UNIVERSITETET I OSLO

Det matematisk-naturvitenskapelige fakultet

Eksamen i:	STK1110 — Statistiske metoder og dataanalyse *Løsningsforslag*
Eksamensdag:	Tirsdag 18. desember 2018
Tid for eksamen:	09.00 – 13.00
Oppgavesettet er på 5 sider.	
Vedlegg:	Tabell over normalfordelingen
Tillatte hjelpemidler:	Formelsamling for STK1100/STK1110. Godkjent kalkulator.

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

Oppgave 1

a

Siden $\mathrm{E}[X] = np$ fås $\mathrm{E}[\hat{p}] = \frac{\mathrm{E}(X)}{n} = \frac{np}{n} = p$, dvs. $\hat{p}$ er forventningsrett for p

Videre has $\mathrm{V}(\hat{p}) = \frac{\mathrm{V}(X)}{n^2} = \frac{np(1-p)}{n} = \frac{p(1-p)}{n}$.

b

Likelihooden er sannsynligheten (eller tettheten) til observasjonene, her X, innsatt observasjonene og sett på som en funksjon av den (de) ukjente parametrene. Siden $\mathrm{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$ når $X \sim \mathrm{Bin}(n, p)$ blir $L(p) = \binom{n}{X} p^X (1-p)^{n-X}$.

Å maksimere $L(p)$ er ekvivalent med å maksimere log-likelihood $l(p) = \ln(L(p)) = \ln(\binom{n}{X}) + X\ln(p) + (n-X)\ln(1-p)$. Vi har $l'(p) = X/p - (n-X)/(1-p)$ og denne satt lik null gir en ligning med løsning $\hat{p} = X/n$.

Fra generell likelihoodteori has at $\hat{p}$ er tilnærmet normalfordelt med forventning p og varians gitt ved $1/I_n(p) = -1/\mathrm{E}[l''(p)]$. Siden $-l''(p) = X/p^2 + (n-X)/(1-p)^2$ som har forventning $-\mathrm{E}[l''(p)] = np/p^2 + n(1-p)/(1-p)^2 = n(1/p + 1/(1-p)) = n/(p(1-p))$ fås (tilnærmet) varians for $\hat{p}$ lik $p(1-p)/n$ som er det samme som ble utledet i punkt a.

c

Det er grunn til å tvile på nullhypotesen $\mathrm{H}_0 : p = p_0$ hvis det er stor forskjell mellom $\hat{p}$ og p_0. Vi må ta hensyn til usikkerheten i $\hat{p}$ og derfor er det rimelig å forkaste hvis $|Z(p_0)| = |\hat{p} - p_0|\sqrt{n}/(p_0(1-p_0))$ er stor. Siden 1.96 er 97.5% persentilen i standardnormalfordelingen vil da, under nullhypotesen, $\mathrm{P}(|Z(p_0)| > 1.96|\mathrm{H}_0) = 0.05 = \alpha$ som er testen fastsatte nivå. Dette betyr

(Fortsettes på side 2.)

at vi har en test med nivå 0.05 for den aktuelle situasjonen ved å forkaste hvis $|Z(p_0)| = |\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}| > 1.96$.

Med $n = 50$ og $X = 15$ blir $\hat{p} = 0.3$ og dermed blir testobservatoren $Z(0.5) = -2.828 < -1.96$, så $\mathrm{H}_0 : p = 0.5$ forkastes på 5% nivå. Testens (tilnærmede) P-verdi blir $2\mathrm{P}(Z < -2.83) = 2 * 0.0023 = 0.0046$ når Z er standard normalfordelt.

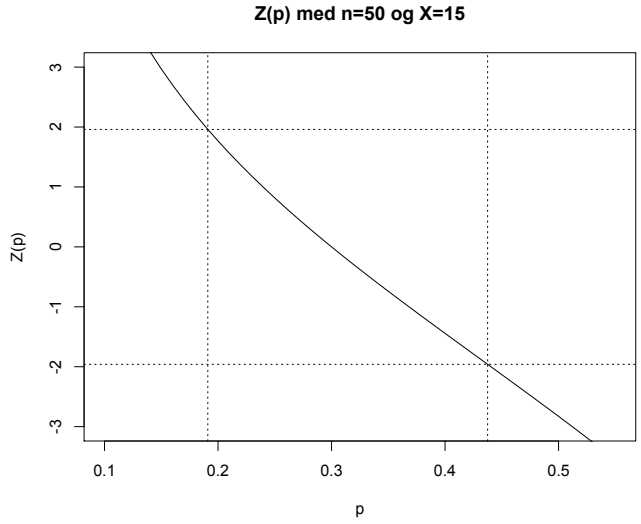
d

En generell sammenheng mellom hypotesetesting (med tosidig alternativ) og konfidensintervall gir at mengden

$$\{\theta_0 : \mathrm{H}_0 : \theta = \theta_0 \text{ ikke forkastes med nivå } \alpha\}$$

er et $(1-\alpha)100\%$ konfidensintervall for θ. I det aktuelle tilfellet er dette $\{p_0 : |Z(p_0)| < 1.96\}$, så dette intervallet blir dermed et tilnærmet 95% konfidensintervall for p.

