

Документ подписан простой электронной подписью

Информация о владельце:

ФИО: Комин Андрей Александрович

Должность: ректор

Дата подписания: 08.04.2019 07:32:42

Уникальный программный ключ:

f6c6d686f0c899fdf76a1ed8b448452ab8cac6fb1af6547b6d40edf1bdc60ae2

Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Приморская государственная сельскохозяйственная академия»
Институт земледелия и природообустройства

Кафедра физики и
высшей математики

Моделирование и статистическая обработка результатов научных
исследований

Методические указания к практическим занятиям и
самостоятельной работе для обучающихся по направлениям
подготовки 35.06.01 Сельское хозяйство; 35.06.02 Лесное
хозяйство; 35.06.04 Технологии, средства механизации и
энергетическое оборудование в сельском, лесном и рыбном
хозяйстве; 36.06.01 Ветеринария и зоотехния; 38.06.01 Экономика

Электронное издание

Уссурийск 2015

Савельева Е.В. Моделирование и статистическая обработка результатов научных исследований: методические указания к практическим занятиям и самостоятельной работе для обучающихся по направлениям подготовки 35.06.01 Сельское хозяйство; 35.06.02 Лесное хозяйство; 35.06.04 Технологии, средства механизации и энергетическое оборудование в сельском, лесном и рыбном хозяйстве; 36.06.01 Ветеринария и зоотехния; 38.06.01 Экономика [Электронный ресурс]: / сост. Е.В. Савельева; ФГБОУ ВПО «Приморская ГСХА». - Уссурийск, 2015. – 71 с. - Режим доступа: www.elib.primacad.ru.

Методические указания составлены в соответствии с учебным планом и рабочей программой дисциплины (модуля).

Включают указания для выполнения практических занятий, указания для выполнения самостоятельной работы.

Предназначены для обучающихся по направлениям 35.06.01 Сельское хозяйство; 35.06.02 Лесное хозяйство; 35.06.04 Технологии, средства механизации и энергетическое оборудование в сельском, лесном и рыбном хозяйстве; 36.06.01 Ветеринария и зоотехния; 38.06.01 Экономика.

Электронное издание

Рецензент: Сидорова Г.М., к. с.-х. н., доцент, доцент кафедры землеустройства

Издается по решению методического совета ФГБОУ ВПО «Приморская ГСХА»

ВВЕДЕНИЕ

Дисциплина «Моделирование и статистическая обработка результатов научных исследований» занимает важное место в системе прикладного математического образования. Она нацелена на формирование мировоззрения научного исследования.

Содержание дисциплины охватывает круг вопросов, связанных с моделированием и статистической обработкой результатов научных исследований.

Преподавание дисциплины предусматривает следующие формы организации учебного процесса: лекции, практические работы, самостоятельная работа аспиранта, консультации.

Целью освоения дисциплины является ознакомление аспирантов с методами обработки и анализа результатов экспериментов и наблюдений с применением информационных технологий для использования их в научно-исследовательской деятельности.

Задачи дисциплины научить

- * методам математической статистики по обработке результатов научных исследований с применением информационных технологий;

- * строить математические модели исследуемых процессов, явлений и объектов, относящихся к области профессиональной деятельности, анализировать и интерпретировать полученные результаты.

Краткое изложение теоретического материала не позволяет осветить все вопросы курса статистики и представить формулы по всем описанным показателям. Поэтому при подготовке к практическим и семинарским занятиям рекомендуется более глубокое изучение тем в основной и дополнительной литературе.

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ ПО ТЕМЕ № 1

«Первичная обработка экспериментальных данных»

Цель занятия: Приобретение навыков первичной обработки эмпирических данных методами математической статистики.

Содержание работы: На основе совокупности опытных данных выполнить следующие задания:

Задание 1. Построить интервальный вариационный ряд распределения.

Задание 2. Построить гистограмму частот интервального вариационного ряда.

Задание 3. Составить эмпирическую функцию распределения и построить график.

Задание 4. Рассчитать основные числовые характеристики вариационного ряда:

- а) моду и медиану;
- б) условные начальные моменты;
- в) выборочную среднюю;
- г) выборочную дисперсию, исправленную дисперсию генеральной совокупности, исправленное среднее квадратичное отклонение;
- д) коэффициент вариации;
- е) асимметрию;
- ж) эксцесс;

Задание 5. Определить границы истинных значений числовых характеристик, изучаемой случайной величины с заданной надёжностью.

Задание 6. Содержательная интерпретация результатов первичной обработки по условию задачи.

Методика выполнения работы

Даны результаты измерения роста (с точностью до см) 60 наудачу отобранных студентов:

178, 160, 154, 183, 155, 153, 167, 186, 163, 155,

157, 175, 170, 166, 159, 173, 182, 167, 171, 169,
179, 165, 156, 179, 158, 171, 175. 173, 164, 172
178, 160, 154, 183, 155, 153, 167, 186, 163, 155,
157, 175, 170, 166, 159, 173, 182, 167, 171, 169,
179, 165, 156, 179, 158, 171, 175. 173, 164, 172

Задание 1. Построить интервальный вариационный ряд распределения

Краткие теоретические сведения

Для того, чтобы статистические данные представить в виде вариационного ряда с равноотстоящими вариантами необходимо:

1. В исходной таблице данных найти наименьшее x_{\min} и наибольшее x_{\max} значения.

2. Определить **размах варьирования**: $x_{\max} - x_{\min}$

3. Определить длину интервала h , если в выборке до 1000 данных, используют формулу: $h = \frac{x_{\max} - x_{\min}}{1 + \log_2 n}$, где n – объем выборки – количество данных в выборке; для вычислений берут $(1 + \log_2 n) \approx (1 + 3,3221 \lg n)$.

Вычисленное отношение округляют до **удобного целого значения**.

4. Определить начало первого интервала для четного числа интервалов рекомендуют брать величину $x_{нач} = x_{\min} - \frac{h}{2}$; а для нечетного числа интервалов $x_{нач} = x_{\min}$.

5. Записать интервалы группировок и расположить их в порядке возрастания границ

$$x_0 - (x_0 + h), (x_0 + h) - (x_0 + 2h), \dots, (x_0 + (s-1)h) - (x_0 + sh),$$

где x_0 - нижняя граница первого интервала. За x_0 берется удобное число не большее x_{\min} , верхняя граница последнего интервала должна быть не меньше x_{\max} . Рекомендуется, чтобы интервалы содержали в себе исходные значения случайной величины и выделять от **5 до 20** интервалов.

6. Записать исходные данные по интервалам группировок, т.е. подсчитать по исходной таблице число значений случайной величины, попадающих в указанные интервалы. Если некоторые значения совпадают с границами интервалов, *то их относят либо только к предыдущему, либо только к последующему интервалу.*

Замечание. Если для некоторых интервалов получены «нулевые», либо малые значения частот n_i , то необходимо перегруппировать данные, укрупняя интервалы (увеличивая шаг h).

Задание 1. Построить интервальный вариационный ряд распределения

Решение

1. $n=60$; $x_{\max} = 186$, $x_{\min} = 153$.

Длина частично интервала:

$$h = \frac{x_{\max} - x_{\min}}{1 + \log_2 n} = \frac{186 - 153}{1 + \log_2 60} = \frac{33}{1 + 3,322 \lg 60} \approx \frac{33}{5,9} \approx 5,59$$

Примем $h = 6$. Начало первого интервала $x_{нач} = x_{\min} - \frac{h}{2} = 153 - 3 = 150$

Исходные данные разбиваем на следующие равные интервалы:

(150,156], (156,162], (162,168], (168,174], (174, 180], (180, 186].

Подсчитаем n_i - число студентов попавших в каждый из полученных промежутков.

Таблица 1

| | | | | | | |
|-----------------|-----------|-----------|-----------|-----------|------------|------------|
| X | [150,156] | (156,162] | (162,168] | (168,174] | (174, 180] | (180, 186] |
| n_i (частота) | 8 | 10 | 12 | 14 | 10 | 6 |

Задание 2. Построить гистограмму частот интервального вариационного ряда

Краткие теоретические сведения

Гистограммой частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длины h , а высоты равны отношению n_i / h (плотность частоты).

Для построения гистограммы на оси Ox отмечают интервалы. На оси Oy -плотность частоты. Над каждым интервалом проводят отрезки параллельные оси Ox

Площадь частичного i -го прямоугольника равна $h(n_i / h) = n_i$, сумме частот вариант, попавших в i -й интервал. Площадь гистограммы частот равна сумме всех частот, т.е. объему выборки n .

Задание 2. Построить гистограмму частот интервального вариационного ряда

Решение

Длина интервала $h = 6$. Найдем плотность частоты n_i / h .

Таблица 2

| | | | | | | |
|-----------------|-------------------|---------------------|--------------|--------------------|---------------------|------------|
| X | [150,156) | [156,162) | [162,168) | [168,174) | [174, 180) | [180, 186] |
| n_i (частота) | 8 | 10 | 12 | 14 | 10 | 6 |
| n_i / h | $8/6 \approx 1,4$ | $10/6 \approx 1,65$ | $12/6 = 2,0$ | $14/6 \approx 2,3$ | $10/6 \approx 1,65$ | $6/6 = 1$ |

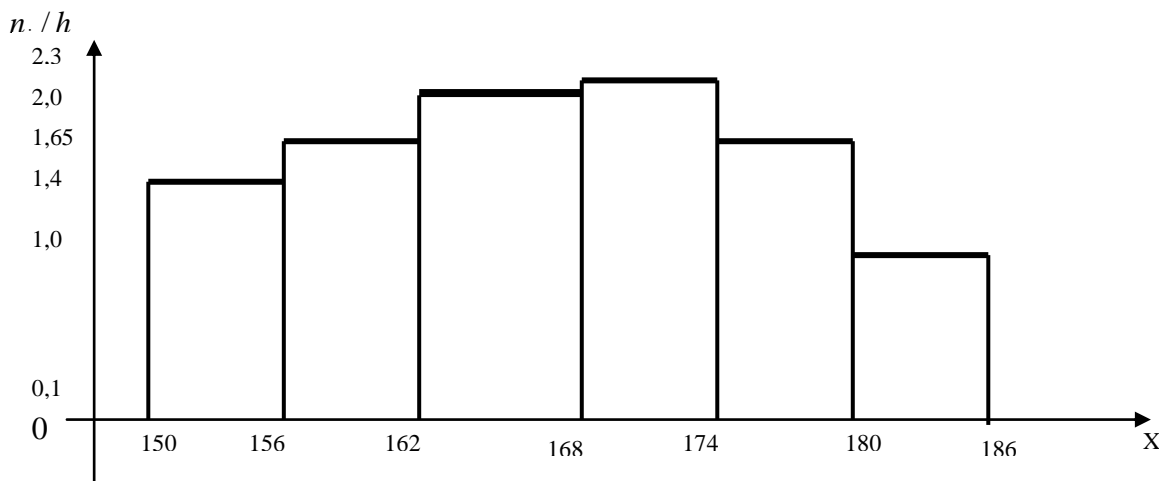


Рисунок 1 - Гистограмма частот вариационного ряда

Задание 3. Записать эмпирическую функцию распределения и построить ее график

Краткие теоретические сведения

Эмпирическая функция распределения $F^*(x)$, определяет для каждого значения x относительную частоту события $X < x$.

$$F^*(x) = \sum_{x < x_i} w_i \text{ где } w_i - \text{ относительная частота (частость)}$$

Для построения графика эмпирической функции распределения переходят к дискретному ряду распределения, для этого в качестве вариант x_i принимают середины частичных интервалов. Частоты при этом не изменяются. На оси Ox отмечают значения x_i , на оси Oy – накопительные частоты.

Задание 3. Записать эмпирическую функцию распределения и построить ее график

Решение

Вычислим:

- середину каждого интервала x_i и запишем эти значения в первую строку таблицы;
- относительные частоты $w_i = n_i / n$ и запишем в третью строку таблицы;
- накопительные частоты $\sum w_i$ и запишем в четвертую строку таблицы.

Таблица 3

| | | | | | | |
|--|----------------------|-----------------------|-----------------------|-----------------------|-----------------------|--------------|
| Значение признака x_i (середина интервала) | 153 | 159 | 165 | 171 | 177 | 183 |
| n_i (частота) | 8 | 10 | 12 | 14 | 10 | 6 |
| $w_i = n_i / n$ (частость) | $8/60 \approx 0,133$ | $10/60 \approx 0,167$ | $12/60 = 0,2$ | $14/60 \approx 0,233$ | $10/60 \approx 0,167$ | $6/60 = 0,1$ |
| Накопительные относительные частоты $\sum w_i$ | 0,133 | $0,133+0,167=0,3$ | $0,133+0,167+0,2=0,5$ | 0,733 | 0,9 | 1 |

$$F^*(x) = \begin{cases} 0 & \text{при } x \leq 153 \\ 0,133 & \text{при } 153 < x \leq 159 \\ 0,3 & \text{при } 159 < x \leq 165 \\ 0,5 & \text{при } 165 < x \leq 171 \\ 0,733 & \text{при } 171 < x \leq 177 \\ 0,9 & \text{при } 177 < x \leq 183 \\ 1 & \text{при } x > 183 \end{cases}$$

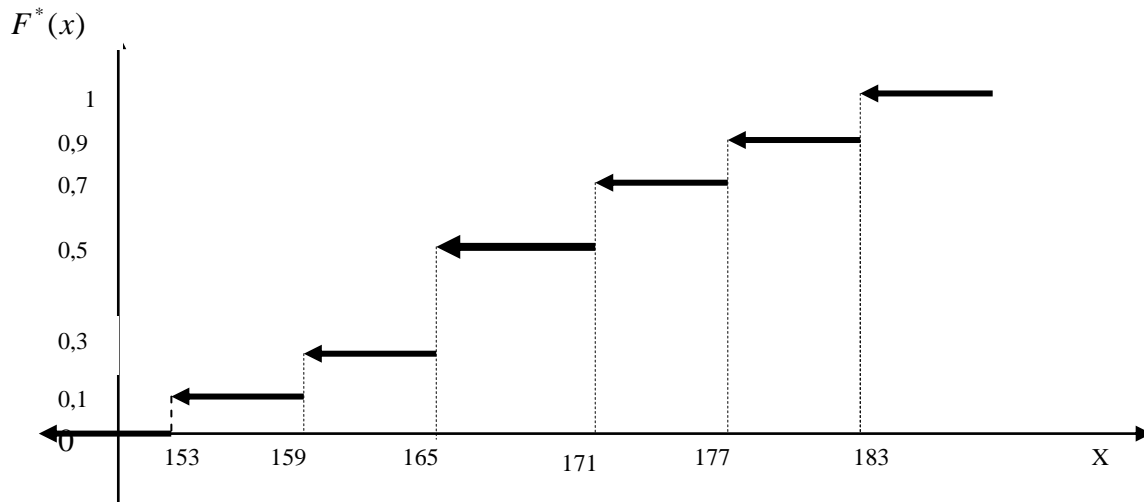


Рисунок 2 - График эмпирической функции распределения

Задание 4. Рассчитать основные числовые характеристики вариационного ряда

Решение

4а) Мода – это варианта, имеющая наибольшую частоту.

По данным Таблицы 3: $M_0(X) = 171$.

Медиана – это варианта, которая делит вариационный ряд на две части, равные по числу вариант. Если число интервалов четное, то медиана определяется как среднее арифметическое срединных значений интервалов, если нечетное, то медиана – это срединное значение медианного (среднего) интервала.

В нашем примере число интервалов четное

По данным Таблицы 3: $(153+159+165+171+177+183)/6=168$,
 $M_e(X) = 168$

Для определения остальных числовых характеристик воспользуемся методом произведений.

Введем **условные варианты**. $u_i = \frac{x_i - C}{h}$, где C – «ложный нуль».

Чаще всего в качестве ложного нуля принимается либо варианта, находящаяся в середине вариационного ряда, либо мода M_0 (варианта x_i , имеющая наибольшую частоту), либо любое другое число, упрощающее расчеты.

Если за C принять какое - либо значение x_i , то соответствующая ему условная варианта u_i будет равна нулю, а слева и справа от нуля будут располагаться соответственно значения $\pm 1, \pm 2, \pm 3, \pm 4$ и т.д.

Примем $C = 171, h = 6$. Составим расчетную таблицу:

Таблица 4

| x_i | n_i | u_i | $n_i \cdot u_i$ | $n_i \cdot u_i^2$ | $n_i \cdot u_i^3$ | $n_i \cdot u_i^4$ | $n_i(u_i+1)^4$ |
|----------|----------------------|-------|----------------------------|-----------------------------|-------------------------------|-----------------------------|-----------------------|
| 153 | 8 | -3 | -24 | 72 | -216 | 648 | 128 |
| 159 | 10 | -2 | -20 | 40 | -80 | 160 | 10 |
| 165 | 12 | -1 | -12 | 12 | -12 | 12 | 0 |
| 171 | 14 | 0 | 0 | 0 | 0 | 0 | 14 |
| 177 | 10 | 1 | 10 | 10 | 10 | 10 | 160 |
| 183 | 6 | 2 | 12 | 24 | 48 | 96 | 486 |
| Σ | $\Sigma n_i =$ 60 | | $\Sigma n_i u_i = -$ 34 | $\Sigma n_i u_i^2 =$ 158 | $\Sigma n_i u_i^3 = -$ 250 | $\Sigma n_i u_i^4 =$ 926 | $\Sigma =$ 798 |

Выполним проверку. $\Sigma n_i u_i^4 + 4 \Sigma n_i u_i^3 + 6 \Sigma n_i u_i^2 + 4 \Sigma n_i u_i + n = \Sigma n_i (u_i + 1)^4$

$$926 + 4 \cdot (-250) + 6 \cdot 158 + 4 \cdot (-34) + 60 = -74 + 948 + 26 = 798$$

4б) Условные начальные моменты найдем по формулам:

$$M_1 = \frac{\Sigma n_i u_i}{n} \quad M_1 = \frac{-34}{60} \approx -0,56$$

$$M_2 = \frac{\Sigma n_i u_i^2}{n} \quad M_2 = \frac{158}{60} \approx 2,63$$

$$M_3 = \frac{\Sigma n_i u_i^3}{n} \quad M_3 = \frac{-250}{60} = -4,2$$

$$M_4 = \frac{\Sigma n_i u_i^4}{n} \quad M_4 = \frac{926}{60} \approx 15,4$$

4в) Выборочная средняя $\bar{x}_g = M_1 \cdot h + C \Rightarrow \bar{x}_g = -0,56 \cdot 6 + 171 = 167,6$

4г) Дисперсия (рассеивание) – характеристика рассеяния значений случайной величины X около ее математического ожидания.

