GeneralizedLinearModel

LSW

2017/4/24

前言

之前介绍了线性回归和方差分析，探索了线性模型能用于预测符合正态分布的响应变量（来自连续型和分类型预测变量）。尽管如此，仍然有很多情况，我们无法合理地假设独立变量符合正态分布，例如：

- 输出变量是分类型。二元变量（0/1，yes / no，成功 / 失败），类别变量，都不属于正态分布。
- 输出变量是个计数（统计交通事故次数），这些变量有值域上的限制，且非负。另外它们的均值和方差是相关联的，而正态分布的均值和方差是独立的，说明它们不是正态分布。

1.广义线性模型

这一节我们主要介绍两种重要的广义线性模型：Logistic regression（分类变量）和Possion regression（计数变量）。广义线性模型是线性模型的推广，它可以写成如下形式:

$g(\mu_r) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$ 这里$g(\mu_r)$是一个条件均值的函数，例如“logit”,“inverse”,“1/mu^2”这些在R包中称为link function.我们理解广义线性模型是通过g函数作用后变为线性模型，因而找到合适的link function g(x)至关重要，下面是常用的link function：

Family	Default link function
binomial	(link = "logit")
gaussian	(link = "identity")
gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log")
quasi	(link = "identity", variance = "constant")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")

- glm()function的常用函数

```
summary()
coefficients();coef()
confint()
residuals()
anova()
plot()
predict()
```

2.Logistic regression

Logistic regression 用于从连续型和分类型预测变量中预测二元输出量，为此我们选取数据集Affairs，它记录了601组婚外情数据，变量包括性别，年龄，婚龄，是否有小孩，宗教信仰程度（1～5），学历、职业和婚姻的自我评价（1～5）。来看一些这个数据集的描述

```
library(AER)
data("Affairs")
attach(Affairs)
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
table(affairs)
## affairs
##   0   1   2   3   7  12
## 451  34  17  19  42  38
```

1).数据准备

简单来看，该数据包括女性315人，男性286人，年龄从17～57岁，其中430人有孩子。从table()函数结果来看，无婚外情包括451人。尽管婚外情次数不等，我们在这里主要关心婚外情是否存在（binary outcome），故将affairs变量转换成因子变量：

```
Affairs$ynaffair[affairs>0]<-1
Affairs$ynaffair[affairs==0]<-0
Affairs$ynaffair<-factor(Affairs$ynaffair,levels = c(0,1),labels = c("No","Yes"))
table(Affairs$ynaffair)
##
##  No Yes
## 451 150
```

2)运用logistic regression 模型:

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

从回归因子对应的p值看出，性别，育儿，教育，职业并不显著，因此，去除这几个变量后重新拟合。

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

```
##
## 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
```

3)剔除变量，调整模型

调整后的模型，每个变量均显著，我们用anova()函数来对比这两个模型的优劣。

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

p值的结果表明，剔除变量的模型和全变量的模型预测的效果类似，即剔除的变量对模型的预测没有影响，故我们偏向于采用更加简单的模型。

4)解释模型的参数

```
coef(fit.remove)
##   (Intercept)           age  yearsmarried religiousness        rating
##    1.93083017   -0.03527112    0.10062274   -0.32902386   -0.46136144
```

给予logistic回归的模型参数，变量都进行了对数化log(odd)，可用指数化还原变量的值：

```
exp(coef(fit.remove))
##   (Intercept)           age  yearsmarried religiousness        rating
##     6.8952321     0.9653437     1.1058594     0.7196258     0.6304248
```

