“药物治疗是否影响癫痫次数？”-基于泊松回归模型的非正态的因变量 (离散计数型数据) 分析

王昊 *(学号: 201821061107)* 指导教师: 段小刚

2019 年 *12* 月 *12* 日

目录

1 摘要

本文探究了“八周中所发生的癫痫次数与药物治疗有何关系”的问题。我们使用广义线性模型分析非正态的因变量-离散的计数型数据，并探讨了对应的解决上述问题的模型：泊松回归模型，还包括拟合后参数的解释以及过度离势的判别和诊断，以及一些针对特殊情况的变种。

2 背景

2.1 数据来源

患有简单或复杂的部分性癫痫的患者被随机分为两组，一组服用抗癫痫药物氟柳双胺，另一组服用安慰剂。在连续四次的麻醉后门诊就诊中，每一次都报告了过去两周内癫痫发作的数量。

2.2 广义线性模型和 `glm()` 函数

2.2.1 广义线性模型

广义线性模型（Generalized linear model, GLM）与标准线性模型的区别主要在于响应变量 Y 无需满足正态分布，只要服从指数分布族的一种分布即可，模型拟合的形式如下：

$$g\left(\mu_Y\right)=\beta_0+\sum_{j=1}^{p}\beta_j X_j$$

其中，$g\left(\mu_Y\right)$ 是条件均值的函数（称为连接函数），连接函数以及响应变量 Y 所满足的概率分布确定以后，广义线性模型通过最大似然估计（最大似然估计）的迭代推导出各个参数。

广义线性模型包含了非正态因变量的分析，其中，泊松回归（因变量为计数型）扩展了之前线性模型的框架。

2.2.2 `glm()` 函数

R 中主要由 `glm()` 函数拟合广义线性模型，格式如下：

```
# glm(formula, family=family(link=function), data=)
```

以下举例说明 `glm()` 函数对泊松回归的拟合。

假设如下：

- 响应变量：Y
- 三个预测变量：X_1 X_2 X_3
- 包含数据的数据框：mydata

泊松回归使用在给定时间内响应变量为事件发生数目的情形。

假设 Y 服从泊松分布，则线性模型的拟合形式为：

$$\log_e(\lambda)=\beta_0+\sum_{j=1}^{p}\beta_j X_j$$

其中 λ 是 Y 的均值，$log(\lambda)$ 为连接函数，概率分布为泊松分布，泊松回归模型拟合代码如下：

```
# glm(Y~X1+X2+X3, family=position(link="log"), data=mydata)
```

值得一提的是，如果令连接函数 $g(\mu_Y) = \mu_Y$，并设定概率分布为正态（高斯）分布，那么此时的线性回归模型拟合代码如下：

```
# glm(Y~X1+X2+X3, family=gaussian(link="identity"), data=mydata)
```

此时生成的结果与下列代码的结果相同：

```
#lm(Y~X1+X2+X3, data=mydata)
```

此时的 `glm()` 函数所模拟的广义线性模型与标准线性模型等价，也就是说，标准线性模型是广义线性模型的一种特例。

综上所述，广义线性模型不直接拟合响应变量 Y 本身的条件均值，而不是拟合响应变量 Y 的条件均值的一个函数，并假设响应变量服从指数分布族中的某个分布，采用极大似然估计而非最小二乘法为推到依据，极大地扩展了线性模型的适用范围。

3 研究问题

八周中所发生的癫痫次数与药物治疗有何关系。患有简单或复杂的部分性癫痫的患者被随机分为两组，一组服用抗癫痫药物 progabide，另一组服用安慰剂。在连续四次的麻醉后门诊就诊中，每一次都报告了过去两周内癫痫发作的数量。我们研究的问题是：药物治疗是否影响癫痫次数？

4 对数据的描述性分析

Breslow 数据的因变量为 sumY (随机化后八周内癫痫发病数)，预测变量为治疗条件 (Trt)、年龄 (Age) 和前八周内的基础癫痫发病数（base）。包含基础癫痫发病数和年龄的原因在于他们对响应变量有潜在影响。我们最终感兴趣的是药物治疗是否能减少癫痫发病数。

