Nichtlineare Effekte in der linearen Regression

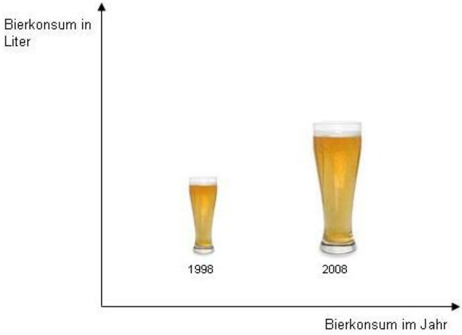
Einführung lineare Regression

Jan-Philipp Kolb

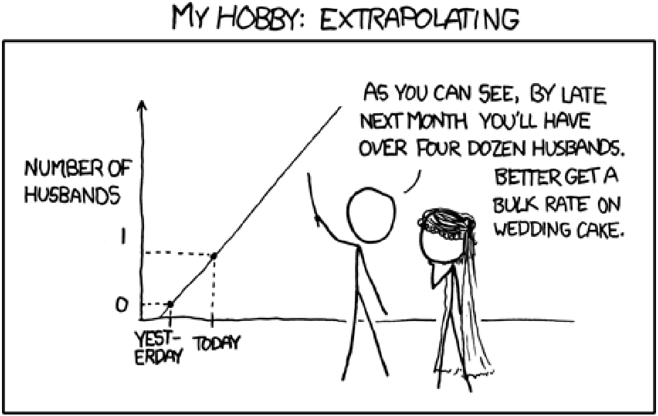
Freitag, 20.06.2014

Inhalt

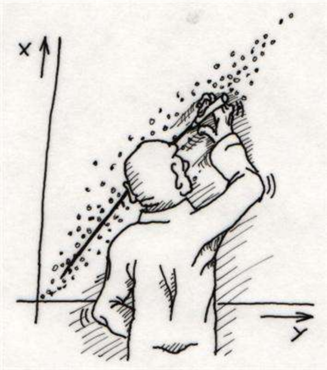
Worum geht es in diesem Abschnitt?



- Lineare Regression dient als Ausgangspunkt für viele weitere Klassen von Methoden.
- Die Grundlagen der linearen Regression sollen wieder ins Gedächtnis gerufen werden.


MY HOBBY: EXTRAPOLATING
AS YOU CAN SEE, BY LATE NEXT MONTH YOU'LL HAVE OVER FOUR DOZEN HUSBANDS.
BETTER GET A BULK RATE ON WEDDING CAKE.
NUMBER OF HUSBANDS
1
0
YEST-ERDAY
TODAY

Die lineare Regression



Quelle: `http://www.trbailey.net/pga/v/post/gm_regression.jpg.html`

Terminologie

X	Y
unabhängige Variable	abhängige Variable
erklärende Variable	zu erklärende Variable
Regressor	Regressand
Kovariable	Response Variable

Zu schätzen gilt es den Niveauparameter α (Achsenabschnitt; Intercept) und den Steigungsparameter β (Slope).

Als Schätzmethoden können herangezogen werden:

- Kleinste-Quadrat-Methode

Terminologie

X	Y
unabhängige Variable	abhängige Variable
erklärende Variable	zu erklärende Variable
Regressor	Regressand
Kovariable	Response Variable

Zu schätzen gilt es den Niveauparameter α (Achsenabschnitt; Intercept) und den Steigungsparameter β (Slope).

Als Schätzmethoden können herangezogen werden:

- Kleinste-Quadrat-Methode
- Maximum-Likelihood-Methode

Terminologie

X	Y
unabhängige Variable	abhängige Variable
erklärende Variable	zu erklärende Variable
Regressor	Regressand
Kovariable	Response Variable

Zu schätzen gilt es den Niveauparameter α (Achsenabschnitt; Intercept) und den Steigungsparameter β (Slope).

Als Schätzmethoden können herangezogen werden:

- Kleinste-Quadrat-Methode
- Maximum-Likelihood-Methode
- Momenten-Methode

Terminologie

X	Y
unabhängige Variable	abhängige Variable
erklärende Variable	zu erklärende Variable
Regressor	Regressand
Kovariable	Response Variable

Zu schätzen gilt es den Niveauparameter α (Achsenabschnitt; Intercept) und den Steigungsparameter β (Slope).

Als Schätzmethoden können herangezogen werden:

- Kleinste-Quadrat-Methode
- Maximum-Likelihood-Methode
- Momenten-Methode
- Jeweils verallgemeinerte Methoden

Terminologie

X	Y
unabhängige Variable	abhängige Variable
erklärende Variable	zu erklärende Variable
Regressor	Regressand
Kovariable	Response Variable

Zu schätzen gilt es den Niveauparameter α (Achsenabschnitt; Intercept) und den Steigungsparameter β (Slope).

