统计学与 R 语言

第 22 讲　泊松回归

张敬信

2022 年 5 月 11 日

哈尔滨商业大学

一. 泊松回归原理

有时候需要对计数因变量建模，比如：

- 某年的摩托车死亡人数是否与某州的头盔法有关?
- 每天就诊的哮喘病的人数是否因空气污染指数的不同而不同?

因变量是单位时间或空间内的计数，大致服从泊松分布，则适合用泊松回归或负二项回归模型。

计数 y 的取值是 $0, 1, 2, \cdots$, 没有负数和小数，通常接一个连接函数：

$$\ln(y) \in (-\infty, \infty)$$

于是，泊松回归就是这样的广义线性模型：

$$\ln(y_i) = \mathrm{x}_i\beta$$

表示条件观测值 $y_i|\mathrm{x}_i$ 服从参数 λ_i 的泊松分布。

注意：泊松回归模型不包含误差项 ε, 因为 λ 完全决定了泊松分布的均值和方差。

泊松回归的模型假设：

- **泊松响应**：因变量是单位时间或空间的计数，可由泊松分布描述；
- **独立性**：观测值必须是相互独立的；
- **均值 = 方差**：根据定义，泊松随机变量的均值必须等于其方差；
- **线性**：均值的对数 $\ln(\lambda)$, 必须是 x 的线性函数。

```
library(tidyverse)
library(broom)
```

二. 泊松回归案例：美国校园犯罪

```
df = read_csv("datas/crime_campus.csv")
df
#> # A tibble: 81 x 6
#>   Enrollment type     nv nvrate enroll1000 region
#>        <dbl> <chr> <dbl>  <dbl>       <dbl> <chr>
#> 1       5590 U        30  5.37         5.59 SE
#> 2        540 C         0  0            0.54 SE
#> 3      35747 U        23  0.643       35.7  W
#> # ... with 78 more rows
```

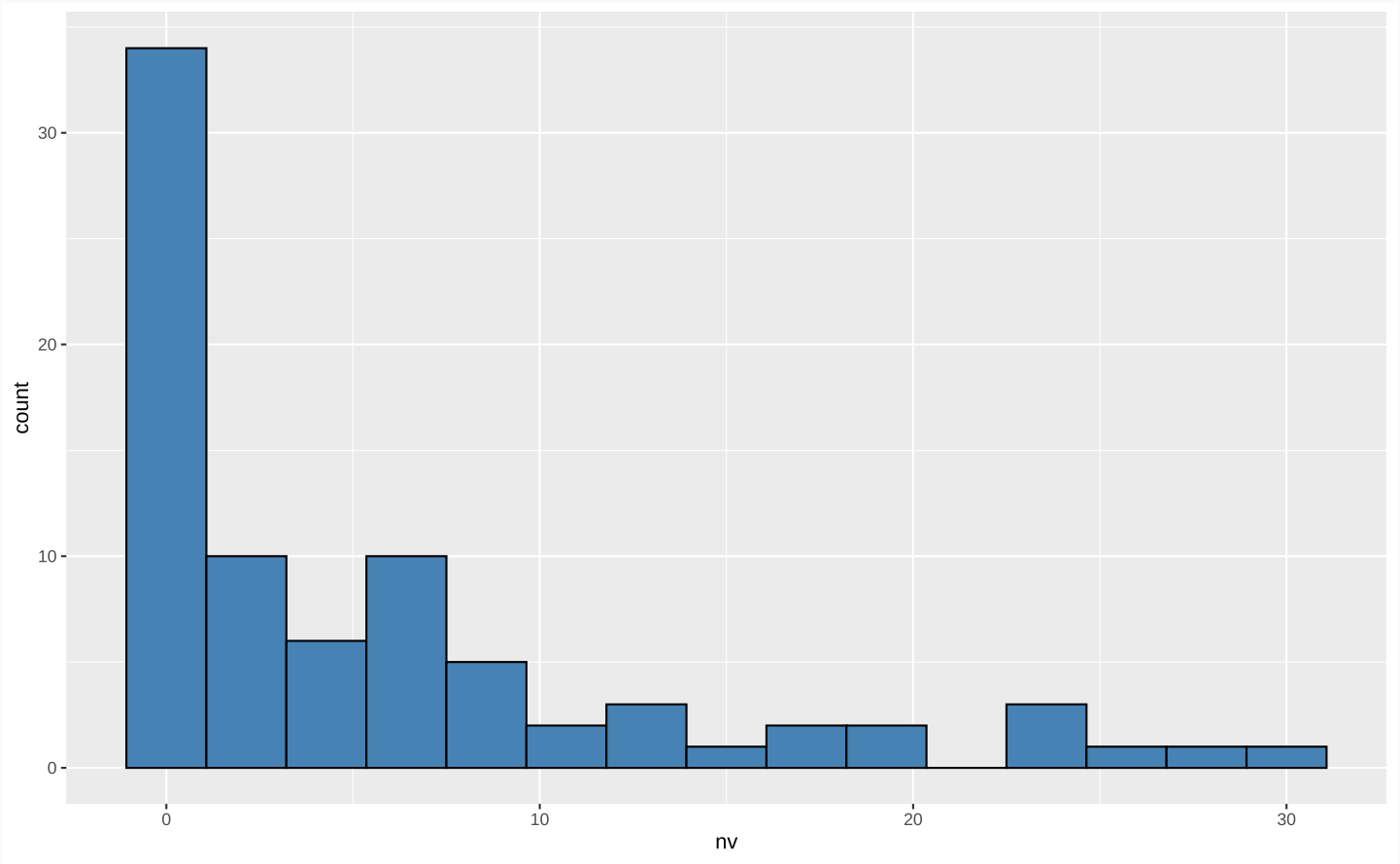
变量 enroll1000 为以千为单位的学生人数，type 为学校类型（学院/大学），region 为学校所在地区，nv 为暴力犯罪人数，nvrate 为暴力犯罪率。

