HW5

Anna Ma

3/23/2022

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

Problem 1

Import data

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

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

```
crab_M1 = glm(Sa ~ W, data = crab, family = poisson(link = log))
summary(crab_M1)
```

```
##
## Call:
## glm(formula = Sa ~ W, 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 ***
## 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
```

Goodness of fit:

```r
#Deviance Analysis D_null-Dm ~ X^2 df=1
pval_dev= 1 - pchisq(crab_M1$null.deviance - crab_M1$deviance, df = 1)
#Pearson
G_crab_M1 = sum(residuals(crab_M1, type = 'pearson')^2)
pval_pear = 1 - pchisq(G_crab_M1, df = 171)
```

From the summary, the null deviance is 632.79 and the model deviance is 567.88. Using deviance analysis, we found that the p-value $= 7.7715612 \times 10^{-16}$. Similarly, using pearson residual test, we have a p-value of 0. Both p-value are less than 0.05, therefore, we reject the null hypothesis and conclude that the model does not fit the data well.

Interpretation:

The log rate ratio of the number of satellites is 0.1640451 given a unit change in the crab's carapace width. That is, the count of satellites changes 1.1782674 times per unit change in carapace width.

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

```r
crab_M2 = glm(Sa ~ W + Wt, data = crab, family = poisson(link = log))
summary(crab_M2)
```

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

Goodness of fit:

```r
#Deviance Analysis
Dev_dif_M12 = crab_M1$deviance - crab_M2$deviance
p_M12 = 1 - pchisq(Dev_dif_M12,df = 1)
```

The p value of the deviance test is $0.0046948 < 0.05$. Therefore, we reject the null hypothesis that M2 fits the data as well as M1, and conclude that M2 is a better fit than M1.

Interpretation:

- The log rate ratio of the number of satellites is 0.045898per unit change in carapace width, holding weight constant. That is, the count of satellites changes 1.0469677 times per unit change in carapace width.
- The log rate ratio of the number of satellites is 0.4474357per unit change in weight, holding carapace width constant.That is, the count of satellites changes 1.5642957 times per unit change in weight.

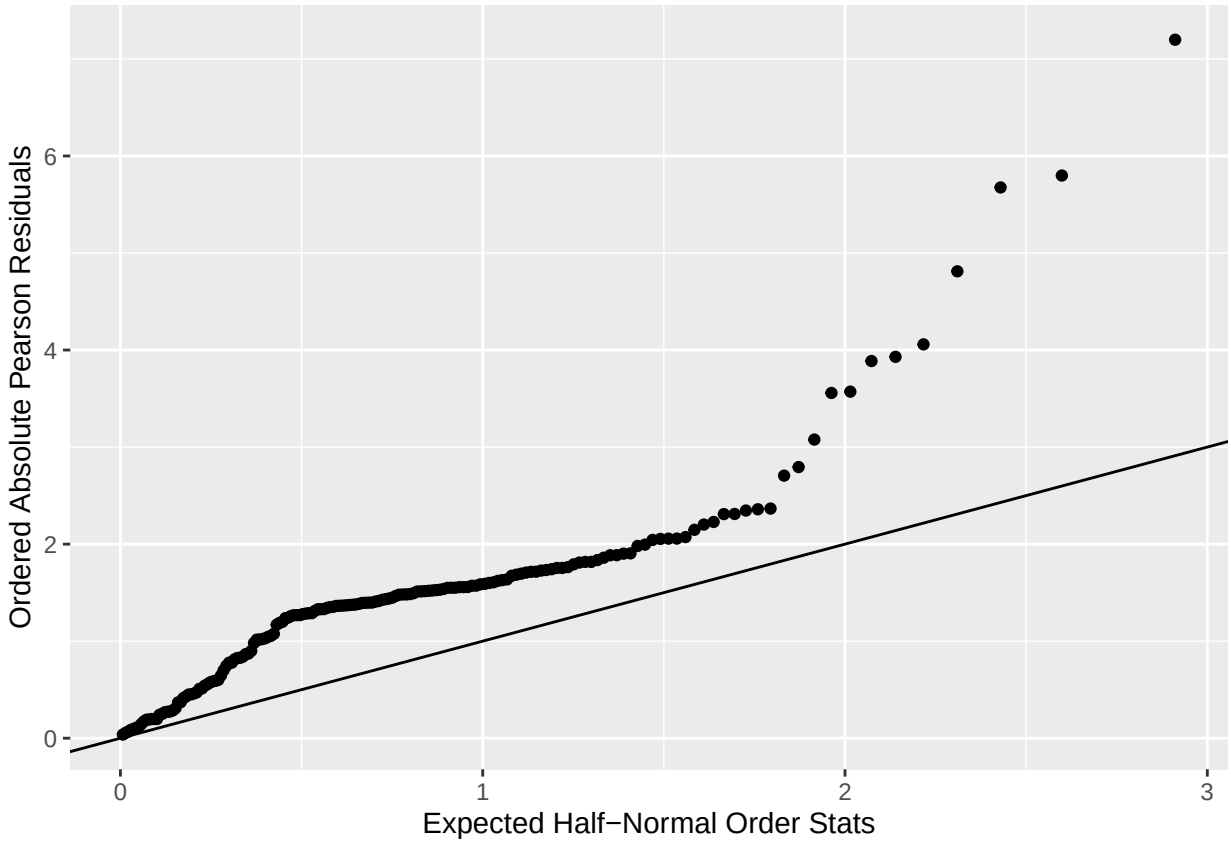
c) Check over dispersion in M2.

```
res_M2 = residuals(crab_M2, type = 'pearson')

G = sum(res_M2 ^ 2)
#Dispersion parameter
phi = G / (173 - 3)

res = tibble(x = qnorm((173 + 1:173 + 0.5) / (2 * 173 + 1.125)),
             y = sort(abs(res_M2)))

res %>% ggplot(aes(x = x,y = y)) +
  geom_point() +
  geom_abline(slope = 1) +
  labs(x = 'Expected Half-Normal Order Stats',
       y = 'Ordered Absolute Pearson Residuals')
```



