# Homework # 5, Due April 6, 2020

BIOS 617

Assigned on: March 23, 2020

1. The correlation coefficient between two variables $X$ and $Y$ in a population is given by

$$R = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{N}(Y_i - \bar{Y})^2}}$$

   (a) Show that $R$ can be written as a function of the following 5 population totals: $T_1 = \sum_{i=1}^{N} X_i$, $T_2 = \sum_{i=1}^{N} Y_i$, $T_3 = \sum_{i=1}^{N} X_i^2$, $T_4 = \sum_{i=1}^{N} Y_i^2$, $T_5 = \sum_{i=1}^{N} X_i Y_i$. **[10 pt]**

   (b) Use the Taylor series approximation to determine a variance estimator for a sample estimate of $R$ if I have 5 unbiased estimators of the population totals above, $t_1, t_2, t_3, t_4, t_5$. **[15 pt]**

2. Consider the weighted regression of dioxin on age, that is, suppose we want to estimate the population model

$$Y_i = B_0 + B_1 A_i + \epsilon_i, \quad \epsilon \sim N(0, \sigma^2), \quad i = 1, \ldots, N$$

   using the weighted least squares estimator

$$b_w = \left(X^\top W X\right)^{-1} X^\top W y$$

   where $X$ consists of the stacked vectors of the intercept and the sampled age and $W$ is the diagonal matrix of sampled weights. We will verify the estimator of $v(U_w(\mathbf{b}_w))$.

   (a) Using the data DIOXIN2.DAT, compute $(X^\top W X)^{-1}$ **[10 pt]**

   (b) Again using the data DIOXIN2.DAT, compute $\sigma^4 v(U_w(\mathbf{b_w}))$ where $\mathbf{b}_w$ is the estimator of the linear regression parameters of log(TCDD) on age and $U$ is the score function. **[10 pt]**

(c) Using the results of (a) and (b), compute $v(\mathbf{b}_w)$ and compare the square root of the diagonal elements of $v(\mathbf{b}_w)$ with the results on the slides **[10 pt]**

3. Show that, for a proportionally stratified sample with a paired selection design, the jackknife estimator of the population mean given by $v_{JRR2}(\hat{\theta}) = \frac{1}{4}\sum_{h=1}^{H}(\bar{y}_{(h1)} - \bar{y}_{(h2)})^2$ is equivalent to $v_{JRR}(\hat{\theta}) = \frac{1}{2}\sum_{h=1}^{H}\sum_{i=1}^{2}\left(\bar{y}_{(hi)} - \bar{y}\right)^2$. **[10 pt]**

4. Prove the result from class notes that plugging in $v_h$ into $C^\star V$ definition yields

$$V \approx \frac{1}{C^\star}\left[\sum_h P_h S_h \sqrt{c_h} + \left(S^2 - \sum_h P_h S_h^2\right)^2 \sqrt{c'}\right].$$

**[15 pt]**

5. Cochran 12.1 (to answer the question "does double sampling produce a gain in precision over single sampling", compute the design effect for a double sampling estimator compared with an SRS design of sample size $n$.) **[20 pt]**