מעבדה בסטטיסטיקה 52568 - 2020-21, מטלה 9. להגשה והצגה ב-3.1

תיאור המשימה:

המעבדה עוסקת בחיזוי שינויים בדפוסי ההצבעה בין שתי מערכות הבחירות. יש להשתמש בקבצי תוצאות הבחירות על פי קלפיות בבחירות מרץ 2020 וספטמבר 2019, וכן בקובץ הדירוג החברתי כלכלי לפי ישובים. בכל השאלות יש להתייחס רק ל-9 או 10 המפלגות הגדולות ב-2 מערכות הבחירות.

1. א. חשבו על פי הנוסחאות הרגילות את סטיית תקן ו-p-value לאברי המטריצה M שנאמדו בשיטת רגרסיה לינארית עם רבועים פחותים בשאלה 1 סעיף א' במעבדה 8, עבור השערת האפס $m_{jk}=0$, ומבחן חד צדדי עם אלטרנטיבה חיובית עבור כל איבר. דווחו על כל ערכי ה-M המובהקים עבור $\alpha = 0.001$.

   ב. חשבו סטיית תקן בשיטת ה-Bootstrap לכל איבר במטריצה M שנאמדה בשיטת הריבועים הפחותים בשאלה 1 סעיף א' במעבדה 8. השתמשו ב-100 איטרציות bootstrap. כעת, השתמשו בקירוב הנורמלי עבור ההתפלגות של כל איבר במטריצה הנאמדת M כאשר הניחו שכל איבר הוא אומד בלתי מוטה של ערך M האמיתי. על סמך קירוב זה חשבו p-value עבור בעיית בדיקת ההשערות מסעיף א. ודווחו על כל ערכי ה-M הסיגניפיקנטיים ברמת מובהקות של 0.001. השוו את תוצאותיכם לתוצאות סעיף א.

   ג. חזרו על סעיף ב. אך הפעם עבור אומדי M עם אילוצי החיוביות ונרמול ל-1 משאלה 3 במעבדה 8, כלומר אמידה באמצעות nnls, איפוס ערכים קטנים מ-0.005 ונרמול כל שורה ל-1. כמו כן, בדקו עבור כל המפלגות פרט לעבודה-גשר ומרץ שהתאחדו האם היה אבדן קולות מובהק, כלומר בדקו את השערת האפס: $m_{jj}=1$ והאלטרנטיבה $m_{jj}<1$ כשהאיבר ה-jj באלכסון מתאר מעבר קולות ממפלגה לעצמה. למה לא ניתן להשתמש בשיטה של סעיף א. עבור שאלה זו? ומה היתרון של שיטת ה-Bootstrap כאן?

   הערה: כשהאומד עבור פרמטר מסויים זהה עבור כל מדגמי ה-bootstrap, סטיית התקן הנאמדת על פי ה-bootstrap היא 0. במקרה זה כדי להמנע מבעיות נומריות נגדיר את ה-p-value להיות 0.5.

2. א. חלקו את הקלפיות המשותפות לשתי מערכות הבחירות פעם אחת באופן אקראי לשתי קבוצות זרות: train ו-test בחלוקה של 80% מהקלפיות ב-trian ו-20% ב-test. אמדו את שלושת המודלים משאלות 1 ב', 2 ו-3 במעבדה 8 על נתוני ה-train וחשבו את השגיאה הריבועית הממוצעת בחיזוי תוצאות בחירות ב על פני כל קלפיות ה-test עבור כל מודל. את השגיאה הריבועית הממוצעת יש לחשב רק עבור התצפיות המתאימות ל-9 המפלגות הגדולות בכל קלפי ב-test במערכת הבחירות של מרץ 2020 - כלומר אין להשתמש בתצפיות המתאימות ל"לא הצביעו", גם אם השתמשתם בתצפיות אלו לצורך אמידת המודל. איזו שיטה נותנת שגיאה ריבועית ממוצעת מינימלית?

   הערה: מספר התצפיות ב-test יהיה מספר קלפיות ה-test כפול 9. עבור כל תצפית כזו מקבלים שגיאת חיזוי ריבועית ויש למצע שגיאות אלו.

   ב. אמדו מודל נוסף החוזה את תוצאות בחירות מרץ 2020 באמצעות בחירות ספטמבר 2019 וכן מדד חברתי כלכלי באופן הבא: חלקו את הקלפיות ל-2 קבוצות על פי אשכולות דמוגרפיים של ישוביהן המופיעים בקובץ החברתי כלכלי ממעבדה 4. קבוצה אחת תהיה קלפיות השייכות לישובים עם מדד חברתי-כלכלי 1-5, והשניה קלפיות השייכות לישובים עם מדד חברתי כלכלי 6-10 וישובים עבורם לא מיפינו את הישוב למדד חברתי כלכלי. בכל קבוצה אמדו את M כל קבוצת ה-train של שאלה 4, לפי אומד ה-nnls של שאלה 3 מעבדה 8 כך שתקבלו 2 מטריצות M שונות.

   בשלב החיזוי השתמשו עבור כל קלפי בקבוצת ה-test במודל המתאים (כלומר במטריצה הנאמדת M המתאימה לישוב של קלפי זו). חשבו את השגיאה הריבועית הממוצעת על קלפיות ה-test והשוו אותה לשגיאות בסעיף א. בנוסף, הציגו את 2 מטריצות ה-M שאמדתם ותארו את ההבדלים הבולטים ביניהן.

3. **בונוס:** התאימו מודלים נוספים לחיזוי תוצאות בחירות 2020 מבחירות ספטמבר 2019 ושפרו את השגיאה הרבועית הממוצעת על ה-test ביחס לשאלה 2. מותר להשתמש במידע דמוגרפי/כלכלי/וכו' נוסף על כל ישוב/קלפי.

   הערות: חשבו על עיצוב הגרפים. תנו כותרת לצירים, שימו לב לאורך הצירים. השתמשו בפונטים גדולים. השתמשו בצבעים, עובי נקודה, וכו׳ כדי להדגיש נקודות חשובות.