutsetninger:

- ε er $N(0,\sigma)$ for en konstant σ
- Når vi trekker n punkter (datasettet), er $\varepsilon_1, \dots, \varepsilon_n$ uavhengige.

$$y_1 = \alpha + \beta x_1 + \varepsilon_1$$
$$\vdots \qquad\qquad \vdots$$
$$y_n = \alpha + \beta x_n + \varepsilon_n$$

Beste estimat for regresjonslinjen: $\hat{y} = \hat{\alpha} + \hat{\beta} x$

$$\hat{\beta} = \frac{S_{XY}}{S_X^2} = R \cdot \frac{S_Y}{S_X}$$
$$\hat{\alpha} = \bar{Y} - \hat{\beta}\cdot\bar{x}$$

← Beste estimat ved minste kvadraters metode, dvs $\varepsilon_1^2 + \varepsilon_2^2 + \dots + \varepsilon_n^2$ er minst mulig.

Statistiske egenskaper:

$\hat{\alpha}, \hat{\beta}$ estimatorer: Stokastiske variabler

$$SE(\hat{\alpha})^2 = \frac{\sigma^2 \cdot \sum x_i^2}{n(\sum x_i^2 - n\bar{x}^2)}$$

$$SE(\hat{\beta})^2 = \frac{\sigma^2}{\sum x_i^2 - n\bar{x}^2}$$

Hvis $\hat{\theta}$ er en estimator for θ, så er:

$E(\hat{\theta}) = \theta$ (forv.rett)

$SE(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$

mest mulig

$SE(\hat{\theta})$ = std. avvik til estimatoren $\hat{\theta}$

Eks: $\bar{X}$ estimator for μ

$E(\bar{x}) = \mu$

$SE(\bar{x}) = \sigma/\sqrt{n}$

når $x_1, \ldots, x_n \sim N(\mu, \sigma)$

Estimat for σ^2:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

der $\hat{y}_i = \hat{\alpha} + x_i\hat{\beta}$

Tolkning av R^2:

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) = e_i + (\hat{y}_i - \bar{y})$$

e_i: feil-ledd, residual

$(\hat{y}_i - \bar{y})$: forklart av regresjonslinjen

Man kan vise:

1) $$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$SS_T = SS_E + SS_R$$

2) $$R^2 = \frac{SS_E}{SS_T} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{\frac{1}{n-1}\sum(\hat{y}_i - \bar{y})^2}{\frac{1}{n-1}\sum(y_i - \bar{y})^2} = \frac{s_{\hat{y}}^2}{s_y^2}$$

Tolkning: R^2 er andelen av variasjonen i y som er forklart av regresjonslinjen

Utledning av S^2: $S^2 = \frac{1}{n-2} \cdot \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{1}{n-2} SS_E$

$R^2 = \frac{SS_R}{SS_T}$

$$S^2 = \frac{1}{n-2} SS_E = \frac{1}{n-2}(SS_T - SS_R)$$

$$= \frac{1}{n-2} \cdot (SS_T - R^2 \cdot SS_T) = \frac{1}{n-2} SS_T \cdot (1-R^2)$$

$$= \frac{1}{n-2} \cdot (n-1) S_y^2 \cdot (1-R^2) = \frac{n-1}{n-2} S_y^2 (1-R^2)$$

Eks: Hypotesetest for β $\begin{cases} H_0: \beta = 0 \\ H_1: \beta \neq 0 \end{cases}$ $(\beta_0 = 0)$

Testobservator: $T = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})} = \frac{\hat{\beta}}{SE(\hat{\beta})}$

t-fordelt med $n-2$ df hvis $\beta = \beta_0$

Forkastningsområde: $|T| > t_{\alpha/2}^{n-2}$

Beregning av $SE(\hat{\beta})$:

$$SE(\hat{\beta})^2 = \frac{\sigma^2}{\sum x_i^2 - n\bar{x}^2} \rightarrow \frac{S^2}{\sum x_i^2 - n\bar{x}^2} =$$

estimer σ^2 ved S^2

$$\rightarrow \frac{\frac{n-1}{n-2} S_y^2 (1-R^2)}{(n-1) S_x^2} = \frac{1}{n-2} \frac{S_y^2}{S_x^2} (1-R^2)$$

