**Chapter 4 – Analyzing a count response**

Chapter 1 and 2 examined a binomial response (a count) arising from a FIXED number of trials. Thus, W, a binomial random variable, could be 0, 1, …, n, where n is the number of trials.

Count responses can also arise from other mechanisms that have nothing to do with Bernoulli trials. Examples include:

* The number of credit cards an individual owns
* The number of arrests for a city per year
* The number of people arriving at an airport on a given day
* The number of cars stopped at the 33rd and Holdrege streets intersection
* The number of people standing in line at Starbucks.

For these settings, a Poisson distribution can be used to model the count responses.

You may also think of particular explanatory variables that could affect these count responses. For example, the number of credit cards an individual owns could be dependent on income level, gender, where they live, … . We will examine how to do this in Section 4.2.

You may also want to control for other items that are not necessarily explanatory variables. If the purpose of the number of arrests example was to make comparisons across cities in the United States, we would want to account for city sizes. For example, it would not make sense to compare Omaha directly to New York City without taking into account their vastly different sizes. We will examine how to do this in Section 4.3.

Next, we begin with the basics of how we model counts through using a Poisson probability distribution.

**Section 4.1 – Poisson model for count data**

Poisson PMF:



for y = 0, 1, 2, …, where

Y is a random variable

y denotes the possible outcomes of Y

μ is a parameter that is greater than 0

I will sometimes write Y ~ Po(μ) as shorthand notation to mean that Y has a Poisson distribution with parameter μ.

Y will denote the number of occurrences of an event.

Properties

Below are characteristics of a Poisson distribution and associated items of interest:

* One can show the mean and variance of Y to be:

E(Y) = μ and Var(Y) = μ

Having the mean and variance BOTH equal to μ is nice, but this is also limiting. Often, the actual variability observed in a sample is GREATER than μ. This is referred to as overdispersion. We will discuss how to handle this in Chapter 5.

* If Y1, …, Yn are independent with distribution Po(μ), then . If each Yk had a different mean, we would have 
* Likelihood function:



* MLE for μ is , i.e., the sample mean
* The estimated variance for  is



Wald confidence interval for μ

What is the interval?

Can this interval have a lower bound less than 0? If so, how can we fix it?

Hypothesis test for μ

H0: μ = μ0 vs. Ha: μ ≠ μ0

Wald test statistic: 

Score test statistic: 

How would you decide on rejection of the null hypothesis?

Score confidence interval for μ



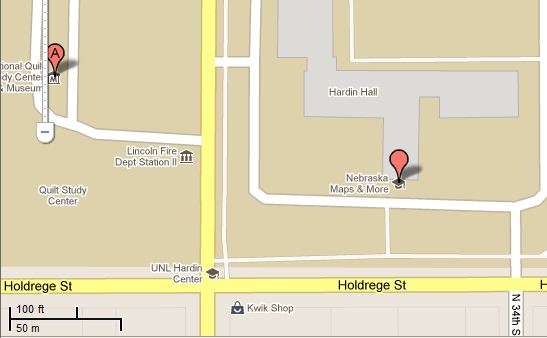
How was this interval derived?

Which interval do you think is better: Wald or Score? Please see an example in Section 4.1 of the book for true confidence level plots.

How else could you find a confidence interval for μ?

Example: 33rd and Holdrege (Stoplight.R, StoplightAdditional.R, Stoplight.csv)

The intersection at 33rd and Holdrege Streets is a typical north-south/east-west, 4-way intersection.



Below is a picture taken from a location northwest of the intersection:



Approximately 150 feet north of the intersection is a fire station located on the west side of the street. A back-up of vehicles at the stoplight waiting to go south could block the fire station's driveway, which would prevent emergency vehicles from exiting the station.

To examine this more closely, I took a sample of 40 consecutive stoplight cycles from 3:25PM to 4:05PM on a non-holiday weekday, and the number of vehicles stopped at the stoplight going south were counted. Below is part of the data:

> stoplight <- read.csv(file = "C:\\data\\stoplight.csv")

> head(stoplight)

Observation vehicles

1 1 4

2 2 6

3 3 1

4 4 2

5 5 3

6 6 3

Note that there were no vehicles remaining in the intersection for more than one stoplight cycle. Why is this important to know?

Is a Poisson distribution appropriate for this data?

> #Summary statistics

> mean(stoplight$vehicles)

[1] 3.875

> var(stoplight$vehicles)

[1] 4.317308

> #Frequencies

> table(stoplight$vehicles)

0 1 2 3 4 5 6 7 8

1 5 7 3 8 7 5 2 2

> rel.freq<-table(stoplight$vehicles) /

length(stoplight$vehicles)

> rel.freq2<-c(rel.freq, rep(0, times = 7))

> #Poisson calculations

> y<-0:15

> prob<-round(dpois(x=y, lambda=mean(stoplight$vehicles)),

4)

