

Homework 5

Yuki Joyama

2024-03-18

```
# data prep
df_crab = read.table("/Users/yukijoyama/Library/CloudStorage/GoogleDrive-jikeyu1995@gmail.com/My Drive/")
df_para = read.table("/Users/yukijoyama/Library/CloudStorage/GoogleDrive-jikeyu1995@gmail.com/My Drive/
```

1

(a)

I will fit a Poisson model (M1) with log link with carapace width (W) as the single predictor.

```
# M1: Poisson model with log link
m1_fit <- glm(Sa ~ W, family = poisson(link = "log"), data = df_crab)
summary(m1_fit)

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

```
exp(m1_fit$coefficients)
```

```
## (Intercept)          W
## 0.03670812  1.17826744
```

The expected number of satellites (Sa) given carapace width (W) = 0 is 0.037. With every unit increase in W, the expected number of Sa has multiplicative effect of 1.178 on $\mu = E(Y)$.

```
# Goodness of fit
G = sum(residuals(m1_fit, type = "pearson") ^ 2)
G
```

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

```
1 - pchisq(G, m1_fit$df.residual)
```

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

Thus, we reject the null, which indicates that the model does not have a good fit.

(b)

Now, I will fit a Poisson model (M2) with log link with carapace width (W) and weight (Wt) as predictors.

```
# M2: W and Wt as predictors
m2_fit <- glm(Sa ~ W + Wt, family = poisson(link = "log"), data = df_crab)
summary(m2_fit)
```

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

```
exp(m2_fit$coefficients)
```

```
## (Intercept)          W          Wt
##    0.274809    1.046968    1.564296
```

The expected number of Sa given W = 0 and Wt = 0 is 0.275. All else being equal, with every unit increase in W, the expected number of Sa has multiplicative effect of 1.047 on $\mu = E(Y)$. Similarly, holding other

variables constant, with every unit increase in Wt, the expected number of Sa has multiplicative effect of 1.564 on $\mu = E(Y)$.

M1 is nested within M2, so I will perform ANOVA to compare the two models.

```
# compare with M1
anova(m1_fit, m2_fit)
```

```
## Analysis of Deviance Table
##
## Model 1: Sa ~ W
## Model 2: Sa ~ W + Wt
##   Resid. Df Resid. Dev Df Deviance
## 1         171      567.88
## 2         170      559.89  1    7.9934
```

The residual deviance appears to be reduced in M2 compared to M1, indicating that M2 has a better fit by adding Wt as a predictor. We need to note that the coefficient in W in M2 is not statistically significant unlike M1.

(c)

```
# over-dispersion in M2
# calculate dispersion parameter
G.stat = sum(residuals(m2_fit, type = 'pearson', data = df_crab) ^ 2) # pearson chisq
G.stat
```

```
## [1] 536.5963
```

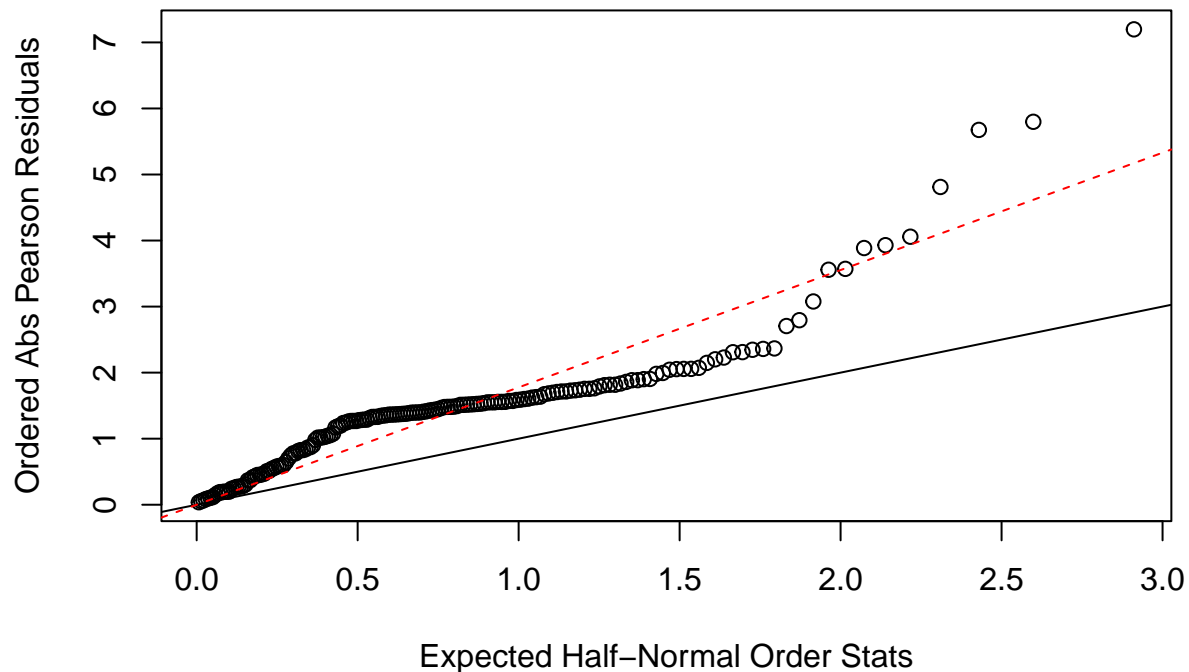
```
phi = G.stat / m2_fit$df.residual
phi
```

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

