Министерство образования и науки Украины

ХАРЬКОВСКИЙ НАЦИОНАЛЬНЫЙ АВТОМОБИЛЬНО-ДОРОЖНЫЙ УНИВЕРСИТЕТ

ОПОРНЫЙ КОНСПЕКТ ЛЕКЦИЙ

по дисциплине
«Экономико-математические методы и модели (Эконометрика)»

для иностранных студентов всех форм обучения
отрасли знаний 0305 – «Экономика и предпринимательство»
направления подготовки 6.030503 – «Международная экономика»

Харьков ХНАДУ 20__

Составитель: Е.П. Болдовская

Кафедра международной экономики

ТЕМА 1
СУЩНОСТЬ И ОСНОВНЫЕ ЗАДАЧИ ЭКОНОМЕТРИИ

1.1 Понятие и предмет эконометрии

В условиях жесткой конкуренции для успешного функционирования крупные компании, предприятия, банки и т.д. испытывают потребность в анализе имеющейся информации и получении обоснованных прогнозов и выводов,

«Эконометрия» как наука возникла на рубеже XIX-XX вв. Ее породила потребность в изучении и классификации изменяющейся экономической действительности.

Буквальный перевод слова «эконометрия» означает «измерение экономики».

Эконометрия – это наука, которая изучает количественные закономерности и взаимосвязи социально-экономических процессов и объектов с помощью математико-статистических методов и моделей.

Эконометрия является инструментом, позволяющим перейти от качественного уровня анализа до уровня, использующего количественные статистические значения исследуемых величин.

Целью дисциплины является формирование системы знаний о методах оценки параметров зависимостей, характеризующих количественные взаимосвязи между экономическими величинами.

Т.о. основной задачей эконометрии является восстановление неизвестных экономико-математических зависимостей по статистическим данным и рассмотрение возможности использования этих моделей в экономических исследованиях.

Модель – это искусственное воссоздание некоторого процесса (экономического) для исследований. В эконометрии под моделью подразумевают математическую модель, т.е. описание экономического процесса с помощью

математических формул.

Эконометрические модели количественно описывают связь между входными факторами экономической системы X и результирующим показателем Y плюс влияние случайной компоненты ε.

Эконометрическое моделирование реальных социально-экономических процессов и систем, как правило, направлено на достижение двух типов конечных прикладных результатов:

- получение прогноза экономических показателей, характеризующих состояние и развитие экономической системы. **Прогноз** – это расчет неизвестного показателя по заданным факторам на основе модели;
- имитирование различных возможных сценариев социально-экономического развития экономической системы.

1.2 Этапы и задачи эконометрического анализа

Классически считается, что эконометрический анализ состоит из таких этапов:

1) Идентификация переменных; сбор и подготовка экономической информации (формирование совокупности наблюдений) – выбор правильных обозначений и единиц измерения для переменных является очень важным при проведении исследования. Например, если речь идет об изменении дохода с течением времени, то функция будет иметь вид $y = f(t)$, где y – доход, t – время. Если речь идет о национальном доходе, то в качестве единиц измерения обычно принимают:

   y – млн. грн., t – год; если речь идет о предприятии – y – грн., t – месяц.
2) Формулирование гипотезы о виде зависимости по статистическим данным в соответствии с набором факторов и разработка (построение) предварительной модели по выбранной зависимости для проверки выдвинутой гипотезы – т.е. выбор конкретного вида функции $y = f(x)$ для некоторого экономического процесса (спецификация модели): необходимо, чтобы все функциональные связи входили в модель в явном виде. Для этого эконометрия может идти путем

от

простого к сложному: начав с самых простых функций, вводить и проверять различные гипотезы и постепенно усложнять характер функциональных связей, исходя из реальных данных. Например, зависимость между доходом и расходом можно описать так: $y = b_0 + b_1 x$.

3) Оценка всех неизвестных параметров выбранной модели на основе имеющихся статистических данных и расчет доверительных интервалов (интервалов, в которые с заданной вероятностью попадет исчисляемая величина).
4) Проверка модели на адекватность, статистические выводы – оценка статистической значимости и качества выбранной модели, т.е. ее простоты, точности описания данных с помощью различных тестов и критериев, а также тестирование

модель с помощью уже известных параметров и сбывшихся прогнозов.

5) Экономический анализ и использование прошедшей проверку модели в экономических исследованиях для принятия решений при помощи сформированных на ее основе прогнозах.

