Эконометрика

Понятие эконометрики.
Основные типы моделей и данных

Понятие эконометрики

Эконометрика как наука сформировалась на основе таких математических дисциплин, как теория вероятностей и статистика, и экономической теории.

В рамках эконометрики

- Экономические связи и зависимости приобретают строгую математическую форму.
- Эконометрические модели предназначены для качественного анализа экономических ситуаций, выявления силы влияния отдельных факторов модели на результирующую характеристику.
- При помощи эконометрических методов можно выявлять новые, ранее не известные связи, уточнять или отвергать гипотезы о существовании определенных связей между экономическими показателями, предполагаемыми экономической теорией.

Эконометрика

– это наука, которая формулирует экономические модели,

основываясь на экономической теории и экспериментальных данных,

оценивает параметры этих моделей,

делает прогнозы с некоторой степенью точности,

которую также можно оценить в рамках данной науки,

и дает рекомендации по экономической политике.

Эконометрика как наука связанна с эмпирическим выводом экономических законов.

Эконометрика

основывается на базовых статистических разработках и методиках.

Следовательно, математические предпосылки эконометрических моделей те же, что и в статистике.

Для обеспечения наивысшей точности эконометрических прогнозов, эти модели требуют большого объема статистической информации.

Кроме того, качество этих моделей зависит от правильного подхода к формированию модели в неразрывной связи с ее экономической интерпретацией.

При формировании эконометрических моделей

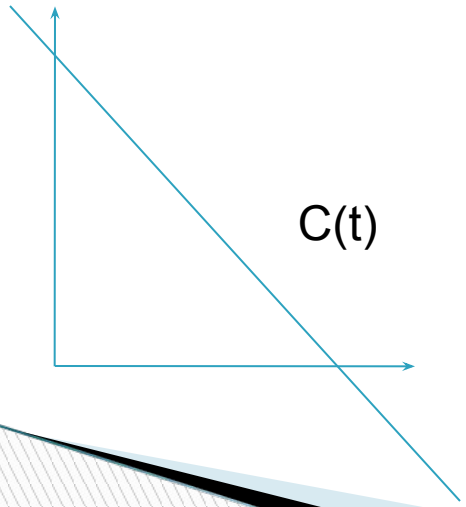
- Требуется четко выделить те факторы, которые будут в нее включены;
- Следует избегать включения в модель несущественных для данной эконометрической зависимости факторов, но и опасаться недооценки влияния некоторых частных показателей;
- Выбор вида связи также оказывает очень существенное воздействие на эконометрическую модель, неправильная интерпретация связи может привести к существенным ошибкам при прогнозировании.

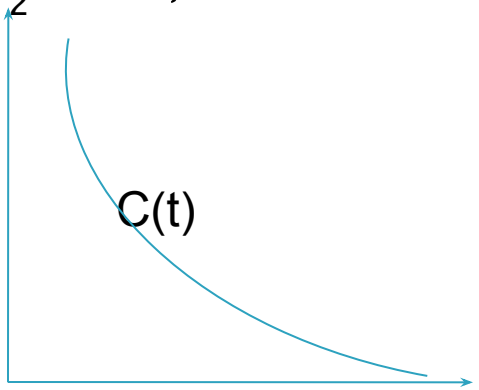
Рассмотрим процесс потребления

- Пусть C – потребление некоторого пищевого продукта на душу населения в некотором году,
- Y – реальный доход на душу населения в этом году, а P – индекс цен на этот продукт, скорректированный (дефлированный) на общий индекс стоимости жизни.
- β_0, β_1, β_2 – константы.
- Поведение потребителя по отношению к покупке данного пищевого продукта можно продемонстрировать при помощи следующей функции:

$$C=\beta_0+\beta_1 Y+\beta_2 P+\varepsilon$$

или $$lnC=\beta_0+\beta_1 lnY+\beta_2 lnP+\varepsilon,$$





Классы эконометрических моделей

- **Модели временных рядов**, которые, в свою очередь бывают:
  - **Моделями тренда** $y(t)=T(t)+\varepsilon_t$, где $T(t)$ - временной ряд, ε_t – случайная составляющая.
  - **Моделями сезонности** $y(t)=S(t)+\varepsilon_t$, где $S(t)$ – сезонная (периодическая) компонента, ε_t – случайная составляющая.
  - **Моделями тренда и сезонности** $y(t)=T(t)+S(t)+\varepsilon_t$ (аддитивная) или $y(t)=T(t)\cdot S(t)+\varepsilon_t$ (мультипликативная) форма модели.
  - **Модели адаптивного прогноза, авторегрессии и скользящего среднего**, общей чертой которых является то, что они объясняют поведение временного ряда, исходя из его предыдущих значений.

