Министерство науки и высшего образования Российской Федерации

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ

ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Институт прикладной математики и компьютерных наук

РЕФЕРАТ

«использование бутстрепа и ресемплинга.»

по дисциплине «Введение в интеллектуальный анализ данных»

Выполнил:

студент группы № \_ 932204 \_

Прокопьев Д. А.\_\_\_\_\_\_\_\_\_\_\_\_

Проверил:

Преподаватель

\_\_\_Замятин А. В.\_\_\_\_\_\_\_\_\_\_\_

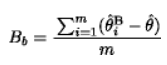
Томск – 2024

Бутстрап — это метод повторной выборки для статистического вывода. Он обычно используется для оценки доверительных интервалов, но его также можно использовать для оценки смещения и дисперсии оценщика или калибровки тестов гипотез. будут проиллюстрированы концепции бутстрапа на простом примере, описаны различные типы бутстрапов и некоторые их теоретические и практические свойства, вычисления и другие детали, а также указаны расширения, которые особенно подходят для данных об окружающей среде. Методы будут проиллюстрированы с использованием данных о концентрации тяжелых металлов в грунтовых водах и концентрации магния в крови.  
  
Статьи, иллюстрирующие разнообразие недавних средовых применений бутстрапа, можно найти в токсикологии, обследованиях рыболовства, моделировании загрязнения грунтовых вод и воздуха, хемометрике, гидрологии, филогенетике, пространственных точечных моделях, экологических индексах и многомерном суммировании.

Я проиллюстрирую концепции бутстрапа с помощью простого примера, опишу различные типы бутстрапов и некоторые из их теоретических и практических свойств, обсужу вычисления и другие детали, а также укажу расширения, которые особенно подходят для средовых данных. Методы будут проиллюстрированы с использованием данных о концентрации тяжелых металлов в грунтовых водах и концентрации магния в крови.

**Концепции бутстрапа**

Рассмотрим оценку средней концентрации тяжелого металла, например меди, в грунтовых водах из почв бассейна в долине Сан-Хоакин в США. Как это обычно бывает с данными по химии окружающей среды, некоторые значения остаются цензурированными. Они сообщаются как «ниже предела обнаружения» с указанным значением предела обнаружения. Часто наблюдения искажены. Точечные оценки среднего значения μ и стандартного отклонения σ можно рассчитать с помощью различных методов. Однако сложнее сделать статистический вывод, например, рассчитать 95% доверительный интервал для среднего значения. Обычный доверительный интервал, основанный на распределении Стьюдента, не подходит из-за цензурирования и асимметрии. Вывод, основанный на оценках максимального правдоподобия, опирается на асимптотическое распределение, которое может не подходить для небольших выборок. Сложность заключается в том, что распределение выборки оценки неизвестно. Бутстрап использует данные и вычислительную мощность для оценки этого неизвестного распределения выборки.

При наличии набора независимых и одинаково распределенных (iid) наблюдений Xi, i = 1, …, n, параметра, который можно определить как некоторую функцию θ = T(x) значений в популяции, и статистики, которая является той же функцией уравнения изображения наблюдений, бутстрап оценивает распределение выборки Fθ(x) этой функции. Данные используются в качестве оценки неизвестной кумулятивной функции распределения (CDF) Fx(x) значений в популяции. Выборки бутстрапа многократно извлекаются из оцененной популяции. Функция (например, среднее) оценивается для каждой выборки бутстрапа, давая набор значений бутстрапа уравнение изображения, i = 1, …, m. Эмпирическое распределение этих значений бутстрапа уравнение изображения оценивает теоретическое распределение выборки Fθ(x).  
  
  
Изображение уравнения распределения бутстрапа используется для оценки смещения, оценки стандартной ошибки (SE) или построения доверительного интервала для интересующей статистики. Оценки смещения бутстрапа, Bb, и SE, sb, являются эмпирическими оценками, рассчитанными из m значений бутстрапа:  
  




Метод процентильного доверительного интервала использует квантили α/2 и 1 − α/2 изображения уравнения в качестве доверительного интервала уровня 1 − α для параметра.