Выборочная дисперсия

$$D_e = [M_2 - M_1^2] \cdot h^2 \quad D_e = [2,63 - (-0,56)^2] \cdot 6^2 = [2,63 - 0,336] \cdot 36 \approx 82,4$$

Среднее квадратичное отклонение также служит для оценки рассеяния случайной величины X вокруг ее среднего значения $\sigma_e = \sqrt{D_e}$

$$\sigma_e = \sqrt{82,4} \approx 9,1$$

Исправленная дисперсия генеральной совокупности

$$s^2 = \frac{n}{n-1} D_e = 60/59 \cdot 82,4 \approx 83,7966$$

Исправленное среднее квадратичное отклонение $s = \sqrt{s^2} \approx 9,15$

4д) Коэффициентом вариации V

$$V = \frac{\sigma_e}{\bar{x}_e} \cdot 100\%; \quad V = \frac{9,1}{167,6} \cdot 100\% \approx 5,4\%$$

4е) Асимметрия (коэффициент асимметрии)

Для вычисления асимметрии и эксцесса найдем **центральные моменты**:

$$\mu_3 = [M_3 - 3M_1M_2 + 2M_1^3] \cdot h^3 \quad \mu_3 = [(-4,2) - 3(-0,56) \cdot 2,63 + 2(-0,56)^3] \cdot 6^3 \approx -34,56$$

$$\mu_4 = [M_4 - 4M_1M_3 + 6M_1^2M_2 - 3M_1^4] \cdot h^4$$

$$\mu_4 = [15,4 - 4(-0,56)(-4,2) + 6(-0,56)^2 \cdot 2,63 - 3(-0,56)^4] \cdot 6^4 \approx 13478,4$$

$$\text{Асимметрия } A_s = \frac{\mu_3}{\sigma_e^3} \quad A_s = \frac{-34,56}{(9,1)^3} = \frac{-34,56}{753,6} \approx 0,05$$

4и) Эксцесс вычисляем по формуле:

$$E_x = \frac{\mu_4}{\sigma_e^4} - 3; \quad E_x = \frac{13478,4}{(9,1)^4} - 3 = \frac{13478,4}{6857,76} - 3 \approx 1,965 - 3 \approx -1,03$$

Задание 5. Определение границ истинных значений числовых характеристик изучаемой случайной величины с заданной надёжностью

Краткие теоретические сведения

Доверительные интервалы находят для оценки неизвестного математического ожидания α нормального распределения по выборочной средней \bar{x}_e при неизвестном σ (или при известном σ) генеральной совокупности.

Доверительный интервал – это интервал, который покрывает неизвестный параметр α с заданной надежностью γ .

Известно, что истинное значение измеряемой величины равно ее математическому ожиданию α . Поэтому задача сводится 1) к оценке математического ожидания α при неизвестном σ , т.к. в задании на лабораторную работу не дано значение σ генеральной совокупности; 2) к оценке среднего квадратичного отклонения нормального распределения.

5а) Доверительный интервал для оценки математического ожидания нормального распределения при неизвестном среднем квадратичном отклонении генеральной совокупности, покрывающий математическое ожидание с заданной надежностью, имеет вид: $\bar{x}_e - t_\gamma \cdot \frac{s}{\sqrt{n}} < \alpha < \bar{x}_e + t_\gamma \cdot \frac{s}{\sqrt{n}}$, где s – исправленное среднее квадратичное отклонение; $t_\gamma = t(\gamma; n)$ – табличное значение, которое зависит от заданной надежности γ и объема выборки см. «Таблицу значений $t_\gamma = t(\gamma; n)$ » Приложение 3 учебника.

5б) Доверительный интервал для оценки среднего квадратичного отклонения нормального распределения, покрывающий σ с заданной надежностью имеет вид:

$s(1 - q) < \sigma_r < s(1 + q)$, где s – исправленное среднее квадратичное отклонение; $q = q(\gamma; n)$ – табличное значение находится по «Таблице значений $q = q(\gamma; n)$ »

Задание 5. Определение границ истинных значений числовых характеристик изучаемой случайной величины с заданной надёжностью

Решение

Найдем доверительные интервалы, покрывающие неизвестный параметр с надежностью $\gamma=0,95$.

5а) Доверительный интервал для оценки математического ожидания нормального распределения при неизвестном среднем квадратичном отклонении генеральной совокупности с заданной надежностью найдем по формуле: $\bar{x}_g - t_\gamma \cdot \frac{s}{\sqrt{n}} < \alpha < \bar{x}_g + t_\gamma \cdot \frac{s}{\sqrt{n}}$.

По условию: $\bar{x}_g = 167,6$, $s = 9,15$, $\gamma = 0,95$, $n = 60$.

По таблице $t_\gamma = t(\gamma; n) = t(0,95; 60) = 2,001$

Подставим все значения в формулу:

$$167,6 - 2,001 \cdot \frac{9,15}{\sqrt{60}} < \alpha < 167,6 + 2,001 \cdot \frac{9,15}{\sqrt{60}}$$

$165,24 < \alpha < 169,96$, округлим до см. $165 < \alpha < 170$

5б) Доверительный интервал для оценки среднего квадратичного отклонения нормального распределения, покрывающий σ с заданной надежностью $\gamma=0,95$ найдем по формуле: $s(1 - q) < \sigma_r < s(1 + q)$.

По условию $s = 9,15$, $\gamma=0,95$, $n=60$. По таблице $q=q(\gamma;n)=0,188$

Следовательно: $9,15 \cdot (1 - 0,188) < \sigma_r < 9,15 \cdot (1 + 0,188) \Rightarrow 7,43 < \sigma_r < 10,87$

Округлим до см. $7 < \sigma_r < 11$

Задание 6. Содержательная интерпретация результатов первичной обработки по условию задачи

Краткие теоретические сведения

1. Интерпретация \bar{x}_g .

Значения случайной величины изменяются в пределах от x_{\min} до x_{\max} . Значение \bar{x}_g является обобщённой характеристикой, средним значением, около которого группируется все другие возможные значения изучаемой случайной величины.

2. Интерпретация $\bar{x}_g - \frac{t \cdot \gamma \cdot s}{\sqrt{n}} < a < \bar{x}_g + \frac{t \cdot \gamma \cdot s}{\sqrt{n}}$.

Доверительный (надежностный) интервал с вероятностью $\gamma=0,95$ утверждает, что значения генеральной средней для выборки данного объема n колеблется в пределах

$$\text{от } \bar{x}_e - \frac{t \cdot \gamma \cdot s}{\sqrt{n}} \text{ до } \bar{x}_e + \frac{t \cdot \gamma \cdot s}{\sqrt{n}},$$

3. Интерпретация σ_e .

Отклонение средней величины \bar{x}_e от отдельно взятого значения x_i составляет в среднем σ_e .

4. Интерпретация $s \cdot (-q) \leq \sigma_r < s \cdot (+q)$.

Доверительный (надежностный) интервал для σ_r указывает с вероятностью γ на то, что отклонения могут принимать значения находящиеся в пределах

$$\text{от } s \cdot (-q) \text{ до } s \cdot (+q).$$

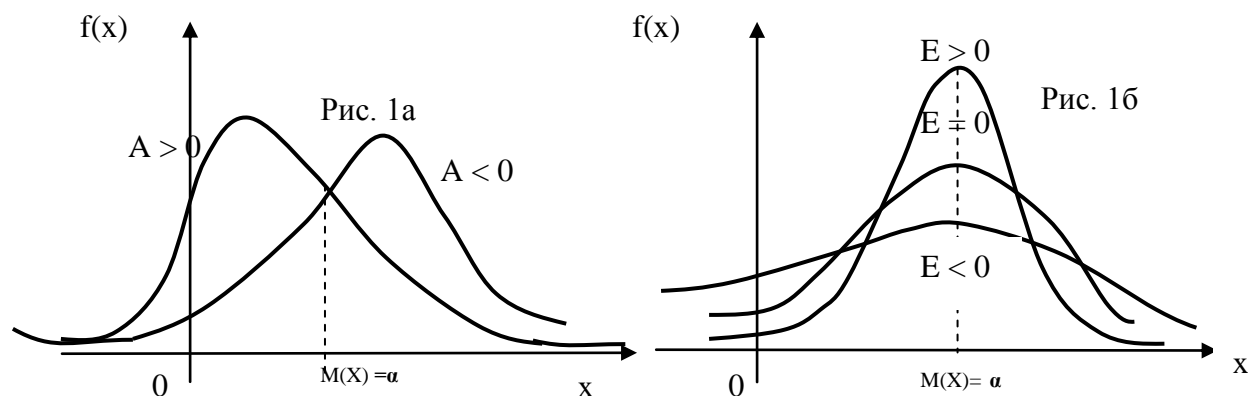
5. Интерпретация коэффициента асимметрии A_s .

Асимметрия (коэффициент асимметрии) – характеризует «скошенность» кривой распределения относительно математического ожидания. Асимметрия равна нулю, если кривая распределения симметрична относительно математического ожидания; асимметрия положительна, если «длинная часть» кривой распределения расположена справа от математического ожидания; асимметрия отрицательна, если «длинная часть» кривой распределения расположена слева от математического ожидания (рис.1а)

6. Интерпретация E_x .

Экссесс – это характеристика большего или меньшего подъема кривой исследуемого распределения по сравнению с нормальной кривой распределения. Если эксцесс равен нулю, то кривая распределения является нормальной кривой; если эксцесс положителен, то кривая распределения имеет более высокую и «острую» вершину, чем нормальная кривая; если

эксцесс отрицателен, то кривая распределения имеет более низкую и «плоскую» вершину, чем нормальная кривая (рис. 1б).



7. Интерпретация значения коэффициента вариации V .

Коэффициент вариации V – это выраженное в процентах отношение выборочного среднего квадратичного отклонения к выборочной средней. Он дает возможность оценить меру рассеивания значений случайной величины вокруг среднего значения. Если $V < 10\%$, то изменчивость считают незначительной, если $10\% < V < 20\%$ то изменчивость считают средней, если $V > 20\%$, то изменчивость значительная.

*Использование коэффициента вариации V как показателя колеблемости (вариации) имеет смысл только **при положительных значениях вариант** и не применимо, если варианты принимают как положительные так и отрицательные значения*

Задание 6. Содержательная интерпретация результатов первичной обработки по условию задачи

Решение

1. Интерпретация \bar{x}_e .

Рост студента – величина случайная, ее выборочные значения изменяются в частности от 153 см. до 186 см, однако среднее значение равно 168 см.

2. Интерпретация доверительного интервала $\bar{x}_e - \frac{t \cdot \gamma \cdot s}{\sqrt{n}} < a < \bar{x}_e + \frac{t \cdot \gamma \cdot s}{\sqrt{n}}$.

Так как $165 < \alpha < 170$ при заданной надежности $\gamma = 0,95$, то можно утверждать с вероятностью 0,95, что средний рост студента будет колебаться в пределах от 165 до 170 см, то есть из 100 студентов примерно 95 будут иметь средний рост в указанных пределах, в интервале (165; 170) будут находиться основные значения роста студентов.

3. Интерпретация среднего квадратичного отклонения σ_e .

Так как $\sigma_e = 9,1$ см., то можно сделать вывод, что отклонение роста отдельно взятого студента в среднем составляет 9,1 см.

4. Интерпретация доверительного интервала $s \cdot (-q) < \sigma_r < s \cdot (+q)$.

По расчетам получили, что $7 < \sigma_r < 11$, следовательно, можно сделать вывод, что возможные отклонения роста студента с вероятностью 0,95 будут составлять значения, заключенные в промежутке (7; 11) см.

5. Интерпретация коэффициента асимметрии A_s .

Значение $A_s \neq 0$, это говорит о том, что изменения роста студентов в сторону увеличения или уменьшения по отношению к среднему значению происходит неодинаково, так как $A_s = 0,05 > 0$, то наблюдается левосторонний скос, т.е. рост выше среднего событие более достоверное.

6. Интерпретация полученного значения эксцесса E_x .

Значение $E_x = -1,03$, т.к. $E_x \neq 0$, то отклонение от среднего значения \bar{x}_e происходит не плавно и наблюдается отклонение от нормы, а так как $E_x = -1,03 < 0$, E_x то отклонения большие, больше нормы, то есть среднее значение резко отличается от других возможных значений случайной величины.

7. Интерпретация значения коэффициента вариации V .

Значение $V = 5,4\%$ - это меньше 10%, следовательно, изменчивость незначительная.

Задания

1. Дана величина заработной платы специалистов компании (в сотнях руб.).

| | | | | | | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 120 | 480 | 175 | 490 | 410 | 425 | 430 | 385 | 335 | 315 | 545 | 265 | 390 | 395 | 360 |
| 475 | 480 | 225 | 445 | 255 | 425 | 375 | 325 | 320 | 160 | 245 | 395 | 390 | 370 | 380 |
| 255 | 265 | 445 | 425 | 265 | 410 | 305 | 330 | 335 | 455 | 220 | 370 | 340 | 340 | |
| 265 | 275 | 435 | 275 | 415 | 310 | 340 | 245 | 215 | 215 | 275 | 325 | 330 | 340 | |
| 355 | 415 | 285 | 265 | 315 | 345 | 345 | 230 | 225 | 285 | 365 | 315 | 255 | 275 | |
| 325 | 335 | 295 | 275 | 330 | 315 | 340 | 345 | 215 | 295 | 375 | 345 | 325 | 325 | |
| 225 | 255 | 370 | 385 | 355 | 360 | 285 | 345 | 305 | 310 | 235 | 325 | 320 | 335 | |

2. Имеются данные социологического опроса студентов о влиянии престижа вуза на выбор профессии (в %):

| | | | | | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 85 | 76 | 80 | 84 | 88 | 89 | 91 | 88 | 75 | 85 | 79 | 82 | 87 | 90 | 83 |
| 76 | 82 | 86 | 89 | 88 | 84 | 90 | 89 | 85 | 91 | 89 | 95 | 89 | 79 | 86 |
| 87 | 81 | 78 | 85 | 88 | 91 | 89 | 87 | 74 | 81 | 85 | 91 | 97 | 90 | |
| 87 | 90 | 88 | 86 | 76 | 84 | 88 | 77 | 88 | 82 | 95 | 91 | 84 | 91 | |
| 85 | 84 | 74 | 80 | 84 | 91 | 99 | 90 | 87 | 77 | 88 | 85 | 81 | 88 | |
| 83 | 89 | 91 | 92 | 88 | 94 | 90 | 88 | 81 | 83 | 90 | 98 | 92 | 86 | |
| 89 | 94 | 96 | 89 | 88 | 95 | 99 | 90 | 86 | 78 | 87 | 90 | 93 | 90 | |

3. Опрашивалось 100 человек о доверии руководству фирмы. Получены данные (в %).

| | | | | | | | | | | | | | | |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 78 | 90 | 90 | 86 | 81 | 77 | 83 | 85 | 92 | 86 | 84 | 84 | 96 | 78 | 84 |
| 73 | 75 | 83 | 78 | 73 | 84 | 85 | 83 | 88 | 76 | 89 | 78 | 92 | 85 | 83 |
| 87 | 85 | 87 | 89 | 83 | 76 | 77 | 84 | 83 | 89 | 79 | 81 | 76 | 86 | |
| 87 | 76 | 82 | 89 | 74 | 89 | 82 | 87 | 71 | 78 | 75 | 80 | 92 | 74 | |
| 85 | 84 | 81 | 83 | 88 | 81 | 83 | 80 | 79 | 82 | 83 | 88 | 81 | 82 | |
| 86 | 74 | 91 | 78 | 93 | 84 | 81 | 76 | 74 | 81 | 82 | 86 | 76 | 90 | |
| 93 | 83 | 92 | 91 | 83 | 79 | 84 | 90 | 80 | 84 | 81 | 78 | 88 | 85 | |

Вопросы для самоконтроля

1. Какая совокупность называется генеральной (выборочной)?
2. Какая выборка называется репрезентативной?
3. Какие способы формирования выборки вы знаете?
4. Когда выборка называется повторной (бесповторной)?
5. Какие вариационные ряды вы знаете?
6. Можно ли от дискретного ряда перейти к интервальному и наоборот?

7. От чего зависит число интервалов группировки?
8. Как от простой статистической таблицы данных перейти к вариационному ряду?
9. Как графически изобразить дискретный (непрерывный) вариационный ряд?
10. Каким свойством обладает выборочное среднее?
11. Как вычислить моду и медиану дискретного (интервального) вариационного ряда?
12. Как вычислить дисперсию (среднее квадратичное отклонение) выборочной совокупности?
13. Какими свойствами обладает дисперсия?
14. Что характеризует коэффициент асимметрии?
15. Что характеризует эксцесс?
16. Какое число принимается в качестве ложного нуля?
17. По каким формулам от числовых характеристик, вычисленных в условных вариантах производится переход к числовым характеристикам в первоначальных вариантах?
18. Какая оценка называется точечной?
19. Что является несмещенной оценкой математического ожидания генеральной совокупности?
20. Как вычисляется несмещенная оценка дисперсии генеральной совокупности?
21. Какая оценка называется интервальной?
22. Какая вероятность называется надежностью? Дайте пояснения на графике нормального закона распределения.
23. По каким формулам находится доверительный (надежностный) интервал для генеральной средней и генерального среднеквадратичного отклонения?

24. Как определить минимальный объем выборки, чтобы с заданной надежностью γ гарантировать определенную точность Δ вычисления генеральной средней?

25. Как ведет себя предельная ошибка $\frac{t \cdot \sigma}{\sqrt{n}}$ (точность оценки), если надежность увеличивается?

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ ПО ТЕМЕ № 2

«Парная регрессия. Метод наименьших квадратов (МНК) для линейной регрессии»

Цель занятия: овладеть навыками корреляционно-регрессивного анализа между признаками X и Y .

Методические указания к проведению занятия

Пример 1. По 10 сельскохозяйственным предприятиям имеются данные о себестоимости молока и средней продуктивности молока (табл. 5).

Таблица 5

| | | | | | | | | | | |
|-----------------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Себестоимость молока, руб./л | 7,5 | 6,0 | 5,2 | 8,3 | 5,8 | 6,9 | 7,8 | 7,0 | 5,9 | 8,0 |
| Средняя продуктивность молока, кг | 197 | 168 | 135 | 259 | 151 | 186 | 233 | 213 | 147 | 234 |

Требуется: Рассчитать параметры уравнения парной линейной регрессии зависимости себестоимости молока от средней продуктивности.

Уравнение парной линейной регрессии имеет вид:

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

где \hat{y}_x – себестоимость молока, руб./л;

x – средняя продуктивность молока, кг;

a, b – параметры уравнения.

Для определения параметров уравнения a и b составим систему нормальных уравнений. Исходное уравнение последовательно умножим на

коэффициенты при неизвестных a и b , и затем каждое уравнение просуммируем:

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

где n – число единиц совокупности.

Для расчетов построим вспомогательную таблицу (табл. 6).