> data.frame(y, prob, rel.freq = rel.freq2)

y prob rel.freq

1 0 0.0208 0.025

2 1 0.0804 0.125

3 2 0.1558 0.175

4 3 0.2013 0.075

5 4 0.1950 0.200

6 5 0.1511 0.175

7 6 0.0976 0.125

8 7 0.0540 0.050

9 8 0.0262 0.050

10 9 0.0113 0.000

11 10 0.0044 0.000

12 11 0.0015 0.000

13 12 0.0005 0.000

14 13 0.0001 0.000

15 14 0.0000 0.000

16 15 0.0000 0.000

> plot(x = y-0.1, y = prob, type = "h", ylab =

"Probability", xlab = "Number of vehicles", lwd = 2,

xaxt = "n")

> axis(side = 1, at = 0:15)

> lines(x = y+0.1, y = rel.freq2, type = "h", lwd = 2, lty

= "solid", col = "red")

> abline(h = 0)

> legend(x = 9, y = 0.15, legend = c("Poisson",

"Observed"), lty = c("solid", "solid"), lwd = c(2,2),

col = c(“black”, “red”), bty = “n”)



Confidence interval calculations:

> alpha<-0.05

> n<-length(stoplight$vehicles)

> mu.hat<-mean(stoplight$vehicles)

> #Wald

> mu.hat + qnorm(p = c(alpha/2, 1-alpha/2))\*sqrt(mu.hat/n)

[1] 3.264966 4.485034

> #Wald using a log() transformation – not in book’s

program

> exp(log(mu.hat) + qnorm(p = c(alpha/2, 1-alpha/2)) \*

sqrt(1/(mu.hat\*n)))

[1] 3.310561 4.535674

> #Score

> (mu.hat + qnorm(p = c(alpha/2, 1-alpha/2))/(2\*n)) +

qnorm(p = c(alpha/2, 1-alpha/2)) \* sqrt((mu.hat +

qnorm(p = 1-alpha/2)/(4\*n))/n)

[1] 3.239503 4.510497

> #Usual t-distribution based interval

> t.test(x = stoplight$vehicles, conf.level = 0.95)

One Sample t-test

data: stoplight$vehicles

t = 11.7949, df = 39, p-value = 1.955e-14

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

3.210483 4.539517

sample estimates:

mean of x

3.875

Note that the Wald interval using the log(μ) transformation is



One can show through a delta-method approximation that .

Suppose all vehicles are 14 feet long with a distance

between cars of 4 feet when stopped at the intersection. This suggests that 9 vehicles (150/18 = 8.3) or more waiting at the intersection could at least partially block the fire station's driveway. Using the estimated Poisson distribution, estimate the probability this will happen.

> prob9<-1 - ppois(q = 8, lambda = mu.hat)

> prob9

[1] 0.01786982

This is the probability for one stoplight cycle. What about this even occur at least once over the 60 stoplight cycles that would occur over one hour?

Let W be a binomial random variable with n = 60 and π = 0.01786982. Then P(W ≥ 1) = 1 – P(W = 0) = 0.66.

> #Not in book’s program

> 1 - dbinom(x = 0, size = 60, prob = prob9)

[1] 0.661044

Questions:

* What do you think will happen to this probability during rush-hour?
* How could we take into account different time periods of the day?
* This is an estimated probability. How could we find a confidence interval for it? The interval is (0.3305, 0.9916). See my additional program for details.

Other intervals for μ exist. Please see the book for the likelihood ratio and exact confidence intervals.

**Section 4.2 – Poisson regression models for count responses**

**Section 4.2.1 – Model for mean: Log link**

Suppose the mean parameter of a Poisson distribution is now dependent on a function of explanatory variables. For example, suppose there is only one explanatory variable x. We could represent this dependence by

μ = β0 + β1x

Depending on the value of the parameters and x, we could obtain a negative value for μ which would not make sense for a count! Instead, we can use

log(μ) = β0 + β1x

which alternatively can be written as

μ = exp(β0 + β1x)

Now, μ is guaranteed to be greater than 0. This is referred to a Poisson regression model.

When needed, we can emphasize that the mean changes as a function of the variable x for the ith observation with

μi = exp(β0 + β1xi)

If there are p explanatory variables, we can write the model as

μ = exp(β0 + β1x1 + βpxp)

or

log(μ) = β0 + β1x1 + βpxp

Generalized linear model

A Poisson regression model is a generalized linear model with the following components:

1. Random: Y has a Poisson distribution
2. Systematic: β0 + β1x1 + βpxp
3. Link: log

**Section 4.2.2 – Parameter estimation and inference**

Maximum likelihood estimation is used again to find the MLEs. Suppose my sample is denoted as (yi, xi1, …, xip) with i = 1, …, n. The likelihood function is

