MATH 8850

Homework Assignment 2

PROBLEM 1: Many applications arise in which testing has to be performed in order to complete classification. In this problem we will investigate a strategy first proposed by Dorfman (1943), which is commonly referred to as group (or pool) testing. The conceptualization of this strategy was centered around the need to screen WWII inductees for syphilis. The premise, at this time, most of the soldiers inducted into the armed forces would not be infected with syphilis; i.e., this disease had a relatively low prevalence. Thus, if subjects were tested one-by-one then most of the observed testing responses would be negative. Question, why spend so much money on testing people who are likely negative? And the simple answer is we have to identify those who are positive. Dorfman's solution, rather than test everyone individually, collect and amalgamate specimens (e.g., blood, urine, saliva, etc.) to form pooled specimens which could be tested. If a pool tests negative, then all contributing specimens (i.e., individuals) are diagnosed negative, at the expense of only one test. If a pool tests positive, further testing would have to be completed in order to identify which individuals in the pool were positive; e.g., simply retest each specimen contributing to positive pools individually. If the disease has a low prevalence within the population, it is relatively easy to see that most pools will test negative, and hence a great deal of savings in testing cost can be realized. Thompson (1967) repurposed the use of group testing as a means to more effectively collect (i.e., at a reduced cost of testing) data for the purposes of estimating population level characteristics; e.g., the prevalence of the binary characteristic of interest. This problem will consider Thompson's approach to data collection and estimation. Let $Y_{ij} = 1$ denote that the ith individual assigned to the jth pool posses the binary characteristic of interest, and $Y_{ij} = 0$ otherwise, for $i = 1, ..., c_j$ and $j = 1, ..., J$, where c_j denotes the size of the jth pool and J denotes the number of pools. We will assume that $Y_{ij} \overset{iid}{\sim} \text{Bernoulli}(p)$, where p denotes the prevalence of the binary characteristic. Our goal is to estimate the prevalence: i.e., p. In order to save cost, we decide not the measure the Y_{ij} directly, but rather decide to pool specimens and take measurements on the pools; i.e., we will observe $Z_j = I\left(\sum_{i=1}^{c} Y_{ij} > 0\right)$, where $Z_j = 1$ denotes the event that the jth pool consists of at least one positive, and $Z_j = 0$ otherwise. Note, we do not get to see the Y_{ij}, they are latent.

(a) Under the simplifying assumption that a common pool size is used (i.e., $c_j = c$ for all j) it is possible to derive the MLE of p, in terms of the Z_j. Derive the MLE of p, establish a central limit theorem without appealing to the MLE theory, and use these results to propose an asymptotic confidence interval for p. Write an R function that accepts 3 inputs: 1) a vector containing the Z_j, a scalar specifying the pool size, and a user specified significance level α. This function should then output the estimated MLE, its asymptotic standard error, and a $(1-\alpha)100\%$ asymptotic confidence interval.

(b) Here we will investigate the performance of the MLE, in terms of being an estimator of p, across a variety of sample sizes J, pool sizes c, and values of p. Specifically, consider $p \in \{0.001, 0.002, ...0.2\}$, $c \in \{4, 6, 8, 10\}$, and $J = \{25, 50, 100, 200\}$. For each combination of (p, c, J) generate 1000 group testing data sets. Hint: for a specified combination of (p, c, J) one can generate a data set using the following code:

```
J<-??
c<-??
p<-??
N<-J*c
stat.mat<-matrix(rbinom(N,1,p),nrow=J,ncol=c)
Zj<-apply(stat.mat,1,max)
```

From the results of this study you will construct 3 figures each of which have 4 panes, one each for the different values of J: Within each pane of Figure 1, plot an estimate of the MLE's bias vs. p, for the different values of c; i.e., there should be 4 curves plotted in this figure, one each for the different values of c. Within each pane of Figure 2, plot the difference in the average estimated asymptotic standard error and the sample standard deviation of the MLEs vs. p, for the different values of c. Within each pane of Figure 3, plot the empirical coverage probabilities for the asymptotic confidence interval (at the $\alpha = 0.05$ significance level) vs. p, for the different values of c. Write a brief discussion of your findings.