建立泊松回归模型，考察暴力犯罪与学校类型、学校所在地区之间的关系。

1. 探索数据

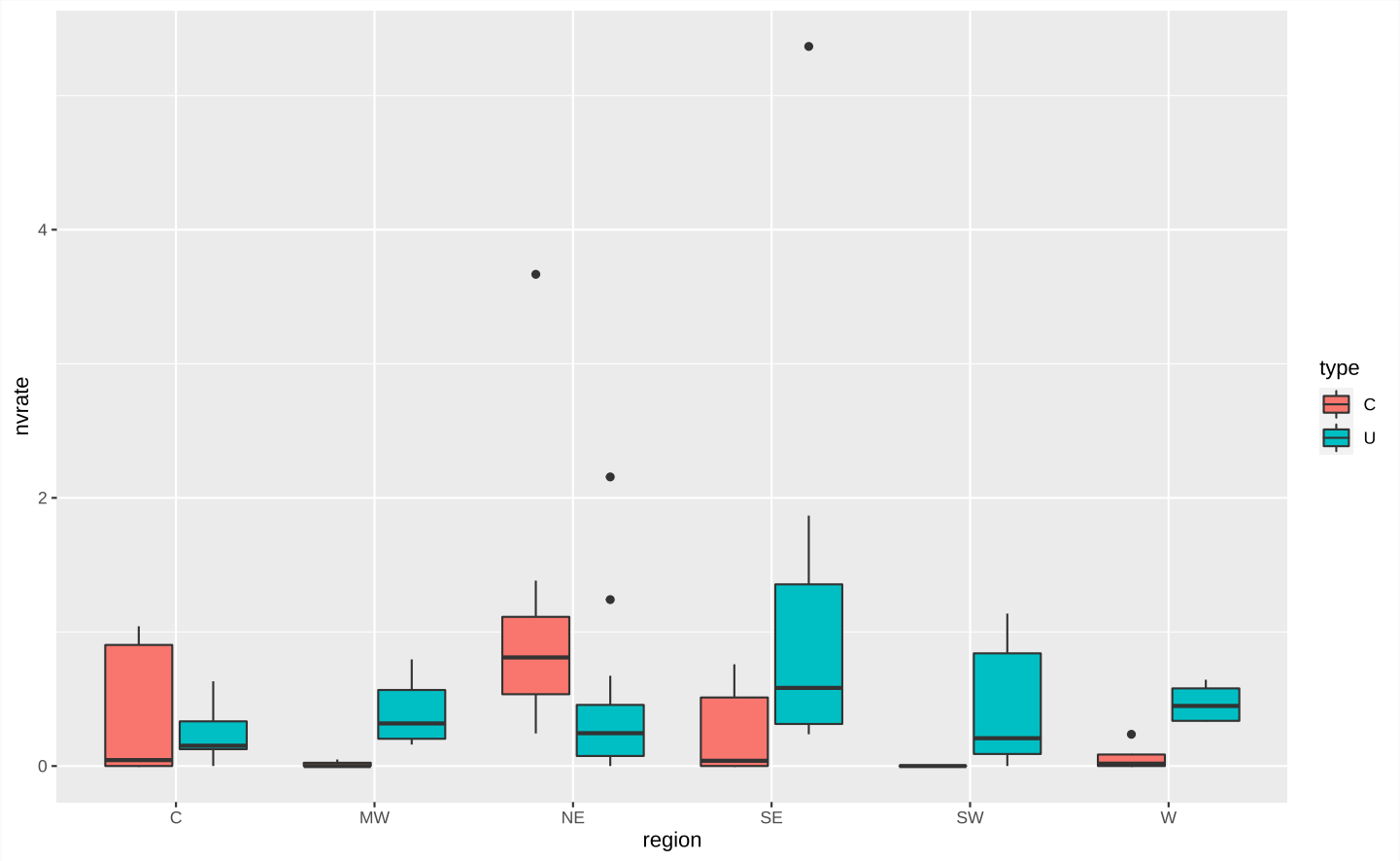
- 探索因变量犯罪数（nv）的分布

```
ggplot(df, aes(nv)) +
  geom_histogram(bins = 15, color = "black",
                 fill = "steelblue")
```



