hw5- Possion regression case study

Cary Ni

2023-03-26

Problem 1

```
crab_df = read.table("./data/hw5_crab.txt", header = TRUE) %>%
  janitor::clean_names() %>%
  mutate(
    c = as.factor(c),
    s = as.factor(s)
  ) %>%
  select(-number)
```

```
# fit model 1 with log link
model_1 = glm(sa ~ w, family = "poisson", data = crab_df)
summary(model_1)
```

```
##
## Call:
## glm(formula = sa ~ w, family = "poisson", 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
```

```
# get pearson G stat
res_1 = residuals(model_1, type = "pearson")
```

```
G_1 = sum(res_1^2)
pval_1 = 1 - pchisq(G_1, df = 171)
pval_1
```

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

(a) The high pearson G statistic with a p-value close to 0 indicates that the model is lack of fit. The summary tells that for every one unit increase in carapace width, the log of expected number of satellites will increase by 0.164 (or to a factor of 1.178 in a standard scale number after exp transformation).

```
# fit model 2 with log link
model_2 = glm(sa ~ w + wt, family = "poisson", data = crab_df)
summary(model_2)
```

```
##
## Call:
## glm(formula = sa ~ w + wt, family = "poisson", 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
```

```
# deviance analysis (without disperson)
test_stat = model_1$deviance - model_2$deviance
pval_2 = 1-pchisq(test_stat, df = 1)
pval_2
```

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

(b) The deviance analysis shows that the model 2 has better fit to the data compared to model 1 with a p-value of 0.0047, suggesting that adding the predictor of weight accounts for more variation in number of satellites.

```
# get dispersion factor from pearson
res_2 = residuals(model_2, type = "pearson")
G_2 = sum(res_2^2)
phi = G_2/170
phi
```

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

```
# get dispersion factor from deviance
model_2$deviance/model_2$df.residual
```

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

```
# model 2 with dispersion factor
summary(model_2, dispersion = phi)
```

```
## 
## Call:
## glm(formula = sa ~ w + wt, family = "poisson", 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
```

```
# gof after adding dispersion factor
1 - pchisq(G_2/phi, df = 170)
```

```
## [1] 0.4855753
```

(c) The dispersion factor calculated from pearson residual is around 3.16 and from deviance it is 3.29. The pearson statistic gives a p value of 0.49, which suggests the model 2 fit is good after considering the dispersion factor. The modified model 2 has larger standard error for the coefficient of carapace width and weight. Holding other variables fixed, for every one unit increase in carapace width, the log of expected number of satellites will increase by 0.046 (or to a factor of 1.047 in a standard scale number after exp transformation). Holding other variables fixed, for every one unit increase in weight, the log of expected number of satellites will increase by 0.447 (or to a factor of 1.564 in a standard scale number after exp transformation).

Problem 2

```
para_df = read.table("./data/hw5_parasite.txt", header = TRUE) %>%
  janitor::clean_names() %>%
  mutate(
    area = as.factor(area),
    year = as.factor(year)
  ) %>%
  select(-sample) %>%
  na.omit()
```

```
# fit model 3 with log link
model_3 = glm(intensity ~ area + year + length, family = "poisson", data = para_df)
summary(model_3)
```

```
##
## Call:
## glm(formula = intensity ~ area + year + length, family = "poisson",
##     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 ***
## 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
```

(a) Holding year and fish length fixed, compared to fish in area 1, fish lives in area 2 has a lower expected number of parasites to a factor of 0.809, fish lives in area 3 has a lower expected number of parasites to a factor of 0.809, fish lives in area 4 has a higher expected number of parasites to a factor of 4.076. Holding area and fish length fixed, compared to year 1999, fish in year 2000 has a higher expected number of parasites to a factor of 1.954 while fish in year 2001 a lower expected number of parasites to a factor of 0.804. Holding year and area fixed, for every one unit increase in length, the log of expected number of satellites will decrease by 0.028 (or to a factor of 0.972 in a standard scale number after exp transformation).

```
# get pearson G stat
res_3 = residuals(model_3, type = "pearson")
G_3 = sum(res_3^2)
pval_3 = 1 - pchisq(G_3, df = 1184)
pval_3
```

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

```
# use deviance
1 - pchisq(model_3$deviance, df = 1184)
```

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

```
# try with constant dispersion factor
phi_3 = G_3/1184
phi_3
```

```
## [1] 35.61231
```

```
pval_4 = 1 - pchisq(G_3/phi_3, df = 1184)
pval_4
```

```
## [1] 0.4945345
```

```
# The model will be a good fit when a constant dispersion factor is introduced indicated by a p value o
```

(b) The test for goodness of fit gives a p value close to 0 for both G stat and D stat, which means the model is lack of fit (without dispersion factor). Thus, it could be say that the variable intensity (number of parasites) does not follow the poisson distribution (without dispersion factor).

```
# fit zero-inflated poisson model
model_z = zeroinfl(intensity ~ area + year + length | area + year + length,
                   data = para_df)
summary(model_z)
```

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

(c) **For logistic model** : Holding year and fish length fixed, compared to fish in area 1, fishes lived in area 2 and area 3 have higher chance in belonging to the group without parasites (insusceptible) while fisher lived in area 4 have a higher chance in belonging to the group with parasites (susceptible).(change the odds of being insusceptible group by 2.054, 1.931, and 0.36 respectively) Holding area and fish length fixed, compared to fish in year 1999, fish in year 2000 has a higher chance in belonging to the group with parasites (susceptible) while fisher in 2001 have a higher chance in belonging to the group without parasites (insusceptible). (change the odds of being insusceptible group by factors of 0.472, 1.578 respectively) Holding year and area fixed, every unit increase in fish length leads to lower odds of being fish group without parasite (insusceptible). (change the odds of being insusceptible group by a factor of 0.99)

(d) **For poisson model** : Within the group of fish with parasites (susceptible), holding year and fish length fixed, compared to fish in area 1, fishes in area 2, 3, and 4 have higher number of parasites (to factors of 1.309 ,1.157, and 2.573 respectively.) Holding area and fish length fixed, compared to year 1999, fish in 2000 has more parasites (to a factor of 1.48) while fish in 2001 has less parasites (to a factor of 0.956). Holding area and year fixed, every unit increase in fish length leads to less number of parasites (to a factor of 0.964).