Løsningsforslag

Oppgavesettet og løsningsforslaget er utarbeidet av Jarle Møen. Takk til Jostein Lillestøl som har bidratt med datasettet og problemstillingen. Det anbefales at sensorene gir inntil 10 poeng per delspørsmål.

a) Vi ser at fordelingen er svært skjevfordelt (høyreskjev). Minimum, 5 %-persentilen og medianen ligger ganske tett på hverandre, mens det er en lang høyrehale. Dette svarer til at responstiden vanligvis er kort, men at det fra tid til annen oppstår flaskehalser som skaper store forsinkelser. Siden normalfordelingen er symmetrisk bør vi være forsiktig med å bruke metoder som krever eksakt normalfordeling.

b) Vi ser at minimumsverdien og medianen forandrer seg svært lite, men at gjennomsnittsverdien og standardavviket faller en del. Variasjonsbredden har falt betydelig og kvartilbredden er nesten halvert. 95 %-persentilen har falt fra 6 til 4. Det ser altså ut til at man har lykkes med å utbedre flaskehalser slik at forsinkelser som skaper irritasjon hos brukerne er blitt mer sjeldent. Kort oppsummert ser fordelingens høyrehale ut til å ha blitt «kortere».

c) $F_{(71,71)}=\sigma_1^2/\sigma_2^2=1.64^2/0,95^2=2,98$. Kritisk grense for denne F-testen med 5 % signifikansnivå og tosidig test 1,60. Ved ensidig test er kritisk grense 1,48. Uansett kan vi klart forkaste en nullhypotese om lik varians og konkludere at variansen til responstiden har gått ned. (Gitt at testens forutsetninger er oppfylte.)

d) F-testen ovenfor forutsetter uavhengige, normalfordelte variabler. Vi har sett at fordelingen til responstiden er ganske skjevfordelt. Følgelig er det fare for at normaltilnærming vil fungere dårlig. Vi kan derfor ikke legge alt for stor vekt på konklusjonen i c). (Merk at F-testen generelt er mer følsom for brudd på normalantagelsen enn T-testen, selv i store utvalg. Dette er derfor en reell bekymring.)

e) Vi ser fra grafen i den deskriptive analysen at trafikken variere sterkt mellom de ulike måletidspunktene. Vi ser også fra korrelasjonsmatrisen at det er klar positiv korrelasjon mellom trafikk og responstid. Ved å matche på måletidspunkt sammenligner vi derfor endring i responstid «betinget» på søketrafikken. Dette kan få den forskjellen i responstid som skyldes programendringen til å tre klarere fram. Trolig er trafikk en så viktig forklaringsvariabel for responstiden at gevinsten ved redusert uforklart variasjon mer enn oppveier tapet i frihetsgrader som matchingen medfører. T-testen for matchede par er derfor å foretrekke. Et argument mot testen for matchede par (konstanteffektmodellen) er at denne forutsetter av de forventede differansene er den samme på alle måletidspunkt. Det er neppe riktig da forbedringene er rettet inn mot tidspunkt med stor trafikk som skaper flaskehalser.

f) Hvis man er trygg på at programendringen ikke kan ha utilsiktede negative effekter bør man bruke ensidig test. Hvis man derimot er redd for at man kan ha gjort programmeringsfeil eller er usikker på effekten av programendringen slik at responstiden kan bli lengre, bør man bruke tosidig test. Man kan alternativt argumentere for ensidig test med utgangspunkt i at det kun er forbedret responstid som er beslutningsrelevant. (Bare da vil man ta i bruk det nye programmet.) Dersom man tolker en evt. negativ effekt som en konsekvens av programmeringsfeil vil man imidlertid neppe skrinlegge prosjektet ved negativt utslag så et argument om at negative utslag ikke er beslutningsrelevant er

ikke vanntett. Argumentasjonen for det ene eller det andre valget er imidlertid viktigere enn selve konklusjonen.

g) T-testene forutsetter at responstiden er normalfordelt. Wilcoxon-testene er fordelingsfrie. I utgangspunktet er derfor Wilcoxon-testene å foretrekke siden vi har sett at responstiden ikke kan være normalfordelt. Siden utvalget er relativt stort vil imidlertid T-testene ha en viss robusthet i forhold til brudd på forutsetningen om normalfordeling, og Willcoxontesten fungerer best når fordelingene har lik form og spredning. Vi har klare indikasjoner på at spredningen er redusert etter programendringen. Valget er derfor ikke helt opplagt.

h) Analyse 3 og 4 tester om medianresponstiden er den samme før og etter reformen. Testene har høye P-verdier og gir over hodet ikke noe grunnlag for å påstå at medianen har forandret seg. Det trenger imidlertid ikke bety at programendringen har vært uten effekt andre steder i fordelingen.

i) Vi må teste om de to andelene er lik hverandre. Testobservatoren er

$$Z = \left((\hat{p}_1 - \hat{p}_2)/\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\hat{p}(1-\hat{p})}\right) = \left(\left(\frac{7}{72}-\frac{1}{72}\right)/\sqrt{\left(\frac{1}{72}+\frac{1}{72}\right)\frac{7+1}{72+72}\left(1-\frac{7+1}{72+72}\right)}\right) =$$

$$\left(0{,}0833/\sqrt{0{,}0278 \cdot 0{,}0556 \cdot 0{,}9444}\right) = 0{,}0833/0{,}0382 = 2{,}18.$$

Kritisk grense for tosidig test på 5 %-nivå er ±1,96. Følgelig kan vi forkaste en nullhypotese om like andeler.

