אקונומטריקה למדע המידע

שיעור 8

שיעור חזרה ועוד

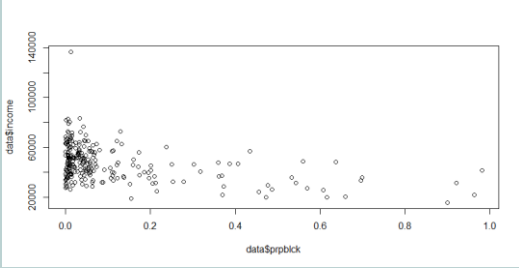
מבוא

- נשתמש בקובץ הנתונים DISCRIM. מדובר בנתונים ברמת המיקוד (zip code) על מחירים של פריטים שונים במסעדות אוכל מהיר, וכן, מאפיינים של האוכלוסייה באזורים אלה בפנסילבניה ובניו-ג'רזי.
- הרעיון הוא לבדוק האם מסעדות אוכל מהיר גובות מחירים גבוהים יותר באזורים שבהם ריכוז השחורים גבוה יותר.

סטטיסטיקה תיאורית

- אלו משתנים יש בקובץ?
- מי מהם יכולים להיות קשור למודל שאנחנו מנסים לבנות?

Income, proportion of black



- ננסה את המודל המסביר את מחיר הסודה, psoda, באמצעות הפרופורציה של השחורים באוכלוסייה וההכנסה החציונית:

$$psoda = \beta_0 + \beta_1 prpblck + \beta_2 income + \epsilon$$

- המודל שהתקבל:
- $psoda = 0.956 + 0.014 * prpblck + 0.0000016 income + u$

משמעות המודל

- משמעות המקדם של $prpblck$: המחיר של סודה יהיה גבוה באזור עם שחורים בלבד ב- 0.014 דולר מאזור ללא שחורים כלל
- או – עליה באחוז אחד של שחורים באזור מסוים תעלה את מחיר הסודה ב-0.0014 שזה 0.14 סנט

מובהקות כלכלית ומובהקות סטטיסטית

- התוצאה אולי מובהקת, אבל...

