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 count of satellites changes by a factor of 1.178 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 count of satellites changes by a factor of 1.047 times per unit change in carapace width.

- The count of satellites changes by a factor of 1.564 times per unit change in weight.

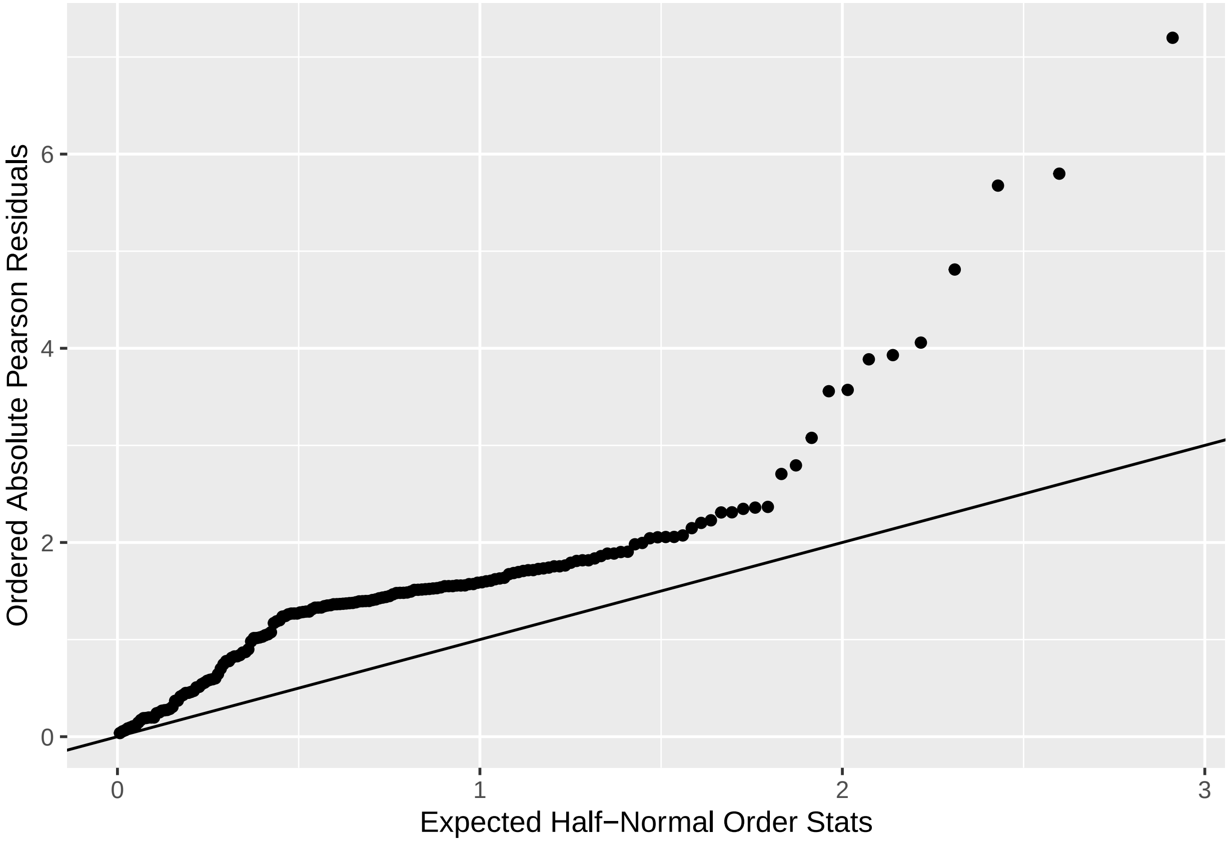
c) Check over dispersion in M2.

```
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
```

```
round(exp(M2_disp$coefficients),3)
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.275      4.942   0.446    1.520
## W              1.047      1.087   1.737    1.787
## Wt             1.564      1.326   4.892    1.119
```

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 count of satellites changes by a factor of 1.047 times per unit change in carapace width.
- The count of satellites changes by a factor of 1.564 times per unit change in carapace width.

