

תרגיל 2

יש להיעזר בקובץ *do* (סטטה) או סקריפט (*R*) ואף צריך להגיש אותו יחד עם קובץ לוג.

1. לשאלה זו יש להשתמש בנתונים שבקובץ *BWGHT_2020.dta* בתיקית קבצי הנתונים. קובץ זה לקוח מעבודה הבודקת את ההשפעה של עישון בזמן ההיריון על משקל התינוק בעת הלידה. ניתן להיעזר בפקודה *des* (סטטה) או בפקודות *str*, *summary* (*R*) בכדי לראות את פירוט המשתנים.

א. כמה נשים יש במדגם (בסטטה ניתן להיעזר בפקודה: *count* או *des*; *R* ניתן להיעזר בפקודות *length* או *nrow*)? כמה מהן מעשנות (בסטטה ניתן להיעזר בפקודה: *tab*; ב-*R* ניתן להיעזר בפקודה *sum*)?

ב. מה הממוצע של מספר הסיגריות שמעשנת אישה ביום בזמן ההיריון? האם הממוצע הוא מדד טוב במקרה זה לאישה הטיפוסית? הסבירו.

ג. מבין הנשים המעשנות, מה הממוצע של מספר הסיגריות שמעשנת אישה ביום בזמן ההיריון? השוו בין תוצאה זו לתוצאה בסעיף הקודם.

ד. מה ממוצע מספר המשקאות השבועי ששותה האם (*drink*)? מדוע הממוצע חושב רק על 1557 תצפיות?

ה. מהו החודש שבו הכי הרבה נשים החלו טיפול (*monpre*) כמה נשים החלו טיפול אחרי החודש השישי (לא כולל החודש השישי)? (החלק הראשון של השאלה דורש שימוש בפקודות שעוד לא למדנו, אין חובה לענות עליו)

ו. הריצו רגרסיה (בסטטה על ידי הפקודה *reg*; ב-*R* על-ידי הפקודה *feols*) בין משקל התינוק למספר הסיגריות שעישנה האם בזמן ההיריון. דווחו את תוצאות הרגרסיה בצורה של משוואה.

מה משקל התינוק החזוי עבור נשים לא מעשנות ומה משקל התינוק החזוי עבור נשים המעשנות קופסה ביום (הניחו שבקופסה יש 20 סיגריות)?

ז. האם ניתן להסיק מתוצאות הרגרסיה שאימהות שמעשנות יותר נוטות להביא ילדים עם משקל לידה נמוך יותר?

האם ניתן להסיק מתוצאות הרגרסיה כי עישון מצד אימהות נוטה להוריד את משקל הלידה של ילדים?

ח. חשבו את מקדם המתאם בין משקל התינוק לבין מספר הסיגריות שעישנה האם בזמן ההיריון. מה הקשר בין מקדם המתאם שקיבלתם לבין האומדן לשיפוע ברגרסיה שאמדתם בסעיף ז' (רמז: השתמשו בנוסחה שלמדנו לאומדן ובנוסחה למקדם המתאם)?

ט. מה צריך להיות הערך של $cigs$ כדי לחזות משקל של 3536 גרם? ממה נובעת תוצאה זו? (שימו לב למספר הלא מעשנות במדגם ולמספר הנשים שילדו תינוקות במשקל נמוך יותר)

2. לשאלה זו יש להשתמש בנתונים שבקובץ `ATTEND_2020.dta` בתיקיית קבצי הנתונים. קובץ זה לקוח מעבודה הבודקת את הקשר בין אחוז ההשתתפות בשיעורים בקולג' (שם המשתנה: `atndtre`) לבין הציון במבחן הכניסה לקולג' (שם המשתנה: `ACT`).

א. מה הערך הגבוה ביותר והנמוך ביותר של `ACT` ו-`atndtre`? (בסטטה ניתן להיעזר בפקודה `sum`; ב-`R` ניתן להיעזר בפקודות `min` ו-`max`)

ב. כיצד ניתן לפרש את β_1 ברגרסיה הבאה?

$$atndtre = \beta_0 + \beta_1 ACT + u$$

האם ניתן לפרש את ההשפעה של β_1 כהשפעה סיבתית? האם ברור מראש מה הסימן של β_1 ?

ג. אמדו את המודל מסעיף ב' והציגו את התוצאות במשוואה. דווחו על ה- R^2 . מה הערך מעיד על איכות האומדים?

ד. מה הפער החזוי באחוז ההשתתפות בין שני סטודנטים שהפער בציון בניהם הוא 20? הסבירו את התוצאה.

ה. מה הטווח של הערך החזוי של שיעור ההשתתפות במשוואה זו (הערך הגבוה ביותר והנמוך ביותר)? (העזרו בפקודה `predict` ליצירת הערך החזוי).