p8131_hw5_mp3745

Matthew Parker

2020-03-09

Problem 1

(a)

Read in the crab data

```
crab_data = read.table("./data/HW5-crab.txt", header = TRUE)
```

Fit a Poisson model (m1) with log link with W as the single predictor

```
m1 = glm(Sa ~ W,
         family = poisson,
         data = crab_data)

# View summary of m1
summary(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
```

```
# Check goodness of fit of M1
m1_dev = sum(residuals(m1, type = 'deviance')^2)

m1_pval = 1 - pchisq(m1_dev, (nrow(crab_data) - 2))
```

Since p- value $= 0 < 0.05$, we reject the null hypothesis of the model being a good fit for the data. The log ratio of the number of satellites for a 1 unit increase in carapace width is 0.1640451. The log of the expected number of satellites for a crab with a zero carapace width is -3.3047572.

(b)

Fit a model (m2) with W and Wt as predictors, then compare to m1

```
m2 = glm(Sa ~ W + Wt,
         family = poisson,
         data = crab_data)

# View summary of m2
summary(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
```

```
# Compare m2 to m1
m1_m2_dev_stat = m1$deviance - m2$deviance
m1_m2_df = (nrow(crab_data) - 2) - (nrow(crab_data) - 3)
m1_m2_pval = 1 - pchisq(m1_m2_dev_stat, df = m1_m2_df)
```

Since the p-value = 0.0046948 < 0.05, we reject the null hypothesis that the smaller model fits the data as well as the larger model. This means the smaller model has a lack of fit and the larger model is better to use. In m2, the log ratio of the number of satellites for a 1 unit increase in carapace width is 0.045898, holding weight constant. The log ratio of the number of satellites for a 1 unit increase in weight is 0.4474357, holding carapace width constant. The log of the expected number of satellites for a crab with a zero carapace width and zero weight is -1.291679.

(c)

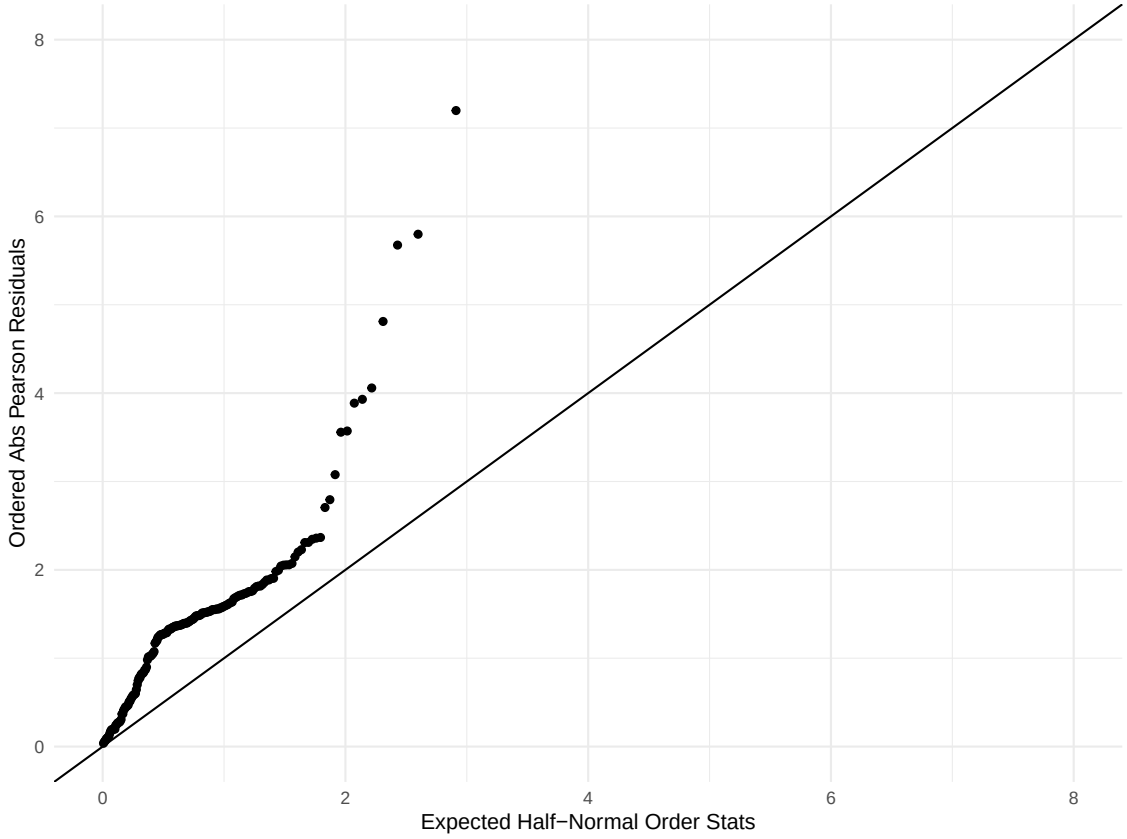
Check over dispersion in M2

```
# Residuals
m2_res = residuals(m2, type = 'pearson')

m2_resid_points = tibble(
  x = qnorm((173 + 1:173 + 0.5)/(2 * 173 + 1.125)),
```

```
  y = sort(abs(m2_res))
)

m2_resid_points %>%
  ggplot(aes(x = x, y = y)) +
  geom_point() +
  geom_abline(slope = 1) +
  scale_x_continuous(limits = c(0, 8)) +
  scale_y_continuous(limits = c(0, 8)) +
  labs(
    x = "Expected Half-Normal Order Stats",
    y = "Ordered Abs Pearson Residuals"
  )
```



