

ТЕМА 1. ПРОСТАЯ ЛИНЕЙНАЯ ЭКОНОМЕТРИЧЕСКАЯ МОДЕЛЬ

Оценивание параметров эконометрической модели методом наименьших квадратов

Простая линейная регрессионная модель устанавливает линейную зависимость между двумя переменными, например, производительностью труда и уровнем механизации производства, товарооборотом и площадью торговой точки, уровнем потребления и уровнем доходов населения и т.п.

В общем виде, как уже отмечалось, уравнение парной регрессии выглядит следующим образом:

$$y = a_0 + a_1x + u,$$

где y – зависимая переменная (регрессант),

x – независимая переменная (регрессор),

a_0, a_1 – оцениваемые параметры,

u – ошибка линии регрессии.

Чтобы иметь явный вид зависимости, необходимо найти (оценить) неизвестные параметры a_0 и a_1 . Для их оценки и построения эконометрической модели необходимо сформировать совокупность наблюдений. Данная совокупность наблюдений может быть отражена в системе координат в виде корреляционного поля (рис. 1.1).

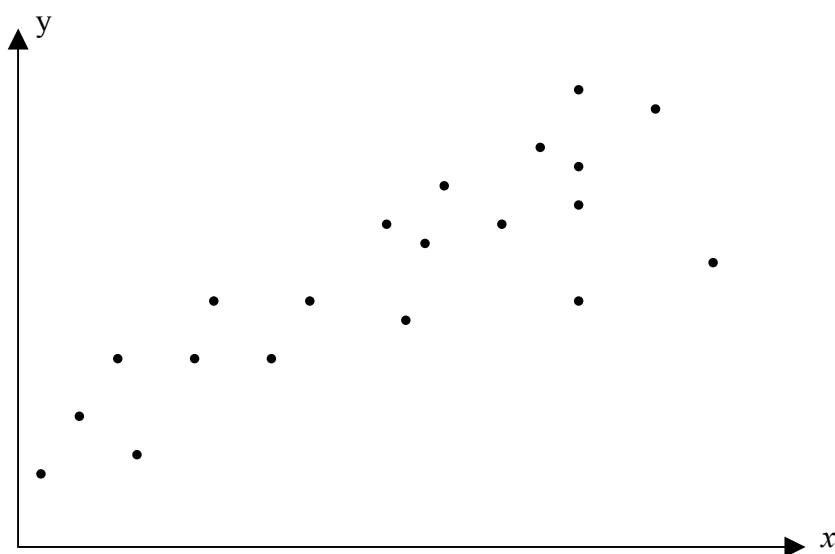


Рис. 1.1. Корреляционное поле

На основе гипотезы о линейной зависимости между x и y , через корреляционное поле можно провести множество прямых линий (рис. 1.2), которые отличаются друг от друга величинами a_0 и a_1 .

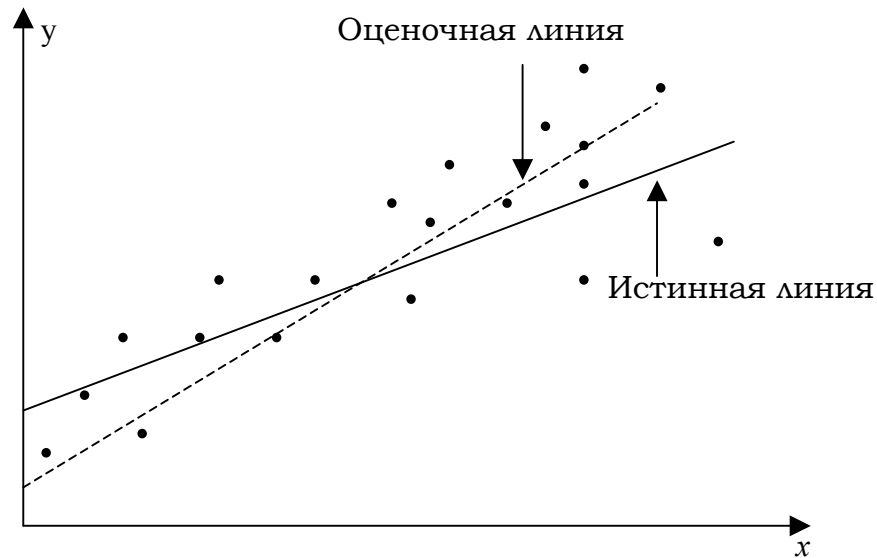


Рис. 1.2. Корреляционное поле и линии регрессии

Даже наиболее точно подобранная линия регрессии будет содержать отклонения истинных наблюдений от расчетных, т.е. лежащих на прямой (рис. 1.3).

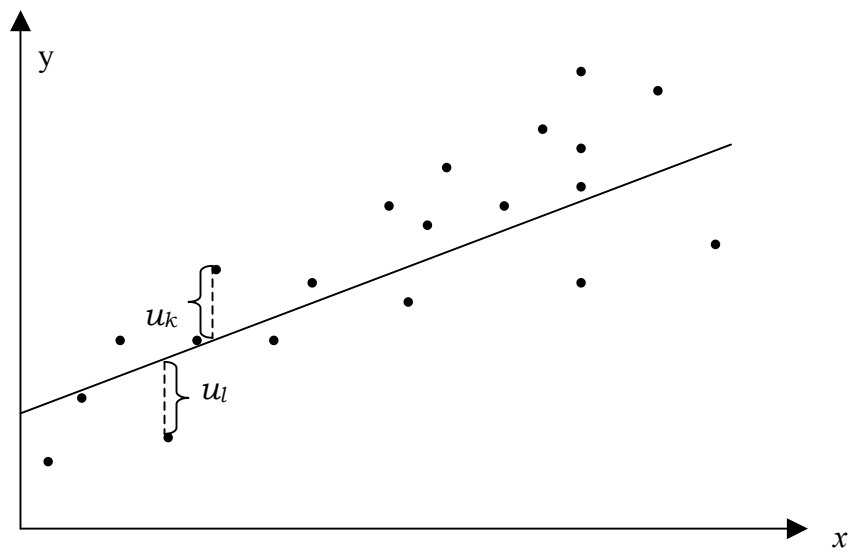


Рис. 1.3. Иллюстрация отклонений

Отклонения от точек, лежащих на прямой, могут быть как положительными (u_k), так и отрицательными (u_l). В связи с этим возникает вопрос, каким критерием воспользоваться, чтобы найденные параметры a_0 и a_1 в наибольшей степени соответствовали истине?

В качестве одного из показателей качества могла бы быть сумма отклонений. Но, учитывая их разнознаковость, указанная сумма не может быть взята в качестве критерия, так как значения отклонений в общей совокупности могут «погасить» друг друга.

Другим критерием качества может служить сумма квадратов отклонений. Здесь все они имеют одинаковый знак, и их сумма может служить индикатором точности рассчитанных значений a_0 и a_1 . Данный принцип и лежит в основе **метода наименьших квадратов (1МНК)**.

Итак, пусть оценка линии регрессии представлена в виде уравнения

$$\hat{y} = \hat{a}_0 + \hat{a}_1 x,$$

где $\hat{y}, \hat{a}_0, \hat{a}_1$ – оценки величин y , a_0 и a_1 .

Данное уравнение должно быть справедливо для всех наблюдений x и y .

