

Точность расчетной модели

Два полезных результата:

$$\bar{e} = 0$$

$$\sum X_i e_i = 0$$

Эта презентация посвящена адекватности (точности) расчетной модели в регрессионном анализе. Рассмотрим два результата. Во-первых, среднее значение остатков должно быть равно нулю.

Точность расчетной модели

Два полезных результата:

$$\bar{e} = 0$$

$$\sum X_i e_i = 0$$

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

Остаток в любом наблюдении определяется разницей между фактическим и расчетным значениями Y для этого наблюдения.

Точность расчетной модели

Два полезных результата:

$$\bar{e} = 0$$

$$\sum X_i e_i = 0$$

$$e_i = Y_i - \hat{Y}_i = Y_i - b_1 - b_2 X_i$$

$$\hat{Y}_i = b_1 + b_2 X_i$$

Вначале заменим расчетное значение выражением для него.

Точность расчетной модели

Два полезных результата:

$$\bar{e} = 0$$

$$\sum X_i e_i = 0$$

$$e_i = Y_i - \hat{Y}_i = Y_i - b_1 - b_2 X_i$$

$$\sum e_i = \sum Y_i - nb_1 - b_2 \sum X_i$$

Просуммируем все наблюдения.

Точность расчетной модели

Два полезных результата: $\bar{e} = 0$ $\sum X_i e_i = 0$

$$e_i = Y_i - \hat{Y}_i = Y_i - b_1 - b_2 X_i$$

$$\sum e_i = \sum Y_i - nb_1 - b_2 \sum X_i$$

$$\frac{1}{n} \sum e_i = \frac{1}{n} \sum Y_i - b_1 - b_2 \frac{1}{n} \sum X_i$$

$$\bar{e} = \bar{Y} - b_1 - b_2 \bar{X}$$

Разделив на n , мы получим среднее значение остатков, выраженное через значения X , Y и коэффициенты регрессии.

Точность расчетной модели

Два полезных результата: $\bar{e} = 0$ $\sum X_i e_i = 0$

$$e_i = Y_i - \hat{Y}_i = Y_i - b_1 - b_2 X_i$$

$$\sum e_i = \sum Y_i - nb_1 - b_2 \sum X_i$$

$$\frac{1}{n} \sum e_i = \frac{1}{n} \sum Y_i - b_1 - b_2 \frac{1}{n} \sum X_i$$

$$\begin{aligned} \bar{e} &= \bar{Y} - b_1 - b_2 \bar{X} \\ &= \bar{Y} - (\bar{Y} - b_2 \bar{X}) - b_2 \bar{X} \\ &= 0 \end{aligned}$$

$$b_1 = \bar{Y} - b_2 \bar{X}$$

Если мы заменим b_1 , выражение будет равно нулю.

Точность расчетной модели

Два полезных результата:

$$\bar{e} = 0$$

$$\sum X_i e_i = 0$$

$$\bar{e} = 0$$

$$\bar{e} = \frac{1}{n} \sum e_i$$

$$\sum e_i = 0$$

Этот результат можно записать в другом виде: сумма остатков должна быть равна нулю.

Точность расчетной модели

Два полезных результата:

$$\bar{e} = 0$$

$$\sum X_i e_i = 0$$

Дальнейшие результаты:

$$\bar{\hat{Y}} = \bar{Y}$$

Из этого результата следует, что среднее значение расчетных значений \hat{Y} равно среднему значению фактических значений Y .

Точность расчетной модели

Два полезных результата:

$$\bar{e} = 0$$

$$\sum X_i e_i = 0$$

Дальнейшие результаты:

$$\hat{Y} = \bar{Y}$$

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

Опять начнем с определения остатка.

Точность расчетной модели

Два полезных результата:

$$\bar{e} = 0$$

$$\sum X_i e_i = 0$$

Дальнейшие результаты:

$$\hat{Y} = \bar{Y}$$

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

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

Просуммируем все наблюдения.

Точность расчетной модели

Два полезных результата:

$$\bar{e} = 0$$

$$\sum X_i e_i = 0$$

Дальнейшие результаты:

$$\hat{Y} = \bar{Y}$$

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

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

$$0 = \frac{1}{n} \sum Y_i - \frac{1}{n} \sum \hat{Y}_i = \bar{Y} - \hat{Y}$$

В левой части сумма остатков равна нулю. Теперь выражение разделим на n .

