

Тема 2. Парная линейная регрессионная модель

0011 0010 1010 1101 0001 0100 1011

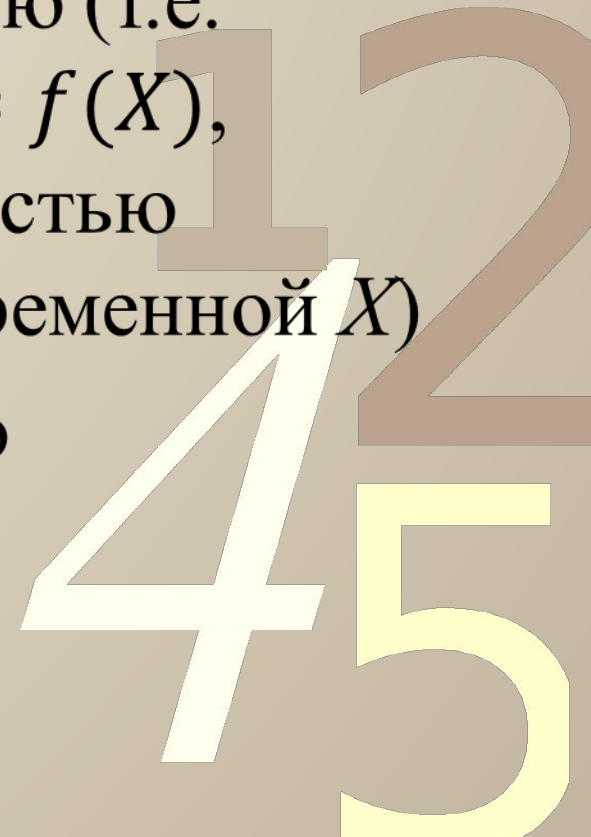
ПЛРМ

1 2
4 5

Две переменные X и Y

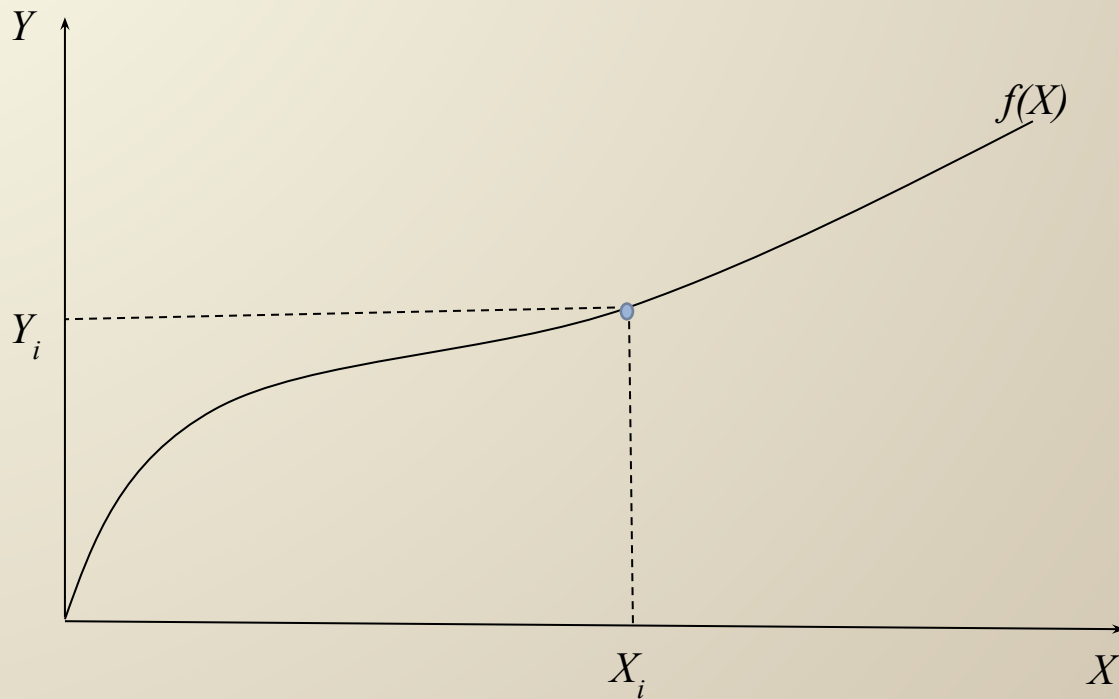
0011 0010 1010 1101 0001 0100 1011

- 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.



Функциональная зависимость

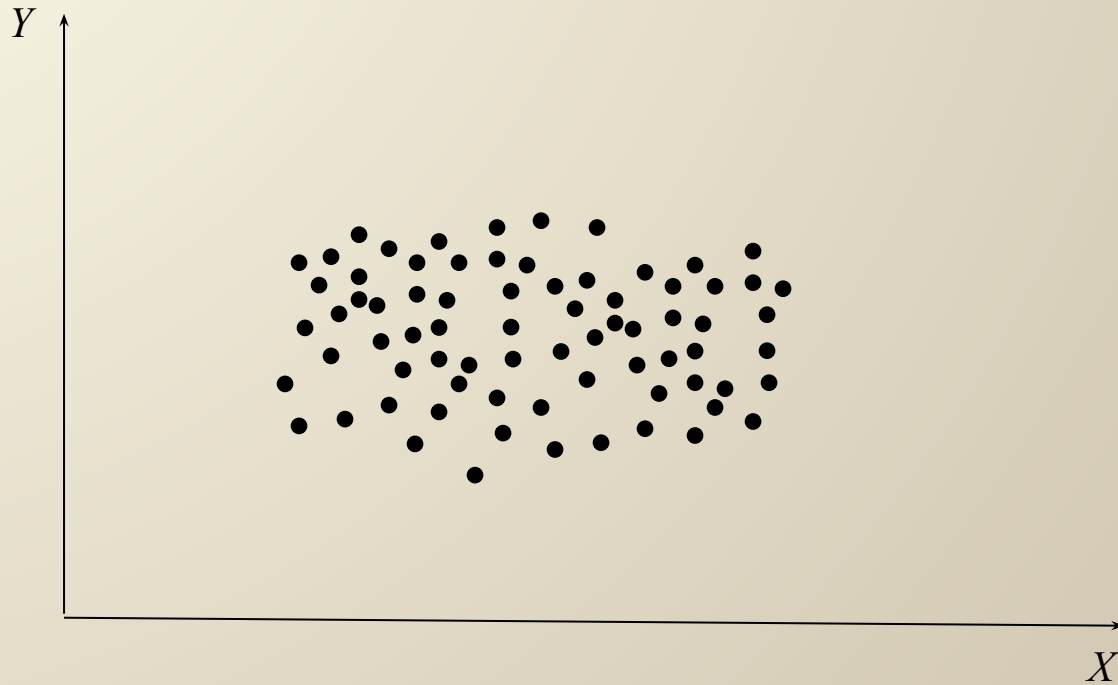
0011 0010 1010 1101 0001 0100 1011



1 2
4 5

Независимость

0011 0010 1010 1101 0001 0100 1011



1 2
4 5



Статистическая зависимость

0011 0010 1010 1101 0001 0100 1011

- Если при изменении X меняется закон распределения случайной величины Y , то говорят, что величины (X, Y) связаны статистической зависимостью.



Статистическая зависимость

0011 0010 1010 1101 0001 0100 1011

Здесь будет красивый рисунок (когда-нибудь)

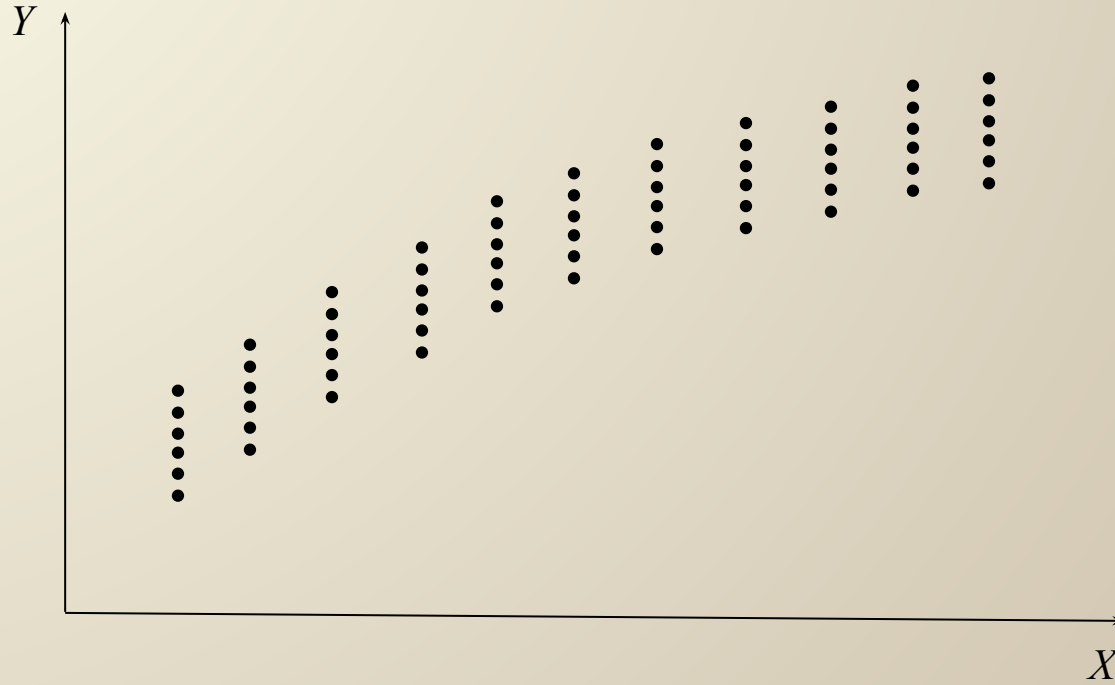


Статистическая зависимость

- Статистическая зависимость называется корреляционной, если при изменении X меняется математическое ожидание случайной величины Y .
- Если при изменении переменной X меняется дисперсия переменной Y , такую зависимость называют гетероскедастичностью.
- Корреляция и гетероскедастичность могут наблюдаться одновременно

