Preliminära lösningar för Tentamen Tillämpad statistik A5 (15hp) 2016-02-13

Statistiska institutionen, Uppsala universitet

Uppgift 1 (22 poäng)

A) (3p) $\hat{\beta}_0 = 6.61$ vilket är det genomsnittliga antalet DNM bland nyfödda barn vars bägge föräldrar är 0 år. Interceptet har i detta fall naturligtvis ingen praktiskt relevant tolkning.

B) (2p) Givet att pappans ålder hålls fix, så ökar antalet DNM i genomsnitt med 0.35 för varje år som mammans ålder ökar.

C) (9p)

- Mål: Beräkna ett 95% konfidensintervall för β_2.
- Parameter: β_2
- Estimator: $\hat{\beta}_2$
- Förusättningar: I uppgiften anges att förutsättning (i)-(v) är uppfyllda.
- Beräkning: Ett konfidensintervall för β_2 ges av

$$\hat{\beta}_2 \pm t_{n-(k+1),\alpha/2}\sqrt{\hat{V}(\hat{\beta}_2)}.$$

  Eftersom $t_{690,0.975} = 1.96$ och medelfelet är $\sqrt{\hat{V}(\hat{\beta}_2)} = 0.08$ får vi efter insättning av värden att intervallet ges av

$$0.35 \pm 1.96 \times 0.08.$$

$$0.35 \pm 15.68.$$

- Svar: Med 95% säkerhet befinner sig β_2 i intervallet 0.19 till 0.51 DNM.

D) (5p) För att jämföra förklaringsgrader i modeller med olika antal variabler måste vi justera förklaringsgraderna. Den ursprungliga modellens justerade förklaringsgrad är av $R_a^2 = 1-(693-1)/(693-(2+1))(1-0.35) = 0.348$, medan den nya modellens förklaringsgrad är $R_a^2 = 1-(693-1)/(693-(4+1))(1-0.38) = 0.376$. Eftersom $0.376 > 0.348$ föredrar vi den nya modellen.

E) (3p) Författarna misstänker multikollinjäritet eftersom föräldrarnas ålder kan vara starkt korrelerade. De redovisar VIF-kolumnen för att stärka resultaten och visa att man tänkt på denna eventuella problematik. VIF-värdena som är 1.99 indikerar att multikollinjäritet inte är något problem.

Uppgift 2 (16 poäng)

Oberoendetest

H_0: Det finns inget samband mellan disponibel inkomst och boendekostnad i populationen.

H_1: Det finns ett samband mellan disponibel inkomst och boendekostnad i populationen.

Signifikansnivå: $\alpha = 0{,}05$

Testfunktion: $\chi^2_{obs} = \sum_{i=1}^{3}\sum_{j=1}^{3}\frac{(O_{ij}-\hat{E}_{ij})^2}{\hat{E}_{ij}}$ $\qquad$ $\hat{E}_{ij} = \frac{R_i K_j}{n}$

Förutsättningar: Stickprovet är utvalt med OSU.
Alla $\hat{E}_{ij} > 5$.

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

		Boendekostnad	
Inkomstklass	Låg	Mellan	Hög	Summa
Låg	26 *32,77*	77 *55,61*	34 *48,62*	137
Mellan	70 *52,63*	88 *89,29*	62 *78,08*	220
Hög	26 *36,60*	42 *62,10*	85 *54,30*	153
Summa	122	207	181	510

Förväntade frekvenser med kursiv stil

$$\chi^2_{obs} = \frac{(26-32{,}77)^2}{32{,}77} + \frac{(77-55{,}61)^2}{55{,}61} + \cdots + \frac{(42-62{,}10)^2}{62{,}10} + \frac{(85-54{,}30)^2}{54{,}30} = 50{,}025$$

$\chi^2_{obs} = 50{,}025 > 9{,}488$ $\qquad$ H_0 förkastas

Testresultatet tyder, på 5 % signifikansnivå, på att det i populationen finns ett samband mellan disponibel inkomst och boendekostnad.

Uppgift 3 (20 poäng)

A) (7p) Modellen formuleras som

$$\ln y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{10} X_{10} + \varepsilon.$$

Skattningen av β_1 är $\hat{\beta}_1 = 0.02792$. Eftersom modellen är en log-linjär modell innebär detta att givet att alla variabler hålls fixa, så leder en ökning i X_1 till $(e^{0.02792} - 1) \times 100\% = 2.83\%$ genomsnittlig ökning av försäljning, Y. Eftersom förändringen är liten är det ok att direkt tolka $\hat{\beta}_2 = 0.02792$ som 2.8% genomsnittlig ökning för en ökning i X_1, givet att alla andra variabler är oförändrade. Vi konstaterar också att sambandet är signifikant eftersom $p < 0.001$.

