Лекция № 16

Корреляционный и регрессионный анализ

Введение

Один из наиболее общих законов объективного мира - закон всеобщей взаимосвязи между явлениями. Естественно, что исследуя явления в самых различных областях, статистика неизбежно сталкивается с зависимостями, как между количественными, так и качественным показателями, признаками. Ее задача - обнаружить (выявить) такие зависимости и дать их количественную характеристику.

Среди взаимосвязанных признаков одни могут рассматриваться как определяющие факторы, влияющие на изменение других, а вторые - как следствие, результат влияния первых. Соответственно, первые, то есть влияющие на изменение других, признаки называют *независимыми (факторными),* а вторые - *зависимыми (результативными).*

Зависимости между признаками могут быть двух видов: функциональные (детерминированные) и статистические (недетерминированную, стохастические).

Зависимость между двумя признаками x и y называется *функциональной,* если определенному значению признака x однозначно соответствует значение признака y. Такие зависимости обычно встречаются в строгих науках - математике, физике и др. Например, известно, что площадь квадрата равна квадрату его стороны $S = a^2$. При увеличении стороны квадрата в 2 раза его площадь увеличится в 4 раза. Функциональные связи можно встретить и в области экономических явлений. Например, при сдельной системе оплаты зависимость между величиной заработка y и количеством изготовленных изделий x, при фиксированной расценке за изделие 5 рублей, легко выразить формулой $y = 5 \cdot x$

Зависимость называется *статистической,* если зависимый признак y

определяется одновременным действием многих факторов и ее значение при фиксированном значении x может варьироваться.

Например, при изучении зависимости урожайности определенной культуры от количества внесенных в поле удобрений урожайность будет являться зависимым признаком, а количество удобрений - независимым. Между ними нет функциональной связи, так как при одном и том же количестве внесенных удобрений урожайность в разных хозяйствах, на разных участках земли будет неодинаковой. Этот очевидный факт объясняется тем, что, помимо удобрений, на урожайность влияет много других факторов - качество семян, густота посева, уход за посевами, своевременность уборки и другие, комбинация которых вызывает вариацию урожайности.

Статистические зависимости можно обнаружить только при массовом наблюдении. Их исследование и составляет предмет корреляционно-регрессионного анализа.

Основными вопросами этого исследования являются:

1) Существует ли зависимость между величинами?

2) Насколько сильной она является?

3) Каков характер этой зависимости - прямой или обратный, линейный или нелинейный?

4) Каково наиболее вероятное значение y при заданном x (прогнозирование)?

Ответ на первые два вопроса дает *корреляционный анализ,* в котором в качестве меры взаимосвязи применяется коэффициент корреляции. Например, известно, что существует много факторов, способствующих возникновению сердечных заболеваний - неподвижный образ жизни, курение, наследственность, возраст, стрессы, питание.

Корреляционный анализ позволяет установить, влияние каких факторов является наиболее значительным.

Если зависимость существует, то коэффициент корреляции позволяет также

установить, является она прямой или обратной.

Прямая зависимость означает, что факторная и результативная переменная возрастают или убывают одновременно.

Например, прямая зависимость существует между ростом и весом человек.

Обратная зависимость существует, в частности, между возрастом и физической силой человека.

Более точное математическое описание зависимости составляет предмет *регрессионного анализа* и осуществляется с помощью построения *регрессионной модели* - функционального соотношения между зависимой и независимой переменной, наилучшим образом описывающего реальную статистическую зависимость. Подобранная функция может быть линейной или нелинейной. В случае, если в рассмотрение включается только один независимый признак, модель называется *парной,* если несколько - то *множественной.* Построенная модель позволит дать ответ и на четвертый вопрос, выполнить прогнозирование. Точность прогноза будет тем выше, чем более сильной является зависимость.

Корреляционный анализ

Диаграмма рассеяния и линейный коэффициент корреляции Пирсона

Простейшим приемом при исследовании зависимости между двумя количественными признаками является построение *диаграммы рассеяния.* Обозначим исследуемые признаки как x и y. Тогда каждая точка на диаграмме рассеяния соответствует одной единице наблюдения, абсцисса точки равна значению признака x, а ордината - признаку y

Пример 1. Построить диаграмму рассеяния для результатов наблюдения за возрастом и артериальным давлением группы людей, табл. 1.