Корреляция

0011 0010 1010 1101 0001 0100 1011

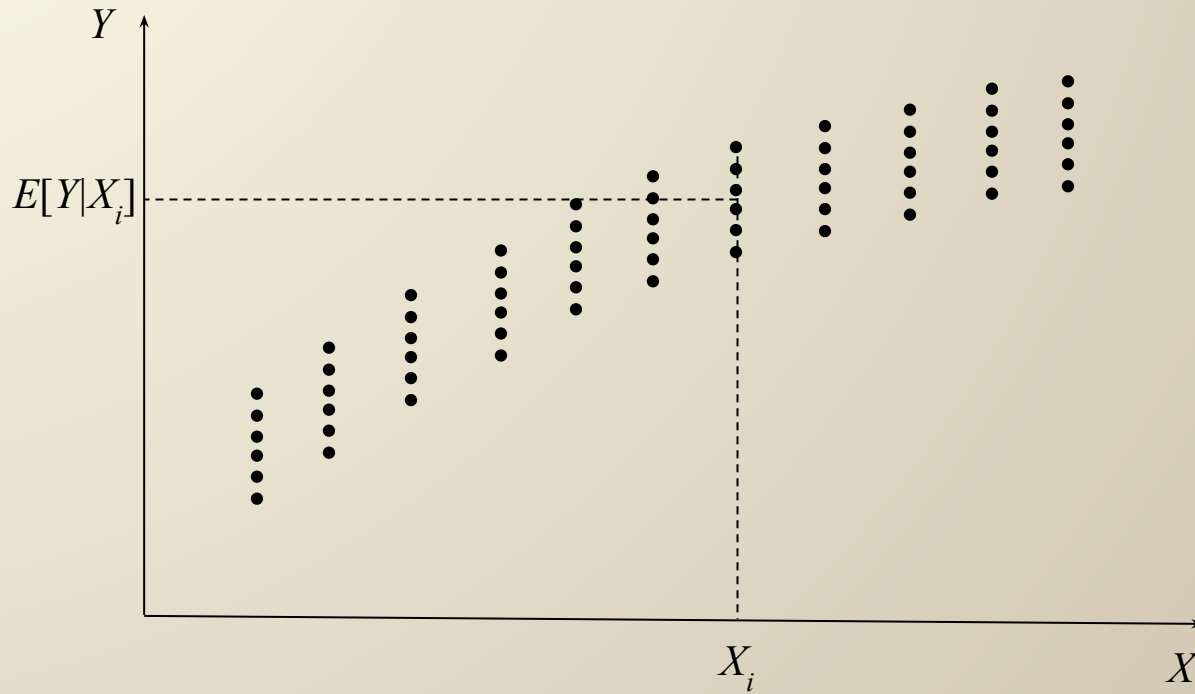


1 2

4 5

Корреляция

0011 0010 1010 1101 0001 0100 1011

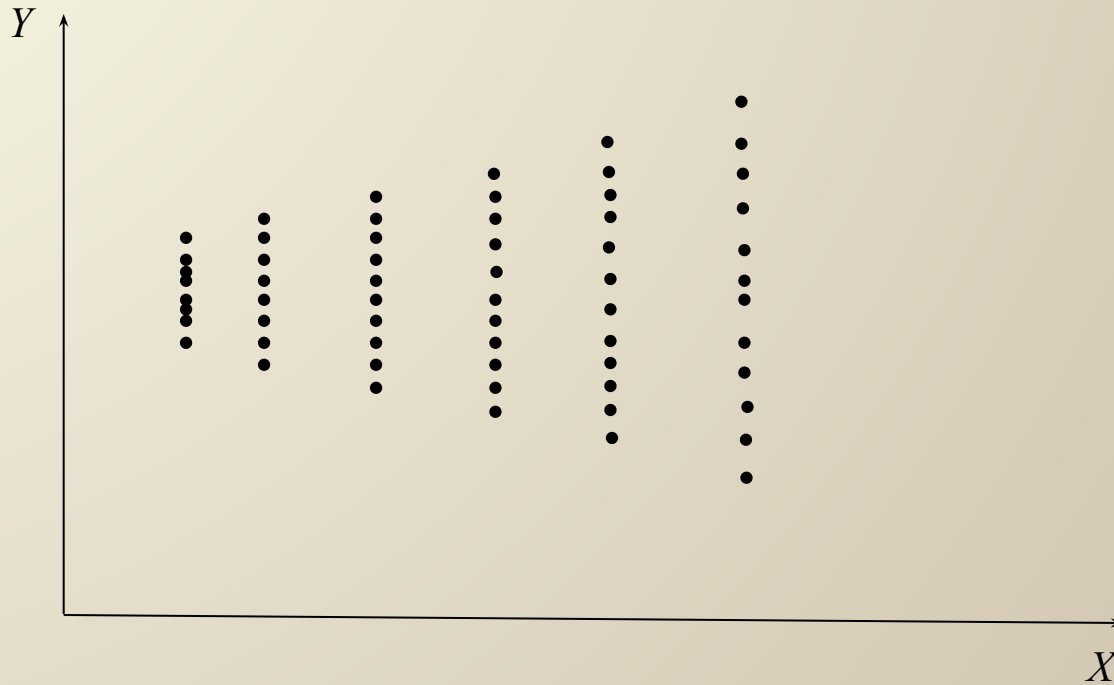


1 2

4 5

Гетероскедастичность

0011 0010 1010 1101 0001 0100 1011



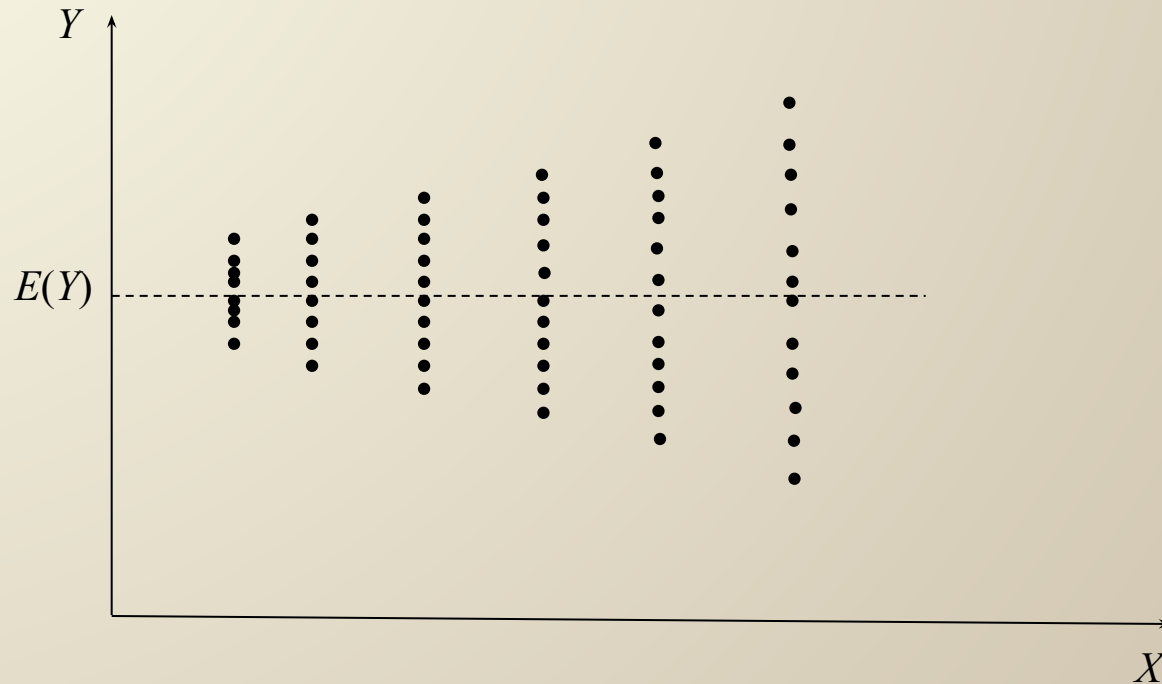
1 2

4 5



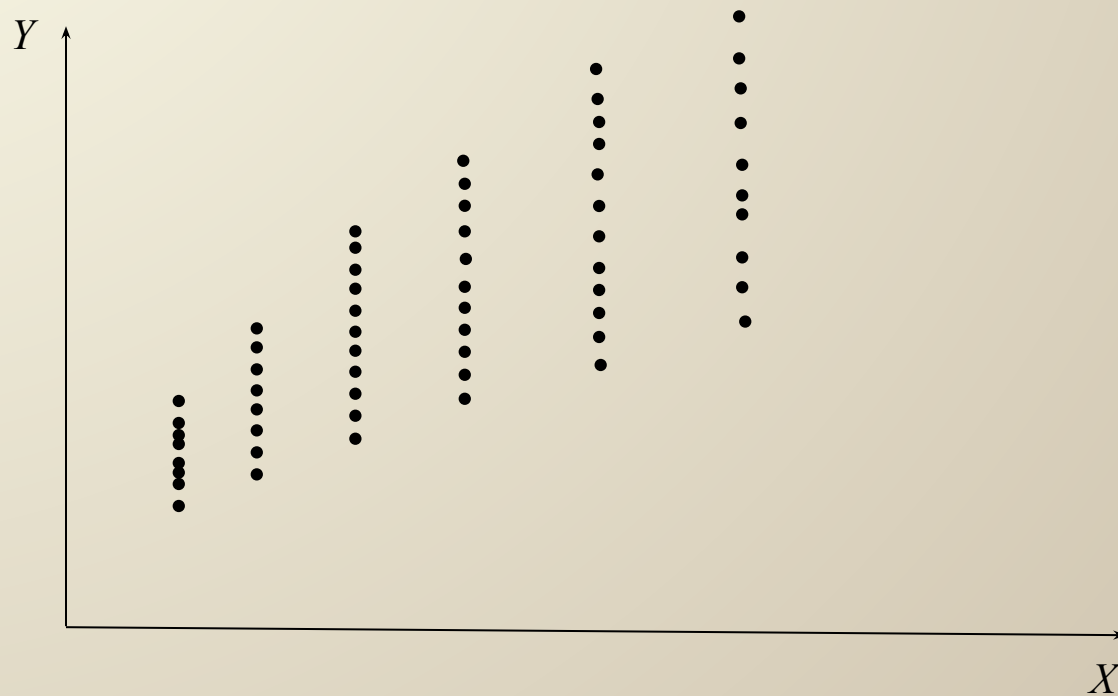
Гетероскедастичность

0011 0010 1010 1101 0001 0100 1011



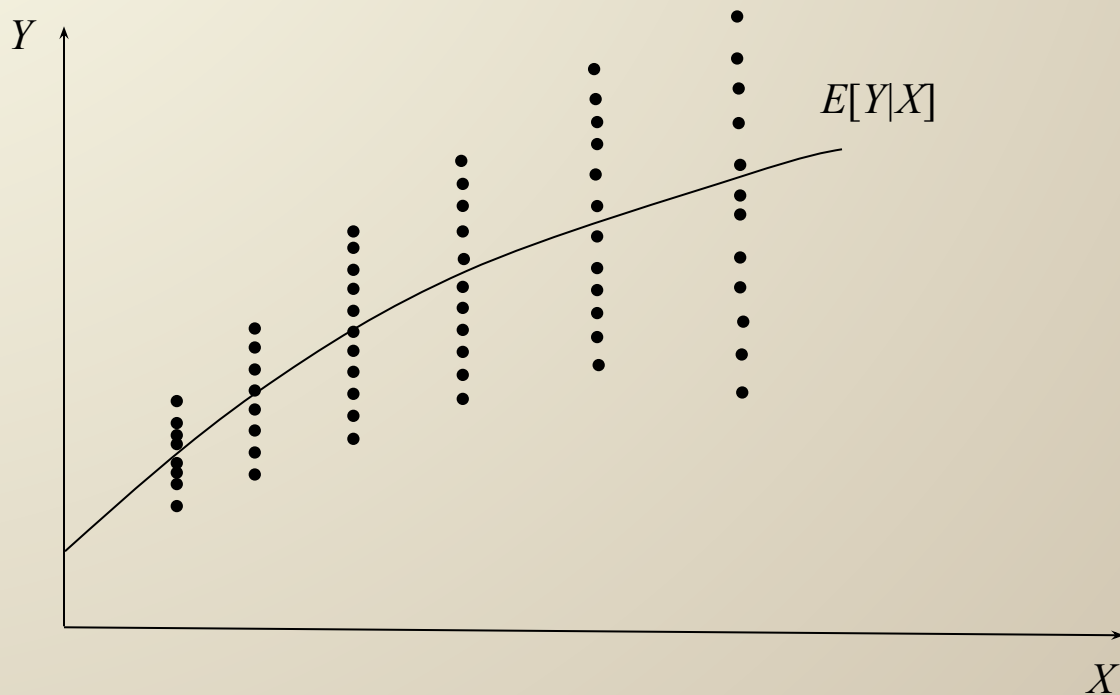
1 2
4 5

Корреляция и гетероскедастичность



1 2
4 5

Корреляция и гетероскедастичность



1 2

4 5

Корреляционная зависимость

Если каждому значению величины X соответствует свое значение $E[Y | X]$ то говорят, что существует регрессионная функция

$$E(Y | X) = f(X)$$

Линию, которую описывает регрессионная функция, называется линия регрессии



Случайная составляющая

0011 0010 1010 1101 0001 0100 1011

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

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



Регрессионное уравнение

0011 0010 1010 1101 0001 0100 1011

Уравнение

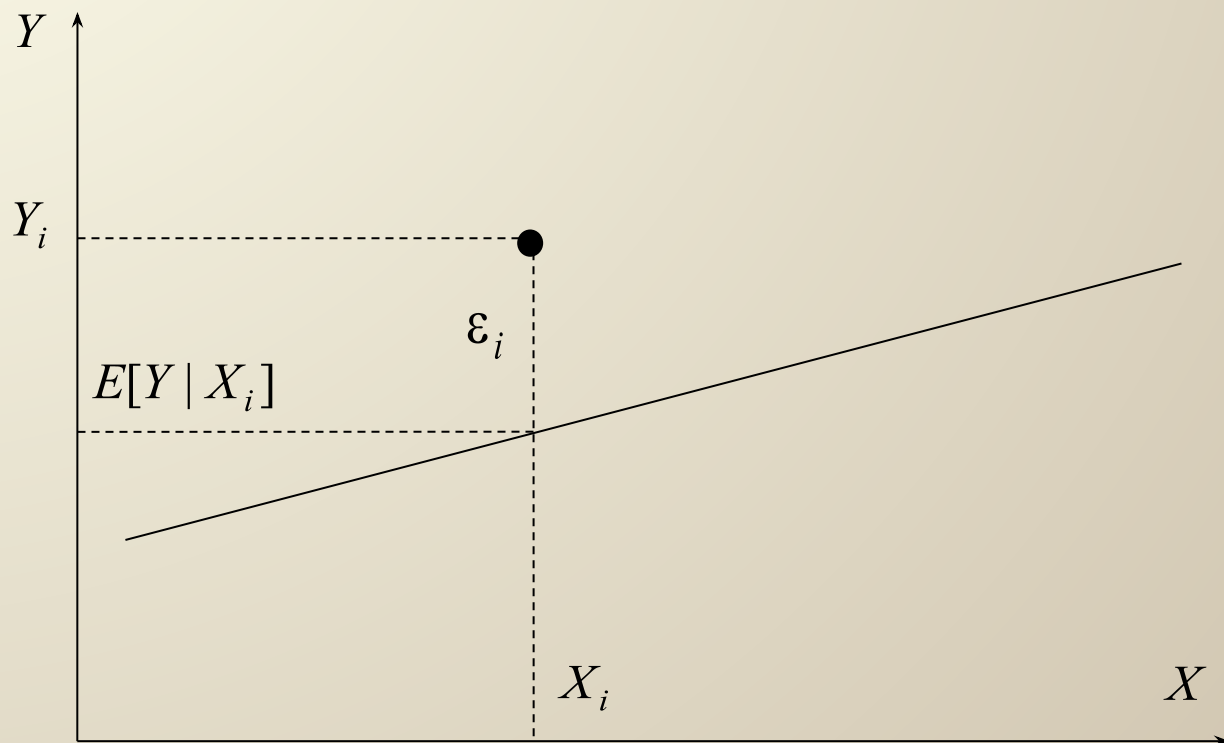
$$Y = f(X) + \varepsilon$$

называется уравнением регрессии
переменной Y на переменную X



Компоненты Y

0011 0010 1010 1101 0001 0100 1011



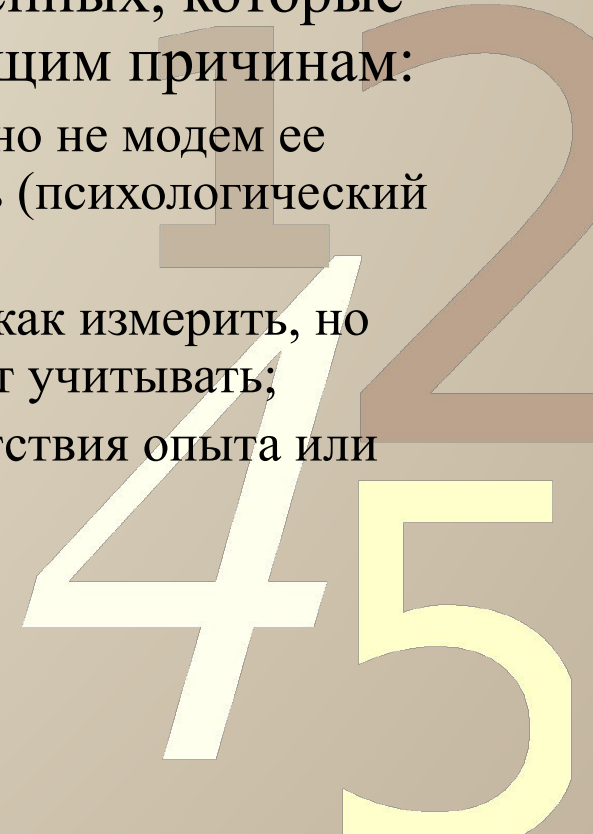
1 2

4 5

Экономический смысл ε

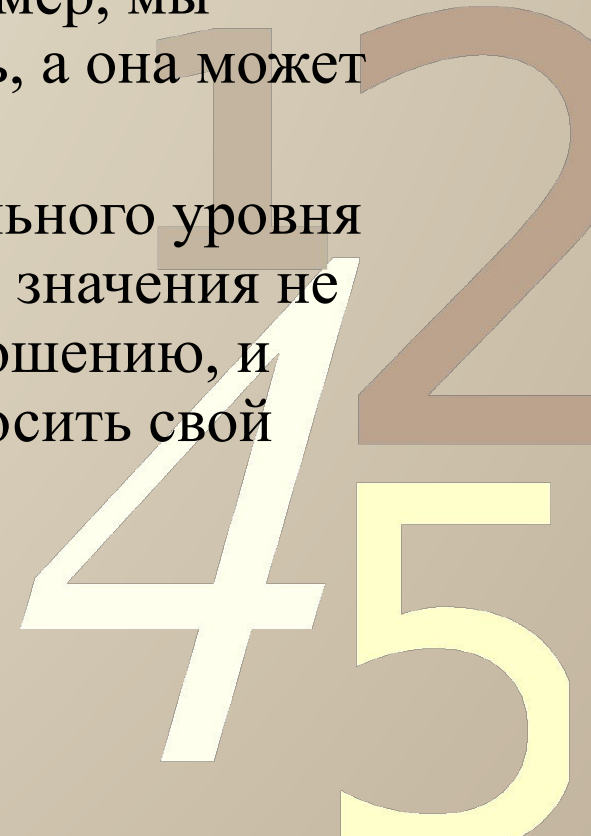
0011 0010 1010 1101 0001 0100 1011

- 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 можем 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 $f(X)$