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
```

```
round(exp(para_M1$coefficients),3)
```

```
## (Intercept)       area2       area3       area4    year2000    year2001
##      14.058       0.809       0.890       4.075       1.955       0.804
##      length
##       0.972
```

Interpretation:

- The intensity of parasite of fish in area 2 is 0.809 times the parasite intensity of fish in area 1, holding year and fish length constant.
- The intensity of parasite of fish in area 3 is 0.89 times the parasite intensity of fish in area 1, holding year and fish length constant.
- The intensity of parasite of fish in area 4 is 4.075 times the parasite intensity of fish in area 1, holding year and fish length constant.
- The intensity of parasite of fish in 2000 is 1.955 times the parasite intensity of fish in 1999, holding area and fish length constant.
- The intensity of parasite of fish in 2001 is 0.804 times the parasite intensity of fish in 1999, holding area and fish length constant.
- The intensity of parasite decrease by 2.8% for every unit increase in length, holding area and year length constant.

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
## -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 ***
## year2000     0.3919828  0.0282952  13.853  < 2e-16 ***
## year2001    -0.0448457  0.0296057  -1.515 0.129831
## length      -0.0368067  0.0009747 -37.762  < 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 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.552579   0.275762   2.004  0.04509 *
## year2000    -0.752121   0.172965  -4.348 1.37e-05 ***
## year2001     0.456533   0.143962   3.171  0.00152 **
## length      -0.009889   0.004629  -2.136  0.03266 *
## 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 ***
## ---
## 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
```

```
round(exp(para_M2$coefficients$count),3)
```

```
## (Intercept)    year2000    year2001      length       area2       area3
##      46.673       1.480       0.956       0.964       1.308       1.158
##       area4
##       2.572
```

```
round(exp(para_M2$coefficients$zero),3)
```

```
## (Intercept)    year2000    year2001      length       area2       area3
##       1.738       0.471       1.579       0.990       2.052       1.930
##       area4
##       0.360
```

Interpretation

1) Poisson Model

- In the fish that is susceptible to parasite, the parasite intensity for fish in 2000 is 1.48 times the intensity of fish in 1999, holding length and area constant.
- In the fish that is susceptible to parasite, the parasite intensity for fish in 2001 is 0.956 times the intensity of fish in 1999, holding length and area constant.
- In the fish that is susceptible to parasite, the parasite intensity is 0.964 times for every unit change in length, holding year and are constant. That is, parasite intensity decrease by 4% with every unit increase in the length of the fish.
- In the fish that is susceptible to parasite, the parasite intensity for fish in area 2 is 1.308 times the intensity of fish in area 1, holding length and year constant.
- In the fish that is susceptible to parasite, the parasite intensity for fish in area 3 is 1.158 times the intensity of fish in area 1, holding length and year constant.
- In the fish that is susceptible to parasite, the parasite intensity for fish in area 4 is 2.572 times the intensity of fish in area 1, holding length and year constant.

2) Binomial model

- The odds ratio of not susceptible to parasite is 0.471 for fish in year 2000 compare to fish in 1999 holding area and length constant.
- The odds ratio of not susceptible to parasite is 1.579 for fish in year 2001 compare to fish in 1999 holding area and length constant.
- The odds ratio of not susceptible to parasite is 0.99 for every unit increase in fish length, holding area and year constant. That is, for every unit increase in length, the odds ratio of not susceptible to parasite decreases by 1%.
- The odds ratio of not susceptible to parasite is 2.052 for fish in area 2 compare to fish in area 1, holding year and length constant.
- The odds ratio of not susceptible to parasite is 1.93 for fish in area 3 compare to fish in area 1, holding year and length constant.
- The odds ratio of not susceptible to parasite is 0.36 for fish in area 4 compare to fish in area 1, holding year and length constant.