Chapter 6 Bootstrapping

6.1 The Bootstrap Principle

- Let X be a random sample. We want to know the distribution of some statistic $T(X)$ without making strong assumptions on the distribution of X.
  - **Example:** Let $X_1, \cdots, X_n$ be i.i.d. samples. We can estimate $\theta = E(X_i)$ by the sample mean $\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_n$. We want to know the distribution of $\overline{X}_n$ (or $\overline{X}_n - \theta$).
  - If we know the distribution of $T(X) - \theta$, we can compute the bias and variance of $T(X)$ as an estimator of θ.
  - If we know the distribution of $T(X) - \theta$, we can construct confidence interval for θ.
  - Let $T(X)$ be some test statistic. If we know the distribution of $T(X)$ under H_0, we can perform test.

6.1 The Bootstrap Principle

- **Example (Efron and Tibshirani, 1993):** A study was done to see if small aspirin doses would prevent strokes and heart attacks in healthy middle-aged men. The data were collected by a controlled, randomized, double-blind study:
  - One half of the subjects received aspirin and the other half received placebo, with no active ingredients.
  - The subjects were randomly assigned to the aspirin or placebo groups.
  - Both the subjects and the supervising physicians were blinded to the assignments.
  - Summary statistics for strokes:

	strokes	subjects
aspirin group:	119	11037
placebo group:	98	11034

6.1 The Bootstrap Principle

- – Assume that the number of strokes in the aspirin group follows a Binomial(n_1, p_1) distribution, and the number of strokes in the placebo group follows a Binomial(n_2, p_2) distribution. Then we can estimate $\theta_s := p_1/p_2$ by

$$\hat{\theta}_s = \frac{119/11037}{98/11034} = 1.21 > 1.$$

  – Summary statistics for heart attacks:

	heart attacks	subjects
aspirin group:	104	11037
placebo group:	189	11034

  The ratio of heart attack rates in the two groups can be estimate by

$$\hat{\theta}_h = \frac{104/11037}{189/11034} = 0.55 < 1.$$

  – How to access the distributions of $\hat{\theta}_s$ and $\hat{\theta}_h$?

6.1 The Bootstrap Principle

- – **Bootstrap:** Consider the ratio for stroke rates in the two groups.
  - ∗ Create two populations (data sets) : the first consisting of 119 ones and 10918(=11037-119) zeros, and the second consisting of 98 ones and 10936(=11034-98) zeros.
  - ∗ For $b = 1, 2, \cdots, B$, *e.g.*, $B = 1000$,
    - · Randomly draw **with replacement** a sample of 11037 items from the first population, and a sample of 11034 items from the second population. Each of these is called a *bootstrap sample.*
    - · Compute

$$\theta_s^{(b)} = \frac{\text{Proportion of ones in bootstrap sample \# 1}}{\text{Proportion of ones in bootstrap sample \# 2}}$$

  - ∗ We can use $\{\theta_s^{(1)}, \cdots, \theta_s^{(B)}\}$ to approximate the distribution of $\widehat{\theta}_s$.

6.1 The Bootstrap Principle

- **Principle of Bootstrap:** Suppose we have one realization $x_1, \cdots, x_n$ of i.i.d. random variables $X_1, \cdots, X_n$ following a population P. We are interested in the distribution of a statistic $T(X_{1:n})$.
  - Assume that the true population P is known.
    * Compute the exact distribution of $T(X_{1:n})$ using P.
    * Draw new sample sets $\{x_{1:n}^{(b)}\}$, $b = 1, \cdots, B$, from P, compute $T(x_{1:n}^{(b)})$, and use $T(x_{1:n}^{(1)}), \cdots, T(x_{1:n}^{(B)})$ to estimate the distribution of $T(X_{1:n})$.
  - The true population P is unknown in most cases.
    * Use limiting theories to develop the asymptotic distribution of $T(X_{1:n})$.
    * **Bootstrap:** Consider $\{x_1, \cdots, x_n\}$ as the "true" population. Draw new sample sets $\{x_1^{*(b)}, \cdots, x_n^{*(b)}\}$, $b = 1, \cdots, B$, from $\{x_1, \cdots, x_n\}$ randomly with replacement, compute $T(x_{1:n}^{*(b)})$, and use $T(x_{1:n}^{*(1)}), \cdots, T(x_{1:n}^{*(B)})$ to estimate the distribution of $T(X_{1:n})$.

