Homework 5

Xinyi Lin

3/5/2019

```
library(readr)
```

Problem 1

Question 1

```
# import data
crab_data <- read_table2("./HW5-crab.txt")
```

Using W as the single predictor, we can fit following model:

$$log(\lambda_i) = \beta_0 + \beta_1 x_i$$

where x_i represents carapace width(W).

By using R, we fit the Posisson model(M1) as follow:

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

```
##
## 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
```

which is:

$$log(\lambda_i) = -3.305 + 0.164 x_i$$

```r
res.p1=residuals(m1.glm,type='pearson',data=crab_data)
G1.stat=sum(res.p1^2)
G1.stat
df = 173-2
pval = 1-pchisq(G1.stat,df=df) # chisq test
pval
```

By using R, we know that the generalize Pearson χ^2 equals to 544.157 and corresponding p-value equals to 0, which means this model doesn't fit the data.

Interpretation: $\beta_1 = 0.164$ means when the width of female crab increases one unit, the log-number of satellites increases 0.164.

Question 2

Using both W and Wt as predictors, we can fit following model:

$$log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

where x_1 represents carapace width(W) and x_2 represents carapace weight(Wt).

By using R, we fit the Posisson model(M2) as follow:

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

```
##
## 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
```

which is:

$$log(\lambda_i) = -1.292 - 0.046 x_{1i} + 0.447 x_{2i}$$

We use deviance to compare two models and null hypothesis is that two model have similar performance.

```
test.stat=m1.glm$deviance-m2.glm$deviance
df=171-170
pval=1-pchisq(test.stat,df=df) # chisq test
pval
```

By using R, we can get p-value equals to 0.00469, which is smaller than 0.05, so we reject null hypothesis and conclude that bigger model(M2) performs far more better.

Interpretation:

$\beta_1 = -0.046$ means that with one unit increase in width, the log-number of satellites for one female crab decreases 0.046 given the weight keeps the same.;

$\beta_2 = 0.447$ means that with one unit increase in weight, the log-number ratio of satellites for one female crab increases 0.447 given the width keeps the same.

Question 3

First, we need to calculate the goodness of fit of M2.

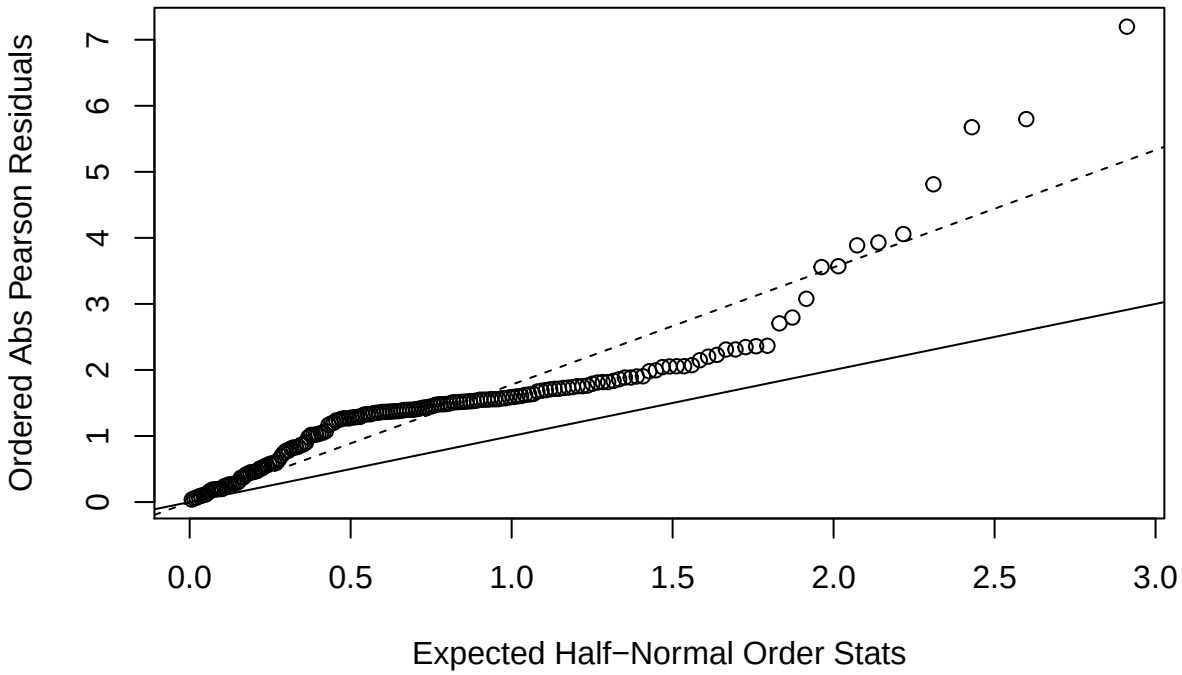
```
res.p2=residuals(m2.glm,type='pearson',data=crab_data)
G2.stat=sum(res.p2^2)
G2.stat
df = 173-3
pval = 1-pchisq(G2.stat,df=df) # chisq test
pval
```

