Statistik II

- GLM (Regression und Varianzanalyse)
- Logistische Regression
- Loglineare Modelle
- Verallgemeinerte Lineare Modelle (GLIM)
  ”Generalised Linear Interactive Models”
- Glättungen — weder linear noch Modelle

Glättung (Smoothing)

- Glättung für empirische Verteilungen
  (Dichteschätzung)
  d.h. Darstellung der Verteilung einer stetigen Variable

- Glättung für Zeitreihen
  —um den Trend herauszuholen

- Glättung für nichtlineare Modellierung
  —weil es keine allgemeine parametrische Familie gibt

Die Daten werden gefiltert.

Warum Glättung?

1. Den Einfluß von Ausreißern entgegenzuwirken.

2. Modelle zu überprüfen
   (z.B. auf Linearität/Nichtlinearität).

3. Strukturänderungen hervorzuheben.

4. Lokale Variabilität zu schätzen.

Was bleibt nach einer Glättung wird der Smooth (Glättung) genannt. Der Rest wird Rough (die Residuen) genannt:

Data = Smooth + Rough

Gleitende Mediane (für Zeitreihen und ?)
(nicht gleitende Durchschnitte)

3-Median

$$z_t = median(x_{t-1}, x_t, x_{t+1})$$

(2m+1)-Median

$$z_t = median(x_{t-m}, x_{t-m+1}, \ldots, x_t, x_{t+1} \ldots, x_{t+m})$$

4-Median

$$z_{t-0.5} = median(x_{t-2}, x_{t-1}, x_t, x_{t+1})$$

wobei wir den Durchschnitt von den zwei Mittleren Werten der vier nehmen. Um einen Glättungswert für den Zeitpunkt t zu haben, nehmen wir

$$z_t = \frac{(z_{t-0.5} + z_{t+0.5})}{2}$$

das heißt einen 2-Median von z

2m-Median

$$z_{t-0.5} = median(x_{t-m}, x_{t-m+1}, \ldots, x_t, \ldots, x_{t+m-1})$$

Kombinationen von Glättungen

4253 bedeutet
4-Median,
dann 2-Median,
dann 5-Median,
dann 3-Median

Nach sovielen robusten Verfahren sollten die Ausreißer (ob einzelne oder Gruppen davon) ausgeschloßen werden. Als letzter Schritt kann man dann einen gleitenden gewichteten Durchschnitt anwenden, wie z. B. Hanning:

$$y_t = (z_{t-1} + 2z_t + z_{t+1})/4$$

Das Endresultat wird mit 4253H bezeichnet.

Manchmal hat das Rough immer noch Struktur. Deswegen macht man reroughing:

$$\begin{aligned} Data &= Smooth_1 + Rough_1 \\ Rough_1 &= Smooth_2 + Rough_2 \\ Data &= (Smooth_1 + Smooth_2) + Rough_2 \end{aligned}$$

So wird “4253H, twice” gemacht (was Data Desk benutzt).

Lowess (statt Regression)
(LOcally WEighted regression Scatterplot Smoothing)

Erste Glättung

1. Für jeden Punkt j nimmt man die f% nächstliegenden Pünkte. Man berechnet die größte Entfernung unter diesen Pünkten:

$$s_j = max(x_j - x_{min}, x_{max} - x_j)$$

2. Tricube Gewichte werden für jeden Punkt i um j berechnet

$$w_i = \begin{cases} (1 - |u_i|^3)^3 & \text{für } |u_i| \leq 1 \\ 0 & \text{sonst} \end{cases}$$

wo

$$u_i = (x_j - x_i)/s_j$$

3. Eine gewichtete lineare Regression um j wird berechnet. Der erste geglättete Wert für j ist die Regressionsvoraussage für j aus der gewichteten Regression um j.

Zweite (robuste) Glättung (nach Data Desk)

4. Nachdem man das für alle Pünkte gemacht hat, wird eine zweite Glättung durchgeführt, um die Wirkung von Extremwerten zu beseitigen. Zuerst werden die Residuen r_j berechnet und dann der Median der absoluten Residuen in der Nachbarschaft des Punkts j:

$$MAD_j = median_j(|r_i|)$$

5. Neue Gewichtsfaktoren werden berechnet

$$w_i = \begin{cases} (1-u_i^2)^2 & \text{für } |u_i| \leq 1 \\ 0 & \text{sonst} \end{cases}$$

wo $u_i = r_i / MAD_i / 6$

6. Neue gewichtete Lokal-Regressionen werden an jedem Punkt berechnet, wo als Gewichtung das Produkt der zwei Gewichte genommen wird. Der endgültige geglättete Wert für Punkt j wird aus dieser entsprechenden Regression genommen.

