

Бутстрэп

Наивный бутстрэп

Краткий план:

- Общая идея бутстрэпа.
- Наивный бутстрэп.

Спасение утопающих

Бутстрэп позволяет не думать о том, как распределены статистики!

Вы всё ещё выбираете степени свободы?

Тогда мы идёт к вам!

Общая логика проверки гипотез

1. Теорема. При верной H_0 , **идеальных условиях** и $n \rightarrow \infty$ статистика $S \rightarrow \chi^2$.

$$S = \{\text{Ужасная формула пугающая студентов}\}$$

2. По имеющимся данным рассчитываем значение S_{obs} .
3. Рассчитываем P -значение — вероятность $\mathbb{P}(S > S_{obs})$.
4. Если P -значение мало, то отвергаем гипотезу H_0 .

А что если...

1. Идеальные условия нарушены.
2. Наблюдений не достаточно, чтобы считать $S \sim \chi^2$.
3. Подходящей теоремы нет.

Вместо χ^2 распределения нужно использовать верное распределение статистики S .

Идея бутстрэпа

При больших n можно **оценить закон распределения** статистики S !

И вместо обещанного теоремой χ^2 -распределения **использовать оценку распределения.**

Предупреждение

Бутстрэп является асимптотическим методом и формально требует $n \rightarrow \infty$.

Часто оказывается, что для хорошей оценки закона распределения S нужно меньшее n , чем для теоремы с идеальными условиями.

Наивный бутстрэп

Доверительный интервал для медианы

Есть случайная выборка y_1, \dots, y_n из непрерывного распределение, n велико.

Посчитали выборочную медиану \hat{m} . Хотим построить доверительный интервал для медианы m .

1. Из исходной выборки y_1, \dots, y_n построим бутстрэп-выборку y_1^*, \dots, y_n^* :
Выберем случайно n наблюдений с повторениями.
2. На базе бутстрэп-выборки посчитаем очередную выборочную медиану \hat{m}_j^* .
3. Повторим первые два шага много раз: $j = 1, \dots, 10000$.

Наивный бутстрэп: формула интервала

Хотим доверительный интервал для истинной медианы m и уже раздобыли 10000 бутстрэп выборочных медиан $\hat{m}_1^*, \dots, \hat{m}_{10000}^*$.

Доверительный интервал

$$q_{\text{left}}(\hat{m}^*) \leq m \leq q_{\text{right}}(\hat{m}^*),$$

где $q_{\text{left}}(\hat{m}^*)$ и $q_{\text{right}}(\hat{m}^*)$ — нужный левый и правый квантили.

Хочу 95% доверительный интервал

1. Отбрасываю 2.5% самых маленьких \hat{m}_j^* и 2.5% самых больших \hat{m}_j^* .
2. Крайние значения оставшихся \hat{m}_j^* и будут границами интервала.

Бутстрэп: проверка гипотез

Скалярный параметр

1. Гипотеза $H_0 : \beta_x = 42$ против $\beta_x \neq 42$.
2. Проверяем входит ли 42 в доверительный интервал.

Бутстрэп t -статистики

Краткий план:

- Бутстрэп t -статистики.
- Сравнение с наивным бутстрэпом.

Задача оценивания вероятности

Доверительный интервал для вероятности $p = \mathbb{P}(y_i > 0)$

Есть случайная выборка y_1, \dots, y_n из непрерывного распределение, n велико.

Нашли выборочную долю положительных наблюдений \hat{p} .

Теория говорит, что $\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$.

Нашли стандартную ошибку $se(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Бутстрэп t -статистики

1. Из исходной выборки y_1, \dots, y_n построим бутстрэп-выборку y_1^*, \dots, y_n^* :

Выберем случайно n наблюдений с повторениями.

2. На базе бутстрэп-выборки посчитаем:

- очередную выборочную долю \hat{p}_j^* ;
- её стандартную ошибку $se(\hat{p}_j^*)$;
- t -статистику

$$t_j^* = \frac{\hat{p}_j^* - \hat{p}}{se(\hat{p}_j^*)}.$$

3. Повторим первые два шага много раз: $j = 1, \dots, 10000$.

Формула доверительного интервала

Хотим доверительный интервал для истинной вероятности p и уже раздобыли 10000 бустрэп t -статистик $t_1^*, \dots, t_{10000}^*$.