0011 0010 1010 1101 0001 0100 1011

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



Эмпирический анализ данных

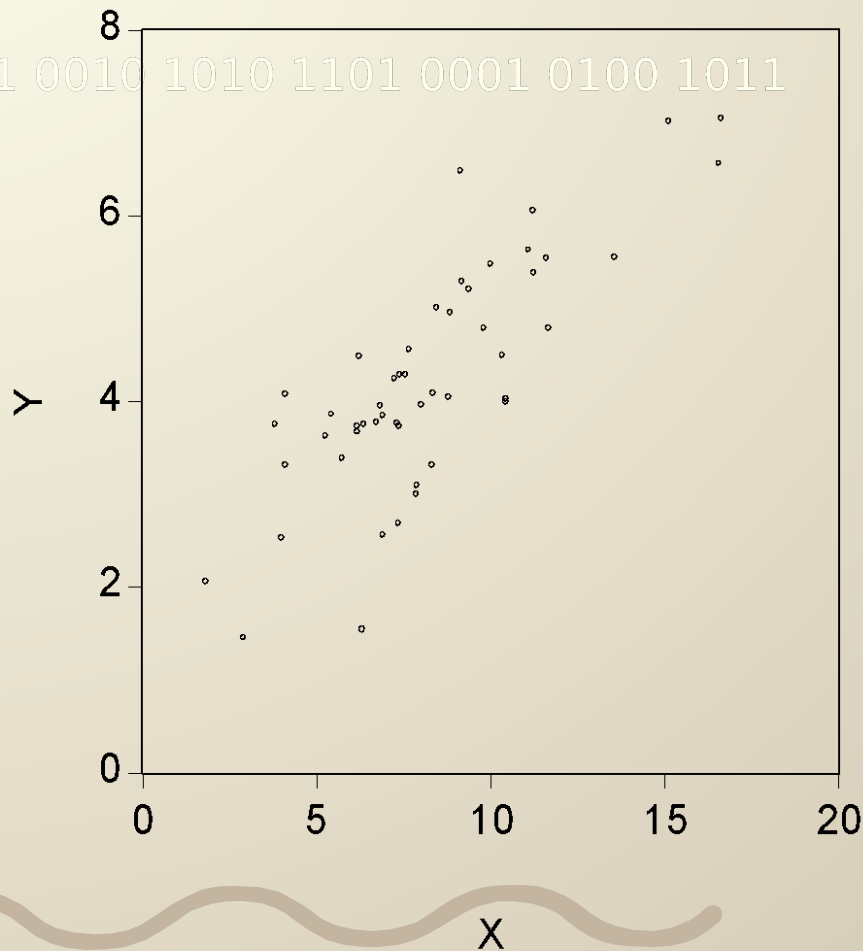
0011 0010 1010 1101 0001 0100 1011

В парном случае материал наблюдений представляет собой набор пар чисел:

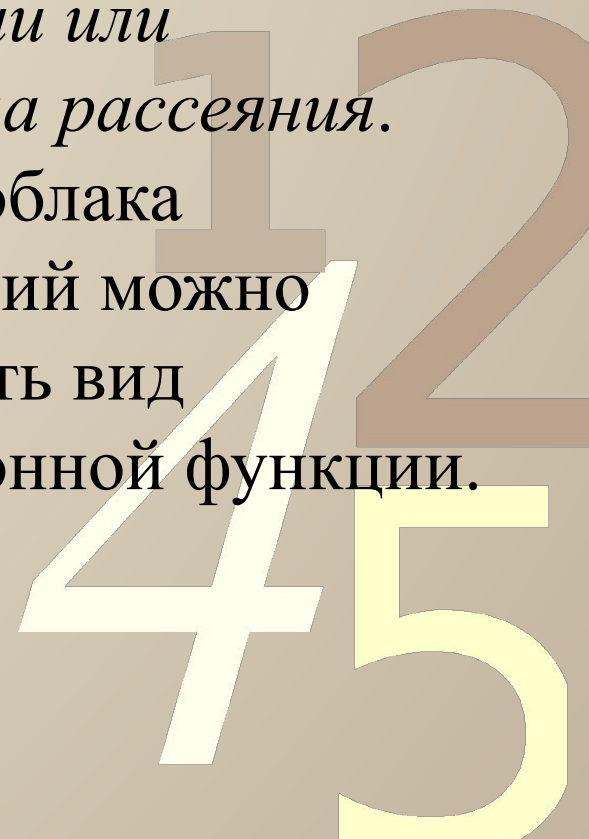
$$(X_i, Y_i) \quad i = 1, \dots, N$$



На плоскости каждому такому наблюдению соответствует точка:



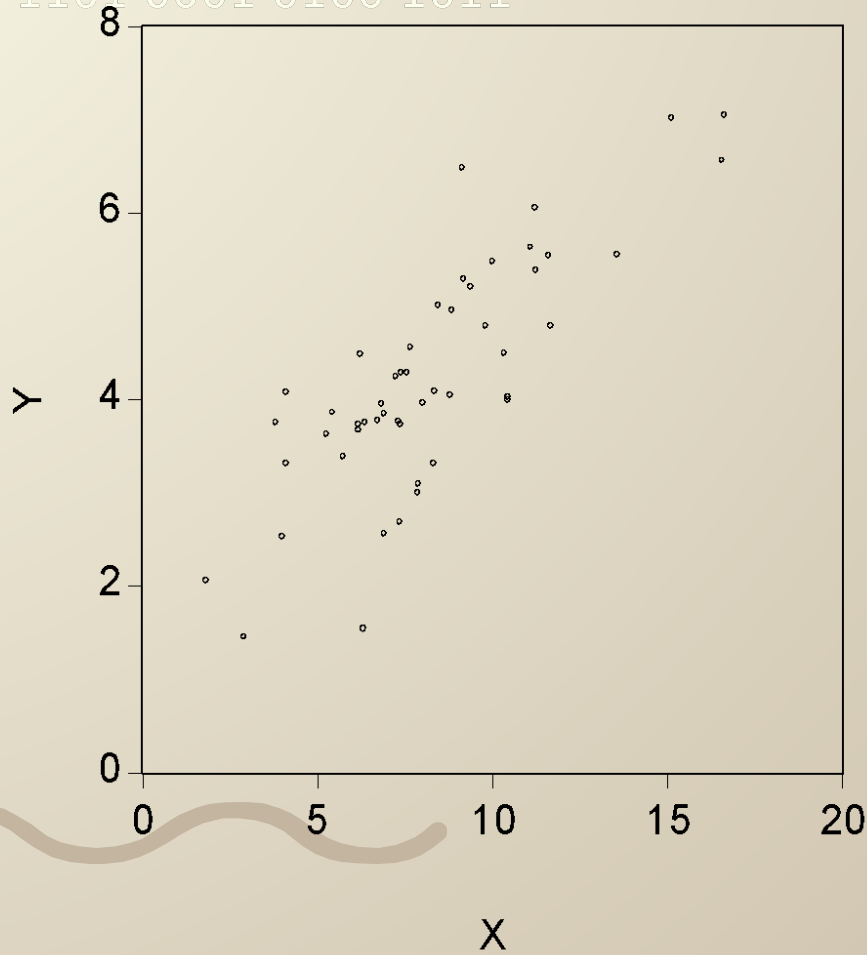
Полученный график называют облако наблюдений, поле корреляции или диаграмма рассеяния. По виду облака наблюдений можно определить вид регрессионной функции.



