

Verallgemeinerte lineare Modelle in R

Achim Zeileis

2009-02-20

1 Syntax

Verallgemeinerte lineare Modelle können in R mit dem Befehl `glm` angepaßt werden. Die wichtigsten Argumente sind

```
glm(formula, data, family, ...)
```

wobei **family** wie schon im linearen Modell eine Formelbeschreibung des Modells ist, etwa $y \sim x$. Im Argument **data** kann dann wieder optional ein Datensatz übergeben werden, aus dem die Variablen **y** und **x** genommen werden. Wichtig ist das Argument **family** das spezifiziert, welche konkrete Verteilung aus der Exponentialfamilie verwendet werden soll. Die Voreinstellung ist dabei **family = gaussian**, was also normalverteilte Beobachtungen unterstellt. Das resultierende Modell ist dann äquivalent zu dem Modell, das von `lm` berechnet wird, jedoch wurde zur Berechnung der Koeffizientenschätzer ein anderer Algorithmus verwendet. Da der Algorithmus in `lm` in aller Regel numerisch stabiler ist, sollte in aller Regel `lm` und nicht `glm` zur Schätzung linearer Regressionsmodelle verwendet werden.

Wenn die abhängige Variable **y** ein binärer Faktor ist, dann kann **family** auch auf **binomial** gesetzt werden. Die Voreinstellung ist dabei implizit **family = binomial(link = logit)** zu verwenden, d.h. eine logistische Regression zu rechnen. Um alternativ den Probit-Link zu verwenden müßte lediglich **family = binomial(link = probit)** gesetzt werden.

Für Zähldaten **y** wird analog **family = poisson** gesetzt. Auch hier könnten andere Link-Funktionen spezifiziert werden, in den meisten Fällen wird man hier jedoch den log-Link (Voreinstellung) verwenden.

Die **binomial** und **poisson** Familien sollen im folgenden anhand von Daten illustriert werden. Dafür laden wir wieder

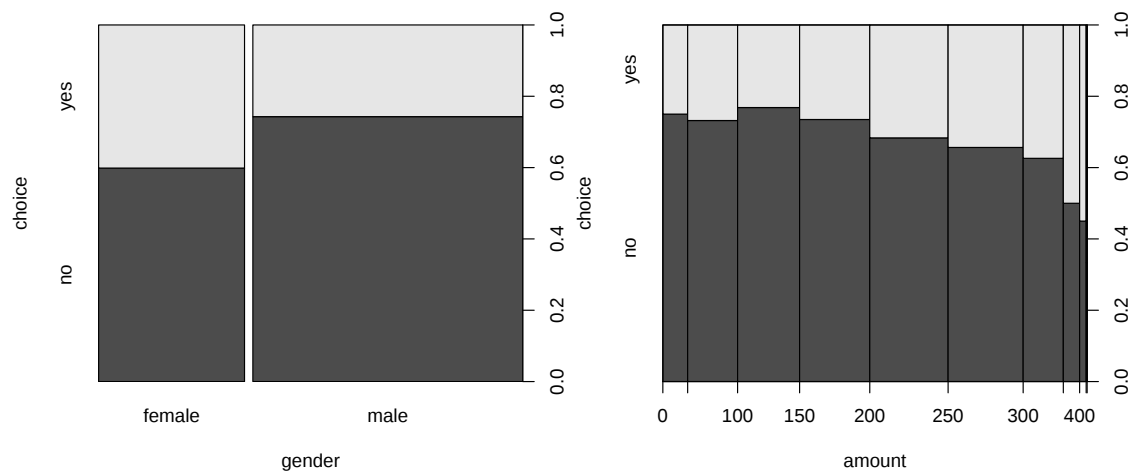
```
R> load("BBBClub.rda")
```

2 Binomialregression

Das Ziel bei der Analyse des **BBBClub** Datensatz war von Anfang an, die Erklärung der abhängigen Variablen **choice** mit Hilfe von allen im Datensatz vorhandenen Variablen. Nun haben wir also das Handwerkszeug beisammen, um dies auch zu tun.

