ИНСТИТУТ ТРАНСПОРТА И СВЯЗИ

Елена АФАНАСЬЕВА

ОЦЕНИВАНИЕ МОДЕЛЕЙ ЛОГИСТИЧЕСКИХ ПРОЦЕССОВ ТРАНСПОРТА НА БАЗЕ ИНТЕНСИВНЫХ КОМПЬЮТЕРНЫХ МЕТОДОВ СТАТИСТИКИ

АВТОРЕФЕРАТ ПРОМОЦИОННОЙ РАБОТЫ

РИГА – 2006

ИНСТИТУТ ТРАНСПОРТА И СВЯЗИ

Елена АФАНАСЬЕВА

ОЦЕНИВАНИЕ МОДЕЛЕЙ ЛОГИСТИЧЕСКИХ ПРОЦЕССОВ ТРАНСПОРТА НА БАЗЕ ИНТЕНСИВНЫХ КОМПЬЮТЕРНЫХ МЕТОДОВ СТАТИСТИКИ

АВТОРЕФЕРАТ ПРОМОЦИОННОЙ РАБОТЫ
представленной на соискание степени доктора инженерных наук

Научная область “Транспорт”
подобласть “Телематика и логистика”

Научный руководитель:
Dr.habil.sc.ing., профессор
Александр Андронов

РИГА - 2006

UDK 656:004
A 941

Институт Транспорта и связи

Афанасьева Е.

A 941 Оценивание моделей логистических процессов транспорта на базе интенсивных компьютерных методов статистики. Автореферат промоционной работы. – Рига: Институт транспорта и связи, 2006. – 52 с.

ISBN 9984-9865-1-9

ПРОМОЦИОННАЯ РАБОТА

ПРЕДСТАВЛЕНА В ИНСТИТУТ ТРАНСПОРТА И СВЯЗИ НА СОИСКАНИЕ УЧЁНОЙ СТЕПЕНИ ДОКТОРА ИНЖЕНЕРНЫХ НАУК

Защита промоционной работы состоится 4 июля 2006 года в 15:00 в промоционном совете Института транспорта и связи по адресу Латвия, г. Рига, ул. Ломоносова 1 в аудитории 4-710, тел. (+371) 7100661, факс. (+371) 7100660.

ОФИЦИАЛЬНЫЕ ОПОНЕНТЫ:

Хаб. док. инж. наук, профессор Евгений Александрович Копытов
Институт транспорта и связи

Хаб. док. инж. наук, профессор Юрий Михайлович Парамонов
Рижский Технический Университет

Хаб. док. инж. наук, профессор Вячеслав Борисович Мелас
Санкт-Перербуржский Государственный Университет

ПОДТВЕРЖДЕНИЕ

Я подтверждаю, что выполнила данную промоционную работу, которая подана на рассмотрение в Институт Транспорта и связи на соискание научной степени доктора инженерных наук. Данная промоционнай работа не была представлена на рассмотрение не в один другой университет или институт для получения учёной степени.

15 мая 2006 года Елена Афанасьева

Промоционная работа написана на английском языке, состоит из шести глав, включает 19 рисунков, 140 формул, 26 таблиц, одно приложение, составляя в общем 135 страниц. Библиография включает 91 литературный источник.

АННОТАЦИЯ

Промоционная работа “Оценивание моделей логистических процессов транспорта на базе интенсивных компьютерных методов статистики” разработана Еленой Николаевной Афанасьвой на получение учёной степени «доктора инженерных наук в области телематики и логистики». Научный руководитель работы Dr.habil.sc.ing., профессор Александр Михайлович Андронов.

Работа посвящена применению современных статистических методов для оценивания логистических моделей транспорта. Основное внимание уделено использованию интенсивного компьютерного метода ресамплинг (resampling). Этот метод является непараметрическим и даёт наиболее эффективные оценки показателей систем в случае малых объёмов исходных выборок. Исследование велось в трёх основных направлениях: прогнозирование и оценивание транспортных моделей, оценивание надёжности и эффективности транспортных средств, а также управление запасами в логистических системах.

Для каждой из поставленных задач в рамках соответствующей математической модели были предложены алгоритмы применения ресамплинг-процедуры, выведены формулы для сравнения эффективности ресамплинга и традиционного подходов. В качестве критерия эффективности оценок использовались смещение, дисперсия или среднеквадратическая ошибка оценок. Получены численные результаты, доказывающие эффективность предлагаемого подхода. Сделаны выводы и даны рекомендации относительно условий, при которых предлагаемый подход является наиболее эффективным. Полученные результаты являются общими, поскольку могут быть применимы также к другим предметным областям.

СОДЕРЖАНИЕ

1. Актуальность работы

Транспорт Латвии – это быстро развивающаяся отрасль, признанная одной из наиболее приоритетных. Данная отрасль требует обширных и дорогих научных исследований, с привлечением информационных технологий, математического аппарата, методов экономики и логистики.

В данной промоционной работе наибольшее внимание уделяется применению современных статистических методов в планировании и организации логистических процессов транспорта. Исследование идёт по трём основным направлениям: прогнозирование и оценивание транспортных потоков; оценивание надёжности и эффективного использования транспортных средств; управление запасами в логистических системах.

Решая практические задачи с применением вероятностных моделей, мы часто сталкиваемся с такой проблемой, как нехватка статистических данных, на основе которых необходимо оценивать неизвестные параметры вероятностных моделей. В такой ситуации использование традиционных параметрических методов оценивания весьма затруднено. Однако интенсивные компьютерные методы позволяют решать эту проблему. Эти методы вычислительной статистики появились сравнительно недавно. Они предполагают многократное использование одних и тех же данных в различных комбинациях, что обеспечивает более полное использование статистической информации. Важной особенностью этого подхода является то, что он является непараметрическим. Непараметричность позволяет избежать ошибок, свойственных традиционным методам и связанных с проверкой гипотез о виде распределения случайных величин. Основное внимание в работе уделено интенсивному компьютерному методу ресамплинг (resampling). Сравнительная схема традиционного плаг-ин (plug-in) и ресамплинг подходов, применительно к имитационному моделированию представлена на рис. 1. Здесь проиллюстрирована основная идея ресамплинга, заключающаяся в том, что случайные величины (СВ) не генерируются специальными генераторами, а берутся непосредственно из исходных выборок. Ресамплинг позволяет оценивать вероятностные характеристики сложных систем, основываясь на сравнительно малых объёмах статистических данных.

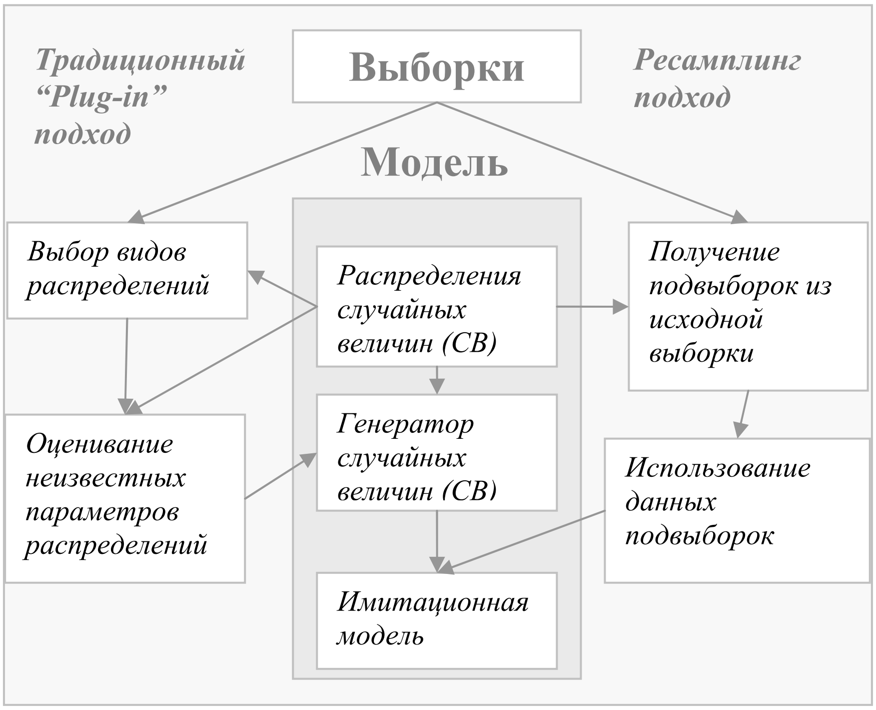

Рис. 1 Сравнительная схема традиционного «plug-in» и ресамплинг подходов

Поскольку ресамплинг метод сравнительно новый, то возникает актуальная проблема исследования эффективности данного подхода для анализа различных математических моделей и решения практических задач.

2. Цель и задачи исследования

Основной целью исследования является разработка алгоритмов использования ресамплинг-метода для анализа различных математических моделей, оценивание эффективности такого подхода, а также применение его для решения практических задач оценивания моделей логистических транспортных процессов.

Основные задачи работы следующие:

- Изучить интенсивные компьютерные методы и их сферы применения.

- Разработать алгоритмы применения ресамплинг подхода для оценки параметров регрессионных моделей.
- Исследовать эффективность ресамплинг метода для построения регрессионных моделей.
- Рассмотреть возможности применения ресамплинг-подхода для прогнозирования и оценивания транспортных потоков на базе регрессионных моделей.
- Предложить алгоритмы использования ресамплинг подхода для оценки параметров процессов массового обслуживания.
- Разработать алгоритмы для оценки эффективности применения рассматриваемого метода в статистических задачах массового обслуживания.
- Рассмотреть возможность применения ресамплинг-подхода для оценки надёжности и эффективности транспортных средств на основе теории массового обслуживания.
- Предложить методику применения ресамплинг подхода для задачи сравнения двух процессов восстановления.
- Оценить эффективность предлагаемого метода для задачи сравнения двух процессов восстановления.
- Применить ресамплинг-метод для управления запасами на основе модели сравнения двух процессов восстановления.
- Разработать комплекс программ для применения и оценки эффективности предлагаемого подхода.
- Используя методологию ресамплинг, провести оценивание различных моделей из области транспорта и применить алгоритмы для оценки эффективности подхода.

3. Степень исследования темы

Интенсивные компьютерные методы или методы вычислительной статистики стали стремительно развиваться с появлением мощной компьютерной техники. Они включают следующие разновидности: метод перекрёстной проверки (cross-validation), метод складного ножа (jackknife), бутстреп (bootstrap) и ресамплинг, которые позволяют решать широкий набор задач. Метод складного ножа предложил М. Кенуй (Quenouille) в 1949 году как оценку, которая является комбинацией оценки, основанной на всех данных и оценок, основанных на части данных. В 1979 году Б. Эфрон (Efron) предложил бутстреп метод, который фактически явился обобщением метода складного ножа. В 1976 году В. Ивницкий предложил использовать ресамплинг для задач оценки надёжности систем методом моделирования. С. Ву (Wu) в 1979 году рассмотрел применение ресамплинг-подхода для регрессионных моделей. Ресамплинг и бутстреп подходы с 1995 года также исследуется под руководством профессора А. Андронова. Были рассмотрены простой и иерархический ресамплинг и применение его в теории надёжности, теории массового обслуживания, задачах оптимизации. При участии Ю. Меркурьева и М. Фиошина рассматривались также свойства ресамплинг для сумм, частично известных распределений и построения доверительных интервалов. При участии Е. Афанасьевой были проведены исследования в области применения ресамплинг метода для оценивания распределений порядковых статистик [4], для оценивания параметров регрессионных моделей [6], [5], теории восстановления, теории массового обслуживания [2], [9]. Некоторые из рассмотренных задач описаны в данной промоционной работе. Е. Афанасьева рассмотрела применение ресамплинг метода для задач сравнения процессов восстановления [2] и применения этого подхода для решения задач теории управления запасами [8]. Данные исследования тоже вошли как глава в промоционную работу. Сейчас работы в этой области продолжаются.

4. Методология и методы исследования

За теоретическую и методическую основу промоционной работы взяты классические работы в области математической и вычислительной статистики, транспорта, логистики, моделирования и компьютерных наук. Учтены последние тенденции развития науки в этих областях.

