Э. И. ДЯМИНОВА, А.С. ФИЛИППОВА

ЛАБОРАТОРНЫЙ ПРАКТИКУМ «СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ С ПОМОЩЬЮ ПРОГРАММНЫХ СРЕДСТВ»

Уфа 2019

Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Башкирский государственный педагогический университет им. М.Акмуллы»

Э. И. ДЯМИНОВА, А. С. ФИЛИППОВА

ЛАБОРАТОРНЫЙ ПРАКТИКУМ «СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ С ПОМОЩЬЮ ПРОГРАММНЫХ СРЕДСТВ»

Методические указания для студентов очной и заочной форм обучения, обучающихся по специальностям/направлениям подготовки УГСН 09.00.00 «Информатика и вычислительная техника»

Уфа 2019

Дяминова Э. И., Филиппова А.С.

Лабораторный практикум «Статистический анализ данных с помощью программных средств» [Электронный ресурс] / Баш. гос. пед. ун-т им. М.Акмуллы. – Уфа: БГПУ им. М.Акмуллы, 2019. – 112 с.

Лабораторный практикум содержит теоретический материал, задания на лабраторные работы, варианты и подробную методику выполнения с использованием программных средств (на примере MS Excel и STATISTICA).

Методические указания предназначены для студентов очной и заочной формы, обучающихся по направлениям подготовки УГСН 09 «Информатика и вычислительная техника», реализуемым в ФГБОУ ВО БГПУ им. М.Акмуллы, при изучении курсов «Эконометрика», «Анализ и моделирование социально-экономических проблем и процессов», «Методы обработки и анализа данных» и других дисциплин, предполагающих освоение методов статистического анализа данных.

Методические указания утверждены на заседании кафедры прикладной информатики 30.08.2019 г., протокол №1.

Табл.___. Ил. ___ . Библиогр.: ___ назв.

Содержание

Лабораторная работа № 1
Линейная регрессия. Коэффициент детерминации. Коэффициент корреляции. Его значимость

1. Цель и задачи лабораторной работы

Цель работы: изучить возможности MS Excel для построения парной линейной регрессии и корреляционного анализа.

Задачи:

– приобрести навыки расчета коэффициента детерминации;

– приобрести навыки расчета коэффициента корреляции и определения его значимости;

– научиться находить коэффициенты регрессии и строить уравнение;

– научиться строить диаграмму рассеяния средствами MS Excel;

– приобрести навык использования статистических функций MS Excel для проведения корреляционного и регрессионного анализа.

2. Теоретическая часть

Парная регрессия – это уравнение связи двух переменных y и x:

$$y = f(x),$$

где y – зависимая (эндогенная) переменная;

x – независимая (экзогенная), объясняющая переменная.

Различают *линейные* и *нелинейные* регрессии.

Линейная регрессия: $\hat{y}_x = a + b \cdot x$.

Построение уравнения регрессии сводится к оценке ее параметров. Для оценки параметров регрессий, линейных по параметрам, используют метод наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака y от теоретических y_x минимальна.

2.1. Определение параметров линейного уравнения регрессии

Для линейных и нелинейных уравнений, приводимых к линейным, решается следующая система относительно a и b:

$$\begin{cases} a \cdot n + b \cdot \sum x = \sum y; \\ a \cdot \sum x + b \cdot \sum x^2 = \sum x \cdot y. \end{cases}$$

Можно воспользоваться готовыми формулами, которые вытекают из этой системы:

$$a = \overline{y} - b \cdot \overline{x}, \qquad b = \frac{cov(x,y)}{\sigma_x^2} = \frac{cov(x,y)}{var(x)},$$

где $cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \overline{xy} - \overline{x}\cdot\overline{y}$ – ковариация признаков *x* и *y*,

$\sigma_x^2 = \overline{x^2} - \overline{x}^2$ – дисперсия признака *x* и

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{y_1 + y_2 + \cdots + y_n}{n},$$

$$\overline{y \cdot x} = \frac{1}{n}\sum_{n} y \cdot x\,, \qquad \overline{x^2} = \frac{1}{n}\sum x^2$$

$$var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \overline{x^2} - \overline{x}^2 = \sigma_x^2,$$

$$var(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \overline{y^2} - \overline{y}^2 = \sigma_y^2,$$

$$\sigma_x = \sqrt{\frac{\sum_{k=1}^{n}(x_k - \overline{x})}{n}} = \sqrt{var(x)}, \qquad \sigma_y = \sqrt{\frac{\sum_{k=1}^{n}(y_k - \overline{y})}{n}} = \sqrt{var(y)}$$

Параметр b называется коэффициентом регрессии. Его величина показывает среднее изменение результата с изменением фактора на одну единицу.

2.2. Расчет коэффициента корреляции

Тесноту связи изучаемых явлений оценивает линейный коэффициент парной корреляции для линейной регрессии ($-1 \le r_{xy} \le 1$):

$$r_{xy} = b \cdot \frac{\sigma_x}{\sigma_y} = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y} = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\sqrt{\left(\overline{x^2} - \overline{x}^2\right)\left(\overline{y^2} - \overline{y}^2\right)}} = \frac{cov(x,y)}{\sqrt{var(x)\cdot var(xy)}}$$

Теснота линейной связи между переменными может быть оценена на основании шкалы Чеддока:

Теснота связи	Значение коэффициента корреляции при наличии	
	прямой связи	обратной связи
Слабая	0,1–0,3	(–0,3) –(–0,1)
Умеренная	0,3–0,5	(–0,5) –(–0,3)
Заметная	0,5–0,7	(–0,7) –(–0,5)
Высокая	0,7–0,9	(–0,9) –(–0,7)
Весьма высокая (сильная)	0,9–1	(–1) –(–0,9)

Положительное значение коэффициента корреляции говорит о положительной связи между *x* и *y*, когда с ростом одной из переменных другая тоже растет. Отрицательное значение коэффициента корреляции означает, с

ростом одной из переменных другая убывает, с убыванием одной из переменной другая растет.

2.3. Оценка значимости коэффициента корреляции

Оценку статистической значимости коэффициента корреляции проводят с помощью *t*-критерия Стьюдента. Выдвигают гипотезу H_0 о статистически незначимом отличии коэффициента от нуля. Оценка значимости коэффициента корреляции с помощью t-критерия Стьюдента проводится путем сопоставления его значения с величиной случайной ошибки:

$$t_r = \frac{r}{m_r}.$$

Стандартная (случайная) ошибка коэффициента корреляции определяется по формуле:

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

Сравнивая фактическое и табличное (критическое) значения *t*-статистики – $t_{табл}$ и $t_{факт}$ – принимает или отвергаем гипотезу H_0.

Если $t_{табл} < t_{факт}$, то гипотеза H_0 отклоняется. Если $t_{табл} > t_{факт}$, то гипотеза H_0 не отклоняется и признается случайная природа формирования коэффициента корреляции.

2.4. Расчет коэффициента детерминации

Коэффициент детерминации характеризует долю дисперсии, объясняемую регрессией, в общей дисперсии результативного признака *y*.

$$R^2 = 1 - \frac{\sum(y_x - y)^2}{\sum(y - \overline{y})^2}.$$

Чем ближе коэффициент детерминации к 1, тем выше качество уравнения регрессии, тем в большей мере оно объясняет поведение эндогенной переменной.

3. Описание оборудования и используемых программных комплексов

При выполнении лабораторной работы необходим специализированный компьютерный класс с минимальными системными требованиями компьютеров:

- Процессор – Intel Pentium IV;
- ОЗУ – 500 Mb;
- видеокарта – 64 Mb.
- Требуемое программное обеспечение:
- Операционная система Microsoft Windows;
- Microsoft Excel версии 2007 и выше.

4. Краткое руководство по эксплуатации оборудования

При использовании оборудования необходимо:

- соблюдать общие правила нахождения в учебных лабораториях, работы с компьютером и использования программных средств;
- осмотреть рабочее место, убрать все мешающие работе предметы;
- визуально проверить правильность подключения ПЭВМ к электросети.

5. Задание

По предприятиям легкой промышленности региона получена информация, характеризующая зависимость объема выпуска продукции (y, млн. руб.) от объема капиталовложений (x, млн. руб.)

Таблица 1.1. Варианты для заданий

№		1	2	3	4	5	6	7	8	9	10
1	x	66	58	73	82	81	84	55	67	81	59
	y	133	107	145	162	163	170	104	132	159	116
2	x	72	52	73	74	76	79	54	68	73	64
	y	121	84	119	117	129	128	102	111	112	98
3	x	38	28	27	37	46	27	41	39	28	44
	y	69	52	46	63	73	48	67	62	47	67
4	x	36	28	43	52	51	54	25	37	51	29
	y	104	77	117	137	143	144	82	101	132	77
5	x	31	23	38	47	46	49	20	32	46	24
	y	38	26	40	45	51	49	34	35	42	24
6	x	33	17	23	17	36	25	39	20	13	12
	y	43	27	32	29	45	35	47	32	22	24
7	x	36	28	43	52	51	54	25	37	51	29
	y	85	60	99	117	118	125	56	86	115	68
8	x	17	22	10	7	12	21	14	7	20	3
	y	26	27	22	19	21	26	20	15	30	13
9	x	12	4	18	27	26	29	1	13	26	5
	y	21	10	26	33	34	37	9	21	32	14
10	x	26	18	33	42	41	44	15	27	41	19
	y	43	28	51	62	63	67	26	43	61	33

По заданной выборке исследовать зависимость результата y от фактора x:

1. Создать таблицу данных.

2. Найти средние значения $\overline{x}, \overline{y}$, выборочные дисперсии S_x^2, S_y^2, исправленные средние квадратические отклонения $\overline{S}_x, \overline{S}_y$.
3. Найти коэффициент корреляции и проверить его значимость.
4. Найти коэффициент детерминации.
5. Найти коэффициенты линейного уравнения регрессии.
6. Дать экономическую интерпретацию значений коэффицента корреляции и параметров уравнения регрессии.
7. Построить диаграмму рассеяния и график уравнения регрессии.

6. Методика выполнения заданий

В табл. 1.2 приведены данные об объеме производства *y* (тыс.ед.) в зависимости от численности занятых *x* (тыс.чел.) некоторой фирмы.

Таблица 1.2. Исходные данные

x	11	13	15	18	20	22	24	25	27
y	15	17	21	20	28	33	34	32	29

1. В диапазоне B3:C11 подготовим исходные данные (рис. 1.1).
2. В ячейках D3:D11 рассчитаем произведение *x* и *y*, в ячейках E3:E11 и F3:F11 квадраты значений *x* и *y*, в ячейках B12:F12 с помощью функции СРЗНАЧ рассчитаем средние значения рассмотренных величин.
3. В ячейках A17 и B17 рассчитаем выборочные дисперсии $S_x^2 = \overline{x^2} - \overline{x}^2$, $S_y^2 = \overline{y^2} - \overline{y}^2$
4. В ячейках A21 и B21 рассчитаем исправленные средние квадратические отклонения $\overline{S}_x, \overline{S}_y$. Для этого воспользуемся функцией СТАНДОТКЛОН. Она оценивает стандартное отклонение по выборке (мера того, насколько широко разбросаны точки данных относительно их среднего).
5. В ячейке E16 рассчитаем коэффициент корреляции. Для этого воспользуемся формулой

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

Из расчетов (рис. 1.2) следует, что коэффициент корреляции r = 0,902. Это свидетельствует о том, что связь между объемом выпуска продукции и численностью занятых весьма высокая и положительная.

	A	B	C	D	E	F	G
1	**Простейшая обработка данных**						
2		x	y	xy	x^2	y^2	
3	1	11	15	165	121	225	
4	2	13	17	221	169	289	
5	3	15	21	315	225	441	
6	4	18	20	360	324	400	
7	5	20	28	560	400	784	
8	6	22	33	726	484	1089	
9	7	24	34	816	576	1156	
10	8	25	32	800	625	1024	
11	9	27	29	783	729	841	
12	среднее значение	19,44	25,44	527,33	405,89	694,33	
13							
14							
15	**Выборочные средние**						
16	S_x^2	S_y^2					
17	27,80	46,91					
18							
19	**Исправленные средние квадратичные**						
20	$S_{x\ испр}$	$S_{y\ испр}$					
21	5,59	7,26					

Рис. 1.1 – Результаты простейшей обработки данных

6. Для проверки значимости коэффициента корреляции введем вспомогательные данные. Ячейка L16 – число предприятий (n): 9; ячейка L17 – уровень значимости: 0,05.
7. В ячейке H20 определим стандатрную ошибку по следующей формуле:

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

8. В ячейке H21 рассчитаем значение t-статистики по формуле:

$$t_r = \frac{r}{m_r}.$$

9. Критическое значение t-статистики определим с помощью функции Excel СТЬЮДРАСПОБР. Она возвращает двустороннее обратное t-распределения Стьюдента.

Синтаксис: ***СТЬЮДРАСПОБР (вероятность,степени_свободы).***

Аргументы:

Вероятность – вероятность, соответствующая двустороннему распределению Стьюдента;

Степени_свободы – число степеней свободы, характеризующее распределение.

В качестве вероятности укажем уровень значимости. Число степеней равно *n*-*m*-1, где *m* – число независимых переменных в модели (в нашем случае она всего одна – *x*)

10. Для наглядного отображения вывода воспользуемся функцией ЕСЛИ и условным форматированием: если рассчетное значение t-статистики больше критического, то коэффициент корреляции значим (выделяем зеленым), в противном случае незначим (выделяем красным).
11. В ячейке H16 рассчитаем коэффициент детерминации как квадрат коэффициента корреляции.

	C	D	E	F	G	H	I	J	K	L
14										
15		**Коэффициент корреляции**			**Коэффициент детерминации**			**Вспомогательные данные**		
16		r_{xy}	0,902		R^2_{xy}	0,814		n		9
17								уровень значимости		0,05
18										
19		**Проверка значимости коэффициента корреляции**								
20		стандартная ошибка				0,1631				
21		t-статистика				5,5315				
22		Критическое значение t-статистики				2,3646				
23		Вывод				Значим				

Рис. 1.2 – Расчет коэффициента корреляции, коэффициента корреляции и анализ его значимости

12. Для определения коэффициентов уравнения линейной регрессии на основе формул (рис. 1.3):

$$b = \frac{\overline{yx} - \overline{y} \cdot \overline{x}}{\overline{x^2} - \overline{x}^2}; \quad a = \overline{y} - b \cdot \overline{x}$$

В нашем примере уравнение имеет вид: $\hat{y} = 2{,}659 + 1{,}172 \cdot x$

Значение коэффициента b=1,172 говорит о том, что при увеличении численности занятых на 1 тыс.чел. объем продукции увеличится на 1,172 тыс.ед.

	G	H	I	J
1				
2		Коэффициенты регрессии		
3		*b*	1,172	
4		*a*	2,659	
5				

Рис. 1.3 – Результаы расчета параметров уравнения регрессии

13. Для построения диаграммы рассеяния выделим диапазон B3:C11. Во вкладке «Вставка» выберем тип диаграммы – «Точечная». На построенной диаграмме выделим нанесенные значения, щелкнув по ним левой кнопкой мыши. Нажав правую кнопку мыши, выведем контекстно-зависимое меню, в котором выберем опцию Добавить линию тренда. В окне Линия тренда по вкладке «Параметры линии тренда» выберем тип функции «Линейная», установим флажок «показывать уравнение на диаграмме» и «поместить на диаграмму величину достоверности аппроксимации (R^2)». В результате на диаграмме появиться вид теоретической кривой – тренда, ее уравнение и коэффициент детерминации (рис.1.4). Добавим подписи осей, заголовок диаграммы и легенду.

