Beispiel (Lineare Regression) Die Regressionsanalyse beschäftigt sich damit, eine oder mehrere abhängige Variablen durch eine oder mehrere unabhängige Variablen modulo des sogenannten Residuums zu erklären. Lineare Regression bildet dabei die einfachste Klasse der Regressionsprobleme, die in ähnlicher Form noch Legendre und Gauß bekannt waren.

Seien $\mathbf{x}_1, \dots, \mathbf{x}_n$ und $y_1, \dots, y_n$ p-variate Zufallsvektoren bzw. (univariate) Zufallszahlen bzgl. eines Wahrscheinlichkeitsraumes $(\Omega, \mathcal{F}, P)$. Definiere die (Zufalls)zeilenmatrizen $\mathbf{X} := (\mathbf{x}_1 \dots \mathbf{x}_n)'$ sowie $\mathbf{Y} := (y_1 \dots y_n)'$. Es gelten die folgenden Annahmen:

1. Es gelte $\operatorname{rang}(\mathbf{X}) = p$ fast sicher bzgl. P.
2. Es existieren Zufallszahlen $\varepsilon_1, \dots, \varepsilon_n$ mit $\boldsymbol{\varepsilon} = (\varepsilon_1 \dots \varepsilon_n)'$, ein Spaltenvektor $\boldsymbol{\beta} \in \mathbb{R}^p$ sowie eine positiv definite Matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ so, dass

$$y_k = \boldsymbol{\beta}' \mathbf{x}_k + \varepsilon_k \text{ für } k = 1, \dots, n \text{ bzw. } \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathrm{E}[\boldsymbol{\varepsilon}|\mathbf{X}] = \mathbf{0}, \quad \mathrm{Cov}[\boldsymbol{\varepsilon}|\mathbf{X}] = \boldsymbol{\Sigma}$$

gilt. Gilt $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_{n \times n}$ für ein $\sigma > 0$, so spricht man von der Homoskedastizität[1].

Die Regressionsaufgabe besteht nun darin, den Parametervektor $\boldsymbol{\beta}$ anhand der Daten zu schätzen. Dies geschieht in der Regel dadurch, dass man die Methode der kleinsten Quadrate anwendet, um die Summe der Quadrate der Residuen zu minimieren. Der resultierende Schätzer lautet dann

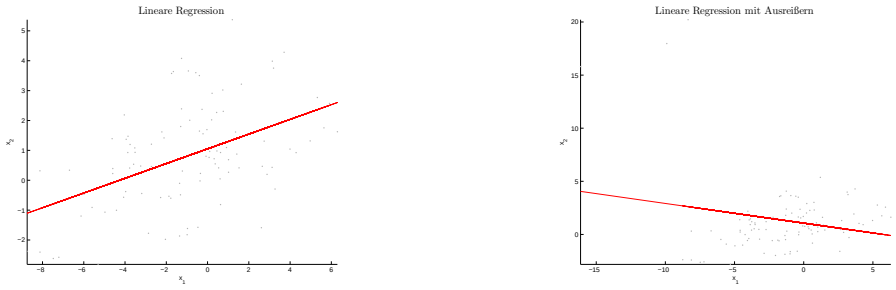
$$\hat{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{k=1}^{n} |\boldsymbol{\beta}' \mathbf{x}_k - y_k|^2 = \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^p} |\mathbf{X}\boldsymbol{\beta} - \mathbf{Y}|^2.$$

Man kann ferner zeigen, dass

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$$

gilt. Unter der Annahme, dass $\mathbf{X}$ fast sicher konstant[2] ist, folgt außerdem, dass der Schätzer erwartungstreu, konsistent, effizient und asymptotisch normal ist:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}\left(0, (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\right).$$



a) 100 Vektoren aus $\mathcal{N}\left(\left(\begin{smallmatrix}1&0\\2&3\end{smallmatrix}\right), \left(\begin{smallmatrix}2\\0\end{smallmatrix}\right)\right)$ b) 100 Vektoren aus $\mathcal{N}\left(\left(\begin{smallmatrix}1&0\\2&3\end{smallmatrix}\right), \left(\begin{smallmatrix}2\\0\end{smallmatrix}\right)\right)$ mit 3 Ausreißern

