R Homework #5

C.J.Anderson

October 15, 2018

Abstract

I have prepared a more in depth description of the code used for this homework set. It was written using the LaTex + R. It required MikTex installed on you computer and uses the packages `knitr` and `rmarkdown`. The R code is run when I compile this. You can also create HTML using this capability. This is one thing that I use RStudio for. I you are interested, I can send you the .Rnw file.

In this problem you were asked to test whether the arrest rate as soccer games is constant for different teams. We will fit the model and then graph the data.

1 Problem 1

First thing to do is read in the data.

```
fights <- read.table("a2007_318_data.txt",header=TRUE)
```

A quick look at top part of the data frame

```
head(fights)
```

```
##            team attend arrests
## 1   Aston_Villa    404     308
## 2 Bradford_City    286     197
## 3  Leeds_United    443     184
## 4   Bournemouth    169     149
## 5     West_Brom    222     132
## 6   Hudderfield    150     126
```

There are (at least) two possible methods to fit a model that corresponds to the desired test: Is the rate of arrest constant at soccer games in different cities in the UK?

1.1 Method I

One way to solve this problems is as a poisson regression model with with no intercept and an identity; namely,

$$E(\text{arrests}) = \beta(\text{attend}).$$

This model is fit to the data in R by the following:

```
method1 <- glm(arrests ~  -1 + attend, data = fights,
                           family = poisson(link="identity"))
```

The "−1" in the equation indicates that no intercept should be included. The results for this model are

```
summary(method1)
```

```
##
## Call:
## glm(formula = arrests ~ -1 + attend, family = poisson(link = "identity"),
##     data = fights)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -12.789   -3.426   -0.938    3.079   10.137
##
## Coefficients:
##        Estimate Std. Error z value Pr(>|z|)
## attend 0.402411   0.008707   46.22   <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:    Inf  on 23  degrees of freedom
## Residual deviance: 669.45  on 22  degrees of freedom
## AIC: 812.62
##
## Number of Fisher Scoring iterations: 3
```

Since the number of arrests are in the thousands, the arrest rate is given by

```
method1$coefficents[1]/1000

## numeric(0)
```

and the effect on the rate of arrests is

```
exp(.0004)

## [1] 1.0004
```

In other words, the rate of arrest for one addition fan attending a game is 1.0004001 times the rate without the addition.

1.2 Method 2

An alternative way to solve this problem is to forumla the model as a Poisson model for rates with no predictor variables and only an intercept; that is,

$$\log(E(\text{arrests/attend})] = \beta_0.$$

In this case, the $\log(arrests)$ is an offset. In R this model is specified by

```
summary(method2 <- glm(arrests ~ offset(log(attend)), data=fights,
                       family=poisson) )

##
## Call:
## glm(formula = arrests ~ offset(log(attend)), family = poisson,
##     data = fights)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -12.789   -3.426   -0.938    3.079   10.137
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.91028    0.02164  -42.07   <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: 669.45  on 22  degrees of freedom
## Residual deviance: 669.45  on 22  degrees of freedom
## AIC: 812.62
##
## Number of Fisher Scoring iterations: 5
```