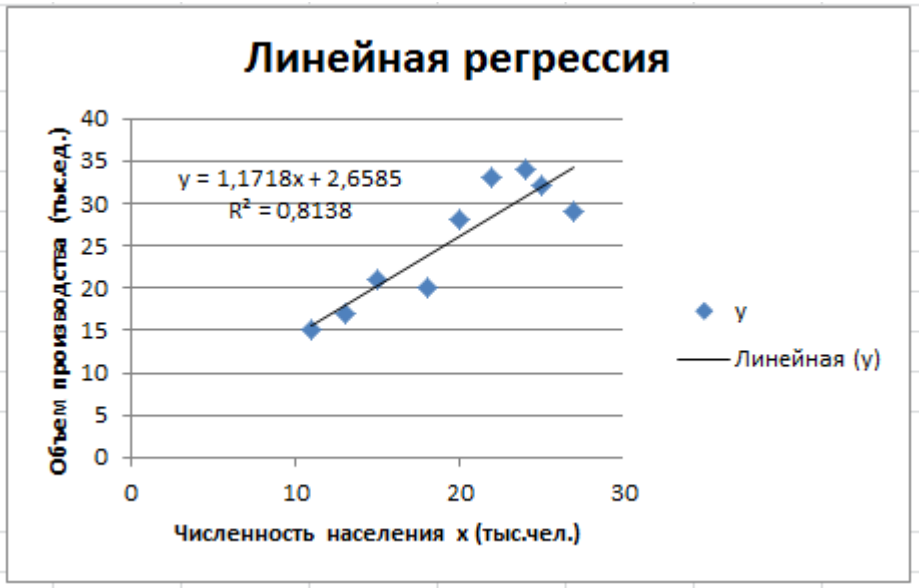


Рис. 1.4 – Графики фактических данных и построенной регрессии

14. Вычисление параметров регрессии с помощью статистических функций Excel:

КОРРЕЛ(массив1;массив2) вычисляет коэффициент корреляции между двумя переменными; значения первой из них приведены в диапазоне массив1, значения второй – в диапазоне массив2;

НАКЛОН(известные_значения_y;известные_значения_x) служит для определения коэффициента b;

ОТРЕЗОК(известные_значения_y;известные_значения_x) служит для определения коэффициента a.

Рассчитаем с их помощью коэффициент корреляции в ячейке E27, параметры *a* и *b* соответственно в ячейках E28 и E29 (рис. 1.5).

	A	B	C	D	E	F	G	H
25	**Расчет параметров регрессии с помощью статистических функций Excel**							
26	**Линейн**							
27	1,171847	2,658526		r_{xy}	0,902118			
28	0,21185	4,268075		*b*	1,171847			
29	0,813817	3,351131		*a*	2,658526			
30	30,59743	7						
31	343,6117	78,61057						

Рис. 1.5 – Расчет параметров регрессии с помощью функций Excel

15. Встроенная статистическая функция ЛИНЕЙН определяет параметры линейной регрессии. Порядок вычислений следующий:

1) выделите ячейку A27, нажмите на кнопку «Вставить функцию» (f_x);

2) в строке Категория (рис.1.6) выберите Статистические, в окне Функция – ЛИНЕЙН. Щелкните ОК.

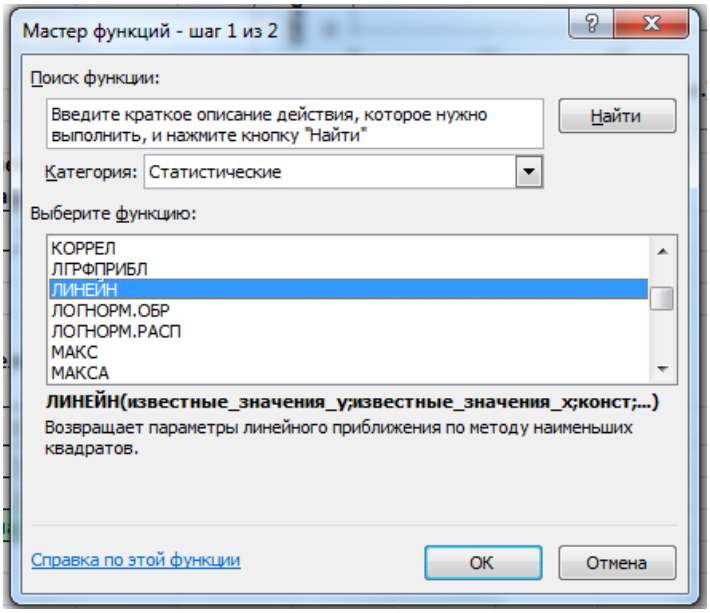


Рис. 1.6 – Диалоговое окно «Мастер функций»

4) Заполните аргументы функции (рис.1.7.):

Известные_значения_y – диапазон, содержащий данные результативного признака;

Известные_значения_x – диапазон, содержащий данные факторов независимого признака;

Константа – логическое значение, которое указывает на наличие или на отсутствие свободного члена в уравнении; если Константа = 1, то свободный член рассчитывается обычным образом, если Константа = 0, то свободный член равен 0.

Статистика – логическое значение, которое указывает выводить дополнительную информацию по регрессионному анализу или нет. Если

Статистика = 1, то дополнительная информация выводится, если Статистика = 0, то выводится только оценки параметров уравнения.

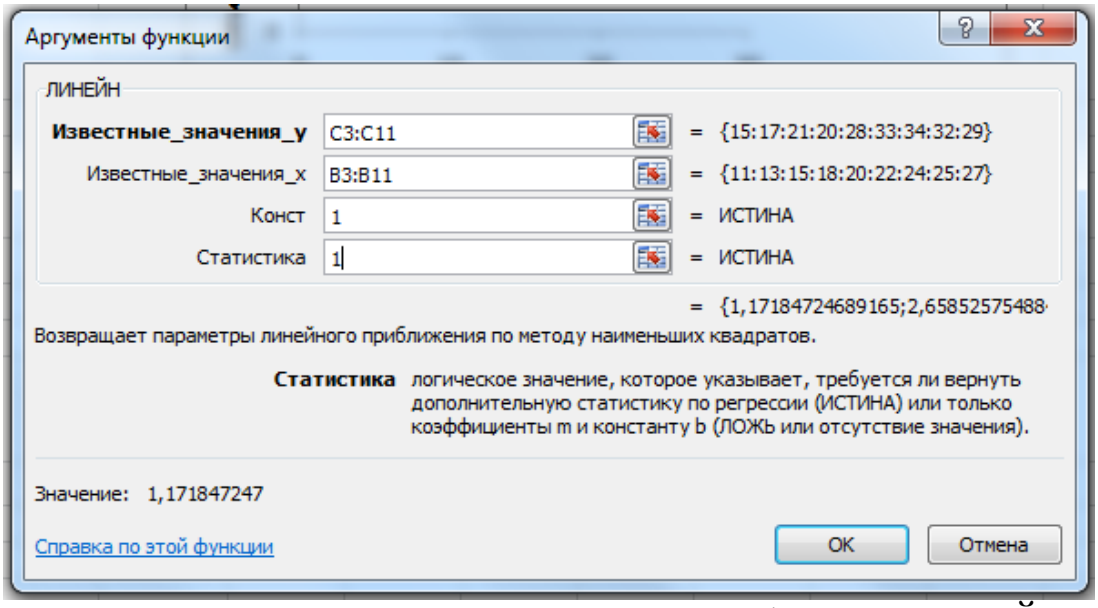


Рис. 1.7 – Диалоговое окно ввода аргументов функции ЛИНЕЙН

5) появится первый элемент итоговой таблицы. Чтобы вывести результаты регрессионной статистики, выделите область пустых ячеек 5x2 (A27:B31). Нажмите на клавишу F2, а затем на комбинацию клавиш CTRL+SHIFT+ENTER. Дополнительная регрессионная статистика будет выводиться в порядке, указанном в следующей схеме:

Значение коэффициента b	Значение коэффициента a
Среднеквадратическое отклонение b	Среднеквадратическое отклонение a
Коэффициент детерминации R^2	Среднеквадратическое отклонение y
F-статистика	Число степеней свободы
Регрессионная сумма квадратов	Остаточная сумма квадратов

Результаты регрессионного анализа представлены на рис.1.5.

7. Контрольные вопросы

1. Какова экономическая интерпретация параметров уравнения регрессии?
2. Что означает отрицательное значение коэффициента корреляции?
3. Назовите диапазон изменения значений коэффициента корреляции и коэффициента детерминации
4. Что является показателем тесноты связи в парной линейной регрессии?
5. Каково значение коэффициента корреляции?
6. Каково значение коэффициента детерминации и что он характеризует?
7. Как оценивается значимость коэффициента корреляции?
8. Какие функции Excel можно использовать для определения параметров линейного уравнения регрессии?
9. Какие функции Excel можно использовать для определения коэффициен корреляции?
10. Для чего используется функция СТЬЮДРАСПОБР?

8. Требования к содержанию отчета

Отчет к лабораторной работе предоставляется в электронном виде и должен содержать:

- название и цель работы;
- номер и исходные данные своего варианта;
- описание хода выполнения заданий, в том числе:
  - скриншоты из MS Excel, отображающие заданные в ячейках формулы и функции, использованные для вычислений;
  - скриншоты из MS Excel с полученными при расчетах результатами;
  - анализ и интерпретация полученных результатов;
- выводы по лабораторной работе.

9. Требования к оформлению отчета

- Все рисунки в отчете должны быть подписаны.
- Все скриншоты должны быть читаемыми в масштабе документа 100%. При необходимости используйте обрезку и пропорциональное изменение размера рисунка.
- Все скриншоты таблиц MS Excel должны содержать системное наименование строк (1, 2, 3….) и столбцов (A, B, C, …)

Лабораторная работа №2
Проверка качества уравнения линейной регрессии. Прогнозирование на основании линейной регрессии

1. Цель и задачи лабораторной работы

Цель работы: изучить возможности MSExcel для проверки качества уравнения линейной регрессии и прогнозирования индивидуальных значений зависимой переменной на основании линейной регрессии.

Задачи:

– научиться проверять статистическую значимость коэффициентов линейной регрессии;

– научиться определять общее качество уравнения линейной регрессии;

– научиться прогнозировать индивидуальные значения зависимой переменной на основании линейной регрессии;

– научиться определять точность прогноза.

2. Теоретическая часть

2.1. Проверка качества уравнения линейной регрессии

Оценку качества построенной модели дает коэффициент (индекс) детерминации r_{xy}^2 (ρ_{xy}^2), а также средняя ошибка аппроксимации.

Средняя ошибка аппроксимации – среднее отклонение расчетных значений от фактических:

$$\overline{A} = \frac{1}{n}\sum \left|\frac{y - y_x}{y}\right| \cdot 100\%.$$

Допустимый предел значений средней ошибки аппроксимации – не более 8–10%.

Согласно основной идее дисперсионного анализа, общая сумма квадратов отклонений переменной *y* от среднего значения $\overline{y}$ раскладывается на две части – «объясненную» и «необъясненную»:

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

где $\sum(\hat{y}_x - \overline{y})^2$ – общая сумма квадратов отклонений; $\sum(\hat{y}_x - \overline{y})^2$ – сумма квадратов отклонений, объясненная регрессией (или факторная сумма квадратов отклонений); $\sum(y - \hat{y}_x)^2$ – остаточная сумма квадратов отклонений, характеризующая влияние неучтенных в модели факторов.

$$\sum (\hat{y}_x - \overline{y})^2 = n\sigma_y^2;$$

$$\sum (\hat{y}_x - \overline{y})^2 = n\sigma_y^2 R^2 = b^2\, n\sigma_x^2;$$

$$\sum (y - \hat{y}_x)^2 = n\sigma_y^2(1 - R^2)$$

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*:

$$F = \frac{S^2_{\text{факт}}}{S^2_{\text{ост}}}.$$

Фактическое значение *F*-критерия Фишера сравнивается с табличным значением $F_{\text{табл}}(\alpha;\ k_1;\ k_2)$ при уровне значимости α и степеняхсвободы $k_1 = m$ и $k_2 = n\text{-}m\text{-}1$. При этом, если фактическое значение *F*-критерия больше табличного, то признается статистическая значимостьуравнения в целом.

Для парной линейной регрессии $m = 1$, поэтому

$$F = \frac{S^2_{\text{факт}}}{S^2_{\text{ост}}} = \frac{\sum(\hat{y}_x - \overline{y})^2}{\sum(y - \hat{y}_x)^2} \cdot (n - 2).$$

Величина *F*-критерия связана с *коэффициентом детерминации* r^2_{xy}, и ее можно рассчитать по следующей формуле:

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

В парной линейной регрессии оценивается значимость не только уравнения в целом, но и отдельных его параметров. С этой целью по каждому из параметров определяется его стандартная ошибка: m_b и m_a.

Стандартная ошибка коэффициента регрессии определяется по формуле:

$$m_b = \sqrt{\frac{S^2_{\text{ост}}}{\sum(x - \overline{x})^2}} = \frac{S_{\text{ост}}}{\sigma_x \cdot n},$$

где $S^2_{\text{ост}} = \frac{\sum(y - \hat{y}_x)^2}{n-2}$ – остаточная дисперсия на одну степень свободы.

Величина стандартной ошибки совместно с *t*–распределением Стьюдента при $n - 2$ степенях свободы применяется для проверки существенности коэффициента регрессии и для расчета егодоверительного интервала.

Для оценки существенности коэффициента регрессии его величинасравнивается с его стандартной ошибкой, т.е. определяется фактическоезначение *t*-критерия Стьюдента:

$$t_b = \frac{b}{m_b},$$

которое затем сравнивается с табличным значением при определенном уровне значимости α и числе степеней свободы $n - 2$.

Доверительный интервал для коэффициента регрессии определяется как $b \pm t_{\text{табл}} \cdot m_b$. Доверительным называют интервал, который покрывает неизвестный параметр с заданной надёжностью.

Стандартная ошибка параметраа определяется по формуле:

$$m_a = \sqrt{S_{\text{ост}}^2 \cdot \frac{\sum x^2}{n \cdot \sum (x - \overline{x})^2}} = S_{\text{ост}} \cdot \frac{\sqrt{\sum x^2}}{\sigma_x \cdot n}.$$

Процедура оценивания существенности данного параметра не отличается от рассмотренной выше для коэффициента регрессии.

Вычисляется *t*-критерий $t_a = \frac{a}{m_a}$,, его величина сравнивается с табличным значением при $n-2$ степенях свободы. Доверительный интервал для коэффициента регрессии определяется как $a \pm t_{\text{табл}} \cdot m_a$.

Если в границы доверительного интервала попадает ноль, т.е. нижняя граница отрицательна, а верхняя положительна, то оцениваемый параметр принимается нулевым, т.к. он не может одновременно принимать и положительное, и отрицательное значения.

2.2. Прогнозирование на основании линейной регрессии

Пусть по заданной выборке объема *n* найдено выборочное уравнение линейной регрессии $y = a + bx$. С помощью этого уравнения можно прогнозировать значениерезультата y_p при определенном прогнозном значении фактора x_p.

Прогнозное значение y_p определяется путем подстановки в уравнение регрессии $y = a + bx$ соответствующего прогнозного значения x_p.

Точное уравнение регрессии нам неизвестно. Поэтому мы не можем делать точный прогноз. Можно только утверждать, что прогнозное значение результата y_p при данном x_p с вероятностью γ попадет в доверительный интервал γ_p. Вероятность γ называется уровнем надежности.

Ошибка прогноза составляет:

$$m_p = S \sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{\sum (x_p - \overline{x})^2}},$$

где $S = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n-2}} = \sqrt{\frac{\sum (y - \hat{y}_x)^2}{n-2}}$ – стандартная ошибка регрессии (дисперсия ошибки или остаточная дисперсия).

Предельная ошибка прогноза, составит:

$$\Delta_p = t_{\text{табл}} \cdot m_p.$$

Доверительный интервал прогноза:

$$\gamma_p = y_p \pm \Delta_p.$$

Точность прогноза можно оценить с помощью относительной ошибки прогноза:

$$\delta_p = \frac{\Delta_p}{|y_p|} \cdot 100\%.$$

3. Описание оборудования и используемых программных комплексов

При выполнении лабораторной работы необходим специализированный компьютерный класс с минимальными системными требованиями компьютеров:

- Процессор – Intel Pentium IV;
- ОЗУ – 500 Mb;
- видеокарта – 64 Mb.
- Требуемое программное обеспечение:
- Операционная система Microsoft Windows;
- MicrosoftExcel версии 2007 и выше.

