P8131_hw5

Shan Jiang

Problem 1 nesting crabs

a). fit a poisson model M1 with logit link W.

```
# import in the crab data
crab = read.table('./HW5-crab.txt', header = T)
# test if there is obs = 0 for W and Wt, no 0 exists for model
nrow(crab)
```

```
## [1] 173
```

```
nrow(subset(crab, crab$W > 0))
```

```
## [1] 173
```

```
crab.M1 <- glm(crab$Sa~W, family = poisson(link = log), data = crab)
summary(crab.M1)
```

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

```
# n = 173, beta0 =-3.30476, beta1 = 0.16405
anova(crab.M1)
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: crab$Sa
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev
## NULL                      172     632.79
## W     1   64.913          171     567.88
```

The response outcome for each female crab is her number of satellites (Sa), with predictor as carapace width (W), estimated model is as following:

$$log(\hat{Sa_i}) = 0.16405 * W_i - 3.30476$$

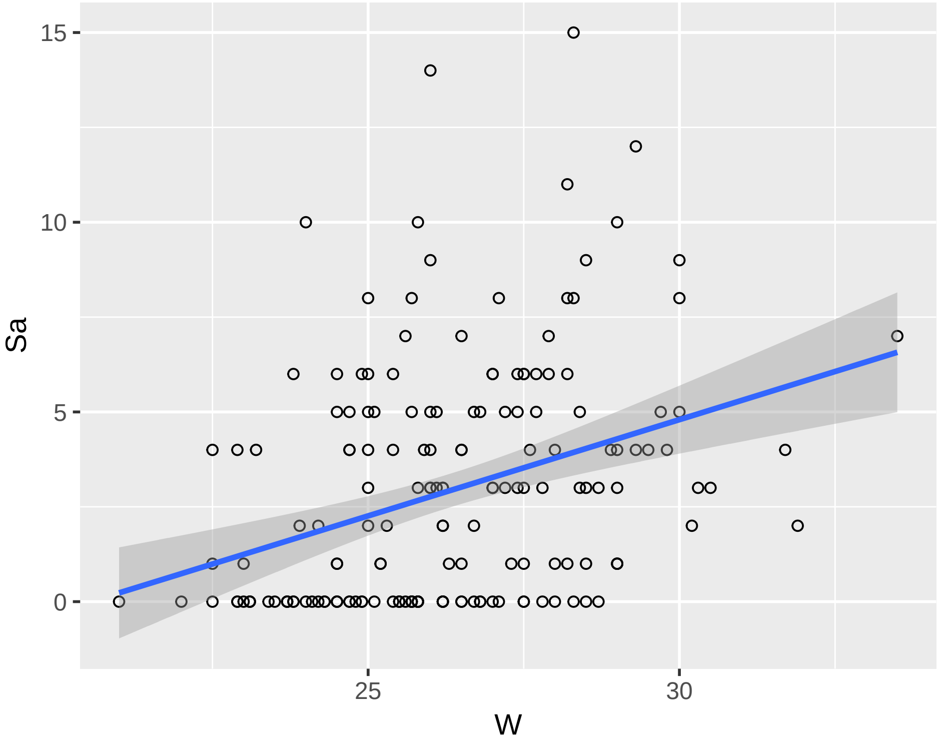
The Std.Error of estimated β_1 = 0.164 is 0.01997 which is small, and the slope is statistically significant given its z-value of 8.216 and its low p-value.

Interpretation: Since estimate of $\beta_1 > 0$, for one unit of increase in the female crab width, the expected number of male satellites will increase, and will be multiplied by e^{β_1} = 1.18 times.

