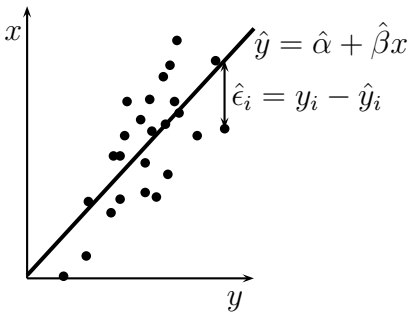
Handout zum Vortrag

Lineare Regression

Problemstellung Es soll der Zusammenhang zwischen zwei Variablen x und y beschrieben werden, deren Wertepaare (x_i, y_i) mit $i \in \{1, \ldots, n\}$ und $n \geq 2$ eine Punktwolke bilden und annähernd auf einer Geraden liegen. Die Variable x wird hierbei als *Regressor*, y als *Regressand* bezeichnet. Gesucht wird die (Regressions-)Gerade, die am nächsten an den Wertepaaren liegt. Dabei versucht man möglichst viel von der Streuung der Punkte zu beschreiben.

Ansatz Da die Punkte annähernd auf einer Gerade liegen, geht man von einem linearen Zusammenhang aus: $y \approx \alpha + \beta x$



Dieser Zusammenhang wird von einer Störgröße ϵ_i überlagert. Es soll nun die additive Konstante $\hat{\alpha}$ und der Regressionskoeffizient $\hat{\beta}$ geschätzt werden, sodass sie möglichst viel der Streuung beschreiben. Die Störgröße $\hat{\epsilon}_i$ erfasst die Streuung die von der Regressions-/Ausgleichsgeraden nicht beschrieben werden kann:

$$\hat{\epsilon}_i = y_i - \hat{\alpha} + \hat{\beta} x_i$$

Um die Werte von $\hat{\alpha}$ und $\hat{\beta}$ zu schätzen, minimiert man die Funktion welche die Summe von $\hat{\epsilon}_i$ beschreibt.

Methode der kleinsten Quadrate Diese Summe lässt sich durch den Abstand der Vektoren $\vec{v} \in \mathbb{R}^n$ mit den Einträgen $v_i = \hat{y}_i$ und $\vec{y} \in \mathbb{R}^n$ mit den Einträgen y_i beschreiben und ist von den zu schätzenden Variablen α und β abhängig:

$$D(\alpha, \beta) = \|\vec{v} - \vec{y}\| = \sqrt{\sum_{i=0}^{n} (v_i - y_i)^2}$$

Die Funktion $D(\alpha, \beta)$ kann durch $d(\alpha, \beta) = D(\alpha, \beta)^2$ vereinfacht werden, da hierbei das Minimum erhalten bleibt. Durch partielle Differentiation von d erhält man die beiden Gleichungen:

$$\begin{aligned} \frac{\partial d}{\partial \alpha} &= \sum_{i=1}^{n} 2(\hat{\alpha} + \hat{\beta} x_i - y_i) \\ \frac{\partial d}{\partial \beta} &= \sum_{i=1}^{n} 2(\hat{\alpha} + \hat{\beta} x_i - y_i) x_i \end{aligned}$$

Nun lässt sich mit Hilfe eines LGS unter der Bedinung $\nabla h = 0$ das Minimum berechnen. Um sicher zu gehen, dass es kein Maximum oder Sattelpunkt ist, müsste die zweite Ableitung mithilfe einer Hesse-Matrix überprüft werden. Die Lösungen des LGS sind:

$$\begin{aligned} \hat{\beta} &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\ \hat{\alpha} &= \bar{y} - m\bar{x} \end{aligned}$$

Güte der Regressionsgeraden Um die Genauigkeit einer Regressiongeraden bestimmen zu können, berechnet man den durch die Gerade erklärten Teil der Streuung (**s**um of **s**quares **e**xplained) und die gesamte Streuung (**s**um of **s**quares **t**otal). Das Verhältnis von erklärter Streuung und gesamter Streuung, bezeichnet man als *Bestimmtheitsmaß* oder *Determinationskoeffizient*. Es berechnet sich durch:

$$R^2 = \frac{SQE}{SQT} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}$$

Pascal Wittmann