# Эконометрическое моделирование

## *Лабораторная работа № 2*

## Корреляционный анализ

## Оглавление

## Понятие корреляционного и регрессионного анализа

Основным аппаратом эконометрики является раздел математической статистики - корреляционно-регрессионный анализ.

*Задача корреляционного анализа* – выявление характера и степени взаимосвязи между экономическими показателями, являющимися случайными величинами.

*Задача регрессионного анализа* – выявление того, насколько изменение одной экономической переменной (фактора) в среднем влияет на изменение другой экономической переменной (результативного признака).

В корреляционном анализе определяется один показатель, характеризующий степень тесноты взаимосвязи экономических показателей. В регрессионном анализе строится модель регрессии в виде математической функции, которая показывает влияние факторов на некоторый экономический показатель.

Теоретически корреляция и регрессия связаны между собой.

Рассмотрим виды зависимости между случайными величинами.

Пусть имеется двумерная (многомерная) случайная величина, например Х, У. Зависимость между случайными величинами может быть следующих видов:

*функциональная* – если значению случайной величины Х по определенному закону ставится в соответствие значение случайной величины У.

*статистическая* (вероятностная) – если значению случайной величины Х ставится в соответствие определенное распределение случайной величины У.

Математическое ожидание У, определенное для каждого значения Х в вероятностной зависимости, называется условным математическим ожиданием.

Статистическая зависимость, в которой при изменении случайной величины Х изменяется условное математическое ожидание случайной величины У, называется *корреляционной зависимостью*. При этом, если условное математическое ожидание меняется по линейному закону, корреляционная зависимость называется *линейной*, если по нелинейному закону – *нелинейной*. Если условное математическое ожидание случайной величины У не изменяется при изменении значений Х, то корреляционной зависимости нет (т.е. любая корреляционная зависимость является статистической, но не всякая статистическая зависимость является корреляционной).

Функция, которая описывает закон изменения условного математического ожидания случайной величины У при изменении другой случайной величины Х, называется *функцией регрессии* Y на Х. Если двумерная случайная величина (Х,У) распределена по нормальному закону, то функция регрессии линейная, корреляционная зависимость тоже линейная.

Из вышесказанного понятна практическая интерпретация корреляционного и регрессионного анализа – выводы исследования можно интерпретировать лишь как некоторые усредненные свойства, обобщенные характеристики экономических процессов. Это необходимо иметь в виду тем, кто применяет корреляционный и регрессионный анализ.

Корреляционный и регрессионный анализ являются методами математической статистики – дисциплины, целью которой является получение научных и практических выводов относительно некоторых явлений в целом на основе исследования корреляционной и регрессионной зависимости в выборке или временном ряду.

Любое конкретное исследование имеет дело с некоторыми выборочными данными из некоторой генеральной совокупности. Задача исследования состоит в том, чтобы на основе анализа выборки сделать выводы о свойствах всей совокупности. Эта задача решается с помощью специальных приемов математической статистики – проверки статистических гипотез.

Принимается с определенной вероятностью гипотеза о наличии (отсутствии) данного свойства в генеральной совокупности. Эта гипотеза проверяется с помощью специального показателя (статистического критерия), который позволяет с заданной вероятностью ошибки сделать вывод о том, значимы ли значения выборочных показателей для выводов о наличии определенных свойств в генеральной совокупности или их значения случайно отличны от нуля и они не значимы. Для проверки статистической значимости различных видов выборочных статистических показателей существуют тесты, использующие свои критерии, каждый из которых является случайной величиной, имеющей то или иное известное вероятностное распределение, зависящее от степеней свободы и выбранного уровня значимости (доверительной вероятности).

## Парный корреляционный анализ. Коэффициент корреляции

*Корреляционный анализ* – метод математической статистики, используемый для изучения, исследования взаимосвязи между (генеральными) экономическими показателями на основе их наблюдаемых статистических (выборочных) аналогов.

