МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА РФ
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ АГРАРНЫЙ УНИВЕРСИТЕТ
ИМ. И. Т. ТРУБИЛИНА»**

КАФЕДРА СТАТИСТИКИ И ПРИКЛАДНОЙ МАТЕМАТИКИ

СТАТИСТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ ДАННЫХ

Методические рекомендации к выполнению контрольных работ студентами магистрами по направлению подготовки 36.04.01 «Ветеринарно – санитарная экспертиза»

Краснодар
2016

Содержание

Введение

В процессе всей своей жизни человек принимает решения: в личной сфере (в какой вуз поступать, с кем общаться, как учиться); в общественной (посещать вечера, театры, митинга, собрания, выборы); в производственной (определение факторов, существенно влияющих на урожайность, качество материалов и т.д.); научной (выдвижение и проверка научных гипотез). Принятие решений обычно преследует одну из целей: характеристика существующего состояния объекта, прогнозирование будущего состояния процесса (объекта); управление (т.е. как следует изменять одни параметры объекта (процесса), чтобы другие параметры приняли желаемое значение; объяснение внутренней структуры объекта (процесса).

Одним из основных подходов к обоснованию и последующему принятию решения является статистический, основанный на использовании статистических методов и приемов при обработке данных по массе явлений.

Статистические методы обработки данных можно разделить:

по способу получения экспериментальных данных:

а) активные эксперименты;

б) пассивные эксперименты (выборочное или сплошное наблюдение);

по цели обработки данных:

а) описательные (получение и сравнение числовых характеристик экспериментальных данных) - анализ вариационных рядов, выборочный метод, проверка статистических гипотез и другие;

б) аналитические (количественная оценка и анализ зависимостей, описывающих изучаемые объекты (процессы) – дисперсионный анализ, регрессионный анализ, анализ рядов динамики и другие).

Цель методических рекомендаций – оказать помощь студентам - магистрам в овладении приемами и методами статистико-математического исследования, в закреплении теоретических знаний, полученных на лекциях и при самостоятельной работе во внеучебное время.

Рекомендации могут быть использованы при самостоятельном изучении курса. Для систематизации и закрепления изучаемого материала даются теоретические пояснения.

По каждой теме предусмотрено выполнение студентами контрольных заданий, с последующей проверкой преподавателем. Выполнение всех заданий должно сопровождаться краткими выводами по результатам расчетов.

Контрольная работа № 1.
Однофакторный корреляционно-регрессионный анализ

Для количественного описания взаимосвязей между переменными, широко используются методы регрессии и корреляции. В зависимости от количества факторов изучаемые признаки подразделяются на факторные (независимые) и результативные (зависимые). Факторные – это признаки, обуславливающие изменение результативного, зависимого признака. Результативный признак или зависимая переменная – это признак, изменяющийся под влиянием факторных признаков, обозначается *Y*. Факторные признаки или независимые переменные обозначаются X_j, $j=1, 2, 3, \ldots p$, где p – число факторов.

Если на изменение результативного признака *Y* оказывает влияние один (основной, доминирующий) фактор, то связь называется парной. Соответственно регрессионный анализ связи *Y* от *X* называется однофакторным.

Уравнение связи между двумя переменными имеет вид $y = f(x)$, где y – зависимая переменная (результативный признак); x – независимая переменная (факторный признак). Реальные наблюдения, в силу неучтенных факторов, отличаются от теоретического значения y на величину ε – случайную ошибку.

Уравнения регрессии подразделяются на линейные и нелинейные. Модель линейной регрессии имеет следующий вид:

$$y = a + bx + \varepsilon, \qquad (1.1)$$

где ε – случайный член, характеризующий отклонение фактически наблюдаемых значений результативного признака от значений, найденных по уровню регрессии;
a и b – параметры уравнения регрессии.

Вид функции $f(x)$ подбирается на основе теоретического анализа сущности изучаемых явлений и процессов. Обычно используют графическое отображение пар наблюдений на плоскости. В исследованиях наиболее часто применяется линейная форма связи между двумя переменными вследствие наглядности и четкой экономической интерпретации параметров уравнения. Модель линейной регрессии имеет вид:

$$y_i = a + bx_i + \varepsilon_i, \text{ где } i = 1, 2, \ldots, n, \qquad (1.2)$$

где y_i – наблюдаемые значения результативного признака по i-й единице совокупности;
x_i – наблюдаемые значения факторного признака по i-й единице совокупности;

n – число единиц изучаемой совокупности объектов или явлений.

На практике часто применяются нелинейные модели, которые подразделяются на два класса.

Нелинейные модели по объясняющим переменным:

$y = b_o + b_1 x + b_2 x^2 + ... + b_к x^к + \varepsilon$ – полином $к$-го порядка;

$y = a + \frac{b}{x} + \varepsilon$ – гипербола;

Нелинейные модели по оцениваемым параметрам:

$y = a \cdot x^b \cdot \varepsilon$ – степенная;

$y = a \cdot e^{bx}$ – экспоненциальная;

$y = a \cdot b^x \cdot \varepsilon$ – показательная;

$y = \frac{1}{a + b_o \cdot b_1^{x+\varepsilon}}$ – логистическая.

Выбор конкретной математической модели осуществляется графическим, аналитическим или экспериментальным методами.

Большинство нелинейных моделей с помощью соответствующих преобразований обычно удается привести к линейной модели, т.е. линеаризовать. Если же модель внутренне нелинейная по параметрам, то для оценки ее параметров используют итеративные методы.

Таблица 1 – Линеаризующие преобразования нелинейных моделей

№ п/п	Функция	Линеаризующие преобразования			
		переменных		выражения для величин a и b	
		y'	x'	a'	b'
1	$y = a + b/x$	y	1 / x	a	b
2	$y = 1/(a + bx)$	1 / y	x	a	b
3	$y = x/(a + bx)$	x / y	x	a	b
4	$y = ab^x$	lg y	x	lg a	lg b
5	$y = ae^{bx}$	ln y	x	ln a	b
6	$y = 1/(a + be^{-x})$	1 / y	e^{-x}	a	b
7	$y = ax^b$	lg y	lg x	lg a	b
8	$y = a + b \lg x$	y	lg x	a	b
9	$y = a/(b + x)$	1 / y	x	b/a	$1/a$
10	$y = ax/(b + x)$	1 / y	$1/x$	b/a	$1/a$
11	$y = ae^{b/x}$	ln y	$1/x$	ln a	b
12	$y = a + bx^n$	y	x^n	a	b

Задачами корреляционно-регрессионного анализа являются: установление типа уравнения регрессии; определение параметров уравнения и оцен-

ка и их значимости; оценка тесноты и значимости связи между переменными; определение точечных и интервальных прогнозных значений зависимой переменной.

Корреляционно-регрессионный анализ проводится в определенной последовательности:

1. Исходя из целей и задач исследования, устанавливаются результативный и факторные признаки, значения которых определяются по совокупности заданных объектов.

2. Выбирается модель уравнения регрессии, обычно графическим способом. Для этого в прямоугольной системе координат строится график зависимости между переменными *X* и *Y*. На оси абсцисс откладываются значения факторного признака *X,* а по оси ординат – результативного признака *У* с соблюдением масштаба. На основе вида корреляционного поля делаются выводы о направлении и возможной функциональной форме связи между факторным и результативным признаками (прямая связь или обратная, линейная или нелинейная).

3. Для линейных и нелинейных уравнений, приводимых к линейному виду, методом наименьших квадратов определяются параметры уравнения регрессии. Для этого составляется и решается следующая система уравнений:

$$\begin{cases} \Sigma y = na + b\Sigma x, \\ \Sigma yx = a\Sigma x + b\Sigma x^2. \end{cases} \qquad (1.3)$$

Параметры уравнения линейной регрессии также можно найти по формулам, вытекающим из системы нормальных уравнений:

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \left(\bar{x}\right)^2}, \qquad a = \bar{y} - b \cdot \bar{x}. \qquad (1.4)$$

Коэффициент регрессии линейного уравнения *b* показывает на сколько единиц в среднем изменится результативный признак *У* при изменении факторного признака *X* на единицу.

4. Качество уравнения регрессии оценивается с помощью средней ошибки аппроксимации:

$$\bar{A} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\ \%; \qquad (1.5)$$

Качество уравнения регрессии считается хорошим, если средняя ошибка аппроксимации составляет менее 10–12 %.

5. Корреляционная зависимость между переменными величинами – это функциональная зависимость между значениями одной из них и групповыми средними другой. Теснота связи при линейной зависимости характеризуется

выборочным коэффициентом корреляции *(r)*, который определяется по формуле:

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y}, \quad -1 \le r \le 1, \tag{1.6}$$

где σ_x и σ_y – средние квадратические отклонения по *X* и по *Y*.

$$\sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2}; \quad \sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2}. \tag{1.7}$$

Чем ближе значение коэффициента корреляции к нулю, тем связь между признаками слабее, а чем ближе к единице, тем связь сильнее. Если $|r| = 1$, то связь линейная и функциональная. Если $r = 0$, то признаки линейно независимы. Коэффициент корреляции также показывает направление связи между признаками. Если $r > 0$, то связь прямая, а если $r < 0$, то – обратная.

При нелинейной зависимости теснота связи между переменными *X* и *Y* определяется с помощью индекса корреляции:

$$R_{xy} = \sqrt{1 - \frac{\sigma^2_{ост}}{\sigma^2_y}}. \tag{1.8}$$

Квадрат коэффициента (индекса) корреляции называется коэффициентом (индексом) детерминации.

$$D = r^2 \cdot 100\ \%. \tag{1.9}$$

Коэффициент детерминации *D* показывает долю влияния фактора *X* на результативную переменную *Y*, а (100 % – *D*) – долю влияния других, неучтенных в модели факторов.

6. Средний коэффициент эластичности определяется по формуле:

$$\overline{Э} = f'(x)\frac{\bar{x}}{\hat{y}}. \tag{1.10}$$

При линейной форме связи он находится по формуле:

$$\overline{Э} = b\frac{\bar{x}}{\bar{y}}, \tag{1.11}$$

где $\bar{x}$ и $\bar{y}$ – средние значения признаков;

b – коэффициент регрессии.

Средний коэффициент эластичности показывает, что при увеличении фактора X на 1 %, результативная переменная Y в среднем изменяется на величину коэффициента эластичности.

7. Оценка статистической значимости построенной модели регрессии в целом производится с использованием критерия F-Фишера. Рассматриваются нулевая гипотеза $H_o : r_z^2 = 0$ и альтернативная ей гипотеза $H_1 : r_2^2 \neq 0$.

