

6. а) Что подразумевается под восстановлением значения по регрессионной модели ?
б) Что такое экстраполяция?
7. Соберите данные о средней дневной температуре в вашем городе за последнюю неделю (10 дней, 20 дней). Оцените (хотя бы на глаз), годится ли использование линейного тренда для описания характера изменения температуры со временем. Попробуйте путем графической экстраполяции предсказать температуру через 2–5 дней.
8. Придумайте свои примеры практических задач, для которых имело бы смысл выполнение восстановления значений и экстраполяционных расчетов.



§ 19

Моделирование корреляционных зависимостей

Регрессионные математические модели строятся в тех случаях, когда известно, что зависимость между двумя факторами существует и требуется получить ее математическое описание. А сейчас мы рассмотрим задачи другого рода. Пусть важной характеристикой некоторой сложной системы является фактор A . На него могут оказывать влияние одновременно многие другие факторы: B , C , D и т. д. Мы рассмотрим два типа задач.

- 1) Оказывает ли фактор B какое-либо заметное регулярное влияние на фактор A ?
- 2) Какие из факторов B , C , D и т. д. оказывают наибольшее влияние на фактор A ?

В качестве примера сложной системы будем рассматривать школу. Пусть для первого типа задач фактором A является средняя успеваемость учащихся школы, фактором B — финансовые расходы школы на хозяйственные нужды: ремонт здания, обновление мебели, эстетическое оформление помещения и т. п. Здесь влияние фактора B на фактор A не очевидно. Наверное, гораздо сильнее на успеваемость влияют другие причины: уровень квалификации учителей, контингент учащихся, уровень технических средств обучения и др.

Специалисты по статистике знают, что для того, чтобы выявить зависимость от какого-то определенного фактора, нужно максимально исключить влияние других факторов. Проще говоря,

собирая информацию из разных школ, нужно выбирать такие школы, в которых приблизительно одинаковый контингент учеников, квалификация учителей и пр., но хозяйственные расходы разные (у одних школ могут быть богатые спонсоры, у других — нет).

Итак, пусть хозяйственные расходы школы выражаются количеством рублей, отнесенных к числу учеников в школе (руб./чел.), потраченных за определенный период времени (например, за последние 5 лет). Успеваемость же пусть оценивается средним баллом учеников школы по результатам окончания последнего учебного года. Еще раз обращаем ваше внимание на то, что в статистических расчетах обычно используются относительные и усредненные величины.

Итоги сбора данных по 20 школам, введенные в электронную таблицу, представлены на рис. 3.7. На рис. 3.8 приведена точечная диаграмма, построенная по этим данным.

А	В	С
№ п/п	Затраты (руб./чел.)	Успеваемость (средний балл)
1	50	3,81
2	345	4,13
3	79	4,30
4	100	3,96
5	203	3,87
6	420	4,33
7	210	4
8	137	4,21
9	463	4,4
10	231	3,99
11	134	3,9
12	100	4,07
13	294	4,15
14	396	4,1
15	77	3,76
16	480	4,25
17	450	3,88
18	496	4,50
19	102	4,12
20	150	4,32

Рис. 3.7. Статистические данные

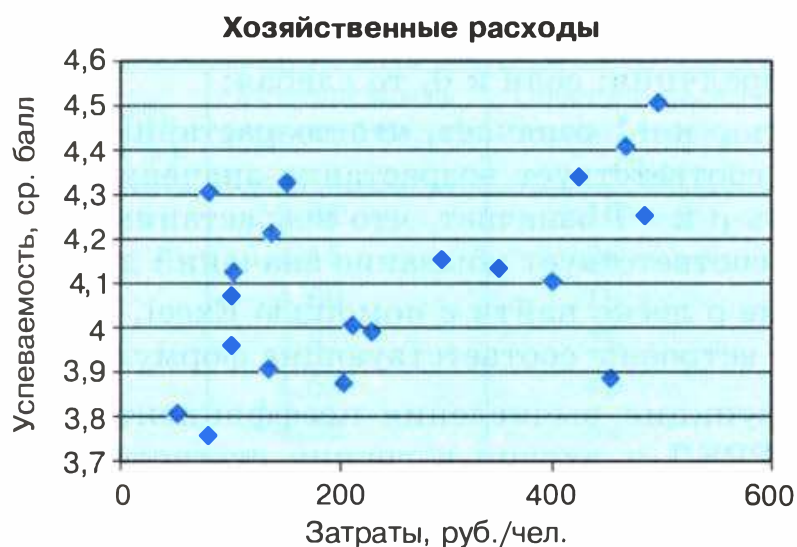


Рис. 3.8. Точечная диаграмма

Значения обеих величин: финансовых затрат и успеваемости учеников — имеют значительный разброс и, на первый взгляд, взаимосвязи между ними не видно. Однако она вполне может существовать.

