



מרכז ארצי לבחינות ולהערכה (ע"ר)
NATIONAL INSTITUTE FOR TESTING & EVALUATION
المركز القطري للامتحانات والتقييم
מיסודן של האוניברסיטאות בישראל

אמידת תוקף ניבוי באוכלוסייה
לפני הליך קבלה ונשירה -
השוואה בין שלוש שיטות לטיפול בנתונים החסרים

דוח מחקר

RR
22-01

ינואר 2022

דביר קלפר | איתמר כהן | תמר קנת-כהן | צור קרליץ

דוח מחקר

RR-22-01

ISBN:978-965-502-224-7

אמידת תוקף ניבוי באוכלוסייה
לפני הליך קבלה ונשירה -
השוואה בין שלוש שיטות לטיפול בנתונים החסרים

דביר קלפר
איתמר כהן
תמר קנת-כהן
צור קרליץ

ינואר 2022

תוכן עניינים

1.	תקציר.....	5
2.	מבוא.....	6
2.1	רקע.....	6
2.2	מנגנונים של נתונים חסרים.....	7
2.2.1	חוסר באופן אקראי לחלוטין.....	7
2.2.2	חוסר באופן אקראי.....	8
2.2.3	חוסר שלא באופן אקראי.....	8
2.2.4	הזיקה בין המנגנונים.....	9
2.3	המחקר הנוכחי.....	9
2.3.1	שיעורי קבלה ושיעורי נשירה באוניברסיטאות בישראל.....	11
2.3.2	תרומת המחקר לספרות על טיפול בנתונים חסרים באמידת תוקף ניבוי.....	12
2.3.3	תרומת המחקר לאמידת תוקף ניבוי בהקשרים יישומיים.....	13
3.	שיטה.....	14
3.1	ניתוח עיקרי – נכונות האמידה (Correctness) כפונקציה של מקור החוסרים.....	14
3.1.1	משתני המחקר.....	14
3.1.2	מערך המחקר.....	15
3.1.3	עיבוד הנתונים.....	17
3.2	ניתוח משני 1 – העמקה: נכונות האמידה כפונקציה של המתאמים עם משתנה העזר.....	18
3.3	ניתוח משני 2 – דיוק האמידה (Precision).....	19
4.	תוצאות.....	20
4.1	ניתוח עיקרי – נכונות האמידה כפונקציה של מקור החוסרים.....	20
4.1.1	נשירה אקראית.....	20
4.1.2	נשירה ודאית של החלשים.....	22
4.1.3	נשירה מתואמת קריטריון.....	23
4.2	ניתוח משני 1 – העמקה: נכונות האמידה כפונקציה של המתאמים עם משתנה העזר.....	25
4.3	ניתוח משני 2 – דיוק האמידה.....	27
5.	סיכום ודיון.....	30
5.1	השוואה בין שלוש השיטות (ללא משתני עזר).....	30
5.2	השוואה בין שלוש השיטות (עם משתני עזר).....	31
5.3	מאפייני משתנה העזר ותרומתם לאמידה.....	32
5.4	דיוק האמידה.....	32
5.5	המלצות מעשיות.....	33
5.6	חשיבות המחקר ומגבלותיו.....	33
5.7	מחקרי המשך.....	34
6.	מקורות.....	35

39.....	7. נספחים
39.....	7.1. נספח 1: השלמות מרובות
42.....	7.2. נספח 2: מנגנוני MNAR ומודל הברירה של הקמן
42.....	7.2.1. מנגנוני MNAR
43.....	7.2.2. מודל הברירה של הקמן
47.....	7.3. נספח 3: קוד R עבור תוכנית המחשב

לוח איורים

- איור 1: דיאגרמת ון עבור הקשרים בין מנגנוני החוסרים 9
- איור 2: מתאם בין X ל- Y כפונקציה של החלק היחסי של הדחייה מתוך החוסרים – נשירה אקראית 21
- איור 3: מתאם בין X ל- Y כפונקציה של החלק היחסי של הדחייה מתוך החוסרים – נשירה ודאית של החלשים 22
- איור 4: מתאם בין X ל- Y כפונקציה של החלק היחסי של הדחייה מתוך החוסרים – נשירה מתואמת קריטריון 24
- איור 5: תוקף נאמד כפונקציה של המתאמים של משתנה העזר עם החזאי והקריטריון עבור שיטת ההשלמות המרובות 25
- איור 6: תוקף נאמד כפונקציה של המתאמים של משתנה העזר עם החזאי והקריטריון עבור מודל הברירה של הקמן 26
- איור 7: טעות התקן האמפירית של אומדן התוקף במקרה של נשירה אקראית 28
- איור 8: גרפי קופסה עבור המתאם הנאמד עם משתנה עזר "יכולת התמדה" ומתאם 0.3 בין החזאי לקריטריון במקרה של נשירה אקראית 29
- איור 9: שלבים 1-6 בתוכנית המחשב 49
- איור 10: שלבים 7-8 בתוכנית המחשב עבור המקרה שאין משתנה עזר 50
- איור 11: שלבים 7-8 בתוכנית המחשב עבור המקרה שיש משתנה עזר 51

1. תקציר

במחקרי תוקף ניבוי משתמשים לרוב בנוסחת התיקון לקיצוץ תחום על מנת לנטרל את ההשפעה של הליך הקבלה על המתאם בין החזאי לקריטריון. נוסחה זו מתאימה למצב שבו קיימת ברירה של תצפיות לפי החזאי, כאשר הנמוכים בחזאי אינם מתקבלים ללימודים ולפיכך אין להם ציון בקריטריון. במקרה שקיימת גם נשירה, כלומר שציוני קריטריון חסרים שלא בגלל ציון בחזאי, הנחות הנוסחה לתיקון לקיצוץ תחום לא בהכרח מתקיימות.

מחקר זה בדק בעזרת סימולציות את איכות אמידת המתאם בין החזאי לקריטריון במצב שבו בנוסף להליך הקבלה קיימת גם נשירה. המחקר השווה בין שלוש שיטות לטיפול בנתונים החסרים: נוסחת התיקון לקיצוץ תחום, שיטת ההשלמות המרובות, שמיועדת לטפל במצב של נתונים שחסרים באופן אקראי (MAR), ומודל הברירה של הקמן, שמיועד לטפל במצב של נתונים שחסרים שלא באופן אקראי (MNAR). המחקר יישם שלוש דרכים למידול נשירה: נשירה אקראית, נשירה ודאית של החלשים בקריטריון, ומצב ביניים של "נשירה מתואמת קריטריון" (מתאם שלילי, אך לא מושלם, בין נשירה לקריטריון). איכות האמידה נבדקה תוך שליטה בפרמטרים כגון גודל האוכלוסייה, המתאם בין החזאי לקריטריון באוכלוסייה, שיעור הנדחים ושיעור הנושרים. בנוסף להצלבות המלאות של כל הפרמטרים לעיל, נבדקה איכות האמידה כאשר יש משתנה עזר – משתנה שמתואם עם הקריטריון או עם הנטייה שיהיה נתון חסר בקריטריון. הגדרנו שתי חלופות למשתנה העזר: משתנה מסוג "יכולת התמדה" שמתואם יותר עם הקריטריון מאשר עם החזאי, ומשתנה מסוג "מהירות" שמתואם יותר עם החזאי מאשר עם הקריטריון.

התוצאות מראות שמעבר לתנאים השונים, התיקון לקיצוץ תחום הינה השיטה העדיפה, במובן שבמרבית התנאים היא מדייקת, וכשלא – היא מספקת אומדן שמרני של המתאם. שיטה זו מדייקת פחות ככל שיש יותר נשירה שאינה אקראית. שיטת ההשלמות המרובות נותנת תוצאות דומות לאלו של התיקון לקיצוץ תחום, ובנסיבות מסוימות (בפרט, כשמשמשים במשתנה עזר מסוג "יכולת התמדה") – היא מדייקת יותר, אך במחיר של מורכבות חישובית. מודל הברירה של הקמן עדיף רק במקרים של אוכלוסייה גדולה, עם נשירה ודאית של החלשים ועם אחוז גבוה יחסית של נשירה לעומת דחייה, אם כי השימוש בו עשוי להיות כרוך באומדן יתר של המתאם.

המחקר כלל גם שני ניתוחים משניים. בראשון, בדקנו את התלות של איכות האמידה במשתנה העזר באופן יותר שיטתי ומקיף. בבדיקה זו נמצא, שכצפוי, המתאם של משתנה העזר עם הקריטריון משפיע על איכות האמידה יותר מאשר המתאם שלו עם החזאי. בניתוח המשני השני, השווינו את הפיזור באומדני שתי שיטות – תיקון לקיצוץ תחום והשלמות מרובות – במצב שבו שתי השיטות אומדות את המתאם באופן כמעט מושלם (נשירה אקראית). בדיקת הפיזור מעבר לסימולציות השונות גילתה יתרון לשיטת ההשלמות המרובות עם משתנה עזר מסוג "יכולת התמדה".

2. מבוא

2.1. רקע

בבדיקת תוקף הניבוי של מבחן שמשמש לברירת מועמדים ללימודים אקדמיים מחשבים מתאם פירסון בין הציון במבחן (לדוגמה, הציון הפסיכומטרי) ובין הציון בקריטריון – המשתנה שהמבחן מיועד לנבא (לדוגמה, ציון הגמר לתואר בוגר). מאחר שמועמדים שהציון הפסיכומטרי שלהם נמוך מסף הקבלה לא מתקבלים ללימודים, אין להם ציון גמר, ולכן הם לא נכללים בחישוב המתאם. הבעיה היא שהמתאם המחושב – המתאם בקרב המתקבלים – נמוך מהמתאם שמעניין אותנו – המתאם בקרב כל המועמדים. כדי להמחיש את האפקט של הערכת החסר, נניח, למשל, שהציון הפסיכומטרי וציון הגמר מתפלגים נורמלית ומתואמים במתאם של 0.50 בקרב המועמדים. נניח גם כי ללימודים מתקבלים אך ורק המועמדים שהציון הפסיכומטרי שלהם גבוה מהמוצע בסטיית תקן אחת. ניתן להראות שהמתאם הנצפה (דהיינו, ברמת המתקבלים ללימודים בלבד) שבין הציון הפסיכומטרי לציון הגמר יהיה כ-0.24 בלבד. כדי לאמוד את המתאם שהיה מתקבל אילו היו כל המועמדים מתקבלים, וכך להעריך את תוקף הניבוי האמיתי של הציון הפסיכומטרי, משתמשים בתיקון סטטיסטי – "תיקון לקיצוץ תחום" (Pearson, 1903). תיקון כזה מומלץ על ידי הסטנדרטים למבחנים חינוכיים ופסיכולוגיים (AERA et al., 2014) ומהווה נוהל שגרתי במחקרי תוקף ניבוי במאל"ו (למשל, קנת-כהן, ברנרן ואורן, 2021) ובעולם (ראו, למשל, Westrick et al., 2019 עבור ה-SAT, ו-Westrick, Le, Robbins, Radunzel, & Schmidt, 2015 עבור ה-ACT). התיקון מבוסס על הנחות התפלגותיות מסוימות שתוארו ונבדקו במחקרים רבים (ראו סקירה אצל Sackett & Yang, 2000) ועל ההנחה שהברירה נבעה רק מהליך הקבלה (שבמקרה הפשוט, שבו נעסוק כאן, התבסס על החזאי שאת התוקף שלו אנחנו בודקים). במילים אחרות, ההנחה היא שערך חסר בציון הקריטריון נובע אך ורק מאי קבלה, כלומר, מכך שהציון בחזאי נמוך מסף הקבלה.

בפועל, הנסיבות שבהן נערכים מחקרי תוקף ניבוי כוללות אפשרות שלחלק מן התצפיות חסר קריטריון לא רק בגלל שלא התקבלו ללימודים, אלא גם בגלל שנשרו במהלך הלימודים, בין אם בגלל כישלון צפוי בקריטריון ובין אם מסיבות שאינן קשורות לקריטריון. במצב כזה הנחת נוסחת התיקון לקיצוץ תחום על כך שהברירה מושפעת רק מהחזאי אינה מתקיימת. המחקר הנוכחי יתמקד בהשלכות של מצבי עולם מגוונים מבחינת הסיבה לנתונים החסרים, ועל התפקוד של הנוסחה לתיקון לקיצוץ תחום ושל שתי שיטות כלליות לטיפול בנתונים חסרים, במצבי העולם האלה. להלן יסקרו מצבי העולם השונים מבחינת הסיבה לחסרות הנתונים, והשיטות לטיפול בנתונים החסרים במצבים האלה.

2.2. מנגנונים של נתונים חסרים

בסיס הנתונים שעומד לרשותנו במחקר תוקף ניבוי טיפוסו הוא אוכלוסיית המועמדים, שבה לכולם יש ציון במבחן ששימש בהליך הקבלה (החזאי, X), אך לא לכולם יש ציון בקריטריון (Y). ערך חסר ב- Y יכול לנבוע מכך שהמועמד לא התקבל (דחייה), או מכך שהתקבל, אך לא השלים את הלימודים (נשירה). את הנשירה עצמה אפשר לסווג לשני מקורות: נשירה "מסיבות אקדמיות", שהיא נשירה שקשורה לביצוע בקריטריון, כלומר, חוסר הצלחה בלימודים שמוביל לנשירה, וכן נשירה "אקראית", שהיא נשירה שאינה קשורה ישירות לקריטריון ומונעת מגורמים אחרים (אישיותיים, בריאותיים, כלכליים, חברתיים, וכדומה).

שלוש הסיבות שמנינו לעיל לנתונים חסרים ב- Y (דחייה ושני סוגים של נשירה) מתארות, למעשה, את מה שמוכר בספרות כמנגנונים של נתונים חסרים (Rubin, 1976). רובין תיאר שלושה מנגנונים של ערכים חסרים שמתארים כיצד ההסתברות לנתון חסר במשתנה כלשהו קשורה למשתני המחקר, אם בכלל. המנגנונים הללו, שיתוארו בקצרה להלן, מתפקדים למעשה כהנחות עבור השיטות השונות לטיפול בחוסרים בנתונים.

2.2.1. חוסר באופן אקראי לחלוטין – (MCAR) Missing Completely At Random

נניח שיש לנו נתונים חסרים במשתנה מסוים Y (הכנסה, לדוגמה). הנתונים יחשבו לחסרים באופן אקראי לחלוטין (MCAR) אם ההסתברות לנתון חסר במשתנה Y אינה קשורה לערכים של Y מקבל או לערכים של משתנים אחרים בבסיס הנתונים. מנגנון ה-MCAR מופר, למשל, אם אנשים שסירבו לתת מידע על ההכנסה שלהם הם (בממוצע) צעירים יותר מאנשים אשר הסכימו לתת מידע על ההכנסה.

המנגנון של MCAR הוא המנגנון היחיד מבין שלושת המנגנונים של נתונים חסרים שניתן לבדוק אם הוא מתקיים. בדוגמה שלמעלה, ניתן לבדוק אם החוסר במשתנה הכנסה הוא מסוג MCAR על ידי חלוקת המדגם לשתי קבוצות: כאלו שהסכימו לתת מידע על ההכנסה שלהם וכאלו שלא, והשוואת ממוצעי הגילאים בין הקבוצות באמצעות מבחן t . הבדל מובהק בין הקבוצות מהווה ראיה לכך שהנתונים אינם MCAR. בדיקה דומה אפשר לעשות על כל המשתנים האחרים בבסיס הנתונים. הוצעה גם הרחבה רב-משתנית למבחני ה- t החד-משתניים (Little, 1988).

עבור חוסרים מסוג MCAR רוב השיטות ה"מסורתיות" לטיפול בנתונים חסרים, כגון: מחיקת שורות מלאות (Listwise Deletion) או מחיקה בזוגות (Pairwise Deletion), מתפקדות היטב.

2.2.2. חוסר באופן אקראי – (MAR) Missing At Random

ההנחה של חוסר באופן אקראי (MAR) היא הנחה חלשה מזו של MCAR. נתונים חסרים במשתנה Y ייחשבו כחסרים באופן אקראי אם ההסתברות לנתון חסר קשורה למשתנה אחר (או למשתנים אחרים), אך לא לערכים של Y עצמו. במילים אחרות, אין קשר בין הנטייה לערך חסר ב- Y ובין הערכים של Y אחרי שמפקחים על השונות של המשתנים האחרים (כלומר, כשמחזיקים את המשתנים האחרים קבועים). כך, למשל, הנחת ה-MAR תתקיים אם ההסתברות של נתונים חסרים בהכנסה (Y) תלויה במשתנה מצב משפחתי (X), אבל בתוך אותו מצב משפחתי ההסתברות לחוסרים אינה תלויה בהכנסה עצמה.

כפי שצינו בסוף הסעיף הקודם שעסק ב-MCAR, אי אפשר לבחון את קיום ההנחה של MAR. הסיבה לכך מובנת מאליה: ההשוואה הנדרשת לבדיקת הנחת ה-MAR היא בין ערכי Y בקרב בעלי ערכים ב- Y לערכי Y בקרב חסרי ערכים ב- Y (תוך פיקוח על X). ומאחר שערכי Y של חסרי ערכים ב- Y אינם ידועים, לא ניתן לעשות השוואה כזאת.

ישנן שתי שיטות חדשות יחסית שמטפלות בנתונים עם חוסרים מסוג MAR: שיטת השלמות מרובות (Multiple Imputation – MI; Rubin, 1976) ושיטת מקסום התוחלת (Expectation Maximization – EM; Dempster, Laird, & Rubin, 1977).

2.2.3. חוסר שלא באופן אקראי – (MNAR) Missing Not At Random

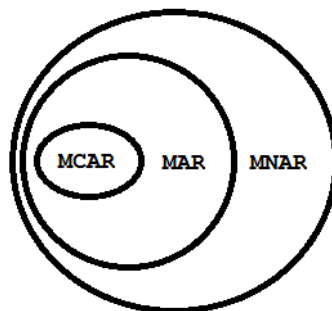
ההנחה של חוסר שלא באופן אקראי (MNAR) היא הנחה חלשה מזו של MAR. נתונים חסרים במשתנה Y ייחשבו לחסרים שלא באופן אקראי אם ההסתברות לנתון חסר קשורה לערכים של Y עצמו, אפילו אחרי שמפקחים על השונות של המשתנים האחרים (כלומר, כשמחזיקים את המשתנים האחרים קבועים). כך, למשל, אם ההסתברות לנתונים חסרים בהכנסה (Y) תלויה בהכנסה עצמה (אנשים עם הכנסה גבוהה נמנעים מלענות, או להיפך) יתקיים מצב של MNAR. כמו במנגנון של MAR, אין דרך לוודא שהחוסרים הם MNAR מבלי לדעת את הערכים החסרים.

במקרה זה, שיטת הטיפול בנתונים החסרים מבוססת על כך שתהליך האמידה כולל גם מרכיב שמתאר (ממדל) את הנטייה לחסרות. שתי שיטות שמטפלות בנתונים עם חוסרים מסוג MNAR הן: מודל הברירה של הקמן (the selection model; Heckman, 1976) ומודל תערובת הדפוסים (the pattern mixture model; Glynn, Laird, & Rubin, 1976).

