HW5

Yangyang Chen

2024-03-18

Nesting Horsehoe Crabs

In a study of nesting horseshoe crabs, each female horseshoe crab had a male crab attached to her in her nest. The study investigated factors that affect whether the female crab had any other males, called satellites, residing near her. Explanatory variables that are thought to affect this included the female crab's color (C), spine condition (S), carapace width (W) and weight (Wt). The response outcome for each female crab is her number of satellites (Sa). There are 173 females in this study. Data are provided in the crab.txt.

(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_df = read_table("HW5-crab.txt")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   number = col_double(),
##   C = col_double(),
##   S = col_double(),
##   W = col_double(),
##   Wt = col_double(),
##   Sa = col_double()
## )
```

```
crab_df |> head()
```

```
## # A tibble: 6 x 6
##   number     C     S     W    Wt    Sa
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1      1     2     3  28.3  3.05     8
## 2      2     3     3  26    2.6      4
## 3      3     3     3  25.6  2.15     0
## 4      4     4     2  21    1.85     0
## 5      5     2     3  29    3        1
## 6      6     1     2  25    2.3      3
```

```
crab_df |> str()
```

```
## spc_tbl_ [173 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ number: num [1:173] 1 2 3 4 5 6 7 8 9 10 ...
##  $ C     : num [1:173] 2 3 3 4 2 1 4 2 2 2 ...
##  $ S     : num [1:173] 3 3 3 2 3 2 3 3 1 3 ...
##  $ W     : num [1:173] 28.3 26 25.6 21 29 25 26.2 24.9 25.7 27.5 ...
```

```
##  $ Wt    : num [1:173] 3.05 2.6 2.15 1.85 3 2.3 1.3 2.1 2 3.15 ...
##  $ Sa    : num [1:173] 8 4 0 0 1 3 0 0 8 6 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   number = col_double(),
##   ..   C = col_double(),
##   ..   S = col_double(),
##   ..   W = col_double(),
##   ..   Wt = col_double(),
##   ..   Sa = col_double()
##   .. )
```

The dataset contains 173 observations and 6 variables with each observation indicating the physiological condition and number of satellites of one female crab.

For the convenience of the model fitting, we let Y denote the number of satellites, and assume that $Y \sim Poisson(\lambda)$.

Fit the M1 model

```
m1 = crab_df |> glm(Sa~W, family=poisson(link=log), data=_)
summary(m1)
```

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

We obtained the poisson log link model M1: $log(\lambda) = -3.3 + 0.16 * W$

- $\beta_0 = -3.3$: The log rate number of female crab's satellites with 0 carapace width is -3.3.
- $\beta_1 = 0.16$: The log rate number of female crab's satellites will increases by $exp(1.164) = 1.17$ with each unit increment in carapace width.

Check goodness-of-fit

```
## deviance
D1 <- sum(residuals(m1, type = "deviance")^2)
D1
```

```
## [1] 567.8786
```

```
## Pearson chi-square
G1 <- sum(residuals(m1, type = "pearson")^2)
G1
```

```
## [1] 544.157
```

```
pvalue1 <- 1 - pchisq(D1, dim(crab_df)[1]-2)
pvalue1
```

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

```
## p1 = 1-pchisq(m1$deviance, df = nrow(crab_df) - 2)
```

- Since $D = 567.8786, df = 171, p - value = 0 < 0.05$, we rejected the model and concluded that the model doesn't fit data well.
- The effect of carapace width on number of satellites is significant.

(b)

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

Fit M2 Model

```
m2 = crab_df |> glm(Sa ~ W + Wt, family=poisson(link=log), data=_)
summary(m2)
```

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

We obtained the poisson log link model M2:

$log(\lambda) = -1.292 + 0.0459W + 0.447Wt.$

- $\beta_0 = -1.292$: The log rate of number of female crab's satellites with 0 carapace width is -1.292.
- $\beta_1 = 0.046$: The log rate of number of female crab's satellites will increases by $exp(0.046) = 1.05$ with each unit increment in carapace width, holding Weight(Wt) unchanged.
- $\beta_2 = 0.447$: The log rate of number of female crab's satellites will increases by $exp(0.447) = 1.56$ with each unit increment in carapace width, holding Width(W) unchanged.

Check goodness-of-fit of M1 and M2

