Корреляционный и регрессионный анализ в использовании факторного планирования экспериментов

В. Д. Поляков
Финансовый университет при Правительстве Российской Федерации (Финуниверситет), Financial University
polyakov-vd@rambler.ru

***Аннотация*. При имитационном моделировании важным этапом является экспериментальное получение информации о свойствах моделируемой системы. К таким свойствам относятся значения средних, дисперсий, корреляционных функций параметров системы. При нахождении значений параметров необходим статистический анализ. Сюда входят задачи определения характера зависимостей, характеризующих данные из моделируемой системы, оценка значений параметров этих зависимостей, нужна также проверка значимости параметров.**

Ключевые слова: корреляция; анализ; регрессия; модель; эксперимент..

Как правило, требуется определение вида распределения случайных величин, для чего применяются методы математической статистики, которые основаны на вычислении параметров экспериментальных распределений. При этом осуществляется проверка статистических гипотез с использованием критериев согласия, и с помощью механизма принятия гипотез проверяют соответствие эмпирических данных известным законам распределения. При этом задается статистически приемлемый доверительный уровень принятия гипотез.

На этапе получения информации о свойствах моделируемой системы может понадобиться проверка наличия резко выделяющихся значений с целью их исключения из дальнейшего рассмотрения, определение ошибок и доверительных интервалов числовых характеристик. Может понадобиться определение доверительного объема измерений, а также сравнение дисперсий и средних значений для данных, полученных при разных условиях. В разных случаях используют различные специфические совокупности и соответствующие критерии: Фишера (F), Стьюдента (t), Бартлетта (B), Кочрена (G). В ряде случаев, когда закон распределения параметра отличается от нормального закона, используют различные непараметрические критерии.

На заключительных этапах имитационного моделирования проводят планирование имитационного компьютерного эксперимента, называемого направленным [1]. При направленном машинном эксперименте на имитационной модели требуется выбрать для применения конкретные аналитические методы обработки результатов исследования на модели. Чаще всего применяются дисперсионный, корреляционный, регрессионный методы обработки. Очень ценными являются методы факторного планирования эксперимента, а также методы оптимизации.

При компьютерных экспериментах на имитационных моделях иногда используют имитационные методы обработки информации, например – численные методы решения задач путем моделирования случайных величин. По результатам машинного эксперимента осуществляется анализ результатов моделирования и принятие решений.

I. Корреляционный и регрессионный анализ

При решении экономических задач анализа и прогнозирования часто используются статистические данные, иногда – отчетные, а иногда – экспериментальные данные, полученные из практики функционирования объектов. Чаще всего эти данные являются реализациями случайных величин.

Случайной величиной называется такая величина, которая случайным образом принимает с некоторой вероятностью различные значения. Основная характеристика случайной величины – закон распределения, который показывает частоту значений в общей совокупности ее значений. Когда исследуют взаимосвязи между экономическими показателями с использованием данных статистики, между ними часто встречают случайную (стохастическую) зависимость. Случайность зависимости имеет место из-за того, что изменение параметров закона распределения одной случайной величины происходит из-за изменения параметров другой. Причем на одну случайную совокупность может влиять не только одна какая-то другая, а несколько совокупностей случайных величин [1].

При обработке информации в процессе изучения зависимостей между совокупностями исходных данных различают:

- корреляционный анализ;
- регрессионный анализ.

Корреляционным анализом называется раздел математической статистики, занимающийся изучением взаимосвязей между случайными величинами. Главная задача корреляционного анализа это выявление характера

и степени связи между зависимыми (выходными) и независимыми (факторными) показателями в данном объекте или процессе. Корреляционная связь обнаруживается лишь при массовом изучении массивов зависимых и независимых показателей. В технических исследованиях для получения исходных данных с целью проведения корреляционного анализа используется понятие «пассивный эксперимент».

