bm2_Hw5

Siyan Chen_sc4456

3/7/2019

Problem 1

a

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

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

```
#gof
res_pearson = residuals(m1, crab_df, type = "pearson")
G = sum(res_pearson^2)
pval = 1-pchisq(G, 171)
pval  # fit bad
```

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

$log\lambda = \beta_0 + \beta_1 * x$
$log\lambda = -3.30 + 0.164 * x$
β_0:The log number of satellites for each female when carapace width is 0, which is not meaningful.
$\beta 1$:The log number of satellites for each female increases 0.164 for one unit increase in carapace width.

This model follows chi-square distribution with degree of 171. The p value is 0, which is significant. Therefore, this model is not good fit.

b

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

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

```
test.stat = m1$deviance - m2$deviance
df = 1
pval = 1 - pchisq(test.stat, df = 1)
pval # reject the null hypothesis, go with the bigger model
```

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

Deviance follows chi-square distribution with degree of 1. P value is significant, so we reject the null hypothesis, go with the bigger model(M2).

c

```
res_pearson2 = residuals(m2, crab_df, type = "pearson")
G2 = sum(res_pearson2^2)
1-pchisq(G2, 170) # not good fit
```

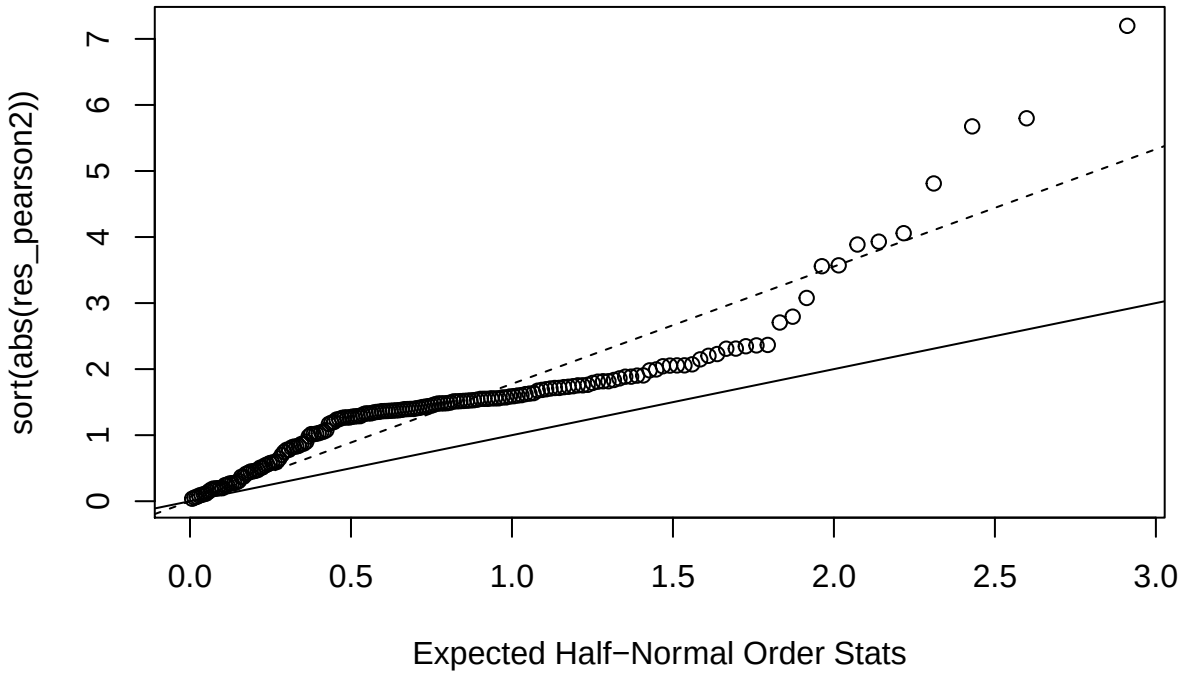
```
## [1] 0
```

```
phi = G2/(170)

summary(m2, dispersion = phi)
```

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

```
## 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
```




```
## integer(0)
```

Based on the plot, there is over dispersion om M2. The estimate of disperion parameter is 3.16. After updating the model, we got model with same coefficients but different variance of each coefficient.

$log\lambda = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2$
$log\lambda = -1.29 + 0.046 * x_1 + 0.447 * x_2$
β_0:The log number of satellites for each female when carapace width(W) is 0 and weight(Wt) is 0, which is not meaningful.
β_1:The log number of satellites for each female increases 0.046 for one unit increase in carapace width(W).
β_2:The log number of satellites for each female increases 0.447 for one unit increase in weight(Wt).

