УНИВЕРСИТЕТ ИТМО

Т.Г. Максимова

И.Н. Попова

ЭКОНОМЕТРИКА

Санкт-Петербург
2018

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

УНИВЕРСИТЕТ ИТМО

Т.Г. Максимова

И.Н. Попова

ЭКОНОМЕТРИКА

РЕКОМЕНДОВАНО К ИСПОЛЬЗОВАНИЮ В УНИВЕРСИТЕТЕ ИТМО
по направлениям подготовки (специальности)
38.04.01 «Экономика» в качестве учебно-методического пособия для реализации основных профессиональных образовательных программ высшего образования магистратуры

УНИВЕРСИТЕТ ИТМО

Санкт-Петербург

2018

Максимова Т.Г., Попова И.Н. Эконометрика: учебно-методическое пособие / Т.Г. Максимова, И.Н. Попова. – СПб.: Университет ИТМО, 2018. – 70 с.

Рецензент: Верзилин Дмитрий Николаевич, д.э.н., профессор, заведующий кафедрой менеджмента и экономики спорта, Национальный государственный университет физической культуры, спорта и здоровья им. П.Ф. Лесгафта, Санкт-Петербург.

Учебно-методическое пособие соответствует образовательному стандарту высшего образования Санкт-Петербургского национального исследовательского университета информационных технологий, механики и оптики по направлению подготовки 38.04.01 «Экономика» (уровень магистратуры).

В пособии подробно рассмотрена парная регрессия, включая расчет параметров, их интерпретация, значимость, определение доверительных интервалов, построение множественной линейной модели, оценка ее качества, отбор значимых факторов, прогнозирование результативного признака и особенности изучения временных рядов: аналитическое выравнивание, устойчивость, сезонность, прогнозирование.

УНИВЕРСИТЕТ ИТМО

Рекомендовано к печати УМС ИМБИП, протокол № 4 от «15» ноября 2017 г.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Традиционный состав эконометрического инструментария представлен стандартным набором математико-статистических методов:

- классическая линейная модель множественной регрессии и классический метод наименьших квадратов;
- обобщенная линейная модель множественной регрессии и обобщенный метод наименьших квадратов;
- некоторые специальные модели регрессии (со стохастическими объясняющими переменными, с переменной структурой, с дискретными зависимыми переменными, нелинейные);
- модели и методы статистического анализа временных рядов;
- анализ систем одновременных эконометрических уравнений.

Решаемые с помощью эконометрики задачи классифицируются в соответствии:

- с конечным прикладным целями исследования:
    - прогноз экономических и социально-экономических показателей, характеризующих экономическую систему,
    - имитация сценариев развития экономической системы;
- с уровнем иерархии анализируемой экономической системы:
    - макроуровень (страна в целом),
    - мезоуровень (регионы, отрасли, корпорации),
    - микроуровень (семьи, предприятия, фирмы);
- с профилем эконометрического моделирования:
    - проблемы рынка;
    - проблемы инвестиционной, финансовой и социальной политики;
    - проблемы ценообразования, спроса и предложения.

Пособие предназначено для самостоятельной практической работы студентов, обучающихся по программам специалитета (направление подготовки 38.05.02 «Таможенное дело») и магистратуры (направление подготовки 38.04.01 «Экономика»), и содержит как краткое изложение ряда теоретических положений эконометрики, так и примеры решения типовых задач с использованием программных продуктов. Существенное внимание уделено регрессионному анализу и анализу временных рядов, как наиболее простым и востребованным практикой методам анализа социально-экономических данных.

В последнем разделе подробно разобран пример эконометрического исследования с использованием многомерных статистических методов, выходящих за рамки традиционного эконометрического инструментария. Пример основан на официальных статистических данных, предоставляемых в открытом доступе Росстатом; может рассматриваться как решение типовой задачи эконометрического анализа социально-экономических данных.

1 ТИПЫ ДАННЫХ, ИХ ПОДГОТОВКА

1.1 Типы данных

Для изучения какого-либо явления необходимо организовать статистическое наблюдение, целью которого является - сбор данные. Сбор данных следует осуществлять в соответствии с определенными переменными, которые могут быть: номинальными; порядковыми; интервальными.

Номинальные (или категориальные) переменные позволяют проводить качественную классификацию объектов, например, по полу, цвету, принадлежности к определенному региону и т.д. Арифметические операции не имеют смысла. Поэтому ни медиана, ни среднее не имеет смысла. Статистикой центральной тенденции является мода.

Порядковые переменные используют для ранжирования (упорядочивания) объектов по степени проявления анализируемого свойства, но не позволяют оценить «на сколько меньше» или «на сколько больше». Например, ранжирование экзаменуемых по степени подготовленности к ответу: удовлетворительно, хорошо, отлично, великолепно. Арифметические операции не имеют смысла. Статистикой центральной тенденции является мода и медиана.

Интервальные переменные позволяют упорядочивать объекты измерения, а также оценивать различия между ними. Чтобы производить сравнения надо ввести единицу измерения и задать начало отсчета. Например, измерение температуры воздуха в градусах Цельсия или Фаренгейта образует интервальную шкалу. Арифметические операции имеют смысл. Статистикой центральной тенденции является мода, медиана, среднее.

Статистические данные — это конкретные численные значения статистических показателей, они могут быть пространственными, временными и панельными

Пространственные данные — данные, полученные в результате статистического наблюдения за несколькими единицами статистической совокупности на один и тот же момент времени или за один и тот же период времени. Такие данные имеют два измерения: признак — объект.

Временные данные (динамические ряды) — данные, полученные в результате статистического наблюдения за одной единицей статистической совокупности за разные моменты или периоды времени. Временные данные имеют два измерения: признак — время.

Панельные данные — данные, полученные в результате статистического наблюдения за несколькими единицами статистической совокупности за разные моменты или периоды времени. Панельные данные имеют три измерения: признак — объект — время. Панельные данные позволяют проводить совместный анализ пространственных выборок и анализ временных рядов.

Существуют также псевдопанельные данные — объединенные по времени независимые одномоментные данные (например, данные, полученные в результате ежегодно повторяющейся случайной выборки).

1.2 Подготовка данных

1. Каждому индивидууму (экспериментальной единице) присваивается идентификатор.
2. Наблюдения над индивидуумами записываются в матричной форме: строки – индивидуумы, столбцы – наблюдаемые для каждого индивидуума значения признаков.
3. Признаки индивидуума надо упорядочивать в близкие по смыслу группы.
4. Коды названий признаков должны быть такие, чтобы было легко понять, что же это за признак.
5. Отсутствующие значения признаков кодируются, как правило, пробелом или числом, которое является невозможным наблюдением для данного признака (например, рост - 9999).
6. Для измерений в шкале наименований и в порядковой шкале лучше присваивать цифры, а не буквы (например, пол: мужчина – 1, женщина – 0, а не м/ж).
7. Не рекомендуется переводить измерения в интервальной шкале в порядковую (например, возраст целесообразно указывать в годах, а не в шкале «1 – менее 14, 2 – от 14 до 21 и т.д.).
8. Целесообразно объединять две переменные в одну, если это не влечет потери информации. Например, переменная1 - наличие в семье детей (0 - нет, 1 - есть) и переменная2 – возраст старшего ребенка могут быть заменены второй (0 – отсутствие детей).

2 ПАРНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

2.1 Основные понятия регрессионного анализа

В регрессионном анализе рассматривается связь между одной зависимой переменной и несколькими другими независимыми переменными. Эта связь представляется с помощью математической модели. Выбор подходящей модели основывается как на статистических доводах, так и на основе содержательного смысла моделируемой зависимости.

Зависимая (объясняемая результирующая, эндогенная) переменная (признак) *y* - характеризует результат функционирования экономической системы. Ее значения формируются под воздействием других переменных и факторов, поэтому результирующая переменная является случайной величиной.

Объясняющие (независимые, предикторные, экзогенные) переменные $X=(x_1, x_2, \ldots, x_m)$ - переменные (признаки), которые описывают условия функционирования экономической системы и в существенной мере определяют значения объясняемой переменной. Независимые переменные могут быть как случайными, так и детерминированными.

Статистические исследования направлены на оценивание параметров регрессии; проверку гипотез о статистической значимости этих параметров; проверку адекватности построенной модели.

Регрессионный анализ используется с двумя целями. Во-первых, для описания зависимости между переменными и определения причинной связи. Во-вторых, для построения прогнозных значений зависимой переменной. Мерой зависимости является величина коэффициента корреляции.

2.2 Парная линейная регрессия: модель

Модель парной линейной регрессии имеет вид:

$$y = \alpha + \beta x + u, \qquad (1)$$

где y - зависимая переменная (объясняемая),

x - независимая переменная (объясняющая),

α, β - параметры модели,

u - случайный остаточный член (случайная ошибка).

Константу α называют также *свободным членом*, а угловой коэффициент β - *регрессионным коэффициентом*.

Оценка параметров модели основана на имеющейся выборке парных наблюдений объема *n*: *(x1 , y1), … (xn , yn).*

Для оценки используется уравнение:

$$\hat{y} = a + bx, \qquad (2)$$

где $\hat{y}$ - прогнозируемое значение объясняемой переменной;

a – статистическая оценка параметра α;

b - статистическая оценка параметра β.

Уравнение (2) задает прямую линию на плоскости.

Разность между фактическим значение зависимой переменной и значением, прогнозируемым по уравнению регрессии, называется *остатком*. Остатки e_i вычисляются по формуле:

$$e_i = y_i - \hat{y}_i . \qquad (3)$$

Принципиальная схема модели парной линейной регрессии приведена на рис. 1.

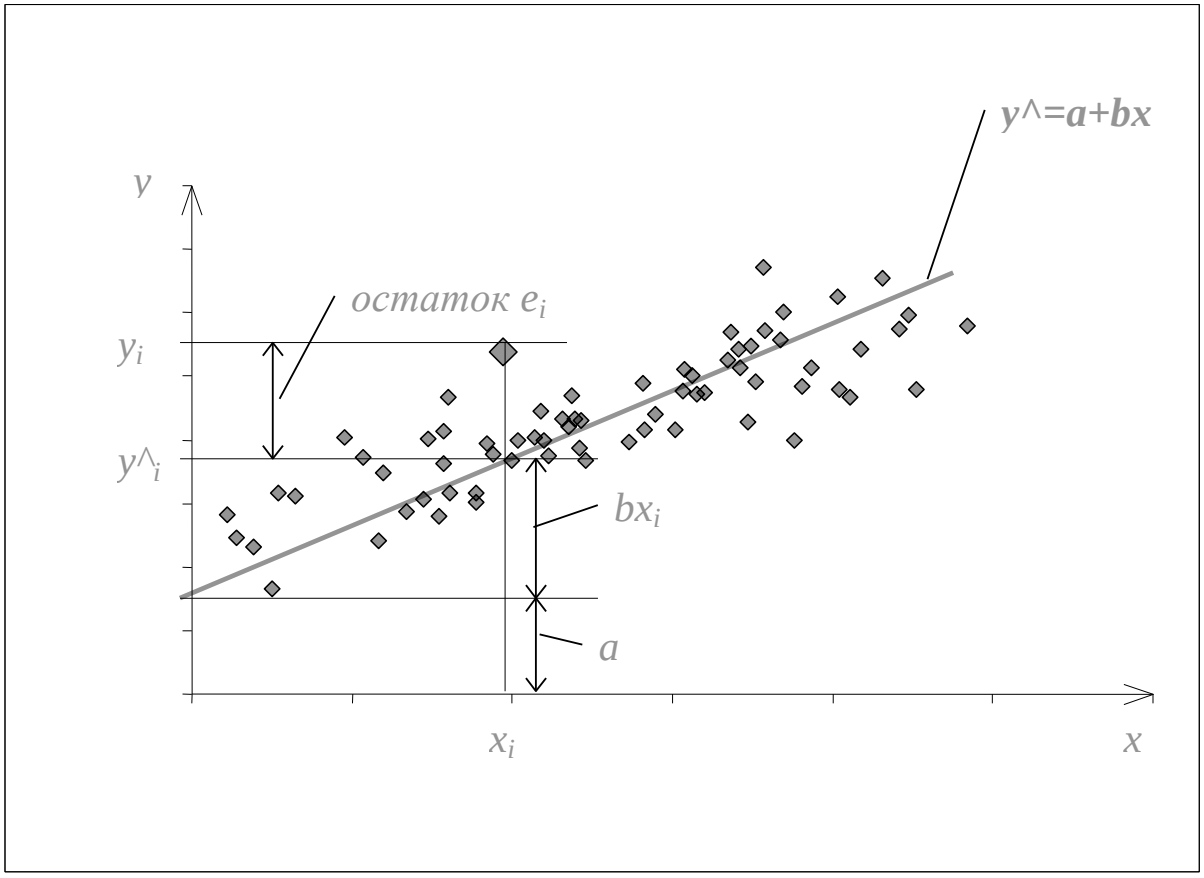

Рис. 1. Принципиальная схема модели парной линейной регрессии

Для определения существования и степени линейной зависимости между переменными необходимо:

- построить диаграммы рассеяния – графическом отображении точек $(x_1 , y_1), \ldots (x_n , y_n)$ на плоскости. Анализируя диаграмму рассеяния, можно оценить, допустимо ли предположение о линейной зависимости между x и y;
- вычислить выборочный коэффициент корреляции r по формуле:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \quad (4)$$

где

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad (5)$$

соответствующие выборочные средние для переменных *x* и *y*.

2.3 Проверка статистической гипотезы о значимости коэффициента парной линейной регрессии

Рассмотрим простую функцию спроса:

$y = \alpha + \beta x + u$, где y – величина спроса (на продукты питания), x – доход. Мы предполагаем, что спрос зависит от дохода.

Обычно строят нулевую гипотезу, которая будет проверяться с помощью альтернативной гипотезы, которая предполагается верной.

В качестве нулевой гипотезы принимается утверждение о том, что спрос *не зависит* от дохода, т.е. y *не зависит* от x, т.е. β=0

Альтернативная гипотеза - β≠0, т.е. x влияет на y, т.е. доход влияет на спрос.

Таким образом H_0 – гипотеза об отсутствии изменений

H_0: β=0

H_1: β≠0

В общем случае для нулевой гипотезы утверждают, что $\beta=\beta_0$, тогда альтернативная гипотеза - $\beta\neq\beta_0$.

H_0: $\beta=\beta_0$

H_1: $\beta\neq\beta_0$

Если H_0 верна, то в – МНК-оценки для β будут иметь распределение с математическим ожиданием и дисперсией:

$$E(b) = \beta_0$$

$$D(b) = \frac{\sigma^2}{\sum\left(x_i - \bar{x}\right)^2}. \quad (6)$$

Если дополнительно

$$u \sim N(0,\sigma\,2) \Rightarrow b \sim N\left(\beta_0, \frac{\sigma^2}{\sum\left(x_i - \bar{x}\right)^2}\right),$$

то МНК-оценки для β распределены по нормальному закону.

Проверка гипотез.

Строим t-статистику для b:

$$t = \frac{b - \beta_0}{s_b} \qquad (7)$$

Если гипотеза H_0 верна, то t-статистика подчиняется распределению Стьюдента (или, как его еще называют, t-распределению), заданному таблично.

Параметром распределения Стьюдента является *v*- число степеней свободы.

Правило определения числа степени свободы.

Оценивание каждого параметра поглощает 1 степень свободы, следовательно, $v = n - 1$, где *n* –количество наблюдений в выборке.

В таблице распределения Стьюдента (см., например, Доугерти, 2003) для различных v заданы критические значения t-статистики, которые обозначаются $t_{кр}$.

Гипотеза t=0 эквивалентна H_0.

Надо проверить: $-t_{кр} < t < t_{кр}$, , если выполнено, то мы не должны отказываться от нулевой гипотезы. Если $t < -t_{кр}$ или $t > t_{кр}$, то H_0 надо отклонить.

Ошибки, возникающие при проверке гипотез, приведены в таблице 1.

Таблица 1 – Ошибки принятия гипотез

Действие	Условие	
	H_0 верна	H_0 не верна
Отвергнуть H_0	Ошибка 1-ого рода, вероятность ошибки α	Верное решение
Принять H_0	Верное решение	Ошибка 2-ого рода вероятность ошибки β

Задача – минимизировать ошибки α и β.

Дилемма: α - убывает, следовательно, β растет,

β - убывает, следовательно, α растет.

Решение дилемы: выбираем малое α и полагаем, что β будет тоже мало. Величину α называют уровнем значимости – это вероятность отвергнуть верную гипотезу H_0, $\alpha = p_r$ (H_0 отвергнута / H_0 верна). Используются значения: α= 0.1, 0.05, 0.01.

Процедура проверки гипотезы:

вычислить t;

задать α= 0.05;

найти $t_{кр;}$

проверить попало t в критическую область или нет;

если попало, то H_0 отвергаем (есть влияние)

если не попало, то H_0 не отвергаем.

Эквивалентная процедура проверки гипотез:

вычислить t;

найти p – значение = $p_г$ ($|t| > |t_{кр}|$) вероятность того, что при выполнении H_0 статистика критерия (t) принимает значение более экстремальное, чем $t_{кр}$;

если p - значение< α, то H_0 отвергаем.

2.4 Доверительные интервалы для параметров уравнения парной линейной регрессии

100(1-α)% -ный доверительный интервал для b:

$$b \pm s_b \times t_{1-\frac{\alpha}{2}}(n-2), \tag{8}$$

100(1-α)% - ный доверительный интервал для a:

$$a \pm s_a \times t_{1-\frac{\alpha}{2}}(n-2). \tag{9}$$

Статистика $t_{1-\alpha/2}(n-2)$ имеет распределение Стьюдента с $(n-2)$ степенями свободы.

Если $\bar{y}$ интерпретируется как наилучшая оценка единственного значения y, соответствующего $x = x_i$, а также число наблюдений достаточно велико (по крайней мере, больше 30), то для $\bar{y}$ может быть построен, так называемый, «быстрый» доверительный интервал.

100(1-α)% - ный доверительный интервал для $\bar{y}$: $\bar{y} \pm t_{1-\frac{\alpha}{2}}(n-2)$, где s – стандартная ошибка оценки.

3 МНОЖЕСТВЕННЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

3.1 Множественная линейная регрессия: модель

Результативная (объясняемая, зависимая, эндогенная) переменная (признак) Y - признак, характеризующий результат функционирования экономической системы.

Значения объясняемой переменной Y формируются под воздействием ряда других переменных. Эти переменные называются объясняющими (факторными или предикторными, экзогенными переменными (признаками или факторами) $X_1, X_2, \ldots, X_k$.

Формализуем проблему предсказания одной переменной Y с помощью *k* переменных $X_1, X_2, \ldots, X_k$. Модель множественной линейной регрессии имеет вид:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \mathrm{u}, \qquad (10)$$

где $\beta_0, \beta_1, \beta_2, \ldots, \beta_m$ - параметры модели;

u - случайная ошибка.

Коэффициент регрессии при каждой переменной *X* дает оценку ее влияния на величину *Y* в случае неизменности влияния на нее всех остальных переменных.

Уравнение для оценки модели множественной регрессии с k объясняющими (независимыми) факторными признаками имеет вид:

$$\hat{Y}_i = a + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_k X_{ki} + \varepsilon_i, \qquad (11)$$

где

$\hat{Y}_i$ – теоретическое (предсказанное) значение результативного (зависимого) признака при i-м наблюдении;

$X_{1i}, X_{2i}, \ldots, X_{ki}$ – значения факторных (независимых) признаков для *i*-ого наблюдения;

a – точечная оценка свободного члена (сдвиг);

$b_1, b_2, \ldots b_k$ – точечные оценки коэффициентов чистой (частной) регрессии;

ξ_i – случайный остаток в i-м наблюдении, равный $(Y_i - \hat{Y}_i)$.

Как и в случае парной регрессии оценка проводится на выборке объема *n*.

Наилучшие оценки получаются при минимизации суммы квадратов остатков между фактическим и прогнозируемым значением зависимой пере-

менной. Для вычисления $b_0, b_1, b_2, ..., b_m$ используется метод наименьших квадратов.

3.2 Общий подход к построению модели множественной регрессии

При построении модели множественной регрессии решаются задачи определения возможности использования факторных признаков $X_1, X_2, ..., X_k$ в уравнении множественной регрессии и определения наиболее существенных факторных признаков, влияющих на значение результативного признака.

При построении уравнения множественной регрессии первоначально на основе качественного анализа необходимо провести отбор наиболее существенных факторных признаков, воздействующих на результативный признак.

Далее анализ дополняется количественными оценками. Для характеристики тесноты связи рассчитываются парные коэффициенты корреляции:

1) парные (линейные) коэффициенты корреляции, характеризующие тесноту связи каждого факторного признака $X_1, X_2, ..., X_k$ с результативным признаком Y по формуле:

$$r_{YX_j} = \frac{\overline{YX_j} - \overline{Y}\,\overline{X_j}}{\sigma_{X_j}\,\sigma_Y}, \quad j = 1, ..., k, \tag{12}$$

где среднеквадратические отклонения определяются по формуле:

$$\sigma_{X_j} = \sqrt{\overline{X_j^2} - (\overline{X_j})^2}; \quad \sigma_Y = \sqrt{\overline{Y^2} - (\overline{Y})^2} \tag{13}$$

2) парные (линейные) коэффициенты корреляции, характеризующие тесноту межфакторной взаимосвязи $X_1, X_2, ..., X_k$, определяются по формуле:

$$r_{X_j X_p} = \frac{\overline{X_j X_p} - \overline{X_j}\,\overline{X_p}}{\sigma_{X_j}\,\sigma_{X_p}}, \quad j = 1, ..., p, ..., k \tag{14}$$

Интерпретация значений парных коэффициентов корреляции дана в таблице 3.

Чем больше абсолютное значение парного коэффициента корреляции, тем более тесная взаимосвязь между данными двумя признаками. При выборе факторных признаков значения парных коэффициентов корреляции между

факторными признаками и результативным признаком должны быть высокими.

Таблица 2 - Интерпретация значений парных коэффициентов корреляции

Значение парного коэффициента корреляции	Характер связи признаков (вид, направление)	Характеристика связи признаков
$r = -1$	функциональная обратная	с ростом одного признака строго функционально уменьшается другой признак
$-1 < r < 0$	корреляционная обратная	с ростом одного признака уменьшается другой признак
$r = 0$	связь отсутствует	–
$0 < r < 1$	корреляционная прямая	с ростом одного признака уменьшается другой признак
$r = 1$	функциональная прямая	с ростом одного признака строго функционально увеличивается другой признак

В случае если парный коэффициент корреляции между объясняющими факторами достаточно высокий (абсолютное значение превышает 0,7), то это свидетельствует о наличии мультиколлинеарности, т.е. о наличии линейно-связанных факторных признаков. При наличии таких связей между факторными признаками один или некоторые из них следует исключить таким образом, чтобы между оставшимися факторными признаками не было тесных связей. Этим обеспечивается существенность воздействия каждого факторного признака на результативный признак и оценка значимости каждого факторного признака.

На основе расчета парных коэффициентов корреляции не только исключаются из дальнейшего рассмотрения коллинеарно-связанные признаки, но и определяется последовательность включения факторных признаков в уравнение.

Далее рассмотрим общий подход к построению модели множественной регрессии, в основе которого находится один из двух рассматриваемых ниже способов последовательного отбора факторных признаков для построения модели.

