Poisson Regression

Vitaly Belik

Institut für Veterinär-Epidemiologie und Biometrie, FU Berlin

2021 January 26

Wiederholung

Poissonregression

Die Poisson-Regression ist nützlich, wenn Sie eine Ergebnisvariable vorhersagen, die die Anzahl aus einer Reihe kontinuierlicher und / oder kategorialer Prädiktorvariablen darstellt.

Wir interessieren uns für den Einfluss einer Behandlung mit Antiepileptika auf die Anzahl der Anfälle, die über einen Zeitraum von acht Wochen nach Beginn der Therapie auftreten [Breslow seizure data (Breslow, 1993)].

Es wurden Daten zum Alter und zur Anzahl der Anfälle gesammelt, die von Patienten gemeldet wurden, die während eines Zeitraums von acht Wochen vor und acht Wochen nach der randomisierten Aufteilung in eine Arzneimittel- oder Placebo-Gruppe an einfachen oder komplexen partiellen Anfällen litten.

SumY (die Anzahl der Anfälle im Zeitraum von acht Wochen nach der Randomisierung) ist die Antwortvariable. Der Behandlungszustand (*Trt*), das Alter in Jahren (*Age*) und die Anzahl der Anfälle, die im Basiszeitraum von acht Wochen (*Basis*) gemeldet wurden, sind die Prädiktorvariablen.

Die Baseline-Anzahl der Anfälle und das Alter sind aufgrund ihrer möglichen Auswirkung auf die Antwortvariable enthalten.

Fragestellung: verringert die medikamentöse Behandlung die Anzahl der Anfälle nach Berücksichtigung der Kovariaten oder nicht?

Beispiel: Breslow seizure data

```
data(breslow.dat, package = "robust")
names(breslow.dat)
```

```
##  [1] "ID"    "Y1"    "Y2"    "Y3"    "Y4"    "Base"  "Age"   "Trt"   "Ysum"
## [10] "sumY"  "Age10" "Base4"
```

```
summary(breslow.dat[c(6, 7, 8, 10)])
```

```
##       Base             Age               Trt          sumY
##  Min.   :  6.00   Min.   :18.00   placebo  :28   Min.   :  0.00
##  1st Qu.: 12.00   1st Qu.:23.00   progabide:31   1st Qu.: 11.50
##  Median : 22.00   Median :28.00                  Median : 16.00
##  Mean   : 31.22   Mean   :28.34                  Mean   : 33.05
##  3rd Qu.: 41.00   3rd Qu.:32.00                  3rd Qu.: 36.00
##  Max.   :151.00   Max.   :42.00                  Max.   :302.00
```

Beispiel: Breslow seizure data

```
ggplot(breslow.dat, aes(x = sumY)) + geom_histogram(fill = "transparent", color = "black")
```

