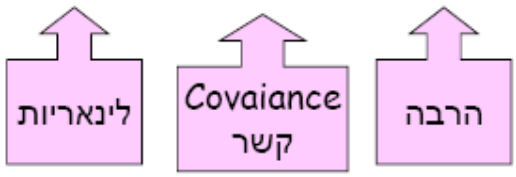
בעיית ה-Multicolinarity

מולטיקולינאריות



מולטיקולניאריות היא מתאם גבוה בין המסבירים!

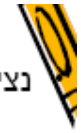
מולטיקולינאריות מושלמת

- מתאם מושלם בין שני מסבירים
- דוגמא: $X_1=2X_2$

זה נקרא קשר מולטיקוליניארי מושלם!

- נאמוד את הרגרסיה

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + u$$



נציב את X_1 ונקבל:

$$Y = \alpha + \beta_1 2X_2 + \beta_2 X_2 + u$$

$$Y = \alpha + X_2 (2\beta_1 + \beta_2) + u$$

- כלומר אנו אומדים משואה עם מסביר אחד בעזרת שני מסבירים!
- המחשב יזהה את בעיית המולטיקולינאריות המושלמת וישמיט אחד מהם

יש כאן משתנה מסביר שהוא מיותר! כיוון שניתן להגדיר באופן מושלם את אחד המסבירים על ידי השני. המחשב ישמיט את המשתנה המיותר.

דוגמא נוספת-משתנה דמי

- נניח שאנו רוצים להסביר את השכר (Y) בעזרת משתנה כמותי-שנות השכלה (EDU) ומשתנה איכותי-מין.

- אם נשתמש בשני משתנה דמי, D1 המקבל 1 עבור גבר ו-0 אחרת ו- D2 המקבל 1 עבור אישה ו-0 אחרת:

נריץ את הרגרסיה

$$Y = \alpha + \beta_1 EDU + \beta_2 D_1 + \beta_3 D_2 + u$$

וידוע ש:

$$D_2 = 1 - D_1$$

כלומר:

$$Y = \alpha + \beta_1 EDU + \beta_2 D_1 + \beta_3 (1 - D_1) + u$$

$$Y = \alpha + \beta_3 + \beta_1 EDU + (\beta_2 - \beta_3) D_1 + u$$

- לכן, תמיד מספר משתנה הדמי יהיה מספר הקטגוריות פחות אחד!

לא חובה שמולטיקוליניאריות תתקיים במשתני דמי. בכל מצב שמתקיים קשר מושלם בין 2 משתנים (כמותיים או איכותניים) תתקיים המולטיקולינאריות.

- לעיתים קיים מצב בו המולטיקולינאריות אינה מושלמת, אך יש מתאם גבוה בין שני מסבירים (מתאם הגבוה מ-0.8)

כאן מתחילים לחשוד במולטיקוליניאריות.

- המולטיקולינאריות תהיה חמורה אם המתאם בין ה-X-ים גבוה מ-0.9

- דוגמא: גיל וניסיון מסבירים את השכר, אך המתאם בין גיל לניסיון יהיה לעיתים גבוה מאוד

- אם נאמוד רגרסיה בעזרת שני המסבירים המתואמים אזי תיוצר בעיה, יש כפילות.

- בדוגמא: גיל וניסיון מסבירים את השכר, אך יש השפעות צולבות. הגיל נשפיע גם במישרין וגם בעקיפין דרך השכר.

נניח כי בחברה מסוימת העובדים מתחילים לעבוד בממוצע בגיל 25- אז יהיה מתאם מושלם בין גיל לניסיון (ניסיון שווה גיל פחות 25). אז הגיל משפיע על הניסיון שמשפיע על השכר **וגם** ניסיון משפיע על גיל שמשפיע על השכר. דבר כזה יוצר לולאה אינסופית שהמחשב לא יכול לחשב.

זיהוי הבעיה:

- בדיקת מתאם בין שני המסבירים

- שני המסבירים לא מובהקים (בעזרת מבחן t)
  אבל מקדם ההסבר (R בריבוע) יוצא גבוה

ב-data יש אופציה של בדיקת מתאם ← covariance. אם זה מעל 0.8 יש מולטי קולינאריות.

אם מריצים רגרסיה רק של 2 המשתנים החשודים במולטיקוליניאריות ומקבלים כי כל משתנה לחוד הוא לא מובהק אך R^2 מאוד גבוה. לא הגיוני כי מקדם ההסבר גבוה והמשתנים המסבירים לא מובהקים. הדרך לבדוק היא להשמיט כל פעם משתנה אחד- אם הם יוצאים לבד מובהקים מכאן שיש מולטי קולינאריות.

