Abstract: Asymptotische Verteilung von vollen Datentiefen

In dem Vortrag am 06.11.18 werden wir uns mit den vollen Datentiefen, eine robusten Maßzahl für Regressionsmodelle beruhend auf Vorzeichen der Residuen, beschäftigen. Für eine vorgegeben Klasse $\Theta \subset \mathbb{R}^p$ von Regressionsmodellen, wird für jeden p-dimensionalen Parametervektor $\vartheta \in \Theta$ eine Regressionsfunktion $g_\vartheta : \mathbb{R} \to \mathbb{R}$ definiert. Für Zufallsvariablen $Y_1, ..., Y_N$ formulieren wir folgende Regressionsgleichung:

$$Y_n = g_{\vartheta^*}(X_n) + E_n, \text{ für } n = 1, ..., N$$
$$\text{wobei } \vartheta^* \text{ der wahre Parameter sei.}$$

Die Zufallsvariablen $E_1, ..., E_N$ wirken als additive (Modell/Mess-)Fehler auf die Regressionsfunktion g_{ϑ^*}. Die Regressoren $X_1, ..., X_N$ können zufällig oder deterministisch sein und sollen fast sicher folgende Anordnungsstruktur erfüllen:

$$X_1 < X_2 < ... < X_N.$$

Dadurch können auch Zeitreihenmodelle mit Explosion wie in Kustosz et al. (2016) formuliert werden, was z.B. am Lehrstuhl *Statistik in den Ingenieurswissenschaften* Anwendung bei der stochastischen Modellierung der Rissentwicklung von Bauteilen findet. Ferner soll für die Modellfehler $E_1, ..., E_N$ gelten:

$$\text{med}(E_n) = 0 \text{ und } P(E_n \neq 0) = 1 \text{ für } n = 1, ..., N. \tag{1}$$

Weiterhin seien $Z_n = (Y_n, X_n)$ für $n = 1, ..., N$ die zufälligen Datenpunkte des Regressionsmodells. Für jedes $\vartheta \in \Theta$ können wir die Residuen definieren:

$$\text{res}(\vartheta, Z_n) := Y_n - g_\vartheta(X_n) \text{ für } n = 1, ..., N,$$

wobei für den wahren Parameter ϑ^* gilt:

$$\text{res}(\vartheta^*, Z_n) = E_n \text{ für } n = 1, ..., N, \tag{2}$$

Das heißt unter dem wahren Parameter ϑ^* wissen wir, wie sich die Vorzeichen der Residuen verhalten. Es gilt mit (1) und (2):

$$P(\text{res}(\vartheta^*, Z_n) > 0) = P(\text{res}(\vartheta^*, Z_n) < 0) = \frac{1}{2}.$$

Ziel der Masterarbeit ist es, robuste statistische Testverfahren beruhend auf der vollen K-Tiefen mit folgenden Hypothesenpaaren konstruieren:

$$H_0 : \vartheta \in \Theta_0 \text{ vs. } H_1 : \vartheta \in \Theta_1,$$

wobei $\Theta_0 \uplus \Theta_1 = \Theta$ gilt. So können wir testen, ob vorgegebene Parameterbereiche innerhalb eines Regressionsmodells als unpassend deklariert werden können. Dazu benötigen wir die asymptotische Verteilung der K-Tiefen. Die volle K-Tiefe ist für einen vorgeschlagenen Parameter-Kandidaten ϑ wie folgt definiert:

$$d_S^K(\vartheta, Z) = \frac{1}{\binom{N}{K}} \sum_{1 \le n_1 < n_2 < \ldots < n_K \le N} \left(\prod_{k=1}^{K} \mathbb{1}\{\text{res}(\vartheta, Z_{n_k})(-1)^k > 0\} + \prod_{k=1}^{K} \mathbb{1}\{\text{res}(\vartheta, Z_{n_k})(-1)^{k+1} > 0\} \right).$$

Sie beschreibt unter allen geordneten K-Tupeln den relativen Anteil mit $(K-1)$ Vorzeichenwechsel (alternierende Residuen). Alleiniges Zählen von Vorzeichen der Residuen liefert zwar einen robusten Ansatz, besitzt allerdings eine schlechte Güte, da viele unpassende Modelle nicht abgelehnt werden, siehe Abbildung 1. Berücksich-