В данных по меди в долине Сан-Хоакин имеется 49 наблюдений, включая 14 цензурированных значений. Поскольку данные довольно асимметричны, среднее значение оценивается с использованием непараметрического оценщика. Оценочное среднее значение составляет 4,33 частей на миллион (ppm). 95% доверительный интервал для среднего значения оценивается с использованием 1000 бутстрап-выборок. Каждая бутстрап-выборка представляет собой простую случайную выборку из 49 значений, выбранных с заменой из исходных наблюдений. Поскольку бутстрап-выборка формируется с заменой, некоторые исходные наблюдения повторяются в бутстрап-выборке более одного раза. Другие наблюдения исключаются из отдельной бутстрап-выборки. Статистика оценивается для каждой бутстрап-выборки. Доверительные интервалы бутстрапа можно вычислить из набора значений бутстрапа различными способами. Самый простой — процентиль доверия бутстрапа, где конечные точки 95% доверительного интервала задаются 25-м и 975-м отсортированными значениями бутстрапа. Для этих данных этот интервал равен (3,05, 5,77).

Процентильный бутстрап, показанный здесь, является одним из самых простых методов доверительного интервала бутстрапа, но он может быть не лучшим методом во всех приложениях. В частности, процентильный интервал может не иметь заявленного покрытия. Покрытие доверительного интервала — это вероятность того, что доверительный интервал включает истинный параметр при повторной выборке из той же базовой популяции. Когда покрытие совпадает с заявленным размером доверительного интервала (например, покрытие = 95% для 95% доверительного интервала), интервалы являются точными. Эмпирические и теоретические исследования покрытия показали, что процентильный интервал является точным в некоторых ситуациях, но не в других.

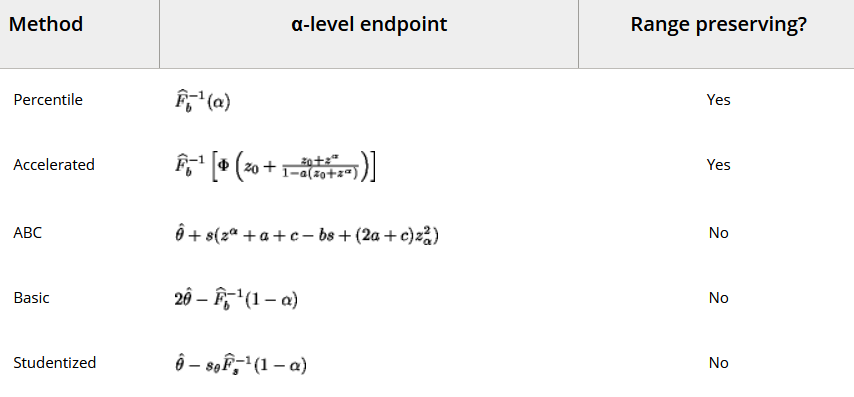
**Зверинец интервалов доверия Bootstrap**

Процентильный бутстрап был расширен многими способами для повышения точности доверия. Разновидности бутстрапов отличаются

1. как рассчитываются конечные точки доверительного интервала (например, процентильный, базовый, ускоренный, стьюдентизированный или скорректированный по смещению, ускоренный (BCA) бутстрап);
2. как аппроксимируется популяция (непараметрический или параметрический бутстрап);
3. как выбираются образцы бутстрапа (обычный, сбалансированный или бутстрап с подвижными блоками).

**Расчет конечных точек доверительного интервала**

Конечные точки процентильного бутстрапа легко вычислить, и они могут хорошо работать, особенно если распределение выборки симметрично. Доверительные интервалы процентильного бутстрапа могут не иметь правильного покрытия, если распределение выборки асимметрично. Другие методы корректируют конечные точки доверительного интервала, чтобы повысить точность покрытия (таблица 1). Одним из сбивающих с толку аспектов этих методов является то, что некоторые методы получили разные названия от разных авторов. Синонимия указана в документации к коллекции макросов SAS JACKBOOT.

Таблица 1. Методы оценки конечных точек доверительных интервалов бутстрапа α-уровня. Уравнение изображения — это наблюдаемая оценка, уравнение изображения — это бутстраповская функция распределения, уравнение изображения — стьюдентизированная бутстраповская функция распределения, Φ(x) — это нормальная функция распределения, уравнение изображения, a — это константа ускорения, а zα — это α-процентиль стандартного нормального распределения  
  
  


(Таблица – 1)

Охват процентильного бутстрапа можно улучшить, скорректировав конечные точки для смещения и непостоянной дисперсии (ускоренный бутстрап) . Вычисление доверительного интервала ускоренного бутстрапа требует оценки коэффициента смещения z0 и коэффициента ускорения a. Оба коэффициента можно оценить непараметрически из данных или теоретически рассчитать для конкретного распределения. Конечные точки доверительного интервала получаются путем инвертирования процентилей распределения бутстрапа. Корректировка смещения и ускорения изменяет процентили, используемые для нахождения конечных точек доверительного интервала. Поскольку конечные точки доверительного интервала получаются путем инвертирования распределения бутстрапа, как процентильный, так и ускоренный бутстрап сохраняют диапазон параметра. Например, если параметр и статистика ограничены диапазоном от 0 до 1, то конечные точки этих доверительных интервалов будут удовлетворять этому ограничению.