2.2.4. הזיקה בין המנגנונים

מצב של חוסרים באופן אקראי לחלוטין (MCAR) הוא המצב הספציפי ביותר, כלומר עם ההנחות החזקות ביותר על מנגנון החוסרים. קצת יותר כללי (הנחות יותר מקלות) הוא המצב של חוסרים באופן אקראי (MAR), והמצב של חוסרים שלא באופן אקראי הוא המצב הכללי ביותר. ניתן להמחיש את הזיקה בין המנגנונים של חוסרים בדיאגרמת ון הבאה:

איור 1: דיאגרמת ון עבור הקשרים בין מנגנוני החוסרים



בהתאם לכך, שיטה שמותאמת לחוסרים של MAR (כגון השלמות מרובות) תעבוד טוב גם עבור חוסרים של MCAR (בהיותה שיטה שדורשת פחות תנאים מראש ולכן כללית יותר), אבל לא תהיה בהכרח טובה מספיק עבור חוסרים שנוצרו עקב MNAR (בהיותה שיטה שמניחה חוסרי MAR, ובהתאם לכך הינה ספציפית לחוסרים מסוג כזה, ולכן היא לא מספיק טובה עבור המקרה הכללי יותר של חוסרי MNAR). בהקשר זה חשוב להדגיש שלמרות שחוסרי MNAR הוא המקרה הכללי ביותר, כדי ששיטות שמתאימות לסוג כזה של חוסרים תעבוד הן צריכות להניח עוד הנחות, ולכן לא מובטח מראש שהן תהיינה יותר טובות משיטות שמותאמות לחוסרים מסוג MAR.

2.3. המחקר הנוכחי

המחקר הנוכחי בודק, באמצעות הדמיה, את איכות האמידה של המתאם בין חזאי לקריטריון באוכלוסייה, כאשר המדגם שעומד לרשותנו הוא תוצר של ברירה משלושה סוגים: דחייה על בסיס החזאי במסגרת הליך הקבלה, נשירה אקראית ונשירה של החלשים בקריטריון. שלושת סוגי הברירה הללו משקפים שלושה מקורות לנתונים חסרים בקריטריון שתואמים לשלושת המנגנונים של נתונים חסרים שתוארו לעיל:

- דחייה בעקבות הליך הקבלה, שמתבסס על החזאי, משקף מנגנון של MAR
- נשירה אקראית משקפת מנגנון של MCAR
- נשירה של החלשים (נשירה ודאית של החלשים או נשירה שמתואמת עם הקריטריון, אך לא באופן מושלם) משקפת מנגנון של MNAR

המונח "ברירה" במחקר זה יתייחס, אם כן, לאפקט הכולל של המקורות השונים לנתונים חסרים: **ברירה = דחייה בעקבות הליך קבלה + נשירה (מסוגים שונים).**

במחקר זה נשווה בין שלוש שיטות שונות¹ לאמידת תוקף הניבוי:

1. נוסחת התיקון לקיצוץ תחום (במקרה זה – קיצוץ תחום ישיר);
2. השלמות מרובות – אחת מהשיטות שמיועדות לטיפול בנתונים חסרים שהם MAR (ראו נספח 1 לתיאור השיטה);
3. מודל הברירה של הקמן – אחת מהשיטות שמיועדות לטיפול בנתונים חסרים שהם MNAR (ראו נספח 2 לתיאור השיטה);

שתי השיטות שמיועדות לטיפול בנתונים חסרים – השלמות מרובות ומודל הברירה של הקמן – יושמו כל אחת בשתי חלופות: עם משתנה עזר ובלעדיו. משתנה העזר הוא משתנה שמתואם עם הקריטריון עצמו או עם הנטייה שיהיה נתון חסר בקריטריון, והוספתו לניתוח מיועדת לשפר את אמידת המתאם בין החזאי לקריטריון. משתנה העזר יכול להיות משתנה דמוגרפי (כמו גיל, למשל) או מדד של תכונת אישיות (כמו יכולת התמדה, למשל), או כל משתנה אחר. חשוב לציין שניתן לכלול יותר ממשתנה עזר אחד בניתוח. לכן משתנה העזר שמודל במחקר זה כמשתנה יחיד יכול לשקף אוסף של משתנים אחדים.

תפקודן של שלוש שיטות האמידה (שמספקות, כפי שהוסבר לעיל, חמישה אומדנים שונים של תוקף הניבוי) נבדק תחת מגוון פרמטרים משתנים, שיתוארו בהמשך בפרק השיטה. שניים מן הפרמטרים – שיעור קבלה ושיעור נשירה – נמצאו ראויים לבדיקה מקדימה, כדי לוודא שייבחרו ערכים מציאותיים להדמיה. בהינתן החידוש של המחקר הנוכחי, של התחשבות בשני מרכיבי הברירה – הליך קבלה (עם תשובת דחייה בעקבותיו) ונשירה – נבדקו השיעורים הטיפוסיים של שני המרכיבים הללו – שיעורי קבלה ושיעורי נשירה – בלימודי התואר הראשון בישראל, כדי שההדמיה שעשינו תשקף מצב עולם רלוונטי למחקרי התוקף שלנו. הממצאים אודות השיעורים הללו יתוארו להלן.

¹ כאמור בחרנו שיטה אחת מכל מנגנון חסרים, כלומר שתי השיטות מקסום התוחלת ותערובת הדפוסים לא יושמו אצלנו במחקר.

2.3.1. שיעורי קבלה ושיעורי נשירה באוניברסיטאות בישראל

שיעורי הקבלה (Acceptance Rates) המדווחים על ידי הלשכה המרכזית לסטטיסטיקה (הלשכה המרכזית לסטטיסטיקה, 2019) מספקים מידע על שיעור המתקבלים ולומדים מתוך המועמדים ועל שיעור כל המתקבלים (מתקבלים ולומדים ומתקבלים שאינם לומדים) מתוך המועמדים. כך, למשל, על פי נתוני המועמדים ב-2019/20 ברפואה שיעור המתקבלים ולומדים נע בטווח של 9%-19% באוניברסיטאות השונות בעוד ששיעור כל המתקבלים נע בטווח של 9%-49%; במשפטים שיעור המתקבלים ולומדים נע בטווח של 33%-62% באוניברסיטאות השונות בעוד ששיעור כל המתקבלים נע בטווח של 51%-86%; בהנדסה ואדריכלות שיעור המתקבלים ולומדים נע בטווח של 42%-56% באוניברסיטאות השונות בעוד ששיעור כל המתקבלים נע בטווח של 45%-79%; ובמדעי הרוח שיעור המתקבלים ולומדים נע בטווח של 57%-65% באוניברסיטאות השונות בעוד ששיעור כל המתקבלים נע בטווח של 70%-92%.

שני המדדים האלה נותנים תמונה שונה למדי בנוגע לשיעורי הקבלה, ובפרט נראה שערך נמוך במדד הראשון מבין השניים ("יחס מועמדים למתקבל ולומד"), שמופיע בפרסומים שונים (הלשכה המרכזית לסטטיסטיקה, 2019; המועצה להשכלה גבוהה, ל.ת.), עשוי להיתפס באופן מוטעה כמרמז על ביקוש גבוה לעומת ההיצע או, לחילופין, על מידת הסלקטיביות של תחום הלימוד, בעוד שערך כזה יכול להיגרם, כשכל יתר הדברים קבועים, דווקא משיעור גבוה של מתקבלים שבחרים שלא ללמוד בתחום הלימוד ובמוסד הנדון. לא ברור איך נכון להתייחס למועמדותם של המתקבלים שבחרים שלא ללמוד. ניתן לראות בהם מועמדים שהחליטו בשלב כלשהו להסיר את מועמדותם – החלטה שלא תועדה מסיבות טכניות/אדמיניסטרטיביות, כך שלצורך חישוב שיעורי קבלה היה נכון אולי שלא לכלול אותם, לא במועמדים ולא במתקבלים. אך מאחר שאין בידינו מידע מקביל לגבי הנדחים בחרנו שלא להפעיל חישוב כזה. כך או כך, אם מתייחסים למדד הפשוט של שיעור כל המתקבלים מתוך המועמדים עולה תמונה של טווח רחב של שיעורי קבלה. על מנת לכסות טווח זה, בחרנו לבדוק במחקר שיעורי קבלה שנעים בין 0% ל-40% (או שיעורי דחייה שבין 0% ל-60%).

אשר לשיעורי נשירה, המונח "נשירה" נחקר תחת הגדרות מגוונות שמשקפות נקודות מבט שונות (Seidman, 2005). אחת הדרכים המקובלות להתייחס לנשירה היא לבחון את שיעור מסיימי הלימודים לתואר ראשון בפרק זמן מוגדר מתחילת הלימודים. נתוני סטודנטים שהחלו ללמוד בשנת 2011/12 מראים ש-86.1% מן הסטודנטים באוניברסיטאות מסיימים את לימודיהם עד שלוש שנים מהזמן התקני (הלשכה המרכזית לסטטיסטיקה, 2020). גם בעניין זה יש הבדל בין תחומי לימוד. כך, למשל, בעוד ששיעור מסיימי התואר באמנויות ובמדעי הרוח הוא נמוך יחסית (69.6% באמנות, אומנויות ואמנות שימושית, 76.8% במדעי הרוח הכלליים, ו-78.0% בשפות, ספרויות ולימודים רגיונליים), במשפטים וברפואה הוא עומד על 97.2% ו-96.3%, בהתאמה. כדי לכסות את הפיזור

שנמצא בשיעור מסיימי הלימודים לתואר בזמן הנקוב, בחרנו לבדוק במחקר שיעורי "התמדה" שבין 100% ל-40% (או שיעורי נשירה שבין 0% ל-60%).

ממצא עקבי שעולה ממחקרים שמתעדים שיעורי נשירה מלימודים בהשכלה הגבוהה – הן בישראל והן בארה"ב – הוא שיותר ממחצית מן הסטודנטים שנושרים מלימודים עושים זאת לפני שנת הלימודים השנייה (קנת-כהן, 2012). להבדיל מהנתונים על שיעור מסיימי הלימודים לתואר, שמדווח בשנתונים הסטטיסטיים באופן שוטף, המידע על שיעור הממשיכים לשנה השנייה, מצוי רק במחקרים חד-פעמיים שמתפרסמים מעת לעת. כך, למשל, מבין הסטודנטים באוניברסיטאות שהחלו ללמוד בשנת 2015/16, 7.4% לא המשיכו לשנה השנייה (ליפלסקי, 2018). במחקר אחר, שהתמקד בסטודנטים לתואר ראשון באוניברסיטה העברית בירושלים, נמצא ש-82.1% סיימו את הלימודים לתואר, ו-17.9% נשרו: 10.2% לא המשיכו לשנה שנייה, 3.2% נוספים לא המשיכו לשנה שלישית, ו-4.4% נוספים נשרו בהמשך (Yair, Rotem, & Shustak, 2020). העובדה שהיקף לא מבוטל מן הנשירה מתרחש לפני שנת הלימודים השנייה היא משמעותית כיוון שפירושה שגם מחקרי תוקף ניבוי שבהם הקריטריון להצלחה בלימודים הוא ציון שנה א' עשויים להיות מושפעים מנשירה (וזאת כיוון שסביר להניח שחלק מן הנשירה שמתרחשת לפני שנה ב', קורית במהלך לימודי שנה א').

2.3.2. תרומת המחקר לספרות על טיפול בנתונים חסרים באמידת תוקף ניבוי

כאמור, למיטב ידיעתנו לא קיימים מחקרים בספרות המדעית שבודקים באופן שיטתי את איכות האמידה של תוקף הניבוי תוך הבאה בחשבון של שני מרכיבי הברירה – הליך הקבלה ונשירה, ובכך המחקר הנוכחי מהווה חידוש. יש ספרות רבה שעוסקת באמידת תוקף הניבוי כאשר הברירה מורכבת רק מהמרכיב המוקדם מבין השניים – הליך הקבלה, כשרוב רובו של המחקר המתודולוגי בתחום עוסק בנוסחאות התיקון לקיצוץ תחום. מאז שזוהו (Pearson, 1903) ההשלכות של הליך הקבלה על המתאם המחושב בין החזאי לקריטריון בקרב המתקבלים, נעשו עבודות רבות בנושא. לצד מחקר אמפירי נרחב שיישם את התיקונים, פותחו במהלך השנים נוסחאות תיקון עבור הליכי קבלה מגוונים (Lawley, 1943) ונעשו מחקרים רבים על הנחות התיקון (ראו סקירה אצל Sackett & Yang, 2000). רוב העבודות המתודולוגיות בנושא נעשו עוד לפני שנות ה-80 של המאה הקודמת, והשימוש בשיטת התיקון לקיצוץ תחום הינו נרחב ומבוסס היטב. לעומת זאת, השיטות הכלליות לטיפול בנתונים חסרים, פותחו ונכנסו לשימוש רק החל משנות ה-70 של המאה הקודמת (Enders, 2010), והאפשרות של יישומן בתחום של תיקוף כלים המשמשים לברירה היא חדשה יחסית.

מספר מצומצם של עבודות מוקדמות יחסית יישמו את מודל הברירה של הקמן בנתונים אמפיריים. כך, למשל, במחקר של לין והסטינגס (Linn & Hastings, 1984) נבדק תוקף הניבוי של הציון

במבחן הקבלה לפקולטה למשפטים (LSAT) ושל ממוצע ציוני התואר הראשון בחיזוי הצלחה בלימודי משפטים, ובשני מחקרים שנערכו בישראל נבדק תוקף הניבוי של הציון הפסיכומטרי וממוצע הבגרות בחיזוי הצלחה בלימודי שנה א' לתואר בוגר באוניברסיטת תל אביב (איילון ויוגב, 1997; קנת-כהן וברונר, 1997). העבודה האחרונה (קנת-כהן וברונר, 1997) כללה גם יישום של המודל בתנאי הדמיה, במטרה להמחיש כיצד יישום שגוי של המודל (איילון ויוגב, 1997) יכול להביא לאומדנים לא נכונים של תוקף הניבוי. יישומים אלה של מודל הברירה נולדו מתוך ההבנה שבאופן טיפוסי תהליך הברירה מושפע מגורמים נוספים מעבר להליך הקבלה – כגון נשירה – ששיטת הטיפול המקובלת בחוסרים (נוסחאות התיקון לקיצוץ תחום) אינה מביאה בחשבון. בנוסף, מביאות העבודות שנימנו לעיל התייחסות נרחבת לחולשות מודל הברירה, אך, כאמור, לא נבדקה בהם באופן שיטתי איכות האמידה שלו. סדרה נרחבת יותר של עבודות חדשות יחסית עסקה ביישום של שיטות כלליות לטיפול בנתונים חסרים בתנאים של חוסרים מסוג MAR. וויברג וסנסטרום (Wiberg & Sunstrom, 2009) הישוו בין נוסחת התיקון לקיצוץ תחום ובין שיטת מקסום התוחלת בנתונים אמפיריים ייחודיים (תיקוף מבחן תאוריה בלימודי נהיגה) שאפשרו בדיקה של איכות האמידה בגישות השונות. פפפל ואחרים (Pfaffel, Kollmayer, Schober, Spiel, 2016; Pfaffel, Schober, Spiel, 2016) הישוו בין נוסחת התיקון לקיצוץ תחום ובין שיטת ההשלמות המרובות באמצעות הדמיות במגוון של תנאים (מתאם בין החזאי לקריטריון במועמדים, שיעור קבלה) הליכי קבלה (קיצוץ תחום ישיר או עקיף) וסוגי קריטריון (רציף או דיכוטומי). מחקר המשך (Pfaffel & Spiel, 2016) התמקד באיכות האמידה של שיטת ההשלמות המרובות במדגמים קטנים עד בינוניים. ולבסוף, צימרמן ואחרים (Zimmermann, Klusmann & Hampe, 2017) הישוו בין נוסחאות התיקון לקיצוץ תחום (ישיר ועקיף) ובין שתי השיטות לטיפול בנתונים חסרים בתנאי MAR – מיקסום התוחלת והשלמות מרובות. עבודתם כללה הן הדמיה והן יישום אמפירי של השיטות (תיקוף מבחן קבלה לבית ספר לרפואה). ככלל, הממצאים בהקשר של חוסרים מסוג MAR מעידים על תפקוד דומה (במונחי נכונות, correctness, של ממוצע האומדנים מעבר לחזרות) של נוסחת התיקון לקיצוץ תחום ושל השיטות הכלליות לטיפול בחוסרים מסוג זה. כן הסתמן, בנסיבות מסוימות, יתרון מבחינת פיזור האומדנים (דיוק, precision) לשיטות הכלליות לטיפול בנתונים חסרים.

2.3.3. תרומת המחקר לאמידת תוקף ניבוי בהקשרים יישומיים

המחקר הנוכחי מאמץ את הגישה השיטתית לבדיקת איכות האמידה של השיטות השונות, שיושמה במחקרים שתוארו לעיל בנתוני MAR, ומיישם אותה בנסיבות מציאותיות יותר של ברירה שמושפעת גם מנשירה. איכות האמידה תיבדק באמצעות הדמיה תחת מגוון של תנאים שיפורטו להלן בפרק השיטה. מעבר לתרומתו התאורטית של המחקר, ממצאיו עשויים להיות רלוונטיים למחקרי תוקף ניבוי בישראל ובעולם. אמנם הקריטריון הרווח במחקרי תוקף ניבוי הוא ציון שנה א' לתואר בוגר

(Zwick, 2019), ומטבע הדברים, הפוטנציאל לנשירה גדול יותר ככל שנתוני הקריטריון נאספים בשלב מאוחר יותר בלימודים (כלומר, כשהקריטריון הוא ציון הגמר לתואר בוגר לעומת ציון שנה א' לתואר בוגר), אך עדיין כפי שעולה מנתוני הנשירה שנסקרו לעיל, קיימת נשירה לא מבוטלת גם במחקרים שעושים שימוש בציון שנה א' כקריטריון. רגישותן של השיטות השונות לאמידת תוקף הניבוי, כשמצב העולם כולל גם נשירה מסוגים שונים, היא תשומה חשובה להסקה על תוקף הניבוי של המדדים המשמשים בהליך הקבלה.

3. שיטה

3.1. ניתוח עיקרי – נכונות האמידה (Correctness) כפונקציה של מקור החוסרים

מחקר זה מבוסס על סימולציות המדמות מצבים שונים שמאפיינים מחקרי תוקף ניבוי. לשם כך, יצרנו הצלבות מלאות בין הפרמטרים השונים, שחלקם מתארים מתאמים בין משתני המחקר (באוכלוסייה לפני ברירה), וחלקם משקפים גורמים שמשפיעים על מקור והיקף החוסרים בנתונים.

