Statistical Modeling
BANA 7042

Lecture 6: Modeling count data

Dr. Dungang Liu

What is count data?

- A type of data in which the observations take non-negative integer values {0, 1, 2, 3, ….} .

- Examples
  - The number of patients who come to the ER of Children's Hospital between 9PM and 1AM.
  - The number of shoppers in Kenwood Towne Centre on a calendar day.
  - The number of Google searches (in a week) for flights to Shanghai right before Lunar New Year.

Compared to other data types …

- In which ways, count data is different from binomial data or rank (ordinal) data?

- Binomial data (e.g., the no. of damages out of 6 O-rings) has an upper bound, whereas count data is unbounded from above.

- Ordinal data (e.g., ratings of a product) only reflects ranks, its values {0,1,2,3} should not be interpreted as numbers. The values of count data {0,1,2,…} are the numbers of the occurrence of a specific event.

What distribution to model count data?

- Poisson distribution

$$\Pr\{Y = y\} = \frac{e^{-\mu}\mu^y}{y!}, \quad y = 0, 1, 2, \ldots$$

- How many parameters used in Poisson distribution?

$$\mu = E(Y) = Var(Y)$$

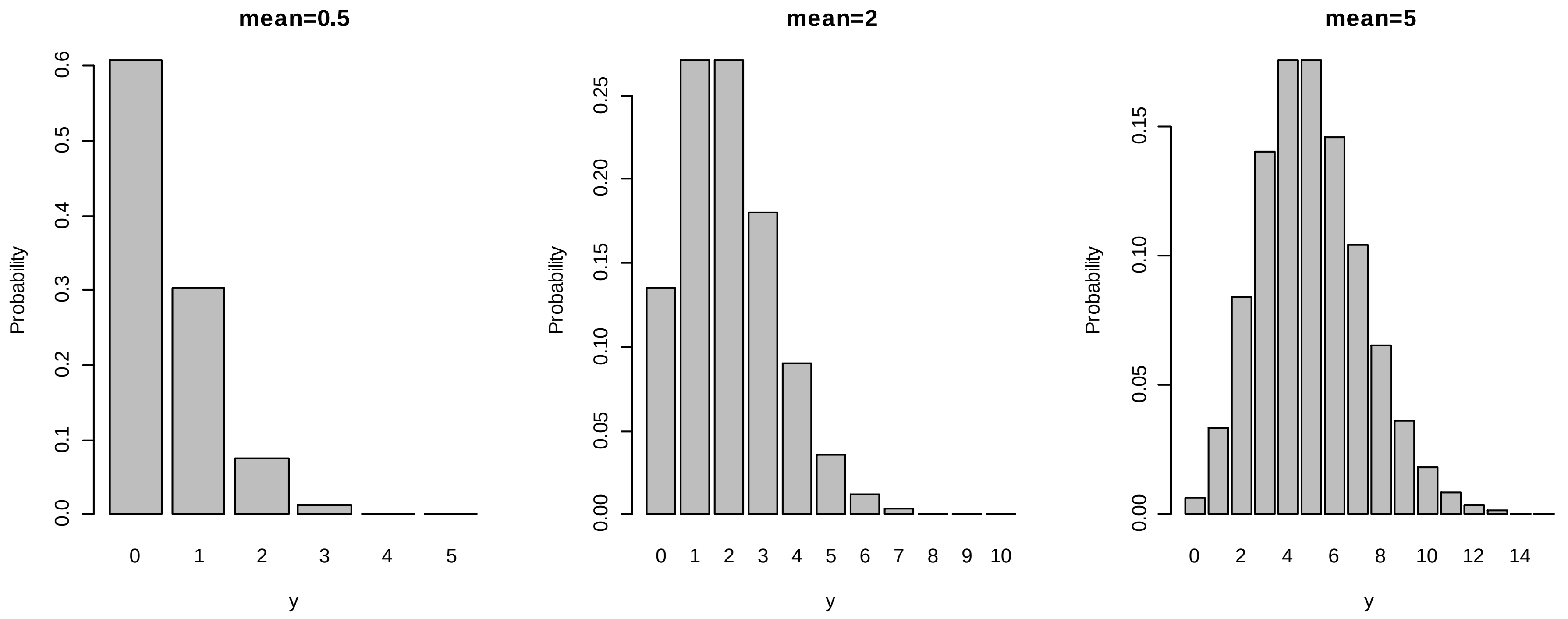
The impact of the mean parameter

```
par(mfrow=c(1,3))

barplot(dpois(0:5,0.5),xlab="y",ylab="Probability",names=0:5,main="mean=0.5")

barplot(dpois(0:10,2),xlab="y",ylab="Probability",names=0:10,main="mean=2")

barplot(dpois(0:15,5),xlab="y",ylab="Probability",names=0:15,main="mean=5")
```