Наблюдаемое (фактическое) значение F – критерия находится по формуле:

$$F_н = \frac{r^2}{1-r^2}\left(\frac{n-m-1}{m}\right); \tag{1.12}$$

где m – число параметров при переменной x;

n – число наблюдений.

Для однофакторного линейного уравнения регрессии расчет $F_н$ производится по формуле:

$$F_н = \frac{r^2}{1-r^2}(n-2). \tag{1.13}$$

Критическое (табличное) значение $F_{кр}(\alpha;\ k_1;\ k_2)$ находится по таблице Фишера-Снедекора, при заданном уровне значимости α и числе степеней свободы $k_1=m;\ k_2=n-m-1$. В случае парной регрессии число степеней свободы $k_1=1;\ k_2=n-2$.

Если $F_н > F_{кр}$, то нулевая гипотеза отклоняется и принимается альтернативная ей гипотеза о статистической значимости уравнения регрессии, в противном случае уравнение регрессии статистически незначимо.

8. Так как обычно регрессионный анализ зависимости между признаками проводится по выборочным данным, то проверяется значимость величины выборочного коэффициента корреляции, а также параметров уравнения регрессии a и b с использованием критерия t-Стьюдента при заданном уровне значимости α.

Находится наблюдаемое значение t для параметров a и b:

$$t_a = \frac{a}{m_a}; \quad m_a = \frac{\sigma_{ост}\sqrt{\sum x^2}}{n\sigma_x}; \tag{1.14}$$

$$t_b = \frac{b}{m_b}; \quad m_b = \frac{б_{ост}}{б_x\sqrt{n}}. \tag{1.15}$$

По таблице значений t-Стьюдента определяется критическое значение $t_{кр.}$. Если $t_н > t_{кр}$, то основную гипотезу отвергаем и принимаем альтернативную гипотезу и коэффициенты уравнения регрессии a и b считаются статистически значимы при заданном уровне значимости α. В противном случае

основную гипотезу о не значимости параметров уравнения регрессии *a* и *b* принимаем.

Для проверки значимости коэффициента корреляции выдвигаем нулевую гипотезу $H_o : r_г = 0$ – коэффициент корреляции в генеральной совокупности равен нулю и изучаемый фактор не оказывает существенного влияния на результативный признак, при альтернативной гипотезе $H_1 : r_г \neq 0$– коэффициент корреляции в генеральной совокупности значительно отличается от нуля при заданном уровне значимости α.

Для проверки нулевой гипотезы применяется критерий

t-Стьюдента и определяется наблюдаемое значение *t*-критерия:

$$t_н = |r|\sqrt{\frac{n-2}{1-r^2}}. \qquad (1.16)$$

Критическое значение $t_{кр}$ находится по таблице распределения *t*-Стьюдента при уровне значимости α и числе степеней свободы $k = n–2$ для двухсторонней критической области.

Сравнивается $t_н$ и $t_{кр}$. Если $t_н > t_{кр}$, то нулевая гипотеза отвергается и коэффициент корреляции *r* существенно отличается от нуля в генеральной совокупности. Если $t_н < t_{кр}$, то принимаем основную гипотезу о не значимости коэффициент корреляции *r*.

При парной линейной зависимости оценка значимости всего уравнения, коэффициентов корреляции и регрессии дает одинаковые результаты, так как $t_b^2 = t_r^2 = F$ (незначительные отличия объясняются ошибками округлений).

9. Рассчитывается доверительный интервал для параметров *a* и *b*. Для этого определяется предельная ошибка для каждого параметра:

$$\Delta_a = t \cdot \hat{m}_a; \qquad \Delta_b = t \cdot \hat{m}_b. \qquad (1.17)$$

Доверительные интервалы для параметров *a* и *b* определяются по следующим формулам:

$$\gamma_{a_{min}} = a - \Delta_a; \gamma_{a_{max}} = a + \Delta_a; \qquad (1.18)$$

$$\gamma_{b_{min}} = b - \Delta_b; \gamma_{b_{max}} = b + \Delta_b. \qquad (1.19)$$

Если ноль попадает в границы доверительных интервалов, т. е. нижняя граница отрицательна, а верхняя положительна, то оцениваемый параметр с заданной доверительной вероятностью является статистически незначимым. В противном случае принимается статистическая значимость оцениваемого параметра.

10. Прогнозное значение результативного признака определяется путем подстановки в построенное парное линейное уравнение регрессии прогнозного значения факторного признака x_p:

$$\hat{y}_p = a + b\,x. \tag{1.20}$$

11. Для нахождения доверительного интервала прогноза определяется средняя и предельная ошибки прогноза:

$$m_{\hat{y}_p} = 6_{ост}\sqrt{\frac{1}{n} + \frac{(\bar{x}-x_p)^2}{\sum(x_i-\bar{x})^2}}. \tag{1.21}$$

Средняя ошибка индивидуального значения y находится по формуле:

$$m_{\hat{y}_p} = 6_{ост}\sqrt{1 + \frac{1}{n} + \frac{(\bar{x}-x_p)^2}{\sum(x_i-\bar{x})^2}}, \tag{1.22}$$

где $6_{ост} = \sqrt{\frac{\sum(y_i-\hat{y})^2}{n-m-1}}$; $\Delta_{\hat{y}_p} = t_{kp}\cdot\hat{m}_{y_p}$. (1.23)

Определим границы доверительного интервала прогноза:

$$\gamma_{yp_{min}} = \hat{y}_p - \Delta_{\hat{y}_p}; \quad \gamma_{yp_{max}} = \hat{y}_p + \Delta_{\hat{y}_p}. \tag{1.24}$$

По доверительному интервалу прогноза оценивается статистическая значимость и надежность прогноза при заданном уровне значимости α.

Замечание. Корреляционно-регрессионный анализ можно осуществить в табличном процессоре *Excel.*

Пример 1. Имеются выборочные данные по 20 сельскохозяйственным предприятиям региона (таблица 2).

Требуется выполнить следующие задания:

1. Построить график зависимости между переменными, по которому необходимо подобрать модель уравнения регрессии, используя следующие наиболее часто применяемые функции:

а) линейную,
б) степенную,
в) экспоненциальную,
г) показательную,
д) равносторонней гиперболы,
е) полиномиальную.

2. Рассчитать параметры выбранного уравнения регрессии методом наименьших квадратов.

3. Оценить качество уравнения регрессии с помощью средней ошибки аппроксимации.

4. Найти коэффициент эластичности.

5. Оценить тесноту связи между переменными с помощью показателей корреляции и детерминации.

6. Оценить значимость коэффициентов корреляции и регрессии по критерию *t*-Стьюдента при уровне значимости $\alpha = 0{,}05$.

7. Охарактеризовать статистическую надежность результатов регрессионного анализа с использованием критерия *F*-Фишера при уровне значимости $\alpha = 0{,}05$.

8. Определить прогнозное значение результативного признака, если возможное значение факторного признака составит 1,25 от его среднего уровня по совокупности.

Таблица 2 – Удой молока от одной коровы и расход концентрированных кормов на одну голову

№ п/п	Расход кормов на 1 гол, ц корм. ед. (*X*)	Продуктивность, ц/гол (*Y*)
1	44,9	37,8
2	69,3	60,7
3	22,7	36,8
4	21,6	41,4
5	26,9	40,0
6	29,5	42,3
7	61,2	42,6
8	59,9	45,2
9	60,1	49,6
10	71,7	57,2
11	48,5	41,7
12	50,3	47,5
13	21,4	30,4
14	40,1	38,0
15	55,0	54,5
16	59,2	44,4
17	40,1	38,2
18	45,3	38,5
19	29,1	35,9
20	21,3	35,7

Решение

1. График зависимости переменных *X* и *Y* строится в прямоугольной системе координат. На оси абсцисс откладываются значения факторного признака *X*, а по оси ординат – результативного признака *У*.

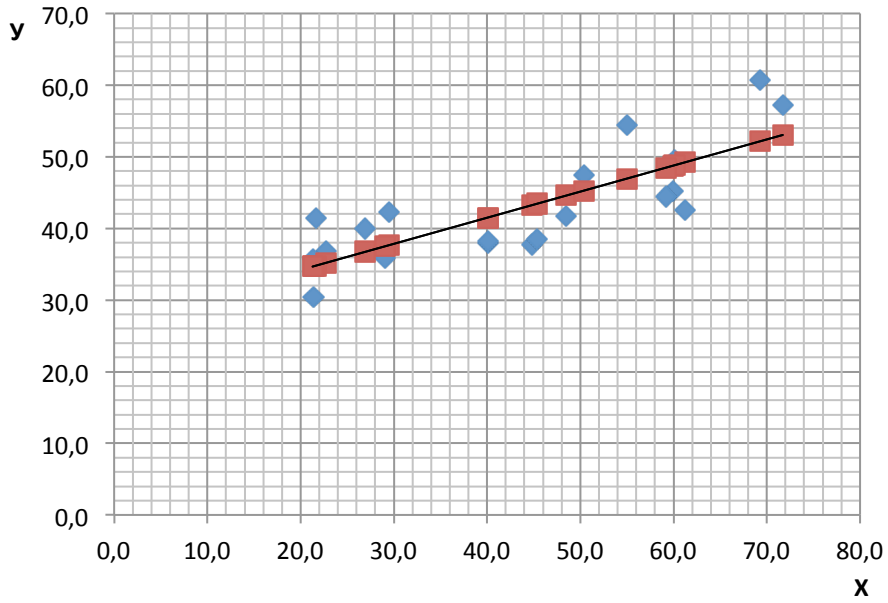

Рисунок 1 – Зависимость продуктивности коров (ц/гол)
от расхода кормов на 1 голову (ц. корм. ед.)

Характер расположения точек на графике показывает, что связь между переменными может выражаться линейным уравнением регрессии:

$$\hat{y}_x = a + bx. \tag{1.25}$$

2. Параметры уравнения регрессии находятся методом наименьших квадратов, путем составления и решения следующей системы нормальных уравнений:

$$\begin{cases} n \cdot a + b \cdot \sum x = \sum y, \\ a \cdot \sum x + b \cdot \sum x^2 = \sum xy. \end{cases} \tag{1.26}$$

Для проведения всех расчетов строится вспомогательная таблица 3, расчеты в которой могут быть проведены как без применения средств вычислительной техники, так и с ее применением.

