

Тема 4. Оценка качества подгонки линии регрессии к имеющимся данным

0011 0010 1010 1101 0001 0100 1011

12
45

Темы лекции.

0011 0010 1010 1101 0001 0100 1011

- Коэффициент детерминации.
- Свойства коэффициента детерминации.
- Скорректированный коэффициент детерминации.
- Свойства скорректированного коэффициента детерминации.



Как понять, насколько наша модель «хороша»

Первое, что приходит на ум – «похожесть» реальных и прогнозных значений, качество подгонки.

Второе – наша модель должна хорошо объяснять имеющиеся данные, мы должны понять, как образуются значения переменной Y .

Первое и второе не одно и тоже.

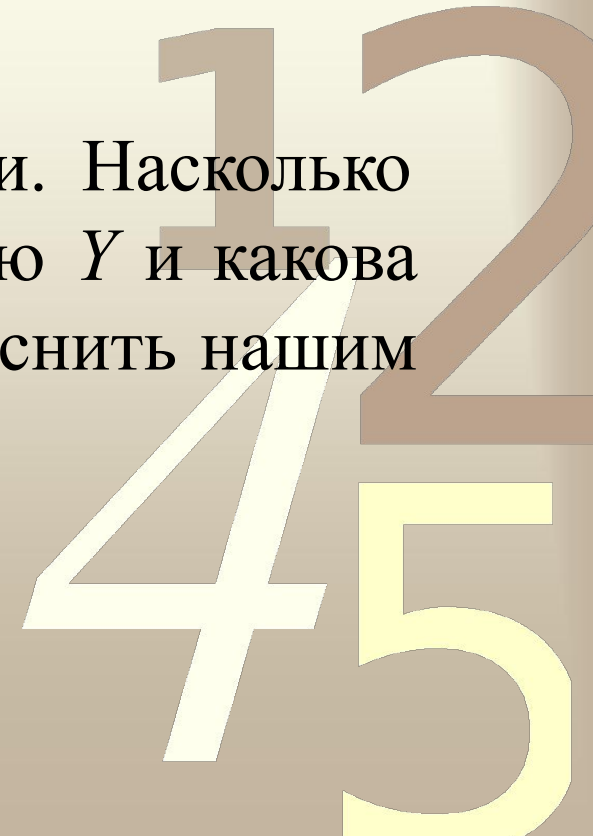
Модель может обладать хорошим качеством подгонки и совсем не обладать объясняющей способностью

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

0011 0010 1010 1101 0001 0100 1011

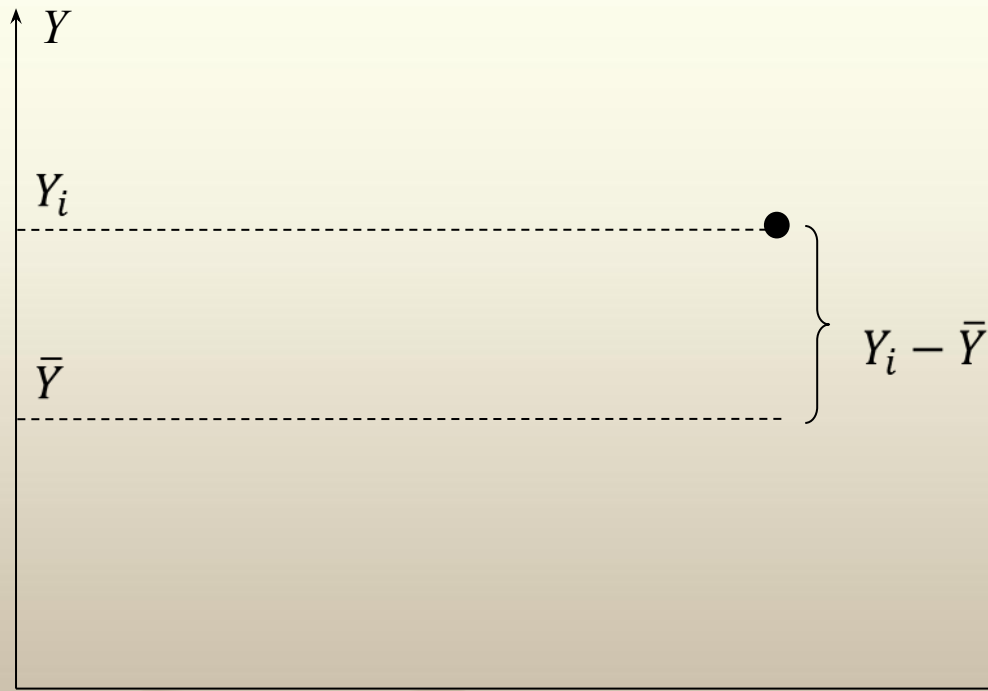
Насколько хорошо нам удалось объяснить изменение переменной Y нашей моделью.