$SE(\hat{\beta})$ estimeres ved $\sqrt{\frac{1}{n-2} \frac{S_y^2}{S_x^2} (1-R^2)}$

$$SE(\hat{\alpha})^2 = \frac{\sigma^2 \sum x_i^2}{n \cdot (\sum x_i^2 - n\bar{x}^2)}$$

estimeres med

$$\frac{s^2 \sum x_i^2}{n(\sum x_i^2 - n\bar{x}^2)} = \frac{\frac{n-1}{n-2} s_y^2 (1-R^2) \sum x_i^2}{n \cdot (n-1) s_x^2} = \frac{1}{n(n-2)} \frac{s_y^2}{s_x^2} (1-R^2) \sum x_i^2$$

Konfidensintervall:

For β: $\hat{\beta} \pm t_{\alpha/2}^{n-2} \cdot SE(\hat{\beta})$

For α: $\hat{\alpha} \pm t_{\alpha/2}^{n-2} \cdot SE(\hat{\alpha})$

(2) Repetisjon:

(A) Sannsynlighetsregning Kap 3 - 5

Stokastisk forsøk

utfallsrom $U = \{w_1, w_2, \ldots\}$

Sannsynlighetsmål: funksjon p som til enhver hendelse A (dvs delmengde av U) tilordner et tall $p(A)$ slik at

i) $p(A) \geq 0$ for enhver hendelse A

ii) $p(U) = 1$

iii) $p(A_1 \cup A_2 \cup \ldots) = p(A_1) + p(A_2) + \ldots$ når A_i, A_j er disjunkte (uten felles utfall) for alle i, j.

Konsekvenser: i) $p(A^c) = 1 - p(A)$

ii) $p(A \cup B) = p(A) + p(B) - p(A \cap B)$

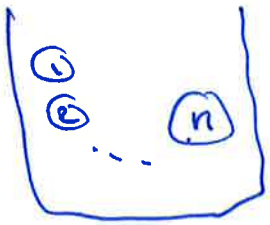
Spesialtilfelle: Uniform sannsynlighet = alle utfall har samme sannsynlighet.

Da har vi

$$p(A) = \frac{\text{antall gunstige utfall}}{\text{antall mulige utfall}}$$

Kombinatorikk: Å telle opp ting

Urnemodell: trekker r blant n kuler fra en urne



i) ordnet, med tilbakelegging: n^r

ii) ordnet, uten tilbakelegging: $n \cdot (n-1) \cdots (n-r+1) = {}_nP_r$

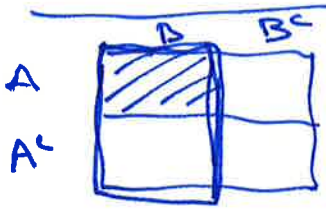
iii) uordnet, uten tilbakelegging: $\frac{n(n-1)\cdots(n-r+1)}{r!} = {}_nC_r$

på kalkulator: ${}_nC_r$

Skrivemåte: $\longrightarrow \binom{n}{r}$

Betinget sannsynlighet: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Sannsynligheten for A, gitt at B inntreffer



Uavhengighet: A og B er uavhengige hvis

$$P(A \cap B) = P(A) \cdot P(B)$$

$$\Updownarrow$$

$$\frac{P(A \cap B)}{P(B)} = P(A)$$

$$P(A|B) = P(A)$$

Stokastiske variabler: X som tar verdier avhengig av utfallet til et stokastisk forsøk

Typisk:
- diskret ← mulige verdier = endelig mengde
- kontinuerlig ← mulige verdier = kont. intervall

i) Diskret:

fordelingsfunksjon = (P)

