Vorlesung

Thursday, April 11, 2019 12:38 PM

Additives Modell:

$$E(Y|x_1,\dots,x_p) = b_0 + f_1(x_1) + f_2(x_2) + \dots + f_p(x_p)$$

bzw. $Y = b_0 + \dots + \varepsilon$

Lineares Modell:

$$E(Y|x_1,\dots,x_p) = b_0 + b_1 x_1 + \dots + b_p x_p$$

Verallg. lineares Modell:

$$E(Y|x_1,\dots,x_p) = f(b_0 + b_1 x_1 + \dots + b_p x_p)$$

Wdh. Schätzer:

$(Y_1,X_1),\dots,(Y_n,X_n)$ iid $\Rightarrow \hat{b}_1$ (wird geschätzt)

$T := g(X_1,Y_1,X_2,Y_2,\dots,X_n,Y_n) : (\Omega,\mathcal{F},P) \to \mathbb{R}$

ϵ wird in Abhängigkeit der Stichprobe berechnet

$\Rightarrow$ variiert je nach Stichprobe

$$E_{b_1}(T) = b_1$$

Univariate lineare Regression:

$Y = b_0 + b_1 X + \varepsilon$, $E(\varepsilon) = 0$, X, ε unabh.

KQ-Schätzer: $\hat{b}_1 = \frac{r_{xy}}{s_x} \cdot s_y$

$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$

$E(\hat{b}_0) = b_0$

$E(\hat{b}_1) = b_1$

$Var(\hat{b}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$

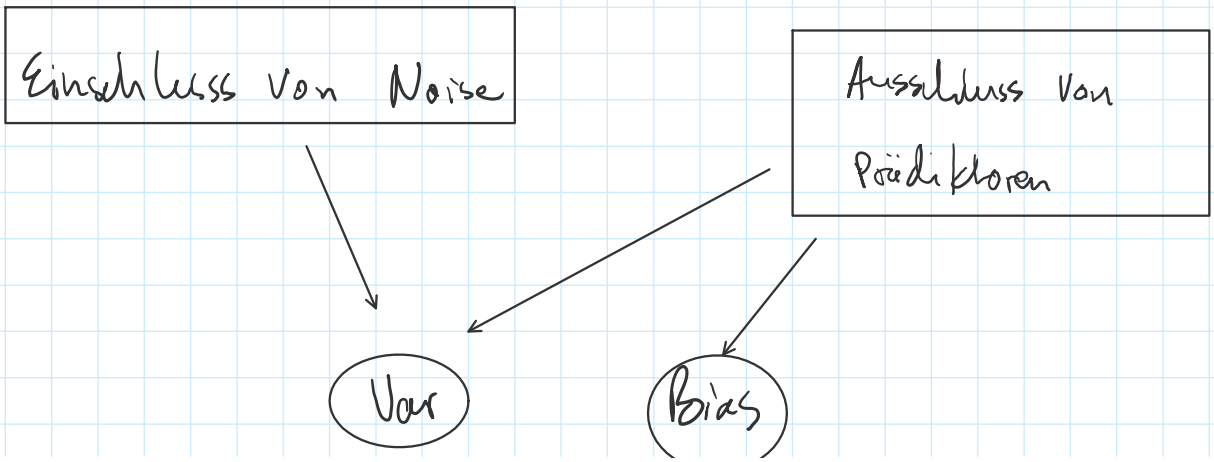
Multivariate lineare Regression:

$Y = b_0 + b_1 X_1 + b_2 X_2 + \varepsilon$

ε-treue ✓

$Var(\hat{b}_1) = \frac{\sigma^2}{(1 - R^2) \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}$ mit R^2 Bestimmtheitsmaß aus $X_1 = \tilde{b}_0 + \tilde{b}_2 X_2 + \varepsilon$

Fehlerquellen:



Var Bias

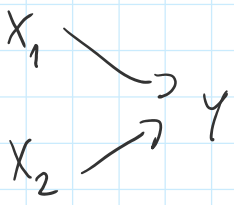
Bsp.: Modellierung einer Noise-Var. im Lin. Regr. Modell

- datengen. Modell: $Y = b_1 X_1 + \varepsilon = b_1 X_1 + 0 \cdot X_2 + \varepsilon$
- Analysemodell: $Y = b_1 X_1 + b_2 X_2 + \varepsilon$
- Bias: $E(\hat{b}_1) = b_1$
- Varianz: $Var(\hat{b}_1) = \frac{\sigma^2}{(1-R^2)\sum_{i=1}^{n}(x_{i1} - \bar{x})^2} \geq \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Bsp.: Fehlende Einflussvariable

- datengen. Modell $Y = b_1 X_1 + \overbrace{b_2 X_2 + \varepsilon}^{:= \tilde{\varepsilon}} = b_1 X_1 + \tilde{\varepsilon}$
- Analysemodell $Y = b_1 X_1 + \varepsilon$

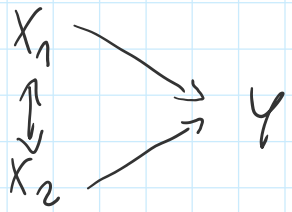
a) unabh. Prädiktoren



Bias: $E(\hat{b}_1) = b_1$

Var: $Var(\hat{b}_1) = \frac{\tilde{\sigma}^2}{\sum_{i=1}^{n}(x_{i1} - \bar{x})^2} \geq \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ mit $\tilde{\sigma}^2 = Var(\tilde{\varepsilon})$

b) asch. Prädiktoren (Bsp.: $X_2 = a \cdot X_1$)



datengen. Modell: $y = b_1 x_1 + b_2 x_2 + \varepsilon = (b_1 + b_2 \cdot a) X_1 + \varepsilon$

