BIOS 6612 Homework 3: Case-control data and other link functions for binary response data

1. (**30 points + 5 extra credit**) Bacterial growth in donated blood can lead to severe infections in transfusion patients. In the absence of limiting factors, bacteria growth follows a power law, with a strain-specific doubling time of θ in some arbitrary time unit.

   Suppose we have $n = 20$ bags of donated blood. Each contains 100 mL of blood and is infused with a common strain of bacteria at time 0. At $t = 8, 12, 16$ hours subsequent to time 0, a 1-mL sample is drawn from each bag and tested for the presence or absence of bacteria. The table below shows the number of bags testing positive at each time point.

Time	Num. bags
(hours)	positive
8	5
12	10
16	19

   Assume that the actual number of bacteria present in a sample of (known) fraction k at time t follows a Poisson distribution with mean $k \cdot 2^{t/\theta}$. Let $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3})'$ be the indicators of presence ($Y_{ij} = 1$) or absence ($Y_{ij} = 0$) of bacteria at each of the three observation times for the ith blood bag.

   (a) Give the likelihood for the data as a function of θ. (**8 points**)

   (b) Show that the MLE of θ can be obtained from a GLM with a binary response, covariate of t_j, complementary log-log link function, offset equal to $\log k$, and no intercept. (**8 points**)

   (c) Fit this model using standard software for generalized linear models; provide your code and model output. Note that $k = 1/100$ since each sample is 1 mL drawn from the original volume of 100 mL in each bag. (**8 points**)

   (d) Using your model output, give the MLE of θ and an associated 95% confidence interval. (**6 points**)

   (e) Compute a measure of model deviance that can be approximated by a χ^2 distribution and assess the fit of the model. (**Extra Credit, 5 points**)

2. (**40 points**) We can use logistic regression to estimate the odds ratio as a measure of association between a binary exposure and a binary outcome regardless of whether cohort or case-control sampling is used. For this question, you will need to conduct a simulation study to verify this result. The parameters needed for this simulation are given below:

   - Population size N =100,000
   - Prevalence of disease
     - 5% in unexposed
     - 10% in exposed
   - Prevalence of exposure is 30%

   (a) Give the parameters needed to simulate the outcome conditional on exposure, assuming a logistic regression model in the population. (**6 points**)

   (b) Generate data on exposure and outcome for the entire population. *Hint:* In R, you can use the function `rbinom(N,1,prob)` to generate `N` independent Bernoulli variables with mean `prob`. (**9 points**)

   (c) Now that you have a population, conduct 1000 simulations where you randomly select $n = 50$ individuals with the exposure and another $n = 50$ without the exposure; this represents a cohort study. For each simulation, fit the appropriate logistic regression model and record the parameter estimates. Give both the mean and median of each the estimates of each parameter across all simulations. Why might the estimates not agree with one another across simulated samples? How do they compare to the true estimate values, and why might they differ? (**10 points**)

   (d) Repeat the previous question for a case-control design: that is, randomly sample $n = 50$ individuals with the disease and $n = 50$ without the disease, then fit the appropriate logistic regression model. As before, record the estimated values of the parameters for each of your 1000 simulations. Give the mean and median of the estimates across all simulations. (**8 points**)

   (e) Both the cohort and case-control designs are estimating the same odds ratio between exposure and disease, but their efficiency might be different. Compare the variability of the estimated log odds ratios between the two designs and explain any differences you find. *Hint:* the `mad()` command in R computes a robust estimate of spread and might be useful if you find large differences between mean and median estimated parameter values. (**7 points**)