Классы эконометрических моделей

- **Модели регрессии** предполагают задание набора факторов модели, оказывающих влияние на результат. Поиск неизвестных параметров этих моделей осуществляется на базе аппарата регрессионного анализа математической статистики.
- В общем виде, регрессионная модель может быть задана в следующей форме:

$$f(x,\beta)=f(x_1,\ldots,x_k,\beta_1,\ldots,\beta_p)+\varepsilon,$$

где $x_1,\ldots,x_k$ – факторы модели, ,$\beta_1,\ldots,\beta_p$ – неизвестные параметры, ε – случайная составляющая.

Классы эконометрических моделей

- **Системы одновременных уравнений** заданы системой регрессионных уравнений в едином временном интервале и тождеств.

Примером может служить модель спроса и предложения на товар широкого потребления:

$Q_{1t}^{D}=\beta_1+\beta_2 P_t+\beta_3 Y_t+u_t$ (спрос),

$Q_{2t}^{S}=\alpha_1+\alpha_2 P_t+\alpha_3 P_{t-1}+\varepsilon_t$ (предложение),

$Q_{1t}^{D}=Q_{2t}^{S}$ (тождество, характеризующее равновесие между спросом и предложением),

где P_t – цена товара в момент времени t, P_{t-1} –цена в предшествующий момент времени, Y_t – совокупный доход населения, u_t и ε_t – случайные составляющие функции предложения и функции спроса.

Типы данных в эконометрике

- Пространственные данные (объемы производства, количество работников, доход в регионе и т. п.);
- Временные данные (спрос, инвестиции и т. п. в привязке ко времени).

Обозначения в эконометрике

- **Модель связи** выражается непосредственно функцией, зависящей от заданных факторов модели и неизвестных параметров регрессии без учета случайной погрешности. В случае парной линейной регрессии модель связи можно записать в следующей форме:

$$y=\alpha+\beta x.$$

- **Модель экспериментальных данных** или **модель наблюдений** позволяет найти значение результирующей переменной в заданной точке. В случае парной линейной регрессии модель данных можно записать в следующей форме

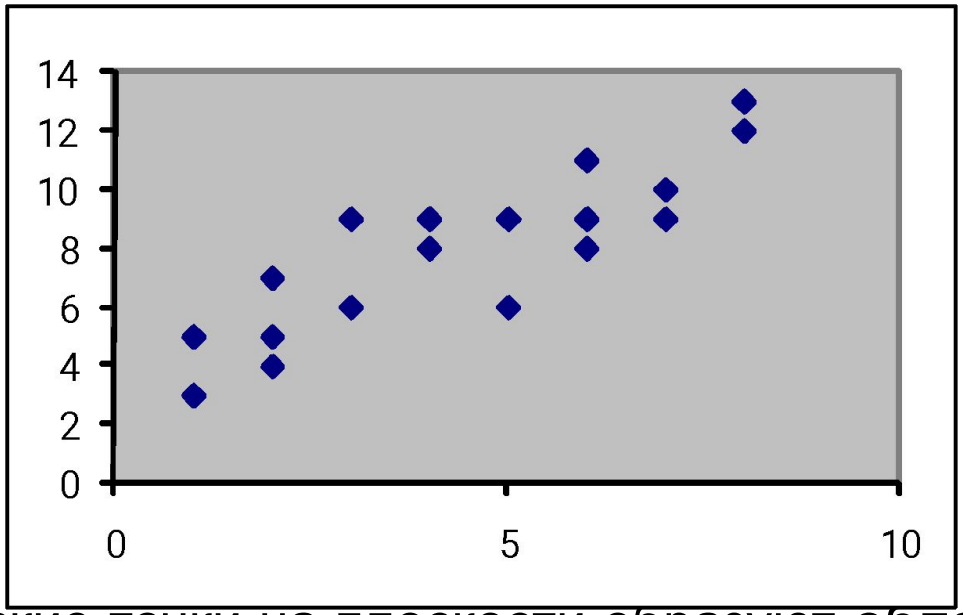
$$y_i=\alpha+\beta x_i+\varepsilon_i.$$

Эконометрика

Парная линейная функция регрессии.

