Stockholms universitet

STOCKHOLMS UNIVERSITET
Statistiska institutionen

Michael Carlson & Maria Anna Di Lucca
2021-11-04

Regressions- och tidsserieanalys

Formelsamling

DESKRIPTIV STATISTIK

Varians: $s_x^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - n\bar{x}^2}{n-1} = \frac{n\sum x_i^2 - (\sum x_i)^2}{n(n-1)}$

Kovarians: $s_{xy} = cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1}$

$= \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n(n-1)}$

Korrelation: $r_{xy} = corr(x,y) = \frac{s_{xy}}{s_x \cdot s_y} = \frac{s_{xy}}{\sqrt{s_x^2 \cdot s_y^2}}$ Inferens: $t_{n-2} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$

ENKEL LINJÄR REGRESSION

Modell: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ Betingat medelvärde för $Y|X = x$: $\mu_{Y|X=x} = \beta_0 + \beta_1 x$

Parameterskattningar och dessas varianser:

$b_1 = \frac{s_{xy}}{s_x^2} = r_{xy} \cdot \frac{s_y}{s_x}$ $s_{b_1}^2 = \frac{s_e^2}{(n-1)s_x^2} = \frac{s_e^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

$b_0 = \bar{y} - b_1\bar{x}$ $s_{b_0}^2 = s_e^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)$

Prediktion och skattat betingat medelvärde:	$\hat{y}_i = \hat{\mu}_{Y\|X=x_i} = b_0 + b_1 x_i$
Prediktionsintervall för prediktionen $\hat{y}_i$ givet $X = x$:	$\hat{y}_i \pm t_{n-2,\alpha/2} \cdot \sqrt{s_e^2\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}\right)}$
Konfidensintervall för betingade medelvärdet $\mu_{Y\|X=x}$ givet $X = x$:	$\hat{\mu}_{Y\|X=x} \pm t_{n-2,\alpha/2} \cdot \sqrt{s_e^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}\right)}$

ICKE-LINJÄR REGRESSION, exempel

Andragradspolynom: $\hat{y}_i = a + b_1 x_i + b_2 x_i^2$

Exponentiell: $\widehat{\ln(y_i)} = a + bx_i$ $\hat{y}_i = \exp(a + bx_i) = (e^a)(e^b)^{x_i} = (a') \cdot (b')^{x_i}$

$\widehat{\log_{10}(y_i)} = a + bx_i$ $\hat{y}_i = (10^a)(10^b)^{x_i} = (a') \cdot (b')^{x_i}$

ENKEL <u>OCH</u> MULTIPEL LINJÄR REGRESSION (sätt $k = 1$ om enkel regression)

Residualvarians: $$s_e^2 = \frac{\sum_{i=1}^n e_i^2}{n-k-1} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-k-1} = \frac{SSE}{n-k-1} = MSE$$

Kvadratsummor: $$SST = \sum_{i=1}^n (y_i - \bar{y})^2 = (n-1)s_y^2 \quad = \quad SSR + SSE$$

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 = (n-k-1)s_e^2$$

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \qquad = [\text{om } \underline{\text{enkel}} \text{ regression}] = b^2 \sum_{i=1}^n (x_i - \bar{x})^2$$

Förklaringsgrad: $$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad R_{\text{adj}}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

Inferens för β_j: KI: $b_j \pm t_{n-k-1,\alpha/2} \cdot s_{b_j}$ Test: $t_{n-k-1} = \dfrac{b_j - \beta_j^*}{s_{b_j}}$

Test för hela modellen: $$F_{k;n-k-1} = \frac{SSR/K}{SSE/(n-K-1)} = \frac{MSR}{MSE}$$

Beräkningsformler för REGRESSIONS- (enkel linjär) och KORRELATIONSKOEFFICIENT

$$b_1 = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})\,/(n-1)}{\sum (x_i - \bar{x})^2\,/(n-1)} = \frac{s_{xy}}{s_x^2} = \frac{s_{xy}}{s_x^2} \cdot \frac{s_x s_y}{s_x s_y} = \frac{s_{xy}}{s_x s_y} \cdot \frac{s_y}{s_x} = r_{xy} \cdot \frac{s_y}{s_x}$$

$$r_{xy} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2} \cdot \sqrt{n\sum y_i^2 - (\sum y_i)^2}} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2} \cdot \sqrt{\sum y_i^2 - n\bar{y}^2}}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \cdot \sqrt{\sum (y_i - \bar{y})^2}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})\,/(n-1)}{\sqrt{\sum (x_i - \bar{x})^2/(n-1)} \cdot \sqrt{\sum (y_i - \bar{y})^2\,/(n-1)}}$$

$$= \frac{s_{xy}}{\sqrt{s_x^2} \cdot \sqrt{s_y^2}} = \frac{s_{xy}}{s_x s_y} = \frac{s_{xy}}{s_x s_y} \cdot \frac{s_x}{s_x} = \frac{s_{xy}}{s_x^2} \cdot \frac{s_x}{s_y} = b_1 \cdot \frac{s_x}{s_y}$$