Разложим вариацию Y на две части. Насколько наше уравнение объясняет вариацию Y и какова часть Y , которую мы не можем объяснить нашим уравнением.



Почему не все Y_i одинаковые?

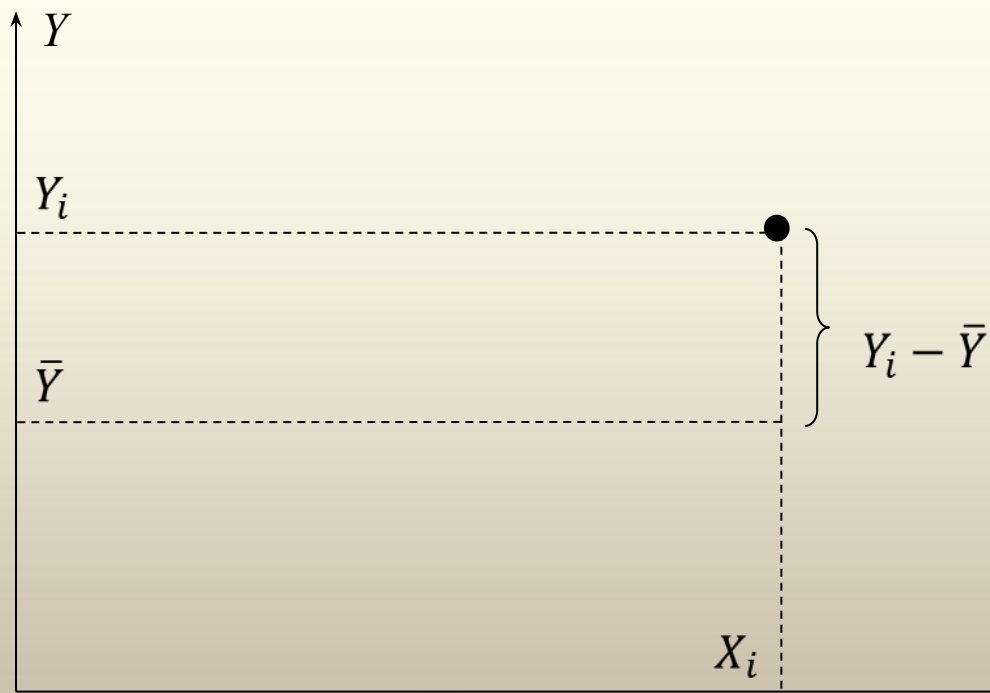
0011 0010 1010 1101 0001 0100 1011



12
45

Может Y зависеть от X ?

