Biost/Stat 533
Emerson, Spr 14
Homework #4
April 30, 2014

Written problems to be handed in Monday, May 5.

Under no circumstances may you refer to a homework key from this or other classes. While you may work with other students to derive a solution, when you write up your final solution, you may not refer to any other source. You must be able to develop your answer as if it were being done in a closed book, closed notes examination. You must provide a signed pledge to that effect:

On my honor I have neither given nor received unauthorized aid on the completion of this homework.

1. Consider again the setting in which $Y_i \sim (\mu_0, \sigma^2)$ for $i = 1, \ldots, n_0$ and $Y_i \sim (\mu_1, \sigma^2)$ for $i = n_0 + 1, \ldots, n = n_0 + n_1 = 2n_0$, except observations within each group are correlated. That is, we have $Cov(Y_i, Y_j) = \rho\sigma^2$ for $i, j = 1, \ldots, n_0; i \neq j$, $Cov(Y_i, Y_j) = \rho\sigma^2$ for $i, j = n_0 + 1, \ldots, n; i \neq j$, and $Cov(Y_i, Y_j) = 0$ for $i = 1, \ldots, n_0; j = n_0 + 1, \ldots, n$. For notational convenience, let $\vec{w}$ be an n-vector such that $w_i = 1$ for $1 \leq i \leq n_0$ and $w_i = 0$ otherwise, and let $\vec{z} = \vec{1}_n - \vec{w}$. Consider linear regression model $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$ with $\mathbf{X} = (\vec{w} \quad \vec{z})$. We are interested in estimating $\vec{a}^T\vec{\beta} = \mu_1 - \mu_0$.

   a. Show that the ordinary least squares estimator $\widehat{\vec{\beta}}$ is equal to the generalized least squares estimator $\widehat{\vec{\beta}}_G$. What is the mean and variance of these estimators?

   b. Provide an estimate of the variance of $\widehat{\vec{\beta}}_G$ and $\vec{a}^T\widehat{\vec{\beta}}_G$ assuming that ρ is known.

   c. Provide an estimate of the variance of $\widehat{\vec{\beta}}$ and $\vec{a}^T\widehat{\vec{\beta}}$ under the assumption that the observations are independent. How do they compare to the answers in b?

2. Now consider the setting in which $Y_i \sim (\mu_0, \sigma^2)$ for $i = 1, \ldots, n_0$ and $Y_i \sim (\mu_1, \sigma^2)$ for $i = n_0 + 1, \ldots, n = n_0 + n_1 = 2n_0$, except observations are paired across groups. That is, we have $Cov(Y_i, Y_i) = \sigma^2$ for $i = 1, \ldots, n$, $Cov(Y_i, Y_{n_0+i}) = \rho\sigma^2$ for $i = 1, \ldots, n_0$, and $Cov(Y_i, Y_j) = 0$ otherwise. For notational convenience, let $\vec{w}$ be an n-vector such that $w_i = 1$ for $1 \leq i \leq n_0$ and $w_i = 0$ otherwise, and let $\vec{z} = \vec{1}_n - \vec{w}$. Consider linear regression model $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$ with $\mathbf{X} = (\vec{w} \quad \vec{z})$. We are interested in estimating $\vec{a}^T\vec{\beta} = \mu_1 - \mu_0$.

   a. Show that the ordinary least squares estimator $\widehat{\vec{\beta}}$ is equal to the generalized least squares estimator $\widehat{\vec{\beta}}_G$. What is the mean and variance of these estimators?

   b. Provide an estimate of the variance of $\widehat{\vec{\beta}}_G$ and $\vec{a}^T\widehat{\vec{\beta}}_G$ assuming that ρ is known.

   c. Provide an estimate of the variance of $\widehat{\vec{\beta}}$ and $\vec{a}^T\widehat{\vec{\beta}}$ under the assumption that the observations are independent. How do they compare to the answers in b?

   d. How does the effect of correlated observations affect an ordinary least squares analysis differ when the correlated observations are within groups sharing the same predictor values versus when the correlated observations have different predictor values?

3. Let $Y_i \sim Bernoulli(p_i), i = 1, \ldots, n$ be independent random variables with $p_i = \vec{x}_i^T\vec{\beta}$ for known predictor vector $\vec{x}_i$.

   a. Is inference about $\vec{\beta}$ using ordinary least squares regression analysis asymptotically valid for this problem? If so, provide justification. If not, are there any situations in which it might be approximately valid?

b. Describe an iterative approach in which weighted least squares might be used to address this problem. What undesirable small sample behavior with respect to the range of estimates $\hat{p}_i$ might persist under this analysis scheme?

4. Consider a linear regression model relating response $\vec{Y}$ to an intercept and two predictor vectors $\vec{W}$ and $\vec{Z}$ (so design matrix $\mathbf{X} = (\vec{1}_n \quad \vec{W} \quad \vec{Z})$ has $X_{i1} \equiv 1$ for $i = 1, \ldots, n$ and $X_{i2} = W_i$ and $X_{i3} = Z_i$ and $\vec{\beta} = (\beta_0, \beta_1, \beta_2)^T$). Assume $E[\vec{\epsilon}] = \vec{0}$ and $var(\vec{\epsilon}) = \sigma^2 \mathbf{I}_n$.

   a. Show that the correlation between OLS estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ is opposite in sign to the sample correlation between $\vec{W}$ and $\vec{Z}$ and that the two slope estimates are uncorrelated if the sample correlation between $\vec{W}$ and $\vec{Z}$ is zero.

   b. Suppose we hold $S_{WW} = (\vec{W} - E[\vec{W}])^T(\vec{W} - E[\vec{W}])$, S_{ZZ}, and σ^2 constant, but we may freely vary $S_{WZ} = (\vec{W} - E[\vec{W}])^T(\vec{Z} - E[\vec{Z}])$. For what value of S_{WZ} do we minimize the variance of $\hat{\beta}_1$ and $\hat{\beta}_2$? What does this suggest about our ability to test for an association between Y and W adjusting for Z when W and Z are correlated?

5. Consider again the linear regression model in which we will assume the true model is

$$\vec{Y} = \beta_0 + \vec{W}\beta_1 + \vec{Z}\beta_2 + \vec{\epsilon}$$

but we want to also consider fitting a model

$$\vec{Y} = \gamma_0 + \vec{W}\gamma_1 + \vec{\epsilon}^*$$

.

   a. Under what conditions is the OLS estimate $\hat{\beta}_1$ equal to the OLS estimate $\hat{\gamma}_1$?

   b. Under what conditions is the standard error of $\hat{\beta}_1$ equal to the standard error of $\hat{\gamma}_1$.

   c. Under what conditions is the <u>estimated</u> standard error of $\hat{\beta}_1$ equal to the <u>estimated</u> standard error of $\hat{\gamma}_1$.

   d. Under what conditions is $\hat{\gamma}_1$ unbiased for β_1?

   e. Under what conditions is $\hat{\gamma}_1$ BLUE for β_1?

   f. Suppose in particular that $\beta_1 = 0$ and $\beta_2 \neq 0$. What is the impact of this situation on the distribution of $\hat{\gamma}_1$, and how would $\hat{\gamma}_1$ compare to $\hat{\beta}_1$ from the full model? Compare this situation to the setting in which $\beta_2 = 0$ and $\beta_1 \neq 0$.