$$Corr(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x) \cdot Var(}}$$

Эконометрическое моделирование

Лабораторная работа № 3

Парная регрессия

Оглавление

Парная регрессия

Парная регрессия - уравнение связи двух переменных y и x:

$y = \hat{f}(x) + \varepsilon$, где $\hat{f}(x)$ – модель регрессии;

y – зависимая переменная (результативный признак);

x – независимая, объясняющая переменная (признак-фактор).

В экономических исследованиях используют линейные и нелинейные модели регрессии.

Линейное однофакторное уравнение имеет вид:

$$y = \hat{\beta}_0 + \hat{\beta}_1 \cdot x + \varepsilon.$$

Из нелинейных моделей регрессии наиболее часто в экономических исследованиях используют полулогарифмические:

$\ln y = \hat{\beta}_0 + \hat{\beta}_1 \cdot x + \varepsilon$, $\quad y = \hat{\beta}_0 + \hat{\beta}_1 \cdot \ln x + \varepsilon.$

Для характеристики нелинейной зависимости могут строиться модели регрессии, нелинейные по объясняющим переменным, например, полиномы разных степеней. Для отображения обратной зависимости между показателями используется уравнение гиперболы. К Регрессиям нелинейным по оцениваемым параметрам относят – степенную, логарифмическую показательную и экспоненциальную.

Метод наименьших квадратов (МНК)

Эконометрика занимается построением моделей на основе полученных экспериментальных данных для объяснения и прогнозирования поведения экономических систем.

При проведении эконометрических исследований на основе моделей экономической теории предлагается гипотетическая параметрическая модель, а расчет количественных значений параметров моделей производится так, чтобы минимизировать расхождение между исходными (y) и вычисленными по модели ($\hat{y}$) значениями показателей. Причём само понятие «расхождение» может выбираться в разных смыслах, в зависимости от ситуации, типа данных и вычислений.

Одним из распространенных в эконометрике методов оценивания параметров моделей является метод наименьших квадратов.

Алгоритм метода наименьших квадратов.

При применении метода наименьших квадратов (МНК) необходимо учитывать следующие обстоятельства:

1. Метод наименьших квадратов применяется для количественного расчета параметров аппроксимирующей функции.

2. МНК применяется для функций, *линейных относительно параметров*. Некоторые функции могут быть приведены к линейному виду относительно параметров обратимыми преобразованиями, например путем логарифмирования.

3. Предварительно выбирается класс функций, который аппроксимирует изучаемые зависимости (например, класс линейных функций $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$)

4. В соответствии с принципом МНК в заданном классе функций находится функция, для которой выполняется условие: сумма квадратов отклонений фактических данных от «теоретических» должна быть минимальной.

5. Это требование записывается следующим образом:

$$S = \sum (y - \hat{y})^2 \to min$$

Далее в выражение S вместо $\hat{y}$ подставляется её аналитическое выражение. В нашем случае это будет выглядеть так:

$$S = \sum \left(y - \hat{\beta}_0 - \hat{\beta}_1 \cdot x\right)^2 \to min$$

Задача сводится к нахождению минимума функции нескольких переменных (по числу неизвестных параметров). В нашем примере:

$$S = f\left(\hat{\beta}_0, \hat{\beta}_1\right) \to min$$

Для нахождения экстремума функции необходимым условием является равенство нулю частных производных функции S по каждому из параметров.

$$\begin{cases} \dfrac{\partial S}{\partial \hat{\beta}_0} = \sum 2 \cdot \left(y - \hat{\beta}_0 - \hat{\beta}_1 \cdot x\right) \cdot (-1) = 0 \\ \dfrac{\partial S}{\partial \hat{\beta}_1} = \sum 2 \cdot \left(y - \hat{\beta}_0 - \hat{\beta}_1 \cdot x\right) \cdot (-x) = 0 \end{cases}$$

Из данной системы получаются формулы для нахождения неизвестных параметров:

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}, \quad \hat{\beta}_1 = \frac{\overline{y \cdot x} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2}.$

Интерпретация уравнения регрессии

Во-первых, можно сказать, что увеличение X на одну единицу (в единицах измерения переменной X) приведет к увеличению значения Y на β_1 единиц (в единицах измерения переменной Y). Параметр β_0 дает прогнозируемое значение Y, если X=0. Это может иметь или не иметь явного смысла в зависимости от контекста.

Средний коэффициент эластичности показывает, на сколько процентов в среднем по совокупности изменится результат y от своей средней величины при изменении фактора x на 1 % от своего среднего значения:

$$\overline{Э} = f'(x)\frac{\bar{x}}{\bar{y}}$$

Оценка качества построенной модели

Оценку качества построенной модели дает коэффициент детерминации R^2, а также средняя ошибка аппроксимации.

$$\bar{O} = \frac{1}{n}\sum \left|\frac{y - \bar{y}}{y}\right| \cdot 100\%$$

Допустимый предел значений средней ошибки аппроксимации составляет 8–10%.