Таблица 6

| Сельскохозяйственное предприятие | Себестоимость молока, руб./л | Средняя продуктивность молока, кг | y^2 | x^2 | xy | \hat{y}_x | $y - \hat{y}_x$ |
|----------------------------------|------------------------------|-----------------------------------|--------|--------|---------|-------------|-----------------|
| | y | x | | | | | |
| 1 | 7,5 | 197 | 56,25 | 38809 | 1477,5 | 6,96 | 0,54 |
| 2 | 6,0 | 168 | 36,00 | 28224 | 1008,0 | 6,25 | -0,25 |
| 3 | 5,2 | 135 | 27,04 | 18225 | 702,0 | 5,44 | -0,24 |
| 4 | 8,3 | 259 | 68,89 | 67081 | 2149,7 | 8,48 | -0,18 |
| 5 | 5,8 | 151 | 33,64 | 22801 | 875,8 | 5,83 | -0,03 |
| 6 | 6,9 | 186 | 47,61 | 34596 | 1283,4 | 6,69 | 0,21 |
| 7 | 7,8 | 233 | 60,84 | 54289 | 1817,4 | 7,84 | -0,04 |
| 8 | 7,0 | 213 | 49,00 | 45369 | 1491,0 | 7,35 | -0,35 |
| 9 | 5,9 | 147 | 34,81 | 21609 | 867,3 | 5,73 | 0,17 |
| 10 | 8,0 | 234 | 64,00 | 54756 | 1872,0 | 7,86 | 0,14 |
| Сумма | 68,4 | 1923 | 478,09 | 385759 | 13544,1 | 68,43 | -0,03 |

Подставим полученные данные в систему уравнений:

$$\begin{cases} 68,4 = 10a + 1923b \\ 13544,1 = 1923a + 385759b \end{cases}$$

Разделим каждый член уравнений на коэффициенты при a (в первом уравнении на 10, во втором на 1923):

$$\begin{cases} 6,84 = a + 192,3b \\ 7,043 = a + 200,6b \end{cases}$$

Вычтем из второго уравнения первое и найдем параметр b :

$$0,203 = 8,3b; \quad b = 0,0245.$$

Подставив значение b в первое уравнение, найдем значение a :

$$a = 6,84 - 192,3 \cdot 0,0245 = 2,13.$$

Параметры уравнения регрессии можно определить и по другим формулам, которые вытекают из системы нормальных уравнений:

$$a = \frac{\Sigma y \Sigma x^2 - \Sigma xy \Sigma x}{n \Sigma x^2 - (\Sigma x)^2} = \frac{68,4 \cdot 385759 - 135441 \cdot 1923}{10 \cdot 385759 - 1923^2} = 2,13$$

$$b = \frac{n \Sigma xy - \Sigma y \Sigma x}{n \Sigma x^2 - (\Sigma x)^2} = \frac{10 \cdot 135441 - 68,4 \cdot 1923}{10 \cdot 385759 - 1923^2} = 0,0245$$

Уравнение регрессии имеет вид: $\hat{y}_x = 2,13 + 0,0245x$.

Пример 2. По данным проведенного опроса восьми групп семей известны данные связи расходов населения на продукты питания с уровнем доходов семьи.

Таблица 7

| | | | | | | | | |
|--|-----|-----|-----|-----|-----|------|------|------|
| Расходы на продукты питания, y , тыс. руб. | 0,9 | 1,2 | 1,8 | 2,2 | 2,6 | 2,9 | 3,3 | 3,8 |
| Доходы семьи, x , тыс. руб. | 1,2 | 3,1 | 5,3 | 7,4 | 9,6 | 11,8 | 14,5 | 18,7 |

Предположим, что связь между доходами семьи и расходами на продукты питания линейная. Для подтверждения нашего предположения построим поле корреляции.

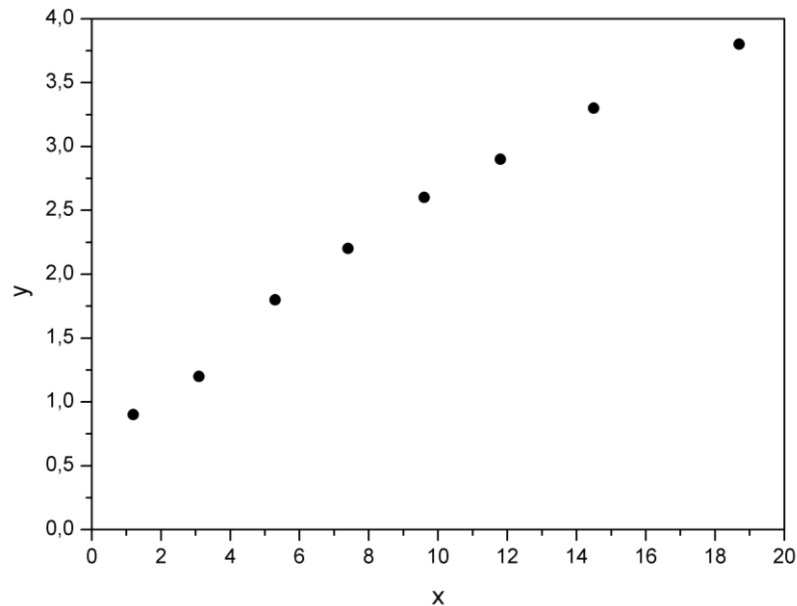


Рисунок 3 - Поле корреляции

По графику видно, что точки выстраиваются в некоторую прямую линию.

Для удобства дальнейших вычислений составим таблицу.

Таблица 8

| | x | y | xy | x^2 | y^2 | $\$y_x$ | $y - \$y_x$ |
|------------------|-------|-------|--------|--------|-------|---------|-------------|
| 1 | 1,2 | 0,9 | 1,08 | 1,44 | 0,81 | 1,038 | -0,138 |
| 2 | 3,1 | 1,2 | 3,72 | 9,61 | 1,44 | 1,357 | -0,157 |
| 3 | 5,3 | 1,8 | 9,54 | 28,09 | 3,24 | 1,726 | 0,074 |
| 4 | 7,4 | 2,2 | 16,28 | 54,76 | 4,84 | 2,079 | 0,121 |
| 5 | 9,6 | 2,6 | 24,96 | 92,16 | 6,76 | 2,449 | 0,151 |
| 6 | 11,8 | 2,9 | 34,22 | 139,24 | 8,41 | 2,818 | 0,082 |
| 7 | 14,5 | 3,3 | 47,85 | 210,25 | 10,89 | 3,272 | 0,028 |
| 8 | 18,7 | 3,8 | 71,06 | 349,69 | 14,44 | 3,978 | -0,178 |
| Итого | 71,6 | 18,7 | 208,71 | 885,24 | 50,83 | 18,717 | -0,017 |
| Среднее значение | 8,95 | 2,34 | 26,09 | 110,66 | 6,35 | 2,34 | - |
| σ | 5,53 | 0,935 | - | - | - | - | - |
| σ^2 | 30,56 | 0,874 | - | - | - | - | - |

Рассчитаем параметры линейного уравнения парной регрессии

$\$y_x = a + b \cdot x$. Для этого воспользуемся формулами:

$$b = \frac{\text{cov } x, y}{\sigma_x^2} = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{26,09 - 8,95 \cdot 2,34}{30,56} = 0,168$$

$$a = \bar{y} - b \cdot \bar{x} = 2,34 - 0,168 \cdot 8,95 = 0,836.$$

Получили уравнение: $\$y_x = 0,836 + 0,168 \cdot x$. Т.е. с увеличением дохода семьи на 1000 руб. расходы на питание увеличиваются на 168 руб.

Задание

При изучении химического состава и реологических характеристик желе из ягод смородины было обследовано 10 образцов и получены следующие результаты о содержании сухих веществ X (%) и прочности студня Y (кг), варианты заданий представлены в таблице 9.

В соответствии с выбранным вариантом произвести статистическую обработку данных:

- построить диаграмму рассеяния;
- полагая, что между признаками X и Y имеет место линейная корреляционная связь, найти выборочное уравнение линейной регрессии.

Используя полученное уравнение регрессии, оценить ожидаемое среднее значение признака Y , когда признак X принимает значение, равное a %.

- построить линию регрессии.

Таблица 9

| Варианты заданий | | | | | | | | | | | |
|----------------------|--------------------------------------|----------------------|--------------------------------------|----------------------|--------------------------------------|----------------------|--------------------------------------|----------------------|--------------------------------------|----------------------|--------------------------------------|
| №1 | | №2 | | №3 | | №4 | | №5 | | №6 | |
| Прочность студня, кг | Содержание сухих веществ в студне, % | Прочность студня, кг | Содержание сухих веществ в студне, % | Прочность студня, кг | Содержание сухих веществ в студне, % | Прочность студня, кг | Содержание сухих веществ в студне, % | Прочность студня, кг | Содержание сухих веществ в студне, % | Прочность студня, кг | Содержание сухих веществ в студне, % |
| 0,11 1 | 59,0 | 0,27 8 | 64,0 | 0,05 0 | 55,0 | 0,32 3 | 66,0 | 0,22 2 | 63,0 | 0,30 1 | 66,0 |
| 0,12 5 | 59,5 | 0,28 4 | 65,0 | 0,08 4 | 56,0 | 0,36 9 | 67,0 | 0,23 7 | 64,0 | 0,32 1 | 67,0 |
| 0,12 6 | 60,0 | 0,29 1 | 66,0 | 0,09 1 | 57,0 | 0,39 5 | 68,0 | 0,26 5 | 65,0 | 0,35 8 | 68,0 |
| 0,16 5 | 61,2 | 0,30 2 | 67,0 | 0,10 1 | 58,0 | 0,40 1 | 69,0 | 0,27 5 | 66,0 | 0,37 9 | 69,0 |
| 0,20 1 | 62,0 | 0,32 3 | 68,0 | 0,12 5 | 59,0 | 0,42 3 | 70,0 | 0,30 3 | 67,0 | 0,40 2 | 70,0 |
| 0,21 0 | 63,0 | 0,37 5 | 69,0 | 0,14 7 | 60,0 | 0,43 7 | 71,0 | 0,34 5 | 68,0 | 0,42 3 | 71,0 |
| 0,24 6 | 64,0 | 0,40 1 | 70,0 | 0,17 5 | 61,0 | 0,47 9 | 72,0 | 0,37 8 | 69,0 | 0,43 5 | 72,0 |
| 0,30 2 | 65,0 | 0,42 1 | 71,0 | 0,19 5 | 62,0 | 0,49 2 | 73,0 | 0,40 1 | 70,0 | 0,46 9 | 73,0 |
| 0,33 3 | 66,0 | 0,52 0 | 72,0 | 0,20 1 | 63,0 | 0,50 2 | 74,0 | 0,41 9 | 71,0 | 0,50 1 | 74,0 |
| 0,35 2 | 67,0 | 0,53 0 | 73,0 | 0,22 2 | 64,0 | 0,50 9 | 75,0 | 0,42 5 | 72,0 | 0,50 9 | 75,0 |
| | $a=61,0$ | | $a=65,5$ | | $a=56,5$ | | $a=67,5$ | | $a=65,5$ | | $a=68,5$ |

Вопросы для самоконтроля

1. Как рассчитываются параметры парной линейной регрессии?
2. Как провести оценку статистической значимости параметров уравнения парной регрессии?
3. Поясните смысл коэффициента корреляции, как оценить его значимость?

4. Как определяется число степеней свободы для факторной и остаточной сумм квадратов?

5. Дайте определение бета–коэффициента. Поясните его смысл.

6. Дайте определение коэффициента эластичности. Поясните его смысл. Как определяется коэффициент эластичности по разным видам регрессионных моделей?

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ ПО ТЕМЕ № 3

«Проверка качества регрессии»

Цель занятия: ознакомить аспирантов с процессом проверки качества уравнения регрессии, показателями качества регрессии.

Методические указания к проведению занятия

Для практического использования моделей регрессии большое значение имеет их адекватность, т.е. соответствие фактическим статистическим данным. При анализе адекватности уравнения регрессии (модели) исследуемому процессу, возможны следующие варианты:

1. Построенная модель на основе **F-критерия Фишера** в целом адекватна и все коэффициенты регрессии значимы. Такая модель может быть использована для принятия решений и осуществления прогнозов.

2. Модель по F-критерию Фишера адекватна, но часть коэффициентов не значима. Модель пригодна для принятия некоторых решений, но не для прогнозов.

3. Модель по F-критерию адекватна, но все коэффициенты регрессии не значимы. Модель полностью считается неадекватной. На ее основе не принимаются решения и не осуществляются прогнозы.

Корреляционный и регрессионный анализ, как правило, проводится для ограниченной по объёму совокупности. Поэтому показатели регрессии и корреляции – параметры уравнения регрессии, коэффициент корреляции и коэффициент детерминации могут быть искажены действием случайных факторов. Чтобы проверить, на сколько эти показатели характерны для всей

генеральной совокупности, не являются ли они результатом стечения случайных обстоятельств, необходимо проверить адекватность построенных статистических моделей.

Проверить значимость уравнения регрессии – значит установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным и достаточно ли включенных в уравнение объясняющих переменных (одной или нескольких) для описания зависимой переменной. Чтобы иметь общее суждение о качестве модели, из относительных отклонений по каждому наблюдению определяют среднюю ошибку аппроксимации. Проверка адекватности уравнения регрессии (модели) осуществляется с помощью средней ошибки аппроксимации, величина которой не должна превышать **10-12%** (рекомендовано).

$$A = \frac{1}{n} \sum \left| \frac{y_i - \hat{y}_x}{y_i} \right| \cdot 100$$

Оценка значимости уравнения регрессии в целом производится на основе **F-критерия Фишера**, которому предшествует **дисперсионный анализ**. В математической статистике **дисперсионный анализ** рассматривается как самостоятельный инструмент статистического анализа. В эконометрике он применяется как вспомогательное средство для изучения качества регрессионной модели. Согласно основной идее **дисперсионного анализа**, **общая сумма квадратов отклонений** переменной (y) от среднего значения ($y_{\text{ср.}}$) раскладывается на две части – **«объясненную»** и **«необъясненную»**:

$$\sum (y - \bar{y})^2 = \sum (\hat{y}_x - \bar{y})^2 + \sum (y - \hat{y}_x)^2$$

Схема дисперсионного анализа имеет следующий вид (n – число наблюдений, m – число параметров при переменной x):

| Компоненты дисперсии | Сумма квадратов | Число степеней свободы | Дисперсия на одну степень свободы |
|----------------------|--------------------------------|------------------------|---|
| Общая | $\sum (y - \bar{y})^2$ | $n - 1$ | $S_{\text{общ}}^2 = \frac{\sum (y - \bar{y})^2}{n - 1}$ |
| Факторная | $\sum (\hat{y}_x - \bar{y})^2$ | m | $S_{\text{факт}}^2 = \frac{\sum (\hat{y}_x - \bar{y})^2}{m}$ |
| Остаточная | $\sum (y - \hat{y}_x)^2$ | $n - m - 1$ | $S_{\text{ост}}^2 = \frac{\sum (y - \hat{y}_x)^2}{n - m - 1}$ |

Определение дисперсии на одну степень свободы приводит дисперсии к сравнимому виду. Сопоставляя факторную и остаточную дисперсии в расчете на одну степень свободы, получим величину **F-критерия Фишера**. Фактическое значение F-критерия Фишера сравнивается с табличным значением $F_{\text{табл.}}(\alpha, k_1, k_2)$ при заданном уровне значимости α и степенях свободы $k_1 = m$ и $k_2 = n - m - 1$. При этом, если фактическое значение F-критерия больше табличного $F_{\text{факт}} > F_{\text{теор}}$, то признается **статистическая значимость уравнения в целом**. Для парной линейной регрессии $m = 1$, поэтому:

$$F = \frac{S_{\text{факт}}^2}{S_{\text{ост}}^2} = \frac{\sum (\hat{y}_x - \bar{y})^2}{\sum (y - \hat{y}_x)^2} \cdot (n - 2)$$

Эта формула в общем виде может выглядеть так:

$$F = \frac{\sigma_{\text{факт}}^2 (n - m)}{\sigma_{\text{ост}}^2 (m - 1)} \Rightarrow \text{сравнить с } F_{\text{табл}}$$

Отношение **объясненной** части дисперсии переменной (y) к **общей дисперсии** называют **коэффициентом детерминации** и используют для характеристики **качества уравнения регрессии** или соответствующей модели связи. Соотношение между **объясненной** и **необъясненной** частями **общей дисперсии** можно представить в альтернативном варианте:

$$R^2 = \frac{\sum (\tilde{y}_x - \bar{y})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum (y_i - \tilde{y}_x)^2}{\sum (y_i - \bar{y})^2}$$

Коэффициент детерминации R^2 принимает значения в диапазоне от нуля до единицы $0 \leq R^2 \leq 1$. **Коэффициент детерминации R^2** показывает, какая часть дисперсии результативного признака (y) объяснена уравнением

регрессии. Чем больше R^2 , тем большая часть дисперсии результативного признака (y) объясняется уравнением регрессии и тем лучше уравнение регрессии описывает исходные данные. При отсутствии зависимости между (y) и (x) коэффициент детерминации R^2 будет близок к нулю. Таким образом, **коэффициент детерминации R^2** может применяться для **оценки качества (точности) уравнения регрессии**. Возникает вопрос, при каких значениях R^2 уравнение регрессии следует считать статистически незначимым, что делает необоснованным его использование в анализе? Ответ на этот вопрос дает **F-критерий Фишера** $F_{\text{факт}} > F_{\text{теор}}$ - делаем вывод о статистической значимости уравнения регрессии. Величина F-критерия связана с коэффициентом детерминации $R^2_{xy}(r^2_{xy})$, и ее можно рассчитать по следующей формуле:

$$F = \frac{r^2_{xy}}{1 - r^2_{xy}} \cdot (n - 2)$$

Либо при оценке значимости **индекса** (аналог **коэффициента**) **детерминации**:

$$i^2(n^2) = 1 - \frac{\sum (y_i - \tilde{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

Использование **коэффициента множественной детерминации R^2** для оценки качества модели, обладает тем недостатком, что включение в модель нового фактора (даже несущественного) автоматически увеличивает величину R^2 . Поэтому, при большом количестве факторов, предпочтительнее использовать, так называемый, улучшенный, скорректированный коэффициент множественной детерминации R^2 , определяемый соотношением:

$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2 : (n - p - 1)}{\sum_{i=1}^n (y_i - \bar{y})^2 : (n - 1)} = 1 - \frac{n - 1}{n - p - 1} (1 - R^2)$$

где p – число факторов в уравнении регрессии, n – число наблюдений. Чем больше величина p, тем сильнее различия между множественным коэффициентом детерминации R^2 и скорректированным R^2 . При

использовании скорректированного R^2 , для оценки целесообразности включения фактора в уравнение регрессии, следует учитывать, что увеличение его величины (значения), при включении нового фактора, не обязательно свидетельствует о его значимости, так как значение увеличивается всегда, когда t -статистика больше единицы ($|t| > 1$). При заданном объеме наблюдений и при прочих равных условиях, с увеличением числа независимых переменных (параметров), скорректированный коэффициент множественной детерминации убывает. При небольшом числе наблюдений, скорректированная величина коэффициента множественной детерминации R^2 имеет тенденцию переоценивать долю вариации результативного признака, связанную с влиянием факторов, включенных в регрессионную модель. Низкое значение коэффициента множественной корреляции и коэффициента множественной детерминации R^2 может быть обусловлено следующими причинами: в регрессионную модель не включены существенные факторы; неверно выбрана форма аналитической зависимости, не реально отражающая соотношения между переменными, включенными в модель.

Для оценки значимости парного коэффициента корреляции (корень квадратный из коэффициента детерминации), при условии линейной формы связи между факторами, можно использовать **t-критерий Стьюдента**:

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

При численности объектов анализа до 30 единиц возникает необходимость проверки значимости (существенности) каждого коэффициента регрессии. При этом выясняют насколько вычисленные параметры характерны для отображения комплекса условий: не являются ли полученные значения параметров результатами действия случайных причин. Значимость коэффициентов простой линейной регрессии (применительно к совокупностям, у которых $n < 30$) осуществляют с помощью **t-критерия**

Стьюдента. При этом вычисляют расчетные (фактические) значения t -критерия для параметров a_0, a_1 :

$$t_{a_0} = \frac{a_0 \sqrt{n-2}}{\sigma_\varepsilon} \quad t_{a_1} = \frac{a_1 \sqrt{n-2}}{\sigma_\varepsilon} \sigma_x$$

$$\sigma_\varepsilon = \sqrt{\frac{\sum(Y_i - \tilde{Y})^2}{n-m}} \quad \sigma_x = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-m}}$$

n -число наблюдений, m -число параметров уравнения регрессии, σ_ε - (остаточное) среднее квадратическое отклонение результативного признака от выровненных значений \hat{y} ; σ_x - среднее квадратическое отклонение факторного признака от общей средней.

