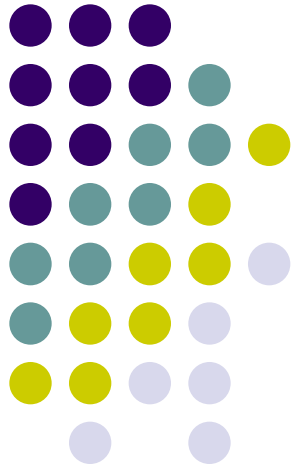
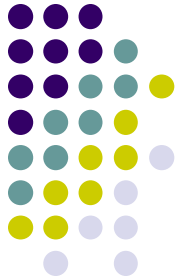
Статистический анализ связей



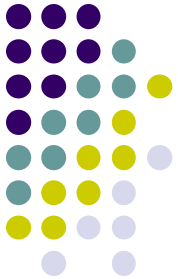
Понятие о статистической и корреляционной связи



Статистика изучает явления через их признаки. Чем более однородна совокупность, тем больше общих признаков имеют ее единицы и тем меньше варьируют их значения.

Признаки различаются **способами их измерения** и другими особенностями, влияющими на **методы статистического изучения**.

Понятие о статистической и корреляционной связи

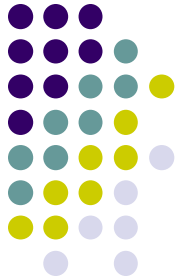


По уровню измерения признаки делятся на **количественные** и **неколичественные**.

Значения **неколичественных** признаков указывают лишь принадлежность единицы к определенной категории.

Из **неколичественных** признаков выделяются **альтернативные** признаки, те, которые могут принимать лишь два значения.

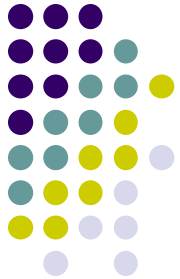
Понятие о статистической и корреляционной связи



Те **неколичественные** признаки, по которым **нельзя. упорядочивать** единицы, называются **номинальными**. Они просто указывают принадлежность единицы к определенной категории.

Те **неколичественные** признаки, по которым **можно упорядочивать** единицы, называются **порядковыми**. Они характеризуют некоторое качество явлений, интенсивность которого выражена по-разному.

Понятие о статистической и корреляционной связи



Для измерения **порядковых** переменных применяется **шкала Ликерта**.

Например, если изучается отношение к труду, то выделяются следующие категории:

+ 1 - Работа нравится.

+ 0,5 - Работа скорее нравится, чем не нравится.

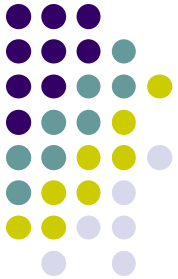
0 - Работа безразлична.

- 0,5 - Работа скорее не нравится, чем нравится.

- 1 - Работа не нравится.

Присваивая цифровые метки категориям, можно ранжировать работников по значениям этих меток.

Понятие о статистической и корреляционной связи



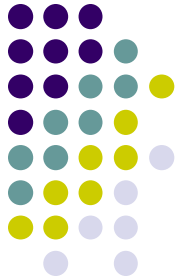
Количественные признаки выражаются числами. Они играют преобладающую роль в экономической статистике.

Количественные признаки могут быть:

- **Дискретные-** это те, значения которых отличаются. не менее чем на единицу измерения признака.
- **Непрерывные** признаки- это те, значения которых у разных единиц могут отличаться на любую сколь угодно малую величину.

Количественные переменные позволяют не только **упорядочивать единицы**, но и определять **интервал**, отделяющий одну единицу от другой.

Понятие о статистической и корреляционной связи

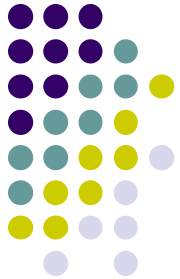


По отнесенности к единице совокупности признаки делятся на **первичные** и **вторичные**.

Первичные признаки характеризуют единицу совокупности в целом. Это **абсолютные** величины. Они могут быть измерены, рассчитаны, взвешены и существуют сами по себе независимо от их статистического изучения.

Вторичные, или **расчетные**, признаки не измеряются непосредственно, а рассчитываются. Они являются продуктами человеческого сознания, результатом познания изучаемого объекта.

