Peters lineare Regression

Peter Stender

2021

1 Einleitung

Mit linearer Regression wird (im eindimensionalen Fall[1]) die Konstruktion einer Geraden bezeichnet, die sinnvoll durch eine Menge von Punkten in der Ebene verläuft. Zu dieser Fragestellungen gibt es reichhaltiges realitätsbezogenes Unterrichtsmaterial für den Mathematikunterricht in der Mittelstufe[2], das gut geeignet ist, sowohl stochastisches Denken als auch wichtige Aspekte funktionalen Denkens zu vertiefen.

Problematisch an der linearen Regression ist, dass die klassische Berechnungsformel komplex ist und in der Mittelstufe nicht und in der Schule insgesamt nur mit großem Aufwand begründet werden kann, da sie partielle Differentialrechnung erfordert: die beiden Parameter der Geraden werden als Lösung einer Minimierung gefunden, was zu einem Lösungsansatz mit Differentiation einer Funktion mit zwei Variablen führt. Hier wird eine **alternative Berechnungsformel für eine Regressionsgerade vorgestellt und mathematisch begründet**. Dies Formel führt nicht zu derselben Geraden wie klassische lineare Regression, erfüllt aber dieselben stochastischen Qualitätsanforderungen. Die hier vorgestellte Regressionsgerade hat dabei zwei zentrale Vorteile: Aus mathematischer Sicht ist der zentrale Vorteil, dass der numerische Aufwand zur Bestimmung der Parameter geringer ist - dies ist ein zentrales Qualitätskriterium der Numerik. Aus fachdidaktischer Sicht ist bedeutsam, dass „Peters Regression“ mit Mitteln der Schulmathematik ab Klasse 7 begründet und berechnet werden kann. Damit kann lineare Regression sinnvoll in Schulcurricula integriert werden.

In diesem Aufsatz wird auf die mathematische Herleitung und den Vergleich der klassischen linearen Regression und „Peters Regression“ fokussiert und nur kurz auf die damit möglichen unterrichtlichen Vorgehensweisen eingegangen: Ist die Mathematik akzeptiert, liegen viele reichhaltige Beispiele für sinnstiftende Unterrichtseinheiten bereits vor[2].

2 Lineare Regression - mathematische Grundlagen

2.1 Sichtweise der Numerik

Gegeben ist eine Punktwolke, also eine Menge von Wertepaaren

$$M = \{(x_i, y_i) \in \mathbb{R}^2 : i \in \{1, 2, \ldots n\}\}$$

Es wird eine affin-lineare Funktion

$$\hat{g} : x \mapsto \hat{a} \cdot x + \hat{b} \tag{1}$$

gesucht, die diese Wertepaare optimal approximiert.

[1] Alle Betrachtungen dazu können analog auf den mehrdimensionalen Fall übertragen werden, dann sucht man affine Unterräume.

[2] Viele sehr schöne Unterrichtskonzepte finden sich hierzu auf http://www.riemer-koeln.de.

Dazu wird die Funktion

$$f_M(a,b) = \sum_{i=1}^{n}(y_i - ax_i - b)^2 = \sum_{i=1}^{n} r_i(a,b)^2 \tag{2}$$

definiert, die die Summe der Quadrate der Abstände r_i der affin-linearen Funktion $g_{a,b} : x \mapsto ax + b$ an den Stellen $x_i \ : \ i \in \{1,2,\dots n\}$ von der Punktwolke in Richtung der y-Achse liefert. Das Minimum dieser Funktion wird gesucht und mit den Parametern $(\hat{a},\hat{b})$ angenommen. Mit den Bezeichnungen

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3}$$

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{4}$$

für die Mittelwerte der Koordinaten lautet die bekannte Formel für $(\hat{a},\hat{b})$:

$$\hat{a} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{5}$$

$$\hat{b} = \overline{y} - \hat{a} \cdot \overline{x} \tag{6}$$

Aus (6) folgt, dass die Gerade $\hat{g}$ durch den Schwerpunkt $(\overline{x},\overline{y})$ der Punktwolke M verläuft.