В таблице 3 все средние значения находятся по формуле средней арифметической простой: $\bar{x} = \sum x : n.$

Таблица 3 – Результаты вычислений в *Excel*

№ п/п	X	Y	X^2	Y^2	XY	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$	$\left\|\frac{Y-\hat{Y}}{Y}\right\|\cdot 100$
1	44,9	37,8	2016,01	1428,84	1697,22	43,2835	-5,4835	30,0688	14,5066
2	69,3	60,7	4802,49	3684,49	4206,51	52,1895	8,5105	72,4286	14,0206
3	22,7	36,8	515,29	1354,24	835,36	35,1805	1,6195	2,6228	4,4008
4	21,6	41,4	466,56	1713,96	894,24	34,7790	6,6210	43,8376	15,9928
5	26,9	40	723,61	1600	1076	36,7135	3,2865	10,8011	8,2163
6	29,5	42,3	870,25	1789,29	1247,85	37,6625	4,6375	21,5064	10,9634
7	61,2	42,6	3745,44	1814,76	2607,12	49,2330	-6,6330	43,9967	15,5704
8	59,9	45,2	3588,01	2043,04	2707,48	48,7585	-3,5585	12,6629	7,8728
9	60,1	49,6	3612,01	2460,16	2980,96	48,8315	0,7685	0,5906	1,5494
10	71,7	57,2	5140,89	3271,84	4101,24	53,0655	4,1345	17,0941	7,2281
11	48,5	41,7	2352,25	1738,89	2022,45	44,5975	-2,8975	8,3955	6,9484
12	50,3	47,5	2530,09	2256,25	2389,25	45,2545	2,2455	5,0423	4,7274
13	21,4	30,4	457,96	924,16	650,56	34,7060	-4,3060	18,5416	14,1645
14	40,1	38	1608,01	1444	1523,8	41,5315	-3,5315	12,4715	9,2934
15	55	54,5	3025	2970,25	2997,5	46,9700	7,5300	56,7009	13,8165
16	59,2	44,4	3504,64	1971,36	2628,48	48,5030	-4,1030	16,8346	9,2410
17	40,1	38,2	1608,01	1459,24	1531,82	41,5315	-3,3315	11,0989	8,7212
18	45,3	38,5	2052,09	1482,25	1744,05	43,4295	-4,9295	24,3000	12,8039
19	29,1	35,9	846,81	1288,81	1044,69	37,5165	-1,6165	2,6131	4,5028
20	21,3	35,7	453,69	1274,49	760,41	34,6695	1,0305	1,0619	2,8866
Итого	878,1	858,4	43919,1	37970,3	39647,0	858,4065	–	412,6699	187,4268
Среднее значение	43,905	42,92	2195,96	1898,52	1982,35	–	–	–	9,3713

Подставим полученные суммы в систему уравнений, учитывая, что $n = 20$.

$$\begin{cases} 20 \cdot a + 878{,}1 \cdot b = 858{,}4, \\ 878{,}1 \cdot a + 43919{,}11 \cdot b = 39646{,}99. \end{cases}$$

Решив систему, например, по формулам Крамера, получим $a = 26{,}892;\ b = 0{,}365$.

Параметры уравнения регрессии также можно найти по формулам, вытекающим из системы нормальных уравнений:

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \left(\bar{x}\right)^2}, \qquad a = \bar{y} - b \cdot \bar{x}. \tag{1.27}$$

Небольшие расхождения в результатах расчетов могут происходить за счет округления средних значений во втором случае.

Таким образом, уравнение регрессии имеет вид:

$$\hat{y}_x = 26{,}892 + 0{,}365 \cdot x.$$

Замечание. Расчет данных в вспомогательной таблице можно осуществить в табличном процессоре *Excel*. Если исходные данные (x и y) приведе-

ны в виде, представленном в таблице 4, то для расчетов можно воспользоваться следующей последовательностью действий (возможны альтернативы):

1) введем для расчета остальных значений таблицы 4 следующие формулы в соответствующие ячейки (ввод – *Enter*): *D*2: =*B*2^2; *E*2: =*C*2^2; *F*2: ==*B*2**C*2;

2) выделим диапазон ячеек *B*2:*F*2 и протащим с помощью маркера заполнения до строки 21;

3) для вычисления сумм введем формулу в ячейку B22:=СУММ(*B*2:*B*21), и протащим с помощью маркера заполнения для диапазона *B*21:*F*21;

4) для расчета средних введем формулу в ячейке *B*23: =*B*22/20 и скопируем с помощью маркера заполнения для диапазона *B*23:*F*23;

5) после расчета параметров уравнения парной регрессии: *G*2: = =0,365**B*2+26,892;

6) *H*2: =*C*2-*G*2;

7) *I*2:=*H*2^2;

8) *J*2: =*ABS*(*H*2/*C*2)*100;

9) выделяется диапазон *G*2:*J*2 и с помощью маркера заполнения копируется до 21 строки;

10) для столбцов *I* и *J* находятся суммы и средние значения (см. выше).

Коэффициент регрессии показывает, при увеличении количества расхода концентрированных кормов на одну голову на 1 ц корм. ед., удой молока увеличивается в среднем на 0,365 ц/гол. Если в уравнение регрессии подставить фактические значения переменной X, то определяются возможные (теоретические) значения переменной $\hat{y}$, которые наносятся на график в виде прямой.

3. Качество уравнения регрессии оценивается с помощью средней ошибки аппроксимации:

$$\bar{A} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100 = \frac{187{,}4268}{20} = 9{,}4\ \%.$$

Значит, фактические значения продуктивности коров от расчетных значений, найденных по уравнению регрессии, в среднем различаются на 9,4%. Полученное уравнение регрессии можно оценить как достаточно хорошее.

4. Средний коэффициент эластичности при линейной форме связи находится по формуле:

$$Э = b \cdot \frac{\bar{x}}{\bar{y}},$$

где $\bar{x}$ и $\bar{y}$ – средние значения признаков.

$$Э = 0{,}365 \cdot \frac{43{,}905}{42{,}92} = 0{,}373.$$

Коэффициент эластичности показывает, что при увеличении расхода кормов на 1 % удой молока от одной коровы в среднем возрастает на 0,37 %.

5. При линейной зависимости теснота связи между переменными *X* и *У* определяется с помощью коэффициента корреляции:

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y}, \qquad (1.28)$$

где σ_x и σ_y – средние квадратические отклонения по *X* и *У*.

$$\sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2} = \sqrt{2195{,}956 - 43{,}905^2} = \sqrt{268{,}307} = 16{,}38;$$

$$\sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2} = \sqrt{1898{,}516 - 42{,}92^2} = \sqrt{56{,}39} = 7{,}509;$$

$$r = \frac{1982{,}35 - 43{,}905 \cdot 42{,}92}{16{,}38 \cdot 7{,}509} = 0{,}796.$$

Так как значение коэффициента корреляции довольно близко к единице, то между признаками связь тесная и прямая.

Коэффициент детерминации $r^2 = 0{,}796^2 = 0{,}634$ показывает, что 63,4 % различий в удое молока от одной коровы между организациями объясняется вариацией расхода концентрированных кормов, а 36,6 % другими, неучтенными факторами.

6. Так как исходные данные обычно являются выборочными, то необходимо оценить существенность или значимость величины коэффициента корреляции. Выдвигаем нулевую гипотезу: коэффициент корреляции в генеральной совокупности равен нулю и изучаемый фактор не оказывает существенного влияния на результативный признак: $H_o : r_г = 0$, при $H_1 : r_г \neq 0$.

Для проверки нулевой гипотезы применим критерий *t*–Стьюдента. Найдем наблюдаемое значение *t*–критерия:

$$t_н = |r| \sqrt{\frac{n-2}{1-r^2}} = 0{,}796 \sqrt{\frac{20-2}{1-0{,}634}} = 5{,}58.$$

Критическое значение t находится по таблицам распределения *t*-Стьюдента при уровне значимости α=0,05 и числе степеней свободы *к = n–2* = 15–2 =13 для двухсторонней критической области, $t_{кр} = 2{,}10$. Сравниваем $t_н$ с $t_{кр}$. Так как $t_н > t_{кр}$, то нулевая гипотеза отвергается, коэффициент корреляции существенно отличен от нуля в генеральной совокупности. Значит, рас-

ход концентрированных кормов оказывает статистически существенное влияние на удой молока от одной коровы.

Статистическая значимость коэффициента регрессии также проводится с использованием критерия *t*-Стьюдента.

Находится наблюдаемое значение критерия:

$$t_н = \frac{b}{m_b};\ m_b = \sqrt{\frac{\sum(y-\hat{y})^2}{(n-2)\cdot\sigma_x^2\cdot n}} = \sqrt{\frac{412{,}6699}{(20-2)\cdot 268{,}307\cdot 20}} \approx 0{,}0654.$$

$$t_н = \frac{0{,}365}{0{,}0654} = 5{,}58.$$

Критическое значение t также равно 2,10. Так как $t_н > t_{кр}$, то коэффициент регрессии статистически значим. Подтверждается вывод о значимости влияния расхода кормов на их продуктивность.

7. Статистическая надежность уравнения регрессии проверяется с использованием критерия *F*-Фишера – рассматривается нулевая гипотеза $H_o : r_г^2 = 0$, при альтернативной $H_1 : r_г^2 \neq 0$ (или нулевая гипотеза $H_o : b = 0$, при $H_1 : b \neq 0$). Наблюдаемое (фактическое) значение *F*-критерия находится по формуле:

$$F_н = \frac{\Sigma(\hat{y}-\bar{y})^2 / m}{\Sigma(y-\hat{y})^2 / (n-m-1)}, \qquad (1.29)$$

где m – число параметров при переменных X;

n – число наблюдений.

Если применяется линейное уравнение регрессии, то расчет $F_н$ упрощается.

$$F_н = \frac{r^2}{1-r^2}(n-2) = \frac{0{,}634}{1-0{,}634}\cdot 18 = 31{,}18.$$

При уровне значимости α = 0,05 и числе степеней свободы $k_1 = m = 1, k_2 = n-m-1 = 20-1-1 = 20-2 = 18$ по таблице находится критическое значение *F*-критерия. $F_{кр} = F_{\alpha=0{,}05}(k_1 = 1, k_2 = 13) = 4{,}67$. Так как $F_н > F_{кр}$, то уравнение регрессии статистически значимое или надежное.

При парной линейной зависимости оценка значимости всего уравнения, коэффициентов корреляции и регрессии дает одинаковые результаты, так как $t_b^2 = t_r^2 = F$ (наблюдаемые отличия объясняются ошибками округлений).

8. Прогнозное значение результативного признака определяется путем подстановки в уравнение регрессии прогнозного или возможного значения факторного признака (x_p).

По условию $x_p = \bar{x} \cdot 1{,}25 = 43{,}905 \cdot 1{,}25 = 54{,}88$.

Тогда прогнозное значение удоя молока составит:

$$\hat{y}_p = 26{,}895 + 0{,}365 \cdot 54{,}88 = 46{,}926.$$

Значит, при расходе концентрированных кормов на 1 голову в 54,9 ц корм. ед. возможная продуктивность молока составит 46,9 ц/гол.

