

מודלים סטטיסטיים ויישומיהם 85251 – תרגיל 50

להגשה עד 8.1.18 בשעה 23:55

1. מצורף קובץ בשם credit.csv עם נתונים בנוגע ל-10000 הלוואות של לקוחות בנק גרמני. המשתנים המסבירים בקובץ (רשימה חלקית מתוך הנתונים המלאים) הם:
 - א. good_credit: משתנה בינארי המציין האם רמת הסיכון של הלוקה טובה (1) או לא (0).
 - ב. term:משך ההלוואה בחודשים.
 - ג. amount: סכום ההלוואה המבוקש (בمارك גרמני).
 - ד. age: גיל הלוקה.
 - ה. land: משתנה בינארי המציין האם הלוקה בעל אדמה (1) או לא (0).המשתנה המסביר יהיה good_credit. בצעו ראשית ניתוח תיאורי של הנתונים והציבו על מאפייני הנתונים השונים (התפלגות, קשרים אפשריים בין משתנים, צפיפות חריגות וכן הלאה). הריצו מודל רגסיה לוגיסטיבית לבדיקת הקשר בין המשתנים המסבירים למשתנה המסביר. יישמו את הכלים בקובץ logcheck1.pdf לבדיקה טיב התאמת המודל ובמידת הצורך הריצו מודל חלופי. דנו בתוצאות.
הערה: אין צורך למצוא מודל אופטימלי, מעבר למודל הבסיסי התאמיו רק מודל חלופי אחד.
2. נניח כי $(Q_m \sim Poi(e^{\alpha + \theta_m}))$ ב"ת עבור $M = 1, \dots, m = 1$, כאשר $\sum_{m=1}^M e^{\theta_m} = 1$ ונסמן $\sum_{m=1}^M Q_m = Q$ בהנחת $n = N$ היא מולטינומית:
$$(Q|N = n) \sim Multinomial(n; e^{\theta_1}, \dots, e^{\theta_M})$$
3. הריצו את מודל הרגסיה הפואסונית שפורסם באתר בקובץ Poisson Regression Example. חשבו אומד עבור $[x = 1, Age = 0.6, Income = 0.9, HScore = 11]$: $E[Y|X = x^*]$.
4. שאלת זו עוסקת במודלים לוג לינאריים
 - א. כתבו פונקציה ב-R מקבלת קלט מערך תלת-ממדי של ערבי π ומחשבת את $\bar{\theta}$ ואת ערבי λ השונים.
 - ב. מערך תלת-ממדי מוגדר ב-R באופן הבא:

```
g <- array(data, c(A,B,C))
```

בארור `data` הוא וקטור המוביל את הנתונים ו- `A, B, C` הם מספרי הרמות התואמים. עבור נתונים הקובץ ex4.csv צרו מערך תלת-ממדי תואם והריצו את הפונקציה אותה כתבתם על `g`.
 - ג. וודאו כי מתקיים $P(A = i, B = j, C = k) = P(A = i)P(B = j)P(C = k)$.