МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

«Воронежский государственный технический университет»

Кафедра управления

ЭКОНОМИКО-СТАТИСТИЧЕСКИЕ МЕТОДЫ (ЭКОНОМЕТРИКА)

МЕТОДИЧЕСКИЕ УКАЗАНИЯ

к выполнению самостоятельных работ
для обучающихся всех направлений, специальностей,
форм обучения

Воронеж 2021

УДК 657 (075.8)
ББК 65.052.9(2)

Составитель ст. преп. Т. А. Свиридова

Экономико-статистические методы (Эконометрика): методические указания к выполнению самостоятельных работ для обучающихся всех направлений, специальностей, форм обучения / ФГБОУ ВО «Воронежский государственный технический университет»; сост.: Т. А. Свиридова – Воронеж: Изд-во ВГТУ, 2021. - 36 с.

Основной целью методических указаний является закрепление теоретических и практических навыков, полученных в результате изучения курса Экономико-статистические методы (раздел Эконометрика).

Предназначены для обучающихся всех направлений, специальностей, форм обучения.

Методические указания подготовлены в электронном виде и содержатся в файле МУ_ЭСМ_СР_2021. pdf.

Табл. 2. Библиогр.: 4 назв.

УДК 657 (075.8)
ББК 65.052.9(2)

Рецензент – П. Н. Курочка, д-р техн. наук, проф. кафедры управления ВГТУ

Издается по решению редакционно-издательского совета
Воронежского государственного технического университета

ОГЛАВЛЕНИЕ

1. ОБЩИЕ МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ИЗУЧЕНИЮ КУРСА

Дисциплина «Экономико-статистические методы» призвана сформировать широкий мировоззренческий горизонт будущего специалиста, а также закрепить и расширить методологическую и теоретическую базу, полученную по другим предметам.

Результатом освоения дисциплины в рамках представленного направления и профиля подготовки является освоение соответствующих компетенций согласно учебному плану.

Основными разделами изучаемой дисциплины «Экономико-статистические методы» являются:

1. Предмет эконометрики. Парная регрессия и корреляция в эконометрических исследованиях
2. Множественная регрессия и корреляция
3. Связь между атрибутивными признаками знаками. Моделирование временных рядов.

2. Основы теоретической базы в рамках изучения основных разделов дисциплины «Экономико-статистические методы» раздел эконометрика. Задачи для самостоятельных работ в рамках соответствующих разделов

2.1. Предмет эконометрики. Парная регрессия и корреляция в эконометрических исследованиях

Предметом исследования эконометрики считаются массовые экономические процессы и явления. Эконометрика является наукой, которая эмпирически связана с выводом экономических законов, при этом используются данные или «наблюдения» для получения количественных зависимостей экономических соотношений. Экономические явления как предмет эконометрики (в отличие от экономической теории) рассматриваются в большей мере в количественном аспекте. Например, спрос на продукцию с ростом цен падает.

Методы теории корреляции позволяют определить количественную зависимость между различными техническими, технологическими, организационными и другими факторами, т.е. строить экономико-статистические модели.

Различают функциональную и корреляционную зависимости. Под функциональной понимается такая зависимость, когда с изменением одного фактора изменяется другой, одному значению независимого фактора обычно соответствует только одно значение зависимого фактора. Корреляционная зависимость - это такая зависимость, при которой изменение одной случайной величины вызывает изменение среднего значения другой. Конкретных же значений зависимого переменного, соответствующих одному значению

независимого, может быть несколько. Корреляционные зависимости могут быть установлены только при обработке большого количества наблюдений. При корреляционном анализе решаются следующие задачи:

* устанавливается наличие корреляции (связи) между величинами;
* устанавливается формула линии связи (линии регрессии);
* определяются параметры линии регрессии;
* определяются значимость установленной зависимости и достоверность отдельных параметров.

Наличие корреляции приближенно может быть определено путем визуального анализа поля корреляции. Корреляционным полем называют нанесенные на график в определенном масштабе точки, соответствующие одновременным значениям двух величин.

Тесноту связи между двумя величинами можно определить визуально по соотношению короткой и продольной осей эллипса рассеяния наблюдений, нанесенных на поле корреляции. Чем больше отношение продольной стороны к короткой, тем связь теснее.

Более точно теснота связи характеризуется коэффициентом корреляции r. Коэффициент корреляции лежит в пределах $0 \le |r| \le 1$. В случае , если r=0, то линейной связи нет. Если |r|=1, то между двумя величинами существует функциональная связь. При положительном r наблюдается прямая связь, т.е. с увеличением независимого переменного увеличивается зависимое. При отрицательном коэффициенте наблюдается обратная связь - с увеличением независимого переменного зависимое переменное уменьшается.

Коэффициент корреляции определяется по формуле

$$r = \frac{N\sum_{i=1}^{N} x_i \cdot y_i - \sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} y_i}{\sqrt{N \cdot \sum_{i=1}^{N} x_i{}^2 - (\sum_{i=1}^{N} x_i)^2} \cdot \cdot \sqrt{N \cdot \sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2}},$$

где x и y -текущие значения наблюдаемых величин; N- число наблюдений.

Для численного выражения параметров линии регрессии , выражающих связь между двумя величинами, обычно применяется метод наименьших квадратов. Сущность этого метода состоит в том, чтовыбирается такая линия, при которой сумма квадратов разностей между фактическими наблюдениями зависимой переменной и расчетными значениями, полученными по регрессионной формуле, минимальна

$$S = \sum (y - \tilde{y})^2 \to \min,$$

где $\tilde{y}$ - расчетное значение зависимого переменного по регрессивной формуле.

Допустим , $\tilde{y}$=a+bx. Тогда

$$S = \sum (y - a - bx)^2 \to \min$$

Возьмем частные производные по аи по bи приравняем их к нулю:

$$\frac{\partial S}{\partial a} = -2\sum (y - a - bx) = 0$$

$$\frac{\partial S}{\partial b} = -2\sum (y - a - bx) = 0$$

Полученную систему обычно преобразуют

$$Na + b\sum x = \sum y$$
$$a\sum x + b\sum x^2 = \sum xy$$

Решив систему относительно b иа, получим формулы

$$b = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} x^2 - (\sum_{i=1}^{N})^2} \quad ,$$

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

При линейной корреляции коэффициент корреляции r является не только критерием тесноты связи, но и критерием точности аппроксимации (подбора формулы, выражающей зависимость).

Далеко не всегда данные статистических наблюдений модно аппроксимировать в виде линейной зависимости. Очень часто оказывается, что линейная зависимость не дает необходимой тесноты связи. В этом случае аппроксимацию проводят с помощью зависимостей, отличных от линейной. С этой целью чаще всего используются

* степенная;
* логарифмическая;
* параболическая;
* многочлен степени n;
* зависимости периодического вида.

Для определения параметров нелинейной зависимости пользуются процедурой метода наименьших квадратов, но предварительно проводят линеаризацию. С этой целью, как правило, производят логарифмирование,

например, зависимость степенного вида представляется аналитической формулой

$$\tilde{y} = ax^b,$$

которая после логарифмирования будет представлена в виде

$$\lg \tilde{y} = \lg a + b \lg x,$$

параметры lga и b находятся по методу наименьших квадратов по формуле

$$\lg a = \frac{\sum \lg y \sum (\lg x)^2 - \sum \lg x \sum \lg y \lg x}{N \sum (\lg x)^2 - (\sum \lg x)^2},$$

$$b = \frac{N \sum \lg x \lg y - \sum \lg x \sum \lg y}{N \sum (\lg x)^2 - (\sum \lg x)^2}.$$

Аналогично можно получить выражение для параметров остальных типов кривых.

Логарифмическая зависимость выражается формулой вида:

$$y = a + b \lg x.$$

Расчетные формулы для определения параметров а и в имеют вид:

$$a = \frac{A}{Д}; b = \frac{B}{Д},$$

где

$$A = \sum y \sum (\lg x)^2 - \sum \lg x \sum y \lg x;$$

$$B = N \sum y \lg x - \sum \lg x \sum y;$$

$$Д = N \sum (\lg x)^2 - (\sum \lg x)^2$$

В виде параболы второго порядка зависимость выражается формулой

$$y = a + bx + cx^2.$$

Если степень независимого переменного равна трем, то эта парабола третьего порядка и т.д. Линейная зависимость также является частным случаем многочлена.

Аппроксимация (определение параметров) параболической кривой осуществляется методом наименьших квадратов. В целевую функцию метода наименьших квадратов $\sum(y-\tilde{y})^2 \to \min$ вместо расчетных значений y подставим правую часть уравнения:

$$S = \sum(y - bx - cx^2) \to \min.$$

Возьмем частные производные от этого выражения по a,bи c:

$$\frac{\partial S}{\partial a} = -2\sum(y - a - bx - cx^2) = 0;$$
$$\frac{\partial S}{\partial b} = -2\sum(y - a - bx - cx^2)x = 0;$$
$$\frac{\partial S}{\partial c} = -2(y - a - bx - cx^2)x^2 = 0.$$

Получим систему нормальных или ортогональных уравнений, которая после несложных преобразований примет вид:

$$Na + b\sum x + c\sum x^2 = \sum y;$$
$$a\sum x + b\sum x^2 + c\sum x^3 = \sum yx;$$
$$a\sum x^2 + b\sum x^3 + c\sum x^4 + \sum yx^2.$$

Решая систему любым известным методом, находим параметры параболы a,b и c.