Определение регрессионных стандартизованных коэффициентов (β-коэффициентов), объяснение их значений и построение уравнение множественной регрессии в стандартизованном виде

В случае если факторные признаки различны по своей природе, имеют разные единицы измерения, то для адекватной оценки влияния на результа-

тивный признак следует перевести все значения факторных признаков в стандартизованный вид по формуле:

$$t_{X_j} = \frac{X_j - \overline{X_j}}{\sigma_{X_j}}, \text{ где } \sigma_{X_j} = \sqrt{\frac{(X_j - \overline{X_j})^2}{n}} \quad (15)$$

Уравнение множественной регрессии в стандартизованном виде:

$$\bar{t}_{1,2,...,k} = \beta_1 t_1 + \beta_2 t_2 + ... + \beta_k t_k, \quad (16)$$

где

$t_{1,2,...,k}$ – стандартизованные значения факторных признаков $X_1, X_2, ..., X_k$;

$\bar{t}_{1,2,...,k}$ – средние значения стандартизованного результативного признака, полученного по уравнению регрессии;

$\beta_1, \beta_2, ..., \beta_k$ – регрессионные стандартизованные коэффициенты.

β-коэффициенты рассчитываются по следующим типовым формулам:

$$\beta_1 = \frac{r_{YX_1} - r_{YX_2} r_{X_1X_2}}{1 - r^2_{X_1X_2}}; \quad \beta_2 = \frac{r_{YX_2} - r_{YX_1} r_{X_1X_2}}{1 - r^2_{X_1X_2}} \quad (17)$$

β_1 (β_2) показывает, что с ростом фактора X_1 (X_2) на величину среднеквадратического отклонения по этому факторному признаку, значение результативного признака Y увеличивается на данное значение β-коэффициента при условии неизменности других включенных в уравнение факторных признаков.

Определение параметров уравнения, объяснение их значений и построение уравнение множественной регрессии в стандартной форме

Параметрами уравнения в обычном виде являются:

1) коэффициенты чистой (частной) регрессии $b_j, j = 1,...,k$, определяемые по формуле:

$$b_j = \beta_j \frac{\sigma_Y}{\sigma_{X_j}}, \; j = 1,...,k \quad (18)$$

Коэффициенты b_j определяются на основе применения метода наименьших квадратов, который обеспечивает наименьшие расхождения между теоретическими и фактическими значениями результативного признака.

Коэффициенты характеризуют изменение среднего значения результативного признака Y при изменении значения факторного признака $X_1, X_2, ..., X_k$ на единицу соответственно, при условии, что все остальные факторные признаки не оказывают влияние на изменение результативного признака.

2) свободный член уравнения (сдвиг) a:

$$a = \overline{Y} - b_1 \overline{X_1} - b_2 \overline{X_2} - ... - b_k \overline{X_k}, \quad (18)$$

Уравнение множественной регрессии в стандартной форме:

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + ... + b_k X_k \quad (19)$$

3.3 Проверка адекватности модели

Расчет и интерпретация множественного коэффициента корреляции и детерминации

Множественный коэффициент корреляции используется при пошаговой регрессии с включением факторных признаков в уравнение. R изменяется в пределах $0 \le R \le 1$ и по мере приближения к единице свидетельствует о повышении тесноты взаимосвязи признаков.

Множественный коэффициент корреляции может быть также рассчитан через стандартизованные коэффициенты (β-коэффициенты) по формуле:

$$R_{X_1, X_2, ..., X_k} = \sqrt{\beta_1 r_{YX_1} + \beta_2 r_{YX_2} + ... + \beta_k r_{YX_k}} \quad (20)$$

,

Если множественный коэффициент корреляции возвести в квадрат, то будет получен множественный коэффициент детерминации:

$$R^2_{X_1, X_2, ..., X_k} = \beta_1 r_{YX_1} + \beta_2 r_{YX_2} + ... + \beta_k r_{YX_k}, \quad (21)$$

Множественный коэффициент детерминации показывает, сколько процентов вариации результативного признака Y объясняется вариацией включенных в уравнение множественной регрессии факторных признаков. Остальная вариация объясняется влиянием факторных признаков, не учтенных в модели.

Оценка существенности (значимости) уравнения множественной регрессии осуществляется на основе расчета F-критерия Фишера и средней ошибки аппроксимации.

Проверка значимости множественной корреляции и всего уравнения в целом осуществляется на основе расчета фактического значения F-критерия Фишера по формуле:

$$F_{факт} = \frac{R^2_{X_1,X_2,\ldots,X_k}}{1 - R^2_{X_1,X_2,\ldots,X_k}} \frac{(n-m)}{(m-1)}, \quad (22)$$

где n – количество наблюдений; m - число параметров в уравнении множественной регрессии.

В случае если рассчитанное фактическое значение F-критерия Фишера превышает критическое значение F-распределения, то следует сделать вывод о значимости коэффициента множественной регрессии и всего уравнения в целом. Критическое значение определяется по таблице F-распределения в зависимости от числа степеней свободы $\nu_1 = m-1$, $\nu_2 = n-m$ и заданного уровня значимости (0,05 (рекомендовано); 0,01; 0,001).

Для оценки величины суммарной ошибки (расхождения теоретических и фактических значений результативного показателя Y) рассчитывается средняя ошибка аппроксимации по формуле:

$$\bar{\varepsilon} = \frac{1}{n} \sum \frac{(Y - \hat{Y})}{Y} 100, \quad (23)$$

Средняя ошибка аппроксимации находится в пределах от 1 до 100%. В случае если средняя ошибка не превышает 10%, то сумма всех отклонений теоретических значений результативного признака от его фактически значений находится в пределах нормы. Следовательно, уравнение множественной регрессии можно использовать для дальнейшего анализа и прогнозирования.

Оценка существенности характеристик связи предполагает расчет значений t-критерия Стьюдента в отношении значения каждого коэффициента регрессии (b_j). Формула расчета и правило интерпретации полученного значения t-критерия Стьюдента были приведены ранее (см. пошаговая регрессия с исключением).

Далее рассмотрим методику определения остатков и их анализ .

Остаток (оценка ошибки e_i) есть разность между фактическим значением Y_i и теоретическим (предсказанным) значением $\hat{Y}_i$ при заданных значениях факторных признаков X_{ij}:

$$e_i = Y_i - \hat{Y}_i \tag{24}$$

Далее следует построить графики и проанализировать распределение остатков в зависимости от предсказанных значений $\hat{Y}_i$ и от всех факторных признаков $X_1, X_2, \ldots, X_k$.

Если построенные графики не имеют ярко выраженной зависимости (остатки примерно одинаково часто принимают отрицательные либо положительные значения), то это свидетельствует о том, что построенная модель множественной регрессии адекватна и может быть использована для анализа и прогнозирования.

О зависимости ошибок свидетельствует автокорреляция. Автокорреляция означает зависимость остатков от значений предыдущих остатков и предполагает сомнительную адекватность построенной модели.

Гипотеза о независимости ошибок может быть проверена на основании измерения автокорреляции посредством расчета статистики Дарбина-Уотсона. Статистика Дарбина-Уотсона дает оценку корреляции соседних остатков:

$$D = \frac{\sum (e_i - e_{i-1})^2}{\sum e_i^2} \tag{25}$$

Для данного уровня значимости, количества наблюдений n и количества факторных признаков k по таблице определяется нижнее d_1 и верхнее d_2 критическое значение статистики Дарбина-Уотсона. Возможны следующие выводы об автокорреляции&

Таблица 3 - Общая интерпретация значений статистики Дарбина-Уотсона следующая:

Значение статистики Дарбина-Уотсона	Характеристика автокорреляции остатков и адекватности модели
$D \to 0$	Между соседними остатками существует положительная автокорреляция, т.е. существуют кластеры остатков, имеющих одинаковый знак. Модель неадекватна (использование метода наименьших квадратов нецелесообразно).
$D \to 2$	Между соседними остатками не существует автокорреляция. Модель адекватна.
$D \to 4$	Между соседними остатками существует отрицательная автокорреляция (встречается редко), т.е. остатки скачкообразно принимают положительные и отрицательные значения. Модель неадекватна.

Таблица 4 - Характеристика автокорреляции остатков

Значение статистики Дарбина-Уотсона	Характеристика автокорреляции остатков
$D < d_1$	Между соседними остатками существует положительная автокорреляция.
$d_1 < D < d_2$	Вывод о наличии либо отсутствии автокорреляции сделать нельзя.
$D > d_2$	Между соседними остатками не существует положительная автокорреляция.

Проверка значимости оценок коэффициентов регрессии b_j находится в основе проведения пошаговой регрессии с исключением факторных признаков из уравнения множественной регрессии (см. Способы отбора). В основе проверки значимости находится расчет значения t-критерия Стьюдента ($t_{факт_{b_j}}$) и его сопоставление с критическим значением. В случае превышения критического значения принимается решение о значимости оценки данного коэффициента регрессии b_j и существовании значимой статистической связи между результативным признаком Y и данным факторным признаком X_j.

Для оценки генеральной совокупности должен быть построен доверительные интервал для коэффициентов регрессии, который вычисляется при данном уровне значимости (например, 0,05) по формуле:

$$b_j - t_{n-m} S_{b_j} \le \beta_j \le b_j + t_{n-m} S_{b_j} \quad (26)$$

Таким образом, можно утверждать, что с вероятностью 0,95 при изменении факторного признака X_j на одну единицу своего значения, результативный признак Y изменится (в большую либо меньшую сторону, в зависимости от знака регрессионного коэффициента β_j) на величину, которая колеблется в пределах рассчитанных верхней и нижней границ. Если доверительный интервал не содержит ноль, это означает, что регрессионный коэффициент β_j имеет статистически значимое влияние на изменение результатного признака Y.

3.4 Предсказание (прогнозирование) значения результативного признака

На основе выводов об отсутствии автокорреляции построенная модель множественной регрессии позволяет прогнозировать значение результативного признака Y. С этой целью следует подставить прогнозируемые значения факторных признаков $X_{1np}, X_{2np}, \ldots, X_{knp}$ в уравнение множественной регрессии и получить предсказанное (прогнозируемое) значение результативного признака $\hat{Y}_{np}$.

Далее определяется средняя ошибка прогноза ($M_{\hat{Y}_{np}}$) по формуле:

$$M_{\hat{Y}_{np}} = S_{YX} \sqrt{\frac{1}{n} + \frac{(X_{1np} - \overline{X_1})^2}{\sum (X_{1i} - \overline{X_1})^2}} \quad (26)$$

Это формула только для одного X.

Предельная ошибка ($\Delta \hat{Y}_{np}$) определяется по формуле:

$$\Delta \hat{Y}_{np} = t_{n-k-1} M_{\hat{Y}_{np}}, \quad (27)$$

где t_{n-k-1} – критическое значение t-критерия Стьюдента с данным уровнем значимости (например, 0,05) и со степенью свободы ($n-k-1$).

Таким образом, доверительный интервал для предсказанного значения будет соответствовать:

(28)

$$\hat{Y}_{np} - \hat{\Delta Y}_{np} \leq Y_{X_{1np}, X_{2np}, \ldots, X_{knp}} \leq \hat{Y}_{np} - \hat{\Delta Y}_{np} \qquad (28)$$

Таким образом, при указанных условиях 95%-ый получен доверительный интервал для среднего значения результативного признака при данных значениях факторных признаков $X_{1np}, X_{2np}, \ldots, X_{knp}$.

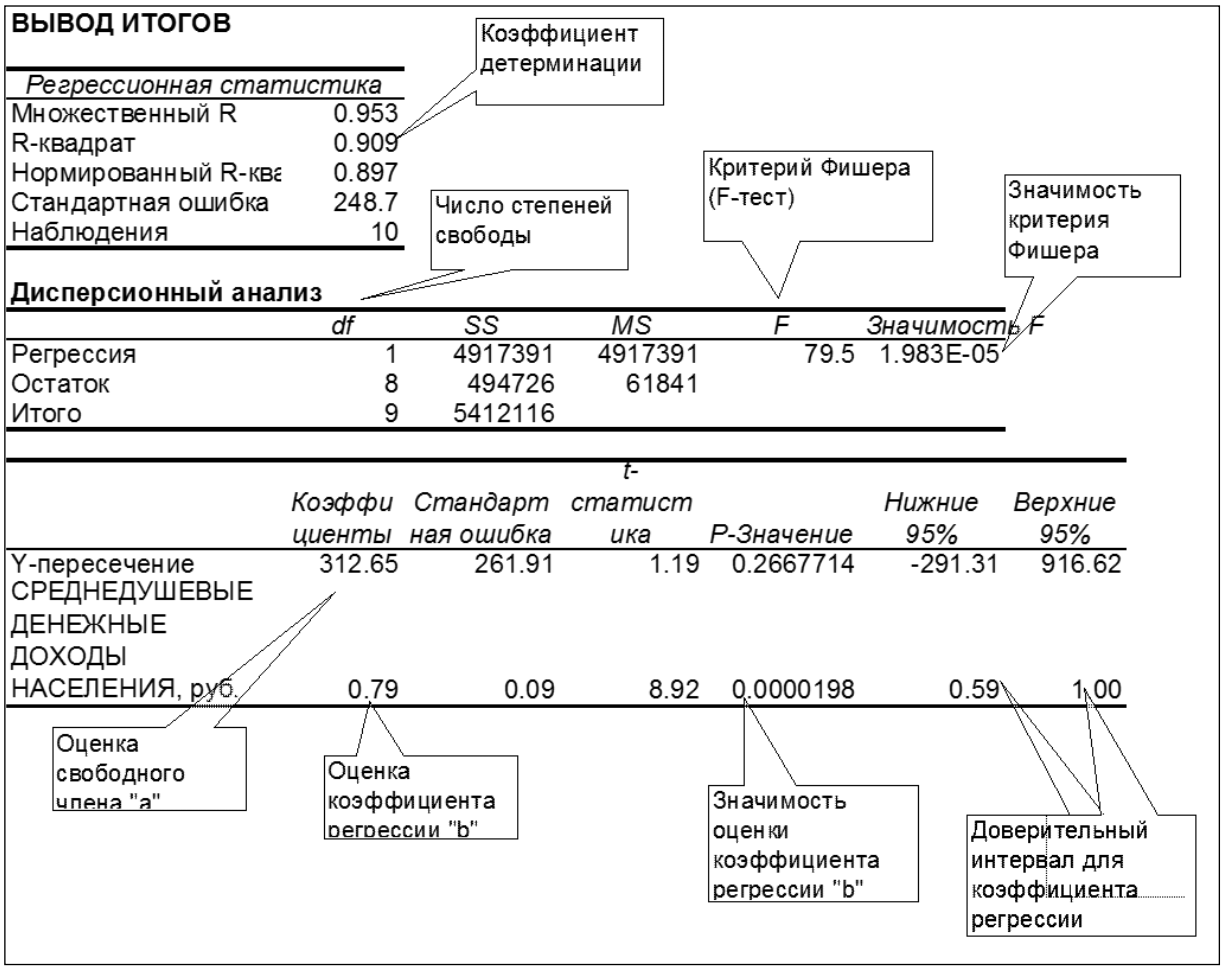

Рис. 2 – Пример таблицы результатов регрессионного и дисперсионного анализа

4 ПОСТРОЕНИЕ МОДЕЛИ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ: ПРИМЕР

4.1 Содержательная постановка задачи

Известны следующие данные. В отношении семей с данным средним количеством человек собраны сведения относительно их общих расходов в неделю в у.е., а также сведения о расходах на питание в неделю в у.е.

Требуется построить модель множественной регрессии и определить характеристики зависимости значений известных признаков.

Таблица 5 - Исходные данные

Общие расходы, у.е. в неделю, $\mathbf{x_1}$	Средний размер семьи, чел, $\mathbf{x_2}$	Расходы на питание, у.е. в неделю, **y**
45	1,52	21,8
75	1,6	33,4
125	1,86	50,3
223	1,83	66,9
92	3,43	47,3
146	3,62	66
227	3,44	81
358	3,47	106
135	5,54	70,3
218	5,44	94,6
331	5,41	119
490	5,33	147,2
175	8,49	92,8
305	8,25	132,8
468	8,14	169
749	7,33	196,9

1) Определить возможность использования факторных признаков в уравнении множественной регрессии. Определить наиболее существенные факторные признаки, влияющие на значение результативного признака.

   Определить регрессионные стандартизованные коэффициенты (β -коэффициенты), объяснить их значения и построить уравнение множественной регрессии в стандартизованном виде.

2) Определить параметры уравнения, объяснить их значения и построить уравнение множественной регрессии в обычном виде.
3) Провести корреляционный анализ. Рассчитать множественный коэффициент корреляции и детерминации. Объяснить полученные значения.
4) Оценить существенность уравнения множественной регрессии и характеристик связи.
5) Определить остатки и провести их анализ.
6) Проверить значимость оценки коэффициентов, построить доверительные интервалы для коэффициентов регрессии.

   Предсказать значение результативного признака Y в зависимости от конкретных значений факторных признаков и определить 95%-доверительный интервал.

4.2 Схема выполнение задания в Excel

В Excel возможно проведение оценки существенности уравнения множественной регрессии и характеристик связи, а также проверка значимости оценок коэффициентов регрессии β_j и построение доверительных интервалов для коэффициентов регрессии.

С этой целью в файле Excel, содержащем исходные данные, следует выбрать опцию «Сервис» - «Анализ данных» - «Регрессия». В качестве результата анализа будут приведены следующие данные.

Во-первых, появляется таблица «*регрессионная статистика»*, которая содержит следующие рассчитанные показатели (таблица 1):

Таблица 6 - Результаты регрессионного анализа (в Excel)

Наименование показателя	Содержание показателя
Множественный R	Множественный коэффициент корреляции R.
R-квадрат	Множественный коэффициент детерминации R^2, учитывающий влияние факторных признаков.

Наименование показателя	Содержание показателя
Нормированный R-квадрат	Множественный нормированный коэффициент детерминации R^2, учитывающий влияние факторных признаков и объем выборки, вычисляемый по формуле: $$R^2_{корр} = 1 - \left((1 - R^2) \frac{n-1}{n-k-1} \right) \quad (29)$$
Стандартная ошибка	Среднеквадратическая ошибка оценки, т.е. стандартное отклонение фактических значений Y от теоретических значений результативного признака: $$S_{YX} = \sqrt{\frac{SS_R}{n-m}} = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n-m}} \quad (30)$$
Наблюдения	количество наблюдений (n).

Во-вторых, формируется таблица «*дисперсионный анализ*» следующего вида (таблица 2). Следует обратить внимание на то, что отношение SS_R и SS_D представляет собой коэффициент множественной детерминации.

В-третьих, формируется таблица, характеризующая качество параметров уравнения множественной регрессии (таблица 3) на основе проверки нулевой и альтернативной гипотез:

$$H_0 : \beta_j = 0 \text{ (нет зависимости)}$$
$$H_1 : \beta_j \neq 0 \text{ (есть зависимость)} \quad (31)$$

Таблица 7 - Результаты дисперсионного анализа (в Excel)

	df	*SS*	*MS*	*F*	*Значимость F*
Вид дисперсии	Количество степеней свободы	Сумма квадратов	Средний квадрат	F - критерий Фишера	Значимость F - критерия
Регрессия (дисперсия теоретических значений Y, сформированная под влиянием факторных признаков)	$\nu_1 = m - 1 = k$	$SS_D = \sum(\hat{Y} - \bar{Y})^2$	$MS_D = \frac{\sum(\hat{Y} - \bar{Y})^2}{\nu_1}$	$F = \frac{MS_D}{MS_R}$	вероятность ошибки 1-го рода*
Остаток (остаточная дисперсия фактических Y, сформированная под влиянием неучтенных в модели факторных признаков)	$\nu_2 = n - m = n - k - 1$	$SS_R = \sum(Y - \hat{Y})^2$	$MS_R = \frac{\sum(Y - \hat{Y})^2}{\nu_2}$	—	—
Итого (общая дисперсия фактических Y, сформированная под влиянием факторных и неучтенных в модели признаков)	$\nu_3 = n - 1$	$SS_T = SS_D + SS = \sum(Y - \bar{Y})^2$	—	—	—

* вероятность ошибки первого рода в данном случае означает вероятность отклонить гипотезу о несущественности уравнения множественной регрессии, при условии, что она правильная (при уровне значимости 0,05).

Таблица 8 - Результаты проверки гипотез о значимости коэффициентов регрессии (в Excel)

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
	Параметры модели	Среднеквадратическая ошибка оценки параметров	t -критерий Стьюдента ($t_{факт}$)	Значимость оценки параметров	Нижняя граница доверительного интервала	Верхняя граница доверительного интервала
Y-пересечение	a			t - значение распределения Стьюдента как функция вероятности и числа степеней свободы.	$a - t_{n-m}S$	$a + t_{n-m}$
Переменная X_1	b_1	$S_{b_1} = \frac{S_{YX}}{\sqrt{SS_{D_1}}}$, $SS'_{D_1} = \sum(X_1 - \overline{X}$	$t_{факт_{b_1}} = \frac{b_1}{\beta_1 = 0}$		$b_1 - t_{n-m}$	$b_1 + t_{n-m}$
Переменная X_2	b_2	$S_{b_2} = \frac{S_{YX}}{\sqrt{SS_{D_2}}}$, $SS'_{D_2} = \sum(X_2 - \overline{X}$	$t_{факт_{b_2}} = \frac{b_2}{\beta_2 = 0}$		$b_2 - t_{n-m}$	$b_2 + t_{n-m}$
…	…	…	…		…	…
Переменная X_k	b_k	$S_{b_k} = \frac{S_{YX}}{\sqrt{SS_{D_k}}}$, $SS'_{D_k} = \sum(X_k - \overline{X}$	$t_{факт_{b_k}} = \frac{b_k}{\beta_k = 0}$		$b_k - t_{n-m}$	$b_k + t_{n-m}$

Значение t-критерия Стьюдента ($t_{факт_{b_j}}$), полученное в таблице 3 для коэффициента регрессии b_j, рассчитывается с уровнем значимости, равным 0,05. В случае если $t_{факт_{b_j}}$ превысит критическое значение t-критерия Стьюдента (функция СТЬЮДРАСПОБР(уровень доверия; количество степеней свободы)), то нулевая гипотеза H_0 для данного коэффициента b_j отклоняется. Значит, при фиксированных значениях других факторных признаков между факторным признаком X_j и результативным признаком Y существует значимая статистическая связь.

Аналогичный вывод формулируется в случае если полученное в таблице 3 p-значение не превышает заданный уровень значимости (в данном случае, 0,05). p-значение также может быть рассчитано по функции СТЬЮДРАСП($t_{факт_{b_j}}$; $n-m$; 2). Т.е. если p-значение почти равно нулю, то это означает, что если бы между факторным признаком X_j и результативным признаком Y не существовало зависимости, то обнаружить ее было бы почти невозможно.

В таблице 3 Excel в целях оценки генеральной совокупности также определяются доверительные интервалы для коэффициентов регрессии, на основании которых строится доверительный интервал. Автоматически определяются нижняя и верхняя границы значений параметров уравнения при уровне значимости 0,05.

