Множественный корреляционный анализ

Пусть имеется набор переменных $X_1, X_2, \ldots, X_n$, тогда можно найти выборочные коэффициенты корреляции $r_{i,j}$ для каждой пары X_i, X_j.

Множественный корреляционный анализ

Пусть имеется набор переменных $X_1, X_2, \ldots, X_n$, тогда можно найти выборочные коэффициенты корреляции $r_{i,j}$ для каждой пары X_i, X_j.

Корреляционная матрица $\mathrm{R} = (r_{i,j})$.

$$\mathrm{R} = \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{pmatrix}.$$

Множественный корреляционный анализ

Пусть имеется набор переменных $X_1, X_2, \ldots, X_n$, тогда можно найти выборочные коэффициенты корреляции $r_{i,j}$ для каждой пары X_i, X_j.

Корреляционная матрица $\mathrm{R} = (r_{i,j})$.

$$\mathrm{R} = \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{pmatrix}.$$

Матрица является симметричной относительно главной диагонали $(r_{ij} = r_{ji})$, причем все диагональные элементы равны единице $(r_{ii} = 1)$.

Регрессионный анализ парной линейной модели

Регрессионный анализ парной линейной модели

Пусть Y - зависимая, X - объясняющая переменные.

Регрессионный анализ парной линейной модели

Пусть Y - зависимая, X - объясняющая переменные.

Ранее вводили функцию регрессии $y = f_r(x) \overset{\text{def}}{=} E\left(Y \mid X = x\right)$.

Регрессионный анализ парной линейной модели

Пусть Y - зависимая, X - объясняющая переменные.

Ранее вводили функцию регрессии $y = f_r(x) \overset{\text{def}}{=} E(Y \mid X = x)$.

Определение. Регрессионная модель – предполагаемый вид регрессионной зависимости, с точностью до неизвестных параметров:

$$y = f_r(x;\beta),$$

где $\beta = (\beta_0, \beta_1, ..., \beta_m)$ - вектор параметров.

Регрессионный анализ парной линейной модели

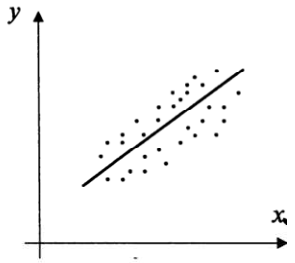
Пусть Y - зависимая, X - объясняющая переменные.

Ранее вводили функцию регрессии $y = f_r(x) \overset{\text{def}}{=} E(Y \mid X = x)$.

Определение. Регрессионная модель – предполагаемый вид регрессионной зависимости, с точностью до неизвестных параметров:

$$y = f_r(x;\beta),$$

где $\beta = (\beta_0, \beta_1, ..., \beta_m)$ - вектор параметров.



Например, линейная регрессионная модель:

$$y = \beta_0 + \beta_1 \cdot x.$$

Регрессионный анализ парной линейной модели

Пусть Y - зависимая, X - объясняющая переменные.

Ранее вводили функцию регрессии $y = f_r(x) \overset{\text{def}}{=} E(Y \mid X = x)$.

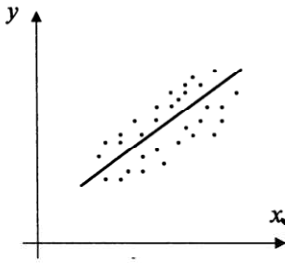
Определение. Регрессионная модель – предполагаемый вид регрессионной зависимости, с точностью до неизвестных параметров:

$$y = f_r(x;\beta),$$

где $\beta = (\beta_0, \beta_1, ..., \beta_m)$ - вектор параметров.

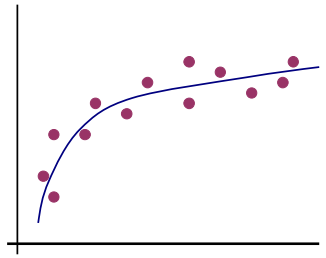
Например, линейная регрессионная модель:

$$y = \beta_0 + \beta_1 \cdot x.$$



Примеры нелинейных моделей:

логарифмическая модель: $y = \beta_0 + \beta_1 \cdot \ln x,$



Регрессионный анализ парной линейной модели

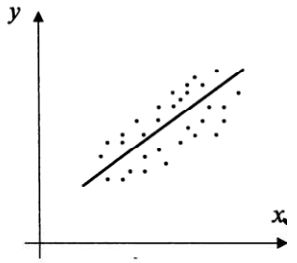
Пусть Y - зависимая, X - объясняющая переменные.

Ранее вводили функцию регрессии $y = f_r(x) \overset{\text{def}}{=} E(Y \mid X = x)$.

Определение. Регрессионная модель – предполагаемый вид регрессионной зависимости, с точностью до неизвестных параметров:

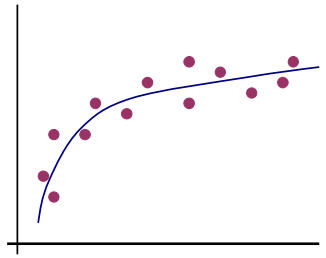
$$y = f_r(x;\beta),$$

где $\beta = (\beta_0, \beta_1, ..., \beta_m)$ - вектор параметров.



Например, линейная регрессионная модель:

$$y = \beta_0 + \beta_1 \cdot x.$$



Примеры нелинейных моделей:

логарифмическая модель: $y = \beta_0 + \beta_1 \cdot \ln x,$

степенная модель: $y = \beta_0 \, x^{\beta_1},$

Регрессионный анализ парной линейной модели

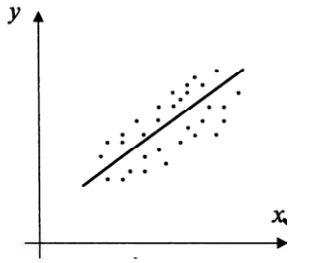
Пусть Y - зависимая, X - объясняющая переменные.

Ранее вводили функцию регрессии $y = f_r(x) \overset{\text{def}}{=} E(Y \mid X = x)$.

Определение. Регрессионная модель – предполагаемый вид регрессионной зависимости, с точностью до неизвестных параметров:

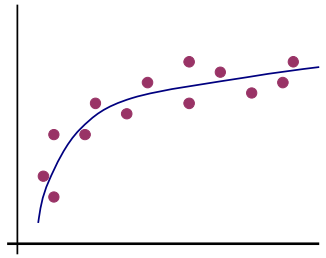
$$y = f_r(x;\beta),$$

где $\beta = (\beta_0, \beta_1, ..., \beta_m)$ - вектор параметров.