*Парный корреляционный анализ* – изучение взаимосвязи между двумя экономическими показателями, описывающими свойства однотипных объектов из некоторой совокупности.

*Шаги корреляционного анализа*

- Постановка задачи.
- Сбор и анализ данных; определение формы корреляционной связи (линейная, криволинейная).
- Вычисление показателя тесноты корреляционной связи.
- Оценка статистической значимости показателя тесноты корреляционной связи.

*Сбор данных*

Сбор данных осуществляется методом случайной выборки некоторого количества наблюдаемых объектов из некоторой однородной совокупности, фиксации для каждого выбранного объекта пары признаков (свойств), взаимосвязь которых будет предметом исследования (Рисунок 1).

*Визуальный анализ данных.*

Визуальная оценка осуществляется на основе графического анализа. Данные на графике можно представить в виде точечной диаграммы в MS Excel (). В результате такой оценки может быть сделана гипотеза о наличии линейной корреляционной связи, о нелинейной корреляционной связи или об отсутствии корреляционной связи.

| | A | B | C | D |
|---|---|---|---|---|
| 1 | № | Y | X | |
| 2 | 1 | 1,5 | 6,5 | |
| 3 | 2 | 3,5 | 6 | |
| 4 | 3 | 0,5 | 6,5 | |
| 5 | 4 | 1,5 | 4,5 | |
| 6 | 5 | 2 | 8 | |
| 7 | 6 | 1,5 | 5,5 | |
| 8 | 7 | 4,5 | 6,5 | |
| 9 | 8 | 3 | 6 | |
| 10 | 9 | 3,5 | 8 | |
| 11 | 10 | 3 | 4,5 | |
| 12 | | | | |

Рисунок 1 – Данные

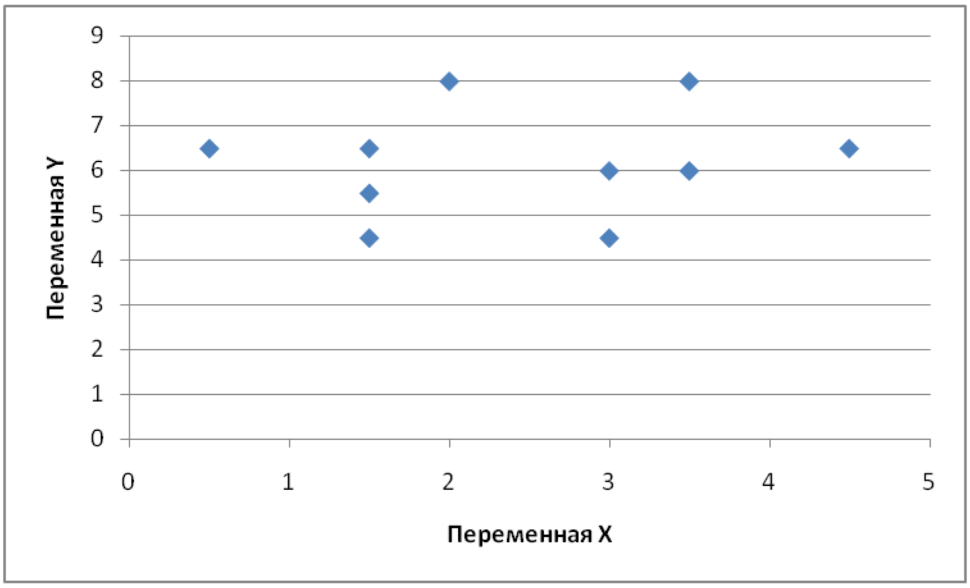


Рисунок 2 – Точечная диаграмма в MS Excel

*Вычисление показателей тесноты корреляционной связи*

Если визуальный анализ позволяет принять гипотезу о линейной форме связи между показателями – для оценки степени тесноты связи применяется *линейный коэффициент корреляции*.