Вычисленные, по вышеприведенным формулам, значения сравнивают с критическими t , которые определяют по таблице значений Стьюдента с учетом принятого уровня значимости α и числа степеней свободы вариации $k(v)=n-2$. В социально-экономических исследованиях уровень значимости α обычно принимают равным 0,05. Параметр признаётся значимым (существенным) при условии, если $t_{расч.} > t_{табл.}$. В этом случае, практически невероятно, что найденные значения параметров обусловлены только случайными совпадениями.

Критерий Дарбина – Уотсона (Durbin - Watson)

Оценивая качество уравнения регрессии, мы предполагаем, что реальная взаимосвязь переменных линейна. Отклонения от регрессионной прямой являются случайными, независимыми друг от друга величинами с нулевым математическим ожиданием и постоянной дисперсией. Если эти предположения не выполняются, то оценки коэффициентов регрессии не обладают свойствами *несмещенности, эффективности и состоятельности*. В этом случае анализ значимости полученных оценок будет неточным.

Статистика **Дарбина—Уотсона** используется для проверки гипотезы о том, что остатки построенной регрессионной модели некоррелированы (корреляции равны нулю), против альтернативы: остатки связаны авторегрессионной зависимостью (первого порядка) вида:

$$\varepsilon_i = \rho \varepsilon_{i-1} + \delta_i$$

На практике для анализа коррелированности отклонений вместо коэффициента корреляции используют тесно с ним связанную статистику Дарбина—Уотсона, рассчитываемую по формуле

$$d = \frac{\sum (e_i - e_{i-1})^2}{\sum e_i^2}.$$

Критические точки статистики Дарбина—Уотсона табулированы для различных α , m , n . При проверке гипотезы об отсутствии автокорреляции остатков используется числовой отрезок, на котором отложены d_l – нижняя граница статистики и d_u – верхняя граница:



Статистика Дарбина—Уотсона

Проверка гипотезы проводится по схеме:

- Если $d < d_l$, то гипотеза H_0 отклоняется, принимается H_1 - значительная положительная автокорреляция остатков;
- Если $d > 4 - d_l$, то гипотеза H_0 отклоняется, принимается H_1 - значительная отрицательная автокорреляция остатков;
- Если $d_u < d < 4 - d_u$, то гипотеза H_0 об отсутствии автокорреляции остатков принимается;
- Если $d_l < d < d_u$, или $4 - d_u < d < 4 - d_l$, то гипотеза об отсутствии автокорреляции не может быть ни принята, ни отклонена.

Не обращаясь к таблице критических точек Дарбина—Уотсона можно воспользоваться «грубым» правилом и считать, что автокорреляция остатков отсутствует, если $1,5 < d < 2,5$. Для более надежных выводов необходимо воспользоваться статистическими таблицами.

Пример. Рассчитать параметры уравнения парной линейной регрессии и проверить качество.

Таблица 10

| | | | | | | | | |
|-----|----|----|----|----|----|----|----|----|
| x | 23 | 29 | 33 | 35 | 38 | 40 | 43 | 45 |
| y | 88 | 83 | 73 | 78 | 68 | 53 | 58 | 48 |

Построим поле корреляции. Как видно из рисунка 4 данные имеют линейную зависимость.

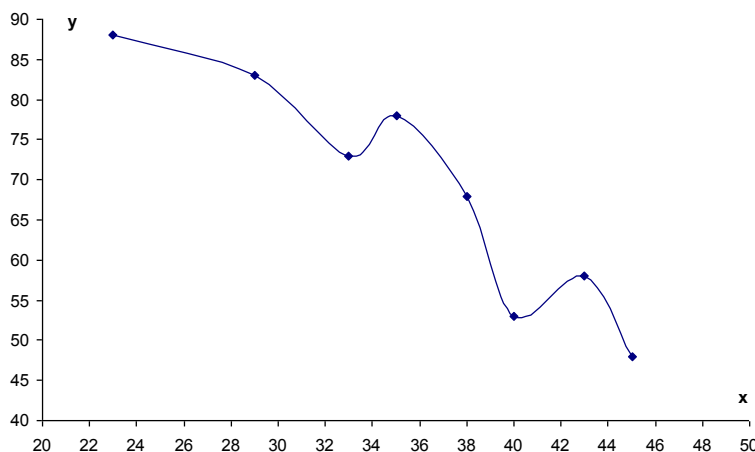


Рисунок 4 - Поле корреляции

Уравнение линейной регрессии имеет вид: $y = a_0 + a_1 \cdot x$.

Составим расчетную таблицу.

Таблица 11

| x | y | xy | x^2 | y^2 | \tilde{y} | $y - \tilde{y}$ | $(y - \tilde{y})^2$ |
|------------|------------|--------------|--------------|--------------|----------------|-----------------|---------------------|
| 23 | 88 | 2024 | 529 | 7744 | 92,225 | -4,225 | 17,850625 |
| 29 | 83 | 2407 | 841 | 6889 | 81,119 | 1,881 | 3,538161 |
| 33 | 73 | 2409 | 1089 | 5329 | 73,715 | -0,715 | 0,511225 |
| 35 | 78 | 2730 | 1225 | 6084 | 70,013 | 7,987 | 63,792169 |
| 38 | 68 | 2584 | 1444 | 4624 | 64,46 | 3,54 | 12,531600 |
| 40 | 53 | 2120 | 1600 | 2809 | 60,758 | -7,758 | 60,186564 |
| 43 | 58 | 2494 | 1849 | 3364 | 55,205 | 2,795 | 7,812025 |
| 45 | 48 | 2160 | 2025 | 2304 | 51,503 | -3,503 | 12,271009 |
| 286 | 549 | 18928 | 10602 | 39147 | 548,998 | 0,002 | 178,493378 |

Получим:

$$\overline{xy} = \frac{18928}{8} = 2366 \quad \overline{y} = \frac{549}{8} = 68,625 \quad \overline{x} = \frac{286}{8} = 35,75$$

$$\overline{x^2} = \frac{10602}{8} = 1325,25 \quad \overline{y^2} = \frac{39147}{8} = 4893,375$$

Следовательно:

$$a_1 = \frac{2366 - 35,75 \cdot 68,625}{1325,25 - (35,75)^2} = \frac{-87,34375}{47,1875} \approx -1,851$$

$$a_0 = 68,625 + 1,851 \cdot 35,75 \approx 134,798$$

Уравнение примет вид: $\tilde{y} = 134,798 - 1,851 \cdot x$.

Рассчитаем коэффициент парной корреляции:

$$r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y}, \quad \text{где } \sigma_x = \sqrt{x^2 - (\bar{x})^2} = \sqrt{1325,25 - (35,75)^2} \approx 6,869$$

$$\sigma_y = \sqrt{y^2 - (\bar{y})^2} = \sqrt{4893,375 - (68,625)^2} = \sqrt{183,984375} \approx 13,564$$

$$\text{Тогда } r_{xy} = \frac{2366 - 35,75 \cdot 68,625}{6,869 \cdot 13,564} \approx -0,937$$

Связь между факторами обратная и сильная.

Коэффициент детерминации $r_{xy}^2 \approx 0,878$, т.е. 87,8 % вариации изменения y объясняется вариацией изменения x .

Ошибка модели рассчитывается следующим образом:

$$\Delta = \frac{\sqrt{\frac{\sum (y_t - \tilde{y}_t)^2}{n - m - 1}}}{y_t} \cdot 100 = \frac{\sqrt{\frac{178,493378}{8 - 1 - 1}}}{68,625} \cdot 100 \approx 7,9\%$$

Модель хорошего качества.

Оценим качество модели с помощью F-критерия Фишера.

$$F_{\text{факт}} = \frac{r_{xy}^2}{1 - r_{xy}^2} (n - 2) \Rightarrow F_{\text{факт}} = \frac{0,878}{1 - 0,878} (8 - 2) = 43,18$$

При уровне значимости $\alpha = 0,05$ (вероятность 95 %) табличное значение равно 5,99. $F_{\text{факт}} > F_{\text{табл}}$ т.е., гипотеза о случайной природе оцениваемых характеристик не принимается, признается их значимость. Следовательно, модель надежна.

Вычислим фактическое значение критерия Дарбина-Уотсона для этой модели (таблица 12):

Таблица 12

Расчет критерия Дарбина-Уотсона

| ε_t | ε_{t-1} | $\varepsilon_t - \varepsilon_{t-1}$ | $(\varepsilon_t - \varepsilon_{t-1})^2$ | ε_t^2 |
|-----------------|---------------------|-------------------------------------|---|-------------------|
| -4,225 | - | - | - | 17,850625 |

| | | | | |
|--------------|--------------|--------------|-------------------|-------------------|
| 1,881 | -4,225 | 6,106 | 37,283236 | 3,538161 |
| -0,715 | 1,881 | -2,596 | 6,739216 | 0,511225 |
| 7,987 | -0,715 | 8,702 | 75,724804 | 63,792169 |
| 3,54 | 7,987 | -4,447 | 19,775809 | 12,5316 |
| -7,758 | 3,54 | -11,298 | 127,644804 | 60,186564 |
| 2,795 | -7,758 | 10,553 | 111,365809 | 7,812025 |
| -3,503 | 2,795 | -6,298 | 39,664804 | 12,271009 |
| 0,002 | 3,505 | 0,722 | 418,198482 | 178,493378 |

$$d_{расч} = \frac{\sum_{t=2}^n (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^n \varepsilon_t^2} = \frac{418,198482}{178,493378} = 2,34$$

Зададим уровень значимости $\alpha = 0,05$. По таблицам значений критерия Дарбина-Уотсона определим для числа наблюдений $n = 8$ и числа независимых переменных модели $k = 1$ критические значения $d_1 = 1,08$ и $d_2 = 1,36$.

$d_{расч} > d_2$, следовательно, форма модели выбрана верно.

Задание

Проверить качество моделей, рассчитанных на занятия №3.

Вопросы для самоконтроля

1. Три группы оценок качества моделирования.
2. Показатели близости и адекватности.
3. Поле корреляции.

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ ПО ТЕМЕ № 4

«Множественная линейная регрессия»

Цель занятия: закрепление теоретического и практического материала по теме множественная регрессия, приобретение навыков построения и анализа многофакторных эконометрических моделей.

Методические указания к проведению занятия

В линейной множественной регрессии $\tilde{y}_x = a + b_1 \cdot x_1 + b_2 \cdot x_2 + K + b_p \cdot x_p$, параметры при x называются коэффициентами «чистой» регрессии. Они характеризуют среднее изменение результата с изменением соответствующего параметра на единицу при неизменном значении других факторов, закрепленных на среднем уровне.

Классический подход к оцениванию параметров линейной модели основан на **методе наименьших квадратов (МНК)**.

Этот метод позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака (y) от расчетных (теоретических) \tilde{y}_x минимальна:

$$\sum_i (y_i - \tilde{y}_{x_i})^2 \rightarrow \min .$$

Чтобы найти минимум функции, надо вычислить производные по каждому из параметров и приравнять их к нулю, т.к. равенство нулю производной – необходимое условие экстремума. В результате получается система уравнений, решение которой и позволяет получить оценки параметров регрессии.

Система нормальных уравнений имеет вид:

$$\begin{cases} \sum y = a \cdot n + b_1 \cdot \sum x_1 + b_2 \cdot \sum x_2 + K + b_p \cdot \sum x_p \\ \sum y \cdot x_1 = a \cdot \sum x_1 + b_1 \cdot \sum x_1^2 + b_2 \cdot \sum x_1 \cdot x_2 + K \cdot \sum x_1 + b_p \cdot \sum x_1 \cdot x_p \\ \text{К К} \\ \sum y \cdot x_p = a \cdot \sum x_p + b_1 \cdot \sum x_p \cdot x_1 + b_2 \cdot \sum x_p \cdot x_2 + K \cdot \sum x_p + b_p \cdot \sum x_p^2 \end{cases}$$

Решение системы может быть осуществлено по одному из известных способов: Метод Гаусса, метод Крамера и т.д.

Пример 1. По четырем предприятиям региона изучается зависимость выработки продукции на одного работника y (тыс. руб.) от ввода в действие новых основных фондов x_2 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_1 (%). Требуется написать уравнение множественной регрессии.

| | | | | |
|-------------------|---|---|---|---|
| Номер предприятия | 1 | 2 | 3 | 4 |
| x_1 , (%) | 1 | 2 | 3 | 5 |

| | | | | |
|-------------------------|---|----|----|----|
| $x_2, (\%)$ | 0 | 1 | 3 | 4 |
| $y, (\text{тыс. руб.})$ | 6 | 11 | 19 | 28 |

Предположим, что зависимость выработки продукции на одного работника характеризуется следующим уравнением: $\tilde{y}_x = a + b_1 \cdot x_1 + b_2 \cdot x_2$.

На основании исходных данных составляем систему уравнений для определения коэффициентов a , b_1 и b_2 .

$$\sum y = 6 + 11 + 19 + 28 = 64;$$

$$\sum x_1 = 1 + 2 + 3 + 5 = 11; \quad \sum x_2 = 0 + 1 + 3 + 4 = 8;$$

$$\sum y \cdot x_1 = 6 \cdot 1 + 11 \cdot 2 + 19 \cdot 3 + 28 \cdot 5 = 225;$$

$$\sum y \cdot x_2 = 6 \cdot 0 + 11 \cdot 1 + 19 \cdot 3 + 28 \cdot 4 = 180;$$

$$\sum x_1^2 = 1^2 + 2^2 + 3^2 + 5^2 = 39; \quad \sum x_2^2 = 0^2 + 1^2 + 3^2 + 4^2 = 26;$$

$$\sum x_1 \cdot x_2 = 1 \cdot 0 + 2 \cdot 1 + 3 \cdot 3 + 5 \cdot 4 = 31.$$

$$\begin{cases} 64 = 4 \cdot a + 11 \cdot b_1 + 8 \cdot b_2 \\ 225 = 11 \cdot a + 39 \cdot b_1 + 31 \cdot b_2 \\ 180 = 8 \cdot a + 31 \cdot b_1 + 26 \cdot b_2 \end{cases}$$

Решим эту систему по методу Крамера. Вычисляем определитель системы:

$$\Delta = \begin{vmatrix} 4 & 11 & 8 \\ 11 & 39 & 31 \\ 8 & 31 & 26 \end{vmatrix} = 4 \cdot 39 \cdot 26 + 11 \cdot 31 \cdot 8 + 11 \cdot 31 \cdot 8 - 8 \cdot 39 \cdot 8 - 31 \cdot 31 \cdot 4 - 11 \cdot 11 \cdot 26 = 26.$$

Аналогично вычисляем частные определители, заменяя соответствующий столбец столбцом свободных членов:

$$\Delta_1 = \begin{vmatrix} 64 & 11 & 8 \\ 225 & 39 & 31 \\ 180 & 31 & 26 \end{vmatrix} = 62; \quad \Delta_2 = \begin{vmatrix} 4 & 64 & 8 \\ 11 & 225 & 31 \\ 8 & 180 & 26 \end{vmatrix} = 88; \quad \Delta_3 = \begin{vmatrix} 4 & 11 & 64 \\ 11 & 39 & 225 \\ 8 & 31 & 180 \end{vmatrix} = 56.$$

Коэффициенты уравнения определяются по формулам:

$$a = \frac{\Delta_1}{\Delta} = \frac{62}{26} \approx 2,4; \quad b_1 = \frac{\Delta_2}{\Delta} = \frac{88}{26} \approx 3,4; \quad b_2 = \frac{\Delta_3}{\Delta} = \frac{56}{26} \approx 2,2.$$

Таким образом, уравнение имеет вид:

$$\tilde{y}_x = 2,4 + 3,4 \cdot x_1 + 2,2 \cdot x_2.$$

Пример 2. В таблице 10 содержатся данные по пяти предприятиям о прибыли y (млн. руб.), выработке продукции на одного работника x_1 (единиц) и доле продукции, производимой на экспорт x_2 (%). Построить уравнение y по x_1, x_2 .

Таблица 10

| | | | | | | |
|-------|----|----|----|----|----|----|
| x_1 | 11 | 10 | 12 | 18 | 15 | 66 |
| x_2 | 3 | 2 | 4 | 10 | 11 | 30 |
| y | 2 | 1 | 3 | 8 | 7 | 21 |

Уравнение с двумя факторами имеет вид: $\hat{y} = a + b_1x_1 + b_2x_2$.

Воспользуемся для вычислений программой Microsoft Excel (Пакет Анализ данных).

| корреляция | Столбец 1 | Столбец 2 | Столбец 3 | | | |
|---------------------------------|--------------------|---------------------------|---------------------|-------------------|---------------------|--------------------|
| Столбец 1 | 1 | | | | | |
| Столбец 2 | 0,91348012 | 1 | | | | |
| Столбец 3 | 0,97666433 | 0,978600184 | 1 | | | |
| Вывод итогов | | | | | | |
| <i>Регрессионная статистика</i> | | | | | | |
| Множественный R | 0,99950104 | | | | | |
| R-квадрат | 0,99900233 | | | | | |
| Нормированный R-квадрат | 0,99800466 | | | | | |
| Стандартная ошибка | 0,13912167 | | | | | |
| Наблюдения | 5 | | | | | |
| Дисперсионный анализ | | | | | | |
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Значимость F</i> | |
| Регрессия | 2 | 38,76129032 | 19,38064516 | 1001,333333 | 0,000997672 | |
| Остаток | 2 | 0,038709677 | 0,019354839 | | | |
| Итого | 4 | 38,8 | | | | |
| | <i>Коэффициент</i> | <i>Стандартная ошибка</i> | <i>t-статистика</i> | <i>P-Значение</i> | <i>Нижние 95%</i> | <i>Верхние 95%</i> |
| Y-пересечение | -4,4129032 | 0,480504418 | -9,183897305 | 0,011649439 | -6,48034687 | -2,345459582 |
| x1 | 0,47580645 | 0,052264038 | 9,103897648 | 0,011851434 | 0,250932446 | 0,700680457 |
| x2 | 0,38870968 | 0,040867289 | 9,511511307 | 0,01087357 | 0,212871927 | 0,564547428 |

Рисунок 5 - Вывод итогов двухфакторной линейной регрессии.

Естественная двухфакторная регрессия принимает форму

$$\hat{y} = -4,413 + 0,476x_1 + 0,389x_2.$$

Оценку математической погрешности приближения реальных значений y_j представляет средняя ошибка $\bar{A}\%$, определяемая на основании расчетов в таблице 13.

Таблица 13

| n | x_1 | x_2 | y | \hat{y} | $(y - \bar{y})^2$ | $(\hat{y} - \bar{y})^2$ | $(y - \hat{y})^2$ | $\left \frac{y_j - \hat{y}_j}{y_j} \right \cdot 100$ |
|----------|-----------|-----------|-----------|-----------|-------------------|-------------------------|-------------------|--|
| 1 | 11 | 3 | 2 | 1,988 | 4,84 | 4,893 | 0,0001 | 0,60 |
| 2 | 10 | 2 | 1 | 1,124 | 10,24 | 9,462 | 0,0154 | 12,40 |
| 3 | 12 | 4 | 3 | 2,852 | 1,44 | 1,817 | 0,0219 | 4,93 |
| 4 | 18 | 10 | 8 | 8,036 | 14,44 | 14,715 | 0,0013 | 0,45 |
| 5 | 15 | 11 | 7 | 7,000 | 7,84 | 7,874 | 0,0000 | 0,00 |
| Σ | 66 | 30 | 21 | 21 | 38,8 | 38,761 | 0,0387 | 18,38 |

Итак, $\bar{A} = \frac{100}{5} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| = \frac{18,38}{5} \cong 3,68\%$;

Величина приемлемой ошибки колеблется в пределах 5÷10%.