Sannsynligheter for ulike utfall

x	$p(x=x)$
x_1	$p(x_1)$
x_2	$p(x_2)$
⋮	⋮
x_n	$p(x_n)$

Forventningsverdi:

$$E(x) = \sum_x x \cdot p(x)$$ forventningsverdi

$$E(g(x)) = \sum_x g(x) \cdot p(x)$$

$$Var(x) = E[(x - E(x))^2]$$ varians

Simultan fordeling: X og Y simultant fordelt $p(x,y) = p(X=x, Y=y)$

	x_1	x_2	...	x_n	
y_1	$p(x_1,y_1)$	·	·	·	$p(Y=y_1)$
y_2	·	·	·	·	
⋮	·	·	·	·	
y_n	·	·	·	·	
	$p(X=x_1)$				

kolonne- og radsummer = marginale fordelingsfunksjoner

Forventningsverdi:

$$E(x) = \sum_x x \cdot p_X(x) \qquad E(y) = \sum_y y \cdot p_Y(y)$$

$$Cov(x,y) = \sum_{x,y} (x - E(x))(y - E(y))$$

$$= E(xy) - E(x) \cdot E(y)$$

Regneregler:

i) $E(aX+b) = aE(X)+b$

ii) $Var(X) = E(X^2) - E(X)^2$

iii) $Var(aX+b) = a^2 \cdot Var(X)$

iv) $Cov(X,Y) = E(XY) - E(X) \cdot E(Y)$

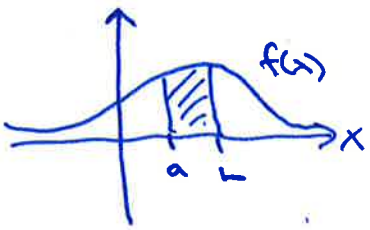
v) $Cov(X_1+X_2, Y_1+Y_2) = Cov(X_1,X_1) + Cov(X_1,Y_2) + Cov(X_2,Y_1) + Cov(X_2,Y_2)$

$Cov(X,Y) = Cov(Y,X)$

$Cov(cX,Y) = Cov(X,cY) = c \cdot Cov(X,Y)$

vi) $Var(X) = Cov(X,X)$

ii) Kontinuerlig: Tetthetsfunksjonen $f(x)$



$P(a \le X \le b) = \int_a^b f(x)\,dx$

punktsannsynlig: $P(X=a) = 0$

Forventning: ∞

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx$$

$$E(g(x)) = \int_{-\infty}^{\infty} g(x) \cdot f(x)\,dx$$

$$Var(X) = E[(X - E(X))^2]$$

Alle regnereglene overfor gjelder også for kontinuerlige stokastiske variabler.

Sannsynlighetsfordelinger:

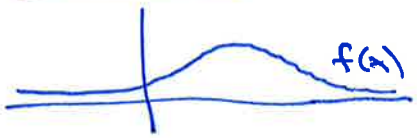
Binomisk: X = antall forekomster av en hendelse A når vi gjør et forsøk der $P(A)=p$ n ganger, og disse er uavhengige

diskret variabel ⟶ parametre: n, p $\quad q=1-p$

Mulige verdier: $0,1,2,\ldots,n$ $\quad P(X=i)=\binom{n}{i}p^i q^{n-i}$, $i=0,1,\ldots,n$

$E(X)=np$ $\quad Var(X)=npq$

Normalfordelt: $N(\mu,\sigma)$ normalfordelt med forventning μ, std.avvik σ



$E(X)=\mu$, $Var(X)=\sigma^2$

$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$

i) X normalfordelt $\Rightarrow$ $aX+b$ normalfordelt
X,Y normalfordelt $\quad X+Y$ — " —

ii) $X_1,\ldots,X_n,\ldots$ uavhengige, identisk fordelte $\Rightarrow$ $\bar{X}_n=\frac{1}{n}(X_1+X_2+\ldots+X_n)$ → normalfordelt når $n\to\infty$

Normaltilnærming til binomisk fordeling:

Hvis X er binomisk fordelt med parametre n,p og n er stor, så er X tilnærmet $N(\mu,\sigma)$ $(npq\geq 5)$ $\quad \mu=E(X)=np$ $\quad \sigma^2=Var(X)=npq$

For å få en mest mulig nøyaktig tilnærming, bruker vi heltallskorrelasjon:

korrelasjon = 0.5

$$p(X \leq a) \simeq \Phi\left(\frac{a+0{,}5-\mu}{\sigma}\right) \quad \text{hvor} \begin{cases} \mu = np \\ \sigma^2 = npq \end{cases}$$