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\left(\overline{x^2} - \bar{x}^2\right)\left(\overline{y^2} - \bar{y}^2\right)}}$$

Границы измерения: $-1 \le r \le 1$.

Если взаимосвязь между показателями обратная (отрицательная) то корреляционная связь отрицательная: $-1 < r < 0$.

Если взаимосвязь между показателями прямая, то корреляционная зависимость положительная: $0 < r < 1$.

Если r = 0, линейная корреляционная зависимость отсутствует.

В крайних случаях |r| = 1 имеется функциональная линейная зависимость между показателями x и y.

Если визуальный анализ не позволяет принять гипотезу о линейной форме связи, коэффициент корреляции в этом случае применять неправомерно.

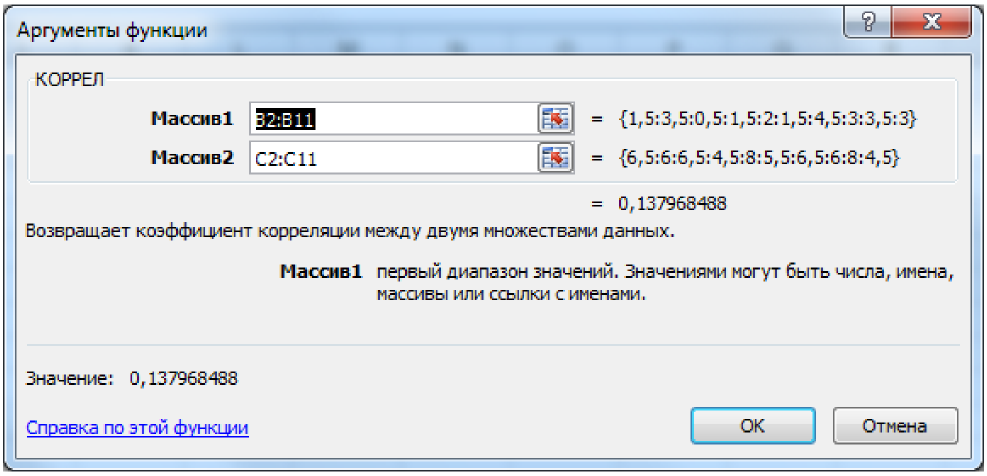


Рисунок 3 – Вычисление коэффициента корреляции в MS Excel

### *Оценка статистической значимости показателя корреляционной связи*

Оценка статистической значимости линейного коэффициента корреляции проводится с помощью теста Стьюдента по t-статистике:

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

Если

$$|t_{\text{набл}}| \geq t_{\text{крит}(\propto, n-2)}$$

то коэффициент корреляции статистически значим.

Таким образом, при определенном уровне значимости проверяется гипотеза о том, что в генеральной совокупности нет корреляционной зависимости между анализируемыми показателями.

В рассматриваемом примере $t_{\text{крит}} = 2{,}31$

$$t_{\text{набл}} = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}138 * \sqrt{10-2}}{\sqrt{1-0{,}138 * 0{,}138}} = 0{,}394$$

поскольку $t_{\text{набл}} < t_{\text{крит}}$, то коэффициент корреляции является незначимым.

## Задание 1. Парный корреляционный анализ

Определите отсутствие или наличие линейной корреляционной взаимосвязи между валовым региональным продуктом и уровнем безработицы. Для анализа используйте данные Федеральной службы государственной статистики.

Уровень безработицы населения по субъектам Российской Федерации, в среднем за год

Валовой региональный продукт по субъектам Российской Федерации в 1998-2012гг.