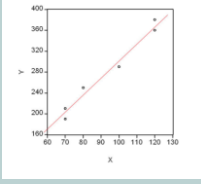
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.412e-01  2.143e-02  43.921  < 2e-16 ***
prpblck     1.481e-01  3.003e-02   4.932 1.33e-06 ***
income      1.865e-06  4.052e-07   4.602 6.14e-06 ***
```

מודלים לא ליניאריים

- **מודל ליניארי:**

$$y = \beta_0 + \beta_1 x + \epsilon$$

- המשמעות: הנגזרת- $\frac{\partial y}{\partial x}$ היא קבועה.
- אם מגדילים את x ביחידה אחת, בכמה יחידות משתנה y ?
- במודל הליניארי - שינוי זה הוא קבוע ושווה ל- β_1



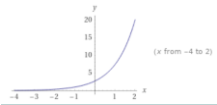
מודלים לא ליניאריים

- בניגוד למודל הליניארי, נציג שלושה המודלים אחרים (המודלים הלוגריתמים) המתארים קשרים שאינם ליניאריים בין x לבין y
- במודלים אלו השינוי השולי (השיפוע) **לא יהיה קבוע**, אלא תלוי במשתנים - x או y או בשניהם

מודל חצי לוגריתמי

- במודל החצי לוגריתמי הקשר בין x לבין y מתואר על ידי הפונקציה הבאה

$$y = e^{\alpha+\beta x+\epsilon}$$



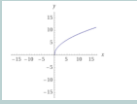
- כפי שניתן לראות, השינוי השולי איננו קבוע אלא תלוי ב-y
- השינוי השולי $\frac{\partial y}{\partial x} = \beta y$
- עליה ביחידה אחת של x , y ישתנה ב-$100 * \beta\%$
- השינוי באחוזים הוא קבוע למרות שהשינוי השולי איננו קבוע

המודל הלוגריתמי הכפול

- במודל הלוגריתמי הכפול הקשר בין בין x לבין y מתואר על ידי הפונקציה

$$y = e^{\alpha} x^{\beta} e^{\epsilon}$$

- כפי שניתן לראות, השינוי השולי איננו קבוע אלא תלוי ב x וגם y: $\frac{\partial y}{\partial x} = \beta \frac{y}{x}$
- משמעות ה-β במודל זה היא **הגמישות** $\beta = \frac{\partial y}{\partial x}\frac{x}{y} = \frac{\frac{\partial y}{y}}{\frac{\partial x}{x}}$
- משמעות הגמישות היא שינוי שולי באחוזים: אם מגדילים את x ב% אחד בכמה % ישתנה y
- במודלים אלו הגמישות היא קבועה למרות שהשינוי השולי איננו קבוע

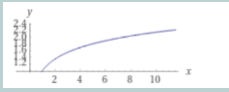


מודל לוג לינארי

- במודל הלוג-לינארי הקשר בין x לבין y מתואר על ידי הפונקציה הבאה:

$$e^{y} = e^{\alpha} x^{\beta} e^{\epsilon}$$

- כפי שניתן לראות, השינוי השולי איננו קבוע אלא תלוי ב-x: $\frac{\partial y}{\partial x} = \frac{\beta}{x}$
- במודל זה אם x עולה ב% אחד y עולה ב-β: $\beta = \frac{\partial y}{\frac{\partial x}{x}}$



טרנספורמציות של המודלים הלא ליניאריים לקו ישר

- בכדי שניתן יהיה לאמוד את המודלים הלא ליניאריים בשיטת OLS, עליהם לעבור טרנספורמציה לקו ישר.
- טרנספורמציה של המודלים לקו ישר תאפשר לתאר את הקשר בין המשתנה המסביר למשתנה המוסבר באופן לינארי.
- טרנספורמציה זו תתבצע על ידי הוצאת ($\ln$ לוג טבעי) משתי צדי המשוואה בכדי לבטל את ה-e

טרנספורמציות של המודלים הלא ליניאריים לקו ישר

- אם נתייחס למשתנה המסביר או המוסבר בתוספת הלוג, ניתן יהיה לתאר את הקשר ביניהם באופן ליניארי