B) (13p)

- Mål: Vi ska jämföra modeller och undersöka om den fulla modellen är mer statistisk användbar än den reducerade modellen.
- Hypoteser $H_0 : \beta_3 = \beta_6 = \beta_9 = 0$ vs H_1 : minst en av β_3, β_6, β_9 är skild från 0.
- Förutsättningar: (i)-(v) måste vara uppfyllda.
- Testfunktion:
$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-(k+1))}$$
som är $F_{k-g,n-(k+1)}$-fördelad om H_0 är sann.
- Beslutsregel: I den fulla modellen finns $k = 10$ och i den reducerade modellen är $g = 7$, dvs vi testar $k-g = 3$ parametrar. Eftersom Total DF $= 99$ så följer att $n = 100$ observationer. Signifikansnivån är 10% dvs $\alpha = 0.1$. F-testet är ensidigt och vi förkastar H_0 om $F_{obs} > F_{3,89,0.1}$. Eftersom vi inte har värden i tabellen för denna kritiska punkt så använder vi det konservativa valet $F_{3,60,0.05} = 2.76$. Skulle vi få ett observerat F-värde är lite mindre än detta konservativa värde så måste vi vara mer noggranna i beräkningarna.
- Beräkning:
$$F_{obs} = \frac{(1.53841 - 0.5152)/3}{0.5152/89} \approx 59$$
- Beslut: Eftersom $F_{obs} \approx 59 >> 2.76 = F_{krit}$ så förkastas H_0 på 10% signifikansnivå.
- Svar: Vi kan på 10% signifikansnivå påvisa att den fulla modellen är mer statistiskt användbar än den reducerade modellen. Minst en av variablerna X_3, X_6 och X_9 är signifikant associerade med försäljning.

Uppgift 4 (24 poäng)

A) (6p)

- Mål: Skatta andelen individer i vården som håller med om journalsystemet är enkelt att lära.
- Parameter: p
- Estimator: $\hat{p}_{st}$
- Förutsättningar: (i) OSU från respektive stratum. (ii) Proportionellt stratifierat urval
- Beräkning: Vi har inte N och stratumspecifika N_j. Däremot vet vi att vi har proportionell allokering vilket innebär att $N_j/N = n_j/n$. Därför skattas p med

$$\hat{p}_{st} = \sum_{j=1}^{4} \left(\frac{n_j}{n}\right) \hat{p}_j = \left(\frac{373}{1019}\right) 0.73 + \cdots + \left(\frac{88}{1019}\right) 0.79 = 0.698.$$

- Svar: Cirka 70% av individerna i vården anser att journalsystemet är enkelt att lära.

B) (12p)

- Mål: Intervallskatta med 95% konfidensgrad andelen individer i vården som håller med om journalsystemet är enkelt att lära.
- Förutsättningar: (i) OSU från respektive stratum. (ii) Stratifierat urval med proportionell allokering, vilket ger oberoende urval. (iii) Ingen ÄK innebär att $V(p_{st})$ skattas med

$$\hat{V}(\hat{p}_{st}) = \left(\frac{n_j}{n}\right)^2 \frac{\hat{p}_j(1-\hat{p}_j)}{n_j - 1}$$

(alternativt $\hat{V}(\hat{p}_{st}) = \left(\frac{n_j}{n}\right)^2 \frac{\hat{p}_j(1-\hat{p}_j)}{n_j}$). För konfidensintervall (vilket förutsätter att CGS gäller) måste $n_j p_j(1-p_j) > 5$ i varje stratum.

- Beräkning: Ett konfidensintervall för p ges av

$$\hat{p}_{st} \pm z_{\alpha/2}\sqrt{\hat{V}(\hat{p}_{st})}.$$

Tvåsidigt KI och 95% konfidensgrad ger $z_{\alpha/2} = 1.96$. Variansen skattas med

$$\hat{V}(\hat{p}_{st}) = \left(\frac{373}{1019}\right)^2 \frac{0.76(1-0.76)}{372} + \cdots + \left(\frac{88}{1019}\right)^2 \frac{0.79(1-0.79)}{87} = 0.000201$$

Insättning av värden ger intervallet

$$0.698 \pm 0.0278$$

Notera att $n_j p_j(1-p_j) > 5$ är uppfyllt.

- Svar: Med 95% säkerhet anser 67 till 73 procent av personalen inom vården att journalsystemet är enkelt att lära.

