УДК 65.012.12:331

Применение статистических методов для априорной обработки информации

Н.Н. Елизарова, А.А. Кузнецова
ФГБОУВПО «Ивановский государственный энергетический университет имени В.И. Ленина», Иваново, Россия
E-mail: elisarova@it.ispu.ru, anyagrunge@mail.ru

Авторское резюме

Состояние вопроса: Статистические методы нашли широкое применение в практической деятельности предприятий, организаций. Применение априорной обработки информации позволяет уменьшить число наблюдений для построения моделей.
Материалы и методы: Использованы корреляционный анализ и метод определения информативности независимых признаков, основанный на сравнении вероятностных характеристик.
Результаты: Предложена методика априорной обработки информации. Рассмотрен метод информативности для отбора входных признаков. Приведен пример реализации методики.
Выводы: Методика априорной обработки направлена не только на отбор значимых переменных, но и на проверку связей между входными переменными, что позволяет уменьшить число требуемых наблюдений.

Ключевые слова: статистические методы, регрессионная модель, результирующий показатель, независимый фактор, математическое ожидание, дисперсия, корреляция, значимость коэффициентов, информативность признаков.

Application of Statistical Methods for Prior Information Processing

N.N. Elizarova, A.A. Kuznetsova
Ivanovo State Power Engineering University, Ivanovo, Russian Federation
E-mail: elisarova@it.ispu.ru, anyagrunge@mail.ru

Abstract

Background: Statistical methods are widely used. The usage of a priori processing allows reducing a number of observations to design models.
Materials and methods: The author used the correlation analysis. The article also considers the informative determination method of independent features, based on the comparison of the probability characteristics.
Results: The author suggests the method of prior processing based on the correlation analysis. For the selection of input features the informative determination method of independent features is considered. As an example, the implementation of the methodology is discussed.
Conclusions: The method of prior processing is aimed not only at selecting the relevant variables, but also at testing the links between input variables. It allows to decrease a number of the required observations.

Key words: statistical methods, regression model, final figure, independent factor, mathematical expectation, variance, correlation, significance of coefficients, informational signs.

Особенностью развития современного общества является переход на рыночные условия хозяйствования, при которых увеличивается динамика показателей экономической деятельности. Это требует существенных изменений в подходах в планировании и ведении хозяйственной деятельности любой организации, предприятия. В этих условиях использование математических методов анализа экономической деятельности приобретает первостепенное значение.

Независимо от уровня экономического развития статистические методы на протяжении многих лет всегда были инструментом управления. Выполняя функции сбора, систематизации и группировки сведений, характеризующих экономическое развитие общества, обработки данных, прогнозирования поведения в будущий период, статистические методы всегда играли ведущую роль в выявлении фактов, закономерностей для нужд управления, научных исследований.

Главной задачей управления является анализ состояния развития предприятий, фирм и предоставление руководству достоверной информации о внешнем окружении (конкурентах, рынке сбыта), о работе предприятия, фирмы. Возрастает потребность в экономической, статистической, социально-демографической информации. Это требует организации информационно-аналитических отделов, центров на предприятии. Необходимость таких центров обусловлена следующим:

1) в условиях рыночной экономики необходимо оперативно предоставлять комплексную, полную, достоверную информацию, используемую при формировании как долгосрочных, так и краткосрочных планов развития предприятий;

2) информация для руководства должна быть предоставлена в виде, удобном для восприятия, с проработкой вариантов стратегических решений и их оценкой;

3) анализ влияния внешних и внутренних факторов на принимаемые решения проводится на основе полученной информации.

Таким образом, должен быть сформирован информационный ресурс, который позволил бы сэкономить другие виды ресурсов (энергетические, трудовые, финансовые).

Статистические методы нашли широкое применение в практической деятельности предприятий, организаций, фирм. Их достоинство состоит в том, что сведения накапливаются в базах данных организации, отчетных документах. Таким образом, имеем дело с пассивными экспериментами (вторичными исследованиями), не требующими много времени и материальных затрат на сбор и подготовку информации. Это дает возможность не только формализованно представить проблему, но и предложить возможные варианты решений, прогнозировать поведение объекта управления в будущем.

