Lecture 20

Poisson Regression II

STAT 8020 Statistical Methods II
October 29, 2020

Whitney Huang
Clemson University

Notes

Species Diversity on the Galapagos Islands Revisited

Recall we are interested in studying the relationship between the **number** of plant species (`Species`) and the following geographic variables: `Area, Elevation, Nearest, Scruz, Adjacent.`

Notes

Data: Species Diversity on the Galapagos Islands

	Species	Area	Elevation	Nearest	Scruz	Adjacent
Baltra	58	25.09	346	0.6	0.6	1.84
Bartolome	31	1.24	109	0.6	26.3	572.33
Caldwell	3	0.21	114	2.8	58.7	0.78
Champion	25	0.10	46	1.9	47.4	0.18
Coamano	2	0.05	77	1.9	1.9	903.82
Daphne.Major	18	0.34	119	8.0	8.0	1.84
Daphne.Minor	24	0.08	93	6.0	12.0	0.34
Darwin	10	2.33	168	34.1	290.2	2.85
Eden	8	0.03	71	0.4	0.4	17.95
Enderby	2	0.18	112	2.6	50.2	0.10
Espanola	97	58.27	198	1.1	88.3	0.57
Fernandina	93	634.49	1494	4.3	95.3	4669.32
Gardner1	58	0.57	49	1.1	93.1	58.27
Gardner2	5	0.78	227	4.6	62.2	0.21
Genovesa	40	17.35	76	47.4	92.2	129.49
Isabela	347	4669.32	1707	0.7	28.1	634.49
Marchena	51	129.49	343	29.1	85.9	59.56
Onslow	2	0.01	25	3.3	45.9	0.10
Pinta	104	59.56	777	29.1	119.6	129.49
Pinzon	108	17.95	458	10.7	10.7	0.03
Las.Plazas	12	0.23	94	0.5	0.6	25.09
Rabida	70	4.89	367	4.4	24.4	572.33
SanCristobal	280	551.62	716	45.2	66.6	0.57
SanSalvador	237	572.33	906	0.2	19.8	4.89
SantaCruz	444	903.82	864	0.6	0.0	0.52
SantaFe	62	24.08	259	16.5	16.5	0.52
SantaMaria	285	170.92	640	2.6	49.2	0.10
Seymour	44	1.84	147	0.6	9.6	25.09
Tortuga	16	1.24	186	6.8	50.9	17.95
Wolf	21	2.85	253	34.1	254.7	2.33

Notes

Poisson Regression Fit

```
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

Number of Fisher Scoring iterations: 5
```


Notes

Wafer Quality and Possible Sampling Schemes

The data shown in the table below were collected as part of a quality improvement study at a semiconductor factory. A sample of wafers was drawn and cross-classified according to whether a particle was found on the die that produced the wafer and whether the wafer was good or bad.

Quality	No Particles	Particles	Total
Good	320	14	334
Bad	80	36	116
Total	400	50	450

Source: Hall, S. (1994). *Analysis of defectivity of semiconductor wafers by contigency*

How the data were collected?

Notes

Possible Sampling Schemes

- We observed the manufacturing process for a certain period of time and observed 450 wafers ⇒ Poisson Model
- We decided to sample 450 wafers. The data were then cross-classified ⇒ Multinomial Model
- We selected 400 wafers without particles and 50 wafers with particles and then recorded the good or bad outcome ⇒ Binomial Model
- We selected 400 wafers without particles and 50 wafers with particles that also included, by design, 334 good wafers and 116 bad ones ⇒ Hypergeometric Model

Notes

Poisson Model: Log-linear Regression

$$Y_{ij} \sim \text{Poi}(\lambda_{ij}), \quad \log \lambda_{ij} = \gamma + \alpha_i + \beta_j, \quad i, j = 1, 2.$$

```
> mod1 <- glm(Freq ~ Quality + Particle, family = "poisson")
> sumary(mod1)
            Estimate Std. Error  z value  Pr(>|z|)
(Intercept)  5.69336    0.05720  99.5350 < 2.2e-16
QualityBad  -1.05755    0.10777  -9.8129 < 2.2e-16
ParticleYes -2.07944    0.15000 -13.8630 < 2.2e-16

n = 4 p = 3
Deviance = 54.03045 Null Deviance = 474.09877 (Difference = 420.06832)
> drop1(mod1, test = "Chi")
Single term deletions

Model:
Freq ~ Quality + Particle
         Df Deviance    AIC    LRT  Pr(>Chi)
<none>          54.03  83.77
Quality   1   164.22 191.96 110.19 < 2.2e-16 ***
Particle  1   363.91 391.66 309.88 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Notes

Multinomial Model

$$Y_{ij} \sim \text{Multi}(n, p_{11}, p_{12}, p_{21}, p_{22})$$

Want to test $H_0 : p_{ij} = p_i p_j$ vs. $H_a : p_{ij} \neq p_i p_j, \quad i, j = 1, 2.$

```
> n = 450
> (pp <- prop.table(xtabs(Freq ~ Particle)))
Particle
       No       Yes
0.8888889 0.1111111
> (qp <- prop.table(xtabs(Freq ~ Quality)))
Quality
     Good       Bad
0.7422222 0.2577778
> (exp <- outer(qp, pp) * n)
       Particle
Quality       No      Yes
   Good 296.8889 37.11111
   Bad  103.1111 12.88889
> (obs <- xtabs(Freq ~ Quality + Particle))
       Particle
Quality  No Yes
   Good 320  14
   Bad   80  36
> (2 * sum(obs * log(obs / exp)))
[1] 54.03045
```


Notes

Binomial Model

$$Y_{11} \sim \text{Bin}(n_1 = 400, p_{11})$$
$$Y_{21} \sim \text{Bin}(n_2 = 50, p_{21})$$

Want to test $H_0 : p_{11} = p_{21}$ vs. $H_a : p_{11} \neq p_{21}$

```
> (m <- matrix(Freq, nrow = 2))
     [,1] [,2]
[1,]  320   80
[2,]   14   36
> (binFit <- glm(m ~ 1, family = binomial))

Call:  glm(formula = m ~ 1, family = binomial)

Coefficients:
(Intercept)  
      1.058  

Degrees of Freedom: 1 Total (i.e. Null);  1 Residual
Null Deviance:      54.03 
Residual Deviance: 54.03        AIC: 66.19
> predict(binFit, type = "response")
        1         2 
0.7422222 0.7422222 
```


Notes

Hypergeometric Model: Fisher's Exact Test

$$Y_{11} \sim \text{Hyper}(N = 450, 400, 334)$$

```
> fisher.test(obs)

        Fisher's Exact Test for Count Data

data:  obs
p-value = 2.955e-13
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  5.090628 21.544071
sample estimates:
odds ratio
  10.21331
```


Notes

Generalized Linear Model

- **Gaussian Linear Model:**

$$Y \sim \text{N}(\mu, \sigma^2), \quad \mu = \boldsymbol{X}^T \boldsymbol{\beta}$$

- **Bernoulli Linear Model:**

$$Y \sim \text{Bernoulli}(\pi), \quad \log(\frac{\pi}{1-\pi}) = \boldsymbol{X}^T \boldsymbol{\beta}$$

- **Poisson Linear Regression:**

$$Y \sim \text{Poisson}(\lambda), \quad \log \lambda = \boldsymbol{X}^T \boldsymbol{\beta}$$

Notes

Notes