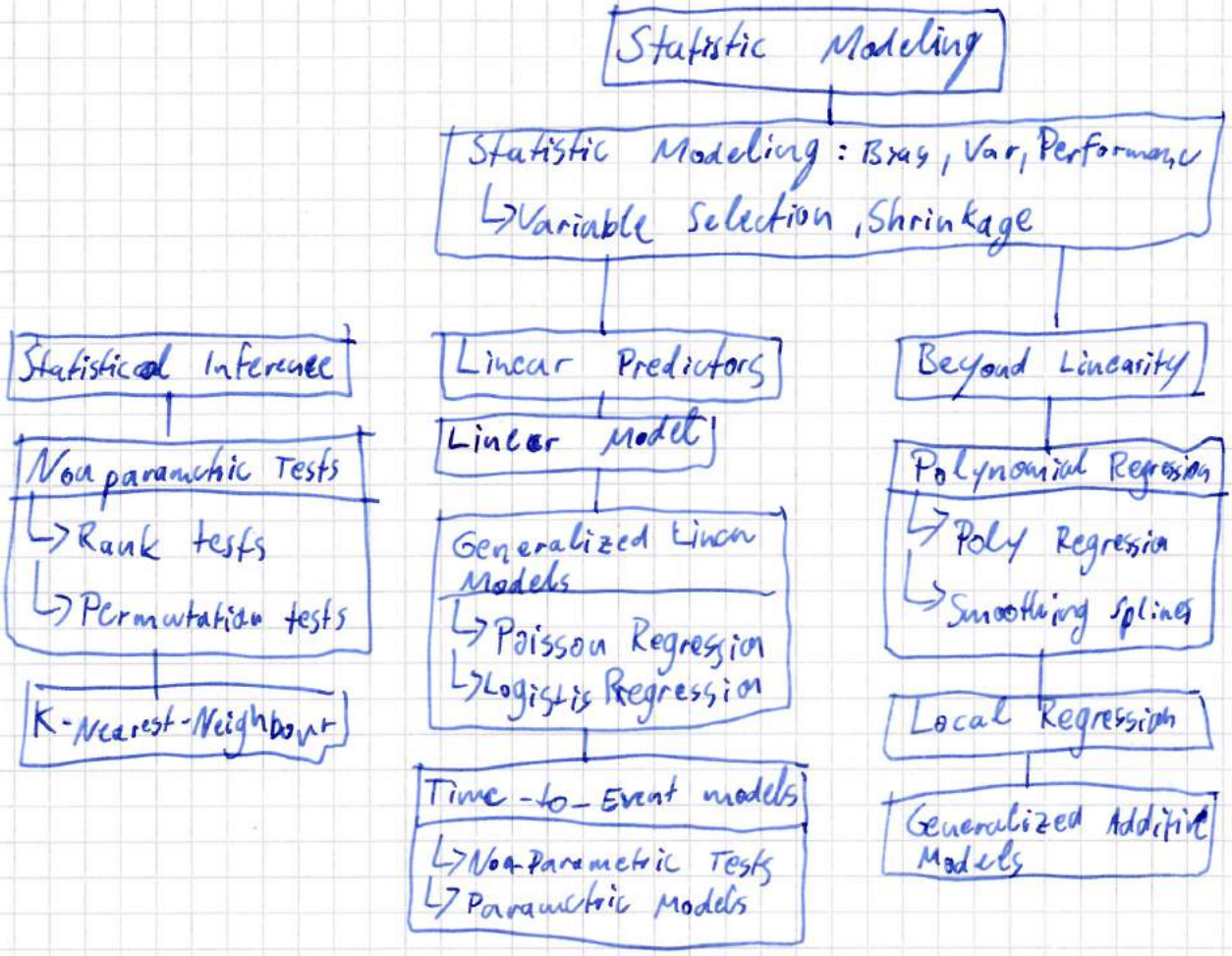
Übersicht


Statistic Modeling
Statistic Modeling : Bias, Var, Performance
↳ Variable Selection, Shrinkage
Statistical Inference
Non parametric Tests
↳ Rank tests
↳ Permutation tests
K-Nearest-Neighbour
Linear Predictors
Linear model
Generalized Linear Models
↳ Poisson Regression
↳ Logistic Regression
Time-to-Event models
↳ Non-Parametric Tests
↳ Parametric Models
Beyond Linearity
Polynomial Regression
↳ Poly Regression
↳ Smoothing Splines
Local Regression
Generalized Additive Models