При необходимости можно получить данные, проведя специальные эксперименты, наблюдения при фиксированных условиях, значениях входных факторов. Это активный эксперимент, или первичные исследования. Чаще всего это проведение социологических опросов для выявления мнений по какой-либо проблеме. Например, исследование удовлетворенностью оказания каких-либо услуг.

Использование статистических методов позволяет не только накопить данные, но и отследить взаимосвязи (корреляционный анализ), установить тесноту взаимосвязи (регрессионный анализ) и спрогнозировать поведение в будущем.

В работе [1] рассмотрены статистические пакеты, которые могут быть использованы для обработки статистической информации на предприятии. Остановимся на предварительных исследованиях данных, направленных на снижение размерности моделей.

Первым этапом любых исследований является сбор статистических данных, либо путем организации специальных испытаний, либо путем использования уже накопленных сведений в базах данных или отчетных документах.

Вторым этапом является выявление зависимостей, построение моделей. Однако использование предварительной обработки в целях выявления взаимосвязи факторов позволит снизить размерность модели и уменьшить число наблюдений.

Например, многофакторный регрессионный анализ служит для отыскания количественной зависимости между результирующим показателем (откликом) Y и входными, независимыми факторами (переменными, признаками) $X = \{x_1, x_2, \ldots, x_k\}$, которые оказывают влияние на Y. Требуется установить зависимость отклика Y от факторов X, т.е. определить зависимость $y = f\{x_1, x_2, \ldots, x_k\}$ при условии независимости факторов X.

В общем виде уравнение регрессии можно представить следующим образом:

$$\tilde{y} = b_0 + \sum_{i=1}^{k} b_i x_i + \sum_{i,j(i\neq j)}^{k} b_{ij} x_i x_j + \sum_{i=1}^{k} b_{ii} x_i^2 + \ldots \qquad (1)$$

Пусть имеется n результатов наблюдений над величиной Y, зависящей от k факторов, причем степень полинома равна d. Тогда число коэффициентов регрессии определяется как число сочетаний C_{k+d}^{d} и должно быть меньше числа опытов n [2].

Для снижения размерности можно использовать корреляционный анализ, с помощью которого решаются две задачи:

1) определяются факторы x_i, оказывающие влияние на исследуемый процесс Y;

2) выявляется наличие корреляции между независимыми переменными X.

Методика априорной обработки включает следующие этапы:

1. Решая первую задачу, вычисляем коэффициент корреляции между Y и x_i для $i = \overline{1,k}$:

$$r_{yx_i} = \frac{\sum_{i=1}^{n}(y_j - \bar{y})(x_{li} - \bar{x}_l)}{\sqrt{\sum_{i=1}^{n}(y_j - \bar{y})^2 \sum_{i=1}^{n}(x_{li} - \bar{x}_l)^2}}, \qquad (2)$$

где $\bar{y}$ – оценка математического ожидания результирующего показателя Y; $\bar{x}_l$ – оценка математического ожидания входного фактора x_l.

2. Проверяем значимость коэффициента r_{yxl}, используя критерий Стьюдента [2]. При проверке гипотезы о том, что коэффициент корреляции равен нулю, вычисляется t-статистика:

$$t = \sqrt{\frac{r_{yx_l}^2 (n-2)}{1 - r_{yx_l}^2}}. \qquad (3)$$

Задавшись уровнем значимости q (вероятность ошибки), для степеней свободы $\nu = n - 2$ по таблицам распределения Стьюдента находим значение $t_{q,\nu}$ [2]. Расчетное значение t сравнивается с табличным значением $t_{q,\nu}$. Если $t < t_{q,\nu}$, это свидетельствует о незначимости коэффициента корреляции, а следовательно, отсутствии связи между Y и x_l. Такие переменные следует исключить из перечня входных факторов. В итоге получаем скорректированный вектор $X = \{x_1, x_2, \ldots, x_k\}$, где $k' \le k$.