Таблица 4.1. Зависимость возраста и давления

Субъект	Возраст x_i	Давление y_i

1	43	128
2	48	120
3	56	135
4	61	143
5	67	141
6	70	152

Результат представлен на рис. 1.

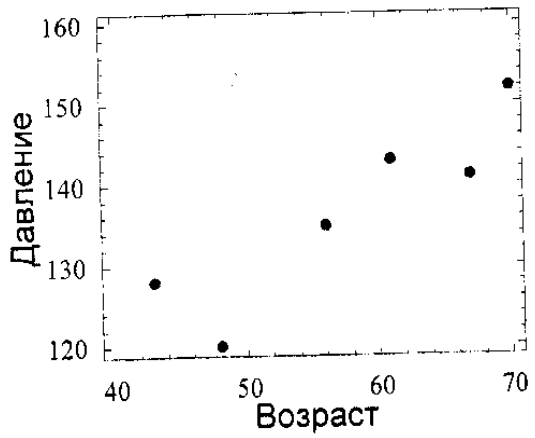


Рис. 4.1. Диаграмма рассеяния для зависимости возраста и давления

Диаграмма рассеяния позволяет получить *качественное* представление о наличии зависимости. Так, в приведенном примере заметно, что с возрастанием факторного признака результативный также возрастает, что указывает на наличие прямой зависимости, а расположение вблизи воображаемой прямой - на зависимость линейного типа.

Наиболее часто употребляемой *количественной* характеристикой линейных зависимостей между признаками является *линейный коэффициент корреляции Пирсона*.

Коэффициент корреляции Пирсона (*r* -Пирсона) применяется для исследования взаимосвязи двух переменных, измеренных в метрических

шкалах на одной и той же выборке. Он позволяет определить, насколько пропорциональная изменчивость двух переменных.

Данный коэффициент разработали Карл Пирсон, Фрэнсис Эджуорт и Рафаэль Уэлдон в 90-х годах XIX века.

Коэффициент корреляции r-Пирсона характеризует существование линейной связи между двумя величинами. Если связь криволинейная (нелинейная), то он не будет работать.

Чтобы приступать к расчетам коэффициента корреляции r-Пирсона необходимо выполнение следующих условий:

1. Исследуемые переменные x и y должны быть распределены нормально.

2. Исследуемые переменные x и y должны быть измерены в интервальной шкале или шкале отношений.

(**Шкала** (от лат. «скале» — лестница) – элемент счетной системы, посредством которого происходит отнесение исследуемого объекта к определенной группе объектов.

Интервальная шкала – количественная шкала. В этой шкале устанавливается единица измерения. В интервальной шкале, например, измеряется температура (по Цельсию или по Фаренгейту).

3. Количество значений в исследуемых переменных x и y должно быть одинаковым.

Слабыми сторонами линейного коэффициента корреляции Пирсона являются:

1. Неустойчивость к выбросам.

2. С помощью коэффициента корреляции Пирсона можно определить только силу линейной взаимосвязи между переменными, другие виды взаимосвязей выявляются методами регрессионного анализа.

Рассмотрим выборку из n единиц, каждая из которых характеризуется значениями признака x_i и y_i. Тогда линейный коэффициент корреляции

определяется следующим образом

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right),$$

где $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ - выборочные средние,

$s_x = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$, $s_x = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$ - выборочные отклонения.

Основные свойства коэффициента корреляции.

1) Изменяется в диапазоне от -1 до +1.

2) Если существует сильная прямая зависимость между признаками, то коэффициент корреляции будет иметь значение вблизи +1.

3) Если существует сильная обратная зависимость между признаками, то коэффициент корреляции будет иметь значение вблизи - 1

4) Если связь между признаками слабая или нелинейная, то коэффициент будет вблизи 0.

Графически диапазон изменений и интерпретация коэффициента корреляции представлена на рис. 2.