המודל	לפני הטרנספורמציה	אחרי הטרנספורמציה
1) לוג-ליניארי	$e^{y}=e^{\alpha}\cdot X^{\beta}\cdot e^{u}$	$Y=\alpha+\beta\ln X+u$
2) לוגריתמי כפול	$Y=e^{\alpha}\cdot X^{\beta}\cdot e^{u}$	$\ln Y=\alpha+\beta\ln X+u$
3) חצי לוגריתמי	$Y=e^{\alpha+\beta k+u}$	$\ln Y=\alpha+\beta X+u$

תרגיל

- על מנת לאמד את התשואה להשכלה בישראל בשנים 1948-1990 נאמדו המודלים הבאים

$$MWAGE_t = 139.547 + 118.628 \cdot SCL_t$$

$$MWAGE_t = -1445.08 + 1239.60 \cdot LN(SCL)_t$$

$$LN(MWAGE)_t = 5.244 + 0.778 \cdot LN(SCL)_t$$

$$LN(MWAGE)_t = 6.292 + 0.070 \cdot SCL_t$$

- הסבירו את המשמעות של β בכל אחד מהמודלים

מודל לוג -לוג

- ייתכן כי מודל עם גמישות מחירים קבועה ביחס להכנסה מתאים כאן יותר

$$\log(psoda) = \beta_0 + \beta_1 prpblck + \beta_2 \log(income) + u$$

```
Call:
lm(formula = lpsoda ~ prpblck + lincome, data = data_1)

Coefficients:
(Intercept)      prpblck      lincome
   -0.96025      0.16022      0.09173
```

מה דעתנו על המודל

- התוצאות תואמות ההשערה שהחוקר התחיל איתה, אבל הניתוח די בסיסי. יכולים להיות גורמים אחרים המתואמים גם עם $income$ וגם $income$ שהם אחראים לתופעה הנצפית.
- לדוגמה, באזורים עניים יותר עשויה להיות צפיפות נמוכה יותר של מסעדות, מה שעלול לגרום למחירים גבוהים בגלל היעדר תחרות.

אי תלות בין הטעות למשתנה המסביר

- אחת מהנחות המודל החשובות ביותר: אי תלות בין ϵ לבין המשתנים המסבירים ברגרסיה
- כלומר, אם - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$
- אז - ϵ לא תלוי ב- x_1 וב- x_2
- מה יש בϵ ? כל הגורמים האחרים שמשפיעים על y אבל לא תלויים ב-x_1 או x_2

דוגמא

- קובץ הנתונים NBASAL מכיל נתונים על משכורות וסטטיסטיקות שונות עבור 269 שחקנים בהתאחדות הלאומית לכדורסל של ארה"ב (NBA- National Basketball Association).
- נאמוד את מודל שמסביר את מספר הנקודות במשחק (*points*) באמצעות שנים בליגה (*exper*)
- רגרסיה פשוטה: $points = \beta_0 + \beta_1 exper + \epsilon$

דוגמא

- רגרסיה פשוטה:

$$points = \beta_0 + \beta_1 exper + \epsilon$$

- **שאלה**: מה המשמעות של β_1?
- **תשובה**: אם נעלה את $exper$ ביחידה אחת **וכל שאר הגורמים יהיו קבועים** אז העלייה ב- $points$ תהיה של β_1
- מה יש ב-ϵ ?

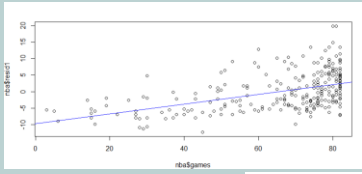
דוגמא

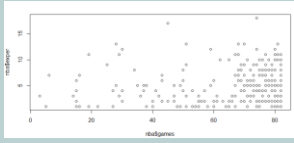
$\beta_2 games$

$$points = \beta_0 + \beta_1 exper + \epsilon$$

- מה יש ב-ϵ ?
- למשל: מספר המשחקים שהשחקן שיחק
- האם שנעלה את $exper$ ביחידה אחת (בממוצע) $games$ יישאר אותו דבר?

דוגמא

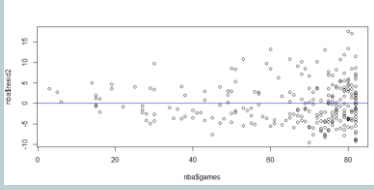




דוגמא

- מה עושים?
- מכניסים את $games$ לתוך הרגרסיה!

$$points = \beta_0 + \beta_1 exper + \beta_2 games + \epsilon$$



דוגמא

