רגרסיה ומודלים לינאריים 52320 תשע"ה 2014־15
בוחן 1 13.05.2015

בבוחן שאלות אמריקאיות ושאלות פתוחות.

משקל כל שאלה הוא 25 נקודות כך שמספר הנקודות הכולל הוא 125. 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 במחברת הטיוטה ולהעתיק את עיקר הפתרון אל הטופס רק לאחר בדיקה.

חומר עזר מותר: מחשבון.

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 $x^T y = \sum_{i=1}^n x_i y_i$.

תזכורת: נגדיר מודל לרגרסיה פשוטה עם חותך: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ עבור נתונים $(x_1, y_1), ..., (x_n, y_n)$. עבור מודל זה שגיאת הרבועים הפחותים $SSE = \sum_{\mathrm{i}=1}^n (y_i - \hat{y}_i)^2 = \sum_{\mathrm{i}=1}^n e_i^2$ ניתנת לכתיבה בצורות הבאות:

$SSE = \left(y - \hat{\beta}_0 - \hat{\beta}_1 x\right)^T \left(y - \hat{\beta}_0 - \hat{\beta}_1 x\right) = (y - \bar{y})^T (y - \bar{y}) - \hat{\beta}_1^2 (x - \bar{x})^T (x - \bar{x})$ כאשר האומדים למקדמים הם: $\hat{\beta}_1 = \frac{(x-\bar{x})^T (y-\bar{y})}{(x-\bar{x})^T (x-\bar{x})},\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

1. עבור מודל הרגרסיה הפשוטה עם חותך לעיל, יהיה $\hat{y}$ וקטור התחזיות ויהיה $e = y - \hat{y}$ וקטור השגיאות. אילו מהגדלים הבאים **תמיד** שווים לאפס? $\sum_{\mathrm{i}=1}^n e_i,\ \sum_{\mathrm{i}=1}^n x_i e_i,\ \sum_{\mathrm{i}=1}^n y_i e_i$

(א) כולם

(ב) אף אחד

(ג) רק $\sum_{\mathrm{i}=1}^n e_i, \sum_{\mathrm{i}=1}^n x_i e_i$

(ד) רק $\sum_{\mathrm{i}=1}^n e_i$

(ה) רק $\sum_{\mathrm{i}=1}^n y_i e_i, \sum_{\mathrm{i}=1}^n x_i e_i$

(ו) אף אחת מהתשובות לעיל אינה נכונה

פתרון:

תשובה (c) היא הנכונה. ראינו בכיתה ובתרגיל ש־$\sum_{\mathrm{i}=1}^n e_i = \sum_{\mathrm{i}=1}^n x_i e_i = 0$. לעומת זאת נניח את הנתונים הבאים עבור $n = 3$ תצפיות: $x_1 = 0, x_2 = 0, x_3 = 1; y_1 = 0, y_2 = 1, y_3 = 0.5$. אומדי הרבועים הפחותים למקדמי הרגרסיה המתקבלים הם $\hat{\beta}_0 = 0.5, \hat{\beta}_1 = 0$ כלומר $\hat{y}_i = 0.5$ ונקבל $e_1 = -0.5, e_2 = 0.5, e_3 = 0$ ולכן $\sum_{\mathrm{i}=1}^n y_i e_i = -0.5 \times 0 + 0.5 \times 1 + 0 \times 0.5 = 0.5 \neq 0$.

2. נניח כעת כי נתונים כל הסכומים הבאים עבור נתונים $(x_1, y_1), ..., (x_n, y_n)$:

$$S_x = \sum_{i=1}^n x_i = 6.1,\ S_y = \sum_{i=1}^n y_i = 42.6,\ S_{xx} = x^T x = 16.45,\ \ S_{yy} = y^T y = 99.02,\ S_{xy} = x^T y = 21.01,\ n = 20$$

חשבו את $\hat{\beta}_0, \hat{\beta}_1$ ואת ה־SSE עבור מודל הרגרסיה עם חותך והנתונים שלעיל.

פתרון:

נחשב תחילה את $\hat{\beta}_1$ ונקבל:

$$\hat{\beta}_1 = \frac{(x-\bar{x})^T (y-\bar{y})}{(x-\bar{x})^T (x-\bar{x})} = \frac{x^T y - \bar{x} S_y - \bar{y} S_x + n\bar{x}\bar{y}}{x^T x - 2\bar{x} S_x + n\bar{x}^2} = \frac{S_{xy} - \frac{1}{n} S_x S_y}{S_{xx} - \frac{1}{n} S_x^2} = \frac{n S_{xy} - S_x S_y}{n S_{xx} - S_x^2} = \frac{20 \times 21.01 - 6.1 \times 42.6}{20 \times 16.45 - 6.1^2} = 0.550$$