Основная задача: подобрать такие $\hat{a}_0, \hat{a}_1$, которые минимизируют сумму квадратов отклонений $\hat{y}$ от y

$$\min \left\{ \sum_{i=1}^n u_i^2 \right\} = f(\hat{a}_0, \hat{a}_1).$$

Следует иметь в виду, что применение метода 1МНК возможно при условии, когда остатки (отклонения) распределены по нормальному закону, т.е. когда их среднее значение равно нулю, а дисперсия – постоянная величина.

Найдем указанную сумму

$$\sum_{i=1}^n u_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{a}_0 - \hat{a}_1 x_i)^2.$$

Для нахождения значений оцениваемых параметров $\hat{a}_0$ и $\hat{a}_1$ возьмем первые производные от найденного выражения по каждому из параметров, приравняем их нулю и решим относительно $\hat{a}_0$ и $\hat{a}_1$.

$$\left\{ \begin{array}{l} \frac{\partial \left(\sum_{i=1}^n u_i^2 \right)}{\partial a_0} = -2 \sum_{i=1}^n (y_i - \hat{a}_0 - \hat{a}_1 x_i) = 0 \\ \frac{\partial \left(\sum_{i=1}^n u_i^2 \right)}{\partial a_1} = -2 \sum_{i=1}^n x_i (y_i - \hat{a}_0 - \hat{a}_1 x_i) = 0 \end{array} \right. \Rightarrow \left\{ \begin{array}{l} \sum_{i=1}^n (y_i - \hat{a}_0 - \hat{a}_1 x_i) = 0 \\ \sum_{i=1}^n x_i (y_i - \hat{a}_0 - \hat{a}_1 x_i) = 0. \end{array} \right.$$

После элементарных преобразования получаем систему так называемых нормальных уравнений

$$\left\{ \begin{array}{l} n\hat{a}_0 + \hat{a}_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ \hat{a}_0 \sum_{i=1}^n x_i + \hat{a}_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i. \end{array} \right.$$

Решаем ее относительно $\hat{a}_0$ и $\hat{a}_1$.

1. Найдем $\hat{a}_0$ из первого уравнения системы

$$\hat{a}_0 = \frac{\sum_{i=1}^n y_i - \hat{a}_1 \sum_{i=1}^n x_i}{n}.$$

2. Подставим полученное выражение во второе уравнение

$$\frac{\sum_{i=1}^n y_i - \hat{a}_1 \sum_{i=1}^n x_i}{n} \sum_{i=1}^n x_i + \hat{a}_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i.$$

3. Находим $\hat{a}_1$

$$\hat{a}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2} \text{ или}$$

$$\hat{a}_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}, \text{ или}$$

$$\hat{a}_1 = \frac{\frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \cdot \bar{y}}{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2}, \text{ или}$$

$$\hat{a}_1 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{cov}(x, y)}{\text{var}(x)}, \text{ или}$$

где $\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ – ковариация между переменными x и y ,

$\text{var}(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ – дисперсия x .

Может быть использована и еще одна формула

$$\hat{a}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

В регрессионной модели коэффициент $\hat{a}_1$ называется **коэффициентом регрессии**.

Коэффициент регрессии в линейной модели показывает, на сколько единиц собственного измерения изменится в среднем зависимая переменная при изменении j -й независимой переменной в среднем на единицу собственного измерения при прочих неизменных условиях.

Зная значение оценки $\hat{a}_1$, можно найти параметр $\hat{a}_0$. Для этого воспользуемся уравнением

$$\hat{a}_0 = \frac{\sum_{i=1}^n y_i - \hat{a}_1 \sum_{i=1}^n x_i}{n}.$$

Отсюда получаем

$$\hat{a}_0 = \bar{y} - \hat{a}_1 \bar{x}.$$

Исходя из этого, имеем особенность уравнения регрессии: линия регрессии проходит через среднюю точку

$$\bar{y} = \hat{a}_0 + \hat{a}_1 \bar{x}.$$

Оценку $\hat{a}_0$ можно также рассчитать, используя формулу

$$\hat{a}_0 = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}.$$

Рассмотрим **пример**. Пусть имеется выборка, содержащая информацию о торговой площади магазина и суточном товарообороте

Наблюдение	Торговая площадь, м ² (x)	Товарооборот, тыс. грн. (y)
1	20	10
2	14	6
3	12	4
4	20	12
5	33	18
6	38	19
7	41	22
8	26	15

Необходимо построить модель, отражающую зависимость суточного товарооборота от размера торговой площади, в предположении о линейной связи между рассматриваемыми показателями.

Решение.

Построим корреляционное поле (рис. 1.4) и убедимся, что между товарооборотом и площадью торгового помещения существует зависимость, близкая к линейной.

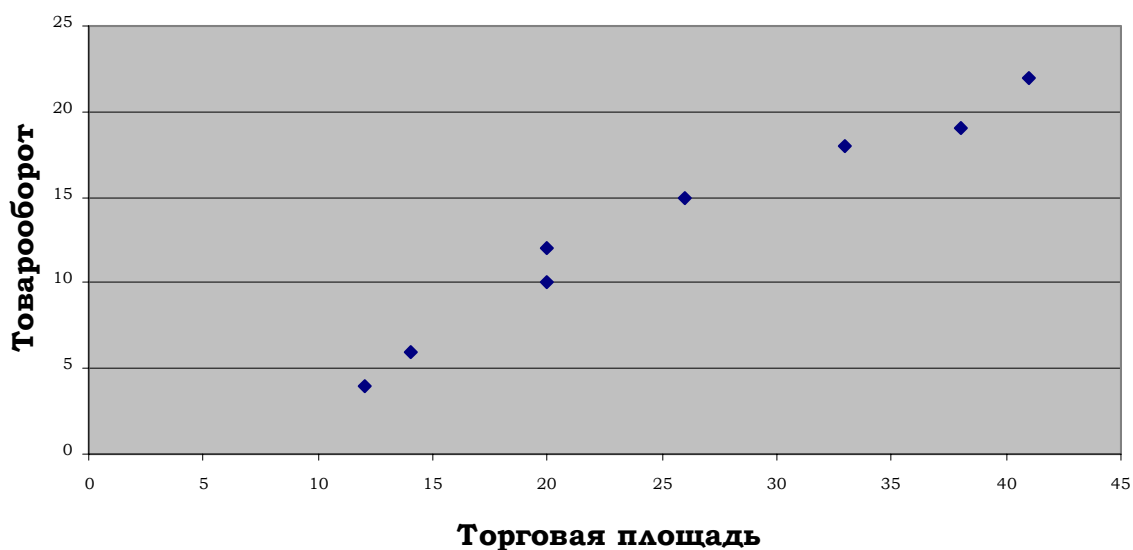


Рис. 1.4. Разброс точек, характеризующих зависимость между рассматриваемыми показателями

Найдем параметры уравнения $\hat{a}_0$ и $\hat{a}_1$, используя формулы

$$\hat{a}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{и} \quad \hat{a}_0 = \bar{y} - \hat{a}_1 \bar{x}$$

Для расчета параметра $\hat{a}_1$ построим вспомогательную таблицу

	x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	x_i^2
	20	10	-5,5	-3,25	17,875	30,25	400
	14	6	-11,5	-7,25	83,375	132,25	196
	12	4	-13,5	-9,25	124,875	182,25	144
	20	12	-5,5	-1,25	6,875	30,25	400
	33	18	7,5	4,75	35,625	56,25	1089
	38	19	12,5	5,75	71,875	156,25	1444
	41	22	15,5	8,75	135,625	240,25	1681
	26	15	0,5	1,75	0,875	0,25	676
Сумма	204	106			477	828	6030
Среднее	25,5	13,25					

$$\hat{a}_1 = \frac{477}{828} = 0,576,$$

$$\hat{a}_0 = 13,25 - 0,576 \cdot 25,2 = -1,44.$$

Таким образом, уравнение регрессии имеет вид:

$$\hat{y} = -1,44 + 0,576 \cdot x$$

Коэффициент регрессии $\hat{a}_1 = 0,576$ показывает, что при увеличении торговой площади в среднем на 1 м² суточный товарооборот возрастет в среднем на 0,576 тыс. грн. при прочих неизменных условиях.

