סטטיסטיקה בעידן המודרני וחשיבותה בעולם הניהול

אילה כהן [1]

תקציר

מטרת מאמר זה היא להציג מספר שיטות סטטיסטיות שפותחו בעידן המודרני, שבו כוח המחשוב גדול בהרבה לעומת זה שהיה בשנים בהן פותחו השיטות הקלאסיות של הסקה סטטיסטית. במאמר מתוארות שתי שיטות שהמוטיבציה לפיתוחן נבעה מהצורך בשיטות ניתוח עבור אוסף נתונים מאוד גדול. השיטה הראשונה היא CART- Classification & Regression Trees, והיא נועדה להתאמת מודל להסבר של משתנה תלוי על ידי משתנים מסבירים. השיטה מאתרת את המסבירים מתוך אוסף גדול מאוד של משתנים אשר כלולים בנתונים שנאספו. המודל המותאם מתאים למקרים הנפוצים, שבהם הקשר בין המשתנים המסבירים למשתנה התלוי לא מתאים לתיאור על ידי רגרסיה ליניארית. שיטה שנייה היא שיטת ההשלמות המרובות Multiple imputation, שבעזרתה משלימים נתונים חסרים באוסף שעבורו רוצים לבצע הסקה סטטיסטית. הצורך בשיטה זו נובע מכך שאומדנים המחושבים מתוך האוסף שנדגם רק על בסיס הנתונים שבהם לא היו חלקים חסרים, הם מוטים ולא מדויקים. בהמשך המאמר נדגים בדיקת מודלים תיאורטיים מורכבים הכוללים משתנים מתווכים ומתערבים. בהצגת הדוגמאות ממאמרים בשטח הניהול, מובאים הסבר והגדרות של המושגים. הדוגמאות ממחישות את החשיבות ביישום השיטות המתוארות להבנת תהליכים מורכבים בשטח הניהול.

מבוא

בשני העשורים האחרונים חלה התקדמות משמעותית בכוח המחשוב, ונפתחו אפשרויות חדשות לפיתוח שיטות סטטיסטיות המנצלות כוח זה.

שיטות ההסקה הסטטיסטית הקלאסיות פותחו על בסיס הנחות, כמו ההנחה שהנתונים מתפלגים נורמלית. איסוף נתונים בעבר לא היה קל, ולכן השיטות התאימו למדגמים לא גדולים (עשרות, מאות, אך לא עשרות ומאות אלפים, כמו בימים אלה). אולם, בפרט בעשור האחרון תודות לזמינות המחשבים עם כוח מחשוב חזק, ניתן היה לפתח שיטות שאינן מתבססות על הנחות חזקות ומתאימות לנתונים במגוון סקאלות ומבנים.

באוניברסיטאות ובפרט בחוגים שאינם החוג לסטטיסטיקה, עדיין ממשיכים ללמד את השיטות הקלאסיות בלי לחשוף את התלמידים לשיטות המודרניות. גם בפרסומים בשטחים מסוימים אנו עדים להרבה פחות שימוש בשיטת החדשות. אבל בהדרגתיות בגלל דרישות עורכי העיתונים המדעיים, החוקרים מאמצים שיטות אלה.

[1] פרופ' אילה כהן שהייתה ראש המעבדה לסטטיסטיקה בטכניון ונפטרה בחודש פברואר 2019.

במאמרם של George,Osinga,Lavie & Scott (2016), מוצגים השינויים בשיטות ניתוח נתונים בניהול, בעקבות הביצועים האפשריים בעזרת מחשבים. השינויים באים לידי ביטוי גם בדרכי איסוף הנתונים וגם בעיבודם.

באוניברסיטאות כמו בטכניון, נפתחות מגמות לימוד הנקראות Data Science ובהן מלמדים שיטות מודרניות בנוסף לקלאסיות. חוקרים בתחומים של Machine Learning, מדעי המחשב והנדסת חשמל מפתחים ומשתמשים בשיטות אלה, כאשר הדגש הוא על ניתוח של Big Data (אוסף נתונים מאוד גדול). המונח Big Data מתייחס לא רק לנפח הנתונים, אלא גם לסוגם. זה כולל גם טקסט חופשי, תמונה או קול.

מושג חשוב בסטטיסטיקה שהשימוש בו עדיין נפוץ במחקרים רבים הוא המובהקות. קל יחסית להגיע למובהקות סטטיסטית כאשר אוסף הנתונים גדול. אבל יש הבדל בינה לבין מובהקות פרקטית. לדוגמה, מחקר שבדק, בין השאר, בקרב מדגם של עובדים ממגזרים שונים את המתאם בין שביעות רצון לבין השכלה.

ערכו של סטטיסטי המבחן לבדיקת מובהקות המתאם, מחושב לפי הנוסחה:

$$\mathbf{t = \frac{r}{\sqrt{(1-r^2)}}\sqrt{n}}$$

כאשר $\mathbf{n}$ הוא גודל המדגם.

ככל שהערך המוחלט של הסטטיסטי גדול יותר, נקבל מובהקות גדולה יותר והחוקר יכתוב שיש מתאם מובהק בין שני המשתנים. אולם, זוהי מובהקות סטטיסטית בלבד ומשמעותה היא, שההשערה שאין מתאם נדחית, וקיים מתאם בין שני המשתנים שאינו אפס.

נניח, שקיבלנו במדגם שגודלו $\mathbf{n}$=40,000 מתאם של 0.02 (ובאוכלוסייה שנבדקה המתאם אכן 0.02, או מאוד קרוב לו). ערך הסטטיסטי המחושב לפי נתונים אלה יהיה 4.00. מאחר וערך זה "גדול", נדחה את ההשערה שהמתאם באוכלוסייה אפס, ונאמר שיש מתאם מובהק **<u>סטטיסטית</u>**. השאלה היא, האם גודל מתאם זה מובהק מעשית כאשר הוא כה קטן.

מחקרים רבים כיום מבוססים על גודלי מדגם ענקיים, שעבורם קשרים בין משתנים מתקבלים מובהקים סטטיסטית והעיתונים מדווחים על מובהקויות אלה. לעתים קרובות, אין לממצאים חשיבות מעשית כלל. בעיקר אנו עדים לכך כאשר מתפרסם מאמר על סקר גדול בנושא רפואי ומוצאים קשר, למשל, בין צריכה של ברוקולי לבין הסיכוי לחלות בסרטן.

בהרבה עיתונים מדעיים, כמו Management Science, העורכים המודעים לתופעה, מבקשים לדווח על "גודל האפקט" (Effect Size). ההגדרה של ערך זה תלויה במה שבודקים. במקרה של מתאם, גודל האפקט הוא המתאם $\mathbf{r}$. במקרה של השוואת ממוצעים, גודל האפקט הוא הפרש הממוצעים מחולק בסטיית התקן ומוכר בשם: Cohen's d $\mathbf{d = \frac{\bar{x}_1 - \bar{x}_2}{s}}$

$$s = \sqrt{\frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 - \sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)^2}{n_1 + n_1 - 2}}$$

הצגת שתי שיטות מודרניות בניתוחים סטטיסטיים

שתי השיטות שבחרתי להציג מבין השיטות המודרניות, הן:

1) CART - שיטה אי-פרמטרית למציאת קשר בין משתנה לבין אוסף רב מאוד של משתנים פוטנציאליים. הקשר מוצג כעץ החלטה.

