B4 - Generalisierte lineare Modelle mit R

Jan-Philipp Kolb

07 August 2018

Agresti - Categorical Data Analysis (2002)

- Sehr intuitiv geschriebenes Buch
- Sehr detailliertes Skript von **Laura A. Thompson**
- Das Buch behandelt grundsätzlich die kategoriale Datenanalyse.

Faraway Bücher zum Thema Regression

- Logistische Regression eingängi erklärt
- Beispiel mit R-Code
  - Faraway - **Extending the linear model with R**
  - Faraway - **Practical Regression and Anova using R**

Importieren des GESIS Panels Datensatzes

```
library(readstata13)
datf <- read.dta13("../data/ZA5666_v1-0-0_Stata14.dta",
                   convert.factors = F)
```

Das Argument `convert.factors`:

- `logical`. Wenn TRUE, werden Faktoren aus dem Stata Werte Labeln erzeugt.

Eine Funktion um fehlende Werte zu rekodieren

```
code_miss <- function(var){
  misvals <- c(-11,-22,-33,-44,-55,-66,-77,-88,-99,-111)
  var[var %in% misvals] <- NA
  return(var)
}
```

Variablen für das glm

- a11d056z: Altersgruppe

```
table(datf$a11d056z)
```

```
##
## -99   1   2   3   4   5   6   7   8   9  10  11  12  13
##   5  31  87 101  91  83 100 163 159 133  64  56 105  44
```

```
age <- code_miss(datf$a11d056z)
```

```
table(age)
```

```
## age
##   1   2   3   4   5   6   7   8   9  10  11  12  13
##  31  87 101  91  83 100 163 159 133  64  56 105  44
```

GP Variable a11d094a: Kinder unter 16 Jahre

Gibt es in Ihrem Haushalt Kinder unter 16 Jahren?

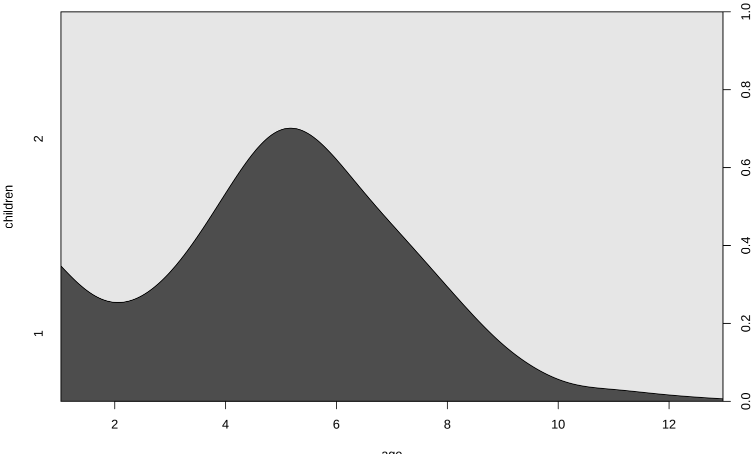
- 1 Ja
- 2 Nein

```
children <- as.factor(code_miss(datf$a11d094a))
table(children)
```

```
## children
##   1   2
## 325 681
```

Conditional Density Plot (GESIS Panel)

```
cdplot(children ~ age, data = dat)
```



Binäre abhängige Variablen im `glm`

- Die **logistische Regression** ghört zur Klasse der generalisierten linearen Modellen (GLM)
- Die Funktion zur Schätzung eines Modells dieser Klasse heißt `glm()`

Ein `glm` spezifizieren

- Formul Objekt
- die Klasse (binomial, gaussian, gamma)
- mit einer Link Funktion (logit, probit, cauchit, log, cloglog)

muss spezifiziert

Logistische Regression mit R

```r
glm_1 <- glm(children ~ age,
                    family = binomial())
```

```r
sum_glm1 <- summary(glm_1)
sum_glm1$coefficients
```

```
##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.7194058 0.16384386 -4.390801 1.129338e-05
## age          0.2225862 0.02376266  9.367056 7.458415e-21
```

