Информационно-вычислительная технология с применением языка «R» в рамках второй ступени высшего образования в медицинских вузах

А. В. Копыцкий,
магистр естественных наук,
Гродненский государственный медицинский университет

Большинство исследований в области медицины не обходится без статистической обработки данных. Адекватное применение статистических методов во многом определяет успешность выполнения исследовательской работы. Но не только правильно выбранный метод влияет на конечный результат научной работы. Одной из проблем в медицинских научных исследованиях является ограниченность объемов выборок, для которых собирается статистика, при большом числе изучаемых показателей.

Согласно образовательному стандарту второй ступени высшего образования, магистранты помимо осуществления учебной деятельности также принимают участие в научно-исследовательской, научно-педагогической, научно-производственной работе. Результатом выполнения научной работы магистранта является его магистерская диссертация.

Особенностью научных работ, выполняемых учащимися второй ступени на базах кафедр медицинских вузов, является то, что зачастую эти работы подразумевают проведение исследования не на одном объекте, а на группе объектов – выборке. В дальнейшем полученные результаты измерений необходимо обработать, используя методы статистического анализа.

В зависимости от профилизации и специальности магистранта он может изучать эти методы в различном объеме в рамках соответствующих дисциплин, таких, например, как «Основы информационных технологий» (для всех специальностей), «Биоинформационный анализ биологических и медицинских данных» (для специальностей 1-31 80 11 «Биохимия» и 1-31 80 12 «Микробиология»), «Введение в программирование на языке “R”» (специальность 1-31 80 01 «Биология») и т. п. То, как магистрант усвоил приемы обработки статистических данных, по сути, проверяется практически тогда, когда магистрант пишет свою диссертацию. Таким образом, можно считать, что организация исследовательской работы магистрантов во многом сопряжена с образовательным процессом в области статистической обработки результатов измерений.

Особенностью исследований в области биологии и медицины является наличие одновременно большого числа параметров, характеризующих один объект. Например, у пациента одновременно измеряются показатели общего и биохимического анализа крови, сатурация, давление, температура, рост, вес и т. д. Но при этом общее число исследуемых объектов невелико, что приводит к ситуации, когда число показателей сопоставимо с объемом выборки.

Обычно исследователя интересуют связи между показателями, их взаимное влияние и интерпретация этих связей. При установлении влияния отдельного

показателя (предиктора) на исследуемый признак-отклик (например, концентрации препарата на гормон, секретируемый организмом) исследователь не испытывает трудностей, однако они появляются, если объектом интереса выступает влияние сразу нескольких показателей на признак, например, не только концентрации препарата на гормон, но и возраста, пола, тяжести заболевания, сопутствующих заболеваний и т. д. В рамках статистического анализа данная проблема решается методами множественной регрессии, и магистрантов учат решать подобного рода задачи.

Так, в учебной программе дисциплины «Биоинформационный анализ биологических и медицинских данных» (специальности 1-31 80 11 «Биохимия»), изучаемой на второй ступени высшего образования в Гродненском государственном медицинском университете, есть разделы «Методы корреляционного и регрессионного анализа биологических и медицинских данных» и «Многомерные методы и модели в биоинформационном анализе биологических и медицинских данных», в которых изучаются статистические модели множественной и обобщенной линейной регрессий.

Традиционно при объяснении этих тем рассматриваются такие методы отбора наилучшего множества предикторов, как пошаговое включение или исключение. Однако эти методы не работают в случаях, когда число предикторов сопоставимо с числом объектов (наблюдений). В этом случае обычно рекомендуется уменьшить число независимых переменных. Для этого можно или провести так называемое уменьшение размерности (используя метод главных компонент или факторный анализ с возможным вращением компонент), или выполнить фильтрацию предикторов (например, с помощью фильтра Борута), или выделить только те предикторы, которые по отдельности статистически значимо связаны с откликом, или оставить только те из них, которые логически связаны с зависимой переменной, или, наконец, вручную перебрать все возможные сочетания предикторов. Все эти методы требуют от магистранта значительных временных затрат на знакомство с ними, освоение и их применение. Кроме этого, при отбрасывании переменных часть информации о связях будет утрачена.