Линейная

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

0011 0010 1010 1101 0001 0100 1011

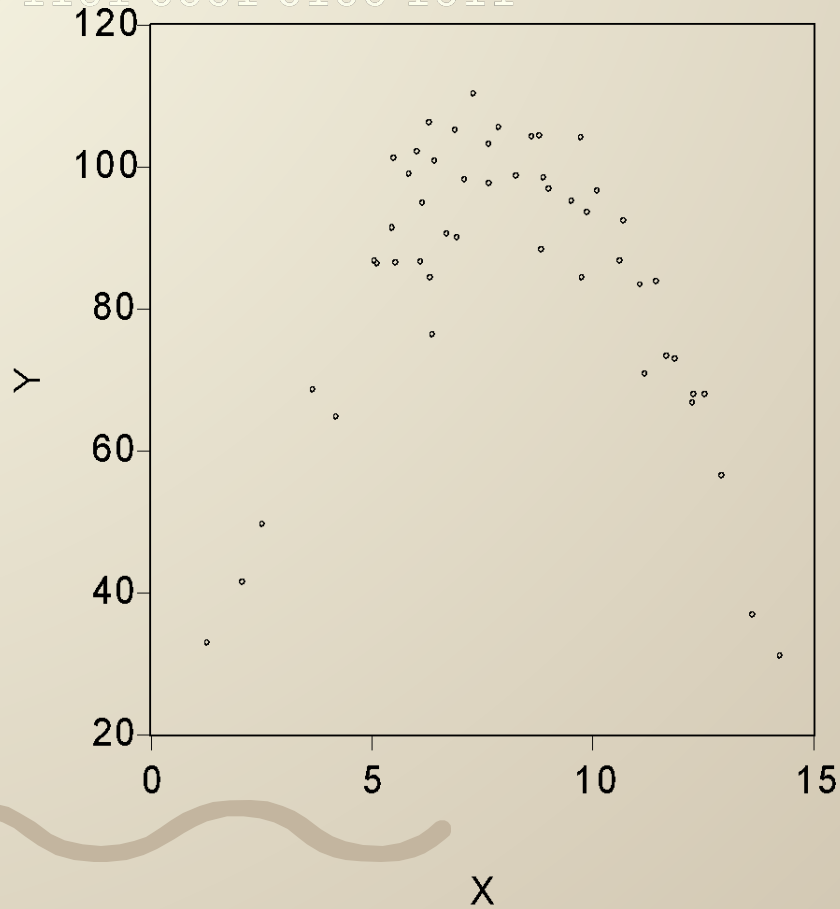


1 2
4 5

Квадратичная

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

0011 0010 1010 1101 0001 0100 1011

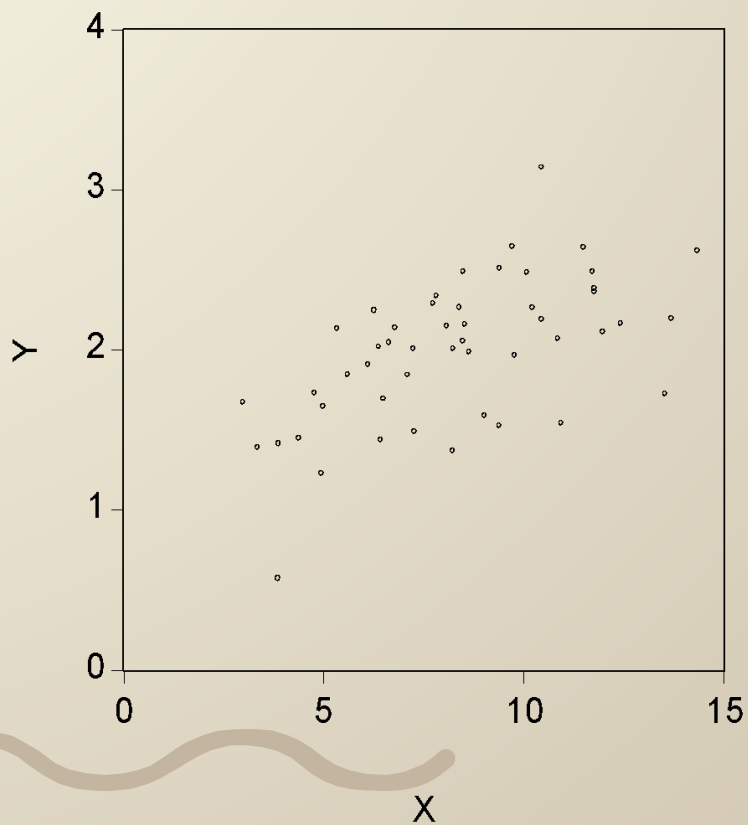


12
45

Показательная

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

0011 0010 1010 1101 0001 0100 1011

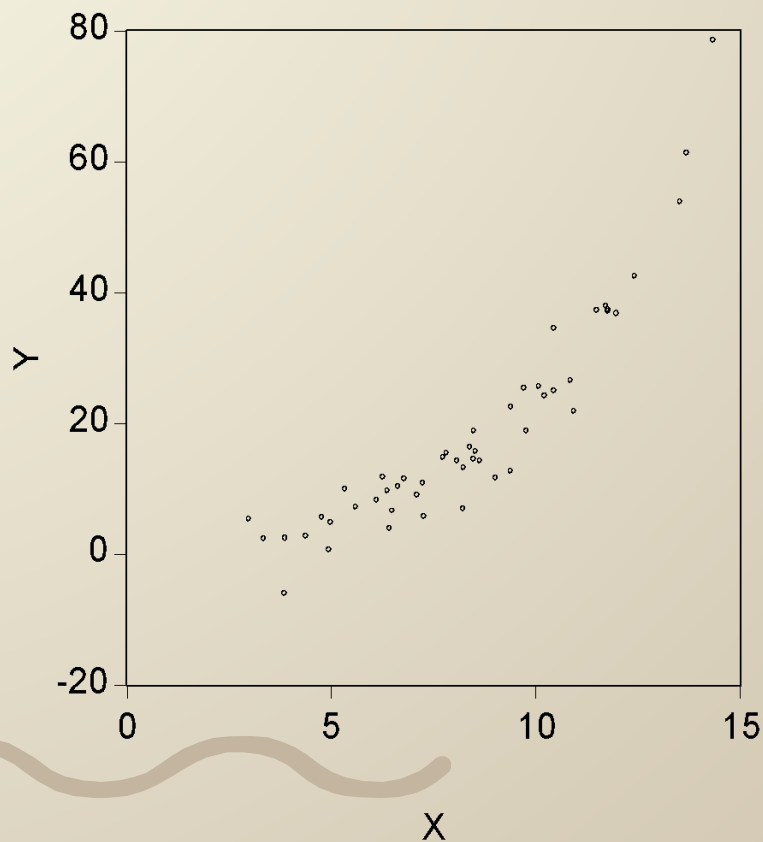


1 2
4 5

Степенная

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

0011 0010 1010 1101 0001 0100 1011

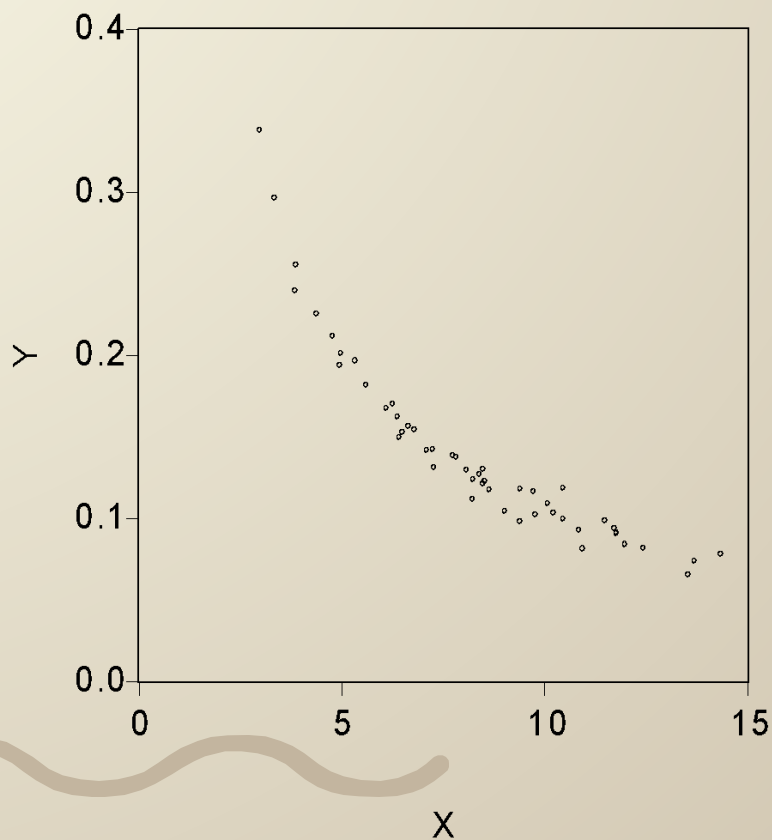


1 2
4 5

Гиперболическая

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

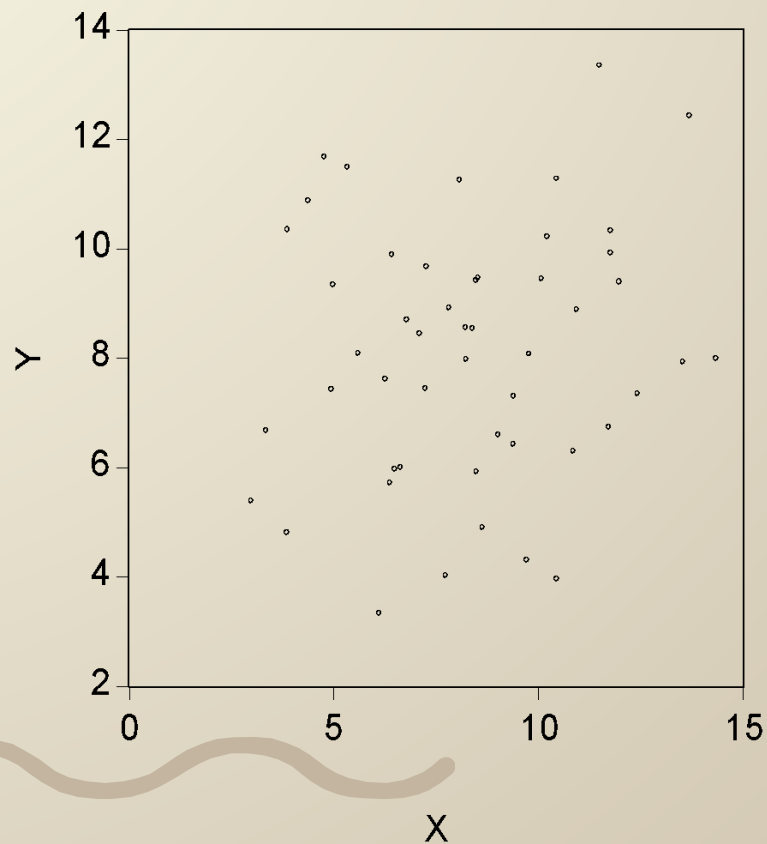
0011 0010 1010 1101 0001 0100 1011



1 2
4 5

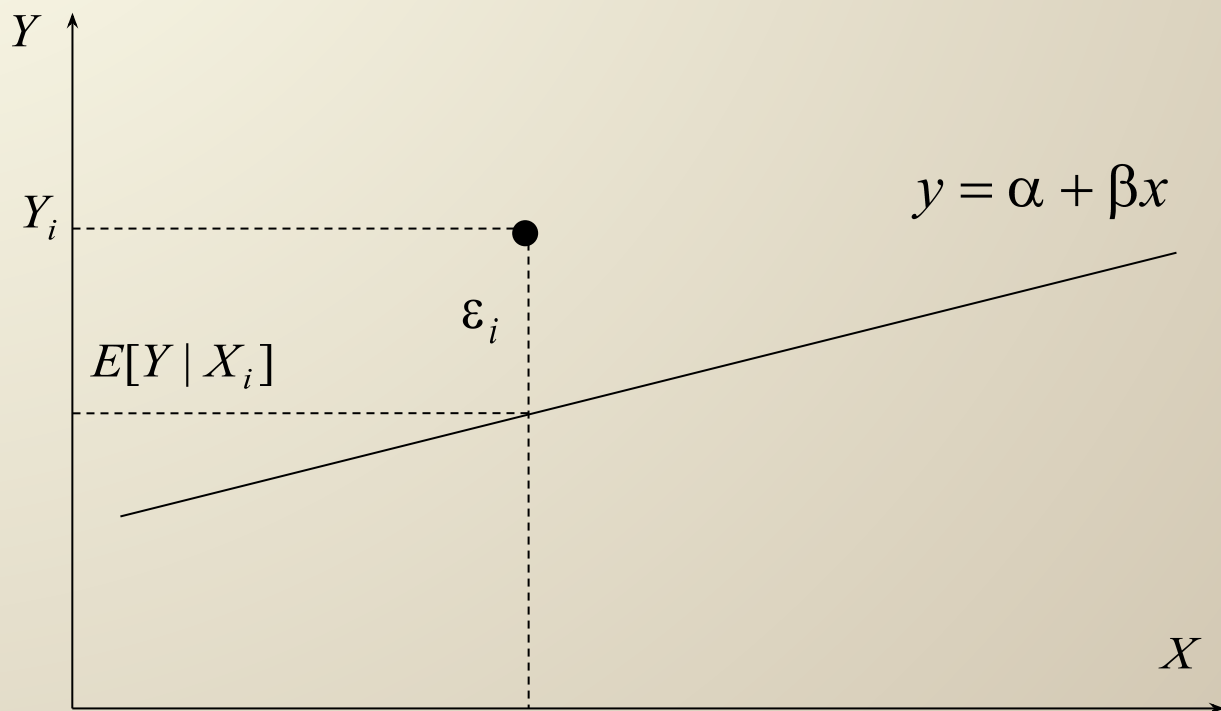
X и Y независимы

0011 0010 1010 1101 0001 0100 1011



1 2
4 5

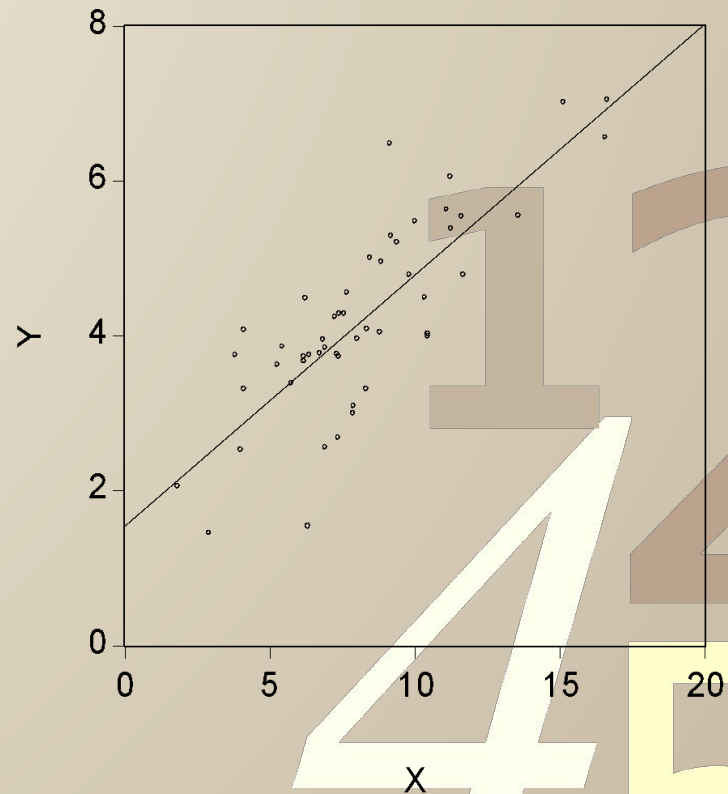
Парная линейная регрессионная модель $Y = \alpha + \beta X + \varepsilon$.



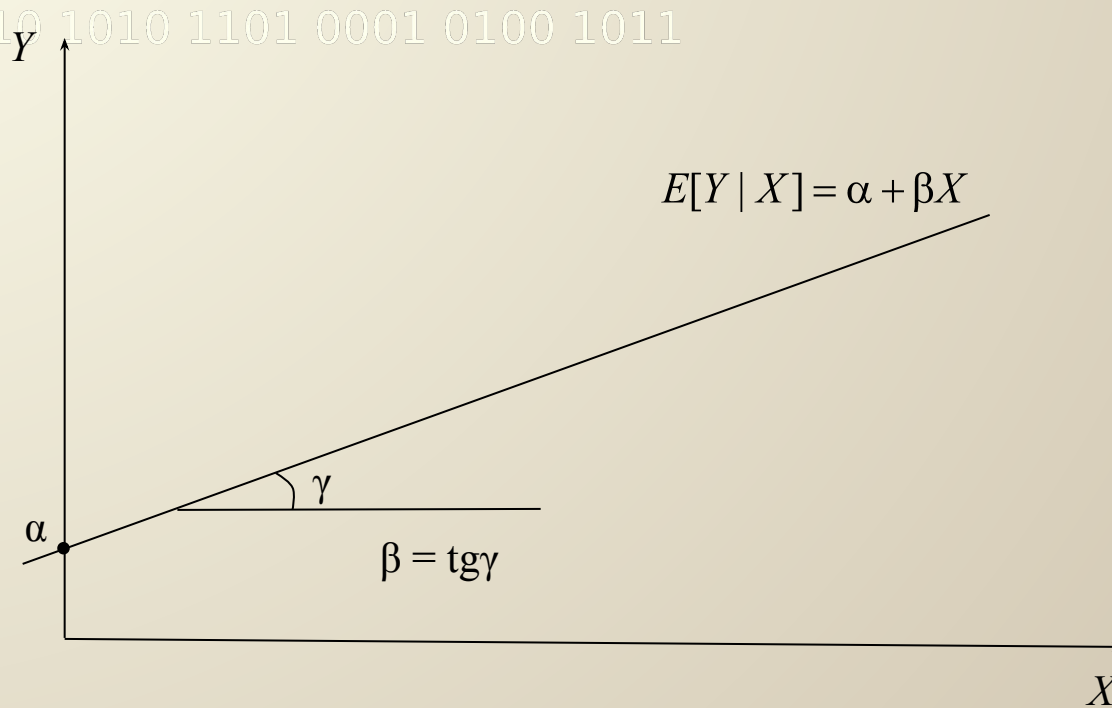
1 2
4 5

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.



Выбор коэффициентов регрессионной прямой

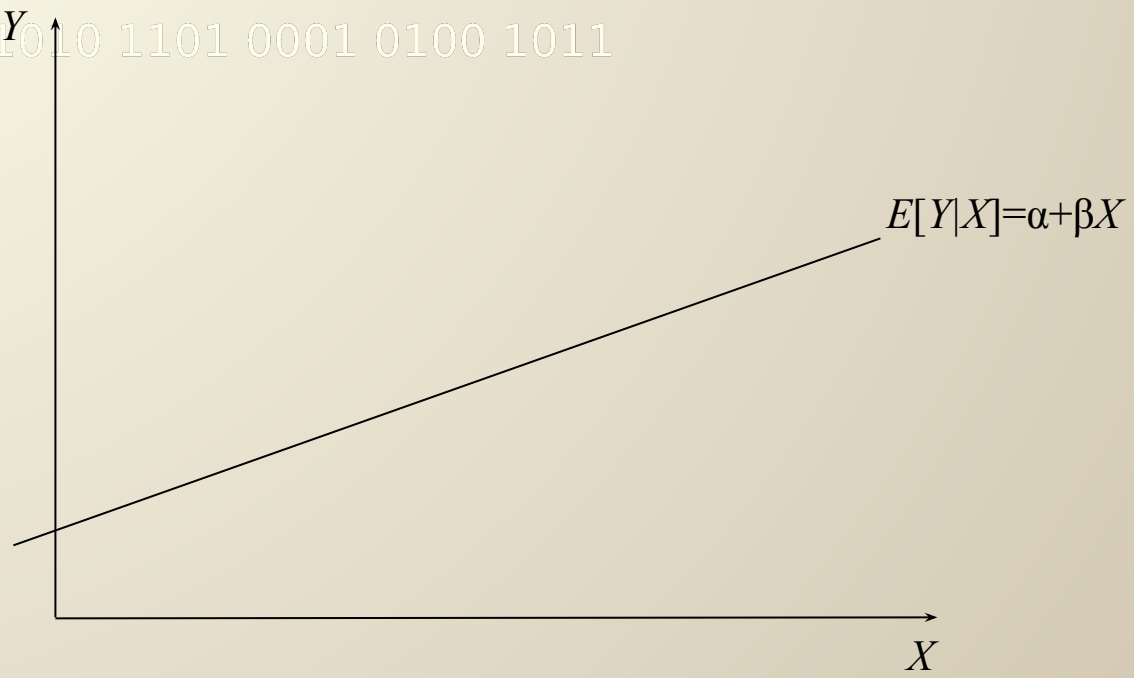


β – коэффициент наклона (slope),
 α – свободный коэффициент (intercept)



