רגרסיה ומודלים סטטיסטיים- תרגיל 3

שאלה 1:

יהי $Z = (Z_1, \dots, Z_n)^T$ וקטור מקרי מהתפלגות משותפת f_Z. עבור המודלים הבאים, מצאו את וקטור התוחלות ומטריצת השונויות של Z:

א. $Z_i \sim N(i, i^2), i = 1, .., n$ ב״ת.

ב. $Z_1 = \eta_1$ ונגדיר $Z_i = 0.5 \cdot \eta_{i-1} + \eta_i, i = 2, \dots, n$ כאשר $\eta_i \sim N(0,1), iid$.

ג. בעבור $j = 1, .. k$ נגדיר

$P(X_j = 1) = p, P(X_j = 2) = q, P(X_j = 3) = 1 - p - q$. וכן:

$Z_i = \sum_{j=1}^{k} 1_{\{X_j = i\}}, i = 1,2,3$.

שאלה 2:

יהיו $Z, W \in R^p$ וקטורים מקריים. הראו שהבאים שקולים:

(1) $\forall v \in R^p, Var(v^T Z) \ge Var(v^T W)$

(2) $B \coloneqq Var(Z) - Var(W)$ היא מטריצה חיובית למחצה.

(3) קיימת המטריצה $B^{\frac{1}{2}}$.

שאלה 3:

יהיו X, Y וקטורים מקריים עם התפלגות משותפת עם תוחלות μ_x, μ_y בהתאמה. נגדיר:

$$\Sigma_X = E[(X - \mu_x)(X - \mu_x)^T]$$

$$Cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)^T]$$

הוכיחו את התכונות הבאות:

(a) $\Sigma_x = \mathbb{E}XX^T - \mu_x \mu_x^T$

(b) $\Sigma_x \ge 0$ (The covariance matrix of X is positive semidefinite)

(c) $cov(AX + b) = A\Sigma_x A^T$

(d) $cov(X, Y) = cov(Y, X)^T$

(e) $cov(X_1 + X_2, Y) = cov(X_1, Y) + cov(X_2, Y)$

(f) $cov(AX, BY) = A[cov(X, Y)]B^T$

שאלה 4:

הניחו מודל רגרסיה לינארית פשוט (תרגיל 1 שאלה 4) בו המשתנה המוסבר הוא לחץ הדם (Y) והמשתנה המסביר הוא הגיל בשנים (X). הניחו שבידיכם דגימות של 100 מטופלים כאשר לכל מטופל מופיע הגיל ולחץ הדם שנמדד עבורו. הראו (מתמטית!) אילו מהנחות המודל הלינארי מופרות בכל אחד מהמקרים הבאים:

א. משיקולי תקציב הנדגמים הגיעו מ-20 משפחות בנות 5 נפשות שנדגמו באופן מקרי מכלל האוכלוסיה. ידוע כי ישנו קשר בין הגנטיקה ובין נטייה למחלות הקשורות ללחץ הדם.
ב. ידוע שהמינון של תרופות <u>המפחיתות</u> לחץ דם עולה עם הגיל.

חשבו את התוחלת והשונות של $\hat{\beta}_{OLS} = (X^TX)^{-1}X^TY$ בכל אחד מהמקרים.

שאלה 5:

בשאלה זו נבחן את הנחות המודל הלינארי דרך סימולציה:

א. ראשית הניחו כי כל הנחות המודל הלינארי מתקיימות. מצאו ביטוי להטייה ולמטריצת השונויות של $\hat{\beta}_{OLS}$.

ב. באמצעות הספרייה MASS והפקודה mvrnorm(num_vectors, mu, Sigma) הגרילו מהתפלגות נורמלית 5-מימדית 500 וקטורים ב"ת, כאשר:
Sigma- מטריצת היחידה ממימד 5.
mu- הוקטור $(0,1,1,2,2)^T$.
לאורך השאלה שמרו את המטריצה המתקבלת קבועה. זו תשמש אותנו להיות המטריצה X.

ג. הגרילו 500 משתנים מקריים נורמליים סטנדרטיים ב"ת (חד מימדיים), והגדירו:

$$Y_i = 2 - 3X_{i1} + 2X_{i2} + X_{i3} + 6X_{i4} - 2X_{i5} + \epsilon_i$$

נסחו את המודל בתצורה מטריציונית. הציגו את 10 השורות הראשונות של המטריצה $X \in R^{n\times(p+1)}$, את וקטור המקדמים (מה הוא $p + 1$?), ואת 10 הכניסות הראשונות של הוקטורים Y ו-ϵ. ללא שימוש בפונקציות מובנות, אלא רק בפעולות פשוטות על מטריצות, אמדו את $\hat{\beta}_{OLS}$, ושמרו את ערכי המקדמים.

ד. חזרו על הסעיף הקודם 10,000 פעמים, כך שברשותכם יהיו 10,000 אומדים ל-$\hat{\beta}_{OLS}$. חשבו את וקטור הממוצעים שלהם, ואת מטריצת השונויות האמפירית. השוו אותם לפרמטרים התיאורטיים שמתקבלים מהחישוב שביצעתם בסעיף א'. הסבירו את התוצאות.

ה. חזרו על סעיפים ג' ו-ד' פעמיים נוספות, אך כעת תחת תוך דגימה של הרעש המקרי באופנים הבאים:

1. $\epsilon_i \sim N\left(0, ||X_i||^2\right)$. כאשר הכוונה היא ש-$X_i \in R^{p+1}$ הוא השורה ה-i של המטריצה X.
2. $\epsilon_i \sim N(1,1)$.

אילו הנחות מופרות בכל אחד מהמקרים? כיצד זה השפיע על התוחלת והשונות של האומדים ביחס לסעיף הקודם? (כדי להשוות את השונות, השתמשו בקריטריון משאלה 2).

הערה: הפקודות מסעיף ב' הן ב-*R* אבל אתם יכולים להשתמש בשפת תכנות לבחירתכם, וכן בכל חומר עזר/מודל שפה ג'נרטיבי. הקפידו לצרף את הקוד <u>והסברים מילוליים משלכם</u>. הגישו קובץ *PDF* אחד הכולל את הפתרון שלכם לכל השאלות וכן את הקוד והתוצאות של השאלה האחרונה.