Dr. Katharina Best
Sommersemester 2011
5. Mai 2011

ulm university universität **uulm**

Reader – Teil 4: Regression

WiMa-Praktikum

In der Regressionsanalyse wird ein funktionaler Zusammenhang

$$y = h(x) + z \tag{1}$$

gesucht. Dabei ist z der Fehler, der als stochastisch angenommen wird. Wir betrachten zwei Datensätze, denen wir diesen Zusammenhang unterstellen, wobei der erste Datensatz als *erklärend*, der zweite als *erklärt* angesehen wird, d. h. wir legen auch eine richtung des Zusammenhangs fest. Aus diesem Grund betrachten wir verbundene Stichproben und untersuchen die folgende Situation.

Ist $\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} X_1, \dots X_n \\ Y_1, \dots, Y_n \end{pmatrix}$ die Stichprobe, so wollen wir eine Funktion $h \in H$ finden, wobei wir die Suche auf die Funktionenmenge H beschränken. Dazu wird die Menge geeignet parametrisiert, d. h. wir befinden uns im Bereich der parametrischen Schätzung.

Des weiteren wird ein geeignetes *Abstandsmaß* bzw. *Fehlerfunktion* für die Abweichung der Stichprobe der erklärten Variable von den theoretischen Werten definiert.

1 Lineare Regression

Bei der linearen Regression haben wir als Grundmenge der Funktionen

$$H = \left\{ x \mapsto \alpha + \beta x : (\alpha, \beta) \in \mathbb{R}^2 \text{ für } x \in \mathbb{R} \right\} \tag{2}$$

gegeben. Für die Zufallsvariablen haben wir dann

$$Y = h(X) + Z = \alpha + \beta x + Z \tag{3}$$

und wir nennen Y die erklärte Variable, x die erklärende Variable, Z der Fehler. Insofern wird häufig x als vorgegeben, Y jedoch als zufällig angesehen. Der Fehler Z wird im Standardmodell als normalverteilt angenommen, d. h.

$$\mathbb{E}Z = 0, \qquad \text{d. h. } \mathbb{E}Y = h(x), \tag{4}$$

$$\mathbb{V}\mathrm{ar}Z = \sigma^2 \qquad \text{(unabhängig von } x\text{)}. \tag{5}$$

Bei der linearen Regression wird mit Hilfe von Kleinste-Quadrate-Schätzern versucht, die Daten an eine Gerade anzupassen, wobei die Quadrate der vertikalen Abstände minimiert werden. Die entsprechende Fehlerfunktion ist dann die Summe der quadratischen Abweichungen

$$Q(h_{\alpha,\beta}) := \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i - h_{\alpha,\beta}(x_i)\Big)^2 = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - (\alpha+\beta x_i)\big)^2 \tag{6}$$

für $h_{\alpha,\beta} \in H$, welche minimiert wird. Damit erhalten wir die Schätzer

$$\begin{aligned}\hat{\alpha}_n(X,Y) &= \overline{Y}_n - \hat{\beta}\overline{x}_n\\ \hat{\beta}_b(X,Y) &= \frac{\sum_{i=1}^{n}(x_i-\overline{x}_n)(Y_i-\overline{Y}_n)}{\sum_{i=1}^{n}(x_i-\overline{x}_n)^2} = r_{x,Y,n}\cdot\frac{\widetilde{S}_Y}{\widetilde{S}_x}\end{aligned}$$

für die Parameter der Geraden, wobei

$$\begin{aligned}r_{x,Y,n} &= \frac{\sum_{i=1}^{n}(x_i-\overline{x}_n)(Y_i-\overline{Y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x}_n)^2\cdot\sum_{i=1}^{n}(Y_i-\overline{Y}_n)^2}}\\ \widetilde{S}_{x,n} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x}_n)^2} \qquad \text{resp.} \qquad \widetilde{S}_{Y,n} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\overline{Y}_n)^2}.\end{aligned}$$