```
> model1nba

Call:
lm(formula = points ~ exper, data = nba)

Coefficients:
(Intercept)        exper
     8.5705       0.3301

> model2nba

Call:
lm(formula = points ~ exper + games, data = nba)

Coefficients:
(Intercept)        exper        games
    -0.7009       0.2062       0.1507
```

לסיכום

- אם ϵ ו- x ב"ת - אז העלייה ב x לא תשנה את ϵ וכך אנחנו "מצליחים" לשמור אותו "קבוע" ולבדוק רק את ההשפעה של x
- אם ϵ ו- x תלויים – אז עלייה ב x תשנה גם את ϵ – בלי שתהיה לנו יכולת לשלוט בזה! – ואז ההשפעה **הנצפית** של x לא תהיה אמתית
- מה אפשר לעשות? להעביר מ ϵ לרגרסיה את מה שתלוי ב- x

רגרסיה לינארית מרובה – המודל

- לכל תצפית $i=1,\ldots,n$ מתקיים:

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ji} + \varepsilon_i$$

- וכן לכל j,i , X_{ji}, ε_i ב"ת ומתקיים:

$$\varepsilon_i \sim N\left(0, \sigma^2\right)$$ ב"ת

- והתפלגות X_{ji} התפלגות כלשהי.
- אנחנו צופים <u>רק</u> ב- $X_{1i},\ldots,X_{ki},Y_i$ אך <u>אין</u> יכולת לצפות ב- ε_i

רגרסיה לינארית מרובה – פירוט הנחות המודל

- נהוג לפרט את הנחות המודל באופן הבא:
  1. המודל לינארי בפרמטרים
  2. המדגם מקרי
  3. אין תלות לינארית מלאה בין המשתנים המסבירים
  4. הנחת התוחלת המותנה: $E\left(\varepsilon \mid X_1,\ldots,X_k\right)=0$
  5. הומוסקדסטיות: $\operatorname{var}\left(\varepsilon \mid X_1,\ldots,X_k\right)=\sigma^2$
  6. הנחת הנורמליות: $\varepsilon \sim N$

הנחת הלינאריות

- על המודל להיות לינארי בפרמטרים, אך אין פירוש הדבר שהמודל בהכרח לינארי במשתנים.
- נתבונן למשל במודל הבא המתאר קשר בין כמות ההון וכמות העבודה, והתוצר השנתי ע"י פונקציית קוב-דוגלס:

$$Y = AK^{\alpha}L^{\beta}e^{\varepsilon}$$

- לכאורה, המודל אינו לינארי בפרמטרים A,α,β אך ע"י טרנספורמציית לוג נקבל את המודל:

הנחת הלינאריות

$$\log Y = \underbrace{\log A}_{\beta_0} + \underbrace{\alpha}_{\beta_1} \log K + \underbrace{\beta}_{\beta_2} \log L + \varepsilon$$

- קיבלנו מודל לינארי בפרמטרים, אך לא לינארי במשתנים.
- משמעות הדבר היא, שבבואנו לבדוק את הקשר בין תשומות הון ועבודה לתוצר, תחת הנחת פונקציית ייצור קוב-דוגלס, עלינו לבצע קודם טרנספורמציית לוג על הנתונים, אחרת נקבל אומדים חסרי משמעות.

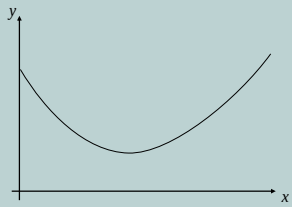
מודל ריבועי

- אחת התוצאות של מודל לינארי, היא השפעה שולית קבועה של המשתנים.
- אך לעיתים, ידוע כי ההשפעה השולית אינה קבועה אלא עולה או יורדת בערכו של המשתנה.
- במקרים כאלה נוכל לשקול את המודל הבא:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

מודל זה מתאר קשר ריבועי בין X ו- Y.

מודל ריבועי



- בגרף מתוארת השפעה שולית עולה.
- בתחילה ההשפעה שלילית, אך היא הולכת ועולה ונהיית לבסוף חיובית.
- במקרה הריבועי ההשפעה השולית היא:

$$\frac{\partial Y}{\partial X} = \beta_1 + 2\beta_2 X$$

- והיא עולה (יורדת) אם β_2 חיובי (שלילי).

מודל ריבועי

- שימו לב שהמודל הנ"ל מקיים את הנחת הלינאריות, כאשר X הוא המשתנה הראשון ו- X^2 הוא המשתנה השני.
- שימו לב שהנחה מס' 3 לא מופרת שכן

$$\left|Corr\left(X, X^2\right)\right| \neq 1$$

דוגמא

- בהמשך לנתונים של ה-NBA נאמוד מודל שמסביר את מספר הנקודות במשחק (*points*) באמצעות שנים בליגה (*exper*), גיל (*age*), ושנות משחק בקולג' (*coll*). בנוסף, נכניסאת המשתנה *exper* בריבוע.
- כאשר מחזיקים את הגיל (*age*), ואת שנות המשחק בקולג' (*coll*) קבוע, אחרי כמה שנים בליגה (או: עבור איזה ערך של *exper*) שנה נוספת למעשה מקטינה את מספר הנקודות במשחק? האם זה הגיוני?

דוגמא

