Тема 5. Эконометрическое моделирование

* Возникновение эконометрики как науки
* Определение эконометрики
* Прикладные цели эконометрики
* Этапы эконометрического моделирования

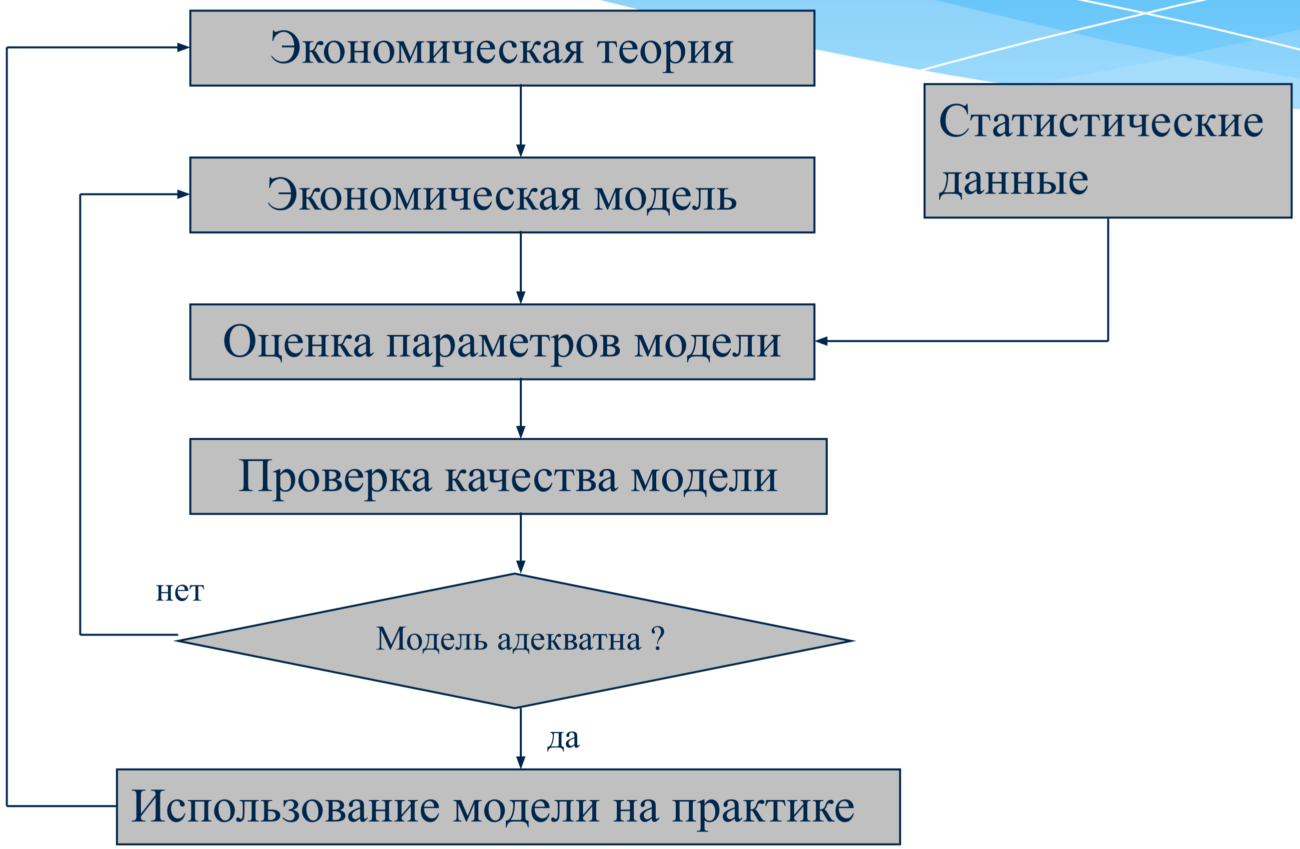
Прикладные цели эконометрики

* вывод экономических законов;
* формулировка экономических моделей, основываясь на экономической теории и эмпирических данных;
* оценка неизвестных величин (параметров) в этих моделях;
* прогнозирование и оценка точности прогноза;
* выработка рекомендаций по экономической политике.

Этапы эконометрического моделирования

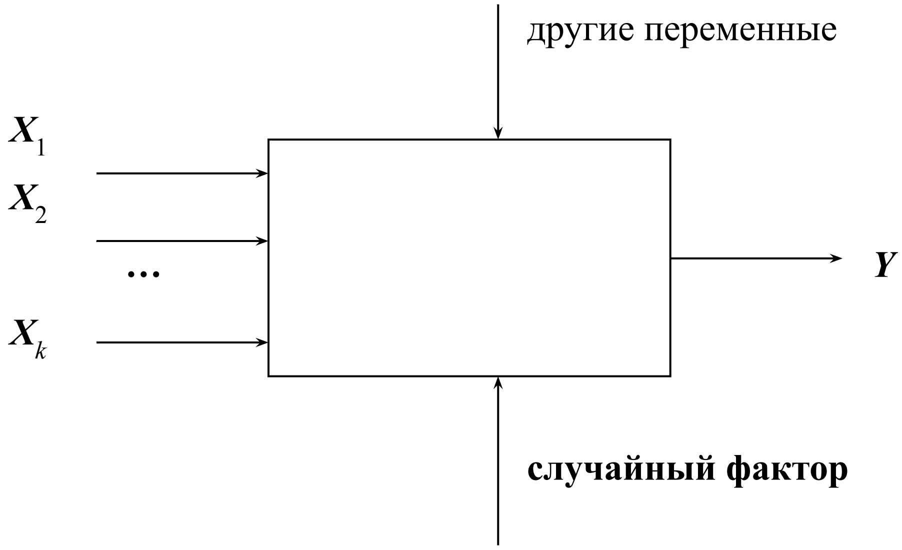
* Осознание того факта, что в экономике многие переменные связаны между собой
* Группировка отдельных соотношений в модель
* Сбор данных
* Идентификация
* Верификация

Этапы эконометрического моделирования



1. Переменные модели

* Переменную, процесс формирования значений которой нас по каким-то причинам интересует, будем обозначать Y и называть зависимой или объясняемой.
* Переменные, которые, как мы предполагаем, оказывают влияние на переменную Y, будем обозначать X_j и называть независимыми или объясняющими.