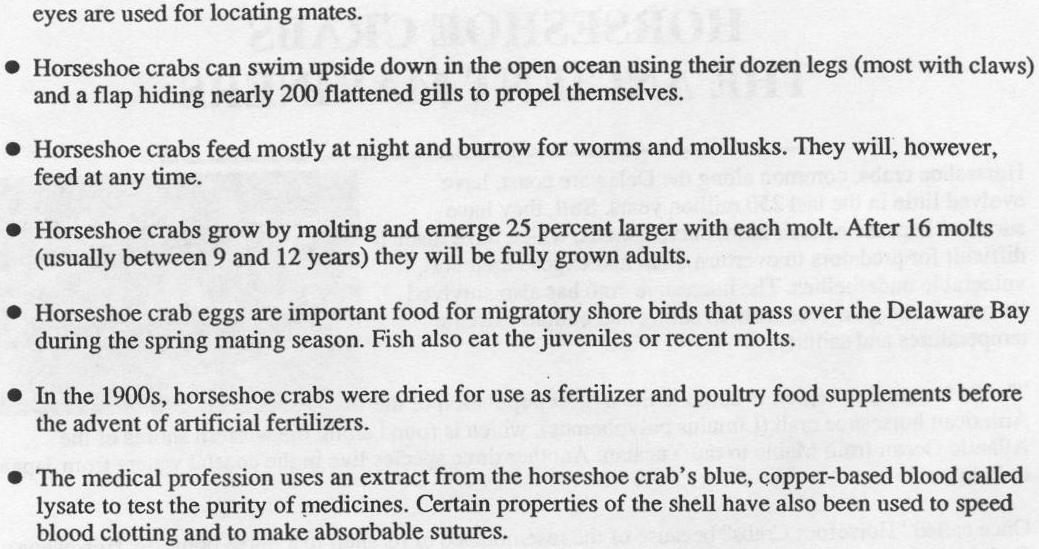
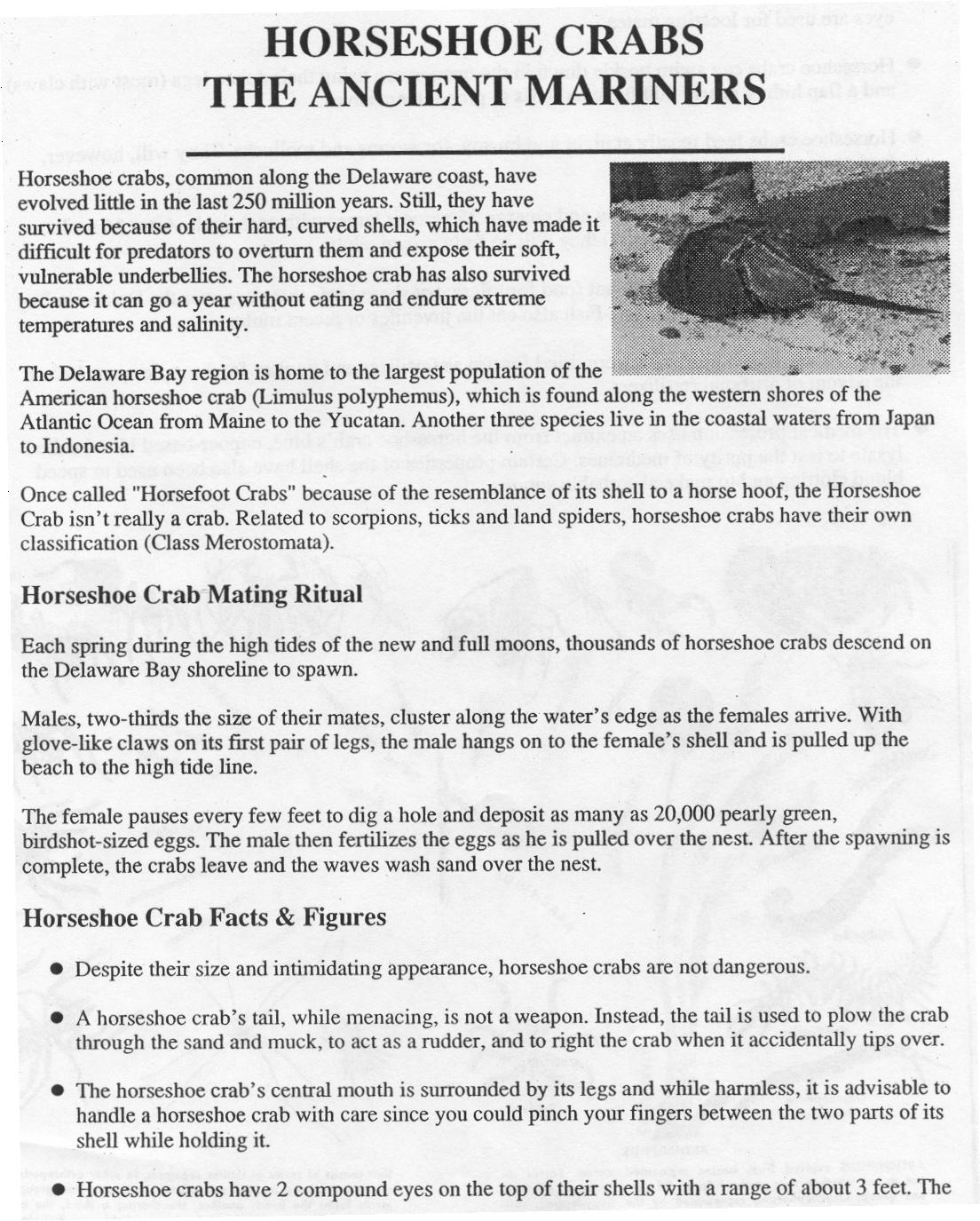


where . For most situations, the likelihood function needs to be maximized using iterative numerical procedures. The glm() function in R completes this maximization where the family argument needs to be given as poisson(link = log).