1.2 Классификация эконометрических моделей и информационная база

Множество существующих моделей традиционно относят к 2 видам:

1. Однофакторные $y = f(x)$.

   1.1. Линейные вида $y = b_0 + b_1 x$;

   1.2. Нелинейные:

      а) сводящиеся к линейным;

      б) существенно нелинейные.

2. Многофакторные $y = f(x_1, x_2, ..., x_p)$.

   2.1. Линейные вида $y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p$;

   2.2. Нелинейные:

      а) сводящиеся к линейным;

      б) существенно нелинейные.

Информационная база эконометрии.

Решение задач эконометрии проводится на базе статистических данных.

Статистические данные – это данные, собранные на реальных экономических объектах, представляющие собой набор определенных характеристик объекта, отмеченных за определенный период времени.

При измерении количественных признаков (характеристик) могут быть получены 2 типа статистических данных:

- динамические (временные) ряды;
- вариационные (перекрестные) ряды.

Временные ряды – это последовательность наблюдений за одним и тем же процессом или явлением в различные (как правило, равные) промежутки времени. Например, данные о динамике уровня инфляции за определенный период. Отдельные значения признака, относящиеся к определенным периодам времени, называют уровнями динамического ряда.

При формировании динамических рядов могут возникать трудности, связанные с недостатком необходимых данных. Один из наиболее распространенных способов его преодоления – выявление закономерностей, которым подчиняется динамический ряд (такие закономерности описывают при помощи математических уравнений) с последующей интерполяцией или экстраполяцией недостающих его уровней.

Интерполяция – нахождение неизвестных показателей (уровней ряда) в рамках имеющегося ряда динамики.

Экстраполяция – нахождение неизвестных показателей за рамками ряда динамики, т.е. в конце или в начале.

Вариационные ряды – последовательность наблюдений по какому-либо экономическому показателю для разных однотипных процессов или объектов. Например, данные об успеваемости всех студентов третьего курса. Все замеры (наблюдения) производятся в одно и то же время. Значения вариационного ряда располагают в порядке возрастания. Отдельные значения признака, относящиеся к определенным объектам наблюдения, называют вариантами.

Обработка информационных данных

Совокупность данных динамических и вариационных рядов обрабатывается по правилам, разработанным в математической статистике.

Генеральная совокупность (N) – все возможные реализации интересующего нас показателя (признака).

На практике мы наблюдаем случайно выбранные значения этого показателя (**выборка**). По генеральной совокупности можно получить точные значения параметров, по выборке – приближенные, или оценки.

Объем выборки (n) – суммарное количество наблюдений.

Объемы выборок могут быть:

- небольшими (n<10),
- большими (n~100)
- очень большими (n~10^4).

Чтоб получить реальный прогноз, необходимо работать с очень большими выборками (а на практике так чаще всего и происходит), поэтому широкое распространение получило использование компьютерной техники в расчетах.

Во всех случаях всю совокупность выборочных данных $x_i\ (i=1...n)$ стараются охарактеризовать некоторыми усредненными параметрами, которые учитывают особенности выборки. По выборкам производится расчет **основных статистических характеристик**:

1. Среднее значение – обобщающая мера вариационного признака, характеризующая его типовой уровень в расчете на единицу однородной совокупности:

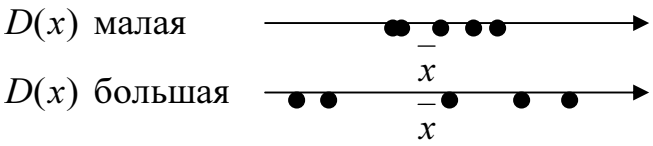
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i\ .$$

2. Вариация (дисперсия) – показатель среднего квадрата отклонения вариант признака от среднего значения.

$$\mathrm{var}(x) = D(x) = \sigma_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}.$$

Дисперсия характеризует, как сильно рассеяны значения выборки

относительно среднего значения.



3. Среднеквадратическое (стандартное) отклонение:

$$\sigma_x = \sqrt{D} = \sqrt{\sigma_x^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}.$$

Эта величина характеризует отклонение выборочных значений в среднем от $\bar{x}$.