Индекс (линейный коэффициент) множественной корреляции вычисляется по одной из формул:

$$R_{yx_1x_2} = \sqrt{1 - \frac{S_{ocm}}{S}}$$

$$R_{yx_1x_2} = \sqrt{1 - \frac{0,0387}{38,8}} = \sqrt{0,500 \cdot 0,977 + 0,523 \cdot 0,979} \cong 0,9995.$$

Направленность связи результата с факторами прямая.

Коэффициент детерминации $R^2 = (0,9995)^2 = 0,9990$ означает долю объясненной дисперсии (вариации) результата по факторам в общей дисперсии результата.

Доля вариации исследуемого признака y , объясненная уравнением регрессии, составляет 99,9 %.

Доля вариации y , объясняемая неучтенными признаками, составляет 0,1%.

Надежность уравнения в целом по общему критерию Фишера может быть проверена с помощью следующих рассуждений.

Наблюдаемое значение общего критерия Фишера вычисляется по формуле

$$F = \frac{R^2(n - m - 1)}{m \cdot (1 - R^2)}$$

F -критерий принимает значение:

$$F_{набл} = \frac{38,7613}{2 \cdot \frac{0,0387}{5-3}} \cong 1001,333.$$

Критическое значение $F_{крит}$ находят в таблице Фишера-Снедекора.

По заданному уровню значимости α (например, $\alpha=0,05$) гипотезы $H_0 : R^2_{yx_1x_2} = 0$ определяется критическая точка из условия для вероятности

$$P(F_{набл} > F_{кр}(\alpha, n_1, n_2)) = \alpha, \quad F_{кр}(0,05, 2, 2) = 19,00;$$

$F_{набл} = 1001,333 > F_{кр}(0,05, n_1 = 2, n_2 = 2) = 19$. Гипотезу H_0 следует отвергнуть и принять $H_1 : R^2_{yx_1x_2} \neq 0$.

Таким образом, уравнение статистически значимо в целом для опытных данных.

Сравнение наблюдаемых значений $T_{I_{набл}}$ и $T_{кр}$ показывает статистическую значимость полученных оценок при выдвинутой гипотезах

$$H_0 : \hat{\lambda} = 0, \quad \alpha = 0,05, \quad \hat{\lambda} = \hat{b}_1, \quad \hat{\lambda} = \hat{b}_2.$$

Случайные ошибки в оценке b_1, b_2 ,

$$m_{\hat{b}_1} = \frac{0,476}{9,165} = 0,0519, \quad m_{\hat{b}_2} = \frac{0,389}{9,487} = 0,0410.$$

Точечный прогноз неизвестных параметров b_1, b_2 представляется в виде $b_1=0,476 \pm 0,052, b_2=0,389 \pm 0,041$.

Задание

1. По данным 30 наблюдений постройте модель множественной регрессии удовлетворительного качества (табл. 14). Рассчитайте прогноз результата, если прогнозные значения независимых факторов будут составлять 112% от их среднего уровня.

Таблица 14

| № | Валовой продукт | Балансовая стоимость | Объем промышленного | Количество занятых |
|-----|-----------------|----------------------|---------------------|--------------------|
| i | y | x_1 | x_2 | x_3 |
| 1 | 1365 | 3938 | 625 | 161 |
| 2 | 1398 | 3002 | 475 | 186 |
| 3 | 1969 | 6269 | 528 | 267 |
| 4 | 1000 | 2270 | 242 | 129 |

| | | | | |
|----|------|------|-----|-----|
| 5 | 761 | 1879 | 197 | 100 |
| 6 | 1253 | 4709 | 407 | 149 |
| 7 | 1590 | 3976 | 702 | 144 |
| 8 | 1425 | 2946 | 521 | 153 |
| 9 | 1127 | 3151 | 369 | 128 |
| 10 | 1595 | 3424 | 447 | 178 |
| 11 | 1636 | 4110 | 645 | 207 |
| 12 | 2110 | 3910 | 717 | 171 |
| 13 | 1131 | 3045 | 299 | 99 |
| 14 | 2005 | 4575 | 464 | 166 |
| 15 | 768 | 2126 | 169 | 105 |
| 16 | 1682 | 4692 | 579 | 167 |
| 17 | 2146 | 5873 | 468 | 255 |
| 18 | 2865 | 6906 | 850 | 299 |
| 19 | 3133 | 8678 | 924 | 458 |
| 20 | 1706 | 3988 | 507 | 182 |
| 21 | 1456 | 3840 | 475 | 179 |
| 22 | 2616 | 6368 | 801 | 244 |
| 23 | 2657 | 6396 | 604 | 304 |
| 24 | 1538 | 3632 | 592 | 170 |
| 25 | 1249 | 2681 | 220 | 111 |
| 26 | 2960 | 6675 | 672 | 306 |
| 27 | 2255 | 5298 | 712 | 241 |
| 28 | 1423 | 3185 | 214 | 133 |
| 29 | 2488 | 7364 | 602 | 245 |
| 30 | 1426 | 3812 | 225 | 116 |

2. Исследуется влияние некоторых показателей социально-экономического положения субъектов Центрального федерального округа России на региональный коэффициент смертности. В таблице приводятся официальные статистические данные по субъектам Центрального федерального округа, где:

- Y — коэффициент смертности в 2006 году (выражается в промилле «‰» и представляет собой число умерших за год на 1000 человек населения);
- X_1 — индекс (темп роста) промышленного производства, в % к 2004 году;
- X_2 — индекс производства продукции сельского хозяйства, в % к 2004 году (для г. Москвы условно принято 100 %);
- X_3 — численность работников малых предприятий, ‰ (чел. на 1000 чел. населения);
- X_4 — среднемесячная номинальная начисленная заработная плата по региону, тыс. руб.;

- X_5 — численность населения на 1 января 2005 года, тыс. чел.

Требуется:

1. Построить матрицу парных коэффициентов линейной корреляции и выявить коллинеарные факторы.
2. Построить линейную регрессионную модель коэффициента смертности, обосновав отбор факторов. Если из-за коллинеарности факторов невозможно построить уравнение регрессии с полным перечнем факторов, то построить несколько моделей.
3. Оценить качество построенных моделей.
4. Дать экономическую интерпретацию параметров лучшего уравнения регрессии и оценить вклад каждого из факторов в вариацию коэффициента смертности с помощью дельта - коэффициентов.
5. Построить три однофакторные нелинейные регрессионные модели зависимой переменной с наиболее подходящим фактором: степенную, гиперболическую и показательную. Сравнить качество моделей. Выбрать лучшую модель.

Примечание. При проверке статистических гипотез уровень значимости α принять равным 0,05.

Таблица 15

| Область | Y | X ₁ | X ₂ | X ₃ | X ₄ | X ₅ |
|--------------|------|----------------|----------------|----------------|----------------|----------------|
| Белгородская | 16,0 | 108,8 | 115,8 | 35,4 | 6,86 | 1512 |
| Брянская | 19,8 | 116,0 | 95,7 | 25,0 | 5,24 | 1346 |
| Владимирская | 20,3 | 100,2 | 113,3 | 43,1 | 6,07 | 1487 |
| Воронежская | 18,8 | 109,6 | 102,1 | 53,3 | 5,60 | 2334 |
| Ивановская | 22,0 | 107,6 | 96,8 | 36,5 | 5,37 | 1115 |
| Калужская | 19,2 | 105,0 | 94,7 | 58,4 | 6,98 | 1022 |
| Костромская | 21,0 | 108,4 | 100,3 | 30,1 | 5,84 | 717 |
| Курская | 19,7 | 104,0 | 101,1 | 29,8 | 5,65 | 1199 |
| Липецкая | 17,9 | 102,5 | 108,2 | 33,6 | 7,19 | 1190 |
| Московская | 17,5 | 129,6 | 101,2 | 61,5 | 9,51 | 6630 |
| Орловская | 18,5 | 110,3 | 101,7 | 28,4 | 5,46 | 842 |
| Рязанская | 20,3 | 106,2 | 100,9 | 49,4 | 6,22 | 1195 |
| Смоленская | 21,5 | 104,3 | 92,3 | 26,3 | 6,30 | 1019 |
| Тамбовская | 19,3 | 102,5 | 110,0 | 25,6 | 5,08 | 1145 |
| Тверская | 23,1 | 104,4 | 93,0 | 34,5 | 6,64 | 1425 |
| Тульская | 22,0 | 105,0 | 102,7 | 36,4 | 6,34 | 1622 |

| | | | | | | |
|-------------|------|-------|-------|-------|-------|-------|
| Ярославская | 19,9 | 104,5 | 105,9 | 43,3 | 7,39 | 1339 |
| г. Москва | 12,4 | 122,4 | 100,0 | 168,9 | 13,74 | 10407 |

Вопросы для самоконтроля

1. Сформулируйте требования, предъявляемые к факторам для включения их в модель множественной регрессии.
2. К каким трудностям приводит мультиколлинеарность факторов и как они могут быть преодолены?
3. Что означает взаимодействие факторов и как оно может быть выражено графически?
4. При каких условиях строится уравнение множественной регрессии с фиктивными переменными?
5. В чем смысл коэффициента детерминации как он соотносится с коэффициентом множественной корреляции?
6. Отличие множественной регрессии от парной регрессии?
7. Какой метод применяют при оценке параметров множественной регрессии?

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ ПО ТЕМЕ № 5

«Нелинейная регрессия»

Цель занятия: закрепление теоретического и практического материала по теме нелинейная регрессия, приобретение навыков построения и анализа моделей.

Методические указания к проведению занятия

При нелинейной зависимости признаков, приводимой к линейному виду, параметры множественной регрессии также определяются по МНК с той лишь разницей, что он используется не к исходной информации, а к преобразованным данным. Так, рассматривая степенную функцию

$$\tilde{y} = a \cdot x_1^{b_1} \cdot x_2^{b_2} \cdot K \cdot x_p^{b_p},$$

мы преобразовываем ее в линейный вид:

$$\lg \tilde{y} = \lg a + b_1 \cdot \lg x_1 + b_2 \cdot \lg x_2 + K + b_p \cdot \lg x_p,$$

где переменные выражены в логарифмах.

Далее обработка МНК та же: строится система нормальных уравнений и определяются неизвестные параметры. Потенцируя значение $\lg a$, находим параметр a и соответственно общий вид уравнения степенной функции.

Вообще говоря, нелинейная регрессия по включенным переменным не таит каких-либо сложностей в оценке ее параметров. Эта оценка определяется, как и в линейной регрессии, МНК. Так, в двухфакторном уравнении нелинейной регрессии

$$\tilde{y} = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 + a_3 \cdot x_1^2 + a_4 \cdot x_2^2$$

может быть проведена линеаризация, введением в него новых переменных $x_3 = x_1^2$, $x_4 = x_2^2$. В результате получается четырехфакторное уравнение линейной регрессии

$$\tilde{y} = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 + a_3 \cdot x_3 + a_4 \cdot x_4.$$

Пример. Исследуется зависимость потребительских расходов на душу населения y (тыс.руб.) и средней заработной платой (и выплатами социального характера) x (тыс.руб.) по данным за 2005-2006 г.

Таблица 16

| i | Средняя заработная плата x (тыс.руб.) | Расходы на душу населения y (тыс.руб.) |
|-----|--|---|
| 1 | 10,4 | 9,0 |
| 2 | 10,6 | 8,5 |
| 3 | 12,3 | 10,7 |
| 4 | 12,5 | 10,1 |
| 5 | 12,5 | 12,0 |
| 6 | 12,5 | 11,5 |
| 7 | 12,8 | 11,8 |
| 8 | 12,8 | 12,0 |
| 9 | 13,0 | 11,1 |
| 10 | 13,2 | 10,0 |
| 11 | 13,5 | 12,7 |
| 12 | 13,6 | 13,1 |
| 13 | 14,4 | 13,9 |
| 14 | 14,6 | 13,5 |

| | | |
|----|------|------|
| 15 | 14,9 | 13,5 |
| 16 | 14,9 | 14,6 |
| 17 | 15,4 | 14,5 |
| 18 | 15,6 | 15,1 |
| 18 | 15,9 | 15,6 |
| 20 | 15,9 | 15,0 |
| 21 | 16,2 | 15,5 |
| 22 | 16,5 | 15,9 |
| 23 | 16,7 | 15,5 |
| 24 | 17,0 | 16,0 |

Уравнение показательной кривой: $y = b_0 \cdot b_1^x$.

Приведем к линейному виду: $\ln y = \ln(b_0 \cdot b_1^x)$

$$\ln y = \ln b_0 + \ln b_1^x$$

$$\ln y = \ln b_0 + x \cdot \ln b_1$$

Пусть $Y = \ln y$, $A = \ln b_0$, $B = \ln b_1$, тогда уравнение имеет вид:
 $Y = A + B \cdot x$.

Составим расчетную таблицу 17.

Получим:

$$\overline{xY} = \frac{867,41765}{24} \approx 36,142 \quad \bar{Y} = \frac{61,1071163}{24} \approx 2,546 \quad \bar{x} = \frac{337,7}{24} \approx 14,071 \quad \overline{x^2} = \frac{4831,95}{24} \approx 201,331$$

Следовательно:

$$B = \frac{36,142 - 14,071 \cdot 2,546}{201,331 - (14,071)^2} = \frac{0,317234}{3,337959} \approx 0,095$$

$$A = 2,546 - 0,095 \cdot 14,071 \approx 1,209$$

$$\text{Тогда } b_0 = e^{1,209} \approx 3,35, \quad b_1 = e^{0,095} \approx 1,1$$

$$\text{Уравнение примет вид: } \tilde{y} = 3,35 \cdot 1,1^x.$$

Вычислим теоретические значения \tilde{y} . Рассчитаем тесноту связи – индекс корреляции:

$$\rho_{xy} = \sqrt{1 - \frac{\sum (y - \tilde{y})^2}{\sum (y - \bar{y})^2}} = \sqrt{1 - \frac{10,92427539}{119,516256}} = \sqrt{0,908595903} \approx 0,953, \quad \text{связь между}$$

факторами очень сильная.

Индекс детерминации $\rho_{xy}^2 \approx 0,9082$, т.е. 90,82 % вариации изменения y объясняется вариацией изменения x .

$$\text{Средний коэффициент эластичности } \bar{\Theta} = f' \cdot \frac{\bar{x}}{\bar{y}} = \bar{x} \cdot \ln b_1 = 14,071 \cdot \ln 1,1 \approx 1,341$$

При увеличении средней заработной платы на 1% расходы на душу населения увеличатся на 1,341 %.

Оценим статистическую значимость параметров регрессии.

Выдвигаем гипотезу H_0 о статистически незначимом отличии показателей от нуля: $b_0 = b_1 = r_{xy} = 0$.

Таблица 17

| n | y | x | Y | xY | x^2 | \bar{y} | $y - \bar{y}$ | $(y - \bar{y})^2$ | $y - y_{cp}$ | $(y - y_{cp})^2$ | A_i |
|--------------|--------------|--------------|--------------------|------------------|----------------|-------------------|---------------------|--------------------|---------------|-------------------|------------------|
| 1 | 9 | 10,4 | 2,197224577 | 22,8511356 | 108,16 | 9,026694256 | -0,026694256 | 0,000712583 | -3,963 | 15,705369 | 0,2966028 |
| 2 | 8,5 | 10,6 | 2,140066163 | 22,68470133 | 112,36 | 9,200411872 | -0,700411872 | 0,490576791 | -4,463 | 19,918369 | 8,2401396 |
| 3 | 10,7 | 12,3 | 2,370243741 | 29,15399802 | 151,29 | 10,81869393 | -0,118693932 | 0,014088249 | -2,263 | 5,121169 | 1,1092890 |
| 4 | 10,1 | 12,5 | 2,312535424 | 28,9066928 | 156,25 | 11,02689836 | -0,92689836 | 0,85914057 | -2,863 | 8,196769 | 9,1772114 |
| 5 | 12 | 12,5 | 2,48490665 | 31,06133312 | 156,25 | 11,02689836 | 0,97310164 | 0,946926801 | -0,963 | 0,927369 | 8,1091803 |
| 6 | 11,5 | 12,5 | 2,442347035 | 30,52933794 | 156,25 | 11,02689836 | 0,47310164 | 0,223825161 | -1,463 | 2,140369 | 4,1139273 |
| 7 | 11,8 | 12,8 | 2,468099531 | 31,591674 | 163,84 | 11,34674192 | 0,453258079 | 0,205442886 | -1,163 | 1,352569 | 3,8411701 |
| 8 | 12 | 12,8 | 2,48490665 | 31,80680512 | 163,84 | 11,34674192 | 0,653258079 | 0,426746117 | -0,963 | 0,927369 | 5,4438173 |
| 9 | 11,1 | 13 | 2,406945108 | 31,29028641 | 169 | 11,56510857 | -0,465108568 | 0,21632598 | -1,863 | 3,470769 | 4,1901672 |
| 10 | 10 | 13,2 | 2,302585093 | 30,39412323 | 174,24 | 11,78767765 | -1,787677654 | 3,195791395 | -2,963 | 8,779369 | 17,876776 |
| 11 | 12,7 | 13,5 | 2,541601993 | 34,31162691 | 182,25 | 12,1295882 | 0,570411804 | 0,325369626 | -0,263 | 0,069169 | 4,4914315 |
| 12 | 13,1 | 13,6 | 2,57261223 | 34,98752633 | 184,96 | 12,2457482 | 0,854251798 | 0,729746135 | 0,137 | 0,018769 | 6,5210060 |
| 13 | 13,9 | 14,4 | 2,63188884 | 37,8991993 | 207,36 | 13,21598306 | 0,68401694 | 0,467879174 | 0,937 | 0,877969 | 4,9209851 |
| 14 | 13,5 | 14,6 | 2,602689685 | 37,99926941 | 213,16 | 13,47032302 | 0,029676978 | 0,000880723 | 0,537 | 0,288369 | 0,2198294 |
| 15 | 13,5 | 14,9 | 2,602689685 | 38,78007631 | 222,01 | 13,86103997 | -0,361039971 | 0,130349861 | 0,537 | 0,288369 | 2,6743701 |
| 16 | 14,6 | 14,9 | 2,681021529 | 39,94722078 | 222,01 | 13,86103997 | 0,738960029 | 0,546061925 | 1,637 | 2,679769 | 5,0613700 |
| 17 | 14,5 | 15,4 | 2,674148649 | 41,1818892 | 237,16 | 14,53758137 | -0,037581366 | 0,001412359 | 1,537 | 2,362369 | 0,2591818 |
| 18 | 15,1 | 15,6 | 2,714694744 | 42,349238 | 243,36 | 14,81735532 | 0,282644676 | 0,079888013 | 2,137 | 4,566769 | 1,8718190 |
| 19 | 15,6 | 15,9 | 2,747270914 | 43,68160754 | 252,81 | 15,24714397 | 0,352856032 | 0,124507379 | 2,637 | 6,953769 | 2,2618976 |
| 20 | 15 | 15,9 | 2,708050201 | 43,0579982 | 252,81 | 15,24714397 | -0,247143968 | 0,061080141 | 2,037 | 4,149369 | 1,6476264 |
| 21 | 15,5 | 16,2 | 2,740840024 | 44,40160839 | 262,44 | 15,68939896 | -0,189398958 | 0,035871965 | 2,537 | 6,436369 | 1,221928 |
| 22 | 15,9 | 16,5 | 2,766319109 | 45,6442653 | 272,25 | 16,14448189 | -0,244481889 | 0,059771394 | 2,937 | 8,625969 | 1,5376219 |
| 23 | 15,5 | 16,7 | 2,740840024 | 45,7720284 | 278,89 | 16,45518045 | -0,955180449 | 0,912369691 | 2,537 | 6,436369 | 6,1624545 |
| 24 | 16 | 17 | 2,772588722 | 47,13400828 | 289 | 16,93247545 | -0,932475455 | 0,869510474 | 3,037 | 9,223369 | 5,8279715 |
| сумма | 311,1 | 337,7 | 61,10711633 | 867,41765 | 4831,95 | 312,027249 | -0,927249006 | 10,92427539 | -0,012 | 119,516256 | 107,07777 |

Коэффициент Стьюдента для числа степеней свободы $df = n - 2 = 24 - 2 = 22$ и $\alpha = 0,05$ составит 2,07.

