TDAB01 Sannolikhetslära och Statistik

Jose M. Peña
IDA, Linköping University, Sweden

Förelsning 12: Linjär Regression

Linjär 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
  - β_0 är intercepten, och
  - β_1 är lutningen.
- Minsta kvdratmetoden eller maximum likelihood metoden för att estimera β_0 och β_1:
  - $b_0 = \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, och
  - $b_1 = \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$.
- $R^2 =$ andel av variation förklarad av modellen $= \frac{SS_{REG}}{SS_{TOTAL}} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$, och $SS_{ERR} =$ andel av variation som modellen inte förklarar $= SS_{TOTAL} - SS_{REG} = \sum_i (y_i - \hat{y}_i)^2$.

Samplingfördelning av 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 β_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. T.ex. $H_0 : \beta_1 = T$ vs $H_A : \beta_1 \neq T$ för att pröva om det finns en linjär relation mellan X och Y.
- $b_0 = \bar{y} - b_1\bar{x} = \frac{\sum_i y_i}{n} - \frac{\sum_i (x_i - \bar{x})y_i\bar{x}}{S_{xx}}$ 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 β_0.

Samplingfördelning och Konfidensintervall 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.
- $E[\hat{y}_*] = E[b_0] + E[b_1]x_* = \beta_0 + \beta_1 x_* = \mu_*$ och då $\hat{y}_*$ är en väntevärdesriktig estimator av μ_*.
- $var[\hat{y}_*] = \sum_i \left(\frac{1}{n} + \frac{\sum_i (x_i - \bar{x})(x_* - \bar{x})}{S_{xx}}\right)^2 var(y_i) = \sigma^2 \left(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}\right)$.
- Konfidensintervall: $\hat{y}_* \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}}$.

Prediktionsintervall för Individuell Responsvariabel

- Obs. att vi inte prediktera Ingre en populations parameter utan en slumpvariabel.
- $E[y - \hat{y}_*] = 0$.
- $sd(y - \hat{y}_*) = \sqrt{var(y) + var(\hat{y}_*)} = \sigma\sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}}$.
- Obs. att $y - \hat{y}_*$ är normal fördelad, eftersom $Y|X = x$ är normal fördelad. Då, $\frac{y - \hat{y}_* - E[y - \hat{y}_*]}{sd(y - \hat{y}_*)}$ är t-fördelad.
- Prediktionsintervall: $\hat{y}_* \pm t_{\alpha/2} s\sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{S_{xx}}}$
- Obs. att prediktionsintervallen är bredare än konfidensintervallen, dvs. prediktera en individuell responsvariabel är svårare än prediktera populations väntevärde.
- Obs. att konfidensintervallen convergerar mot 0 när n ökar, eftersom S_{xx} ökar också. Prediktionsintervallen convergerar inte mot 0.
- Obs. att prediktionsintervallen är smalare om x_* ligger nära $\bar{x}$.