Истинная линия регрессии, определяемая коэффициентами α и β

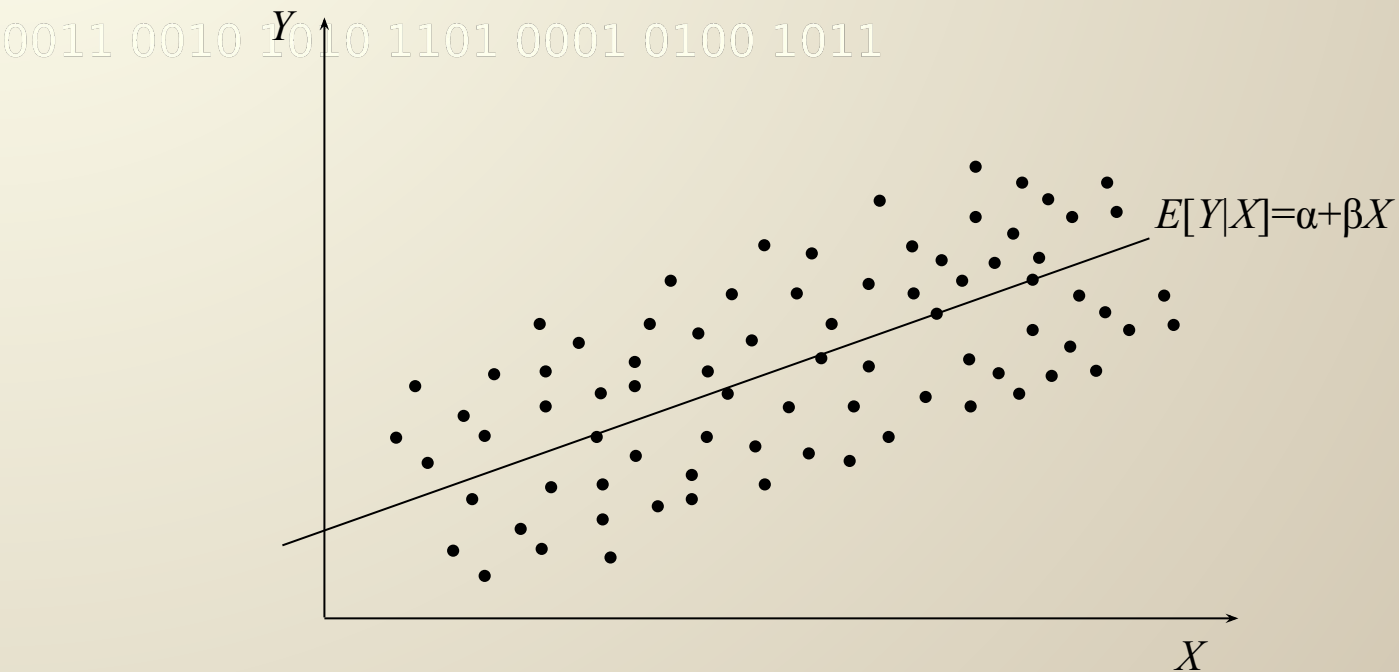
0011 0010 1010 1101 0001 0100 1011



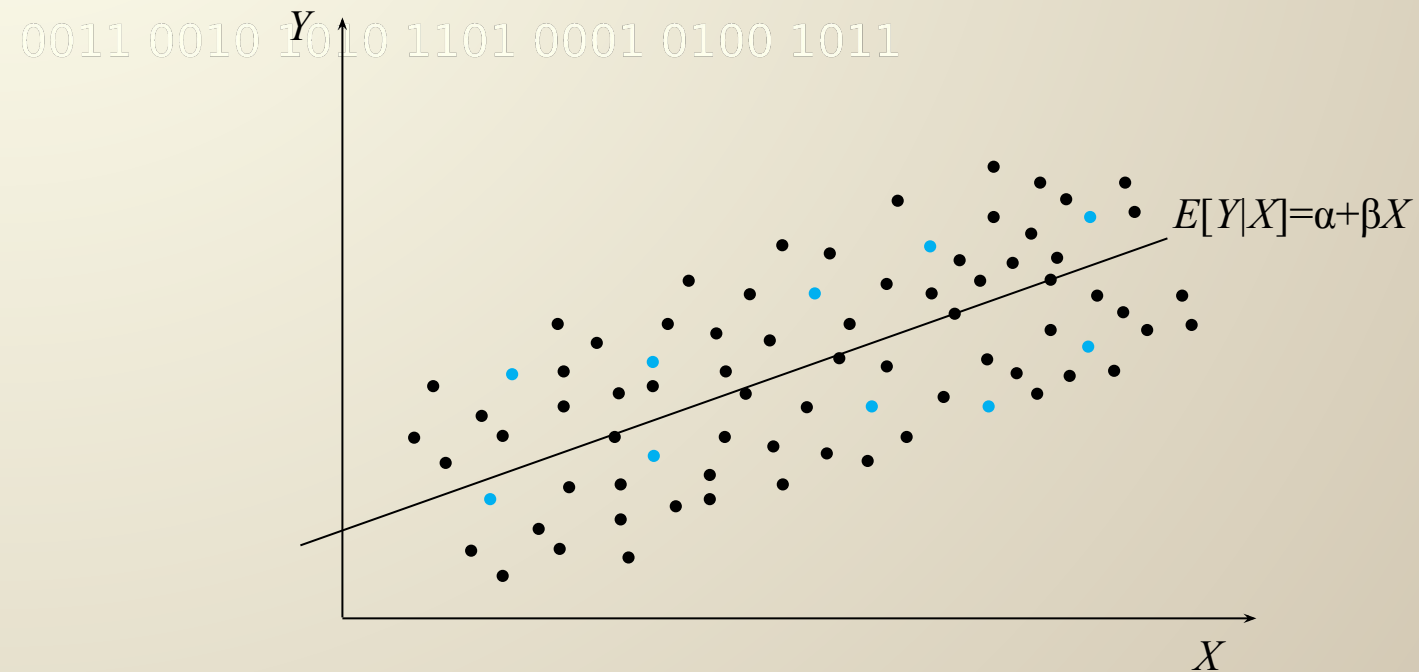
1 2
4 5



Точки наблюдений разбросаны вокруг этой линии. Их бесконечность.

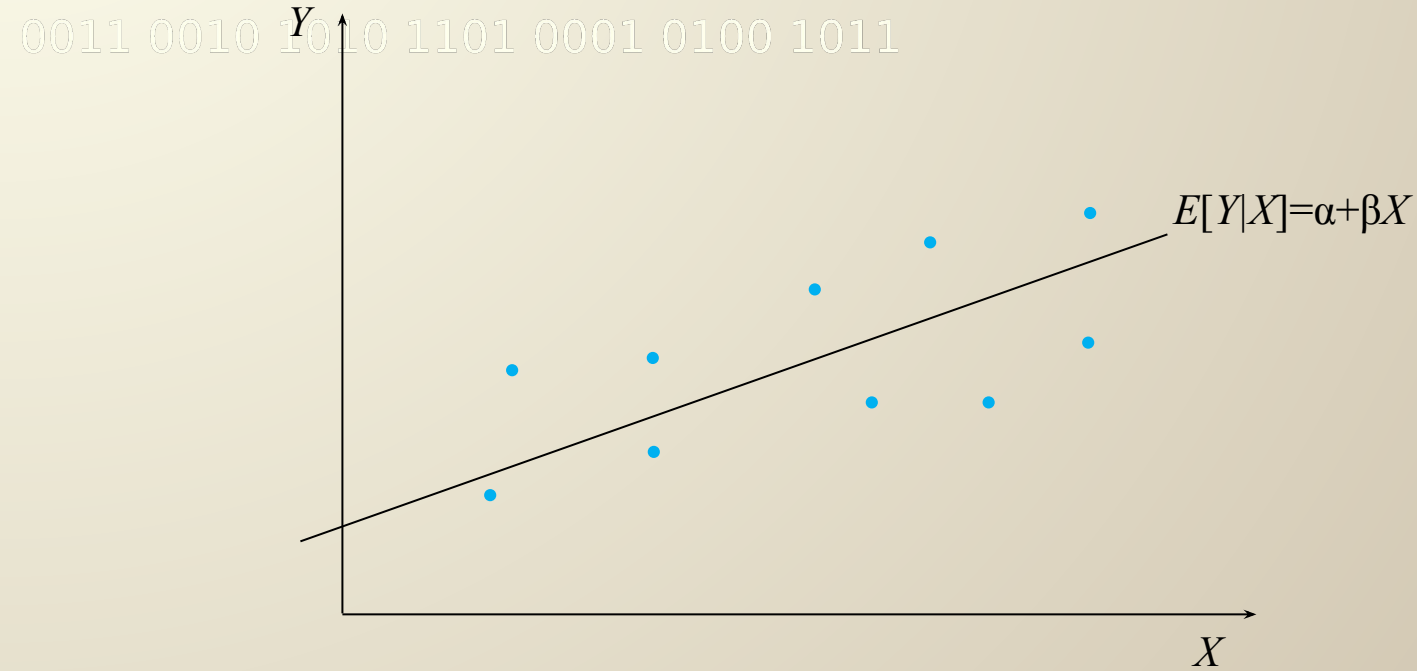


В выборку попадает только их часть



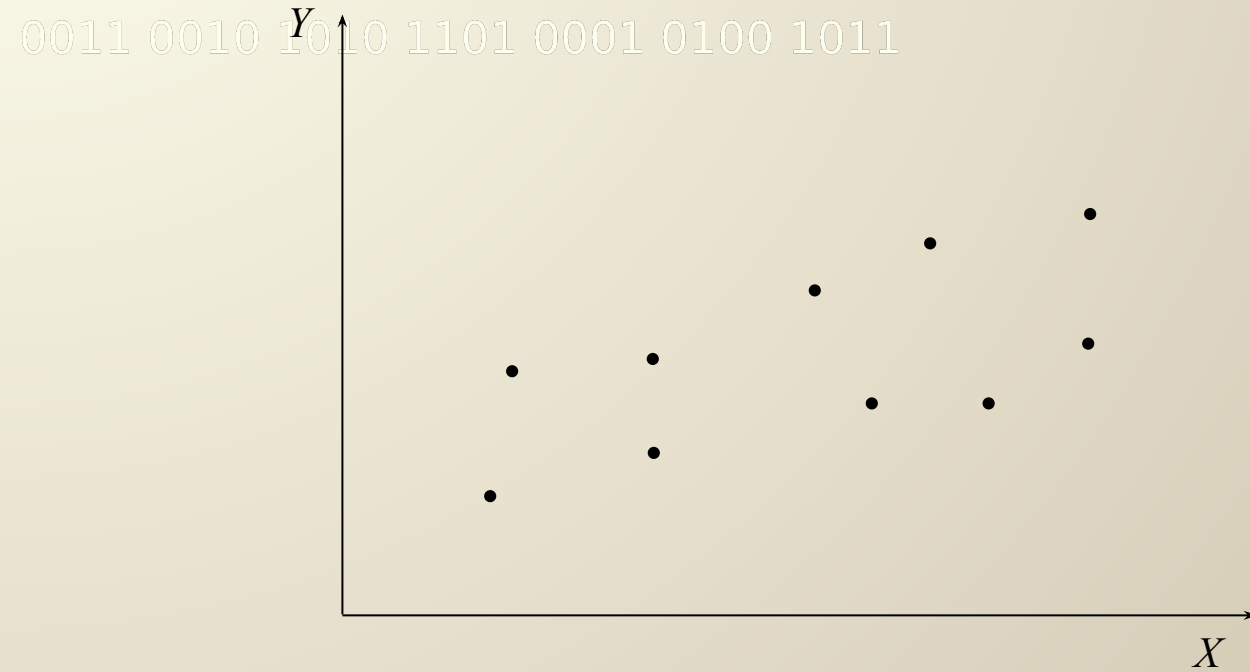
1 2
4 5

В выборку попадает только их часть



1 2
4 5

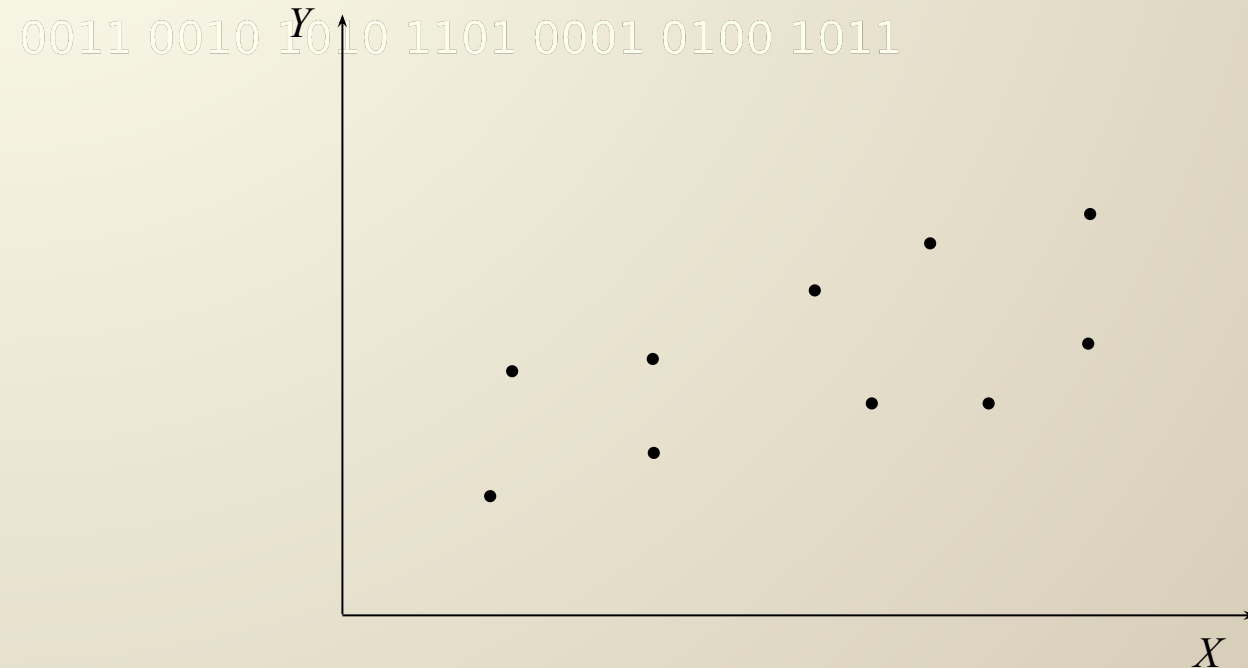
И что мы наблюдаем



Всего мы наблюдаем N точек

1 2
4 5

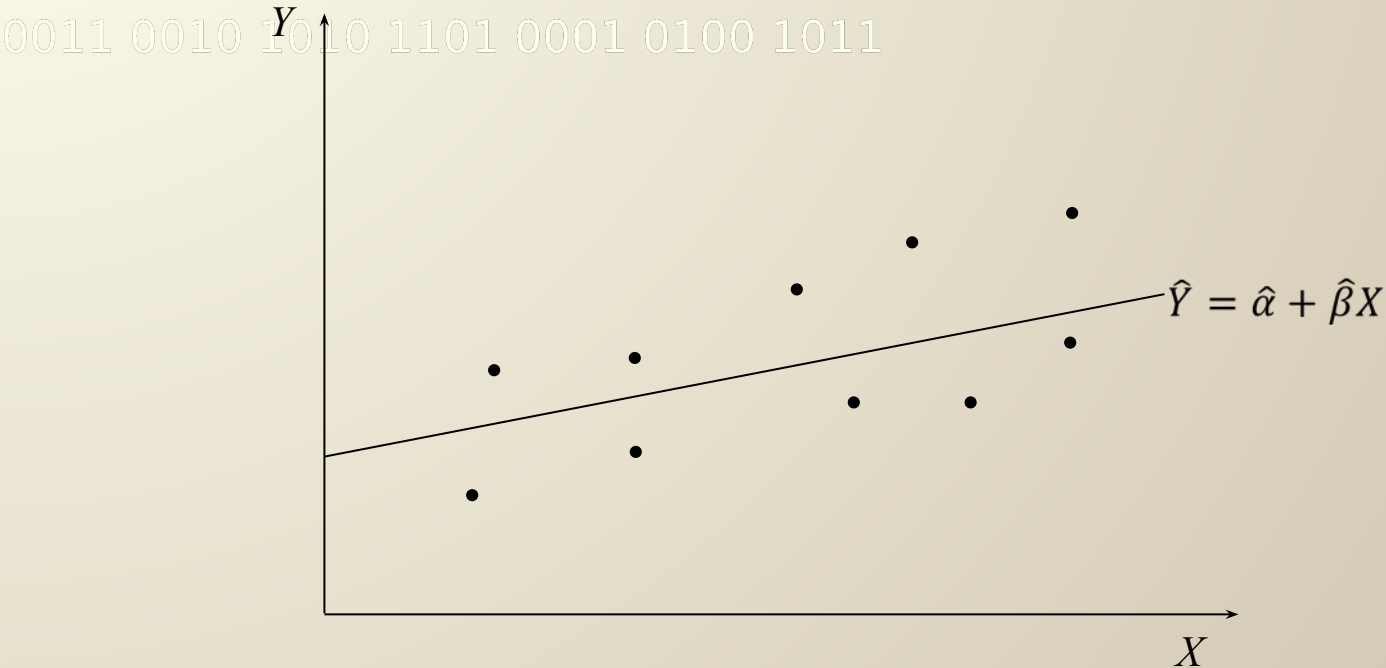
Линия, которую мы проводим



Мы используем эти N точек для того, чтобы построить аппроксимацию неизвестной линии регрессии $E[Y|X] = \alpha + \beta X$. Обозначим эту линию $E[\widehat{Y|X}] = \hat{Y} = \hat{\alpha} + \hat{\beta}X$, где $\hat{\alpha}$ - статистическая оценка α , а $\hat{\beta}$ - оценка β . \hat{Y} называется прогнозным значением переменной Y

