

# P8131 HW5

Brian Jo Hsuan Lee

3/15/2022

Load packages

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

## Problem 1: Crab Satellite Count

Import and tidy data

```
# txt file read in using read_delim(), separated grouped values, and corrected column types
crab_data =
  read_delim("HW5-crab.txt", delim = "\t") %>%
  mutate(
    number = str_trim(number, side = c("both"))
  ) %>%
  separate(number, c("number", "C", "S", "W", "Wt", "Sa"), sep = " +") %>%
  mutate(
    across(where(is.character), as.numeric)
  )
```

a) Fit a simple Poisson model, check the goodness of fit and interpret the model

```
# m1 model fit
crab_m1 = glm(Sa ~ W, family = poisson, data = crab_data)
summary(crab_m1)

##
## Call:
## glm(formula = Sa ~ W, family = poisson, data = crab_data)
##
## 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
# both deviance residual and pearson's residual goodness of fit tests,
# with df = 173 observations - 2 predictors = 171
crab_m1_deviance_pval = 1 - pchisq(crab_m1$deviance, 171)
crab_m1_pchisq = sum(residuals(crab_m1, 'pearson')^2)
crab_m1_pchisq_pval = 1 - pchisq(crab_m1_pchisq, 171)

ifelse(crab_m1_deviance_pval > 0.05 | crab_m1_pchisq_pval > 0.05,
       'Failed to reject the null, since no significant evidence suggest the poisson fit is not good',
       'Reject the null with significant data suggesting the poisson fit is not good')

## [1] "Reject the null with significant data suggesting the poisson fit is not good"
```

Fit M1 shows the log count of a female horseshoe crab's satellite increases by 0.164 per unit increase of its carapace width. The coefficient for carapace width is significant at p-value < 2e-16, despite the simple poisson model does not provide a good fit to the data.

- b) **Fit a Poisson model with 2 predictors, compare it with the previous model and interpret it**

```
# m2 model fit
crab_m2 = glm(Sa ~ W + Wt, family = poisson, data = crab_data)
summary(crab_m2)

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

Fit M2 shows the log count of a female horseshoe crab's satellite increases by 0.0459 per unit increase of its carapace width while adjusting for its weight, and it increases by 0.447 per unit increase of its weight, holding carapace width fixed. Only the weight predictor is significant.

```
# use chisq test and evaluate the nested models m1 and m2,
# with df = 171 m1 predictors - 170 m2 predictors = 1
m1_m2_stat = crab_m1$deviance - crab_m2$deviance
m1_m2_pval = 1 - pchisq(m1_m2_stat, df = 171-170)
ifelse(m1_m2_pval > 0.05,
      'Failed to reject the null, since no significant evidence suggest the larger model has a better
      'Reject the null with significant evidence suggesting the larger model fits the data better')
```

```
## [1] "Reject the null with significant evidence suggesting the larger model fits the data better"
```

Per the results of the deviance analysis, M2 has a significantly better fit and is preferred to M1. Likewise, M2 also has a lower AIC than M1, suggesting its better fit.

### c) Estimate overdispersion and interpret under the assumption of overdispersion

```
# obtain dispersion paramater using m3's pearson's chisq residual
# with df = 173 observations - 3 predictors = 170
crab_m2_pchisq = sum(residuals(crab_m2, 'pearson')^2)
phi = crab_m2_pchisq/170; phi
```

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

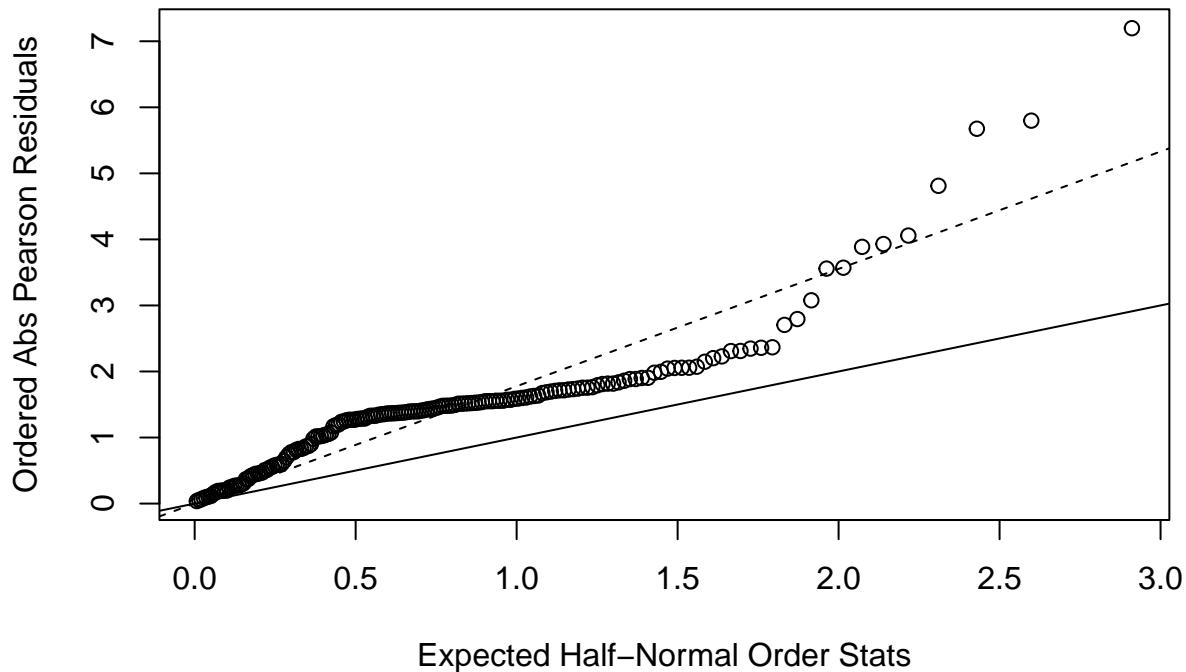
```
# the following code yields a similar phi estimate
## alt_phi = crab_m2$deviance/crab_m2$df.residual; alt_phi
```

For count models, overdispersion occurs when  $Var[X] > E[X] = \lambda$ , and it is shown with  $\phi = 3.16 > 1$ .