другие переменные
X_1
X_2
...
X_k
Y
случайный фактор

Другая классификация переменных

* Переменные, значения которых объясняются в рамках нашей модели, называются *эндогенными*.
* Переменные, значения которых нашей моделью не объясняются, являются для нее внешними, ничего о том, как формируются эти значения, мы не знаем, называются *экзогенными*

2. Спецификация модели

* определение цели моделирования;
* определения списка экзогенных и эндогенных переменных;
* определение форм зависимостей между переменными;
* формулировка априорных ограничений на случайную составляющую, что важно для свойств оценок и выбора метода оценивания;
* формулировка априорных ограничений на коэффициенты

Виды эконометрических моделей

* Модели временных рядов.
* Регрессионные модели с одним уравнением.
* Системы одновременных уравнений.

Модели временных рядов.

* Такие модели объясняют поведение переменной, меняющейся с течением времени, исходя только из ее предыдущих значений. К этому классу относятся модели тренда, сезонности, тренда и сезонности (аддитивная и мультипликативная формы) и др.

Регрессионные модели с одним уравнением.

* В таких моделях зависимая (объясняемая) переменная представляется в виде функции от независимых (объясняющих) переменных и параметров. В зависимости от вида функции модели бывают линейными и нелинейными.

Системы одновременных уравнений.

* Ситуация экономическая, поведение экономического объекта описывается системой уравнений. Системы состоят из уравнений и тождеств, которые могут содержать в себе объясняемые переменные из других уравнений (поэтому вводят понятия экзогенных и эндогенных переменных).

3. Сбор данных.

* cross-sectional data – пространственные данные – набор сведений по разным экономическим объектам в один и тот же момент времени;
* time-series data – временные ряды – наблюдение одного экономического параметра в разные периоды или моменты времени. Эти данные естественным образом упорядочены во времени.
* panel data – панельные данные – набор сведений по разным экономическим объектам за несколько периодов времени (данные переписи населения).

4. Идентификация.

- Идентификация модели – статистический анализ модели и, прежде всего – статистическое оценивание параметров. Выбор метода оценивания сюда тоже входит. Зависит от особенностей модели.

5. Верификация.

* Верификация модели – сопоставление реальных и модельных данных, проверка оцененной модели с тем, чтобы прийти к выводу о достаточной реалистичности получаемой с ее помощью картины объекта, либо признать необходимость оценки другой спецификации модели.

Вопросы для самопроверки

* Кто первый ввел в употребление термин «Эконометрика».
* В каком году был основан журнал «Econometrics».
* Каких вы знаете лауреатов нобелевской премии по экономике за достижения в эконометрических методах.
* На каких «трех китах» базируется современная экономическая теория.
* Приведите определение эконометрики, отражающее современный взгляд на эту науку.
* Каковы прикладные цели эконометрики.
* Перечислите основные этапы эконометрического моделирования.
* Что входит в спецификацию модели.
* Что происходит на этапе идентификации модели.
* Какие основные типы экономических данных вы знаете.
* Основные типы эконометрических моделей.
* Как происходит верификация модели

Тема 6. Парная линейная регрессионная модель

ПЛРМ

Две переменные X и Y

могут быть связаны

* функциональной зависимостью (т.е. существует функция f что $Y = f(X)$, значения переменной Y полностью определяются значениями переменной X)
* статистической зависимостью
* независимы.

Статистическая зависимость

* Если при изменении X меняется закон распределения случайной величины Y, то говорят, что величины (X,Y) связаны статистической зависимостью.
* Статистическая зависимость называется корреляционной, если при изменении X меняется математическое ожидание случайной величины Y.

Корреляционная зависимость

Если каждому значению величины *X* соответствует свое значение $M(Y \mid X)$

то говорят, что существует регрессионная функция

$$M(Y \mid X) = f(X)$$

Случайная составляющая

Отклонение переменной Y от математического ожидания для соответствующего значения переменной X называется ошибкой и обозначается ε

$$\varepsilon(X) = Y(X) - f(X)$$

Регрессионное уравнение

Уравнение

$$Y = f(X) + \varepsilon$$

называется уравнением регрессии переменной Y на переменную X

Экономический смысл ε

* невключение объясняющих переменных в уравнение. На самом деле на переменную *Y* влияет не только переменная *X*, но и ряд других переменных, которые не учтены в нашей модели по следующим причинам:
  * мы знаем, что другая переменная влияет, но не модем ее учесть, потому как не знаем, как измерить (психологический фактор, например);
  * существуют факторы, которые мы знаем, как измерить, но влияние их на Y так слабо, что их не стоит учитывать;
  * существенные переменные, но из-за отсутствия опыта или знаний мы их таковыми не считаем.

Экономический смысл ε (продолжение)

* Неправильная функциональная спецификация. Функциональное соотношение между *Y* и *X* может быть определено неправильно. Например, мы предположили линейную зависимость, а она может быть более сложной.
* Ошибки наблюдений (занижение реального уровня доходов). В этом случае наблюдаемые значения не будут соответствовать точному соотношению, и существующее расхождение будет вносить свой вклад в остаточный член.

Способы определения регрессионной функции $f(X)$

* параметрический – предполагаем, что вид регрессионной функции известен, неизвестны параметры функции
* непараметрический – предполагаем, что вид регрессионной функции неизвестен и мы составляем алгоритм расчета значений функции в каждой точке

Выбор вида $f(X)$

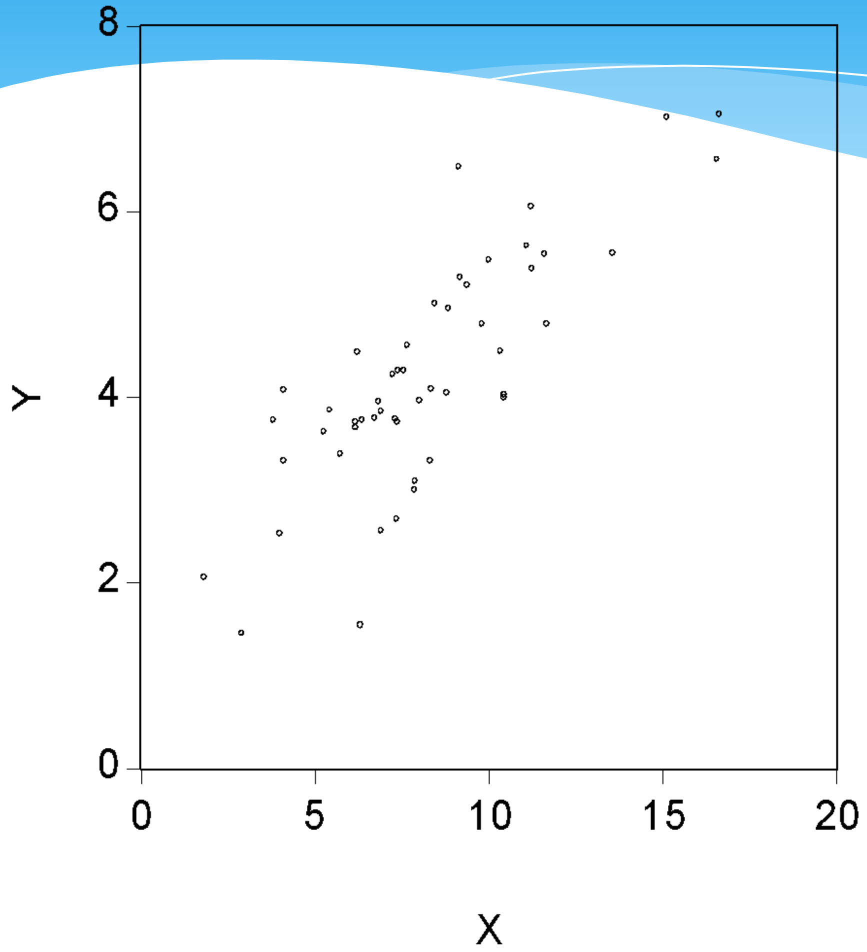
* экономическая теория
* опыт, интуиция исследователя
* эмпирический анализ данных

