# Ekstaoppgave i STK1110 høsten 2007

**Oppgave E5**

Anta at vi har par av observasjoner $(y_i, x_i)$, for $i = 1, 2, \ldots, n$, og at vi har tilpasset en rett linje til punktene med minste kvadraters metode (slik det er beskrevet i avsnitt 14.1 i boka til Rice). Formålet med denne oppgaven er å studere et mål for hvor godt regresjonslinja beskriver dataene.

Vi lar $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ være de tilpassede verdiene, og innfører kvadratsummene

$$SS_{\text{tot}} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$
$$SS_{\text{reg}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$
$$SS_{\text{res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Den totale kvadratsummen $SS_{\text{tot}}$ angir hvor stor variasjon det er i $y_i$-ene. $SS_{\text{reg}}$ kalles kvadratsum for regresjon, og den angir hvor stor variasjon det er i de tilpassede verdiene. $SS_{\text{res}}$ kalles residual kvadratsum, og den angir hvor mye observasjonene avviker fra de tilpassede verdiene.

a) Vis at $SS_{\text{tot}} = SS_{\text{reg}} + SS_{\text{res}}$. Den totale kvadratsummen kan altså deles opp i en del som kan "forklares" av regresjonslinja og en del som ikke blir "forklart" av denne.

*Vink:* Ved direkte regning finner du at

$$SS_{\text{tot}} = SS_{\text{res}} + SS_{\text{reg}} + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}).$$

For å vise at den siste summen er lik null, bruker du at $\hat{y}_i = \bar{y} + \hat{\beta}_1(x_i - \bar{x})$ (hvorfor?) og uttrykket for $\hat{\beta}_1$ gitt på side 544, linje 13 nedenfra, i boka til Rice.

b) Et mål som er mye brukt for å angi hvor godt regresjonslinja beskriver dataene, er

$$R^2 = \frac{SS_{\text{reg}}}{SS_{\text{tot}}}$$

Vis at

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

og diskuter hvorfor $R^2$ kan brukes som et mål for hvor mye av variasjonen i $y_i$-ene som blir "forklart" av regresjonslinja.

c) Vis at $R^2$ er kvadratet av korrelasjonskoeffisienten mellom $x_i$-ene og $y_i$-ene (gitt på side 561, linje 5 ovenfra, i boka til Rice).