4. Вариационный размах – разница между наибольшим и наименьшим значениями вариант:

$$R_x = x_{\max} - x_{\min}.$$

5. Коэффициент вариации:

$$V_x = \frac{\sigma_x}{\bar{x}} \cdot 100\%.$$

Коэффициент вариации дает возможность:

- сравнить вариацию одного и того же признака в разных группах объектов;
- выявить степень отличия одного признака в одной группе за разные промежутки времени;
- сравнить вариацию разных признаков в одинаковых группах объектов.

ТЕМА 2
ПРОСТАЯ ВЫБОРОЧНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ. ПОДБОР ПАРАМЕТРОВ ЛИНЕЙНОЙ РЕГРЕССИИ ПО МЕТОДУ НАИМЕНЬШИХ КВАДРАТОВ (МНК).

2.1 Однофакторная линейная регрессия

Изучение зависимостей экономических показателей начнем со случая двух переменных: независимой (факторной, объясняющей) переменной X и зависимой (результативной) переменной Y:

$$Y = f(X). \qquad (2.1)$$

Этот метод наиболее прост и может быть представлен графически.

Для начала нужно установить, связаны ли эти переменные между собой, и, если да, то определить форму, направление и тесноту (силу) связи.

Для анализа данные представляют в виде таблицы:

X	Y
x_1	y_1
x_2	y_2
…	…
x_n	y_n

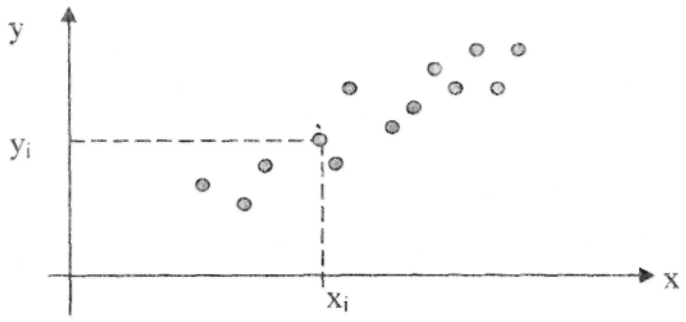


Рис. 2.1.

По таблице строим корреляционное поле (диаграмму рассеивания).

Корреляционным полем (диаграммой рассеивания) будем называть систему точек (x_i, y_i) $i = 1,..., n$, изображенную на координатной плоскости $X0Y$.

Точка с координатами $(\bar{x}, \bar{y})$ называется **центром рассеяния**.

По виду корреляционного поля можно предположить, является ли зависимость между X и Y линейной или нелинейной (форма связи), прямой или обратной (направление связи) и сильной или слабой (теснота связи).

При этом, значения среднеквадратического отклонения σ (большие или

малые) еще не дают характеристику того, есть ли связь между X и Y.

На рис. 2.2, 2.3, 2.4 показаны ситуации, когда σ_x, σ_y малы, но в случае рис. 2.2 зависимости $Y = f(X)$ нет, в случае рис. 2.3 зависимость есть и она линейная, в случае рис. 2.4 есть явно нелинейная зависимость.

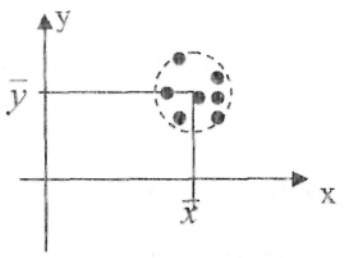


Рис. 2.2.

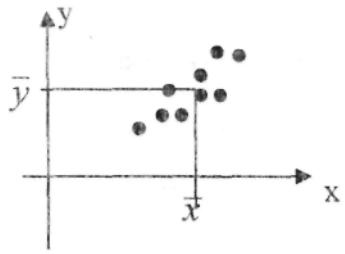


Рис. 2.3.

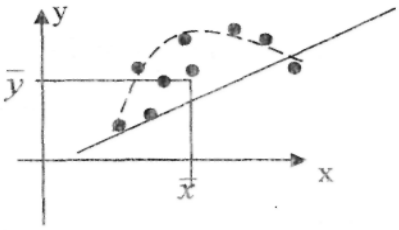


Рис. 2.4.

Для получения ответа на вопрос, есть ли линейная зависимость между переменными X и Y и если да, то насколько значительным является влияние X на Y, вводится еще одна статистика – **коэффициент корреляции,** который дает количественную оценку связи между двумя показателями.

