# Angewandte Stochastik

Prof. Dr. Evgeny Spodarev | Vorlesungskurs |

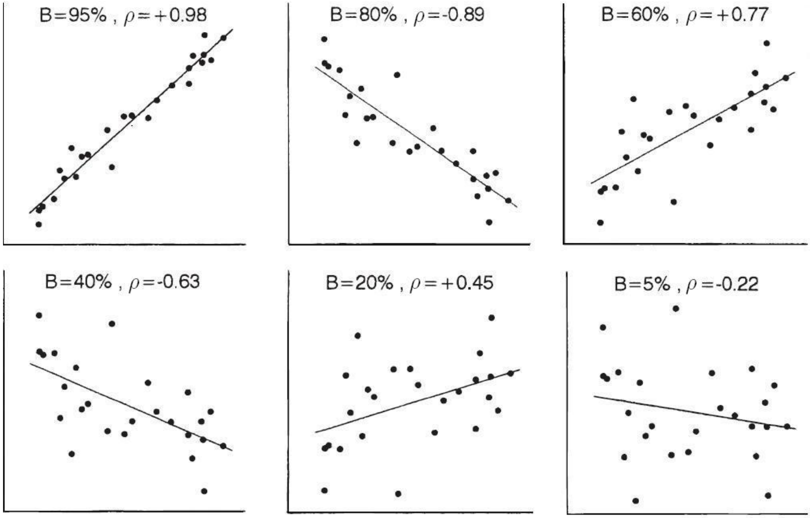
6. Thema

# Heutiges Thema

- Einfache lineare Regression

# Einfache lineare Regression

Wenn man den Zusammenhang von Merkmalen $X$ und $Y$ mit Hilfe von Streudiagrammen visualisiert, wird oft ein linearer Trend erkennbar, obwohl der Bravais-Pearson-Korrelationskoeffizient einen Wert kleiner als 1 liefert, z.B. $\varrho_{xy} \approx 0,6$ (vgl. Abb. auf der nächsten Folie).


B=95% , ρ=+0.98
B=80% , ρ=-0.89
B=60% , ρ=+0.77
B=40% , ρ=-0.63
B=20% , ρ=+0.45
B=5% , ρ=-0.22

# Einfache lineare Regression

- Dies ist der Fall, weil die Datenpunkte $(x_i, y_i)$, $i = 1, \ldots, n$ oft um eine Gerade streuen und nicht exakt auf einer Geraden liegen.
- Um solche Situationen stochastisch modellieren zu können, nimmt man den Zusammenhang der Form

$$Y = f(X) + \varepsilon$$

  an.
- $\varepsilon$ ist die sogenannte Störgröße, die auf mehrere Ursachen wie z.B. Beobachtungsfehler (Messfehler, Berechnungsfehler, usw.) zurückzuführen sein kann.
- Dabei nennt man die Zufallsvariable $Y$ *Zielgröße* oder *Regressand*, die Zufallsvariable $X$ *Einflussfaktor*, *Regressor* oder *Ausgangsvariable*.