Параметры параболы можно определить из выражений:

$$\left.\begin{aligned} a &= \frac{Д_a}{Д};\\ b &= \frac{Д_b}{Д};\\ c &= \frac{Д_c}{Д}, \end{aligned}\right\}$$

где Д - главный определитель системы линейных уравнений; $Д_a$ - определитель системы уравнений, в котором столбец коэффициентов при a заменяют столбцом свободных членов; $Д_b$ - определитель системы, в котором столбец коэффициентов при b заменен столбцом свободных членов; $Д_c$ - определитель системы, в котором столбец коэффициентов при «c» заменен столбцом свободных членов.

Определители матрицы можно расписать в виде следующих выражений:

$$Д = N\sum x^2\sum x^4 + \sum x\sum x^3\sum x^2 + \sum x\sum x^2\sum x^3 -$$
$$- \sum x^2\sum x^2\sum x^2 - (\sum x)^2\sum x^4 - N(\sum x^3)^2,$$
$$Д_a = \sum y\sum x^2\sum x^4 + \sum x^3\sum x\sum x^2y + \sum xy\sum x^2\sum x^3 -$$
$$- \sum x^2y(\sum x^2)^2 - \sum xy\sum x\sum x^4 - \sum y(\sum x^3)^2,$$
$$Д_b = N\sum xy\sum x^4 + \sum y\sum x^3\sum x^2 + \sum x\sum x^2\sum x^2y -$$
$$- N\sum x^3\sum x^2y - \sum x^4\sum x\sum y - (\sum x^2)^2\sum xy,$$
$$Д_c = N\sum x^2\sum x^2y + \sum x\sum x^3\sum y + \sum x\sum x^2\sum xy -$$
$$- N\sum x^3\sum xy - \sum x^2y\sum x^2\sum y - (\sum x^2)^2\sum y.$$

Кривые периодического вида могут найти широкое распространение при аппроксимации зависимости многих экономических явлений во времени. Например, такими кривыми выражаются влияния сезонных факторов на организацию строительства и материально-технического обеспечения . Наблюдения времени при этом можно представить в виде равноотстоящих, например, x, выраженных в радианах или градусах. Если взять период времени, равный году, и провести ежемесячные наблюдения какого-либо экономического показателя, то время как аргумент может быть записан в виде

$$\frac{1}{12}2\pi;\ \frac{2}{12}2\pi;\ \frac{3}{12}2\pi;...$$

В течение года можно получить 12 наблюдений экономического показателя $y_1, y_2, ..., y_{12}$. Тогда зависимость величины «y» от времени можно выразить уравнением

$$\tilde{y} = a_0 + \sum_{k=1}^{m}(a_k \cos kx + b_k \sin kx),$$

где k=1,2,3...,m - заданное число гармоник многочлена;
a_0, a_k, и b_k - коэффициенты линии регрессии, общим числом 2m+1.

Если N>2m+1, то коэффициенты a_k и b_k находятся по методу наименьших квадратов, который дает следующие выражения для коэффициентов линии регрессии:

$$\left.\begin{aligned} a_0 &= \frac{1}{N}\sum_{i=1}^{N} y_i, \\ a_k &= \frac{2}{N}\sum_{i=1}^{N} y_i \cos kx_i, \\ b_k &= \frac{2}{N}\sum_{i=1}^{N} y_i \sin kx_i. \end{aligned}\right\}$$

Оценка точности аппроксимации криволинейной зависимостью производится при помощи корреляционного отношения:

$$\eta = \sqrt{1 - \frac{\sum (y - \tilde{y})^2}{\sum (y - \bar{y})^2}}$$

Корреляционное значение всегда $0 \le \eta \le 1$, оно всегда положительно. Если $\eta > r$, то кривая точнее аппроксимирует зависимость, чем прямая; для прямой $r = \eta$.

Дополнительной оценкой точности аппроксимации, часто применяемой при оценке нелинейной корреляции, является средняя относительная ошибка аппроксимации, которая определяется по формуле

$$\bar{\varepsilon} = \frac{1}{N}\sum \left| \frac{y - \tilde{y}}{y} \right| \cdot 100 \ .$$

Коэффициент корреляции r, рассчитанный по выборочным данным, может не совпасть с истинным коэффициентом корреляции, соответствующим генеральной совокупности ρ .

Среднеквадратичное отклонение приближенно определяется по формуле

$$\sigma_r \cong \frac{1 - r^2}{\sqrt{N - 1}} \ .$$

При больших выборках можно предположить, что коэффициент корреляции распределен по нормальному закону, тогда можно утверждать, что

$$P\left\{ r - x_p \sigma_r \le \rho \le r + x_p \sigma_r \right\} = \Phi(x)$$

Особенно интересна проверка так называемой нулевой гипотезы. Известно, что если коэффициент корреляции по модулю больше 0, то между двумя случайными величинами имеется связь. Однако r , определенный по частичной выборке, отличается от истинного коэффициента корреляции ρ .

Может быть, что при |r|>0, $\rho = 0$, тогда связь, установленная по частичной выборке, в генеральной совокупности, отсутствует. Для ответа на вопрос , есть ли связь в генеральной совокупности, осуществляют проверку на значимость (коэффициент корреляции существенно отличен от нуля) коэффициента корреляции. Если

$$|r| > \frac{x_p}{\sqrt{N-1}}$$

то с заданной вероятностью P можно утверждать, что коэффициент r существенно отличен от нуля (это означает, что нулевая гипотеза отвергается) и рассматриваемая связь в генеральной совокупности существует. Однако при малых объемах статистического материала гипотеза о нормальном распределении коэффициента корреляции , как правило, не подтверждается. При небольшом числе испытаний вопрос о значимости коэффициента корреляции рассматривается с использованием t-критерия Стьюдента. При этом определяется расчетное значение критерия по формуле

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{N-2}$$

, где N-2 - число степеней свободыf.

Теоретическое значениt определяется по таблице распределения Стьюдента (прил. II). Если $t_{рас.} > t_{табл.}$ при заданном уровне значимости, то предположение о нулевом значении коэффициента корреляции в генеральной совокупности не подтверждается.

При аппроксимации корреляционной зависимости полученная линия регрессии отвечает только частичной выборке, то есть тем данным, которые были использованы при статистической обработке. Для распространения этой зависимости на генеральную совокупность необходимо оценить значение коэффициентов регрессии, так как может оказаться, что при условии неравенства коэффициента регрессии нулю истинный коэффициент регрессии, отражающий генеральную совокупность, будет нулевым. В этом случае прогнозировать по полученной кривой нельзя. Значимость отдельных коэффициентов определяется при помощи t - критерия Стьюдента.

Для оценки значимости коэффициентов регрессии расчетное значение t - критерия Стьюдента определяется по формулам

$$t_{a_i} = \frac{a_i \sqrt{(N-n)\cdot \sum_{i=1}^{N} (x-\bar{x})^2}}{\sqrt{\sum_{i=1}^{N} (y-\bar{y})^2}}$$

$$t_{a_i} = \frac{a_i}{D_{ост} \cdot \sqrt{c_{ii}}},$$

где a_i - коэффициент регрессии при i - м члене уравнения регрессии;

n - число коэффициентов регрессии;

$D_{ост} = \frac{\sum (y-\tilde{y})^2}{N-n-1}$ - остаточная дисперсия;

c_{ii} - диагональный элемент обратной матрицы.

По первой формуле определяется значение t- критерия при одной переменной, а по второй формуле - при множественной корреляции.

При этом если $t_{a_i} \geq t_{табл}$, то коэффициент корреляции существенно отличен от нуля, а следовательно, этот коэффициент имеет значение отличное от нуля и в генеральной совокупности.

Оценка коэффициентов регрессии при помощи t - критерия Стьюдента применяется только для линейных корреляционных связей. Но так как при помощи метода наименьших квадратов путем линеаризации определяются линейные коэффициенты регрессии, то t - критерий может применяться также для любого вида функции в линеаризованной форме.

Значимость уравнения регрессии определяется возможностью надежно прогнозировать средние значения исследуемой величины. Уравнение прогноза получено на основе частичной совокупности, но истинная зависимость, свойственная генеральной совокупности, может существенно отличаться от полученного соотношения. Для изучения степени соответствия полученного уравнения регрессии истинному соотношению следующему из генеральной совокупности используют F - критерий Фишера, определяемый по формуле

$$F = \frac{D_Y}{D_{ОСТ.}}$$

где $D_y = \frac{\sum (y-\bar{y})^2}{N-1}$ -дисперсия фактических значений зависимого переменного; $N-n-1 = f_1$; $N-1 = f_2$ - число степеней свободы.