Таблицы распределений

Для расчетов используйте данные за один год по всем субъектам РФ в соответствии с вариантами:

| Вариант | Год | Вариант | год | Вариант | Год |
|---|---|---|---|---|---|
| 1 | 2000 | 8 | 2007 | 15 | 2001 |
| 2 | 2001 | 9 | 2008 | 16 | 2002 |
| 3 | 2002 | 10 | 2009 | 17 | 2003 |
| 4 | 2003 | 11 | 2010 | 18 | 2004 |
| 5 | 2004 | 12 | 2011 | 19 | 2005 |
| 6 | 2005 | 13 | 2012 | 20 | 2006 |
| 7 | 2006 | 14 | 2000 | 21 | 2007 |

Этапы выполнения задания:

1. Постройте точечную диаграмму и выдвинете гипотезу о характере связи между рассматриваемыми переменными.
2. Рассчитайте коэффициент корреляции.
3. Проверьте значимость коэффициента корреляции.
4. Сделайте выводы.

## Множественный корреляционный анализ

В отличие от парной корреляции и регрессии, где исследовалась рассматривает взаимосвязь двух переменных, в множественном корреляционном и регрессионном анализе рассматриваются взаимосвязи многих показателей, что более соответствует экономическим реалиям, системным взаимосвязям в экономических процессах.

В эконометрических исследованиях множественный корреляционный анализ, оставаясь вспомогательным инструментом, играет важную аналитическую роль при исследовании отдельных специфических проблем эконометрики, в то же время исследователь, применяя множественный корреляционный анализ, может получить дополнительную полезную информацию о взаимосвязях экономических показателей.

В экономике значение изучаемого показателя достаточно часто складывается под влиянием не одного, а многих факторов. Взаимосвязь между множеством экономических показателей изучает множественная корреляция. При этом могут исследоваться две проблемы:

1) влияние на один какой-либо показатель совокупности факторов;

2) анализ взаимосвязи двух каких-либо факторов при исключении влияния на них обоих других факторов.

Множественный корреляционный анализ основывается на парной корреляции. Пусть имеется n экономических показателей. Для проведения множественного корреляционного анализа вычисляются парные коэффициенты корреляции между каждой парой экономических показателей и на их основе составляется корреляционная матрица.

$$K = \begin{bmatrix} 1 & r_{12} \cdots\cdots & r_{1n} \\ r_{21} & 1 \ \cdots\cdots & r_{2n} \\ r_{n1} & r_{n2} \cdots\cdots & 1 \end{bmatrix}$$

### *Коэффициент множественной корреляции*

Для измерения интенсивности совместного влияния всех факторов на изучаемый признак используют коэффициент множественной корреляции, который рассчитывается на основе следующего соотношения:

$$R_{1.23\ldots n} = \sqrt{1 - \frac{D}{D_{11}}}$$

где D – определитель полной матрицы корреляции,

D11 – определитель подматрицы полной матрицы корреляции, содержащий все элементы за исключением элементов первой строки и первого столбца.

Границы изменения коэффициента множественной корреляции от 0 до 1. Чем ближе его значение к 1, тем теснее связь изучаемого признака со всем набором факторов.

Проверка статистической значимости *R* осуществляется на основе критерия Фишера:

$$F_{\text{набл}} = \frac{n-k-1}{k} \cdot \frac{R^2_{1.23\ldots n}}{1 - R^2_{1.23\ldots n}}$$

где $R^2_{1.23\ldots n}$ – квадрат множественного коэффициента корреляции,

n – число наблюдений,

k – число факторов, влияние которых изучается.

$F_{\text{набл}}$ сравнивают с $F_{\text{критич}}$ для уровня значимости α и числа степеней свободы $m_1 = k$ и $m_2 = n - k - 1$.

Если $F_{\text{набл}} > F_{\text{критич}}$, то $R^2_{1.23\ldots n}$ статистически значим. В противном случае ($F_{\text{набл}} \le F_{\text{критич}}$) – статистически незначим.