Определим случайные ошибки:

$$m_{b_0} = \sqrt{\frac{\sum (y - \tilde{y})^2}{n-2} \cdot \frac{\sum x^2}{n \sum (x - \bar{x})^2}} = \sqrt{\frac{10,92427539}{24-2} \cdot \frac{483,195}{24 \cdot 80,229584}} \approx 1,116$$

$$m_{b_1} = \sqrt{\frac{\sum (y - \tilde{y})^2 / (n-2)}{\sum (x - \bar{x})^2}} = \sqrt{\frac{10,92427539 / 22}{80,229584}} \approx 0,079.$$

$$m_{r_{xy}} = \sqrt{\frac{1 - r_{xy}^2}{n-2}} = \sqrt{\frac{1 - 0,9082}{24-2}} \approx 0,065$$

тогда $t_{b_0} = \frac{b_0}{m_{b_0}} = \frac{3,35}{1,116} \approx 3,002$

$$t_{b_1} = \frac{b_1}{m_{b_1}} = \frac{1,1}{0,079} \approx 13,924$$

$$t_{r_{xy}} = \frac{r_{xy}}{m_{r_{xy}}} = \frac{0,953}{0,065} = 14,662$$

Получим:

$$|t_{b_0}| = 3,002 > t_{табл} = 2,07$$

$$|t_{b_1}| = 13,924 > t_{табл} = 2,07$$

$$|t_r| = 14,662 > t_{табл} = 2,07$$

Т.к. фактические значения t-статистики больше табличного значения, то гипотеза отклоняется, т.е. b_0, b_1, r_{xy} не случайно отличны от нуля и являются статистически значимыми.

Построим доверительные интервалы для коэффициентов регрессии.

$$\gamma_{b_0} = b_0 \pm \Delta_{b_0} = b_0 \pm t_{табл} \cdot m_{b_0} = 3,35 \pm 2,07 \cdot 1,116 = 3,35 \pm 2,31, \text{ т.е. } b_0 \in [1,04; 5,66]$$

$$\gamma_{b_1} = b_1 \pm \Delta_{b_1} = b_1 \pm t_{табл} \cdot m_{b_1} = 1,1 \pm 2,07 \cdot 0,079 = 1,1 \pm 0,164, \text{ т.е. } b_1 \in [0,936; 1,264]$$

Оценим качество модели с помощью F-критерия Фишера.

$$F_{факт} = \frac{\rho_{xy}^2}{1 - \rho_{xy}^2} (n - 2)$$

$$F_{\text{факт}} = \frac{0,9082}{1 - 0,9082} (24 - 2) = 217,651$$

При уровне значимости $\alpha = 0,05$ (вероятность 95 %) табличное значение равно 4,3. $F_{\text{факт}} > F_{\text{табл}}$, следовательно, гипотеза о случайной природе оцениваемых характеристик не принимается, признается их значимость, следовательно модель надежна.

Средняя ошибка аппроксимации составит:

$$\bar{A} = \frac{1}{n} \cdot A_i = \frac{1}{n} \sum \left| \frac{y - \tilde{y}}{y} \right| \cdot 100\%$$

$$\bar{A} = \frac{1}{24} \cdot 107,0777763 \approx 4,462\%$$

Качество модели хорошее, т.к. не превышает значение 8 – 10 %.

Задание

1. По территории Северного, Северо-Западного и Центрального районов известны данные:

Таблица 18

| Район | Потребительские расходы на душу населения, тыс. руб., y | Денежные доходы на душу населения, тыс. руб., x |
|-----------------------|---|---|
| Республика Карелия | 596-N | 913-N |
| Республика Коми | 417+N | 1095-N |
| Архангельская область | 354+N | 606+N |
| Вологодская область | 526+N | 876+N |
| Мурманская область | 934-N | 1314-N |
| Ленинградская область | 412+N | 593+N |
| Новгородская область | 525-N | 754-N |
| Псковская область | 367+N | 528+N |
| Брянская область | 364+N | 520+N |
| Владимирская область | 336+N | 539+N |
| Ивановская область | 409-N | 540+N |
| Калужская область | 452-N | 682+N |
| Костромская область | 367+N | 537+N |
| Московская область | 328+N | 589+N |
| Орловская область | 460-N | 626-N |
| Рязанская область | 380+N | 521+N |
| Смоленская область | 439-N | 626-N |
| Тверская область | 344+N | 521+N |
| Тульская область | 401-N | 658-N |
| Ярославская область | 514-N | 746-N |

Требуется:

1. Рассчитайте параметры уравнений степенной, показательной, гиперболической парной регрессии.

2. Оцените тесноту связи каждого уравнения с помощью показателей корреляции и детерминации.

3. С помощью среднего коэффициента эластичности дайте сравнительную оценку силы связи фактора с результатом (для каждого уравнения).

4. Оцените с помощью средней ошибки аппроксимации качество уравнений.

5. С помощью F-критерия Фишера оцените статистическую надежность результатов регрессионного моделирования.

6. По значениям характеристик, рассчитанных в пп. 2–5 и данном пункте, выберите лучшее уравнение регрессии и дайте его обоснование.

7. Рассчитайте прогнозное значение результата по линейному уравнению регрессии, если прогнозируется увеличение значения фактора на 10% от среднего уровня. Определите доверительный интервал прогноза для уровня значимости $\alpha = 0,05$.

2. По семи территориям Уральского района известны значения двух признаков (табл. 19.).

Таблица 19.

| Район | Расходы на покупку продовольственных товаров в общих расходах, %, у | Среднедневная заработная плата одного работающего, руб., х |
|-------------------|---|--|
| Удмуртская респ. | $68,8 + N/2$ | $45,1 - K/2$ |
| Свердловская обл. | $61,2 + M/2$ | $59,0 - N/2$ |
| Башкортостан | $59,9 + K/2$ | $57,2 - M/2$ |
| Челябинская обл. | $56,7 + N/2$ | $61,8 - K/2$ |
| Пермская обл. | $55,0 + K/2$ | $58,8 - N/2$ |
| Курганская обл. | $54,3 + M/2$ | $47,2 - K/2$ |
| Оренбургская обл. | $49,3 + K/2$ | $55,2 - M/2$ |

Требуется:

1. Для характеристики зависимости y от x рассчитать параметры следующих функций: степенной, показательной, равнобочной гиперболы.
2. Оценить каждую модель через среднюю ошибку аппроксимации \bar{A} и F-критерий Фишера.

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ ПО ТЕМЕ № 6

«Гетероскедастичность и автокорреляция»

Цель занятия: закрепить пройденный материал данной теме, проверить модель на наличие автокорреляции и гетероскедастичности, используя различные методы.

Методические указания к проведению занятия

В результате построения с помощью МНК уравнения регрессии получается не точное значение, а отличающееся от точного на некоторую величину ε :

$$y = \tilde{y}_x + \varepsilon = a + b_1 \cdot x_1 + b_2 \cdot x_2 + \dots + b_p \cdot x_p + \varepsilon.$$

После того как проведена оценка параметров модели, рассчитывая разности фактических и теоретических значений $y - \tilde{y}_x$ можно получить оценки случайной составляющей ε . В задачу регрессионного анализа входит не только построение самой модели, но и исследование остаточных величин.

Необходимость этого объясняется тем, что при использовании МНК предполагалось, что остатки представляют собой независимые случайные величины и их среднее значение равно 0; они имеют одинаковую (постоянную) дисперсию.

Таким образом, исследование остатков предполагают проверку наличия следующих предпосылок МНК

Случайных характер остатков

Для проверки строится график зависимости остатков ε_i от теоретических значений результативного признака. Если на графике получена горизонтальная полоса, то остатки ε_i представляют собой

случайные величины и МНК оправдан, а теоретические значения \tilde{y}_x хорошо аппроксимируют фактические значения y . Пример случайности остатков приведен на рисунке:

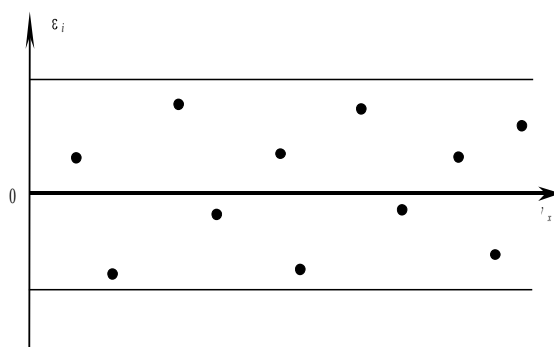


Рисунок 5

Возможны различные случаи зависимости остатков от теоретических значений \tilde{y}_x . Приведем примеры

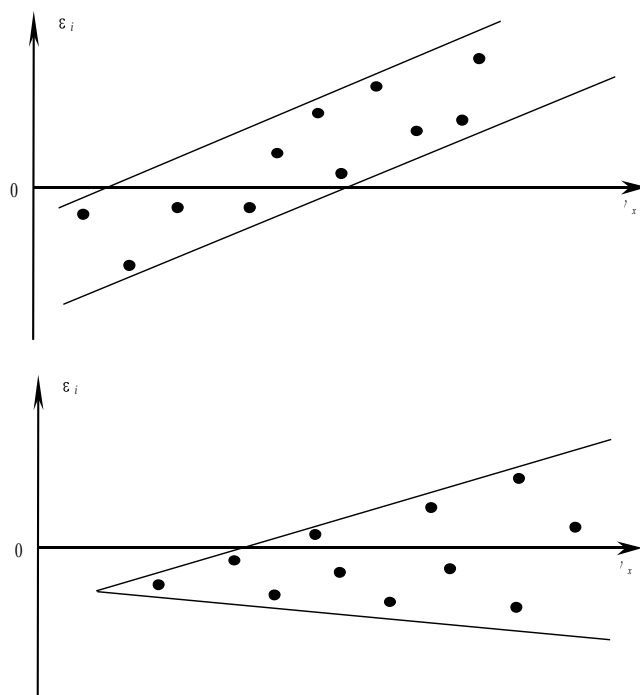


Рисунок 6

Нулевая средняя величина остатков, не зависящая от x_j

Эта предпосылка означает, что $\sum_{k=1}^n (y - \tilde{y}_x) = 0$. Это условие выполнимо для линейных моделей. Для определения независимости величины остатков от x_j , как и в случае определения независимости ε_i от \tilde{y}_x , строится график

ε_i от x_j . Если остатки на графике расположены в виде горизонтальной полосы, то они независимы от значений x_j . Если же зависимость присутствует, то модель является неадекватной.

Гомоскедастичность

Гомоскедастичность остатков означает, что дисперсия каждого отклонения одинакова для всех значений x . Если это условие не соблюдается, то имеет место **гетероскедастичность**. Наличие гетероскедастичности можно наглядно видеть из поля корреляции (смотри рисунок).

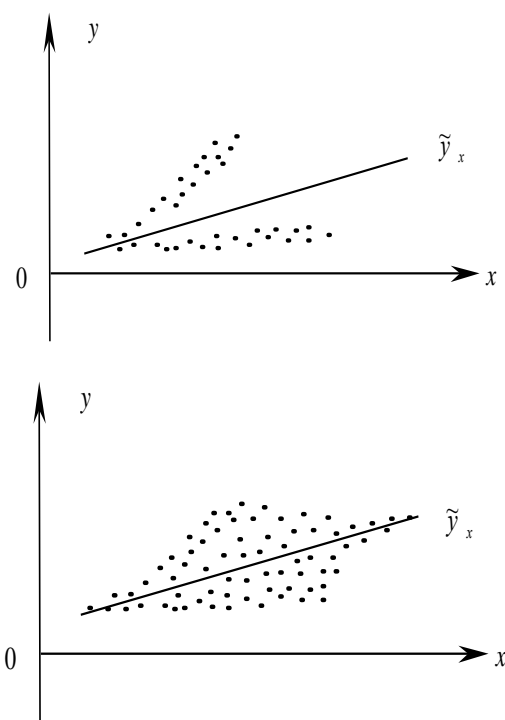


Рисунок 7

Т.к. дисперсия характеризует отклонение то из рисунков видно, что в первом случае дисперсия остатков растет по мере увеличения x , а во втором – дисперсия остатков достигает максимальной величины при средних значениях величины x и уменьшается при минимальных и максимальных значениях x . Наличие гетероскедастичности будет сказываться на уменьшении эффективности оценок параметров уравнения регрессии. Наличие гомоскедастичности или гетероскедастичности можно определять также по графику зависимости остатков от теоретических значений \tilde{y}_x .

Отсутствие автокорреляции остатков

Под **автокорреляцией** остатков понимают зависимость распределения значений остатков ε_i друг от друга. Автокорреляция остатков означает наличие корреляции между остатками текущих и предыдущих (последующих) наблюдений. Оценить эту зависимость можно вычислив коэффициент корреляции между этими остатками по формуле

$$r_{\varepsilon_i \varepsilon_j} = \frac{\overline{\varepsilon_i \varepsilon_j} - \bar{\varepsilon}_i \cdot \bar{\varepsilon}_j}{\sigma_{\varepsilon_i} \sigma_{\varepsilon_j}}.$$

Если этот коэффициент окажется существенно отличным от нуля, то остатки автокоррелированы.

Пример. Проверить для уравнения регрессии, полученного ранее, выполнение предпосылок МНК.

Вычисляем теоретические значения по уравнению регрессии полученному ранее, а остатки по формуле $\varepsilon = y - \tilde{y}_x$ и записываем в таблицу

| Номер предприятия | 1 | 2 | 3 | 4 |
|-----------------------------|------|-------|-------|-------|
| x_1 , (%) | 1 | 2 | 3 | 5 |
| x_2 , (%) | 0 | 1 | 3 | 4 |
| y , (тыс. руб.) | 6 | 11 | 19 | 28 |
| \tilde{y}_x , (тыс. руб.) | 5,79 | 11,31 | 19,07 | 27,87 |
| ε , (тыс. руб.) | 0,21 | -0,31 | -0,07 | 0,13 |

Теперь для проверки случайного характера остатков построим график их зависимости от теоретических значений \tilde{y}_x .

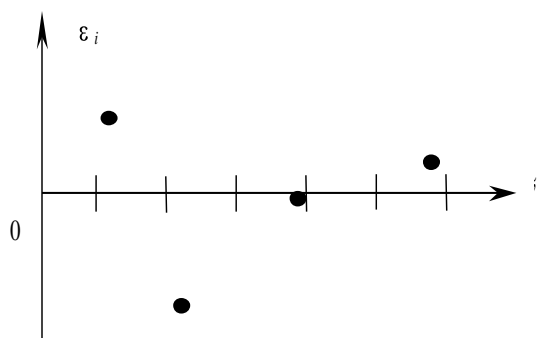


Рисунок 8

Хотя по четырем точкам судить трудно, но в целом можно сделать вывод, что остатки распределены случайно. Из этого же рисунка можно сделать вывод о гомоскедастичности остатков, т. к. дисперсия каждого отклонения одинакова для всех значений x .

Вычислим теперь величину суммарного отклонения:

$$\sum_{k=1}^n (y - \tilde{y}_x) = 0,21 - 0,31 - 0,07 + 0,13 = -0,04.$$

По малости этой величины можно сделать вывод о практически нулевой средней величине остатков.

Коэффициент автокорреляции остатков находим по следующим рядам данных:

| | | | |
|-----------------------------------|-------|-------|-------|
| ε_i , (тыс. руб.) | -0,31 | -0,07 | 0,13 |
| ε_{i-1} , (тыс. руб.) | 0,21 | -0,31 | -0,07 |

$$\bar{\varepsilon}_i = \frac{-0,31 - 0,07 + 0,13}{3} = -0,083;$$

$$\bar{\varepsilon}_{i-1} = \frac{0,21 - 0,31 - 0,07}{3} = -0,057;$$

$$\overline{\varepsilon_i \varepsilon_{i-1}} = \frac{(-0,31) \cdot 0,21 + (-0,07) \cdot (-0,31) + 0,13 \cdot (-0,07)}{3} = -0,018;$$

$$\sigma_{\varepsilon_i} = \sqrt{\frac{\sum_{k=1}^3 (\varepsilon_{ik} - \bar{\varepsilon}_i)^2}{3}} = \sqrt{\frac{(-0,31 + 0,083)^2 + (-0,07 + 0,083)^2 + (0,13 + 0,083)^2}{3}} = 0,18$$

$$\sigma_{\varepsilon_{i-1}} = \sqrt{\frac{\sum_{k=1}^3 (\varepsilon_{i-1k} - \bar{\varepsilon}_{i-1})^2}{3}} = \sqrt{\frac{(0,21 + 0,057)^2 + (-0,31 + 0,057)^2 + (-0,07 + 0,057)^2}{3}} = 0,212$$

Отсюда находим

$$r_{\varepsilon_i \varepsilon_{i-1}} = \frac{\overline{\varepsilon_i \varepsilon_{i-1}} - \bar{\varepsilon}_i \cdot \bar{\varepsilon}_{i-1}}{\sigma_{\varepsilon_i} \sigma_{\varepsilon_{i-1}}} = \frac{-0,018 - (-0,083) \cdot (-0,057)}{0,18 \cdot 0,212} = -0,348$$

Коэффициент корреляции не так велик, и его можно считать приемлемым. Таким образом мы установили, что у нас были все

предпосылки к тому, чтобы применять МНК и линейное уравнение регрессии к исходным данным.

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ ПО ТЕМЕ № 7

«Фиктивные переменные»

Цель занятия: Целью данной работы является ознакомление с методами построения регрессионных моделей, включающих в себя фиктивные переменные.

Методические указания к проведению занятия

До сих пор в качестве факторов рассматривались экономические переменные, принимающие количественные значения в некотором интервале. Вместе с тем может оказаться необходимым включить в модель факторы, которые представляют собой различные атрибутивные признаки. Такими признаками, например, являются профессия, пол, образование, климатические условия и т.п. Чтобы ввести такие переменные в регрессионную модель, им должны быть присвоены те или иные цифровые метки, т.е. качественные переменные преобразовать в количественные. Такого вида сконструированные переменные в эконометрике принято называть *фиктивными переменными*.

Рассмотрим применение фиктивных переменных для функции спроса. Предположим, что по группе лиц мужского и женского пола изучается линейная зависимость потребления кофе от цены. В общем виде для совокупности обследуемых уравнение регрессии имеет вид: $y = a + b \cdot x + \varepsilon$, где y – количество потребляемого кофе; x – цена кофе.

