Вероятность, статистика и прикладные исследования в аграрном университете

Серия основана в 2012 году

Министерство сельского хозяйства Российской Федерации
ФГБОУ ВПО «Кубанский государственный аграрный университет»

Практикум по эконометрике

Учебно-практическое пособие для бакалавров

Допущено Министерством сельского хозяйства Российской Федерации в качестве учебно-практического пособия для студентов высших аграрных учебных заведений, обучающихся по направлению 080100.62 «Экономика»

Под редакцией профессора П.С. Бондаренко

Серия: Вероятность, статистика и прикладные исследования в аграрном университете
Под редакцией: профессора А.И. Трубилина, профессора И.А. Кацко

Краснодар
2014

УДК 519.2+681.3
ББК 65в6я73
П69

Р е ц е н з е н т ы:

кафедра математической статистики, эконометрики и актуарных расчётов
Ростовского государственного экономического университета (РИНХ),
заведующая кафедрой - заслуженный деятель науки РФ,
доктор экономических наук,
профессор **Л.И. Ниворожкина**

С.Г. Чефранов – доктор экономических наук, профессор, заведующий
кафедрой организации и технологии защиты информации
(Майкопский государственный технологический университет)

Авторский коллектив:
П.С. Бондаренко, И.А. Кацко, В.И. Перцухов, А.Е. Сенникова,
А.Е. Жминько, Т.В. Соловьёва, Е.Д. Стеганцова, Т.Ю. Чернобыльская

П69 **П.С. Бондаренко [и др.]; Практикум** по эконометрике: учеб.-практ. пособие для бакалавров / под ред. П.С. Бондаренко. – Краснодар: Кубанский ГАУ, 2014. – 164 с., ил. (Серия: Вероятность, статистика и прикладные исследования в аграрном университете)

Эконометрика является одной из ведущих дисциплин при подготовке специалистов по направлению 080100.62 «Экономика» и в настоящее время данный учебный курс является обязательным, включен в качестве федерального компонента при подготовке бакалавров третьего поколения. В современных условиях развития экономики повышается роль специалистов, обладающих знаниями и навыками использования методов эконометрического анализа при изучении социально-экономических явлений.

Учебное пособие призвано оказать помощь студентам в овладении приемами и методами эконометрического анализа данных с использованием компьютера. В нем содержатся начальные сведения работы с часто употребляющимися средствами (*MS EXCEL, Statistica, Stata*), примеры выполнения заданий и задания для самостоятельного решения по основным разделам эконометрики.

Учебное пособие предназначено для оказания методической и практической помощи в соответствии с программой курса для подготовки бакалавров по направлению 080100.62 «Экономика», профили подготовки «Экономика предприятий и организаций», «Бухгалтерский учет, анализ и аудит», «Финансы и кредит», «Производственный менеджмент», «Мировая экономика», «Налоги и налогообложение».

УДК 519.2+681.3
ББК 65в6я73

ISBN 978-5-94672-617-7

ОГЛАВЛЕНИЕ

ПРЕДИСЛОВИЕ

В последние десятилетия XX века существенно ускорились социально-экономические процессы в обществе, усложнились связи между общественными явлениями, поэтому с 1970-х годов во всем мире считается, что теоретическое изучение экономических процессов и явлений, необходимое для адекватного реагирования (прогнозирования, управления) в банковском деле, финансах, бизнесе (аграрном, промышленном), сфере услуг на уровне организаций и государства должно основываться на эконометрических моделях, полученных по ретроспективным данным.

Эконометрика – это наука, которая позволяет придавать количественное выражение качественным закономерностям и связям в экономических явлениях и процессах. Она базируется на экономической теории, экономической и математической статистике и обычно предполагает вероятностную природу изучаемых данных. При подготовке бакалавров-экономистов уделяется большее внимание эконометрическому моделированию связей и закономерностей социально-экономических явлений и процессов.

Предметом эконометрики являются социально-экономические явления и процессы, их связи и зависимости.

Государственные образовательные стандарты третьего поколения включают эконометрику, наряду с микроэкономикой и макроэкономикой в качестве обязательной дисциплины для обучения, как бакалавров, так и магистров.

Разная глубина изучения и степень детализации данного предмета позволяет рассматривать перечисленные разделы как на уровне бакалавриата, так и на уровне магистратуры.

Данное методическое пособие посвящено формированию практических навыков эконометрического моделирования с использованием пакетов программ *Excel* и *Statistika10*, которые вполне пригодны для решения большинства эконометрических задач и, при необходимости, позволяют легко перейти к профессиональным эконометрическим пакетам (например, *Stata*).

Изучение этой дисциплины предполагает приобретение студентами опыта построения эконометрических моделей, принятия решений об идентификации модели, выбора метода оценки параметров модели, интерпретацию результатов и дальнейшего прогнозирования.

Особое внимание уделяется построению эконометрических моделей на основе пространственных данных и временных рядов.

Большое число задач составлено так, чтобы обеспечить индивидуальную работу студентов, предусмотрена возможность различных комбинаций объясняющих переменных, выбор различных зависимых переменных, предлагаются дифференцированные задания.

Следует отметить некоторое отличие классических эконометрических методов от методов, использующихся в бизнес-аналитике, основанной на разведочном анализе данных, нашедшем свое воплощение в информационных технологиях *Data Mining* и *KDD (Knowledge Discoveryin Databases).*

В настоящее время в соответствии с классификацией методов анализа данных можно выделить два подхода к изучению предметной области:

а) верификации гипотез о процессах в предметной области, которая основывается на использовании традиционных статистических методов;
б) открытии новых знаний с использованием методов *DataMining* (машинного обучения), способных автоматически обнаруживать ранее неизвестные связи, скрытые в данных.

Эконометрика постулирует первый подход. В бизнес-аналитике комбинируются и применяются оба подхода.

В бизнес-аналитике жажда прибыли позволяет жертвовать пониманием модели, в эконометрике понимание – необходимый атрибут использования модели экономистами на практике (при прогнозировании и управлении).

Авторы полагают, что реализованный ими в учебном пособии подход позволит студентам получить необходимые навыки построения и анализа эконометрических моделей социально-экономических явлений и процессов.

Учебно-практическое пособие предназначено для выполнения лабораторных заданий студентами очной и заочной форм обучения.

Цель пособия – оказать помощь студентам в овладении приемами и методами статистического анализа данных с использованием компьютера. Содержатся начальные сведения работы с часто употребляющимися средствами (*MS EXCEL, Statistika, Stata*).

Предлагаемые лабораторные работы охватывают стандартные разделы многомерного статистического анализа: метод визуализации, поиска зависимостей (дисперсионный анализ, корреляционно-регрессионный анализ и анализ временных рядов), классификации и методы снижения размерности изучаемого признакового пространства (факторный анализ).

Задания могут быть использованы при самостоятельном изучении курса.

1. Парная регрессия и корреляция (однофакторный корреляционно-регрессионный анализ)

Ученые много столетий наблюдают, что произойдет с интересующим их явлением (*y*–результативным признаком или функцией отклика), если изменить независимую переменную (фактор *x*). Обычно неявно предполагается, что изменение фактора x является причиной изменения y. Считается, что факторы должны иметь достаточный уровень вариабельности, который может характеризоваться коэффициентом вариации (обычно не менее 0,1).

Предполагается, что между результативным и факторным признаками существует зависимость типа $y=f(x)$, но реальные наблюдения в силу неучтенных факторов отличаются от теоретического значения y на величину ε – случайную ошибку:

$$y=f(x)+\varepsilon. \qquad (1.1)$$

Таким образом, в основе корреляционного анализа лежит некоторая вероятностно-статистическая модель.

Вид функции $f(x)$ при изучении парной регрессии подбирается на основе графического отображения пар наблюдений на плоскости.

Таблица 1.1 – Различные виды функций

№ п/п	Функции	Нормальные уравнения
1	$y = a + bx + \varepsilon$ линейная	$an + b\sum x = \sum y$ $a\sum x + b\sum x^2 = \sum (xy)$
2	$\lg y = a + bx$ или $y = 10^{a+bx}$ показательная	$an + b\sum x = \sum \lg y$ $a\sum x + b\sum x^2 = \sum (x \lg y)$
3	$y = a + b \lg x$ полулогарифмическая	$an + b\sum \lg x = \sum y$ $a\sum \lg x + b\sum (\lg x)^2 = \sum (y \lg x)$
4	$\lg y = a + b \lg x$ или $y = 10^{a+b\lg x}$	$an + b\sum \lg x = \sum \lg y$ $a\sum \lg x \cdot b\sum (\lg x)^2 = \sum (\lg x \lg y)$
5	$y = ab^x$ или $\lg y = \lg a + x \lg b$ показательная	$n\lg a + \lg b\sum x = \sum \lg y$ $\lg a\sum x + \lg b\sum x^2 = \sum (\lg x \lg y)$
6	$y = a + bx + cx^2$ парабола второго порядка	$an + b\sum x + c\sum x^2 = \sum y$ $a\sum x + b\sum x^2 + c\sum x^3 = \sum xy$ $a\sum x^2 + b\sum x^3 + c\sum x^4 = \sum (x^2 y)$
7	$y = ax^b$ или $lny=lna+blnx$ степенная	$n\lg a + b\sum \lg x = \sum \lg y$ $\lg a\sum \lg x + b\sum (\lg x)^2 = \sum (\lg x \lg y)$
8	$y = \frac{a}{1 + b \cdot e^{-cx}}$ Логистическая	(решается численно приближенными методами)

В эконометрике чаще всего рассматривают различные типы функций *f*, параметры которых находят с использованием решения соответствующих систем нормальных уравнений.

Ряд нелинейных зависимостей можно идентифицировать путем соответствующих нелинейных преобразований.

Таблица 1.2 – Преобразование функций в линейный вид

№ п/п	Функция	Линеаризующие преобразования			
		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 его параметров и оценки их значимости; оценка тесноты и направления связи между переменными; определение прогнозных значений зависимой переменной и оценка полученного прогноза.

Корреляционно-регрессионный анализ проводится в следующей последовательности:

1. Исходя из целей и задач исследования зависимости устанавливается результативный и факторные признаки. По совокупности объектов определяются значения результативного и факторных признаков.

2. Обосновывается (обычно графическим способом) модель уравнения регрессии (рисунок 1.1). 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 *y* с соблюдением масштаба. На основе корреляционного поля делаются выводы о направлении и возможной функциональной форме связи между факторным и результативным признаками (прямая или обратная, линейная или нелинейная).

3. Методом наименьших квадратов определяются параметры уравнения регрессии.

Для линейных и нелинейных уравнений, приводимых к линейному виду, для этого решается следующая система уравнений:

$$\begin{cases} \Sigma y = na + b\Sigma x, \\ \Sigma yx = a\Sigma x + b\Sigma x^2. \end{cases} \quad (1.2)$$

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

$$b = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \left(\overline{x}\right)^2}, \qquad a = \overline{y} - b \cdot \overline{x}. \quad (1.3)$$

Коэффициент регрессии линейного уравнения b есть абсолютный показатель силы связи, характеризующий среднее абсолютное изменение результата при изменении факторного признака на единицу.

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

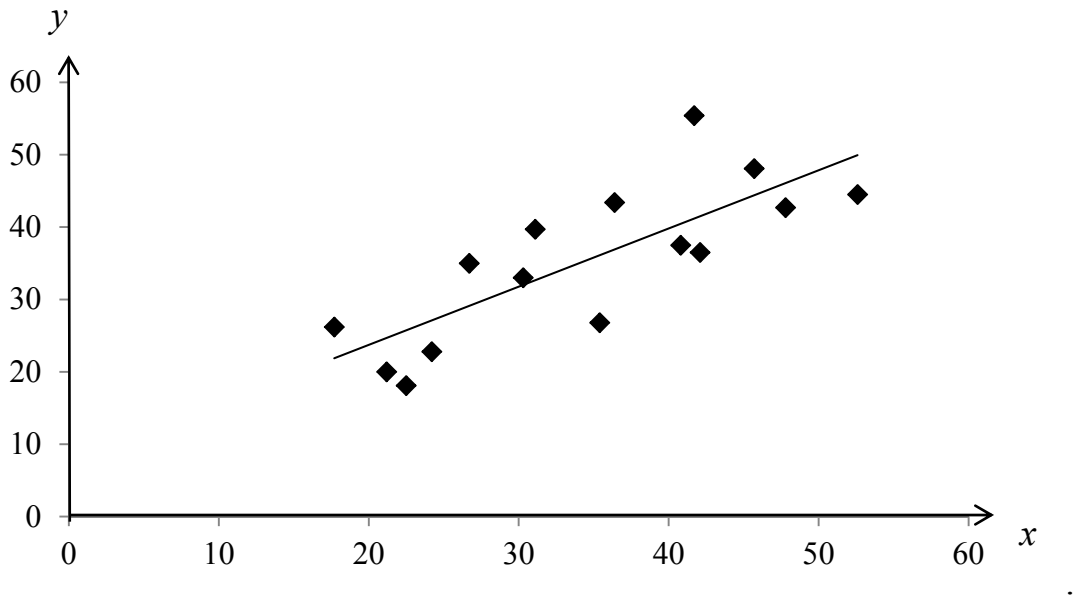


Рисунок 1.1– Корреляционное поле

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

$$\bar{A} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\,\%. \quad (1.4)$$

Допустимый предел значений средней ошибки аппроксимации 10-12%. В этом случае качество уравнения регрессии считается хорошим.

5. При линейной зависимости теснота связи между переменными x и y. определяется с помощью коэффициента корреляции:

$$r = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sigma_x \cdot \sigma_y}, \quad (1.5)$$

где σ_x и σ_y – средние квадратические отклонения по x и по y.

$$\sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2}; \quad \sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2}; \quad (1.6)$$

Корреляционная зависимость между переменными величинами – это та функциональная зависимость, которая существует между значениями одной из них и групповыми средними другой. (Корреляционные зависимости *y* на *x* и *x* на *y* обычно не совпадают).

Корреляционная связь чаще всего характеризуется выборочным коэффициентом корреляции r, который характеризует степень линейной функциональной зависимости между СВ *x* и *y*.

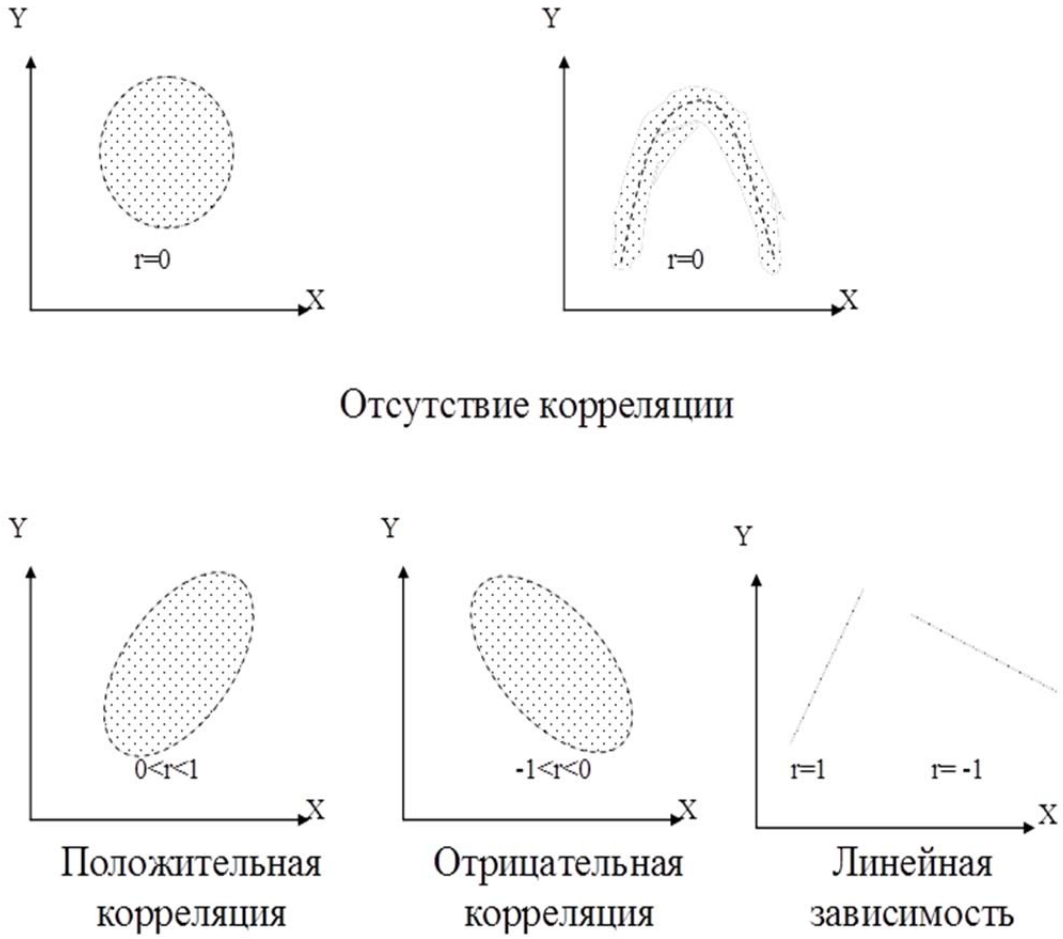


Рисунок 1.2 – Виды корреляционных полей

Чем ближе значение коэффициента корреляции к нулю, тем степень линейной связи между признаками слабее. Чем ближе к единице, тем связь сильнее. Если r=1, то связь функциональная. Если r=0, то связи между признаками нет.

Коэффициент корреляции также показывает направление связи между признаками. Если r> 0, то связь прямая. Если r<0, то связь обратная.

При нелинейной зависимости теснота связи между переменными *x* и *y* определяется с помощью индекса корреляции:

$$R_{xy} = \sqrt{1 - \frac{\sigma^2_{ост}}{\sigma^2_y}}\,. \qquad (1.7)$$

Квадрат коэффициента (индекса) корреляции называется коэффициентом (индексом) детерминации. Коэффициент детерминации *D* показывает долю влияния фактора *x* на результативную переменную *y*, а (1-*D*) – доля влияния других, неучтенных в модели факторов (толерантность – в англоязычных источниках).

$$D = r^2 \cdot 100\%. \qquad (1.8)$$

6. Средний коэффициент эластичности определяется по формуле:

$$\overline{Э} = f'(x)\frac{\overline{x}}{\overline{y}}. \qquad (1.9)$$

Таблица 1.3 – Коэффициенты эластичности для ряда математических функций

Функция, y	Первая производная, y'_x	Коэффициент эластичности, $Э = y'_x \cdot \frac{x}{y}$
Линейная $y = a + b \cdot x + \varepsilon$	b	$Э = \frac{b \cdot x}{a + b \cdot x}$
Парабола второго порядка $y = a + b \cdot x + c \cdot x^2 + \varepsilon$	$b + 2 \cdot c \cdot x$	$Э = \frac{(b + 2 \cdot c \cdot x) \cdot x}{a + b \cdot x + c \cdot x^2}$
Гипербола $y = a + \frac{b}{x} + \varepsilon$	$-\frac{b}{x^2}$	$Э = \frac{-b}{a \cdot x + b}$
Показательная $y = a \cdot b^x \cdot \varepsilon$	$\ln b \cdot a \cdot b^x$	$Э = x \cdot \ln b$
Степенная $y = a \cdot x^b \cdot \varepsilon$	$a \cdot b \cdot x^{b-1}$	$Э = b$
Полулогарифмическая $y = a + b \cdot \ln x + \varepsilon$	$\frac{b}{x}$	$Э = \frac{b}{a + b \cdot \ln x}$
Логистическая $y = \frac{a}{1 + b \cdot e^{-cx+\varepsilon}}$	$\frac{a \cdot b \cdot c \cdot e^{-cx}}{(1 + b \cdot e^{-cx})^2}$	$Э = \frac{c \cdot x}{\frac{1}{b} \cdot e^{cx} + 1}$
Обратная $y = \frac{1}{(a + b \cdot x)^2}$	$\frac{-b}{(a + b \cdot x)^2}$	$Э = \frac{-b \cdot x}{a + b \cdot x}$

При линейной форме связи он находится по формуле:

$$\overline{Э} = b\frac{\bar{x}}{\bar{y}}, \qquad (1.10)$$

где $\overline{x}$ и $\overline{y}$ - средние значения признаков;

b – коэффициент регрессии.

Средний коэффициент эластичности для линейной модели показывает, что при увеличении фактора x на 1 %, результативная переменная y в среднем возрастает на величину коэффициента эластичности.

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

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

$$F_{\text{н}} = \frac{r^2}{1-r^2}\left(\frac{n-m-1}{m}\right); \qquad (1.11)$$

где m – число параметров при переменной x(для одной переменной x, m=1);

n – число пар наблюдений.

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

$$F_{\text{н}} = \frac{r^2}{1-r^2}(n-2); \qquad (1.12)$$

Для определения табличного значения используем таблицу Фишера-Снедекора. При уровне значимости α=0,05 и числе степеней свободы k_1=m; k_2=n-m-1;

$$F_{kp} = F_{\alpha=0,05}(\alpha;\ k_1;\ k_2);$$

В случае парной регрессии число степеней свободы большей дисперсии k_1=1, а число степеней свободы меньшей дисперсии k_2=n-2.

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

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

Находится наблюдаемое значение для параметров a и b:

$$t_a = \frac{a}{m_a}; m_a = \frac{\sigma_{\text{ост}}\sqrt{\sum x^2}}{n\sigma_x}; \qquad (1.13)$$

$$t_b = \frac{b}{m_b}; m_b = \frac{б_{\text{ост}}}{б_x\sqrt{n}}; \qquad (1.14)$$

Определяем критическое значение $t_{кр.}$ по таблице критических значений t – Стьюдента. Если $t_н > t_{кр}$, то основную гипотезу отвергаем и принимаем аль-

тернативную гипотезу и коэффициенты уравнения регрессии *a* и *b* статистически значимы при заданном уровне значимостиα. 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 применяем критерий *t* – Стьюдента и определяем наблюдаемое значение *t* – критерия:

$$t_н = |r|\sqrt{\frac{n-2}{1-r^2}}; \qquad (1.15)$$

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

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

Рассчитаем доверительный интервал для параметров *a* и *b*.

Для этого определим предельную ошибку для каждого параметра:

$$\Delta_a = t \cdot \hat{m}_a; \qquad \Delta_b = t \cdot \hat{m}_b; \qquad (1.16)$$

Доверительные интервалы для параметров a и b определим по следующим формулам:

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

$$\gamma_{b_{min}} = b - \Delta_b; \qquad \gamma_{b_{max}} = b + \Delta_b; \qquad (1.17)$$

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

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

$$\hat{y}_p = a + bx_p; \qquad (1.18)$$

10. Для оценки значимости прогноза определяют стандартную и предельную ошибки прогноза:

$$m_{\hat{y}_p} = \sigma_{\text{ост}} \sqrt{1 + \frac{1}{n} + \frac{\left(\bar{x} - \overline{x_p}\right)^2}{\sum (x_i - \bar{x})^2}}, \qquad (1.19)$$

где
$$\sigma_{\text{ост}} = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n - m - 1}}; \qquad (1.20)$$

$$\Delta_{\hat{y}_p} = t \cdot \hat{m}_{y_p};$$

Доверительный интервал прогноза:

$$\gamma_{yp_{min}} = \hat{y}_p - \Delta_{\hat{y}_p};\ \gamma_{yp_{max}} = \hat{y}_p + \Delta_{\hat{y}_p}; \qquad (1.21)$$

По доверительному интервалу прогноза можно оценить статистическую значимость и надежность прогноза при заданном уровне значимости α.

Точность прогноза оценим с помощью диапазона прогноза:

$$D_{\hat{y}_p} = \frac{\gamma_{\text{ypmax}}}{\gamma_{\text{pmin}}}; \qquad (1.22)$$

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

11. Дисперсионный анализ модели регрессии.

После построения уравнения регрессии возникает вопрос о качестве решения. Пусть при исследовании n пар наблюдений (x_i, y_i) получено уравнение регрессии *Y* на *X*.

Рассмотрим тождество:

$$y_i - \hat{y}_i = y_i - \bar{y}_i - (\hat{y}_i - \bar{y}_i). \qquad (1.23)$$

Если переписать предыдущее уравнение в виде:

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i), \qquad (1.24)$$

возвести обе части в квадрат и просуммировать по *i*, то получим:

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2, \qquad (1.25)$$

(можно показать, что $2\sum (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$).

Уравнение является основополагающим в дисперсионном анализе.

Выражение можно переписать в виде:

$$SS_1 = SS_2 + SS_3 \qquad (1.26)$$

Для сумм квадратов обычно вводятся названия:

$SS_1 = \sum (y_i - \bar{y})^2$ – сумма квадратов отклонений *i*-го наблюдения от общего среднего, или сумма квадратов отклонений относительно среднего наблюдений (*скорректированная сумма квадратов Y*-ов, в *Excel*[1]*:SS* - итог);

[1] Здесь и далее "в Excel" понимается инструмент **Регрессия** в *Пакете анализа*

$SS_2 = \sum (y_i - \hat{y}_i)^2$ – сумма квадратов отклонений *i*-го наблюдения от предсказанного, или сумма квадратов, обусловленная регрессией (в *Excel:SS* - остаток);

$SS_3 = \sum (\hat{y}_i - \bar{y})^2$ – сумма квадратов отклонений предсказанного значения от общего среднего, или сумма квадратов относительно регрессии (в *Excel:SS* - регрессия);

$(\sum y_i)^2$ - не скорректированная сумма квадратов *Y*-ов;

$\frac{(\sum y_i)^2}{n} = SS_3(b_0)$ - коррекция на среднее суммы квадратов *Y*-ов.

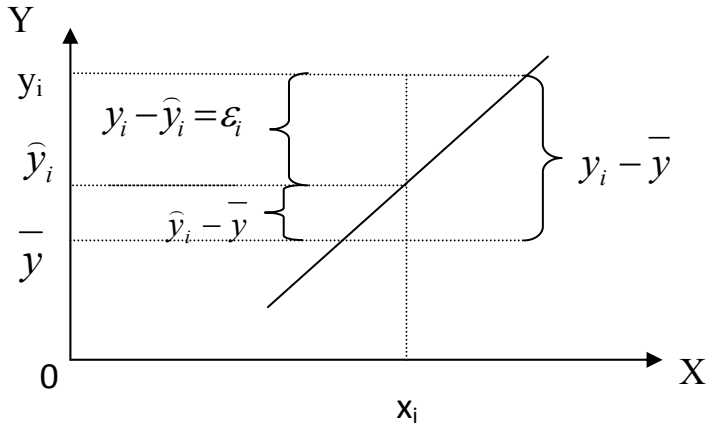


Рисунок 1.3 – Иллюстрация формулы (1.24)

Адекватность линии регрессии зависит от того, какая часть суммы квадратов относительно среднего обусловлена суммой квадратов относительно регрессии, а какая суммой квадратов обусловленной регрессией.

Суммы квадратов связаны с некоторым числом - числом их степеней свободы ν=*df*. Это число показывает, сколько независимых элементов информации (из n чисел $y_1, y_2, \ldots y_n$) необходимо для образования данной суммы квадратов. Например, для $\sum (y_i - \bar{y})^2$, *df=(n-1)*.

Действительно, из *n* разностей $(y_1 - \bar{y}), (y_2 - \bar{y}), \ldots, (y_n - \bar{y})$ только (*n-1*) независимы (или иначе, для образования рассматриваемой суммы из $y_1, y_2, \ldots, y_n$ достаточно (n - 1) значение, так как оставшееся можно определить, зная $\bar{Y}$). Аналогично для $\sum (\hat{y}_i - \bar{Y})^2$, *df=1*; для $\sum (y_i - \hat{y}_i)^2$, *df = (n-2)* (число степеней свободы определяется как *n* минус число оцениваемых параметров).

Для построения таблицы дисперсионного анализа необходимо получить средние квадраты (*MS*), для этого каждая сумма SS делится на соответствующие число степеней свободы *df* ($MS_R = \frac{SS_{рег.}}{1}$, $S^2 = \frac{SS}{n-2}$).

Если в уравнении регрессии ($y=b_0+b_1x$) $b_1=0$, то величина $F = \frac{MS_R}{S^2}$ распределена по распределению Фишера с ($1, n-2$) степенями свободы. Этот факт используется для проверки гипотезы H_0: $b_1=0$ ($r^2=0$) с уровнем значимости α (риском ошибиться не более чем в $\alpha 100\%$ случаев), против альтернативы H_1: $b_1 \neq 0$, *с*.

Обобщим все в таблице дисперсионного анализа в таблице 1.4.

Если $F_{расч.} > F_{кр.}$ при заданном уровне значимости α и соответствующих числах степеней свободы, то гипотеза H_0: $b_1=0$ ($r^2 \neq 0$)отбрасывается с риском ошибиться не более, чем в $\alpha 100\%$ случаев (и уравнение регрессии считается статистически значимым).

Таблица 1.4 – Основное разложение дисперсионного анализа

Источник вариации	Число степеней свободы, *df*	Суммы квадратов, *SS*	Средние квадраты, *MS*	$F_{расч.}$	$F_{кр.}$
Обусловленный регрессией	*1*	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	MS_R	$F = \frac{MS_R}{S^2}$	$F_\alpha(1, n-2)$
Относительно регрессии (остаток)	*n-2*	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$S^2 = \frac{SS}{n-2}$		
Общий, скорректированный на среднее *Y*	*n-1*	$\sum_{i=1}^{n}(y_i - \bar{y})^2$			

Доля суммы квадратов, объясняемая регрессией называется коэффициентом детерминации (квадратом коэффициента корреляции r):

$$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}, \qquad -1 \le r \le 1. \tag{1.27}$$

Значимость r^2 для уравнения определяется по *F*-критерию. $F_{кр.} = F_\alpha(1, n-2)$, $F_н = \frac{r^2}{1-r^2}(n-2)$. Если $F_{расч.} > F_{кр.}$, то гипотезу H_0:$r=0$ отвергают и связь между *x* и *y* считают статистически значимой.

Выбор структуры уравнения наилучшей регрессии (наиболее точно описывающей исследуемый процесс) можно осуществить, используя r^2 - квадрат коэффициента корреляции или дисперсионный анализ. Структура уравнения регрессии усложняется (например, в полиномиальном случае повышается степень многочлена) до тех пор, пока увеличение соответствующего критерия не станет пренебрежительно малым. Однако, вывод о корректности модели по условию $r^2 \approx 1$ не всегда верен, т.к. результата $r^2 \approx 1$ можно добиться, увеличи-

вая число оцениваемых параметров β_j и в случае «насыщенности» r^2 будет равен 1, но модель при этом не обязательно корректна.

Пример 1.1. Имеются следующие выборочные данные по 15 хозяйствам центральной зоны Краснодарского края за 2011 г.

Таблица 1.5 - Фондообеспеченность и производство продукции

№	Фондообеспеченность на 1 га сельхозугодий, тыс. руб., (x)	Стоимость валовой продукции на 1 га сельхозугодий, тыс. руб., (y)
1	38,4	62,3
2	24,2	30,1
3	29,2	47,3
4	23,0	29,9
5	18,2	37,2
6	33,2	46,1
7	14,1	22,3
8	26,2	43,0
9	20,1	34,1
10	35,0	49,2
11	31,7	41,4
12	24,4	37,4
13	18,9	28,2
14	27,1	37,0
15	17,0	26,1

Требуется:

1. Построить график зависимости между переменными, по которому необходимо подобрать модель уравнения регрессии. Используя следующие функции:

а) линейную;

б) степенную;

в) экспоненциальную;

г) показательную.

2. Рассчитать параметры уравнения регрессии методом наименьших квадратов.

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

4. Найти коэффициент эластичности.

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

6. Оценить, для линейной функции, значимость коэффициентов корреляции и регрессии по критерию *t* – Стьюдента при уровне значимости α= 0,05.

7. Охарактеризовать статистическую надежность результатов регрессионного анализа с использованием критерия *F* – Фишера при уровне значимости α= 0,05 .

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

Решение:

а) Регрессия в виде линейной функции имеет вид:

$$\hat{y} = a + bx$$

1. Построим график зависимости переменных x и y в прямоугольной системе координат.

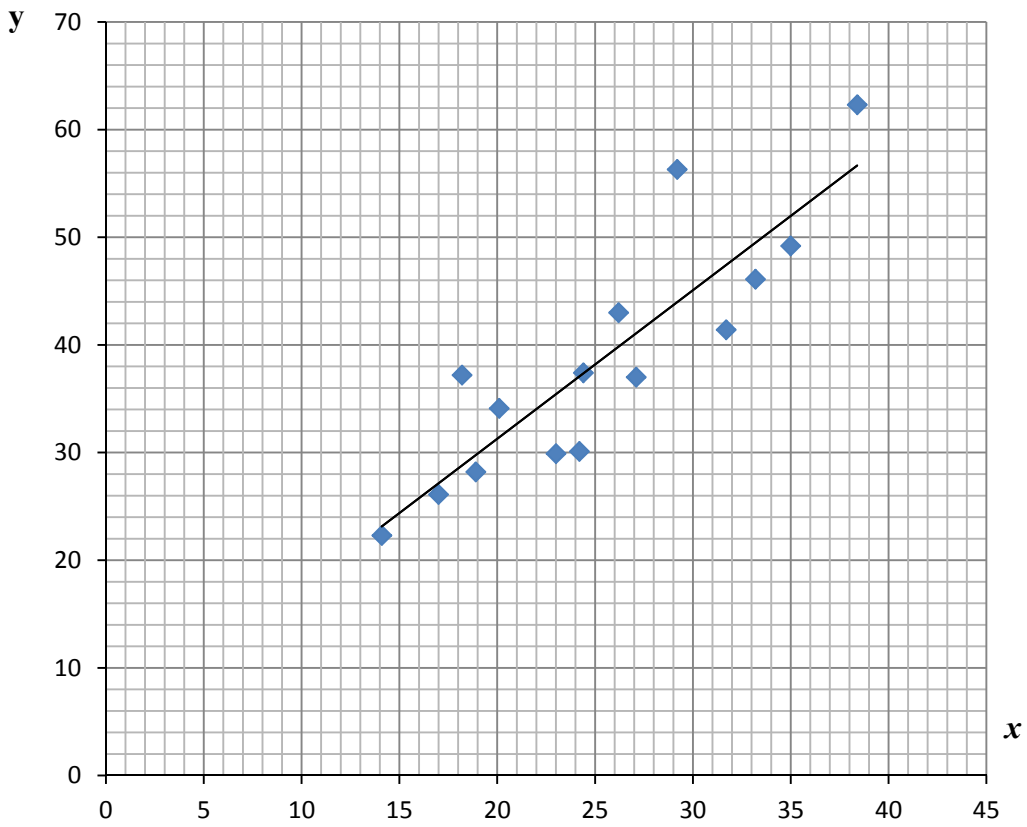


Рисунок 1.4 – Зависимость стоимости валовой продукции (тыс. руб.) от фондообеспеченности на 1 га сельхозугодий (тыс. руб.)

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

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

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

Для проведения всех расчетов построим вспомогательную таблицу 10.6.

В таблице все средние находятся по формуле средней арифметической простой: $\overline{x} = \Sigma x : n$.

Таблица 1.6 - Вспомогательная таблица регрессионного анализа для уравнения линейной регрессии

№ п/п	x	y	x^2	y^2	xy	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$	$A=\left\|\frac{y-\hat{y}}{y}\right\|\cdot 100\%$
1	38,4	62,3	1474,56	3881,29	2392,32	55,436	6,864	47,109	11,017
2	24,2	30,1	585,64	906,01	728,42	36,536	-6,436	41,425	21,383
3	29,2	47,3	852,64	2237,29	1381,16	43,191	4,109	16,882	8,687
4	23,0	29,9	529,00	894,01	687,70	34,939	-5,039	25,392	16,853
5	18,2	37,2	331,24	1383,84	677,04	28,550	8,650	74,819	23,252
6	33,2	46,1	1102,24	2125,21	1530,52	48,515	-2,415	5,833	5,239
7	14,1	22,3	198,81	497,29	314,43	23,093	-0,793	0,629	3,557
8	26,2	43,0	686,44	1849,00	1126,60	39,198	3,802	14,454	8,841
9	20,1	34,1	404,01	1162,81	685,41	31,079	3,021	9,126	8,859
10	35,0	49,2	1225,00	2420,64	1722,00	50,911	-1,711	2,928	3,478
11	31,7	41,4	1004,89	1713,96	1312,38	46,519	-5,119	26,201	12,364
12	24,4	37,4	595,36	1398,76	912,56	36,802	0,598	0,357	1,598
13	18,9	28,2	357,21	795,24	532,98	29,482	-1,282	1,643	4,546
14	27,1	37,0	734,41	1369,00	1002,70	40,396	-3,396	11,533	9,179
15	17,0	26,1	289,00	681,21	443,70	26,953	-0,853	0,728	3,268
Итого	380,7	571,6	10370,45	23315,56	15449,92	-	-	279,058	142,119
Среднее значение	25,38	38,107	691,36	1554,37	1029,99	-	-	18,604	9,475

Замечание.

Расчет вспомогательной таблицы можно осуществить в табличном процессоре *Excel*.

Для этого, если исходные данные (x и y) приведены в виде, представленном на рисунке 1.4:

а) Введем для расчета остальных значений таблицы 1.2 следующие формулы в соответствующие ячейки (*ввод – Enter*) :

*D2: =B2^2; E2: =C2^2; F2: ==B2*C2.*

б) Выделим диапазон ячеек и протащим с помощью маркера заполнения 2 до строки 16.

в) Для вычисления сумм введем формулу в ячейку:

B17:=СУММ(B2:B16),

и протащим с помощью маркера заполнения для диапазона *B17:F17*.

г) Для расчета средних введем формулу в ячейке *B18: =B17/15* и скопируем с помощью маркера заполнения для диапазона *B18:F18*.

д) После расчета параметров уравнения парной регрессии: *G2: = =1,331*B2+4,326.*

е) *H2: =C2-G2*.

ж) *I2:=H2^2*.

з) *J2: =ABS(H2/C2)*100*.

и) Выделяется диапазон *G2:J2* и с помощью маркера заполнения копируется до 16 строки.

к) Для столбцов *I* и *J* находятся суммы и средние (см.выше).

Результаты вычисления округлены.

Подставим полученные суммы в систему уравнений, учитывая, что n=15.

$$\begin{cases} 571,6 = 15a + 380,7b, \\ 15449,92 = 380,7a + 10370,45b. \end{cases}$$

Решим систему, например, по формулам Крамера:

$$\Delta = \begin{vmatrix} 15 & 380,7 \\ 380,7 & 10370,45 \end{vmatrix} = 15 \cdot 10370,45 - 380,7^2 = 10624,26;$$

$$\Delta_a = \begin{vmatrix} 571,1 & 380,7 \\ 15449,92 & 10370,45 \end{vmatrix} = 571,1 \cdot 10370,45 - 380,7 \cdot 380,7 = 45964,676;$$

$$\Delta_b = \begin{vmatrix} 15 & 571,1 \\ 380,7 & 15449,92 \end{vmatrix} = 15 \cdot 15449,92 - 380,7 \cdot 380,7 = 14140,68;$$

[2]**Маркер заполнения** - небольшой черный квадрат в углу выделенного диапазона. Попав на маркер заполнения, указатель принимает вид черного креста. Чтобы скопировать содержимое выделенного диапазона в соседние ячейки или заполнить их подобными данными (например, днями недели), нажмите левую кнопку мыши и перемещайте мышь в нужном направлении. Чтобы вывести на экран контекстное меню с параметрами заполнения, перетаскивайте маркер заполнения, нажав и удерживая правую кнопку мыши. (Данные справочной системы *Excel.)*

$$a = \frac{\Delta_a}{\Delta} = \frac{45964{,}676}{10624{,}26} = 4{,}326;$$

$$b = \frac{\Delta_b}{\Delta} = \frac{14140{,}68}{10624{,}26} = 1{,}331,$$

получим: $a = 4{,}326$; $b = 1{,}331$.

	A	B	C	D	E	F	G	H	I	J
1	№ п/п	x	y	x^2	y^2	xy	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$	$A=\left\|\frac{y-\hat{y}}{y}\right\|\cdot 100\%$
2	1	38,4	62,3	1474,56	3881,29	2392,32	55,436	6,864	47,109	11,017
3	2	24,2	30,1	585,64	906,01	728,42	36,536	-6,436	41,425	21,383
4	3	29,2	47,3	852,64	2237,29	1381,16	43,191	4,109	16,882	8,687
5	4	23,0	29,9	529,00	894,01	687,70	34,939	-5,039	25,392	16,853
6	5	18,2	37,2	331,24	1383,84	677,04	28,550	8,650	74,819	23,252
7	6	33,2	46,1	1102,24	2125,21	1530,52	48,515	-2,415	5,833	5,239
8	7	14,1	22,3	198,81	497,29	314,43	23,093	-0,793	0,629	3,557
9	8	26,2	43,0	686,44	1849,00	1126,60	39,198	3,802	14,454	8,841
10	9	20,1	34,1	404,01	1162,81	685,41	31,079	3,021	9,126	8,859
11	10	35,0	49,2	1225,00	2420,64	1722,00	50,911	-1,711	2,928	3,478
12	11	31,7	41,4	1004,89	1713,96	1312,38	46,519	-5,119	26,201	12,364
13	12	24,4	37,4	595,36	1398,76	912,56	36,802	0,598	0,357	1,598
14	13	18,9	28,2	357,21	795,24	532,98	29,482	-1,282	1,643	4,546
15	14	27,1	37,0	734,41	1369,00	1002,70	40,396	-3,396	11,533	9,179
16	15	17,0	26,1	289,00	681,21	443,70	26,953	-0,853	0,728	3,268
17	Итого	380,7	571,6	10370,45	23315,56	15449,92			279,058	142,119
18	Среднее значение	25,380	38,107	691,363	1554,371	1029,995			18,604	9,475

Рисунок 1.5 – Скриншот вспомогательной таблицы регрессионного анализа в *MSExcel*

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

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

$$a = \overline{y} - b\cdot\overline{x}.$$

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

Таким образом, уравнение регрессии имеет вид:

$$\hat{y} = 4{,}326 + 1{,}331x.$$

Коэффициент регрессии показывает, что при увеличении фондообеспеченности на 1 тыс. руб. стоимость валовой продукции в среднем увеличивается на 1,331 тыс. руб.

Если в уравнение регрессии подставить фактические значения переменной x, то определяются возможные (теоретические) значения переменной $\hat{y}$, которые наносятся на график в виде уравнения прямой.

3. Качество уравнения регрессии оценивается с помощью средней ошибки аппроксимации

$$\overline{A} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| .100\ \%;$$

$$\overline{A} = \frac{147{,}707}{15} = 9{,}847\%$$

Значит, фактические значения стоимости валовой продукции на 1 га от расчетных по уравнению регрессии в среднем различаются на 9,847 %.

Качество уравнения регрессии считается хорошим, если ошибка аппроксимации не превышает 8-10 %. Полученное уравнение регрессии можно оценить как вполне хорошее.

4. При линейной форме связи, средний коэффициент эластичности находится по формуле:

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

где $\overline{x}$ и $\overline{y}$ - средние значения признаков.

$$\overline{Э} = 1{,}331 \cdot \frac{25{,}380}{38{,}107} = 0{,}886$$

Коэффициент эластичности показывает, что при увеличении фондообеспеченности на 1 га на 1 % стоимость валовой продукции на 1 га в среднем возрастает на 0,886 %.

5. При линейной зависимости теснота связи между переменными x и y определяется с помощью коэффициента корреляции:

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

где σ_x и σ_y – средние квадратические отклонения по x и y.

$$\sigma_x = \sqrt{\overline{x^2} - \overline{x}^2} = \sqrt{691{,}363 - 38{,}107^2} \approx 6{,}872;$$

$$\sigma_y = \sqrt{\overline{y^2} - \bar{y}^2} = \sqrt{1554{,}371 - 38{,}107^2} \approx 10{,}112;$$

$$r = \frac{1029{,}995 - 25{,}380 \cdot 38{,}107}{6{,}872 \cdot 10{,}112} \approx 0{,}904.$$

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

Коэффициент детерминации $r^2 = 0{,}904^2 = 0{,}817$ показывает, что 81,7 % различий в стоимости валовой продукции объясняется вариацией фондообеспеченности, а 18,3 % другими, неучтенными факторами (различными статьями затрат и т.д.).

6. Так как исходные данные 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: $H_0: r_г = 0$, при $H_1: r_г \neq 0$.

Для проверки нулевой гипотезы применим критерий t – Стьюдента.

Найдем наблюдаемое значение t – критерия

$$t_н = |r| \sqrt{\frac{n-2}{1-r^2}} = 0{,}904 \sqrt{\frac{15-2}{1-0{,}904^2}} \approx 7{,}62.$$

Критическое значение t находится по таблицам распределения t – Стьюдента при уровне значимости $\alpha = 0{,}05$ и числе степеней свободы $к = n\text{-}2 = 15\text{-}2 = 13$ для двухсторонней критической области, $t_{кр} = 2{,}16$.

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, фондообеспеченность на 1 га сельхозугодий оказывает статистически существенное влияние на стоимость валовой продукции.

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

Находится наблюдаемое значение критерия:

$$t_н = \frac{b}{m_b};\ m_b = \sqrt{\frac{\sum(y - \hat{y})^2}{(n-2) \times \sum(x - \bar{x})^2}} = \sqrt{\frac{\sum(y - \hat{y})^2}{(n-2) \times \sigma_x^2 \times n}}$$

$$= \sqrt{\frac{286{,}374}{(15-2) \times 6{,}872^2 \times 15}} \approx 0{,}176;$$

$$t_н = \frac{1,331}{0,176} \approx 7,56.$$

Критическое значение t также равно 2,16. Так как $t_н > t_{кр}$, то коэффициент регрессии статистически значим. Подтверждается вывод о значимости влияния фондообеспеченности на стоимость валовой продукции.

7. Статистическая надежность уравнения регрессии проверяется с использованием критерия *F*–Фишера – рассматривается нулевая гипотеза $H_0: r^2 = 0$, при альтернативной $H_0: r^2 \neq 0$ (или нулевая гипотеза $H_0: b=0$, при $H_1: b \neq 0$). Наблюдаемое (фактическое) значение *F*–критерия находится по формуле:

$$F_н = \frac{\sum(\hat{y} - \bar{y})^2 / m}{\sum(y - \hat{y})^2 / (n - m - 1)};$$

где m – число параметров при переменных x;

n – число наблюдений.

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

$$F_н = \frac{r^2}{1 - r^2} \cdot (n - 2) = \frac{0,817}{1 - 0,817} \cdot 13 = 58,03$$

При уровне значимости α=0,05 и числе степеней свободы $k_1 = m = 1$, $k_2 = n - m - 1 = 15 - 1 - 1 = 15 - 2 = 13$ 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 (*xp*).

По условию: $x_p = \bar{x} \cdot 1,1 = 25,380 \cdot 1,1 = 27,918.$

Тогда прогнозное значение стоимости валовой продукции составит

$$\hat{y}_p = a + bx_p = 4,326 + 1,331 \times 27,918 = 41,485.$$

Значит, при фондообеспеченности в 27,918 тыс. руб. возможная стоимость валовой продукции составит 41,485 тыс. руб. (на 1 га сельхозугодий).

б*) Регрессия в виде степенной функции имеет вид:*

$$y = a \cdot x^b \cdot \varepsilon.$$

Для оценки параметров модель линеаризуется путем логарифмирования:

$$lny=lna+blnx.$$

Таким образом, если ввести соответствующие обозначения (*Y=lny, X=lnx, A=lna)*, то получим модель в линейном виде *Y=A+bX.*

Система нормальных уравнений в этом случае примет вид:

$$\begin{cases} 54{,}0861 = 15A + 47{,}9389b, \\ 173{,}8480 = 47{,}9389A + 154{,}3802b \end{cases}$$

Отсюда:

$$\Delta = \begin{vmatrix} 15 & 47{,}9389 \\ 47{,}9389 & 154{,}3802 \end{vmatrix} = 15 \cdot 154{,}3802 - 47{,}9389^2 = 17{,}5649;$$

$$\Delta_A = \begin{vmatrix} 54{,}0861 & 47{,}9380 \\ 173{,}8480 & 154{,}3802 \end{vmatrix} = 54{,}0861 \cdot 154{,}3802 - 47{,}9380 \cdot 173{,}8480 = 15{,}7410;$$

$$\Delta_b = \begin{vmatrix} 15 & 54{,}0861 \\ 47{,}9380 & 173{,}8480 \end{vmatrix} = 15 \cdot 173{,}8480 - 47{,}9380 \cdot 54{,}0861 = 14{,}8919;$$

$$A = \frac{\Delta_A}{\Delta} = \frac{15{,}7410}{17{,}5649} = 0{,}8962;$$

$$b = \frac{\Delta_b}{\Delta} = \frac{14{,}8919}{17{,}5649} = 0{,}8478.$$

Уравнение регрессии:

$$ln(\hat{y}) = 0{,}8962 + 0{,}8478ln(x).$$

Выполнив потенцирование, получим:

$$\hat{y}_x = 2{,}4507 \cdot x^{0{,}8478}.$$

Параметр *b*=0,8478.

