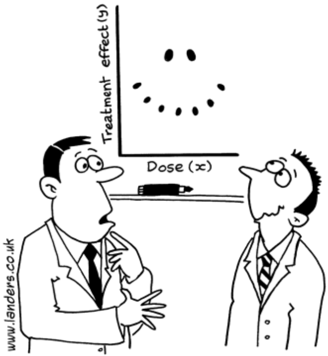
Nichtlineare Effekte in der linearen Regression

Transformationen

Jan-Philipp Kolb

Freitag, 20.06.2014

Regression - nicht ganz linear



"It's a non-linear pattern with outliers.....but for some reason I'm very happy with the data."

Quelle: `http://pn.bmj.com/content/7/4/259.extract`

Annahmen der Linearen Einfachregression

Annahmen bezüglich des Störterms

1. Die Störterme haben den Erwartungswert 0:
$$E(\varepsilon_i|x_i) = 0$$

2. Die Störterme weisen eine konstante Varianz auf (Homoskedastizität)
$$\text{var}(\varepsilon_i|x_i) = \sigma_\varepsilon^2$$

3. Die Störvariablen sind unkorreliert
$$\text{cov}\left(\varepsilon_i, \varepsilon_j|x_i\right) = 0 \qquad \text{für alle } i \neq j$$

4. Normalverteilungsannahme
$$\varepsilon_i|x_i \sim N(0; \sigma_\varepsilon^2)$$

Transformationen sind notwendig wenn...

- ... Annahmen des linearen Regressionsmodells verletzt sind (BSP: keine lineare Beziehung zwischen Variablen, keine konstante Varianz der Fehlerterme)
- ... eine theoretisch abgeleitete Formel mit den Daten in Einklang gebracht werden soll.
- ...

Unterschied zwischen Transformation der

- abhängigen (bspw. Box-Cox-Transformation) und der
- unabhängigen Variable (bspw. broken stick regression/segmented regression).

Änderung von Maßeinheiten

- Multipliziert man die abhängige Variable y mit einer Konstanten c, so werden a und b auch mit dieser Konstanten multipliziert.
- Multipliziert man die unabhängige Variable x mit einer Konstanten c, so ändert sich an a nichts. b hingegen wird mit dieser Konstanten c dividiert.
- Der Wert des Bestimmtheitsmaßes bleibt in jedem Fall unverändert, es ist unabhängig von Veränderungen der Einheiten.

Logarithmische Transformation

- Häufigste Transformation bei biologischen Daten
- Rechtslastigkeit kann durch Logarithmierung aufgehoben werden
- Den Daten zugrunde liegende Exponentialverteilung kann mittels log-Transformation linearisiert werden
- Beziehung zwischen Varianz und Mittelwert kann durch diese Transformation getilgt werden.

Problem - wenn Datensatz Nullen enthält,

Annahmen bei der linearen Einfachregression

Linearer Zusammenhang

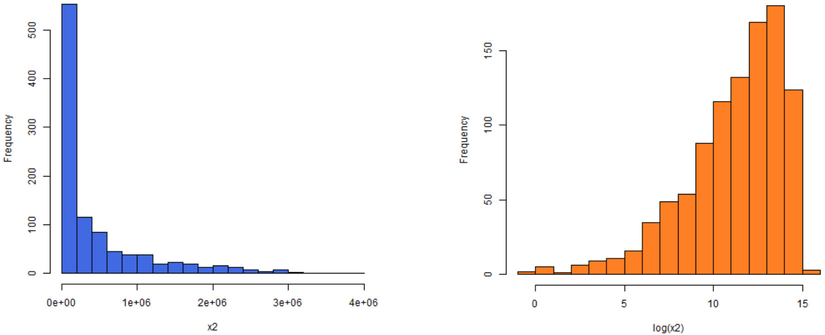
Unterstellt wurde ein linearer (linear in den Parametern α und β) Zusammenhang. In ausgewählten Fällen können nichtlineare Modelle in ein lineares Modell überführt werden.

Beachte: Unterschiedliche Interpretation der Koeffizienten und des Störterms.

Linearisierung der unabhängigen Variable

- Oft kein linearen Zusammenhang bei empirischen Beobachtungen
  Beispiel: Geschwindigkeit und Bremsweg eines PKW
- Variable wird transformiert $\rightarrow$ linearer Zusammenhang $\rightarrow$ lineare Regression $\rightarrow$ Rücktransformation
- x wird im Modell mit $f(x)$ ersetzt.