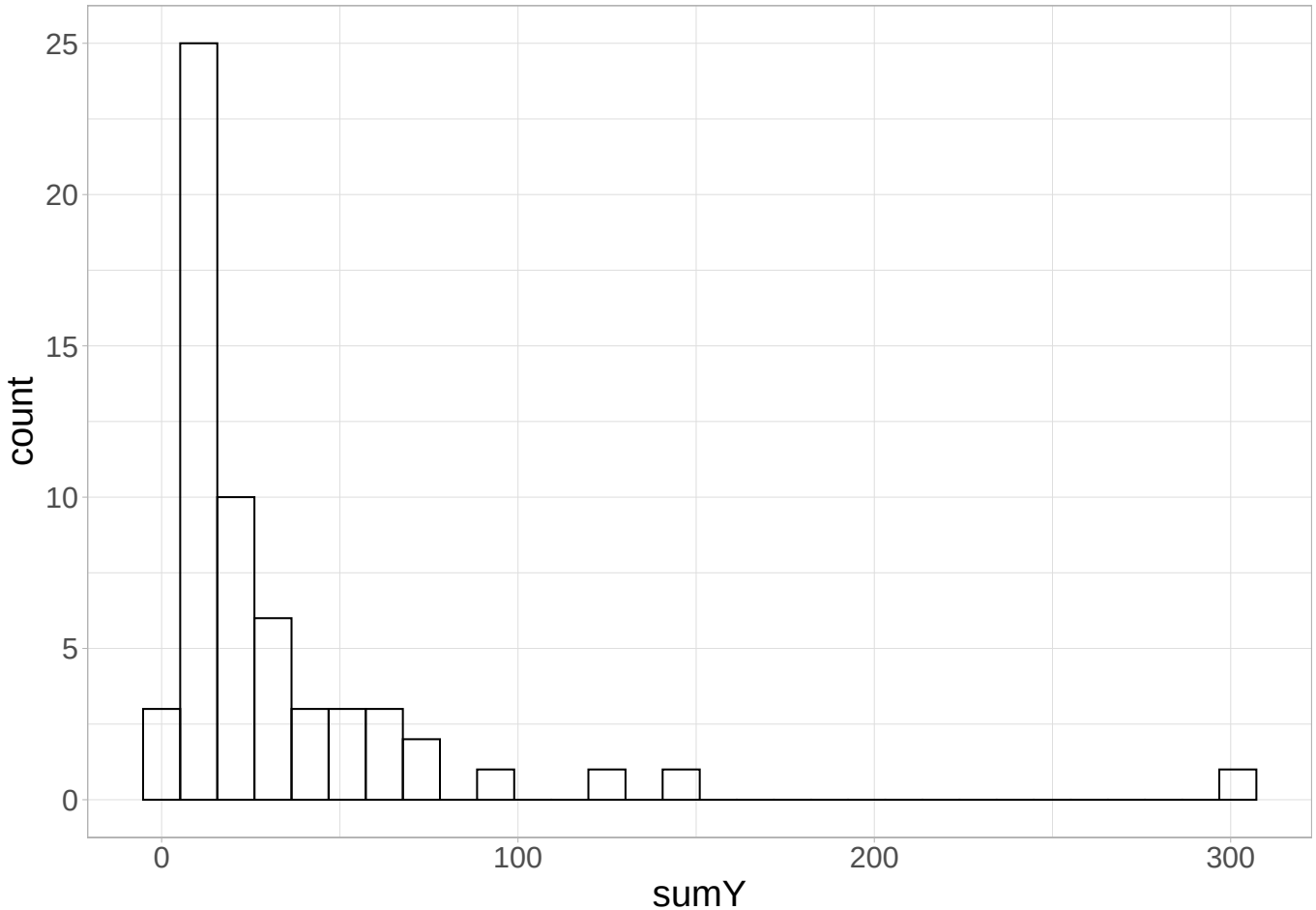
