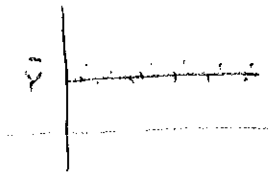
שיעור מס' 7 :

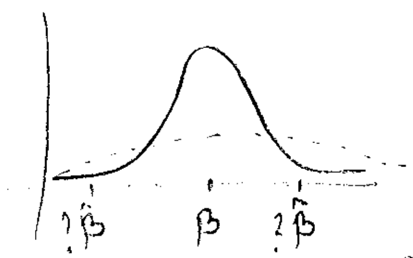
חזרה על R^2. מה קורה כאשר מוסיפים משתנה רק עם חותך?

* $R^2 = 0$,



הסקה סטטיסטית - בעיקר השערות: (המטרה של מודל הרגרסיה)

* הרצנו רגרסיה וקיבלנו אומדים: $\hat{\alpha}, \hat{\beta}$.
* $\hat{\beta}$ - מספר. אני לא יודע אם זה מה רוצה. $\hat{\beta}$ מתקבל מתוך התפלגות מסוימת מסביב ל-β.



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

אם ההתפלגות שטוחה ← יש סיכוי גבוה שאני רחוק מהתוחלת.

* בעולם קיבלנו אומד נקודתי.
* מה שאנחנו צריכים לדעת כדי להסיק מסקנות לגבי β:

① האומד לסטיית התקן של האומד ← $\hat{\sigma}_{\hat{\beta}}$

② מה צורת ההתפלגות של $\hat{\beta}$ (נקבל את זה מההנחות הקלאסיות על ε_i.) **

* המטרה שלנו: לומר ש $Prob(\text{משהו} - \hat{\beta} < \beta < \text{משהו} + \hat{\beta}) = 1-\alpha$, $0 \le \alpha \le 1$

אף פעם לא נוכל לומר משהו מבוטח לחלוטין אז בביטחון של $1-\alpha$.
אומרים שבהסתברות מסוימת β נמצא בטווח מסוים.

דוגמא פשוטה:

המקרה של $\bar{Y}$. הראינו בשבוע הראשון ש:

* $\sigma^2_{\bar{Y}} = \frac{\sigma^2_Y}{n}$ ← השונות של $\bar{Y}$.

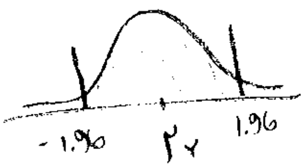
* $\bar{Y}$ אומד חסר הטיה ל-μ_Y

$1 = Prob(-\infty < Y < +\infty)$

$\bar{Y} \sim AN\left(\mu_Y, \frac{\sigma^2_Y}{n}\right)$ → בעלת התפלגות ידועה

בשביל מבחני השערות אני צריך טבלה סטנדרטית שיהיה אפשר להשתמש בה תמיד. לכן אני עושה טרנספורמציה לאומד:

$$Z = \underbrace{\frac{\bar{Y} - \mu_Y}{\sigma_{\bar{Y}}}}_{\text{מרחק במונחי סטיות תקן.}} \sim N(0, 1)$$ → עכשיו אפשר להשתמש בטבלה של התפלגות נורמלית סטנדרטית.



$$Prob\left[-1.96 < \frac{\bar{Y} - \mu_Y}{\sigma_{\bar{Y}}} < 1.96\right] = 1 - 0.05 = 0.95$$

$-1.96\sigma_{\bar{Y}} - \bar{Y} < -\mu < 1.96\sigma_{\bar{Y}} - \bar{Y}$

$\bar{Y} + 1.96\sigma_{\bar{Y}} > \mu > \bar{Y} - 1.96\sigma_{\bar{Y}}$

$$Prob\left[\bar{Y} - 1.96\sigma_{\bar{Y}} \le \mu_Y \le \bar{Y} + 1.96\sigma_{\bar{Y}}\right] = \underbrace{0.95}_{\text{ביטחון גבוה כמה שאפשר}}$$

הטווח הוא קטן כמה שאפשר

<u>למה?</u>

<u>שים לב:</u>

$\sigma_{\bar{Y}} = \sqrt{\frac{\sigma^2_Y}{n}}$ תלוי ב-n ← n גדל, השונות קטנה

* יש תמיד מתח בין הרצון לתת מובהקות הסתברות גבוהה לעומת טווח קטן כמה שיותר.