2) השלמות מרובות לטיפול בנתונים חסרים.

הראשונה מייצגת שיטות לעיבוד נתונים, והשנייה להכנתם לניתוח.

Classification and Regression -CART

השיטה פותחה על ידי Breiman, Friedman, Olshen & Stone (1984) כבר לפני למעלה מ-30 שנים, אבל רק בעשור האחרון הפכה השיטה הרבה יותר פופולרית בתחומים שונים, כולל ניהול. שיטה זו שייכת למשפחת שיטות הנקראות חלוקה רקורסיבית, או עצי החלטה.

אוסף נתונים מאוד גדול - Big Data - כולל בדרך כלל הרבה מאוד משתנים, לעתים אף מאות. למשל, כאשר בחקר שווקים נאספים נתונים על הרגלי צריכה של צרכנים, או כאשר בנק אוסף נתונים על ניהול הכספים של כל לקוח. במחקרים כיום התיאוריות מתבססות על מסקנות מניתוח הנתונים, ולא כפי שהיה נהוג בעבר שתחילה פיתחו תיאוריות ולאחר מכן על בסיס נתונים בחנו את אמיתותן.

כאשר המטרה היא למצוא מודל המסביר משתנה תלוי, כמו למשל סיכוי של לווה משכנתא לעמוד בתשלום, מספר המשתנים הנדרש להסבר הוא עצום. השיטה של רגרסיה בצעדים לא ישימה, שכן היא אינה מניבה את האוסף האופטימלי להסבר. כמו כן, למודל של רגרסיה ליניארית יש מספר מוגבלויות. הוא מתאים לקשרים ליניאריים ואינו מתאים כאשר יש אינטראקציות מסדר גבוה. כלומר, כאשר צורת הקשר בין משתנה X למשתנה Y תלויה גם בערכי המסבירים האחרים שבמודל. לפיכך, פותחו שיטות שונות וביניהן השיטה של CART. לשיטה זו כמה יתרונות. היא מאפשרת קשר בין המשתנה התלוי למסבירים שיכולים להיות קטגוריים וגם נומריים. המודל המוסבר מתאים לקשרים לא לינאריים, וכאשר יש אינטראקציות. אין בעיה בהתאמת המודל כאשר יש ערכים חסרים לחלק מהמשתנים. זהו יתרון חשוב. כי כאשר יש אוסף משתנים גדול, סביר שתהיינה תצפיות חסרות לחלק מהמשתנים.

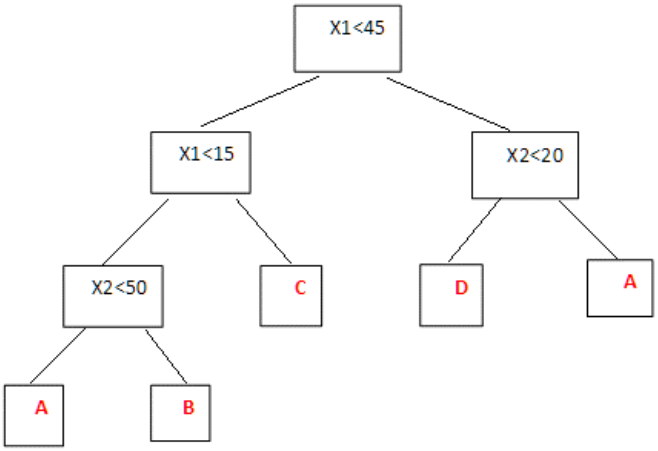
לעתים, משתמשים בשיטת CART כצעד ראשון לאיתור המשתנים המתאימים מתוך האוסף הגדול של מסבירים פוטנציאליים, וכאשר נמצא האוסף המתאים מתאימים מודל רגרסיה פרמטרי (כמו רגרסיה ליניארית או לוגיסטית) עם האוסף המצומצם, כמשתנים המסבירים במודל.

אסביר תחילה באפן כללי על מבנה העץ המסכם את תוצאות ניתוח CART. ייתכן כי ההסבר לא יהיה לגמרי מובן. אבל הדוגמאות אשר בהמשך תבהרנה את ההסבר. מומלץ לקרוא את ההסבר לפני וגם אחרי קריאת הדוגמאות.

תחילת העץ הוא השורש. על ידי פיצול בינארי המוגדר על ידי משתנה מסביר אשר נבחר מתוך האוסף הנתון של המסבירים, נוצרים צמתים. את כל אחת משתי תוצאות הפיצול מהצומת מכנים בשם "ילד" (child), כי זו תולדה של הפיצול לשניים מהצומת. כל מסלול בעץ מסתיים בעלה. המהלך מהשורש עד לעלה יוצר מסלול, וכל מסלול מגדיר תת-קבוצה על ידי צירוף תחומי ערכים של משתנים שנבחרו מהאוסף. תת-קבוצה זו כוללת את התצפיות מהאוסף שהן יחסית הומוגניות בערך של המשתנה התלוי. השורש כולל את כל התצפיות, ואילו כל עלה כולל תת-קבוצה "הומוגנית" של התצפיות שעבורן מנובא אותו ערך של המשתנה המוסבר. כל תצפית שייכת בסוף לאחד, ורק לאחד מהעלים. האלגוריתם בנוי כך שבחירת המשתנה המסביר שמגדיר את חלוקת הקבוצה בכל צומת, הוא זה שהכי טוב מחלק את תצפיות הקבוצה לחלוקה של שתי קבוצות הדומות ביותר בערך של המשתנה המוסבר והכי שונות ביניהן. בכל צומת, המשתנה שנבחר לחלוקת תת-הקבוצה לחלוקה נוספת של שתי תת-קבוצות, נבחר מתוך אוסף **כל** המשתנים. המשתנה אשר נבחר, ולפי ערכיו מוגדרת החלוקה, הוא המשתנה שחלוקתו לתת-הקבוצות נותנת את ההומוגניות הכי טובה עבור המשתנה התלוי בתוך כל אחת משתי תת-הקבוצות, והטרוגניות בין שתי תת-הקבוצות. משתנה יכול לשמש כבסיס לחלוקה מספר פעמים בחלוקות שונות. סדר הופעת המסבירים מעיד על עוצמת חשיבותם כמנבאים. הראשון הוא זה שהכי טוב מפריד לשתי תת-קבוצות, וכך בהמשך. בדרך כלל, בהצגת העץ, בכל צומת אם הכלל המתואר בגרף מתקיים, פונים שמאלה ואם לא, אז ימינה. השיטה מתאימה הן למשתנה תלוי נומרי והן לקטגורי, לכן הכינוי **קלסיפיקציה** מתאים למשתנה קטגורי, והכינוי **רגרסיה** מתאים למשתנה נומרי. מדד ההטרוגניות יכול להיות שונות, או מדד אחר הנתון גם לעתים לבחירת המשתמש.

להבהרת השיטה, אתן דוגמה פשוטה של עץ שכולל רק שני מסבירים: X1, X2. המשתנה התלוי בדוגמה זו הוא קטגורי ומקבל ערכים: A, B, C, D.