Здесь показателем тесноты связи выступает индекс корреляции:

$$R = \sqrt{1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}}.$$

Величина $\sum(y-\hat{y})^2$, рассчитана в таблице 1.7, а величина $\sum(y-\bar{y})^2 = n \times \sigma_y^2 = 15 \times (10{,}112)^2$ =1533,788.

$$R = \sqrt{1 - \frac{289{,}991}{1533{,}788}} = 0{,}9005.$$

Коэффициент детерминации равен: $R^2 = 0{,}9005^2 = 0{,}811$. Значит 81,1% вариации результативного признака, объясняется вариацией факторного признака, а на долю прочих факторов приходится 18,9%.

Таблица 1.7– Вспомогательная таблица регрессионного анализа для уравнения степенной регрессии

№ п/п	X	Y	X^2	Y^2	XY	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$	$A=\left\|\frac{y-\hat{y}}{y}\right\|\cdot 100\%$
1	3,6481	4,1320	13,3086	17,0734	15,0739	54,0041	8,296	68,822	13,316
2	3,1864	3,4045	10,1531	11,5906	10,8481	36,5116	-6,412	41,109	21,301
3	3,3742	3,8565	11,3852	14,8726	13,0126	42,8132	4,487	20,131	9,486
4	3,1355	3,3979	9,8314	11,5457	10,6541	34,9695	-5,070	25,700	16,955
5	2,9014	3,6163	8,4181	13,0776	10,4923	28,6745	8,526	72,684	22,918
6	3,5025	3,8308	12,2675	14,6750	13,4174	47,7328	-1,633	2,666	3,542
7	2,6462	3,1046	7,0024	9,6385	8,2154	23,0957	-0,796	0,633	3,568
8	3,2658	3,7612	10,6654	14,1466	12,2833	39,054	3,946	15,571	9,177
9	3,0007	3,5293	9,0042	12,4560	10,5904	31,193	2,907	8,451	8,525
10	3,5553	3,8959	12,6402	15,1780	13,8511	49,9181	-0,718	0,516	1,460
11	3,4563	3,7233	11,9460	13,8630	12,8688	45,8994	-4,499	20,245	10,868
12	3,1946	3,6217	10,2055	13,1167	11,5699	36,7663	0,634	0,402	1,694
13	2,9392	3,3393	8,6389	11,1509	9,8149	29,6083	-1,408	1,983	4,994
14	3,2995	3,6109	10,8867	13,0386	11,9142	40,1859	-3,186	10,150	8,611
15	2,8332	3,2619	8,0270	10,6400	9,2416	27,0635	-0,964	0,928	3,692
Итого	47,9389	54,0861	154,3802	196,0632	173,8480	-	-	289,991	140,106
В среднем	3,196	3,606	10,292	13,071	11,590	-	-	-	9,340

F-Фишера составит:

$$F=\frac{R^2}{1-R^2}(n-2)=\frac{0{,}811}{1-0{,}811}(15-2)=55{,}78.$$

Выше было сказано, что при уровне значимости α=0,05 и числе степеней свободы $k_1=m=1,\ k_2=n-m-1=15-1-1=15-2=13$ табличное критическое значение F–критерия составляет:

$$F_{кр}=F_{\alpha=0{,}05}(k_1=1,k_2=13)=4{,}67.$$

Так как $F_н>F_{кр}$, то уравнение регрессии статистически значимое или надежное.

Из таблицы 1.7 видно, что средняя ошибка аппроксимации составила 9,340%. Значит, фактические значения стоимости валовой продукции на 1 га от расчетных по уравнению регрессии в среднем различаются на 9,340%. Полученное уравнение регрессии можно оценить как вполне хорошее.

в) Регрессия в виде экспоненты имеет вид:

$$y=a\cdot e^{bx}\cdot\varepsilon.$$

Для ее линеаризации прологарифмируем и получим уравнение в виде:

$$lny=lna+bx \text{ или } Y=A+bx,$$

где $Y=lny$, $A=lna$.

Система нормальных уравнений в этом случае примет вид:

$$\begin{cases} 54{,}0861 = 15A + 380{,}7b \\ 1397{,}25 = 380{,}7A + 10370{,}45b \end{cases}$$

Отсюда:

$$\Delta = \begin{vmatrix} 15 & 380{,}7 \\ 380{,}7 & 10370{,}45 \end{vmatrix} = 15 \cdot 1370{,}45 - 380{,}7^2 = 10624{,}26;$$

$$\Delta_A = \begin{vmatrix} 54{,}0861 & 380{,}7 \\ 1397{,}25 & 10370{,}45 \end{vmatrix} = 54{,}0861 \cdot 10370{,}45 - 1397{,}25 \cdot 380{,}7 = 28963{,}8162;$$

$$\Delta_b = \begin{vmatrix} 15 & 54{,}0861 \\ 380{,}7 & 10370{,}45 \end{vmatrix} = 15 \cdot 10370{,}45 - 380{,}7 \cdot 54{,}0861 = 368{,}1837;$$

$$A = \frac{\Delta_A}{\Delta} = \frac{28963{,}8162}{10624{,}26} = 2{,}7262;$$

$$b = \frac{\Delta_b}{\Delta} = \frac{368{,}1837}{10624{,}26} = 0{,}0347,$$

$$Y=2{,}7262+0{,}0347x$$

Выполнив потенцирование, получим:

$$a= e^{2{,}7262}=15{,}275;\ \ b=0{,}0347;$$

Степенное уравнение регрессии будет иметь следующий вид:

$$\hat{y} = e^{2{,}7262+0{,}0347x}=15{,}275e^{0{,}0347x}$$

где $e^{0{,}0347} = 0{,}8478$.

Здесь также показателем тесноты связи выступает индекс корреляции:

$$R = \sqrt{1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}}.$$

Величина $\sum(y-\hat{y})^2$, рассчитана в таблице 1.8, а величина $\sum(y-\bar{y})^2 = n \times \sigma_y^2 = 15 \times (10{,}112)^2 = 1533{,}788$.

$$R = \sqrt{1 - \frac{248{,}462}{1533{,}788}} = 0{,}915.$$

Таблица 1.8 – Вспомогательная таблица регрессионного анализа для уравнения экспоненциальной регрессии

№ п/п	x	Y	x^2	Y^2	xY	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$	$A=\left\|\frac{y-\hat{y}}{y}\right\|100\%$
1	38,4	4,1320	1474,56	17,0734	158,6688	57,8978	4,402	19,379	7,066
2	24,2	3,4045	585,64	11,5906	82,3889	35,3727	-5,273	27,801	17,517
3	29,2	3,8565	852,64	14,8726	112,6098	42,0744	5,226	27,307	11,048
4	23,0	3,3979	529,00	11,5457	78,1517	33,93	-4,030	16,241	13,478
5	18,2	3,6163	331,24	13,0776	65,8167	28,7242	8,476	71,839	22,784
6	33,2	3,8308	1102,24	14,6750	127,1826	48,3391	-2,239	5,013	4,857
7	14,1	3,1046	198,81	9,6385	43,7749	24,915	-2,615	6,838	11,726
8	26,2	3,7612	686,44	14,1466	98,5434	37,9147	5,085	25,860	11,826
9	20,1	3,5293	404,01	12,4560	70,9389	30,6818	3,418	11,684	10,024
10	35,0	3,8959	1225,00	15,1780	136,3565	51,4546	-2,255	5,083	4,583
11	31,7	3,7233	1004,89	13,8630	118,0286	45,8874	-4,487	20,137	10,839
12	24,4	3,6217	595,36	13,1167	88,3695	35,619	1,781	3,172	4,762
13	18,9	3,3393	357,21	11,1509	63,1128	29,4305	-1,230	1,514	4,363
14	27,1	3,6109	734,41	13,0386	97,8554	39,1175	-2,118	4,484	5,723
15	17,0	3,2619	289,00	10,6400	55,4523	27,5527	-1,453	2,110	5,566
Итого	380,7	54,0861	10370,45	196,0632	1397,251	-	-	248,462	146,163
Среднее значение	25,38	3,6060	691,363	13,071	93,150	-	-	-	9,744

Коэффициент детерминации равен: $R^2 = 0{,}915^2 = 0{,}837$. Значит 83,7% вариации результативного признака, объясняется вариацией факторного признака, а на долю прочих факторов приходится 16,3%.

F-Фишера составит:

$$F = \frac{R^2}{1-R^2}(n-2) = \frac{0{,}837}{1-0{,}837}(15-2) = 66{,}75.$$

При уровне значимости α=0,05 и числе степеней свободы $k_1 = m = 1$, $k_2 = n - m - 1 = 15 - 1 - 1 = 15 - 2 = 13$ табличное критическое значение *F*–критерия составляет:

$$F_{кр} = F_{\alpha=0{,}05}(k_1 = 1, k_2 = 13) = 4{,}67.$$

Так как $F_н > F_{кр}$, то уравнение регрессии статистически значимое или надежное.

Из таблицы 1.8 видно, что средняя ошибка аппроксимации составила 9,744%. Значит, фактические значения стоимости валовой продукции на 1 га от

расчетных по уравнению регрессии в среднем различаются на 9,744%. Полученное уравнение регрессии можно оценить как вполне хорошее.

г) Показательная регрессия имеет вид:

$$y = a \cdot b^x \cdot \varepsilon.$$

Прологарифмировав, получим линейное уравнение относительно параметров:

$$Y=A+Bx,$$

где $A=lna;$
$B=lnb;$
$Y=lny.$

Таблица 1.9 – Вспомогательная таблица регрессионного анализа для уравнения показательной регрессии

№ п/п	x	y	Y	x^2	Y^2	xy	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$	$A=\left\|\frac{y-\hat{y}}{y}\right\| 100\%$
1	38,4	62,3	4,1320	1474,56	17,0734	158,6688	57,8782	4,422	19,552	7,098
2	24,2	30,1	3,4045	585,64	11,5906	82,3889	35,3651	-5,265	27,721	17,492
3	29,2	47,3	3,8565	852,64	14,8726	112,6098	42,0636	5,236	27,420	11,071
4	23,0	29,9	3,3979	529,00	11,5457	78,1517	33,9231	-4,023	16,185	13,455
5	18,2	37,2	3,6163	331,24	13,0776	65,8167	28,7196	8,480	71,917	22,797
6	33,2	46,1	3,8308	1102,24	14,6750	127,1826	48,3249	-2,225	4,950	4,826
7	14,1	22,3	3,1046	198,81	9,6385	43,7749	24,9119	-2,612	6,822	11,713
8	26,2	43,0	3,7612	686,44	14,1466	98,5434	37,9060	5,094	25,949	11,847
9	20,1	34,1	3,5293	404,01	12,4560	70,9389	30,6763	3,424	11,722	10,040
10	35,0	49,2	3,8959	1225,00	15,1780	136,3565	51,4387	-2,239	5,012	4,550
11	31,7	41,4	3,7233	1004,89	13,8630	118,0286	45,8745	-4,475	20,021	10,808
12	24,4	37,4	3,6217	595,36	13,1167	88,3695	35,6113	1,789	3,199	4,783
13	18,9	28,2	3,3393	357,21	11,1509	63,1128	29,4255	-1,226	1,502	4,346
14	27,1	37,0	3,6109	734,41	13,0386	97,8554	39,1081	-2,108	4,444	5,698
15	17,0	26,1	3,2619	289,00	10,6400	55,4523	27,5485	-1,449	2,098	5,550
Итого	380,7	571,6	54,086	10370,45	196,06	1397,26	-	-	248,514	146,072
В среднем	25,380	38,107	3,606	691,363	13,071	93,150	-	-	-	9,738

Итак, параметр а=15,275; $b = 0{,}8478$.

$$\begin{cases} 54{,}0861 = 15A + 380{,}7B \\ 1397{,}25 = 380{,}7A + 10370{,}45B \end{cases}$$

Отсюда:

$$\Delta = \begin{vmatrix} 15 & 380{,}7 \\ 380{,}7 & 10370{,}45 \end{vmatrix} = 15 \cdot 1370{,}45 - 380{,}7^2 = 10624{,}26;$$

$$\Delta_A = \begin{vmatrix} 54{,}0861 & 380{,}7 \\ 1397{,}25 & 10370{,}45 \end{vmatrix} = 54{,}0861 \cdot 10370{,}45 - 1397{,}25 \cdot 380{,}7 = 28959{,}277;$$

$$\Delta_B = \begin{vmatrix} 15 & 54{,}0860 \\ 380{,}7 & 1397?26 \end{vmatrix} = 15 \cdot 1397?26 - 380{,}7 \cdot 54{,}0861 = 368{,}36;$$

$$A = \frac{\Delta_A}{\Delta} = \frac{28959{,}277}{10624{,}26} = 2{,}726;$$

$$B = \frac{\Delta_B}{\Delta} = \frac{368{,}36}{10624{,}26} = 0{,}035,$$

$$Y=2{,}726+0{,}035x$$

Выполнив потенцирование, получим:

$$a = e^{2{,}7262} = 15{,}275; \quad b = e^{0{,}035} = 0{,}8478;$$

Показательное уравнение регрессии будет иметь следующий вид:

$$\hat{y} = 15{,}275 \cdot 0{,}8487^{x}$$

Индекс корреляции будет равен:

$$R = \sqrt{1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}};$$

$$R = \sqrt{1 - \frac{248{,}514}{1533{,}788}} = 0{,}915.$$

Коэффициент детерминации равен: $R^2 = 0{,}915^2 = 0{,}837$. Значит 83,7% вариации результативного признака, объясняется вариацией факторного признака, а на долю прочих факторов приходится 16,3%.

F-Фишера составит:

$$F = \frac{R^2}{1-R^2}(n-2) = \frac{0{,}837}{1-0{,}837}(15-2) = 66{,}75.$$

Табличное критическое значение *F*–критерия составляет:

$$F_{кр} = F_{\alpha=0{,}05}(k_1 = 1, k_2 = 13) = 4{,}67.$$

Так как $F_н > F_{кр}$, то уравнение регрессии статистически значимое или надежное.

Из таблицы 1.9 видно, что средняя ошибка аппроксимации составила 9,738%. Значит, фактические значения стоимости валовой продукции на 1 га от расчетных по уравнению регрессии в среднем различаются на 9,738%. Полученное уравнение регрессии можно оценить как вполне хорошее.

Замечание.

Важнейшим методом анализа данных является визуализация (представление данных в виде таблиц, диаграмм, кросс-таблиц, кросс-диаграмм, графиков). Рассмотрим применение диаграммы рассеяния.

Выделим в *Excel* диапазон B2:C16 (рисунок 1.5), выполним команду:

Вставка - Точечная – Точечная с маркерами.

В результате получим рисунок 1.6.

Важность графического представления данных заключается в возможности увидеть возможные ошибки, допущенные при вводе данных (артефакты – объекты созданные человеком) или неоднородные значения признаков - выбросы – явно не принадлежащие изучаемой совокупности.

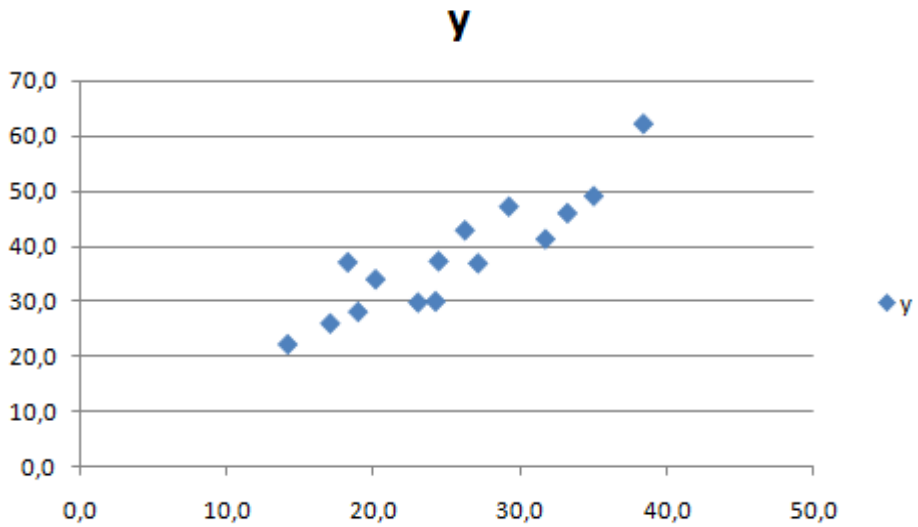


Рисунок 1.6 - Диаграмма рассеяния

Например, при вводе исходных данных мы вместо 62,3 ввели 623. Построим соответствующую диаграмму рассеяния (рисунок 1.6) из которой видно, что есть наблюдение, отличающееся от других данных.

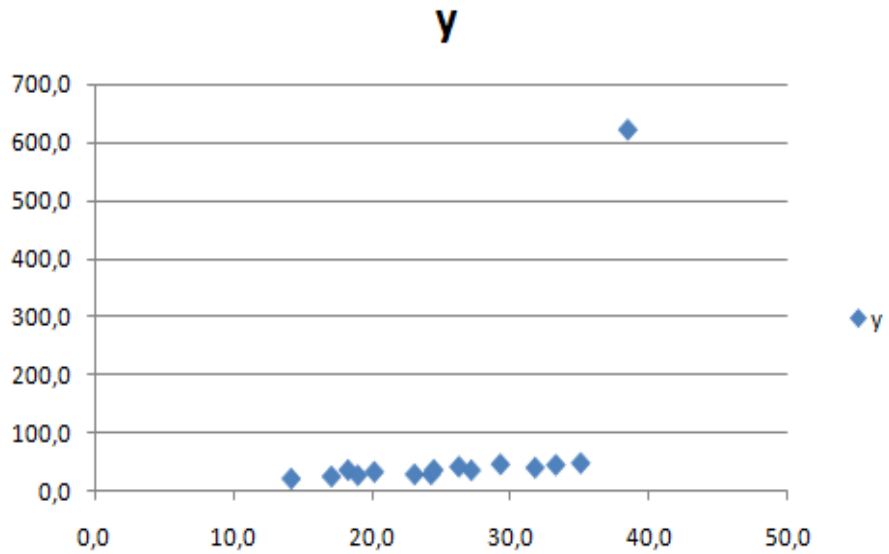


Рисунок 1.7 – Диаграмма рассеяния с артефактом (или выбросом)

Важным методом анализа данных в *Excel* являются диаграммы. Выделим на рисунке 1.7 щелчком левой клавиши мыши маркеры наблюдений; с помощью правой клавиши откроем контекстное меню (рисунок 1.8) и выберем одну из перечисленных линий трендов (рисунок 1.9):

- Линейная;
- Логарифмическая;
- Полиномиальная;
- Степенная;
- Экспоненциальная;
- Линейная фильтрация (скользящая средняя).

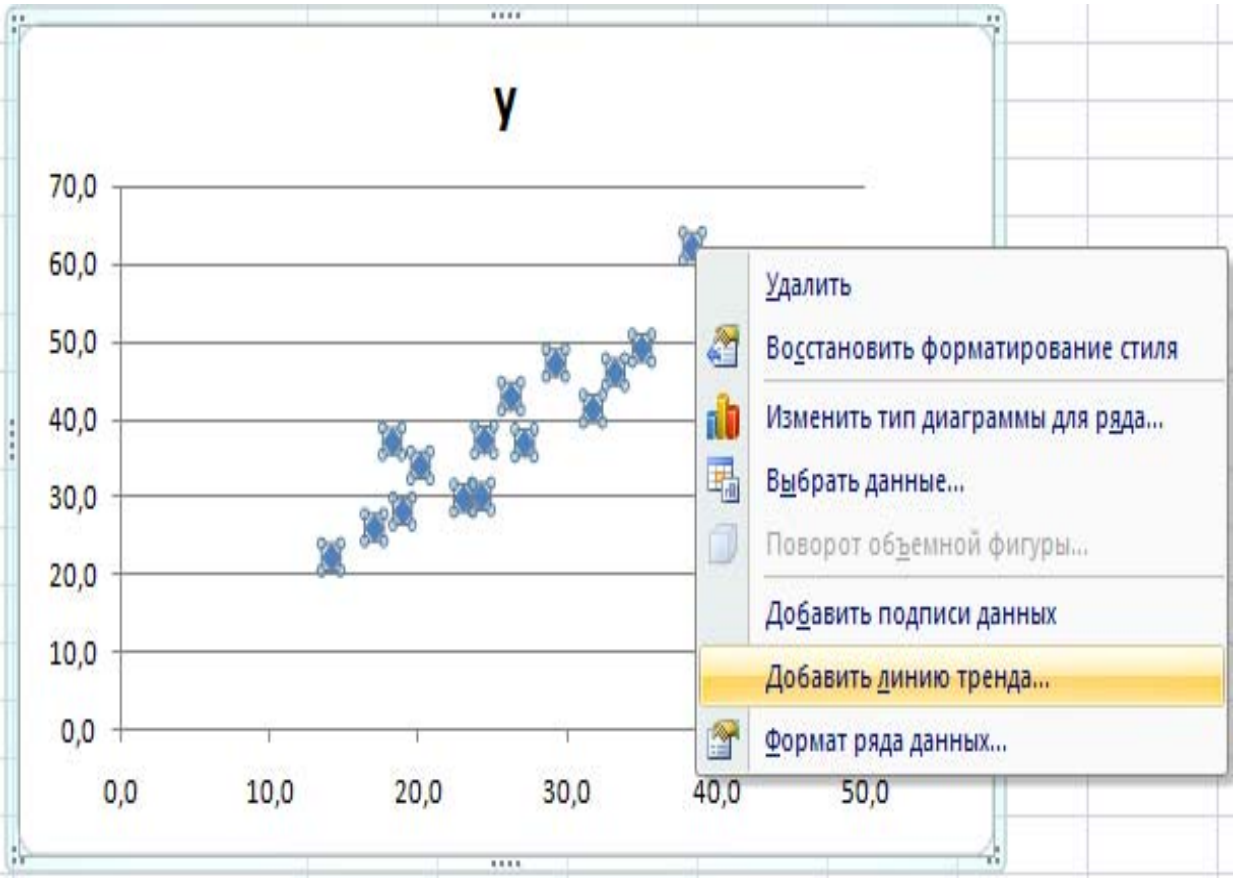


Рисунок 1.8 – Контекстное меню выделенных точек наблюдений

После выбора одного из трендов, например, линейного - выберем и заполним вкладку Параметры диалогового окна (рисунок 1.9).

Можно выбрать название (назвать тренд самостоятельно) или оставить автоматически предлагаемое *Excel*; для прогноза согласно выбранной линии тренда на 5 лет вперёд выберем соответствующее значение в диалоговом окне; для отображения на диаграмме уравнения тренда и коэффициента детерминации отметим соответствующие элементы вкладки Параметры (рис. 1.9). Далее выберем *OK*.

MSExcel позволяет проиллюстрировать основное свойство коэффициента корреляции – характеристики «степени линейности» наблюдаемой совокупности пар данных.

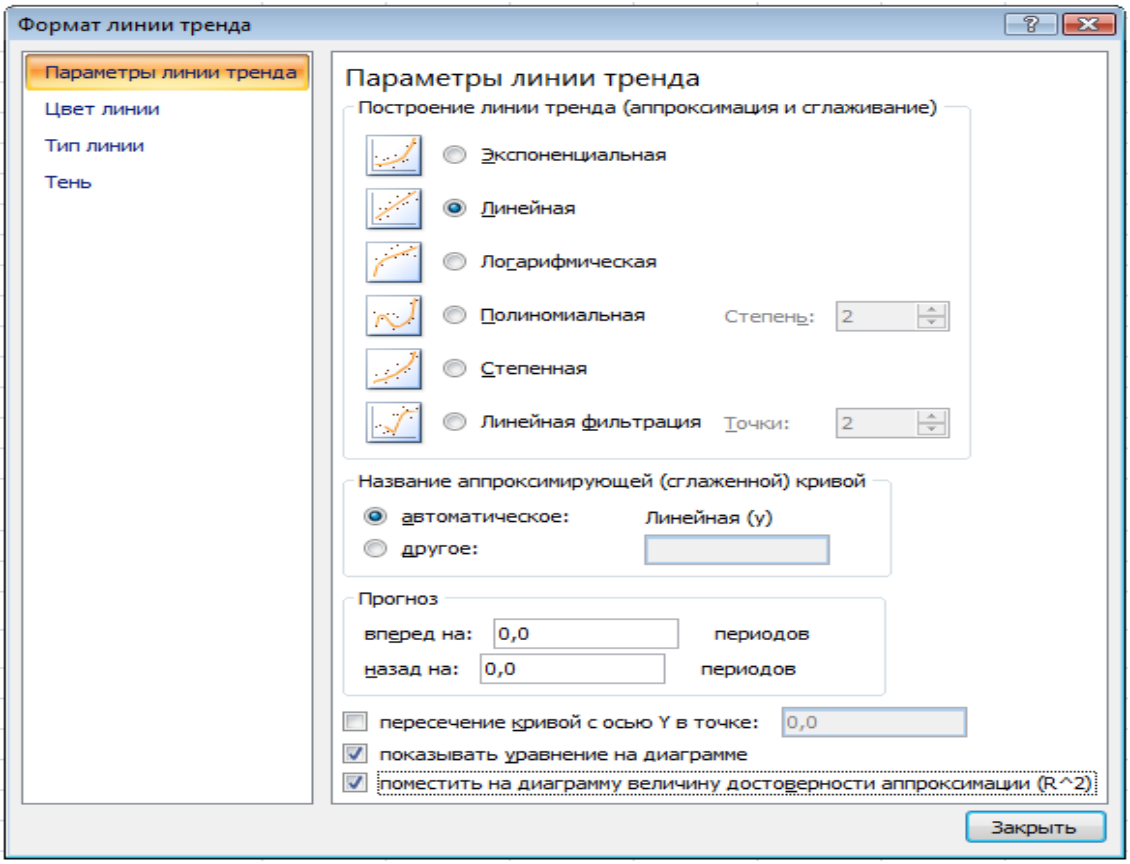


Рисунок 1.9 – Диалоговое окно выбора линии тренда

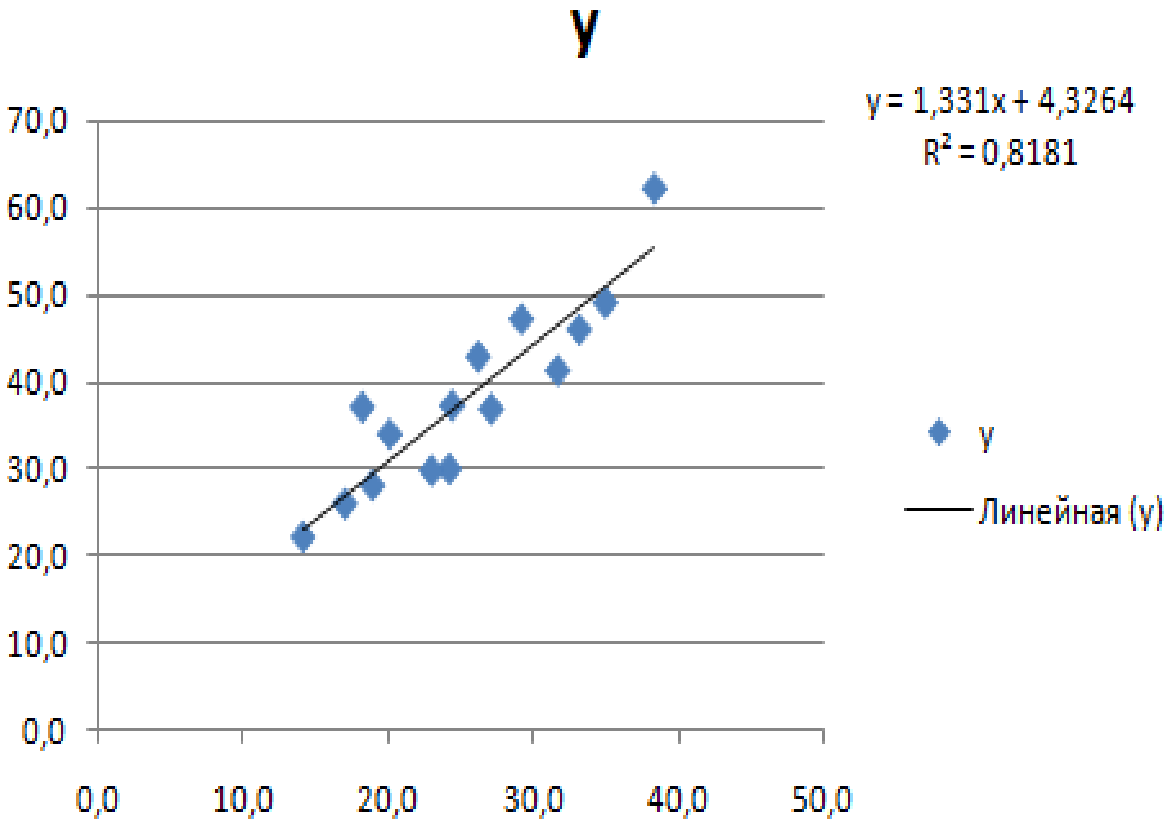


Рисунок 1.10 – График линейного уравнения

Предлагается провести имитационный эксперимент:

1. Ввести 100 пар значений (x_i, y_i), подчиняющихся зависимости $y=2x+3(x=\{1, 2, \ldots, 100\})$.

2. Сгенерировать 100 значений случайной величины ε, подчиняющейся равномерному закону распределения на промежутке [-10; 10](генератор случайных чисел в пакете анализа).

3. Наложить значения случайной величины ε на переменную Y: $V=Y+\varepsilon$.

4. Построить зависимость V от X.

5. Изменить границы равномерного закона на [-50; 50].

6. Построить зависимость V от X.

7. Изменить знак перед переменой хв уравнении зависимости:$y= - 2x+3$. и наложить шум (ε).

8. Сделать выводы о свойствах коэффициентов корреляции и регрессии.

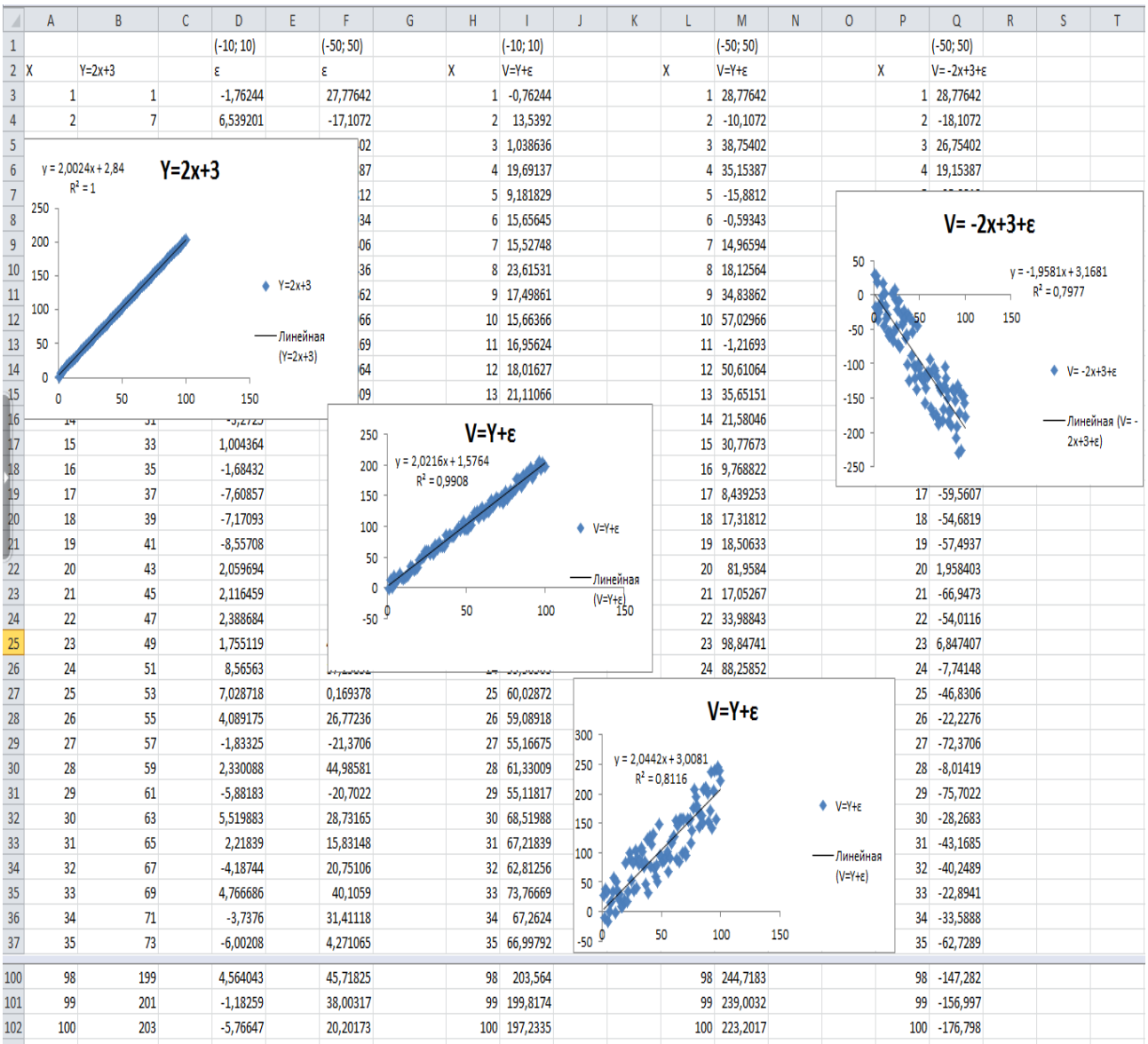


Рисунок 1.11 – Возможные результаты имитационного эксперимента

В новой русифицированной версии программы *STATISTICA 10*: выбор команды Основные статистики и таблицы – Г1 (2) позволяет визуально представить результаты первичного анализа. Рассмотрим применительно к примеру 1.1.

Графики ствол и листья являются альтернативой Гистограмм. Как и гистограммы, графики ствол и листья могут быть построены для всех выбранных переменных.

Так же как и в гистограмме, на графике каждый ствол означает интервал, но на гистограмме откладываются вертикальные столбцы (число наблюдений), а на графиках этого типа реальные значения отражаются в виде листьев.

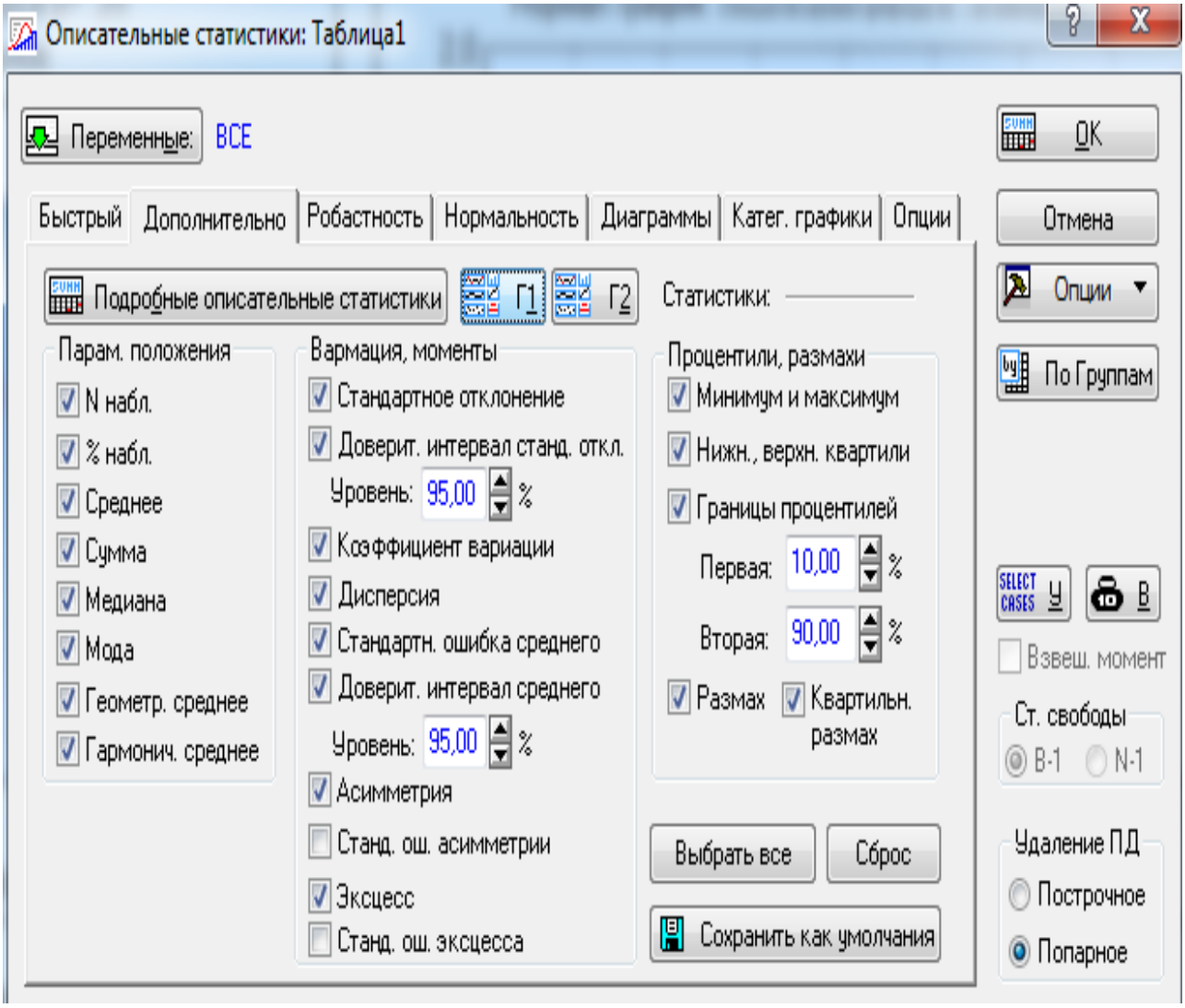


Рисунок 1. 12 – Описательные статистики

Введя команду*Графика-Диаграмма рассеяния – 2М Графики – Диаграммы рассеяния,* получим рисунок 1.15.

Гр. ствол и листья: *Фондообеспеченность на 1 га сельхозугодий, тыс. руб., (x)* (Таблица1)
Фондообеспеченность на 1 га сельхозугодий, тыс. руб., (x)
1 лист=1 набл.

ствол°лист (лист. ед.=1,000000, напр.,6°5 = 6,500000)	Класс n	Процентили
10°	0	
11°	0	
12°	0	
13°	0	
14° 1	1	
15°	0	
16°	0	
17° 0	1	
18° 29	2	25%
19°	0	
20° 1	1	
21°	0	
22°	0	
23° 0	1	
24° 24	2	медиана
25°	0	
26° 2	1	
27° 1	1	
28°	0	
29° 2	1	
30°	0	
31° 7	1	75%
32°	0	
33° 2	1	
34°	0	
35° 0	1	
36°	0	
37°	0	
38° 4	1	
39°	0	
мин = 14,10000 макс= 38,40000	Всего N: 15	

Рисунок 1.13 - Диаграмма ствол и листья для переменной *Фондообеспеченность на 1 га*

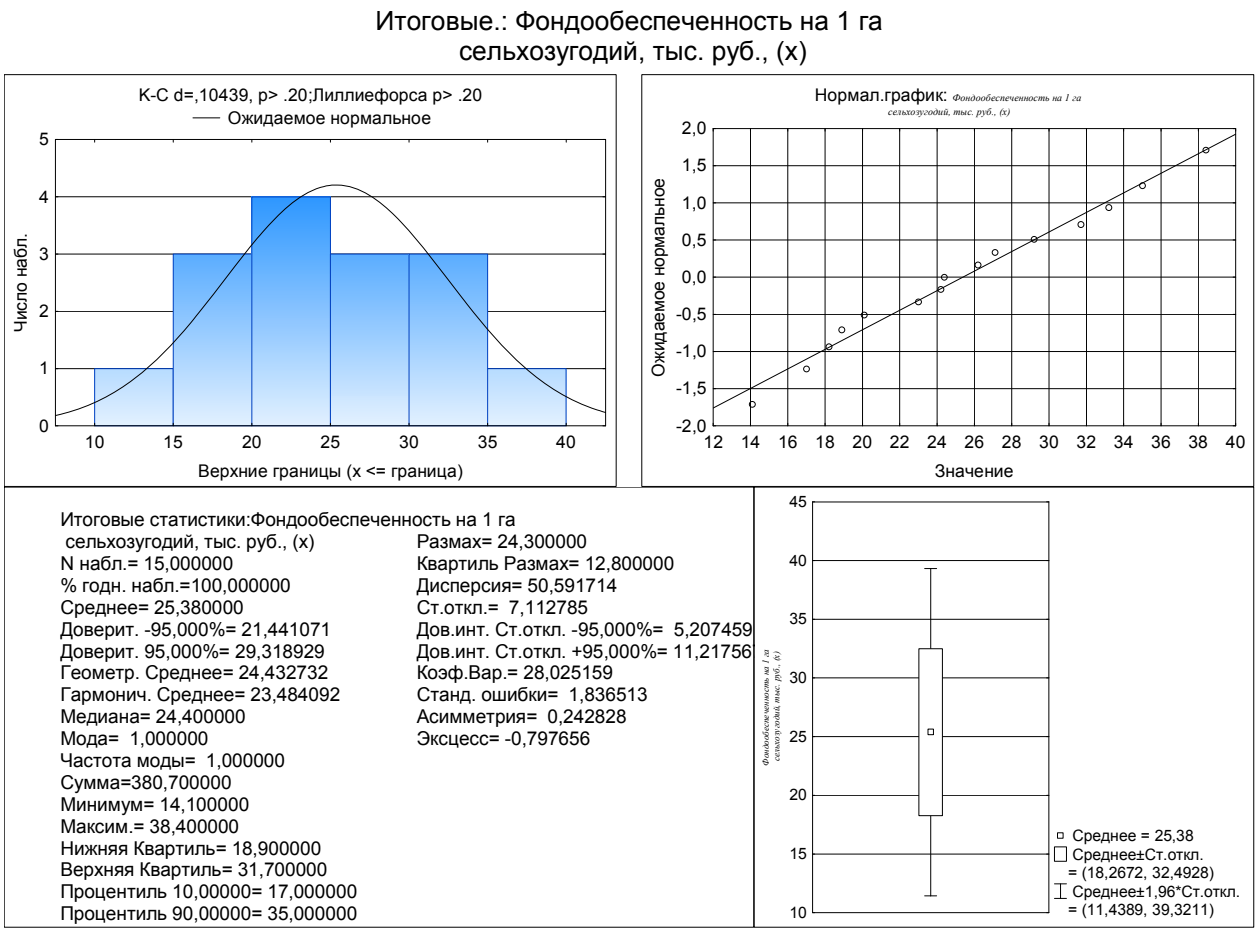


Рисунок 1.14 - Итоги применения инструмента *Описательная статистика* по переменной фондообеспеченность

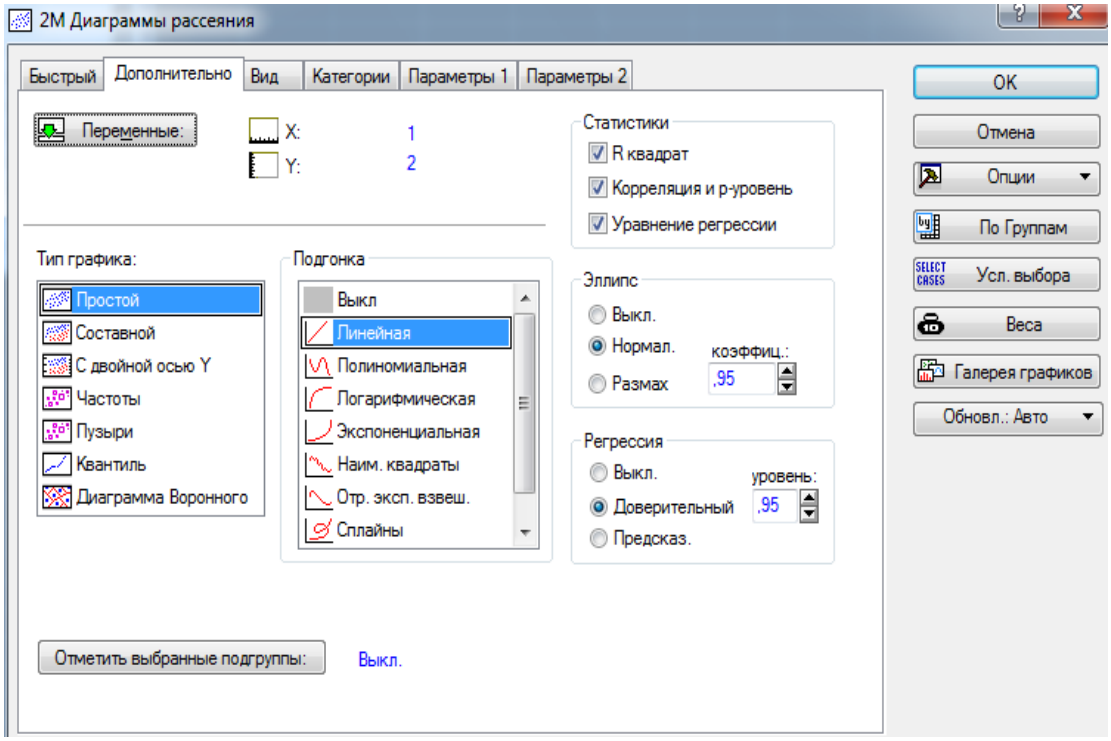


Рисунок 1.15 –Окно «Диаграмма рассеяния»

Итоговые.: Стоимость валовой продукции на 1 га сельхозугодий, тыс. руб., (y)

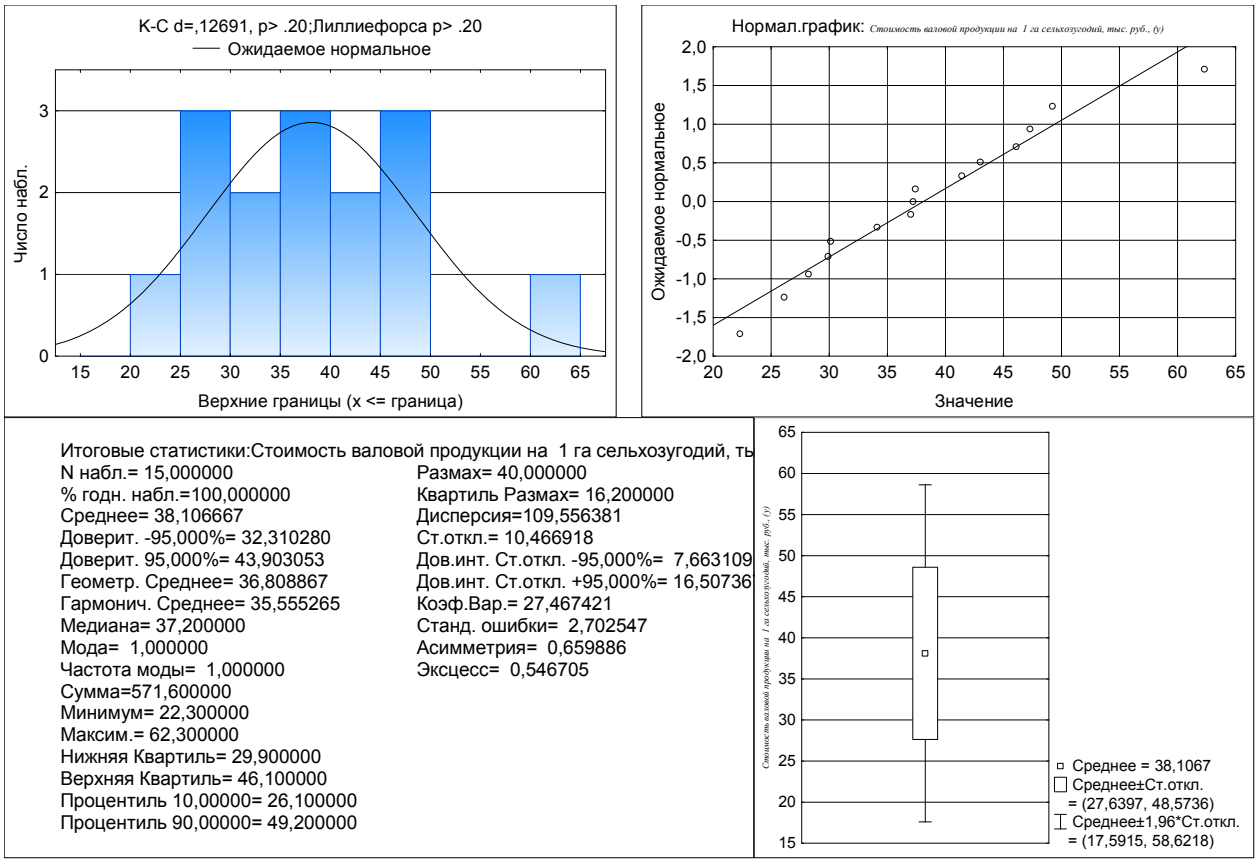


Рисунок 1.16 - Итоги применения инструмента *Описательная статистика* по переменной стоимость валовой продукции на 1 га угодий

Диаграмма рассеяния для *Стоимость валовой продукции на 1 га сельхозугодий, тыс. руб., (y)* и *Фондообеспеченность на 1 га сельхозугодий, тыс. руб., (x)* y = 4,3264 + 1,331*x; r = 0,9045; p = 0,00000; r^2 = 0,8181

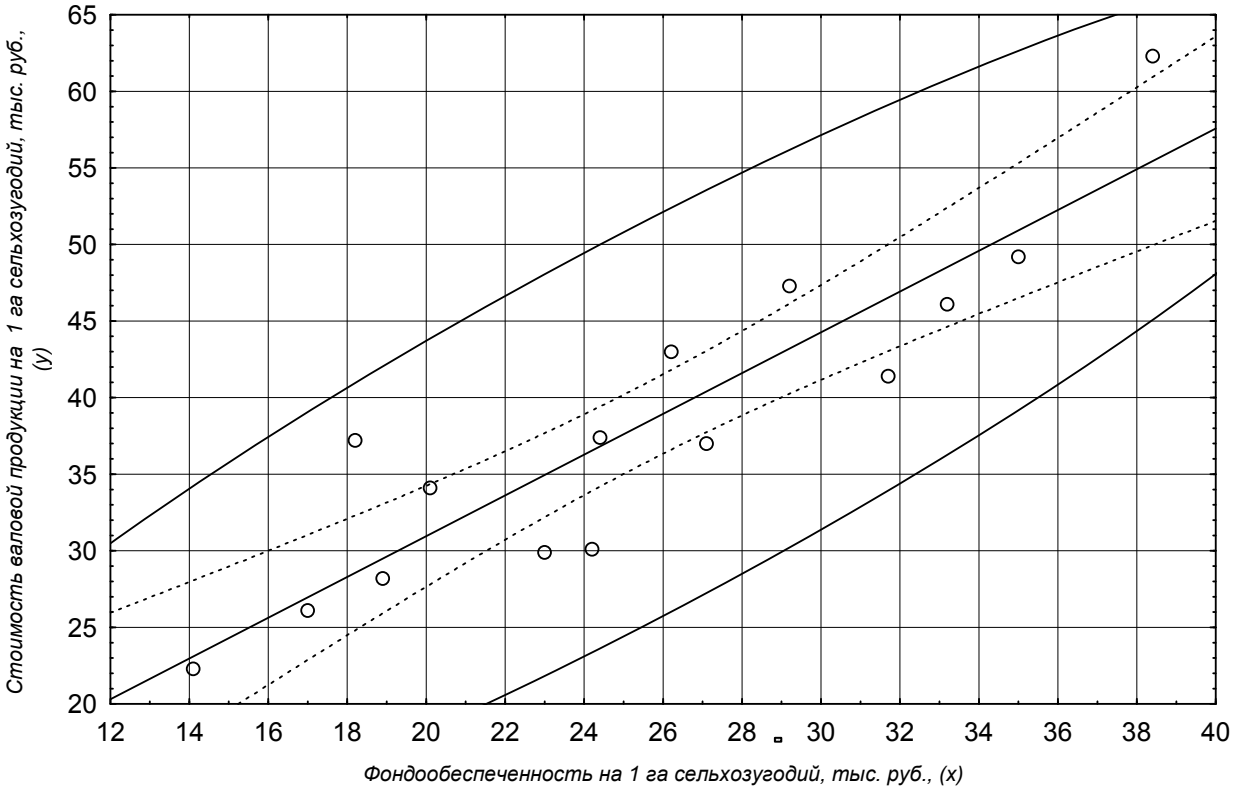


Рисунок 1.17 – Диаграмма рассеяния

Задача 1.1

Получены функции:

1. $y = a + bx^c + \varepsilon$,
2. $lny = a + bx + \varepsilon$,
3. $ln\, y = a + b\, ln\, x + \varepsilon$,
4. $y^a = a + bx^3 + \varepsilon$,
5. $y = a + b\frac{X}{20} + \varepsilon$,
6. $y = 10 + a(x^b - 5) + \varepsilon$.