4. Краткое руководство по эксплуатации оборудования

При использовании оборудования необходимо:

- соблюдать общие правила нахождения в учебных лабораториях, работы с компьютером и использования программных средств;
- осмотреть рабочее место, убрать все мешающие работе предметы;
- визуально проверить правильность подключения ПЭВМ к электросети.

5. Задание

Используя данные к лабораторной работе №1, найти уравнение линейной регрессии и проверить:

1. Значимость коэффициента b. Для этого надо найти:
   1) Сумму квадратов остатков: $\sum e_i^2 = \sum (y - \hat{y}_x)^2$;
   2) Сумму квадратов отклонений: $\sum (x - \overline{x})^2$;
   3) Стандартную ошибку параметра b:

$$m_b = \sqrt{\frac{\sum (y - \hat{y}_x)^2}{(n-2)\sum (x - \overline{x})^2}};$$

   4) Наблюдаемое значение t-статистики параметра b: $t_b = \frac{b}{m_b}$;
   5) Число степеней свободы $k = n-2$ и критическое значение $t_{\text{табл}} = t_{\alpha;k}$;

6) Сделать вывод о значимости коэффициента b: если $t_{\text{табл}} < |t_b|$, то параметр регрессии b статистически значим, а в противном случае статистически незначим.

2. Значимость коэффициента a. Для этого надо найти:
   1) Сумму квадратов: $\sum x^2$;
   2) Стандартную ошибку параметра a:
$$m_a = m_b \sqrt{\frac{\sum x^2}{n}};$$
   3) Наблюдаемое значение t-статистики параметра a: $t_a = \frac{a}{m_a}$;
   4) Сделать вывод о значимости коэффициента a: если $t_{\text{табл}} < |t_a|$, то параметр регрессии a статистически значим, а в противном случае статистически незначим.
3. Общее качество уравнения регрессии. Для этого надо найти:
   1) Сумму квадратов отклонений: $\sum(y - \overline{y})^2$;
   2) Коэффициент детерминации:
$$R^2 = 1 - \frac{\sum e_i^2}{\sum(y - \overline{y})^2};$$
   3) Наблюдаемое значение F-статистики:
$$F_{\text{факт}} = \frac{R^2}{1 - R^2} \cdot (n - 2);$$
   4) Число степеней свободы критерия Фишера-Снедекора: $k_1 = 1;\ k_2 = n - 2$ и критическое значение этого критерия $F_{\text{табл}} = F_{\alpha;k_1;k_2}$;
   5) Сделать вывод о значимости уравнения регрессии: если $F_{\text{факт}} > F_{\text{табл}}$, то уравнение регрессии статистически значимо и надежно, если $F_{\text{факт}} < F_{\text{табл}}$ признается статистическая незначимость, ненадежность уравнения регрессии.
4. Общее качество уравнения регрессии с помощью средней ошибки аппроксимации. Для этого надо найти:
   1) Отклонения: $y - y_x$;
   2) Ошибки аппроксимации:
$$A_i = \left|\frac{y - y_x}{y}\right| \cdot 100\%;$$
   3) Среднюю ошибку аппроксимации $\overline{A}$:
$$\overline{A} = \frac{1}{n} \cdot \sum A_i;$$
   4) Сделать вывод о качестве уравнения регрессии: если $\overline{A}$ не превышает предела значений 8-10%, то качество модели хорошее.

5. Представить результаты с помощью инструмента анализа данных Регрессия ППП Excel.
6. Используя построенную модель, выполнить прогнозирование:
   1) Определить прогнозное значение y_p при $x_p = 20$;
   2) Рассчитать доверительный интервал прогноза для уровня надежности 90%;
   3) Найти относительную ошибку прогноза;
   4) Построить графики линии регрессии с доверительными границами;

6. Методика выполнения заданий

1. В диапазоне A2:C11 подготовить исходные данные (рис. 2.1).
2. Введем следующие вспомогательные данные: в ячейку C16 – число предприятий, C17 – уровень значимости (0,05), в ячейках C18 иC19 рассчитаем значения параметров a и b, в ячейках C20 иC21 – соответственно средние значения фактора и результата (рис. 2.1).

	A	B	C	D	E	F	G
1	Проверка качества уравнения линейной регрессии						
2		x	y	y_x	e^2	$(x-x_{ср.})^2$	x^2
3	1	11	15	15,55	0,30	71,31	121
4	2	13	17	17,89	0,80	41,53	169
5	3	15	21	20,24	0,58	19,75	225
6	4	18	20	23,75	14,08	2,09	324
7	5	20	28	26,10	3,63	0,31	400
8	6	22	33	28,44	20,80	6,53	484
9	7	24	34	30,78	10,35	20,75	576
10	8	25	32	31,95	0,00	30,86	625
11	9	27	29	34,30	28,07	57,09	729
12	Среднее	**19,44**		Сумма	**78,61**	**250,22**	**3653**
13							
14							
15	**Вспомогательные данные**						
16	Число предприятий	9					
17	Уровень значимости	0,05					
18	Параметр *a*	2,659					
19	Параметр *b*	1,172					
20	Среднее значение *x*	19,444					
21	Среднее значение *y*	25,444					

Рис. 2.1 – Исходные и вспомогательные данные

Тогда линейное уравнение регрессии имеет вид: $\hat{y}_i = 2{,}659 + 1{,}172 \cdot x_i$

3. Проверим значимость параметра b (рис. 2.2):

3.1. Используя полученное уравнение линейной регрессии, рассчитаем в ячейках D3:D11 оцененное значение $\hat{y}_i$;

3.2. В ячейках E3:E11рассчитаем квадраты отклонений оцененных значений от исходных данных $e_i^2 = (y_i - \hat{y}_i)^2$ и просуммируем их в ячейке E12;

3.3. В ячейках F3:F11 рассчитаем квадраты отклонений независимой переменной от ее среднего значения $(x_i - \overline{x})^2$ и просуммируем их в ячейке F12;

3.4. В ячейке B24 определим стандартную ошибку параметра b по формуле

$$m_b = \sqrt{\frac{\sum(y - \hat{y}_x)^2}{(n-2)\sum(x - \overline{x})^2}};$$

3.5. В ячейке D25 рассчитаем t-статистику параметра b как отношение величины этого параметра к его стандартной ошибке:

$$t_b = \frac{b}{m_b};$$

3.6. С помощью функции СТЬЮДРАСПОБР в ячейке D26 определим критическое значение t-статистики для числа степеней свободы n-2 и уровня значимости α = 0,05;

3.7. Для получения автоматического вывода о значимости параметра b в ячейке D27 воспользуемся функцией ЕСЛИ и условным форматированием;

3.8. Для расчета доверительного интервала определим в ячейке D28 предельную ошибку $t_{\text{табл}} \cdot m_b$

3.9. В ячейках D29 и D30 определим нижнюю и верхнюю границы доверительного интервала: $b \pm t_{\text{табл}} \cdot m_b$

	A	B	C	D
23	**Значимость *b***			
24	стандартная ошибка			0,212
25	t-статистика			5,531
26	Критическое значение t-статистики			2,365
27	Вывод			Значим
28	Предельная ошибка			0,501
29	Нижняя граница			0,671
30	Верхняя граница			1,673

Рис. 2.2 – Оценка значимости параметра b

4. Проверим значимость параметра a (рис. 2.3):

4.1. Рассчитаем в ячейках G3:G11 квадраты значений x и просуммируем в ячейке G12;
4.2. В ячейке D33 определим стандартную ошибку параметра a по формуле:

$$m_a = m_b \sqrt{\frac{\sum x^2}{n}};$$

4.3. Аналогично параметру b определим для параметра a в ячейке D34 значение t-статистики, в ячейке D35 критическое значение t-статистики, в ячейке D36 сделаем вывод о значимости параметра a, в ячейке D37 рассчитаем предельную ошибку, в D38 и D39 вычислим границы доверительного интервала.

	A	B	C	D
32	**Значимость *a***			
33	стандартная ошибка			4,2680755
34	t-статистика			0,6228863
35	Критическое значение t-статистики			2,365
36	Вывод			Незначим
37	Предельная ошибка			10,092395
38	Нижняя граница			-7,434
39	Верхняя граница			12,751

Рис. 2.3 – Оценка значимости параметра a

5. Оценим общее качество уравнения регрессии с помощью F-теста (рис. 2.4):
5.1. В ячейке D42 определим коэффициент детерминации как квадрат коэффициента корреляции (можно воспользоваться функцией КОРРЕЛ);
5.2. В ячейке D43 рассчитаем наблюдаемое значение F-статистики по формуле:

$$F_{\text{факт}} = \frac{R^2}{1 - R^2} \cdot (n - 2);$$

5.3. В ячейке D44 рассчитаем критическое значение $F_{\text{табл}}$ для уровня значимости $\alpha = 0{,}05$ и $k_1 = 1$; $k_2 = n - 2$. Для этого воспользуемся функцией FРАСПОБР (она возвращает значение обратное распределению вероятностей Фишера).
5.4. В ячейке D45 сделаем вывод о значимости уравнения регрессии

	A	B	C	D
41	**Значимость уравнения регрессии**			
42	Коэффициент детерминации			0,8138171
43	F-статистика			30,597433
44	Критическое значение F-статистики			5,5914478
45	Вывод			Значимо

Рис. 2.4 – Оценка значимости уравнения с помощью F-теста

6. Оценим общее качество уравнения регрессии с помощью средней ошибки аппроксимации (рис. 2.5):
   6.1. В ячейках H3:H11 найдем отклонения: $y - y_x$;
   6.2. В ячейках I3:I11 рассчитаем ошибки аппроксимации по формуле:

$$A_i = \left|\frac{y - y_x}{y}\right| \cdot 100\%;$$

   6.3. В ячейке I12 вычислим среднюю ошибку аппроксимации $\overline{A}$:

$$\overline{A} = \frac{1}{n} \cdot \sum A_i$$

	A	B	C	D	E	F	G	H	I
1	**Проверка качества уравнения линейной регрессии**								
2		x	y	y_x	e^2	$(x-x_{ср.})^2$	x^2	$y-y_x$	A
3	1	11	15	15,55	0,30	71,31	121	-0,55	3,66
4	2	13	17	17,89	0,80	41,53	169	-0,89	5,25
5	3	15	21	20,24	0,58	19,75	225	0,76	3,64
6	4	18	20	23,75	14,08	2,09	324	-3,75	18,76
7	5	20	28	26,10	3,63	0,31	400	1,90	6,80
8	6	22	33	28,44	20,80	6,53	484	4,56	13,82
9	7	24	34	30,78	10,35	20,75	576	3,22	9,46
10	8	25	32	31,95	0,00	30,86	625	0,05	0,14
11	9	27	29	34,30	28,07	57,09	729	-5,30	18,27
12	Среднее	**19,44**		Сумма	**78,61**	**250,22**	**3653**	Среднее	**8,87**

Рис. 2.5 – Оценка общего качества уравнения регрессии с помощью средней ошибки аппроксимации

7. Представим результаты с помощью инструмента анализа данных Регрессия ППП Excel.

Инструмент анализа данных Регрессия применяется для подбора графика для набора наблюдений с помощью метода наименьших квадратов. Используется для анализа воздействия на отдельную зависимую переменную значений одной или нескольких независимых переменных. С помощью Регрессии можно получить результаты регрессионной статистики,

дисперсионного анализа и доверительных интервалов, остатки и графики подбора линии регрессии, остатков и нормальной вероятности.

7.1. Для запуска инструмента необходимо перейти к закладке Данные и нажать на кнопку Анализ данных. В случае отсутствия кнопки на панели необходимо подключить пакет анализа:

1) Нажать на кнопку «Office» ( ) в левом верхнем углу окна, затем нажать на «Параметры Excel».
2) В открывшемся окне «Параметры Excel» выбрать строку «Надстройки» и нажать на кнопку «Перейти…» в нижней части окна (рис. 2.6)
3) В открывшемся окне «Надстройки» отметить галочкой пункт «Пакет анализа» и нажать ОК (рис. 2.7)
4) Выполните инструкции программы установки, если это необходимо.

