רגרסיה ומודלים לינאריים 52320 תשע"ו 2015-16
בוחן 1 28.03.2016

בבוחן שאלות אמריקאיות ושאלות פתוחות.

ליד כל **סעיף** מצויין מספר הנקודות המקסימלי עבור **פתרון מלא** . מספר הנקודות הכולל הוא 120. בכל מקרה, ציון הבוחן הוא 100 לכל היותר.

שימו לב שהשאלות הן בדרגת קושי שונה כך שמומלץ לא להתעכב יתר על המידה על שאלה מסויימת.

אנא הקפידו על ההנחיות הבאות:

- כתבו את ת.ז. (**לא את השם!**) בראש כל עמוד של טופס הבחינה.
- אין לצרף לטופס דפים נוספים.
- אין לתלוש דפים מטופס הבחינה.

לתשומת לבכם לגבי השאלות הפתוחות:

- **תשובה סופית ללא דרך לא תזכה בניקוד כלשהו (ציון 0).**
- בשאלות הפתוחות יש לכתוב את הפתרון רק במקום המוקצה לכך, מעל לכל קו כתבו שורה אחת בלבד בכתב יד קריא. (השאלות נכתבו כך שניתן לכתוב פתרון תמציתי לכל סעיף).
- מגבלת המקום תאכף באופן קפדני. פתרונות אשר יחרגו מהמקום המותר, יהיו בכתב קטן מכדי שיהיה קריא, ו/או יכללו יותר משורת כתב אחת לכל קו לא ייבדקו.
- מומלץ מאוד לפתור תחילה את השאלה בדפי טיוטה ולהעתיק את עיקר הפתרון אל הטופס רק לאחר בדיקה.

חומר עזר מותר: מחשבון. דף-נוסחאות אחד דו-צדדי.

משך הבוחן: שעה

בהצלחה!

סימונים: נכתוב משתנים בכתיב וקטורי, כאשר $x, y, ..$ הם וקטורי עמודה. x_i מסמן את האיבר ה־i של וקטור x ו־$\bar{x}$ מסמן את הממוצע של וקטור x. עבור שני וקטורים x, y באורך n המכפלה הסקלרית שלהם היא $x^T y = \sum_{i=1}^n x_i y_i$. עבור וקטור x הנורמה בריבוע היא: $||x||^2 = x^T x = \sum_{i=1}^n x_i^2$.

1. [30 נק'] נתונים כל הסכומים הבאים עבור מדגם $\mathcal{D} =\{(x_1,y_1), ..., (x_n, y_n)\}$:

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2} = 3.8,\ s_y = \sqrt{\frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2} = 14.9, \bar{x} = -12.3,\ \bar{y} = -4.8,\ R^2 = 0.46,\ n = 30$$

חשבו את **כל** הערכים האפשריים עבור $\hat{\beta}_0, \hat{\beta}_1$ וה־SSE עבור רגרסיה לינארית פשוטה עם חותך לנתונים שלעיל. הסבירו את תשובותיכם

$R^2 = 1 - \frac{SSE}{SST} \Rightarrow SSE = SST(1-R^2) = n s_y^2 (1-R^2) = 30 \cdot 14.9^2 \cdot (1-0.46) = 3596.6$

$|r_{xy}| = \sqrt{R^2} = 0.678 \Rightarrow r_{xy} = \pm 0.678.$

$\hat{\beta}_1 = \frac{s_y}{s_x} r_{xy} = \pm \frac{14.9}{3.8} 0.678 = \pm 2.66$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -4.8 \pm 12.3 \cdot 2.66 = \{27.92$ או $-37.52\}$

2. עבור מדגם $\mathcal{D} =\{(x_1,y_1), ..., (x_n, y_n)\}$ מבצעים רגרסיה לינארית פשוטה (עם חותך). כעת הניחו שמבצעים טרנספורמציות לינאריות על x ועל y : $x' = ax,\ y' = cy$ עבור סקלרים $a, c \neq 0$ לקבלת מדגם חדש $\mathcal{D}' =\{(ax_1,cy_1), ..., (ax_n, cy_n)\}$ ועושים רגרסיה לינארית פשוטה של y' מול x'. יהיו $\hat{\beta}_0, \hat{\beta}_1$ אומדי הריבועים הפחותים של המדגם הראשון ו־$\hat{\beta}'_0, \hat{\beta}'_1$ של המדגם החדש D'.