Контрольная работа № 2.
Множественный корреляционно-регрессионный анализ

В статистических исследованиях результативный признак Y формируется, как правило, под влиянием не одного, а нескольких факторных признаков $X_1, X_2, \ldots, X_p$. Уравнение множественной регрессии в таком случае имеет вид:

$$y = f(x_1, x_2, \ldots, x_p, e).$$

В зависимости от вида функции используются как линейные, так и нелинейные модели. Линейная модель множественной регрессии с несколькими переменными имеет вид:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e, \quad (2.1)$$

где $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ – параметры модели,
e – случайные отклонения или остаток,
p – количество переменных.

Наиболее распространенным методом оценивания параметров линейных эконометрических моделей является метод наименьших квадратов. Его идея сводится к выбору таких значений оценок $b_0, b_1, \ldots, b_p$ структурных параметров $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$, при которых сумма квадратов отклонений наблюдаемых значений зависимой переменной (y_i) от ее теоретических значений $(\hat{y}_i)$, рассчитанных по уравнению регрессии в натуральном масштабе $\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}$, оказывается наименьшей. Это условие записывается в виде:

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2 \rightarrow min. \quad (2.2)$$

Чтобы найти экстремум функции, необходимо определить частные производные по параметрам и приравнять их к нулю:

$$\frac{\partial S}{\partial b_0} = 0, \ \frac{\partial S}{\partial b_1} = 0, \ldots, \frac{\partial S}{\partial b_p} = 0. \quad (2.3)$$

Параметры линейного уравнения множественной регрессии находятся путем составления и решения следующей системы нормальных уравнений:

$$\begin{cases} \Sigma y = nb_o + b_1\Sigma x_1 + b_2\Sigma x_2 + \ldots + b_p\Sigma x_p, \\ \Sigma yx_1 = b_o\Sigma x_1 + b_1\Sigma x_1^2 + b_2\Sigma x_1x_2 + \ldots + b_p\Sigma x_1x_p, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \Sigma yx_p = b_o\Sigma x_p + b_1\Sigma x_1x_p + b_2\Sigma x_2x_p + \ldots + b_p\Sigma x_p^2. \end{cases} \quad (2.4)$$

В матричной форме уравнение множественной регрессии имеет вид:

$$Y = XB + \varepsilon, \quad (2.5)$$

где Y – вектор-столбец размерности (n×1);

X – матрица наблюдаемых значений факторных признаков размерности (n × (p+1));

B – вектор-столбец коэффициентов регрессии ((p+1)×1);

ε – вектор-столбец остаточных значений результативного признака размерности (n ×1).

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; B = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix}; \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}; X = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \ldots & \ldots & \ldots & \ldots & \cdots \\ 1 & x_{1n} & x_{2n} & \ldots & x_{pn} \end{bmatrix}. \quad (2.6)$$

Тогда оценки параметров линейной модели b находятся из выражения:

$$b = (X^TX)^{-1}X^TY, \quad (2.7)$$

где X^T – транспонированная матрица X;

$(X^TX)^{-1}$ – обратная матрица.

Множественный коэффициент регрессии (b_j) показывает, на сколько единиц изменяется в среднем результативный признак $У$, если j-й факторный признак X_j увеличить на единицу при условии, что все другие факторы в линейной модели закреплены на постоянном, обычно среднем, уровне.

Уравнение множественной регрессии может быть построено в стандартизованном масштабе, когда единицей измерения признаков принимается их среднее квадратическое отклонение:

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \cdots + \beta_p t_{x_p}. \quad (2.8)$$

$$t_y = \frac{y-\bar{y}}{\sigma_y},\ t_{x_j} = \frac{x_j-\bar{x}_j}{\sigma_{x_j}},$$

где β_j – стандартизованные коэффициенты регрессии.

Параметры уравнения регрессии определяются методом наименьших квадратов путем составления и решения следующей системы уравнений:

$$\begin{cases} r_{yx_1} = \beta_1 + \beta_2 r_{x_1x_2} + \cdots + \beta_p r_{x_1x_p} \\ r_{yx_2} = \beta_1 r_{x_1x_2} + \beta_2 + \cdots + \beta_p r_{x_2x_p} \\ \cdots \\ \cdots \\ r_{yx_p} = \beta_1 r_{x_1x_p} + \beta_2 r_{x_2x_p} + \cdots + \beta_p \end{cases} \quad (2.9)$$

Зная стандартизованные коэффициенты можно получить множественные коэффициенты регрессии:

$$b_j = \beta_j \cdot \frac{\sigma_y}{\sigma_{x_j}},\quad b_0 = \overline{y} - b_1\overline{x}_1 - b_2\overline{x}_2 - \cdots - b_p\overline{x}_p. \quad (2.10)$$

По абсолютной величине -коэффициентов судят об относительной силе влияния факторов на изменение результативного признака. Для характеристики силы влияния факторов на результативный признак используется также коэффициент эластичности, который представляет отношение прироста результативного признака У к приросту факторного признака X_j:

$$Э_{x_j} = \frac{dy}{\hat{y}} : \frac{dx_j}{x_j} = \frac{dy}{dx_j} \cdot \frac{x_j}{\hat{y}}\ . \quad (2.11)$$

По линейной модели множественной регрессии коэффициент эластичности определяется по формуле:

$$Э_{x_j} = b_j \frac{x_j}{b_0 + b_1x_{i1} + b_2x_{i2} + \cdots + b_px_{ip}}. \quad (2.12)$$

Если в этой формуле значения факторов взять на среднем уровне, то будет получен средний коэффициент эластичности, который показывает, на сколько процентов в среднем изменится результативный признак, если *j*-й фактор увеличить на один процент, при условии что все другие факторы закреплены на среднем уровне.

$$\overline{Э}_{x_j} = b_j \frac{\overline{x}_j}{\overline{y}}. \quad (2.13)$$

Для оценки тесноты связи между признаками применяются парные, частные и множественные коэффициенты (индексы) корреляции и детерминации.

Множественный коэффициент (индекс) корреляции ($R_{yx_1x_2\ldots x_p}$) характеризует совместное влияние всех факторов, включенных в уравнение регрессии. Он рассчитывается по следующим формулам:

$$R_{yx_1x_2\ldots x_p} = \sqrt{1-\frac{\sigma^2_{ост.}}{\sigma^2_y}} = \sqrt{\frac{\sigma^2_{рег.}}{\sigma^2_y}} = \sqrt{1-\frac{\sum_i^n (y-\hat{y}_{x_1x_2\ldots x_p})^2}{\sum_i^n (y-\overline{y})^2}} = \sqrt{1-\frac{SS_{ост.}}{SS_{общ.}}}, \qquad (2.14)$$

где σ^2_y – общая дисперсия результативного признака,

$\sigma^2_{рег.}$ – дисперсия, объяснимая регрессией,

$\sigma^2_{ост.}$ – остаточная дисперсия, причем

$$\sigma^2_y = \sigma^2_{рег.} + \sigma^2_{ост.};\ \sigma^2_y = \frac{\sum(y-\overline{y})^2}{n};\ \sigma^2_{рег.} = \frac{\sum(\hat{y}-\overline{y})^2}{n};\ \sigma^2_{ост.} = \frac{\sum(y-\hat{y})^2}{n}.$$

$$SS_{общ.} = SS_{факт.} + SS_{ост.};$$

$$\sum(y-\overline{y})^2 = \sum(\hat{y}-\overline{y})^2 + \sum(y-\hat{y})^2, \qquad (2.15)$$

где $SS_{общ.}$ – общая сумма квадратов отклонений результативного признака;

$SS_{факт.}$ – факторная сумма квадратов отклонений;

$SS_{ост.}$ – остаточная сумма квадратов отклонений.

Квадрат множественного коэффициента (индекса) корреляции называется множественным коэффициентом (индексом) детерминации. Он показывает, какая часть вариации результативного признака объясняется влиянием факторов, включенных в уравнение регрессии. Если используется линейное уравнение множественной регрессии в стандартизованном масштабе, то множественный коэффициент детерминации рассчитывается по формуле:

$$R^2_{yx_1x_2\ldots x_p} = \beta_1 r_{yx_1} + \beta_2 r_{yx_2} + \ldots + \beta_p r_{yx_p} = \sum \beta_j r_{x_j}. \qquad (2.16)$$

Частные коэффициенты корреляции, характеризующие тесноту связи между фактором x_j и результативным признаком, при исключении влияния других факторов, включенных в модель, определяется по формулам:

$$r_{yx_j \cdot x_1 x_2 \ldots x_{j-1} x_{j+1} \ldots x_p} = \sqrt{1 - \frac{1 - R^2_{yx_1 x_2 \ldots x_j \ldots x_p}}{1 - R^2_{yx_1 x_2 \ldots x_{j-1} x_{j+1} \ldots x_p}}}, \quad (2.17)$$

$$r_{yx_j \cdot x_1 x_2 \ldots x_{j-1} x_{j+1} \ldots x_p} =$$

$$= \frac{r_{yx_j \cdot x_1 x_2 \ldots x_{j-1} x_{j+1} \ldots x_{p-1}} - r_{yx_p \cdot x_1 x_2 \ldots x_{j-1} x_{j+1} \ldots x_{p-1}} \cdot r_{x_j x_p \cdot x_1 \ldots x_{j-1} x_{j+1} \ldots x_{p-1}}}{\sqrt{(1 - r^2_{yx_p \cdot x_1 x_2 \ldots x_{j-1} x_{j+1} \ldots x_{p-1}})(1 - r^2_{x_j x_p \cdot x_1 \ldots x_{j-1} x_{j+1} \ldots x_{p-1}})}} \quad (2.18)$$

В формуле частные коэффициенты корреляции j-го порядка рассчитываются через частные коэффициенты корреляции ($j–1$)-го порядка. Значения частных коэффициентов корреляции изменяются от –1 до 1. Они могут быть использованы при отсеве несущественно влияющих факторов.

С учетом поправки на число степеней свободы рассчитывается скорректированный коэффициент (индекс) множественной корреляции:

$$R^2_{ск} = 1 - \frac{\sum (y - \hat{y})^2 : (n - m - 1)}{\sum Y - \overline{y})^2 : (n - 1)}, \quad (2.19)$$

$$R^2_{ск} = 1 - (1 - R^2) \cdot \frac{(n-1)}{(n-m-1)}, \quad (2.20)$$

где m – число параметров уравнения регрессии без учета свободного члена.

В линейном уравнении m = p.

Оценка значимости множественного уравнения регрессии производится с помощью F-критерия Фишера-Снедекора.

