

Tentamen, Generaliserade linjära modeller (GLM) 2024-12-04

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. Frågan 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 att 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 godkänt krävs minst 50 utav de totalt 100 poängen. För detaljerade betygskriterier, se kursbeskrivningen.

Good luck!

Del A: besvara samtliga frågor

Problem 1. (14 poäng)

I samband med valet i USA den 5:e november, undersökte en forskare vad som predicerade att en väljare röstade på Trump.

	Definition
Y_i	Röstade för Donald Trump (1) eller Kamala Harris (0)
x_{i1}	Har college degree
x_{i2}	Är kvinna (1) eller inte (0)
x_{i3}	Äger ett vapen (1) eller inte (0)
x_{i4}	Tilltro till vetenskap, från ingen (0) till helt och hållet (10)

Antag följande model

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$$

och en logistisk länkfunktion.

Baserat på skattningar från ett (riktigt) stickprov, antag att

$$\beta_0 = 2.25, \beta_1 = -0.53, \beta_2 = -0.29, \beta_3 = 0.97, \text{ och, } \beta_4 = -0.43$$

och betrakta dessa som givna.

(a) Vad är den predicerade sannolikheten att en kvinna som tror helt och hållet på vetenskapen, har en college examen och inte äger ett vapen skall rösta på Kamala?

(b) En specifik man som inte har en college examne och inte äger ett vapen har en predicerad sannolikhet att rösta på Trump som är 0.525. Vad är hans tilltro till vetenskapen?

(c) Antag att ett lager i 'Midwest' har 30 anställda som alla äger sitt eget vapen, har noll tro på vetenskapen, och är alla män. Bara en av dem, deras lagerchef, har en college examen. Vad är sannolikheten att *ingen* röstar på Kamala Harris?

Problem 2. (14 poäng)

För samma datamaterial som siffrorna ovan är baserade på, ger en skattning med *Probit* länk

$$\beta_0 = 1.31, \beta_1 = -0.32, \beta_2 = -0.16, \beta_3 = 0.58, \text{ and, } \beta_4 = -0.25$$

och betrakta dessa som givna (därav inga hattar).

(a) Vad är den predicerade sannolikheten att en kvinna som tror helt och hållet på vetenskapen, har en college examen och inte äger ett vapen skall rösta på Kamala?

(b) Två män som inte har college examen och inte äger vapen, har predicerade sannolikheter på 0.524 respektive 0.905. Hur stor skillnad är det på deras tilltro till vetenskapen?

(c) För 10 män exakt som den första mannen i (b), vad är sannolikheten att exakt 4 röstar för Trump?

Problem 3. (14 poäng)

Göteborg har Sveriges största hamn. Antalet kontainrar Y_i som hamnen hanterar varje dag i , åren 2019-2021, ges av en $Po(\lambda_i)$ som med kanonisk länkfunktion har den linjära prediktorn

$$\eta_i = \beta_0 + \beta_1 Q2_i + \beta_2 Q3_i + \beta_3 Q4_i + \beta_4 \text{Year}20_i + \beta_5 \text{Year}21_i$$

där $Q2_i$, $Q3_i$, och $Q4_i$, är indikatorer för om dagen är i kvartal 2, 3, eller 4, respektive. För dagar under 2020 är $\text{Year}20_i = 1$, och noll annars. Samma gäller för $\text{Year}21_i$. Följande parametervärden är givna (betrakta dessa som om de vore sanna):

β_0	7.651
β_1	0.036
β_2	0.025
β_3	0.065
β_4	0.044
β_5	0.112

(a) Ge en prediktion för antal kontainrar hamnen hanterar den 3:e januari, 2020. Predicera med det betingade väntevärdet.

(b) För samma dag, finn det antal kontainrar k , så att sannolikheten att fler än k kontainrar är 0.05. Approximera Poissonfördelningen med en lämplig normalfördelning (dvs, med matchande väntevärde och varians).

(c) Stockholms hamn hanterade endast 4246 kontainrar under Januari månad 2020. Antag att alla 31 dagar i Januari $Y_i \sim Po(\lambda)$, oberoende. Om du antar att alla 24 timmar på ett dygn oberoende har samma fördelning för antal kontainrar, vad är sannolikheten att det inte skulle komma en enda kontainer i ett slumpvis valt entimmesintervall?