В работе использовались книги в отмеченных областях, тематические материалы периодических изданий, материалы международных конференций, статистические сборники.

Все выводы базируются на применении классических аппаратов теории случайных процессов, теории вероятностей и математической статистики.

В процессе исследования были проанализированы различные примеры из области транспортной логистики, для которых было проиллюстрировано применение предлагаемой методики. В примерах использовались как реальные, так гипотетические данные, которые в полной мере отражают специфику решаемых задач и эффективность предлагаемого подхода. В качестве критериев эффективности используемого метода были взяты: смещение, дисперсия и среднеквадратическая ошибка оценок. Были проанализированы тенденции изменения эффективности в зависимости от различных факторов.

Для решения поставленных задач в работе применялись как аналитические, так и экспериментальные методы. С помощью аналитических методов получены выражения для анализа эффективности метода в различных ситуациях. С помощью экспериментальных методов подсчитаны значения критериев эффективности применяемой методики для конкретных численных примеров, которые позволяют сделать выводы о влиянии различных факторов на результаты.

Результаты применения предлагаемой методики сравниваются с результатами применения классических подходов. Делаются выводы о необходимых условиях, при которых ресамплинг-подход оказывается эффективнее.

5. Научная новизна работы

Можно говорить о двух аспектах научной новизны работы. Во-первых, в математическом плане. Это применение ресамплинг-метода в различных стстистических задачах и анализ эффективности этого применения. Во-вторых, в прикладном плане. До сих пор применение ресамплинг-метода к транспортным системам в данной постановке не анализировалось.

В данной промоционной работе проанализировано применение современных статистических методов в планировании и организации логистических процессов транспорта в трёх основных направлениях: прогнозирование и оценивание транспортных потоков; оценивание надёжности и эффективности транспортных средств, управление запасами в логистических системах.

Математической основой для решения первой задачи, связанной с прогнозированием транспортных потоков, явилась регрессионная модель. В рамках данной части исследования ресамплинг-подход применялся для оценивания параметров регрессионной модели. Алгоритмы применения и алгоритмы расчёта эффективности являются новыми научными результатами.

Математической моделью для второй части исследования, связанной с оцениванием надёжности и эффективности транспортных средств, является бесконечно линейная система массового обслуживания. Было рассмотрено применение ресамплинг-подхода в рассмотренной ситуации, выведены формулы для анализа эффективности. Это является научной новизной второй части исследования.

Наконец, в рамках последней части исследования ресамплинг подход был применён для решения задач теории управления запасами. Математическим аппаратом анализа явились процессы восстановления. Решалась задача оценки вероятности дефицита в конкретной задаче управления запасами, которая могла быть математически описана, как задача сравнения двух процессов восстановления. Помимо алгоритмов применения ресамплинг-подхода в данной ситуации, были также получены выражения для оценки эффективности данного подхода, составляющие новые научные результаты последней части.

6. Практическая ценность, реализация и применение работы

Практическими результатами являются применения предлагаемых подходов к решению задач транспортной логистики. Первая рассматриваемая задача – это прогнозирование и оценивание транспортных процессов как моделей некоторых макроэкономических показателей. В рамках этой главы иллюстрируется решение таких задач как:

- Прогнозирование пассажирского потока на авиационном транспорте в странах Европейского союза на определённый год, в зависимости от различных факторов: территория страны, население, средняя заработная плата, валовой внутренний продукт на душу населения.
- Прогнозирование спроса на транспортную услугу в различных регионах, в зависимости от уровня урбанизации, уровня образования и уровня зарплаты.

Второй рассматриваемый класс задач связан с оцениванием надёжности и эффективности транспортных средств. В рамках данной главы приводится пример решения следующей практической задачи:

- Имеется поток повреждений на авиационной конструкции. Имеется статистика возникновения повреждений до развития опасной ситуации. Необходимо оценить вероятность того, что в рассматриваемом интервале времени не произойдёт отказа.

Третья сфера применения данной методики – это теория управления запасами. Подход иллюстрируется на примере следующей задачи:

- Рассматривается процесс эксплуатации парка некоторых изделий. Известен начальный запас этих изделий, статистика по интервалам между их поставками, а также статистика по интервалам между поставкой и выходом из строя каждого изделия. Необходимо оценить вероятность того, что в момент спроса на данное изделие дефицита не возникнет, а также найти оптимальный уровень запаса при заданных параметрах прибыли от использования, издержках на хранение и потерях от дефицита.

Полученные численные результаты позволяют судить об эффективности рассматриваемого похода для решения описанных задач. Анализ результатов позволяет сделать вывод о целесообразности применения ресамплинг-подхода в аналогичных задачах транспортной логистики. Данный подход является общим, поэтому он применим для решения практических задач в других предметных областях.

7. Публикации

Результаты работы были рассмотрены в 10 публикациях в научных журналах, сборниках международных конференций и сборниках научных трудов. Рассматриваемые в публикациях проблемы были освещены на международных научных конференциях в Латвии, Польше, Израиле и Франции.

8. Структура работы

В первой главе работы рассматривается применение статистических методов в планировании и организации логистических процессов транспорта. Во второй главе рассмотрены современные интенсивные компьютерные методы (ИКМ) статистики. В третьей главе сделан обзор применений ресамплинг-метода для различных математических моделей. Каждая из последующих глав рассматривает применение ресамплинг-подхода для решения конкретной задачи. Это – задачи прогнозирования транспортных процессов на основе регрессионных моделей, оценивания надёжности и эффективности транспортных средств, управления запасами. Каждая глава имеет содержательную постановку задачи, математическую модель, анализ традиционных подходов к решению данной задачи и мотивацию предлагаемого подхода. Далее излагается содержание предлагаемого подхода с соответствующими алгоритмами и выражениями для подсчёта эффективности. Каждая глава заканчивается численным анализом эффективности подхода на основе гипотетических данных, которые в полной мере отражают специфику и особенности рассматриваемой задачи. Далее следует пример применения из области транспорта, содержащий постановку задачи, алгоритм применения предлагаемого подхода и численные результаты. В конце каждой главы сделаны выводы относительно целесообразности применения предлагаемого подхода в рассмотренной ситуации. Произведено сравнение с результатами применения традиционных методов, даны рекомендации, при каких условиях предлагаемый подход даёт лучшие результаты. Работа заканчивается заключением, содержащим общие выводы по работе, и списком использованной литературы.

9. Обзор основных результатов исследования

9.1 Статистические методы в планировании и организации логистических процессов транспорта

Необходимость применения математических методов и информационных технологий на транспорте вызвана увеличением объёма перевозок, всевозрастающей сложностью технологических процессов, непрерывно усиливающимися связями между различными звеньями транспортной системы. Необходимо создать такой математический аппарат, который позволит анализировать эффективность работы транспортных средств, моделировать процесс перевозок, прогнозировать спрос на перевозки, разрабатывать оптимальные варианты развития транспортной системы. В свою очередь, логистика является одним из основных направлений научного и технологического развития транспортного сектора Европейского союза. Это развивающаяся сфера экономики и новое научное направление.

В данной работе предлагается применять современные статистические методы для решения некоторых задач планирования и организации логистических процессов транспорта. Актуальность вызвана тем, что управлять сложными экономическими системами нельзя, не располагая достоверной статистической информацией об объектах исследования.

С помощью анализа соответствующих моделей на транспорте могут решаться следующие задачи: планирование наиболее эффективных маршрутов для лучшей организации транспортных и пассажирских потоков; анализ статистики о дорожных происшествиях для выяснения факторов, влияющих на происшествия; моделирование планов поддержания надлежащего состояния дорожного покрытия; прогнозирование возможного ремонта дорог; прогнозирование спроса на услуги различных видов транспорта; формирование сети автодорог или авиалиний; планирование резервов парка транспортных средств. Ошибки здесь приводят к неэффективному использованию производственных фондов и к неполному удовлетворению спроса.

Зачастую получение данных является сложной задачей. Описанная ситуация весьма характерна для задач надёжности некоторых видов транспортной техники. В связи с высокой надёжностью, например в авиационной технике, имеется очень маленькая статистика отказов.

В данной работе применяются следующие математические модели и статистические методы: регрессионные модели, методы оценивания неизвестных параметров, проверка статистических гипотез, теория восстановления, теория управления запасами, теория случайных процессов, теория массового обслуживания. Применяются ресамплинг-методы оценивания неизвестных характеристик систем. Эти методы сравниваются с классическими аналогами.

9.2 Современные интенсивные компьютерные методы статистики

Развитие классических (традиционных) статистических методов обычно сдерживалось ограничением на объём вычислений, которые ранее производились вручную. Поэтому основные классические результаты были связаны с асимптотическим подходом, для больших размеров выборки. Для маленьких объёмов выборки можно было использовать и обычный перебор. Оставался открытым вопрос как действовать, когда объём выборки средний. Коренное изменение пришло с появлением компьютеров. Это привело к развитию методов вычислительной статистики или ИКМ.

Традиционные статистические методы и традиционные статистические модели часто основываются на большом количестве допущений (линейные зависимости, классы распределений, стационарность и др.). С одной стороны, такие допущения упрощают модель и позволяют использовать аналитические методы. С другой стороны, на практике, такие допущения ведут к неполным и некорректным моделям. Это делает результаты, полученные с применением таких методов, неточными.

Трудным является также применение традиционных методов для анализа сложных систем, нестационарных систем, в случае, когда классы распределений отличаются от классических. Конструирование аналитических моделей для таких систем сложно и зачастую невозможно. В таких случаях лучше использовать ИКМ статистики.

Отметим также, что интенсивные компьютерные методы не такие точные, как классические. Но это справедливо только в случае, когда выполняются допущения классических методов. В случае невыполнения соответствующих допущений, интенсивные компьютерные методы являются гораздо более точным инструментом. Таким образом, можно говорить о том, что ИКМ могут решать многие проблемы, которые ранее не могли быть решены приемлемым образом.

Мы должны помнить, что существует также потенциальный риск использования ИКМ. Нужно понимать, что большое число вычислений не гарантирует, что информация используется корректно и, что результаты будут лучше. Во многих случаях применение классического метода для таких задач даёт более точные результаты, чем результаты ИКМ, полученные после целого дня вычислений.

Область вычислительной статистики в настоящее время включает большое количество методов. Эти методы позволяют рассматривать данные с различных перспектив, анализировать различные подмножества данных. Из-за большого числа различных комбинаций входных данных может потребоваться большое количество вычислений.

Методы Монте-Карло (Monte Carlo) - первые методы, требующие большого объёма вычислений, появились ещё в 60-ые годы. Эти методы представляют собой рандомизацию условий испытаний при принципиально неполном переборе, причём сами испытания могут быть не только физическими экспериментами, но и расчётами на ЭВМ. Методы Монте-Карло используются, например, для проверки определенной статистической гипотезы, определяя уровень значимости некоторый статистики, рассчитанной для набора данных, а также для оценивания неизвестных параметров систем.

К методам рандомизации относятся метод *перекрёстной проверки* (cross-validation) и *метод складного ножа* (jackknife). Перестановки элементов, получение подвыборок или другие манипуляции с выборкой не могут дать нам дополнительной информации. Однако, изменяя размещение в массиве данных мы можем получить информацию о том, насколько хорошо данные отвечают выдвинутой гипотезе. Это является основной идеей рандомизации. *Метод перекрёстной проверки* – это процедура, в которой мы делим выборку на два множества: «оцениваемое множество», множество по которому строится модель и «проверочное множество», множество, на котором затем проверяется модель. Оцениваемое множество используется для оценки интересующих нас параметров, а проверочное множество используется для оценки качества построенной модели. *Метод складного ножа* в 1949 году предложил М. Кенуй (M. Quenouille). Идея заключалась в том, чтобы последовательно исключать из рассмотрения по одному наблюдению, обрабатывать всю оставшуюся информацию и предсказывать результат в исключенной точке. Совокупность расхождений, полученных таким образом по всем точкам, несет в себе информацию о смещении, которой можно воспользоваться. Дж. Тьюки (Tukey), принимавший активное участие в совершенствовании этого метода, назвал его методом складного ножа.