Квадратичные доверительные интервалы ABC являются приближением к ускоренному бутстрапу, которое не требует множества симуляций бутстрапа, что может быть полезно, когда оценка параметров требует значительных вычислений. Три требуемых коэффициента a, b и c (таблица 1) вычисляются либо из наблюдений, либо из модели. Конечные точки доверительного интервала вычисляются с помощью аппроксимации ряда Тейлора для Fb(x). Из-за аппроксимации конечные точки могут не удовлетворять ограничениям на пространстве параметров, в отличие от первых трех методов.

Базовый и стьюдентизированный бутстрапы вычисляют конечные точки путем инвертирования тестов гипотез. В обоих случаях верхний квантиль распределения бутстрапа используется для расчета нижней границы доверия, а нижний квантиль используется для расчета верхней границы. Когда распределение бутстрапа симметрично относительно оценки из исходных данных, т. е. изображения уравнения, базовый бутстрап выдает те же конечные точки, что и процентильный бутстрап. Когда распределение асимметрично, конечные точки двух методов различаются. Ни базовый, ни стьюдентизированный бутстрап не ограничивают конечные точки доверительного интервала попаданием в ограниченное пространство параметров.

Стьюдентизированный бутстрап основан на другом распределении бутстрапа, чем другие бутстрапы. Оценка, уравнение изображение и его SE, уравнение изображение, из каждой выборки бутстрапа используются для вычисления стьюдентизированных оценок уравнение изображение, где уравнение изображение — это оценка, рассчитанная из исходного набора данных. Квантили 1 − α/2 и α/2 из соответствующего распределения уравнение изображение используются для вычисления доверительного интервала. Конечные точки стьюдентизированного бутстрапового доверительного интервала имеют естественную интерпретацию. Они похожи на «обычные» доверительные интервалы, основанные на статистике Стьюдента t, за исключением того, что данные используются для оценки более подходящего распределения для статистики t. Распределение стьюдентизированного бутстрапа требует SE для каждой выборки бутстрапа. Использование повторной выборки методом складного ножа или второго, вложенного бутстрапа, может быть использовано, если SE невозможно оценить каким-либо другим способом.

Использование стьюдентизированного бутстрапа несколько спорно. Некоторым конечные точки интервалов кажутся слишком широкими, а метод, по-видимому, чувствителен к выбросам. Для других стьюдентизированный бутстрап, по-видимому, является единственным бутстрапом с разумным покрытием доверительного интервала в сложных задачах.

Доверительные интервалы бутстрапа могут не иметь заявленного покрытия при вычислении из небольших выборок. Детали, например, является ли эмпирическое покрытие слишком большим или слишком маленьким и лучше ли покрытие в одном хвосте, чем в другом, зависят от оцениваемой статистики и характеристик выборочной совокупности. Многочисленные исследования оценивали покрытие бутстрапа для конкретных случаев. Итерация бутстрапа обеспечивает способ улучшения покрытия доверительного интервала за счет дополнительных вычислений.

**Аппроксимация популяции**

В основе бутстрапа лежит концепция, что распределение интересующей статистики Fθ(x) может быть аппроксимировано оценками из повторных выборок из аппроксимации неизвестной популяции. Популяцию можно аппроксимировать разными способами, каждый из которых приводит к разному типу бутстрапа. Наиболее распространенные аппроксимации приводят к параметрическому и непараметрическому бутстрапу. Менее часто используемые аппроксимации приводят к сглаженному и обобщенному бутстрапу.

Параметрический бутстрап предполагает, что Fx(x) известно, за исключением, возможно, одного или нескольких неизвестных параметров ψ. Например, может быть известно (или предполагаться), что Fx(x) из логнормального распределения с неизвестными параметрами μ и σ2. Уравнение изображения аппроксимируется путем подстановки оценок уравнения изображения вместо неизвестных параметров ψ. Часто эти оценки являются оценками максимального правдоподобия, но могут использоваться и другие оценки. Обобщенный бутстрап — это параметрический бутстрап, где Fx(x) — это гибкое распределение со многими (часто четырьмя) параметрами, например, обобщенное лямбда-распределение.