6.1 The Bootstrap Principle

- **Example:** Suppose that $X_1, \cdots, X_n$ are i.i.d. with $\mathrm{Var}(X_i) = \sigma_X^2$ and $Y_1, \cdots, Y_m$ are i.i.d. with $\mathrm{Var}(Y_i) = \sigma_Y^2$. We want to estimate $r = \sigma_X^2/\sigma_Y^2$.
  - We can estimate r by
  $$\hat{r} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2}{\frac{1}{m-1}\sum_{i=1}^{m}\left(Y_i - \overline{Y}_m\right)^2},$$
  where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\overline{Y}_m = \frac{1}{m}\sum_{i=1}^{m} Y_i$.
  - **How to determine the bias and variance of $\hat{r}$?**
  - **Bootstrap:** For $b = 1, \cdots, B$,
    * Randomly draw $\{X_1^{*(b)}, \cdots, X_n^{*(b)}\}$ with replacement from $\{X_1, \cdots, X_n\}$ and draw $\{Y_1^{*(b)}, \cdots, Y_m^{*(b)}\}$ with replacement from $\{Y_1, \cdots, Y_m\}$.
    * Compute
    $$\tilde{r}^{(b)} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i^* - \overline{X}_n^*\right)^2}{\frac{1}{m-1}\sum_{i=1}^{m}\left(Y_i^* - \overline{Y}_m^*\right)^2}.$$

6.1 The Bootstrap Principle

- – We can show that given $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$, the distribution of $\tilde{r}^{(b)} - \hat{r}$ is approximately the same as the distribution of $\hat{r} - r$. So we have

$$\text{bias}(\hat{r}) = E(\hat{r}) - r \approx \frac{1}{B} \sum_{b=1}^{B} \left(\tilde{r}^{(b)} - \hat{r}\right)$$

and

$$\text{Var}(\hat{r}) \approx \frac{1}{B-1} \sum_{b=1}^{B} \left(\tilde{r}^{(b)} - \overline{r}\right)^2,$$

where $\overline{r} = \frac{1}{B} \sum_{b=1}^{B} \tilde{r}^{(b)}$.

6.1 The Bootstrap Principle

- **Jackknife (Quenouille, 1949, 1956):** Suppose that $X_1, \cdots, X_n$ are i.i.d. samples. We want to estimate the *bias and variance* of an estimator $T(X_{1:n})$ for θ.
  - Define $X_{(-i)} = (X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)$ and
$$\tilde{\theta}_{(-i)} = T\big(X_{(-i)}\big)$$
for $i = 1, 2, \cdots, n$.
  - We can use $\tilde{\theta}_{(-i)}$, $i = 1, \cdots, n$, to estimate the bias and variance of $\hat{\theta} = T(X_{1:n})$.

- **Jackknife for Bias:**

$$\text{bias}\big(\hat{\theta}\big) \approx (n-1)\cdot\left[\frac{1}{n}\sum_{i=1}^{n}\big(\tilde{\theta}_{(-i)}-\hat{\theta}\big)\right] \triangleq \widehat{\text{bias}}_{Jack}\big(\hat{\theta}\big).$$

- **Example:** Suppose $X_1, \cdots, X_n$ are i.i.d. with $\text{Var}(X_i) = \sigma^2 < \infty$. Consider $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\big(X_i - \overline{X}_n\big)^2$.

  – The true bias is $E\big(\hat{\sigma}^2\big) - \sigma^2 = (n-1)\sigma^2/n - \sigma^2 = -\sigma^2/n$.

  – Note that

$$\begin{aligned} E\big(\tilde{\sigma}^2_{(-i)} - \hat{\sigma}^2\big) &= E\big(\tilde{\sigma}^2_{(-i)} - \sigma^2\big) + E\big(\sigma^2 - \hat{\sigma}^2\big) \\ &= -\sigma^2/(n-1) + \sigma^2/n = -\sigma^2/[n(n-1)]. \end{aligned}$$

  Then

$$E\left[\widehat{\text{bias}}_{Jack}\big(\hat{\theta}\big)\right] = (n-1)\cdot(-\sigma^2)/[n(n-1)] = -\sigma^2/n = E\big(\hat{\sigma}^2\big) - \sigma^2.$$

- **Jackknife for Variance:**

$$\mathrm{Var}\big(\hat{\theta}\big) \approx (n-1) \cdot \left[\frac{1}{n}\sum_{i=1}^{n}\Big(\tilde{\theta}_{(-i)} - \bar{\tilde{\theta}}\Big)^2\right] \triangleq \widehat{\mathrm{Var}}_{Jack}\big(\hat{\theta}\big),$$

where $\bar{\tilde{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\tilde{\theta}_{(-i)}$.