- 探索自变量 region 和 type 对犯罪率（nvrate）的影响关系

```
ggplot(df, aes(region, nvrate, fill = type)) +
  geom_boxplot()
```



- 检查犯罪数、犯罪率的分组均值和方差是否大致相等

```
df %>%
  group_by(region, type) %>%
  summarise(across(starts_with("nv"),
                   list(mean= mean, var = var)), n = n())
#> # A tibble: 12 x 7
#> # Groups:   region [6]
#>   region type  nv_mean nv_var nvrate_mean nvrate_var     n
#>   <chr>  <chr>   <dbl>  <dbl>       <dbl>      <dbl> <int>
#> 1 C      C       1.6    3.3         0.398    0.278        5
#> 2 C      U       4.75  30.9         0.222    0.0349      12
#> 3 MW     C       0.333  0.333       0.0163   0.000793     3
#> # ... with 9 more rows
```

泊松回归是对计数建模，比起犯罪数更合理的是用犯罪率（剔除校园人数的影响），方法就是将校园人数作为**偏移量**加入模型。因为

$$\ln(\frac{y}{\text{enroll1000}}) = X\beta$$

等价于：

$$\ln(y) = X\beta + \ln(\text{enroll1000}))$$

2. 初始模型

```
mdl0 = glm(nv ~ type + region, data = df, family = poisson,
           offset = log(enroll1000))
glance(mdl0)
#> # A tibble: 1 x 8
#>   null.deviance df.null logLik   AIC   BIC deviance df.res
#>           <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>
#> 1          491.      80  -322.  658.  675.     426.
```

```
tidy(mdl0)
#> # A tibble: 7 x 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)  -1.60       0.171    -9.36  8.34e-21
#> 2 typeU         0.340      0.132     2.57  1.02e- 2
#> 3 regionMW      0.0994     0.178     0.560 5.75e- 1
#> 4 regionNE      0.781      0.153     5.10  3.33e- 7
#> 5 regionSE      0.877      0.153     5.72  1.04e- 8
#> 6 regionSW      0.503      0.185     2.72  6.63e- 3
#> 7 regionW       0.273      0.187     1.46  1.45e- 1
```

Northeast 和 South 与 Central 地区有显著的不同 (P 值)；回归系数 0.778 意味着 Northeast 每千人的犯罪率是参照地区（Central）的接近 $e^{0.778} = 2.2$ 倍，进一步计算其置信区间为 $0.778 \pm 1.96 * 0.153 = (1.61, 2.94)$.