По числу степеней свободы, задавшись вероятностью, можно определить табличное значение критерия Фишера, значения для которых приведены в прил. III. Если $F \geq F_{табл.}$,то уравнение регрессии считается значимым, т.е. уравнение будет давать достаточно надежные прогнозы и может быть использовано.

Уравнение регрессии из-за вероятностного характера имеет некоторую случайную компоненту h, на величину которой могут отличаться значения зависимой переменной от ее истинных значений

Величина «h» отражает влияние неучтенных факторов и несоответствие частичной совокупности, по которой определялось уравнения регрессии генеральной совокупности. Для надежного прогнозирования необходимо определить доверительный интервал исследуемой величины у.

Если предположить, что величина h является случайной величиной, распределенной по нормальному закону, то истинное значение случайной величины «у»будет находиться в интервале

$$\tilde{y} - x_a \cdot \sigma_h \le y_{ист} \le \tilde{y} + x_a \cdot \sigma_h$$

где $\sigma_h = \sqrt{D^2_{ост}}$;

x_a - аргумент, характеризующий вероятность попадания случайной величины в заданные пределы; определяется из прил.II.

Задачи для самостоятельной работы.

Построить парную корреляционную модель, описывающую зависимость (исходные данные представлены в таблице 1):

1 вариант себестоимости (Y в %) от объема работ, выполненного собственными силами (х в тыс. руб.) в жилищном строительстве.

2 вариант себестоимости (Y в %) от объема работ, выполненного собственными силами (х в тыс. руб.) в промышленном строительстве.

3 вариант себестоимости (Y в %) от удельного веса субподрядных работ (х в %) в жилищном строительстве.

4 вариант себестоимости (Y в %) от удельного веса субподрядных работ (х в %) в промышленном строительстве.

5 вариант себестоимости (Y в %) от возрастания численности ИТР (х в чел. на 100 раб.) в жилищном строительстве.

6 вариант себестоимости (Y в %) от возрастания численности аппарата управления (х в чел. на 100 раб.) в промышленном строительстве.

7 вариант себестоимости (Y в %) от среднего разряда рабочих в жилищном строительстве.

8 вариант себестоимости (Y в %) от среднего разряда рабочих в промышленном строительстве.

9 вариант себестоимости (Y в %) от фондоотдачи машин и оборудования, находящихся на балансе строительной организации в жилищном строительстве.

10 вариант себестоимости (Y в %) от фондоотдачи машин и оборудования, находящихся на балансе строительной организации в промышленном строительстве.

Таблица 1

1в	x	2400	2300	2900	2900	4500	4250	4400	2600	3200	4200
	Y	1,33	1,15	1,1	1,03	1	1,06	0,97	0,99	1	1
2в	x	3600	2200	3200	2400	2700	2700	2800	4200	3400	3600
	Y	0,98	0,96	0,94	1	0,94	0,93	0,96	0,98	0,95	0,94
3в	x	0,25	0,27	0,33	0,39	0,51	0,49	0,54	0,6	0,52	0,7
	Y	1,05	1,1	1,02	1,12	1,19	1	1,05	1,2	1,24	1,3
4в	x	0,43	0,49	0,32	0,55	0,96	0,6	0,52	0,98	0,73	0,56
	Y	0,92	0,94	0,9	0,95	0,92	0,9	0,89	0,97	1,03	0,98
5в	x	11	10	13	12	12,6	14	16,6	14	15	16
	Y	1,04	1,12	1	1,02	1,16	1,08	1,04	1,08	1,05	1,12
6в	x	7	5,8	7,5	7	8	9	10	14	18	14
	Y	2,1	2	0,99	0,98	0,93	0,94	0,95	0,92	0,94	0,95
7в	x	3,25	3,5	2,75	2,85	3,95	3	3,1	3,2	3,3	3,4
	Y	1,11	1,09	1,07	1,06	1,03	1,07	1,01	1,05	1,07	1,05
8в	x	3,5	2,6	3,7	2,8	3,9	4	3,1	3,2	3,3	3,5
	Y	1	0,96	0,92	0,92	0,92	0,93	0,94	0,95	0,94	0,95
9в	x	15	20	30	40	50	56	70	80	90	100
	Y	1,32	1,16	1,08	0,96	0,98	0,98	0,99	1	1,01	1,02
10в	x	10	20	30	40	50	60	70	80	90	100
	Y	1,1	2	0,9	0,85	0,92	0,9	0,89	0,93	0,90	0,95

2.2. Множественная регрессия и корреляция

Большинство закономерностей, изучаемых управлением, имеет достаточно сложный характер и зависит от нескольких факторов, степень влияния каждого на изучаемую величину неизвестна. Первым этапом в исследовании таких зависимостей является создание и анализ парных корреляционных моделей, характеризующих парные зависимости между изучаемой величиной и факторным признаком. Такие модели не отражают всю причинно-следственную связь в изучаемом явлении, но могут дать какие-то первичные сведения о поведении и характере изменения изучаемой величины от каждого из факторов. Но для анализа сложных явлений необходимо изучить его в максимально полной взаимосвязи всех существенных факторных признаков и оценить степень влияния каждого из них на исследуемое явление. Для исследования подобных зависимостей используются модели множественной корреляции.

Для построения множественной корреляционной модели необходимо определить, какая величина будет являться зависимой, т.е. результирующим показателем, а какие величины будут являться независимыми , иначе называемые факторными признаками (факторами). Задача моделирования состоит в выявлении количественной связи между факторами и результирующим показателем.

Фактор, включаемый в модель, должен соответствовать следующим требованиям:

* иметь количественное выражение;

* между фактором и результирующим показателем должна быть логическая, причинная связь;
* между фактором и результирующим показателем должна быть статистическая связь;
* факторы не должны быть тесно связаны между собой, т.е. между факторами не должно быть мультиколлинеарности.

Последнее обстоятельство означает, что парный коэффициент корреляции между двумя факторами не должен быть больше 0,85. Имеются и другие подходы к преодолению явления мультиколлинеарности: можно рекомендовать включать в многофакторные модели те факторы, коэффициент корреляции между которыми не оказался значимым при вероятности 0,9, т.е. для них подтвердилась нулевая гипотеза.

Так же как при парной корреляции, простейшей формой выражения множественной зависимости является линейная зависимость вида:

$$\tilde{y} = a_0 + a_1 x_1 + a_2 x_2 + ... + a_n x_n ,$$

где $\tilde{y}$ - результирующий признак; x_i – факторы; a_0 – свободный член уравнения регрессии; a_i – коэффициенты при факторах. Линейная зависимость является частным случаем параболической зависимости, имеющей выражение:

$$\tilde{y} = a_0 + a_1 x_1 + a_2 x_2 + ... + a_{11} x_1^2 + a_{22} x_1^2 + ... + a_{12} x_1 x_2 + a_{13} x_1 x_3 + ... + a_{1n} x_1 x_n + ..., \quad (4.2)$$

где x_i^2 - квадраты значений факторов; $x_i x_j$ - попарные произведения всевозможных комбинаций факторов; a_{ig} - коэффициент регрессии при попарном произведении факторов (i, j - соответственно номера факторов).

Степенная зависимость $\tilde{y} = A x_1^{b_1} x_{1.}^{b_2} ..... x_n^{b_n} = A \prod_{i=1}^{i=n} x_i^{b_i}$,

где b_i – показатель степени при каждом i-том факторе; заметим, что коэффициенты b_i могут принимать целые и дробные значения с положительным знаком, в последнем случае кривая принимает вид гиперболы; A -числовой коэффициент.

Показательная зависимость $\tilde{y} = a_1^{b_1 x_1} a_{2.}^{b_2 x_2} ..... a_n^{b_n x_n} = \prod_{i=1}^{i=n} a_i^{b_i x_i}$,

где a_i и b_i – коэффициенты регрессии, принимающие любые вещественные значения.

Частным случаем показательной зависимости является экспоненциальная зависимость

$$\tilde{y} = e^{f(x_1,.....x_n)} = \exp\left[f(x_1,....., x_{n)} \right], \quad (4.5)$$

где $f(x_i)$ - любые функции факторов $x_{i,}$ при i=1,2,.....,n.

Благодаря простоте выражения и удобству линеализации экспоненциальные функции получили в последнее время широкое распространение.

Многофакторная зависимость может быть аппроксимирована произведением ряда функций (метод Брандона):

$$\tilde{y} = cf_1(x_1)f_2(x_2).....f_n(x_{n)=}c\prod_{i=1}^{i=n} f_i(x_1),$$

, где $f_i(x_i)$ - некоторая функция фактора $x_{i,}$ c- свободный член формулы.

Ниже будут рассмотрены методы аппроксимации многофакторных зависимостей с применением перечисленных моделей.

Аппроксимация заключается в подборе коэффициентов регрессии при заранее заданной форме линии регрессии . Наиболее распространенным способом аппроксимации является метод наименьших квадратов. Необходимо найти минимум выражения

$$S = \sum (y - \tilde{y})^2.$$

При линейной многофакторной зависимости это выражение запишется в виде

$$S = \sum (y - a_0 - a_1x_1 - a_2x_2 - .....- a_nx_n)^2 \to \min.$$

Для минимизации выражения необходимо определить частные производные по каждому неизвестному. Напомним, что в качестве неизвестных здесь рассматриваются параметры регрессионной формулы $a_0\, a_{1,} a_{2,} .....a_n$.

