רגרסיה ושיטות ניתוח רב- משתניות

אוניברסיטת בן-גוריון

**תרגיל מס. 2:**

**תאריך הגשה: 4.1.2023 בשעה 23:59**

**העבודה תוגש בזוגות בלבד (הגשה ביחידים באישור מיוחד)**

**אופן ההגשה: כקובץ וורד (.docx) +קובץ R, בתיבת ההגשה הייעודית.**

**היקף העבודה:** עד 10 שורות לשאלה. רווח כפול, פונט David, גודל פונט 12. אין הגבלה על מספר שורות הקוד בR.

**שימו לב!** בסעיפים אשר דורשים בניית קוד/ גרפים, עליכם להעתיק את הקוד/הגרף הרלוונטי ולהוסיף את אלו לקובץ הוורד שאתם מגישים. אין צורך להוסיף פלטים. שימו לב- אי הוספה של הקוד/ גרף הרלוונטיים תגרור הורדה בניקוד.

**שאלות נכון לא נכון**

1. חישוב של מתאם חלקי נעשה (בין היתר) בעזרת מתאם פירסון.
2. שמעון בנה מודל רגרסיה בו הוא ניבא את Y באמצעות שלושה מנבאים (Xיקסים). לאחר שבנה את מודל הרגרסיה שמעון שם לב כי כל השיפועים במודל שבנה שונים באופן מובהק מאפס, ובנוסף הוא הבחין כי הביטא של כל אחד מהמשתנים שווה למתאם פירסון בין אותו משתנה לבין Y. שמעון טען שסיבה אפשרית לכך היא שההשפעה הישירה של כל אחד מהמשתנים על Y היא אפס. שמעון צודק.
3. כאשר שווה לאפס, אין סיבה לבנות מודל רגרסיה על מנת לשפר את הניבוי של מודל הממוצע.
4. שנהב רצתה לבנות מודל רגרסיה מרובה. שנהב ביצעה תיקוף צולב ומצאה כי הR בריבוע של סט האימון גדול מהR בריבוע של סט הבחינה. לכן, שנהב הסיקה כי המודל שבנתה על סט האימון סובל מהתאמת יתר (over- fitting). שנהב צודקת.

**שאלה פתוחה**

לשאלה זו קובץ נתונים הרלוונטי לניתוחים השונים (movies\_big\_five). הקובץ נמצא במודל.

המחקר שלפניכם בדק את הקשר בין אישיות ומשתנים נוספים לבין דירוגי סרטים על ידי נבדקים. הנבדקים גויסו מתוך פלטפורמה שבה משתמשים נותנים דירוג לסרטים שהם כבר צפו בהם בעבר, ועל בסיס הדירוגים הנבדקים קיבלו רשימה המכילה 12 סרטים שהם לא צפו בהם בעבר, כך שלכל נבדק רשימה אחרת המבוססת על הדירוגים שעשה בעבר. כל נבדק צפה בכל 12 הסרטים שהופיעו ברשימה שהותאמה עבורו, ודירג כל סרט בנפרד. בנוסף, הנבדקים עברו מדידה של ה-Big Five שלהם (מדדי אישיות) ושני מדדים נוספים שבהם דיווחו טרם הצפייה עד כמה הם חושבים שיהנו מהצפייה ברשימת הסרטים שקיבלו, ואת המידה שבה חשו שהרשימה הותאמה אישית עבורם.

במחקר השתתפו 178 משתתפים, ונאספו עבורם הנתונים הבאים:

* מספר הנבדק/ת (userid)
* Big Five – ציונים אישיים מ-1 עד 7 בתכונות Openness, Agreeableness, Emotional Stability, Conscientiousness, Extraversion.
  + Openness – פתיחות מחשבתית, פתיחות לחוויה, יצירתיות וסקרנות
  + Agreeableness – נעימות, אדיבות, סימפתיה, שיתוף פעולה
  + Emotional Stability – רוגע, קור רוח
  + Conscientiousness – מצפוניות, מוכוונות, נכונות לעבודה קשה ולאחריות
  + Extraversion – נטייה להפניית אנרגיה כלפי חוץ, מעורבות חברתית, פעלתנות, אסרטיביות והרפתקנות
* התאמה אישית (Is\_personalized) – עד כמה הנבדקים חשו שרשימת הסרטים שקיבלו הותאמה עבורם אישית, כלומר, עד כמה חשו שרשימת הסרטים שנתנו להם התבססה על סרטים קודמים שהם צפו בהם ואהבו. משתנה זה נמדד ע"י מידת ההסכמה של הנבדקים מ-1 (מאוד לא מסכימ/ה) עד 5 (מאוד מסכימ/ה) עם המשפט "This list is personalized for me".
* ציפייה לגבי הנאה (Enjoy\_watching) – עד כמה הנבדקים חשבו לפני הצפייה שהם יהנו בסך הכל מהצפייה בסרטים שברשימה – מידת ההסכמה של הנבדקים מ-1 עד 5 עם המשפט "This list contains movies I think I would enjoy watching".
* הדירוג הממוצע (Avg\_rating) – הדירוג הממוצע של כל נבדק על הסרטים שראה, בטווח שבין 0.5 ל-5 כוכבים. כלומר, נבדק דירג את כל אחד מהסרטים ברשימה בנפרד (דירוג עבור סרט יחיד יכול היה להיות בטווח שבין 0.5 ל-5 כוכבים), ומשתנה זה הוא הממוצע של אותם דירוגים.

