Бутстреп

Браженко Дмитрий

December 19, 2017

План

1. **Непараметрический бутстреп**

2. **Оценка доверительных интервалов на основе бутстрепа**
   - Нормальный интервал
   - Центральный интервал
   - Метод складного ножа

Стандартная постановка задачи

Идея

- Есть некоторая выборка $\{X_i\}_{i=1}^n \in R$, порожденная некоторым распределением.
- Имеется некоторая статистика $T_n = T_n(X_1, X_2, ..., X_n)$
- Цель – определить дисперсию $V_F(T_n)$

Тогда

- $T_n = \overline{X_n}$
- $V_F(T_n) = \frac{\sigma^2}{n}$
- $\sigma^2 = \int (x-\mu)^2 dF(x)$, а $\mu = \int x dF(x)$

Идея бутстрепа

To pull oneself over a fence by one's bootstraps

Как на самом деле

$F \to X_1, \ldots, X_n \to T_n = g(X_1, X_2, \ldots, X_n)$

С точки зрения бутстрепа

$\hat{F}_n \to X_1^*, \ldots, X_n^* \to T_n^* = g(X_1^*, X_2^*, \ldots, X_n^*)$

Как получить $X_1^*, \ldots, X_n^*$

Из имеющейся выборки выбрать сделайть новую выборку такой же длины, при этом каждый раз элемент каждый раз попадает в новую выборку равно вероятно

Пример про МНК-оценку

- Имеется два наблюдения: $(x_1, y_1) = (1, 1)$, $(x_2, y_2) = (2, 3)$
- Требуется оценить МНК-модель $y_i = \theta x_i + \epsilon_i$
- МНК-оценка: $\hat{\theta} = 1.2$
- $$(x_1, y_1), (x_2, y_2) = \begin{cases} (1,1),(1,1) & p = 1/4 \\ (1,1),(2,3) & p = 1/4 \\ (2,3),(1,1) & p = 1/4 \\ (2,3),(2,3) & p = 1/4 \end{cases}$$
- Получаются МНК-оценки

$$\hat{\theta}^* = \begin{cases} 1 & p = 1/4 \\ 1.2 & p = 1/2 \\ 1.5 & p = 1/4 \end{cases}$$

Алгоритм

1. Выбираем $X_1^*, ..., X_n^* \sim \hat{F}_n$
2. Вычисляем $T_n^* = g(X_1^*, ..., X_n^*)$
3. Повторяем шаги 1-2, пока не получится $T_{n,1}^*, ..., T_{n,B}^*$
4. Получаем $v_{boot} \approx \frac{1}{B}\sum_{b=1}^{B}(T_{n,b}^* - \frac{1}{b}\sum_{r=1}^{B} T_{n,r}^*)^2$

Нормальный интервал

Если можно предположить, что данные распределены нормально, то можно воспользоваться нормальным доверительным интервалом.

$$(T_n - z_{\alpha/2}\hat{se}, T_n + z_{\alpha/2}\hat{se})$$

- $z_\alpha : F_{N(0,1)(z_\alpha)=1-\alpha}$
- $\hat{se} = \sqrt{v_{boot}}$

Оценка параметра в распределении

- $X_1, ..., X_n \sim t_3$, $n = 25$
- $\Theta = T(F) = \frac{q_{0.75} - q_{0.25}}{1.34}$
- Доверительный интервал $(0.71, 1.78)$
- Реальное значение 1.14

Исследование влияния реагентов на осадки, 1975

Требеутся оценить влияния распыления реагентов на вероятность дождя.

- Имеется 26 облаков, на каждое распылили реагент. Также контрольная выборка из 26 других облаков.
- После распыления измерили "объем выпавшего дождя" (acre-foot)
- Как оценить влияние реагента?
- θ – разность в средних значения выпадения осадков.
- $\hat{\theta} = 277.9$
- $\theta \in (10.6, 545.2)$
- Величина эффекта не так очевидна.