Erwünschte Eigenschaften von Glättungen

1. Eine Glättung sollte durch die Mitte der Daten gehen.

2. Eine Glättung sollte nie die Richtung scharf ändern.

3. Eine Glättung sollte lange Wellen haben und nicht viel wackeln.

4. Ausreißer und kurze, scharfe Änderungen sollen nicht beachtet werden.

Wie soll man glätten?

$$y_i = m(x_i) + \epsilon_i$$

mit

$$\begin{aligned} m(x) &= E[Y|X] \\ E[\epsilon|X] &= 0 \\ V[\epsilon|X] &= \sigma^2(x) \end{aligned}$$

$$\begin{aligned} => m(x) &= \int y f(y|x) dy \\ &= \frac{\int y f(x,y)}{f_X(x)} \end{aligned}$$

$f(y|x)$ ist die bedingte Dichte von Y gegeben X
$f(x,y)$ ist die gemeinsame Dichte von X und Y
$f_X(x)$ ist die marginelle Dichte von X

Die Dichten sind unbekannt und müssen geschätzt werden. Ein Kernschätzer für die eindimensionale Dichte einer Variable ist

$$f(x) = \frac{\sum_{i=1}^{n} K(\frac{x-x_i}{h})}{nh}$$

Die Kernfunktion K und die Bandbreite h müssen bestimmt werden.

Die Form eines Kernschätzers.

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} K(\frac{x-x_i}{h})}{nh}$$

$\hat{f}(x)$ ist eine Dichte $\Rightarrow$

$$\begin{aligned}
\int \hat{f}(x)dx &= 1 \\
\int \sum_{i=1}^{n} K(\frac{x-x_i}{h})dx &= nh \\
\int K(\frac{x-x_i}{h})dx &= h \qquad \forall i \\
\int K(v)dv &= 1 \qquad \text{für } v = (\frac{x-x_i}{h})
\end{aligned}$$

Beispiele von Kernfunktionen

1. **Gleichverteilt**

$$K(u) = \begin{cases} 0.5 & \text{für } |u_i| \leq 1 \\ 0 & \text{sonst} \end{cases}$$

2. **Epanechnikov**

$$K(u) = \begin{cases} 0.75(1-u^2) & \text{für } |u_i| \leq 1 \\ 0 & \text{sonst} \end{cases}$$

3. **Biweight**

$$K(u) = \begin{cases} \frac{15}{16}(1-u^2)^2 & \text{für } |u_i| \leq 1 \\ 0 & \text{sonst} \end{cases}$$

4. **Gauß**

$$K(u) = (2\pi)^{-\frac{1}{2}} e^{-\frac{u^2}{2}}$$

Eigenschaften von Kernfunktionen

$$\begin{aligned} K(u) &\geq 0 & \\ \int K(u)du &= 1 & \text{(Gewicht)} \\ \int uK(u)du &= 0 & \text{(Bias)} \end{aligned}$$

Wie gut ist eine Kernfunktion?

Wir können die L_2 Norm als Kriterium nehmen:

$$ISE = \int [\hat{f}(x) - f(x)]^2 dx$$

Es gibt auch die L_1 Norm oder die L_∞ Norm

$$sup|\hat{f}(x) - f(x)|$$

ISE könnte für einen bestimmten Datensatz und bekannten $f(x)$ interessant sein. Für unbekannte $f(x)$ betrachtet man statt dessen

$$E_f[ISE] = MISE$$

$$
\begin{aligned}
MSE(x) &= E[(\hat{f}(x) - f(x))^2] \\
&= Var(\hat{f}(x)) + Bias^2(\hat{f}(x)) \\
Bias(\hat{f}(x)) &= E[\hat{f}(x)] - f(x) \\
E[\hat{f}(x)] &= \int_{x-h}^{x+h} K(\frac{x-y}{h}) f(y) dy \\
&= \int_{-h}^{+h} K(\frac{u}{h}) f(x-u) du
\end{aligned}
$$

Für h klein ersetzen wir $f(x-u)$ mit

$$
f(x-u) = f(x) - uf'(x) + \frac{u^2}{2} f''(x) \ldots
$$

Für $n \to \infty$ und $h \to 0$

$$
\begin{aligned}
E[\hat{f}(x)] &= f(x) + \frac{h^2}{2} f''(x) \sigma_K^2 + O(h^4) \\
\sigma_K^2 &= \int_{-h}^{+h} u^2 K(\frac{u}{h}) du
\end{aligned}
$$

$$
\Rightarrow Bias(\hat{f}(x)) \simeq \frac{h^2}{2} f''(x) \sigma_K^2
$$

Auf ähnliche Weise kann man zeigen, dass

$$
Var(\hat{f}(x)) = \frac{f(x) \int K^2(u) du}{nh} + O(n^{-1})
$$