BIOS 6612 Homework 2: Logistic Regression
Solutions

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.

   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.**

(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**)

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.

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

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

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.

(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**)

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:

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

(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**)

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.

(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**)

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

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

(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**)

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 $(0.9525164, 1.3491139)$.

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

Answer: Although these models are technically nested (so we can use a likelihood ratio test to compare them), the rationale for this is something we won't get to in detail. We can always use AIC to compare models, however, regardless of nesting. The 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.

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**) Yes, test statistic is 7.308759 and p-value is 0.006861932

(b) Provide a likelihood ratio test statistic and p-value to determine whether COPD is significantly associated with BMI squared. (**5 points**) Yes, test statistic is 7.952939 and p-value is 0.004800934

(c) Based on your answers to the previous questions, is there evidence that COPD has a quadratic relationship with BMI? (**3 points**) Yes, first two answers suggest that the effect of BMI squared on odds of COPD is significant.

(d) Why do you think the BMI variable was centered? (**2 points**) To avoid multicollinearity because BMI is likely to be highly correlated with BMI squared.

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**) The odds that adolescent in the HIV educational program obtains condoms is estimated to be 4.04 times that for an adolescent not in the program, controlling for gender, SES, and lifetime number of partners. The interpretation of the confidence interval is that there is a 95% chance that the interval shown contains the true odds ratio, but NOT a 95% chance that the true odds ratio is within this interval. The true odds ratio is fixed, so it is either within this interval or not (0% or 100% chance). However, if we repeated this experiment or study 100 times and calculated a 95% confidence interval for this odds ratio each time, then we would expect that 95 such intervals would contain the true odds ratio. This is referred to as coverage probability, and is often used in simulations to evaluate performance of an estimation procedure: you want this to be as close to the nominal level (95%) as possible.

(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**)

$$\hat{\beta}_1 = \log(4.04) = 1.40$$
$$\hat{\beta}_2 = \log(1.38) = 0.32$$
$$\hat{\beta}_3 = \log(5.82) = 1.76$$
$$\hat{\beta}_4 = \log(3.22) = 1.17$$

(c) What additional piece of information would you need to obtain an estimate for the intercept β_0? (**2 points**) Since the intercept is the logit of the probability that $Y_i = 1$ for someone with a zero vector for the covariates, we would be able to calculate the intercept estimate if we knew the sample proportion of adolescents not in the education group, with female gender, low SES, and 0 lifetime partners.

(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**) The Wald 95% confidence interval for the group effect is $(\log(1.17), \log(13.9)) = (0.157, 2.632)$. The width of the CI on the log odds scale is $2 \times 1.96 \times \text{SE}(\hat{\beta}_1)$, so we have $2 \times 1.96 \times \text{SE}(\hat{\beta}_1) = 2.632 - 0.157 = 2.475$. Therefore, $\text{SE}(\hat{\beta}_1) = 2.475/(2 \times 1.96) = 0.63$.

(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**) The Wald 95% confidence interval for the gender effect is $(\log(1.23), \log(12.88)) = (0.207, 2.556)$. For 95% Wald confidence intervals, the estimate should be in the center of the interval *on the log-odds scale*. The center of the confidence interval for the gender effect is $(2.556 - 0.207)/2 + 0.207 = 1.38$. However, from the table the regression coefficient estimate is $\log 1.38 = 0.32$, which is not in the center of the interval. If the confidence interval is correct, then the estimated odds ratio for gender would be $exp(1.38) = 3.97$.