Med tallene $X = 15$ og $n = 50$ i punkt c får vi følgende plott av $Z(p)$ mot p:



95% konfidensintervallet kan avleses som (0.19,0.44) (ca.).

Vi har at

$$|Z(p_0)| < 1.96 \Leftrightarrow (\hat{p}-p)^2 n < 1.96^2 p(1-p) \Leftrightarrow p^2 - 2\hat{p}p + \hat{p}^2 < 1.96^2 p/n - 1.96^2 p^2/n$$

som er en 2. gradsuliket i p.

e

Løsningen på 2.gradsuliketen er oppgitt som

$$\frac{\hat{p} + 1.96^2/2n}{1 + 1.96^2/n} \pm \frac{1.96}{\sqrt{n}} \frac{\sqrt{\hat{p}(1-\hat{p}) + 1.96^2/4n}}{1 + 1.96^2/n}$$

Hvis vi fjerner ledd av typen a/n finner vi at intervallet kan tilnærmes med $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$.

(Fortsettes på side 3.)

Med $\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})}}\sqrt{n}$ tilnærmet standardnormalfordelt har vi

$$0.95 \approx \mathrm{P}(-1.96 < \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})}}\sqrt{n} < 1.96).$$

Men $-1.96 < \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})}}\sqrt{n} < 1.96 \Leftrightarrow -1.96\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} < \hat{p}-p < 1.96\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$ som igjen er ekvivalent med $\hat{p} - 1.96\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} < p < \hat{p} + 1.96\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$, dvs. $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$ er et tilnærmet 95% konfidensintervall.

Oppgave 2

a

Estimatene er $\hat{\beta}_0 = 30.511906, \hat{\beta}_1 = -0.082898$ for β_0 og β_1 og $s = 2.983$ for σ. Det første parameterestimatet $\hat{\beta}_0$ er et estimat for den forventede reduksjonen når $x_{i1} = 0$ (men siden alle målte temperaturer er betraktelig høyere bør denne verdien fortolkes med forsiktighet). Videre er $\hat{\beta}_1$ et estimat for hvor mye reduksjonen endres når x_{i1} endres med en enhet. Endelig er s et estimat for standard avviket σ for feilleddet ε_i og for variationen i rundt regresjonslinja.

Minste kvadraters estimatene $\hat{\beta}_0$ og $\hat{\beta}_1$ bestemmes som de verdiene (b_0, b_1) som minimerer kvadratsummen $\sum_{i=1}^{n}(Y_i - b_0 - b_1 x_{1i})^2$

t verdien for $\hat{\beta}_1$ beregnes som $t_1 = \hat{\beta}_1/se(\hat{\beta}_1) = -0.082898/0.007515 = -11.03$ der standardfeilen $\hat{\beta}_1$ er gitt som $0.007515 = se(\hat{\beta}_1)$ $(= s/\sqrt{\sum_{i=1}^{n}(x_{i1} - \bar{x_1})^2}.)$ P-verdien for Temperatur beregnes som $2\mathrm{P}(T > 14.87)$ der $14.87 = |t_1|$ og T er t-fordelt med $n - 2 = 126$ frihetsgrader.

Siden dette er en enkel lineær regresjonsmodell har vi at R^2 er lik kvadratet av korrelasjonen mellom Y_i og forklaringsvariablen x_{i1}, derfor blir korrelasjonen $-\sqrt{R^2} = -\sqrt{0.4913} = -0.701$ der minustegnet følger siden $\hat{\beta}_1$ er negativ.

b

Grunnen til at $\hat{\beta}_1$ har samme verdi i den enkle lineære regresjonen i punkt a) og den multiple lineære regresjonen i dette punktet er at designet er balansert, vi har like mange observasjoner av hver verdi av x_{i1} for hver verdi x_{i2}. Da vil x_{i1} ikke gi noen informasjon om x_{i2} og de to forklaringsvariablene må være ukorrelerte. Når vi har to ukorrelerte forklaringsvariable vil det å utelate en av dem ikke endre estimert regresjonskoeffisient for den andre.

Selv om det ikke var endring i $\hat{\beta}_1$ er det likevel endring i estimatet s for σ og tilsvarende for σ^2 estimert som $s^2 = \frac{1}{n-3}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ der $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$ er den predikerte verdien ved å bruke begge forklaringsvariable. Dette estimatet er mindre enn s^2 fra den enkle lineære regresjonen siden begge forklaringsvariable er viktige prediktorer for Y_i.

Dette er også grunnen til at den nye $R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$ er større enn den i den enkle lineære regresjonen.

(Fortsettes på side 4.)

Standardfeilen se_1 for $\hat{\beta}_1$ er proporsjonal med s og derfor har dette anslaget blitt mindre i denne situasjonen der forklaringsvariablene er ukorrelert. Dette igjen førte til at t-verdien $\hat{\beta}_1/se_1$ hadde større avvik fra 0.