При интерпретации оценок необходимо помнить

1. Прочие неизменные условия;
2. Что каждая числовая величина представляет собой оценку, которая, как правило, несет в себе ошибку;
3. Единицы измерения переменных, которые рассматриваются.

Уравнение регрессии можно использовать при принятии решений. Например, можно прогнозировать товарооборот торговой точки, которая должна быть построена с торговой площадью 30 м².

$$\hat{y} = -1,44 + 0,576 \cdot x = -1,44 + 0,576 \cdot 30 = 15,84 \text{ тыс. грн.}$$

Таким образом, оцененный (спрогнозированный) на основе уравнения регрессии товарооборот должен быть равен 15,84 тыс. грн.

С помощью регрессионного уравнения можно также оценить, какое значение должна иметь переменная x , чтобы при прочих неизменных условиях было получено определенное значение зависимой переменной y .

Так, пусть $\hat{y} = 20$ тыс.грн. Определим, какой должна быть площадь торговой точки, чтобы мог быть достигнут запланированный товарооборот. Для этого решим найденное уравнение относительно x и, подставив заданное значение $\hat{y}$, найдем искомую величину

$$x = \frac{\hat{y} + 1,44}{0,576} = \frac{20 + 1,44}{0,576} = 37,2 \text{ м}^2.$$

Таким образом, торговая площадь должна составлять 37,2 м².

Как уже отмечалось, уравнение регрессии обладает тем свойством, что оно справедливо для всех без исключения наблюдений переменных. То есть при неизменных коэффициентах регрессии значение каждого конкретного (i -го) наблюдения зависимой переменной является функцией i -го наблюдения независимой переменной (данное утверждение справедливо и для многофакторных моделей). В условиях нашего примера мы можем записать

$$\hat{y}_1 = -1,44 + 0,576 \cdot x_1,$$

$$\hat{y}_2 = -1,44 + 0,576 \cdot x_2,$$

$$\vdots$$

$$\hat{y}_8 = -1,44 + 0,576 \cdot x_8.$$

Естественно, что числовые характеристики коэффициентов регрессии, полученные с помощью метода наименьших квадратов, описывают в среднем зависимость между зависимой и независимыми переменными. В связи с этим имеют место отклонения u_i , соответствующие каждому наблюдению переменной:

$$u_i = y_i - \hat{y}_i,$$

где y_i – фактическое значение i -го наблюдения зависимой переменной, $\hat{y}_i$ – оценка i -го наблюдения зависимой переменной, полученная с помощью уравнения регрессии.

Оценка отклонений (ошибок) будет осуществлена ниже.

Коэффициенты эластичности, парной корреляции, детерминации

Наряду с коэффициентом регрессии, одним из показателей, характеризующим количественную связь между зависимой и независимой переменными, является **коэффициент эластичности**. Для линейной парной регрессионной модели данный коэффициент рассчитывается по формуле

$$E_{y/x} = \frac{\partial y}{\partial x} \cdot \frac{\bar{x}}{\bar{y}}$$

Он показывает, на сколько процентов в среднем изменится зависимая переменная при изменении независимой переменной в среднем на 1 процент при прочих неизменных условиях.

Рассчитаем коэффициент эластичности

$$\frac{\partial y}{\partial x} = \hat{a}_1 = 0,576,$$

$$E_{y/x} = \frac{\partial y}{\partial x} \cdot \frac{\bar{x}}{\bar{y}} = 0,576 \cdot \frac{25,5}{13,25} = 1,1.$$

Таким образом, при увеличении торговой площади в среднем на 1% товарооборот в среднем увеличится на 1,1% при прочих равных условиях.

В экономическом анализе очень важно знать, насколько тесно связаны между собой изучаемые показатели. Для оценки тесноты и направления связи между двумя показателями используется **коэффициент парной корреляции**. Он рассчитывается по формуле

$$r_{yx} = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)\text{var}(y)}} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Коэффициент корреляции, в отличие от коэффициента регрессии, является показателем относительной меры связи между двумя факторами. Значения коэффициента корреляции всегда находятся в пределах между -1 и +1

$$-1 \leq r_{yx} \leq 1.$$

Положительное значение коэффициента свидетельствует о прямой связи, а отрицательное – об обратной. При приближении коэффициента парной корреляции по абсолютной величине к 1 ($|r_{yx}| \rightarrow 1$) считается, что связь между показателями тесная, при приближении к 0 ($|r_{yx}| \rightarrow 0$) – связь отсутствует.

Очень тесной считается связь, когда ($|r_{yx}| \geq 0,7$), средней – когда ($0,5 \leq |r_{yx}| \leq 0,7$) и слабой, когда ($|r_{yx}| \leq 0,5$).

Коэффициент корреляции характеризует тесноту и направление связи между переменными.

Знак коэффициента корреляции **всегда** совпадает со знаком коэффициента регрессии.

Связь между коэффициентом корреляции и коэффициентом регрессии $\hat{a}_1$ выражается следующим соотношением:

$$r_{yx} = \hat{a}_1 \frac{\sigma_x}{\sigma_y},$$

$$r_{yx} = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)\text{var}(y)}} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}, \quad \hat{a}_1 = \frac{\text{cov}(x, y)}{\text{var}(x)} = \frac{\text{cov}(x, y)}{\sigma_x^2}.$$

Рассчитаем коэффициент корреляции. Для этого найдем дисперсии переменных, их среднеквадратические отклонения и ковариацию

$$\sigma_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad \sigma_x^2 = 103,5, \quad \sigma_x = 10,17349,$$

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2, \quad \sigma_y^2 = 35,6875, \quad \sigma_y = 5,973902,$$

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad \text{cov}(x, y) = 59,625.$$

С учетом этого имеем

$$r_{yx} = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{59,625}{10,17349 \cdot 5,973902} = 0,981.$$

Величина коэффициента парной корреляции, равная 0,981, говорит о прямой и тесной связи между показателями товарооборота и торговой площади

В простой эконометрической модели коэффициент корреляции служит одним из показателей ее качества.

Зададимся теперь вопросом: как вариация (дисперсия) зависимой переменной объясняется исходя из оцениваемой линии регрессии?

Сравнение расчетных значений с фактическими позволяет исследователю оценить степень адекватности гипотетической регрессионной модели. Различные статистические показатели могут быть использованы для оценки того, в какой степени $\bar{y}_i$ аппроксимирует соответствующее выборочное наблюдение y_i , но все они делают акцент на то, на сколько регрессионное уравнение, полученное с помощью метода МНК, объясняет значение y лучше, чем «примитивный» оценщик, каковым является выборочное среднее $\bar{y}$.

Вот почему эконометрики используют квадрат отклонений y_i от их средних как меру количества вариации y , объясняемой регрессией. Это расчетное количество обычно называется общей суммой квадратов.

Рассмотрим декомпозицию (разложение) отклонений y с помощью иллюстративного примера (рис. 1.5).