Определить, какие из представленных выше функций линейны по переменным, линейны по параметрам, не линейны ни по переменным, ни по параметрам.

Задача 1.2

Исследуя спрос на телевизоры марки *N*, аналитический отдел компании ABC по данным, собранным по 25 торговым точкам компании, выявил следующую зависимость:

$$ln\, y = \underset{(2,1)}{9{,}5} - \underset{(-3,0)}{0{,}7}\, ln\, x + \varepsilon\ ,$$

где *y* – объем продаж телевизоров марки *N* в отдельной торговой точке;

x – средняя цена телевизора в данной торговой точке;

В скобках приведены фактические значения *t*-критерия Стьюдента для параметров уравнения регрессии.

До проведения этого исследования администрация компании предполагала, что эластичность спроса по цене для телевизоров марки N составляет -0,85. Подтвердилось ли предположение администрации результатами исследования?

Задача 1.3

Для трех видов продукции A, B и C модели зависимости удельных постоянных расходов от объема выпускаемой продукции выглядят следующим образом:

$$y_A = 200,$$
$$y_B = 70 + 0{,}3x,$$
$$y_C = 30x^{0{,}2}.$$

Задание:

а) Определите коэффициенты эластичности по каждому виду продукции и поясните их смысл.

б) Сравните при x = 100 эластичность затрат для продукции B и C.

в) Определите, каким должен быть объем выпускаемой продукции, чтобы коэффициенты эластичности для продукции B и C были равны.

Задача 1.4

Пусть имеется следующая модель регрессии, характеризующая зависимость y от x:

$y = 10 - 5x + \varepsilon$,

Известно также, что $r_{xy} = -0{,}6$, n = 25.

Задание:

1)Постройте доверительный интервал для коэффициента регрессии в этой модели:

а) с вероятностью 90%,

б) с вероятностью 99%.

2) Проанализируйте результаты, полученные в п.1 и поясните причины их различия.

Задача 1.5

Изучается зависимость потребления материалов *Y* от объема производства продукции *X*. По 20 наблюдениям были получены следующие варианты уравнения регрессии:

1. $y = 4 + \underset{(5{,}32)}{3}x + \varepsilon,$

2. $\ln y = 3{,}5 + \underset{(5{,}21)}{0{,}1} \ln x + \varepsilon, \;\; r^2 = 0{,}7,$

3. $\ln y = 2{,}3 + \underset{(5{,}0)}{0{,}7} \ln x + \varepsilon, \;\; r^2 = 0{,}69,$

4. $y = 4 + \underset{(3{,}0)}{1{,}2}x + \underset{(2{,}8)}{0{,}3}x^2 + \varepsilon, \;\; r^2 = 0{,}72.$

В скобках указаны фактические значения t-критерия

Задание:

1.Определите коэффициент детерминации 1-го уравнения.

2. Запишите функции, характеризующие зависимость у от х во 2-м и 3-м уравнениях.

3. Определите коэффициенты эластичности для каждого из уравнений.

4. Выберите наилучший вариант уравнения регрессии.

Задача 1.6

Изучается зависимость потребления продукта А (Y) от среднедушевого дохода (X) по данным 20 семей. При оценке регрессионной модели были получены результаты:

$$\sum(y_j - \hat{y}_x)^2 = 1{,}2;$$

$$\sum(y_j - \bar{y})^2 = 6{,}3.$$

Задание:

1) Какой показатель корреляции можно определить по этим данным.

2) Постройте таблицу дисперсионного анализа для расчёта значения F-критерия Фишера.

3) Сравните фактическое значение F-критерия.

Задача 1.7.

Зависимость среднемесячной производительности труда от возраста рабочих характеризуется моделью: $y=a+bx+cx^2$. Её использование привело к результатам, представленным в таблице.

№ п/п	Производительность труда рабочих, тыс. руб., у	
	Фактическая	Расчетная
1	22	20
2	18	20
3	23	23
4	25	24
5	26	25
6	21	22
7	22	23
8	19	20
9	21	20
10	19	19

Оцените качество модели, определив ошибку аппроксимации, индекс корреляции и *F*-критерий Фишера.

Задача 1.8.

Моделирование зависимости розничного товарооборота (млн.руб.) магазинов от среднесписочного числа работников по уравнению $y=ab^x$ привело к результатам, представленным ниже.

Розничный товарооборот y, млн. руб.		
№ магазина	Фактический	Расчетный
1	0,5	0,43
2	0,7	0,66
3	0,9	0,99
4	1,1	1,24
5	1,4	1,37
6	1,4	1,45
7	1,7	1,60
8	1,9	1,85

Оцените качество модели:

– определить среднюю ошибку аппроксимации;

– найти показатель тесноты связи с исследуемым в модели фактором. Сделать выводы.

– рассчитайте F-критерий Фишера. Сделайте выводы.

Задача 1.9

Изучалась зависимость вида $Y = aX^b$. Для преобразованных в логарифмах переменных получены следующие данные:

$$\sum xy = 2{,}1927; \qquad \sum x^2 = 1{,}1693; \qquad \sum x = 2{,}0792;$$

$$\sum y = 4{,}9255. \qquad \sum\left(Y - \hat{Y}_x\right)^2 = 1{,}2;$$

1. Найдите параметр b.
2. Найдите показатель корреляции, предполагая $\sigma_Y = 2{,}45$.
3. Оцените его значимость, если n=5.

Задача 1.10

Зависимость объема производства y (тыс.ед) от численности занятых x (чел.) по 20 заводам концерна характеризуется следующим образом:

уравнение регрессии: $y = 25 - 0{,}3x + 0{,}06x^2$,

доля остаточной дисперсии в общей: 20%.

Определите:

а) индекс корреляции;

б) значимость уравнения регрессии;

в) коэффициент эластичности, предполагая, что численность занятых составляет 35 человек.

Задача 1.11

По группе 10 заводов, производящих однородную продукцию, получено уравнение регрессии себестоимости единицы продукции y (тыс. руб.) от уровня технической оснащённости x (тыс. руб.): $y = 30 + \frac{600}{x}$.

Доля остаточной дисперсии в общей составила 0,16.

Определите:

1) коэффициент эластичности, предполагая, что стоимость активных производственных фондов составляет 250 тыс. руб.;

2) индекс корреляции;

3) *F*-критерий Фишера. Сделайте выводы.

Задача 1.12

Зависимость спроса на товар К от его цены характеризуется по 20 наблюдениям уравнением: $lg\, y = 2{,}15 - 0{,}45lgx$. Доля остаточной дисперсии в общей составила 15%.

1. Запишите данное уравнение в виде степенной функции.

2. Оцените эластичность спроса на товар в зависимости от его цены.

3. Определите индекс корреляции.

4. Оцените значимость уравнения регрессии через *F*-критерий Фишера. Сделайте выводы.

Задача 1.13

По 20 семьям получена информация, представленная в таблице:

Показатель	Среднее значение	Коэффициент вариации
Годовой расход на личное потребление, у.е.	3378	3,2
Годовой располагаемый доход, у.е.	3508	2,8

Фактическое значение *F*-критерия Фишера составило72.

1. Определите линейный коэффициент детерминации.

2. Постройте уравнение линейной регрессии.

3. Найдите средний коэффициент эластичности.

4. С вероятностью 0,95 укажите доверительный интервал ожидаемого значения годового расхода в предположении роста годового дохода на 10% от своего среднего уровня.

Задача 1.14

Для двух видов продукции А и В зависимость расходов предприятия y (тыс. руб.) от объёма производства x (шт.) характеризуется данными, представленными в таблице.

Уравнение регрессии	Показатели корреляции	Число наблюдений
$y_A = 120 + 0{,}6x$	0,80	35
$y_B = 40x^{0{,}5}$	0,75	30

1. Поясните смысл величин 0,6 и 0,5 в уравнениях регрессии.

2. Сравните эластичность расходов от объёма производства для продукции А и В при выпуске продукции А в 400 единиц.

3. Определите, каким должен быть выпуск продукции А, чтобы эластичность её расходов совпадала с эластичностью расходов на продукцию В.

4. Оцените значимость каждого уравнения регрессии с помощью *F*-критерия Фишера.

Задача 1.15

Зависимость объёма продаж y (тыс. долл.) от расходов на рекламу *x* (тыс. долл.) характеризуется по 12 предприятиям концерна следующим образом:

уравнение регрессии: $y = 58{,}5 + 2{,}4x$,

среднее квадратическое отклонение *x:* $\sigma_x = 2{,}9$,

среднее квадратическое отклонение *y:* $\sigma_y = 8{,}1$.

1. Определите коэффициент корреляции.

2. Постройте таблицу дисперсионного анализа для оценки значимости уравнения регрессии в целом.

3. Найдите стандартную ошибку оценки коэффициента регрессии.

4. Оцените значимость коэффициента регрессии через *t*-критерий Стьюдента.

5. Определите доверительный интервал для коэффициента регрессии с вероятностью 0,95 и сделайте экономический вывод.

Задача 1.16

По 20 регионам страны изучается зависимость уровня безработицы y (%) от индекса потребительских цен x (% к предыдущему году). Информация о логарифмах исходных показателей представлена в таблице.

Показатель	*ln x*	*ln y*
Среднее значение	0,5	0,9
Среднее квадратичное отклонение	0,3	0,2

Известно также, что коэффициент корреляции между логарифмами исходных показателей составил $r_{lnxlny} = 0{,}7$.

1. Постройте уравнение регрессии зависимости уровня безработицы от индекса потребительских цен в степенной форме.

2. Дайте интерпретацию коэффициента эластичности данной модели регрессии.

3. Определите значение коэффициента детерминации и поясните смысл.

2 Множественный корреляционно - регрессионный анализ

Парная регрессия применяется в случае, когда результативный признак (*У*) определяется влиянием одного, доминирующего фактора (*Х*). Но в большинстве экономических исследованиях результативный признак формируется, как правило, под влиянием не одного, а нескольких факторных признаков $X_1, X_2, \ldots, X_p$. К таким задачам можно отнести изучение спроса, потребления, урожайности сельскохозяйственных культур и продуктивности животных, производительности труда, себестоимости продукции, рентабельности, уровня жизни населения, безработицы, производства продукции и т. п.

Уравнение множественной регрессии имеет вид

$$\hat{y} = f(x_1, x_2, x_3, \ldots, x_p). \quad (2.1)$$

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

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon, \quad (2.2)$$

где $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ – параметры модели,

ε – свободный член или остаток,

p – количество переменных.

Свободный член в классической нормальной линейной регрессионной модели удовлетворяет условиям Гаусса – Маркова:

а) свободный член (остаток) является случайной величиной;

б) математическое ожидание свободного члена ε в любом наблюдении равно нулю, $M(\varepsilon_i) = 0, i = 1,2, \ldots, n$;

в) дисперсия свободного члена постоянна для любого наблюдения $D(\varepsilon_i) = \sigma^2$, это предположение называется условием гомоскедастичности;

г) случайные члены ε_i и ε_j не коррелированы, cov($\varepsilon_i, \varepsilon_j$) = 0;

д) остаткиε_i распределяются по нормальному закону.

Если не выполняется пятое условие, то модель называется классической линейной моделью множественной регрессии.

Пусть из генеральной совокупности взята случайная выборка объема n. По выборке определены значения результативного и факторных признаков по i - тому наблюдению $(y_i, x_{i1}, x_{i2}, \ldots, x_{ip}), i = 1,2,\ldots,n$. Тогда линейная модель множественной регрессии будет иметь вид:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i. \quad (2.3)$$

Параметры модели определяются методом наименьших квадратов.

По выборочным данным находятся оценки $b_0, b_1, b_2, \ldots, b_p$ неизвестных параметров $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.

Эти оценки являются несмещенными, эффективными и состоятельными оценками параметров генеральной совокупности. Параметры линейной модели множественной регрессии$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip} + \varepsilon_i$ определяются исходя из условия минимизации суммы квадратов отклонений наблюдаемых значений результативного признака (y_i) от теоретических значений результативного признака, найденных по уравнению регрессии $(\hat{y}_i)$, т. е. минимизации суммы квадратов остатков. Условие имеет вид:

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2 \rightarrow min. \qquad (2.4)$$

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

$$\frac{\partial S}{\partial b_0} = 0, \frac{\partial S}{\partial b_1} = 0, \ \ldots, \frac{\partial S}{\partial b_p} = 0 \qquad (2.5)$$

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

$$\begin{cases} \Sigma y = n b_o + b_1 \Sigma x_1 + b_2 \Sigma x_2 + \ldots + b_p \Sigma x_p, \\ \Sigma y x_1 = b_o \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2 + \ldots + b_p \Sigma x_1 x_p, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \Sigma y x_p = b_o \Sigma x_p + b_1 \Sigma x_1 x_p + b_2 \Sigma x_2 x_p + \ldots + b_p \Sigma x_p^2. \end{cases} \qquad (2.6)$$

В матричной форме уравнение множественной регрессии имеет вид:

$$Y = XB + \varepsilon, \qquad (2.7)$$

где Y – вектор-столбец размерности $(n \times 1)$;

X– матрица наблюдаемых значений факторных признаков размерности $(n \times (p+1))$;

B – вектор-столбец коэффициентов регрессии $((p+1) \times 1)$;

ε– вектор-столбец остаточных значений результативного признака размерности $(n \times 1)$.

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; B = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix}; \ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}; \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix}. \qquad (2.8)$$

Тогда оценки параметров линейной модели *b*находятся из выражения:

$$b = (X^T X)^{-1} X^T Y, \quad (2.9)$$

где X^T – транспонированная матрица *X;*
$(X^T X)^{-1}$ – обратная матрица.

При построении уравнения множественной регрессии обычно используются следующие нелинейные функции:

степенная $y = b_o \cdot x_1^{b_1} \cdot x_2^{b_2} \ldots x_p^{b_p} \cdot \varepsilon$; (2.10)

экспонента $y = e^{b_o + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p + \varepsilon}$; (2.11)

гипербола $y = b_0 + \frac{b_1}{x_1} + \frac{b_2}{x_2} + \ldots + \frac{b_p}{x_p} + \varepsilon$; (2.12)

логлинейная $lny = b_0 + b_1 lnx_1 + b_2 lnx_2 + \cdots + b_p lnx_p + \varepsilon$. (2.13)

Довольно часто применяются и другие виды функций, например

$$y = b_o + b_1 x_1 + b_2 x_2 + b_3 x_1^2 + b_4 x_2^2 + b_5 x_1 x_2 + \varepsilon . \quad (2.14)$$

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

Множественный коэффициент регрессии (b_j) показывает, на сколько единиц изменяется в среднем результативный признак *y*, если *j* – тый 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:

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \ldots + \beta_p t_{x_p}, \; t_y = \frac{y - \bar{y}}{\sigma_y}, t_j = \frac{x_j - \bar{x}_j}{\sigma_{x_j}}, \quad (2.15)$$

где β_j - стандартизованные коэффициенты регрессии;

σ_y и σ_{xj} – средние квадратические отклонения по переменным x_j и *y*.

Параметры уравнения регрессии (2.15) определяются методом наименьших квадратов путем составления и решения следующей системы уравнений:

$$\begin{cases} r_{yx_1} = \beta_1 + \beta_2 r_{x_1 x_2} + \cdots + \beta_p r_{x_1 x_p} \\ r_{yx_2} = \beta_1 r_{x_1 x_2} + \beta_2 + \cdots + \beta_p r_{x_2 x_p} \\ \cdots \\ \cdots \\ r_{yx_p} = \beta_1 r_{x_1 x_p} + \beta_2 r_{x_2 x_p} + \cdots + \beta_p \end{cases} . \quad (2.16)$$

Зная стандартизованные коэффициенты можно получить множественные коэффициенты регрессии:

$$b_j = \beta_j \cdot \frac{\sigma_y}{\sigma_{x_j}}, b_0 = \overline{y} - b_1\overline{x}_1 - b_2\overline{x}_2 - \ldots - b_p\overline{x}_p. \quad (2.17)$$

По абсолютной величине β– 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 относительного изменения результативного признака *у*котносительному изменению факторного признака x_j:

$$\mathcal{Э}_{x_j} = \frac{dy}{\hat{y}} : \frac{dx_j}{x_j} = \frac{dy}{dx_j} \cdot \frac{x_j}{\hat{y}}. \quad (2.18)$$

По линейной модели множественной регрессии коэффициент эластичности для заданного *i*-го значения фактора x_jопределяется по формуле:

$$\mathcal{Э}_{x_j} = b_j \frac{x_j}{b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}}. \quad (2.19)$$

Если в формуле (2.19) значения факторов принять на среднем уровне, то будет получен средний коэффициент эластичности, который показывает, на сколько процентов в среднем изменится результативный признак, если *j*-ый фактор увеличить на один процент, при условии что все другие факторы закреплены на среднем уровне.

$$\overline{\mathcal{Э}}_{x_j} = b_j \frac{\overline{x}_j}{\overline{y}}. \quad (2.20)$$

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:

$$R_{yx_1x_2\ldots x_p} = \sqrt{1 - \frac{\sigma^2_{\text{ост.}}}{\sigma^2_y}} = \sqrt{\frac{\sigma^2_{\text{рег.}}}{\sigma^2_y}} = \sqrt{1 - \frac{\sum_i^n (y - \hat{y}_{x_1x_2\ldots x_p})^2}{\sum_i^n (y - \overline{y})^2}} = \sqrt{1 - \frac{SS_{\text{ост.}}}{SS_{\text{общ.}}}}, \quad (2.21)$$

где σ^2_y– общая дисперсия результативного признака,

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

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

причем $\sigma_y^2 = \sigma_{\text{рег.}}^2 + \sigma_{\text{ост.}}^2$; $\sigma_y^2 = \frac{\sum(y-\overline{y})^2}{n}$; $\sigma_{\text{рег.}}^2 = \frac{\sum(\hat{y}-\overline{y})^2}{n}$; $\sigma_{\text{ост.}}^2 = \frac{\sum(y-\hat{y})^2}{n}$.(2.22)

$$SS_{\text{общ.}} = SS_{\text{факт.}} + SS_{\text{ост.}} \quad \text{или} \quad \sum(y-\overline{y})^2 = \sum(\hat{y}-\overline{y})^2 + \sum(y-\hat{y})^2, \qquad (2.23)$$

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

$SS_{\text{факт.}}$– факторная сумма квадратов отклонений (обусловленная регрессией);

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

(*SS –SummSquare*).

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 (2.15), то множественный коэффициент детерминации рассчитывается по формуле:

$$R^2_{yx_1x_2\ldots x_p} = \beta_1 r_{yx_1} + \beta_2 r_{yx_2} + \ldots + \beta_p r_{yx_p} = \sum \beta_j r_{x_j}. \qquad (2.24)$$

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

$$r_{yx_j \cdot x_1x_2\ldots x_{j-1}x_{j+1}\ldots x_p} = \sqrt{1 - \frac{1-R^2_{yx_1x_2\ldots x_j\ldots x_p}}{1-R^2_{yx_1x_2\ldots x_{j-1}x_{j+1}\ldots x_p}}}, \qquad (2.25)$$

или

$$r_{yx_j \cdot x_1x_2\ldots x_{j-1}x_{j+1}\ldots x_p} = \frac{r_{yx_j \cdot x_1x_2\ldots x_{j-1}x_{j+1}\ldots x_{p-1}} - r_{yx_p \cdot x_1x_2\ldots x_{j-1}x_{j+1}\ldots x_{p-1}} \cdot r_{x_jx_p \cdot x_1\ldots x_{j-1}x_{j+1}\ldots x_{p-1}}}{\sqrt{(1-r^2_{yx_p \cdot x_1x_2\ldots x_{j-1}x_{j+1}\ldots x_{p-1}})(1-r^2_{x_jx_p \cdot x_1\ldots x_{j-1}x_{j+1}\ldots x_{p-1}})}} \qquad (2.26)$$

В формуле (2.26) 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:

$$R^2_{\text{ск}} = 1 - \frac{\sum(y-\hat{y})^2 : (n-m-1)}{\sum Y-\overline{y})^2 : (n-1)}, \qquad (2.27)$$

$$R^2_{\text{ск}} = 1 - (1-R^2) \cdot \frac{(n-1)}{(n-m-1)}, \qquad (2.28)$$

где m – число параметров уравнения регрессии без учета свободного члена. В линейном уравнении $m=p$.

Адекватность линии регрессии зависит от того, какая часть суммы квадратов относительно среднего обусловлена суммой квадратов относительно регрессии, а какая – суммой квадратов, обусловленной регрессией. Суммы квадратов связаны с числом степеней свободы $\nu=df$. Это число показывает, сколько независимых элементов информации (из n чисел $y_1, y_2,...y_n$) необходимо для образования данной суммы квадратов. Например, для $\sum(y_i-\overline{Y})^2$, $df=(n-1)$. Действительно, из n разностей $(y_1-\overline{Y}),(y_2-\overline{Y}),...,(y_n-\overline{Y})$ только (n-1) независимы (или иначе, для образования рассматриваемой суммы из $y_1, y_2,\ldots,y_n$ достаточно (n - 1) значение, так как оставшееся можно определить, зная $\overline{Y}$). Аналогично для $\sum(\hat{y}_i-\overline{Y})^2$, $df=m-1$; для $\sum(y_i-\hat{y}_i)^2$, $df=n-m-1$.

Для построения таблицы дисперсионного анализа необходимо рассчитать средние квадраты (*MS*) (*MeamSquare*), для этого каждая сумма *SS* делится на соответствующие число степеней свободы $df=k$ ($MS_R=\frac{SS_{рег.}}{m}, S^2=\frac{SS_{ост.}}{n-m-1}$).

Если в уравнении регрессии $(y=b_0+b_1x_1+...+b_mx_m)$ $b_1=b_2=...=b_m=0$ (или $R^2=0$), то величина $F=\frac{MS_R}{S^2}$ распределена по распределению Фишера с (m, n-m-1) степенями свободы. Этот факт используется для проверки гипотезы $H_0: b_1=b_2=...=b_m=0$ с уровнем значимости α, против альтернативы H_1: хотя бы одно $b_j\neq0$ $(j=1,...,m)$.

В результате составляется таблица дисперсионного анализа (табл.2.1).

Таблица 2.1 - Дисперсионный анализ

Источник вариации	Число степеней свободы, *df*	Суммы квадратов, *SS*	Средние квадраты, *MS*	$F_н$	$F_{кр.}$
Обусловленный регрессией	*m*	$SS_{факт.}=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2$	$MS_R=\frac{SS_{факт.}}{m}$	$F=\frac{MS_R}{S^2}$	$F_\alpha(m, n-m-1)$
Относительно регрессии (остаток)	*n-m-1*	$SS_{ост.}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$S^2=\frac{SS_{ост.}}{n-m-1}$		
Общий, скорректированный на среднее *Y*	*n-1*	$SS_{общ.}=\sum_{i=1}^{n}(y_i-\overline{y})^2$			

Оценка значимости множественного уравнения регрессии производится с помощью *F* – критерия Фишера-Снедекора. Определяется наблюдаемое значение критерия по следующей формуле:

$$F_{\text{н}} = \frac{SS_{\text{факт.}}}{m} : \frac{SS_{\text{ост.}}}{n-m-1} = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}. \quad (2.29)$$

При заданном уровне значимости α и числе степеней свободы факторной $(k_1 = m)$и остаточной дисперсий$(k_2 = n - m - 1)$по таблицам находится критическое значение критерия Фишера-Снедекора. Сравнивается наблюдаемое и критическое значения критерия. Если $F_{\text{н}} < F_{\text{кр}}$, то нулевая гипотеза о незначимости уравнения регрессии принимается. Если $F_{\text{н}} > F_{\text{кр}}$, то нулевая гипотеза отвергается и принимается альтернативная гипотеза о статистической значимости всего множественного уравнения регрессии.

Доля суммы квадратов, объясняемая регрессией называется множественным коэффициентом детерминации (квадратом множественного коэффициента корреляции R):

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}, 0 \le \text{R} \le 1. \quad (2.30)$$

Если *Y* зависит только от одной переменной *X*, то *R*=*r*– парному коэффициенту корреляции.

Оценка значимости параметров множественного линейного уравнения регрессии производится с помощью *t* – критерия Стьюдента. Выдвигается основная гипотеза о равенстве нулю параметров уравнения регрессии $(\text{H}_0: \beta_j = 0)$, при конкурирующей гипотезе, что параметры уравнения отличны от нуля $(\text{H}_0: \beta_j \neq 0)$. Наблюдаемое значение *t* – критерия для параметра уравнения b_j определяется по формуле:

$$t_{b_j} = \frac{b_j}{s_{b_j}}, \quad s_{b_j} = \sqrt{\frac{SS_{\text{ост.}}}{n-p-1}[(X^T X)^{-1}]_{jj}}, \quad (2.31)$$

где $s_{b_j} -$ стандартная ошибка параметра уравнения регрессии b_j,

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

Стандартная ошибка множественного коэффициента регрессии b_jможет быть найдена также по формуле:

$$s_{b_j} = \frac{\sigma_y}{\sigma_{x_j}} \sqrt{\frac{1-R^2_{yx_1x_2\ldots x_p}}{\left(1-R^2_{x_j x_1 x_2 \ldots x_{j-1} x_{j+1} \ldots x_p}\right)(n-m-1)}}, \quad (2.32)$$

где $\sigma_y -$ 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 $\left|t_{b_j}\right| > \left|t_{\text{кр}}\right|$, то параметр уравнения статистически значим. Если$\left|t_{b_j}\right| < \left|t_{\text{кр}}\right|$, то параметр уравнения статистически не значим и *j*– ая переменная ис-

ключается из уравнения регрессии. Доверительные интервалы для коэффициентов линейного уравнения регрессии находятся по формуле:

$$b_j \pm t_{\text{кр}} s_{b_j}. \quad (2.33)$$

Пример 2.1.

По данным 20 сельскохозяйственных предприятий центральной зоны Краснодарского края за 2011 год исследовать зависимость объема реализованной продукции с единицы земельной площади от обеспеченности основными фондами, рабочей силой и земельными ресурсами.

Результативным признаком (y) является стоимость реализованной продукции на 1 га сельскохозяйственных угодий, тыс. руб.

Факторные признаки:

x_1 – среднегодовая стоимость основных фондов на 1 га сельскохозяйственных угодий, тыс. руб.;

x_2 – среднегодовая численность работников, занятых в сельскохозяйственном производстве на 100 га сельскохозяйственных угодий, чел.;

x_3 – площадь сельскохозяйственных угодий на одно предприятие, га;

x_4 – энергетические мощности на 1 га сельскохозяйственных угодий, л. с.

Определить:

а) обобщающие статистические характеристики по каждой переменной;

б) парные коэффициенты корреляции между всеми переменными;

в) наличие или отсутствие мультиколлинеарности между факторами;

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

д) средние коэффициенты эластичности для каждого фактора;

е) коэффициенты частной и множественной корреляции;

ж)значимость множественного уравнения регрессии в целом с помощью общего критерия F – Фишера;

з) значимость множественных коэффициентов регрессии с использованием критериев Фишера и Стьюдента;

и) доверительные интервалы множественных коэффициентов регрессии при уровне доверительной вероятности 0,95.

Решение.

Рассмотрим применение пакета анализа данных в *Excel MS Office 2007*для решения задачи. Исходные данные введем на листе *MSExcel* в виде, представленном таблицей 2.2.

Таблица 2.2 – Исходные данные для регрессионного анализа в *MSExcel*

№ п/п	y	x_1	x_2	x_3	x_4
1	27,71042	24,29720	3,62838	4327	2,48902
2	26,01304	18,04788	2,57281	5597	1,66964

Продолжение таблицы 2.2

№ п/п	y	x_1	x_2	x_3	x_4
3	34,14714	22,40120	1,94072	2834	1,46824
4	30,12978	33,62928	2,89235	3976	2,95221
5	38,15353	29,27760	2,34034	7093	1,17947
6	34,48997	24,83687	3,29015	8723	2,35619
7	33,94965	25,16215	3,86916	12871	1,66763
8	61,87436	48,82279	6,13314	9734	3,56482
9	22,30012	8,66436	1,42344	7517	2,06466
10	32,04532	27,64991	2,46715	7458	1,01502
11	38,15472	34,18335	3,14768	22874	2,69559
12	49,48773	52,45759	4,34499	7664	3,81654
13	45,05932	49,40448	3,47968	4282	3,30009
14	50,81574	51,11557	2,90657	5780	3,77775
15	31,24106	24,85636	5,06338	14042	3,50534
16	23,50313	20,88697	4,27607	3999	2,68067
17	21,86536	31,61165	1,59496	2696	1,36498
18	62,93002	65,00027	8,31354	10946	4,52019
19	22,70692	25,25396	2,32941	7899	1,98152
20	32,15883	36,11076	1,91235	3765	1,92404

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 на вкладке *Данные* ленты появится инструмент *Пакет анализа*.

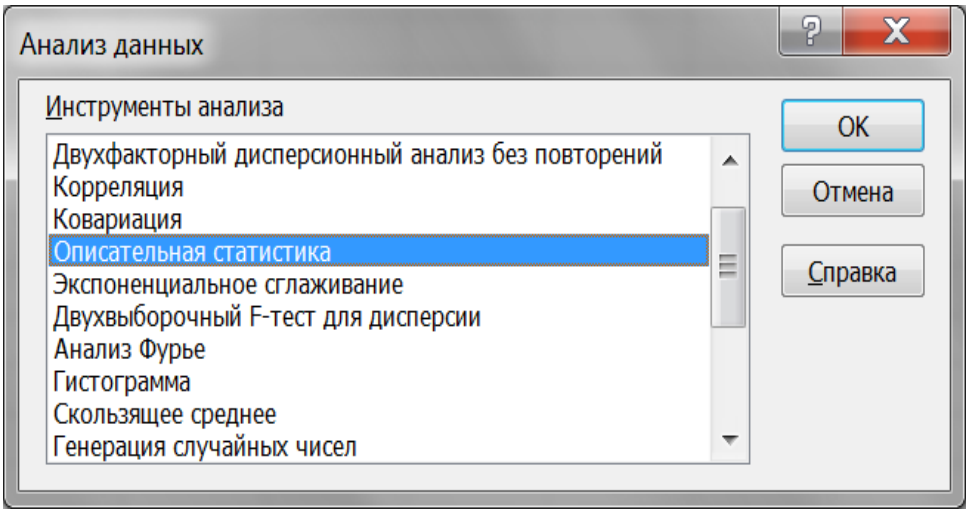


Рисунок 2.9 – Диалоговое окно пакета анализа данных

Выберем в *Пакете анализа* инструмент *Описательная статистика* и заполним параметры диалогового окна (рисунок 2.9). В результате будут рассчитаны описательные статистики по каждому признаку (таблица 2.3).

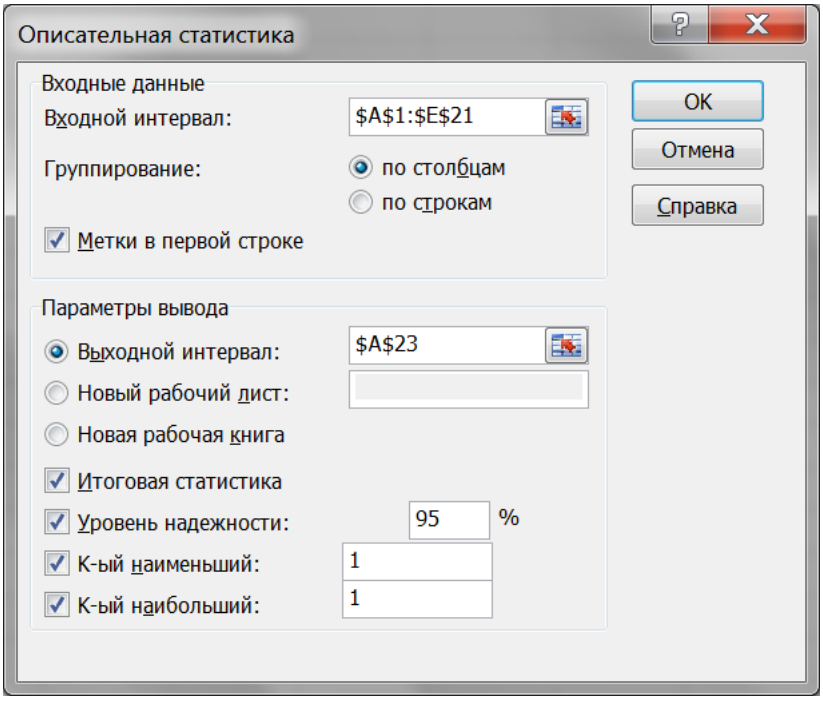


Рисунок 2.10 – Диалоговое окно описательной статистики

Таблица 2.3 – Описательные статистики признаков
по совокупности сельскохозяйственных организаций

Показатель	y	x_1	x_2	x_3	x_4	Принятые обозначения
Среднее значение	35,94	32,68	3,40	7703	2,50	$\overline{X} = \sum x_i n_i / n$
Стандартная ошибка	2,75	3,12	0,37	1072	0,22	$s_{\overline{X}} = s / \sqrt{n}$
Медиана	33,05	28,46	3,03	7276	2,42	M_e
Мода	н/д	н/д	н/д	н/д	н/д	M_o
Стандартное отклонение	12,30	13,97	1,66	4792	1,00	s
Дисперсия выборки	151,24	195,22	2,75	2296659	1,00	$s^2 = \sum (x_i - \overline{X})^2 n_i /(n-1)$
Эксцесс	0,31	0,13	3,05	4,34	-0,87	$Ex = \sum ((x_i - \overline{X}) / S)^4 n_i / n - 3$
Асимметричность	1,01	0,74	1,56	1,82	0,34	$Sk = \sum ((x_i - \overline{X}) / S)^3 n_i / n$
Интервал	41,06	56,33	6,89	20178	3,50	$W = x_{max} - x_{min}$
Минимум	21,86	8,66	1,42	2696	1,02	x_{min}
Максимум	62,93	65,00	8,31	22874	4,52	x_{max}
Сумма	718,74	653,67	67,93	154077	49,99	Σx_i
Счет	20	20	20	20	20	$n = \Sigma n_i$
Наибольший(1)	62,93	65,00	8,31	22874	4,52	-
Наименьший(1)	21,86	8,66	1,42	2696	1,02	-
Уровень надежности(95,0%)	5,76	6,54	0,78	2242	0,47	$\Delta = t_{\alpha;n-1} s_{\overline{X}}$

Данные таблицы 2.3 показывают, что по совокупности предприятий средняя стоимость реализованной продукции на 1 га сельскохозяйственных угодий составила 35,94 тыс. руб. и в среднем между предприятиями она варьирует в границах 35,94 ± 12,30 тыс. руб., т.е. от 23,64 до 48,24 тыс. руб. Коэффициент вариации составил 34,2 %, что свидетельствует об очень больших различиях в выручке от реализации продукции на 1га сельскохозяйственных угодий между предприятиями. По значению медианы видно, что половина предприятий имеет размер выручки до 33,05 тыс. руб./га, а половина более. Распределение предприятий по данному признаку имеет правостороннюю асимметрию ($К_{ас}$=1,01) и является средневершинным ($Э$=0,31). Наименьшее значение выручки на 1 га сельскохозяйственных угодий составило 21,86, а наибольшее – 62,93 тыс. руб.

Аналогичные выводы можно сделать и по факторным признакам x_1, x_2, x_3 и x_4.

Для нахождения парных коэффициентов корреляции применим инструмент пакета анализа «*Корреляция*», для этого заполним параметры диалогового окна как на рисунке 2.11.

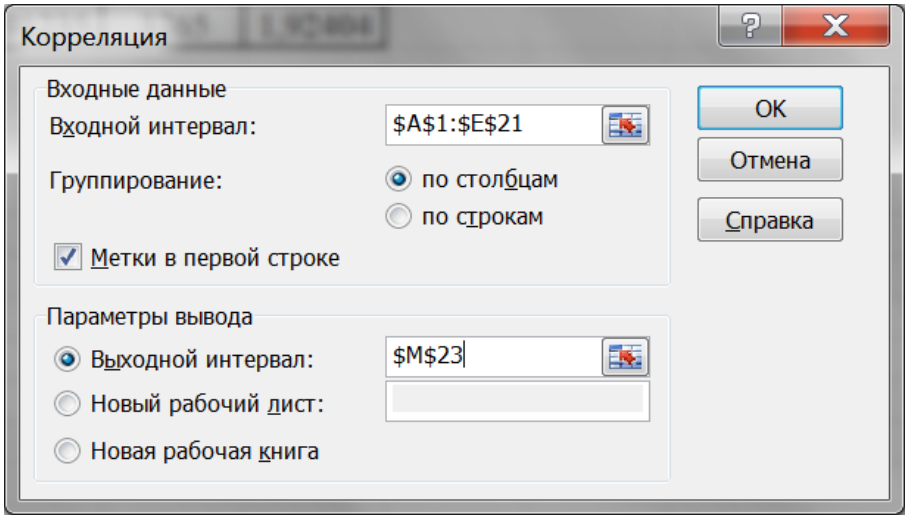


Рисунок 2.11 – Диалоговое окно «Корреляция»

В результате будет получена матрица парных коэффициентов корреляции между всеми изучаемыми переменными (таблица 2.4).

Таблица 2.4 – Парные коэффициенты корреляции между признаками

	y	x_1	x_2	x_3	x_4
y	1,0000				
x_1	0,8778	1,0000			
x_2	0,7098	0,6196	1,0000		
x_3	0,2518	0,0927	0,3564	1,0000	
x_4	0,6974	0,7229	0,7614	0,2407	1,0000

Значит: r_{yx_1}= 0,8778; r_{yx_2}= 0,7098; r_{yx_3}= 0,2518, r_{yx_4} = 0,6974. Парные коэффициенты корреляции показывают, что связь между выручкой от реализации на 1 га сельскохозяйственных угодий, с одной стороны, фондооснащенностью, трудообеспеченностью и энергооснащенностью с другой, довольно тесная, между выручкой от реализации на 1 га сельскохозяйственных угодий и площадью сельскохозяйственных угодий на одно предприятие – слабая.

Наблюдается сильная связь между факторными признаками x_1 и x_2, x_1 и x_4, x_2 и x_4. Высокие значения парных коэффициентов корреляции между этими факторами могут свидетельствовать о наличии мультиколлинеарности, когда более чем два фактора связаны линейной зависимостью. Для оценки мультиколлинеарности факторов найдем определитель матрицы парных коэффициентов корреляции между факторами.

Таблица 2.5 – Парные коэффициенты корреляции между факторами

	x_1	x_2	x_3	x_4
x_1	1,0000	0,6196	0,0927	0,7229
x_2	0,6196	1,0000	0,3564	0,7614
x_3	0,0927	0,3564	1,0000	0,2407
x_4	0,7229	0,7614	0,2407	1,0000

С помощью функции *МОПРЕД* найдем определитель матрицы: $det|R_1|$= 0,331. Наблюдаемое значение статистики Фаррара – Глоубера вычислим по формуле:

$$FG_{\text{н}} = -\left[n - 1 - \frac{1}{6}(2m + 5)\right] \cdot \ln(det|R_1|); \qquad (2.34)$$

$$FG_{\text{н}} = -\left[20 - 1 - \frac{1}{6}(2 \cdot 4 + 5)\right] \ln(0{,}331) = 18{,}62.$$

Статистика Фаррара – Глоубера имеет приближенное распределение χ^2с $k = \frac{1}{2}m(m-1)$ степенями свободы. При уровне значимости α = 0,05 и числе степеней свободы $k = \frac{1}{2} \cdot 4 \cdot 3 = 6$, критическое значение χ^2– критерия равно 12,59. Сравнивается наблюдаемое и критическое значения критерия. Так как наблюдаемое значение критерия *хи*-квадрат больше критического, то нулевая гипотеза об отсутствии мультиколлинеарности отвергается. Значит, доказано наличие значимой мультиколлинеарности.

В данной задаче фондооснащенность и энергооснащенность являются общим и частным показателями оснащенности производства основными фондами. Поэтому следует исключить оснащенность энергетическими ресурсами, как менее связанной с выручкой и более тесно с другими факторами. Таким образом целесообразно построить множественное уравнение регрессии результативного признака y с факторами x_1, x_2и x_3.

Линейное уравнение множественной регрессии в натуральной формеимеет вид: $Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$.

Найдем параметры этого уравнения, используя инструмент:

Пакета анализа – Регрессия.

Заполним параметры диалогового окна (рисунок 2.12).

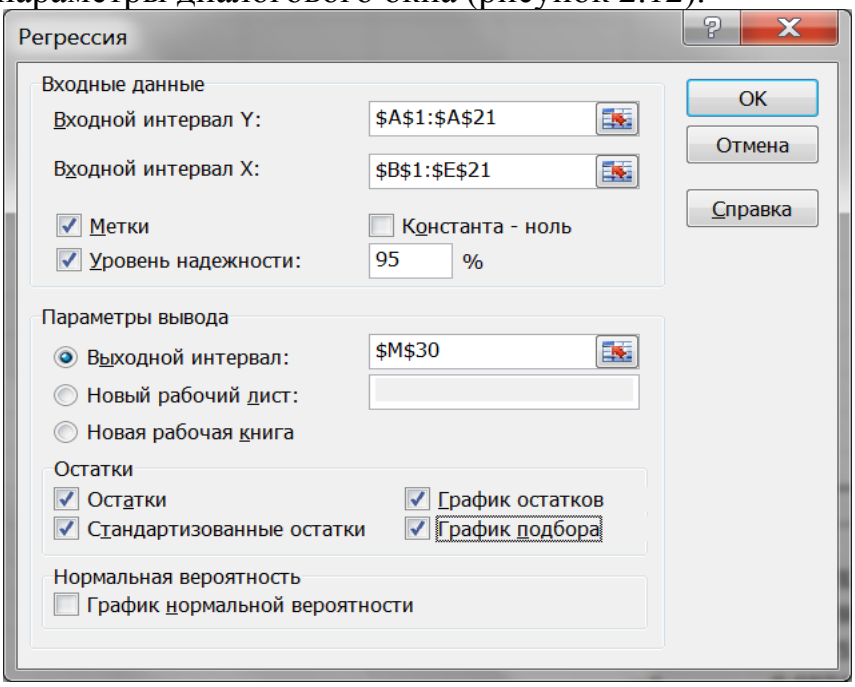


Рисунок 2.12– Диалоговое окно «Регрессия»

Регрессионная статистика	
Множественный R	0,908187
R-квадрат	0,824803
Нормированный R-квадрат	0,791954
Стандартная ошибка	5,609427
Наблюдения	20

Дисперсионный анализ						
	df	*SS*	*MS*	*F*	Значи-мость *F*	
Регрессия	3	2370,182	790,0607	25,108	2,72E-06	
Остаток	16	503,451	31,4657			
Итого	19	2873,633				
	Коэффи-циенты	Стандарт-ная ошибка	*t*-статистика	*P*-Значение	Нижние 95%	Верхние 95%
y-пересечение	7,262886	3,737278	1,94336	0,06977	-0,65979	15,1855
x_1	0,645231	0,119178	5,41402	5,74E-5	0,392586	0,89787
x_2	1,615269	1,070309	1,50916	0,15075	-0,65369	3,88422
x_3	0,000273	0,000292	0,93401		-0,00035	0,00089

Рисунок 2.13– Вывод итогов регрессионного анализа

Линейное уравнение множественной регрессии имеет вид:

$$y=\underset{1,94}{7,263}+\underset{5,41}{0,645x_1}+\underset{1,509}{1,615x_2}+\underset{0,934}{0,000272x_3}$$

Коэффициенты множественной регрессии показывают, что при увеличении среднегодовой стоимости основных фондов на 1 га сельскохозяйственных угодий на 1 тыс. руб. выручка от реализации продукции на 1 га сельскохозяйственных угодий в среднем увеличивается на 645 руб. (при исключении влияния факторов x_2 и x_3), при росте численности работников 100 га сельскохозяйственных угодий на одного работника выручка в среднем возрастает на 1615 руб./га, при увеличении площади сельскохозяйственных угодий на одно предприятие на 100 га выручка от реализации продукции на 1 га возрастает на27,3 руб. Не все факторы, включенные в уравнение регрессии, могут оказывать статистически значимое влияние на изменение результативного признака. Выдвигается нулевая гипотеза о равенстве нулю множественных коэффициентов регрессии. Она проверяется с применением критерия Стьюдента. В приведенной выше таблице приведены наблюдаемые значения критерия. Критическое значение критерия Стьюдента при уровне значимости $\alpha = 0{,}05$ и числе степеней свободы $к$ = 16 составляет 2,12. Наименьшее фактически наблюдаемое значение критерия (0,934) по фактору x_3 меньше критического значения, поэтому фактор x_3 следует исключить из множественного уравнения регрессии.

Аналогично найдем параметры уравнения, используя инструмент *Пакет анализа – Регрессия.* Результаты расчета приведены ниже в таблице.

Регрессионная статистика						
Множественный R	0,902916					
R-квадрат	0,815258					
Нормированный R-квадрат	0,793523					
Стандартная ошибка	5,588229					
Наблюдения	20					
Дисперсионный анализ						
	df	*SS*	*MS*	*F*	Значимость *F*	
Регрессия	2	2342,752	1171,376	37,5101	5,83E-07	
Остаток	17	530,8812	31,2283			
Итого	19	2873,633				
	Коэффициенты	Стандартная ошибка	*t*-статистика	*P*-значение	Нижние 95%	Верхние 95%
y-пересечение	8,699397	3,393022	2,563908	0,020125	1,5405	15,85805
x_1	0,625794	0,116902	5,353153	5,27E-05	0,3793	0,872436
x_2	1,997537	0,985169	2,027608	0,058573	-0,0819	4,076062

Рисунок 2.14 – Вывод итогов регрессионного анализа

После исключения фактора x_3, изменилась оценка значимости оставшихся параметров уравнения регрессии. При уровне значимости 0,05 существенно отличен от нуля свободный член уравнения регрессии и коэффициент регрессии при факторе x_1 ($t_{05,17}$=2,11). Коэффициент регрессии при фактореx_2 статистически значим при уровне значимости 0,058, что вполне приемлемо, учитывая небольшое число наблюдений и экономическую значимость этого фактора.