5)评估预测变量对输出结果概率的影响

利用predict()函数进行变量对模型预测结果的影响。首先创建一组随机的数据集，其中的age，yearsmarried，religiousness设为原数据集的均值，rating设置为1~5。

```
testdata<-data.frame(rating=c(1,2,3,4,5),age=mean(Affairs$age),yearsmarried=mean(A
ffairs$yearsmarried),religiousness=mean(Affairs$religiousness))
testdata$prob<-predict(fit.remove,newdata = testdata,type="response")
testdata
##   rating      age yearsmarried religiousness      prob
## 1      1 32.48752     8.177696      3.116473 0.5302296
## 2      2 32.48752     8.177696      3.116473 0.4157377
## 3      3 32.48752     8.177696      3.116473 0.3096712
## 4      4 32.48752     8.177696      3.116473 0.2204547
## 5      5 32.48752     8.177696      3.116473 0.1513079
```

固定其他四个变量之后，我们发现rating的值从1～5变化，导致婚外情概率从0.53降到0.15。 同样的，我们研究年龄的变化对结果的影响：

```
testdata<-data.frame(rating=mean(Affairs$rating),age=seq(17,57,10),yearsmarried=me
an(Affairs$yearsmarried),religiousness=mean(Affairs$religiousness))
testdata$prob<-predict(fit.remove,newdata = testdata,type="response")
testdata
##    rating age yearsmarried religiousness      prob
## 1 3.93178  17     8.177696      3.116473 0.3350834
## 2 3.93178  27     8.177696      3.116473 0.2615373
## 3 3.93178  37     8.177696      3.116473 0.1992953
## 4 3.93178  47     8.177696      3.116473 0.1488796
## 5 3.93178  57     8.177696      3.116473 0.1094738
```

我们发现，年龄从17～57变化，婚外情概率从0.335降到0.11。可见rating的影响最大，其他的变量也可如此推断。

6)超散布性(overdispersion)

在数据分析和建模的过程中，我们通常需要假设数据变量服从某个分布，再利用数据和估计方法对参数进行估计，当分布被确定后，均值和方差也被确定，若此时观测数据的方差系统地大于分布假设条件下的方差，就出现了“超散布性”，若小于系统方差，则出现了“超聚集性”。

- 一种用于检测超散布性的方法是比较残差偏离值(Residual deviance)和自由度的比率，如果 $\phi = Residualdeviance/Residualdf$ 大于1，则说明数据超散步性。 下面利用Affair数据及进行演示：

```
fit.remove$deviance/fit.remove$df.residual
## [1] 1.03248
```

值很接近1，说明“超散布性”不存在。

- 另一种方法是，拟合模型两次，第一次family=binomial,第二次family=quasibinomial。然后进行卡方检

验，原假设为$\phi = 1$。

```
fit.od<-glm(ynaffair~age+yearsmarried+religiousness+rating,family = binomial(),dat
a=Affairs)
fit.new<-glm(ynaffair~age+yearsmarried+religiousness+rating,family = quasibinomial
(),data=Affairs)
pchisq(summary(fit.new)$dispersion*fit.od$df.residual,fit.od$df.residual,lower=F)
## [1] 0.340122
```

P值结果为0.34，表明无显著性，证明“超散布性”不存在。

7)logistic 回归的扩展模型

- 稳健logistic回归：glmRob()函数用于拟合稳健的广义线性模型，适用于拟合模型中数据存在离群点和强影响点。
- 多项式分布回归：响应变量包含两个以上的无序类（例如已婚，寡居，离婚）时，可使用mlogit()函数拟合多项logistic回归。
- 序数logistic回归：当响应变量是一组有序的类别（例如信用为好，良，差）时，可使用rms包中的lrm()函数拟合序数logistic回归。

