Введение

В студенческие годы индуктивная статистика давалась мне, как и многим другим студентам, с большим трудом. Здесь не идет речь об элементарных понятиях статистики — средних, медианах или доверительных интервалах. Я имею в виду вещи, с которыми редко приходится сталкиваться за пределами учебной аудитории, такие как рандомизация выборок и дисперсионный анализ.

Я ненавидел их. Я их не понимал. Лекторы и учебники засыпали нас градом формул, в которых мы видели мало смысла. Мы должны были применять эти формулы к данным, но постижение смысла результатов с загадочными названиями наподобие "внутригрупповая дисперсия" было для нас непосильной задачей. Нередко казалось, будто формулы никак не связаны с понятиями, для получения количественной оценки которых они предназначаются. Уже позднее я понял, что это были так называемые "вычислительные формулы", позволяющие ускорить вычисления и уменьшить погрешности по сравнению с интуитивно понятными формулами, определяющими соответствующие величины.

В конечном итоге ко мне постепенно пришло понимание того, почему для вычисления различий между средними используется дисперсионный анализ, или ANOVA, но все эти межгрупповые и внутригрупповые дисперсии и степени свободы оставались для меня загадкой. Я знал, что от меня требовалось вычислять их, и знал, как это делается, но не понимал главного: зачем это нужно.

Как-то у меня в руках оказалась одна книга по регрессионному анализу. Ее порекомендовал другой студент — в ней он нашел объяснение многих вопросов, которые до этого смущали его и по-прежнему смущали меня. В той книге, которая давно уже не переиздавалась, обсуждение дисперсионного и ковариационного анализа велось в терминах регрессионного анализа. Когда это имело смысл, автор приводил

результаты компьютерных расчетов. Основной упор в книге делался на коэффициенты корреляции и общую дисперсию. Главным образом именно с их помощью демонстрировалась взаимосвязь традиционного анализа Фишера и регрессионного подхода.

Суть концепций начала проясняться для меня, и я осознал, что всегда наблюдал их, просто они скрывались за таинственными формулами дисперсионного анализа. Соответствующие формулы приводились преподавателями и использовались для расчетов, поскольку они были разработаны еще в начале XX века, когда современные вычислительные мощности были не просто дефицитным ресурсом, а их тогда вообще не существовало. В то время было гораздо легче рассчитывать суммы квадратов (особенно посредством вычислительных формул), чем такие элементы регрессионного анализа, как коэффициенты множественной корреляции и квадраты получастных корреляций. При таком подходе не приходилось находить обратные матрицы в рамках традиционного дисперсионного анализа, как это иногда требуется в регрессии. (Вычисление вручную обратных матриц размером более чем 3×3 превращается в сущий кошмар.)

В настоящее время все эти возможности заложены в рабочих листах Excel, что делает понятия, лежащие в основе дисперсионного анализа, гораздо более прозрачными для пользователя. Кроме того, Excel значительно облегчает выполнение любых вычислений, о чем в книге, которую я читал, ничего не было сказано. Та книга была написана задолго до выхода первых версий Excel, и даже сама мысль о том, что однажды мне понадобится обратная матрица и я буду вынужден каким-то способом выбирать ее элементы, чтобы получить результат, меня ужасала. Сегодня те же результаты можно легко получить, просто комбинируя надлежащим образом фиксированные и относительные адреса ячеек.

Дисперсионный и ковариационный анализ все еще очень востребованы в самых разных областях, число которых постоянно растет: это и медицина, и фармакология, финансовый анализ, эконометрика, сельскохозяйственные эксперименты, исследование операций и т.п. Но во всех этих случаях глубокое понимание концептуальных основ применяемых методик играет важнейшую роль, и я придерживаюсь той точки зрения, что при рассмотрении задач сквозь призму регрессии, а не традиционного дисперсионного анализа, такое понимание приходит намного скорее.

Однако, по моему мнению, еще важнее то, что понимание основных концепций регрессионного анализа существенно облегчает изучение более сложных методов, таких как логистическая регрессия или факторный анализ. Эти методики позволяют включать в анализ одновременно несколько зависимых переменных. Они также обеспечивают возможность проведения исследований в областях, включающих скрытые, ненаблюдаемые переменные и переменные, подчиняющиеся полиномиальным зависимостям. Для тех, кто предварительно не ознакомился с понятием объясненной дисперсии, кривая изучения метода главных компонент будет гораздо более крутой.

Именно поэтому я и принял решение написать данную книгу. Работая в области индуктивной статистики сначала исполнителем, а затем независимым консультантом, я накопил богатый практический опыт, позволяющий обоснованно судить о том, насколько мощным инструментом она может быть, если правильно ее применять. Я использую Excel вот уже более двадцати лет. Некоторые не относятся серьезно к Excel как к приложению для численного анализа. Я думаю, что они неправы. С другой стороны, история компании Microsoft как издателя Excel, скажем так, довольно пестрая. Недавно коллега

переслал мне электронное письмо, в котором его автор интересовался, с заметными нотками беспокойства, "безопасно" ли использовать статистические функции Excel. К тому времени я уже почти закончил писать эту книгу, значительная часть которой была посвящена применению функции Excel ЛИНЕЙН(). Вот что я написал в ответ:

> Вопрос о том, "безопасно" ли использовать Excel для статистического анализа, довольно каверзный. Частично вина лежит на Microsoft, а те, кто громогласно заявляет об опасности применения Excel в статистическом анализе, разделяют эту вину.
>
> С 1995 года компания Microsoft не сделала ровным счетом ничего для улучшения надстройки Анализ данных (также известной под названием Пакет анализа), разве что переписала ее, использовав VBA вместо устаревшего языка макросов V4. Это стыдно, поскольку с данной надстройкой связано множество проблем, которые можно было бы легко устранить. Впрочем, нельзя ведь и отождествлять саму надстройку с Excel. Я видел немало статей, опубликованных как в блогах, так и в авторитетных журналах, в которых справедливо критиковались статистические инструменты этой надстройки, и все они возлагали вину на само приложение.
>
> Примерно в 2003 или 2007 году (не помню точно, когда) проявились две главные проблемы, связанные с функцией ЛИНЕЙН(). Одна из них касалась способа расчета регрессионной и остаточной сумм квадратов в условиях, когда третий аргумент функции ЛИНЕЙН(), *конст*, задается равным ЛОЖЬ. Об этом было известно еще в 1995 году, но компания Microsoft исправила эту ошибку с огромным опозданием.
>
> Суть другой проблемы, связанной с функцией ЛИНЕЙН(), заключается в том, что так называемые "нормальные уравнения" решаются в ней с использованием матричной алгебры, которая издавна применялась в статистических приложениях в качестве предпочтительного метода. Но изредка возникают ситуации, когда мультиколлинеарность (наличие сильной корреляции между предикторными переменными) приводит к матрице с нулевым определителем. Для такой матрицы не существует обратной матрицы, что делает невозможным возврат функцией ЛИНЕЙН() обычных результатов. В 2003 или 2007 году компания Microsoft внесла соответствующее исправление, использовав так называемое *QR-разложение*.
>
> Теперь при наличии коллинеарности функция ЛИНЕЙН() возвращала код ошибки #ЧИСЛО!, но зато вычисления не могли пойти по неправильному пути. Что касается проблемы, связанной с третьим аргументом, *конст*, то она, например, приводила к таким результатам, как отрицательные значения R^2. Корректные вычисления не могут давать такие значения, и это служило признаком того, что где-то что-то не так.
>
> Наконец, всевозможные прочие статистические функции Excel, используемые в повседневной работе, были значительно усовершенствованы в плане повышения точности при работе с по-настоящему экстремальными значениями. Этот шаг был чрезвычайно полезным — бо́льшая точность всегда лучше меньшей точности. Но это пример того, что Фрейд назвал "нарциссизмом малых различий". Если бы я занимался статистическими исследованиями в области

биологии и меня попросили принять решение, основываясь на различии между величинами 10^{-16} и 10^{-17}, то я обязательно повторил бы эксперимент. Как с технической точки зрения, так и по существу, различие между этими величинами слишком мало, чтобы на него можно было опираться при принятии важных решений, какое бы приложение для этого ни использовалось: SAS, R или Excel.

И здесь я вновь возвращаюсь к тем паникерам, которые пугают людей своим пессимизмом, звучащим в их рассуждениях на эту тему. Политика опорочивания вполне нормальных статистических приложений имеет свою давнюю бесчестную историю. Еще когда я учился в колледже, другие студенты, которые в поте лица изучали приложение под названием BMD, доказывали, что использовать иные приложения — плохая идея. Они аргументировали это неточностью конкурирующих приложений, но их истинным мотивом было желание не допустить разрушения их ореола незаменимых специалистов. (Тем другим, конкурирующим приложением было SPSS.)

Если уж и говорить о возможных опасностях использования Excel для статистического анализа, то речь может идти исключительно об опасности применения точного инструмента теми, кто не отдает себе отчета в том, что они делают, независимо от того, касается это индуктивной статистики или Excel.

На этом я заканчиваю длинное вступление, написанное в 2016 году. Надеюсь, чтение книги доставит вам такое же удовольствие, какое мне доставляют встречи с давними друзьями.

Сайт книги

Файлы примеров книги (в том числе русифицированные) доступны на сайте издательства по следующему адресу:

`http://www.williamspublishing.com/Books/978-5-9908462-7-2.html`

Авторские файлы примеров доступны на сайте издательства Que Publishing:

`http://www.quepublishing.com/store/regression-analysis-microsoft-excel-9780789756558`

Файлы Excel с расширением `*.xlsm` содержат макросы VBA. Для работы с ними нужно либо разрешить использование макросов при открытии файла, либо сохранить файлы в надежном расположении.

Ждем ваших отзывов!

Вы, читатель этой книги, и есть главный ее критик. Мы ценим ваше мнение и хотим знать, что было сделано нами правильно, что можно было сделать лучше и что еще вы хотели бы увидеть изданным нами. Нам интересны любые ваши замечания в наш адрес.

Мы ждем ваших комментариев и надеемся на них. Вы можете прислать нам бумажное или электронное письмо либо просто посетить наш сайт и оставить свои замечания там. Одним словом, любым удобным для вас способом дайте нам знать, нравится ли вам эта книга, а также выскажите свое мнение о том, как сделать наши книги более интересными для вас.

Отправляя письмо или сообщение, не забудьте указать название книги и ее авторов, а также свой обратный адрес. Мы внимательно ознакомимся с вашим мнением и обязательно учтем его при отборе и подготовке к изданию новых книг.

Наши электронные адреса:

E-mail: `info@dialektika.com`
WWW: `http://www.dialektika.com`

Наши почтовые адреса:

в России: 195027, Санкт-Петербург, Магнитогорская ул., д. 30, ящик 116
в Украине: 03150, Киев, а/я 152