Точность расчетной модели

Два полезных результата:

$$\bar{e} = 0$$

$$\sum X_i e_i = 0$$

Дальнейшие результаты:

$$\hat{Y} = \bar{Y}$$

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

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

$$0 = \frac{1}{n} \sum Y_i - \frac{1}{n} \sum \hat{Y}_i = \bar{Y} - \bar{\hat{Y}}$$

$$\bar{\hat{Y}} = \bar{Y}$$

Следовательно, среднее значение расчетных значений равно среднему значению фактических значений.

Точность расчетной модели

Два полезных результата:

$$\bar{e} = 0$$

$$\sum X_i e_i = 0$$

Дальнейшие результаты:

$$\widehat{Y} = \bar{Y}$$

Далее покажем, что сумма произведений значений X и остатков равна нулю.

Точность расчетной модели

Два полезных результата: $\bar{e} = 0$ $\sum X_i e_i = 0$

Дальнейшие результаты: $\hat{Y} = \bar{Y}$

$$\sum X_i e_i = \sum X_i (Y_i - b_1 - b_2 X_i)$$

$$e_i = Y_i - \hat{Y}_i = Y_i - b_1 - b_2 X_i$$

Начнем с замены остатка его выражением через Y и X.

Точность расчетной модели

Два полезных результата: $\bar{e} = 0$ $\sum X_i e_i = 0$

Дальнейшие результаты: $\hat{Y} = \bar{Y}$

$$\begin{aligned}\sum X_i e_i &= \sum X_i (Y_i - b_1 - b_2 X_i) \\ &= \sum X_i Y_i - b_1 \sum X_i - b_2 \sum X_i^2\end{aligned}$$

Упростим выражение.

Точность расчетной модели

Два полезных результата: $\bar{e} = 0$ $\sum X_i e_i = 0$

Дальнейшие результаты: $\hat{Y} = \bar{Y}$

$$\begin{aligned}\sum X_i e_i &= \sum X_i (Y_i - b_1 - b_2 X_i) \\ &= \sum X_i Y_i - b_1 \sum X_i - b_2 \sum X_i^2 = 0\end{aligned}$$

Выражение равно нулю. Одним из способов продемонстрировать это: заменить b_1 и b_2 и показать, что все слагаемые сокращаются.

Точность расчетной модели

Два полезных результата: $\bar{e} = 0$ $\sum X_i e_i = 0$

Дальнейшие результаты: $\hat{Y} = \bar{Y}$

$$\begin{aligned}\sum X_i e_i &= \sum X_i (Y_i - b_1 - b_2 X_i) \\ &= \sum X_i Y_i - b_1 \sum X_i - b_2 \sum X_i^2 = 0\end{aligned}$$

$$\frac{\partial RSS}{\partial b_2} = 0 \Rightarrow 2b_2 \sum X_i^2 - 2 \sum X_i Y_i + 2b_1 \sum X_i = 0$$

Точный способ - вспомнить условие первого порядка для b_2 при выводе коэффициентов регрессии. Вы можете видеть, что так оно и есть.

Точность расчетной модели

Два полезных результата: $\bar{e} = 0$ $\sum X_i e_i = 0$

Дальнейшие результаты: $\bar{\hat{Y}} = \bar{Y}$ $\sum \hat{Y}_i e_i = 0$

Наконец, побочным результатом нашего последнего расчета, является равенство нулю суммы произведений расчетных значений \hat{Y} и остатков.

Точность расчетной модели

Два полезных результата: $\bar{e} = 0$ $\sum X_i e_i = 0$

Дальнейшие результаты: $\bar{\hat{Y}} = \bar{Y}$ $\sum \hat{Y}_i e_i = 0$

$$\sum \hat{Y}_i e_i = \sum (b_1 + b_2 X_i) e_i \quad \hat{Y}_i = b_1 + b_2 X_i$$

Сначала подставим расчетные значения \hat{Y} .

Точность расчетной модели

Два полезных результата: $\bar{e} = 0$ $\sum X_i e_i = 0$

Дальнейшие результаты: $\hat{Y} = \bar{Y}$ $\sum \hat{Y}_i e_i = 0$

$$\begin{aligned}\sum \hat{Y}_i e_i &= \sum (b_1 + b_2 X_i) e_i \\ &= \sum b_1 e_i + \sum b_2 X_i e_i \\ &= b_1 n \bar{e} + b_2 \sum X_i e_i\end{aligned}$$

$$\uparrow$$
$$\sum e_i = n \bar{e}$$

Производим расчеты.