2.Poisson回归

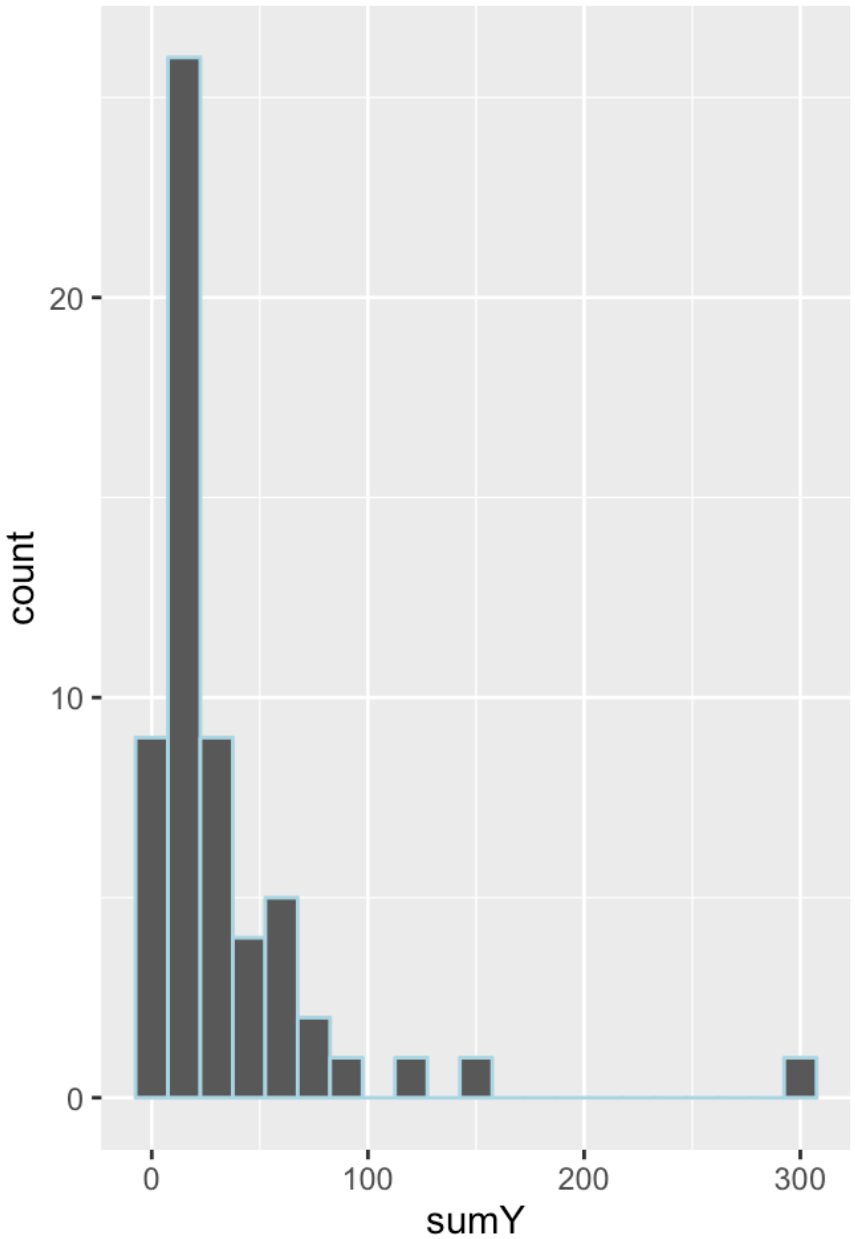
Poisson回归对于预测变量是连续型和分类型，响应变量是计数型的模型很适用。他有两个假设条件：一是具备相同特征和同时的不同对象的人时风险时同质的，其次，当样本量越来越大时，频数的均值趋近于方差。调用Poisson模型的公式如下：

