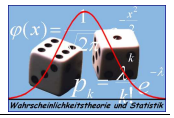
Institut für Mathematik,
Lehrstuhl für Wahrscheinlichkeitstheorie und Statistik



Regressionsanalyse

Problemstellung: $X,\ Y$ zufälliger Vektor über $[\Omega, \mathcal{F}, P]$. Gesucht ist eine funktionale Abhängigkeit zwischen X und Y (z. B. Druck und Temperatur, Geschwindigkeit und Spritverbrauch). Gesucht ist also eine Funktion $g : \mathbb{R} \longrightarrow \mathbb{R}$ mit

$$\mathbb{E}(Y - g(X))^2 = \inf_f \mathbb{E}(Y - f(X))^2.$$

Diese Funktion heißt *Regressionsfunktion*. Im Idealfall gilt $Y = g(X)$.

Praxis: Es wird die konkrete Stichprobe $(x_1, y_1), \ldots, (x_n, y_n)$ gemessen. Man ermittle eine Funktion $f : \mathbb{R} \longrightarrow \mathbb{R}$, so daß $\sum_{k=1}^n (y_k - f(x_k))^2$ sehr klein wird.

Theoretisches Intermezzo:[1] **1.** $(X,\ Y)$ diskreter zufälliger Vektor über $[\Omega, \mathcal{F}, P]$ mit $X(\Omega) = \{x_1, x_2, \ldots\}$ bzw. $Y(\Omega) = \{y_1, y_2, \ldots\}$.
Definition: Für $y_m \in Y(\Omega)$ ist $X|Y = y_m$ die Zufallsgröße mit Verteilung

$$P(X = x_k | Y = y_m) = \frac{P(X = x_k, Y = y_m)}{P(Y = y_m)}$$

Der Erwartungswert von $X|Y = y_m$ heißt *bedingter Erwartungswert von X unter der Bedingung $Y = y_m$*. Es gilt für $y_m \in Y(\Omega)$

$$\mathbb{E}(X|Y = y_m) = \sum_k x_k \cdot P(X = x_k | Y = y_m)$$

Die Funktion $\Psi : Y(\Omega) \Longrightarrow \mathbb{R},\ \Psi(y) := \mathbb{E}(X|Y = y)$ heißt *Erwartungswertfunktion* von $X|Y$.

2. Sei $(X,\ Y)$ stetiger zufälliger Vektor über $[\Omega, \mathcal{F}, P]$ mit Dichte $f(x, y)$. Für $y \in \mathbb{R}$ mit $f_Y(y) > 0$ bezeichnet $X|Y = y$ die zufällige Größe mit Dichte

$$f_{X|Y=y}(x) := \frac{f(x, y)}{f_Y(y)}$$

Der Erwartungswert von $X|Y = y$ heißt *bedingter Erwartungswert von X unter der Bedingung $Y = y$*. Es gilt für y mit $f_Y(y) > 0$

$$\mathbb{E}(X|Y = y) = \int_{-\infty}^{\infty} \frac{f(x, y)}{f_Y(y)} dx$$

Die Funktion $\Psi : Y(\Omega) \Longrightarrow \mathbb{R},\ \Psi(y) := \mathbb{E}(X|Y = y)$ heißt *Erwartungswertfunktion* von $X|Y$.

[1] Im folgenden wird die Existenz der auftretenden Erwartungswerte stets stillschweigend vorausgesetzt.