Um die Güte eines Regressionsmodells zu beurteilen, bezeichnen wir den Fehler

$$\hat{Z}_i = Y_i - \hat{Y}_i$$

als *Residuum*, welches wir weiter untersuchen. Damit können wir die Streuung zerlegen. Wir sagen

$$\underbrace{SQT}_{\text{total sum of squares}} = \underbrace{SQE}_{\text{explained sum of squares}} + \underbrace{SQR}_{\text{residual sum of squares}} \tag{7}$$

wobei

$$\begin{aligned}SQT &= \sum_{i=1}^{n}(Y_i-\overline{Y}_n)^2 = n\cdot\widetilde{S}_{Y,n}^2\\ SQE &= \sum_{i=1}^{n}(\hat{Y}_i-\overline{Y}_n)^2\\ SQR &= \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2\end{aligned}$$

Das *Bestimmtheitsmaß*

$$R^2 = \frac{SQE}{SQT} = 1 - \frac{SQR}{SQT}$$

beschreibt den Anteil der durch das Modell erklärten Varianz. Im Standardmodell ist

$$\hat{\sigma}^2(X,Y) := \frac{SQR}{n-2}$$

ein erwartungstreuer Schätzer für σ^2.

2 Multiple lineare Regression

Die multiple lineare Regression verallgemeinert den obigen Fall. Die erklärte Variable y hängt dann nicht von einer erklärenden Variable x, sondern von $x_1, \ldots, x_r$, die untereinander unabhängig sind. Nach wie vor suchen wir jedoch eine lineare Anpassung, d. h. die Grundmenge der Funktionen ändert sich zu

$$H = \left\{ (x_1, \ldots, x_r) \mapsto \alpha_0 + \alpha_1 x_1 + \ldots + \alpha_r x_r : (\alpha_0, \ldots, \alpha_r) \in \mathbb{R}^{r+1} \text{ für } x_i \in \mathbb{R} \right\}. \tag{8}$$

Wir wollen nun die Parameter $\alpha_0, \ldots, \alpha_r$ schätzen. Dazu erweitern wir das obige Verfahren zu einer vektoriellen Schreibweise und setzen

$$\boldsymbol{Y} := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \boldsymbol{\alpha} := \begin{pmatrix} \alpha_0 \\ \vdots \\ y_r \end{pmatrix} \qquad \text{und} \qquad \boldsymbol{X} := \begin{pmatrix} 1 & x_{11} & \ldots & x_{1r} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nr} \end{pmatrix} \tag{9}$$

Mit dem euklidischen Abstand $\|v\| = (\boldsymbol{v}^T \boldsymbol{v})^2 = \sqrt{v_1^2 + \ldots + v_p^2}$ für ein $\boldsymbol{v} \in \mathbb{R}^p$ untersuchen wir die Fehlerfunktion

$$Q(h_{\boldsymbol{\alpha}}) := \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - h_{\boldsymbol{\alpha}}(x_{i1}, \ldots, x_{in}) \right)^2 = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \alpha_0 - \alpha_1 x_{i1} - \ldots - \alpha_r x_{ir} \right)^2 = \frac{1}{n} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha}\|^2 \tag{10}$$

Sind die Spalten von $\boldsymbol{X}$ linear unabhängig (vorausgesetzt $r \leq n - 1$), so hat $\boldsymbol{X}^T \boldsymbol{X}$ den vollen Rang $r + 1$ und das Gleichungsystem

$$\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\alpha} = \boldsymbol{X}^T \boldsymbol{Y} \tag{11}$$

wird von

$$\hat{\boldsymbol{\alpha}} = \left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{Y} \tag{12}$$

gelöst. Die Streuungszerlegung und das Bestimmtheismaß ergeben sich wie oben.

3 Quantile Regression

Bei der quantilen Regression minimieren wir über eine Fehlerfunktion, welche den absoluten Abstand einbezieht, deswegen auch als *LAD (least absolute deviation)* bezeichnet. Wir haben

$$L(h_{\alpha,\beta}) := \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - h_{\alpha,\beta}(x_i) \right| = \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - (\alpha + \beta x_i) \right| = \frac{1}{n} \sum_{i=1}^{n} \left| Y_i - \hat{Y}_i \right| \tag{13}$$

Die Lösung des Minimierungsproblems ist dann der Median. Entsprechend können auch andere Quantile geschätzt werden

Wegen ihrer Robustheit gewinnt dieses Verfahren immer mehr an Bedeutung.