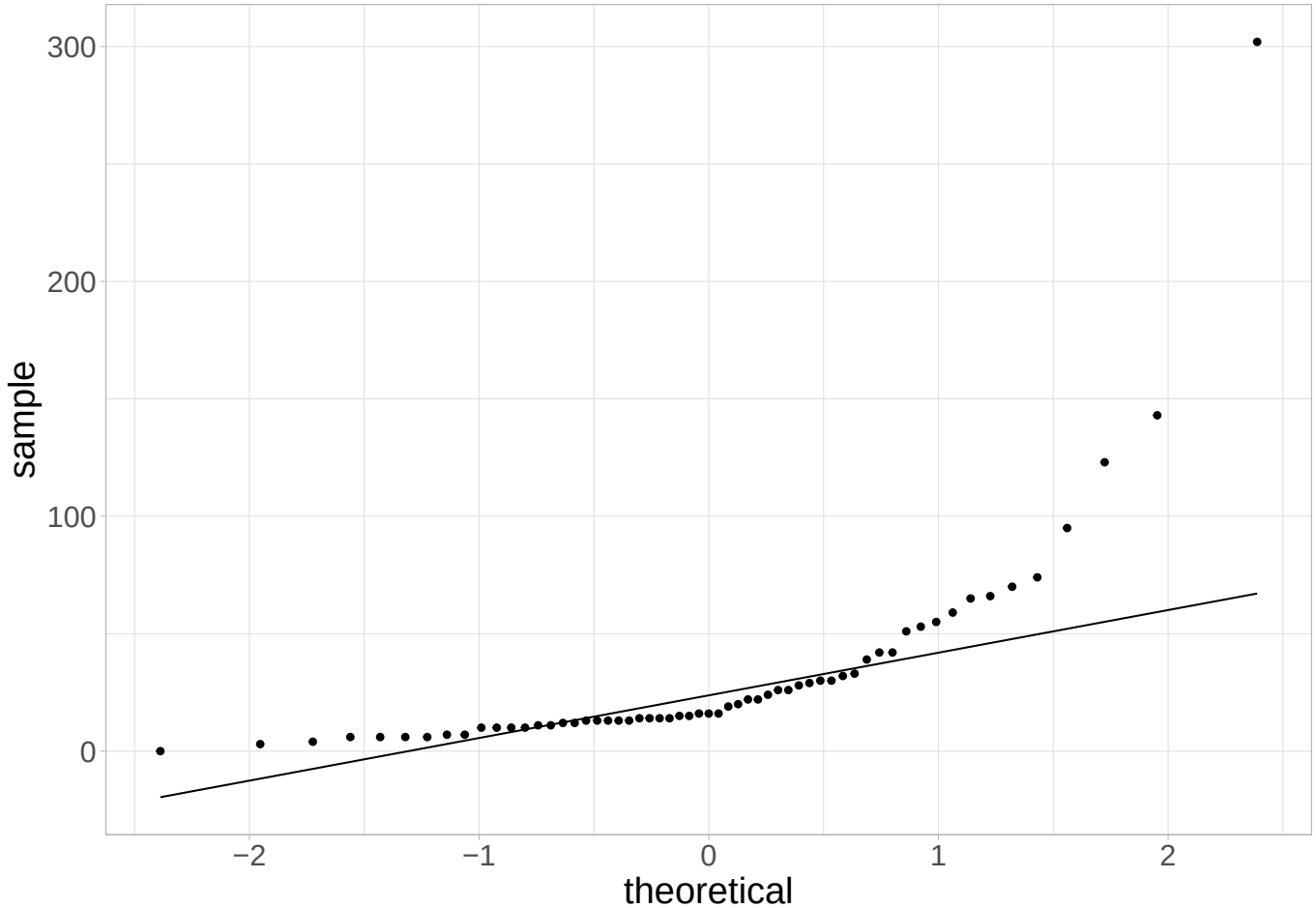