2.2 Stochastik

Aus Sicht der Stochastik werden an die lineare Regression zunächst Voraussetzungen formuliert: gegeben ist ein affin-linearer Zusammenhang zwischen zwei Größen x und y durch

$$\tilde{g} : x \mapsto \tilde{a} \cdot x + \tilde{b} \tag{7}$$

wobei die Parameter $\tilde{a}$ und $\tilde{b}$ unbekannt sind und auf Grundlage von Messdaten geschätzt werden sollen.

Es werden Datenpaare

$$M = \{(x_i,y_i) \ \in \ \mathbb{R}^2 \ : i \in \{1,2,\dots n\}\}$$

zu den Größen erhoben, wobei typischerweise (Mess-)Fehler gemacht werden oder eine sachbasierte Streuung der Messwerte vorliegt[3]. Die Fehler (Residuen) lauten

$$r_i = y_i - \tilde{a} \cdot x_i - \tilde{b} \tag{8}$$

[3]Der Aspekt der Messfehler ist in der Physik stark präsent. In der Schulphysik tritt dies zum Beispiel bei Experimenten auf, die proportionale Zusammenhänge untersuchen. Sachbasierte Streuungen treten beispielsweise auf beim Zusammenhang zwischen Körpergröße und Schuhgröße: es existiert ein linearer Zusammenhang, der jedoch durch individuelle Ausprägungen überlagert wird.

Die zentrale Annahme für die Residuen ist, dass diese voneinander stochastisch unabhängig sind und Erwartungswert Null haben mit jeweils gleicher Varianz. Genauer wird diese Annahme so formuliert, dass die Residuen unabhängig normalverteilt $\mathcal{N}(0,\sigma^2)$ sind. Damit ist die Summe von n Residuen wieder normalverteilt mit $\mathcal{N}(0,n\sigma^2)$ und der Mittelwert der Residuen ist normalverteilt mit $\mathcal{N}(0,\frac{1}{n}\sigma^2)$ (Faltung der Normalverteilung).

Unter dieser Voraussetzung (alle Residuen $\mathcal{N}(0,\sigma^2)$) wird mit Hilfe des Maximum Likelihood Schätzers (5) und (6) hergeleitet. Damit sind $\hat{a}$ und $\hat{b}$ Maximum Likelihood Schätzer für $\tilde{a}$ und $\tilde{b}$ (Stahel, 2002, S. 262, Stahel, 2013, S. 55).

Für den Schwerpunkt von M und analog für die Schwerpunkte von Teilmenge $T \subset M$ von M gilt:

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{n} \left(\tilde{a}\cdot x_i + \tilde{b} + r_i\right)}{n} = \tilde{a}\frac{\sum_{i=1}^{n} x_i}{n} + \frac{n\cdot\tilde{b}}{n} + \frac{\sum_{i=1}^{n} r_i}{n} = \tilde{a}\overline{x} + \tilde{b} + \frac{\sum_{i=1}^{n} r_i}{n} \quad (9)$$

Damit liegt der Erwartungswert des Schwerpunkts von M aber auch jeder Teilmenge $T \subset M$ auf $\tilde{g}$ mit Varianz $\frac{1}{n}\sigma^2$ (bei Teilmengen muss statt n die Mächtigkeit der Teilmenge verwendet werden.).

3 Spezialfall

Betrachtet werden zwei disjunkte gleich mächtige Punktwolken ($n = 2k$ sei gerade),

$$T_l = \{(x_i,y_i) : i \in \{1,2,\ldots,k\}\}\ T_r = \{(x_i,y_i) : i \in \{k+1,\ldots,n\}\} \quad (10)$$

also $T_l \cup T_r = M$. Dabei sei die Nummerierung so gewählt, dass $x_i \leq x_j$ für $i < j$ und $x_k \leq x_{\text{Median}} \leq x_{k+1}$. Die Punktwolke wird also durch den Median der x_i in eine linke und eine rechte Teilwolke zerlegt.

