Лекция 2

1.3. Коэффициент корреляции

В эконометрическом исследовании вопрос о наличии или отсутствии зависимости между анализируемыми переменными решается с помощью методов корреляционного анализа. Только после утвердительного ответа на этот вопрос имеет смысл определять вид зависимости. Корреляционный анализ подробно изучается в курсе математической статистики, напомним некоторые его положения.

Корреляционный анализ *– метод, применяемый тогда, когда данные наблюдений или эксперимента можно считать случайными и выбранными из совокупности, распределенной по многомерному нормальному закону*.

В корреляционном анализе исследуют следующие варианты зависимостей.

1. ***Парную корреляцию*** – связь между двумя признаками (результативным и факторным или двумя факторными).

2. ***Частную корреляцию*** – зависимость между результативным и одним факторным признаком при фиксированных значениях других факторных признаков.

3. ***Множественную корреляцию*** – зависимость между результативным и двумя и более факторными признаками.

Основная задача корреляционного анализа *состоит в выявлении связи между случайными переменными путем точечной и интервальной оценок различных (парных, частных, множественных) коэффициентов корреляции*.

Для оценки тесноты связей количественных признаков чаще всего используются следующие показатели: коэффициент линейной корреляции, эмпирическое корреляционное отношение, теоретическое корреляционное отношение (индекс корреляции), коэффициент множественной корреляции, частные коэффициенты корреляции, коэффициент детерминации. Для оценки тесноты связей качественных признаков используются: коэффициент ассоциации, ранговые коэффициенты Спирмена и Кендалла, коэффициент конкордации. Большинство перечисленных показателей будут использоваться в данном лекционном курсе. В данном параграфе напомним определение и свойства коэффициента корреляции.

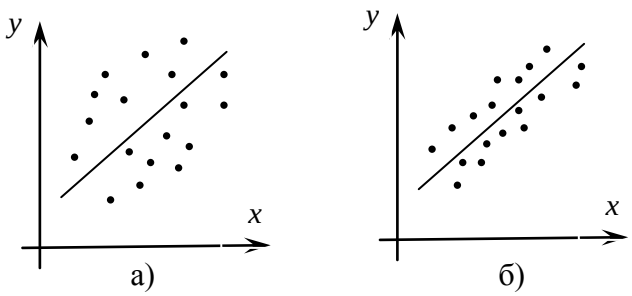
Перейдем к оценке тесноты парной корреляционной зависимости с помощью коэффициента корреляции. Рассмотрим наиболее важный для практики случай линейной зависимости. В теории вероятностей показателем тесноты линейной зависимости являлся коэффициент корреляции, в математической статистике таким показателем является выборочный коэффициент корреляции.

Выборочным коэффициентом (линейной) корреляции называется величина, рассчитываемая по формуле

$$r = r_{xy} = r_B = r(x, y) = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y}, \qquad (1.2)$$

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} y_i x_i$ – выборочные средние, $\sigma_x = \sqrt{\overline{x^2} - \bar{x}^2}$, $\sigma_y = \sqrt{\overline{y^2} - \bar{y}^2}$ – выборочные средние квадратические отклонения, полученные по наблюдаемым значениям x и y соответственно.

Выборочный коэффициент линейной корреляции r является показателем тесноты связи признаков в **линейной** форме (на фоне влияния остальных признаков, входящих в модель). На рисунке 1.2 приведены две корреляционные зависимости переменной y по x. Очевидно, что в случае а) зависимость между переменными менее тесная и коэффициент корреляции должен быть меньше, чем в случае б), так как точки корреляционного поля а) дальше отстоят от линии регрессии, чем точки поля б).



а) б)

Рисунок 1.2

Отметим основные свойства выборочного коэффициента корреляции (при достаточно большом объеме выборки n), аналогичные свойствам коэффициента корреляции для двух случайных величин.

1. Коэффициент корреляции принимает значения на отрезке $[-1,1]$, то есть $-1 \le r \le 1$.