The covariance matrix for the parameter estimators follows from using standard likelihood procedures as outlined in Appendix B. Wald and LR-based inference methods are performed in the same ways as for likelihood procedures in earlier chapters.

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

See the video for more on horseshow crabs. Below are some additional details:



An NPR story on horseshoe crabs is at

[www.npr.org/templates/story/story.php?storyId=106489695](http://www.npr.org/templates/story/story.php?storyId=106489695)

The purpose of this example is to determine if the shell width of a female (x) is related to the number of satellites (Y) she has around her. A simple Poisson regression model:

log(μ) = β0 + β1x

where

Y = Number of satellites

x = Shell width of female (measured in cm)

can be used to estimate the mean number of satellites given a shell width.

Below is how I read in the data:

> crab<-read.table(file = "c:\\data\\horseshoe.txt", header

= TRUE)

> head(crab)

satellite width

1 8 28.3

2 0 22.5

3 9 26.0

4 0 24.8

5 4 26.0

6 0 23.8

Below is how I estimate the model:

> mod.fit<-glm(formula = satellite ~ width, data = crab,

family = poisson(link = log))

> summary(mod.fit)

Call:

glm(formula = satellite ~ width, family = poisson(link = log), data = crab)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.8526 -1.9884 -0.4933 1.0970 4.9221

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: 632.79 on 172 degrees of freedom

Residual deviance: 567.88 on 171 degrees of freedom

AIC: 927.18

Number of Fisher Scoring iterations: 6

The estimated Poisson regression model is



The model could also be written as:



Questions:

* What happens to the estimated mean number of satellites as the width increases?
* What type of female crabs do male crabs prefer?

Now that we have the estimated model, many of the basic types of analyses performed in Chapters 2 and 3 can be performed here. The R code used is very similar as well. Because of the similarity, I would like you to tell me how to do the following:

1. Perform a Wald test for a β1.
2. How can we perform a LRT for an explanatory variable?
3. How can we estimate the expected number of satellites when the shell width is 23?



1. How can we find a Wald confidence interval for μ? Are there any worries about interval limits being outside of the appropriate numerical range?
2. How can we find a Profile LR confidence interval for μ?

> K<-matrix(data = c(1, 23), nrow = 1, ncol = 2)

> K

[,1] [,2]

[1,] 1 23

> #Calculate -2log(Lambda)

> linear.combo<-mcprofile(object = mod.fit, CM = K)

> #CI for beta\_0 + beta\_1 \* x

> ci.logmu.profile<-confint(object = linear.combo, level

= 0.95)

> ci.logmu.profile

mcprofile - Confidence Intervals

level: 0.95

adjustment: single-step

Estimate lower upper

C1 0.468 0.284 0.647

> ci.logmu.profile$confint

lower upper

C1 0.2841545 0.6471545

> exp(ci.logmu.profile)

mcprofile - Confidence Intervals

level: 0.95

adjustment: single-step

Estimate lower upper

C1 1.6 1.33 1.91

The 95% interval is 1.33 < μ < 1.91, which is quite similar to the Wald interval.

1. How could you estimate the covariance matrix and print it in R?
2. How could you include some type of transformation of an explanatory variable(s) in the model?

When there is only one explanatory variable in the model, we can easily examine the estimated model through a plot:

> plot(x = crab$width, y = crab$satellite, xlab = "Width

(cm)", ylab = "Number of satellites", main = "Horseshoe

crab data set \n with Poisson regression model fit",

panel.first =grid())

> curve(expr = exp(mod.fit$coefficients[1] +

mod.fit$coefficients[2]\*x), col = "red", add = TRUE,

lty = "solid")

> #Could also use this to plot the model:

> #curve(expr = predict(object = mod.fit, newdata =

data.frame(width = x), type ="response"), col = "red",

add = TRUE, lty = 1)

> #Function to find confidence interval

> ci.mu<-function(newdata, mod.fit.obj, alpha) {

lin.pred.hat<-predict(object = mod.fit.obj, newdata =

newdata, type = "link", se = TRUE)

lower<-exp(lin.pred.hat$fit - qnorm(1 - alpha/2) \*

lin.pred.hat$se)

upper<-exp(lin.pred.hat$fit + qnorm(1 - alpha/2) \*

lin.pred.hat$se)

list(lower = lower, upper = upper)

}

> #Test

> ci.mu(newdata = data.frame(width = 23), mod.fit.obj =

mod.fit, alpha = 0.05)

$lower

1

1.332135

$upper

1

1.915114

> #Add confidence interval bands

> curve(expr = ci.mu(newdata = data.frame(width = x),

mod.fit.obj = mod.fit, alpha = 0.05)$lower, col =

"blue", add = TRUE, lty = "dotdash")

> curve(expr = ci.mu(newdata = data.frame(width = x),

mod.fit.obj = mod.fit, alpha = 0.05)$upper, col =

"blue", add = TRUE, lty = "dotdash")

> legend(x = 21, y = 14, legend = c("Poisson regression

model", "95% individual C.I."), bty = "n", lty =

c("solid", "dotdash"), col = c("red", "blue"))



The data shows somewhat of an upward trend. The model captures this through displaying similar qualities.

