

ФГБОУ ВО «БРЯНСКИЙ ГОСУДАРСТВЕННЫЙ
АГРАРНЫЙ УНИВЕРСИТЕТ»

Экономический факультет

Кафедра коммерции и экономического анализа

А.В. Раевская, Н.А. Каширина

Корреляционно-регрессионный анализ

Учебно-методическое пособие для
студентов экономического факультета

Брянская область
2015

УДК311.1(075.8)
ББК 60.6я73

Раевская А.В., Каширина Н.А. Корреляционно-регрессионный анализ: Учебно-методическое пособие (доп. и перераб.). – Брянск: БГАУ, 2015. – 16 с.

Рецензент: к.э.н., доцент кафедры
коммерции и экономического анализа
Иванюга Т.В.

Учебно-методическое пособие предназначено для оказания помощи студентам в изучении темы «Статистическое изучение связи показателей (корреляционно-регрессионный анализ)», выполнении индивидуального задания (типового расчета) по указанной теме и написании курсовой работы по статистике.

Пособие предназначено для студентов направлений подготовки 38.03.01 Экономика и 09.03.03 Прикладная информатика.

Рекомендовано к переизданию учебно-методическим советом экономического факультета (протокол № 1 от 3 сентября 2015 г.).

© Раевская А.В., Каширина Н.А., 2015
© Брянский ГАУ, 2015

ВВЕДЕНИЕ

Изучение реальной действительности показывает, что практически каждое общественное явление находится в тесной связи и взаимодействии с другими явлениями, какими бы случайными они не казались на первый взгляд. Так, например, уровень урожайности сельскохозяйственных культур зависит от множества природных и экономических факторов, тесно связанных между собой.

Исследования и измерения взаимосвязей и взаимозависимостей социально-экономических явлений является одной из важнейших задач статистики.

Среди методов статистики особое место занимает метод корреляции, который является логическим продолжением таких методов как аналитическое группировки, дисперсионный анализ и сопоставление параллельных рядов. В сочетании с этими методами он предоставляет статистическому анализу законченный, завершённый характер.

Корреляционный анализ помогает решить целый ряд новых задач в анализе экономических данных; расчеты на основе корреляционных моделей повышают степень точности анализа, часто выявляют недостатки предварительного анализа. Преимущество этого метода состоит также и в том, что он дает возможность решать задачи, которые нельзя решить с помощью других методов анализа.

Изучение данной темы и практическое решение задач позволяет студентам направления подготовки 38.03.01 Экономика освоить следующие компетенции:

ОК-1: владеет культурой мышления, способен к обобщению, анализу, восприятию информации, постановке цели и выбору путей ее достижения.

ОК-4: способен анализировать социально-значимые проблемы и процессы, происходящие в обществе, и прогнозировать возможное их развитие в будущем.

ПК-4: способен осуществлять сбор, анализ и обработку данных, необходимых для решения поставленных экономических задач.

ПК-5: способен выбрать инструментальные средства для обработки экономических данных в соответствии с поставленной задачей, проанализировать результаты расчетов и обосновать полученные выводы.

ПК-6: способен на основе описания экономических процессов и явлений строить стандартные теоретические и эконометрические модели,

анализировать и содержательно интерпретировать полученные результаты.

ПК-10: способен использовать для решения аналитических и исследовательских задач современные технические средства и информационные технологии.

Изучение данной темы и практическое решение задач позволяет студентам направления подготовки 09.03.03 Прикладная информатика овладеть следующими компетенциями:

ОК-3: способностью использовать основы экономических знаний в различных сферах деятельности.

ОК-5: способностью к коммуникации в устной и письменной формах на русском и иностранном языках для решения задач межличностного и межкультурного взаимодействия.

ПК-3: способностью проектировать ИС в соответствии с профилем подготовки по видам обеспечения.

КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

Цель: изучение видов взаимосвязей явлений (функциональная, корреляционная); видов корреляционной зависимости; способов определения тесноты связи; усвоение смысла показателей регрессии.

После изучения вы сможете: определять силу тесноты связи между социально – экономическими показателями, давать интерпретацию параметров управления регрессии.

