Generalized Linear Models

Vitaly Belik

Institut für Veterinär-Epidemiologie und Biometrie, FU Berlin

15/12/2020

Wiederholung

Generalized Linear Models

Generalized Linear Models

In der Statistik ist das *generalisierte lineare Modell* (GLM) eine flexible Verallgemeinerung der gewöhnlichen linearen Regression, die Antwortvariablen mit anderen (exponenziellen) Fehlerverteilungsmodellen als einer Normalverteilung ermöglicht.

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i,\ y \sim \mathcal{N}(\mu, \sigma)$$

Das GLM verallgemeinert die lineare Regression, indem das lineare Modell über eine *Verknüpfungsfunktion* (*Link function*) mit der Antwortvariablen in Beziehung gesetzt wird und die Größe der Varianz jeder Messung von ihrem vorhergesagten Wert abhängt.

- Linearität nur in den Koeffizienten

$$g(y) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i$$

Generalized Linear Models

- binomiale (dichotomische) logistische Regression
- multinominale logistische Regression
- Poissonsche Regression (count data)

Table 1: Link functions

Datentyp	Transformation	Verteilung
kontinuierlich	$\log(x)$	Log-normal
Anzahl	$\sqrt{x}$ oder $\log(x + 0.5)$	Poisson, Negative binomiale, ...
Verhältnis	$\arcsin \sqrt{x}$ oder $\text{logit} = \log \frac{x}{1-x}$	Bernoulli / binomiale, Beta binomiale, ...

Logistische Regression

Es gibt Situationen wann die *Antwortvariable* nicht normal verteilt ist. Z.B. kann sie kategoriell und *binomial* oder *multinomial* sein.

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i X_i.$$

Dabei ist $\pi = \mu_Y$ ein bedingter Mittelwert (d.h. die Wahrscheinlichkeit, dass $Y = 1$ vorausgesetzt die vorhandenen X-Werte).

$\frac{\pi}{1-\pi}$ ist das Odds-Ratio, dass $Y = 1$.

$\log\left(\frac{\pi}{1-\pi}\right)$ ist *log odds* oder *logit*.

Logistische Regression: Beispiel

Welche persönliche, demographische, und Beziehungsvariablen können Untreue vorhersagen?

Table 2: Auszug aus dem Datensatz über Untreueverhalten [nach Kabacoff / Green&Fair]

	affairs	gender	age	yearsmarried	children	religiousness	education	occupation	rating
4	0	male	37	10.00	no	3	18	7	4
5	0	female	27	4.00	no	4	14	6	4
11	0	female	32	15.00	yes	1	12	1	4
16	0	male	57	15.00	yes	5	18	6	5
23	0	male	22	0.75	no	2	17	6	3
29	0	female	32	1.50	no	2	17	5	5
44	0	female	22	0.75	no	2	12	1	3
45	0	male	57	15.00	yes	2	14	4	4
47	0	female	32	15.00	yes	4	16	1	2
49	0	male	22	1.50	no	4	14	4	5
50	0	male	37	15.00	yes	2	20	7	2
55	0	male	27	4.00	yes	4	18	6	4
64	0	male	47	15.00	yes	5	17	6	4
80	0	female	22	1.50	no	2	17	5	4
86	0	female	27	4.00	no	4	14	5	4

Logistische Regression: Beispiel(1)

```
summary(Affairs)
```

```
##     affairs          gender         age         yearsmarried    children
##  Min.   : 0.000   female:315   Min.   :17.50   Min.   : 0.125   no :171
##  1st Qu.: 0.000   male  :286   1st Qu.:27.00   1st Qu.: 4.000   yes:430
##  Median : 0.000                Median :32.00   Median : 7.000
##  Mean   : 1.456                Mean   :32.49   Mean   : 8.178
##  3rd Qu.: 0.000                3rd Qu.:37.00   3rd Qu.:15.000
##  Max.   :12.000                Max.   :57.00   Max.   :15.000
##  religiousness     education       occupation        rating
##  Min.   :1.000   Min.   : 9.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:14.00   1st Qu.:3.000   1st Qu.:3.000
##  Median :3.000   Median :16.00   Median :5.000   Median :4.000
##  Mean   :3.116   Mean   :16.17   Mean   :4.195   Mean   :3.932
##  3rd Qu.:4.000   3rd Qu.:18.00   3rd Qu.:6.000   3rd Qu.:5.000
##  Max.   :5.000   Max.   :20.00   Max.   :7.000   Max.   :5.000
```

Logistische Regression: Beispiel(1)

```
knitr::kable(table(Affairs$affairs))
```

Var1	Freq
0	451
1	34
2	17
3	19
7	42
12	38

Logistische Regression: Beispiel(2)

- Transformation zu binären Variablen

```
Affairs$ynaffair[Affairs$affair > 0] <- 1
Affairs$ynaffair[Affairs$affair == 0] <- 0
Affairs$ynaffair <- factor(Affairs$ynaffair, levels = c(0, 1), labels = c("No", "yes"))
knitr::kable(table(Affairs$ynaffair))
```

