

# TDAB01 Sannolikhetslära och Statistik

Jose M. Peña  
IDA, Linköpings Universitet

Föreläsning 12

- ▶ Enkel regression, minsta kvadratmetoden, och  $R^2$
- ▶ Konfidensintervall och hypotestest för intercept och lutning
- ▶ Konfidensintervall och hypotestest för prediktion
- ▶ Prediktionsintervall för individuell responsvariabel

## Enkel regression, minsta kvadratmetoden, och $R^2$

- ▶ Regression: Prediktera  $E[Y|X = x]$  där  $Y$  är en slumpvariabel (responsvariabel eller beroende variabel) och  $X = x$  är en observation (förklarandevariabel eller oberoende variabel).
- ▶ Obs. att vi vill prediktera en populations parameter.
- ▶ Linjär regression: Antag  $Y|X = x \sim \mathcal{N}(\mu(x), \sigma^2)$  och  $E[Y|X = x] = \mu(x) = \beta_0 + \beta_1 x$  där
  - ▶  $\beta_0$  är intercepten, och
  - ▶  $\beta_1$  är lutningen.
- ▶ Minsta kvadratmetoden eller maximum likelihood metoden för att estimerar  $\beta_0$  och  $\beta_1$ :
  - ▶  $b_0 = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ , och
  - ▶  $b_1 = \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$ .
- ▶  $SS_{TOTAL} = \sum_i (y_i - \bar{y})^2$  = den **totala** variationen av  $Y$  i strickprovet.
- ▶  $SS_{REG} = \sum_i (\hat{y}_i - \bar{y})^2$  = den variationen **förklarad** av modellen.
- ▶  $SS_{ERR} = \sum_i (y_i - \hat{y}_i)^2$  = den variationen **inte** förklarad av modellen  
=  $SS_{TOTAL} - SS_{REG}$ .
- ▶  $R^2 = \frac{SS_{REG}}{SS_{TOTAL}}$  = **andelen** av den totala variationen förklarad av modellen.

## Konfidensintervall och hypotestest för intercept och lutning

- ▶ **Trick:**  $\sum_i (x_i - \bar{x}) = \sum_i x_i - n\bar{x} = 0$  och då

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i (x_i - \bar{x})y_i - \bar{y} \sum_i (x_i - \bar{x}) = \sum_i (x_i - \bar{x})y_i$$

- ▶  $b_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_i (x_i - \bar{x})y_i}{S_{xx}}$  och då  $b_1$  är en linjär funktion av  $y_i$  och då **normal** fördelad.
- ▶  $E[b_1] = \frac{\sum_i (x_i - \bar{x})E[y_i]}{S_{xx}} = \frac{\sum_i (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{S_{xx}} = \frac{\beta_1 \sum_i (x_i - \bar{x})x_i}{S_{xx}} = \beta_1$  och då  $b_1$  är en **väntevärdesriktig** estimator av  $\beta_1$ .
- ▶  $var[b_1] = \frac{\sum_i (x_i - \bar{x})^2 var[y_i]}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}$ .
- ▶ Nu kan vi bygga konfidensintervall och hypotestest för lutningen baserade på  $t$ -fördelningen, eftersom  $\sigma^2$  brukar vara okänd.
  - ▶  $(1 - \alpha)100\%$  konfidensintervall:

$$b_1 \pm t_{\alpha/2} \frac{s}{\sqrt{S_{xx}}}$$

där  $t$ -fördelningen har  $n - 2$  frihetsgrader, och  $s^2 = SS_{ERR}/(n - 2)$ . För tekniskt varför  $n - 2$  istället för  $n - 1$ . Se sida 371 i Baron.

- ▶ Hypotestest  $H_0: \beta_1 = B$  vs  $H_A: \beta_1 \neq B$ :

$$t = \frac{b_1 - B}{s/\sqrt{S_{xx}}}$$

som har en  $t$ -fördelningen har  $n - 2$  frihetsgrader. Ta  $B = 0$  för att pröva om det finns en linjär relation mellan  $X$  och  $Y$ .

## Konfidsensintervall och hypotestest för intercept och lutning

- ▶  $b_0 = \bar{y} - b_1 \bar{x} = \frac{\sum_i y_i}{n} - \frac{\sum_i (x_i - \bar{x}) y_i \bar{x}}{S_{xx}}$  och då  $b_0$  är en linjär funktion av  $y_i$  och då **normal** fördelad.
- ▶  $E[b_0] = \frac{\sum_i E[y_i]}{n} - E[b_1] \bar{x} = \frac{\sum_i (\beta_0 + \beta_1 x_i)}{n} - \beta_1 \bar{x} = \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0$  och då  $b_0$  är en **väntevärdesriktig** estimator av  $\beta_0$ .