Например, линейная регрессионная модель:

$$y = \beta_0 + \beta_1 \cdot x.$$



Примеры нелинейных моделей:

логарифмическая модель: $y = \beta_0 + \beta_1 \cdot \ln x$,

степенная модель: $y = \beta_0 \, x^{\beta_1}$,

экспоненциальная модель $y = e^{\beta_0 + \beta_1 x}$.

Замечание 1. Некоторые нелинейные модели можно преобразовать к линейному виду (заменой переменных и т.д.).

Замечание 1. Некоторые нелинейные модели можно преобразовать к линейному виду (заменой переменных и т.д.).

Замечание 2. Модель зависимости можно представить в виде:

$$Y_i = f_r(x_i;\beta) + \varepsilon_i \ ,$$

где ε_i - случайная ошибка, $i = 1,...,n$

(сумма неслучайной и случайной компоненты).

Замечание 1. Некоторые нелинейные модели можно преобразовать к линейному виду (заменой переменных и т.д.).

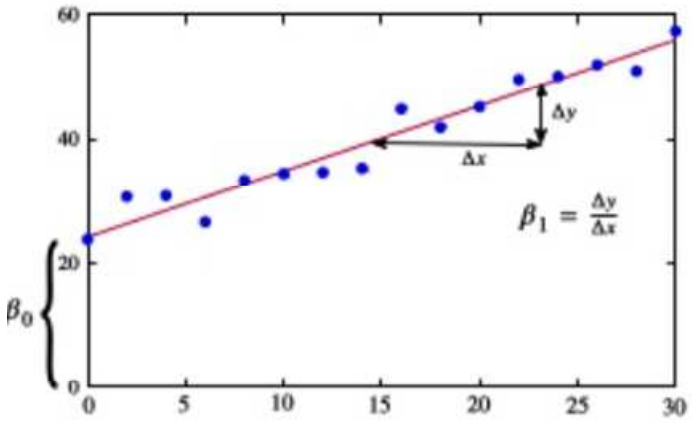
Замечание 2. Модель зависимости можно представить в виде:

$$Y_i = f_r(x_i;\beta) + \varepsilon_i \ ,$$

где ε_i - случайная ошибка, $i = 1,...,n$

(сумма неслучайной и случайной компоненты).

$$Y^{(i)} = \beta_0 + \beta_1 x^{(i)} + \varepsilon^{(i)}$$



простейшая (одномерная) линейная регрессия

β_0 коэффициент пересечения (с осью X=0)

β_1 коэффициент наклона

Оценивание параметров модели

Принцип наименьших квадратов: параметры модели подбираются так, чтобы сумма квадратов ошибок (отклонений наблюдаемых и теоретических значений Y) была минимальной.

Оценивание параметров модели

Принцип наименьших квадратов: параметры модели подбираются так, чтобы сумма квадратов ошибок (отклонений наблюдаемых и теоретических значений Y) была минимальной.

Пусть $S(\beta) = \sum_{i=1}^{n} \left(y_i - f\left(x_i; \beta\right)\right)^2$, тогда значение $b = (b_0, b_1)$ такое,

$$\text{что } S\left(b\right) = \min_{\beta \in \mathbf{R}^2} S(\beta),$$

называется оценкой параметров методом наименьших квадратов (МНК-оценкой).

Оценивание параметров модели

Принцип наименьших квадратов: параметры модели подбираются так, чтобы сумма квадратов ошибок (отклонений наблюдаемых и теоретических значений Y) была минимальной.

Пусть $S(\beta) = \sum_{i=1}^{n} \left(y_i - f\left(x_i; \beta\right)\right)^2$, тогда значение $b = (b_0, b_1)$ такое,

$$\text{что } S\left(b\right) = \min_{\beta \in \mathbf{R}^2} S(\beta),$$

называется оценкой параметров методом наименьших квадратов (МНК-оценкой).

Замечание. В общем случае S может быть многоэкстремальной (требуется найти глобальный минимум).

МНК-оценка параметров линейной модели

Рассмотрим сумму квадратов ошибок:

$$S(\beta_0, \beta_1) = \sum_i (y_i - (\beta_0 + \beta_1 x_i))^2 .$$

МНК-оценка параметров линейной модели

Рассмотрим сумму квадратов ошибок:

$$S(\beta_0, \beta_1) = \sum_i (y_i - (\beta_0 + \beta_1 x_i))^2 .$$

Необходимое условие $\min$ - равенство нулю частных производных:

$$\begin{cases} S'_{\beta_0} = -2\sum_i (y_i - \beta_0 - \beta_1 x_i) = 0 \\ S'_{\beta_1} = -2\sum_i (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \end{cases}$$

МНК-оценка параметров линейной модели

Рассмотрим сумму квадратов ошибок:

$$S(\beta_0, \beta_1) = \sum_i (y_i - (\beta_0 + \beta_1 x_i))^2 .$$

Необходимое условие $\min$ - равенство нулю частных производных:

$$\begin{cases} S'_{\beta_0} = -2\sum_i (y_i - \beta_0 - \beta_1 x_i) = 0 \\ S'_{\beta_1} = -2\sum_i (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \end{cases}$$

$$\begin{cases} \sum_i y_i - n\beta_0 - \beta_1 \sum_i x_i = 0 \\ \sum_i y_i x_i - \beta_0 \sum_i x_i - \beta_1 \sum (x_i)^2 = 0 \end{cases}$$

$\overline{y}, \overline{x}$ - средние значения x и y $\Rightarrow$

$$\begin{cases} n\,\overline{y} - n\,\beta_0 - \beta_1\, n\,\overline{x} = 0 \\ \sum_i y_i x_i - \beta_0 n\,\overline{x} - \beta_1 \sum_i (x_i)^2 = 0 \end{cases}$$

$\overline{y}, \overline{x}$ - средние значения x и $y \Rightarrow$

$$\begin{cases} n\,\overline{y} - n\,\beta_0 - \beta_1\, n\,\overline{x} = 0 \\ \sum_i y_i x_i - \beta_0 n\,\overline{x} - \beta_1 \sum_i (x_i)^2 = 0 \end{cases}$$

$$\beta_0 = \boxed{\overline{y} - \beta_1\,\overline{x} = b_0}$$

$\overline{y}, \overline{x}$ - средние значения x и $y \Rightarrow$

$$\begin{cases} n\,\overline{y} - n\,\beta_0 - \beta_1\, n\,\overline{x} = 0 \\ \sum_i y_i x_i - \beta_0 n\,\overline{x} - \beta_1 \sum_i (x_i)^2 = 0 \end{cases}$$

$$\beta_0 = \boxed{\overline{y} - \beta_1\,\overline{x} = b_0}$$

$$\sum_i y_i x_i - (\overline{y} - \beta_1 \overline{x})\, n\,\overline{x} - \beta_1 \sum_i (x_i)^2 = 0$$