Vorhersage Fehler

□ Bias-Variance-TradeOFF

↳ Expected Test Error = Bias² + Var + Irreducable Error

↳ $MSE = Bias[\hat{f}(x)] + Var[\hat{f}(x)] + \sigma^2$

$E[(y - \hat{f}(x))^2]$ ↑ Modellfehler

↑ Datenfehler

↑ $Var(y)$ ↑ deswegen ist der auch irreducable

↳ Varianz + ⇒ Overfitting

↳ Bias + ⇒ Underfitting

□ Anpassungsgüte

↳ $AIC = -2LL + 2(p+1)$

↳ $BIC = -2LL + p\log(n)$

Einer der

$L(b) = \prod_{i=1}^{n} P(y=y_i | b, x_i)$

$\Leftarrow LL(b) = \sum_{i=1}^{n} y_i (b^T x_i) - \exp(b^T x_i) - \log(y_i!)$

↳ Der AIC ist immer nur auf den gleichen Daten vergleichbar, da dieser ein Vergleichswert und kein Richtwert ist (wächst mit der Anzahl an Daten)

↳ AIC ist ein Maß für die Anpassungsgüte und bestraft Overfitting

↳ In der Praxis nimmt man das Modell mit dem geringsten AIC bei etwa gleichem MSE auf den Testdaten

Nichtparametrisch

Nicht Parametrisch	Parametrisch
▯ Verteilung nicht bekannt	▯ Für jeden Parameter Θ ist die Verteilung der ZV bekannt
▯ Nullhypothese nicht über Verteilung definierbar	▯ Nullhypothese ist Punkt/Intervall und somit ist die Verteilung bekannt
↳ Ablehnbereich über kritische Werte	↳ Ablehnbereiche über Quantile der Verteilung

Nichtparametrische Tests

↳ Rangtests: Werte bekommen Ränge zugewiesen und über diese wird getestet

↳ Permutationstests: Beobachtete Werte werden als fest betrachtet und die Verteilung zufällig behandelt

↳ Monte-Carlo-Permutationstest: Simulationstest über Permutationen

① Einstichprobentests / Verbundene Stichproben

$P(X \leq x_{med}) \geq P(X \geq x_{med}) = \frac{1}{2}$

▯ Vorzeichentest

↳ X_i uiv mit stetiger Verteilung

↳ $H_0: \{x_{med} = \delta_0\}$ vs $H_1: \{x_{med} \neq \delta_0\}$ zu α

↳ $Y_i = \begin{cases} 1, & x < x_{med} \\ 0, & x \geq x_{med} \end{cases} \Rightarrow Y_i \overset{H_0}{\sim} B(1, \frac{1}{2}) \Rightarrow T = \sum Y_i \overset{H_0}{\sim} B(n, \frac{1}{2})$

↳ $x_i = \delta_0$ können wegen stetig theoretisch nicht auftreten $\overset{\text{weg}}{\Rightarrow}$ n reduzieren

↳ nominal und ordinal möglich

↳ $n \geq 30$: $B(n, \frac{1}{2}) \sim N(\frac{1}{2}n, \frac{1}{4}n)$

▯ Wilcoxon-Vorzeichen-Rangtest

↳ X_i uiv symmetrisch mit stetiger Verteilung

↳ $H_0: \{x_{med} = \delta_0\}$ vs $H_1: \{x_{med} \neq \delta_0\}$

↳ $D_i = X_i - \delta_0 \Rightarrow$ Bilde Ränge der $|D_i|$ (ordinalskala ab 1 ↗)

↳ $W^+ = \sum_{D_i > 0} rk(|D_i|)$, $W^- = \sum_{D_i < 0} rk(|D_i|) \Rightarrow T = \min(W^+, W^-)$

↳ $W^+ \overset{H_0}{\approx} W^- \Rightarrow$ Ablehnbereich für T über n

↳ Für $X_i = X_j$ bekommen die Werte den durchschnittlichen Rang

↳ $n \geq 30$: Approximation über Normalverteilung möglich

▯ Permutationstest

↳ X_i uiv mit $D_i = X_i - Y_i$ symmetrisch um Θ verteilt, $M \subset \{1, \dots, n\}$ beliebig

↳ $H_0: \{\Theta = 0\}$ vs $H_1: \{\Theta \neq 0\}$

↳ $(X_i - Y_i) \overset{H_0}{\sim} (Y_i - X_i) \Rightarrow t = \frac{1}{n} \sum D_i \sim T_M = \frac{1}{n}\left(\sum_{i \in M} x_i - y_i + \sum_{i \notin M} y_i - x_i\right)$

⇒ Test über alle M. Wahrscheinlichkeit mit Monte-Carlo zufällig probieren.