Die Koeffizienten interpretieren

Wir betrachten das logistische Modell der Kinder im Haushalt als eine Funktion des Alters.

```
sum_glm1$coefficients
```

```
##                Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.7194058 0.16384386 -4.390801 1.129338e-05
## age          0.2225862 0.02376266  9.367056 7.458415e-21
```

- Die Schätzungen und Standardfehler werden mit Log Odds angegeben, nicht mit der Wahrscheinlichkeit.
- Die p-Werte bedeuten das Gleiche, wie bei der linearen Regression.

Der inverse Logit

```
sum_glm1$coefficients
```

```
##                Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -0.7194058 0.16384386 -4.390801 1.129338e-05
## age          0.2225862 0.02376266  9.367056 7.458415e-21
```

- Die Koeffizienten können nicht so einfach wie "die Kinder im Haushalt in der Altersgruppe 0" interpretiert werden. Wir müssen den inversen Logit verwenden, um etwas auszusagen.

Werte für die Log-odds von -0.7194058 sind das Gleiche, wie die Wahrscheinlichkeit: 0.3275238.

```
library(faraway)
ilogit(sum_glm1$coefficients[1,1])
```

About the intercept in a logistic model

- It is possible to get an intercept of less than 0.
- This means that the log-odds are negative, NOT the probability.
- E.g. a log-odds of 0 translates to a probability of 0.5.

Log-odds and the probability

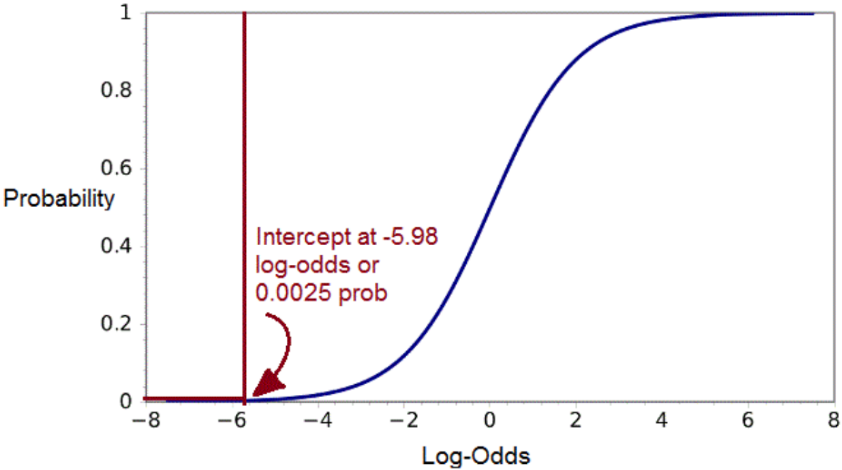
- Log-odds always increases as probability increases.

Therefore...

- A positive slope coefficient means that the response increases with the associated explanatory variable.
- In this case, the probability of children in the household increases with age.

Plotting the result

but it increases by the sigmoid curve, not at a constant rate.



Logistic regression model formula

Logistic models have regression formulas. This model's formula is:

Log-Odds(Children) = -0.7194058 + 0.2225862(Age) + error

We can plug age values into this formula to get predicted log-odds at different ages.

