# רגרסיה ומודלים סטטיסטיים - בוחן 2

אביב 2021

## שאלה 1

רגרסיית רידג׳ (Ridge) היא אלטרנטיבה לשיטת הריבועים הפחותים עבור אמידה במודל הליניארי. רגרסיית רידג׳ מתאימה במיוחד לאמידה של $\beta$ הקיצוני שבו יש ממש תלות ליניארית בין עמודות של מטריצת ה-X, ושנית, השונות של האומד רגישה הרבה פחות למולטיקוליניאריות בהשוואה לאומדי הריבועים הפחותים, כך שהוא יכול להשיג שגיאת אמידה נמוכה משמעותית במצבים כאלה. לאורך כל השאלה נניח שמתקיים המודל הליניארי הכללי

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon} \tag{1}$$

כאשר $\mathbb{E}[\boldsymbol{\epsilon}] = 0$ ו- $Cov(\boldsymbol{\epsilon}) = \sigma^2 I$. נזכיר כי אומד הריבועים הפחותים נתון על ידי

$$\hat{\beta}^{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{2}$$

כלומר, הוא מקיים

$$\hat{\beta}^{OLS} = \text{arg}\min_b ||\boldsymbol{Y} - \boldsymbol{X}b||_2^2 \tag{3}$$

האומד ברגרסית Ridge מוגדר על ידי

$$\hat{\beta}^{Ridge} = \text{arg}\min_b \mathcal{H}(b) \tag{4}$$

עבור

$$\mathcal{H}(b) = ||\boldsymbol{Y} - \boldsymbol{X}b||_2^2 + \lambda||b||_2^2 \tag{5}$$

כאשר $\lambda > 0$ הוא גודל שנבחר (כקבוע מראש או כפונקציה של הנתונים) ע"י המשתמש . כלומר, אם משווים את פונקציות המטרה ב- (4) ו- (5) , אז Ridge מוסיף "קנס" פרופורציונלי לנורמה הריבועית של $b$.

1. נניח כי ל-$\boldsymbol{X}$ יש מספר עמודות תלויות ליניארית, כלומר קיים $c \neq 0 \in \mathbb{R}^{p+1}$ כך שמתקיים $\boldsymbol{X}c \approx 0$. הראו שבמצב זה $\boldsymbol{X}^T\boldsymbol{X}$ לא הפיכה.
   *הערה* - אפשר להשתמש בעובדה שמטריצה הפיכה אם ורק אם הגרעין שלה לא טריוויאלי.
2. הוכיחו כי לכל $\lambda > 0$ המטריצה $\boldsymbol{X}^T\boldsymbol{X} + \lambda I$ הפיכה.

*הדרכה* - ניתן להיעזר בעובדה שמטריצה $A$ הפיכה אם ורק אם הפתרון למשוואה $Ax = 0$ הוא הפתרון הטריוויאלי $x = 0$.

3. הראו כי

$$\hat{\beta}^{Ridge} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda I)^{-1}\boldsymbol{X}^T\boldsymbol{Y} \tag{6}$$

*הדרכה* - אפשר להשתמש בהשלמה לריבוע, כלומר בעובדה שעבור וקטור $x$ ומטריצה סימטרית $A$ מתקיים

$$x^TAx + x^Tc + d = (x-h)^TA(x-h) + k$$

כאשר

$$h = -\frac{1}{2}a^{-1}c, \quad k = d - \frac{1}{4}c^TA^{-1}c$$

4. הראו כי ניתן לכתוב

$$\hat{\beta}^{Ridge} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda I)^{-1}\boldsymbol{X}^T\hat{\boldsymbol{Y}} \tag{7}$$

עבור $\hat{\boldsymbol{Y}} = P_X\boldsymbol{Y}$ ו- $P_X$ זה מטריצת ההטלה למרחב שנפרש ע"י העמודות של $\boldsymbol{X}$.

5. חשבו את $\mathbb{E}[\hat{\beta}^{Ridge}]$ ואת $Cov(\hat{\beta}^{Ridge})$.

6. עבור אומד כלשהו ל- $\beta$ שנסמנו ב- $\hat{\beta}$, נגדיר את ההטייה של $\hat{\beta}$ להיות $Bias(\hat{\beta}) = \mathbb{E}[\hat{\beta}] - \beta$. חשבו את $Bias(\hat{\beta}^{Ridge})$.

7. בצורה דומה לסעיף הקודם, נגדיר את תוחלת סכום השגיאות הריבועיות באמידה,

$$MSE[\hat{\beta}] = \mathbb{E}\left[\sum_{j=0}^{p}(\hat{\beta}_j - \beta_j)^2\right] \tag{8}$$

הראו שלכל $j \in \{0, \ldots, p\}$ מתקיים

$$\mathbb{E}[(\hat{\beta}_j - \beta_j)^2] = Var(\hat{\beta}_j) + Bias^2(\hat{\beta}_j) \tag{9}$$

כלומר, לשגיאה הריבועית תורמות השונות וההטייה (בריבוע) של האומד.