(c) Making things more realistic. In many applications, the use of a common pool size is not practical (i.e., $c_j \neq c_{j'}$ for $j \neq j'$). In this situation, an analytical expression for the MLE of p is not available. Further, the observed testing responses can be subject to measurement error; i.e., a specimen tests positive when it is truly negative and vice versa. Typically, the accuracies of a test are quantified through two quantities sensitivity (S_e) and (S_p), where S_e (S_p) denotes the probability that a truly positive (negative) specimen will test positive (negative). To acknowledge that testing error exists, we denote the observed testing responses as $\tilde{Z}_j$, where $\tilde{Z}_j = 1$ denotes the event that the jth pool **tests** positive, and $\tilde{Z}_j = 0$ otherwise. Derive the likelihood for the observed data $\{(\tilde{Z}_j, c_j), j = 1, ..., J\}$. Using the likelihood, in R code a function that can be used to find the MLE of p, an estimate of its asymptotic variance, and an asymptotic confidence interval. Hint: it might be useful to look at the optimize function in R. This function should then output the estimated MLE, its asymptotic standard error, and a $(1 - \alpha)100\%$ asymptotic confidence interval.

d.) Using your results from part (c), perform the simulation study described in (b), with two alterations: 1) for each data set randomly generate pool sizes c_j as 2 plus a Poisson random variable whose mean is λ, take $\lambda \in \{2, 4, 6\}$; 2) take the testing accuracies to be $S_e = 0.95$ and $S_p = 0.98$. Hint: given a vector of the trues statuses of the pools the following code can be used to generate the error laden observed testing responses:

```
Se<-??
Sp<-??
tZj<-rbinom(N,1,(Se*Zj + (1-Sp)*(1-Zj)))
```

Using the results of this study, construct the 3 figure discussed in (b) and discuss your findings.

PROBLEM 2: Vansteelandt et al. (2000) extended the seminal work of Thompson (1967) to the regressions setting; i.e., instead of simply estimating a population level prevalence, we are now going to be interested in relating covariate information to the infection statuses of the individuals. As before, let $Y_{ij}|\mathbf{x}_i \overset{ind}{\sim} \text{Bernoulli}\{p(\mathbf{x}_{ij})\}$, where $p(\mathbf{x}_{ij}) = \exp(\mathbf{x}'_{ij}\boldsymbol{\beta})/\{1+\exp(\mathbf{x}'_{ij}\boldsymbol{\beta})\}$, $\mathbf{x}_i = (1, x_{ij1}, ..., x_{ijp})'$ is a vector of predictor variables, and $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ is the corresponding vector of regression coefficients. Note, the primary change here is that each individual has their own, predictor specific, probability of being positive. Our goal is to estimate $\boldsymbol{\beta}$. In order to save cost, we decide not the measure the Y_{ij} directly, but rather decide to pool specimens and take measurements on the pools.

(a) Derive the likelihood of the observed data, your likelihood should be general enough to accommodate imperfect testing and pool sizes that vary from group to group. Using your likelihood, right and R function that accepts as inputs a vector of pool testing responses, the usual design matrix, a vector identifying the pool sizes, a scalar identifying the sensitivity of the test, a scalar specifying the specificity of the test, and the significance level α. Be careful to make sure you associate the covariates to the testing responses in the correct fashion. Based on these inputs, your function should return the following:

   (i) The MLEs of the regression parameters, their estimated asymptotic standard errors, and a $(1-\alpha)100\%$ asymptotic confidence interval for each of the regression coefficients.

   (ii) A table that summarizes the test statistics (both likelihood ratio and Wald) and p-values associated with testing $H_0 : \beta_k = 0$ vs. $H_1 : \beta_k \neq 0$, for $k = 0, 1, ..., p$, when all other predictors are in the model.

   (iii) Test statistics (both likelihood ratio and Wald) and their associated p-value for the test of $H_0 : \beta_1 = ... = \beta_p = 0$ vs. $H_1 :$ at least one $\beta_k \neq 0$.

(b) Derive a $(1-\alpha)100\%$ asymptotic confidence interval for the probability that an individual with covariates $\mathbf{x}_{ij}$ will be truly positive; i.e., a confidence interval for $p(\mathbf{x}_{ij})$. Then, write an R function that accepts 3 objects as inputs: 1) a matrix whose rows are the $\mathbf{x}_{ij}$ at which we wish to create confidence intervals for $p(\mathbf{x}_{ij})$, 2) the significance level α, and 3) the output from the function you wrote in part (a). The output from this function should be a matrix whose rows consist of the MLE of $p(\mathbf{x}_{ij})$ and the endpoints of the corresponding $(1-\alpha)100\%$ asymptotic confidence interval.

(c) Develop and conduct a simulation study that examines the performance of the proposed regression methodology, be succinct in summarizing your study. Hints: 1) you are asking the data to paint a much bigger picture and as such you may want to consider sample sizes of say $J > 200$, 2) group testing is only effective in situations in which the population level prevalence is small (e.g., if prevalence is 0.80 then you will likely only see positive pools), keep this in mind when you are choosing the regression parameters and covariate distributions, and 3) test error rates are usually near 1; e.g., $S_e = 0.95$ and $S_p = 0.98$.

(d) Using the techniques developed and vetted in parts (a)-(c), perform an analysis of the Chlamydia data (see the course webpage).