- **Example:** Suppose $X_1, \cdots, X_n$ are i.i.d. with $E(X_i) = \theta$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Consider $\hat{\theta} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
  - The true variance of $\hat{\theta} = \overline{X}_n$ is σ^2/n.
  - Note that $\tilde{\theta}_{(-i)} = \frac{1}{n-1}\sum_{j\neq i} X_i = \frac{1}{n-1}\big(n\overline{X}_n - X_i\big)$ and $\bar{\tilde{\theta}} = \overline{X}_n$. Then

$$\begin{aligned} E\left[\widehat{\mathrm{Var}}_{Jack}\big(\hat{\theta}\big)\right] &= \frac{n-1}{n} E\left[\sum_{i=1}^{n}\left(\frac{n\overline{X}_n - X_i}{n-1} - \overline{X}_n\right)^2\right] \\ &= \frac{n-1}{n} E\left[\sum_{i=1}^{n}\left(\frac{\overline{X}_n - X_i}{n-1}\right)^2\right] = \frac{n-1}{n}\cdot\frac{\sigma^2}{n-1} = \sigma^2/n. \end{aligned}$$

6.1 The Bootstrap Principle

- **Example:** Let $X_1, \cdots, X_n$ be i.i.d. samples with finite expectation. We can write $X_i = \theta + \varepsilon_i$, where $\theta = E(X_i)$ and $\varepsilon_i = X_i - E(X_i)$.
  - We can estimate θ by $\hat{\theta} := \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. How to find a 95% *confidence interval* for θ?
  - **Confidence Interval (CI):** A pair of statistics $L(X_{1:n}) < U(X_{1:n})$ construct a level $100(1-\alpha)\%$ *confidence interval* for the parameter θ if

$$P_\theta\Big(L(X_{1:n}) < \theta < U(X_{1:n})\Big) > 1 - \alpha \quad \text{for all } \theta.$$

    * Note that θ is deterministic, but the interval $\big(L(X_{1:n}), U(X_{1:n})\big)$ is random.
    * When a realization $x_{1:n} = (x_1, \cdots, x_n)$ is observed, θ is either in $\big(L(x_{1:n}), U(x_{1:n})\big)$ or not. There is no uncertainty.
    * For a given confidence level, there are many different CI's. Usually, we want to find the CI with the shortest length.

6.1 The Bootstrap Principle

- – Given that $\varepsilon_i \sim N(0,1)$, then $\hat{\theta} - \theta = \frac{1}{n}\sum_{i=1}^n \varepsilon_i \sim N(0, 1/n)$. We have

$$P\Big(-1.96/\sqrt{n} < \hat{\theta} - \theta < 1.96/\sqrt{n}\Big) = 0.95$$
$$\Rightarrow P\Big(\hat{\theta} - 1.96/\sqrt{n} < \theta < \hat{\theta} + 1.96/\sqrt{n}\Big) = 0.95.$$

  So a level 95% confidence interval for θ is $\Big(\overline{X}_n - 1.96/\sqrt{n},\ \overline{X}_n + 1.96/\sqrt{n}\Big)$.

  – Usually, the distribution of ε_i is unknown. We consider using the bootstrap to construct a level 95% confidence interval for θ, especially when the sample size n is not very large.

  – **Bootstrap:** For $b = 1, \cdots, B$,

    * Randomly draw $\{X_1^{*(b)}, \cdots, X_n^{*(b)}\}$ with replacement from $\{X_1, \cdots, X_n\}$.
    * Compute $\tilde{\theta}^{(b)} = \frac{1}{n}\sum_{i=1}^n X_i^{*(b)}$.

  – We can show that given $X_1, \cdots, X_n$, the distribution of $\tilde{\theta}^{(b)} - \hat{\theta}$ is approximately the same as the distribution of $\hat{\theta} - \theta$. So we can use $\tilde{\theta}^{(1)} - \hat{\theta}$, $\cdots$, $\tilde{\theta}^{(B)} - \hat{\theta}$ to estimate the distribution of $\hat{\theta} - \theta$.