Таким образом, применение традиционных методик изучения классических статистических подходов к регрессии в данной ситуации не представляется возможным. Навыки работы с известными компьютерными программами Statistica и SPSS, которые врачи получают в вузе, не позволяют находить оптимальные модели и использовать преимущества статистического моделирования: обобщать знания, находить и объяснять связи между явлениями, прогнозировать значения показателей. Поэтому для биомедицинских исследований мы предлагаем на начальном этапе рассмотреть методы построения моделей множественной регрессии для выборок ограниченного объема и внедрить метод прямого автоматизированного перебора в методику преподавания посредством информационно-вычислительной технологии.

На рис. 1 представлен алгоритм выбора метода подбора оптимальной модели в регрессионном анализе, приводящий к нашей информационно-вычислительной технологии в случае малых выборок.

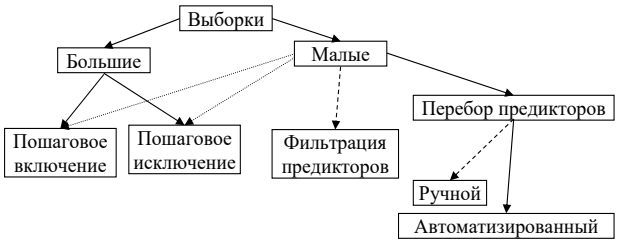

Рис. 1. **Алгоритм выбора метода подбора модели множественной регрессии (сплошная линия – оптимальная, точечная – нереализуемая, штриховая – неоптимальная)**

На начальном этапе освоения теоретического материала при регрессионном анализе исследователь должен понимать, что при выборе модели важным является не только то, насколько хороши ее математические свойства, но и какова ее интерпретация, насколько логичными являются выявленные связи. Поэтому при объяснении тем, связанных с корреляционным и регрессионным анализом, внимание магистрантов заостряется на причинах наличия статистически значимых корреляционных и ассоциативных связей между показателями. Исследуемые показатели могут изменяться из-за существования между ними причинно-следственной связи; два показателя могут зависеть от одного предиктора, явно не измеренного в эксперименте; и наконец, связь между величинами может носить случайный характер.

Таким образом, с нашей точки зрения, оптимальным для учебного процесса при подготовке магистрантов медицинских вузов в контексте проведения ими статистической обработки результатов исследований на ограниченных выборках является программное решение, позволяющее анализировать совокупное воздействие предикторов на некоторый показатель (отклик). Это же решение будет актуальным и на лекционных занятиях, и при проведении практических занятий в рамках изложения материала, описывающего построение моделей множественной регрессии, для демонстрации того, как меняются математические характеристики модели по мере ее усложнения (включения новых предикторов).

Нами предлагается программное решение, реализованное на языке программирования «R», реализующее прямой перебор моделей обобщенной линейной регрессии, а также моделей выживаемости. Данное решение направлено на достижение двух целей:

1) уменьшение временных затрат магистрантов при проведении множественного регрессионного анализа результатов измерений, полученных в рамках их научного диссертационного исследования;

2) наглядная демонстрация подходов в построении регрессионных моделей в условиях часто встречающихся ситуаций ограниченности объемов выборок.

Идея прямого перебора моделей не нова. Принципиально новым является предлагаемый нами подход к реализации метода перебора моделей и внедрения его в учебный процесс.

Анализ научных статей на эту тему показал, что уже существуют некоторые решения и алгоритмы перебора: пакет расширения «leaps» (США, Новая Зеландия) [1] языка программирования «R» [2]; решение, предложенное И. М. Митасовым и А. Н. Завьялкиным (Россия) [3]; решение, приводимое С. Э. Мастицким и В. К. Шитиковым (Беларусь, Россия)) [4].

Решения для обобщенных линейных моделей, имеющиеся в пакетах расширения «bestglm» (Канада) [5] и «glmulti» (Франция) [6], имеют ряд существенных ограничений при малых объемах выборок. Популярные коммерческие статистические пакеты «Statistica» (США) [7] и «SPSS» (США) [8] не имеют готовых имплементированных решений поиска оптимальной модели при ограниченном объеме выборки. Кроме того, пакет «Statistica» не допускает пропущенных значений при пошаговом включении или исключении предикторов. Также в перечисленных программах и пакетах расширений не решается проблема мультиколлинеарности (зависимости предикторов друг от друга).

Таким образом, существующие на сегодняшний день решения данной задачи эффективно работают только в случае, когда число предикторов m значительно меньше объема n наблюдений, т. е. $m << n$.

