סיכום חזרה - סוף סמסטר: (בשילוב עם רמה של שאלות/פתרון משאלות ממבחנים)

(1) רגרסיה מרובת משתנים: k משתנים מסבירים

$$Y_i = \beta_0 + \beta_1 X_{1i} + \dots \beta_k + \varepsilon_i$$

β_0 ← התוחלת של Y כאשר כל ה-X-ים שווים לאפס.

β_j ← ההשפעה השולית של X_j על Y כאשר יתר המשתנים המסבירים מוחזקים קבועים.

למה לא רגרסיה פשוטה? אם מריצים $Y_i = \beta_0 + \beta_1 X_{ji} + \varepsilon_i$

ומשמיטים יתר משתנים מסבירים אחרים ← המתואמים עם X_j

אז: האומדים יהיו מוטים (הטיה)

תכונות ברגרסיה מרובה:

תנאי משוואות:

$\sum \hat{\varepsilon}_i = 0$ → טעויות מאופסות בממוצע

$\sum \hat{\varepsilon}_i X_{ji} = 0$ → השגיאות לא מתואמות עם כל אחד מה-X-ים.

במקרה של 2 משתנים מסבירים:

$$\hat{\beta}_1 = \frac{M_{22} M_{Y1} - M_{12} M_{Y2}}{M_{11} M_{22} - M_{12}^2} \qquad \hat{\beta}_2 = \frac{M_{11} M_{Y2} - M_{12} M_{Y1}}{M_{11} M_{22} - M_{12}^2}$$

אם אין מתאם בין X_1 ל-X_2 ← חוזרים ל-SRM

גם כאן ראינו ש:

$$\hat{\beta}_1 = \beta_1 + \sum \square \varepsilon_i$$

→ אומד חסר הטיה

בוחנת לשאלה ← 11.2.09 אומר את

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

טענה: אם המתאם בין X_1 ל Y_i במדגם $=0$, אזי $\hat{\beta}_1 = 0$

שאלה 5:

טענה A:

$$b_1 = \frac{\widehat{\text{cov}}(X_1,Y)\,\widehat{\text{var}}(X_2) - \widehat{\text{cov}}(X_2,Y)\,\widehat{\text{cov}}(X_1,X_2)}{\widehat{\text{var}}(X_1)\,\widehat{\text{var}}(X_2) - \left[\widehat{\text{cov}}(X_1,X_2)\right]^2}$$

. אם המתאם בין X_{1i} ל- Y_i במדגם שווה ל-0

נשארנו עם הביטוי: $b_1 = \dfrac{-\text{cov}(X_2,Y)\,\text{cov}(X_1,X_2)}{\text{var}(X_1)\,\text{var}(X_2) - \left[\text{cov}(X_1,X_2)\right]^2}$ **שאינו** שווה לאפס. האינטואיציה כאן שונה מאשר במצב שבו יש משתנה מסביר אחד בלבד כיוון שכאן אנחנו מחפשים את השונות של Y שמוסברת באמצעות 1X ושאינה מוסברת באמצעות 2X. לכן מתאם פשוט ששוה לאפס אינו אומר שהאומד להשפעה של 1X יצא גם כן אפס. **לכן טענה A אינה נכונה.**

טענה B:

מאותו שיקול שיתכן ונקבל אומד שאינו שווה לאפס, כך גם הפרמטר אינו בהכרח שווה לאפס בתודעת הגם שהמתאם הפשוט בין 1X לבין Y שווה לאפס. שוב --צריך לנקות מהמתאם את ההשפעה העקיפה של 1X דרך 2X. אם היא מתקזזת אז נקבל מתאם פשוט ששוה לאפס למרות שעדיין קיימת השפעה של 1X שאינה שווה לאפס (השפעה אמיתית. כלומר – פרמטר שונה מאפס). **לכן טענה B נכונה.**

תשובה ב. נכונה

השונות של $\hat{\beta}_j$:

$$Var(\hat{\beta}_j) = \frac{\sigma^2_\varepsilon}{\sum(x_{ij}-\bar{x}_j)^2} \cdot \frac{1}{1-R^2_{1,2}} \quad i \neq j$$

אם מוסיפים משתנים מסבירים ← השונות האמיתית תגדל

← הסיבה - שני כוחות:

$\hat{\sigma}^2_\varepsilon \downarrow \quad , \quad \frac{1}{1-R^2_{1,2}} \uparrow$

הוספה של משתנים מסבירים מקשה על אמידה של השפעה נפרדת.

