מועד א' - שאלות על שבועות 1-4

שאלה קשה

שאלה זו עוסקת במודל לינארי בו הרעש אינו בעל התפלגות גאוסיינית.
עבור המודל הלינארי הרגיל מהצורה

$$y_i = \sum_{j=1}^{p} x_{ji}\beta_j + \epsilon_i,\ i = 1, \dots, n$$

נתבונן במקרה בו $\epsilon_i \overset{iid}{\sim} Laplace(0, \sigma)$.
למשתנה מקרי $W \sim Laplace(\mu, \sigma)$ יש צפיפות

$$f_W(w) = \frac{1}{2\sigma} \exp\left(-\frac{|w-\mu|}{\sigma}\right),\ w \in \mathbb{R}$$

כמו כן, אם $Z \sim Laplace(0, \sigma)$ אז עבור כל $a \in \mathbb{R}$ מתקיים $W = Z + a \sim Laplace(a, \sigma)$

א. מצאו את ההתפלגות והצפיפות של y_i

ב. מצאו באמצעות אינטגרציה את $E[y_i]$

ג. הראו, תחת המודל שתואר, שאומד נראות מרבית ל- $\beta \in \mathbb{R}^p$ שקול לפתרון בעיית האופטימיזציה

$$\arg\min_{\beta} \sum_{i=1}^{n} |y_i - \sum_{j=1}^{p} x_{ji}\beta_j| = \arg\min_{\beta} ||Y - X\beta||_1$$

ד. הסבירו מהו הקושי בפתרון הבעיה שתוארה בסעיף ג' והציעו פתרונות אפשריים.

ה. תחת המודל הרגיל שתואר בקורס (כלומר הרעש בעלת התפלגות גאוסיינית), הסבירו מה הם היתרונות של מציאת אומדים ל β באמצעות השיטה שתוארה בסעיף ג'.

א. מהרמז מתקבל ישירות כי

$$y_i \sim Laplace(\sum_{j=1}^{p} x_{ji}\beta_j, \sigma)$$

וכן

$$f_{y_i}(y) = \frac{1}{2\sigma} \exp\left(-\frac{|y - \sum_{j=1}^{p} x_{ji}\beta_j|}{\sigma}\right),\ y \in \mathbb{R}$$

ב. נסמן $\sum_{j=1}^{p} x_{ji}\beta_j = X_i\beta$

$$
\begin{aligned}
Ey_i =& \frac{1}{2\sigma}\int_{-\infty}^{\infty} y e^{-\frac{|y-X_i\beta|}{\sigma}} dy \\
=& \begin{bmatrix} t = y - X_i\beta \\ dt = dy \end{bmatrix} \\
=& \frac{1}{2\sigma}\int_{-\infty}^{\infty} (t + X_i\beta) e^{-\frac{|t|}{\sigma}} dt \\
\overset{*}{=}& \frac{1}{2\sigma}\int_{-\infty}^{\infty} t e^{-\frac{|t|}{\sigma}} dt + X_i\beta \\
=& \frac{1}{2\sigma}\int_{0}^{\infty} t e^{-\frac{t}{\sigma}} dt + \frac{1}{2\sigma}\int_{-\infty}^{0} t e^{\frac{t}{\sigma}} dt + X_i\beta \\
=& -\frac{1}{2\sigma}\int_{-\infty}^{0} k e^{\frac{k}{\sigma}} dk + \frac{1}{2\sigma}\int_{-\infty}^{0} t e^{\frac{t}{\sigma}} dt + X_i\beta \\
=& X_i\beta
\end{aligned}
$$

כאשר * נובע מהגדרת צפיפות של התפלגות Laplace.

ג. פונ' הנראות היא

$$L(Y;\beta) = \prod_{i=1}^{n} \frac{1}{2\sigma} \exp\left(-\frac{|y_i - X_i\beta|}{\sigma}\right)$$

לכן לוג-נראות היא

$$\ell(Y;\beta) = -n\log(2\sigma) - \sum_{i=1}^{n} \frac{|y_i - X_i\beta|}{\sigma}$$

מכיוון שאנו רוצים למקסם את הנראות והרכיב השני בנראות הוא תמיד אי שלילי (כי $\sigma > 0$) אז נקבל שבעית מקס' נראות שקולה לבעית המינימזציה של הרכיב השני, כלומר שקולה לפתרון

$$\text{minimize} \sum_{i=1}^{n} |y_i - X_i\beta| = \text{minimize } ||Y - XB||_1$$

ד. הקושי המרכזי הוא שפונ' ערך מוחלט איננה גזירה בנק' 0. זה גורם לכך שלא נוכל למצוא ביטוי סגור לפתרון הבעיה כמו שהיה לנו בOLS. פתרונות אפשריים זה קירובים נומרים כלשהם לפתרון המינימזציה (הם ראו ניוטון-רפסון בקורס, מספיק לציין דרך כללית לא אלג' מא' עד ת'). אפשר לקבל תשובות עם אחרות עם הסבר הגיוני כאן.

