STOCKHOLM UNIVERSITY
Department of Statistics
Johan Koskinen

Tentamen, Generaliserade linjära modeller (GLM)
2024-04-26

Tid för examination: 14.00-18.00

Tillåtna hjälpmedel: Miniräknare, tilldelad tabellsamling, och medtaget formelblad (1 dubbel-sidigt A4)

Observera att inget formelblad kommer utdelas

Tentamen består av två delar. Del A: frågor 1-5; Del B: 6a och 6b. Frågorna i del A är tillsammans värda SJUTTIO (70) poäng. Fågan i del B är värd TRETTIO (30) poäng. Besvara **alla** frågor i del A. Besvara **endast en** av frågorna i del B. Du väljer vilken fråga i del B som du vill besvara och lämna in. Endast **ett** (1) svar kommer at rättas! Om du lämnar in svaret på två B-frågor förbehåller examinatorn sig rätten att inte rätta någon och du får noll (0) poäng på del B.

För godkännt krävs minst 50 utav det totalt 100 poängen. För detaljerade betygskriterier, se kursbeskrivningen.

Good luck!

Del A: besvara samtliga frågor

Problem 1. (14 poäng)
Några utbildningsekonomer har studerat vilka sökande som blir antagna till en grupp av de mest prestigefulla colleges i Oxford. För en delmängd av sökande $i = 1, \ldots, n$ med likvärdiga betyg, noterar de om den sökande blir antagen ($y_i = 1$), eller inte ($y_i = 0$), är man ($x_{i1} = 1$) eller inte ($x_{i1} = 0$). Dessutom har de ett centrerat mått på ekonomisk bakgrund x_{i2}, där ett högt värde betyder att den sökande kommer fån ett förmöget hem.

De skattar logistisk regression med linjär prediktor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}$$

och får skattningarna

$$\hat{\beta}_0 = -1,\ \hat{\beta}_1 = -.05,\ \hat{\beta}_2 = 0.15,\ \text{ samt } \hat{\beta}_3 = 0.6.$$

(a) Enligt den skattade modellen, vad är sannolikheten att en kvinna med $x_{i2} = 0$ blir antagen?

(b) Vad är oddskvoten (alltså kvoten mellan odds) mellan en man och en icke-man som båda har $x_{i2} = 1$

(c) För vilket värde på bakgrund är en icke-man lika trolig att bli antagen som en man?

Problem 2. (14 poäng)
Varje kväll j har Pete $i = 1$ och John $i = 2$ nyttofunktionen för att gå ut och klubba

$$U_{ij} = -0.5 + 0.1 m_{ij} + 0.1_d j - 0.5 a_{ij} + \epsilon_{ij}$$

där m_{ij} är hur många tusen SEK i har på bankkontot, d_j är 1 om Pete och Johns favorit-DJ spelar (noll annars), och a_{ij} är 1 i var ute kvällen innan. Vi antar att slumpfelen är oberoende och fördelade $\epsilon_{ij} \sim N(0, 1)$.

Givet vad som hände igår, går person i ut kväll j om $U_{ij} > 0$, men stannar annars hemma.

(a) Det är fredag, Pete och John var inte ute kvällen innan och deras favorit-DJ spelar. Pete har 11 tusen SEL kvar på kontot, vad är sannolikheten att han går ut?

(b) Vi vet nu att samma kväll är sannolikheten att John går ut 0.8. Hur mycket pengar har John på kontot?

(c) Hur mycket pengar måste John låna ut till Pete så att båda har samma sannolikhet att gå ut?

Problem 3. (14 poäng)
En studie utförs på det totala antalet 'likes' en TikTok-post attraherar för ett sample av

TikTok-poster som har bedömts rikta sig mot Indonesien och som är relaterade till mobiltelefon märken. För varje post i, räknas antalet 'likes' y_{ij}, och för att förklara en posts popularitet, kodas indikatorer (dummy variabler) av posten är en 'review', innehåller dans, har en foto-slideshow, animationer, music, ljudeffekter (andra än musik), har en beskrivning med hashtag, samt om innehållet visar eller innehåller referenser till en 'influencer'.

Maximum likelihood-skattningarna för en Poisson-regressions med kanonisk länk på materialer ges i tabellen nedan

j	Variable	$\hat{\beta}_j$
0	intercept	7.44
1	review	1.02
2	dance	0.32
3	photo	-0.75
4	animation	-0.56
5	music	0.60
6	sound	0.28
7	hastag	-0.93
8	influencer	0.09

(a) Vad är sannolikheten att en post som inte är en review, med endast en foto slideshow och med en hashtag (dvs utan de andra elementen), inte får ett ända 'like'?

(b) Om du till posten, såsom definerad i (a), lägger till en käck liten dans, och lite musik, vad blir den predicerade sannolikheten att inte få någon 'like'?

(c) För en 'post' som är en review, inte har dans eller slideshow, har animation och har musik men inget övrigt ljud, och pratar om en influencer, är det predicerade värdet $\hat{y}_i = 765.095$. Har posten en hashtag?