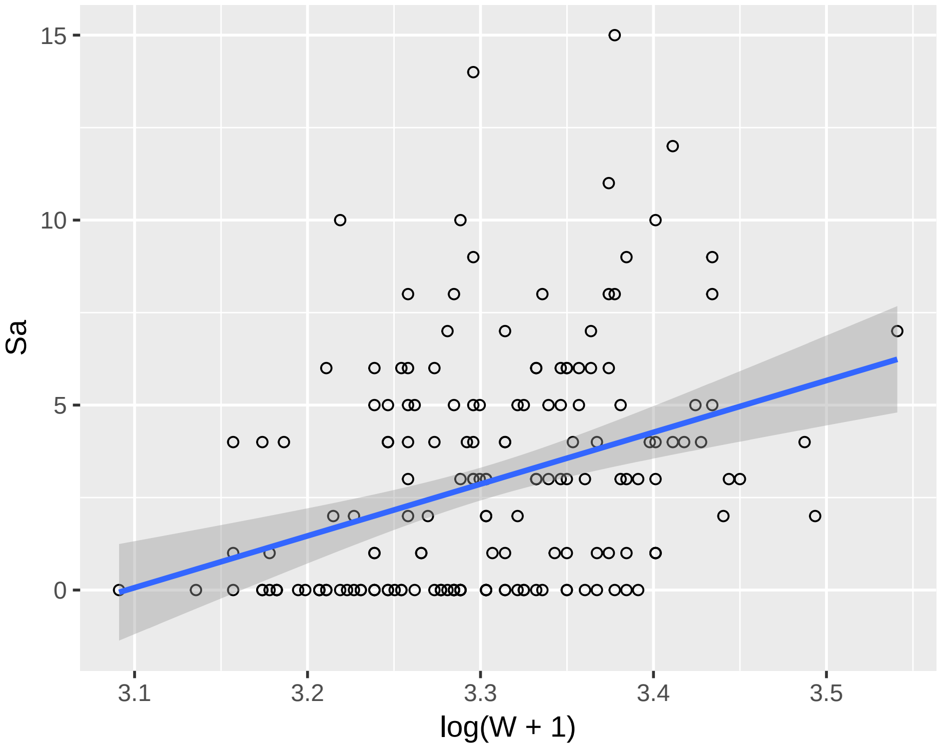
Goodness of fit

(1) Graphs test.

```
## we may suspect some outliers from the simple Y-X scatter plot;
ggplot(data = crab, aes(x = W, y = Sa )) +
  geom_point(shape = 1)  +
  geom_smooth(method = glm)
```



```
## fit the log model:
ggplot(data = crab, aes(x = log(W + 1), y = Sa )) +
  geom_point(shape = 1)  +
  geom_smooth(method = glm)
```



From two graphs, it looks like there is no significant difference for the log model compared with the simple linear regression model, we may consider more about the outliers aligned in the bottom part and the top grid of the panel.

(2) Model diagnostic: Goodness of Fit

```
## Methods 1: Genearalized Pearson Chi-square
res.p1 = residuals(crab.M1, type = 'pearson',data = crab)
G1 = sum(res.p1^2) # calc dispersion param based on full model
## [1] 544.157

df = 173 - 2
1 - pchisq(G1, df = df) ##
```

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

```
## [2] df = 171, p-value = 0


## Alternative Methods: Deviance Analysis
D = deviance(crab.M1)
## [1] 567.8786
1 - pchisq(D, df = df)
```

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

```
## [2] p-value = 0, means
```

The Pearson's chi-squared and deviance both are around 540 and 560, and maintained a p-value of $0 < 0.05$, meaning this model doesn't fit well, we need an adjustment.

b). fit a poisson model M2

```
## There is offset in this dataset
crab.M2 <- glm(Sa ~ Wt + W, family = poisson(link = log), data = crab)
summary(crab.M2)
```

```
## 
## Call:
## glm(formula = Sa ~ Wt + W, family = poisson(link = log), data = crab)
## 
## 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   
## Wt           0.44744    0.15864   2.820  0.00479 **
## W            0.04590    0.04677   0.981  0.32640   
## ---
## 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
```

After adding a new predictor Wt, we have a larger model: The response outcome for each female crab is her number of satellites (Sa), with predictors as carapace width (W), and weight(Wt) in estimated model is as following:

$$log(\hat{S}a_i) = 0.44744W_t + 0.04590 * W - 1.29168$$

Interpretation: * Since estimate of β_1> 0, for one unit of increase in the female crab weight, the expected number of male satellites will increase, and will be multiplied by $e^{\beta_1} = 1.5643$ times.

- Although estimate of β_2 is also positive, the coefficient didn't past the Z-test which indicates that there is no significant influence of female crab width on the expected number of male satellites.