Figure 1: Verteilung von Anzahl der Anfälle nach der Behandlung.

```
ggplot(breslow.dat, aes(sample = sumY)) + stat_qq() + stat_qq_line()
```



```
ggplot(breslow.dat, aes(y = sumY, x = Trt)) + geom_boxplot(fill = "transparent",
    color = "black")
```

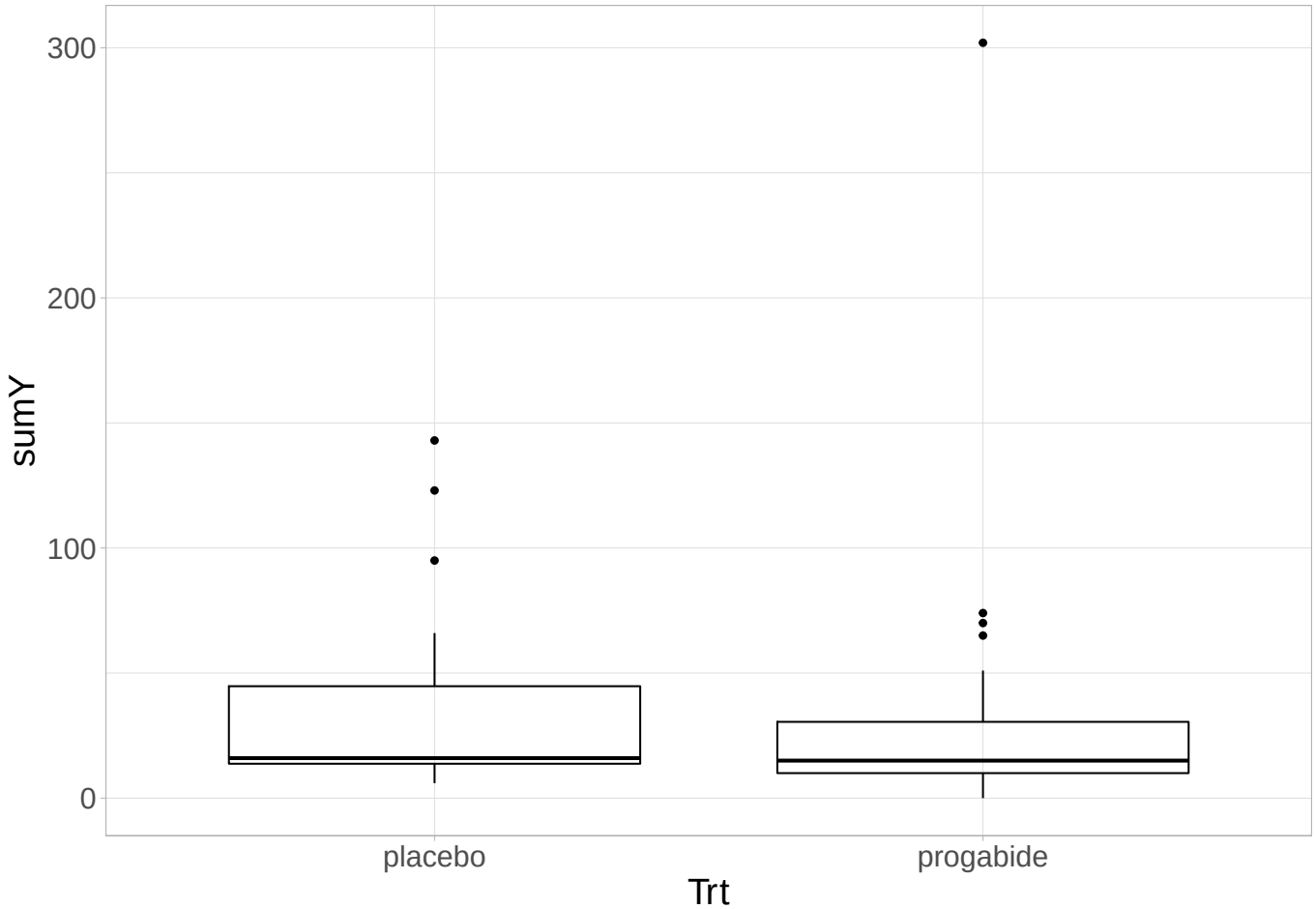


Figure 2: Anzahl der Anfälle vor und nach der Behandlung.

```
ggplot(breslow.dat, aes(y = log10(sumY), x = Trt)) + geom_boxplot(fill = "transparent",
    color = "black")
```