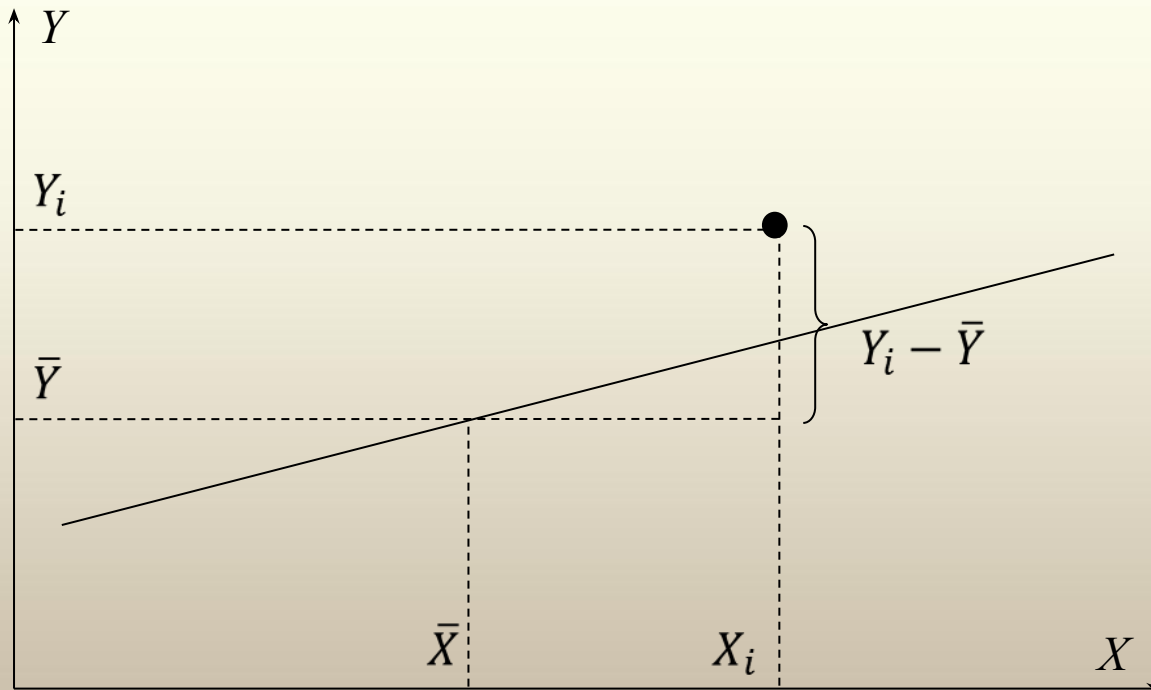
0011 0010 1010 1101 0001 0100 1011



12
45

И эта зависимость линейная?

