**Section 4.3 – Poisson rate regression**

Rate data consists of the rate that a number of events occur for some time period or for some other baseline measure. Examples include:

* The number of times a computer crashes during unequal time periods,
* The number of melanoma cases per city size, and
* The number of arrivals at airports over unequal time periods.

The time period or baseline measure needs to be incorporated into the analysis. One way to do this is to model Y/t instead of just Y, where Y is the number of events and t is the time period or baseline measure.

Suppose there is only one explanatory variable x. The Poisson regression model becomes:

log(μ/t) = β0 + β1x

where μ = E(Y). This expression can be simplified to

log(μ) – log(t) = β0 + β1x ⇒ log(μ) = β0 + β1x + log(t).

The log(t) term is an offset. Notice the effect that an offset has on μ:



Thus, t helps to adjust the “usual” mean () by the time period or baseline measure.

One can emphasize the model for each observation with

log(μi) = β0 + β1xi + log(ti)

for i = 1, …, n.

Estimation, inference, and model interpretation proceed in a similar manner as before.

Example: Horseshoe crabs and satellites (Horseshoe.R, Horseshoe.csv)

This is not necessarily the best example where one would want to use Poisson regression for rate data; however, it gives a nice illustration of the relationship between a Poisson model for rate data and for the data format seen previously. Please see the book for another example where a rate data format is needed.

Suppose the data was given by the number of satellites per distinct width. Let Y be the number of satellites for a distinct width. Let t be the number of female crabs observed for a distinct width. For example, there are t = 3 crabs with a width of 22.9 cm, and they have a total of Y = 4 + 0 + 0 = 4 satellites.

The data set originally had this form:

|  |  |  |
| --- | --- | --- |
| Observation number | Satellites | Width |
| 1 | 8 | 28.3 |
| 2 | 0 | 22.5 |
| 3 | 9 | 26.0 |
| 4 | 0 | 24.8 |
|  |  |  |

We want to convert the data set to this form:

| # of crabs (t) | Total satellites (Y) | Width |
| --- | --- | --- |
| 1 | 0 | 21.0 |
| 1 | 0 | 22.0 |
| 3 | 5 | 22.5 |
| 3 | 4 | 22.9 |
|  |  |  |

Below is the code to make the conversion:

> # Number of satellites per unique width

> total.sat <- aggregate(x = satellite ~ width, data = crab, FUN = sum)

> # Number of crabs per unique width

> numb.crab <- aggregate(x = satellite ~ width, data = crab, FUN = length)

> rate.data <- data.frame(total.sat, numb.crab = numb.crab$satellite)

> head(rate.data)

width satellite numb.crab

1 21.0 0 1

2 22.0 0 1

3 22.5 5 3

4 22.9 4 3

5 23.0 1 2

6 23.1 0 3

We will use the glm() function to estimate the model:



The offset() function is used in the formula argument of glm() to let R know NOT to estimate a parameter for log(numb.crab):

> mod.fit.rate <- glm(formula = satellite ~ width + offset(log(numb.crab)), data = rate.data, family = poisson(link = log))

> summary(mod.fit.rate)

Call:

glm(formula = satellite ~ width + offset(log(numb.crab)), family = poisson(link = log), data = rate.data)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.8003 -1.4515 -0.3788 0.6619 4.7586

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.30476 0.54224 -6.095 1.1e-09 \*\*\*

width 0.16405 0.01997 8.216 < 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: 254.94 on 65 degrees of freedom

Residual deviance: 190.03 on 64 degrees of freedom

AIC: 402.52

Number of Fisher Scoring iterations: 5

The estimated model is



Notice the regression parameter estimates are the same as before!

Estimated values of μ are found in a similar manner as before, where now the value of t needs to be specified:

> predict(object = mod.fit.rate, newdata = data.frame(width

= c(23, 23), numb.crab = c(1, 2)), type = "response")

1 2

1.597244 3.194488

where .

