Homework 2

Tim Vigers

14 February 2019

1. Model 1 equation estimates

Strike 1

$$ln(\text{odds}_{\text{strike 1}}) = ln(\frac{619}{1797}) = -1.066 = \hat{\beta}_0$$

Strike 2

$$ln(\text{odds}_{\text{strike 2}}) = ln(\frac{355}{416}) = -0.159 = \hat{\beta}_0 + \hat{\beta}_1$$

$$\hat{\beta}_1 = -0.159 - \hat{\beta}_0 = -0.159 - (-1.066) = 0.907$$

Strike 3

$$ln(\text{odds}_{\text{strike 3}}) = ln(\frac{162}{569}) = -1.256 = \hat{\beta}_0 + \hat{\beta}_2$$

$$\hat{\beta}_2 = -1.256 - \hat{\beta}_0 = -1.256 - (-1.066) = -0.190$$

Model 1

$$\text{logit P (misconduct violation)} = \hat{\beta}_0 + \hat{\beta}_1 * \text{strikes 2} + \hat{\beta}_2 * \text{strikes 3} = -1.066 + 0.907 * \text{strikes 2} - 0.190 * \text{strikes 3}$$

2. Log-likelihood for Model 1

strikes	y	n	sum
1	619	1797	2416
2	355	416	771
3	162	569	731

$$p1 = \frac{619}{2416}$$
$$p2 = \frac{355}{771}$$
$$p3 = \frac{162}{731}$$

$$LL = 619 * ln(p1) + 1797 * ln(1 - p1) + 355 * ln(p2) + 416 * ln(1 - p2) + 162 * ln(p3) + 569 * ln(1 - p3) = -2293.492$$

3. Log-likelihood for Model 0

Calculate p estimate at the MLE

$$\hat{p} = \frac{\text{number with misconduct}}{\text{total n}} = \frac{1136}{3918} = 0.290$$

Calculate log-likelihood estimate

$$LL = \text{total number with misconduct} * ln(\hat{p}) + \text{total number without misconduct} * ln(1 - \hat{p}) = 1136 * ln(0.290) + 2782 * ln(0.710) = -2359.0$$

4. Perform a likelihood ratio test comparing Model 1 with Model 0

Calculate the LRT statistic

$$\text{LRT statistic} = 2(LL_{\text{model 1}} - LL_{\text{model 0}}) = 2(-2293.492 - (-2359.033)) = 131.082$$

This is a very high number for a chi square distribution with two degrees of freedom, so we can reject the null hypothesis. In this test, the null hypothesis is that $\beta_1 = \beta_2 = 0$ and our alternative hypothesis is that at least one of the coefficients is not equal to 0. In other words, model 1 is better than a model with just an intercept (model 0).

5. Consider a model for this data where strikes enters as a linear term

$$\hat{p} = \frac{e^{-0.99461+0.0627*\text{strike}}}{1+e^{-0.99461+0.0627*\text{strike}}}$$

$$\text{So, for a prisoner with 1 strike:}$$

$$\hat{p} = \frac{e^{-0.99461+0.0627*1}}{1+e^{-0.99461+0.0627*1}} = 0.283$$

$$\text{And for a prisoner with 3 strikes:}$$

$$\hat{p} = \frac{e^{-0.99461+0.0627*3}}{1+e^{-0.99461+0.0627*3}} = 0.309$$

6. Relative odds using model 2

$$\hat{OR} = \frac{e^{-0.99461+0.0627*3}}{e^{-0.99461+0.0627*1}} = e^{0.0627*(3-1)} = 1.134$$

$$95\%\text{ CI lower} = e^{0.0627*2-1.96(0.04439*2)} = 0.953$$
$$95\%\text{ CI upper} = e^{0.0627*2+1.96(0.04439*2)} = 1.349$$

An increase of two strikes (from 1 to 3) raises the risk of a misconduct violation 1.13-fold (95% CI: 0.953,1.349).