Таким образом, множественное линейное уравнение регрессии имеет вид:

$$y = \underset{2,56}{8,699} + \underset{5,35}{0,626x_1} + \underset{2,03}{1,998x_2}$$

При увеличении стоимости основных фондов на 1 га сельскохозяйственных угодий на 1 тыс. руб. выручка от реализации продукции на 1 га сельскохозяйственных угодий в среднем возрастает на 626 руб., при исключении влияния обеспеченности рабочей силой. С ростом численности работников на 100 га сельскохозяйственных угодий на одного работника выручка от реализации продукции на 1 га сельскохозяйственных угодий в среднем растет на 1198 руб., при исключении влияния фондооснащенности производства.

В стандартизованной форме уравнение регрессии имеет вид:

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

Найдем β – коэффициенты, используя их связь с коэффициентами b_jуравнения регрессии в нормальной форме:

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

$$\beta_1 = b_1 \frac{\sigma_{x_1}}{\sigma_y} = 0,626 \cdot \frac{13,97}{12,30} = 0,711;$$

$$\beta_2 = b_2 \frac{\sigma_{x_2}}{\sigma_y} = 1,998 \cdot \frac{1,66}{12,30} = 0,270.$$

β – коэффициенты, можно также найти с помощью парных коэффициентов корреляции по формулам:

$$\beta_1 = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{1 - r_{x_1x_2}^2} = \frac{0.8778 - 0.7098 \cdot 0.6196}{1 - 0.6196^2} = 0,711;$$

$$\beta_2 = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{1 - r_{x_1x_2}^2} = \frac{0.7098 - 0.8778 \cdot 0.6196}{1 - 0.6196^2} = 0,270.$$

Линейное уравнение множественной регрессии в стандартизованном масштабе имеет вид:

$$t_y = 0,711\ t_{x_1} + 0,270\ t_{x_2}.$$

По абсолютной величине β – коэффициентов можно сделать вывод об относительной силе влияния факторов на изменение результативного признака. На выручку от реализации продукции значительно более сильное влияние оказывает обеспеченность основными фондами и значительно меньшее – обеспеченность рабочей силой.

1. Средние коэффициенты эластичности находятся по формуле:

$$Э_{yx_j} = b_j \cdot \frac{\overline{x}_j}{\overline{y}};$$

$$Э_{yx_1} = b_1 \cdot \frac{\overline{x}_1}{\overline{y}} = 0{,}626 \cdot \frac{32{,}68}{35{,}94} = 0{,}569;$$

$$Э_{yx_2} = b_2 \cdot \frac{\overline{x}_2}{\overline{y}} = 1{,}998 \cdot \frac{3{,}40}{35{,}94} = 0{,}189.$$

Значит, при увеличении обеспеченность основными фондами на1% выручка от реализации продукции на 1 га сельскохозяйственных угодий увеличивается в среднем на 0,569 %, исключив влияние второго фактора. Если увеличить численность работников на 100 га сельхозугодий на 1 %, то выручка от реализации в среднем возрастет на 0,189 %, исключив влияние фондооснащенности.

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

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{(1-r_{yx_2}^2)} \cdot (1-r_{x_1x_2}^2)} = \frac{0{,}8778 - 0{,}7098 \cdot 0{,}6196}{\sqrt{(1-0{,}7098^2)(1-0{,}6196^2)}} = 0{,}792;$$

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{(1-r_{yx_1}^2)} \cdot (1-r_{x_1x_2}^2)} = \frac{0{,}7098 - 0{,}8778 \cdot 0{,}6196}{\sqrt{(1-0{,}8778^2)(1-0{,}6196^2)}} = 0{,}441;$$

$$r_{x_1x_2 \cdot y} = \frac{r_{x_1x_2} - r_{yx_1} \cdot r_{yx_2}}{\sqrt{(1-r_{yx_1}^2)} \cdot (1-r_{yx_2}^2)} = \frac{0{,}6196 - 0{,}8778 \cdot 0{,}7098}{\sqrt{(1-0{,}8778^2)(1-0{,}7098^2)}} = -\ 0{,}010.$$

Коэффициенты частной корреляции характеризуют тесноту связи между двумя переменными, исключив влияние третьей переменной. Значит, связь между фондооснащенностью и выручкой от реализации прямая и тесная, между трудообеспеченностью и выручкой от реализации также прямая и довольно слабая. Связь между факторами x_1 и x_2 очень слабая и обратная. Можно считать, что факторы линейно независимые.

Коэффициент множественной корреляции находится по формуле:

$$R_{y_{x_1x_2}} = \sqrt{\beta_1 \cdot r_{yx_1} + \beta_2 \cdot r_{yx_2}} = \sqrt{0{,}711 \cdot 0{,}8778 + 0{,}270 \cdot 0{,}7098} =$$

$$= \sqrt{0{,}6241 + 0{,}1916} = \sqrt{0{,}8157} = 0{,}903;$$

$$R_{y_{x_1x_2}} = \sqrt{\frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2}} = \sqrt{1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \overline{y})^2}} = \sqrt{0{,}81526} = 0{,}903.$$

Величина коэффициента множественной корреляции показывает, что связь между выручкой и обоими факторами очень тесная, причем 81,6 % вариации выручки от реализации продукции объясняется влиянием фондо- и трудообеспеченности, из которой на долю фондообеспеченности приходится 62,4% вариации, а трудообеспеченности – 19,2%. Некоторые расхождения в результатах объясняются округлением значений при промежуточных расчетах.

4. Оценим значимость уравнения регрессии и множественного коэффициента детерминации R^2 с помощью критерия F – Фишера. Выдвигается 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:

$$F_н = \frac{R^2_{yx_1x_2}}{1 - R^2_{yx_1x_2}} : \frac{m}{n - m - 1}, \qquad (2.35)$$

где m – число факторов в линейном уравнении регрессии;
n – число единиц наблюдения.

$$F_н = \frac{0{,}81526}{1 - 0{,}81526} : \frac{2}{20 - 2 - 1} = 37{,}51.$$

При уровне значимости α=0,05 и числе степеней свободы $к_1 = m = 2$, $к_2 = n – m – 1 = 20 – 2 – 1 = 17$, 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 множественный коэффициент детерминации 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 критерияt – Стьюдента.

$$F_{нx_1} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1 - R^2_{yx_1x_2}} \cdot \frac{n - m - 1}{1} = \frac{0{,}81526 - 0{,}7098^2}{1 - 0{,}81526} \cdot \frac{20 - 2 - 1}{1} = 28{,}66.$$

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 послеx_2. Фактор x_1 оказывает статистически значимое влияние на y.

$$F_{нx_2} = \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1 - R^2_{yx_1x_2}} \cdot \frac{n - m - 1}{1} = \frac{0{,}81526 - 0{,}8778^2}{1 - 0{,}81526} \cdot \frac{20 - 2 - 1}{1} = 4{,}12$$

В этом случае наблюдаемое значение критерия Фишера несколько меньше критического, что свидетельствует о статистической не значимости влияния фактора x_2 и не целесообразности включения его в уравнение множествен-

ной регрессии. В данной задаче на выручку от реализации продукции статистически значимое влияние оказывает первый фактор. Но из экономических соображений желательнов множественном уравнении регрессии оставить оба фактора.

Пример 2.2.

Предположим зависимость между переменными, представленными в примере 2.1, выражается степенным уравнением множественной регрессии, которое имеет вид:

$$Y= b_0 \cdot X_1^{b_1} \cdot X_2^{b_2} \cdot X_3^{b_3} \cdot \varepsilon \,. \qquad (2.36)$$

Определить:

а) наблюдаемые значения переменных для расчета параметров степенного уравнения регрессии;

б) парные коэффициенты корреляции между всеми переменными;

в) наличие или отсутствие мультиколлинеарности между признаками;

г) параметры множественного уравнения регрессии в натуральной форме;

д) коэффициенты множественной корреляции и детерминации;

е) значимость множественного уравнения регрессии с помощью общего критерия *F* – Фишера;

ж)значимость множественных коэффициентов регрессии с использованием критерия Стьюдента.

Решение.

Для нахождения параметров степенного уравнения регрессии методом наименьших квадратов необходимо его привести к линейному виду путем логарифмирования уравнения (2.36):

$$lg\, Y = lg\, b_0 + b_1\, lg\, X_1 + b_2\, lg\, X_2 + b_3\, lg\, X_3. \qquad (2.37)$$

С помощью функции *log10* прологарифмируем значения переменных в таблице 2.3, которые представим в таблице 2.7.

Таблица 2.7 – Исходные данные для определения параметров степенного уравнения регрессии

№ п/п	*lg Y*	lgX_1	lgX_2	lgX_3
1	1,442643	1,385556	0,559713	3,636187
2	1,415191	1,256426	0,410407	3,747955
3	1,533354	1,350271	0,287963	3,4524
4	1,478996	1,526718	0,461251	3,599446
5	1,581535	1,466535	0,369278	3,85083
6	1,537693	1,395097	0,517216	3,940666
7	1,530835	1,400748	0,587617	4,109612
8	1,791511	1,688623	0,787683	3,988291
9	1,348307	0,937737	0,153339	3,876045

Продолжение таблицы 2.7

№ п/п	*lg Y*	lgX_1	lgX_2	lgX_3
10	1,505765	1,441694	0,392195	3,872622
11	1,581548	1,533815	0,49799	4,359342
12	1,694498	1,719808	0,637989	3,884455
13	1,653785	1,693766	0,54154	3,631647
14	1,705998	1,861784	0,463381	3,761928
15	1,494726	1,395438	0,704441	4,147429
16	1,371126	1,319875	0,631045	3,601951
17	1,339757	1,499847	0,202749	3,43072
18	1,798858	1,812915	0,919786	4,039255
19	1,356158	1,402329	0,367246	3,897572
20	1,507300	1,557637	0,281568	3,575765

Парные коэффициенты корреляции найдем с помощью инструмента пакета анализа «*Корреляция*», для этого заполним параметры диалогового окна как на рисунке 2.14.

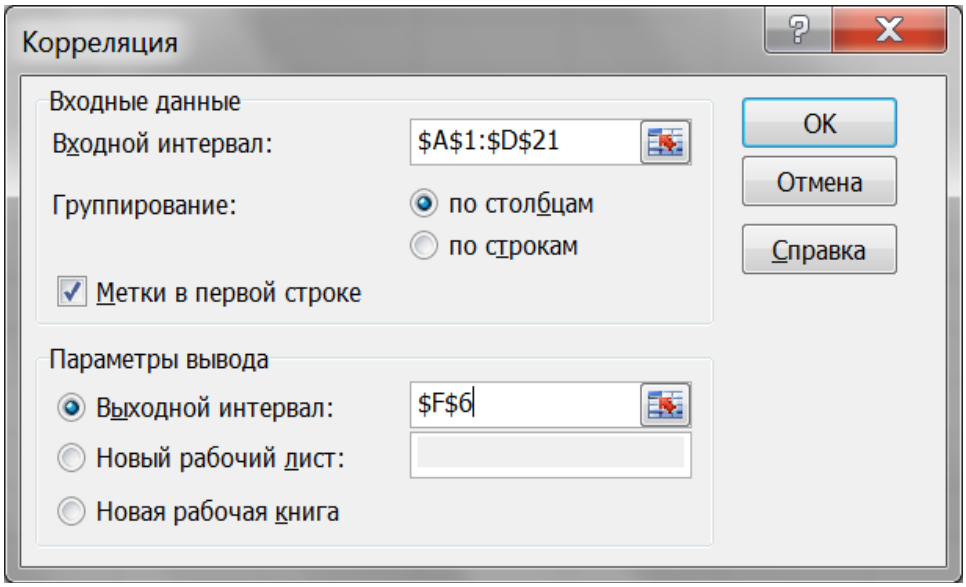


Рисунок 2.14 – Диалоговое окно «Корреляция»

В результате будет получена матрица парных коэффициентов корреляции между всеми изучаемыми переменными (таблица 2.8).

Таблица 2.8 – Парные коэффициенты корреляции между признаками

	lgY	lgX_1	lgX_2	lgX_3
lg Y	1			
lgX_1	0,799931	1		
lgX_2	0,650133	0,523639	1	
lgX_3	0,347594	0,072018	0,47438	1

Парные коэффициенты корреляции показывают, что наблюдается сильная связь между выручкой от реализации продукции на 1 га сельскохозяйственных угодий и фондооснащенностью, довольно тесная с трудообеспеченностью и слабая – с размерами землепользования. Между факторами значения коэффициентов корреляции меньше 0,6, что свидетельствует об отсутствии мультиколлинеарности факторов. Это подтверждается также значением определителя матрицы межфакторной корреляции, составившего 0,762. Сопоставление парных коэффициентов корреляции при линейной и степенной формам связи, показывает, что связь между выручкой от реализации продукции и факторами ближе к линейной форме, чем степенной.

Найдем параметры степенного уравнения, используя инструмент *Пакета анализа – Регрессия.*

Регрессионная статистика		
Множественный R	0,862936	
R-квадрат	0,744658	
Нормированный R-квадрат	0,696782	
Стандартная ошибка	0,076801	
Наблюдения	20	

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	3	0,275225	0,091742	15,55	5,3E-05
Остаток	16	0,094374	0,005898		
Итого	19	0,369599			

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхнее 95%
y-пересечение	0,338139	0,35325	0,957225	0,352695	-0,41072	1,08700
$lg\ x_1$	0,453576	0,101292	4,477921	0,000381	0,238847	0,66830
$lg\ x_2$	0,140829	0,125056	1,126129	0,276719	-0,12428	0,40594
$lg\ x_3$	0,118885	0,085185	1,395612	0,1819	-0,0617	0,29947

Рисунок 2.15 – Вывод итогов регрессионного анализа

Множественное уравнение регрессии имеет вид:

$$lg\ Y = lg\ 0{,}338139 + 0{,}453576\ lg\ X_1 + 0{,}140829\ lg\ X_2 + 0{,}118885\ lg\ X_3,$$

а в естественной форме*:* $y = 2{,}1784 \cdot X_1^{0{,}4536} \cdot X_2^{0{,}1408} \cdot X_3^{0{,}1189}$.

Коэффициенты регрессии в степенном уравнении являются коэффициентами эластичности, которые показывают, что при увеличении фондо- и трудо-

обеспеченности сельскохозяйственных предприятий на 1% выручка от реализации продукции на 1 га сельскохозяйственных угодий в среднем возрастает на 0,453 и 0,141% соответственно. При увеличении площади сельскохозяйственных угодий на одно предприятие на 1% выручка от реализации увеличивается на 0,119%, при исключении влияния других факторов. Сумма коэффициентов эластичности меньше единицы, что говорит о снижающейся эффективности факторов производства (0,4536 + 0,1408 + 0,1189 = 0,7133). Уравнение регрессии объясняет 74,5% различий в выручке от реализации продукции между предприятиями влиянием отобранных трех факторов. Так как наблюдаемое значение *F*-критерия больше критического значения ($F_{05,3,16} = 3,24$), то в целом уравнение регрессии является статистически значимым.

Проверка значимости множественных коэффициентов регрессии проводится с помощью *t*-статистики Стьюдента. Критическое значение критерия при уровне значимости 0,05 и 16 степенях свободы составляет 2,12. Сравнение фактически наблюдаемых значений критерия с критическим показывает, что влияние логарифмов факторов x_2 и x_3 оказалось статистически не значимым (1,126 и 1,396 меньше 2,12). Наименьшее из этих значений соответствует логарифму фактора x_2, поэтому исключим этот фактор из множественного уравнения регрессии. Найдем параметры нового степенного уравнения, используя инструмент*Пакет анализа – Регрессия.*

Регрессионная статистика	
Множественный R	0,851129
R-квадрат	0,72442
Нормированный R-квадрат	0,691999
Стандартная ошибка	0,077404
Наблюдения	20

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	0,26774	0,133873	22,34401	1,75E-05
Остаток	17	0,10185	0,005991		
Итого	19	0,36959			

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%
Y-пересечение	0,124381	0,300258	0,414246	0,683874	-0,50911	0,757869
lg X_1	0,517164	0,084753	6,102029	1,17E-05	0,338351	0,695977
lg X_3	0,168182	0,073650	2,283520	0,035536	0,012793	0,323571

Рисунок 2.16 – Вывод итогов регрессионного анализа

Исключение логарифма фактора x_2 привело к незначительному снижению множественных коэффициентов корреляции и детерминации. Уравнение с двумя факторами объясняет 72,4% вариации выручки от реализации продукции на 1 га сельскохозяйственных угодий. Уравнение регрессии статистически значимо и имеет следующий вид:

$$lg\,Y = lg\,0{,}12438 + 0{,}51716\,lg\,X_1 + 0{,}16818\,lg\,X_3$$

или

$$Y = 1{,}3316 \cdot X_1^{0.51716} \cdot X_3^{0.16818}.$$

Видно, что при увеличении фондооснащенности на 1% выручка от реализации продукции на 1 га сельскохозяйственных угодий в среднем растет на 0,517%, исключив влияние размера землепользования. При увеличении площади сельскохозяйственных угодий на 1% выручка от реализации растет в среднем на 0,168%, исключив влияние фондооснащенности. При уровне значимости 0,05 и 17 степенях свободы критическое значение *t*-критерия Стьюдента составляет 2,11. Так как наблюдаемые значения критерия по обоим факторам больше критического значения, то коэффициенты регрессии статистически значимо отличаются от нуля, поэтому фондооснащенность и площадь сельскохозяйственных угодий оказывают статистически существенное влияние на величину выручки от реализации продукции на 1 га сельскохозяйственных угодий.

Рассмотрим решение примера 2.1 с использованием пакета *STATISTICA10*. Исходное меню множественной регрессии представлено на рисунке 2.17.

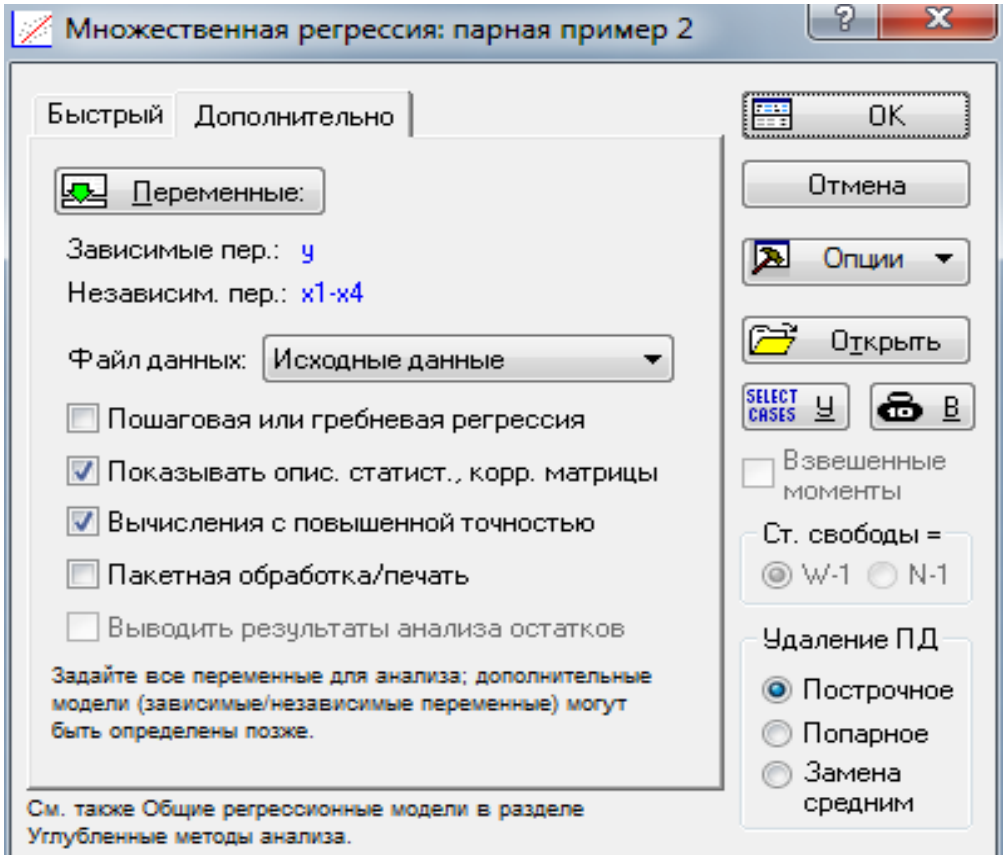


Рисунок 2.17– Главное меню *Множественной регрессии*

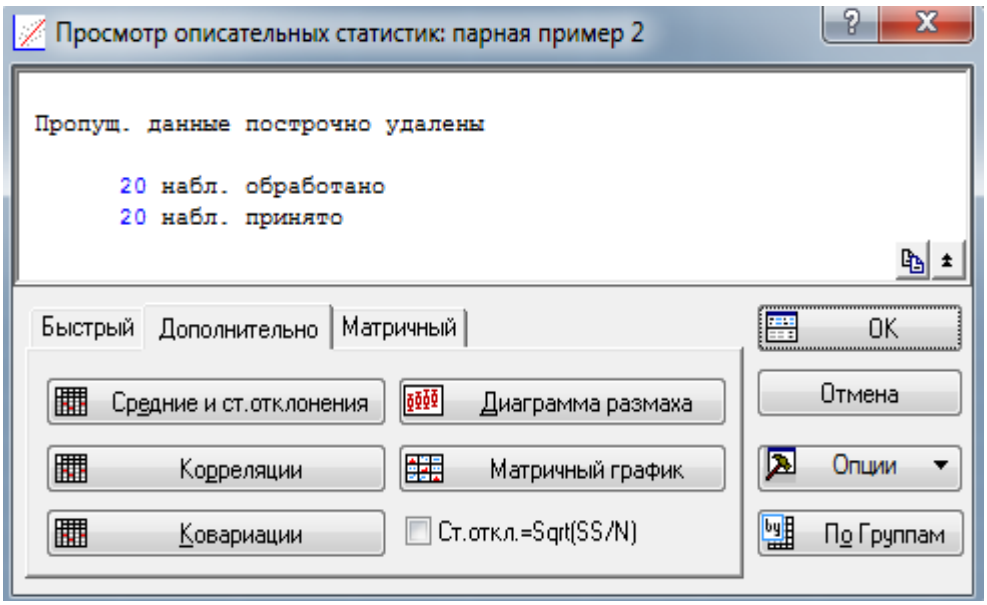


Рисунок 2.18 – Множественная регрессия – описательные статистики

После выбора *Ок* в основном меню и вкладки *Дополнительно* можно получить основные описательные статистики для выбранных переменных:

Таблица 2.9 - Средние и стандартные отклонения

	Средние	Стандартные отклонения	№
X1	32,684	13,972	20
X2	3,396	1,658	20
X3	7703,850	4792,365	20
X4	2,500	1,000	20
Y	35,937	12,298	20

Таблица 2.10 – Корреляционная матрица

	X1	*X2*	*X3*	*X4*	*Y*
X1	1,000000	0,619615	0,092743	0,722917	0,877827
X2	0,619615	1,000000	0,356418	0,761414	0,709820
X3	0,092743	0,356418	1,000000	0,240665	0,251800
X4	0,722917	0,761414	0,240665	1,000000	0,697439
Y	0,877827	0,709820	0,251800	0,697439	1,000000

Переменная X_4достаточно сильно коррелирует с переменными X_1и X_2,что говорит о наличии мультиколлинеарности, переменную X_4следует удалить из рассмотрения в качестве факторной.

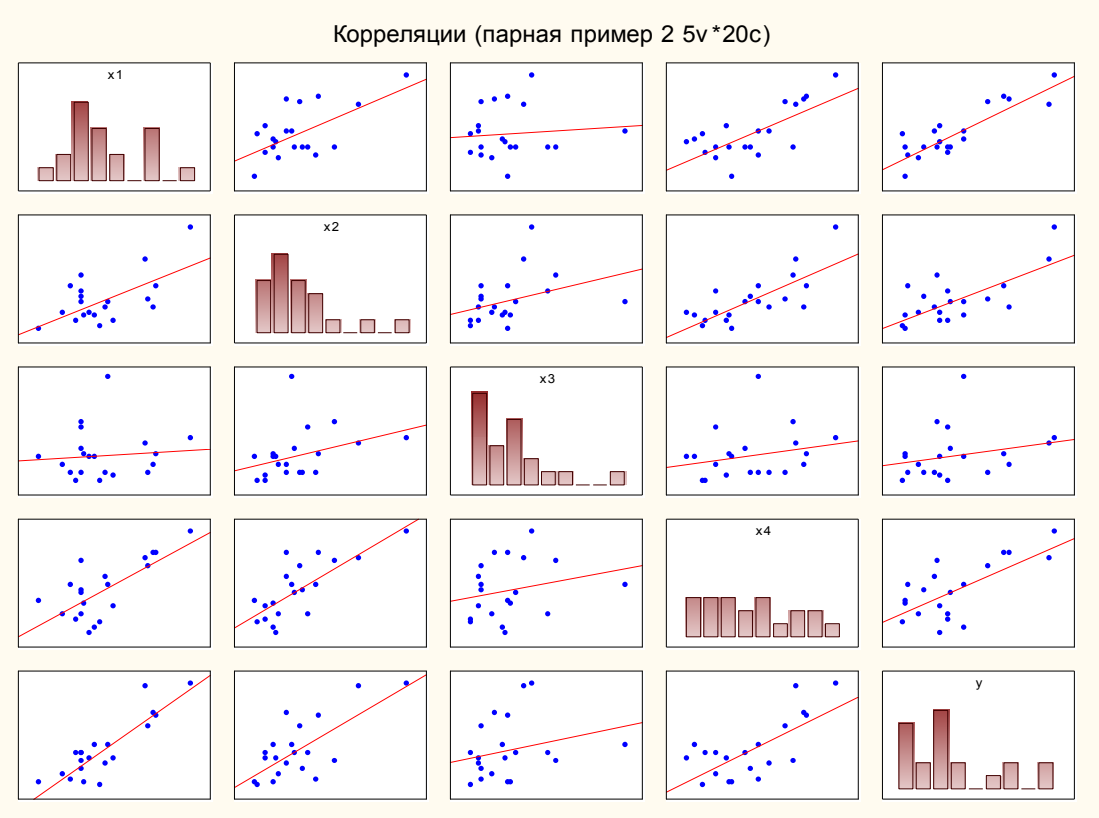


Рисунок 2.19 – Матрица корреляционных полей
между парами переменными

На рисунке наглядно показана попарная связь между изучаемыми переменными. По характеру расположения точек можно судить о направлении и в определенной мере о тесноте связи между переменными.

Таблица 2.11 – Итоги регрессии

R= ,90818688 R^2= ,82480341 Скоррект. R^2= ,79195405 F(3,16)=25,109 p

	БЕТА	Ст.ош. - БЕТА	B	Ст.ош. - B	t(16)	p-знач.
Св. член			7,262883	3,737278	1,943362	0,069778
X1	0,733051	0,135399	0,645231	0,119178	5,414026	0,000057
X2	0,217759	0,144291	1,615270	1,070309	1,509162	0,150751
X3	0,106201	0,113745	0,000273	0,000292	0,933678	0,364347

Итоги регрессии (табл. 2.11) демонстрируют, что фактор X_3не является статистически значимым, после отбрасывания переменной X_3мы получим зависимость $Y=8,699394+0,625794X_1+1,997538X_2$(таблица 2.12).

Таблица 2.12 – Итоги регрессии

R= ,90291631 R^2= ,81525787 Скоррект. R^2= ,79352350 F(2,17)=37,510 p

	БЕТА	Ст.ош. - БЕТА	B	Ст.ош. - B	*t*(17)	*p*-знач.
Св.член			8,699394	3,393022	2,563907	0,020125
X1	0,710969	0,132813	0,625794	0,116902	5,353154	0,000053
X2	0,269293	0,132813	1,997538	0,985169	2,027609	0,058573

Таблица 2.13 - Дисперсионный анализ

Дисперсионный анализ; ЗП: y (парная пример 2)

	Сумма квад-ратов	Степени свободы	Среднее квадратическое отклонение	F	p-знач.
Регрессия	2342,752	2	1171,376	37,51008	0,000001
Остатки	530,881	17	31,228		
Итого	2873,633				

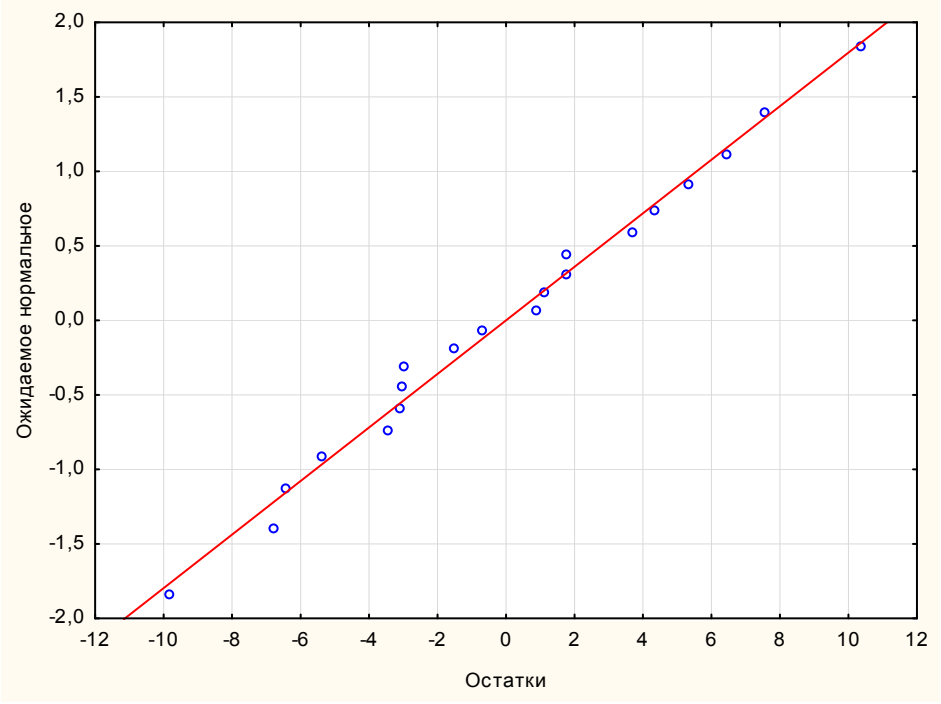


Рисунок 2.20 – Нормальный вероятностный график остатков

Изучение остатков (рисунок 2.20) по графику «вероятностной бумаги» показывает выполнение основных предположений использования *МНК* – нормальность распределения остатков и отсутствие выбросов.

Задача 2.1.

По данным приложения Б требуется определить:

1. Обобщающие статистические характеристики по каждой переменной;
2. Парные коэффициенты корреляции между всеми переменными;
3. Наличие или отсутствие мультиколлинеарности между признаками.

Задача 2.2.

Используя данные приложения Бпо одному варианту определить:

1. Параметры линейного множественного уравнения регрессии в натуральной и стандартизованной форме;
2. Средние коэффициенты эластичности и $\beta -$ коэффициенты для каждого фактора;
3. Коэффициенты частной и множественной корреляции;
4. Значимость множественного уравнения регрессии в целом с помощью общего критерия *F* – Фишера;
5. Значимость множественных коэффициентов регрессии с использованием критериев Фищера и Стьюдента;
6. Доверительные интервалы множественных коэффициентов регрессии при уровне доверительной вероятности 0,95;

Дать оценку полученным результатам, которые оформить в виде кратких выводов.

По всем вариантам в качестве зависимой переменной (*y*) взять выручку от реализации продукции на 1 га пашни, тыс. руб.

Варианты заданий по данным приложения Б

Вариант	1	2	3	4
Переменные	$x_1;x_2;x_6;x_7$	$x_1;x_2;x_5;x_7$	$x_1;x_2;x_4;x_7$	$x_1;x_2;x_4;x_6;x_7$
Вариант	5	6	7	8
Переменные	$x_2;x_3;x_5;x_6$	$x_2;x_3;x_5;x_7$	$x_2;x_3;x_4;x_6;x_7$	$x_3;x_4;x_6;x_7$

Задача 2.3.

По данным, представленным в приложении В, изучить влияние совокупности факторов на производственную и коммерческую себестоимость производства молока (y_1 и y_2) в сельскохозяйственных предприятиях северной зоны Краснодарского края за 2011 г.

1. С помощью инструмента анализа данных *Описательная статистика* рассчитать обобщающие характеристики вариационных рядов, написав выводы по каждой переменной.

2. С помощью инструмента анализа данных *Корреляция* построить матрицу парных коэффициентов корреляции. Оценить наличие или отсутствие мультиколлинеарности между факторами.

3. С помощью инструмента анализа данных *Регрессия*рассчитать параметры множественного уравнения регрессии с включением всех факторов.

4. Рассчитать частные коэффициенты корреляции, эластичности и стандартизованные коэффициенты регрессии.

5. Оценить значимость множественного уравнения регрессии и множественных коэффициентов регрессии с помощью критериев Фишера и Стьюдента.

6. Определить среднюю ошибку аппроксимации.

7. Провести последовательный отсев статистически не значимых факторов и получить модель себестоимости производства молока со всеми статистически значимо влияющими факторами. Рассчитать частные коэффициенты корреляции, эластичности и стандартизованные коэффициенты регрессии. Оценить значимость множественного уравнения регрессии и множественных коэффициентов регрессии с помощью критериев Фишера и Стьюдента. Определить среднюю ошибку аппроксимации.

8. Написать выводы по результатам множественного регрессионного анализа себестоимости производства молока.

Задача 2.4.

По данным 41 сельскохозяйственного предприятия северной зоны Краснодарского края изучается зависимость продуктивности коров от влияющих на нее факторов.

Зависимая переменная (y) – годовой надой молока на среднегодовую корову, кг.

Объясняющие переменные:

x_1 – производственные затраты на среднегодовую корову, тыс. руб.;

x_2 – затраты на корма на среднегодовую корову, тыс. руб.;

x_3– прямые затраты труда на среднегодовую корову, чел.-ч;

x_4– среднегодовое поголовье коров на предприятие, гол.;

x_5 – затраты по оплате труда на 1 чел.-ч, руб.;

x_6 – доля молока в выручке от реализации продукции животноводства, %.

В таблицах 2.14 и 2.15 приведены статистические характеристики по изучаемой совокупности предприятий и парные коэффициенты корреляции между переменными.

Таблица 2.14 – Статистические характеристики выборочной совокупности сельскохозяйственных предприятий

Показатель	y	x_1	x_2	x_3	x_4	x_5	x_6
Среднее значение	5383,5	70,04	31,74	130,58	961,34	135,34	70,68
Стандартная ошибка	228,24	2,83	1,51	6,33	100,29	9,23	2,61

Продолжение таблицы 2.14

Показатель	y	x_1	x_2	x_3	x_4	x_5	x_6
Медиана	5178,91	68,85	31,61	127,69	736,00	124,38	74,18
Среднее квадратическое отклонение	1461,47	18,15	9,66	40,54	642,16	59,09	16,73
Дисперсия выборки	2135883	329,27	93,29	1643,71	412372,98	3491,44	279,95
Эксцесс	0,480	3,700	-0,146	-0,432	1,753	1,067	0,160
Асимметричность	0,739	1,134	0,316	0,452	1,486	0,973	-0,867
Интервал	6936,9	98,9	43,7	154,5	2763,0	267,7	66,7
Минимум	2718,30	39,27	14,07	60	229	39,217	28,446
Максимум	9655,23	138,15	57,74	214,5	2992	306,921	95,167
Сумма	220722	2872	1302	5354	39415	5549	2898

Таблица 2.15 – Парные коэффициенты корреляции между переменными

	y	x_1	x_2	x_3	x_4	x_5	x_6
y	1						
x_1	0,8334	1					
x_2	0,6471	0,8165	1				
x_3	-0,1378	-0,1519	-0,0285	1			
x_4	0,2858	0,2166	0,1392	0,0212	1		
x_5	0,3106	0,3061	0,0705	-0,4933	0,0920	1	
x_6	0,2418	0,2456	0,3239	-0,1536	-0,3065	-0,0968	1

Варианты заданий к задаче 2.4

Вариант	1	2	3	4	5	6
Переменные	$y; x_1; x_4$	$y; x_1; x_5$	$y; x_1; x_6$	$y; x_2; x_3$	$y; x_2; x_4$	$y; x_2; x_5$
Вариант	7	8	9	10	11	12
Переменные	$y; x_2; x_6$	$y; x_4; x_5$	$y; x_4; x_6$	$y; x_5; x_6$	$y; x_3; x_4$	$y; x_3; x_5$

1. По одному варианту составить матрицу парных коэффициентов корреляции между тремя переменными.

2. Определить параметры множественного уравнения регрессии в стандартизированной и естественной форме.

3. Рассчитать частные коэффициенты эластичности.

4. Рассчитать частные и множественный коэффициенты корреляции и детерминации.

5. Оценить значимость множественного уравнения регрессии с помощью F-критерия Фишера, для чего составить таблицу дисперсионного анализа.

6. С помощью частных F-критериев Фишера оценить целесообразность включения фактора x_1 после x_2 и фактора x_2 после x_1.

7. Оценить значимость множественных коэффициентов регрессии с помощью *t*-критерия Стьюдента.

8. Написать выводы по результатам расчетов.

Задача 2.5.

По данным, приведенным в приложении В, изучить влияние факторов x_2, x_3, x_4, x_5, x_6 на годовой удой на среднегодовую корову (y).

1. Определить параметры множественного уравнения регрессии с включением в линейную модель всех факторов.

2. Оценить значимость множественных коэффициентов регрессии с помощью *t*-критерия Стьюдента.

3. Провести последовательный отсев статистически не значимых факторов.

4. По уравнению со всеми значимо влияющими факторами рассчитать частные и множественный коэффициенты корреляции и детерминации.

5. Оценить значимость множественного уравнения регрессии с помощью *F*-критерия Фишера, для чего составить таблицу дисперсионного анализа.

6. Написать выводы по результатам расчетов.

3. Оценивание систем одновременных уравнений

В условиях быстро изменяющейся техногенной, политической, экологической, экономической и т.д. обстановки на первое место выходит системный подход к изучению окружающего мира.

Любая реальная система требует решения целого комплекса задач, часто взаимоисключающих друг друга и изменение одного факторного признака не может происходить при абсолютной неизменности других. Следовательно, отдельно взятое уравнение множественной регрессии не в состоянии охарактеризовать истинные влияния отдельных признаков на вариацию результатирующей переменной. Поэтому важное место занимает проблема описания структуры связей между переменными системы так называемых одновременных уравнений или называемых структурными уравнениями, которые чаще всего рассматриваются на уровне макроэкономических исследований.

Система уравнений в эконометрических исследованиях может быть построена по разному.

1. *Система независимых уравнений*. В данной системе каждая зависимая переменная (*y)* рассматривается как функция одного и того же набора факторов (*x*):

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1m}x_m + \varepsilon_1, \\ y_2 = a_{21}x_1 + a_{22}x_2 + ... + a_{2m}x_m + \varepsilon_2, \\ ..... \\ y_n = a_{n1}x_1 + a_{n2}x_2 + ... + a_{nm}x_m + \varepsilon_n. \end{cases} \qquad (3.1)$$

Набор факторов x_i в каждом уравнении может варьироваться. Отсутствие того или иного фактора в уравнении системы может быть следствием как экономической целесообразности его включения в модель, так и несущественности его влияния на результативный признак (незначимо значение *t*-критерия или частного *F*-критерия для данного фактора). Примером такой модели может служить *модель экономической эффективности сельскохозяйственного производства*, где в качестве зависимых переменных выступают показатели, характеризующие эффективность сельскохозяйственного производства – продуктивность коров, себестоимость 1 ц молока, а в качестве факторов – специализация хозяйства, количество голов на 100 га пашни, затраты труда и т.д.

2. С*истемы рекурсивных уравнений*. В такой системе зависимая переменная *(y)* одного уравнения выступает в виде фактора *(x)* в другом уравнении:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1m}x_m + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + ... + a_{2m}x_m + \varepsilon_2, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2 + ... + a_{3m}x_m + \varepsilon_3, \\ ..... \\ y_n = b_{n1}y_1 + b_{n2}y_2 + ... + b_{n(n-1)}y_{n-1} + a_{n1}x_1 + a_{n2}x_2 + ... + a_{nm}x_m + \varepsilon_n. \end{cases} \qquad (3.2)$$

Примером такой системы может служить *модель производительности труда и фондоотдачи вида*:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \varepsilon_2, \end{cases} \quad (3.3)$$

где y_1 – производительность труда;

y_2 – фондоотдача;

x_1 – фондовооруженность труда;

x_2 – энерговооруженность труда;

x_3 – квалификация рабочих.

В этих двух видах систем уравнений каждое уравнение может рассматриваться самостоятельно и его параметры определяются методом наименьших квадратов.

3. *Система взаимозависимых уравнений,* получившая наибольшее распространение в эконометрических исследованиях. В ней одни и те же зависимые переменные в одних уравнениях входят в левую часть, а в других уравнениях – в правую часть системы:

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + ... + b_{1n}y_n + a_{11}x_1 + a_{12}x_2 + ... + a_{1m}x_m + \varepsilon_1, \\ y_2 = b_{21}y_1 + b_{23}y_3 + ... + b_{2n}y_n + a_{21}x_1 + a_{22}x_2 + ... + a_{2m}x_m + \varepsilon_2, \\ ..... \\ y_n = b_{n1}y_1 + b_{n2}y_2 + ... + b_{n(n-1)}y_{n-1} + a_{n1}x_1 + a_{n2}x_2 + ... + a_{nm}x_m + \varepsilon_n. \end{cases} \quad (3.4)$$

Систему взаимозависимых уравнений называют *системой совместных, одновременных уравнений*. В эконометрике эта система называется *структурной формой модели*. В системе одновременных уравнений одни и те же переменные (y) одновременно рассматриваются как зависимые в одних уравнениях и как независимые в других. В отличие от предыдущих систем каждое уравнение не может рассматриваться самостоятельно и МНК для нахождения параметров неприменим. Для этого используются специальные приемы оценивания.

Примером системы одновременных уравнений может служить модель динамики цены и заработной платы:

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3 + \varepsilon_2, \end{cases} \quad (3.5)$$

где y_1 – темп изменения месячной заработной платы,

y_2 – темп изменения цен,

x_1 – процент безработных,

x_2 – темп изменения постоянного капитала,

x_3 – темп изменения цен на импорт сырья.

Система совместных, одновременных уравнений (или структурная форма модели) обычно содержит эндогенные и экзогенные переменные.

Эндогенные переменные (y) - это зависимые переменные, число которых равно числу уравнений в системе.

Экзогенные переменные (x) - это предопределенные переменные, влияющие на эндогенные переменные, влияющие на эндогенные переменные, но не зависящие от них.

Простейшая *структурная форма модели* имеет вид:

$$\begin{cases} y_1 = b_{12} y_2 + a_{11} x_1 + \varepsilon_1, \\ y_2 = b_{21} y_1 + a_{22} x_2 + a_{23} x_3 + \varepsilon_2, \end{cases} \qquad (3.6)$$

где y – эндогенные переменные,

x – экзогенные переменные.

Классификация переменные на эндогенные и экзогенные зависит от теоретической концепции модели. Экономические переменные могут выступать в одних моделях как эндогенные, а в других как экзогенные переменные. Внеэкономические переменные (например, климатические условия) входят в систему как экзогенные переменные.

Структурная форма модели позволяет увидеть влияние изменений любой экзогенной переменной на значение эндогенной переменной. Целесообразно в качестве экзогенных переменных выбирать такие переменные, которые могут быть объектом регулирования, меняя и управляя которыми можно иметь целевые значения эндогенных переменных.

Коэффициенты b_i - при эндогенной переменной и a_i - при экзогенной переменной называются *структурными коэффициентами модели.* Для определения структурных коэффициентов модели структурная форма модели преобразуется в *приведенную форму модели*.

Приведенная форма модели – система линейных функций эндогенных переменных от экзогенных:

$$\begin{cases} \hat{y}_1 = \delta_{11} x_1 + \delta_{12} x_2 + ... + \delta_{1m} x_m + \varepsilon_1, \\ \hat{y}_2 = \delta_{21} x_1 + \delta_{22} x_2 + ... + \delta_{2m} x_m + \varepsilon_2, \\ ..... \\ \hat{y}_n = \delta_{n1} x_1 + \delta_{n2} x_2 + ... + \delta_{nm} x_m + \varepsilon_n. \end{cases} \qquad (3.7)$$

где δ_i – коэффициенты приведенной формы модели

Для структурной модели вида: $\begin{cases} y_1 = b_{12} y_2 + a_{11} x_1, \\ y_2 = b_{21} y_1 + a_{22} x_2, \end{cases}$

приведенная форма модели имеет вид $\begin{cases} y_1 = \delta_{11} x_1 + \delta_{12} x_2, \\ y_2 = \delta_{21} x_1 + \delta_{22} x_2, \end{cases}$

где
$$\delta_{11} = \frac{a_{11}}{1 - b_{12} b_{21}}, \quad \delta_{12} = \frac{a_{22} b_{12}}{1 - b_{12} b_{21}}$$
$$\delta_{21} = \frac{a_{11} b_{21}}{1 - b_{12} b_{21}}, \quad \delta_{22} = \frac{a_{22}}{1 - b_{12} b_{21}}$$

По своему виду приведенная форма модели не отличается от системы независимых уравнений, параметры которой оцениваются традиционным МНК. Применяя МНК, можно оценить δ, а затем оценить значения эндогенных переменных через экзогенные. Приведенная форма модели аналитически уступает структурной форме модели, т.к. в ней отсутствуют оценки взаимодействия эндогенных переменных.

При переходе от приведенной формы модели к структурной возникает проблема идентификации.

Идентификация – это единственность соответствия между приведенной и структурной формами модели.

С точки зрения идентифицируемости структурные модели подразделяют на три вида:

- идентифицируемые;
- неидентифицируемые;
- сверхидентифицируемые.

Модель идентифицируема, если все структурные коэффициенты определяются однозначно, единственным образом по коэффициентам приведенной формы модели, т.е. число параметров структурной модели равно числу параметров приведенной формы. Тогда структурные коэффициенты модели оцениваются через параметры приведенной формы.

Модель неидентифицируема, если число приведенных коэффициентов меньше числа структурных коэффициентов. Тогда структурные коэффициенты не могут быть оценены через коэффициенты приведенной формы.

Модель сверхидентифицируема, если число приведенных коэффициентов больше числа структурных коэффициентов. В этом случае на основе коэффициентов приведенной формы можно получить два или более значений одного структурного коэффициента.

Структурная модель всегда представляет собой систему совместных уравнений, каждое из которых требуется проверять на идентификацию. Модель считается идентифицируемой, если каждое уравнение системы идентифицируемо. Чтобы уравнение было идентифицируемо, необходимо, чтобы число предопределенных (экзогенных) переменных *(x)*, отсутствующих в данном уравнении, но присутствующих в системе, было равно числу эндогенных переменных *(y)* без одного в данном уравнении.

Обозначим число эндогенных переменных в *j*-ом уравнении системы через *H*, а число экзогенных (предопределенных) переменных, содержащихся в системе, но не входящих в данное уравнение, - через *D*, то условие идентифицируемости модели можно записать в виде счетного правила:

D+1=H - уравнение идентифицируемо;
D+1<H - уравнение неидентифицируемо;
D+1>H - уравнение сверхидентифицируемо.

Данное счетное правило есть необходимое, но недостаточное условие идентификации. Более точно условия идентификации определяются, если накладывать ограничения на коэффициенты матриц параметров структурной модели. Уравнение идентифицируемо, если из коэффициентов отсутствующих в нем переменным (эндогенным и экзогенным), можно в других уравнениях системы получить матрицу, определитель которой не равен нулю, а ранг матрицы не меньше числа эндогенных переменных в системе без одного. Целесообразность этой проверки в том, возможна ситуация, когда для каждого уравнения системы счетное правило выполняется, а определитель матрицы равен нулю. В этом случае соблюдается лишь необходимое, но не достаточное условие идентификации.

Пример 3.1.

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2. \end{cases}$$

Проверим каждое уравнение:

Для первого уравнения H=3 (y_1,y_2,y_3) и D= 2 (x_3 и x_4 отсутствуют) , т.е. $D+1=2+1=3=H$ счетное правило – необходимое условие выполнено. Для проверки достаточного условия составим таблицу:

Уравнения	Переменные	
	X_3	X_4
2	a_{23}	a_{24}
3	0	0

Определитель равен нулю, следовательно, достаточное уравнение не выполнено

Для второго уравнения $H=2$ (y_1,y_2), $D=1$ (отсутствует x_1).
$D+1=1+1=2=H$. Необходимое условие выполнено.

Уравнения	Переменные	
	y_3	X_1
1	b_{13}	A_{11}
3	-1	A_{31}

Определитель матрицы не равен нулю, а ранг равен 2 , т.е. равен числу эндогенных переменных (*y*) без одного, т.е. достаточное условие выполнено.

Для третьего уравнения H=3 и D=2, т.е.). D+1=2+1=3=H. Необходимое условие выполнено.

Уравнения	Переменные	
	X_3	X_4
1	*0*	*0*
2	$0a_{23}$	A_{24}

Определитель равен нулю → Достаточное условие не выполняется.

Следовательно, рассматриваемая в целом структурная модель, идентифицируемая по счетному правилу, не может считаться идентифицируемой по достаточному условия идентификации.