```
tilde.phi = m2_fit$deviance / m2_fit$df.residual
tilde.phi
```

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

```
# test over-dispersion (half normal plot)
res = residuals(m2_fit, type = 'pearson')
plot(qnorm((173 + 1: 173 + 0.5)/(2 * 173 + 1.125)), sort(abs(res)), xlab = 'Expected Half-Normal Order Statistic',
     abline(a = 0, b = 1)
     abline(a = 0, b = sqrt(phi), lty = 2, col = 'red'))
```



There is a linear deviation from the reference line in the half normal plot, suggesting that the response variance of the data exceeds the μ assumed by the model.

Hence, we can say that there is over-dispersion in the original model.

The estimate of dispersion parameter: $\hat{\phi} = 3.16$

```
# adjust for over-dispersion
summary(m2_fit, dispersion = phi)
```

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

After adjusting for over-dispersion, The coefficient values remain the same. However, the standard errors of each variable differ from the unadjusted model. Now, all coefficients are not statistically significant.

2

a

I will fit a Poisson model with log link with area, year, and leangth of the fish as predictors. Area and Year are treated as categorical variables.

```
# Poisson model with log link
poi_fit <- glm(Intensity ~ factor(Area) + factor(Year) + Length, family = poisson(link = "log"), data =
summary(poi_fit)
```

```
##
## Call:
## glm(formula = Intensity ~ factor(Area) + factor(Year) + Length,
##      family = poisson(link = "log"), data = df_para)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.6431709   0.0542838  48.692 < 2e-16 ***
## factor(Area)2   -0.2119557   0.0491691  -4.311 1.63e-05 ***
## factor(Area)3   -0.1168602   0.0428296  -2.728 0.00636 **
## factor(Area)4    1.4049366   0.0356625  39.395 < 2e-16 ***
## factor(Year)2000 0.6702801   0.0279823  23.954 < 2e-16 ***
## factor(Year)2001 -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
```

```
exp(poi_fit$coefficients)
```

```
##      (Intercept)      factor(Area)2      factor(Area)3      factor(Area)4
##      14.0577088          0.8090006          0.8897096          4.0752685
## factor(Year)2000 factor(Year)2001          Length
##      1.9547848          0.8040134          0.9719773
```

The expected number of parasites given Area = 1, Year = 1999, and Length = 0 is 14.058. All else being equal, the expected number of parasites has multiplicative effect of 0.809 on $\mu = E(Y)$ when Area = 2; 0.89 when Area = 3; 4.075 when Area = 4. Similarly, holding other variables constant, the expected number of parasites has multiplicative effect of 1.955 on $\mu = E(Y)$ when Year = 2000; 0.804 when Year = 2001. Finally, with every unit increase in Length, the expected number of parasites has multiplicative effect of 0.972 on $\mu = E(Y)$ with other predictors unchanged.

b

```
# Goodness of fit
G = sum(residuals(poi_fit, type = "pearson") ^ 2)
G
```

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

```
1 - pchisq(G, poi_fit$df.residual)
```

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

Given the chi-squared goodness of fit statistic and its p-value, we reject the null and conclude that the model does not have a good fit.

c

```
# check zero-inflation
check_zeroinflation(poi_fit)
```

```
## # Check for zero-inflation
##
##      Observed zeros: 651
##      Predicted zeros: 84
##              Ratio: 0.13
```

I will refit the model using the same predictors accounting for the zero-inflation issue.

```
# fit zero-inflated poisson model
zip_fit <- zeroinfl(Intensity ~ factor(Area) + factor(Year) + Length, data = df_para) # child and campe
summary(zip_fit)
```

```
##
## Call:
## zeroinfl(formula = Intensity ~ factor(Area) + factor(Year) + Length,
##          data = df_para)
##
## 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 ***
## factor(Area)2    0.2687835   0.0500467    5.371 7.85e-08 ***
## factor(Area)3    0.1463173   0.0439485    3.329 0.000871 ***
## factor(Area)4    0.9448068   0.0368342   25.650 < 2e-16 ***
## factor(Year)2000 0.3919831   0.0282952   13.853 < 2e-16 ***
```

```

## factor(Year)2001 -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 *
## factor(Area)2  0.718676   0.189552   3.791  0.00015 ***
## factor(Area)3  0.657708   0.167402   3.929 8.53e-05 ***
## factor(Area)4 -1.022868   0.188201  -5.435 5.48e-08 ***
## factor(Year)2000 -0.752119   0.172965  -4.348 1.37e-05 ***
## factor(Year)2001  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

```

All of the predictors in both the count and zero-inflation model are statistically significant except Year = 2001 in count model.

The zero-inflation model tells us that the fish in Area 2, 3 (versus Area 1), and fish in 2001 (versus 1999) are more susceptible to parasites. The count model indicates that for fish with more than one parasites, Area 2, 3, 4 (versus Area 1), and 2001 (versus 1999) tend to have higher number of parasites.