# Biostat 200C Homework 2

Due Apr 26 @ 11:59PM

## Q1. Beta-Binomial

Let $Y_i$ be the number of successes in $n_i$ trials with

$$Y_i \sim \text{Bin}(n_i, \pi_i),$$

where the probabilities $\pi_i$ have a Beta distribution

$$\pi \sim \text{Be}(\alpha, \beta)$$

with density function

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad x \in [0,1], \alpha > 0, \beta > 0.$$

### 1.1

Find the mean and variance of $\pi$.

Solution: See the pdf file attached.

### 1.2

Find the mean and variance of $Y_i$ and show that the variance of $Y_i$ is always larger than or equal to that of a Binomial random variable with the same batch size and mean.

Solution: See the pdf file attached.

## Q2. Poisson regression log-likelihood

Let $Y_1, \ldots, Y_n$ be independent random variables with $Y_i \sim \text{Poisson}(\mu_i)$ and $\log \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$, $i = 1, \ldots, n$.

### 2.1

Write down the log-likelihood function.

Solution: See the pdf file attached.

### 2.2

Derive the gradient vector of the log-likelihood function with respect to the regression coefficients $\boldsymbol{\beta}$, i.e. taking derivative with respect to each $\beta_j$.

Solution: See the pdf file attached.

**2.3**

Show that for the fitted values $\widehat{\mu}_i$ from maximum likelihood estimates

$$\sum_i \widehat{\mu}_i = \sum_i y_i.$$

Therefore the deviance reduces to

$$D = 2\sum_i y_i \log \frac{y_i}{\widehat{\mu}_i}.$$

Solution: See the pdf file attached.

## Q3. Simpson's paradox

The dataset `death` contains data on murder cases in Florida in 1977. The data is cross-classified by the race (black or white) of the victim, of the defendant and whether the death penalty was given.

**3.1**

Consider the frequency with which the death penalty is applied to black and white defendants, both marginally and conditionally, with respect to the race of the victim. Is this an example of Simpson's paradox? Are the observed differences in the frequency of application of the death penalty statistically significant?

```
library(faraway)
```

```
## Warning: package 'faraway' was built under R version 4.1.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data(death)

m.death <-xtabs(y~ penalty + defend, data = death)
m.death
```

```
##        defend
## penalty   b   w
##     no  149 141
##     yes  17  19
```

```
summary(m.death)
```

```
## Call: xtabs(formula = y ~ penalty + defend, data = death)
## Number of cases in table: 326
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 0.22145, df = 1, p-value = 0.6379
```

Because the p-value is 0.6379, which is greater than the significant level $\alpha$ 0.05, we do not reject the null hypothesis and conclude that there is no association between death penalty and the race of the defendants if we consider them marginally.

```
c.death = xtabs(y~ penalty + defend + victim, data = death)
c.death
```

```
## , , victim = b
##
##        defend
## penalty   b   w
##     no   97   9
##     yes   6   0
##
## , , victim = w
##
##        defend
## penalty   b   w
##     no   52 132
##     yes  11  19
```

```
summary(c.death)
```

```
## Call: xtabs(formula = y ~ penalty + defend + victim, data = death)
## Number of cases in table: 326
## Number of factors: 3
## Test for independence of all factors:
##  Chisq = 122.4, df = 4, p-value = 1.642e-25
```

Because the p-value $< 0.001$, which is smaller than the significant level $\alpha$ 0.05, we reject the null hypothesis and conclude that there is a association between death penalty and the race of the defendants if we consider them conditionally on the race of the victim.

```
marginal = (149 * 19) / (17 * 141)
marginal
```

```
## [1] 1.18106
```

```
victim.b = (97.5 * 0.5) / (6.5 * 9.5)
victim.b
```

```
## [1] 0.7894737
```

```r
victim.w = (52 * 19) / (11 * 132)
victim.w
```

```
## [1] 0.6804408
```

Additionally, we can see that the marginal odds ratio is 1.18, and two conditional odds ratios are 0.79 and 0.68 (smaller than 1).

In this case, it is the Simpson's Paradox, when the marginal association contradicts the conditional association between death penalty and the race of the defendants.