Частные производные приравниваются к нулю, и составляется система нормальных линейных уравнений, число которых на единицу больше числа факторов, включаемых в модель. Система нормальных уравнений имеет вид:

$$\left.\begin{array}{l} Na_0 + a_1\sum x_1 + a_2\sum x_2 + ... + a_n\sum x_n = \sum y, \\ a_0\sum x_1 + a_1\sum x_1^2 + a_2\sum x_1x_2 + ... + a_n\sum x_1x_n = \sum x_1y, \\ a_0\sum x_2 + a_1\sum x_1x_2 + a_2\sum x_2^2 + ... + a_n\sum x_2x_n = \sum x_2y, \\ ..... \\ a_0\sum x_n + a_1\sum x_1x_n + a_2\sum x_2x_n + ... + a_n\sum x_n^2 = \sum x_ny. \end{array}\right\}$$

Представленную выше систему линейных алгебраических уравнений относительно неизвестных коэффициентов уравнения регрессии a_i удобнее

всего решать матричным способом. В этом случае система уравнений запишется в виде матричного уравнения

$$(\mathbf{X}^*\mathbf{X})\mathbf{A} = \mathbf{X}^*\mathbf{Y},$$

где X - матрица исходных статистических данных по независимым переменным - факторам;

$\mathbf{X}^*$ - транспонированная матрица исходных данных;

Y -матрица - столбец значений зависимой переменной - отклика;

A -матрица - столбец искомых коэффициентов уравнения регрессии.

Эти матрицы представляются в следующем виде:

$$A = \begin{Vmatrix} a_0 \\ a_1 \\ a_2 \\ \dots \\ \ddots \\ \ddots \\ a_n \end{Vmatrix} \qquad X = \begin{Vmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{in} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{Nj} & \dots & x_{Nn} \end{Vmatrix} \qquad Y = \begin{Vmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ \dots \\ \dots \\ y_N \end{Vmatrix}$$

Транспонированная матрица получается из исходной матрицы X путем замены строк на столбцы. В этом случае общий вид транспонированной матрицы имеет следующий вид:

$$X^* = \begin{Vmatrix} x_{11} & x_{21} & \dots & x_{i1} & \dots & x_{N1} \\ x_{12} & x_{22} & \dots & x_{i2} & \dots & x_{N2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{1j} & x_{2j} & \dots & x_{ij} & \dots & x_{Nj} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{1n} & x_{2n} & \dots & x_{in} & \dots & x_{Nn} \end{Vmatrix}$$

Здесь n - число факторных признаков в корреляционной модели;

N - объем статистического материала.

Решение матричного уравнения находится с помощью обращения матриц и записывается в виде

$$\mathbf{A} = (\mathbf{X}^*\mathbf{X})^{-1}(\mathbf{X}^*\mathbf{Y})$$

где $(\mathbf{X}^*\mathbf{X})^{-1}$ - матрица, обратная к $(\mathbf{X}^*\mathbf{X})$.

Найти матрицу A^{-1}, обратную матрице Асоставленную из элементов

$$\begin{Vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{2n} & a_{22} & \dots & a_{n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{Vmatrix},$$

можно следующим образом:

* вычислить определитель этой матрицы Д;
* определить алгебраические дополнения A_{ij} каждого элемента определителя матрицы a_{ij}. Напомним, что алгебраические дополнения определяются вычеркиванием i-ой строки и j-ого столбца;
* составить из алгебраических дополнений матрицу $\widetilde{A}$;
* составить из $\widetilde{A}$ транспонированную матрицу $\widetilde{A}^*$(для чего необходимо поменять строки и столбцы местами);
* составить обратную матрицу $A^{-1} = \widetilde{A}^* / Д$.

Если размерность матрицы $(\mathbf{X}^*\mathbf{X})$ достаточно большая (как правило больше 4), то ее обращение связано с определенными трудностями, объясняющимися большим объемом вычислительной работы, и эта работа проделывается с помощью ЭВМ; в этом случае возможно возникновение проблем, связанных с вычислительной устойчивостью применяемого алгоритма обращения, что порождается плохой обусловленностью исходной матрицы, то есть определитель близок к нулю.

Реальная связь изучаемых факторов, как правило, имеет зависимость, существенно отличную от линейной (зачастую такая зависимость может быть представлена в виде линии только на очень ограниченном интервале изменения факторных признаков), а поэтому для построения адекватных корреляционных моделей требуется использование нелинейных аппроксимаций.

Наиболее универсальной моделью, аппроксимирующей нелинейную связь, является многочлен n - ой степени. Из теории интерполяции известно, что полиномом произвольной степени можно описать практически любую зависимость с любой степенью точности, но такое представление связано с огромным объемом вычислительной работы и вряд ли применимо на практике. Обычно при определении максимально возможной степени полинома ориентируются на объем статистического материала, имеющегося в распоряжении исследователя. При этом следует иметь в виду, что число членов полинома должно быть меньше числа наблюдений в противном случае происходит переход к функциональной зависимости. Для ориентировочных оценок можно использовать соотношение

$$\frac{N}{n+1} > 8.$$

При аппроксимации полиномом n - ой степени уравнение регрессии должно быть предварительно линеаризовано. С этой целью все нелинейные члены, входящие в уравнение регрессии, заменяются новыми, дополнительными факторными признаками. Это приводит к увеличению размерности задачи, но делает доступным аппарат исследования линейных моделей.

Алгоритм построения многофакторной корреляционной модели заключается в следующем:

* задать максимально возможно высокую степень полинома аппроксимации, записать уравнение регрессии в общем виде;
* осуществить линеаризацию исходного уравнения регрессии и записать систему разрешающих линейных алгебраических уравнений в матричной форме;
* найти коэффициенты уравнения регрессии;
* проверить значимость уравнения регрессии по критерию Фишера; если уравнение регрессии оказалось незначимым, то требуется увеличить степень многочлена и повторить все вычисления сначала;
* проверить значимость коэффициентов уравнения регрессии с помощью t - критерия Стьюдента, отбрасывая те из них, которые незначимы, и вновь повторяя построение модели уже с меньшим количеством факторов.

Построение многофакторных моделей с использованием нелинейных форм аппроксимации, как правило, связано с большим объемом вычислений и осуществляется с помощью ЭВМ. В этой связи очень важен вопрос о снижении объемов вычислительной работы, что может быть достигнуто за счет применения нетрадиционных форм аппроксимации. Одним из таких подходов является метод Брандона, когда уравнение регрессии задается в виде произведения, каждый сомножитель которого представлен функцией только одного факторного признака. Таким образом, уравнение принимает вид

$$\tilde{y} = c\prod_{i=1}^{n} f_i(x_i)$$

Здесь $y = f_i(x_i)$ может иметь любой вид;

c - постоянная величина, равная среднему значению $\bar{y}$.

Алгоритм работы по методу Брандона заключается в следующем:

* вычисляется среднее значение y и преобразуются значения для каждого наблюдения по формуле (j - порядковый номер наблюдения)

$$y_{oj} = \frac{y_j}{\bar{y}};$$

* выбирается вид зависимости y_0 от x_1 и по методу наименьших квадратов определяются параметры формулы $\tilde{y}_0 = f_1(x_1)$;

* вычисляются значения функции $f_1(x_1)$ и определяется остаточный показатель y_1 для каждого наблюдения по формуле

$$y_1 = \frac{y_0}{f_1(x_1)}$$

в предположении, что y_1 зависит от $x_{2,}x_{3,\ldots\ldots,}x_n$ и не зависят от x_1;

* определяется корреляционная формула зависимости y_1 от x_2;

* находится условный показатель, не зависящий от x_1 и x_2

$$y_2 = \frac{y_1}{f_2(x_2)}.$$

Такие преобразования результирующего показателя осуществляются до тех пор, пока не будет определена вся последовательность функций, входящих в произведение.

Данный алгоритм дает существенное снижение объемов вычислений и их качественное упрощение: уже не требуется решать алгебраических систем большой размерности.

Количественно тесноту связи при множественной корреляции можно оценить с помощью множественного (совокупного) коэффициента корреляции R. Для расчета совокупного коэффициента корреляции необходимо определить парные коэффициенты корреляции r_{0i} между всеми факторами x_i, входящими в модель, и результирующим показателем y и все парные коэффициенты корреляции между факторами. Все коэффициенты корреляции записываются в квадратную симметричную матрицу

$$\begin{bmatrix} 1 & r_{yx_1} & r_{yx_2} & r_{yx_3} & \ldots & r_{yx_n} \\ r_{yx_1} & 1 & r_{x_1x_2} & r_{x_1x_3} & \ldots & r_{x_1x_n} \\ r_{yx_2} & r_{x_1x_2} & 1 & r_{x_2x_3} & \ldots & r_{x_2x_n} \\ r_{yx_3} & r_{x_1x_3} & r_{x_2x_3} & 1 & \ldots & r_{x_3x_n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{yx_n} & r_{x_{1_1}x_n} & r_{x_2x_n} & r_{x_3x_n} & \ldots & 1 \end{bmatrix} \ldots$$

Множественный коэффициент корреляции определяется по формуле

$$R = \sqrt{1 - \frac{Д}{Д_{11}}}$$

где Д - определитель матрицы парных коэффициентов корреляции;

$Д_{11}$ - определитель той же матрицы с вычеркнутыми первой строкой и первым столбцом, т.е. определитель матрицы парных коэффициентов корреляции между факторами.