4.3 Пример выполнение задания в системе STATISTICA

1. Выбрать из меню «*Statistics*» («Статистика») «*Multiple Regression*» («Множественная регрессия»). Появится окно вида:

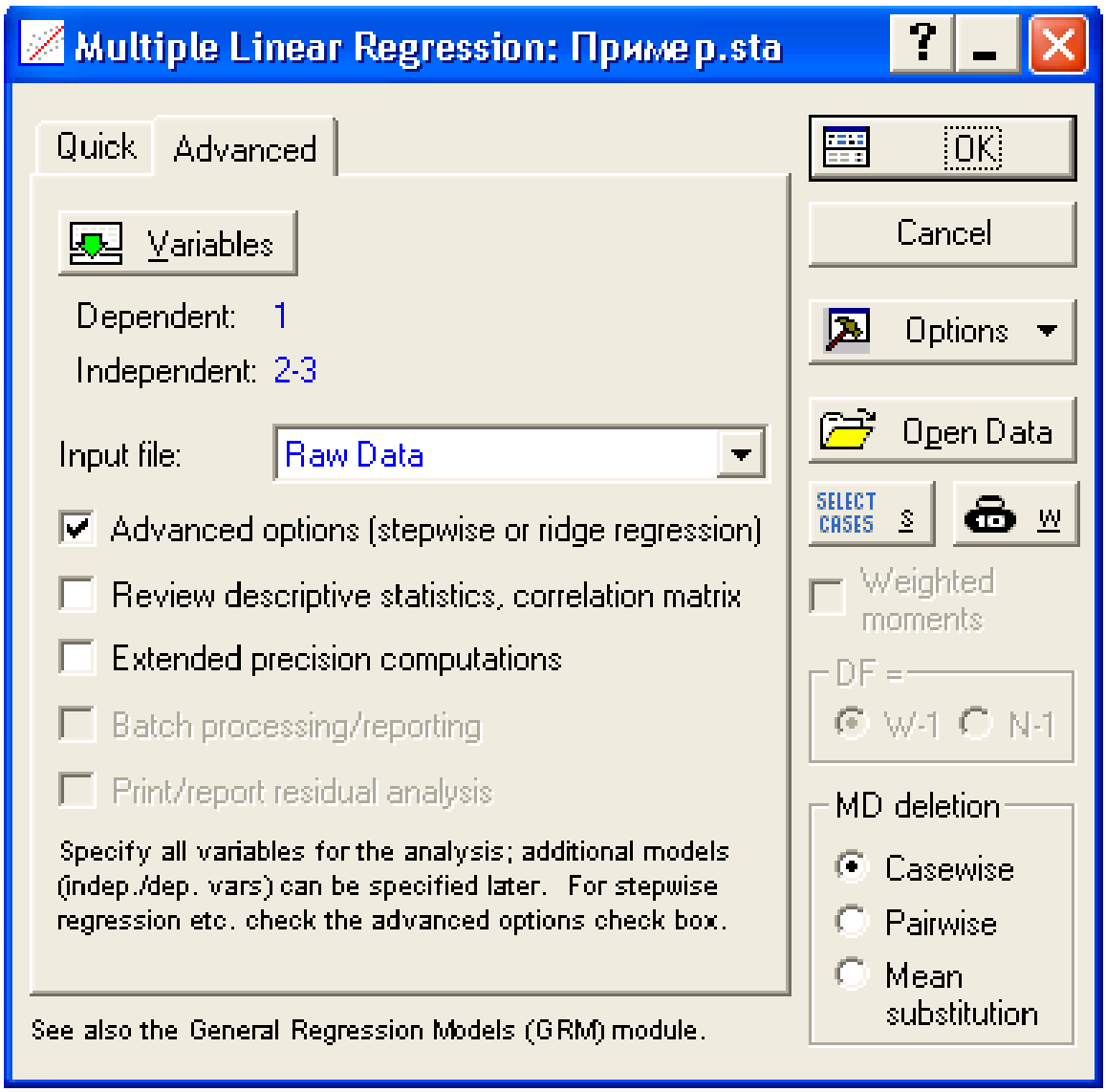

Рис. 3. Диалоговое окно «Множественная регрессия»

В окне «*Multiple Linear Regression*» («Множественная линейная регрессия») выбрать переменные, нажав кнопку «*Variables*» («Переменные»):

- «*Dependent*» («Зависимые»): расходы на питание, долл. в неделю, y;

- «*Independent*» («Независимые»): общие расходы, долл. в неделю, x1; Средний размер семьи, чел, x2.

Поставить «галочку» «*Advanced Options*». Нажать кнопку «*Ok*».

2. В появившемся окне «*Model Definition*» («Определение модели»)

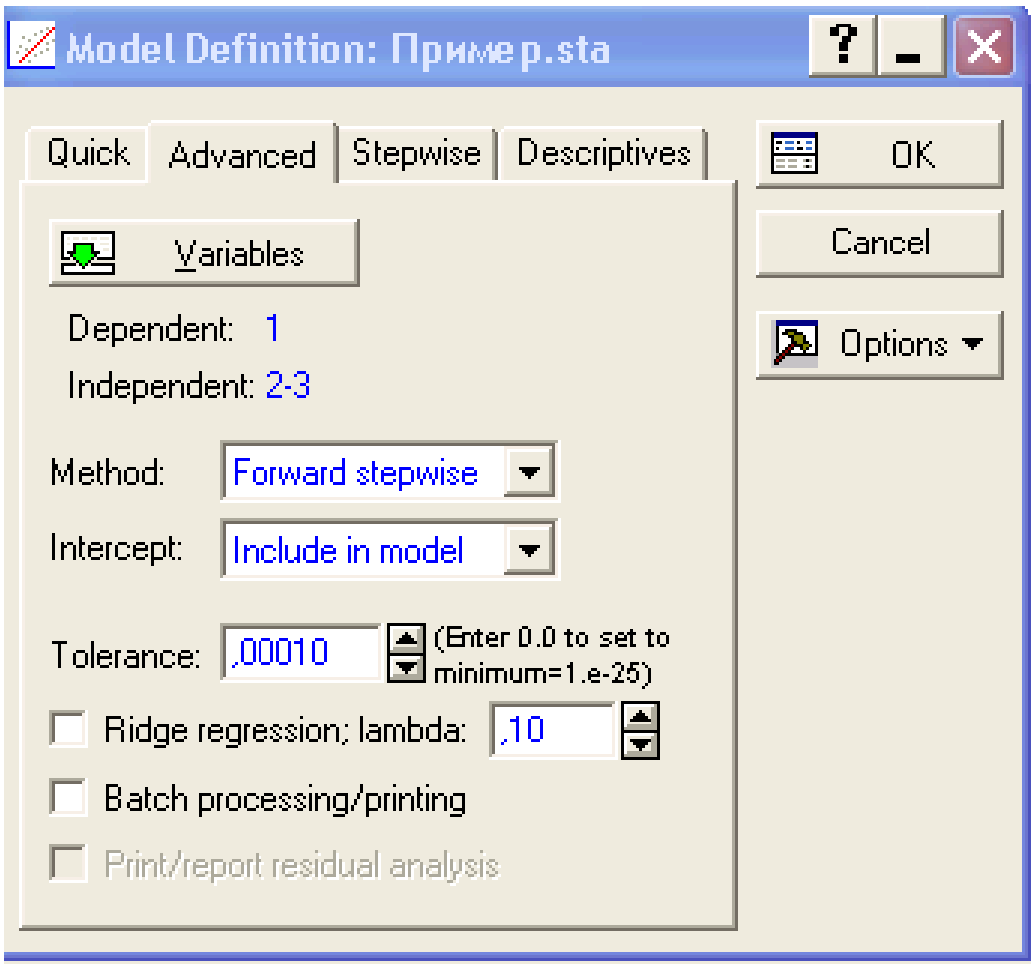

Рис. 4. Диалоговое окно для выбора модели

выбрать на вкладке «*Advanced*» выбрать метод отбора факторных признаков «*Method*»: «*Forward stepwise*» («Пошаговая регрессия с включением») либо «*Backward stepwise*» («Пошаговая регрессия с исключением»).

Дальнейшие действия выполняются либо на основе применения метода «*Forward stepwise*», либо на основе применения метода «*Backward stepwise*».

Выберем, например, метод «*Forward stepwise*».

На вкладке «*Stepwise*» автоматически будут установлены (могут быть изменены пользователем) численности включаемых или исключаемых переменных за один шаг (F to enter и F to remover) и число шагов («*Number of Steps*»).

Отображение результатов пошаговой регрессии («*Display results*») может быть как окончательным («*Summary only*»), так и пошаговым («*At each step*»). Нажать кнопку «*Ok*».

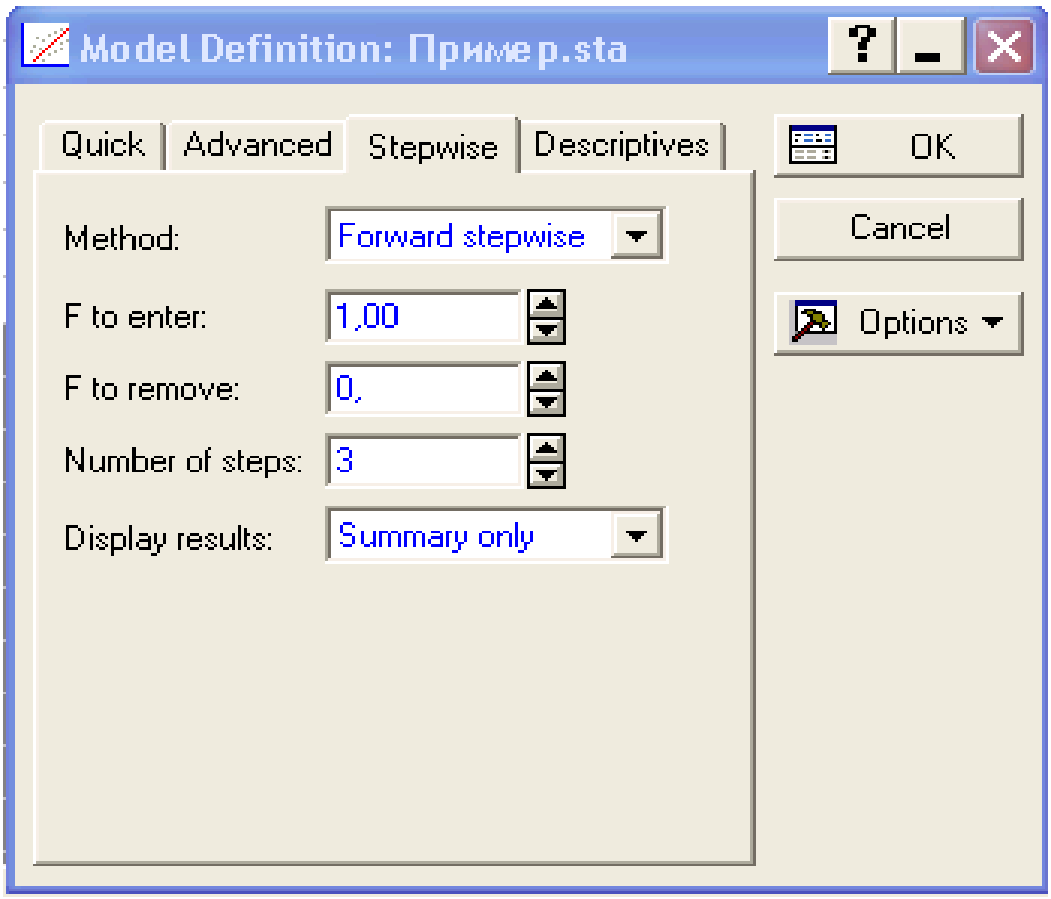

Рис. 5. Диалоговое окно для выбора метода

3. В появившемся окне «*Multiple Regression Results*» («Результаты множественной регрессии») указаны следующие основные результаты:
Step, число достаточных реализовавшихся шагов, равно 2, т.е решение найдено за две итерации;
Dependent, зависимая переменная (результативный признак) Y, расходы на питание, долл. в неделю;
No. of cases, количество наблюдений n, равное 16;
Multiple R, множественный коэффициент корреляции R, равный 0,99;
R2, множественный коэффициент детерминации R^2, равный 0,98, свидетельствует о достаточно высокой зависимости признаков;

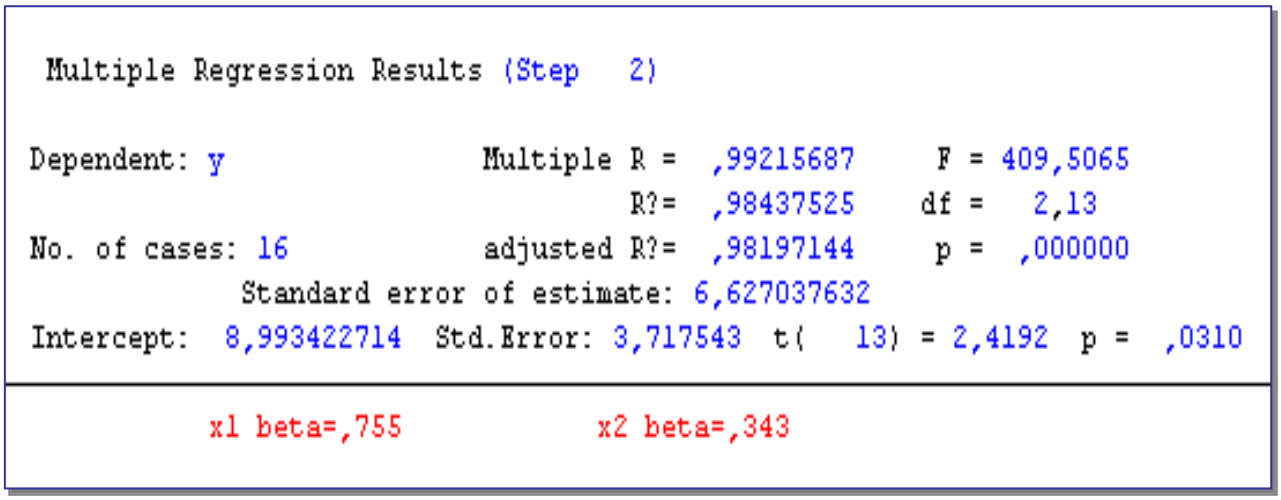

Рис. 6. Результаты

adjusted R2, множественный нормированный коэффициент детерминации R^2, равный 0,98;
Standard error of estimate, стандартная ошибка, т.е. среднеквадратическая ошибка оценки (стандартное отклонение фактических значений Y от теоретических значений результативного признака), равная 6,62;
F , F -критерий Фишера, равное 409,51;
df, количество степеней свободы, равное 2,13;
p , p -значение для F -критерия Фишера, равное 0;
Intercept , свободный член уравнения (сдвиг) a, равный 8,99;
Std.Error, стандартная ошибка a, равная 3,72;
t(13), значение t -критерия Стьюдента для a, равное 2,42;
p, p -значение для a, равное 0,03;
x1 beta, β -коэффициент при факторном признаке x1 равен 0,755;
x2 beta, β -коэффициент при факторном признаке x1 равен 0,343.

4. На вкладке «Advanced» окна «Multiple Regression Results» нажать кнопку «Summary: Regression Results».

В окне «*Regression Summary for Dependent Variable*» появится таблица, в которой отражены результаты проверки гипотез о значимости коэффициентов регрессии, где:

1) *Beta* - β -коэффициенты;

2) *Std.Err. of Beta* - стандартная ошибка β -коэффициентов;

3) *B* – параметры уравнения;
4) *Std.Err. of B* - стандартная ошибка оценки параметров;

5) *t( )* - t -критерий Стьюдента ($t_{факт}$);

6) *p-level* - p -значение (значимость оценки параметров).

Regression Summary for Dependent Variable: y (пример_1.sta)
R= ,99215687 R?= ,98437525 Adjusted R?= ,98197144
F(2,13)=409,51 p<,00000 Std.Error of estimate: 6,6270

N=16	Beta	Std.Err. of Beta	B	Std.Err. of B	t(13)	p-level
Intercept			8,993423	3,717543	2,41918	0,030952
x1	0,754870	0,042260	0,200311	0,011214	17,86243	0,000000
x2	0,343497	0,042260	6,930512	0,852656	8,12815	0,000002

Рис. 7. Результаты

Таким образом, получены значения параметров уравнения множественной регрессии в нормальном виде и проведена оценка их значимости. Красным шрифтом автоматически выделяются статистически значимые значения.

5. С целью проведения анализа остатков на вкладке «Residuals/assumptions/predictions» окна «*Multiple Regression Results*» нажать кнопку «*Perform Residual analysis*».

Появится окно «*Residual Analysis*».

На вкладке «*Advanced*» нажав кнопку («*Summary: Residuals&predicted*» получаем данные об остатках.

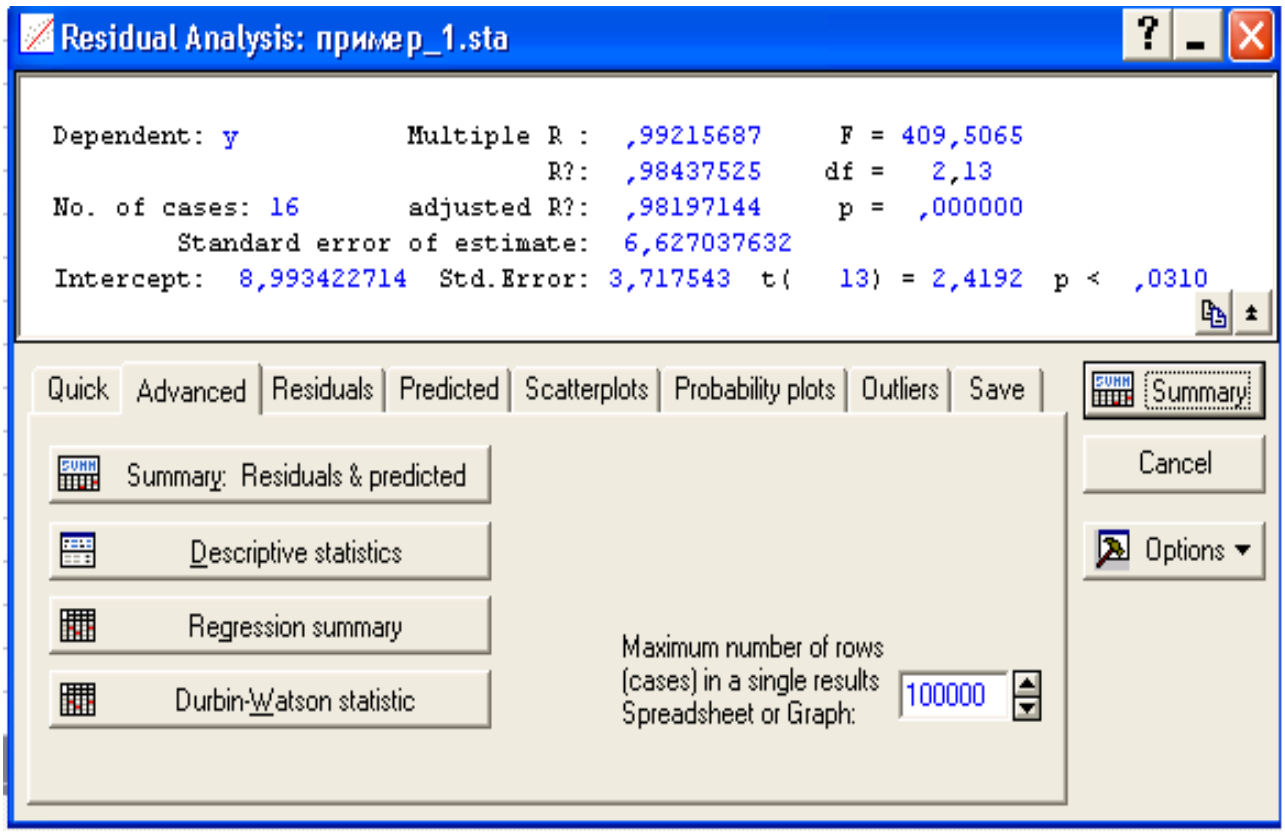

Рис. 8. Диалоговое окно, результаты

Case No.	Predicted & Residual Values (пример_1.sta) Dependent variable: y					
	Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val
1	21,8000	28,5418	-6,7418	-1,32563	-1,01732	2,892267
2	33,4000	35,1056	-1,7056	-1,19159	-0,25737	2,772824
3	50,3000	46,9231	3,3769	-0,95026	0,50956	2,574110
4	66,9000	66,3457	0,5543	-0,55363	0,08364	2,761226
5	47,3000	51,1937	-3,8937	-0,86305	-0,58755	2,266618
6	66,0000	63,3273	2,6727	-0,61527	0,40330	1,968124
7	81,0000	78,3051	2,6949	-0,30940	0,40666	1,880267
8	106,0000	104,7538	1,2462	0,23071	0,18805	2,505462
9	70,3000	74,4305	-4,1305	-0,38853	-0,62328	2,541212
10	94,6000	90,3633	4,2367	-0,06316	0,63931	1,937976
11	119,0000	112,7906	6,2094	0,39483	0,93698	1,789273
12	147,2000	144,0856	3,1144	1,03391	0,46995	2,836500
13	92,8000	102,8880	-10,0880	0,19261	-1,52224	4,222474
14	132,8000	127,2651	5,5349	0,69042	0,83520	3,250373
15	169,0000	159,1535	9,8465	1,34162	1,48580	3,008025
16	196,9000	209,8273	-12,9273	2,37643	-1,95069	4,870496
Minimum	21,8000	28,5418	-12,9273	-1,32563	-1,95069	1,789273
Maximum	196,9000	209,8273	9,8465	2,37643	1,48580	4,870496
Mean	93,4562	93,4562	-0,0000	0,00000	-0,00000	2,754827
Median	86,9000	84,3342	1,9594	-0,18628	0,29567	2,667668

Рис. 9. Результаты (наблюдаемое, предсказанное значения, остатки

На вкладке «*Advanced*» кнопка «*Durbin-Watson statistic*» позволяет рассчитать статистику Дарбина-Уотсона (**Ошибка! Источник ссылки не айден.**).

	Durbin-Watson d (пример_1.sta) and serial correlation of residuals	
	Durbin-Watson d	Serial Corr.
Estimate	2,088112	-0,325488

Рис. 10. Результат расчета статистики Дарбина-Уотсона

На вкладке «*Residuals*» возможен просмотр графического изображения остатков («*Histogram of Residuals*») (**Ошибка! Источник ссылки е найден.**).