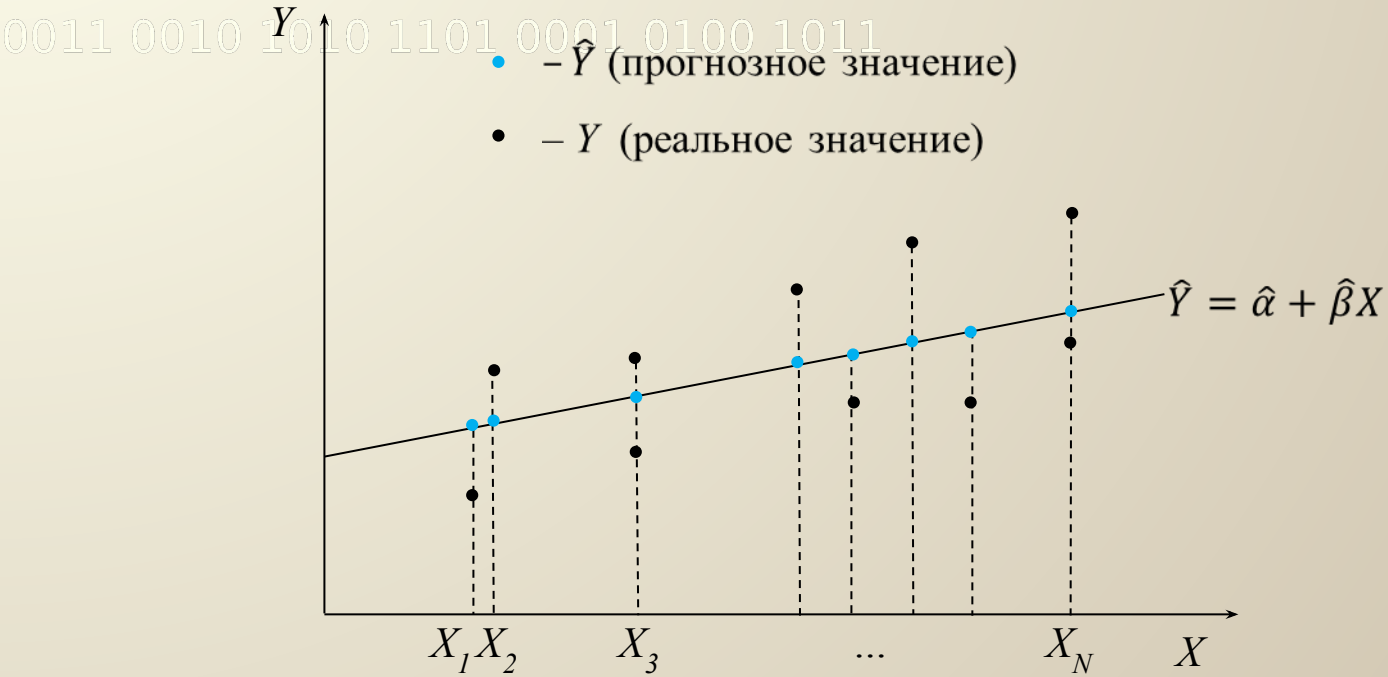


Линия, которую мы проводим



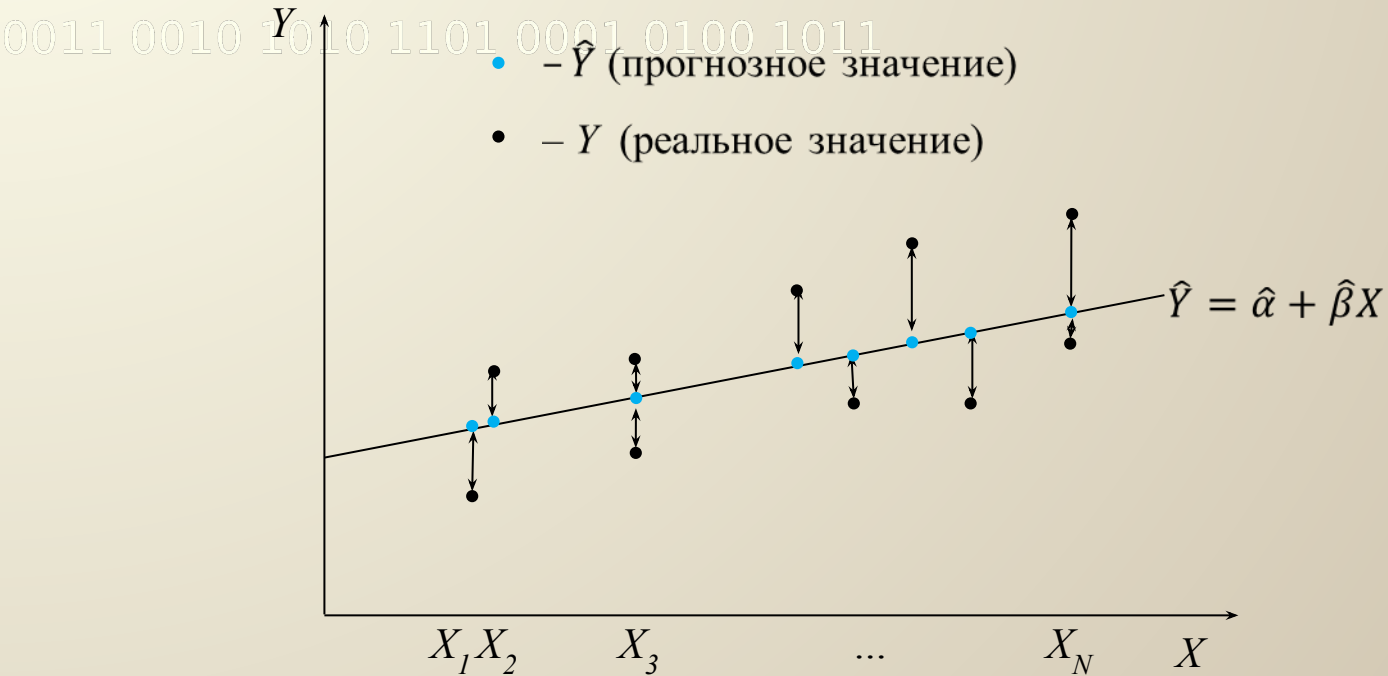
Проводим прямую через центр скопления точек облака наблюдений, т. е. таким образом, чтобы точки облака наблюдений были одновременно к этой линии близки.

Реальные и прогнозные значения



1 2
4 5

Разницу между реальным и прогнозным значением назовем остатком



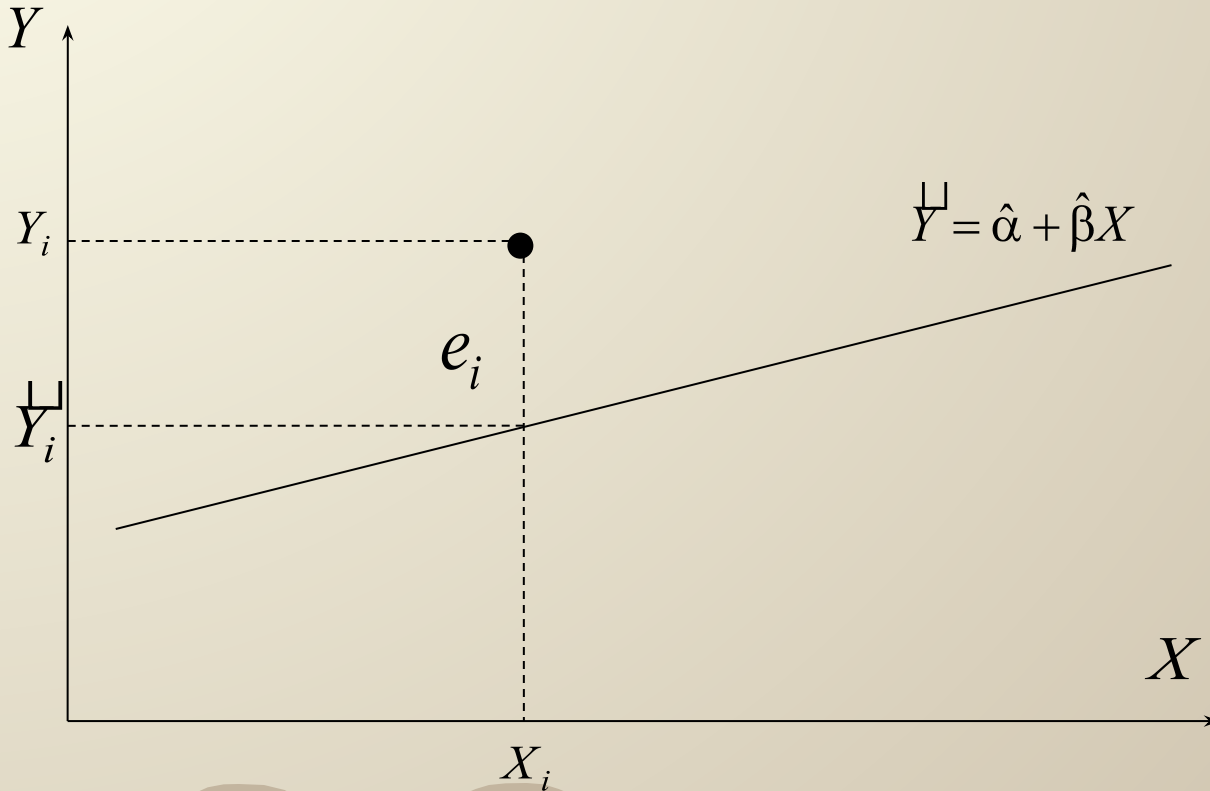
$$e_i = Y_i - \hat{Y}_i$$

остатки могут быть положительными и отрицательными



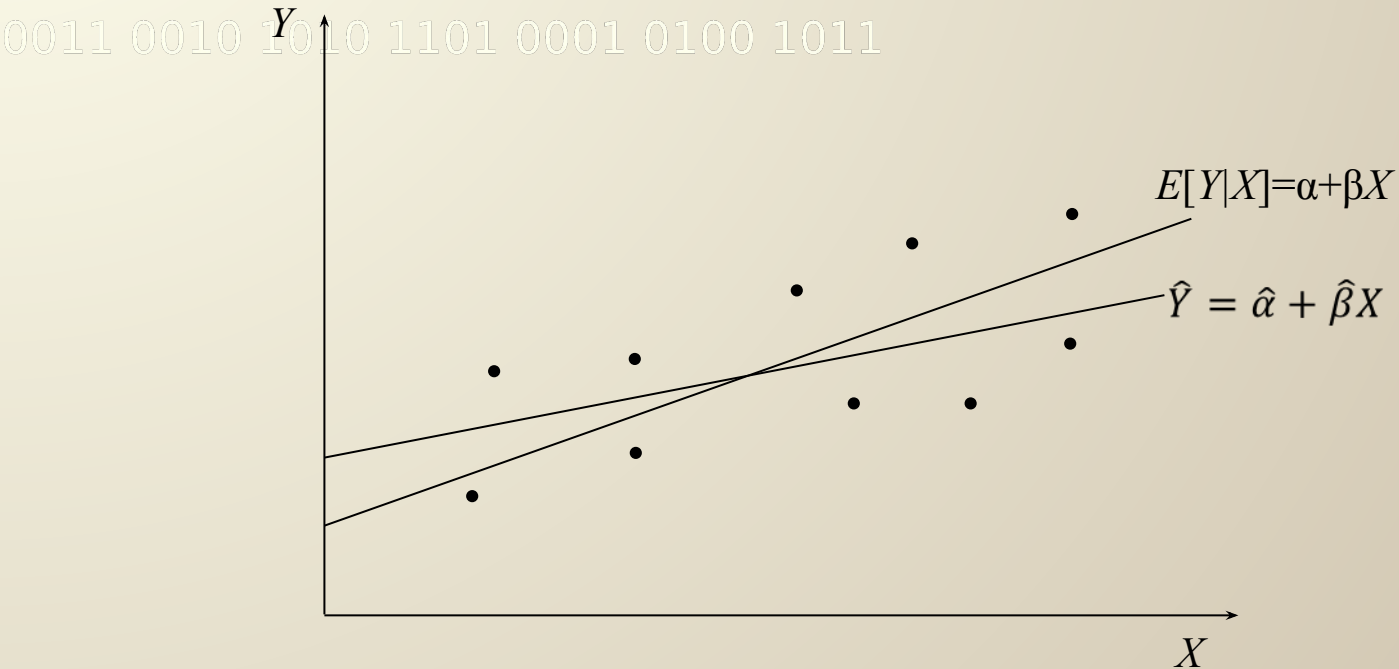
Рассмотрение остатков на графике

0011 0010 1010 1101 0001 0100 1011



1 2
4 5

Истинная и оцененная линия регрессии



Мы надеемся, что построенная линия регрессии не очень сильно отличается от истинной, в частности, чем больше объем выборки, тем шансы на то, что линии похожи, возрастают (состоятельность).



Грусть печаль

0011 0010 1010 1101 0001 0100 1011

Метод наименьших квадратов не всегда состоятельный

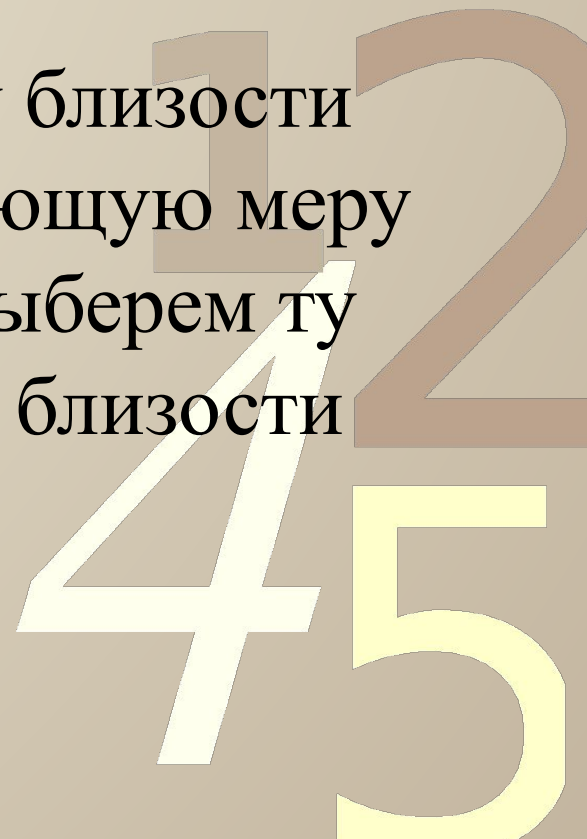


1 2
4 5

Как найти «наилучшую» прямую аналитически?