Эмпирический анализ данных

В парном случае материал наблюдений представляет собой набор пар чисел:

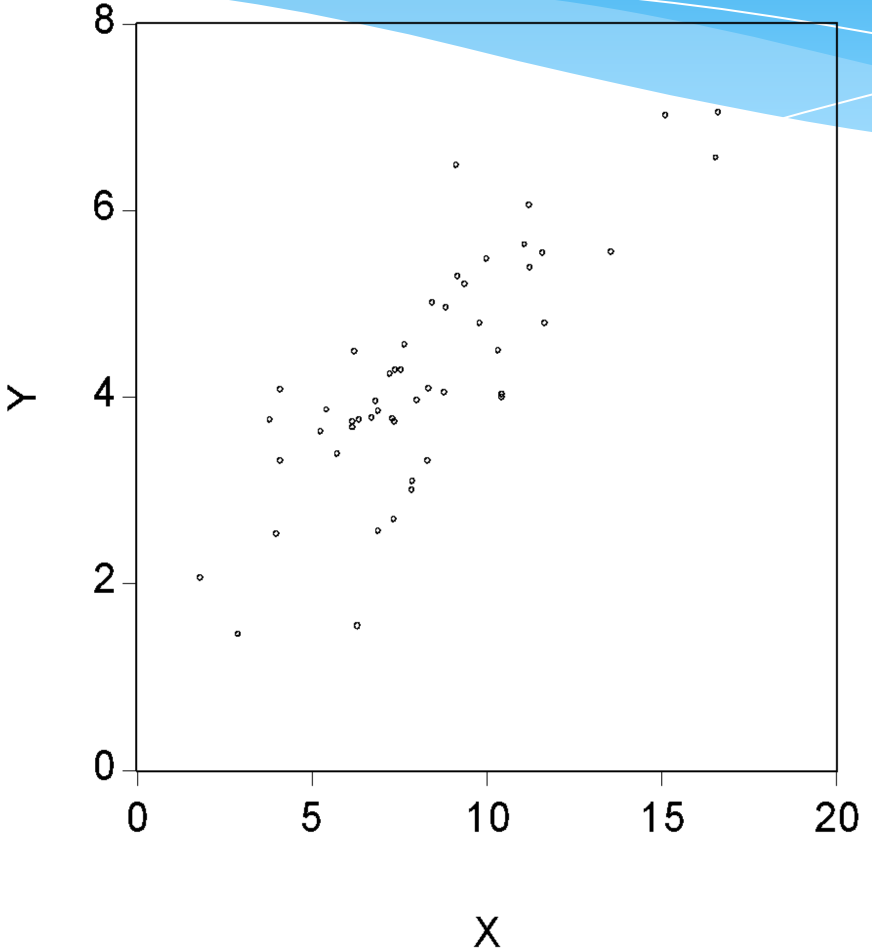
$$(X_i, Y_i) \quad i = 1, \ldots, N$$

На плоскости каждому такому наблюдению соответствует точка:



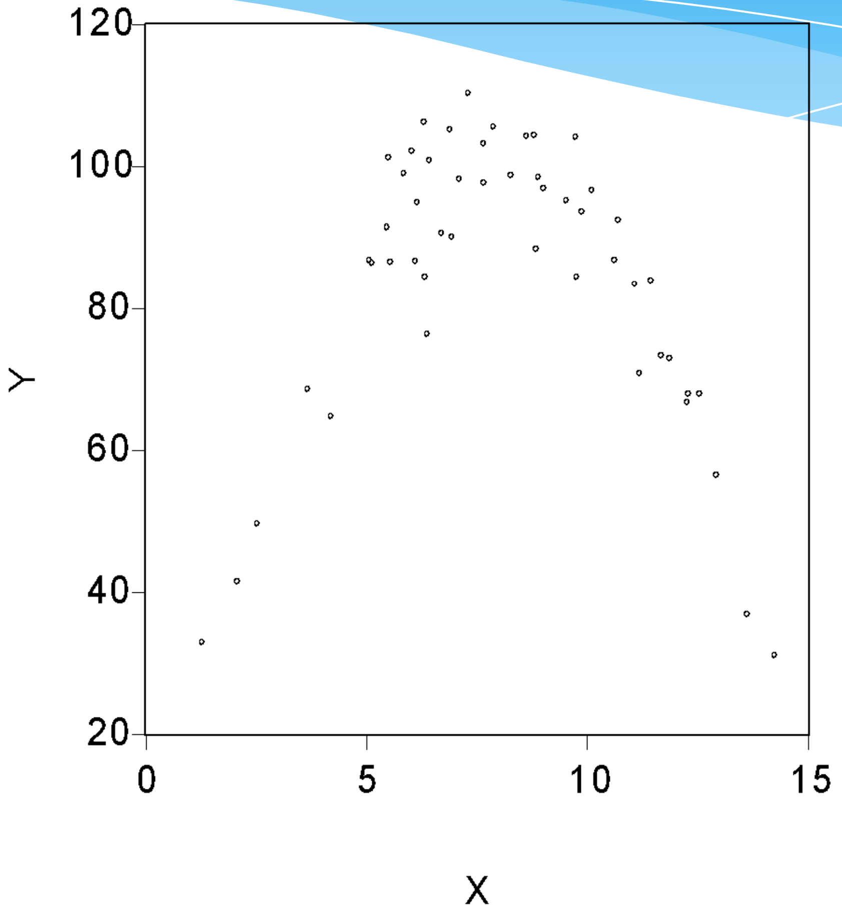
Полученный график *называют облако наблюдений, поле корреляции или диаграмма рассеяния.* По виду облака наблюдений можно определить вид регрессионной функции.