Парная линейная регрессия

- Наличие линейной связи между данными можно подозревать по форме поточечной диаграммы, построенной по экспериментальным данным.
- Такая поточечная диаграмма носит название **диаграмма рассеяния :**



- Эмпирические точки на плоскости образуют **облако рассеяния** произвольной формы.

Вид парной линейной зависимости

- Если облако рассеяния имеет вытянутую в некотором направлении форму, то это позволяет предполагать наличие линейной зависимости между фактор-признаком и результатом.
- В случае парной зависимости вытянутая форма облака рассеяния позволяет предполагать наличие линейной связи $y=\alpha+\beta x$.
- Соотношение между результирующим фактором и фактор-признаком приобретает следующий вид $y_i=\alpha+\beta x_i+\varepsilon_i$.
- При этом ε_i представляет собой отклонение реально наблюдаемых результатов от значений результирующего признака, предсказываемого гипотетической линейной моделью связи $y=\alpha+\beta x$. При этом отклонения $\varepsilon_i=y_i-(\alpha+\beta x_i)$.

Экономический смысл парной линейной модели

На основе данных о размерах располагаемого дохода x_i и расходов на личное потребление C_i для семейных хозяйств можно построить парную линейную модель связи между располагаемым доходом и расходами на личное потребление:

$$C=\alpha+\beta x,$$

где – некоторая постоянная величина ($0<\beta<1$), характеризующая в данном круге семейных хозяйств их склонность к потреблению, связанную с традициями и привычками,

а α – коэффициент автономного потребления.

Степень выраженности линейной связи показывает

Выборочный коэффициент корреляции:

$$r_{xy} = \frac{Cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Величина стоящая в числителе, определяется соотношением:

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

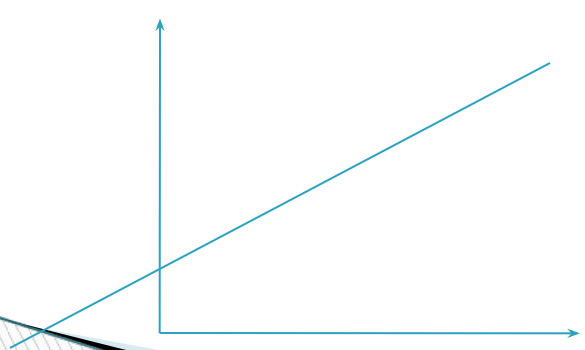
называется **выборочной ковариацией** переменных и , при этом,

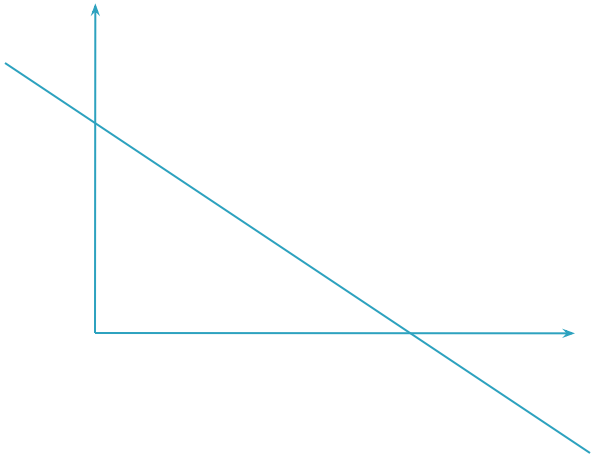
$$Cov(x,x) = Var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad Cov(y,y) = Var(y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

$$Var(x)\cdot\frac{n-1}{n} = \sigma^2{}_x = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \qquad Var(y)\cdot\frac{n-1}{n} = \sigma^2{}_y$$

Выраженность линейной связи

- Если тенденция линейной связи выражена на диаграмме рассеяния довольно ясно, то значени r_{xy} по абсолютной величине близки к единице (т. е. значения близки к +1 или к –1).
- Если же наличие линейной тенденции связи обнаруживается на диаграмме рассеяния с трудом, то тогда значения близки к нулю. r_{xy}
- Значения r_{xy} не зависят от выбора шкал измерения переменных *x* и *y* (если, конечно, эти шкалы линейны).
- Близость коэффициента корреляции к +1 означает наличие прямой линейной связи, а к –1 – обратной.





