רגרסיה ומודלים סטטיסטיים - בוחן 1

אביב 2021

הערה כללית: כאשר אתם מתבקשים למצוא "ביטוי מפורש", הכוונה לביטוי שאפשר להציב בו את הנתונים ולקבל פתרון (אם צריך, הנוסחה בהחלט יכולה לערב מטריצת). למשל $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ זה ביטוי מפורש עבור $\hat{\boldsymbol{\beta}}$.

שאלה 1 (30 נקודות)

נניח את המודל הלינארי הכללי

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim (0, \sigma^2\boldsymbol{I}_n)$$

כאשר $\boldsymbol{X} \in \mathbb{R}^{n\times(p+1)}$ מטריצה קבועה (לא מקרית) שעמודותיה $\boldsymbol{X}^{(j)} \in \mathbb{R}^n, 0 \le j \le p$ ב"ת לינארית, והעמודה הראשונה שלה היא $\boldsymbol{X}^{(0)} = \boldsymbol{1}_n = (1,\dots,1)^T \in \mathbb{R}^n$.

תזכורת: הסימן $\boldsymbol{\epsilon} \sim (0, \sigma^2\boldsymbol{I}_n)$ משמעו שהוקטור המקרי $\boldsymbol{\epsilon}$ בעלת תוחלת $\mathbb{E}[\boldsymbol{\epsilon}] = 0$ ומטריצת שונויות $Cov[\boldsymbol{\epsilon}] = \sigma^2\boldsymbol{I}_n$.

א. ראינו שאומד הריבועים הפחותים עבור $\boldsymbol{\beta}$ נתון על ידי $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. הראו שמתקיים גם $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y}$ כאשר $\boldsymbol{P}_{\boldsymbol{X}} \in \mathbb{R}^{n\times n}$ היא מטריצת ההיטל על $Im(\boldsymbol{X})$.

ב. מצאו אומד לינארי (ב-$\boldsymbol{Y}$) וחסר-הטייה בעל שונות מינימלית עבור $\theta = \beta_2 - \beta_1$. כלומר, מצאו אומד $\hat{\theta}$ אשר מקיים $\mathbb{E}[\hat{\theta}] = \theta$ והשונות שלו $Var(\hat{\theta})$ נמוכה מהשונות של כל אומד חסר הטייה אחר. יש לנמק מדוע האומד אכן בעל שונות מבין כל האומדים הלינאריים חסרי-ההטייה.

ג. עבור $1 \le i \le n$, נגדיר $h_i := Cov(\hat{Y}_i, Y_i)$, כאשר $\hat{Y}_i$ הוא הערך החזוי בשיטת הריבועים הפחותים עבור התצפית ה- i. מצאו ביטוי מפורש עבור h_i.

שאלה 2 (10 נקודות)

יהי $\boldsymbol{Y} = (Y_1,\dots,Y_n)^T \in \mathbb{R}^n$ וקטור מקרי, ונסמן $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{\mu}$ עבור התוחלת של $\boldsymbol{Y}$. כמו כן, עבור כל וקטור $v \in \mathbb{R}^n$, נגדיר $Q(v) = v^T\boldsymbol{A}v$, כאשר $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ מטריצה סימטרית קבועה (לא מקרית). הראו שמתקיים

$$\mathbb{E}[Q(\boldsymbol{Y})] = Q(\boldsymbol{\mu}) + \mathbb{E}[Q(\boldsymbol{Y} - \boldsymbol{\mu})]$$

הדרכה: עבור ההוכחה, יהיה נוח להשתמש בייצוג $\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}, \quad \mathbb{E}[\boldsymbol{\epsilon}] = 0$.

שאלה 3 (30 נקודות)

אליס ובוב צופים ב- n תצפיות מדגם $(x_i, Y_i)_{i=1}^n$, כאשר $x_i \in \mathbb{R}$ מספרים קבועים (לא מקריים). הם מעוניינים ללמוד על הקשר בין x_i ל- Y_i. אליס משתמשת במודל

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} (0, \sigma^2) \tag{A}$$

עבור הנתונים, ואילו בוב משתמש במודל

$$Y_i = \alpha + \beta x_i + \gamma x_i^2 + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} (0, \sigma^2) \tag{B}$$

תזכורת: הסימן $\epsilon_i \overset{iid}{\sim} (0, \sigma^2)$ משמעו ש- ϵ_i הם מ"מ ב"ת ושווי-התפלגות בעלי תוחלת $\mathbb{E}[\epsilon_i] = 0$ ושונות $Var[\epsilon_i] = \sigma^2$.

א. מצאו במפורש את אומדי הריבועים הפחותים עבור α ו- β תחת המודל (A), ואת אומדי הריבועים הפחותים עבור α, β ו- γ תחת המודל (B).