### *Рассмотрим пример.*

Пусть имеется 4 экономических показателя x, y, z, v. Известна матрица парных коэффициентов корреляции между ними, рассчитанных по выборке, состоящей из 20 значений.

| | y | x | z | v |
|---|---|---|---|---|
| y | 1 | 0,69 | 0,58 | 0,55 |
| x | 0,69 | 1 | 0,46 | 0,50 |
| z | 0,58 | 0,46 | 1 | 0,41 |
| v | 0,55 | 0,50 | 0,41 | 1 |

Необходимо определить совместное влияние на переменную y всех остальных факторов. Рассчитаем множественный коэффициент корреляции.

Найдем определить полной матрицы корреляции в MS Excel:

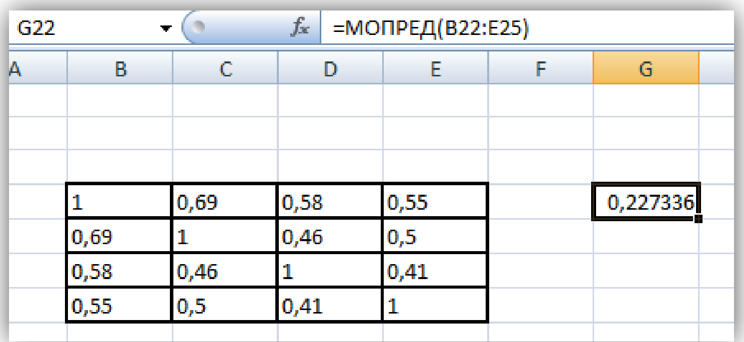


Рисунок 4 – Расчет определителя матрицы

Аналогично найдем определитель подматрицы:

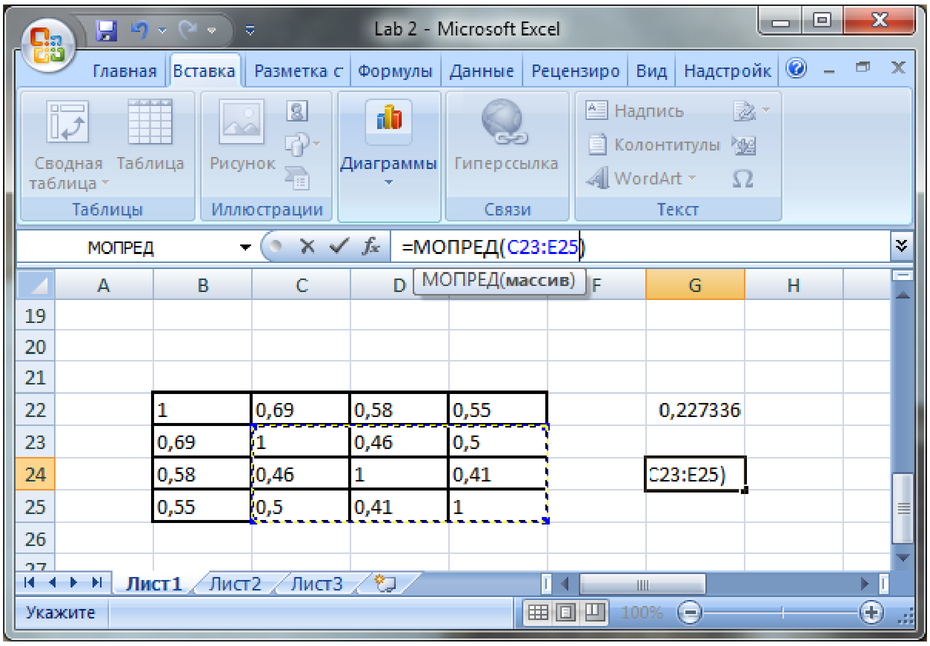


Рисунок 5 – Расчет определителя подматрицы

Рассчитаем коэффициент множественной корреляции:

$$R_{1.23\dots n} = \sqrt{1 - \frac{D}{D_{11}}} = \sqrt{1 - \frac{0{,}227336}{0{,}5589}} \approx 0{,}77$$