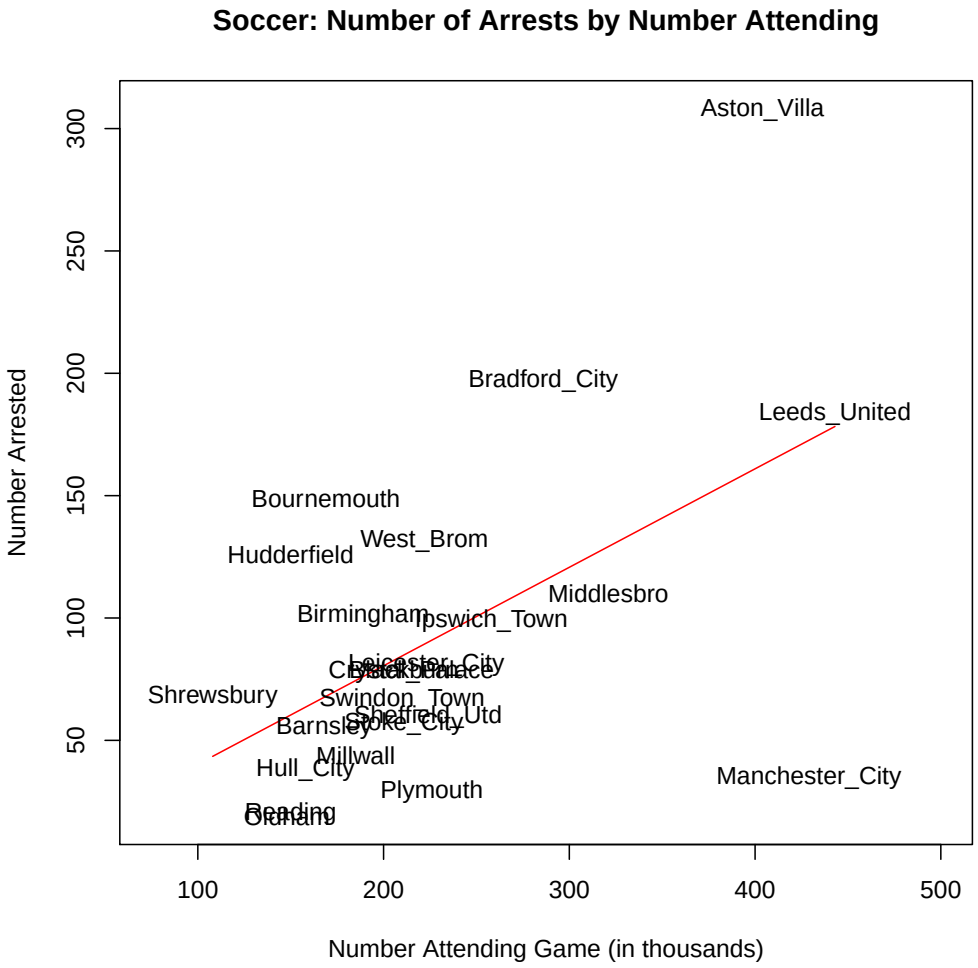
To put the coefficient into the same scale as data we first take the exponential of the coefficient 0.4024115 and then divide by 1000, which gives up the same value as method I, 4.0241145×10^{-4}.

1.3 Graphing Data and Results

In this plot we could just graph points; however, in the example, using the city name in the graph would be useful. We start with a basic graph for these data