```
# ## deviance
# D2 <- sum(residuals(m2, type = "deviance")^2)
# D2
# ## Pearson chi-square
# G2 <- sum(residuals(m2, type = "pearson")^2)
# G2
# pvalue2 <- 1 - pchisq(D2, dim(crab_df)[1]-2)
# pvalue2
p2_1 = 1 - pchisq(m1$deviance - m2$deviance, df = 1)
```

Since $D = D_1 - D_2 = 7.99, df = 1, p - value = 0.004 < 0.05$, we rejected the null hypothesis and concluded that the model doesn't fit data well.

(c)

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

Check goodness-of-fit of M2

```
p2 = 1 - pchisq(m2$deviance, df = nrow(crab_df) - 3)
p2
```

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

Since $D = 559.885, df = 170, p = 0 < 0.05$, we rejected the model M2 and concluded that M2 is also not a good fit for the data. Therefore, we suspected there exists over-dispersion:

Estimate over-dispersion parameter

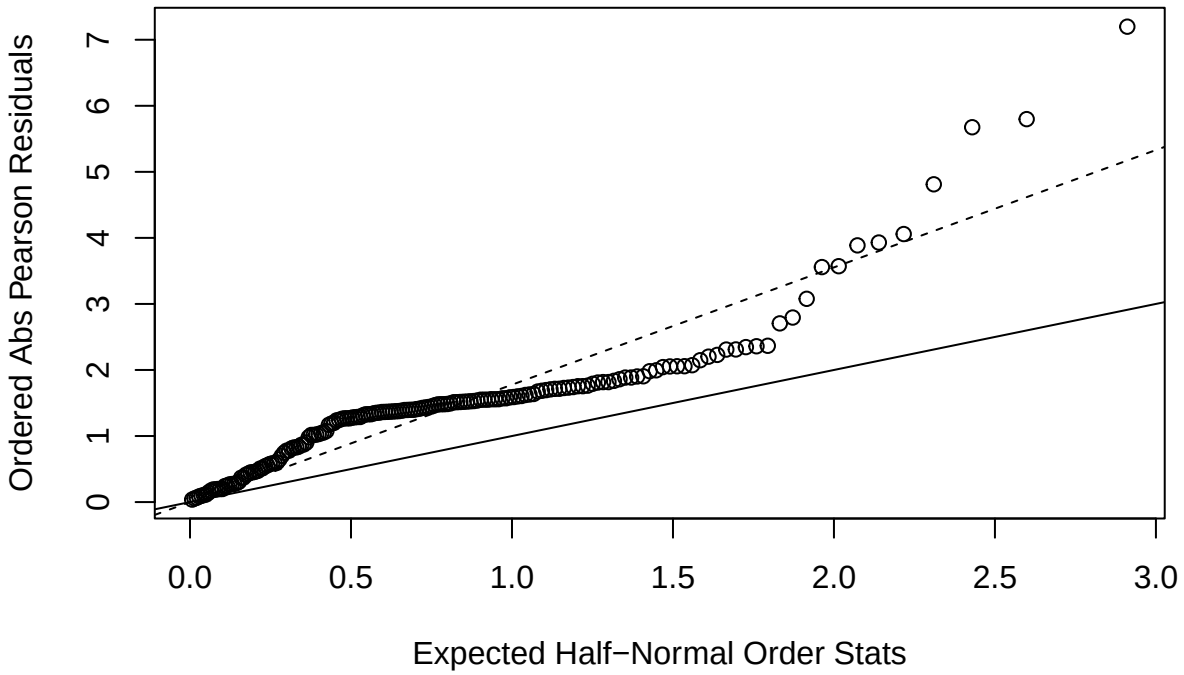
```
res.p1 = crab_df |> residuals(m2,type='pearson',data=_)
G1=sum(res.p1^2) # calc dispersion param based on full model
pval=1-pchisq(G1,df = dim(crab_df)[1]-3) # lack of fit
phi=G1/(dim(crab_df)[1]-3)
phi
```

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

Over dispersion parameter $\hat{\phi} = \frac{G_1}{n-p} = 3.156$.

Half-normal plot