Бутстреп (bootstrap) первоначально был предложен Б. Эфроном в 1977 году, после чего был обобщён и другими авторами. Суть метода заключается в непосредственном использовании имеющихся статистических данных в процессе моделирования. Предполагается многократная обработка различных частей одних и тех же данных. Бутстреп – непараметрический метод, который первоначально был разработан для того, чтобы бороться со смещением, обусловленным малым размером выборки. Для решения этой проблемы нужно выбирать такие модели, для которых результаты слабо зависят от действительной ситуации. Это прием непараметрической статистики, который привёл к развитию таких подходов, как *непараметрические и робастные методы*. Затем выяснилось, что его можно применять для оценивания выборочной дисперсии, построения доверительных интервалов и проверки гипотез.

Метод *Трейс-Дривен(Trace-Driven) моделирование* применяется для валидации построенной модели по отношению к реальной. *Трейс-Дривен моделирование* обсуждается в работах Дж. Кляйнена (Kleijnen) с 1998 года. Этим же автором было рассмотрено применение данного метода для валидации транспортной модели.

9.3 Ресамплинг-метод в статистике и его развитие

Как уже отмечалось ранее, трудно провести грань между различными ИКМ. Во многих литературных источниках все подобные методы просто называются методами ресамплинг.

Ресамплинг является ИКМ, который применяется в статистике и моделировании. Это альтернативный метод для имитационного моделирования, который может быть эффективно использован в случае малых выборок, а также в случае, когда для различных переменных используется одна общая выборка. Этот подход к моделированию систем был рассмотрен ранее в работах А. Андронова. Там отмечалось, что ресамплинг метод позволяет избежать смещения и уменьшить дисперсию оценок. Основная идея применения ресамплинг-методов заключается в том, что в процессе имитационного моделирования значения СВ не вырабатываются генераторами случайных чисел, а извлекаются из соответствующих выборочных совокупностей. При этом может оказаться, что объёмы исходных выборочных совокупностей неодинаковы и слишком малы, чтобы осуществить достаточно представительное моделирование. Эта трудность преодолевается путём применения ресамплинг-метода, при котором одни и те же выборочные данные для одной случайной величины многократно используются в различных комбинациях с данными для других случайных величин.

Итак, рассматривается известная функция ϕ многих независимых СВ $X_1, X_2, ..., X_m$: $\phi(X_1, X_2, ..., X_m)$. Предполагается, что функция распределения $F_i(.)$ СВ X_i неизвестна, но для каждой случайной величины предполагается известной выборочная совокупность:

$$H_i = \{X_{i1}, X_{i2}, \ldots, X_{in_i}\},\ |H_i| = n_i,\ i=1,..,m. \qquad (1)$$

Нашей задачей является найти оценку математического ожидания:

$$\Theta = E(\phi(X_1, X_2, \ldots, X_m)). \qquad (2)$$

При этом подходе значения аргументов X_i функции ϕ случайно извлекаются из соответствующих выборок $\{H_i\}$. Другими словами, $j(l) = (j_1(l), \ldots, j_m(l))$, $l = 1,2,\ldots$, являются случайными выборками из $\{H_i\}$. Обычно говорят, что $j(l)$ и $X(l) = (X_{1j_2(l)}, X_{2j_2(l)}, \ldots, X_{mj_m(l)})$ - это l-ая ресамплинг выборка. Тогда ресамплинг-оценка Θ примет вид:

$$\Theta^* = \frac{1}{r}\sum_{l=1}^{r}\phi(X(l)). \qquad (3)$$

Обычно число повторений ресамплинг-процедуры r значительно меньше, чем число всех возможных комбинаций исходных данных.

Пример использования ресамплинг метода показан на рис. 2. Здесь видно, что вектор $j(l)$ содержит индексы тех элементов выборки, которые были выбраны на l-ой реализации. Так для l=1, j(1)=(2,1,4,2), что означает, что из первой выборки H_1 взяли второй элемент, из второй выборки H_2 первый элемент, из третьей H_3 четвёртый элемент и из четвёртой H_4 второй элемент.

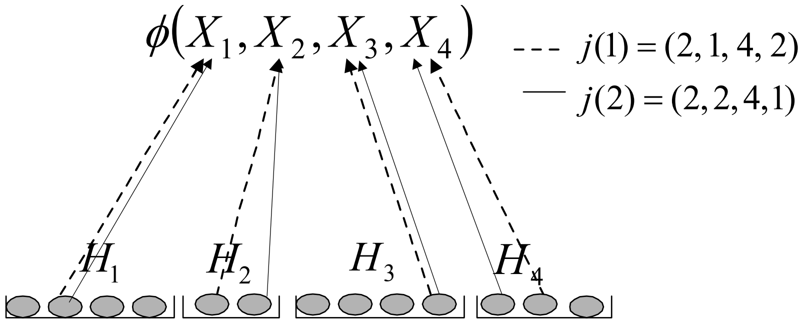

Рис. 2 Пример процедуры ресамплинг

Выбор элементов в ресамплинг-выборку на некоторой реализации является консервативным. Это означает, что у всех выборок $X(l)$ одно и то же маргинальное распределение, которое можно получить из входного распределения $\{F_i\}$, и у выборок $(X(l), X(l'))$ одинаковое совместное распределение при $l \neq l'$.Очевидно, что для такого консервативного плана искомая оценка является несмещённой: $E(\Theta^*) = \Theta.$

Консервативный план предполагает наличие различных видов ресамплинга: простой ресамплинг, управляемый и иерархический ресамплинг, ресамплинг стохастических процессов, ресамплинг для оценивания параметров регрессионных моделей. Все эти разновидности метода, не используют полный перебор всех возможных комбинаций исходных данных.

В качестве критериев эффективности ресамплинг-оценки используют: смещение, дисперсию или среднеквадратическую ошибку.

Введём следующие обозначения для моментов:

$$\mu_\nu = E(\phi(X(l))^\nu), \quad \nu = 1,2,\ldots \tag{4}$$

$$\mu_{11} = E(\phi(X(l)), \phi(X(l'))), \quad l \neq l'. \tag{5}$$

Тогда дисперсия ресамплинг оценки имеет следующий вид:

$$Var\,\Theta^* = E\Theta^{*2} - \mu^2, \;\; E\Theta^{*2} = \frac{1}{r}\left(\mu_2 + (r-1)\mu_{11}\right). \tag{6}$$

Отметим, что значение смешанного момента μ_{11} зависит от способа случайного выбора элементов для ресамплинг-выборки (с повторением, без повторения).

9.4 Регрессионные модели транспортных процессов

Одной из важнейших первоначальных процедур формирования долгосрочного плана является прогнозирование. Прогнозы помогают определить, какие процессы целесообразно развивать опережающими темпами, позволяют подготовить альтернативы их развития и более или менее объективно сравнить их.

Основным математическим аппаратом для моделей прогнозирования являются статистические методы анализа временных рядов и теории регрессии.

Рассмотрим случай прогнозирования на основе регрессионной модели и опишем процедуру применения ресамплинг-метода для оценки параметров регрессии [6].

Как известно, модели линейной регрессии являются одними из наиболее популярных статистических моделей. Они имеют следующий вид:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}, \tag{7}$$

где $\mathbf{X}$ – матрица независимых переменных размера $n\times m$, n – число наблюдений, m – число независимых переменных, $\mathbf{Y}$ – вектор зависимых переменных размерности $n\times 1$, $\mathbf{Z}$ – вектор размерности $n\times 1$, компоненты которого Z_1, Z_2, ... , Z_n являются независимыми, одинаково распределёнными СВ с нулевым средним и дисперсией σ^2, таким образом, $Cov(\mathbf{Z}) = \sigma^2\mathbf{I}_n$, и $\boldsymbol{\beta}$ – вектор размерности $m\times 1$ параметров регрессионной модели, подлежащий оценке.

Классический (традиционный) подход

Классическая оценка параметра $\boldsymbol{\beta}$ методом наименьших квадратов хорошо известна:

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}. \tag{8}$$

В математической статистике доказывается, что в рамках модели линейной регрессии наилучшими, т.е. наиболее эффективными оценками являются оценки, получаемые методом наименьших квадратов.

Пользуясь оценкой (8) мы можем прогнозировать значение зависимой переменной Y_d для выбранного наблюдения с индексом d, представленного вектором $\mathbf{x}_d$ размерности $1\times m$ из значений независимых переменных по следующей формуле:

$$\hat{Y}_d = \mathbf{x}_d \cdot \hat{\boldsymbol{\beta}}. \qquad (9)$$

Оценка (8) может быть успешно применена, если выборка не содержит фальшивых наблюдений, в противном случае полученная оценка может быть не такой эффективной и даже смещённой. Для улучшения качества получаемых оценок предлагается использовать ресамплинг-подход вместо классического. Исследуем эффективность применения предлагаемого метода, принимая смещение оценки в качестве критерия эффективности.

Ресамплинг-подход

Ресамплинг-подход работает следующим образом. Пусть $\mathbf{X} = \left(\mathbf{x}_1^T\ \mathbf{x}_2^T \ldots \mathbf{x}_n^T\right)^T$, где $\mathbf{x}_i$ - i-ая строка матрицы $\mathbf{X}$, которая соответствует i-ому наблюдению, и $N = \{1, 2, \ldots, n\}$ множество целых чисел 1, 2, ... , n. Мы выполняем цикл из r шагов. На текущем (например, l-ом) шаге мы формируем подвыборку путём выбора (без возвращения) k (целое число, $m \le k < n$) чисел из N: $J(l, 1), J(l, 2), \ldots, J(l, k)$. Таким образом, вектор $\mathbf{J}(l) = (J(l, 1)\ J(l, 2) \ldots J(l, k))$ определяет номера наблюдений (строки матриц $\mathbf{X}$ и $\mathbf{Y}$), которые были выбраны на l-ом шаге. Далее мы формируем l-ую подвыборку:

$$\mathbf{X}(\mathbf{J}(l)) = \left(\mathbf{x}_{J(l,1)}^T\ \mathbf{x}_{J(l,2)}^T \ldots \mathbf{x}_{J(l,k)}^T\right)^T,$$

$$\mathbf{Y}(\mathbf{J}(l)) = \left(Y_{J(l,1)}\ Y_{J(l,2)} \ldots Y_{J(l,k)}\right)^T.$$

Теперь мы можем подсчитать оценку $\boldsymbol{\beta}^*$ параметра модели $\boldsymbol{\beta}$ в соответствии с формулой (8):

$$\boldsymbol{\beta}^*(\mathbf{J}(l)) = \left(\mathbf{X}(\mathbf{J}(l))^T \mathbf{X}(\mathbf{J}(l))\right)^{-1} \mathbf{X}(\mathbf{J}(l))^T \mathbf{Y}(\mathbf{J}(l)). \qquad (10)$$

Далее, после r повторений описанной процедуры, мы получаем последовательность оценок:

$$\boldsymbol{\beta}^*(\mathbf{J}(1)), \boldsymbol{\beta}^*(\mathbf{J}(2)), \ldots, \boldsymbol{\beta}^*(\mathbf{J}(r)). \tag{11}$$

Покомпонентное среднее арифметическое элементов данной последовательности даёт ресамплинг-оценку параметра $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}^*(\mathbf{J}) = \frac{1}{r}\sum_{l=1}^{r}\boldsymbol{\beta}^*(\mathbf{J}(l)). \tag{12}$$

В этой главе исследуется устойчивость классических и ресамплинг оценок в случае *нарушенной модели*, используя в качестве критерия эффективности смещение полученных оценок.