TIDSSERIEANALYS

– Komponenter

Additiv modell: $Y_t = T_t + S_t + C_t + E_t$ Multiplikativ modell: $Y_t = T_t \cdot S_t \cdot C_t \cdot E_t$

där T = trend, S = säsong, C = cyklisk/konjunktur samt E = slumpkomponent

- Skattning av trendkomponenten:

– med glidande medelvärden utan säsongvariation, exempel:

3-punkter centrerat: $\hat{T}_t = \frac{1}{3} \cdot y_{t-1} + \frac{1}{3} \cdot y_t + \frac{1}{3} \cdot y_{t+1}$

5-punkter centrerat: $\hat{T}_t = \frac{1}{5} \cdot y_{t-2} + \frac{1}{5} \cdot y_{t-1} + \frac{1}{5} \cdot y_t + \frac{1}{5} \cdot y_{t+1} + \frac{1}{5} \cdot y_{t+2}$

– med centrerade glidande medelvärden med säsongvariation, exempel:

halvårsdata: $\hat{T}_t = \frac{1}{4} \cdot y_{t-1} + \frac{1}{2} \cdot y_t + \frac{1}{4} \cdot y_{t+1}$

kvartalsdata: $\hat{T}_t = \frac{1}{8} \cdot y_{t-2} + \frac{1}{4} \cdot y_{t-1} + \frac{1}{4} \cdot y_t + \frac{1}{4} y_{t+1} + \frac{1}{8} \cdot y_{t+2}$

månadsdata: $\hat{T}_t = \frac{1}{24} \cdot y_{t-6} + \frac{1}{12} \cdot y_{t-5} + \cdots + \frac{1}{12} \cdot y_{t+5} + \frac{1}{24} \cdot y_{t+6}$

– med regressionsanalys, linjär trend och exponentiell trend:

Modell: $Y_t = \beta_0 + \beta_1 t + \varepsilon_1$ Skattad modell: $\hat{y}_t = b_0 + b_1 t = \hat{T}_t$

$\ln Y_t = \beta_0 + \beta_1 t + \varepsilon_1$ $\hat{y}_t = \exp(b_0 + b_1 t) = \hat{T}_t$

- Justering av säsongsindex $\bar{S}_j$ med p säsonger (halvår, kvartal el. månader osv.):

Additiv modell: $S_j^+ = \bar{S}_j - \left(\frac{\sum \bar{S}_i}{p}\right)$ Multiplikativ modell: $S_j^+ = \frac{\bar{S}_j}{(\sum \bar{S}_i / p)}$

- Trend- och säsongsrensning:

Additiv modell: $y_t - \hat{T}_t$ resp. $y_t - S_t^+$ Multiplikativ modell: $y_t / \hat{T}_t$ resp. y_t / S_t^+

LOGISTISK REGRESSION och ODDS

Odds för en händelse A: Oddskvot för händelsen A mot B:	$\text{Odds}(A) = \dfrac{P(A)}{P(\bar{A})} = \dfrac{P(A)}{1 - P(A)} \iff P(A) = \dfrac{\text{Odds}(A)}{1 + \text{Odds}(A)}$ $\text{OR} = \dfrac{\text{Odds}(A)}{\text{Odds}(B)}$

- Logistisk regression:

Enkel modell:	$P(Y = 1\|x) = \dfrac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} = \dfrac{1}{1 + \exp(-\beta_0 - \beta_1 x)}$ $P(Y = 0\|x) = 1 - P(Y = 1\|x) = \dfrac{1}{1 + \exp(\beta_0 + \beta_1 x)}$ $\text{Odds}(Y = 1\|x) = \exp(\beta_0 + \beta_1 x)$ $\text{LogOdds}(Y = 1\|x) = \beta_0 + \beta_1 x$
Multipel modell:	$\text{LogOdds}(Y = 1\|x_1, \dots x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$

Intercept β_0:	$P(Y = 1\|x_1 = \cdots = x_k = 0) = \dfrac{\exp(\beta_0)}{1 + \exp(\beta_0)}$
Oddskvot för $Y = 1$ när $X_j = x_j + 1$ mot $X_j = x_j$:	$\text{OR}(X_j) = \dfrac{\text{Odds}\left(Y = 1\middle\|x_j + 1, \text{allt annat lika}\right)}{\text{Odds}\left(Y = 1\middle\|x_j, \text{allt annat lika}\right)} = \exp(\beta_j)$
KI för $\text{OR}(X_j)$:	$\left(\exp(b_j - z_{\alpha/2} \cdot s_{b_j})\ ;\ \exp(b_j + z_{\alpha/2} \cdot s_{b_j})\right)$