Существенная черта корреляционных зависимостей – случайный характер как входных, так и выходных величин. Вид и характер связи между величинами наглядно, хотя и приблизительно, виден по корреляционному полю. Корреляционное поле – график, образуемый точками с координатами X_i и Y_i (сопряженными парами), когда эти значения, полученные из пассивного эксперимента, регистрируются в виде пар. По расположению точек можно заключить следующее: чем гуще и кучнее располагаются точки, тем теснее корреляционная связь. Численно теснота связи определяется с помощью рассчитываемого по специальной формуле коэффициента корреляции. Этот показатель лежит в пределах от -1 до +1. Когда величина коэффициента корреляции находится в пределах от 1 до 0,7 по абсолютному значению, то корреляционная зависимость считается сильной. Если величина коэффициента корреляции находится в границах от 0,7 до 0,3, то считается, что корреляционная зависимость – средняя. Когда же значение коэффициента корреляции лежит в пределах от -0,3 до 0,3, то корреляционная зависимость считается слабой. При значении коэффициента корреляции, равном нулю или очень близким к нулю, то говорят о полном отсутствии корреляции. Корреляционный анализ предназначен для определения характера и тесноты связи между случайными массивами.

Целью регрессионного анализа служит нахождение уравнения регрессии, а также в это понятие включается статистическая, вероятностная, оценка его параметров. Уравнение регрессии есть модель, связывающая зависимую переменную, с независимой переменной или с несколькими независимыми переменными.

На графике благодаря анализу множества точек (т.е. множества статистических данных), ищется линия, которая достаточно точно отражает имеющуюся в этих данных закономерность, или тенденцию. Это – линия регрессии.

По количеству факторов различают однофакторные и многофакторные регрессионные зависимости.

По характеру связи однофакторные регрессионные зависимости бывают:

а) линейные: $y = a + bx$;

где x – независимая переменная, y – зависимая, выходная переменная; a, b – параметры, или коэффициенты регрессии;

б) степенные: $y = ax^b$;

в) показательные: $y = ab^x$;

г) прочие (логарифмические, гиперболические). Причем следует заметить, что на практике применяют методы линеаризации, позволяющие свести нелинейные модели к линейным. Например, зависимости типа б) и в) линеаризуются логарифмированием.

Наиболее ценным можно считать многофакторный регрессионный анализ, а именно факторное планирование эксперимента.

II. Использование факторного планирования экспериментов.

При планировании и построении активных экспериментов, включая и модельные, мы имеем дело с двумя типами переменных, которые называются соответственно факторами x и откликами y. Факторы называют независимыми переменными, или управляемыми переменными, а отклики – выходными, или зависимыми переменными.

Планирование эксперимента при имитационном моделировании, как и другие проблемы планирования, требует, во-первых, выбора плана эксперимента, для чего необходимо:

- выбрать выходной параметр (отклик), количество варьируемых факторов, количество уровней варьирования, требуемое количество измерений выходной переменной или количество повторностей при проведении опытов;
- составить экспериментальную регрессионную модель;
- сравнить полученную модель с известными моделями, со стандартными планами и выбрать наилучший план.

Процедуру получения плана эксперимента можно разбить на три этапа: построение структурной модели; построение функциональной модели; построение экспериментальной модели.

Структурная модель отличается: количеством факторов; числом уровней варьирования факторов.

В зависимости от целей эксперимента выбирается структурная модель, причем выбор параметров структурной модели определяется точностью измерений факторов, необходимостью учитывать нелинейные эффекты и т. п. Структурная модель эксперимента имеет следующий вид:

$$N_s = f(k, q_i)$$

где N_s – число элементов эксперимента, другими словами, количество строк в матрице планирования; k – количество факторов в эксперименте; q_i количество уровней факторов, где i=1, 2,..., k.

В функциональной модели определяется число компонентов структурной модели, которые будут работать при построении отклика. Функциональные модели бывают

совершенными, или несовершенными. Функциональная модель называется совершенной, если в генерировании выходного параметра участвуют все ее элементы, т. е. $Nf = Ns$. В идеальном случае структурная модель совпадает с функциональной, однако в имитационном машинном эксперименте часто бывает ограничение на ресурсы, что может привести к выбору несовершенной функциональной модели. Следовательно, требуется реализовать компромисс между существующими ресурсами и пожеланиями исследователя: $N = pq^k$, где: p – число повторностей в эксперименте; q – количество уровней факторов; k – число факторов.

С учетом ограничений на ресурсы нужно определить q, k, p.