You may be alarmed by the number of plotting points far from the estimated model. However, remember that the model is trying to estimate the “average” number of satellites given the width. The plot below examines this more closely where I have added the average number of satellites for a “width group”.

> #One way to put the data into groups

> groups<-ifelse(test = crab$width<23.25, yes = 1, no =

ifelse(test = crab$width<24.25, yes = 2, no =

ifelse(test = crab$width<25.25, yes = 3, no =

ifelse(test = crab$width<26.25, yes = 4, no =

ifelse(test = crab$width<27.25, yes = 5, no =

ifelse(test = crab$width<28.25, yes = 6, no =

ifelse(test = crab$width<29.25, yes = 7, no =

8)))))))

> crab.group<-data.frame(crab,groups)

> head(crab.group)

satellite width groups

1 8 28.3 7

2 0 22.5 1

3 9 26.0 4

4 0 24.8 3

5 4 26.0 4

6 0 23.8 2

> #Find the average number of satellites per group and plot

> ybar<-aggregate(formula = satellite ~ groups, data =

crab, FUN = mean)

> xbar<-aggregate(formula = width ~ groups, data = crab,

FUN = mean)

> data.frame(ybar, xbar$width)

groups satellite xbar.width

1 1 1.000000 22.69286

2 2 1.428571 23.84286

3 3 2.392857 24.77500

4 4 2.692308 25.83846

5 5 2.863636 26.79091

6 6 3.875000 27.73750

7 7 3.944444 28.66667

8 8 5.142857 30.40714

> points(x = xbar$width, y = ybar$satellite, pch = 17, col

= "darkgreen", cex = 2)

> #New legend (re-run all of the plot code without the

first legend() call

> legend(x = 21, y = 14, legend = c("Poisson regression

model", "95% individual C.I.", "Sample mean"), bty =

"n", lty = c("solid", "dotdash", NA), col = c("red",

"blue", "darkgreen"), pch = c(NA, NA, 17))



Notice how the red line goes through the middle of the green triangles (group means).

Comments:

* Does the model fit the data well? This is a difficult question to answer solely based on this plot. At the very least, it looks like the model is doing what it is supposed to do.
* What if we used different groupings? My program gives details on how to use a more general method (based on quantiles of the data) to break the data up into groups. Below is the corresponding plot:



This plot does not look as good as the previous plot. There is more variability of the green triangles around the red line. Overall, this helps to illustrate that different groupings can produce somewhat different results, and this method should not be used alone to judge if a model is fitting well.

* You could construct a similar type of plot to provide an ad-hoc assessment of how well a logistic regression model fits the data. This is especially useful in situations when the explanatory variable is continuous.

Model interpretation

We are no longer modeling the log-odds anymore so we will not use odds ratios to interpret the effect of an explanatory variable on the response ☹

Consider a model with one explanatory variable again:

μ(x) = exp(β0 + β1x)

where I am using “μ(x)” here to emphasize we are evaluating the model as particular numerical value of x. The model evaluated at a c-unit increase in the explanatory variable is

μ(x+c) = exp(β0 + β1(x + c))

If we examine the ratio of the two cases, we obtain:



This leads to a convenient way to interpret the effect of x:

The percentage change in the mean response that results from a c-unit change in x is PC = 

For example, suppose . This leads to PC = 10%.

Comments:

* This interpretation is not dependent on the original value of x!
* Choose a value of c appropriate for the data.
* The estimate of  is .
* Wald and LR confidence intervals can be found using the usual methods
* If there is more than one explanatory variable in the model, the same result holds (make sure you can show it). You also need to add “holding the other variables in the model constant” to the interpretation.
* If there are interactions or transformations of explanatory variables or categorical explanatory variables, similar types of adjustments need to be made as in Chapter 2 for odds ratios. Make sure you are ready for questions about this on a test!

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

A c = 1 cm increase in width results in:

> exp(mod.fit$coefficients[2])

width

1.178267

> 100\*(exp(mod.fit$coefficients[2]) - 1)

width

17.82674

A 1 unit change in width leads to an estimated 17.83% change in the mean number of satellites.

To help emphasize the type of change, we could phrase our interpretation as:

* A 1 unit INCREASE in width leads to an estimated 17.83% INCREASE in the mean number of satellites.
* The mean number of satellites is estimated to increase by 17.83% for every 1 cm increase in the width of the shell.

Is c = 1 appropriate? This is difficult to answer without knowing more about the crabs. When there is not an obvious choice for c, you can resort to using one standard deviation:

> c.unit<-sd(crab$width)

> c.unit

[1] 2.109061

> 100\*exp(c.unit\*mod.fit$coefficients[2]) - 1

width

41.33759

A 2.11 cm increase in width leads to an estimated 41.34% increase in the mean number of satellites.