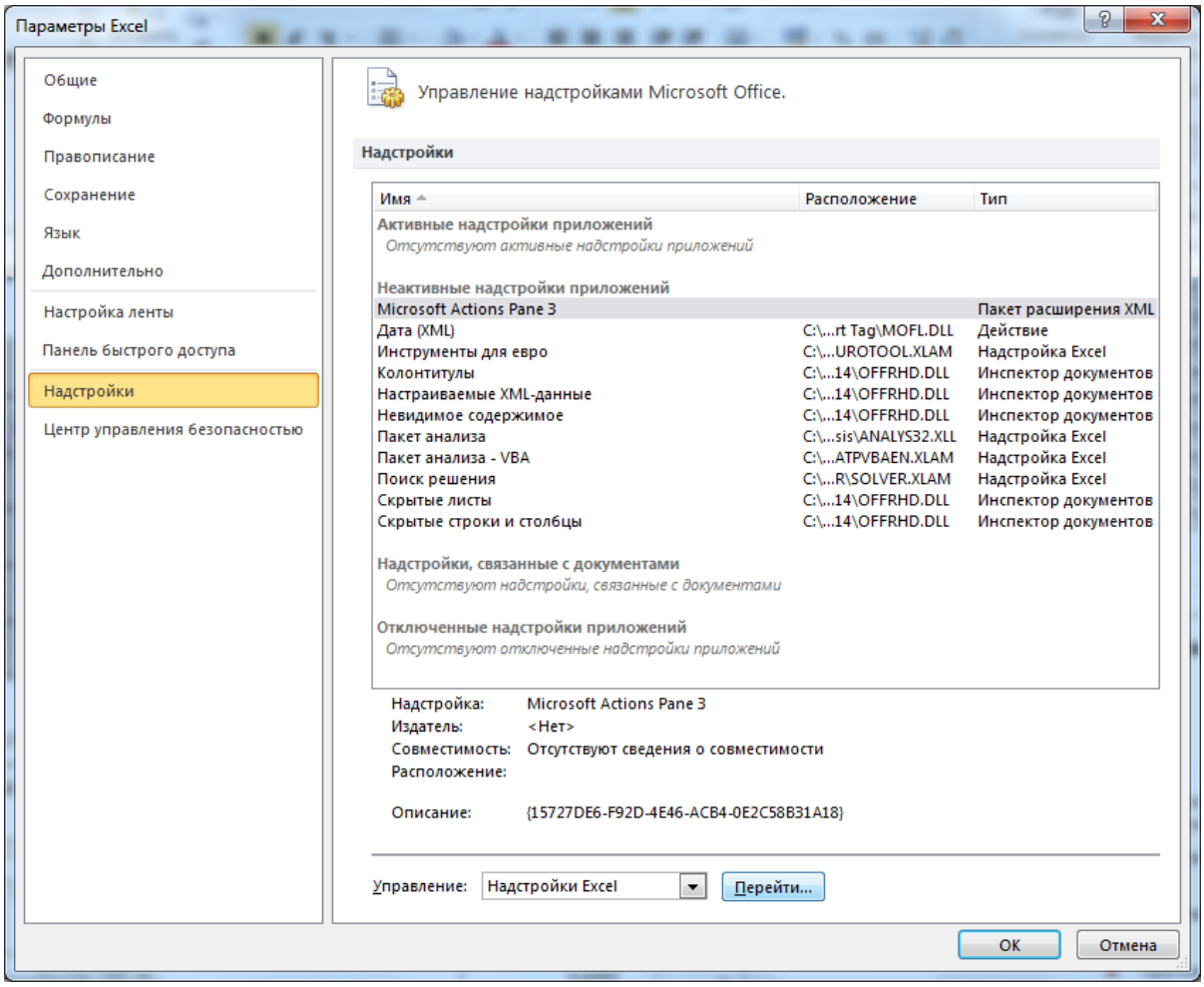


Рис. 2.6 – Подключение пакета анализа

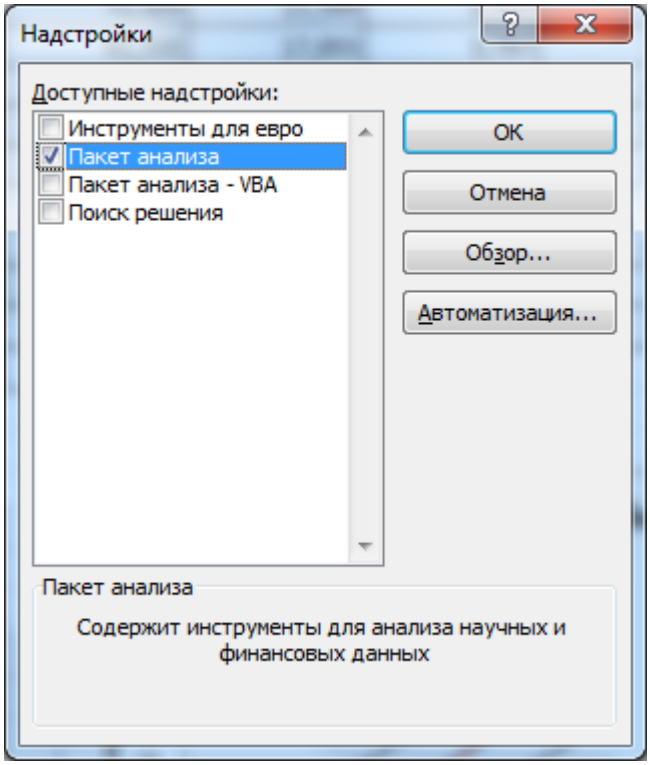


Рис. 2.7 – Окно Надстройки

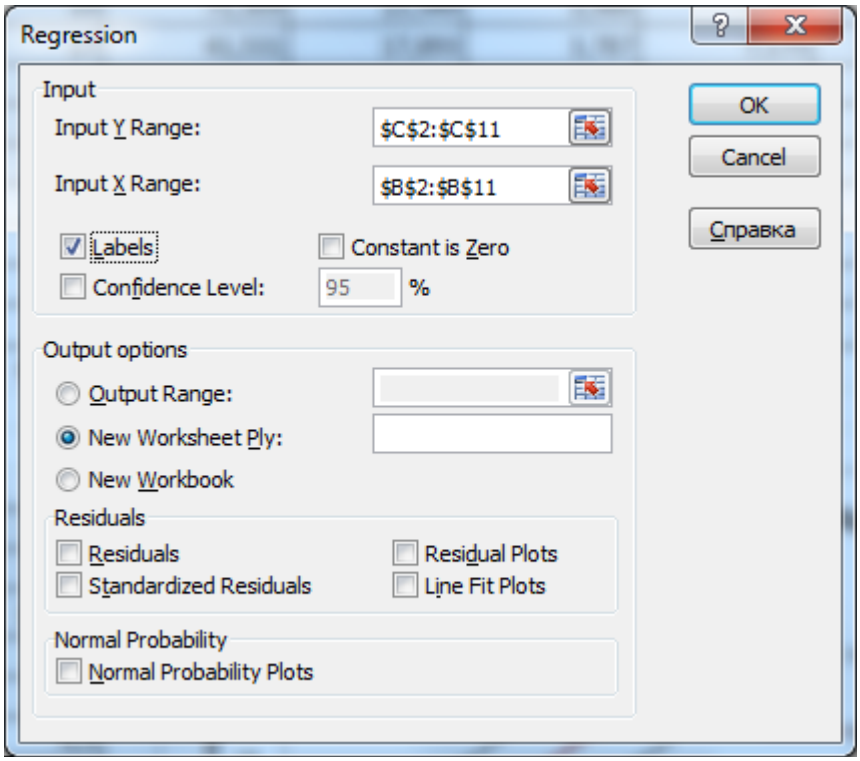


Рис. 2.8 – Параметры данных Регрессии

7.2. Заполните диалоговое окно ввода данных и параметров вывода (рис. 2.8):

Входной интервал Y – диапазон, содержащий данные результативного признака;

Входной интервал X – диапазон, содержащий данные независимого признака;

Метки – флажок, который указывает, содержит ли первая строка названия столбцов или нет (ОБРАТИТЕ ВНИМАНИЕ! Если, задавая входные данные, Вы не включили заголовки столбцов, то ставить флажок около «Метки» не нужно!);

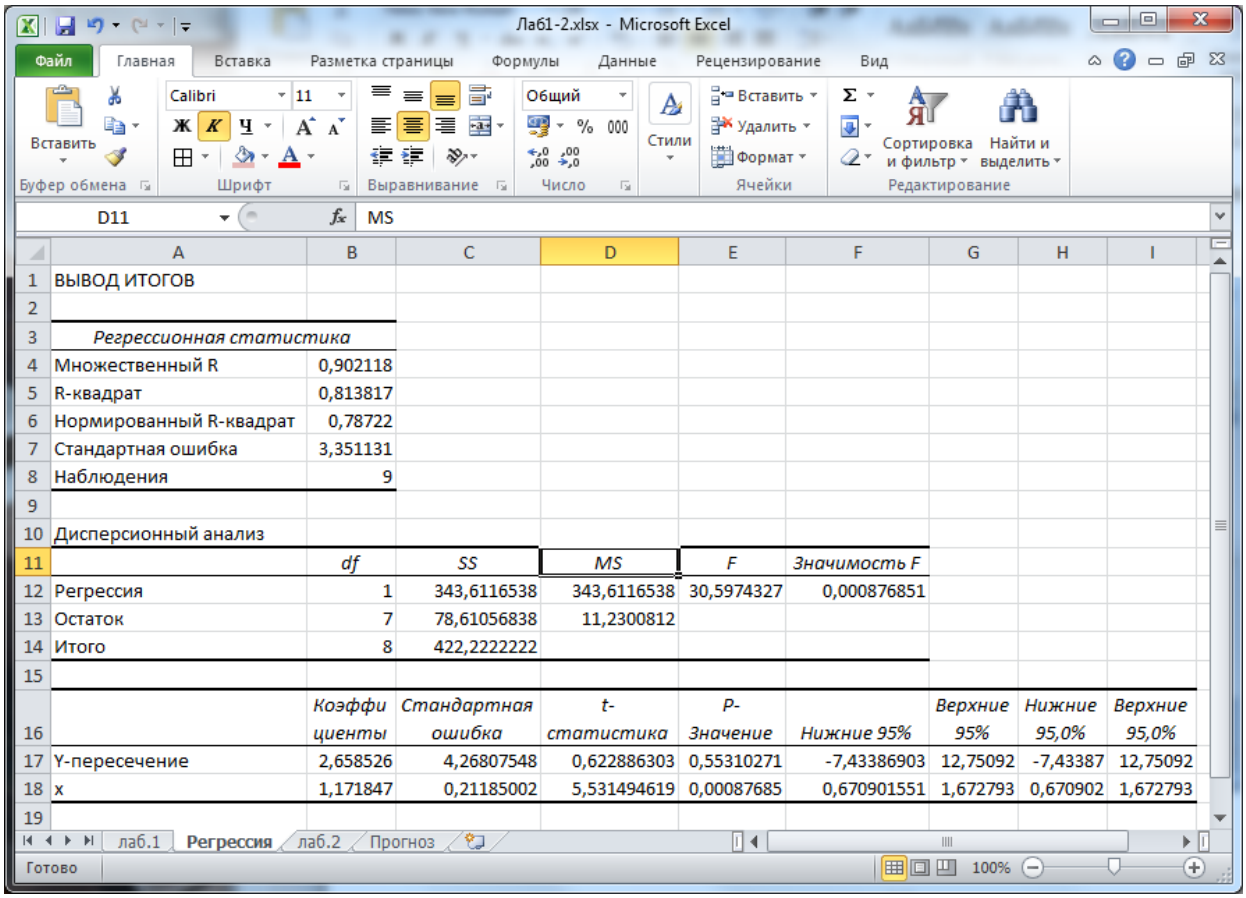
Константа – ноль – флажок, указывающий на наличие или отсутствие свободного члена в уравнении;

Выходной интервал – достаточно указать левую верхнюю ячейку будущего диапазона;

Новый рабочий лист – можно задать произвольное имя нового листа.

Если необходимо получить информацию и графики остатков, установите соответствующие флажки в диалоговом окне. Нажмите ОК.

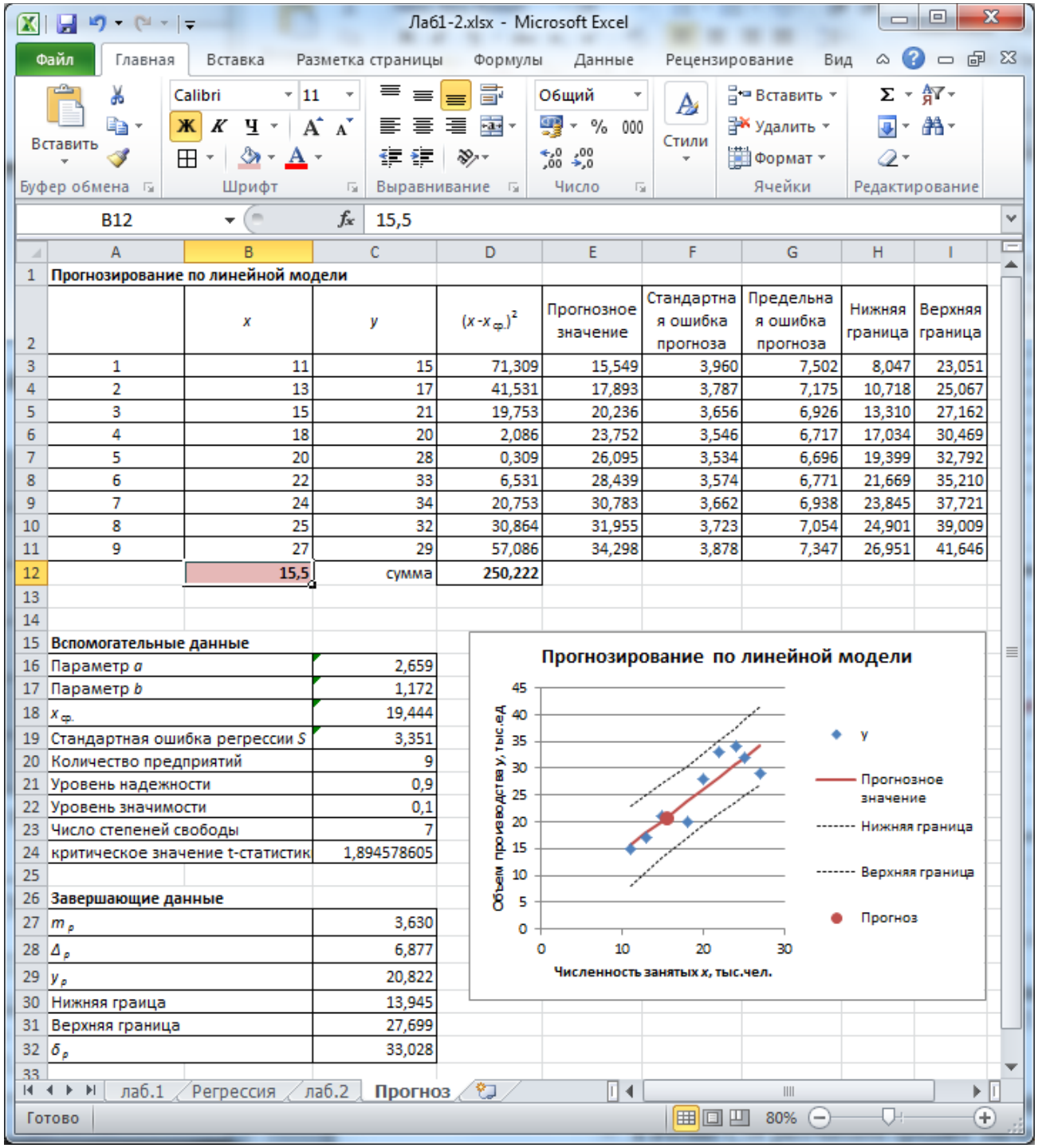
7.3. На новом рабочем листе появляются данные регрессионного анализа (рис. 2.9)



	A	B	C	D	E	F	G	H	I
1	ВЫВОД ИТОГОВ								
2									
3	*Регрессионная статистика*								
4	Множественный R	0,902118							
5	R-квадрат	0,813817							
6	Нормированный R-квадрат	0,78722							
7	Стандартная ошибка	3,351131							
8	Наблюдения	9							
9									
10	Дисперсионный анализ								
11		*df*	*SS*	*MS*	*F*	*Значимость F*			
12	Регрессия	1	343,6116538	343,6116538	30,5974327	0,000876851			
13	Остаток	7	78,61056838	11,2300812					
14	Итого	8	422,2222222						
15									
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
17	Y-пересечение	2,658526	4,26807548	0,622886303	0,55310271	-7,43386903	12,75092	-7,43387	12,75092
18	x	1,171847	0,21185002	5,531494619	0,00087685	0,670901551	1,672793	0,670902	1,672793
19									

Рис. 2.9 – Результаты регрессионного анализа

8. Расчеты для определения прогнозного значения будем проводить на отдельном листе (рис. 2.10):



	A	B	C	D	E	F	G	H	I
1	**Прогнозирование по линейной модели**								
2		*x*	*y*	$(x-x_{ср.})^2$	Прогнозное значение	Стандартная ошибка прогноза	Предельная ошибка прогноза	Нижняя граница	Верхняя граница
3	1	11	15	71,309	15,549	3,960	7,502	8,047	23,051
4	2	13	17	41,531	17,893	3,787	7,175	10,718	25,067
5	3	15	21	19,753	20,236	3,656	6,926	13,310	27,162
6	4	18	20	2,086	23,752	3,546	6,717	17,034	30,469
7	5	20	28	0,309	26,095	3,534	6,696	19,399	32,792
8	6	22	33	6,531	28,439	3,574	6,771	21,669	35,210
9	7	24	34	20,753	30,783	3,662	6,938	23,845	37,721
10	8	25	32	30,864	31,955	3,723	7,054	24,901	39,009
11	9	27	29	57,086	34,298	3,878	7,347	26,951	41,646
12		15,5	сумма	250,222					
13									
14									
15	**Вспомогательные данные**								
16	Параметр *a*		2,659						
17	Параметр *b*		1,172						
18	$x_{ср.}$		19,444						
19	Стандартная ошибка регрессии *S*		3,351						
20	Количество предприятий		9						
21	Уровень надежности		0,9						
22	Уровень значимости		0,1						
23	Число степеней свободы		7						
24	критическое значение t-статистик		1,894578605						
25									
26	**Завершающие данные**								
27	m_p		3,630						
28	Δ_p		6,877						
29	y_p		20,822						
30	Нижняя граица		13,945						
31	Верхняя граница		27,699						
32	δ_p		33,028						
33									

Рис. 2.10 – Прогнозирование на основании линейной модели

8.1. В диапазоне A2:C11 подготовим исходные данные.

8.2. В ячейку B12 запишем значение x_p=15,5, для которого необходимо спрогнозировать значение результата y_p.

8.3. Введем вспомогательные данные:

- в ячейках C16 и C17 определим коэффициенты регрессии a и b;
- в ячейке C18 рассчитаем среднее значение $\overline{x}$;
- в ячейке C19 найдем стандартную ошибку регрессии S с помощию функции СТОШХУ;
- в ячейке C20 укажем число наблюдений (количество предприятий) n;
- в ячейке C21 зададим уровень надежности $\gamma = 0{,}9$;
- в ячейке C22 рассчитаем уровень значимости, равный $1 - \gamma$;
- в ячейке C23 укажем число степеней свободы $n - 2$;
- в ячейке C24 с помощью функции СТЬЮДРАСПОБР рассчитаем критическое значение t-статистики;

8.4. В ячейках D3: D11 найдем квадраты отклонений фактора x от его среднего значения $(x - \overline{x})^2$ и просуммируем в ячейке D12;

8.5. В ячейке C27 рассчитаем стандартную ошибку прогноза по формуле:

$$m_p = S\sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{\sum(x_p - \overline{x})^2}};$$

8.6. В ячейке C28 определим предельную ишибку прогноза по формуле $\Delta_p = t_{\text{табл}} \cdot m_p$;

8.7. В ячейке C29 рассчитаем прогнозное значение y_p, для этого подставим значение x_p в уравнение регрессии;

8.8. В ячейках C30 и C31 определим нижнюю и верхнюю границы доверительного интервала по формулам: $\gamma_p = y_p \pm \Delta_p$;

8.9. В ячейке C32 рассчитаем относитеную ошибку прогноза по формуле:

$$\delta_p = \frac{\Delta_p}{|y_p|} \cdot 100\%;$$

8.10. Подготовим данные для построения графика, для этого рассчитаем для всех значений x_i:

- в ячейках E3:E11 – оцененные значения $\hat{y}_i$;
- в ячейках F3:F11 – стандартные ошибки прогноза;
- в ячейках G3:G11 – предельные ошибки прогноза;
- в ячейках H3:H11 и I3:I11 – значения нижней и верхней границы доверительного интревала.

