Count Data, Poisson Regression

Hana Akbarnejad

3/10/2020

Problem 1

1-a

In this part, I want to fit a Poisson model (M1) with log link with W(carapace width) as the single predictor:

```
# fit poisson model
m1 = glm(sa ~ w, family=poisson(link = "log"), data=crab_data)
m1_summary = summary(m1)
m1_summary
```

```
##
## Call:
## glm(formula = sa ~ w, family = poisson(link = "log"), data = crab_data)
##
## 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 ***
## w            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 model shows that the log rate ratio of the number of satellites residing near female crabs is 0.164 per one unit increase in female crab's carapace width. The p-value for this covariate is much smaller than 0.05 and shows this covariate is highly significant in the model. We can also see that the coeffcicient is positive which means that by increase in the width it is likely that on average there are more number of crabs with the female.

Now, I would like to Check the goodness of fit and interpret the model:

```r
# gof
deviance_m1 = m1$deviance
df_m1 =  m1$df.residual              # df=n-p  173-(1+1)=171
pval1=1-pchisq(deviance_m1,df=df_m1) # chisq test
pval1                                # pvalue is 0 this shows that the fit is not good
```

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

Assessing the goodness of fit of this model, we can see that the deviance is 567.879. Comparing this value with χ^2 with degree of freedom 171, the p-value is 0. This means that this model does not fit the data well.

1-b

In this part I am going to fit a poisson model using both **weight** and **carapace width**:

```r
# fit poisson model
m2 = glm(sa ~ w+wt, family=poisson(link = "log"), data=crab_data)
m2_summary = summary(m2)
m2_summary
```

```
##
## Call:
## glm(formula = sa ~ w + wt, family = poisson(link = "log"), data = crab_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9308  -1.9705  -0.5481   0.9700   4.9905
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.29168    0.89929  -1.436  0.15091
## w            0.04590    0.04677   0.981  0.32640
## wt           0.44744    0.15864   2.820  0.00479 **
## ---
## 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: 559.89  on 170  degrees of freedom
## AIC: 921.18
##
## Number of Fisher Scoring iterations: 6
```

The model shows that the log rate ratio of the number of satellites residing near female crabs is 0.046 per one unit increase in female crab's carapace width, holding their weight constant. The p-value for this covariate is 0.326 which shows this covariate is not significant in the model anymore.

Also, we can observe that the log rate ratio of the number of satellites residing near female crabs is 0.447 per one unit increase in female crab's weight, holding their carapace width constant. The p-value for this covariate is 0.005 which shows this covariate is significant in the model.

Next I want to compare the model I have fitted with the model that I had fit using only **carapace width** as covariate. We do not consider over dispersion in this part yet.

```r
# compare m_2 and m_1
deviance_m2 = m2$deviance
df_m2 =  m2$df.residual              # df=n-p   173-3=170

test_stat =m1$deviance - m2$deviance
df =df_m1 - df_m2
pval=1-pchisq(test_stat,df=df)       # chisq test
pval
```

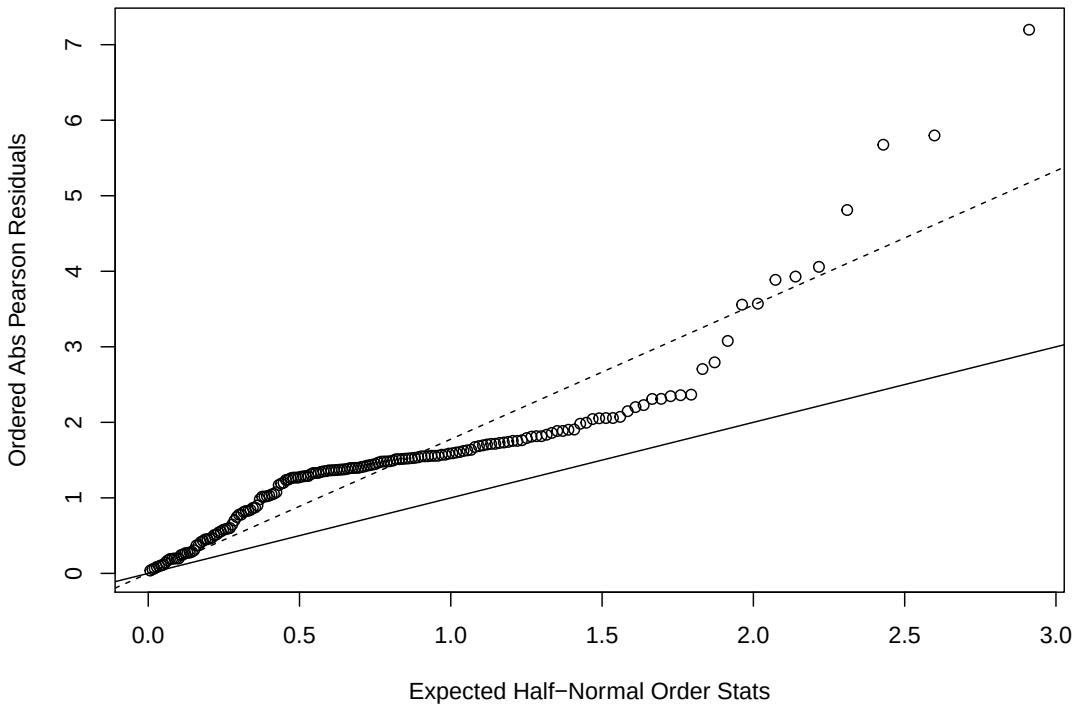
```
## [1] 0.004694838
```

Ignoring whether or not there is overdispersion in any of the models, when comparing the small model using only carapace width as predictor and the bigger model using both carapace width and weight as predictors, we can see that deviance is 559.885. Comparing this devianve value with χ^2 distribution with degree of freedom 1 (the difference between the bigger model and df of smaller model), we observe that the p-value is 0.005 This shows that we reject null hypothesis (stating that the smaller moddel is legible), and conclude that the bigger model is better, so we will consider **M2** as our model and move to the next part.

1-c

First I graph a half-normal plot to check if there is any over dispersion in model2.