Определение нарушенной модели

Рассмотрим случай, когда среди всех наблюдений присутствуют фальшивые. Будем обозначать «правильные» наблюдения индексом t и «фальшивые» наблюдения индексом f. Без потери общности мы можем предположить, что правильные наблюдения соответствуют первым n - h строкам матриц $\mathbf{X}$, $\mathbf{Y}$ и $\mathbf{Z}$:

$$\mathbf{Y} = \begin{bmatrix}\mathbf{Y}_t \\ \mathbf{Y}_f\end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix}\mathbf{X}_t \\ \mathbf{X}_f\end{bmatrix}, \quad \mathbf{Z} = \begin{bmatrix}\mathbf{Z}_t \\ \mathbf{Z}_f\end{bmatrix},$$

где

$$\mathbf{Y}_t = (Y_1\ Y_2\ \ldots Y_{n-h})^T, \qquad \mathbf{Y}_f = (Y_{n-h+1}\ Y_{n-h+2}\ \ldots\ Y_n)^T,$$

$$\mathbf{X}_t = (\mathbf{x}_1^T\ \mathbf{x}_2^T\ \ldots\ \mathbf{x}_{n-h}^T)^T, \qquad \mathbf{X}_f = (\mathbf{x}_{n-h+1}^T\ \mathbf{x}_{n-h+2}^T\ \ldots\ \mathbf{x}_n^T)^T,$$

$$\mathbf{Z}_t = (Z_1\ Z_2\ \ldots Z_{n-h})^T, \qquad \mathbf{Z}_f = (Z_{n-h+1}\ Z_{n-h+2}\ \ldots\ Z_n)^T,$$

$$\mathbf{Y}_t = \mathbf{X}_t\boldsymbol{\beta} + \mathbf{Z}_t, \qquad \mathbf{Y}_f = \mathbf{X}_f\boldsymbol{\beta}_f + \mathbf{Z}_f,$$

$$\begin{gathered} E\mathbf{Z}_t = \mathbf{0}, \quad Cov(\mathbf{Z}_t) = \sigma^2\mathbf{I}_{n-h}, \quad E\mathbf{Z}_f = \mathbf{0}, \\ Cov(Z_f) = \sigma^2\mathbf{I}_h, \quad Cov(\mathbf{Z}_t, \mathbf{Z}_f) = \mathbf{0}. \end{gathered} \tag{13}$$

Такую модель мы будем называть *нарушенной моделью*.

В процессе исследования были получены оценки обоих подходов для нарушенной модели с одним фальшивым наблюдением. Математическое ожидание классической оценки нарушенной модели, представлено следующей формулой:

$$E\,\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} - \frac{1}{1+\mathbf{x}\left(\mathbf{X}_t^T\mathbf{X}_t\right)^{-1}\mathbf{x}^T}\left(\mathbf{X}_t^T\mathbf{X}_t\right)^{-1}\mathbf{x}^T\mathbf{x}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_f\right). \tag{14}$$

Математическое ожидание ресамплинг оценки нарушенной модели представлено следующей формулой:

$$E\,\boldsymbol{\beta}^* = \boldsymbol{\beta} - \binom{n}{k}^{-1} \times$$

$$\times\left[\sum_{\mathbf{u}\in\mathbf{L}_t(k-1)} \frac{1}{1+\mathbf{x}\left(\mathbf{X}^T(\mathbf{u})\mathbf{X}(\mathbf{u})\right)^{-1}\mathbf{x}^T}\cdot\left(\mathbf{X}^T(\mathbf{u})\mathbf{X}(\mathbf{u})\right)^{-1}\right]\mathbf{x}^{\mathbf{T}}\mathbf{x}\left(\boldsymbol{\beta}-\boldsymbol{\beta}_{\mathbf{f}}\right). \tag{15}$$

Оба подхода дают смещённые оценки. Смещение подробно исследуется на численных примерах. В работе также был рассмотрен случай для произвольного числа фальшивых наблюдений.

Ресамплинг-медианные оценки регрессионной модели

Этот подход рассматривался в работе [4].

Обсудим процедуру получения ресамплинг-медианных оценок для регрессионной модели. Мы не можем найти медиану непосредственно для последовательности (11), поскольку элементы последовательности являются векторами, которые невозможно упорядочить обычным образом. Поэтому мы будем применять этот подход к оценке прогнозов. Рассматриваем прогнозирование зависимой переменной для некоторого прошлого или будущего момента времени. Так для d-ого элемента последовательности (11) $\mathbf{x}_d$ (вектор-строка) мы получаем оценку прогноза значения Y по формуле (9) для текущей реализации:

$$Y_d^*(\mathbf{J}(l)) = \mathbf{x}_d\boldsymbol{\beta}^*(\mathbf{J}(l)),\ l = 1,\dots,r \tag{16}$$

Таким образом, проделывая описанную процедуру для каждого элемента последовательности (11) мы получается следующая последовательность оценок прогнозов:

$$Y_d^*(\mathbf{J}(1)), Y_d^*(\mathbf{J}(2)),\dots,Y_d^*(\mathbf{J}(r)) \tag{17}$$

Далее, после упорядочивания по возрастанию, получаем:

$$Y_{d(1)}^*, Y_{d(2)}^*,\dots,Y_{d(s+1)}^*,\dots Y_{d(r)}^*, \tag{18}$$

где r=2s+1. Оценку ресамплинг-выборки, которая дала серединное значение $Y^*_{d(s+1)}$ в упорядоченной последовательности (18), будет называться *ресамплинг-медианной оценкой* параметров регрессии.

Как известно, среднее является несмещённой оценкой математического ожидания. Медианные оценки не обладают таким свойством, но имеют другие преимущество. Рассмотрим следующие особенности, включающие как достоинства, так и недостатки, использования медианного подхода к оцениванию $\boldsymbol{\beta}$, вместо нахождения среднего арифметического оценок (12). Первая особенность – это то, что несмещённость как критерий имеет некоторые недостатки. Дело в том, что несмещенные оценки могут иметь неправильный знак, большую дисперсию и давать неустойчивые результаты в случае зашумлённых данных (наш случай). Второй особенностью является то, что если мы берём $E(\hat{\Theta}-\Theta)^2$ в качестве критерия эффективности, смещённые оценки могут давать лучшие результаты. Последняя особенность – это то, что медианные оценки обладают таким преимуществом, что они являются медианно несмещенными устойчивыми оценками, если мы отклоняемся от допущений статистической гипотезы.

Численный анализ эффективности обоих подходов

Для иллюстрации эффективности предлагаемого подхода, рассмотрим численный пример *прогнозирования потребления некоторого транспортного товара или услуги*. Данные взяты из книги Дрейпера и Смита, где было n=9 наблюдений и m=3 независимых переменных или факторов.

Мы подсчитали квадратическое расстояние Махалонобиса, которое является мерой расстояния от каждого наблюдения до центра:

D=(2.62 5.281 1.95 0.524 3.686 3.695 1.379 1.205 3.659).

Анализируя их, мы сделали вывод, что наблюдения с номерами 2, 6, 5 и 9 из-за своего большого удаления от центра являются наиболее вероятными кандидатами на выбросы.

Сравнение классических и ресамплинг оценок

Был проведён ряд экспериментов для сравнения качества классического и ресамплинг подходов к оцениванию $\boldsymbol{\beta}$. Очевидно, что если мы не имеем выбросов в наших данных, классические оценки (8) будут лучшими в классе линейных несмещённых оценок. Напротив, если в модели присутствуют выбросы, классические оценки имеют значительное смещение. Подсчитаем смещение классической оценки (8) обозначаемое, как $Bias\,\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} - E\,\hat{\boldsymbol{\beta}}$ и смещение ресамплинг-оценки (12) обозначаемое как $Bias\,\boldsymbol{\beta}^* = \boldsymbol{\beta} - E\,\boldsymbol{\beta}^*$. Проверим все возможные комбинации наблюдений, предполагая по очереди каждое наблюдение фальшивым. Соответствующие численные значения представлены в таблице 1. Здесь первая колонка обозначает индекс текущего фальшивого наблюдения, вторая колонка обозначает оцениваемый параметр, третья колонка содержит смещение классической оценки (8). Следующие четыре колонки содержат смещение $Bias\,\boldsymbol{\beta}^*$ ресамплинг оценки (12), причём каждая колонка соответствует своему размеру выборки k=3, 5, 6, 7.

Анализ результатов показывает, что ресамплинг подход даёт лучшие оценки, когда фальшивыми приняты наблюдения с индексами 2 и 6. Отметим, что именно эти наблюдения имели наибольшее расстояния Махалонобиса. Это позволяет нам сформулировать следующие практические рекомендации. Перед началом регрессионного анализа необходимо подсчитать расстояние Махалонобиса для всех наблюдений, с целью выявления наиболее выделяющихся наблюдений по этому критерию. Если существует возможность, что эти выявленные наблюдения могут быть ложными, рекомендуется использовать ресамплинг подход для оценки параметров регрессионной модели.

Ресамплинг-медианные оценки

Для тех же данных мы предприняли ещё одну серию экспериментов. Для рассматриваемой модели за единственное ложное наблюдение мы взяли наблюдение под номером 2, поскольку оно имело наибольшее расстояние от центра. Мы получили оценки прогнозов для некоторого наблюдения (имеющегося или будущего).

Таблица 1

Сравнение смещений классических и ресамплинг оценок параметров регрессионной модели

i	$\boldsymbol{\beta}$	$Bias\ \hat{\boldsymbol{\beta}}$	$Bias\ \boldsymbol{\beta}^*$			
			k=3	k=5	k=6	k=7
1	β_1	0,018	0,268	0,033	0,020	0,015
	β_2	-0,433	-1,500	-0,563	-0,477	-0,436
	β_3	0,263	0,328	0,294	0,278	0,268
2	β_1	1,895	0,989	1,403	0,020	0,015
	β_2	-7,615	-3,900	-5,586	-0,477	-0,436
	β_3	0,158	0,055	0,097	0,278	0,268
3	β_1	0,879	1,561	1,310	1,170	1,027
	β_2	-4,126	-6,844	-5,873	-5,310	-4,729
	β_3	0,397	0,420	0,432	0,425	0,413
4	β_1	-0,350	-1,138	-0,538	-0,452	-0,394
	β_2	1,455	4,757	2,233	1,874	1,635
	β_3	0,042	-0,039	0,026	0,034	0,039
5	β_1	0,171	0,243	0,194	0,178	0,171
	β_2	-1,190	-1,771	-1,363	-1,258	-1,204
	β_3	0,307	0,414	0,339	0,322	0,312
6	β_1	-0,541	0,063	-0,122	-0,266	-0,266
	β_2	3,527	0,678	1,646	2,311	2,311
	β_3	-0,576	-0,364	-0,465	-0,510	-0,510
7	β_1	-0,978	-1,773	-1,395	-1,240	-1,099
	β_2	4,386	7,636	6,125	5,477	4,886
	β_3	-0,172	-0,278	-0,235	-0,210	-0,188
8	β_1	0,056	0,981	0,421	0,259	0,139
	β_2	0,046	-3,401	-1,296	-0,706	-0,262
	β_3	-0,053	-0,058	-0,065	-0,057	-0,054
9	β_1	-1,067	-1,111	-1,224	-1,168	-1,116
	β_2	5,286	5,681	6,013	5,751	5,510
	β_3	-0,410	-0,521	-0,467	-0,444	-0,426

По формуле (18) были получены ресамплинг-медианные оценки прогноза и параметров регрессии. Эти результаты сравнивались с результатами классического подхода, принимая в качестве критерия эффективности смещение получаемых оценок.

С целью изучения влияния размера ресамплинг-выборки k на свойства оценок мы варьировали этот параметр в диапазоне $m \le k < n$. Заметим, что в нашем случае число повторений r ресамплинг-процедуры для получения ресамплинг-оценки было равно количеству всех сочетаний k из n. В общем случае число ресамплинг выборок значительно меньше, чем количество возможных сочетаний.

Мы получили ресамплинг-медианные оценки прогноза зависимой переменной для всех имеющихся наблюдений и сравнили их с классическими оценками прогноза, а также с «чистыми» оценками, полученными без влияния фальшивого наблюдения. Последнее означало, что мы удалили фальшивое наблюдение из нашей модели и использовали классическую оценку наименьших квадратов (8) для своих целей. Соответствующие результаты представлены в таблице 2.