$\overline{y}, \overline{x}$ - средние значения x и $y \Rightarrow$

$$\begin{cases} n\,\overline{y} - n\,\beta_0 - \beta_1\, n\,\overline{x} = 0 \\ \sum_i y_i x_i - \beta_0 n\,\overline{x} - \beta_1 \sum_i (x_i)^2 = 0 \end{cases}$$

$$\beta_0 = \boxed{\overline{y} - \beta_1\,\overline{x} = b_0}$$

$$\sum_i y_i x_i - (\overline{y} - \beta_1 \overline{x})\, n\,\overline{x} - \beta_1 \sum_i (x_i)^2 = 0$$

$$\sum_i y_i\, x_i - \overline{y}\, n\,\overline{x} + \beta_1\, n\,(\overline{x})^2 - \beta_1 \sum_i (x_i)^2 = 0$$

$\overline{y}, \overline{x}$ - средние значения x и $y \Rightarrow$

$$\begin{cases} n\,\overline{y} - n\,\beta_0 - \beta_1\, n\,\overline{x} = 0 \\ \sum_i y_i x_i - \beta_0 n\,\overline{x} - \beta_1 \sum_i (x_i)^2 = 0 \end{cases}$$

$$\beta_0 = \boxed{\overline{y} - \beta_1\,\overline{x} = b_0}$$

$$\sum_i y_i x_i - (\overline{y} - \beta_1 \overline{x})\, n\,\overline{x} - \beta_1 \sum_i (x_i)^2 = 0$$

$$\sum_i y_i\, x_i - \overline{y}\, n\,\overline{x} + \beta_1\, n\,(\overline{x})^2 - \beta_1 \sum_i (x_i)^2 = 0$$

$$\sum_i y_i\, x_i - n\,\overline{y}\,\overline{x} = \beta_1 \sum_i (x_i)^2 - \beta_1\, n\,(\overline{x})^2$$

$\overline{y}, \overline{x}$ - средние значения x и $y \Rightarrow$

$$\begin{cases} n\,\overline{y} - n\,\beta_0 - \beta_1\, n\,\overline{x} = 0 \\ \sum_i y_i x_i - \beta_0 n\,\overline{x} - \beta_1 \sum_i (x_i)^2 = 0 \end{cases}$$

$$\beta_0 = \boxed{\overline{y} - \beta_1\,\overline{x} = b_0}$$

$$\sum_i y_i x_i - (\overline{y} - \beta_1 \overline{x})\, n\,\overline{x} - \beta_1 \sum_i (x_i)^2 = 0$$

$$\sum_i y_i\, x_i - \overline{y}\, n\,\overline{x} + \beta_1\, n\,(\overline{x})^2 - \beta_1 \sum_i (x_i)^2 = 0$$

$$\sum_i y_i\, x_i - n\,\overline{y}\,\overline{x} = \beta_1 \sum_i (x_i)^2 - \beta_1\, n\,(\overline{x})^2$$

$$\beta_1 = \boxed{\frac{\sum_i y_i x_i - n\,\overline{y}\,\overline{x}}{\sum_i (x_i)^2 - n\,(\overline{x})^2} = b_1}$$

$$\boxed{b_1 = \frac{s_{xy}}{s_x^2}},$$

где $s_{xy} = \overline{xy} - \overline{x} \cdot \overline{y}$, $s_x^2 = \overline{x^2} - (\overline{x})^2$.

$$\boxed{b_1 = \frac{s_{xy}}{s_x^2}},$$

где $s_{xy} = \overline{xy} - \overline{x} \cdot \overline{y}$, $s_x^2 = \overline{x^2} - (\overline{x})^2$.

Другая форма записи: $b_1 = \dfrac{\sum_i (y_i - \overline{y})(x_i - \overline{x})}{\sum_i (x_i - \overline{x})^2}$.

$$\boxed{b_1 = \frac{s_{xy}}{s_x^2}},$$

где $s_{xy} = \overline{xy} - \overline{x} \cdot \overline{y}$, $s_x^2 = \overline{x^2} - (\overline{x})^2$.

Другая форма записи: $b_1 = \dfrac{\sum_i (y_i - \overline{y})(x_i - \overline{x})}{\sum_i (x_i - \overline{x})^2}$.

Можно показать, что точка минимума (b_0, b_1) единственна.

$$\boxed{b_1 = \frac{s_{xy}}{s_x^2}},$$

где $s_{xy} = \overline{xy} - \overline{x} \cdot \overline{y}$, $s_x^2 = \overline{x^2} - (\overline{x})^2$.

Другая форма записи: $b_1 = \dfrac{\sum_i (y_i - \overline{y})(x_i - \overline{x})}{\sum_i (x_i - \overline{x})^2}$.

Можно показать, что точка минимума (b_0, b_1) единственна.

Качество модели определяется несмещенной оценкой σ^2 (дисперсии ошибки:

$$\boxed{s^2 = \frac{1}{n-2} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}$$

$$\boxed{b_1 = \frac{s_{xy}}{s_x^2}},$$

где $s_{xy} = \overline{xy} - \overline{x} \cdot \overline{y}$, $s_x^2 = \overline{x^2} - (\overline{x})^2$.

Другая форма записи: $b_1 = \dfrac{\sum_i (y_i - \overline{y})(x_i - \overline{x})}{\sum_i (x_i - \overline{x})^2}$.

Можно показать, что точка минимума (b_0, b_1) единственна.

Качество модели определяется несмещенной оценкой σ^2 (дисперсии ошибки:

$$\boxed{s^2 = \frac{1}{n-2} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2}$$

где $\hat{y}_i = b_0 + b_1 x_i$ - прогноз для x_i,

$$\boxed{b_1 = \frac{s_{xy}}{s_x^2}},$$

где $s_{xy} = \overline{xy} - \overline{x} \cdot \overline{y}$, $s_x^2 = \overline{x^2} - (\overline{x})^2$.

Другая форма записи: $b_1 = \dfrac{\sum_i (y_i - \overline{y})(x_i - \overline{x})}{\sum_i (x_i - \overline{x})^2}$.

Можно показать, что точка минимума (b_0, b_1) единственна.

Качество модели определяется несмещенной оценкой σ^2 (дисперсии ошибки:

$$\boxed{s^2 = \frac{1}{n-2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

где $\hat{y}_i = b_0 + b_1 x_i$ - прогноз для x_i, $e_i = y_i - \hat{y}_i$ - остаток,

$$\boxed{b_1 = \frac{s_{xy}}{s_x^2}},$$

где $s_{xy} = \overline{xy} - \overline{x} \cdot \overline{y}$, $s_x^2 = \overline{x^2} - (\overline{x})^2$.