C) (6p) Vi har att $N = 900$, $\sigma^2 = 52$ samt att felmarginalen får max 2 enheter. Målet är att skatta μ med $\bar{x}$ och vid OSU-UÅ ges felmarginalen av

$$felmarginal = 1.96 \times \sqrt{\left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}}$$

Detta är givet att CGS gäller samt 95% konfidensgrad. Om felmarginalen får vara max 2 eneheter får vi efter insättning av värden att felmarginalen är

$$2 \geq 1.96 \times \sqrt{\left(\frac{900-n}{900-1}\right)\frac{52}{n}}$$

Genom ekvationslösning får vi att

$$2 \geq 1.96 \times \sqrt{\left(\frac{900-n}{899}\right)\frac{52}{n}}$$

$$4 \geq 1.96^2 \times \left(\frac{900-n}{899}\right)\frac{52}{n}$$

$$\frac{4}{1.96^2} \geq \left(\frac{900-n}{899}\right)\frac{52}{n}$$

$$\frac{4}{52 \cdot 1.96^2} \geq \left(\frac{900-n}{899}\right)\frac{1}{n}$$

$$\frac{4}{52 \cdot 1.96^2} \geq \frac{900-n}{899n}$$

$$\frac{899 \cdot 4}{52 \cdot 1.96^2} \geq \frac{900-n}{n}$$

$$\frac{899 \cdot 4}{52 \cdot 1.96^2} \geq \frac{900}{n} - 1$$

$$\frac{899 \cdot 4}{52 \cdot 1.96^2} + 1 \geq \frac{900}{n}$$

$$n \geq \frac{900}{\frac{899 \cdot 4}{52 \cdot 1.96^2} + 1} = 47.36$$

Vi avrundar uppåt och sätter $n = 48$. Eftersom $n > 30$ gäller också CGS vilket gör att våra formler för felmarginalen stämmer. Rätt ges också för om du istället för ekvationslösning iterativt provar dig fram till rätt n.

Uppgift 5 (20 poäng)

A) Modellen som har skattats är

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t$$

för $t = 1, 2, \ldots, 15$ och $y_t =$ Livsmedelsförsäljning.

B) Intervallet ges av

$$\hat{y} \pm t_{n-(k+1),\alpha/2}\sqrt{(\hat{V}(\hat{y}) + s_\epsilon^2)}.$$

Punktskattningen får vi genom

$$\hat{y}_{16} = \hat{\beta}_0 + \hat{\beta}_1 \times 16 = 111704 + 2551 \times 16 = 152520.$$

Eftersom $n = 15$ och $k = 1$ har vi t-koefficienten

$$t_{13,0.025} = 2.16.$$

För att beräkna $\hat{V}(\hat{y})$ använder vi

$$\hat{V}(\hat{y}) = s_\epsilon^2 \left(\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}} \right).$$

I vårt fall så kallar vi x för t, så vi använder SS_{tt} istället för SS_{xx}. Prognosen vill vi göra för $t = 16$, så vi ersätter x_p med 16. Vi har också $\bar{t}$ istället för $\bar{x}$. Från den givna informationen vet vi att $\bar{t} = 120/15 = 8$.

$$SS_{tt} = \sum_{t=1}^{n} t^2 - n\bar{t}^2 = 1240 - 15 \times 8^2 = 1240 - 15 \times 64 = 280.$$

Från utskriften har vi $s_\epsilon^2 = 1852,19^2 = 3430608$. Därmed får vi att

$$\hat{V}(\hat{y}) = 3430608 \times \left(\frac{1}{15} + \frac{(16-8)^2}{280} \right) = 1012846.$$

Prediktionsinverallet blir därför

$$152520 \pm 2.16 \times \sqrt{1012846 + 3430608}$$
$$152520 \pm 4553,$$

dvs prediktionsintervallet säger att med 95 % säkerhet kommer 2015 års livsmedelsförsäljning ligga i intervallet (147967, 157073) miljoner kronor.

C) Koefficienterna är signifikanta, modellen är signifikant och förklaringsgraden är hög (vilket är vanligt vid trendande serier).

I residualfigurerna kan vi se problem. Det är tydligt att residualerna är

korrelerade, då positiva värden följs av fler positiva värden och vice versa. Det är därför tydligt att vi har problem med autokorrelation. En av förutsättningarna för korrekt inferens är att feltermerna är oberoende och här finner vi stöd för motsatsen. Med anledning av det är medfelen inkorrekt skattade, vilket påverkar vår inferens. Punktskattningarna är fortfarande väntevärdesriktiga, men huruvida de är signifikanta eller ej kan vi inte svara på med den information vi har tillgänglig. Dessutom innebär det faktum att vi har ett litet stickprov att vi inte kan förlita oss på centrala gränsvärdessatsen. Vi kan därför skatta modellen, men det blir problematiskt att göra inferens, vilket bl a innebär att prediktionsintervallen inte är tillförlitliga.