ה. היתרון המרכזי הוא שנורמה ℓ_1 פחות רגישה לערכים חריגים (משהו בין huber לOLS).

שאלות בינוניות

1. שאלה זו עוסקת ב- Bias-Variance tradeoff
עבור אומד $\hat{\theta}$ לפרמטר θ נגדיר

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$$

ובנוסף נגדיר

$$bias(\hat{\theta}) = E[\hat{\theta}] - \theta$$

א. הראו כי מתקיים

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + bias^2(\hat{\theta})$$

ב. מציאת אומדים ברגרסית Ridge היא הפתרון של הבעיה

$$\hat{\beta}_{Ridge} = \arg\min_{\beta} ||Y - XB||_2^2 + \lambda||\beta||_2^2$$

עבור $\lambda \geq 0$.
בבוחן 1 ראינו כי האומד שמתקבל הינו

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

עבור המקרה של $p = 1$, כלומר $X^T X$ הינו סקלר, הראו כי מתקיים

i

$$bias(\hat{\beta}_{Ridge}) \neq 0$$

ii

$$Var(\hat{\beta}_{Ridge}) \leq Var(\hat{\beta}_{OLS})$$

כאשר $\hat{\beta}_{OLS}$ זה האומד הרגיל של הרגרסיה הלינארית.

ג. הסבירו, בהתבסות על סעיף א', את המשמעות של התוצאות שהתקבלו בסעיף ב'.

א.

$$\begin{aligned}
MSE(\hat{\theta}) =& E[(\hat{\theta} - \theta)^2] \\
=& E[((\hat{\theta} - E[\hat{\theta}]) + (E[\hat{\theta}] - \theta))^2] \\
=& E[(\hat{\theta} - E[\hat{\theta}])^2] + E[(E[\hat{\theta}] - \theta)^2] + 2E[(\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta)] \\
=& Var(\hat{\theta}) + E(bias(\hat{\theta})^2) + 2(E[\hat{\theta}] - \theta) \cdot E[\hat{\theta} - E[\hat{\theta}]] \\
=& Var(\hat{\theta}) + bias(\hat{\theta})^2 + 0 \\
=& Var(\hat{\theta}) + bias(\hat{\theta})^2
\end{aligned}$$

ב. i.

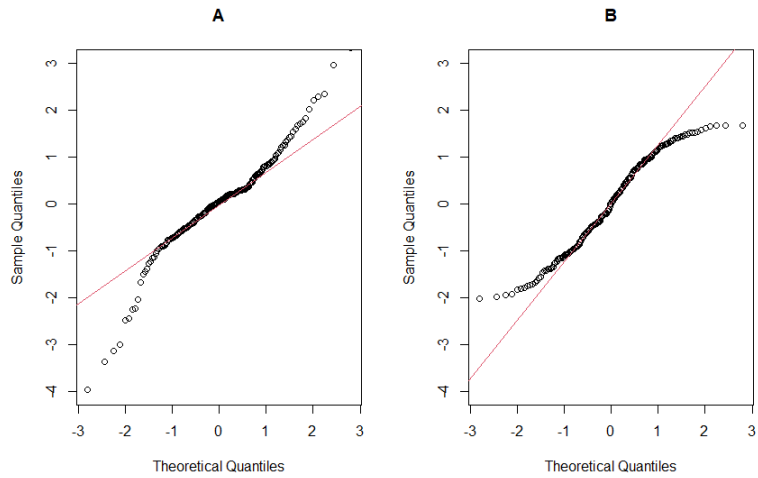
$$\begin{aligned}
E\hat{\beta}_{Ridge} =& E(X^T X + \lambda I)^{-1} X^T Y \\
=& (X^T X + \lambda I)^{-1} X^T (X\beta + E\epsilon) \\
=& (X^T X + \lambda I)^{-1} X^T X\beta \\
=& \frac{X^T X}{X^T X + \lambda}\beta \neq \beta
\end{aligned}$$

ii.