Таблица 2

Смещение ресамплинг-медианного и классического подхода для оценок прогноза для данного примера

Назв. парам етра	Смещение оценок прогноза						
	классиче ский	ресамплинг					
		k=3	k=4	k=5	k=6	k=7	k=8
Y_1	0.115	0.617	0.43	0.332	0.139	**0.034**	0.333
Y_2	5.793	7.57	7.441	5.802	**4.85**	**5.037**	**5.631**
Y_3	1.957	5.131	3.165	2.301	**1.702**	**1.682**	1.967
Y_4	0.497	2.571	**0.108**	**0.243**	0.545	0.739	0.772
Y_5	0.129	1.012	0.636	0.282	**0.024**	**0.042**	**0.066**
Y_6	1.481	4.385	3.53	**1.242**	**0.523**	**0.517**	**1.246**
Y_7	1.519	**0.416**	**0.681**	**0.959**	**1.204**	1.572	1.935
Y_8	1.199	1.557	1.523	**0.111**	**0.073**	**0.539**	**0.749**
Y_9	0.499	2.857	0.956	**0.261**	**0.493**	0.543	0.778

Первая колонка содержит названия прогнозируемых параметров. Вторая колонка содержит смещение, т.е. разницу между математическим ожиданием классической оценки и математическим ожиданием чистой оценки для всех наблюдений. В следующих колонках мы видим результаты применения ресамплинг-подхода для различных объёмов k исходной ресамплинг выборки. Здесь находится смещение, т.е. разница между математическим ожиданием ресамплинг-медианных оценок и чистых оценок для всех наблюдений. Анализируя полученные результаты, мы можем сделать вывод, что для всех наблюдений есть ресамплинг-медианные оценки с меньшим значением смещения, чем у классических, при выборе надлежащего размера выборки. Особенно хорошие результаты были получены для размера ресамплинг-выборки равного 6. Например, для наблюдения с индексом 7, классический подход даёт смещение равное 1.519, а ресамплинг-медианный подход даёт смещения в интервале (0.416-1.572) в зависимости от размера исходной ресамплинг выборки.

Пример регрессионной модели прогнозирования пассажирских перевозок воздушным транспортом.

Рассмотрим пример регрессионной модели прогнозирования пассажирских перевозок воздушным транспортом. Прогнозирование осуществляется на основе следующих факторов:

- x_1 - площадь территории страны (км^2);
- x_2 - численность населения страны (чел.);
- x_3- валовой внутренний продукт на душу населения (EUR);
- x_4 - средняя месячная заработная плата одного жителя (EUR).

Регрессионная модель запишется так:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\varepsilon,$$

где

y - число перевезённых авиатранспортом пассажиров определенной страной за год, ε - случайная составляющая, имеющая нормальное распределение с нулевым средним и постоянной дисперсией.

Исходные данные для прогнозирования получены на официальном статистическом сайте Европейского Союза. Часть данных использовалась для составления модели прогноза. Другая часть данных не участвовала в построении модели и использовалась только для прогноза.

Для анализа корректности предлагаемой классической модели использовались следующие критерии: критерий Фишера для проверки гипотезы о незначимости регрессии (уровень значимости α=5%); критерий Стьюдента для проверки гипотезы о незначимости *i*-ой сопутствующей переменной (α=5%); коэффициент множественной детерминации R^2.

Далее в моделях 2-3 было решено произвести преобразования над переменными, используя комбинации факторов. Прогнозируется индивидуальная подвижность населения (доля перевезённых пассажиров, по отношению к общему числу жителей) страны y/x_2.

Коэффициенты всех моделей были оценены классическим и ресамплинг методом. Сами модели, их параметры и коэффициенты их множественной детерминации представлены в следующей таблице 3.

Таблица 3

Различные модели прогнозирования

Nr.	Регрессионная модель	R^2
1	$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\varepsilon$, Коэффициенты по важности: β_3, β_2, β_4.	0.85
2	$y/x_2=\beta_0+\beta_1x_4+\beta_2(x_1/x_2)+\beta_3(x_3/x_2)+\beta_4 x_3/(x_2*\sqrt{x_1})+$ $+\beta_5 x_3/(x_2*x_1)+\varepsilon$, Коэффициенты по важности: β_1, β_3, β_5, β_4, β_2, β_0.	0.90
3	(Исключён Люксембург) $y/x_2=\beta_0+\beta_1 x_3+\beta_2 x_4+\beta_3(x_1/x_2)+\beta_4(x_3/x_2)+\beta_5(x_4/x_2)+$ $+\beta_6\sqrt{x_1}+\beta_7 x_3/(x_2*\sqrt{x_1})+\beta_8 x_3/(x_2*x_1)+\varepsilon$, Коэффициенты по важности: β_7, β_3 ,β_8, β_2.	0.92

Ресамплинг подход практически везде давал сравнимые и даже лучшие оценки прогнозов, чем традиционный подход, принимая в качестве критерия эффективности несмещённость оценок прогноза. Результаты этого сравнения в процентных соотношениях, представленных в таблице 4.

Таблица 4

Сравнение результатов прогнозов для классического и ресамплинг подходов

Nr.	Данные модели (ресамплинг/классический)	Данные проверки (ресамплинг/классический)
1	24/14 (63/37)%	17/7 (71/29)%
2	21/17 (55/45)%	15/9 (62.5/37.5)%
3	25/10 (71/29)%	5/14 (26/74)%

Выводы

Рассмотрено применение ресамплинг подхода для оценивания модели линейной регрессии, когда среди наблюдений присутствуют фальшивые наблюдения, не принадлежащие данному уравнению регрессии. Было подсчитано значение смещения оценок параметров классического и ресамплинг подходов. Рассмотрены условия, при которых ресамплинг подход даёт оценки с меньшим смещением. Показано, что применение ресамплинг подхода для нарушенной регрессионной модели ведёт к лучшим результатам. Анализ численных результатов показывает также, что предлагаемый ресамплинг-медианный подход даёт лучшие оценки, чем классический, если мы используем смещение оценки прогноза $E(Y_d)$ как критерий эффективности.

Применение ресамплинг подхода было проиллюстрировано для прогнозирования объёмов перевозок пассажиров авиационным транспортом.

9.5 Об одной задаче оценивания надёжности и эффективности транспортных средств

Содержательная постановка задачи

Надёжность – это вероятность того, что устройство выполняет свои функции в соответствии с предъявленными требованиями в течение заданного интервала времени. Как известно, для транспорта этот показатель имеет особенное значение. Проблемы, связанные с надёжностью транспортных систем (транспортных средств), могут зачастую стать жизненно важными. Проблемы, связанные с надёжностью поставок в логистических цепочках, могут привести к огромным финансовым потерям компаний. Всё это требует привлечения математического аппарата оценки и прогнозирования, способствующие выработке правильных управленческих решений. Для этих задач характерным является ограниченность объёмов исходных статистических данных, например об отказах, что ограничивает применение классических методов оценивания. Это требует привлечения новых интенсивных компьютерных методов, способных преодолеть эти проблемы. Результаты этого исследования были опубликованы в работах [2], [9].

Математическая модель и методы её анализа

Рассмотрим следующую математическую модель теории массового обслуживания, характерную для теории надёжности. Модель предполагает существование *повреждений*, которые со временем перерождаются в *отказ*. Повреждения возникают в соответствии с однородным процессом Пуассона с интенсивностью λ. Каждое повреждение перерастает в отказ через случайное время *B*. Таким образом, если повреждение появилось в момент τ_i, оно переродится в отказ в момент времени $B_i + \tau_i$. Отказы и соответствующие повреждения мгновенно устраняются. Мы предполагаем, что $\{B_i\}$ являются взаимно-независимыми СВ, не зависящими также от $\{\tau_i\}$. Пусть $F(x)$ – функция распределения СВ *B*. Мы интересуемся числом повреждений $X(t)$ в момент времени *t* (которые не переросли в отказ) и числом отказов $Y(t)$, которые произошли до момента *t*. Пусть $\Lambda(t) = EX(t)$ и $\widetilde{\Lambda}(t) = EY(t)$ будут соответствующими математическими ожиданиями, а $P_i(t) = P\{X(t) = i\}$ и $R_i(t) = P\{Y(t) = i\}$ соответствующими вероятностными распределениями, $i = 0, 1, \dots$.

Поскольку хорошо известно, что $X(t)$ и $Y(t)$ являются взаимно-независимыми СВ, то:

$$\Lambda(t) = \lambda \int_0^t \left(1 - F(x)\right) dx\,,\ \widetilde{\Lambda}(t) = \lambda \int_0^t F(x)\,dx\,, \tag{19}$$

$$P_i(t) = \frac{1}{i!} (\Lambda(t))^i \exp\left(-\Lambda(t)\right),\ i = 0, 1, \dots \tag{20}$$

Вероятность $R_i(t)$ вычисляется аналогично по формуле (20), где $\Lambda(t)$ заменяется на $\widetilde{\Lambda}(t)$.

Фактически интенсивность λ и функция распределения $F(x)$ не известны. Нужно оценить $\Lambda(t)$, $\widetilde{\Lambda}(t)$, $R_i(t)$ и $P_i(t)$, используя выборку интервалов между возникновением повреждений $A_1, A_2, \dots, A_k$ и выборку $B_1, B_2, \dots, B_l$.

Мы рассматриваем два метода оценивания: традиционный «плаг-ин» и ресамплинг. Исследуется свойства математического ожидания и дисперсии оценок, полученных обоими подходами. Оценки сравниваются, используя в качестве критерия эффективности смещение и дисперсию. Показывается, что ресамплинг-оценки имеют некоторые преимущества для малых объёмов выборок k и l.

Сущность традиционного «плаг-ин» подхода

Традиционный плаг-ин (plug-in) метод оценивание предполагает использование оценок $\hat{\lambda}$ и $\hat{F}(t)$ вместо неизвестных λ и $F(t)$. Здесь $\hat{F}(t)$ является эмпирической функцией распределения СВ B, которая подсчитывается по выборке $B_1, B_2, \dots, B_l$, $\hat{\lambda}$ является точечной оценкой интенсивности λ:

$$\hat{\lambda} = \left(\frac{1}{k} \sum_{i=1}^{k} A_i \right)^{-1}. \tag{21}$$

В этом случае оценки параметров $\Lambda(t)$ и $P_i(t)$ следующие:

$$\hat{\Lambda}(t) = \hat{\lambda}\int_0^t \left(1 - \hat{F}(x)\right)dx, \tag{22}$$

$$\hat{P}_i\ (t) = \frac{1}{i!}\hat{\Lambda}(t)^i \exp\left(-\hat{\Lambda}(t)\right),\ i = 0,\ 1,\ \ldots\ . \tag{23}$$

Оценки параметров $\widetilde{\Lambda}(t)$ и $R_i(t)$ вычисляются по аналогии.

Для того, чтобы изучать статистические свойства этих оценок необходимо знать распределение СВ $\hat{\lambda}$ и $\int_0^t (1 - \hat{F}(u))du$. Эти распределения вычислены в работе. Были получены формулы для математического ожидания $E\hat{\Lambda}(t)$ и дисперсии $Var\,\hat{\Lambda}(t)$ плаг-ин оценок (22), (23). Среднеквадратическая ошибка оценки $\hat{\Lambda}(t)$ может быть найдена последующей формуле: $MSE\hat{\Lambda}(t) = Var\hat{\Lambda}(t) + \left(E\hat{\Lambda}(t) - \Lambda(t)\right)^2$.

Сущность ресамплинг-подхода

Ресамплинг-подход предполагает процедуру обычного имитационного моделирования с той лишь разницей, что мы не генерируем необходимые значения СВ, а берём их случайным образом непосредственно из имеющихся выборочных совокупностей $\{A_1, A_2, ..., A_k\}$ и $\{B_1, B_2, ..., B_l\}$. Пусть $k \le l$. Мы делаем r независимых реализаций процесса имитационного моделирования. На некоторой q-ой реализации мы выбираем элементы из $\{A_1, A_2, ..., A_k\}$ без возвращения, формируем последовательность временных интервалов между появлениями повреждений $A(q) = \{A_{i_1(q)}, A_{i_2(q)}, ..., A_{i_{N_t}(q)}\}$ и подсчитываем $\tau_i(q) = \sum_{u=1}^{i} A_{i_u(q)}$, $i = 1,2,\ \ldots, N_t(q)$, где $N_t(q)$ число повреждений до момента t на q-ой реализации:

$$N_t(q) = \begin{cases} \max\{j : \tau_j(q) \le t\} & если \quad \tau_k(q) \ge t, \\ k & иначе. \end{cases} \tag{24}$$

Аналогично, мы получаем последовательность $B(q) = \{B_{j_1(q)}, B_{j_2(q)}, ..., B_{j_{N_t}(q)}\}$ временных интервалов перерастания повреждения в отказ. Потом мы строим последовательность $\{\tau_1(q) + B_{j_1(q)}, \tau_2(q) + B_{j_2(q)}, ..., \quad \tau_{N_t}(q) + B_{j_{N_t}(q)}\}$ моментов отказов для q-ой реализации.