6.1 The Bootstrap Principle

- – Let $\tilde{q}_{0.025}$ and $\tilde{q}_{0.975}$ be the 2.5% and 97.5% quantiles of the set $\{\tilde{\theta}^{(1)}, \cdots, \tilde{\theta}^{(B)}\}$, respectively. We have

$$\begin{aligned} 0.95 &\approx P\big(\hat{q}_{0.025} < \tilde{\theta}^{(b)} < \hat{q}_{0.975} \,|\, X_1, \cdots, X_n\big) \\ &= P\big(\hat{q}_{0.025} - \hat{\theta} < \tilde{\theta}^{(b)} - \hat{\theta} < \hat{q}_{0.975} - \hat{\theta} \,|\, X_1, \cdots, X_n\big) \\ &\approx P\big(\hat{q}_{0.025} - \hat{\theta} < \hat{\theta} - \theta < \hat{q}_{0.975} - \hat{\theta}\big) \\ &= P\big(\hat{\theta} - (\hat{q}_{0.975} - \hat{\theta}) < \theta < \hat{\theta} - (\hat{q}_{0.025} - \hat{\theta})\big). \end{aligned}$$

  The level 95% bootstrap confidence interval for θ is

$$\big(2\hat{\theta} - \hat{q}_{0.975},\, 2\hat{\theta} - \hat{q}_{0.025}\big).$$

  – Note that $\big(2\hat{\theta} - \hat{q}_{0.975},\, 2\hat{\theta} - \hat{q}_{0.025}\big)$ may not be the same as $(\hat{q}_{0.025}, \hat{q}_{0.975})$, unless $\hat{\theta} = (\hat{q}_{0.025} + \hat{q}_{0.975})/2$.

6.1 The Bootstrap Principle

- – If $\hat{q}_{0.975} - \hat{\theta} \approx \hat{\theta} - \hat{q}_{0.025}$ (or when $\hat{\theta}$ is an unbiased estimator for θ), we can also use $(\hat{q}_{0.025}, \hat{q}_{0.975})$ as the confidence interval for θ.

  – When the sample size n is large, we can often obtain that

  $$\sqrt{n}\big(\hat{\theta} - \theta\big) \xrightarrow{d} N(0, \sigma^2),$$

  where σ^2/n is the (asymptotic) variance of $\hat{\theta}$, which can be estimated by

  $$\hat{\sigma}^2/n \approx \frac{1}{B-1}\sum_{b=1}^{B}\big(\tilde{\theta}^{(b)} - \overline{\theta}\big)^2,$$

  where $\overline{\theta} = \frac{1}{B}\sum_{b=1}^{B}\tilde{\theta}^{(b)}$. A level 95% confidence interval can also be constructed as

  $$\big(\hat{\theta} - 1.96\sqrt{\hat{\sigma}^2/n},\, \hat{\theta} + 1.96\sqrt{\hat{\sigma}^2/n}\big).$$

6.2 Bootstrap for Linear Regression

- **Linear Regression:** Consider a linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \varepsilon_i = X_i'\beta + \varepsilon_i,$$

where Y_i is the response varaible and $X_i = (1, X_{i,1}, \cdots, X_{i,p})$ are the corresponding covariates. The error ε_i satisfies $E(\varepsilon_i \mid X_i) = 0$ and $\mathrm{Var}(\varepsilon_i \mid X_i) < \infty$. Given the observations (Y_i, X_i'), $i = 1, \cdots, n$, we want to make inference of the linear coefficients $\beta = (\beta_0, \beta_1, \cdots, \beta_p)'$.

  – *Ordinary Least Square (OLS) Estimator*: We can estimate β by

$$\begin{aligned}\hat{\beta} &= \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i'\beta)^2 \\ &= \arg\min_{\beta} (\boldsymbol{Y} - \boldsymbol{X}\beta)'(\boldsymbol{Y} - \boldsymbol{X}\beta) \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y}),\end{aligned}$$

  where $\boldsymbol{Y} := (Y_1, \cdots, Y_n)'$ is a $n \times 1$ vector and $\boldsymbol{X} := (X_1, \cdots, X_n)'$ is a $n \times (p+1)$ matrix (We consider the case $n \gg p$).

6.2 Bootstrap for Linear Regression

- – Note that $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_p)'$. We have

$$\begin{aligned} \hat{\beta} - \beta &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\big[\boldsymbol{X}'(\boldsymbol{X}\beta + \boldsymbol{\varepsilon})\big] - \beta \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}. \end{aligned}$$

  – If we consider $\boldsymbol{X}$ as a constant matrix, and assume that $\varepsilon_1, \cdots, \varepsilon_p$ are i.i.d. following the $N(0, \sigma^2)$ distribution, then

$$\hat{\beta} - \beta \sim N(0, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

  We can use this distribution (with an estimated $\hat{\sigma}^2$) to construct confidence interval for each β_j, $j = 0, 1, \cdots, p$.

  – If we don't know the distribution of ε_i, we can use the bootstrap method (or limiting theories) to construct confidence interval for β_j.