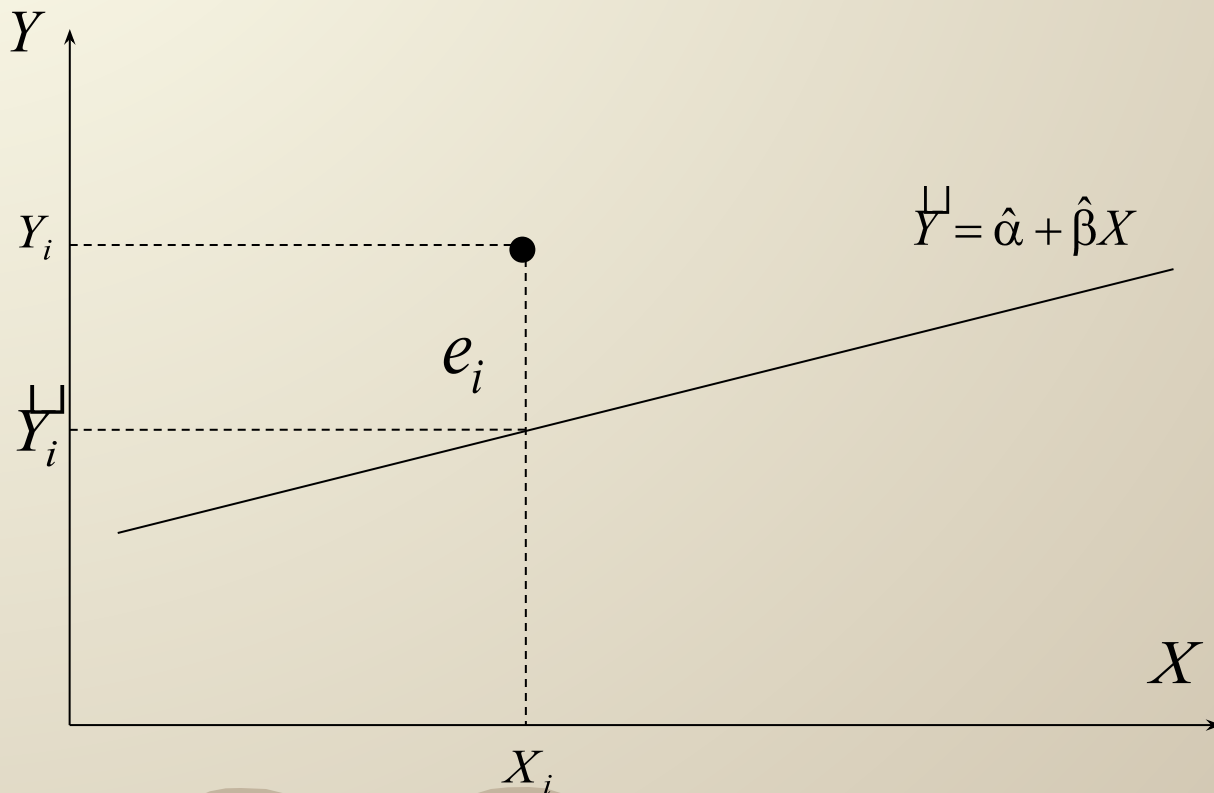
0011 0010 1010 1101 0001 0100 1011

- Выберем меру близости одной точки к прямой.
- Построим интегральную меру близости всех точек к прямой, учитывающую меру близости отдельных точек и выберем ту прямую, для которой эта мера близости минимальна.



Мера близости одной точки к прямой – остаток.

0011 0010 1010 1101 0001 0100 1011



$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta}X_i$$



Интегральная мера близости

0011 0010 1010 1101 0001 0100 1011

$$\sum_{i=1}^N e_i^2 = \sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2 = RSS \rightarrow \min_{(\hat{\alpha}, \hat{\beta})}$$

$$\sum_{i=1}^N |e_i| = \sum_{i=1}^N |Y_i - \hat{\alpha} - \hat{\beta}X_i| \rightarrow \min_{(\hat{\alpha}, \hat{\beta})}$$



Интегральная мера близости

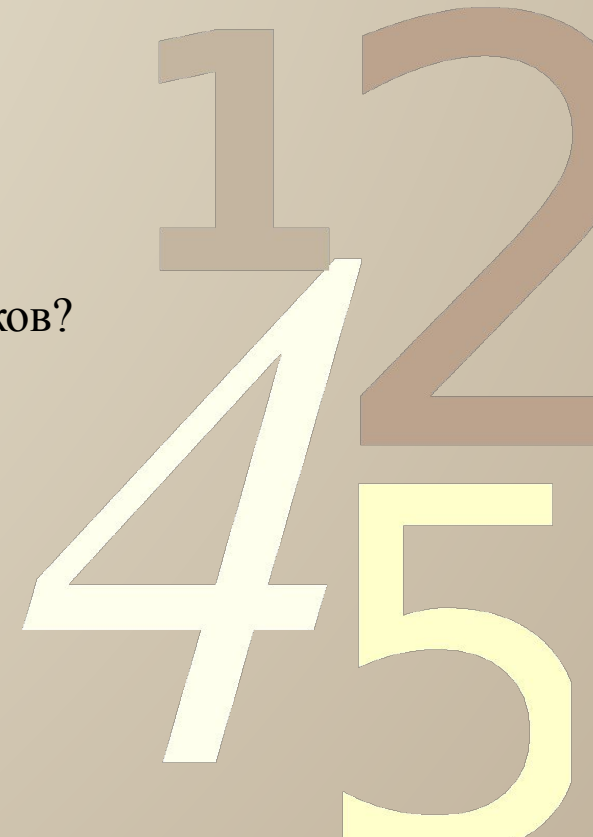
0011 0010 1010 1101 0001 0100 1011

$$\sum_{i=1}^N e_i^2 = \sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2 = RSS \rightarrow \min_{(\hat{\alpha}, \hat{\beta})}$$

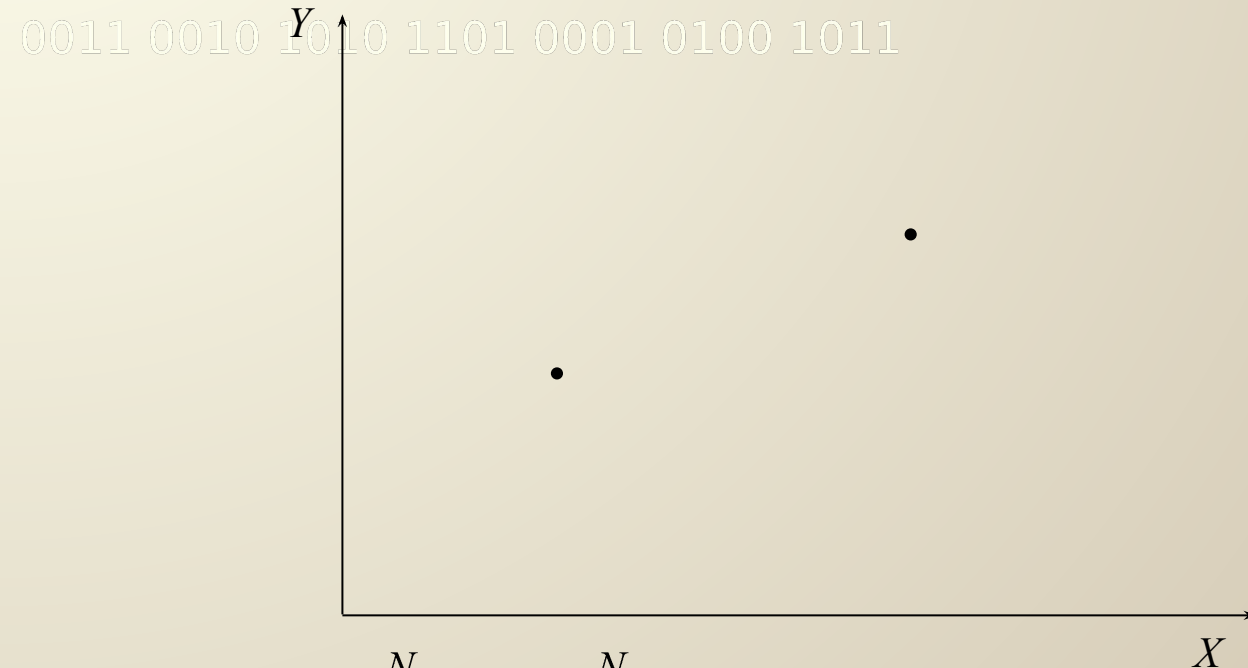
$$\sum_{i=1}^N |e_i| = \sum_{i=1}^N |Y_i - \hat{\alpha} - \hat{\beta}X_i| \rightarrow \min_{(\hat{\alpha}, \hat{\beta})}$$

почему бы не минимизировать просто сумму остатков?

$$\sum_{i=1}^N e_i = \sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta}X_i) \rightarrow \min_{(\hat{\alpha}, \hat{\beta})}$$



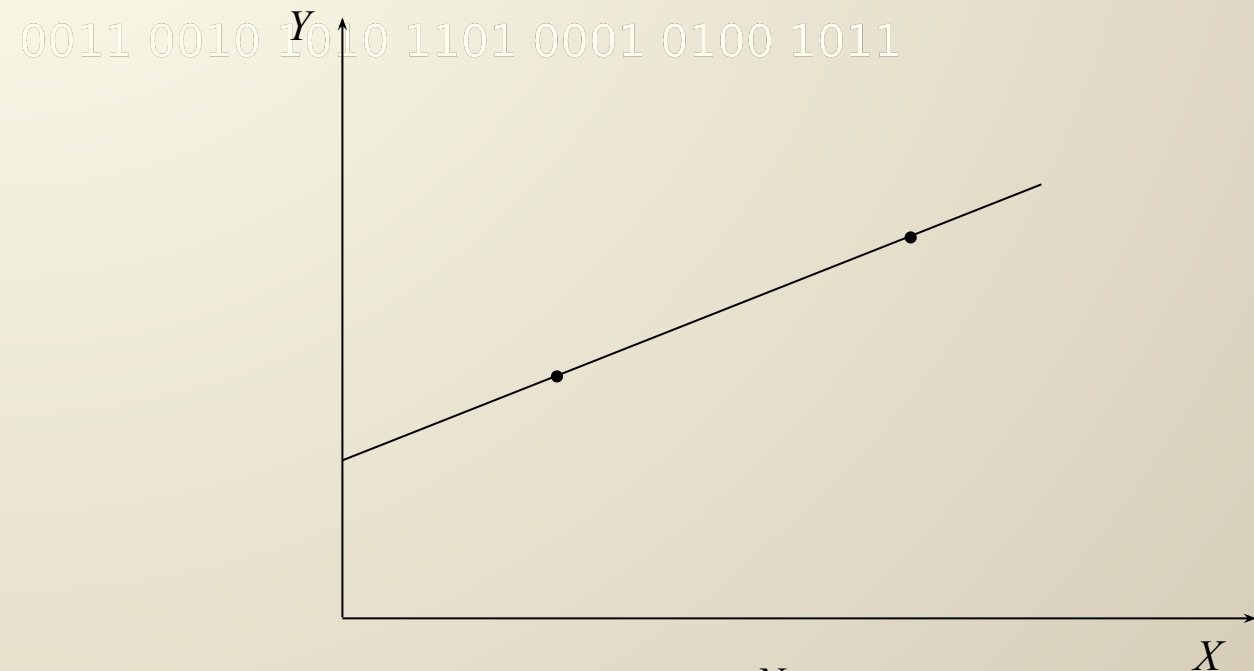
Для какой прямой сумма остатков равна 0?



$$\sum_{i=1}^N e_i = \sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta}X_i) \rightarrow \min_{(\hat{\alpha}, \hat{\beta})}$$

1 2
4 5

для такой



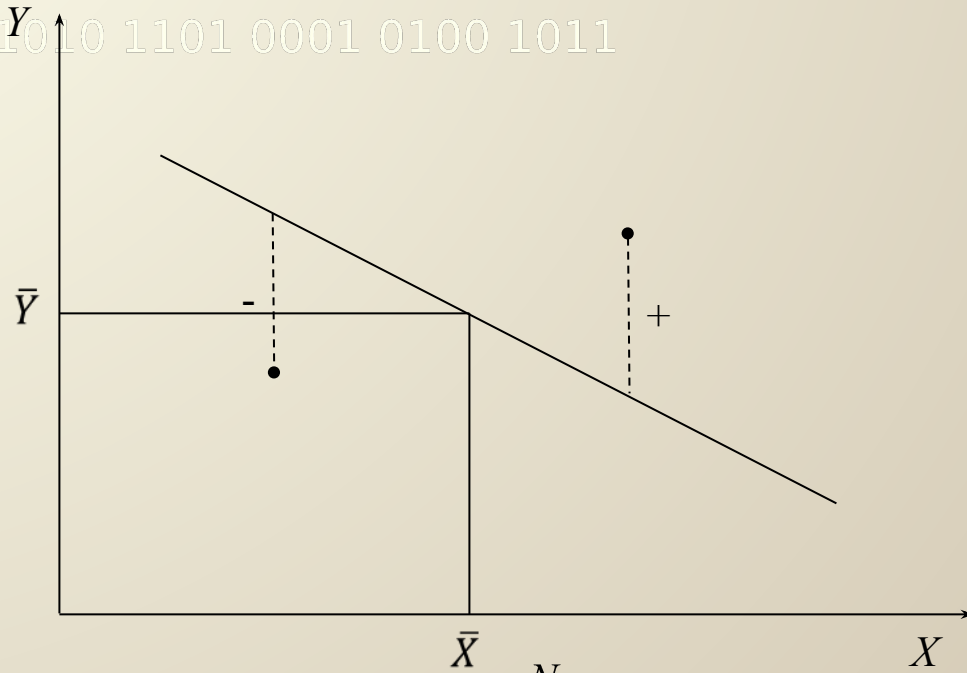
$$e_1 = e_2 = 0$$

$$\sum_{i=1}^N e_i = e_1 + e_2 = 0$$

1 2
4 5

и для такой

0011 0010 1010 1101 0001 0100 1011



$$e_1 = -e_2$$

$$\sum_{i=1}^N e_i = e_1 + e_2 = 0$$

1 2
4 5

Метод наименьших квадратов

0011 0010 1010 1101 0001 0100 1011

$$\sum_{i=1}^N e_i^2 \rightarrow \min_{(\hat{\alpha}, \hat{\beta})}$$

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



Минимизация

$$\frac{\partial S}{\partial \hat{\alpha}} = -2 \sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0$$

$$\frac{\partial S}{\partial \hat{\beta}} = -2 \sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta} X_i) X_i = 0$$

$$\begin{cases} \sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0 \\ \sum_{i=1}^N (Y_i - \hat{\alpha} - \hat{\beta} X_i) X_i = 0 \end{cases}$$

или

$$\begin{cases} \sum_{i=1}^N e_i = 0 \\ \sum_{i=1}^N X_i e_i = 0 \end{cases}$$



Система нормальных уравнений

0011 0010 1010 1101 0001 0100 1011

$$\left\{ \begin{array}{l} N\hat{\alpha} + \hat{\beta} \sum_{i=1}^N X_i = \sum_{i=1}^N Y_i \\ \hat{\alpha} \sum_{i=1}^N X_i + \hat{\beta} \sum_{i=1}^N X_i^2 = \sum_{i=1}^N X_i Y_i \end{array} \right.$$