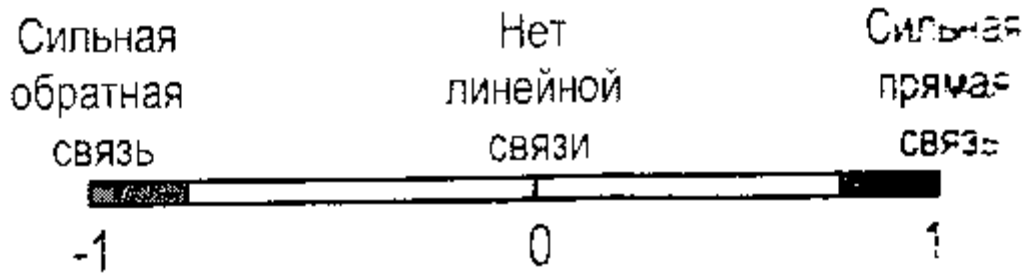
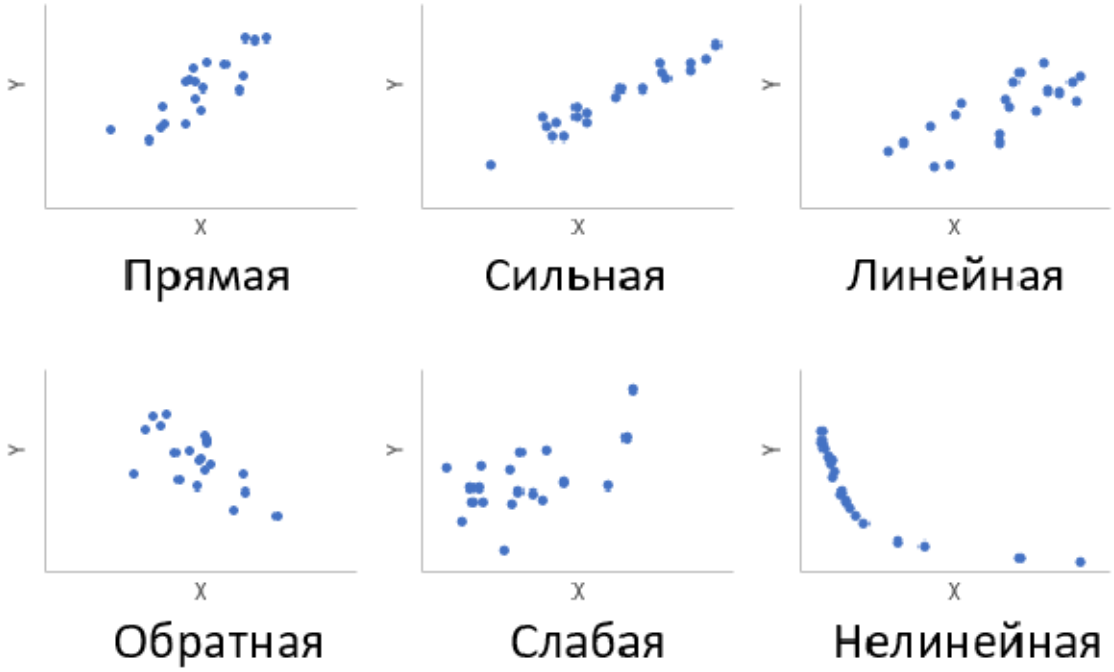


Рис. 2. Диапазон *значений коэффициента корреляции*



Для упрощенного расчета линейного коэффициента корреляции часто применяют преобразованную формулу, позволяющую избежать промежуточного расчета средних и отклонений:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y}) \sum_{i=1}^{n}(y_i - \bar{y})^2}} =$$

В следующих примерах для расчета коэффициента корреляции будет использована первоначальная формула, что позволит лучше понять суть этой величины.

Пример 1а. Вычислить линейный коэффициент корреляции и определить тип зависимости для данных примера 1.

Решение. Исходные данные и результаты расчетов представлены в табл. 2. В третьем и четвертом столбцах таблицы приведены нормированные отклонения от среднего для возраста x и давления y. Следует обратить внимание на знаки этих величин - во всех случаях они совпадают. Если возраст

наблюдаемого субъекта больше среднего, то и его давление оказывается выше среднего. Математически это означает, что произведение отклонений будет положительным и они будут давать положительный вклад в коэффициент корреляции. В конечном итоге значение коэффициента окажется близко к +1, указывая на сильную прямую связь между величинами.