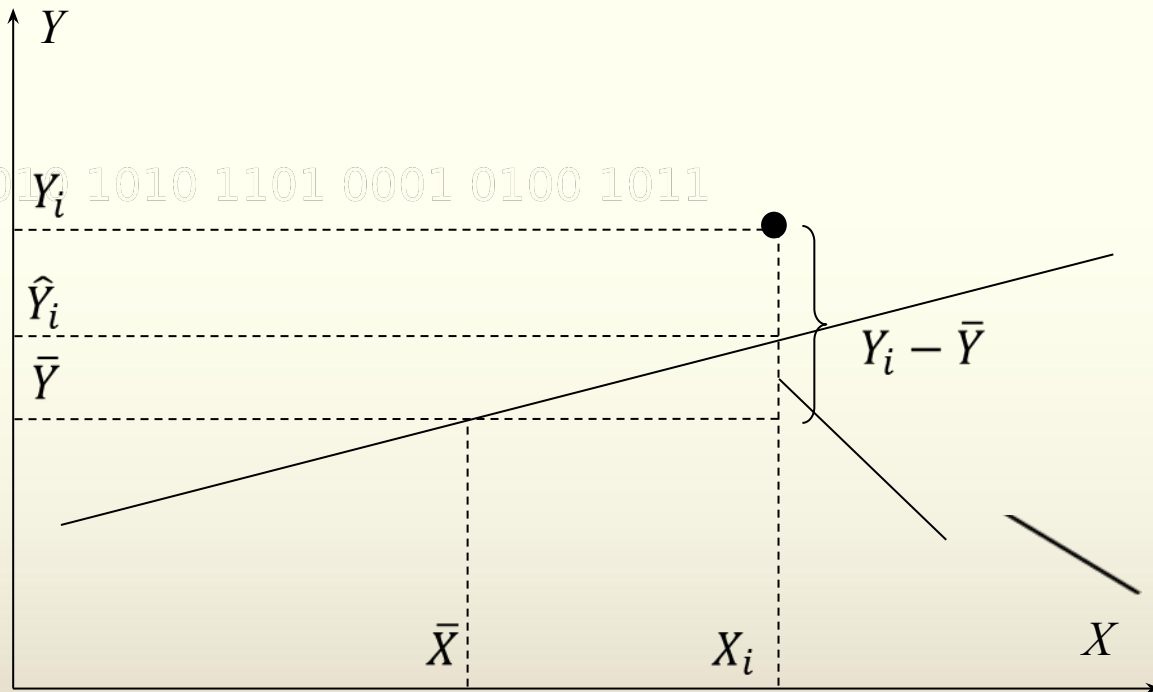
0011 0010 1010 1101 0001 0100 1011



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

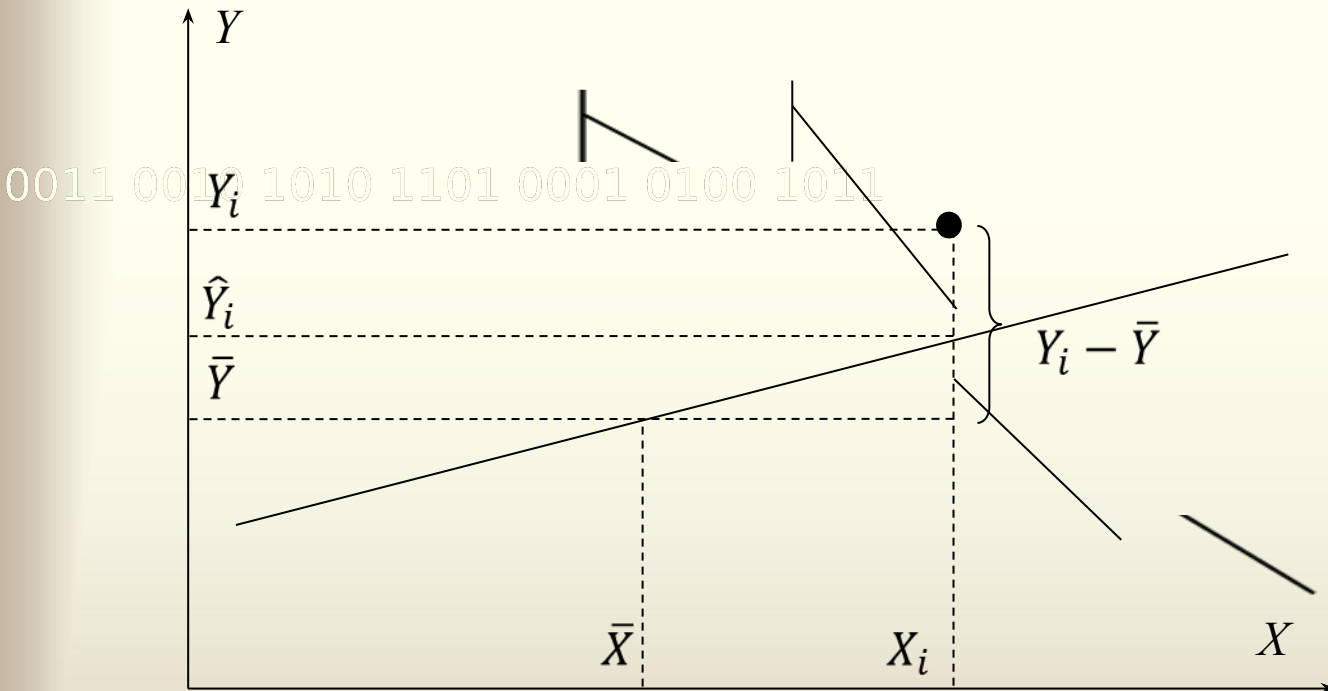
12
45

0011 0010 1010 1101 0001 0100 1011



Согласно модели, для данного значения переменной X переменная Y должна быть равна $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ и отклонение от среднего должно быть $\hat{Y}_i - \bar{Y}$

12
45



Согласно модели, для данного значения переменной X переменная Y должна быть равна $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$ и отклонение от среднего должно быть $\hat{Y}_i - \bar{Y}$

Модель знает не все.

Ошибка модели для данного наблюдения $e_i = Y_i - \hat{Y}_i$

12
45

Отклонение значения переменной Y от среднего

$$Y_i - \bar{Y}$$

0011 0010 1010 1101 0001 0100 1011

Раскладывается на две части

$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

объясненная рассматриваемой моделью +
необъяснённая часть (остаток)

чем модель лучше, тем остаток меньше.

12
45

$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

Возведем обе части в квадрат и просуммируем по всем наблюдениям

$$\sum_{i=1}^N (Y_i - \bar{Y})^2 = \sum_{i=1}^N ((\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i))^2$$

Раскроем скобки

$$\begin{aligned} \sum_{i=1}^N (Y_i - \bar{Y})^2 &= \sum_{i=1}^N (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 = \\ &= \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 - 2 \sum_{i=1}^N (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) + \sum_{i=1}^N (\hat{Y}_i - \bar{Y})^2 \end{aligned}$$

I II III

В этой сумме II = 0, если в уравнении есть свободный коэффициент



Разложение общей вариации переменной Y

$$\sum_{i=1}^N (Y_i - \bar{Y})^2 = \sum_{i=1}^N (\bar{Y} - \bar{Y})^2 + \sum_{i=1}^N (Y_i - \bar{Y})^2$$

TSS ESS RSS

$$TSS = ESS + RSS$$

$$\sum_{i=1}^N (Y_i - \bar{Y})^2 = \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^N (\hat{Y}_i - \bar{Y})^2$$

TSS

RSS

ESS

- TSS – total sum of squares – вся дисперсия или вариация Y , характеризует степень случайного разброса значений функции регрессии около среднего значения Y
- RSS – residual sum of squares – есть сумма квадратов остатков регрессии, та величина, которую мы минимизируем при построении прямой, часть дисперсии, которая нашим уравнением не объясняется
- ESS – equation sum of squares – объясненная часть общей вариации