泊松回归通常用于通过一系列连续型和/或类别型预测变量来预测计数型结果变量。以下将采用 `robust` 包中的 Breslow 数据阐述泊松回归的应用。

4.1 安装 robust 包：

```
#install.packages("robust")
```

4.2 查看数据集的统计汇总信息：

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

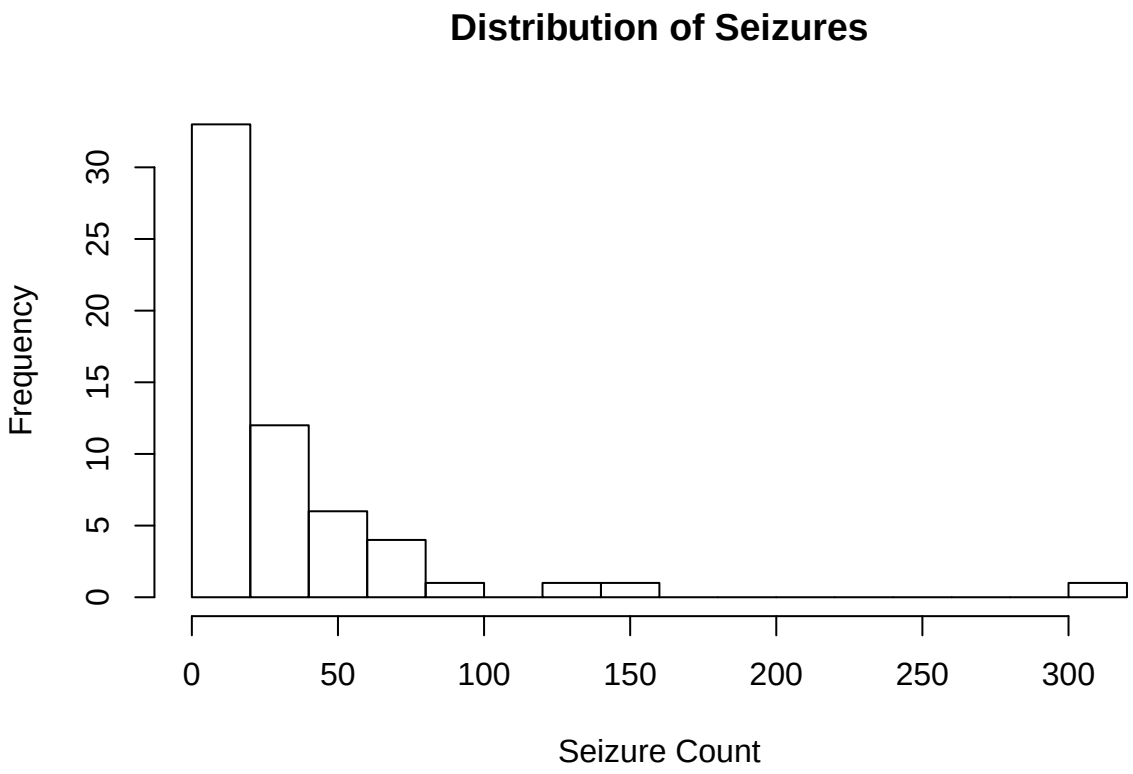
```
##  [1] "ID"    "Y1"    "Y2"    "Y3"    "Y4"    "Base"  "Age"   "Trt"   "Ysum"
## [10] "sumY"  "Age10" "Base4"
```