```
> model3nba

Call:
lm(formula = points ~ exper + age + coll + exper2, data = nba)

Coefficients:
(Intercept)        exper          age         coll       exper2
   34.98355      2.33702     -1.06174     -1.27589     -0.07577
```

משתני דמי

- משתנה דמי הוא משתנה המקבל את הערכים 0 ו-1 בלבד.
- משמעות הערך 1 היא קיום תכונה מסוימת והערך 0 מסמן אי-קיום התכונה.
- לדוגמה, המודל המתאר את הקשר בין שכר להשכלה

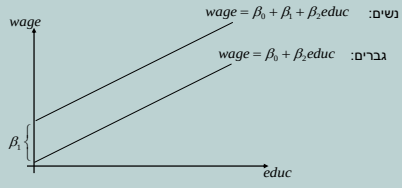
$$wage = \beta_0 + \beta_1 F + \beta_2 educ + \varepsilon$$

- כאשר:

$$F = \begin{cases} 1 & \text{if } i \text{ is female} \\ 0 & \text{if } i \text{ is male} \end{cases}$$

משתני דמי - דוגמה

- במקרה זה, נקבל שני מודלים שונים, אחד עבור גברים, ואחד עבור נשים:
  - גברים: $wage = \beta_0 + \cancel{\beta_1 \cdot 0} + \beta_2 educ + \varepsilon$
  - נשים: $wage = (\beta_0 + \beta_1 \cdot 1) + \beta_2 educ + \varepsilon$



משתני דמי - דוגמה

- נשים לב שבמודל הנ"ל קיימת הנחה שההבדל בין גברים לנשים הוא רק בחותך, ואילו השיפוע בשני המינים זהה.
- משמעות הדבר היא הבדל קבוע בשכר גברים ונשים ללא תלות בהשכלה.
- כמו כן, ע"פ המודל, תוספת השכר בגין כל שנת לימוד נוספת <u>זהה</u> בין גברים לנשים.
- הנחה זו עלולה להיות לא נכונה וכדי לאפשר הבדל גם בשיפוע, עלינו להוסיף משתנה נוסף.

משתני דמי - אינטראקציה

- ע"מ לאפשר הבדל גם בשיפוע עלינו להוסיף משתנה אינטראקציה והמודל יהיה:

$$wage = \beta_0 + \beta_1 F + \beta_2 educ + \beta_3 F \cdot educ + \varepsilon$$

- כך ש:
  - עבור גברים:

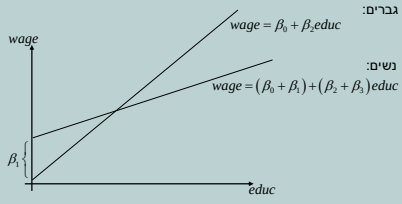
$$wage = \beta_0 + \cancel{\beta_1 \cdot 0} + \beta_2 educ + \cancel{\beta_3 \cdot 0 \cdot educ} + \varepsilon$$

  - ועבור נשים:

$$wage = (\beta_0 + \beta_1 \cdot 1) + (\beta_2 + \beta_3 \cdot 1) \cdot educ + \varepsilon$$

משתני דמי - אינטראקציה

- במקרה זה, נקבל שני קוים בעלי חותך <u>ושיפוע</u> שונים:



- ציור זה מתייחס למקרה שבו $\beta_1 > 0, \beta_3 < 0$.

משתני דמי - אינטראקציה

- שימו לב שתמיד נוכל לבדוק את נחיצות האינטראקציה ע"י בדיקת ההשערה: $H_0 : \beta_3 = 0$
- וכדי לבדוק האם קיים הבדל **כלשהו** בין גברים לנשים במשוואת השכר, נבדוק את ההשערה המורכבת:

$$H_0 : \beta_1 = \beta_3 = 0$$

דוגמא

$$\widehat{sat} = \underset{(6.29)}{1{,}028.1} + \underset{(3.83)}{19.3}hsize - \underset{(.53)}{2.19}hsize^2 - \underset{(4.29)}{45.09}\,female - \underset{(12.71)}{169.81}black + \underset{(18.15)}{62.31}\,female \cdot black$$

$$n = 1{,}191,\ R^2 = .0493$$

המשתנה sat הוא ציון ה-SAT, $hsize$ גודל הכיתה של התלמיד בשנה האחרונה של התיכון (במאות), $female$ הוא משתנה דמי למגדר ו- $black$ הוא משתנה דמי לגזע השווה ל- 1 עבור שחורים ול- 0 אחרת.

(1) כאשר $hsize$ מוחזק קבוע, מהו ההבדל הנאמד בין ציוני ה- SAT של נשים לא-שחורות לזה של גברים לא-שחורים? עד כמה מובהק הבדל זה מבחינה סטטיסטית?

