BIOS 6612 Homework 2: Logistic Regression
Solutions

The California Department of Corrections (CDC) has developed a "classification score" to predict whether a prisoner will commit misconduct violations during incarceration. A study of 3918 inmates was performed to examine whether this classification score, determined at sentencing, is associated with subsequent misconduct violations during the first year of incarceration. Seven hundred thirty (730) of the 3918 inmates were incarcerated in maximum security prisons. In addition, the number of felony convictions or "strikes" was recorded for each prisoner. A "1 Strike" inmate is a prisoner who is serving time for a first felony conviction. A "2 Strikes" inmate is a prisoner who is serving time for a second felony and who was sentenced under a California law mandating sentence length enhancements. A "3 Strikes" inmate is a prisoner who is serving time for a third felony, in which case that same law mandated a life sentence.

We will work with these variables:

- `strikes`: number of felony convictions ("strikes": 1, 2, or 3)
  - `strikes2`: inmate had 2 strikes (0 = No, 1 = Yes)
  - `strikes3`: inmate had 3 strikes (0 = No, 1 = Yes)
- `misconduct`: Committed a misconduct violation during the first year of incarceration (0 = No, 1 = Yes)

The following table provides the number of prisoners with misconduct violations during the first year of incarceration by the number of felony convictions or "strikes" against them.

`strikes`	`misconduct=1`	`misconduct=0`
1	619	1797
2	355	416
3	162	569

Answer the following questions, showing your calculations; you may check your work using SAS or R.

Note: I have left numerical answers with more digits than you should to make it easier to compare answers; **make it a practice to round to 3 or 4 decimal places when reporting results.**

1. (6 points) Calculate estimates of $\beta_0, \beta_1, \beta_2$ for the logistic regression model

$$\text{logit}\, P(\text{misconduct violation}) = \beta_0 + \beta_1 \times \texttt{strikes2} + \beta_2 \times \texttt{strikes3}.$$

This is **Model 1**.

Answer: The intercept estimate is the log odds of misconduct in the reference group (strike 1 group); the two slope parameter estimates can be found by taking differences of log odds between the strike 2 and strike 3 groups with the strike 1 group.

$$\hat{\beta}_0 = \log(619/1797) = -1.0658$$
$$\hat{\beta}_1 = \log(355/416) - \log(619/1797) = -0.1586 - (-1.0658) = 0.9072$$
$$\hat{\beta}_2 = \log(162/569) - \log(619/1797) = -1.2563 - (-1.0658) = -0.1905$$

2. (3 points) Calculate the log-likelihood for Model 1.

Answer: There are two ways to do this, grouped and ungrouped. The data in tabular form above are grouped, so their log-likelihood may be calculated based on the binomial probability mass functions as

$$\log\left\{\binom{2416}{619}\left(\frac{619}{2416}\right)^{619}\left(1-\frac{619}{2416}\right)^{2416-619}\right\}+$$
$$\log\left\{\binom{771}{355}\left(\frac{355}{771}\right)^{355}\left(1-\frac{355}{771}\right)^{771-355}\right\}+$$
$$\log\left\{\binom{731}{162}\left(\frac{162}{731}\right)^{162}\left(1-\frac{162}{731}\right)^{731-162}\right\}=$$
$$-3.985140 + (-3.546823) + (-3.338017) = -10.86998.$$

The combinatoric terms don't contain unknown parameters, so can be dropped without affecting their estimation, leading to the ungrouped log-likelihood

$$\log\left\{\left(\frac{619}{2416}\right)^{619}\left(1-\frac{619}{2416}\right)^{2416-619}\right\}+$$
$$\log\left\{\left(\frac{355}{771}\right)^{355}\left(1-\frac{355}{771}\right)^{771-355}\right\}+$$
$$\log\left\{\left(\frac{162}{731}\right)^{162}\left(1-\frac{162}{731}\right)^{731-162}\right\}=$$
$$-1374.8339 + (-532.0009) + (-386.6577) = -2293.492.$$

SAS seems to typically report the ungrouped version, while R will report the grouped version if it is given data in grouped format. The two forms are identical up to an additive constant.