Veranschaulichung - Logarithmische Transformation



Linearisierung von Regressionsmodellen

Log-lineare-Modell (*constant elasticity model*)

$$y_i = \alpha \cdot x_i^{\beta} \cdot \varepsilon_i \quad \Rightarrow \quad \ln y_i = \ln a + b \cdot \ln x_i + \ln e_i$$

Die abhängige und die unabhängige Variablen werden logarithmisch transformiert.
β misst hierbei die Elastizität von y im Bezug auf x (daher auch *Modell mit konstanten Elastizitäten* genannt).

Fragestellung: Um wieviel Prozent ändert sich y, wenn x um ein Prozent erhöht wird?

Marginaler Effekt:

$$\frac{d\ \ln(y)}{d\ \ln(x)} = \beta$$

Linearisierende Transformationen:		
Transformation	S-Formel	Modell
$\log(x)$	`Ozone ~ log(Temp)`	$\texttt{Ozone}_i = \beta_0 + \beta_1 \log(\texttt{Temp}_i) + \varepsilon_i$
$\sqrt{x}$	`Ozone ~ sqrt(Temp)`	$\texttt{Ozone}_i = \beta_0 + \beta_1 \sqrt{\texttt{Temp}_i} + \varepsilon_i$
x^2	`Ozone ~ I(Temp^2)`	$\texttt{Ozone}_i = \beta_0 + \beta_1 \ (\texttt{Temp}_i)^2 + \varepsilon_i$
$\exp(x)$	`Ozone ~ exp(Temp)`	$\texttt{Ozone}_i = \beta_0 + \beta_1 \exp(\texttt{Temp}_i) + \varepsilon_i$
$1/x$	`Ozone ~ I(1/Temp)`	$\texttt{Ozone}_i = \beta_0 + \beta_1 \ 1/\texttt{Temp}_i + \varepsilon_i$
$\exp(-x)$	`Ozone ~ exp(-Temp)`	$\texttt{Ozone}_i = \beta_0 + \beta_1 \exp(-\texttt{Temp}_i) + \varepsilon_i$

Quelle: Eichner (2011): Einführung in die lineare Regression. S.197

`http://www.uni-giessen.de/cms/fbz/fb07/fachgebiete/mathematik/mathematik/arbeitsgruppen/stoch/stochpers/eichnerdateien/skriptenfiles/r1sum10.r4win11`

Aufgabe 3 - Linearisierende Transformationen

- Laden Sie den Datensatz airquality aus dem R-Paket datasets ein.
- Rechnen Sie eine Regression mit der Variable Ozone als abhängiger Variable und der Variable Temp als unabhängige Variable.
- Versuchen Sie die Variable Temp geeignet zu linearisieren und visualisieren Sie das Ergebnis.

Semi-log-Modell

$$y_i = \exp(a + b \cdot x_i + e_i) \quad \Rightarrow \quad \ln y_i = a + b \cdot x_i + e_i$$

Die abhängige Variable ist logarithmisch transformiert. β misst hierbei die Elastizität von y im Bezug auf x.

Fragestellung: Um wieviel Prozent ändert sich y, wenn x um eine Einheit erhöht wird?
(prozentuale Aussagen)

Marginaler Effekt:

$$\frac{d \ln(y)}{d\, x} = \beta$$

Lin-log–Modell

$$e^{y_i} = \alpha \cdot x_i^{\beta} \cdot e_i \quad \Rightarrow \quad y_i = \ln \alpha + \beta \cdot \ln x_i + \ln e_i$$

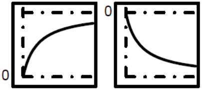
Die unabhängige Variable ist logarithmisch transformiert.

Reziprokes Modell

$$y_i = \alpha + \beta \cdot \left(\frac{1}{x_i} \right) + \varepsilon_i$$

Substitution von $x_i^* = \frac{1}{x_i}$ liefert:

$$y_i = \alpha + \beta \cdot x_i^* + \varepsilon_i$$

Spezielle linearisierbare Regressionsfunktionen:		
Linearisierbare Funktion	Transformation und Modell	Mögliche y-Graphen und **S**-Formel
$y = \theta_0 x^{\theta_1}$ (Cobb-Douglas-Funktion; Ökonomie)	$\log(y) = \log(\theta_0) + \theta_1 \log(x)$ $\log(y_i) = \beta_0 + \beta_1 \log(x_i) + \varepsilon_i$	 `log(Y) ~ log(x)`
$y = \dfrac{\theta_0 x}{\theta_1 + x}$ (Bioassay- oder Dosis-Response-Modell; Biologie)	$\dfrac{1}{y} = \dfrac{1}{\theta_0} + \dfrac{\theta_1}{\theta_0}\dfrac{1}{x}$ $\dfrac{1}{y_i} = \beta_0 + \beta_1 \dfrac{1}{x_i} + \varepsilon_i$	 `1/Y ~ I(1/x)`