Метод наименьших квадратов

- Определение неизвестных параметров функции регрессии в эконометрике осуществляется на основе стандартного **метода наименьших квадратов** (МНК), метода, основанного на принципе наименьших квадратов.
- **Принцип наименьших квадратов** утверждает, что выбор параметров функции регрессии является оптимальным в случае, когда сумма квадратов отклонений эмпирических значений результирующей переменной от теоретических значений этой переменной, рассчитанной по функции регрессии, является минимальной. Математически принцип наименьших квадратов можно записать следующим образом:

$$Q(\alpha,\beta)=\sum_{i=1}^{n}\left(y_i-\hat{y}_i(\alpha,\beta)\right)^2\rightarrow \min$$

где: $\hat{y}_i$ –расчетное значение тренда, y_i – фактическое значение тренда из ретроспективного ряда, n –число наблюдений.

МНК применяется в следующих предпосылках

1. Случайные ошибки имеют нулевую среднюю (отсутствуют систематические ошибки), конечные дисперсию и ковариацию;
2. Каждое измерение случайной погрешности характеризуется нулевым средним, не зависящим от значений наблюдаемых переменных;
3. Дисперсии каждой случайной ошибки одинаковы, а их величины независимы от значений наблюдаемых переменных (**гомоскедастичность**);
4. Отсутствует **автокорреляция ошибок**, то есть значения ошибок различных наблюдений независимы друг от друга;
5. Случайные погрешности имеют нормальное распределение;
6. Значение тренда (внутренней переменной) свободны от ошибок измерения и имеют конечные средние значения и дисперсии.

Невыполнение этих предпосылок

Может сделать применение метода некорректным или привести к чрезмерным ошибкам прогноза!

В соответствии с принципом Лагранжа

$$Q(\alpha,\beta)=\sum_{i=1}^{n}\left(y_i-(\alpha+\beta x)\right)^2\to\min$$

$$\begin{cases}\dfrac{\partial Q}{\partial \alpha}=0\\ \dfrac{\partial Q}{\partial \beta}=0\end{cases}$$

Эту систему линейных уравнений можно решить относительно неизвестных α и β.

Возьмем производные

$$\begin{cases} \dfrac{\partial Q}{\partial \alpha} = \dfrac{\partial}{\partial \alpha}\sum_{i=1}^{n}(y_i-(\alpha+\beta x_i))^2 = \sum_{i=1}^{n}\dfrac{\partial}{\partial \alpha}(y_i-(\alpha+\beta x_i))^2 = \\ \dfrac{\partial Q}{\partial \beta} = \dfrac{\partial}{\partial \beta}\sum_{i=1}^{n}(y_i-(\alpha+\beta x_i))^2 = \end{cases}$$

$$= 2\sum_{i=1}^{n}\left(y_i-(\alpha+\beta x_i)\right)\cdot(-1) = 0$$

$$= \sum_{i=1}^{n}\frac{\partial}{\partial \beta}(y_i-(\alpha+\beta x_i))^2 = \quad 2\sum_{i=1}^{n}\left(y_i-(\alpha+\beta x_i)\right)\cdot(-x_i) = 0$$

Умножим каждое уравнение на ½.