Линейная Y=α+βX+ε.

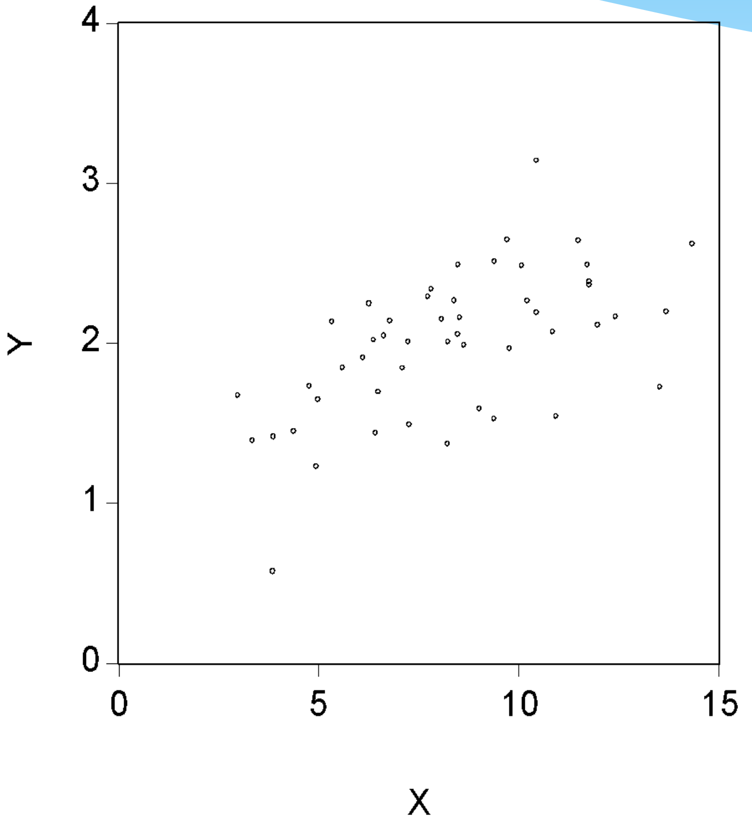


Квадратичная

$$Y = \alpha + \beta X + \gamma X^2 + \varepsilon$$

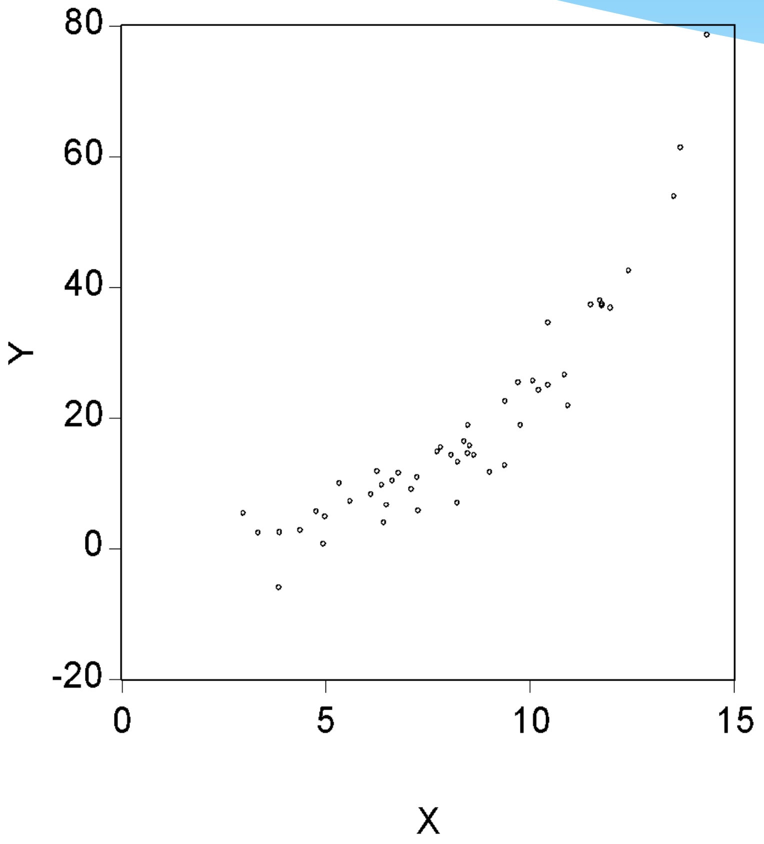


Показательная $Y = \alpha X^{\beta}\varepsilon$



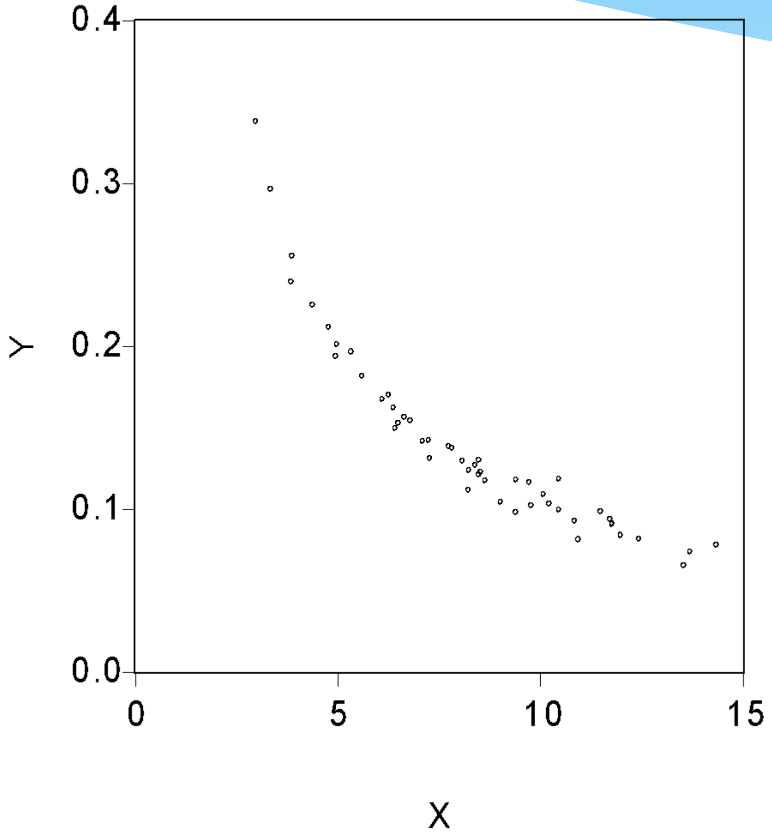
Степенная