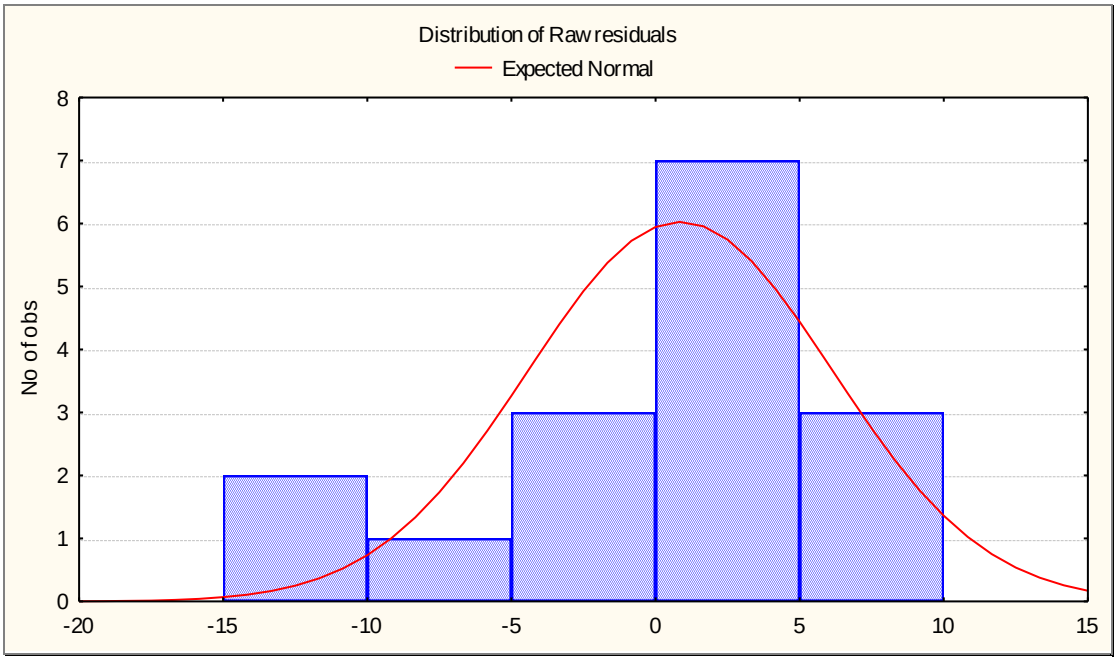

Рис. 11. Результат проверки нормальности остатков

6. Для предсказания значения результативного признака на вкладке «Residuals/assumptions/predictions» окна «*Multiple Regression Results*» нажать кнопку «*Predict dependent variable*», предварительно ниже установив уровень значимости для доверительного интервала (Alpha), например, 0,5.

В появившемся окне «Specify values for independent variables» следует задать прогнозируемые значения факторных признаков.

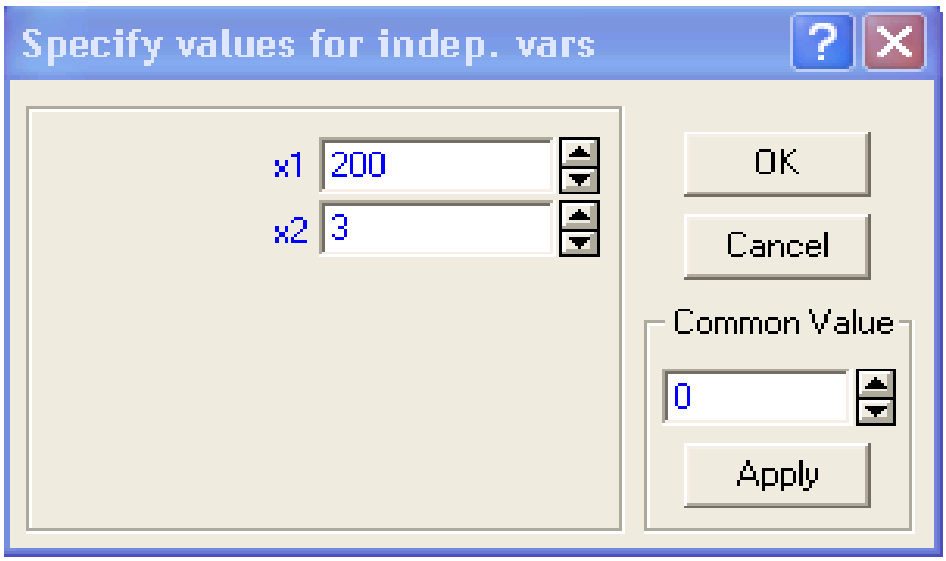

Рис. 12. Диалоговое окно для задания значений признаков

В окне «*Predicting Values*» будет указано предсказанное значение результативного признака (Predicted), а также нижняя (-) и верхняя (+) границы доверительного интервала с установленным уровнем значимости.

Variable	Predicting Values for (пример_1.sta) variable: y		
	B-Weight	Value	B-Weight * Value
x1	0,200311	200,0000	40,06228
x2	6,930512	3,0000	20,79153
Intercept			8,99342
Predicted			69,84724
-95,0%CL			65,45855
+95,0%CL			74,23592

Рис. 13. Результаты предсказания

7. Итоговое уравнение множественной регрессии будет иметь вид:

$$\hat{Y} = 8{,}993 + 0{,}2X_1 + 6{,}931X_2. \qquad (34)$$

5 АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

5.1 Аналитическое выравнивание временных рядов

При изучении изменений на основе фактических данных временного ряда (time series) уровни ряда рассматриваются как функция от времени, а задача выравнивания сводится к определению вида функции, отысканию ее параметров по эмпирическим данным и расчету теоретических уровней по выбранной формуле.

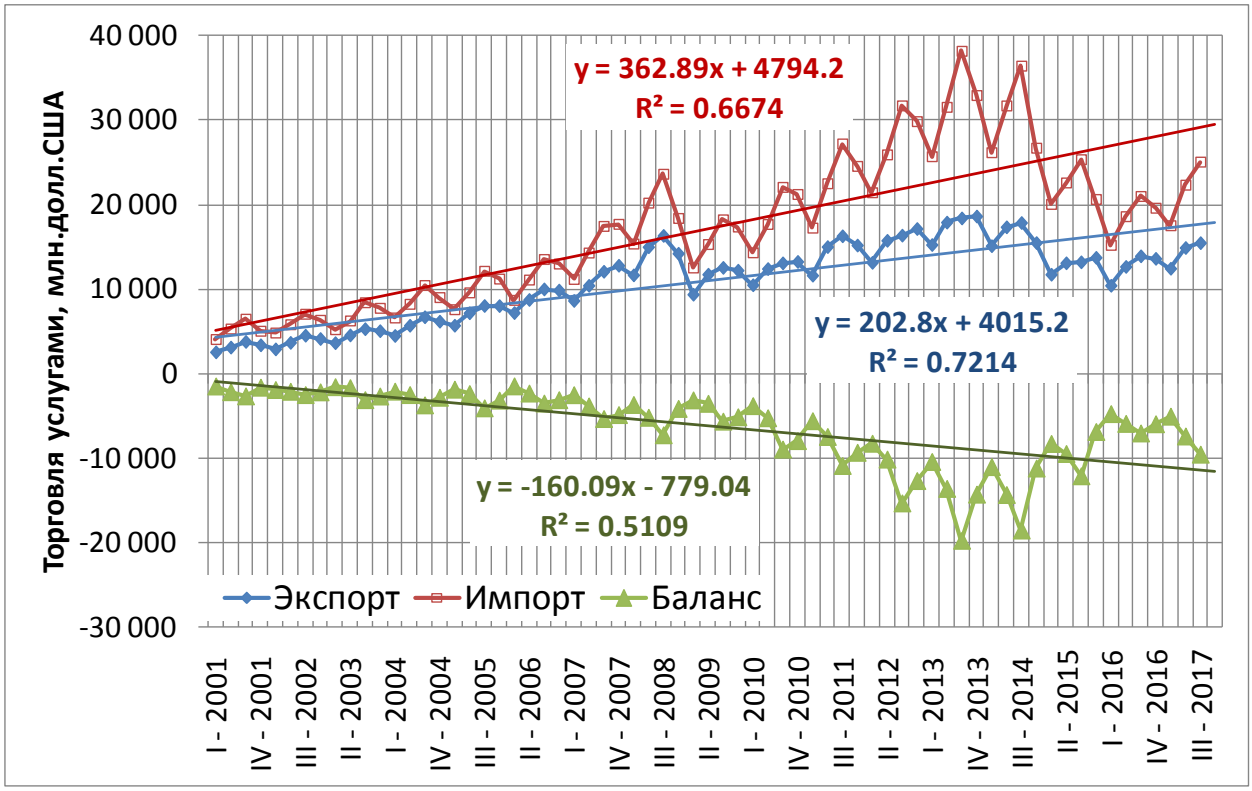

Рис. 14. Пример временного ряда, построенного поданным ЦРБ России (http://www.cbr.ru/statistics/?PrtId=svs) о поквартальной динамике внешней торговли РФ коммерческими услугами. Определены линейные тренды экспорта, импорта услуг и торгового баланса услугами.

Основными функциями, используемыми для выявления тенденции развития, являются следующие.

1. Линейная зависимость (применяется, если цепные абсолютные изменения относительно стабильны, не имеют отчетливой тенденции к росту или снижению, т.е. уровень изменяется с достаточно постоянной абсолютной скоростью):

$$\hat{y}_i = a + bt_i \,. \qquad (35)$$

2. Показательная функция (экспонента) (применяется в случае, когда цепные коэффициенты изменения относительно стабильны, т.е. уровень ряда изменяется приблизительно с постоянной относительной скоростью):

$$\hat{y}_i = ab^t \quad (36)$$

3. Парабола 2-го порядка (применяется в случае, когда абсолютные приросты равномерно возрастают или снижаются):

$$\hat{y}_i = a + bt_i + ct_i^2 \quad (37)$$

4. Гипербола:

$$\hat{y}_i = a + \frac{b}{t_i} \quad (38)$$

Кроме перечисленных функций, могут быть подобраны другие математические формулы, адекватно отражающие временные изменения анализируемого показателя. Зависимость уровней динамического ряда от фактора времени можно считать частным случаем корреляционной зависимости. Параметры уравнения, как правило, определяются методом наименьших квадратов (см.выше)

Это обеспечивает нахождение именно той заданной кривой, которая наилучшим образом отражает динамику фактического ряда. Использование уравнений позволяет выявить направление основной тенденции и дать ее числовую характеристику. Как правило, строится несколько моделей, из них выбирается та, которая наиболее адекватно отражает существующую динамику. Модели более высоких порядков будут точнее описывать исходный временной ряд. Однако при выборе трендовой модели из достаточно схожих уравнений стоит отдавать предпочтение более простой линии, т.к. надежность ее параметров, как правило, выше. Построенная динамическая модель может использоваться для прогнозирования дальнейшего развития изучаемого явления.

Самым простым уравнением и с точки зрения построения, и с точки зрения интерпретации является уравнение прямой. Построение линейного тренда

$$\hat{y}_i = a + bt_i \quad (39)$$

сводится к определению двух параметров: свободного члена a и коэффициента b. Расчет этих параметров осуществляется решением следующей системы уравнений:

$$\begin{cases} na + b\sum t_i = \sum y_i \\ a\sum t_i + b\sum t_i^2 = \sum y_i t_i \end{cases}, \quad (40)$$

где n – число уровней ряда;

ti – порядковый номер периода времени;

yi – эмпирические значения уровней ряда.

Если отсчет времени начинается с первого уровня ряда, то свободный член a – это значение показателя в начале отсчета. Коэффициент регрессии b показывает, на сколько в среднем в единицу времени происходит изменение уровней временного ряда в его единицах измерения. При выполнении расчетов вручную для их упрощения используют следующий прием – за начало отсчета выбирают серединный уровень ряда. Тогда сумма пронумерованных периодов становится равной нулю, и система уравнений для нахождения параметров уравнения существенно упрощается:

$$\begin{cases} na = \sum y_i \\ b\sum t_i^2 = \sum y_i t_i \end{cases}. \quad (41)$$

При этом свободный член a становится средним уровнем ряда.

Под параболическим трендом понимается, как правило, парабола второго порядка, но могут использоваться и параболы более высоких порядков. Чаще всего данное уравнение используется, если уровни ряда со временем увеличиваются или уменьшаются с ускорением. Обычно это какой-то ограниченный этап развития явления. Общий вид уравнения параболического тренда следующий:

$$\hat{y}_i = a + bt_i + ct_i^2. \quad (42)$$

Для нахождения параметров уравнения методом наименьших квадратов следует решить следующую систему нормальных уравнений:

$$\begin{cases} na + b\sum t_i + c\sum t_i^2 = \sum y_i \\ a\sum t_i + b\sum t_i^2 + c\sum t_i^3 = \sum y_i t_i \\ a\sum t_i^2 + b\sum t_i^3 + c\sum t_i^4 = \sum y_i\, t_i^2 \end{cases} \quad (43)$$

В случае выбора началом отсчета серединного уровня ряда расчеты могут быть существенно упрощены.

Показательная функция. Данный вид тренда используется в случае изменения уровней ряда в геометрической прогрессии, т.е. когда цепные коэффициенты изменения приблизительно стабильны. Общий вид уравнения:

$$\hat{y_i} = ab^{t_i}. \quad (44)$$

Если прологарифмировать обе части формулы, то получим линейную функцию

$$lg\hat{y_i} = lga + t_i lgb. \quad (45)$$

Отсюда, заменив уровни ряда их логарифмами, можно найти параметры a и b через их логарифмы с помощью системы уравнений:

$$\begin{cases} nlga + lgb\sum t_i = \sum lgy_i \\ lga\sum t_i + lgb\sum t_i^2 = \sum t_i lgy_i \end{cases} \quad (46)$$

При отсчете от серединного уровня, когда $\sum t_i = 0$, расчет параметров существенно упрощается:

$$\begin{cases} nlga = \sum lgy_i \\ lgb\sum t_i^2 = \sum t_i lgy_i \end{cases} \quad (47)$$

Рассмотрим выравнивание временного ряда по показательной функции на примере динамики экспорта РФ за 2002-2008 гг. В этот период наблюдался рост объема экспорта, причем от года к году цепные коэффициенты были

приблизительно одинаковы. В таблице 9 даны промежуточные вычисления. Итоговые значения подставим в систему уравнений для нахождения параметров показательной функции.

Таблица 9 - Расчет параметров показательной функции

Годы	y_i	$k_{цеп}$	$\lg y_i$	t_i	t_i^2	$t_i lg y_i$	$\hat{y}_i$
2002	106,7	-	2,028164	1	1	2,028164	108,5
2003	133,7	1,25	2,126131	2	4	4,252263	138,7
2004	181,6	1,36	2,259116	3	9	6,777348	177,4
2005	241,5	1,33	2,382917	4	16	9,531669	226,8
2006	301,2	1,25	2,478855	5	25	12,39427	290,1
2007	351,9	1,17	2,546419	6	36	15,27852	370,9
2008	468,1	1,33	2,670339	7	49	18,69237	474,3
Итого	1784,7		16,49194	28	140	68,9546	1786,7

$$\begin{cases} 7\lg a + 28\lg b = 16{,}49 \\ 28\lg a + 140\lg b = 68{,}95 \end{cases} \quad (48)$$

Отсюда, $\lg b = 0{,}10678$, т. е. $b = 1{,}2787$; $\lg a = 1{,}92859$, т. е. $a = 84{,}837$.

Следовательно, $\hat{y}_i = 84{,}8379 \cdot 1{,}2787^{t_i}$. (82)

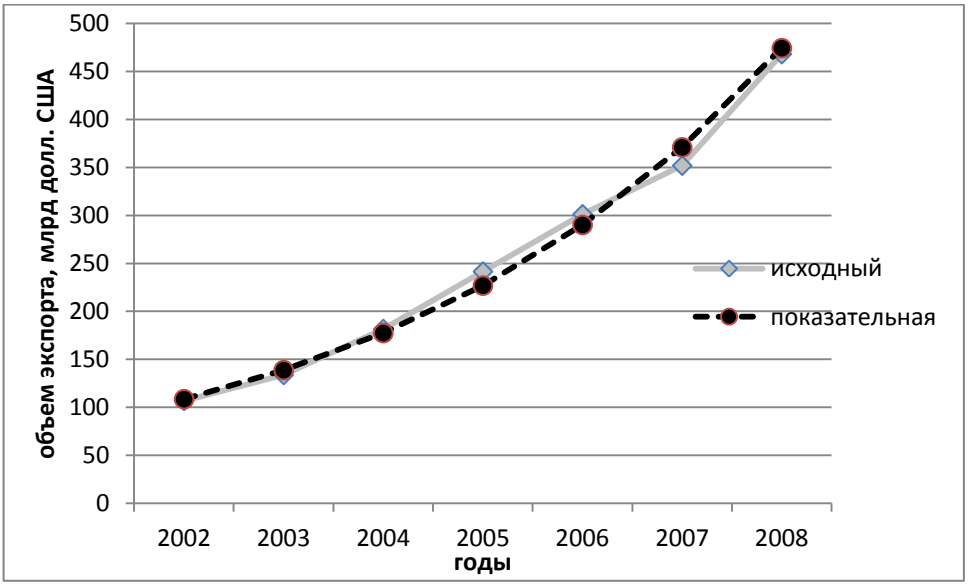

Рис. 15. Динамика экспорта РФ за 2002-2008 гг.

График исходного и выравненного динамического ряда наглядно подтверждают соответствие выбранной функции.

Применение уравнения гиперболы в качестве трендовой модели дает наилучшие результаты, в случаях рассмотрения этапов развития явления, когда анализируемый показатель сначала резко снижается, а затем продолжает

уменьшаться меньшими темпами. При этом этот показатель не может принимать нулевые или отрицательные значения.

Для определения параметров уравнения гиперболы

$$\hat{y}_i = a + b\frac{1}{t_i} \quad (49)$$

необходимо решить систему нормальных уравнений:

$$\begin{cases} na + b\sum\frac{1}{t_i} = \sum y_i \\ a\sum\frac{1}{t_i} + b\sum\left(\frac{1}{t_i}\right)^2 = \sum\frac{y_i}{t_i} \end{cases} \quad (50)$$

Рассмотрим построение гиперболического тренда на условном примере динамики расхода материала на одно изделие (м2).

Таблица 10 - Расчет параметров гиперболической функции

Годы	yi	ti	1/ti	(1/ti)2	yi/ti	$\hat{y}_i$
2005	140	1	1,00	1,00	140,0	138,6
2006	96	2	0,50	0,25	48,0	101,0
2007	86	3	0,33	0,11	28,7	88,5
2008	82	4	0,25	0,06	20,5	82,2
2009	80	5	0,20	0,04	16,0	78,4
2010	78	6	0,17	0,03	13,0	75,9
2011	77	7	0,14	0,02	11,0	74,1
Итого	639	28	2,59	1,51	277,2	638,8

$$\begin{cases} 7a + 2{,}59b = 639 \\ 2{,}59a + 1{,}5b = 277{,}2 \end{cases}. \quad (51)$$

Отсюда, $\hat{y}_i = 63{,}4 + 75{,}2\frac{1}{t_i}$. Графическое изображение исходного ряда и гиперболического тренда представлено на рис. 6.

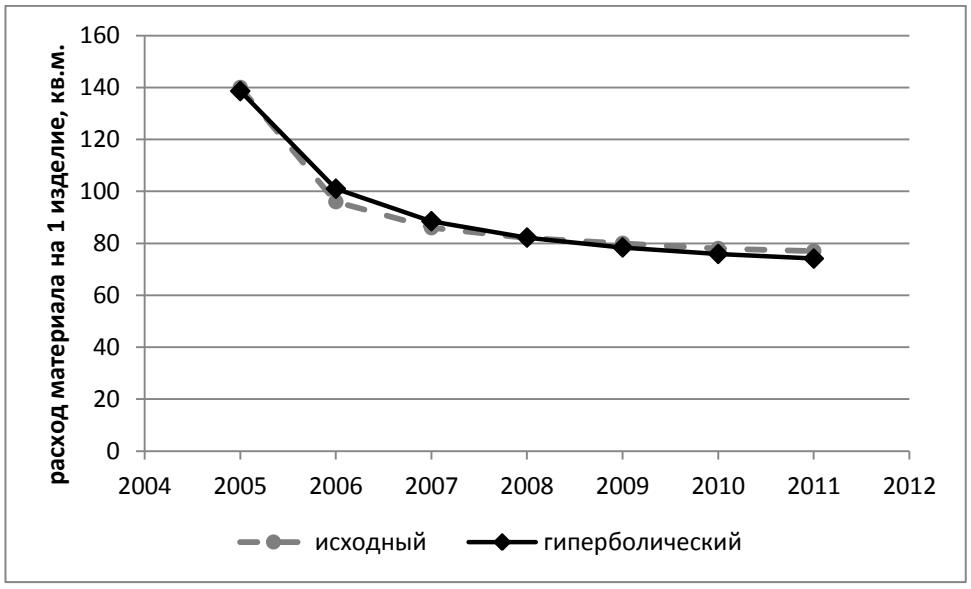

Рис. 16. Динамика расхода материала на одно изделие

Следует обратить внимание на то, что значение свободного члена в уравнении является предельным значением для данного показателя, т. е. расход материала на одно изделие не может быть меньше 63,4 м2.

5.2 Устойчивость временного ряда

Первоначально следует определить, что именно включается в понятие устойчивости временного ряда. При анализе динамики исследователь стремится выделить изменения показателя под влиянием основополагающих причин и под влиянием случайных факторов. Однако это разделение – условный прием. Динамика показателя, как правило, включает в себя направленные изменения (тенденцию) и случайные колебания вокруг этого тренда. Отсюда возникает необходимость отдельного рассмотрения устойчивости уровней ряда и устойчивости тенденции динамики. Можно сказать, что устойчивость временного ряда обеспечивается наличием тренда, тенденции изменения и минимизацией колебаний фактических уровней временного ряда около этого тренда. Достижение устойчивого роста результативных показателей часто является основной задачей в экономике и других областях человеческой деятельности.

Самой простой оценкой устойчивости уровней временного ряда, по аналогии с размахом вариации, является размах колеблемости средних уровней за благоприятные и неблагоприятные периоды времени. К благоприятным периодам относятся те периоды, когда уровни были выше трендовых, к неблагоприятным – периоды с уровнями ниже трендовых.

$$R = \bar{y}_{\text{благ}} - \bar{y}_{\text{неблаг}}, \quad (52)$$

где $\bar{y}_{\text{благ}}$ - средняя величина из уровней за благоприятные периоды;

$\bar{y}_{\text{неблаг}}$- средняя величина из уровней за неблагоприятные периоды.

В качестве оценки устойчивости уровней может использоваться и соотношение средних уровней за благоприятные и неблагоприятные периоды. Этот показатель называется индексом устойчивости уровней динамического ряда, чем ближе его значение к единице, тем меньше колеблемость, а, значит, выше устойчивость.

$$i_{\bar{y}} = \frac{\bar{y}_{\text{благ}}}{\bar{y}_{\text{неблаг}}} \quad (53)$$

К обобщающим абсолютным показателям отклонений фактических уровней от тренда относят среднее линейное отклонение и среднее квадратическое отклонение:

$$\bar{l}_t = \frac{\sum|y_i - \hat{y}_i|}{n-p}; \quad (54)$$

$$\sigma_t = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-p}}, \quad (55)$$

где yi — фактический уровень;

$\hat{y}_i$ — выравненный уровень;

n — количество уровней ряда;

p — число параметров уравнения.

Среднее квадратическое отклонение часто называют точностью модели. Оба показателя являются абсолютными величинами, характеризующими колеблемость фактических уровней около тренда, имеющими те же единицы измерения, что и сам признак. Для сравнения степени колеблемости по показателям с разными единицами измерения используются относительные показатели. Они рассчитываются соотношением абсолютных значений со средним уровнем временного ряда:

коэффициент линейной колеблемости

$$V_{\bar{l}} = \frac{\bar{l}_t}{\bar{y}} \cdot 100\% , \quad (56)$$

коэффициент колеблемости

$$V_\sigma = \frac{\sigma_t}{\bar{y}} \cdot 100\% . \quad (57)$$

На основе коэффициента колеблемости определяют коэффициент устойчивости:

$$K_{уст} = 100 - V_\sigma .$$

Если коэффициент колеблемости составил 10%, то коэффициент устойчивости соответственно равен 90%. Это означает, что среднее колебание относительно среднего уровня составляет 10%. При этом следует помнить, что вероятность того, что конкретные колебания не превысят среднеквадратического отклонения составляет 68,3%, если распределение колебаний по их величине близко к нормальному распределению.