אם בודקים לדוגמא את המצב הבריאותי ע"י משתנה מסביר עיקרי שעות ריצה ומשתנה זה לא יוצא מובהק יש לחשוד אולי הכנסתי משתנה "שעות אימון" ויש מולטי קולינאריות בין 2 מסבירים אלו.

פתרון הבעיה

- השמטת אחד המסבירים המתואמים, המשתנה המסביר השני צריך לצאת מובהק יותר. המשתנה שנשמיט הוא זה שנותן לי פחות אינפורמציה (לדוגמא: גודל הדירה במ"ר ומספר חדרים, ניקח את גודל הדירה במ"ר).
- אם השינוי היה זניח אחרי ההשמטה סימן שאין קשר בין המסבירים.
- הוספת תצפיות, אז ייתכן שהקשר בין המשתנים המסבירים ייחלש.

<u>דוגמא- שאלה לדוגמא ממבחן</u>

חוקר מעוניין להסביר את מחירי הדירות הוא אסף נתונים על מחירי הדירות ועל המשתנים המסבירים הבאים:

1. גודל הדירה במ"ר.
2. מיקום הדירה (מחוז).
3. מרחק הדירה ממרכז העיר.
4. מספר חדרים.
5. קומה.
6. האם יש מעלית?
7. האם יש חניה?

א. האם יכולה להיווצר בעיה של מולטיקולינאריות מושלמת או חמורה אם החוקר השתמש בכל המשתנים המסבירים?
ב. במידה ונוצרת בעיה איזה משתנה מסביר תשמיט?

פתרון

א. יכולה להיווצר מולטיקולינאריות חמורה (לא מושלמת) בין גודל הדירה במ"ר לבין מספר החדרים (חדר לפי הגדרה הוא 9 מ"ר). המולטיקולינאריות היא מושלמת ולא חמורה כיוון שיכולות להיות דירות עם חדרים בגדלים שונים. אם כל החדרים בדירות היו באותו גודל היתה פה מולטיקולינאריות **מושלמת**.

ב. עדיף להשמיט את המשתנה המסביר שנותן פחות אינפורמציה, במקרה זה הוא מספר החדרים (משתנה גודל הדירה מכניס גם את הגודל הפיזי שלה וניתן להסיק ממנה את מספר החדרים).

<u>טרנספורמציה של משתנים</u>

כל טרנספורמציה על משתנה שכוללת שינוי של יחידות מדידה.

$$Y = a + b*x$$

$$Y' = k_1*y$$

$$X' = k_2*x$$

הסימון של טאג הוא לא לנגזרת הוא ל: X ו-Y החדשים לאחר הרגרסיה.

X ו-Y נאמדו בש"ח ואני ממירה אותם לדולרים ₪ $=4 ← $k_1 = k_2 = 1/4$

- b ← הוא מודד את היחס בין המשתנים, בגלל שהכפלנו אותם באותו גודל היחס לא משתנה ← $b' = (k_1/k_2) * b$
- a ← הוא משתנה כפול k, כלומר, $a'=k_1*a$.
- R^2 ← גם לא ישתנה, עדיין הרגרסיה תסביר את אותו אחוז מהשונות.
- (את הנוסחאות בצהוב יש לזכור לבחינה, לא יופיעו בדפי נוסחאות).

<u>דוגמא</u>
חוקר אמד את הרגרסיה הבאה: Income = a + b*Month
- Income- הכנסה שנתית
- Month- חודשי העבודה.

הוא קיבל: Income = 1000 + 6000* Month

הוחלט להמיר את חודשים העבודה למספר הימים בשנה שהפרט עובד. בהנחה שבחודש 30 יום, מהם המקדמים החדשים?

פתרון
K_2= 30 ← המקדם שמשנה את X
K_1=1 ← המקדם שמשנה את Y (אין שינוי ולכן זה 1).

$a' = k_1 * a = 1*1000 = 1000$

$b' = (k_1/k_2)*b = (1/30)*6000 = 200$

השיפוע יותר נמוך כי התרומה לשכר של יום היא נמוכה מהתרומה לשכר של חודש (יחידות המדידה השתנו מחודשים לימים).

<u>דוגמא 2</u>
חוקר אומד את הקשר בין התל"ג לבין הייצוא: Exp = 2,000,000 + 0.2GNP. הוחלט להמיר את יחידות המדידה מש"ח לדולרים NIS 3=$.

א. מה יהיו המקדמים החדשים?
ב. מה יקרה ל-e? (סטיות מקו הרגרסיה).

פתרון

<u>סעיף א'</u>
$k_1=k_2=1/3$

$a' = k_1 * a = 666{,}667$

$b' = (k_1/k_2)*b = 0.2$

<u>סעיף ב'</u>
e ו-a מתנהגים בנוסחא הלינארית אותו דבר (שניהם מספרים קבועים), מכאן ש:
$e' = k_1 * e$ (מתנהג בדיוק כמו a).