Var1	Freq
No	451
yes	150

Logistische Regression: Beispiel(3)

```
fit.full <- glm(ynaffair ~ gender + age + yearsmarried + children + religiousness +
    education + occupation + rating, data = Affairs, family = binomial())
summary(fit.full)
```

```
##
## Call:
## glm(formula = ynaffair ~ gender + age + yearsmarried + children +
##     religiousness + education + occupation + rating, family = binomial(),
##     data = Affairs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5713  -0.7499  -0.5690  -0.2539   2.5191
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.37726    0.88776   1.551 0.120807
## gendermale     0.28029    0.23909   1.172 0.241083
## age           -0.04426    0.01825  -2.425 0.015301 *
## yearsmarried   0.09477    0.03221   2.942 0.003262 **
## childrenyes    0.39767    0.29151   1.364 0.172508
## religiousness -0.32472    0.08975  -3.618 0.000297 ***
## education      0.02105    0.05051   0.417 0.676851
## occupation     0.03092    0.07178   0.431 0.666630
## rating        -0.46845    0.09091  -5.153 2.56e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 675.38  on 600  degrees of freedom
## Residual deviance: 609.51  on 592  degrees of freedom
## AIC: 627.51
##
## Number of Fisher Scoring iterations: 4
```

Beispiel: reduziertes Modell

```
fit.reduced <- glm(ynaffair ~ age + yearsmarried + religiousness + rating, data = Affairs,
    family = binomial())
summary(fit.reduced)
```

```
##
## Call:
## glm(formula = ynaffair ~ age + yearsmarried + religiousness +
##     rating, family = binomial(), data = Affairs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6278  -0.7550  -0.5701  -0.2624   2.3998
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.93083    0.61032   3.164 0.001558 **
## age           -0.03527    0.01736  -2.032 0.042127 *
## yearsmarried   0.10062    0.02921   3.445 0.000571 ***
## religiousness -0.32902    0.08945  -3.678 0.000235 ***
## rating        -0.46136    0.08884  -5.193 2.06e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 675.38  on 600  degrees of freedom
## Residual deviance: 615.36  on 596  degrees of freedom
## AIC: 625.36
##
## Number of Fisher Scoring iterations: 4
```

Beispiel: Modellvergleich (χ^2)

```r
anova(fit.reduced, fit.full, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: ynaffair ~ age + yearsmarried + religiousness + rating
## Model 2: ynaffair ~ gender + age + yearsmarried + children + religiousness +
##     education + occupation + rating
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       596     615.36
## 2       592     609.51  4   5.8474   0.2108
```

Es wird getestet ob die Reduzierung der Restsumme der Quadrate statistisch signifikant ist oder nicht.

Beispiel: Interpretation der Koeffezienten

Regressionskoeffizienten geben die Veränderung (in *log(odds)*) in der Antwortvariable, wenn alle weiteren Variablen konstant bleiben.

```
coef(fit.reduced)
```

```
##   (Intercept)           age  yearsmarried religiousness        rating
##    1.93083017   -0.03527112    0.10062274   -0.32902386   -0.46136144
```

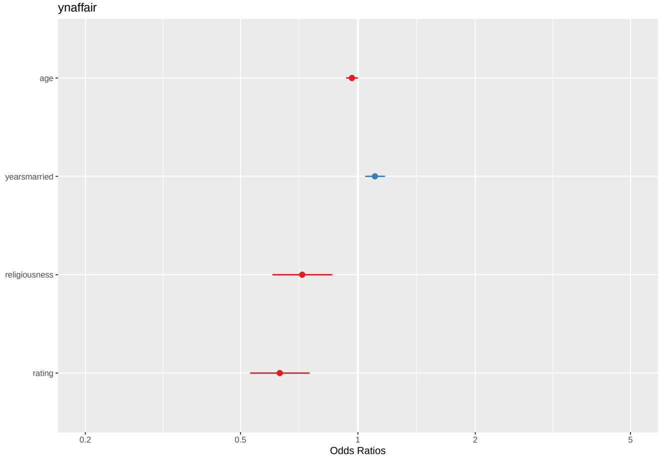
```
exp(coef(fit.reduced))
```

```
##   (Intercept)           age  yearsmarried religiousness        rating
##     6.8952321     0.9653437     1.1058594     0.7196258     0.6304248
```

```
exp(confint(fit.reduced))
```

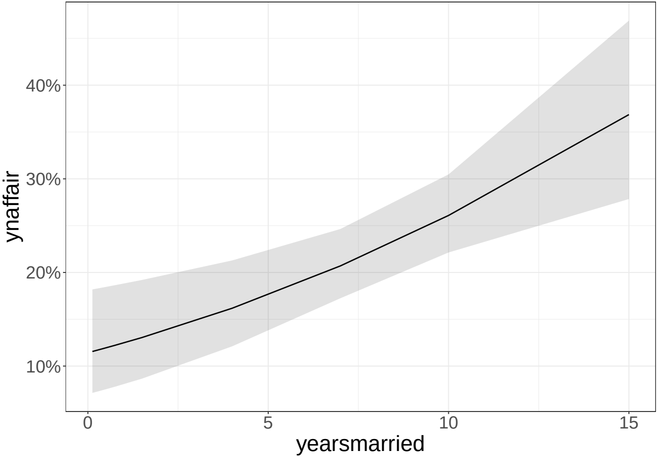
```
##                   2.5 %     97.5 %
## (Intercept)   2.1255764 23.3506030
## age           0.9323342  0.9981470
## yearsmarried  1.0448584  1.1718250
## religiousness 0.6026782  0.8562807
## rating        0.5286586  0.7493370
```

```r
library(sjPlot)
library(sjlabelled)
library(sjmisc)
plot_model(fit.reduced, axis.lim = c(0.5, 2))
```