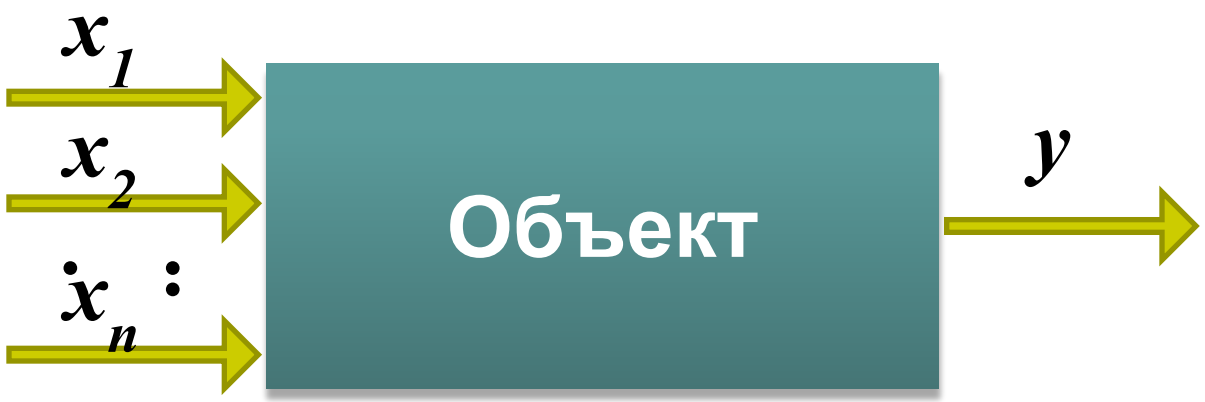
Понятие о статистической и корреляционной связи



Если значение признака зависит от **интервала времени**, к которому он относится, то признак называется **интервальным**. В определение такого признака входит время. Например, продукция, выпущенная за месяц или за год; душевой доход за месяц; расходы на транспорт за неделю, за день и т. д.

Понятие о статистической и корреляционной связи

При **статистической** связи **разным** значениям одной переменной (**фактора**, ***x***) соответствуют разные **распределения** другой переменной (**результата**, ***y***)

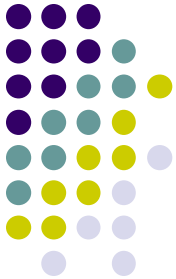


Понятие о статистической и корреляционной связи

Проявление статистической и корреляционной связи

Значения фактора	Количество единиц в группе	Распределение значений результата	Среднее значение результата
x_1	k	$y_{11}, y_{12}, \ldots, y_{1k}$	$\overline{y}_1$
x_2	m	$y_{21}, y_{22}, \ldots, y_{2m}$	$\overline{y}_2$
x_3	p	$y_{31}, y_{32}, \ldots, y_{3p}$	$\overline{y}_3$

Понятие о статистической и корреляционной связи



Корреляционная связь - частный случай **статистической связи**, при котором разным **значениям** переменной соответствуют разные **средние значения** другой переменной.

Корреляционная связь предполагает, что изучаемые переменные имеют **количественное** выражение.

Статистическая связь- более широкое понятие, она не включает ограничений на уровень измерения переменных. Переменные, связь между которыми изучается, могут быть как **количественными**, так и **неколичественными**.

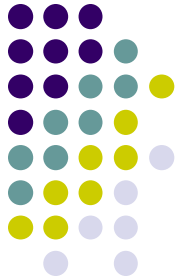
Если изучается связь между двумя признаками, налицо **парная** корреляция. Если изучается связь между многими признаками – **множественная** корреляция.

Понятие о статистической и корреляционной связи

Методы изучения связей

Шкала измерения переменной у	Шкала измерения переменной х		
	Номинальная	Порядковая	Интервальная
Номинальная	Таблицы сопряженности, коэффициенты взаимной сопряженности		
Порядковая		Ранговая корреляция	
Интервальная	Аналитическая группировка, эмпирическое корреляционное отношение		Коэффициенты (индексы) корреляции, уравнения регрессии