Merk: Testen ovenfor bruker normaltilnærming. For at tilnærmingen skal være god bør *np* og *n(1-p)* begge være større enn 5. Her er *np*=8 så vi tilfredsstiller kravet, men ikke med særlig god margin. (Testen forutsetter også at søk med mer enn fem sekunders responstid er binomisk fordelt. Da må slike forsinkelser opptre uavhengig av hverandre. Siden perioder med høy søketrafikk trolig vedvarer i mer enn fem minutter kan dette problematiseres. I praksis er dette problemet løst ved at vi bruker samme måletidspunkt på de to dagene slik at vi har like mange måletidspunkt med høy og lav trafikk i begge utvalgene.)

j) Vi tar som utgangspunkt at antall forsinkelser, F, er binomisk fordelt (n,p) med n=72 og vårt estimat for p lik de observerte andelene.

Da blir E[0,5(F1+F2)]=0,5(E[F1]+ E[F2])= 0,5(np_1+ np_2)= 0,5(7+1)=<u>4</u>.

Videre får vi at Var[0,5(F1+F2)]=$0{,}5^2$[Var(F1) +Var(F2)]= 0,25[$np_1(1-p_1)$ + $np_2(1-p_2)$]= 0,25[72*7/72(1-7/72) + 72*1/72(1-1/72]= 0,25[7*(65/72) + 1*(71/72] = <u>1,826</u>.

k) Vi ser at forventet responstid er 0,36 sekunder kortere etter programendringen og at denne forskjellen er nær signifikant på 5 %-nivå. Videre ser vi at responstiden avhenger klart av søketrafikken, men at denne sammenhengen er ikke-lineær. Søketiden øker overproporsjonalt med trafikken slik man kan forvente om det «korker seg» når trafikken overstiger visse terskler. Vi ser for øvrig at den lineære komponenten ikke er signifikant. Vi kunne derfor vurdere å droppe denne fra analysen. Forklaringsgraden er brukbar med en justert R^2 på 36 %.

l) Trafikk inngår som et annengradspolynom. Når vi øker *trafikk* med én enhet forandrer derfor både *trafikk* og *trafikk_sq* seg. Effekten vil derfor variere avhengig av verdien på *trafikk*. Deriverer vi regresjonsligningen med hensyn på trafikkvariabelen får vi

$$\frac{\partial responstid}{\partial trafikk} = \beta_{trafikk} + 2\beta_{sq\_trafikk} \cdot trafikk = -0.109611 + 0{,}01783342 \cdot trafikk$$

Dersom *trafikk* for eksempel øker med én fra sitt gjennomsnitt på 14,31 øker predikert responstid med 0,146.

Man kan også bruke den estimerte ligningen direkte og få

dr = 2.10-0.36*d-0.11*(t+1)+0.01*(t+1)*(t+1)
  -(2.10-0.36*d-0.11*t+0.01*t*t
  = 0+0-0.11*1+0.01*(2t+1)
  = -0.10+0.02*t
og t = 14.31 => dr = 0.155 (avviket fra svaret over skyldes avrundinger)

m) Minste kvadraters metode forutsetter konstant varians (homoskedastisitet). Hvis variansen er mindre etter reformen (når forklaringsvariabelen *dag*=1) har vi brudd på denne forutsetningen (heteroskedastisitet). Koeffisientene er fortsatt forventningsrette, men estimatene kan i teorien forbedres og inferensen er ikke gyldig.

n) Vi ser at Durbin-Watson observatoren er 1,173. Dette er en god del lavere enn 2 og indikerer dermed positiv autokorrelasjon i residualene. Slår vi opp i tabell finner vi at 5 % kritiske grenser for DW med n = 144 og k = 3 er dL = 1,6854 og dU = 1,7704. Vi kan dermed forkaste en nullhypotese om ingen autokorrelasjon. (Merk at tabellen i boka bare går til 100 med dL = 1,61 og dU = 1,74, men konklusjonen blir den samme.)

o)

(i) Tilfeldig gang er ikke en god modell. Selv om det muligens kan være sterk tidsavhengighet i responstiden kan ikke responstiden i systemet variere fritt og verdien på tidspunkt *t* er neppe den beste prediksjonen for alle framtidige tidspunkt.

(ii) Hvit støy kan være en bedre tilnærming siden denne prosessen har konstant forventning og varians og gjennomsnittlig responstid kan være en brukbar prediksjon for framtidige responstider. Men det er neppe uavhengighet i responstiden siden trafikken varierer systematisk gjennom dagen og det påvirker responstiden.

(iii) En AR(1) prosess er både stasjonær og tillater avhengighet i responstiden over tid slik at responstiden på tidspunkt *t+1* kan avhenge av responstiden på tidspunkt *t*. Dette synes klart å være den mest realistiske modellen blant disse tre, gitt det vi vet om prosessen.