Profile LR confidence interval with c = 1:

> #Profile likelihood interval

> beta.ci<-confint(object = mod.fit, parm = "width", level =

0.95)

Waiting for profiling to be done...

> 100\*(exp(beta.ci) - 1)

2.5 % 97.5 %

13.28362 22.50566

> #Profile LR interval using mcprofile

> library(package = mcprofile)

> K<-matrix(data = c(0, 1), nrow = 1, ncol = 2)

> #Calculate -2log(Lambda)

> linear.combo<-mcprofile(object = mod.fit, CM = K)

> ci.beta<-confint(object = linear.combo, level = 0.95)

> 100\*(exp(ci.beta$confint) - 1)

lower upper

C1 13.28365 22.50525

With 95% confidence, a 1 unit INCREASE in width leads to a 13.3% to 22.5% INCREASE in the mean number of satellites.

Wald confidence interval with c = 1:

> beta.ci<-confint.default(object = mod.fit, parm = "width",

level = 0.95)

> beta.ci

2.5 % 97.5 %

width 0.1249137 0.2031764

> exp(beta.ci)

2.5 % 97.5 %

width 1.133051 1.225289

> 100\*(exp(beta.ci) - 1)

2.5 % 97.5 %

width 13.30507 22.52887

> #Calculation without confint.default

> vcov(mod.fit) #Var^(beta^\_1) is in the (2,2) element of

the matrix

(Intercept) width

(Intercept) 0.29402590 -0.0107895239

width -0.01078952 0.0003986151

> beta.ci<-mod.fit$coefficients[2] + qnorm(p = c(0.025,

0.975))\*sqrt(vcov(mod.fit)[2,2])

> 100\*(exp(beta.ci) - 1)

[1] 13.30507 22.52887

**Section 4.2.3 – Categorical explanatory variables**

You can handle these variables in a similar manner as shown in Chapter 2! Suppose an explanatory variable has levels of A, B, C, and D. Three indicator variables can be used to represent the explanatory variable in a model:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Indicator variables | | |
| Levels | x1 | x2 | x3 |
| A | 0 | 0 | 0 |
| B | 1 | 0 | 0 |
| C | 0 | 1 | 0 |
| D | 0 | 0 | 1 |

A Poisson regression model with these indicator variables is



Focusing on β1, we have PC =  representing the percentage change in the mean response for level B when compared to level A. Similar comparisons can be made for C to A and D to A. To compare level B to C, we have PC = .

**Section 4.2.4 – Poisson regression for contingency tables: loglinear models**

Poisson regression models are frequently used to model counts in a contingency table. For this setting, each cell corresponds to an independent Poisson random variable. The explanatory variables in the model are simply categorical explanatory variables representing the rows and columns!

Example: Larry Bird (bird\_ch.R)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Second | |  |
|  |  | Made | Missed | Total |
| First | Made | 251 | 34 | 285 |
| Missed | 48 | 5 | 53 |
|  | Total | 299 | 39 | 338 |

The data can be converted to a data frame:

> c.table<-array(data = c(251, 48, 34, 5), dim = c(2,2),

dimnames = list(First = c("made", "missed"),

Second = c("made", "missed")))

> bird<-as.data.frame(as.table(c.table))

> bird

First Second Freq

1 made made 251

2 missed made 48

3 made missed 34

4 missed missed 5

Using bird$Freq as the Poisson response variable, we can form the following model:

log(μ) = β0 + β1FirstMissed + β2SecondMissed

where

FirstMissed = 1 when the first free throw is missed and = 0 for made

SecondMissed = 1 when the second free throw is missed and = 0 for made

Notice that this model means there is “independence” between the row and column variable because they do not “interact” in how they affect the response.

Below is how we can estimate this model under independence:

> mod.fit1<-glm(formula = Freq ~ First + Second, data =

bird, family = poisson(link = log))

> summary(mod.fit1)

Call:

glm(formula = Freq ~ First + Second, family = poisson(link = log), data = bird)

Deviance Residuals:

1 2 3 4

-0.0703 0.1623 0.1934 -0.4659

Coefficients:

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

(Intercept) 5.52989 0.06241 88.61 <2e-16 \*\*\*

Firstmissed -1.68220 0.14959 -11.25 <2e-16 \*\*\*

Secondmissed -2.03688 0.17025 -11.96 <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: 402.05553 on 3 degrees of freedom

Residual deviance: 0.28575 on 1 degrees of freedom

AIC: 28.212

Number of Fisher Scoring iterations: 3

The estimated model is

log() = 5.53 - 1.68FirstMissed - 2.04SecondMissed

The estimated mean counts are

> predict(object = mod.fit1, newdata = bird, type =

"response")

1 2 3 4

252.115385 46.884615 32.884615 6.115385

> Pearson.test<-chisq.test(x = c.table, correct = FALSE)

> Pearson.test$expected

Second

First made missed

made 252.11538 32.884615

missed 46.88462 6.115385

Notice these counts are the same as the expected cell counts we saw in Chapter 3!

