רגרסיה ומודלים סטטיסטיים ־ בוחן 1

<u>הנחיות כלליות</u>: הפרידו בצורה ברורה את התשובות לכל אחד מהסעיפים. הסבירו ופרטו, תשובות לא מנומקות לא יזכו בניקוד. בפרט, הקפידו על ההנחיות המופיעות בקובץ Code Submission Instructions הן לגבי ההגשה והן לגבי הקוד.

1. (30 נקודות) בשאלה זו נעסוק במודל WLS תחת ההנחות הבאות:

$$Y_i = \sum_{j=0}^{p} \beta_j X_{ij} + \epsilon_i$$
$$\epsilon_i \sim N\left(0, v_i\sigma^2\right)$$
$$\text{Cov}\left(\epsilon_{i_1}, \epsilon_{i_2}\right) = 0 \; \forall i_1 \neq i_2$$

כאשר $w_i = \frac{1}{v_i}$ ו־v_i הם מספרים חיוביים ידועים.

א. נסחו במופרש את משפט גאוס מרקוב עבור האומדים הממזערים את סכום ריבועי הטעויות הממשוקל $s_w(b)$.

ב. הוכיחו את המשפט שניסחתן בסעיף א'.

2. <u>הקדמה:</u>

בשאלה זו נעסוק ב־ Ridge Regression. השימוש ב ־Ridge Regression נעשה לרוב באחד משני מצבים: לטיפול במצב שבו ישנה מולטיקולינאריות חלקית חזקה בין משתנים מסבירים, או במצב שבו מספר המשתנים המסבירים גדול משמעותית ממספר התצפיות.

לצורך שאלה זו, אנו נתמקד בטיפול במולטיקולינאריות. Ridge Regression היא הדוגמה הפשוטה ביותר מקבוצת ה־ penalized regression methods ונעשה בה שימוש בעיקר כאשר אמידת המודל נעשית במטרה לייצר תחזיות טובות, להבדיל ממצב בו אנו נתעניין בהסברים על ההשפעה של כל משתנה מוסבר על המשתנה המסביר.

המודל שנעסוק בו הוא מהצורה שבה עסקנו בקורס עד כה, כלומר

$$Y = X\beta + \epsilon$$

עם p משתנים מסבירים, כאשר מניחים כי $E\left[\epsilon\right]=0$ ו־ $\text{Cov}\left(\epsilon\right)=\sigma^{2}I$. במודל OLS ראינו כי האומד מתקבל על ידי

$$\hat{\beta}_{\text{OLS}}=\text{argmin}_{\beta}\left\Vert Y-\hat{Y}\right\Vert _{2}^{2}=\text{argmin}_{\beta}\left\Vert Y-X\beta\right\Vert _{2}^{2}$$

כלומר על ידי מזעור סכום ריבועי הטעויות, שכן כפי שראינו בתרגול חזרה הראשון:

$$\left\Vert w\right\Vert _{2}^{2}=\left(\sum_{k}w_{k}^{2}\right)^{1/2}.$$

ב־ Ridge Regression נרצה (במקום למזער את סכום ריבועי הטעויות), למזער את הביטוי

$$\mathcal{H}\left(\beta\right)=\left\Vert Y-X\beta\right\Vert _{2}^{2}+\alpha\left\Vert \beta\right\Vert _{2}^{2}$$

כלומר

$$\hat{\beta}_{\text{Ridge}}=\text{argmin}_{\beta}\mathcal{H}\left(\beta\right)=\text{argmin}_{\beta}\left(\left\Vert Y-X\beta\right\Vert _{2}^{2}+\alpha\left\Vert \beta\right\Vert _{2}^{2}\right)$$

כאשר α היא ערך קבוע שנקבע על ידי מי שאומד את הרגרסיה.

כאשר אומדים Ridge Regression, נהוג לנרמל את המשתנים המסבירים כך שממוצע כל אחד מהמשתנים המסבירים יהיה 0 והשונות המדגמית 1. כמו כן, נהוג לנרמל את המשתנה המוסבר Y באופן דומה ובעקבות כך להשמיט את החותך β_0 מהמודל. בשאלה זו אנו נניח כי השינויים הללו נעשו.

<u>נוסחה חלופית לאומד:</u>

נייצג את המטריצה X^TX על ידי הפירוק הספקטרלי שלה:

$$X^{T}X=V\Lambda V^{T}$$

מכיוון ש־X^TX היא מטריצה חיובית לחלוטין (ראו תזכורת בתרגול חזרה הראשון אם יש צורך), כל הערכים של Λ הם חיוביים. הניחו כי הם מסודרים בסדר יורד. נסמן ב־$\Lambda^{1/2}$ את המטריצה האלכסונית שאבריה באלכסון הם השורשים הריבועיים של איברי האלכסון של Λ. נגדיר:

$$\begin{aligned}G&=\Lambda^{1/2}V^{T}\\ \theta&=G\beta\\ \tilde{X}&=XG^{-1}\\ \tilde{Y}&=\tilde{X}^{T}Y\end{aligned}$$

שימו לב כי $\tilde{Y}$ הוא וקטור באורך p.