Первая задача – выбор переменной отклика, или другими словами, целевой функции, или выходного параметра. Этот выбор зависит от цели исследования.

К параметру оптимизации предъявляются следующие требования: он должен быть эффективным при достижении цели эксперимента; должен быть достаточно универсальным; должен быть количественным; должен быть эффективным по статистическим меркам, т.е. точным; должен иметь физический смысл и быть легко вычисляемым; должен существовать при различных условиях проведения эксперимента.

Вторая задача: выделение существенных факторов.

После выбора выходных параметров, следует определить факторы, которые оказывают заметное влияние на эти переменные. Количество таких факторов может быть большим, поэтому надо найти самые существенные. К сожалению, чем меньше информации об исследуемой системе, тем больше можно найти факторов, которые, вроде бы, способны влиять на выходной параметр. Облегчает эту задачу при имитационном моделировании такая предварительная процедура, как анализ чувствительности имитационной модели к факторам [1].

После задания переменных отклика и определения существенных факторов необходимо эти факторы классифицировать по их отношению к машинному эксперименту. Факторы могут участвовать в эксперименте в трех видах: быть постоянным; быть переменным, но неуправляемым; быть измеряемыми и управляемыми. Для вхождения в план эксперимента нужны именно факторы измеряемые и управляемые.

Основные требования к факторам: управляемость, что нужно для проведения активного эксперимента, и однозначность.

Существуют важные требования к ансамблю факторов:

- задаваемая группа факторов должна быть достаточно полным множеством;
- должна быть высокой точность фиксации факторов;
- необходимы совместимость факторов друг с другом и требуется отсутствие линейной корреляции между факторами;
- факторы должны устанавливаться независимо от уровней остальных факторов.

Следующий шаг при разработке плана эксперимента – определение уровней варьирования. Более просто для математической обработки – использовать для всех факторов одинаковое число уровней. От выбора уровней зависит точность результатов. Также надо иметь в виду, следует ли учесть нелинейные эффекты. Минимальное количество уровней факторов – два. При двухуровневом варьировании можно получить линейную регрессионную модель или неполную квадратическую модель, включающую эффекты взаимодействия типа $b_{ij}x_ix_j$. Для получения полной квадратической модели число уровней должно быть 3 – 5. Число строк в матрице планирования вычисляется так: $N = q^k$.

Уровни могут быть: качественные или количественные; фиксированные или случайные. Чаще используются количественные уровни факторов, значения которых могут быть измерены с использование некоторой шкалы. Также обычно применяют фиксированные значения уровней, т.е. их выбирают заранее. Обработка данных упрощается, если сделать уровни равноотстоящими от середины интервала, границами которых они являются. Такое расположение обеспечивает ортогональность плана и этим упрощает определение коэффициентов полинома. Т. е. две наиболее удаленные точки на оси в области изменения количественной переменной, интересующей исследователя, устанавливают как два крайних уровня, и остальные уровни помещают внутри этого диапазона. Расстояние от центра варьируемого фактора на числовой оси до крайнего уровня называется интервалом варьирования. Минимальное значение интервала варьирования можно найти в предварительном эксперименте, определив значения выходного параметра y при крайних уровнях фактора x_i, с последующей проверкой значимости различий между этими значениями $|\bar{Y}_1 - \bar{Y}_2|$ с помощью критерия Стьюдента:

$$t_R = \frac{|\bar{Y}_1 - \bar{Y}_2|}{S\{|\bar{Y}_1 - \bar{Y}_2|\}}$$

где $S\{|\bar{Y}_1 - \bar{Y}_2|\}$ – среднее квадратическое отклонение разности $|\bar{Y}_1 - \bar{Y}_2|$; t_R – расчетное значение критерия Стьюдента, которое надо сравнить с табличным значением, взятым при доверительной вероятности $P_Д$ = 0,95 и числом степеней свободы f, равным числу степеней свободы дисперсии $S^2|\bar{Y}_1 - \bar{Y}_2|$.

Вероятностные модели не требуют дополнительной интерпретации, поскольку они основаны на аксиоматической теории, а именно теории вероятностей. Следовательно, использование байесовского подхода в данном контексте также отражено необходимостью машинного переобучения и использования новых данных для пересчета вероятности появления риска с

увеличением точности показателя при увеличении объема данных.

