HW5_yq2378

Qi Yumeng

2024-03-17

Contents

Crab

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(pscl)
```

```
## Classes and Methods for R originally developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University (2002-2015),
## by and under the direction of Simon Jackman.
## hurdle and zeroinfl functions by Achim Zeileis.
```

```
crab = readxl::read_excel("/Users/luchen/Documents/P8131 Biostatistics Method 2/P8131 HW/Data/HW5-crab.
parasite = read.delim("/Users/luchen/Documents/P8131 Biostatistics Method 2/P8131 HW/Data/HW5-parasite.
  mutate(Year = factor(Year,levels = c(1999,2000,2001),),
         Area = factor(Area, levels = c(1,2,3,4)))
```

a

```
# ggplot(crab, aes(Sa)) + geom_density()
# fit Poisson log linear model
crab_m1 <- glm(Sa~W, family=poisson(link=log), data=crab)
crab_s1 <- summary(crab_m1)
crab_s1
```

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

```
res.p1=residuals(crab_m1,type='pearson',data=crab)  # exactly the same as pearson residual for wave.glm
G1=sum(res.p1^2) # calc dispersion param based on full model
pval=1-pchisq(G1,df=171)
pval
```

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

Coefficients

(Intercept): The estimated intercept is -3.305 with a standard error of 0.542. The z value is -6.095, and the p-value is very small, indicating that the intercept is significantly different from zero.

W: The estimated coefficient for variable W is 0.164 with a standard error of 0.02. This means, for a one-unit increase in carapace width (W), the expected change in the log of satellites is 0.164. The z value is 8.216, and the p-value is extremely small (< 2e-16), indicating a very strong evidence against the null hypothesis (which would be that this coefficient is zero), suggesting that W has a significant positive effect on satellites(Sa).

Model Fit

Under the assumption of the mean and variance of the distribution are equal, the residual deviance if 567.879 on 171 degrees of freedom. Compared to the null deviance, there is little deduction. Also, the deviance is relatively large compared to the degree of freedom. If we calculate the dispersion parameter based on M1, we have a near zero pvalue, indicating M1 is lack of fit. What's more, the AIC IS 927.176, quite large. All in all, M1 model doesn't seem like a good fit.

b

```
crab_m2 <- glm(Sa~W + Wt, family=poisson, data=crab)
crab_s2 <- summary(crab_m2)
crab_s2
```

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

Coefficients

W: The estimated coefficient for variable W is 0.046 with a standard error of 0.047. This means, for a one-unit increase in carapace width (W), the expected change in the log of satellites is 0.046, holding Wt. Compared to M1, the coefficient is smaller and the standard error is larger, indicating M1 might have over-dispersion.

The z value is 0.981, and the p-value is 0.3264, indicating the influence of W on Sa is not significantly different from 0 anymore.

Wt: The estimated coefficient for variable Wt is 0.447 with a standard error of 0.159. This means, for a one-unit increase in weight (Wt), the expected change in the log of satellites is 0.447, holding W. The z value is 2.82, and the p-value is 0.0048, indicating the influence of Wt on Sa is significantly different from 0. Also, compared to W, Wt has a stronger impact on Sa.

Model Fit

```
## deviance analysis (ignoring the over dispersion)
test.stat=crab_m1$deviance-crab_m2$deviance
df=1
pval=1-pchisq(test.stat,df=df) # chisq test
pval # rej, go with the bigger model
```

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

If we compared M1 and M2 with deviance analysis and ignore the over dispersion. We reject the null hypothesis and conclude that M2 should be reserve instead of M1.

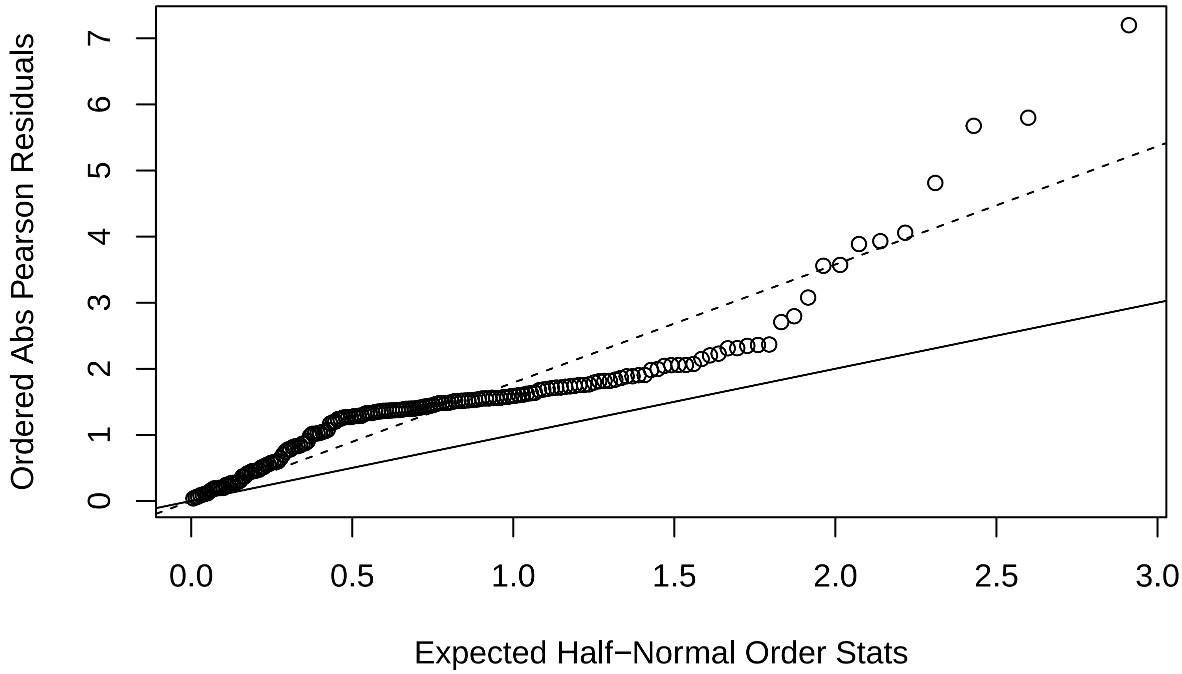
c