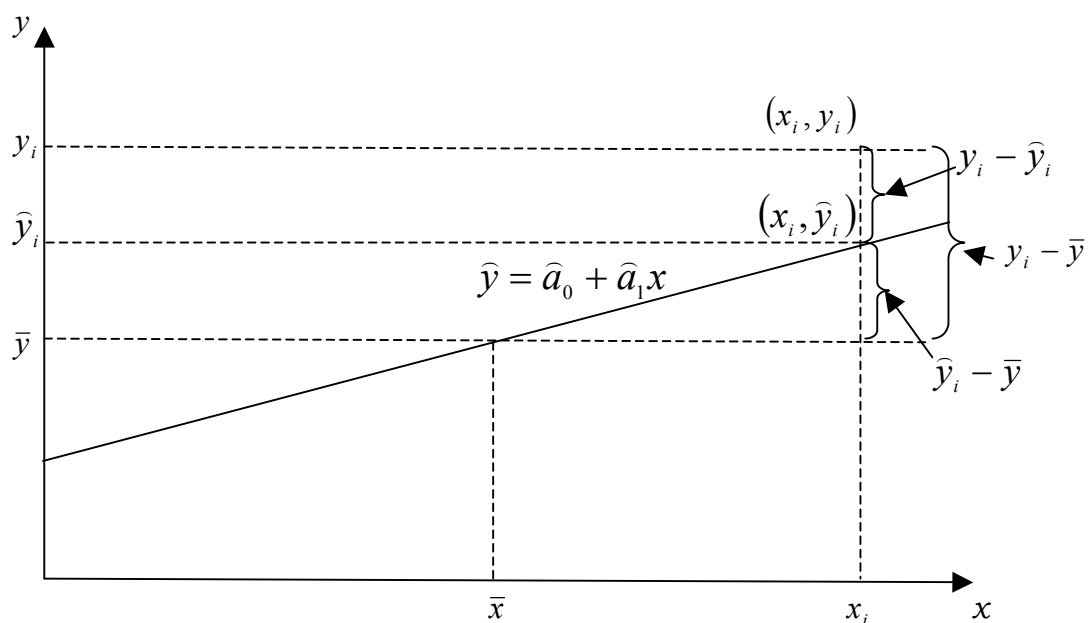


Рис. 1.5. Иллюстрация отклонений значений зависимой переменной

Отклонение от точки, лежащей на линии регрессии, определяется как $(y_i - \hat{y}_i)$, разность $(y_i - \bar{y})$ называется общим отклонением и, наконец, $(\hat{y}_i - \bar{y})$ представляет собой отклонение, связанное с линией регрессии.

Так как $\bar{y}$ – величина постоянная, то значение последнего отклонения может быть всегда найдено, исходя из уравнения регрессии, поскольку каждому значению x_i соответствует значение $\hat{y}_i$.

$(y_i - \hat{y}_i)$ – отклонение, которое нельзя объяснить исходя из регрессионной прямой (необъяснимое отклонение). Учитывая тот факт, что при изменении x_i изменяются обе величины – и y_i , и $\hat{y}_i$, то регрессионная прямая не дает возможности объяснить данное отклонение.

Таким образом, общее отклонение раскладывается на две величины: отклонение, которое объясняется исходя из линии регрессии и отклонение, которое нельзя объяснить исходя из этого, так называемое, необъяснимое отклонение.

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i).$$

Возведем в квадрат данное выражение и просуммируем по всем i

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

Рассмотрим среднее выражение в правой части

$$\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^n (\hat{a}_0 + \hat{a}_1 x_i - \hat{a}_0 - \hat{a}_1 \bar{x}) u_i = \hat{a}_1 \sum_{i=1}^n (x_i - \bar{x}) u_i.$$

Так как $\sum_{i=1}^n (x_i - \bar{x}) = 0$, то и все выражение $\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$. Отсюда имеем

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Величина $\sum_{i=1}^n (y_i - \bar{y})^2$ называется общей суммой квадратов отклонений

(SST), $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ – суммой квадратов необъяснимых отклонений (ошибок)

(SSE) и $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ – суммой квадратов отклонений, объясняемая регрессией

(SSR).

Таким образом,

$$SST = SSE + SSR.$$

Разделив обе части выражения на n , получим дисперсии

$$\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n} + \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{n}$$

или

$$\sigma_{\text{общ.}}^2 = \sigma_{\text{ош.}}^2 + \sigma_{\text{регр.}}^2,$$

где $\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n} = \sigma_{\text{общ.}}^2$ – общая дисперсия,

$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n} = \sigma_{\text{ош.}}^2$ – дисперсия ошибок,

$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{n} = \sigma_{\text{регр.}}^2$ – дисперсия, объясняемая регрессией.

Поделим обе части равенства на $\sigma_{\text{общ.}}^2$.

$$1 = \frac{\sigma_{\text{ош.}}^2}{\sigma_{\text{общ.}}^2} + \frac{\sigma_{\text{регр.}}^2}{\sigma_{\text{общ.}}^2}.$$

Первое слагаемое в правой части $\frac{\sigma_{\text{ош.}}^2}{\sigma_{\text{общ.}}^2}$ – доля ошибок в общей дисперсии, а

второе $\frac{\sigma_{\text{регр.}}^2}{\sigma_{\text{общ.}}^2}$ – составная часть дисперсии, которая объясняется регрессией.

Часть дисперсии, которая объясняет регрессию, называется **коэффициентом детерминации**.

Коэффициент детерминации обозначается через R^2 . Таким образом,

$$R^2 = \frac{\sigma_{\text{регр.}}^2}{\sigma_{\text{общ.}}^2}$$

или

$$R^2 = \frac{SSR}{SST}.$$

Величина коэффициента детерминации всегда изменяется в пределах от нуля до единицы

$$0 \leq R^2 \leq 1.$$

Как уже отмечалось, коэффициент детерминации является одним из критериев адекватности модели. Если значение R^2 близко к единице, то модель адекватна, если близко к нулю, то неадекватна. Данный коэффициент позволяет ответить на вопрос, является ли связь между переменными линейной и не влияют ли на изменение y в большей степени случайные факторы?

Связь между коэффициентом корреляции и коэффициентом детерминации

Так как рассмотренные коэффициенты – корреляции и детерминации, служат критериями качества модели, то между ними должна быть объективная связь. Рассмотрим снова коэффициент детерминации

$$R^2 = \frac{\sigma_{\text{регр.}}^2}{\sigma_{\text{общ.}}^2} = \frac{SSR}{SST}.$$

Числитель выражения представим как

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2,$$

а знаменатель как

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2.$$

Преобразуем SSR

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^n (\hat{a}_0 + \hat{a}_1 x_i - (\hat{a}_0 + \hat{a}_1 \bar{x}))^2 = \hat{a}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2.$$

Подставим полученное выражение в формулу для R^2 , умножив и числитель, и знаменатель на $\frac{1}{n}$

$$R^2 = \frac{SSR}{SST} = \frac{\hat{a}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \cdot \left(\frac{1}{n}\right)}{\sum_{i=1}^n (y_i - \bar{y})^2 \cdot \left(\frac{1}{n}\right)} = \hat{a}_1^2 \frac{\sigma_x^2}{\sigma_y^2}.$$