# Einfache lineare Regression

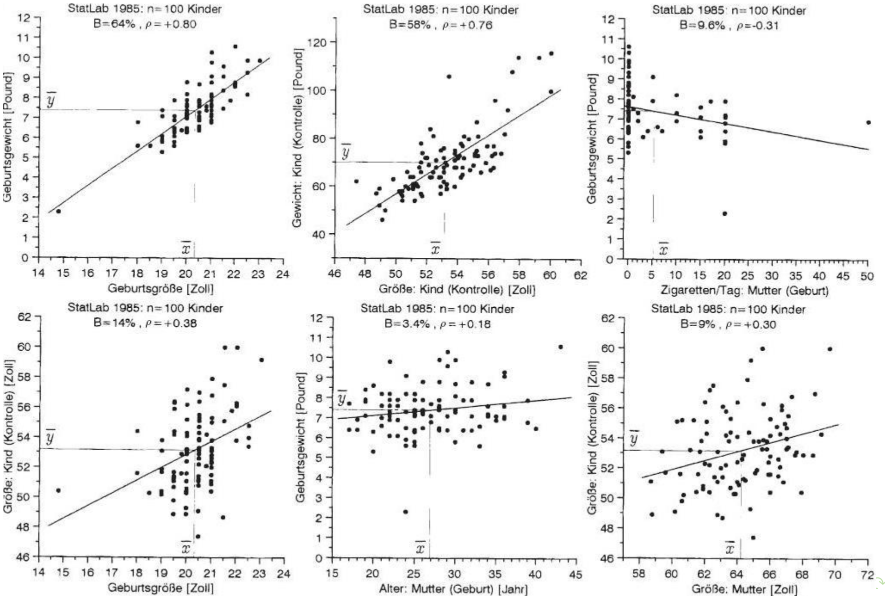
- Der Zusammenhang $Y = f(X) + \varepsilon$ wird *Regression* genannt, wobei man oft über $\varepsilon$ voraussetzt, dass $\mathsf{E}\varepsilon = 0$ (kein systematischer Beobachtungsfehler).
- Wenn $f(x) = \alpha + \beta x$ eine lineare Funktion ist, so spricht man von der *einfachen linearen Regression*.
- Es sind aber durchaus andere Arten der Zusammenhänge denkbar, wie z.B.

$$f(x) = \sum_{i=0}^{n} \alpha_i x^i$$

  (*polynomiale Regression*), usw. Beispiele für mögliche Ausgangs- bzw. Zielgrößen sind in der folgenden Tabelle zusammengefasst, einige Beispiele in der darauffolgenden Abbildung.

| $X$ | $Y$ |
|---|---|
| Geschwindigkeit | Länge des Bremswegs |
| Körpergröße des Vaters | Körpergröße des Sohnes |
| Produktionsfaktor | Qualität des Produktes |
| Spraydosen-Verbrauch | Ozongehalt der Atmosphäre |
| Noten im Bachelor-Studium | Noten im Master-Studium |

Table: Beispiele möglicher Ausgangs- und Zielgrößen


StatLab 1985: n=100 Kinder
B=64% , ρ=+0.80
Geburtsgewicht [Pound]
Geburtsgröße [Zoll]
StatLab 1985: n=100 Kinder
B=58% , ρ=+0.76
Gewicht: Kind (Kontrolle) [Pound]
Größe: Kind (Kontrolle) [Zoll]
StatLab 1985: n=100 Kinder
B=9.6% , ρ=-0.31
Geburtsgewicht [Pound]
Zigaretten/Tag: Mutter (Geburt)
StatLab 1985: n=100 Kinder
B=14% , ρ=+0.38
Größe: Kind (Kontrolle) [Zoll]
Geburtsgröße [Zoll]
StatLab 1985: n=100 Kinder
B=3.4% , ρ=+0.18
Geburtsgewicht [Pound]
Alter: Mutter (Geburt) [Jahr]
StatLab 1985: n=100 Kinder
B=9% , ρ=+0.30
Größe: Kind (Kontrolle) [Zoll]
Größe: Mutter [Zoll]

# Einfache lineare Regression

- Auf Modellebene ist folgende Fragestellung gegeben:
- Es gebe Zufallsstichproben von Ziel- bzw. Ausgangsvariablen $(Y_1, \ldots, Y_n)$ und $(X_1, \ldots, X_n)$, zwischen denen ein verrauschter linearer Zusammenhang $Y_i = \alpha + \beta X_i + \varepsilon_i$ besteht, wobei $\varepsilon_i$ Störgrößen sind, die nicht direkt beobachtbar und uns somit unbekannt sind.
- Annahme: $\mathrm{E}\, \varepsilon_i = 0 \quad \forall\, i = 1, \ldots, n$ und $\mathrm{Cov}\, (\varepsilon_i, \varepsilon_j) = \sigma^2 \delta_{ij}$, d.h. $\varepsilon_1 \ldots \varepsilon_n$ sind unkorreliert mit $\mathrm{Var}\ \varepsilon_i = \sigma^2$.
- Wenn wir über die Eigenschaften der Schätzer für $\alpha$, $\beta$ und $\sigma^2$ reden, gehen wir davon aus, dass die $X$-Werte nicht zufällig sind, also $X_i = x_i \quad \forall i = 1, \ldots, n$.

