P8131 Assignment 5

Yihan Feng

2021/3/6

Problem 1

```
crab.df = read.csv("./data/HW5-crab.txt", sep = "") %>%
  janitor::clean_names()
```

(a) Fit a Poisson model (M1) with log link with W as the single predictor. Check the goodness of fit and interpret your model.

```
crab.m1 = glm(sa ~ w,
              family = poisson(link = log),
              data = crab.df)
summary(crab.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
```

Goodness of fit:

```
G1 = sum(residuals(crab.m1, type = 'pearson')^2)
1 - pchisq(G1, df = 171)
```

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

P value is less than 0.05, so it rejects the null hypothesis at 0.05 significance level, and conclude that model 1 is poorly fitted.

Model interpretation:

The relative risk of satellites number per unit change increases the carapace width is 1.1782674, holding the other variables fixed.

(b) Fit a model (M2) with W and Wt as predictors. Compare it with the model in (a). Interpret your results.

```
crab.m2 = glm(sa ~ w + wt,
              family = poisson(link = log),
              data = crab.df)
summary(crab.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
```

Goodness of fit (M1 and M2):

```
cm12.dtest = crab.m1$deviance - crab.m2$deviance
cm12.dtest.p = 1 - pchisq(cm12.dtest, 1)
```

Since p value of deviance test is $0.0046948 < 0.05$, at 95% confidence, we are able to reject the null hypothesis, and conclude that model 2 is better fitted.

Model interpretation:

- The relative risk of satellites number per unit change increases the carapace width is 1.0469677, holding the other variables fixed.
- The relative risk of satellites number per unit change increases the weight is 1.5642957, holding the other variables fixed.

(c) Check over dispersion in M2. Interpret the model after adjusting for over dispersion.

```
phi=G1 / (173 - 3)
phi
```

```
## [1] 3.200924
```

As ϕ is 3.2009236, there exists over dispersion in M2.

```
summary(crab.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    1.60893  -0.803    0.422
## w            0.04590    0.08367   0.549    0.583
## wt           0.44744    0.28382   1.576    0.115
##
## (Dispersion parameter for poisson family taken to be 3.200924)
##
##     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
```

Model interpretation:

- The relative risk of satellites number per unit change increases the carapace width is 1.0469677, holding the other variables fixed.
- The relative risk of satellites number per unit change increases the weight is 1.5642957, holding the other variables fixed.

Problem 2

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

(a) Fit a Poisson model with log link to the data with area, year, and length as predictors. Interpret each model parameter.

```
para.m1 = glm(intensity ~ area + length + year,
              family = poisson(link = log),
              data = para.df)
summary(para.m1)
```

```
##
## 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
```

Model interpretation:

- The relative risk of parasite intensity in area 2 vs. area 1 is 0.8090006, holding the other variables fixed.
- The relative risk of parasite intensity in area 3 vs. area 1 is 0.8897096, holding the other variables fixed.
- The relative risk of parasite intensity in area 4 vs. area 1 is 4.0752685, holding the other variables fixed.
- The relative risk of parasite intensity per unit change increases the length is 0.9719773, holding the other variables fixed.
- The relative risk of parasite intensity in year 2000 vs. year 1999 is 1.9547848, holding the other variables fixed.
- The relative risk of parasite intensity in year 2001 vs. year 1999 is 0.8040134, holding the other variables fixed.

(b) Test for goodness of fit of the model in (a) and state conclusions.

```
G2 = sum(residuals(para.m1, type = 'pearson')^2)
1 - pchisq(G2, df = 171)
```

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

P value is less than 0.05, so it rejects the null hypothesis at 0.05 significance level, and conclude that model 1 is poorly fitted.

(c) Researchers suspect that there may be two strains of fish, one that is susceptible to parasites and one that is not. Without knowing which fish are susceptible, this could be regarded as a zero-inated model. Building on the model in (a) (using the same predictors), fit an appropriate model to the data that can account for extra zeros. Provide an interpretation for each model parameter in terms of the problem.

```
para.m2 <- zeroinfl(intensity ~ length + year + area, data = para.df)
summary(para.m2)
```

```
##
## Call:
## zeroinfl(formula = intensity ~ length + year + area, 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 ***
## 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
## 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 *
## 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 **
## 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
```

Poisson model:

- In the fish that are susceptible to parasites, the relative risk of parasite intensity per unit change increases the length is 0.9638624, holding the other variables fixed.
- In the fish that are susceptible to parasites, the relative risk of parasite intensity in year 2000 vs. year 1999 is 1.4799122, holding the other variables fixed.
- In the fish that are susceptible to parasites, the relative risk of parasite intensity in year 2001 vs. year 1999 is 0.956145, holding the other variables fixed.
- In the fish that are susceptible to parasites, the relative risk of parasite intensity in area 2 vs. area 1 is 1.3083722, holding the other variables fixed.
- In the fish that are susceptible to parasites, the relative risk of parasite intensity in area 3 vs. area 1 is 1.1575635, holding the other variables fixed.
- In the fish that are susceptible to parasites, the relative risk of parasite intensity in area 4 vs. area 1 is 2.572317, holding the other variables fixed.

Zero-inflation model:

- In the fish that are susceptible to parasites, the odds ratio of parasite intensity per unit change increases the length is 0.9901599, holding the other variables fixed.
- In the fish that are susceptible to parasites, the odds ratio of parasite intensity in year 2000 vs. year 1999 is 0.4713656, holding the other variables fixed.
- In the fish that are susceptible to parasites, the odds ratio of parasite intensity in year 2001 vs. year 1999 is 1.578591, holding the other variables fixed.
- In the fish that are susceptible to parasites, the odds ratio of parasite intensity in area 2 vs. area 1 is 2.0517222, holding the other variables fixed.
- In the fish that are susceptible to parasites, the odds ratio of parasite intensity in area 3 vs. area 1 is 1.9303664, holding the other variables fixed.
- In the fish that are susceptible to parasites, the odds ratio of parasite intensity in area 4 vs. area 1 is 0.3595635, holding the other variables fixed.