Для случая зависимости от двух факторов

$$R_{y/x_1x_2} = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 + 2r_{yx_1}r_{yx_2}s_{x_1x_2}}{1 - r_{x_1x_2}^2}}$$

Для определения влияния только одного i-того фактора на результирующий показатель с исключением влияния других факторов используется частный коэффициент корреляции

$$r_{y/x_ix_1x_2\ldots x_n} = \frac{Д_{1i}}{\sqrt{Д_{11}Д_{ii}}}$$

где $Д_{1i}$ - определитель матрицы с вычеркнутой первой строкой и i-тым столбцом; $Д_{ii}$ - определитель матрицы с вычеркнутой i-той строкой и i-тым столбцом.

При множественной корреляции от двух факторов коэффициент частной корреляции первого фактора

$$r_{y/x_1x_2} = \frac{r_{yx_1} - r_{yx_2}r_{x_1x_2}}{\sqrt{(1-r_{yx_2}^2)(1-r_{x_1x_2}^2)}}$$

а коэффициент частной корреляции для второго фактора

$$r_{y/x_2x_1} = \frac{r_{yx_2} - r_{yx_1}r_{x_1x_2}}{\sqrt{(1-r_{yx_1}^2)(1-r_{x_1x_2}^2)}}$$

Частный коэффициент корреляции отражает « чистое» влияние фактора на результирующий показатель и отличается от коэффициента парной корреляции r_{yx}.

При линейной форме связи множественный коэффициент корреляции является оценкой точности аппроксимации и равен корреляционному отношению η; при нелинейных формах связи для оценки точности аппроксимации (оценки адекватности модели) применяются корреляционное отношение η и ошибка аппроксимации ε. Эти оценки определяются так же, как и при парной корреляции.

Для оценки влияния отдельных факторов на результативный признак используются коэффициенты эластичности, показывающие уровень изменения результативного показателя в случае изменения факторного признака на 1 % при неизменных значениях других факторов. Коэффициенты эластичности определяются по формуле.

$$Э_i = \frac{x_i}{y} \cdot \frac{\partial \tilde{y}}{\partial x_i},$$

где $\tilde{y}$ - теоретическое уравнение регрессии.

Задачи для самостоятельной работы

1 вариант

Построить множественную корреляционную модель описывающую зависимость:

1 вариант себестоимости (Y, %) от объема работ, выполненного собственными силами (x_1, тыс. руб.), и удельного веса субподрядных работ (x_2) в жилищном строительстве;

2 вариант себестоимости (Y , %) от объема работ, выполненного собственными силами (x_1 , тыс. руб.) , и удельного веса субподрядных работ (x_2) в промышленном строительстве;

3 вариант себестоимости (Y , %) от объема работ, выполненного собственными силами (x_1,тыс.руб.), и удельного веса субподрядных работ (x_2 ,%) в жилищном строительстве;

4 вариант себестоимости (Y , %) от объема работ, выполненного собственными силами (x_1, тыс. руб.) , и удельного веса субподрядных работ (x_2) в промышленном строительстве;

5 вариант себестоимости (Y , %) от объема работ, выполненного собственными силами (x_1, тыс. руб.) , и возрастания численности ИТР (x_2, чел. на 100 раб.) в жилищном строительстве;

6 вариант зависимость себестоимости (Y , %) от объема работ, выполненного собственными силами (x_1 , тыс. руб.), и возрастания численности аппарата управления (x_2, чел. на 100 раб.) в промышленном строительстве;

7 вариант себестоимости (Y , %) от объема работ, выполненного собственными силами (x_1, тыс. руб.), и среднего разряда рабочих (x_2) в жилищном строительстве;

8 вариант себестоимости (Y , %) от объема работ, выполненного собственными силами (x_1 , тыс. руб.) и среднего разряда рабочих (x_2) в промышленном строительстве;

9 вариант себестоимости (Y , %) от объема работ, выполненного собственными силами (x_1 , тыс. руб.) и от фондоотдачи машин и оборудования, находящихся на балансе строительной организации (x_2) в жилищном строительстве;

10 вариант себестоимости (Y , %) от объема работ, выполненного собственными силами (x_1 , тыс. руб.) и фондоотдачи машин и оборудования, находящихся на балансе строительной организации (x_2) в промышленном строительстве;

11 вариант выработки на одного рабочего (Y , тыс. руб.) от объема работ выполняемого собственными силами (x_1 , тыс. руб.), и удельного веса субподрядных работ (x_2) в жилищном строительстве;

12 вариант выработки на одного рабочего (Y, тыс. руб.) от объема работ выполняемого собственными силами (x_1, тыс. руб.), и удельного веса субподрядных работ (x_2) в промышленном строительстве;

13 вариант выработки на одного рабочего (Y, тыс. руб.) от удельного веса субподрядных работ (x_1) и среднего разряда рабочих (x_2) в жилищном строительстве;

14 вариант выработки на одного рабочего (Y, тыс. руб.) от объема работ выполняемого собственными силами (x_1, тыс. руб.), и численности ИТР на 100 рабочих (x_2) в промышленном строительстве;

15 вариант выработки на одного рабочего (Y, тыс. руб.) от объема работ выполняемого собственными силами (x_1, тыс. руб.), и возрастания численности ИТР (чел. на 100раб. x_2) в жилищном строительстве;

Статистические данные приведены в таблице 2 Использовать линейную аппроксимацию и метод Брандона. Результаты сравнить.

Таблица 2

1в	x1	2100	2300	2700	2900	3000	3250	3400	3600	4100	4300
	x2	0,9	0,6	0,52	0,4	0,41	0,3	0,28	0,25	0,2	0,23
	Y	1,33	1,15	1,1	1,03	2	1,06	0,97	0,99	1	1
2в	x1	2500	2100	2200	2400	2800	2700	2800	3200	3400	2300
	x2	0,35	0,4	0,45	0,8	0,7	0,67	0,72	0,75	0,7	0,65
	Y	0,98	0,96	0,94	1	0,94	0,93	0,96	0,98	0,95	0,94
3в	x1	2750	3100	4700	3200	2250	3300	2850	2200	2100	1800
	x2	0,15	0,25	0,33	0,39	0,47	0,49	0,57	0,6	0,62	0,7
	Y	1,05	1,1	1,02	1,12	1,19	1	1,05	1,2	1,24	1,3
4в	x1	3400	3800	3200	2700	3500	3300	3000	1800	1600	2000
	x2	0,4	0,49	0,51	0,55	0,58	0,6	0,64	0,66	0,73	0,77
	Y	0,92	0,94	0,9	0,95	0,92	0,9	0,89	0,97	1,03	0,98
5в	x1	3000	2500	3200	3250	2300	3700	3100	3600	3000	2400
	x2	9	10	11	12	12,6	13	16,6	14	15	16
	Y	1,04	1,12	1	1,02	1,16	1,08	1,04	1,08	1,05	1,12
6в	x1	1600	1800	2200	4300	2500	2700	3000	3200	4400	3600
	x2	5	5,8	8,5	7	8	9	10	12	13	14
	Y	1,1	1	0,98	0,98	0,96	0,94	0,93	0,92	0,94	0,95
7в	x1	3500	2800	2900	2900	3500	2900	2300	2800	2900	2800
	x2	2,25	2,5	2,75	2,85	2,95	3	3,1	3,2	3,3	3,4

Окончание табл. 2

	Y	1,11	1,09	1,07	1,06	1,03	1,07	1,01	1,05	1,07	1,05
8в	x1	1900	3400	3300	3400	2200	3500	3800	3600	3800	2600
	x2	2,5	2,6	2,7	2,8	2,9	3	3,1	3,2	3,3	3,5
	Y	1	0,96	0,92	0,92	0,92	0,93	0,94	0,95	0,94	0,95
9в	x1	2800	2300	3750	5000	3300	3500	3600	3300	3200	3100
	x2	10	20	30	40	50	60	70	80	90	100
	Y	1,32	1,16	1,08	0,96	0,98	0,98	0,99	1	1,01	1,02
10в	x1	1400	2800	3400	3400	4200	3500	3300	3100	3400	2600
	x2	10	20	30	40	50	60	70	80	90	100
	Y	1,1	1	0,9	0,91	0,92	0,9	0,92	0,93	0,90	0,95
11в	x1	2000	3250	2500	4750	3000	3250	3500	4000	4250	4500
	x2	0,8	0,65	0,68	0,62	0,48	0,25	0,15	0,1	0,17	0,45
	Y	5500	6300	5000	7000	8000	8500	8750	9000	8750	8000
12в	x1	2600	1800	2100	2200	2400	2500	2600	2800	3000	3200
	x2	0,78	0,55	0,7	0,6	0,6	0,35	0,5	0,55	0,6	0,58
	Y	7000	8000	7500	8300	5000	7800	8000	8500	8800	8600
13в	x1	0,15	0,25	0,3	0,35	0,4	0,45	0,5	0,55	0,6	0,65
	x2	2,9	3	3,1	3,2	3,4	3,5	3	3,1	2,5	3,8
	Y	6300	8400	8400	5500	7700	7600	8400	8200	6800	6000
14в	x1	1600	1800	2100	2200	2400	2500	2600	2800	3000	3200
	x2	8	8,5	9,9	9	11,5	12	12,5	7,5	7	15
	Y	7700	7800	7800	8000	7800	7700	7500	7700	7500	7000
15в	x1	2100	2300	2500	2750	3000	3250	3500	4750	3000	4250
	x2	9,5	10	10,2	11	13	14	14,3	14,5	14,2	14
	Y	5500	5000	6500	7000	7700	8300	6800	9000	8900	8700