Другая форма записи: $b_1 = \dfrac{\sum_i (y_i - \overline{y})(x_i - \overline{x})}{\sum_i (x_i - \overline{x})^2}$.

Можно показать, что точка минимума (b_0, b_1) единственна.

Качество модели определяется несмещенной оценкой σ^2 (дисперсии ошибки:

$$\boxed{s^2 = \frac{1}{n-2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

где $\hat{y}_i = b_0 + b_1 x_i$ - прогноз для x_i, $e_i = y_i - \hat{y}_i$ - остаток,

$s = \sqrt{s^2}$ - средняя ошибка регрессии.

Пример (продолжение). По данным примера предыдущей лекции найти МНК-оценки параметров и среднюю ошибку регрессии.

x	2	3	5	2	6
y	1	4	3	2	5

Пример (продолжение). По данным примера предыдущей лекции найти МНК-оценки параметров и среднюю ошибку регрессии.

x	2	3	5	2	6
y	1	4	3	2	5

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{1.8}{2.6} = 0.69,$$

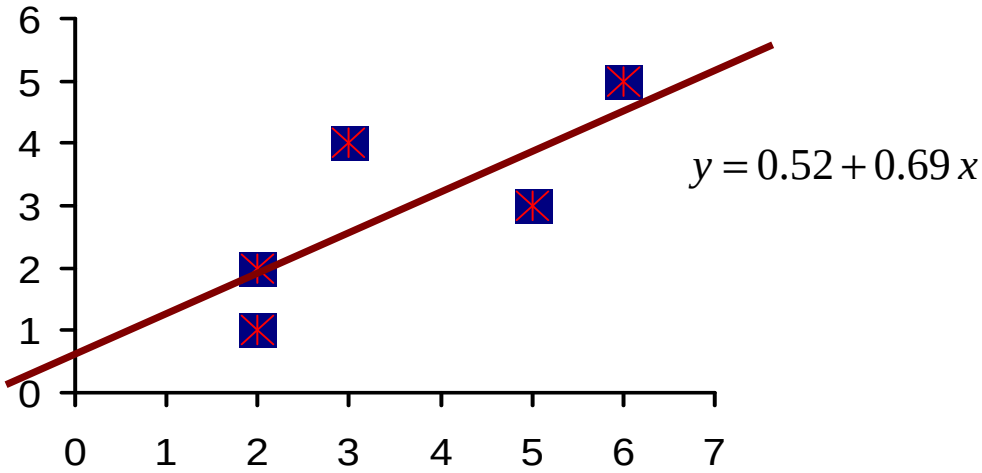
$$b_0 = \overline{y} - b_1 \overline{x} = 3 - 0.69 \cdot 3.6 = 0.52$$

Пример (продолжение). По данным примера предыдущей лекции найти МНК-оценки параметров и среднюю ошибку регрессии.

x	2	3	5	2	6
y	1	4	3	2	5

$$b_1 = \frac{s_{xy}}{s_x^2} = \frac{1.8}{2.6} = 0.69,$$

$$b_0 = \bar{y} - b_1 \bar{x} = 3 - 0.69 \cdot 3.6 = 0.52$$



Вычислим прогноз для каждого наблюдения и остатки:

x	2	3	5	2	6
y	1	4	3	2	5
$\hat{y}$	1.9	2.6	3.95	1.9	4.63
$y-\hat{y}$	-0.9	1.4	-0.95	0.1	0.37
$(y-\hat{y})^2$	0.81	1.96	0.9	0.01	0.137

Вычислим прогноз для каждого наблюдения и остатки:

x	2	3	5	2	6
y	1	4	3	2	5
$\hat{y}$	1.9	2.6	3.95	1.9	4.63
$y-\hat{y}$	-0.9	1.4	-0.95	0.1	0.37
$(y-\hat{y})^2$	0.81	1.96	0.9	0.01	0.137

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \frac{1}{3}\cdot 3.86 = 1.286,$$

Вычислим прогноз для каждого наблюдения и остатки:

x	2	3	5	2	6
y	1	4	3	2	5
$\hat{y}$	1.9	2.6	3.95	1.9	4.63
$y-\hat{y}$	-0.9	1.4	-0.95	0.1	0.37
$(y-\hat{y})^2$	0.81	1.96	0.9	0.01	0.137

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \frac{1}{3}\cdot 3.86 = 1.286,$$

$$s = 1.13.$$

Классическая модель множественной линейной регрессии

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_m x_{i,m} + \varepsilon_i,$$

Классическая модель множественной линейной регрессии

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_m x_{i,m} + \varepsilon_i,$$

где Y_i - случайное значение Y, соответствующее i-му наблюдению,

Классическая модель множественной линейной регрессии

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_m x_{i,m} + \varepsilon_i,$$

где Y_i - случайное значение Y, соответствующее i-му наблюдению, $x_{i,j}$ - i-е значение неслучайной переменной X_j; $j = 1, \ldots, m$;

Классическая модель множественной линейной регрессии

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_m x_{i,m} + \varepsilon_i,$$

где Y_i - случайное значение Y, соответствующее i-му наблюдению, $x_{i,j}$ - i-е значение неслучайной переменной X_j; $j = 1, \ldots, m$; ε_i - случайная ошибка, $i = 1, \ldots, n$; $\beta_0, \ldots, \beta_m$ — параметры модели.

Классическая модель множественной линейной регрессии

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_m x_{i,m} + \varepsilon_i,$$

где Y_i - случайное значение Y, соответствующее i-му наблюдению, $x_{i,j}$ - i-е значение неслучайной переменной X_j; $j = 1, \ldots, m$; ε_i - случайная ошибка, $i = 1, \ldots, n$; $\beta_0, \ldots, \beta_m$ — параметры модели.

Параметр β_j - ожидаемое изменение Y при изменении X_j на одну единицу измерения (при прочих неизменных значениях переменных).

Классическая модель множественной линейной регрессии

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_m x_{i,m} + \varepsilon_i,$$

где Y_i - случайное значение Y, соответствующее i-му наблюдению, $x_{i,j}$ - i-е значение неслучайной переменной X_j; $j = 1, \ldots, m$; ε_i - случайная ошибка, $i = 1, \ldots, n$; $\beta_0, \ldots, \beta_m$ — параметры модели.

Параметр β_j - ожидаемое изменение Y при изменении X_j на одну единицу измерения (при прочих неизменных значениях переменных).

Предполагается, что случайная ошибка имеет **нулевое математическое ожидание и постоянную дисперсию** σ^2; ε_i, ε_j **независимы**.