Die Schwerpunkte $(\overline{x_l},\overline{y_l})$ von T_l und $(\overline{x_r},\overline{y_r})$ von T_r und der Schwerpunkt $(\overline{x},\overline{y})$ liegen auf einer Geraden, da sich der Gesamtschwerpunkt gleich großer Punktemengen als Mittelwert der Schwerpunkte der Teilmengen ergibt.

$$\frac{(\overline{x_l},\overline{y_l}) + (\overline{x_r},\overline{y_r})}{2} = (\overline{x},\overline{y}) \quad (11)$$

Nach (9) liegen die Erwartungswerte der Schwerpunkte der beiden Teilwolken auf $\tilde{g}$ mit kleiner Varianz $\frac{1}{k}\sigma^2$. Daher kann man $\tilde{g}$ schätzen als Gerade $\breve{g}$ durch die beiden Schwerpunkte $(\overline{x_l},\overline{y_l})$ und $(\overline{x_r},\overline{y_r})$. Wegen (11) verläuft $\breve{g}$ durch den Schwerpunkt von M und hat damit diese wichtige Eigenschaft der klassischen Regressionsgeraden $\hat{g}$.

Es gilt dann für die geschätzten Parameter $\breve{a},\breve{b}$

$$\breve{a} = \frac{\overline{y_r} - \overline{y_l}}{\overline{x_r} - \overline{x_l}} \quad (12)$$

$$\breve{b} = \overline{y_r} - \breve{a}\cdot\overline{x_r} = \overline{y_l} - \breve{a}\cdot\overline{x_l} = \overline{y} - \breve{a}\cdot\overline{x} \quad (13)$$

Die Schätzung für die Parameter des zugrunde liegenden affin-linearen Zusammenhangs sollte aufgrund der kleinen Varianz der Mittelwerte der Residuen gut sein und möglicherweise nicht schlechter als die klassische Schätzung der Regressionsgeraden durch (5) und (6).

4 Fehlerbetrachtungen

Zur Betrachtung der Fehler der Schätzer ($\hat{a}$ bzw. $\breve{a}$, es wird hier nur die Steigung betrachtet) wird zunächst das Koordinatensystem, in dem die Punktwolke, liegt so verschoben, dass $\overline{x} = 0 = \overline{y}$, das Koordinatensystem liegt also im Schwerpunkt der Punktwolke. Weiterhin sei n gerade und damit haben T_l und T_r beide $k = \frac{n}{2}$ Elemente.

Damit gilt:

$$\hat{a} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \tag{14}$$

$$= \frac{\sum_{i=1}^{n} x_i(\tilde{a}x_i + \tilde{b} + r_i)}{\sum_{i=1}^{n} x_i^2} = \frac{\sum_{i=1}^{n} x_i \tilde{a} x_i}{\sum_{i=1}^{n} x_i^2} + \frac{\sum_{i=1}^{n} x_i \tilde{b}}{\sum_{i=1}^{n} x_i^2} + \frac{\sum_{i=1}^{n} x_i r_i}{\sum_{i=1}^{n} x_i^2} \tag{15}$$

$$= \tilde{a} + \tilde{b}\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} + \frac{\sum_{i=1}^{n} x_i r_i}{\sum_{i=1}^{n} x_i^2} \tag{16}$$

