8.1 The Basic Bootstrap Method

- We would like to have a measure of how close the estimate is to the population value.

8.1.1 Mean Squared Error and Margin of Error

Mean Squared Error (MSE):

$MSE = E[(\hat{\theta}-\theta)^2]$, where θ is a population parameter, and $\hat{\theta}$ is a statistical estimate of θ

Margin of Error:

$\text{Margin of Error} = z\sqrt{MSE}$, and this can be used for calculating Chebyshev-Markov inequality, which is $P(|\hat{\theta}-\theta| \leq k\sqrt{MSE}) \geq 1-\frac{1}{k^2}$

* The problem of the above formulas is that in most cases, the population parameter θ is unknown so we cannot proceed with the above process. This is where the bootstrap sampling comes in.

8.1.2 The Bootstrap Estimate of MSE

- When the population distribution is not known, the data may be used as a substitute for the population, and we may resample the data as if we were sampling the population. (with replacement)

Process of Bootstrap Sampling

1. Compute $\hat{\theta}$ from the original data.
2. Take a number of bootstrap samples of size n from the data. Let REP denote the number of such bootstrap samples. Typically, $REP \geq 1000$
3. Compute $\hat{\theta}_{b,i}$, the estimate of θ obtained from the i^{th} bootstrap sample.
4. Obtain the bootstrap MSE,

$$\hat{MSE} = \frac{1}{REP}\sum_{i=1}^{REP}(\hat{\theta}_{b,i}-\hat{\theta})^2$$

* Coefficient of Variation

Bootstrap Variance and Bias

Bias : the difference between the expected value of an estimate and the quantity being estimated.

$B = E(\hat{\theta}) - \theta$, and let "Var" be the variance of the estimate, then

$MSE = Var + B^2$

* Instead of calculating $\widehat{MSE}$ in the last step of Bootstrap process, we may compute,

$$\hat{E} = \frac{1}{REP}\sum_{i=1}^{REP}\hat{\theta}_{b,i},$$

$$\hat{B} = \hat{E} - \hat{\theta},$$

$$Var = \frac{1}{REP}\sum_{i=1}^{REP}(\hat{\theta}_{b,i} - \hat{E})^2$$

* Usually, 1000 bootstrap samples are recommended. If no concerns about the computation, 5000 is also recommended.

Nonparametric	versus	Parametric Bootstrap
- no assumptions are made about the functional form of the population distribution		- assumptions are made about the form of the population distribution.

8.2 Bootstrap Intervals for Location-Scale Models

- Let $g(x) = \frac{1}{\nabla} f\left(\frac{x-\mu}{\nabla}\right)$, where $f(z) \sim N(0,1)$. Then $X = \mu + \nabla z$.

8.2.1 Interval Estimate of the Mean

- If $f(z)$ follows a normal distribution, the t-statistic defined by $t = \frac{\bar{x}-\mu}{S/\sqrt{n}}$ has a t-distribution with $n-1$ degrees of freedom. Then,

$$P\left(-t_{0.975} < \frac{\bar{x}-\mu}{S/\sqrt{n}} < t_{0.975}\right) = 0.95$$

$\Rightarrow$ $1-\alpha$ Confidence Interval

$$\bar{x} - t_{1-\alpha/2}\left(\frac{S}{\sqrt{n}}\right) < \mu < \bar{x} + t_{1-\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$$

Pivot Quantity :

- a function of observations and parameters whose probability distribution does not depend on the parameters (any quantity not depending on unknown parameters)

Steps to Obtain a Bootstrap Confidence Interval for μ

1. Compute the mean $\bar{x}$ and standard deviation S of the original data.
2. Obtain a bootstrap sample of size n from the data. Compute the mean $\bar{x}_b$ and standard deviation S_b of the bootstrap sample, and compute the bootstrap t-pivot quantity

$$t_b = \frac{\bar{x}_b - \bar{x}}{S_b/\sqrt{n}}$$

3. Repeat step 2 a number of times to obtain a bootstrap distribution of the t_b's.
4. For a 95% confidence interval, let $t_{b,0.025}$ and $t_{b,0.975}$ be the 2.5th and 97.5th percentiles of the bootstrap distribution.

$$\bar{x} - t_{b,0.975}\left(\frac{S}{\sqrt{n}}\right) < \mu < \bar{x} - t_{b,0.025}\left(\frac{S}{\sqrt{n}}\right)$$

* Asymmetric Bootstrap Distribution → bootstrap confidence interval using t-pivot
  Symmetric Bootstrap Distribution → $(1-\alpha)100\%$ confidence intervals using t-distribution with $n-1$ degrees of freedom.