Зависимости между величинами, каждая из которых подвергается не контролируемому полностью разбросу, называются **корреляционными зависимостями**.

Раздел математической статистики, который исследует такие зависимости, называется **корреляционным анализом**. Корреляционный анализ изучает усредненный закон поведения каждой из величин в зависимости от значений другой величины, а также меру такой зависимости.

Оценку корреляции величин начинают с высказывания гипотезы о возможном характере зависимости между их значениями. Чаще всего допускают наличие линейной зависимости. В таком случае мерой корреляционной зависимости является величина, которая называется **коэффициентом корреляции**. Как и прежде, мы не будем писать формулы, по которым этот коэффициент вычисляется; их написать нетрудно, гораздо труднее понять, почему они именно такие. На данном этапе достаточно знать следующее:

- коэффициент корреляции (обычно обозначаемый греческой буквой ρ) есть число из диапазона от -1 до $+1$;

- если это число по модулю близко к 1, то имеет место сильная корреляция; если к 0, то слабая;
- близость ρ к +1 означает, что возрастанию значений одного набора соответствует возрастание значений другого набора, близость ρ к -1 означает, что возрастанию значений одного набора соответствует убывание значений другого набора;
- значение ρ легко найти с помощью Excel, так как в эту программу встроены соответствующие формулы.

В Excel функция вычисления коэффициента корреляции называется КОРРЕЛ и входит в группу статистических функций. Покажем, как ею воспользоваться. На том же листе Excel, где находится таблица, представленная на рис. 3.7, надо установить курсор на любую свободную ячейку и запустить функцию КОРРЕЛ. Она запросит два диапазона значений. Укажем, соответственно, B2:B21 и C2:C21. После их ввода будет выведен ответ: $\rho = 0,500273843$. Эта величина говорит о среднем уровне корреляции.

Наличие зависимости между хозяйственными затратами школы и успеваемостью нетрудно понять. Ученики с удовольствием ходят в чистую, красивую, уютную школу, чувствуют там себя, как дома, и поэтому лучше учатся.

В следующем примере проводится исследование по определению зависимости успеваемости учащихся старших классов от двух факторов: обеспеченности школьной библиотеки учебниками и оснащения школы компьютерами. И та, и другая характеристика количественно выражается в процентах от нормы. Нормой обеспеченности учебниками является их полный комплект, т. е. такое количество, когда каждому ученику выдаются из библиотеки все нужные ему для учебы книги. Нормой оснащения компьютерами будем считать такое их количество, при котором на каждом из четырех старшеклассников в школе приходится один компьютер. Предполагается, что компьютерами ученики пользуются не только на информатике, но и на других уроках, а также во внеурочное время.

В таблице, изображенной на рис. 3.9, приведены результаты измерения обоих факторов в 11 разных школах. Напомним, что влияние каждого фактора исследуется независимо от других (т. е. влияние других существенных факторов должно быть приблизительно одинаковым).

Обеспечение учебного процесса				
№	Обеспеченность учебниками (%)	Успеваемость (средний балл)	Обеспеченность компьютерами (%)	Успеваемость (средний балл)
1	50	3,81	10	3,98
2	78	4,15	25	4,01
3	94	4,69	19	4,34
4	65	4,37	78	4,41
5	99	4,53	45	3,94
6	87	4,23	32	3,62
7	100	4,73	90	4,6
8	63	3,69	21	4,24
9	79	4,08	34	4,36
10	94	4,2	45	3,99
11	93	4,32	67	4,5
$\rho = 0,780931$			$\rho = 0,572465$	

Рис. 3.9. Сравнение двух корреляционных зависимостей

Для обеих зависимостей получены коэффициенты линейной корреляции. Как видно из таблицы, корреляция между обеспеченностью учебниками и успеваемостью сильнее, чем корреляция между компьютерным обеспечением и успеваемостью (хотя и тот, и другой коэффициенты корреляции не очень большие). Отсюда можно сделать вывод, что пока еще книга остается более значительным источником знаний, чем компьютер.

Система основных понятий



Корреляционные зависимости	
Это зависимости между величинами, каждая из которых подвергается неконтролируемому разбросу	
Корреляционный анализ дает возможность:	
определить, оказывает ли один фактор существенное влияние на другой фактор	выбрать из нескольких факторов наиболее существенный
Коэффициент корреляции ρ : количественная мера корреляции	
ρ по модулю близко к единице — сильная корреляция	ρ близко к нулю — слабая корреляция
Расчет ρ возможен в Microsoft Excel с помощью функции КОРРЕЛ	