

PLSC 503: Solutions to Problem Set 8

Thad Dunning

Department of Political Science

Yale University

Spring 2010

1 Theoretical/Conceptual Exercises

Question 1:

The bootstrap. An analyst assumes the following regression model:

$$Y = X\beta + \epsilon, \quad (1)$$

where Y is an $n \times 1$ vector of observable random variables. Here, X is a fixed $n \times p$ matrix with a vector of 1's as the first column, and ϵ is mean-zero vector of i.i.d. random variables with $\text{var}(\epsilon_i) = \sigma^2$. The OLS estimator for this model is $\hat{\beta} = (X'X)^{-1}X'Y$. The residuals from the OLS fit are $e = Y - X\hat{\beta}$.

Suppose the analyst uses the procedure described by Freedman (2009, Chapter 8) to bootstrap the regression model. In particular, for the k th bootstrap replicate, she samples at random with replacement from the vector e to produce an $n \times 1$ vector of bootstrap errors, $\epsilon_{(k)} = \{\epsilon_{(k)1}, \dots, \epsilon_{(k)n}\}'$. For each bootstrap replicate, she then constructs $Y_{(k)} = X\hat{\beta} + \epsilon_{(k)}$ and fits the OLS estimator, $\hat{\beta}_{(k)} = (X'X)^{-1}X'Y_{(k)}$. There are 100 bootstrap replicates. Finally, let

$$\begin{aligned} \hat{\epsilon}_{(k)} &= Y_{(k)} - X\hat{\beta}_{(k)} \\ s_k^2 &= \frac{\hat{\epsilon}_{(k)}'\hat{\epsilon}_{(k)}}{n-p} \\ \widehat{\text{cov}}(\hat{\beta}_{(k)}) &= s_k^2(X'X)^{-1} \\ \hat{\beta}_{\text{ave}} &= \frac{1}{100} \sum_{k=1}^{100} \hat{\beta}_{(k)} \\ V &= \frac{1}{100} \sum_{k=1}^{100} [\hat{\beta}_{(k)} - \hat{\beta}_{\text{ave}}][\hat{\beta}_{(k)} - \hat{\beta}_{\text{ave}}]'. \end{aligned}$$

Say whether the following statements are true or false, and most importantly, explain your answers:

(a): $E(\epsilon_{(k)}) = 0_{n \times 1}$.

Solution: True. The bootstrap errors, $\epsilon_{(k)}$, are an i.i.d. sample from a box with mean zero. In more detail, the original OLS residual vector e has mean zero because there is an intercept in the model, and $e \perp X$. We are sampling at random with replacement from this vector. Thus, the expected value of each observed bootstrap error is the mean of the box.

(b): $E(\hat{\beta}_{(k)}) = \hat{\beta}$.

Solution: True. By (a), the bootstrap errors have an expected value of zero, and they are independent of X (which here is held fixed). The bootstrap OLS estimator $\hat{\beta}_{(k)}$ is therefore an unbiased estimator of $\hat{\beta}$. In more detail,

$$\begin{aligned}\hat{\beta}_{(k)} &= (X'X)^{-1}X'Y_{(k)} \\ &= (X'X)^{-1}X'(X\hat{\beta} + \epsilon_{(k)}) \\ &= \hat{\beta} + (X'X)^{-1}X'\epsilon_{(k)}\end{aligned}$$

so $E(\epsilon_{(k)}) = 0_{n \times 1}$ implies that $E(\hat{\beta}_{(k)}) = \hat{\beta}$.

(c): $E(s_k^2) = \sigma^2$.

Solution: False. Here, s_k^2 is an unbiased estimator for the variance of the box from which the bootstrap errors are drawn—namely, the box of original residuals e . But the variance of the box of residuals is not σ^2 , at least not exactly.

In more detail, s_k^2 is the analogue of $\hat{\sigma}^2$, where σ^2 is the variance of the error term in the usual regression model. See Freedman (2009: Theorem 4, pp. 47-48) for a proof that $E(\hat{\sigma}^2|X) = \sigma^2$. For the same reason, $E(s_k^2)$ is the variance of the residuals from the original regression fit.

(d): $E(s_k^2) = \frac{1}{n}e'e$.

Solution: True: see (c). The variance of the box of original residuals is $\frac{1}{n}e'e$, and s_k^2 is an unbiased estimator for the variance of this box. (Remember that the residuals have mean zero, so $(e - \bar{e})'(e - \bar{e}) = e'e$, where \bar{e} is the average of the residuals).

(e): $E(s_k^2) = \frac{1}{n-p}e'e$.

Solution: False. The variance of the box of original residuals is $\frac{1}{n}e'e$, not $\frac{1}{n-p}e'e$: see (d).

(f): $E(V) = \sigma^2(X'X)^{-1}$

Solution: False. Under the OLS model, the theoretical variance-covariance matrix of the bootstrap replicates is the variance of the bootstrap “error term” times $(X'X)^{-1}$. Here, the error term is given by i.i.d. draws from the vector e —and *not* draws of the unobservable distribution of ϵ . Thus, in (f), σ^2 should be replaced by $\frac{1}{n}e'e$ —that is, the theoretical variance of the bootstrap errors.

In more detail, the bootstrap errors are i.i.d. random variables, because we are drawing at random with replacement from the vector e . The variance of the vector e is $\frac{1}{n}e'e$ (see d). Thus, the variance of the random variables is $\frac{1}{n}e'e$. Finally, we are generating the bootstrap data according to the OLS model, and so the theoretical variance-covariance matrix of the bootstrap replicates is $\frac{1}{n}e'e(X'X)^{-1}$.

(g): The square roots of the diagonal elements of V are the bootstrap standard errors.

Solution: True. The variance-covariance matrix of the bootstrap replicates is V , and the square roots of the diagonal elements are the bootstrap standard errors.

(h): The sample SD of the $\hat{\beta}_{(k)}$'s is a good approximation to the SE of $\hat{\beta}$.

Solution: True. This is the bootstrap principle at work.

(i): $\hat{\epsilon}_{(k)} \perp X$ for all k .

Solution: True. The $\hat{\epsilon}_{(k)}$ are the residuals from the OLS fit to the k th bootstrapped data set. Mechanically, these residuals are orthogonal to X – that is what regression does.

(j): The bootstrap can provide evidence that the original data were produced according to equation (1).

Solution: False. This claim is essentially untestable. The bootstrap *assumes* the original data were produced according to equation (1), with i.i.d. errors, $E(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \sigma^2$, and then evaluates the sampling distribution of estimators such as $\hat{\beta}$. That is where the bootstrap principle comes in.

(k): The bootstrap can provide evidence that $E(\hat{\beta}) = \beta$ if the original data were produced according to equation (1), with i.i.d. errors, $E(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \sigma^2$.

Solution: True. See the solution to (j).