Определяется наблюдаемое значение критерия по следующей формуле:

$$F_н = \frac{SS_{факт.}}{m} : \frac{SS_{ост.}}{n-m-1} = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}. \quad (2.21)$$

При заданном уровне значимости α и числе степеней свободы факторной (k_1) и остаточной дисперсий (k_2) по таблицам находится критическое значение критерия Фишера-Снедекора. Сравнивается наблюдаемое и критическое значения критерия. Если $F_н < F_{кр}$, то нулевая гипотеза о незначимости уравнения регрессии принимается. Если $F_н > F_{кр}$, то нулевая гипотеза отвергается и принимается альтернативная гипотеза о статистической значимости всего уравнения регрессии.

Оценка значимости параметров множественного линейного уравнения регрессии производится с помощью t-критерия Стьюдента. Выдвигается основная гипотеза о равенстве нулю параметров уравнения регрессии

($H_0: \beta_j = 0$), при конкурирующей гипотезе, что параметры уравнения отличны от нуля ($H_0: \beta_j \neq 0$). Наблюдаемое значение *t*-критерия для параметра уравнения b_j определяется по формуле:

$$t_{b_j} = \frac{b_j}{s_{b_j}}, \quad s_{b_j} = \sqrt{\frac{SS_{ост.}}{n-p-1}\left[(X^T X)^{-1}\right]_{jj}}, \tag{2.22}$$

где s_{b_j} – стандартная ошибка параметра уравнения регрессии b_j,

$\left[(X^T X)^{-1}\right]_{jj}$ – диагональный элемент матрицы $(X^T X)^{-1}$.

Стандартная ошибка множественного коэффициента регрессии b_j может быть найдена также по формуле:

$$s_{b_j} = \frac{\sigma_y}{\sigma_{x_j}} \sqrt{\frac{1-R^2_{yx_1x_2\ldots x_p}}{\left(1-R^2_{x_j x_1 x_2 \ldots x_{j-1} x_{j+1} \ldots x_p}\right)(n-m-1)}}, \tag{2.23}$$

где σ_y – среднее квадратическое отклонение результативного признака;

σ_{x_j} – среднее квадратическое отклонение факторного признака x_j.

Критическое значение t находится по таблице значений *t*-критерия Стьюдента.

При уровне значимости α и числе степеней свободы $k = n - m - 1$.

Если $\left|t_{b_j}\right| > \left|t_{кр}\right|$, то параметр уравнения статистически значим.

Если $\left|t_{b_j}\right| < \left|t_{кр}\right|$, то параметр уравнения статистически не значим и *j*-ая переменная исключается из уравнения регрессии.

Доверительные интервалы для коэффициентов линейного уравнения регрессии находятся по формуле:

$$b_j \pm t_{кр} s_{b_j}. \tag{2.24}$$

Условия, необходимые для получения несмещенных, состоятельных эффективных оценок, представляют собой предпосылки метода наименьших квадратов, соблюдение которых желательно для получения достоверных результатов регрессии:

1) остаток является случайной величиной;

2) математическое ожидание случайного остатка равно 0:

$$M(\varepsilon_i) = 0, i = 1,2, \dots, n;$$

3) дисперсия случайных остатков постоянна для любого наблюдения $D(\varepsilon_i) = \sigma^2$, это предположение называется условием гомоскедастичности;

4) случайные остатки ε_i *и* ε_j не коррелированы,

$$cov(\varepsilon_i, \varepsilon_j) = 0,\ i \neq j;$$

5) остатки распределены по нормальному закону.

Первые 4 условия известны как условия Гаусса-Маркова.

При построении уравнения множественной регрессии обычно используются следующие нелинейные функции:

степенная $y = b_o \cdot x_1^{b_1} \cdot x_2^{b_2} \dots x_p^{b_p} \cdot \varepsilon$; (2.25)

экспонента $y = e^{b_o + b_1 x_1 + b_2 x_2 + \dots + b_p x_p + \varepsilon}$; (2.26)

гипербола $y = b_0 + \frac{b_1}{x_1} + \frac{b_2}{x_2} + \dots + \frac{b_p}{x_p} + \varepsilon$; (2.27)

логлинейная $ln\ y = b_0 + b_1 ln x_1 + b_2 ln x_2 + \dots + b_p ln x_p + \varepsilon.$ (2.28)

Довольно часто применяются и другие виды функций, например:

$$y = b_o + b_1 x_1 + b_2 x_2 + b_3 x_1^2 + b_4 x_2^2 + b_5 x_1 x_2 + \varepsilon.$$ (2.27)

Если уравнение регрессии нелинейное, то оно вначале приводится путем соответствующего преобразования к линейному виду.

Пример 2. Исследовать влияние расхода кормов на 1 голову и затрат труда на 1 голову на удой молока от одной коровы по сельскохозяйственным организациям региона.

Результативным признаком (*Y*) - Удой молока от одной коровы, ц/гол.

Факторные признаки: X_1 – Расход кормов на 1 гол, ц корм. ед.; X_2 – Затраты труда на 1 гол, ч – час.

Требуется определить:

1) параметры множественного уравнения регрессии в натуральной и стандартизованной форме;

2) средние коэффициенты эластичности для каждого фактора;

3) коэффициенты частной и множественной корреляции;

4) общий и частные критерии *F*-Фишера.

Решение

Связь между результативным признаком Y и факторами X_1 *и* X_2 выражается множественным линейным уравнением регрессии, которое имеет вид:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2. \quad (2.28)$$

Рассмотрим применение пакета анализа данных в *Excel MS Office* для решения задачи. Исходные данные введем на листе *MS Excel* в виде, представленном таблице 4.

Таблица 4 – Исходные данные для регрессионного анализа в *MS Excel*

№ п/п	Удой молока от одной коровы, ц/гол (*У*)	Расход кормов на 1 гол, ц корм. ед. (X_1)	Затраты труда на 1 гол, ч - час (X_2)
1	37,8	44,9	3,02
2	60,7	69,3	6,87
3	36,8	22,7	2,55
4	41,4	21,6	4,82
5	40,0	26,9	4,55
6	42,3	29,5	1,92
7	42,6	61,2	3,29
8	45,2	59,9	6,81
9	49,6	60,1	6,77
10	57,2	71,7	5,84
11	41,7	48,5	3,16
12	47,5	50,3	5,93
13	30,4	21,4	1,54
14	38,0	40,1	2,83
15	54,5	55,0	5,97
16	44,4	59,2	4,99
17	38,2	40,1	1,58
18	38,5	45,3	5,02
19	35,9	29,1	4,04
20	35,7	21,3	2,93

Для проведения анализа предварительно установим пакет анализа, выполнив последовательно действия: кнопка *Office – Параметры Excel – Надстройки – Пакет анализа – Перейти* (выделим в окне доступных надстроек *Пакет анализа*), после этого во вкладке *Данные* ленты появится инструмент *Пакет анализа*.

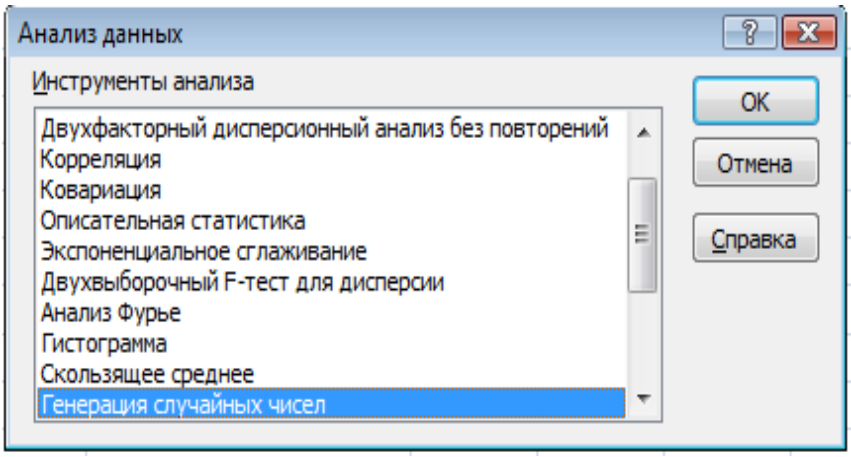

Рисунок 2 – Анализ данных

Выберем в *Пакете анализа* инструмент *Описательная статистика* и заполним параметры диалогового окна (рисунок 3).

В результате будут рассчитаны обобщающие характеристики по каждому признаку (таблица 5).

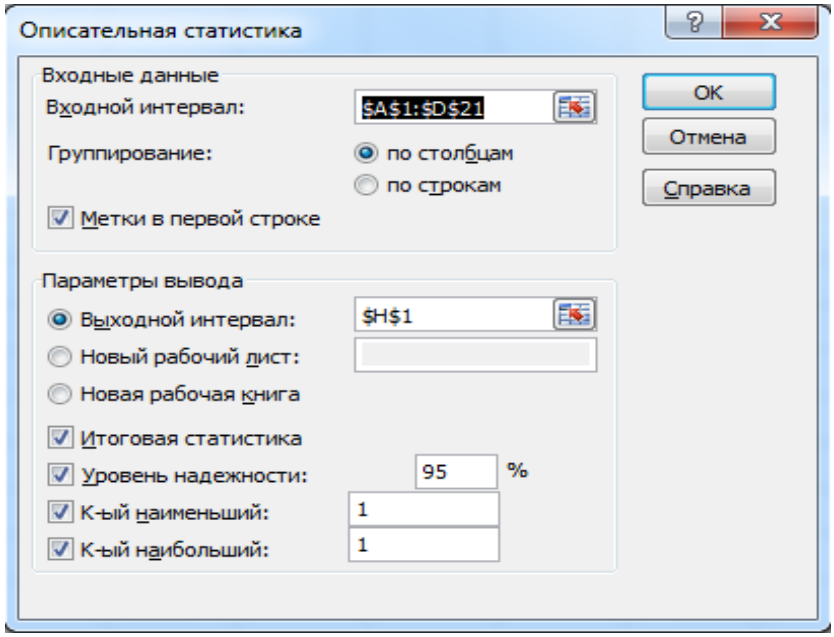

Рисунок 3 – Диалоговое окно описательной статистики

Данные таблицы 5 показывают, что по совокупности предприятий средний удой от одной коровы составил 42,9 ц/гол и в среднем между предприятиями продуктивность коров колеблется в границах 42,9 ± 7,7 ц/гол, т. е. от 35,2 до 50,6 ц/гол. Коэффициент вариации составил 17,9 %, что свидетельствует о больших различиях в удое молока от одной коровы между предприятиями. По значению медианы видно, что половина предприятий имеет размер удоя до 41,5 ц/гол, а половина более. Распределение предприятий по данному признаку является несимметричными (*Ка* = 0,909) и остро-

вершинным (Э = – 0,471). Наименьшее значение удоя молока от одной коровы составило 30,4, а наибольшее – 60,7 ц/гол.