Проверим статистическую значимость полученного коэффициента:

$$F_{\text{набл}} = \frac{n-k-1}{k} \cdot \frac{R^2_{1.23\dots n}}{1 - R^2_{1.23\dots n}} = \frac{20-3-1}{3} \cdot \frac{0{,}77^2}{1 - 0{,}77^2} = 7{,}579$$

По таблице значений F критерия Фишера для уровня значимости α=0,01 и числа степеней свободы $m_1$=3 и $m_2$=16 находим $F_{\text{крит}}$ = 5,29. Таким образом $F_{\text{набл}} > F_{\text{крит}}$, следовательно полученный коэффициент множественной корреляции является статистически значимым.

## Задание 2. Расчет множественного коэффициента корреляции

Имеются данные о пяти экономических показателях: инфляция, денежная масса, процентная ставка по кредитам, индекс ММВБ, цена на нефть.

Данные.

Определите совместное влияние четырех оставшихся показателей на показатель, соответствующий вашему варианту:

| Вариант | Год |
|---|---|
| 1, 6, 11, 16, 21 | Инфляция |
| 2, 7, 12, 17 | Денежная масса |
| 3, 8, 13, 18 | Процентная ставка |
| 4, 9, 14, 19 | Индекс ММВБ |
| 5, 10, 15, 20 | Цена на нефть |

Для получения парных коэффициентов корреляции можно использовать MS Excel/ Анализ данных/ Кореляция.

## Частные коэффициенты корреляции

Кроме коэффициента множественной корреляции весьма полезным в сфере экономических исследований являются частные (чистые, парциальные) коэффициенты корреляции, оценивающие степень связи одного признака с одним фактором при исключении влияния всех прочих факторов.

Частные коэффициенты корреляции позволяют выявить «чистую» зависимость признака от одного из факторов и установить, каково было бы влияние этого фактора на величину признака при условии, что влияние других (другого) факторов на этот признак исключается.

Частные коэффициенты могут быть разных порядков. Порядок коэффициента определяется числом факторов, влияние которых исключается.

Так, частные коэффициенты корреляции $r_{yz\bullet z}$, $r_{yz\bullet x}$ – это коэффициенты первого порядка; $r_{yz\bullet zv}$, $r_{yz\bullet vx}$ – это коэффициенты второго порядка и т.д.

Расчет частных коэффициентов корреляции может осуществляться на основе алгебраических дополнений:

$$r_{12\cdot 34\ldots n} = -\frac{D_{12}}{\sqrt{D_{11}\cdot D_{22}}}$$

где $D_{11}$, $D_{12}$, $D_{22}$ – алгебраические дополнения соответственно к элементам $r_{11}$, $r_{12}$, $r_{22}$ корреляционной матрицы.

Пределы изменения частных коэффициентов корреляции $-1 < r_{част} < 1$ и их интерпретаций такая же, как и у парных коэффициентов корреляции.

Проверка статистической значимости частных коэффициентов корреляции осуществляется на основе t - критерия:

$$t_{набл} = \frac{r_{част}\cdot\sqrt{n-2-l}}{\sqrt{1-r_{част}^2}}$$

где $(n - 2 - l)$ – число степеней свободы,

$l$ – число факторов, влияние которых исключается,

$n$ – объем выборки.

Если $r_{набл} > r_{критич}$ для уровня значимости α и числа степеней свободы $(n - 2 - l)$, то $r_{част}$ статистически значим. В противном случае ($r_{набл} \leq r_{критич}$) – статистически незначим.

Для выявления «чистой» зависимости между экономическими показателями и влияния на неё исключаемых факторов необходимо сравнить частные и соответствующие парные коэффициенты корреляции.