3.1.1. משתני המחקר

הוגדרו שלושה משתנים ברמת כל תצפית:

- חזאי (למשל, ציון פסיכומטרי) – X
- קריטריון (למשל, ציון גמר לתואר בוגר) – Y
- משתנה עזר – Z

כאמור, משתנה העזר הוא משתנה שמתואם עם הקריטריון עצמו או עם הנטייה שיהיה נתון חסר בקריטריון, והוספתו לניתוח מיועדת לשפר את אמידת המתאם בין החזאי לקריטריון. משתנה העזר אינו רלוונטי לנוסחה לתיקון לקיצוץ תחום, אלא רק לשתי השיטות הכלליות לטיפול בנתונים חסרים – השלמות מרובות ומודל הברירה של הקמן. לכן, כל אחת משתי השיטות האלה תיושם פעם אחת ללא משתנה העזר ופעם אחת עם משתנה העזר (כך שבסך הכל, עבור כל מצב עולם יהיו לנו 5 אומדנים שונים לתוקף). כאמור, משתנה זה אמור לעזור לנו למדל את החלק של הנשירה בנתונים החסרים (ובפרט של נשירה הקשורה לערכים של Y עצמו), ולכן ברור שהמתאם של משתנה העזר עם משתנה ה- Y הוא המרכיב החשוב במודל. מאחר ש- X נכלל במידול בכל מקרה, התרומה האפקטיבית של Z לניבוי Y ,

היא התרומה לניבוי מעבר ל- X . לכן יש להביא בחשבון גם את המתאם בין Z ל- X . לפיכך, מודלו המתאמים של Z עם Y ועם X , וזאת בהתאם לאופציות שנפרט בהמשך.

3.1.2. מערך המחקר

הפרמטרים שבהם שלטנו במחקר הנוכחי הינם:

- גודל האוכלוסייה לפני ברירה (N)
 - קטן – 80
 - בינוני – 300
- מתאם של החזאי עם הקריטריון באוכלוסייה לפני הליך קבלה ונשירה (ρ_{XY}^2)
 - בינוני – 0.3
 - גבוה – 0.6
- שיעור החוסרים ומקורם (דחייה ו/או נשירה)
 - באופן טבעי ניתן למדל את שיעור החוסרים על ידי שני משתנים: שיעור הדחייה (בעקבות הליך הקבלה) ושיעור הנשירה, כאשר שיעור הדחייה הוא אחוז הלא מתקבלים מבין המועמדים ושיעור נשירה הוא אחוז הנושרים מבין המתקבלים. בחרנו למדל את שיעור החוסרים באופן שיאפשר הצגה רציפה של האפקט הכולל שלהם. לשם כך הגדרנו שתי רמות של אחוז נתונים חסרים (%Miss):
 - נמוך: 10% מהנתונים חסרים
 - גבוה: 60% מהנתונים חסרים
 - בנוסף, הגדרנו את החלק היחסי של הנתונים החסרים הנובע מדחייה בתהליך הקבלה, כמשתנה רציף (pmr=proportion missing by rejection) שערכיו נעים בין 0 ל-1 בקפיצות של 0.1 (סה"כ 11 ערכים). לפיכך, כאשר pmr=0, כל הנתונים החסרים נובעים מנשירה, וכאשר pmr=1 כל הנתונים החסרים נובעים מדחייה בהליך הקבלה.

² היות שזהו פרמטר חשוב (הערך שאליו מושווים האומדנים המתקבלים), דגמנו פרמטר זה ברמת דיוק של 0.005 (כלומר, בדיוק של 2 ספרות אחרי הנקודה, שהיא הרמה שבה מקובל להציג מתאמים). כלומר, כאשר ייצרנו נתונים עבור תנאי כלשהו במחקר, השתמשנו רק בסטים של נתונים בהם ערך המתאם שהתקבל היה בטווח של לכל היותר 0.005 מעל או מתחת לערך הרצוי.

ההיגיון שעומד בבסיס ההמשגה של שיעור הדחייה ושיעור הנשירה על ידי שני המשתנים %Miss ו-pmr הוא שבאופן טיפוסי במסגרות סלקטיביות, כלומר כאלה שמאופיינות בשיעור דחייה גבוה (כמו רפואה) יש פחות נשירה. ולהיפך.

- סוג נשירה

- נשירה אקראית
- נשירה ודאית של החלשים בקריטריון
- נשירה מתואמת קריטריון (מצב אמצע בין נשירה אקראית לנשירה ודאית של החלשים)

לצורך מידול הנשירה הוגדר משתנה Y_p (P מייצג persistence, התמדה) שהוא משתנה סמוי שמבטא את הנטייה להתמיד (שלא לנשור). למשתנה זה הוגדרו המתאמים הבאים עם משתנה הקריטריון (ρ_{Y_pY}) , בהתאמה לשלושת סוגי הנשירה:

- $\rho_{Y_pY} = 0$ – מתאים לנשירה אקראית
- $\rho_{Y_pY} = 1$ – מתאים לנשירה ודאית של החלשים בקריטריון
- $\rho_{Y_pY} = 0.7$ – מתאים למצב אמצע בין נשירה אקראית לוודאית³

חשוב לציין שעבור המקרים של נשירה אקראית ונשירה ודאית של החלשים אין באמת צורך במשתנה הסמוי Y_p . אפשר היה למדלם ללא שימוש בו. המשתנה הסמוי נחוץ עבור מקרה האמצע, אולם השתמשנו בו גם עבור המקרים האחרים על מנת שמידול הנשירה יעשה באופן אחיד בכל תנאי המחקר.

- מתאמים של החזאי והקריטריון עם משתנה העזר (באוקלוסייה לפני ברירה) (ρ_{XZ}, ρ_{YZ}) אשר ימודלו יחד באופן הבא⁴:

- משתנה עזר "יכולת התמדה" – $\rho_{XZ} = 0.3, \rho_{YZ} = 0.6$
זהו משתנה עזר שמתואם יותר עם הקריטריון מאשר עם החזאי, כדוגמת "יכולת התמדה" שאמורה להשפיע יותר על הישגים מצטברים לאורך זמן, כמו בלימודים לתואר, מאשר על הצלחה במטלה חד-פעמית כמו הבחינה הפסיכומטרית.

- משתנה עזר "מהירות" – $\rho_{XZ} = 0.6, \rho_{YZ} = 0.3$

³ מצב אמצע מתקבל עם מתאם של 0.7 (ולא 0.5), כיוון שהאמצע מוגדר ביחס לאחוז השונות המוסברת, כלומר לריבוע המתאם ($0.7^2 = 0.49 \sim 0.5$). למעשה השתמשנו בערך המדויק (0.707) שריבועו הוא 0.5.
⁴ מסיבות פרקטיות של זמני הרצת האלגוריתם, מתאמים אלו נדגמו בדיוק של 0.02. כלומר, ביצירת המדגמים, השתמשנו רק באלו שערך המתאם בהם היה לכל היותר 0.02 מעל או מתחת לערך הרצוי.

זהו משתנה עזר שמתואם יותר עם החזאי מאשר עם הקריטריון, כדוגמת "מהירות" שאמורה להשפיע יותר על ביצועים במבחנים סטנדרטיים כמו הבחינה הפסיכומטרית, מאשר על מבחנים ועבודות בלימודים לתואר.

3.1.3. עיבוד הנתונים

את הסימולציה ביצענו בהתאם לשלבים הבאים:

1. דגמנו N שלשות X, Y, Z מהתפלגות הרב נורמלית הבאה:

$$(X, Y, Z) \sim MVN \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{XY} & \rho_{XZ} \\ \rho_{XY} & 1 & \rho_{YZ} \\ \rho_{XZ} & \rho_{YZ} & 1 \end{pmatrix} \right)$$

2. זרקנו דגימות שלא עמדו ברמות הדיוק שקבענו, כלומר זרקנו דגימה אם לא התקיים אחד משלושת התנאים הבאים:

$$|\rho_{XY} - r_{XY}| < 0.005, |\rho_{XZ} - r_{XZ}| < 0.02, |\rho_{YZ} - r_{YZ}| < 0.02$$

כאשר r (עם אינדקס מתאים) מסמל את המתאם שהתקבל בפועל במדגם הספציפי.

3. חושב משתנה סמוי Y_p לפי הנוסחה:

$$Y_p = \rho_{Y_p Y} Y + \left(\sqrt{1 - \rho_{Y_p Y}^2} \right) \cdot \mathbb{N}$$

כאשר $\mathbb{N}$ הינו משתנה בעל התפלגות נורמלית סטנדרטית והינו בלתי תלוי ב- Y . נוסחה זו מבטיחה שהמתאם של Y_p עם Y הוא $\rho_{Y_p Y}$, ו- Y_p הינו גם נורמלי עם ממוצע 0 ושונות 1.

4. ביצענו "הליך קבלה" על ידי כך שמחקנו $pmr * N * \frac{\%miss}{100}$ מהנמוכים ב- Y לפי הנמוכים

במשתנה ה- X .

5. ביצענו "נשירה" על ידי כך שמחקנו מהנמוכים ב- Y (בקרב המתקבלים, כלומר אלו שלא נדחו

בשלב הקודם) $(1 - pmr) * N * \frac{\%miss}{100}$ לפי משתנה Y_p .

6. חישבנו את המתאם הנצפה על ידי חישוב מתאם פירסון בין משתנה X לבין Y (עבור

התצפיות שלא נמחקו בשלב 4 או 5).

7. חישבנו את המתאם המתוקן לתיקון לקיצוץ תחום בהתאם לנוסחה הבאה (Gulliksen,)

(1987):

$$\widehat{\rho_{XY}} = \frac{r_{XY} \left(\frac{S_X}{S_x} \right)}{\sqrt{1 + r_{XY}^2 \cdot \left(\left(\frac{S_X}{S_x} \right)^2 - 1 \right)}}$$

כאשר S_X הינו סטיית התקן של X בקרב המועמדים (כלומר מבוסס על N תצפיות), ו- S_x הינו

סטיית התקן של X בקרב המסיימים (כלומר מבוסס על $N * (1 - \frac{\%miss}{100})$ תצפיות).

8. ביצענו השלמות מרובות בשתי גרסאות: ללא משתנה העזר Z ועם משתנה העזר Z. כל גרסה

כללה 10 השלמות שונות, כאשר בכל השלמה שכזו חושב המתאם על בסיס הנתונים המלא

לאחר ההשלמות. אומדן המתאם בשיטה זו הינו ממוצע המתאמים (לאחר טרנספורמציות

פישר) מעבר ל-10 ההשלמות (ראו פירוט בנספח 1).

9. יישמנו את מודל הברירה של הקמן בשתי גרסאות: ללא משתנה העזר Z ועם משתנה העזר

Z. האמידה נעשתה בשלב אחד (ראו פירוט בנספח 2).

המחקר כלל בסך הכל 528 תנאים שונים שהתקבלו מהצלבת כל הפרמטרים (2^* גודל קבוצת המועמדים, 2^* מתאם בין החזאי לקריטריון, 2^* שיעור חוסרים, 11^* pmr, 3^* סוג נשירה ו- 2^* מתאמים עם משתנה העזר). יצרנו 5,000 סטים של נתוני הדמיה עבור כל אחד מהתנאים. התוצאות יוצגו בגרף של המתאם הנאמד כפונקציה של pmr עבור כל אחד מצירופי הפרמטרים האחרים. המתאם שמחושב ומוצג עבור כל תנאי, הינו ממוצע אריתמטי פשוט מעבר ל-5,000 הסטים. הסימולציות השונות ובעיקר הפעלת שתי השיטות לטיפול בנתונים חסרים הורצו בתוכנת R (ראו בנספח 3 את הקוד המלא שהופעל ביישום שתי השיטות הנ"ל – ובעיקר החלק של השלמות מרובות ומודל הברירה של הקמן).

3.2. ניתוח משני 1 – העמקה: נכונות האמידה כפונקציה של המתאמים עם משתנה העזר

הניתוח המרכזי התמקד באיכות האמידה של המתאמים כפונקציה של הפרמטרים השונים, ובעיקר – כפונקציה של החלק היחסי של הנתונים החסרים שנובע מדחייה בעקבות הליך הקבלה או מנשירה (pmr הוא המשתנה הבלתי תלוי ששינינו את ערכיו בצורה רציפה). את השפעת המתאמים של משתנה העזר עם החזאי והקריטריון בדקנו רק תחת שני תנאים (שקראנו להם "יכולת התמדה" ו- "מהירות"). בנוסף לשני התנאים הללו בדקנו עוד שני מצבים: $\rho_{xz} = 0.3$, $\rho_{yz} = 0.3$ ו-

$\rho_{xz} = 0.1, \rho_{yz} = 0.6$. תוצאות המקרה הראשון יצאו זהות למקרה של משתנה מסוג "מהירות", ותוצאות המקרה השני יצאו זהות למקרה של " יכולת התמדה". לכאורה משתמעת מההשוואות האלה המסקנה שרק המתאם של משתנה העזר עם הקריטריון משפיעה על התוצאות, ושאינן חשיבות למתאם של משתנה העזר עם החזאי. כדי לבדוק שאלה זו באופן שיטתי, בחרנו צירוף אחד של תנאים אשר עבורו שינינו את זוג המתאמים הללו בצורה רציפה ובלתי תלויה. הצירוף שנבחר הוא המקרה שבו גודל האוכלוסייה לפני ברירה הוא $N=300$, המתאם של החזאי עם הקריטריון הוא $\rho_{xy} = 0.6$, שיעור החוסרים הוא $\text{Miss}=60\%$, החלק של חוסרים שנובע מדחייה הוא $\text{pmr}=0.1$, וקיימת נשירה ודאית של החלשים בקריטריון (התנאי הזה נבחר כיוון שבו הפער באומדנים בין שני התנאים של המתאמים של משתנה העזר הוא הגדול ביותר).

בתנאי שנבחר שינינו באופן רציף ובלתי תלוי את שני המתאמים של משתנה העזר: המתאם של Z עם X והמתאם של Z עם Y . בחרנו למדל את שני המתאמים הללו כמשתנים רציפים, ודגמנו אותם בקפיצות של 0.1 בין הערכים 0.1-0.7. תוצאות החלק הזה יוצגו כגרפים תלת-ממדיים – המתאם הנאמד כפונקציה של ρ_{xz} ו- ρ_{yz} – גרף אחד עבור כל אחת משתי השיטות שמשמשות במשתנה העזר: השלמות מרובות ומודל הברירה של הקמן.

גם בחלק זה חזרנו על כל תנאי 5,000 פעמים. נציין שהפעם הדגימה של המשתנים הבלתי תלויים ρ_{xz} ו- ρ_{yz} נעשתה בדיוק של 0.01 (בניגוד לניתוח המרכזי שבו היא נעשתה ברמת דיוק של 0.02), וזאת כיוון שהפעם משתנים אלה מוצגים על הצירים ולכן חשוב להשיג דיוק גדול יותר.

3.3. ניתוח משני 2 – דיוק האמידה (Precision)

בפרק התוצאות ניתן לראות שעבור המקרה של נשירה אקראית ישנן שתי שיטות (השלמות מרובות והתיקון לקיצוץ תחום) שאומדות את המתאם באופן כמעט מושלם, כלומר: בנכונות מלאה. בנסיבות כאלה מעניין לבדוק אם יש הבדל בין השיטות בדיוק האמידה, כלומר: בפיזור של האומדנים. מודל הברירה של הקמן נותן במקרה כזה אומדנים פחות טובים, לכן אין טעם לבדוק את פיזור האומדנים עבורו – אם אין נכונות באמידה, סוגיית הדיוק באמידה היא פחות קריטית. לצורך זה בדקנו את ההתפלגות של התוצאות, כלומר את הפיזור של האומדנים מעבר ל-5,000 החזרות של הסימולציות. התוצאות של חלק זה יוצגו בשני אופנים: כגרפים במתכונת הגרפים שהוצגו עבור הניתוח המרכזי רק שהפעם ציר ה- y הינו סטיית התקן של האומדן, וכגרף קופסה המציג את הפיזור המלא של האומדנים.

4. תוצאות

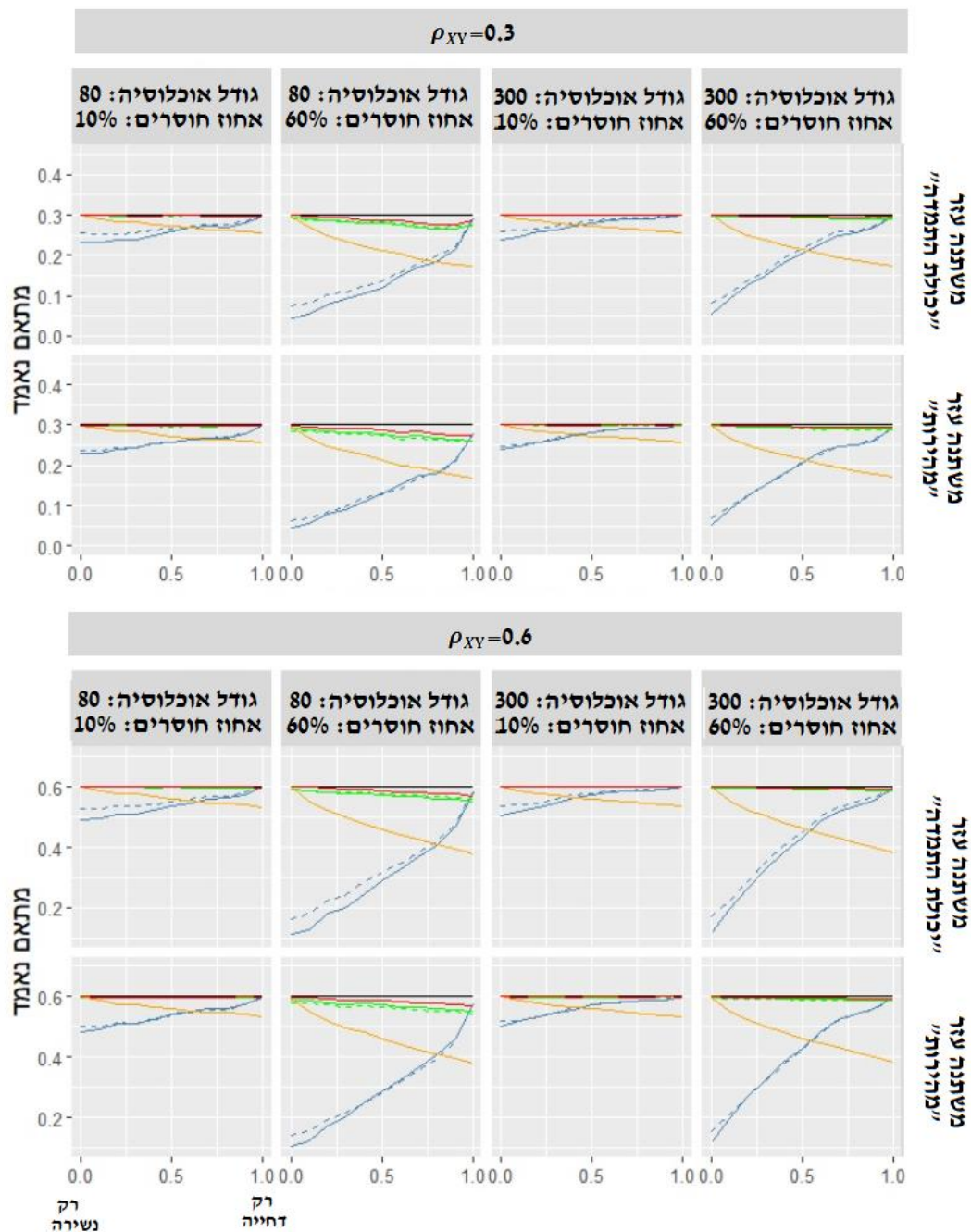
4.1. ניתוח עיקרי – נכונות האמידה כפונקציה של מקור החוסרים

בחלק זה נציג את תוצאות הסימולציות השונות שערכנו. לצורך התצוגה נחלק את התוצאות לשלושה חלקים על פי סוג הנשירה. כל חלק מכיל שני סטים של גרפים, בהתאם למתאם של החזאי עם הקריטריון, כאשר בתוך כל סט השורה הראשונה תהיה עבור משתנה עזר מסוג "יכולת התמדה" והשנייה תהיה עבור משתנה עזר מסוג "מהירות". בשתי העמודות הראשונות (משמאל) מוצגות תוצאות עבור גודל אוכלוסייה 80, בשתי העמודות הימניות מוצגות תוצאות עבור גודל אוכלוסייה של 300, ובתוך העמודות ישנה חלוקה לפי היקף החוסרים בנתונים. בכל אחד מהגרפים מוצגות 6 עקומות: המתאם הנצפה (כתום), אומדן המתאם עפ"י תיקון לקיצוץ תחום (אדום), 2 אומדנים של המתאם עפ"י השלמות מרובות (ירוק רציף עבור המקרה ללא משתנה עזר וירוק מקווקו עבור המקרה עם משתנה עזר), ו-2 אומדנים של המתאם עפ"י מודל הברירה של הקמן (כחול רציף עבור המקרה ללא משתנה עזר וכחול מקווקו עבור המקרה עם משתנה עזר). כל עקומה מציגה את המתאם הרלוונטי כפונקציה של מקור החוסרים (החלק היחסי של הנתונים החסרים שנובע מדחייה).