3. (3 points) Calculate the log-likelihood for the null model (i.e., a model with only an intercept, β_0; this is **Model 0**).

   Answer: The MLE for this model is calculated by ignoring the groupings based on strikes. This is the marginal probability of misconduct:

$$\frac{619+355+162}{619+355+162+1797+416+569} = \frac{1136}{3918} = 0.2899438.$$

   As before, we can calculate either the grouped

$$\begin{aligned} \log\left\{\binom{2416}{619}(0.2899438)^{619}(1-0.2899438)^{2416-619}\right\}+ \\ \log\left\{\binom{771}{355}(0.2899438)^{355}(1-0.2899438)^{771-355}\right\}+ \\ \log\left\{\binom{731}{162}(0.2899438)^{162}(1-0.2899438)^{731-162}\right\} = \\ -10.82830+(-53.50318)+(-12.07935) = -76.41084 \end{aligned}$$

   or ungrouped forms:

$$1136\log\left(\frac{1136}{3918}\right)+(3918-1136)\log\left(1-\frac{1136}{3918}\right) = -2359.033$$

4. (6 points) Perform a likelihood ratio test comparing Model 1 with Model 0. Describe what this is testing: what is the null hypothesis, and what does it mean to reject the null hypothesis?

   Answer: This can be done with either the grouped or ungrouped log-likelihoods for each model. With the grouped data, the likelihood ratio statistic is

$$2(-10.86998+76.41084) = 131.0817,$$

   while with the ungrouped data this is

$$2(-2293.492+2359.033) = 131.082.$$

   We see from this that the likelihood ratio test statistic is the same regardless of whether we have grouped or ungrouped binary data. Now we need to compare this with a reference chi-square distribution. The degrees of freedom for the test is equal to the difference in dimension of the two models: we have three parameters in Model 1 and 1 in Model 2, so we need to look at the chi-square with 2 degrees of freedom. This has 5% critical value of 5.991465, so with the observed value of 131.0817, we reject the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ and conclude that the number of felony convictions is significantly associated with the odds of misconduct violations at the 0.05 significance level.

5. Consider a model for this data where `strikes` enters as a linear term rather than categorical; this is **Model 2**. This model fit produces the following R output:

```
Call:
glm(formula = cbind(y, n - y) ~ strikes, family = binomial,
    data = miscond)

Deviance Residuals:
     1       2       3
-2.903   9.647  -5.254

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.99461    0.07872 -12.635   <2e-16 ***
strikes      0.06270    0.04439   1.413    0.158
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 131.08  on 2  degrees of freedom
Residual deviance: 129.10  on 1  degrees of freedom
AIC: 154.84

Number of Fisher Scoring iterations: 4
```

(4 points) Using Model 2, what is the predicted probability of a misconduct violation during the first year in prison for a prisoner with 1 strike? With 3 strikes?

Answer: The predicted probabilities can be calculated using the coefficient estimates and our knowledge of the logistic link function. The probability of misconduct with 1 strike is

$$\frac{\exp(-0.9946 + 1 \times 0.0627)}{1 + \exp(-0.9946 + 1 \times 0.0627)} = 0.2825394.$$

With 3 strikes the probability of misconduct is

$$\frac{\exp(-0.9946 + 3 \times 0.0627)}{1 + \exp(-0.9946 + 3 \times 0.0627)} = 0.3086368.$$

6. (5 points) Using Model 2, what are the relative odds of a misconduct violation during the first year in prison for a prisoner with 3 strikes compared to a prisoner with 1 strike? Calculate a 95% confidence interval for this estimate.

Answer: This question is asking us to calculate the odds ratio comparing a prisoner with 3 strikes with one with 1 strike. This is equal to

$$\exp((3-1) \times 0.0627) = \exp(2 \times 0.0627) = 1.133613.$$

To get a 95% confidence interval, we start on the log-odds scale since this is the scale the model is estimated on. On this scale, the 95% confidence interval for the regression parameter is equal to $0.0627 \pm 1.96 \times 0.0444 = (-0.024324, 0.149724)$. We want to apply the function $\exp(2\cdot)$ to our coefficient estimate, so we apply this to the endpoints of this confidence interval to get the confidence interval for the odds ratio as $(0.9525164, 1.3491139)$.

