Preliminära lösningar till tentamensskrivning på kursen Tillämpad statistik A5 (15hp) 2014-08-26 /BW, AG, RP

Uppgift 1

A) $b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{10}{10} = 1$, $a = \bar{Y} - b\bar{X}$=8-1*3=5

B) e_1=6-(5+1*1)= 0, e_2=9-(5+1*2)=2, e_3=4-(5+1*3)=-4, e_4=11-(5+1*4)=2, e_5=10-(5+1*5)= 0

SSE=2^2+$(-4)^2$+2^2=24

C) SST =$\sum(Y - \bar{Y})^2$ = $(-2)^2$+1^2+$(-4)^2$+3^2+2^2= 34

D) R^2=1 – SSE/SST =1 – 24/34 = 0,294 (29,4%)

E) $s_b = \frac{s_e}{\sqrt{\sum(x_i - \bar{x})^2}} = \frac{\sqrt{24/3}}{\sqrt{10}} = \sqrt{0,8}$ = 0,894

F) Residualerna är noll för observationerna ett och fem, vilket innebär att residualkvadratsummorna för observationerna 1-4, 2-5, 2-4 är minimerade.

Uppgift 2

A) Förklaringsgraden blir 50%, eftersom X_1 förklarar 50% och X_2 inget ytterligare. Skattningen av variansen för b_{12} blir avsevärt större än skattningen av variansen för b beroende på multikollinearitet. Se formeln för VIF.
B) Förklaringsgraden ökar med 50% ($0,71^2$=0,5) till 75% (X_3 kapar hälften av vad som återstår att förklara).
C) Eftersom residualkvadratsumman halveras minskar residualvariansen med 50%, om vi kan bortse från olika antal frihetsgrader. Formeln för VIF ger då en minskning av variansen med 50%.

Uppgift 3

A) Modellspecifikation: Andel sysselsatta = $\alpha + \beta_1$tid + β_2Dummy(15-74) + β_3Kv2 + β_4Kv3 +β_5Kv4 +ε

H_0: β_2=0

H_1: $\beta_2 \neq 0$

Signifikansnivå: 5%

Testfunktion: $t = \frac{b_2 - \beta_{2,H_0}}{s_{b_2}}$, t-fördelad med n-k-1 = 66 f.g..Föruts: ε NID(0,σ_ε^2).

Beslutsregel: Förkasta H_0 om t_{obs} <-2,00 eller t_{obs}> 2,00 (eller p-värde<0,005)

Resultat: t_{obs}=-5,32, p-värde = 0,000.

Signifikant resultat. Resultatet tyder på att effekten av definitionsförändringen är olika för män och kvinnor.

B) Durbin-Watson statistikan är mycket låg (0,326), vilket indikerar förekomst av autokorrelation. Förutsättningen om okorrelerade störningar skulle därmed inte vara uppfylld.

Preliminära lösningar till tentamensskrivning på kursen Tillämpad statistik A5 (15hp) 2014-08-26 /BW, AG, RP

Uppgift 4

A. Använda formler är:

$$S_t = \alpha Y_t + (1-\alpha)(S_{t-1} + T_{t-1})$$

$$T_t = \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1}$$

S_{72}=0,969919*0,477453+(1-0,969919)*(0,477336+(-0,0000854))=0,477447
T_{72}=0,025261*(0,477447-0,477336)+(1-0,025261)*(-0,0000854)=-0,0000805

Prognos för t=73: 0,477447+1*(-0,0000805) = 0,4774h

t=74: 0,477447+2*(-0,0000805) =0,4773

t=75: 0,477447+3*(-0,0000805) = 0,4772

B. Insättning av α=1 i ekvationerna i uppgift A ger

$$S_t = 1Y_t + (1-1)(S_{t-1} + T_{t-1}) = Y_t$$

$$T_t = \beta(Y_t - Y_{t-1}) + (1-\beta)Y_{t-1}$$

D v s enkel exponentiell utjämning av den diffade tidsserien (Y_t-Y_{t-1}) med utjämningskonstanten β. Den utjämnade förändringen läggs sen på det senaste observerade värdet för att erhålla prognos en tidpunkt framåt.

Preliminära lösningar till tentamensskrivning på kursen Tillämpad statistik A5 (15hp) 2014-08-26 /BW, AG, RP

Uppgift 5

Finns det något samband mellan typ av däck och väglag?

H_0: Inget samband finns mellan typ av däck på det förolyckade fordonet och väglag.

H_1: Det finns ett samband mellan typ av däck på det förolyckade fordonet och väglag.

Signifikansnivå: 5%

Testfunktion: $\chi^2_{obs} = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{\left(O_{ij}-\hat{E}_{ij}\right)^2}{\hat{E}_{ij}}$ där $\hat{E}_{ij} = \frac{R_i K_j}{n}$

r = antal rader, k = antal kolumner R_i = radsumma och K_j = kolumnsumma

Antal frihetsgrader: $(r-1)(k-1) = (3-1)(3-1) = 4$

Förutsättningar:
Alla förväntade frekvenser ska vara minst 5 (uppfyllt här). Stickprovet är utvalt med OSU.

