Парная линейная регрессионная модель

В экономических исследованиях одной из основных задач является анализ зависимостей между переменными. Зависимость может быть строгой (функциональной) либо статистической. Алгебра и математический анализ занимаются изучением функциональных зависимостей, то есть зависимостей, заданных в виде точных формул. Но любая такая зависимость в определенной степени является абстракцией, поскольку в окружающем мире, частью которого является экономика, значение конкретной величины не определяется неизменной формулой ее зависимости от некоторого набора других величин. Всегда есть несколько величин, которые определяют главные тенденции изменения рассматриваемой величины, и в экономической теории и практике ограничиваются тем или иным кругом таких величин (объясняющих переменных). Однако всегда существует и воздействие большого числа других, менее важных или трудно идентифицируемых факторов, переменной от конкретной формулы ее связи с объясняющих переменных, построение формул зависимости и оценка их параметров являют не только одним из важнейших разделов математической статистики. Это своего рода искусство, учитывающее в каждой конкретной области знаний (в частности, в экономике, о которой идет речь), ее внутренние законы и потребности. Но это также и наука, поскольку выбираемый и оцениваемый вид формулы должен быть объяснен в терминах данной области знаний.

Пусть требуется оценить связь между переменными (например, связь показателей безработицы и инфляции в данной стране за определенный период времени). В частности, может стоять вопрос, связаны ли между собой эти показатели, и при положительном ответе на него, естественно, встает задача нахождения формулы этой связи. Основой для ответа на этот вопрос являются статистические данные о динамике этих показателей (годовые, квартальные, месячные и т.п.). Эти данные представляют собой некоторую, предположительно – случайную, выборку из генеральной совокупности, то есть из совокупности всех возможных сочетаний

показателей инфляции и безработицы в сложившихся условиях. Таким образом, вывод о наличии связи для всей генеральной совокупности нужно делать по выборочным данным, что само по себе уже делает ответ на поставленный вопрос небезусловным. Более того, по данным выборки ответить на вопрос в приведенной постановке, то есть о наличии связи «вообще», невозможно. Действительно, через любые n точек на плоскости всегда можно провести полином степени m и объявить, что найдена точная формула связи. Однако опыт подсказывает, что если бы мы получили еще одну точку-наблюдение, то она наверняка не удовлетворяла бы найденной формуле. Поэтому вопрос о наличии связи между переменными (в частности – экономическими) следует ответить на вопрос о наличии конкретной формулы (спецификации) такой связи, устойчивой к изменению числа наблюдений. При этом нужно понимать, что ответ на этот вопрос по данным выборки не может быть однозначным и категоричным.

Простейшей формой зависимости между переменными является линейная зависимость, и проверка наличия такой зависимости, оценивание ее индикаторов и параметров является одним из важнейших направлений приложения математической статистики.

Рассмотрим вначале вопрос о линейной связи двух переменных:

1) Связаны ли между собой линейно переменные?
2) Какова формула связи переменных?

В первом случае переменные выступают как равноправные, здесь нет независимой и зависимости одной переменной от другой, например об оценивании формулы $y = a + bx$ (a и b - неизвестные коэффициенты такой зависимости). В этом случае переменная X является независимой (объясняющей), а переменная Y - зависимой (объясняемой). Вопрос о нахождении формулы зависимости можно ставить после положительного ответа на вопрос о существовании такой зависимости, но эти два вопроса можно решать и одновременно.

Для ответа на поставленные вопросы существуют специальные статистические методы и, соответственно, показатели, значения которых определенным образом (и с определенной вероятностью) свидетельствуют о наличии или отсутствии линейной связи между переменными. В первом случае это коэффициент корреляции величин во втором случае – коэффициенты линейной регрессии их стандартные ошибки - статистики, по значениям которых проверяется гипотеза об отсутствии связи величин.

Вначале объясним логику появления такого показателя, как коэффициент корреляции. Предположим, что между переменными существует линейная связь. Наличие такой связи можно интерпретировать следующим образом. Если переменная принимает значения большие, чем ее среднее значение, и связь положительна (на языке формул это означает, что коэффициент положителен), то значение переменной также должно быть больше ее среднего значения и соотношение отклонений от их средних значений должно быть постоянным. Если связь переменных отрицательна, то положительное отклонение от среднего значения должно сочетаться с отрицательным отклонением от ее средней, а отрицательное отклонение от среднего значения – с положительным отклонением от ее средней – при постоянном соотношении этих отклонений. Если линейной связи между переменными нет, то положительные отклонения переменной от ее среднего значения могут (хотя и не обязательно будут) сочетаться как с положительными, так и с отрицательными отклонениями от ее среднего, то же можно сказать и про отрицательные отклонения от среднего.

r>0 r<0 r=0

у у у

Рис 9.1. Виды корреляционной связи.

В качестве меры для степени линейной связи двух переменных используется коэффициент их корреляции:

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

По формуле коэффициента корреляции видно, что он будет положителен, если отклонения переменных от своих средних значений имеют, как правило, одинаковый знак, и отрицательным – если разные знаки.

