BIOS 6612 Homework 2: Logistic Regression

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.

1. 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**.

2. Calculate the log-likelihood for Model 1.

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

4. 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?

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
```

   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?

6. 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.

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

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 Value	misconduct	Total Frequency
1	1	1136
2	0	2782

Probability modeled is misconduct=1.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald 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

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.