(2) מהו ההבדל הנאמד בין ציוני ה- SAT של גברים לא-שחורים לזה של גברים שחורים? בדקו את השערת האפס שאין הבדל בין הציונים שלהם כנגד ההשערה האלטרנטיבית שיש הבדל.

(3) מהו ההבדל הנאמד בין ציוני ה- SAT של נשים לא-שחורות לזה של נשים שחורות? מה אתם צריכים כדי לבדוק אם הבדל זה מובהק מבחינה סטטיסטית?

דוגמא

- הרגרסיה הבאה התקבלה באמצעות OLS תוך שימוש בנתונים מתוך קובץ GPA2RAW:

$$\widehat{sat} = \underset{(6.29)}{1{,}028.1} + \underset{(3.83)}{19.3}hsize - \underset{(.53)}{2.19}hsize^2 - \underset{(4.29)}{45.09}\,female - \underset{(12.71)}{169.81}black + \underset{(18.15)}{62.31}\,female \cdot black$$

$$n = 1{,}191,\ R^2 = .0493$$

- המשתנה sat הוא ציון ה- SAT , hsize הוא גודל הכיתה של התלמיד בשנה האחרונה של התיכון (במאות) , female הוא משתנה דמי למגדר ו-black הוא משתנה דמי לגזע השווה ל-1 עבור שחורים ו-0 אחרת

דוגמא

$$\widehat{sat} = 1{,}028.1 + 19.3hsize - 2.19hsize^2 - 45.09female - 169.81black + 62.31female \cdot black$$
$$(6.29) \quad (3.83) \quad (.53) \quad (4.29) \quad (12.71) \quad (18.15)$$
$$n = 1{,}191, R^2 = .0493$$

- כאשר hsize מוחזק קבוע , מהו ההבדל הנאמד בין ציוני ה-SAT של נשים לא שחורות לזה של גברים לא שחורים? עד כמה מובהק הבדל זה מבחינה סטטיסטית?
- **<u>תשובה</u>**:
- נשים לא שחורות: $female = 1\ black = 0$
- גברים לא שחורים: $female = 0\ black = 0$
- כלומר, הפער הזה נמדד על ידי המקדם של $female$ שהוא - 45.09
- ה- t statistic שווה ל- 10.51- שזה כמובן מובהק

דוגמא

$$\widehat{sat} = 1{,}028.1 + 19.3hsize - 2.19hsize^2 - 45.09female - 169.81black + 62.31female \cdot black$$
$$(6.29) \quad (3.83) \quad (.53) \quad (4.29) \quad (12.71) \quad (18.15)$$
$$n = 1{,}191, R^2 = .0493$$

- מהו ההבדל הנאמד בין ציוני ה-SAT של גברים לא שחורים לזה של גברים שחורים? בדקו את השערת האפס שאין הבדל בין הציונים שלהם כנגד ההשערה האלטרנטיבית שיש הבדל
- **<u>תשובה</u>**:
- גברים לא שחורים: $female = 0\ black = 0$
  גברים שחורים $female = 0\ black = 1$
- כלומר, הפער הזה נמדד על ידי המקדם של $black$ – כמעט 170 נקודות פער בציון
- ה- t statistic גדול מ-13, כלומר, מובהק

דוגמא

$$\widehat{sat} = 1{,}028.1 + 19.3hsize - 2.19hsize^2 - 45.09female - 169.81black + 62.31female \cdot black$$
$$(6.29) \quad (3.83) \quad (.53) \quad (4.29) \quad (12.71) \quad (18.15)$$
$$n = 1{,}191, R^2 = .0493$$

- מהו ההבדל הנאמד בין ציוני ה-SAT של נשים לא שחורות לזה של נשים שחורות? מה אתם צריכים כדי לבדוק אם הבדל זה מובהק מבחינה סטטיסטית?
- **<u>תשובה</u>**:
- נשים לא שחורות $female = 1\ black = 0$
- נשים שחורות $female = 1\ black = 1$
- כלומר הפער הזה נמדד על ידי סכום המקדמים של $black$ ו - $\text{female} * black$ שהוא: −107.50 = 62.31 + −169.81
- מהנתונים שיש לנו, והצורה הספיצפית הזו של הרגרסיה – לא ניתן לחשב את הT סטטיסטי