In einem ersten Schritt, kann man sich einen Überblick über die Daten verschaffen, indem man sich (diskretisierte) Mosaikplots für die abhängige Variable gegen jede der potentiellen Erklärungsvariablen anschaut. Dies ist wird hier einmal exemplarisch für **gender** und **amount** getan:

```
R> plot(choice ~ gender, data = BBBClub)
R> plot(choice ~ amount, data = BBBClub)
```



und funktioniert für die übrigen Variablen natürlich analog. Alle diese Visualisierungen geben aber natürlich immer nur einen bestimmten Blick auf die Daten, das gesamte Bild läßt sich bei so vielen Variablen im Allgemeinen nicht einfach visualisieren.

Um nun ein geeignetes Modell für die Daten zu finden, passen wir in einem ersten Schritt das Haupteffektmodell an, das alle potentiellen Erklärungsvariablen enthält.

```
R> fm1 <- glm(choice ~ gender + amount + freq + last + first + child +
+           youth + cook + diy + art, data = BBBClub, family = binomial)
```

Durch eine AIC-basierte Modellselektion wollen wir nun versuchen ein reduziertes Modell zu finden. Als untere Schranke geben wir das triviale Modell an und als obere Schranke das Modell, in dem separate logistische Regressionsmodelle für beide Geschlechter angepaßt werden:

```
R> fm2 <- step(fm1, scope = list(lower = choice ~ 1, upper = choice ~
+           gender * (amount + freq + last + first + child + youth +
+           cook + diy + art)))
```

Start: AIC=1244.28

```
choice ~ gender + amount + freq + last + first + child + youth +
+       cook + diy + art
```

	Df	Deviance	AIC
- first	1	1222.5	1242.5
+ gender:diy	1	1219.2	1243.2
<none>		1222.3	1244.3
+ gender:cook	1	1220.7	1244.7
+ gender:last	1	1221.4	1245.4
+ gender:freq	1	1221.8	1245.8
+ gender:youth	1	1222.0	1246.0
+ gender:amount	1	1222.1	1246.1
+ gender:art	1	1222.1	1246.1
+ gender:first	1	1222.1	1246.1
+ gender:child	1	1222.2	1246.2
- amount	1	1227.4	1247.4
- youth	1	1243.9	1263.9

- art	1	1244.2	1264.2
- freq	1	1245.2	1265.2
- gender	1	1259.6	1279.6
- last	1	1263.9	1283.9
- diy	1	1267.3	1287.3
- child	1	1275.2	1295.2
- cook	1	1282.0	1302.0

Step: AIC=1242.46

choice ~ gender + amount + freq + last + child + youth + cook +
diy + art

	Df	Deviance	AIC
+ gender:diy	1	1219.3	1241.3
<none>		1222.5	1242.5
+ gender:cook	1	1220.8	1242.8
+ gender:last	1	1221.5	1243.5
+ gender:freq	1	1222.0	1244.0
+ gender:youth	1	1222.2	1244.2
+ gender:amount	1	1222.3	1244.3
+ first	1	1222.3	1244.3
+ gender:art	1	1222.3	1244.3
+ gender:child	1	1222.4	1244.4
- amount	1	1227.5	1245.5
- youth	1	1244.2	1262.2
- art	1	1244.2	1262.2
- gender	1	1259.9	1277.9
- diy	1	1267.9	1285.9
- last	1	1275.7	1293.7
- child	1	1276.0	1294.0
- cook	1	1283.7	1301.7
- freq	1	1303.0	1321.0

Step: AIC=1241.32

choice ~ gender + amount + freq + last + child + youth + cook +
diy + art + gender:diy

	Df	Deviance	AIC
<none>		1219.3	1241.3
- gender:diy	1	1222.5	1242.5
+ gender:art	1	1218.6	1242.6
+ gender:cook	1	1218.6	1242.6
+ gender:youth	1	1218.6	1242.6
+ gender:amount	1	1218.7	1242.7
+ gender:freq	1	1218.8	1242.8
+ first	1	1219.2	1243.2
+ gender:child	1	1219.3	1243.3
+ gender:last	1	1219.3	1243.3
- amount	1	1224.4	1244.4
- art	1	1241.0	1261.0
- youth	1	1241.2	1261.2
- child	1	1272.7	1292.7
- last	1	1272.9	1292.9
- cook	1	1280.3	1300.3

```
- freq          1    1298.9 1318.9
```

Dabei wird also die Variable `first` aus dem Modell eliminiert und die Interaktion `gender:diy` in das Modell aufgenommen. Für das gewählte Modell ist aus der Zusammenfassung ablesbar, welchen Einfluß die einzelnen Variablen haben:

```
R> summary(fm2)
```

