Eco311: Bootstrap Standard Error

(Jing Li, Miami University)

1. Recall: (1) the fundamental idea of statistics is using sample to understand population; (2) there are many samples; (3) consequently, results from using different samples are random; (4) standard error measures the uncertainty associated with using sample

2. Statistically speaking, reporting result without standard error is unacceptable. People may wonder, do you get that result by chance? Does your result still hold after trying different samples? In short, is your result statistically significant?

3. For a regression model, the key unknown parameter is slope coefficient β_1, which measures the marginal effect of x on y. The basic idea of regression analysis is using sample to estimate β_1. OLS method is the most popular one in regression analysis.

4. The OLS estimate $\hat{\beta}_1$ is random since it varies from sample to sample. Under certain assumptions, we can mathematically derive its variance as

$$var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} \tag{1}$$

   The square root is called standard error

$$se(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_i (x_i - \bar{x})^2}} = \frac{\sigma}{\sqrt{(n-1)S_x^2}} \tag{2}$$

   Then a t statistic can be computed as

$$t = \frac{\hat{\beta}_1 - c}{se}$$

   The hypothesis $H_0 : \beta_1 = c$ is rejected if the t value exceeds 1.96 in absolute value.

5. Big variance or standard error implies that the estimate is imprecise since it may change a lot if sample changes. In general we prefer a precise estimate that comes with small standard error

6. We use simulated data as an illustration

```
> set.seed(12345)
> n = 100
> x = rnorm(n)
> y = 2*x + rnorm(n)
> cat("se of beta1hat from formula (2) is ", 1/sqrt(var(x)*99), "\n")

se of beta1hat from formula (2) is  0.09015969

> m = lm(y~x)
> summary(m)$coef

              Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 0.02205339  0.1035259  0.2130228 8.317517e-01
x           2.09453503  0.0911382 22.9819673 2.975126e-41
```

Note that OLS estimate based on this particular sample $\hat{\beta}_1 = 2.09453503$ is close to the true value $\beta_1 = 2$; the se reported by R 0.0911382 is close to the answer 0.0901596 obtained from formula (2)

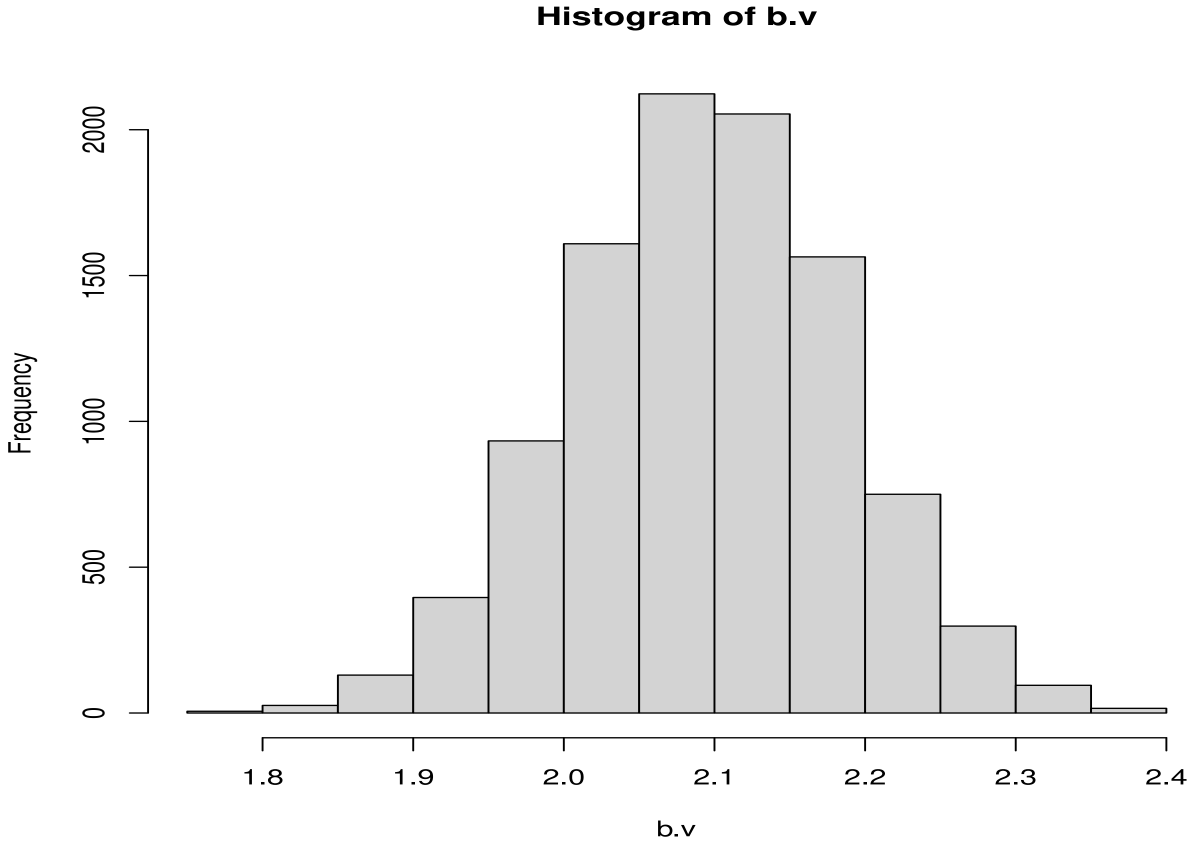
7. Next we use bootstrap to demonstrate that we can obtain different $\hat{\beta}_1$ from different (bootstrap) samples.

```
> uhat = resid(m)
> eb0 = summary(m)$coef[1,1]; eb1 = summary(m)$coef[2,1]
> nboot=10000; b.v=rep(0,nboot)
> for (i in 1:nboot) {
+ inde = sample(seq(1,n),n,replace=T)
+ booty = eb0 + eb1*x  + uhat[inde]
+ b.v[i] = summary(lm(booty~x))$coef[2,1]
+ }
> cat("bootstrap se is ", sd(b.v), "\n")

bootstrap se is  0.09012332
```

8. Basically we hold x fixed. Then we (i) re-sample residual with replacement; (ii) using the re-sampled residual, orginal x, and OLS estimated coefficients to generate bootstrap y; (iii) regress bootstrap y onto x and obtain new $\hat{\beta}_1$. We repeat steps (i, ii, iii) many times. In the end, the standard error of all those bootstrap estimates .09012332 is very close to the se reported by R lm function.

9. Note that the distribution of bootstrap $\hat{\beta}_1$ is bell-shaped, a finding implied by the central limit theorem


Histogram of b.v
Frequency
b.v