גרף מספר 1: עץ החלטה בדוגמה עם שני מסבירים

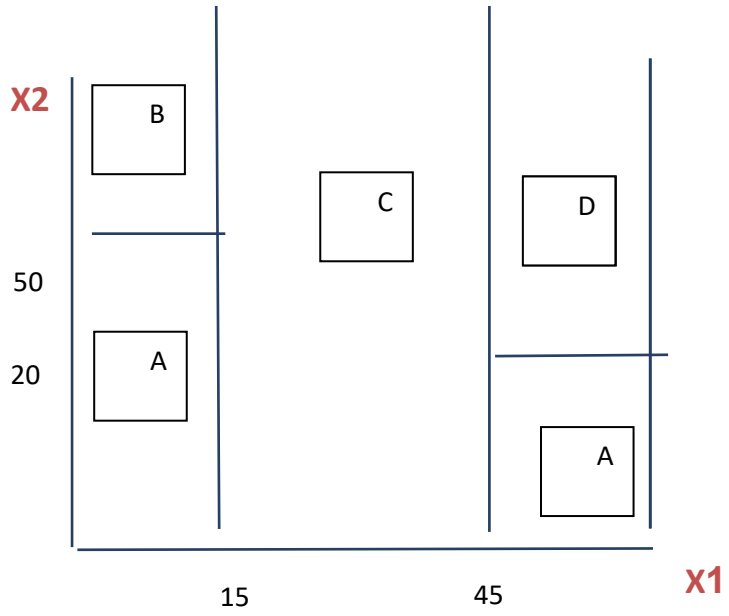


לפי המתואר בעץ, אם הערך של X1 קטן מ-15 ואם גם X2 קטן מ-50, אזי ננבא A.

ההגדרה של תת-הקבוצות שמתקבלות בסוף בעלים, ניתנת לתיאור במרחב הדו-ממדי של שני המשתנים.

גרף מספר 2: חלוקת אזורים בדוגמה עם שני מסבירים

אזור ראשון שעבורו ננבא A, כולל תצפיות שעבורן X1<15 , X2<50 ; גם עבור האזור הכולל תצפיות שעבורן X1>45 , X2<20, ננבא A.



בתוכנות רבות קיימת פרוצדורה לביצוע CART, וצורת הצגת העץ משתנה בין התוכנות. אציג כמה לדוגמה. בחרתי אותן כדי להראות את מגוון צורות הצגת הממצאים. לא כולן מהתחום של ניהול.

<u>דוגמה ראשונה</u> לקחתי מאתר

www.washburn.edu/faculty/boncella/.../Lecture%207%20-%20CART.ppt

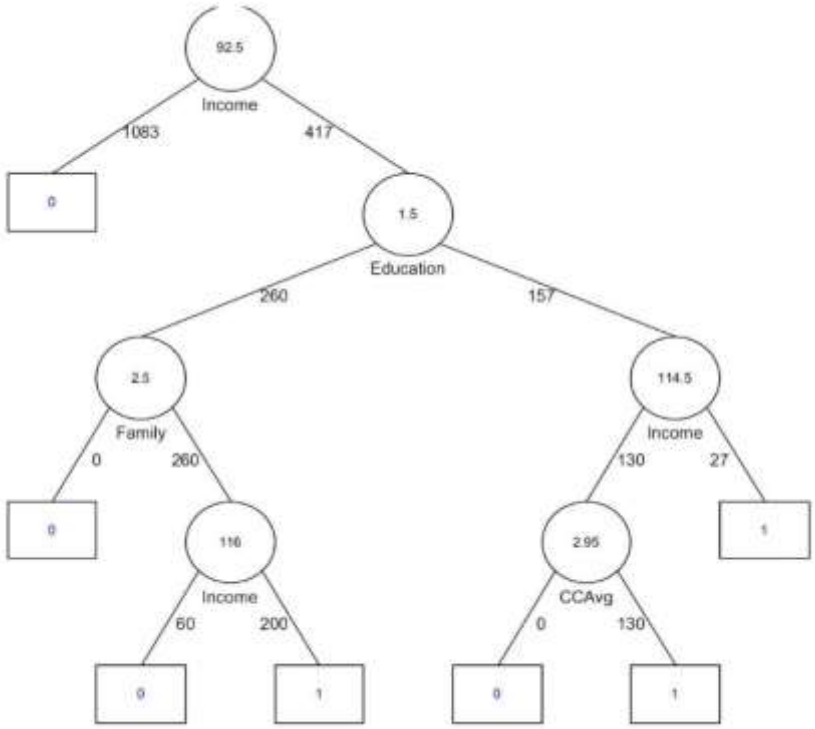
המשתנה התלוי בינארי: מאשרים כרטיס אשראי כן/לא. אם מאשרים, התוצאה מסומנת ב-1 ואם לא, אז ב-0.

בדוגמה זו, הצגת הגדרת המסלול בעץ, היא על ידי הכלל שאם ערך המשתנה המגדיר את החלוקה לתת-הקבוצות קטן מהערך בעיגול, אזי פונים שמאלה ואחרת ימינה. לכן, למשל, לפי התוצאה שמראה העץ, אם ההכנסה גדולה מ-92.5, אבל רמת החינוך קטנה מ-1.5, ואם ערך המשפחה גדול מ-2.5, וההכנסה קטנה מ-116, אזי הניבוי הוא שלא יינתן אשראי. הקבוצה שקיימה את הכלל כללה 60 אנשים. לעומת זאת, לפי התוצאה שמראה העץ, אם ההכנסה גדולה מ-92.5, אבל רמת החינוך קטנה מ-1.5, ואם ערך המשפחה גדול מ-2.5, וההכנסה גדולה מ-116, אזי הניבוי הוא שיינתן אשראי. הקבוצה שקיימה את הכלל הזה כללה 200 אנשים. לפי מה שמראה העץ, בראשיתו אוסף הנתונים כלל 417 ועוד 1,083 נבדקים, סה"כ 1,500. מתוכם, אם ההכנסה קטנה מ-92 (קבוצה שכללה 1,083 נבדקים), הניבוי הוא לשלול אשראי.

כאשר המשתנה המוסבר הוא קטגורי, הניבוי לכל תת-קבוצה הוא בהתאם לקטגוריה הכי שכיחה בתת-הקבוצה. אולם המשתמש יכול לתת משקל שונה לכל קטגוריה, ואז הניבוי אינו רק לפי השכיחות היחסית.