Для оценки устойчивости тенденции динамики чаще всего используется коэффициент рангов Спирмена, рассчитываемый по формуле:

$$\rho = 1 - \frac{6\sum \Delta^2}{n^3 - n} , \quad (58)$$

где Δ – разница рангов; n – количество уровней ряда.

Прежде чем использовать формулу необходимо, пронумеровать периоды времени и уровни ряда в порядке возрастания, т. е. каждому периоду времени и каждому уровню присваивается номер в порядке возрастания. В случае совпадения значений уровней ряда им присваивается ранг, равный частному от деления суммы рангов, приходящихся на эти значения, на число совпадающих значений.

Коэффициент рангов может принимать значения от 0 до 1 по абсолютному значению. Значение коэффициента +1 означает, что ранги периодов и ранги уровней совпадали, т.е. с ростом номеров периодов увеличивались ранги уровней, следовательно, имеет место устойчивый, непрерывный рост. Если коэффициент рангов равен нулю, то это свидетельство отсутствия устойчивого роста. При коэффициенте -1 следует говорить об устойчивом снижении показателя. Реальные значения коэффициента рангов Спирмена находятся между этими значениями. По их приближенности к единице (или минус единице) следует делать вывод об устойчивом росте (или снижении).

Для оценки устойчивости тенденции может быть использован индекс корреляции:

$$I_r = \sqrt{1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}}, \quad (59)$$

где y_i – фактические уровни; $\hat{y}_i$ – выравненные уровни; $\bar{y}$ – средний уровень ряда.

Индекс корреляции показывает степень сопряженности колебаний фактических уровней с колебаниями теоретических уровней, происходящих под влиянием комплекса основополагающих факторов.

Важно отметить, что при оценке устойчивости временного ряда следует одновременно использовать показатели, характеризующие устойчивость уровней ряда, и показатели устойчивости тенденции динамики. На практике даже при полной устойчивости роста может присутствовать колеблемость уровней (коэффициент устойчивости меньше 100%).

5.3 Сезонность и ее измерение

Одним из компонентов изменения уровней динамического ряда является кратковременное систематическое движение. Чаще всего это сезонные колебания, характеризующиеся постоянным повышением и понижением уровней ряда в одни и те же периоды года (месяцы, кварталы). Сезонность свойственна многим явлениям, особенно заметно она проявляется в сельском хозяйстве, туристическом бизнесе, а также в торговле. Для стабильности развития желательно снижение сезонных колебаний, однако, это часто невозможно, поэтому задача сводится к тому, чтобы учитывать эти колебания, а, следовательно, уметь их измерять. Поэтому выявление степени колеблемости уровней ряда по сезонам с целью ее учета в практической деятельности представляет собой актуальную задачу исследования временных изменений.

Выявление наличия сезонных колебаний может начинаться с графического представления исходных данных. Для этого используют линейный график в системе координат. По нему визуально делается предположение о наличии сезонности. Кроме того, графически сезонные колебания можно изобразить радиальной диаграммой. Этот вид графика хорошо демонстрирует смещение значений показателя по сезонам относительно среднего (стабильного) уровня.

Существуют различные способы измерения сезонных колебаний. В основе оценки степени сезонности лежит сопоставление фактических уровней со средним или сглаженными уровнями. Основные показатели сезонности можно подразделить на два вида: 1) показатели формы сезонной волны; 2) показатели силы сезонных колебаний.

К первой группе относятся:

- абсолютные отклонения уровня каждого отдельного месяца от среднемесячного уровня за год:

(60)

(61)

$$\Delta_i = y_i - \bar{y};$$

- отношение уровня отдельного месяца к среднему уровню за год, выраженное в процентах:

$$i_i = \frac{y_i}{\bar{y}} \cdot 100\%;$$

этот показатель называется индексом сезонности (существуют другие способы его расчета).

Силу сезонных колебаний показывает коэффициент сезонности:

$$V_c = \frac{\sigma_c}{\bar{y}} \cdot 100\%$$

, где $\sigma_c = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n}}$,

где V_c – коэффициент сезонности;

σ_c – среднеквадратическое отклонение фактических значений от среднего уровня;

n – количество уровней ряда.

Принято считать, что

Коэффициент сезонности, %	Колеблемость по сезонам
меньше 10	слабая
10-20	средняя
20-40	сильная
больше 40	очень сильная

Сезонные изменения могут быть наглядно представлены радиальной диаграммой (рис. 7), где отчетливо наблюдается смещение фактических значений вниз вправо, что означает увеличение количества проданных велосипедов в весенние и летние месяцы.

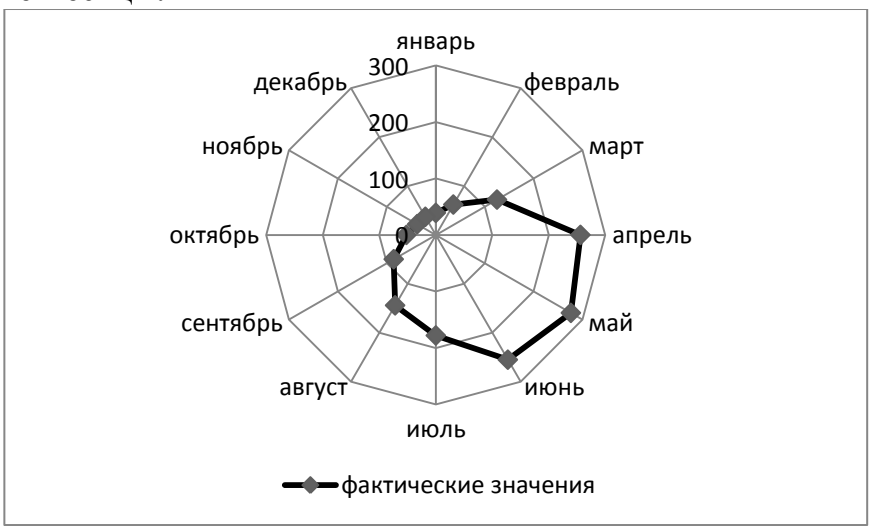

Рис. 17. Динамика объема продаж роликовых коньков в магазинах области, тыс. шт.

5.4 Аналитическое выравнивание сезонных колебаний с помощью ряда Фурье

При выявлении периодических, в частности сезонных колебаний, может использоваться метод аналитического выравнивания рассмотренный в п.12.3.3. В этом случае наиболее подходящей формулой для моделирования временных изменений является ряд Фурье. Функция зависимости уровней временного ряда от времени может быть представлена следующим образом:

$$\hat{y}_t = a_0 + \sum_{k=1}^{m}(a_k \cos kt + b_k \sin kt), \quad (63)$$

где $\hat{y}_t$ – выравненные значения;

t – показатель времени;

k – число гармоник;

$a_0, \ldots, a_m$ – параметры уравнения.

При построении ряда Фурье получают гармоники – синусоиды, представляющие собой отражение гармонических колебаний. Число гармоник может быть различным. Как правило, строится несколько гармоник, из которых выбирается та, которая наиболее точно отражает исследуемые сезонные колебания (исходный ряд данных). Параметры уравнения определяются методом наименьших квадратов.

$$a_0 = \frac{\sum y}{n}, \quad (64)$$

$$a_k = \frac{2\sum y \cos kt}{n}, \quad (65)$$

$$b_k = \frac{2\sum y \sin kt}{n}. \quad (66)$$

Процедура выравнивания по ряду Фурье заключается в построении нескольких гармоник, накладывающихся одна на другую, до получения удовлетворяющей поставленной задаче.

При k = 1 ряд Фурье будет иметь следующий вид:

$$\hat{y}_{1t} = a_0 + a_1 \cos t + b_1 \sin t; \quad (67)$$

при k = 2:

$$\hat{y}_{2t} = a_0 + a_1 \cos t + b_1 \sin t + a_2 \cos 2t + b_2 \sin 2t; \quad (68)$$

при k = 3:

$$\hat{y}_{3t} = a_0 + a_1 \cos t + b_1 \sin t + a_2 \cos 2t + b_2 \sin 2t + a_3 \cos 3t + b_3 \sin 3t; \quad (69)$$

и т. д.

Значения времени t задают следующим образом. Если весь цикл изменения составляет 360°, то прирост времени от периода к периоду должен составлять 360°/n или 2π/n (где n – число уровней ряда). Т.е. при n=12 значения t будут принимать следующие значения:

0; π/6; π/3; π/2; 2π/3; 5π/6; π; 7π/6; 4π/3; 3π/2; 5π/3; 11π/6. (70)

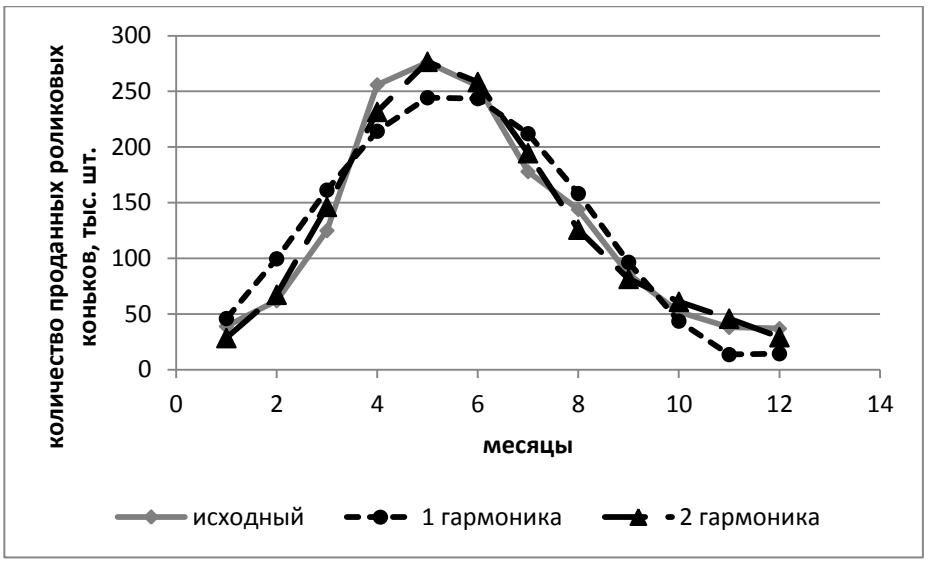

Рис. 18. Выравнивание по ряду Фурье

Более надежная оценка сезонности получается, когда анализируют данные за несколько лет. В этом случае индексы сезонности рассчитывают, задавая отношение среднего уровня за каждый месяц и среднемесячного уровня за весь анализируемый период.

Если уровни ряда растут или снижаются, т. е. наблюдается тенденция к увеличению (уменьшению) уровней ряда от года к году, то расчет индексов сезонности осуществляется иначе. Необходимо учесть и изменения внутри годичные (сезонные), и изменения, связанные с направлением тренда. Сначала производится выравнивание исходного динамического ряда по выбранному уравнению. Через соотношение фактических уровней и уровней, выравненных по уравнению тренда, получают индексы сезонности для каждого месяца анализируемого периода. Затем по одноименным месяцам из полученных индексов рассчитывают средний индекс сезонности, произведение которого на теоретическое значение уровня дает выравненное значение показателя с учетом сезонности.

5.5 Прогнозирование на основе экстраполяции тренда

Наиболее обоснованный подход — использование динамических (трендовых) моделей. Сама построенная модель дает оценку тенденции изменения показателя за исследуемый период времени. Предполагая, что эта тенденция сохранится в будущем, полученная трендовая модель используется для расчета прогнозируемого значения. Следует обратить внимание на то, что при прогнозировании предпочтение отдается простым моделям, содержащим меньшее количество параметров. Чаще всего используется линейный тренд. Заметим, что получаемое точечное значение вряд ли будет достигнуто в точности, поскольку фактические уровни, как правило, не совпадают с трендом, а колеблются около него. Именно поэтому прогноз может даваться с учетом этих колебаний: в интервале от прогнозируемого значения минус

точность модели до прогнозируемого значения плюс точность модели. Такой прогноз учитывает и направление тренда, и колеблемость уровней вокруг него.

Рассмотрим процедуру выполнения прогноза на примере урожайности зерновых в РФ.

Таблица 11 - Расчет линейного тренда урожайности зерновых

Годы	y_i	t_i	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	2	3	4	5
2000	15,6	1	17,4	3,3
2001	19,4	2	17,8	2,5
2002	19,6	3	18,2	2,0
2003	17,8	4	18,6	0,6
2004	18,8	5	19,0	0,0
2005	18,5	6	19,4	0,8
2006	18,9	7	19,8	0,8
2007	19,8	8	20,2	0,1
2008	23,8	9	20,6	10,5
2009	22,7	10	21,0	3,0
2010	18,3	11	21,4	9,3
Итого	213,2		213,2	33,0

Сначала на основе данного динамического ряда (табл. 6) рассчитаем параметры линейного уравнения. В результате вычислений получим уравнение прямой:

$$\hat{y}_i = 17{,}015 + 0{,}3945 \cdot t_i. \quad (71)$$

Затем рассчитаем теоретические значения урожайности зерновых (графа 4 табл.6), квадраты их отклонений от фактических значений (графа 5 табл.6) и среднее квадратическое отклонение фактических значений от теоретических (точность модели).

$$\sigma_t = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - p}} = \sqrt{\frac{33}{11 - 2}} = 1{,}91 \quad (72)$$

.

Точечное прогнозируемое значение урожайности зерновых на 2011 г. составит:

$$\hat{y}_{2011} = 17{,}015 + 0{,}3945 \cdot 12 = 21{,}749 \approx 21{,}75 \text{ ц/га} \quad (73)$$

Прогнозируемый диапазон с учетом точности модели:

$$\hat{y}_{2011} - \sigma_{y(t)} \leq \hat{y}_{\text{прогн}} \leq \hat{y}_{2011} + \sigma_{y(t)}, \quad (74)$$

$$21{,}75 - 1{,}91 \leq \hat{y}_{\text{прогн}} \leq 21{,}75 + 1{,}91, \quad (75)$$

$$19{,}84 \leq \hat{y}_{\text{прогн}} \leq 23{,}66. \quad (76)$$

Таким образом, урожайность зерновых в РФ в 2011 г. составит значение из интервала от 19,84 до 23,66 ц/га.

Существенное влияние на достоверность прогноза оказывает длина исходного временного ряда, на основе которого строится прогноз, а также удаленность прогнозируемого года от исходного ряда. Как правило, прогноз рассчитывается на длину, не превышающую трети длины исходного ряда.

В рассмотренном примере доверительный интервал прогноза урожайности зерновых в РФ вычисляется следующим образом:

средняя ошибка прогноза положения линейного тренда

$$m_{\hat{y}_k} = \sigma_t \cdot \sqrt{\frac{1}{n} + \frac{t_k^2}{\sum t_i^2}} = 1{,}91 \cdot \sqrt{\frac{1}{11} + \frac{12^2}{506}} = 1{,}1651 \quad (77)$$

,

значение t-критерия Стьюдента при 9-ти степенях свободы 9 (11-2) и 0,05 уровне значимости составляет 2,262

предельная ошибка отсюда $\Delta_{\hat{y}_k} = m_{\hat{y}_k} t = 1{,}1651 \cdot 2{,}262 = 2{,}6354 \approx 2{,}64$ (78)

значит, доверительный интервал прогноза показателя следующий:

$$19{,}11 \leq \hat{y}_{\text{прогн}} \leq 24{,}39 \quad (79)$$

Таким образом, с вероятностью 0,95 можно утверждать, что урожайность зерновых в РФ в 2011 г. будет составлять значение из интервала от 19,11 до 24,39 ц/га, если тенденция динамики данного показателя останется неизменной. Фактически урожайность зерновых в 2011 г. составила 22,4 ц/га, что входит в рассчитанный доверительный интервал прогноза.

Чем удаленнее прогнозируемый уровень от последнего известного значения динамического ряда, тем менее точным становится прогноз, рамки доверительного интервала раздвигаются. Если на основе того же временного ряда рассчитать доверительный интервал прогноза урожайности на 2012 год, то будет получен следующий диапазон:

$$\hat{y}_{2012} = 17{,}015 + 0{,}3945 \cdot 13 = 22{,}14 \text{ ц/га}, \quad (80)$$

$$m_{\hat{y}_k} = \sigma_t \cdot \sqrt{\frac{1}{n} + \frac{t_k^2}{\sum t_i^2}} = 1{,}91 \cdot \sqrt{\frac{1}{11} + \frac{13^2}{506}} = 1{,}2415 \quad (81)$$

, (82)

$$\Delta_{\hat{y}_k} = m_{\hat{y}_k} t = 1{,}2415 \cdot 2{,}262 = 2{,}808273 \approx 2{,}81, \quad (83)$$

$$19{,}33 \leq \hat{y}_{\text{прогн}} \leq 24{,}95. \quad (84)$$

Как видно из расчетов, ошибка прогноза увеличивается, и доверительный интервал становится шире. С вероятностью 0,95 можно утверждать, что урожайность зерновых в РФ в 2012 г. будет составлять значение из интервала от 19,33 до 24,95 ц/га, если тенденция динамики данного показателя останется неизменной. Однако этот прогноз не оправдался. 2012 г. оказался неблагоприятным с точки зрения погодных условий, фактическая урожайность зерновых составила 18,3 ц/га, что является одним из самых низких значений показателя за весь рассмотренный период.

6 ЭКОНОМЕТРИЧЕСКИЙ АНАЛИЗ ДИФФЕРЕНЦИАЦИИ РЕГИОНОВ РФ ПО ПОКАЗАТЕЛЯМ ЗАНЯТОСТИ И БЕЗРАБОТИЦЫ НАСЕЛЕНИЯ

Цель приведенного ниже исследования состояла в выявлении и анализе социально-экономических факторов занятости и безработицы в регионах России с использованием многомерных статистических методов.

Использованы статистические данные, предоставляемые Федеральной службой государственной статистики[1].. Методики расчета анализируемых статистических показателей можно найти в официальных комментариях Росстата [2].

6.1 Построение линейных трендов для показателей уровня безработицы по федеральным округам и отдельным регионам РФ

Создадим электронную таблицу в пакете STATISTICA с 8 столбцами и 8 случаями. Столбцы назовем соответственно по названиям федеральных округов и первый столбец Year. Для каждого из них выберем тип – числовой (Number), а количество цифр после запятой (Decimal places) – 0, 1, 1, 1, 1, 1, 1, 1. Прежде чем строить регрессионную модель посмотрим на соответствующие графики зависимости уровня безработицы от времени.

Проанализируем динамику уровня безработицы для РФ в целом, федеральных округов и отдельных регионов РФ за период с 2000 по 2008 годы (рис. 1).

Для того чтобы построить график зависимости в пакете STATISTICA, выполним следующую операцию. Построим вначале график зависимости уровня безработицы в центральном федеральном округе от времени: Graphs → 2D Graphs → Line Plots (Variables) → Advanced → Graph Type: XY Trace → Variables: X = Year, Y = ЦФО → OK → OK.

[1] www.gks.ru

[2]Методологические пояснения. - Росстат, http://www.gks.ru/free_doc/2007/metod_rus_fig/05-35.htm

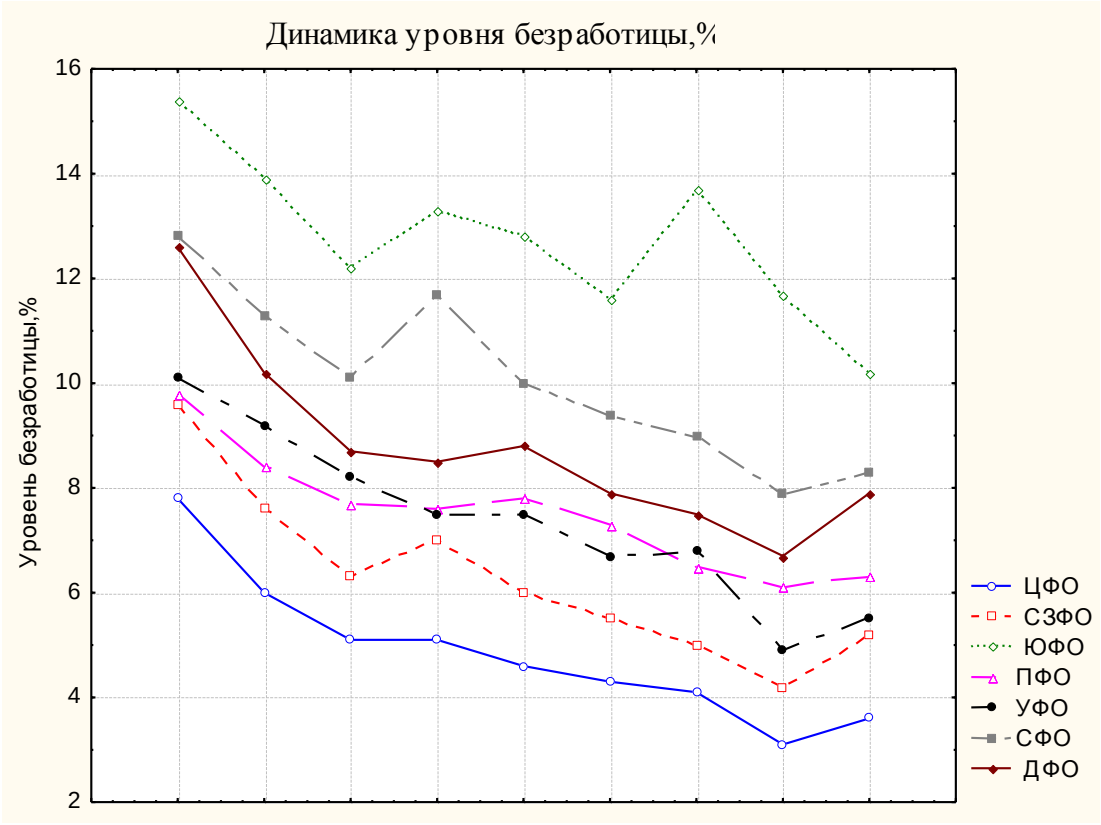

Рис. 19. Динамика уровня безработицы, %

Чтобы убрать номера наблюдений Case 1, Case 2, …, необходимо кликнуть по одному из них дважды и убрать в появившемся меню галочку напротив надписи Text labels → OK. Теперь нужно поместить в то же окно еще 6 графиков – зависимость уровня безработицы в других федеральных округах от времени. Правая кнопка мыши (ПКМ) → Graph Data Editor → (ПКМ) → Add Plot → Type: Line Plot → OK. Копируем столбец Year в буфер и вставляем этот столбец на место столбца New1. В столбец New2 вставляем столбец СЗФО и повторяем эту операцию, добавляя на график оставшиеся федеральные округа.

Для изучения связи между уровнем безработицы в отдельных федеральных округах и временем построим диаграмму рассеивания. Graphs → 2D Graphs → Scatterplots → Variables: X = Year, Y = ЦФО → OK → OK. Диаграмма рассеивания приведена на рисунке. Там же представлена соответствующая регрессионная линия.