* במציאות אין לנו את $\sigma_{\bar{Y}}$ אלא את $\hat{\sigma}_{\bar{Y}}$ ← לכן הטרנספורמציה היא

$$\frac{\bar{Y} - \mu_Y}{\hat{\sigma}_{\bar{Y}}} \sim t$$

נחזור לעולם של הרגרסיה:

$$\hat\beta = \beta + \frac{\sum (X_i - \bar X)\varepsilon_i}{\sum (X_i - \bar X)^2} \longrightarrow \hat\beta \sim N \leftarrow$$

צירוף ליניארי של שגיאות שמתפלגים נורמלית

$$\frac{\hat\beta - E(\beta)}{\sigma_{\hat\beta}} \sim N(0,1)$$

ואם משתמשים באומד לסטיית התקן $\rightarrow$

$$\frac{\hat\beta - \beta}{\hat\sigma_{\hat\beta}} \sim t_{N-K}$$

$K \downarrow$ מספר הפרמטרים במודל (מספר המשתנים המסבירים + חותך)

להראות טבלת t, $n \to \infty$ ההתפלגות מאוד דומה ל-N.

$$Prob\left(-t_{crit} < \frac{\hat\beta - \beta}{\hat\sigma_{\hat\beta}} < +t_{crit}\right) = 1 - \alpha$$

t_{crit} – נקבע לפי α

α – סיכוי לטעות מסוג ראשון

$$Prob\left(\hat\beta - \hat\sigma_{\hat\beta} < \beta < \hat\beta + t_{\frac{\alpha}{2}} \cdot \hat\sigma_{\hat\beta}\right) = 0.95$$

$$Prob\left(\hat\alpha - t_{\frac{\alpha}{2}} \hat\sigma_{\hat\alpha} < \alpha < \hat\alpha + t_{\frac{\alpha}{2}} \cdot \hat\sigma_{\hat\alpha}\right)$$

$\leftarrow$ רווח סמך

מבחני t $\leftarrow$ השערות וכו'.

נניח אני רוצה לשאול שאלה ספציפית, האם סביר ש-$\beta = 0$? $\beta =$ מספר מסוים?

מחשבים t סטטיסטי:

$$t_{stat} = \frac{\hat\beta - \beta^{H_0}}{\hat\sigma_\beta}$$

$\leftarrow$ השערה ספציפית על הפרמטר

אם $t_{crit} < t_{stat}$ $\leftarrow$ דוחים את השערת האפס. לא סביר ש-$\beta = \beta^{H_0}$.

$H_0: \beta = \_$

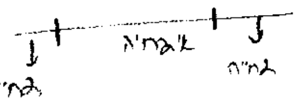
$H_1: \beta \neq \_$

* משמעות כלכלית.
* להראות בתוכנה.
* המבחן הזה שקול לבדיקה בעזרת רו"ס: מסתכלים אם β^{H_0} נמצא בטווח.

בעצם יש 2 דרכים לגשת לענין:

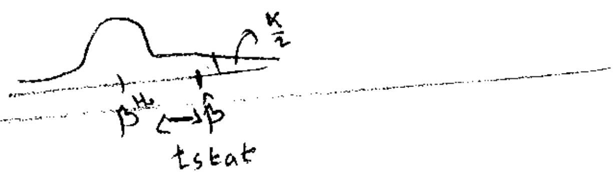
(א) קיבלנו אומד מסוים ← נבנה סביבו את הרווח $\hat\beta \mp t\cdot\hat\sigma_{\hat\beta}$

אם β_{H_0} לא ברווח "אי דחיה" ← נדחה את H_0.



α ← הסתברות לטעות מסוג ראשון ← הסיכוי לדחות את H_0 בהינתן שהיא נכונה.

(ב) לעשות מבחן t. ← לראות בפלט את המשמעות של מבחן.



בדיקת השערות חד צדדיות:

לפעמים אני יודע את הכיוון של הסטייה האפשרית (נניח שלא הבנו לזה תשובה להשפעה שלילית).

$H_0: \beta \le 0$

$H_1: \beta > 0$

עכשיו עבור $\alpha = 5\%$ יהיה t_{crit} קטן יותר, כלומר יהיה לי קל יותר לדחות

* לחשוב איך להסביר אינטואיטיבית

בדיקת השערות לגבי קומבינציות של פרמטרים:

$H_0: \alpha = 2\beta \iff \alpha - 2\beta = 0 \longrightarrow \hat\alpha - 2\hat\beta \sim N$

$H_1: \alpha \neq 2\beta$

$$t_{stat} = \frac{\hat\alpha - 2\hat\beta - 0}{\sqrt{\hat{Var}(\hat\alpha - 2\hat\beta)}}$$

$= Var(\hat\alpha) + 4Var(\hat\beta) - 2\cdot 2Cov(\hat\alpha, \hat\beta)$

להראות בתרגול

שיעור מס' 4: רווח בר סמך לתחזית:

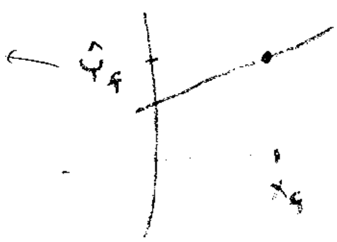
היינו רוצים לשאול מהו הערך של Y בהינתן ש $x = x_f$ (עתידי)

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

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

נניח יש לי מכונית מסוים שנותן לשוטר (התנאי יכול להיות שער הריבית)
← נשאל מה יהיה לאינפלציה? לשער החליפין? ← מתאים למכונית.

