STATISTICAL MODELLING Part IIC, Michaelmas 2019

Practical 8: Contingency tables and Gamma regression By courtesy of Dr. S. A. Bacallado and Dr. R. Shah

Studying the effect of sociability

Suppose we want to test the effect of sociability, as measured by the variables `go_out_f` and `go_out_m`. In particular, we are interested in the hypothesis that opposites attract. To do this, we can represent the ordinal variables `go_out_f` and `go_out_m` as numbers between 1 and 7, and include a predictor in the Binomial regression model which is the absolute value of their difference.

```
file_path <- "https://raw.githubusercontent.com/AJCoca/SM19/master/"
SD_data <- read.csv(paste0(file_path, "SD_match.csv"))

levels(SD_data$subject_m) <- c("Arts+Humanities", "Econ+Law", "Econ+Law", "Sciences")
levels(SD_data$subject_f) <- c("Arts+Humanities", "Econ+Law", "Econ+Law", "Sciences")

levels(SD_data$go_out_m)
```

```
## [1] "> 2/week"     "1/month"      "1/week"       "2/month"
## [5] "2/week"       "almost never" "several/yr"
```

```
levels(SD_data$go_out_m) <- c("7","3","5","4","6","1","2")
levels(SD_data$go_out_f) <- c("7","3","5","4","6","1","2")
SD_data$diff = abs(as.numeric(SD_data$go_out_m)-as.numeric(SD_data$go_out_f))

# Fit the GLM
mod2bin <- glm(match~subject_m+subject_f+diff,data=SD_data,family="binomial")
summary(mod2bin)
```

```
##
## Call:
## glm(formula = match ~ subject_m + subject_f + diff, family = "binomial",
##     data = SD_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7138  -0.6364  -0.5978  -0.5683   1.9865
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.63055    0.11880 -13.725   <2e-16 ***
## subject_mEcon+Law  0.14730    0.11807   1.248   0.2122
## subject_mSciences -0.03510    0.12735  -0.276   0.7829
## subject_fEcon+Law  0.24580    0.11352   2.165   0.0304 *
## subject_fSciences  0.06448    0.11086   0.582   0.5608
## diff              -0.03705    0.02931  -1.264   0.2062
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3226.4  on 3517  degrees of freedom
## Residual deviance: 3216.6  on 3512  degrees of freedom
##   (92 observations deleted due to missingness)
```

```
## AIC: 3228.6
##
## Number of Fisher Scoring iterations: 4
```

There does not seem to be evidence in favor of the hypothesis that opposites attract. By contrast, note, e.g., that the total sociability of the pair has a statistically significant, negative effect on the chance of a match:

```
SD_data$addSoc = as.numeric(SD_data$go_out_m) + as.numeric(SD_data$go_out_f)

# Fit the GLM
mod3bin <- glm(match~subject_m+subject_f+addSoc,data=SD_data, family="binomial")
summary(mod3bin)
```

```
##
## Call:
## glm(formula = match ~ subject_m + subject_f + addSoc, family = "binomial",
##     data = SD_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7622  -0.6330  -0.5910  -0.5431   2.1238
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.30077    0.15963  -8.149 3.68e-16 ***
## subject_mEcon+Law  0.11388    0.11867   0.960   0.3372
## subject_mSciences  0.01027    0.12820   0.080   0.9362
## subject_fEcon+Law  0.22147    0.11389   1.945   0.0518 .
## subject_fSciences  0.02867    0.11148   0.257   0.7971
## addSoc            -0.06100    0.01854  -3.290   0.0010 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3226.4  on 3517  degrees of freedom
## Residual deviance: 3207.3  on 3512  degrees of freedom
##   (92 observations deleted due to missingness)
## AIC: 3219.3
##
## Number of Fisher Scoring iterations: 4
```

Gamma regression

```
Drinks <- read.table(paste(file_path, "drinks.txt", sep =""), header = TRUE)
GammaMod1 <- glm(Time ~ Distance + Cases, family = Gamma, data=Drinks)
summary(GammaMod1)
```

```
##
## Call:
## glm(formula = Time ~ Distance + Cases, family = Gamma, data = Drinks)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.56304  -0.18654  -0.00856   0.10545   0.49732
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.132e-02  4.154e-03  17.171 3.14e-14 ***
## Distance    -6.428e-06  1.039e-05  -0.618  0.54269
## Cases       -1.728e-03  5.093e-04  -3.393  0.00261 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.06801047)
##
##     Null deviance: 7.7060  on 24  degrees of freedom
## Residual deviance: 1.5431  on 22  degrees of freedom
## AIC: 157.29
##
## Number of Fisher Scoring iterations: 4
```

The variable that appears not to be significant is `Distance`, so we obtain the second fit through:

```
GammaMod0 <- glm(Time ~ Cases, family = Gamma, data=Drinks)
```

We can verify that the F-test computed by the `anova` is the same as the one described in the practical sheet.

```
Ftest <- (GammaMod0$dev - GammaMod1$dev) / summary(GammaMod1)$dispersion
pf(Ftest, df1 = 1, df2 = 25 - 3, lower.tail = FALSE)
```

```
## [1] 0.5351263
```

```
anova(GammaMod0,GammaMod1,test="F")
```

```
## Analysis of Deviance Table
##
## Model 1: Time ~ Cases
## Model 2: Time ~ Distance + Cases
##   Resid. Df Resid. Dev Df Deviance     F Pr(>F)
## 1        23     1.5701
## 2        22     1.5431  1 0.027001 0.397 0.5351
```

There are at least two disadvantages to using the canonical link function. First, since the mean μ_i is positive, it restricts $x_i^T\beta$ to be positive. Second, it lacks the convenient interpretation of the coefficients furnished by the log-link, i.e. a unit increase in the jth predictor multiplies the mean of the response by $\exp(\beta_j)$.

We can compare the fit of the models `GammaMod1` and `GammaMod2` through their AIC.

```
GammaMod2 <- glm(Time ~ Distance + Cases, family = Gamma(link = log), data=Drinks)
AIC(GammaMod1,GammaMod2)
```

```
##           df      AIC
## GammaMod1  4 157.2891
## GammaMod2  4 134.1222
```

The second model has a better quality of fit as expected from looking at `summary(GammaMod2)`.