8.11. Нанесем данные на график:

- выделим одновременно диапазоны B2:C11, E2:E11, H2:I11 (поскольку эти диапазоны несмежные, при этом должна быть нажата клавиша Ctrl);
- перейдем к вкладке Вставка и в группе Диаграммы выберем тип Точечная\Точечная с маркерами;
- отформатируем диаграмму: добавим название, подписи осей, изменим параметры рядов (для этого выделим нужный ряд, вызовем конткстное меню и выберем пункт Формат ряда данных; другой вариант – перейти к закладке Макет, в группе Текущий фрагмент выбрать нужный ряд из ракрывающегося списка, нажать кнопку Формат выделенного);

8.12. Добавим на диаграмму прогнозируемое значение:

- вызовем контекстное меню, щелкнув правой кнопкой мыши по облати диаграммы, выберем строку Выбрать данные;
- в окне Выбор источника данных нажмем кнопку Добавить;
- в окне Изменение ряда заполним поля: Имя − Прогноз, Значения X – B12, Значения Y – C29.
- нажмем кнопку ОК;
- изменим параметры нового ряда.

7. Контрольные вопросы

1. Перечислите показатели, используемые для оценки качества регрессии?
2. Как оценивается значимость параметров уравнения регрессии?
3. С помощью какой функции Excel можно рассчитать критическое значение $F_{табл}$?
4. Как определить число степеней свободы для уравнения парной линейной регрессии?
5. Что такое доверительный интервал? Как рассчитывается доверительный интервал для параметров регрессии?
6. Каким образом осуществляется проверка значимости уравнения в целом?
7. Каким образом осуществляется проверка качества уравнения регрессии?
8. В чем смысл средней ошибки аппроксимации? Каковы допустимые значения ошибки аппроксимации?
9. Для каких целей используется инструмент анализа данных Excel «Регрессия»? В чем удобство этого инструмента?

10. Каким образом осуществляется прогнозирование с помощью уравнения регрессии?
11. Как оценить точность прогноза?
12. С помощью какой фунции можно определить стандартную ошибку регрессии? Каков ее синтаксис и параметры?

8. Требования к содержанию отчета

Отчет к лабораторной работе предоставляется в электронном виде и должен содержать:

- название и цель работы;
- номер и исходные данные своего варианта;
- описание хода выполнения заданий, в том числе:
  - скриншоты из MS Excel, отображающие заданные в ячейках формулы и функции, использованные для вычислений;
  - скриншоты из MS Excel с полученными при расчетах результатами;
  - анализ и интерпретация полученных результатов;
- выводы по лабораторной работе.

9. Требования к оформлению отчета

- Все рисунки в отчете должны быть подписаны.
- Все скриншоты должны быть читаемыми в масштабе документа 100%. При необходимости используйте обрезку и пропорциональное изменение размера рисунка.
- Все скриншоты таблиц MS Excel должны содержать системное наименование строк (1, 2, 3….) и столбцов (A, B, C, …)

Лабораторная работа №3
Нелинейные модели

1. Цель и задачи лабораторной работы

Цель работы: изучить возможности MSExcelдля построения нелинейных моделей.

Задачи:

– научиться приводить нелинейные уравнения к линейному виду средствами MSExcel;

– научиться определять индекс корреляции и коэффициент детерминации для нелинейных моделей средствами MSExcel;

– научиться проводить сравнительный анализ уравнений регрессии различного вида.

2. Теоретическая часть

2.1. Виды нелинейных уравнений регрессии

Если между экономическими явлениями существуют нелинейные соотношения, то они выражаются с помощью соответствующих нелинейных функций.

Различают два класса нелинейных регрессий:

1. Регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам, например:

– полиномы различных степеней:$\hat{y}_x = a + b \cdot x + c \cdot x^2$;

– равносторонняя гипербо□ла: $\hat{y}_x = a + b/x$;

– полулогарифмическая функция: $\hat{y}_x = a + b \cdot \ln x$

2. Регрессии, нелинейные по оцениваемым параметрам, например:

– степенная:$\hat{y}_x = a \cdot x^b$;

– показательная: $\hat{y}_x = a \cdot b^x$

– экспоненциальная: $\hat{y}_x = a \cdot e^{bx}$.

Регрессии нелинейные по включенным переменным приводятся клинейному виду простой заменой переменных, а дальнейшая оценкапараметров производится с помощью метода наименьших квадратов.

Несколько иначе обстоит дело с регрессиями нелинейными пооцениваемым параметрам, которые делятся на два типа: нелинейныемодели внутренне линейные (приводятся к линейному виду с помощьюсоответствующих преобразований, например, логарифмированием) инелинейные модели внутренне нелинейные (к линейному виду неприводятся).

К внутренне линейным моделям относятся, например, степеннаяфункция $\hat{y}_x = a \cdot x^b$, показательная $\hat{y}_x = a \cdot b^x$, экспоненциальная $\hat{y}_x = a \cdot e^{bx}$, обратная$\hat{y}_x = \frac{1}{a+b \cdot x}$.

К внутренне нелинейным моделям можно, например, отнести следующие модели: $\hat{y}_x = a + b \cdot x^c, \hat{y}_x = a \cdot \left(1 - \frac{1}{1-x^b}\right)$.

Приведем формулы для расчета параметров наиболее часто используемых типов уравнений регрессии (табл. 3.1):

Таблица 3.1. Формулы для расчета параметров нелинейных уравнений

Вид функции, y	Линеаризация	Параметры уравнения	Искомое уравнение
Степенная $\hat{y}_x = a \cdot x^b$	$X = \ln x, Y = \ln y,$ $A = \ln a, B = b$	$b = \frac{\overline{YX} - \overline{Y} \cdot \overline{X}}{\overline{X^2} - (\overline{X})^2}$ $A = \overline{Y} - b \cdot \overline{X}$	$y = e^A \cdot x^b$
Показательная $\hat{y}_x = a \cdot b^x$	$X = x, Y = \ln y,$ $A = \ln a, B = \ln b$	$B = \frac{\overline{Yx} - \overline{Y} \cdot \overline{x}}{\overline{x^2} - (\overline{x})^2}$ $A = \overline{Y} - B \cdot \overline{x}$	$y = (e^A) \cdot (e^B)^x$
Обратная $\hat{y}_x = \frac{1}{a + b \cdot x}$	$X = x, Y = 1/y,$ $A = a, B = b$	$b = \frac{\overline{Yx} - \overline{Y} \cdot \overline{x}}{\overline{x^2} - (\overline{x})^2}$ $A = \overline{Y} - b \cdot \overline{x}$	$y = \frac{1}{a + b \cdot x}$
Полулогарифмическая $\hat{y}_x = a + b \cdot \ln x$	$X = \ln x, Y = y,$ $A = a, B = b$	$b = \frac{\overline{yX} - \overline{y} \cdot \overline{X}}{\overline{X^2} - (\overline{X})^2}$ $a = \overline{y} - b \cdot \overline{X}$	$y = a + b \cdot \ln x$
Равносторонняя гипербо□ла $\hat{y}_x = a + b/x$	$X = 1/x, Y = y,$ $A = a, B = b$	$b = \frac{\overline{yX} - \overline{y} \cdot \overline{X}}{\overline{X^2} - (\overline{X})^2}$ $a = \overline{y} - b \cdot \overline{X}$	$y = a + b/x$
Экспоненциальная $\hat{y}_x = a \cdot e^{bx}$	$X = x, Y = \ln y,$ $A = \ln a, B = b$	$b = \frac{\overline{Yx} - \overline{Y} \cdot \overline{x}}{\overline{x^2} - (\overline{x})^2}$ $A = \overline{Y} - b \cdot \overline{x}$	$y = e^A \cdot e^{bx}$

2.2. Индекс корреляции и коэффициент детерминации

В случае нелинейной зависимости тесноту связи между величинами оценивают по величине корреляционного отношения (индекс корреляции):

$$\rho_{xy} = \sqrt{1 - \frac{\sum(y - y_x)^2}{\sum(y - \overline{y})^2}} = \sqrt{1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2_y}}.$$

Интервал изменения корреляционного отношения: $0 \leq \rho_{xy} \leq 1$.

Оценку качества построенной модели дает индекс детерминации ρ^2_{xy}.

Коэффициент детерминации $R^2 = \rho^2_{xy}$– квадрат индекса корреляции – характеризует долю дисперсии, объясняемую регрессией, в общей дисперсии результативного признака *y*.

$$R^2 = 1 - \frac{\sum(y - y_x)^2}{\sum(y - \overline{y})^2}$$

Чем ближе коэффициент детерминации к 1, тем выше качество уравнения регрессии, тем в большей мере оно объясняет поведение отклика.

3. Описание оборудования и используемых программных комплексов

При выполнении лабораторной работы необходим специализированный компьютерный класс с минимальными системными требованиями компьютеров:

- Процессор – IntelPentium IV;
- ОЗУ – 500Mb;
- видеокарта – 64Mb.
- Требуемое программное обеспечение:
- Операционная система MicrosoftWindows;
- MicrosoftExcel версии 2007 и выше.

4. Краткое руководство по эксплуатации оборудования

При использовании оборудования необходимо:

- соблюдать общие правила нахождения в учебных лабораториях, работы с компьютером и использования программных средств;
- осмотреть рабочее место, убрать все мешающие работе предметы;
- визуально проверить правильность подключения ПЭВМ к электросети.

5. Задание

Используя данные лабораторной работы №1, построить линейную, степенную, показательную, экспоненциальную, полулогарифмическую, гиперболическую и обратную модели и с помощью коэффициента детерминации и средней ошибки аппроксимации сравнить эти модели. Для уравнения каждого вида необходимо:

1. Найти уравнение регрессии.

2. Найти параметры регрессии с помощью расчетных формул из таблицы 3.1 и статистической функцииЛИНЕЙН.
3. Найти общую сумму квадратов отклонений и остаточную суммуквадратов отклонений.
4. Найти коэффициент детерминации.
5. Найти среднюю ошибку аппроксимации.
6. Нанести полученные значения на график.

6. Методика выполнения заданий

Создадим новую рабочую книгу с восемью листами: линейная, степенная, показательная, обратная, полулогарифмическая, гиперболическая, экспоненциальная, сравнение.

Будем использовать данные из лабораторной работы №1.

6.1. Линейная регрессия.

1. Оформим лист Линейная, как показано на рис. 3.1.

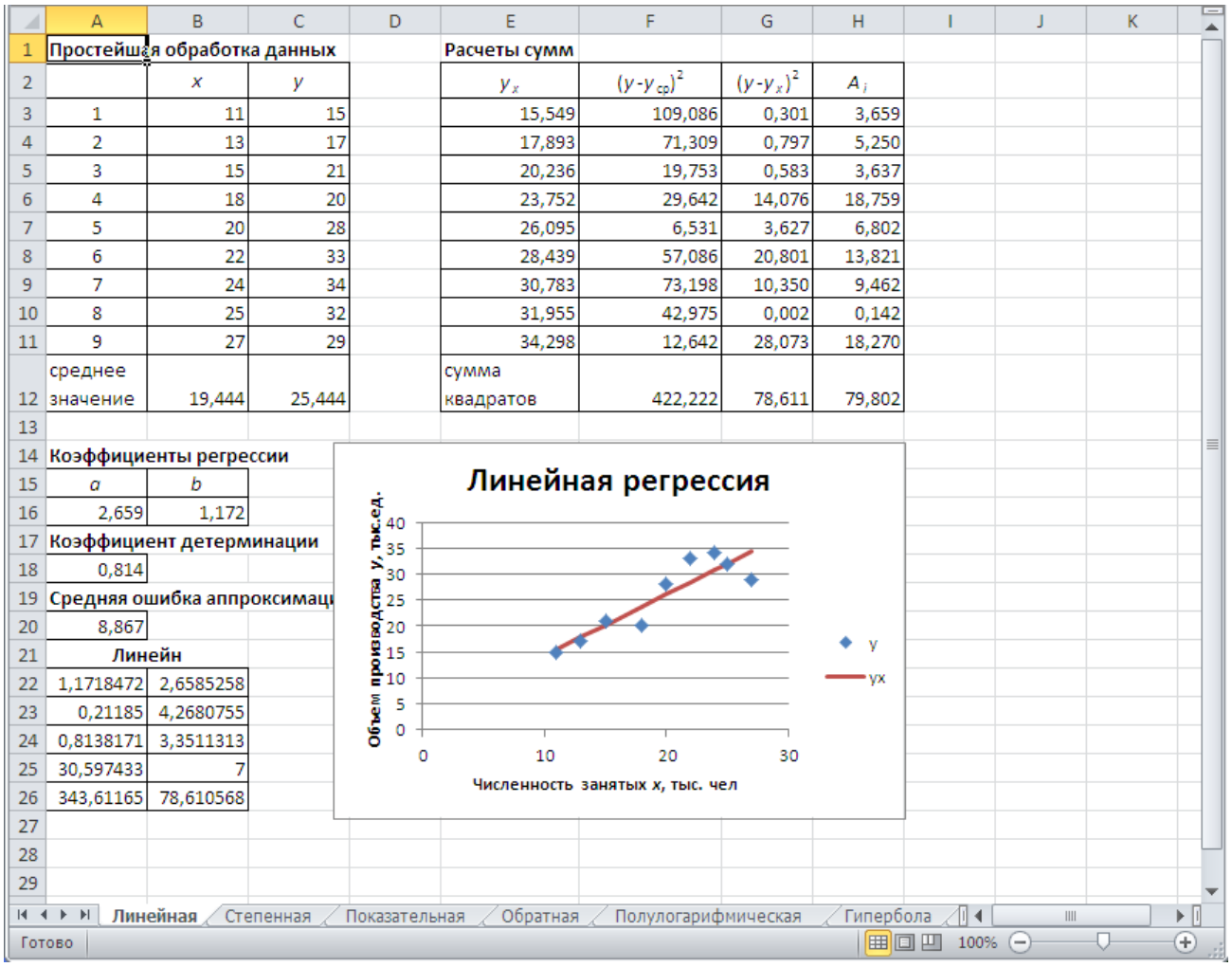


Рис. 3.1 – Лист «Линейная»

2. В ячейках A16 и B16 с помощью статистических функций ОТРЕЗОК и НАКЛОН определим коэффициенты линейной регрессииa и b. Получим уравнение линейной регрессии $\hat{y} = 0{,}506 + 0{,}919 \cdot x$.

3. В ячейках E3:E11 рассчитаем теоретические значение эндогенной переменной – $\hat{y}_i$, подставив значения x в найденное уравнение.

4. Рассчитаем вспомогательные данные для определения коэффициента детерминации и средней ошибки аппроксимации:

– в ячейках F3:F11 рассчитаем квадраты отклонений эндогенной переменной от ее среднего значения и просуммируем в ячейкеF12;

– в ячейках G3:G11рассчитаем квадраты отклонений эндогенной переменной от ее теоретического значения и просуммируем в ячейкеG12;

– в ячейках H3:H11рассчитаем ошибку аппроксимации по формуле $A_i = \left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\%$и просуммируем в ячейкеH12.

5. В ячейке A18, используя вспомогательные данные, определим коэффициент детерминации по формуле $R^2 = 1 - \frac{\sum(y - y_x)^2}{\sum(y - \overline{y})^2}$.

6. В ячейке A20, используя вспомогательные данные, определим среднюю ошибку аппроксимации по формуле$\overline{A} = \frac{1}{n}\sum\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\%$.

7. Нанесем исходные данные и полученные значения на диаграмму и отформатируем ее как это показано на рис. 3.1.