התשובה:



זה הערך התחזי. אבל זה לא מספיק. זה רק תחזית נקודתית. מה שחשוב הוא הרב"ס לתחזית. אם הייתי בונה קו רגרסיה ממדגם אחר, אז הייתי מקבל מספר אחר. אני רוצה טווח תחזית שאני אהיה בטוח בו. כיוון ש $\hat{\alpha}$, $\hat{\beta}$ הם משתנים מקריים. לכן רוצים לדעת מהי שונות התחזית.

$$Var(\hat{Y}_f) = Var(\hat{\alpha} + \hat{\beta}\hat{x}_f) = \underbrace{Var(\hat{\alpha})}_{\frac{\sigma_\varepsilon^2 \sum x_i^2}{n\sum(x_i-\bar{x})^2}} + x_f^2 \underbrace{Var(\hat{\beta})}_{\frac{\sigma_\varepsilon^2}{\sum(x_i-\bar{x})^2}} + 2\hat{x}_f \underbrace{Cov(\hat{\alpha},\hat{\beta})}_{=\,?\, = -\frac{\bar{x}\sigma_\varepsilon^2}{\sum(x_i-\bar{x})^2}}$$

ראשית נחשב $Cov(\hat{\alpha},\hat{\beta})$

$$E\left(\frac{\sum(x_i-\bar{x})\varepsilon_i}{\sum(x_i-\bar{x})^2}\right)\left(\sum\left(\frac{1}{n} - \frac{\bar{x}(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}\right)\varepsilon_i\right) = E\left(\sum w_i\varepsilon_i\right)\left(\sum h_i\varepsilon_i\right)$$

$$= E\left(\sum w_i h_i \varepsilon_i^2 + \sum\sum w_i h_j \underbrace{\varepsilon_i\varepsilon_j}_{Cov(\varepsilon_i\varepsilon_j)=0}\right) = \sigma_\varepsilon^2 E\left(\sum w_i h_i\right)$$

$$w_i = \frac{(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}$$

$$h_i = \frac{1}{n} - \bar{x}\cdot w_i$$

$$\Rightarrow w_i h_i = \frac{w_i}{n} - \bar{x} w_i^2$$

$$\Rightarrow \sum w_i h_i = \underbrace{\frac{\sum w_i}{n}}_{0} - \bar{x}\sum w_i^2 = -\frac{\bar{x}\cancel{\sum(x_i-\bar{x})^2}}{\left(\sum(x_i-\bar{x})^2\right)^{\cancel{2}}}$$

המשך ש:

$$Var(\hat{Y}_f)=\frac{\sigma_\varepsilon^2\sum x_i^2}{n\sum(x_i-\bar{x})^2}+\frac{x_f^2\sigma_\varepsilon^2}{\sum(x_i-\bar{x})^2}-2x_f\cdot\frac{\bar{x}\,\sigma_\varepsilon^2}{\sum(x_i-\bar{x})^2}$$

$$=\sigma_\varepsilon^2\left(\frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}+\frac{x_f^2}{\sum(x_i-\bar{x})^2}-\frac{2x_f\bar{x}}{\sum(x_i-\bar{x})^2}\right)$$

$$(x_f-\bar{x})^2-\bar{x}^2$$

$$=\frac{\sigma_\varepsilon^2}{\sum(x_i-\bar{x})^2}\left(\frac{\sum x_i^2}{n}+(x_f-\bar{x})^2-\bar{x}^2\right)$$

$$\sum(x_i-\bar{x})^2=\sum x_i^2+\underbrace{\sum\bar{x}^2}_{n\bar{x}^2}-2\bar{x}\underbrace{\sum x_i}_{n\bar{x}}=\boxed{\sum(x_i^2)+n\bar{x}^2-2n\bar{x}^2}=\sum(x_i^2)-n\bar{x}^2$$

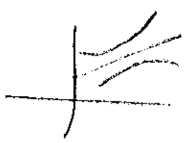
$$=\frac{\sigma_\varepsilon^2}{\sum(x_i-\bar{x})^2}\left(\frac{\sum(x_i^2)-n\bar{x}^2}{n}+(x_f-\bar{x})^2\right)$$

$$\Rightarrow Var(\hat{Y}_f)=\sigma_\varepsilon^2\left(\frac{\cancel{\sum(x_i-\bar{x})^2}}{n\cancel{\sum(x_i-\bar{x})^2}}+\frac{(x_f-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)$$

נשתמש באומד סטיית התקן $\Rightarrow$

$$\hat{\sigma}_{\hat{Y}_f}=\hat{\sigma}_\varepsilon\sqrt{\frac{1}{n}+\frac{(x_f-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$$

התחזית הכי מדויקת בנקודת הממוצע.



* צריך להיזהר מזה

הכל כאמור בהנחה ש: $E(Y_f)=\alpha+\beta x_f$

* שימו לב שאנחנו לא נותנים תחזית ל ε. אנחנו מניחים שהמודל תקף.

* באופן כללי - עדיף לתת תחזית עבור x-ים שיש לי בהם נתונים (באמצע עם שונות מינימלית ותחום שווי). יש להיות שהפרמטרים של המודל משתנים ב x-ים שאין עבורם תצפית.