עדי טענה שמה שמשפיע על דירוג הסרטים הממוצע זו בעיקר האישיות של הנבדקים מכיוון שאנשים הגבוהים ב-Agreeableness נוטים להיות יותר חמימים ופחות ביקורתיים, ואנשים שגבוהים ב-Openness פתוחים לחוויות חדשות ואוהבים אומנות. לכן עדי שיערה כי אנשים הגבוהים ב-Agreeableness או ב-Openness יתנו דירוגים גבוהים יותר לסרטים בממוצע. לעומת זאת, שירין טענה שמה שמשפיע על דירוגי הסרטים הם לא דפוסי אישיות אלא הסיטואציה, ולכן המידה שבה נבדקים הרגישו שרשימת הסרטים הותאמה אליהם אישית (Is\_personalized) היא זו שתביא לדירוג סרטים ממוצע גבוה יותר.

1. כתבו את משוואות הרגרסיה עבור המודל של עדי ועבור המודל של שירין. הריצו את מודלי הרגרסיה המתאימים ב-R כדי לחלץ את המקדמים השונים במשוואה.
2. עבור כל חוקרת, האם ההשערות אוששו?
3. בהמשך לסעיפים א' ו-ב', חלצו ב-R את ערכי ה-b וה- של כל אחד מהמודלים וכתבו אותם. הסבירו מהו ההבדל באופן כללי בין b ל- , וכתבו באופן ספציפי מהי המשמעות של ערכי ה-b וה- שקיבלתם ביחס לממוצע דירוגי הסרטים.
4. רועי בחן את סט הנתונים ומצא שהמתאם בין Emotional stability לממוצעי דירוגי הסרטים הוא מאוד חלש. רועי החליט שבשל המתאם החלש, לא שווה להכניס את Emotional stability למודל שמטרתו להסביר את ממוצעי דירוגי הסרטים. האם רועי צודק? הסבירו מדוע באמצעות ההיגיון שעליו מתבססת הסכימה של רייט.
5. נופר סברה כי המידה שבה הנבדקים חשו שהרשימה הותאמה להם אישית (is\_personalized) וגם הדיווח של הנבדקים בנוגע לעד כמה הם ציפו לחוות הנאה מהצפייה ברשימת הסרטים (enjoy\_watching) משפיעים על דירוגי הסרטים. עם זאת, נופר עדיין חשבה שגם מאפייני אישיות אמורים להשפיע בצורה משמעותית על דירוגי הסרטים ושיש להכניס את ה-Big Five למודל. הסבירו כיצד יכולה נופר לבחון את ההשפעה של ה-Big Five מעבר למשתנים האחרים שהחליטה להכניס. בצעו את הניתוח המתאים ב-R וכתבו מהן המסקנות.
6. דביר החליט לנבא את ממוצע דירוגי הסרטים בעזרת כל רכיבי ה-Big Five ורק איתם. לאחר מכן, דביר החליט ש-conscientiousness אינה תכונה רלוונטית לניבוי ממוצע דירוגי סרטים מבחינה תיאורטית, והסיר את Conscientiousness מהמודל. כתוצאה מכך, ה-R בריבוע של המודל החדש ירד בהשוואה למודל הקודם. האם R בריבוע נמוך יותר במודל החדש של דביר בהשוואה למודל עם כל ה-Big Five יכול להוות עדות נגדית לטענתו של דביר שמודל בלי Conscientiousness הוא עדיף? הסבירו.
7. חברת הסרטים דיסמי פרסמה תחרות שבה המטרה של המשתתפים היא לבנות מודל אשר ינבא במדויק את הדירוג של כל צופה בסרט החדש שיצרו, כאשר כל צופה מדרג את הסרט בסולם של 1 עד 5 כוכבים. המודל הזוכה יהיה המודל שניבא את הדירוג המדויק הכי הרבה פעמים (למשל, ניבוי מדויק הוא כאשר נבדק מדרג את הסרט ב-3 כוכבים, המודל ינבא שהנבדק ידרג את הסרט בשלושה כוכבים). אתם החלטתם להשתתף בתחרות של דיסמי, אולם הנתונים היחידים שבידכם הם דירוגים של מאה צופים שצפו בסרט החדש של דיסמי ודירגו את הסרט בסולם המדובר (1-5 כוכבים). מהו המודל שתציעו לדיסמי? פרטו.
8. דימה הביט על המתאם בין הציפייה לגבי הנאה (Enjoy\_watching) לבין ממוצע דירוגי הסרטים של כל נבדק. דימה טען כי ייתכן שחלק מהציפייה של הנבדקים שיהנו מרשימת הסרטים קשורה גם למידה שבה הנבדקים חשו שהרשימה מותאמת אישית עבורם (Is\_personalized). בהתאם לכך, בצעו ב-R בקרה סטטיסטית כך שתוכלו לבחון את הקשר של הציפייה לגבי הנאה על ממוצע דירוגי הסרטים, בניכוי המידה שבה הנבדקים חשו שהרשימה מותאמת אישית עבורם. הוסיפו הסבר קצר על התהליך. מהי המשמעות של התוצאות שמצאתם?