8.2.2 Confidence Intervals for the Variance and Standard Deviation

χ^2-pivot : a pivot quantity for the variance, $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$

- If the observations come from a normal distribution, then this pivot quantity has a χ^2 distribution with $n-1$ degrees of freedom.

$$P\left(\chi^2_{0.025} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{0.975}\right) = 0.95 \text{, if normality assumptions hold}$$

$$\frac{(n-1)S^2}{\chi^2_{0.975}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{0.025}} \text{, if normality assumptions do not hold.}$$

Steps to Obtain a Bootstrap Confidence Interval for σ^2

1. Compute the sample variance S^2 for the original data.
2. Obtain a bootstrap sample of size n. Compute the sample variance S_b^2 of the bootstrap sample, and compute the bootstrap χ^2-pivot quantity $\chi^2_b = \frac{(n-1)S_b^2}{S^2}$

3. Repeat step 2 a number of times to obtain a bootstrap distribution of the χ^2_b's.

4. For a 95% confidence interval, let $\chi^2_{b,0.025}$ and $\chi^2_{b,0.975}$ be the 2.5th and 97.5th percentiles of the bootstrap distribution. The bootstrap 95% confidence interval is $\frac{(n-1)S^2}{\chi^2_{0.975}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{0.025}}$

8.2.3 Coverage Percentages

8.2.4 Derivation of Pivotal Quantities

음 뭔소린지 1도 모르겠음

8.3 BCA and Other Bootstrap Intervals

8.3.1 Percentile and Residual Methods

Percentile Method:

1. Draw a specified number of bootstrap samples of size n from the data, and for each bootstrap sample, compute the estimate $\hat{\theta}_b$ of θ.
2. For a 95% confidence interval for θ, find the 2.5th and 97.5th percentiles $\hat{\theta}_{b,0.025}$ and $\hat{\theta}_{b,0.975}$ of the bootstrap distribution.

Residual Method:

1. Compute $\hat{\theta}$ from the data.
2. Draw a bootstrap sample of size n from the data. Compute $\hat{\theta}_b$ and the residual $e_b = \hat{\theta}_b - \hat{\theta}$
3. Repeat step 2 a specified number of times to obtain the bootstrap distribution of the e_b's
4. For a 95% confidence interval, obtain the 2.5th and 97.5th percentiles, $e_{b,0.025}$ and $e_{b,0.975}$ of the bootstrap distribution, $\hat{\theta} - e_{b,0.975} \le \theta < \hat{\theta} - e_{b,0.025}$

or

$\hat{\theta} - (\hat{\theta}_{b,0.975} - \hat{\theta}) \le \theta < \hat{\theta} - (\hat{\theta}_{b,0.025} - \hat{\theta})$

or

$2\hat{\theta} - \hat{\theta}_{b,0.975} \le \theta < 2\hat{\theta} - \hat{\theta}_{b,0.025}$

8.3.2 BCA Method

- Percentile and residual methods may not provide an appropriate interval for the true parameter.
- Bias Corrected and Accelerated method adjust for these problems. BCA interval is corrected for bias and skewness
- If the bootstrap distribution is symmetric about $\hat{\theta}$, then the BCA method, percentile method, and residual method give the same endpoints.

A Sketch of the BCA Method

- The idea of the BCA method is to assume that there is a transformation of $\hat{\theta}$ whose distribution is normal and whose mean and standard deviation depend in a particular way on θ. A confidence interval is made on the transformed parameter and then the interval is inverted to obtain an interval for θ. The inversion can be done without knowledge of the explicit form of the transformation using the bootstrap method.