Аналогичные уравнения могут быть найдены отдельно для лиц мужского пола: $y_1 = a_1 + b_1 \cdot x_1 + \varepsilon_1$ и женского пола: $y_2 = a_2 + b_2 \cdot x_2 + \varepsilon_2$. Если сила влияния цены на количество потребления кофе одинакова как для мужчин, так и для женщин ($b_1 \approx b_2 \approx b$), то становится возможным построение общего уравнения регрессии с включением в него фактора «пол»

в виде фиктивной переменной. Это уравнение может быть записано в виде:
 $y = a_1 \cdot z_1 + a_2 \cdot z_2 + b \cdot x + \varepsilon$, где z_1, z_2 - фиктивные переменные, принимающие значения:

$$z_1 = \begin{cases} 1 - \text{мужской пол} \\ 0 - \text{женский пол} \end{cases}; \quad z_2 = \begin{cases} 0 - \text{мужской пол} \\ 1 - \text{женский пол} \end{cases}.$$

Следует отметить, что применение МНК для оценивания параметров a_1 и a_2 приводит к вырожденной матрице исходных данных, а следовательно, и к невозможности получения их оценок.

Выходом из создавшегося положения может явиться переход к уравнению $y = A + A_1 \cdot z_1 + b \cdot x + \varepsilon$, т.е. уравнению, включающему только одну фиктивную переменную. Предположим, что МНК были получены оценки параметров этого уравнения, тогда теоретические значения размера потребления кофе для мужчин будут получены из уравнения $\tilde{y} = A + A_1 + b \cdot x$.

Для женщин соответствующие значения получим из уравнения $\tilde{y} = A + b \cdot x$.

Пример 1. Согласно условию задания 1, дана зависимость объема выпускаемой продукции (шт.) на участке механической обработки от времени (часы), которое включает в себя ремонт, техническое обслуживание оборудования. Исходя из этого, примем объем продукции за результативную переменную, а время простоев оборудования за объясняющую переменную. Выдвинем гипотезу о том, что соответствие сроков выполнения технического обслуживания повлияет на объем выпуска продукции. Тогда введем фиктивную переменную d , которая будет характеризовать качественный признак – соответствие сроков выполнения технического обслуживания. При выполнении сроков обслуживания $d=1$; при невыполнении $d=0$.

Таблица 20

| у (шт) | х (час) простои | d |
|--------|-----------------|---|
| 3520 | 10 | 1 |
| 3460 | 19 | 1 |
| 3100 | 16 | 1 |

| | | |
|------|----|---|
| 3320 | 26 | 1 |
| 3540 | 4 | 1 |
| 3310 | 14 | 1 |
| 3360 | 21 | 1 |
| 3350 | 10 | 1 |
| 3150 | 22 | 1 |
| 3440 | 8 | 1 |
| 3110 | 29 | 1 |
| 3290 | 15 | 1 |
| 3190 | 3 | 0 |
| 3060 | 12 | 0 |
| 3270 | 17 | 0 |
| 3370 | 14 | 0 |
| 3230 | 18 | 0 |
| 3300 | 11 | 0 |
| 3200 | 14 | 0 |
| 3460 | 9 | 0 |

Абстрагируясь от качественного признака во время построения уравнения регрессии, получим следующие данные:

Таблица 21

Результаты моделирования зависимости выпускаемого объема физической продукции от времени простоев по техническим причинам без учета своевременности выполнения технического обслуживания оборудования.

| Регрессионная статистика | | Дисперсионный анализ | | Коэффициенты | | Р-Значение |
|--------------------------|--------|----------------------|--------|---------------|-----------|------------|
| Множественный R | 0,3817 | Значимость F | 0,0968 | Y-пересечение | 3417,6797 | 1,357E-20 |
| | | | | x (час) | -7,958 | 0,04841 |
| R-квадрат | 0,1457 | F | 3,0691 | | | |

Задействуем фиктивную переменную и сравним полученные результаты:

Таблица 22

Результаты моделирования зависимости выпускаемого объема физической продукции от времени простоев по техническим причинам с учетом своевременности выполнения технического обслуживания оборудования.

| Регрессионная статистика | | Дисперсионный анализ | | Коэффициенты | | Р-Значение |
|--------------------------|--------|----------------------|----------|---------------|----------|------------|
| Множественный R | 0,5363 | Значимость F | 0,055995 | Y-пересечение | 3386,984 | 6,3635E-20 |
| | | | | x (час) | -10,366 | 0,0165 |
| Нормированный R-квадрат | 0,2038 | F | 3,4315 | d | 109,767 | 0,0416 |

Сравнивая результаты моделирования до и после использования фиктивной переменной, можно отметить, что во втором случае наблюдается более высокая точность описания регрессионной моделью эмпирических данных, о чем свидетельствует сопоставление коэффициентов детерминации: без использования фиктивной переменной $R^2=14,57\%$, после использования фиктивной переменной скорректированный $R^2=20,38\%$. Что касается адекватности регрессионной модели, то в первом случае при уровне значимости $\alpha_{кр} = 10\%$ модель является адекватной, т.к. $\alpha_{кр} > \alpha_{факт}$ ($0,1 > 0,96$). Во втором случае модель является адекватной уже при $\alpha_{кр} = 6\%$, поскольку $0,060 > 0,055$. Если рассматривать значимость коэффициентов, то и в первом, и во втором случае они окажутся отличными от нуля при $\alpha_{кр} = 5\%$, т.к. $0,05 > 6.36E-20$; $0,05 > 0,0165$; $0,05 > 0,042$. Тем самым, подтверждается значимость и фиктивной переменной, которая отражает качественный признак. Уравнение множественной регрессии с использованием фиктивной переменной будет иметь вид: $Y = 3386,98 - 10,37 \times X + 109,77 \times d$

На основе проведенного анализа можно сделать вывод о том, что своевременное техническое обслуживание оборудования влияет на ход производственного процесса, а именно на объем выпускаемой продукции: при своевременном техническом обслуживании наблюдается рост объема производства на 109,77 шт. Также важно отметить, что фиктивная переменная является переменной сдвига, поскольку при $d=1$ происходит увеличение объема производства на 109,77 шт., что наглядно продемонстрировано на графике подбора.

Пример 2. Задана зависимость доходности портфеля ценных бумаг от доходности рынка в целом, которая отражается формулой: $R_p = A + B \times (R_m - R_f)$, где R_p - доходность портфеля ценных бумаг; R_m - доходность рынка; R_f - доходность базисного актива.

Нам необходимо определить, насколько качественно управляющий портфелем осуществляет свою деятельность. Высокий уровень управления

портфелем характеризуется превышением доходности портфеля над доходностью рынка и изменением стратегии при переходе от положительной доходности рынка к отрицательной. Для этого введем фиктивную переменную d , которая будет отражать качественный признак – качество управления портфелем ценных бумаг. При положительной доходности на рынке $d=0$; при отрицательной доходности на рынке $d=1$. Тогда уравнение зависимости портфеля ценных бумаг от доходности рынка в целом будет иметь вид: $R_p = A + B \times (R_m - R_f) + C \times (R_m - R_f) \times d$, где d – фиктивная переменная.

Таблица 23

| Rp,% | Rm,% | Rf,% | d | Rm-Rf | d*(Rm-Rf) |
|------|------|------|---|-------|-----------|
| 4 | 7 | 4 | 0 | 3 | 0 |
| 16 | 12 | 4 | 0 | 8 | 0 |
| 28 | 23 | 4 | 0 | 19 | 0 |
| 44 | 32 | 4 | 0 | 28 | 0 |
| 27 | 21 | 4 | 0 | 17 | 0 |
| 12 | 9 | 4 | 0 | 5 | 0 |
| -3 | 2 | 4 | 0 | -2 | 0 |
| 10 | -9 | 4 | 1 | -13 | -13 |
| -3 | -23 | 4 | 1 | -27 | -27 |
| 2 | -12 | 4 | 1 | -16 | -16 |
| 1 | -5 | 4 | 1 | -9 | -9 |
| 5 | 0 | 4 | 0 | -4 | 0 |
| 10 | 12 | 4 | 0 | 8 | 0 |

Построим регрессионные модели для первого и второго случая и проведем их сравнительный анализ.

Таблица 24

Результаты моделирования зависимости доходности портфеля ценных бумаг от рынка в целом без учета качества управления портфелем.

| Регрессионная статистика | | Дисперсионный анализ | | Коэффициенты | | Р-Значение |
|--------------------------|--------|----------------------|----------|---------------|--------|------------|
| Множественный R | 0,8598 | Значимость F | 0,000164 | Y-пересечение | 10,753 | 0,0001 |
| | | | | Rm-Rf | 0,7773 | |
| R-квадрат | 0,7392 | F | 31,18204 | | | |

Результаты моделирования зависимости доходности портфеля ценных бумаг от рынка в целом с учетом качества управления портфелем.

| Регрессионная статистика | | Дисперсионный анализ | | Коэффициенты | | Р-Значение |
|--------------------------|---------|----------------------|----------|---------------|---------|------------|
| Множественный R | 0,95944 | Значимость F | 3,17E-06 | Y-пересечение | 4,05402 | 0,0262 |
| | | | | Rm-Rf | 1,33451 | 1E-06 |
| Нормированный R-квадрат | 0,90462 | F | 57,9094 | d*(Rm-Rf) | -1,194 | 0,0004 |

При использовании фиктивной переменной наблюдается более высокая точность описания регрессионной моделью эмпирических данных, чем без ее использования. Так скорректированный R^2 при значении 90,46% говорит о том, что регрессионная модель, отражающая качество управления портфелем ценных бумаг, описывает эмпирические данные более точно, чем это делает модель, в которой не придается значения качеству управления портфелем ($R^2=73,92\%$). Обе модели оказались адекватными при уровне значимости $\alpha_{кр} = 3\%$, поскольку в первом случае $0,03 > 0,000164$; во втором случае $0,03 > 3,17E-06$. Все коэффициенты в обоих случаях при уровне значимости $\alpha_{кр} = 3\%$ оказались значимыми (отличными от нуля), так как $0,03 > 0,0001$; $0,03 > 8E-05$ и $0,03 > 0,026$; $0,03 > 1E-06$; $0,03 > 0,0004$, что говорит о том, что качество управления портфелем влияет на конечную доходность портфеля ценных бумаг и деятельность управляющего может быть оценена положительно. Уравнение регрессии с использованием фиктивной переменной будет иметь вид: $Y = 4,054 + 1,3345 \times (R_m - R_f) - 1,194 \times (R_m - R_f) \times d$

Задание

1. Необходимо оценить влияние качественного признака d на результативную переменную Y . Для этого сравним регрессионную модель, не учитывающую качественный признак с регрессионной моделью, учитывающей качественный признак.

Таблица 26

| x | D (фикт) | y |
|----|----------|----------|
| 1 | 0 | 6,191552 |
| 2 | 0 | 8,235765 |
| 3 | 0 | 11,48288 |
| 4 | 0 | 14,40931 |
| 5 | 0 | 16,97195 |
| 6 | 0 | 19,79557 |
| 7 | 0 | 22,6074 |
| 8 | 0 | 24,76633 |
| 9 | 0 | 26,98018 |
| 10 | 0 | 28,76968 |
| 11 | 0 | 30,52625 |
| 12 | 1 | 34,92994 |
| 13 | 1 | 36,18605 |
| 14 | 1 | 37,3445 |
| 15 | 1 | 38,5371 |
| 16 | 1 | 39,4185 |
| 17 | 1 | 40,52482 |
| 18 | 1 | 41,17279 |
| 19 | 1 | 41,96986 |
| 20 | 1 | 42,73582 |

2. Необходимо исследовать зависимость между результатами ЕГЭ и курсового экзамена по математике. Получены данные о числе решенных задач (задание – 18 задач) на ЕГЭ и на курсовом экзамене 16 студентами, а также распределение этих студентов по категории «пол»: y_i – число решенных задач на курсовом экзамене; x_i – число решенных задач на ЕГЭ:

Таблица 27

| № | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|-------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| x_i | 18 | 15 | 9 | 18 | 12 | 12 | 18 | 6 | 18 | 12 | 16 | 12 | 18 | 6 | 17 | 15 |
| y_i | 16 | 15 | 3 | 18 | 6 | 6 | 16 | 4 | 18 | 6 | 16 | 6 | 17 | 2 | 15 | 9 |
| пол | М. | М. | Ж. | М. | Ж. | Ж. | М. | Ж. | М. | Ж. | М. | Ж. | М. | Ж. | М. | Ж. |

1. Составить модель множественной регрессии изменения зависимой переменной y_i от переменной x_i при учете пола студента с помощью фиктивной переменной, взяв женский пол в качестве эталонной категории.

2. Построить выборочное уравнение регрессии и с доверительной вероятностью 0,95 оценить значимость коэффициентов уравнения регрессии и самой регрессии.

3. Из найденного выборочного уравнения регрессии вывести отдельные уравнения для юношей и девушек и дать их графическую иллюстрацию.

4. С помощью критерия Чоу для уровня значимости $\alpha=0,05$ проверить, являются ли выборки для юношей и девушек однородными в регрессионном смысле и можно ли их объединить.

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ ПО ТЕМЕ № 8

«Моделирование одномерных рядов»

Цель занятия: научиться определять тип тенденции и выявлять наличие периодической составляющей во временном ряду, а также оценивать уровень сезонности, осуществлять фильтрацию периодических составляющих временного ряда и их моделирование. Получить навыки идентификации и применения различных моделей стационарных временных рядов для прогнозирования экономических процессов в профессиональной деятельности.

Методические указания к проведению занятия

Обычно эконометрические модели строятся на основе двух типов исходных данных:

- данные, характеризующие совокупность различных объектов в определенном момент (период) времени;
- данные, характеризующие один объект за ряд последовательных моментов (периодов) времени.

Модели, построенные по данным первого типа, называются **пространственными моделями**. Модели, построенные на основе второго типа данных, называются **моделями временных рядов**.

Временной ряд – совокупность значений какого-либо показателя за несколько последовательных моментов или периодов времени. Каждый уровень временного ряда формируется под воздействием большого числа факторов, которые условно можно подразделить на три группы:

- факторы, формирующие тенденцию ряда (например, инфляция влияет на увеличение размера средней заработной платы);
- факторы, формирующие циклические колебания ряда (например, уровень безработицы в курортных городах в зимний период выше по сравнению с летним);
- случайные факторы.

Очевидно, что реальные данные чаще всего содержат все три компоненты. Модель, в которой временной ряд представлен как сумма перечисленных компонент, называется **аддитивной моделью** временного ряда. Если же временной ряд представлен как их произведение, то такая модель называется **мультипликативной**.

При наличии в временном ряде тенденции и циклических колебаний значения каждого последующего уровня ряда зависят от предыдущих. Корреляционную зависимость между последовательными уровнями временного ряда называют уровнями автокорреляции уровней ряда. Количественно эту зависимость с помощью коэффициента корреляции между уровнями исходного временного ряда и уровнями этого ряда, сдвинутого на несколько шагов во времени.

Пример 1. Пусть имеются условные данные о средних расходах на конечное потребление (y_t , денежных единиц) за 8 лет.

Таблица 28

| t | y_t | y_{t-1} | $y_t - \bar{y}_1$ | $y_{t-1} - \bar{y}_2$ | $(y_t - \bar{y}_1) \cdot (y_{t-1} - \bar{y}_2)$ | $(y_t - \bar{y}_1)^2$ | $(y_{t-1} - \bar{y}_2)^2$ |
|----------|-------|-----------|-------------------|-----------------------|---|-----------------------|---------------------------|
| 1 | 7 | - | - | - | - | - | - |
| 2 | 8 | 7 | -3,39 | -3 | 9,87 | 10,8241 | 9 |
| 3 | 8 | 8 | -3,29 | -2 | 6,58 | 10,8241 | 4 |
| 4 | 10 | 8 | -1,29 | -2 | 2,58 | 1,6641 | 4 |
| 5 | 11 | 10 | -0,29 | 0 | 0,00 | 0,0841 | 0 |
| 6 | 12 | 11 | 0,71 | 1 | 0,71 | 0,5041 | 1 |
| 7 | 14 | 12 | 2,71 | 2 | 5,42 | 7,3441 | 4 |
| 8 | 16 | 14 | 4,71 | 4 | 18,84 | 22,1841 | 16 |
| Σ | 86 | 70 | -0,03 | 0 | 44,0 | 53,4287 | 38 |

По формулам

$$\bar{y}_1 = \frac{\sum_{t=2}^n y_t}{n-1}; \quad \bar{y}_2 = \frac{\sum_{t=2}^n y_{t-1}}{n-1}$$

ВЫЧИСЛЯЕМ

$$\bar{y}_1 = \frac{8+8+10+11+12+14+16}{7} = \frac{79}{7} = 11,29,$$

$$\bar{y}_2 = \frac{7+8+8+10+11+12+14}{7} = \frac{70}{7} = 10.$$

Далее, заполняем таблицу и используя формулу для вычисления линейного коэффициента корреляции, получаем

$$r_1 = \frac{\sum_{t=2}^n (y_t - \bar{y}_1) \cdot (y_{t-1} - \bar{y}_2)}{\sqrt{\sum_{t=2}^n (y_t - \bar{y}_1)^2 \cdot (y_{t-1} - \bar{y}_2)^2}} = \frac{44}{\sqrt{53,4287 \cdot 38}} = 0,976.$$

Полученное значение свидетельствует об очень тесной зависимости между расходами на конечное потребление текущего непосредственно предшествующего годов и, следовательно, о наличии во временном ряде расходов на конечное потребление сильной линейной тенденции.

Нами был посчитан коэффициент автокорреляции для смещения на один год. Такой коэффициент называется коэффициентом первого порядка. При смещении на два года получим коэффициент второго порядка и так далее. Число периодов (в данном случае лет), по которым рассчитывается коэффициент автокорреляции, называется *лагом*.

Одним из наиболее распространенных способов моделирования тенденции временного ряда является построение аналитической функции, характеризующей зависимость уровней ряда от времени. Поскольку зависимость может принимать различные формы, то ее формализации можно использовать различные виды функций: линейную, гиперболическую, параболическую, степенную и т.п. Параметры каждой из перечисленных моделей могут быть найдены по МНК.

Пример 2. Известны статистические данные наблюдений за некоторое количество моментов или периодов времени.

Требуется:

1. Построить график динамики уровней ряда.
2. Рассчитать значения сезонных компонент методом скользящей средней.
3. Устранить сезонную компоненту из исходных уровней ряда. Построить уравнение, моделирующее динамику трендовой компоненты.
4. Найти прогноз фактора y .

Таблица 29

| Объем продаж | Номер квартала |
|--------------|----------------|
| 3974 | 1 |
| 4134 | 2 |
| 4054 | 3 |
| 4200 | 4 |
| 3995 | 5 |
| 4162 | 6 |
| 4091 | 7 |
| 4224 | 8 |
| 4038 | 9 |
| 4191 | 10 |
| 4125 | 11 |
| 4259 | 12 |
| 4079 | 13 |
| 4240 | 14 |
| 4181 | 15 |
| 4320 | 16 |
| 4126 | 17 |
| 4285 | 18 |
| 4213 | 19 |
| 4359 | 20 |
| 4143 | 21 |
| 4311 | 22 |
| 4240 | 23 |
| 4390 | 24 |

Построим график динамики уровней ряда.