Satz: $(X,\ Y)$ zufälliger Vektor über $[\Omega, \mathcal{F}, P]$, $\Psi(y) = \mathbb{E}(X|Y = y)$, $\Phi(x) = \mathbb{E}(Y|X = x)$. $\Longrightarrow$

$$\mathbb{E}(Y - \Phi(X))^2 = \inf_f \mathbb{E}(Y - f(X))^2$$

$$\mathbb{E}(X - \Psi(Y))^2 = \inf_f \mathbb{E}(X - f(Y))^2.$$

Beispiel: (X, Y) zufälliger Vektor mit $X \geq 0,\ Y \geq 0$ und Dichte

$$f(x, y) = \frac{a^p}{\Gamma(p)} \cdot y^p e^{-(a+x)y} \qquad (x \geq 0, y \geq 0),$$

wobei $a > 0,\ p > 0$.
$\Longrightarrow$
$Y \sim \mathrm{Gam}(a, p)$ und X hat eine sog. *Pareto*-Verteilung: für $x \geq 0$ gilt

$$f_X(x) = \frac{p \cdot a^p}{(a + x)^{p+1}}$$

$\Longrightarrow$ für $x \geq 0,\ y \geq 0\ \ f_{X|Y=y}(x) = y \cdot e^{-yx}$, d. h. $X|Y = y \sim \mathrm{Exp}(y)$.
Man erhält die *Regressionsfunktion* $\Psi(y) = \mathbb{E}(X|Y = y) = \frac{1}{y}$

Definition: (X, Y) zufälliger Vektor mit $\sigma(X) = \sqrt{\mathrm{Var}(X)} > 0,\ \sigma(Y) = \sqrt{\mathrm{Var}(Y)} > 0$. Die Zahl

$$\varrho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma(X)\sigma(Y)}$$

heißt *Korrelationskoeffizient*.

Satz: $|\varrho(X, Y)| = 1 \Longleftrightarrow \exists a,\ b \in \mathbb{R} \quad P(Y = aX + b) = 1$

Folgerung: $|\varrho(X, Y)| = 1 \Longrightarrow \Psi(x) = ax + b$.

Definition: (X, Y) zufälliger Vektor. Die zufällige Größe $\alpha X + \beta$ heißt *Regressionsgerade* von Y bezüglich X, falls

$$\mathbb{E}(Y - (\alpha X + \beta))^2 = \inf_{a,\ b n \mathbb{R}} \mathbb{E}(Y - (aX + b)^2$$

Satz:

$$\beta = \mathbb{E}Y - \alpha \mathbb{E}X, \qquad \alpha = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} = \frac{\mathbb{E}(XY) - \mathbb{E}X \cdot \mathbb{E}Y}{\mathbb{E}(X^2) - (\mathbb{E}X)^2} = \varrho \cdot \frac{\sigma(Y)}{\sigma(X)}$$

Praktische Berechnung:
(X,Y) zufälliger Vektor,
$(X_1,Y_1),\ldots,(X_n,Y_n)$ Stichprobe, $(x_1,y_1),\ldots,(x_n,y_n)$ konkrete Stichprobe.
Im obigen Satz ersetze man

$$\begin{aligned}
\mathbb{E}X \quad &\text{durch} \quad \overline{x} = \frac{1}{n}\sum_{k=1}^{n} x_k \\
\mathbb{E}Y \quad &\text{durch} \quad \overline{y} = \frac{1}{n}\sum_{k=1}^{n} y_k \\
\sigma^2(X) \quad &\text{durch} \quad s_x^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})^2 \\
\sigma^2(Y) \quad &\text{durch} \quad s_y^2 = \frac{1}{n}\sum_{k=1}^{n}(y_k-\overline{y})^2 \\
\mathrm{Cov(X,Y)} \quad &\text{durch} \quad s_{xy} = \frac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y}) \\
\varrho(X,Y) \quad &\text{durch} \quad \frac{s_{xy}}{s_x\cdot s_y} \\
\mathbb{E}(XY) \quad &\text{durch} \quad \overline{xy} = \frac{1}{n}\sum_{k=1}^{n} x_k y_k \\
\mathbb{E}(X^2) \quad &\text{durch} \quad \overline{x^2} = \frac{1}{n}\sum_{k=1}^{n} x_k^2
\end{aligned}$$

Es gilt

$$s_x^2 = \overline{x^2} - (\overline{x})^2 \qquad s_{xy} = \overline{xy} - \overline{x}\cdot\overline{y}$$

