

Statistik II

- GLM (Regression und Varianzanalyse)
- Logistische Regression
- Loglineare Modelle
- Verallgemeinerte Lineare Modelle (GLIM)
"Generalised Linear Interactive Models"
- Glättungen
 - Lowess
 - Dichteschätzer (mit Kernfunktionen)
 - Regressionsglättungen
 - Glättungen und Modelle

Einfluß von Punkten in Glättungen

Jeder Punkt x_i beeinflusst die eigene Nachbarschaft $(x_i - h, x_i + h)$. Mit den meisten Kernfunktionen (d.h. außer der Gleichverteilungskernfunktion) schwächt dieser Einfluß mit zunehmender Entfernung ab.

Andersherum wird die geschätzte Dichte $\hat{f}(x)$ am Punkt x von allen Punkten in der Nachbarschaft $(x - h, x + h)$ beeinflusst.

Der Wert von $\hat{f}(x)$ hängt teils davon ab, wieviele Punkte in der Nachbarschaft liegen und teils davon ab, wie nah sie an x liegen.

Momente von Kerndichteschätzern

Gegeben sei der Datensatz $\{x_1, x_2, \dots, x_n\}$ mit Schätzern

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\hat{\sigma}^2 = s^2 = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2$$

Eine approximierende Normalverteilung $N(\mu, \sigma^2)$ hätte diese Stichprobenwerte als Parameter nehmen können. Was sind μ und σ^2 von einem Kerndichteschätzer?

$$\hat{\mu} = \int x \hat{f}(x) dx$$

$$\hat{f}(x) = \frac{1}{nh} \sum K\left(\frac{x - x_i}{h}\right)$$

$$\hat{\mu} = \frac{1}{n} \sum \frac{1}{h} \int x K\left(\frac{x - x_i}{h}\right) dx$$

Nach der Transformation $u = \frac{x - x_i}{h}$ gilt

$$\hat{\mu} = \frac{1}{n} \sum \int_{-1}^1 (x_i + hu) K(u) du$$

$$\Rightarrow \hat{\mu} = \bar{x}$$

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n} \sum \frac{1}{h} \int (x - \bar{x})^2 K\left(\frac{x - x_i}{h}\right) dx \\
&= \frac{1}{n} \sum \int_{-1}^1 (x_i + hu - \bar{x})^2 K(u) du \\
&= \frac{1}{n} \sum \int_{-1}^1 [(x_i - \bar{x}) + hu]^2 K(u) du \\
&= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{1}{n} h^2 \sum \sigma_K^2 \\
&= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 + h^2 \sigma_K^2
\end{aligned}$$

Streudiagramme und Glättungen bzw. Dichteschätzer

Wir müssen unterscheiden zwischen Modellglättungen für das Verhältnis zwischen Y und X (wie Lowess oder Nadaraya-Watson) und bivariate Dichteschätzer.

Für ein Modell schätzen wir eine $1 - d$ Funktion (z.B. Nadaraya-Watson):

$$E[Y|X] = m_{NW}(x) = \frac{\sum y_i K_x\left(\frac{x-x_i}{h_x}\right)}{\sum K_x\left(\frac{x-x_i}{h_x}\right)}$$

Für eine $2 - d$ Dichte könnten wir verschiedene Schätzer einsetzen, z.B. (wie in der Ableitung von Nadaraya-Watson) die Produktkernfunktion:

$$\frac{\sum K_x\left(\frac{x-x_i}{h_x}\right) K_y\left(\frac{y-y_i}{h_y}\right)}{nh_x h_y}$$

Wenn genug Daten zur Verfügung stehen, könnten wir echte $2 - d$ Dichteschätzer einsetzen:

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

wo H eine $d \times d$ nichtsinguläre Matrix ist und $K : \mathbb{R}^d \rightarrow \mathbb{R}^1$ eine Kernfunktion.

Binning

Um große Datenmengen in Streudiagrammen darzustellen, wird Binning vorgeschlagen. Der Plot wird in kleinen Gebieten ("Bins") aufgeteilt und die Anzahl Punkte in jedem Bin gezählt.

Die Bins werden nach den Häufigkeiten durch Grauschattierungen oder Farbe gezeichnet, um die zweidimensionale Dichte darzustellen. Binning ist schnell, aber grob. Die künstlichen Binsgrenzen werden oft unnötwendigerweise hervorgehoben.

Um die Resultate von Binningverfahren zu glätten, hat Scott eine "**Average Shifted**" (durchschnittlich verschoben) Methode vorgeschlagen. Sei (x_0, y_0) die untere linke Ecke des ersten Bins.

$$x_0 = \min x - b_x \text{ und } y_0 = \min y - b_y$$

Für rechteckige Bins haben alle andere Bins untere linke Ecke der Form $(x_0 + ib_x, y_0 + jb_y)$.

Binning wird D^2 mal durchgeführt, wo der erste Ursprung in der $d = d_1 * d_2$ ten Schätzung

$$(x_0 + b_x * (d_1 - 1)/D, y_0 + b_y * (d_2 - 1)/D)$$

wird. Man nimmt den Durchschnitt der D^2 Dichteschätzungen an jedem Punkt.

(Nach Stigler (1986) hat Galton das schon Ansatzweise 1886 (!) in seiner Tabelle von Größen von Kindern und ihren Eltern gemacht.)