- Let T is a strictly increasing transformation, $T(\hat{\theta})$ has a mean and standard deviation,

$$E[T(\hat{\theta})] = T(\theta) - z_0[1 + aT(\theta)]$$

$$V[T(\hat{\theta})] = 1 + aT(\theta)$$, then the $(1-\alpha)100\%$ confidence interval would be,

$$-z_p < \frac{T(\hat{\theta}) - T(\theta)}{1 + aT(\theta)} + z_0 < z_p \Rightarrow \frac{T(\hat{\theta}) + z_0 - z_p}{1 - a(z_0 - z_p)} < T(\theta) < \frac{T(\hat{\theta}) + z_0 - z_p}{1 - a(z_0 + z_p)}$$

The Upper and Lower Limit of the BCA: (when referred to the standard normal distribution)

$$z_U = \frac{z_0 + z_p}{1 - a(z_0 + z_p)} + z_0 \quad , \quad z_L = \frac{z_0 - z_p}{1 - a(z_0 - z_p)} + z_0 \quad , \text{ where } z_0 = \Phi^{-1}\left\{\frac{1}{REP}\sum_{b=1}^{REP}(\hat{\theta}_b < \hat{\theta})\right\}$$

and $$a = \frac{\sum_{i=1}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{-i})^3}{6\left[\sum_{i=1}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{-i})^2\right]^{\frac{3}{2}}}$$, where $\hat{\theta}_{(\cdot)}$ is the mean of $\hat{\theta}_{-i}$'s, and $\hat{\theta}_{-i}$ is $\hat{\theta}$ without i^{th} bootstrap sample.

- and obtain the percentages of z_U and z_L, and this would be the BCA $(1-\alpha)100\%$ confidence limits for θ

8.4 Correlation and Regression

Bivariate Sampling and Fixed-X Sampling

8.4.1 Bivariate Bootstrap Sampling

- a random sample with replacement of the pairs (X_i, Y_i), $i = 1, 2, 3, \ldots, n$.

n^S equally likely possible bootstrap samples

Bootstrap Confidence Interval:

1. Draw a specified number of bivariate samples of size n from the data.
2. Compute the bootstrap Pearson correlation coefficient r_b for each bootstrap sample.
3. Obtain a confidence interval from the distribution of the r_b's using the BCA method or the percentile method, with the BCA method preferred. The residual method has the undesirable property of possibly producing limits that are beyond the bounds of -1 to 1 for correlation

- We may obtain limits developed for the bivariate normal distribution. These limits are based on the transformation $Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$, which has an approximate normal distribution with mean $Z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ and standard deviation $\frac{1}{\sqrt{n-3}}$.
- This transformation gives the 95% confidence interval $\frac{(1+r)e^c - (1-r)}{(1+r)e^c + (1-r)}$, where

$c = \frac{2(z_{1-\alpha_2})}{\sqrt{n-3}}$. The BCA interval in particular gives a smaller value for the lower limit and a wider interval than the normal-theory method.

8.4.2 Fixed-X Bootstrap Sampling

- assume $h(x)$ is a function of the independent variable, and Y may be expressed in terms of the regression model. $Y = h(x) + \varepsilon$

1. Compute an estimate $\hat{h}(x)$ of the mean function $h(x)$.
2. Obtain the observed errors $e_i = Y_i - \hat{h}(x_i)$.
3. Select n values of the e_i's at random with replacement. Denote these values as $e_{1,b}, e_{2,b}, \ldots e_{n,b}$.
4. Compute $Y_i = \hat{h}(x_i) + e_{i,b}$. The pairs (x_i, Y_i) comprise a fixed-X bootstrap sample of size n.