Долю дисперсии, объясняемую регрессией, в общей дисперсии результативного признака y характеризует коэффициент детерминации R^2:

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{ESS}{TSS}$$

где R^2 – коэффициент детерминации;

$\sum(y_i - \bar{y})^2 = TSS$ – общая сумма квадратов отклонений;

$\sum(\hat{y}_i - \bar{y})^2 = ESS$ – сумма квадратов отклонений, обусловленная регрессией (объясненная сумма квадратов);

$\sum(y_i - \hat{y}_i)^2 = RSS$ – остаточная сумма квадратов отклонений, TSS = ESS + RSS.

F-тест – оценивание качества уравнения регрессии состоит в проверке гипотезы H_0 об отсутствии между величинами *X* и *Y*. Для этого выполняется сравнение фактического $F_{расч}$ ($F_{набл}$) и критического (табличного) $F_{крит}$ значений критерия Фишера.

$$F_{расч} = \frac{R^2}{1-R^2}\frac{n-k}{k-1} = \frac{R^2}{1-R^2}\frac{(n-1-m)}{m}$$

где n – объем выборки, k – число коэффициентов β, m – число независимых переменных.

Если $F_{расч} < F_{крит}$, то гипотеза H_0 не отклоняется (принимается), то есть мы делаем вывод о том, что все независимые переменные x не оказывает значимого влияния на переменную y. В этом случае уравнение называют незначимым. В противном случае гипотеза H_0 не принимается (отклоняется).

Для оценки статистической значимости коэффициентов регрессии рассчитываются t-критерий Стьюдента и доверительные интервалы. Оценка значимости коэффициента регрессии с помощью t-критерия Стьюдента проводится путем сопоставления их значений с величиной случайной ошибки:

$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{SE_{\hat{\beta}_1}},\ t_{\hat{\beta}_0} = \frac{\hat{\beta}_0}{SE_{\hat{\beta}_0}}.$$

Случайные ошибки параметров линейной регрессии определяются по формулам:

$$SE_{\hat{\beta}_1} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2/(n-2)}{\sum(x_i - \bar{x})^2}}$$

$$SE_{\hat{\beta}_0} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{(n-2)} \cdot \frac{\sum x^2}{n\sum(x_1 - \bar{x})^2}}$$

Сравнивая фактическое (наблюдаемое) и критическое (табличное) значения t-статистики принимаем гипотезу H_0 – параметр β_k является не значимым (переменная $x^{(k)}$ не оказывает значимого влияния на переменную y).

Если $|t_{расч}| < t_{крит}$ то гипотеза H0 не отклоняется (принимается), то есть мы делаем вывод о том, что переменная $x^{(k)}$ не оказывает значимого влияния на переменную y. В этом случае коэффициент при переменной $x^{(k)}$ называют незначимым.

В противном случае гипотеза H0 не принимается (отклоняется).

Формулы для расчета доверительных интервалов имеет следующий вид:

$$(\hat{\beta}_k - t_{n-k} \cdot SE_{(\hat{\beta}_k)}, \hat{\beta}_k + t_{n-k} \cdot SE_{(\hat{\beta}_k)})$$

Доверительный интервал – это границы, в которых с вероятностью (1-α) находятся значения истинных параметров регрессии.

Если в границы доверительного интервала попадает ноль, т.е. нижняя граница отрицательна, а верхняя положительна, то оцениваемый параметр принимается нулевым, так как он не может одновременно принимать и положительное и отрицательное значение.

Прогнозное значение y_p определяется путем подстановки в уравнение регрессии соответствующего прогнозного значения x_p. Вычисляется средняя стандартная ошибка прогноза:

$$SE_{\hat{y}_p} = \sqrt{\frac{\sum(y_i - \hat{y})^2}{(n-m-1)}} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

И строится доверительный интервал:

$$\left(\hat{y}_p - t_{n-k} \cdot SE_{\hat{y}_p}, \hat{y}_p + t_{n-k} \cdot SE_{\hat{y}_p}\right)$$

Задание 1. Сбор статистических данных

Подберите статистические данные для анализа взаимосвязи между количеством занятых и поступлением налогов (в разрезе субъектов РФ). Для подбора данных воспользуйтесь Единой межведомственной информационно-статистической системой.

Год	Вариант
2006	1, 9, 17
2007	2, 10, 18
2008	3, 11, 19
2009	4, 12, 20
2010	5, 13, 21
2011	6, 14, 22
2012	7, 15, 23
2013	8, 16, 24

Задание 2. Построение регрессионной модели в Excel

Изучите функцию ЛИНЕЙН. Воспользуйтесь ей для построения регрессионной модели.

Задание 3. Оценка качества построенной модели.

Проведите оценку качества построенной модели. Дайте интерпретацию полученных результатов.