Таблица 5 – Обобщающие характеристики исследуемых признаков по совокупности сельскохозяйственных организаций

Показатель	*У*	X_1	X_2	Принятые обозначения
Среднее значение	42,921	43,908	4,222	$\overline{X} = \sum x_i n_i / n$
Стандартная ошибка	1,722	3,760	0,397	$s_{\overline{X}} = s / \sqrt{n}$
Медиана	41,548	45,108	4,295	M_e
Мода	н/д	н/д	н/д	M_o
Стандартное отклонение	7,700	16,816	1,774	s
Дисперсия выборки	59,289	282,768	3,148	$s^2 = \sum (x_i - \overline{X})^2 n_i / (n-1)$
Эксцесс	0,471	- 1,321	- 1,289	$Ex = \sum ((x_i - \overline{X})/S)^4 n_i / n - 3$
Асимметричность	0,909	0,022	0,053	$Sk = \sum ((x_i - \overline{X})/S)^3 n_i / n$
Интервал	30,3	50,426	5,33	$W = x_{max} - x_{min}$
Минимум	30,4	21,3	1,54	x_{min}
Максимум	60,7	71,7	6,87	x_{max}
Сумма	858,42	878,2	84,43	Σx_i
Счет	20	20	20	$n = \Sigma n_i$
Наибольший (1)	60,7	71,7	6,87	-
Наименьший (1)	30,4	21,3	1,54	-
Уровень надежности (95,0%)	3,61	7,87	0,83	$\Delta = t_{\alpha;n-1} s_{\overline{X}}$

Аналогичные выводы можно сделать и по факторным признакам X_1 *и* X_2.

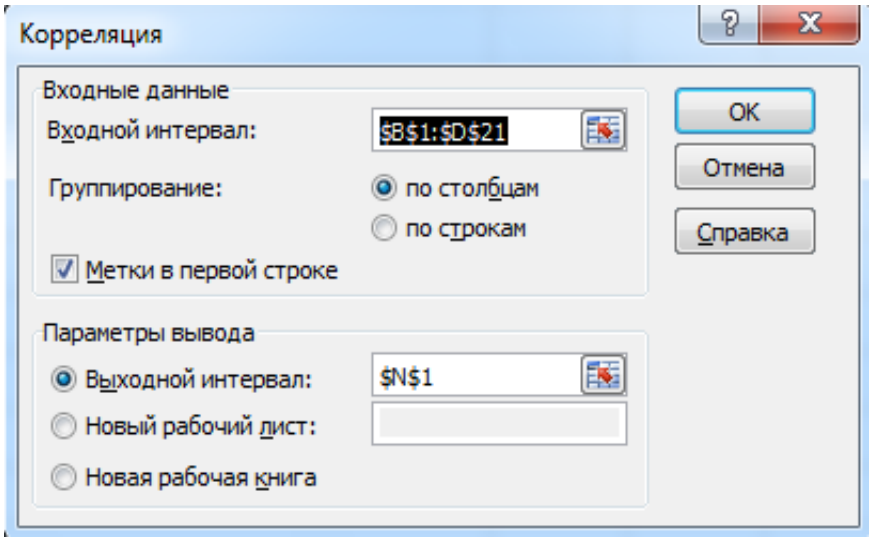

Рисунок 4 – Диалоговое окно «Корреляция»

Для нахождения парных коэффициентов корреляции применим инструмент пакета анализа *Корреляция*, для этого заполним параметры диалогового окна как на рисунке 4.

В результате будет получена матрица парных коэффициентов корреляции между всеми изучаемыми переменными (таблица 6).

Таблица 6 – Парные коэффициенты корреляции между признаками

	У	X_1	X_2
У	1	0,7963	0,7692
X_1	0,7963	1	0,6586
X_2	0,7692	0,6586	1

Значит: r_{yx_1} = 0,7963; r_{yx_2}= 0,7692; $r_{x_1x_2}$= 0,6586. Парные коэффициенты корреляции показывают, что связь между удоем молока от одной коровы, расходом концентрированных кормов на 1 гол и затратами труда на 1 гол довольно тесная, а между факторными признаками X_1 и X_2 – средняя.

Линейное уравнение множественной регрессии в натуральной форме имеет вид:

$$y = b_o + b_1x_1 + b_2x_2.$$

Найдем параметры этого уравнения, используя инструмент *Пакета анализа – Регрессия.* Заполним параметры диалогового окна (рисунок 5).

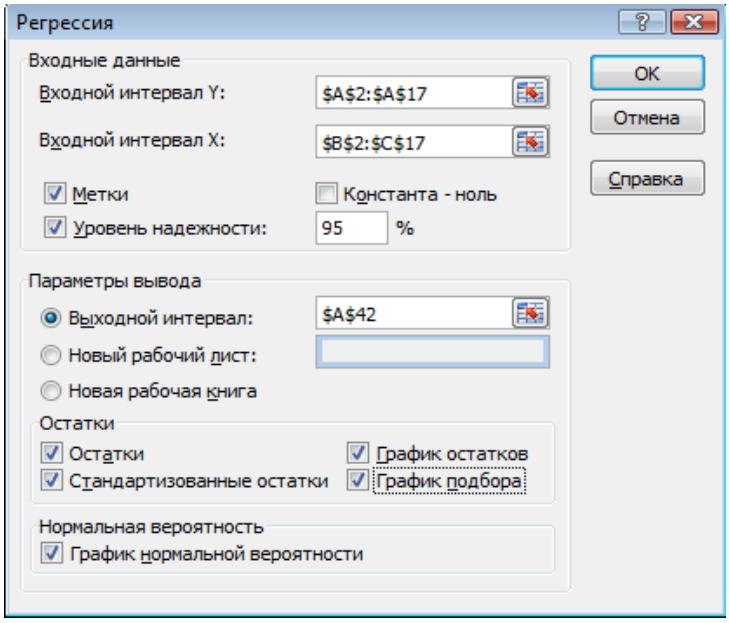

Рисунок 5 – Диалоговое окно «Регрессия»

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,8602
R-квадрат	0,7399
Нормированный R-квадрат	0,7093
Стандартная ошибка	4,1536
Наблюдения	20

Дисперсионный анализ						
	df	*SS*	*MS*	*F*	*Значимость F*	
Регрессия	2	834,4994099	417,249705	24,18487	1,0669E-05	
Остаток	17	293,2925901	17,2525053			
Итого	19	1127,792				
Коэффициенты		*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	24,699	2,78480	8,8693	8,718E-08	18,8239	30,5747
X1	0,2345	0,0753	3,1125	0,0063	0,0756	0,3934
X2	1,8770	0,71356	2,6304	0,0175	0,3715	3,3825

Рисунок 6 – Вывод итогов регрессионного анализа

Получим линейное уравнение множественной регрессии:

$$y = 24{,}699 + 0{,}234x_1 + 1{,}877x_2.$$

Коэффициенты множественной регрессии показывают, что при увеличении расхода концентрированных кормов на 1 гол на 1 ц корм ед. удой молока от одной коровы в среднем увеличивается на 0,234 ц/гол (при исключении влияния второго фактора X_2), а при росте затрат труда на 1 гол на 1 ч-час он в среднем возрастает на 1,877 ц/гол

В стандартизованной форме уравнение регрессии имеет вид:

$$t_y = \beta_1 \cdot t_{x_1} + \beta_2 t_{x_2},\ t_y = \frac{y - \overline{y}}{\sigma_y}; t_{x_1} = \frac{x_1 - \overline{x}_1}{\sigma_{x_1}}; t_{x_2} = \frac{x_2 - \overline{x}_1}{\sigma_{x_1}}.$$

Найдем β-коэффициенты, используя их связь с коэффициентами b_j уравнения регрессии в нормальной форме:

$$\beta_j = b_j \frac{\sigma_{x_j}}{\sigma_y},$$

$$\beta_1 = 0{,}234 \cdot \frac{16{,}816}{7{,}7} = 0{,}512,$$

$$\beta_2 = 1{,}877 \cdot \frac{1{,}774}{7{,}7} = 0{,}432.$$

B-коэффициенты, можно также найти с помощью парных коэффициентов корреляции по формулам:

$$\beta_1 = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{1 - r^2_{x_1x_2}} = \frac{0{,}7963 - 0{,}7692 \cdot 0{,}6568}{1 - 0{,}6568^2} = 0{,}512,$$

$$\beta_2 = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{1 - r^2_{x_1x_2}} = \frac{0{,}7692 - 0{,}7963 \cdot 0{,}6568}{1 - 0{,}6568^2} = 0{,}432.$$

Линейное уравнение множественной регрессии в стандартизованном масштабе имеет вид:

$$t_y = 0{,}512 t_{x_1} + 0{,}432\, t_{x_2}.$$

По абсолютной величине β-коэффициентов можно сделать вывод об относительной силе влияния факторов на изменение результативного признака. Видно, что на продуктивность коров большее влияние оказывает расход концентрированных коромов и меньшее – затраты труда.

2. Средние коэффициенты эластичности находятся по формуле:

$$Эyx_1 = b_j \cdot \frac{\bar{x}_j}{\bar{y}},$$

$$Эyx_1 = b_1 \cdot \frac{\bar{x}_1}{\bar{y}} = 0{,}234 \cdot \frac{43{,}908}{42{,}921} = 0{,}239,$$

$$Эyx_2 = b_2 \cdot \frac{\bar{x}_2}{\bar{y}} = 1{,}877 \cdot \frac{4{,}222}{42{,}921} = 0{,}185.$$

Значит, при увеличении расхода кормов на 1 % удой молока увеличивается в среднем на 0,239 %, исключив влияние второго фактора. Если увеличить затраты труда на 1 гол на 1 %, то удой молока от одной коровы в среднем возрастет на 0,185 %, исключив влияние первого фактора.

3. Коэффициенты частной корреляции определяются через парные коэффициенты корреляции по формулам:

$$r_{yx_1\text{-}x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{(1 - r^2_{yx_2})(1 - r^2_{x_1x_2})}} = \frac{0{,}7963 - 0{,}7692 \cdot 0{,}6568}{\sqrt{(1 - 0{,}7692^2)(1 - 0{,}6568^2)}} = 0{,}604;$$

$$r_{yx_2\text{-}x_1} = \frac{r_{yx_2} - r_{yx_1}r_{x_1x_2}}{\sqrt{(1-r_{yx_1}^2)(1-r_{x_1x_2}^2)}} = \frac{0{,}7692 - 0{,}7963 \cdot 0{,}6568}{\sqrt{(1-0{,}7963^2)(1-0{,}6568^2)}} = 0{,}540;$$

$$r_{x_1x_2\text{-}y} = \frac{r_{x_1x_2} - r_{yx_1}r_{yx_2}}{\sqrt{(1-r_{yx_1}^2)(1-r_{yx_2}^2)}} = \frac{0{,}6568 - 0{,}7963 \cdot 0{,}7692}{\sqrt{(1-0{,}7963^2)(1-0{,}7692^2)}} = 0{,}115.$$

Коэффициенты частной корреляции характеризуют тесноту связи между двумя переменными, исключив влияние третьей переменной. Значит, связь между расходами концентрированных кормов и удоем молока довольно тесная и прямая, между затратами труда на 1 гол и продуктивностью коров прямая и средняя. А связь между факторами x_1 и x_2 слабая и также прямая.