Так как коэффициент корреляции рассчитывается по формуле

$$r_{yx} = \hat{a}_1 \frac{\sigma_x}{\sigma_y},$$

то делаем окончательное заключение: коэффициент детерминации в уравнении парной регрессии – это квадрат коэффициента парной корреляции

$$R^2 = r_{yx}^2.$$

Используя данные примера, рассчитаем коэффициент детерминации. Для определения общей дисперсии ($\sigma_{\text{общ.}}^2$) и дисперсии, объясняющей регрессию ($\sigma_{\text{регр.}}^2$), найдем оценки переменной y ($\hat{y}_i$).

y_i	$\hat{y}_i$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$	$u_i = y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$	u_i / y_i
10	10,08152	-3,25	10,5625	-3,16848	10,03927	-0,08152	0,006645	-0,00815
6	6,624997	-7,25	52,5625	-6,625	43,89066	-0,625	0,390622	-0,10417
4	5,472824	-9,25	85,5625	-7,77718	60,48447	-1,47282	2,169209	-0,36821
12	10,08152	-1,25	1,5625	-3,16848	10,03927	1,918481	3,680569	0,159873
18	17,57065	4,75	22,5625	4,32065	18,66801	0,42935	0,184342	0,023853
19	20,45108	5,75	33,0625	7,201084	51,85562	-1,45108	2,105646	-0,07637
22	22,17935	8,75	76,5625	8,929345	79,73321	-0,17935	0,032165	-0,00815
15	13,53804	1,75	3,0625	0,288041	0,082968	1,461959	2,137324	0,097464
Сумма			285,5		274,7935		10,70652	-0,28386

Общая дисперсия составит

$$\sigma_{\text{общ.}}^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n} = \frac{285,5}{8} = 35,6875.$$

Дисперсия, объясняющая регрессию равна

$$\sigma_{\text{регр.}}^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{n} = \frac{274,7935}{8} = 34,34919.$$

С учетом этого находим коэффициент детерминации

$$R^2 = \frac{\sigma_{\text{регр.}}^2}{\sigma_{\text{общ.}}^2} = \frac{34,34919}{35,6875} = 0,9625.$$

Значение коэффициента детерминации позволяет сделать вывод о том, что 96,25% вариации зависимой переменной (величины товарооборота) объясняется вариацией независимой переменной (торговой площадью).

Полученное значение коэффициента детерминации можно было бы получить, зная коэффициент корреляции

$$R^2 = r_{yx}^2 = (0,981)^2 = 0,9625.$$

Для определения доли вариации y за счет неучтенных в модели факторов рассчитывается **коэффициент остаточной детерминации**

$$1 - R^2 = 1 - 0,9625 = 0,0375.$$

Таким образом, 3.75% вариации товарооборота объясняется неучтенными в модели факторами.

Степени свободы для сумм квадратов

В анализе регрессионной модели и оценке ее адекватности реальному процессу, а также адекватности ее параметров, важным понятием является понятие числа степеней свободы.

Число степеней свободы некоторой величины – это разность между числом ее исследований (наблюдений) и числом констант, которые установлены (найденны) в результате этих исследований, независимо друг от друга.

Применительно к общим суммам квадратов (SST, SSR, SSE), которые использовались при расчете коэффициента детерминации, **степень свободы** ($DF - Degree of Freedom$) – число, которое показывает, сколько независимых элементов информации, полученных из элементов $y_1, y_2, \dots, y_n$, необходимо для расчета данной суммы квадратов.

Найдем число степеней свободы для каждой из составляющих выражения

$$SST = SSE + SSR.$$

Для нахождения SST рассматриваются разности $y_1 - \bar{y}, y_2 - \bar{y}, \dots, y_n - \bar{y}$. Их сумма равна нулю: $\sum_{i=1}^n (y_i - \bar{y}) = 0$, что говорит о зависимости всех разностей. Если одну (любую) исключим из суммы, то получим выражение, не равное нулю: $\sum_{i=1}^{n-1} (y_i - \bar{y}) \neq 0$. То есть $n-1$ разность независима. Таким образом, для расчета SST необходимо $n-1$ независимое число разностей. Отсюда $n-1$ является числом степеней свободы для SST : $DF_{SST} = n - 1$.

Для нахождения SSR необходимо иметь только один параметр $\hat{a}_1$:

$$\hat{y}_i - \bar{y} = \hat{a}_0 + \hat{a}_1 x_i - \hat{a}_0 - \hat{a}_1 \bar{x} = \hat{a}_1 (x_i - \bar{x}),$$

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \hat{a}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2.$$

Отсюда SSR можно получить, используя одну единицу независимой информации, т.е. $\hat{a}_1 : DF_{SSR} = 1$.

Число степеней свободы в данном случае совпадает с числом независимых переменных.

Наконец для SSE имеем $DF_{SSE} = n - 2$, т.е. число степеней свободы равно разности между числом наблюдений n и числом оцениваемых параметров $\hat{a}_0$ и $\hat{a}_1$ (в дальнейшем число степеней свободы будем обозначать буквой ν).

Для простой линейной регрессии можно записать

$$n - 1 = 1 + (n - 2).$$

Как уже отмечалось, степени свободы используются в анализе дисперсий и при проверке адекватности модели. Для этих целей определим понятие среднего квадрата.

Средний квадрат – сумма квадратов, поделенная на соответствующее число степеней свободы.

Отсюда, средний квадрат ошибок (MSE) будет равен:

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2},$$

а средний квадрат, объясняющий регрессию (MSR):

$$MSR = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{1} = SSR.$$

Для общей суммы квадратов средний квадрат не рассчитывается.

Проверка адекватности модели и ее параметров

Проверка адекватности модели по F -критерию Фишера

Чтобы быть уверенным в том, что построенная модель действительно пригодна для анализа исследуемого процесса или явления, необходимо дать оценку как самой модели и входящих в нее параметров, так и сопутствующих ей показателей.

Начнем с того, что оценим адекватность самой модели. Для этих целей применим F -критерий Фишера. Из курса теории вероятностей известно, что величина

$$F = \frac{MSR}{MSE}$$

имеет распределение Фишера с $\nu_1 = 1$ и $\nu_2 = n - 2$ степенями свободы при условии, что наклон линии регрессии равен нулю, т.е. $\hat{a}_1 = 0$.

Если это действительно так, то выражение для $\hat{y}_i$ может быть записано следующим образом

$$\hat{y}_i = \hat{a}_0 + \hat{a}_1 x_i = \bar{y} - \hat{a}_1 \bar{x} + \hat{a}_1 x_i = \bar{y} + \hat{a}_1 (x_i - \bar{x}) = \bar{y}.$$

С учетом полученного результата данные лучше аппроксимировать на основе среднего значения y , т.е. $\bar{y}$, а не на основе уравнения регрессии

$$\hat{y}_i = \hat{a}_0 + \hat{a}_1 x_i.$$

И, следовательно, линейная связь между x и y отсутствует.

Таким образом, F -критерий служит для проверки нулевой гипотезы $H_0 : \hat{a}_1 = 0$, что наклон прямой равен нулю ($\hat{a}_1 = 0$). Если данная гипотеза подтверждается, то рассматриваемые данные лучше аппроксимировать с помощью средней величины ($\hat{y}_i = \bar{y}$).

