(F מבחן / התאמה + adj R^2)

משתני דמי: ①

עד עכשיו המשתנים ~~[illegible]~~ שלנו היו רציפים. יש משתנים שמייצגים הבחנה כמו קיים / אי קיים של תכונה מסוימת:

(נקבה) (חזק) (שחור) (יורה Rock)

(זכר) (נשוי) (נשחמה) (יורה HipHop)

$D \to$ קיים / אי קיים

דוגמא: נניח משתנה דמי לגבר / אישה

$D \to$ אישה 1, גבר 0

אפשר לשאול האם יש הבדל מובהק בציון בסטטיסטיקה בין גברים לנשים.

$$Grade_i = \beta_0 + \beta_1 \overset{?}{D_i} + \varepsilon_i \qquad \hat{\beta}_1 = \frac{\hat{Cov}(D_i, Grade_i)}{\hat{Var}(D_i)}$$

המשמעות של β_1:

$$E(Grade \mid male) = \beta_0$$
$$E(Grade \mid female) = \beta_0 + \beta_1$$

המקדם β_1 הוא ההפרש בתוחלת הציון בין גבר לאישה

באותו אופן היה אפשר להפוך את זה ונקבל מקבלים: $\beta_1 \to -\beta_1$ (-5)

אפשר גם להרחיב את המודל ולהוסיף משתנים מסבירים רציפים: כמו מספר שעות לימוד: ← נרצה שהאומדים יתאימו לאמת יותר.

$$Grade_i = \beta_0 + \beta_1 D_i + \beta_2 Hours_i + \varepsilon_i$$

עכשיו הפרשנות היא ההבדל בין גברים לנשים בהינתן כמה שעות.

שים לב שהמודל הבסיסי מניח ש... התשואה לשעת לימוד זהה לגברים ולנשים.

אם רוצים להפטר מההנחה הזו, מכניסים מונח של $D_i \times Hours$

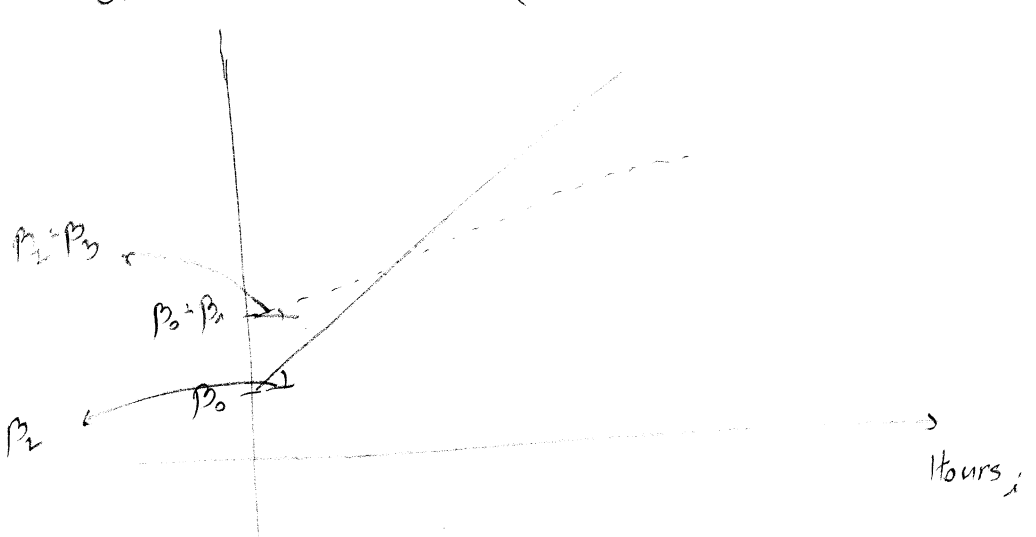
$$Grade_i = \underbrace{\beta_0}_{\text{תוחלת הציון של גבר שלא לומד}} + \underbrace{\beta_1 D_i}_{\text{הפרש בין גבר ואישה ברמה של שלא לומדי}} + \underbrace{\beta_2 Hours_i}_{\text{תשואה ללימוד עבור גבר}} + \underbrace{\beta_3 D_i \cdot Hours_i}_{\text{ההבדל בתשואה ללימוד בין אישה לגבר}} + \varepsilon_i$$

עדיף לחשוב ב2 פונקציות

( ) + ( )

אם זה נראה גרפית:

(2 קווי רגרסיה בגרף אחד)



איך יוצרים משתנה כזה בתוכנה?