Непараметрический бутстрап — это бутстрап, описанный ранее. Популяция Fx(x) аппроксимируется эмпирическим распределением наблюдаемых значений, по сути, мультиномиальным распределением. Непараметрические выборки бутстрапа включают повторы многих наблюдений, что может привести к несостоятельным оценкам. Сглаженный бутстрап аппроксимирует Fx(x) как сглаженную версию эмпирической CDF. Сглаженные выборки бутстрапа генерируются путем выборки наблюдений с заменой и дрожанием каждого наблюдения бутстрапа путем добавления небольшого количества случайного шума. Обычно распределение шума является нормальным со средним нулевым и небольшой дисперсией. Увеличение дисперсии увеличивает степень сглаживания. Когда выборочное пространство ограничено, немного другая процедура сглаживания может генерировать наблюдения бутстрапа, которые удовлетворяют ограничению.

**Выбор выборок бутстрапа**

В обычном непараметрическом бутстрапе, описанном выше, каждая выборка бутстрапа представляет собой простую случайную выборку с заменой наблюдений. Выборки бутстрапа представляют собой подмножество всех возможных выборок размера n из конечной популяции с n копиями каждого наблюдения. Следовательно, оценки бутстрапа смещения, SE и конечных точек доверительного интервала являются случайными величинами. Их дисперсию можно уменьшить, увеличив количество выборок бутстрапа или используя более сложные методы выборки.

Сбалансированный бутстрап — это альтернативный метод выборки, который может повысить точность смещения бутстрапа и SE. Сбалансированный бутстрап заставляет каждое наблюдение появляться в общей сложности nB раз в коллекции nB выборок бутстрапа. Это не заставляет каждую выборку бутстрапа содержать все наблюдения; первое наблюдение может появиться дважды в первой выборке бутстрапа и ни разу во второй, в то время как второе наблюдение может появиться один раз в каждой выборке. Сбалансированные бутстрап-выборки могут быть получены путем построения популяции с n копиями каждого из n наблюдений, затем случайным образом переставляя эту популяцию. Первые n переставленных значений являются первой бутстрап-выборкой, вторые n переставленных значений являются второй выборкой и т. д. Хотя балансировка часто уменьшает дисперсию оцененного смещения и SE, она, по-видимому, менее полезна для оценки конечных точек доверительного интервала.

Бутстрапы с подвижными блоками и подвижными плитками расширяют бутстрап на коррелированные данные. Обычный бутстрап предполагает, что наблюдения независимы, что может не подходить для данных временных рядов, пространственных данных или других коррелированных данных. В бутстрапе с подвижными блоками для данных временных рядов ряд наблюдений делится на b неперекрывающихся блоков из l последовательных наблюдений. Бутстрап-выборка создается путем случайной выборки b блоков с заменой и объединения их в ряд из bl наблюдений. Корреляция между наблюдениями предполагается самой сильной внутри блока и относительно слабой между блоками. Выбор l имеет решающее значение. Если l велико, то b мало, и может быть очень мало уникальных выборок бутстрепа. Если l мало, то наблюдения в разных блоках могут не быть независимыми. Даже если l выбрано правильно, корреляция между наблюдениями в выборке бутстрепа меньше, чем в исходной выборке, поскольку блоки считаются независимыми. Бутстреп пространственных данных с использованием движущихся плиток аналогичен. Стационарный бутстреп является вариантом бутстрепа с движущимися блоками со случайной длиной блоков. Подходы на основе моделей к бутстреппингу коррелированных данных описаны в следующем разделе.

**Расширения для не-iid данных**

Методы бутстрапа в предыдущих двух разделах подходят для одной выборки наблюдений iid. Многие проблемы связаны с наблюдениями, которые не являются iid. К ним относятся проблемы регрессии, временно или пространственно коррелированные данные и иерархические проблемы.

Регрессия и многовыборочные данные

Одним из распространенных источников не-iid наблюдений является ситуация, когда предполагается, что они исходят из некоторой линейной или нелинейной модели с аддитивными ошибками. Это включает в себя проблемы с двумя выборками, проблемы регрессии или более сложные модели для разработанных экспериментов. Интересующими величинами могут быть разница двух средних значений, наклон или отсекаемый элемент регрессии, некоторый параметр в модели или функция любого из них. Одним из приложений для измерения окружающей среды является использование бутстрапа для оценки RI50, концентрации токсичного вещества, которая снижает репродуктивный выход на 50%. Это функция параметров полиномиальной регрессии; бутстрап может быть использован для оценки доверительного интервала для RI50.