Call:

```
glm(formula = choice ~ gender + amount + freq + last + child +  
    youth + cook + diy + art + gender:diy, family = binomial,  
    data = BBBClub)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.6154	-0.7340	-0.4611	0.6678	2.6988

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.1149091	0.2248196	-0.511	0.6093
gendermale	-0.7501591	0.1643403	-4.565	5.00e-06 ***
amount	0.0018984	0.0008479	2.239	0.0252 *
freq	-0.0909503	0.0112399	-8.092	5.88e-16 ***
last	0.6215875	0.0887741	7.002	2.53e-12 ***
child	-0.8866662	0.1280970	-6.922	4.46e-12 ***
youth	-0.7131437	0.1580585	-4.512	6.43e-06 ***
cook	-0.9845525	0.1326676	-7.421	1.16e-13 ***
diy	-0.7992877	0.1909804	-4.185	2.85e-05 ***
art	0.6435696	0.1401890	4.591	4.42e-06 ***
gendermale:diy	-0.3906355	0.2230318	-1.751	0.0799 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1604.8 on 1299 degrees of freedom
Residual deviance: 1219.3 on 1289 degrees of freedom
AIC: 1241.3

Number of Fisher Scoring iterations: 5

Daraus ist ersichtlich, dass die Variablen `amount`, `last` und `art` einen positiven Einfluß auf den Kauf des Kunstbandes haben. Die Chancen erhöhen sich jeweils um

```
R> exp(100 * coef(fm2)[3])
```

```
amount  
1.209061
```

```
R> exp(coef(fm2)[5])
```

```
last  
1.861881
```

```
R> exp(coef(fm2)[10])
```

```
      art  
1.903263
```

d.h.

- um rund 21% Prozent bei einem Unterschied in den Ausgaben von USD 100,
- um 86% mit jedem weiteren Monat, die der letzte Kauf zurückliegt und
- um über 90% mit jedem weiteren Kunstbuch, das vorher bereits gekauft wurde.

Alle anderen Variablen haben einen negativen Einfluß auf die Kaufwahrscheinlichkeit, d.h. insbesondere daß Frauen den Kunstband eher kaufen als Männer (Chancenverhältnis etwa 2:1) und daß das Interesse an einem Kunstband mit allen anderen Interessen (Kinder-, Jugend-, Koch- und Heimwerkerbücher) sinkt.

Das Logit-Modell soll jetzt noch mit einem Probit-Modell verglichen werden:

```
R> fm3 <- glm(choice ~ gender + amount + freq + last + child + youth +  
+      cook + diy + art + gender:diy, data = BBBClub, family = binomial(link = probit))
```

Auch wenn die konkreten Koeffizientenschätzer der beiden Modelle unterschiedlich ist (aufgrund der anderen Link-Funktion), so ist doch die Anpassung beider Modelle nahezu identisch. Sowohl die log-Likelihood

```
R> logLik(fm2)
```

```
'log Lik.' -609.6615 (df=11)
```

```
R> logLik(fm3)
```

```
'log Lik.' -609.596 (df=11)
```

als auch völlig äquivalent das AIC

```
R> AIC(fm2)
```

```
[1] 1241.323
```

```
R> AIC(fm3)
```

```
[1] 1241.192
```

für beide Modelle ist fast gleich. Dies ist nicht untypisch: Wenn nicht sehr extreme Wahrscheinlichkeiten nahe 0 oder 1 auftreten, so liefern beide Modelle meist sehr vergleichbare Ergebnisse.

Des weiteren ist noch zu sagen, daß die Funktionen, die wir für "lm"-Objekte kennengelernt haben, auch für "glm"-Objekte funktionieren, wenn auch manchmal mit leichten Veränderungen. Exemplarisch sollen hier kurz `anova` und `residuals` illustriert werden, die Interpretation läuft in beiden Fällen im wesentlichen analog zum linearen Modell.

Um beispielsweise das triviale Modell mit dem gewählten Modell zu vergleichen, können wir wieder `anova` verwenden, müssen jedoch noch spezifizieren, welchen Test wir verwenden möchten.