4.1.1. נשירה אקראית

באיור 2 מוצגות התוצאות עבור המצב של נשירה אקראית. ראשית, ניתן להיווכח שנשירה אקראית כשלעצמה לא משפיעה על המתאם, ולכן כאשר כל החוסרים מקורם בנשירה (כלומר, בצד שמאל של הציר האופקי, כש- $pmr=0$) המתאם הנצפה הוא כמו המתאם באוכלוסייה. לעומת זאת, דחייה (ברירה על פי X) כן משפיעה על המתאם, והוא אכן קטן כאשר pmr גדל. התיקון לקיצוץ תחום ושתי השיטות של השלמות מרובות מצליחים בתנאי נשירה זה לתקן את המתאם כמעט עד לערכו האמיתי, כאשר המקרה היחיד שהתיקון הינו קצת פחות טוב הוא המקרה של אוכלוסייה קטנה (80) ואחוז חוסרים גבוה (60%). לעומתם, מודל הברירה של הקמן לא אומד טוב את המתאם ומדייק רק במצב שבו כל החוסרים מקורם בדחייה. הסבר לכך שהקמן לא עובד הוא שבמקרה זה של נשירה אקראית כול החוסרים בנתונים הם במנגנון MAR (ועבור המקרה של נשירה בלבד הם אפילו במנגנון MCAR), והמודל של הקמן לא מיועד למצבים כאלה. נשים לב שעבור משתנה עזר "יכולת התמדה", המודל של הקמן עם משתנה העזר עובד קצת יותר טוב מאשר המצב ללא משתנה העזר. עוד נציין שפרט לערך המתאם, אין ממש הבדל בהתנהגות הגרפים בין $\rho_{XY} = 0.3$ לבין $\rho_{XY} = 0.6$ (מסקנה זו מתקיימת גם ביחס לסוגי הנשירה האחרים, שיוצגו להלן).

איור 2: מתאם בין X ל- Y כפונקציה של החלק היחסי של הדחייה מתוך החוסרים – נשירה אקראית



החלק היחסי של החוסרים הנובע מהליך הקבלה = pmr

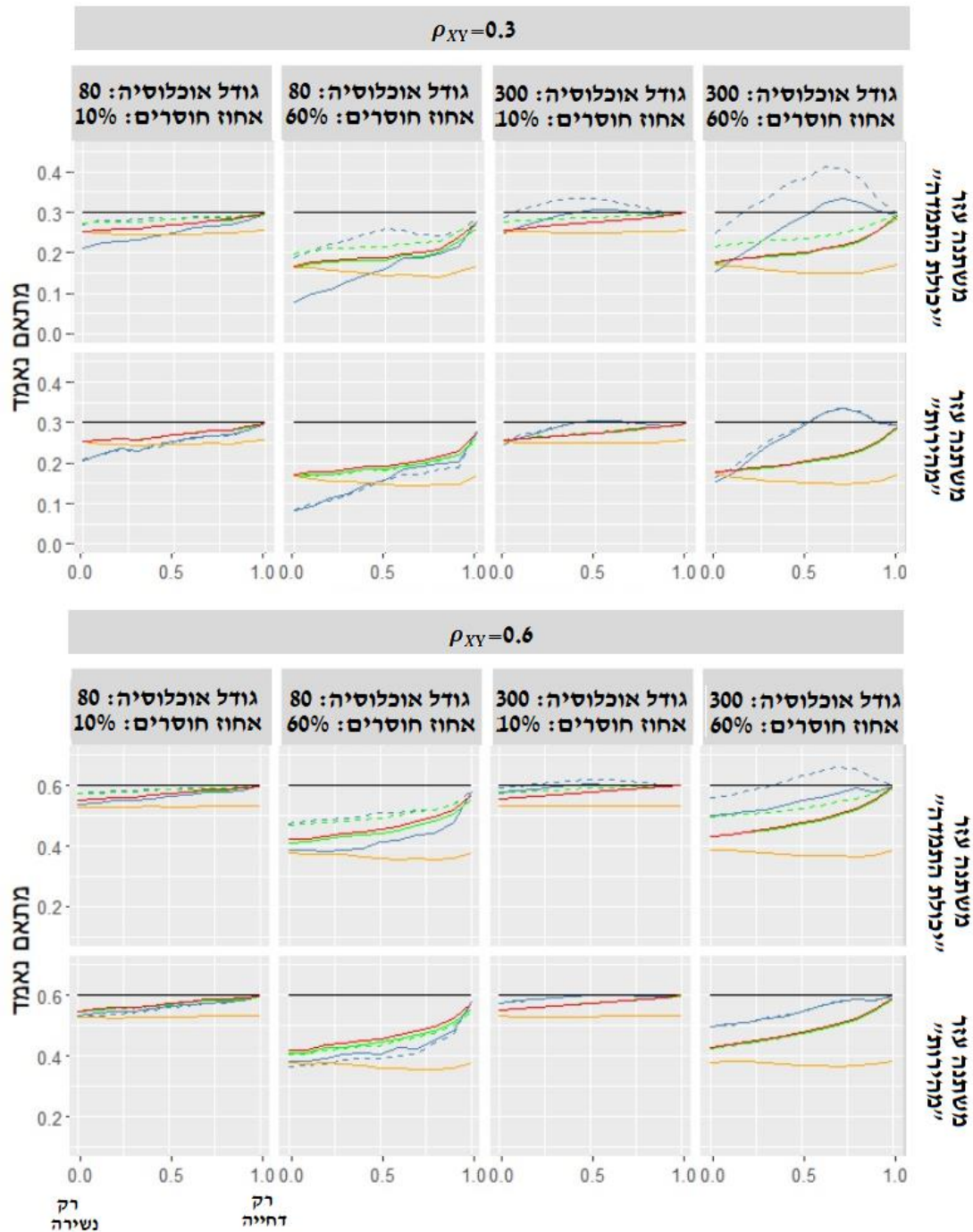
מקרא

- תיקון לקיצוץ תחום
- השלמות מרובות ללא משתנה עזר
- הקמן ללא משתנה עזר
- השלמות מרובות עם משתנה עזר
- מתאם נצפה לאחר הליך קבלה ונשירה
- הקמן עם משתנה עזר

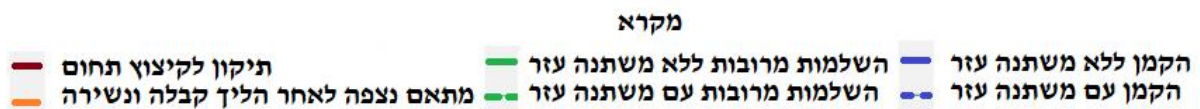
4.1.2. נשירה ודאית של החלשים

באיור 3 מוצגות תוצאות הסימולציה עבור המצב של נשירה ודאית של החלשים בקריטריון.

איור 3: מתאם בין X ל-Y כפונקציה של החלק היחסי של הדחייה מתוך החוסרים – נשירה ודאית של החלשים



החלק היחסי של החוסרים הנובע מהליך הקבלה = pmr



במקרה זה העקומה הכתומה שמייצגת, כאמור, את המתאם הנצפה היא אופקית למדי, כלומר שבמצב שבו הנשירה של החלשים בקריטריון היא ודאית, אזי מקור החוסרים (דחייה או נשירה) אינו משפיע עליו. עבור אוכלוסייה בינונית (300), שני היישומים של המודל של הקמן נותנים במרבית המצבים מתאמים גבוהים יותר מאלה של השיטות האחרות, כאשר בחלק מהמקרים הם אף נמצאים במצב של אומדן יתר (למשל, במצב של $\rho_{XY} = 0.3$, עם 60% חוסרים, משתנה עזר מסוג "יכולת התמדה" ו-pmr מעל 0.25). עבור אוכלוסייה קטנה התמונה יותר מורכבת: המתאם עפ"י מודל הקמן ללא משתנה עזר הוא תמיד הכי נמוך ולכן הכי פחות טוב⁵, ולעומתו המתאם עפ"י מודל הקמן עם משתנה עזר מסוג "יכולת התמדה" הוא הכי גבוה (אם כי אינו אומדן יתר), ולכן הוא הכי טוב. המתאם עפ"י מודל הקמן עם משתנה עזר מסוג "מהירות" דומה למתאם ללא משתנה עזר. אשר להשלמות מרובות, כאשר אין משתנה עזר ובמקרה של משתנה עזר מסוג "מהירות" האומדנים זהים לאומדנים שמתקבלים מתיקון לקיצוץ תחום. לעומת זאת, כאשר משתנה העזר הינו מסוג "יכולת התמדה", שיטת ההשלמות המרובות נותנת אומדן טוב יותר מזה שמתקבל בתיקון לקיצוץ תחום.

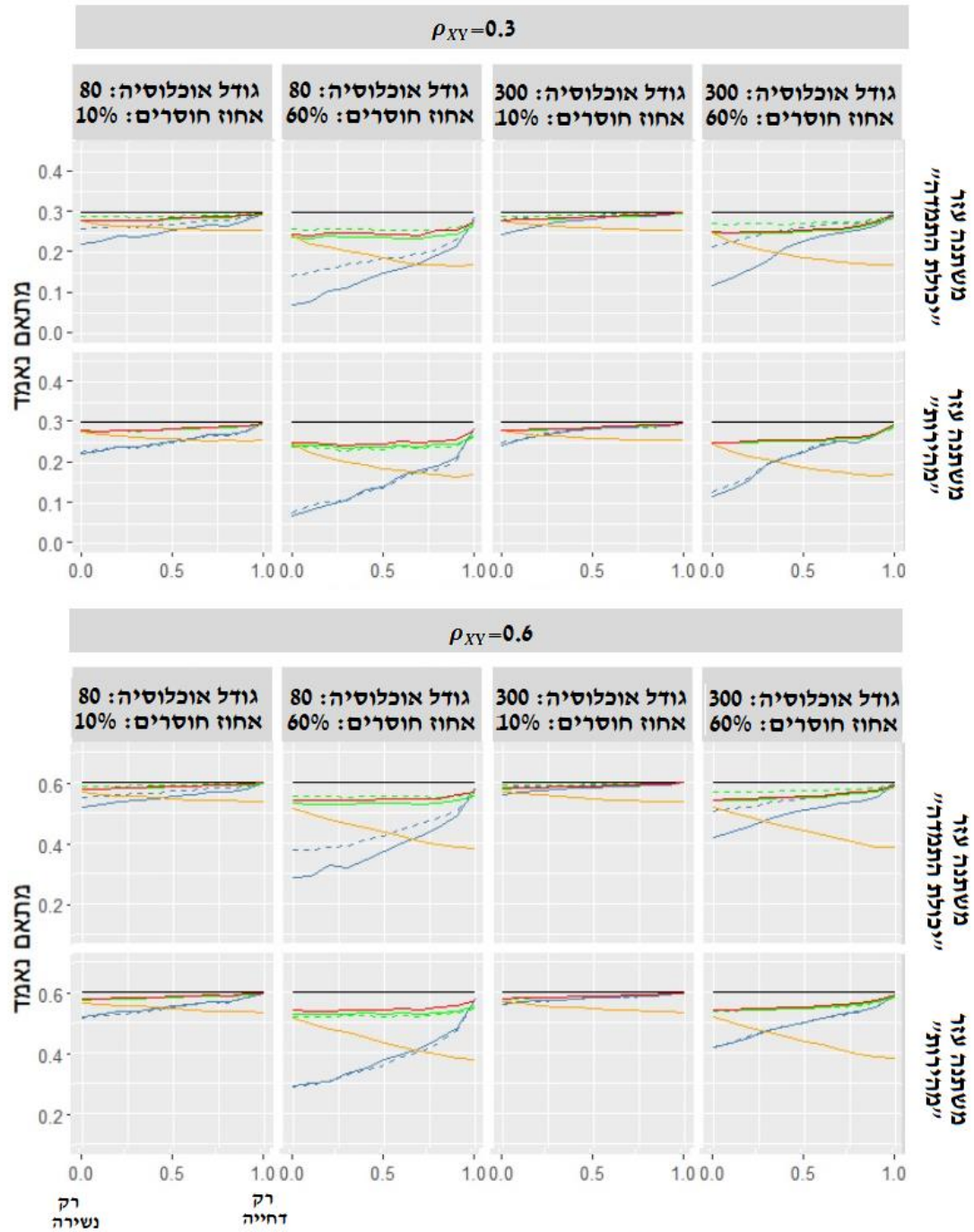
4.1.3. נשירה מתואמת קריטריון

באיור 4 מוצגות תוצאות הסימולציה עבור המצב של נשירה מתואמת קריטריון שאמור להיות מצב האמצע בין נשירה אקראית לבין נשירה ודאית של החלשים בקריטריון.

ניתן לראות שהתמונה המתקבלת היא של מצב ביניים, שדומה יותר למצב של נשירה אקראית מאשר למצב של נשירה ודאית של החלשים. נציין, עם זאת, שני הבדלים בין נשירה מתואמת קריטריון לנשירה אקראית: ההבדל הראשון הוא שבעוד שבמצב של נשירה אקראית, שיטת ההשלמות המרובות (בלי ועם משתנה עזר) והתיקון לקיצוץ תחום תפקדו באופן דומה (ודי מושלם), בנשירה מתואמת קריטריון, השיטות מתפקדות באופן שונה, כאשר היישום של השלמות מרובות עם משתנה עזר "יכולת התמדה" עוזר (במעט), ובכל מקרה, פרט למצב של אוכלוסייה קטנה (80) עם חוסרים גדולים (60%) הוא לא מזיק. ההבדל השני נוגע למודל של הקמן שבנשירה מתואמת קריטריון מתפקד טוב יותר מאשר בנשירה אקראית (אבל עדיין פחות טוב מהשיטות האחרות), כאשר משתנה העזר לא מזיק ובמקרה של משתנה עזר מסוג "יכולת התמדה" הוא אפילו עוזר.

⁵ זה קורה אפילו במצב של נשירה בלבד (pmr=0) למרות שמצב של נשירה ודאית של החלשים הוא בדיוק המצב שבו המידול של הקמן אמור לעבוד הכי טוב (כי זה מצב של חוסרים מסוג MNAR טהור). אחת הסיבות לכך כנראה היא שמודל הקמן מסתמך מאוד על הנחות הנורמליות ועבור אוכלוסייה קטנה זהו בדיוק המצב בהן הנחות אלו מופרות.

איור 4: מתאם בין X ל- Y כפונקציה של החלק היחסי של הדחייה מתוך החוסרים – נשירה מתואמת קריטריון



החלק היחסי של החוסרים הנובע מהליך הקבלה = pmr

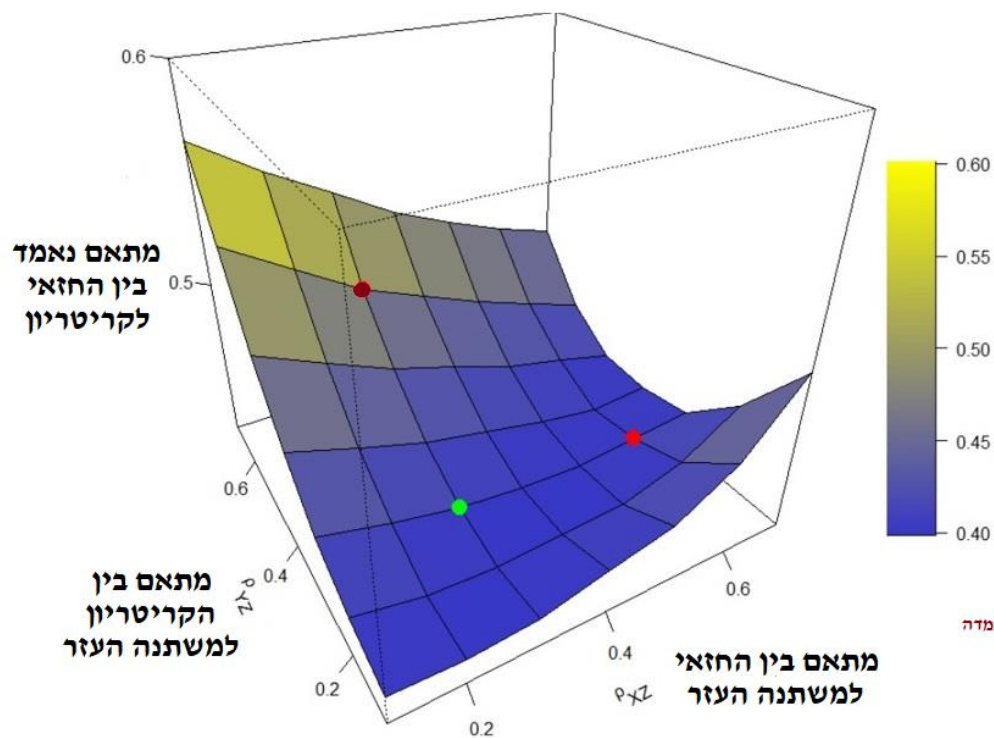
מקרא

- תיקון לקיצוץ תחום
- השלמות מרובות ללא משתנה עזר
- הקמן ללא משתנה עזר
- השלמות מרובות עם משתנה עזר
- מתאם נצפה לאחר הליך קבלה ונשירה
- הקמן עם משתנה עזר

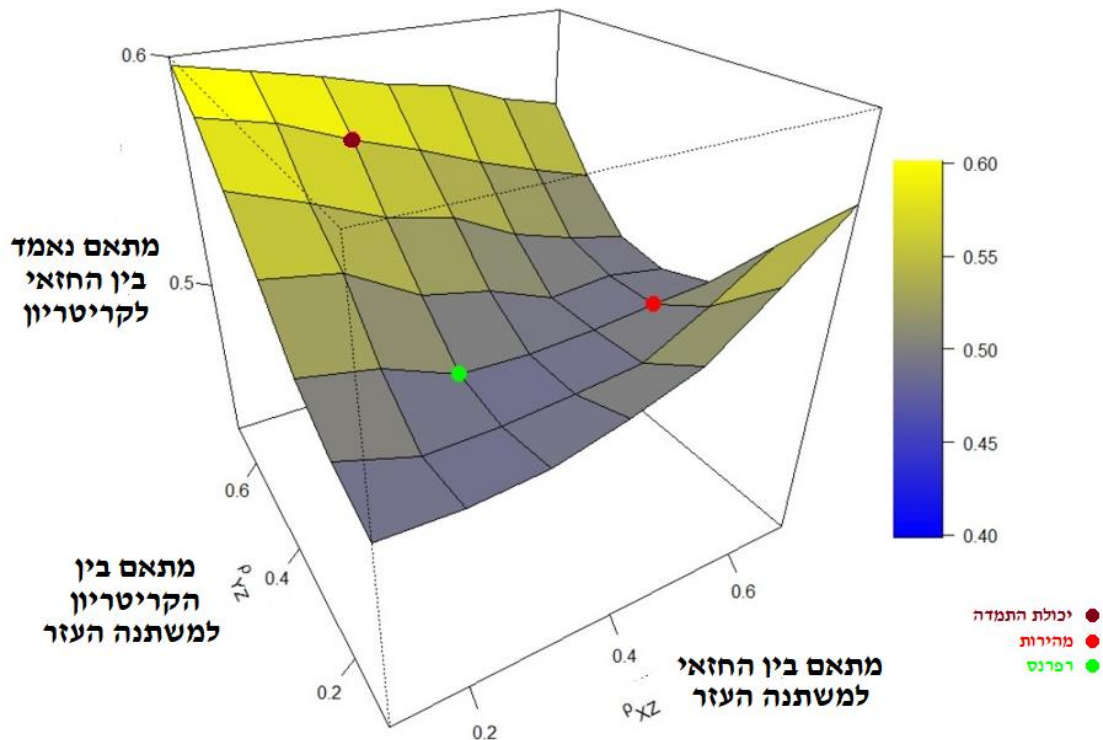
4.2. ניתוח משני 1 – העמקה: נכונות האמידה כפונקציה של המתאמים עם משתנה העזר

עבור המקרה של נשירה ודאית של החלשים, אוכלוסייה בינונית (300), $\rho_{XY} = 0.6$, ו-60% חוסרים, כשהחלק של חוסרים שנובע מדחייה הוא 0.1, בדקנו את נכונות האמידה כפונקציה של המתאמים עם משתנה העזר באופן יותר שיטתי. באיור 5 ובאיור 6 מוצגים הגרפים התלת-ממדיים של המתאם הנאמד ($\hat{\rho}_{XY}$) כפונקציה של שני משתנים בלתי תלויים: המתאם של Z עם X (ρ_{XZ}) והמתאם של Z עם Y (ρ_{YZ}), וזאת עבור השלמות מרובות עם משתנה עזר ומודל הקמן עם משתנה עזר, בהתאמה.