## Konfidsensintervall och hypotestest för prediktion

- ▶  $\mu_* = \mu(x_*) = E[Y|X = x_*] = \beta_0 + \beta_1 x_*$  som kan estimeras av  
 $\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_* = \bar{y} - b_1 \bar{x} + b_1 x_* = \bar{y} + b_1 (x_* - \bar{x}) = \frac{\sum_i y_i}{n} + \frac{\sum_i (x_i - \bar{x}) y_i (x_* - \bar{x})}{S_{xx}} =$   
 $\sum_i \left( \frac{1}{n} + \frac{\sum_i (x_i - \bar{x})(x_* - \bar{x})}{S_{xx}} \right) y_i$  och då  $\hat{y}_*$  är en linjär funktion av  $y_i$  och då **normal** fördelad.
- ▶ Obs. att vi predikterar en **populations parameter**.
- ▶  $E[\hat{y}_*] = E[b_0] + E[b_1]x_* = \beta_0 + \beta_1 x_* = \mu_*$  och då  $\hat{y}_*$  är en **väntevärdesriktig** estimator av  $\mu_*$ .
- ▶  $var[\hat{y}_*] = \sum_i \left( \frac{1}{n} + \frac{\sum_i (x_i - \bar{x})(x_* - \bar{x})}{S_{xx}} \right)^2 var(y_i) = \sigma^2 \left( \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}} \right)$ .
- ▶  $(1 - \alpha)100\%$  konfidsensintervall:

$$\hat{y}_* \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}}$$

där  $t$ -fördelningen har  $n - 2$  frihetsgrader, och  $s^2 = SS_{ERR}/(n - 2)$ .

## Prediktionsintervall för individuell responsvariabel

- ▶ Ett konfidensintervall för  $\hat{y}_*$  representerar osäkerheten om **populationens väntevärde** vid  $X = x_*$ . Men hur ser osäkerheten för ett faktiskt  $Y$ -värde ut om  $X = x_*$  ?
- ▶ Obs. att vi **inte** längre predikterar en populations parameter utan en **slumpvariabel**, dvs vi predikterar inte väntevärdet för stoppsträckan när jag kör 50 km/t, utan stoppsträckan när jag kör 50 km/t, dvs genomsnittet av många körningar vs utfallet av en körning.
- ▶ 95%-igt **prediktionsintervall** för  $Y$ -värdet är ett intervall  $[a, b]$  sådant att

$$P(a \leq Y \leq b | X = x_*) = 0.95$$

där  $a$ ,  $b$  och  $Y$  är slumpvariabler, dvs  $Y$  också !

- ▶ **Prediktera**  $Y = \hat{y}_*$ . Obs. att  $y - \hat{y}_*$  är normal fördelad, eftersom  $Y|X = x$  är normal fördelad. Dessutom,

$$E[y - \hat{y}_*] = 0 \text{ och } sd(y - \hat{y}_*) = \sqrt{\text{var}(y) + \text{var}(\hat{y}_*)} = \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}}$$

Då,  $\frac{y - \hat{y}_* - E[y - \hat{y}_*]}{sd(y - \hat{y}_*)}$  är  $t$ -fördelad. Då,

$$\hat{y}_* \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}}$$

där  $t$ -fördelningen har  $n - 2$  frihetsgrader, och  $s^2 = SS_{ERR}/(n - 2)$ .

## Prediktionsintervall för individuell responsvariabel

- ▶  $(1 - \alpha)100\%$  konfidensintervall:

$$\hat{y}_* \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}}$$

- ▶  $(1 - \alpha)100\%$  prediktionsintervall:

$$\hat{y}_* \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}}$$

- ▶ Prediktionsintervallet är bredare än konfidensintervallet, dvs prediktera en individuell responsvariabel är svårare än prediktera populationens väntevärde.
- ▶ Konfidensintervallet konvergerar mot 0 när  $n$  ökar, eftersom  $S_{xx}$  ökar också. Prediktionsintervallet konvergerar inte mot 0.
- ▶ Prediktionsintervallet är smalare om  $x_*$  ligger nära  $\bar{x}$ , dvs lättare att prediktera under "normala" omständigheter.



- ▶ Enkel regression, minsta kvadratmetoden, och  $R^2$
- ▶ Konfidensintervall och hypotestest för intercept och lutning
- ▶ Konfidensintervall och hypotestest för prediktion
- ▶ Prediktionsintervall för individuell responsvariabel