2. Чем ближе значение $|r|$ к единице, тем более тесная ***линейная*** зависимость между изучаемыми величинами. В зависимости от того, насколько $|r|$ приближается к единице, различают *слабую*, *умеренную*, *заметную*, *достаточно тесную* и *весьма тесную* линейную связь.

3. Нетрудно видеть, что r совпадает по знаку с $\hat{\beta}_1$. Если $r > 0$ ($\hat{\beta}_1 > 0$), то корреляционная связь между переменными называется *прямой*, а если $r < 0$ ($\hat{\beta}_1 < 0$) – *обратной*. При прямой связи увеличение одной из переменных ведет к увеличению условной средней другой, при обратной наоборот.

4. Если все значения переменных увеличить (уменьшить) на одно и то же число или в одно и то же число раз, то величина коэффициента корреляции не изменится. Коэффициент корреляции есть безразмерная характеристика тесноты линейной связи.

5. При $r = \pm 1$ корреляционная связь представляет линейную функциональную зависимость, при этом все точки поля корреляции лежат на одной прямой. И наоборот, если x и y связаны линейной функциональной зависимостью, то $|r| = 1$.

6. Парный коэффициент корреляции является симметричной характеристикой, то есть $r_{xy} = r_{yx}$.

7. При $r = 0$ ***линейная*** корреляционная связь отсутствует, а величины x и y называют ***некоррелированными***. Но это не означает отсутствие вообще корреляционной, а тем более статистической зависимости. Например, нелинейная корреляционная связь может быть очень тесной.

8. Если случайные величины x и y статистически независимы, то $r_{xy} = 0$. Обратное верно не всегда. *В случае нормального распределения из некоррелированности x и y, когда $r_{xy} = 0$, следует их независимость*.

Как всякая выборочная характеристика, (1.5) является точечной оценкой генерального коэффициента корреляции $r_г$. Так как r вычисляется по значениям переменных, случайно попавшим в выборку из генеральной совокупности, то в отличие от параметра $r_г$ параметр r – случайная величина.

Пусть вычисленное значение $r \neq 0$. Возникает вопрос, объясняется ли это действительно существующей линейной корреляционной связью между переменными x и y в генеральной совокупности или является следствием случайности отбора переменных в выборку (т.е. при другом отборе возможно, например, $r = 0$ или изменение знака r).

Для ответа на вопрос о значимости коэффициента корреляции проверяют нулевую гипотезу $H_0: r_г = 0$ о равенстве нулю генерального коэффициента корреляции. Если гипотеза принимается, то это означает, что между x и y нет линейной корреляционной зависимости, в противном случае линейная зависимость признается значимой.

Для того чтобы при уровне значимости α проверить нулевую гипотезу при конкурирующей $H_1: r_г \neq 0$, надо вычислить наблюдаемое значение критерия

$$t_{набл} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

и по таблице критических точек распределения Стьюдента, по заданному уровню значимости α и числу степеней свободы $\nu = n-2$ найти критическую точку $t_{кр} = t_{\alpha;n-2}$ двухсторонней критической области. Если $|t_{набл}| < t_{\alpha;n-2}$ – нет оснований отвергнуть нулевую гипотезу. Если $|t_{набл}| > t_{\alpha;n-2}$ – нулевую гипотезу отвергаем и корреляционную связь между переменными признаем значимой.

В случае анализа зависимости компонент m-мерного случайного вектора $x = (x_1, x_2, ..., x_m)$ используется матрица парных коэффициентов корреляции

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} & \dots & r_{1m} \\ r_{21} & 1 & r_{23} & \dots & r_{2m} \\ r_{31} & r_{32} & 1 & \dots & r_{3m} \\ \dots & \dots & \dots & \dots & \dots \\ r_{m1} & r_{m2} & r_{m3} & \dots & 1 \end{pmatrix},$$

где $r_{ij} = r(x_i, x_j)$ – парный коэффициент корреляции между признаками x_i, x_j.

Матрица парных коэффициентов корреляции обладает свойствами, вытекающими из свойств парного коэффициента корреляции.

1. Симметрична относительно главной диагонали, так как $r_{ij} = r_{ji}$.
2. На главной диагонали стоят единицы, так как $r(x, x) = 1$.
3. Если компоненты вектора x попарно независимы, то R – единичная матрица.

