STAT/BIOST 571: Homework 5

Philip Pham

February 21, 2019

Problem 1: Sandwich and bootstrap standard error estimates (10 points)

As on slide 2.76, fit the model

$$EY_{ij} = \beta_0 + \beta_1(Age_{ij} - 8) + \beta_2 Gender_i + \beta_3(Age_{ij} - 8) \times Gender_i$$

to the dental data by using REML, but use a homoscedastic covariance models with no correlation.

(a) Calculate sandwich-based standard error estimates for $\hat{\beta}_3$ that account for clustering by subject. Write your own code for this, using matrix algebra.

	Estimate	Standard Error
$\hat{\beta}_0$	22.615610	0.472075
$\hat{\beta}_1$	0.784380	0.126167
$\hat{\beta}_2$	-1.406521	0.739599
$\hat{\beta}_3$	-0.304834	0.197666

Table 1: Parameter estimates using REML with a homoscedastic covariance models with no correlation.

Solution: The REML estimates can be found in Table 1. Standard errors were calculated assuming that covariance model is specified correctly by taking the square root of the diagonal $\left(\sum_{i=1}^n X_i^\intercal \hat{\Sigma}_{\text{REML}}^{-1} X_i\right)^{-1}$, where $\hat{\Sigma}_{\text{REML}} = \hat{\alpha} I_m$, since cluster sizes are equal and there is only one covariance parameter on the diagonal.

Sandwich covariance estimates can be obtained by

$$\hat{\Sigma}_{\text{Sandwich}} = \left(\sum_{i=1}^n X_i^\intercal \hat{\Sigma}_{\text{REML}}^{-1} X_i\right)^{-1} \left(\sum_{i=1}^n X_i^\intercal \hat{\Sigma}_{\text{REML}}^{-1} \hat{\Sigma}_{\text{Empirical}} \hat{\Sigma}_{\text{REML}}^{-1} X_i\right) \left(\sum_{i=1}^n X_i^\intercal \hat{\Sigma}_{\text{REML}}^{-1} X_i\right)^{-1},$$

where the empirical covariance estimate is $\hat{\Sigma}_{\text{Empirical}} = \frac{1}{n} \sum_{i=1}^n \left(Y_i - X_i \hat{\beta}\right) \left(Y_i - X_i \hat{\beta}\right)^\intercal$ since we assume each cluster has the same covariance structure.

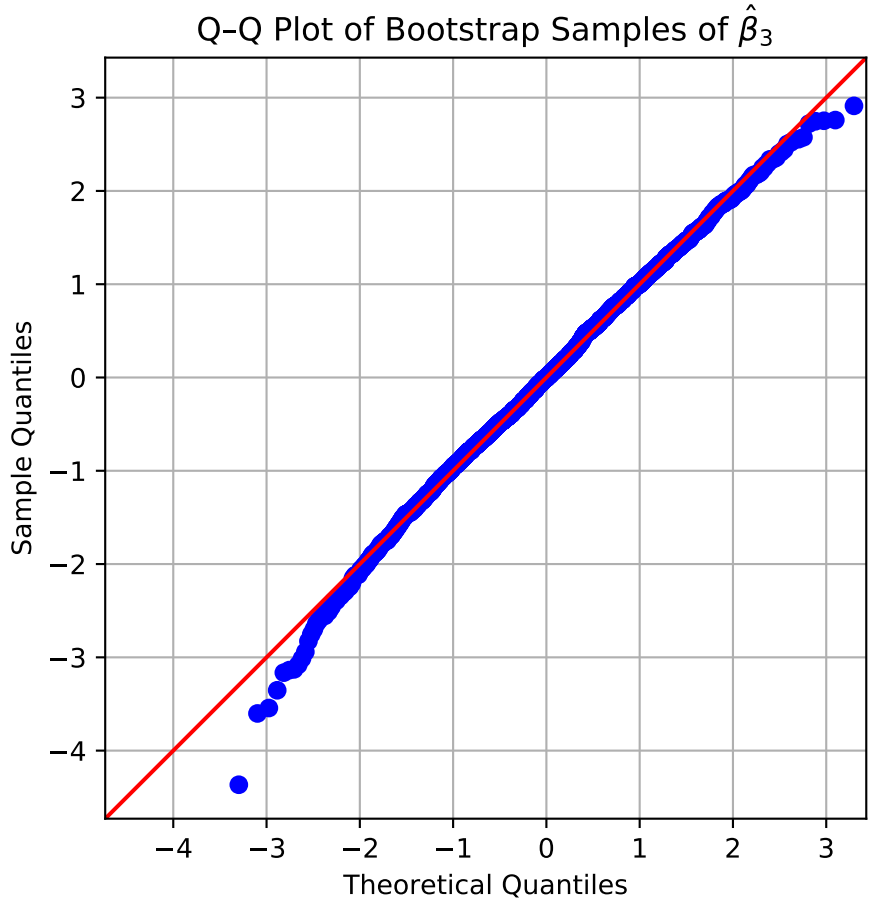

Figure 1: Bootstrap Q–Q plot for $\hat{\beta}_3$ when resampling clusters.