8. היעזרו בסעיפים הקודמים כדי למצוא ביטוי ל- $MSE[\hat{\beta}^{Ridge}]$.

כעת, נבצע מספר סימולציות אשר ימחישו את התוצאות התיאורטיות שראינו. בתיבת ההגשה במודל יש קובץ נתונים $quiz2\_df.csv$ וקובץ R בשם $Quiz2.R$. קבצים אלו ישמשו אותנו בסעיפים הבאים. בקובץ הנתונים יש משתנים מסבירים X ומשתנה מוסבר Y . בשביל לקבל תוצאות הגיוניות ברגרסית Ridge יש לעשות תקנון (Scale) של X לפי כל עמודה (כלומר, כל עמודה של X , למעט החותך, צריכה להיות עם תוחלת 0 וסטיית תקן 1).

9. השתמשו בכלים שנלמדו בכיתה כדי לבדוק את מידת הקוליניאריות במטריצת X, ודונו בקצרה בתוצאות.

10. בקובץ $Quiz2.R$ נתונים הערכים האמיתיים של $\beta, \sigma^2$ וכן שני וקטורים של ערכי $\lambda$ שנרצה לבדוק. בנוסף, בקובץ יש את הפונקציה $ridge\_aux\_functions()$ עם הסברים לגבי הקלטים והפלטים שלה. תשלימו את הנדרש בהסבר הפונקציה, כלומר תכתבו את הפונקציה לפי ההדרכה הנמצאת בקובץ R הנתון.

11. עבור כל אחד מערכי $\lambda$ בוקטור $lambda\_seq$ , חשבו את (8) עבור $\hat{\beta}^{OLS}$ ו- $\hat{\beta}^{Ridge}$. את התוצאות שהתקבלו הציגו בגרף שבציר ה-X יש את ערכי $\lambda$ בסדר עולה (משמאל לימין) ובציר ה-Y את ערכי ה- MSE שהתקבלו עבור כל מודל ועבור כל $\lambda$. דונו בתוצאות.

12. עבור כל אחד מערכי $\lambda$ בוקטור $lambda\_seq\_mod$ , חשבו את $\hat{\beta}_j^{Ridge},\ j = 0, \ldots, p$. הציגו את התוצאות בגרף שבו בציר ה-x יש את ערכי $\lambda$ ובציר ה-Y את ערכי $\hat{\beta}_j^{Ridge}$ השונים. לכל משתנה $j \in \{0, \ldots, p\}$ הציגו את הגרף בצבע שונה. דונו בקצרה בתוצאות.

## שאלה 2

הקובץ המצורף $eagles\_sim.csv$ כולל תצפיות (מסימולציה מבוססת נתונים אמיתיים) על העיט הקירח Bald eagle בארה״ב, זן של נשר שהיה בסכנת הכחדה בסביבות מחצית המאה העשרים. שינוי במדיניות לגבי שימוש בחומרי הדברה שנכנס לתוקף ב-1972 הביא לתפנית ועלייה חדה בהתרבות הזן הנכחד, ומדענים מעוניינים לנתח את קצב הגידול באוכלוסיית הנשרים במשך העשורים העוקבים. בקובץ מופיעים מספר זוגות נצפים של נשרים (y) לפי מספר השנים שחלפו מאז 1950 (x) .

1. צרו תרשים פיזור של y כנגד x , ושימו לב שהנתונים אכן רומזים על קשר אקספוננציאלי בין משתנה התוצאה והמשתנה המסביר. הציעו טרנספורמציה מתאימה ששתעביר את y למשתנה חדש z , בנסיון לשחזר מודל ליניארי.

2. עבור רגרסיה של z על x , חשבו את השאריות המתוקננות, והציגו תרשימים מתאימים לבדיקת כל אחת מהנחות הליניאריות, שוויון-השונויות והנורמליות של השגיאות: עבור כל אחת מההנחות, ציינו אילו מהגרפים מתאימים לבדיקתה, והבחינו שאין אינדיקציה חזקה להפרה של ההנחות.

3. נניח שישנה תצפית משנת 1977 $(x = 27)$ שהושמטה מקובץ הנתונים המקורי. השתמשו בנתונים המקוריים כדי לבנות רווח-חיזוי ברמה $90\%$ למספר זוגות הנשרים עבור התצפית החסרה. השתמשו בעובדה ש-y ו-z קשורים דרך טרנספורמציה מונטונית כדי להצדיק את התוקף של רווח-החיזוי שבניתם, כלומר, הוכיחו שהרווח שבניתם אכן בעל רמת כיסוי $90\%$ . בסעיף זה הניחו שהמודל הליניארי הנורמלי תקף עבור רגרסיה של z על x .