```
R> fm0 <- glm(choice ~ 1, data = BBBClub, family = binomial)  
R> anova(fm0, fm2, test = "Chisq")
```

Analysis of Deviance Table

Model 1: choice ~ 1

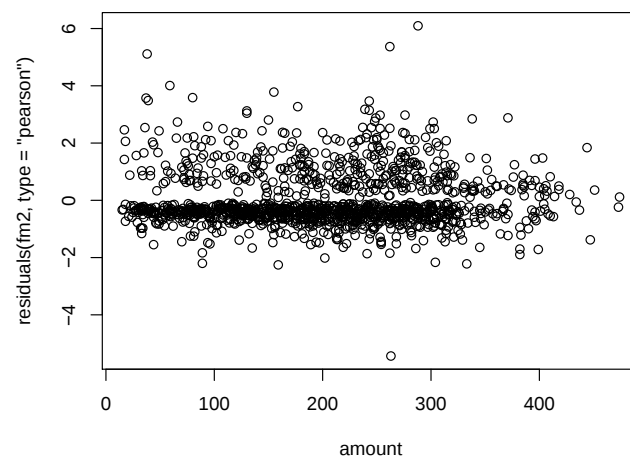
Model 2: choice ~ gender + amount + freq + last + child + youth + cook +
diy + art + gender:diy

	Resid. Df	Resid. Dev	Df	Deviance	P(> Chi)
1	1299	1604.83			
2	1289	1219.32	10	385.51	1.141e-76

Der dabei gewählte χ^2 -Test ist eine Verallgemeinerung des F -Tests, den wir im linearen Modell verwendet haben. Das Ergebnis ist natürlich hochsignifikant.

Oder wir können uns auch einen Residualplot anschauen, bspw. gegen die Erklärungsvariable **amount**.

```
R> plot(residuals(fm2, type = "pearson") ~ amount, data = BBBClub)
```



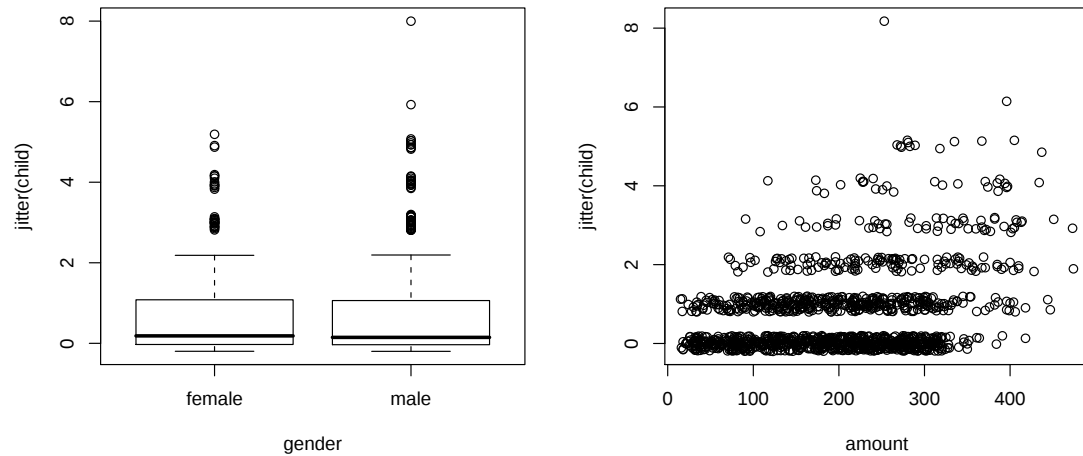
Hier tragen wir also die Pearson-Residuen gegen den Regressor **amount** ab und können feststellen, daß es keine leicht erkennbaren Strukturen mehr in den Residuen gibt (wie etwa Heteroskedastizität oder funktionale Form).

3 Poissonregression

Schließlich soll auch noch die Poissonregression illustriert werden. Als Beispiel versuchen wir die Anzahl der gekauften Kinderbücher **child** durch die Erklärungsvariablen **gender**, **amount**, **first** und **last** zu beschreiben. Dabei beginnen wir wieder mit der explorativen Analyse und visualisieren die Abhängigkeit der numerischen Variablen **child** von den Erklärungsvariablen durch Boxplots bzw. Streudiagramme. Speziell bei letzteren hilft es häufig die Variablen etwas zu ‘verwackeln’ (englisch: to jitter), da es bei Zählvariablen typischerweise viele Bindungen, d.h. identische Werte, gibt:

```
R> plot(jitter(child) ~ gender, data = BBBClub)
```

```
R> plot(jitter(child) ~ amount, data = BBBClub)
```



Hier beginnen wir mit getrennten Regressionmodellen nach Geschlecht

```
R> fm1 <- glm(child ~ gender * (amount + first + last + freq), data = BBBClub,
+             family = poisson)
```

und versuchen dieses Modell durch eine schrittweise Suche zu vereinfachen.