Помимо линейных и нелинейных моделей для анализа связей между величинами, на наш взгляд, также нецелесообразно использовать машинное обучение. Такие методы машинного обучения, как «деревья» или «леса» классификации, метод «ближайших соседей», нейронные сети и т. д. [9], являются довольно эффективными при работе с большими объемами выборок, когда счет наблюдений идет на тысячи, десятки тысяч и более. А при малых объемах и при большом числе предикторов их точность сопоставима с точностью методов прямого перебора. Для биологических и медицинских данных эти методы сильно проигрывают по такому важному качеству, как объяснимость модели, т. е. возможность логично интерпретировать полученные результаты с теоретической точки зрения. Модели же с линейным предиктором такой возможностью обладают, поэтому и являются одними из лучших [10].

«Ядром» информационно-вычислительной технологии является разработанное программное решение, обеспечивающее прямой автоматизированный перебор всех возможных регрессионных моделей, которые могут быть получены для данного набора предикторов при данной целевой переменной-отклике. Универсальные модели, описывающие широкий круг зависимостей, могут быть получены в рамках обобщенной линейной регрессии, включающей в себя линейную, логит-, пробит-регрессии и т. д. Так как перебор всех возможных комбинаций предикторов является экстенсивным («жадным») алгоритмом, то количество возможных комбинаций может быть существенно уменьшено, если исходить из того, что число предикторов в модели не должно быть слишком велико. Для этого у нас есть следующие основания:

1. Модель должна быть относительно простой (т. е. должна быть легко интерпретируемой).

2. Большинство оценок качества подгонки моделей, исходя из первого основания, вводят штраф за избыточное количество предикторов в модели. Чаще всего в медицинских и биологических исследованиях число предикторов в модели ограничено 8–10 предикторами, поэтому время перебора моделей может быть относительно небольшим: от нескольких минут до нескольких дней, что намного меньше времени, затрачиваемого на сбор и подготовку медицинской и биологической информации.

В качестве языка программирования нашего решения был выбран язык «R», специализированный на статистических расчетах, имеющий готовые стандартные функции для построения моделей регрессии – функции «lm» и «glm» из стандартной библиотеки «stats». Структура программы представлена на рис. 2.

Как видно из рис. 2, в разработанной нами программе можно выделить пять модулей. Рассмотрим подробнее каждый из них.

Модуль 1 «Ввод данных». В этом модуле обеспечивается ввод данных пользователя. Подразумевается, что это электронная структурированная таблица, состоящая из m столбцов (из которых максимальное число предикторов m – 1) и n строк (наблюдений).

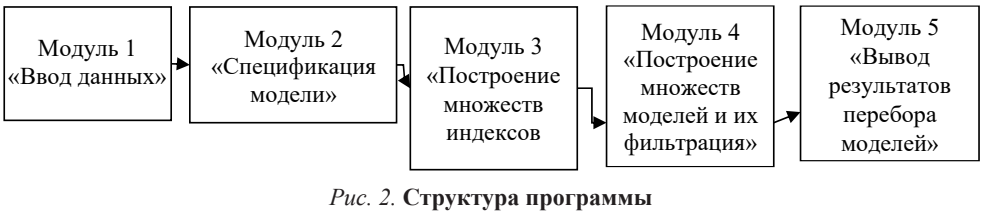

Рис. 2. **Структура программы**

Тут же осуществляется разметка данных: определяется их тип – числовые или факторные переменные.

Модуль 2 «Спецификация модели». В этом модуле пользователь определяет специфику модели: указывает индекс или имя переменной-отклика; указывает, какие переменные будут использоваться как потенциальные предикторы модели; выбирает тип модели – линейная или обобщенная линейная, модель выживаемости, модель мультиномиальной регрессии; здесь же накладываются ограничения на модели.

Модуль 3 «Построение множеств индексов». В этом модуле перебираются все возможные сочетания из k предикторов, отобранных из множества m потенциальных предикторов, выбранных в модуле 2. При необходимости пользователь может сузить полученный набор матриц, указав в модуле 2, какие предикторы обязательно должны входить в итоговый набор.

Модуль 4 «Построение множеств моделей и их фильтрация». По сути, это ядро программы, где анализируются все модели, заданные в модуле 2, с индексами предикторов, полученными в модуле 3. Здесь же при необходимости происходит фильтрация моделей, из которых оставляются только те, что удовлетворяют критериям, описанным в модуле 2.