Коэффициенты структурной модели, в зависимости от вида системы, могут оцениваться разными методами. Наибольшее распространение получили:

– Косвенный метод наименьших квадратов *(КМНК)*;

– Двухшаговый метод наименьших квадратов *(ДМНК)*;

– Трехшаговый метод наименьших квадратов *(ТМНК)*;

– Метод максимального правдоподобия с полной информацией *($ММП_f$)*;

– Метод максимального правдоподобия с ограниченной информацией *($ММП_s$)*.

При использовании $ММП_f$ для нахождения статистических оценок неизвестных параметров распределения выбирают те, при которых данные результаты наблюдений «наиболее вероятны». При нормальном распределении признаков результаты $ММП_f$ совпадают с *МНК*.

Если число уравнений системы достаточно велико, что приводит к сложным вычислительным процедурам, то применяется $ММП_s$. В данном методе с параметров, связанных с функционированием системы в целом, снимаются ограничения, что приводит к упрощению решения задачи, но трудоемкость процесса остается высокой.

Более просты в использовании *КМНК* и *ДМНК*.

КМНК применяется для идентифицируемой системы одновременных уравнений. Данный метод предполагает выполнение следующих этапов:

– Структурная модель преобразовывается в приведенную форму;

– Для каждого уравнения приведенной формы модели обычным *МНК* оценивают приведенные коэффициенты δ_{ij};

– Полученные коэффициенты транспортируются в параметры структурной модели.

Если изучаемая система сверхидентифицируема, то применяется *ДМНК*, получивших свое название из-за применения МНК дважды. На первом шаге он применяется при определении приведенной формы модели и нахождении на ее основе оценок теоретических значений эндогенной переменной. На втором ша-

ге *МНК* применяется к структурному сверхидентифицируемому уравнению пр определении структурных коэффициентов модели по полученным на первом шаге значениям эндогенных переменных.

Если все уравнения системы сверхидентифицируемые, то *ДМНК* используется для оценки структурных коэффициентов каждого уравнения. Если в системе есть точно идентифицируемые уравнения, то структурные коэффициенты по ним находятся из системы приведенных уравнений.

ТМНК является дальнейшим развитием *ДМНК*.

Задание к задачам 3.1 – 3.19

1. Применив необходимое и достаточное условие идентификации, определите, идентифицировано ли каждое из уравнений модели.
2. Определите метод оценки параметров модели.
3. Запишите приведенную форму модели.

Задача 3.1

Модель денежного рынка:

$R_t = a_1 + b_{11} M_t + b_{12} \cdot Y_t + \varepsilon_1;$

$Y_t = a_2 + b_{21} \cdot R_t + b_{22} \cdot I_t + \varepsilon_1;$

где R – процентная ставка;

Y - ВВП;

M - денежная масса;

I - внутренние инвестиции;

t - текущий период.

Задача 3.2

Модель Менгеса:

$Y_t = a_1 + b_{11} \cdot Y_{t-1} + b_{12} \cdot I_t + \varepsilon_1;$

$I_t = a_2 + b_{21} \cdot Y_t + b_{22} \cdot Q_t + \varepsilon_2;$

$C_t = a_3 + b_{31} \cdot Y_t + b_{32} \cdot C_{t-1} + b_{33} \cdot P_t + \varepsilon_3;$

$Q_t = a_4 + b_{41} \cdot Q_{t-1} + b_{42} \cdot R_t + \varepsilon_4,$

где Y - национальный доход;

C - расходы на личное потребление;

I- чистые инвестиции;

Q - валовая прибыль экономики;

P- индекс стоимости жизни;

R - объем продукции промышленности;

t - текущий период;

$t-1$ - предыдущий период.

Задача 3.3

Одна из версий модифицированной модели Кейнса имеет вид

$C_t = a_1 + b_{11} \cdot Y_t + b_{12} \cdot Y_{t-1} + \varepsilon_1;$

$I_t = a_2 + b_{21} \cdot Y_t + b_{22} \cdot Y_{t-x} + \varepsilon_2$;

$Y_t = C_t + I_t + G_t$,

где C - расходы на потребление;

Y - доход;

I - инвестиции;

G - государственные расходы;

t - текущий период;

$t-1$ - предыдущий период.

Задача 3.4

Модель мультипликатора-акселератора:

$C_t = a_1 + b_{11} \cdot R_t + b_{12} \cdot C_{t-1} + \varepsilon_1$;

$I_t = a_2 + b_{21} \cdot (R_t - R_{t-1}) + \varepsilon_2$;

$R_t = C_t + I_t$,

где C - расходы на потребление;

R - доход;

I - инвестиции;

t - текущий период;

$t-1$ - предыдущий период.

Задача 3.5

Конъюнктурная модель имеет вид

$C_t = a_1 + b_{11} \cdot Y_t + b_{12} \cdot C_{t-1} + \varepsilon_1$;

$I_t = a_2 + b_{21} \cdot r_t + b_{22} \cdot I_{t-1} + \varepsilon_2$;

$r_t = a_3 + b_{31} \cdot Y_t + b_{32} \cdot M_t + \varepsilon_3$;

$Y_t = C_t + I_t + G_t$,

где C - расходы на потребление;

Y - ВВП;

I - инвестиции;

r - процентная ставка;

M - денежная масса;

G - государственные расходы;

t - текущий период;

$t-1$ - предыдущий период.

Задача 3.6

Модель протекционизма Сальватора (упрощенная версия):

$M_t = a_1 + b_{12} \cdot N_t + b_{13} \cdot S_t + b_{14} \cdot E_{t-1} + b_{15} \cdot M_{t-1} + \varepsilon_1$;

$N_t = a_2 + b_{21} \cdot M_t + b_{23} \cdot S_t + b_{26} \cdot Y_t + \varepsilon_2$;

$S_t = a_3 + b_{31} \cdot M_t + b_{32} \cdot N_t + b_{37} \cdot X_t + \varepsilon_3$,

где M - доля импорта в ВВП;

N - общее число прошений об освобождении от таможенных пошлин;

S - число удовлетворенных прошений об освобождении от таможенных пошлин;

E - фиктивная переменная, равная 1 для тех лет, в которые курс рубля на международных валютных рынках был искусственно завышен, и 0 - для всех остальных лет;

Y- реальный ВВП;

X - реальный объем чистого экспорта;

t - текущий период;

t-1 - предыдущий период.

Исходные данные

Текущий период *t*	Реальный ВВП *Y*	Доля импорта ВВП *M*	Общее число прошений об освобождении от таможенных пошлин *N*	Число удовлетворенных прошений об освобождении от таможенных пошлин *S*	Фиктивная переменная *E*	Реальный объем чистого экспорта *X*
1	1 398,5	0,129471	900	800	1	185,6
2	19 005,5	0,482632	2 200	1 500	1	11 847
3	171 509,5	0,304956	5 500	4 000	1	65 524
4	610 745,2	0,232131	10 500	6 000	1	169 534
5	1 524 049,0	0,242857	20 350	18 000	1	426 735
6	2 145 655,5	0,205965	20 000	15 000	0	532 239
7	2 478 594,1	0,209359	30 000	20 000	0	592 332
8	2 741 051,2	0,234951	45 000	35 000	0	840 596
9	4 757 233,7	0,268908	50 000	45 000	0	2 090 687
10	7 063 392,8	0,249292	45 000	40 000	0	3 232 388

Задача 3.7

Макроэкономическая модель (упрощенная версия модели Клейна):

$C_t = a_1 + b_{11} Y_t + b_{13} T_t + \varepsilon_1$;

$I_t = a_2 + b_{21} \cdot Y_t + b_{24} \cdot K_{t-1} + \varepsilon_2$;

$Y_t = C_t + I_t$

где *C* - потребление;

I - инвестиции;

Y - доход;

T — налоги;

K — запас капитала;

t - текущий период;

t-1 - предыдущий период.

Задача 3.8

Макроэкономическая модель экономики России (одна из версий):

$C_t = a_1 + b_{11} \cdot Y_t + b_{12} \cdot C_{t-1} + \varepsilon_1$ (функция потребления);

$I_t = a_2 + b_{21} \ Y_t + b_{23} \ r_t +$ (функция инвестиций);

$r_t = a_3 + b_{31} \cdot Y_t + b_{34} \cdot M_t + b_{35} \cdot r_{t-1} + \varepsilon_3$ (функция денежного рынка);

$Y_t = C_t + I_t + G_t$ (тождество дохода),

где C - потребление;

Y - ВВП;

I - инвестиции;

r - процентная ставка;

M - денежная масса;

G - государственные расходы;

t - текущий период;

$t-1$ - предыдущий период.

Задача 3.9

Модель Кейнса (одна из версий):

$C_t = a_1 + b_{11} \cdot Y_t + b_{12} \cdot Y_{t-1} + \varepsilon_1$ (функция потребления);

$I_t = a_2 + b_{21} \cdot Y_t + \varepsilon_2$ (функция инвестиций);

$Y_t = C_t + I_t + G_t$ (тождество дохода),

где C - потребление;

Y - ВВП;

I - валовые инвестиции;

G - государственные расходы;

t - текущий период;

$t-1$ - предыдущий период.

Задача 3.10

Модель денежного и товарного рынков:

$R_t = a_1 + b_{11} \cdot Y_t + b_{13} \cdot r_t + \varepsilon_1$ (функция товарного рынка);

$Y_t = a_2 + b_{21} \cdot R_t + b_{23} \cdot I_t + b_{25} \cdot G_t + \varepsilon_2$ (функция денежного рынка);

$I_t = a_3 + b_{31} \cdot R_t + \varepsilon_2$ (функция инвестиций),

где R - процентные ставки;

Y - реальный ВВП;

M - денежная масса;

I - внутренние инвестиции;

G - реальные государственные расходы.

Задача 3.11

Для прогнозирования спроса на свою продукцию предприятия используется следующая модель, характеризующая общую экономическую ситуацию:

$Q_t = a_1 + b_{11} \ Y_t +$;

$C_t = a_1 + b_{21} \ Y_t +$;

$I_t = a_3 + b_{32}\ (У_{t-1} - K_{t-1}) +$;

$Y_t = C_t + I_t$,

где Q - реализованная продукция в период t;

Y-ВДС;

C - конечное потребление;

I - инвестиции;

K - запас капитала;

t - текущий период;

$t-1$ - предыдущий период.

Исходные данные

Текущий период t	ВДС региона Y	Инвестиции I	Конечное потребление C	Реализованная продукция в период t Q	Запас капитала K
	560,5	210,5	450	52	325
	12 170	2 670	7 500	250	4 550
	77 725	271 25	40 600	790	34 965
	292 810	108 810	124 000	1390,4	133 209
	476 974	266 974	310 000	7318,2	327 941
	735 998	375 998	260 000	7524,4	454 369
	698 797	408 797	390 000	7323,6	482 451
	797 086	407 086	490 000	8804,5	485 452
	2 160 439	970 439	990 000	13130,3	766 672
	2415 181	1 165 181	1 650 000	14874,2	1 293 750

Задача 3.12

Модифицированная модель Кейнса:

$C_1 = a_1 + b_{11} \cdot Y_t + \varepsilon_1$;

$I_t = a_2 + b_{21} \cdot Y_t + b_{22} \cdot Y_{t-1} + \varepsilon_2$;

$Y_t = C_t + I_t + G_t$,

где C - расходы на потребление;

Y- доход;

I - инвестиции;

G - государственные расходы;

t - текущий период;

$t-1$ - предыдущий период.

Задача 3.13

Макроэкономическая модель:

$C_t = a_1 + b_{11} \cdot D_t + \varepsilon_1;$

$I_t = a_2 + b_{22} \cdot Y_t + b_{23} \cdot Y_{t-1} + \varepsilon_2;$

$Y_t = D_t + T_t;$

$D_t = C_t + I_t + G_t,$

где C - расходы на потребление;
Y- чистый национальный продукт;
D - чистый национальный доход;
I - инвестиции;
T - налоги;
G - государственные расходы;
t - текущий период;
$t-1$ - предыдущий период.

Задача 3.14

Дана следующая структурная форма модели:

$C_t = b_1 + b_2 \cdot S_t + b_3 \cdot P_t + \varepsilon_1;$

$S_t = a_1 + a_2 \cdot R_t + a_3 \cdot R_{t-1} + a_4 \cdot t + \varepsilon_2;$

$R_t = S_t + P_t,$

где C_t - личное потребление в период t;
S_t - зарплата в период t;
P_t - прибыль в период t;
R_t - национальный доход в период t;
R_{t-1} - национальный доход в период t-1;
$t-1$ - предыдущий период.

Задача 3.15

Эконометрическая модель имеет вид

$C_t = a + b \cdot Y_t + \varepsilon;$

$Y_t = C_t + I_t,$

где C - расходы на потребление;
Y - доход;
I - инвестиции;
t - текущий период.

Задача 3.16

Гипотетическая модель экономики:

$C_t = a_1 + b_{11} \cdot Y_t + b_{12} \cdot J_t + \varepsilon_1;$

$I_t = a_2 + b_{2l} \cdot Y_{t-1} + \varepsilon_2;$

$T_t = a_3 + b_{31} \cdot Y_t + \varepsilon_3;$

$Y_t = C_t + J_t + G_t,$

где C - совокупное потребление в период t;
Y - совокупный доход в период t;

J - инвестиции в период t;
T - налоги в период t;
G - государственные расходы в период t.

Задача 3.17

Эконометрическая модель имеет вид

$C_t = a + b \cdot Y_t + \varepsilon$;

$Y_t = C_t + I_t + G_t$;

где C - расходы на потребление;
Y- доход;
I - инвестиции;
G - государственные расходы;
t - текущий период.

Задача 3.18

Модель спроса и предложения на деньги:

$R_t = a_1 + b_{11} \cdot M_t + b_{12} \cdot Y_t + \varepsilon_1$;

$Y_t = a_2 + b_{21} \cdot R_t + \varepsilon_2$,

где R - процентные ставки в период t;
Y- ВВП в период t;
M- денежная масса в период t;
t - текущий период.

Задача 3.19

Модель денежного рынка:

$R_t = a_1 + b_{11} \quad M_t + b_{12} \quad Y_t + \quad ;$

$Y_t = a_2 + b_{21} \quad R_t + b_{22} \quad I_t + \quad ;$

$I_t = a_3 + b_{31} \quad R_t + \quad ,$

где R - процентные ставки;
Y - ВВП;
M - денежная масса;
I- внутренние инвестиции;
t - текущий период.

Задача 3.20

Имеется следующая модель кейнсианского типа:

$C_t = a_1 + b_{11} \cdot Y_t + b_{12} \cdot T_{t-1} + \varepsilon_1$ (функция потребления);

$I_t = a_2 + b_{21} \cdot Y_{t-1} + \varepsilon_2$ (функция инвестиций);

$T_t = a_3 + b_{31} \cdot Y_t + \varepsilon_3$ (функция денежного рынка);

$Y_t - C_t + I_t + G_t$ (тождество дохода),

где C - совокупное потребление в период t;
Y- совокупный доход в период t;

I - инвестиции в период времени *t;*

T - налоги в период времени *t;*

G - государственные расходы в период времени *t;*

Y_{t-1} - совокупный доход в период *t*-1;

t - текущий период;

t-1 - предыдущий период.

В этой модели переменные *C, I, T* и Y являются эндогенными.

Определите:

1. Каким методом вы будете оценивать структурные параметры этой мо дели?
2. Выпишите приведенную форму модели.
3. Кратко охарактеризуйте методику расчета параметров первого и второго структурного уравнения модели.

4. Фиктивные переменные

При изучении зависимостей между экономическими явлениями результативный признак принимает количественные значения, а факторы могут быть как количественными, так и качественными, т. е. не принимающие численных значений. Так, например, продуктивность сельскохозяйственных животных зависит от величины скармливаемых кормов, производственных затрат на одно животное, размеров поголовья и т. п. Эти переменные являются количественными. В то же время продуктивность животных зависит от качества кормов, породного состава, условий содержания животных. Эти переменные являются качественными. Для включения в уравнение регрессии качественных переменных необходимо заменить качественные категории числовыми значениями. В эконометрике такая замена производится с помощью фиктивных переменных, которые являются диахотомическими (бинарными) переменными. Под бинарной переменной понимают переменную, которая принимает значение 0 или 1.

Если на зависимую переменную (y) оказывает влияние один количественный фактор (x) и один не количественный (z), то линейная модель примет следующий вид:

$$Y = b_0 + b_1 x + c_1 z_1 + \varepsilon. \qquad (4.1)$$

Бинарная переменная z принимает два категориальных значения: 0 – если данная единица принадлежит первой категории, являющейся «эталонной», и 1 – если единица наблюдения принадлежит второй категории. В модели (4.1) предполагается, что коэффициент регрессии одинаков для обеих подвыборок, соответствующих выделенным категориям. Пусть y – сбережения на одного члена семьи в течении года, x – величина полученных доходов на одного члена семьи в течение года, а z– состав семьи. Тогда z=0, для семей, имеющих детей и z=1, для семей, не имеющих детей. Если предположить, что сбережения изменяются пропорционально величине дохода, то коэффициент регрессии будет один и тот же для семей бездетных и семей, имеющих детей. Тогда уравнение регрессии для семей, имеющих детей, будет иметь вид:

$$\hat{y} = b_0 + b_1 x, \qquad (4.2)$$

для семей, имеющих детей, где $z = 0$;

$$\hat{y} = b_0 + b_1 x + c = (b_0 + c) + b_1 x, \qquad (4.3)$$

для бездетных семей, т. к. $z = 1$.

Уравнения (4.2) и (4.3) отличаются величиной свободного члена, причем «c» будет показывать дополнительную величину сбережений в семьях, не имеющих детей. Графически уравнения (4.2) и (4.3) изображаются в виде двух параллельных прямых с одинаковым углом наклона.

Если коэффициенты регрессии по категориям различны, то используется следующая модель:

$$Y = b_0 + b_1 x + c_1 z_1 + c_2 z_1 x + \varepsilon \quad (4.4)$$

Если z=0, то линейное уравнение регрессии имеет вид:

$$\hat{y} = b_0 + b_1 x. \quad (4.5)$$

Если z=1, тогда линейное уравнение регрессии примет вид:

$$\hat{y} = b_0 + b_1 x + c_1 z_1 + c_2 x z_1 = b_0 + b_1 x + c_1 + c_2 x = (b_0 + c_1) + (b_1 + c_2) x,$$

$$\hat{y} = d_0 + d_1 x, \quad (4.6)$$

где $d_0 = b_0 + c_1, d_1 = b_1 + c_2$.

Графически уравнения (4.5) и (4.6) изображаются в виде двух прямых с разным углом наклона, отображающих влияние количественной переменной при разных значениях качественной переменной.

Качественная переменная может принимать несколько категориальных или альтернативных значений. В таких случаях в эконометрическую модель вводится число фиктивных переменных меньших на единицу числа категорий или альтернативных значений. В рассматриваемом примере семьи могут не иметь детей, семьи с одним, двумя, тремя, четырьмя и большим числом детей. Если выделяется пять категорий семей, то в регрессионную модель необходимо включить четыре фиктивных переменных:

$$y = b_0 + b_1 x + c_1 z_1 + c_2 z_2 + c_3 z_3 + c_4 z_4 \varepsilon. \quad (4.7)$$

В матрице исходных данных значения качественной переменной отображаются следующим образом:

Категория семей	Фиктивная переменная			
	z_1	z_2	z_3	z_4
Не имеющие детей	0	0	0	0
Имеющие одного ребенка	1	0	0	0
Имеющие двух детей	0	1	0	0
Имеющие трех детей	0	0	1	0
Имеющие четырех и более детей	0	0	0	1

Часто возникает необходимость проверки гипотезы на однородность изучаемой совокупности наблюдений в регрессионном смысле. Предположим, что совокупность, состоящая из n наблюдений, является неоднородной в отношении изучаемых зависимостей. Эту совокупность можно разбить на две одно-

родные выборки. Возникает вопрос – строить регрессионную модель по всей совокупности наблюдений или отдельные модели по каждой из двух выборок?

Эта задача решается с помощью теста Чоу. Находятся параметры уравнений регрессии по всей совокупности и по отдельным выборкам. Для каждого уравнения определяются остаточные суммы квадратов отклонений $SS_{\text{ост}}$по формулам (тема «Множественный корреляционно-регрессионный анализ»):

$$SS_{\text{ост}}=\sum \varepsilon^2=\sum(y-\hat{y})^2;\ SS_1=\sum \varepsilon_1^2=\sum(y_1-\hat{y}_1)^2;\ SS_2=\sum \varepsilon_2^2=\sum(y_2-\hat{y}_2)^2. \quad (4.8)$$

Выдвигается нулевая гипотеза, что коэффициенты регрессии в уравнениях по выборкам принимают одинаковые значения и эти выборки можно объединить в одну совокупность. Данная гипотеза проверяется с помощью *F*-критерия Фишера. Фактически наблюдаемое значение критерия определяется по формуле:

$$F_{\text{н}}=\frac{(SS_{\text{общ}}-SS_1-SS_2)}{(SS_1+SS_2)}\cdot\frac{n-m_1-m_2-2}{m_1+m_2+1-m}, \quad (4.9)$$

где m_1и m_2 – число коэффициентов при переменных в уравнениях, построенных по первой и второй выборкам;
m– число коэффициентов при переменных в уравнении, построенном по всей совокупности наблюдений.

Критическое значение находится при заданном уровне значимости α, числе степеней свободы$k_1=m_1+m_2+1-m$, $k_2=n-m_1-m_2-2$.

Если наблюдаемое значение критерия больше критического, то нулевая гипотеза отвергается и целесообразно строить уравнения регрессии по каждой выборке отдельно, т. е. с учетом фиктивной переменной. Если наблюдаемое значение критерия меньше критического, то нулевая гипотеза принимается, тогда вся совокупность считается однородной, по которой строится единое уравнение регрессии.

Задача 4.1.

По статистическим данным сельскохозяйственных предприятий Краснодарского края в разрезе муниципальных образований (Приложение Г)изучается влияние доз вносимых минеральных удобрений на урожайность озимой пшеницы.

1. С помощью инструмента анализа данных *Описательная статистика* рассчитать обобщающие характеристики вариационных рядов урожайности и доз вносимых минеральных удобрений, написав выводы по каждой переменной.

2. Провести парный регрессионный анализ влияния доз минеральных удобрений на урожайность озимой пшеницы.

3. Считая, что урожайность озимой пшеницы зависит от размещения посевов культуры по природно-экономическим зонам Краснодарского края, ввести в уравнение парной регрессии фиктивные переменные, отражающие зональные различия в урожайности.

4. Оценить значимость множественных коэффициентов регрессии с помощью *t*-критерия Стьюдента. Провести исключение несущественно влияющих переменных на изменение урожайности.

5. Оценить значимость множественного уравнения регрессии с помощью *F*-критерия Фишера, для чего составить таблицу дисперсионного анализа.

Написать выводы по результатам расчетов. Сравнить результаты регрессионного анализа по обеим вариантам расчетов.

6. Построить уравнения регрессии для районов: северной и западной зон; Анапо-Таманской и Южно-Предгорной зон.

7. Используя критерий Чоу, выяснить, можно ли выразить одним уравнением и охарактеризовать зависимость между урожайностью озимой пшеницы и количеством внесенных минеральных удобрений на 1 га посева.

5. Модели с дискретной зависимой переменной

При изучении социально-экономических процессов с использованием регрессионного анализа зависимая переменная может быть не только непрерывной, но и дискретной – бинарной или множественной.

Современный подход позволяет объединить все подходы в одной модели – обобщенной регрессии.

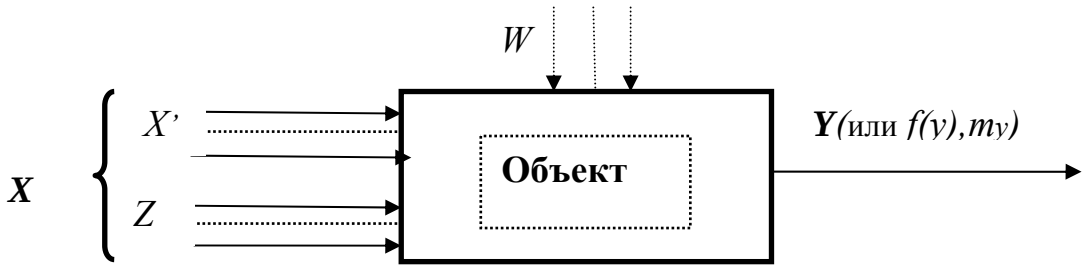


Рис. 5.1 – Регрессионная модель

Здесь $X = (X',Z)$ – факторы модели (вектор входных переменных); X'– управляемые независимые переменные; Z – контролируемые, но не управляемые факторы; Y – отклик системы («показатель качества управления», «выход» и т.п), $f(y)$– закон распределения, m_y–математическое ожидание случайной величины Y; W– помехи.

Таблица 5.1 – Таблица классификации

Шкалы измерений		Модель
независимых переменных, X	зависимой переменной, Y	
количественные, качественные	количественная	Линейная или нелинейнаярегрессия
количественные, качественные	дихотомическая (бинарная)	Бинарная регрессия (пробит-модель или логит-модель)
категориальные, количественные	категориальная (номинальная)	Модель множественного выбора с неупорядоченными альтернативами (мультиномиальная логистическая регрессия)
категориальные, количественные	категориальная (порядковая)	Модель множественного выбора с упорядоченными альтернативами (порядковая регрессия)

Бинарные зависимые переменные

В случае, когда зависимая переменная *y* принимает два значения 1 (успех) или 0 (неудача) обычно моделируется вероятность успеха:

$$P(y=1|x)=F(x); \tag{5.1}$$

$$P(y=0|x)=1-F(x), \tag{5.2}$$

где *F* – интегральная функция распределения.

Чаще всего в качестве *F* рассматривается функция логистического распределения

$$L(z)=\frac{1}{1+exp(-z)}, \tag{5.3}$$

или Функция Лапласа *Ф(x)* – функция распределения стандартной нормальной величины.

Первая из моделей называется логит-моделью (логистической регрессией), вторая – пробит-моделью.

1. Логит-модель. Бинарная логистическая регрессия обычно рассматривается, когда речь идет о событии, которое может произойти или не произойти, она позволяет оценить вероятность наступления в зависимости от значений независимых переменных.

Бинарная логистическая регрессия *(logit regression)*– это разновидность множественной регрессии, которая принимает два значения и имеет следующий вид:

$$P = \frac{1}{1+e^{-y}}, \tag{5.4}$$

где $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$

P – вероятность того, что, например, безработный не найдет работу в течение 60 дней.

2.Пробит-модель (*probit regression*) предполагает, что результативная переменная подчиняется нормальному закону распределения, а случайныевеличины являются нормально распределенными.

В обобщенном виде функцию стандартного нормального распределения можно представить в виде:

$$F(w) = \Phi(w) = \int_{-\infty}^{w} \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}t^2\right\} dt \tag{5.5}$$

3. *Модель множественного выбора с неупорядоченными альтернативами* применяется в том случае, если дискретная величина имеет некоторое количество неупорядоченных значений, то есть когда нереально предполагать, что между латентной переменной (т.е. переменной, у которой отсутствуют явные наблюдаемые значения) и наблюдаемыми исходами существу-

ет монотонное соотношение. В качестве примера можно привести способ транспортировки грузов (пешком, велосипедом, самолетом, железнодорожным транспортом или самолетом). Тогда вероятность наступления какого-либо события будет описываться моделями:

$$p_{ij} = P(y_i=j|x_i) \sim G(x_i^T\beta_j), \quad (5.6)$$

где $j=1,\ldots., m-1, x_i^T\beta_j=\beta_0+\beta_1x_1+\ldots+\beta_kx_k$

$$p_{ij} = P(y_i=j|x_i)=\frac{G(x_i^T\beta_j)}{1+\sum_{k=1}^{m-1}G(x_i^T\beta_k)}\ , \quad (5.7)$$

$$p_{im} = P(y_i=m|x_i)=\frac{1}{1+\sum_{k=1}^{m-1}G(x_i^T\beta_k)}. \quad (5.8)$$

Показателем, характеризующим данные модели, является величина полезности, которая каждую альтернативную величину представляет линейной функцией, состоящих из наблюдаемых переменных и аддитивных (суммарных) остатков. При анализе с использованием модели множественного выбора с неупорядоченными альтернативами логично предположить, что выбор выпадет на значения, обладающие наибольшей полезностью.

Предположим, что необходимо сделать выбор между несколькими вариантами: $j = 1, 2, 3,\ldots, N,$ а показатель полезности субъективного выбора примем за U_{ij}(*Utiety*) (то есть, субъект *i*выбирает вариант *j*). Тогда показатель полезности можно представить в виде: $U_{ij} = max\ \{U_{i1}, U_{i2}, U_{iN}\}$.

Следует отметить, что для определения величины полезности, необходимо выдвинуть ряд предположений, для объективной постановки задачи.

*4. Модель множественного выбора с упорядоченными альтернативами*применяется в том случае, если имеются ранжированные значения зависимой переменной. Данная модель предполагает наличие некой латентной величины, которая представляет собой логически упорядоченные альтернативные признаки.

Например, при поиске работы субъект может ее найти через разные (упорядоченные) промежутки времени: до 30 дней, от 30 до 60 дней, от 60 дней и более. Срок трудоустройства зависит от ряда важных причин (факторов). Введем условные обозначения: $y_i=1$– трудоустройство до 30 дней; $y_i=2$– трудоустройство от 30 до 60 дней; $y_i=3$ – трудоустройство более, чем через 60 дней. Видно, что в данном случае существует логический порядок при оценке сроков поиска работы. Соответственно, латентная переменная y' может интерпретироваться как «желание найти работу». Модель множественного выбора с упорядоченными альтернативами будет выглядеть следующим образом:

$$y'=x'_i\beta+\varepsilon_i$$
$$y_i=1, \text{если } y'\le 0;$$
$$y_i=2, \text{ если } 0< y'\le\gamma$$
$$y_i=3, \text{ если } y'>\gamma$$

где γ – неизвестный параметр, анализ которого происходит совместно с коэффициентами β, который в представленном примере характеризуется с двух позиций: например, положительный коэффициент интерпретируется как желание быстрей трудоустроиться и, соответственно, наоборот.

$$P\{y_i = 1|x_i\} = P\{y'_i \leq 0|x_i\} = \Phi(-x'_i\beta), \quad (5.9)$$

$$P\{y_i = 3|x_i\} = P\{y'_i > \gamma|x_i\} = 1 - \Phi(\gamma - x'_i\beta), \quad (5.10)$$

$$P\{y_i = 2|x_i\} = \Phi(\gamma - x'_i\beta) - \Phi(-x'_i\beta). \quad (5.11)$$

Если предположить, что в моделях β – коэффициент является положительным, то, соответственно, переменная y' будет увеличиваться, тем самым повышая вероятность положительного исхода $y_i=3$ и снижая вероятность положительного решения для $y_i=1$. Но, вероятность наступления положительного исхода для $y_i=2$ непредсказуема, то есть она может в равной степени как возрастать, так и убывать.

Логистическая регрессия успешно решает задачу линейной классификации, для оценки, качества которой используется инструмент *ROC*-анализа (*Receiver Operator Characteristic Curve*).

У нас положительное событие – в течение 60 дней безработный не нашел работу (z_1), а– отрицательное, «нашел» (z_0). Результаты классификации представляются в виде таблицы классификации (таблица 5.2).

Таблица 5.2 - Таблица классификации

Прогноз	Фактически	
	Положительно (H_0)	Отрицательно (H_1)
Положительно (H_0)	*TP*	*FP*
Отрицательно (H_1)	*FN*	*TN*

В таблице 5.2 отражено количество примеров, полученных в результате применения логистической модели:

TP (True Positives) – верно классифицированных положительных примеров (истинно положительные случаи);

TN (True Negatives) – верно классифицированных отрицательных примеров (истинно отрицательные случаи);

FN (False Negatives) – положительных примеров, классифицированных как отрицательные (ошибка I рода). Это так называемый «ложный пропуск» – когда интересующее нас событие ошибочно не обнаруживается (ложно отрицательные примеры);

FP (False Positives) – отрицательные примеры, классифицированные как положительные (ошибка II рода). Это ложное обнаружение, т.к. при отсутствии события ошибочно выносится решение о его присутствии (ложно положительные случаи).

Таблица 5.3 - Основные понятия *ROC*-анализа

Формулы	Понятия
$TPR = \frac{TP}{TP + FN} \cdot 100\%$	*Доля истинно положительных примеров* (*TruePositivesRate*)
$FPR = \frac{FP}{TN + FP} \cdot 100\%$	*Доля ложно положительных примеров* (*FalsePositivesRate*)
$Se = TPR = \frac{TP}{TP + FN} \cdot 100\%$	*Чувствительность* (*Sensitivity*)– доля истинно положительных случаев
$Sp = 100 - FPR = \frac{TN}{TN + FP} \cdot 100$	*Специфичность* (*Specificity*) – доля истинно отрицательныхслучаев, которые были правильно идентифицированы моделью

Модель с высокой чувствительностью часто дает истинный результат при наличии положительного исхода (обнаруживает положительные примеры). Наоборот, модель с высокой специфичностью чаще дает истинный результат при наличии отрицательного исхода (обнаруживает отрицательные примеры).

В системе координат с абсциссой (*FPR=100%–Sp) и* ординатой *Se* строится *ROC*-кривая – множество пар точек (*Sp, Se*), полученных для порога отсечения (*optimalcut-offvalue*)с определенным шагом (например, 0,01). Чем ближе *ROC*-кривая к диагонали (*y=x*), тем она хуже, чем ближе к левому углу – тем лучше. Сравнение моделей между собой можно проводить с использованием показателя площади под кривой -*AUC* (*AreaUnderCurve*).

Значение порога отсечения, влияющего на соотношение *Se* и *Sp* соответствует стратегии исследования:

$ocov_1$ – максимальная специфичность (чувствительность) предполагает обеспечить долю истинно отрицательных случаев не ниже определенной границы (например, 90%);

$ocov_2$ –максимальная суммарная чувствительность и специфичность модели, $C = \max_k (Se_k + Sp_k)$;

$ocov_3$ – баланс между чувствительностью и специфичностью, т.е. когда $Se \approx Sp$: $C = \min_k |Se_k - Sp_k|$.

Пример. Рассмотрим результаты применения теории логистической регрессии к задаче оценки безработных с низким уровнем активности при поиске работы (не нашедших работу в течение 60 дней с момента постановки на учет в службу занятости).

Характеристика исходных данных. Имеется 12429 наблюдения зарегистрированных безработных в службе занятости одного из районов Краснодарского края в период 1998-2012 гг.:

*t*1–период 1998-1998 гг.;

*t*2– период 1999-2008 гг.;

*t*3 – период 2009-2012 гг.;

*edu*1 – нет общего образования;

*edu*2 – имеет основное общее образование;
*edu*3 – имеет среднее общее образование;
*edu*4 – имеет общее профессиональное образование;
*edu*5 – имеет среднее профессиональное образование;
*edu*6 – имеет высшее профессиональное образование;
lnW – логарифм заработной платы на последнем месте работы в ценах 2012 года;
*exp*0 – стаж на последнем рабочем месте;
exp – общий стаж работы;
*exp*2 – общий стаж в квадрате;
age – возраст;
*age*2 – возраст в квадрате;
gen – пол (1 – мужской, 0 – женский);
city – место жительства (1–город, 0 – сельская местность).

Вообще говоря, современные информационные технологии позволяют формировать базу данных о безработных, классифицировать их и использовать для решения задач диагностики.

Класс безработных, зарегистрированных в службе занятости не нашедших работу в течение двух месяцев – класс с положительными исходами (истинно положительные примеры), класс безработных, нашедших работу за 60 дней – с отрицательными исходами (ложно отрицательные примеры). *ROC*–кривая *(Receiver Operator Characteristic Curve*) показывает зависимость количества верно классифицированных положительных примеров от количества неверно классифицированных отрицательных примеров. Меняя порог отсечения (от 0 до 1) можно получать разные классификаторы с различными ошибками *I и II рода*.

Выходная переменная логистической регрессии – «устройство на работу в течение 60 дней», принимающая два значения: отрицательное – «устроился», положительное – «не устроился», полученные через два месяца после регистрации в службе занятости. Ниже рассмотрен вариант входных переменных свариантом порога (точки) отсечения – баланс между чувствительностью и специфичностью, т.е. $Se \approx Sp$.

Чтобы оценить качество модели используются аналоги R^2, так называемые *pseudo* R^2 и *McFadder* R^2, имеющие несколько иной смысл, чем R^2. Если логарифмическую функцию правдоподобия обозначить l, а ограниченную функцию правдоподобия $\bar{l}$ (где свободный член равен нулю), тогда $l \geq \bar{l}$. Качество модели можно оценить исходя из величины различий между моделями. Модель считается тем лучше, чем больше имеются различия между показателями *pseudo* R^2 и *McFadder* R^2, которые можно определить следующим образом.

$$pseudoR^2 = 1 - \frac{1}{1 + \frac{2(l - \bar{l})}{N}}$$

где N – объем выборки;

$$McFadderR^2 = 1 - \frac{l}{\bar{l}}$$

Для получения модели логистической регрессии используем команду:

. logistic z1 t3 edu1 edu2 edu3 edu4 edu5 exp0 lnwexp exp2 gen age age2 city

Таблица 5.4 –Результат анализа данных с использованием логит-модели

```
Logistic regression                               Number of obs   =      12429
                                                  LR chi2(14)     =     462.26
                                                  Prob > chi2     =     0.0000
Log likelihood = -6569.4857                       Pseudo R2       =     0.0340
```

z1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
t3	.8360824	.0390686	-3.83	0.000	.7629113	.9162714
edu1	.7342966	.1239974	-1.83	0.067	.5273913	1.022375
edu2	.7048194	.0570756	-4.32	0.000	.6013791	.8260519
edu3	.8327539	.0600941	-2.54	0.011	.7229218	.9592726
edu4	.8477809	.0605353	-2.31	0.021	.7370622	.9751314
edu5	1.110692	.0716964	1.63	0.104	.9786959	1.260491
exp0	1.023715	.0050743	4.73	0.000	1.013817	1.033708
lnw	1.094996	.0272257	3.65	0.000	1.042914	1.149679
exp	1.010294	.0107616	0.96	0.336	.9894203	1.031608
exp2	1.000273	.0002814	0.97	0.332	.9997218	1.000825
gen	.7400853	.0345618	-6.45	0.000	.675353	.811022
age	.9972096	.0197267	-0.14	0.888	.9592858	1.036633
age2	1.000143	.0002539	0.57	0.572	.999646	1.000641
city	.7595576	.0338738	-6.17	0.000	.6959851	.8289369
_cons	1.440496	.5494483	0.96	0.339	.6820851	3.042187

Из полученной таблицы итогов логистической регресии очевидно, что ряд переменных необходимо отбросить из-за проблем значимости (age;age2).

После преобразования диапозона входных переменных, введем следующую команду:

. logistic z1 t3 edu1 edu2 edu3 edu4 edu5 exp0 lnwexp exp2 gen city

Таблица 5.5 –Результат логит-регрессии после корректировки

```
Logistic regression                                     Number of obs   =      12429
                                                        LR chi2(12)     =     457.87
                                                        Prob > chi2     =     0.0000
Log likelihood = -6571.6825                             Pseudo R2       =     0.0337
```

z1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
t3	.8531128	.0389228	-3.48	0.000	.7801369	.9329151
edu1	.7569398	.1273284	-1.66	0.098	.5443481	1.052558
edu2	.7169611	.0577347	-4.13	0.000	.6122812	.8395377
edu3	.8432599	.0606093	-2.37	0.018	.7324556	.9708264
edu4	.8556521	.0609054	-2.19	0.029	.7442324	.9837525
edu5	1.111034	.071462	1.64	0.102	.9794401	1.260308
exp0	1.022825	.0050426	4.58	0.000	1.012989	1.032756
lnw	1.090108	.0268714	3.50	0.000	1.038693	1.144068
exp	1.01746	.0069305	2.54	0.011	1.003967	1.031134
exp2	1.000299	.0001985	1.51	0.132	.99991	1.000688
gen	.7427234	.0344587	-6.41	0.000	.6781653	.8134271
city	.7560923	.033635	-6.29	0.000	.6929611	.8249749
_cons	1.494663	.3361494	1.79	0.074	.9618519	2.322621

Полученные результаты можно считать удовлетворительными, за исключением нескольких переменных (edu1, edu5, exp2).

Для получения модели логистической регрессии отвечающей нашим целям (получения адекватного прогноза) мы выберем вариант значений входных переменных с вариантом порога отсечения – баланс между чувствительностью и специфичностью, т.е. $Se \approx Sp$. Для выбора порога отсечения введем команду:

. lsens, recast(scatter)

На рисунке 5.2 наглядно показано пересечение кривых, одна из которых представляет собой долю истинно положительных случаев (*чувствительность)*, а другая – долю истинно отрицательных случаев *(специфичность),* построенных по данным службы занятости одного из районов Краснодарского края.

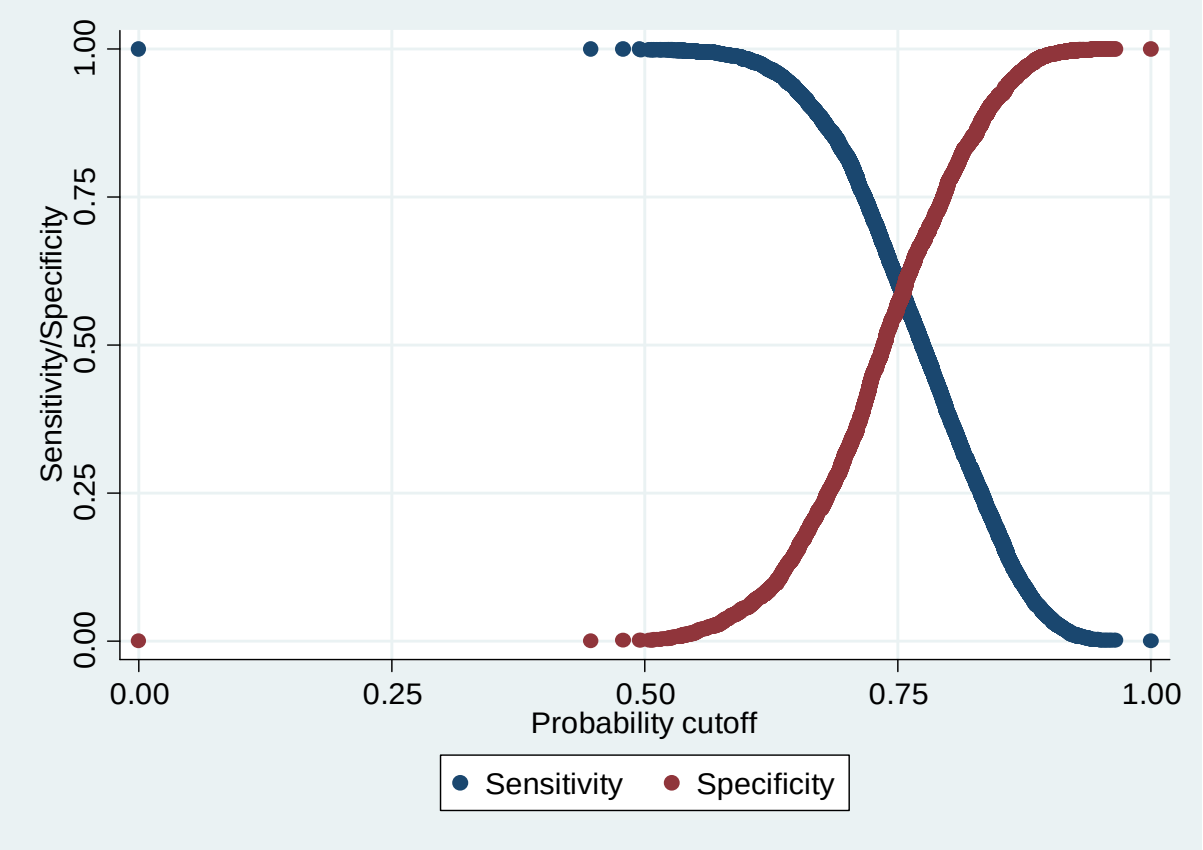


Рисунок 5.2 – Вероятность точки отсечения

Из рисунка 5.2 видно, что точка отсечения, при которой ***Se*≈*Sp***, равна 0,75.

Для получения матрицы классификаций при заданном значении точки отсечения (0,75) зададим команду:

. estatclassification, cutoff(0.75)

В результате получим таблицу 5.6, которая показывает, что из 12429 наблюдений истинно положительных 5775 случаев, а верно классифицированных отрицательных примеров – 1666.

Негативных событий, которые классифицировались как положительные – 1275 случаев и положительные примеры, классифицируемые как отрицательные, составили 3713 показателей.

В нашем случае примерное равенство между чувствительностью и специфичностью (*Se*≈*Sp*) соблюдено (60,87%≈56,65%).

Величина корректно классифицированных показателей составляет 59,87 %.

Таблица 5.6 – Матрица классификаций с учетом точки отсечения

Logistic model for z1

Classified	True D	True ~D	Total
+	5775	1275	7050
–	3713	1666	5379
Total	9488	2941	12429

Classified + if predicted Pr(D) >= .75
True D defined as z1 != 0

Sensitivity	Pr(+| D)	60.87%
Specificity	Pr(–|~D)	56.65%
Positive predictive value	Pr(D| +)	81.91%
Negative predictive value	Pr(~D| –)	30.97%
False + rate for true ~D	Pr(+|~D)	43.35%
False – rate for true D	Pr(–| D)	39.13%
False + rate for classified +	Pr(~D| +)	18.09%
False – rate for classified –	Pr(D| –)	69.03%
Correctly classified		59.87%

Можно уточнить модель, отбросив менее значимые предикторные переменные и введя команду:

. logistic z1 t3 edu1 edu2 edu3 edu4 exp0 lnwexp gen city

Таблица 5.7 – Уточненная модель

Logistic regression — Number of obs = 12429; LR chi2(10) = 452.61; Prob > chi2 = 0.0000; Log likelihood = -6574.3089; Pseudo R2 = 0.0333

z1	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
t3	.8531186	.0388214	-3.49	0.000	.7803246	.9327033
edu1	.712637	.1160943	-2.08	0.038	.5178462	.9806997
edu2	.6736803	.0475515	-5.60	0.000	.5866405	.7736342
edu3	.7884176	.0474864	-3.95	0.000	.7006296	.8872054
edu4	.7969126	.0475777	-3.80	0.000	.7089111	.8958383
exp0	1.023167	.0050078	4.68	0.000	1.013399	1.03303
lnw	1.081013	.0263331	3.20	0.001	1.030614	1.133877
exp	1.02734	.0023471	11.81	0.000	1.02275	1.03195
gen	.7453942	.034437	-6.36	0.000	.6808646	.8160397
city	.7558	.0335986	-6.30	0.000	.6927352	.8246061
_cons	1.647185	.3495852	2.35	0.019	1.086652	2.496862

Качество логистической регрессии обычно оценивается с помощью *ROC* – анализа.

ROC curve или *кривая ошибок* представляет собой график, построенный по двум значениям: количество верно классифицированных показателей и неверно классифицированных признаков с учетом установленной точки отсечения. Качество модели оценивается путем измерения площади кривой *AUC* (*AreaUnderCurve*). Данная величина является довольно условной, но принято считать, что с увеличением площади возрастает и качество модели. Этот показатель используют для сравнения различных моделей классификаций. Так, значение 0,5 характеризует плохую классификацию, а более, чем 0,75 является хорошей.

Для проведения *ROC* – анализа введем команду, результаты которой представлены на рисунке 5.2.

. lroc, recast(scatter)

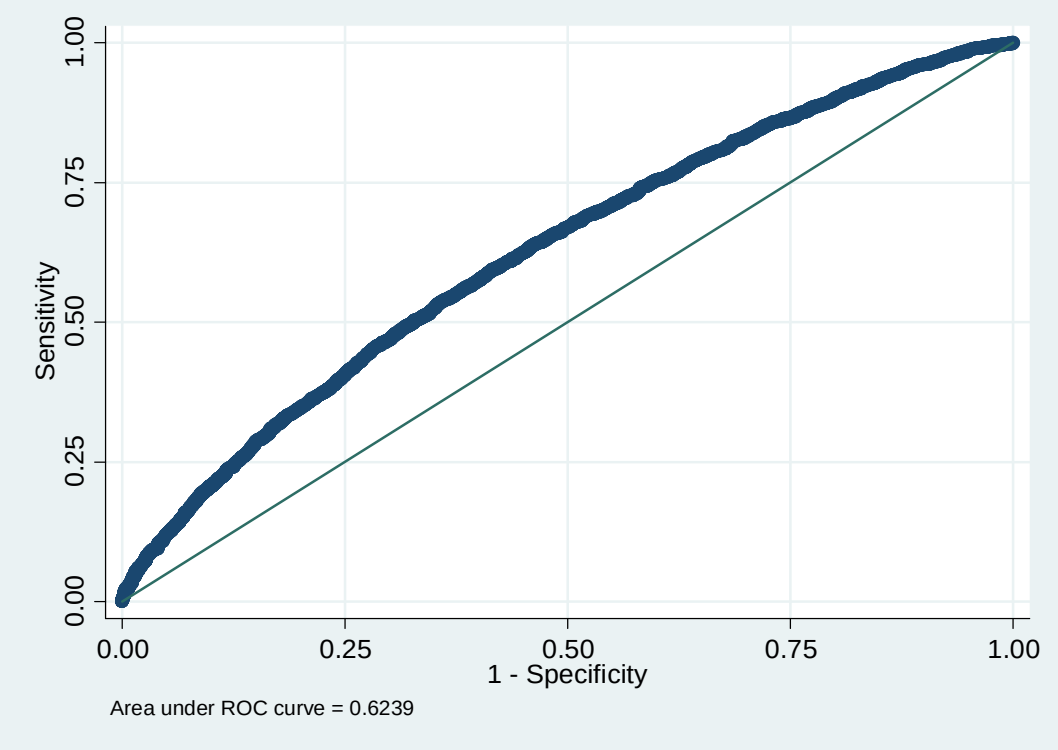


Рисунок 5.2 – Результат ROC – анализа

В нашем примере данная величина составляет 0,6239, что характеризует удовлетворительное качество полученной модели, но при анализе реальных данных зачастую получаются небезукоризненные результаты.