איור 5: תוקף נאמד כפונקציה של המתאמים של משתנה העזר עם החזאי והקריטריון עבור שיטת ההשלמות המרובות



איור 6: תוקף נאמד כפונקציה של המתאמים של משתנה העזר עם החזאי והקריטריון עבור מודל הברירה של הקמן



ראשית נציין שפרט לערכים שונים על ציר הגובה (ציר ה-z), שני הגרפים מציגים התנהגות דומה מאוד. באופן גורף, המתאמים המופקים בשיטת הקמן גבוהים יותר מהמתאמים המופקים בשיטת ההשלמות המרובות.

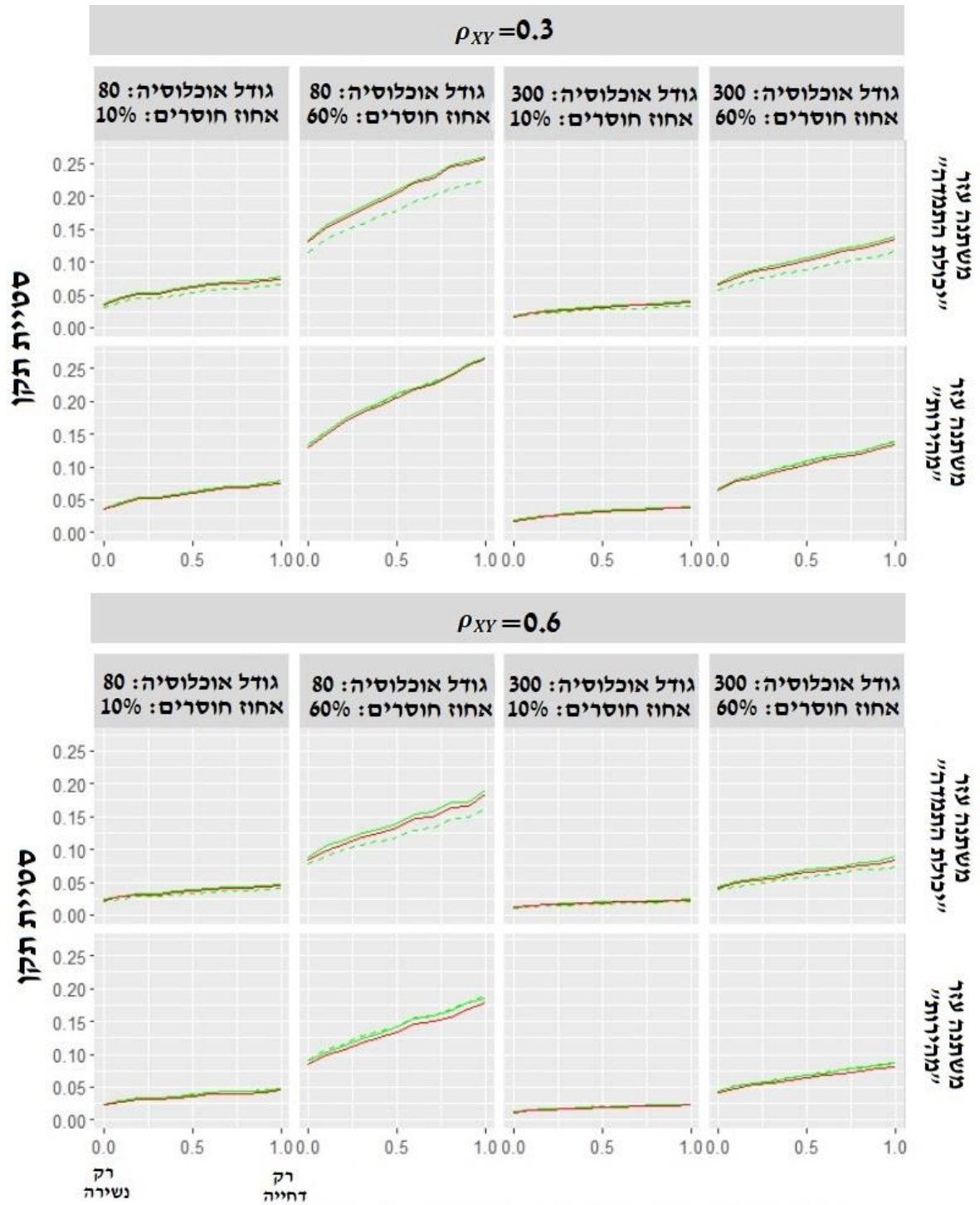
על שני הגרפים מצוירות 3 נקודות: שתי הנקודות האדומות מתאימות למשתני עזר "יכולת התמדה" (השמאלית הכהה יותר) ו"מהירות" (הימנית הבהירה יותר) והן למעשה מתאימות לערכים שהוצגו באיור 3. הנקודה הירוקה מתאימה למתאם של 0.3 הן עם החזאי והן עם הקריטריון (שכאמור הייתה אחת מהערכים הנוספים שבדקנו ונמצאת בגרפים לצורך נקודת השוואה בין שני המצבים); ניתן לראות בשני הגרפים שהנקודה הירוקה דומה יותר למקרה של משתנה עזר מסוג "מהירות". עוד נזכיר, נקודה נוספת שנבדקה, זו של מתאם של 0.6 עם הקריטריון ומתאם של 0.1 עם החזאי (הנקודה לא מוצגת בגרפים והיא למעשה הפינה השמאלית העליונה בגרפים) שטענו שתמיד מביאה לתוצאות דומות לזו של משתנה העזר "יכולת התמדה", ואכן ניתן לראות זאת בנקל בשני הגרפים וזאת כמובן רק עבור המקרה הספציפי שאנו בודקים.. בחלק השיטה טענו שלכאורה עולה מכך שרק המתאם של משתנה העזר עם הקריטריון רלוונטי לאמידת המתאם בין החזאי לקריטריון, אך ניתן לראות שהתמונה העולה משני הגרפים מורכבת יותר. מהגרפים עולה ששני המתאמים של משתנה העזר הם רלוונטיים, כשהמתאם של משתנה העזר עם הקריטריון משפיע על גודל האומדן יותר מהמתאם שלו עם החזאי. מסקנה זו נובעת מהתבוננות בשיפועים של הגרף בכיווני הצירים המתאימים: השינוי בצבעים, שמבטא את השינוי בגובה המתאם הנאמד, ניכר יותר כשמתקדמים

בציר השמאלי (המתאם בין משתנה העזר לקריטריון) מאשר כשמתקדמים בציר הימני (המתאם בין משתנה העזר לחזאי). בנוסף, מתקיימת אינטראקציה בין האפקט של המתאם של משתנה העזר עם הקריטריון לאפקט של המתאם שלו עם החזאי: בערכים נמוכים של מתאם בין משתנה העזר לקריטריון (עד 0.3 בערך) האומדן משתפר ככל שהמתאם בין משתנה העזר לחזאי גבוה יותר, בעוד שבערכים גבוהים של מתאם בין משתנה העזר לקריטריון, האומדן יורד באיכותו ככל שהמתאם בין משתנה העזר לחזאי גבוה יותר.

4.3. ניתוח משני 2 – דיוק האמידה

בחלק זה יוצגו הפיזורים מעבר ל-5,000 ההרצות השונות עבור המקרה של נשירה אקראית, וזאת רק עבור השיטה של השלמות מרובות (עם ובלי משתנה עזר) ועבור תיקון לקיצוץ תחום שכאמור יוצאות כמעט מושלמות במקרה זה. באיור 7 נציג גרפים במתכונת דומה לזו שבאיורים 2-4, רק שהפעם במקום ממוצע המתאם הנאמד מוצגת בציר ה-Y סטיית התקן שלו, שהיא טעות התקן האמפירית של אומדן התוקף. איור 8 מתמקד בפאנל העליון שבאיור 7 (כלומר: בתנאי של מתאם בין החזאי לקריטריון של 0.3 ומשתנה עזר מסוג "יכולת התמדה") עבור מקרי הקצה של נשירה בלבד ($\rho_{mr}=0$) ודחייה בלבד ($\rho_{mr}=1$) על ידי הצגתם בגרף קופסה. גרפי קופסה מאפיינים את הפיזור של האומדנים באופן הבא: קצוות הקופסה הם רבעונים 1 ו-3 (כלומר, אחוזונים 25 ו-75 בהתאמה), קו האמצע הוא החציון (כלומר אחוזון 50), וה"שפמים" (whiskers) הם ערכי הקצה (מינימום ומקסימום) של הנתונים ללא חריגים. החריגים מוצגים כנקודות ונקבעים לפי כלל שלא נפרט פה.

איור 7: טעות התקן האמפירית של אומדן התוקף במקרה של נשירה אקראית

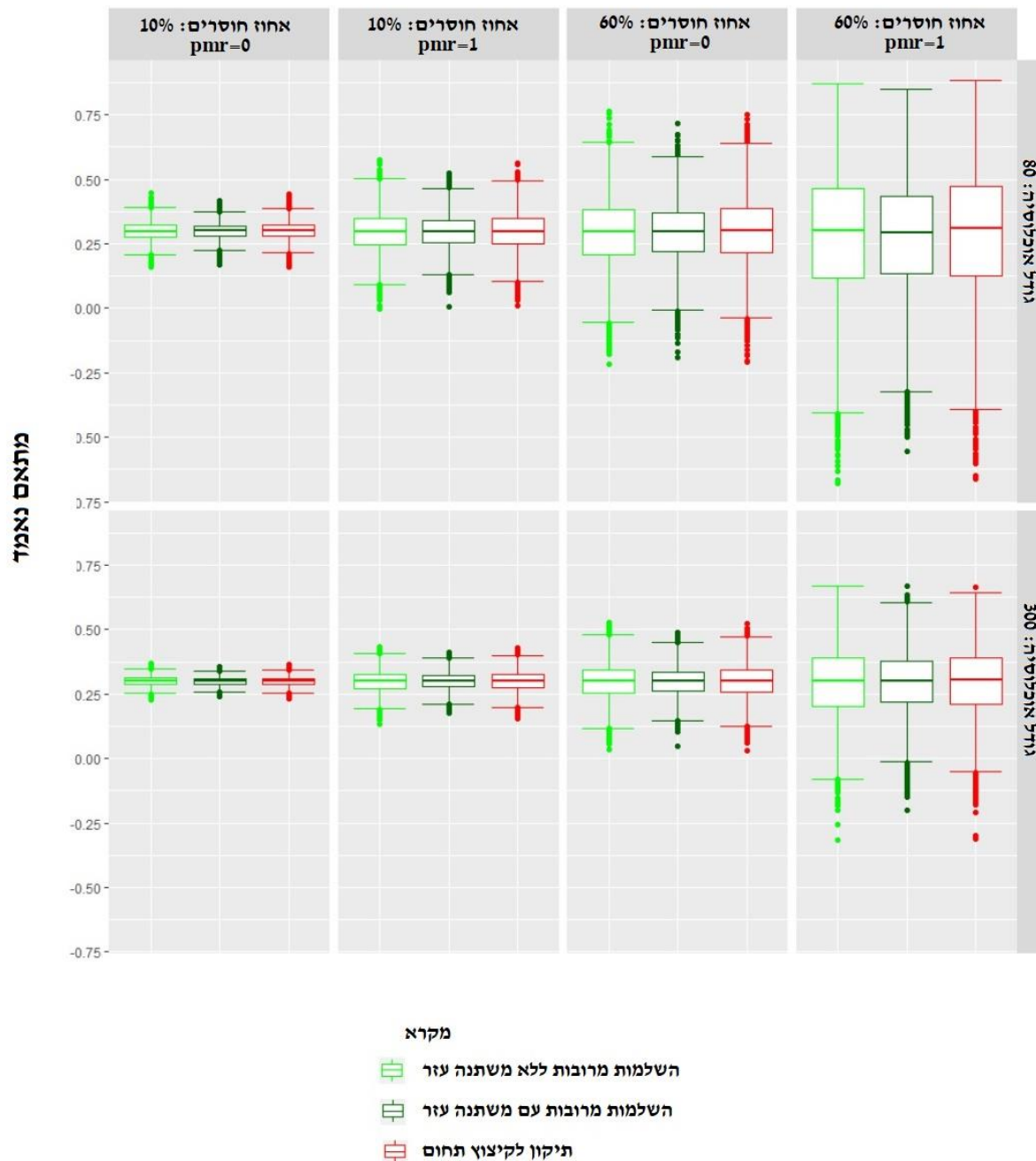


החלק היחסי של החוסרים הנובע מהליך הקבלה = $\rho_{X|Y}$

מקרא

- השלמות מרובות ללא משתנה עזר
- - - השלמות מרובות עם משתנה עזר
- תיקון לקיצוץ תחום

איור 8: גרפי קופסה עבור המתאם הנאמד עם משתנה עזר "יכולת התמדה" ומתאם 0.3 בין החזאי לקריטריון במקרה של נשירה אקראית



באיור 7 רואים שטעות התקן של אומדן המתאם קטנה יותר ככל שהאוכלוסייה גדולה יותר, ככל שאחוז החוסרים קטן יותר, וככל שהמתאם של החזאי עם הקריטריון גדול יותר. עוד ניתן לראות שכל הגרפים עולים (משמאל לימין), כלומר שכל שחלקה של הדחייה (שכמובן איננה אקראית) גדול יותר בחוסרים, כך גדלה טעות התקן. לבסוף ניתן לראות שעבור המקרה של משתנה עזר מסוג "יכולת התמדה", טעות התקן שמתקבלת בהשלמות מרובות עם משתנה עזר היא הנמוכה ביותר, כשהאפקט הזה גדול יותר במקרה של $\rho_{XY} = 0.3$. עבור משתנה עזר מסוג "מהירות", שימוש

בהשלמות מרובות עם משתנה עזר לא הקטין את הפיזור, אך גם לא הגדיל אותו (למעט מקרה אחד של אוכלוסייה קטנה וחוסרים גדולים).

מסקנות דומות עולות מאיור 8. באיור זה בולטת העובדה שעבור חוסרים גדולים יתכנו מצבים בהם המתאם הנאמד הינו שלילי; הדבר בולט במצבים של דחייה בלבד (בלי קשר לגודל המדגם) ובמצבים של נשירה בלבד בגודל אוכלוסייה קטן. עוד נשים לב שהקופסה האמצעית בכל תנאי (שצבועה בירוק כהה ומייצגת כאמור את שיטת ההשלמות המרובות עם משתנה העזר), הינה הקטנה ביותר וגם "השפם" שלה הינו הקטן ביותר, כלומר הפיזור במקרה הזה הינו הקטן ביותר. לפיכך נתוני הפיזור מספקים "שובר שוויון", במובן שהן מגלים את היתרון של שיטת ההשלמות המרובות עם משתנה עזר, בנסיבות האלה (נשירה אקראית), שבהן שלוש השיטות שמוצגות אומדות בממוצע היטב את המתאם. לבסוף נשים לב שהקו האופקי בתוך הקופסה, שמייצג כאמור את החציון, נמצא באמצע הקופסה, מה שאומר שההתפלגויות של המתאמים הן פחות או יותר סימטריות סביב החציון.

5. סיכום ודיון

5.1 השוואה בין שלוש השיטות (ללא משתני עזר)

ממצאי המחקר מראים שהתיקון לקיצוץ תחום אומד היטב את המתאם בין החזאי לקריטריון באוכלוסייה כאשר הנשירה היא אקראית (או כשאין נשירה כלל, וכל הנתונים החסרים נובעים מדחייה). הוא פחות מדויק כאשר הנשירה אינה אקראית. במצבים שבהם התיקון אינו מדויק, הוא מספק אומדן שמרני של המתאם.

שיטת ההשלמות המרובות (ללא משתנה עזר) מתפקדת באופן דומה לזה של התיקון לקיצוץ תחום (מבחינת ההיבט העיקרי של איכות האמידה – נכונות). ממצא זה אינו מפתיע שכן בדומה להשלמות מרובות, גם התיקון לקיצוץ תחום מניח מנגנון חוסרים מסוג MAR, למרות שהנחה זו לא מכונה כך באופן מפורש – בין היתר כיוון שהנחות התיקון נוסחו (Pearson, 1903) הרבה לפני שרובין הציג בספרות (Rubin, 1976) את המנגנונים של נתונים חסרים.