6.2 Bootstrap for Linear Regression

- **Empirical Bootstrap:** Assume that $(Y_1, X_1), \cdots, (Y_n, X_n)$ are i.i.d., and ε_i is independent of X_i.
  - For $b = 1, \cdots, B$,
    * Randomly draw $\{(Y_i^{*(b)}, X_i^{*(b)})\}_{i=1}^n$ with replacement from $\{(Y_i, X_i)\}_{i=1}^n$.
    * Compute

$$\tilde{\beta}^{(b)} = \left[\left(\boldsymbol{X}^{*(b)}\right)'\boldsymbol{X}^{*(b)}\right]^{-1}\left[\left(\boldsymbol{X}^{*(b)}\right)'\boldsymbol{Y}^{*(b)}\right],$$

where $\boldsymbol{X}^{*(b)} := \left(X_1^{*(b)}, \cdots, X_n^{*(b)}\right)'$ and $\boldsymbol{Y}^{*(b)} := \left(Y_1^{*(b)}, \cdots, Y_n^{*(b)}\right)'$.

  - A 95% confidence interval for β_j is constructed as

$$\left(2\hat{\beta}_j - \tilde{q}_{j,0.975},\, 2\hat{\beta}_j - \tilde{q}_{j,0.025}\right),$$

where $\tilde{q}_{j,0.025}$ and $\tilde{q}_{j,0.975}$ are the 2.5% and 97.5% quantiles of the set $\{\tilde{\beta}_j^{(1)}, \cdots, \tilde{\beta}_j^{(B)}\}$, respectively.

6.2 Bootstrap for Linear Regression

- **Residual Bootstrap:** When there are some influential observations (outliers) in X_i's, the empirical bootstrap may lead to a bad result (we can not assume that X_i's are i.i.d.). Define $\hat{\varepsilon}_i := Y_i - X_i'\hat{\beta}$.
  - For $b = 1, \cdots, B$,
    * Randomly draw $\{\hat{\varepsilon}_i^{*(b)}\}_{i=1}^n$ with replacement from $\{\hat{\varepsilon}_i\}_{i=1}^n$.
    * Calculate $Y_i^{*(b)} = X_i\hat{\beta} + \hat{\varepsilon}_i^{*(b)}$, $i = 1, \cdots, n$.
    * Compute

$$\tilde{\beta}^{(b)} = \big(\boldsymbol{X}'\boldsymbol{X}\big)^{-1}\big(\boldsymbol{X}'\boldsymbol{Y}^*\big),$$

      where $\boldsymbol{Y}^{*(b)} = \big(Y_1^{*(b)}, \cdots, Y_n^{*(b)}\big)'$.
  - A 95% confidence interval for β_j is constructed as

$$\big(2\hat{\beta}_j - \tilde{q}_{j,0.975},\, 2\hat{\beta}_j - \tilde{q}_{j,0.025}\big),$$

    where $\tilde{q}_{j,0.025}$ and $\tilde{q}_{j,0.975}$ are the 2.5% and 97.5% quantiles of the set $\{\tilde{\beta}_j^{(1)}, \cdots, \tilde{\beta}_j^{(B)}\}$, respectively.

6.2 Bootstrap for Linear Regression

- **Wild Bootstrap:** When ε_i is not independent of X_i, for example, $E(\varepsilon_i \mid X_i) = 0$, but $\mathrm{Var}(\varepsilon_i \mid X_i)$ depends on X_i, we need to use the *wild bootstrap* method.
  - For $b = 1, \cdots, B$,
    * Draw $R_1, \cdots, R_n$ i.i.d. from a distribution with zero mean and unit variance, for example, $N(0,1)$.
    * Calculate $Y_i^{*(b)} = X_i\hat{\beta} + R_i\hat{\varepsilon}_i$, $i = 1, \cdots, n$.
    * Compute

$$\tilde{\beta}^{(b)} = \big(\boldsymbol{X}'\boldsymbol{X}\big)^{-1}\big(\boldsymbol{X}'\boldsymbol{Y}^*\big),$$

    where $\boldsymbol{Y}^{*(b)} = \big(Y_1^{*(b)}, \cdots, Y_n^{*(b)}\big)'$.
  - A 95% confidence interval for β_j is constructed as

$$\big(2\hat{\beta}_j - \tilde{q}_{j,0.975},\ 2\hat{\beta}_j - \tilde{q}_{j,0.025}\big),$$

  where $\tilde{q}_{j,0.025}$ and $\tilde{q}_{j,0.975}$ are the 2.5% and 97.5% quantiles of the set $\{\tilde{\beta}_j^{(1)}, \cdots, \tilde{\beta}_j^{(B)}\}$, respectively.