2. 考虑地区与学校类型的交互影响

```
mdl1 = glm(nv ~ type * region, data = df,
           family = poisson, offset = log(enroll1000))
glance(mdl1)
#> # A tibble: 1 x 8
#>   null.deviance df.null logLik   AIC   BIC deviance df.res
#>           <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>
#> 1          491.      80  -281.  587.  615.     345.
```

```
tidy(mdl1)
#> # A tibble: 12 x 5
#>    term           estimate std.error statistic   p.value
#>    <chr>             <dbl>     <dbl>     <dbl>     <dbl>
#>  1 (Intercept)      -1.47      0.354   -4.17   0.0000305
#>  2 typeU             0.196     0.378    0.519  0.604
#>  3 regionMW         -1.98      1.06    -1.86   0.0624
#>  4 regionNE          1.55      0.382    4.07   0.0000477
#>  5 regionSE          0.251     0.486    0.516  0.606
#>  6 regionSW        -15.5     737.      -0.0210 0.983
#>  7 regionW          -1.83      0.791   -2.32   0.0204
#>  8 typeU:regionMW    2.20      1.08     2.04   0.0413
#>  9 typeU:regionNE   -1.07      0.420   -2.55   0.0109
#> 10 typeU:regionSE    0.694     0.512    1.36   0.175
#> 11 typeU:regionSW   16.1     737.       0.0218 0.983
#> 12 typeU:regionW     2.41      0.814    2.96   0.00306
```

可见，地区与学校类型的交互影响是显著的，那么两个模型是否有显著差异呢?

```
anova(mdl0, mdl1, test = "Chisq")
#> Analysis of Deviance Table
#>
#> Model 1: nv ~ type + region
#> Model 2: nv ~ type * region
#>   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
#> 1        74        426
#> 2        69        345  5     81.3  4.5e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

$\chi^2 = 81.312$, P 值几乎为 0, 表明两个模型有显著差异。

写出泊松回归模型：

$$\begin{aligned}\ln(nv) = &-1.47 + 0.196 * typeU + \cdots + 2.41 * typeU * regionW \\ &+ \ln(enroll1000)\end{aligned}$$

Cameron 和 Trivedi（2009）建议对参数估计值使用稳健标准误，以控制对方差等于均值的分布假设的轻微违反：

```
library(sandwich)
cov1 = vcovHC(mdl1, type = "HC0")
std.err = sqrt(diag(cov1))
r.est = cbind(Estimate= coef(mdl1), `Robust SE` = std.err,
              `P 值` = 2 * pnorm(abs(coef(mdl1) / std.err),
                                 lower.tail = FALSE),
               LL = coef(mdl1) - 1.96 * std.err,
               UL = coef(mdl1) + 1.96 * std.err)
```

```
r.est
#>                 Estimate Robust SE      P 值      LL
#> (Intercept)       -1.474     0.713 3.86e-02  -2.871  -0.077
#> typeU              0.196     0.748 7.93e-01  -1.270   1.661
#> regionMW          -1.977     0.853 2.05e-02  -3.649  -0.304
#> regionNE           1.553     0.783 4.74e-02   0.018   3.087
#> regionSE           0.251     0.898 7.80e-01  -1.510   2.011
#> regionSW         -15.463     1.134 2.26e-42 -17.685 -13.241
#> regionW           -1.834     0.878 3.68e-02  -3.555  -0.112
#> typeU:regionMW     2.196     0.902 1.48e-02   0.429   3.963
#> typeU:regionNE    -1.070     0.872 2.20e-01  -2.779   0.639
#> typeU:regionSE     0.694     0.976 4.77e-01  -1.219   2.608
#> typeU:regionSW    16.084     1.214 4.77e-40  13.704  18.463
#> typeU:regionW      2.411     0.918 8.66e-03   0.611   4.210
```

注：想要更好解释，可以取出非 P 值列，再作用 exp().

模型预测新数据

```
library(modelr)
newdat = data_grid(df, type, region) %>%
  mutate(nv = mean(df$nv), enroll1000 = mean(df$enroll1000))
rslt = bind_cols(newdat,
                 predict(mdl1, newdata = newdat,
                         type = "response", se = TRUE))
```