7. Which model is better, Model 2 or Model 1?

Grouped LL and AIC for model 1

$$LL_{\text{grouped}} = ln\binom{2416}{619} + ln\binom{771}{355} + ln\binom{731}{162} - 2293.492 = -10.870$$

This needs to be calculated using R's lchoose() function

```
lchoose(2416,619) + lchoose(771,355) + lchoose(731,162) - 2293.492
```

```
## [1] -10.86956
```

Check with logLik() function.

```
mod1 <- glm(cbind(y,n) ~ strikes,dat,family=binomial)
logLik(mod1)
```

```
## 'log Lik.' -10.86998 (df=3)
```

$$\text{AIC}_{\text{model 1}} = 2k - 2LL = 6 - (2*(-10.870)) = 27.74$$

Check with R:

```
summary(mod1)
```

```
##
## Call:
## glm(formula = cbind(y, n) ~ strikes, family = binomial, data = dat)
##
## Deviance Residuals:
## [1]  0  0  0
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.06577    0.04660 -22.868   <2e-16 ***
## strikes2     0.90720    0.08598  10.551   <2e-16 ***
## strikes3    -0.19052    0.10051  -1.895    0.058 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.3108e+02  on 2  degrees of freedom
## Residual deviance: 6.3061e-14  on 0  degrees of freedom
## AIC: 27.74
##
## Number of Fisher Scoring iterations: 2
```

Model 1 is better than model 2 based on AIC, because it has the lower AIC value and the difference is large enough to be considered significant (127.1).

8. Multiple covariate model interpretation

Exponentiate the coefficient and SE.

$$\text{OR}_{\text{nomaxsec}} = e^{\beta_{\text{score}}} = e^{0.0300} = 1.030$$
$$\text{CI lower, no max security} = e^{0.0300-1.96(0.00315)} = e^{0.023826} = 1.024$$
$$\text{CI upper, no max security} = e^{0.0300+1.96(0.00315)} = e^{0.036174} = 1.037$$
$$\text{OR}_{\text{maxsecurity}} = e^{\beta_{\text{score}}+\beta_{\text{scoremaxsecurity}}} = e^{0.0300-0.0356} = 0.994$$
$$\text{CI lower, max security} = e^{0.0300-0.0356-1.96(\sqrt{(9.923E-6+0.000052+2(-9.92E-6))})} = e^{-0.8574474} = 0.982$$
$$\text{CI upper, max security} = e^{0.0300-0.0356+1.96(\sqrt{(9.923E-6+0.0.000052+2(-9.92E-6))})} = e^{0.8462474} = 1.007$$

There is a significant association between classification score and misconduct violations in the first year of incarceration (p <0.0001), and a significant interaction between score and whether or not someone is incarcerated in a maximum security prison (p <0.0001). On average, for someone not incarcerated in a maximum security prison, the odds of a violation increase 1.03 times (95% CI: 1.024,1.037) for each 1 unit increase in classification score. For someone in a maximum security prison, the odds of a violation increase 0.994 times (95% CI: 0.982,1.007) for each 1 unit increase in classification score. Because the CI contains 1, this relationship is not statistically significant.

Code

```
# Long data
dat_long <- reshape(dat,direction='long',varying=c('y','n'),
                    v.names='count',timevar='misconduct',times=1:0)
# Replicated data
dat_longrep <- dat_long[rep(1:6,dat_long$count),c('strikes','misconduct')]
# Check log-likelihood for model 1 (ungrouped)
ungroup_mod1 <- glm(misconduct ~ strikes,dat_longrep,family = binomial)
logLik(ungroup_mod1)
```

```
## 'log Lik.' -2293.492 (df=3)
```

```
# LRT
ungroup_mod0 <- glm(misconduct ~ 1,dat_longrep,family = binomial)
anova(ungroup_mod0,ungroup_mod1,test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: misconduct ~ 1
## Model 2: misconduct ~ strikes
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      3917     4718.1
## 2      3915     4587.0  2   131.08 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
1 - pchisq(131.082,2)
```

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

```
# Check model 2 predictions
mod2 <- glm(formula = cbind(y,n) ~ as.numeric(strikes), family = binomial,
            data = dat)
predict(mod2, newdata=list(strikes=c(1,3)),type = "response")
```

```
##         1         2
## 0.2825393 0.3086388
```