Раскроем суммы

$$\begin{cases} -\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \alpha + \sum_{i=1}^{n} \beta x_i = 0 \\ -\sum_{i=1}^{n} y_i x_i + \sum_{i=1}^{n} \alpha x_i + \sum_{i=1}^{n} \beta x_i^{\,2} = 0 \end{cases}$$

$$\sum_{i=1}^{n} \alpha = \alpha \sum_{i=1}^{n} 1 = \alpha(1 + \ldots + 1) = n\alpha$$

$$\begin{cases} n\alpha + \beta \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^{\,2} = \sum_{i=1}^{n} y_i x_i \end{cases}$$

$$\begin{cases} \alpha = \frac{1}{n} \sum_{i=1}^{n} y_i - \frac{\beta}{n} \sum_{i=1}^{n} x_i \\ \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^{\,2} = \sum_{i=1}^{n} y_i x_i \end{cases}$$

$$\begin{cases} \alpha = \frac{1}{n} \sum_{i=1}^{n} y_i - \frac{\beta}{n} \sum_{i=1}^{n} x_i \\ \left(\frac{1}{n} \sum_{i=1}^{n} y_i - \frac{\beta}{n} \sum_{i=1}^{n} x_i \right) \cdot \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^{\,2} = \sum_{i=1}^{n} y_i x_i \end{cases}$$

Получим β

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

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

Решение:

Решение этой системы уравнений дает оценки параметров линейной функции регрессии по МНК (крышечка означает МНК-оценку). Оценка α равна:

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\hat{\beta}\ \sum_{i=1}^{n} x_i = \bar{y} - \hat{\beta}\ \bar{x}\,.$$

При этом точка $(\bar{x}, \bar{y})$ лежит на прямой $y = \hat{\alpha} + \hat{\beta}\, x$

Подстановка выражения для во второе уравнение системы дает:

$$\frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \hat{\beta} + \left(\sum_{i=1}^{n} x_i^2\right)\hat{\beta} = \sum_{i=1}^{n} y_i x_i\,,$$

откуда:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} y_i x_i - n\overline{yx}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Для доказательства проведем преобразование:

Рассмотрим:

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-2\bar{x}\cdot\sum_{i=1}^{n}x_i+n\bar{x}^2=\sum_{i=1}^{n}x_i^2-2\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\cdot\frac{n}{n}+n\bar{x}^2=$$

$$=\sum_{i=1}^{n}x_i^2-2n\bar{x}^2+n\bar{x}^2=\sum_{i=1}^{n}x_i^2-n\bar{x}^2,$$

$$\sum_{i=1}^{n}\bar{x}^2=\bar{x}^2\sum_{i=1}^{n}1=\bar{x}^2(1+\ldots+1)=n\bar{x}^2$$

$$\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})=\sum_{i=1}^{n}y_ix_i-\bar{y}\left(\sum_{i=1}^{n}x_i\right)\cdot\frac{n}{n}-\bar{x}\left(\sum_{i=1}^{n}y_i\right)\cdot\frac{n}{n}+n\overline{yx}=$$

$$=\sum_{i=1}^{n}y_ix_i-n\overline{yx}-n\overline{yx}+n\overline{yx}=\sum_{i=1}^{n}y_ix_i-n\overline{yx}.$$

Последние соотношения позволяют получить более употребительную форму записи выражения для $\hat{\beta}$ (в отклонениях от средних значений):

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

Решение системы уравнений парной линейной регрессии будет:

$$\begin{cases} \hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x} \\ \hat{\beta} = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \end{cases}$$

Такое решение может существовать только, если:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \neq 0\,,$$

что равносильно отличию от нуля определителя системы.

Действительно этот определитель равен:

$$n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2 = n\left[\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right] = n\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 .$$

Последнее условие называется **условием идентифицируемости** модели наблюдений $y_i=\left(\alpha+\beta\cdot x_i\right)+\varepsilon_i,\ i=1,\ldots,n$ и означает попросту, что не все значения $x_1,\ldots,x_n$ совпадают между собой. При нарушении этого условия все точки $(x_i,y_i),\ i=1,\ldots,n$ лежат на одной вертикальной прямой $x=\bar{x}$.

Вспоминая определения ковариации и вариации, видим, что:

$$\hat{\beta}=\frac{Cov(x,y)}{Var(x)}.$$

Отсюда видно, что значения $\hat{\beta}$ близки к нулю, если ковариация между наблюдаемыми значениями переменных *x* и *y* близка к нулю, кроме того, знак $\hat{\beta}$ совпадает со знаком ковариации $Cov(x,y)$, поскольку

$$Var(x)>0 .$$

Эконометрика

Коэффициент детерминации.