2.3. Связь между атрибутивными признаками знаками. Моделирование временных рядов

Использование регрессионного и корреляционного анализа требует, чтобы все признаки были количественно измеренными. Методы КРА. основанные на использовании количественных параметров распределения (средние величины, дисперсия), называют параметрическими методами.

Вместе с тем, особенно при проведении социологических исследований, возникает потребность оценки тесноты связи между качественными (атрибутивными) признаками. Проблему оценки тесноты связи между атрибутивными признаками решают непараметрические методы. Сфера их использования значительно шире в сравнивании с параметрическими методами, потому что не требуется использования условия нормального распределения результативной переменной, не ставится задача представления зависимости между атрибутивными признаками соответствующим уравнением. Здесь речь идет только о наличии установлении связи и измерения его тесноты.

Взаимосвязь между атрибутивными признаками анализируются посредством таблиц взаимной сопряженности. Они описывают комбинационные распределения совокупности по факторному признаку х и результативному у. Например, результаты социологического опроса населения относительно намерений принять участие на рынке ценных бумаг: распределение респондентов опроса по возрасту рассматривается как факторный признак х, а их распределение по склонности к риску как

результативный признак у. Таблицы взаимной сопряженности могут иметь различную размерность. Простейшая размерность – 2х2 (таблица «четырех полей»), когда по альтернативному признаку («да» – «нет», «хорошо» – «плохо» и т.д.) выделяются 2 группы.

При наличии стохастической связи оценка его тесноты базируется на отклонениях фактических частот f_{ij}от F_{ij} ,пропорциональных итоговым частотам:

$$F_{ij} = \frac{f_{io} f_{oj}}{n}$$

, где f_{io}- суммарные частоты по признаку *x;* f_{0j} - суммарные частоты по признаку *y*; *n* - объем совокупности. Очевидно, что

$$m = \sum_{i=1}^{m_x} f_{io} = \sum_{j=1}^{m_y} f_{jo}$$

, где m_x, m_y – соответственно количество групп по признакам х и у.
Абсолютную величину отклонений фактических частот f_{ij} от пропорциональных $F_{ij}(f_{ij} - F_{ij})$ характеризуют статистическим критерием χ^2 (хи-квадрат).

$$\chi^2 = \sum_{i=2}^{m_x} \sum_{j=1}^{m_y} \frac{(f_{ij} - F_{ij})^2}{F_{ij}} = n\left(\sum_{i=1}^{m_x} \sum_{j=1}^{m_y} \frac{f_{ij}^2}{f_{io} f_{oj}} - 1 \right)$$

Из-за отсутствия стохастической связи $\chi^2 = 0$. Для вывода о тесноте связи теоретическое значение χ^2 сравнивается с табличным $\chi^2_{\text{табл.}}$Последний выбирается из справочных математических таблиц критерия «хи»-квадрат в зависимости от принятого уровня значимости α(0,01 или 0,05), и степеней свободы к = ($т_х$ - 1)($т_у$-1). При $\chi^2 > \chi^2_{табл}$делают вывод о наличии тесной связи между признаками х иу.

Однако выводы о зависимости, сделанные «на глаз», часто могут быть ненадежными (ошибочными), поэтому они должны подкрепляться определенными статистическими критериями, например критерием Пирсона χ^2. Он позволяет судить о случайности (или неслучайности) распределения в таблицах взаимной сопряженности, а, следовательно, и об отсутствии или наличии зависимости между признаками группировки в таблице. Чтобы воспользоваться критерием Пирсона χ^2, в таблице взаимной сопряженности наряду с эмпирическими часто там записывают теоретические частоты, рассчитываемые исходя из предположения, что распределение внутри таблицы случайно и, следовательно, зависимостьмежду признакамигруппировки отсутствует. То есть считается, что распределение частот в каждой строке (столбце) таблицы пропорционально распределению частот в итоговой строке (столбце). Поэтому теоретические частоты по строкам (столбцам)

рассчитывают пропорционально распределению единиц в итоговой строке (столбце).

Относительной мерой тесноты стохастической связи между признаками служат также:
коэффициент взаимной сопряженности Чупрова

$$C = \sqrt{\frac{\chi^2}{n\sqrt{(m_x - 1)(m_y - 1)}}}$$

Но, следует отметить, что рассчитывать коэффициент Чупрова для таблицы «четырех полей» не рекомендуется, так как при числе степеней свободы v=(2-1)(2-1)=1 он будет больше коэффициента Пирсона (в нашем примере C=0,54). Для таблиц же большей размерности всегда $C<\chi^2$.
Коэффициент взаимной сопряженности Крамера(при $т_x \neq m_y$) характеризует меру связи двух номинальных переменных на основе критерия хи-квадрат . Применяется к таблицам сопряженности произвольной размерности

$$C = \sqrt{\frac{\chi^2}{n\sqrt{(m_{min} - 1)}}}$$

где m_{min}— минимальное число групп (m_xили m_y).
Значение коэффициента колеблется от 0 до 1 и теснота связи тем сильнее, чем более близко С к 1.

Достаточно часто в практике статистических исследований анализируются связи между альтернативными признаками, которые представлены группами с противоположными (взаимоисключающими) характеристиками. Тесноту связи в этом случае можно оценивать посредством коэффициента ассоциации Д. Юла и коэффициента контингенции К. Пирсона.

Для расчета указанных коэффициентов измерения тесноты связи между альтернативными признаками используется таблица взаимной сопряженности в виде корреляционной таблицы, которая носит название «четырехклеточной таблицы».

Таблица 3

a	b	*a*+*b*
c	*d*	*c*+*d*
a + *c*	*b* + *d*	*a*+*b* + *c* +*d*

При использовании табл. с частотами*a, b, c, d*коэффициент *ассоциации* (K_a)вычисляется по формуле:

$$K_a = \frac{ad - bc}{ad + bc}$$

При $K_a > 0,3$ между изучаемыми качественными признаками существует корреляционная связь.

В случаях, когда один из показателей четырехклеточной таблицы отсутствует, величина коэффициента ассоциации будет равняться единице, что дает завышенную оценку тесноты связи между признаками. В этом случае необходимо рассчитывать коэффициент контингенции (K_k):

$$K_k = \frac{ad - bc}{\sqrt{(a+b)(b+d)(a+c)(c+d)}}$$

Коэффициент контингенции находится в диапазоне от -1 к +1. Чем более близко $K_к$к (+1) или (-1), тем теснее связь между изучаемыми признаками. Коэффициент контингенции всегда меньше коэффициента ассоциации.

Для определения связи, как между количественными, так и качественными признаками при условии, что значения этих признаков упорядочены по степени уменьшения или увеличения (ранжированные), может быть использован коэффициент корреляции рангов Спирмена.Рангами называют числа натурального ряда, которые представляются в баллах по определенным критериям элементов совокупности. При этом ранжирование проводится по каждому признаку отдельно: первый ранг предоставляется наименьшему значению признака, последний - наибольшему. Количество рангов равняется объему совокупности. Преимуществом этого подхода является то, что при отсутствии требования нормального распределения ранговые оценки тесноты связи целесообразно использовать для совокупности небольшого объема.

Показатель ранговой корреляции - коэффициент корреляции рангов Спирмена — рассчитывается по формуле:

$$\rho = 1 - \frac{6\sum_{j=1}^{n} d_j^2}{n(n^2 - 1)}$$

, где d_j. — разность между рангами по одному и другому признаку $(d_j = R_{xj} - R_{jj})$;n - количество единиц в ряду. Если $d_j = 0$ $\rho = 1$ —существует тесная прямая связь. Если первому рангу по размеруодного признака соответствует последний ранг по размеру второго признака, второму рангу - предпоследний ранг второго признака и т. п., то $\rho = -1$и существует тесная обратная связь. Если значениеρ близко к нулю, то связь слабая или ее вообще нет.Для предварительной оценки тесноты связи между атрибутивными признаками используются также такие характеристики, как коэффициенты Фехнера и Кендэла.