Siden forklaringsvariablenene er ukorrelerte har vi at den multiple R2-verdien kan skrives som $R^2 = r_1^2 + r_2^2$ der r_j er korrelasjonen mellom Y_i-ene og x_{ji}-ene. Derfor fås $r_2^2 = R^2 - r_1^2 = 0.722 - 0.491 = 0.231$. Videre blir da korrelasjonen mellom Y_i-ene og x_{2i}-ene lik $-\sqrt{(}0.231) = -0.48$ (hvor igjen minus følger av at $\hat{\beta}_2 < 0$..

c

Det første plottet viser residualene $e_i = Y_i - \hat{Y}_i$ mot tilpassede verdier $\hat{Y}_i$. Den glattede linjen gjennom punktene $(\hat{Y}_i, e_i)$ viser klar kurvatur. Dette indikerer at transformasjoner av forklaringsvariablene eller inklusjon av kvadratledd x_{i1}^2 og x_{i2}^2 og muligens også interaksjonersledd $x_{i1}x_{i2}$ kan forbedre tilpasningen.

Det andre plottet er et qqplot av de ordnede (standardiserte) residualene e_i^* mot persentiler i standardnormalfordelingen. Når disse punktene ikke ligger nær en rett linje har vi en indikasjon på at feilleddene ikke er normalfordelt. Dette igjen kan indikere at å bruke t-fordelingen for å beregene P-values ikke er optimalt (Likelvel, en nærmere kikk på plottet viser at halene i feilleddenes fordeling er lettere enn halene i normalfordelingen og fra dette perspektivet er ikke avviket veldig alvorlig).

Det tredje plottet viser $\sqrt{|e_i^*|}$ mot $\hat{Y}_i$ og er konstruert for undersøke heteroskedastisitet, dvs. om variansen til ε_i avhenger av $\mathrm{E}[Y_i]$. Selv om kurven ikke er helt rett har den verdier mellom 0.7 og 1.2 hvilket ikke er betraktet å være så veldig mye så avviket fra konstant varians er neppe alvorlig.

Dette betyr at den delen av modellen som må forbedres er den lineære strukturen ved å inkludere kvadratledd eller transformere forklaringsvariable.

Oppgave 3

a

Siden $SSE/\sigma^2 = (n-k-1)s^2/\sigma^2 \sim \chi^2_{n-k-1}$ blir

$$\mathrm{E}[s^2] = \sigma^2/(n-k-1)\mathrm{E}[(n-k-1)s^2/\sigma^2] = (\sigma^2/(n-k-1))(n-k-1) = \sigma^2,$$

så s^2 er forventningsrett.

Dessuten blir $V(s^2) = (\sigma^4/(n-k-1)^2)\mathrm{V}((n-k-1)s^2/\sigma^2) = (\sigma^4/(n-k-1)^2)2(n-k-1) = 2\sigma^4/(n-k-1)$ som går mot 0 når $n-k$ går mot uendelig. Ved Tsjebysjeff's ulikhet får vi dermed at for alle $\epsilon > 0$ så vil

$$\mathrm{P}(|s^2 - \sigma^2| > \epsilon) < \frac{\mathrm{V}(s^2)}{\epsilon^2} = \frac{2\sigma^4}{(n-k-1)\epsilon^2} \to 0$$

når $n-k$ går mot uendelig. Dette betyr at s^2 konvergerer (i sannsynlighet) mot σ^2, dvs. er konsistent for σ^2.

(Fortsettes på side 5.)

Når $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ blir $\mathrm{E}[Y] = \beta_0 = \mu$, dvs. alle forventninger blir like. Dermed er $\mathrm{SST}/(n-1) = \sum_{i=1}^{n}(Y_i - \bar{Y})^2/(n-1)$ en forventningsrett estimator for σ^2. Siden også $(n-1)\mathrm{SST}/\sigma^2 \sim \chi^2_{n-1}$ under forutsetningen vil $\mathrm{V}(\mathrm{SST}/(n-1)) = 2\sigma^4/(n-1)$ og også gå mot 0 når $n-k$ og dermed n går mot uendelig.
Ved å benytte Tsjebysjeff igjen blir også $\mathrm{SST}/(n-1)$ konsistent for σ^2.

b

Vi får

$$\mathrm{R}^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{n-k-1}{n-1}\frac{\mathrm{SSE}/(n-k-1)}{\mathrm{SST}/(n-1)} \to 1 - (1-\rho) = \rho$$

siden $\frac{\mathrm{SSE}/(n-k-1)}{\mathrm{SST}/(n-1)} \to \frac{\sigma^2}{\sigma^2} = 1$ ved å bruke hintet og $(n-k-1)/(n-1) \to 1-\rho$ når $n \to \infty$ og $k/n \to \rho$.

Tilsvarende for justert R^2 definert som $\mathrm{R}^2_{adj} = 1 - \frac{\mathrm{SSE}/(n-k-1)}{\mathrm{SST}/(n-1)}$ får vi at teller og nevner begge går mot σ^2 når n og $n-k$ går mot uendelig. Dermed går brøken mot 1 og R^2_{adj} mot null.

SLUTT