By using R, we know that the generalize Pearson χ^2 equals to 536.596, corresponding p-value equals to 0, which means this model doesn't fit the data and there might be over dispersion.

```
phi=G2.stat/(173-3)
phi
m2.glm$deviance/m2.glm$df.residual
```

By using R, we get dispersion parameter $\phi = 3.156$ and half normal plot is as follow:

```
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)
```



By using R, we get adjusted model as follow:

```
summary(m2.glm,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
```

which is:

$$log(\lambda_i) = -1.292 - 0.046x_{1i} + 0.447x_{2i}$$

Interpretation:

$\beta_1 = -0.046$ means that with one unit increase in width, the log-number of satellites for one female crab decreases 0.046 given the weight keeps the same.;

$\beta_2 = 0.447$ means that with one unit increase in weight, the log-number ratio of satellites for one female crab increases 0.447 given the width keeps the same.

Estimated values of parameters are the same in the adjusted model, while variances increase.

Problem 2

Question 1

```
# import data
parasite_data <- read_table2("./HW5-parasite.txt")
parasite_data = na.omit(parasite_data)
parasite_data$Year = as.factor(parasite_data$Year)
parasite_data$Area = as.factor(parasite_data$Area)
```

With predictors area, year and length, we can fit following model:

$$log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i}$$

where x_1 is a indicator of area2, x_2 is a indicator of area3, x_3 is a indicator of area4, x_4 is a indicator of year 2000, x_5 is a indicator of year 2001, x_6 is a indicator of length.

```
parasite.glm = glm(Intensity ~ Area + Year + Length, family=poisson, data=parasite_data)
summary(parasite.glm)
```

```
##
## Call:
## glm(formula = Intensity ~ Area + Year + Length, family = poisson,
##     data = parasite_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
## AIC: 21089
##
## Number of Fisher Scoring iterations: 7
```

which is:

$$log(\lambda_i) = 2.643 - 0.212x_{1i} - 0.117x_{2i} + 1.405x_{3i} + 0.6703x_{4i} - 0.2181x_{5i} - 0.0284x_{6i}$$

Interpretation:

$\beta_1 = -0.212$ means comparing to area 1, the log-number of parasites in fish living in area 2 decreases 0.212 given year and length are the same.

$\beta_2 = -0.117$ means comparing to area 1, the log-number of parasites in fish living in area 3 decreases 0.117 given year and length are the same.

$\beta_3 = 1.405$ means comparing to area 1, the log-number of parasites in fish living in area 4 increases 1.405 given year and length are the same.

$\beta_4 = 0.6703$ means comparing to year 1999, the log-number of parasites in fish increases 0.6703 in year 2000 given area and length are the same.

$\beta_5 = -0.2181$ means comparing to year 1999, the log-number of parasites in fish decrease 0.2181 in year 2001 given area and length are the same.

$\beta_6 = -0.0284$ means with one unit increases in length, the log-number of parasites in fish decrease 0.0284 given area and year are the same.

Question 2

```
res.p=residuals(parasite.glm,type='pearson',data=parasite_data)
G.stat=sum(res.p^2)
G.stat
df = 1191-6
pval = 1-pchisq(G2.stat,df=df) # chisq test
pval
```

By using R, we know that the generalize Pearson χ^2 equals to 42164.97, corresponding p-value equals to 0, which means this model doesn't fit the data.

Question 3

We assume whether a fish is susceptible to parasites depends on area, year and length and how many parasites in one fish depends on area, year and length.

Let Z_i be a latent binary variable that generates structural zeros

$$P(Z_i = 0) = \pi_i$$

.

The response satisfies

$$Y_i|(Z_i = 0) = 0$$
$$Y_i|(Z_i = 1) \sim Poisson(\lambda_i)$$

Then we get corresponding models:

$$\log(\frac{\pi_i}{1-\pi_i}) = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{4i} + \alpha_5 x_{5i} + \alpha_6 x_{6i}$$

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i}$$

where x_1 is a indicator of area2, x_2 is a indicator of area3, x_3 is a indicator of area4, x_4 is a indicator of year 2000, x_5 is a indicator of year 2001, x_6 is a indicator of length.