Этапы проверки:

1. Рассчитываем величину F -критерия:

$$F_{\text{расч.}} = \frac{MSR}{MSE};$$

2. Задаем уровень значимости α – в скольких случаях из 100 может быть допущена ошибка (чаще всего используются 5% и 1% уровни значимости);
3. Находим по таблицам F -распределения Фишера с $(1, n - 2)$ степенями свободы и уровнем значимости α табличное (критическое) значение F -критерия ($F_{(1, n-2, \alpha)}$);
4. Если рассчитанное значение F -критерия превышает табличное ($F_{\text{расч.}} > F_{(1, n-2, \alpha)}$), то гипотеза о равенстве $\hat{a}_1$ нулю отвергается, т.е. модель адекватна, если ($F_{\text{расч.}} \leq F_{(1, n-2, \alpha)}$), то гипотеза принимается.

Проверим значимость уравнения регрессии по критерию Фишера. Рассчитаем F -статистику Фишера:

$$F_{\text{расч.}} = \frac{MSR}{MSE}.$$

Средний квадрат, объясняющий регрессию (MSR)

$$MSR = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{1} = \frac{274,7935}{1} = 274,7935.$$

Средний квадрат ошибок (MSE)

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2} = \frac{10,70652}{8 - 2} = 1,78442.$$

С учетом этого F -статистика Фишера

$$F_{\text{расч.}} = \frac{MSR}{MSE} = \frac{274,7935}{1,78442} = 153,996.$$

Зададимся уровнем значимости $\alpha = 0,05$ и найдем табличное значение критерия Фишера для числа степеней свободы $\nu_1 = 1$ и $\nu_2 = n - 2 = 8 - 2 = 6$. Оно равно $F_{(1,6,0.05)} = 5,99$. Так как

$$F_{\text{расч.}} = 153,996 > F_{(1,6,0.05)} = 5,99,$$

то делаем вывод о значимости уравнения регрессии.

Проверка гипотезы о значимости коэффициента корреляции

Следующим этапом оценки качества модели и ее параметров является проверка значимости коэффициента корреляции. Данная процедура необходима для того, чтобы сделать заключение: имеется или отсутствует связь между независимой и зависимой переменными.

С этой целью выдвигается гипотеза $H_0 : r_{yx} = 0$ – коэффициент корреляции равен нулю и связь отсутствует.

Для проверки гипотезы рассчитывается t -статистика Стьюдента:

$$t_{\text{расч.}} = r_{yx} \cdot \sqrt{\frac{n-2}{1-r_{yx}^2}} \quad \text{или} \quad t_{\text{расч.}} = r_{yx} \cdot \sqrt{\frac{n-2}{1-R^2}},$$

где $n - 2$ – число степеней свободы.

Для заданного уровня значимости α и числа степеней свободы ν находится табличное значение $t_{(\nu, \alpha)}$. Если

$$|t_{\text{расч.}}| > t_{(\nu, \alpha)},$$

то гипотеза $H_0 : r_{yx} = 0$ об отсутствии корреляционной связи между переменными отвергается, если $|t_{\text{расч.}}| \leq t_{(\nu, \alpha)}$ – то принимается.

Проверим гипотезу о значимости коэффициента корреляции с помощью критерия Стьюдента. Рассчитаем t -статистику Стьюдента

$$t_{\text{расч.}} = r_{yx} \cdot \sqrt{\frac{n-2}{1-r_{yx}^2}} = 0,981 \cdot \sqrt{\frac{8-2}{1-0,9625}} = 12,4096482,$$

зададимся уровнем значимости $\alpha = 0,05$. Табличное значение критерия Стьюдента для числа степеней свободы $\nu = n - 2 = 8 - 2 = 6$ и уровня значимости $\alpha = 0,05$ равно $t_{(6, 0,05)} = 2,227$.

Так как рассчитанное нами значение критерия Стьюдента больше табличного

$$t_{\text{расч.}} = 12,4 > t_{(6, 0,05)} = 2,227,$$

то делаем вывод о значимости коэффициента корреляции.

Критерии оценки качества линейной модели

Наряду с рассмотренными выше для оценки качества модели используются и другие критерии. Вот некоторые из них:

1. Средняя ошибка прогноза

$$ME = \bar{u} = \frac{1}{n} \sum_{i=1}^n u_i.$$

Величина $ME \rightarrow 0$ при $n \rightarrow \infty$.

2. Дисперсия и стандартное отклонение ошибок

$$\text{var}(u) = \frac{1}{n} \sum_{i=1}^n (u_i - \bar{u})^2 = \sigma_u^2,$$

$$\sigma = \sqrt{\text{var}(u)} = \sqrt{\frac{1}{n} \sum_{i=1}^n (u_i - \bar{u})^2}.$$

3. Среднее абсолютное отклонение

$$MAD(u) = \frac{1}{n} \sum_{i=1}^n |u_i - \bar{u}|.$$

4. Средний квадрат ошибки

$$MSE(u) = \frac{1}{n} \sum_{i=1}^n u_i^2.$$

Иногда вместо данного показателя используется сумма квадратов ошибок

$$SSE = \sum_{i=1}^n u_i^2.$$

5. Средняя абсолютная процентная ошибка

$$MAPE = \frac{1}{n} \sum_{i=1}^n \frac{|u_i|}{y_i} 100\%.$$

6. Средняя процентная ошибка (аппроксимации)

$$MPE = \frac{1}{n} \sum_{i=1}^n \frac{u_i}{y_i} 100\%.$$

7. Средняя абсолютная ошибка

$$MAE(u) = \frac{1}{n} \sum_{i=1}^n |u_i|$$

Для условий нашего примера рассчитаем среднюю процентную ошибку аппроксимации

$$MPE = \frac{1}{n} \sum_{i=1}^n \frac{u_i}{y_i} \cdot 100\% = \frac{-0,28386}{8} \cdot 100\% = -3,54824\%.$$

В процентном отношении по абсолютной величине ошибка невелика. Таким образом, можно сделать вывод о хорошо подобранной модели.

Как уже отмечалось, для оценки достоверности модели используются дисперсии остатков σ_u^2 . Несмещенная оценка истинного значения σ_u^2 для парной зависимости рассчитывается по той же формуле, что и MSE

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^n u_i^2}{n-2}.$$

Для многофакторной модели

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^n u_i^2}{n-k},$$

где $n-2$ и $n-k$ – число степеней свободы.

Для определения уровня корреляции между наблюдениями y_i и рассчитанными по модели значениями $\hat{y}_i$, используется **индекс корреляции** (применяется также для проверки линейности связи между y и x)

$$R = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}} = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}} = \sqrt{1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}} = \sqrt{1 - \frac{SSE}{SST}}$$

$$R \leq 1$$

Если все $u_i^2 = 0$, то теоретические значения $\hat{y}_i$ и значения наблюдений y_i сходятся.

Если все $\hat{y}_i = \bar{y}$, то изменения y не связаны с изменениями x и $R = 0$.

Надежные границы индекса корреляции определяются через оценки среднеквадратического отклонения

$$\sigma_R = \frac{1-R}{\sqrt{n}}, \quad \Delta R = \pm t_\alpha \sigma_R.$$

Условия применения метода наименьших квадратов

1. Математическое ожидание остатков (ошибок) равно нулю

$$M(u_i) = 0.$$

Факторы, не включенные в модель, не влияют систематически на математическое ожидание y .

2. Случайные величины u_i имеют одинаковую дисперсию для всех $i = \overline{1, n}$ (гомоскедастичность)

$$\text{var}(u_i) = \sigma_u^2, \quad i = \overline{1, n}.$$

3. Случайная величина u распределена по нормальному закону с математическим ожиданием, равным нулю и постоянной дисперсией

$$u \sim N(0, \sigma_u^2).$$

4. Между случайными величинами u отсутствует автокорреляция

$$\text{cov}(u_i, u_j) = M \{ [u_i - M(u_i)] \cdot [u_j - M(u_j)] \} = M(u_i u_j) = 0, \quad i \neq j.$$

Это означает, что случайные величины должны быть независимыми между собой, или, что то же самое, корреляция между u_i и u_j отсутствует.