```
summary(crab_m2, dispersion = phi)
```

```
##
## Call:
## glm(formula = Sa ~ W + Wt, family = poisson, data = crab_data)
##
## 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
res = residuals(crab_m2, type='pearson')
plot(qnorm((173+1:173+0.5)/(2*173+1.125)),
     sort(abs(res)),
     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 estimated intercept and coefficients for the new fit are equivalent to those of the original fit, as shown in the summaries, and therefore have the same interpretation as for the original fit. The half-normal plot visualizes  $\phi > 1$  by displaying the data deviance (linear regression as the dashed line) from the solid  $x = y$  reference line in a  $\phi^{-1}(\frac{n+i+0.5}{2n+1.125})$  vs  $|r_{(i)}|$  plot, where  $r$  is Pearson's residuals. It shows that the dispersion is not constant and the dashed line does not truly capture the trend.

## Problem 2: Fish Parasite Count

```
# txt file read in using read_delim() and dropped 'NA' rows and o 'omit' columns
para_data =
  read_delim("HW5-parasite.txt", delim = "\t") %>%
  dplyr::select(c('Intensity', 'Year', 'Length', 'Area')) %>%
  mutate(
    Year = factor(Year),
    Area = factor(Area)
  ) %>%
  drop_na()
```

a) Fit a simple Poisson model, check the goodness of fit and interpret the model

```
# m1 model fit
para_m1 = glm(Intensity ~ Year + Area + Length, family = poisson, data = para_data)
summary(para_m1)

##
## Call:
## glm(formula = Intensity ~ Year + Area + Length, family = poisson,
##      data = para_data)
##
## 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 ***
## Year2000      0.6702801  0.0279823  23.954 < 2e-16 ***
## Year2001     -0.2181393  0.0287535  -7.587 3.29e-14 ***
## 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 ***
## ---
## 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
```

The fit shows the log count of parasites is 2.64 in year 1999, in area 1 and at fish length 0. The count increases by  $e^{0.670} = 1.95$  times of year 1999 in 2000, but is  $e^{-0.218} = 0.804$  times of year 1999 in 2001 while adjusting for area and body length; the count response is  $e^{-0.212} = .809$  and  $e^{-0.117} = 0.890$  times of area 1 in area 2 and area 3, respectively, and  $e^{-0.140} = 4.06$  times area 1 in area 4 while adjusting for year and body length; the count increases by  $1 - e^{0.0284} = 0.03$  in percentage per unit increase in length while adjusting for year and area. The predictors, including the intercept and all coefficients, are significant at  $\alpha = 0.05$ .

#### b) Goodness of fit and conclusions

```
# both deviance residual and pearson's residual goodness of fit tests,
# with df = 1191 observations - 7 predictors = 1184
para_m1_deviance_pval = 1 - pchisq(para_m1$deviance, 1184)
para_m1_pchisq = sum(residuals(para_m1, 'pearson')^2)
para_m1_pchisq_pval = 1 - pchisq(para_m1_pchisq, 1184)

ifelse(para_m1_deviance_pval > 0.05 | para_m1_pchisq_pval > 0.05,
       'Failed to reject the null, since no significant evidence suggest the poisson fit is not good',
       'Reject the null with significant data suggesting the poisson fit is not good')
```

```
## [1] "Reject the null with significant data suggesting the poisson fit is not good"
```

Though the coefficients are significant, it is a poor fitting model for the data. We may speculate the issue to be that the data actually falls in a zero-inflated, zero-truncated, or multi-modal poisson distribution, and question if the overdispersion paramter equals 1.

#### c) Fit a zero-inflated poisson model and interpret it

```
# m2 model fit
para_m2 = zeroinfl(Intensity ~ Year + Length | Area, data = para_data)
summary(para_m2)

##
## Call:
## zeroinfl(formula = Intensity ~ Year + Length | Area, data = para_data)
```

```

##
## 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

```

Assuming fish in different areas may exhibit a difference in their susceptibility to parasites, and the parasite intensity depends on year and body length, fit M2 is hence created. The model first evaluates whether a fish could be susceptible to parasite infection under a binomial distribution depending on the area, and then models the parasite count under a poisson distribution. Contextually speaking, in areas 2, 3, and 4, the respective odds of a fish is susceptible to parasite infection is  $e^{0.747} = 2.11$ ,  $e^{0.681} = 1.96$ , and  $e^{-0.883} = 0.414$  times that of fish in area 1. If a fish is susceptible to infection in the eyes of the model, the parasite count on a fish in year 2000 and 2001 is  $e^{0.421} = 1.52$  and  $e^{0.099} = 1.10$  times a fish in year 1999, respectively, holding body length fixed. When adjusting for the year, a fish with an increased unit body length has  $1 - e^{-0.0439} = 0.957$  times the parasite count.