Рецепт

Находим p из неравенства

$$q_{\text{left}}(t^*) \leq \frac{\hat{p} - p}{se(\hat{p})} \leq q_{\text{right}}(t^*)$$

Получаем

$$\hat{p} - se(\hat{p})q_{\text{right}}(t^*) \leq p \leq \hat{p} - se(\hat{p})q_{\text{left}}(t^*)$$

Не пугайтесь минуса справа!

Скорее всего $q_{\text{left}}(t^*)$ меньше нуля.

Аналогия

Классика	Бутстрэп
Параметр p	Оценка \hat{p}
Исходная выборка	Бутстрэп выборки
Оценка \hat{p}	Бутстрэп оценки \hat{p}_j^*
Стандартная ошибка $se(\hat{p})$	Стандартные ошибки $se(\hat{p}_j^*)$
Статистика $t = (\hat{p} - p)/se(\hat{p})$	Статистики $t_j^* = (\hat{p}_j^* - \hat{p})/se(\hat{p}_j^*)$

Сравнение с наивным бутстрэпом

1. Любой бутстрэп лучше, чем отсутствие.
2. Бутстрэп t -статистики лучше, чем наивный.
3. Бутстрэп t -статистики требует формулы для $se(\hat{\theta})$.
4. В качестве $se(\hat{\theta})$ можно использовать приближение.
5. Можно рассчитать $se(\hat{\theta})$ с помощью бутстрэпа в бутстрэпе.

Рекомендация

1. Используйте бутстрэп t -статистики.
2. Если нет готовой формулы для $se(\hat{\theta})$, придумайте приближенную — **бутстрэп сам поправит!**

Пример приближения стандартной ошибки

Хочу использовать бутстрэп t -статистики при построении интервала для медианы m . Не знаю никакой формулы для $se(\hat{m})$.

Я: Выборочная медиана примерно похожа на выборочное среднее. Возьму в t -статистике стандартную ошибку среднего!

$$t_j^* = \frac{\hat{m}_j^* - \hat{m}}{se(\bar{y}^*)}, \quad se(\bar{y}^*) = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i^* - \bar{y}^*)^2 / (n - 1)}.$$

Бутстрэп:

Это неправильная формула, но я сам подправляю квантили!

Бутстрэп в бутстрэпе

Краткий план:

- Как оценить $se(\hat{\theta}_j^*)$, если нет готовой формулы?
- Свойства бутстрэпа в бутстрэпе.

Пример бутстрэп-стандартной ошибки

Хочу использовать бутстрэп t -статистики при построении интервала для медианы m . Не знаю никакой формулы для $se(\hat{m})$.

Сформировали на базе исходной выборки y_1, \dots, y_n очередную бутстрэп-выборку y_1^*, \dots, y_n^* .

Нашли бутстрэп выборочную медиану \hat{m}_j^* .

Как бы найти её стандартную ошибку $se(\hat{m}_j^*)$?

Запустим **бутстрэп второго уровня**!

Бутстрэп второго уровня

Перед нами очередная бутстрэп-выборка y_1^*, \dots, y_n^* .

Алгоритм:

1. Из бутстрэп выборки y_1^*, \dots, y_n^* построим бутстрэп-выборку **второго уровня** $y_1^{**}, \dots, y_n^{**}$:
Выберем случайно n наблюдений с повторениями.
2. Посчитаем бутстрэп медиану второго уровня \hat{m}_k^{**} .
3. Повторим первые два шага много раз: $k = 1, \dots, 1000$.
4. Имея выборку $\hat{m}_1^{**}, \dots, \hat{m}_{1000}^{**}$ оценим стандартную ошибку

$$se(\hat{m}_j^*) = \sqrt{\sum_{k=1}^{1000} (\hat{m}_k^{**} - \bar{\hat{m}}^{**})^2 / (1000 - 1)}.$$

Резюме про бутстрэп в бутстрэпе

- **Минус:** медленный.

Если организовать 10000 бутстрэп-выборок первого уровня, а для каждой из них 1000 бутстрэп-выборок второго уровня, то получится 10 000 000 выборок.

- **Плюс:** Чаше точнее наивного.

При том же n номинальная доверительная вероятность ближе к фактической.

Параметрический бутстрэп

Краткий план:

- Добавим модель и предикторы!
- Вариации параметрического бутстрэпа.

Постановка задачи

Модель $y_i = \beta_1 + \beta_x x_i + \beta_w w_i + u_i$ и сомнения в предпосылках на распределение u_i .

Применили обычный МНК и получили оценки $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_x, \hat{\beta}_w)$ и наивную оценку дисперсии $\hat{\sigma}_u^2 = RSS/(n - k)$.

