Løsningsforslag INT010 vår 2009

Oppgave 1

a) Overgangsmatrise: Like mange rekker som kolonner og hver rekke består av en sannsynlighetsvektor som summerer seg til 1. P(arbeidsløs$_t$ → i arbeid$_{t+1}$) = **0,4**. P(arbeidsløs$_t$ → i arbeid$_{t+2}$) = 0,6 · 0,4 + 0,4 · 0,7 = 0,24 + 0,28 = **0,52**.

b) Vi finner tilstandsfordelingen etter ett år ved å multiplisere tilstandsfordelingen i utgangspunktet med overgangsmatrisen: $[0{,}8 \quad 0{,}2] \cdot \begin{bmatrix} 0{,}7 & 0{,}3 \\ 0{,}4 & 0{,}6 \end{bmatrix} = [0{,}64 \quad 0{,}36]$. Dvs. etter ett år et **64% i arbeid**, mens **36% er arbeidsløs**.

c) Fra ligningssystemet $\boldsymbol{\pi} \cdot \mathbf{P} = \boldsymbol{\pi}$ får vi to ligninger (den stasjonære fordelingen er bestemt at dette systemet): $0{,}7\,\pi_1 + 0{,}4\,\pi_2 = \pi_1$ og $0{,}3\,\pi_1 + 0{,}6\,\pi_2 = \pi_2$. Første ligning gir: $\pi_2 = 0{,}75\,\pi_1$. Innsetter dette i $\pi_1 + \pi_2 = 1$, som gir $\pi_1 + 0{,}75\,\pi_1 = 1$, da blir $\boldsymbol{\pi_1 = 0{,}57}$. Dermed er $\boldsymbol{\pi_2 = 0{,}43}$. På lang sikt er altså 57% i arbeid, mens 43% arbeidsløse.

Oppgave 2

a) Testobservator $t = (x_1 - x_2) / \sqrt{[s_p^2\,(1/n_1 + 1/n_2)]} = (60{,}9 - 70{,}2) / \sqrt{[\,66{,}7 \cdot (1/27 + 1/27)]} =$ **-4,18**. Kritisk grense på 5% nivå er $t < -\,t_{0{,}05,\,52} \simeq -\,\mathbf{t_{0{,}05,\,50} = -\,1{,}676}$. **Forkast H_0** og påstå at ledere på østsiden har lavere lønn enn ledere på vestsiden.

b) $\mathbf{E(T)} = n_1\,(n_1 + n_2 + 1) / 2 = 27\,(27 + 27 + 1)/\,2 =$ **742,5**. $\sigma_T = \sqrt{[\,n_1 n_2 (n_1 + n_2 + 1) / 12\,]} = \sqrt{[729 \cdot 55 / 12]} =$ **57,8**. Testobservator $Z = T - E(T) / \sigma_T = (537{,}5 - 742{,}5) / 57{,}8 =$ **-3,5**. Kritisk verdi er $Z < -\,z_{0{,}05} = -\,1{,}645$. **Forkast H_0**. p-verdi = $P(Z < -\,3{,}5) \simeq 0{,}5 - 0{,}5 =$ **0.**

c) **H_0: $\mu_1 = \mu_2 = \mu_3$ vs. H_1: minst et av områdene er forskjellig**. Testobservatøren er F = 21,83 (fra Minitab). Kritisk verdi på 5% nivå er $F > F_{0{,}05,k-1,n-k} = \mathbf{F_{0{,}05,2,78} \simeq 3{,}10}$, siden k = 3 og n = 27*3=81. **Forkast H_0. (evt. på bakgrunn av lav p-verdi).** Mao er det forskjell i lønn i de ulike områdene. Konfidensintervallene for forventningene viser at i område vest har lederne høyere lønn enn i de andre områdene.

d) Hver metode gir oss tre intervaller med en nedre og øvre grense (parvis sammenligning). Om 0 er inneholdt i intervallet, er det ikke signifikant forskjell mellom forventningene til dette paret (dvs. ikke forskjell i forventet lønn). Begge metodene gir oss samme svar (intervallene til Tukeys metode i parentes). For **senter og vest** er forskjellen **signifikant** [8,4; 18,3], **senter og øst** er **ikke signifikant** [-0,87; 9,0], **vest og øst** er **signifikant** forskjellig [-14,2; -4,3].

e) Basert på p-verdiene som er lik null for faktorene område og bransje, vil vi påstå at det eksisterer effekter for hver av disse to faktorene. **Hypotesen om lik forventning blir forkastet.** P-verdien for en test om **ingen samspill** mellom område og jobbtype **beholdes**.