Таблица 2. *Расчет коэффициента корреляции*

Возраст x_i	Давление y_i	$z_{x,i} = \left(x_i - \overline{x}\right)/s_x$	$z_{y,i} = \left(y_i - \overline{y}\right)/s_y$	$z_{x,i} \cdot z_{y,i}$
43	128	-1.50	-0.82	1.22
48	120	-0.98	-1,59	1.56
56	135	-0.16	-0.14	0.02
61	143	0.36	0.62	0.23
67	141	0.98	0.43	0.42
70	152	1.29	1.49	1.92
$\overline{x} = 57.5$ $s_x = 9.65$	$\overline{y} = 136.5$ $s_y = 10.37$			$r = \frac{\sum_{i=1}^{6} z_{x,i} \cdot z_{y,i}}{6} = 0.897$

Пример 2. Построить диаграмму рассеяния, вычислить коэффициент корреляции Пирсона и определить тип зависимости для двух признаков - количество пропущенных занятий и итоговая (суммарная) оценки. Данные наблюдений для группы учащихся представлены в табл. 3, колонки 1 и 2.

Решение. Для каждой строки таблицы отложим на оси абсцисс значение количества пропусков, а по оси ординат - итоговую оценка, рис. 3. Получившийся график указывает на возможную обратную связь - чем больше пропущено занятий, тем ниже итоговая оценка, хотя в области малого числа пропусков это соотношение проявляется слабо. Вычислим коэффициент корреляции, который позволит численно охарактеризовать предполагаемую зависимость. Сравнивая значения нормированных отклонений в колонках 3 и 4,

легко заметить, что в большинстве случаев их знаки противоположны - если число пропусков ниже среднего, то итоговый бал выше, и наоборот. При перемножении двух чисел с разными знаками произведения получается отрицательным, колонка 5. При расчете коэффициента корреляции произведения суммируются и дают отрицательное значение, близкое к -1. Между двумя переменными существует сильная обратная связь.

***Таблица** 3. Расчет коэффициента корреляции для числа пропусков и итоговой оценки*

Число пропусков x_i	Итоговая оценка y_i	$z_{x,i} = \left(x_i - \overline{x}\right)/s_x$	$z_{y,i} = \left(y_i - \overline{y}\right)/s_y$	$z_{x,i} \cdot z_{y,i}$
6	82	-0.53	0.58	-0.31
2	86	-1.52	0.84	-1.27
15	43	1.69	-1.93	-3.27
9	74	0.21	0.06	0.01
12	58	0.95	-0.97	-0.92
5	90	-0.78	1.09	-0.85
8	7*8	-0.04	0.32	-0.01
$\overline{x} = 8.14$ $s_x = 4.05$	$\overline{y} = 73$ $s_y = 15.53$			$r = \frac{\sum_{i=1}^{7} z_{x,i} \cdot z_{y,i}}{7}$ = =-0.944