6.2. Степенная регрессия.

Регрессия в виде степенной функции имеет вид: $y = ax^b$.

Для нахождения параметров регрессии $y = ax^b$ необходимо провести ее линеаризацию:

– прологарифмируем левую и правую части равенства $y = ax^b$, получим$\ln y = \ln(ax^b) = \ln a + b \cdot \ln x$;

– сделаем замены$Y = \ln y$, $X = \ln x$, $A = \ln a$, тогда уравнение примет линейный вид: $Y = A + bX$.

1. Составляем вспомогательную таблицу для преобразованных данных (рис. 3.2):

– в ячейках D3:D11 и E3:E11 рассчитаем значения $X = \ln x$ и $Y = \ln y$ соответственно, определим их средние в ячейках D12 и E12;

– в ячейках F3:F11 найдем произведения $X \cdot Y$ и определим их среднее в ячейкеF12;

– в ячейках G3:G11 возведем значения X в квадрат и определим их среднее в ячейке G12;

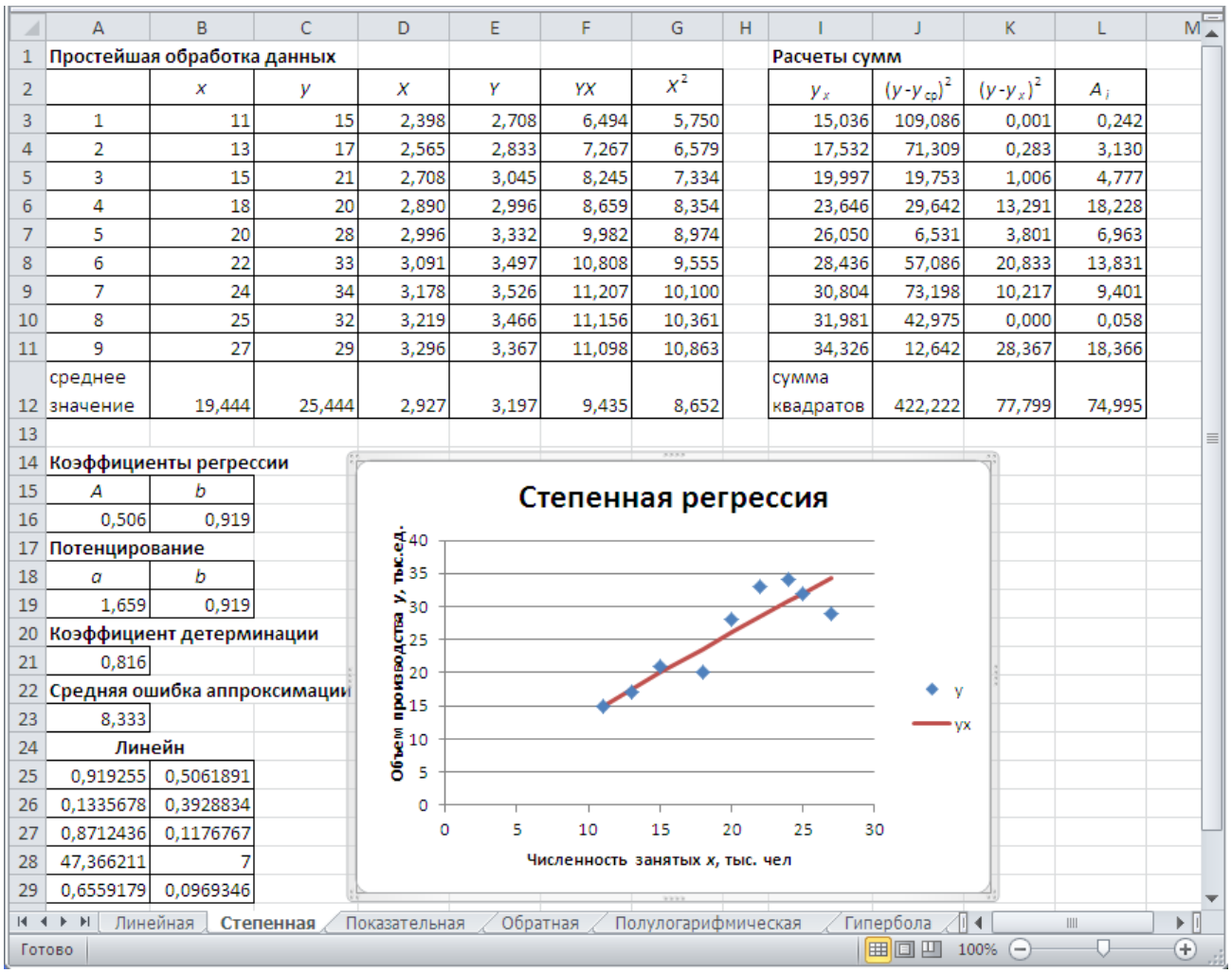


Рис. 3.2 – Лист «Степенная»

2. В ячейках A16 и B16 определим значение параметров линейного уравнения $Y = A + bX$ по формулам: $b = \frac{\overline{YX} - \overline{Y} \cdot \overline{X}}{\overline{X^2} - (\overline{X})^2}$; $A = \overline{Y} - b \cdot \overline{X}$. Проверим правильность расчетов с помощью статистическойфункции ЛИНЕЙН. Получим линейное уравнение следующего вида: $Y = 0{,}506 + 0{,}919 \cdot X$.

3. Чтобы найти параметры степенного уравнения, необходимо провести потенцирование (процедура, обратная логарифмированию):

– из $A = \ln a$ следует, что $a = e^A$. Рассчитаем это значение в ячейке A19, используя функцию Excel EXP (возвращает экспоненту заданного числа); получаем $a = e^{0{,}506} = 1{,}659$;

– параметр $b = 0{,}919$ остается без изменений;

– таким образом, степенное уравнение будет иметь вид: $\hat{y} = 1{,}659 \cdot x^{0{,}919}$.

4. В ячейках I3:I11 рассчитаем теоретические значение эндогенной переменной – $\hat{y}_i$, подставив значения x в найденное уравнение.

4. В ячейках J3:L12 рассчитаем вспомогательные данные для определения коэффициента детерминации и средней ошибки аппроксимации (аналогично линейной регрессии).

5. В ячейкахA21 иA23аналогично линейной регрессии определим коэффициент детерминации среднюю ошибку аппроксимации (аналогично линейной регрессии).

7. Нанесем исходные данные и полученные значения на диаграмму (рис. 3.2).

6.3. Сравнение результатов.

1. Расчеты на остальных листах во многом повторяют расчеты, произведенные на листе Степенная, поэтому остальные листы лучшевсего получить копированием листа Степенная.

Для этого необходимо:

– Находясь на листе Степенная, выделить его полностью,щелкнув мышью на пересечении названий столбцов и строк; с помощью кнопки Копировать скопировать лист в Буферобмена;

– Перейти на следующий лист и выделив ячейку A1, щелкнуть мышью по кнопкеВставить.

– Изменить значения в ячейках D3:E11, A19, B19,I3:I11 в соответствии с формулами линеаризации из табл. 3.1.

– Изменить заголовок и исходные данные диаграммы, щелкнув по ней правой кнопкой мыши и выбрав из контекстного меню пункт Выбрать данные. Далее в обоих рядах необходимо заменить диапазоны значений степенной функции на диапазоны значений текущего листа.

Получим следующие результаты (рис. 3.3-3.7).

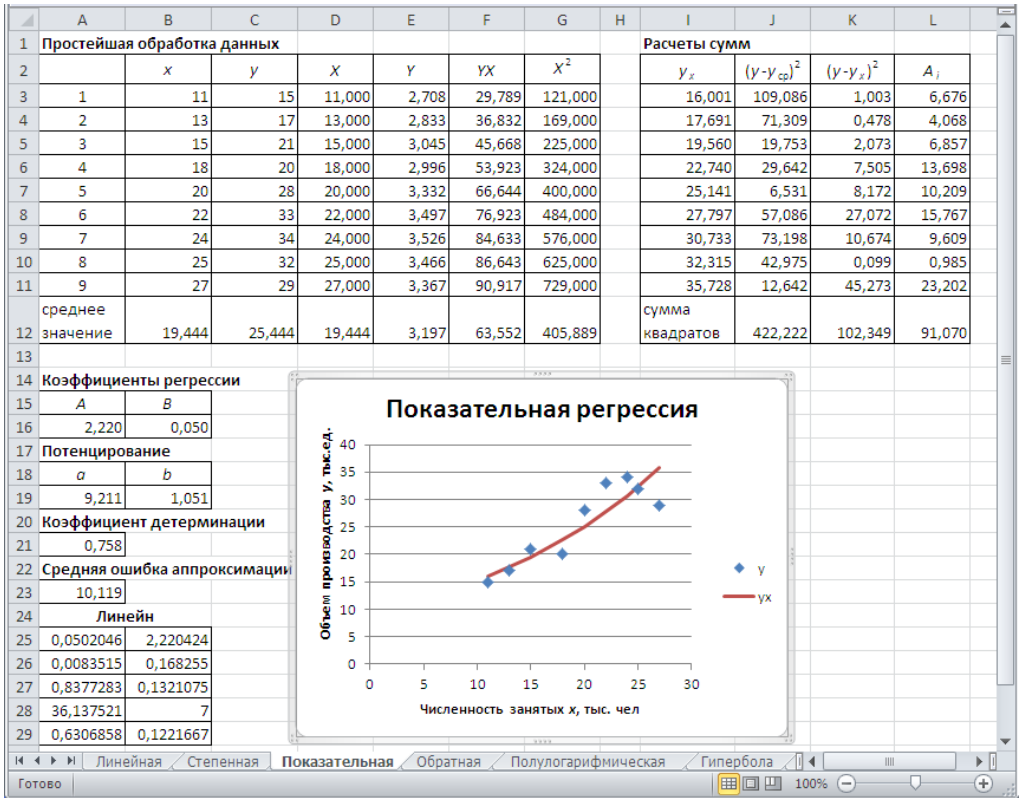
2. Выберем наилучшую модель, для чего объединим результатыпостроения парных регрессий в одной таблице (рис. 3.8).Для наглядности с помощью условного форматирования выделим:

– зеленым цветом максимальный коэффициент детерминации,

– зеленым цветом наименьшую среднюю ошибку аппроксимации,

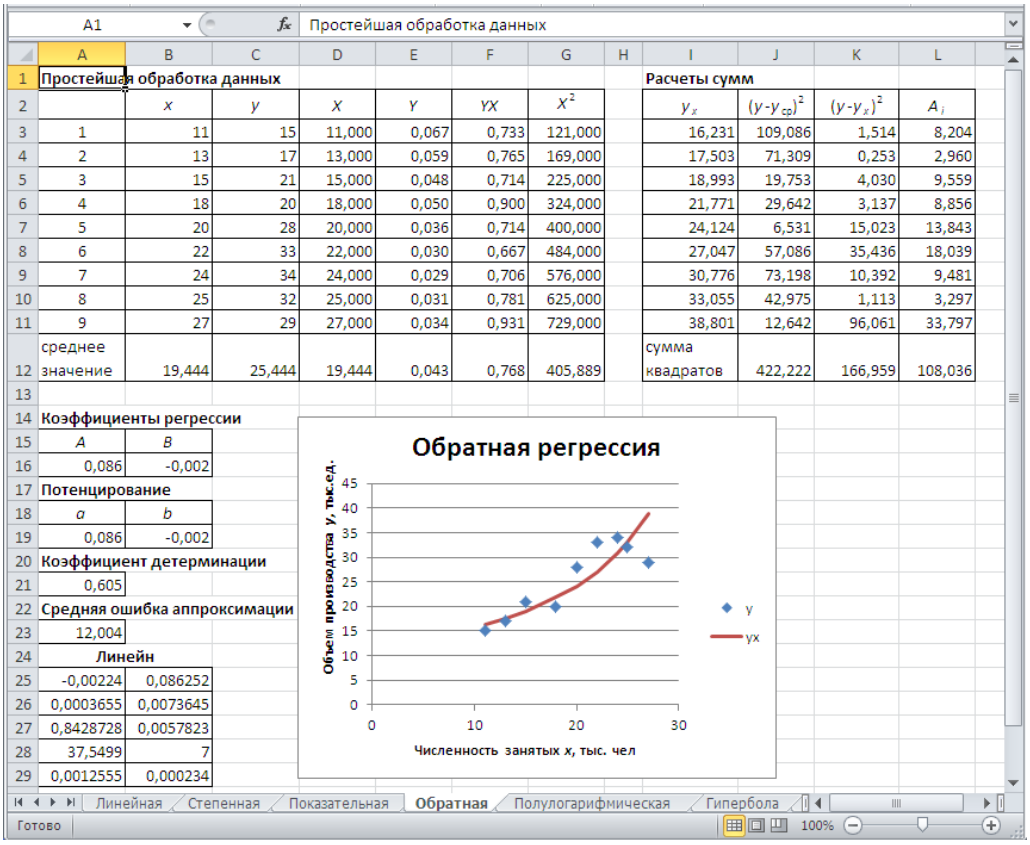
– красным цветом уравнения, для которых средняя ошибка аппроксимации превышает допустимый порог 10%.

Согласно коэффициенту детерминации все уравнения регрессиидостаточно хорошо описывают исходные данные. Однако для показательной, обратной и экспоненциальной регрессии средняя ошибка аппроксимации превышает 10%, а их коэффициенты детерминации наименьшие по сравнению с остальными функциями. Некоторое предпочтение можно отдать степенной (наименьшая средняя ошибка аппроксимации) или полулогарифмической (наибольший коэффициент детерминации) функциям.



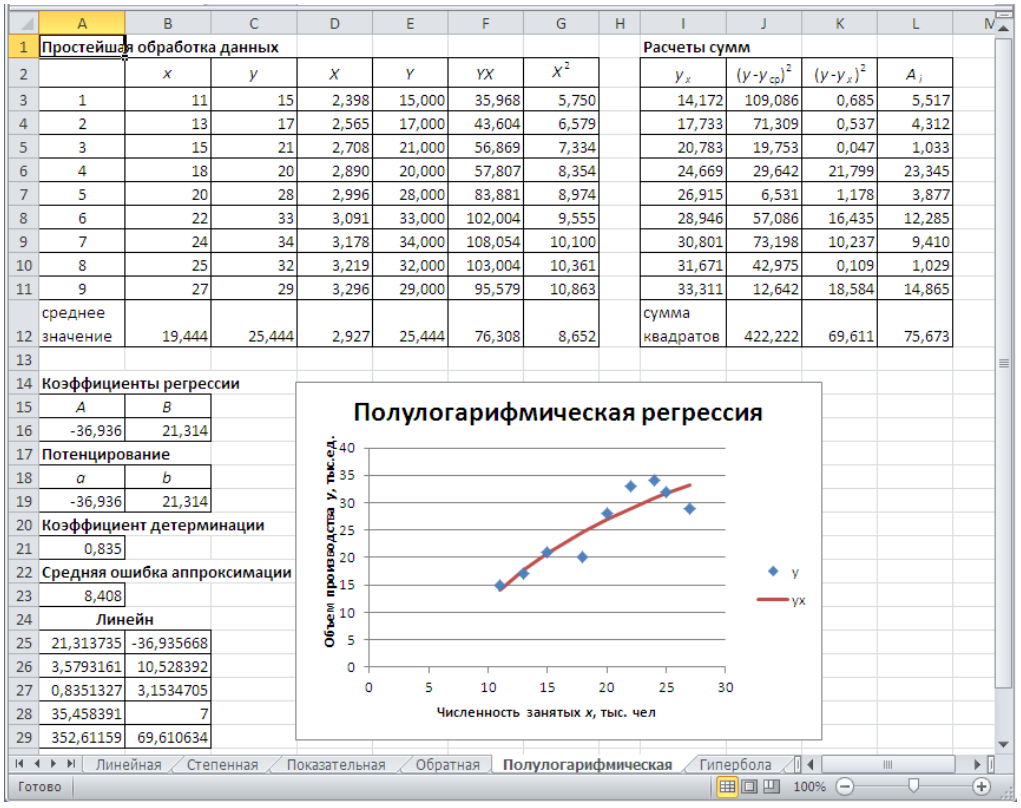
	A	B	C	D	E	F	G	H	I	J	K	L
1	Простейшая обработка данных								Расчеты сумм			
2		x	y	X	Y	YX	X^2		y_x	$(y-y_{ср})^2$	$(y-y_x)^2$	A_i
3	1	11	15	11,000	2,708	29,789	121,000		16,001	109,086	1,003	6,676
4	2	13	17	13,000	2,833	36,832	169,000		17,691	71,309	0,478	4,068
5	3	15	21	15,000	3,045	45,668	225,000		19,560	19,753	2,073	6,857
6	4	18	20	18,000	2,996	53,923	324,000		22,740	29,642	7,505	13,698
7	5	20	28	20,000	3,332	66,644	400,000		25,141	6,531	8,172	10,209
8	6	22	33	22,000	3,497	76,923	484,000		27,797	57,086	27,072	15,767
9	7	24	34	24,000	3,526	84,633	576,000		30,733	73,198	10,674	9,609
10	8	25	32	25,000	3,466	86,643	625,000		32,315	42,975	0,099	0,985
11	9	27	29	27,000	3,367	90,917	729,000		35,728	12,642	45,273	23,202
12	среднее значение	19,444	25,444	19,444	3,197	63,552	405,889		сумма квадратов	422,222	102,349	91,070
13												
14	Коэффициенты регрессии											
15	A	B										
16	2,220	0,050										
17	Потенцирование											
18	a	b										
19	9,211	1,051										
20	Коэффициент детерминации											
21	0,758											
22	Средняя ошибка аппроксимации											
23	10,119											
24	Линейн											
25	0,0502046	2,220424										
26	0,0083515	0,168255										
27	0,8377283	0,1321075										
28	36,137521	7										
29	0,6306858	0,1221667										

