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. However, 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 weight, and increases by 0.447 per unit increase of its weight while adjusting for carapace width. Only the coefficient for weight is significant.

```r
# 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 f
       '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"
```

M2 has a significantly better fit and is preferred to M1

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

```r
# 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
```

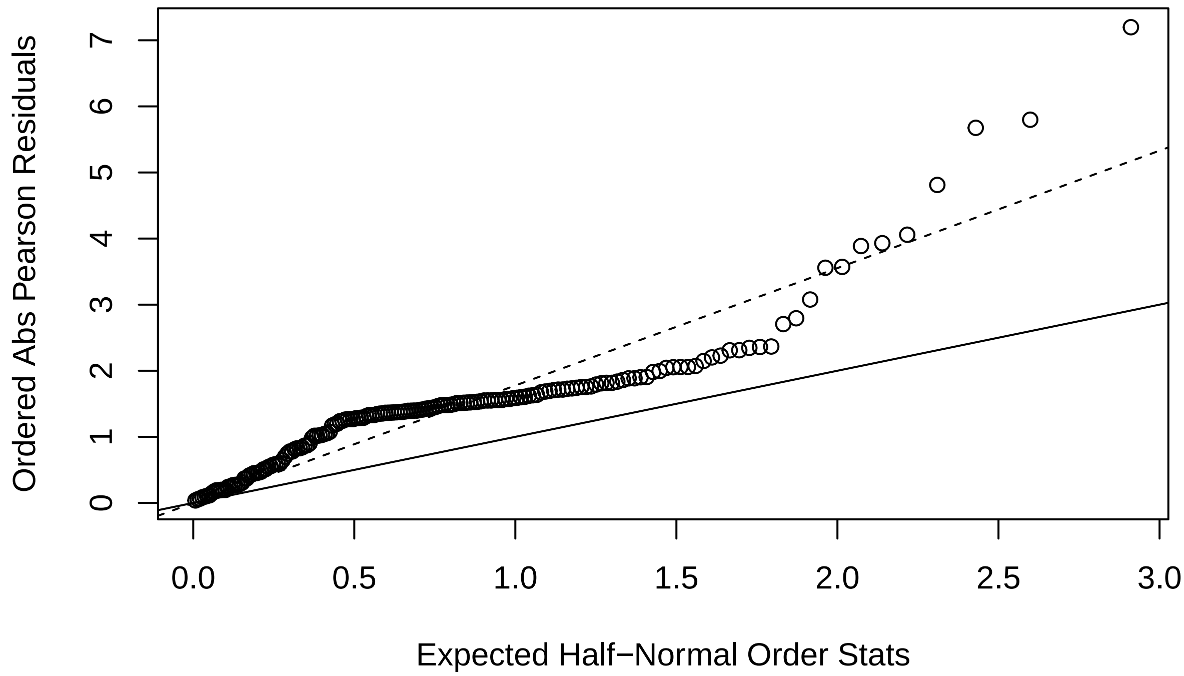
```r
# the following code yields a similar phi estimate
## alt_phi = crab_m2$deviance/crab_m2$df.residual; alt_phi

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
```

Estimated betas don't change

```r
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)
```



Problem 2:

```r
# txt file read in using read_delim() and dropped 'NA' rows and o 'omit' columns
para_data =
  read_delim("HW5-parasite.txt", delim = "\t") %>%
  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**

```r
# 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 length 0. The response increases by 0.670 in 2000, but decreases by 0.218 in 2001 when compared to year 1999 while adjusting for areas and fish body length; the response decreases by 0.212 and 0.117 in area 2 and area 3, respectively, and increases by 1.40 in area 4 when compared to area 1, while adjusting for year and fish body length; the response decreases by 0.0284 per unit increase in length while adjusting for year and area. The intercept and all coefficients are significant at $\alpha = 0.05$.

b) **Goodness of fit and conclusions**

```r
# both deviance residual and pearson's residual goodness of fit tests,
# with df = 1191 observations - 4 predictors = 1187
para_m1_deviance_pval = 1 - pchisq(para_m1$deviance, 1187)
para_m1_pchisq = sum(residuals(para_m1, 'pearson')^2)
para_m1_pchisq_pval = 1 - pchisq(para_m1_pchisq, 1187)

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"
```

Despite the coefficients are significant, the model does not provide a good fit to the data. We may speculate the issue be that the data actually falls in a zero-inflated, zero-truncated, or multi-modal poisson distribution.

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

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

```
##
## Call:
## zeroinfl(formula = Intensity ~ Year + Area | Length, data = para_data)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.9095 -0.8288 -0.7946 -0.2150 29.5318
##
## Count model coefficients (poisson with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.92502    0.03353  57.416  < 2e-16 ***
```

```
## Year2000     0.23180    0.02788   8.316  < 2e-16 ***
## Year2001    -0.22788    0.02940  -7.751 9.14e-15 ***
## Area2        0.31024    0.04998   6.207 5.39e-10 ***
## Area3        0.28528    0.04353   6.553 5.64e-11 ***
## Area4        1.24006    0.03631  34.153  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.298749   0.227678   1.312    0.189
## Length      -0.002113   0.004109  -0.514    0.607
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 11
## Log-likelihood: -7781 on 8 Df
```