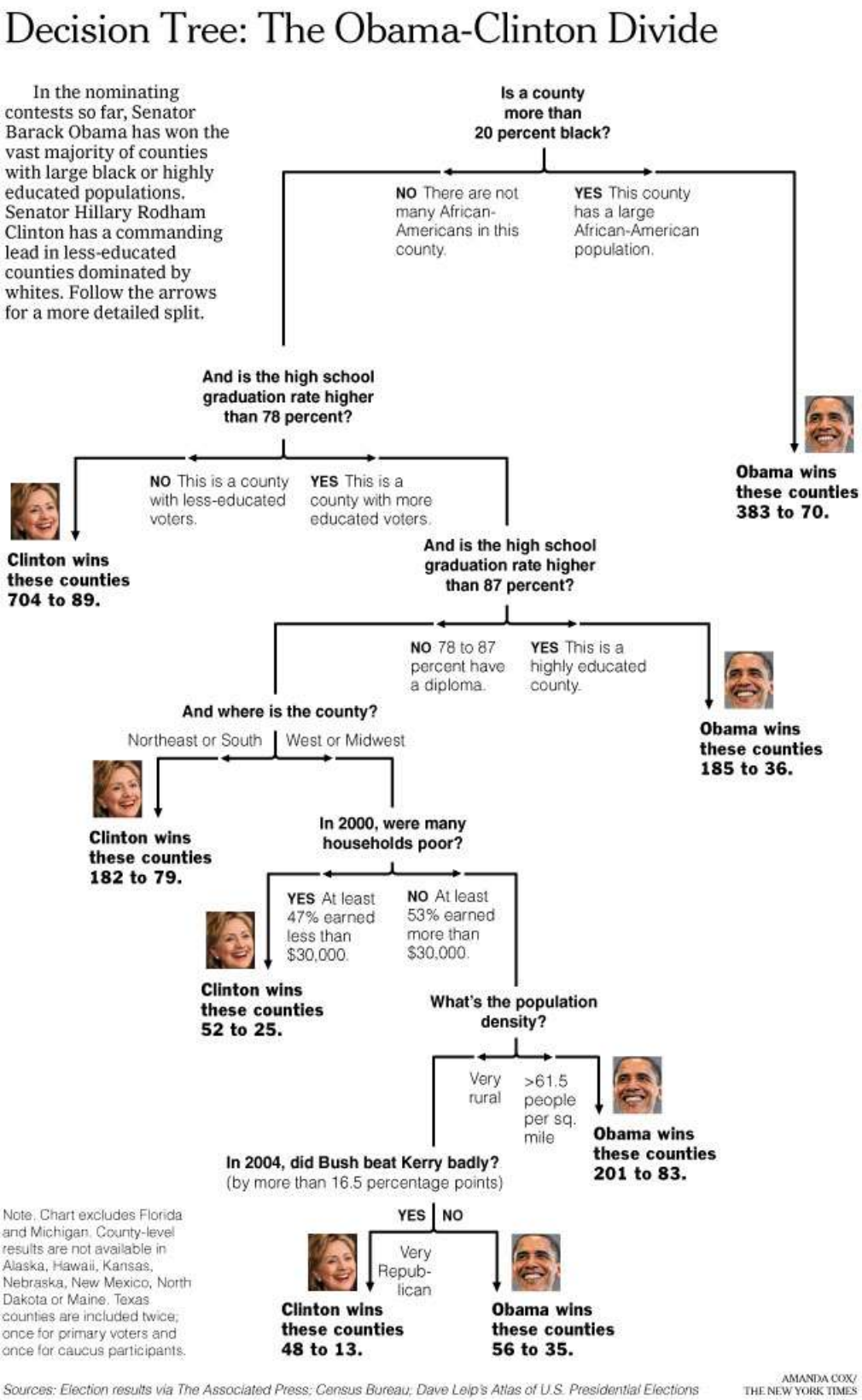
באתר לא הוצגו הנתונים על השכיחויות שהיו בפועל בכל תת-קבוצה.

גרף מספר : עץ ההחלטה לדוגמה 1



הדוגמה השנייה כפי שפורסמה ב-NY TIMES, מוצגת בצורה הברורה לקורא שלא בהכרח מכיר את השיטה, ולא אוסיף לכך הסבר.

גרף מספר 4: עץ ההחלטה לדוגמה 2

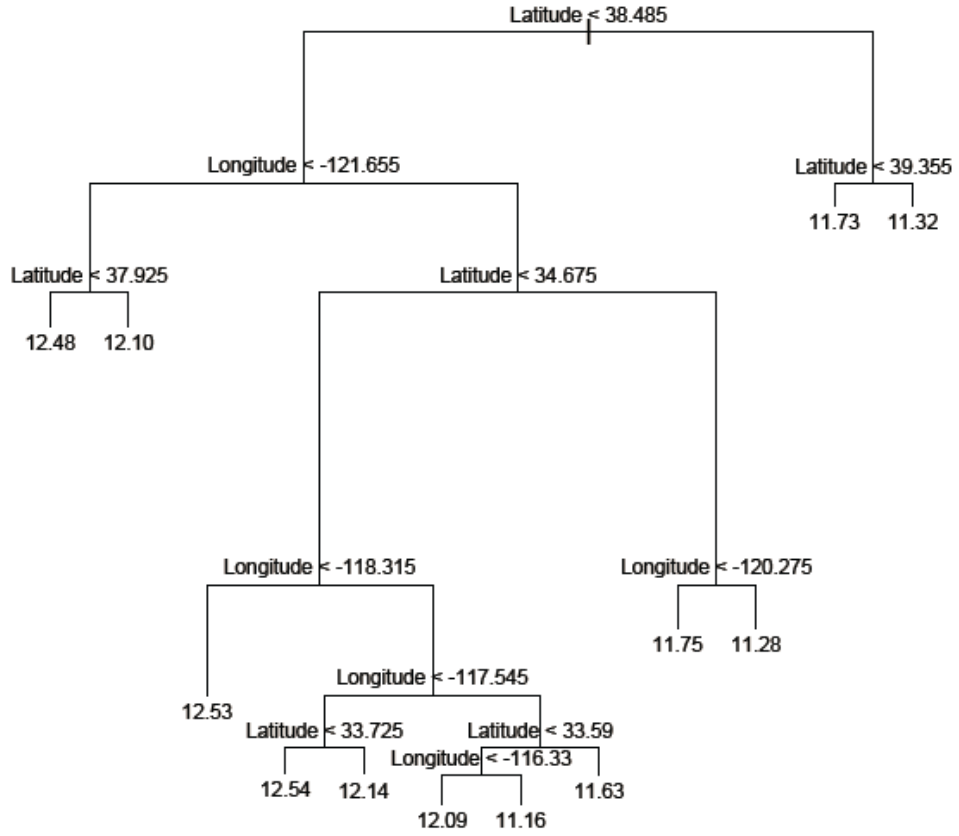


הדוגמה הבאה (מספר 3) מובאת מתוך:
https://www.stat.cmu.edu/~cshalizi/350/lectures/22/lecture-22.pdf

המשתנה המוסבר הוא מחירי נדל"ן בקליפורניה, כאשר המסבירים הם קו הרוחב והאורך של הנכס. נעקוב לדוגמה, אחר המסלול השמאלי הקיצוני. לפיו, אם קו הרוחב קטן מ-38.45 וקו האורך קטן מ-121.655, אבל קו הרוחב קטן יותר מאשר 38.45 והוא קטן מ-37.925, אזי הערך המנובא למחיר הנדל"ן הוא 12.48. לעומת זאת, אם קו הרוחב נע בין 38.485 לבין 37.925, אזי הערך המנובא למחיר הנדל"ן הוא 12.1. הערך המנובא בכל עלה שווה לממוצע ערכי הנדל"ן של תת-הקבוצה באוסף הנתונים אשר קיימה את הכלל של המסלול.

