שאלה 3 (18 נקודות)

יהי $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, 2, 3 \dots n$
וכרגיל, ε_i מתפלג נורמלית עם תוחלת 0 ושונות σ_ε^2.

תמי רוצה לבצע רגרסיה כדי למצוא את הקשר בין x ל-y. לצערה של תמי, היא לא יכולה למדוד את ערכי ה-x באופן מדויק, ומקום זאת היא יכולה למדוד רק "גירסה רועשת" שלהם:

$$z_i = x_i + \delta_i$$

כאשר δ_i הינו משתנה מקרי שמתפלג נורמלית עם תוחלת 0 ושונות σ_δ^2.

תמי, שאיננה יכולה לדעת את ערכי x האמיתיים (וגם לא את ערכי δ), טוענת:

"מכיוון δ_i הינו משתנה מקרי חסר-הטיה בעל תוחלת 0, הרי שזה כאילו מוסיפים שגיאה חסרת-הטיה ל-ε . בהתאם, ניתן להשתמש בערכי z במקום ערכי x ולמצוא כרגיל את אומדי הריבועים הפחותים עבור המודל

$$y_i = \beta_0 + \beta_1 z_i + \tilde{\varepsilon}_i \qquad i = 1, 2, 3 \dots n$$

והאומדים שנקבל כך יהיו הערכים הרצויים (עבור מדגם גדול לפחות)"

א. עבור $\widehat{\beta_1}$, אומד הריבועים הפחותים לשיפוע, האם תמי צודקת? בפרט, מה יקרה עבור גודל מדגם גדול (אם תמי תשתמש בערכי z במקום ערכי x)?
הדרכה: סמנו את השונות של x בעזרת σ_x^2 (אליה השונות המדגמית של x מתכנסת), ניתן להעזר גם במשפט סלוצקי (ראו למטה).

ב. תנו דוגמה והסבירו מתי/איך עשויים לדעת את σ_δ^2 , או לקבל אומדן שלו. מה תמליצו לתמי לעשות במקרה כזה?

משפט סלוצקי (חלקי):
יהי $\{a_n\}$ ו- $\{b_n\}$ סדרות של משתנים מקריים.
אם a_n מתכנס בהתפלגות ל-A, ו-b_n מתכנס בהסתברות לקבוע b (אשר שונה מ-0), אזי:

$$\frac{a_n}{b_n} \longrightarrow \frac{A}{b}$$

(כאשר ההתכנסות הינה בהתפלגות).

פיתרון:

א. אומד הריבועים הפחותים לשיפוע, כאשר משתמשים ב-z במקום ב-x, הינו פשוט:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2}$$

עבור n (גודל מידגם) גדול מאוד המכנה מתכנס לשונות של z:

$$Var[z] = Var[x + \delta] = Var[x] + Var[\delta] = \sigma_x^2 + \sigma_\delta^2$$

באופן דומה, עבור n (גודל מידגם) גדול מאוד המונה מתכנס לשונות-המשותפת של z ו-y:

$$Cov[y, z] = Cov[\beta_0 + \beta_1 x + \varepsilon,\ x + \delta] = Cov[\beta_1 x, x] = \beta_1 \sigma_x^2$$

ולפי משפט סלוצקי, האומד עצמו מתכנס למנה:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})^2} \longrightarrow \frac{\beta_1 \sigma_x^2}{\sigma_x^2 + \sigma_\delta^2} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\delta^2}\ \beta_1$$

ומכיוון שהשבר $\frac{\sigma_x^2}{\sigma_x^2+\sigma_\delta^2}$ הינו קטן ממש מ-1, הרי שהאומד איננו עקיב ונקבל אומד שונה מ-β_1 האמיתי ושיפוע יותר "שטוח" (אפילו אם נגדיל את המידגם)!

ב. יש כמה מצבים שונים בהם עשויים לדעת את σ_δ^2 , או לקבל אומדן שלו.
דוגמה 1: כאשר אנו יודעים מה ההתפלגות של שגיאת המדידה; למשל כאשר מודדים משקל בעזרת מכשיר בעל רזולוציה גסה של קילוגרם (תכלס מכשיר די סטנדרטי) הרי שהמשקל האמיתי איננו המשקל המדווח **השלם**. המישקל האמיתי היננו המשקל המדווח + "רעש עיגול" בין 0.5 ק"ג ל 0.5- ק"ג. וכעת ניתן להניח שהשונות של "רעש העיגול" היא השונות של $U_c[-0.5, 0.5]$, ואת זה קל לחשב ולקבל את σ_δ^2.
דוגמה 2: לפעמים ניתן לקבל חזרות עבור ערך x_i מסוים (גם אם את y_i עצמו לא ניתן למדוד שוב). בעזרת חזרות אלו ניתן לאמוד את השונות של δ_i וכך לקבל אומדן עבור σ_δ^2.

אם תמי יודעת את σ_δ^2 (או אומדן שלו), אז היא יכולה גם לדעת (או לאמוד) σ_x^2 את כי $Var[z] = \sigma_x^2 + \sigma_\delta^2$.
כעת, כל מה שעליה לעשות זה לקחת את $\hat{\beta}_1$ שמצאה (ראו למעלה) ולהכפיל אותו בגורם התיקון ההופכי $\frac{\sigma_x^2+\sigma_\delta^2}{\sigma_x^2}$ ואז תקבל אומד עקיב.