Пусть $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \hline \beta_m \end{pmatrix}$ - вектор параметров,

Пусть $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \hline \beta_m \end{pmatrix}$ - вектор параметров, $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \hline y_n \end{pmatrix}$ - вектор

наблюдений зависимой переменной Y,

Пусть $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \hline \beta_m \end{pmatrix}$ - вектор параметров, $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \hline y_n \end{pmatrix}$ - вектор

наблюдений зависимой переменной Y,

$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,m} \\ 1 & x_{2,1} & \ldots & x_{2,m} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n,1} & \ldots & x_{n,m} \end{pmatrix}$ – «расширенная» матрица наблюдений

объясняющих переменных,

Пусть $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \hline \beta_m \end{pmatrix}$ - вектор параметров, $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \hline y_n \end{pmatrix}$ - вектор

наблюдений зависимой переменной Y,

$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,m} \\ 1 & x_{2,1} & \ldots & x_{2,m} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n,1} & \ldots & x_{n,m} \end{pmatrix}$ – «расширенная» матрица наблюдений

объясняющих переменных, $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \hline \varepsilon_n \end{pmatrix}$ — вектор ошибок.

Пусть $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \hline \beta_m \end{pmatrix}$ - вектор параметров, $\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \hline y_n \end{pmatrix}$ - вектор

наблюдений зависимой переменной Y,

$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,m} \\ 1 & x_{2,1} & \ldots & x_{2,m} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n,1} & \ldots & x_{n,m} \end{pmatrix}$ – «расширенная» матрица наблюдений

объясняющих переменных, $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \hline \varepsilon_n \end{pmatrix}$ — вектор ошибок.

Модель в матричном виде: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$.

Нахождение оценок параметров: метод наименьших квадратов:

$$S(\beta_0, \beta_1, \ldots, \beta_m) = \sum_i \left(y_i - \sum_{j=0}^{m} \beta_j x_{i,j} \right)^2 \to \min_{\beta_0, \beta_1, \ldots, \beta_m}.$$

Нахождение оценок параметров: метод наименьших квадратов:

$$S(\beta_0, \beta_1, \ldots, \beta_m) = \sum_i \left(y_i - \sum_{j=0}^{m} \beta_j x_{i,j} \right)^2 \to \min_{\beta_0, \beta_1, \ldots, \beta_m}.$$

$$\frac{\partial S}{\partial \beta_l} = \sum_i (y_i - \sum_{j=0}^{n} \beta_j x_{i,j}) x_{i,l} = 0, \qquad l = 0,1,\ldots,m$$

Нахождение оценок параметров: метод наименьших квадратов:

$$S(\beta_0, \beta_1, ..., \beta_m) = \sum_i \left(y_i - \sum_{j=0}^{m} \beta_j x_{i,j} \right)^2 \to \min_{\beta_0, \beta_1, ..., \beta_m}.$$

$$\frac{\partial S}{\partial \beta_l} = \sum_i (y_i - \sum_{j=0}^{n} \beta_j x_{i,j}) x_{i,l} = 0, \qquad l = 0,1,...,m$$

$$\sum_i y_i x_{i,l} - \sum_i \sum_{j=0}^{m} \beta_j x_{i,j} x_{i,l} = 0, \quad l = 0,1,...,m$$

Нахождение оценок параметров: метод наименьших квадратов:

$$S(\beta_0, \beta_1, ..., \beta_m) = \sum_i \left(y_i - \sum_{j=0}^{m} \beta_j x_{i,j} \right)^2 \to \min_{\beta_0, \beta_1, ..., \beta_m}.$$

$$\frac{\partial S}{\partial \beta_l} = \sum_i (y_i - \sum_{j=0}^{n} \beta_j x_{i,j}) x_{i,l} = 0, \qquad l = 0,1,...,m$$

$$\sum_i y_i x_{i,l} - \sum_i \sum_{j=0}^{m} \beta_j x_{i,j} x_{i,l} = 0, \quad l = 0,1,...,m$$

$$\sum_{j=0}^{m} \beta_j \sum_i x_{i,j} x_{i,l} = \sum_i y_i x_{i,l}, \quad l = 0,1,...,m$$

- система нормальных уравнений.

В матричном виде:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\boldsymbol{\beta} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

$\mathbf{X}^{\mathrm{T}}$ - транспонированная матрица.

В матричном виде:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\boldsymbol{\beta} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

$\mathbf{X}^{\mathrm{T}}$ - транспонированная матрица.

Предположим, что существует обратная матрица $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ (ранг(X)=m+1).

В матричном виде:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\boldsymbol{\beta} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

$\mathbf{X}^{\mathrm{T}}$ - транспонированная матрица.

Предположим, что существует обратная матрица $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ (ранг(X)=m+1). Тогда

$$\boldsymbol{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{b},$$

В матричном виде:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\boldsymbol{\beta} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

$\mathbf{X}^{\mathrm{T}}$ - транспонированная матрица.

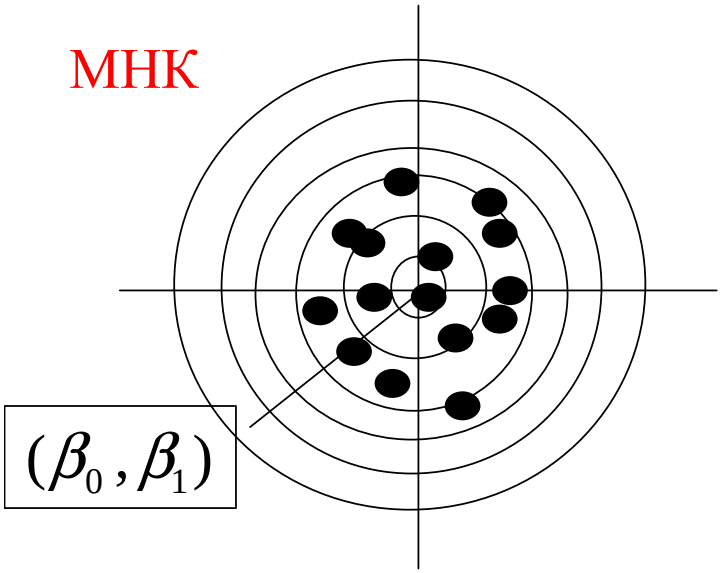
Предположим, что существует обратная матрица $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ (ранг(X)=m+1). Тогда