# Einfache lineare Regression

- Wenn man von einer konkreten Stichprobe $(y_1, \ldots, y_n)$ für $(Y_1, \ldots, Y_n)$ ausgeht, so sollen anhand von den Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ Regressionsparameter $\alpha$ ( *Regressionskonstante* ) und $\beta$ ( *Regressionskoeffizient* ) sowie die *Regressionsvarianz* $\sigma^2$ geschätzt werden.
- Dabei verwendet man die sogenannte *Methode der kleinsten Quadrate* , die den mittleren quadratischen Fehler von den Datenpunkten $(x_i, y_i)_{i=1,\ldots,n}$ des Streudiagramms zur *Regressionsgeraden* $y = \alpha + \beta x$ minimiert:

$$(\hat{\alpha}, \hat{\beta}) = argmin_{\alpha,\beta\in\mathbb{R}} e(\alpha, \beta) \quad \text{mit} \quad e(\alpha, \beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 .$$

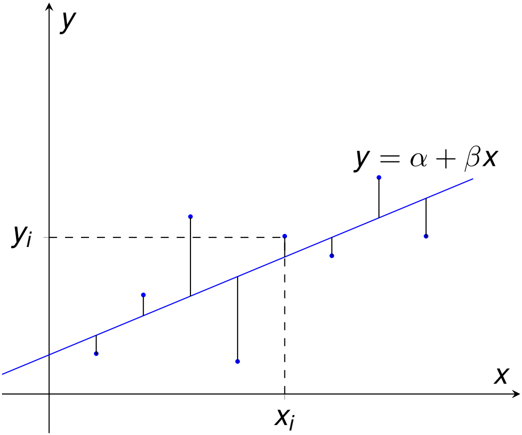



Figure: Methode kleinster Quadrate

# Einfache lineare Regression

- Diese Methode wurde 1809 von C.F. Gauß in seinem Werk ”Theoria motus corporum coelestium” verwendet, um die Laufbahnen der Himmelskörper an Hand von Beobachtungen zu bestimmen.
- Die Bezeichnung ”Methode der kleinsten Quadrate” stammt allerdings vom französischen Mathematiker A.M. Legendre (1752-1832), der sie unabhängig von Gauß entdeckt hat.
- Da die Darstellung $y_i = \alpha + \beta x_i + \varepsilon_i$ gilt, kann man $e(\alpha, \beta) = {}^1/_n \sum_{i=1}^n \varepsilon_i^2$ schreiben. Es ist der vertikale mittlere quadratische Abstand von den Datenpunkten $(x_i, y_i)$ zur Geraden $y = \alpha + \beta x$ (vgl. Abb. auf der letzten Folie).

# Einfache lineare Regression

Das Minimierungsproblem $e(\alpha, \beta) \mapsto \min$ löst man durch das zweifache Differenzieren von $e(\alpha, \beta)$. Somit erhält man $\hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n$, wobei

$$\hat{\beta} = \frac{S_{xy}^2}{S_{xx}^2}, \quad \bar{x}_n = \frac{1}{n}\sum_{i=1}^n x_i, \quad \bar{y}_n = \frac{1}{n}\sum_{i=1}^n y_i,$$

$$S_{xy}^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n), \quad S_{xx}^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \bar{x}_n)^2.$$

# Einfache lineare Regression

- Die Varianz $\sigma^2$ schätzt man durch $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{\varepsilon}_i^2$ , wobei $\hat{\varepsilon}_i = y_i - \hat{\alpha} - \hat{\beta} x_i \, , \quad i = 1, \ldots, n$ die sogenannten *Residuen* sind.
- Die Gründe, warum $\hat{\sigma}^2$ diese Gestalt hat, können an dieser Stelle nicht angegeben werden, weil wir noch nicht die Maximum-Likelihood-Methode kennen.