3. Среди оставшихся переменных X проводим проверку на независимость их друг от друга, так как нет необходимости включать в уравнения факторы, объясняющие друг друга.

Для проверки независимости входных факторов вычисляем парные коэффициенты корреляции между факторами x_i и x_j ($i = \overline{1,k'}$; $j = \overline{2,k'}$) по формуле

$$r_{ij} = \frac{\sum_{g=1}^{n}(x_{ig} - \overline{x}_i)(x_{jg} - \overline{x}_j)}{\sqrt{\sum_{g=1}^{n}(x_{ig} - \overline{x}_i)^2 \sum_{g=1}^{n}(x_{jg} - \overline{x}_j)^2}}. \quad (4)$$

4. Аналогично п.2 проверяем значимость каждого коэффициента корреляции. Если $t \geq t_{q,\nu}$, то факторы x_i и x_j признаются связанными и нет необходимости включать в модель оба фактора.

5. В случае обнаружения коррелированных факторов проводится анализ их информативности.

Для решения этой задачи можно использовать следующие методы [3]:

– метод, основанный на сравнении апостериорных вероятностей;

– метод, основанный на сравнении вероятностных характеристик;

– метод, основанный на определении количества информации.

Для анализа информативности признаков рассмотрим метод, основанный на сравнении вероятностных характеристик. Метод может быть применен и в том случае, когда законы распределений $f_i(x)$ неизвестны, но при этом известны математические ожидания m_{ij} и дисперсии D_{ij}. Оценка, основанная на использовании этих данных, возможна, так как признаки объектов x_i могут быть условно подразделены на две группы:

1) признаки, значения которых мало изменяются при переходе от одного объекта данного класса к другому и заметно изменяются при переходе от объектов одного класса к объектам другого;

2) признаки, которые чувствительны к переходам от объекта к объекту данного класса и лишь незначительно изменяются при переходах от одного класса к другому.

Последовательность расчета информативности признаков следующая:

1. Определяется математическое ожидание признаков:

– средние значения для каждого класса признака по формуле равномерного распределения [4]

$$m_{ij} = \frac{a_{ij} + b_{ij}}{2}, \quad (5)$$

где a_{ij}, b_{ij} – верхняя и нижняя границы i-го класса j-го признака;

– математическое ожидание j-го признака

$$M(x_j) = \sum_{i=1}^{m} m_{ij} P(\Omega_i). \quad (6)$$

2. Рассчитывается математическое ожидание дисперсии признаков:

– дисперсия для каждого класса признака по формуле равномерного распределения

$$D_{ij} = \frac{(b_{ij} - a_{ij})^2}{12}; \quad (7)$$

– математическое ожидание дисперсии для признаков

$$M(D_j) = \sum_{i=1}^{m} D_{ij} P(\Omega_i). \quad (8)$$

3. Определяется дисперсия математического ожидания распределений признаков при переходе от класса к классу по формуле

$$\overline{D}_j = M\{[m_{ij} - M(x_j)]^2\}. \quad (9)$$

4. Вычисляется критерий сравнительной оценки для каждого признака:

$$K_j = \frac{M[D_j]}{\overline{D}_j}. \quad (10)$$

При этом наилучшим признаком является тот, который имеет минимальное значение критерия:

$$K(X^*) = \min_j K_j. \quad (11)$$

Таким образом, следует выбрать наиболее информативный признак.

Рассмотрим методику на примере анализа экономической деятельности предприятий пищевой промышленности в одном из регионов. Для анализа прибыльности предприятий Y (балансовая прибыль предприятия, тыс. руб.) задан набор признаков:

x_1 – электровооруженность труда, кВт·ч/чел.;

x_2 – энерговооруженность труда, кВт·ч/чел.;

x_3 – расход тепловой энергии;

x_4 – фондовооруженность труда, руб/чел.;

x_5 – рентабельность продукции, %;

x_6 – численность работников, чел.;

x_7 – фонд времени без потерь, ч;

x_8 – коэффициент сменности рабочих.