Problem 4. (14 poäng)
För exemplet i problem 3, skattar vi också en model där vi antar istället att Y_i följer en negativ Binomialfördelning, med $k = 3$. Du använder inte den kanoniska länfunktionen, utan

$$g(E(Y_i \mid x_{i1}, x_{i2}, x_{i3})) = \log(E(Y_i \mid x_{i1}, x_{i2}, x_{i3})) = \eta_i$$

Du får samma parameterskattningar som in problem 3.

(a) För en post med innehåll som i 3(b), ger regressionen som antar negativ Binomial en högre eller lägre predicerad variance, $V(Y_i \mid x_{i1}, \ldots, x_{i8})$ än Poissonmodellen i 3?

(b) För en post med innehåll som i 3(b), ge ett intervall $[a, b]$ så att

$$\Pr(a < Y_i < b \mid x_{i1}, \ldots, x_{i8}) \approx 0.95$$

med hjälp av att approximera fördelningen med en lämplig normalfördelning.

Problem 5. (14 poäng)
En restaurang som har öppen 6 kvällar i veckan har 4 huvidrätter. Vi ger de fem möjliga valen en kund kan göra, dvs, ingen huvudrätt, huvudrätt A, B, C, eller D, etiketterna $j = 0, 1, 2, 3, 4$, respektive, och definierar de fem variablerna

$$Y_{ij} = \begin{cases} 1, & \text{om } i \text{ gör val } j \\ 0, & \text{annars} \end{cases},$$

för $i = 1, \ldots, n$. Vi antar en länk funktion

$$g(\Pr(Y_{ij} = 1 \mid x_{i1}, x_{i2})) = \log\left(\frac{\Pr(Y_{ij} = 1 \mid x_{i1}, x_{i2})}{\Pr(Y_{i0} = 1 \mid x_{i1}, x_{i2})}\right) = \eta_{ij}.$$

Restaurangen är öppen Tisdag tom Söndag och $x_{i1} = 1$ för en kund på en Tisdag kväll, $x_{i2} = 1$ för en kund på en Onsdag kväll, osv. Den lijära prediktorn antas vara

$$\eta_{ij} = \beta_{0j} + \sum_{h=1}^{6} \beta_{hj} x_{ih}$$

Restaurangchefen skattar parametrarna

	$\hat{\beta}_{hj}$			
	Rätt j			
Dag h	1	2	3	4
1	2.50	1.45	1.00	1.20
2	1.00	1.10	1.13	1.05
3	0.90	1.01	1.55	1.55
4	1.70	1.12	0.01	0.01
5	1.15	1.05	1.20	1.22
6	1.20	1.15	2.20	2.09

samt,

$$\hat{\beta}_{01} = 1, \ \hat{\beta}_{02} = 1.2, \ \hat{\beta}_{03} = 0.9, \ \text{ samt } \hat{\beta}_{04} = 0..$$

(a) Enligt modellen, vad är sannolikheten att en slumpvis vald kund på en Lördag väljer rätt D?

(b) Givet att restaurangen har 50 kunder en Lördag, och att de väljer rätter oberoende av varandra, hur många rätter kan chefskocken förvänta sig att de skall laga?

Del B: besvara endast <u>en</u> av frågorna 6a och 6b

En fråga är värd 30 poäng

Problem 6.a
Antag att du har grupper $j = 1, \ldots, m$ och att i varje grupp har vi individer $i = 1, \ldots, n_j$. För varje individ i i grupp j, låt

$$\eta_{ij} = \beta_0 + \beta_1 x_{ij} + u_j$$

där, oberoende för $j = 1, \ldots, m$

$$u_j \sim N(0, \tau^2).$$

För varje individ i i grupp j, definierar vi

$$Y_{ij}^* = \eta_{ij} + \epsilon_{ij},$$

där, oberoende för $j = 1, \ldots, m$ och i, $\epsilon_{ij} \sim N(0, 1)$.
(i) I termer av β_0, β_1, och u_j, vad har Y_{ij}^* för fördelsning betingat på u_j?
(ii) Definiera

$$Y_{ij} = \begin{cases} 1, & \text{om } Y_{ij}^* > 0 \\ 0, & \text{om } Y_{ij}^* \leq 0 \end{cases}.$$

Vad är $\Pr(Y_{ij} = 1 \mid x_{ij})$ i termer av β_0, β_1, och τ.
(iii) Obetingat av u_j, men givet τ^2, är Y_{ij} och Y_{hj} oberoende?

Problem 6.b

För Y_i givet x_i, antar du en model som tillhör exponentialfamiljen och har den linjära prediktorn

$$\eta_i = \beta_0 + \beta_1 x_i$$

men en kanonisk länkfunktion, där x_i är binär. För $n = 10$ observationer, får beräknar Anna

$$\sum_{i=1}^{1} 0 y_i = 26, \text{ och } \sum_{i=1}^{1} 0 y_i x_i = 19.$$

Vidare skattar hon, medelst maximum likelihood, parametrarna till

$$\hat{\beta}_0 = 0.3365, \text{ och } \hat{\beta}_1 = 0.9985.$$

Vilken fördelning har Anna antagit för Y_i givet x_i?