$$Y = \alpha e^{\beta X} \varepsilon$$

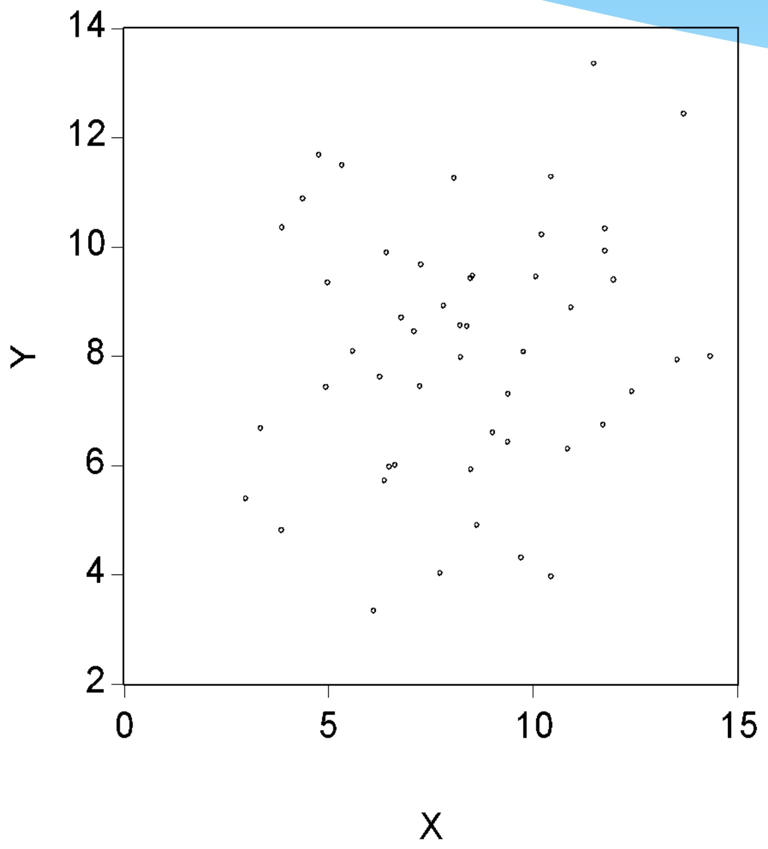


Гиперболическая

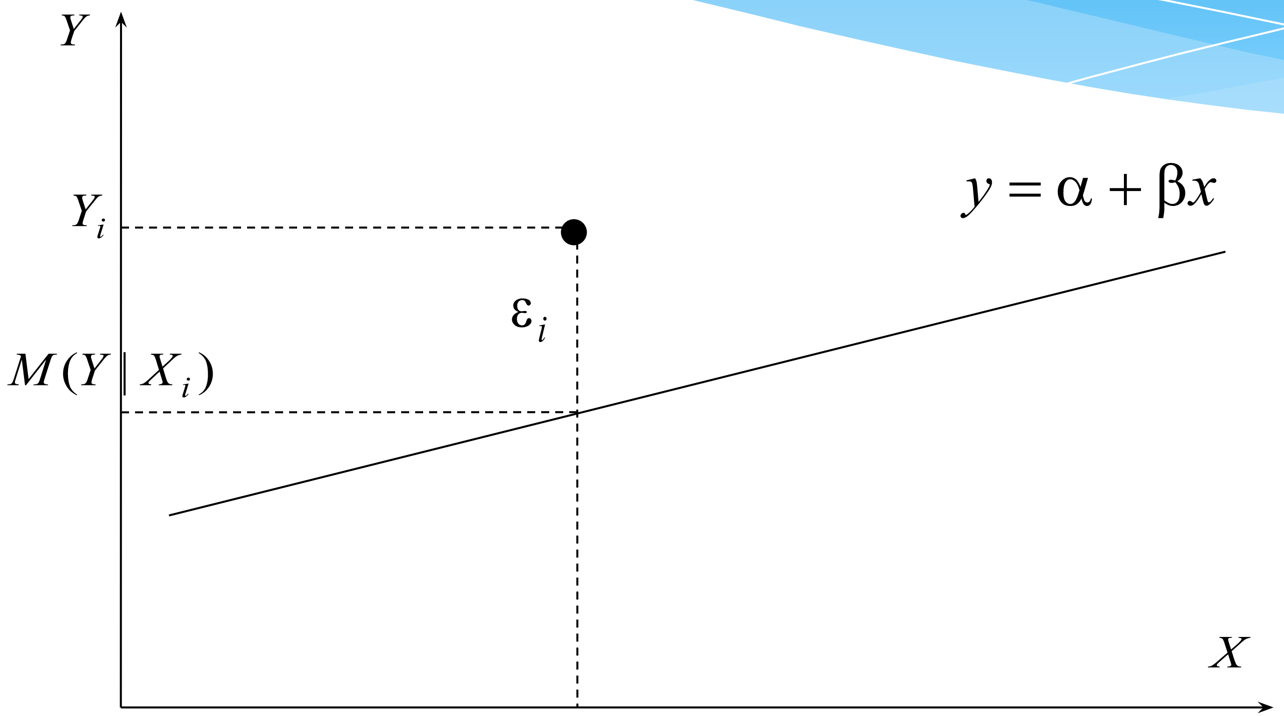
$$Y = \alpha + \frac{\beta}{X} + \varepsilon$$



X и Y независимы

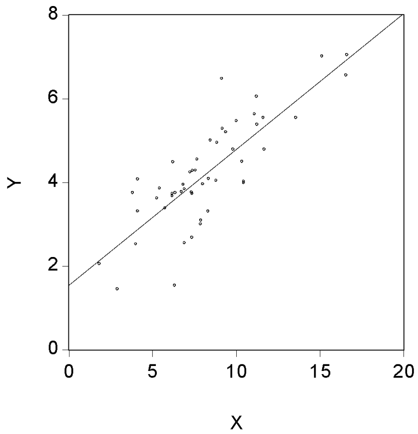


Парная линейная регрессионная модель Y=α+βX+ε.