Рис. 3.3 – Лист «Показательная»



	A	B	C	D	E	F	G	H	I	J	K	L
1	Простейшая обработка данных								Расчеты сумм			
2		x	y	X	Y	YX	X^2		y_x	$(y-y_{ср})^2$	$(y-y_x)^2$	A_i
3	1	11	15	11,000	0,067	0,733	121,000		16,231	109,086	1,514	8,204
4	2	13	17	13,000	0,059	0,765	169,000		17,503	71,309	0,253	2,960
5	3	15	21	15,000	0,048	0,714	225,000		18,993	19,753	4,030	9,559
6	4	18	20	18,000	0,050	0,900	324,000		21,771	29,642	3,137	8,856
7	5	20	28	20,000	0,036	0,714	400,000		24,124	6,531	15,023	13,843
8	6	22	33	22,000	0,030	0,667	484,000		27,047	57,086	35,436	18,039
9	7	24	34	24,000	0,029	0,706	576,000		30,776	73,198	10,392	9,481
10	8	25	32	25,000	0,031	0,781	625,000		33,055	42,975	1,113	3,297
11	9	27	29	27,000	0,034	0,931	729,000		38,801	12,642	96,061	33,797
12	среднее значение	19,444	25,444	19,444	0,043	0,768	405,889		сумма квадратов	422,222	166,959	108,036
13												
14	Коэффициенты регрессии											
15	A	B										
16	0,086	-0,002										
17	Потенцирование											
18	a	b										
19	0,086	-0,002										
20	Коэффициент детерминации											
21	0,605											
22	Средняя ошибка аппроксимации											
23	12,004											
24	Линейн											
25	-0,00224	0,086252										
26	0,0003655	0,0073645										
27	0,8428728	0,0057823										
28	37,5499	7										
29	0,0012555	0,000234										

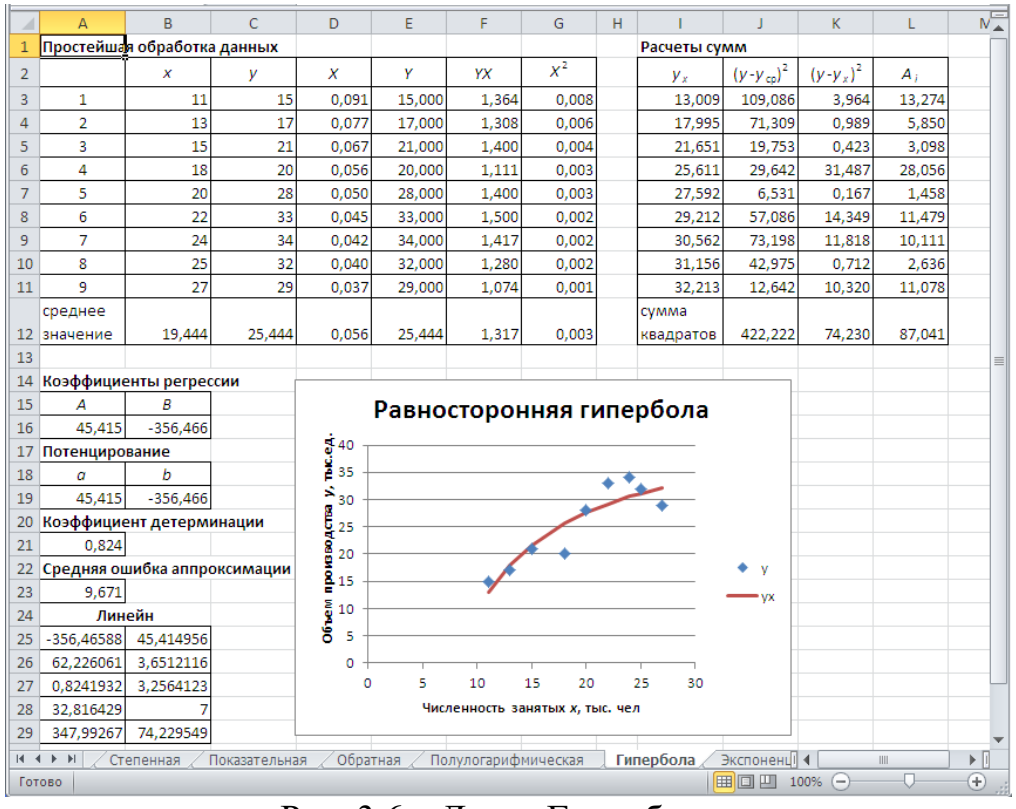
Рис. 3.4 – Лист «Обратная»



	A	B	C	D	E	F	G	H	I	J	K	L
1	Простейшая обработка данных								Расчеты сумм			
2		x	y	X	Y	YX	X^2		y_x	$(y-y_{ср})^2$	$(y-y_x)^2$	A_i
3	1	11	15	2,398	15,000	35,968	5,750		14,172	109,086	0,685	5,517
4	2	13	17	2,565	17,000	43,604	6,579		17,733	71,309	0,537	4,312
5	3	15	21	2,708	21,000	56,869	7,334		20,783	19,753	0,047	1,033
6	4	18	20	2,890	20,000	57,807	8,354		24,669	29,642	21,799	23,345
7	5	20	28	2,996	28,000	83,881	8,974		26,915	6,531	1,178	3,877
8	6	22	33	3,091	33,000	102,004	9,555		28,946	57,086	16,435	12,285
9	7	24	34	3,178	34,000	108,054	10,100		30,801	73,198	10,237	9,410
10	8	25	32	3,219	32,000	103,004	10,361		31,671	42,975	0,109	1,029
11	9	27	29	3,296	29,000	95,579	10,863		33,311	12,642	18,584	14,865
12	среднее значение	19,444	25,444	2,927	25,444	76,308	8,652		сумма квадратов	422,222	69,611	75,673

	A	B
13		
14	Коэффициенты регрессии	
15	A	B
16	-36,936	21,314
17	Потенцирование	
18	a	b
19	-36,936	21,314
20	Коэффициент детерминации	
21	0,835	
22	Средняя ошибка аппроксимации	
23	8,408	
24	Линейн	
25	21,313735	-36,935668
26	3,5793161	10,528392
27	0,8351327	3,1534705
28	35,458391	7
29	352,61159	69,610634

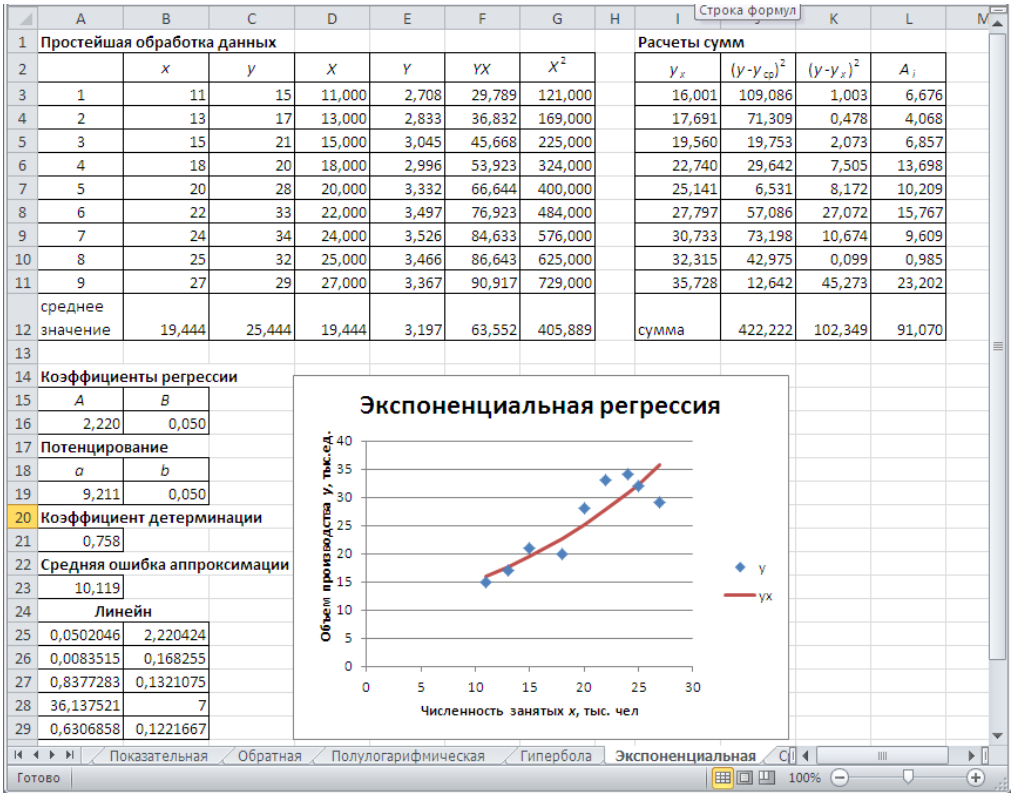
Рис. 3.5 – Лист «Полулогарифмическая»



	A	B	C	D	E	F	G	H	I	J	K	L
1	Простейшая обработка данных								Расчеты сумм			
2		x	y	X	Y	YX	X^2		y_x	$(y-y_{ср})^2$	$(y-y_x)^2$	A_i
3	1	11	15	0,091	15,000	1,364	0,008		13,009	109,086	3,964	13,274
4	2	13	17	0,077	17,000	1,308	0,006		17,995	71,309	0,989	5,850
5	3	15	21	0,067	21,000	1,400	0,004		21,651	19,753	0,423	3,098
6	4	18	20	0,056	20,000	1,111	0,003		25,611	29,642	31,487	28,056
7	5	20	28	0,050	28,000	1,400	0,003		27,592	6,531	0,167	1,458
8	6	22	33	0,045	33,000	1,500	0,002		29,212	57,086	14,349	11,479
9	7	24	34	0,042	34,000	1,417	0,002		30,562	73,198	11,818	10,111
10	8	25	32	0,040	32,000	1,280	0,002		31,156	42,975	0,712	2,636
11	9	27	29	0,037	29,000	1,074	0,001		32,213	12,642	10,320	11,078
12	среднее значение	19,444	25,444	0,056	25,444	1,317	0,003		сумма квадратов	422,222	74,230	87,041

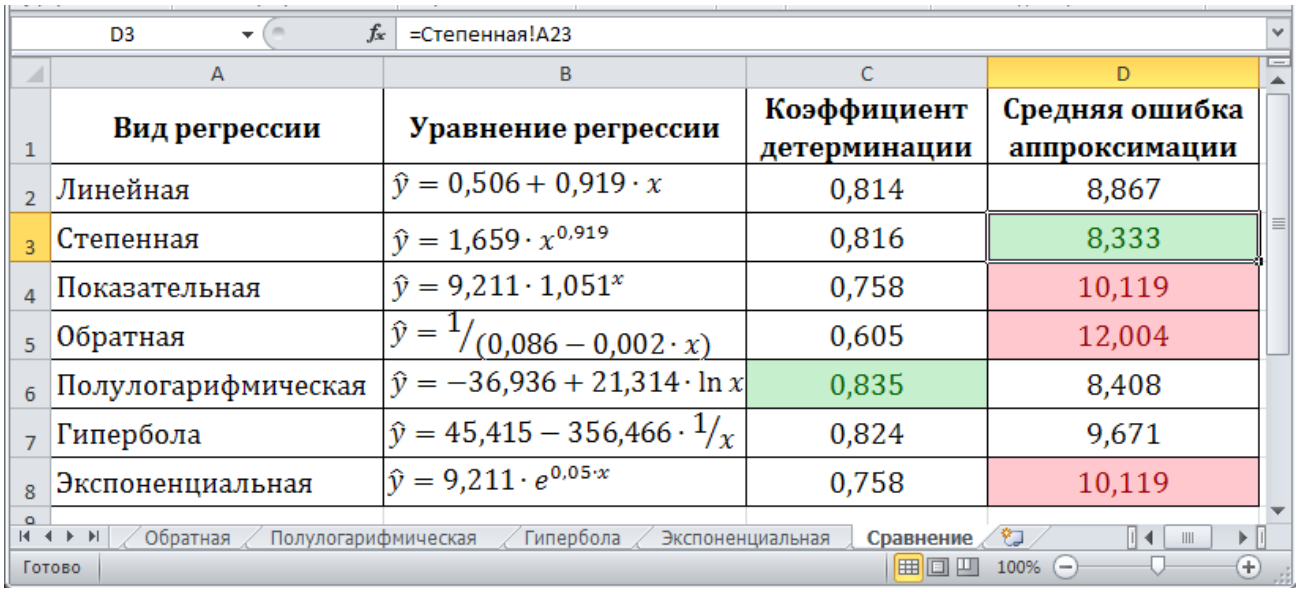
	A	B
13		
14	Коэффициенты регрессии	
15	A	B
16	45,415	-356,466
17	Потенцирование	
18	a	b
19	45,415	-356,466
20	Коэффициент детерминации	
21	0,824	
22	Средняя ошибка аппроксимации	
23	9,671	
24	Линейн	
25	-356,46588	45,414956
26	62,226061	3,6512116
27	0,8241932	3,2564123
28	32,816429	7
29	347,99267	74,229549

Рис. 3.6 – Лист «Гипербола»



	A	B	C	D	E	F	G	H	I	K	L	M
1	Простейшая обработка данных								Расчеты сумм			
2		x	y	X	Y	YX	X^2		y_x	$(y-y_{ср})^2$	$(y-y_x)^2$	A_i
3	1	11	15	11,000	2,708	29,789	121,000		16,001	109,086	1,003	6,676
4	2	13	17	13,000	2,833	36,832	169,000		17,691	71,309	0,478	4,068
5	3	15	21	15,000	3,045	45,668	225,000		19,560	19,753	2,073	6,857
6	4	18	20	18,000	2,996	53,923	324,000		22,740	29,642	7,505	13,698
7	5	20	28	20,000	3,332	66,644	400,000		25,141	6,531	8,172	10,209
8	6	22	33	22,000	3,497	76,923	484,000		27,797	57,086	27,072	15,767
9	7	24	34	24,000	3,526	84,633	576,000		30,733	73,198	10,674	9,609
10	8	25	32	25,000	3,466	86,643	625,000		32,315	42,975	0,099	0,985
11	9	27	29	27,000	3,367	90,917	729,000		35,728	12,642	45,273	23,202
12	среднее значение	19,444	25,444	19,444	3,197	63,552	405,889		сумма	422,222	102,349	91,070
13												
14	Коэффициенты регрессии											
15	A	B										
16	2,220	0,050										
17	Потенцирование											
18	a	b										
19	9,211	0,050										
20	Коэффициент детерминации											
21	0,758											
22	Средняя ошибка аппроксимации											
23	10,119											
24	Линейн											
25	0,0502046	2,220424										
26	0,0083515	0,168255										
27	0,8377283	0,1321075										
28	36,137521	7										
29	0,6306858	0,1221667										

Рис. 3.7 – Лист «Экспоненциальная»



D3 =Степенная!A23

	A	B	C	D
1	**Вид регрессии**	**Уравнение регрессии**	**Коэффициент детерминации**	**Средняя ошибка аппроксимации**
2	Линейная	$\hat{y} = 0{,}506 + 0{,}919 \cdot x$	0,814	8,867
3	Степенная	$\hat{y} = 1{,}659 \cdot x^{0{,}919}$	0,816	8,333
4	Показательная	$\hat{y} = 9{,}211 \cdot 1{,}051^{x}$	0,758	10,119
5	Обратная	$\hat{y} = {}^{1}/_{(0{,}086 - 0{,}002 \cdot x)}$	0,605	12,004
6	Полулогарифмическая	$\hat{y} = -36{,}936 + 21{,}314 \cdot \ln x$	0,835	8,408
7	Гипербола	$\hat{y} = 45{,}415 - 356{,}466 \cdot {}^{1}/_{x}$	0,824	9,671
8	Экспоненциальная	$\hat{y} = 9{,}211 \cdot e^{0{,}05 \cdot x}$	0,758	10,119

Рис. 3.8 – Сравнение уравнений регрессии разных видов

7. Контрольные вопросы

11. Приведите пример регрессий, нелинейных относительно включенных в анализ объясняющих переменных, но линейных по оцениваемым параметрам.
12. Приведите пример регрессий, нелинейных по оцениваемым параметрам.
13. Назовите типы уравнений регрессии, нелинейных по параметрам.
14. Как осуществляется линеаризация модели?
15. Приведите пример внутренне нелинейныхмоделей.
16. Каким образом измеряется теснота связи между величинамив случае нелинейной зависимости?
17. Каким образом связаны коэффициент детерминации и индекс корреляции?
18. Что такое потенцирование?
19. Каким образом можно выбрать наилучшую для исследуемых данных модель?
20. Что характеризует коэффициент детерминации?