Коэффициент детерминации

0011 0010 1010 1101 0001 0100 1011

Коэффициентом детерминации или долей объясненной нашим уравнением дисперсии называется величина

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$



Свойства коэффициента детерминации

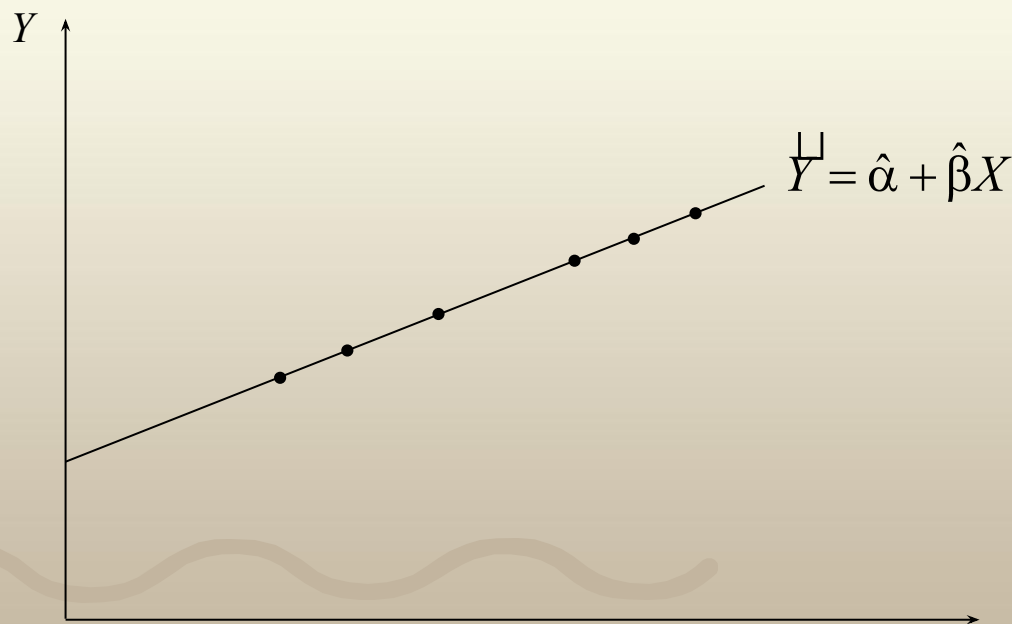
$$0 \leq R^2 \leq 1 \quad \text{в силу определения}$$



$$R^2 = 1$$

0011 0010 1010 1101 0001 0100 1011

в этом случае все точки (X_i, Y_i) лежат на одной прямой (RSS = 0).



Новые точки будут лежать на этой прямой?

$$R^2 = 0$$

0011 0010 1010 1101 0001 0100 1011

в этом случае $ESS = 0$,

$$\sum_{i=1}^N (Y_i - \bar{Y})^2 = 0$$

$$\forall i = 1 \dots N \quad Y_i - \bar{Y} = 0$$

$$Y_i = \bar{Y}$$

наша регрессия ничего не объясняет, ничего не дает по сравнению с тривиальным прогнозом