Существует два общих подхода для таких данных: бутстрапирование наблюдений, также называемое повторной выборкой случаев; и бутстрапирование остатков, также называемое повторной выборкой ошибок. Рассмотрим набор наблюдений, предположительно возникающих из линейной модели, Yi = Xiβ + ɛi. Каждое наблюдение представляет собой вектор значений ковариатов и отклик (Xi, Yi)T. Если наблюдения подвергаются бутстрапу, то весь вектор подвергается повторной выборке с заменой. Моменты и распределение значений ковариатов не фиксированы во всех выборках бутстрапа. Когда данные группируются, как в задаче с двумя выборками, принято обуславливать количество наблюдений в каждой группе. Бутстрапирование наблюдений требует отдельной повторной выборки каждой группы наблюдений.

Бутстрапирование остатков — это трехэтапный процесс. Изображение уравнения остатков вычисляется для каждого наблюдения. Затем формируется выборка бутстрапа изображения уравнения остатков с заменой из наблюдаемых остатков. Выборка бутстрапа наблюдений создается путем добавления случайно выбранного остатка к исходному прогнозируемому значению для каждого наблюдения: изображение уравнения.

Бутстрапирование наблюдений и бутстрапирование остатков не эквивалентны в малых выборках, но они асимптотически эквивалентны. Выбор бутстрапа зависит от цели и контекста анализа. Бутстрапирование остатков сохраняет структуру ковариатов, но вывод бутстрапа предполагает, что исходная модель (используемая для расчета остатков) является подходящей. Бутстрапирование наблюдений повторяет некоторые значения ковариатов и опускает другие. Это обычный выбор, когда анализ включает в себя какой-либо аспект выбора модели.  
  
  
**Коррелированные данные**

Дихотомия между бутстрапом наблюдений и бутстрапом остатков повторяется с временными рядами и пространственными данными. Бутстрапы с подвижными блоками и подвижными плитками, обсуждавшиеся выше, аналогичны бутстрапу наблюдений. Ни один из них не предполагает определенной модели для данных. Бутстрап остатков требует подгонки модели. Для данных временных рядов модель часто является моделью авторегрессионного скользящего среднего (ARMA), но это может быть модель пространства состояний. Для пространственных данных модель определяет среднюю и корреляционную структуру. Рассмотрим пространственные данные, которые, как предполагается, следуют модели



Один из подходов заключается в оценке изображения уравнения, матрицы дисперсии-ковариации ошибок ɛ, а затем вычислении изображения уравнения разложения Холецкого таким образом, что изображение уравнения ɛ = LL’. Изображение уравнения оцененных ошибок можно отбелить, умножив предварительно на L−1, т. е. изображение уравнения. Выборка бутстрепа пространственно коррелированных наблюдений строится путем составления выборки бутстрепа отбеленных остатков {eB}, введения структуры корреляции и восстановления среднего значения, т. е. изображение уравнения. Распределение интересующей статистики оценивается с помощью эмпирического распределения статистики во многих выборках бутстрепа.

**Иерархические данные**

Экологические данные часто включают несколько источников вариации, которые можно описать с помощью иерархического байесовского метода. Когда модель достаточно проста (например, линейная смешанная модель, в которой все случайные эффекты имеют нормальное распределение с постоянной дисперсией), данные можно выразить в виде 3 с матрицей дисперсии-ковариации Σ, которая зависит от компонентов дисперсии. Описанный выше бутстрап остатков на основе модели можно использовать для генерации бутстрап-выборок. Если, кроме того, указаны распределения всех случайных эффектов, то можно использовать параметрический бутстрап. Одним из примеров иерархического параметрического бутстрапа для сложной модели является построение доверительной области для эволюционной траектории.

Непараметрический бутстрап иерархических данных осложняется необходимостью оценки эмпирических функций распределения для двух (или более) случайных величин. Хотя процедуры могут быть выведены для конкретных случаев, в настоящее время не существует общего непараметрического метода для бутстреппинга иерархических данных.

**Теория бутстрапа**

Теория бутстрапа — это активная область статистических исследований. Я даю краткое введение, в теорию для одного параметра, оцененного по одной выборке наблюдений iid.