Below is a plot of the model:

> #Find the number of unique values of t and put into a

vector

> plot.char.numb <-

as.numeric(names(table(rate.data$numb.crab)))

> plot(x = rate.data$width, y = rate.data$satellite, xlab =

"Width (cm)", ylab = "Number of satellites", type =

"n", panel.first = grid(), main = "Horseshoe crab data

set \n with Poisson regression model fit (rate data)")

> #Put observed values and estimated model on plot by

values of t

> for(t in plot.char.numb) {

width.t <- rate.data$width[rate.data$numb.crab ==

plot.char.numb[t]]

satellite.t <- rate.data$satellite[rate.data$numb.crab

== plot.char.numb[t]]

points(x = width.t, y = satellite.t, pch =

as.character(plot.char.numb[t]), cex = 0.5, col = t)

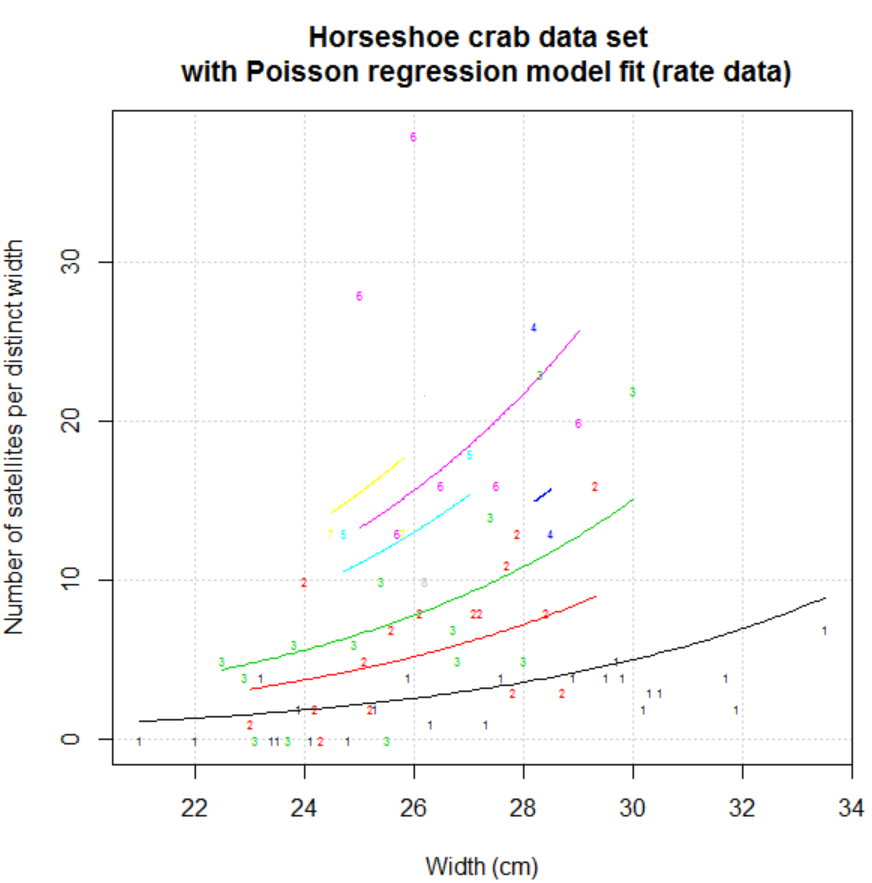
curve(expr = t \* exp(mod.fit.rate$coefficients[1] +

mod.fit.rate$coefficients[2]\*x), xlim =

c(min(width.t), max(width.t)), lty = "solid", col =

t, add = TRUE)

}



It would be nice to include group means for width groups too, but there may not be enough observations per distinct width for this addition to be beneficial.

One difference between a Poisson regression model for rate data and a Poisson regression model for “regular” data is the residual deviances most likely will be different:

|  |  |  |
| --- | --- | --- |
| **Data** | **-2log(Λ)** | **DF** |
| Regular | 567.88 | 171 |
| Rate | 190.03 | 64 |

However, the -2log(Λ) statistic will be the same when testing the regression parameters:

> Anova(mod.fit)

Analysis of Deviance Table (Type II tests)

Response: satellite

LR Chisq Df Pr(>Chisq)

width 64.913 1 7.828e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> Anova(mod.fit.rate)

Analysis of Deviance Table (Type II tests)

Response: satellite

LR Chisq Df Pr(>Chisq)

width 64.913 1 7.828e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Where this difference between residual deviances can play a role is when assessing the “goodness of fit” for a model, which is discussed in a later chapter.

Note that this same type of difference between residual deviances occurs for Bernoulli vs. Binomial response data in logistic regression.