P8131-HW5-yz4184

Yunlin Zhou

3/24/2022

```
library(tidyverse)
library(pscl)
```

Import the data

```
crab_df = read.csv("./HW5-crab.txt", sep = "")

para_df = read.csv("./HW5-parasite.txt", sep = "")%>%
  janitor::clean_names()%>%
  mutate(area = as.factor(area),
         year = as.factor(year)) %>%
  select(year, intensity, length, area)
```

Question 1

(a)

Fit a Poisson model (M1) with log link with W as the single predictor.

```
M1 <- glm(Sa ~ W,
          family = poisson(link = log),
          data=crab_df)

summary(M1)
```

```
##
## Call:
## glm(formula = Sa ~ W, family = poisson(link = log), data = crab_df)
##
## 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 log rate ratio of satellites number would increase 0.16405 as carapace width increases 1 unit. The deviance of this model is 567.88 with 171 df. Carapace width is statistically significant.

goodness of fit

```
# pearson residual
P1 = sum(residuals(M1, type = 'pearson')^2)

p_val1 = 1 - pchisq(P1, df = 171)
p_val1
```

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

Since the p-value of the goodness fit is less than 0.5, we reject the null hypothesis that the M1 is a good fit.

(b)

Fit a model (M2) with W and Wt as predictors.

```
M2 <- glm(Sa ~ W + Wt,
          family = poisson(link = log),
          data=crab_df)

summary(M2)
```

```
##
## Call:
## glm(formula = Sa ~ W + Wt, family = poisson(link = log), data = crab_df)
##
## 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 log rate ratio of satellites number would increase 0.04590 as carapace width increases 1 unit, holding other variables fixed.

The log rate ratio of satellites number would increase 0.44744 as weight increases 1 unit, holding other variables fixed.

The deviance of this model is 559.89 with 170 df. Weight is statistically significant.

Compare M1 with M2

```
# analysis of deviance
D1 = M1$deviance - M2$deviance

p_val2 = 1 - pchisq(D1, 1)
p_val2
```

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

Since the p-value of the goodness fit is less than 0.5, we reject the null hypothesis that the M1 is better.

(c)

Check over dispersion in M2.

```
# Pearson residual
P2=sum(residuals(M2, type = 'pearson')^2)

phi=P2/170
phi
```

```
## [1] 3.156449
```

```
summary(M2, dispersion = phi)
```

```
##
## Call:
## glm(formula = Sa ~ W + Wt, family = poisson(link = log), data = crab_df)
```

```
##
## 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
```

Since phi is 3.156449 which is greater than 1, we can conclude that there is over-dispersion in this model.

After adjusting for over dispersion, the standard error of carapace width changes from 0.04677 to 0.08309; the standard error of weight changes from 0.15864 to 0.28184.

The log rate ratio of satellites number would increase 0.04590 as carapace width increases 1 unit, holding other variables fixed.

The log rate ratio of satellites number would increase 0.44744 as weight increases 1 unit, holding other variables fixed.

The deviance of this model is 559.89 with 170 df.

Question 2

(a)

Fit a Poisson model with log link to the data with area, year, and length as predictors.

```
M3 = glm(intensity ~ area + length + year,
              family = poisson(link = log),
              data = para_df)

summary(M3)
```

```
##
## Call:
## glm(formula = intensity ~ area + length + year, family = poisson(link = log),
##     data = para_df)
##
## 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 ***
## length      -0.0284228  0.0008809 -32.265  < 2e-16 ***
## year2000     0.6702801  0.0279823  23.954  < 2e-16 ***
## year2001    -0.2181393  0.0287535  -7.587 3.29e-14 ***
## ---
## 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
```

The log rate ratio of intensity would decrease 0.2119557 as the area changes from 1 to 2, holding other variables fixed.

The log rate ratio of intensity would decrease 0.1168602 as the area changes from 1 to 3, holding other variables fixed.

The log rate ratio of intensity would increase 1.4049366 as the area changes from 1 to 4, holding other variables fixed.

The log rate ratio of intensity would decrease 0.0284228 as length increases 1 unit, holding other variables fixed.

The log rate ratio of intensity would increase 0.6702801 as the year changes from 1999 to 2000, holding other variables fixed.

The log rate ratio of intensity would decrease 0.2181393 as the year changes from 1999 to 2001, holding other variables fixed.

(b)

Test for goodness of fit of the model.

```
P3 = sum(residuals(M3, type = 'pearson')^2)

p_val3 = 1 - pchisq(P3, df = 1184)
p_val3
```

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

Since the p-value of the goodness fit is less than 0.5, we reject the null hypothesis that the M3 is a good fit.

(C)

Fit zero-inflated poisson regression model

```
M4 <- zeroinfl(intensity ~ area + length + year ,
               data = para_df)
summary(M4)
```

```
##
## Call:
## zeroinfl(formula = intensity ~ area + length + year, data = para_df)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -2.1278 -0.8265 -0.5829 -0.1821 25.4837
##
## Count model coefficients (poisson with log link):
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.8431720  0.0583793  65.831  < 2e-16 ***
## area2        0.2687838  0.0500467   5.371 7.84e-08 ***
## area3        0.1463174  0.0439485   3.329 0.000871 ***
## area4        0.9448070  0.0368342  25.650  < 2e-16 ***
## length      -0.0368067  0.0009747 -37.762  < 2e-16 ***
## year2000     0.3919828  0.0282952  13.853  < 2e-16 ***
## year2001    -0.0448457  0.0296057  -1.515 0.129831
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.552579   0.275762   2.004  0.04509 *
## area2        0.718680   0.189552   3.791  0.00015 ***
```

```
## area3        0.657710   0.167402   3.929 8.53e-05 ***
## area4       -1.022864   0.188201  -5.435 5.48e-08 ***
## length      -0.009889   0.004629  -2.136  0.03266 *
## year2000    -0.752121   0.172965  -4.348 1.37e-05 ***
## year2001     0.456533   0.143962   3.171  0.00152 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 17
## Log-likelihood: -6950 on 14 Df
```

Within the susceptible fish, we use the poisson model:

The log rate ratio of intensity would increase 0.2687838 as the area changes from 1 to 2, holding other variables fixed.

The log rate ratio of intensity would increase 0.1463174 as the area changes from 1 to 3, holding other variables fixed.

The log rate ratio of intensity would increase 0.9448070 as the area changes from 1 to 4, holding other variables fixed.

The log rate ratio of intensity would decrease 0.0368067 as length increases 1 unit, holding other variables fixed.

The log rate ratio of intensity would increase 0.3919828 as the year changes from 1999 to 2000, holding other variables fixed.

The log rate ratio of intensity would decrease 0.0448457 as the year changes from 1999 to 2001, holding other variables fixed.

Since we donot know which fish are susceptible, we assume that all the variables are realted to binomial model.

The log rate ratio of intensity would increase 0.718680 as the area changes from 1 to 2, holding other variables fixed.

The log rate ratio of intensity would increase 0.657710 as the area changes from 1 to 3, holding other variables fixed.

The log rate ratio of intensity would decrease 1.022864 as the area changes from 1 to 4, holding other variables fixed.

The log rate ratio of intensity would decrease 0.009889 as length increases 1 unit, holding other variables fixed.

The log rate ratio of intensity would decrease 0.752121 as the year changes from 1999 to 2000, holding other variables fixed.

The log rate ratio of intensity would increase 0.456533 as the year changes from 1999 to 2001, holding other variables fixed.