כאמור, כאשר הנשירה אינה אקראית שתי השיטות סיפקו אומדן חסר של המתאם – מה שנחשב ליתרון, כיוון שעדיף לקבל אומדן שמרני של איכות כלי המיון. עם זאת, חשוב להדגיש, שהממצא של אומדן חסר נובע מכך שבמחקר זה מידלנו נשירה לא אקראית כנשירה שמתואמת באופן שלילי עם הקריטריון. הנחה זו – שהנטייה לנשור חזקה יותר ככל שציוני הקריטריון נמוכים יותר היא, כמובן, הנחה סבירה, שנדונה ואוששה בספרות (קנת-כהן וברונר, 1997; Linn, 1983). עם זאת, צריך להביא בחשבון שיש הקשרים שבהם הנושרים הם בעלי ציוני הקריטריון הגבוהים יותר. מצב כזה

עשוי להתקיים בחוגים (למשל, ביולוגיה) שסטודנטים נוטים להירשם אליהם במטרה להשיג ציוני שנה א' גבוהים ובכך לשפר את סיכויי קבלתם לחוגים סלקטיביים יותר (רפואה). כאמור, בהעדר אפיון מדויק יותר של הנשירה בהקשרים ספציפיים, הנחת המתאם השלילי היא סבירה, ולפיכך הממצא של אומדן חסר של המתאם הוא ממצא אופייני, אך חשוב לזכור שהוא תלוי הנחה או נסיבות ולא תכונה קבועה של השיטות.

מודל הברירה של הקמן (ללא משתנה עזר) עדיף על שתי השיטות האחרות רק כאשר האוכלוסייה גדולה יחסית, רוב החוסרים נובעים מנשירה (ולא מדחייה), והנשירה היא נשירה ודאית של החלשים בקריטריון. במילים אחרות, מודל זה מתפקד טוב יותר משתי השיטות האחרות רק בנסיבות שבהן הנחת ה-MAR שלהן היא לא נכונה בעליל (נשירה לא אקראית, כשרוב החוסרים נובעים מנשירה) ובתנאי שהאוכלוסייה גדולה יחסית. הסיבה לתפקודו החלש של המודל של הקמן קשורה לביקורת לא מבוטלת שהועלתה כלפי מודל זה, הן משיקולים תאורטיים והן על סמך מחקרים שבוצעו על נתונים סימולטיביים – למשל, העובדה שהמודל נשען על הנחות התפלגותיות שלא ניתן לבדוק (סקירת ספרות ביקורתית ניתן למצוא אצל Enders, 2010). בהתאם לכך, ההמלצה הרווחת היא להשתמש במודל הברירה (ובמודלי MNAR אחרים) רק לצורך ניתוחי רגישות שמיישמים הנחות שונות על אותם נתונים, ולא לצורך אמידה יחידה במחקר אמפירי.

5.2 השוואה בין שלוש השיטות (עם משתני עזר)

השימוש במשתני עזר רלוונטי, כמובן, רק לשיטת ההשלמות המרובות ולמודל הברירה של הקמן. ביחס להשלמות מרובות משתנה העזר משפר את האמידה רק כאשר הוא משתנה מסוג "יכולת התמדה", כלומר, כאשר הוא מתואם יותר עם הקריטריון מאשר עם החזאי. שיפור כזה מושג רק כאשר הנשירה לא אקראית – במידה רבה יותר כאשר יש נשירה ודאית של החלשים, ובמידה פחותה במצב הביניים של נשירה מתואמת קריטריון. בכל מקרה, שימוש במשתנה העזר בהשלמות מרובות כמעט אף פעם לא מזיק. השלכות השימוש במשתנה עזר במודל של הקמן הן מעורבות, ובאופן כללי השימוש במשתנה העזר לא גורר יתרון של המודל של הקמן על פני השיטות האחרות. המסקנה היא, אם כן, שמבחינת נכונות האמידה, כאשר יש משתנה עזר מסוג "יכולת התמדה" שיטת ההשלמות המרובות עדיפה על שתי השיטות האחרות.

5.3 מאפייני משתנה העזר ותרומתם לאמידה

כפי שתואר לעיל, מאפייניו של משתנה העזר, ובפרט, המתאמים שלו עם החזאי והקריטריון, משפיעים על תרומתו לאמידת המתאם. חקירה שיטתית של הקשר בין המתאמים של משתנה העזר עם הקריטריון ועם החזאי ובין איכות האמידה הובילה למסקנה שהמתאם של משתנה העזר עם הקריטריון משפיע על גודל האומדן יותר מהמתאם שלו עם החזאי. אפשר להסביר ממצא זה בכך שמשתנה העזר מיועד מלכתחילה לעזור במידול הקריטריון (או במידול הנטייה שיהיה נתון חסר בקריטריון). לפיכך, המתאם שלו עם הקריטריון משפיע על תרומתו. הנקודה היא שהתרומה הרלוונטית היא תרומתו של משתנה העזר למידול הקריטריון **מעבר לחזאי**, שנכלל בכל מקרה במודלים השונים. לכן עוצמת הקשר בין משתנה העזר לקריטריון מייצגת את ההשפעה הכוללת של משתנה העזר על הקריטריון, והמתאם בין משתנה העזר לחזאי מייצג את החלק שיש לנכות מהשפעה הכוללת, כיוון שהוא כבר קיים במידול. במובן זה עוצמת הקשר בין משתנה העזר לקריטריון מייצגת השפעה גולמית ואילו עוצמת הקשר בין משתנה העזר לחזאי היא השפעה משנית שיש להביא בחשבון כדי להפיק את התרומה הנקייה של משתנה העזר.

ממצא נוסף שעלה מניתוח הקשר בין המתאמים של משתנה העזר עם הקריטריון ועם החזאי ובין איכות האמידה הוא קיומה של אינטראקציה בין האפקט של המתאם של משתנה העזר עם הקריטריון למתאם שלו עם החזאי: כאשר המתאם בין משתנה העזר לקריטריון הוא בינוני ומעלה, ככל שהמתאם של משתנה העזר עם החזאי גבוה יותר, תרומתו לשיפור האמידה קטנה יותר. ממצא זה מתיישב עם הניתוח שהובא לעיל (מנכים מרכיב גדול יותר מתרומתו הגולמית של משתנה העזר). לעומת זאת, כאשר המתאם בין משתנה העזר לקריטריון הוא נמוך, אזי ככל שהמתאם שלו עם החזאי גבוה יותר, תרומתו לשיפור האמידה גדלה יותר! במקרה הזה משתנה העזר מתפקד כמשתנה מדכא (suppressor variable) – הוא מדכא בשונות של החזאי את אותם חלקים שאינם רלוונטיים לקריטריון (למשל, מהירות) ובכך תורם למידול הקריטריון.

5.4 דיוק האמידה

הסיכום עד כה התייחס למדד העיקרי לאיכות התפקוד של השיטות השונות – ממוצע האומדנים (מעבר ל-5,000 החזרות), מה שמכונה בספרות "נכונות" (correctness). כאשר הנכונות היא מלאה, מעניין להשוות בין השיטות גם בפיזור של האומדנים (מעבר ל-5,000 החזרות), מה שמכונה בספרות "דיוק" (precision). השוואה כזאת עשינו עבור השלמות מרובות (עם ובלי משתנה עזר) ותיקון לקיצוץ תחום במקרה של נשירה אקראית, שבו האמידה של השיטות האלה הייתה בנכונות מלאה. הממצאים הראו שבמצב כזה יש יתרון לשיטת ההשלמות המרובות עם משתנה עזר "יכולת התמדה". ההסבר לממצא הוא שבהשלמות מרובות עם משתנה עזר (שרלוונטי יותר לקריטריון) נוסף

מרכיב במידול הקריטריון שהוא קבוע בחזרות השונות – בכל החזרות עבור תצפית נתונה ערכו הקבוע של משתנה העזר מייצב את ערך הקריטריון שהמודל משלים ולכן האמידה כולה יציבה יותר.

5.5 המלצות מעשיות

המסקנה המעשית – יישומית שעולה מן הממצאים השונים היא שמעבר לתנאים השונים, התיקון לקיצוץ תחום היא השיטה העדיפה, במובן שבמרבית התנאים היא מדייקת, וכשלא – היא מספקת אומדן שמרני של המתאם. כמו כן, נמצא יתרון מסוים לשיטת ההשלמות המרובות (כאשר משתמשים במשתנה עזר מסוג "יכולת התמדה") הן בנכונות האמידה (כאשר יש נשירה לא אקראית) והן בדיוק האמידה (כאשר יש נשירה אקראית). אבל היתרון הלא גדול של השלמות מרובות במקרים האלה מושג במחיר של סיבוכיות חישובית, לכן בשקלול כולל של יתרונות וחסרונות אנו ממליצים על השימוש בנוסחת התיקון לקיצוץ תחום.

מעבר לסיבוכיות החישובית, האתגר העיקרי של שיטת ההשלמות המרובות (עם משתנה עזר) הוא באיתור ובמדידה של משתנה עזר כזה. שלושה כיוונים אפשריים עולים בהקשר הזה. כיוון אחד הוא לזהות משתנים דמוגרפיים (כמו גיל או SES) שעשויים לתרום להסבר השונות בציוני הקריטריון מעבר לחזאי. כיוון שני הוא לאסוף מידע על משתנים פוטנציאליים במסגרת שאלונים שונים שבדקים מוכנות ללימודים אקדמיים. שאלונים כאלה (לדוגמה, שאלון ה-PRC; פיליפס-ברנשטיין וגתי, 2021) מודדים משתנים פסיכו-חברתיים שמוכרים כקורלטים של הצלחה והתמדה בלימודים, כגון משמעת עצמית וחוסן אישי (ראו ניתוח-על נרחב שסקר את המחקר על הקשר בין משתנים אלה לנשירה מלימודים אקדמיים אצל Robbins et al., 2004). כיוון שלישי הוא להשתמש בהישגים לימודיים בשלבים מוקדמים של הלימודים כדי לנבא את הקריטריון. כך, למשל, במחקר שבו הקריטריון הוא ציון הגמר לתואר בוגר (למשל, קלפר, טורוול ואורן, 2014) ניתן להשתמש בציון שנה א' כמשתנה עזר. המלצה זו מתיישבת עם ממצא חוזר ונשנה בספרות העולמית (ראו סקירה אצל Pascarella & Terenzini, 2005), בדבר התפקיד המרכזי שממלא ציון שנה א' בחיזוי נשירה או סיום תואר. מחקר שנערך לאחרונה באוניברסיטה העברית בירושלים (Yair et al., 2020) שחזר ממצא זה, והציג מדדים נוספים של ביצוע אקדמי בשלבים המוקדמים יחסית של הלימודים (מספר נקודות זכות או מספר כישלונות בשנה הראשונה) שעשויים להיות רלוונטיים.

5.6 חשיבות המחקר ומגבלותיו

מחקר זה בדק באופן שיטתי את טיב התיקון של הנוסחה לתיקון לקיצוץ בתחום בנסיבות מציאותיות שבהן, בנוסף להליך הקבלה, מתרחשת גם נשירה בקרב המתקבלים. ההשלכות של קיומה של

נשירה נידונו בספרות הפסיכומטרית (כמקרה פרטי של המצב הכללי שבו יש משתנים נוספים – שבאופן טיפוסי אינם ידועים ו/או נמדדים – שמשפיעים על הברירה, מלבד החזאי) וכאמור, אף נבדקו במחקרים אמפיריים ספורים. המחקר הנוכחי מספק תרומה ייחודית לספרות באמצעות חקירה שיטתית של הסוגייה. בנוסף, המחקר משלב יישום ובדיקה של שיטות מתקדמות לטיפול בנתונים חסרים, והנגשתן לעולם התוכן של תיקוף כלי ברירה ומיון היא תרומה חשובה.

לצד תרומותיו, חשוב להפנות את תשומת הלב למגבלות המחקר. אלו הן במידה רבה מגבלות טיפוסיות של מחקרי הדמיה: כל השאלות נבדקות תחת הנחות מסוימות, ולכן כל מסקנה שמתקבלת היא תלויה בהנחות שנעשו והחוסן של הממצאים לנוכח הפרתן. כך, למשל, אחת ההנחות המרכזיות של המחקר היא שקיים קשר שלילי בין הקריטריון לנשירה, ובהתאם, מסקנת המחקר שהתיקון לקיצוץ תחום נותן אומדן שמרני של התוקף (במצב של נשירה לא אקראית) היא תוצר של הנחה זו. כפי שצוין לעיל, זוהי הנחה סבירה ומקובלת שמתיישבת עם הנסיבות הטיפוסיות, אך קיימות גם נסיבות הפוכות של קשר חיובי בין הקריטריון לנשירה, בפרט בחוגים או מוסדות לימוד לא סלקטיביים, שחלק מהלומדים בהם עושים זאת כדי לשפר את נתוני הקבלה ולעבור למסגרות סלקטיביות יותר. הנחות נוספות שעליהן התבסס המידול (למשל, התפלגות רב-נורמלית של המשתנים) עשויות גם הן להגביל את יכולת ההכללה של ממצאי המחקר.

5.7 מחקרי המשך

ניתן להציע מחקרי המשך שיתמקדו בכיוונים הבאים:

1. יש מקום לבדיקה שיטתית של הנחות אפשריות נוספות (למשל, הנחות התפלגותיות). לאור ממצאי המחקר הנוכחי ניתן במחקרים בעתיד לדייק, ולצמצם את תנאי המידול ביחס לחלק מהמשתנים (למשל, מתאם בין החזאי לקריטריון), שנמצא שאינם משפיעים על איכות האמידה, ולהרחיבם ביחס למשתנים אחרים (למשל, מתאמים בין משתנה העזר לחזאי ולקריטריון) שנמצאו קריטיים יותר.
2. אנו בדקנו את הפיזור עבור המקרה של נשירה אקראית בלבד היות ועבור מקרה זה חלק מהמודלים יוצאים כמעט מדויקים. במחקר עתידי יש מקום לבדוק את הפיזור גם עבור המקרים האחרים היות ופיזור גדול מאוד יכול להפוך אומדן טוב לאומדן פחות טוב יחסית לאומדן בעל פיזור נמוך בהרבה.
3. במחקר זה מודלו החוסרים במשתנה הקריטריון כנגרמים מדחייה ומנשירה. אופציה נוספת שקיימת לחוסרים, והפעם במשתנה החזאי, היא קבלה אוטומטית (למשל לפי ממוצע הבגרות כאשר אין ציון פסיכומטרי). שאלה ראויה למחקר עתידי היא מה קורה לאמידת התוקף כתוצאה, למשל, ממצב של שילוב בין קבלה אוטומטית לקבלה המבוססת על צירוף של שני החזאים וכיצד השיטות השונות מתפקדות במקרה כזה.

6. מקורות

איילון, ח', ויוגב, א' (1997). **גורמי הטיה בחיזוי ההצלחה בלימודים באוניברסיטה** (נייר דיון מס' 97-3). תל אביב: המרכז לפיתוח על שם פנחס ספיר ליד אוניברסיטת תל אביב.

הלשכה המרכזית לסטטיסטיקה (2019). **מועמדים ללימודים במוסדות להשכלה גבוהה, תשס"ח-תש"פ (2019/20-2007/8)**.
<https://www.cbs.gov.il/he/publications/Pages/2019/תשסח-תשעז-201617-20078.aspx>

הלשכה המרכזית לסטטיסטיקה (2020). **שנתון סטטיסטי לישראל 2020 – מספר 71**.
<https://www.cbs.gov.il/he/publications/Pages/2020/לישראל-2020-מספר-71.aspx>

המועצה להשכלה גבוהה (ל.ת.). **קבצי נתונים סטטיסטיים אודות ההשכלה הגבוהה בישראל**.
<https://che.org.il/קבצי-נתונים-סטטיסטיים-אודות-ההשכלה-הג>

ליפלבסקי, ס. (2018). **נשירה לאחר שנת הלימודים הראשונה במערכת ההשכלה הגבוהה**. הכנס כנס של הלשכה המרכזית לסטטיסטיקה, מודיעין.
https://www.cbs.gov.il/he/Events/DocLib/kenes/kns_2_45_20.pdf

פיליפס-ברנשטיין, מ' וגתי, א' (2021). **מניעה מותאמת אישית: גישה חדשנית לצמצום הנשירה מההשכלה הגבוהה**. **אדם ועבודה**, 17, 22-34.

קלפר, ד', טורוול, א', ואורן, כ' (2014). **תוקף הניבוי של כלי המיון לאוניברסיטאות בישראל מול ממוצע תואר בוגר** (דוח 403). ירושלים: מרכז ארצי לבחינות ולהערכה.

קנת-כהן (2012). **הקשר בין כישורים לימודיים ודפוסי קבלת החלטה בבחירת לימודים ובין נשירה במערכת ההשכלה הגבוהה** (עבודת דוקטורט שלא פורסמה). האוניברסיטה העברית בירושלים, ירושלים.

קנת-כהן, ת', וברונר, ש' (1997). **תוקף מרכיבי מערכת מיון התלמידים להשכלה גבוהה: תיקון עבור הטיה הנובעת מברירה באמצעות המודל של הקמן** (דוח 236). ירושלים: מרכז ארצי לבחינות ולהערכה.

קנת-כהן, ת', ברנרן, מ', ואורן, כ' (2021). **תוקף והוגנות מערכת המיון לאוניברסיטאות בישראל** (דוח RR 21-05). ירושלים: מרכז ארצי לבחינות ולהערכה.

AERA (American Educational Research Association), APA (American Psychological Association), & NCME (National Council on Measurement in Education) (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B.* 39 (1): 1–38. JSTOR 2984875. MR 0501537.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Galimard, J. E., Chevret, S., Curis, E., & Resche-Rigon, M. (2018). Heckman imputation models for binary or continuous MNAR outcomes and MAR predictors. *BMC Medical Research Methodology*, 18:90.

Glynn, R. J., Laird, N. M., & Rubin, D. B. (1993). Multiple Imputation in Mixture Models for Nonignorable Nonresponse With Follow-ups. *Journal of the American Statistical Association*, 88(423), 984–993. <https://doi.org/10.2307/2290790>.

Gulliksen, H. (1987). *Theory of mental tests*. Hillsdale, NJ: Erlbaum. (Original work published 1950).

Heckman, J. J. (1976). The common structure of statistical models of truncated sample selection and limited dependent variables, and a simple estimator of such models. *Annals of Economic and Social Measurement*, 5, 475-492.

Lawley, D. N. (1943). A note on Karl Pearson's selection formulae. *Proceedings of the Royal Society of Edinburgh, Section A*, 62(1), 28–30.

Linn, R. L. (1983). Pearson selection formulas: Implications for studies of predictive bias and estimates of educational effects in selected samples. *Journal of Educational Measurement*, 20, 1-15.

Linn, R. L., & Hastings, C. N. (1984). A meta-analysis of the validity of predictors of performance in law school. *Journal of Educational Measurement*, 21, 245–259.

Little, R. J. A. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association*, 83, 1198-1202.

Little, R. J. A., & Rubin, D. B. (1987). Statistical analysis with missing data. New York: Wiley.

Pascarella, E. T., & Terenzini, P. T. (2005). *How college affects students: A third decade of research, Volume 2*. San Francisco: Jossey-Bass.

Pearson, K. (1903). Mathematical contributions to the theory of evolution—XI. On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions*, CC.-A 321, 1-66.

Pfaffel, A., Schober, B., & Spiel, C. (2016). A Comparison of Three Approaches to Correct for Direct and Indirect Range Restrictions: A Simulation Study. *Practical Assessment, Research, and Evaluation: Vol. 21*, Article 6. DOI: <https://doi.org/10.7275/x4ep-fv42>.