Парная корреляция



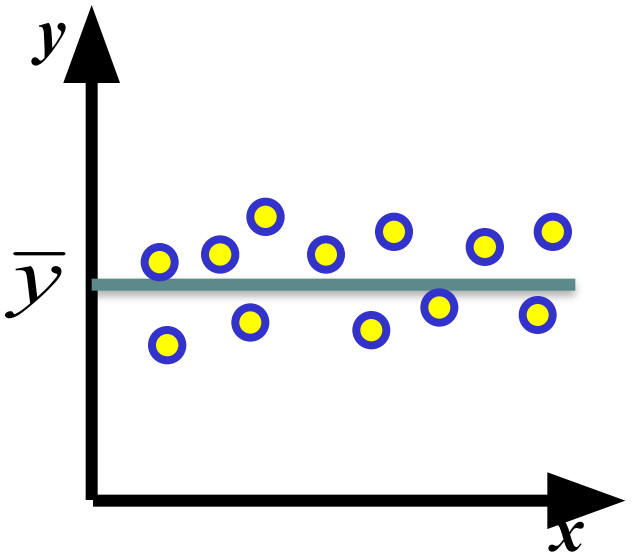
Парная корреляция - это изучение корреляционной связи между **двумя** переменными.

Прежде всего, чтобы проверить, как проявляется связь между двумя переменными, нужно построить график - **поле корреляции**.

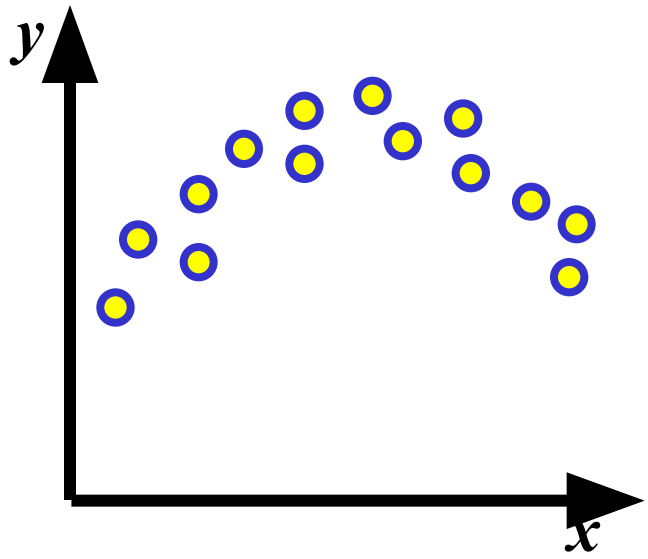
Поле корреляции - это поле точек, на котором каждая точка соответствует единице совокупности; ее координаты определяются значениями признаков x и y.