Модуль 5 «Вывод результатов перебора моделей». По окончании работы модуля 4 результаты перебора должны быть выведены пользователю в этом модуле. В простейшем случае характеристики полученных моделей будут выводиться в итоговый файл в виде электронной таблицы.

Одним из важных методических аспектов, на наш взгляд, является то, что помимо перебора моделей наша программа проводит и их фильтрацию, отбрасывая модели, содержащие предикторы, имеющие статистически значимые корреляционные или ассоциативные связи.

Для определения факта наличия связи между переменными выполняются следующие действия:

1) определяется тип пары предикторов. Возможны следующие пары: «фактор – фактор», «ковариата – фактор», «ковариата – ковариата»;

2) в зависимости от сочетания предикторов определяется показатель связи.

Так, для пары «фактор – фактор» строится таблица сопряженности, для которой индикатором связи выступают χ^2-статистика или статистика точного теста Фишера.

Для определения наличия связи в паре «ковариата – фактор» могут быть использованы как параметрические подходы (критерий Уэлча или однофакторный дисперсионный анализ), так и непараметрические критерии (Манна – Уитни или Краскела – Уоллиса).

Для пары «ковариата – ковариата» мерами связи могут быть значения коэффициентов корреляции (Пирсона, Спирмена или Кендалла).

Вместо статистик критериев, проверяющих гипотезы об отсутствии связи между переменными в генеральной совокупности, можно использовать так называемые размеры эффектов и их максимальное значение, превышение которого можно рассматривать как наличие связи. Для нашего решения это бисериальный и рангово-бисериальный коэффициенты корреляции, коэффициент детерминации, коэффициент ε^2, сами коэффициенты корреляции.

Благодаря данной фильтрации удается решить параллельно две задачи: исключить модели, содержащие пары связанных переменных, и одновременно снизить количество возможных моделей, что приводит к повышению скорости расчетов.

Готовая программа может быть упакована в пакет расширения языка «R» и, таким образом, устанавливаться на любом ПК с данным языком.

Результат работы программы возвращается в виде электронной таблицы. Каждая строка таблицы соответствует конкретной модели и ее характеристикам. В строке содержится информация о предикторах и отклике, об объеме выборки, коэффициенте детерминации, информационном критерии Акаике, девиансе и остаточном девиансе модели. По умолчанию данная таблица отсортирована по убыванию коэффициента детерминации.

Нами также намечены пути оптимизации производительности программного продукта, который применим в рамках информационно-вычислительной технологии. Вместе с этим частично опробованы методы параллелизации перебора и предварительно установлено, что на двух физических ядрах не удается добиться существенного повышения производительности, однако на трех и более физических ядрах скорость перебора повышается на 40–60 %.

Разработан черновой вариант модуля перебора множеств моделей на удаленном ПК, который через локальную сеть соединен с головным компьютером. Задача последнего – разбиение множества комбинаций предикторов на подмножества, запись индексов этих подмножеств в таблицу с заданиями для удаленных компьютеров, агрегирование результатов перебора удаленными компьютерами. Такой подход позволяет использовать технологию параллельных вычислений, при которой несколько десятков сравнительно маломощных ПК, соединенных в одну сеть, значительно сокращают время решения задачи.

При испытаниях данного метода нами получено сокращение времени решения конкретного задания: перебора множеств моделей, имеющих максимум 5 предикторов при общем числе предикторов 41 и при объеме выборки в 120 испытуемых. С 14 часов на одном ПК с частотой процессора 4 ГГц и 4 GB оперативной памяти время решения сетью из 10 ПК (с частотами 1,5 ГГц и оперативной памятью 2 GB каждый) уменьшилось до двух часов. Перспективным, на наш взгляд, является

использование вычислительных кластеров или суперкомпьютеров для ускорения перебора.

Предлагаемая технология, основной частью которой является программа, может быть использована как для методического сопровождения занятия, связанного с моделями множественной регрессии и их интерпретацией, так и для самостоятельного решения с целью изучения связей между откликом и набором предикторов.

Программа была апробирована в 2020 г. в Гродненском государственном медицинском университете на занятиях по дисциплине «Биоинформационный анализ биологических и медицинских данных» второй ступени высшего образования специальности 1-31 80 11 «Биохимия». Магистранты на реальном примере увидели, каким образом меняются характеристики статистических моделей регрессии при добавлении предикторов. Им были показаны ограничения, возникающие на малых объемах выборок в программах «Statistica» и «SPSS» при попытках построения моделей; продемонстрировано, что методы прямого перебора, несмотря на временные издержки, позволяют получить логично интерпретируемые связи.