Model comparison:

Since the model M1 is nested in model M2, we can have a deviance analysis for comparing these two models.

```
D.stat = crab.M1$deviance - crab.M2$deviance
df.test =  171 - 170
pval.test = 1 - pchisq(D.stat, df = df.test) # chisq test
pval.test
```

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

df for M1: 173 - 2 = 171 df for M2: 173 - 3 = 170

Compared with model M1, the coefficient for female crab width became smaller and not significant at all(p-value = 0.32), which alters the original conclusion, while the weight of male crab significantly affects the number of satellites in the M2 model. Because the p-value is $0.00469 < 0.05$, so we need to reject the null hypothesis, and say that the M2 model is a better fit than M1.

c). Check the over-dispersion.

(1) estimate the dispersion parameter

```
# the traditional way of calc constant dispersion parameter
res.p2 = residuals(crab.M2, type = 'pearson', data = crab)
G1 = sum(res.p2^2)
# Generalized-chi-square: 536.5963

## Estimate the dispersion parameter
phi = G1/(173 - 3)
phi # 3.15644
```

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

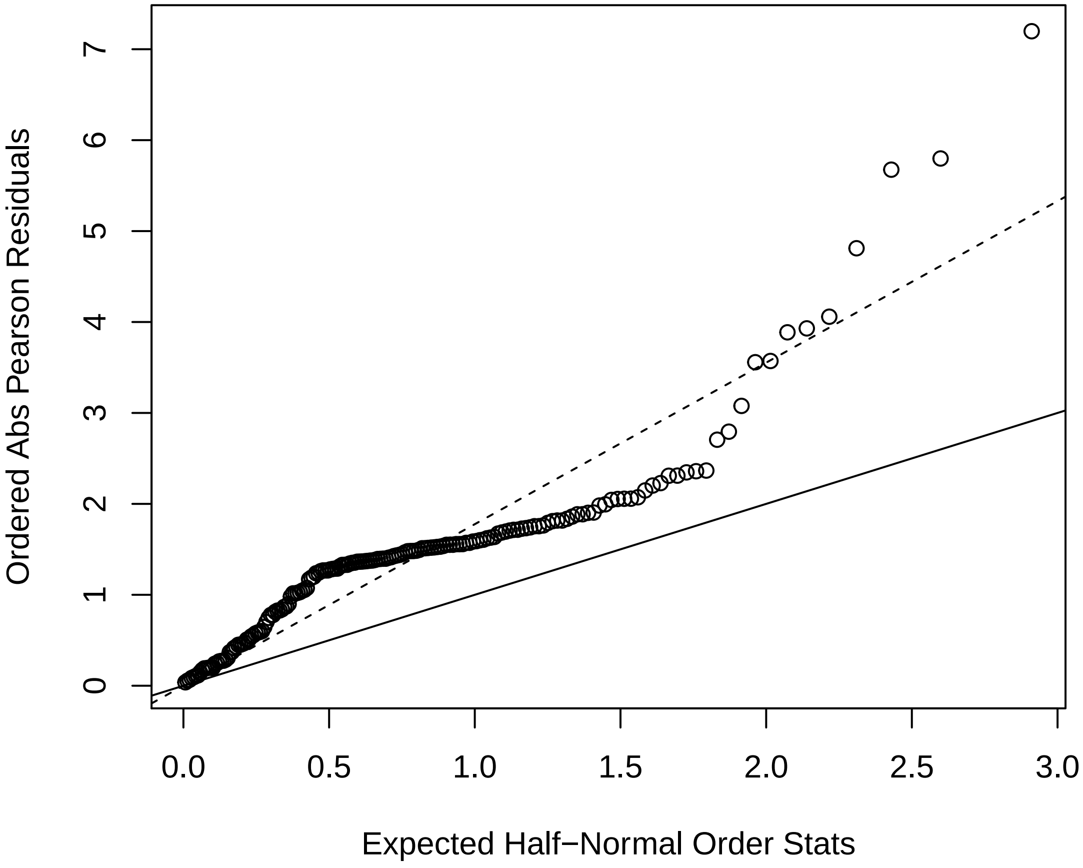
```
crab.M2$deviance/crab.M2$df.residual
```

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

```
# 3.293442
```

(2) Half-normal Plot

```
plot(qnorm((173+ 1:173+0.5)/(2*173+ 1.125)),sort(abs(res.p2)),
     xlab= 'Expected Half-Normal Order Stats', ylab= 'Ordered Abs Pearson Residuals')
abline(a=0,b=1)
abline(a=0,b=sqrt(phi),lty=2)
```



Conclusion: * This one captured a few points at the end, most of points fall around the reference line. * As the linear deviation revealed that the overdispersion that leads to lack of fit.