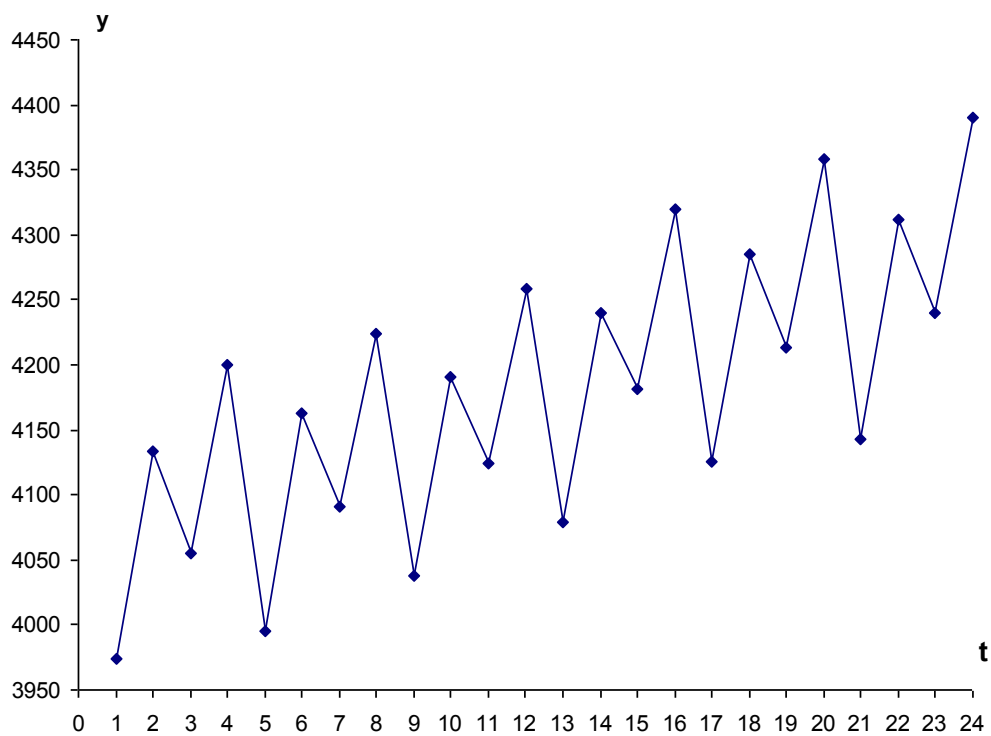


Рисунок 9 - График динамики

Изучая график, заметим наличие сезонных колебаний с периодом 4 квартала и возрастающую тенденцию в уровнях ряда.

Построим мультипликативную модель временного ряда $Y = T \cdot S \cdot E$.

Рассчитаем компоненты.

1) Проведем выравнивание исходных уровней ряда, которое полностью совпадает выравниванием для построения мультипликативной модели (табл. 30).

Таблица 30

Расчет оценок сезонной компоненты в мультипликативной модели

| Номер квартала, t | Объем продаж | Итого за четыре квартала | Скользящая средняя | Центрированная скользящая средняя | Оценка сезонной компоненты |
|-------------------|--------------|--------------------------|--------------------|-----------------------------------|----------------------------|
| 1 | 3974 | | | | |
| 2 | 4134 | 16362 | 4090,5 | | |
| 3 | 4054 | 16383 | 4095,75 | 4093,125 | 0,990441289 |
| 4 | 4200 | 16411 | 4102,75 | 4099,25 | 1,024577667 |
| 5 | 3995 | 16448 | 4112 | 4107,375 | 0,972640677 |
| 6 | 4162 | 16472 | 4118 | 4115 | 1,011421628 |
| 7 | 4091 | 16515 | 4128,75 | 4123,375 | 0,992148422 |
| 8 | 4224 | 16544 | 4136 | 4132,375 | 1,02217248 |
| 9 | 4038 | 16578 | 4144,5 | 4140,25 | 0,975303424 |
| 10 | 4191 | 16613 | 4153,25 | 4148,875 | 1,010153355 |
| 11 | 4125 | 16654 | 4163,5 | 4158,375 | 0,991974028 |
| 12 | 4259 | 16703 | 4175,75 | 4169,625 | 1,021434781 |
| 13 | 4079 | 16759 | 4189,75 | 4182,75 | 0,975195744 |
| 14 | 4240 | 16820 | 4205 | 4197,375 | 1,010155156 |
| 15 | 4181 | 16867 | 4216,75 | 4210,875 | 0,992905275 |
| 16 | 4320 | 16912 | 4228 | 4222,375 | 1,023120874 |
| 17 | 4126 | 16944 | 4236 | 4232 | 0,974952741 |
| 18 | 4285 | 16983 | 4245,75 | 4240,875 | 1,010404692 |
| 19 | 4213 | 17000 | 4250 | 4247,875 | 0,991790013 |
| 20 | 4359 | 17026 | 4256,5 | 4253,25 | 1,02486334 |
| 21 | 4143 | 17053 | 4263,25 | 4259,875 | 0,972563749 |
| 22 | 4311 | 17084 | 4271 | 4267,125 | 1,010282099 |
| 23 | 4240 | | | | |
| 24 | 4390 | | | | |
| Итого | 100334 | 351131 | 87782,75 | 83602 | 19,99850143 |

2) Найдем оценки сезонной компоненты как частное от деления фактических уровней ряда на центрированные скользящие средние.

Используем эти оценки для расчета значений сезонной компоненты S (табл. 31).

Таблица 31

Расчет сезонной компоненты

| год | 1 квартал | 2 квартал | 3 квартал | 4 квартал | | |
|-------------------|-----------|-------------|-------------|-------------|----------------------------|-------------|
| 1 | 0 | 0 | 0,990441289 | 1,024577667 | | |
| 2 | 0,9726407 | 1,011421628 | 0,992148422 | 1,02217248 | | |
| 3 | 0,9753034 | 1,010153355 | 0,991974028 | 1,021434781 | | |
| 4 | 0,9751957 | 1,010155156 | 0,992905275 | 1,023120874 | | |
| 5 | 0,9749527 | 1,010404692 | 0,991790013 | 1,02486334 | | |
| 6 | 0,9725637 | 1,010282099 | 0 | 0 | | |
| итого | 4,8706563 | 5,052416931 | 4,959259027 | 5,116169141 | | |
| Средняя оценка | 0,8117761 | 0,842069488 | 0,826543171 | 0,852694857 | Сумма | 3,333083572 |
| Скорректированная | 0,9742043 | 1,010559106 | 0,991926129 | 1,023310503 | Корректирующий коэффициент | 1,200089921 |

Для данной модели имеем:

$$0,8117761 + 0,84206949 + 0,826543171 + 0,852694857 = 3,333083572$$

Определим корректирующий коэффициент:

$$k = \frac{4}{3,333083572} = 1,200089921$$

Рассчитаем скорректированные значения сезонной компоненты:

$$S_1 = 0,9742043$$

$$S_2 = 1,01055911$$

$$S_3 = 0,991926129$$

$$S_4 = 1,023310503$$

3) Разделим каждый уровень исходного ряда на соответствующие значения сезонной компоненты. Получим величины $T \cdot E = Y : S$.

4) Определим компоненту T данной модели. Для этого проведем аналитическое выравнивание ряда $T \cdot E$ с помощью линейного тренда.

Получим линейный тренд: $T = 4061,189 + 9,534 \cdot t$.

Подставляя в это уравнение значения $t = 1, 2, \dots, 24$ найдем уровни T для каждого момента времени.

5) Найдем уровни ряда по мультипликативной модели, умножив уровни T на значения сезонной компоненты для соответствующих кварталов, т.е. ряд $T \cdot S$.

б) Рассчитаем ошибку: $E = Y / (T \cdot S)$. Абсолютные ошибки определяются как $E' = Y - (T \cdot S)$. Доля объясненной дисперсии уровней ряда равна 99,53 %.

| Вывод итогов | | | | | |
|---------------------------------------|-------------------|--------------------|---------------------|----------------------|-------------------------------------|
| <i>Регрессионная статистика</i> | | | | | |
| Множественный R | 0,993727 | | | | |
| R-квадрат | 0,987493 | | | | |
| Нормированный R-квадрат | 0,986925 | | | | |
| Стандартная ошибка | 7,761887 | | | | |
| Наблюдения | 24 | | | | |
| <i>Дисперсионный анализ</i> | | | | | |
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>значимость F</i> |
| Регрессия | 1 | 104651,7 | 104651,7 | 1737,048 | 1,98E-22 |
| Остаток | 22 | 1325,431 | 60,24688 | | |
| Итого | 23 | 105977,1 | | | |
| <i>Коэффициент стандартной ошибки</i> | | | | | |
| <i>Р-Значение</i> | <i>нижние 95%</i> | <i>верхние 95%</i> | <i>нижние 95,0%</i> | <i>верхние 95,0%</i> | |
| Y-пересечение | 4061,189 | 3,270475 | 1241,774 | 8,38E-55 | 4054,407 4067,972 4054,407 4067,972 |
| t | 9,53947 | 0,228886 | 41,6779 | 1,98E-22 | 9,06479 10,01415 9,06479 10,01415 |

Рисунок 10 - Протокол выполнения регрессионного анализа

Построим аддитивную модель временного ряда $Y = T + S + E$.

Рассчитаем компоненты.

1) Проведем выравнивание исходных уровней ряда методом скользящей средней, найдем оценки сезонной компоненты как разность между фактическими уровнями ряда и центрированными скользящими средними (табл. 32).

Таблица 32

Расчет оценок сезонной компоненты в аддитивной модели

| Номер квартала, t | Объем продаж | Итого за четыре квартала | Скользящая средняя | Центрированная скользящая средняя | Оценка сезонной компоненты |
|-------------------|--------------|--------------------------|--------------------|-----------------------------------|----------------------------|
| 1 | 3974 | | | | |
| 2 | 4134 | 16362 | 4090,5 | | |
| 3 | 4054 | 16383 | 4095,75 | 4093,125 | -39,125 |
| 4 | 4200 | 16411 | 4102,75 | 4099,25 | 100,75 |

| | | | | | |
|-------|--------|--------|----------|----------|----------|
| 5 | 3995 | 16448 | 4112 | 4107,375 | -112,375 |
| 6 | 4162 | 16472 | 4118 | 4115 | 47 |
| 7 | 4091 | 16515 | 4128,75 | 4123,375 | -32,375 |
| 8 | 4224 | 16544 | 4136 | 4132,375 | 91,625 |
| 9 | 4038 | 16578 | 4144,5 | 4140,25 | -102,25 |
| 10 | 4191 | 16613 | 4153,25 | 4148,875 | 42,125 |
| 11 | 4125 | 16654 | 4163,5 | 4158,375 | -33,375 |
| 12 | 4259 | 16703 | 4175,75 | 4169,625 | 89,375 |
| 13 | 4079 | 16759 | 4189,75 | 4182,75 | -103,75 |
| 14 | 4240 | 16820 | 4205 | 4197,375 | 42,625 |
| 15 | 4181 | 16867 | 4216,75 | 4210,875 | -29,875 |
| 16 | 4320 | 16912 | 4228 | 4222,375 | 97,625 |
| 17 | 4126 | 16944 | 4236 | 4232 | -106 |
| 18 | 4285 | 16983 | 4245,75 | 4240,875 | 44,125 |
| 19 | 4213 | 17000 | 4250 | 4247,875 | -34,875 |
| 20 | 4359 | 17026 | 4256,5 | 4253,25 | 105,75 |
| 21 | 4143 | 17053 | 4263,25 | 4259,875 | -116,875 |
| 22 | 4311 | 17084 | 4271 | 4267,125 | 43,875 |
| 23 | 4240 | | | | |
| 24 | 4390 | | | | |
| Итого | 100334 | 351131 | 87782,75 | 83602 | -6 |

2) Используем эти оценки для расчета значений сезонной компоненты S (табл. 33).

Таблица 33

Расчет сезонной компоненты

| год | 1 квартал | 2 квартал | 3 квартал | 4 квартал | | |
|-------------------|------------|-----------|--------------|-------------|----------------------------|-------|
| 1 | 0 | 0 | -39,125 | 100,75 | | |
| 2 | -112,375 | 47 | -32,375 | 91,625 | | |
| 3 | -102,25 | 42,125 | -33,375 | 89,375 | | |
| 4 | -103,75 | 42,625 | -29,875 | 97,625 | | |
| 5 | -106 | 44,125 | -34,875 | 105,75 | | |
| 6 | -116,875 | 43,875 | 0 | 0 | | |
| итого | -541,25 | 219,75 | -169,625 | 485,125 | | |
| Средняя оценка | -90,208333 | 36,625 | -28,27083333 | 80,85416667 | Сумма | -1 |
| Скорректированная | -89,958333 | 36,875 | -28,02083333 | 81,10416667 | Корректирующий коэффициент | -0,25 |

Для данной модели имеем:

$$-90,208333+36,625-28,27083333+80,85416667=-1$$

Определим корректирующий коэффициент:

$$k = \frac{-1}{4} = -0,25$$

Рассчитаем скорректированные значения сезонной компоненты:

$$S_1 = -89,958333$$

$$S_2 = 36,875$$

$$S_3 = -28,02083333$$

$$S_4 = 81,10416667$$

3) Элиминируем влияние сезонной компоненты, вычитая ее значение из каждого уровня исходного временного ряда. Получим величины $T + E = Y - S$.

4) Определим компоненту T данной модели. Для этого проведем аналитическое выравнивание ряда $T + E$ с помощью линейного тренда.

Получим линейный тренд: $T = 4058,397 + 9,775 \cdot t$.

Подставляя в это уравнение значения $t = 1, 2, \dots, 24$ найдем уровни T для каждого момента времени.

5) Найдем значения уровней ряда, полученные по аддитивной модели. Для этого прибавим к уровням T значения сезонной компоненты для соответствующих кварталов, т.е. ряд $T + S$.

6) Рассчитаем ошибку: $E = Y - (T + S)$. Эта абсолютная ошибка составляет ≈ 0 . Доля объясненной дисперсии уровней ряда равна 98,13 %.

| Вывод итогов | | | | | | | | | |
|--|-----------|-----------|-----------|----------|---------------------|----------|----------|----------|--|
| <i>Регрессионная статистика</i> | | | | | | | | | |
| Множественный R | 0,976879 | | | | | | | | |
| R-квадрат | 0,954293 | | | | | | | | |
| Нормированный R-квадрат | 0,952216 | | | | | | | | |
| Стандартная ошибка | 15,4667 | | | | | | | | |
| Наблюдения | 24 | | | | | | | | |
| <i>Дисперсионный анализ</i> | | | | | | | | | |
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>значимость F</i> | | | | |
| Регрессия | 1 | 109880,8 | 109880,8 | 459,3317 | 3,13E-16 | | | | |
| Остаток | 22 | 5262,814 | 239,2188 | | | | | | |
| Итого | 23 | 115143,6 | | | | | | | |
| <i>Коэффициент стандартная статистика P-Значение нижние 95% верхние 95% нижние 95,0% верхние 95,0%</i> | | | | | | | | | |
| Y-пересечение | 4058,397 | 6,516901 | 622,7495 | 3,29E-48 | 4044,882 | 4071,912 | 4044,882 | 4071,912 | |
| t | 9,774891 | 0,456088 | 21,43202 | 3,13E-16 | 8,829023 | 10,72076 | 8,829023 | 10,72076 | |

Рисунок 11 - Протокол выполнения регрессионного анализа

ВЫВОД: Лучше всего описывает процесс мультипликативная модель, т.к. доля объясненной дисперсии больше, чем в аддитивной модели.

Задание

1. Имеются поквартальные данные по розничному товарообороту (в % к предыдущему периоду). Постройте мультипликативную модель временного ряда. Рассчитайте прогноз розничного товарооборота на 1 квартал 2016 года.

Таблица 34

| № квартала | t | y_t |
|---------------|-----|-------|
| 1 кв. 2011 г. | 1 | 113,1 |
| 2 кв. 2011 г. | 2 | 95,9 |
| 3 кв. 2011 г. | 3 | 98 |
| 4 кв. 2011 г. | 4 | 101,8 |
| 1 кв. 2012 г. | 5 | 107,8 |
| 2 кв. 2012 г. | 6 | 96,3 |
| 3 кв. 2012 г. | 7 | 95,7 |
| 4 кв. 2012 г. | 8 | 99,8 |
| 1 кв. 2013 г. | 9 | 104 |
| 2 кв. 2013 г. | 10 | 95,8 |
| 3 кв. 2013 г. | 11 | 95,5 |
| 4 кв. 2013 г. | 12 | 99,3 |
| 1 кв. 2014 г. | 13 | 104 |
| 2 кв. 2014 г. | 14 | 96,2 |
| 3 кв. 2014 г. | 15 | 95,1 |
| 4 кв. 2014 г. | 16 | 97,5 |
| 1 кв. 2015 г. | 17 | 101 |
| 2 кв. 2015 г. | 18 | 93,5 |
| 3 кв. 2015 г. | 19 | 92 |
| 4 кв. 2015 г. | 20 | 94,6 |

2. Имеются поквартальные данные о разрешениях на строительство нового частного жилья, выданных в США в 1990 - 1994 гг в % к уровню 1987 года. Постройте аддитивную модель временного ряда. Рассчитайте прогноз на 1 квартал 1995 года.

Таблица 35

| № квартала | t | y_t |
|---------------|-----|-------|
| 1 кв. 1990 г. | 1 | 68,3 |
| 2 кв. 1990 г. | 2 | 61,9 |
| 3 кв. 1990 г. | 3 | 65,1 |

| | | |
|---------------|----|------|
| 4 кв.1990 г. | 4 | 74,1 |
| 1 кв. 1991 г. | 5 | 67,5 |
| 2 кв. 1991 г. | 6 | 66,8 |
| 3 кв. 1991 г. | 7 | 65,5 |
| 4 кв. 1991 г. | 8 | 73,6 |
| 1 кв. 1992 г. | 9 | 73,7 |
| 2 кв. 1992 г. | 10 | 61,0 |
| 3 кв. 1992 г. | 11 | 71,4 |
| 4 кв. 1992 г. | 12 | 82,3 |
| 1 кв. 1993 г. | 13 | 79,0 |
| 2 кв. 1993 г. | 14 | 72,7 |
| 3 кв. 1993 г. | 15 | 73,9 |
| 4 кв. 1993 г. | 16 | 85,3 |
| 1 кв. 1994 г. | 17 | 83,4 |
| 2 кв. 1994 г. | 18 | 76,2 |
| 3 кв. 1994 г. | 19 | 81,6 |
| 4 кв. 1994 г. | 20 | 89,3 |

Вопросы для самоконтроля

1. Дайте понятие временного ряда. Перечислите его основные характеристики.
2. В чем экономическое содержание составляющих временного ряда?
3. Как выделяется случайная составляющая временного ряда?
4. Для чего строится математическая модель временного ряда?
5. Что такое автокорреляция уровней временного ряда и как ее можно оценить количественно?
6. Перечислите основные виды трендов.
7. Выпишите общий вид аддитивной и мультипликативной моделей временного ряда.
8. Перечислите этапы построения модели временного ряда.
9. С какими целями проводится выявление и устранение сезонного эффекта?

Савельева Екатерина Владимировна

Моделирование и статистическая обработка результатов научных исследований: методические указания к практическим занятиям и самостоятельной работе для обучающихся по направлениям подготовки 35.06.01 Сельское хозяйство; 35.06.02 Лесное хозяйство; 35.06.04 Технологии, средства механизации и энергетическое оборудование в сельском, лесном и рыбном хозяйстве; 36.06.01 Ветеринария и зоотехния; 38.06.01 Экономика

ЭЛЕКТРОННОЕ ИЗДАНИЕ

ФГБОУ ВПО «Приморская ГСХА»
Адрес: 692510, г. Уссурийск, пр-т Блюхера, 44