```
R> fm2 <- step(fm1)
```

Start: AIC=2498.31

```
child ~ gender * (amount + first + last + freq)
```

	Df	Deviance	AIC
- gender:first	1	1122.5	2496.3
- gender:freq	1	1122.6	2496.5
- gender:last	1	1123.1	2497.0
<none>		1122.4	2498.3
- gender:amount	1	1126.9	2500.8

Step: AIC=2496.35

```
child ~ gender + amount + first + last + freq + gender:amount +
gender:last + gender:freq
```

	Df	Deviance	AIC
- gender:freq	1	1122.6	2494.5
- first	1	1122.8	2494.7
<none>		1122.5	2496.3
- gender:last	1	1125.8	2497.7
- gender:amount	1	1126.9	2498.8

Step: AIC=2494.49

```
child ~ gender + amount + first + last + freq + gender:amount +
gender:last
```

	Df	Deviance	AIC
--	----	----------	-----

```

- first          1    1122.9 2492.8
- freq           1    1123.4 2493.3
<none>           1122.6 2494.5
- gender:last    1    1126.1 2496.0
- gender:amount  1    1127.2 2497.1

```

Step: AIC=2492.78

```
child ~ gender + amount + last + freq + gender:amount + gender:last
```

```

          Df Deviance    AIC
- freq          1    1123.4 2491.3
<none>          1122.9 2492.8
- gender:last    1    1126.6 2494.5
- gender:amount  1    1127.5 2495.3

```

Step: AIC=2491.28

```
child ~ gender + amount + last + gender:amount + gender:last
```

```

          Df Deviance    AIC
<none>          1123.4 2491.3
- gender:last    1    1127.1 2493.0
- gender:amount  1    1128.0 2493.9

```

In dem gewählten Modell verbleiben dann die Haupteffekte **gender**, **amount** und **last**, sowie jeweils deren Interaktion mit **gender**. Die Zusammenfassung liefert wieder noch mehr Details:

```
R> summary(fm2)
```

Call:

```
glm(formula = child ~ gender + amount + last + gender:amount +
    gender:last, family = poisson, data = BBBClub)
```

Deviance Residuals:

```

      Min       1Q   Median       3Q      Max
-2.4835  -0.8878  -0.7878   0.7478   2.9481

```

Coefficients:

```

              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4012805  0.1415438  -9.900  <2e-16 ***
gendermale     0.0887673  0.1764595   0.503  0.6149
amount         0.0013068  0.0006196   2.109  0.0349 *
last           0.1899378  0.0161246  11.779  <2e-16 ***
gendermale:amount -0.0016843  0.0007878  -2.138  0.0325 *
gendermale:last  0.0386908  0.0200657   1.928  0.0538 .
---

```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for poisson family taken to be 1)

```

Null deviance: 1810.1  on 1299  degrees of freedom
Residual deviance: 1123.4  on 1294  degrees of freedom
AIC: 2491.3

```

Number of Fisher Scoring iterations: 5

Hieraus ist abzulesen, daß der dominierende Effekt (sowohl qualitativ – kleinster p -Wert – als auch quantitativ – größter Koeffizient) von **last** ausgeht. Dies ist ein Phänomen, das bei der Analyse dieses Datensatzes, schon an mehreren Stellen auftrat. Wenn der letzte Einkauf lange zurückliegt, so wurden bei diesem Einkauf generell viele Bücher erworben, und offensichtlich umfaßt dies auch Kinderbücher. Bei Männern ist dieser Effekt sogar noch stärker.

Außerdem ist ein gewisser Effekt bzgl. **amount** zu beobachten. Dabei ist interessant, daß dieser Effekt i.W. nur in der Frauengruppe vorliegt, und in der Männergruppe praktisch auf 0 sinkt. Frauen, die bei ihren Einkäufen beim Buchclub viel Geld ausgeben, kaufen auch eher Kinderbücher. Bei Männern hingegen scheint dies nicht der Fall zu sein.