Log-odds for age group 5

-0.7194058 + 0.2225862*(5) = 0.3935251

Children probability in age group 5

```r
ilogit(0.3935251)
```

```
## [1] 0.597131
```

Interpreting the results

```r
anova(glm_1, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: children
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                  1000       1259
## age   1   98.956       999       1160 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Mc Fadden's R^2

```r
library(pscl)
pR2(glm_1)
```

```
##          llh      llhNull           G2     McFadden
## -580.02210772 -632.93066002 105.81710461   0.08359297
##         r2CU
##   0.13978426
```

llh	The log-likelihood from the fitted model
llhNull	The log-likelihood from the intercept-only restricted model
G2	Minus two times the difference in the log-likelihoods
McFadden	McFadden's pseudo r-squared
r2ML	Maximum likelihood pseudo r-squared
r2CU	Cragg and Uhler's pseudo r-squared

Distance between residential area and large city

How far is it from where you live to the center of the nearest large city?

- 1 - In the center of a big city
- 6 - 60 km and more

```
region <- code_miss(datf$bczd001a)
table(region)
```

```
## region
##   1   2   3   4   5   6
##  87 191 279 157 126 165
```

Satisfaction life in place of residence

How satisfied are you - all in all - with your life in [place of residence] at the moment?

- 1 - Very satisfied
- 5 - Very dissatisfied

```r
satisfactionplace <- datf$a11c019a
table(satisfactionplace)
```

```
## satisfactionplace
##   1   2   3   4   5
## 553 534  99  30   6
```

Another model

```
glm_2 <- glm(children ~ age + satisfactionplace*region,
                          family = binomial())
```

```
pseudor2 <- pR2(glm_2)
pseudor2["McFadden"]
```

```
## McFadden
## 0.258121
```

Another variable in the Gesis Panel data

- Number of tattoos:

```
Tatoos <- code_miss(datf$bdao067a)
Tatoos[Tatoos==97]<-0
```

```
table(Tatoos)
```

```
## Tatoos
##   0   1   2   3   4   5   6
## 871  56  28  13   7   4   8
```

Generalized regression with R - more functions

- Logistic model with Probit link:

```
probitmod <- glm(children ~ age,
    family=binomial(link=probit))
```

- Regression with count data:

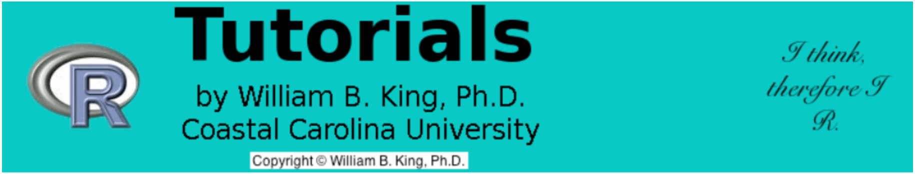
```
modp <- glm(Tatoos ~ age,family=poisson)
```

- Proportional odds logistic regression in library `MASS`:

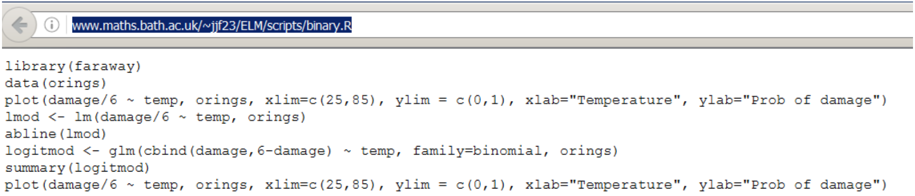
```
library("MASS")
mod_plr<-polr(a11c020a ~ a11d096b ,data=dat)
```

Linklist - logistic regression

- Introduction to **logistic regression**



- **Code for the book of Faraway**


www.maths.bath.ac.uk/~jjf23/ELM/scripts/binary.R

```
library(faraway)
data(orings)
plot(damage/6 ~ temp, orings, xlim=c(25,85), ylim = c(0,1), xlab="Temperature", ylab="Prob of damage")
lmod <- lm(damage/6 ~ temp, orings)
abline(lmod)
logitmod <- glm(cbind(damage,6-damage) ~ temp, family=binomial, orings)
summary(logitmod)
plot(damage/6 ~ temp, orings, xlim=c(25,85), ylim = c(0,1), xlab="Temperature", ylab="Prob of damage")
```

- **Categorical data: - How to perform a Logistic Regression in R**