Abbildung 1

Bemerkung (Robustheit) Weicht die tatsächliche Situation von den Modellannahmen ab, was in der Praxis fast immer der Fall ist, so kann ein parametrischer Schätzer unbrauchbare Resultate liefern. Ersetzt man nur 3 von 100 Punkten aus der auf der Abbildung 1 a) dargestellten Stiprobe durch Ausreißer, so stellt man eine völlig andere Korrelationsstruktur fest. Es lässt sich sogar beweisen, dass der Bruchpunkt des Schätzers $\frac{1}{n}$ beträgt, d.h., ein einziger Ausreißer ist ausreichend, um den Schätzer beliebig zu verzerren.

Daher sind die meisten medizinischen, psychologischen, soziologischen und manchmal sogar ökonomischen oder ökologischen Studien nur mit sehr viel Vorsicht zu genießen, da diese mit wenigen Ausnahmen auf parametrischen nichtrobusten statistischen Methoden basieren.

Sollte man sich also als Anwender doch für ein parametrisches Modell entscheiden, muss man sich unbedingt ein robustes Verfahren zu Nutze machen (vgl. [3], [5]). In den letzten Jahren haben sich auch semiparametrische Methoden (s. [2]) oder eine Kombination aus parametrischen und nichtparametrischen Methoden (vgl. [4]) als sehr effektiv erwiesen. Im Falle der linearen Regression wären dabei das neuere JP1 Verfahren (viz. [4]), das altbewahrte MCD Verfahren (oder dessen neueres Analogon – FSRMCD) oder das MVE Verfahren (obwohl seine Effizienz sowie Konvergenzrate niedriger sind) besonders zu empfehlen.

[1] gr.: *ὁμός* – gleich, *σκεδαστός* – zerstreut, verteilt.
[2] Diese Situation wird als „schwache Exogenität" bezeichnet.

Beispiel (Spline-Regression) Der in der Aufgabe 7.3 beschriebene Zugang zur penalisierten nichtparametrischen Regression im Ganzraum wird in der Praxis für kompakte Menge implementiert. Diese Methode zählt zu den einfacheren nichtparametrischen Methoden und wird oft im $\mathbb{R}$ als ,,Spline Smoothing" (auch De Boorsche Methode genannt) bzw. im $\mathbb{R}^n$ als ,,Elastic net regularization" oder ,,Elastic map" bezeichnet. Der Ansatz lautet: Minimiere das Funktional

$$J(u) = p\sum_{k=1}^{n}(u(x_k) - y_k)^2 + (1-p)\int_\Omega |u''(x)|^2 \mathrm{d}x$$

über alle $u \in H^2(\Omega)$, $\Omega := (x_{(1)}, x_{(n)})$ mit $u''' = u''$ auf $\partial\Omega$ (,,free boundary conditions"), wobei $x_{(k)}$ die k-te Ordnungsstatistik bezeichnet. Insbesondere gilt also $x_{(1)} = \min\limits_{k=1,\dots,n} x_k$, $x_{(n)} = \max\limits_{k=1,\dots,n} x_k$.

Beispiel (Engel-Kurve[3] & Motorcycle Data[4]) Abschließend werden zwei Anwendungsbeispiele vorgestellt.

- 1961 bis 2001 hat das Office of Population Censuses and Surveys der Britischen Regierung im Rahmen von *Family Expenditure Survey (FES)*[5] Einkommen und Ausgaben britischer Haushalte an aufeinander folgenden 14 Tagen statistisch erfasst. Diese können verwendet werden, um die sogenannte Engel-Kurve zu bestimmen. Als Engel-Kurve bezeichnet man in der Volkswirtschaftslehre und dort speziell in der Mikroökonomik eine mathematische Funktion, die – bezogen auf ein bestimmtes Gut – für jedes Einkommensniveau angibt, wie viele Einheiten ein Konsument optimalerweise von diesem Gut nachfragen sollte (s. [1]).