1.4. Парная линейная регрессия. Оценка параметров

Модель парной линейной регрессии является наиболее распространенным видом зависимости. Пусть имеются n наблюдений над переменными x и y, то есть пары (x_i, y_i), $i = 1,2,...,n$.

Пример 1.1. Исследовать зависимость расходов на покупку продовольственных товаров y (% к общему объему расходов) от размера среднемесячной заработной платы одного работающего x (у.е.). Опытные данные, за 20XX г. по десяти районам области представлены в таблице 1.1.

Таблица 1.1

№	x (%)	y (у.е.)
1	4,5	68,8
2	5,9	58,3
3	5,7	62,6
4	7,2	52,1
5	6,2	54,5
6	6,0	57,1

7	7,8	51,0
8	7,5	50,7
9	8,1	48,6
10	7,9	49,1

Отобразим пары наблюдений точками на графике, рисунок 1.3.

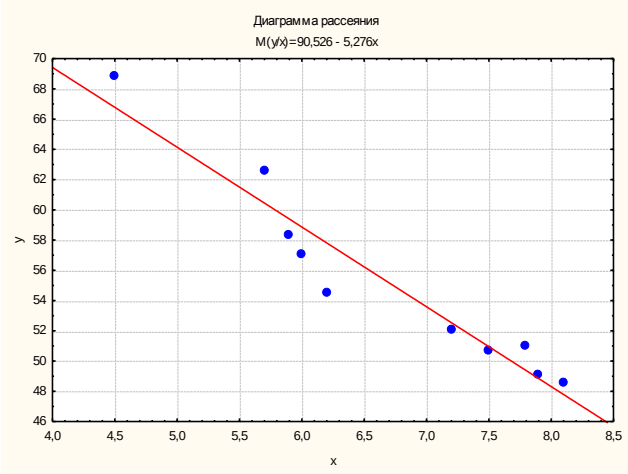


Рисунок 1.3. Диаграмма рассеяния

На основании анализа диаграммы рассеяния можем сделать предположение, что в среднем y есть линейная функция от x, т. е. имеет место уравнение парной линейной регрессии

$$M(y/x)=\beta_0+\beta_1 x,$$

где $M(y/x)$ – условное математическое ожидание случайной величины y при заданном x. Объясняющая переменная x рассматривается как неслучайная величина, β_0, β_1 – неизвестные параметры генеральной совокупности, которые подлежат оценке по результатам выборочных наблюдений.

Для отражения факта, что каждое индивидуальное значение y_i отклоняется от соответствующего математического ожидания, необходимо ввести случайное слагаемое ε, и тогда для наблюдений (x_i, y_i) уравнение регрессии имеет вид

$$y_i=M(y/x=x_i)+\varepsilon_i=\beta_0+\beta_1 x_i+\varepsilon_i,\quad i=\overline{1,n}. \qquad (1.3)$$

Соотношение (1.3) – ***теоретическая линейная регрессионная модель***, β_0, β_1 – ***теоретические параметры регрессии***, ε_i – ***случайные отклонения***. В общем виде теоретическую линейную модель будем представлять в виде:

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

Для определения значений теоретических коэффициентов регрессии необходимо знать и использовать все значения переменных y и x генеральной совокупности, что практически невозможно. В этом случае речь может идти об оценке (приближенном выражении) по выборке функции регрессии. Таким образом, **задача линейного регрессионного анализа** состоит в том, чтобы по имеющимся статистическим данным (x_i, y_i) объема n построить ***эмпирическое уравнение регрессии***

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i, \qquad (1.4)$$

где $\hat{y}_i$ – оценка условного математического ожидания $M(y/x=x_i)$, $\hat{\beta}_0$, $\hat{\beta}_1$ – оценки неизвестных параметров β_0, β_1 или ***эмпирические коэффициенты регрессии***. Следовательно

$$y_i=\hat{\beta}_0+\hat{\beta}_1 x_i+e_i,$$

где e_i – оценка ε_i.