Чтобы ввести коэффициент детерминации

Введем для любой точки (x_i, y_i) на диаграмме рассеяния разложение:

$$y_i - \bar{y} = \left(y_i - \hat{y}_i \right) + \left(\hat{y}_i - \bar{y} \right),$$

где $\hat{y}_i = \hat{\alpha} + \hat{\beta}\, x_i$. Возведя обе части последнего представления в квадрат и просуммировав левые и правые части полученных для каждого i равенств, получаем:

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

Входящая в правую часть сумма называется **остаточной суммой квадратов:**

$$\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 = \sum_{i=1}^{n} e_i^2$$

Докажем, что третья сумма равно нулю

Раскроем третье слагаемое:

$$\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)\left(\hat{y}-\bar{y}\right)=\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)\left(\hat{\alpha}+\hat{\beta}\,x_i\right)-\bar{y}\sum_{i=1}^{n}e_i=\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)\hat{y}_i-\bar{y}\sum_{i=1}^{n}e_i=$$

$$=\hat{\alpha}\sum_{i=1}^{n}e_i+\hat{\beta}\sum_{i=1}^{n}(y_i-\hat{y}_i)x_i-\bar{y}\sum_{i=1}^{n}e_i=0 \qquad y_i-\hat{y}_i=e_i, \qquad \sum_{i=1}^{n}e_i=0$$

Тем самым, останется:

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

Слева – полная сумма квадратов, справа – сумма квадратов, объясненная моделью регрессии и остаточная сумма квадратов.

Исходя из теории

Тенденция линейной связи между x и y выражена в максимальной степени, если $\sum_{i=1}^{n} e_i^2 = 0$.

Тенденция линейной связи между переменными x и y не обнаруживается вовсе, если $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2$.

Таким образом, в качестве «меры выраженности» линейной связи между переменными можно использовать величину:

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

называемую **коэффициентом детерминации**. Этот коэффициент изменяется в пределах

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

Исходя из разложения полной суммы квадратов,

Коэффициент детерминации можно представить в форме:

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

поэтому вторая форма коэффициента детерминации будет:

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

то есть отношение суммы квадратов, объясненной моделью регрессии к полной сумме квадратов.

Таким образом, значение R^2 тем выше, чем больше доля объясненной моделью суммы квадратов по отношению к полной сумме квадратов. Вследствие этого, привлечение информации о значениях переменной не дает ничего нового для объяснения изменений значений y от наблюдения к наблюдению.

Коэффициент детерминации

Оценивается следующими способами:

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

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

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

Степень изменчивости результирующей переменной

характеризуется значением выборочной дисперсии:

$$Var(y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$$

При этом вариацию можно разложить на объясненную регрессией и остаточную вариацию:

$$\frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{n-1} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{\hat{y}}\right)^2}{n-1} = Var(\hat{y}),$$

где $\hat{y}$ – переменная, принимающая в i-м наблюдении значение $\hat{y}_i$. (Здесь мы использовали тот факт, что

$$\sum_{i=1}^{n} e_i = 0, \text{ так что } \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right) = 0, \quad \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$

и $\bar{y} = \bar{\hat{y}}$

Тем самым,

$$\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{n-1}=\frac{\sum_{i=1}^{n}e_i^2}{n-1}=\frac{\sum_{i=1}^{n}\left(e_i-\bar{e}\right)^2}{n-1}=Var(e),$$

где e – переменная, принимающая в i-м наблюдении значение e_i.

В итоге, мы получаем разложение:

$$Var(y)=Var(\hat{y})+Var(e)$$

показывающее, что изменчивость переменной y (степень которой характеризуется значением $Var(y)$) частично

объясняется изменчивостью переменной $\hat{y}$

(степень которой характеризуется значением $Var(\hat{y})$)

Не объясненная

часть изменчивости переменной y соответствует изменчивости переменной e (степень которой выражается $Var(e)$).

Таким образом, вспомогательная переменная $\hat{y}$ берет на себя объяснение некоторой части изменчивости значений переменной y, и эта объясненная часть будет тем больше, чем выше значение коэффициента детерминации R^2, который мы теперь можем записать также в виде:

$$R^2 = \frac{Var(\hat{y})}{Var(y)} = 1 - \frac{Var(e)}{Var(y)}$$

Поскольку переменная $\hat{y}$ получается линейным преобразованием переменной x, то изменчивость $\hat{y}$ однозначно связана с изменчивостью x, так что, в конечном счете, построенная модель объясняет часть изменчивости переменной y изменчивостью переменной x. Поэтому, y называют **объясняемой** переменной, а x – **объясняющей** переменной.