Мы можем сравнить качество классификаций двух моделей логистической регрессии.

. estat classification, cutoff(0.75)

Logistic model for z1

Classified	True D	True ~D	Total
+	5879	1310	7189
-	3609	1631	5240
Total	9488	2941	12429

Classified + if predicted Pr(D) >= .75
True D defined as z1 != 0

Sensitivity	Pr(+| D)	61.96%
Specificity	Pr(-|~D)	55.46%
Positive predictive value	Pr(D| +)	81.78%
Negative predictive value	Pr(~D| -)	31.13%
False + rate for true ~D	Pr(+|~D)	44.54%
False - rate for true D	Pr(-| D)	38.04%
False + rate for classified +	Pr(~D| +)	18.22%
False - rate for classified -	Pr(D| -)	68.87%
Correctly classified		60.42%

Если в качестве результативной переменной ввести переменную *z,* принимающей два значения: 1 – работа найдена в течение 60 дней, 0 – работа не найдена в течение 60 дней. То есть по сравнению с переменной *z1* выбраны противоположные обозначения. Построим логистическую регрессию с теми же предикторными переменными, что и ранее и выходной переменной *z:*

. logistic z t3 edu1 edu2 edu3 edu4 exp0 lnwexp gen city

Logistic regression — Number of obs = 12429
LR chi2(10) = 452.61
Prob > chi2 = 0.0000
Log likelihood = -6574.3089 — Pseudo R2 = 0.0333

z	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
t3	1.17217	.0533399	3.49	0.000	1.072152	1.281518
edu1	1.403239	.228599	2.08	0.038	1.01968	1.931075
edu2	1.484384	.1047748	5.60	0.000	1.292601	1.704621
edu3	1.268363	.0763935	3.95	0.000	1.127135	1.427288
edu4	1.254843	.0749173	3.80	0.000	1.116273	1.410614
exp0	.9773573	.0047836	-4.68	0.000	.9680264	.9867781
lnw	.9250583	.0225341	-3.20	0.001	.8819301	.9702957
exp	.9733879	.0022238	-11.81	0.000	.969039	.9777563
gen	1.341572	.0619802	6.36	0.000	1.225431	1.468721
city	1.323101	.0588177	6.30	0.000	1.2127	1.443553
_cons	.6070963	.1288452	-2.35	0.019	.4005028	.9202582

Полученные результаты в точности повторяют предыдущие, то есть когда использовалась зависимая переменная *z1*.

Построив график зависимости между чувствительностью и специфичностью ($Se \approx Sp$) по переменной *z*, определим точку отсечения, которая, как и можно было ожидать, равна 0,25.

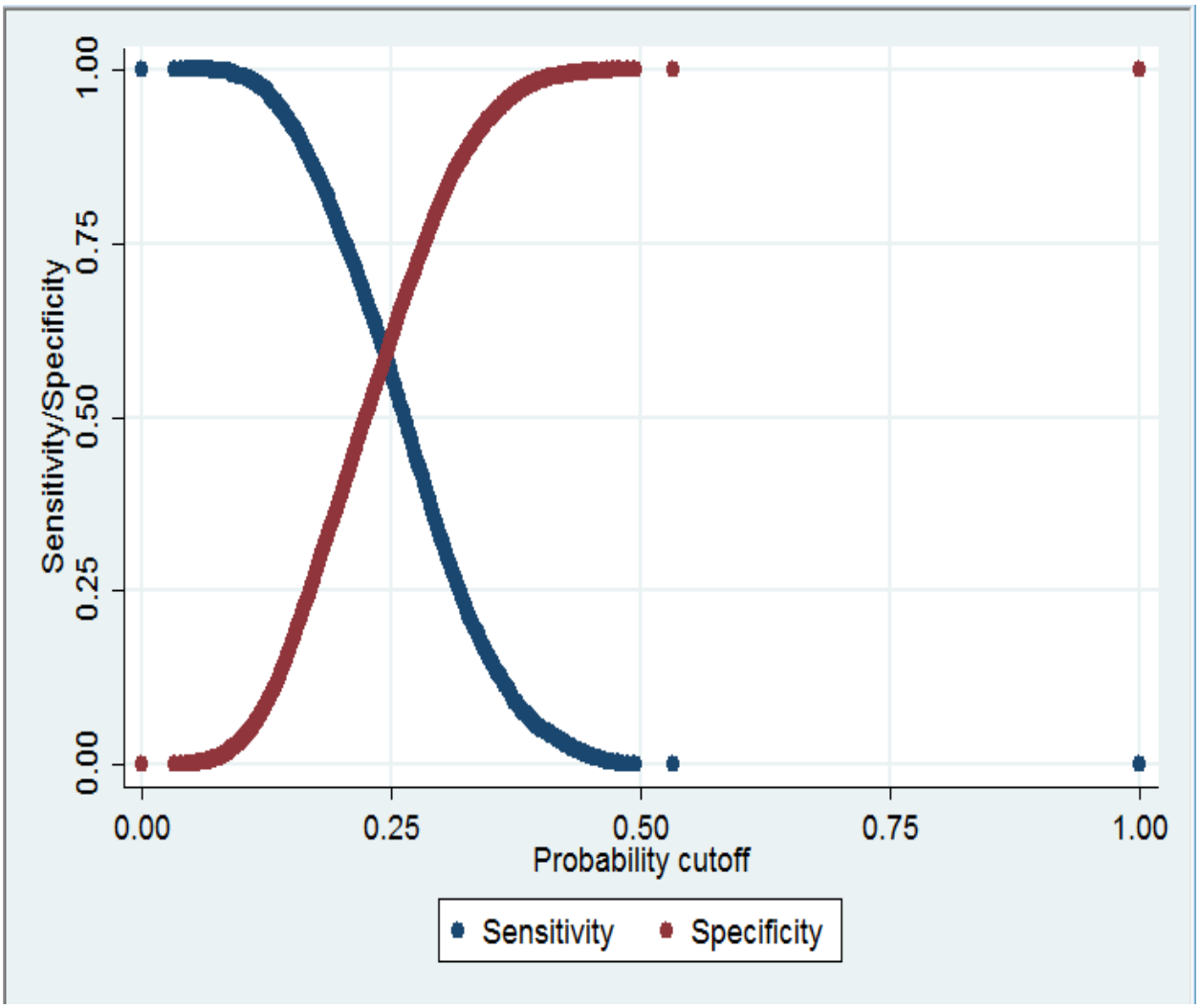


Рисунок 5.3 – Точка отсечения

Построим модель пробит-регрессии при помощи команды:

. probit z1 t3 edu1 edu2 edu3 edu4 exp0 lnwexp gen city

Полученные результаты являются идентичными вышеописанным моделям.

Таблица 5.8 – Результат анализа пробит-модели

Probit regression — Number of obs = 12429; LR chi2(10) = 451.55; Prob > chi2 = 0.0000; Log likelihood = -6574.8401; Pseudo R2 = 0.0332

z	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
t3	.0914999	.0265125	3.45	0.001	.0395363	.1434634
edu1	.195403	.0949011	2.06	0.039	.0094002	.3814059
edu2	.2261887	.0422489	5.35	0.000	.1433823	.3089951
edu3	.131609	.0352433	3.73	0.000	.0625334	.2006846
edu4	.1272193	.0347687	3.66	0.000	.0590739	.1953646
exp0	-.0126633	.0026782	-4.73	0.000	-.0179125	-.0074142
lnw	-.0417853	.013938	-3.00	0.003	-.0691033	-.0144674
exp	-.0158539	.0013122	-12.08	0.000	-.0184258	-.013282
gen	.1746244	.0271042	6.44	0.000	.1215011	.2277477
city	.1628857	.0258301	6.31	0.000	.1122597	.2135117
_cons	-.3492799	.1220573	-2.86	0.004	-.5885078	-.110052

Задача.

По имеющимся данным о зарегистрированных безработных построить альтернативные модели, отличающиеся набором переменных (от приведенных в тексте показателей), и выбрать ту, которая лучше описывает исходные данные.

6. Временные ряды

6.1 Модели временных рядов по функционированию экономических объектов на микро- и мезоуровне

Экономические процессы и явления, их связи и значимость могут рассматриваться как в пространстве, так и во времени, путем построения и анализа одного или нескольких временных рядов.

Временной ряд- это совокупность значений изучаемого показателя за последовательные моменты или периоды времени. Он состоит из значений или уровней временного ряда (*y*) и периодов или моментов времени (*t*).

t	*1*	*2*	*3*	. . .	n
Y_t	Y_1	Y_2	Y_3	. . .	Y_n

(6.1)

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

Каждый уровень временного ряда формируется под воздействием большого числа факторов, которые условно можно подразделить на три группы:

- факторы, формирующие тенденцию ряда – трендовая компонента (*T*);
- факторы, формирующие циклические или сезонные колебания ряда – циклическая компонента (*S*);
- случайные факторы - случайная компонента (ε).

Учитывая перечисленные три компоненты, рассматривают две модели временных рядов:

- $Y_t = T + S + \varepsilon$ - аддитивную;
- $Y_t = T * S * \varepsilon$-мультипликативную.

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 ряде тенденции и циклических колебаний значения каждого последующего уровня ряда зависят от предыдущих уровней. Корреляционная 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 несколько шагов во времени, называемого коэффициентом автокорреляции.

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

$$r_1 = \frac{\sum_{t=2}^{n}(y_t - \overline{y_1})\cdot(y_{t-1} - \overline{y_2})}{\sqrt{\sum_{t=2}^{n}(y_t - \overline{y_1})^2 \cdot \sum_{t=2}^{n}(y_{t-1} - \overline{y_2})^2}}, \qquad (6.2)$$

где:

$$\overline{y_1} = \frac{\sum_{t=2}^{n} y_t}{n-1}; \qquad \overline{y_2} = \frac{\sum_{t=2}^{n} y_{t-1}}{n-1}.$$

Коэффициент автокорреляции уровней ряда второго порядка, смещенных на две единицы времени:

$$r_2 = \frac{\sum_{t=3}^{n}(y_t - \overline{y}_3)\cdot(y_{t-2} - \overline{y}_4)}{\sqrt{\sum_{t=3}^{n}(y_t - \overline{y}_3)^2 \cdot \sum_{t=3}^{n}(y_{t-2} - \overline{y}_4)^2}}, \qquad (6.3)$$

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

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 больше абсолютное значение коэффициента автокорреляции первого порядка, тем более выражена линейная тенденция. Для некоторых временных рядов, имеющих сильную нелинейную тенденцию, коэффициент автокорреляции уровней исходного ряда может приближаться к нулю.

Последовательность 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 втмоментов 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 не линейную 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 временного ряда. Тенденция во времени может принимать разные формы, для ее формализации используют следующие функции:

- линейная:$\hat{y}_t = a + bt$;
- гипербола: $\hat{y}_t = a + \frac{b}{t}$;
- экспонента: $\hat{y}_t = e^{a+bt}$;
- степенная $\hat{y}_t = at^b$;
- показательная: $\hat{y}_t = ab^t$;
- парабола k-ого порядка: $\hat{y}_t = a + b_1 t + b_2 t^2 + \cdots + b_k t^k$;

Параметры каждой из перечисленных выше функций определяются методом наименьших квадратов (МНК), 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 6.1).

Таблица 6.1 - Линеаризующие преобразования

Функция	Преобразования переменных	
	y	*t*
$y_t = a + b / t$	*y*	*1/t*
$y_t = e^{a+bt}$ или $y_t = a \cdot b^t$	*ln y*	*t*
$y_t = a\, t^b$	*lgy*	*lgt*
$y_t = a+b_1t+b_2t^2+\ldots+b_kt^k$	*y*	$t_1=t,\ t_2=t^2,\ldots t_k=t^n$

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 экспоненциального тренда:*a*– начальный уровень временного ряда в момент времени *t = 0*; e^b или *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.

Очень часто моделирование рядов динамики с помощью перечисленных выше функций не дает удовлетворительных результатов, так как имеют место периодические колебания вокруг общей тенденции. В таких случаях используют спектральный анализ или, как частный случай, выравнивание по ряду Фурье (гармоники ряда Фурье):

$$\hat{y}_t = a + \sum b_k \sin kt + \sum c_k \cos kt. \qquad (6.4)$$

Параметры b_kи c_k находятся с помощью МНК, в результате применения которого, получим:

$$a = \frac{1}{n}\sum_{t=1}^{n} y_{t,} c_k = \frac{2}{n}\sum_{t=1}^{n} y_t \cos kt, b_k = \frac{2}{n}\sum_{t=1}^{n} y_t \sin kt. \qquad (6.5)$$

Применение этого метода позволяет определить закон, по которому можно достаточно точно спрогнозировать значения уровней ряда

Пример 6.1. Имеются данные о валовом сборе винограда в хозяйствах Краснодарского края.

Таблица 6.2 - Валовой сбор винограда в хозяйствах Краснодарского края

Год	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Валовой сбор, тыс. т	112	205	138	168	85	137	122	137	132	202

Требуется:

а) построить график временного ряда;

б) рассчитать коэффициент автокорреляции первого порядка;

в) обосновать выбор типа уравнения тренда и рассчитать его параметры;

г) дать интерпретацию параметров тренда и сделать выводы по задаче.

Решение.

а) Рассмотрим систему координат y_0t, где y_t – валовой сбор, t_e – порядковый номер года и нанесем в нее данные (рисунок 6.1):

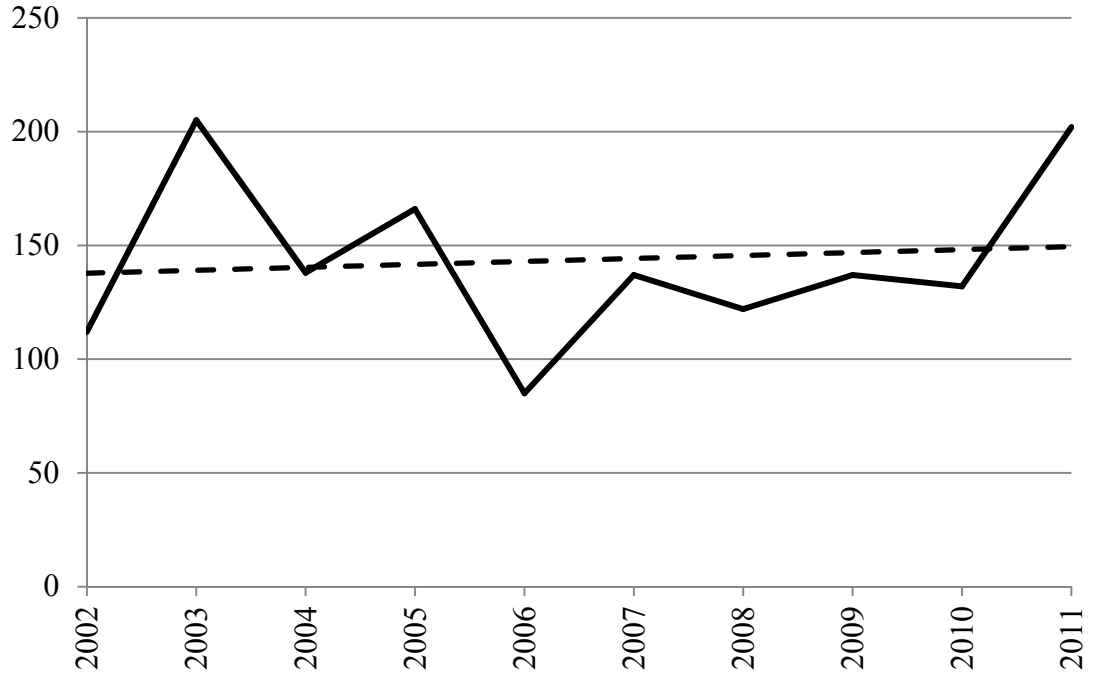


Рисунок 6.1 – Валовой сбор винограда в хозяйствах Краснодарского края

б) Определим коэффициент автокорреляции первого порядка, для этого заполним вспомогательную таблицу 6.3.

Таблица 6.3 - Вспомогательная таблица для расчета коэффициента автокорреляции

T	y_t	y_{t-1}	$y_t - \overline{y}_1$	$y_{t-1} - \overline{y}_2$	$(y_t - \overline{y}_1) \cdot (y_{t-1} - \overline{y}_2)$	$(y_t - \overline{y}_1)^2$	$(y_{t-1} - \overline{y}_2)^2$
1	112	-	-	-	-	-	-
2	205	112	57,6667	-25,3333	-1460,8889	3325,4444	641,7778
3	138	205	-9,3333	67,6667	-631,5556	87,1111	4578,7778
4	168	138	20,6667	0,6667	13,7778	427,1111	0,4444
5	85	168	-62,3333	30,6667	-1911,5556	3885,4444	940,4444
6	137	85	-10,3333	-52,3333	540,7778	106,7778	2738,7778
7	122	137	-25,3333	-0,3333	8,4444	641,7778	0,1111
8	137	122	-10,3333	-15,3333	158,4444	106,7778	235,1111
9	132	137	-15,3333	-0,3333	5,1111	235,1111	0,1111
10	202	132	54,6667	-5,3333	-291,5556	2988,4444	28,4444
Сумма	1438	1236	-	-	-3569,0000	11804,0000	9164,0000

$$\overline{y_1} = \frac{\sum_{t=2}^{n} y_t}{n-1} = \frac{1438 - 112}{10 - 1} = 147,3333;$$

$$\overline{y_2} = \frac{\sum_{t=2}^{n} y_{t-1}}{n-1} = \frac{1236}{9} = 137,3333.$$

$$r_1 = \frac{\sum_{t=2}^{n} (y_t - \overline{y_1}) \cdot (y_{t-2} - \overline{y_2})}{\sqrt{\sum_{t=2}^{n} (y_t - \overline{y_1})^2 \cdot \sum_{t=2}^{n} (y_{t-1} - \overline{y_2})^2}} = \frac{-3569}{\sqrt{11804 \cdot 9164}} = -0,34315$$

в) Полученное значение коэффициента автокорреляции и графическое изображение временного ряда позволяют сделать вывод о том, что ряд валового сбора винограда содержит тенденцию близкую к линейной. Поэтому для моделирования его тенденции используем линейную функцию *y=a+bt.*

Таблица 6.4 - Вспомогательная таблица для расчета параметров линейного тренда

Год	t	y	$y \cdot t$	t^2	$\hat{y}_t$
2002	1	112	112	1	138,1273
2003	2	205	410	4	139,3879
2004	3	138	414	9	140,6485
2005	4	168	672	16	141,9091
2006	5	85	425	25	143,1697
2007	6	137	822	36	144,4303

Продолжение таблицы 6.4

Год	t	y	$y \cdot t$	t^2	$\hat{y}_t$
2008	7	122	854	49	145,6909
2009	8	137	1096	64	146,9515
2010	9	132	1188	81	148,2121
2011	10	202	2020	100	149,4727
Сумма	55	1438	8013	385	-
Среднее значение	5,5	143,8	801,3	38,5	-

Для расчета параметров линейного тренда *a*и *b* используем метод наименьших квадратов:

$$b = \frac{\overline{yt} - \bar{y} \cdot \bar{t}}{\overline{t^2} - \bar{t}^2} = \frac{801,1 - 5,5 \cdot 143,8}{38,5 - 5,5^2} = 1,261;$$

$$a = \bar{y} - b\bar{t} = 143,8 - 1,262 \cdot 5,5 = 136,8667 \Rightarrow$$

$$\hat{y}_t = 136,8667 + 1,261$$

Таким образом, в среднем ежегодно валовой сбор винограда во всех категориях хозяйств Краснодарского края за 2002 – 2011 гг. увеличивался на 1,26 тыс. тонн.

Применяется несколько способов определения типа тенденции. Одним из них является *метод последовательных разностей.*Он заключается в следующем:

- если первые разности *(*$\Delta_t = y_t - y_{t-1}$) равны между собой, ряд содержит линейный тренд;
- если вторые разности ($\Delta_t^2 = \Delta_t - \Delta_{t-1}$) равны между собой, ряд содержит параболический тренд;
- если между собой равны отношения уровней ряда (y_t / y_{t-1}), ряд содержит экспоненциальный тренд.

Однако наиболее распространенным является качественный анализ изучаемого процесса, построение и визуальный анализ графика зависимости уровней ряда от времени. Выбор наилучшего уравнения в случае, если ряд содержит нелинейную тенденцию, можно осуществить путем перебора основных форм тренда, расчета по каждому уравнению скорректированного коэффициента детерминации R^2 и выбора уравнения тренда с максимальным значением скорректированного коэффициента детерминации. Данный метод относительно легко реализуется при компьютерной обработке данных.

Пример 6.2. Имеются данные о валовом сборе чайного листа в Краснодарском крае, тонн (таблица 6.5).

Таблица 6.5 – Валовой сбор чайного листа в Краснодарском крае

Год	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Валовой сбор, т	1239	1288	871	1240	1261	633	815	630	373	267

Требуется обосновать выбор типа уравнения тренда и рассчитать его параметры.

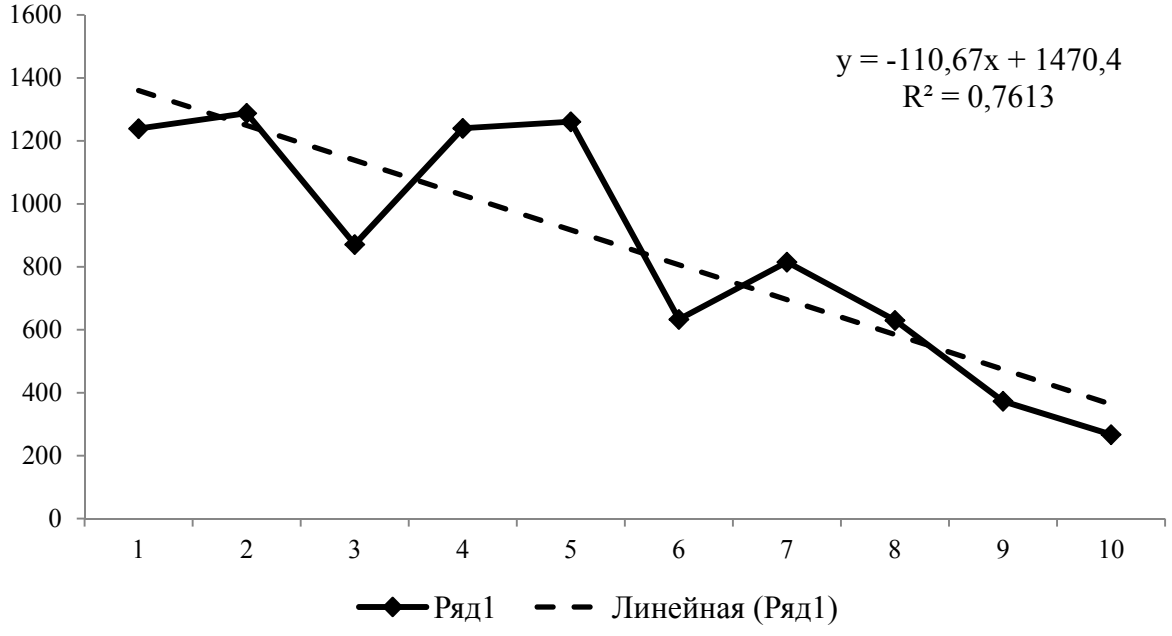
Решение.

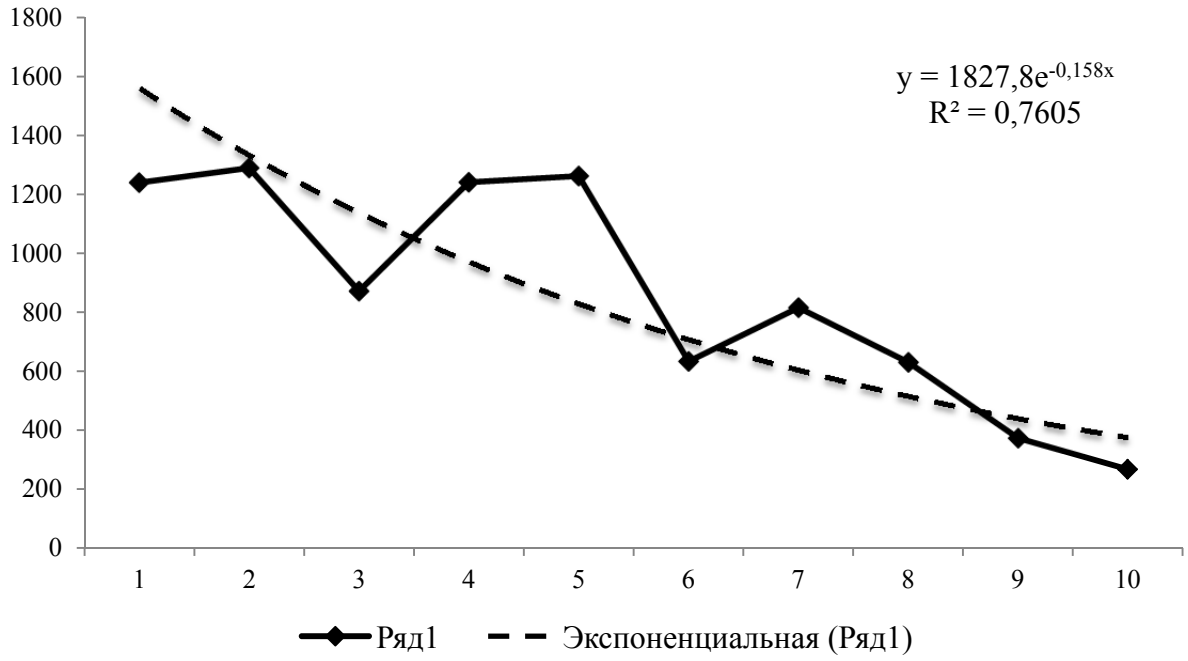
Решим данную задачу с использованием *ППП MSExcel*.

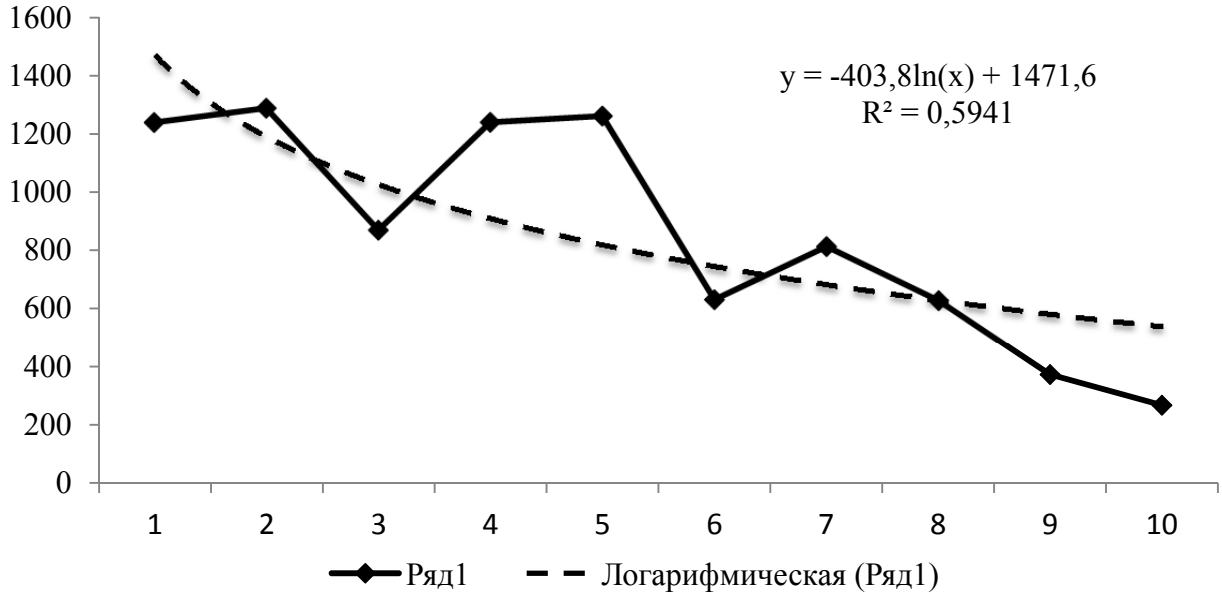
1. Введем исходные данные:
   - в главном меню выберем *Вставка*;
   - в окне *Диаграммы*выберем *Точечная*;
   - выделив область построения, заполним параметры диаграммы;

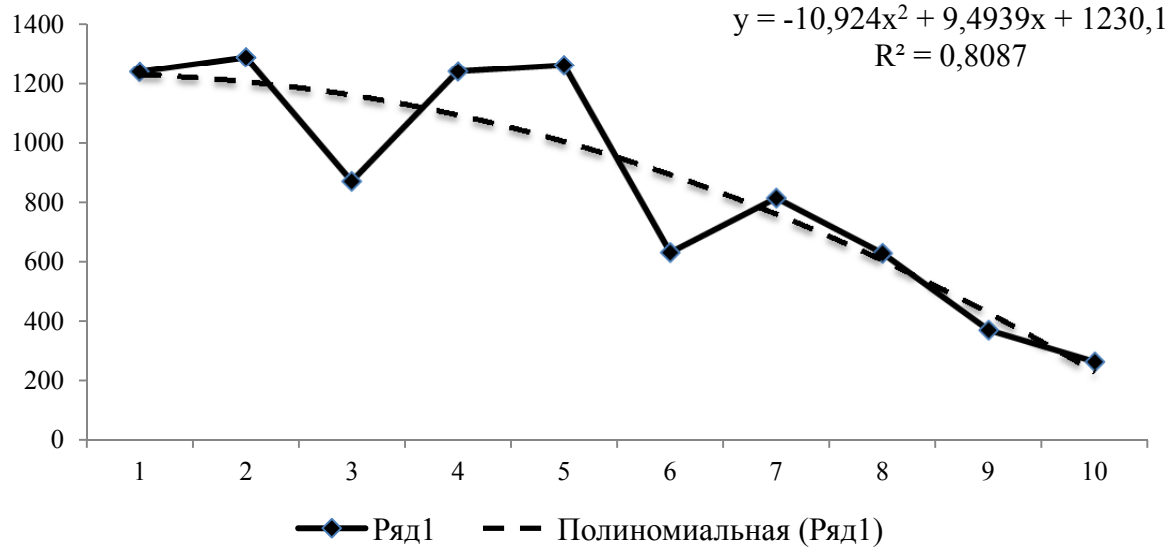
2. Добавим в диаграмму линии основных трендов и параметры уравнений данных трендов. Для этого, выделив область построения диаграммы:
   - в главном меню выберем *Макет/Анализ/Линия тренда*;
   - в появившемся диалоговом окне выберем вид линии тренда и в закладке *готовых линий тренда* установим флажки на строках «показывать уравнение на диаграмме» и «поместить на диаграмму величину достоверности аппроксимации (R^2).

На рисунке 6.2 представлены различные виды трендов, описывающие исходные данные задачи.




y = 1827,8e-0,158x
R² = 0,7605
1800
1600
1400
1200
1000
800
600
400
200
0
1 2 3 4 5 6 7 8 9 10
Ряд1
Экспоненциальная (Ряд1)


y = -403,8ln(x) + 1471,6
R² = 0,5941
1600
1400
1200
1000
800
600
400
200
0
1 2 3 4 5 6 7 8 9 10
Ряд1
Логарифмическая (Ряд1)


y = -10,924x² + 9,4939x + 1230,1
R² = 0,8087
1400
1200
1000
800
600
400
200
0
1 2 3 4 5 6 7 8 9 10
Ряд1
Полиномиальная (Ряд1)

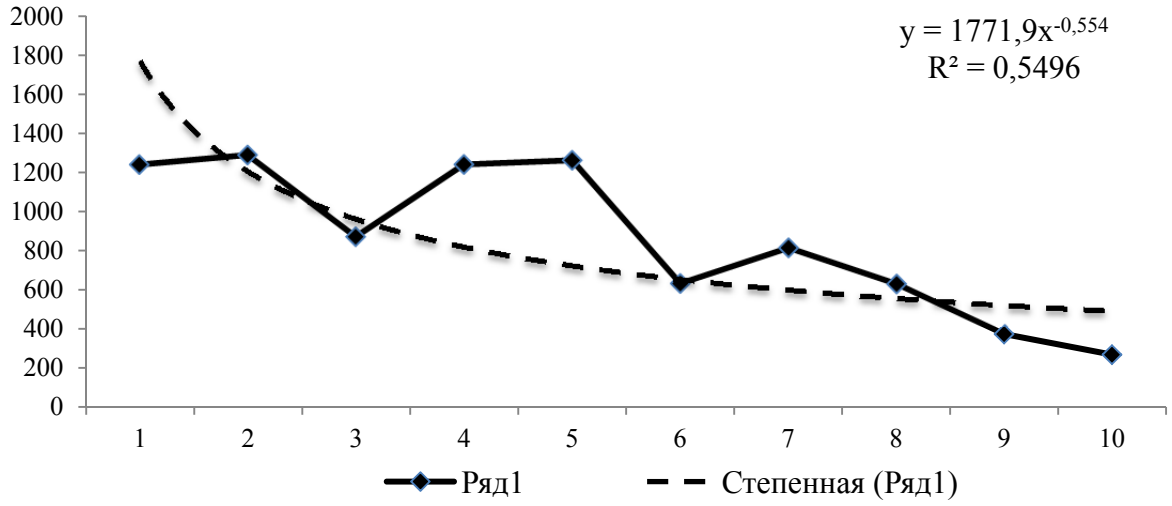


Рисунок 6.2 - Различные виды трендов, описывающие динамику валового сбора чайного листа в Краснодарском крае

3. Сравним значения R^2 по различным уравнениям трендов:

- линейный - R^2= 0,7613;
- степенной - R^2= 0,5496;
- экспоненциальный - R^2 = 0,7605;
- парабола 2-ого порядкаR^2 = 0,8087;
- логарифмический R^2 = 0,5941.

Так как наиболее высокое значение имеет скорректированный коэффициент детерминации уравнения тренда параболы 2-ого порядка, то именно данный тренд лучше всего описывает тенденцию валового сбора чайного листа в Краснодарском крае. Следовательно, уравнение параболы 2-ого порядка следует использовать для расчета прогнозных значений.

На формирование временного ряда оказывают влияние три группы факторов: факторы, формирующие тенденцию рядаT(тренд); факторы, формирующие циклические или сезонные колебания S; случайные факторы ε. Как правило, реальные экономические данные содержат все три компоненты, а значит, в большинстве случаев, фактический уровень ряда можно представить как сумму или произведение трендовой, циклической (или сезонной) и случайной компонент временного ряда.

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 ($Y=T\cdot S\cdot \varepsilon$).

Выбор одной из двух моделей осуществляется на основе анализа структуры сезонных колебаний. Если амплитуда колебаний приблизительно постоянна, строят аддитивную модель, в которой значения сезонных компонент предполагаются постоянными для различных циклов. Если амплитуда колебаний возрастает или уменьшается, строят мультипликативную модель, которая ставит уровни ряда в зависимость от значений сезонной компоненты. Построение ад-

дитивной и мультипликативной моделей сводится к расчету значенийT, S и ε для каждого уровня ряда. Процесс построения модели включает в себя следующие шаги:

а) выравнивание исходного ряда методом скользящей средней;

б) расчет значений сезонной компоненты S;

в) устранение сезонной компоненты

г) из исходных уровней ряда и получение выровненных данных (Т+ε) в аддитивной или ($T \cdot \varepsilon$) в мультипликативной модели;

д) аналитическое выравнивание уровней ($T+\varepsilon$) или ($T \cdot \varepsilon$) и расчет значений Т с использованием полученного уравнения тренда;

е) расчет полученных по модели значений ($T+S$) или ($T \cdot S$);

ж) расчет абсолютных и (или) относительных ошибок.

Метод скользящей средней величины позволяет вместо данных каждого периода получить средние из уровней рядом стоящих. Такая средняя величина называется скользящей, т.к. из нее постоянно вычитается один член и добавляется другой, т.е. границы осреднения все время меняются.

Метод аналитического выравнивания состоит в подборе для данного ряда такой теоретической линии, которая выражает основные черты или закономерности изменения уровней явления.

Пример 6.3. Имеются поквартальные данные по заявленной предприятиями потребности в работниках за 4 года по Краснодарскому краю, тыс. чел.:

Таблица 6.6 – Потребность в работниках по Краснодарскому краю

№ квартала, t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Потребность в работниках, $У$	27,6	31,7	28,8	22,6	27,6	37,7	37,1	24,7	29,6	34,7	33,4	25,8	30,5	40,2	38,1	26,4

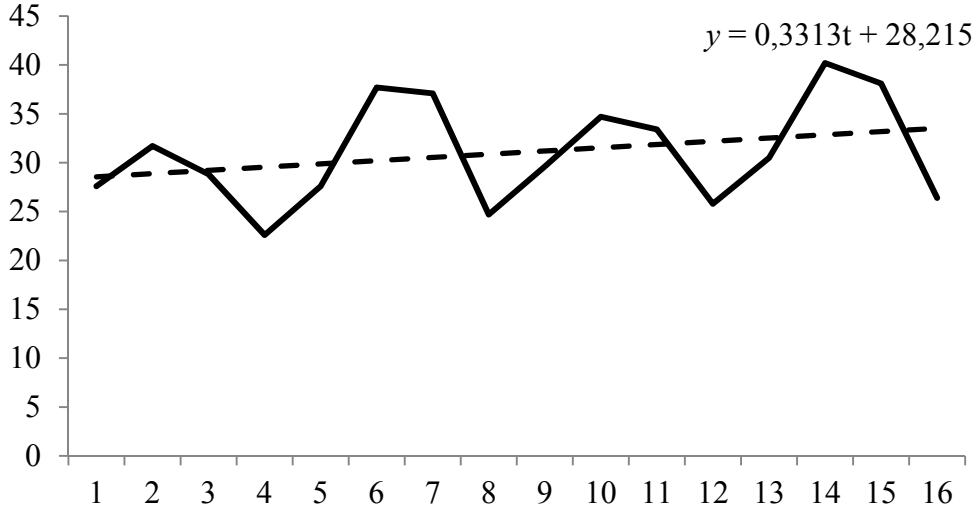


Рисунок 6.4 - Данные потребности в работниках за 4 года

График рассматриваемого временного ряда позволяет сделать вывод о приблизительно постоянной амплитуде колебаний (рисунок 6.3). Следовательно, для анализа структуры временных рядов построим аддитивную модель.

Шаг 1. Производится выравнивание исходных уровней ряда методомцентрированной скользящей средней, так как периодичность составляет 4 квартала (таблица 6.7):

Определятся сумма и средняя за первые четыре периода, затем сумма и средняя за четыре периода, начиная со второго и т.д.

$$\overline{y}_1 = \frac{y_1+y_2+y_3+y_4}{4}; \overline{y}_2 = \frac{y_2+y_3+y_4+y_5}{4}; \overline{y}_3 = \frac{y_3+y_4+y_5+y_6}{4}$$ и т.д.

Запишем суммы и средние уровни, начиная с третьего. Так как суммы и средние располагаютсямежду датами, то их необходимо центрировать и относить центрированные средние к определенному периоду.

Таблица 6.7– Расчет скользящих средних величин и оценка сезонной компоненты

№ квартала, *t*	Потребность в работниках, тыс.чел., *У*	Итого за4 квартала	Скользящая средняя за 4 квартала	Центрированная скользящая средняя	Оценка сезонной компоненты, *S*
1	27,6				
2	31,7				
		110,7	27,675		
3	28,8			27,675	1,125
		110,7	27,675		
4	22,6			28,425	-5,825
		116,7	29,175		
5	27,6			30,2125	-2,6125
		125	31,25		
6	37,7			31,5125	6,1875
		127,1	31,775		
7	37,1			32,025	5,075
		129,1	32,275		
8	24,7			31,9	-7,2
		126,1	31,525		
9	29,6			31,0625	-1,4625

Продолжение таблицы 6.7

№ квартала, t	Потребность в работниках, тыс.чел., y	Итого за4 квартала	Скользящая средняя за 4 квартала	Центрированная скользящая средняя	Оценка сезонной компоненты, S
		122,4	30,6		
10	34,7			30,7375	3,9625
		123,5	30,875		
11	33,4			30,9875	2,4125
		124,4	31,1		
12	25,8			31,7875	-5,9875
		129,9	32,475		
13	30,5			33,0625	-2,5625
		134,6	33,65		
14	40,2			33,725	6,475
		135,2	33,8		
15	38,1				
16	26,4				

Шаг 2. Находятся оценки сезонной компоненты как разности между фактическими уровнями ряда и центрированными скользящими средними (последняя колонка таблицы 6.7). Эти оценки используются для расчета значений сезонной компоненты S (таблица 6.8).

Таблица 6.8 – Определение скорректированной оценки сезонной компоненты

	Год	Номер квартала			
		I	II	III	IV
Данные колонки «Оценка сезонной компоненты	1	-	-	1,125	-5,825
	2	-2,6125	6,1875	5,075	-7,2
	3	-1,4625	3,9625	2,4125	-5,9875
	4	-2,5625	6,475	-	-
Итого за i-й квартал (по всем данным)		-6,6375	16,625	8,6125	-19,0125
Средняя оценка сезонной компоненты для i-ого квартала, $\overline{S_i}$		-2,2125	5,5417	2,8708	-6,3375
Скорректированная сезонная компонента, S_i		-2,178125	5,576075	2,905175	-6,303125

В моделях с сезонной компонентой сезонные воздействия за период взаимопогашаются. В аддитивной модели это означает, что сумма значений сезонной компоненты по всем кварталам должна быть равна нулю. Для нашей модели:-2,2125+5,5417+2,8708-6,3375=-0,1375.

Определяется корректирующий коэффициент: k=-0,1375/4= -0,034375.
Заполняется последняя строка в таблице 6.8:

$$S_i = \overline{S}_i - k$$

Проверим равенство нулю: -2,178125+5,576075+2,905175-6,303125=0

Шаг 3. Элиминируется влияние сезонной компоненты, вычитая ее значение из каждого уровня исходного временного ряда и получая величины$T+\varepsilon = Y-S$ (таблица 6.9).

Определяется компонентаT данной модели. Для этого проводится аналитическое выравнивание ряда$T+\varepsilon$ с помощью линейного тренда.
Трендовое уравнение объема продаж: T=28,215+0,3313t.

Подставляя в тренд значения tот 1 до 16, заполним колонку в таблице 6.9.

Таблица 6.9 – Определение компонент временного ряда по аддитивной модели

№ квартала	Потребтребность в работниках, Y	Сезонная компонента S_i	$Y-S=T+\varepsilon$	T	$T+S$	$\varepsilon=Y-(T+S)$	ε^2
1	27,6	-2,178125	29,778125	28,5463	26,36815	1,231825	1,51739
2	31,7	5,576075	26,123925	28,8776	34,453675	-2,753675	7,58272
3	28,8	2,905175	25,894825	29,2089	32,114075	-3,314075	10,9830
4	22,6	-6,303125	28,903125	29,5402	23,237075	-0,637075	0,40586
5	27,6	-2,178125	29,778125	29,8715	27,693375	-0,093375	0,00871
6	37,7	5,576075	32,123925	30,2028	35,778875	1,92112	3,69072
7	37,1	2,905175	34,194825	30,5341	33,439275	3,660725	13,4009
8	24,7	-6,303125	31,003125	30,8654	24,562275	0,137725	0,27545
9	29,6	-2,178125	31,778125	31,1967	29,018575	0,581425	0,33805
10	34,7	5,576075	29,123925	31,528	37,104075	-2,404075	5,77957
11	33,4	2,905175	30,494825	31,8593	34,764475	-1,364475	1,86179
12	25,8	-6,303125	32,103125	32,1906	25,887475	-0,087475	0,00765
13	30,5	-2,178125	32,678125	32,5219	30,343775	0,156225	0,02440
14	40,2	5,576075	34,623925	32,8532	38,429275	1,770725	3,13546
15	38,1	2,905175	35,194825	33,1845	36,089675	2,010325	4,04140
16	26,4	-6,303125	32,703125	33,5158	27,212675	-0,812675	0,66044
Сумма	496,5	-	-	-	-	-	53,7136

Шаг 4. Находим значения уровней ряда, полученные по аддитивной модели: *T+S* и заполняем колонку в таблице 6.9.

Шаг 5. Проводится расчет ошибки: $\mathcal{E}$*=У-(T+S)* и заполняется колонка таблицы 6.9. Это абсолютные ошибки. По аналогии с моделью регрессии для оценки качества построения модели применяется сумма квадратов полученных абсолютных ошибок, которая равна 53,71367. Находится общая сумма квадратов отклонений уровней ряда от его среднего уровня *(ППП MSExcel /Формулы/Вставить формулу/статистические/КВАДРОТКЛ):*

$$\sum(y_i - \overline{y})^2 = 425{,}6944 \;\rightarrow R^2 = 1 - \frac{53{,}71367}{425{,}6944} = 0{,}874.$$

Следовательно, можно сказать, что аддитивная модель объясняет 87,4% общей вариации уровней временного ряда.

Построение мультипликативной модели проводится аналогично, только в шаге 2 учитывается, что сумма значений сезонной компоненты должна быть равна числу периодов в цикле (в рассмотренном примере этот цикл равен 4) и, начиная с шага 3,выполняются расчеты с уравнением $Y=T\cdot S\cdot \mathcal{E}$.

Прогнозирование по аддитивной модели. Прогнозное значение F_tесть сумма трендовой и сезонной компонент: $F_t=S+T$.

T_{17}= 28,215+0,3313*17=33,8471 тыс.чел. за квартал.

Соответствующая сезонная компонента равна -2,178125, т.к. это 1-ый квартал. Следовательно, прогноз на этот квартал 33,8471 - 2,178125≈31,669 тыс.чел.

Задача 6.1.

По имеющимся данным о поголовье крупного рогатого скота в сельскохозяйственных предприятиях: а) построить уравнение линейного тренда; б) дать интерпретацию параметров линейного тренда; в) определить коэффициент детерминации для линейного тренда.

Год , *t*	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Поголовье скота, тысяч голов, у	733	729	673	582	520	490	478	470	460	440

Задача 6.2.

Используя данные по валовому надою молока в сельскохозяйственных организациях Краснодарского края обосновать выбор уравнения тренда, рассчитать параметры выбранного тренда и дать прогноз о количестве крупного рогатого скота в процентах в общей структуре поголовья скота в 2015 году.

Год , *t*	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Валовой надой молока, тыс. тонн, у	1004	977	908	891	885	869	853	905	870	851

Задача 6.3.

По данным по производству мяса (в убойном весе) в сельскохозяйственных хозяйствах всех категорий Краснодарского края построить экспоненциальный тренд, дать интерпретацию его параметров.

Год , *t*	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Мясо , тыс. тонн *y*	287	305	301	297	360	368	378	375	394	429

Задача 6.4.

По данным по производству меда в сельскохозяйственных предприятиях Краснодарского края рассчитать коэффициента автокорреляции 1 – ого и 2 – ого порядков, обосновать выбор уравнения тренда, дать прогноз производства меда на 2016 год.

Год , *t*	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Мед, тонн	3474	3505	3508	3605	3570	3475	2999	3059	2853	2586

Задача 6.5.

Имеются поквартальные данные по численности незанятого населения, зарегистрированного в органах государственной службы занятости за 3 года по Краснодарскому краю, тыс. чел: 18,3; 16,9; 18,4; 21; 21,6; 18,2; 16,5; 17,4; 17,5; 16,4; 21,7; 24,2. Обосновать выбор модели временного ряда и рассчитать значение компонент *T, S* и *ε*. Дать прогноз на четвертый год по кварталам численности незанятого населения.

Задача 6.6.

По данным об урожайности сельскохозяйственных культур (приложение А), по одному варианту:

а) построить график временного ряда;

б) рассчитать коэффициент автокорреляции 1-ого порядка;

в) анализируя предыдущие результаты, обосновать выбор типа уравнения тренда и рассчитать его параметры;

г) дать интерпретацию параметров тренда и сделать выводы по задаче.

Задача 6.7.

В приложении Б приведен временной ряд урожайности зерновых и зернобобовых культур в Краснодарском крае за 1961-2011гг. в хозяйствах всех категорий и в сельскохозяйственных организациях в расчете на 1 га посевной площади. Провести эконометрическое моделирование урожайности одной из сельскохозяйственных культур за весь период.

1. Определить статистические характеристики урожайности культуры за весь период и по десятилетним периодам: средние значения; средние квадратические отклонения; коэффициенты вариации; коэффициенты устойчивости.