Коэффициент корреляции является безразмерной величиной (так как размерности числителя и знаменателя есть размерности произведения) ; его величина не зависит от выбора единиц измерения обеих переменных. Величина коэффициента корреляции меняется от – 1 в случае строгой линейной отрицательной связи до +1 в случае строгой линейной положительной связи. Близкая к нулю величина коэффициента корреляции говорит об отсутствии линейной связи переменных, но не об отсутствии связи между ними вообще. Последнее вытекает из того, что каждой паре одинаковых отклонений переменной от ее среднего значения соответствуют равные по абсолютной величине положительное и отрицательное отклонения переменной от ее среднего. Соответственно, произведения этих отклонений «гасят» друг друга в числителе формулы коэффициента корреляции, и он оказывается близким к нулю. Заметим, что в числителе формулы для выборочного коэффициента корреляции величин стоит их показатель ковариации: $\mathrm{cov}(x,y) = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})$.

Этот показатель, как и коэффициент корреляции, характеризует степень линейной связи величин и он также равен нулю, если эти величины независимы. Однако, в отличие от коэффициента корреляции, показатель ковариации не нормирован – он имеет размерности, и его величина зависит

от единиц измерения величин. В статистическом анализе показатель ковариации сам по себе используется редко; он фигурирует обычно как промежуточный элемент расчета коэффициента корреляции.

Далее, в анализе коэффициента корреляции возникает следующий вопрос. Если он равен нулю для генеральной совокупности, это вовсе не значит, что он в точности будет равен нулю для выборки. Наоборот, он обязательно будет отклоняться от истинного значения, но чем больше такое отклонение, тем менее оно вероятно при данном объеме выборки. Таким образом, при каждом конкретном значении коэффициента корреляции величин для генеральной совокупности выборочный коэффициент корреляции является случайной величиной. Следовательно, случайной величиной является также любая его функция, и требуется указать такую функцию, которая имела бы одно из известных распределений, удобное для табличного анализа. Для выборочного коэффициента корреляции такой функцией является статистика, рассчитываемая по $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$.

Число степеней свободы меньше числа наблюдений на 2, поскольку в формулу выборочного коэффициента корреляции входит средние выборочные значения, для расчета которых используются две линейные формулы их зависимости от наблюдений случайных величин. Сразу уточним, что для коэффициента корреляции будет проверяться нулевая гипотеза, то есть гипотеза о равенстве его нулю в генеральной совокупности. Эта гипотеза отвергается, если выборочный коэффициент корреляции слишком далеко отклонился от нулевого значения, то есть произошло событие, которое было бы маловероятным в случае $\rho_{xy} = 0$.

Здесь, конечно, очень важно понять, что конкретно значат слова «слишком далеко» и «маловероятное событие». В последнем случае нужно задать вероятность такого события, которая называется в статистике «уровень значимости». Чаще всего задается уровень значимости 1% или 5%. Если для некаторого показателя проверяется гипотеза, то она отвергается в том случае,

если оценка показателя по данным выборки такова, что вероятность получения такого или большего (по модулю) ее значения меньше, чем 1% или 5% соответственно.

9.3. Парный регрессионный анализ

Коэффициент корреляции показывает, что две переменные связаны друг с другом, однако он не дает представления о том, каким образом они связаны. Рассмотрим более подробно те случаи, что одна переменная зависит от другой.

$$Y = \alpha + \beta x + u$$

Здесь $\alpha + \beta x$ - неслучайная составляющая, где x выступает как объясняющая переменная, u – случайный член.

Точки P – это единственные точки отражающие реальные значения переменных. Фактические значения α и β и положения точки Q неизвестны, так же как и фактические значения случайного члена.

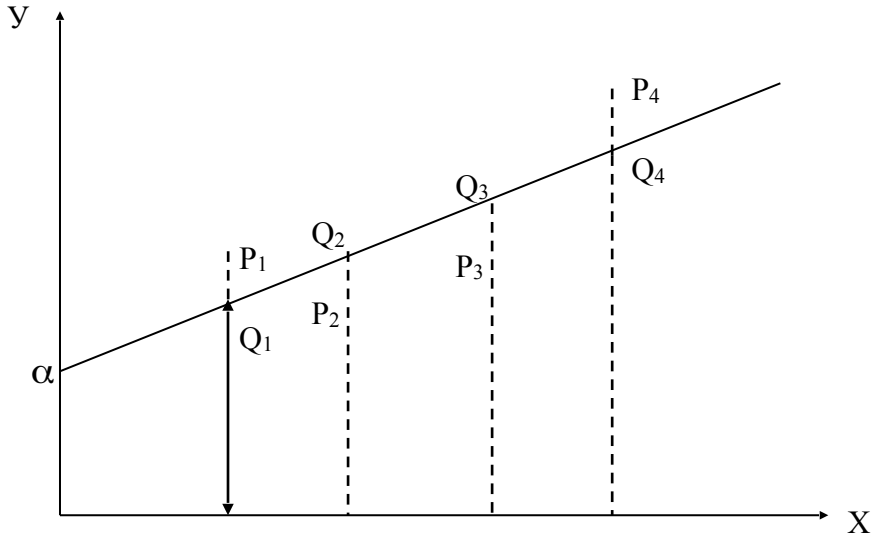

Рис.1

Задача регрессионного анализа состоит в получении оценок α и β и следовательно, в определении положения прямой по точкам P.

Очевидно, что чем меньше значения u, тем легче эта задача.