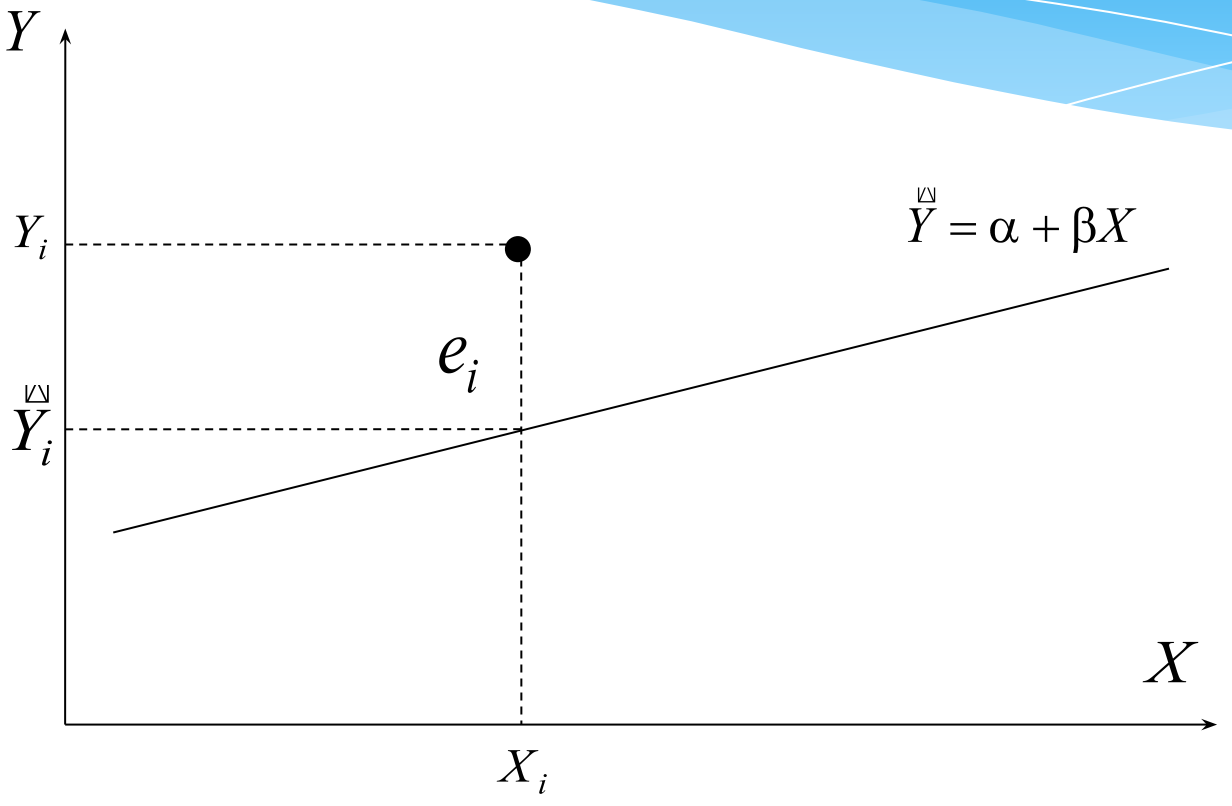


Выбор коэффициентов регрессионной прямой

Из всех возможных прямых мы хотим выбрать ту, чтобы она «наилучшим образом» подходила к нашим данным, т. е. отражала бы линейную зависимость *Y* от *X*. Иными словами, чтобы каждое Y_i лежало бы как можно ближе к прямой. Можно сказать, мы хотим, чтобы желаемая прямая *была бы в центре скопления наших данных.*



Рассмотрение остатков на графике



Интегральная мера близости

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - a - bX_i)^2 = S(a,b) \to \min_{(a,b)}$$

$$\sum_{i=1}^{N} \left| e_i \right| = \sum_{i=1}^{N} \left| Y_i - a - bX_i \right| = S(a,b) \to \min_{(a,b)}$$

Метод наименьших квадратов

$$\sum_{i=1}^{N} e_i^2 \to \min_{(a,b)}$$

Среди всех возможных прямых выбираем ту, для которой сумма квадратов остатков минимальна

Минимизация

$$\frac{\partial S}{\partial a} = -2\sum_{i=1}^{N}(Y_i - a - bX_i) = 0$$

$$\frac{\partial S}{\partial b} = -2\sum_{i=1}^{N}(Y_i - a - bX_i)X_i = 0$$

$$\begin{cases} \sum_{i=1}^{N}(Y_i - a - bX_i) = 0 \\ \sum_{i=1}^{N}(Y_i - a - bX_i)X_i = 0 \end{cases} \quad \text{или} \quad \begin{cases} \sum_{i=1}^{N} e_i = \mathbf{0} \\ \sum_{i=1}^{N} X_i e_i = \mathbf{0} \end{cases}$$

Система нормальных уравнений

$$\begin{cases} Na + b\sum_{i=1}^{N} X_i = \sum_{i=1}^{N} Y_i \\ a\sum_{i=1}^{N} X_i + b\sum_{i=1}^{N} X_i^2 = \sum_{i=1}^{N} X_i Y_i \end{cases}$$

МНК-коэффициенты ПЛРМ

$$\hat{\beta} = \frac{\dfrac{\sum_{i=1}^{N} X_i Y_i}{N} - \dfrac{\sum_{i=1}^{N} X_i}{N}\dfrac{\sum_{i=1}^{N} Y_i}{N}}{\dfrac{\sum_{i=1}^{N} X_i^2}{N} - \left(\dfrac{\sum_{i=1}^{N} X_i}{N}\right)^2}$$

- коэффициент наклона

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$$

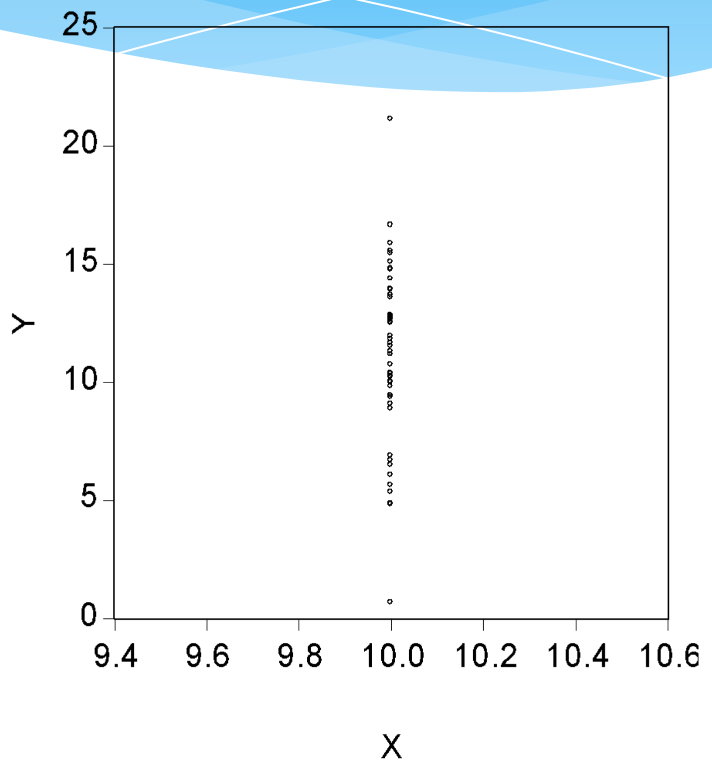
- свободный коэффициент

Другие формы записи коэффициента наклона

$$\hat{\beta} = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N} (X_i - \bar{X})^2} = \frac{Cov(X, Y)}{\sigma_X^2}$$

Замечания

* Линия регрессии проходит через точку $\left(\overline{X}, \overline{Y}\right)$
* Мы предполагаем, что среди X_i есть разные, тогда $\sigma_X \neq 0$. В противном случае, оценок по методу наименьших квадратов не существует.