ככל שהמתאם גבוה יותר ← מולטיקולינאריות עולה.

מולטיקולינאריות חלקית

דוגמא מבחן 7.10.08

$$\log(wage) = \alpha_0 + \alpha_1 Tenure + \alpha_2 Tenure^2 + \varepsilon_i$$

$$\log(wage) = \beta_0 + \beta_1 \log(Tenure) + \beta_2 \log(Tenure^2) + \varepsilon_i$$

מולטיקולינאריות מלאה!

במבחן

למבחני השערות: כשיש מולטיקולינאריות גבוהה (מעל $R^2_{1,2} = 0.75$) ←

נהוג להשתמש במבחן מובהקות הרגרסיה:

$H_0: \beta_1 = \dots = \beta_k = 0$ → בודק האם המודל הכולל מסביר את המשתנה התלוי.

לא כולל חותך

טעויות ספציפיקציה:

כל מה שאנחנו מסיקים מהאומדנים מתוך הנחה שהנוסחה של המודל
היא נכון ←

(א) השמטת משתנה רלוונטי:

מודל אמיתי ← $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

מודל נאמד ← $Y_i = \beta_0' + \beta_1' X_{1i} + \underbrace{\varepsilon_i'}_{\beta_2 X_{2i} + \varepsilon_i}$

ברמה התיאורטית (פרמטרים)

הראינו כי:

$$E(\hat{\beta}_1') = \beta_1 + \underbrace{\beta_2 \cdot \frac{M_{12}}{M_{11}}}_{\text{ההטיה (מעלה/מטה)}}$$

* אם $V_{1,2}$ לא מתואמים אז אין הטיה

כיוון ההטיה תלוי ב 2 גורמים:

1. הקשר בין X_2 לבין Y
2. הקשר בין X_2 לבין X_1
3. גם החותך יהיה מוטה.
4. השמטה יכולה לגרום גם הגדרה פונק' לא נכונה $x, x^2, \ldots$

ברמה הסטטיסטית (אומדים לפרמטרים ← הקשר במדגם)

דוגמא לשגיאה ממבחן (השאלה עם משתני דמי)

מבחן אוגוסט 3/08/08

שאלה 14:

בניסיון לחקור את הקשר בין השכלה לבין שכר. חוקר הריץ את הרגרסיות הבאות על אותו מדגם (להלן הפלטים):

רגרסיה 1:

Dependent Variable: LN_WAGE_HR

Variable	Coefficient	Std. Error	t-Statistic	Prob.
SCLYRS	0.06	0.00	102.70	0.00
C	2.57	0.01	318.01	0.00

רגרסיה 2:

Dependent Variable: LN_WAGE_HR

Variable	Coefficient	Std. Error	t-Statistic	Prob.
SCLYRS	0.07	0.00	108.79	0.00
GENDER	-0.26	0.00	-55.55	0.00
C	2.63	0.01	330.92	0.00

כאשר:

sclyrs -- שנות הלימוד של הפרט

Gender – מקבל את הערך 0 באם מדובר בגבר ומקבל את הערך 1 באם מדובר באישה

Ln_wage_hr – לוגריתם השכר לשעה

א. במדגם של החוקר נשים יותר משכילות מגברים

ב. במדגם של החוקר גברים יותר משכילים מנשים

ג. חסרים נתונים על מנת לקבוע מי הקבוצה בעלת ההשכלה הגבוהה יותר במדגם

ד. יתכן ומדד טיב ההתאמה גבוה יותר עבור הרגרסיה הראשונה

שאלה 14:

(1) $\ln w\_h = a_0 + a_1 \, sclyrs_i + e_{i1}$

(2) $\ln w\_h = b_0 + b_1 \, sclyrs_i + b_2 \, Gender + e_{i2}$

~~ידוע כי:~~

$$a_1 = b_1 + b_2 \cdot (c_2)$$

כאשר:

$$Gender = c_0 + c_2 \cdot sclyrs_i$$

$\frac{\hat{cov}(s,g)}{\hat{Var}(s)} \leftarrow$ קובע את כיוון המתאם בין Gender ל sclyrs

מכיוון ש $a_1 < b_1$

וגם כן b_2 שלילי

ת"ב להתקיים כי c_2 חיובי $\Leftarrow$ מתאם חיובי בין Gender (נשים) לבין sclyrs.

טעות ספציפיקציה – הכנסת משתנים לא רלוונטיים:

אמיתי $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$

מה שנאמד $Y_i = \beta_0' + \beta_1' X_{1i} + \beta_2' X_{2i} + \varepsilon_i'$

↓ הראינו כבר

$Var(\hat{\beta}_1') > Var(\hat{\beta}_1)$ ← שונות לא מינימלית (לא Blue)

← האומד עדיין חסר הטיה

← כל עוד אין מולטיקולינאריות מלאה

שאלה קשה מבחן אמצע 2007 ← משהו שלא עסקנו בו ← "טעות מדידה":

(סוג אחר של טעות ספציפיקציה)

הרבה פעמים אין לנו את המשתנים שאנו מעוניינים בהם (X , Y) אלא ערכים שהם המקבילה עם רעש:

לדוגמא: יש כל מיני מבחנים אינטליגנציה ← הבנה כללי

נסיון ← Proxy

Beta ← יש מניות הנאמדות משנים קודמות

הכנסה מדווחת , נתונים חשבונאיים ← לא מעידים לחלוטין , לא מדויקים לחלוטין

מה האם האומד מוטה?

לאיזה כיוון?

* טעות מדידה יכולה להיות במשתנה התלוי Y / בלתי X

אומד OLS

חסר הטיה / עקבי

השונות גדולה יותר. אבל אין מה לעשות

(1) $Y_i = \beta_0 + \beta_1 X_i^* + u_i$

לא נצפה

לא מתואמים

אנחנו רואים את X_i : $X_i = X_i^* + e_i$

ונניח ש: (II) $\boxed{Cov(e_i, X_i^*) = 0}$

$Cov(u_i, X_i^*) = 0$

$Cov(u_i, X_i) = 0$

$Cov(u_i, e_i) = 0$

ראשית נראה ש x (המשתנה הנצפה) ו e מתואמים:

$$Cov(X_i, e_i) = E(X_i e_i) = E(X_i^* e_i) + \underset{\uparrow\ \mathrm{Var}\, e_i}{E(e_i^2)} = 0 + \sigma^2_{e_i} \neq 0$$

וזה עושה לנו בעיות. האומד $\hat\beta_1$ יהיה מוטה

x^* מציבים

המודל שאומדים:

$$Y_i = \beta_0 + \beta_1 (X_i - e_i) + u_i$$

$$= \beta_0 + \beta_1 X_i + [u_i - \beta_1 e_i]$$

מתואמים

כיוון ההטיה:

$$\mathrm{plim}(\hat\beta_1) = \beta_1 + \frac{Cov(X_i, u_i - \beta_1 e_i)}{Var(X_i)}$$

$-\beta_1 \sigma_e^2$

(II) $Var\, X^* + Var(e)$

$$= \beta_1 - \frac{\beta_1 \sigma_e^2}{\sigma_{x^*}^2 + \sigma_e^2} = \beta_1 \left(1 - \frac{\sigma_e^2}{\sigma_{x^*}^2 + \sigma_e^2}\right) = \beta_1 \left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2}\right)$$

← לכן ההטיה לכיוון האפס.

אומד אמיתי מוטה/מוחלש ← "attenuated"