Хотим доверительный интервал для β_x с корректной вероятностью накрытия.

Параметрический бутстрэп

1. Для бутстрэп выборки сохраняем полностью исходную матрицу регрессоров X .
2. Генерируем бутстрэп выборку для ошибок:

$$u_i^* \sim \mathcal{N}(0; \hat{\sigma}_u^2);$$

3. Генерируем бутстрэп выборку для зависимой переменной:

$$y_i^* = \hat{\beta}_1 + \hat{\beta}_x x_i + \hat{\beta}_w w_i + u_i^*$$

Используем $\hat{\beta}$ и $\hat{\sigma}_u^2$ **исходной регрессии**.

4. Считаем очередную бутстрэп оценку коэффициента $\hat{\beta}_{xj}^*$ или t -статистику $t_j^* = (\hat{\beta}_{xj}^* - \hat{\beta}_x) / se(\hat{\beta}_{xj}^*)$.
5. Повторяем шаги два, три и четыре много раз:
 $j = 1, \dots, 10000$.

Интервал: наивный вариант

Хотим доверительный интервал для β_x .

Есть 10000 штук бутстрэп оценок $\hat{\beta}_{x,1}^*, \dots, \hat{\beta}_{x,10000}^*$.

Наивный вариант

$$q_{\text{left}}(\hat{\beta}_x^*) \leq \beta_x \leq q_{\text{right}}(\hat{\beta}_x^*)$$

Интервал: вариант с t -статистикой

Хотим доверительный интервал для β_x .

Есть 10000 штук бутстрэп t -статистик $t_1^*, \dots, t_{10000}^*$.

Вариант с t -статистикой

Находим β_x из неравенства

$$q_{\text{left}}(t^*) \leq \frac{\hat{\beta}_x - \beta_x}{se(\hat{\beta}_x)} \leq q_{\text{right}}(t^*)$$

Получаем

$$\hat{\beta}_x - se(\hat{\beta}_x)q_{\text{right}}(t^*) \leq \beta_x \leq \hat{\beta}_x - se(\hat{\beta}_x)q_{\text{left}}(t^*)$$

Дикий бутстрэп

Краткий план:

- Добавим модель и предикторы!
- Вариации дикого бутстрэпа.

Постановка задачи

Модель $y_i = \beta_1 + \beta_x x_i + \beta_w w_i + u_i$ и сомнения в предположениях на распределение u_i .

Применили обычный МНК и получили оценки $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_x, \hat{\beta}_w)$ и наивную оценку дисперсии $\hat{\sigma}_u^2 = RSS/(n - k)$.

Хотим доверительный интервал для β_x с корректной вероятностью накрытия.

Дикий бутстрэп

1. Для бутстрэп выборки сохраняем полностью исходную матрицу регрессоров X .
2. Генерируем бутстрэп выборку ошибок:

$$u_i^* \sim \dots$$

3. Генерируем бутстрэп выборку зависимой переменной:

$$y_i^* = \hat{\beta}_1 + \hat{\beta}_x x_i + \hat{\beta}_w w_i + u_i^*$$

Используем $\hat{\beta}$ исходной регрессии.

4. Считаем очередную бутстрэп оценку коэффициента $\hat{\beta}_{xj}^*$ или t -статистику $t_j^* = (\hat{\beta}_{xj}^* - \hat{\beta}_x) / se(\hat{\beta}_{xj}^*)$.
5. Повторяем шаги два, три и четыре много раз:
 $j = 1, \dots, 10000$.

Дикий бутстрэп: детали

Генерирование бутстрэп выборки ошибок u_1^*, \dots, u_n^* .

1. Рассчитываем отмасштабированные остатки исходной регрессии

$$\hat{u}_i^{sc} = \frac{\hat{u}_i}{\sqrt{1 - H_{ii}}}, \quad H = X(X'X)^{-1}X'.$$

Это действие **приравнивает** дисперсии остатков при гомоскедастичности.

2. Домножаем отмасштабированные ошибки на плюс или минус единицу

$$u_i^* = \hat{u}_i^{sc} \cdot v_i^*, \quad v_i^* \in \{-1, +1\} \text{ равновероятно.}$$

Теорема. При гомоскедастичности ошибок u_i дисперсия остатка \hat{u}_i пропорциональна $1 - H_{ii}$.

Интервал: наивный вариант

Хотим доверительный интервал для β_x .