Расчет парных и частных коэффициентов корреляции и их последующее сравнение может привести к одному из следующих выводов:

1) $r_{yx} > r_{yx\bullet z\ldots v}$ – факторы z, …, v искажают взаимосвязь между Y и X в сторону ее увеличения;
2) $r_{yx} < r_{yx\bullet z\ldots v}$ – факторы z, …, v искажают взаимосвязь между Y и X в сторону ее уменьшения;
3) $r_{yx} \approx r_{yx\bullet z\ldots v}$ – факторы z, …, v практически не искажают взаимосвязь между Y и X.

*Рассмотрим пример.*

Пусть имеется 4 экономических показателя x, y, z, v. Известна матрица парных коэффициентов корреляции между ними, рассчитанных по выборке, состоящей из 20 значений.

| | y | x | z | v |
|---|---|---|---|---|
| y | 1 | 0,69 | 0,58 | 0,55 |
| x | 0,69 | 1 | 0,46 | 0,50 |
| z | 0,58 | 0,46 | 1 | 0,41 |
| v | 0,55 | 0,50 | 0,41 | 1 |

Необходимо вычислить частный коэффициент корреляции второго порядка $r_{yx\bullet xv}$.

Найдем алгебраическое дополнение $D_{yx}$:

$$D_{yx} = (-1)^{1+2}\begin{vmatrix} 0{,}69 & 0{,}46 & 0{,}50 \\ 0{,}58 & 1 & 0{,}41 \\ 0{,}55 & 0{,}41 & 1 \end{vmatrix} = -0{,}026$$

Расчеты определителей можно осуществить с помощью MS Excel:

| | | | | |
|---|---|---|---|---|
| Dyx | 0,69 | 0,46 | 0,5 | 0,254841 |
| | 0,58 | 1 | 0,41 | |
| | 0,55 | 0,41 | 1 | |
| | | | | |
| Dxx | 1 | 0,58 | 0,55 | 0,45458 |
| | 0,58 | 1 | 0,41 | |
| | 0,55 | 0,41 | 1 | |
| | | | | |
| Dyy | 1 | 0,46 | 0,5 | 0,5589 |
| | 0,46 | 1 | 0,41 | |
| | 0,5 | 0,41 | 1 | |

Рисунок 6 – Вспомогательные расчеты

$$r_{yx\cdot zv} = -\frac{D_{yx}}{\sqrt{D_{yy}\cdot D_{xx}}} = \frac{-0{,}26}{\sqrt{0{,}45\cdot 0{,}55}} \approx 0{,}505$$

Проверим статистическую значимость этого показателя:

$$t_{\text{набл}} = \frac{r_{\text{част}}\cdot\sqrt{n-2-l}}{\sqrt{1-r_{\text{част}}^2}} = \frac{0{,}505\sqrt{20-2-2}}{\sqrt{1-0{,}505^2}} = 2{,}32$$

$t_{\text{крит}}(0{,}05,\ 20\text{-}2\text{-}2) = 2{,}12$

Таким образом, полученный показатель является значимым на 5% уровне.

Так как $r_{yx} = 0{,}69$, что больше чем $r_{yx\bullet zv} = 0{,}505$, то факторы z и v Y искажают взаимосвязь между Y и X в сторону ее увеличения.

### Задание 3. Расчет частных коэффициентов корреляции

На основе данных из предыдущего задания рассчитайте частные коэффициенты корреляции в соответствии с вариантом:

| Вариант | Частный коэффициент корреляции между |
|---|---|
| 1, 11, 21 | Инфляция и денежная масса |
| 2, 12 | Инфляция и процентная ставка |

| | |
|---|---|
| 3, 13 | Инфляция и Индекс ММВБ |
| 4, 14 | Инфляция и цена нефти |
| 5, 15 | Денежная масса и процентная ставка |
| 6, 16 | Денежная масса и индекс ММВБ |
| 7, 17 | Денежная масса и цена нефти |
| 8, 18 | Процентная ставка и индекс ММВБ |
| 9, 19 | Процентная ставка и цена нефти |
| 10, 20 | Индекс ММВБ и цена нефти |

Сделайте выводы по полученным результатам.