BIOS 6612 Homework 3: Logistic Regression

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.

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

   (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.

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

   (e) Why do you think the BMI variable was centered?

   (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?

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.

(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

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

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

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

(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.