1 2
4 5

МНК-коэффициенты ПЛРМ

0011 0010 1010 1101 0001 0100 1011

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

- коэффициент наклона

$$\alpha = \bar{Y} - \beta \bar{X}$$

- свободный коэффициент



Другие формы записи коэффициента наклона

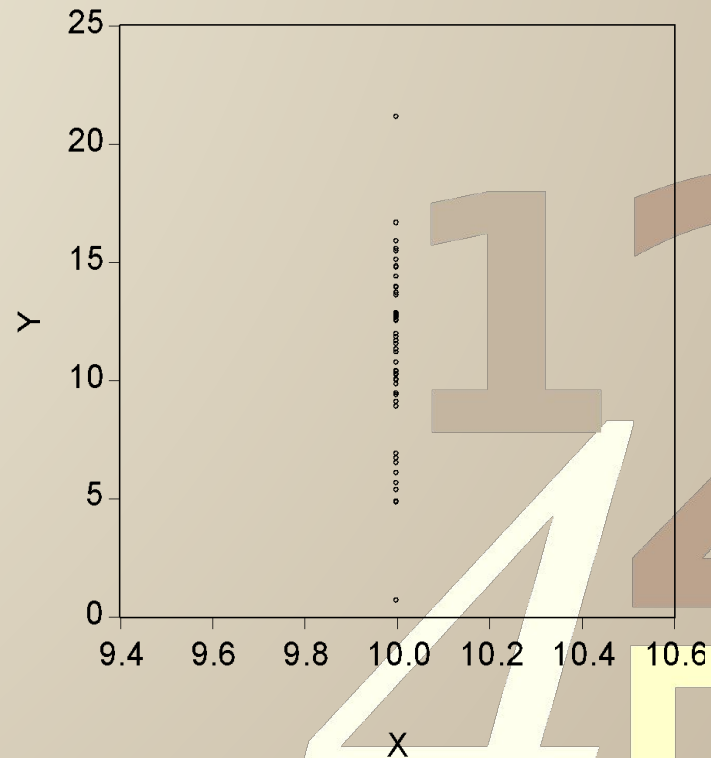
$$\beta = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^N (X_i - \bar{X})^2} = \frac{Cov(X, Y)}{Var(X)}$$

1 2
4 5

Замечания

0011 0010 1010 1101 0001 0100 1011

- Линия регрессии проходит через точку (\bar{X}, \bar{Y})
- Мы предполагаем, что среди X_i есть разные, тогда $d_X \neq 0$. 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):

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

Вспомним теоретический коэффициент корреляции

$$\text{corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) * \text{Var}(Y)}}$$

$$r_{xy} = \frac{\frac{\sum_{i=1}^N X_i Y_i}{N} - \bar{X}\bar{Y}}{\sqrt{\left(\frac{\sum_{i=1}^N X_i^2}{N} - (\bar{X})^2\right) \left(\frac{\sum_{i=1}^N Y_i^2}{N} - (\bar{Y})^2\right)}} = \frac{\hat{\text{Cov}}(X, Y)}{\sqrt{\hat{\text{Var}}(X) * \hat{\text{Var}}(Y)}}$$

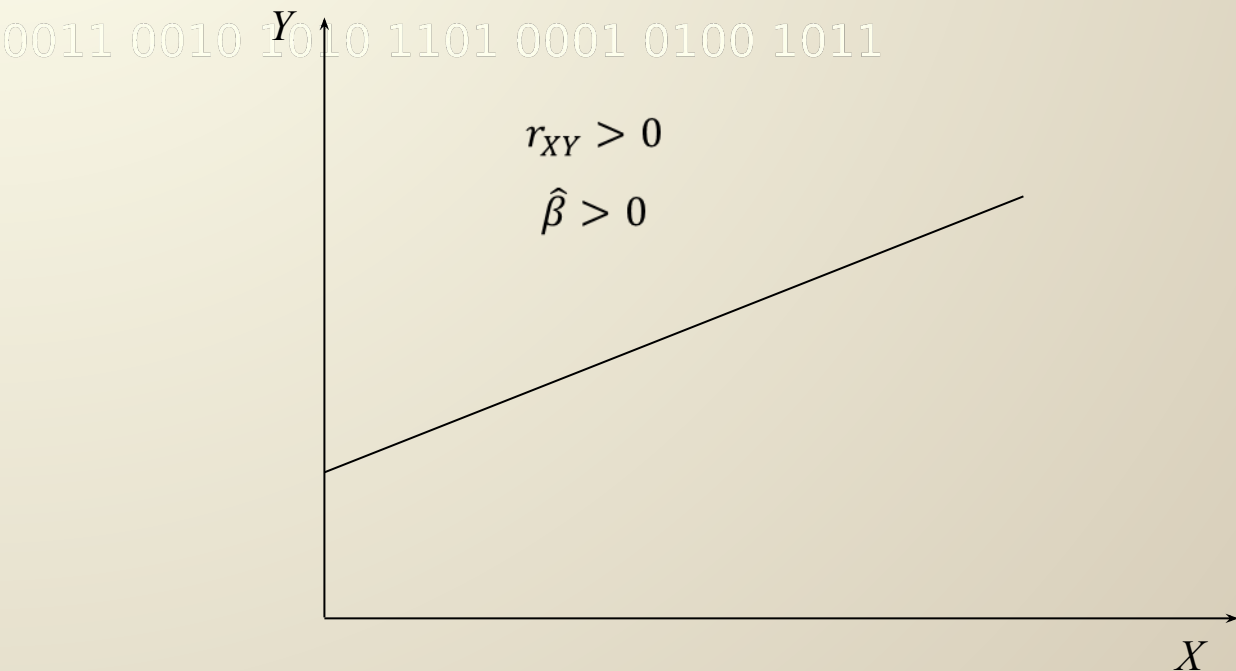
Связь между коэффициентом корреляции и коэффициентом наклона

$$\beta = r_{XY} \frac{d_Y}{d_X} \quad r_{XY} = \beta \frac{d_X}{d_Y}$$

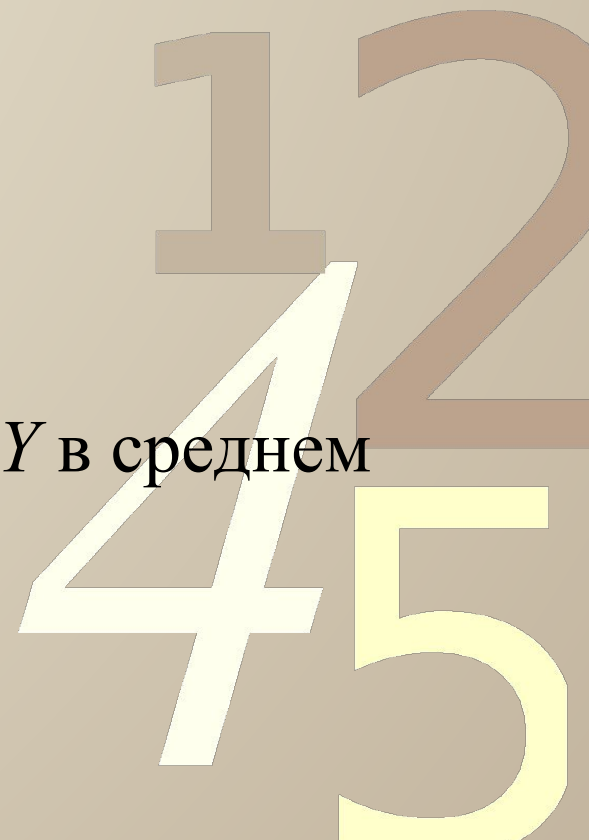
Знак коэффициента наклона линии регрессии и коэффициента корреляции совпадают



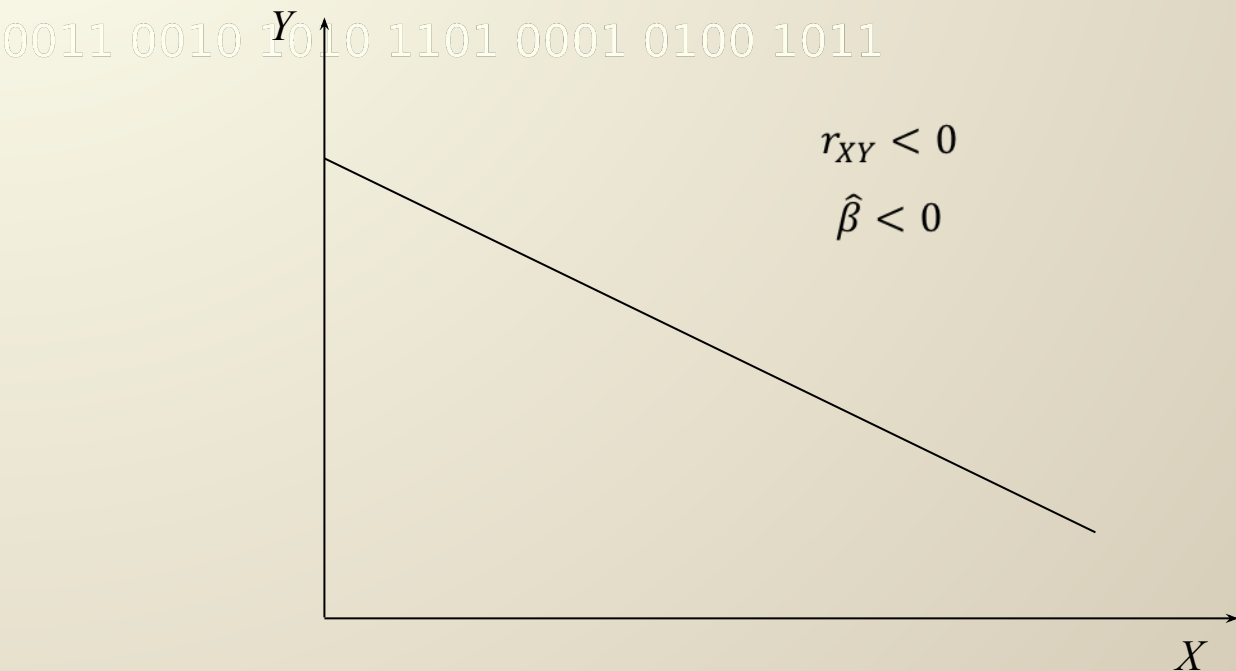
Положительная корреляция



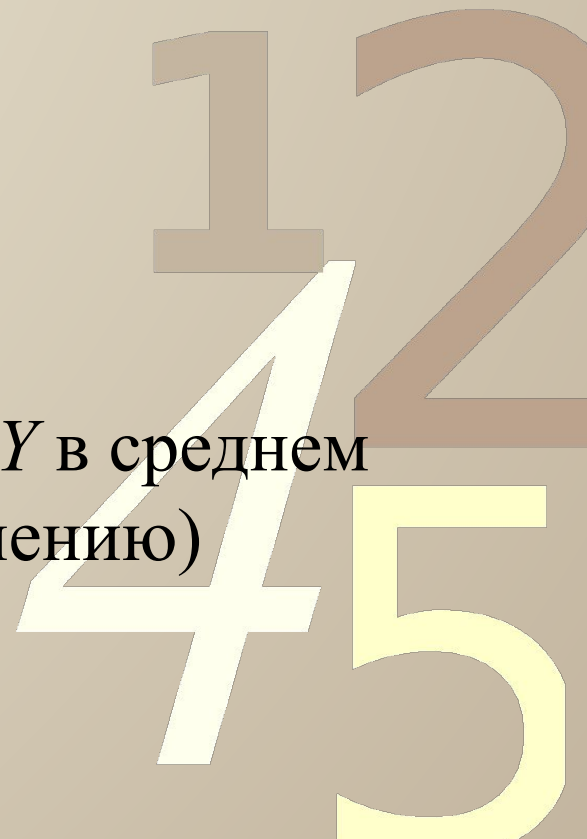
С ростом переменной X переменная Y в среднем растет (имеет тенденцию к росту)



Отрицательная корреляция



С ростом переменной X переменная Y в среднем убывает (имеет тенденцию к уменьшению)



Свойства коэффициента корреляции

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

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



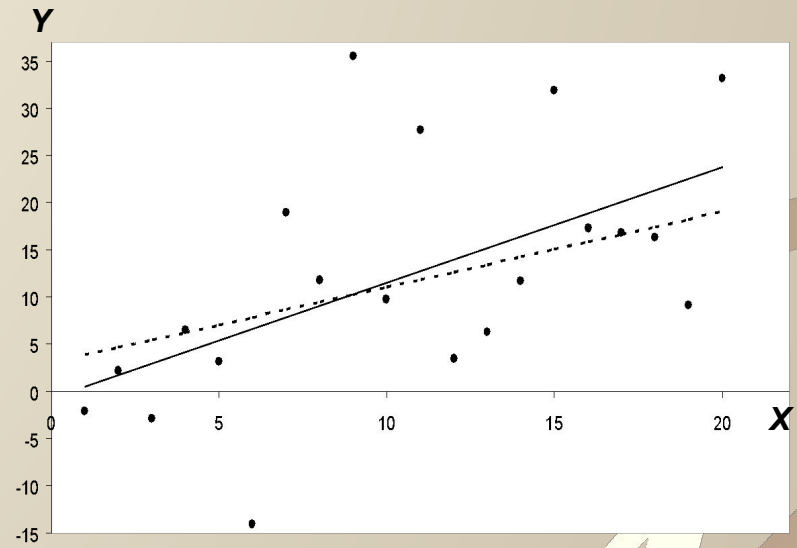
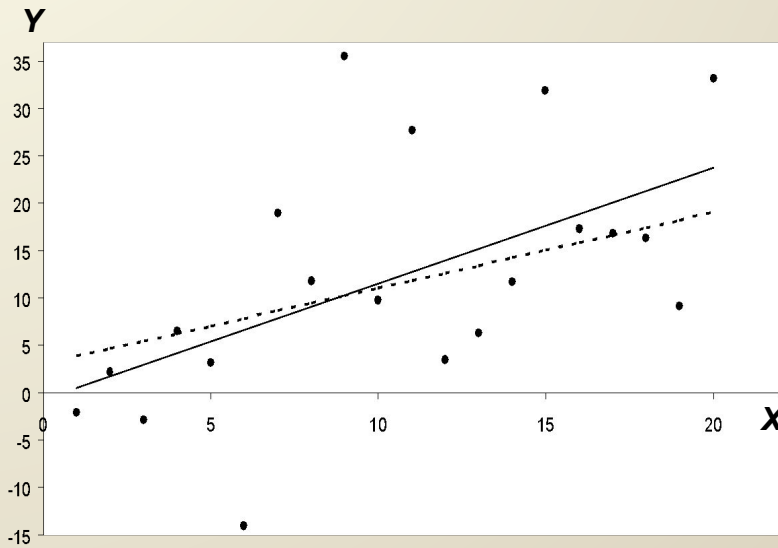
Свойства коэффициента корреляции (продолжение)

$r_{XY} = 0$ 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

0011 0010 1010 1101 0001 0100 1011



$$Y = 3.0 + 0.8X$$

45

Вопросы для самопроверки

0011 0010 1010 1101 0001 0100 1011

- 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.
- Формы записи МНК коэффициента наклона регрессионной 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.