Der Fehler des Schätzers ist somit[4]:

$$\hat{a} - \tilde{a} = \tilde{b}\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2} + \frac{\sum_{i=1}^{n} x_i r_i}{\sum_{i=1}^{n} x_i^2} = \tilde{b}\frac{\frac{\sum_{i=1}^{n} x_i}{n}}{\frac{\sum_{i=1}^{n} x_i^2}{n}} + \frac{\frac{\sum_{i=1}^{n} x_i r_i}{n}}{\frac{\sum_{i=1}^{n} x_i^2}{n}} = \frac{\frac{\sum_{i=1}^{n} x_i r_i}{n}}{\frac{\sum_{i=1}^{n} x_i^2}{n}} \tag{17}$$

Für die Analoge Betrachtung von $\breve{a}$ wird zur Vereinfachung der Schreibweise zunächst eine Indikatorfunktion definiert:

$$I(i) = \begin{cases} -1, & i \in \{1, 2, \dots k\} \\ 1, & i \in \{k+1, k+2, \dots n\} \end{cases} \tag{18}$$

[4] Auf Betragsstriche wird hier verzichtet, da sie für die weiteren Überlegungen keine Relevanz haben.

Dann folgt:

$$\breve{a} = \frac{\overline{y_r} - \overline{y_l}}{\overline{x_r} - \overline{x_l}} = \frac{\frac{\sum_{i=k+1}^{n} y_i}{k} - \frac{\sum_{i=1}^{k} y_i}{k}}{\frac{\sum_{i=k+1}^{n} x_i}{k} - \frac{\sum_{i=1}^{k} x_i}{k}} = \frac{\sum_{i=k+1}^{n} y_i - \sum_{i=1}^{k} y_i}{\sum_{i=k+1}^{n} x_i - \sum_{i=1}^{k} x_i} \tag{19}$$

$$= \frac{\sum_{i=k+1}^{n} (\tilde{a}x_i + \tilde{b} + r_i) - \sum_{i=1}^{k} (\tilde{a}x_i + \tilde{b} + r_i)}{\sum_{i=k+1}^{n} x_i - \sum_{i=1}^{k} x_i} \tag{20}$$

$$= \frac{\sum_{i=k+1}^{n} \tilde{a}x_i + \sum_{i=k+1}^{n} \tilde{b} + \sum_{i=k+1}^{n} r_i - \sum_{i=1}^{k} \tilde{a}x_i - \sum_{i=1}^{k} \tilde{b} - \sum_{i=1}^{k} r_i}{\sum_{i=k+1}^{n} x_i - \sum_{i=1}^{k} x_i} \tag{21}$$

$$= \frac{\sum_{i=k+1}^{n} \tilde{a}x_i + \sum_{i=k+1}^{n} r_i - \sum_{i=1}^{k} \tilde{a}x_i - \sum_{i=1}^{k} r_i}{\sum_{i=k+1}^{n} x_i - \sum_{i=1}^{k} x_i} \tag{22}$$

$$= \frac{\sum_{i=k+1}^{n} I(i)\tilde{a}x_i + \sum_{i=k+1}^{n} I(i)r_i + \sum_{i=1}^{k} I(i)\tilde{a}x_i + \sum_{i=1}^{k} I(i)r_i}{\sum_{i=k+1}^{n} I(i)x_i + \sum_{i=1}^{k} I(i)x_i} \tag{23}$$

$$= \frac{\sum_{i=1}^{n} I(i)\tilde{a}x_i + \sum_{i=1}^{n} I(i)r_i}{\sum_{i=1}^{n} I(i)x_i} = \frac{\sum_{i=1}^{n} I(i)\tilde{a}x_i}{\sum_{i=1}^{n} I(i)x_i} + \frac{\sum_{i=1}^{n} I(i)r_i}{\sum_{i=1}^{n} I(i)x_i} \tag{24}$$

$$= \tilde{a} + \frac{\sum_{i=1}^{n} I(i)r_i}{\sum_{i=1}^{n} I(i)x_i} \tag{25}$$

$$\tag{26}$$

Damit ergibt sich für den Fehler dieses Schätzers:

$$\breve{a} - \tilde{a} = \frac{\sum_{i=1}^{n} I(i)r_i}{\sum_{i=1}^{n} I(i)x_i} = \frac{\frac{\sum_{i=1}^{n} I(i)r_i}{n}}{\frac{\sum_{i=1}^{n} I(i)x_i}{n}} \tag{27}$$