We use R and fit following models:

```
library(pscl)
zero.model <- zeroinfl(Intensity ~ Area + Year + Length, data = parasite_data)
summary(zero.model)
```

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

which are:

$$\log(\frac{\pi_i}{1-\pi_i}) = 0.553 + 0.719x_{1i} + 0.658x_{2i} - 1.023x_{3i} - 0.752x_{4i} + 0.457x_{5i} - 0.010x_{6i}$$

$$\log(\lambda_i) = 3.843 + 0.269x_{1i} + 0.146x_{2i} + 0.945x_{3i} + 0.392x_{4i} - 0.045x_{5i} - 0.037x_{6i}$$

Interpretation:

$\alpha_0 = 0.553$ means the log odd of a fish is not susceptible to parasites is 0.553;

$\alpha_1 = 0.719$ means the log odd ratio of a fish is not susceptible to parasites in area 2 vs area 1 is 0.719 given length and year are the same;

$\alpha_2 = 0.658$ means the log odd ratio of a fish is not susceptible to parasites in area 3 vs area 1 is 0.658 given length and year are the same;

$\alpha_3 = 1.023$ means the log odd ratio of a fish is not susceptible to parasites in area 4 vs area 1 is 1.023 given length and year are the same;

$\alpha_4 = -0.752$ means the log odd ratio of a fish is not susceptible to parasites in year 2000 vs year 1999 is -0.752 given area and length are the same;

$\alpha_5 = 0.457$ means the log odd ratio of a fish is not susceptible to parasites in year 2001 vs year 1999 is 0.457 given area and length are the same;

$\alpha_6 = -0.010$ means the log odd ratio of a fish is not susceptible to parasites is 0.658 with one unit increases in length given area and year are the same.

$\beta_1 = 0.269$ means comparing to area 1, the log-number of parasites in fish living in area 2 increases 0.269 given year and length are the same when fishes are susceptible to parasites;

$\beta_2 = 0.146$ means comparing to area 1, the log-number of parasites in fish living in area 3 increases 0.146 given year and length are the same when fishes are susceptible to parasites;

$\beta_3 = 0.945$ means comparing to area 1, the log-number of parasites in fish living in area 4 increases 0.945 given year and length are the same when fishes are susceptible to parasites;

$\beta_4 = 0.392$ means comparing to year 1999, the log-number of parasites in fish increases 0.392 in year 2000 given area and length are the same when fishes are susceptible to parasites;

$\beta_5 = -0.045$ means comparing to year 1999, the log-number of parasites in fish decrease 0.045 in year 2001 given area and length are the same when fishes are susceptible to parasites;

$\beta_6 = -0.037$ means with one unit increases in length, the log-number of parasites in fish decrease 0.037 given area and year are the same when fishes are susceptible to parasites.

Appendix Code

```
library(readr)
# import data
crab_data <- read_table2("./HW5-crab.txt")
m1.glm = glm(Sa~W, family=poisson, data=crab_data)
summary(m1.glm)
res.p1=residuals(m1.glm,type='pearson',data=crab_data)
G1.stat=sum(res.p1^2)
G1.stat
df = 173-2
pval = 1-pchisq(G1.stat,df=df) # chisq test
pval
m2.glm = glm(Sa~W + Wt, family=poisson, data=crab_data)
summary(m2.glm)
test.stat=m1.glm$deviance-m2.glm$deviance
df=171-170
pval=1-pchisq(test.stat,df=df) # chisq test
pval
res.p2=residuals(m2.glm,type='pearson',data=crab_data)
G2.stat=sum(res.p2^2)
G2.stat
df = 173-3
pval = 1-pchisq(G2.stat,df=df) # chisq test
pval
phi=G2.stat/(173-3)
phi
```

```
m2.glm$deviance/m2.glm$df.residual
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)
summary(m2.glm,dispersion=phi)
# import data
parasite_data <- read_table2("./HW5-parasite.txt")
parasite_data = na.omit(parasite_data)
parasite_data$Year = as.factor(parasite_data$Year)
parasite_data$Area = as.factor(parasite_data$Area)
parasite.glm = glm(Intensity ~ Area + Year + Length, family=poisson, data=parasite_data)
summary(parasite.glm)
res.p=residuals(parasite.glm,type='pearson',data=parasite_data)
G.stat=sum(res.p^2)
G.stat
df = 1191-6
pval = 1-pchisq(G2.stat,df=df) # chisq test
pval
library(pscl)
zero.model <- zeroinfl(Intensity ~ Area + Year + Length, data = parasite_data)
summary(zero.model)
```