Coding

Bohao Tang

February 25, 2018

6.18

Here we do a similiar thing as in Example 6.3.2. The basic main effect model is:

$$\log \frac{\pi_{ij}}{\pi_{i1}} = \beta_{j0} + \beta_{j1}s_i + \beta_{j2}g_i + \beta_{j3}z_i^O + \beta_{j4}z_i^T + \beta_{j5}z_i^G$$

For $j = 2, 3, 4, 5$. Here the notation is the same as in Eample 6.3.2. And g_i is the gender for alligator i. Then we fit the model:

```
library("VGAM")

Alligators = read.table("Alligators2.txt", header = TRUE)

fit <- vglm(formula = cbind(y2,y3,y4,y5,y1) ~ size + gender + factor(lake),
            family = multinomial, data = Alligators)

summary(fit)
```

```
##
## Call:
## vglm(formula = cbind(y2, y3, y4, y5, y1) ~ size + gender + factor(lake),
##     family = multinomial, data = Alligators)
##
##
## Pearson residuals:
##                         Min      1Q   Median     3Q   Max
## log(mu[,1]/mu[,5]) -1.3218 -0.4611  0.01054 0.3810 1.866
## log(mu[,2]/mu[,5]) -0.7033 -0.5751 -0.35511 0.2610 2.064
## log(mu[,3]/mu[,5]) -1.1985 -0.5478 -0.22421 0.3678 3.478
## log(mu[,4]/mu[,5]) -1.6945 -0.2893 -0.10807 1.1236 1.367
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept):1      -2.9477     0.6741  -4.373 1.23e-05 ***
## (Intercept):2      -1.7295     0.7928  -2.182  0.02914 *
## (Intercept):3      -1.1266     0.6696  -1.682  0.09248 .
## (Intercept):4      -0.9547     0.5295  -1.803  0.07138 .
## size:1              1.3363     0.4112   3.250  0.00116 **
## size:2             -0.5570     0.6466  -0.861  0.38898
## size:3             -0.7302     0.6523  -1.120  0.26292
## size:4              0.2906     0.4599   0.632  0.52751
## gender:1           -0.4630     0.3955  -1.171  0.24180
## gender:2           -0.6276     0.6853  -0.916  0.35978
## gender:3           -0.6064     0.6888  -0.880  0.37867
## gender:4           -0.2526     0.4663  -0.542  0.58810
## factor(lake)2:1     2.6937     0.6693   4.025 5.70e-05 ***
## factor(lake)2:2     1.4008     0.8105   1.728  0.08393 .
## factor(lake)2:3    -1.1256     1.1923  -0.944  0.34513
## factor(lake)2:4  -0.7405     0.7421  -0.998  0.31837
## factor(lake)3:1   2.9363     0.6874   4.272 1.94e-05 ***
## factor(lake)3:2   1.9316     0.8253   2.340  0.01926 *
## factor(lake)3:3   0.6617     0.8461   0.782  0.43415
## factor(lake)3:4   0.7912     0.5879   1.346  0.17840
## factor(lake)4:1   1.7805     0.6232   2.857  0.00428 **
## factor(lake)4:2  -1.1295     1.1928  -0.947  0.34369
## factor(lake)4:3  -0.5753     0.7952  -0.723  0.46943
## factor(lake)4:4  -0.7666     0.5686  -1.348  0.17756
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  4
##
## Names of linear predictors:
## log(mu[,1]/mu[,5]), log(mu[,2]/mu[,5]), log(mu[,3]/mu[,5]), log(mu[,4]/mu[,5])
##
## Residual deviance: 50.2637 on 40 degrees of freedom
##
## Log-likelihood: -73.3221 on 40 degrees of freedom
##
## Number of iterations: 5
##
## No Hauck-Donner effect found in any of the estimates
##
## Reference group is level  5  of the response
```

The residual deviance 50.2637 with $df = 40$ doesn't give much evidence against our model, because the p value is not small:

```
1 - pchisq(50.2637,40)
```

```
## [1] 0.1281848
```

Drop `size` or `gender` will cause a significant loss while no significant loss to drop `gender`. But dropping `gender` we will get the same result as in Example 6.3.2. Therefore here we accept our original model.