$$R^2 = 0$$

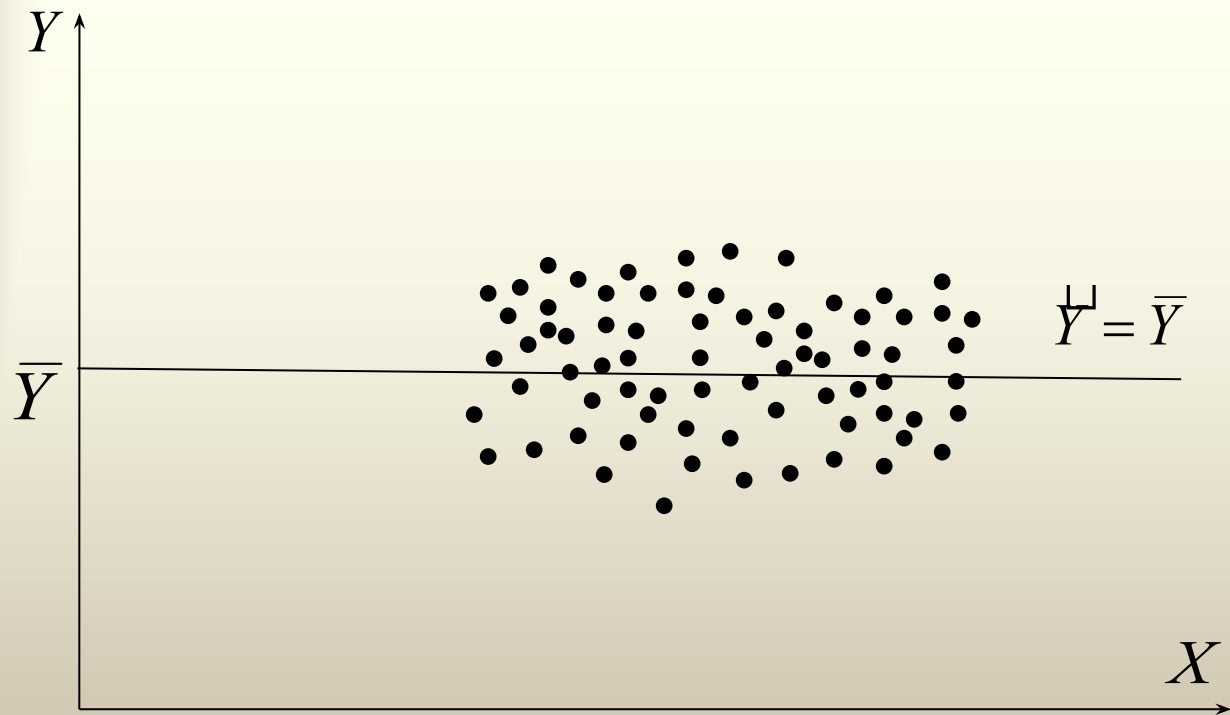
0011 0010 1010 1101 0001 0100 1011



12
45

Х и Y независимы

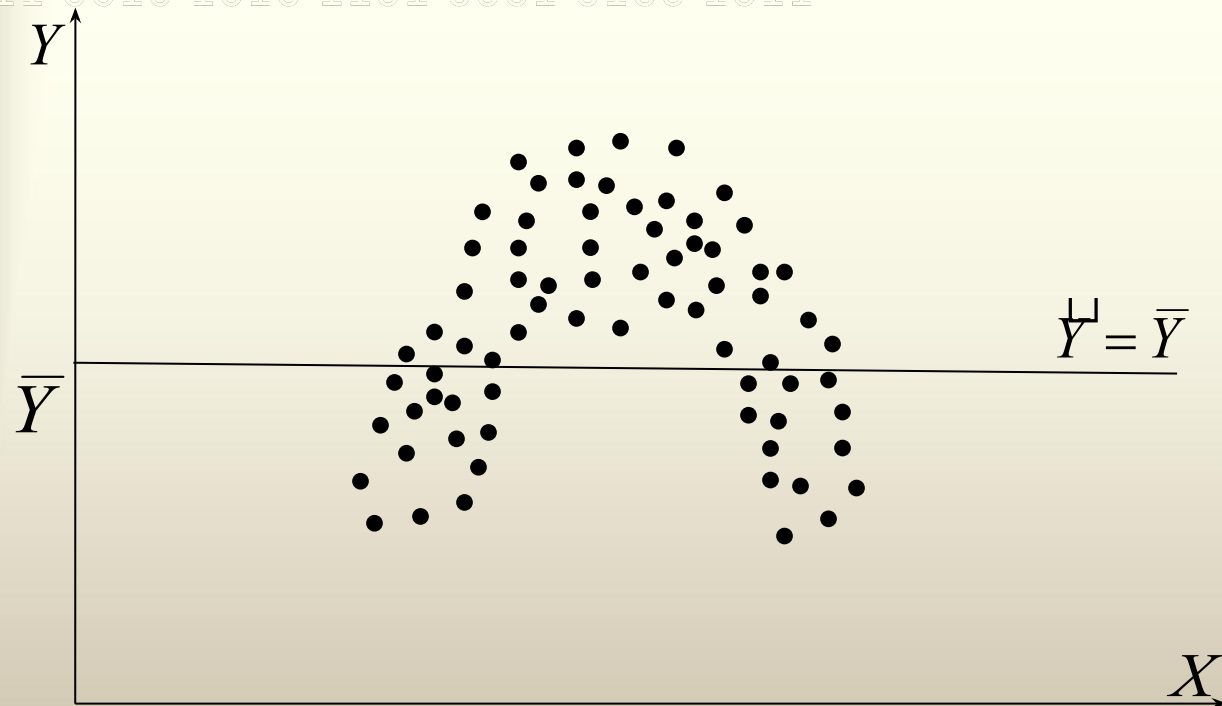
0011 0010 1010 1101 0001 0100 1011



12
45

Нелинейная корреляция

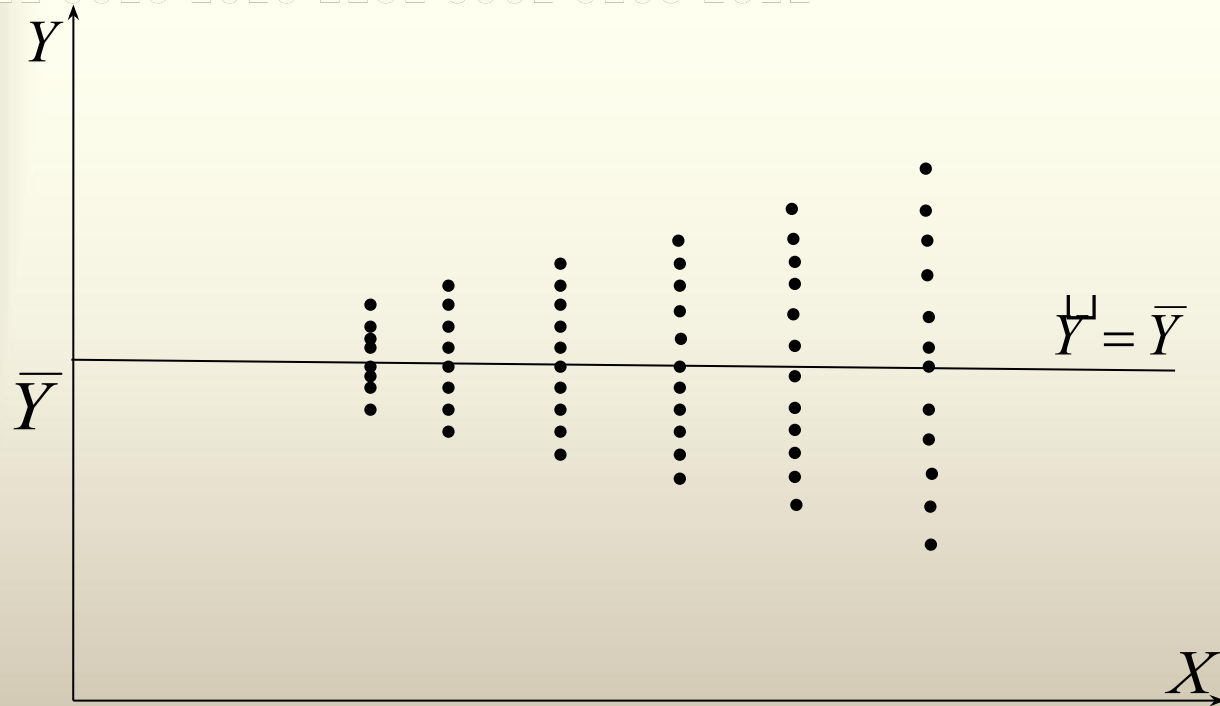
0011 0010 1010 1101 0001 0100 1011



12
45

Другая статистическая связь

0011 0010 1010 1101 0001 0100 1011



12
45

$$0 < R^2 < 1$$

0011 0010 1010 1101 0001 0100 1011