```
rslt
#> # A tibble: 12 x 7
#>    type  region    nv enroll1000          fit   se.fit res
#>    <chr> <chr>  <dbl>      <dbl>        <dbl>    <dbl>
#>  1 C     C       5.94       13.9  3.18          1.13
#>  2 C     MW      5.94       13.9  0.441         0.441
#>  3 C     NE      5.94       13.9 15.0           2.17
#>  4 C     SE      5.94       13.9  4.09          1.36
#>  5 C     SW      5.94       13.9  0.000000613 0.000452
#>  6 C     W       5.94       13.9  0.509         0.360
#>  7 U     C       5.94       13.9  3.87          0.513
#>  8 U     MW      5.94       13.9  4.82          0.618
#>  9 U     NE      5.94       13.9  6.28          0.715
#> 10 U     SE      5.94       13.9  9.96          0.929
#> 11 U     SW      5.94       13.9  7.20          0.989
#> 12 U     W       5.94       13.9  6.89          0.975
```

关于**过分散**，即观测到的响应变量的方差大于期望的泊松分布的方差。过分散会导致奇异的标准误检验和不精确的显著性检验。当出现过分散时，仍可用 `glm()` 函数拟合泊松回归，但此时需要将泊松分布改为准泊松分布 `(family = quasipoisson)`。

另一种处理过离散的方法是，改用负二项回归。

三. 负二项回归

泊松分布的参数 λ，既是均值又是方差，而实际中的计数分布通常会有一个不等于其均值的方差（欠分散或过分散），此时更适合采用负二项分布。

从数学上来说，负二项分布可以看作是，其参数 λ 为服从 Γ 分布随机变量的泊松分布，即

$$Y|\lambda \sim \text{Poisson}(\lambda), \qquad \lambda \sim \text{Gamma}(r, \frac{1-p}{p})$$

则 $Y \sim \text{NegBinom}(r, p)$.

负二项分布，也是描述非负整数出现的概率，其方差和均值并不相等，其方差是均值的函数，并有一个额外的参数 r，称为分散参数：

$$E(Y) = \frac{pr}{1-p} = \mu, \qquad Var(Y) = \frac{pr}{(1-p)^2} = \mu + \frac{\mu^2}{r}$$

- 校园犯罪数据的负二项回归建模：

```
library(MASS)
mdl2 = glm.nb(nv ~ type + region + region:type
              + offset(log(enroll1000)), data = df)
glance(mdl2)
#> # A tibble: 1 x 8
#>   null.deviance df.null logLik     AIC   BIC deviance df.r
#>           <dbl>   <int> <logLik> <dbl> <dbl>    <dbl>
#> 1          133.      80 -198      422.  453.     81.2
```

```
tidy(mdl2)
#> # A tibble: 12 x 5
#>    term           estimate   std.error  statistic p.value
#>    <chr>             <dbl>       <dbl>      <dbl>   <dbl>
#>  1 (Intercept)      -1.00        0.504 -1.99       0.0471
#>  2 typeU            -0.429       0.575 -0.745      0.456
#>  3 regionMW         -2.51        1.28  -1.96       0.0495
#>  4 regionNE          1.13        0.598  1.89       0.0587
#>  5 regionSE         -0.274       0.732 -0.375      0.708
#>  6 regionSW        -30.9   1333002.     -0.0000232 1.00
#>  7 regionW          -2.23        0.961 -2.32       0.0204
#>  8 typeU:regionMW    3.01        1.35   2.23       0.0261
#>  9 typeU:regionNE   -0.551       0.708 -0.778      0.437
#> 10 typeU:regionSE    1.84        0.832  2.21       0.0268
#> 11 typeU:regionSW   31.6   1333002.      0.0000237 1.00
#> 12 typeU:regionW     2.91        1.09   2.67       0.00749
```

负二项回归模型的写法、估计稳健标准误、预测新数据与泊松回归是一样的(略)。

注 1：若因变量计数数据中有很多 0（真实的和多余的），适合构建零膨胀泊松/负二项回归模型（估计一个计数模型，一个多余 0 点模型）。

注 2：若因变量数据是删减的，不缺少数据，只是高于某阈值都变成该阈值，适合构建 Tobit 回归模型。

注 3：若因变量数据是截断的，相当于缺少一部分分布曲线的数据，适合构建截断回归模型。

本篇主要参阅 (张敬信, 2022), Beyond Multiple Linear Regression，模板感谢 (黄湘云, 2021), (谢益辉, 2021).

参考文献

张敬信 (2022). *R* 语言编程：基于 *tidyverse*. 人民邮电出版社, 北京.

谢益辉 (2021). *rmarkdown: Dynamic Documents for R*.

黄湘云 (2021). *Github: R-Markdown-Template*.