UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1110 — Statistiske metoder og dataanalyse - **Løsningsforslag**
Eksamensdag:	4. desember - 2024
Tid for eksamen:	15.00 – 19.00.
Oppgavesettet er på 4 sider.	
Vedlegg:	Ingen
Tillatte hjelpemidler:	Godkjent kalkulator Formelsamling for STK1110

Kontroller at oppgavesettet er komplett før
du begynner å besvare spørsmålene.

Oppgave 1

(a) En finner momentestimatoren ved å løse ligningen $\bar{X} = \frac{\tilde{\alpha}}{\tilde{\alpha}+1}$ hvilket gir $\tilde{\alpha} = \frac{\bar{X}}{1-\bar{X}}$.

(b) For $Y_{ri} = X_i^r$ blir ligningen $\bar{Y}_r = \frac{\tilde{\alpha}_r}{\tilde{\alpha}_r+r}$ hvilket gir $\tilde{\alpha}_r = \frac{r\bar{Y}_r}{1-Y_r} = \frac{r\sum_{i=1}^n X_i^r}{n-\sum_{i=1}^n X_i^r}$

Det er altså ikke en entydig momentestimator for α når vi betrakter ulike $Y_{ir} = X_i^r$ og det uklart hvilken av disse estimatorene som er best (minst varians, f.eks.)

(c) Likelihooden for X_i-ene blir $L(\alpha) = \prod_{i=1}^n \alpha X_i^{\alpha-1} = \alpha^n (\prod_{i=1}^n X_i)^{\alpha-1}$.

Dermed får vi log-likelihood $l(\alpha) = n\ln(\alpha) + (\alpha-1)\sum_{i=1}^n \ln(X_i)$ som har derivert

$$l'(\alpha) = \frac{n}{\alpha} + \sum_{i=1}^n \ln(X_i)$$

Ved å løse $l'(\hat{\alpha}) = 0$ finner vi maximum likelihood estimatoren (MLE) $\hat{\alpha} = \frac{n}{-\sum_{i=1}^n \ln(X_i)}$ for α. Man sjekker at dette er et (entydig) maksimum ved at $l''(\alpha) < 0$ for alle α (se neste punkt).

(d) Når n er tilstrekkelig stor er MLE $\hat{\alpha}$ tilnærmet normalfordelt med forventning α og varians gitt som invers av forventet informasjon $I_n(\alpha)$ for hele datasettet $X_1, X_2, \ldots, X_n$. Vi finner denne informasjonen som

$$I_n(\alpha) = -\mathrm{E}(l''(\alpha)) = -\mathrm{E}(-\frac{n}{\alpha^2}) = \frac{n}{\alpha^2}$$

altså blir variansen i den tilnærmede fordelingen for $\hat{\alpha}$ lik $1/I_n(\alpha) = \frac{\alpha^2}{n}$

(Fortsettes på side 2.)

(e) Vi har

$$0.95 \approx \mathrm{P}(-1.96 < Z = \frac{\hat{\alpha} - \alpha}{\hat{\alpha}/\sqrt{n}} < 1.96) = \mathrm{P}(\hat{\alpha} - 1.96\hat{\alpha}/\sqrt{n} < \alpha < \hat{\alpha} + 1.96\hat{\alpha}/\sqrt{n})$$

Dermed blir $\hat{\alpha} \pm 1.96\hat{\alpha}/\sqrt{n}$ et tilnærmet 95% konfidensintervall for α.

Med $n = 20$ og $\sum_{i=1}^{n} \ln(X_i) = -9.82$ blir $\hat{\alpha} = \frac{20}{9.82} = 2.03$.

Dette gir at 95% konfidensintervallet $2.03 \pm 1.96 * 2.03/\sqrt{20} = [1.14, 2.93]$

(f) Man kan gi konklusjonen på en tosidig test som $\mathrm{H}_0 : \alpha = 1$ mot $\mathrm{H}_0 : \alpha \neq 1$ med angitt nivå ved å sjekke om $\alpha = 1$ er innenfor intervallet.

I vårt tilfelle er $\alpha = 1$ ikke innenfor intervallet og vi forkaster da nullhypotesen på 5% nivå (siden vi sjekket med et 95% intervall).