Вначале рассчитывается **ковариация x, y** – $\mathrm{cov}(x, y)$ или σ_{xy} (аналог вариации – совместная вариация):

$$\mathrm{cov}(x, y) = \sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \qquad (2.2)$$

Ковариация обладает тем свойством, что для случаев рис. 2.2 и рис. 2.4 равна 0, а для случая рис. 2.3 ≠ 0, и тем больше по модулю, чем ближе корреляционное поле к прямой.

Если корреляционное поле начинает размываться (рис.2.5), ковариация уменьшается.

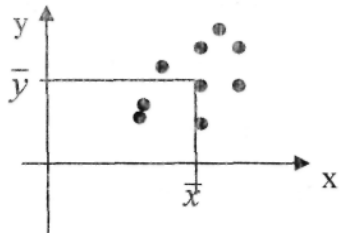


Рис. 2.5.

Для удобства работы ковариацию делят на произведение $\sigma_x \cdot \sigma_y$ и называют **<u>коэффициентом корреляции</u>** (обозначают r_{xy}).

Коэффициент корреляции между переменными X и Y вычисляется по формуле:

$$r_{xy} = \frac{\operatorname{cov}(x, y)}{\sqrt{\operatorname{var}(x)\operatorname{var}(y)}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}. \tag{2.3}$$

Расширенная запись имеет вид:

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \cdot \sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} \text{ или} \tag{2.4}$$

$$r_{xy} = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\sqrt{\overline{X^2} - \left(\overline{X}\right)^2} \cdot \sqrt{\overline{Y^2} - \left(\overline{Y}\right)^2}}, \text{ где } \overline{XY} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i ;\ \overline{Y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2 ;\ \overline{X^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 .$$

(2.5)

Коэффициент корреляции является *показателем плотности* (тесноты) <u>линейной</u> взаимосвязи.

<u>Свойства коэффициента корреляции</u>:

- значения коэффициента всегда находятся в пределах $-1 \le r_{xy} \le 1$;
- если $r_{xy} > 0$, то зависимость между фактором X и Y прямая, т.е. с ростом X показатель Y также возрастает;
- если $r_{xy} < 0$, то зависимость между фактором X и Y обратная;
- если $\left|r_{xy}\right| \to 1$, то плотность связи между X и Y велика – связь почти линейная (рис.

2.3);

- если $|r_{xy}| \to 0$, либо связи нет (рис. 2.2), либо связь резко нелинейная (рис. 2.4).

Плотность линейной взаимосвязи оценивают по следующей таблице:

Значение r_{xy}	Плотность линейной связи
0,9-1,0	тесная
0,6-0,9	достаточная
0,3-0,6	слабая
< 0,3	нет связи

Обычно строят корреляционную таблицу (корреляционную матрицу) связи между переменными *X* и *Y*.

	X	*Y*
X	1	r_{xy}
Y	r_{xy}	1

2.2 Метод наименьших квадратов

Парная (однофакторная) линейная регрессия – линейная зависимость $y = b_0 + b_1 x$ между зависимым показателем *Y* и независимым фактором *X*.

Можно попытаются описать связь между *X* и *Y* зависимостью (2.1):

$$\hat{y} = b_0 + b_1 x. \qquad (2.6)$$

Чтобы получить явный вид зависимости (например, $\hat{y} = 20 - 5x$), необходимо найти (оценить) неизвестные параметры b_0 и b_1 этой модели.

В силу случайных влияний показатель y_i является случайным и может быть записан:

$$y_i = b_0 + b_1 x_i + \varepsilon_i \ \ (i = 1,..., n), \qquad (2.7)$$

где ε_i – случайное отклонение.

Отклонение (ошибка) исходных данных y_i (эмпирических) от рассчитанных по модели $\hat{y}_i = y(x_i)$ (теоретических) вычисляется по формуле:

$$\varepsilon_i = \hat{y}_i - y_i = (b_0 + b_1 x_i) - y_i. \qquad (2.8)$$

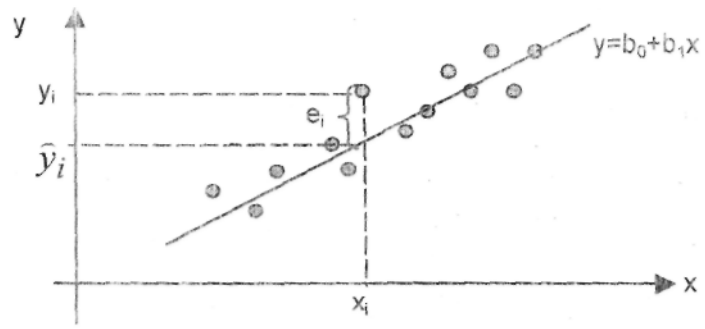


Рис. 2.6.