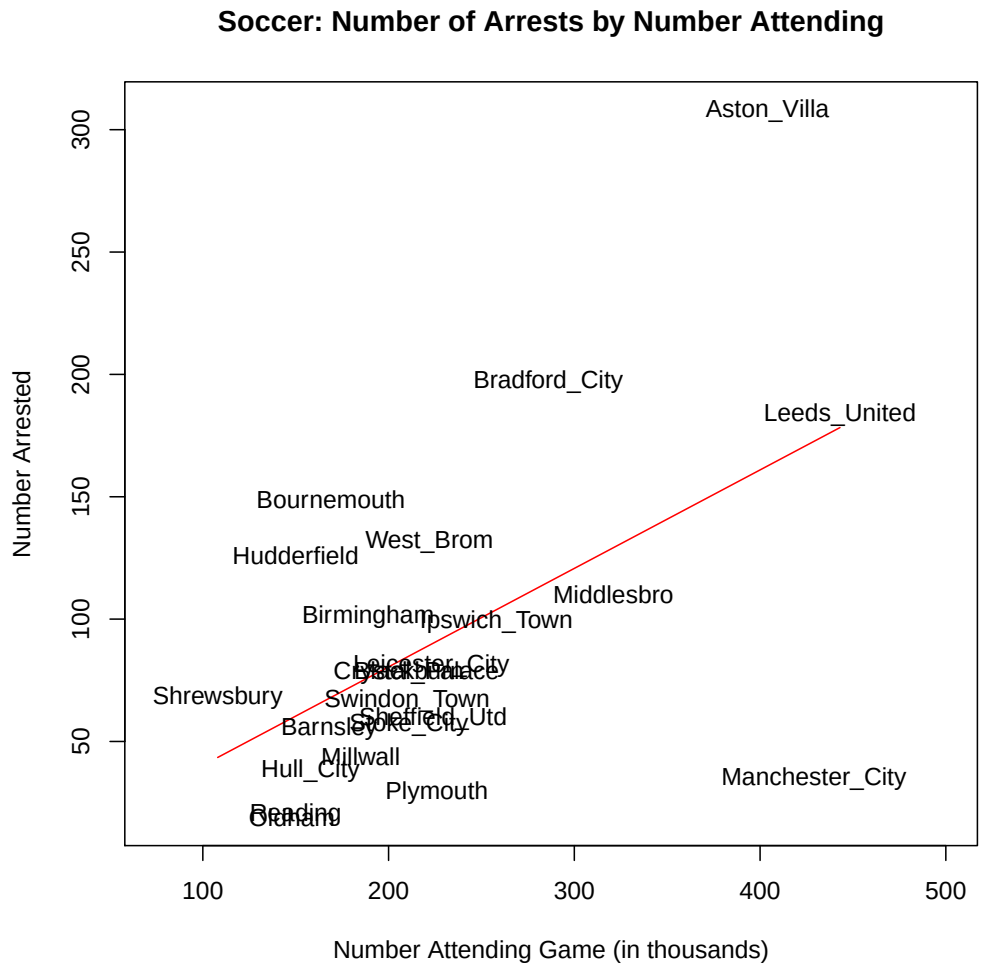
```
# Graph
plot(fights$attend,fights$arrests,
  type="p",
  col="white",
        xlim=c(75,500),
        xlab="Number Attending Game (in thousands)",
        ylab="Number Arrested",
  main="Soccer: Number of Arrests by Number Attending"
```

```
  )
j <- order(fights$attend)
lines(fights$attend[j],method2$fitted[j],type="l",col="red")
text(fights$attend,fights$arrests,fights$team)
```



In this plot we could just graph points; however, in the example, using the city name in the graph would be useful. We start with a basic graph for these data

```r
# Graph
plot(fights$attend,fights$arrests,
  type="p",
  col="white",
        xlim=c(75,500),
        xlab="Number Attending Game (in thousands)",
        ylab="Number Arrested",
  main="Soccer: Number of Arrests by Number Attending"
  )
j <- order(fights$attend)
lines(fights$attend[j],method2$fitted[j],type="l",col="red")
text(fights$attend,fights$arrests,fights$team)
```



Notice that I indicated type='p' but gave made them white, which means they are not visible. This does create the set up for our graph. The next step was to get the order of the number who attended and use this as index for both attendence and the fitted values. The lines() command added the regression lines and the text() command added text at coordianate (fights$attend[j], method2$fitted[j]) and the text to put at these cooridnates is the team name, fights$team. Unfortunately the figure does not show up right after the code—this is a LaTex thing and needs some finesing to get figures where you want them.

Negative Binomial There is yet another model that might be reasonable

to these data. You can use either method 1 or 2, but use a negative bionomial distribution. You need to use the `MASS` package for this model.

```
library(MASS)
```

Although this can be done with either methods 1 or 2, I used method 2 with an offset. My model is specified in R by

```
summary(nb <- glm.nb(arrests ~ offset(log(attend)),data=fights))

##
## Call:
## glm.nb(formula = arrests ~ offset(log(attend)), data = fights,
##     init.theta = 3.135631071, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2049  -0.7464  -0.1857   0.6129   1.5568
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.9052     0.1200  -7.546 4.49e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(3.1356) family taken to be 1)
##
##     Null deviance: 24.15  on 22  degrees of freedom
## Residual deviance: 24.15  on 22  degrees of freedom
## AIC: 244.24
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  3.136
##           Std. Err.:  0.920
##
##  2 x log-likelihood:  -240.236
```

The dispersion parameter in the output is ϕ (SAS give $1/\phi$).