② Zweistichprobentests / Unverbundene Tests

□ Wilcoxon-Rangsummen-Test
↳ X_i, Y_i unabhängig. X_i stetig mit F_X verteilt. Y_i stetig mit F_Y verteilt.
  ↳ Es gelte: $F_X(x) = F_Y(x+\theta)$
↳ $H_0: \theta = 0$ vs $H_1: \theta \neq 0$ ⇒ Wahlweise $<$/$>$, um die Lage zu testen
↳ Bilde Ränge über X_i, Y_i ⇒ $T = \sum_{k=1}^{n} R_k(X_i)$
↳ $x_i = x_j$ ⇒ zufälliger Rang; $x_i = y_j$ Durchschnitt
↳ $n \geq 30$: Normalverteilungsapproximation

□ Permutationstest
↳ $Z = (X_1, \dots, X_n, Y_1, \dots, Y_m) \Rightarrow Z_i$
↳ wie Wilcoxon ohne Stetigkeitsannahme
↳ $Z \underset{H_0}{\sim} Z_\pi$ für alle Permutationen $Z_\pi = Z_{\pi(1)}, \dots, Z_{\pi(n+m)}$
↳ Zweiseitige Testentscheidung mit $\frac{\alpha}{2}$

Durchschnittsrang

D_i	1.5	2.5	3.5	3.5	3.5	4.5
Rang ohne Durchschnitt	1	2	3	4	5	6
Rang mit Durchschnitt	1	2	4	4	4	6

$(3+4+5)/3$

Verallgemeinerte Lineare Modelle (GLM)

Idee: Die Fehlerverteilung der abhängigen Variablen (Response variable) bekommt eine andere Verteilung anstatt einer Normalverteilung

Bisher: $E(Y_i|x_i) = b^T x_i$ ← Transformation der Erwartenden Verteilung

↳ Nun: mit Link Funktion $\eta_i = g(E(Y_i|x_i))$ diskret!

Poisson-Regression

↳ $Y_1, \dots, Y_n$ (bedingt) unabhängig aus $\mathbb{N}$ $\Rightarrow$ $Y_i | x_i = x_i \sim Poi(\lambda_i)$

mit $\lambda_i = \exp(b^T x_i)$ $\Rightarrow$ $E(Y_i|x_i) = \lambda_i = \exp(b^T x_i)$

↳ Optimierungsansatz

↳ $L(b) = \prod P(Y_i = y_i | b, x_i) = \prod \exp(-\lambda_i) \frac{\lambda_i^{y_i}}{y_i!}$ ← einsetzen

$LL(b) = \sum y_i (b^T x_i) - \exp(b^T x_i) - \log(y_i!)$

$[LL(b)]' = \Delta l = \sum x_i (y_i - \exp(b^T x_i)) = s(b) = 0$ $\Rightarrow$ GLS

↳ Modellannahmen: Pearson-Residuen streuen mit Var=1 um E=0

3 Kernelemente der GLM

1. Exponentielle Familie von W-Verteilungen
2. Linear Predictor $\eta = b^T x$
3. Link Function $g(E(Y_i|x_i)) = \eta_i \Rightarrow E(Y_i|x_i) = g^{-1}(\eta_i)$

Weitere GLM

↳ Änderung der Link Funktion g. Notwendig: g invertierbar.

Deviaz

$\hat{b}_M$ ML-Schätzer

↳ $L_M(\hat{b}_M)$, $LL_M(\hat{b}_M) = \log(L_M)$

↳ $D_M = 2(LL_{opt} - LL_M)$ ← Deviaz

⇓

Likelihood-Ratio-Test

$x_1, \dots, x_p, x_{p+1}, \dots, x_{p+q}$ Kovariaten

↳ $H_0: \{b_{p+1} = \dots = b_{p+q} = 0\}$

↳ $LR = D_{M^-} - D_M = 2(LL_M - LL_{M^-}) \underset{H_0}{\sim} \chi^2_q$

Basiserweiterung in X

Modell: $Y = f(X) + \varepsilon$ mit $E(\varepsilon) = 0$, $V(\varepsilon) = \sigma^2$

Idee: $f(X) = b_0 + b_1 h_1(X) + \dots + b_m h_m(X) = b_0 + \sum_{i=1}^{m} b_i h_i(X)$

↳ $Z = \begin{pmatrix} 1 & h_1(x_1) & \dots & h_m(x_1) \\ \vdots & & & \\ 1 & h_1(x_n) & \dots & h_m(x_n) \end{pmatrix}$ Designmatrix