We have:

$$\log \frac{\hat{\pi}_{i2}}{\hat{\pi}_{i1}} = -2.9477 + 1.3363 s_2 - 0.463 g_2 + 2.6937 z_2^O + 2.9363 z_2^T + 1.7805 z_2^G$$

For a given lake and gender, for small alligators the estimated odds that primary food choice was invertebrates instead of fish are exp(1.3363) = 3.8 times the estimated odds for large alligators. The estimated effect is imprecise, as the Wald 95% confidence interval is exp[1.3363 ± 1.96(0.4412)] = (1.602, 9.034). For a given lake and size, for gender = 0 alligators the estimated odds that primary food choice was invertebrates instead of fish are exp(0.463) = 1.5888 times the estimated odds for gender=1 alligators. The estimated effect is also imprecise, as the Wald 95% confidence interval is exp[0.463 ± 1.96(0.3955)] = (0.7318, 3.44933). The lake effects indicate that the estimated odds that the primary food choice was invertebrates instead of fish are relatively higher at lakes Ocklawaha, Trafford and George than they are at Lake Hancock.

7.31

a.

We fit the model like below:

$$\log \mu_i = \beta_0 + \beta_1 x_i$$

```
Homicide = read.table("Homicide.txt", header = TRUE)

fit.poi <- glm(count ~ race, family=poisson, data=Homicide)

summary(fit.poi)
```

```
##
## Call:
## glm(formula = count ~ race, family = poisson, data = Homicide)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0218  -0.4295  -0.4295  -0.4295   6.1874
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.38321    0.09713  -24.54   <2e-16 ***
## race         1.73314    0.14657   11.82   <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: 962.80  on 1307  degrees of freedom
## Residual deviance: 844.71  on 1306  degrees of freedom
## AIC: 1122
##
## Number of Fisher Scoring iterations: 6
```

```
logLik(fit.poi)
```

```
## 'log Lik.' -558.9949 (df=2)
```

Then the $\hat{\beta} = 1.73314$ means the estimated log ratio of mean for black and white people is 1.73314.

b.

Maybe the time interval to collect the response for each people is not fixed but rather a random variable. Then the Poisson assumption is inadequate.

Here we fit the corresponding negative binomial model:

```
library("MASS")

fit.nb <- glm.nb(count ~ race, data=Homicide)

summary(fit.nb)
```

```
##
## Call:
## glm.nb(formula = count ~ race, data = Homicide, init.theta = 0.2023119205,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7184  -0.3899  -0.3899  -0.3899   3.5072
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.3832     0.1172 -20.335  < 2e-16 ***
## race          1.7331     0.2385   7.268 3.66e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.2023) family taken to be 1)
##
##     Null deviance: 471.57  on 1307  degrees of freedom
## Residual deviance: 412.60  on 1306  degrees of freedom
## AIC: 1001.8
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.2023
##           Std. Err.:  0.0409
##
##  2 x log-likelihood:  -995.7980
```

Here, the AIC for NB2 is much smaller than that for Poisson (1001.8 vs 1122). The much larger SE(0.2385 vs 0.14657) also the estimated NB2 dispersion parameter is $1/0.2023 = 4.943$ is much larger than 1. We can notice that Poisson model had overdispersion.

c.

Here we are calculating the Wald 95% CI for β. We have:

For Poisson model:

```
print(exp(1.7331+0.14657*c(qnorm(0.025),qnorm(0.975))))
```

```
## [1] 4.245366 7.541129
```

For Negative Binomial model:

```
print(exp(1.7331+0.2385*c(qnorm(0.025),qnorm(0.975))))
```

```
## [1] 3.545391 9.029991
```

Since we have enough evidence that Poisson GLM have an overdispersion, we rather believe the larger interval for $\hat{\beta}$. Also the AIC shows that NB GLM is a better model, so we choose (3.5,9.0) to be our CI.

overdispersion

The only thing we need to do is to estimate the overdispersion term σ^2 for $var(Y_i) = \sigma^2 \mu_i$.

We have:

```
n = length(Homicide$Obs)
p = 1
sigma_2 = 1 / (n-p) * sum((Homicide$count - fitted(fit.poi))^2 / fitted(fit.poi))
sigma_2
```

```
## [1] 1.744356
```

Therefore the ture estimate variance for $\hat{\beta}$ is $0.14657 * 1.744356 = 0.2556703$. So the CI is

```r
print(exp(1.7331+0.2556703*c(qnorm(0.025),qnorm(0.975))))
```

```
## [1] 3.428063 9.339050
```