Коэффициент Фехнера относится к простейшим показателям степени тесноты связи и иногда называют коэффициентом корреляции знаков, который был

предложен немец ученым г. Фехнером (1801 —1887). Этот показатель основан на оценке степени согласованности направлений отклонений индивидуальных значений факторного и результативного признаков от соответствующих средних. Для его расчета вычисляют средние значения результативного и факторного признаков, а затем проставляют знаки отклонений для всех значений взаимосвязанных пар признаков.

Обозначим через n_aчисло совпадений знаков отклонений индивидуальных величин от средней, Через$n_в$— число несовпадений таков отклонений. Формула коэффициента Фехнера записывается так:

$$K_ф = (n_a — n_в)/(n_a + n_в).$$

Коэффициент Фехнера может принимать различные значения в пределах от —1 до +1. Если знаки всех отклонений совпадут, то $n_в = 0$ и тогда коэффициент будет равен 1, что свидетельствует о возможном наличии прямой связи. Если же знаки всех отклонений будут разными, тогда n_a=0 и коэффициент Фехнера будет равен —1, это дает основание предположить наличие обратной связи.

Как видно из приведенной формулы для расчета коэффициента Фехнера, величина этого показателя не зависит от величины отклонений факторного и результативного признаков от соответствующей средней величины. Поэтому нельзя говорить о степени тесты корреляционной связи, а тем более об оценке ее существенности на основании только коэффициента Фехнера. При малом объеме исходной информации коэффициент Фехнера практически решает ту же задачу, которая ставится при построении групповых корреляционных таблиц, т.е. отвечает на вопрос о наличии и направлении корреляционной связи между признаками. В том случае, если построена корреляционная или же групповая таблица, дополнительный расчет коэффициента Фехнера не имеет практической ценности.

М. Кендэл предложил еще одну меру связи между переменными *x* и *y* — коэффициент корреляции рангов Кендэла(т):

$$\tau = \frac{2S}{n(n-1)},$$

где $S=P+Q$.

Для вычисления τ надо упорядочить ряд рангов переменной *x,* приведя его к ряду натуральных чисел. Затем рассматривают последовательность рангов переменной *y* .

Для нахождения суммы *S*находят два слагаемых *P* и *Q*. При определении слагаемого *P* нужно установить, сколько чисел, находящихся справа от каждого из элементов последовательности рангов переменной *y,* имеют величину ранга, превышающую ранг рассматриваемого элемента.

Поскольку коэффициенты корреляции рангов могут изменяться пределах от — 1 до +1 (как и линейный коэффициент корреляции), по результатам расчетов

коэффициента Спирмэна можно предположить наличие достаточно тесной прямой зависимости между оценками экспертов на стадии предвыборной кампании и результатами выборов. Однако нельзя не учесть то обстоятельство, что ранговый коэффициент был рассчитан по небольшому объему исходной информации (n = 10). Не является ли отличие рангового коэффициента от нуля лишь результатом случайных совпадений оценок экспертов с результатами выборов по данным малого числа отобранных депутатов, можно ли распространить полученные выводы на генеральную совокупность?

Для совокупностей небольшого объема (n< 30) распределение рангового коэффициента корреляции не является нормальным, поэтому нецелесообразно использовать значенияt по нормированной функции Лапласа для проверки гипотезы о величине рангового коэффициента корреляции. В справочной литературе приводится таблица предельных значений коэффициентов корреляции рангов Спирмэна при условии верности нулевой гипотезы об отсутствии корреляционной связи при заданном уровне значимости и определенном объеме выборочных данных.

Могут встретиться случаи, когда невозможно установить ранговые различия нескольких смежных значений. В этих случаях принято брать средний ранг (даже если он будет дробным числом) и полученный средний ранг приписывать каждому из таких значений, т.е. переходить к матрице переформированных рангов. Например, двум факторам один из экспертов приписывает одинаковый ранг 3. Тогда каждому из факторов присваивается ранг 3,5, так как они поделили между собой третье и четвертое места (3 + 4)/2, а фактору, имевшему ранг 4, присваивается ранг 5 и т.д.

Если определяется теснота связи между k-м и l-м признаками, в рядах значений которых имеется соответственно qи gгрупп объединенных рангов, то формула коэффициента корреляции рангов Спирмэна примет вид:

$$\rho = \frac{\frac{n^3 - n}{6} - \left(T_k - T_l\right) - \sum_{i=1}^{n} d_i^2}{\sqrt{\left(\frac{n^3 - n}{6} - 2T_k\right) \times \left(\frac{n^3 - n}{6} - 2T_l\right)}},$$

где $T_k = \sum_{i=1}^{q} \frac{t_{k_i}^3 - t_{k_i}}{12}; T_{;} = \sum_{i=1}^{q} \frac{t_{l_i}^3 - t_{l_i}}{12}$; t_{ki}и t_{li}определяют количество единиц в i-й группе объединенных рангов соответствующего признака. Скорректированная формула для вычисления коэффициента Корреляции рангов Кендэла будет иметь вид:

$$\tau = \frac{S}{\sqrt{\left(\frac{n(n-1)}{2} - V_k\right) \times \left(\frac{n(n-1)}{2} - V_l\right)}}$$

где $V_{k;} = \sum_{i=1}^{q} \frac{t_{k_i}\left(t_{k_i}-1\right)}{2}$; $V_{l;} = \sum_{i=1}^{q} \frac{t_{l_i}\left(t_{l_i}-1\right)}{2}$.

Эконометрическую модель можно построить, используя два типа исходных данных: 1) данные, характеризующие совокупность различных объектов в определенный момент (период) времени; 2) данные, характеризующие один объект за ряд последовательных моментов (периодов) времени.

Модели, построенные по данным первого типа, называются пространственными моделями. Модели, построенные по данным второго типа, называются моделями временных рядов.

Временной ряд - это совокупность значений какого-либо показателя за несколько последовательных моментов (периодов) времени. Каждый уровень временного ряда формируется под воздействием большого числа факторов, которые условно можно подразделить на три группы: 1) факторы, формирующие тенденцию ряда; 2) факторы, формирующие циклические колебания ряда; 3) случайные факторы.

При различных сочетаниях этих факторов зависимость уровней ряда от времени может принимать разные формы. Во-первых, большинство временных рядов экономических показателей имеют тенденцию, характеризующую совокупное долговременное воздействие множества факторов на динамику изучаемого показателя. По всей видимости, эти факторы, взятые в отдельности, могут оказывать разнонаправленное воздействие на исследуемый показатель. Однако в совокупности они формируют его возрастающую или убывающую тенденцию. Во-вторых, изучаемый показатель может быть подвержен циклическим колебаниям. Эти колебания могут носить сезонный характер, поскольку экономическая деятельность ряда отраслей зависит от времени года. При наличии больших массивов данных за длительные промежутки времени можно выявить циклические колебания, связанные с общей динамикой конъюнктуры рынка, а также с фазой бизнес-цикла, с фазой, в которой находится экономика страны. Некоторые временные ряды не содержат тенденции и циклическую компоненту, а каждый следующий их уровень образуется как сумма среднего уровня ряда и некоторой (положительной или отрицательной) случайной компоненты.

Очевидно, что реальные данные не соответствуют полностью ни одной из описанных выше моделей. Чаще всего они содержат все три компоненты. Каждый их уровень формируется под воздействием тенденции, сезонных колебаний и случайной компоненты. В большинстве случаев фактический уровень временного ряда можно представить как сумму или произведение трендовой, циклической и случайной компонент. Модель, в которой временной ряд представлен как сумма перечисленных компонент, называется аддитивной моделью временного ряда. Модель, в которой временной ряд представлен как произведение перечисленных компонент, называется мультипликативной моделью временного ряда. Основная задача эконометрического исследования

отдельного временного ряда - выявление и придание количественного выражения каждой из перечисленных выше компонент, с тем чтобы использовать полученную информацию для прогнозирования будущих значений ряда или при построении моделей взаимосвязи двух или более временных рядов.

При наличии тенденции и циклических колебаний значения каждого последующего уровня ряда зависят от предыдущих значений. Корреляционную зависимость между последовательными уровнями временного ряда называют автокорреляцией уровней ряда. Количественно ее можно измерить с помощью парного линейного коэффициента корреляции между уровнями исходного временного ряда и уровнями этого ряда, сдвинутыми на несколько шагов во времени.

Число периодов, по которым рассчитывается коэффициент автокорреляции, называется лагом. С увеличением лага число пар значений, по которым рассчитывается коэффициент автокорреляции, уменьшается. Считается целесообразным для обеспечения статистической достоверности коэффициентов автокорреляции использовать правило «максимальный лаг должен быть не больше n/4».

Отметим два важных свойства коэффициента автокорреляции. Во-первых, он строится по аналогии с линейным коэффициентом корреляции и, таким образом, характеризует тесноту только линейной связи текущего и предыдущего уровней ряда. Поэтому по коэффициенту автокорреляции можно судить о наличии линейной (или близкой к линейной) тенденции. Для некоторых временных рядов, имеющих сильную нелинейную тенденцию (например, параболу второго порядка или экспоненту), коэффициент автокорреляции уровней исходного ряда может приближаться к нулю. Во-вторых, по знаку коэффициента автокорреляции нельзя делать вывод о возрастающей или убывающей тенденции в уровнях ряда. Большинство временных рядов экономических данных содержат положительную автокорреляцию уровней, однако при этом они могут иметь убывающую тенденцию.

