HG
April 2010

(Det vil være en fordel om du har gjort deg kjent med notatet **regresjon II** som er lagt ut på nettet før eller samtidig som du løser ekstraoppgave 4 og 5.)

Ekstraoppave 4

La $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ representere modellen for *n* observasjoner av to variable (x, Y), der x_i-ene er faste (ikke-stokastiske) tall mens $Y_1, Y_2, \ldots, Y_n$ er uavhengige stokastiske variable som oppfyller

(1) $\quad E(Y_i) = \alpha + \beta x_i \quad$ og $\quad \text{var}(Y_i) = \sigma^2 \quad$ for $\quad i = 1, 2, \ldots, n$

(*Dette er den enkle regresjonsmodellen i Løvås uten forutsetningen om at Y_i-ene er normalfordelte, som vi ikke trenger i denne oppgaven.*)

Minste kvadraters estimator (mkv) for regresjonskoeffisienten, β, er gitt ved

$$\hat{\beta} = \frac{S_{xy}}{s_x^2} \quad \text{der} \quad S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}) \quad \text{og} \quad s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

(a) Vis at S_{xy} og s_x^2 kan skrives

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})Y_i \quad \text{og} \quad s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})x_i$$

[Hint: Du kan ha nytte av regnereglene for summer som beskrevet i appendiks **A1** i regresjon-I-notatet.]

(b) Vis (ved hjelp av regel 4.12 og 4.17 i Løvås) at $\hat{\beta}$ er forventningsrett og at variansen er

$$\text{var}(\hat{\beta}) = \frac{\sigma^2}{(n-1)s_x^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

[**Hint:** Husk at x_i-ene, og derfor også s_x^2, er konstanter i denne modellen.]

Ekstraoppgave 5

(Fortsettelse av ekstraoppgave 3 om kvinnehøyder i Norge.)

Innledning. Vi tar utgangspunkt i den enkle regresjonsmodellen i Løvås. La *Y* være høyden for en tilfeldig valgt kvinne i Norge, og la *x* betegne høyden til hennes mor. Vi antar i modellen at *Y* er en stokastisk variabel (som uttrykk for at kvinnen er trukket tilfeldig fra populasjonen) mens *x* antas å være en fast (ikke-stokastisk) verdi[1]. Regresjonsfunksjonen, $\mu(x)$, er gitt ved

$$E(Y) = \mu(x) = \alpha + \beta x$$

der $\mu(x)$ er forventet høyde for en tilfeldig kvinne med en mor som er *x* cm høy (tolket som gjennomsnittshøyden i (del-)populasjonen av kvinner i Norge som alle har en mor som er *x* cm høy[2]).

For de $n = 42$ observasjons-parene i utvalget antar vi således bakgrunns-modellen: La $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ representere *n* observasjoner av (x, Y), der x_i -ene (mødrenes høyder) antas faste (ikke-stokastiske) tall mens $Y_1, Y_2, \ldots, Y_n$ (døtrenes høyder) antas uavhengige og normalfordelte stokastiske variable som oppfyller

(1) $\quad Y_i = \alpha + \beta x_i + e_i \quad \text{og} \quad \text{var}(Y_i) = \sigma^2 \quad \text{for} \quad i = 1, 2, \ldots, n$

der de ikke-observerbare restleddene, $e_1, e_2, \ldots, e_n$, er *uid* og normalfordelte, $e_i \sim N(0, \sigma)$, $i = 1, 2, \ldots, n$.

(Dette impliserer at $E(Y_i) = \mu(x_i)$ og $\text{var}(Y_i) = \text{var}(e_i) = \sigma^2$ for $i = 1, 2, \ldots, n$.)

Oppgave.

(a) Basert på resultatene i ekstraoppgave 3, estimer regresjonskoeffisienten β og beregn et 95% konfidensintervall for denne. Du vil antakelig finne at verdien $\beta = 0$ *ikke* er med i konfidensintervallet. Kommenter dette funnet.

(b) Estimer gjennomsnittshøyden for kvinner i Norge som har en mor på 172 cm (dvs estimer $\mu(172)$) og beregn et 95% konfidensintervall for denne. (Jfr. regresjon-II-notatet på nettet for beregning av standardfeil med mer..).

(c) Genetikere rundt 1900 observerte et populasjonsfenomen de kalte "regresjon" – nemlig hvis foreldrene avviker betydelig fra populasjonsgjennomsnittet av et kjennetegn, for eksempel høyde, er det en tendens til at avkommets kjennetegn ligger nærmere populasjonsgjennomsnittet enn foreldrene. Mener du resultatet i **(b)** bekrefter dette? (Dette er den historiske bakgrunnen for at betegnelsen "regresjon" er knyttet til modeller av den typen som er brukt her.)

---

[1] Dette virker som en selvmotsigelse siden *x* vitterlig også er et resultat av den tilfeldige trekningen. At vi likevel kan tolke *x* som om den var valgt (og ikke trukket) på forhånd skyldes egenskaper ved simultane (og såkalte betingete) sannsynlighetsfordelinger som vil bli utdypet og presisert senere i Stat2 og økonometri.

[2] Det er her den implisitte forutsetningen om *representativitet* kommer inn. Tolkningen er berettiget hvis utvalget er representativt. Utvalget er *representativt* hvis det kan argumenteres at utvalget er trukket *som om* alle mulige utvalg på *n* kvinner i Norge hadde samme sannsynlighet for å bli trukket ut.