7. (4 points) Which model is better, Model 2 or Model 1? Justify your answer.

   Answer: We should answer this using AIC; the likelihood ratio test does not apply here because these models are not nested. They would only be nested if we were adding the categorical version of `strikes` to a model that also included the linear version (and one reason not to do this is multicollinearity). AIC is given as output here for Model 2, as 154.84. For Model 1 above, it is $-2 \times -10.86998 + 2 \times 3 = 27.73996$, much lower than for Model 2, showing a significantly better fit for the categorical version of `strikes`. The effect of the number of strikes on log-odds of a misconduct violation therefore should not be treated as linear.

8. For this question, you will need to interpret results from a model including some different covariates. One of these is `score`, the CDC classification score, which ranges from 0 to 80 and is used to predict whether prisoners will have misconduct violations. The other is `maxsecurity`, an indicator variable equal to 1 if the inmate was incarcerated in a maximum security prison and 0 otherwise. Model output from SAS appears below.

```
                The LOGISTIC Procedure

                       Model Information

         Data Set                        WORK.PRISON
         Response Variable               misconduct
         Number of Response Levels       2
         Model                           binary logit
         Optimization Technique          Fisher's scoring

            Number of Observations Read        3918
            Number of Observations Used        3918


                        Response Profile

              Ordered                          Total
                Value      misconduct      Frequency
                    1               1           1136
                    2               0           2782

              Probability modeled is misconduct=1.
```

```
                  Analysis of Maximum Likelihood Estimates

                                          Standard          Wald
Parameter              DF     Estimate       Error    Chi-Square    Pr > ChiSq
Intercept               1      -1.6532      0.0864      366.2619        <.0001
score                   1       0.0300     0.00315       90.8322        <.0001
maxsecurity             1       1.3356      0.4346        9.4431        0.0021
score*maxsecurity       1      -0.0356     0.00721       24.3318        <.0001
```

```
                         Estimated Covariance Matrix

Parameter            Intercept        score    maxsecurity    scoremaxsecurity
Intercept             0.007463     -0.00024       -0.00746            0.000241
score                 -0.00024     9.923E-6       0.000241            -9.92E-6
maxsecurity           -0.00746     0.000241       0.188901            -0.00296
scoremaxsecurity      0.000241     -9.92E-6       -0.00296            0.000052
```

(10 points) **Use the results of this model fit to provide a complete interpretation of the association between classification score and misconduct violations during the first year of incarceration.**

Answer: Note that there is a significant interaction between classification score and whether or not the felon was in a maximum security prison, so the effect of classification score should be interpreted separately for those in a maximum security prison and those not in a maximum security prison.

- Effect of classification score for prisoners NOT in Maximum Security
  - Odds ratio is $\exp(0.030) = 1.030$
  - 95% CI is $\exp(0.030 \pm 1.96 \times 0.00315) = (1.024, 1.037)$
  - p-value is < 0.0001
- Effect of classification score for prisoners in Maximum Security
  - Odds ratio is $\exp(0.030 - 0.0356) = \exp(-0.0056) = 0.994$
  - Standard error is $\sqrt{0.00315^2 + 0.00721^2 + 2 \times -0.00000992} = 0.006486$ (needs to account for the variance of both coefficient estimates plus their covariance)
  - 95% CI is $\exp(-0.0056 \pm 1.96 \times 0.006486) = (0.982, 1.007)$
  - Z score is $.0056/.006486 = 0.863$, so $p > 0.05$

Conclusion: The association between classification score and misconduct violations during the first year of incarceration depends on whether the prisoner is in a maximum security prison (interaction $p < 0.0001$). For prisoners in maximum security prisons, the classification score is not significantly associated with the odds of a misconduct violation during the first year of incarceration ($p > 0.05$). For prisoners not in maximum security prisons, the classification score is significantly associated with the odds of a misconduct violation during the first year of incarceration ($p < 0.0001$); the odds of a misconduct violation increases 3.0% (95% CI: 2.4 to 3.7%) for every 1 point increase in classification score.