5. Значения независимых переменных x_i не зависят от значений случайных величин u_i , т.е. имеет место нулевая ковариация между ними

$$\begin{aligned} \text{cov}(u_i, x_i) &= M \{ [u_i - M(u_i)] \cdot [x_i - M(x_i)] \} = \\ &= M \{ u_i [x_i - M(x_i)] \} = \\ &= M(u_i x_i) - M(u_i) M(x_i) = \\ &= M(u_i x_i) = 0, \end{aligned}$$

так как $M(u_i) = 0$ и $M(x_i) = \text{const}$.

6. Независимые переменные модели создают линейно независимую систему векторов, т.е. независимые переменные не должны быть мультиколлинеарными.

7. Регрессионная модель построена (специфицирована) правильно (отсутствует ошибка): правильно подобрана линия регрессии, в нее включены только существенно влияющие переменные, модель адекватна экономическому процессу.

8. Количество наблюдений больше числа независимых переменных (регрессоров) ($n > m$).

9. Регрессоры – детерминированные величины.

10. Набор переменных X не содержит ошибок наблюдений.

11. Набор переменных X содержит все существенно влияющие на регрессант факторы.

12. Коэффициенты регрессии – детерминированные величины, неизменные во времени и пространстве.

Распределение независимой переменной y

Единственным источником изменения y является случайная величина u . Естественно, что и распределение переменной y зависит от распределения случайной величины u , т.е. является нормальным.

Математическое ожидание переменной y определяется как

$$M(y_i) = M(a_0 + a_1 x_i + u_i) = M(a_0 + a_1 x_i) + M(u_i).$$

А так как $M(u_i) = 0$ и, учитывая, что a_0 и a_1 – константы и x_i принимают фиксированные значения, то

$$M(y_i) = M(a_0 + a_1 x_i) = a_0 + a_1 x_i.$$

Дисперсия зависимой переменной равняется дисперсии случайной величины u

$$D(y_i) = M[y_i - M(y_i)]^2 = \sigma_u^2.$$

Здесь и далее в расчетах используется несмещенная оценка дисперсии $\hat{\sigma}_u^2$.

Математическое ожидание и дисперсия распределения параметров a_0 и a_1

Параметры a_0 и a_1 можно выразить как линейные функции от случайной величины u . Из предположения 3 о нормальности распределения остатков u вытекает нормальность распределения параметров a_0 и a_1 .

Найдем математическое ожидание и дисперсию для a_0 :

$$M(a_0) = \hat{a}_0,$$

$$D(a_0) = M[a_0 - M(a_0)]^2 = M(a_0 - \hat{a}_0)^2 = \sigma_u^2 \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}.$$

Оценка дисперсии a_0 :

$$\hat{\sigma}_{a_0}^2 = \hat{\sigma}_u^2 \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}.$$

Аналогично рассчитаем данные показатели для a_1 :

$$M(a_1) = \hat{a}_1,$$

$$D(a_1) = M[a_1 - M(a_1)]^2 = M(a_1 - \hat{a}_1)^2 = \sigma_u^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Оценка дисперсии a_1 :

$$\hat{\sigma}_{a_1}^2 = \hat{\sigma}_u^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

Рассчитанные значения дисперсий параметров a_0 и a_1 используются для проверки их значимости и расчета доверительных интервалов.

Проверка значимости параметров a_0 и a_1 по критерию Стьюдента и построение доверительных интервалов

Для проверки значимости параметров модели рассчитывают t -статистику Стьюдента (t_{p_i}) для каждого из них по следующей формуле:

$$t_{p_i} = \frac{p_i - p_i^*}{\sqrt{\hat{\sigma}_{p_i}^2}}$$

с $(n - k)$ степенями свободы, где

p_i – оценка рассматриваемого i -го параметра;

p_i^* – гипотетическое значение, которое должен принять параметр p_i ($H_0 : p_i = p_i^*$);

$\hat{\sigma}_{p_i}^2$ – оценка дисперсии параметра p_i (по регрессии);

n – размер выборки;

k – число оцениваемых параметров модели (для парной регрессии $k = 2$).

Для оценки параметра $\hat{a}_1$ выдвигается гипотеза $H_0 : \hat{a}_1 = 0$. t -статистика Стьюдента принимает вид:

$$t_{\hat{a}_1} = \frac{\hat{a}_1}{\hat{\sigma}_{a_1}}.$$

Сравниваем рассчитанное значение $t_{\hat{a}_1}$ с табличным ($t_{(n-k, \alpha)}$) для числа степеней свободы $(n - k)$ и уровня значимости α .

Если $|t_{\hat{a}_1}| > t_{(n-k, \alpha)}$, то гипотеза $H_0 : \hat{a}_1 = 0$ отвергается, если $|t_{\hat{a}_1}| \leq t_{(n-k, \alpha)}$, то принимается.

Расчетная формула для оценки значимости параметра $\hat{a}_0$ аналогична:

$$t_{\hat{a}_0} = \frac{\hat{a}_0}{\hat{\sigma}_{a_0}}.$$

Оценим значимость параметров модели $\hat{a}_1$ и $\hat{a}_0$. Для этого сначала рассчитаем оценку дисперсии ошибок $\hat{\sigma}_u^2$:

$$\hat{\sigma}_u^2 = \frac{\sum_{i=1}^n u_i^2}{n-2} = \frac{10,70652}{8-2} = 1,78442.$$

Найдем оценку дисперсии a_1 ($\hat{\sigma}_{a_1}^2$) и среднеквадратическое отклонение ($\hat{\sigma}_{a_1}$):

$$\hat{\sigma}_{a_1}^2 = \hat{\sigma}_u^2 \cdot \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} = 1,78442 \cdot \frac{1}{828} = 0,002155,$$

$$\hat{\sigma}_{a_1} = \sqrt{\hat{\sigma}_{a_1}^2} = 0,046423.$$

Рассчитаем t -статистику Стьюдента:

$$t_{\hat{a}_1} = \frac{\hat{a}_1}{\hat{\sigma}_{a_1}} = \frac{0,57608696}{0,046423} = 12,40951.$$

Как и в случае с коэффициентом корреляции, найдем для $\alpha = 0,05$ и $n-2 = 8-2 = 6$ табличное значение критерия Стьюдента и сравним рассчитанное значение t -статистики с табличным:

$$t_{\hat{a}_1} = 12,4 > t_{(6,0.05)} = 2,227.$$

Таким образом, значение коэффициента $\hat{a}_1$ является значимым.

Найдем оценку дисперсии a_0 ($\hat{\sigma}_{a_0}^2$) и среднеквадратическое отклонение ($\hat{\sigma}_{a_0}$):

$$\hat{\sigma}_{a_0}^2 = \hat{\sigma}_u^2 \cdot \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} = 1,78442 \cdot \frac{6030}{8 \cdot 828} = 1,624404,$$

$$\hat{\sigma}_{a_0} = \sqrt{\hat{\sigma}_{a_0}^2} = 1,274521.$$

Рассчитаем t -статистику Стьюдента:

$$t_{a_0} = \frac{\widehat{a}_0}{\widehat{\sigma}_{a_0}} = \frac{-1,44022}{1,274521} = -1,13001.$$

Сравним рассчитанное значение t -статистики по модулю с табличным для $\alpha = 0,05$ и $\nu = n - 2 = 8 - 2 = 6$:

$$|t_{\widehat{a}_0}| = |-1,13001| < t_{(6,0,05)} = 2,227.$$

Таким образом, для выбранного уровня доверия α значение коэффициента $\widehat{a}_0$ не является значимым.