Die Fehler in (17) und (27) haben ähnliche Struktur. Beide Ausdrücke sind normierte gewichtete Mittelwerte der Residuen. In (17) sind die Gewichte die x_i, in (27) sind es die Werte der Indikatorfunktion. Keiner der beiden Ausdrücke ist per se für alle Konstellationen kleiner: treten große Residuen bei großen x_i auf, so wird (17) groß, dies ist die bekannte Sensibilität der linearen Regression gegenüber Ausreißern. Andererseits werden große Residuen bei kleinen x_i durch die Gewichtung in (17) weniger zu einem Fehler beitragen als in (27). Damit ist der Schätzer $\breve{a}$ für $\tilde{a}$ dem traditionellen Schätzer nicht unterlegen sondern in manchen Konstellationen sogar besser.

Dieser Aspekt wurde mit numerischen Experimenten auf Basis von Zufallszahlen untersucht:

Es wurden mit $n = 200$ x_i gleichverteilt im Intervall $[-10, 10]$ erzeugt und Residuen normalverteilt mit $\sigma^2 = 1$. Mit $\tilde{g} : x \mapsto \tilde{a}x + \tilde{b}$ ($\tilde{a} = 0.5$, $\tilde{b} = 0$) wurden dann $y_i = \tilde{g}(x_i) + r_i$ berechnet. Auf Grundlage dieser Punktwolken wurden die Schätzer $(\hat{a}, \hat{b})$ sowie $(\breve{a}, \breve{b})$ bestimmt sowie die absoluten Fehler bezogen auf $\tilde{a}, \tilde{b}$. Die Differenzen der Schätzer lagen beide in der Größenordung von bis zu 1% der wahren Parameter. Bei einem Durchlauf von hundert

aufeinanderfolgender Neuberechnungen war die Differenz zum wahren Wert bei $(\breve{a},\breve{b})$ in 24 Fällen bei beiden Parametern kleiner und in 23 Fällen bei beiden Parametern größer als bei $(\hat{a},\hat{b})$. In 53 Fällen wurde ein Parameter besser und einer schlechter geschätzt. Ein Durchlauf mit 1000 Neuberechnungen ergab mit $(222, 479, 299)$ ($(\breve{a},\breve{b})$ besser, unentschieden, schwächer) nur einen leichten Vorteil für $(\hat{a},\hat{b})$. $(\breve{a},\breve{b})$ ist damit bei geringerem Rechenaufwand von ähnlicher Qualität wie der traditionelle Schätzer.

5 Konsequenzen für die Schule

Die Parameter der Regressionsgeraden können mit dem Schätzer $(\breve{a},\breve{b})$ bestimmt werden. Damit kann lineare Regression in der Form „Peters Regression“ ab Jahrgang 7 in der Schule elementar begründet, hergeleitet und berechnet werden.

Die Begründung/Herleitung geschieht dann verbal in folgenden Schritten:

Situation: Es liegt eine Punktwolke von Messwerten in der Ebene vor und es soll eine sinnvolle Gerade durch diese Punktwolke gelegt werden.

Erste Überlegung: die gesuchte Gerade sollte sinnvollerweise durch den Mittelpunkt der Punktwolke verlaufen, wobei die Koordinaten des Mittelpunkts die Mittelwerte der x-Koordinaten und der y-Koordinaten sind ((3), (4)).

Zweiter Schritt: Da dieses *eine* Kriterium nicht genügt, um die *zwei* Parameter der Geraden zu bestimmen („die Gerade kann sich noch um den Schwerpunkt drehen“), wird dasselbe Argument für die linke und die rechte Hälfte der Punktwolke verwendet: gesucht ist die Gerade, die durch den Schwerpunkt der linken Hälfte der Punktwolke und durch den Schwerpunkt der rechten Hälfte der Punktwolke verläuft.