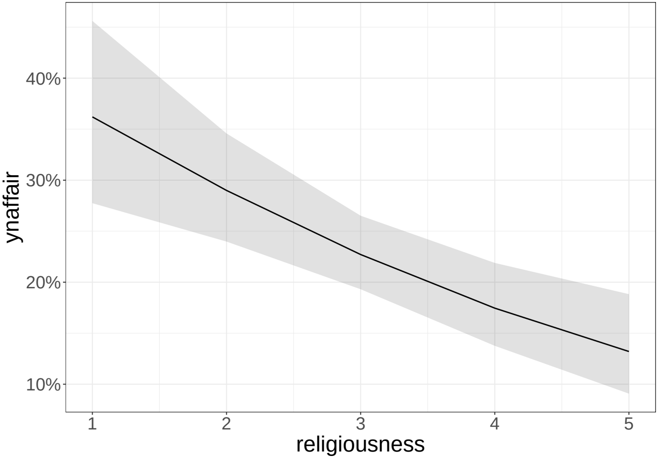
Visualization of the model

```r
library(ggeffects)
library(ggthemes)
library(ggplot2)
plot(ggpredict(fit.reduced, "yearsmarried")) + theme_bw() + theme(text = element_text(size = 24)) +
    labs(title = NULL)
```



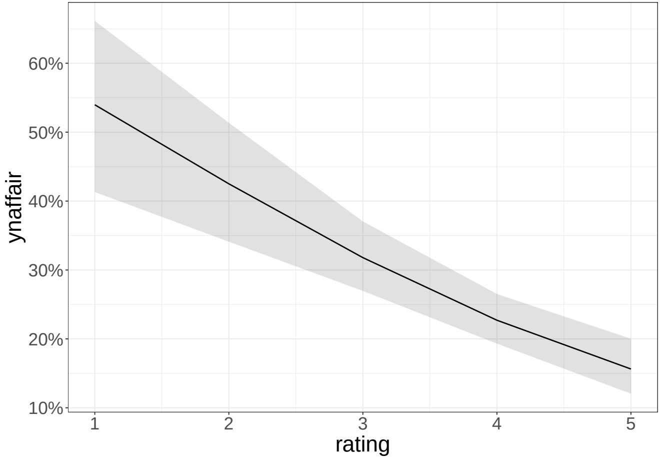
Visualization of the model

```r
plot(ggpredict(fit.reduced, "religiousness")) + theme_bw() + theme(text = element_text(size = 24)) +
    labs(title = NULL)
```



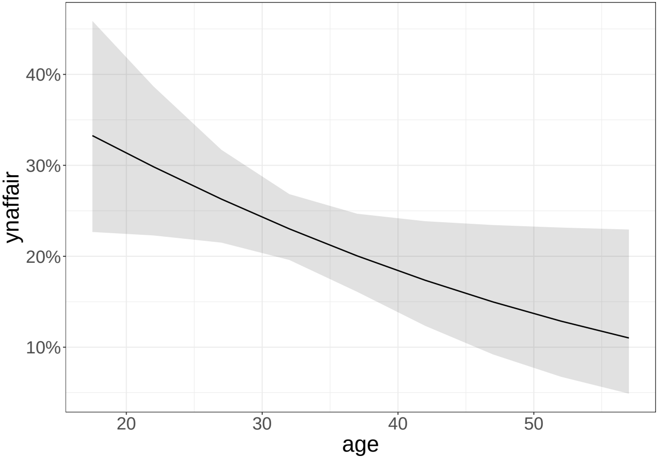
Visualization of the model

```
plot(ggpredict(fit.reduced, "rating")) + theme_bw() + theme(text = element_text(size = 24)) +
    labs(title = NULL)
```



Visualization of the model

```
plot(ggpredict(fit.reduced, "age")) + theme_bw() + theme(text = element_text(size = 24)) +
    labs(title = NULL)
```



Overdispersion

$$\sigma^2 = n\pi(1 - \pi).$$

Overdispersion findet dann statt, wenn die beobachtete Varianz von der Zielvariablen größer ist als die nach der Binonialverteilung zu erwartende Varianz.

```
fit <- fit.reduced
fit.od <- glm(ynaffair ~ age + yearsmarried + religiousness + rating, data = Affairs,
    family = quasibinomial())
pchisq(summary(fit.od)$dispersion * fit$df.residual, fit$df.residual, lower = F)
```

```
## [1] 0.340122
```

Der Test ist nicht signifikant, also sind unsere Daten nicht "overdispersed".

–>