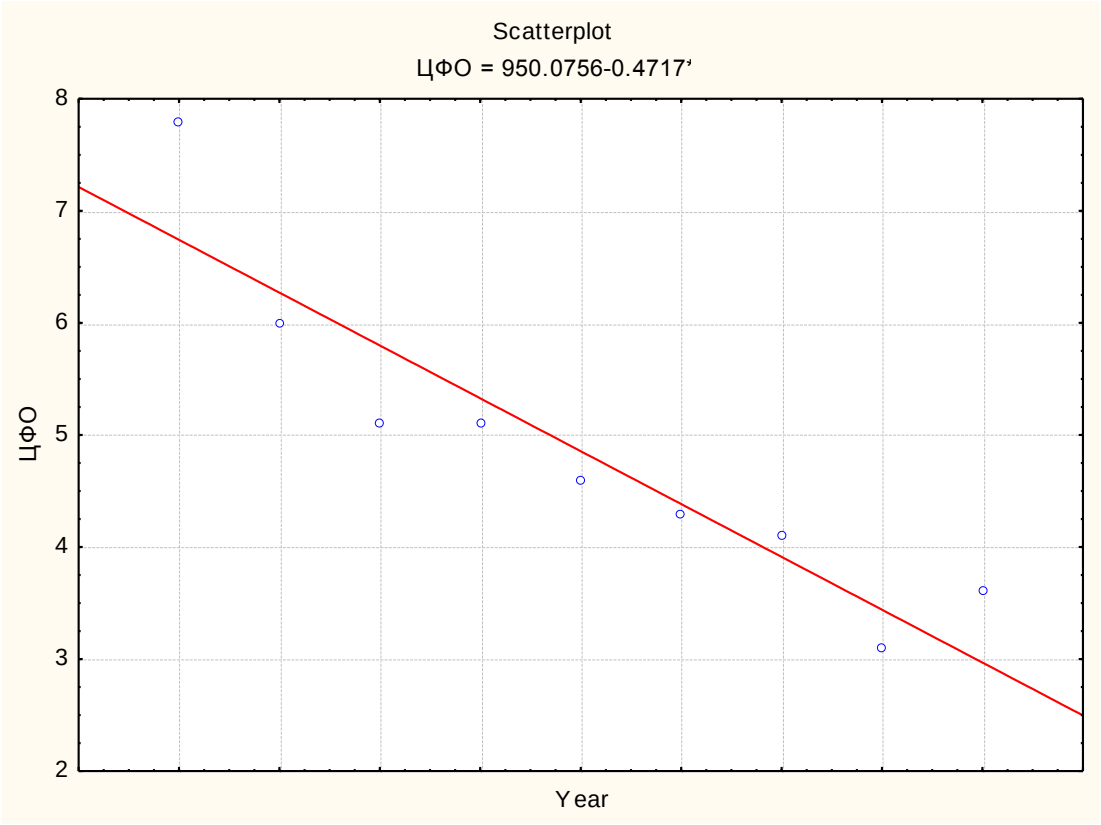

Рис. 20Диаграмма рассеивания и регрессионная прямая

Для всех трендов (по федеральным округам) значения коэффициентов регрессии отрицательны, значения коэффициентов детерминации выше 0,7. Для Республики Ингушетия отмечается устойчивый рост уровня безработицы, характеризующийся положительным линейным трендом.

Для построения диаграммы рассеивания для двух линейных трендов необходимо осуществить следующие действия: Graphs → 2D Graphs → Scatterplots → Graph Type: Multiple → Variables: X = Year, Y = Ing, Dag → OK → OK. Определим значимость коэффициентов регрессионной модели:

Ing = -5943,6267+2,99*x

Для определения значимости коэффициентов регрессионной модели обратимся к модулю множественной регрессии в пакете STATISTICA: Statistics → Multiple Regression → Variables: Dependent var.: Ing, Independent variable list: Year → OK → OK → меню «Multiple Regression Results…» → Summary: Regression results. Итоговое меню приведено в таблице.

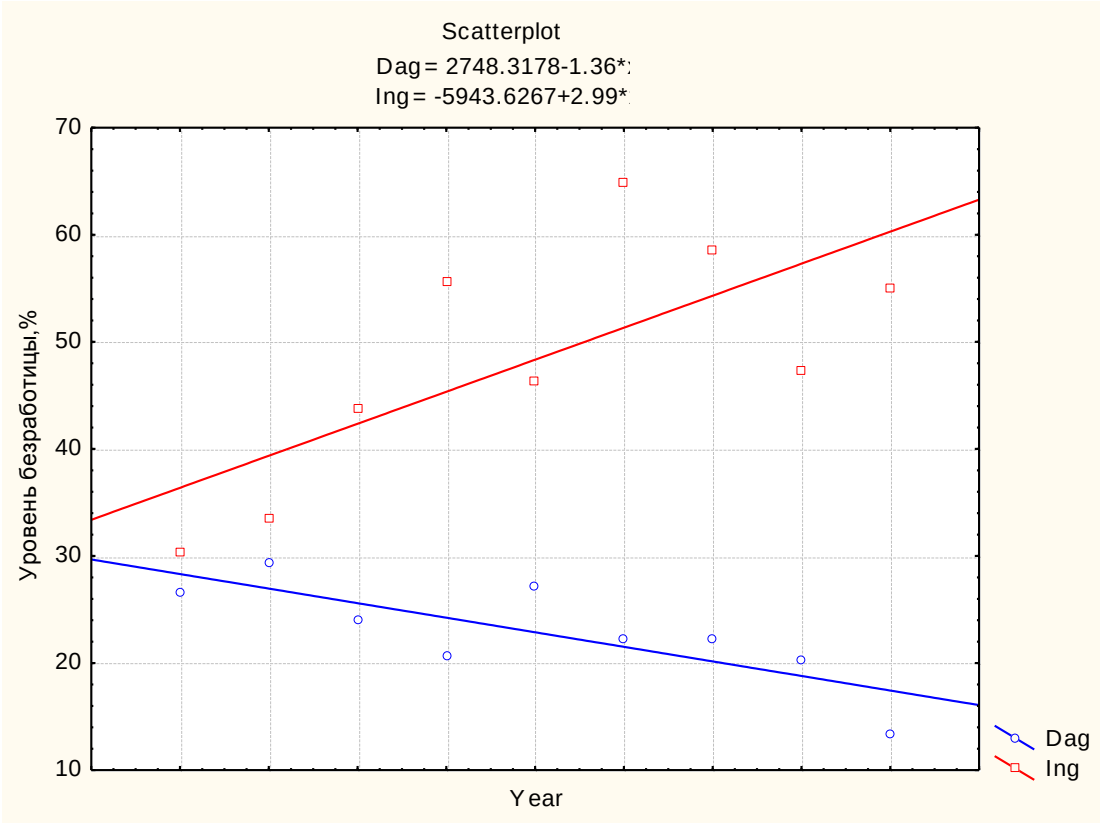

Рис. 21. Динамика уровня безработицы в отдельных регионах РФ

Regression Summary for Dependent Variable: Ing (Spreadsheet7)
R = ,71476713 R?= ,51089205 Adjusted R?= ,44101949
F (1,7)=7,3118 p<,03046 Std.Error of estimate: 8,5652

	Beta	Std.Err.	B	Std.Err.	t(7)	p-level
Intercept			-5943,63	2215,940	-2,68221	0,031438
Year	0,714767	0,264334	2,99	1,106	2,70403	0,030461

Рис. 22. Итоговая таблица регрессионного анализа

Прежде всего, обратим внимание на столбец p-level, где приведены минимальные уровни значимости $\alpha_{min,a}$, $\alpha_{min,b}$ коэффициентов регрессионной модели a и b. По умолчанию компьютер проверяет две нулевые гипотезы вида: H0: a = 0 и H0: b = 0 на уровне значимости 0,05. На этом уровне значимости оба коэффициента отличаются от нуля, поскольку $\alpha_{min,a} = 0{,}031438 < 0{,}05$ и $\alpha_{min,b} < 0{,}05$. В пакете STATISTICA данные окрашены в красные цвет, что означает отказ от нулевых гипотез и статистическую значимость как коэффициента a, так и коэффициента b.

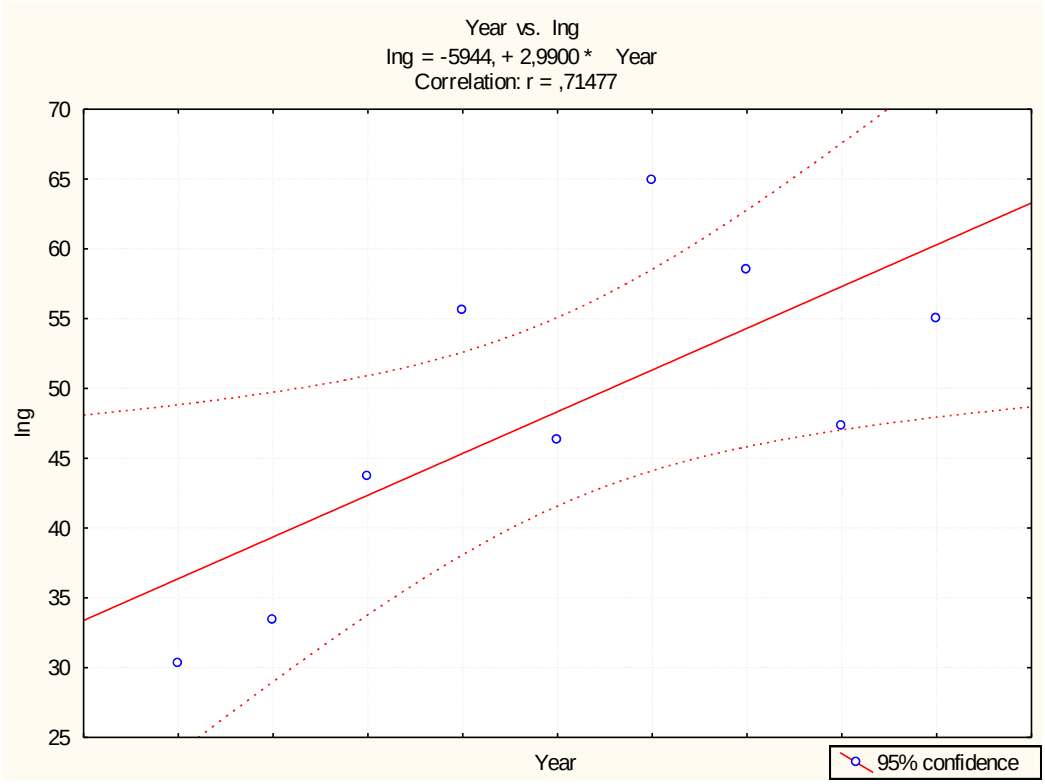

Рис. 23. Линейная регрессионная модель динамики уровня безработицы в Республике Ингушетия

Наличие высокой значимости коэффициентов регрессионной модели еще не означает, что связь в целом значима. Коэффициент детерминации оказался равен $R^2 = 0,51$, это означает, что 51,09 % всей изменчивости статистики уровня безработицы связано с фактором времени. Оставшаяся, необъясненная часть дисперсии приходится на прочие факторы.

6.2 Регрессионно-корреляционный анализ факторов занятости и безработицы в регионах РФ

Исходя из цели работы, выделена группа социально-экономических факторов, которые могут оказывать прямое или косвенное влияние на уровень безработицы в регионах РФ, и соответствующих им статистических показателей.

Таблица 12 - Статистические показатели для описания социально-экономических факторов, сопутствующих безработице

Используемое обозначение	Статистический показатель
bezr	Численность безработных по субъектам Российской Федерации, человек на 10000 населения
zanyt	Среднегодовая численность занятого населения (ОКВЭД), человек на 10000 населения
zarpl	Среднемесячная номинальная начисленная заработная плата в расчете на одного человека, рубль
urbezr	Уровень безработицы населения по субъектам Российской Федерации, %
potr	Потребность организаций в работниках для замещения вакантных рабочих мест по субъектам Российской Федерации (по данным выборочного обследования организаций), человек на 10000 населения
potr%	Потребность организаций в работниках для замещения вакантных рабочих мест по субъектам Российской Федерации (по данным выборочного бследования организаций) в % к общему числу рабочих мест по субъекту Российской Федерации
prinyto	Принято работников с начала года (ОКВЭД с 2005 г.), человек на 10000 нас., всего, январь-декабрь
vibilo	Выбыло работников с начала года (ОКВЭД с 2005 г.), человек на 10000 нас., всего, январь-декабрь
prest	Число зарегистрированных преступлений в расчете на 100 тыс. чел. населения, значение показателя за год
ptu	Численность студентов государственных (муниципальных) учреждений среднего профессионального образования, человек на 10000 нас, за год
gosvuz	Численность студентов государственных (муниципальных) высших учебных заведений, человек на 10000 нас., значение показателя за год
negos	Численность студентов негосударственных высших учебных заведений, человек на 10000 нас., значение показателя за год
stud	Численность студентов государственных и негосударственных высших учебных заведений, человек на 10000 нас., значение показателя за год
invest	Инвестиции в основной капитал на душу населения в фактически действующих ценах (до 1998 года-тыс.рублей), рубль, значение показателя за год
trud	Доля постоянного населения в трудоспособном возрасте (16-59 для мужчин), (16-54 для женщин), на 1 января 2009г.
gor	Доля постоянного городского населения в трудоспособном возрасте (16-59 для мужчин), (16-54 для женщин), на 1 января 2009г.
sel	Доля постоянного сельского населения в трудоспособном возрасте (16-59 для мужчин), (16-54 для женщин), на 1 января 2009г.
migr	Миграционный прирост населения, человек на 10000 населения за год

Проанализированы пространственные данные за один год наблюдения. При статистическом анализе данные для таких регионов, как Чечн-

ская Республика, Республика Ингушетия, Коми-Пермяцкий авт. округ, Таймырский (Долгано-Ненецкий) авт. округ, Эвенкийский авт. округ, Усть-Ордынский Бурятский авт. округ, Корякский авт. округ, Агинский Бурятский авт. округ, интерпретированы как аномальные наблюдения.

Прежде всего, необходимо исключить аномальные наблюдения. В первую очередь удалим регионы, у которых значение по уровню безработицы отсутствует, это такие регионы как: Коми-Пермяцкий авт. округ, Таймырский (Долгано-Ненецкий) авт. округ, Эвенкийский авт. округ, Усть-Ордынский Бурятский авт. округ, Корякский авт. округ, Агинский Бурятский авт. округ.

Далее построим диаграмму рассеивания, главным преимуществом которой является возможность находить «выбросы» (аномальные, нетипичные данные), которые влияют на значение коэффициента корреляции. Построим диаграмму, используя следующие команды Graphs → 2D Graphs → Scatterplots → Graph Type: Multiple → Variables: X = sel, Y = urbezr → OK → OK (рис. 5).

Средство Brushing (закрашивание) интерактивно удаляет выбросы, при этом можно непосредственно наблюдать за изменением аппроксимирующей функции или линии регрессии. На графике видно, что две точки лежать за пределами совокупности данных. Сначала необходимо определить какому региону соответствует эта точка. Для этого необходимо выбрать кнопку Brushing на панели инструментов. После того как откроется окно, в рамке Action (операция) на вкладке Normal выделить операцию Label. В рамке Selection brush → Point. Подвести «прицел» к точке и щелкнуть левой кнопкой мыши. Если выделен режим Auto Update, то появится метка с обозначением региона. Если режим Auto Update не выделен, то необходимо нажать кнопку Update, которая находится в верхней части окна. Если необходимо исключить эту точку из графика, на вкладке Normal выделите функцию Turn OFF, подведите «прицел» к точке, щелкните левой кнопкой мыши и нажмите кнопку Update. Точка исчезнет.

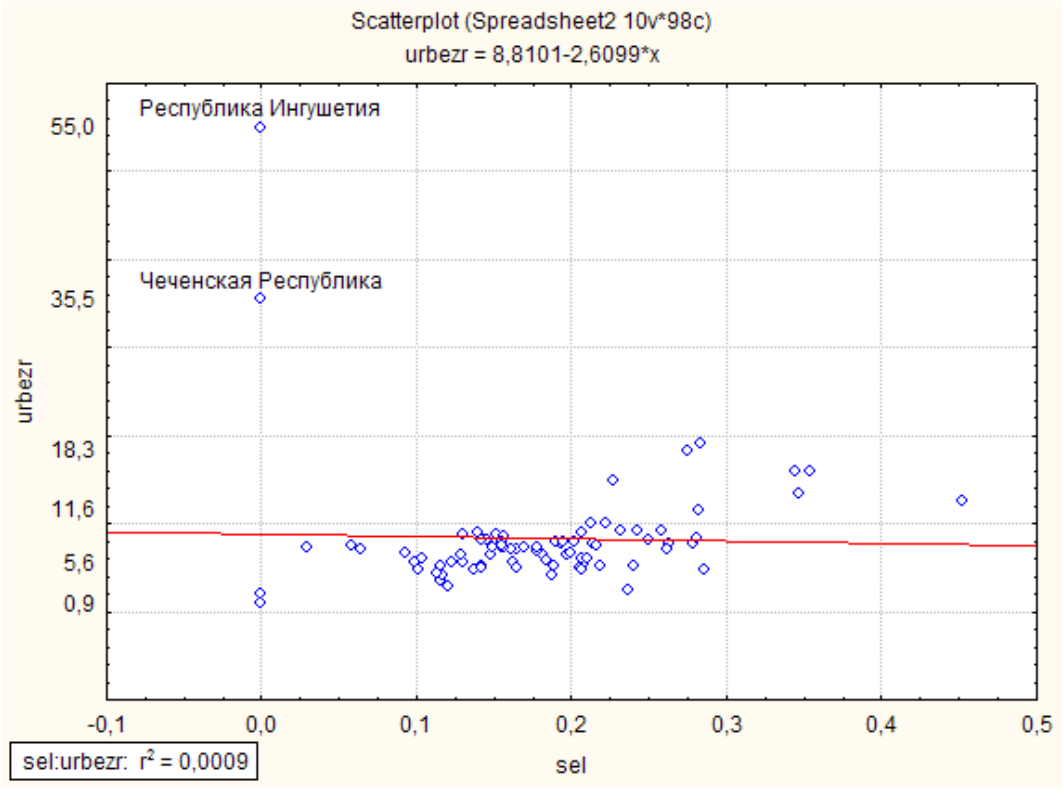

Рис. 24. Регрессионная зависимость между уровнем безработицы в регионах РФ и долей сельского населения трудоспособного возраста

После удаления аномальных точек на графике появится новое уравнение регрессии, изменится положение прямой на плоскости и корреляция существенно возрастет по абсолютной величине и составит – 0,6083.

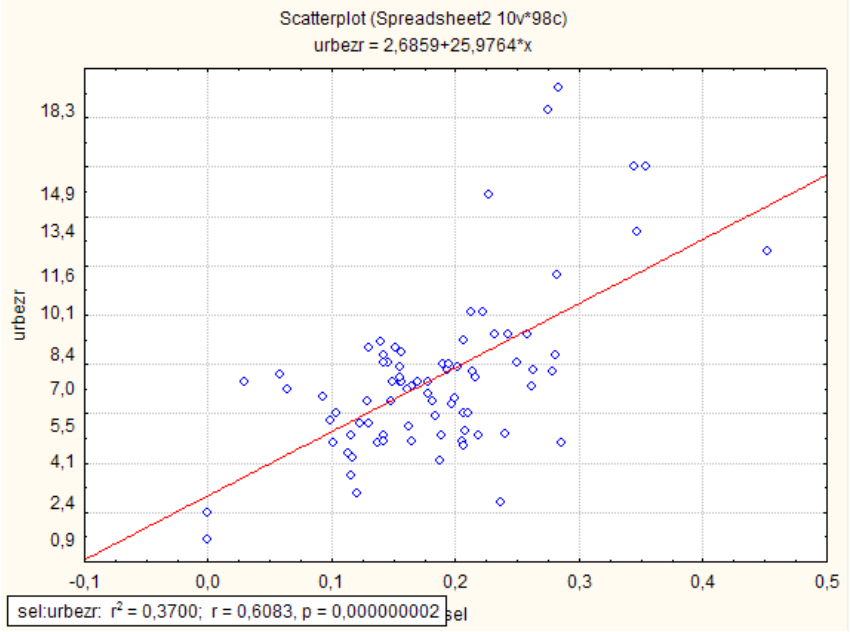

Рис. 25. Регрессионная зависимость между уровнем безработицы в регионах РФ и долей сельского населения трудоспособного возраста с исключением аномальных единиц

Дальнейший анализ проведен для 81 территориального образования.

Результаты корреляционного анализа представлены ниже.

Для показателей с наиболее высокими значениями коэффициентов корреляции построены регрессионные зависимости.