# Bemerkung

- Die angegebenen Schätzer für $\alpha$ und $\beta$ sind nicht symmetrisch bzgl. Variablen $x_i$ und $y_i$.
- Wenn man also die *horizontalen* Abstände (statt vertikaler) zur Bildung des mittleren quadratischen Fehlers nimmt (was dem Rollentausch $x \leftrightarrow y$ entspricht), so bekommt man andere Schätzer für $\alpha$ und $\beta$, die mit $\hat{\alpha}$ und $\hat{\beta}$ nicht übereinstimmen müssen:

$$d_i = y_i - \alpha - \beta x_i \mapsto d_i' = x_i - \frac{(y_i - \alpha)}{\beta}\,.$$

- Ein Ausweg aus dieser asymmetrischen Situation wäre es, die orthogonalen Abstände $o_i$ von $(x_i, y_i)$ zur Geraden $y = \alpha + \beta x$ zu betrachten (vgl. Abb. auf der nächsten Folie).

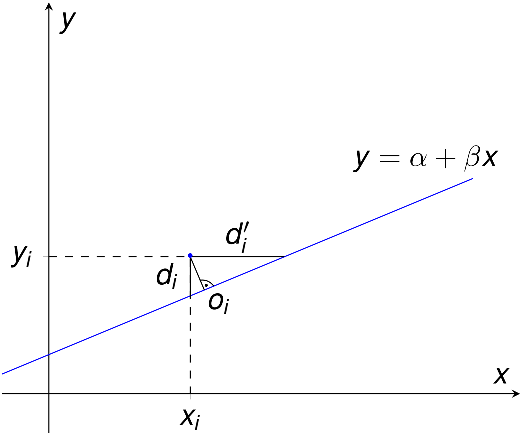



Figure: Orthogonale Abstände

- Diese Art der Regression, die ”errors-in-variables regression” genannt wird, hat aber eine Reihe von Eigenschaften, die sie zur Prognose von Zielvariablen $y_i$ durch die Ausgangsvariablen $x_i$ unbrauchbar machen.
- Sie sollte zum Beispiel nur dann verwendet werden, wenn die Standardabweichungen für $X$ und $Y$ etwa gleich groß sind.

# Beispiel

- Ein Kinderpsychologe vermutet, dass sich häufiges Fernsehen negativ auf das Schlafverhalten von Kindern auswirkt.
- Um diese Hypothese zu überprüfen, wurden 9 Kinder im gleichen Alter befragt, wie lange sie pro Tag fernsehen dürfen, und zusätzlich die Dauer ihrer Tiefschlafphase gemessen.
- So ergibt sich der Datensatz in folgender Tabelle und die Regressionsgerade aus der darauffolgenden Abbildung.

| Kind $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Fernsehzeit $x_i$ | 0,3 | 2,2 | 0,5 | 0,7 | 1,0 | 1,8 | 3,0 | 0,2 | 2,3 |
| Tiefschlafdauer $y_i$ | 5,8 | 4,4 | 6,5 | 5,8 | 5,6 | 5,0 | 4,8 | 6,0 | 6,1 |

Table: Daten von Fernsehzeit und korrespondierender Tiefschlafdauer

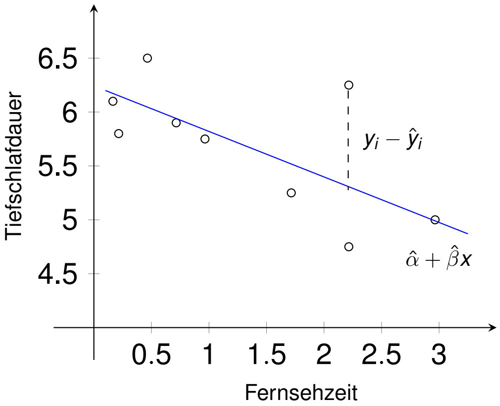



Figure: Streudiagramm und Ausgleichsgerade zur Regression der Dauer des Tiefschlafs auf die Fernsehzeit