```r
# check overdispersion using half-normal plot
p_res=residuals(m2,type='pearson',data=crab_data)
G = sum(p_res^2)
phi=G/170
plot(qnorm((173+1:173+0.5)/(2*173+1.125)),sort(abs(p_res)),xlab='Expected Half-Normal Order Stats',ylab=
abline(a=0,b=1)
abline(a=0,b=sqrt(phi),lty=2)
```



The above plot shows that there is probably a source of overdispersion in the model because $\phi \neq 1$. Now, I will calculate the dispersion parameter and refit the model taking this parameter into account.

Equivalently, I could fit a negative binomial model to account for overdispersion without the need to calculate dispersion paarmeter and refitting the model.

I am going to do it both ways and show the results:

```
# calc dispersion parameter based on full model
pval_disp=1-pchisq(G,df=170)
m2$deviance/m2$df.residual
```

```
## [1] 3.293442
```

```
m2_overdisp = summary(m2,dispersion=phi)
m2_overdisp
```

```
##
## Call:
## glm(formula = sa ~ w + wt, family = poisson(link = "log"), data = crab_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9308  -1.9705  -0.5481   0.9700   4.9905
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept) -1.29168    1.59771  -0.808    0.419
## w            0.04590    0.08309   0.552    0.581
## wt           0.44744    0.28184   1.588    0.112
##
## (Dispersion parameter for poisson family taken to be 3.156449)
##
##     Null deviance: 632.79  on 172  degrees of freedom
## Residual deviance: 559.89  on 170  degrees of freedom
## AIC: 921.18
##
## Number of Fisher Scoring iterations: 6
```

```
m2_nb=glm.nb(sa ~ w+wt,data=crab_data) ## glm.nb is a function from MASS to model negative binomial
summary(m2_nb)
```

```
##
## Call:
## glm.nb(formula = sa ~ w + wt, data = crab_data, init.theta = 0.9323857725,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8381  -1.3999  -0.3214   0.4884   2.1225
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.38857    1.82569  -0.761   0.4469
## w            0.02889    0.09682   0.298   0.7654
## wt           0.66340    0.35182   1.886   0.0593 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.9324) family taken to be 1)
##
##     Null deviance: 216.60  on 172  degrees of freedom
## Residual deviance: 196.24  on 170  degrees of freedom
## AIC: 756.57
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.932
##           Std. Err.:  0.168
##
##  2 x log-likelihood:  -748.572
```

The dispersion parameter is 3.156 for M2. After adjusting for over dispersion, it can be observed that the coefficients do not differ that much. However the covariate **weight** which was significant before adjusting for overdispersion becomes insignificant after adjusting for overdispersion. We can also see that the overalll results are similar using negative binomial model.

After adjusting for overdispersion, we can observe that the log rate ratio of the number of satellites residing near female crabs is 0.046 per one unit increase in female crab's carapace width, holding their weigt constant.

The p-value for this covariate after adjusting for overdispersion is 0.581 which shows this covariate is not significant in the model.

Also, after adjusting for overdispersion we can observe that the log rate ratio of the number of satellites residing near female crabs is 0.447 per one unit increase in female crab's weight, holding their carapace width constant. The p-value for this covariate after adjusting for overdispersion is 0.112 which shows this covariate is not significant in the model.

Problem 2

2-a

In this part, I fitted a Poisson model with log link function to the data with area, year, and length as predictors.

