# Assignment 9

Instructor: Xiner Zhou

March 13, 2023

## Bootstrapping Regression Models

### 0.1 Some Regression Basics

**Assumed Model (Scalar Form)** Consider the linear regression model of the form

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$$

for $i = 1, \dots, n$, where $y_i$ is the response variable for the $i$-th observation, $x_{ij}$ is the $j$-th predictor for the $i$-th observation, $\beta_0$ is the regression intercept parameter, $\beta_j$ is the regression slope for the $j$-th predictor, and $\varepsilon_i \overset{\text{iid}}{\sim} \left(0, \sigma^2\right)$ is the error term for the $i$-th observation. This implies that the conditional mean of $y_i$ given $\mathbf{x}_i = \left(x_{i1}, \dots, x_{ip}\right)^\top$ has the form

$$E(y_i \mid \mathbf{x}_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

and the conditional variance of $y_i$ given $\mathbf{x}_i$ is $\text{Var}(y_i \mid \mathbf{x}_i) = \sigma^2$.

**Assumed Model (Matrix Form)** In matrix form, the linear regression model has the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{y} = (y_1, \dots, y_n)^\top$ is the observed response vector, $\mathbf{X}$ is the $n \times p + 1$ design matrix with $i$-th row equal to $\left(1, x_{i1}, \dots, x_{ip}\right)$, $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \dots, \beta_p\right)^\top$ is the regression coefficient vector, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$ is the error vector. Writing out the vectors and matrices, the regression model has the form

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The assumptions from the previous section imply that $E(\mathbf{y} \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{Cov}(\mathbf{y} \mid \mathbf{X}) = \sigma^2 \mathbf{I}_n$, where $\mathbf{I}_n$ denotes the identity matrix of order $n$.

**Coefficient Estimates and Fitted Values** The ordinary least squares coefficient estimates are the coefficients that minimize the least squares loss function

$$\sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

which can be written more compactly in matrix form such as

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

where $\|\cdot\|$ denotes the Euclidean norm. It is well known that the ordinary least squares coefficient estimates have the form

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

and the corresponding fitted values have the form

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is the "hat matrix" for the linear model.

**Statistical Inference** If the model assumptions are correct, the expectation of the estimated coefficient vector has the form

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E\Big(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}\Big) \\ &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top E(\mathbf{y}) \\ &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

which reveals that the coefficient estimates are unbiased. The covariance matrix of the coefficients have the form

$$\begin{aligned} \mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= \mathrm{Cov}\Big(\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}\Big) \\ &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top \mathrm{Cov}(\mathbf{y})\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \\ &= \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\left(\sigma^2\mathbf{I}_n\right)\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \\ &= \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1} \end{aligned}$$

Similar arguments can be used to show that the expected value and covariance matrix for the fitted values have the form

$$E(\hat{\mathbf{y}}) = \mathbf{X}\boldsymbol{\beta} \quad \text{and} \quad \mathrm{Cov}(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}$$

which is due to the fact that $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

## 0.2 Need for the Nonparametric Bootstrap

The expectations and covariance matrices derived in the previous section depend on the assumption that the model is correctly specified, i.e., that

1. $E(y_i \mid \mathbf{x}_i) = \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}$
2. $\mathrm{Var}(y_i \mid \mathbf{x}_i) = \sigma^2$
3. the $\varepsilon_i$ are independent of one another

Note that even these assumptions are not enough to conduct statistical inference on the coefficient estimates and/or fitted values. If we add the assumption of normality, i.e., that $\varepsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right)$, then the sampling distribution of the coefficient estimates is known to be Gaussian, i.e., $\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\right)$. This is the approach that is commonly used for inference in regression models, e.g., in the *lm()* function in R. However, in practice, the normality assumption may not be reasonable, which implies that a more general (nonparametric) approach is needed for inference in regression.

In the following subsections, we will talk through how nonparametric bootstrap can be used for inference in regression, by doing 3 exercises. The statistic of interest is assumed to be the regression coefficient vector $\boldsymbol{\beta}$. However, it should be noted that this does not have to be the case. In some applications, it may be more useful to use some function of the coefficient vector as the statistic.