ב. מצאו את $Var[\hat{\beta}]$, $Var[\hat{\alpha}]$ ואת $Cov(\hat{\alpha}, \hat{\beta})$ תחת המודל (A).

ג. מצאו ביטוי מפורש עבור $Var[\hat{\gamma}]$ תחת המודל (B).

ד. מצאו רווח-סמך ברמת ביטחון $1 - \alpha$ עבור β תחת המודל (A), כאשר מוסיפים את ההנחה ש- ϵ_i בעלי התפלגות נורמלית (כלומר $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$).

ה. מצאו מבחן דו-צדדי עבור השערת האפס $H_0 : \gamma = 0$ תחת המודל (B), כאשר מוסיפים את ההנחה ש- ϵ_i בעלת התפלגות נורמלית (כלומר $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$).

שאלה 4 (30 נקודות)

בשאלה זו נבנה ידנית פונקציה אשר מבצעת את מרבית הפעולות של הפונקציה המובנית $lm()$ ב-R. לאחר מכן, ניישם את הפונקציה שבנינו לניתוח קובץ הנתונים $Startups.csv$ אשר מכיל נתונים אודות ההוצאות השונות, המדינה והרווח של סטארט-אפים אמריקינים. אנחנו נרצה לחזות את רווח החברה בהינתן המאפיינים השונים הנתונים לנו.

א. בתיקייה המצורפת במודל יש קובץ R בשם $Quiz\_1.R$ שבו שלד לפונקציה $My\_lm(\boldsymbol{X}, \boldsymbol{Y})$ אשר מקבלת כקלט את מטריצת המשתנים המסבירים $\boldsymbol{X}$ ואת וקטור המשתנים המוסברים $\boldsymbol{Y}$. הפונקציה צריכה לאמוד את הפרמטרים ברגרסיה לינארית מהצורה

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 I_n)$$

ולהחזיר אומדנים שונים כפי שמוסבר בהרחבה בקובץ ה-R המצורף. בסעיף זה עליכם לכתוב את הקוד הנדרש לקבלת הפלט בפונקציה. אסור להשתמש בפונקציה $lm()$ או כל פונקציה אחרת אשר אומדת רגרסיה לינארית.

ב. קראו את הקובץ $Startups.csv$. אנו נרצה לאמוד מודל לינארי בו $Profit$ הוא המשתנה המוסבר ויתר המשתנים הם המשתנים המסבירים. בעזרת הפונקציה מסעיף א', תבצעו את השלבים הבאים

(i) הציגו סטטיסטיקה תיאורית בסיסית עבור כל אחד מהמשתנים במודל.

(ii) בדקו את ההתפלגות האמפירית של $Profit$ באמצעת שימוש ב- $Histogram$. מה ניתן ללמוד מהתפלגות זו?

(iii) עבור כל אחד מהמשתנים המסבירים הרציפים, הציגו את ה- $Scatter-plot$ השולי שלו מול המשתנה $Profit$ וחשבו את הקורלציה ביניהם.

(iv) בצעו טרנספומציות רלוונטיות על המשתנים, אמדו את המודל הלינארי המבוקש וחשבו את $\hat{\beta}$.

(v) עבור כל אחד מ- $0 \leq j \leq p$ חשבו את הסטטיסטי $T_j^{H_0} = \frac{\hat{\beta}_j - \beta_j^{H_0}}{S.E.(\hat{\beta}_j)}$ תחת השערת האפס $0 : \beta_j = 0$, כאשר $S.E.(\hat{\beta}_j)$ הוא האומד לסטיית התקן של $\hat{\beta}_j$.

(vi) עבור כל אחד מ- $0 \leq j \leq p$ חשבו את רווח הסמך ל- β_j ברמת מובהקות $1 - \alpha$.

(vii) עבור כל אחד מ- $0 \leq j \leq p$ קבעו אם נדחה את השערת האפס $\beta_j = 0$ ברמת מובהקות $\alpha = 0.05$.

ג. הקובץ $Startups\_test.csv$ מכיל מידע על סטארט-אפים אמריקנים נוספים. בעזרת המודל שאמדתם בסעיף ב':

(i) עבור כל תצפית בקובץ $Startups\_test.csv$ מצאו את ערך $Profit$ החזוי.

(ii) נסמן ב- $\hat{Y}_i^*$ את ערך $Profit$ החזוי שחישבתם בסעיף הקודם, וכן ב- Y_i^* את הערך האמיתי של $Profit$ אשר מופיע בקובץ. חשבו את שורש הסטייה הריבועית הממוצעת של התחזית, כלומר חשבו את

$$RMSE(Y^*, \hat{Y}^*) = \sqrt{\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (Y_i^* - \hat{Y}_i^*)^2}$$

כאשר n_{test} זה מספר התצפיות בקובץ $Startups\_test.csv$.