→ להביא פלט של התשואה המשוערת

משמעות הנגזרות:

$$\frac{dGrade}{dHours}\Big|_{\text{גבר}} = \beta_2 \qquad \frac{dGrade}{dHours}\Big|_{\text{אישה}} = (\beta_2+\beta_3)$$

מה קורה כשיש משתנים שהם לא כמותיים ← כלומר יותר מ2 קטגוריות

$$Liquor = f(\text{מצב משפחתי})$$ → אין משמעות למספרים

רווק	נשוי	אלמן	פרוד / גרוש
1	2	3	4

הדוגמה של [illegible]

(התגרש כי היה פרוד)

<u>הכלל:</u> אם יש m קטגוריות צריך להגדיר m-1 משתני דמי וכולם צריכים להופיע במשוואה.

	D_1	D_2	D_3
רווק	0	0	0
נשוי	1	0	0
אלמן	0	1	0
גרוש	0	0	1

← קבוצת הבסיס היא "רווקים" לא מקבל משתני דמי. כל הפרשנות תהיה ביחס לקבוצת הבסיס (תוספת או גריעה)

<u>המודל שנוצר:</u>

$$Liquor = \beta_0 + \beta_1 \cdot D_1 + \beta_2 \cdot D_2 + \beta_3 D_3 + \varepsilon_i$$

אפשר גם להפטר מהחותך ולהציג:

$$\beta_0 \cdot D_0$$

0 אחרת

1 גרוש

~~הפרשנות~~ β המקדמים כבר לא בהפרשים, אלא פשוט תוחלת β הקבוצה עצמה.

4 מה קורה עם 2 מדומים של קטגוריות (2 סוגים של Dummy):

תרחיש הזוגיות והבנות:

$D_1 \to 0$ בלי זוגיות, $D_1 \to 1$ זוגיות

$D_2 \to 0$ בלי הבנות, $D_2 \to 1$ עם הבנות

$$Fun_i = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 (D_1 \times D_2) + \varepsilon_i$$

כ-ף

משתנה אינטראקציה
אני שואל האם השילוב
של זוגיות עם הבנות
משנה את התוצאה במשהו

דוגמה נוספת:

+10, -3, -2, -1

שכר שבועה = $\beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 (D_1 D_2) + \varepsilon_i$

Income

D_1: 0 עיר, 1 עיירת פיתוח

D_2: 0 מרכז, 1 דרום

$D_1 D_2$: עיירת פיתוח בדרום

(ננית מדרימים) יותר כספים לעיירות פיתוח בדרום.

לעבור עם טבלה

משכורת קבועה הבסיס

	עיר	עיירת פיתוח
מרכז	10	7
דרום	8	4

-3, -2, -1

$\beta_2 = -2$

* חשוב: כשיש יותר ממימד אחד בשאלה הרלוונטית
צריך לשים לב שקטגוריית הבסיס היא הבסיס בכל המימדים
(בכל המימד הבסיסי יהיה במימד של חישוב)

* נניח → חברה ציבורית, לא נסחרת ב S and P.

באותו אופן אפשר לבנות מודל עם שלושה שונים.
לשים לב → מה קורה כאשר שמים את כל הקטגוריות כי
החות כמויות

מאוד טוב כשבא מבחני השערה

מבחן צ'או פשר ← ~~גרסת~~ גרסת משתני הדמי:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 D_i X_i + \varepsilon_i$$

אפשר באמצעות מבחן t לשאול במה האוכלוסיות שונות

מבחן F a צ'או פשר הוא כמובן:

$H_0: \beta_1 = \beta_3 = 0 \longrightarrow$ אוכלוסיות זהות

$H_1:$ otherwise בהתנהגות שונה

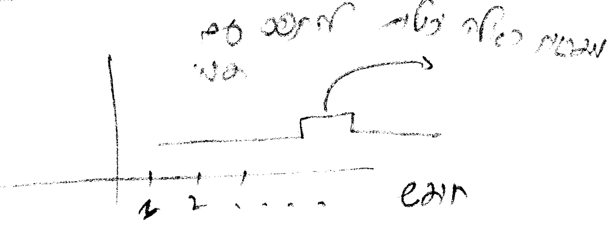
האם יש שקילות בין צ'או פשר לפי דמי ו-2 רגרסיות? כן ולא.

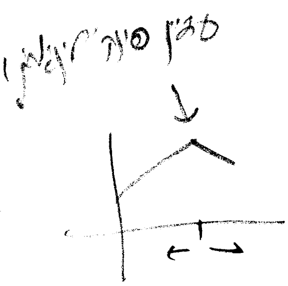
כן ← כי התוצאה תהיה זהה

לא ← כי יש הנחה סמויה ברגרסיה לפי דמי: ש σ^2_ε זהה ל-2 קבוצות האוכלוסיה (אין בעיה a שונות שונה ל-2 אוכלוסיות).