יש תוכנות שמציגות לא רק את הממוצע אלא גם את סטיית התקן, או אפילו היסטוגרמה של תתי-הקבוצות בעלים, והיתרון במידע נוסף זה הוא ברור.

גרף מספר 5: עץ ההחלטה לדוגמה 3



הגרף הבא מציג את האזורים השונים שהתקבלו כ"עלים" ומסומנים במספרים מ-1 עד 12. האזורים מסומנים לפי סדר העלים בעץ משמאל, וימינה.
בעץ הערכים עבור ערכי LONGITUDE שליליים, ובמפה חיוביים. לכן, לדוגמה, כאשר בעץ מופיע התנאי: Longitude<-121.655, אזי במפה האזור מתאים לקו אורך הגדול מ-121.655.

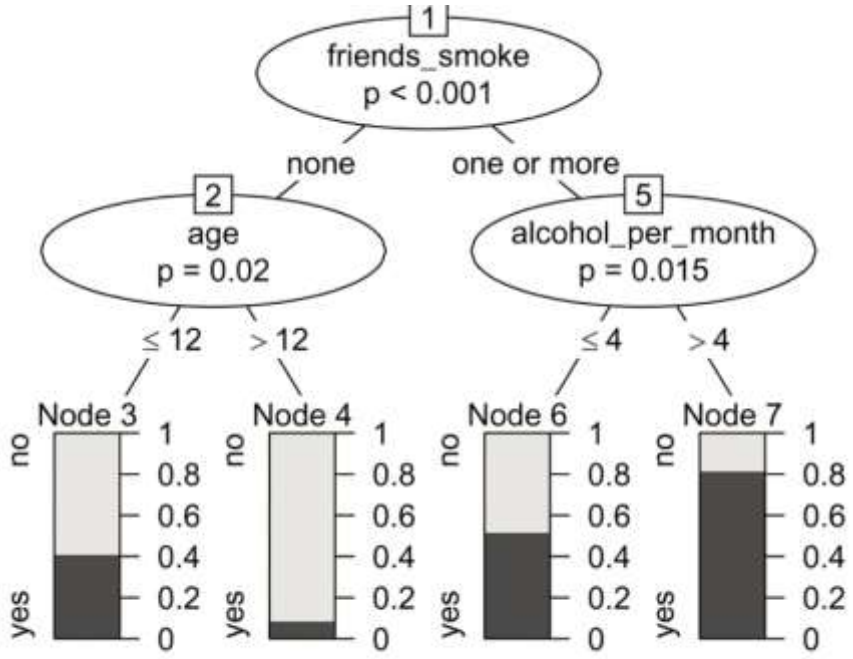
האזור המסומן במפה כמספר 12, הוא בעץ העלה הקיצוני ביותר בצד ימין. עבורו התנאי הוא:

Latitude>39.355, Longitude<-121.655. האזור המסומן 11 במפה, הוא העלה השני הקיצוני הימני ביותר בעץ, ועבורו התנאי הוא 38.485<Latitude<39.355. אזור מספר 1 במפה, הוא העלה הקיצוני ביותר בצד שמאל בעץ, ועבורו התנאי הוא: Latitude<37.925, Longitude<-121.655.

גרף מספר 6: מפת האזורים לדוגמה 3

בדוגמה האחרונה המשתנה המוסבר הוא, עישון של נער בשנה הקרובה (משתנה בינארי), מתוך המאמר Strobl, Malley, & Tutz (2009).

גרף מספר 7: עץ ההחלטה לדוגמה 4



הדוגמאות שהבאתי, ממחישות את מגוון צורות התצוגה הזמינות בתוכנות שונות. בדוגמה האחרונה, כל עלה מתואר על ידי הצגת השכיחויות של אלה שקיימו את הכללים במסלול המוביל לעלה. העונים "כן" (בשחור), לעומת העונים "לא", כשחלק המלבן המתאים הוא לבן. לכן לדוגמה, ב"עלה מספר 3", המסומן כ-Node 3, עבור אלה במדגם שחבריהם לא עישנו והיו בגיל פחות מ-12, שכיחות התשובה "כן" הייתה 0.4 (40%), ו-60% מקבוצה זו ענו בשלילה. מאחר ואחוז העונים בחיוב היה נמוך לעומת אלה שענו בשלילה, הניבוי עבור נערים שחבריהם לא מעשנים והם קטנים או שווים לגיל 12, יהיה שהם לא יעשנו. התצוגה גם כוללת בכל צומת את מובהקות מבחן ההשערה על ההבדל בערך המשתנה התלוי בין שתי תת-הקבוצות בפיצול.

ביצוע מבחן השערה הוא ברוח השיטות הקלאסיות ופחות רלוונטי כאשר אוסף הנתונים מאוד גדול וקל להגיע למובהקות סטטיסטית. ברגרסיה הקלאסית בצעדים, המובהקות הסטטיסטית מהווה קריטריון להחלטה מתי להפסיק לכלול משתנים. אולם ב-CART, יש שיטות מתאימות יותר להחלטה באיזה שלב לחתוך את הענפים ולא להמשיך בפיצולים (חיתוך זה נקרא בספרות העוסקת ב-CART בשם **pruning**).

נותרות בוודאי עוד שאלות שלא נתתי להם מענה בסקירה זו. כמו למשל, מהו הקריטריון לפיו אומדים הומוגניות של קבוצה, מתי נקבע סוף המסלול, כלומר השלב שבו לא מפצלים יותר, ומה המשמעות של כך שמשתנים באוסף לא נכללו בעץ. שאלות אלו דומות לשאלות העולות בעת התאמת מודל רגרסיה מרובה

שבה יש קריטריונים שונים, כגון אלו מסבירים יש לכלול במודל וכיצד להימנע מ"התאמת יתר" (overfitting).

מטרת המאמר הייתה רק ליידע את הקורא בשיטה ובעקרונותיה. כיום, יש ספרות עשירה ברמות שונות על הנושא. השיטה שוכללה והורחבה לשיטה הנקראת "שדות אקראיים" (random forests), שאותה לא אתאר במאמר זה.

לסיום, אציג שני מחקרים מתוך רבים, בהם יישמתי את השיטה בתחום הניהול. האחד היה, אפיון לקוחות של בנק שלקחו משכנתא. המשתנה התלוי היה קטגורי (אם המשכנתא הוחזרה בכל חודש כנדרש, אם היה פיגור במשך השנה אבל בסופה התשלום היה מלא, אם היה פיגור שנמשך). במחקר אחר המשתנה התלוי היה, מידת ההסכמה של עובדים בתוך צוותים על האקלים הארגוני בחברה.[2]

השלמות מרובות Multiple Imputation -MI

בארגונים רבים כמו במחקרי שווקים, נהוג להשתמש בשאלונים (באינטרנט, בטלפון, בראיונות) כדי להסיק מסקנות מעשיות על מדיניות החברה. בעיבוד אוסף הנתונים נוצרת בעיה גדולה עקב נתונים חסרים. לעתים קרובות, למעלה מ-50% מהתשובות שממלא נבדק אינן מלאות. סביר שתופעה זו בחלקה נובעת מהעובדה, שכל נבדק נתקל פעמים רבות בבקשה לענות על שאלון. לעתים, כאשר השאלון מופץ לצוותים, קורה שאחד או יותר מהצוות כלל לא ענה על השאלון, ואילו כאשר נבדק ממלא שאלון, הוא לא עונה על אחת השאלות, אם מתוך כוונה, או כי לא שם לב.