Точность расчетной модели

Два полезных результата: $\bar{e} = 0$ $\sum X_i e_i = 0$

Дальнейшие результаты: $\hat{Y} = \bar{Y}$ $\sum \hat{Y}_i e_i = 0$

$$\begin{aligned}\sum \hat{Y}_i e_i &= \sum (b_1 + b_2 X_i) e_i \\ &= \sum b_1 e_i + \sum b_2 X_i e_i \\ &= b_1 n \bar{e} + b_2 \sum X_i e_i = 0\end{aligned}$$

\uparrow \uparrow

$\bar{e} = 0$ $\sum X_i e_i = 0$

Выражение равно нулю, учитывая ранее полученные результаты.

$$\sum (Y_i - \bar{Y})^2$$

Мерой вариации Y является сумма его квадратов отклонений от среднего значения выборки. Это называется общей суммой квадратов TSS .

$$\sum (Y_i - \bar{Y})^2$$

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

Мы разложим общую сумму квадратов, используя тот факт, что фактическое значение Y в любом наблюдении равно сумме его расчетного значения и остатка.

$$\sum (Y_i - \bar{Y})^2 = \sum ([\hat{Y}_i + e_i] - \bar{Y})^2$$

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

Подставим Y_i

$$\begin{aligned}\sum (Y_i - \bar{Y})^2 &= \sum ([\hat{Y}_i + e_i] - \bar{Y})^2 \\ &= \sum ([\hat{Y}_i - \bar{Y}] + e_i)^2\end{aligned}$$

Перегруппируем члены.

$$\begin{aligned}\sum (Y_i - \bar{Y})^2 &= \sum ([\hat{Y}_i + e_i] - \bar{Y})^2 \\ &= \sum ([\hat{Y}_i - \bar{Y}] + e_i)^2 \\ &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2 + 2 \sum ([\hat{Y}_i - \bar{Y}] e_i)\end{aligned}$$

Разложим квадрат в правой части уравнения.

$$\begin{aligned}\sum (Y_i - \bar{Y})^2 &= \sum ([\hat{Y}_i + e_i] - \bar{Y})^2 \\ &= \sum ([\hat{Y}_i - \bar{Y}] + e_i)^2 \\ &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2 + 2 \sum ([\hat{Y}_i - \bar{Y}] e_i) \\ &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2 + 2 \sum \hat{Y}_i e_i - 2 \bar{Y} \sum e_i\end{aligned}$$

Разложим третий член в правой части уравнения.

$$\begin{aligned}
 \sum (Y_i - \bar{Y})^2 &= \sum ([\hat{Y}_i + e_i] - \bar{Y})^2 \\
 &= \sum ([\hat{Y}_i - \bar{Y}] + e_i)^2 \\
 &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2 + 2 \sum ([\hat{Y}_i - \bar{Y}] e_i) \\
 &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2 + 2 \sum \hat{Y}_i e_i - 2\bar{Y} \sum e_i
 \end{aligned}$$

$\sum \hat{Y}_i e_i = 0$

$\sum e_i = 0$

Последние два члена равны нулю, учитывая ранее полученные результаты.

$$\begin{aligned}
 \sum (Y_i - \bar{Y})^2 &= \sum ([\hat{Y}_i + e_i] - \bar{Y})^2 \\
 &= \sum ([\hat{Y}_i - \bar{Y}] + e_i)^2 \\
 &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2 + 2 \sum ([\hat{Y}_i - \bar{Y}] e_i) \\
 &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2
 \end{aligned}$$

$$TSS = ESS + RSS$$

$$\sum (Y_i - \bar{Y})^2 = TSS, \text{ total sum of squares}$$

$$\sum (\hat{Y}_i - \bar{Y})^2 = ESS, \text{ explained sum of squares}$$

$$\sum e_i^2 = RSS, \text{ residual sum of squares}$$

Мы показали, что TSS, общая сумма квадратов Y может быть разложена на ESS, объяснённую сумму квадратов, и RSS, сумму квадратов остатков.

$$\begin{aligned}
 \sum (Y_i - \bar{Y})^2 &= \sum ([\hat{Y}_i + e_i] - \bar{Y})^2 \\
 &= \sum ([\hat{Y}_i - \bar{Y}] + e_i)^2 \\
 &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2 + 2 \sum ([\hat{Y}_i - \bar{Y}] e_i) \\
 &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2
 \end{aligned}$$

$$TSS = ESS + RSS$$

$$\sum (Y_i - \bar{Y})^2 = TSS, \text{ total sum of squares}$$

$$\sum (\hat{Y}_i - \bar{Y})^2 = ESS, \text{ explained sum of squares}$$

$$\sum e_i^2 = RSS, \text{ residual sum of squares}$$

Слова, «объясненные» и «необъяснимые», заключены в кавычки, потому что объяснение может быть ложным. Y может действительно зависеть от некоторой другой переменной Z , а X может выступать в качестве замены для Z .