שימושים נוספים אפשריים למשתני דמי:

* נניח שאני רוצה לבדוק עליית מחירים → למבחן אינפלציה. בודקים מחירים של שנים. ← אם אני חושב שבק"פ זה עליה ← זה לא אומר שבהכרח היתה אינפלציה ← יכול להיות פשוט שבק"פ הביקוש עולה.





* שימוש נוסף: אומדה של התנהגות עם שבירה:

20/12 שיעור 6

ביטול ההנחות הקלאסיות

1. $E(\varepsilon_i)=0$
2. x לא סטוכסטי
3. $cov(\varepsilon_i,\varepsilon_j)=0$
4. $Var(\varepsilon_i)=\sigma_\varepsilon^2$
5. $\varepsilon_i \sim N$

ביטול הנחה 5

→ לא צריכים להנחה זו עבור גאוס-מרקוב $\varepsilon_i \sim N$

→ הנחה זו משמעותית רק במבחני ובמבחני רווח סמך.

כדי לומר ש- $\hat{\beta} \sim N$:

$$\hat{\beta} = \beta + \frac{\sum (x_i-\bar{x})\varepsilon_i}{\sum (x_i-\bar{x})^2}$$

$\hat{\beta}$ סכום של משתנים שמתפלגים נורמלית הם $\varepsilon_1,\varepsilon_2 \dots \varepsilon_n$ מתפלג נורמלית.

משפט הגבול המרכזי:

סדרה של משתנים מקריים שמקיימים אותה התפלגות בלתי תלויים ביניהם ⇐ הסכום שלהם מתפלג נורמלית.

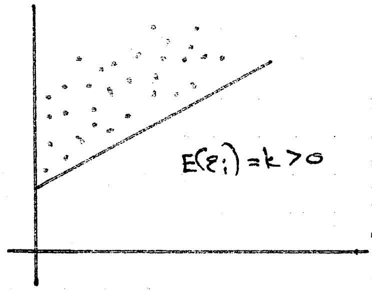
נניח שהנחה 5 לא מתקיימת אבל לכל ה-ε_i יש אותה התוחלת אותה השונות ⇐ הם בעלי אותה התפלגות!
התנאי השני שהם בלתי תלויים (שאותו נקבל מהנחה 3).
בגלל שאנחנו משתמשים רק בסכום $\sum(x_i-\bar{x})\varepsilon_i$ ניתן להשתמש במשפט הגבול המרכזי ויבטל לנו את הצורך בהנחה 5 (כי לא משתמשים אף פעם ב- ε_1, ε_2, ε_3 בנפרד אלא בסכום!).

→ ביטול הנחה קלאסית ראשונה: $E(\varepsilon_i)=0$

2 מקרים:
1) $E(\varepsilon_i)=K$
2) $E(\varepsilon_i)=\varepsilon_i$

טעות מכוונת של חוקר השדה! האומדים מוטים
(ואז גם השונות לא נכ' אבל העיקר האומדים מוטים, לא)
אכפת לי מהשונות

$E(\varepsilon_i)=k$



1) החותך שלי מוטה. ⇐

2) השיפוע שלו יצא ללא הט"ה

1) אפשר לפתור את הבעיה ע"י הזזת החותך ב-k.

2) $E(\varepsilon_i)=\varepsilon_i$

מודל אמיתי: $E(\varepsilon_i)=0$ $Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\varepsilon_i$

מודל נאמד:
(השמטנו משתנה מסביר) $Y_i=\beta_0+\beta_1X_{1i}+\varepsilon_i$

הכל "יכנס" ל-

$E(\varepsilon_i)=X_{2i}$ כל מה שמשפיע על Y, ולא התחשבנו בו.

השמטת משתנה רלוונטי ⇐ הטעות משתנה מתצפית לתצפית.

מפר את הנחה מס' 1.

(בגלל זה האומדים מוטים)

β_1 כעת חו"ז מוטה, למעט המקרה שבו (X_1,X_2) בלתי מתואמים! ואם נרצה לדעת את כיוון ההט"ה:
$cov(X_1,X_2)$ $cov(X_2,Y)$

לסיכום:

1) $E(\varepsilon_i)=k$ האומד לשיפוע לא מוטה והשונות מינימלית.

2) $E(\varepsilon_i)=\varepsilon_i$ האומד לשיפוע מוטה, השונות לא נכונה ⇐ אין מה לעשות עם ההערכה!

* יוצא דופן: $cov(X_1,X_2)=0$

ביטול הנחה קלאסית 2:

2 מקרים: 1) $cov(X_i,\varepsilon_i)=0$
2) $cov(X_i,\varepsilon_i)\neq0$

באומד עם חוסר הט"ה ועם שונות מינימלית.
בעולם האמיתי אין אומדים חסרי הט"ה.
יש הבדלה בין אומדים עקביים ללא עקביים.