<u>השאלה</u> (35 נקודות):

א. הוכיחו כי

$$\hat{\beta}_{\text{Ridge}} = \left(X^T X + \alpha I\right)^{-1} X^T Y$$

ב. חשבו את $E\left[\hat{\beta}_{\text{Ridge}}\right]$ ואת $\text{Cov}\left(\hat{\beta}_{\text{Ridge}}\right)$.

ג. כעת נתבונן בהצגה החלופית. חשבו את $\tilde{X}^T \tilde{X}$

ד. רשמו את פונקציית המטרה $\mathcal{H}\left(\beta\right)$ כפונקציה של θ וסמנו אותה ב־ $\mathcal{H}\left(\theta\right)$

ה. הוכיחו כי איברי הערך θ הממזער את $\mathcal{H}\left(\theta\right)$ ניתנים על ידי

$$\hat{\theta}_j = \frac{\tilde{Y}_j}{1 + \alpha \lambda_j^{-1}}$$

ו. הגדירו $\hat{Y} = X\hat{\beta}_{\text{Ridge}}$ ורשמו ביטוי מפורש למטריצה S_α המקיימת $\hat{Y} = S_\alpha Y$. מטריצה זו נקראת smoother matrix.

ז. הראו כי

$$\sum_{i=1}^{n} \text{Var}\left(\hat{Y}_i\right) = \sigma_\epsilon^2 \text{trace}\left(S_\alpha^2\right)$$

3. (35 נקודות) בשאלה זו נחזור לעסוק באמידת OLS לרגרסיה מרובת משתנים:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots \beta_p X_{pi} + \epsilon_i$$

באתר הקורס מופיע קובץ הנתונים GM_cars.csv.

קובץ זה מכיל נתונים לגבי מעל 800 מכוניות של General Motors. כמו כן מופיע קובץ GM_cars_desc שבו תמצאו את תיאור המשתנים בקובץ. המטרה שלנו תהיה לחזות את מחיר המכונית בהינתן יתר המשתנים המסבירים ולפתח מודל שמתאר את ההשפעה של כל אחד מן המשתנים על המחיר.

<u>מודל עם המשתנה מסביר יחיד</u>: אמדו תחילה את המחיר כפונקציה של Mileage בלבד.

א1. התבוננו באומדים שקיבלתם. איזו השפעה יש למרחק שמכונית עוברת על מחיר המכונית? מה יהיה ההבדל במחירן של 2 מכוניות זהות מלבד כך שמכונית אחת עברה $50,000$ מיילים יותר?

א2. התבוננו בתוצאות של מבחן t, מה ניתן להסיק מהן?

א3. התבוננו במדדים R^2, adjusted R^2, C_p, AIC, BIC. הסבירו את משמעתם.

א4. האם קיימת סתירה בין התוצאות של מבחן t ומבחן R^2? הסבירו.

ב1. בצעו scatter plot של משתנה המחיר (Price) אל מול המרחק שעברה המכונית (Mileage). מה ניתן ללמוד ממנו על ההשפעה של המרחק על המחיר?

ב2. בגרף אתן תראו תצפיות עם מחירים חריגים. האם מותר להסיר אותן? נסו להסביר מדוע מחירן חריג.

ב3. האם אתם מבחינים בהשפעה שונה של המרחק על המחיר עבור תצפיות אלו מיתר התצפיות?

<u>מודל מרובה משתנים:</u> כעת בנוסף למשתנה המסביר Mileage הוסיפו למודל גם את המשתנים המסבירים הבאים: Cylinder, Doors, Cruise, Sound, Leather, Liter.

ג1. כיצד השתנה האומד להשפעת המשתנה Mileage? מדוע?

ג2. התבוננו בערכי R^2, adjusted R^2, C_p, AIC, BIC והשוו לסעיף א3'. מה ניתן ללמוד מכך?

ג3. התבוננו בתוצאות של מבחן t של כל אחד מהאומדים. על מה תוצאות אלו יכולות להעיד?

ג4. האם אתן חושבות שיש לבצע שינוי במודל בעקבותיהם? אם כן, פרטו איזה (אין צורך לבצע).

ד1. בצעו ניתוח שאריות. לשם כך בצעו היסטוגרמה ו QQ-Plot. אם אתם רואים צורך אתם יכולים להיעזר גם ביתר הגרפים שלמדנו. האם השאריות מתפלגות נורמלית? האם יש צורך לנקוט פעולות כלשהן? אם כן, אילו?

ד2. האם נראה כי קיימת הטרוסקדסטיות ($\mathrm{Var}(\epsilon_i)$ אינו זהה לכל i) במודל? הסבירו.

ד3. בחנו האם קיימת מולטיקולינאריות בין המשתנים במודל. אם כן, הסבירו בין אילו משתנים ואיך זה לדעתכן השפיע על התוצאות בסעיפים הקודמים.