Das Bestimmen einer Geraden durch zwei gegebene Punkte ist den Schülerinnen und Schüler entweder bereits bekannt oder muss in dieser Unterrichtssituation zur Bewältigung der Fragestellung entwickelt werden. Die Summenformeln ((12), (13)) werden in der Schule nicht benötigt sondern durch eine verbale Beschreibung des Rechenweges ersetzt.

6 Konsequenzen für die Mathematik

Die Berechnung der Regressionsgeraden mit dem 2-Schwerpunkte-Schätzer verursacht gegenüber dem traditionellen Schätzer einen geringeren Rechenaufwand bei vergleichbarer Qualität. Dies ist angesichts aktueller Rechenleistungen in erster Linie relevant bei der Bewältigung sehr großer Datenmengen, es ist jedoch grundsätzlich ein numerischer Qualitätsvorteil.

Das Verfahren kann offensichtlich analog im $\mathbb{R}^n$ oder für Regressionspolynome realisiert werden: Für Regressionspolynome vom Grad n wird die Punktwolke in $n+1$ gleich große Teilmengen zerlegt und das Interpolationspolynom durch diese $n+1$ Punkte bestimmt. Im $\mathbb{R}^n$ muss die Punktwolke ebenfalls in Teilmengen zerlegt werden. Die Anzahl der Teilmengen ist um eins größer als die Dimension des gesuchten affinen Unterraumes. Dieser wird durch

die Schwerpunkte der Teilmengen gelegt. Dabei kann möglicherweise die Segmentierung der Punktwolke die Qualität der Regression beeinflussen.

Die Lage des Schwerpunktes von Teilmengen der Punktwolke kann auch zur Qualitätsprüfung einer linearen Regression verwendet werden. Abbildung 2 zeigt die Schwerpunkte von Teilemengen von M. Dabei wurde so segmentiert, dass die x_i in den Quartilen liegen (kurz: Schwerpunkte über den Quartilen). Hier wurden normalverteilte Residuen zu einer quadratischen Funktion addiert. Die traditionelle Regressionsgerade wird gezeigt. Die Schwerpunkte über den Quartile liegen nicht mehr auf der Regressionsgeraden, jedoch sehr nahe der zugrundeliegenden Parabel.

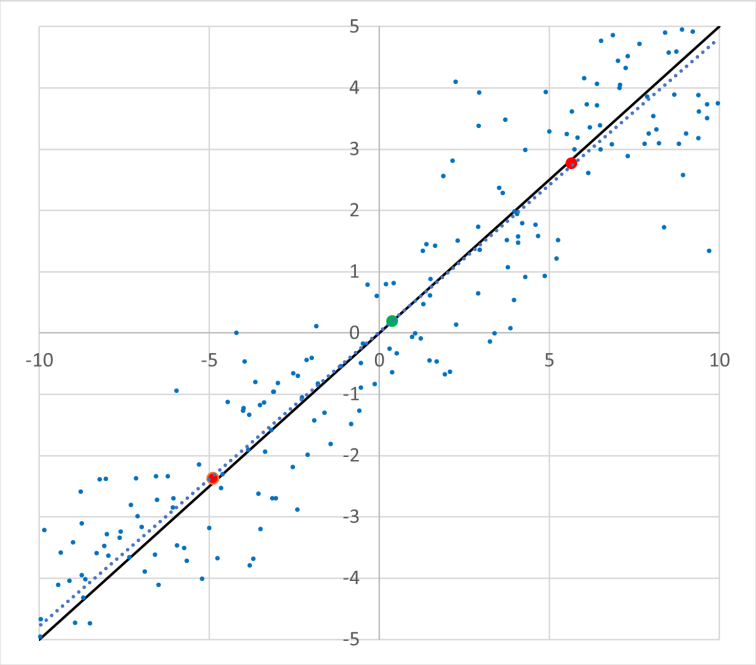

Abbildung 1: Mit Zufallszahlen erzeugte Punktwolke, Regressionsgerade und Schwerpunkte