האומד לשיפוע ברגרסיה
עם משתנה מסביר I

$$\hat\beta = \beta + \frac{\Sigma(x_i-\bar{x})\varepsilon_i}{\Sigma(x_i-\bar{x})^2}$$

$$E(\hat\beta) = \beta + E\left[\frac{\Sigma(x_i-\bar{x})\varepsilon_i}{\Sigma(x_i-\bar{x})^2}\right]$$ → לא ניתן להוציא את ה-E החוצה כי לא קבוע יותר...

$$= \beta + E\left[\frac{\overbrace{\Sigma(x_i-\bar{x})(\varepsilon_i-\bar{\varepsilon})}^{\widehat{cov}(x_i,\varepsilon_i)}}{(x_i-\bar{x})^2}\right] \xrightarrow{N\to\infty} = \beta + \frac{cov(x_i,\varepsilon_i)}{V(x_i)}$$

← בשביל שהאומד $\widehat{cov}$ יהיה שווה ל- cov : נפעל

תוחלת $\widehat{cov} = cov$ $E(\widehat{cov}) = cov$ אם $N\to\infty$ ⇐ עקיבות!!!

← $\hat\beta=\beta$ יקרה במצב היחידי כשיהיה לי נתונים על כל האוכלוסיה.

עקיבות ויעילות אסימפטוטית:

עקיבות: מקבילה חוסר הטיה. $E(\hat\theta)\xrightarrow[N\to\infty]{}\theta$

← האומד מוטה $E(\hat\theta)\neq\theta$

וההטייה קטנה עם מספר התצפיות.

← עבורנו $N\geq 30$, ההטייה נהיית שולית.

ולכן אומד עקיב הוא אומד חסר הטיה.

יעילות אסימפטוטית: שונות מינ' $Var(\hat\theta)\xrightarrow[N\to\infty]{}0$

---

1) אם $cov(x_i,\varepsilon_i)=0$:
$E(\hat\beta)\xrightarrow[N\to\infty]{}\beta$ ⇐ האומדים עקיבים! (חסרי הטייה)

2) אם $cov(x_i,\varepsilon_i)\neq 0$!
$E(\hat\beta)\xrightarrow[N\to\infty]{}\beta+\frac{Cov}{Var}$ ⇐ האומדים מוטים ולא עקיבים!

כאשר $cov(x,\varepsilon)=0$ האומדים הם יעילים אסימפטוטית:

במקור שונות: $Var(\hat\beta) = E(\hat\beta - \underbrace{E(\hat\beta)}_{\text{המאומד: }\beta})^2$

נעשה באופן אחר: $Var(\hat\beta) = E(\hat\beta-E(\hat\beta))^2 \xrightarrow[n\to\infty]{} E(\hat\beta-\beta)$

⊕ אי אפשר לעלות (או לבדוק) הכל כש $cov(x_i,\varepsilon_i)\neq 0$!

אי אפשר לבדוק או למדוד $cov(x_i,\varepsilon_i)=0$.

$$\varepsilon_i = \alpha + \beta x_i + u_i$$

אין נתונים על ε_i!

$$\hat{\varepsilon}_i = \alpha + \beta x_i + u_i$$

$cov(x_i,\hat{\varepsilon}_i)=0$ הוא [illegible] מתוך המשוואות הנורמליות!!

← נגיד שאנחנו אומדים את השכר באמצעות כל מה שניתן לחשוב עליו:

שכר השכלה ניסיון גיל

$$Y = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \varepsilon_i$$

אמביציה

פרט אמביציוני יגיע יותר השכלה. $\rightarrow cov(x_{1i},\varepsilon_i)\neq 0$

מוטה ולא עקיב $\hat{\beta}_1 \leftarrow$

השמטתי את המשתנה של האמביציה ← נכנס ל-ε_i

$\leftarrow cov(x_i,\varepsilon_i)\neq 0$

← אומדי תרומת שכר לנשואים מול שכר לא נשואים (נשוי):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_z X_{zi} + \varepsilon_i \qquad X_{zi} \begin{cases} 0 & \text{לא נשוי} \\ 1 & \text{נשוי} \end{cases}$$

האם הפרט נשוי

שאפתנות, רצינות, מחויבות...

$\hat{\beta}_z = 0.3$ אנשים נשואים מקבלים 30% יותר מאנשים לא נשואים.

$\beta_z = 0$ מגלה שנשואים יותר שאפתניים, מחויבים... ותכונות אלה משפיעות על אנשים נשואים.

$$\hat{\beta}_z = \beta_z + \frac{cov(x,\varepsilon)\,cov(y,\varepsilon)}{var(x)}$$

לסיכום:

$cov(x,\varepsilon_i)=0$ האומדים עקיבים, נמשיך רגיל.

$cov(x,\varepsilon_i)\neq 0$ האומד מוטה, לא ניתן לעשות כלום.