נסמן כי s_x', s_y', r_{xy}' את ס"ת התקן של x', y'

ומקדם המתאם של x', y'. אז:

$$\begin{cases} s_x' = |a| s_x \\ s_y' = |c| s_y \\ r_{xy}' = \mathrm{sign}(ac) r_{xy} \end{cases}$$

(א) [15 נק'] $\hat{\beta}'_1$ שווה ל:

i. $\hat{\beta}_1$

ii. $\frac{a}{c}\hat{\beta}_1$

iii. $a\hat{\beta}_1$

iv. $\frac{c}{a}\hat{\beta}_1$

v. $c\hat{\beta}_1$

(ב) [15 נק'] $\hat{\beta}'_0$ שווה ל:

נקבל:

$\hat{\beta}_1' = \frac{s_y'}{s_x'} r_{xy}' = \frac{|c| s_y}{|a| s_x} \mathrm{sign}(ac) r_{xy} = \frac{c}{a} \cdot \frac{s_y}{s_x} r_{xy} = \frac{c}{a} \hat{\beta}_1$

$\hat{\beta}_0' = \bar{y}' - \hat{\beta}_1' \bar{x}' = c\bar{y} - \frac{c}{a}\hat{\beta}_1 a \bar{x} = c(\bar{y} - \hat{\beta}_1 \bar{x}) = c\hat{\beta}_0$

i. $\hat{\beta}_0$

ii. $\frac{a}{c}\hat{\beta}_0$

iii. $a\hat{\beta}_0$

iv. $\frac{c}{a}\hat{\beta}_0$

v. $c\hat{\beta}_0$

(ג) [15 נק'] ה-R'^2 של המדגם החדש D' במונחים של R^2 עבור המדגם המקורי D שווה ל:

$$R'^2 = r'^2_{xy} = \text{sign}(ac)^2 \, r_{xy}^2 = r_{xy}^2 = R^2$$

i. R^2
ii. $\frac{a}{c}R^2$
iii. $\frac{a^2}{c^2}R^2$
iv. $\frac{c}{a}R^2$
v. $\frac{c^2}{a^2}R^2$

3. [20 נק'] עבור n תצפיות מרגרסיה לינארית פשוטה עם המודל $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, יהי ϵ וקטור השגיאות ויהי $e = y - \hat{y}$ וקטור השאריות עבור **אותם הנתונים**. סמנו את התשובה הנכונה

(1 מסמן וקטור אחדות)

לפי הגדרת ריבועים פחותים:

$$\|e\|^2 = \|y - \hat\beta_0 1 - \hat\beta_1 x\|^2 = \min_{\beta_0', \beta_1'} \|y - \beta_0' 1 - \beta_1' x\|^2 \le \|y - \beta_0 1 - \beta_1 x\|^2 = \|\epsilon\|^2$$

עבור $n=2$ תמיד $\|e\|^2 = 0$

אבל בדר"כ השגיאות ϵ: לא מתאפסים ולכן $\|\epsilon\|^2 > 0$.

(א) תמיד $||\epsilon||^2 \le ||e||^2$ ויש מקרים בהם אי השוויון חזק
(ב) תמיד $||\epsilon||^2 \ge ||e||^2$ ויש מקרים בהם אי השוויון חזק
(ג) תמיד $||e||^2 = ||\epsilon||^2$
(ד) אף אחת מהתשובות לעיל איננה נכונה

4. [25 נק'] ידוע שבמודל רגרסיה לינארית **פשוטה ללא חותך** עם וקטורי נתונים x, y, אומד הרבועים הפחותים לשיפוע הוא $\hat\beta_1 = \frac{x^T y}{||x||^2}$. נניח כעת שיש לנו נתונים $\mathbf{X}, y$ עבור רגרסיה לינארית **מרובה** עם p משתנים, כך שהמטריצה $\mathbf{X}$ היא אורתוגונלית, כלומר $\mathbf{X^T X} = I_p$. מבצעים רגרסיה לינארית מרובה ומקבלים את אומד הרבועים הפחותים הרגיל $\hat\beta$. הוכיחו שבמקרה זה כל ערך $\hat\beta_j$ שהתקבל שווה לערך שהיה מתקבל עבור השיפוע אם היינו עושים רגרסיה פשוטה ללא חותך של y מול המשתנה X_j.

ברגרסיה מרובה: $\hat\beta = [X^T X]^{-1} X^T y \overset{\text{אורתוגונליות}}{=} X^T y$

לכן: $\hat\beta_j = [X^T y]_j = x_j^T y$ כאשר x_j מסמן עמודה j של X.

ברגרסיה פשוטה של x_j מול y: $\hat\beta = \frac{x_j^T y}{\|x_j\|^2} = x_j^T y$

כאשר $\|x_j\|^2 = x_j^T x_j = 1$ בגלל האורתוגונליות של X. לכן האומדים ב.ב. הם זהים.