Естественным требованием при аппроксимации (выравнивании) опытных данных путем построения гипотетической прямой, является сведение к минимуму ошибок при спецификации (определении аналитического вида) формы связи между переменными.

Суть метода наименьших квадратов (МНК) состоит в том, чтобы минимизировать отклонения ε_i в совокупности путем правильного подбора коэффициентов b_0 и b_1. Т.к. отклонение может иметь случайный знак («+» или «-»), то рассматривают квадраты отклонений.

Другими словами, согласно МНК неизвестные параметры b_0 и b_1 подбираются таким образом, чтобы сумма квадратов отклонений эмпирических значений y_i от теоретических значений $\hat{y}_i$, найденных по уравнению регрессии, была минимальной:

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} (b_0 + b_1 x_i - y_i)^2 \to \min . \qquad (2.9)$$

Сумма S является функцией двух неизвестных параметров $S = S(b_0, b_1) = f(b_0, b_1)$. Из математического анализа известно, что необходимое условие минимума функции S от двух переменных – это равенство нулю первых частных производных по каждому из параметров b_0 и b_1:

$$\begin{cases} \dfrac{\partial S}{\partial b_0} = 2\sum\limits_{i=1}^{n} (b_0 + b_1 x_i - y_i) = 0 \\ \dfrac{\partial S}{\partial b_1} = 2\sum\limits_{i=1}^{n} (b_0 + b_1 x_i - y_i) x_i = 0 \end{cases} . \qquad (2.10)$$

После преобразований получаем:

$$\begin{cases} nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases} \tag{2.11}$$

Разделив оба уравнения на n:

$$\begin{cases} b_0 + b_1 \dfrac{\sum_{i=1}^{n} x_i}{n} = \dfrac{\sum_{i=1}^{n} y_i}{n} \\ b_0 \dfrac{\sum_{i=1}^{n} x_i}{n} + b_1 \dfrac{\sum_{i=1}^{n} x_i^2}{n} = \dfrac{\sum_{i=1}^{n} x_i y_i}{n} \end{cases} \Rightarrow \begin{cases} b_0 + b_1 \bar{x} = \bar{y} \\ b_0 \bar{x} + b_1 \bar{x}^2 = \overline{xy} \end{cases}. \tag{2.12}$$

Получилась систему двух линейных уравнений от двух неизвестных. Такая система имеет единственное решение.

Выразив коэффициенты b_0 и b_1, сделав арифметические преобразования, получим выражения для определения этих коэффициентов:

$$b_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = r_{xy} \frac{\sigma_y}{\sigma_x} = \left[\frac{\sigma_{xy}}{\sigma_x \sigma_y}\right] \cdot \left[\frac{\sigma_y}{\sigma_x}\right] = \frac{\sigma_{xy}}{\sigma_x^2}, \tag{2.13}$$

(отношение ковариации к дисперсии)

$$b_0 = \bar{y} - b_1 \bar{x} = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \sum x_i \sum x_i}. \tag{2.14}$$

Затем коэффициенты подставляются в исходное уравнение регрессии.

Коэффициент b_0 – точка пересечения прямой регрессии с осью ординат – постоянная регрессии.

Коэффициент b_1 – выборочный коэффициент регрессии (или просто коэффициент регрессии) Y по X – характеризует наклон прямой регрессии к оси абсцисс: b_1 равняется тангенсу угла этого наклона; это также коэффициент эластичности линейной модели – мера, которая в среднем указывает влияние независимой переменной X на зависимую Y, т.е. показывает на сколько единиц в среднем изменится переменная Y при изменении (увеличении) переменной X на 1 единицу.

2.3 Декомпозиция дисперсий. Понятие о коэффициенте детерминации

Наряду с коэффициентом корреляции используется еще один критерий, при помощи которого также измеряется плотность связи между двумя или более показателями и проверяется адекватность построенной регрессионной модели реальной действительности. Т.е. дается ответ на вопрос, действительно ли изменение значения *Y* линейно зависит именно от изменения значения *X*, а не происходит под влиянием различных случайных факторов. Таким критерием является **коэффициент детерминации.**

Перед рассмотрением сущности данного коэффициента и его связи с коэффициентом корреляции, рассмотрим вопрос о декомпозиции дисперсий, являющийся одним из центральных в статистике.