Коэффициент множественной корреляции находится по формуле:

$$R_{yx_1x_2} = \sqrt{\beta_1 r_{yx_1} + \beta_2 r_{yx_2}} = \sqrt{0{,}512 \cdot 0{,}7963 + 0{,}432 \cdot 0{,}7692} = \sqrt{0.408 + 0.332} = \sqrt{0{,}74} = 0{,}86.$$

Величина коэффициента множественной корреляции показывает, что связь между продуктивностью коров и обоими факторами очень тесная, причем 73,9 % вариации продуктивности объясняется влиянием обоих факторов, из которой на долю первого приходится 36,5 % вариации, а второго – 29,2 %.

4. Оценим значимость уравнения регрессии и множественного коэффициента детерминации R^2 с помощью критерия F-Фишера. Фактически рассматривается нулевая гипотеза $H_0: R^2=0, (b_1=b_2=0)$ и альтернативная гипотеза $H_1: R^2\neq 0, (b_1\neq 0, b_2\neq 0)$.

Наблюдаемое значение критерия находится по формуле:

$$F_н = \frac{R^2_{yx_1x_2}}{1-R^2_{yx_1x_2}} : \frac{m}{n-m-1},$$

где m – число факторов в линейном уравнении регрессии;

n – число единиц наблюдения.

$$F_н = \frac{0{,}737}{1-0{,}737} : \frac{2}{20-2-1} = 24{,}19.$$

При уровне значимости α = 0,05 и числе степеней свободы $k_1=m=2$, $k_2=n-m-1=20-2-1=17$, по таблице значений критерия F-Фишера критическое значения составляет 3,59, т. е. $F_{кр}=3{,}59$. Сравниваем $F_н$ с $F_{кр}$. Так как $F_н > F_{кр}$,

то нулевую гипотезу о незначимости величины R^2 отклоняем, т.е. уравнение множественной регрессии и R^2 статистически значимы.

В уравнении множественной регрессии не все факторы могут оказывать статистически существенное влияние на изменение результативного признака. Оценка значимости факторов в уравнении регрессии может быть дана с помощью частного *F*-критерия или критерия *t*-Стьюдента.

$$F_{нx_1} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1 - R^2_{yx_1x_2}} \cdot \frac{n-m-1}{1} = \frac{0{,}74-0{,}5917}{1-0{,}74} \cdot \frac{20-2-1}{1} = 9{,}70.$$

При α = 0,05, $k_1 = 1$, $k_2 = 17$, $F_{кр} = 4{,}45$. Так как $F_{нx_1} > F_{кр}$, то в уравнение регрессии целесообразно включение фактора X_1 после X_2. Фактор X_1 оказывает статистически значимое влияние на *У*.

$$F_{нx_2} = \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1 - R^2_{yx_1x_2}} \cdot \frac{n-m-1}{1} = \frac{0{,}737-0{,}6341}{1-0{,}74} \cdot \frac{20-2-1}{1} = 6{,}92.$$

В этом случае также наблюдаемое значение критерия Фишера больше критического, это свидетельствует о статистической значимости влияния фактора X_2 и целесообразности включения его в уравнение множественной регрессии. В данной задаче на удой молока статистически значимое влияние оказывают оба фактора. Небольшие расхождения в результатах расчетов в компьютерном и ручном вариантах расчетов обусловлено округлением расчетных значений.

Контрольная работа № 3.
Временные ряды

Явления, их связи и зависимости могут рассматриваться как в пространстве, так и во времени, путем построения и анализа одного или нескольких временных рядов.

Временной ряд – это совокупность числовых значений изучаемого показателя в последовательные моменты или периоды времени. Он состоит из значений или уровней временного ряда (Y) и периодов или моментов времени (t). Первый член временного ряда называют начальным (y_1), а последний конечным (y_n). Тогда временной ряд имеет вид:

$$\begin{array}{l} t: \ 1 \quad 2 \quad 3 \ldots n \\ Y_t: \ Y_1 \ Y_2 \ Y_3 \ldots Y_n \end{array} \qquad (3.1)$$

Модели, построенные по данным, характеризующим один объект за ряд последовательных моментов или периодов времени, называются моделями временных рядов.

Уровни временного ряда формируются под воздействием большого числа факторов, которые условно можно подразделить на три группы:

– факторы, формирующие тенденцию изменения уровней временного ряда – трендовая компонента (T);

– факторы, формирующие циклические или сезонные колебания уровней ряда – циклическая компонента (S);

– случайные факторы – случайная компонента (ε).

Модель, в которой временной ряд представлен как сумма перечисленных выше компонент, называется аддитивной моделью временного ряда ($Y=T+S+\varepsilon$).

Если временной ряд представлен как произведение компонент, то она называется мультипликативной моделью временного ряда ($Y=T\cdot S\cdot \varepsilon$).

Основная задача эконометрического исследования отдельного временного ряда – выявление и количественная оценка каждой из компонент с целью использования полученной информации для анализа и прогнозирования будущих значений ряда.

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

несколько периодов или моментов времени, называемого коэффициентом автокорреляции.

Коэффициент автокорреляции уровней ряда первого порядка, смещенных на одну единицу времени, определяется по формуле:

$$r_1 = \frac{\sum_{t=2}^{n}(y_t - \overline{y}_1)\cdot(y_{t-1}\overline{y}_2)}{\sqrt{\sum_{t=2}^{n}(y_t - \overline{y}_1)^2 \cdot \sum_{t=2}^{n}(y_{t-1} - \overline{y}_2)^2}},$$

где $\overline{y}_1 = \frac{\sum_{t=2}^{n} y_t}{n-1}$; $\overline{y}_2 = \frac{\sum_{t=2}^{n} y_{t-1}}{n-1}$; (3.2)

Коэффициент автокорреляции уровней ряда второго порядка:

$$r_2 = \frac{\sum_{t=3}^{n}(y_t - \overline{y}_3)\cdot(y_{t-2} - \overline{y}_4)}{\sqrt{\sum_{t=3}^{n}(y_t - \overline{y}_3)^2 \cdot \sum_{t=3}^{n}(y_{t-2} - \overline{y}_4)^2}}, \quad (3.3)$$

где $\overline{y}_3 = \frac{\sum_{t=3}^{n} y_t}{n-2}$; $\overline{y}_4 = \frac{\sum_{t=3}^{n} y_{t-2}}{n-2}$.

Аналогично можно определить коэффициенты автокорреляции более высоких порядков.

Так как коэффициент автокорреляции строится по аналогии с линейным коэффициентом корреляции, то по нему можно судить о наличии линейной или близкой к линейной тенденции. Чем ближе коэффициент автокорреляции первого порядка к единице, тем более выражена линейная тенденция. Для некоторых временных рядов, имеющих сильную нелинейную тенденцию, коэффициент автокорреляции уровней исходного ряда может приближаться к нулю.

Последовательность значений коэффициентов автокорреляции уровней первого, второго и т. д. порядков называют автокорреляционной функцией временного ряда. Если наиболее высоким оказался коэффициент автокорреляции первого порядка, исследуемый ряд содержит только тенденцию. Если наиболее высоким оказался коэффициент автокорреляции порядка τ, то ряд содержит циклические или сезонные колебания с периодичностью в τ моментов времени. Если ни один коэффициент не является значимым, можно сделать вывод о том, что либо ряд не содержит тенденции и циклических колебаний, либо содержит сильную нелинейную тенденцию.

Число периодов или моментов времени, по которым рассчитывается коэффициент автокорреляции, называют лагом.

Построение аналитической функции для моделирования тенденции (тренда) временного ряда называют аналитическим выравниванием времен-

ного ряда. Тенденция во времени может принимать разные формы, для ее формализации используют следующие функции:

- линейная: $y_t = a + bt$;
- степенная: $y_t = at^b$;
- гипербола: $y_t = a + b/t$;
- экспонента: $y_t = e^{a=bt}$;
- показательная: $y_t = ab^t$;
- полином k– ого порядка: $y_t = a + b_1t + b_2t^2 + \ldots + b_k t^k$;
- логическая: $y_t = \frac{1}{1+be^{-ct}}$;
- Гомперца: $log_c f(x) = a - bc^t$, где 0<c<1.

Параметры каждой из перечисленных выше функций определяются методом наименьших квадратов, используя в качестве независимой переменной время (t), а в качестве зависимой переменной – фактические уровни временного ряда (y_t). Для нелинейных трендов предварительно проводят стандартную процедуру линеаризации (таблица 8).

При выборе конкретной функции предпочтение отдается той, которая имеет меньшую сумму квадратов отклонений фактических уровней временного ряда от теоретических, найденных по уравнениям тренда.

Таблица 7 – Линеаризующие преобразования

Функция	Преобразования переменных	
	y	t
$y_t = a + b / t$	y	$1/t$
$y_t = e^{a+bt}$	$ln\ y$	t
$y_t = a\ t^b$	$lg\ y$	$lg\ t$
$y_t = a+b_1t+b_2t^2+\ldots+b_kt^k$	y	$t_1=t,\ t_2=t^2,\ldots t_к=t^к$
$y_t = ab^t$	$lg\ y$	t

Наиболее простую экономическую интерпретацию имеют параметры линейного и экспоненциального трендов. Для линейного тренда: a – начальный уровень временного ряда в момент времени $t = 0$; b – средний за единицу времени абсолютный прирост уровней ряда. Для показательного тренда:

a – начальный уровень временного ряда в момент времени $t = 0$;

b – средний за единицу времени коэффициент роста уровней ряда.

Критерием отбора наилучшей формы тренда является значение скорректированного коэффициента детерминации: чем выше его значение, тем лучше форма тренда отражает тенденцию изменения уровней ряда.

Пример 3. Имеются данные о продуктивности коров, кг/гол.

Таблица 8 – Продуктивность коров, кг/гол.