This document's purpose is mainly to show you the R commands and what they produce. **You still have to interprete and explain the results as given in the answer key.**

2 Problem 2 (3.13 from Agresti (2007))

As always we need some setting up,

```
library(MASS)
crabs<- read.table("crab_data.txt",header=TRUE)
```

It is good pratice to get basic descriptive statistics and look at the data before we start to analysis it. To get a few statistics, which what the are should be self-explanatory.

```
mean(crabs$satell)
```

```
## [1] 2.919075
```

```
var(crabs$satell)
```

```
## [1] 9.912018
```

```
length(crabs$satell)
```

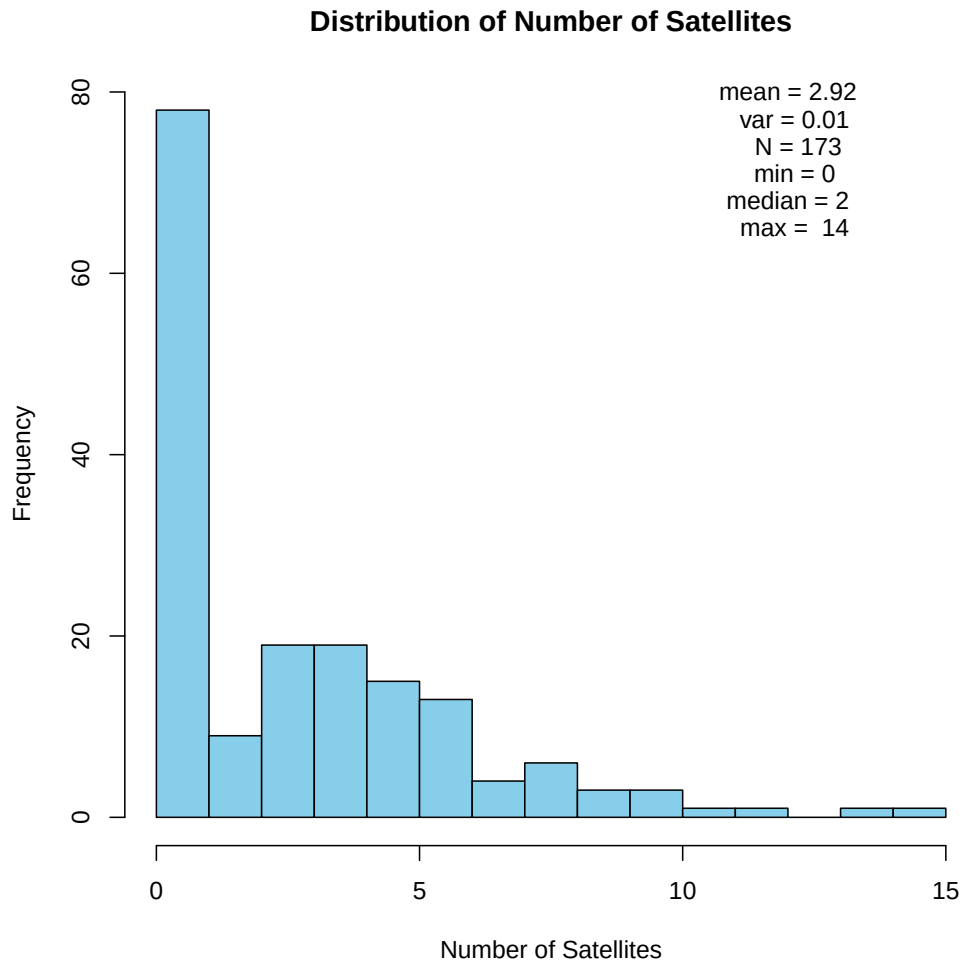
```
## [1] 173
```

```
summary(crabs$satell)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   2.000   2.919   5.000  15.000
```

Now for a look at the distribution of the data, including the sample statistics in a histrogram.

```r
hist(crabs$satell,
breaks=16,
xlab="Number of Satellites",
main="Distribution of Number of Satellites",col="skyblue"
)
text(12,80,"mean = 2.92")
text(12,77,"  var = 0.01")
text(12,74,"   N = 173")
text(12,71,"  min = 0")
text(12,68,"median = 2")
text(12,65,"  max =  14")
```