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

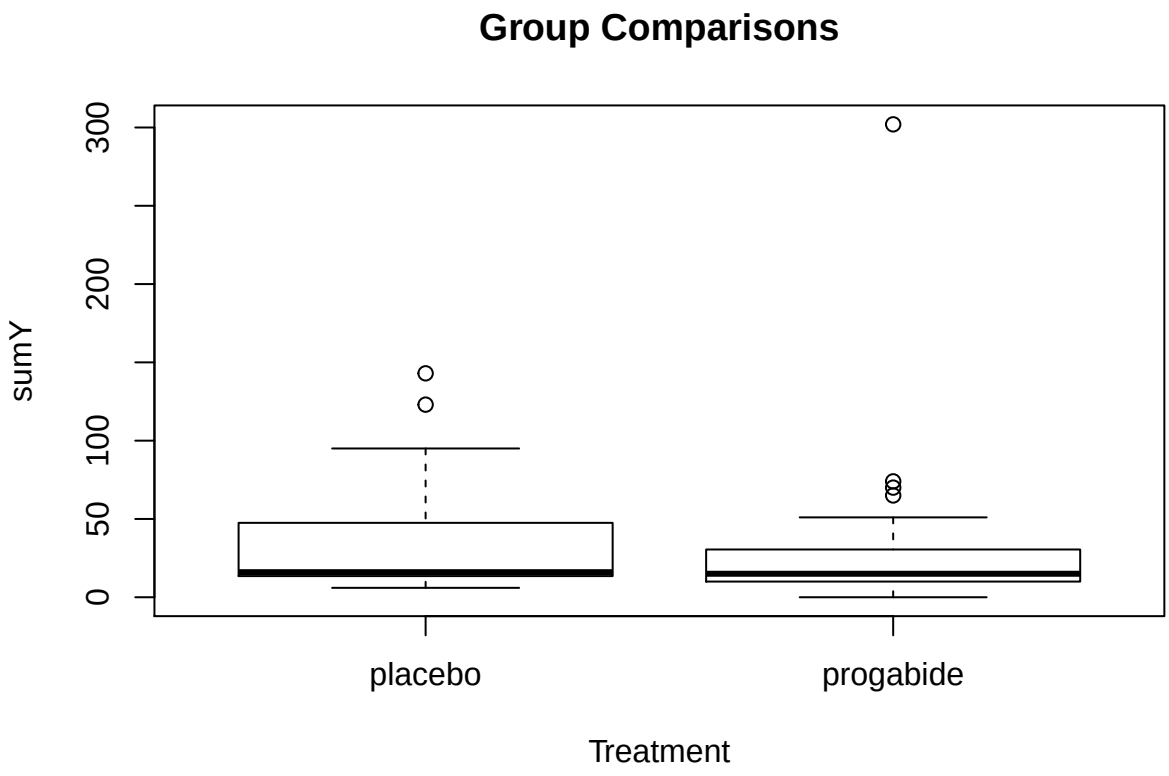
```
##       Base             Age               Trt          sumY
##  Min.   :  6.00   Min.   :18.00   placebo  :28   Min.   :  0.00
##  1st Qu.: 12.00   1st Qu.:23.00   progabide:31   1st Qu.: 11.50
##  Median : 22.00   Median :28.00                  Median : 16.00
##  Mean   : 31.22   Mean   :28.34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

虽然数据集一共有 12 个变量，但是我们只关心之前描述的四个变量。对于因变量，我们用如下代码生成下列图形：

```
opar <- par(no.readonly = TRUE)
par(mfow=c(1,2))
attach(breslow.dat)
hist(sumY, breaks = 20,
       xlab = "Seizure Count",
       main = "Distribution of Seizures")
```




```
boxplot(sumY~Trt,
        xlab="Treatment",
        main = "Group Comparisons")
```




```
par(opar)
```

上图清楚地展示了因变量的偏倚特性和可能存在的离群点，而且在药物治疗下癫痫发病数似乎变小了，且方差也变小了。

5 统计模型的具体形式

下面用泊松回归进行拟合：

```
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
```

6 模型的结果分析与解释

6.1 解释模型参数

Logistic 回归是对数优势比，在泊松回归中，因变量以条件均值的对数形式 $\log_e(\lambda)$ 来建模，响应的初始模型参数为对数均值，同样也可以进行指数化以探求因变量的初始尺度上解释回归系数。以下代码给出了着两者的参数：