(3) Change the Model: Adjusting for Overdispersion

```
## Adjusting for Overdispersion: phi parameter
dispersion.M = summary(crab.M2, dispersion = phi)
```

The new model is here after adjusted for over-dispersion:

$$log(\hat{Sa_i}) = 0.44744W_t + 0.04590 * W - 1.29168$$

Interpretation: * The coefficients stay the same as before the adjustment for over-dispersion, while for the Std. Error it changes after adding up the phi parameter. * Since estimate of $\beta_1 = 0.44744 > 0$, for one unit of increase in the female crab weight, the expected number of male satellites will increase, and will be multiplied by $e^{0.44744}$ = 1.5643 times.

- Although estimate of $\beta_2 = 0.045$ is also positive, the coefficient didn't past the Z-test which indicates that there is no significant influence of female crab width on the expected number of male satellites.

Goodness of fit: Overdispersion

```
## Methods 1: Genearalized Pearson Chi-square
res.p1 = residuals(crab.M2, type = 'pearson',data = crab)
G2 = sum(res.p1^2) # calc dispersion param based on full model
## [1] 536.5963

df2 = 173 - 3
1 - pchisq(G2, df = df2) ##
```

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

```
## df = 170, p-value = 0
```

The Generalized chi-square test results show us that the p-value is $0 < 0.05$, we can reject the null-hypothesis, meaning that there is overdispersion in the model.

Problem 2

a) Fit a poisson model

```
library(MASS)
parasite_raw = read.table('./HW5-parasite.txt', header = T)
parasite_raw = parasite_raw %>%
  mutate(Area = as.factor(Area),
         Year = as.factor(Year) )

##test missing values in N.
colSums(is.na(parasite_raw))
```

```
##    Sample Intensity      omit      Year    omit.1    omit.2    Length
##         0        57         0         0         0         6         6
##    omit.3    omit.4    omit.5      Area
##         0         0         0         0
```

```
parasite_raw = na.omit(parasite_raw)

ps.glm = glm(Intensity ~ Length + Area + Year, data = parasite_raw, family = poisson(link = log))
summary(ps.glm)
```

```
##
## Call:
## glm(formula = Intensity ~ Length + Area + Year, family = poisson(link = log),
##     data = parasite_raw)
##
## 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 ***
## Length      -0.0284228  0.0008809 -32.265  < 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 ***
## ---
## 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
```

$$log(Par\hat{a}sites) = 2.6432 - 0.0284 * Length - 0.21195 * Area2 - 0.1169 * Area3 + 1.4049 * Area4 + 0.6703 * Year2000 - 0.2181 * Year2001$$

Interpretation: As all coefficients have significant p-values with them, we can have interpretations as,

- Intercept: For fish being in Year 1999 and Area 1 with zero Length, the expected number of parasites is $\exp(2.6432) = 14.057$.
- For a fish in certain Area and year, with one unit increase of **Length** of fish, the expected number of parasites will significantly decrease, and will be multiplied by $e^{\beta_1} = 0.9719$ times.
- Being in **Area2** significantly decreases the expected number of parasites for fish when holding other variables constant, and will be multiplied by $e^{\beta_2} = 0.8090$ times.
- Being in **Area3** significantly decreases the expected number of parasites for fish when holding other variables constant, and will be multiplied by $e^{\beta_3} = 0.8897$ times.
- Being in **Area4** significantly increases the expected number of parasites for fish when holding other variables constant, and will be multiplied by $e^{\beta_4} = 4.075$ times.
- Being in **Year2000** significantly increases the expected number of parasites for fish when holding other variables constant, and will be multiplied by $e^{\beta_5} = 1.9547$ times.
- Being in **Year2001** decreases the expected number of parasites for fish when holding other variables constant, and will be multiplied by $e^{\beta_6} = 0.8040$ times.