```
plot(qnorm((dim(crab_df)[1]+1:dim(crab_df)[1]+0.5)/(2*dim(crab_df)[1]+1.125)),
     sort(abs(res.p1)),
     xlab='Expected Half-Normal Order Stats',
     ylab='Ordered Abs Pearson Residuals')
abline(a=0,b=1)
abline(a=0,b=sqrt(phi),lty=2)
```



The linear deviation from the reference line indicatess constant over-dispersion.

```
# fit model with constant over-dispersion
summary(m2, dispersion = phi)
```

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

We obtained the poisson log link model M2 (with dispersion parameter):

$log(E(y)) = -1.292 + 0.0459W + 0.447Wt.$

- $\beta_0 = -1.292$: The log rate ratio of female crab's satellites with 0 carapace width is -1.292.
- $\beta_1 = 0.046$: The log rate ratio of female crab's satellites will increase by $exp(0.046) = 1.05$ with each unit increment in carapace width, holding Weight(Wt) unchanged.
- $\beta_2 = 0.447$: The log rate ratio of female crab's satellites will increase by $exp(0.447) = 1.56$ with each unit increment in carapace width, holding Width(W) unchanged.

Checking goodness-of-fit of M2 with dispersion parameter

```
# deviance analysis
pval2 = 1 - pchisq(m2$deviance/phi, df = nrow(crab_df)-3)
pval2
```

```
## [1] 0.3334054
```

Since $p - value = 0.333 > 0.05$, we failed to reject the model and concluded that after considering over-dispersion, the model fits the data well.

Prevalence of Parasites

Researchers examined a large number of fish to determine the prevalence of parasites. The dataset (parasite.txt) includes the variables Intensity (i.e., the number of parasites), Area (a categorical variable), Year (to be treated as categorical), and Length of the fish.

(a)

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

```
paras_df = read_table("HW5-parasite.txt") |>
  janitor::clean_names() |>
  mutate(
  year = as.character(year),
  area = as.character(area)
)
```

```
## Warning: Duplicated column names deduplicated: 'omit' => 'omit_1' [5], 'omit'
## => 'omit_2' [6], 'omit' => 'omit_3' [8], 'omit' => 'omit_4' [9], 'omit' =>
## 'omit_5' [10]
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Sample = col_double(),
##   Intensity = col_double(),
##   omit = col_double(),
##   Year = col_double(),
##   omit_1 = col_double(),
##   omit_2 = col_double(),
##   Length = col_double(),
##   omit_3 = col_double(),
##   omit_4 = col_double(),
##   omit_5 = col_double(),
##   Area = col_double()
## )
```

```
paras_df |> head()
```

```
## # A tibble: 6 x 11
##   sample intensity  omit year  omit_1 omit_2 length omit_3 omit_4 omit_5 area
##    <dbl>     <dbl> <dbl> <chr>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl> <chr>
## 1      1         0     0 1999     220    148     26      0      0      0 2
## 2      2         0     0 1999     220    144     26      0      0      0 2
## 3      3         0     0 1999     220    146     27      0      0      0 2
```

```
## 4      4         0     0 1999      220     138      26       0       0       0 2
## 5      5         0     0 1999      220      40      17       0       0       0 2
## 6      6         0     0 1999      220      68      20       0       0       0 2
```

Fit the model

```
pois.fit = paras_df |> glm(intensity~area+year+length, family=poisson(link=log), data=_)
summary(pois.fit)
```