$$\begin{aligned}
 \sum (Y_i - \bar{Y})^2 &= \sum ([\hat{Y}_i + e_i] - \bar{Y})^2 \\
 &= \sum ([\hat{Y}_i - \bar{Y}] + e_i)^2 \\
 &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2 + 2 \sum ([\hat{Y}_i - \bar{Y}] e_i) \\
 &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2
 \end{aligned}$$

$$TSS = ESS + RSS$$

$$\sum (Y_i - \bar{Y})^2 = TSS, \text{ total sum of squares}$$

$$\sum (\hat{Y}_i - \bar{Y})^2 = ESS, \text{ explained sum of squares}$$

$$\sum e_i^2 = RSS, \text{ residual sum of squares}$$

Правильнее, «явно объясненные» вместо «объясненные».

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2 \quad TSS = ESS + RSS$$

$$R^2 = \frac{ESS}{TSS} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

Основным критерием точности расчетной модели является коэффициент детерминации R^2 , определяемый как отношение ESS к TSS, то есть часть дисперсии Y , объясняемая уравнением регрессии.

Точность расчетной модели

```
. reg EARNINGS S
```

Source	SS	df	MS			
Model	19321.5589	1	19321.5589	Number of obs =	540	
Residual	92688.6722	538	172.283777	F(1, 538) =	112.15	
Total	112010.231	539	207.811189	Prob > F =	0.0000	
				R-squared =	0.1725	
				Adj R-squared =	0.1710	
				Root MSE =	13.126	

EARNINGS	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
S	2.455321	.2318512	10.59	0.000	1.999876	2.910765
_cons	-13.93347	3.219851	-4.33	0.000	-20.25849	-7.608444

Статистика для регрессии почасового заработка по годам обучения.

Точность расчетной модели

```
. reg EARNINGS S
```

Source	SS	df	MS			
Model	19321.5589	1	19321.5589	Number of obs =	540	
Residual	92688.6722	538	172.283777	F(1, 538) =	112.15	
				Prob > F =	0.0000	
				R-squared =	0.1725	
				Adj R-squared =	0.1710	
Total	112010.231	539	207.811189	Root MSE =	13.126	

EARNINGS	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
S	2.455321	.2318512	10.59	0.000	1.999876	2.910765
_cons	-13.93347	3.219851	-4.33	0.000	-20.25849	-7.608444

$$R^2 = \frac{ESS}{TSS} = \frac{19,322}{112,010} = 0.1725$$

Заголовок столбца «SS» обозначает суммы квадратов. ESS, названная как «модельная» сумма квадратов, составляет 19322. TSS составляет 112010.

Точность расчетной модели

```
. reg EARNINGS S
```

Source	SS	df	MS	Number of obs = 540		
Model	19321.5589	1	19321.5589	F(1, 538)	=	112.15
Residual	92688.6722	538	172.283777	Prob > F	=	0.0000
Total	112010.231	539	207.811189	R-squared	=	0.1725

EARNINGS	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
S	2.455321	.2318512	10.59	0.000	1.999876	2.910765
_cons	-13.93347	3.219851	-4.33	0.000	-20.25849	-7.608444

$$R^2 = \frac{ESS}{TSS} = \frac{19,322}{112,010} = 0.1725$$

Разделив ESS на TSS, мы имеем $R^2 = 19,322 / 112,010 = 0.1725$, как указано в верхнем правом углу слайда.

Точность расчетной модели

```
. reg EARNINGS S
```

Source	SS	df	MS			
Model	19321.5589	1	19321.5589	Number of obs =	540	
Residual	92688.6722	538	172.283777	F(1, 538) =	112.15	
				Prob > F =	0.0000	
				R-squared =	0.1725	
				Adj R-squared =	0.1710	
Total	112010.231	539	207.811189	Root MSE =	13.126	

EARNINGS	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
S	2.455321	.2318512	10.59	0.000	1.999876	2.910765
_cons	-13.93347	3.219851	-4.33	0.000	-20.25849	-7.608444

$$R^2 = \frac{ESS}{TSS} = \frac{19,322}{112,010} = 0.1725$$

Низкий R^2 частично объясняется тем, что в модели отсутствуют важные переменные, такие как опыт работы.