Несколькими годами ранее прототип нашей программы был аналогичным образом продемонстрирован слушателям курса «Математическая статистика в медицинских исследованиях» повышения квалификации в Гродненском государственном медицинском университете, вызвав интерес у некоторых из них, ранее сталкивавшихся с проблемой ограниченности объемов выборок и невозможностью выбора наилучшего множества предикторов. Наша программа смогла решить эти проблемы, что позволило получить модели, которые легли в основу методов прогнозирования результатов лечения пациентов.

Таким образом, нами разработаны информационно-вычислительная технология и алгоритм ее применения в образовательном процессе второй ступени высшего образования. Основой предлагаемой технологии является программа, написанная на языке программирования «R», позволяющая определять характеристики обобщенных линейных моделей с различными функциями связи. Программа проста в использовании и не требует дополнительных навыков от обучаемых. Основная задача, которая ставится перед магистрантом по завершении работы программы, – интерпретация полученных в электронной таблице результатов перебора моделей.

Список использованных источников

1. *Miller, T. L.* Regression Subset Selection: leaps [Electronic resource] / T. L. Miller based on Fortran code by A. Miller. – Mode of access: https://CRAN.R-project.org/package=leaps. – Date of access: 27.11.2020.

2. R Core Team. R: A Language and Environment for Statistical Computing: R [Electronic resource]. – Mode of access: https://www.r-project.org/about.html. – Date of access: 01.05.2020.

3. *Митасов, И. М.* Метод полного перебора в задаче многофакторного регрессионного анализа / И. М. Митасов, А. Н. Завьялкин // Вестник Сибирского государственного аэрокосмического университета имени академика М. Ф. Решетнева. – 2009. – Т. 1, № 1. – С. 19–22.

4. *Мастицкий, С. Э.* Статистический анализ и визуализация данных с помощью R / С. Э. Мастицкий, В. К. Шитиков. – М.: ДМК Пресс, 2015. – 496 с.

5. Yuanhao, Lai. Best Subset GLM and Regression Utilities: bestglm [Electronic resource] / A. I. McLeod, Changjiang Xu, Yuanhao Lai. – Mode of access: https://CRAN.R-project.org/package=bestglm. – Date of access: 27.11.2020.

6. *Calcagno, V*. Model Selection and Multimodel Inference Made Easy: glmulti [Electronic resource] / V. Calcagno. – Mode of access: https://CRAN.R-project.org/package=glmulti. – Date of access: 27.11.2020.

7. STATISTICA: Data Mining, анализ данных, контроль качества, прогнозирование, обучение, консалтинг [Электронный ресурс]. – Режим доступа: http://statsoft.ru/. – Дата доступа: 27.11.2020.

8. SPSS Software IBM [Electronic resource]. – Mode of access: https://www.ibm.com/analytics/spss-statistics-software. – Date of access: 27.11.2019.

9. *Гелиг, А. Х.* Введение в математическую теорию обучаемых распознающих систем и нейронных сетей. Прикладная математика и информатика: учеб. пособие / А. Х. Гелиг, А. С. Матвеев. – СПб.: Изд-во СПбГУ, 2014. – 224 с.

10. Открытый курс машинного обучения. Тема 4. Линейные модели классификации и регрессии [Электронный ресурс]. – Режим доступа: https://habr.com/ru/company/ods/blog/323890/. – Дата доступа: 27.11.2020.

Аннотация

В статье описана информационно-вычислительная технология построения статистических моделей прямым перебором для обработки результатов биомедицинских исследований, основой которой является разработанная на языке программирования «R» программа. Приведен алгоритм применения технологии как в образовательном процессе в рамках второй ступени высшего образования в медицинских вузах при рассмотрении тем, связанных с множественной регрессией, так и в качестве самостоятельного инструмента.

Abstract

The paper describes an information-computational technology for constructing statistical models by direct search for processing the results of biomedical research, the basis of which is a program developed in the programming language “R”. An algorithm for the application of technology both in the educational process within the framework of the second stage of higher education in medical universities when considering topics related to multiple regression, and as an independent tool is presented.