Рассмотрим рис. 2.8, на котором представлена декомпозиция отклонений фактических значений зависимой переменной *Y* от теоретических – значений, которые находятся на построенной регрессионной прямой:

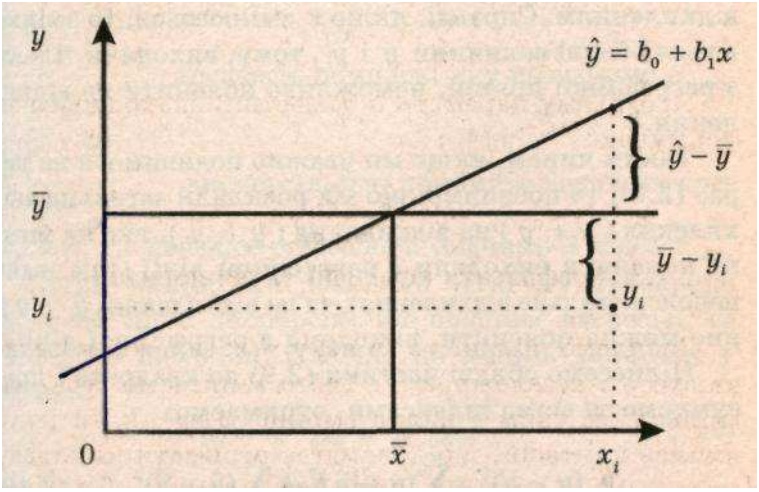


Рис. 2.8.

Как видим, такие отклонения можно записать в виде:

$$(\hat{y}_i - y_i) = (\hat{y}_i - \bar{y}) + (\bar{y} - y_i), \qquad (2.15)$$

умножив обе части равенства на -1: $-(\hat{y}_i - y_i) = -(\hat{y}_i - \bar{y}) - (\bar{y} - y_i)$,

выполнив необходимые преобразования: $(y_i - \hat{y}_i) = (y_i - \bar{y}) - (\hat{y}_i - \bar{y})$,

и переписав следующим образом получим $(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$.

В статистике разницу $(y_i - \bar{y})$ принято называть общим отклонением. Разницу $(\hat{y}_i - \bar{y})$ – отклонением, которое можно объяснить, исходя из регрессионной прямой (действительно, если x_i изменяется, то можно всегда найти значение этого отклонения, имея только регрессионную прямую, т.к. $\bar{y}$ всегда остается неизменной величиной). Разницу $(y_i - \hat{y}_i)$ называют отклонением, которое нельзя объяснить, исходя из регрессионной прямой, или необъясненным отклонением (действительно, если x_i изменяется, то изменяются обе величины y_i и ε_i, поэтому, исходя только из регрессионной прямой, невозможно пояснить это отклонение).

Т.о., если внимательно рассмотреть выражение $(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$, то окажется, что мы разложили общее отклонение $(y_i - \bar{y})$ на отклонение $(y_i - \hat{y}_i)$, которое нельзя пояснить из регрессионной прямой, так называемое, необъясненное отклонение, и на отклонение $(y_i - \hat{y}_i)$, которое можно пояснить, исходя из регрессионной прямой.

Возведем обе части выражения $(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$ в квадрат, просуммируем по всем индексам и, выполнив все преобразования, получим:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 , \qquad (2.16)$$

где $\sum_{i=1}^{n}(y_i - \bar{y})^2$ – общая сумма квадратов;

$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ – сумма квадратов ошибок;

$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ – сумма квадратов, объясняющая регрессию.

Если разделить последнее выражение на n, получим выражение для дисперсий:

$$\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} + \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{n}, \qquad (2.17)$$

где $\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}$ – общая дисперсия, которую обозначим $\sigma^2_{общ}$;

$\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$ – дисперсия ошибок (остаточная дисперсия), которую обозначим

$\sigma^2_{ош}$;

$\frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{n}$ – дисперсия, которую принято называть дисперсией, объясняющей регрессию (факторная дисперсия), обозначим ее через $\sigma^2_{регр}$.