Below is how we can estimate the model that includes the interaction:

> mod.fit2<-glm(formula = Freq ~ First + Second +

First:Second, data = bird, family = poisson(link =

log))

> summary(mod.fit2)

Call:

glm(formula = Freq ~ First + Second + First:Second, family = poisson(link = log), data = bird)

Deviance Residuals:

[1] 0 0 0 0

Coefficients:

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

(Intercept) 5.52545 0.06312 87.540 <2e-16 \*\*\*

Firstmissed -1.65425 0.15754 -10.501 <2e-16 \*\*\*

Secondmissed -1.99909 0.18275 -10.939 <2e-16 \*\*\*

Firstmissed:Secondmissed -0.26267 0.50421 -0.521 0.602

---

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

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 4.0206e+02 on 3 degrees of freedom

Residual deviance: 4.3077e-14 on 0 degrees of freedom

AIC: 29.926

Number of Fisher Scoring iterations: 3

The estimated model is

log() = 5.52 - 1.65FirstMissed – 2.00SecondMissed

– 0.2627FirstMissed×SecondMissed

Notice the residual deviance is 0! This is because we are estimating a saturated model! There are four parameters and four Poisson observations.

The estimated mean counts are

> predict(object = mod.fit2, newdata = bird, type =

"response")

1 2 3 4

251 48 34 5

> bird$Freq

[1] 251 48 34 5

The estimated values are the same as the observed!

Below is the information needed for a LRT involving the interaction (H0:β3 = 0 vs. Ha:β3 ≠ 0):

> Anova(mod.fit2, test = "LR")

Analysis of Deviance Table (Type II tests)

Response: Freq

LR Chisq Df Pr(>Chisq)

First 174.958 1 <2e-16 \*\*\*

Second 226.812 1 <2e-16 \*\*\*

First:Second 0.286 1 0.593

---

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

> anova(mod.fit1, mod.fit2, test = "Chisq")

Analysis of Deviance Table

Model 1: Freq ~ First + Second

Model 2: Freq ~ First + Second + First:Second

Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 1 0.28575

2 0 0.00000 1 0.28575 0.593

> library(package = vcd)

Loading required package: colorspace

> assocstats(c.table)

X^2 df P(> X^2)

Likelihood Ratio 0.28575 1 0.59296

Pearson 0.27274 1 0.60150

Phi-Coefficient : 0.028

Contingency Coeff.: 0.028

Cramer's V : 0.028

The hypothesis test for the interaction term results in   
-2log(Λ) = 0.286 with a p-value of 0.593. Notice that this is exactly the same as we would find using Section 3.1’s methods!

Loglinear models

It is very common in categorical data analysis books to differentiate between Poisson regression models when

* Only categorical explanatory variables are present (Larry Bird data) and
* At least one continuous explanatory variable is present (Horseshoe crab data)

For example, Agresti (2007) would give the following models for the Larry Bird data:

log(μij) = λ +  +  for i = 1, 2 and j = 1, 2

under independence and

log(μij) = λ +  +  +  for i = 1, 2 and j = 1, 2

under dependence, where

* μij is the mean response for row i and column j
* , , and  are parameters, and they have constraints such as , , and  for (i,j) = (1,1), (1,2), and (2,1) so that we do not have more parameters than Poisson observations (rows of our data set); in other classes, you may hear these constraints referred to as “identifiability” restrictions**.**

The models for the Larry Bird data are often referred to as loglinear models. A “Poison regression” model would generally refer to a model where at least one explanatory variable is continuous.

This differentiating between models with and without categorical explanatory variables is unnecessary. For example, the estimated model for the Larry Bird data under independence was

log() = 5.53 - 1.68FirstMissed - 2.04SecondMissed

Using Agresti’s notation, this model would be written instead as



where  and . Essentially, the loglinear model is an ANOVA model representation of a regression model. Remember that ANOVA models are just a special case of regression models!

**Section 4.2.5 – Large loglinear models**

Contingency tables can have additional “dimensions” to them. For example, a three-dimensional (three-way) contingency table would have a form:

Z=1 Z = 2

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Y | |  |  |  |  | Y | |  |
|  |  | 1 | 2 |  |  |  |  | 1 | 2 |  |
| X | 1 | n111 | n121 | n1+1 |  | X | 1 | n112 | n122 | n1+2 |
| 2 | n211 | n221 | n2+1 |  | 2 | n212 | n222 | n2+2 |
|  |  | n+11 | n+21 | n++1 |  |  |  | n+12 | n+22 | n++2 |

where Z represents the “strata” or “layer” dimension. Similar to the last section, these dimensions are represented as explanatory variables in a Poisson regression model. Main effects, two-way interactions, and three-way interactions can be included.

Poisson and multinomial discussion

In contingency table settings, one can perform very similar analyses using a multinomial regression model (or a logistic regression model when there are only two response categories) and a Poisson regression model. This occurs for the following reason:

The joint probability distribution of IJ independent Poisson random variables CONDITIONAL on their sum is a multinomial distribution.