Есть 10000 штук бутстрэп оценок $\hat{\beta}_{x,1}^*, \dots, \hat{\beta}_{x,10000}^*$.

Наивный вариант

$$q_{\text{left}}(\hat{\beta}_x^*) \leq \beta_x \leq q_{\text{right}}(\hat{\beta}_x^*)$$

Интервал: вариант с t -статистикой

Хотим доверительный интервал для β_x .

Есть 10000 штук бутстрэп t -статистик $t_1^*, \dots, t_{10000}^*$.

Вариант с t -статистикой

Находим β_x из неравенства

$$q_{\text{left}}(t^*) \leq \frac{\hat{\beta}_x - \beta_x}{se(\hat{\beta}_x)} \leq q_{\text{right}}(t^*)$$

Получаем

$$\hat{\beta}_x - se(\hat{\beta}_x)q_{\text{right}}(t^*) \leq \beta_x \leq \hat{\beta}_x - se(\hat{\beta}_x)q_{\text{left}}(t^*)$$

Парный бутстрэп

Краткий план:

- Парный бутстрэп.
- Практические рекомендации.

Парный бутстрэп — это просто!

На примере модели $y_i = \beta_1 + \beta_x x_i + \beta_w w_i + u_i$.

Есть исходные наблюдения (x_i, w_i, y_i) , где $i \in \{1, \dots, n\}$

1. Генерируем очередную бутстрэп-выборку (x_i^*, w_i^*, y_i^*) , где $i \in \{1, \dots, n\}$.

Случайно выберем n наблюдений из исходной выборки с повторениями.

2. Считаем очередную бутстрэп оценку коэффициента $\hat{\beta}_{xj}^*$ или t -статистику $t_j^* = (\hat{\beta}_{xj}^* - \hat{\beta}_x) / se(\hat{\beta}_{xj}^*)$.
3. Повторим первые два шага много раз: $j = 1, \dots, 10000$.

Доверительный интервал

Наивный вариант

$$q_{\text{left}}(\hat{\beta}_x^*) \leq \beta_x \leq q_{\text{right}}(\hat{\beta}_x^*)$$

Вариант с t -статистикой

Находим β_x из неравенства

$$q_{\text{left}}(t^*) \leq \frac{\hat{\beta}_x - \beta_x}{se(\hat{\beta}_x)} \leq q_{\text{right}}(t^*)$$

Получаем

$$\hat{\beta}_x - se(\hat{\beta}_x)q_{\text{right}}(t^*) \leq \beta_x \leq \hat{\beta}_x - se(\hat{\beta}_x)q_{\text{left}}(t^*)$$

Бутстрэп: рекомендации

Общие

1. Используйте бутстрэп!
2. Берите большое количество (10000) бутстрэп выборок.

Бутстрэп — идея, а не конкретный метод. Какой выбрать?

Без регрессоров

1. Смело берите бутстрэп t -статистики.
2. Если формулы для стандартных ошибок нет, попробуйте наивный бутстрэп или бутстрэп в бутстрэпе.

С регрессорами

1. Смело берите дикий бутстрэп t -статистики.
2. Если матрица регрессоров X не фиксирована, попробуйте парный бутстрэп.

Источники мудрости

1. Tim Hestenberg, What Teachers Should Know about the Bootstrap.
2. James MacKinnon, Bootstrap Methods in Econometrics.

Резюме: бутстрэп до регрессии

- Бутстрэп: оценка распределения вместо теорем.
- Наивный бутстрэп: сгенерируем много значений величины \hat{m}_j^* .
- Бутстрэп t -статистики: сгенерируем много значений

$$t_j^* = \frac{\hat{m}_j^* - \hat{m}}{se(m_j^*)}$$

- Бутстрэп в бутстрэпе: способ получить $se(m_j^*)$, если нет явной формулы.

Резюме: бутстрэп и регрессия

- Параметрический бутстрэп:

$$y_i^* = \hat{\beta}_1 + \hat{\beta}_x x_i + \hat{\beta}_w w_i + u_i^*, \quad u_i^* \sim \mathcal{N}(0; \hat{\sigma}^2)$$

- Дикий бутстрэп:

$$y_i^* = \hat{\beta}_1 + \hat{\beta}_x x_i + \hat{\beta}_w w_i + \hat{u}_i^{sc} v_i^*, \quad v_i^* \in \{-1, +1\}$$

- Парный бутстрэп: выбираем случайные наблюдения с повторениями.

Следующая лекция: причинно-следственные связи.