6.2 Bootstrap for Linear Regression

- **Remarks:**

  – When ε_i is not independent of X_i, we can not break the (X_i, ε_i) pairs in the bootstrap.

  – It is easy to verify that

$$E\big(R_i\varepsilon_i \,|\, X_i\big) = E\big(R_i \,|\, X_i\big)E\big(\varepsilon_i \,|\, X_i\big) = 0$$

  and

$$\begin{aligned}
\mathrm{Var}\big(R_i\varepsilon_i \,|\, X_i\big) &= E\big(R_i^2\varepsilon_i^2 \,|\, X_i\big) - 0 \\
&= E\big(R_i^2 \,|\, X_i\big)E\big(\varepsilon_i^2 \,|\, X_i\big) \\
&= E\big(\varepsilon_i^2 \,|\, X_i\big) \\
&= \mathrm{Var}\big(\varepsilon_i \,|\, X_i\big).
\end{aligned}$$

6.2 Bootstrap for Linear Regression

- **Bootstrap for Logistic Regression:** Consider a logistic regression model

$$P(Y_i = 1; X_i, \beta) = \frac{\exp\{X_i^T \beta\}}{1 + \exp\{X_i^T \beta\}},$$

where $Y_i \in \{0, 1\}$ and X_i, $i = 1, \cdots, n$, are p-dimensional covariates. We can use the bootstrap to make inference of β.

  – We can find the MLE $\hat{\beta}$ using the observed data (Y_i, X_i), $n = 1, \cdots, n$.

  – We can define

$$\begin{aligned} \hat{\varepsilon}_i &:= Y_i - \hat{E}(Y_i \,|\, X_i) \\ &= Y_i - \frac{\exp\{X_i^T \hat{\beta}\}}{1 + \exp\{X_i^T \hat{\beta}\}}. \end{aligned}$$

  However, we can not use $\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n$ for the bootstrap. (**Why?**)

6.2 Bootstrap for Linear Regression

- – **Bootstrap:** For $b = 1, \cdots, B$,
  - ∗ For $i = 1, \cdots, n$, draw $Y_i^{*(b)}$ from the distribution with

$$Y_i^{*(b)} = \begin{cases} 1, & \text{with probability } \frac{\exp\{X_i^T \hat{\beta}\}}{1+\exp\{X_i^T \hat{\beta}\}}; \\ 0, & \text{with probability } \frac{1}{1+\exp\{X_i^T \hat{\beta}\}}. \end{cases}$$

  - ∗ Calculate the MLE $\tilde{\beta}^{(b)}$ using the data $\{(Y_i^*, X_i)\}_{i=1}^n$.

  – Then we can use $\{\tilde{\beta}^{(1)} - \hat{\beta}, \cdots, \tilde{\beta}^{(B)} - \hat{\beta}\}$ to approximate the distribution of $\hat{\beta} - \beta$ and make inference of β.

6.3 Bootstrap for Tests

- **Hypothesis Test:** Suppose $X_1, \cdots, X_n$ are i.i.d. from a population $P \in \mathcal{P}$, where $\mathcal{P}$ is a family of populations. Let $\mathcal{P}_0$ and $\mathcal{P}_1$ be two complementary subsets of $\mathcal{P}$, *i.e.*, $\mathcal{P}_0 \cup \mathcal{P}_1 = \mathcal{P}$ and $\mathcal{P}_0 \cap \mathcal{P}_1 = \emptyset$. We want to test

$$H_0 : P \in \mathcal{P}_0 \quad \text{versus} \quad H_1 : P \in \mathcal{P}_1$$

based on the observations $X_1 = x_1, \cdots, X_n = x_n$.

  - *Test Statistic:* Usually, we perform test through a test statistic $T(X_{1:n})$, that is, we reject the *null hypothesis* H_0 if $T(X_{1:n}) \in C$ and accept H_0 if $T(X_{1:n}) \notin C$. Here C is called the *critical region* or *rejection region*.
  - *p-Value:* The p-value for the observed $T(x_{1:n})$ is the **probability** that the test statistic $T(X_{1:n})$ is **more extreme** than $T(x_{1:n})$ **under** H_0. For example, if we reject H_0 when $T(X_{1:n}) > c$, then the p-value is $P\big(T(X_{1:n}) > T(x_{1:n})\big)$ when $P \in \mathcal{P}_0$.
  - Given a significance level α, we reject H_0 if the p-value is less than α.

