

GENERALISIERTE LINEARE MODELLE MIT R

Jan-Philipp Kolb

11 Juni, 2019



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

Extending the Linear Model with R

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

```
library(readstata13)
datname <- "ZA5666_v1-0-0_Stata14.dta"
pathname <- "D:/Daten/GitLab/IntroDataAnalysis/data/"
datf <- read.dta13(paste0(pathname,datname),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)  
}
```

- 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
```

Leben 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)
```



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

EIN GLM SPEZIFIZIEREN

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

muss spezifiziert

LOGISTISCHE REGRESSION MIT R

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

```
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 interpretiert werden
- Wir müssen den inversen Logit verwenden, um etwas auszusagen.

Werte für die Log-odds von 0.2225862 sind das Gleiche, wie die Wahrscheinlichkeit: 0.5554179.

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

```
## [1] 0.3275238
```

ZUM AXSENABSCHNITT IN EINEM LOGISTISCHEN MODELL

- Es ist möglich, dass der Schätzwert für den Achsenabschnitt kleiner als null ist.
- Das bedeutet, dass die log-odds negativ sind und NICHT die Wahrscheinlichkeit.
- Ein Log-Odd Wert von 0 bedeutet eine Wahrscheinlichkeit von 0.5.

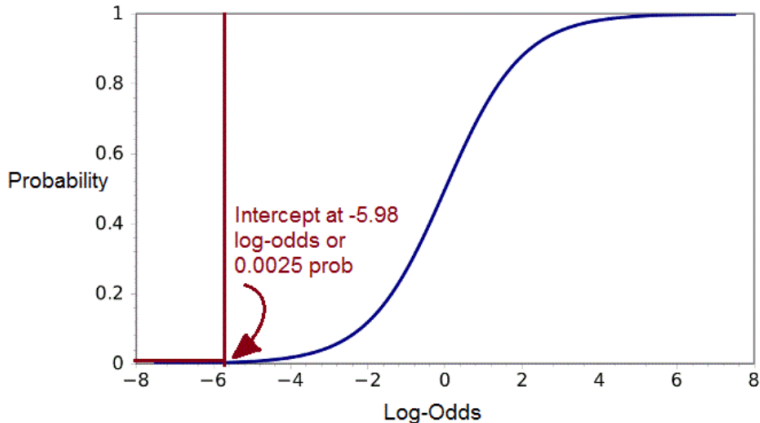
- Die Log-odds steigen an, wenn die Wahrscheinlichkeit auch ansteigt.

Daraus folgt. . .

- Ein positiver Steigungskoeffizient bedeutet, dass der Response-Wert mit zunehmendem Wert für die erklärende Variable auch zunimmt.
- In unserem Fall heißt das: Die Wahrscheinlichkeit, dass sich im Haushalt Kinder befinden steigt mit dem Alter des Befragten.

DAS ERGEBNIS GRAPHISCH DARSTELLEN

Es resultiert eine Sigmoid-Kurve, anstatt einer Gerade mit konstanter Steigungsrate wie bei der linearen Regression.



LOGISTISCHE REGRESSIONSFORMEL

DAS MODELL ALS FORMEL:

$\text{Log-Odds(Children)} = -0.7194058 + 0.2225862(\text{Age}) + \text{Fehler}$

- Wir können Werte in die Formel einsetzen um die vorhergesagten Log-Odds für unterschiedliche Altersklassen zu bekommen.

BEISPIEL: LOG-ODDS FÜR DIE ALTERSGRUPPE 5

$-0.7194058 + 0.2225862 \cdot (5) = 0.3935251$

Wahrscheinlichkeit für Kinder in der Altersgruppe 5

```
ilogit(0.3935251)
```

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


Die Ergebnisse interpretieren

```
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 ' ' 1
```

ABWEICHUNG VOM IDEALWERT

- Zweimal die Differenz zwischen der maximalen Log-Likelihood $\ell^{(M)}$ und dem Wert für das angepasste Modell
- Je niedriger die Devianz, desto besser.

```
sum_glm1 <- summary(glm_1)  
sum_glm1$deviance
```

```
## [1] 1160.044
```

MC FADDEN'S R^2

```
library(psc1)
pR2(glm_1)
```

```
##              llh              llhNull              G2              McFadden
## -580.02210772 -632.93066002  105.81710461  0.08359297  0.
##              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
```

WIE WEIT IST ES VON IHRER WOHNUNG BIS INS ZENTRUM DER NÄCHSTEN GROSSSTADT?

- 1 - Im Großstadtzentrum
- 6 - 60 km und mehr

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

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

ZUFRIEDENHEIT LEBEN IN WOHNORT

ALLES IN ALLEM, WIE ZUFRIEDEN SIND SIE MIT DEM LEBEN IN [WOHNORT]?

- 1 - Sehr zufrieden
- 5 - Sehr unzufrieden

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

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

EIN ANDERES MODELL

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

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

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

EINE WEITERE VARIABLE AUS DEM GESIS PANEL DATENSATZ

- Anzahl 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
```

GENERALISIERTE REGRESSION MIT R - MEHR FUNKTIONEN

- Logistisches Modell mit einem Probit Link:

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

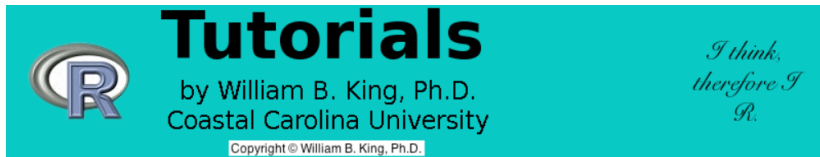
- Regression mit Zähldaten:

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

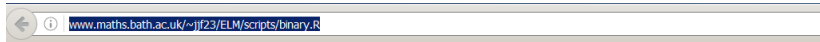
- Proportional Odds logistic Regression aus dem Paket MASS:

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


- Einführung in die **logistische Regression**



- **Code zum Buch von Faraway**



```
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")
```

- **Kategoriale Daten: - Wie man eine logistische Regression in R durchführt**