Problem 2

```
par_df = read.delim("./data/HW5-parasite.txt", header = TRUE, sep = "") %>%
  mutate(Year = as.factor(Year),
         Area = as.factor(Area))
```

a

```
p_model = glm(Intensity ~ I(Area) + I(Year) + Length, family = poisson, data = par_df)
summary(p_model)
```

```
##
## Call:
## glm(formula = Intensity ~ I(Area) + I(Year) + Length, family = poisson,
##     data = par_df)
##
## 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 ***
## I(Area)2    -0.2119557  0.0491691  -4.311 1.63e-05 ***
## I(Area)3    -0.1168602  0.0428296  -2.728  0.00636 **
## I(Area)4     1.4049366  0.0356625  39.395  < 2e-16 ***
## I(Year)2000  0.6702801  0.0279823  23.954  < 2e-16 ***
## I(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
```

$$log(\lambda) = \beta_0 + \beta_1 * I(area2) + \beta_2 * I(area3) + \beta_3 * I(area4) + \beta_4 * I(year2000) + \beta_5(year2001)$$

β_0 is not meaningful to interpret.
β_1:Log intensity rate of parasite on fish in area 2 decreases 0.212 compared to area 1, when other variables are constant.
β_2:Log intensity rate of parasite on fish in area 3 decreases 0.117 compared to area 1, when other variables are constant.
β_3:Log intensity rate of parasite on fish in area 4 increases 1.405 compared to area 1, when other variables are constant.
β_4:Log intensity rate of parasite on fish in year 2000 increases 0.670 compared to year 1999, when other variables are constant.

β_5:Log intensity rate of parasite on fish in year 2001 decreases 0218 compared to year 1999, when other variables are constant.
β_6:Log intensity rate of parasite on fish decreases 0.0284 for one unit increase in length when other variables are constant.

b

```
r = residuals(p_model, data = par_df, type = "pearson")
G_stats = sum(r^2)
pval = 1-pchisq(G_stats, 1184)
pval
```

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

The pearson residuals follows chi-square distribution with degree of 1184(G = 42164.97). P value is 0 which is significant. Therefore, we reject the null hypothesis and this model is not good fit.

c

```
i_model = zeroinfl(Intensity ~ I(Area) + I(Year) + Length, data = par_df)
summary(i_model)
```

```
##
## Call:
## zeroinfl(formula = Intensity ~ I(Area) + I(Year) + Length, data = par_df)
##
## 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 ***
## I(Area)2     0.2687835  0.0500467   5.371 7.85e-08 ***
## I(Area)3     0.1463173  0.0439485   3.329 0.000871 ***
## I(Area)4     0.9448068  0.0368342  25.650  < 2e-16 ***
## I(Year)2000  0.3919831  0.0282952  13.853  < 2e-16 ***
## I(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 *
## I(Area)2     0.718676   0.189552   3.791  0.00015 ***
## I(Area)3     0.657708   0.167402   3.929 8.53e-05 ***
## I(Area)4    -1.022868   0.188201  -5.435 5.48e-08 ***
## I(Year)2000 -0.752119   0.172965  -4.348 1.37e-05 ***
## I(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
```

We use zero-inflated Possion model to classify the response at risk and response not at risk. Z_i is a latent binary variable that generates structural zeros. $P(Z_i = 0) = \pi_i$
The response satisfies: $Y_i|(Z_i = 0) = 0$
$Y_i|(Z_i = 0)$~possion(λ_i)
Therefore, we get $log(\lambda_i) = \beta_0 + \beta_1 * I(area2) + \beta_2 * I(area3) + \beta_3 * I(area4) + \beta_4 * I(year2000) + \beta_5(year2001)$
$log(\frac{\pi_i}{1-\pi_i}) = z_i\gamma = \alpha_0 + \alpha_1 * I(area2) + \alpha_2 * I(area3) + \alpha_3 * I(area4) + \alpha_4 * I(year2000) + \alpha_5(year2001)$

β_0 is not meaningful to interpret.
β_1:Log intensity rate of parasite on fish in area 2 increases 0.27 compared to area 1, when other variables are constant.
β_2:Log intensity rate of parasite on fish in area 3 increases 0.15 compared to area 1, when other variables are constant.
β_3:Log intensity rate of parasite on fish in area 4 increases 0.94 compared to area 1, when other variables are constant.
β_4:Log intensity rate of parasite on fish in year 2000 increases 0.29 compared to year 1999, when other variables are constant.
β_5:Log intensity rate of parasite on fish in year 2001 decreases 0.045 compared to year 1999, when other variables are constant.
β_6:Log intensity rate of parasite on fish decreases 0.037 for one unit increase in length when other variables are constant.
α_0:not meaningful for length = 0.
α_1: Log odd ratio of fish not at risk of parasite vs. at risk is 0.72 for fish in area 2 versus in area 1, holding other variables constant.
α_2:Log odd ratio of fish not at risk of parasite vs. at risk is 0.66 for fish in area 3 versus in area 1, holding other variables constant.
α_3:Log odd ratio of fish not at risk of parasite vs. at risk is -1.02 for fish in area 4 versus in area 1, holding other variables constant.
α_4:Log odd ratio of fish not at risk of parasite vs. at risk is -0.75 for fish in year 2000 versus in year 1999, holding other variables constant.
α_5:Log odd ratio of fish not at risk of parasite vs. at risk is 0.45 for fish in year 2001 versus in year 1999, holding other variables constant.
α_6:Log odd ratio of fish not at risk of parasite vs. at risk decreases 0.010 for one unit increas in length, holding other variables constant.