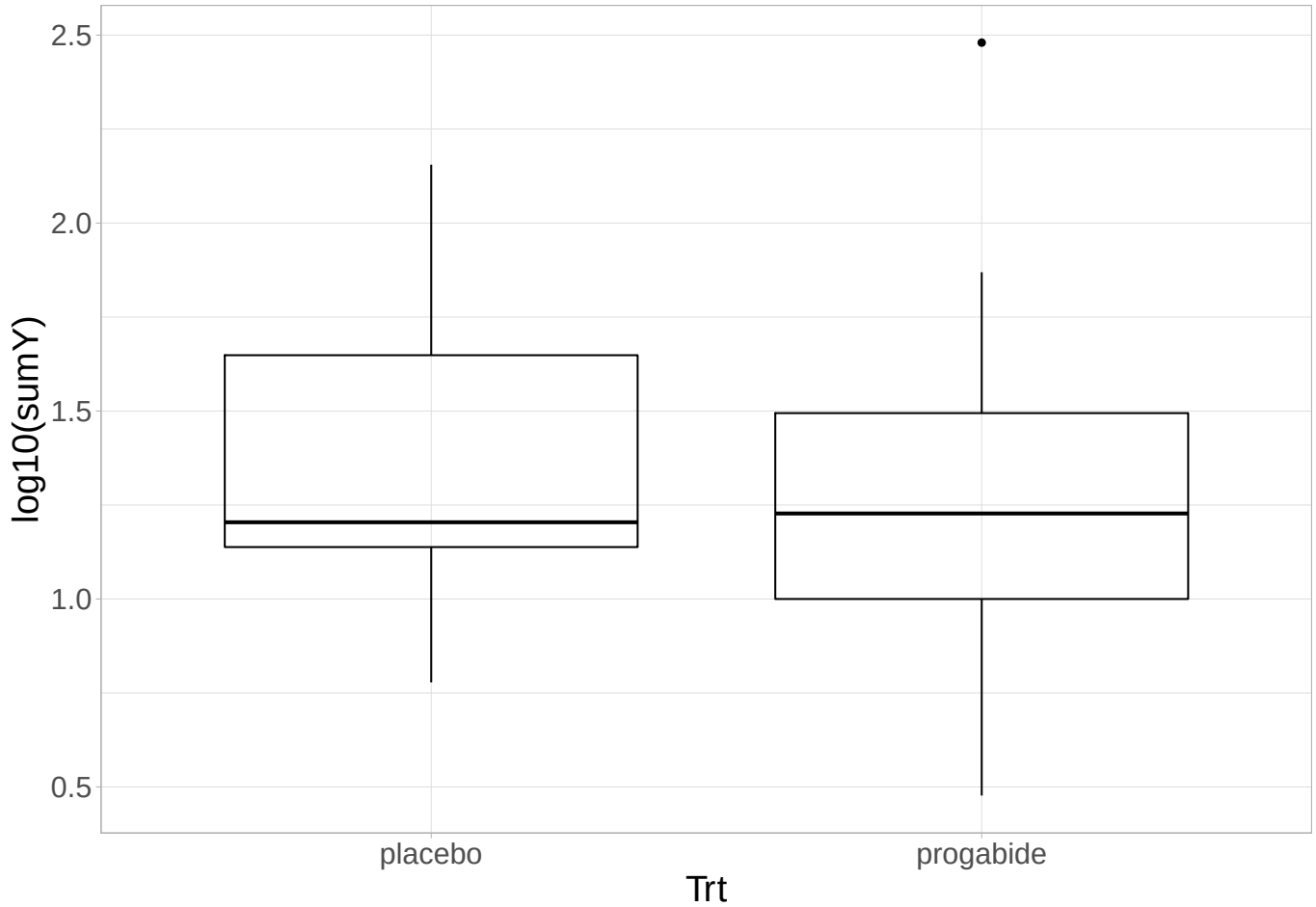


Figure 3: Anzahl der Anfälle vor und nach der Behandlung.

Sie können die Verzerrung der abhängigen Variablen und das mögliche Vorhandensein von Ausreißern deutlich erkennen.

Auf den ersten Blick scheint die Anzahl der Anfälle im Arzneimittel-Gruppe geringer zu sein und weist eine geringere Varianz auf.

Bei Poisson-verteilten Daten geht eine geringere Varianz mit einen kleineren Mittelwert einher.

Poisson regression I

```
fit <- glm(sumY ~ Base + Age + Trt, data = breslow.dat, family = poisson())
summary(fit)
```

```
##
## Call:
## glm(formula = sumY ~ Base + Age + Trt, family = poisson(), data = breslow.dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.0569  -2.0433  -0.9397   0.7929  11.0061
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.9488259  0.1356191  14.370  < 2e-16 ***
## Base          0.0226517  0.0005093  44.476  < 2e-16 ***
## Age           0.0227401  0.0040240   5.651 1.59e-08 ***
## Trtprogabide -0.1527009  0.0478051  -3.194   0.0014 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2122.73  on 58  degrees of freedom
## Residual deviance:  559.44  on 55  degrees of freedom
## AIC: 850.71
##
## Number of Fisher Scoring iterations: 5
```

```
anova(fit)
```

Poisson regression II

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: sumY
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev
## NULL                    58    2122.73
## Base  1  1508.11        57     614.62
## Age   1    44.96        56     569.66
## Trt   1    10.21        55     559.44
```

Poisson regression: theory

- Poisson-Regression

$$\log\left(\mathbb{E}(Y|x)\right) = \beta_0 + \sum_i \beta_i x_i$$

- Poissonsche Wahrscheinlichkeitsdiche:

$$P_{\lambda t}(k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}$$

Wahrscheinlichkeit k Ereignisse in der Zeit t zu sehen. Der Erwartungswert (Mittelwert) und die Varianz von der Poissonverteilung ist durch $\lambda = \mathbb{E}(Y|x)$ gegeben.