7 Illustrierende Abbildungen

Abbildung (1) zeigt ein Beispiel für eine Punktwolke bei den beschriebenen numerischen Experimenten. In diesem Beispiel ist der Schätzer $(\breve{a}, \breve{b})$ geringfügig besser, was daran sichtbar wird, dass die Schwerpunkte der Teilwolken zwischen der wahren Geraden und der klassischen Regressionsgeraden liegen.

Abbildung (2) zeigt eine Punktwolke, die mit Zufallszahlen erzeugt wurde, die um eine vorgegebene quadratische Funktion $q : x \mapsto 0.05x^2 + 0.5x - 2$ normalverteilt streuen. Hier ist es zunächst eher überraschend, dass die Schwerpunkte der Teilwolken fast auf der Regressi-

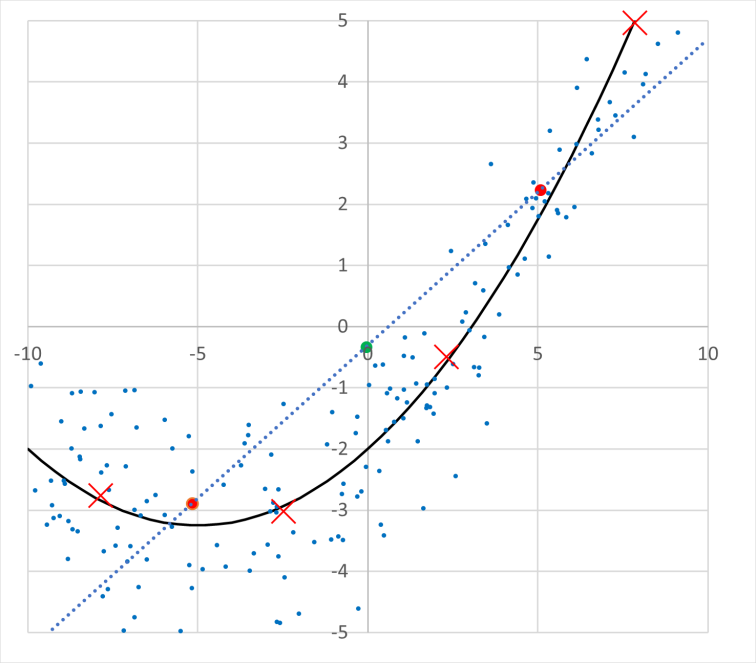

Abbildung 2: Zufallspunkte streuen um quadratische Funktion

onsgeraden liegen. Klassische Regression und „Peters Regression“ würden hier fast dieselben Parameter der Regressionsgeraden schätzen, obwohl kein linearer Zusammenhang vorliegt[5].

Die Schwerpunkte über den Quartilen treffen die Parabel sehr gut und liegen fern von der klassischen Regressionsgeraden, so dass die Annahme eines linearen Zusammenhang durch Betrachtung der Schwerpunkten dieser Teilmengen auf Grundlage sehr einfacher Rechnungen verworfen werden kann.

Literatur

Stahel, W. A. (2002). *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler* (4., verbesserte Auflage). Vieweg+Teubner Verlag. https://doi.org/10.1007/978-3-322-96962-0

Stahel, W. A. (2013). Lineare Regression: Seminar für Statistik, ETH Zürich.

[5]Es wurde bisher nicht untersucht, welchen schwächeren Voraussetzungen als r_i sind unabhängig $\mathcal{N}(0,\sigma^2)$ genügen, damit Peters Regression sehr ähnliche Ergebnisse liefert, wie die klassische lineare Regression.