## 0.3 Procedure for Bootstrapping Regression Models

The multivariate data are assumed to have the form $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]^\top$ where $\mathbf{z}_i = \left(y_i, x_{i1}, \ldots, x_{ip}\right)^\top$ is the $i$-th observation's vector of length $p+1$. This implies that $\mathbf{Z}$ is a data matrix (or data frame) of dimension $n \times p+1$, where $n$ is the number of observations and $p+1$ is the total number of variables ( $p$ predictors plus 1 response).

The procedure is as follows:

1. Independently sample $\mathbf{z}_i^b$ with replacement from $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ for $i = 1, \ldots, n$.

2. Calculate the statistic $\hat{\boldsymbol{\beta}}^b = \left(\mathbf{X}^{b\top}\mathbf{X}^b\right)^{-1}\mathbf{X}^{b\top}\mathbf{y}^b$ where the $i$-th row of $\mathbf{X}^b$ is defined as $\mathbf{x}_i^b = \left(1, x_{i1}^b, \ldots, x_{ip}^b\right)^\top$ and $\mathbf{y}^b = \left(y_1^b, \ldots, y_n^b\right)$.
3. Repeat steps $1-2$ a total of $B$ times to for the bootstrap distribution of $\hat{\boldsymbol{\beta}}$.

## Exercise 1

In this example, we will generated $n = 100$ observations from a simple linear regression model where the error terms satisfy $\varepsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right)$. Note that the true regression coefficients are $\beta_0 = 0$ and $\beta_1 = 1$, and the error variance is $\sigma^2 = 1$.

```
# generate 100 observations
 n <- 100
 set.seed(1)
 x <- seq(0, 1, length.out = n)
 y <- x + rnorm(n)
 data <- data.frame(x = x, y = y)
```

(1) Use Bootstrap to estimate the standard error and construct 95% confidence interval for $\beta_1$.

(2) compare them with the reported standard error and confidence interval from *lm()* function in R. Comment on whether they agree with each other, if not, why do they differ and which one is more trustworthy?

## Exercise 2

This exercise is the same as the previous example, except that now the error terms have different variances, i.e., $\varepsilon_i \overset{\text{ind}}{\sim} N\left(0, \sigma_i^2\right)$ where $\sigma_i^2 = 1 + 4(x_i - 0.5)^2$.

```
# generate 100 observations
 n <- 100
 set.seed(1)
 x <- seq(0, 1, length.out = n)
 y <- x + rnorm(n, sd = sqrt(1 + 4 * (x - 0.5)^2))
 data <- data.frame(x = x, y = y)
```

(1) Use Bootstrap to estimate the standard error and construct 95% confidence interval for $\beta_1$.

(2) compare them with the reported standard error and confidence interval from *lm()* function in R. Comment on whether they agree with each other, if not, why do they differ and which one is more trustworthy?

## Exercise 3

We will use the SAT and College GPA example. The dataset contains information from $n = 105$ students that graduated from a state university with a B.S. degree in computer science.

- high GPA = High school grade point average
- math SAT= Math SAT score
- verb SAT=Verbal SAT score
- comp GPA = Computer science grade point average
- univ GPA = Overall university grade point average

*R* code to read-in and look at the data

```
# read-in data
sat <- read.table("http://onlinestatbook.com/2/case_studies/data/sat.txt",
                  header = TRUE)
head(sat)
```

```r
# Plot the data.
 plot(sat$high_GPA, sat$univ_GPA,
      xlab = "High school GPA", ylab = "University GPA")
 abline(lm(univ_GPA ~ high_GPA, data = sat))
```

**Is University GPA Linearly Related to High School GPA?** Consider the simple linear regression model

$$Y = \beta_0 + X\beta_1 + \varepsilon$$

where $Y$ is the University GPA and $X$ is the high school GPA.

(1) Use Bootstrap to estimate the standard error and construct 95% confidence interval for $\beta_1$.

(2) Compare them with the reported standard error and confidence interval from *lm()* function in R. Comment on whether they agree with each other, if not, why do they differ and which one is more trustworthy?