משתנים קטגוריאלים

- לעיתים קיים משתנה שמי, שעשוי לקבל יותר משני ערכים אפשריים כמו במקרה של משתנה דמי.
- במקרה זה, נצטרך "לקודד" את ערכי המשתנה בעזרת כמה משתני דמי.
- לדוגמה, נניח שאין לנו נתונים על שנות הלימוד, אלה רק על סוג ההשכלה שיכולה להיות, יסודית, תיכונית או אקדמית.
- במקרה זה ניתן ליצור 3 משתני דמי, אחד עבור כל סוג השכלה: E, H ו- A המקבלים את הערך 1 עבור השכלה יסודית, תיכונית ואקדמית בהתאמה.

משתנים קטגוריאלים - דוגמה

- במידה ונכניס את כל שלושת המשתנים לרגרסיה, תהיה הפרה של הנחה מס' 3 שכן תמיד מתקיים:

$$E + H + A = 1$$

- כלומר, משתנה אחד, הוא פונקציה לינארית של השניים האחרים.
- לכן, נבחר שני משתנים בלבד מתוך השלושה ונבנה את המודל:

$$wage = \beta_0 + \beta_1 H + \beta_2 A + \varepsilon$$

משתנים קטגוריאלים - דוגמה

- במקרה זה נקבל ש:
  - עבור בעלי השכלה יסודית:

$$wage = \beta_0 + \varepsilon$$

  - עבור בעלי השכלה תיכונית:

$$wage = \beta_0 + \beta_1 + \varepsilon$$

  - עבור בעלי השכלה אקדמית:

$$wage = \beta_0 + \beta_2 + \varepsilon$$

- כלומר, אפשרנו מודל שונה לכל קבוצת אוכלוסייה.

משתנים קטגוריאלים - דוגמה

- משמעות הבחירה במודל זה היא שהשכלה יסודית מהווה את קבוצת הבסיס וכל יתר הקבוצות נמדדות ביחס אליה.
- כך, β_1 הוא ההבדל בין השכלה תיכונית ליסודית.
- β_2 הוא ההבדל בין השכלה אקדמית להשכלה תיכונית.
- שימו לב כי $\beta_2 - \beta_1$ הוא ההבדל בין השכלה אקדמית לתיכונית.

הנחת אי תלות לינארית מלאה בין המשתנים המסבירים

- כזכור, משמעות המקדם של משתנה מסויים היא השינוי הצפוי ב-Y כאשר הוא עולה ביח' וכאשר כל שאר המשתנים המסבירים ללא שינוי.
- במידה ובנתונים קיימת תלות לינארית מלאה, כלומר ניתן לבטא משתנה אחד כקומבינציה לינארית של היתר, לא ניתן לאמוד את המקדם הנ"ל לאור העובדה שהתופעה פשוט לא קורית!

הנחת אי תלות לינארית מלאה בין המשתנים המסבירים

- נשים לב לנוסחה שהוצגה בשיעור הקודם:

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} e_{(j)i}\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} e_{(j)i}^{\ 2}} = \frac{\operatorname{cov}\left(e_{(j)}, Y\right)}{\operatorname{var}\left(e_{(j)i}\right)}$$

- כאשר $e_{(j)}$ הן השאריות מהרגרסיה של X_j על יתר המשתנים המסבירים.
- במידה והנחה 3 מופרת, המכנה פשוט שווה ל-0 והביטוי אינו מוגדר.

מולטי-קולינאריות

- לעיתים, למרות שהנחה 3 מתקיימת, קיים קשר לינארי חזק, אך לא מלא, בין המשתנים המסבירים.
- תופעה זו נקראת מולטי-קולינאריות והיא גוררת לעיתים בעיות חישוביות.
- בעיה נוספת שמתעוררת היא בטיב האמידה, קרי בשונות האומדים.

מולטי-קולינאריות

- כזכור, שונות האומדים היא:

$$\frac{\sigma^2}{n\,\mathrm{var}(X)\left(1-R_{(j)}^2\right)}$$

- כאשר קיים קשר לינארי חזק בין המשתנים המסבירים, $R_{(j)}^2$ קרוב מאוד ל-1 ולכן השונות המתקבלת גבוהה מאוד מה שמקשה מאוד על הסקת מסקנות מהאמידה.
- פתרון אפשרי יהיה בד"כ הגדלת גודל המדגם.

דוגמא

- נשקול להכניס גם את