Сформируем группы, в которые входят общие признаки, по экономической сущности исходных переменных следующим образом:

A: x_1, x_2, x_3 – энерготопливные ресурсы предприятия;

B: x_4, x_5, x_6 – результативность (эффективность) работы предприятия;

С: x_7, x_8 – наличие и использование трудовых ресурсов предприятия.

Для десяти предприятий (n = 10) получены количественные величины этих параметров (табл. 1). Для построения линейной регрессионной модели (d = 1) при k = 8 необходимо как минимум $C_9^1 = 9$ опытов. Для построения полиномиальной модели второй степени – $C_{10}^2 = 45$ опытов. Таким образом, в случае неадекватности линейной модели десять наблюдений будет недостаточно для выявления зависимости $Y = f(x)$.

1) Вычислим парные коэффициенты корреляции по формуле (2) и критерий Стьюдента (3) (табл. 2).

Таблица 1. **Исходные данные**

№ п/п	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	878	2,58	178	12	1,37	0,49	57	919	1
2	1595	2,65	301	16	1,58	1,44	213	3433	1,32
3	1067	2,1	315	14	1,36	0,51	158	934	1,43
4	1303	2,14	362	11	1,66	0,61	136	3159	1,42
5	1017	1,31	280	12	0,98	0,76	124	1998	2,07
6	711	3,53	204	8	1,4	0,49	52	838	1,37
7	1350	1,4	325	17	1,71	1,21	115	1854	1
8	712	3,59	302	7	1,14	0,53	53	854	1
9	696	0,87	147	9	1,28	0,46	48	774	1
10	812	3,03	258	12	0,96	0,64	65	1048	1,44

Таблица 2. **Коэффициенты корреляции и критерий Стьюдента**

	r_{yx1}	r_{yx2}	r_{yx3}	r_{yx4}	r_{yx5}	r_{yx6}	r_{yx7}	r_{yx8}
r_{yxi}	-0,249	0,663	0,817	0,664	0,837	0,900	0,883	0,140
t	0,726	2,504	4,012	2,511	4,320	5,827	5,309	0,400

2) Для уровня значимости q = 0,05 и числа степеней свободы $\nu = n - 2 = 8$ по статистическим таблицам находим критическое значение $t_{q,\nu}$ = 2,31.

3) Сравним расчетное значение критерия Стьюдента t с критическим для каждого фактора. Согласно данным табл. 2, факторы x_1 и x_8 являются не связанными с y, поэтому нет необходимости включать их в уравнение.

4) Для каждой группы признаков проверим их независимость (табл. 3, 4).

Для группы А: $r_{x_2x_3}$ = 0,423326; t_{23} = 1,322 < $t_{q,\nu}$, следовательно, факторы x_2 и x_3 можно признать независимыми.

Для группы В: в табл. 3 приведены парные коэффициенты корреляции $r_{x_ix_j}$, в табл. 4 – критерии Стьюдента t_{ij}.

Анализ результатов расчетов показывает, что связь между x_5 и x_6 следует признать значимой и нет необходимости включать в модель обе переменные. Необходимо определить какая из этих переменных более информативна.

Таблица 3. **Коэффициенты корреляции**

Признаки	x_4	x_5	x_6
x_4	1		
x_5	0,455171	1	
x_6	0,438848	0,699093	1

Таблица 4. **Критерии Стьюдента**

Признаки	x_4	x_5
x_5	1,446	
x_6	1,381	2,765

5) Проведем сравнение информативности двух признаков x_5 и x_6.

Начальные значения признаков для двух классов приведены в табл. 5, результаты расчета (формулы (5)–(11)) сведены в табл. 6.