Correlations (bezrab.sta)
Marked correlations are significant at p < ,05000
N=80 (Casewise deletion of missing data)

Variable	bezr	zanyt	zarpl	urbezr	potr	potr%	prinyto	vibilo	prest	ptu	gosvuz	negos	stud	invest	trud	gor
bezr	1.00	-0.32	-0.12	0.97	-0.22	-0.33	-0.48	-0.22	0.12	0.13	-0.21	-0.26	-0.26	0.00	0.23	-0.42
zanyt	-0.32	1.00	0.82	-0.44	0.73	0.60	0.19	0.90	0.21	-0.05	-0.12	0.16	-0.05	0.63	0.53	0.58
zarpl	-0.12	0.82	1.00	-0.22	0.66	0.57	0.24	0.83	0.21	-0.22	-0.16	0.09	-0.10	0.63	0.77	0.62
urbezr	0.97	-0.44	-0.22	1.00	-0.29	-0.39	-0.45	-0.33	0.04	0.15	-0.20	-0.26	-0.25	-0.05	0.09	-0.52
potr	-0.22	0.73	0.66	-0.29	1.00	0.90	0.12	0.70	0.06	-0.27	-0.27	-0.01	-0.23	0.32	0.45	0.37
potr%	-0.33	0.60	0.57	-0.39	0.90	1.00	0.32	0.53	0.11	-0.35	-0.10	0.16	-0.03	0.14	0.37	0.48
prinyto	-0.48	0.19	0.24	-0.45	0.12	0.32	1.00	0.09	0.09	-0.17	0.43	0.55	0.54	-0.04	0.02	0.41
vibilo	-0.22	0.90	0.83	-0.33	0.70	0.53	0.09	1.00	0.27	0.02	-0.28	-0.06	-0.25	0.75	0.53	0.50
prest	0.12	0.21	0.21	0.04	0.06	0.11	0.09	0.27	1.00	0.46	0.11	-0.04	0.08	0.04	0.29	0.26
ptu	0.13	-0.05	-0.22	0.15	-0.27	-0.35	-0.17	0.02	0.46	1.00	0.05	-0.15	-0.01	0.07	-0.16	-0.24
gosvuz	-0.21	-0.12	-0.16	-0.20	-0.27	-0.10	0.43	-0.28	0.11	0.05	1.00	0.37	0.95	-0.31	-0.08	0.24
negos	-0.26	0.16	0.09	-0.26	-0.01	0.16	0.55	-0.06	-0.04	-0.15	0.37	1.00	0.63	-0.13	-0.01	0.34
stud	-0.26	-0.05	-0.10	-0.25	-0.23	-0.03	0.54	-0.25	0.08	-0.01	0.95	0.63	1.00	-0.30	-0.07	0.31
invest	0.00	0.63	0.63	-0.05	0.32	0.14	-0.04	0.75	0.04	0.07	-0.31	-0.13	-0.30	1.00	0.31	0.14
trud	0.23	0.53	0.77	0.09	0.45	0.37	0.02	0.53	0.29	-0.16	-0.08	-0.01	-0.07	0.31	1.00	0.49
gor	-0.42	0.58	0.62	-0.52	0.37	0.48	0.41	0.50	0.26	-0.24	0.24	0.34	0.31	0.14	0.49	1.00
sel	0.53	-0.49	-0.46	0.61	-0.27	-0.42	-0.45	-0.39	-0.20	0.22	-0.29	-0.38	-0.37	-0.06	-0.25	-0.96
migr	-0.50	-0.26	-0.30	-0.42	-0.34	-0.13	0.44	-0.27	-0.08	-0.12	0.32	0.08	0.29	-0.14	-0.50	0.02

Рис. 26. Корреляционная матрица между статистическими показателями для описания социально-экономических факторов, сопутствующих безработице

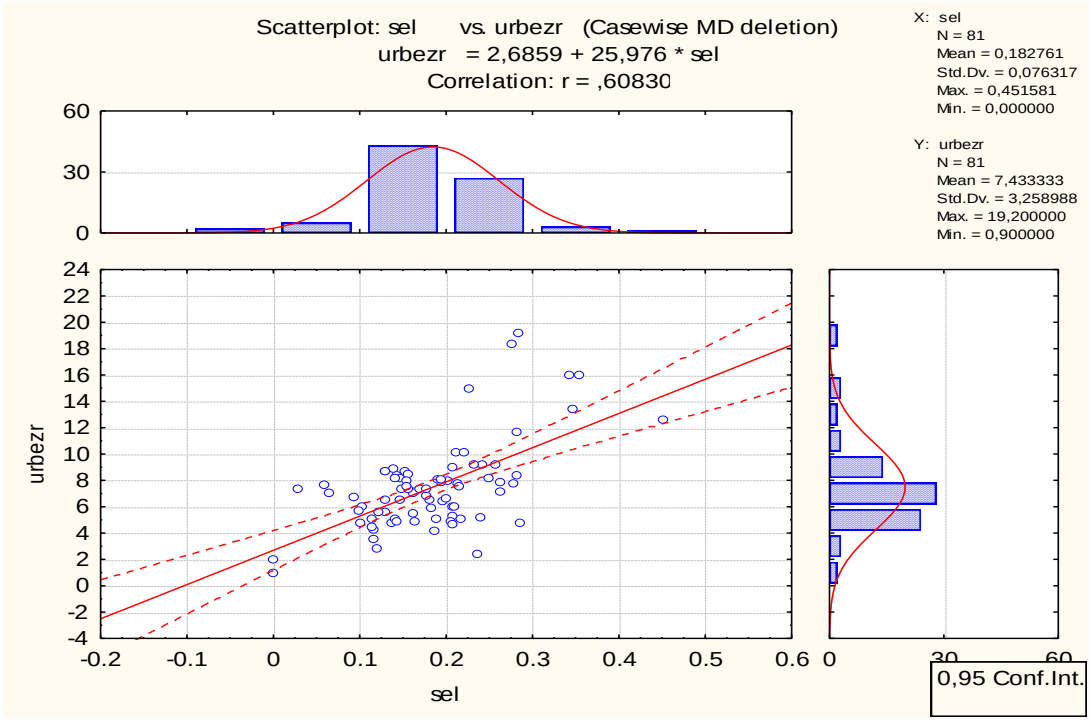

Рис. 27. Регрессионная зависимость между уровнем безработицы в регионах РФ и долей сельского населения трудоспособного возраста

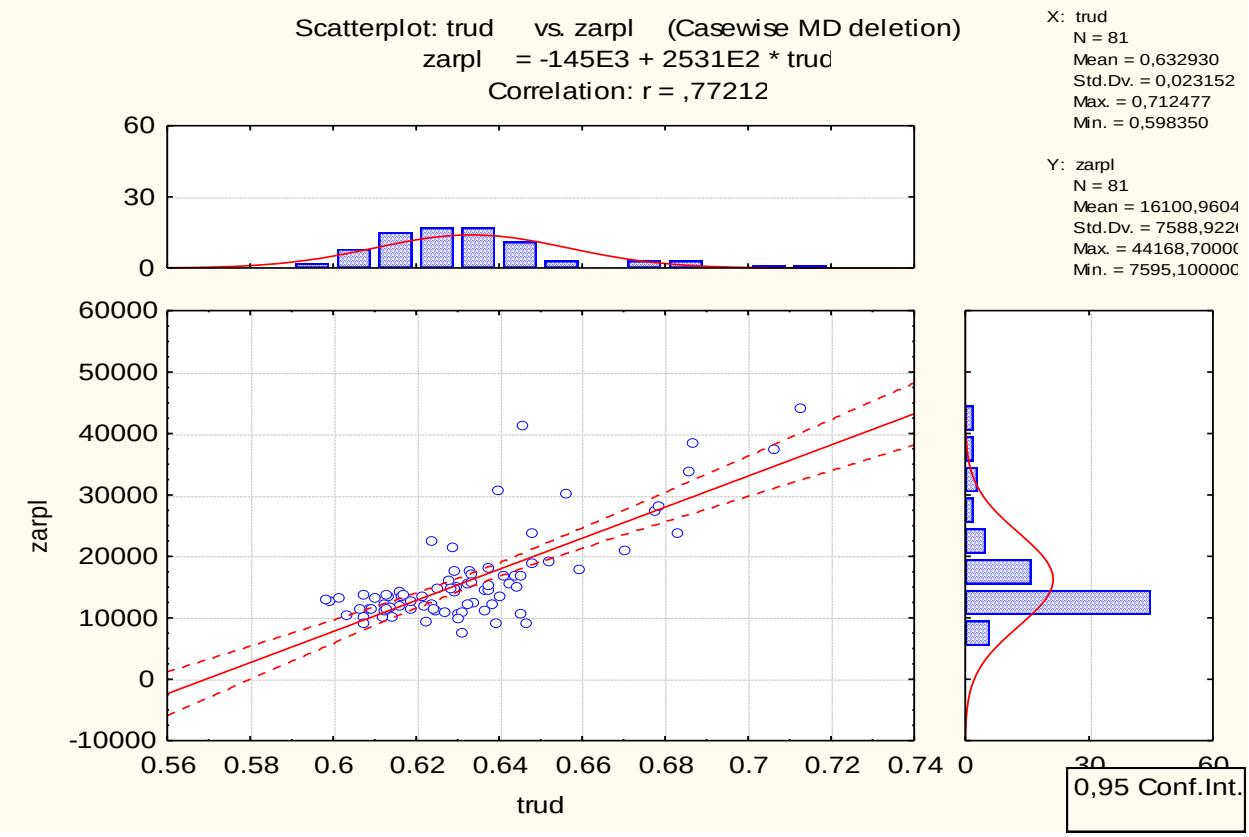

Рис. 28. Регрессионная зависимость между среднемесячной номинальной начисленной заработной платой в расчете на одного человека и долей населения трудоспособного возраста в регионах РФ

6.3 Многомерный статистический анализ факторов безработицы и занятости

Следующий этап исследования – выявление из множества наблюдаемых показателей латентных факторов, в достаточной степени информативно характеризующих дифференциацию регионов, то есть объясняющих значительную долю вариабельности регионов. На этом этапе применяются методы многомерной статистики. Для этого используется метод главных компонент. Применен вариант процедуры определения латентных факторов с вращением главных компонент по методу варимакс.

Результаты факторного анализа приведены ниже на рисунках и в таблицах. Использована программная система статистической обработки данных STATISTICA 8.0.

Перейдем непосредственно к описанию факторного анализа. Меню Statistics → Multivariate Exploratory Techniques → Factor Analysis. В поле Input File (файл входящих данных) необходимо указать тип исходного файла, с которым предстоит работать. В модуле возможны следующие типы исходных данных:

- Correlation Matrix (корреляционная матрица);
- Raw Data (исходные данные).

Выберем Raw Data рабочий файл данных. В правом нижнем углу за всеми функциональными кнопками находится поле MD deletion (обработка пропущенных значений). В этом поле необходимо задать один из способов, которым будут обрабатываться пропущенные значения.

- *Casewise* – способ состоит в том, что в электронной таблице, содержащей данные игнорируются все строки (наблюдения), в которых имеется хотя бы одно пропущенное значение. Недостаток заключается в том, что если в каждой строке таблицы есть по одному пропущенному значению, то в таблице не остается ни одной не исключенной строки, или очень мало строк для проведения анализа и получения достоверного результата;
- *Pairewise* – в этом способе игнорируется не вся строка таблицы, в которой есть пропущенные данные, а только ячейки. Очевидно, что в способе Pairewise остается больше наблюдений для обработки, чем в способе Casewise, так как игнорируются наблюдения только для столбцов, в которых есть пустые ячейки;
- *Mean Substitution* – способ, который предполагает при выполнении анализа заполнение пустых ячеек средними значениями элементов столбца, в котором содержатся пустые ячейки.

Выберем метод *Casewise*. После этого перейдем к выбору переменных, для которых будет проводиться факторный анализ. Select the variables for the factor analysis → выбираем с 4-21 → OK. Программа начнет анализ выбранных переменных. В информационной части окна Define Method of Factor Extraction сообщается, что пропущенные значения обработаны методом *Casewise*. Обработан 81 случай и 80 случаев принято для дальнейших вычислений. Корреляционная матрица вычислена для 18 переменных. Перейдем на вкладку Descriptive, так как факторный анализ необходимо начинать с корреляционной матрицы. Ее анализ позволит оценить степень коррелированности переменных между собой. И если эта степень окажется высокой, то данные переменные можно объединить в один фактор. Review correlation, means, standard deviations → Quick (Advanced) → Correlations. В матрице коэффициенты корреляции между переменными принимают как малые, так и большие значения. Этот факт отразится на результатах последующих этапов факторного анализа.

	Correlations (bezrab.sta) Marked correlations are significant at p < ,05000 N=80 (Casewise deletion of missing data)															
Variable	bezr	zanyt	zarpl	urbezr	potr	potr%	prinyto	vibilo	prest	ptu	gosvuz	negos	stud	invest	trud	gor
bezr	1.00	-0.32	-0.12	0.97	-0.22	-0.33	-0.48	-0.22	0.12	0.13	-0.21	-0.26	-0.26	0.00	0.23	-0.42
zanyt	-0.32	1.00	0.82	-0.44	0.73	0.60	0.19	0.90	0.21	-0.05	-0.12	0.16	-0.05	0.63	0.53	0.58
zarpl	-0.12	0.82	1.00	-0.22	0.66	0.57	0.24	0.83	0.21	-0.22	-0.16	0.09	-0.10	0.63	0.77	0.62
urbezr	0.97	-0.44	-0.22	1.00	-0.29	-0.39	-0.45	-0.33	0.04	0.15	-0.20	-0.26	-0.25	-0.05	0.09	-0.52
potr	-0.22	0.73	0.66	-0.29	1.00	0.90	0.12	0.70	0.06	-0.27	-0.27	-0.01	-0.23	0.32	0.45	0.37
potr%	-0.33	0.60	0.57	-0.39	0.90	1.00	0.32	0.53	0.11	-0.35	-0.10	0.16	-0.03	0.14	0.37	0.48
prinyto	-0.48	0.19	0.24	-0.45	0.12	0.32	1.00	0.09	0.09	-0.17	0.43	0.55	0.54	-0.04	0.02	0.41
vibilo	-0.22	0.90	0.83	-0.33	0.70	0.53	0.09	1.00	0.27	0.02	-0.28	-0.06	-0.25	0.75	0.53	0.50
prest	0.12	0.21	0.21	0.04	0.06	0.11	0.09	0.27	1.00	0.46	0.11	-0.04	0.08	0.04	0.29	0.26
ptu	0.13	-0.05	-0.22	0.15	-0.27	-0.35	-0.17	0.02	0.46	1.00	0.05	-0.15	-0.01	0.07	-0.16	-0.24
gosvuz	-0.21	-0.12	-0.16	-0.20	-0.27	-0.10	0.43	-0.28	0.11	0.05	1.00	0.37	0.95	-0.31	-0.08	0.24
negos	-0.26	0.16	0.09	-0.26	-0.01	0.16	0.55	-0.06	-0.04	-0.15	0.37	1.00	0.63	-0.13	-0.01	0.34
stud	-0.26	-0.05	-0.10	-0.25	-0.23	-0.03	0.54	-0.25	0.08	-0.01	0.95	0.63	1.00	-0.30	-0.07	0.31
invest	0.00	0.63	0.63	-0.05	0.32	0.14	-0.04	0.75	0.04	0.07	-0.31	-0.13	-0.30	1.00	0.31	0.14
trud	0.23	0.53	0.77	0.09	0.45	0.37	0.02	0.53	0.29	-0.16	-0.08	-0.01	-0.07	0.31	1.00	0.49
gor	-0.42	0.58	0.62	-0.52	0.37	0.48	0.41	0.50	0.26	-0.24	0.24	0.34	0.31	0.14	0.49	1.00
sel	0.53	-0.49	-0.46	0.61	-0.27	-0.42	-0.45	-0.39	-0.20	0.22	-0.29	-0.38	-0.37	-0.06	-0.25	-0.96
migr	-0.50	-0.26	-0.30	-0.42	-0.34	-0.13	0.44	-0.27	-0.08	-0.12	0.32	0.08	0.29	-0.14	-0.50	0.02

Рис. 29. Корреляционная матрица между статистическими показателями для описания социально-экономических факторов, сопутствующих безработице

Исключим факторы, дублирующие переменные и после этого вернемся в исходное окно Define Method of Factor Extraction → Advanced → Maximum no. of factors (10) → Minimum eigenvalue (0) → OK. В методе главных компонент по умолчанию предполагается, что дисперсии всех переменных равны 1. Тогда общая дисперсия равна общему числу переменных (в исследовании число переменных равно 10). Это означает, что наибольшая изменчивость, которая потенциально может быть выделена, равна 10. Максимально возможное число выделяемых факторов равно числу переменных. Каждому фактору соответствует дисперсия, объясненная этим фактором. Дисперсии, соответствующие факторам, называются собственными значениями. Далее откроем таблицу с собственными числами: Factor Analysis Results → Eigenvalues.

Таблица 13 - Таблица собственных чисел и доли объясненной дисперсии

	Eigenvalue	% Total	Cumulative Eigenvalue	Cumulative
1	3,967739	39,67739	3,96774	39,6774
2	2,548919	25,48919	6,51666	65,1666
3	1,225043	12,25043	7,74170	77,4170
4	0,758699	7,58699	8,50040	85,0040
5	0,495751	4,95751	8,99615	89,9615
6	0,423930	4,23930	9,42008	94,2008
7	0,229929	2,29929	9,65001	96,5001
8	0,194552	1,94552	9,84456	98,4456
9	0,093511	0,93511	9,93807	99,3807
10	0,061927	0,61927	10,00000	100,0000

Во втором столбце таблицы приведены дисперсии выделенных факторов – собственные числа. В третьем столбце для каждого фактора приводится процент от общей дисперсии. Как видно из таблицы, первый фактор объясняет 39,7% общей дисперсии, второй фактор – 25,5% и т.д. Четвертый столбец содержит накопленную и кумулятивную дисперсию. Получив результаты, необходимо решить, сколько факторов следует оставить.

Критерий Кайзера. Согласно данному критерию необходимо отобрать факторы с собственными значениями, большими 1. Это говорит о том, что если фактор не выделяет дисперсию, эквивалентную, по крайне мере, дисперсии одной переменной, то он опускается. В нашем исследовании можно выделить 3 фактора, так как остальные не подходят под условие, наложенное на собственные значения (табл. 4).

Критерий каменистой осыпи. Данный критерий является графическим методом, впервые предложенным Кэттелем. На графике необходимо изобразить собственные значения и найти такое место, где их убывание слева направо максимально замедляется. Выполним следующие действия: Factor Analysis Results → Explained variance → Scree plot.

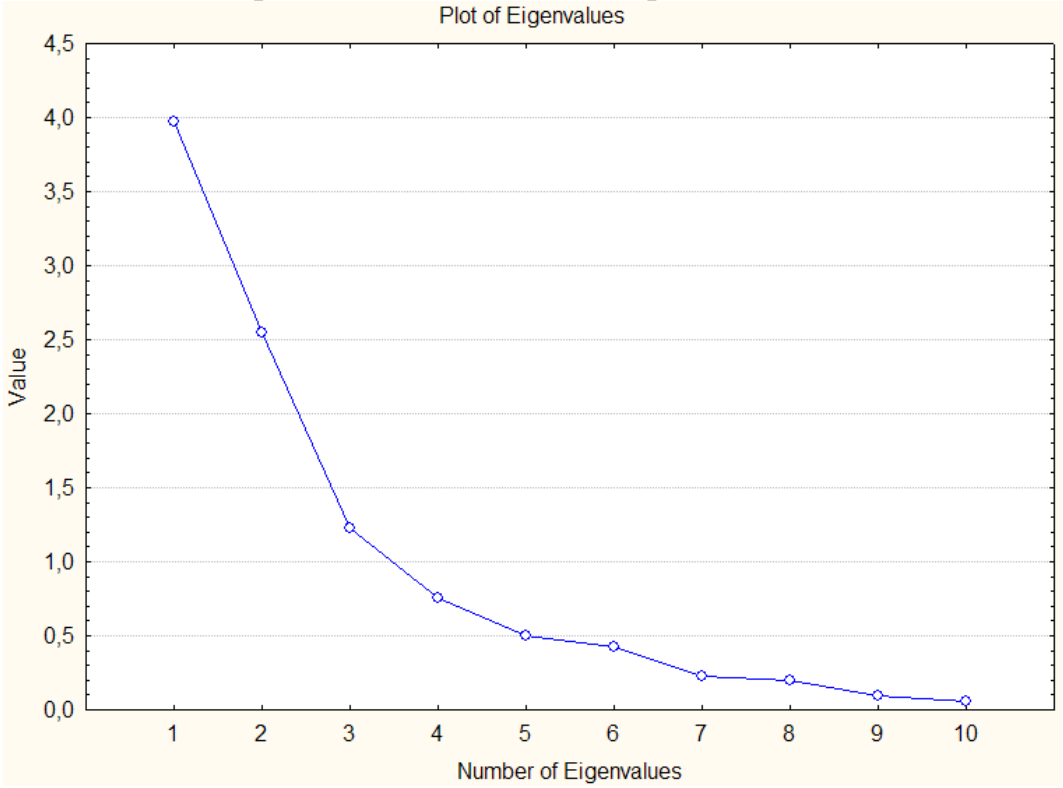

Рис. 30. График собственных значений переменных

Из построенного графика видно, что в соответствии с этим критерием можно выделить 3 или 4 фактора. Различные методы выделения факторов расположены на вкладке Advanced окна Define Method of Factor Extraction. Выберем опцию Principal components. Назначим число факторов равное – 3. Define Method of Factor Extraction → Advanced → Maximum no. of factors (3) → Minimum eigenvalue (0) → OK. Для того чтобы проинтерпретировать ре-

шение, необходимо применить повороты осей, которые достигаются вращением факторов. В окне Factor Analysis Results программа предлагает несколько способов вращения. Выберем метод варимакс, который предназначен для максимизации дисперсий квадратов исходных факторных нагрузок по переменным для каждого фактора, что эквивалентно максимизации дисперсий в столбцах матрицы квадратов исходных факторных нагрузок. Factor Analysis Results → Quick → Factor rotation → Varimax raw → Summary: Factor loadings.

Таблица 14. Факторные нагрузки (коэффициенты корреляции исходных показателей с латентными факторами).

	Factor 1	Factor 2	Factor 3
bezr	-0,242807	-0,286996	-0,793482
zarpl	0,930501	0,188580	-0,152966
potr	0,766346	-0,012432	-0,047401
prinyto	0,169216	0,682749	0,472082
vibilo	0,951570	-0,102354	0,040429
stud	-0,244450	0,822376	0,149660
invest	0,723640	-0,381559	0,093390
trud	0,647880	0,287542	-0,591288
gor	0,587084	0,640593	0,063475
migr	-0,276842	0,173938	0,816470
Expl.Var	3,870661	1,939891	1,931149
Prp.Totl	0,387066	0,193989	0,193115

Из фрагмента таблицы факторных нагрузок (табл. 5) следует, что Factor 1 имеет высокие факторные нагрузки по переменным zarpl, potr, vibilo, invest, trud. У Factor 2 факторные нагрузки по переменным stud, prinyto, gor. Factor 3 – bezr, migr. Исходя из анализа коэффициентов корреляции между выявленными факторами и исходными показателями, а также из содержательного состава показателей предложена интерпретация факторов: первый фактор – фактор экономической стабильности региона и положительной трудовой мобильности населения, второй фактор – фактор наличия резерва квалифицированных трудовых ресурсов, третий фактор – трудовой и миграционной привлекательности региона.

Для выявления и интерпретации закономерностей можно использовать графическое представление факторных нагрузок. Factor Analysis Results → Loadings → Plot of loadings 2D → Factor 1 и Factor 2. (Factor 2 и Factor 3, Factor 1 и Factor 3). Графики иллюстрируют соотношение между факторами и группами переменных.

В результате факторного анализа выявлены три латентных фактора, в значительной степени определяющие вариабельность регионов по оцениваe-

мым социально-экономическим показателям. Доля общей дисперсии, объясняемая совокупностью факторов, составляет 77%. Каждый из рассмотренных показателей (кроме показателя доли трудоспособного населения) имеет высокие по абсолютной величине коэффициенты корреляции с одним и только одним фактором.

Исходя из анализа коэффициентов корреляции между выявленными факторами и исходными показателями, а также из содержательного состава показателей предложена интерпретация факторов: первый фактор – фактор экономической стабильности региона и положительной трудовой мобильности населения, второй фактор – фактор наличия резерва квалифицированных трудовых ресурсов, третий фактор – трудовой и миграционной привлекательности региона.