8.4.3 Bootstrap Inferences for the Slope of a Regression Line

- BCA, percentile, or residual method

t-Pivot for the Slope :

Estimated Standard Error : $SE(\hat{\beta}_1) = \sqrt{\frac{MSE}{(n-1)S_x^2}}$

Standard Deviation of X's : $t = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)}$

- When the errors have a normal distribution, the t-pivot has a t-distribution with $n-2$ degrees of freedom. $-t_{0.975} < \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} < t_{0.975}$

Bootstrap Distribution of the t-pivot :

- Suppose we cannot assume the normality of ε's, apply the least squares procedure to the pairs (x_i, e_i).

1. Compute the least squares estimates and the residuals $e_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$
2. Take a sample of size n of the e_i's with replacement, and form the bootstrap sample $(x_i, e_{i,b})$. Compute the least squares estimate of the slope $\hat{\beta}_{1e}$, the estimated standard error $SE(\hat{\beta}_{1e})$, and the bootstrap t-statistic $t_e = \frac{\hat{\beta}_{1e}}{SE(\hat{\beta}_{1e})}$
3. Repeat step 2 a number of times to obtain the bootstrap distribution of the t_e's. Find the desired percentiles of the distribution of the t_e's and use these to construct a confidence interval or conduct a test of hypothesis for β_1.

Adjusted Errors :

- Leverage of X, $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$, and the variance of e_i is given $Var(e_i) = \sigma^2(1 - h_i)$. Then, we define r_i as $r_i = \frac{e_i}{\sqrt{1-h_i}}$, and the adjusted errors are $e_i^* = r_i - \bar{r}$.

Derivation of the Distribution of the t-pivot :

딱히 중요한 내용은 아닌듯

8.5 Two-Sample Inference

Homogeneity of variances vs. Heterogeneity of variances

8.5.1 t-pivot Method Assuming Equality of Error Distributions,

- Suppose the distributions of the errors are the same for the two populations. Let,

$$t = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \text{ where } S_p^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\frac{(Y_{ij} - \bar{Y}_i)^2}{n_1 + n_2 - 2}$$

$$t_\varepsilon = \frac{\bar{\varepsilon}_1 - \bar{\varepsilon}_2}{S_{\varepsilon p}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad S_{\varepsilon p}^2 = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\frac{(\varepsilon_{ij} - \bar{\varepsilon}_i)^2}{n_1 + n_2 - 2}$$

- If the ε_{ij}'s have a normal distribution, then t_ε and t have a t-distribution with $n_1 + n_2 - 2$ degrees of freedom. We may obtain a 95% confidence interval for the difference of the population means by sovling the inequality

$$-t_{0.975} < \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{0.975}$$

Bootstrap Interval

1. Compute the observed errors $e_{ij} = Y_{ij} - \bar{Y}_i$
2. Randomly select n_1 errors with replacement from the set of all errors and assign them to the first sample, and similarly select n_2 errors and assign them to the second sample. Denote these as $e_{ij,b}$. Compute t_ε and the $e_{ij,b}$'s instead of the ε_{ij}'s and denote this bootstrap statistic as t_ε.
3. Repeat step 2 a number of times to obtain the bootstrap distribution of the t's.
4. For a 95% confidence interval, let $t_{e,0.025}$ and $t_{e,0.975}$ denote the 2.5^{th} and 97.5^{th} percentiles of the bootstrap distribution. The 95% bootstrap confidence interval is found by solving the inequality

$$-t_{e,0.025} < \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{e,0.975}$$

$$\rightsquigarrow \bar{Y}_1 - \bar{Y}_2 - t_{e,0.975}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < \bar{Y}_1 - \bar{Y}_2 - t_{e,0.025}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Unequal Distributions of the Errors

$$Z = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} \rightsquigarrow Z_\varepsilon = \frac{\bar{\varepsilon}_1 - \bar{\varepsilon}_2}{\sqrt{S_{\varepsilon 1}^2/n_1 + S_{\varepsilon 2}^2/n_2}}, \quad S_{\varepsilon i}^2 = \sum_{j=1}^{n_i} \frac{(\varepsilon_{ij} - \bar{\varepsilon}_i)^2}{n_i - 1}$$

