HW5_yq2378

Qi Yumeng

2024-03-17

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 errors
```

```
crab = readxl::read_excel("/Users/luchen/Documents/P8131 Biostatistics Method 2/P8131 HW/Data/HW5-crab.x
parasite = read.delim("/Users/luchen/Documents/P8131 Biostatistics Method 2/P8131 HW/Data/HW5-parasite.t
```

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

Under the assumption of the mean and variance of the distribution are equal, the residual deviance if 559.885 on 170 degrees of freedom. Compared to the null deviance and the M1 deviance, there is little deduction. Also, the deviance is still relatively large compared to the degree of freedom. What's more, the AIC IS 921.183, still quite large. To sum up, M2 slightly improves the interpretability of the model.

```
## 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.

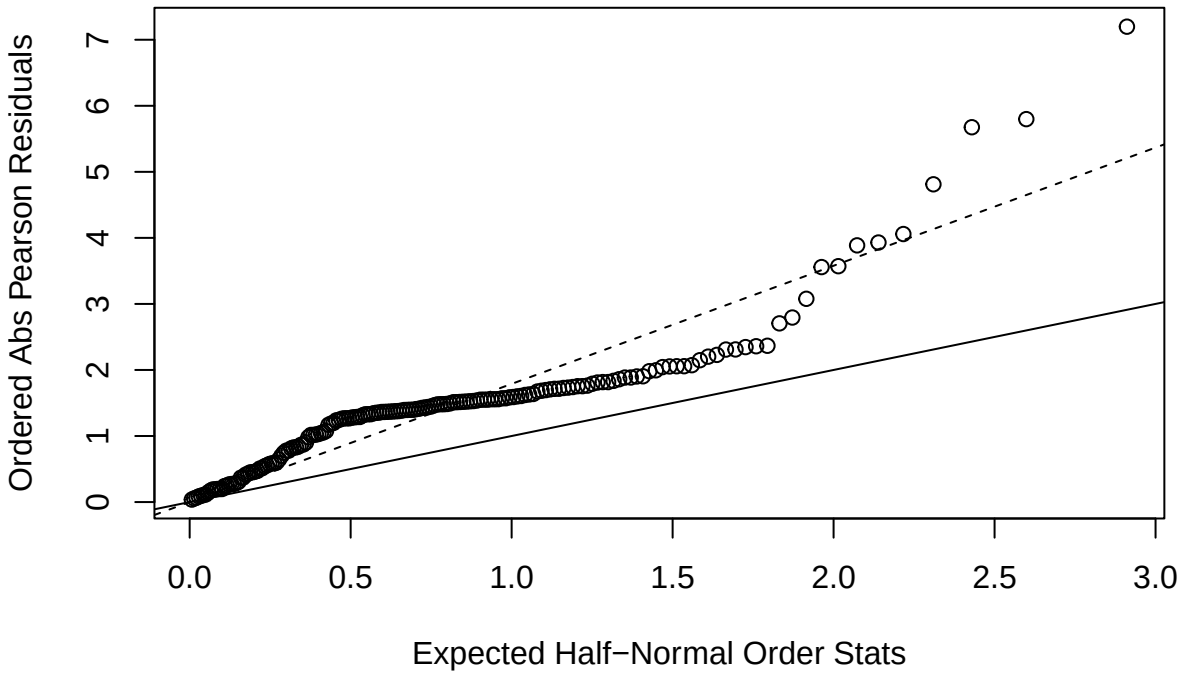
c

```
### 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

area, year, and length as predictors. Interpret each model parameter.

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

```
##
## Call:
## glm(formula = Intensity ~ Area + Year + Length, family = poisson(link = log),
##     data = parasite)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.957e+02  2.594e+01   7.544 4.55e-14 ***
## Area         6.201e-01  1.298e-02  47.767  < 2e-16 ***
## Year        -9.710e-02  1.297e-02  -7.485 7.18e-14 ***
## Length      -2.972e-02  8.513e-04 -34.912  < 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: 21312  on 1187  degrees of freedom
##   (63 observations deleted due to missingness)
## AIC: 23242
##
## Number of Fisher Scoring iterations: 7
```