```
##
## Call:
## glm(formula = intensity ~ area + year + length, family = poisson(link = log),
##     data = paras_df)
##
## 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
##   (63 observations deleted due to missingness)
## AIC: 21089
##
## Number of Fisher Scoring iterations: 7
```

The poisson log link model is:

$M : log(E(Intensity)) = \beta_0 + \beta_1 I(Area = 2) + \beta_2(Area = 3) + \beta_3 I(Area = 4) + \beta_4 I(Year = 2000) + \beta_5 I(Year = 2001) + \beta_6 I(Length)$

We obtained the poisson log link model (without dispersion parameter):

$log(E(Intensity)) = 2.643 - 0.211x_1 - 0.116x_2 + 1.405x_3 + 0.670x_4 - 0.218x_5 - 0.028I(length)$.

The interpretation of the coefficients is as follows:

- $\beta_0 = 2.643$: The log rate ratio of parasite's intensity within Area = 1 (reference group), Year = 1999 (reference group), and Length = 0 is 2.643.
- $\beta_1 = -0.211$ indicates the relative log rate ratio of parasite's intensity for Area = 2 is $exp(-0.211) = 0.81$ times that for the parasite intensity in Area = 1, holding year and length unchanged.
- $\beta_2 = -0.116$ indicates the relative log rate ratio of parasite's intensity for Area = 3 is $exp(-0.116) = 0.89$ times that for the parasite intensity in Area = 1, holding year and length unchanged.
- $\beta_3 = 1.404$ indicates the relative log rate ratio of parasite's intensity for Area = 3 is $exp(1.404) = 4.07$ times that for the parasite intensity in Area = 1, holding year and length unchanged.

- $\beta_4 = 0.670$ indicates the relative log rate ratio of parasite's intensity for Year = 2000 is $exp(0.670) = 1.95$ times that for the parasite intensity in Year = 1999, holding area and length unchanged.
- $\beta_5 = -0.218$ indicates the relative log rate ratio of parasite's intensity for Year = 2001 is $exp(-0.218) = 0.80$ times that for the parasite intensity in Year = 1999, holding area and length unchanged.
- $\beta_6 = -0.028$ indicates the increment of relative log rate ratio of parasite's intensity is $exp(-0.028) = 0.97$ with every unit increase of length of fish, holding area and year unchanged.

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

```
# # goodness of fit
# # Deviance
# D3 <- sum(residuals(pois.fit, type = "deviance")^2)
# D3
# Pearson Chi-square statistics
G3 <- sum(residuals(pois.fit, type = "pearson")^2)
G3
```

```
## [1] 42164.97
```

```
# # model fitting test
# pvalue3 <- 1 - pchisq(deviance(pois.fit), df.residual(pois.fit))
# pvalue3
p = 1 - pchisq(pois.fit$deviance, df = pois.fit$df.residual)
p
```

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

- $Pearson - \chi^2 statistic : X^2 = \sum X_i^2 = 42164.97$ *and* $Deviance\ D = \sum d_i^2 = 1.91528 * 10^4$.
- Comparing X^2 and D with $\chi^2(1184)$, since $p - value = 0 < 0.05$, we rejected the model and concluded that the model isn't a good fit for the data.

(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-inflated 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.

Let Z_i be a latent binary variable that generates structural zeros: $P(Z_i = 0) = \pi_i$.

Then,

$$P(Y_i|Z_i = 0) = 0, P(Y_i|Z_i = 1) \sim Poisson(\lambda_i)$$

.

Here, we want to fit two models:

- Count Model:

  $$log(\lambda) = \beta_0 + \beta_1 I(Area = 2) + \beta_2 (Area = 3) + \beta_3 I(Area = 4) + \beta_4 I(Year = 2000) + \beta_5 I(Year = 2001) + \beta_6 I(Length).$$

- Binomial Model: $log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 I(Area = 2) + \beta_2 (Area = 3) + \beta_3 I(Area = 4) + \beta_4 I(Year = 2000) + \beta_5 I(Year = 2001) + \beta_6 I(Length)$.