```
coef(fit)
```

```
##  (Intercept)         Base          Age Trtprogabide
##   1.94882593   0.02265174   0.02274013  -0.15270095
```

```
confint(fit)
```

```
##                     2.5 %      97.5 %
## (Intercept)   1.68188804  2.21355355
## Base          0.02165019  0.02364694
## Age           0.01483716  0.03061253
## Trtprogabide -0.24646125 -0.05904066
```

Der Regressionsparameter 0,0227 für Alter zeigt an, dass ein Anstieg des Alters um ein Jahr mit einem Anstieg der logarithmischen mittleren Anzahl von Anfällen um 0,02 verbunden ist, wobei die Baseline-Anzahl der Anfälle und der Behandlungsart konstant gehalten werden.

```
exp(coef(fit))
```

```
##  (Intercept)         Base          Age Trtprogabide
##    7.0204403    1.0229102    1.0230007    0.8583864
```

```
exp(confint(fit))
```

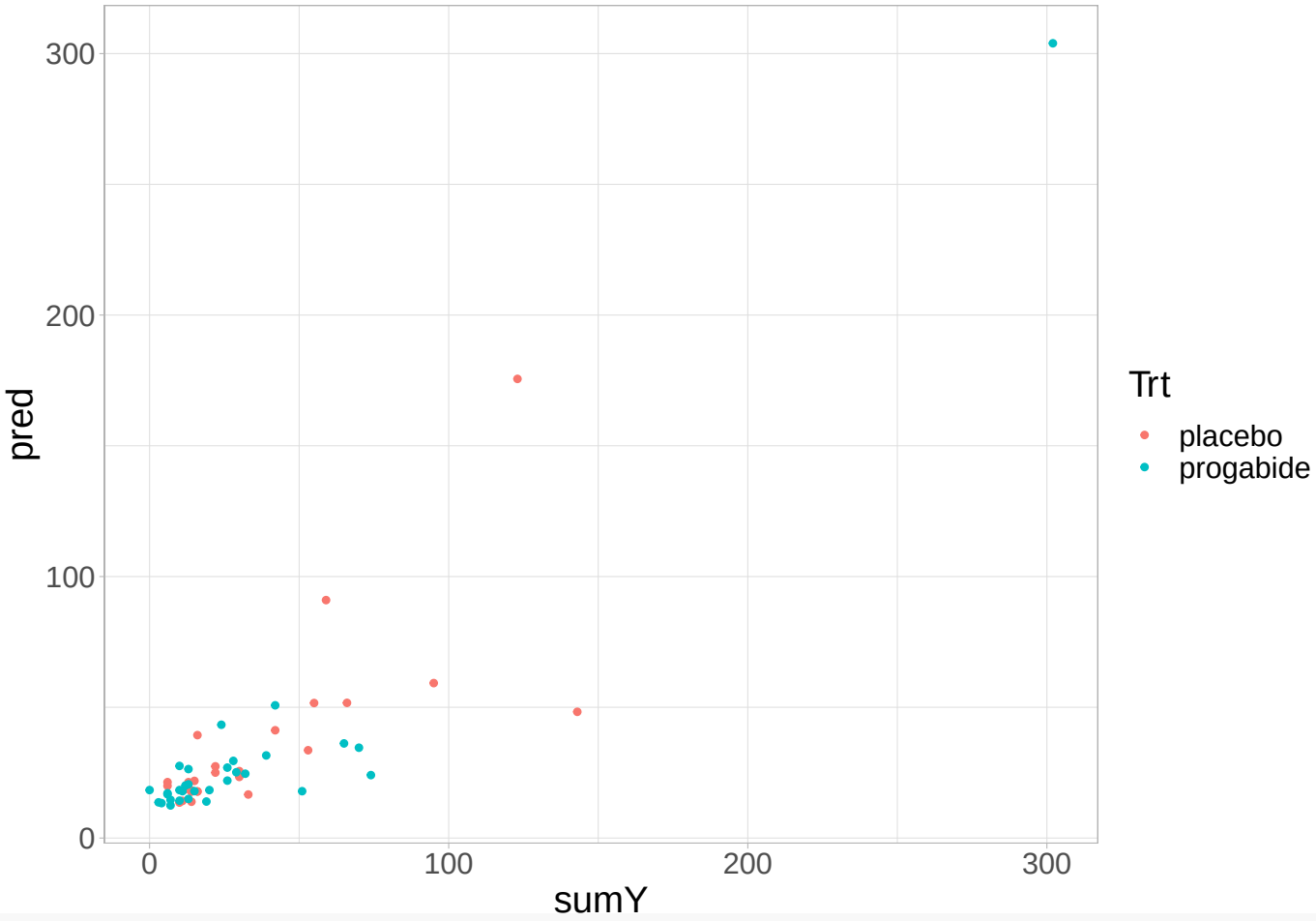
```
##                   2.5 %    97.5 %
## (Intercept)  5.3756959 9.1481671
## Base         1.0218863 1.0239287
## Age          1.0149478 1.0310859
## Trtprogabide 0.7815616 0.9426684
```

Ein Anstieg des Alters um ein Jahr multipliziert die erwartete Anzahl von Anfällen mit 1,023, wobei die anderen Variablen konstant bleiben.

Noch wichtiger ist, dass eine Änderung von *Trt* um eine Einheit (d.h. der Wechsel von Placebo zu Progabid) die erwartete Anzahl von Anfällen mit 0,86 multipliziert. Es ist somit einen Rückgang der Anzahl der Anfälle in der Arzneimittelgruppe um 20% im Vergleich zur Placebogruppe zu erwarten, wobei die Anzahl der Anfälle und das Alter konstant bleiben.

Die potenzierten Parameter im Poisson-Modell wirken sich multiplikativ (wie in der logistischen Regression) und nicht additiv auf die Antwortvariable aus.

Visualisierung




