Лекция 8.

Тема: **МОДЕЛЬ ПАРНОЙ РЕГРЕССИИ**

Большинство процессов и явлений в экономике являются многофакторными и, как правило, в них на поведение исследуемого экономического показателя влияет множество других факторов. В экономической действительности хоть и реже, но все - таки встречаются ситуации, в которых на поведение результирующего признака доминирующим образом влияет один другой экономический показатель, а влияние остальных является слабым и случайным. Для описания таких зависимостей используется модель парной регрессии.

Часто для представления регрессионной зависимости подходят одновременно несколько функций, поэтому выбор окончательного вида зависимости нужно осуществлять на альтернативной основе. Естественно построение различных моделей нужно начинать с самой простой – линейной. Тем более, что она является наиболее удобной в плане ее построения, анализа, использования и интерпретации. Если после построения линейной модели и ее статистического анализа выяснится, что она является достаточно точной (адекватной), необходимость в построении других моделей само собой отпадет. Поэтому разбор процесса построения, анализа, интерпретации и применения моделей парной регрессии начнем с линейной.

1.1. Модель парной линейной регрессии

Пусть имеется выборка наблюдений (x_i, y_i) $i = \overline{1,n}$ и на основе этой выборки нужно определить вид зависимости и построить выборочное уравнение регрессии. Есть предположение, что зависимость между выборочными значениями переменных *X* и *Y* есть **линейная**. Чтобы подтвердить или опровергнуть это предположение, вычисляется коэффициент парной корреляции между выборочными значениями переменных *X* и *Y* и проверяется на значимость.

Выборочный коэффициент парной корреляции между переменными *X* и *Y*, определяемый по выборке из *n* наблюдений, вычисляется по формуле:

$$r_{xy} = \frac{\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\overline{x})\cdot(y_i-\overline{y})}{\sqrt{\frac{1}{n}\cdot\sum_{i=1}^{n}(x_i-\overline{x})^2\cdot\frac{1}{n}\cdot\sum_{i=1}^{n}(y_i-\overline{y})^2}}. \qquad (2.1)$$

Значимость коэффициента парной корреляции проверяется следующим образом: вычисляется статистика

$$t_r = r_{xy} \cdot \sqrt{\frac{(n-2)}{1-r_{xy}^2}} \,. \qquad (2.2)$$

Она подчиняется распределению Стьюдента с числом степеней свободы $\nu = n - 2$. Выдвигается нулевая гипотеза H_0: $\rho_{xy} = 0$. Далее для заданного уровня значимости α и числа степеней свободы $\nu = n - 2$ по таблице распределения критерия Стьюдента определяется $t_{кр} = t(\alpha, \nu)$.

Если $|t_r| > t_{кр}$, то нулевая гипотеза об отсутствии линейной зависимости между переменными X и Y отвергается, в этом случае переменные X и Y считаются коррелированными, то есть между выборочными значениями этих переменных существует значимая **линейная зависимость.**

Если $|t_r| < t_{кр}$, то нет оснований для того, чтобы отвергнуть нулевую гипотезу, в этом случае необходимо признать, что между переменными X и Y не существует значимой линейной зависимости т. е. они не коррелированы.

Далее рассматривается случай, когда выборочный коэффициент парной корреляции r_{xy} является значимым. В этом случае необходимо строить выборочное уравнение регрессии вида:

$$\hat{Y} = a + b \cdot X \,, \qquad (2.3)$$

где a, b – оценки параметров α и β зависимости по генеральной совокупности:

$$Y = \alpha + \beta \cdot X + \varepsilon, \qquad (2.4)$$

где X – неслучайная величина; Y и ε – случайные величины, Y – объясняемая переменная, X – объясняющая переменная, α и β – параметры зависимости.

Наличие случайной величины ε в зависимости (2.4) диктуется возможным влиянием на объясняемую переменную y, не включенных в уравнение объясняющих переменных, а также с возможной нелинейностью модели и ошибками измерения.

Определяется остаток e_i в i – ом наблюдении как разность между наблюдаемым и расчетным значениями объясняемой переменной т. е.

$$e_i = y_i - \hat{y}_i. \qquad (2.5)$$

Неизвестные значения коэффициентов a и b определяются методом наименьших квадратов (МНК).

Для фиксированной выборки наблюдений (x_i, y_i) $i = \overline{1,n}$ – известные значения; а a, b – неизвестные.

Методом подстановки позволяет получить следующие выражения для коэффициентов a и b:

$$b = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2} = \frac{\operatorname{cov}(x, y)}{\operatorname{var}(x)}, \tag{2.7}$$

$$a = \overline{y} - b \cdot \overline{x}. \tag{2.8}$$

Подстановка в (2.4) выражения для a из (2.8) позволяет получить:

$$\hat{y} - \overline{y} = b \cdot (x - \overline{x}). \tag{2.9}$$

Из (2.9) следует, что линия регрессии проходит через точку $(\overline{x}, \overline{y})$, и выполняются равенства: $\overline{e} = 0,\ \overline{y} = \overline{\hat{y}}$.

Коэффициент b есть угловой коэффициент регрессии, он показывает, на сколько единиц, в среднем, изменяется переменная Y при изменении переменной X на одну единицу.

Коэффициент a дает прогнозируемое значение объясняемой переменной Y при $X = 0$. Это может иметь смысл в зависимости от того, далеко ли находится $X = 0$ от выборочных значений X.