Точность расчетной модели

```
. reg EARNINGS S
```

Source	SS	df	MS			
Model	19321.5589	1	19321.5589	Number of obs =	540	
Residual	92688.6722	538	172.283777	F(1, 538) =	112.15	
				Prob > F =	0.0000	
				R-squared =	0.1725	
				Adj R-squared =	0.1710	
Total	112010.231	539	207.811189	Root MSE =	13.126	

EARNINGS	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
S	2.455321	.2318512	10.59	0.000	1.999876	2.910765
_cons	-13.93347	3.219851	-4.33	0.000	-20.25849	-7.608444

$$R^2 = \frac{ESS}{TSS} = \frac{19,322}{112,010} = 0.1725$$

Это также частично объясняется тем фактом, что ненаблюдаемые характеристики важны для определения зарплаты, R^2 редко намного превышает 0,5 даже в хорошо определенной модели.

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2 \quad TSS = ESS + RSS$$

$$R^2 = \frac{ESS}{TSS} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

Мы хотим построить уравнение регрессии так, чтобы точность была максимально возможной согласно R^2 . Возможно ли это при определении b_1 и b_2 с помощью метода наименьших квадратов?

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2 \quad TSS = ESS + RSS$$

$$R^2 = \frac{ESS}{TSS} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

Чтобы увидеть это, перепишите выражение для R^2 в терминах RSS.

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2 \quad TSS = ESS + RSS$$

$$R^2 = \frac{ESS}{TSS} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

При МНК коэффициенты регрессии выбираются таким образом, чтобы минимизировать сумму квадратов остатков. Из этого автоматически следует, что они максимизируют R^2 .

$$r_{Y,\hat{Y}} = \frac{\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{Y})^2}}$$

Другим критерием точности является корреляция между фактическими и расчетными значениями Y . Если для расчета коэффициентов регрессии используется МНК, то точность расчетной модели становится максимальной.

$$r_{Y,\hat{Y}} = \frac{\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{Y})^2}}$$

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

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

Заменяем фактическое значение Y в первом сомножителе.

$$r_{Y,\hat{Y}} = \frac{\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{Y})^2}}$$

$$\begin{aligned}\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) &= \sum ([\hat{Y}_i + e_i] - \bar{Y})(\hat{Y}_i - \bar{Y}) \\ &= \sum ([\hat{Y}_i - \bar{Y}] + e_i)(\hat{Y}_i - \bar{Y})\end{aligned}$$

Делаем перестановки.

$$r_{Y,\hat{Y}} = \frac{\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{Y})^2}}$$

$$\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) = \sum ([\hat{Y}_i + e_i] - \bar{Y})(\hat{Y}_i - \bar{Y})$$

$$= \sum ([\hat{Y}_i - \bar{Y}] + e_i)(\hat{Y}_i - \bar{Y})$$

$$= \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i \hat{Y}_i - \bar{Y} \sum e_i$$

$$\sum \hat{Y}_i e_i = 0$$

$$\sum e_i = 0$$

Разложим выражение. Последние два члена равны нулю.

$$r_{Y,\hat{Y}} = \frac{\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{Y})^2}} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sqrt{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{Y})^2}}$$

$$\begin{aligned} \sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) &= \sum ([\hat{Y}_i + e_i] - \bar{Y})(\hat{Y}_i - \bar{Y}) \\ &= \sum ([\hat{Y}_i - \bar{Y}] + e_i)(\hat{Y}_i - \bar{Y}) \\ &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i \hat{Y}_i - \bar{Y} \sum e_i \\ &= \sum (\hat{Y}_i - \bar{Y})^2 \end{aligned}$$

Таким образом, числитель равен сумме квадратов отклонений расчетных значений \hat{Y} от среднего значения \bar{Y} .

$$\begin{aligned}
 r_{Y,\hat{Y}} &= \frac{\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{Y})^2}} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sqrt{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{Y})^2}} \\
 &= \frac{\sqrt{\sum (\hat{Y}_i - \bar{Y})^2}}{\sqrt{\sum (Y_i - \bar{Y})^2}}
 \end{aligned}$$

Мы имеем то же выражение под квадратным корнем в знаменателе. Следовательно, это выражение под квадратным корнем остается в числителе.

$$\begin{aligned}
 r_{Y,\hat{Y}} &= \frac{\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})}{\sqrt{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{Y})^2}} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sqrt{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{Y})^2}} \\
 &= \frac{\sqrt{\sum (\hat{Y}_i - \bar{Y})^2}}{\sqrt{\sum (Y_i - \bar{Y})^2}} = \sqrt{\frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}} = \sqrt{R^2}
 \end{aligned}$$

Таким образом, коэффициент корреляции является квадратным корнем из R^2 . Что и требовалось доказать.