↳ Nichtlineare Modelle mit linearen Methoden verarbeiten

□ Polynomregression: $Y = b_0 + b_1 X + b_2 X^2 + \dots + b_d X^d + \varepsilon$

□ Stückweise Polynomial:

↳ $(-\infty, \xi_1), [\xi_1, \xi_2), \dots, [\xi_{K-1}, \xi_K), [\xi_K, \infty)$ Intervalle ← K Knoten ξ_k, ← K+1 Intervalle

↳ $Y = f_k(X) + \varepsilon$ Polynom in jedem Intervall

↳ Anzahl freie Parameter: $(d+1)(K+1)$ ← Anzahl der Polynome, ↑ Komplexität pro Polynom

□ Regression Splines

↳ d-1 mal stetig db ⇒ Ableitungen und Punkte stimmen in Knoten überein

↳ $d + K + 1$ Freiheitsgrade

↳ Bsp: $h_i(x) = x^{i-1}, i = 1, \dots, d+1$; $h_i(x) = (x - \xi_{i-d-1})_+^d$ $i = d+2, \dots, d+K+1$

$f(X) = \sum_{m=1}^{d+1} b_m X^{m-1} + \sum_{k=1}^{K} b_{d+1+k} (X - \xi_k)_+^d = \sum_{m=1}^{d+K+1} b_m h_m(X)$

Truncated power basis

↳ Knotenanzahl z.B. äquidistante Quantile und CV für die Anzahl

↳ Natürlicher Spline: $S_0''(x_0) = S_{n-1}''(x_n) = 0$ oder $(-\infty, \xi_1], [\xi_K, \infty)$ linear verlauf

↳ Anzahl freie Parameter: $K - d + 3$

□ Smoothing-Regression-Spline

↳ $RSS(f, \lambda) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int (f''(t))^2 dt$

↳ f minimiert: natural cubic spline mit Knoten in x_i

↳ λ mittels CV

□ Effective Degrees of freedom

↳ Linear: $\hat{y}_i = (Hy)_i$, $H = X(X^T X)^{-1} X^T$ ⇒ $p = \text{spur}(H)$ ← Anzahl zu schätzende Parameter

↳ RSS: $\hat{y}_i = (S_\lambda y)_i$ ⇒ $p = \text{spur}(S_\lambda)$

Hängt nur von λ ↑ x ab

Anzahl zu schätzende Parameter = Effective Degrees of freedom

Kernel Smoother

KNN (K-Nearest-Neighbor)

↳ Idee: Moving average
↳ $N_K(x) = \{x_{(1)}, \dots, x_{(K)}\}$ mit $|x - x_i|$ K nächste Nachbarn (kleinste Δ)
↳ $\hat{f}(x) = \frac{1}{K} \sum y_i$ für i von $x_i \in N_K(x)$

$$= \underset{b_0}{\operatorname{argmin}} \sum_{i=1}^{n} \underbrace{K(x, x_i)}_{\text{Kernel}} (y_i - b_0)^2 \quad \text{mit } K(x, x_i) = I(x_i \in N_K(x))$$

⇒ Lokale Polynomiale Regression von Grad 0

⇒ Nicht glatt sondern konstant auf Intervallen

Kernel Average Smoother

(Verbesserung)

↳ Idee: $K(x, x_i)$ durch eine Gewichtsfunktion ersetzen
↳ $K_\lambda : \mathbb{R}^2 \to \mathbb{R}^+$ mit Fensterbreite λ ⇒ $\int_{-\infty}^{\infty} K_\lambda(x, x_i)\, dx = 1$
↳ K_λ stetig ⇒ λ als Smoothing Parameter

Lokale Lineare Regression

↳ $\hat{f}(x) = \hat{b}_0(x) + \hat{b}_1(x) x$
↳ $(\hat{b}_0(x), \hat{b}_1(x)) = \underset{(b_0, b_1) \in \mathbb{R}^2}{\operatorname{argmin}} \sum_{i=1}^{n} K_\lambda(x, x_i) (y_i - b_0 - b_1 x_i)^2$

↳ Anstatt quadratische Differenz wird nun über die optimale Verbindungsgerade geglättet

Lokale Polynomiale Regression

↳ $\hat{f}(x) = \hat{b}_0(x) + \hat{b}_1(x) x + \hat{b}_2(x) x^2$

↳ Wie Linear nur Erweiterung um Polynome höherer Ordnung

↳ $d > 2$ möglich, aber unüblich.

↳ Grund: Quadrat erweitert das Modell und ermöglicht kurvige Glättung. Höhere Ordnungen würden Wendepunkte ermöglichen, nur diese sind aber durch die Fensterbreite λ abzufangen als durch den Kernel