Содержание темы: исследование объективно существующих связей; количественные оценки тесноты связи; регрессионный анализ; показатели тесноты связи (параметрические, ранговые).

При статистических исследованиях корреляционных связей одной из главных задач является определение формы корреляционной связи, т.е. *построение модели связи*.

Для аналитических целей корреляционную связь представляют при помощи математических функций, т.е. придают ей функциональную форму. Под формой связи понимают тенденцию, которая проявляется в изменении результативного признака в связи с изменением признака-фактора.

Построение и анализ корреляционной модели связи осуществляются с помощью корреляционно-регрессионного анализа, который состоит из следующих этапов:

- предварительного априорного анализа;
- сбора информации и ее первичной обработки;
- построения модели (уравнения регрессии);
- оценки и анализа модели.

Все этапы связаны между собой, границы их часто переплетаются и носят условный характер.

Форма корреляционной связи может быть выражена различными математическими функциями. Выбор формы связи решается на основе теоретического анализа сущности изучаемых явлений и исследования эмпирических данных.

Эмпирическое исследование формы связи включает построение графиков корреляционных полей, эмпирических линий регрессии, а также анализ параллельных рядов. Изучение эмпирического материала дает возможность установить направление и форму связи.

Для определения видов функции необходимо применять ком-

плекс приемов: экономический, логический, графический и математический.

Линейная форма связи может быть выражена уравнением прямой: $\tilde{y}_x = a + bx$

Нелинейная форма связи показана:

1) уравнением параболы второго порядка $\tilde{y}_x = a + bx + cx^2$;

2) уравнением гиперболы $\tilde{y}_x = a + \frac{b}{x}$;

3) показательной функцией $\tilde{y}_x = a + b^x$;

4) степенной функцией $\tilde{y}_x = a \cdot x^b$.

и другими функциями.

Главной проблемой при построении модели связи является определение вида аналитической функции, которая отразит механизм связи между факторным и результативным признаками и даст количественную оценку этой связи.

Наиболее часто для определения формы корреляционной связи используют уравнение прямой: $\tilde{y}_x = a + bx$,

где \tilde{y}_x – теоретические значения результативного признака;

x – факторный признак;

a – параметр уравнения (свободный член);

b – параметр уравнения (коэффициент регрессии).

Уравнением связи называется *уравнение регрессии*, а анализ, производимый с помощью уравнения регрессии, называется *регрессионным анализом*.

После установления вида функции для модели связи определяются параметры уравнения регрессии a и b . Параметры уравнения регрессии определяются методом наименьших квадратов, суть которого состоит в том, что теоретическая линия регрессии должна быть проведена так, чтобы сумма квадратов отклонений эмпирических данных от теоретических была величиной минимальной. Исчисляя первые производные по a и b от функции $\sum (y - a - bx)^2 \rightarrow \min$ и приравнявая их к нулю, получаем систему нормальных уравнений вида:

$$\begin{cases} \sum y = na + b \sum x \\ \sum yx = a \sum x + b \sum x^2 \end{cases}$$

Решая систему нормальных уравнений, определяем значения параметров a и b :

$$b = \frac{n \sum yx - \sum x \sum y}{n \sum x^2 - \sum x \sum x}$$
$$a = \bar{y} - b\bar{x}$$

Параметр b показывает изменения результативного признака при изменении факторного признака на единицу. Параметр a не имеет экономического содержания, так как может принимать отрицательные значения.

Очень часто исследуемые признаки имеют разные единицы измерения, поэтому для оценки влияния факторного признака на результативный применяется *коэффициент эластичности*. Он вычисляется для каждой точки и в среднем для всей совокупности.

Средний коэффициент эластичности для уравнения прямой вычисляется следующим образом: $\mathcal{E} = b \frac{\bar{x}}{\bar{y}}$.

Коэффициент эластичности показывает, на сколько процентов изменяется результативный признак при изменении факторного признака на один процент.

Уравнение регрессии имеет практическое значение. Сравнивая фактический объем продукции у отдельных предприятий с теоретическим, мы получаем возможность его оценки с точки зрения средних условий существующих в данной совокупности предприятий. Регрессионную модель можно использовать для прогноза выпуска продукции в зависимости от изменения основных фондов тогда, когда не изменяются условия формирования уровней исследуемого признака.