Based on the above plot, there may be overdispersion. There is clear deviance from the reference line.

m2 with overdispersion

```
# Estimate overdispersion parameter
g = sum(m2_res^2)

phi = g / (173 - 3)

# m2 with overdispersion parameter
summary(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
```

In the updated model, our estimates and their interpretations are still valid. In the updated model, the std. error has been inflated by 3.1564486. This has caused the z value and Pr(>|z|) to change. The coefficients for the both carapace width and weight in the model have now clearly become insignificant, indicating that:

- there may not be a significant association between carapace width with number of satellites, when accounting for weight, and
- there may not be a significant association between weight with number of satellites, when accounting for carapace width.

Problem 2

(a)

Read in the parasite data

```
par_data = read.table("./data/HW5-parasite.txt", header = TRUE) %>%
  janitor::clean_names() %>%
  mutate(
    area = as_factor(area),
    year = as_factor(year)
  )
```

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

```
par_fit = glm(intensity ~ area + year + length,
         family = poisson,
         data = par_data)

# View summary of par_fit
summary(par_fit)
```

```
##
## Call:
## glm(formula = intensity ~ area + year + length, family = poisson,
##     data = par_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 ***
## 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 log ratio of the number of parasites for a fish with area equal to 2 vs area equal to 1 is -0.2119557, holding year and length constant. The log ratio of the number of parasites for a fish with area equal to 3 vs area equal to 1 is -0.1168602, holding year and length constant. The log ratio of the number of parasites for a fish with area equal to 4 vs area equal to 1 is 1.4049366, holding year and length constant. The log ratio of the number of parasites for a fish of year 2000 vs year 1999 is 0.6702801, holding area and length constant. The log ratio of the number of parasites for a fish of year 2001 vs year 1999 is -0.2181393, holding area and length constant. The log ratio of the number of parasites for a 1 unit increase in length is -0.0284228, holding area and year constant. The log of the expected number of parasites for a fish with area equal to 1 and of year 1999 with length equal to zero is 2.6431709.

(b)

Test for goodness of fit

```r
# Check goodness of fit of par_fit
par_fit_dev = sum(residuals(par_fit, type = 'deviance')^2)

par_fit_pval = 1 - pchisq(par_fit_dev, (nrow(par_data) - 7))
```

Since p-value $= 0 < 0.05$, we reject the null hypothesis of the model being a good fit for the data.

(c)

Fit zero-inflated poisson regression model based on the assumption that whether a fish is susceptible to parasites depends on the area of the fish and how many parasites a fish has (if they are susceptible) depends on length and year.

```r
zip_par_fit = zeroinfl(intensity ~ length + year | area, data = par_data)
```

```r
# View summary
summary(zip_par_fit)
```

```
##
## Call:
## zeroinfl(formula = intensity ~ length + year | area, data = par_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)  4.6630528  0.0459573 101.465  < 2e-16 ***
## length      -0.0438777  0.0009298 -47.193  < 2e-16 ***
## year2000     0.4214742  0.0278972  15.108  < 2e-16 ***
## year2001     0.0988372  0.0286162   3.454 0.000553 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.001796   0.121809   0.015    0.988
## area2        0.746780   0.183065   4.079 4.52e-05 ***
## area3        0.680876   0.161795   4.208 2.57e-05 ***
## area4       -0.882655   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: 15
## Log-likelihood: -7563 on 8 Df
```

The log odds ratio of a fish not being susceptible to parasites for a fish with area equal to 2 vs 1 is 0.7467803. The log odds ratio of a fish not being susceptible to parasites for a fish with area equal to 3 vs 1 is 0.680876. The log odds ratio of a fish not being susceptible to parasites for a fish with area equal to 4 vs 1 is -0.8826545. The log odds of a fish not being susceptible to parasites for a fish with area equal to 1 is 0.0017965.

The log ratio of the number of parasites for a 1 unit increase in length is -0.0438777, holding year constant, given the fish is susceptible to parasites. The log ratio of the number of parasites for a fish of year 2000 vs year 1999 is 0.4214742, holding length constant, given the fish is susceptible to parasites. The log ratio of the number of parasites for a fish of year 2001 vs year 1999 is 0.0988372, holding length constant, given the fish is susceptible to parasites. The log of the expected number of parasites for a fish of year 1999 with length equal to zero is -0.0438777, given the fish is susceptible to parasites.