Bias: $E(\hat{b}_1) = b_1 + b_2 \cdot a \Rightarrow R^2 \to 1$

Var: $\mathrm{Var}(\hat{b}_1) = \dfrac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} < \dfrac{\sigma^2}{(1-R^2)\sum_{i=1}^{n} (x_i - \bar{x})^2}$ ($1-R^2 \downarrow 0$)

Nichtparametrische Tests:

Vergleich: Gauss-Test (bekannter parametrischer Test)

$X_1, \dots, X_n$ iid. ZV $X_i \sim N(\mu, \sigma^2)$, σ^2 bekannt

$H_0: \{\mu = 0\}$ vs $H_1: \{\mu \neq 0\}$

$$T = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i}{\sigma} \cdot \sqrt{n} \underset{H_0}{\sim} N(0,1)$$

Jetzt:

Einstichprobentest:

Rangtests

$X_1, \dots, X_n$ iid. ZV, X_i stetig, x_{med} = Median von X_i

$H_0: \{ x_{med} = S_0 \}$ vs $H_1: \{ x_{med} \neq S_0 \}$, Sign. niveau α

Unter H_0 gilt: $P(X_i < S_0) = P(X_i \geq S_0) = 0.5$

$$Y_i := \begin{cases} 0, & x < S_0 \\ 1, & x_i \geq S_0 \end{cases} \Rightarrow \text{unter } H_0 \text{ ist } Y_i \sim B(1, 0.5)$$

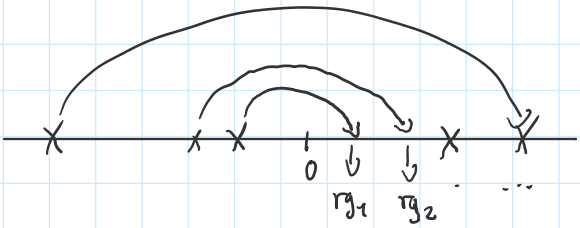
$$T := \sum_{i=1}^{n} Y_i \underset{H_0}{\sim} B(n, 0.5)$$

Wilcoxon-Vorzeichen-Rang-Test:

$X_1, \dots, X_n$ iid. ZV, stetig, symmetrisch, x_{med} = Median der X_i

$H_0: \{ x_{med} = S_0 \}$ vs. $H_1: \{ x_{med} \neq S_0 \}$

$$D_i := X_i - S_0$$



bilde Ränge der $|D_i|$, $rg(|D_i|)$

$$W^+ := \sum_{i, D_i > 0} rg(|D_i|) \qquad W^- := \sum_{i, D_i < 0} rg(|D_i|)$$

Es gilt: $W^+ + W^- = \frac{n(n+1)}{2}$. Unter H_0 $E(W^+) = \frac{n(n+1)}{4}$.

$$T = \min(W^+, W^-)$$

$$T = \min(W^+, W^-)$$

Zweistichprobentest:

Wilcoxon-Rangsummentest:

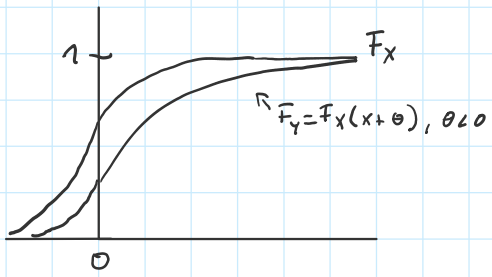
$X_1, \dots, X_n$ iid.

$Y_1, \dots, Y_m$ "

$X_1, \dots, X_n, Y_1, \dots, Y_m$ unabh.

X_i stetig mit Verfkt. F_X, Y_i stetig mit Verfkt. F_Y

Es gelte das Lokations-Shift-Modell: $F_X(x) = F_Y(x+\theta) \quad \forall x$



Frage: "Sind die Verteilungen gleich?"

$H_0: \{\theta = 0\}$ vs. $H_1: \{\theta \neq 0\}$

Bilde Ränge $rg(X_1), rg(X_2), \dots, rg(Y_1), \dots, rg(Y_m)$

$$T = \sum_{i=1}^{n} rg(X_i)$$

$$T = \sum_{i=1}^{n} g(X_i)$$

Permutationstests:

Verb. Stichproben: $(X_i, Y_i)_{i=1,\dots,n}$ iid. ZV

$D_i := X_i - Y_i$ symmetrisch um θ

Fragestellung: „Gibt es systematische Tendenzen?"

$H_0: \{\theta = 0\}$ vs $H_1: \{\theta \neq 0\}$

Unter H_0 gilt: $X_i - Y_i \sim Y_i - X_i$

$$T := \frac{1}{n}\sum_{i=1}^{n} D_i = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i) \sim \frac{1}{n}\left(\sum_{i\in M}(X_i - Y_i) + \sum_{i\notin M}(Y_i - X_i)\right) = T_M$$

für $M \subset \{1,\dots,n\}$

$$p = P_0(T \geq t) = \frac{|\{M \subset \{1,\dots,n\},\ T_M \geq t\}|}{2^n}$$

2 unabh. Stichproben (wie bei Rangsummentest):

$$Z := (X_1,\dots,X_n,Y_1,\dots,Y_m) = (Z_1,\dots,Z_{n+m})$$

Unter H_0 gilt: $Z \sim Z_\pi = (Z_{\pi(1)}, Z_{\pi(2)},\dots, Z_{\pi(n+m)})$

$$T := \frac{1}{n}\sum_{i=1}^{n} X_i \underset{H_0}{\sim} \frac{1}{n}\sum_{i=1}^{n} Z_{\pi(i)} = T_\pi$$

$$p = P_0(T \geq t) = \frac{|\{\pi : T_\pi \geq t\}|}{\binom{n+m}{n}}$$