Парная корреляция

Основные типы корреляции между двумя переменными



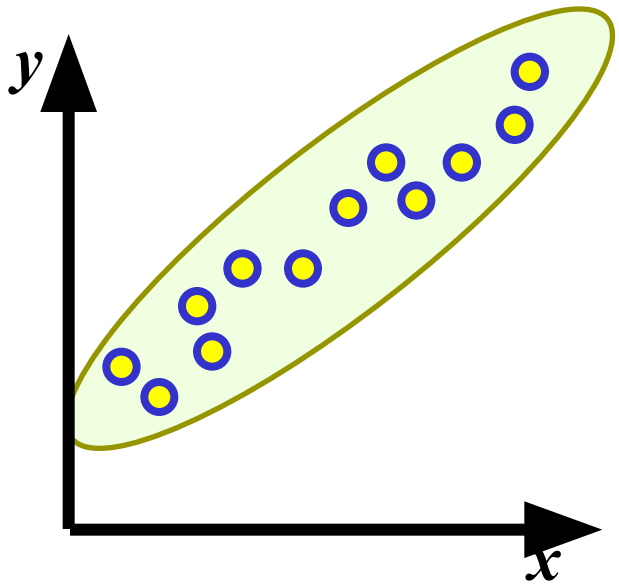
а) связь между x и y отсутствует



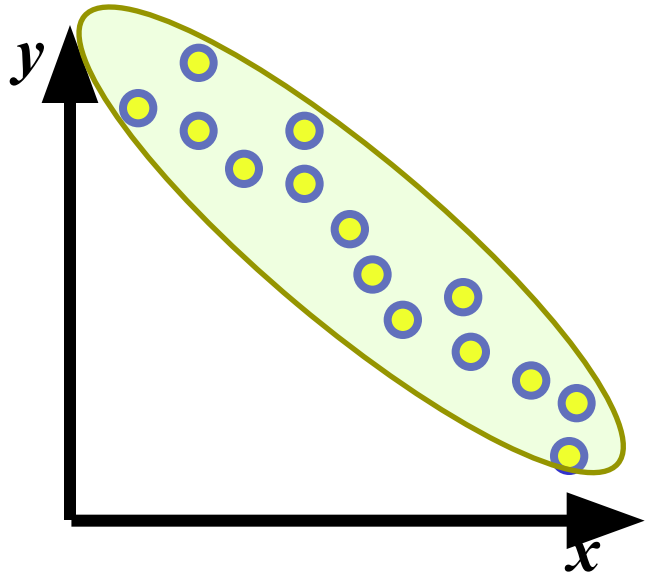
а) связь между x и y нелинейная

Парная корреляция

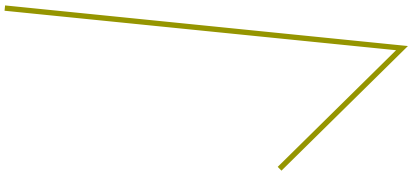
Основные типы корреляции между двумя переменными



а) связь между x и y
линейная прямая



а) связь между x и y
линейная обратная



корреляционный
эллипс

Парная корреляция

Пример. Изучается, зависимость цены товара от дальности его перевозки по 7 фирмам. Данные представлены в табл.

Исходные данные

Номер фирмы	Дальность перевозки, км (x)	Цена товара, руб. (y)
1	10	45
2	17	50
3	15	55
4	25	70
5	19	62
6	20	65
7	8	45
В среднем	16,3	56

Парная корреляция

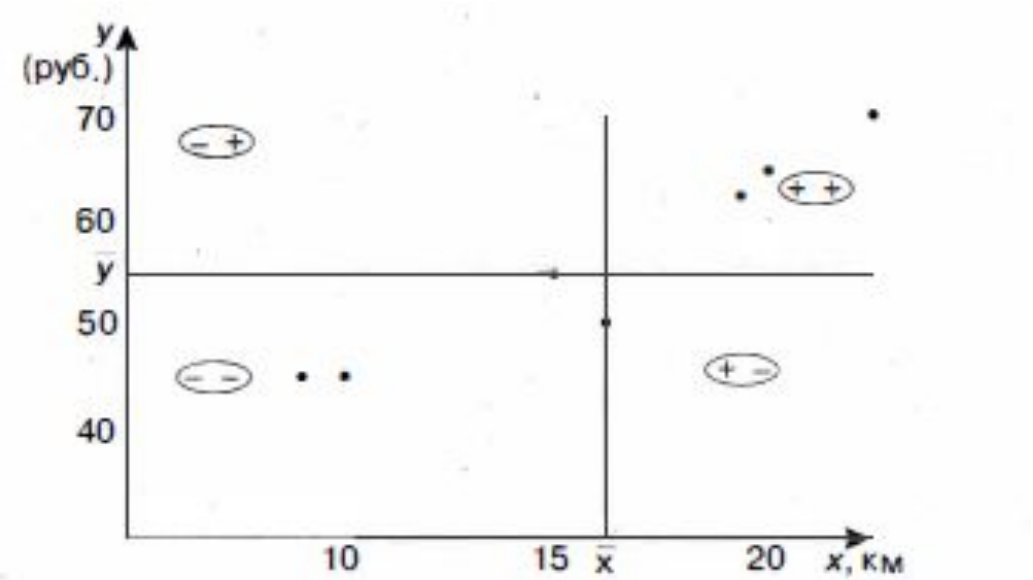


Рис. 7.2. Поле корреляции

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

Парная корреляция

Если знаки отклонений от средних совпадают, то связь прямая ($r_{xy} > 0$), если знаки отклонений не совпадают, то связь обратная ($r_{xy} < 0$). Разделив числитель и знаменатель на n (число наблюдений), получим:

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n\sigma_x\sigma_y}$$

или

$$r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y},$$

Коэффициент парной корреляции измеряется от -1 (случай полной обратной связи) до 1 (случай полной прямой связи). По абсолютной величине: $0 \le |r_{xy}| \le 1$. Чем ближе значение r_{xy} к единице, тем теснее связь, чем ближе значение r_{xy} к нулю, тем слабее связь.