2. Построить график временного ряда урожайности, по которому визуально оценить общую тенденцию изменения урожайности, выделить, по возможности, периоды устойчивого роста, снижения или стабильности уровня. Подобрать вид уравнения тренда, отражающего тенденцию изменения урожайности.

3. Рассчитать автокорреляционную функцию, которую графически изобразить в виде коррелограммы.

4. Методом наименьших квадратов определить параметры уравнения тренда. Оценить адекватность модели тренда с помощью критерия Дарбина – Уотсона.

5. Построить модели урожайности с применением фиктивных переменных сдвига и наклона, учитывающих характерные этапы изменения, а также модели авторегрессии.

6. Провести сравнительный анализ полученных моделей с выбором наиболее приемлемой модели для целей прогнозирования.

7. Определить точечные и интервальные прогнозы урожайности на период до 2016 года.

Задача 6.8

Имеются следующие данные о величине дохода на одного члена семьи и расхода на товар А

Показатель	2007 г.	2008 г.	2009 г.	2010 г.	2011 г.	2012 г.
Расходы на товар А , у.е.	33	37	40	43	48	53
Доход на одного члена семьи, % к 2005 г.	100,0	103,0	105,0	107,0	110,0	113,0

Требуется:

1. Определить ежегодные абсолютные приросты доходов и расходов и сделать выводы о тенденции развития каждого ряда.

2. Перечислить основные пути устранения тенденции для построения модели спроса на товар А в зависимости от дохода.

3. Построить линейную модель спроса, используя первые разности уровней исходных динамических рядов.

4. Пояснить экономический смысл коэффициента регрессии.

5. Построить линейную модель спроса на товар А, включив в нее фактор времени. Интерпретировать полученные параметры.

Задача 6.9

Имеются следующие данные о величине дохода на одного члена семьи и расхода на товар А.

Показатель	2002г	2003г	2004г	2005г	2006г	2007г	2008г	2009г	2010г	2011г
Расходы на товар А , у.е.	30	34	37	40	45	50	53	55	58	60
Доход на одного члена семьи, % к 2002 г.	100	103	105	107	110	113	115	117	120	122

Требуется:

1. Определить ежегодные абсолютные приросты доходов и расходов и сделать выводы о тенденции развития каждого ряда.
2. Перечислить основные пути устранения тенденции для построения модели спроса на товар Ав зависимости от дохода.
3. Построить линейную модель спроса, используя первые разности уровней исходных динамических рядов.
4. Пояснить экономический смысл коэффициента регрессии.
5. Построить линейную модель спроса на товар А, включив в нее фактор времени. Интерпретировать полученные параметры.

6.2. Модели временных рядов по функционированию экономических объектов на макроуровне

Гиперболический рост уровня технологического развития Мир-Системы. Гипотеза Кузнеца-Кремера: абсолютные темпы технологического роста на момент t пропорциональны уровню развития технологий и численности населения на данный момент времени. В математической форме она записывается в виде дифференциального уравнения 1-го порядка

$$\frac{dT}{dt} = bNT,$$

гдеT – уровень технологического развития Мир-Системы;

N – численность населения;

b-коэффициент, соответствующей средней продуктивности изобретательской работы одного обитателя Мир-Системы.

Задача 6.10

По данным приведенным в работе [Марков, Коротаев] провести эмпирическую проверку гипотезы Х. фон Ферстера, П. Мора, Л. Амиота (1960 г.), которые показали, что между 1 и 1958 г. н.э. динамика численности населения может быть «гиперболическим законом», описанным уравнением:

$$N = \frac{C}{t_0 - t'},$$

где N – численность населения в момент времени t;

C и t_0 – константы.

Результаты прогнозов оказались в пределах допустимой погрешности.

М. Кремер прослеживает эту тенденцию с 1 млн. лет до н.э., а С.П. Капица – с 4 млн. лет до н.э. Дайте оценку параметров модели Ферстера по данным, приведенным в таблице 6.10.

Таблица 6.10 - Численность населения мира
(http://ru.wikipedia.org/wiki/Численность населения мира)

Год	Численность населения, тыс. чел.
4000 до н. э.	7 000
1000 до н. э.	50 000
500 до н. э.	100 000
1 н. э.	300 000
1000	400 000
1750	791 000
1800	978 000
1850	1 262 000
1900	1 650 000
1950	2 518 629
1955	2 755 823
1960	3 021 475
1965	3 334 874
1970	3 692 492
1975	4 068 109
1980	4 434 682
1985	4 830 979
1990	5 263 593
1995	5 674 380
2000	6 070 581
2005	6 343 628
2010	6 500 000
2011	7 000 000

Задача 6.11.

В 2001 г. А. Мэддисон показал, что $G(t)$ – мировой ВВП (в млрд. долл. 1990г. в паритетах покупательной способности) за 1-1973 гг. аппроксимируется квадратично-гиперболическим уравнением вида:

$$G(t) = \frac{C}{(t_0 - t)^2},$$

где C и t_0 – константы.

По данным о величине мирового *ВВП* на душу населения, в ценах 1990 г, *$*[1] найти параметры указанной модели.

Таблица 6.11 – Величина мирового *ВВП* в ценах 1990 г, *$*

Год	Мировой ВВП на душу населения	Год	Мировой *ВВП* на душу населения
1820	704	1971	3 662
1870	876	1972	3 763
1913	1 531	1973	3 941
1940	2 171	1974	3 953
1950	2 024	1975	3 944
1951	2 108	1976	4 063
1952	2 166	1977	4 155
1953	2 234	1978	4 267
1954	2 265	1979	4 341
1955	2 368	1980	4 348
1956	2 433	1981	4 361
1957	2 475	1982	4 336
1958	2 503	1983	4 377
1959	2 569	1984	4 505
1960	2 665	1985	4 582
1961	2 721	1986	4 666
1962	2 801	1987	4 766
1963	2 858	1988	4 887
1964	3 007	1989	4 961
1965	3 104	1990	4 978
1966	3 211	1991	4 971
1967	3 266	1992	4 996
1968	3 376	1993	5 034
1969	3 488	1994	5 133
1970	3 593	1995	5 273

[1] http://www.ggdc.net/maddison/maddison-project/home.htm

Продолжение таблицы 6.11

Год	Мировой ВВП на душу населения	Год	Мировой ВВП на душу населения
1996	5 369	2004	6 546
1997	5 504	2005	6 757
1998	5 527	2006	7 019
1999	5 654	2007	7 234
2000	5 856	2008	7 346
2001	5 952	2009	7 220
2002	6 085	2010	7 224
2003	6 301	2011	7229

Задача 6.12.

Закон Мура. «*Закон Мура — основной лейтмотив нашей деятельности в области конвергенции вычислительных и коммуникационных возможностей*», — заявил глава корпорации *Intel Крейг* Барретт, открывая весенний (2012 г.) Форум *Intel* для разработчиков, Гордон Мур сумел предугадать темпы развития микроэлектроники на несколько десятилетий вперед и предсказать, что количество транзисторов на чипе ежегодно будет удваиваться, потом этот период (удвоения увеличился до двух лет, а в настоящее время – полтора года). Процесс идет за счет совершенствования технологии производства чипов, размеры которых в 2012 г. достигли отметки 22 нанометров и к 2017 г. достигнут 7 нанометров. Считается, что в ближайшее время – при размерах чипов меньше 10 нанометров наступает теоретический предел, при котором квантовые эффекты делают дальнейшее уменьшение транзистора невозможным. Электроны начнут просачиваться сквозь слой толщиной в несколько атомов – явление, в физике называющееся «туннелированием»,что возможно положит начало практической реализации «квантовых» компьютеров.

Таблица 6.12 – Число транзисторов на чипе

Год выпуска	Микропроцессор	Число транзисторов, тыс. шт., (y)
1971	Intel 4004	2,300
1972	Intel 8008	2,500
1974	Intel 8080	5,000
1976	Zilog Z80	8,500
1978	Intel 8086	29,000
1982	Intel 286	120,000
1985	Intel 386™ processor	275,000
1989	Intel 486™ processor	1180
1993	Intel® Pentium® processor	3100
1997	Intel® Pentium® II processor	7500
1999	Intel® Pentium® III processor	24000
2000	Intel® Pentium® 4 processor	42000
2002	Intel® Itanium® processor	220000
2003	Intel® Itanium® 2 processor	410000
2003	AMD K8	105900

Продолжение таблицы 6.12

Год выпуска	Микропроцессор	Число транзисторов, тыс. шт., (y)
2006	Core 2 Duo	291000
2007	AMD K10	463000
2008	AMD K10	758000
2008	Core i7 (Quad)	731000
2009	Six-CoreOpteron 2400	904000
2010	16-Core SPARC T3	1000000
2010	Six-Core Core i7 (Gulftown)	1170000
2010	8-core POWER7	1200000
2010	Quad-core z196	1400000
2011	Six-Core Core i7 (Sandy Bridge-E)	2270000
2012	GeForce GTX670	3540000
2013	GeForce Titan	7100000
2014 прогноз		14000000

Построить зависимость вида $Ln(y)=C + kt$,
где t – порядковый номер года.

Дать оценку изменения k на разных временных промежутках.

7. Анализ взаимосвязи временных рядов

При изучении развития явления во времени часто возникает необходимость оценить *степень взаимосвязи* в изменениях уровней 2-х или более рядов динамики различного содержания, но связанных между собой. Предварительный этап такого анализа состоит в выявлении структуры изучаемых временных рядов, т.к. каждый уровень временного ряда содержит три основных компоненты: тенденцию, циклические или сезонные колебания и случайную компоненту. Наличие циклической или сезонной компоненты приводит к завышению истинных показателей силы и тесноты связи, если оба ряда содержат колебания одинаковой периодичности, либо к снижению этих показателей, если сезонные или циклические колебания содержит только один из рядов или в случае, если периодичность колебаний рассматриваемых рядов разная. Поэтому перед проведением исследования взаимосвязи необходимо устранить сезонную компоненту из уровней каждого ряда в соответствии с методикой, рассмотренной на примерах построения аддитивной или мультипликативной моделей временного ряда. Дальнейший анализ проводится, считая, что ряды не содержат периодических колебаний.

Пусть необходимо изучить зависимость между рядами *y и x*. Для количественной характеристики зависимости между ними используется линейный коэффициент корреляции. Если изучаемые ряды зависят от времени или имеют тенденцию, то коэффициент корреляции по абсолютной величине будет высоким, однако это не значит, что *x* причина *y* или наоборот. Для получения коэффициентов корреляции, характеризующих причинно-следственную связь между изучаемыми рядами, необходимо избавиться от ложной корреляции, вызываемой наличием тенденции в каждом ряде. Для этого используют несколько методов, сущность которых заключается в том, чтобы устранить или зафиксировать воздействие фактора времени на формирование уровней ряда. Основные методы исключения тенденции делятся на две группы:

- методы, основанные на преобразовании уровней исходного ряда в новые переменные, не содержащие тенденции. Эти методы предполагают непосредственное устранение трендовой компоненты из каждого уровня временного ряда (метод последовательных разностей и метод отклонения от тренда);
- методы, основанные на изучении взаимосвязи исходных уровней временного ряда при элиминировании (исключении или удалении) воздействия фактора времени на зависимую или независимую переменные (метод включения в модель регрессии по временным рядам фактора времени).

Метод отклонения от тренда.

Пусть имеются два временных рядаx_tиy_t, каждый из которых содержит трендовую компонентуT и случайную компоненту ε. Проведение аналитического выравнивания по каждому из этих рядов позволяет найти параметры соответствующих уравнений трендов и определить расчетные по тренду уровни $\hat{x}_t$ и $\hat{y}_t$ соответственно. Эти расчетные значения можно принять за оценку трен-

довой компонентыT каждого ряда. Поэтому, влияние тенденции можно устранить путем вычитания расчетных значений уровней ряда из фактических

$$(\Delta_x = x_t - \hat{x}_t \; и \; \Delta_y = y_t - \hat{y}_t). \qquad (7.1)$$

Дальнейший анализ взаимосвязи рядов проводится с использованием не исходных уровней, а отклонений от тренда $\Delta_x = x_t - \hat{x}_t \; и \; \Delta_y = y_t - \hat{y}_t$ при условии, что последние не содержат тенденции (это проверяется путем расчета коэффициента автокорреляции первого порядка для отклонений от тренда каждого временного ряда). Коэффициент корреляции по отклонениям от тренда $r_{\Delta x \Delta y}$ характеризует направление и тесноту взаимосвязи между временными рядами x_t и y_t. Интерпретация параметров этой модели затруднена, но ее можно использовать для прогнозирования.

Метод последовательных разностей.

Если временной ряд содержит ярко выраженную линейную тенденцию, ее можно устранить путем замены исходных уровней ряда цепными абсолютными приростами (первыми разностями).Пусть$y_t = \hat{y}_t + \varepsilon_t$,

где ε_t – случайная ошибка;

$$\hat{y}_t = a + bt \, . \qquad (7.2)$$

Тогда $\Delta_t = y_t - y_{t-1} = a + bt + \varepsilon_t - (a + b(t-1) + \varepsilon_{t-1}) = b + (\varepsilon_t - \varepsilon_{t-1})$. (7.3)

Коэффициент b – константа, не зависящая от времени. При наличии сильной линейной тенденции остатки ε_t достаточно малы и носят случайный характер. Поэтому первые разности уровней ряда Δ_t не зависят от переменной времени, их можно использовать для дальнейшего анализа.

Если временной ряд содержит тенденцию в форме параболы второго порядка, то для ее устранения используют вторые разности.

Если временной ряд содержит тенденции, соответствующие экспоненциальному или степенному трендам, метод последовательных разностей применятся не к исходным уровням ряда, а к их логарифмам. Временные ряды последовательных разностей проверяются на автокорреляцию и если они не содержат тенденцию, рассчитав коэффициент корреляции меду этими рядами, мы получим направление и тесноту связи. Параметры уравнения регрессии последовательных разностей ($\hat{\Delta}_t y \; и \; \hat{\Delta}_t x$) легко поддаются интерпретации.

Однако, при всей своей простоте этот метод имеет два существенных недостатка. Во-первых, его применение связано с сокращением числа пар наблюдений, по которым строится уравнение регрессии и, следовательно,с потерей числа степеней свободы. Во-вторых, использование вместо исходных уровней ряда их приростов или ускорений (экспоненциальный или степенной тренд) приводит к потере информации, содержащейся в исходных данных.

Метод включения в модель регрессии фактора времени.

В корреляционно-регрессионном анализе устранить воздействие какого-либо фактора можно, зафиксировав воздействие этого фактора на результат и другие включенные в модель факторы. Этот прием широко используется в анализе временных рядов, когда тенденция фиксируется через включение фактора времени в модель в качестве независимой переменной.

Модель вида $y_t = a + b_1 x_t + b_2 t + \varepsilon_t$ относится к группе моделей, включающих фактор времени, т.е. число независимых переменных в такой модели может быть больше единицы (в данном случае каждое значение исходного ряда зависит от конкретного времени). Преимущество данной модели в том, что она позволяет учесть всю информацию исходных данных, поскольку значения y_t *и* x_t есть уровни исходных временных рядов. Параметры *a и b*определяются обычным МНК.

Автокорреляция в остатках.

При построении по двум временным рядам x_t, y_t уравнения парной линейной регрессии вида $y_t = a + bx_t + \varepsilon_t$ влияние фактора времени, не учтенного непосредственно в модели, выражается корреляционной зависимостью между значениями остатков $\acute{\varepsilon}_t$ за текущий и предыдущий моменты времени. Такая корреляционная зависимость называется «автокорреляцией в остатках». Автокорреляция в остатках – это нарушение одной из основных предпосылок МНК – 0 случайности остатков, полученных по уравнению регрессии и может быть вызвана несколькими причинами:

- наличием ошибок измерения в значениях результативного признака исходных данных;
- формулировкой модели: модель может не включать фактор, оказывающий существенное воздействие на результат, влияние которой отражается в остатках. Часто этим фактором является фактор времени. Либо модель не учитывает несколько второстепенных факторов, совокупное влияние которых на результат существенно из-за совпадений тенденций их изменения или фаз циклических колебаний. Эту ситуацию не следует путать с неправильной спецификацией функциональной формы модели.

Существуют два наиболее распространенных метода определения автокорреляции остатков. Первый метод – это построение графика зависимости остатков от времени и визуальное определение наличия или отсутствия автокорреляции.

Второй метод – использование критерия Дарбина-Уотсона и расчет величины отношения суммы квадратов разностей последовательных значений остатков к остаточной сумме квадратов по модели регрессии:

$$d = \frac{\sum_{t=2}^{n} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n} \varepsilon_t^2} \qquad (7.6)$$

Между критерием Дарбина–Уотсона и коэффициентом автокорреляции остатков первого порядка существует следующее соотношение: $d \approx 2(1 - r_1^{\varepsilon})$, где

$$r_1^{\varepsilon} = \frac{\sum_{t=2}^{n}(\varepsilon_t - \bar{\varepsilon}_1)(\varepsilon_{t-1} - \bar{\varepsilon}_2)}{\sqrt{\sum_{t=2}^{n}(\varepsilon_t - \bar{\varepsilon}_1)^2 \sum_{t=2}^{n}(\varepsilon_{t-1} - \bar{\varepsilon}_2)^2}}, \quad \bar{\varepsilon}_1 = \frac{\sum_{t=2}^{n}\varepsilon_t}{n-1}; \quad \bar{\varepsilon}_2 = \frac{\sum_{t=2}^{n}\varepsilon_{t-1}}{n-1} \tag{7.7}$$

Таким образом, если в остатках существует полная положительная автокорреляция и $r_1^{\varepsilon} = 1$, то $d = 0$. Если в остатках полная отрицательная автокорреляция, то $r_1^{\varepsilon} = -1$ и $d = 4$. Если автокорреляция остатков отсутствует, то $r_1^{\varepsilon} = 0$ и $d = 2$,следовательно $0 \le d \le 4$.

Алгоритм выявления автокорреляции остатков на основе критерия Дарбина-Уотсона следующий. Выдвигается гипотеза H_0 об отсутствии автокорреляции остатков. Альтернативные гипотезы H_1 и H_1^* состоят соответственно в наличии положительной или отрицательной автокорреляции в остатках. Далее по специальным таблицам определяются критические значения критерия Дарбина-Уотсона d_L *и* d_U для заданного числа наблюдений *n*, числа независимых переменных модели *k* и уровня значимости α. По этим значениям числовой промежуток [0;4] разбивают на пять отрезков. Принятие или отклонение каждой из гипотез с вероятностью α–1 рассмотрим на схеме:

Есть положительная автокорреляция остатков. H_0 отклоняется. Принимается H_1	Зона неопределенности	Нет оснований отклонять H_0 (автокорреляция остатков отсутствует)	Зона неопределенности	Есть отрицательная автокорреляция остатков. H_0 отклоняется. Принимается H_1^*
0	d_L d_U	2	$4-d_U$ $4-d_L$	

Если фактическое значение критерия Дарбина-Уотноса попадает в зону неопределенности, то на практике предполагают существование автокорреляции остатков и отклоняют гипотезу H_0.

Существуют ограничения на применение этого критерия:

– Он неприменим к моделям авторегрессии, где в качестве независимых переменных лаговые значения результативного признака.

– Этот критерий используется только для выявления автокорреляции остатков первого порядка.

– Критерий Дарбина-Уотсона дает достоверные результаты только для больших выборок.

–

Оценивание параметров уравнения регрессии при наличии автокорреляции в остатках.

Если остатки по исходному уравнению регрессии содержат автокорреляцию, то для оценки параметров уравнения используют обобщенный МНК. Обобщенный МНК аналогичен методу последовательных разностей. Однако мы вычитаем из y_t (или x_t) не все значения предыдущего уровня y_{t-1}(или x_{t-1}), а некоторую его долю - $r_1^{\varepsilon} \cdot y_{t-1}$ *или* $r_1^{\varepsilon} \cdot x_{t-1}$:

$$y_t^{'} = y_t - r_1^{\varepsilon} \cdot y_{t-1}$$

$$x_t^{'} = x_t - r_1^{\varepsilon} \cdot x_{t-1}$$

Если $r_1^{\varepsilon} = 1$, данный метод есть метод первых разностей. Если $r_1^{\varepsilon} \to 0$, применение метода первых разностей также вполне обоснованно. Если $r_1^{\varepsilon} = -1$- в остатках наблюдается полная отрицательная автокорреляция, то мы имеем модель:

$(y_t + y_{t-1})/2 = a + b(x_t + x_{t-1})/2 + u_t/2$, *где* $u_t = \varepsilon_t - r_1^{\varepsilon} \cdot \varepsilon_{t-1}$ *случайная ошибка*

В сущности, в данной модели мы определяем средние за два периода уровни каждого ряда, а затем по полученным усредненным уровням обычным МНК рассчитываем параметры *a* и *b*. Данная модель называется *моделью регрессии по скользящим средним.*

Задача 7.1.

Изучается динамика урожайности озимых зерновых культур и цен реализации зерна (без кукурузы) в сельскохозяйственных организациях Краснодарского края за 2000-2012 годы:

Цена за 1 ц, руб *Yt*	179	180	155	254	254	239	324	528	496	437	511	559	1006
Урожайность с 1 га, ц *Xt*	38,2	42,7	45,0	29,7	41,9	44,6	42,6	45,1	54,7	47,3	51,2	55,3	39,6

1. Постройте уравнения линейного тренда по каждой переменной и дайте интерпретацию их параметров

2. Определите коэффициенты корреляции и детерминации по линейным трендам.

3. Постройте уравнения регрессии и оцените тесноту и силу связи между уровнями временных рядов, между первыми разностями

4. Постройте уравнения множественной регрессии влияния урожайности на уровень цены зерна с включением фактора времени и урожайности предыдущего года. Сравните полученные модели и выберете лучшее из них.

8. Анализ панельных данных

Панельные данные представляют собой *двумерные массивы, одна из размерностей которых – время.* Обследование большого числа объектов на протяжении некоторого периода времени.

Сегодня исследователи видят следующие преимущества панельных данных:

а) увеличивается число наблюдений, что повышает число степеней свободы и снижает коллинеарность между объясняющими переменными, что приводит к улучшению эффективности оценок;

б) появляется возможность изучать экономические процессы и в пространстве, и во времени;

в) предотвращается смещение агрегированности (во времени – изменение усредненного «репрезентативного» объекта; в пространстве – не учитываются ненаблюдаемые индивидуальные характеристики объекта);

г) появляется возможность моделировать индивидуальную динамику изменения объектов во времени.

Традиционный источник панельных данных – результаты маркетинговых, социологических обследований, репрезентативных опросов индивидуумов, домохозяйств, предприятий.

Панельная регрессия.

Традиционно в эконометрике рассматриваются одномерные данные в пространстве (i – люди, фирмы, районы, регионы и т. д. в один момент времени) или во времени (t – наблюдения, упорядоченные во времени). Рассмотрение экономических данных с двумя измерениями (i, t) приводит к одномоментным данным за разные временные периоды двух типов:

- *независимое объединение* (разные единицы, независимые выборки),
- *панельные данные* (одни и те же единицы в динамике).

Использование панельных данных ценится в экономических исследованиях, так как их правильный анализ позволяет избавиться от ненаблюдаемых индивидуальных особенностей объектов. Схема панельных данных представлена на рисунке 8.1.

Если данные следующего периода не зависят от других периодов (попадают туда случайно), то:

$$(y_{it}, Z_{it}), \quad i = \overline{1, N_t}, \quad t = \overline{1, T}, \tag{8.1}$$

где y_{it}– объясняемая переменная;

Z_{it} – объясняющие факторы.

Когда для всех объектов наблюдения имеются данные в каждый момент времени, то панель является *сбалансированной*. Если для некоторых i или t наблюдения отсутствуют, то панель считается *несбалансированной*. Если в различные моменты времени наблюдаются различные объекты, то в этом случае имеем дело с *псевдо панелью*.

$$y_{it} = Z_{it}\alpha + \varepsilon_{it}, \qquad (8.2)$$

где ε_{it} – независимы, имеют нулевое математическое ожидание, постоянную дисперсию и некоррелированы с факторами Z_{it}.

Модель (8.2) называется *объединенной моделью регрессии* (*pooled regression model*), коэффициенты которой находятся с помощью *МНК*.

time	*Y*	X_1	X_2	...	X_i	...	X_m
1	$y1(t_1)$	$x_{11}(t_1)$	$x_{12}(t_1)$		$x_{1i}(t_1)$		$x_{1m}(t_1)$
1	$y1(t_2)$	$x_{11}(t_2)$	$x_{12}(t_2)$		$x_{1i}(t_2)$		$x_{1m}(t_2)$
1	$y1(t_3)$	$x_{11}(t_3)$	$x_{12}(t_3)$		$x_{1i}(t_3)$		$x_{1m}(t_3)$
.							
1	$y_1(t_T)$	$x_{11}(tx)$	$x_{12}(t_T)$		$x_{1i}(t_T)$		$x_{1m}(t_T)$
2	$y_2(t_1)$	$x_{21}(tl)$	$x_{22}(t_1)$		$x_{2i}(t_1)$		$x_{2m}(t_1)$
2	$y_2(t_2)$	$x_{21}(t_2)$	$x_{22}(t_2)$		$x_{2i}(t_2)$		$x_{2m}(t_2)$
2	$y_2(t_3)$	$x_{21}(t_3)$	$x_{22}(t_3)$		$x_{2i}(t3)$		$x_{2m}(t_3)$
.							
2	$y_2(t_T)$	$x_{21}(t_T)$	$x_{22}(t_T)$		$x_{2i}(t_T)$		$x_{2m}(t_T)$
3	$y_3(t_1)$	$x_{31}(t_1)$	$x_{32}(t_1)$		$x_{3i}(t_1)$		$x_{3m}(t_1)$
3	$y_3(t_2)$	$x_{31}(t_2)$	$x_{32}(t_2)$		$x_{3i}(t_2)$		$x_{3m}(t_2)$
.							
k	$y_k(t_1)$	$x_{k1}(t_1)$	$x_{k2}(t_1)$		$x_{ki}(t_1)$		$x_{km}(t_1)$
k	$y_k(t_2)$	$x_{k1}(t_2)$	$x_{k2}(t_2)$		$x_{ki}(t_2)$		$x_{km}(t_2)$
.							
k	$y_k(t_t)$	$x_{k1}(t_t)$	$x_{k2}(t_t)$		$x_{ki}(t_t)$		$x_{km}(t_T)$
.							
k	$y_k(t_{T-1})$	$x_{k1}(t_{T-1})$	$x_{k2}(t_{T-1})$		$x_{ki}(t_{T-l})$		$x_{km}(t_{T-1})$
k	$y_k(t_T)$	$x_{k1}(t_T)$	$x_{k2}(t_T)$		$x_{ki}(t_T)$		$x_{km}(t_T)$
.							
.							
n	$y_n(t_1)$	$x_{n1}(t_1)$	$x_{n2}(t_1)$		$x_{ni}(t_1)$		$x_{nm}(t_1)$
.							
n	$y_n(t_{T-1})$	$x_{n1}(t_{T-1})$	$x_{n2}(t_{T-1})$		$x_{ni}(t_{T-1})$		$x_{nm}(t_T\text{-}l)$
n	$y_n(t_T)$	$x_{n1}(t_T)$	$x_{n2}(t_T)$		$x_{ni}(t_T)$		$x_{nm}(t_T)$

Рисунок 8.1 – Схема панельных данных

Регрессионную модель для панельных данных (*panel data*) можно представить в виде, учитывающем индивидуальные эффекты f_i ($\nu_{it} = f_i + \varepsilon_{it}$, ε_{it} – остаточное возмущение)

$$y_{it} = Z_{it}\alpha + f_i + \varepsilon_{it}. \tag{8.3}$$

Если считается, что эффекты f_i представляют собой N фиксированных неизвестных параметров модели, то (9.3) – *модель с фиксированными эффектами* (*fixed effects model – fe*), считается, что f_i – мешающий параметр, который следует устранить.

Если эффекты f_i – случайные величины, коррелированные с ε_{it}, то (8.3) – модель со случайными эффектами (*random effects model – re*).

Вычислительные аспекты рассмотренных выше моделей излагаются в соответствующей литературе. В практических исследованиях внимание уделяется адекватности полученных моделей. Выбор между рассмотренными моделями может ориентироваться как на содержательные соображения, так и на стандартную технику проверки статистических гипотез:

1) *тест Вальда* (сравнивается объединенная регрессия (7.2) и модель с фиксированными эффектами (9.3) *fe*). Нулевая гипотеза: $f_1 = f_2 = \cdots = f_N$, т. е. истинна модель (9.2). Альтернативная гипотеза – верна модель с фиксированными эффектами – *fe*. Рассчитывается F-статистика по суммам квадратов в двух сравниваемых регрессиях.

2) *тест Бройша-Пагана* позволяет ответить на вопрос – есть ли в данных панельная структура (нужно ли использовать панельные методы). При выборе между обычной регрессией (9.2) и моделью регрессии со случайными эффектами – *re* нулевая гипотеза – верна модель обычной объединенной регрессии, ($H_0: \sigma_f^2 = 0$). Альтернативная гипотеза – верна модель *re*. Статистика множителей Лагранжа:

$$LM = \frac{NT}{2(T-1)}\left(\frac{T^2\sum_{i=1}^{N}(\hat{\varepsilon}_{i\cdot})^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}(\hat{\varepsilon}_{it})^2} - 1\right)^2, \tag{8.4}$$

где $\hat{\varepsilon}$ – остатки обычной регрессии (8.2).

Если нулевая гипотеза верна, то *LM* распределена как χ_1^2.

3) *тест Хаусмана* позволяет выбрать одну из моделей – *fe* или *re*. Нулевая: f_i – случайные эффекты; альтернативная гипотеза: f_i – детерминированные эффекты.

Для повышения адекватности полученных моделей можно предложить использование бутстреп-метода.

Традиционно для отечественных эконометрических исследований подобные исследования проводят на основании данных РМЭЗ – Российского мониторинга экономического положения и здоровья населения, представляющего собой единственное в России представительное панельное обследование семей.

Далее приводится пример анализа панельных данных, полученных в результате наблюдения за деятельностью сельскохозяйственных предприятий Краснодарского края.

8.1 Моделирование процесса оценки ресурсного обеспечения сельскохозяйственных предприятий для поддержки принятия управленческих решений

В 2011 году на кафедре статистики и прикладной математики проводились исследования оценки ресурсного обеспечения сельскохозяйственных предприятий для формирования политики государственной поддержки ресурсного обеспечения предприятий АПК.

Характеристика данных. Источником исходных данных для анализа являлась база данных, основанная на статистической отчетности за 2006 – 2010 гг. 100 сельхозпредприятий 17-и районов Центральной зоны Краснодарского края и данных Федерального бюджетного учреждения «*Кадастровая палата*» по *Краснодарскому краю.* Было отобрано 30 показателей, характеризующих эффективность использования основных производственных ресурсов, которые можно разделить на три группы: трудовые, земельные и основные фонды (капитал).

Трудовые ресурсы:

y1 - стоимость валовой продукции (СВП) в сопоставимых ценах на 1 работника, тыс. руб.;

y2 - СВП в текущих ценах на 1 работника, тыс. руб.;

y3 - СВП на 1 чел.-ч, руб. (*уз*);

y4 - валовой доход на 1 работника, тыс. руб.;

x1 - продолжительность рабочего дня, часов;

x2 - доля затрат на оплату труда в общих затратах, %;

x3 - среднегодовая начисленная заработная плата 1 работника, тыс. руб.;

x4 - удельный вес работников сельского хозяйства в общей численности, %;

x5 - энерговооруженность, л. с.;

x6 - фондовооруженность, тыс. руб.;

x7 - приходится специалистов на 100 постоянных работников, чел.;

x8 - годовая выплаченная заработная плата 1 работника, тыс. руб.;

x9 - оплата за 1 чел.-ч, руб.;

x10 - площадь сельскохозяйственных угодий на 1 работника сельского хозяйства, га;

x11- отработано за год 1 работником, дней;

x12- затратоемкость;

x13- производственные затраты (ПЗ) на 1 га сельскохозяйственных угодий, тыс. руб.;

Земельные ресурсы:

y5 - СВП растениеводства на 1 га пашни в текущих ценах, тыс. руб.;

x14- производственные затраты растениеводства на 100 га пашни, тыс. руб.;

x15 - фондообеспеченность на 1 га, тыс. руб.;

x16 - начисленная заработная плата на 1 га сельскохозяйственных угодий, тыс. руб.;

x17- нагрузка пашни на 1 трактор, га;

x18 - коэффициент использования пашни;

x19 - численность работников сельского хозяйства на 100 га сельскохозяйственных угодий, чел.;

x20 - кадастровая стоимость 1 га сельскохозяйственных земель, тыс. руб.;

Основные фонды:

y6- стоимость валовой продукции в текущих ценах в расчете на 1 га сельскохозяйственных угодий, тыс. руб.;

y7 - СВП на 1 руб. основных производственных фондов, руб.;

x21 – фондоемкость;

x22- удельный вес машин и оборудования в основных производственных фондах, %.

Замечание. Необходимость анализа панельных данных часто иллюстрируется тем, что не учет индивидуальных особенностей наблюдаемых объектов, вызванных влиянием ненаблюдаемых факторов, приводит к неэффективности оценок регрессии.

Анализ в пакете STATA 12

Загрузим исходные данные для анализа, находящиеся в файле *svp.dta* (*File-Open-…-svp.dta*).

Если имеются выбросы или артефакты, то их можно идентифицировать с использованием диаграммы размах Дж. Тьюки (1970 г.), полученной с помощью команды:

. **graph box y6, over (year)**

Из полученной диаграммы размаха (рис. 8.2) для данных стоимости валовой продукции за 2006–2010 гг. в сопоставимых ценах видно, что данные неоднородны и имеются выбросы.

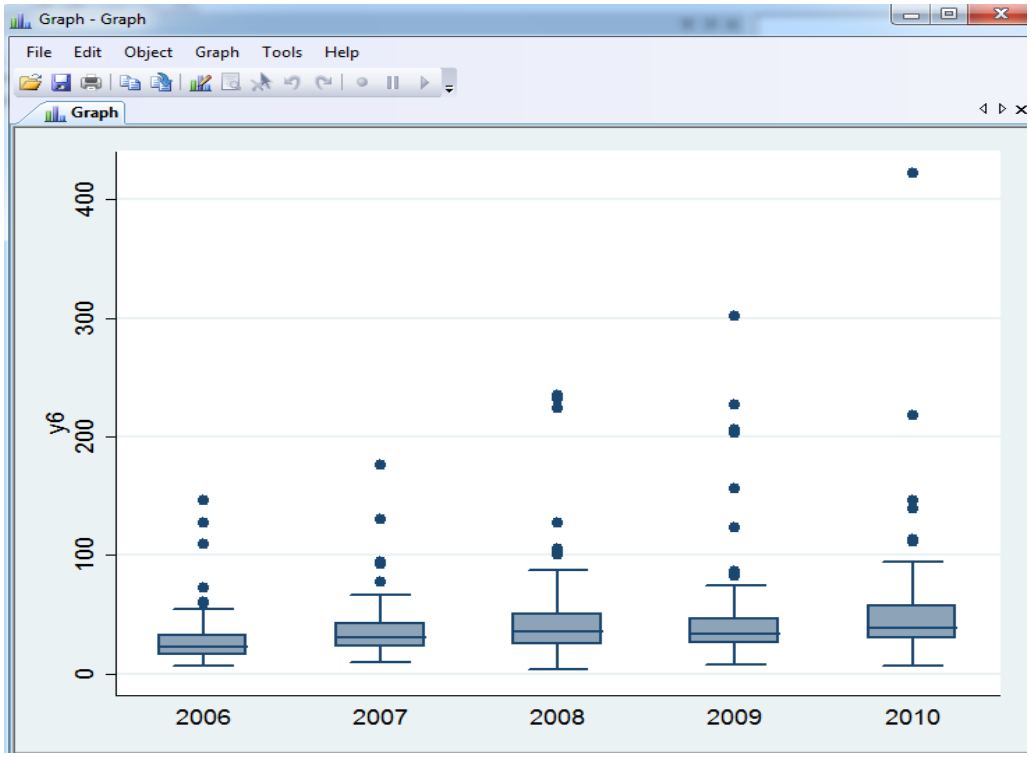


Рисунок 8.2 – Диаграмма размаха для данных, *y6*

Например, если есть значения переменных y6>100; x1>15; x5>150 и т.д., не интерпретирующийся содержательно), то их можно исключить с помощью команды:

.drop if y6>100 // x1>15 // x5>150.

Исследуемая совокупность содержит наблюдения по 100 одним и тем же предприятиям за пять лет, что соответствует идеологии сбора и изучения панельных данных.

Для устранения асимметрии распределения эконометрических величин переходят к их логарифмам, что обычно позволяет в большинстве случаев считать распределение остатков регрессии близким к нормальному.

В эконометрическую модель эффективности использования ресурсов включим логарифмы всех переменных, выбранных после проведения предварительного анализа:

ln y6 – логарифм стоимости валовой продукции в текущих ценах в расчете на 1 га сельскохозяйственных угодий, тыс. руб.,

ln x1 – логарифм продолжительность рабочего дня, часов;

ln x5 - логарифм энерговооруженность, л. с.;

ln x14- логарифм производственных затрат растениеводства в расчете на 100 га пашни, тыс. руб.;

ln x15 - логарифм фондообеспеченности, тыс. руб.;

ln x17 - логарифм нагрузки пашни на 1 трактор, га;

ln x19 - логарифм численности работников на 100 га сельскохозяйственных угодий, чел.;

ln x20 - логарифм кадастровой оценки 1 га земли сельскохозяйственного назначения, тыс. руб.;

ln x22 - логарифм удельного веса машин и оборудования в стоимости основных фондов, %.

Для создания новых переменных можно использовать команду:

. generate имя переменной = выражение

Например:

. generate lnx1=ln(x1)

Кроме того, в модели можно рассматривать фиктивные переменные: *k* – фиктивная переменная, которая принимает значение 0, если рассматриваются данные до экономического кризиса 2008 г. (2006-2009 гг.); 1 – в противном случае (2010 г.).

Построим описательные статистики наших переменных для фиксированного года, например, для 2010 (*year*==2010), используя команду ***if***:

. sum y6 x1 x5 x14 x15 x17 x19 x20 x22 if year ==2010

В результате получим данные, представленные в таблице 7.1.

Таблица 8.1 – Описательные статистики

Variable	Obs	Mean	Std. Dev.	Min	Max
y6	100	50.5184	47.92625	6.320008	422.1921
x1	100	7.428563	.9453052	1	9.26087
x5	100	58.47834	26.01077	5.138614	164.3385
x14	100	42.24655	46.4324	6.062762	421.5163
x15	100	43.99765	41.13608	.9671198	246.1652
x17	100	139.6369	86.46893	4.416667	724.2
x19	100	5.290143	5.325294	.7057117	36.87822
x20	100	109.9902	17.87144	61.215	126.951
x22	100	43.78628	16.39633	5.99789	90.36858

Средняя стоимость валовой продукции на 1 га сельскохозяйственных угодий (*y6*) по совокупности 100 организаций в 2010 г. составила 50,52 тыс. руб. Колебания в среднем составляют 50,52 ±47,93 тыс. руб. согласно средне-

го квадратического отклонения, что говорит о разнородности изучаемой совокупности.

Продолжительность рабочего дня (*x1*) в среднем 7,4 часа, что ниже установленной продолжительности рабочего дня. Согласно среднего квадратического отклонения, изменение колеблется в пределах 7,4 ± 0,9 часа.

Средняя энерговооруженность - 58,5 л.с., при этом имеют место значительные колебания.

Производственные затраты в растениеводстве (*x14*) составили 42,45 тыс. руб. В результате разнородной продукции размер колебаний по среднему квадратическому отклонению превышает среднее значение показателя.

Что касается фондоотдачи, также наблюдаются значительные колебания показателей в результате различного уровня использования сельскохозяйственными организациями основными фондами.

Нагрузка пашни на 1 трактор составила 139,6±86,5 га. Разница в значениях показателей очень велика, так как ряд организаций практически не имея собственной сельскохозяйственной техники, вынуждены прибегать к услугам сторонних организаций.

Трудообеспеченность (*x19*) в среднем по совокупности организаций составила 5 чел. на 1 га.

Кадастровая стоимость сельскохозяйственных земель изменяется незначительно и соответствует в среднем 109,99 тыс. руб. за 1 га.

Доля машин и оборудования по изучаемым организациям центральной зоны Краснодарского края в среднем изменяется от 27,39% до 60,19%.

Рассчитанные коэффициенты вариации на основании среднего квадратического отклонения показывают, что изучаемая совокупность является однородной лишь по показателям продолжительности рабочего дня (*x1*) и кадастровой оценки 1 га земли сельскохозяйственного назначения (*x20*).

Дополнительная опция *detail* выводит характерные квантили (*p*% - ые точки распределения), большие (*Largest*) и малые (*Smallest*) значения вариационного ряда, коэффициенты ассиметрии (*Skewness*) и эксцесса (*Kurtosis*).
Например:

. **sum y6 if year==2010, detail**

Квантили, делят ранжированную совокупность на определенные группы в процентном отношении, пример представлен в таблице 8.2.

Процентиль *(percentiles)* – условное значение ранжированного вариационного ряда, меньше которого находится *p%* данных.

Таблица 8.2 – Описательные порядковые статистики переменной (y6)

y6

	Percentiles	Smallest		
1%	6.505407	6.320008		
5%	18.38227	6.690806		
10%	23.91335	13.03572	Obs	100
25%	29.47702	14.83883	Sum of Wgt.	100
50%	38.75381		Mean	50.5184
		Largest	Std. Dev.	47.92625
75%	58.02803	139.2604		
90%	74.58625	146.1226	Variance	2296.925
95%	112.4422	218.1542	Skewness	5.308774
99%	320.1731	422.1921	Kurtosis	38.82979

Для остальных лет и переменных по годам результаты можно получить аналогично.

Описательные порядковые статистики, построенные в соответствии с таблицей 8.3, позволяют оценить однородность данных и наличие выбросов. *P*% квантили показывают значение, ниже которого находится p% данных. Соответствующие малые значения изменяются достаточно равномерно. Последние процентили (75%, 90%, 95%, 99%) и большие значения указывают на наличие выбросов, что соответствует рисунку 8.1. Коэффициент вариации составляет 94,9% и характеризует совокупность, как разнородную, т.е. имеются достаточно большие различия между признаками.

Таким образом, при анализе децилей видно, что у 10% предприятий стоимость валовой продукции на 1 га сельскохозяйственных угодий не превышает 23,91 тыс. руб., а 90% организаций получают с 1 га сельхозугодий более 23,91 тыс. руб. валовой продукции, выраженной в денежном эквиваленте.

Четвертая часть изучаемого ранжированного ряда (25% организаций) с 1 га получают менее 29,48 тыс. руб., а остальные организации (75%) имеют более, чем 29,48 тыс. руб. стоимости валовой продукции в расчете на 1 га.

Половина предприятий с единицы площади сельскохозяйственных угодий получают меньше, чем 38,75 тыс. руб., а другая половина больше 38,75 тыс. руб. стоимости валовой продукции.

Коэффициент ассиметрии, составляющий 5,309 является достаточно значительным, что говорит о существенной правосторонней ассиметрии вариационного признака.

Таблица 8.3 – Квантили вариационного ряда

p	x_p	x_{min}
1,0%	$x_{1,0\%}$	x_1
5,0%	$x_{5,0\%}$	x_2
10,0%	$x_{10,0\%}$	x_3
50,0%	-	-
75,0%	$x_{75,0\%}$	x_{n-3}
90,0%	$x_{90,0\%}$	x_{n-2}
95,0%	$x_{95,0\%}$	x_{n-1}
99,0%	$x_{99,0\%}$	x_n

Сквозная регрессия нами рассматривается по всем годам и не учитывает панельную структуру данных.

Для построения модели линейной регрессии введем команду

. reg lny6 lnx1 lnx5 lnx14 lnx17 lnx19 lnx20 lnx22 k

Оценка ее параметров в пакете *STATA* 12 с использованием обычного метода наименьших квадратов приведена в таблице 8.4.

Следует отметить, что наибольшее влияние на результативный признак оказывают производственные затраты растениеводства на 100 га пашни, кадастровая оценка земли, а также удельный вес машин и оборудования (активной части) в стоимости основных фондов.

Таблица 8.4 – Результаты оценки сквозной регрессии

Source	SS	df	MS		Number of obs =	500
					F(8, 491) =	935.92
Model	182.370872	8	22.796359		Prob > F =	0.0000
Residual	11.9594094	491	.024357249		R-squared =	0.9385
					Adj R-squared =	0.9375
Total	194.330282	499	.389439443		Root MSE =	.15607

lny6	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnx1	-.053463	.048997	-1.09	0.276	-.1497327	.0428066
lnx5	-.0080585	.0122328	-0.66	0.510	-.0320936	.0159767
lnx14	.9484192	.0182323	52.02	0.000	.9125962	.9842422
lnx17	-.0142536	.0181556	-0.79	0.433	-.0499259	.0214187
lnx19	.0059945	.0218413	0.27	0.784	-.0369194	.0489083
lnx20	.0802483	.0387762	2.07	0.039	.0040605	.156436
lnx22	.0401685	.0165823	2.42	0.016	.0075874	.0727496
k	.0432879	.018881	2.29	0.022	.0061904	.0803855
_cons	-4.352946	.2856713	-15.24	0.000	-4.914235	-3.791656

После отбрасывания незначимых переменных в результате применения команды

. **reg lny6 lnx14 lnx20 lnx22 k**

получим результаты, представленные в таблице 8.5.

Таблица 8.5 – Результаты анализа сквозной регрессии после отбрасывания незначимых переменных

Source	SS	df	MS	Number of obs =	500
				F(4, 495) =	1875.56
Model	182.301922	4	45.5754805	Prob > F =	0.0000
Residual	12.02836	495	.024299717	R-squared =	0.9381
				Adj R-squared =	0.9376
Total	194.330282	499	.389439443	Root MSE =	.15588

lny6	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnx14	.9623459	.0116031	82.94	0.000	.9395485	.9851433
lnx20	.0849971	.0376108	2.26	0.024	.0111005	.1588936
lnx22	.0350452	.0160228	2.19	0.029	.003564	.0665264
k	.0387274	.0176925	2.19	0.029	.0039658	.0734891
_cons	-4.662623	.2116917	-22.03	0.000	-5.078548	-4.246698

Построение сквозной регрессии без учета статистически незначимых факторов позволило получить результаты, представленные в таблице 8.6.

Таблица 8.6 – Результаты эконометрической оценки модели сквозной регрессии с исключением незначимых факторов (зависимая переменная: ln – стоимость валовой продукции в текущих ценах в расчете на 1 га сельскохозяйственных угодий, тыс. руб.)

Переменная	Коэффициент	Стандартная ошибка
ln x14	0,9623***	0,0116
ln x20	0,0850**	0,0376
ln x22	0,0350**	0,0160
k	0,0387**	0,1769
cons	-4,6626***	0,2117
Скорректированный коэффициент детерминации	0,9376	
***, **, * - уровни значимости 1%, 5% и 10 % соответственно		

На основании данных таблицы 8.6 можно сделать вывод о том, что с увеличением производственных затрат растениеводства в расчете на 100 га

пашни (*x14*) на 1% наблюдается рост на 0,96% стоимости валово1 продукции на 1 га сельхозугодий.

Повышение кадастровой оценки 1 га сельхозугодий (*x20*) на 1% приводит к увеличению результата на 0,08%.

Рост доли активной части основных фондов (*x22*) на 1% обуславливает повышение стоимости валовой продукции на 1 га в размере 0,03%.

Незначительная величина стандартной ошибки всех факторов говорит о достаточной достоверности данных.

1. *Регрессия «between»* позволяет оценить исходную модель при усредненных по времени значениях переменных.

Для анализа панельных данных в пакете *Stata* используется префикс **xt**, обозначающий наличие как структурной, так и временной компоненты. Зададим временную компоненту *t*:

. **tis year**

Зададим пространственную компоненту *i:*

. **iis id**

Для построения «between»-регрессии используем команду

. xtreg lny6 lnx1 lnx5 lnx14 lnx15 lnx17 lnx19 lnx20 lnx22 k, be

Фиктивная переменная *k* была отброшена из-за проблем мультиколлинеарности: *note: k omitted because of collinearity*

Таблица 8.7 – Результаты оценки «*between*»-регрессии

```
Between regression (regression on group means)  Number of obs      =       500
Group variable: id                              Number of groups   =       100

R-sq:  within  = 0.8467                         Obs per group: min =         5
       between = 0.9741                                        avg =       5.0
       overall = 0.9330                                        max =         5

                                                F(8,91)            =    428.20
sd(u_i + avg(e_i.))=  .0883137                  Prob > F           =    0.0000
```

lny6	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnx1	-.1807155	.1079508	-1.67	0.098	-.3951465	.0337154
lnx5	-.0113948	.0220098	-0.52	0.606	-.0551146	.032325
lnx14	.841477	.0397278	21.18	0.000	.7625625	.9203914
lnx15	.0311222	.0220211	1.41	0.161	-.0126199	.0748644
lnx17	.0084319	.0288774	0.29	0.771	-.0489295	.0657933
lnx19	.0904245	.0390451	2.32	0.023	.0128661	.1679829
lnx20	.0486276	.0512154	0.95	0.345	-.0531055	.1503608
lnx22	.0796499	.0308388	2.58	0.011	.0183923	.1409075
k	0	(omitted)				
_cons	-3.563538	.4758456	-7.49	0.000	-4.508747	-2.618329

Достаточно большое значение *R-sq between* = 0,9741, характеризующее качество подгонки регрессии, показывает, что изменение средних по време-

ни показателей оказывает более существенное влияние на каждую переменную, нежели временные колебания этих показателей, относительно средних.

sd(u_i + avg(e_i.)) = .0883137 - стандартное отклонение случайной составляющей для «*between*»-регрессии.