4 Kategoriale Regression

Wollen wir uns kategoriale Daten einer Regession unterziehen, d. h. gehen wir davon aus, dass sich die Wahrscheinlichkeit einer Zufallsvariablen Z, einen bestimmten Wert anzunehmen durch eine Reihe von Effekten erklären lässt, benötigen wir eine Anpassung der bereits vorgestellten Modelle.

Dazu müssen wir zuerst die Ausgänge der erklärenden Zufallsvariablen kodieren. Die häufigste, wenn auch nicht die einzige Art der Kodierung ist die *Dummy-Kodierung*. Bei binären Zufallsvariablen, also bei nur zwei möglichen Ausgängen, werden diese selbst mit 0 resp. 1 kodiert. Besitzt die Zufallsvariable X_A, die das Merkmal A angibt, etwa r verschiedene Ausprägungen, so werden aus der Zufallsvariablen $r-1$ binäre Zufallsvariable X_j^A gemacht,

$$X_j^A(\omega) = \begin{cases} 1, & \text{falls } X_A(\omega) \text{ die } j\text{-te Ausprägung aufweist,} \\ 0, & \text{sonst} \end{cases} \tag{14}$$

für $1 \leq j \leq r-1$. Wir haben dann den Zufallsvektor $(X_1^A, \ldots, X_{r-1}^A)^T$. Entsprechend lässt sich bei k erklärenden Variablen $X^{A_1}, \ldots, X^{A_k}$ der Vektor

$$\left(X_1^{A_1}, \ldots, X_{r_{A_1}-1}^{A_1}, \ldots, X_1^{A_k}, \ldots, X_{r_{A_k}-1}^{A_k}\right)^T$$

definieren. Um die Analogie zu oben zu erhalten, betrachten wir sogar den Vektor

$$X := \left(1, X_1^{A_1}, \ldots, X_{r_{A_1}-1}^{A_1}, \ldots, X_1^{A_k}, \ldots, X_{r_{A_k}-1}^{A_k}\right)^T. \tag{15}$$

Der Vektor $\boldsymbol{\alpha}$ der *Haupteffekte* ist entsprechend notiert,

$$\boldsymbol{\alpha} = \left(\alpha_0, \alpha_1^{A_1}, \ldots, \alpha_{r_{A_1}-1}^{A_1}, \ldots, \alpha_1^{A_k}, \ldots, \alpha_{r_{A_k}-1}^{A_k}\right)^T$$

und wir haben

$$Y := \mathbb{P}(Z = 1|X) = X^T\boldsymbol{\alpha} \tag{16}$$

als das *lineare Haupteffekt-Modell*.

Für die n beobachteten Realisationen von X definieren wir

$$\boldsymbol{X} := \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \text{und} \quad \boldsymbol{Y} := \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \mathbb{P}(Z = 1|X = x_1) \\ \vdots \\ \mathbb{P}(Z = 1|X = x_n) \end{pmatrix} \tag{17}$$

und erhalten als Analogie zu oben

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\alpha}. \tag{18}$$

Diese lineare Modell lässt sich mittels einer *Linkfunktion* $g : (0,1) \to \mathbb{R}$, definiert durch

$$g(Y) = X^T\boldsymbol{\alpha}$$

weiter verallgemeinern. Schätzen wir $\boldsymbol{Y}$ mittels der relativen Häufigkeiten der Eintritte der Merkmalsausprägungen an den Beobachtungen, benannt $\hat{\boldsymbol{Y}}$, so lässt sich mit Hilfe der Kovarianzmatrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}$ von $\hat{\boldsymbol{Y}}$ der *verallgemeinerte Kleinste-Quadrate-Schätzer*

$$\hat{\boldsymbol{\alpha}} = \left(\boldsymbol{X}^T\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}}}^{-1}\hat{\boldsymbol{Y}} \tag{19}$$

definieren.