Average Shifted Histogramme

Die Idee des Verschiebens kann man leichter im eindimensionalen Fall sehen. Betrachten wir eine Reihe von Histogrammschätzern $\hat{f}_1, \hat{f}_2, \dots, \hat{f}_m$, jede mit Binbreite h , aber mit Ankerpunkten

$$t_0 = 0, \frac{h}{m}, \frac{2h}{m}, \dots, \frac{(m-1)h}{m}$$

Dann wird unser ASH (Average Shifted Histogramm) Schätzer

$$\hat{f}_{ASH}(x) = \frac{1}{m} \sum_{i=1}^m \hat{f}_i(x)$$

Es bleiben immer noch Sprünge im Schätzer, in der Tat mehr, aber kleinere. $\hat{f}_{ASH}(x)$ ist stückweise konstant über die Intervalle $[k\delta, (k+1)\delta]$ für $\delta = h/m$. Es ist sinnvoll, dieses kleinere Intervall als den "Bin" zu nehmen und die Häufigkeiten der Daten für diese Bins zu berechnen, ν_k für $B_k = [k\delta, (k+1)\delta]$. Daraus ergibt sich

$$\begin{aligned}\hat{f}(x; m) &= \frac{1}{m} \sum_{i=1-m}^{m-1} \frac{(m - |i|)\nu_{k+i}}{nh} \\ &= \frac{1}{nh} \sum_{i=1-m}^{m-1} \left(1 - \frac{|i|}{m}\right) \nu_{k+i} \quad \text{für } x \in B_k\end{aligned}$$

Die Gewichte für die Häufigkeiten nehmen eine gleichschenkelige Dreiecksform an. Andere Gewichtungen wären möglich.

2 – d Dichteschätzer

Verwirklichungen in Software

In R oder in deren Erweiterungen gibt es Produktkernfunktionmethoden und Binning mit FFT (Fast Fourier Transform). Statische Darstellungen mit Konturen oder nach der geschätzten Dichte schattierten Plots werden angeboten.

Die Software MANET benutzt eine kreisförmige Gleichverteilungskernfunktion, um eine graphische Darstellung einer $2-d$ Dichte anzubieten. An jedem Pixel hängt die Helligkeit von der Dichte ab. (In der gegenwärtigen Version gibt es eine obere Grenze von $L(\leq 32)$ Punkten. Höhere Dichten können nicht diskriminiert werden.) Die Größe des Kreises kann interaktiv geändert werden.

[Der Kreis ist selbstverständlich kein richtiger Kreis, da wir mit Bildschirmpixeln arbeiten müssen. Was sollte passieren, wenn ein Pixel vergrößert wird?]

MONDRIAN bietet ähnliches aber mit Einsatz des α Kanals. Binning wird auch teils angeboten.

GAM

Eine Vereinfachung der allgemeinen Form, die gleichzeitig nichtlineare Funktionen erlaubt, bilden GAM Modelle.

Generalised Additive Models

$$y = m(x) = \alpha + \sum f_j(x_j)$$

d.h. keine Interaktion zwischen den erklärenden Variablen.

Die $f_j(x_j)$ werden iterativ und zyklisch mit einem univariaten Glättungsverfahren über die Residuen geschätzt, zuerst $f_1(x_1)$, dann $f_2(x_2)$ über $f_p(x_p)$, um dann wieder bei $j = 1$ anzufangen, und weiter bis Konvergenz.

Die Vorteile dieser Verallgemeinerung sind klar.

Es gibt aber Nachteile:

- 1) Der Prozeß ist rechenintensiv.
- 2) Die Behandlung von "Ausreißern" ist unklar.
- 3) Es wird kein analytisches Modell gefunden.

Ein anderer Ansatz ist multivariater Lowess

Hier werden Interaktionsterme zwischen den Variablen berücksichtigt, aber die Form der lokalen Funktionen wird stark eingeschränkt (linear mit einfachen Interaktionen der Art $x_j x_k$).

Die Entfernung zwischen Punkten wird Euklidisch gemessen, nachdem man standardisiert hat (meistens mit einer robusten Schätzung der Standardabweichungen). In allen solchen multivariaten Verfahren ist es wichtig, dass die Skalierung der Variablen vergleichbar gemacht wird.

Modellkriterien und Glättungen

Der Vergleich zwischen Modellen erweist sich als schwierig. Wie sollen Ausreißer berücksichtigt werden? Wieviele Parameter sollte man einer Glättung zuschreiben, um Freiheitsgrade zu bestimmen?

Graphische Überprüfung von Modellen.

Gegeben sind Daten

$$\{y_i; \{x_{ij}\}_{j=1,\dots,p}\}_{i=1,\dots,n}$$

und ein Modell

$$y = g(\{x_{ij}\})$$

Die Grundidee ist, dass, wenn ein Modell in allen 2-dimensionalen Projektionen gut angepasst wird, wird es den Daten in den $(p + 1)$ Dimensionen gut anpassen.

Glättungsmethode zur Modellüberprüfung

Die empfohlene Methode ist dann, viele Projektionen schnell durchzulaufen.

In jeder Projektion wird

1. y gegen eine lineare Kombination der x Variablen, $\sum a_j x_j$ ($\sum a_j^2 = 1$), geplottet.
2. eine Glättung (z.B. Lowess) berechnet.
3. \hat{y} gegen die selbe lineare Kombination geplottet.
4. eine Glättung für den \hat{y} Plot berechnet.

Dann vergleicht man die Glättungen. In Data Desk kann man das besonders schön mit Sliders für die Koeffizienten $\{a_j\}$ machen. Es wäre noch besser, wenn beide Glättungen im selben plot gezeigt werden könnten.