. xtreg lny6 lnx14 lnx19 lnx22 k, be

Результаты представим в таблице 8.8.

Таблица 8.8 – Результаты оценки «*between*»-регрессии после отбрасывания незначимых факторов

```
Between regression (regression on group means)  Number of obs      =       500
Group variable: id                              Number of groups   =       100

R-sq:  within  = 0.8512                         Obs per group: min =         5
       between = 0.9724                                        avg =       5.0
       overall = 0.9328                                        max =         5

                                                F(3,96)            =   1125.92
sd(u_i + avg(e_i.))=  .0888568                  Prob > F           =    0.0000
```

lny6	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnx14	.8618851	.0347155	24.83	0.000	.7929754	.9307948
lnx19	.0995669	.0338419	2.94	0.004	.0323913	.1667425
lnx22	.0620104	.029513	2.10	0.038	.0034276	.1205932
k	0	(omitted)				
_cons	-3.714428	.2356924	-15.76	0.000	-4.182274	-3.246582

Вид модели при переходе от сквозной к «*betwen*»-регрессии сильно не изменился, но логарифм кадастровой оценки 1 га земли сельскохозяйственного назначения заменился на логарифм численности работников на 100 га сельскохозяйственных угодий.

Это объясняется тем, что кадастровая стоимость угодий является незначимой величиной и это обоснованно, так как цена 1 га зависит от ряда факторов, одним из которых является месторасположение участка. Соответственно, чем ближе сельскохозяйственные угодья к городам и районным центрам, тем выше кадастровая оценка земли. Хотя этот фактор не улучшает качество почвы и не способствует росту получаемой продукции, а, следовательно, и повышению стоимости валовой продукции с 1 га сельскохозяйственных угодий. Зависимость между переменными значительно не изменилась (таблица 8.9).

Таблица 8.9 – Результаты эконометрической оценки модели «*between*»-регрессии (зависимая переменная: *ln* (стоимость валовой продукции в текущих ценах в расчете на 1 га сельскохозяйственных угодий, тыс. руб.)

Переменная	Коэффициент	Стандартная ошибка
ln X14	0,8619***	0,0347
ln X19	0,0996***	0,0338
ln X22	0,0620**	0,0295
cons	-3,7144***	0,2357
Коэффициент детерминации «*between*»-модели	0,9724	
***, **, * - уровни значимости 1%, 5% и 10 % соответственно		

«*between*»-регрессия показывает, что рост на 1% производственных затрат в растениеводстве в расчете на 100 га пашни *(x14)* приводит к повышению на 0,86% стоимости валовой продукции на 1 га земель сельхозназначения.

Увеличение на 1% численности работников в расчете на 100 га сельхозугодий (*x19*) и доли машин и оборудования в стоимости основных производственных фондов *(x22)* ведет к повышению результативного признака на 0,09% и 0,06% соответственно.

2. *Регрессия «within»* – позволяет оценить коэффициенты регрессионной модели с детерминированными индивидуальными эффектами (*fe*).

Используя команду:

. xtreg lny6 lnx1 lnx5 lnx15 lnx17 lnx19 lnx20 lnx22 k, fe

получим следующие результаты, представленные в таблице 8.10.

На основании таблицы 8.10 проводим отбор более значимых факторов модели. Незначимыми оказались признаки, не изменяющиеся во времени: *lnx1, lnx15, lnx17, lnx19, lnx20.*

Таблица 8.10 – Результаты оценки «*within*» - регрессии

```
Fixed-effects (within) regression               Number of obs      =       500
Group variable: id                              Number of groups   =       100

R-sq:  within  = 0.8638                         Obs per group: min =         5
       between = 0.8773                                        avg =       5.0
       overall = 0.8709                                        max =         5

                                                F(9,391)           =    275.51
corr(u_i, Xb)  = -0.2776                        Prob > F           =    0.0000
```

lny6	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnx1	.0091761	.0547184	0.17	0.867	-.098403	.1167552
lnx5	-.0290018	.0173441	-1.67	0.095	-.0631012	.0050976
lnx14	1.000148	.0280756	35.62	0.000	.94495	1.055346
lnx15	.0151036	.0224438	0.67	0.501	-.029022	.0592293
lnx17	-.0530778	.0305275	-1.74	0.083	-.1130964	.0069407
lnx19	-.0406407	.0370205	-1.10	0.273	-.1134248	.0321434
lnx20	-.8215857	1.994855	-0.41	0.681	-4.74357	3.100399
lnx22	.048975	.0231687	2.11	0.035	.0034243	.0945258
k	.0258174	.0185125	1.39	0.164	-.0105791	.0622138
_cons	-.4065846	9.358599	-0.04	0.965	-18.80605	17.99289
sigma_u	.19456044					
sigma_e	.14115494					
rho	.65515334	(fraction of variance due to u_i)				

```
F test that all u_i=0:     F(99, 391) =     2.10                 Prob > F = 0.0000
```

После отбрасывания незначимых переменных, используя команду:

. xtreg lny6 lnx14 lnx22 k, fe

получим результаты, представленные в таблице 8.11.

Достаточно большое значение *R-sq within* = 0,8619, характеризующее качество подгонки регрессии, показывает, что в рамках нашей модели межиндивидуальные различия проявляются сильнее, чем динамические, что говорит о необходимости учета индивидуальных эффектов и позволяет выдвинуть гипотезу против модели сквозного оценивания.

Таблица 8.11 - Результаты эконометрической оценки модели *«within»* - регрессии

```
Fixed-effects (within) regression               Number of obs      =       500
Group variable: id                              Number of groups   =       100

R-sq:  within  = 0.8619                         Obs per group: min =         5
       between = 0.9699                                        avg =       5.0
       overall = 0.9374                                        max =         5

                                                F(3,397)           =    825.82
corr(u_i, Xb)  = -0.2923                        Prob > F           =    0.0000
```

lny6	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnx14	1.012196	.0216493	46.75	0.000	.9696348	1.054758
lnx22	.0350032	.0217855	1.61	0.109	-.0078262	.0778326
k	.0274314	.0166349	1.65	0.100	-.0052722	.060135
_cons	-4.657772	.2017774	-23.08	0.000	-5.054458	-4.261086
sigma_u	.09705762					
sigma_e	.14105879					
rho	.32131308	(fraction of variance due to u_i)				

```
F test that all u_i=0:     F(99, 397) =     2.16                 Prob > F = 0.0000
```

Полученное противоречие между моделями ниже разрешается с использованием идеологии проверки статистических гипотез.

Таблица 8.12 – Результаты эконометрической оценки модели «within»-регрессии (зависимая переменная: *ln* (стоимость валовой продукции в текущих ценах в расчете на 1 га сельскохозяйственных угодий, тыс. руб.)

Переменная	Коэффициент	Стандартная ошибка
ln X_{14}	1,0122 ***	0,0216
ln X_{22}	0,0350*	0,0218
k	0,0274*	0,0166
cons	-4,6578***	0,2018
Коэффициент детерминации «*within*»-модели	0,8619	
***, **, * - уровни значимости 1%, 5% и 10 % соответственно		

3. В панельной модели *со случайными эффектами* предполагается, что индивидуальные различия имеют случайный характер.

Используя команду:

. xtreg lny6 lnx1 lnx5 lnx14 lnx15 lnx17 lnx19 lnx20 lnx22 k, re

получим следующие данные (таблица 8.13).

Таблица 8.13 – Результаты оценки модели со случайными эффектами

```
Random-effects GLS regression                   Number of obs      =       500
Group variable: id                              Number of groups   =       100

R-sq:  within  = 0.8628                         Obs per group: min =         5
       between = 0.9712                                        avg =       5.0
       overall = 0.9385                                        max =         5

                                                Wald chi2(9)       =   5750.45
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000
```

lny6	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lnx1	-.0305438	.0490839	-0.62	0.534	-.1267465	.0656589
lnx5	-.0158212	.0136082	-1.16	0.245	-.0424928	.0108504
lnx14	.9587887	.0219183	43.74	0.000	.9158296	1.001748
lnx15	.0137431	.0152324	0.90	0.367	-.0161119	.0435982
lnx17	-.0179366	.0207257	-0.87	0.387	-.0585582	.022685
lnx19	-.0177964	.0238687	-0.75	0.456	-.0645783	.0289855
lnx20	.0753259	.0502513	1.50	0.134	-.0231648	.1738167
lnx22	.0421858	.018019	2.34	0.019	.0068693	.0775023
k	.034193	.0178034	1.92	0.055	-.000701	.069087
_cons	-4.423823	.3374342	-13.11	0.000	-5.085182	-3.762464

sigma_u	.06176052	
sigma_e	.14115494	
rho	.16067862	(fraction of variance due to u_i)

Довольно высокое значение коэффициента регрессии в данной модели *(R-sq overall=0,9385)* показывает, что изменение зависимой переменной достаточно сильно зависит от индивидуальных различий.

Регрессия со случайными эффектами значима, что и показывает большое значение статистики Вальда – *Wald chi2(9)* = 5750,45. Регрессоры данной модели не коррелированны с ненаблюдаемыми случайными эффектами, что подтверждается значением выражения *corr (u_i, x) = 0 (assumed)*.

После исключения статистически незначимых переменных при помощи команды:

. xtreg lnx14 lnx22 k, re

получим (таблица 8.14).

Оценка регрессии «*between*» имела вспомогательный характер.

Когда необходимо проводить оценку модели при работе с реальными данными, всегда возникает вопрос о выборе адекватной модели. Различия между моделями можно рассматривать следующим образом. Так как при оценке факторов обычной регрессией предполагается, что индивидуальных различий у отдельных единиц совокупности нет. Но при наличии данных различий необходимо решить вопрос о том, носят они случайный или фиксированный характер.

Таблица 8.14 – Результаты эконометрической оценки модели со случайными эффектами (зависимая переменная: ln (стоимость валовой продукции в текущих ценах в расчете на 1 га сельскохозяйственных угодий, тыс. руб.)

```
Random-effects GLS regression                   Number of obs      =       500
Group variable: id                              Number of groups   =       100

R-sq:  within  = 0.8618                         Obs per group: min =         5
       between = 0.9699                                        avg =       5.0
       overall = 0.9375                                        max =         5

                                                Wald chi2(3)       =   5563.39
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000
```

lny6	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lnx14	.975057	.0135921	71.74	0.000	.948417	1.001697
lnx22	.0315985	.0172247	1.83	0.067	-.0021613	.0653583
k	.0354011	.0161897	2.19	0.029	.00367	.0671323
_cons	-4.351651	.1369188	-31.78	0.000	-4.620006	-4.083295
sigma_u	.06737354					
sigma_e	.14105879					
rho	.18575254	(fraction of variance due to u_i)				

Рассмотрим попарное сравнение полученных моделей для выбора между тремя основными регрессиями (сквозной, регрессией с фиксированными индивидуальными эффектами и регрессией со случайными индивидуальными эффектами).

Выбор адекватной модели

В результате мы получили три регрессии: сквозную (без учета времени…), с фиксированными эффектами, со случайными эффектами. Сравним попарно, полученные модели (таблица 8.15).

Сравним модель с фиксированными эффектами со сквозной регрессией. Согласно тесту Вальда *F test that all u_i=0: F(99,397)=2,16 Prob > F =*

0,0000. Так как уровень значимости не превышает 0,01, то нулевая гипотеза о равенстве «0» всех индивидуальных эффектов отвергается, и, следовательно, модель с фиксированными эффектами лучше подходит для описания данных, чем модель сквозной регрессии.

Таблица 8.15– Результат сравнения моделей

Сравниваемые модели, (тест)	Результаты
reg..., fe *reg ...* (Вальда)	*F test that all u_i=0:* *F(99,397) = 2,16* *Prob > F = 0,0000*
reg..., re *reg ...* (Бройша-Пагана)	*Breusch and Pagan Lagrangian multiplier test for random effects* *lny6[id,t] = Xb + u[id] + e[id,t]* *Estimated results:* *| Var sd = sqrt(Var)* *lny6 | .3894394 .6240508* *e | .0843764 .2904762* *u | .0288725 .169919* *Test: Var(u) = 0* *chibar2(01) = 57,85* *Prob > chibar2 = 0,0000*

Тест Бройша-Пагана, позволяющий сравнить модель сквозной регрессии с моделью со случайными эффектами, показал, что с уровнем значимости не более 0,01, нулевая гипотеза об отсутствии случайного индивидуального эффекта отвергается, т. е. модель регрессии со случайными эффектами лучше описывает наши данные, чем модель простой (сквозной) регрессии.

Тест Хаусмана:

reg…, re

reg…, fe (Хаусмана)

При помощи теста Хаусмана есть возможность выбора между описанными выше моделями (***fe*** и ***re***).

Если случайные эффекты не коррелируют с регрессорами, то

$$H_0: corr(u_i, x_{it}) = 0$$

(или u_i – случайные эффекты)

$$H_1: corr(u_i, x_{it}) \neq 0$$

(или u_i – детерминированные эффекты).

Тест Хаусмана основан на разности оценок $\hat{q} = \widehat{b_{fe}} - \widehat{b_{re}}$ ($\widehat{b_{fe}}$ представляет собой оценку, которая получена для модели с фиксированными эффектами для основной и альтернативной гипотезы, а $\widehat{b_{re}}$ −оценка модели со случайными эффектами при условии основной гипотезы), тест предполагает, что переменные в моделях с фиксированными и случайными эффектами должны быть одинаковы.

Тест Хаусмана позволяет сравнивать оценки коэффициентов одновременно присутствующие в моделях с фиксированными и случайными эффектами (таблица 8.16). Для проведения теста необходимо ввести команды, фиксирующие последние модели (*fe и re*):

.xtreg lny6 lnx14 lnx22 k, fe
.estimates store fixed
. xtreg lny6 lnx14 lnx22 k, re
.hausman fixed

Таблица 8.16 - Результаты проведения сравнительной оценки моделей при помощи теста Хаусмана

	Coefficients			
	(b) fixed	(B) .	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
lnx14	1.012196	.975057	.0371394	.0168507
lnx22	.0350032	.0315985	.0034047	.0133386
k	.0274314	.0354011	-.0079698	.003823

b = consistent under Ho and Ha; obtained from xtreg
B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test: Ho: difference in coefficients not systematic

chi2(3) = (b-B)'[(V_b-V_B)^(-1)](b-B)
= 4.86
Prob>chi2 = 0.1825
(V_b-V_B is not positive definite)

Так как *p-уровень=0,1825>0,01*, то основная гипотеза принимается и в нашем случае подходит модель со случайными индивидуальными эффектами. Этого и следовало ожидать, так как, несмотря на то, что выбирались одни и те же сельхозорганизации, однако результаты их деятельности существенно зависят от природно-климатических факторов, которые имеют случайный характер.

Выбор между ***fe*** и ***re*** моделями, осуществленный с помощью теста Хаусмана, привел к выводу о том, что модель со случайными индивидуальными эффектами лучше позволяет описать исходные данные, что содержательно обусловлено агроклиматическими условиями.

Таблица 8.17 – Результаты эконометрической оценки модели регрессии со случайными эффектами (зависимая переменная: ln (стоимость валовой продукции в текущих ценах в расчете на 1 га сельскохозяйственных угодий, тыс. руб.)

Переменная	Коэффициент	Стандартная ошибка
ln X14	0,9750***	0,0168
ln X22	0,0316***	0,0133
k	-0,0354**	0,0038
Коэффициент детерминации	0,9375	
***, **, * - уровни значимости 1%, 5% и 10 % соответственно		

Таким образом, модель со случайными эффектами является самой адекватной, причем согласно анализируемым данным подтверждается первоначальная гипотеза о статистически существенном влиянии на результативный фактор (логарифм стоимости валовой продукции в текущих ценах в расчете на 1 га сельскохозяйственных угодий, тыс. руб.). В частности результаты эконометрической оценки показывают, что наиболее значимым фактором в формировании результатов использования производственного потенциала (стоимость валовой продукции в текущих ценах на 1 га) является уровень является уровень интенсификации производства, выражающийся в уровне производственных затрат растениеводства на 100 га пашни (*x14*); коэффициент эластичности составил 0,975. Это можно объяснить тем, что данный показатель является интеграционным, так как в нем отражены и наличие, и использование основных фондов (через амортизационные отчисления) и рабочей силы через оплату труда.

Значительна роль активной части основных фондов – машин и оборудования (*x22*); увеличение их доли на 1 % обеспечивает прирост стоимости валовой продукции на 0,0316 %; *k* – период после кризиса 2008 г. добавляет 0,0354%.

Скорректированный множественный коэффициент детерминации (0,9375) достаточно значителен, что подтверждает правильность выбора факторных признаков и проведенного анализа.

Следовательно, использование моделей со случайными и детерминированными эффектами позволяет сделать вывод о том, что вышеуказанные факторы являются наиболее значимыми при регрессионном анализе.

Выводы: Анализ деятельности сельскохозяйственных предприятий Краснодарского края за 2010 г. показал, что существует связь между обеспеченностью ресурсами и объемами выпуска сельскохозяйственной продукции, что позволяет рассматривать ресурсы как систему факторов, которые формируют результаты производства. Отсюда возникает необходимость изучать зависимость между наличием ресурсов и объемами производства продукции.

В проведенном анализе изучения сельскохозяйственных предприятий подтверждается очевидная гипотеза их неоднородности и многообразия.

Задача.

По данным об оценке ресурсного обеспечения сельскохозяйственных организаций построить две альтернативные модели, отличающиеся набором переменных (от приведенных в тексте показателей), построить панельные регрессии и выбрать ту, которая лучше описывает исходные данные.

ПРИЛОЖЕНИЕ А

Основные показатели производства в сельскохозяйственных предприятиях Краснодарского края

№ п/п	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
1	351	155709	115278	31388	198701	172058	7886	15104	101641	182352
2	397	159544	135162	32089	223402	201735	7892	14331	129699	191211
3	222	110479	74741	20480	168581	111554	4687	19394	114006	207506
4	150	64543	55616	11877	108155	83009	3731	6708	60930	109689
5	536	182344	157580	46616	340653	235195	11434	8514	63782	147802
6	185	134062	88813	19438	184493	130608	5268	9854	77096	167947
7	341	127618	66723	23901	117282	99587	4802	17252	88500	185313
8	810	158417	87152	116717	368281	133362	14816	57884	87124	193423
9	370	169838	92409	25415	193640	137924	7755	16293	67305	135554
10	375	91918	77838	32484	134143	116176	6223	22152	91036	170156
11	385	113593	77680	17026	127807	115951	8831	17790	71814	125449
12	269	49333	140870	42887	269571	110254	8915	23435	87661	198650
13	435	130132	101667	25746	165201	101742	9659	16252	91199	143591
14	274	70377	66040	23289	111445	98564	5894	14500	78899	153258
15	124	30287	18710	11721	61314	50323	3334	10738	87038	161278
16	266	87057	113638	14005	196313	129610	8860	7838	95918	113760
17	377	168755	109097	8292	242429	132832	9949	12530	80585	188352
18	199	63826	52203	12307	108462	77916	5372	8719	41672	94372
19	345	123960	127635	19546	271314	140501	9860	9645	122010	204773
20	187	52210	42992	12113	77242	64168	5221	5711	68500	125284
21	324	146978	173814	20005	295199	59424	13281	9741	56606	102753
22	373	151084	121184	23391	200798	160873	8748	9906	91895	171009
23	971	204709	195510	66206	381457	191807	9955	21592	104805	224359
24	267	62740	58384	13335	114455	86896	4309	13007	90075	170289
25	120	59723	37280	3989	70426	59818	2646	6514	69542	120331
26	365	172447	104377	28154	241243	103508	8818	8888	51608	110182
27	381	77964	60249	26837	148307	92914	6276	7205	85401	143357
28	151	51636	31335	8764	75242	45503	3362	6321	55658	101375
29	340	127370	98445	28525	205838	137098	5551	7584	88695	156110
30	209	87083	93695	18276	145841	126111	5580	6020	63184	106813
31	302	84315	48868	20387	99349	83092	4869	7098	84426	150572
32	305	113236	117980	24848	192617	91121	4272	5935	54727	104732
33	813	178065	152741	55071	227871	80784	11668	4797	34934	65152
34	445	175573	148184	39496	325100	74995	8341	6949	45545	85678
35	194	44751	24690	10452	42787	9656	2230	5148	52770	104882
36	330	87611	94517	22972	180826	140795	5964	6507	63452	127642
37	420	138217	143627	38979	234873	144983	6058	14608	94938	168107
38	344	92015	83660	21431	125092	112532	4133	10194	76502	124575
39	465	222142	98457	63898	548580	177836	14399	8022	87503	152886
40	933	208908	142093	75978	236328	88493	9265	4718	30407	66743
41	331	216454	104582	116505	383380	152175	18363	14200	83460	167659
42	293	114757	130935	21667	225647	197022	5787	14097	108290	203382
43	727	131150	208454	53721	393764	219192	11421	10230	63329	225722

Продолжение приложения А

№ п/п	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
44	973	205782	151796	59795	128594	149364	10220	6298	31242	125012
45	422	173453	199096	31393	305305	147401	6518	18430	39374	71357
46	427	173900	151911	39733	296392	102209	10817	19680	61676	128083
47	767	233139	253323	29725	262997	106642	9286	49625	78251	169494
48	375	155911	93839	23033	196138	114611	9260	25382	51038	107438
49	355	91975	73342	10212	163378	134668	7530	26278	55500	120685
50	306	101823	157211	26361	228932	143776	4657	12931	76567	136979
51	571	163920	149779	28428	259681	213923	15214	28204	143591	229678
52	300	94839	58736	19094	129160	128514	4222	17604	144980	195485
53	454	127553	46808	23095	143124	101733	7406	20082	109086	188944
54	526	115270	118642	40039	230098	171913	6699	25210	110159	173455
55	367	105472	103608	16935	160955	154640	7233	17360	122945	201787
56	587	113959	144273	32325	265700	165331	6600	26069	107084	164328
57	147	40754	35300	7418	61218	51914	2924	6998	65500	120058
58	317	121082	92393	14015	190188	137800	8139	7013	82489	176470
59	450	112869	138572	18935	251787	86825	6983	21480	31390	67834
60	267	106658	90415	36312	174460	134949	6014	15181	109225	196469

Обозначения:

x_1 – Численность работников, чел.;
x_2 – Затраты на реализованную продукцию, тыс. руб.;
x_3 – Материальные затраты, тыс. руб.;
x_4 – Затраты по оплате труда, тыс. руб.;
x_5 – Валовая продукция, тыс. руб.;
x_6 – Затраты на производство, тыс. руб.;
x_7 – Сельскохозяйственные угодья, га;
x_8 – Энергетические мощности, л. с.;
x_9 – Стоимость основных фондов, млн. руб.;
x_{10} – Реализованная продукция, тыс. руб.

ПРИЛОЖЕНИЕ Б

Статистические данные по сельскохозяйственным организациям центральной зоны Краснодарского края, 2011 г.

№	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	27,710	24,297	3,628	21,416	160,259	2,027	0,113	4327
2	26,177	18,161	2,589	22,596	179,419	1,680	0,116	5562
3	30,482	34,023	2,926	12,823	91,395	2,895	0,079	3930
4	38,154	29,278	2,340	28,185	283,720	1,179	0,046	7093
5	35,357	41,915	3,373	22,881	137,242	2,415	0,183	8509
6	34,494	25,565	3,931	24,860	204,323	1,694	0,122	12668
7	63,345	49,983	6,279	38,721	111,859	4,701	0,239	9508
8	49,898	52,892	4,381	35,559	88,384	6,479	0,415	7601
9	45,939	50,369	3,548	31,143	161,538	3,365	0,198	4200
10	51,214	51,517	2,929	31,651	136,548	3,807	0,225	5735
11	23,759	21,114	4,323	20,040	188,381	2,710	0,145	3956
12	64,655	66,782	8,541	52,642	88,783	4,644	0,403	10654
13	24,901	27,694	2,554	22,169	160,067	2,173	0,072	7203
14	36,235	61,056	4,940	19,719	66,667	4,411	0,217	5000
15	52,942	76,518	4,456	43,572	79,794	4,162	0,120	5027
16	29,450	30,346	3,841	25,878	203,607	2,339	0,182	5701
17	49,391	39,839	3,681	34,537	120,314	2,791	0,223	4211
18	33,007	38,946	3,242	21,392	129,762	2,279	0,102	8175
19	21,635	25,615	4,015	22,075	197,216	2,039	0,115	7297
20	26,800	17,595	2,675	20,925	191,273	1,699	0,086	14728
21	33,721	30,670	3,477	21,540	165,357	4,760	0,159	4630
22	48,311	40,346	4,632	34,806	110,000	4,918	0,173	4620
23	40,411	58,586	4,766	27,405	127,865	3,876	0,214	12275
24	33,269	19,080	6,438	29,172	116,650	3,634	0,357	11665
25	38,302	51,770	4,807	29,039	121,297	3,734	0,267	14313
26	46,369	47,300	5,369	37,039	104,860	3,469	0,226	10486
27	46,831	26,440	4,840	39,902	133,000	5,194	0,275	2128
28	31,036	19,000	4,343	24,081	109,906	3,406	0,147	5825
29	32,972	34,227	5,656	28,434	70,719	3,930	0,284	4031
30	43,606	50,436	5,432	32,072	105,850	3,419	0,275	4234
31	42,909	32,950	2,816	34,491	254,347	1,785	0,084	12463
32	38,745	59,374	3,814	19,736	53,980	3,103	0,141	2753
33	55,514	75,247	5,844	39,211	114,471	2,810	0,187	9959
34	38,842	58,883	7,519	34,806	98,390	3,923	0,355	12102
35	29,319	38,990	3,807	28,874	127,045	2,095	0,217	13975

Обозначения:

y– выручка от реализации продукции на 1 га пашни, тыс. руб.;

x_1 – основные средства на 1 га пашни, тыс. руб.;

x_2 – среднегодовая численность работников на 100 га пашни, чел.;

x_3 – затраты на реализованную продукцию 1 га пашни, тыс. руб.;

x_4 – площадь пашни на 1 трактор, га;

x_5 – энергетические мощности на 100га пашни, л. с.;

x_6 – потреблено электроэнергии на 1 га пашни, тыс. квт.-ч;

x_7 – площадь пашни на одно предприятие, га.

ПРИЛОЖЕНИЕ В

Данные по производству молока в сельскохозяйственных предприятиях северной зоны Краснодарского края, 2011 г.

№ п/п	y_1	y_2	x_1	x_2	x_3	x_4	x_5	x_6
1	1242	1260	5488	3,18	700	118,0	646,7	74,5
2	1543	1560	3800	5,22	257	68,5	468,7	49,9
3	1421	1422	5453	1,69	500	162,3	610,6	78,6
4	1154	1214	6231	1,43	450	276,7	613,5	79,5
5	1366	1366	4616	3,61	1440	64,8	705,2	63,0
6	1746	1741	4939	2,62	575	119,1	549,7	64,2
7	1118	1166	5040	2,83	736	141,9	386,9	69,4
8	1381	1433	5201	2,49	1150	110,0	730,5	61,0
9	1424	1474	4636	3,11	500	156,9	717,0	74,0
10	1375	1413	5179	2,82	787	168,4	810,2	74,2
11	1520	1540	3732	3,17	600	113,0	661,0	64,1
12	1419	1419	5633	3,64	1340	124,4	644,3	81,1
13	1117	1194	3632	1,65	300	144,2	440,0	84,5
14	1207	1310	6711	1,47	2500	155,6	628,5	60,1
15	1284	1444	5602	1,52	2000	251,5	402,4	34,8
16	1458	1462	5741	2,15	1750	128,6	704,1	67,4
17	855	875	5527	3,88	2000	70,6	351,5	37,3
18	1120	1121	7407	1,88	1900	108,0	590,9	81,0
19	1193	1326	7561	1,80	823	106,0	650,5	90,9
20	1263	1264	4208	3,18	836	98,4	525,3	66,2
21	1177	1249	6520	2,06	358	157,7	619,9	72,5
22	1175	1212	5195	1,73	1200	98,2	583,0	84,9
23	1465	1510	4367	3,20	229	124,6	612,0	88,2
24	1359	1386	4539	2,51	800	206,1	652,0	77,8
25	1387	1396	4071	5,11	620	39,2	626,3	93,7
26	1058	1078	3714	2,94	403	115,4	379,0	80,9
27	1431	1436	9655	1,06	1233	162,4	598,0	90,8
28	2502	2582	2718	7,53	918	62,9	1475,4	49,5
29	1305	1305	6902	2,32	905	110,6	621,3	83,2
30	1291	1295	5904	2,36	640	104,8	669,0	79,0
31	1235	1318	8003	1,02	622	124,4	509,1	95,2
32	1624	1628	3431	3,63	632	58,8	610,6	28,4
33	1324	1391	4914	2,46	2992	159,7	461,9	74,2
34	1186	1186	7434	1,28	746	207,9	501,1	85,9
35	1274	1284	4455	1,40	595	123,9	568,0	81,3
36	1136	1136	5145	3,54	500	58,8	418,1	72,8
37	1773	1773	3883	4,11	670	89,3	774,3	61,0
37	1178	1165	7236	1,57	2108	306,9	436,9	39,8
39	1088	1001	7329	2,30	600	169,1	390,2	45,7
40	1493	1496	4410	2,56	850	110,3	794,7	69,2
41	1555	1556	4562	2,80	650	90,7	789,3	87,9

Обозначения:

x_1 – годовой надой молока на среднегодовую корову, кг;

x_2 –прямые затраты труда на 1 ц, чел.-ч;

x_3 – среднегодовое поголовье коров на предприятии, гол.;

x_4 – затраты по оплате труда на 1 ц, руб.;

x_5 – затраты на корма на 1 ц молока, руб.;

x_6 – доля молока в выручке от реализации продукции животноводства, %.

ПРИЛОЖЕНИЕ Г

Урожайность озимой пшеницы и количество внесенных минеральных удобрений на 1 га посева в сельскохозяйственных предприятиях

№ п/п	Муниципальное образование	Урожайность с 1 га, ц	Внесено минеральных удобрений на 1 га посева, кг д. в.	Природно-экономическая зона
1	г. Анапа	46,6	43,5	Анапо-Таманская
2	г. Армавир	56,0	113,1	Центральная
3	г. Краснодар	62,5	169,4	Центральная
4	Абинский район	43,8	113,2	Южно-предгорная
5	Белоглинский район	53,2	173,9	Северная
6	Белореченский район	38,7	101,2	Южно-предгорная
7	Брюховецкий район	58,4	138,2	Центральная
8	Выселковский район	67,6	236,1	Центральная
9	Гулькевичский район	61,4	151,2	Центральная
10	Динской район	62,9	115,4	Центральная
11	Ейский район	53,2	118,0	Северная
12	Кавказский район	62,1	118,3	Центральная
13	Калининский район	61,8	115,0	Западная
14	Каневский район	64,2	196,5	Северная
15	Кореновский район	57,2	106,1	Центральная
16	Красноармейский район	60,2	166,4	Западная
17	Крыловский район	52,6	142,9	Северная
18	Крымский район	37,5	88,1	Южно-предгорная
19	Курганинский район	55,8	120,4	Центральная
20	Кущевский район	52,1	64,0	Северная
21	Лабинский район	47,7	163,5	Южно-предгорная
22	Ленинградский район	58,3	162,3	Северная
23	Мостовский район	38,6	89,8	Южно-предгорная
24	Новокубанский район	59,3	145,0	Центральная
25	Новопокровский район	51,4	103,3	Северная
26	Отрадненский район	42,6	96,8	Южно-предгорная
27	Павловский район	55,8	167,5	Северная
28	Приморско-Ахтарский район	58,3	116,7	Центральная
29	Северский район	29,8	100,9	Южно-предгорная
30	Славянский район	44,3	104,6	Западная
31	Староминский район	54,3	145,8	Северная
32	Тбилисский район	61,2	97,1	Центральная
33	Темрюкский район	34,2	66,6	Анапо-Таманская
34	Тимашевский район	59,8	162,6	Центральная
35	Тихорецкий район	56,9	147,6	Северная
36	Успенский район	54,6	185,3	Южно-предгорная
37	Усть-Лабинский район	59,2	133,8	Центральная
38	Щербиновский район	54,4	145,4	Северная

ПРИЛОЖЕНИЕ Д

Урожайность сельскохозяйственных культур в хозяйствах Краснодарского края, ц с 1 га

№	Культура	2003 г.	2004 г.	2005 г.	2006 г.	2007 г.	2008 г.	2009 г.	2010 г.	2011г.	2112г.
1.	Зерновые культуры	29,5	42,7	44,5	41,0	38,3	54,3	45,9	48,7	54,5	41,9
2.	Пшеница озимая	33,6	44,2	48,2	42,7	45,1	57,4	47,0	51,1	55,9	39,9
3.	Пшеница яровая	11,9	22,5	27,5	23,8	22,5	29,4	23,8	31,0	28,4	24,8
4.	Рожь озимая	28,9	23,3	29,5	22,2	26,4	47,9	32,0	30,6	47,5	30,9
5.	Кукуруза на зерно	27,1	48,1	44,1	40,2	21,8	52,8	38,0	36,4	51,1	43,8
6.	Ячмень озимый	35,2	46,5	42,0	43,5	47,5	53,5	49,1	51,7	55,4	38,0
7.	Ячмень яровой	14,6	23,4	24,2	25,8	18,7	40,1	30,2	28,9	36,6	28,1
8.	Овес	13,2	26,3	26,6	25,2	21,0	36,8	25,7	25,1	31,4	26,4
9.	Просо	6,2	9,9	13,4	14,4	12,5	25,8	11,9	13,3	23,8	19,9
10.	Гречиха	3,8	4,3	6,4	6,6	4,5	7,3	4,1	4,1	6,9	7,7
11.	Рис	32,9	39,7	44,4	47,1	48,3	50,4	60,1	61,8	61,1	63,5
12.	Зернобобовые	8,8	23,3	19,7	22,9	14,4	33,8	23,6	23,9	28,6	42,3
13.	Горох	8,7	23,6	19,8	23,3	14,6	34,5	24,0	24,1	28,7	22,1
14.	Сахарная свекла	218	396	327,9	359,6	262,4	447,6	394,4	369,4	448,0	432,0
15.	Масличные культуры	13,8	18,0	19,3	18,7	16,4	23,6	21,3	20,322	22,5	21,9
16.	Подсолнечник	15,0	18,2	20,9	20,7	18,9	25,3	22,4	22,1	24,1	24,2
17.	Соя	10,1	18,0	15,1	12,8	9,1	16,4	18,2	15,818,9	18,9	18,6
18.	Картофель	38	86	88,0	89,6	78,9	194,9	132,0	126,1	163,4	153,2
19.	Рапс озимый	9,5	16,6	15,0	15,5	15,9	19,1	17,9	20,1	20,9	16,4
20.	Овощи	47	91	100,1	93,4	79,7	126,9	117,4	97,8	125,0	93,9
21.	Кукуруза на силос	124	188	155,9	171,3	126,3	190,7	166,3	141,1	193,9	155,3
22.	Кормовые корнеплоды	233	351	325,6	247,5	196,8	361,9	285,6	285,3	234,7	209,1
23.	Сено многолетних трав	24,1	28,5	24,8	33,3	16,5	33,0	23,3	29,5	31,7	45,0
24.	Сено однолетних трав	30,6	17,9	29,9	34,9	21,9	32,1	25,7	24,5	33,0	27,3
25.	Плоды и ягоды	86,6	58,1	77,5	63,3	55,5	89,7	93,8	73,7	96,0	109,2
26.	Виноград	84,3	57,1	69,6	49,1	80,7	74,8	86,1	81,7	116,2	76,3
27.	Чайный лист	6,1	8,7	8,9	8,1	4,5	5,7	4,7	2,7	2,2	1,0
28.	Бахчи кормовые	86	127	143,7	135,8	122,3	181,4	176,8	149,0	155,6	164,1

ПРИЛОЖЕНИЕ Е

Урожайность зерновых и зернобобовых культур в Краснодарском крае с 1 га, ц

Год	Хозяйства всех категорий							Сельскохозяйственные организации						
	Зерновые культуры	Озимая пшеница	Озимый ячмень	Ячмень яровой	Кукуруза на зерно	Рис	Бобовые культуры	Зерновые культуры	Озимая пшеница	Озимый ячмень	Ячмень яровой	Кукуруза на зерно	Рис	Бобовые культуры
1961	20,5	19,5	21,2	9,7	27,8	30,9	11,3	20,2	19,5	21,3	9,8	27,2	30,9	11,3
1962	24,8	26,9	25,7	18,7	21,5	30,1	13,3	24,9	26,9	25,8	19,0	21,1	30,0	13,4
1963	25,8	27,6	25,9	20,6	22,4	35,0	18,8	25,8	27,5	26,1	20,7	21,8	35,0	18,8
1964	23,0	21,2	24,3	21,8	33,5	26,7	15,2	22,7	21,1	24,3	22,0	33,5	26,6	15,1
1965	26,0	25,6	29,2	16,3	35,2	35,0	16,3	25,8	25,6	29,3	16,5	33,9	35,0	16,3
1966	29,1	29,5	30,1	19,8	32,7	34,4	18,1	28,8	29,5	30,3	20,3	30,7	34,4	18,0
1967	26,8	25,7	28,0	21,9	33,9	39,3	19,0	26,5	25,7	28,0	22,1	32,6	39,3	18,9
1968	30,0	30,1	30,7	13,7	34,8	48,5	15,6	30,0	30,1	30,9	13,9	35,4	48,5	15,7
1969	20,7	22,0	22,7	16,5	24,3	40,8	13,8	20,6	22,0	22,5	16,5	24,2	40,8	13,7
1970	35,2	36,6	36,6	27,9	27,7	45,1	27,2	35,3	36,6	36,6	27,9	25,8	45,1	27,2
1971	35,2	37,2	37,0	24,5	22,9	48,2	19,6	35,6	37,2	37,0	24,5	23,7	48,2	19,8
1972	25,2	25,5	30,9	20,4	25,3	47,0	16,1	25,1	25,5	30,9	20,4	24,0	47,0	16,2
1973	32,5	30,9	37,0	27,1	37,0	46,1	20,6	32,3	30,9	37,0	27,1	36,2	46,1	20,6
1974	34,3	34,3	39,5	27,8	32,9	47,0	24,3	34,3	34,3	39,5	27,8	32,1	47,0	24,5
1975	27,1	25,5	29,3	19,5	28,2	50,0	17,5	27,0	25,5	29,3	19,5	26,8	50	17,7
1976	34,0	33,5	37,1	31,3	37,0	43,2	21,1	34,0	33,5	37,1	31,3	36,7	43,2	21,2
1977	33,9	32,8	35,7	25,9	42,3	44,7	24,1	33,6	32,8	35,7	25,9	41,3	44,7	24,2
1978	34,7	39,7	41,0	29,8	31,5	39,8	22,7	37,6	39,7	41,0	29,8	32,5	39,8	22,8
1979	30,9	31,3	35,3	19,8	26,2	43,8	13,2	30,9	31,3	35,3	19,8	23,8	43,8	13,3
1980	31,6	31,1	33,4	23,3	26,4	45,9	19,8	31,6	31,1	33,4	23,3	25,1	45,9	20,0
1981	31,1	32,4	35,8	23,8	24,9	39,0	12,6	31,0	32,4	35,8	23,7	22,4	39,0	12,6
1982	34,2	33,1	34,7	24,3	42,2	40,1	20,8	34,0	33,1	34,7	24,3	42,2	40,1	20,9
1983	32,2	31,6	36,4	25,9	33,5	43,3	20,0	32,1	31,6	36,4	25,9	32,5	43,3	20,1

Продолжение приложения Е

Год	Хозяйства всех категорий							Сельскохозяйственные организации						
	Зерновые культуры	Озимая пшеница	Озимый ячмень	Ячмень яровой	Кукуруза на зерно	Рис	Бобовые культуры	Зерновые культуры	Озимая пшеница	Озимый ячмень	Ячмень яровой	Кукуруза на зерно	Рис	Бобовые культуры
1984	37,1	37,6	39,9	30,1	38	45,5	25,3	37,1	37,6	39,9	30,1	38,2	45,5	25,5
1985	29,9	28,8	26,7	24,6	34,8	41,4	22,5	29,3	28,8	26,7	24,6	31,3	41,4	22,5
1986	41,4	43,2	47,3	30,1	27,8	52,2	14,1	41,7	43,2	47,3	30,1	26,8	52,2	14,1
1987	38,6	39,5	40,8	32,1	40,2	47,7	18,0	38,6	39,5	40,8	32,1	39,3	47,7	18,1
1988	37,7	37,5	42,5	27,2	39,4	50,2	17,6	37,7	37,5	42,5	27,2	39,8	50,2	17,7
1989	40,2	42,5	43,5	31,3	42,2	37,0	23,3	41,6	43,7	45,1	31,3	42,6	42,5	23,4
1990	49,4	56,4	57,7	37,4	35,3	35,6	24,5	49,7	55,0	56,0	37,4	35,9	35,6	24,6
1991	39,5	42,7	42,4	30,4	36,3	33,5	16,6	39,5	42,7	42,4	30,4	36,3	33,5	16,6
1992	36,7	38,7	39,3	29,5	33,3	38,1	15,7	37,0	38,9	39,5	30,4	33,5	38,0	15,7
1993	38,2	40,9	42	30,4	35,4	36,8	24,0	38,6	41,3	42,3	31,3	34,2	36,8	24,3
1994	30,9	35,2	31,7	30,8	16,4	36,6	23,7	32,2	36,2	32,1	31,6	15,4	36,5	23,9
1995	30,4	32,0	31,5	24,9	31,3	34,4	16,7	30,9	32,7	31,8	25,6	30,9	34,4	17,0
1996	25,7	28,1	34,2	18,4	14,6	27,1	19,1	27,1	28,7	34,5	19,0	15,8	27,1	19,4
1997	30,8	33,6	34,9	13,9	35,1	23,5	8,5	31,4	34,6	35,4	15,2	35,2	23,5	8,6
1998	24,1	29,0	32,1	19,0	12,9	34,3	16,1	25,9	30,3	33,0	20,0	13,3	34,3	16,4
1999	33,7	37,8	41,7	25,7	20,1	29,7	17,1	35,8	38,8	42,6	28,3	21,4	29,8	17,5
2000	34,5	38,8	42,5	22,0	22,0	41,7	18,6	37,2	40,2	43,6	24,8	24,5	41,8	19,0
2001	37,9	44,0	43,2	27,0	12,0	39,6	23,6	41,3	45,6	44,5	30,4	12,9	39,6	24,1
2002	41,5	47,2	46,1	22,3	28,8	39,6	19,5	44,2	48,7	47,6	25,5	31,3	39,6	19,9
2003	29,6	33,4	35,0	16,0	31,4	32,9	8,8	29,5	33,6	35,2	14,6	27,1	32,9	8,8
2004	41,0	43,1	44,9	22,3	43,9	39,8	23,1	42,7	44,2	46,5	23,4	48,1	39,7	23,3
2005	42,5	46,7	41,2	23,6	40,4	44,3	19,5	44,5	48,2	42,0	24,2	44,1	44,4	19,7
2006	41,0	42,7	43,5	25,8	40,3	47,1	22,9	42,7	43,7	44,9	28,0	43,3	46,9	23,3
2007	38,3	45,1	47,5	18,7	22,0	48,3	14,4	41,4	46,5	49,2	21,5	23,8	48,4	14,8
2008	51,9	55,3	51,4	36,9	49,5	50,7	32,9	54,3	57,4	53,5	40,1	52,8	50,4	33,8
2009	43,2	45,7	46,9	27,0	33,8	60,3	23,1	45,9	47,0	49,1	30,2	38,0	60,1	23,6
2010	46,1	49,7	49,2	25,0	33,8	62,1	23,7	48,7	51,1	51,7	28,9	36,4	61,8	23,9
2011	52,6	55,1	53,8	33,8	47,7	61,0	27,9	54,5	55,9	55,4	36,6	51,1	61,1	28,6
2012	40,8	39,8	37,1	27,4	41,9	64,3	41,1	41,9	39,9	38,0	28,1	43,8	63,5	42,3

Список использованной литературы

1. Бондаренко П.С. Чуприна Н.В. Основы теории вероятностей и математической статистики. Часть 1. Теория вероятностей. КГАУ. Краснодар, 1997.
2. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике: Учеб. пособие для бакалавров. / В.Е. Гмурман. –М.: Юрайт, 2013. 405 с.
3. Гмурман В.Е. Теория вероятностей и математическая статистика. Учебное пособие. М.: Высшая школа, 2003, 2009. – 479с. М.: Юрайт, 2013. – 479с.
4. Горелова Г.В., Кацко И.А. Теория вероятностей и математическая статистика в примерах и задачах с применением Excel, Ростов н/Д.: Феникс,2006. – 475 с.
5. Калинина В. Н., Панкин В.Ф. Математическая статистика М.: Дрофа, 2002.– 336 с.
6. Колемаев В.А. Теория вероятностей и математическая статистика: учебник / В.А. Колемаев, В.Н. Калинина. – 3-е изд., перераб. и доп. – М.: КНОРУС, 2009.– 384 с.
7. Кремер Н.Ш. Теория вероятностей и математическая статистика: Учебник для вузов. - М.: ЮНИТИ-ДАНА, 2007. – 573с.
8. Общий курс высшей математики для экономистов: Учебник / Под общ. ред. В.И. Ермакова. – М.: ИНФРА-М. 2008. – 656с.
9. Солодовников А.С. Математика в экономике: учебник. В 3-х ч. Ч.3. Теория вероятностей и математическая статистика / А.А.Солодовников, В.А. Бабайцев, А.И. Браилов. – М.: Финансы и статистика, 2008. – 464с.
10. Теория статистики с основами теории вероятностей./ И.И. Елисеева, В.С. Князевский, Л.И. Новорожкина, Э.А. Морозова; Под ред. И.Н. Елисеевой.-М.: ЮНИТИ-ДАНА, 2001. – 446с.
11. Боровиков В. П. STATISTICA. Искусство анализа данных на компьютере: Для профессионалов. 2-е изд. / В. П. Боровиков – СПб.: Питер, 2003. – 688 с.: ил.
12. Дрейпер И. Прикладной регрессионный анализ: В 2-х кн.; пер. с англ. – 2-е изд. перераб. и доп. / И. Дрейпер, Г. Смит – М. : Финансы и статистика, 1986. – Кн. 1. – 366 с., ил; 1987. Кн. 2. – 351 с., ил.
13. Елисеева И. И. Эконометрика: учебник / И. И. Елисеева, С. В. Курышева, Т. В. Костева и др., под ред. И. И. Елисеевой. – 2-е изд., перераб. и доп. – М. : Финансы и статистика, 2005. – 576 с.: ил.
14. Кацко И. А. Практикум по анализу данных на компьютере / И. А. Кацко, Н. Б. Палкин; под ред. Г. В. Гореловой. – М. : КолосС, 2009. – 278 с.
15. Магнус Я. Р. Эконометрика. Начальный курс: учебник / Я. Р. Магнус, П. К. Катышев, А. А. Пересецкий – М.: Дело, 2004. – 576 с.

16. Орлов А. И. Прикладная статистика / А. И. Орлов. – М. : Экзамен, 2006. – 611 с.
17. Орлов А. И. Организационно-экономическое моделирование: учеб-ник: в 3-х ч.; Ч. 1 Нечисловая статистика / А. И. Орлов. – М. : Изд-во МГТУ им. Н. Э. Баумана, 2009. – 541 с.
18. Орлов А.И. Эконометрика: учебник / А.И. Орлов – М. : Экзамен, 2004. – 412 с.
19. Паклин, Н. Б. Бизнес-аналитика: от данных к знаниям: учеб. пособие. 2-е изд., перераб. и доп./ Н. Б. Паклин, В. И. Орешков. – СПб. : Питер, 2010. – 704 с.
20. Ратникова Т. А. Введение в эконометрический анализ панельных данных (рус.) / Т. А. Ратникова // Экономический журнал ВШЭ. – 2006. – № 2. – С. 267–316.
21. Ратникова Т. А. Анализ панельных данных в пакете «STATA» / Т. А. Ратникова // Методические указания к комьютерному практикуму по курсу «Эконометрический анализ панельных данных». – М. : ГУ ВШЭ.– 2004.

Учебное издание

Бондаренко Петр Сергеевич
Кацко Игорь Александрович
Перцухов Виктор Иванович
Сенникова Алина Евгеньевна
Жминько Альбина Евгеньевна
Соловьева Татьяна Владимировна
Стеганцова Екатерина Дмитриевна
Чернобыльская Татьяна Юрьевна

ПРАКТИКУМ ПО ЭКОНОМЕТРИКЕ

Учебно-практическое пособие для бакалавров

В авторской редакции

Подписано в печать 09.10.13 . Формат 60 × 84 $^{1}/_{16}$.
Тираж 500 экз. Усл. печ. л. - 10,2. Уч.-изд. л. – 9,3.
Заказ № 652

Типография Кубанского государственного аграрного университета.
350044, Краснодар, ул. Калинина, 13