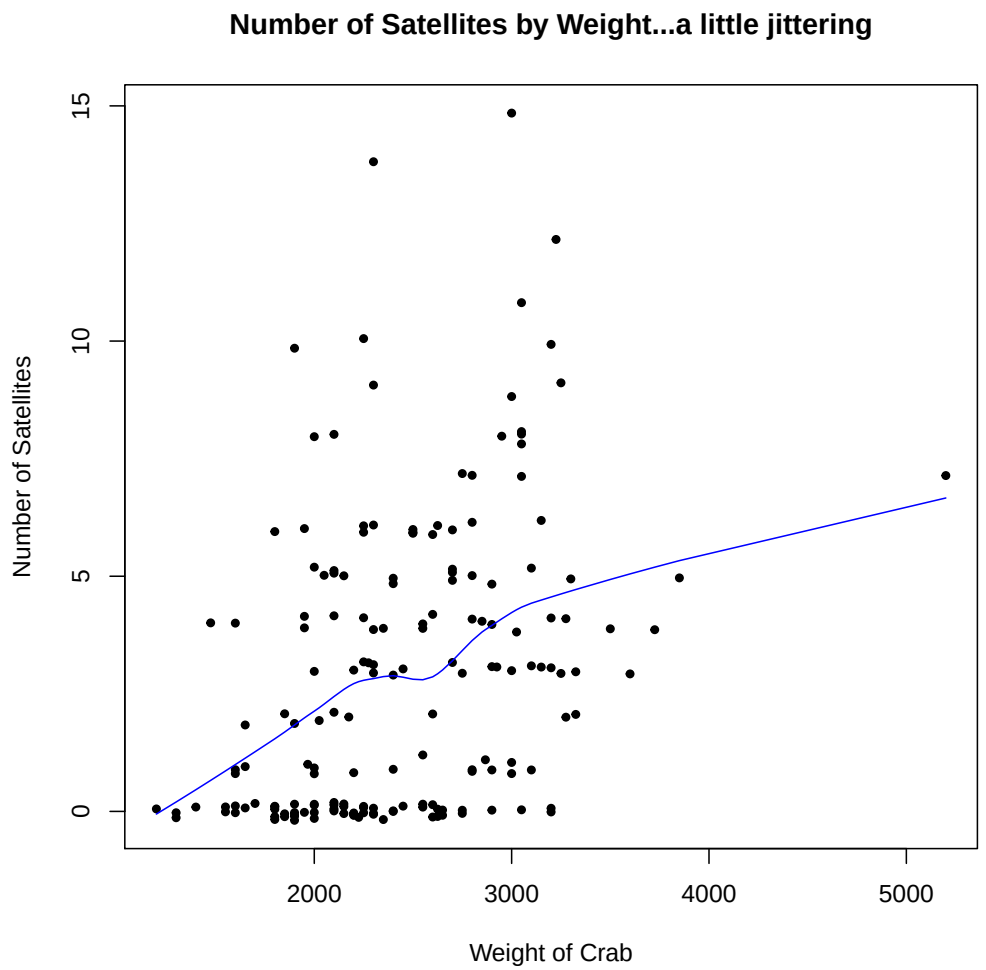


Note that I used trial and error to descide on the number of breaks. When I had it go with the default, there were too few and I wanted it to look like what is in the lecture notes.

Since we will be modeling number of satellites by width, we should look at a scatter plot. I also added a smooth (loess) curve, which may suggest a good model for the data.

```
plot(crabs$weight,jitter(crabs$satell),
   type="p",
         pch=20,
```

```
         main="Number of Satellites by Weight...a little jittering",
         xlab="Weight of Crab",
         ylab="Number of Satellites")
smooth <- loess(satell ~ weight,data=crabs)
j <- order(crabs$weight)
lines(crabs$weight[j],smooth$fitted[j],col="blue")
```



Note that I jittered the points (in a vertical direction). I did this because, there are some crabs that have 0 satellites and how many crabs with 0 doesn't show up by well without the jittering.