נהוג לסווג נתונים חסרים לפי שלוש קטגוריות. הקטגוריה הראשונה: Missing Completely at - MCAR Random, היא הכי פחות מגבילה בהמשך העיבודים. במקרה זה, לנתונים החסרים אין כל קשר עם המשתנים שנבדקו בפרויקט. לדוגמה, אם חלק מהשאלונים שהיו במשרד ללא סדר ספציפי נפלו במקרה לפח ונגרסו. סביר שאין שום קשר בין אובדנם לבין ערך המשתנים שנבדקו. רוב המקרים של נתונים חסרים אינם כאלה. הקטגוריה השנייה היא: Missing at Random - MAR. במקרה זה ההסתברות של נתון להיות חסר תלויה במשתנה מסוים. אבל למשתנה זה אין קשר עם המשתנה התלוי שעבורו נעשה המחקר. לדוגמה, במחקר על הכנסה יש סבירות גבוהה יותר שנבדקים מאזור גיאוגרפי מסוים והטרוגני מבחינת הכנסה לא ענו בגלל סיבות שונות הקשורות בלוגיסטיקה של איסוף הנתונים, אך אין הסתברות שונה שנתון יהיה חסר כאשר נשווה אנשים ברמות הכנסה שונות. הקטגוריה השלישית הכי פחות טובה לחוקר היא: MNA - Missing NOT at Random. במקרה זה חוסר הנתונים קשור לערך המשתנה הנבדק. לדוגמה, בשאלון על שביעות רצון של עובדים, אלה הפחות מרוצים נוטים לא לענות על שאלה זו. בכך נוצרת הטיה לגבי ממוצע ערך אי-שביעות הרצון, וגם מתקבלת שונות נמוכה מזו אשר בפועל.

מרבית הפיתוחים לטיפול בנתונים חסרים מבוססים על ההנחה שהנתונים שייכים לקטגוריה הראשונה או השנייה. אבל לעתים די קרובות, הנתונים שייכים לקטגוריה השלישית. למרות זאת, באופן מעשי משתמשים

[2] מומלץ לקרוא את המאמר:

George, G., Osinga, E. C., Lavie, D., & Scott, B. A. (2016). Big data and data science methods for management research,

המפרט שיטות סטטיסטיות כמו CART, כולל תיאור של מכלול השינויים באיסוף וניתוח נתונים בעידן המודרני.

בשיטת ההשלמות המרובות, שכן היא עדיין עדיפה מהשיטות שבהן מתעלמים מהבעיה בעיבוד, או כאשר משתמשים בשיטה של השמטת תצפיות עם נתונים חסרים.

בעבר הלא רחוק פותחו שיטות להשלמת נתונים, כאשר כל נתון חסר הושלם על ידי נתון אחד. אולם, שיטה זו נמצאה כלא טובה מספיק, ולכן כיום נהוג להשתמש בשיטת ההשלמות המרובות (Rubin, 1987). בשיטה זו, כאשר ערך מסוים חסר, יוצרים מספר ערכים סבירים להשלמתו על ידי מודל המבוסס על ערכי משתנים שנצפו. דרך זו של השלמה עם מספר ערכים, משקפת את חוסר הביטחון שהערכים שנוצרו שווים בדיוק לערך החסר. תהליך הניתוח בשיטת ההשלמות המרובות בנוי משני שלבים. בשלב הראשון, שהוא שלב ההשלמות, מבצעים מספר פעמים יצירה של ערך משלים עבור הערך החסר. בשלב שני, עבור כל אחד מהערכים המשלימים יש עתה אוסף ללא ערכים חסרים ועל בסיסו מבצעים את הניתוח המבוקש. לבסוף, ממצעים את התוצאות מכל אוסף שהושלם ומחשבים את השונות. רוב התוכנות הפופולאריות, כמו SAS ו-R, כוללות פרוצדורה לביצוע ההשלמות המרובות .

שיטות סטטיסטיות נוספות שנעשו פופולאריות בספרות המקצועית בנושא ניהול

מאמר זה מציג רק חלק מזערי מהשיטות שפותחו לאחרונה ושפתחו אפשרויות חדשות רבות להבנת קשרים ולבדיקת מודלים תיאורטיים בניהול.

בעוד רוב השיטות בעבר התאימו רק לניתוח נתוני מדגם של תצפיות בלתי תלויות, כיום יש שיטות רבות לניתוח נתונים בעלי מבנה היררכי, כמו צוותים בתוך מפעלים בתוך ארצות שונות; מורים בתוך בתי ספר בתוך מגזרים שונים; אחיות בתוך בתי חולים.

כמו כן, פותחו שיטות לבדיקת מודלים שבהם נבדקו קשרים בין משתנים שכוללים משתנה מתערב ומשתנה מתווך, ועל כך אפרט.

משתנה מתווך

משתנה מתווך (mediator) הוא משתנה שדרכו מועבר האפקט של משתנה מסביר על משתנה תלוי. במלים אחרות, תיווך מתייחס למכניזם של תהליך המקשר בין משתנה מקדים לתוצאה.

דוגמאות:

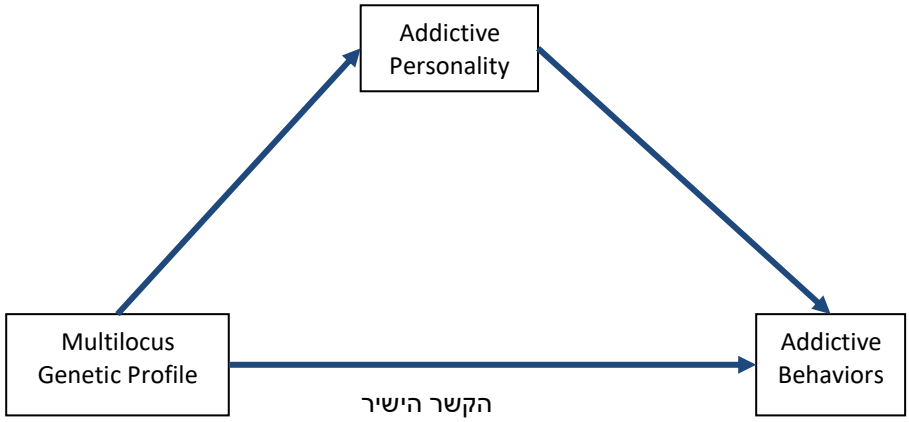
האפקט של סביבה תחרותית משפיע על האסטרטגיה של החברה, ולזה יש השפעה על ביצועי החברה.
המשתנה המסביר הוא הסביבה התחרותית, המתווך הוא האסטרטגיה של החברה, והמשתנה התלוי הוא ביצועי החברה.

