BIOS 6612 Homework 2: Logistic Regression

1. (**40 points**) 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 R.

   (a) 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. (**6 points**)

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

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

(d) 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? (**6 points**)

(e) 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? (**4 points**)

(f) 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. (**4 points**)

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

2. (**15 points**) The Genetic Epidemiology of COPD (COPDGene) Study is a multi-center case/control study designed to identify genetic factors associated with COPD and to characterize COPD-related phenotypes. The study recruited COPD cases and smoking controls ages 45 to 80 with at least 10 pack-years of smoking history.

The `copd.txt` file contains the COPD status (`copd=1` if the subject has COPD and `0` otherwise), age, gender (`gender=0` for males and `1` for females), current smoking status (`smoker=1` if the subject is a current smoker and `smoker=0` if the subject is a former smoker), mean centered BMI (labeled `BMI`), mean centered BMI squared (labeled `BMIsquared`).

Answer the questions below and provide the relevant code and output in the appendix at the end of the assignment. Do not include all of the output, only the output that pertains to the questions below.

Note: All models should include age, gender, current smoking status, and BMI as covariates; you will need to evaluate the inclusion of BMI squared.

(a) Provide a Wald test statistic and p-value to determine whether COPD is significantly associated with BMI squared. (**5 points**)

(b) Provide a likelihood ratio test statistic and p-value to determine whether COPD is significantly associated with BMI squared. (**5 points**)

(c) Based on your answers to the previous questions, is there evidence that COPD has a quadratic relationship with BMI? (**3 points**)

(d) Why do you think the BMI variable was centered? (**2 points**)

3. (**25 points**) Rickert et al. (*Clinical Pediatrics* 1992; p. 205) designed a study to evaluate whether an HIV educational program makes sexually active adolescents more likely to obtain condoms ($Y = 1$ if the adolescent obtained condoms and 0 otherwise). Adolescents were randomly assigned to different groups, according to whether education in the form of a lecture and video about the transmission of the HIV virus was provided. In a logistic regression model, factors observed to influence a teenager's probability of obtaining condoms were gender, socioeconomic status, lifetime number of partners, and the experimental condition (treatment variable). Results from a single model were summarized in a table such as the following. *This table contains at least one mistake.*

Variable	OR	95% Wald CI
`group` (none [ref.] vs. education)	4.04	$(1.17, 13.9)$
`gender` (female [ref.] vs. male)	1.38	$(1.23, 12.88)$
`SES` (low [ref.] vs. high)	5.82	$(1.87, 18.28)$
Lifetime number of `partners`	3.22	$(1.08, 11.31)$

(a) Interpret the odds ratio and the corresponding confidence interval for group. (**5 points**)

(b) Calculate the parameter estimates for the fitted logistic regression model. That is, find $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$ for the model

$$\text{logit}\, P(Y_i = 1) = \beta_0 + \beta_1 \texttt{group}_i + \beta_2 \texttt{gender}_i + \beta_3 \texttt{SES}_i + \beta_4 \texttt{partners}_i.$$

(**4 points**)

(c) What additional piece of information would you need to obtain an estimate for the intercept β_0? (**2 points**)

(d) Based on the corresponding Wald 95% confidence interval for the log odds ratio, determine the standard error for the `group` effect, i.e., $\text{SE}(\hat{\beta}_1)$. (**5 points**)

(e) Argue that either the estimate of 1.38 for the odds ratio for gender or the corresponding confidence interval is incorrect. Show that, if the reported interval is correct, 1.38 is actually the log odds ratio and the estimated odds ratio approximately equals 3.97. (**7 points**)