### 3.2

Determine the most appropriate dependence model between the variables.

```r
glm(y ~ (penalty + victim + defend)^2, family = poisson, data = death) %>%
  step()
```

```
## Start:  AIC=50.38
## y ~ (penalty + victim + defend)^2
##
##                   Df Deviance     AIC
## - penalty:defend  1    1.882  49.563
## <none>                 0.701  50.382
## - penalty:victim  1    7.910  55.592
## - victim:defend   1  131.458 179.140
##
## Step:  AIC=49.56
## y ~ penalty + victim + defend + penalty:victim + victim:defend
##
##                   Df Deviance     AIC
## <none>                 1.882  49.563
## - penalty:victim  1    8.132  53.813
## - victim:defend   1  131.680 177.361

##
## Call:  glm(formula = y ~ penalty + victim + defend + penalty:victim +
##     victim:defend, family = poisson, data = death)
##
## Coefficients:
##        (Intercept)          penaltyyes             victimw             defendw
##             4.5797             -2.8717             -0.5876             -2.4375
## penaltyyes:victimw     victimw:defendw
##             1.0579              3.3116
##
## Degrees of Freedom: 7 Total (i.e. Null);  2 Residual
## Null Deviance:       395.9
## Residual Deviance: 1.882     AIC: 49.56
```

```r
glm(y ~ penalty*victim + victim*defend, family = poisson, data = death) %>%
  summary()
```

```
##
## Call:
## glm(formula = y ~ penalty * victim + victim * defend, family = poisson,
##     data = death)
##
## Deviance Residuals:
##        1         2         3         4         5         6         7         8
## -0.47967   0.18976  -0.98198   0.16368   0.70243  -0.29660   0.20237  -0.04887
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)           4.5797     0.1011  45.314  < 2e-16 ***
## penaltyyes           -2.8717     0.4196  -6.843 7.75e-12 ***
## victimw              -0.5876     0.1639  -3.586 0.000336 ***
## defendw              -2.4375     0.3476  -7.013 2.34e-12 ***
## penaltyyes:victimw    1.0579     0.4635   2.282 0.022471 *
## victimw:defendw       3.3116     0.3786   8.748  < 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: 395.9153  on 7  degrees of freedom
## Residual deviance:   1.8819  on 2  degrees of freedom
## AIC: 49.563
##
## Number of Fisher Scoring iterations: 4
```

By the Stepwise Algorithm, we choose the model by AIC and get the model which suggests that given victim, penalty and defend are independent.

The conditional independence model has the smallest AIC and residual deviance, thus, it is the most appropriate model.

### 3.3

Fit a binomial regression with death penalty as the response and show the relationship to your model in the previous question.

```
deathu = cbind(matrix(death$y, ncol = 2, byrow = TRUE),
               unique(death %>% select(victim, defend)))
colnames(deathu)[1:2] = c('yes','no')
glm(cbind(yes, no) ~ victim * defend, family = binomial, data = deathu) %>%
  step(test = "Chi") %>%
  summary()
```

```
## Start:  AIC=20.31
## cbind(yes, no) ~ victim * defend
##
##                 Df Deviance    AIC     LRT Pr(>Chi)
## - victim:defend  1  0.70074 19.015 0.70074   0.4025
## <none>              0.00000 20.314
##
```

```
## Step:  AIC=19.01
## cbind(yes, no) ~ victim + defend
##
##          Df Deviance    AIC    LRT Pr(>Chi)
## - defend  1   1.8819 18.196 1.1812 0.277121
## <none>        0.7007 19.015
## - victim  1   7.9102 24.224 7.2094 0.007252 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Step:  AIC=18.2
## cbind(yes, no) ~ victim
##
##          Df Deviance    AIC    LRT Pr(>Chi)
## <none>        1.8819 18.196
## - victim  1   8.1316 22.445 6.2497  0.01242 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Call:
## glm(formula = cbind(yes, no) ~ victim, family = binomial, data = deathu)
##
## Deviance Residuals:
##       1        3        5        7
## -0.5158  -0.9955   0.7625   0.2082
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.8717     0.4196  -6.843 7.75e-12 ***
## victimw       1.0579     0.4635   2.282   0.0225 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8.1316  on 3  degrees of freedom
## Residual deviance: 1.8819  on 2  degrees of freedom
## AIC: 18.196
##
## Number of Fisher Scoring iterations: 4
```

We fit the binomial model with death penalty as the response. The final model has `victim` as the single predictor.

The residual deviance is the same as that from the **conditional independence** model, meaning the two models are equivalent.