Оценить параметры β_0, β_1 в данном случае означает выбрать «наилучшие» значения параметров, при которых линия регрессии (1.4) будет ближайшей к точкам наблюдений по их совокупности, см. рисунок 1.4.

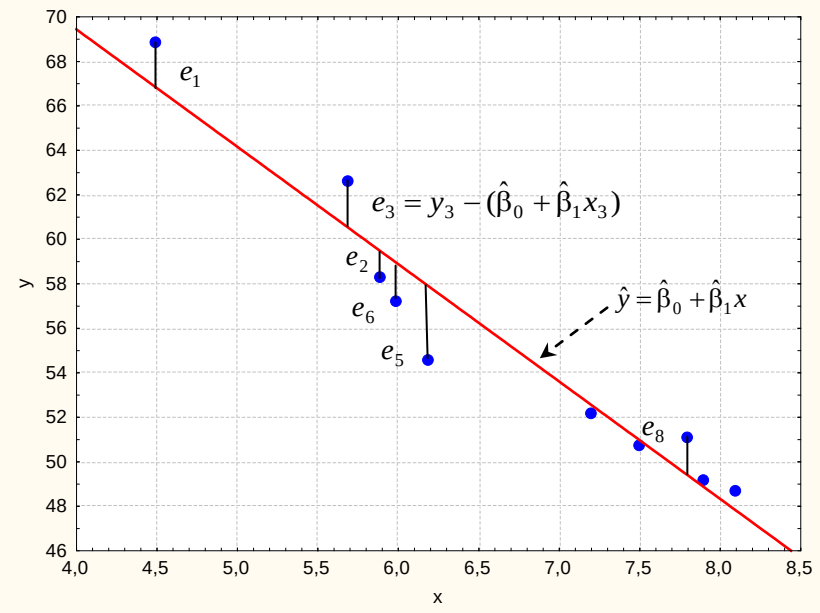


Рисунок 1.4

Например, коэффициенты $\hat{\beta}_0$, $\hat{\beta}_1$ могут быть найдены из условий минимизации одной из следующих сумм:

$$1.\ \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i),$$

$$2.\ \sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} |y_i - \hat{y}_i| = \sum_{i=1}^{n} |y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i|,$$

$$3.\ \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2,$$

Первая сумма не может быть мерой качества найденных оценок в силу того, что существует бесчисленное множество прямых (в частности $y = \bar{y}$), для которых $\sum_{i=1}^{n} e_i = 0$.

На минимизации второй суммы основан метод наименьших модулей. Плюсом данного метода является робастность, т.е. нечувствительность к выбросам. К минусам – сложность вычислительной процедуры, неоднозначность выводов.

На минимизации третьей суммы основан метод наименьших квадратов. Этот метод является наиболее простым с вычислительной точки зрения. Кроме того его оценки обладают рядом оптимальных статистических свойств. Простота математических выводов делает возможным построить развитую теорию, позволяющую провести тщательную проверку различных статистических гипотез. Минусом данного метода является чувствительность к «выбросам».

Среди других методов следует отметить метод моментов (ММ) и метод максимального правдоподобия (ММП), рассмотренные в курсе математической статистики.

1.5. Метод наименьших квадратов (МНК)

Согласно методу наименьших квадратов (МНК) в качестве оценок неизвестных параметров β_0, β_1 *следует брать такие значения* $\hat{\beta}_0$, $\hat{\beta}_1$, *которые минимизирует сум-*

му квадратов отклонений фактических значений результативного признака y_i *от значений* $\hat{y}_i$*, рассчитанных по уравнению регрессии (1.4)*, см. рисунок 1.3, т.е. из условия

$$Q(\hat{\beta}_0,\ \hat{\beta}_1)=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2\longrightarrow\min_{\hat{\beta}_0,\,\hat{\beta}_1}.$$