From the half normal plot, we can see that dispersion exists in M2. The dispersion parameter phi is calculated to be $\phi = 3.156$

Adjust for dispersion

```
M2_disp = summary(crab_M2, dispersion = phi)
M2_disp
```

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

The estimate of the adjusted model is the same with the original model without adjustment of dispersion. Therefore, the interpretation stays the same that

- The log rate ratio of the number of satellites is 0.045898per unit change in carapace width, holding weight constant.That is, the count of satellites changes 1.0469677 times per unit change in carapace width.
- The log rate ratio of the number of satellites is 0.4474357per unit change in weight, holding carapace width constant. That is, the count of satellites changes 1.5642957 times per unit change in weight.

Problem 2

import data

```
parasite = read.csv("HW5 data/HW5-parasite.txt", sep = "") %>%
  janitor::clean_names() %>%
  mutate(year = factor(year), area = factor(area)) %>%
  drop_na()
```

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

```
para_M1 = glm(intensity ~ area + year + length, data = parasite, family = poisson(link = log))
summary(para_M1)
```

```
##
## Call:
```

```
## glm(formula = intensity ~ area + year + length, family = poisson(link = log),
##     data = parasite)
##
## 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
## AIC: 21089
##
## Number of Fisher Scoring iterations: 7
```

Interpretation:

- The log count of parasites is 2.6431709 in year 1999, area 1, and fish length 0. That is, the count of parasite for a fish of 0 length in year 1999 at area 1 is 14.0577088
- The log rate ratio of parasite intensity is -0.2119557 given area change from 1 to 2, holding year and fish length constant.That is, the counts of parasite for fish in area 2 is 0.8090006 times the counts of parasite for fish in area 1.
- The log rate ratio of parasite intensity is -0.1168602 given area change from 1 to 3, holding year and fish length constant.That is, the counts of parasite for fish in area 3 is 0.8897096 times the counts of parasite for fish in area 1.
- The log rate ratio of parasite intensity is 1.4049366 given area change from 1 to 4, holding year and fish length constant.That is, the counts of parasite for fish in area 4 is 4.0752685 times the counts of parasite for fish in area 1.
- The log rate ratio of parasite intensity is 0.6702801 given year change from 1999 to 2000, holding area and fish length constant.That is, the counts of parasite in 2000 is 1.9547848 times the counts of parasite in 1999.
- The log rate ratio of parasite intensity is -0.2181393 given year change from 1999 to 2001, holding area and fish length constant. That is, the counts of parasite in 2001 is 0.8040134 times the counts of parasite in 1999.
- The log rate ratio of parasite intensity is -0.0284228 for 1 unit change in length, holding area and year constant. That is, the counts of parasite changes by 0.9719773 times for every unit change in fish length.

b) Test for goodness of fit of the model

```
#deviance
pval_para_dev = 1 - pchisq(para_M1$deviance, nrow(parasite) - 7)
#pearson
G_para_M1 = sum(residuals(para_M1, type = 'pearson')^2)
pval_para_G = 1 - pchisq(G_para_M1, 1187)
```

The deviance test gives a p value of 0, and the pearson test gives a p value of 0. Both p-values are less than 0.05, therefore, we reject the null hypothesis and conclude that the model does not fit the data well.

c) Take consideration of zero-inflation and fit appropriate model.

```
para_M2 = zeroinfl(intensity ~ year + length | area, data = parasite)
summary(para_M2)
```

```
##
## Call:
## zeroinfl(formula = intensity ~ year + length | area, data = parasite)
##
## 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)  4.6630522  0.0459573 101.465  < 2e-16 ***
## year2000     0.4214742  0.0278972  15.108  < 2e-16 ***
## year2001     0.0988373  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)  0.001797   0.121809   0.015    0.988
## area2        0.746780   0.183065   4.079 4.52e-05 ***
## area3        0.680875   0.161795   4.208 2.57e-05 ***
## area4       -0.882654   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: 11
## Log-likelihood: -7563 on 8 Df
```

Interpretation

1) Poisson Model

- In the fish that is susceptible to parasite, the log rate ratio of parasite intensity is 0.421 given year change from 1999 to 2000, holding length constant.That is, the count of parasite in 2000 is 1.5234843 times the count of parasite in 1999.
- In the fish that is susceptible to parasite, the log rate ratio of parasite intensity is 0.098 given year change from 1999 to 2001, holding length constant.That is, the count of parasite in 2001 is 1.1029628 times the count of parasite in 1999.

- In the fish that is susceptible to parasite, the log rate ratio of parasite intensity is -0.044 given a unit change in the fish length, holding year constant. That is, the count of parasite changes by 0.9570497 times for a unit change in fish length.

2) Binomial model

- The log odds ratio of whether a fish is susceptible to parasite is 0.747 given area change from 1 to 2. That is, the odds of a susceptible for fish in area 2 is 2.1106585 times compare to fish in area 1.
- The log odds ratio of whether a fish is susceptible to parasite is 0.681 given area change from 1 to 3. That is, the odds of a susceptible for fish in area 3 is 1.9758526 times compare to fish in area 1.
- The log odds ratio of whether a fish is susceptible to parasite is -0.883 given area change from 1 to 4. That is, the odds of a susceptible for fish in area 4 is 0.4135404 times compare to fish in area 1.