Теснота линейной корреляционной связи

В качестве меры близости данных наблюдений к линии регрессии служит выборочный коэффициент парной линейной корреляции (парный линейный коэффициент корреляции):

$$r_{xy} = \frac{\dfrac{\sum_{i=1}^{N} X_i Y_i}{N} - \overline{X}\,\overline{Y}}{\sqrt{\dfrac{\sum_{i=1}^{N} X_i^2}{N} - \left(\overline{X}\right)^2}\sqrt{\dfrac{\sum_{i=1}^{N} Y_i^2}{N} - \left(\overline{Y}\right)^2}}$$

Связь между коэффициентом корреляции и коэффициентом наклона

$$\hat{\beta} = r_{XY} \frac{\sigma_Y}{\sigma_X} \qquad r_{XY} = \hat{\beta} \frac{\sigma_X}{\sigma_Y}$$

Знак коэффициента наклона линии регрессии и коэффициента корреляции совпадают

Свойства коэффициента корреляции

$$|r_{xy}| \le 1$$

$r_{xy} = \pm 1$ - необходимое и достаточное условием того, что все наблюдаемые значения (X_i, Y_i) лежат на прямой регрессии

Свойства коэффициента корреляции (продолжение)

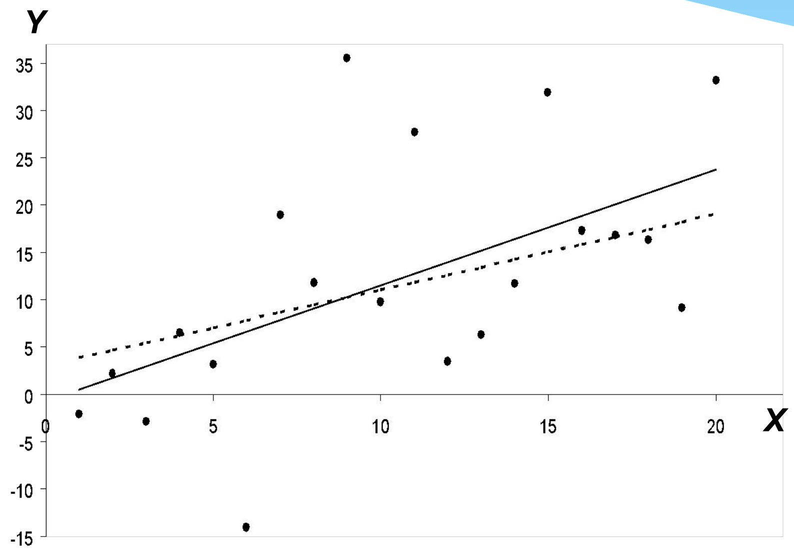
$$r_{XY} = 0$$

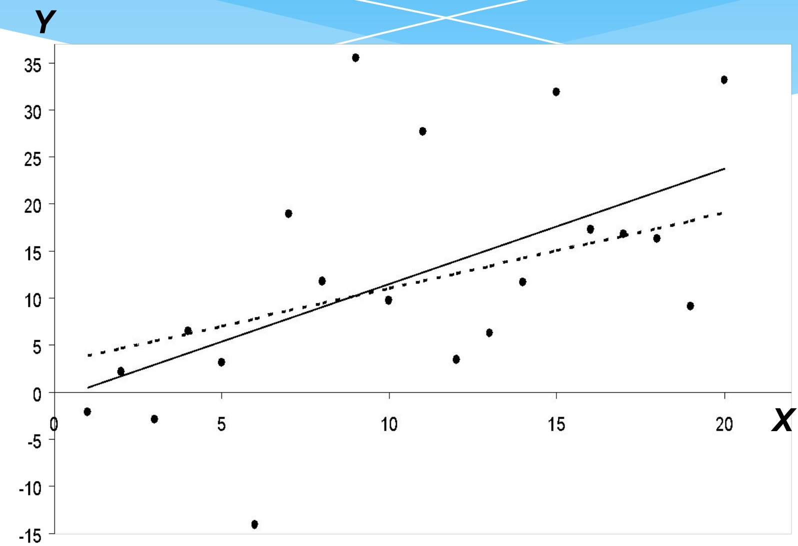
переменные не связаны линейной корреляционной связью. Линия регрессии проходит горизонтально.

$$0 < |r_{xy}| < 1$$

между переменными существует линейная корреляционная связь, которая тем лучше (ближе к линейной функциональной), чем ближе коэффициент корреляции по модулю к 1

Уравнение одно, коэффициенты корреляции разные





$Y = 3.0 + 0.8X$

Вопросы для самопроверки

* Что такое функциональная зависимость между переменными.
* Что такое статистическая зависимость.
* Что такое корреляционная зависимость.
* Дайте определение независимых переменных.
* Что такое линия регрессии.
* Какова основная идея метода наименьших квадратов.
* Какие меры близости точек к линии регрессии вы знаете.
* Почему мы называем расчетные коэффициенты линии регрессии «статистическими оценками».
* Как выбрать функциональную форму линии регрессии.
* Форы записи МНК коэффициента наклона ергрессионной прямой.
* В чем заключается экономический смысл случайной составляющей регрессионного уравнения.
* Для чего нужен коэффициент корреляции.
* Как связан коэффициент корреляции и коэффициент наклона линии регрессии.
* Перечислите свойства коэффициента корреляции.
* В каком случае линии регрессии по методу наименьших квадратов не существует.

Тема_7. Множественная линейная регрессионная модель

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$$

Темы лекции

* Множественная линейная регрессионная модель
* Метод наименьших квадратов оценки коэффициентов МЛРМ.
* Матричное выражение МНК-оценок коэффициентов МЛРМ.

Множественные регрессионные модели

Независимая переменная Y характеризует состояние или поведение экономического объекта. Набор переменных $X_1,\dots,X_k$, характеризуют этот экономический объект качественно или количественно. Предполагаем, что переменные X оказывают влияние на переменную Y, т. е. реализации переменной Y выступают в виде функции, значения которой определяются. правда, с некоторой погрешностью, значениями объясняющих переменных, выступающих в роли аргументов этой функции, т. е.

$$Y = f(X_1,\dots,X_k) + \varepsilon,$$

где ε - случайная компонента

МЛРМ

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$$

Пример

$$Q^D = \beta_0 + \beta_1 P + \beta_2 X + \beta_3 PM + \varepsilon$$

где Q^D – объем спроса на масло,

X – доход,

P – цена на масло,

PM – цена на мягкое масло.