Функция $Q(\hat{\beta}_0,\ \hat{\beta}_1)$ является квадратичной функцией параметров $\hat{\beta}_0$, $\hat{\beta}_1$ (x_i, y_i – известные данные наблюдений). Так как $Q(\hat{\beta}_0,\ \hat{\beta}_1)$ – непрерывна, выпукла и ограничена снизу ($Q(\hat{\beta}_0,\ \hat{\beta}_1)\ge 0$), то она имеет минимум. Необходимым условием существования минимума функции двух переменных является равенство нулю ее частных производных по неизвестным параметрам $\hat{\beta}_0$ и $\hat{\beta}_1$.

$$\begin{cases}\dfrac{\partial Q}{\partial\hat{\beta}_0}=-2\sum\limits_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)=0,\\ \dfrac{\partial Q}{\partial\hat{\beta}_1}=-2\sum\limits_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)x_i=0.\end{cases}\Rightarrow\begin{cases}n\hat{\beta}_0+\hat{\beta}_1\sum\limits_{i=1}^{n}x_i=\sum\limits_{i=1}^{n}y_i,\\ \hat{\beta}_0\sum\limits_{i=1}^{n}x_i+\hat{\beta}_1\sum\limits_{i=1}^{n}x_i^2=\sum\limits_{i=1}^{n}x_iy_i.\end{cases}$$

Данная система уравнений называется ***системой нормальных уравнений МНК***. Разделив оба уравнения последней системы на n, имеем:

$$\begin{cases}\hat{\beta}_0+\hat{\beta}_1\bar{x}=\bar{y},\\ \hat{\beta}_0\bar{x}+\hat{\beta}_1\overline{x^2}=\overline{xy}.\end{cases}$$

Решая систему относительно $\hat{\beta}_0$ и $\hat{\beta}_1$, получим:

$$\begin{cases}\hat{\beta}_1=\dfrac{\overline{xy}-\bar{x}\cdot\bar{y}}{\overline{x^2}-\bar{x}^2}=\dfrac{\operatorname{cov}(x,y)}{\sigma_x^2},\\ \hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x},\end{cases}\qquad(1.5)$$

где $\bar{x}=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i$, $\bar{y}=\dfrac{1}{n}\sum\limits_{i=1}^{n}y_i$, $\overline{xy}=\dfrac{1}{n}\sum\limits_{i=1}^{n}y_ix_i$, $\overline{x^2}=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i^2$.

Таким образом, по МНК оценки параметров $\hat{\beta}_0$ и $\hat{\beta}_1$ определяются по формулам (1.5). Отметим, что, как известно из курса математической статистики, в случае нормального закона распределения случайных величин ε_i, оценки МНК и ММП совпадают.

Зная выборочный коэффициент линейной корреляции $r=r_{xy}=r(x,y)=\dfrac{\overline{xy}-\bar{x}\cdot\bar{y}}{\sigma_x\sigma_y}=\dfrac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y}$, формулу для $\hat{\beta}_1$ можно записать в виде:

$$\hat{\beta}_1=\frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_x}=\frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y}\cdot\frac{\sigma_y}{\sigma_x}=r\frac{\sigma_y}{\sigma_x};\quad\sigma_x=\sqrt{\overline{x^2}-\bar{x}^2},\ \sigma_y=\sqrt{\overline{y^2}-\bar{y}^2}.$$

Коэффициент $\hat{\beta}_1$ при x называется ***эмпирическим (выборочным) коэффициентом регрессии***. Его величина показывает среднее изменение результата с изменением фактора на одну единицу. Знак при коэффициенте регрессии $\hat{\beta}_1$ показывает направление связи: $\hat{\beta}_1>0$ – связь прямая, $\hat{\beta}_1<0$ – связь обратная.

Коэффициент $\hat{\beta}_0$ – ***свободный член уравнения регрессии***, указывает на значение результирующего признака при нулевом значении фактора. Это важный показатель для выбора вида уравнения регрессии. Например, если в результате вычислений коэффициент $\hat{\beta}_0$ оказался отрицательным, а экономический смысл задачи диктует положительность или равенство нулю показателя $\hat{\beta}_0$, значит, выбор вида уравнения был неудачен. Например, в регрессионной модели производительности труда о каком производстве может идти речь, если равны нулю производственные площади или число рабочих.