Асимптотические свойства могут быть выведены для многих различных типов статистик. Один из наиболее общих подходов рассматривает статистики, которые могут быть выражены как функция T эмпирического распределения n наблюдений Fn, т. е. статистики, которые могут быть записаны как Tn = T(Fn). Параметр равен θ = T(F), где F — функция распределения совокупности. При условии соответствующей дифференцируемости функции T и ограниченного второго момента для функции влияния T распределение бутстрапа FB(x) является состоятельной оценкой истинного распределения выборки Fθ(x). При условии дополнительного ограничения на хвосты распределения Tn оценка бутстрапа дисперсионного уравнения image является состоятельной оценкой дисперсии выборки параметра θ.

Точность покрытия, где покрытие — это вероятность того, что доверительный интервал включает θ, является важным свойством для процедуры доверительного интервала. Нижняя и верхняя границы рассматриваются отдельно, но их асимптотические свойства схожи. Методы бутстрепа доверительного интервала отличаются своими асимптотическими свойствами. Процентильные интервалы имеют точность первого порядка, т. е. изображение уравнения, где изображение уравнения — это предполагаемая нижняя граница двустороннего доверительного интервала 1 − 2α% [[13], стр. 187]. Как стьюдентизированные, так и BCa интервалы имеют точность второго порядка.

Другое сравнение процедур доверительного интервала — это связь между Fθ(x) и нормальным распределением. Если распределение выборки Fθ(x) является нормальным с известной дисперсией, то доверительные интервалы, основанные на z-оценках, имеют желаемое покрытие, и бутстреп не нужен. Пределы процентильного бутстрепа верны, если Fθ(x) можно преобразовать в нормальное распределение. Другими словами, существует некоторая монотонная g(x) такая, что уравнение image, где ϕ = g(θ), а τ2 — постоянная дисперсия. Другие процедуры доверительного интервала бутстрапа верны при более общих моделях распределения уравнения image. Например, интервалы BCa верны, если уравнение image, где τϕ = 1 + aϕ, z0 — коэффициент коррекции смещения, а a — коэффициент ускорения.

**Вычисление**

Бутстрап может быть реализован везде, где есть возможность генерировать однородные случайные числа и составлять случайную выборку наблюдений. Макросы и функции в различных статистических пакетах включают более сложные вычисления доверительных интервалов. К ним относятся макрос JACKBOOT в SAS и различные библиотеки функций S‐PLUS. Все макросы и библиотеки могут бутстрапить одну выборку наблюдений и вычислять смещение, SE и различные доверительные интервалы. Некоторые пакеты можно легко расширить для задач с несколькими выборками.

Пример

Этот расширенный пример проиллюстрирует множество различных типов доверительных интервалов бутстрапа и связь между повторной выборкой складного ножа и бутстрапом. Данные являются частью исследования нагрузки тяжелыми металлами у детей, целью которого является описание связи между концентрациями креатинина и магния в моче. Концентрации креатинина и магния были измерены у 38 детей из относительно загрязненной области (Капфенберг) и 52 детей из менее загрязненной области (Книттельфельд) Штирии, Австрия. Анализ складного ножа в этой записи учитывает четыре статистики: ρ — корреляция между креатинином и магнием, β1 и β2 — наклоны регрессии магния на креатинин для каждой области и β1/β2 — их отношение. Наклоны β1 и β2 определяются гетероскедастической линейной регрессией с различными параметрами для каждой группы детей.