Pfaffel, A., & Spiel, C. (2016). Accuracy of Range Restriction Correction with Multiple Imputation in Small and Moderate Samples: A Simulation Study. *Practical Assessment, Research & Evaluation* 21(10), 1-20.

Pfaffel A., Kollmayer M., Schober B., & Spiel C. (2016). A Missing Data Approach to Correct for Direct and Indirect Range Restrictions with a Dichotomous Criterion: A Simulation Study. *PLoS ONE* 11(3): e0152330. DOI:10.1371/journal.pone.0152330.

Puhani, P. (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys*, 14, 53–68.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, 130, 261-288.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

- Sackett, P. R., & Yang, H. (2000). Correction for range restriction: An expanded typology. *Journal of Applied Psychology*, 85, 112-118.
- Schafer, J. L. (1998). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.
- Seidman, A. (Ed.). (2005). *College Student Retention: Formula for Student Success*. Westport, CT: ACE/Praeger.
- Westrick, P. A., Le, H., Robbins, S. B., Radunzel, J. M. R., & Schmidt, F. L. (2015). College performance and retention: A meta-analysis of the predictive validities of ACT scores, high school grades, and SES. *Educational Assessment*, 20, 23-45.
- Westrick, P. A., Marini, J. P., Young, L., NG, H., Shmueli, D., & Shaw, E. J. (2019). *Validity of the SAT for predicting first-year grades and retention to the second year*. (College Board Statistical Report). New York: The College Board.
- Wiberg, M., & Sundström, A. (2009). A comparison of two approaches to correction of restriction of range in correlation analysis. *Practical Assessment, Research, and Evaluation: Vol. 14*, Article 5. DOI: <https://doi.org/10.7275/as0k-tm88>.
- Yair, G., Rotem, N., & Shustak, E. (2020). The riddle of the existential dropout: Lessons from an institutional study of student attrition. *European Journal of Higher Education*, 1–18. doi:10. 1080/21568235.2020.1718518.
- Zimmermann, S., Klusmann, D., & Hampe, W. (2017). Correcting the predictive validity of a selection test for the effect of indirect range restriction. *BMC Med. Educ.* 17, 246 <https://doi.org/10.1186/s12909-017-1070-5>.
- Zwick, R. (2019a). Assessment in American higher education: The role of admission tests. *The ANNALS of the American Academy of Political and Social Science*, 683, 130–148.

7. נספחים

7.1. נספח 1: השלמות מרובות (Rubin, 1976)

שיטת ההשלמות המרובות (Multiple Imputation - MI) מפיקה אומדנים שהינם עקביים, יעילים וסימפטוטית, ואסימפטוטית נורמלים כאשר הנתונים הינם מסוג MAR.

הרעיון הבסיסי מאחורי השיטה הוא להשלים את הנתונים החסרים מספר פעמים, וזאת על ידי שימוש ברגרסיה בה המשתנים הקיימים מנבאים את המשתנים החסרים (אותו הליך נעשה לכל דפוס חוסרים, כאשר מנבאים כל משתנה חסר על ידי אלו הקיימים בדפוס). עבור כל אחת מההשלמות מחשבים את האומדן הרצוי⁶, ואז ממצעים את האומדנים השונים שמתקבלים לקבלת האומדן הסופי⁷. במידה והיינו מעוניינים גם בטעות התקן של האומדן⁸, הנוסחה לחישוב קצת יותר מסובכת, היות והיא צריכה לקחת בחשבון שקיימת גם טעות/אי ודאות בגלל שאנו משלימים נתונים. בהנחה שמבצעים M השלמות⁹ הנוסחה לטעות התקן עבור ההשלמות המרובות הינה:

$$S.E.(\bar{r}) = \sqrt{\frac{1}{M} \sum_k S_k^2 + \left(1 + \frac{1}{M}\right) \cdot \left[\left(\frac{1}{M-1}\right) \cdot \sum_k (r_k - \bar{r})^2\right]}$$

כאשר k הוא אינדקס ההשלמות (נע מ-1 עד ל-M), r_k הינו האומדן מכל אחת מההשלמות, $\bar{r}$ הינו הממוצע האריתמטי הפשוט של האומדנים שמתקבלים בכל אחת מההשלמות, ו- S_k^2 הינו טעות התקן של האומדן בכל אחת מההשלמות. הנוסחה שלעיל למעשה אומרת שיש למצע את שונות התקן, להוסיף לזה את שונות האומדנים (עם תיקון קטן שבא לידי ביטוי במכפלה ב- $1+1/M$) ולקחת שורש לקבלת טעות התקן.

⁶ במקרה של המחקר הנוכחי מדובר באומדן שהינו מתאם פירסון.
⁷ במקרה של המחקר הנוכחי, היות ואנו מעוניינים באומדן שהינו מתאם פירסון, ביצענו קודם טרנספורמציה פישר לכל אחד מהמתאמים שקיבלנו כדי לקבל את ערך ה-z של המתאם, וזאת בהתאם לנוסחה הבאה:

$$z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right) = \operatorname{artanh}(r)$$

את ערכי ה-z שהתקבלו מצענו. ולבסוף את הערך הממוצע $\bar{z}$ הצבנו בטרנספורמציה ההפוכה של פישר לקבלת המתאם הסופי בהתאם לנוסחה הבאה:

$$r = \frac{\exp(2\bar{z} - 1)}{\exp(2\bar{z} + 1)} = \tanh(\bar{z})$$

⁸ הנוסחה נמצאת כאן לצורך שלמות ההצגה ובהתאם מוצגת כאן עבור האומדן מתאם. נוסחה דומה קיימת עבור כל אומדן אחר.
⁹ מקובל ערך של לפחות 5 עבור מספר ההשלמות, אבל רצוי יותר. אנו למשל, מטעמי זמני הרצה, הסתפקנו במחקר זה ב-10 השלמות.

הנקודה העקרונית שיש להסביר היא כיצד לקבל מספר השלמות שונות, שהרי לו היינו משתמשים באותה הרגרסיה עבור ההשלמות השונות, כל ההשלמות היו יוצאות אותו הדבר. עושים זאת על ידי כך שאנו מכניסים גורם של רעש לרגרסיה.

אפשרות פשוטה יחסית היא להוסיף לרגרסיה את הטעות, כך שלמשל אם אנחנו עושים רגרסיה של Y על X ומתקבלת (עבור המקרים בהם אין חוסרים) משוואת הרגרסיה הבאה $\hat{y}_i = a + bx_i$, נשלים את הערכים החסרים ב- Y בהתאם למשוואה הבאה: $\tilde{y}_i = a + bx_i + S_{y \cdot x} u_i$, כאשר u_i הינו דגימה אקראית מההתפלגות הנורמלית הסטנדרטית ו- $S_{y \cdot x}$ הוא האומדן של סטיית התקן של איבר הטעות (שורש ממוצע ריבועי הטעות - RMSE).

הבעיה בגישה הפשוטה הזו היא שהיא מטפלת ב- $a, b, S_{y \cdot x}$ כאילו הם היו הפרמטרים האמיתיים ולא אומדנים. ברור שאנחנו לא יכולים לדעת מהם הערכים האמיתיים, אבל כדי לבצע השלמות מרובות בצורה מדויקת, כל השלמה צריכה להיות מבוססת על סט שונה של ערכים אלו. ביתר דיוק, ערכים אלו צריכים להיות דגימות אקראיות מההתפלגות הפוסטיריורית הבייסיאנית של הפרמטרים. רק בדרך זו שיטת ההשלמות המרובות מגלמת בתוכה בצורה נכונה את אי-הודאות ביחס לפרמטרים הבלתי ידועים.

אם כן, כיצד מקבלים דגימות אקראיות מההתפלגות אפוסטריות של הפרמטרים?

ישנן מספר גישות לבצע את זה, כאשר אנו נתאר את אחת השיטות הסטנדרטיות: שיטת הרחבת נתונים (Data Augmentation). שיטה זו מניחה שהמשתנים בבסיס הנתונים הינם מהתפלגות רב נורמלית¹⁰, אך עם זאת, באופן פרקטי מבצעת עבודה טובה של השלמות גם כאשר התפלגות חלק מהמשתנים אינה נורמלית (Schafer, 1997)¹¹.

שיטה זו הינה אלגוריתם ממשפחת שרשרת מרקוב מונטה קרלו (Markov Chain Monte Carlo - MCMC), שהינו משפחה של שיטות כלליות למציאת התפלגות אפוסטריריורית, אלגוריתם שהינו מאוד פופולארי בתחום הסטטיסטיקה הבייסיאנית.

להלן פירוט מלא של השלבים בשיטה:

1. בחר ערכי התחלה עבור הערכים של הפרמטרים, כאשר במקרה של המודל הרב-נורמלי, הפרמטרים הינם הממוצעים ומטריצת השונות המשותפת. ערכים אלו יכולים להתקבל למשל מנוסחאות רגילות עבור פרמטרים אלו תוך שימוש בשיטות סטנדרטיות שמטפלות בנתונים חסרים כמו מחיקת שורה (listwise deletion) או מחיקה בזוגות (pairwise deletion).

¹⁰ שזה אומר למעשה שכל המשתנים הם בעלי התפלגות נורמלית, וכל משתנה יכול להיות מוצג כפונקציה לינארית של כל המשתנים האחרים, ביחד עם איבר טעות שמתפלג גם הוא בצורה נורמלית והינו הומוסקדסטי (שונות קבועה).
¹¹ וזאת בניגוד למודל הברירה של הקמן ששם הנחת הנורמליות הינה יותר משמעותית.

אפשרות נוספת (ואפילו עדיפה) היא להשתמש כערכי התחלה באלגוריתם מקסום התוחלת (EM).

2. השתמש בערכים העדכניים של הפרמטרים (שהם כאמור הממוצעים ומטריצת השונות) כדי לקבל את אומדני מקדמי הרגרסיה עבור המשוואות בהן כל משתנה עם נתונים חסרים נחזה על ידי המשתנים הקיימים, זאת עבור כל דפוס של חוסרים.
3. השתמש באומדני מקדמי הרגרסיה כי לייצר ערכים חזויים לכל הערכים החסרים. לכל ערך חזוי, הוסף שליפה אקראית מהשארית הנורמלית ברגרסיה עבור אותו משתנה.
4. על ידי שימוש בבסיס הנתונים השלם (שבנוי כאמור מערכים נצפים וערכים שהושלמו), חשב מחדש את הממוצע ומטריצת השונות המשותפת על ידי שימוש בנוסחאות הרגילות (ללא חוסרים) עבורם.
5. בהתבסס על הממוצעים ומטריצת השונות המשותפת החדשה שחושבה, בצע שליפה אקראית מההתפלגות האפוסטריורית של הממוצעים ומטריצת השונות.
6. בהשתמש בשליפה האקראית של הממוצעים ומטריצת השונות המשותפת, חזור לשלב 2 והמשך במחזוריות לבצע את סדרת הצעדים עד שהתכנסות מושגת. ההשלמה שמתקבלת בשלב האחרון של האיטרציה הינה זו שמשתמשים בה להמשך החישובים.

כמה הערות לאלגוריתם:

- כדי לבצע את שלב 5, כלומר כדי לשלוף באופן אקראי מההתפלגות האפוסטריורית של הפרמטרים, יש צורך בפריור. בדרך כלל נהוג לקחת פריור לא אינפורמטיבי (noninformative) הן עבור התפלגות הממוצעים והן עבור התפלגות מטריצת השונות המשותפת.
- בשלב 6 נאמר שיש לחזור על התהליך עד להתכנסות. מדובר על התכנסות של ההתפלגות האפוסטריורית של הנתונים ושל הפרמטרים (בהינתן הנתונים הנצפים). על ידי כך שאנחנו עוברים הלך ושוב בין דגימה של הפרמטרים (בהינתן הנתונים הנצפים והחזויים) לבין דגימה אקראית של הנתונים החסרים (בהינתן הפרמטרים העדכניים) אנחנו מצליחים לעשות זאת לבסוף.

נדגיש שהשיטה לא דורשת לקבוע מיהו המשתנה התלוי וכל המשתנים נכנסים אליה באופן שווה. לבסוף, שיטת ההשלמות המרובות יושמה בשתי חלופות: עם ובלי משתנה עזר. כאשר לא השתמשנו במשתנה העזר Z , אזי ההשלמה בוצעה לפי המשתנים X ו Y (כאשר רק Y כאמור חסר). כאשר השתמשנו במשתנה העזר Z , ההשלמות בוצעה לפי 3 המשתנים X, Y, Z (ושוב, חוסר קיים רק במשתנה Y).

7.2. נספח 2: מנגנוני MNAR ומודל הברירה של הקמן (Heckman, 1976)

7.2.1. מנגנוני MNAR

שיטת ההשלמות המרובות הינה שיטה שמניחה שמגנון החוסרים הינו בר-התעלמות, כאשר הכוונה שלנו בכך היא שאין צורך למדל את ההליך שבגללו הנתונים חסרים. המרכיב המרכזי במנגנון בר-התעלמות הוא שהנתונים חסרים באקראיות (MAR), כלומר שההסתברות לחוסרים בנתונים עבור משתנה מסוים איננה תלויה בערכים של משתנה זה (אחרי לקיחה בחשבון של משתנים אחרים באנליזה).

האסטרטגיה הבסיסית לטיפול במנגנון בר-התעלמות היא כדלקמן: התאם את כל ההבדלים הנצפים בין המקרים עם ובלי נתונים חסרים, והנח כי כל שאר ההבדלים אינם שיטתיים. זו כמובן אסטרטגיה מוכרת שרגרסיה רגילה מיועדת בדיוק בשבילה.

לרוע המזל, לפעמים ישנן סיבות חזקות לחשוד שהנתונים לא חסרים באקראיות (Missing - MANR - Not at Random). למשל, על פי ההיגיון הפשוט, אנשים שהיו בבית-כלא פחות סביר שיענו על השאלה "האם היית בבית כלא?" וזאת יחסית לכאלה שלא נאסרו. דוגמא נוספת היא אנשים עם הכנסה גבוהה שפחות סביר שידווחו על גובה ההכנסות שלהם. ולבסוף, במחקרים קליניים על תרופות, אנשים אשר מצבם מדרדר יותר סביר שינשרו מהמחקר, לעומת לאנשים שמצבם הולך ומשתפר.

מה צריך להיעשות במקרים שכאלה? ישנם מודלים ושיטות לטיפול במצבים של חוסרים שאינם ברי-התעלמות, וכמובן שזה טבעי לרצות ולהפעיל אותם. אבל הבעיה הבסיסית היא שבהינתן מודל של הנתונים, יש אומנם רק מכניזם בר-התעלמות אחד, אבל הרבה מנגנונים שאינם ברי-התעלמות. לכן, טכניקות כאלה קשות יותר ליישום, ויש חשיבות גדולה לבחירה של המודל הנכון, ולו משום שבחירה שכזאת דורשת ידע מפורט ומדויק של התופעה שאנו חוקרים. גרוע יותר, אין עדיין שום דרך אמפירית כדי להבחין בין מודל שאינו בר-התעלמות אחד ממשנהו (ואפילו לא להבחין בינו לבין מודל בר-התעלמות).

הרעיון הבסיסי במידול חוסרים שהם במנגנון MANR הוא לפרק את פונקציית הצפיפות. נדגים זאת עבור משתנה יחיד עם ערכים חסרים. יהי Y משתנה שכזה ויהי R משתנה דמה עם ערך 1 אם Y נצפה ועם ערך 0 אם Y חסר. יהי $f(Y, R)$ פונקציה הצפיפות המשותפת עבור שני משתנים אלו. בחירת מודל משמעה בחירת מפרט מופרש עבור פונקציה זו. את פונקציית הצפיפות המשותפת ניתן לפרק בשני דרכים שונות (Little & Rubin, 1987), כאשר במקרה של מודלי ברירה אנו עושים את הפירוק הבא:

$$f(Y, R) = \Pr(R|Y) f(Y)$$

כאשר $f(Y)$ הוא הצפיפות השולית של Y ו- $\Pr(R|Y)$ הוא ההסתברות המותנית של R בהינתן ערך כלשהוא של Y . במילים, אנחנו קודם כל ממדלים את Y כאילו לא היו נתונים חסרים, ואז, בהינתן הערך של Y , אנו ממדלים קיום או אי-קיום של נתונים חסרים. מידול שכזה הינו מזוהה (identified) ואפשר לאומדו למשל בעזרת שיטת הנראות המקסימאלית.

פירוק אלטרנטיבי לצפיפות המשותפת מתאים למידול בשיטת ערבוב הדפוסים והינו בהתאם ל:

$$f(Y, R) = f(Y|R) \Pr(R)$$

כאשר $f(Y|R)$ הוא הצפיפות של Y מותנה אם Y חסר או לא. מידול שכזה אינו מזוהה, ולכן אינו בר-אמידה מבלי שעושים עוד הגבלות על הפרמטרים.

ראוי לציין שמידול בשיטת ערבוב הדפוסים נראה לכאורה כדרך לא טבעית לחשוב על מנגנון חוסר בנתונים. באופן טבעי, אנחנו מניחים שהערכים במשתנים (במקרה הזה Y) נקבעים מראש. ואז, בהתחשב בפרוצדורה שבה הנתונים נאספו, הערכים של Y יכולים להשפיע על האם הנתונים עליו קיימים או חסרים. דרך זו של חשיבה מתאימה למידול ברירה. מידול בשיטת ערבוב דפוסים למעשה הופך את כיוון הסיבתיות, בכך שהוא מאפשר לחוסרים להשפיע על ההתפלגות של המשתנה בעל העניין.

7.2.2. מודל הביררה של הקמן

מודל הביררה של הקמן (Heckman's Model for Sample Selection Bias) מיועד למצבים בהם המשתנה התלוי ברגרסיה לינארית חסר בחלק מהמקרים, אבל כמובן שלא בכלם. דוגמא אופיינית היא רגרסיה שחוזת שכר של נשים, כאשר השכר חסר בעיקר עבור נשים שאינן בשוק העבודה. זה טבעי לחשוב שהסבירות של נשים להיכנס לשוק העבודה היא נמוכה יותר אם השכר שלהן נמוך. לכן, הנתונים אינם חסרים באקראיות.

להלן התיאור המתמטי של המודל:

עבור Y_i משתנה רציף (שהוא למעשה משתנה הקריטריון/המשתנה התלוי) ו- X_i וקטור p -מימדי של משתנים מסבירים עבור אנשים $i = 1, \dots, n$, נניח שמתקיימת הרגרסיה הלינארית הפשוטה הבאה:

$$Y_i = X_i * \beta + \varepsilon_i$$

כאשר β הוא וקטור p -מימדי של אפקטים קבועים (מקדמי הרגרסיה). למשוואה זו נקרא משוואת התוצאה (outcome). בהנחה שהמנגנון עבור החוסרים ב- Y הוא MNAR, נמדל את החוסרים (שלא

באקראיות) על ידי הצגת משוואה נוספת, שלה נקרא משוואת הברירה (Selection), שהיא למעשה מידול פרוביט¹²:

$$P(R_{yi} = 1 | X_i^S) = \Phi(X_i^S * \beta^S)$$

כאשר R_{yi} הוא אינדיקטור לחוסרים ב- Y_i (כלומר הוא שווה ל-1 אם Y_i נצפה ושווה ל-0 אם Y_i חסר), X_i^S הוא וקטור q מימדי של משתנים נצפים (אשר באופן אידיאלי קשורים עם מנגנון החוסרים), Φ הינו הפונקציה המצטברת של ההתפלגות הנורמלית הסטנדרטית, ו- β^S הוא וקטור בלתי ידוע ממימד q של מקדמים.