mean=0.5
Probability
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
1
2
3
4
5
y
mean=2
Probability
0.00
0.05
0.10
0.15
0.20
0.25
0
1
2
3
4
5
6
7
8
9
10
y
mean=5
Probability
0.00
0.05
0.10
0.15
0
2
4
6
8
10
12
14
y

How can we model a count response using a number of covariates?

Model the number of species on the Galapagos Islands


Galapagos Islands
90°W
Isla Pinta
Isla Marchena
Isla Genovesa
EQUATOR
0°
Isla Santiago
Isla Baltra
Isla Fernandina
Pinzon
Isla Isabela
Isla Santa Cruz
Puerto Ayora
Isla Santa Fe
Isla San Cristobal
Puerto Baquerizo Moreno
Puerto Villamil
Pacific Ocean
Isla Santa Maria
Isla Espanola
50 mi
50 km
GALAPAGOS
LOW / HILLS / MOUNTAINS
NORTH AMERICA
Atlantic Ocean
Caribbean Sea
Pacific Ocean
0°
EQUATOR
Galapagos Islands
SOUTH AMERICA
23°5'S
Pacific Ocean
Atlantic Ocean
90°W

Species diversity on the Galapagos Islands

There are 30 Galapagos islands and 7 variables in the dataset. The relationship between the number of plant species and several geographic variables is of interest.

```
library("faraway")
data(gala)
str(gala)
?gala
gala<-gala[,-2] ### Remove the second variable
summary(gala)
```

`Species`

the number of plant species found on the island

`Endemics`

the number of endemic species

`Area`

the area of the island (km^2)

`Elevation`

the highest elevation of the island (m)

`Nearest`

the distance from the nearest island (km)

`Scruz`

the distance from Santa Cruz island (km)

`Adjacent`

the area of the adjacent island (square km)

Source

M. P. Johnson and P. H. Raven (1973) "Species number and endemism: The Galapagos Archipelago revisited" Science, 179, 893-895

Galapagos Islands
90°W
Isla Pinta
Isla Marchena
Isla Genovesa
EQUATOR
0°
Isla Santiago
Isla Baltra
Isla Fernandina
Pinzon
Isla Isabela
Isla Santa Cruz
Puerto Ayora
Isla Santa Fe
Isla San Cristobal
Puerto Baquerizo Moreno
Puerto Villamil
Pacific Ocean
Isla Santa Maria
Isla Espanola
50 mi
50 km
GALAPAGOS
LOW / HILLS / MOUNTAINS
NORTH AMERICA
Atlantic Ocean
Caribbean Sea
Pacific Ocean
0°
EQUATOR
Galapagos Islands
SOUTH AMERICA
23°5'S
Pacific Ocean
Atlantic Ocean
90°W