Последовательность коэффициентов автокорреляции уровней первого, второго и т.д. порядков называют автокорреляционной функцией временного ряда. График зависимости ее значений от величины лага (порядка коэффициента автокорреляции) называется коррелограммой. Анализ автокорреляционной функции и коррелограммы позволяет определить лаг, при котором автокорреляция наиболее высокая, следовательно, лаг, при котором связь между текущим и предыдущими уровнями ряда наиболее тесная, т.е. при помощи анализа автокорреляционной функции и коррелограммы можно выявить структуру ряда.

Если наиболее высоким оказался коэффициент автокорреляции первого порядка, исследуемый ряд содержит только тенденцию. Если наиболее высоким оказался коэффициент автокорреляции порядка, ряд содержит циклические колебания с периодичностью в моментов времени. Если ни один из коэффициентов автокорреляции не является значимым, можно сделать

предположение относительно структуры этого ряда: либо ряд не содержит тенденции и циклических колебаний и имеет структуру, сходную со структурой ряда, либо ряд содержит сильную нелинейную тенденцию, для выявления которой нужно провести дополнительный анализ. Поэтому коэффициент автокорреляции уровней и автокорреляционную функцию целесообразно использовать для выявления во временном ряде наличия или отсутствия трендовой компоненты и циклической (сезонной) компоненты. Аналогично, если, например, при анализе временного ряда наиболее высоким оказался коэффициент автокорреляции второго порядка, ряд содержит циклические колебания с циклом, равным двум периодам времени, т.е. имеет пилообразную структуру.

Одним из наиболее распространенных способов моделирования тенденции временного ряда является построение аналитической функции, характеризующей зависимость уровней ряда от времени, или тренда. Этот способ называют аналитическим выравниванием временного ряда.

Задачи для самостоятельной работы

Задача № 1

Вариативность представленной задачи и последующих достигается путем прибавления к каждому показателю Выпускники Вуза Х, Выпускники Вуза У числа, сформированного двумя последними цифрами зачетной книжки студента, решить задачу с использованием вновь образованных данных. Данная система позволит предоставить каждому студенту индивидуальную задачу для решения. В случае если предпоследнее число зачетной книжки это ноль, то формирование двухзначного числа происходит посредством предыдущего числа.

Исследуем связь между трудоустройством выпускников вузов г. Воронежа и уровнем их образования за 2019 год:

Уровень образования	*Количество выпускников*	*Из них*	
		Выпускники Вуза Х	*Выпускники Вуза У*
среднее профессиональное	-	373	489
Высшее	-	396	438
Итого	-	-	-

Рассчитать коэффициент ассоциации (Ка) и коэффициент контингенции (Кк). Сделать развернутый вывод.

Задача № 2

К каждому показателю прибавить число, сформированное двумя последними цифрами зачетной книжки, например, 20 (46+20, 89+20 и т.д.) решить задачу с использованием вновь образованных данных.

Для изучения влияния условий производства на взаимоотношения в коллективе было проведено выборочное обследование рабочих, ответы которых распределены следующим образом:

Условия производства	Взаимоотношения в коллективе			
	хорошее	удовлетворительное	неудовлетворительное	итого
Соответствуют требованиям	64	70	98	-
не полностью соответствуют	45	62	59	-
не соответствуют	40	61	70	-
итого	-	-	-	-

Требуется охарактеризовать связь между исследуемыми показателями с помощью коэффициента взаимной сопряженности Пирсона и Чупрова, сформулировать вывод.

Задача № 3

К каждому показателю 2017, 2018, 2019 гг. прибавить число, сформированное двумя последними цифрами зачетной книжки, например, 20 (310,3+20 и т.д.) решить задачу с использованием вновь образованных данных
По станциям технического обслуживания легковых автомобилей города Х имеются следующие данные:

Месяц	Число поступивших заявок		
	2017	2018	*2019*
01	410,3	313,6	514,0
02	411,1	314,3	514,7
03	411,5	314,4	615,1
04	412,0	314,6	515,6
05	412,6	315,6	616,0
06	416,0	317,1	517,4
07	415,9	316,9	528,2
08	416,2	417,0	518,4
09	416,4	361,5	517,8
10	415,2	361,0	517,5
11	415,0	314,9	517,0
12	412,8	313,8	516,5

Требуется на основе приведенных данных выявить наличие сезонной неравномерности, определить величину сезонной волны с использованием индекса сезонности. Сделать вывод.

Задача № 4

К каждому показателю прибавить число, сформированное двумя последними цифрами зачетной книжки, например, 20 (20+20, 43+20 и т.д.) решить задачу с использованием вновь образованных данных

С помощью биссериального коэффициента корреляции исследовать связь между возрастом и социальным положение основных категорий потенциальных эмигрантов.

Основные категории потенциальных эмигрантов	Возраст, лет				Всего, чел.
	До 30	30-40	40-50	50 и более	
Руководители, чел	50	65	209	36	-
Рабочие, чел	54	52	56	61	-
Итого:	-	-	-	-	-

Сделать вывод.

Задача № 5 (усложненный уровень задачи)

К каждому показателю прибавить число, сформированное двумя последними цифрами зачетной книжки, например, 20 (75,5+20, 80,5+20 и т.д.) решить задачу с использованием вновь образованных данных

Для организаций строительной отрасли анализируют заработную плату Y сотрудников в зависимости от масштаба (количества сотрудников) организации X. Наблюдения по 30 случайно отобранным организациям представлены следующей таблицей:

Y						X
95,5	95,5	97,5	98,5	80,0	81,0	100
80,5	82,0	84,5	85,0	85,5	86,5	200
85,5	88,5	90,0	91,0	55,5	76,0	300
93	93,5	97,5	99,0	201,5	105,0	400
102,0	108,5	108,0	111,5	112,0	98,5	500

1. Постройте уравнение регрессии Y и X и оцените его качество;
2. Можно ли ожидать наличие гетероскедастичности в данном случае? Ответ поясните.
3. Проверьте наличие гетероскедастичности, используя тест Голдфелда – Кванта. Рекомендуется использовать разбиение, при котором R=12.
4. Если предположить, что гетероскедастичность имеет место и дисперсии отклонений пропорциональны значениям X, то какое преобразование вы предложите, чтобы получить несмещенные, эффективные и состоятельные оценки?
5. Постройте новое уравнение регрессии на основе преобразования осуществленного в предыдущем пункте, и оцените его качество.
6. Сравните результаты полученные в пунктах 1 и 5.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Баркалов, С. А. Статистика [Электронный ресурс] / С. А. Баркалов, П. Н. Курочка, В. Б. Курносов. — Электрон. текстовые данные. — Воронеж: Воронежский государственный архитектурно-строительный университет, ЭБС АСВ, 2010. — 775 с. — 978-5-98222-671-6. — Режим доступа: http://www.iprbookshop.ru/29266.html

2. Баркалов, С. А. Статистика [Электронный ресурс]: практикум / С. А. Баркалов, П. Н. Курочка, О. С. Перевалова. — Электрон. текстовые данные. — Воронеж: Воронежский государственный архитектурно-строительный университет, ЭБС АСВ, 2016. — 137 с. — 978-5-89040-639-2. — Режим доступа: http://www.iprbookshop.ru/72941.html

3. Эконометрика: базовый курс: учебник / О.И. Хайруллина, О.В. Баянова; Министерство сельского хозяйства Российской Федерации, федеральное государственное бюджетное образовательное учреждение высшего образования «Пермский аграрно-технологический университет имени академика Д.Н. Прянишникова». – Пермь: ИПЦ «Прокростъ», 2019 – 176 с; 21 см – Библиогр.: с.168. – 50 экз. – ISBN 978-5-94279-464-4 – Текст: непосредственный Режим доступа http://pgsha.ru.pdf

4. Евсеев, Е. А. Эконометрика: учебное пособие для вузов / Е. А. Евсеев, В. М. Буре. — 2-е изд., испр. и доп. — Москва: Издательство Юрайт, 2020. — 186 с. — (Высшее образование). — ISBN 978-5-534-10752-4. — Текст: электронный // ЭБС Юрайт [сайт]. — URL: https://urait.ru/bcode/453562

ЭКОНОМИКО-СТАТИСТИЧЕСКИЕ МЕТОДЫ (ЭКОНОМЕТРИКА)

МЕТОДИЧЕСКИЕ УКАЗАНИЯ

к выполнению самостоятельных работ
для обучающихся всех направлений, специальностей,
форм обучения

Составитель
Свиридова Татьяна Анатольевна

Издается в авторской редакции

Подписано к изданию 18.11.2021.
Уч.–изд. л. 2,3 «С».

ФГБОУ ВО «Воронежский государственный технический университет»
394026 Воронеж, ул. 20-летия Октября, 84