Таблица 5. **Граничные значения признаков по классам**

Класс	$P(\Omega_i)$	x_5		x_6	
		нижняя граница	верхняя граница	нижняя граница	верхняя граница
Ω_1 (неоптимальный)	0,4	0	0,6	0	80
Ω_2 (оптимальный)	0,6	0,5	1,5	70	220

Таблица 6. **К расчету сравнительной оценки информативности**

Признаки x_j	Класс Ω_i	m_{ij}	$M(x_j)$	D_{ij}	$M(D_j)$	$\overline{D}_j$	K_j
x_5	Ω_1	0,3	0,72	0,03	0,062	0,1176	0,5272
	Ω_2	1		0,083333			
x_6	Ω_1	40	103	533,33	1338,33	2646	0,5057
	Ω_2	145		1875			

Так как $K_5 > K_6$, то качество признака x_6 выше, чем качество признака x_5. Следовательно, для дальнейшего исследования следует выбрать признак x_6.

Таким образом, для построения модели следует рассматривать следующие признаки:

x_2 – энерговооруженность труда;

x_3 – расход тепловой энергии;

x_4 – фондовооруженность труда;

x_6 – численность работников;

x_7 – фонд времени без потерь.

Для построения линейной регрессионной модели (d = 1) для пяти признаков (k = 5) минимально необходимо $C_5^1 = 5$ опытов. Для построения полиномиальной модели второй степени (d = 2) – $C_{10}^2 = 21$ опыт.

Следовательно, априорная обработка информации помогает существенно снизить объем необходимых исходных данных при построении моделей и уменьшает трудоемкость построения моделей.

Заключение

Рассмотренная методика априорной обработки направлена на использование корреляционного анализа не только для отбора значимых факторов, но и для проверки наличия или отсутствия связей между входными переменными. Это приводит к уменьшению набора исследуемых переменных, следовательно, к уменьшению и числа требуемых наблюдений.

Список литературы

1. **Елизарова Н.Н.** Использование программных средств статистической обработки данных при формировании информационного обеспечения управления // Вестник ИГЭУ. – 2009. – Вып. 3. – С. 76–80.

2. **Белов А.А., Баллод Б.А., Елизарова Н.Н.** Теория вероятностей и математическая статистика: учебник. – Ростов н/Д: Феникс, 2008. – 318 с.

3. **Горелик А.Л., Скрипкин В.А.** Методы распознавания. Изд. 2-е. – М.: Высш. шк., 1989. – 231с.

4. **Айвазян С.А., Мхитарян В.С.** Прикладная статистика и основы эконометрики: учебник для вузов. – М.: ЮНИТИ, 1998. – 1022 с.

References

1. Elizarova, N.N. *Vesnik IGEU*, 2009, issue 3, pp. 76–80.

2. Belov, A.A., Ballod, B.A., Elizarova, N.N. *Teoriya veroyatnostey i matematicheskaya statistika: uchebnik* [Probability Theory and Mathematical Statistics]. Rostov n/D, Feniks, 2008. 318 p.

3. Gorelik, A.L., Skripkin, V.A. *Metody raspoznavaniya* [Recognition Methods]. Moscow, Vysshaya shkola, 1989. 231 p.

4. Ayvazyan, S.A. Mkhitaryan, V.S. *Prikladnaya statistika i osnovy ekonometriki: uchebnik dlya vuzov* [Applied Statistics and Econometrics Fundamentals]. Moscow, YuNITI, 1998. 1022 p.

Елизарова Надежда Николаевна,
ФГБОУВПО «Ивановский государственный энергетический университет имени В.И. Ленина»,
кандидат технических наук, доцент кафедры информационных технологий,
телефон (4932) 26-98-55,
e-mail: elisarova@it.ispu.ru

Кузнецова Анна Адольфовна,
ФГБОУВПО «Ивановский государственный энергетический университет имени В.И.Ленине»,
студент,
телефон (4932) 26-98-55,
e-mail: anyagrunge@mail.ru