To see this result, consider a 2×2 contingency table using the notation of Section 3.1:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Y | |  |
|  |  | 1 | 2 |  |
| X | 1 | n11 | n12 | n1+ |
| 2 | n21 | n22 | n2+ |
|  |  | n+1 | n+2 | n++ |

Please note that I am using a little different notation here to help make the connection to Chapter 3. Suppose nij ~ independent Po(μij) for i = 1, 2 and j = 1, 2. Then the joint probability distribution is



Let n++ = n11 + n12 + n21 + n22 (note: n++ = n in our more commonly used notation). Because each random variable is independent, we have n++ ~ . Then









This is a multinomial distribution for n++ trials and probabilities .

The bird\_ch4.R program contains some additional code to help demonstrate the equivalence between the multinomial and Poisson approaches.

Is there an advantage to using the Poisson model over the multinomial model? There can be one – when you do not want to specify having ONLY one response variable. Below are two examples to explain this:

Example: One could think of the Larry Bird data as having TWO response variables: First free throw outcome and second free throw outcome.

Example: In the fiber data set from Chapter 3, the fiber source was an “input” and we were interested in the bloating “output” that resulted from it. Thus, bloating was the only response variable.

Interestingly, you will still obtain the same analysis results even if you used a Poisson regression model when a multinomial regression model may seem more appropriate and vice versa. This is because of the equivalence that occurs between the Poisson and multinomial distributions.

Odds ratios

Odds ratios can be helpful to interpret the fit of a model in contingency table settings. Simply, find the predicted counts for each cell and use odds ratios to interpret relationships between variables. For example, the estimated odds ratio for the Larry Bird data example was 1 when estimating the model without the interaction.

Ordinal categorical variables

One can take advantage of ordinal properties of a categorical explanatory variable in a Poisson regression setting. Please see Section 4.2.6 of my book for a discussion.

**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).

log(t) 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.txt)

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 “regular” data as seen in Section 4.2. Please see my 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:

> total.sat<-aggregate(formula = satellite ~ width, data =

crab, FUN = sum) #Number of satellites per unique width

> numb.crab<-aggregate(formula = satellite ~ width, data =

crab, FUN = length) #Number of crabs per unique width

> 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

The glm() function is used to estimate the model where we now use the offset() function 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 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)

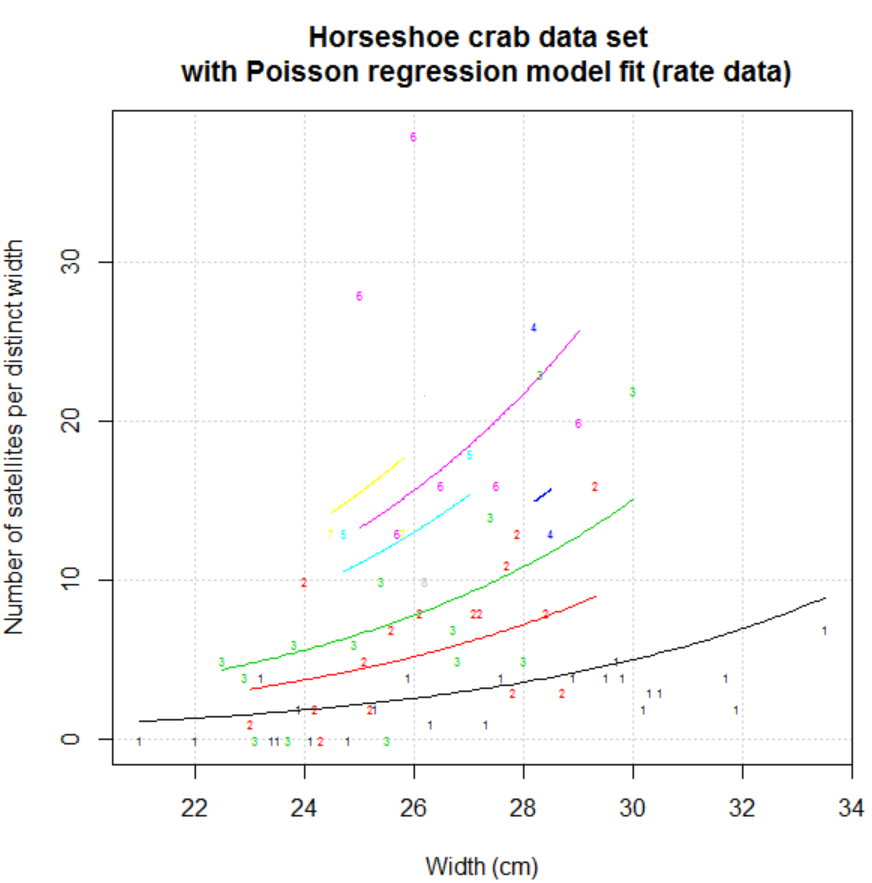
curve(expr = t \* exp(mod.fit.rate$coefficients[1] +

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 significance of 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 Chapter 5.

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

**Section 4.4 – Zero inflation**

There are times when more 0 counts occur than expected for a Poisson distribution. In this setting, we can account for the “zero inflation” by using a “mixture” distribution for Y rather than a Poisson distribution alone:

Y ~ Po(μ) with probability 1 – π

Y = 0 with probability π

We can allow both μ and π to be a function of explanatory variables.

Please see my book for more information.