Т.о. мы разложили общую дисперсию на две части: дисперсию, объясняющую регрессию, и дисперсию ошибок:

$$\sigma^2_{общ} = \sigma^2_{регр} + \sigma^2_{ош}. \quad (2.18)$$

Разделив обе части на $\sigma^2_{общ}$, получим:

$$1 = \frac{\sigma^2_{регр}}{\sigma^2_{общ}} + \frac{\sigma^2_{ош}}{\sigma^2_{общ}}. \quad (2.19)$$

Как видим, первая часть $\sigma^2_{регр} \big/ \sigma^2_{общ}$ является пропорцией (т.е. частью) дисперсии, которую можно объяснить, исходя из регрессии в общей дисперсии, а вторая часть – является пропорцией дисперсии ошибок в общей дисперсии $\sigma^2_{ош} \big/ \sigma^2_{общ}$, т.е. представляет собой часть дисперсии, которую нельзя объяснить через регрессионную связь.

Часть дисперсии, объясняющая регрессию, называется **коэффициентом детерминации** и обозначается R^2. Коэффициент детерминации используется как критерий адекватности модели, поскольку является мерой объясняющей силы независимой переменной *X*.

$$R^2 = \frac{\sigma^2_{регр}}{\sigma^2_{общ}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sigma^2_{ош}}{\sigma^2_{общ}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \quad (2.20)$$

Т.о., это отношение указывает, какая часть общего рассеяния значений зависимой переменной *Y* обусловлена изменчивостью независимой переменной *X*.

Другими словами, числовое значение коэффициента детерминации характеризует, в какой степени вариация зависимой переменной *Y* определяется вариацией независимой переменной *X*.

Из определения коэффициента детерминации как относительной части, очевидно, что он всегда заключен в пределах от 0 до 1: $0 \le R^2 \le 1$, при этом чем больше его значение приближается к единице, тем лучше определена функция регрессии.

Связь между коэффициентом корреляции и углом наклона b_1

Коэффициент корреляции и параметр простой линейной регрессии b_1 определяются по известным формулам: $r_{xy} = \dfrac{\operatorname{cov}(x, y)}{\sqrt{\operatorname{var}(x)\operatorname{var}(y)}} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$,

$b_1 = \dfrac{\operatorname{cov}(x, y)}{\operatorname{var}(x)} = \dfrac{\sigma_{xy}}{\sigma_x^2}$,

Умножив числитель и знаменатель формулы расчета коэффициента корреляции на σ_x и выполнив некоторые преобразования, получим:

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \cdot \frac{\sigma_x}{\sigma_x} = \left[\frac{\sigma_{xy}}{\sigma_x^2}\right] \times \left[\frac{\sigma_x}{\sigma_y}\right] \Rightarrow r_{xy} = b_1 \frac{\sigma_x}{\sigma_y}. \qquad (2.21)$$

Из того, что оба значения σ_x и σ_y – положительные, следует, что знак коэффициента корреляции r_{xy} всегда совпадает со знаком параметра b_1.

Связь между коэффициентом корреляции и коэффициентом детерминации

$$R^2 = \frac{\sigma_{регр}^2}{\sigma_{общ}^2} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

Учитывая, что $\hat{y} = b_0 + b_1 x$, выражение факторной дисперсии можно переписать следующим образом:

$$\sigma_{регр}^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} \left[b_0 + b_1 x_i - (b_0 + b_1 \bar{x})\right]^2 = b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 .$$

Отсюда имеем: $R^2 = \dfrac{\sigma_{регр}^2}{\sigma_{общ}^2} = \dfrac{b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \left(\frac{1}{n}\right)}{\sum_{i=1}^{n} (y_i - \bar{y})^2 \cdot \left(\frac{1}{n}\right)} = \dfrac{b_1^2 \sigma_x^2}{\sigma_y^2}$.

Ранее было показано, что $r_{xy} = b_1 \frac{\sigma_x}{\sigma_y}$.

Т.о. $R^2 = r_{xy}^2$ коэффициент детерминации равен квадрату коэффициента корреляции.

Подставив вместо параметра b_1 его выражение, полученное ранее, и учитывая определение дисперсий σ_x^2 и σ_y^2, а также средних $\bar{x}$ и $\bar{y}$, получим:

$$R^2 = \frac{\left[n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \right]^2}{\left[n\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i \right] \cdot \left[n\sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} y_i \right]} = \frac{\left[n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \right]^2}{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right] \cdot \left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i \right)^2 \right]}$$

или $R^2 = \frac{\left[\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \right]^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}$.