При $|r| < 0{,}30$ связь считается слабой, при $|r| = 0{,}3 - 0{,}7$ – средней, при $|r| > 0{,}7$ – сильной, или тесной.

Коэффициент корреляции – симметричная мера связи, т. е. это мера взаимосвязи между x и y. Поэтому $r_{xy} = r_{yx}$.

Парная корреляция

Расчетная таблица

№ п/п	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	-6,3	-11	69,3	39,69	121
2	0,7	-6	-4,2	0,49	36
3	-1,3	-1	1,3	1,69	1
4	8,7	14	121,8	75,69	196
5	2,7	6	16,2	7,29	36
6	3,7	9	33,3	13,69	81
7	-8,3	-11	91,3	68,89	121
Σ	X	X	329,0	207,43	592

$$r_{xy} = \frac{329}{\sqrt{207{,}43 \cdot 592}} = \frac{329}{350{,}37} = 0{,}939.$$

Полученное значение коэффициента корреляции показывает, что связь между ценой данного товара и дальностью его перевозки является очень тесной.

Парная корреляция

Интерпретация значения коэффициента корреляции зависит и от объема выборки. Существуют таблицы критических значений коэффициентов корреляции для разных объемов выборки (разного количества наблюдений). Так, при 3 наблюдениях можно утверждать наличие корреляционной связи лишь при очень высоких значениях коэффициента корреляции ($r_{yx} \geq 0{,}997$), а при 100 наблюдениях то же утверждение можно делать при $r_{yx} \geq 0{,}19$.

Квадрат коэффициента корреляции представляет собой *коэффициент детерминации*:

Коэффициент детерминации = r^2.

Коэффициент детерминации часто более предпочтителен для измерения связи, так как он может быть использован для измерения не только линейных, но и нелинейных связей. Коэффициент детерминации может быть выражен в процентах. В рассматриваемом примере $r^2 = 0{,}881$, или, иначе говоря, на 88,1% цена данного товара зависит от дальности его транспортировки. Конечно, нужно осторожно относиться к такого рода выводам и иметь в виду, что вряд ли полученное значение отражает в чистом виде зависимость цены от дальности перевозки. Здесь сказывается влияние и тех факторов, которые связаны с расстоянием доставки.

Коэффициент детерминации принимает значения в интервале [0, 1]. Чем ближе значение к 1, тем теснее связь, и наоборот.

Уравнение парной регрессии

Если изучается связь между двумя переменными, причем их можно рассматривать как фактор и результат, т. е. вероятно наличие зависимости, то эту зависимость целесообразно представить в математическом виде. С этой целью подбирают функцию $y = f(x)$, которая наилучшим образом соответствует исходным данным, иначе говоря, обеспечивает наилучшую аппроксимацию поля корреляции. При выборе типа функции руководствуются характером расположения точек на поле корреляции, а также содержанием изучаемой связи. Так, например, при изучении зависимости себестоимости единицы продукции (y) от объема производства (x) теоретический анализ показывает, что такая зависимость должна описываться уравнением гиперболы: $\hat{y} = a + \frac{b}{x}$, поскольку при увеличении объема производства себестоимость снижается до определенного предела, по достижении которого ее дальнейшего снижения не происходит. Однако расположение точек на поле корреляции может показать, что наилучшим образом исходным данным соответствует линейная функция $\hat{y} = a - bx$.

Математически описание зависимости в среднем изменений переменной y от переменной x называется **уравнением парной регрессии**.

Чаще всего используется *линейное уравнение парной регрессии*:

$$\hat{y}_x = a - bx,$$

где $\hat{y}_x$ – среднее значение результативного признака при определенном значении факторного признака x; a – свободный член уравнения регрессии; b – коэффициент регрессии, который показывает, на сколько единиц в среднем изменится результативный признак при изменении факторного признака на одну единицу его измерения.

Уравнение парной регрессии

При такой интерпретации коэффициента регрессии предполагается, что сила воздействия x на y постоянна при любых значениях x.

Знак при коэффициенте регрессии соответствует направлению зависимости y от x:

$b > 0$ – зависимость прямая;

$b < 0$ – зависимость обратная.

Если в исходных данных имеется нулевое значение x, то свободный член a показывает среднее значение y при $x=0$.