Центральный интервал

- $\Theta = T(F)$, $\hat{\Theta}_n = T(\hat{F}_n)$, а $R_n = \Theta_n - \Theta$
- $H(r) = P_F\{R_n < r\}$
- $C_n^* = (a, b), a = \hat{\Theta}_n - H^{-1}(1 - \alpha/2)$, $b = \hat{\Theta}_n - H^{-1}(\alpha/2)$
- $P(a < \Theta < b) = P(a - \Theta_n < \Theta - \Theta_n < b - \Theta_n) =$
  $= P(\Theta_n - b < \Theta_n - \Theta < \Theta_n - a) = P(\Theta_n - b < R_n < \Theta_n - a) =$
  $= H(H^{-1}(1 - \alpha/2)) - H(H^{-1}(\alpha/2)) = 1 - \alpha$

Центральный интервал

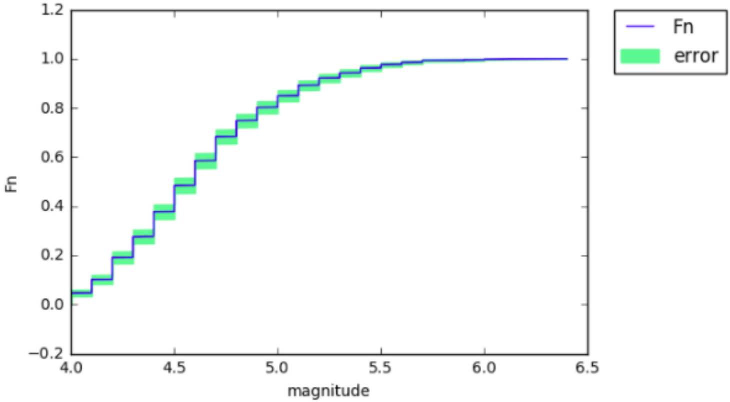
Требеутся оценить функцию распределения

$$\hat{H}(r) = \frac{1}{B}\sum_{b=1}^{B} I(\hat{\Theta}^{*}_{n,b} - \hat{\Theta}_n < r)$$

Далее берется нижняя граница и верхняя граница, получается доверительный интервал.

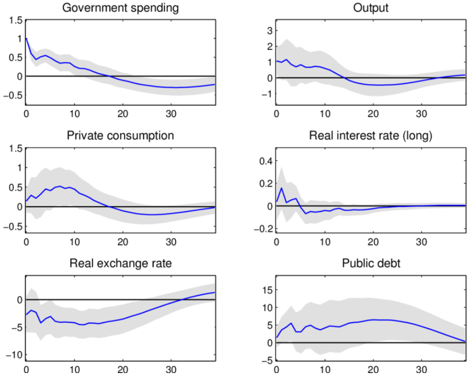
Пример про землетрясения у острова Фиджи

Требеутся оценить функцию распределения магнитуды землетрясений.



Оценка отклика в VAR-моделях

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix} + \begin{bmatrix} \alpha_{11}^1 & \alpha_{12}^1 \\ \alpha_{21}^1 & \alpha_{22}^1 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \cdots + \begin{bmatrix} \alpha_{11}^p & \alpha_{12}^p \\ \alpha_{21}^p & \alpha_{22}^p \end{bmatrix} \begin{bmatrix} x_{t-p} \\ y_{t-p} \end{bmatrix} + \begin{bmatrix} \epsilon_{x,p} \\ \epsilon_{y,p} \end{bmatrix}$$



Метод складного ножа

- $T_n = (X_1, ..., X_n)$. Рассматривается n подвыборок:
- $T_{-1} = \frac{1}{n-1} \sum_{j \neq i} X_j$
- $\overline{T_n} = \frac{\sum i{=}1^n T_{-1}}{n}$
- $v_{jack} = \frac{n-1}{n} \sum_{i=1}^{n} (T_{-i} - \overline{T_n})^2$
- $se_{jack} = v_{jack}$
- $\frac{v_{jack}(T)}{v(T)} \xrightarrow{\mathsf{P}} 1$