Als Schätzmethoden können herangezogen werden:

- Kleinste-Quadrat-Methode
- Maximum-Likelihood-Methode
- Momenten-Methode
- Jeweils verallgemeinerte Methoden
- Empirical Likelihood

Terminologie

X	Y
unabhängige Variable	abhängige Variable
erklärende Variable	zu erklärende Variable
Regressor	Regressand
Kovariable	Response Variable

Zu schätzen gilt es den Niveauparameter α (Achsenabschnitt; Intercept) und den Steigungsparameter β (Slope).

Als Schätzmethoden können herangezogen werden:

- Kleinste-Quadrat-Methode
- Maximum-Likelihood-Methode
- Momenten-Methode
- Jeweils verallgemeinerte Methoden
- Empirical Likelihood

Lineare Regression

x und y sind stetige metrische Variablen
x ist die unabhängige Variable
y ist die abhängige Variable

Unterstellt wird ein lineares Modell

$$Y = \alpha + \beta \cdot X$$

Modell der linearen Einfachregression

Unterstellt wird ein lineares Regressionsmodell

$$Y = \alpha + \beta \cdot X + \varepsilon \quad ,$$

oder

$$Y \sim \mathrm{N}(\alpha + \beta \cdot X, \sigma_\varepsilon^2) \quad ,$$

Annahme Standardmodell: Absolutglied im Modell enthalten

Alle weiteren Faktoren, die neben X Zielvariable Y beeinflussen, werden in stochastischen Störterm ε zusammengefasst.

Da ε als nichtsystematische Komponente definiert wird, muss $E(\varepsilon) = 0$ gelten.

Lineare Regression in R - Beispieldatensatz

Lawn Roller Data

Description

The `roller` data frame has 10 rows and 2 columns. Different weights of roller were rolled over different parts of a lawn, and the depression was recorded.

Usage

`roller`

Format

This data frame contains the following columns:

weight

a numeric vector consisting of the roller weights

depression

the depth of the depression made in the grass under the roller

```
library(DAAG)
data(roller)
?roller
```

Das lineare Regressionsmodell in R

Schätzen eines Regressionsmodells:

```
roller.lm <- lm(depression ~ weight, data = roller)
```

So bekommt man die Schätzwerte:

```
summary(roller.lm)
```

Falls das Modell ohne Intercept geschätzt werden soll:

```
lm(depression ~ -1 + weight, data = roller)
```

Summary des Modells

```
summary(roller.lm)
```

```
Call:
lm(formula = depression ~ weight, data = roller)

Residuals:
   Min     1Q Median     3Q    Max 
-8.180 -5.580 -1.346  5.920  8.020 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  -2.0871     4.7543  -0.439  0.67227   
weight        2.6667     0.7002   3.808  0.00518 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.735 on 8 degrees of freedom
Multiple R-squared:  0.6445, Adjusted R-squared:  0.6001 
F-statistic:  14.5 on 1 and 8 DF,  p-value: 0.005175
```

R arbeitet mit Objekten

- `roller.lm` ist nun ein spezielles Regressions-Objekt
- Auf dieses Objekt können nun verschiedene Funktionen angewendet werden

```
predict(roller.lm) # Vorhersage
resid(roller.lm) # Residuen
```

Behandlung von nominalen Variablen

In vielen empirischen Anwendungen spielen qualitative Variablen eine Rolle.
$\Rightarrow$ Um qualitative Merkmale in Regressionsmodellen zu berücksichtigen werden Dummy Variablen verwendet.

Erklärende Variable vs. Regressor

- Bei quantitativen Variablen gilt: Erklärende Variable = Regressor.
- Eine qualitative erklärende Variable kann mehrere Kategorien haben (Beispiel: Geschlecht).
- Eine Dummy Variable ist hingegen ein Regressor, der für eine Ausprägung der erklärenden Variablen steht

Dichotome Variable

Beispiel: Betrachtet wird ein Modell, in welchem der Lohn (Y) auf die Anzahl der Ausbildungsjahre (X) und auf das Geschlecht regressiert wird.

$$y_i = \beta_0 + \beta_1 \cdot x_i + \delta \cdot D_i + \varepsilon_i$$

Mit

$$D_i = \begin{cases} 1 & \text{falls } i\text{−te Beobachtung eine Frau ist} \\ 0 & \text{falls } i\text{−te Beobachtung eine Mann ist} \end{cases}$$

als Dummy–Variable für das Geschlecht.