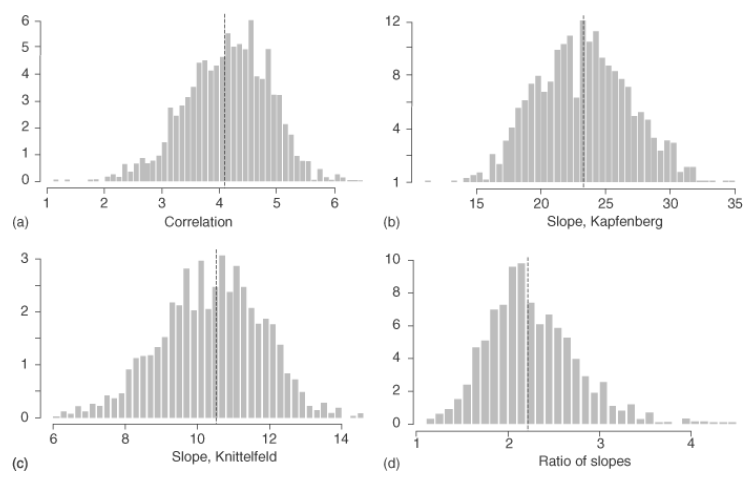






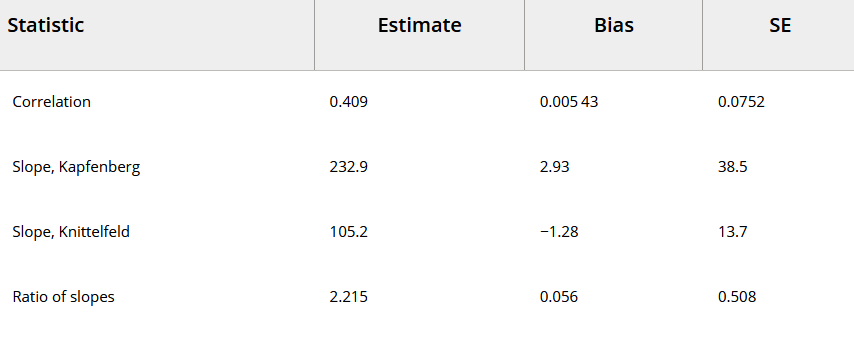
где Mi и Ci — концентрации магния и креатинина в крови. Модель может быть подогнана с использованием функции glm() в S-PLUS. Здесь анализ повторяется с использованием бутстрепа. Для вычислений использовалась библиотека функций boot()в S-PLUS.

Выборочная корреляция между креатинином и магнием, рассматривающая всех детей как одну выборку из 90 наблюдений, составляет 0,409. Распределение бутстрепа изображения уравнения коэффициента корреляции (рисунок 1a) оценивается по 1000 выборкам бутстрепа, каждая из которых содержит 90 наблюдений. Это распределение очень слегка скошено. Оцененные смещение и SE (таблица 2) аналогичны тем, которые вычисляются с использованием различных форм складного ножа (сравните с таблицей 1 в записи о повторной выборке складного ножа). Девяносто пять процентов доверительных интервалов для коэффициента корреляции были построены с использованием четырех методов бутстрепа (таблица 3). Интервалы стьюдентизированного бутстрапа не были рассчитаны для ρ, поскольку оценка дисперсии методом складного ножа была очень ресурсоемкой. Конечные точки довольно похожи друг на друга, хотя конечные точки из метода BCa немного отличаются от других. Я бы выбрал интервал BCa, поскольку он делает самые общие предположения, имеет лучшие асимптотические свойства, а набор данных достаточно большой, чтобы обеспечить разумную оценку “a”, константы ускорения.



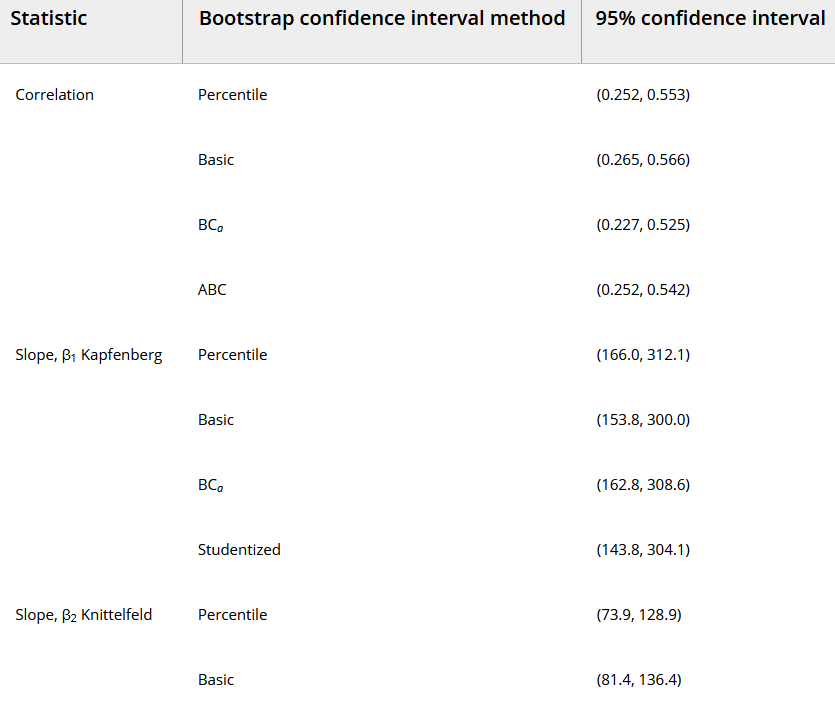
Распределения бутстрепа (a) корреляции между концентрациями креатинина в крови и магния; (b) наклона линейной регрессии магния на концентрации креатинина в крови для 38 детей из относительно загрязненной области (Капфенберг); (c) наклона для 52 детей из менее загрязненной области (Книттельфельд); и (d) отношения двух наклонов. Наблюдаемое значение отмечено пунктирной вертикальной линией на всех четырех панелях.