```
# ggplot(breslow.dat, aes(x = Age, y = sumY, colour = Trt)) + geom_point() +
# geom_smooth(method='glm', family='poisson', se=TRUE)
```

Overdispersion

In einer Poisson-Verteilung sind Varianz und Mittelwert gleich. Eine Überdispersion (Overdispersion) tritt bei der Poisson-Regression auf, wenn die beobachtete Varianz der Antwortvariablen größer ist als durch die Poisson-Verteilung vorhergesagt. Beim Umgang mit Zähldaten eine Überdispersion tritt häufig auf und kanns sich negativ auf die Interpretation der Ergebnisse auswirken.

Gründe für Überdispersion

- Das Weglassen einer wichtigen Prädiktorvariablen kann zu einer Überdispersion führen.
- Überdispersion kann auch durch ein Phänomen verursacht werden, das als Zustandsabhängigkeit bezeichnet wird. Innerhalb von Beobachtungen wird angenommen, dass jedes Ereignis in einer Zählung unabhängig ist. Für die Anfallsdaten würde dies bedeuten, dass für jeden Patienten die Wahrscheinlichkeit eines Anfalls unabhängig voneinander ist. Diese Annahme ist jedoch oft unhaltbar. Für eine bestimmte Person ist es unwahrscheinlich, dass die Wahrscheinlichkeit eines ersten Anfalls mit der Wahrscheinlichkeit eines 40. Anfalls übereinstimmt, vorausgesetzt es gab schon 39 Anfälle.
- In Längsschnittstudien kann eine Überdispersion durch die Clusterbildung verursacht werden die in wiederholten Messungen vorkommt.
- Wenn eine Überdispersion vorliegt und diese im Modell nicht berücksichtigt wird, führt es zu sehr kleinen Standardfehlern und Konfidenzintervallen sowie zu sehr liberalen Signifikanztests (d.h. man findet Effekte, die nicht vorhanden sind).

Overdispersion: Beispiel und Test

```
deviance(fit)/df.residual(fit)
```

```
## [1] 10.1717
```

```
library(qcc)
qcc.overdispersion.test(breslow.dat$sumY, type = "poisson")
```

```
##
## Overdispersion test Obs.Var/Theor.Var Statistic p-value
##        poisson data          62.87013  3646.468       0
```

Der Signifikanztest hat einen p-Wert von weniger als 0,05, was stark auf das Vorhandensein einer Überdispersion hinweist.

"Residual deviance" größer als "residual degress of freedom" deutet auf Überdispersion.

Quasi-Poissonische Regression I

Um die Overdispersion zu berücksichtigen kann man quasipoissonische Regression durchführen.