Bedingten Erwartungswerte:

$$\mathrm{E}\left(y_i|X, D=1\right) = \beta_0 + \beta_1 \cdot x_i + \delta$$
$$\mathrm{E}\left(y_i|X, D=0\right) = \beta_0 + \beta_1 \cdot x_i$$

Dummy Variable bewirkt Parallelverschiebung der Regressionsgeraden

Interpretation von δ

$$\delta = \mathrm{E}\left(y_i|X, D=1\right) - \mathrm{E}\left(y_i|X, D=0\right)$$

gibt Lohnunterschied zwischen Männer und Frauen an

Interaktion mit Dummy–Variablen

Interaktion zwischen metrischer und Dummy–Variable

Eine weitere erklärende Variable als das Produkt einer Dummy–Variable und stetigen Variable (hier: Ausbildungsjahre) in Modell aufgenommen.

⇒Es resultiert ein Modell, das je nach Kategorie (Mann / Frau) eine unterschiedliche Steigung aufweist.

⇒Diese neue Variable wird daher häufig auch *Steigungsvariable* genannt.

Das Regressionsmodell lautet:

$$y_i = \beta_0 + \beta_1 x_i + \delta D_i + \gamma x_i D_i + \varepsilon_i \quad .$$

- Die neue Variable ist zwar eine Funktion von X und D, jedoch keine lineare Funktion (kein Multikollinearitätsproblem).
- Durch Umsortierung erhält man die äquivalente Darstellung

$$y_i = (\beta_0 + \delta D_i) + (\beta_1 + \gamma D_i)\, x_i + \varepsilon_i$$

- für die bedingten Erwartungswerte ergeben sich

$$\begin{aligned} \mathrm{E}\left(y_i | x_i, D_i = 0\right) &= \beta_0 + \beta_1 x_i \\ \mathrm{E}\left(y_i | x_i, D_i = 1\right) &= (\beta_0 + \delta) + (\beta_1 + \gamma) x_i \end{aligned}$$

Interpretation der Parameter

- β_0 ist der Niveauparameter für Männer
- β_1 ist der Steigungsparameter für Männer
- δ ist der Unterschied im Achsenabschnitt zwischen Männer und Frauen
- γ ist der Unterschied in der Steigung zwischen Männer und Frauen

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.20050    0.84356   0.238    0.812
Jahre             0.53948    0.06422   8.400 4.24e-16 ***
Geschl.          -1.19852    1.32504  -0.905    0.366
Jahre:Geschl.    -0.08600    0.10364  -0.830    0.407
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Residual standard error: 3.186 on 522 degrees of freedom
Multiple R-Squared: 0.2598,     Adjusted R-squared: 0.2555
F-statistic: 61.07 on 3 and 522 DF,  p-value: < 2.2e-16
```

Dummy Variablen in R

So erzeugt man Dummies:

```
DumW <- rep(0,length(roller$weight))

DumW[roller$weight>6] <- 1
```

Dummy Variablen in R

Dummies können auch mit `library(dummies)` erzeugt werden:

```
data(NonResponse,package="vcd")

dummy(NonResponse$residence)
```

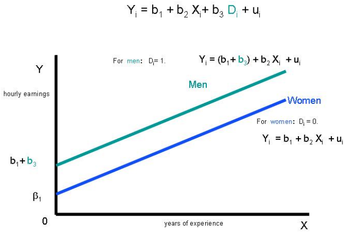
Für Regressionen kann das sehr nützlich sein.

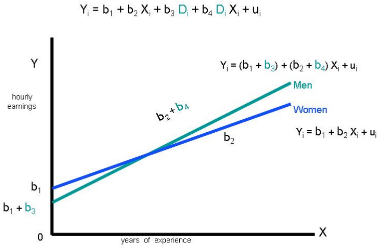
Dummy Variablen in R

```
> NonResponse
   Freq  residence response gender status
1   306 Copenhagen      yes   male      1
2   264 Copenhagen      yes female      1
3    49 Copenhagen       no   male      0
4    76 Copenhagen       no female      0
5   609       City      yes   male      1
6   627       City      yes female      1
7    77       City       no   male      0
8    79       City       no female      0
9   978    Country      yes   male      1
10  947    Country      yes female      1
11  103    Country       no   male      0
12  114    Country       no female      0
```