```
### estimate the dispersion parameter (from the additive model)
# the traditional way of calc constant dispersion parameter
res.p2=residuals(crab_m2,type='pearson',data=crab)  # exactly the same as pearson residual for wave.glm
G2=sum(res.p2^2) # calc dispersion param based on full model
pval=1-pchisq(G2,df=170) # lack of fit
phi=G1/170

crab_m2$deviance/crab_m2$df.residual
```

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

```
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)  # controversial?
```



We first prove that M2 is also lack of fit and the dispersion parameter is around 3. The half normal plot could further prove the over dispersion.

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

```
##
## Call:
## glm(formula = Sa ~ W + Wt, family = poisson, data = crab)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.29168    1.60893  -0.803    0.422
## W            0.04590    0.08367   0.549    0.583
## Wt           0.44744    0.28382   1.576    0.115
##
## (Dispersion parameter for poisson family taken to be 3.200924)
##
##     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
```

If we add the over dispersion parameter in the M2, both the coefficients of W and Wt have larger standard errors and also, their z values are no more significant.

Parasite

a

```
parasite_m1 <- glm(Intensity~ Area + Year + Length, family=poisson(link=log), data=parasite)
parasite_s1 <- summary(parasite_m1)
parasite_s1
```

```
##
## Call:
## glm(formula = Intensity ~ Area + Year + Length, family = poisson(link = log),
##     data = parasite)
##
## 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
```

Coefficients

The model takes Area1 and Year1999 as baseline.

Area: The estimated coefficients for variable Area2,3,4 are -0.212, -0.117, 1.405 with standard errors of 0.049, 0.043, 0.036. This means, being Area 2, Area 3 will decrease the expected change in the log of Intensity, while being Area 4 will increase it, holding Year and Length the same. The p-values are less than 0.05, indicating the coefficients are significantly different from 0.

Year: The estimated coefficients for variable Year2000, 2001 are 0.67 and -0.218, with standard errors of 0.028, 0.029. This means, being Year 2000 will increase the log of Intensity while being Year 2002 will decrease the outcome, holding Year1999 as benchmark and other variables the same. The p-values are less than 0.05, indicating the coefficients are significantly different from 0.

Length: The estimated coefficient for variable Year is -0.028 with a standard error of 0.001. This means, for a one-unit increase in Length, the expected change in the log of Intensity is -0.028, holding the rest variables the same. The z value is -32.265, and the p-value is also near zero, indicating the influence of Length on Intensity is significantly different from 0 under the significance value of 0.05.

b

```r
res.p3=residuals(parasite_m1,type='pearson',data=parasite)  # exactly the same as pearson residual for
G3=sum(res.p3^2) # calc dispersion param based on full model
pval=1-pchisq(G3,df=1184) # lack of fit
pval
```

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

Under the assumption of the mean and variance of the distribution are equal, the residual deviance if 1.9152798×10^4 on 1184 degrees of freedom. Compared to the null deviance, there is little deduction. If we further test the goodness of fit of the model, we have a near zero pvalue, indicating the model is lack of fit. Also, the deviance is still relatively large compared to the degree of freedom. What's more, the AIC IS 2.1088733×10^4, still quite large.

c

```r
parasite_m2 <- zeroinfl(Intensity~ Area + Year + Length, data = parasite)

summary(parasite_m2)
```

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

Count Model Coefficients (Poisson Part)

Intercept (3.8432): The log of the expected count of parasites for a fish from Area1 in the base year 1999 with a length of 0.

Area2 (0.2688), Area3 (0.1463), Area4 (0.9448): These coefficients represent the log difference in the expected count of parasites for fish in Areas 2, 3, and 4, respectively, compared to Area1, all else being equal. Area4 has a significantly higher expected parasite count.

Year2000 (0.3920), Year2001 (-0.0448): Indicates the log difference in the expected count of parasites for the years 2000 and 2001 compared to the base year, respectively. There's an increase in 2000 and a slight, non-significant decrease in 2001.

Length (-0.0368): For each unit increase in fish length, there's a log decrease in the expected count of parasites. Larger fish have fewer parasites, all else being equal.

Zero-inflation Model Coefficients (Binomial Part)

Intercept (0.5526): The log-odds of a fish being from the zero-inflated (not susceptible to parasites) group for a fish from Area1 in the base year with a length of 0.

Area2 (0.7187), Area3 (0.6577), Area4 (-1.0229): The log-odds ratio of being in the zero-inflated group for fish in these areas compared to Area1. Fish in Area4 are less likely to be in the zero-inflated group.

Year2000 (-0.7521), Year2001 (0.4565): The change in log-odds of being in the zero-inflated group for these years compared to the base year. There's a decrease in 2000, suggesting more fish were susceptible to parasites, and an increase in 2001.

Length (-0.0099): For each unit increase in fish length, the log-odds of being in the zero-inflated group slightly decreases.