# Регрессия и корреляция

# Зависимость

Основная задача регрессионного и корреляционного анализа состоит в выявлении связи между случайными переменными. Например, на свободном рынке обычно наблюдается большая степень корреляции между размером урожая и рыночными ценами на соответствующую продукцию сельского хозяйства. Часто корреляция привлекает наше внимание к причинно-следственным связям, существующим между изучаемыми двумя рядами величин. В области естественных и общественных наук установление существенной корреляции часто заставляет нас искать возможные связи между явлениями, которые в противном случае могли остаться незамеченными.

В экономике в большинстве случаев между переменными величинами существуют зависимости, когда каждому значению одной переменной соответствует не какое-то определённое, а множество возможных значений другой переменной. Иначе говоря, каждому значению одной переменной соответствует определённое условное распределение другой переменной. Такая зависимость получила название статистической.

Возникновение понятия статистической связи обусловливается тем, что зависимая переменная подвержена влиянию неконтролируемых или неучтённых факторов, а также тем, что измерение значений переменных неизбежно сопровождается некоторыми случайными ошибками.

# Зависимость

Статистическая зависимость между двумя переменными, при которой каждому значению одной переменной соответствует определённое условное математическое ожидание (среднее значение) другой, называется корреляционной.

Функциональная зависимость представляет собой частный случай корреляционной. При функциональной зависимости с изменением значений некоторой переменной ходнозначно изменяется определенное значение переменной y, при корреляционной – определённое среднее значение (математическое ожидание) y, а при статистической – определённое распределение переменной y. Каждая корреляционная зависимость является статистической, но не каждая статистическая зависимость является корреляционной.

Статистические связи между переменными можно изучать методами корреляционного и регрессионного анализа. Основной задачей корреляционного анализа является выявление связи между случайными переменными и оценка её степени. Основной задачей регрессионного анализа является установление формы и изучение зависимости между переменными.

# Корреляция

Корреляция определяет степень, с которой значения двух переменных «пропорциональны» друг другу. Пропорциональность означает просто линейную зависимость. Корреляция высокая, если на графике зависимость «можно представить» прямой линией (с положительным или отрицательным углом наклона). Таким образом, это простейшая регрессионная модель, описывающая зависимость одной переменной от одного фактора.

В производственных условиях обычно информации, полученной из диаграмм рассеяния при условии их корректного построения, бывает достаточно для того, чтобы оценить степень зависимости y от x. Но в ряде случаев требуется дать количественную оценку степени связи между величинамих и y. Такой оценкой является коэффициент корреляции.

# Отметим основные характеристики этого показателя.

- Он может принимать значения от –1 до +1. Знак «+» означает, что связь прямая (когда значения одной переменной возрастают, значения другой переменной также возрастают), «–» означает, что связь обратная.
- Чем ближе коэффициент к |1|, тем теснее линейная связь. При величине коэффициента корреляции менее 0,3 связь оценивается как слабая, от 0,31 до 0,5 – умеренная, от 0,51 до 0,7 – значительная, от 0,71 до 0,9 – тесная, 0,91 и выше – очень тесная.
- Если все значения переменных увеличить (уменьшить) на одно и то же число или в одно и то же число раз, то величина коэффициента корреляции не изменится.
- При r=±1 корреляционная связь представляет линейную функциональную зависимость. При этом все наблюдаемые значения располагаются на общей прямой. Её ещё называют линией регрессии.
- При r=0 линейная корреляционная связь отсутствует. При этом групповые средние переменных совпадают с их общими средними, а линии регрессии параллельны осям координат.

# Корреляция

Основываясь на коэффициентах корреляции, вы не можете строго доказать причинной зависимости между переменными, однако можете определить ложные корреляции, т.е. корреляции, которые обусловлены влияниями «других», остающихся вне вашего поля зрения переменных. Лучше всего понять ложные корреляции на простом примере. Известно, что существует корреляция между ущербом, причиненным пожаром, и числом пожарных, тушивших пожар. Однако эта корреляция ничего не говорит о том, насколько уменьшатся потери, если будет вызвано меньше число пожарных. Причина в том, что имеется третья переменная (начальный размер пожара), которая влияет как на причинённый ущерб, так и на число вызванных пожарных. Если вы будете учитывать эту переменную, например, рассматривать только пожары определённой величины, то исходная корреляция между ущербом и числом пожарных либо исчезнет, либо, возможно, даже изменит свой знак. Основная проблема ложной корреляции состоит в том, что вы не знаете, кто является её носителем. Тем не менее, если вы знаете, где искать, то можно воспользоваться частные корреляции, чтобы контролировать (частично исключённое) влияние определённых переменных.

Корреляция, совпадение или необычное явление сами по себе ничего не доказывают, но они могут привлечь внимание к отдельным вопросам и привести к дополнительному исследованию. Хотя корреляция прямо не указывает на причинную связь, она может служить ключом к разгадке причин. При благоприятных условиях на её основе можно сформулировать гипотезы, проверяемые экспериментально, когда возможен контроль других влияний, помимо тех немногочисленных, которые подлежат исследованию.

Иногда вывод об отсутствии корреляции важнее наличия сильной корреляции. Нулевая корреляция двух переменных может свидетельствовать о том, что никакого влияния одной переменной на другую не существует, при условии, что мы доверяем результатам измерений.

# Регрессионный анализ

Регрессионный анализ является одним из наиболее распространённых методов обработки экспериментальных данных при изучении зависимостей в физике, биологии, экономике, технике и других областях.

Исследование объективно существующих связей между явлениями – важнейшая задача общей теории статистики. Регрессионный анализ заключается в определении аналитического выражения, в котором изменение одной величины (называемой зависимой или результативным признаком) y обусловлено влиянием одной или нескольких независимых величин (факторов) x1, x2,…, xn, а множество всех прочих факторов, также оказывающих влияние на зависимую величину, принимается за постоянные и средние значения.

# Регрессионный анализ

Регрессия может быть однофакторной (парной) и многофакторной (множественной). Для простой (парной) регрессии в условиях, когда достаточно полно установлены причинно-следственные связи, можно использовать графическое изображение. При множественности причинных связей невозможно чётко разграничить одни причинные явления от других. В этом случае наиболее приемлемым способом определения зависимости (уравнения регрессии) является метод перебора различных уравнений, реализуемый с помощью компьютера.

После выбора вида регрессионной модели, используя результаты наблюдений зависимой переменной и факторов, нужно вычислить оценки (приближённые значения) параметров регрессии, а затем проверить значимость и адекватность модели результатам наблюдений.

# Порядок проведения регрессионного анализа следующий:

- выбор модели регрессии, что заключает в себе предположение о зависимости функций регрессии от факторов;
- оценка параметров регрессии в выбранной модели методом наименьших квадратов;
- проверка статистических гипотез о регрессии.