לאבטלה של עובד יש אפקט על איכות תפקודו כהורה, וזה מביא לבעיות התנהגותיות של הילד.
המשתנה המסביר הוא אבטלה של עובד, המתווך הוא איכות תפקודו כהורה, והמשתנה התלוי הוא בעיות התנהגותיות של הילד.
למאפייני תעסוקה של עובד יש אפקט פסיכולוגי שמשפיע על המוטיבציה שלו.
המשתנה המסביר הוא אפיון תעסוקה, המתווך הוא האפקט הפסיכולוגי והמשנה התלוי הוא המוטיבציה.

תיווך יכול להיות מלא או חלקי. כאשר תיווך הוא מלא, כל האפקט של המשתנה המסביר על התלוי הוא דרך המשתנה המתווך. כאשר תיווך הוא חלקי, חלק מהאפקט של המשתנה המסביר על התלוי הוא דרך המשתנה המתווך, אבל חלק מהאפקט הוא ישיר.

נהוג במאמרים לתאר בגרף את המודל הכולל משתנה מתווך. כמו לדוגמה, גרף מספר 7 המתאר מודל פשוט עם משתנה מסביר, משתנה מוסבר ומתווך, ולקוח מתוך המאמר של Davis & Loxton (2013). הקשר העקיף בין הפרופיל הגנטי לבין התנהגות ממכרת הוא דרך אישיות מתמכרת. במאמר מראים המחברים כי נמצא תיווך מלא.

גרף מספר 8: מודל עם משתנה מסביר, מוסבר ומתווך



משתנה מתערב (moderator) - W הוא משתנה מתערב בין משתנה X לבין משתנה Y, דהיינו זהו משתנה שהקשר בין X ו-Y תלוי בערכים שלו. בלשון הסטטיסטיקאים, יש אינטראקציה בין W לבין X.
לדוגמה, הקשר בין אסטרטגיה או מבנה של חברה לרמת התוצר של החברה תלוי במוצר או בשירותים שהיא נותנת. מוצרי החברה/השירותים מהווים משתנה מתערב בקשר שבין לחץ האסטרטגיה והמבנה לבין רמת התוצר.

את הממצאים במחקר על האפקט של משתנה מתערב רציף, נהוג להציג בגרף הנקרא **simple slopes analysis** (Aiken & West, 1991). השיטה אמנם פורסמה כבר בשנת 1991, אולם רק בעשור האחרון הפכה להיות כלי פופולארי בספרות המקצועית של ניהול. להלן הסבר קצר:

מודל רגרסיה ליניארית עם משתנה מתערב W ומשתנה מסביר X הוא:

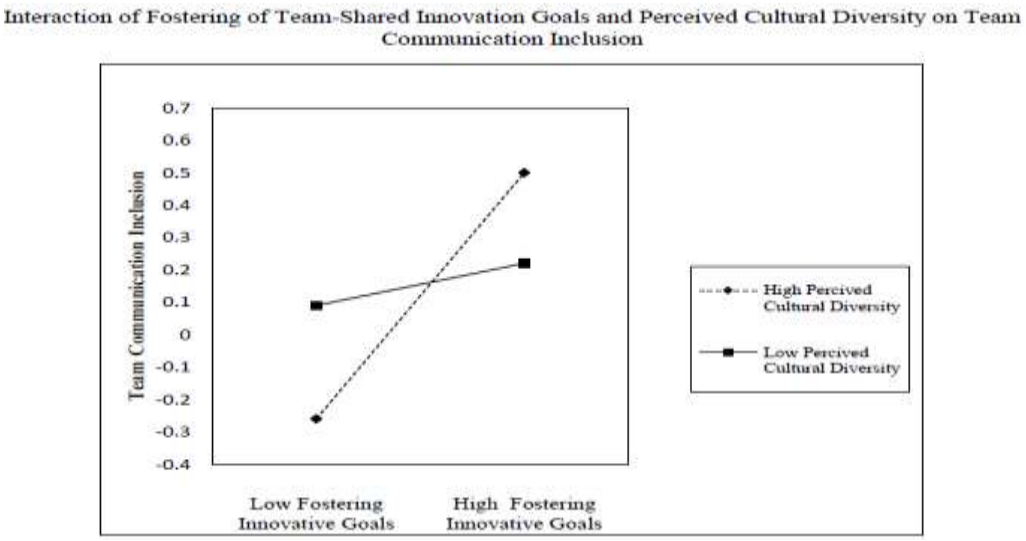
$$\mathbf{Y = \beta_0 + \beta_X X + \beta_W W + \beta_{XW} XW + \varepsilon =}$$
$$\mathbf{= \beta_0 + (\beta_X + \beta_{XW} W) X + \beta_W W + \varepsilon}$$

כלומר, השיפוע בקשר הליניארי בין X ל-Y הוא פונקציה של ערכי W.

כאשר W בינארי (לדוגמה, עובד בכיר או זוטר; זכר או נקבה; עבר השתלמות או לא עבר), אזי קל להציג בצורה גרפית את ממצאי הניתוח, מכיוון שעבור כל ערך מהמשתנה המתערב הבינארי ניתן להציג את הקו המתאר את הקשר בין X ל-Y. אבל כאשר W רציף, יש אינסוף קווים המתאימים לתיאור בהתאם לערכי W. לפי השיטה של Aiken & West, מתארים שלושה קווים המתאימים לערכי W, שהם: הממוצע של W, הממוצע פלוס סטיית התקן, והממוצע מינוס סטיית התקן. לרוב, מתארים רק שני קווים המתאימים לערכים של W "גבוה" ו"נמוך".

דוגמה לגרף עם שני קווים מוצגת בגרף מספר 8, הלקוח מתוך המאמר של Lisak, Erez, Sui & Lee (2016). במאמר זה מציינים החוקרים את שני הקווים כמתאימים לערך הגבוה והנמוך של המשתנה המתערב, אך לא מפורט האם אלה מקסימום ומינימום.

גרף מספר 9: תיאור גרפי של Simple Slopes



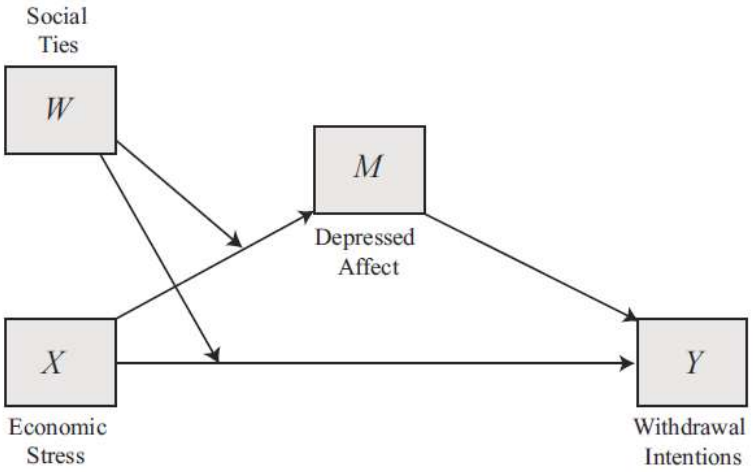
מרבית המודלים התיאורטיים כוללים משתנים מתערבים ומשתנים מתווכים.

משתנה מתערב ומתווך

דוגמה יחסית פשוטה של משתנה מתווך אחד ומשתנה מתערב אחד קיימת במאמר של Pollack, VanEpps &Hayes (2012). המודל מוצג בגרף מספר 9.