Пусть $\zeta_j(t)$ - индикаторная функция события: “j-ое повреждение произошло, но ещё не привело к отказу до момента t ”:

$$\zeta_{j,q}(t) = \begin{cases} 1 & если \quad \tau_j(q) \le t < \tau_j(q) + B_{j_j(q)}, \\ 0 & иначе. \end{cases} \tag{25}$$

Тогда число повреждений $\mathrm{X}_q(t)$, которые не переросли в отказ до момента t для q-ой реализации:

$$\mathrm{X}_q(t) = \sum_{j=1}^{N_t(q)} \zeta_{j,q}(t) = \sum_{j=1}^{k} \zeta_{j,q}(t). \tag{26}$$

Ресамплинг оценка $\Lambda(t)$ следующая:

$$\Lambda^*(t) = \frac{1}{r} \sum_{q=1}^{r} X_q(t). \tag{27}$$

Аналогично, число повреждений $Y_q(t)$, которые переросли в отказ до момента t на q-ой реализации подсчитывается по формуле (26), соответствующей заменой функции $\zeta_{j,q}(t)$. Ресамплинг-оценка $\widetilde{\Lambda}^*(t)$ может быть найдена по формуле (27) заменой $X_q(t)$ на $Y_q(t)$.

Теперь подсчитаем ресамплинг-оценки вероятностей $P_i(t)$ и $R_i(t)$. Пусть $\Phi_i(X(t))$ - индикаторная функция события $\{X(t)=i\}$. Тогда ресамплинг-оценка вероятности $P_i(t)$ следующая:

$$P_i^*(t) = \frac{1}{r}\sum_{q=1}^{r}\Phi_i(\mathrm{X}_q(t)). \tag{28}$$

Ресамплинг-оценка $R_i^*(t)$ может быть найдена по формуле (28) заменой $\Phi_i(X(t))$. Подсчитаем математическое ожидание ресамплинг оценок. Очевидно, что $EP_i^*(t) = E\Phi_i(t)$. Следовательно, математическое ожидание $E\Lambda^*(t)$ ресамплинг-оценки $\Lambda^*(t)$ вычисляется следующим образом:

$$E\Lambda^*(t) = q_1\sum_{j=1}^{k} j\,\frac{(\lambda t)^j}{j!}\cdot\exp(-\lambda t) + q_1 k\cdot\sum_{j=k+1}^{\infty}\frac{(\lambda t)^j}{j!}\cdot\exp(-\lambda t). \tag{29}$$

Мы также можем найти математическое ожидание $EP_i^R(t)$:

$$\begin{aligned} EP_i^*(t) = {} & \sum_{j=i}^{l}\frac{(\lambda t)^j}{j!}\cdot\exp(-\lambda t)\cdot\binom{j}{i}\cdot q_1{}^i\cdot(1-q_1)^{j-i} + \\ & + \binom{l}{i}\cdot q_1{}^i\cdot(1-q_1)^{l-i}\cdot\sum_{j=l+1}^{\infty}\frac{(\lambda t)^j}{j!}\cdot\exp(-\lambda t),\ \ i = 0,1,2,..., \end{aligned} \tag{30}$$

где $\binom{j}{i}$ - биноминальный коэффициент и q_1 – вероятность того, что до момента времени t рассматриваемое повреждение не перерастёт в отказ.

Математические ожидания $ER_i^*(t)$ и $E\tilde{\Lambda}^*(t)$ оценок $R_i^*(t)$ и $\tilde{\Lambda}^*(t)$ подсчитываются аналогично. Были получены также выражения для дисперсии $Var\Lambda^*(t)$ ресамплинг-оценок (27). Среднеквадратическая ошибка может быть подсчитана по следующей формуле: $MSE\Lambda^*(t) = Var\Lambda^*(t) + \left(E\Lambda^*(t) - \Lambda(t)\right)^2$.

Численный анализ эффективности подхода
Пример: “Треугольное время перерастания неисправности в отказ”.

Рассмотрим следующую систему массового обслуживания: входной Пуассоновский поток требований с интенсивностью λ=0.4; треугольное время обслуживания с параметрами a=2.

В таблице 5 показаны искомые математические ожидания оценок традиционного (Trad.) $E\hat{P}_i(t)$ и ресамплинг (Res.) $EP_i^*(t)$ подходов вероятности соответствующей длины очереди в фиксированный момент времени t=5, с различными объёмами ресамплинг выборок l, в сравнении с истинными вероятностями.

Мы рассматриваем случай, когда объёмы исходных выборок l и k одинаковые. Здесь мы видим, что с увеличением l смещение становится всё меньше и меньше, а у математического ожидания ресамплинг-оценки практически исчезает.

На рисунках 3-4 мы можем легко сравнить эти оценки. На рисунке 4 видно, что кривая истинной вероятности и математического ожидания ресамплинг оценки сливаются в одну линию.

Таблица 5

Математические ожидания классических оценок $E\hat{P}_i(5)$ и ресамплинг-оценок $EP_i^*(5)$

i	*l=3*		*l=4*		*l=5*		*l=8*		Истинная
	Plug.	Res.	Plug.	Res.	Plug.	Res.	Plug.	Res.	
0	.346	.379	.348	.370	.350	.368	.352	.368	.368
1	.291	.392	.307	.374	.317	.369	.334	.368	.368
2	.169	.189	.176	.189	.180	.186	.186	.184	.184
3	.088	.040	.087	.058	.085	.062	.080	.061	.061
4	.045		.041	.009	.037	.014	.031	.015	.015
5	.024		.019		.016	.002	.011	.003	.003
6	.013		.010		.007		.004		.001
7	.008		.005		.003		.001		
8	.005		.003		.002		.001		
9	.003		.001		.001				
10	.002		.001						

В таблице 6 находятся математические ожидания $E\hat{\Lambda}(t)$, $E\Lambda^*(t)$, дисперсии $Var\hat{\Lambda}(5)$, $Var\Lambda^*(5)$ и среднеквадратические ошибки $MSE\ \hat{\Lambda}(5)$, $MSE\ \Lambda^*(5)$.

В процессе исследования рассматривались различные примеры с различными видами распределения и значениями параметров. Анализ результатов подтвердил только что отмеченные тенденции. Во всех случаях ресамплинг-подход давал лучшие результаты, чем традиционный. Это было особенно заметно, когда объём выборки превышал число 3, поскольку такой объём выборки, не мог дать достоверных результатов ресамплинг подходу. Мы не могли получить вероятности длины очереди больше, чем *l*, хотя в реальной жизни, такие случаи бывают. Ресамплинг-подход даёт в большинстве случаев меньшее смещение, чем традиционный, а когда объём выборки достаточно большой (больше 8), смещение ресамплинг-оценок практически пропадает.

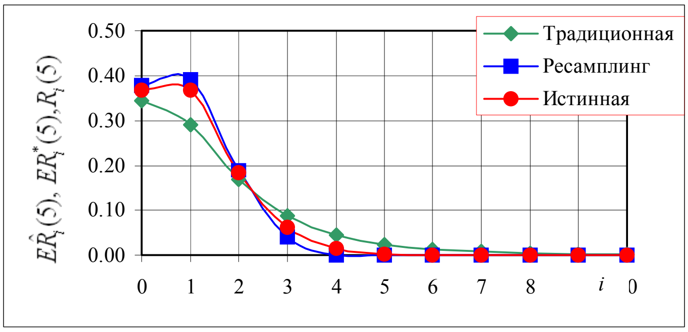

Рис. 3 Математическое ожидание традиционной оценки $E\hat{P}_i\ (5)$ и ресамплинг-оценки $EP_i^*(5)$, *l*=3, *t*=5.

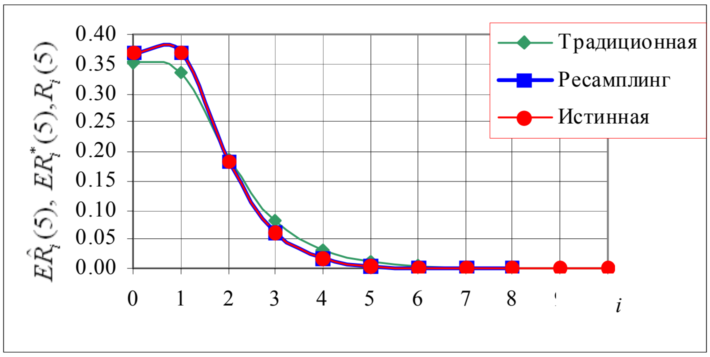

Рис. 4 Математическое ожидание традиционной оценки $E\hat{P}_i(5)$ и ресамплинг-оценки $EP_i^*(5)$, *l*=8, *t*=5.

Таблица 6

Математические ожидания, дисперсии, среднеквадратические ошибки плаг-ин и ресамплинг оценок $\Lambda(t)$

i	*l*=3	*l*=4	*l*=5	*l*=6	*l*=7	*l*=8
$E\hat{\Lambda}(5)$	1.41	1.32	1.25	1.21	1.19	1.16
$E\Lambda^*(5)$	0.89	0.96	0.99	0.997	0.99	0.99
$Var\hat{\Lambda}(5)$	1.52	0.79	0.51	0.38	0.30	0.24
$Var\Lambda^*(5)$	0.58	0.55	0.49	0.43	0.36	0.31
$MSE\ \hat{\Lambda}(5)$	1.69	0.89	0.57	0.42	0.34	0.27
$MSE\ \Lambda^*(5)$	0.59	0.55	0.49	0.43	0.36	0.31

Выводы

Предлагаемый ресамплинг-подход является хорошей альтернативой традиционному для рассмотренной задачи надёжности. Особенно примечательно, что с увеличением размера выборки скорость сходимости у ресамплинг-оценок к истинным значениям значительно быстрее, чем у традиционных. Единственным недостатком предлагаемого подхода является то, что мы не можем получить ресамплинг-оценки $EP_i^*(t)$ при $i > l$, поэтому в таких случаях лучше использовать традиционный подход. Лучшим решением в данной ситуации будет комбинация этих двух подходов: использовать ресамплинг-оценки, для $i \le l$, и традиционные оценки для $i > l$. Мы можем использовать такую комбинацию со специальной нормализацией, чтобы сумма вероятностей была равна 1.

Важно подчеркнуть, что был рассмотрен случай, когда вид распределения был выбран правильно. В реальной жизни мы часто сталкиваемся с ситуациями, когда доступные выборки слишком малы (как в нашем случае) и мы не можем правильно выбрать вид распределения. Таким образом, может быть построить совсем другая модель, результатам которой вообще нельзя верить. В таких ситуациях особенно выгодно и эффективно использовать ресамплинг-подход. Поскольку ресамплинг подход составлял конкуренцию традиционному и в менее выигрышных ситуациях, то в последнем случае его преимущества ещё более очевидны.

9.6 Управление запасами логистических систем

Формулировка и математическая постановка задачи

Если рассматривать запасы как объект управления, то ключевыми являются такие вопросы, как расчет оптимального уровня запасов, принимая во внимание риски возникновения дефицита.

Рассмотрим следующую задачу. Предположим, что имеются два простых независимых процесса восстановления $\{X_i, i=1,2,...\}$ и $\{Y_i, i=1,2,...\}$, где $\{X_i\}$ и $\{Y_i\}$ являются последовательностью неотрицательных независимых СВ, каждая последовательность со своим одним и тем же распределением. Пусть $D_m = \sum_{i=1}^{m} X_i$ и $S_m = \sum_{i=1}^{m} Y_i$ - моменты m-ого восстановления соответствующих процессов. Функции распределения последовательностей $\{X_i\}$ и $\{Y_i\}$ неизвестны, но имеются соответствующие выборочные совокупности объёмов n_X и n_Y. Нашей целью является оценить вероятность $P\{D_m > S_k\}$, где $n_X \ge 2m$ и $n_Y \ge 2k$.