For at et tilsvarende konfidensintervall skal inneholde $\alpha = 1$ må vi ha $\hat{\alpha} - z\hat{\alpha}/\sqrt{20} = 1$ for egnet kvantil z i $\mathrm{N}(0, 1)$, dvs. at vi må ha $z = (\hat{\alpha} - 1)\sqrt{20}/\hat{\alpha} = 2.28$. Fra tabellen finner vi da at P-verdien ligger i intervallet [0.02,0.03] siden testen er to-sidig.

Sannsynlighet	0.001	0.005	0.010	0.015	0.020	0.025	0.050	0.100	0.200
Kvantil q_γ	-3.09	-2.58	-2.33	-2.17	-2.05	-1.96	-1.64	-1.28	-0.84

Oppgave 2

(a) Vi har at $\mathrm{E}(Y_i) = \beta_0 + \beta_1 x_i$ og dermed dessuten at $\mathrm{E}(\bar{Y}) = \beta_0 + \beta_1 \bar{x}$. Da får vi videre at

$$\begin{aligned}
\mathrm{E}(\hat{\beta}_1) &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})\mathrm{E}(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i - \beta_0 - \beta_1 \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\beta_1 \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1
\end{aligned}$$

og $\hat{\beta}_1$ er altså forventningsrett for β_1.

Fra den andre representasjonen $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ fås at

$$V(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 V(Y_i)}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} = \frac{\sigma^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(b) Modellens estimerte parametre er $\hat{\beta}_0 = -2.29, \hat{\beta}_1 = 0.0314$ og $\hat{\sigma} = 0.197$. Dette betyr at når høyden endres med 1 cm så endres forventet FVC med 0.0314. Hvis vi ekstrapolerer til høyde = 0 cm (ikke så realistisk) så predikerer modellen en negativ (-2.29) forventet FVC (heller ikke spesielt realistisk). Standardavviket σ til Y_i-ene estimeres med $\hat{\sigma} = 0.197$.

t-verdien for høyde beregnes som $t_1 = \hat{\beta}_1/se_1 = 29.74$ der $se_1 = \hat{\sigma}/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0.001057$. Under nullhypotesen $\mathrm{H}_0 : \beta_1 = 0$ skal denne være trukket fra en t-fordeling

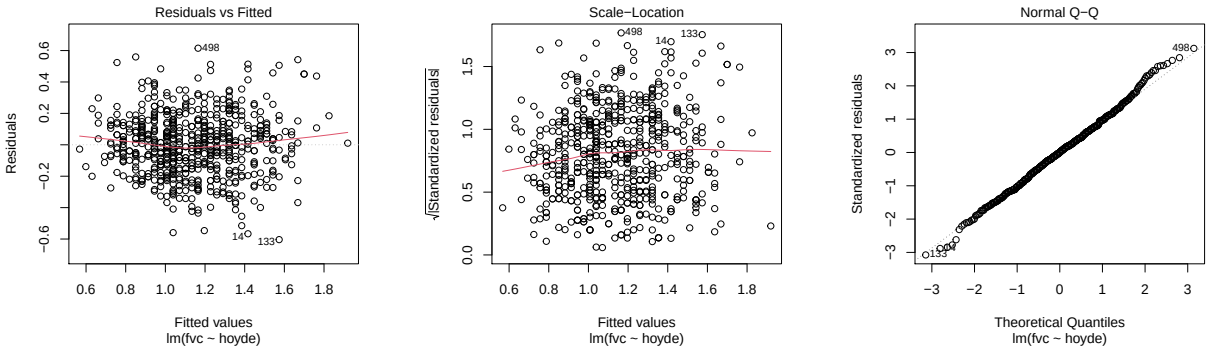
(Fortsettes på side 3.)

med 598=600-2 frihetsgrader. En så stor t-verdi gir en soleklar forkastning på H_0 også med tosidig alternativ $H_a : \beta_1 \neq 0$.