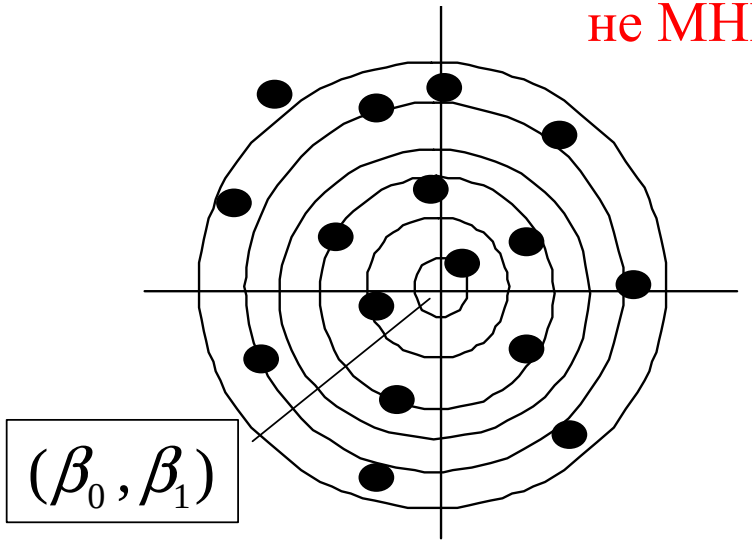
$$\boldsymbol{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{b},$$

где $\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \cdots \\ b_m \end{pmatrix}$ — вектор оценок параметров.

Теорема Гаусса-Маркова. Пусть выполняются условия классической модели. Тогда МНК-оценки являются наиболее эффективными, т.е. обладают наименьшей дисперсией среди всех линейных несмещенных оценок.

Теорема Гаусса-Маркова. Пусть выполняются условия классической модели. Тогда МНК-оценки являются наиболее эффективными, т.е. обладают наименьшей дисперсией среди всех линейных несмещенных оценок.





Анализ качества модели:

- определение степени соответствия модели и наблюдений (дисперсионный анализ);

Анализ качества модели:

- определение степени соответствия модели и наблюдений (дисперсионный анализ);

- проверка гипотез о значимости оценок параметров и модели в целом.

Цель **дисперсионного анализа** регрессионной модели - проверить, насколько вариация (изменчивость) зависимой переменной объясняется включенными в модель факторами.

Цель **дисперсионного анализа** регрессионной модели - проверить, насколько вариация (изменчивость) зависимой переменной объясняется включенными в модель факторами.

Пусть

$$e_i = y_i - \hat{y}_i$$

i-й остаток, где

$$\hat{y}_i = b_0 + b_1 x_{i,1} + \ldots + b_m x_{i,m}$$

прогноз для i-го наблюдения.

Цель **дисперсионного анализа** регрессионной модели - проверить, насколько вариация (изменчивость) зависимой переменной объясняется включенными в модель факторами.

Пусть

$$e_i = y_i - \hat{y}_i$$

i-й остаток, где

$$\hat{y}_i = b_0 + b_1 x_{i,1} + \ldots + b_m x_{i,m}$$

прогноз для i-го наблюдения.

Остаточная вариация (residual sum of squares)

$$RSS = \sum_{i=1}^{n} (e_i)^2;$$

Цель **дисперсионного анализа** регрессионной модели - проверить, насколько вариация (изменчивость) зависимой переменной объясняется включенными в модель факторами.

Пусть

$$e_i = y_i - \hat{y}_i$$

i-й остаток, где

$$\hat{y}_i = b_0 + b_1 x_{i,1} + \ldots + b_m x_{i,m}$$

прогноз для i-го наблюдения.

Остаточная вариация (residual sum of squares)

$$RSS = \sum_{i=1}^{n} (e_i)^2;$$

Стандартная ошибка (несмещенная оценка дисперсии ошибки):

$s^2 = RSS / (n - m - 1).$

Цель **дисперсионного анализа** регрессионной модели - проверить, насколько вариация (изменчивость) зависимой переменной объясняется включенными в модель факторами.

Пусть

$$e_i = y_i - \hat{y}_i$$

i-й остаток, где

$$\hat{y}_i = b_0 + b_1 x_{i,1} + \ldots + b_m x_{i,m}$$

прогноз для i-го наблюдения.

Остаточная вариация (residual sum of squares)

$$RSS = \sum_{i=1}^{n} (e_i)^2;$$

Стандартная ошибка (несмещенная оценка дисперсии ошибки):

$s^2 = RSS / (n - m - 1).$

Общая вариация $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2;$

Цель **дисперсионного анализа** регрессионной модели - проверить, насколько вариация (изменчивость) зависимой переменной объясняется включенными в модель факторами.

Пусть

$$e_i = y_i - \hat{y}_i$$

i-й остаток, где

$$\hat{y}_i = b_0 + b_1 x_{i,1} + \ldots + b_m x_{i,m}$$

прогноз для i-го наблюдения.

Остаточная вариация (residual sum of squares)

$$RSS = \sum_{i=1}^{n} (e_i)^2;$$

Стандартная ошибка (несмещенная оценка дисперсии ошибки):

$s^2 = RSS / (n - m - 1)$.

Общая вариация $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$;

Вариация, объясненная регрессией $ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$.

Основное тождество вариации:

$$\boxed{TSS = ESS + RSS}$$ (если $\beta_0 \neq 0$).

Основное тождество вариации:

$$\boxed{TSS = ESS + RSS}$$ (если $\beta_0 \neq 0$).

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

Основное тождество вариации:

$$\boxed{TSS = ESS + RSS}$$ (если $\beta_0 \neq 0$).

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

$$\sum_i (y_i - \bar{y})^2 = \sum_i \left((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\right)^2$$

Основное тождество вариации:

$$\boxed{TSS = ESS + RSS} \quad \text{(если } \beta_0 \neq 0\text{)}.$$

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

$$\sum_i (y_i - \bar{y})^2 = \sum_i \left((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})\right)^2$$

$$\sum_i (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2\sum_i (y_i - \hat{y}_i)\cdot(\hat{y}_i - \bar{y})$$

RSS

ESS

TSS

=0 если $\beta_0 \neq 0$

Коэффициент детерминации

$$R^2 = 1 - RSS/TSS = ESS/TSS; \quad R^2 \in [0,1]$$

показывает степень подгонки модели к наблюдаемым значениям *Y* (чем ближе к 1, тем лучше).

Коэффициент детерминации

$$R^2 = 1 - RSS/TSS = ESS/TSS; \quad R^2 \in [0,1]$$

показывает степень подгонки модели к наблюдаемым значениям *Y* (чем ближе к 1, тем лучше).

