

משימת פרויקט 4

מטרת משימה 4 היא ליישם את שיטת Bootstrap במקרים בהם קשה לאמוד את השונות וכן ליישם את שיטת מבחני הפרמוטציות לבדיקת השערות אי-פרמטריות.

ודאו כי ההסברים שלכם מלווים בתוצאות המוצגות בקובץ ההגשה, במידת הצורך ניתן להציג בשנית תוצאות ממשימות קודמות. הפעילו שיקול דעת ובחרו בעצמכם כיצד להציג את התוצאות של הסעיפים השונים.

במשימה זו, בכל השאלות על Bootstrap אין להשתמש בפקודות של פייתון לחישוב אוטומטי תוך שימוש בשיטה, והגדירו $B = 400$. בכל סעיף שבו אינכם מתבקשים במפורש לבצע Bootstrap או מבחן פרמוטציות ניתן להשתמש בכל פונקציה בפייתון.

חלק א': רגרסיה לינארית

בחלק זה נמשיך את חלקה הראשון של משימה 3. השתמשו באותה שאלת מחקר, אותם משתנים מסבירים ואותו תת-מדגם בגודל 200 שבחרתם במשימה 3 חלק ראשון.

1. חשבו רווחי סמך לכל אחד מהמקדמים $\beta_0^*, \dots, \beta_k^*$ בשיטות השונות:
 - א. רווח סמך על סמך הקירוב הנורמלי כאשר האומד ל-s.e. מבוסס על מטריצת השונות של וקטור המקדמים $\hat{\beta}$.
 - ב. רווח סמך על סמך הקירוב הנורמלי כאשר האומד ל-s.e. מבוסס על Bootstrap.
 - ג. רווח סמך על סמך Bootstrap בשיטת הפיווט.
 - ד. רווח סמך על סמך Bootstrap בשיטת האחוזונים.
 2. עבור כל אחד מהמקדמים $\beta_0^*, \dots, \beta_k^*$ השוו בין אורך רווח הסמך שהתקבל בכל אחת מהשיטות ובדקו האם רווח הסמך מכיל את המקדמים של הוקטור $\hat{\beta}$ המבוסס על המדגם של כלל הנתונים.
 3. עבור 100 נקודות מהמדגם המקורי שאינן מופיעות בתת-מדגם בגודל 200, נחשוב על כל נקודה כזו כ- (X_{new}, Y_{new}) :
 - א. חזו את ערך הנקודה בהינתן X_{new} .
 - ב. חשבו רווח סמך לתוחלת $E[Y_{new}|X_{new}]$ ברמת סמך מקורבת של 95% בעזרת שיטת Bootstrap לפי בחירתכם.
 - ג. חשבו את אחוז רווחי הסמך שמכילים את הערך האמיתי Y_{new} . האם הוא קרוב ל-95%?
 - ד. כיצד הייתם מחשבים את רווח הסמך לנקודה Y_{new} ולא לתוחלת שלה בעזרת Bootstrap? אין צורך לחשב רק לתת הסבר.
- הדרכה: שימו לב שהשונות של Y_{new} מורכבת משני גורמים, אחד התלוי ברעש ואחד התלוי באמידה. העזרו בתרגול 5 שאלה 1.

חלק ב': מבחנים

בחלק זה השתמשו בשאלת מבחן שהשתמשתם במשימת פרויקט 2: האם הערך של משתנה X שונה בין קטגוריות שונות של משתנה בינארי Y . אם יש לכם יותר משתי קטגוריות, הגבילו את עצמכם לשתי קטגוריות לטובת המשימה הבאה.

1. נסחו את שאלת המחקר כשאלה האם ההתפלגות של ערכי X בקטגוריה אחת גדולים מערכי X בקטגוריה השנייה.
2. בחרו באופן אקראי תת-מדגם בגודל 200. אלו הנתונים שאיתם נעבוד בחלק זה.
3. הגדירו μ_1, μ_2 להיות התוחלות בקטגוריה הראשונה והשנייה בהתאמה.
 - א. אמדו את μ_1, μ_2 ואת $\delta = \mu_1 - \mu_2$. האם האומד שקיבלתם הוא אומד נראות מרבית?
 - ב. מצאו רווח סמך ל- δ .
 - ג. אמדו את δ על סמך כל הנתונים. האם האומד של δ על סמך כל הנתונים נמצא ברווח הסמך של הסעיף הקודם?
 - ד. בחנו את ההשערה של שאלה 1 בעזרת מבחן וולד ובעזרת מבחן פרמוטציות. ניתן להשתמש בדגימה במבחן הפרמוטציות.
4. הגדירו m_1, m_2 להיות החציונים (שברון 0.5) בקטגוריה הראשונה והשנייה.
 - א. אמדו את m_1, m_2 ואת $\delta = m_1 - m_2$. האם האומד שקיבלתם הוא אומד נראות מרבית?
 - ב. מצאו רווח סמך ל- δ .
 - ג. אמדו את δ על סמך כל הנתונים. האם האומד של δ על סמך כל הנתונים נמצא ברווח הסמך של הסעיף הקודם?
 - ד. בחנו את ההשערה של שאלה 1 בעזרת מבחן פרמוטציות. ניתן להשתמש בדגימה.
 - ה. אם ניתן להשתמש במבחן וולד? במבחן t בסעיף ד'? נמקו.
5. הגדירו S_1 להיות סכום הדרגות בקטגוריה הראשונה. בחנו את ההשערה של שאלה 1 בעזרת מבחן פרמוטציות. האם ניתן להשתמש בקירוב נורמלי? אם כן, השתמשו בו. השוו לתוצאה בעזרת דגימה.
6. דונו בקצרה בתוצאות שאלות 3-5. האם התוצאות שקיבלתם מסתדרות עם ההתפלגויות שהצגתם בהיסטוגרמות של המשתנים?