  Auf der Abbildung 2 a) werden die Daten (eigentlich nur zwei Komponenten davon) aus dem Jahre 1973 dargestellt (s. [2, pp. 85–86]). Es wird die Engel-Kurve mit Hilfe der De Boorschen Methode für drei verschiedene Werte des Parameters p berechnet, wobei der Fall $p \nearrow 1$ der linearen Regression entspricht. Das ,,beste" Resultat wird im Fall $p = 0.95$ erzielt, obwohl auch der lineare Regressor befriedigendes Resultat produziert. Es sei aber angemerkt, dass die lineare Funktion dem Engelschen Gesetz widerspricht, da Letzteres nichtmonotone Funktionen vorschreibt.

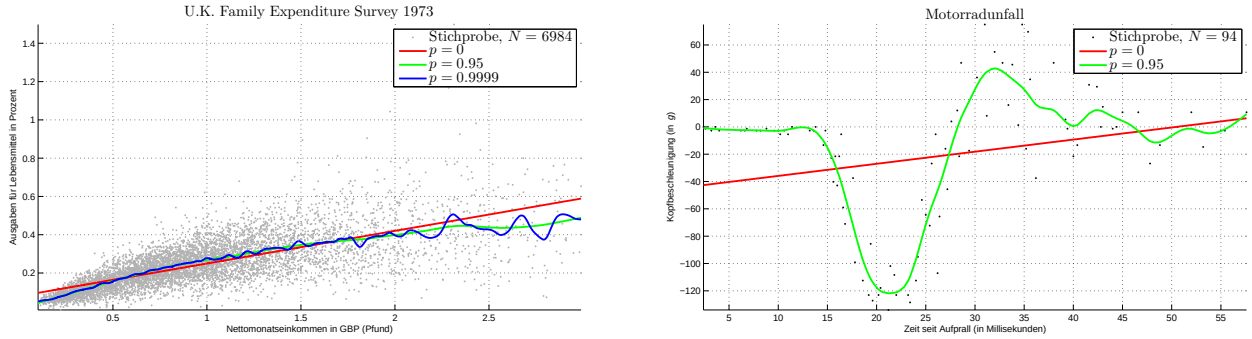


Abbildung 2

- In [6] hat Silverman einen simulierten vierdimensionalen Datensatz präsentiert, dessen erste Variable die Zeit nach dem Aufprall in einem Motorradunfall beschreibt (,,erklärende Variable"), während die zweite Variable die Kopfbeschleunigung (,,abhängige Variable") bemisst. Die Abhängigkeit wird mit Hilfe der Spline-Regression geschätzt. Dabei ist deutlich zu erkennen, dass lineare Regression ein nicht zufriedenstellendes Ergebnis liefert.

Literatur

[1] Breyer, F.: Mikroökonomik. Eine Einführung. 5. Aufl. Springer, Heidelberg u.a. (2011)

[2] Härdle, W., Müller, M., Sperlich, S., Werwatz, A. Nonparametric and Semiparamteric Models, Springer (2004)

[3] Huber, P.J. Robust Statistics, Wiley, (1981)

[4] Jobe, J.M., Pokojovy, M. A Multistep, Cluster-Based Multivariate Chart for Retrospective Monitoring of Individuals, Journal of Quality Technology, Vol. 41, No. 4, pp. 323-339, (2009)

[5] Rousseeuw, P.J., Leroy, A.M., Robust Regression and Outlier Detection, Wiley, (1987)

[6] Silverman, B.W., Some aspects of the spline smoothing approach to non-parametric curve fitting. Journal of the Royal Statistical Society, B, 47, pp. 1-52 (1985)

[3] Quelle: `http://de.wikipedia.org/wiki/Engel-Kurve`
[4] Quelle: `http://vincentarelbundock.github.io/Rdatasets/doc/boot/motor.html`
[5] `http://discover.ukdataservice.ac.uk/catalogue/?sn=3050&type=Data%20catalogue`