b). Goodness of fit

```r
## Genearalized Pearson Chi-square
res.p3 = residuals(ps.glm, type = 'pearson')
G3 = sum(res.p3^2) # calc dispersion param based on full model
G3
```

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

```r
df3 = (1254 - 63) - (4 - 1) - (3 - 1) - 1  ## (n - missing) - p

1 - pchisq(G3, df = df3)
```

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

```r
## Alternative Methods: Deviance
D3 = deviance(ps.glm)

1 - pchisq(D3, df = df3)
```

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

Conclusions:

Based on the generalized pearson chi-square statistics and Deviance methods, we find that we cannot reject the null hypothesis, meaning that this model fits well.

c). 0-inflation model

Thus, the zip model has two parts, a poisson count model and the logit model for predicting excess zeros, both contained three predictors for modelling.

```r
## Test zeros in dataset.
nrow(subset(parasite_raw, parasite_raw$Intensity == 0))
```

```
## [1] 651
```

```r
m1 <- zeroinfl( Intensity ~  Length + Area + Year, data = parasite_raw)
# child and camper for poisson, persons for binary
summary(m1)
```

```
## 
## Call:
## zeroinfl(formula = Intensity ~ Length + Area + Year, data = parasite_raw)
## 
## 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 ***
## Length      -0.0368067  0.0009747 -37.762  < 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    
## 
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  0.552585   0.275762   2.004  0.04509 *  
## Length      -0.009889   0.004629  -2.136  0.03266 *  
## 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 ** 
## ---
## 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
```

Model processes

In these 2 processes both give zeros: * True zero: fish is not susceptible to parasites, captured in the binary model; * psedo zero: fish is susceptible to parasites but have no parasite, captured in the poisson model;

Interpretation of *zero-inflation model* coefficients:

- Being in **Area2** increases the odds of being susceptible to parasites by 105% (exp(0.7186)= 2.0517), and this is statistically significant ($p = 0.00015 < 0.05$);
- Being in **Area3** increases the odds of being susceptible to parasites by 93% (exp(0.6577) = 1.9303), and this is statistically significant ($p = 8.53e\text{-}05 < 0.05$);
- Being in **Area4** decreases the odds of being susceptible to parasites by 64.04% (exp(-1.0229) = 0.3596, 1 - 0.3596 = 0.6404), and this is statistically significant ($p = 5.48e\text{-}08 < 0.05$);
- Being in **Year2000** decreases the odds of having the opportunity of susceptible to parasites by 52.86% (exp(-0.7521)= 0.4714, 1 - 0.4714 = 0.5286), and this is statistically significant ($p = 1.37e\text{-}05 < 0.05$);
- Being in **Year2001** increases the odds of having the opportunity of susceptible to parasites by 57.86% (exp(0.4565) = 1.5786), and this is statistically significant ($p = 0.00152 < 0.05$);
- With one unit increase of fish **length**, the odds of being susceptible strains to parasites decreases by 0.98% (exp(-0.009889) = 0.9901, 1-0.9901 = 0.0098).

Interpretation of *poisson model* coefficients:

- Among the fish species who are susceptible to parasites, The expected rate of parasites intensity the fish had for one unit increase of **length** is 0.9638 (exp(-0.0368067) = 0.9638), holding other variables constant, and this is statistically significant ($p<.0001$);
- Among the fish species who are susceptible to parasites, being in **Area2** increases the expected rate of parasites intensity by 30.85% (exp(0.2689) = 1.3085), holding other variables constant, and this is statistically significant ($p<.0001$).
- Among the fish species who are susceptible to parasites, being in **Area3** increases the expected rate of parasites intensity by 93.03% (exp(0.1463) = 1.9303), holding other variables constant, and this is statistically significant ($p<.0001$);
- Among the fish species who are susceptible to parasites, being in **Area4** significantly increases the expected rate of parasites by 157% (exp(0.9448) = 2.5722);
- Among the fish species who are susceptible to parasites, being in **year 2000** increases the expected rate of parasites by 47.98% (exp(0.3919) = 1.4798);
- Among the fish species who are susceptible to parasites, being in **year 2001** decreases the expected rate of parasites intensity by 4.38% (exp(-0.0448) = 0.9562, 1- 0.9562 = 0.0438).