Förkastelseområde: H_0 förkastas om $\chi^2_{obs} > \chi^2_{(r-1)(k-1);\alpha} = \chi^2_{4:0,05} = 9,488$

De förväntade frekvenserna anges inom parentes i tabellen:

	A	B	C	Totalt
Dubbdäck	38 (43,22)	30 (28,53)	34 (30,25)	**102**
Friktionsdäck	31 (36,02)	26 (23,77)	28 (25,21)	**85**
Sommardäck	31 (20,76)	10 (13,70)	8 (14,53)	**49**
Totalt	**100**	**66**	**70**	**236**

$$\chi^2_{obs} = \sum\sum\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}} = \frac{(38-43,22)^2}{43,22} + \frac{(30-28,53)^2}{28,53} + \ldots\ldots\ldots\ldots + \frac{(8-14,53)^2}{14,53} = 11,37$$

$\chi^2_{obs} = 11,37 > 9,488$ H_0 förkastas Signifikant resultat på 5% signifikansnivå

Resultatet tyder på 5% signifikansnivå, på att det finns ett samband mellan typ av däck på det förolyckade fordonet och väglag.

(Om förväntade och observerade frekvenser i tabellen jämförs, ser vi att sommardäck är överrepresenterade vid incidenter på isgata.)

Lösning Uppgift 6, Tentamen A5, 2014-08-26

A.

Eftersom

$$V(\bar{X}) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$$

har vi efter insättning av värden följande olikhet

$$5 \geq \sqrt{\left(\frac{2000-n}{1999}\right)\frac{30^2}{n}}.$$

Genom att lösa ut n algebraiskt eller genom att använda en iterativ lösning får vi att $n \geq 36$.
Svar: Stickprovsstorleken måste vara minst 36 individer.

B.

a)

Den genomsnittliga energiåtgången per hushåll i populationen är

$$\mu = \frac{N_A}{N}\mu_A + \frac{N_B}{N}\mu_B + \frac{N_C}{N}\mu_C,$$

som vi skattar den med

$$\bar{x}_{st} = \left(\frac{1510}{5040}\right)1100 + \left(\frac{987}{5040}\right)950 + \left(\frac{2543}{5040}\right)1200 = 1121,1.$$

Förutsättning 1: För att skattningen ska vara väntevärdesriktig krävs ett OSU från respektive stratum.

Svar: Den genomsnittliga energianvändningen hushållen per vecka är 1091,1 watt.

b)

Medelfelet i undersökningen beräknas med

$$\begin{aligned}
&\sqrt{\hat{V}(\bar{X}_{st})} \\
=&\sqrt{\left(\frac{N_A}{N}\right)^2\left(1-\frac{n_A}{N_A}\right)\frac{s_A^2}{n_A} + \left(\frac{N_B}{N}\right)^2\left(1-\frac{n_B}{N_B}\right)\frac{s_B^2}{n_B} + \left(\frac{N_C}{N}\right)^2\left(1-\frac{n_C}{N_C}\right)\frac{s_C^2}{n_C}} \\
=&\sqrt{\left(\frac{1510}{5040}\right)^2\left(1-\frac{200}{1510}\right)\frac{140^2}{200} + \left(\frac{987}{5040}\right)^2\left(1-\frac{200}{987}\right)\frac{220^2}{200} + \left(\frac{2543}{5040}\right)^2\left(1-\frac{200}{2543}\right)\frac{123^2}{200}} \\
=&5,737.
\end{aligned}$$

Förutsättning 2: För att skatta medelfelet krävs att stickproven i respektive stratum är dragna oberoende av varandra (eftersom variansen för stickprovsmedelvärdet vid stratifierat urval består av en summa av varianser).

Svar: Medelfelet i undersökningen är 5,737.

c)

Ett konfidensintervall för μ ges av $\bar{x}_{st} \pm z_{\alpha/2}\sqrt{\hat{V}(\bar{X}_{st})}$. Eftersom konfidensgraden är 95% sätter vi $z_{\alpha/2} = 1,96$ och efter insättning av värden ovan får vi att konfidensintervallet är

$$1109,8 < \mu < 1132,3.$$

Förutsättning 3: För att kunna hämta z-värdet från den standardiserade normalfördelningen krävs att $\bar{X}_{st}$ är approximativt normalfördelad. Eftersom stickproven i respektive stratum är stora, samt utifrån tidigare antaganden om OSU och oberoende stickprov, så gäller CGS och därmed är detta normalfördelningsantagande uppfyllt.

Svar: Med 95% säkerhet är den genomsnittliga energianvändningen för hushållen i populationen mellan 1080 och 1102 watt i veckan.