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$$

Здесь нам неизвестны коэффициенты β и параметры распределения ε.
Для их оценки имеется выборка из *N* наблюдений над переменными Y и $X_1,\ldots,X_k$.
Для каждого наблюдения должно выполнятся следующее равенство:

$$Y_i = \beta_{\mathbf{0}} + \beta_{\mathbf{1}} X_{\mathbf{1}i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

Матричная форма записи МЛРМ

$$Y = X\beta + \varepsilon$$

где

$$Y = \begin{bmatrix} Y_1 \\ \dots \\ Y_N \end{bmatrix} \quad X = \begin{bmatrix} 1\, X_{11} & \dots & X_{k1.} \\ . & . & . \\ 1\, X_{N1} & \dots & X_{kN} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \dots \\ \beta_k \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_N \end{bmatrix}$$

Метод наименьших квадратов

$$\sum_{i=1}^{N} e_i^2 \to \min_{(a,b)}$$

Среди всех возможных гиперплоскостей выбираем ту, для которой сумма квадратов остатков минимальна

Что будем минимизировать

$$\hat{Y}_i = b_0 + b_1 X_{1i} + \ldots + b_k X_{ki}$$

$$e_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_1 X_{1i} - \ldots - b_k X_{ki}$$

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_{1i} - \ldots - b_k X_{ki})^2 =$$

$$= S(b_0, b_1, \ldots, b_k) \to \min_{(b_0, b_1, \ldots, b_k)}$$

Минимизация

$$\begin{cases} \dfrac{\partial S}{\partial b_0} = 0 \\ \dfrac{\partial S}{\partial b_1} = 0 \\ \dfrac{\partial S}{\partial b_2} = 0 \\ \ldots \\ \dfrac{\partial S}{\partial b_k} = 0 \end{cases} \quad \text{или} \quad \begin{cases} -2\sum\limits_{i=1}^{N}(Y_i - b_1 X_{1i} - \ldots - b_k X_{ki}) = 0 \\ -2\sum\limits_{i=1}^{N}(Y_i - b_1 X_{1i} - \ldots - b_k X_{ki})X_{1i} = 0 \\ -2\sum\limits_{i=1}^{N}(Y_i - b_1 X_{1i} - \ldots - b_k X_{ki})X_{2i} = 0 \\ \ldots \\ -2\sum\limits_{i=1}^{N}(Y_i - b_1 X_{1i} - \ldots - b_k X_{ki})X_{ki} = 0 \end{cases}$$

Система нормальных уравнений

$$
\begin{cases}
Nb_0 + b_1 \sum_{i=1}^{N} X_{1i} + b_2 \sum_{i=1}^{N} X_{2i} + \ldots + b_k \sum_{i=1}^{N} X_{ki} = \sum_{i=1}^{N} Y_{1i} \\
b_0 \sum_{i=1}^{N} X_{1i} + b_1 \sum_{i=1}^{N} X_{1i}^2 + b_2 \sum_{i=1}^{N} X_{1i} X_{2i} + \ldots + b_k \sum_{i=1}^{N} X_{1i} X_{ki} = \sum_{i=1}^{N} Y_{1i} X_{1i} \\
b_0 \sum_{i=1}^{N} X_{2i} + b_1 \sum_{i=1}^{N} X_{1i} X_{2i} + b_2 \sum_{i=1}^{N} X_{2i}^2 + \ldots + b_k \sum_{i=1}^{N} X_{2i} X_{ki} = \sum_{i=1}^{N} Y_{1i} X_{2i} \\
\ldots \\
b_0 \sum_{i=1}^{N} X_{ki} + b_1 \sum_{i=1}^{N} X_{1i} X_{ki} + b_2 \sum_{i=1}^{N} X_{2i} X_{ki} + \ldots + b_k \sum_{i=1}^{N} X_{ki}^2 = \sum_{i=1}^{N} Y_{1i} X_{ki}
\end{cases}
$$

Вывод формулы для нахождения коэффициентов в матричном виде

$$\sum_{i=1}^{N} e_i^2 = e'e$$

$$e = \begin{bmatrix} e_1 \\ \dots \\ e_N \end{bmatrix}$$

$$e = Y - X\hat{\beta}$$

Вывод формулы для нахождения коэффициентов в матричном виде

$$e'e = (Y - X\hat{\beta})'(Y - X\hat{\beta}) = Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} =$$
$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

$$\frac{\partial(e'e)}{\partial\hat{\beta}} = -2X'Y + 2X'X\hat{\beta}$$

$$2X'Y = 2X'X\hat{\beta}$$

$$X'Y = X'X\hat{\beta}$$

Итог

$$\hat{\beta} = (X'X)^{-1}X'Y$$

МНК оценки коэффициентов МЛРМ

Полная мультиколлинеарность

Коэффициенты по методу наименьших квадратов существуют не всегда, а только в том случае, когда определитель матрицы (X'X) отличен от нуля.

Определитель будет равен нулю в случае, если столбцы матрицы X линейно зависимы. Такое может произойти, если между независимыми переменными существует точное линейное соотношение.

Пример

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 W + \varepsilon$$

где

Y - средняя оценка на экзамене состоящую из трех объясняющих переменных:

I – доход родителей,

D – среднее число часов, затраченных на обучение в день,

W – среднее число часов, затраченных на обучение в неделю.

Очевидно, что W=7D.

Устранение полной мультиколлинеарности

Случай полной мультиколлинеарности отследить легко, поскольку в этом случае невозможно построить оценки по методу наименьших квадратов. Если в модели присутствует полная мультиколлинеарность, следует удалить из регрессионного уравнения одну из переменных, которые входят в линейное соотношение.

Вопросы для самопроверки

* Система нормальных уравнений для нахождения коэффициентов по МНК.
* В каком случае линии регрессии по методу наименьших квадратов не существует
* Приведите примет модели, в которой присутствует полная мультиколлинеарность.
* Укажите размерности матриц, участвующих в формуле МНК-коэффициентов.
* .Как устранить проблему полной мультиколлинеарности.
* Выведите систему нормальных уравнений.
* Выведите матричную формулу МНК коэффициентов.
* Приведите пример ситуации, когда линейной зависимости между объясняющими переменными нет, а коэффииценты МЛРМ не существуют.
* Как влияют выбросы на результаты оценивания.
* Как исследовать устойчивость результатов оценивания.