```
> dummy(NonResponse$residence)
      residenceCopenhagen residenceCity residenceCountry
 [1,]                   1             0                0
 [2,]                   1             0                0
 [3,]                   1             0                0
 [4,]                   1             0                0
 [5,]                   0             1                0
 [6,]                   0             1                0
 [7,]                   0             1                0
 [8,]                   0             1                0
 [9,]                   0             0                1
[10,]                   0             0                1
[11,]                   0             0                1
[12,]                   0             0                1
```

Visualisierung Effekt Dummy Variablen





Interactions in formulas

Let `a`, `b`, `c` be categorical variables and `x`, `y` be numerical variables. Interactions are specified in R as follows[3]:

Formula	Description
`y~a+x`	no interaction
`y~a:x`	interaction between variables `a` and `x`
`y~a*x`	the same and also includes the main effects
`y~a/x`	interaction between variables `a` and `x` (nested)
`y~(a+b+c)^2`	includes all two-way interactions
`y~a*b*c-a:b:c`	excludes the three-way interaction
`I()`	to use the original arithmetic operators

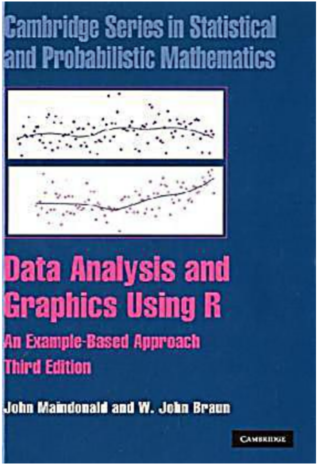
[3] Adapted from Kleiber and Zeileis, 2008.

http://scc.stat.ucla.edu/page_attachments/0000/0140/reg_2.pdf

Linkliste - lineare Regression

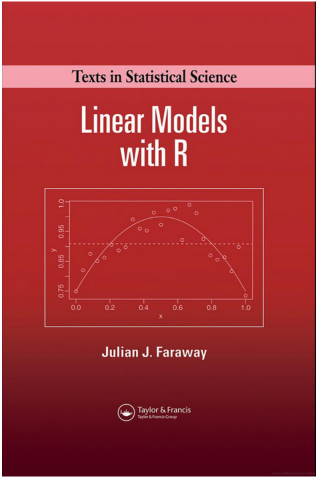
- Auf dem Kurs an der Uni Leipzig von Verena Zuber basieren auch viele der Aufgaben in diesem Workshop: `http://www.uni-leipzig.de/~zuber/teaching/ws09/r-kurs/theorie/Kurs9.pdf`
- Eine der vielen interessanten Blogs auf r-bloggers: `http://www.r-bloggers.com/r-tutorial-series-simple-linear-regression/`
- Komplettes Buch von Faraway (sehr intuitiv geschrieben): `http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf`
- Gute Einführung auf Quick-R: `http://www.statmethods.net/stats/regression.html`

Literatur Regression



1. Einführung in R
2. Datenanalyse
3. Statistische Modelle
4. Inferenzkonzepte
5. Regression mit einem Prädiktor
6. Multiple lineare Regression
7. Ausweitung des linearen Modells
8. ...

Literatur



- Lineare Regression gut erklärt
- Beispiele mit R-code

Pakete - Regression

Paket	Für was?
base{lm}	Einfache lineare Regression
base{glm}	Generalisierte Lineare Modelle
tsDyn	Autoregressive Modelle (Zeitreihen)
robustbase	Robuste Regressionen
crs	Nichtparametrische Regression
glmnet	Lasso Verfahren

Aufgabe B1 - lineare Regression

Datensatz toycars - Paket DAAG

Beschrieben wird Wegstrecke, dreier Spielzeugautos die in unterschiedlichen Winkeln Rampe herunterfuhren.

- `angle`: Winkel der Rampe
- `distance`: Zurückgelegte Strecke des Spielzeugautos
- `car`: Autotyp (1, 2 oder 3)

Quelle: `http://www.uni-leipzig.de/~zuber/teaching/ws09/r-kurs/praxis/U9.pdf`

Aufgabe B1 - lineare Regression

(a) Installieren und laden Sie das Paket `DAAG`.

(b) Speichern Sie den Datensatz "*toycars*" in einem dataframe `data` ab und wandeln Sie die Variable "`car`" des Datensatzes in einen Faktor (`as.factor`) um.

(c) Erstellen Sie drei Boxplots, die die zurückgelegte Strecke getrennt nach dem Faktor "`car`" darstellen.

(d) Schätzen Sie für **jedes** der 3 Autos **separat** die Parameter des folgenden linearen Modells mit Hilfe der Funktion "`lm()`"

$$\text{distance}_i = \beta_0 + \beta_1 \cdot \text{angle}_i + \varepsilon_i$$

(e) Überprüfen Sie deskriptiv den Fit der drei Modelle, indem Sie die Regressiongerade in einen Plot von *distance* gegen *angle* einfügen. Deutet das R^2 jeweils auf eine gute Modellanpassung hin?

(f) Führen Sie weitere deskriptive Diagnosen mit Hilfe der `plot.lm()` Funktion durch. Besteht ein linearer Zusammenhang? Sind die Residuen normalverteilt? Haben die Fehler gleiche Varianz?

Quelle: `http://www.uni-leipzig.de/~zuber/teaching/ws09/r-kurs/praxis/U9.pdf`