```
fit.od <- glm(sumY ~ Base + Age + Trt, data = breslow.dat, family = quasipoisson())
summary(fit.od)
```

```
##
## Call:
## glm(formula = sumY ~ Base + Age + Trt, family = quasipoisson(),
##     data = breslow.dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.0569  -2.0433  -0.9397   0.7929  11.0061
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.948826   0.465091   4.190 0.000102 ***
## Base          0.022652   0.001747  12.969  < 2e-16 ***
## Age           0.022740   0.013800   1.648 0.105085
## Trtprogabide -0.152701   0.163943  -0.931 0.355702
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 11.76075)
##
##     Null deviance: 2122.73  on 58  degrees of freedom
## Residual deviance:  559.44  on 55  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

Quasi-Poissonische Regression II

Beachten Sie, dass die Parameterschätzungen im Quasi-Poisson-Ansatz mit denen des Poisson-Ansatzes identisch sind.

Die Standardfehler sind jedoch viel größer. In diesem Fall haben die größeren Standardfehler zu p-Werten für Trt (und Alter) geführt, die größer als 0,05 sind.

Wenn Sie die Überdispersion berücksichtigen, gibt es keine ausreichenden Beweise dafür, dass das Medikamentenschema die Anzahl der Anfälle stärker reduziert als die Einnahme eines Placebos, nachdem die Anfallsrate und das Alter zu Studienbeginn kontrolliert wurden.

Poissonregression mit variablen Zeitintervallen

Unsere Diskussion über die Poisson-Regression beschränkte sich auf Antwortvariablen, die eine Zählung über einen festgelegten Zeitraum messen (z. B. Anzahl der Anfälle in einem Zeitraum von acht Wochen, Anzahl der Verkehrsunfälle im vergangenen Jahr oder Anzahl des sozialen Verhaltens an einem Tag). Die Zeitdauer ist über die Beobachtungen hinweg konstant. Sie können jedoch Poisson-Regressionsmodelle anpassen, bei denen der Zeitraum für jede Beobachtung variieren kann. In diesem Fall ist die Antwortvariable eine Rate.

$$\log\left(\frac{\mathbb{E}(Y|x)}{time}\right) = \beta_0 + \sum_i \beta_i x_i$$

$$\log\left(\mathbb{E}(Y|x)\right) = \log(time) + \beta_0 + \sum_i \beta_i x_i$$

```
glm(sumY ~ Base + Age + Trt, data = breslow.dat, offset = log(time), family = poisson)
```

Zero inflated models

Es gibt Fälle, in denen die Anzahl der Nullen in einem Datensatz größer ist als vom Poisson-Modell vorhergesagt. Dies kann auftreten, wenn es eine Untergruppe der Grundgesamtheit gibt, die sich niemals auf das gezählte Verhalten (z.B. Rauchen) einlassen würde.

In solchen Fällen können Sie die Daten mithilfe eines Ansatzes analysieren, der als *zero-inflated Poisson regression* bezeichnet wird. Stellen Sie sich dies als ein Modell vor, das eine logistische Regression (zur Vorhersage struktureller Nullen) und ein Poisson-Regressionsmodell (zur Vorhersage von Zählungen für Beobachtungen, die keine strukturellen Nullen sind) kombiniert. Die zero inflated Poisson-Regression kann z.B. mithilfe der Funktion *zeroinfl()* im *pscl*-Paket angepasst werden.

Literatur

1. Peter Bruce, Andrew Bruce & Peter Gedeck. Practical Statistics for Data Scientists.
2. Kabakoff, R in Action (2d Edition).