## Beispiel

- Es ergibt sich für die oben genannten Stichproben $(x_1, \ldots, x_9)$ und $(y_1, \ldots, y_9)$

$$\bar{x}_9 = 1,33\,, \quad \bar{y}_9 = 5,56\,, \quad \hat{\beta} = -0,45\,, \quad \hat{\alpha} = 6,16\,.$$

- Somit ist

$$y = 6,16 - 0,45x$$

  die Regressionsgerade, die eine negative Steigung hat, was die Vermutung des Kinderpsychologen bestätigt.
- Außerdem ist es mit Hilfe dieser Geraden möglich, Prognosen für die Dauer des Tiefschlafs für vorgegebene Fernsehzeiten anzugeben.
- So wäre z.B. für die Fernsehzeit von 1 Stunde der Tiefschlaf von $6,16 - 0,45 \cdot 1 = 5,71$ Stunden plausibel.

## Bemerkung

1. Es gilt sgn $(\hat{\beta}) = $ sgn $(\rho_{xy})$, was aus $\hat{\beta} = \frac{s_{xy}^2}{s_{xx}^2}$ folgt. Dies bedeutet (falls $s_{yy}^2 > 0$):
   (a) Die Regressionsgerade $y = \hat{\alpha} + \hat{\beta}x$ steigt an, falls die Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ positiv korreliert sind.
   (b) Die Regressionsgerade fällt ab, falls sie negativ korreliert sind.
   (c) Die Regressionsgerade ist konstant, falls die Stichproben unkorreliert sind.

   Falls $s_{yy}^2 = 0$, dann ist die Regressionsgerade konstant $(y = \bar{y}_n)$.

# Bemerkung

2. Die Regressionsgerade $y = \hat{\alpha} + \hat{\beta} x$ verläuft immer durch den Punkt $(\bar{x}_n, \bar{y}_n)$: $\hat{\alpha} + \hat{\beta}\bar{x}_n = \bar{y}_n$.
3. Seien $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$, $i = 1, \ldots, n$. Dann gilt

$$\overline{\hat{y}_n} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \bar{y}_n \quad \text{und somit} \quad \sum_{i=1}^{n} (\underbrace{y_i - \hat{y}_i}_{\hat{\varepsilon}_i}) = 0\,.$$

Dabei sind $\hat{\varepsilon}_i$ die schon vorher eingeführten Residuen. Mit ihrer Hilfe ist es möglich, die Güte der Regressionsprognose zu beurteilen.

# Residualanalyse und Bestimmtheitsmaß

## Definition

Der relative Anteil der Streuungsreduktion an der Gesamtstreuung $S_{yy}^2$ heißt das *Bestimmtheitsmaß* der Regressionsgeraden:

$$R^2 = \frac{S_{yy}^2 - \frac{1}{n-1}\sum_{i=1}^n \hat{\varepsilon}_i^2}{S_{yy}^2} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_n)^2}.$$

Es ist nur im Fall $S_{xx}^2 > 0$, $S_{yy}^2 > 0$ definiert, d.h., wenn nicht alle Werte $x_i$ bzw. $y_i$ übereinstimmen.

Warum $R^2$ in dieser Form eingeführt wird, zeigt folgende Überlegung, die *Streuungszerlegung* genannt wird:

## Lemma

Die Gesamtstreuung ("sum of squares total") $SQT = (n-1)S_{yy}^2 = \sum_{i=1}^n (y_i - \bar{y}_n)^2$ lässt sich in die Summe
• der sogenannten erklärten Streuung "sum of squares explained" $SQE = \sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2$ und
• der Residualstreuung "sum of squared residuals" $SQR = \sum_{i=1}^n \hat{\varepsilon}_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$
zerlegen:

$$SQT = SQE + SQR$$

bzw.

$$\sum_{i=1}^n (y_i - \bar{y}_n)^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 .$$

Die erklärte Streuung gibt die Streuung der Regressionsgeradenwerte um $\bar{y}_n$ an. Sie stellt damit die auf den linearen Zusammenhang zwischen $X$ und $Y$ zurückgeführende Variation der $y$-Werte dar. Das oben eingeführte Bestimmtheitsmaß ist somit der Anteil dieser Streuung an der Gesamtstreuung:

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

Es folgt aus dieser Darstellung, dass $R^2 \in [0, 1]$ ist.