Для парной линейной модели $R^2 = r_{xy}^2$ (коэффициент корреляции)

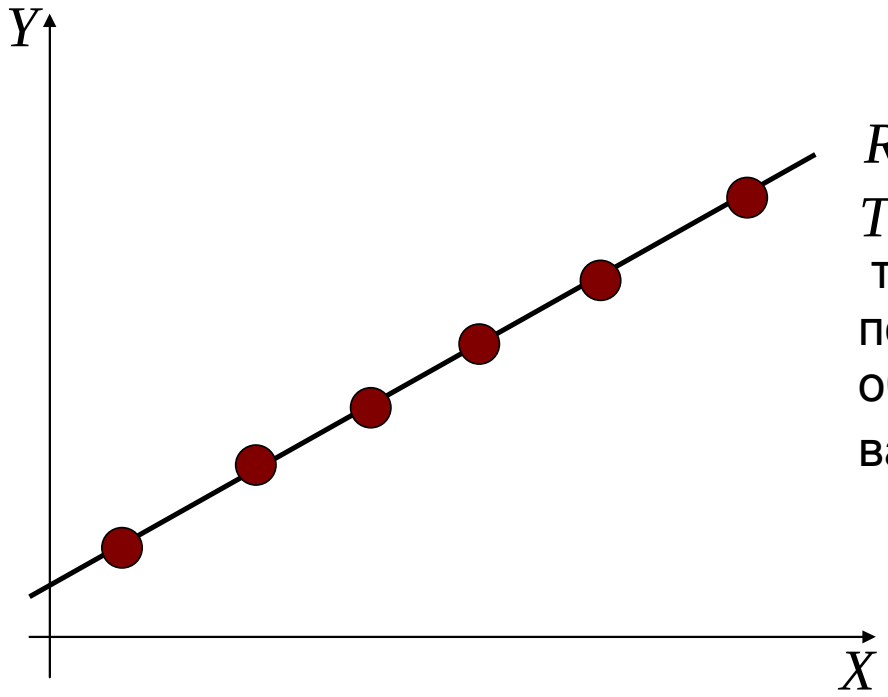
Коэффициент детерминации

$$R^2 = 1 - RSS/TSS = ESS/TSS; \quad R^2 \in [0,1]$$

показывает степень подгонки модели к наблюдаемым значениям Y (чем ближе к 1, тем лучше).

Для парной линейной модели $R^2 = r_{xy}^2$ (коэффициент корреляции)

Пример.



$RSS = 0,$

$TSS = ESS$

т.е. регрессия полностью объясняет вариацию Y

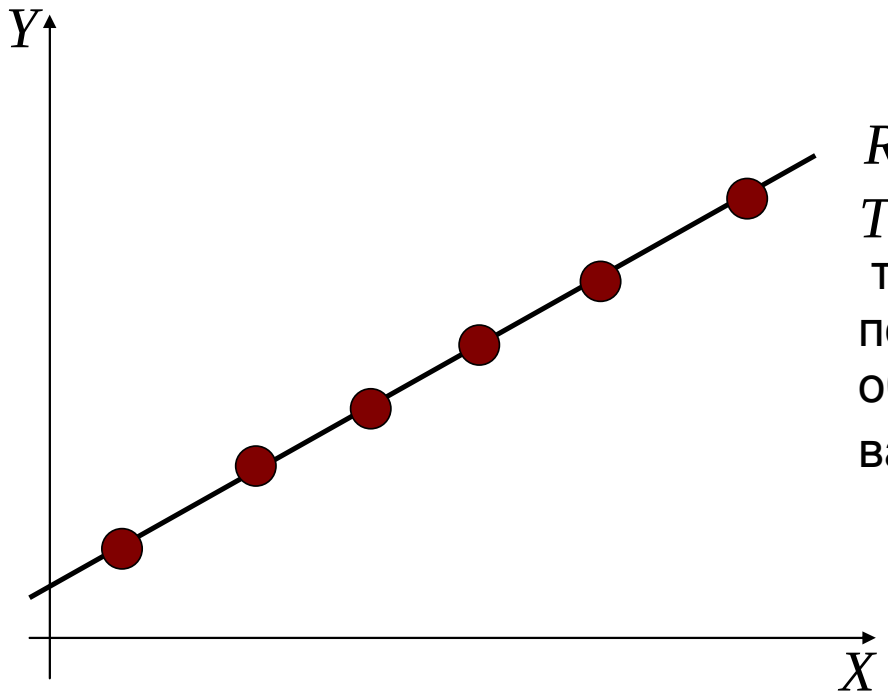
Коэффициент детерминации

$$R^2 = 1 - RSS/TSS = ESS/TSS; \quad R^2 \in [0,1]$$

показывает степень подгонки модели к наблюдаемым значениям Y (чем ближе к 1, тем лучше).

Для парной линейной модели $R^2 = r_{xy}^2$ (коэффициент корреляции)

Пример.



$RSS = 0,$

$TSS = ESS$

т.е. регрессия полностью объясняет вариацию Y

- Остаточная вариация – «необъясненная»

R^2 можно использовать для сравнения моделей по качеству (степени соответствия наблюдениям)

R^2 можно использовать для сравнения моделей по качеству (степени соответствия наблюдениям)

Недостаток R^2 - автоматически увеличивается при включении в модель дополнительных переменных (даже если они незначимы).

R^2 можно использовать для сравнения моделей по качеству (степени соответствия наблюдениям)

Недостаток R^2 - автоматически увеличивается при включении в модель дополнительных переменных (даже если они незначимы).

Нормированный (скорректированный, adjusted) коэффициент детерминации:

$$R^2_{adj} = 1 - \frac{RSS/(N-n-1)}{TSS/(N-1)}.$$

R^2 можно использовать для сравнения моделей по качеству (степени соответствия наблюдениям)

Недостаток R^2 - автоматически увеличивается при включении в модель дополнительных переменных (даже если они незначимы).

Нормированный (скорректированный, adjusted) коэффициент детерминации:

$$R^2_{adj} = 1 - \frac{RSS/(N-n-1)}{TSS/(N-1)}.$$

Свойства:

1. $R^2_{adj} = 1-(1-R^2)\frac{N-1}{N-n-1}$;

R^2 можно использовать для сравнения моделей по качеству (степени соответствия наблюдениям)

Недостаток R^2 - автоматически увеличивается при включении в модель дополнительных переменных (даже если они незначимы).

Нормированный (скорректированный, adjusted) коэффициент детерминации:

$$R^2_{adj} = 1 - \frac{RSS/(N-n-1)}{TSS/(N-1)}.$$

Свойства:

1. $R^2_{adj} = 1-(1-R^2)\frac{N-1}{N-n-1}$;
2. $R^2_{adj} \le R^2$ при $n > 2$;

R^2 можно использовать для сравнения моделей по качеству (степени соответствия наблюдениям)

Недостаток R^2 - автоматически увеличивается при включении в модель дополнительных переменных (даже если они незначимы).