Quelle: Eichner (2011): Einführung in die lineare Regression. S.198

`http://www.uni-giessen.de/cms/fbz/fb07/fachgebiete/mathematik/mathematik/arbeitsgruppen/stoch/stochpers/eichnerdateien/skriptenfiles/r1sum10.r4win11`

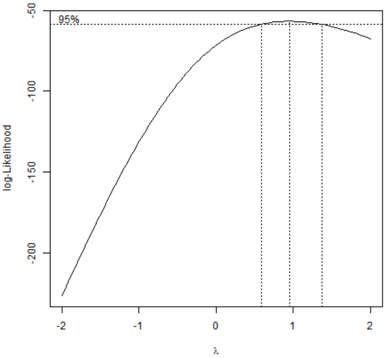
Box-cox-Transformation

- Transformation wird angewendet wenn Fehlerterme nicht normalverteilt sind
- Werte für λ werden so berechnet, dass die Fehlerterme normalisiert werden
- Mit der Transformation soll eine Stabilisierung der Varianz erreicht werden soll

http://www.r-bloggers.com/veterinary-epidemiologic-research-linear-regression-part-3-box-cox-and-matrix-representation/

Box-cox-Transformation

```
g <- lm(sr ~ pop15+pop75+dpi+ddpi,savings)
boxcox(g,plotit=T)
```



Segmented regression

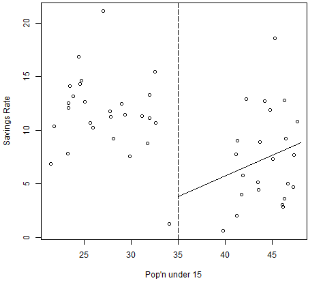
- Wird angewendet, wenn verschiedenen Regionen der Daten unterschiedliche Regressionsmodelle passen

```
g1 <- lm(sr ~ pop15, savings, subset=(pop15 < 35))
g2 <- lm(sr ~ pop15, savings, subset=(pop15 > 35))
```

Der passende Graph zu den Regressionen:

```
plot(sr ~ pop15,savings,xlab="Pop'n under 15",ylab="Savings
abline(v=35,lty=5)
segments(20,g1$coef[1]+g1$coef[2]*20,35,g1$coef[1]+
g1$coef[2]*35)
segments(48,g2$coef[1]+g2$coef[2]*48,35,g2$coef[1]+
g2$coef[2]*35)
```

Segmented regression



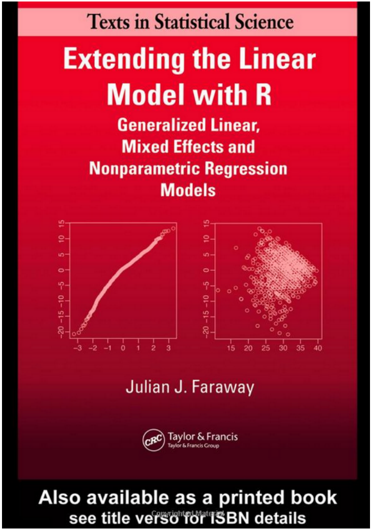
Segmented regression in einem Modell:

```
lhs <- function(x) ifelse(x < 35,35-x,0)
rhs <- function(x) ifelse(x < 35,0,x-35)
gb <- lm(sr ~ lhs(pop15) + rhs(pop15), savings)
```

Literaturhinweise

- Hackl, P. (2008): *Einführung in die Ökonometrie*, Band 7118. Pearson Education.
- Holtmann, D. (2005): *Grundlegende multivariate Modelle der sozialwissenschaftlichen Datenanalyse*
- Hübler, O. (2005): *Einführung in die empirische Wirtschaftsforschung: Probleme, Methoden und Anwendungen*, Oldenbourg Wissenschaftsverlag.

Basisliteratur



- Chapter 1
  Introduction
- Transformationen
  S. 11 ff