Oppgave 3

a) I en regresjonsmodell kan vi ikke uten videre påstå at en endring i en forklarende variabel vil forårsake en endring i responsen, mao vi kan **ikke** påstå at det eksisterer en **kausal sammenheng.** Skal en regresjonsmodell avdekke en kausal sammenheng må

produksjonskostnadene være uavhengig av feilleddet. Feilleddet vil inneholde alle uobserverbare forhold som påvirker bruttoinntekten, bl.a. markedsføringskostnadene som introduseres i neste regresjon. Vi ser der at koeffisenten til produksjonskostnadene faller vesentlig. Men selv med markedsføringskostnader inkludert kan det være gjenværende utelatte variabler som er korrelert med produksjonskostnadene, for eksempel manuskriptets kvalitet (potensial for suksess). Generelt må man være forsiktig med kausale tolkninger når man står overfor observasjonsdata der variablene er bestemt av optimerende aktører.

b) Prediksjonsintervallet er $\bar{Y} \pm t_{\alpha/2,n-2} s_\varepsilon \sqrt{1+1/n+(x_g-\bar{x})^2/(n-1)s_X^2}$. Det er opplyst at x_g = 9,9. Fra Minitab er s_ε = 9,506. Fra tabell er $t_{0,025,18}$ = 2,101. Fra regresjonsligningen er $\bar{Y} = 5{,}07 + 5{,}53 \cdot 9{,}9 = 59{,}8$. Prediksjonsintervallet er da 59,8 ± 20,8, dvs. **[39,0; 80,6]**.

c) Konfidensintervallet er $\bar{Y} \pm t_{\alpha/2,n-2} s_\varepsilon \sqrt{1/n+(x_g-\bar{x})^2/(n-1)s_X^2}$. Beregningen blir som over, bortsett fra leddet under rottegnet. Konfidensintervallet blir 59,8 ± 6,0, dvs. **[53,8; 65,8]**. Prediksjonsintervallet i b) er videre, fordi å predikere bruttoinntekten for en film er vanskeligere enn å estimere den forventede bruttoinntekten for mange filmer.

d) R^2 er høy, så det betyr at mye av variasjonen i bruttoinntekter kan forklares med variasjon i de tre forklarende variablene. Merk at vi ikke har observasjoner i nærheten av null for de forklarende variablene, så derfor gir en tolkning av konstantleddet liten mening. En økning i produksjonskostnadene med \$1 million vil føre til en økning i gjennomsnittlige bruttoinntekter på **\$2,85 millioner** (en nedgang fra den enkle modellen), mens en økning i markedsføringskostnadene på \$1 million fører til en gjennomsnittlig økning i brutto-inntekter på **\$2,28 millioner**. Om en film er basert på en bok vil gjennomsnittlig økning i bruttoinntekter være **\$7,17 millioner.** Alle disse tre koeffisientene er signifikante (lav p-verdi).

e) H_0: Alle koeffisienter i modellen (unntatt konstantleddet) er lik null
H_A: Minst en forklaringsvariabel (utenom konstantleddet) har koeffisient ulik null
Antall frihetsgrader er k=**3**, og n-k-1=20-3-1=**16** i hhv. teller og nevner. Kritisk verdi på 5% nivå er **3,24**. **Forkast** dermed nullhypotesen.

f) Grafene på venstre side viser at residualene er tilnærmet normalfordelte. Grafen øverst i høyre hjørne viser residualene plottet mot predikert produksjon, og det er ingen tendens at variansen til residualene er økende (dvs. ingen antydning til heteroskedastisitet). Grafen nederst til høyre antyder ingen autokorrelasjon, siden dette er tverrsnittsdata forventer vi heller ikke dette. Mao. er **forutsetningene** for å benytte minste kvadraters metode, samt å gjøre inferens **oppfylt**.

g) Fremgangsmåten er nøyaktig som i b). Først beregnes punktprediksjonen ($\hat{Y}$) ved å innsette de gitte verdier av forklaringsvariablene i regresjonsligningen. Deretter beregnes det estimerte standardavviket til prediksjonsfeilen, og vi finner prediksjonsintervallet som vanlig ved $\hat{Y} \pm 1.96 \cdot S\{Y - \hat{Y}\}$ når vi antar stor n (evt. $t_{0,025,n-k-1}$). [I denne situasjonen er det utregningen av standardavviket som er mer komplisert enn i b), her må nemlig matriseoperasjoner benyttes da vi har flere enn en forklaringsvariabel.]