```
mod1 = glm(intensity ~ area+year+length, family=poisson(link = "log"), data=parasite_data)
summary(mod1)
```

```
##
## Call:
## glm(formula = intensity ~ area + year + length, family = poisson(link = "log"),
##     data = parasite_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -9.3632  -2.7158  -2.0142  -0.4731  30.2492
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.6431709  0.0542838  48.692  < 2e-16 ***
## area2       -0.2119557  0.0491691  -4.311 1.63e-05 ***
## area3       -0.1168602  0.0428296  -2.728  0.00636 **
## area4        1.4049366  0.0356625  39.395  < 2e-16 ***
## year2000     0.6702801  0.0279823  23.954  < 2e-16 ***
## year2001    -0.2181393  0.0287535  -7.587 3.29e-14 ***
## length      -0.0284228  0.0008809 -32.265  < 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: 25797  on 1190  degrees of freedom
## Residual deviance: 19153  on 1184  degrees of freedom
##   (63 observations deleted due to missingness)
## AIC: 21089
##
## Number of Fisher Scoring iterations: 7
```

-0.212 is the log rate ratio of the number of parasites in area 2 versus area 1, holding year and length of the fish constant.

-0.117 is the log rate ratio of the number of parasites in area 3 versus area 1, holding year and length of the fish constant.

1.405 is the log rate ratio of the number of parasites in area 4 versus area 1, holding year and length of the fish constant.

0.67 is the log rate ratio of the number of parasites in year 2000 versus 1999, holding area and length of the fish constant.

-0.218 is the log rate ratio of the number of parasites in year 2001 versus 1999, holding area and length of the fish constant.

-0.028 is the log rate ratio of the number of parasites per each one unit increase in the length of the fish, holding area and year constant.

Note that all these variables have p-value of smaller than 0.05, which shows these variables are significant in our model.

2-b

Now I am going to test the model I have built in previous part for goodness of fit.

```
deviance_mod1 = mod1$deviance
df_mod1 =  mod1$df.residual
pval_mod1=1-pchisq(deviance_mod1,df=df_mod1)
```

To assess the good of fit of this model, I used deviance analysis and observed that the model has deviance of 1.9152798×10^4 which is very high. Comparing this result with χ^2 with degree of freedom of 1184 (63 observations deleted due to missingness), the p-value is 0. This means that we should reject the null and conclude that the model does not fit the data well.

2-c

In this part I am interested in refit the model in part **a** that can account for extra zeros.

Note: In this model, I assumed that the presence or absence of parasites depend on the area that the fish is living in, and intensity of parasites (if they have any) depends on the year and length of the fish

From 1254 rows in dataset, there are 654 observations with intensity of zero. We can consider two types of zeros in the context of this problem:

- True zeros: the strains are not susceptible to parasites
- Pseudo zeros: the strains that are susceptible to parasites but the parasites have not been detected.

This said, I will fit a zero-inflated model.

I conditioned the model on area because the presence of parasite depends highly on the area we are investigating and I prefered to fix that when looking at the the difference between fish strains.

```
mod2 = zeroinfl(intensity ~ year+length | area, data=parasite_data)
summary(mod2)
```

```
##
## Call:
## zeroinfl(formula = intensity ~ year + length | area, data = parasite_data)
##
```

```
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.5077 -0.7131 -0.6447 -0.2369 26.2175
##
## Count model coefficients (poisson with log link):
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.6630528  0.0459573 101.465  < 2e-16 ***
## year2000     0.4214742  0.0278972  15.108  < 2e-16 ***
## year2001     0.0988372  0.0286162   3.454 0.000553 ***
## length      -0.0438777  0.0009298 -47.193  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.001796   0.121809   0.015    0.988
## area2        0.746780   0.183065   4.079 4.52e-05 ***
## area3        0.680876   0.161795   4.208 2.57e-05 ***
## area4       -0.882655   0.180987  -4.877 1.08e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 15
## Log-likelihood: -7563 on 8 Df
```

The model built has two parts: Count model and Zero-inflation model.

For the count model:

Given that the fish strain is susceptible to parasites, 0.421 is the log rate ratio of the number of parasites in year 2000 versus 1999, holding length constant.

Given that the fish strain is susceptible to parasites, 0.099 is the log rate ratio of the number of parasites in year 2001 versus 1999, holding length constant.

Given that the fish strain is susceptible to parasites, -0.044 is the log rate ratio of the number of parasites per 1 unit increase in the length of the fish, holding year constant.

For the zero-inflated model:

0.747 is the log odds ratio of being the fish strain not being susceptible to parasites in area 2 compared to area 1.

0.681 is the log odds ratio of being the fish strain not being susceptible to parasites in area 3 compared to area 1.

-0.883 is the log odds ratio of being the fish strain not being susceptible to parasites in area 4 compared to area 1.

We can see that all predictors have p-va;ue smaller than 0.05 which shows that the predictors are statistically significance in the model.