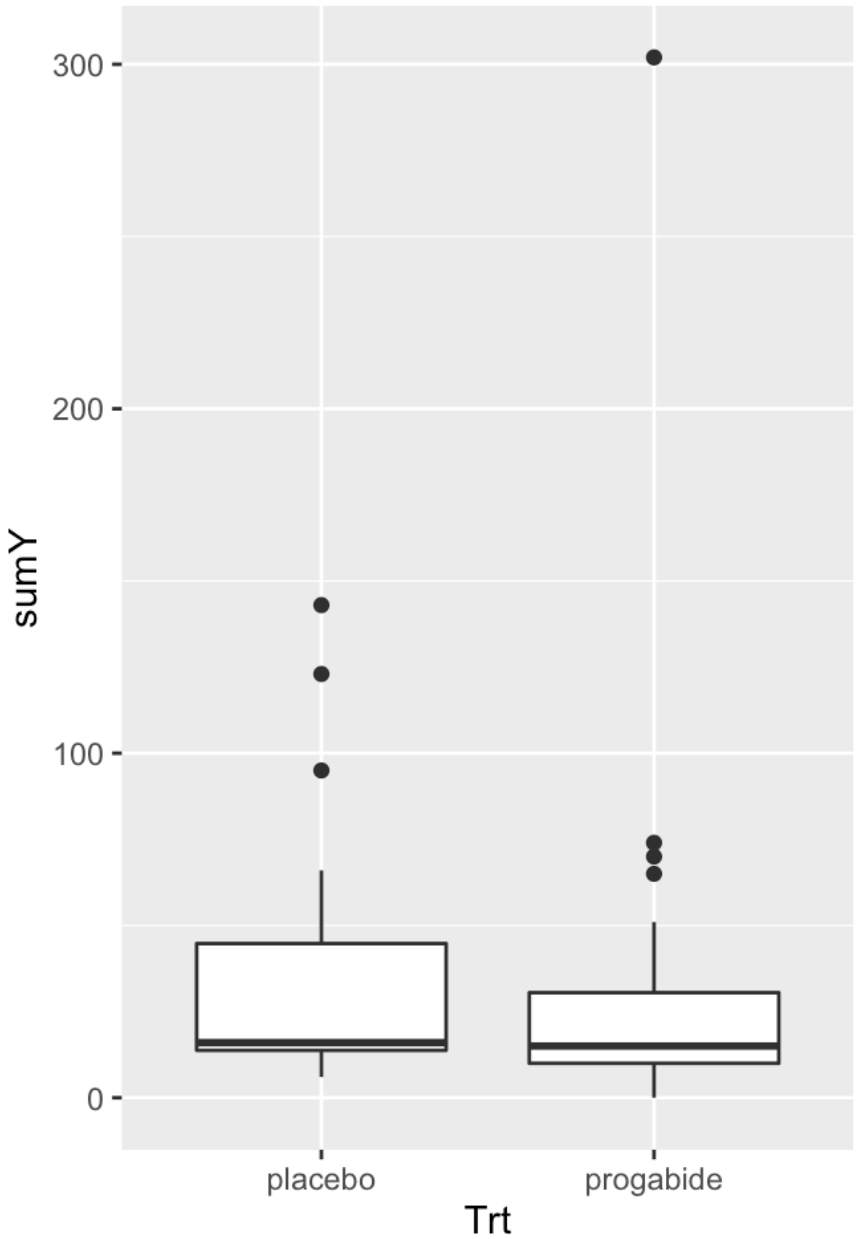
```
myfit<-glm(y~x1+x2+...+xn,data=,family=possion)
```

1)数据准备

robust包中Breslow癫痫数据记录了治疗初期八周内，抗癫痫药物对癫痫发病数的影响。响应变量为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），在这个数据集中，我们感兴趣的是药物治疗能否减少癫痫发病数。

```
data(breslow.dat,package = "robust")
library(ggplot2)
g1<-ggplot(breslow.dat,aes(sumY))+geom_histogram(color="lightblue",binwidth = 15)
g2<-ggplot(breslow.dat,aes(Trt,sumY))+geom_boxplot()
library(gridExtra)
grid.arrange(g1,g2,nrow=1)
```





从图中可以清楚的看到因变量的偏移特性及可能的离群点。药物治疗下癫痫的发病数似乎变小，且方差也变小了。

2)构建模型

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

结果说明治疗药物对癫痫的发病数有改善。

- 关于低方差数据的Poisson建模，可参考统计之都的文章Poisson分布低方差数据建模(https://cos.name/tag/纯生过程模型)。