Summary statistics

```
> summary(gala)
    Species            Area            Elevation          Nearest
 Min.   :  2.00   Min.   :   0.010   Min.   :  25.00   Min.   : 0.20
 1st Qu.: 13.00   1st Qu.:   0.258   1st Qu.:  97.75   1st Qu.: 0.80
 Median : 42.00   Median :   2.590   Median : 192.00   Median : 3.05
 Mean   : 85.23   Mean   : 261.709   Mean   : 368.03   Mean   :10.06
 3rd Qu.: 96.00   3rd Qu.:  59.237   3rd Qu.: 435.25   3rd Qu.:10.03
 Max.   :444.00   Max.   :4669.320   Max.   :1707.00   Max.   :47.40
     Scruz           Adjacent
 Min.   :  0.00   Min.   :   0.03
 1st Qu.: 11.03   1st Qu.:   0.52
 Median : 46.65   Median :   2.59
 Mean   : 56.98   Mean   : 261.10
 3rd Qu.: 81.08   3rd Qu.:  59.24
 Max.   :290.20   Max.   :4669.32
```

The idea of modeling a binary response

- Suppose the binary response

$$Y \sim \text{Bernoulli}(p)$$

- We will make two additional assumptions on top of this assumption:
  1. Y relies on X only through its mean, e.g $Y \sim \text{Bernoulli}(p(X))$
  2. The logit transformation of the parameter p

$$\text{logit}(p) = \log(\tfrac{p}{1-p}) = \beta_1 + \beta_2 X$$

How to extend the idea to modeling a count response?

- Suppose the count response

$$Y \sim \text{Poisson}(\mu)$$

- We will make two additional assumptions on top of this assumption:
  1. Y relies on X only through its mean, e.g $Y \sim \text{Poisson}(\mu(X))$
  2. The log transformation of the parameter μ

$$\log(\mu) = \beta_1 + \beta_2 X$$

Fit a GLM

```
modp<-glm(Species ~ ., family=poisson, gala)
summary(modp)
step(modp)
drop1(modp,test="Chisq")
```

```
> summary(modp)

Call:
glm(formula = Species ~ ., family = poisson, data = gala)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.2752  -4.4966  -0.9443   1.9168  10.1849

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.155e+00  5.175e-02  60.963  < 2e-16 ***
Area        -5.799e-04  2.627e-05 -22.074  < 2e-16 ***
Elevation    3.541e-03  8.741e-05  40.507  < 2e-16 ***
Nearest      8.826e-03  1.821e-03   4.846 1.26e-06 ***
Scruz       -5.709e-03  6.256e-04  -9.126  < 2e-16 ***
Adjacent    -6.630e-04  2.933e-05 -22.608  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
AIC: 889.68
```

```
> step(modp)
Start:  AIC=889.68
Species ~ Area + Elevation + Nearest + Scruz +
Adjacent

            Df Deviance     AIC
<none>           716.85  889.68
- Nearest    1   739.41  910.24
- Scruz      1   813.62  984.45
- Area       1  1204.35 1375.18
- Adjacent   1  1341.45 1512.29
- Elevation  1  2389.57 2560.40
```

```
> drop1(modp,test="LRT")
Single term deletions

Model:
Species ~ Area + Elevation + Nearest + Scruz + Adjacent
          Df Deviance     AIC     LRT  Pr(>Chi)
<none>          716.85  889.68
Area       1   1204.35 1375.18  487.51 < 2.2e-16 ***
Elevation  1   2389.57 2560.40 1672.72 < 2.2e-16 ***
Nearest    1    739.41  910.24   22.57 2.031e-06 ***
Scruz      1    813.62  984.45   96.77 < 2.2e-16 ***
Adjacent   1   1341.45 1512.29  624.61 < 2.2e-16 ***
```

Check the correlation

```
> round(cor(gala),2)
           Species  Area Elevation Nearest Scruz Adjacent
Species       1.00  0.62      0.74   -0.01 -0.17     0.03
Area          0.62  1.00      0.75   -0.11 -0.10     0.18
Elevation     0.74  0.75      1.00   -0.01 -0.02     0.54
Nearest      -0.01 -0.11     -0.01    1.00  0.62    -0.12
Scruz        -0.17 -0.10     -0.02    0.62  1.00     0.05
Adjacent      0.03  0.18      0.54   -0.12  0.05     1.00
```

