BIOS 6612 Homework 3: Logistic Regression
Solutions

1. **Background**: 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. An article detailing the COPDGene study design is included in the HW3 folder on Canvas (`COPDGene.pdf`).

   Dataset: The `hw3.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) Another way to look at the value of a covariate in a regression model is to assess its influence on predictive accuracy. AUC (area under the ROC curve), also known as the c index or concordance index, is one way to measure predictive accuracy. Calculate the AUC for each of these models. Note: This is given by default in SAS as part of the model fit summary; if you are using R, then the function `auc()` in the `pROC` package will give very similar results. (**5 points**) Area under the curve for reduced model: 0.705. Area under the curve for full model: 0.7121

   (d) Based on your answers to the previous questions, is there evidence that COPD has a quadratic relationship with BMI? (**2 points**) Yes, first two answers suggest

that the effect of BMI squared on odds of COPD is significant, last suggests that the model predictions are better with BMI squared included

(e) 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.

(f) Using the full model (that is, including BMI squared), calculate and interpret the estimated odds ratio for the effect of BMI on COPD **for a patient with average BMI**. Construct a 95% confidence interval for this odds ratio using both the Wald and likelihood ratio procedures; are these confidence intervals similar to one another? Why or why not? (**9 points**) For patients at average BMI, the coefficient of the quadratic term will be irrelevant because centered BMI squared is equal to 0, so the effect will be captured entirely by the coefficient of the linear term. This odds ratio is estimated as $\exp(-0.045133) = 0.9558707$, so there is approximately a 5% decrease in odds of COPD for each one-unit increase in BMI for someone with average BMI. This means that higher BMI has a slight protective effect with respect to COPD. Exponentiate the endpoints of the confidence intervals on the log odds scale: Wald CI is 0.9279670 - 0.9846134, LR CI is 0.9276001 - 0.9843026; these are similar because we have a large sample, so asymptotic theory holds.

2. 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)

$$\hat{\beta}_1 = \log(4.04) = 1.04$$
$$\hat{\beta}_2 = \log(1.38) = 0.32$$
$$\hat{\beta}_3 = \log(5.82) = 1.76$$
$$\hat{\beta}_4 = \log(3.22) = 1.17$$

(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. **(8 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 $exp(1.38) = 3.97$.