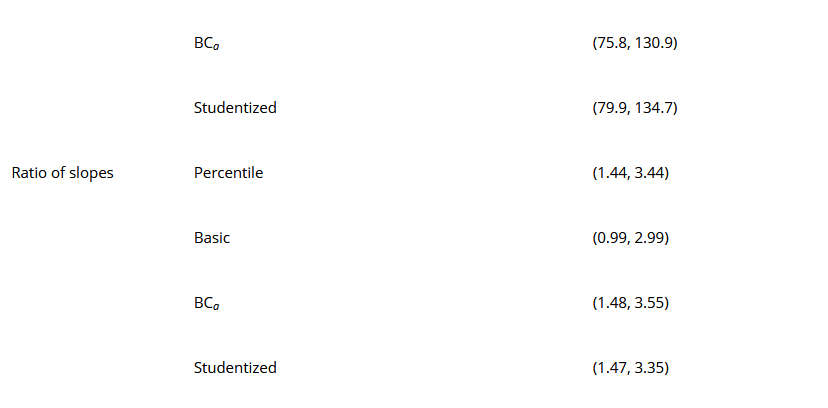
Таблица 2. Оценки параметров, бутстреп-оценка смещения и бутстреп-оценка SE для четырех величин, описывающих связь между концентрацией креатинина в моче и магния



(Таблица – 2)

Таблица 3. Конечные точки 95% доверительных интервалов для четырех величин, описывающих связь между концентрацией креатинина в моче и магния. Доверительные интервалы вычисляются с использованием методов процентиля, базового и BCa. Студентизированный бутстрап включается, когда его можно легко рассчитать





(Таблица – 3)

Существует много способов бутстрапа в задаче регрессии. Соответствующий выбор зависит от того, как были собраны данные. Я предположил, что количество детей в каждой области было зафиксировано в плане, но распределение значений X (уровень креатинина в крови) — нет. Следовательно, уместно бутстрапировать наблюдения (не остатки) и указать слои, чтобы заставить каждую бутстрап-выборку включить 38 детей из Капфенберга и 52 из Книттельфельда.

Бутстрап-распределения β1 (рисунок 1b) и β2 (рисунок 1c) достаточно симметричны. Опять же, бутстрап-оценки смещения и SE (таблица 2) довольно близки к оценкам удаления 1 складного ножа (сравните с таблицей 3 в записи о повторной выборке складного ножа). Бутстрап-SE немного (примерно на 5%) меньше, чем складные ножевые SE, возможно, потому, что бутстрап-выборки вынуждены иметь 38 и 52 потомка из двух областей. Конечные точки четырех процедур доверительных интервалов довольно похожи (таблица 3). Хотя стьюдентизированные интервалы для β1 немного шире, чем три других интервала для β1, стьюдентизированные интервалы для β2 немного короче, чем три других интервала для β2. Я бы выбрал стьюдентизированные интервалы здесь, но между любыми интервалами мало практической разницы.

Бутстрап-распределение отношения β1/β2 скошено (рисунок 1d), поэтому можно было бы ожидать найти различия между процедурами доверительных интервалов. Как нижняя, так и верхняя конечные точки для базового интервала намного меньше, чем для других трех интервалов. Конечные точки BCa и стьюдентизированных интервалов почти идентичны. Я бы выбрал любой из этих интервалов.

Хотя может показаться, что бутстрап творит чудеса, в том смысле, что он позволяет делать статистические выводы в очень общих обстоятельствах, он не заменяет качественные данные. Эффективность бутстрапа зависит от размера выборки. Невозможно рекомендовать минимальные размеры выборки, поскольку каждая проблема индивидуальна. Однако увеличение числа повторов бутстрапа или использование более сложной процедуры бутстрапа не компенсирует недостаточность данных. Все, что может сделать бутстрап, — это (приблизительно) количественно оценить неопределенность в заключении.