```
coef(fit)
```

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

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

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

6.1.1 年龄

年龄的回归参数为 0.0227，表明保持其他预测变量不变, 年龄增加一岁，癫痫发病数的对数均值将相应增加 0.03。

6.1.2 截距项

截距项为预测变量都为 0 时，发病数的对数均值，显然年龄不可能为 0，而且调查对象的基础发病数也都不为 0，因此截距项没有实际意义。

6.1.3 Trtprogabide

从指数化的系数可以看出，保持其他变量不变, 年龄增加一岁，期望的癫痫发病数将乘以 1.023。这意味着年龄的增加与较高的癫痫发病数相关联。更为重要的是, 一单位 Trt 的变化 (即从安慰剂到治疗组), 期望的癫痫发病数将乘以 0.86，换句话说，保持基础癫痫发病数和年龄不变，服药组相对于安慰剂组癫痫发病数降低了 20%。

需要注意的是，与 Logistic 回归中的指数化参数相似, 泊松模型中的指数化参数对响应变量的影响都是成倍增加的，而不是线性相加。

6.2 过度离势

泊松分布的期望和方差相等。当因变量观测的方差比依据泊松分布预测的方差大时，泊松回归可能发生过度离势。可能发生过度离势的原因有如下几个：

- 遗漏了某个重要的预测变量；
- 可能因为事件相关，在在泊松分布的观测中, 计数中每次事件都被认为是独立发生的；
- 在纵向数据分析中，重复测量的数据由于内在群聚特性可导致过度离势。

如果过度离势发生了，在模型中你无法进行解释，那么有可能你会发现并不真实存在的效应。

判断过度离势是否发生的准则依然是残差偏差与残差自由度的比例，如果远远大于 1 则表明存在过度离势。对于上述癫痫发病数数据，它的比例是：

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

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

很显然，比例远大于 1，存在过度离势。

qcc 包提供了一个对泊松模型过度离势的检验方法。

```
#install.packages("qcc")
```

```
library(qcc)
```

```
## Package 'qcc' version 2.7
```

```
## Type 'citation("qcc")' for citing this R package in publications.
```

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

```
##
```

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

显著性检验的 p<0.05，进一步确认存在过度离势。

同样的，可以用类泊松分布代替泊松分布来尝试解决这个问题。

```
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
```

凡事都有两面性，使用类泊松 (quasi-Poisson) 方法所得的参数估计与泊松方法相同，但标准误变大了许多，并且标准误越大将会导致 Trt (和 Age) 的 p 值越大于 0.05。当考虑过度离势, 并控制基础癫痫数和年龄时，并没有充足的证据表明药物治疗相对于使用安慰剂能显著降低癫痫发病次数。

6.3 扩展的泊松模型变种

6.3.1 时间段变化的泊松回归

需要引入一个记录每个观测的时间长度的变量，并将模型从 $\log_e(\lambda) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$ 修改为：$\log_e\left(\frac{\lambda}{\text{time}}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$。同时，为了拟合新模型，需要使用 glm() 函数当中的 offset 选项。以癫痫数据为例：

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

6.3.2 零膨胀的泊松回归

零膨胀的泊松回归主要解决 0 计数的数目比泊松模型预测的数目多的问题，以婚外情数据为例，它将同时拟合两个模型：一个用来预测哪些人又会发生婚外情，另外一个用来预测排除了婚姻忠诚者后的调查对象会发生多少次婚外情。pscl 包中的 `zeroinfl()` 函数可以做零膨胀泊松回归。

6.3.3 稳健泊松回归

robust 包中的 `glmRob()` 函数可以拟合稳健广义线性模型，包括稳健泊松回归，主要应对存在离群点和强影响点的问题。

7 参考文献

[1]KABACOFF R. R in Action: Data Analysis and Graphics with R[M]. Manning, 2015.

8 附录 (代码)

(将以文件形式附在压缩包内)