$$
\begin{aligned}
Var(\hat{\beta}_{Ridge}) =& (X^TX+\lambda I)^{-1}X^T \cdot (Var(Y)) \cdot X(X^TX+\lambda I)^{-1} \\
=& \sigma^2 (X^TX+\lambda I)^{-1}X^TX(X^TX+\lambda I)^{-1} \\
=& \frac{X^TX}{(X^TX+\lambda)^2}\sigma^2 \\
\leq& \frac{X^TX}{(X^TX)^2}\sigma^2 \\
=& \frac{1}{X^TX}\sigma^2 = Var(\hat{\beta}_{OLS})
\end{aligned}
$$

ג. פירוש רגיל של Bias-Variance, tradeoff כלומר הגדלנו הטייה אבל הקטנו שונות ולבסוף ייתכן (ובמקרה זה יש) תחלופה ביניהם כך שMSE של האומד יקטן, כלומר הקטנו את הסיכון הריבועי באמידה.

2. שאלה זו עוסקת בניתוח התפלגות השאריות.
נתונים שני גרפים אשר משווים את השארית המתוקננת להתפלגות נורמלית סטנדרטית בשני מודלים של רגרסיה לינארית מרובת משתנים.



א. הסבירו מה ניתן ללמוד מגרפים כאלה על הנחות המודל

ב. עבור כל אחד מהגרפים הסבירו מה היא הסטייה מהנחות המודל שהתבצעה

ג. עבור כל אחד מהמודלים הציעו שיטה שתשפר את התאמת המודל והסבירו כיצד ומדוע שיטה זו תעבוד.

ד. הסבירו האם ניתן להסיק מגרפים אלו על קיום מולטיקולינאריות בין המשתנים המסבירים במודל

א. גרפים אלו מלמדים בעיקר על ההנחה של התפלגות השאריות, כלומר סטיה מהקו $y = x$ תוביל לסטיה מסוימת מהנחות המודל (עם רעש מסוים כמובן)

ב. בגרף A השארית המתוקננת בעלת זנבות (ימינים ושמאליים) עבים יותר מההתפלגות הנורמלית הסטנדרטית, שהרי מצד אחד בחלק הימני (שגיאה חיוביות) מקבלים אחוזונים גבוהים יותר מהצפוי ובחלק השמאלי (שגיאה שלילית) מקבלים אחוזונים נמוכים יותר מהצפוי. גרף B הוא בדיוק ההפך, זנבות דקים יותר מהתפלגות נורמלית.
לצורך העניין, בגרף A הטעויות האמיתיות מתפלגות t ובגרף B מתפלגות אחיד ומכאן ההבדל.

ג. עבור A אפשר לעשות טרנספורמציה קעורה וB אפשר לעשות טרנספורמציה קמורה על Y, זה יגרום ל QQ Plot להראות קרוב יותר לנדרש.

ד. לא.

שאלה קלה

נתון הפלט הבא שהתקבל מהרצת מודל לינארי רגיל בתוכנה R.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.131225   0.283877  -0.462    0.644
x1          -0.002499   0.007238  -0.345    0.730
x2           1.887137    A        18.622  < 2e-16 ***
x3          -1.199656   0.264541  -4.535 1.01e-05 ***
x4              B       0.143885  34.227  < 2e-16 ***
x5           3.446345   4.421579   0.779     C
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.892 on 194 degrees of freedom
Multiple R-squared:  0.8856,    Adjusted R-squared: D
F-statistic: 300.3 on E and 194 DF,  p-value: < 2.2e-16
```

מצאו את הערכים המספריים של כל הגדלים המסומנים באדום .A,B,C,D,E

פתרון

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.131225   0.283877  -0.462    0.644
x1          -0.002499   0.007238  -0.345    0.730
x2           1.887137   0.101342  18.622  < 2e-16 ***
x3          -1.199656   0.264541  -4.535 1.01e-05 ***
x4           4.924741   0.143885  34.227  < 2e-16 ***
x5           3.446345   4.421579   0.779    0.437
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.892 on 194 degrees of freedom
Multiple R-squared:  0.8856,    Adjusted R-squared:  0.8826
F-statistic: 300.3 on 5 and 194 DF,  p-value: < 2.2e-16
```