Enkel lineær regresjon.

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ uavh.

MKE $\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}} = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}$, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$

Forv.: $\hat{\beta}_1$ med varians $\frac{\sigma^2}{S_{xx}}$ og $\hat{\beta}_0$ med $\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$

Spesielt er $\bar{Y}$ og $\hat{\beta}_1$ ukorrelert

Dessuten fant vi at $t_1 = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{n-2}$ fordelt under $H_0: \beta_1 = 0$, da $\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{1}{n-2} \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

$\Rightarrow F = t_1^2 \sim F_{1, n-2}$

Rapporteres ofte ANOVA-tabellen for ugyr. analysen.

Kvadratsums oppspaltning, nå uint

$SST = \sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2 + \hat{\beta}_1^2 S_{xx}$

da $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{Y} + \hat{\beta}_1 (x_i - \bar{x})$

Dessuten, husk at $\hat{\beta}_1^2 S_{xx} = \sum (\hat{Y}_i - \bar{Y})^2 = SSR$ = kvadratsum av avvik til regr.

Altså $SST = SSR + SSE$

Til SST svarer $n-1$ frihetsgrader

SSE " $n-2$ —"—

SSR " 1 —"—

okt 24-12:16

Dessuten - når $\beta_1 = 0$, $\hat{\beta}_1 \sim N(0, \sigma^2/S_{xx})$

$\Rightarrow E[\hat{\beta}_1^2 S_{xx}] = Var(\hat{\beta}_1 \sqrt{S_{xx}}) = \sigma^2$

" $E[SSR]$

Anova tabell

Kilde	SS	df	MS	F	P-verdi
x	SSR	1	SSR/1 = SSR	SSR/MSE	
Residual	SSE	$n-2$	MSE = $\frac{SSE}{n-2}$		
Total	SST	$n-1$			

okt 24-12:35

Prediksjon + KI / PI

Ex) x_i = barnets alder

Y_i = lunge funksjon (FEV$_1$, PEF, ...)

Ny verdi: x^* for forklaringsvariabel

Predikert verdi $\hat{Y}_{x^*} = \hat{\beta}_0 + \hat{\beta}_1 x^* = \begin{cases} \text{Estimat } E Y_{x^*} \\ \text{Predikert } Y_{x^*} \end{cases}$

Sen fokus på estimeringsusikkerheten

$Var(\hat{Y}_{x^*}) = Var(\hat{\beta}_0 + \hat{\beta}_1 x^*) = Var(\bar{Y} + \hat{\beta}_1 (x^* - \bar{x}))$

$= Var(\bar{Y}) + (x^* - \bar{x})^2 Var(\hat{\beta}_1)$

$= \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right)$

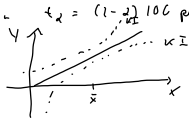
Husk at $\frac{\hat{Y}_{x^*} - E(\hat{Y}_{x^*})}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}} = \frac{\hat{Y}_{x^*} - (\beta_0 + \beta_1 x^*)}{\hat{\sigma}\sqrt{\quad}} \sim t_{n-2}$

når $\varepsilon_i \sim N(0, \sigma^2)$ uavh

Dette leder til at $\hat{Y}_{x^*} \pm t_{\alpha/2} \cdot \hat{\sigma}\sqrt{\quad}$

er et $(1-\alpha)100\%$ KI for $\beta_0 + \beta_1 x^*$.

når $t_\alpha = (1-\alpha)100$ persentil i t_{n-2}.



okt 24-12:43

For prediksjonsintervall rundt vi på

$Y_{x^*} = \hat{Y}_{x^*} + \varepsilon^*$, $\varepsilon^* \sim N(0, \sigma^2)$ uavh $\varepsilon_1, \ldots, \varepsilon_n$

$E[Y_{x^*}] = E\hat{Y}_{x^*} + \underbrace{E(\varepsilon^*)}_{=0} = \beta_0 + \beta_1 x^*$

$Var(Y_{x^*}) = Var(\hat{Y}_{x^*}) + Var(\varepsilon^*)$

$= \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right) + \sigma^2$

$= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right)$

Tilsv. $\frac{Y_{x^*} - (\beta_0 + \beta_1 x^*)}{\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}} \sim t_{n-2}$

$\Rightarrow \hat{Y}_{x^*} \pm t_{\alpha/2} \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$ er et $[1-\alpha]100\%$ prediksjonsintervall for en ny verdi Y_{x^*}

okt 24-13:17

14.6 Modellsjekk

Vi skriver den lineære regresjonsmodellen som

$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma^2)$ uavh.

Modellen kan brytes opp i 4 elementer

(1) Linearitet $E(Y_i) = \beta_0 + \beta_1 x_i$

(2) Konstant varians $Var(Y_i) = Var(\varepsilon_i) = \sigma^2$

(3) Uavhengighet

(4) Normalitet, $\varepsilon_i \sim N(0, \sigma^2)$, $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

Elementene er satt opp etter viktighet

Hvis (1) svikter - alvorlig - er hele modellen unødvendig.

Hvis (1) holder, men (2) eller (3) svikter så er parameterestimatene $\hat{\beta}_0$ og $\hat{\beta}_1$ fortsatt forv.rette, men standardfeil kan være galt estimert. $\Rightarrow$ Problemer med inferens.

Hvis (1), (2) og (3) holder, (4) svikter, og datasettet er stort (n stor), så vil teorien i KI likevel være tilnærmet OK. Likevel: Pass på ved utliere.

okt 24-13:31

Modellsjekkene sjekkes ved å beregne

(1) Predikerte verdier $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

(2) Residualene $e_i = Y_i - \hat{Y}_i$

Noen ganger brukes i stedet for e_i

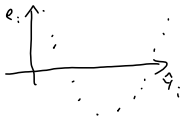
Standardiserte residualer $e_i^* = \frac{e_i}{\hat{\sigma}\sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}}}$

Man finner de standardiserte residualene ved å beregne $Var(e_i) = Var(\hat{Y}_i - Y_i) = \sigma^2 \left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right)$

Skal vises.

okt 24-13:40

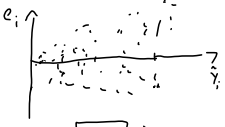
Sjekk (1) linearitet: Plott residualene e_i mot predikerte verdier $\hat{Y}_i$.



Kurvatur tyder på ikke linearitet.

Sjekk (2) Konstant varians

Samme plott



Hvis store/mindre variasjon i e_i med store $\hat{Y}_i$ indikerer dette at variansen endres med $E(Y_i) = \beta_0 + \beta_1 x_i$

evt. Plott $(\hat{Y}_i, \sqrt{|e_i^*|})$

Tyder også på ikke - konstant varians.

okt 24-13:50

Sjekk (3) uavhengighet

Tidsserie Y_i = obs. dag i

(a) Plott (i, e_i^*) I ikke automatisk i R

(b) Plott (e_{i-1}^*, e_i^*)

Sjekk (4) normalitet

QQ plott, boxplott, histogram over e_i (e_i^*).

okt 24-13:58