1. Compute the observed errors $e_{ij} = Y_{ij} - \bar{Y}_i$
2. Randomly select n_1 errors with replacement from the set of all errors and assign them to the first sample, and similarly select n_2 errors and assign them to the second sample. Denote these as $e_{i1,b}$. Compute Z_e.
3. Repeat step 2 a number of times to obtain the bootstrap distribution of Z_e.
4. For a 95% confidence interval,

$$Z_{e,0.025} < \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}} < Z_{e,0.975}$$

Other Methods

넘어가도 될듯

8.6 Bootstrap Sampling from Several Populations
8.6.1 Equal Error Distributions
8.6.2 Unequal Error Distributions
8.6.3 Regression with Unequal Error Variances

} 안바쁜 내용

8.7 Bootstrap Sampling for Multiple Regression

- Under the assumption that the errors have a normal distribution, F_M has an F-distribution with degrees of freedom k for the numerator and $n-k-1$ for the denominator, and F_j has an F-distribution with degrees of freedom 1 for the numerator and $n-k-1$ for the denominator.

※ $F_j = t_j^2$

8.7.1 The Bootstrap Procedure for Testing $H_{0(M)}$ and $H_{0(j)}$

1. Fit the regression model to the data and obtain the observed errors e_i, where $e_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki})$.
2. Take a random sample of size n of the e_is with replacement, and let $e_{i,b}$ denote these values. Form the fixed-X bootstrap sample and $e_{i,b}$. Apply

the computational formulas for the statistics F_M and F_j to these data, and denote the results F_{Me} and F_{je}, respectively.

3. Repeat step 2 a sufficient number of times to generate the bootstrap distributions of the F_{Me}'s and F_{je}'s

4. Let $F_{M,obs}$ and $F_{j,obs}$ denote the observed values of F_M and F_j, respectively, from the original data. The bootstrap p-value for $F_{M,obs}$ is the fraction of the F_{Me}'s that are greater than or equal to $F_{M,obs}$. The bootstrap p-value for $F_{j,obs}$ is the fraction the F_{je}'s that are greater than or equal to $F_{j,obs}$.

erence distribution, often agree quite well with the actual levels. The bootstrap methodology gives a way to verify robustness for a particular problem. If the bootstrap percentiles agree with those based on normal theory, then the use of normal-theory methodology would be supported. However, if there is substantial disagreement between percentiles, then one would have to look carefully at the data to see whether the violation of normality assumptions might be severe enough to cause concern. If so, the bootstrap *p*-values would present an alternative to normal-theory *p*-values in carrying out tests of hypotheses.

8.7.2 A Confidence Interval for β_j

- The steps for obtaining a bootstrap approximation of the t-distributions are the same as those for F_j, the only difference being the statistic,

$$t_{e,0.025} < \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} < t_{e,0.975}$$

8.7.3 Theoretical Development 뭔진 모르겠지만

$$C = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \quad F = \frac{(C\hat{\beta})^T [C(X^TX)^{-1}C^T]^{-1} C\hat{\beta}}{q\,MSE}$$

$$MSE = \frac{(Y - X\hat{\beta})^T (Y - X\hat{\beta})}{n-k-1}$$

- If the ε's have a normal distribution, then under the null hypothesis $H_0: C\beta = 0$, F has an F-distribution with q-degrees of freedom for the numerator and n-k-1 degrees of freedom for the denominator.

Distribution of F Under H_0

8.7.4 Other Methods for Regression Analysis

Permutation Tests

Multivariate Bootstrap Sampling:

- a direct extension of bivariate sampling, may use the percentile method, or the BCA method

1. Take a bootstrap multivariate sample and compute estimates of the coefficients, denoted $\hat{\beta}_{j,b}$, $j = 0, 1, \ldots, k$. The method of estimation may be least squares or it may be other methods as discussed in Section 10.3. Also compute any functions of the coefficients, such as linear combinations, for which confidence intervals are desired.
2. Form the bootstrap distributions of the $\hat{\beta}_{j,b}$'s and other functions of interest by taking repeated bootstrap samples of the data.
3. Apply the BCA method, percentile method, or residual method to the bootstrap distributions.