המשתנה המתערב בקשר בין המשתנה המסביר "לחץ כלכלי" לבין המתווך "דיכאון", הוא הקשרים החברתיים. המשתנה המתווך הוא בין "הלחץ הכלכלי" ל"כוונה לפרוש". המחקר נעשה על יזמים.

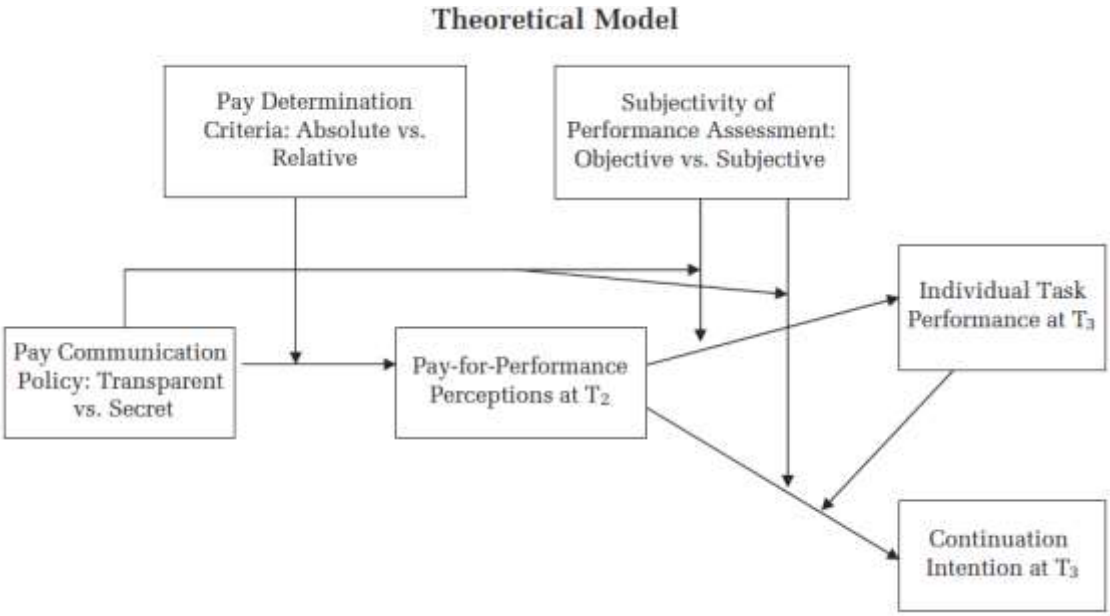
גרף מספר 10: מודל עם משתנה מסביר, מוסבר, מתערב ומתווך



משתנה מתערב יכול להיות בכל אחד מהמעברים בין שלושת המשתנים X ,Y,W. בדוגמה לעיל הוא קיים בקשר שבין X ל-M ובקשר הישיר שבין X לבין Y.

במרבית המאמרים המודל מורכב מכמה משתנים מתערבים ומתווכים, כפי שניתן לראות למשל בגרף מספר 10 הלקוח מהמאמר של Belogolovsky & Bamberger (2014).

גרף מספר 11: מודל תיאורטי (מתוך המאמר של Belogolovsky & Bamberger, 2014)



במאמר מציעים ובוחנים המחברים מודל על האפקט של המדיניות הנהוגה של "לא לחשוף מידע על השכר עבור ביצוע". המודל נבחן על ידי ניסוי במעבדה, כשהמשתנה המתווך עבור הביצוע הוא:

Pay for performance perceptions at T2 הוא Individual task performance at T3

והמשתנים מתערבים הם:

Pay determination criteria: Absolute vs. Relative

Subjectivity of performance assessment: Objective vs. Subjective

מאמר מפורט המסביר על משתנה מתווך ומתערב עם דוגמאות מתחום הניהול, הוא מאמרם של Aguinis, Edwards & Bradley (2017).

רק בעשור האחרון, תודות לכוח המחשוב הזמין, ניתן היה לפתח תוכנות מתאימות להתאמת המודלים המורכבים. בעבר, באמידת האפקט העקיף של משתנה מתווך השתמשו בקירוב, תחת הנחות לא נכונות שמכפלת משתנים נורמליים מתפלגת נורמלי. תודות לשימוש בשיטה של Bootstrap יש היום תוכנות רבות שבעזרתן ניתן לאמוד אפקט עקיף וישיר של משתנה מתווך.

לקורא שלא מכיר את שיטת ה-Bootstrap, רק אציין שזו שיטה שבנויה על העיקרון של שימוש בסימולציות במחשב ומאפשרת אמידה של רווחי סמך ומובהקות עבור מרבית הפרמטרים, גם אם התפלגות הדגימה של האמד שלהם אינה ידועה.

סיכום

תקצר היריעה מלתאר את שפע הנושאים שיכולתי לכסות במאמר שכותרתו "סטטיסטיקה בעידן המודרני וחשיבותה בעולם הניהול". בחרתי שתיים מהשיטות היותר חדשניות והמתאימות במיוחד לאוסף נתונים מהסוג של Big Data. בנוסף, הצגתי והסברתי שיטות שמתאימות גם למדגמים לא גדולים, ושרק לאחרונה נעשו פופולאריות ושאינן בהכרח מוכרות לחוקרים, סטודנטים ואנשי מקצוע בניהול.

REFERENCES

Aguinis, H., Edwards, J.R., & Bradley, K.J. (2017). Improving Our Understanding of Moderation and Mediation in Strategic Management Research. *Organizational Research Methods 20*(4), 665-685.

Aiken, L.S, & West, S.G. (1991). Multiple regression: Testing and Interpreting interactions. Newbury Park: Sage.

Beloolovsky, E. & Bamberger, P.A. (2014). Signaling in secret: Pay for performance and the incentive and sorting effects of pay secrecy. *Academy of Management Journal*, *57(6)*, 1706-1733.

Breiman, L., Friedman, J. H., Olshen, R. A., &Stone, C. J. (1984). Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software. ISBN 978-0-412-04841-

Davis, C.A., &Loxton, J.B. (2013). Addictive behaviors and addiction-prone personality traits: Associations with a dopamine multilocus genetic profile. *Addictive Behaviors*, *38*, 2306–2312.

George, G., Osinga, E.C., Lavie, D., & Scott, B.A. (2016). Big data and data science methods for management research. *Academy of Management Journal, 59(5),* 1493-1507.

Lisak, A., Erez, M., Sui,Y.,& Lee,C. (2016). The positive role of global leaders in enhancing multicultural team innovation. *Journal of International Business Studies,47(6)*, 655-673.

Pollack, J.M., VanEpps, E.M., & Hayes, A.F. (2012). The moderating role of social ties on entrepreneurs' depressed affect and withdrawal intentions in response to economic stress. *Journal of Organizational Behavior*, *33(6)*, 745-863.

Rubin, D.B. (1987). Multiple imputation for nonresponse in surveys. Wiley: New York.

Strobl, C., Malley, J., & Tutz, G. (2009). An Introduction to Recursive Partitioning: Rationale, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychological Methods, 14(4),* 323–348.