Год	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Продуктивности коров, кг/гол.	2635	2331	3105	5186	5101	4212	3962	5255	7413	7033

Требуется:

1. Построить график динамики продуктивности коров.
2. Рассчитать коэффициент автокорреляции первого порядка.
3. Обосновать выбор типа уравнения тренда и рассчитать его параметры.
4. Дать интерпретацию параметров тренда и сделать выводы по результатам решения.

Решение

1. Рассмотрим систему координат Y0t, где Y_t – продуктивности коров, ц/гол; t – порядковый номер года и нанесем в ней данные примера 3 на график.

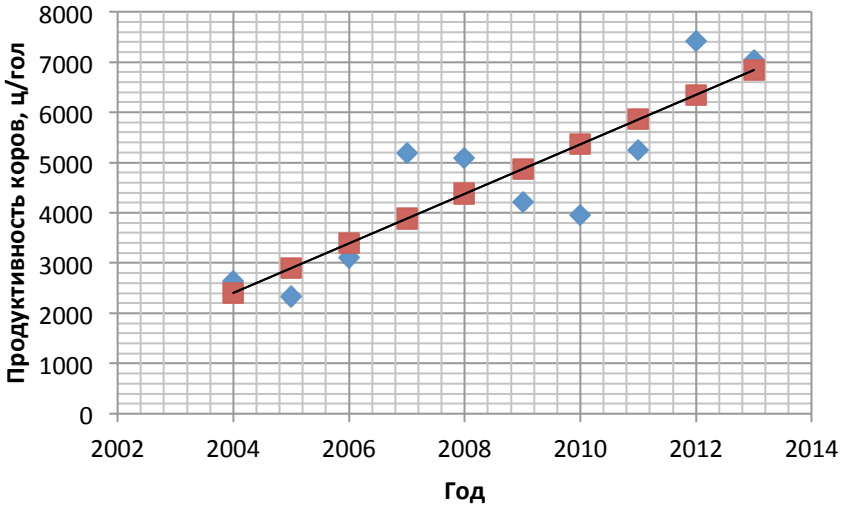

Рисунок 6 – Динамика продуктивности коров

2. Определим коэффициент автокорреляции первого порядка, характеризующего степень тесноты связи между последовательными уровнями временного ряда продуктивности коров, сдвинутыми на один год. Заполним вспомогательную таблицу 10.

Таблица 10 – Вспомогательная таблица для расчета коэффициента автокорреляции

t	y_t	y_{t-1}	$y_t - \bar{y}_1$	$y_{t-1} - \bar{y}_2$	$(y_t - \bar{y})(y_{t-1} - \bar{y})$	$(y_t - \bar{y}_1)^2$	$(y_{t-1} - \bar{y}_2)^2$
1	2635	–	–	–	–	–	–
2	2331	2635	-2513,22	-1720,56	4324145,8	6316274,8	2960326,7
3	3105	2331	-1739,22	-2024,56	3521155,2	3024886,2	4098843,2
4	5186	3105	341,78	-1250,56	-427416,4	116813,6	1563900,3
5	5101	5186	256,78	830,44	213240,4	65936,0	689630,6
6	4212	5101	-632,22	745,44	-471282,1	399702,1	555680,8
7	3962	4212	-882,22	-143,56	126651,5	778312,1	20609,5
8	5255	3962	410,78	-393,56	-161666,6	168740,2	154889,5
9	7413	5255	2568,78	899,44	2310463,5	6598630,7	808992,3
10	7033	7413	2188,78	3057,44	6692063,5	4790757,9	9347939,4
Сумма	46233	39200	0,02	-0,04	16127354,8	22260053,6	20200812,3

$$\overline{y_1} = \frac{\sum_{t=2}^{n} y_t}{n-1} = \frac{46233 - 2635}{9} = 4844{,}22;$$

$$\overline{y_2} = \frac{\sum_{t=2}^{n} y_{t-1}}{n-1} = \frac{39200}{9} = 4355{,}56;$$

$$r_1 = \frac{\sum_{t=2}^{n} (y_t - \overline{y_1}) \cdot (y_{t-2} - \overline{y_2})}{\sqrt{\sum_{t=2}^{n} (y_t - \overline{y_1})^2 \cdot \sum_{t=2}^{n} (y_{t-1} - \overline{y_2})^2}} = \frac{16127354{,}8}{\sqrt{22260053{,}6 \cdot 20200812{,}3}} = 0{,}7605.$$

3. Полученное значение коэффициента автокорреляции и графическое изображение временного ряда позволяют сделать вывод о том, что ряд продуктивности коров содержит тенденцию, близкую к линейной. Поэтому для моделирования его тенденции используем линейную функцию

$$y = a + bt.$$

Для расчета параметров линейного тренда a и b используем метод наименьших квадратов, для чего составим и решим следующую систему:

$$\begin{cases} na + b\Sigma t = \Sigma y, \\ a\Sigma t + b\Sigma t^2 = \Sigma yt. \end{cases}$$

Для составления и решения системы уравнений заполним таблицу 11.

Таблица 11 – Вспомогательная таблица для расчета параметров тренда

№ п/п	y	t	Yt	t^2	$\hat{y}_t$	$(y-\hat{y}_t)^2$
1	2635	1	2635	1	2404,8	52992,04
2	2331	2	4662	4	2897,8	321262,24
3	3105	3	9315	9	3390,8	81681,64
4	5186	4	20744	16	3883,8	1695724,84
5	5101	5	25505	25	4376,8	524465,64
6	4212	6	25272	36	4869,8	432700,84
7	3962	7	27734	49	5362,8	1962240,64
8	5255	8	42040	64	5855,8	360960,64
9	7413	9	66717	81	6348,8	1132521,64
10	7033	10	70330	100	6841,8	36557,44
Сумма	46233	55	294954	385	46233	6601107,6
Среднее значение	4623,3	5,5	29495,4	38,5		

Воспользуемся формулами, вытекающими из системы:

$$b = \frac{\overline{yt} - \bar{y}\cdot\bar{t}}{\overline{t^2} - \bar{t}^2} = \frac{29495{,}4 - 4623{,}3\cdot 5{,}5}{38{,}5 - 5{,}5^2} = 493;$$

$$a = \bar{y} - b\bar{t} = 4623{,}3 - 493\cdot 5{,}5 = 1911{,}8 \Rightarrow \mathrm{Y} = 1911{,}8 + 463t.$$

Таким образом, в среднем ежегодно за 2004–2013 гг. продуктивность коров увеличивалась на 493 ц/гол. Парный коэффициент корреляции между *У* и *t* составил 0,867, что свидетельствует о достаточно хорошем отражении тенденции роста продуктивности коров линейным уравнением.

Рассчитаем прогнозное значение средней продуктивности коров на 2016 г. путем подстановки в уравнение линейного тренда значения $t_{\text{пр}}$=13:

$$\hat{Y}_{пр} = 1911{,}8 + 493\cdot 13 \;\rightarrow \hat{Y}_{пр} = 8320{,}8 \text{ ц/гол}$$

Однако точечный прогноз нереален и он дополняется расчетом интервальной оценки $\hat{y}^*_{пр}$ с учетом 95 %-й доверительной вероятности:

$$\hat{Y}_{пр} - t_\alpha \cdot m_{\hat{y}_t} \le y^*_{пр} \le \hat{Y}_{пр} + t_\alpha \cdot m_{\hat{y}_t},$$

где: t_α – критическое значение *t*-критерия Стьюдента при уровне значимости α и числе степеней свободы $df = n$-2;

$m_{\hat{y}_t} = \sqrt{S^2_{ост}\left(\frac{1}{n} + \frac{(t_{пр}-\bar{t})^2}{\sum(t-\bar{t})^2}\right)}$ – стандартная ошибка прогноза.

$$S^2 = \frac{\sum(y-\hat{y}_t)^2}{n-2} = \frac{6601107,6}{8} = 825138,45,$$

$$m_{\hat{y}_t=} \sqrt{825138,45 \cdot \left(\frac{1}{10} + \frac{(13-5.5)^2}{82.5}\right)} = 803,2$$

при $t_{\alpha=0.05;df=8} = 2,3.$

$$8320,8 - 2,3 \cdot 803,2 \leq y_{np}^* \leq 8320,8 + 2,3 \cdot 803,2,$$
$$6473,4 \leq y_{np}^* \leq 10168,2.$$

Следовательно, с доверительной вероятностью 0,95 можно утверждать, что продуктивность коров в 2016 г. будет находиться в интервале от 6473,2 до 10168 ц/гол

Список литературы для самостоятельного изучения

1. Крянев А.В. Метрический анализ и обработка данных [Электронный ресурс]/ Крянев А.В., Лукин Г.В., Удумян Д.К.– Электрон.текстовые данные.— М.: ФИЗМАТЛИТ, 2012.— 280 с.— Режим доступа: http://www.iprbookshop.ru/33374.— ЭБС «IPRbooks», по паролю

2. Лисицин Д.В. Устойчивые методы оценивания параметров статистических моделей [Электронный ресурс]: учебное пособие/ Лисицин Д.В.— Электрон. текстовые данные.— Новосибирск: Новосибирский государственный технический университет, 2013.— 76 с.— Режим доступа: http://www.iprbookshop.ru/45452.— ЭБС «IPRbooks», по паролю

3. Общая и прикладная статистика: учебник / П.Ф. Аскеров, Р.Н. Пахунова, А.В. Пахунов. – ИНФРА-М, 2014. – 271 с. + Доп. Материалы [Электронный ресурс; режим доступа http://www.znanium.com]

4. Статистика : учеб. пособие / Иода Е.В. - М.: Вуз. учеб. : ИНФРА -М, 2012. – 302 с.

5. Статистика : учеб. пособие / Шумак О.А., Гераськин А.В. – М.: РИОР : ИНФРА-М, 2012.

6. Теория статистики [Электронный ресурс]: учебник/ Р.А. Шмойлова [и др.].— Электрон.текстовые данные.— М.: Финансы и статистика, 2014.— 656 с.— Режим доступа: http://www.iprbookshop.ru/18846.— ЭБС «IPRbooks»

7. Теоретико-вероятностные и статистические методы и модели анализа внешнеэкономической деятельности предприятий [Электронный ресурс]/ И.Н. Абанина [и др.].— Электрон. текстовые данные.— М.: Московская государственная академия делового администрирования, 2014.— 215 с.— Режим доступа: http://www.iprbookshop.ru/30548.— ЭБС «IPRbooks», по паролю

8. Федин Ф.О. Анализ данных. Часть 1. Подготовка данных к анализу [Электронный ресурс]: учебное пособие/ Федин Ф.О., Федин Ф.Ф.— Электрон. текстовые данные.— М.: Московский городской педагогический университет, 2012.— 204 с.— Режим доступа: http://www.iprbookshop.ru/26444.— ЭБС «IPRbooks», по паролю