Во всех остальных случаях a – доводка, обеспечивающая следующее равенство:

$$\bar{y} = a + b\bar{x}.$$

В этом случае значение a не интерпретируется. Знак при свободном члене a зависит от соотношения между интенсивностью вариации (V) переменных x и y:

если $V_y > V_x$, то $a < 0$;

если $V_y < V_x$, то $a > 0$.

Оценка значимости коэффициентов уравнения регрессии

Оценка практической значимости синтезированной модели

Уравнение зависимости времени поставки от расстояния поставки имеет вид (в скобках указаны стандартные ошибки):

$$\hat{y} = 5{,}91 + 2{,}66 * x$$
$$(0{,}88) \quad (0{,}28)$$

При проверке качества регрессионной модели целесообразно оценить *значимость коэффициентов регрессии.* Эта оценка проводится по t-статистике Стьюдента путём проверки гипотезы о равенстве нулю k-того коэффициента регрессии ($k = 1,2,..., m$). Расчётное значение t-критерия с числом степеней свободы $(n - m - 1 = 10 - 1 - 1 = 8)$ находят путём деления k-того коэффициента регрессии на среднеквадратическое отклонение этого коэффициента (m- количество факторов, n- количество наблюдений). Расчетное значение сравнивается с табличным значением критерия Стьюдента при заданном уровне значимости, и если оно больше табличного значения, коэффициент регрессии считается значимым. В противном случае соответствующий данному коэффициенту регрессии фактор следует исключить из модели, при этом качество модели не ухудшится. По таблице находим $t_{0.05,8} = 2{,}306$. Рассчитанное значение критерия $(9{,}45)$ больше, чем $2{,}306$ (коэффициент регрессии $b = 2{,}66$ превосходит свою случайную ошибку в $9{,}5$ раза). Следовательно, коэффициент корреляции в генеральной совокупности не равен нулю и между временем и расстоянием существует линейная связь. Следовательно, коэффициент регрессии будем считать значимым. Свободный член a также можно считать значимым (см. значения t-статистики при 5% уровне принятия решений)

Оценка адекватности уравнения регрессии

Для оценки адекватности уравнения (модели) применяется F-**критерий Фишера** (F-статистика). Величина F_R сравнивается с критическим значением F_k, который определяется по таблице F- **критерия** с учётом принятого уровня значимости α и числа степеней свободы.

Если F_R больше, чем F-критическое, то взаимосвязь между переменными имеется. Следовательно, полученное регрессионное уравнение имеет практическую значимость.

F- критическое можно получить из таблицы F-критических значений в любом справочнике по математической статистике (Приложение 1).

В нашем примере $F_R = 89{,}91$, число степеней свободы $k_1 = 2-1=1$ и $k_2 = 10-2=8$, табличное значение $F_k = 5{,}32$. Следовательно, величина индекса корреляции признаётся существенной.

Оценка адекватности и качества уравнения регрессии

Индекс детерминации (причинности) R^2 показывает долю изменения (вариации) результативного признака $\hat{y}$ под влиянием факторного признака x. В нашем случае коэффициент детерминации высок $R^2 = 0{,}918 = 91{,}8\%$ и показывает процент общей вариации времени поставки, который зависит от расстояния. Не объясняется лишь $8{,}2\%$ вариации времени поездки. Эта часть вариации обусловлена всеми остальными факторами, влияющими на время поездки, но не включёнными в модель.

Показателем тесноты связи между признаками x и y служит показатель, который называется **индекс корреляции** $R = \sqrt{R^2}$ (в таблице это обозначено как **Множественный** R).

При функциональной (однозначной) связи значения $\hat{y}$ полностью совпадают с соответствующими индивидуальными значениями y. Для получения выводов о практической значимости показателю тесноты связи можно дать качественную оценку. Это осуществляется на основе шкалы Чеддока.

Таблица 1. Шкала Чеддока

Показания тесноты связи	0.1-0.3	0.3-0.5	0.5-0.7	0.7-0.9	0.9-0.99
Характеристика силы связи	слабая	умеренная	заметная	высокая	весьма высокая

Полученная в нашем примере величина $R = 0{,}958$ означает, что в соответствие со шкалой Чеддока установленная по уравнению регрессии связь между временем поставки и расстоянием весьма высокая.