	Eigenvalues (bezrab.sta) Extraction: Principal components			
Value	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	3.97	39.68	3.97	39.68
2	2.55	25.49	6.52	65.17
3	1.23	12.25	7.74	77.42

Рис. 31. Собственные числа и доля объясненной дисперсии

	Factor Loadings (Varimax normalized) (bezrab. Extraction: Principal components (Marked loadings are >,700000)		
Variable	Factor 1	Factor 2	Factor 3
bezr	-0.25	-0.34	-0.77
zarpl	0.90	0.27	-0.21
potr	0.76	0.06	-0.09
prinyto	0.12	0.72	0.44
vibilo	0.96	-0.00	-0.00
stud	-0.32	0.80	0.13
invest	0.76	-0.30	0.07
trud	0.59	0.33	-0.63
gor	0.52	0.70	0.01
migr	-0.26	0.18	0.82
Expl.Var	3.74	2.07	1.93
Prp.Totl	0.37	0.21	0.19

Рис. 32. Факторные нагрузки (коэффициенты корреляции исходных показателей с латентными факторами)

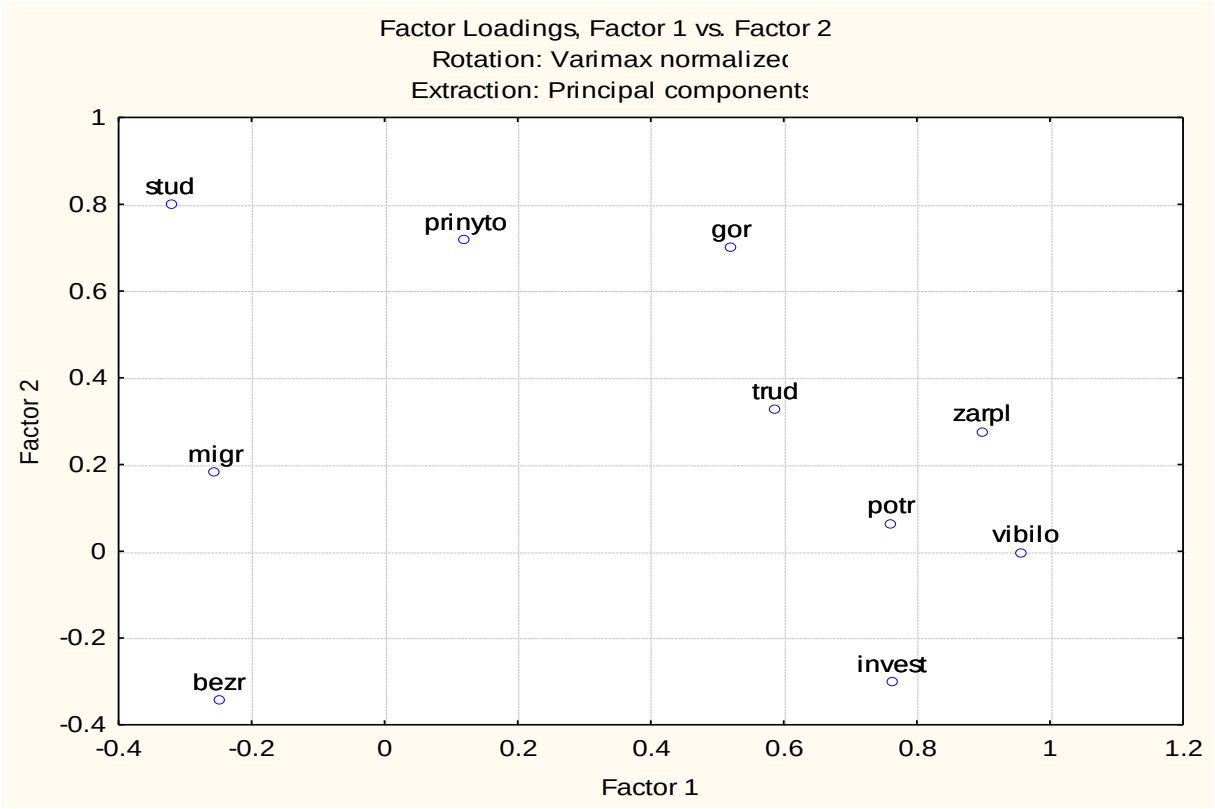

Рис. 33. Факторные нагрузки для факторов 1 и 2 (коэффициенты корреляции исходных показателей с первым и вторым латентными факторами)

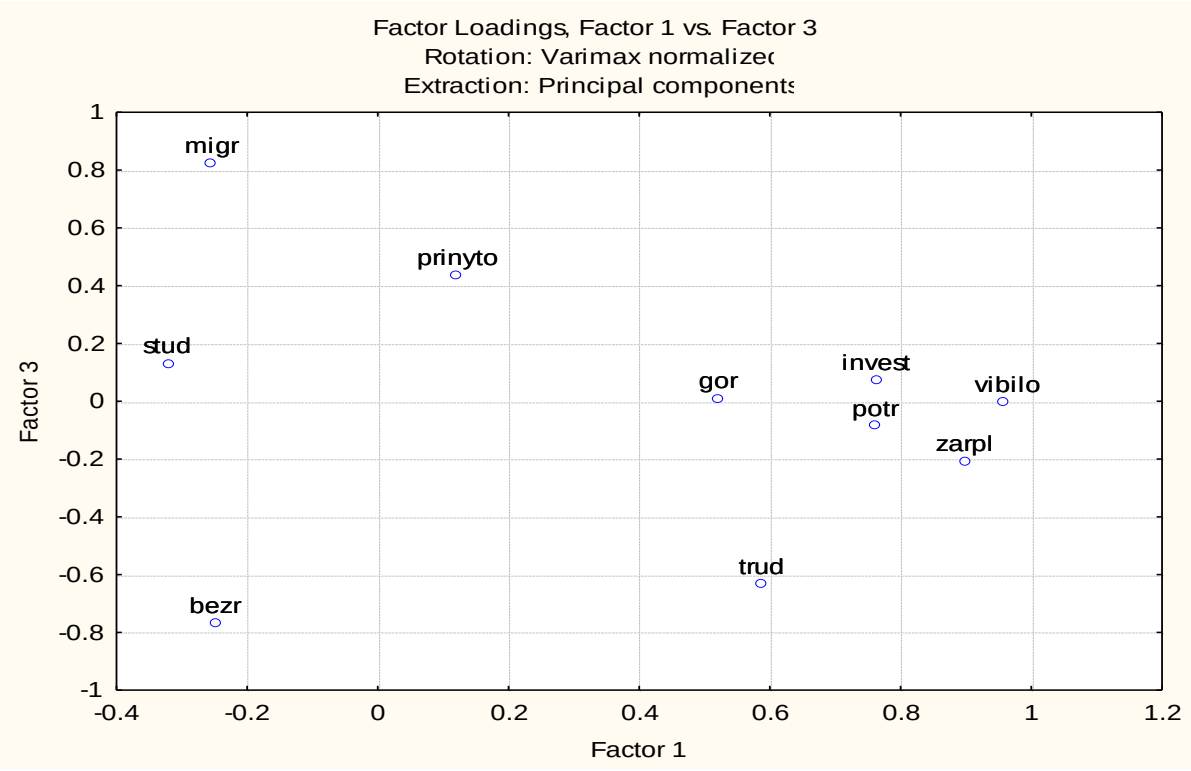

Рис. 34. Факторные нагрузки для факторов 1 и 3 (коэффициенты корреляции исходных показателей с первым и третьим латентными факторами)

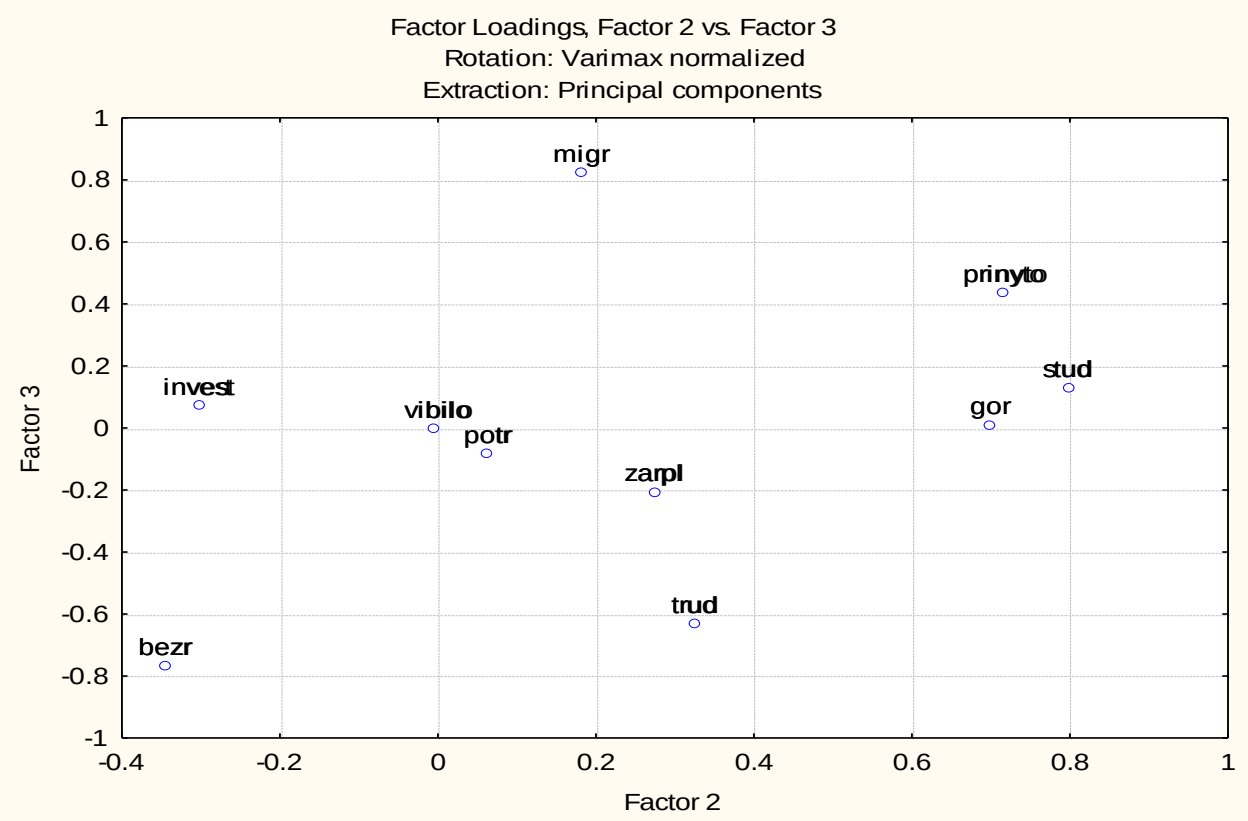

Рис. 35. Факторные нагрузки для факторов 2 и 3 (коэффициенты корреляции исходных показателей с вторым и третьим латентными факторами)

Далее выведем коэффициенты для определения значений латентных факторов. Factor Analysis Results → Scores → Factor Score coefficients.

Ниже представлены коэффициенты для определения значений латентных факторов (19) и полученные значения латентных факторов для регионов РФ (10).

	Factor Score Coefficients (bezrab.sta Rotation: Varimax normalized Extraction: Principal components		
Variable	Factor 1	Factor 2	Factor 3
bezr	-0.130	-0.023	-0.430
zarpl	0.216	0.101	-0.071
potr	0.210	-0.026	0.026
prinyto	0.015	0.305	0.144
vibilo	0.286	-0.098	0.111
stud	-0.161	0.453	-0.109
invest	0.266	-0.258	0.189
trud	0.068	0.241	-0.376
gor	0.086	0.335	-0.064
migr	0.002	-0.028	0.434

Рис. 36. Коэффициенты для определения значений латентных факторов

Для того чтобы вывести значения латентных факторов для регионов РФ необходимо нажать кнопку Factor Scores. По этим значениям можно су-

дить о привлекательности регионов (положительные значения) или экономической нестабильности субъектов (отрицательные значения).

	Factor 1	Factor 2	Factor 3
1	-0,44837	-0,67020	-1,03030
2	-0,62113	-0,03013	-0,34160
3	-0,23972	-0,28483	-0,70441
		…	
79	1,48284	0,71512	0,16528
80	-0,24542	0,55046	0,31368
81	4,06690	1,77757	-0,60548

Рис. 37. Значения латентных факторов для регионов РФ

6.4 Кластерный анализ регионов РФ по латентным факторам занятости и безработицы

Далее обобщенные характеристики состояния занятости и безработицы в регионах (латентные факторы), используются для построения типологии регионов с помощью кластерного анализа.

Для запуска модуля кластерного анализа необходимо осуществить следующие действия Statistics → Multivariate Exploratory Techniques → Cluster Analysis. На вкладке Quick находится список методов кластерного анализа. Это Joining tree clustering (древовидная кластеризация), k-means clustering (метод k-средних), Two-way joining (двухвходовая кластеризация). Выберем метод k-средних. Cluster Analysis → k-means clustering → Advanced → Variables (Factor 1, Factor 2, Factor 3). В поле Cluster необходимо выбрать объекты для кластеризации. Так как цель исследования – типологизация регионов, которые являются в файле данных наблюдениями, необходимо выбрать Cases (rows). В поле Number of cluster запишем число групп равное 4. В поле Number of iterations задается максимальное число итераций, используемых при построении классов, например 10 → OK. На вкладке Advanced содержится подробная информация о результатах анализа.

Summary: Cluster means & Euclidean distance предназначена для вывода таблиц.

Результаты кластерного анализа представлены ниже на рисунках и в таблицах. Выделены четыре кластера.

Cluster Number	Euclidean Distances between Clusters (Spreadsheet Distances below diagonal Squared distances above diagonal			
	No. 1	No. 2	No. 3	No. 4
No. 1	0.000	1.875	2.481	3.233
No. 2	1.369	0.000	0.978	0.667
No. 3	1.575	0.989	0.000	1.737
No. 4	1.798	0.816	1.318	0.000

Рис. 38 Межкластерные расстояния

Variable	Analysis of Variance (Spreadsheet55)					
	Between SS	df	Within SS	df	F	signif. p
FACTOR1	33.5	3	46.5	77	18.5	0.00000
FACTOR2	23.9	3	56.1	77	11.0	0.00000
FACTOR3	54.4	3	25.6	77	54.7	0.00000

Рис. 39. Проверка статистической значимости различий между кластерами по критерию Фишера

Variable	Cluster Means (Spreadsheet55)			
	Cluster No. 1	Cluster No. 2	Cluster No. 3	Cluster No. 4
FACTOR1	1.4569	-0.16776	-0.84764	0.26100
FACTOR2	0.6230	0.36042	-0.72217	-0.68108
FACTOR3	-1.4689	0.23900	-0.90103	1.09406

Рис. 40. Средние значения латентных факторов для сформированных кластеров

Описательная статистика для сформированных четырех кластеров приведена на следующих четырех рисунках. Определены среднее, стандартное отклонение, дисперсия, количество регионов для каждого из трех латентных факторов, описывающих регионы РФ, входящих в соответствующий кластер.

Variable	Descriptive Statistics for Cluster 1 (Spreadsheet5 Cluster contains 10 cases		
	Mean	Standard Deviation	Variance
FACTOR1	1.457	1.255	1.576
FACTOR2	0.623	0.865	0.749
FACTOR3	-1.469	0.664	0.441

Рис. 41. Описательная статистика для кластера 1 (среднее, стандартное отклонение, дисперсия)

Variable	Descriptive Statistics for Cluster 2 (Spreadsheet5 Cluster contains 41 cases		
	Mean	Standard Deviation	Variance
FACTOR1	-0.168	0.302	0.091
FACTOR2	0.360	0.941	0.885
FACTOR3	0.239	0.530	0.281

Рис. 42. Описательная статистика для кластера 2 (среднее, стандартное отклонение, дисперсия)

Variable	Descriptive Statistics for Cluster 3 (Spreadsheet5 Cluster contains 14 cases		
	Mean	Standard Deviation	Variance
FACTOR1	-0.848	0.353	0.125
FACTOR2	-0.722	0.472	0.223
FACTOR3	-0.901	0.741	0.549

Рис. 43. Описательная статистика для кластера 3 (среднее, стандартное отклонение, дисперсия)

Variable	Descriptive Statistics for Cluster 4 (Spreadsheet5 Cluster contains 16 cases		
	Mean	Standard Deviation	Variance
FACTOR1	0.261	1.342	1.801
FACTOR2	-0.681	0.857	0.734
FACTOR3	1.094	0.462	0.214

Рис. 44. Описательная статистика для кластера 4 (среднее, стандартное отклонение, дисперсия)

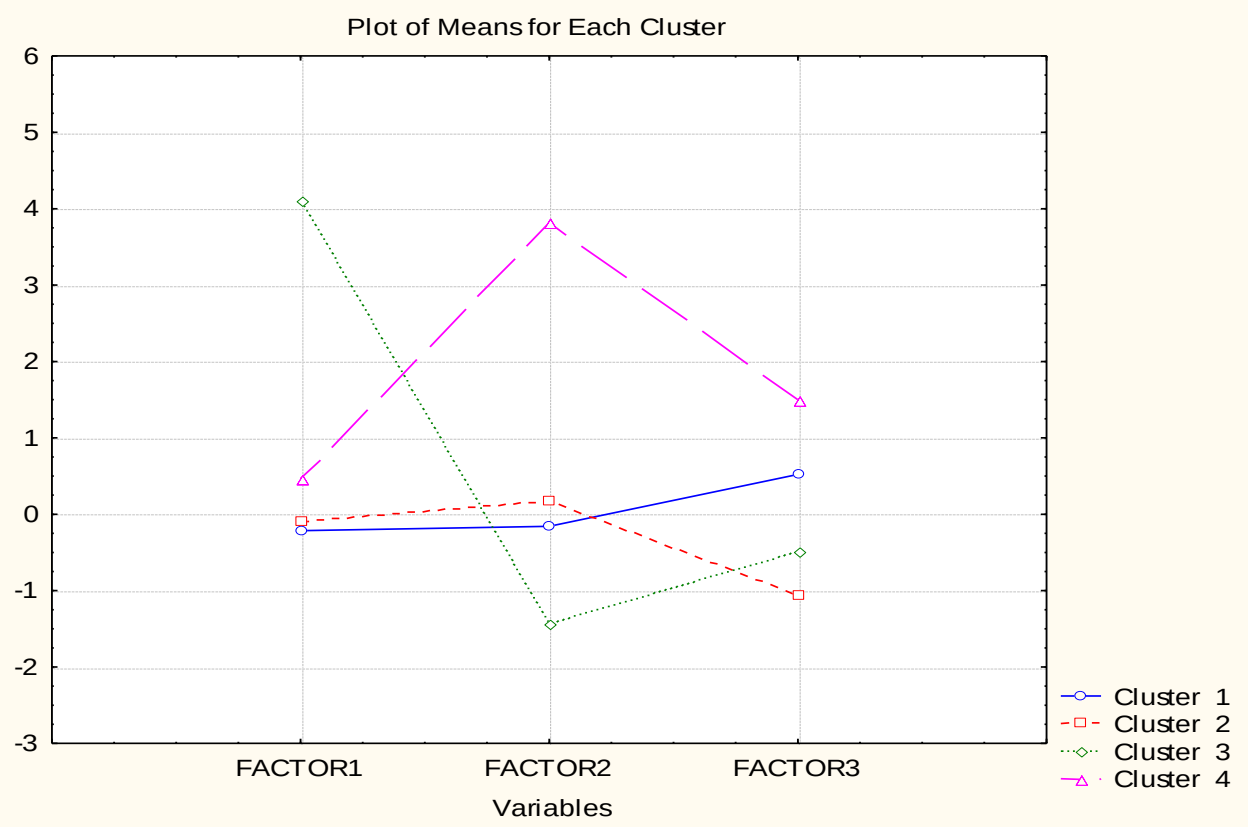

Рис. 45. График средних значений латентных факторов для сформированных кластеров

Далее переходим к содержательной интерпретации состава кластеров. Первый и второй кластеры различаются по фактору трудовой и миграционной привлекательности региона, значения факторов экономической стабильности региона и положительной трудовой мобильности населения и наличия резерва квалифицированных трудовых ресурсов близки к среднему уровню. Регионы первого кластера – это относительно благополучные регионы. Регионы второго кластера характеризуются отрицательной трудовой и миграционной привлекательностью. Для регионов третьего кластера существенно высоки значения фактора экономической стабильности региона и положительной трудовой мобильности населения. Для регионов четвертого кластера выражен фактор наличия резерва квалифицированных трудовых ресурсов.

Список литературы

Основная литература

1. Айвазян, С.А, Фантаццини Д. Эконометрика - 2. Продвинутый курс с приложениями в финансах / С.А. Айвазян, Д.Фантаццини. Учебник – М.: Инфра-М, 2015, - 994 с.
2. Елисеева, И.И. Эконометрика. Учебник для магистровИ. И. Елисеева ; под ред. И. И. Елисеевой. — М. : Издательство Юрайт, 2014. — 449 с.
3. Кремер, Н. Ш. Эконометрика: учебник и практикум для академического бакалавриата / Н. Ш. Кремер, Б. А. Путко ; под ред. Н. Ш. Кремера. — 4-е изд., испр. и доп. — М. : Издательство Юрайт, 2018. — 354 с.
4. Магнус, Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс: Учебник. / Я.Р. Магнус, П.К. Катышев, А.А.Пересецкий – М.: Дело, 2007. – 400 с.
5. Методы и модели эконометрики. Часть 2. Эконометрика пространственных данных: учебное пособие/под ред. А.Г. Реннера; Оренбургский гос. ун-т. – Оренбург: ОГУ, 2015. – 435 с.

Дополнительная литература

6. Гладилин А.В., Герасимов А.Н., Громов Е.И. Практикум по эконометрике – г. Ростов-на Дону: Феникс, 2011, - 326 с.
7. Доугерти К. Введение в эконометрику: Пер. с англ. – М.: ИНФРА-М, 2003. – 402 с.
8. Елисеева И.И. Практикум по эконометрике. Учебное пособие – М.: Финансы и статистика, 2003, - 192 с.
9. Кремер, Н. Ш. Математика для экономистов: от арифметики до эконометрики. Учебно-справочное пособие : для академического бакалавриата / Н. Ш. Кремер, Б. А. Путко, И. М. Тришин ; под общ. ред. Н. Ш. Кремера. — 4-е изд., перераб. и доп. — М. : Издательство Юрайт, 2017. — 724 с.
10. Носко В.П. Эконометрика. Книга 1. Части 1 и 2 – М.: Издательский дом "Дело" РАНХиГС, 2011, - 672 с.
11. Подкорытова, О. А. Анализ временных рядов : учебное пособие для бакалавриата и магистратуры / О. А. Подкорытова, М. В. Соколов. — 2-е изд., перераб. и доп. — М. : Издательство Юрайт, 2018. — 267 с.
12. Прикладная статистика. Основы эконометрики: Учебник для вузов: В 2-х т. – Т. 1. Айвазян С.А., Мхитарян В.С. Теория вероятностей и прикладная статистика. – М: ЮНИТИ-ДАНА, 2006. – 656 с.
13. Теория статистики с элементами эконометрики в 2 ч. Часть 1 : учебник для академического бакалавриата / В. В. Ковалев [и др.] ; отв. ред. В. В. Ковалев. — М. : Издательство Юрайт, 2017. — 333 с.

14. Теория статистики с элементами эконометрики в 2 ч. Часть 2 : учебник для академического бакалавриата / В. В. Ковалев [и др.] ; отв. ред. В. В. Ковалев. — М.: Издательство Юрайт, 2017. — 348 с.

Программное обеспечение и Интернет-ресурсы

1. Программные продукты: электронные таблицы Open Office Calc, MS Office Excel, студенческие версии программных продуктов EViews, Statistica, PcGive, Gretl.
2. Официальный сайт Росстата [Электронный ресурс] Режим доступа: http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/science_and_innovations/science/
3. Сборники по статистике науки, инноваций и информационного общества, выпущенные Росстат совместно с НИУ ВШЭ [Электронный ресурс] Режим доступа: https://www.hse.ru/org/hse/primarydata/
4. Единая межведомственная информационно-статистическая система [Электронный ресурс] Режим доступа: https://fedstat.ru/
5. Всемирная организация интеллектуальной собственности [Электронный ресурс] Режим доступа: http://ipstats.wipo.int/ipstatv2/keyindex.htm -
6. Евостат. Официальные данные. [Электронный ресурс] Режим доступа: http://ec.europa.eu/eurostat/data/browse-statistics-by-theme

Максимова Татьяна Геннадьевна

Попова Ирина Николаевна

Эконометрика

Учебно-методическое пособие

В авторской редакции
Редакционно-издательский отдел Университета ИТМО
Зав. РИО Н.Ф. Гусарова
Подписано к печати
Заказ №
Тираж
Отпечатано на ризографе

Редакционно-издательский отдел
Университета ИТМО
197101, Санкт-Петербург, Кронверкский пр., 49