Before fittined the model, I put the data into a different scale:

```r
# First rescale to kg
crabs$weight <- crabs$weight/1000
```

The model that we will fit is a Poisson regression, with a log link (the default and a good choice), so

$$E(\text{number of satellites} = \beta_0 + \beta_1(\text{weight}))$$

```r
# 1. Fit poisson log-linear model
summary( mod1 <- glm(satell ~ weight,data=crabs,family=poisson) )
```

```
##
## Call:
## glm(formula = satell ~ weight, family = poisson, data = crabs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9307  -1.9981  -0.5627   0.9298   4.9992
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.42841    0.17893  -2.394   0.0167 *
## weight       0.58930    0.06502   9.064   <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: 560.87  on 171  degrees of freedom
## AIC: 920.16
##
## Number of Fisher Scoring iterations: 5
```

In the equation are you asked to esimate the mean if weight=2.44. The following does this

```r
# 2. Estimated mean if weight=2.44
(muhat_2.44<- exp(mod1$coefficients[1] + mod1$coefficients[2]*2.44))

## (Intercept)
##     2.74422
```

I did double check which coefficient was which: `mod1$coefficients[1]` is β_0 and `mod1$coefficients[2]` is β_1.

```r
# 3. For 1 unit change
exp(mod1$coefficients[2])

##   weight
## 1.802734

#  CI for beta using estimates
(lowerB <-  mod1$coefficients[2] - 1.96*0.06502)

##    weight
## 0.4618649

(upperB <-  mod1$coefficients[2] + 1.96*0.06502)

##    weight
## 0.7167433

#   CI for mean effect
exp(lowerB)

##   weight
## 1.587031

exp(upperB)

##   weight
## 2.047753
```

```r
# 4. Wald
# either report z or
9.064**2
```

```
## [1] 82.1561
```

```r
1 - pchisq(82.16,1)
```

```
## [1] 0
```

```r
# 5. Likelihood ratio test
anova(mod1)
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: satell
##
## Terms added sequentially (first to last)
##
##
##        Df Deviance Resid. Df Resid. Dev
## NULL                     172     632.79
## weight  1   71.925       171     560.87
```

```r
1-pchisq(71.925,1)
```

```
## [1] 0
```

```r
#######################################################
# 3.14 from Agresti                                   #
#######################################################

summary( mod.nb <- glm.nb(satell ~ weight, data=crabs))
```

```
##
## Call:
## glm.nb(formula = satell ~ weight, data = crabs, init.theta = 0.9310592338,
```

```
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8394  -1.4122  -0.3247   0.4744   2.1279
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.8647     0.4048  -2.136   0.0327 *
## weight        0.7603     0.1578   4.817 1.45e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.9311) family taken to be 1)
##
##     Null deviance: 216.43  on 172  degrees of freedom
## Residual deviance: 196.16  on 171  degrees of freedom
## AIC: 754.64
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.931
##           Std. Err.:  0.168
##
##  2 x log-likelihood:  -748.644

#sas reports 1/phi, which for these data is
1/0.9310

## [1] 1.074114

#  CI for beta using estimates
# (there is another was to get this--see R code under log-linear models)
(lowerB <-  mod1$coefficients[2] - 1.96*0.06502)

##    weight
## 0.4618649
```

```r
(upperB <-  mod1$coefficients[2] + 1.96*0.06502)
```

```
##    weight
## 0.7167433
```

```r
#   CI for mean effect
exp(lowerB)
```

```
##   weight
## 1.587031
```

```r
exp(upperB)
```

```
##   weight
## 2.047753
```

```r
# 4. Wald
# either report z or
4.817**2
```

```
## [1] 23.20349
```

```r
1 - pchisq(23.20,1)
```

```
## [1] 1.459973e-06
```

```r
#####################################################
# Zero inflated model                               #
#####################################################

library(ggplot2)
library(pscl)
```

```
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
```

```
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis

summary(zip <- zeroinfl(satell ~ weight | width , data = crabs))
```