Эта задача имеет множество приложений. Например, в теории управления запасами, она возникает в следующей ситуации.

Предположим, что задан начальный уровень запаса K, где K известное целое число. Уровень запаса возрастает в соответствии с предложением и уменьшается в соответствии со спросом. Предполагается также, что если спрос превышает предложение, возникает дефицит. Нашей целью является оценить вероятность отсутствия дефицита в момент спроса на m-ое изделие.

Формальная постановка задачи

Описанный пример может быть рассмотрен в терминах процессов восстановления следующим образом. Пусть спрос соответствует первому процессу восстановления $\{X_i,\ i=1,2,...\}$ и момент m-ого восстановления является моментом запроса на m-ое изделие. Пусть предложение соответствует второму процессу восстановления $\{Y_i,\ i=1,2,...\}$ и момент m-ого восстановления является моментом поставки m-ого изделия. Тогда интересующая нас вероятность отсутствия дефицита – это вероятность того, что m-ое изделие потребуется раньше, чем произойдёт $m-K$-ая поставка: $D_m>S_{m-K}$. Предполагается также, что начальный уровень запаса K известен. Нашей задачей является изучение некоторых свойств различных оценок искомой вероятности.

Функции распределения последовательностей $\{X_i\}$ и $\{Y_i\}$ соответственно $F_X^1(x)$ и $F_Y^1(x)$ неизвестны, но имеются соответствующие выборки $H_X=\{X_1,X_2,\ldots,X_{n_X}\}$ и $H_Y=\{Y_1,Y_2,\ldots,Y_{n_Y}\}$, где $|H_X|=n_X$ и $|H_Y|=n_Y$.

Мы интересуемся моментами m-ого и $m-K$-ого восстановлений: $D_m=\sum_{i=1}^{m}X_i$, $S_{m-K}=\sum_{i=1}^{m-K}Y_i$. Нашей задачей является оценить вероятность: $P\{D_m>S_{m-K}\}$, что m-ое восстановление процесса спроса $\{X_i\}$ произойдёт позже, чем $m-K$-ое восстановление процесса предложения $\{Y_i\}$.

Рассмотрим индикаторную функцию $\Psi(\mathbf{x},\mathbf{y})$, где $\mathbf{x}=(x_1,x_2,\ldots,x_{m_X})$ и $\mathbf{y}=(y_1,y_2,\ldots,y_{m_Y})$ вектора из вещественных чисел:

$$\Psi(\mathbf{x},\mathbf{y}) = \begin{cases} 1 & \textit{если} \ \sum_{i=1}^{m_X} x_i > \sum_{i=1}^{m_Y} y_i \\ 0 & \textit{иначе.} \end{cases} \tag{31}$$

Предположим, что мы имеем два вектора СВ $\mathbf{X} = (X_1, X_2, \ldots X_{m_X})$ и $\mathbf{Y} = (Y_1, Y_2, \ldots Y_{m_Y})$, $m_X = m$, $m_Y = m - K$. Искомая вероятность может быть представлена, как математическое ожидание индикаторной функции (31): $\Theta = E(\Psi(\mathbf{X},\mathbf{Y}))$. Мы будем оценивать Θ, используя два различных подхода: классический и ресамплинг. Классический плаг-ин, параметрический подход широко известен. Предлагаемый ресамплинг-метод является альтернативным непараметрическим подходом.

Традиционный подход

Традиционный подход предполагает точечное оценивание параметров соответствующих распределений, если известны виды распределений и начальные выборки H_i, $i=\{X, Y\}$

Пример: Экспоненциальное распределение

Рассмотрим случай, когда СВ X и Y имеют экспоненциальное распределение с параметрами λ и ν соответственно. Как известно, сумма экспоненциально распределённых СВ имеет распределение Эрланга. Интересующая нас вероятность запишется как $\Theta = P\{D_{m_X} > S_{m_Y}\}$.

Подставив, вместо параметров λ и ν их точечные оценки, получим следующее выражение для оценки искомой вероятности:

$$\hat{\Theta} = \sum_{i=0}^{m_X - 1} \frac{\hat{\nu}^{m_Y}}{(\hat{\lambda} + \hat{\nu})^{m_Y + i}} \frac{\hat{\lambda}^i}{i!} \prod_{p=0}^{i-1} (m_Y + p), \tag{32}$$

где $\prod_{p=0}^{-1} = 1$, $\hat{\lambda} = n_X / D_{n_X}$ и $\hat{\nu} = n_Y / S_{n_X}$.

Теперь мы можем подсчитать математическое ожидание и дисперсию $\hat{\Theta}$.

Пример: Нормальное распределение

Рассмотрим теперь случай когда СВ X и Y имеют нормальное распределение, соответственно $N(\mu_X,\sigma_X)$ и $N(\mu_Y,\sigma_Y)$. Реальная искомая вероятность имеет следующий вид: $\Theta = P\{D_{m_X} > S_{m_Y}\} = P\{D_{m_X} - S_{m_Y} > 0\}$.

При использовании традиционного подхода для оценки этой вероятности, предполагается оценивание параметров $\boldsymbol{\mu} = (\mu_X, \mu_Y)$, $\boldsymbol{\sigma} = (\sigma_X, \sigma_Y)$, по имеющимся выборочным совокупностям. Мы имеем следующие оценки:

$$\hat{\Theta} = \hat{\Theta}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}) = 1 - \Phi\left(\frac{0 - (m_X \hat{\mu}_X - m_Y \hat{\mu}_Y)}{\sqrt{m_X \hat{\sigma}_X^2 + m_Y \hat{\sigma}_y^2}}\right). \qquad (33)$$

Теперь мы можем подсчитать математическое ожидание $E\,\hat{\Theta}$, дисперсию $Var\,\hat{\Theta}$ и среднеквадратическую ошибку $MSE\,\hat{\Theta} = Var\,\hat{\Theta} + (\Theta - E\,\hat{\Theta})^2$ оценки $\hat{\Theta}$.

Ресамплинг - подход

Этот метод в отличие от традиционного, не предполагает оценивание параметров распределений или конструирование эмпирических функций распределений с целью найти искомые характеристики. В данном подходе мы используем непосредственно данные выборок в различных комбинациях, что способствует получению несмещённых оценок и уменьшению дисперсии. Ресамплинг-подход предполагает выполнение следующих шагов. Мы выбираем случайно m_X элементов из выборки H_X и m_Y элементов из выборки H_Y. Элементы выбираются без возвращения, вспоминая что $n_X \ge 2m_X$, $n_Y \ge 2m_Y$. Далее мы можем подсчитать соответствующее значение функции $\Psi(\mathbf{x}, \mathbf{y})$, используя формулу (31). Затем мы возвращаем выбранные элементы в выборки и повторяем описанную процедуру в течение r реализаций.

Пусть $j_d^i(l)$, d=1,..,m_i индексы элементов из выборки H_i, $i \in \{X, Y\}$, которые выбраны на l-ой реализации. Тогда для l-ой реализации мы получаем следующие вектора:

$$\mathbf{X}(l)=(X_{j_1^X(l)},X_{j_2^X(l)},\dots,X_{j_{m_X}^X(l)})\,,\ \mathbf{Y}(l)=(Y_{j_1^Y(l)},Y_{j_2^Y(l)},\dots,Y_{j_{mY}^Y(l)})\,.$$

Ресамплинг-оценка Θ^*, являющаяся среднем арифметическим по r реализациям, вычисляется по формуле (3), заменив $\phi(X(l))$ на $\Psi(\mathbf{X}(l),\mathbf{Y}(l))$.

Очевидно, что эта оценка является несмещённой. Мы интересуемся дисперсией этой оценки. Моменты и дисперсия могут быть подсчитаны по формулам (4-6), заменив функцию $\phi(X(l))$ функцией $\Psi(\mathbf{X}(l),\mathbf{Y}(l))$.

Для того, чтобы оценить дисперсию оценки, мы должны сначала найти выражение для смешенного момента μ_{11} по формуле (5). В нашем случае для этого необходимо ввести нотацию α-пар.

Обозначим $W_i(l)$, l=1,...,r, $i\in\{X,Y\}$, подмножество выборки H_i, которое было использовано для получения значений векторов $\mathbf{X}(l)$ и $\mathbf{Y}(l)$ соответственно, $W_i(l)\subset H_i$. Обозначим $M_i=\{0,1,...,m_i\}$, $M=M_X\times M_Y$. Пусть $\boldsymbol{\alpha}=(\alpha_X,\ \alpha_Y)$ – элемент множества M, $\boldsymbol{\alpha}\in M$. Мы говорим, что $W_i(l)$ и $W_i(l')$ образуют **α**-пару, тогда и только тогда, когда $W_i(l)$ и $W_i(l')$ имеют ровно α_i общих элементов: $|W_i(l)\cap W_i(l')|=\alpha_i$.

Пусть $A_{ll'}(\boldsymbol{\alpha})$ обозначает событие “подвыборки $(\mathbf{X}(l),\mathbf{Y}(l))$ и $(\mathbf{X}(l'),\mathbf{Y}(l'))$ образуют **α**-пару”, а $\mathrm{P}_{ll'}(\boldsymbol{\alpha})$ – есть вероятность этого события: $\mathrm{P}_{ll'}(\boldsymbol{\alpha})=\mathrm{P}\{A_{ll'}(\boldsymbol{\alpha})\}$. Поскольку все реализации l=1,..,r являются статистически эквивалентными, мы можем опустить нижний индекс ll' и писать $\mathrm{P}(\boldsymbol{\alpha})$.

Пусть

$$\mu_{11}(\boldsymbol{\alpha})=E\big(\Psi(\mathbf{X}(l),\mathbf{Y}(l))\Psi(\mathbf{X}(l'),\mathbf{Y}(l'))\,|\,A_{ll'}(\boldsymbol{\alpha})\big)\,, \tag{34}$$

тогда

$$\mu_{11}=\sum_{\boldsymbol{\alpha}\in M}P(\boldsymbol{\alpha})\mu_{11}(\boldsymbol{\alpha})\,. \tag{35}$$

Следовательно, нам нужно подсчитать $\mathrm{P}\{\boldsymbol{\alpha}\}$ и $\mu_{11}(\boldsymbol{\alpha})$ для всех $\boldsymbol{\alpha}\in M$. Вероятность $\mathrm{P}\{\boldsymbol{\alpha}\}$ может быть подсчитана в соответствии с гипергеометрическим распределением. Формула для $\mu_{11}(\boldsymbol{\alpha})$, $\forall\boldsymbol{\alpha}\in M$ была также получена.

Численный анализ эффективности подхода

Пример: Нормальное распределение

Пусть СВ *X* и *Y* имеют нормальное распределение с параметрами $\mu_X = \mu_Y = 2$, $\sigma_X = \sigma_Y = 1$. Размеры исходных выборок одинаковы $n = n_X = n_Y$.Мы интересуемся отмеченной вероятностью в момент спроса на *m*-ое изделие при различных уровнях начального запаса *K*=0..3. Все вычисления проводились для *r* = 1000 реализаций. Сравним дисперсию оценок ресамплинг-подхода со среднеквадратической ошибкой традиционного подхода. Примем во внимание, что ресамплинг-подход даёт несмещённые оценки, в отличие от традиционного подхода.

Рассмотрим таблицу 7.