1. $R^2 = 0$ bedeutet $SQE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_n)^2 = 0$ und somit $\hat{y}_i = \bar{y}_n \; \forall i$. Dies weist darauf hin, dass das lineare Modell in diesem Fall schlecht ist, denn aus $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i = \bar{y}_n$ folgt $\hat{\beta} = \frac{S_{xy}^2}{S_{xx}^2} = 0$ und somit $S_{xy}^2 = 0$. Also sind die Merkmale $X$ und $Y$ unkorreliert.

2. $R^2 = 1$ bedingt $SQR = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = 0$. Somit liegen alle $(x_i, y_i)$ perfekt auf der Regressionsgeraden. Dies bedeutet, dass die Daten $x_i$ und $y_i$, $i = 1, \ldots, n$ perfekt linear abhängig sind.

   *Faustregel* zur Beurteilung der Güte der Anpassung eines linearen Modells an Hand von Bestimmtheitsmaß $R^2$:
   $R^2$ ist deutlich von Null verschieden (d.h. es besteht noch ein linearer Zusammenhang), falls $R^2 > \frac{4}{n+2}$, wobei $n$ der Stichprobenumfang ist.

Allgemein gilt folgender Zusammenhang zwischen dem Bestimmtheitsmaß $R^2$ und dem Bravais-Pearson-Korrelationskoeffizienten $\varrho_{xy}$:

$$R^2 = \varrho_{xy}^2$$

# Folgerung

1. Der Wert von $R^2$ ändert sich bei einer Lineartransformation der Daten $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$ nicht.
2. Da $R^2 = \varrho_{xy}^2$, ist der Wert von $R^2$ symmetrisch bzgl. der Stichproben $(x_1, \ldots, x_n)$ und $(y_1, \ldots, y_n)$:

$$\varrho_{xy}^2 = R^2 = \varrho_{yx}^2 \qquad \text{bzw.} \qquad R_{xy}^2 = R_{yx}^2 \,,$$

   wobei $R_{xy}^2$ das Bestimmtheitsmaß bezeichnet, das sich aus der normalen Regression ergibt und $R_{yx}^2$ das mit vertauschten Achsen.

# Güte der Modellanpassung

Grafisch kann man die Güte der Modellanpassung bei der linearen Regression folgendermaßen überprüfen:

- Man zeichnet Punktepaare $(\hat{y}_i, \hat{\varepsilon}_i)_{i=1,\dots,n}$ als Streudiagramm (der sogenannte *Residualplot*).
- Falls diese Punktewolke gleichmäßig um Null streut, so ist das lineare Modell gut gewählt worden.
- Falls das Streudiagramm einen erkennbaren Trend aufweist, bedeutet das, dass die Annahme des linearen Modells für diese Daten ungeeigenet sei (vgl. folgende Abb.)

# Güte der Modellanpassung

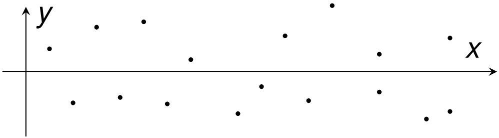



Figure: Gute Übereinstimmung mit dem linearen Modell

# Güte der Modellanpassung

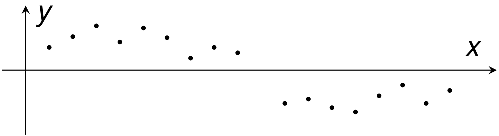



Figure: Schlechte Übereinstimmung mit dem linearen Modell