Расчет доверительных интервалов для параметров осуществляется в несколько этапов:

1. Рассчитываем t -статистику (t_{p_i}) Стьюдента:

$$t_{p_i} = \frac{p_i - p_i^*}{\sqrt{\widehat{\sigma}_{p_i}^2}};$$

2. Выбираем уровень значимости α ;

3. По таблице критерия Стьюдента находим значения $\pm t_{\alpha/2}$ с $(n - k)$ степенями свободы (для парной зависимости $k = 2$). С учетом этого получаем:

$$P(-t_{\alpha/2} \leq t_{p_i} \leq t_{\alpha/2}) = 1 - \alpha$$

или

$$P\left(-t_{\alpha/2} \leq \frac{p_i - p_i^*}{\sqrt{\widehat{\sigma}_{p_i}^2}} \leq t_{\alpha/2}\right) = 1 - \alpha.$$

Отсюда имеем:

$$p_i^* = p_i \pm t_{\alpha/2} \cdot \widehat{\sigma}_{p_i}.$$

Для параметра $\widehat{a}_1$ получаем:

$$a_1^* = \widehat{a}_1 \pm t_{\alpha/2} \cdot \widehat{\sigma}_{a_1}.$$

Для параметра $\widehat{a}_0$:

$$a_0^* = \widehat{a}_0 \pm t_{\alpha/2} \cdot \widehat{\sigma}_{a_0}.$$

Найдем доверительные интервалы для параметров $\widehat{a}_1$ и $\widehat{a}_0$.

Для параметра $\widehat{a}_1$ имеем:

$$a_1^* = \widehat{a}_1 \pm t_{\alpha/2} \cdot \widehat{\sigma}_{a_1}.$$

Пусть $\alpha = 0,05$. Для $\nu = n - 2 = 8 - 2 = 6$ степеней свободы находим табличное значение $t_{\alpha/2} = t_{0,025} = 2,227$. С учетом этого

$$a_1^* = \hat{a}_1 \pm t_{0,025} \cdot \hat{\sigma}_{a_1} = 0,57608696 \pm 2,227 \cdot 0,046423 = 0,57608696 \pm 0,103384.$$

Таким образом, значение $\hat{a}_1$ находится в пределах между 0,472703 и 0,679471.

Аналогичные действия проделываем для $\hat{a}_0$:

$$a_0^* = \hat{a}_0 \pm t_{0,025} \cdot \hat{\sigma}_{a_0} = -1,44022 \pm 2,227 \cdot 1,274521 = -1,44022 \pm 2,838358.$$

Как видим, диапазон изменения параметра $\hat{a}_0$ достаточно большой: от -4,27836 до 1,398358.

Прогнозирование по модели простой линейной регрессии

Различают точечный и интервальный прогнозы зависимой переменной.

Прогнозную величину $\hat{y}_{n+1}$ получаем из уравнения регрессии, подставляя в него значение переменной x_i для $i = n + 1$, т.е. x_{n+1} :

$$\hat{y}_{n+1} = \hat{a}_0 + \hat{a}_1 x_{n+1}.$$

Расчет прогнозной величины $\hat{y}_{n+1}$ сопровождается расчетом ошибки прогноза u_{n+1} и ее дисперсии (оценки дисперсии).

Значение ошибки прогноза u_{n+1} , равное

$$u_{n+1} = y_{n+1} - \hat{y}_{n+1}$$

позволяет рассчитать ее дисперсию:

$$\sigma_{u_{n+1}}^2 = \sigma_u^2 \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right].$$

Оценка дисперсии $\hat{\sigma}_{u_{n+1}}^2$, определяемая из выражения

$$\hat{\sigma}_{u_{n+1}}^2 = \hat{\sigma}_u^2 \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right],$$

позволяет рассчитать доверительные интервалы прогноза. t -статистика Стьюдента для переменной $\hat{y}_{n+1}$ определяется как

$$t_{y_{n+1}} = \frac{\hat{y}_{n+1} - y_{n+1}^*}{\hat{\sigma}_u \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}}.$$

Исходя из этого, имеем доверительный интервал для прогнозного $(n+1)$ -го значения переменной y :

$$y_{n+1}^* = \hat{y}_{n+1} \pm t_{\alpha/2} \cdot \hat{\sigma}_u \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

На практике построение доверительных интервалов чаще осуществляется для математического ожидания $\hat{y}_{n+1}$ (так называемого, точечного прогноза):

$$M(y_{n+1}) = a_0 + a_1 x_{n+1}.$$

Величина ошибки равна:

$$u_{n+1} = M(y_{n+1}) - \hat{y}_{n+1}$$

Дисперсия ошибок в этом случае определяется следующим образом:

$$\sigma_{u_{n+1}}^2 = \sigma_u^2 \left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right],$$

а ее оценка

$$\hat{\sigma}_{u_{n+1}}^2 = \hat{\sigma}_u^2 \left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right].$$

С учетом этого выражение для расчета доверительного интервала немного видоизменяется:

$$M(y_{n+1}^*) = \hat{y}_{n+1} \pm t_{\alpha/2} \cdot \hat{\sigma}_u \sqrt{\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

Рассчитаем прогноз величины $\hat{y}_{n+1}$ и доверительные интервалы прогноза для $n+1=8+1=9$. Для этого определим значение $x_9 = 20$.

Найдем $\hat{y}_9$:

$$\hat{y}_9 = -1,44022 + 0,57608696 \cdot x_9 = -1,44022 + 0,57608696 \cdot 20 = 10,0815192.$$

Определим доверительные интервалы для уровня значимости $\alpha/2 = 0,025$. Так, для интервального прогноза имеем

$$\begin{aligned} y_9^* &= \hat{y}_9 \pm t_{\alpha/2} \cdot \hat{\sigma}_u \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_9 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} = \\ &= 10,0815192 \pm 2,227 \cdot 1,335821844 \cdot \sqrt{1 + \frac{1}{8} + \frac{(20 - 25,5)^2}{828}} = \\ &= 10,0815192 \pm 3,206156 \end{aligned}$$

Значение $\hat{y}_9$ в соответствии с произведенными расчетами находится в пределах между 6,875363 и 13,28768.

Аналогично произведем вычисления доверительных интервалов для математического ожидания $\hat{y}_9$ (для $\alpha/2 = 0,025$):

$$\begin{aligned} M(y_9^*) &= \hat{y}_9 \pm t_{\alpha/2} \cdot \hat{\sigma}_u \cdot \sqrt{\frac{1}{n} + \frac{(x_9 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} = \\ &= 10,0815192 \pm 2,227 \cdot 1,335821844 \cdot \sqrt{\frac{1}{8} + \frac{(20 - 25,5)^2}{828}} = \\ &= 10,0815192 \pm 1,19564. \end{aligned}$$

Таким образом, истинное значение $\hat{y}_9$ находится в пределах между 8,885879 и 11,27715. Можно заметить, что доверительный интервал для математического ожидания переменной $\hat{y}_9$ имеет меньший диапазон, чем для отдельной точки.