Problem 4. (14 poäng)

Antag samma förutsättningar som i Problem 3, med skillnaden att $Y_i \sim NB(k = 4, p_i)$. Du använder inte den kanoniska länkfunktionen, utan länkfunktionen

$$\log(E(Y_i | Q2_i, Q3_i, Q4_i, \text{Year}20_i, \text{Year}21_i)) = \eta_i$$

Antag samma linjära prediktor, med samma parametervärden, som i Problem 3.

(a) För 3:e Januari 2020 finn det antal kontainrar c , så att sannolikheten att fler än c kontainrar är 0.05. Approximera Negativ Binomial fördelningen med en lämplig normalfördelning (dvs, med matchande väntevärde och varians).

(b) Vad blir ditt svar på frågan i (a) om vi antar att $k = 3$ istället för 4?

Problem 5. (14 poäng)

I en stor representativ amerikansk studie är vi intresserade av att förklara jaktvanor. Variabeln Y_i är respondentens svar på frågan: ‘Do you/Do you or does your spouse go hunting?’. Svarsalternativen är

1. Ja, respondenten gör det
2. Ja, respondentens partner gör det
3. Ja, både respondenten *och* deras partner gör det
4. Nej, varken respondenten eller deras partner gör det

För varje respondent have vi följande kovariat

x_{i1}	Kvinna (1) eller inte (0)
x_{i2}	Växte du upp på landsbygden (1) eller inte (0)?
x_{i3}	Har du en college degree (1) eller inte (0)?
x_{i4}	‘Ras vit’: Anser du dig vit (1) eller inte (0)?

Vi kodar variabeln som

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

för $j = 1, 2, 3, 4$ och $i = 1, \dots, n$. Vi antar en länk funktion

$$g(\Pr(Y_{ij} = 1 \mid x_{i1}, x_{i2}, x_{i3}, x_{i4})) = \log \left(\frac{\Pr(Y_{ij} = 1 \mid x_{i1}, x_{i2}, x_{i3}, x_{i4})}{\Pr(Y_{i1} = 1 \mid x_{i1}, x_{i2}, x_{i3}, x_{i4})} \right) = \eta_{ij},$$

där

$$\eta_{ij} = \beta_{0j} + \sum_{k=1}^4 \beta_{kj} x_{ik}.$$

I följande uppgift, använd skattningarna från studien som parametervärden. Dessa är

j	2	3	4
β_{0j}	-5.018	-2.253	2.358
β_{1j}	5.004	1.455	1.649
β_{2j}	-0.467	-0.169	-1.137
β_{3j}	0.302	-0.081	0.330
β_{4j}	0.548	0.403	-0.755

(a) Givet att du har en vit man som växte upp på landet och inte har college utbildning, vad är sannolikheten att han *och* hans partner båda jagar?

(b) I det lokala softball laget, där alla är vita, utav de 9 spelarna, vad är sannolikheten att det är exakt två män som jagar med sina respektive partners? Alla i laget har college utbildning och växte upp på landsbygden. (Antag att alla är gifta/har en partner).

Del B: besvara endast en av frågorna 6a och 6b

En fråga är värd 30 poäng.

Problem 6.a

Antag att du har samma data som i Problem 1 och 2 men att du istället för logit eller probit länkar antar den *komplimentära log-log* länkfunktionen

$$\log \{-\log(1 - p_i)\} = \eta_i.$$

Betrakta de skattade parametrarna som givna. De är:

$$\beta_0 = 1.01, \beta_1 = -0.33, \beta_2 = -0.26, \beta_3 = 0.71, \text{ och, } \beta_4 = -0.29.$$

För de 29 anställda på lagret i Problem 1 (c) som inte har college utbildning, vad är sannolikheten att minst 3 röstar på Trump?

Problem 6.b

Antag att vi har två kategorier av respondenter som indikeras medelst $x_i = 0$ eller $x_i = 1$. Vi har en utfallsvariabel Y_i som tar värden $\{0, 1, 2, \dots\}$, och vi antar att $Y_i \mid x_i \sim NB(k = 3, p_i)$.

(i) Härled den *kanoniska* länkfunktionen (alltså inte $\log(\mu_i) = \eta_i$).

(ii) Givet ett dataset där $n = 2000$, $\sum_{i=1}^n x_i = 1000$, samt

$$\sum_{i=1}^n y_i = 3514, \text{ och } \sum_{i=1}^n y_i x_i = 2367,$$

verifiera att

$$\hat{\beta}_0 = -1.501, \text{ och } \hat{\beta}_1 = 0.511$$

maximerar likelihood-funktionen.