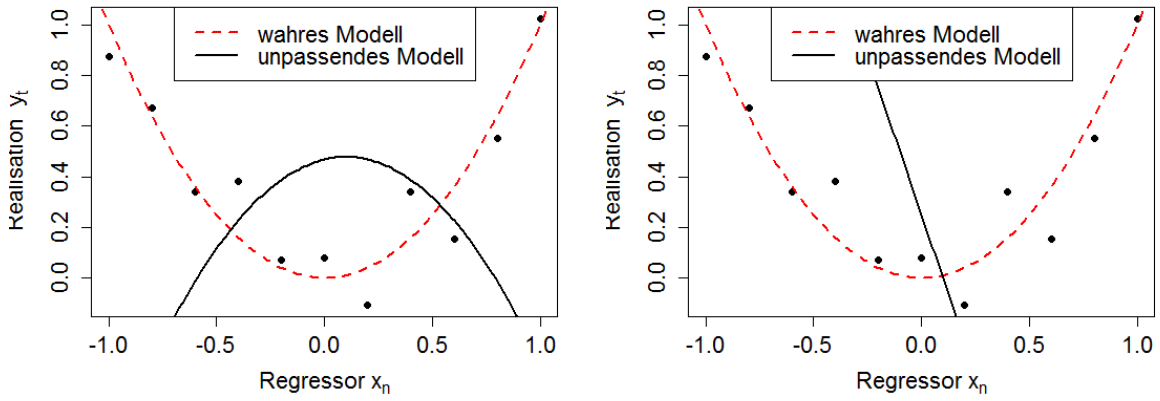

Abbildung 1: Vergleich: wahres Modell und unpassendes Modell und die Vorzeichen der Modellabweichungen

tigt man zusätzlich die Reihenfolge der Vorzeichenstrukturen, so verbessert sich die Modellgüte. In dem Vortrag soll die Herleitung der asymptotischen Verteilung von

$$N\left(d_S^K(\vartheta^*, Z) - \left(\frac{1}{2}\right)^{K-1}\right) \text{ für } K = 3$$

skizziert werden. Die Herleitung beruht auf den Resultaten in Kustosz et al. (2016). Allerdings werden in der Masterarbeit einige Aussagen innerhalb der Rechnung präzisiert, wodurch insbesondere sich die volle Dreier-Tiefe in linearer statt kubischer Laufzeit bestimmen lässt. Ferner wird in der Masterarbeit die Analyse der Asymptotik durch die direkte Anwendung des Invarianzprinzips von Donsker stark vereinfacht, da in Kustosz et al. (2016) der Satz von Donsker für einen Spezialfall für die Herleitung der Asymptotik der vollen Dreier-Tiefe gezeigt wird. In der Masterarbeit müssen Eigenschaften der Skorohod-Topologie diskutiert werden, welche in unserer vorliegenden Situation auf die uniforme Topologie zurückgeführt werden kann.
Außerdem ist in der Masterarbeit die asymptotische Verteilung für den Fall $K = 4$ vom Autor selbst hergeleitet worden. Neben der Asymptotik gewinnen wir auch hier eine Darstellung der vollen Vierer-Tiefe, die sich in linearer Laufzeit bestimmen lässt.
Ein besonderes Resultat des Autors ist die Darstellung der vollen K-Tiefen durch die Funktion $\Phi(x) := \mathbb{1}\{x < 0\} - \mathbb{1}\{x > 0\}$ für $x \neq 0$. Der Autor hat folgenden Satz gefunden und ihn in seiner Masterarbeit bewiesen.

Satz 0.1 (Φ-Darstellung von allgemeinen vollen K-Tiefen).
Für Zufallsvariablen $E_{n_1}, ..., E_{n_K}$ mit $P(E_{n_i} \neq 0) = 1$ für $i = 1, ..., K$ gilt für $K \in \mathbb{N}$:

$$\begin{aligned}&\prod_{i=1}^{K} \mathbb{1}\{E_{n_i}(-1)^i > 0\} + \prod_{i=1}^{K} \mathbb{1}\{E_{n_i}(-1)^{i+1} > 0\} - \frac{1}{2^{K-1}} \\ =&\frac{1}{2^{K-1}} \sum_{L=1}^{\lfloor \frac{K}{2} \rfloor} \sum_{\substack{m_1<...<m_{2L} \\ \subset\{n_1,...,n_K\}}} \prod_{j \in \mathcal{N}(m_1,...,m_{2L})} (-1)^j \prod_{i=1}^{2L} \Phi(E_{m_i}).\end{aligned}$$

wobei $\mathcal{N}(m_1, ..., m_{2L}) = \{j \in \{1, .., K\} \mid \exists i \in \{1, ..., 2L\} : m_i = n_j\}$ eine von den Laufindizes abhängige Menge und $\Phi(x) := \mathbb{1}\{x < 0\} - \mathbb{1}\{x > 0\}$ sind.

Für den Fall $K = 2$ wird dies bereits in Müller (2005) gezeigt und in Kustosz et al. (2016) wird eine äquivalente Formulierung für $K = 3$ verwendet. Die Darstellung für allgemeine K legt den Grundbaustein zur Herleitung der asymptotischen Verteilung für ein $K \geq 5$.

Literatur

Kustosz, C., Leucht, A. and Müller, C. (2016): Tests based on simplicial depth for AR(1) models with explosion. *Journal of Time Series Analysis* 37, 763-784.

Müller, C. (2005): Depth estimators and tests based on the likelihood principle with application to regression. *Journal of Multivariate Analysis* 95, 153-181.