Эконометрика

СВОЙСТВА ВЫБОРОЧНЫХ ДИСПЕРСИИ, КОВАРИАЦИИ, КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ И ПАРАМЕТРА РЕГРЕССИИ.

Рассмотрим свойства выборочной функции ковариации

Пусть a – некоторая постоянная, а x_i, y_i, z_i переменные, в i–м наблюдении, $i = 1, \ldots, n$ (n количество наблюдений). Тогда a можно рассматривать как переменную, значения которой в i-м наблюдении равно a. Можно выделить ряд свойств ковариации. Итак,

1. $Cov(x,a) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(a_i - \bar{a}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(a - a),$

так что $Cov(x, a) = 0$.

2. $Cov(x, y) = Cov(y, x)$ $\quad Cov(x, x) = Var(x)$.

3. $Cov(ax, y) = \frac{1}{n-1}\sum_{i=1}^{n}(ax_i - a\bar{x})(y_i - \bar{y}) = a\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$

так что $Cov(ax, y) = a\ Cov(x, y)$.

Свойства ковариации

4. $$Cov(x, y+z) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i + z_i - (\bar{y} + \bar{z})) =$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left((y_i - \bar{y}) + (z_i - \bar{z})\right) =$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) + \frac{1}{n\text{-}1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(z_i - \bar{z}\right),$$

поэтому:

$$Cov(x, y+z) = Cov(x, y) + Cov(x, z)\ .$$

Свойства выборочной вариации

1. Постоянная не обладает изменчивостью: $Var(a) = 0$
2. При изменений единицы измерения переменной в a раз, во столько же раз изменяется и величина стандартного отклонения этой переменной:
   $$Var(ax) = a^2 Var(x),$$
3. Сдвиг начала отсчета не влияет на изменчивость переменной: $Var(x + a) = Var(x)$
4. Дисперсия суммы двух переменных отличается от суммы дисперсий этих переменных на величину, равную удвоенному значению ковариации между этими переменными:

$$Var(x + y) = Cov(x + y, x + y) =$$
$$= Cov(x, x) + Cov(x, y) + Cov(y, x) + Cov(y, y),$$

т.е.

$$Var(x + y) = Var(x) + Var(y) + 2Cov(x, y)$$

Свойства выборочного коэффициента корреляции

При изменении начала отсчета и единицы измерения коэффициента корреляции r_{xy} он остается инвариантен относительно изменения системы координат (выбора единиц изменения и начала отсчета переменных x и y).

Если изменяется начало переменной x и вместо переменных $x_1, \ldots, x_n$ будут значения:

$$\tilde{x}_i = a + bx_i, \quad i = 1, \ldots, n, \quad (b > 0), \text{ т.е. } \tilde{x} = a + bx$$

Тогда:

$$r_{\tilde{x}y} = \frac{Cov(\tilde{x}, y)}{\sqrt{Var(\tilde{x})}\sqrt{Var(y)}} = \frac{Cov(a+bx, y)}{\sqrt{Var(a+bx)}\sqrt{Var(y)}} =$$

$$= \frac{Cov(bx, y)}{\sqrt{Var(bx)}\sqrt{Var(y)}} = \frac{bCov(x, y)}{\sqrt{b^2 Var(x)}\sqrt{Var(y)}} = r_{xy}.$$

Свойства параметра регрессии

Оценка $\hat{\beta}_x$ параметра регрессии β в модели наблюдений

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n.$$

инвариантна относительно изменения системы координат не будет .

Если перейти к новой системе координат , так что $\tilde{x} = a + bx$,

то $$\hat{\beta}_{\tilde{x}} = \frac{Cov(\tilde{x}, y)}{Var(\tilde{x})} = \frac{Cov(a + bx, y)}{Var(a + bx)} = \frac{bCov(x, y)}{b^2 Var(x)} = \frac{1}{b}\beta_x .$$

Таким образом, изменяя единицу измерения переменной x (или переменной y), мы можем получать существенно различные значения β от сколь угодно малых, до сколь угодно больших. (Желательно выбирать единицы измерения таким образом, чтобы сравниваемые переменные имели одинаковый порядок.). Близость значений $\hat{\beta}$ к нулю всегда должна интерпретироваться с оглядкой на используемые единицы измерения переменных x и y.