```
zero.fit = paras_df |>
  zeroinfl(intensity ~ area + year + length, data = _)
summary(zero.fit)
```

```
##
## Call:
## zeroinfl(formula = intensity ~ area + year + length, data = paras_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)  3.8431714  0.0583793  65.831  < 2e-16 ***
## area2        0.2687835  0.0500467   5.371 7.85e-08 ***
## area3        0.1463173  0.0439485   3.329 0.000871 ***
## area4        0.9448068  0.0368342  25.650  < 2e-16 ***
## year2000     0.3919831  0.0282952  13.853  < 2e-16 ***
## year2001    -0.0448455  0.0296057  -1.515 0.129833
## 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)  0.552585   0.275762   2.004  0.04509 *
## area2        0.718676   0.189552   3.791  0.00015 ***
## area3        0.657708   0.167402   3.929 8.53e-05 ***
## area4       -1.022868   0.188201  -5.435 5.48e-08 ***
## year2000    -0.752119   0.172965  -4.348 1.37e-05 ***
## year2001     0.456535   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: 22
## Log-likelihood: -6950 on 14 Df
```

From the count model results, the interpretation of the coefficients is as follows:

- $\beta_0 = 3.843$: The log rate ratio of parasite's intensity within Area = 1 (reference group), Year = 1999 (reference group), and Length = 0 is 3.843.
- $\beta_1 = 0.268$ indicates the relative log rate ratio of parasite's intensity for Area = 2 is $exp(0.268) = 1.307$ times that for the parasite intensity in Area = 1, holding year and length unchanged.
- $\beta_2 = 0.146$ indicates the relative log rate ratio of parasite's intensity for Area = 3 is $exp(0.146) = 1.157$ times that for the parasite intensity in Area = 1, holding year and length unchanged.
- $\beta_3 = 0.944$ indicates the relative log rate ratio of parasite's intensity for Area = 3 is $exp(0.944) = 2.57$ times that for the parasite intensity in Area = 1, holding year and length unchanged.
- $\beta_4 = 0.392$ indicates the relative log rate ratio of parasite's intensity for Year = 2000 is $exp(0.392) = 1.479$ times that for the parasite intensity in Year = 1999, holding area and length unchanged.
- $\beta_5 = -0.045$ indicates the relative log rate ratio of parasite's intensity for Year = 2001 is $exp(-0.045) = 0.956$ times that for the parasite intensity in Year = 1999, holding area and length unchanged.
- $\beta_6 = -0.037$ indicates the increment of relative log rate ratio of parasite's intensity is $exp(-0.037) = 0.963$ with every unit increase of length of fish, holding area and year unchanged.

From the zero-inflation model results, the interpretation of the coefficients is as follows:

- $\beta_0 = 0.552$: The log rate ratio of being insusceptible to parasite within Area = 1 (reference group), Year = 1999 (reference group), and Length = 0 is 0.552.
- $\beta_1 = 0.268$ indicates the relative log rate ratio of being insusceptible to parasite for Area = 2 is $exp(0.268) = 1.307$ times that for the parasite intensity in Area = 1, holding year and length unchanged.
- $\beta_2 = 0.146$ indicates the relative log rate ratio of being insusceptible to parasite for Area = 3 is $exp(0.146) = 1.157$ times that for the parasite intensity in Area = 1, holding year and length unchanged.
- $\beta_3 = 0.944$ indicates the relative log rate ratio of being insusceptible to parasite for Area = 3 is $exp(0.944) = 2.57$ times that for the parasite intensity in Area = 1, holding year and length unchanged.
- $\beta_4 = 0.392$ indicates the relative log rate ratio of being insusceptible to parasite for Year = 2000 is $exp(0.392) = 1.479$ times that for the parasite intensity in Year = 1999, holding area and length unchanged.
- $\beta_5 = -0.045$ indicates the relative log rate ratio of being insusceptible to parasite for Year = 2001 is $exp(-0.045) = 0.956$ times that for the parasite intensity in Year = 1999, holding area and length unchanged.
- $\beta_6 = -0.037$ indicates the increment of relative log rate ratio of being insusceptible to parasite is $exp(-0.037) = 0.963$ with every unit increase of length of fish, holding area and year unchanged.