HG
April 2013

Ekstraoppgave 3

Tabellen viser kornavling (i 1000 tonn) for bygg og hvete i Akershus (med Oslo) for perioden 2001-2008 (kilde: SSB).

År	Hvete	Bygg
2001	46.2	101.6
2002	44.7	90.6
2003	65.2	89.9
2004	78.9	91.4
2005	75.2	87.9
2006	68.3	68.6
2007	85.4	68.5
2008	93.5	68.8

Slike data der indeksen er relatert til tiden, kalles tidsserier. La variabelen for tid være *t* med verdier $1,2,\ldots,8$ for årene 2001,…,2008 h.h.v.

La y_t betegne bygg-avling for år *t*, $t = 1,2,\ldots,8$. Vi ønsker å beregne mkv trend-linje[1] for bygg, $\hat{y} = a + bt$. Variabelen *t* spiller altså rollen som *x* i "Notat til kapittel 4", med observasjoner, $t_1 = 1, t_2 = 2,\ldots,t_8 = 8$, og dataene har formen, $(t_1, y_1),(t_2, y_2),\ldots,(t_8, y_8)$ eller $(1, y_1),(2, y_2),\ldots,(8, y_8)$.

A. Lag et spredningsplott (kalt tidsserieplott i denne sammenhengen) for *y* med hensyn på *t*. (Det er vanlig – men ikke nødvendig- å trekke linjer mellom observasjonspunktene i et slik tidsserieplott.)

B. Beregn de fem grunnleggende størrelsene, $\overline{t}, \overline{y}, s_t^2, s_y^2, s_{ty}$ som i notatet. [Benytt regel 1 i notatet. Hvis du mister tålmodigheten med kalkulator, er det lov å bruke Excel.] Beregn også den empiriske korrelasjonskoeffisienten, *r*, som altså her (jfr. notatet) ikke har naturlig tolkning som korrelasjonskoeffisient siden verdiene av *t* ikke er tilfeldige. Hvor stor del av variasjonen i *y* blir forklart av trendlinja?

C. Beregn a og b i trend-linja og tegn inn trendlinja i tidsserieplottet ditt.

D. Lag også spredningsplott (tidsserieplott) for hveteproduksjonen m.h.p. *t*. Hvis du bruker Excel, be Excel om å legge inn trendlinja (som anvist i notatet).

[1] Kalt "deterministisk trend" i tidsserie-litteraturen.

E. Lag et spredningsplott for bygg-avlingene med hensyn på hvete-avlingene (dette er ikke noe tidsserieplott!), og beregn den empiriske korrelasjonskoeffisienten, *r*.

[**Merknad.** Du vil antakelig finne en sterk negativ korrelasjon her. Så spørsmålet oppstår om denne korrelasjonskoeffisienten gir god mening. Her bør vel alt være i skjønneste orden siden begge variable gir opphav til tilfeldige observasjoner? Svaret er dessverre nei. Problemet er at det mellom tidsserier som inneholder trender kan oppstå sterke korrelasjoner (dvs sterke statistiske sammenhenger) uten at det behøver å være noen som helst kausal sammenheng mellom dem. Slike sammenhenger kalles i litteraturen for "spuriøse" (fra engelsk "spurious") og er årsaken til at tidsserieanalyse hadde dårlig rykte blant økonomer/økonometrikere tidligere. Dette varte til 70-tallet omtrent da man greide å slå hull på byllen og fant måter å kontrollere for slike fenomener (et stikkord her er "kointegrasjonsanalyse"). Siden da har tidsserier med trender vært et hot tema i økonometri.]

Ekstraoppgave 4

Du trenger modulen "Data analysis" for å løse oppgaven. Sjekk at Data analysis ligger på data-menyen. Hvis ikke, må den legges til ("add in"): I siste Excel versjon: Start fra "office button" (en sirkel øverst til venstre på Excel-arket i Excel 2007. I Excel 2010 klikker du på "file" i stedet.) Klikk så på "excel options" helt nederst på menyen som kommer fram. Og videre:

file (eller office button i Excel2007) → excel options → add-ins → marker "Analysis toolpack" → Klikk "Go.." → merk av "Analysis toolpack" → klikk OK.

(I eldre Excel: Fra menyen: tools → add-ins → merk av "Analysis toolpack" → klikk OK.)

På forelesningen 26. februar 2013 ble det samlet inn 36 kvinnehøyder fra studentene. Hver ble bedt om også å oppgi mors høyde. La *x* betegne "mors høyde" og *y* "datters høyde". Vi ønsker bl.a. å undersøke i hvilken grad datters høyde kan forklares av mors høyde (i gjennomsnitt). Vi ønsker også å kombinere informasjonen fra disse dataene med informasjonen fra tilsvarende datasett samlet inn i 2010 - 2013 – til sammen 161 observasjonspar.

1) Last ned kvinnehøydedataene i en Excel-fil fra fra http://folk.uio.no/haraldg/. Lag et nytt datasett i Excel-fila der alle dataene fra 2010-2013 er slått sammen.

2) Beregn deskriptive størrelser (med "Descriptive Statistics" fra "Data analysis") for både *x* og *y*. Lag også et spredningsdiagram for *y* med hensyn på *x*.

3) Beregn $\overline{x}, \overline{y}, s_x^2, s_y^2, s_{xy}$ [Fås fra "Descriptive Statistics" eller fra "Covariance" [2], begge rutiner i "Data analysis"]. Beregn den empiriske korrelasjonskoeffisienten, *r*, mellom *x* og *y*. Under en viss forutsetning kan *r* betraktes som et anslag (estimat) på den ukjente populasjons-korrelasjonskoeffisienten, ρ. Hvilken forutsetning er det? Spiller det noen

[2] En liten modifikasjon: "Covariance" gir ikke eksakt s_x^2, s_y^2, s_{xy}, men s_x^2, s_y^2, s_{xy} multiplisert med $(n-1)/n$ som innebærer at Excel der deler summene på *n* istedenfor $n-1$. For å få fram "våre" s_x^2, s_y^2, s_{xy}, bør man altså multiplisere tallene i "Covariance" med $n/(n-1)$. s_x^2 og s_y^2 fra "Descriptive Statistics" (under navnet "sample variance") derimot er definert som hos oss.

rolle for *r* om vi bruker Løvås-varianten av s_x^2, s_y^2, s_{xy} med $n-1$ i nevneren, eller om man bruker Excel-varianten fra "Covariance" med *n* i nevneren?

4) Beregn a, b, SS_T, SS_R, SS_E fra de fem verdiene i 3). Og tegn inn mkv-regresjonslinja i spredningsdiagrammet (jfr. bruksanvisning i notatet til kapittel 4). Hvor mange % av variasjonen i døtrenes høyde er forklart av mødrenes høyde?

5) Kjør Excels regresjonsrutine ("regression" i "Data analysis") og identifiser a, b, SS_T, SS_R, SS_E i utskriften.

6) Det er velkjent at gjennomsnittshøyden for mennesker har økt jevnelig opp igjennom historien. For eksempel gjennomsnitthøyden for menn har økt anslagsvis 8-9 cm fra ca 1900. Er det noe i data som tyder på at gjennomsnittshøyden i kvinne-populasjonen fortsetter å øke i vår tid? Eller mener du kanskje at utvalget er for lite til at man kan besvare det spørsmålet? Diskuter. (Det kreves ikke noe presist svar her - som uansett ville kreve mer statistisk teori fra kapittel 6 og muligens Stat 2 – bare litt sunn fornuft og intuisjon.)