Prediction

```
> modp$y
     Baltra   Bartolome    Caldwell    Champion     Coamano Daphne.Major Daphne.Minor       Darwin
         58          31           3          25           2           18           24           10
       Eden     Enderby    Espanola  Fernandina    Gardner1     Gardner2     Genovesa      Isabela
          8           2          97          93          58            5           40          347
   Marchena      Onslow       Pinta      Pinzon  Las.Plazas       Rabida SanCristobal  SanSalvador
         51           2         104         108          12           70          280          237
  SantaCruz     SantaFe  SantaMaria     Seymour     Tortuga         Wolf
        444          62         285          44          16           21


> round(predict(modp,type="response"),1)
     Baltra   Bartolome    Caldwell    Champion     Coamano Daphne.Major Daphne.Minor       Darwin
       78.7        20.4        25.7        21.4        17.0         36.6         32.1         10.9
       Eden     Enderby    Espanola  Fernandina    Gardner1     Gardner2     Genovesa      Isabela
       29.8        26.8        27.9        87.7        15.9         38.2         25.0        370.8
   Marchena      Onslow       Pinta      Pinzon  Las.Plazas       Rabida SanCristobal  SanSalvador
       55.8        20.3       212.6       121.4        32.2         53.1        218.8        371.0
  SantaCruz     SantaFe  SantaMaria     Seymour     Tortuga         Wolf
      297.3        60.9       158.2        36.9        35.5         18.1
```

Goodness of fit measure

- Pearson's X^2 statistic

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

```
gof<-sum(residuals(modp,type="pearson")^2)
pchisq(gof,df.residual(modp),lower=F)
```

```
> gof  ### Peason’s goodness of fit statistic
[1] 761.9792
> df.residual(modp)    ### Degrees of freedom
[1] 24
> pchisq(gof,df.residual(modp),lower=F) ### P-value
[1] 2.18719e-145
```

The model does not fit the data well. Why?

The mean and variance parameters are not separatable

$$Y \sim \text{Poisson}(\mu)$$

- The mean is $E(Y) = \mu$
- The variance is $Var(Y) = \mu$

- Once the mean is specified by the Poisson regression, the variance is determined at the same time! This is different from the case of linear regression models.

What would you do?

$$Y \sim \text{Poisson}(\mu)$$

- Suppose the mean $E(Y) = \mu$ is correctly captured by the Poisson regression model.
- But the data suggest that the variance is consistently **greater** than

$$Var(Y) = \mu$$

Dispersion parameter

- Suppose the count response

$$Y \sim \text{Poisson}(\mu)$$

- We will make three additional assumptions on top of this assumption:
  1. Y relies on X only through its mean, e.g $Y \sim \text{Poisson}(\mu(X))$
  2. The log transformation of the parameter μ

$$\log(\mu) = \beta_1 + \beta_2 X$$

  3. The variance $Var(Y) = \phi\mu$, where the **dispersion parameter** ϕ allows one more layer of flexibility of the model.

Estimating the dispersion parameter

$$\hat{\sigma}^2 = \frac{X^2}{n-q}$$

- Here, X^2 is the usual Pearson goodness-of-fit statistic, n is the number of sample cases (number of rows in the dataset we are modeling), and q is the number of parameters.

- Please try:
- `dp<-gof/modp$df.res`

Updating the model

```
> dp
[1] 31.74914


> summary(modp,dispersion=dp)

Call:
glm(formula = Species ~ ., family = poisson, data = gala)


Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)  3.1548079  0.2915897  10.819  < 2e-16 ***
Area        -0.0005799  0.0001480  -3.918 8.95e-05 ***
Elevation    0.0035406  0.0004925   7.189 6.53e-13 ***
Nearest      0.0088256  0.0102621   0.860    0.390    
Scruz       -0.0057094  0.0035251  -1.620    0.105    
Adjacent    -0.0006630  0.0001653  -4.012 6.01e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 31.74914)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
AIC: 889.68
```

Further reading

- Zero inflated count models
- Rate models