Коэффициент корреляции наблюдаемого значения результата и оценки по МНК-методу

Рассмотрим теперь коэффициент корреляции $r_{y\hat{y}}$ между переменными y и $\hat{y}$, где $\hat{y} = \hat{\alpha} + \hat{\beta} x$, а $\hat{\alpha}$ и $\hat{\beta}$ оценки параметров регрессии α и β.

Замечая, что $y = \hat{y} + e$ (т.к. $e_i = y_i - \hat{y}_i$ по определению), находим:

$$r_{y\hat{y}} = \frac{Cov(y, \hat{y})}{\sqrt{Var(y)}\sqrt{Var(\hat{y})}} = \frac{Cov(\hat{y} + e, \hat{y})}{\sqrt{Var(y)}\sqrt{Var(\hat{y})}}$$

$$= \frac{Cov(\hat{y}, \hat{y}) + Cov(e, \hat{y})}{\sqrt{Var(y)}\sqrt{Var(\hat{y})}}.$$

Но ранее было доказано, что

Соотношение $\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)\left(\hat{y}_i-\bar{y}\right)=0$, а с учетом того, что $\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)=0$

приходим к утверждению: $\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)\hat{y}_i=0\Rightarrow Cov(e,\hat{y})=Cov(y-\hat{y},\hat{y})=0$

Следовательно: $r_{y\hat{y}}=\frac{Var(\hat{y})}{\sqrt{Var(y)}\sqrt{Var(\hat{y})}}=\sqrt{\frac{Var(\hat{y})}{Var(y)}}$,

так что $r^2_{y\hat{y}}=\frac{Var(\hat{y})}{Var(y)}=R^2$.

Следовательно, коэффициент детерминации равен квадрату коэффициента корреляции между переменными y и $\hat{y}$. При достаточно сильно выраженной линейной связи между переменными x и y, что соответствует значению R^2, близкому к 1, тоже стремится к 1. $r_{y\hat{y}}$ называют $r^2_{y\hat{y}}$ **множественным коэффициентом корреляции** и обозначают символом R.

Отметим также, что переменная $\hat{y}$ измеряется в тех же единицах, что и переменная y, и при изменении масштаба измерения переменной y значение $r_{y\hat{y}}$ не изменяется. Отсюда вытекает, что коэффициент детерминации R^2 **инвариантен относительно изменения масштаба и начала отсчета** переменных x и y.

В результате получим

Множественный коэффициент корреляции в виде:

$$r_{y\hat{y}} = \frac{Cov(y,\hat{y})}{\sqrt{Var(y)}\sqrt{Var(\hat{y})}} = \frac{Cov(y,\hat{\alpha}+\hat{\beta}\ x)}{\sqrt{Var(y)}\sqrt{Var(\hat{\alpha}+\hat{\beta}\ x)}}$$

$$= \frac{\hat{\beta}\ Cov(y,x)}{\sqrt{Var(y)}\sqrt{\hat{\beta}^{\,2}Var(x)}} = \frac{\text{sign}(\hat{\beta})\cdot Cov(y,x)}{\sqrt{Var(y)}\sqrt{Var(x)}}.$$

(здесь *sign*(*z*)=–1 для *z*<0, *sign*(*z*)=0 для *z*=0, sign(*z*)=1 для *z*>0)

Поскольку же: $\hat{\beta} = \frac{Cov(x,y)}{Var(x)}$, то $\text{sign}(\hat{\beta}) = \text{sign}(Cov(x,y))$, и

$$r_{y\hat{y}} = \text{sign}(Cov(x,y))\cdot r_{xy}\ ,$$

Т.е. $r^2_{y\hat{y}} = r^2_{xy} = R^2$,

и мы можем установить значение R^2 еще до построения модели линейной связи.

В заключении

Коэффициент детерминации необходимо оценивать тремя различными способами:

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

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

$$r^2_{y\hat{y}} = r^2_{xy} = R^2 .$$