6.3 Bootstrap for Tests

- **One-Sample Mean Test:** Suppose $X_1, \cdots, X_n$ are i.i.d with $E(X_i) = \mu$. We want to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$ using the observed data $x_1, \cdots, x_n$.
  - Consider the test statistic
  $$T_n := T(X_{1:n}) = \frac{\sqrt{n}(\overline{X}_n - \mu_0)}{S_n},$$
  where $S_n = \left[\frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2\right]^{1/2}$ is the *sample standard deviation.*
  - We reject H_0 if $T_n > c$ and accept H_0 if $T_n \leq c$, where c is called the *critical value.*
  - The p-value is $P\big(T_n > T(x_{1:n})\big)$ when $H_0 : \mu = \mu_0$ is true.
  - If $X_i \sim N(\mu, \sigma^2)$, then $T_n \sim t(n-1)$ under H_0. Here $t(n-1)$ denotes the Student's t distribution with *degrees of freedom* $n-1$. We can use the $t(n-1)$ distribution to compute $P\big(T_n > T(x_{1:n})\big)$.

6.3 Bootstrap for Tests

- – When we don't know the distribution of X_i, we can use the bootstrap to compute the p-value for $T(x_{1:n})$.

  – Note that we need to know the **distribution of T_n under H_0**, however, the observe data $x_1, \cdots, x_n$ may not satisfy $E(X_i) = \mu_0$.

  – Let $\tilde{x}_i = x_i - \overline{x}_n + \mu_0$, then $\tilde{x}_1, \cdots, \tilde{x}_n$ satisfies $E(\tilde{X}_i) = \mu_0$.

  – **Bootstrap:** For $b = 1, \cdots, B$,

    * Randomly draw $\{\tilde{x}_1^{*(b)}, \cdots, \tilde{x}_n^{*(b)}\}$ with replacement from $\{\tilde{x}_1, \cdots, \tilde{x}_n\}$.
    * Compute $\tilde{T}_n^{(b)} = T\big(\tilde{x}_{1:n}^{*(b)}\big)$.

  – The bootstrap p-value is

$$\hat{P}\big(T_n > T(x_{1:n})\big) = \frac{1}{B}\sum_{b=1}^{B} I\big(\tilde{T}_n^{(b)} > T(x_{1:n})\big),$$

  where $I(\cdot)$ is the indicator function. **Note that we compare $\tilde{T}_n^{(b)}$ with $T(x_{1:n})$, but not $T(\tilde{x}_{1:n})$.**

6.3 Bootstrap for Tests

- **Two-Sample Mean Test:** Suppose that $X_1, \cdots, X_n$ are i.i.d. with $E(X_i) = \mu_X$, $\mathrm{Var}(X_i) = \sigma_X^2$, and $Y_1, \cdots, Y_m$ are i.i.d. with $E(Y_i) = \mu_Y$, $\mathrm{Var}(Y_i) = \sigma_Y^2$, . We want to test $H_0 : \mu_X = \mu_Y$ versus $H_1 : \mu_X \neq \mu_Y$ based on observed $x_1, \cdots, x_n$ and $y_1, \cdots, y_m$.
  - **Case 1: $\sigma_X^2 = \sigma_Y^2$.** Consider the test statistic

$$T_{n,m} := T(X_{1:n}, Y_{1:m}) = \frac{\left(\overline{X}_n - \overline{Y}_m\right)/\sqrt{1/m + 1/n}}{\sqrt{\frac{1}{m+n-2}\left[\sum_{i=1}^m (X_i - \overline{X}_n)^2 + \sum_{i=1}^n (Y_i - \overline{Y}_m)^2\right]}}.$$

  - We reject H_0 when $|T(x_{1:n}, y_{1:m})| > c$. The p-value for $\{x_{1:n}, y_{1:m}\}$ is $P\big(|T_{n,m}| > |T(x_{1:n}, y_{1:m})|\big)$ when H_0 is true.
  - If $X_i \sim N(\mu_X, \sigma^2)$ and $Y_j \sim N(\mu_Y, \sigma^2)$, $T_{n,m}$ follows a $t(n + m - 2)$ distribution when $H_0 : \mu_X = \mu_Y$ holds. We can use the $t(n + m - 2)$ distribution to compute the p-value $P\big(|T_{n,m}| > |T(x_{1:n}, y_{1:m})|\big)$.