Таблица 7

Экспериментальные результаты для традиционных $\hat{\Theta}$ и ресамплинг Θ^* оценок

		K=0	*K*=1	*K*=2	*K*=3
n=10 *m*=5	*Var* $\hat{\Theta}$	.061	.043	.015	.002
	Bias $\hat{\Theta}$	0	.028	.029	.013
	MSE $\hat{\Theta}$	.061	.044	.015	.002
	Var Θ^*	.087	.055	.014	.001
n=10 *m*=4	*Var* $\hat{\Theta}$	.053	.032	.006	---
	Bias $\hat{\Theta}$	0	.021	.018	---
	MSE $\hat{\Theta}$	.053	.033	.007	---
	Var Θ^*	.069	.039	.005	---
n=12 *m*=6	*Var* $\hat{\Theta}$	.06	.045	.019	.004
	Bias $\hat{\Theta}$	0	.028	.033	.019
	MSE $\hat{\Theta}$	.06	.046	.02	.004
	Var Θ^*	.085	.058	.02	.002

В таблице 7 мы можем видеть дисперсию ресамплинг-оценки $Var\,\Theta^*$ в сравнении с дисперсией $Var\,\hat{\Theta}$, смещением $Bias\,\hat{\Theta}$ и среднеквадратической ошибкой $MSE\,\hat{\Theta}$ традиционной оценки. Таблица 7 показывает, как изменяются результаты при различных объёмах исходных выборок *n*, моментах спроса на *m*-ое изделие и начальном уровне запаса *K*. Мы можем прийти к выводу, что дисперсия и соответствующая среднеквадратическая ошибка оценок обоих подходов уменьшается при возрастании размеров выборок *n*, *m* и уровня начального запаса *K*. Дисперсия ресамплинг-оценки почти всегда близка к дисперсии традиционной. Однако, поскольку ресамплинг-оценка является несмещённой, то принимая в качестве критерия эффективности среднеквадратическую ошибку, ресамплинг при больших *K* даёт лучшие результаты.

Был также рассмотрен случай, когда СВ *X* и *Y* имеют экспоненциальное распределение с параметрами λ=0.3, ν=0.7. Для нахождения той же вероятности были так же, как и в прошлом примере рассмотрены два подхода. Критерии оценки эффективности были те же. Ресамплинг-подход также в некоторых случаях показал лучшие результаты..

Пример из области логистики

Рассмотрим пример определения среднего дохода от поступивших изделий со следующими исходными данными:

- *K* - первоначальный запас;
- $f(k)$ – стоимость первоначального заказа (например, $f(k)=b_0+b_1\cdot k$);
- c_d – доход от удовлетворения одного запроса;
- c_s – штраф от задержки своевременного удовлетворения заказа;

Средний доход от удовлетворения *m* заказов:

$$\Pi(m,K) = m\cdot c_d - \left[f(K) + c_s \cdot \sum_{i=1}^{m} P_i(K) \right], \qquad (36)$$

где $P_i(K)$ - вероятность того, что при начальном запасе равном *K* в момент спроса на *i* –ое изделие возникнет дефицит. Поскольку ресамплинг подход даёт несмещённые оценки оцениваемой функции, то математическое ожидание оценки этой вероятности совпадает с истинным значением. В предыдущем пункте была найдена вероятность отсутствия дефицита в момент спроса на i - ое изделие, что является вероятностью противоположного события $1 - P_i(K)$.

Средний убыток в процессе удовлетворения m заказов:

$$\Xi(m,K) = f(K) + c_s \cdot \sum_{i=1}^{m} P_i(K). \tag{37}$$

Таким образом: $\Pi(m,K) = m \cdot c_d - \Xi(m,K)$. Задача сводится к максимизации прибыли или к минимизации издержек:

$$\max_K \Pi(m,K) = m \cdot c_d - \Xi(m,K) \text{ или } \min_K \Xi(m,K) = f(K) + c_s \cdot \sum_{i=1}^{m} P_i(K).$$

Проанализируем изменение этой функции от K при фиксированных параметрах $c_d=2$, $c_s=5$, $b_0=0$, $b_1=0.2$, при нормально распределённом спросе $N(2,1)$, при нормально распределённом предложении $N(2.5,0.2)$.

Понятно, что в реальной ситуации распределения спроса и предложения нам не известны. Однако мы можем их оценить на основе имеющихся выборок, а также оценить искомую вероятность дефицита и средний доход от поступивших изделий. В данной главе описывались традиционный “plug-in” и ресамплинг методы для соответствующего оценивания.

Было отмечено, что при малых объёмах исходных выборок целесообразней использовать ресамплинг подход, поскольку он даёт несмещённые оценки параметров и обладает меньшей среднеквадратической ошибкой. В таблице 8 приведены соответствующие результаты (П- реальный доход и $\hat{\Pi}$ - классическая оценка) при изменении начального объёма запаса K и времени (момента спроса на m-ое изделие). Из данных результатов видно, что при поиске оптимального размера начального запаса мы можем определить неправильный оптимум из-за смещения классической оценки.

На рисунке 5 показано, что в момент спроса на 5-ое изделие оптимальный объём начального запаса составляет 3 единицы, в то время как классическая оценка достигает оптимума при начальном запасе равном 4-ём единицам. Поскольку ресамплинг подход даёт несмещенные оценки параметров, то он позволяет избежать только что отмеченных недостатков классического подхода.

Таблица 8

Смещение классической оценки среднего дохода от реального значения

		K=0	K=1	K=2	K=3	K=4
m=1	$\Pi(1,K)$	-2.816	1.8	1.6	1.4	1.2
	$\hat{\Pi}(1,K)$	-2.872	1.8	1.6	1.4	1.2
m=3	$\Pi(3,K)$	-9.723	2.781	5.444	5.4	5.2
	$\hat{\Pi}(3,K)$	-11.197	2.709	5.375	5.4	5.2
m=5	$\Pi(5,K)$	-17.624	0.462	8.02	9.285	9.197
	$\hat{\Pi}(5,K)$	-19.9	0.34	7.604	9.144	9.18
m=7	$\Pi(7,K)$	-26.139	-4.109	8.532	12.572	13.129
	$\hat{\Pi}(7,K)$	-28.814	-3.995	7.653	11.954	12.945

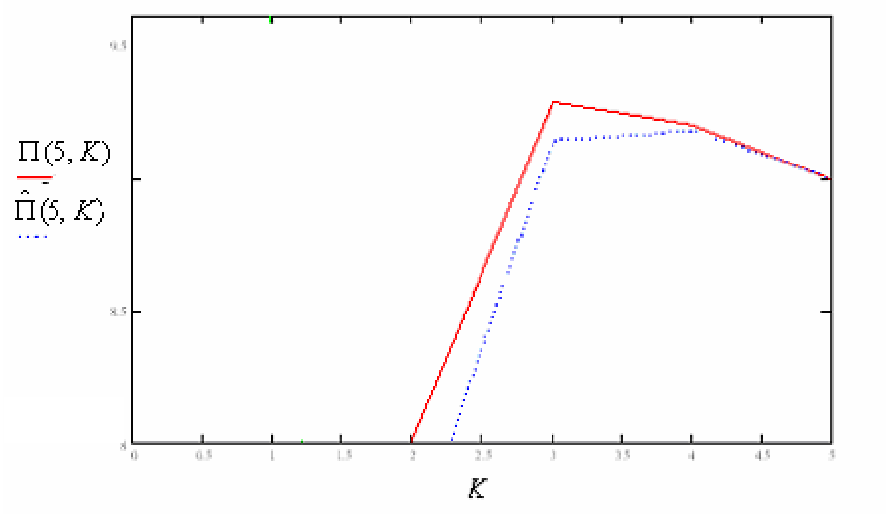

Рис. 5 Оптимальное значение среднего дохода в зависимости от начального запаса *K*

Выводы

Ресамплинг-подход может быть успешно использован для получения оценок интересующих нас параметров процессов восстановления. Полученные формулы позволяют подсчитать дисперсию оценок ресамплинг и традиционных подходов. Численные примеры показывают эффективность предлагаемого подхода, используя как критерий эффективности среднеквадратическую ошибку. Всё вышесказанное позволяет говорить, что предлагаемый подход может служить хорошей альтернативой традиционному и его можно применять для управления запасами высоконадёжных и дорогостоящих изделий.

Заключение

1. Работа посвящена вопросам совершенствования организации и управления логистическими процессами транспорта на основе применения интенсивных компьютерных методов статистики. Её актуальность обусловлена тем, что адекватное описание любого процесса является необходимым условием его эффективной организации и управления.
2. Ресамплинг, являющийся основным из интенсивных компьютерных методов статистики, по сравнению с традиционными методами имеет следующие преимущества: он является непараметрическим, в связи с чем он свободен от ошибок в выборе видов распределений случайных величин; он базируется на прямой имитации рассматриваемого процесса, в связи с чем применим в ситуациях, когда аналитическое описание и анализ процесса невозможны.
3. Практическое применение ресамплинга отличается исключительной простой, так что основная цель работы состояла в исследовании его эффективности в задачах, типичных для транспорта и логистики. Это и составляет теоретическую часть работы. А именно, проанализирована эффективности ресамплинг-подхода по сравнению с традиционными подходами в следующих задачах математической статистики:
    - робастность оценок для параметров многомерной линейной регрессии;
    - оценивание показателей эффективности бесконечно-линейной системы массового обслуживания, широко применяемой в надёжности и страховании;
    - сравнение двух различных процессов восстановления.

   Аналитические исследования и экспериментальные данные говорят о том, что ресамплинг-подход является более предпочтительным при малых выборках.
4. Целесообразность применения ресамплинг-подхода проиллюстрирована на следующих практических задачах логистики и транспорта:
    - прогнозирование спроса на авиационный транспорт в странах Европейского Союза, как функции таких факторов,

как территории, численность населения, валовой внутренний продукт на душу населения, средняя месячная заработная плата одного жителя.

- прогнозирование надёжности изделий, подверженных накоплению и развитию повреждений;
- расчёт оптимального первоначального уровня запасов в условиях случайного расходования и пополнения изделий.

5. Полученные теоретические результаты являются новыми и не встречались ранее в литературе. Методика и результаты прогнозирования спроса на авиаперевозки пассажиров была запрошена для включения в Отчёт по Второму проекту министерства Образования и Науки: Разработка и оценивание математических методов для прогнозирования пассажирских и грузовых потоков в Балтийском регионе (2. Izglītības un zinātnes ministrijas projektu: Matemātisko metožu izstrādāšana un novērtēšana pasažieru un krāvu plūsmu prognozēšanai Baltijas reģionā*)*.

6. Результаты работы докладывались на девяти Международных научных конференциях и семинарах, были опубликованы в десяти статьях. В четырёх из них соискатель является единственным автором.

Публикации с участием автора

1. Afanasyeva H. *Statistical Analysis of Air Traffic in Latvian Region* //In proceedings of the Second International Conference Simulation, Gaming, Training and Business Process Reengineering in Operations. -Riga: RTU, 2000, pp. 125-129.
2. Afanasyeva H. *The Resampling-estimator of Queuing Length Nonstationary Distribution for the Queuing System M/G/∞.* //Transport and Telecommunication, Vol. 3. N.1, - Riga: TTI, 2002, pp. 89-94.
3. Андронов А.М., Афанасьева Е.Н. *Компьютеризация преподавания дисциплин прикладной математики*, Transport and Telecommunication, - Riga, TTI, 2004, pp. 5-8.
4. Andronov A., Afanasyeva H. *Resampling Based Non-Parametric Statistical Inferences about Distribution, Moments and Quantiles of Order Statistics*. //In transactions of *The XX1V International Seminar on Stability Problems for Stochastic Models,* -Riga: TTI, 2004, pp. 300-307.
5. Afanasyeva H. *Resampling Median Estimators for Linear Regression Model*. //In Proceedings of RelStat'04 Transport and Telecommunication, -Riga: TTI, 2005, pp. 90-94.
6. Afanasyeva H., Andronov A. *On robustness of resampling estimators for linear regression models.* //In proceedings of the International Symposium on Stochastic Models in Reliability, Safety and Logistics, Beer Sheva, Israel, 2005, pp. 6-11.
7. Afanasyeva H. *Resampling-approach to a task of comparison of two renewal processes*. //In proceedings of the 12th International Conference on Analytical and Stochastic Modelling Techniques and Applications, -Riga: RTU, 2005, pp. 94-100.
8. Afanasyeva H. *A task of the storage control theory in transport systems using resampling-method.* //In proceedings

of the 5-th International Conference “Transport Systems Telematics”, Katowice-Ustron, Poland, 2005, pp. 13-21.

9. Andronov A., Afanasyeva H. and Fioshin M. *Statistical Estimation for a Failure Model with the Accumulation of Damages*. //In proceedings of the international Conference on Degradation, Damage, Fatigue and Accelerated Life Models in Reliability Testing, Angers, France, 2006, pp. 75-81.
10. Afanasyeva H., Andronov A. *On robustness of resampling estimators for linear regression models.* //Communications in Dependability and Quality Management: An international Journal, Vol. 9, N. 1, 2006, pp. 5-11.