в этом случае чем ближе R^2 к 1, тем лучше
качество подгонки кривой к нашим
данным, тем точнее аппроксимирует Y

12
45

Недостаток коэффициента детерминации

0011 0010 1010 1101 0001 0100 1011

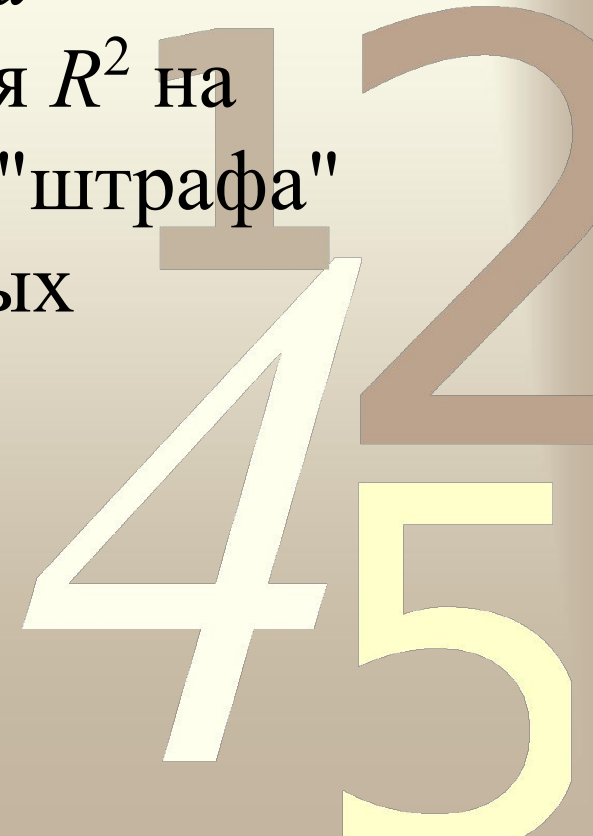
R^2 , вообще говоря, возрастает при добавлении еще одного регрессора, поэтому для выбора между несколькими регрессионными уравнениями не следует полагаться только на R^2