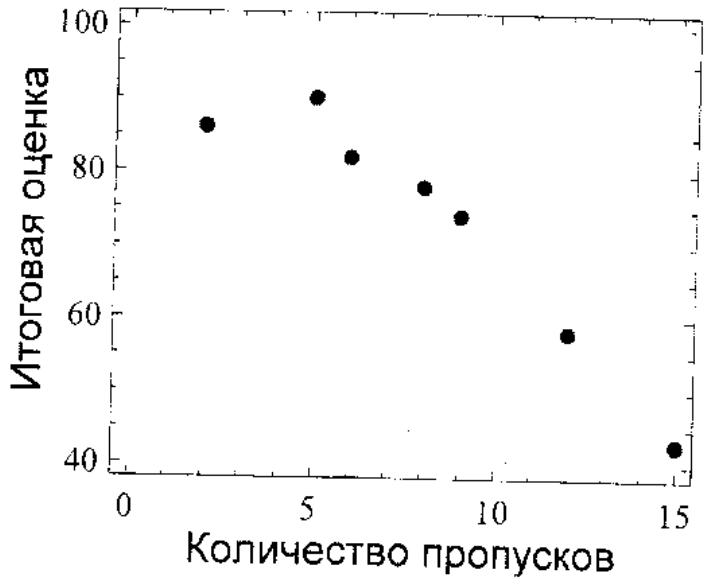


Рис.3. *Диаграмма рассеяния зависимости количества пропусков и итоговой оценки*

Пример. Построить диаграмму рассеяния, вычислить коэффициент корреляции Пирсона и определить тип зависимости для возраста субъекта и количества выкуриваемых им в день сигарет. Данные наблюдений для группы курящих представлены в табл. 4, колонки 1,2.

Решение. Диаграмма рассеяния для данного примера представлена на рис. 4. Единицы наблюдения на диаграмме расположены достаточно хаотично. Данные по расчету коэффициента корреляции, которые представлены в колонках 3 — 5 таблицы, также указывают на отсутствие связи. Знаки отклонений z_x и z_y соотносятся произвольно, их произведения также имеют различные знаки и после суммирования получается значение вблизи 0. Значение коэффициента корреляции равно -0.035.

Таблица 4. Расчет коэффициента корреляции зависимости возраста и числа выкуренных сигарет

Возраст x_i	Число сигарет y_i	$z_{x,i} = \left(x_i - \overline{x}\right)/s_x$	$z_{y,i} = \left(y_i - \overline{y}\right)/s_y$	$z_{x,i} \cdot z_{y,i}$
27	6	-0.92	-0.71	0.65
64	10	1.38	0.27	0.37
36	9	-0.36	0.02	-0.01
42	18	0.01	2.21	0.02
31	7	-0.67	-0.46	0.31
18	12	-1.48	0.75	-1.11
53	5	0.70	-0.95	-0.66
64	12	1.38	0.75	1.03
58	3	1.01	-1.44	-1.45
25	7	-1.04	-0.46	0.48
$r = \frac{\sum_{i=1}^{10} z_{x,i} \cdot z_{y,i}}{10} = -0.035$				

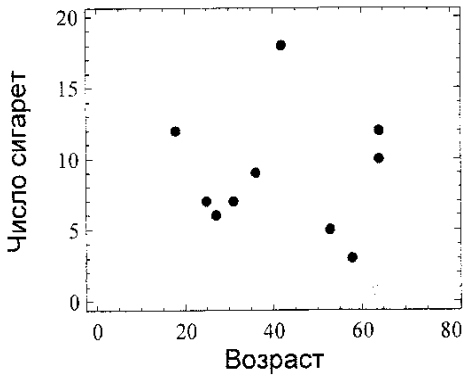


Рис. 4. *Диаграмма рассеяния зависимости возраста и числа выкуренных сигарет*

Замечание. Значение коэффициента корреляции вблизи нуля может означать не только отсутствие зависимости между величинами, но и то, что эта зависимость нелинейная. Иллюстрацией этого факта служит совокупность данных, представленных на рис. 5, для которой можно предположить существование квадратичной зависимости между признаками, тогда как коэффициент корреляции Пирсона равен 0.047.

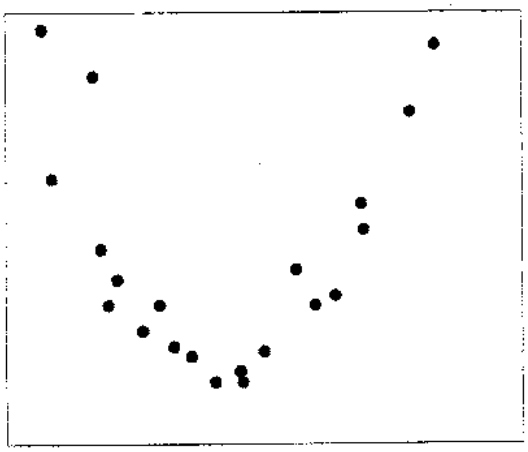


Рис. 4.5. *Диаграмма рассеяния с возможной нелинейной зависимостью*

На следующей лекции будет:

1) Проверка значимость коэффициента корреляции

2) Интерпретация значимой зависимости между величинами

3) Коэффициент корреляции рангов Спирмана