Parameter der Regressionsgerade:

$$\beta = \overline{y} - \alpha\overline{x}, \qquad \alpha = \frac{\overline{xy}-\overline{x}\cdot\overline{y}}{\overline{x^2}-(\overline{x})^2} = \frac{s_{xy}}{s_x^2}$$

Beispiel: *Galton* benutzt 1889 erstmals den Begriff Regression, und zwar im sog. *Galtonschen Vererbungsgesetz:* Eltern, die hinsichtlich eines Merkmals vom Durchschnitt abweichen, haben Kinder, die in der gleichen Richtung abweichen, aber abgeschwächt (Selbstheilungskräfte der Natur - s. auch *Mendelsche Gesetze*). *Pearson* (1903) untersuchte 1 078 (Vater, Sohn)–Paare hinsichtlich der Körpergröße:
X - Körpergröße Vater, Y - Körpergröße Sohn. Es ergibt sich die folgende Gerade, die er erstmalig *Regressionsgerade* nannte:

$$y = 0,516x + 85,674 \qquad [cm]$$

$$\begin{aligned}
x &= 200 \Longrightarrow y = 189 \\
x &= 170 \Longrightarrow y = 173 \\
x &= 150 \Longrightarrow y = 163 \\
x &= 177 \Longrightarrow y = 177
\end{aligned}$$

Messung der Güte der Anpassung:
Sei $\varphi : \mathbb{R} \longrightarrow \mathbb{R}$, (X,Y) zufälliger Vektor.

$$U = \frac{\mathbb{E}(Y-\varphi(X))^2}{\mathrm{Var}(Y)}$$

ist Maß für die Abweichung zwischen Y und $\varphi(X)$.

Definition:

$$U = \frac{\mathbb{E}(Y-(\alpha X+\beta))^2}{\mathrm{Var}(Y)}$$

heißt *Unbestimmtheitsmaß* der Regressionsgerade $y = \alpha x + \beta$,
$B = 1 - U$ heißt *Bestimmtheitsmaß.*

Satz:

$$B = \varrho^2(X,Y) = \frac{\mathrm{Cov}^2(X,Y)}{\sigma^2(X)\sigma^2(Y)}$$

Definition:

$$r = \frac{s_{xy}}{s_x\cdot s_y} = \frac{\overline{xy}-\overline{x}\cdot\overline{y}}{\sqrt{(\overline{x^2}-\overline{x}\cdot\overline{x})(\overline{y^2}-\overline{y}\cdot\overline{y})}}$$

heißt *empirischer Korrelationskoeffizient.*

Anmerkung: r ist ein Maß für die Güte der Anpassung der Regressionsgerade:
$|r|$ "nahe" an 1 $\Longrightarrow$ gute Anpassung,
r "nahe" an 1 $\Longrightarrow$ gute Anpassung, Regressionsgerade wachsend,
r "nahe" an -1 $\Longrightarrow$ gute Anpassung, Regressionsgerade fallend,
$r = 0$ $\Longrightarrow$ X, Y unkorreliert.

Nichtlineare Regression

Häufig verwendete Regressionsfunktionen:

1) lineare Regression	$f(x) = ax + b$
2) polynomiale Regression	$f(x) = a_0x^m + a_1x^{m-1} + \ldots + a_m$
3) exponentielle Regression	$f(x) = a\cdot b^x$
4) geometrische Regression	$f(x) = a\cdot x^b$
5) logarithmische Regression	$f(x) = a\cdot\ln x + b$

Beispiel: Sei $\tilde{x} = \ln x$. Bestimme die Regressionsgerade $y = \alpha\cdot\tilde{x} + \beta$. $\Longrightarrow$

$$\beta = \overline{y} - \alpha\cdot\overline{\ln x} \qquad \alpha = \frac{\overline{y\cdot\ln x} - \overline{y}\cdot\overline{\ln x}}{\overline{\ln^2 x} - \overline{\ln x}\cdot\overline{\ln x}}$$