(c) Tilpassede (predikerte) verdier $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ og residualene $e_i = y_i - \hat{y}$. For å sjekke modellantagelsene grafisk lager vi gjerne

- spredningsplott av $(\hat{y}_i, e_i)$: Hvis man ser trender som først positive, deretter negative og tilslutt positive residualer (evt. omvendt) tyder det på ikke-linearitet
- spredningsplott $(\hat{y}_i, |e_i|^{0.5})$: Hvis man f.eks. ser en økende (avtagende) trend i disse tyder det på ikke-konstant varians
- qq-plot over sorterte e_i mot kvantiler i $N(0, 1)$: Hvis disse ikke avtegner en rett linje tyder dette på avvik fra normalitet

Her er det få problemer med modellantagelsene, plottene er omtrent som man ønsker, men en kan ane en svak ikke-linearitet ved linja gjennom det første plottet. Muligens er det også litt mindre varians for de aller minste predikerte verdiene, men neppe grunn til bekymring for om at variansestimater er rimelige.



(d) Vi finner estimerte regresjonsparametre ved å partiellderivere kvadratsummen $g(b_0, b_1, b_2) = \sum_{i=1}^{n}(Y_i - b_0 - b_1 x_{i1} - b_2 x_{i2})^2$ mhp hhv b_0, b_1, b_2 og løse ligningene der disse er satt lik null. De partiellderiverte blir

$$\frac{\partial g(b_0,b_1,b_2)}{\partial b_0} = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 x_{i1} - b_2 x_{i2})$$

$$\frac{\partial g(b_0,b_1,b_2)}{\partial b_1} = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 x_{i1} - b_2 x_{i2})x_{i1}$$

$$\frac{\partial g(b_0,b_1,b_2)}{\partial b_2} = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1 x_{i1} - b_2 x_{i2}))x_{i2}$$

Dette gir da tre lineære ligninger med tre ukjente $\hat{\beta}_0, \hat{\beta}_1$ og $\hat{\beta}_2$

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}$$

$$\sum_{i=1}^{n} Y_i x_{i1} = \hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} x_{i1}$$

$$\sum_{i=1}^{n} Y_i x_{i2} = \hat{\beta}_0 \sum_{i=1}^{n} x_{i2} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} x_{i2} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}^2$$

(Fortsettes på side 4.)

Med vektoren av responser $Y = (Y_1, Y_2, \cdots, Y_n)'$ og designmatrisen

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix}$$

kan vi skrive om disse ligningene til matriseformen $X'X\hat{\beta} = X'Y$ som har løsning $\hat{\beta} = (X'X)^{-1}X'Y$.

For at denne løsningen skal være entydig må $X'X$ være inverterbare som er det samme som at X har full rang, ingen kolonner kan skrives som linærkombinasjoner av andre kolonner.

(e) I den enkle lineære regresjon var $\hat{\beta}_1 > 0$ og i de nye analysen har både høyde og vekt positive $\hat{\beta}_j$. Høyde og vekt er innbyrdes positivt korrelert. Den estimerte effekten av høyde alene (i den enkle lineære regresjonen) er da en kombinert effekt av høyde pluss en "mediert" effekt av vekt i den multiple regresjonen. Det er dermed rimelig at estimatet $\hat{\beta}_1$ i den enkle lineære regresjonen er høyere enn tilsvarende $\hat{\beta}_1$ i den multiple lineære regresjonen.

Størrelsen `Multiple R-squared` gis ved $R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$. Telleren $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ er knyttet til variasjonen (generelt) og inngår i variansestimatet for de aktuelle modellene, mens nevneren tilsvarende refererer til variasjonen (variansen) over y_i-ene uten å ta hensyn til forklaringsvariable. Derfor kan man oppfatte R^2 forklart andel varians.

Siden kvadratsummer må bli mindre når vi inkluderer en ny forklaringsvariabel må R^2 bli større. Her ser vi riktignok bare en moderat økning fra 0.5966 til 0.6154, så selv om det er klart signifikant betydning av både høyde og vekt i den simultane modellen er forklaringsgraden av å legge inn vekt ikke spesielt stor.

For enkel lineær regresjon har vi at R^2 bare er kvadratet av korrelasjonen mellom x_i-er og y_i-er. Men siden $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ bare er en lineær transformasjon av x_i-ene vil den også være korrelasjonen mellom $\hat{y}_i$-er og y_i-er.

Denne siste karakteriseringen av R^2 gjelder også for multippel lineær regresjon.

SLUTT