כעת נציב ונחשב את $\hat{\beta}_0$ ונקבל: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \frac{42.6}{20} - 0.550 \times \frac{6.1}{20} = 1.962$
והשגיאה הרבועית היא:
$SSE = (y-\bar{y})^T(y-\bar{y}) - \hat{\beta}_1^2(x-\bar{x})^T(x-\bar{x}) = S_{yy} - \frac{1}{n}S_y^2 - \hat{\beta}_1^2\left(S_{xx} - \frac{1}{n}S_x^2\right) = 99.02 - \frac{42.6^2}{20} - 0.550^2\left(16.45 - \frac{6.1^2}{20}\right) = 3.869$

3. נגדיר כעת מודל רגרסיה **ללא** חותך: $y_i = \gamma_1 x_i + \epsilon_i$ עבור נתונים $(x_1, y_1), ..., (x_n, y_n)$. הוכיחו שבמודל זה אומד הרבועים הפחותים $\hat{\gamma}_1$ עבור γ_1 שווה ל-$\frac{x^T y}{x^T x}$.

פתרון:

נחשב את השגיאה הרבועית כפונקציה של הפרמטר γ_1 ונקבל:

$$F(\gamma_1) = e^T e = (y - \gamma_1 x)^T (y - \gamma_1 x) \tag{1}$$

כעת נגזור לפי γ_1 ונקבל:

$$F'(\gamma_1) = \left(\sum_{i=1}^{n}(y_i - \gamma_1 x_i)^2\right)' = 2\sum_{i=1}^{n}(y_i - \gamma_1 x_i)(-x_i) = 0 \tag{2}$$

נעביר אגפים ונקבל:

$$\sum_{i=1}^{n} x_i y_i = \gamma_1 \sum_{i=1}^{n} x_i^2 \Rightarrow \gamma_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \frac{x^T y}{x^T x} \tag{3}$$

גזירה שניה מראה כי מדובר במינימום:

$$F''(\gamma_1) = \left(2\sum_{i=1}^{n}(y_i - \gamma_1 x_i)(-x_i)\right)' = 2\sum_{i=1}^{n} x_i^2 > 0 \tag{4}$$

לכן זהו מינימום מקומי. מכיוון שזהו המינימום היחיד והפונקציה לא חסומה כאשר $\gamma_1 \to \pm\infty$ אז זה חייב להיות המינימום הגלובלי.

4. חשבו את ערכו של $\hat{\gamma}_1$ ואת השגיאה הרבועית $SSE_1 \equiv \sum_{\mathrm{i}=1}^{n}(y_i - \hat{y}_i)^2$ כאשר $\hat{y}_i = \hat{\gamma}_1 x_i$ עבור המודל **בלי** החותך עם הנתונים בשאלה 2 לעיל

פתרון:

נחשב את $\hat{\gamma}_1$ על פי הנוסחא לעיל, ונקבל: $\hat{\gamma}_1 = \frac{S_{\mathrm{xy}}}{S_{\mathrm{xx}}} = \frac{21.01}{16.45} = 1.278$. כדי לחשב את השגיאה הרבועית, נצטרך להביע אותה באמצעות הגדלים הנתונים לנו.

$$SSE_1 = e^T e = (y-\hat{\gamma}_1 x)^T (y-\hat{\gamma}_1 x) = y^T y - 2\hat{\gamma}_1 y^T x + \hat{\gamma}_1^2 x^T x = 99.02 - 2\times 1.278 \times 21.01 + 1.278^2 \times 16.45 = 72.19 \tag{5}$$

5. באופן כללי, עבור רגרסיה לינארית פשוטה, נסמן ב־SSE את השגיאה הרבועית במודל עם חותך ואת SSE_1 את השגיאה הרבועית במודל ללא חותך עבור **אותם הנתונים**. סמנו את התשובה הנכונה

(א) תמיד $SSE_1 \leq SSE$ ויש מקרים בהם אי השוויון חזק

(ב) תמיד $SSE_1 \geq SSE$ ויש מקרים בהם אי השוויון חזק

(ג) תמיד $SSE_1 = SSE$

(ד) אף אחת מהתשובות לעיל איננה נכונה

פתרון:

תשובה (b) היא הנכונה - במודל ללא חותך אנחנו עושים התאמה למודל עם פרמטר אחד פחות. לכן טיב ההתאמה לא יכול להשתפר. עבור כל ערך של $\hat{\gamma}_1$ במודל ללא חותך ניתן לקבל בדיוק את אותו מודל, ולכן אותה שגיאה ריבועית, אם נבחר $\hat{\beta}_1 = \hat{\gamma}_1, \hat{\beta}_0 = 0$ במודל עם החותך. בניסוח מתמטי:

$$SSE = min_{\beta_0,\beta_1\in\mathbb{R}}(y-\beta_0-\beta_1 x)^T(y-\beta_0-\beta_1 x)^T \leq min_{\beta_1\in\mathbb{R}}(y-0-\beta_1 x)^T(y-0-\beta_1 x)^T = min_{\gamma_1\in\mathbb{R}}(y-\gamma_1 x)^T(y-\gamma_1 x)^T = SSE_1 \tag{6}$$