- – **Bootstrap:** Consider the distribution of $T_{n,m}$ under H_0. Let $\widetilde{x}_i = x_i - \overline{x}_n$ and $\widetilde{y}_j = y_j - \overline{y}_m$. Then $E(\tilde{X}_i) = E(\tilde{Y}_j)$ (H_0 holds).
  - $*$ For $b = 1, \cdots, B$,
    - $\cdot$ Randomly draw $\{\tilde{x}_1^{*(b)}, \cdots, \tilde{x}_n^{*(b)}\}$ with replacement from $\{\tilde{x}_1, \cdots, \tilde{x}_n\}$ and $\{\tilde{y}_1^{*(b)}, \cdots, \tilde{y}_m^{*(b)}\}$ with replacement from $\{\tilde{y}_1, \cdots, \tilde{y}_m\}$.
    - $\cdot$ Compute $\tilde{T}_{n,m}^{(b)} = T\big(\tilde{x}_{1:n}^{*(b)}, \tilde{y}_{1:m}^{*(b)}\big)$.
  - $*$ The bootstrap p-value is

$$\hat{P}\big(|T_{n,m}| > |T(x_{1:n}, y_{1:m})|\big) = \frac{1}{B}\sum_{b=1}^{B} I\big(|\tilde{T}_{n,m}^{(b)}| > |T(x_{1:n}, y_{1:m})|\big).$$

Note that we compare $\tilde{T}_{n,m}^{(b)}$ with $T(x_{1:n}, y_{1:m})$, but not $T(\tilde{x}_{1:n}, \tilde{y}_{1:m})$.

- – **Case 2: $\sigma_X^2 \neq \sigma_Y^2$.** Consider the test statistic

$$R_{n,m} := R(X_{1:n}, Y_{1:m}) = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{S_X^2/n + S_Y^2/m}},$$

  where $S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$ and $S_Y^2 = \frac{1}{m-1}\sum_{j=1}^{m}\left(Y_j - \bar{Y}_m\right)^2$.

  – We reject H_0 when $|R(x_{1:n}, y_{1:m})|$ is large. The p-value for $\{x_{1:n}, y_{1:m}\}$ is $P\big(|R_{n,m}| > |R(x_{1:n}, y_{1:m})|\big)$ when H_0 is true.

  – Note that even when X_i and Y_j are normally distributed, $R_{n,m}$ no longer follows a t-distribution.

6.3 Bootstrap for Tests

- – **Bootstrap:** Consider the distribution of $R_{n,m}$ under H_0. Let $\widetilde{x}_i = x_i - \overline{x}_n$ and $\widetilde{y}_j = y_j - \overline{y}_m$. Then $E(\tilde{X}_i) = E(\tilde{Y}_j)$ (H_0 holds).
  - * For $b = 1, \cdots, B$,
    - · Randomly draw $\{\tilde{x}_1^{*(b)}, \cdots, \tilde{x}_n^{*(b)}\}$ with replacement from $\{\tilde{x}_1, \cdots, \tilde{x}_n\}$ and $\{\tilde{y}_1^{*(b)}, \cdots, \tilde{y}_m^{*(b)}\}$ with replacement from $\{\tilde{y}_1, \cdots, \tilde{y}_m\}$.
    - · Compute $\tilde{R}_{n,m}^{(b)} = R\big(\tilde{x}_{1:n}^{*(b)}, \tilde{y}_{1:m}^{*(b)}\big)$.
  - * The bootstrap p-value is

$$\hat{P}\big(|R_{n,m}| > |R(x_{1:n}, y_{1:m})|\big) = \frac{1}{B}\sum_{b=1}^{B} I\big(|\tilde{R}_{n,m}^{(b)}| > |R(x_{1:n}, y_{1:m})|\big).$$

Note that we compare $\tilde{R}_{n,m}^{(b)}$ with $R(x_{1:n}, y_{1:m})$, but not $R(\tilde{x}_{1:n}, \tilde{y}_{1:m})$.

Homework

1. Let $X_1, \cdots, X_{100}$ be i.i.d. from the Bernoulli(p) distribution. Suppose that we observed 24 ones and 76 zeros in one realization of $X_1, \cdots, X_{100}$ and estimated p by $\hat{p} = 0.24$.

   (1) Use the bootstrap to find the variance of $\hat{p} - p$.

   (2) Use the bootstrap to construct a 95% confidence interval for p.

2. Suppose that (X_{1i}, X_{2i}, Y_i), $i = 1, \cdots, n$ are i.i.d. following the logistic model

$$P(Y_i = 1 \mid X_{1i}, X_{2i}; \beta) = \frac{\exp\{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}\}}{1 + \exp\{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}\}}.$$

   Use the bootstrap to find the 95% confidence intervals for β_0, β_1 and β_2, respectively.