Using $\hat{\Sigma}_{\text{Sandwich}}$ for to get the standard error of $\hat{\beta}_3$, we obtain a smaller standard error $\boxed{0.11686716}$ since we can exploit within cluster correlation to get a better estimate.

(b) Calculate bootstrap standard error estimates for $\hat{\beta}_3$ by resampling clusters. Describe the results of some basic diagnostics you can do to provide confidence that bootstrap intervals are valid for this dataset and that you have simulated a sufficient number of draws to accurately approximate true bootstrap intervals?

Solution: The bootstrap standard error for $\hat{\beta}_3$ when resampling clusters is $\boxed{0.11997406,}$ which is similar to the sandwich estimate.

This was calculated by taking the square root of the sample variance of the of the boostrap samples for $\hat{\beta}_3$. 2,048 samples were taken. Normality of the samples was checked by using a Q–Q plot in Figure 1 and a histogram in Figure 2. The distribution in the histrogram does look normal, and the fit in the Q–Q plot is quite good, so we can be confident that the distribution of the samples is indeed normal.

Indeed, if $\hat{\beta}_3$ is our REML estimate, $\hat{\sigma}$ is our bootstrap standard error, and Φ is the CDF for the standard normal, then the interval $\left[\hat{\beta}_3 - \Phi^{-1}(0.975)\,\hat{\sigma}, \hat{\beta}_3 + \Phi^{-1}(0.975)\,\hat{\sigma}\right]$ contains

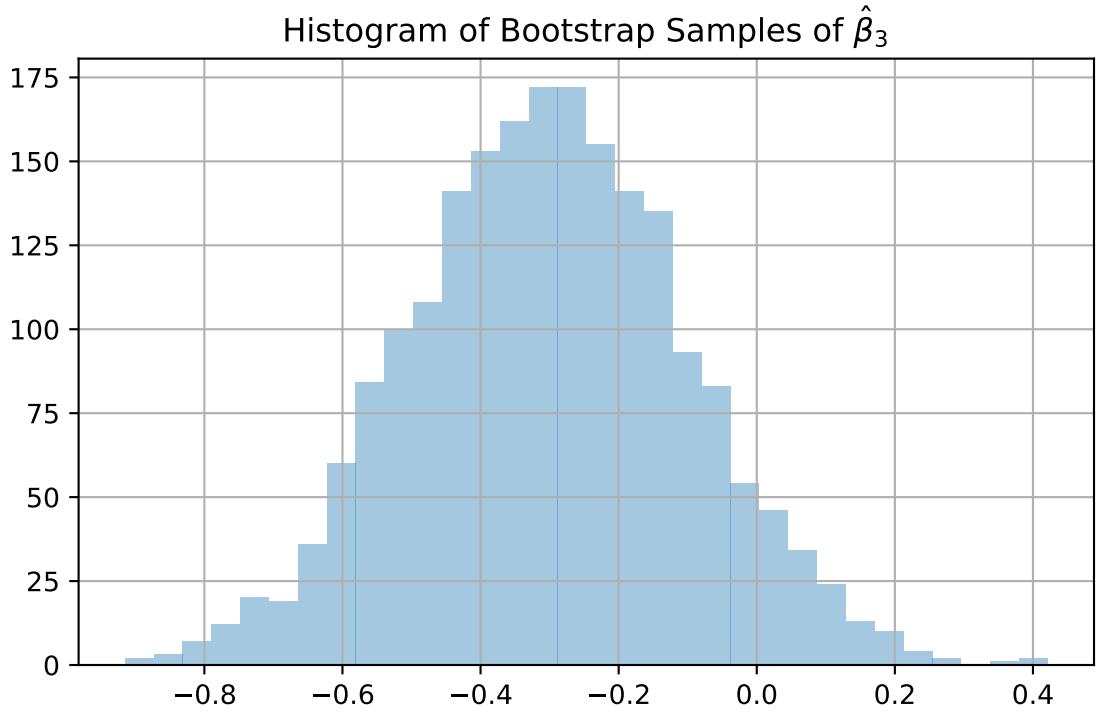

Figure 2: Bootstrap histogram for $\hat{\beta}_3$ when resampling clusters.

94.97% of the bootstrap samples, so the interval is quite accurate.

(c) Calculate bootstrap standard error estimates for $\hat{\beta}_3$ based on resampling observations without regard to cluster and resampling both clusters and observations within clusters.

(d) Discuss any differences between your sandwich standard error estimates and the three versions of bootstrap standard errors.