Измерение тесноты корреляционной связи. Важное место в анализе регрессионной модели занимает *оценка тесноты корреляционной связи* между изучаемыми признаками.

Для измерения тесноты корреляционной связи между признаками при линейной форме связи применяется *линейный коэффициент корреляции*:

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

Он изменяется в пределах от -1 до +1 и показывает тесноту и направление корреляционной связи. Чем ближе коэффициент корреляции к 1 (по модулю), тем связь теснее. Отрицательное значение свидетельствует об обратной связи между признаками. Коэффициент корреляции можно вычислять и по формулам:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{n \sigma_x \sigma_y}$$

$$r = a_1 \cdot \frac{\sigma_x}{\sigma_y}$$

При любой форме связи для измерения тесноты корреляционной связи применяются *теоретическое корреляционное отношение* и *индекс корреляции*. Теоретическое корреляционное отношение определяется по формуле:

$$\eta = \sqrt{\frac{\sigma_{y_x}^2}{\sigma_y^2}}$$

Факторная дисперсия, характеризующая вариацию результативного признака под влиянием вариации признака-фактора определяется по следующей формуле:

$$\sigma_{y_x}^2 = \frac{\sum (y_x - \bar{y})^2}{n}$$

Общая дисперсия, характеризующая вариацию результативного признака под влиянием всех факторов, вызывающих эту вариацию, определяется по формуле

$$\sigma_y^2 = \frac{\sum (y - \bar{y})^2}{n}$$

Представим индекс корреляции:

$$R = \sqrt{1 - \frac{\sigma_{y-y_x}^2}{\sigma_y^2}}$$

Остаточная дисперсия, характеризующая вариацию результативного признака под влиянием прочих неучтенных факторов, определяется по формуле

$$\sigma_{y-y_x}^2 = \frac{\sum (y_x - \hat{y}_x)^2}{n}$$

Индекс корреляции и теоретическое корреляционное отношение изменяются от 0 до 1 и показывают не только тесноту связи, но и степень пригодности подобранных функций связи.

η^2 и $R^2 (r^2)$ – называются коэффициентами детерминации, которые показывают долю вариации результативного признака под влиянием вариации признака-фактора. Коэффициент детерминации используют в качестве критерия оценки подбора наилучшей модели связи.

Показатели тесноты корреляционной связи используются не только для оценки уже построенной модели связи (уравнения регрессии), но и для выбора оптимального варианта формы связи. Если теоретический анализ не дает возможности дать однозначный ответ о форме связи, то необходимо строить уравнения регрессии с различными формами связи – линейные и нелинейные. Оценка пригодности модели связи осуществляется путем анализа коэффициента детерминации или индекса корреляции. Наилучшей считается модель с наибольшими значениями этих показателей.

При линейной форме связи теоретическое корреляционное отношение и линейный коэффициент корреляции равны.

Решение задачи с помощью программы STRAZ

Задача. Установить зависимость трудоемкости производства картофеля от двух факторов.

Для построения уравнения регрессии трудоемкости используем следующие показатели:

y – трудоемкость на производство картофеля, чел.-час.;

x_1 – электровооруженность на производство картофеля, л.с.;

x_2 – урожайность картофеля, ц/га.

Общий вид уравнения регрессии: $\hat{y} = a + b_1x_1 + b_2x_2$.

Уравнение регрессии имеет вид:

МНОГОФАКТОРНЫЙ КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

Исходная матрица:

	1	2	3
1	13.400	206.300	29.500
2	1.000	848.800	257.900
3	3.300	171.200	100.000
4	.700	490.200	280.400
5	.500	184.500	200.000
6	1.700	254.300	66.500
7	40.000	111.800	5.900
8	.800	282.300	175.900
9	.100	437.300	364.000
10	6.000	236.000	58.300
11	20.600	196.200	34.100
12	2.300	84.600	24.400
13	.900	302.400	145.200
14	1.200	334.500	283.100
15	5.700	129.000	48.200
16	.800	625.900	75.800
17	2.300	51.700	216.700
18	.900	301.600	345.100

Корреляционная матрица:

	1	2	3
1	1.000	-.358	-.539
2	-.358	1.000	.464
3	-.539	.464	1.000

Множественная регрессия и одномерные статистики

## переменных	Множественная регрессия		Одномерные статистики	
	Козфф. регр.	Станд. ош.	Средние	Станд. откл.
Зависимая	5.6778	10.0662
Независимая 1	-.0070	.0123	293.8112	199.0206
2	-.0406	.0208	150.6111	117.8487
Свободный член уравнения регрессии =			13.8320	
Множественный коэфф. корреляции =			.5523	
Множественный коэфф. детерминации =			.3051	
Стандартная ошибка ур. регр. =			8.9334	
F-значение =			3.2923	
Число степеней свободы для воспр. дисперсии =			2	
для остат. дисперсии =			15	
Бета-коэффициенты:				
-.1377				-.4748
Значения Т-Стьюдента:				
-.5667				-1.9537
Коэффициенты отдельного определения:				
.0493				.2557
Коэффициенты эластичности:				
-.3605				-1.0757

$$\hat{y} = 13,832 - 0,007x_1 - 0,041x_2 .$$

Из уравнения видно, что при повышении электровооруженности на 1 л. с. трудоемкость снижается на 0,007 чел.-час. при фиксированном значении второго фактора.

При повышении урожайности на 1 ц/га трудоемкость сокращается на 0,041 чел.-час. при фиксированном значении первого фактора.

Таким образом, оба фактора оказывают обратное влияние на изменение трудоемкости производства картофеля.

Множественный коэффициент корреляции $R = 0,552$ показывает, что между трудоемкостью картофеля и факторами наблюдается средняя связь, а это говорит о том, что в анализ включены достаточно существенные факторы.

Коэффициент множественной детерминации $R^2 = 0,305$, говорит о том, что вариация трудоемкости картофеля в данной совокупности районов на 30,5% зависит от влияния электровооруженности и урожайности, а на 69,5% под влиянием факторов, не включенных в анализ.

Коэффициент множественной детерминации подразделяется на составляющие части, которые получили название коэффициенты отдельного определения. В нашем случае коэффициенты отдельного определения показывают долю влияния каждого фактора на трудоемкость продукции.

$$O_{x_1} = 0,049 ; O_{x_2} = 0,256 .$$

Из 30,5% общего объема вариации трудоемкости картофеля на долю электровооруженности приходится 4,9% и на долю урожайности картофеля 25,6%.

Оценку статистической значимости параметров регрессии проводим с помощью t -критерия Стьюдента. При уровне значимости $\alpha=0,05$ (с вероятностью 95%) и числе степеней свободы $k = n - m$, где m – число параметров уравнения регрессии, n – число единиц совокупности.

В нашем случае $k = n - m = 18 - 3 = 15$.

По таблицам критических значений t -критерия необходим значение $t_{табл(15;0,05)} = 2,13$.

Затем сравниваем $t_{факт}$ с $t_{табл}$.

Для параметра b_1 $|t_{факт}| = 0,57$.

Для параметра $b_2 \left| t_{\text{факт}} \right| = 1,95$.

Так как $t_{\text{факт}} < t_{\text{табл}}$, то параметры уравнения признаются статистически незначимыми и ненадежными.

Статистическая значимость уравнения регрессии и показателя тесноты связи (коэффициента корреляции) проверяем с помощью F -критерия Фишера.

По таблицам значений F -критерия Фишера при $\alpha=0,05$ и двух степенях свободы: $k_1 = m$ и $k_2 = n - m - 1$ определяем $F_{\text{табл}(2;15;0,05)} = 3,68$.

Сравниваем значение $F_{\text{расч}} = 3,29$ со значением $F_{\text{табл}}$.

По расчетам $F_{\text{расч}} = 3,29 < F_{\text{табл}}$. Следовательно, полученное уравнение регрессии и коэффициент корреляции с вероятностью 95% являются статистически незначимыми и ненадежными.

Для оценки резервов снижения трудоемкости продукции используют β – коэффициенты:

$$\beta_{x_1} = -0,138; \beta_{x_2} = -0,475.$$

Оба фактора оказывают обратное влияние на изменение трудоемкости производства. Но так как $\beta_{x_1} < \beta_{x_2}$ можно сделать вывод, что наибольшие резервы снижения трудоемкости производства заложены в увеличении урожайности картофеля.