Важной характеристикой качества уравнения парной линейной регрессии служит квадрат коэффициента корреляции, который в данном случае носит название ***коэффициент детерминации*** и обозначается $R^2 = r^2$. Коэффициент детерминации служит для оценки качества подбора линейной функции. Он характеризует долю дисперсии результативного признака y объясняемую построенной регрессией в общей дисперсии. Соответственно величина $1 - R^2$ характеризует долю дисперсии y, вызванную влиянием остальных неучтенных в модели факторов. Например, если $R^2 = 0{,}98$, то построенное уравнение регрессии объясняет 98% дисперсии y, а на долю прочих неучтенных в уравнении факторов приходится лишь 2% дисперсии y.

Пример 1.2. Для данных примера 1.1 рассчитать коэффициент линейной корреляции, оценить параметры линейного уравнения регрессии y на x, рассчитать сумму квадратов остатков, коэффициент детерминации. Сделать выводы.

Построим в Excel расчетную таблицу 1.2.

Таблица 1.2

№	x_i	y_i	x_i^2	$x_i y_i$	y_i^2	$\hat{y}_i$	e_i	e_i^2
1	4,5	68,8	20,25	309,6	4733,44	66,78243	2,017568	4,070582
2	5,9	58,3	34,81	343,97	3398,89	59,39555	-1,09555	1,200227
3	5,7	62,6	32,49	356,82	3918,76	60,45082	2,149182	4,618984
4	7,2	52,1	51,84	375,12	2714,41	52,5363	-0,4363	0,190358
5	6,2	54,5	38,44	337,9	2970,25	57,81265	-3,31265	10,97362
6	6	57,1	36	342,6	3260,41	58,86791	-1,76791	3,125521
7	7,8	51	60,84	397,8	2601	49,37049	1,629506	2,65529
8	7,5	50,7	56,25	380,25	2570,49	50,9534	-0,2534	0,06421
9	8,1	48,6	65,61	393,66	2361,96	47,78759	0,81241	0,660009
10	7,9	49,1	62,41	387,89	2410,81	48,84286	0,257141	0,066121
Сумма	66,8	552,8	458,94	3625,61	30940,42		1,49E-13	27,62492
Среднее	6,68	55,28	45,894	362,561	3094,042			

Тогда коэффициент линейной корреляции равен

$$r = \frac{362{,}56 - 6{,}68 \cdot 55{,}28}{\sqrt{45{,}89 - 6{,}68^2}\sqrt{3094{,}04 - 55{,}28^2}} = -0{,}96,$$

следовательно, можем сделать вывод о существовании весьма тесной обратной линейной зависимости между размером среднемесячной заработной платы и расходами на покупку продовольственных товаров. Оценим параметры линейного уравнения регрессии y на x:

$$\begin{cases} \hat{\beta}_1 = \dfrac{362{,}561 - 6{,}68 \cdot 55{,}28}{45{,}894 - 6{,}68^2} = -5{,}276, \\ \hat{\beta}_0 = 55{,}28 + 5{,}276 \cdot 6{,}68 = 90{,}526. \end{cases}$$

Значит, эмпирическое уравнение регрессии, описывающей зависимость между среднемесячной заработной платой и расходами на покупку продовольственных товаров, имеет вид

$$\hat{y}_i = 90{,}526 - 5{,}276 x_i,$$

отрицательное значение коэффициента $\hat{\beta}_1$ говорит об обратной зависимости признаков. Добавим в таблицу столбец значений $\hat{y}_i$. Рассчитаем остатки $e_i = y_i - \hat{y}_i = y_i - 90{,}526 + 5{,}276 x_i$, столбец остатков занесем в таблицу. Рассчитаем в таблице столбец квадратов остатков и его сумму:

$$\sum_{i=1}^{n} e_i^2 = 27{,}6249.$$

Учитывая данные значения переменной y, значение рассчитанной суммы можно считать небольшим.

Коэффициент детерминации для полученной модели составляет

$$R^2 = (-0{,}96)^2 = 0{,}9276,$$

следовательно, уравнение регрессии описывает $92{,}76\%$ вариации зависимой переменной, что говорит о хорошем качестве модели.