בהתאם למידול פרוביט, נגדיר משתנה סמוי (latent) R^i כמשתנה נורמלי הקשור למשתנה הדיכוטומי R , כך שמתקיים $R_{yi} = 1$ אם $R_{yi}' > 0$ ו- $R_{yi} = 0$ אחרת. המודל של הקמן מקשר בין משוואת התוצאה ובין משוואת הברירה על ידי כך שהוא יוצר קשר בשגיאות משוואות הרגרסיה של שניהם. ביתר פירוט, המודל המשותף לתוצאה ולברירה מוגדר להיות:

$$\begin{pmatrix} \varepsilon^S \\ \varepsilon \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma_\varepsilon \\ \rho\sigma_\varepsilon & \sigma_\varepsilon^2 \end{pmatrix} \right) \quad \text{כאשר} \quad \begin{aligned} R_{yi}' &= X_i^S \beta^S + \varepsilon_i^S \\ Y_i &= X_i \beta + \varepsilon_i \end{aligned}$$

כאשר ρ הינו מקדם המתאם בין הטעות במשוואה הברירה (ε_i^S) והטעות במשוואת התוצאה (ε_i), ו- σ_ε הוא השונות של איבר הטעות (ε_i).

כאשר ρ שווה ל-0, הברירה והתוצאה בלתי תלויים, $E(R_{yi}^i | Y_i, X_i, X_i^S)$ בלתי תלוי ב- Y , והמנגנון של החוסרים הינו MAR. במקרה ש ρ אינו שווה ל-0, $E(R_{yi}^i | Y_i, X_i, X_i^S)$ תלוי ב- Y , והמנגנון של החוסרים הוא MNAR. ככל שערכו של ρ גדול יותר, מנגנון ה-MNAR הינו גדול יותר.

אמידת הפרמטרים במודל של הקמן ($\beta, \beta^S, \rho, \sigma_\varepsilon$) יכולה להיעשות בשתי דרכים: בשלב אחד או בשני שלבים.

אמידה בשלב אחד:

יש למקסם את פונקציה הנראות הבאה לקבלת אומדן נראות מקסימלי¹³:

¹² מידול רגרסיה פרוביט (probit) דומה מאוד למידול של רגרסיה לוגיסטית רק במקום להשתמש בפונקציה הלוגיסטית משתמשים בפונקציה המצטברת של ההתפלגות הנורמלית הסטנדרטית. נשים לב שלשתי הפונקציות הללו צורה של S מתוח, ולמעשה האמידה בשניהם בדרך כלל יוצאת דומה מאוד.

¹³ למעשה אנו רושמים את לוג פונקציה הנראות.

$$l = \sum_{\{i:R_{yi}=0\}} \log(-X_i^s \beta^s) + \sum_{\{i:R_{yi}=1\}} \left[\log \phi \left(\frac{X_i^s \beta^s + \frac{\rho}{\sigma_\varepsilon} (Y_i - X_i \beta)}{\sqrt{1 - \rho^2}} \right) - \frac{1}{2} \log 2\pi - \log \sigma_\varepsilon - \frac{1}{2} \frac{(Y_i - X_i \beta)^2}{\sigma_\varepsilon^2} \right]$$

אמידה בשני שלבים:

בהתאם למידול מתקבלת הנוסחה הבאה עבור התוחלת המותנה של התוצאה בהינתן שלא חסר נתונים:

$$E(Y_i | X_i, X_i^s, R_{yi} = 1) = X_i \beta + \rho \sigma_\varepsilon \lambda_i$$

כאשר $\lambda_i = \frac{\varphi(X_i^s \beta^s)}{\Phi(X_i^s \beta^s)}$ נקרא "יחס מילס ההופכי" (Inverse Mills Ratio - IMR); φ, Φ הינן פונקציית הצפיפות/מצטברת הנורמלית הסטנדרטית בהתאמה. היות ורכיב ה-IMR עבור כל תצפית מתאים לאיבר הטעות שנובע ממודל הפרוביט (משוואת הברירה), הקמן הציע את ההליך הבא¹⁴:

1. אמוד את פרמטרי הבחירה $\hat{\beta}^s$ על ידי שימוש בנראות מקסימלית עבור משוואת הברירה (כלומר בצע רגרסיה פרוביט).

2. עבור כל תצפית i , חשב את האומדן של יחס מילס ($\hat{\lambda}_i$) בהשתמש ב- $\hat{\beta}^s$ שחושב בשלב הקודם

3. אמוד את $\hat{\beta}$ ואת הזוג $\rho \sigma_\varepsilon$ (ביחד), בהתאם לנוסחאת לעיל של התוחלת המותנה של התוצאה כרגרסיה בשני משתנים (X -ו- λ). שונות הטעות באמידת רגרסיה זו הינה σ_ε שבעזרתה אפשר לחלץ גם את האומדן של ρ .

במידה ואומדים בשיטה הדו-שלבית, וכדי שלא נצטרך להסתמך בזיהוי על החלק הלא לינארי ביחס ההופכי של מילס (λ) ולהסתכן בכך בקולינאריות, רצוי לכלול לפחות משתנה נוסף עבור משוואת הברירה מעבר לאלו שמופיעים במשוואת התוצאה (Puhani, 2000)¹⁵.

¹⁴ בהליך יש שני שלבי אמידה (3+1 להלן), ובהתאם לכך ההליך נקרא אמידה בשני שלבים.
¹⁵ היות ואצלנו במחקר, לפחות במקרה שאין משתנה עזר, תנאי זה לא יכול להתקיים, בחרנו לעבוד רק עם האמידה בשלב אחד. נציין עוד שבשנת 1976 המחשבים שעמדו לרשות הקמן היו איטיים, ולכן הקמן הציע את הקירוב לפתרון על ידי אמידה בשני שלבים שהינו מהיר בהרבה מאמידה בשלב אחד. במחשבים שקיימים היום אין כזאת בעיה, ולכן בכל מקרה כאשר מתקיימים התנאים שבמידול מומלץ לאמוד בשלב אחד. ראוי לציין שאמידה בשלב אחד רגישה מאוד להנחת הנורמליות בעוד שאמידה בשני שלבים פחות רגישה לתנאי הנורמליות.

תחת מנגנון ה-MAR, השלמות הנתונים משתמשות כאמור בהתפלגות המותנית של המשתנה הנצפה Y בהינתן שאר המשתנים כדי להשלים ערכים חסרים ב- Y . אבל, במודל של הקמן, התוחלת הנצפית של התצפיות הקיימות ואלו החסרות ב- Y שונות אחד מהשני. רשמנו (בחלק של האמידה בשני שלבים) את הנוסחה של התוחלת של Y כאשר אין חוסרים. נוסחה מעט שונה קיימת עבור המקרה בו יש חוסרים. להלן הנוסחה שמשמשת להשלמת הנתונים החסרים ב- Y במקרה הזה¹⁶:

$$E(Y_i | X_i, X_i^S, R_{yi} = 0) = X_i \beta + \rho \sigma_\varepsilon \frac{-\varphi(X_i^S \beta^S)}{\Phi(-X_i^S \beta^S)}$$

כאמור, המודל של הקמן יושם בשתי חלופות (שתיהן באמידה חד-שלבית): עם ובלי משתנה עזר. כאשר לא השתמשנו במשתנה העזר Z , ביצענו רגרסיה של Y על X בשני המודלים – מודל הברירה ומודל התוצאה. כאשר השתמשנו במשתנה העזר Z , השתמשנו הן במשוואת הברירה והן במשוואת התוצאה בשני המשתנים X ו- Z כחזאים של Y .

¹⁶ במחקר זה השתמשנו בנוסחה זו כדי להשלים את הנתונים החסרים ב- Y , ואז חישבנו את המתאם בין X ל- Y על הנתונים המלאים. נוסחה זו פותחת פתח לביצוע מספר השלמות שונות, על ידי כך שמכניסים רעש לאמידה בנוסחה זו הן על ידי דגימה של המקדמים של הרגרסיה והן על ידי הוספת טעות אקראית (דבר דומה נעשה בשיטת ההשלמות המרובות). רעיון זה נחקר לראשונה על ידי גלימרד ושות' (Galimard, Chevret, Curis, & Resche-Rigon, 2018) ונקרא על ידו "מידול בשיטת השלמות הקמן". שם הוא גם מראה שעבור מקרים בהם יש חוסרים במשתנים נוספים (שזה לא המקרה שלנו) יש לסיבוכן נוסף זה יתרון גדול באמידה. אנו הכנסנו סיבוכן פשוט יותר על ידי כך שהכנסנו רק טעות אקראית מתפלגת נורמלית. 50 ההשלמות השונות מוצעו תוך כדי שימוש בטרנספורמציה פשוטת בדומה להשלמות המרובות.

7.3. נספח 3: קוד R עבור תוכנית המחשב

בחלק זה נתאר בקצרה את הקוד בו השתמשנו בתוכנת R כדי ליצור את הסימולציות במחקר זה ובעיקר זה שמפעיל את מודל ההשלמות המרובות ואת המודל של הקמן. חשוב לציין שעבור הפעלת המודלים אנו משתמשים בחבילות מוכנות, כלומר בקוד מוכן ולא מבצעים את המידול בעצמנו. הפונקציה העיקרית בתוכנית המחשב נקראת על ידנו `cor.calc2`. הפונקציה מקבלת מספר פרמטרים שיפורטו בהמשך ובהתאם לכך, דוגמת מדגם מההתפלגות הרב נורמלית עם שלושה משתנים שכוללים ציון במבחן קבלה (למשל ציון פסיכומטרי), קריטריון (ציון גמר לתואר) ומשתנה עזר, הופכת את חלק מהערכים בקריטריון לערכים חסרים ואומדת את המתאם בין הקריטריון למבחן הקבלה בהתאם לערכי ציון הקבלה והנתונים שנשארו בקריטריון, לפני ואחרי יישום שיטות לטיפול בחסרים¹⁷.

פרמטרים לפונקציה:

N – גודל המדגם.

missing – אחוז הנתונים שנשארים לאחר ברירה ונשירה (משתנה Miss % בתיאור המחקר).

accept – אחוז הברירה מתוך החוסרים. 0 – מייצג נשירה בלבד, 1 – מייצג ברירה בלבד (משתנה pmr בתיאור המחקר).

r_{xy}, r_{xz}, r_{yz} – המתאמים בין שלושת המשתנים במדגם הרב נורמלי. x – מייצג את המבחן הממייין (החזאי), y – את הקריטריון, z – את משתנה העזר.

r_{y2y} – מתאם בין הקריטריון למשתנה נוסף y_2 שאחראי בפועל על הנשירה (משתנה Y_p בתיאור המחקר). $r_{y2y}=0$ מייצג מקרה בו כל מקרי הנשירה (במידה וקיימים) לא תלויים כלל בקריטריון, $r_{y2y}=1$ מייצג מקרה בו כל הנשירה היא של התלמידים החלשים ביותר (ציון קריטריון הכי נמוך מבין התלמידים שנשארו לאחר הברירה).

מבנה תוכנית המחשב:

1. בשלב הראשון דוגמים 3 משתנים מהתפלגות רב נורמלית לפי המתאמים שהוגדרו. ממשיכים לדגום עד שהמתאם בפועל בין x ל-y קרוב עד כדי 0.005 מעל או מתחת למתאם התיאורטי ושני המתאמים האחרים קרובים עד כדי 0.02 למתאמים התיאורטיים. שלושת המשתנים נכנסים לתוך וקטור בשם `samp`.

¹⁷ נעיר שמטעמי נוחות התכנות, חלק מהמשתנים בתוכנית המחשב אינם זהים לאלו שמופיעים במסמך זה.

2. מחשבים כמה אנשים צריכים להישאר אחרי כל שלב (אחרי הברירה – n ואחרי הברירה והנשירה – n_{final}).
3. מחשבים את המשתנה הסמוי שגורם בפועל לנשירה (y_2) לפי הנוסחה:

$$new_y = samp[2]*ry_2y + (\sqrt{1-ry_2y^2})*rnorm(N)$$
 כאשר $samp[2]$ זה המשתנה השני של הרב נורמלי (הקריטריון) ו $rnorm(N)$ זה משתנה נורמלי בלתי תלוי. מחברים את המשתנה החדש למטריצה הרב נורמלית.
4. ממיינים את המטריצה בסדר יורד של המשתנה הראשון (x-פסיכומטרי), ומציבים ערכים חסרים משורה $n+1$ עד שורה N (במקרה בו אחוז הברירה גדול מ-0) במשתנה השני (y -קריטריון) וכן במשתנה הרביעי (מסביר נשירה).
5. ממיינים את המטריצה לפי המשתנה הרביעי בסדר יורד (כאשר שמים את הערכים החסרים בסוף) ומציבים ערכים חסרים משורה $n_{final}+1$ עד N במשתנה השני (קריטריון).
6. מחשבים את השונות של x ללא חסרים ועם השונות בהם יש נתון ב- y בלבד ואת המתאם הנצפה בין x ל- y , ואז מציבים את הנתונים בנוסחה לחישוב קיצוץ תחום.
7. מחשבים את המתאם לפי שיטת השלמות מרובות עם הפונקציה $mice$: מתחילים עם מטריצת אפסים של $2*2$ (או $3*3$ אם כוללים משתנה עזר), עושים 10 השלמות ומחשבים את מטריצת המתאמים עם טרנספורמצית פישר. בכל שלב מחברים את המטריצה למטריצת האפסים ההתחלתית ובסוף מחלקים ב-10 ועושים טרנספורמצית פישר הפוכה. $mi1 -$ זה המטריצה ללא משתנה עזר ו $mi2$ זה המטריצה עם משתנה עזר.
8. עבור שיטת הקמן מחשבים את מקדמי הרגרסיה בעזרת הפונקציה $selection$ ולאחר מכן משלימים את הנתונים החסרים לפי הנוסחה האחרונה בנספח 2, כאשר הדגימה של הרעש המקרי נעשית 50 פעמים, ולבסוף עושים ממוצע (עם טרנספורמצית פישר ואז טרנספורמציה הפוכה).

הקוד עבור שלבים 1-6 נמצא בנספח 3.1.

הקוד עבור שלבים 7-8 נמצא בנספחים 3.2-3.3, כאשר האיור הראשון מבניהם מתייחס למידול ללא משתנה עזר והשני בניהם הוא עבור מידול עם משתנה עזר.

נספח 3.1: שלבים 1-6 בתוכנית המחשב

```

cor.calc2 = function(N,missing,accept,rx,rxz,ryz,ry2y,samp){
  sig = matrix(c(1,rx,rxz,rx,1,ryz,ryz,1),nrow=3) # full correlation matrix
  n_final = round(N*missing) #number of full observations after selection and drop

  n = round(N-(N-n_final)*accept) # number of observations after selection
  repeat{ # sample until correlations are close to real correlation
    samp = mvrnorm(N,rep(0,3),sig)
    if(abs(cor(samp[,1],samp[,2])-rxy)<0.005&&abs(cor(samp[,1],samp[,3])-rxz)<0.02&&abs(cor(samp[,2],samp[,3])-ryz)<0.02)break
  }

  samp=samp[order(samp[,1],decreasing = T),] # sort by x

  new_y = samp[,2]*ry2y+(sqrt(1-ry2y^2))*rnorm(N) # new variable explaining drop

  samp = cbind(samp,new_y) # add new variable to mvnnonrm matrix

  Sx = var(samp[,1]) # variance of x

  samp = samp[order(samp[,1],decreasing = T),] # sort variables by x
  #insert missing values to y and y_new in the position of lowest x values:
  if(n<N) samp[(n+1):N,2]=NA
  if(n<N) samp[(n+1):N,4]=NA

  samp[1:n,] = samp[1:n,][order(samp[1:n,4],decreasing = T),] #sort matrix by y_new

  samp[(n_final+1):N,2] =NA #insert missing values to y in the position of lowest y_new values

  rxy.trim2 = cor(samp[,1],samp[,2],use = "complete.obs") #observed correlation between x and y after selection and drop
  sx= var(samp[1:n_final,1]) # variance of x where y values are not missing

  trim.cor.fixed2 = sqrt(Sx)*rxy.trim2/(sqrt(Sx*rxy.trim2^2+sx-sx*rxy.trim2^2))#correlation between x and y after trim correlation

```

```
# mi calculated with x and y:
corMatrix = matrix(0,nrow=2,ncol=2)
m=mice(samp[,1:2], meth=c("norm","norm"), m = 10, seed = NA)
for(k in 1:10){
  corMatrix = corMatrix + fisherz(cor(complete(m,k)))
}
mi1 = fisherz2r(corMatrix/10)
# heckman calculated with x and y:
l=selection(!is.na(samp[,2]))~samp[,1],samp[,2]~samp[,1],data =environment()
b0s= l$estimate[1]
b1s= l$estimate[2]
b0 = l$estimate[3]
b1 =beta= l$estimate[4]
sigma = l$estimate[5]
rho = l$estimate[6]
pred1 = b0s+b1s*samp[,1]
heck.fixed=0
for(i in 1:50){
  y.hat = b0+b1*samp[,1]-rho*sigma*dnorm(pred1)/pnorm(-pred1)+rnorm(N,0,sigma)
  y.hat[1:n_final]=samp[1:n_final,2]
  heck.fixed=heck.fixed+fisherz(cor(samp[,1],y.hat))
}
heck.fixed1=fisherz2r(heck.fixed/50)
```

נספח 3.3: שלבים 7-8 בתוכנית המחשב עבור המקרה שיש משתנה עזר

```
# mi calculated with x,y and z:
corMatrix = matrix(0,nrow=3,ncol=3)
m=mice(samp[,1:3], meth=c("norm","norm", "norm"), m = 10, seed = NA)
for(k in 1:10){
  corMatrix = corMatrix + fisherz(cor(complete(m,k)))
}
mi2 = fisherz2r(corMatrix/10)
# heckman calculated with x,y and z:
l=selection(!is.na(samp[,2]))~samp[,1]+samp[,3], samp[,2]~samp[,1]+samp[,3], data =environment())
b0s= l$estimate[1]
b1s= l$estimate[2]
b2s = l$estimate[3]
b0 = l$estimate[4]
b1 = l$estimate[5]
b2= l$estimate[6]
sigma = l$estimate[7]
rho = l$estimate[8]
pred1 = b0s+b1s*samp[,1]+b2s*samp[,3]
heck.fixed=0
for(i in 1:50){
  y.hat = b0+b1*samp[,1]+b2*samp[,3]-rho*sigma*dnorm(pred1)/pnorm(-pred1)+rnorm(N,0,sigma)
  y.hat[1:n_final]=samp[1:n_final,2]
  heck.fixed=heck.fixed+fisherz(cor(samp[,1],y.hat))
}
heck.fixed2=fisherz2r(heck.fixed/50)
#return results:
data.frame(rxy.observed.after.drop=rxy.trim2,rxy.after.trim.correction=trim.cor.fixed2,mi.xy=mi1[1,2],
           mi.xyz=mi2[1,2],heckman.xy=heck.fixed1,heckman.xyz=heck.fixed2)
}
```

RR-22-01
ינואר 2022