Коэффициент эластичности показывает, насколько процентов изменяется трудоемкость производства при увеличении фактора на 1%.

$$\varepsilon_{x_1} = -0,36; \varepsilon_{x_2} = -1,08.$$

Следовательно, при повышении электровооруженности на 1% трудоемкость понижается на 0,36% при фиксированном значении второго фактора.

При росте урожайности на 1% трудоемкость производства картофеля снижается на 1,08%, при фиксированном значении электровооруженности.

Таблица значений F-критерия Фишера при уровне значимости $\alpha = 0,05$

$k_1 \backslash k_2$	1	2	3	4	5	6	8
1	161,5	199,5	215,7	224,6	230,2	233,9	238,9
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28

Критические точки распределения t – критерия Стьюдента

Число степеней свободы k	Уровень значимости α			
	0,1	0,05	0,02	0,01
1	6,3138	12,7062	31,8205	63,6567
2	2,9200	4,3027	6,9646	9,9248
3	2,3534	3,1824	4,5407	5,8409
4	2,1318	2,7764	3,7469	4,6041
5	2,0150	2,5706	3,3649	4,0321
6	1,9432	2,4469	3,1427	3,7074
7	1,8946	2,3646	2,9980	3,4995
8	1,8595	2,3060	2,8965	3,3554
9	1,8331	2,2622	2,8214	3,2498
10	1,8125	2,2281	2,7638	3,1693
11	1,7959	2,2010	2,7181	3,1058
12	1,7823	2,1788	2,6810	3,0545
13	1,7709	2,1604	2,6503	3,0123
14	1,7613	2,1448	2,6245	2,9768
15	1,7531	2,1314	2,6025	2,9467
16	1,7459	2,1199	2,5835	2,9208
17	1,7396	2,1098	2,5669	2,8982
18	1,7341	2,1009	2,5524	2,8784
19	1,7291	2,0930	2,5395	2,8609
20	1,7247	2,0860	2,5280	2,8453
21	1,7207	2,0796	2,5176	2,8314
22	1,7171	2,0739	2,5083	2,8188
23	1,7139	2,0687	2,4999	2,8073
24	1,7109	2,0639	2,4922	2,7969
25	1,7081	2,0595	2,4851	2,7874
26	1,7056	2,0555	2,4786	2,7787
27	1,7033	2,0518	2,4727	2,7707
28	1,7011	2,0484	2,4671	2,7633
29	1,6991	2,0452	2,4620	2,7564
30	1,6973	2,0423	2,4573	2,7500
35	1,6896	2,0301	2,4377	2,7238
40	1,6839	2,0211	2,4233	2,7045
45	1,6794	2,0141	2,4121	2,6896
50	1,6759	2,0086	2,4033	2,6778

Значения коэффициента корреляции

Значение коэффициента корреляции	Степень тесноты связи	Направление связи
$-0,2 \leq r \leq 0,2$	связь отсутствует	
$-0,4 \leq r < -0,2$	слабая	обратная
$0,2 < r \leq 0,4$	слабая	прямая
$-0,6 \leq r < -0,4$	средняя	обратная
$0,4 < r \leq 0,6$	средняя	прямая
$-0,8 \leq r < -0,6$	тесная	обратная
$0,6 < r \leq 0,8$	тесная	прямая
$-0,9 \leq r < -0,8$	высокая	обратная
$0,8 < r \leq 0,9$	высокая	прямая
$-0,99 \leq r < -0,9$	очень высокая	обратная
$0,9 < r \leq 0,99$	очень высокая	прямая
$r = -1$	функциональная	обратная
$r = 1$	функциональная	прямая

Учебное издание

Анна Васильевна Раевская

Надежда Алексеевна Каширина

Учебно-методическое пособие для
студентов экономического факультета

Редактор Павлютина И.П.

Подписано к печати 12.09.2015 г. Формат 60x84 Бумага печатная
Усл. п.л. 0,98 Тираж 50 экз. Издат. № 3198.

Издательство Брянской государственной сельскохозяйственной академии
243365 Брянская обл., Выгоничский район, с. Кокино, Брянская БГАУ