Лучшее распределение понимается как взвешенное среднее между знанием о параметрах до того, как будут наблюдаться данные (которые представлены априорным распределением), и информация о параметрах, содержащихся в наблюдаемых данных (представленных функцией правдоподобия). После того, как было получено апостериорное распределение, возможно вычислить точечные и интервальные оценки параметров, а также получить прогнозный вывод для вероятностной оценки гипотез. Как только модель данных (модель вероятности) выбрана, байесовский анализ требует утверждения априорного распределения для неизвестных параметров модели. Априорное распределение можно рассматривать как представляющее текущее состояние знаний или текущее описание неопределенности о ранее указанных параметрах модели к наблюдаемым данным. Подходы к выбору априорного распределения делятся на две основные категории.

Первый подход включает в себя выбор информативного предварительного распределения. Статистик использует свои знания о существенной проблеме возможно, основываясь на других данных, а также, если это возможно, привлекая экспертное мнение, чтобы построить предварительное распределение, которое должным образом отражает его (и экспертные) убеждения о неизвестных параметрах. Понятие информативного предварительного распределения может показаться, прежде всего, чрезмерно субъективным и ненаучным. Также, следует отметить, что выбор модели данных также является субъективным выбором.

Второй основной подход к выбору априорного распределения состоит в том, чтобы построить неинформативное предварительное распределение. Помимо неинформативного, этот тип распределения также называется объективным, расплывчатым и диффузным. Выбор неинформативного предварительного распределения является попыткой объективности посредством такого действия, будто перед наблюдением не существовало определенных предварительных знаний о параметрах. Это реализуется путем присвоения равной вероятности всем значениям параметра (или, по крайней мере, примерно равной вероятности по локализованным диапазонам параметра).

Даже несмотря на доказанную в прикладном применении повышенную точность использования байесовских методов для выражения вероятности реализации той или иной гипотезы, необходимо учитывать объем первоначальной информации и возможную субъективность изначального подхода к оценке ситуации. Это должно непосредственно учитываться в оценке риска при новом производстве, когда данных о предыдущих неисправностях не предоставлено и возможное априорное распределение базируется на экспертной оценке.

В связи с развитием ЭВМ сфера использования имитации в экономике увеличилась. Она действенна как для решения задач внутри предприятий, так и для моделирования макроэкономических процессов.

Статистический анализ может помочь конкретизировать выводы имитационного эксперимента, необходимый для построения прогнозных моделей и сценариев. Моделирование – это построение математической модели, требующее наличие четкого представления о целях функционирования исследуемой экономической системы и информации об ограничениях, определяющие область допустимых значений данных переменных. Анализ такой модели может помочь найти способ наилучшего воздействия на объект управления при выполнении всех условий.

Сложность таких систем очень затрудняет определение целей и ограничений в аналитическом виде. Несмотря на то, что существует большое количество переменных и ограничений, на которые стоит обратить внимание при анализе ситуаций, только малая их часть требуется при описании исследуемых систем. Следовательно, для моделирования систем нужно определить приоритетные ограничения, переменные и параметры.

СПИСОК ЛИТЕРАТУРЫ

[1] Звягин Л.С. Применение экспертного прогнозирования в системно-аналитических задачах// Экономика и управление: проблемы, решения. 2017. Т. 4. № 1. С. 86-93.

[2] Звягин Л.С. Применение системно-аналитических методов в области экспертного прогнозирования// Экономика и управление: проблемы, решения. 2017. Т. 3. № 9. С. 47-50.

[3] Лычкина Н.Н. Имитационное моделирование экономических процессов: уч. пос. / Н.Н. Лычкина. М.: Академия Ай-Ти, 2005. 164 с.

[4] Мичасова О.В. Имитационное моделирование экономических систем: уч.-метод. пос. / О.В. Мичасова. Нижний Новгород: Нижегородский госуниверситет, 2014. 186 с.

Спиридонов А.А. Планирование эксперимента при исследовании технологических процессов / А.А. Спиридонов. М.: Машиностроение, 1981. 184 с.