Нормированный (скорректированный, adjusted) коэффициент детерминации:

$$R^2_{adj} = 1 - \frac{RSS/(N-n-1)}{TSS/(N-1)}.$$

Свойства:

1. $R^2_{adj} = 1-(1-R^2)\frac{N-1}{N-n-1}$;
2. $R^2_{adj} \le R^2$ при $n > 2$;
3. $R^2_{adj} \le 1$, но может быть $R^2_{adj} < 0$.

Проверка гипотезы о значимости регрессии:

$$H_0: "\beta_1 = \beta_2 = \ldots = \beta_m = 0".$$

Проверка гипотезы о значимости регрессии:

$$H_0: "\beta_1 = \beta_2 = \ldots = \beta_m = 0".$$

F-критерий:

$$F = \frac{R^2}{1-R^2}\frac{n-m-1}{m},$$

где R^2 - коэффициент детерминации.

Проверка гипотезы о значимости регрессии:

$$H_0: "\beta_1 = \beta_2 = \ldots = \beta_m = 0".$$

F-критерий:

$$F = \frac{R^2}{1-R^2}\frac{n-m-1}{m},$$

где R^2 - коэффициент детерминации.

H_0 отвергается на уровне значимости α, если

$$F_{набл} > F_{кр}(\alpha; m, n-m-1),$$

где $F_{кр}(\alpha; m, n-m-1)$ определяется из таблицы F-распределения.

Проверка гипотезы о значимости регрессии:

$$H_0: "\beta_1 = \beta_2 = \ldots = \beta_m = 0".$$

F-критерий:

$$F = \frac{R^2}{1-R^2}\frac{n-m-1}{m},$$

где R^2 - коэффициент детерминации.

H_0 отвергается на уровне значимости α, если

$$F_{набл} > F_{кр}(\alpha;\, m,\, n–m–1),$$

где $F_{кр}(\alpha;\, m,\, n–m–1)$ определяется из таблицы F-распределения.

p-value: $\mathrm{P}[F > F_{набл} \mid H_o]$ - используется в компьютерных пакетах статистического анализа.

Гипотеза о значимости переменной

(насколько существенно влияние X_j на Y) : H_{0j} =”$\beta_j = 0$” .

Гипотеза о значимости переменной

(насколько существенно влияние X_j на Y) : H_{0j} =”$\beta_j = 0$” .

Проверка: t- критерий Стьюдента.

Гипотеза о значимости переменной

(насколько существенно влияние X_j на Y) : H_{0j} =”$\beta_j = 0$” .

Проверка: t- критерий Стьюдента.

Пусть $t_{j\ набл} = \frac{b_j}{s_j}$, s_j - стандартная ошибка j -го параметра:

$s_j = s\sqrt{q_j}$, $q_j - j$ -й диагональный элемент матрицы $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$,

Гипотеза о значимости переменной

(насколько существенно влияние X_j на Y) : H_{0j} =”$\beta_j = 0$” .

Проверка: t- критерий Стьюдента.

Пусть $t_{j\ набл} = \dfrac{b_j}{s_j}$, s_j - стандартная ошибка j -го параметра:

$s_j = s\sqrt{q_j}$, $q_j - j$ -й диагональный элемент матрицы $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$,

Определим табличное значение $t_{кр}(\alpha,\ n-m-1)$,

где α - заданный уровень значимости; $n-m-1$ - число степеней свободы.

Гипотеза о значимости переменной

(насколько существенно влияние X_j на Y) : H_{0j} =”$\beta_j = 0$” .

Проверка: t- критерий Стьюдента.

Пусть $t_{j\ набл} = \dfrac{b_j}{s_j}$, s_j - стандартная ошибка j -го параметра:

$s_j = s\sqrt{q_j}$, $q_j - j$ -й диагональный элемент матрицы $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$,

Определим табличное значение $t_{кр}(\alpha,\ n-m-1)$,

где α - заданный уровень значимости; $n-m-1$ - число степеней свободы.

Если $|t_{j\ набл}| > t_{кр}$, то гипотеза отвергается.

Гипотеза о значимости переменной

(насколько существенно влияние X_j на Y) : H_{0j} =”$\beta_j = 0$” .

Проверка: t- критерий Стьюдента.

Пусть $t_{j\ набл} = \dfrac{b_j}{s_j}$, s_j - стандартная ошибка j -го параметра:

$s_j = s\sqrt{q_j}$, $q_j - j$ -й диагональный элемент матрицы $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$,

Определим табличное значение $t_{кр}(\alpha,\ n-m-1)$,

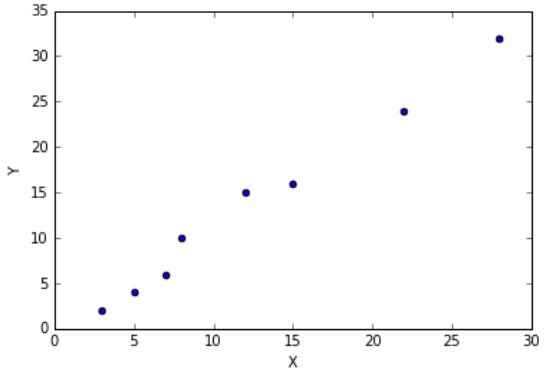
где α - заданный уровень значимости; $n-m-1$ - число степеней свободы.

Если $|t_{j\ набл}| > t_{кр}$, то гипотеза отвергается.

p-value: $\mathrm{P}[T_j > t_{j\ набл} \mid H_{oj}]$, где T_j величина с распределением Стьюдента;

- чем меньше p-value, тем более значима X_j.

Пример




```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import statsmodels.formula.api as smf
4
5 dat = pd.DataFrame({'X':[3,5,7,8,12,15,22,28],
6                     'Y':[2,4,6,10,15,16,24,32]});
7 dat.plot('X','Y',kind='scatter')
8
9 lm = smf.ols(formula="X ~ Y", data=dat).fit()
10
```

Результаты

```
ipdb> lm.summary()
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:                      X   R-squared:                       0.987
Model:                            OLS   Adj. R-squared:                  0.985
Method:                 Least Squares   F-statistic:                     454.5
Date:                Sun, 23 Apr 2017   Prob (F-statistic):           6.95e-07
Time:                        00:04:06   Log-Likelihood:                -10.793
No. Observations:                   8   AIC:                             25.59
Df Residuals:                       6   BIC:                             25.74
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      1.0926      0.657      1.664      0.147        -0.514     2.699
Y              0.8372      0.039     21.320      0.000         0.741     0.933
------------------------------------------------------------------------------
```