12
45

Скорректированный коэффициент детерминации

0011 0010 1010 1101 0001 0100 1011

Попыткой устранить эффект, связанный с ростом R^2 при увеличении числа регрессоров, является коррекция R^2 на число регрессоров - наложение "штрафа" за увеличение числа независимых переменных.



Скорректированный коэффициент детерминации

$$R_{adj}^2 = 1 - \frac{RSS / (N - k - 1)}{TSS / (N - 1)}$$

12
45

Свойства скорректированного коэффициента детерминации

$$R_{adj}^2 = 1 - (1 - R^2) \frac{N - 1}{N - k - 1}$$

$$R^2 > R_{adj}^2$$

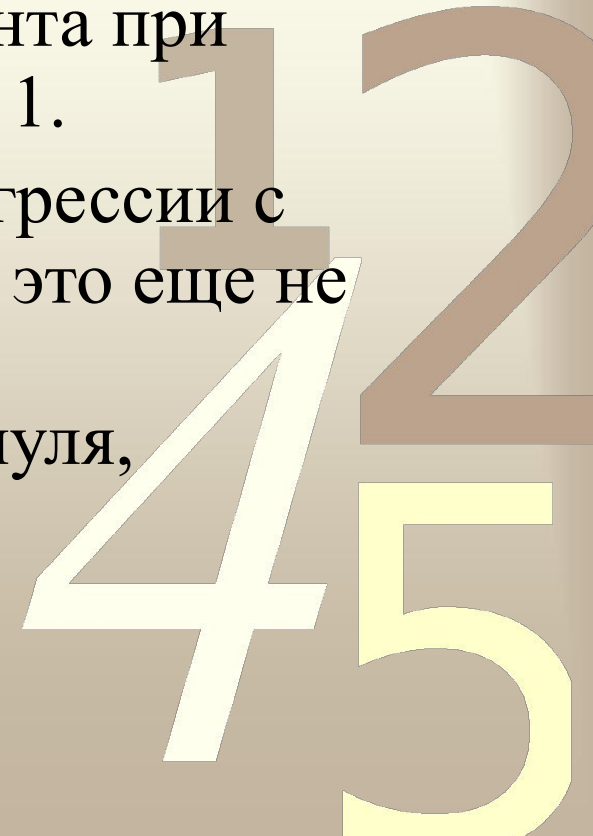
$$R_{adj}^2 \leq 1, \text{ но может быть и } < 0$$

Упражнение

0011 0010 1010 1101 0001 0100 1011

Показать, что статистика F увеличится при добавлении новой переменной тогда и только тогда, когда t -статистика коэффициента при этой переменной по модулю больше 1.

Следовательно, если в результате регрессии с новой переменной R^2_{adj} увеличился, это еще не означает, что коэффициент при этой переменной значимо отличается от нуля, поэтому мы не можем сказать, что спецификация модели улучшилась



Вопросы для самопроверки

0011 0010 1010 1101 0001 0100 1011

- Для чего нужен коэффициент детерминации.
- Основная идея построения характеристики качества подгонки линии регрессии к имеющимся данным.
- Как связаны между собой коэффициент детерминации и коэффициент корреляции в парной модели.
- В каком случае коэффициент детерминации имеет смысл.
- Докажите, что второе слагаемое в разложении общей вариации равно нулю.
- Какие вы знаете свойства коэффициента детерминации
- В каких случаях нельзя использовать коэффициент детерминации для сравнения моделей.
- Что такое скорректированный коэффициент детерминации.
- Всегда ли скорректированный коэффициент детерминации увеличивается при добавлении новых переменных.
- Перечислите свойства скорректированного коэффициента детерминации