```
Call:
lm(formula = lpsoda ~ prpblck + lincome + prppov, data = data_1)

Coefficients:
(Intercept)      prpblck      lincome       prppov
    -1.5689       0.1093       0.1467       0.3501
```

מולטיקולינאריות

- חשבו את המתאם הליניארי (מקדם המתאם של פירסון) בין log(income) לבין prppov. האם זה פחות או יותר מה שציפינו?

```
> cor(data_1$prppov,data_1$lincome)
[1] -0.8468176
```

הנחת התוחלת המותנה

- הנחה מס' 4, $E(\varepsilon \mid X_1,\ldots,X_k)=0$ היא הנחה קריטית לתכונת חוסר ההטיה של אומדי OLS.
- במידה וההנחה מופרת, לא נוכל לומר שהאומדים חסרי הטיה ולכן גם לא נוכל להשתמש בסטטיסטים T, ו-F לצורך בדיקת השערות.
- בהמשך נתאר סיבות אפשריות להפרת ההנחה וניתן דרך אפשרית להתמודדות עם הבעיה.

השמטת משתנה

- הפרת של הנחה 4 עלולה להיגרם עקב השמטת משתנה מהמודל כפי שיודגם להלן.
- נניח כי המודל האמתי, המקיים את הנחות הרגרסיה הוא:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\varepsilon$$

- אך מסיבה מסוימת אין לנו נתונים על X_2 ולכן אנחנו אומדים את המודל:

$$Y=\beta_0+\beta_1X_1+u$$

- כאשר:

$$u=\beta_2X_2+\varepsilon$$

השמטת משתנה

- המשמעות של "העברת" X_2 לטעות, היא חוסר היכולת "להחזיק" אותו קבוע כאשר X_1 משתנה.
- אם נניח כי X_1 ו- X_2 מקיימים את הקשר הבא:

$$\overset{\alpha_1}{X_2} = \alpha_0 + \alpha_1 \overset{1}{X_1} + v$$

- נקבל שההשפעה הנצפית של X_1 היא:

$$\overset{\beta_1+\alpha_1\beta_2}{Y} = \beta_0 + \beta_1 \overset{1}{X_1} + \underbrace{\beta_2 \overset{\alpha_1}{X_2} + \varepsilon}_{u}$$

השמטת משתנה

- המסקנה, השמטת עלולה לגרום להפרה של הנחה מס' 4 ולכן להטיית האומד.
- החדשות הטובות הן שבמקרים מסוימים נוכל להעריך את כיוון ההטיה.
- במקרה שתואר בשקופית הקודמת, ההטיה הייתה: $\alpha_1\beta_2$
- בהתאם לסימן הפרמטרים נוכל להעריך האם ההטיה כלפי מעלה או כלפי מטה.
- שימו לב, השמטת משתנה לא תסב שום נזק במידה ואחד מהפרמטרים הנ"ל שווים ל-0.

סימולטאניות

- סיבה אפשרית נוספת להפרת הנחה מס' 4 היא סימולטאניות.
- במצב של סימולטאניות, ערכם של X ו-Y נקבע בו זמנית ע"י שתי משוואות/מודלים:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$X = \alpha_0 + \alpha_1 Y + u$$

סימולטאניות - דוגמה

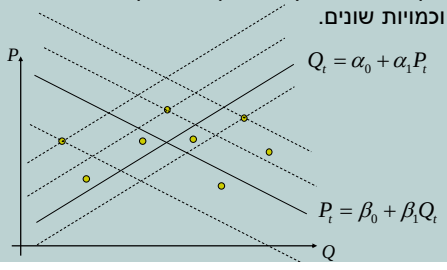
- דוגמה קלאסית לסימולטאניות היא הקשר בין מחירי מוצרים (P) והכמות הנמכרת שלהם (Q).
- ע"פ התיאוריה הכלכלית, המחיר והכמות נקבעים לפי ביקוש והיצע.
- פירוש הדבר הוא שהמחיר והכמות של מוצר מסוים בתקופה מסוימת נקבעים ע"י פתרון המשוואות:

ביקוש: $P_t = \beta_0 + \beta_1 Q_t + \varepsilon_t$

היצע: $Q_t = \alpha_0 + \alpha_1 P_t + u_t$

סימולטאניות - דוגמה

- בכל תקופה t, המשתנים ε_t, u_t מקבלים ערכים מקיים שונים ולכן בכל תקופה מתקבלים מחירים וכמויות שונים.



סימולטאניות - דוגמה

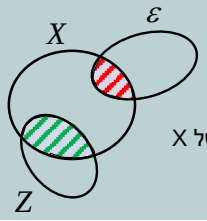
- מהאיור בשקופית הקודמת ברור כי לא נוכל להסיק מסקנות, לא על משוואת הביקוש, ולא על משוואת ההיצע מהנתונים.
- הסיבה לכך היא מתאם בין הטעות למשתנה המסביר.
- כך למשל, כמות גבוהה עשויה לרמז על כך שהטעות במשוואת הביקוש הייתה גבוהה וכד'.
- וזוהי כמובן הפרה של הנחה מס' 4.

משתני עזר

- דרך אחת להתמודדות עם אי-קיום הנחה מס' 4 היא שימוש במשתני עזר (IV).
- הבעיה כאמור היא קיום מתאם בין אחד המשתנים המסבירים לטעות. משתנה כזה נקרא משתנה אנדוגני.
- משתנה שלא מתואם עם הטעות נקרא משתנה אקסוגני.

משתני עזר

- משתנה עזר (Z) למשתנה אנדוגני (X) הוא משתנה המקיים את שני התנאים להלן:
  - $\operatorname{cov}(Z,\varepsilon)=0$
  - $\operatorname{cov}(Z,X)\neq 0$
- נמחיש ע"י תרשים:
- נעזר ב-Z כדי "לצפות" בחלק של X שאינו מתואם עם הטעות.



משתני עזר - דוגמה

- במשוואת שכר כפונקציה של השכלה.
- השכלה עשויה להיות מתואמת עם "כשרון" המושמט מהרגרסיה.
- משתני עזר אפשריים הם:
  - השכלת ההורים.
  - מס' אחים.
  - (בארה"ב) רבעון שבו נולד הפרט.

TSLS

- אמידה בעזרת משתנה עזר נעשית בשני שלבים:
- בשלב הראשון אומדים את הרגרסיה:

$$X = \alpha_0 + \alpha_1 Z + u$$

- ומחשבים בעזרתה את התחזית של X באמצעות Z:

$$\hat{X} = \hat{\alpha}_0 + \hat{\alpha}_1 Z$$

TSLS

- בשלב השני אומדים את הרגרסיה:

$$Y = \beta_0 + \beta_1 \hat{X} + \varepsilon$$

- שימו לב שמובטח לנו ש:

$$\operatorname{cov}\left(\hat{X}, \varepsilon\right) = 0$$

- כיוון ש- $\hat{X}$ הוא טרנספורמציה לינארית של Z.

TSLS

- אומדי TSLS אינם חסרי הטיה אך הם עקיבים, כלומר מתקיים:

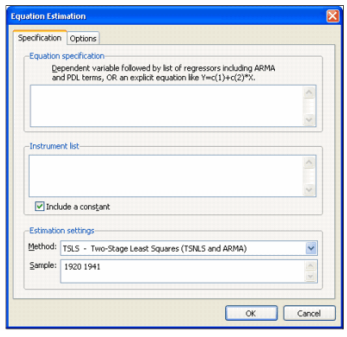
$$\hat{\beta}_1^{IV} \xrightarrow{n\to\infty} \beta_1$$

- ולכן שיטה זו ישימה עבור מדגמים גדולים בלבד.
- שונות האומד במקרה הפשוט תהיה:

$$\frac{\sigma^2}{n \operatorname{var}\left(X\right) R_{X,Z}^2}$$

- כלומר, ככל ש-X ו-Z מתואמים יותר כך שונות האומד תהיה קטנה יותר.

TSLS

- ניתן להשתמש בשיטה זו גם עבור מספר גדול יותר של משתני עזר.
- במידה והמודל כולל משתנים נוספים אקסוגניים, יש להשתמש גם בהם בשלב הראשון של TSLS.

TSLS ב-EViews