Лабораторная работа №4
Построение многофакторной линейной регрессии с помощью пакета Анализ данных MS Excel. Анализ остатков.

1. Цель и задачи лабораторной работы

Цель работы: изучить возможности MS Excel для построения многофакторной линейной регрессии, оценки ее качества и анализа остатков.

Задачи:

– научиться проверять факторы на мультиколлинеарность;

– приобрести навыки нахождения параметров уравнения многофакторной линейной регрессии в естественной и стандартизованной форме;

– научиться ранжироватьфакторы по силе их воздействия на результат;

– научиться проверять качество уравнения многофакторной линейной регрессии;

– научиться проводить анализ остатков на выполнение пяти предпосылок метода наименьших квадратов.

2. Теоретическая часть

Множественная регрессия – уравнения связи с несколькими независимыми переменными:

$$y = f(x_1, x_2, \dots, x_m),$$

где y – зависимая переменная (результативный признак);

x_1, x_2,…, x_m – независимые переменные (факторы).

Множественная регрессия применяется в ситуациях, когда из множества факторов, влияющих на результативный признак, нельзя выделить один доминирующий фактор и необходимо учитывать влияние нескольких факторов.

Основная цель множественной регрессии – построить модель с большим числом факторов, определив при этом влияние каждого из них в отдельности, а также совокупное их воздействие на моделируемый показатель.

2.1. Отбор факторов при построении множественной регрессии

Включение в уравнение множественной регрессии того или иного набора факторов связано, прежде всего, с представлением исследователя о природе взаимосвязи моделируемого показателя с другими экономическими явлениями.

Факторы, включаемые во множественную регрессию, должны отвечать следующим требованиям:

1. Факторы должны быть количественно измеримы.

2. Факторы не должны быть взаимно коррелированы. Если между факторами существует высокая корреляция, то нельзя определить их

изолированное влияние на результативный показатель, и параметры уравнения регрессии оказываются неинтерпретируемыми.

Коэффициенты интеркорреляции (т. е. корреляции между объясняющими переменными) позволяют исключать из модели дублирующие факторы. Считается, что две переменные явно коллинеарные, т. е. находятся между собой в линейной зависимости, если $\left|r_{x_i x_j}\right| \geq 0{,}7$.

Если факторы явно коллинеарны, то они дублируют друг друга и один из них рекомендуется исключить из регрессии. Предпочтение при этом отдается тому фактору, который при достаточно тесной связи с результатом имеет наименьшую тесноту связи с другими факторами.

3. Факторы должны иметь заметную связь с результирующей переменной, т.е. $\left|r_{yx_j}\right| \geq 0{,}4$.

2.2. Оценка параметров уравнения множественной регрессии

Для оценки параметров уравнения множественной регрессии применяют метод наименьших квадратов (МНК). В результате мы получаем линейное уравнение регрессии в естественном виде:

$$y = b_0 + b_1 x_1 + \cdots + b_m x_m + u.$$

Для определения значимости факторов и повышения точности результата используется уравнение множественной регрессии в стандартизованном масштабе:

$$t_y = \beta_0 + \beta_1 t_{x_1} + \cdots + \beta_m t_{x_m} + u,$$

где $t_y,\ t_{x_1}, \ldots, t_{x_m}$ - стандартизованные переменные

$$t_y = \frac{y - \overline{y}}{\sigma_y}, t_{x_i} = \frac{x_i - \overline{x_i}}{\sigma_{x_i}},$$

для которых среднее значение равно нулю $\overline{t}_y = \overline{t}_{x_i} = 0$, а среднее квадратическое отклонение равно единице$\sigma_y = \sigma_{x_i} = 1$.

Величины β_i называются стандартизованными коэффициентами регрессии. Они показывают, на сколько сигм (средних квадратических отклонений) изменится в среднем результат, если соответствующий фактор x_i изменится на одну сигму при неизменном среднем уровне других факторов. В силу того, что все переменные заданы как центрированные и нормированные, стандартизованные коэффициенты регрессии β_i сравнимы между собой. Сравнивая их друг с другом, можно ранжировать факторы по силе их воздействия на результат. В этом основное достоинство стандартизованных коэффициентов регрессии в отличие от коэффициентов регрессии в естественном виде, которые несравнимы между собой.

Связь коэффициентов множественной регрессии b_i со стандартизованными коэффициентами β_i описывается соотношением

$$b_i = \beta_i \frac{\sigma_y}{\sigma_{x_i}}.$$

Параметр b_0 определяется из соотношения: $b_0 = \overline{y} - b_1\overline{x}_1 - \cdots - b_m\overline{x}_m$.

Средние коэффициенты эластичности для линейной множественной регрессии рассчитываются по формуле

$$\overline{Э}_{yx_j} = b_j \frac{\overline{x}_j}{\overline{y}}$$

и показывают, на сколько процентов в среднем по совокупности изменится результат *y* от своей величины при изменении фактора *x* на 1 % от своего значения при неизменных значениях других факторов.

2.3. Множественная корреляция

Практическая значимость уравнения множественной регрессии оценивается с помощью линейного коэффициента множественной корреляции и его квадрата – коэффициента детерминации.

1. *Линейный коэффициент множественной корреляции*оценивает тесноту факторов на результат и может быть рассчитан по формуле:

$$R_{yx_1x_2\ldots x_m} = \sqrt{1 - \frac{\sum\left(y - \hat{y}_{yx_1x_2\ldots x_m}\right)^2}{\sum(y - \overline{y})^2}}$$

Низкое значение коэффициента множественной корреляции означает, что в регрессионную модель не включены существенные факторы – с одной стороны, а с другой стороны – рассматриваемая форма связи не отражает реальные соотношения между переменными, включенными в модель.

В этом случае требуются дальнейшие исследования по улучшению качества модели и увеличению ее практической значимости.

2. *Коэффициент детерминации* – ещё один показатель качества подгонки. $0 \leq R^2 \leq 1$, чем ближе R^2 к 1, тем лучше регрессионное уравнение (т.е. качество подгонки).

3. *Скорректированный коэффициент детерминации*. В многофакторном регрессионном уравнении добавление дополнительных объясняющих увеличивает коэффициент детерминации. Следовательно, коэффициент детерминации должен быть скорректирован с учетом числа независимых переменных:

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

Чем больше величина *m*, тем сильнее различия $\overline{R}^2$ и R^2.

2.4. Оценка надежности результатов множественной регрессии и корреляции

Значимость уравнения множественной регрессии в целом, так же как и в парной регрессии, оценивается с помощью *F-критерия Фишера*. Он состоит в проверке гипотезы H_0 о статистической незначимости уравнения регрессии и показателя тесноты связи. Для этого выполняется сравнение фактического $F_{факт}$ и критического (табличного) $F_{табл}$ значений F-критерия Фишера. $F_{факт}$ определяется из соотношения значений факторной иостаточной дисперсий, рассчитанных на одну степень свободы:

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

Число степеней свободы критерия Фишера: $k_1=m$; $k_2=n-m-1$ и критическое значение этого критерия $F_{кр} = F_{\alpha;k_1;k_2}$

Если $F_{факт}>F_{табл}$, то гипотеза H_0 о случайной природе оцениваемых характеристик отклоняется и признается их статистическая значимость и надежность. Если $F_{факт}<F_{табл}$, то гипотеза H_0 не отклоняется и признается статистическая незначимость, ненадежность уравнения регрессии.

Оценка значимости коэффициентов регрессии производится с помощью *t-критерия Стьюдента*. Вычисляются наблюдаемые значения *t*-статистики:

$$t_j = \frac{b_j}{m_{b_j}}$$

где m_{b_j} – средняя квадратическая ошибка коэффициента регрессии b_j, она может быть определена по следующей формуле:

$$m_{b_j} = \frac{\sigma_y \cdot \sqrt{1-R^2_{yx_1\ldots x_m}}}{\sigma_{x_j} \cdot \sqrt{1-R^2_{x_j x_1\ldots x_m}}} \cdot \frac{1}{\sqrt{n-m-1}}$$

Критическое значение t-статистики: $t_{кр}= t_{\alpha;n-m-1}$ где $k = n-m-1$ – число степеней свободы, m – число факторов; Если $|t_{набл}| > t_{кр}$, то коэффициент регрессии статистически значим; в противном случае – статистически незначим.

2.5. Анализ остатков

Исследование остатков u_i предполагает проверку наличия следующих пяти предпосылок МНК:

1) Случайный характер остатков;
2) Нулевая средняя величина остатков, не зависящая от x_i;
3) Гомоскедастичность;
4) Отсутствие автокорреляции остатков;
5) Остатки подчиняются нормальному закону распределения.

Для проверки *первой предпосылки* – случайного характера остатков – строится график зависимости остатков u_i от теоретического значения результативного признака $\hat{y}_x$. Если на графике получена горизонтальная полоса (рис. 4.1а), внутри которой остатки расположены случайным образом, то первая предпосылка выполняется.

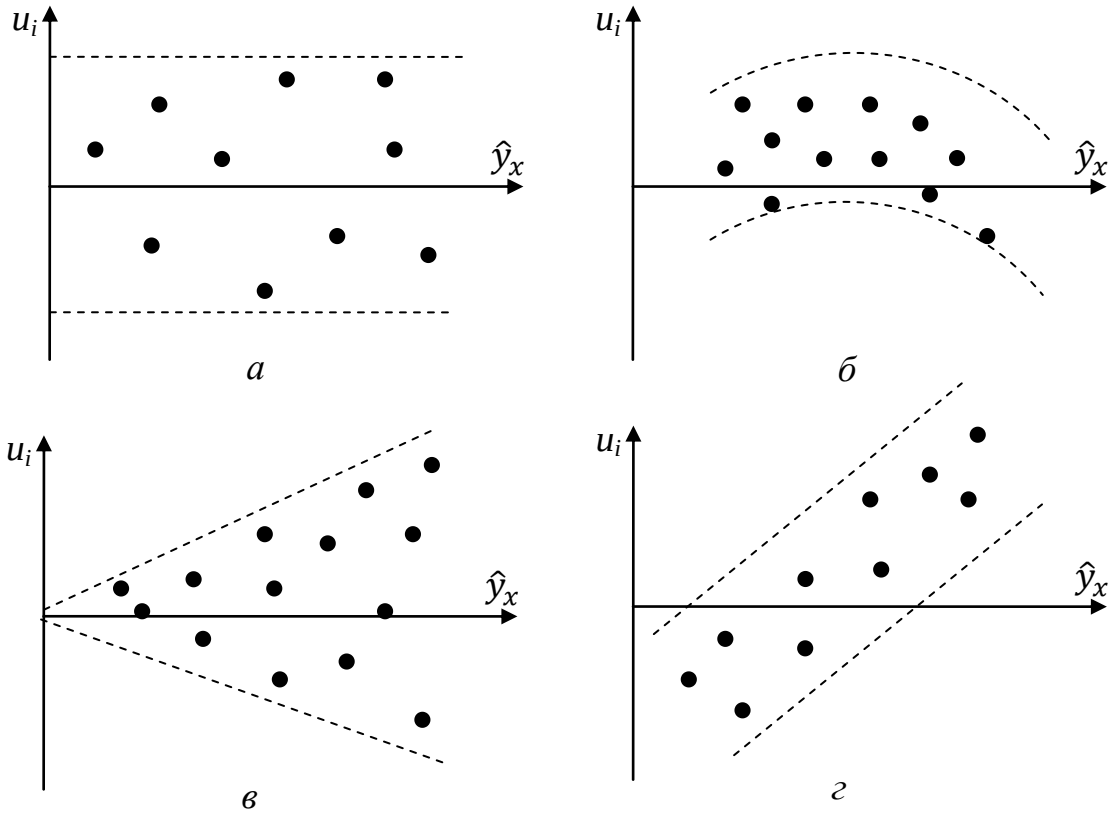


Рис. 4.1. Зависимость случайных остатков u_i от теоретических значений $\hat{y}_x$

Возможны следующие случаи: если u_i зависит от $\hat{y}_x$, то:

- Остатки u_i не случайны (рис. 4.1б);
- Остатки u_i не имеют постоянной дисперсии (рис. 4.1в);
- Остатки u_i носят систематический характер (рис. 4.1г).

В этих случаях необходимо либо применить другую функцию, либо вводить дополнительную информацию и заново строить уравнение регрессии.

Вторая предпосылка МНК заключается в равенстве нулю средних значений остатков и независимости их от факторов. Она обеспечивает несмещенность оценок. Для ее проверки строится график зависимости случайных остатков u_i от факторов, включенных в регрессию – x_i. Если на графике получена горизонтальная полоса, внутри которой остатки расположены случайным образом, то вторая предпосылка выполняется. Если же график показывает наличие зависимости u_i от x_i, то модель неадекватна.

В соответствии с *третьей предпосылкой МНК* требуется, чтобы дисперсия остатков была гомоскедастичной. Это означает, что для каждого значения фактора остатки имеют одинаковую дисперсию. Если это условие не соблюдается, то имеет место гетероскедастичность. Наличие гомоскедастичности или гетероскедастичности можно видеть по графику зависимостиостатков u_i от теоретического значения результативного признака $\hat{y}_x$. Так на рис. 4.1а остатки гомоскедастичны, а на рис. 4.1в – гетероскедастичны.

Четвертая предпосылка МНК – отсутствие автокорреляции остатков означает, что остатки u_i распределены независимо друг от друга. Автокорреляция остатков означает наличие корреляции между остатками текущих и предыдущих наблюдений. Для проверки выполнения четвертой предпосылки можно воспользоваться критерием Дарбина-Уотсона:

- Выдвигается гипотеза H_0 об отсутствии автокорреляции в остатках.
- Рассчитывается статистика DW по формуле:

$$DW = \frac{\sum_{i=2}^{n}(u_i - u_{i-1})^2}{\sum_{i=1}^{n} u_i^2}; 0 \le DW \le 4$$

- По специальным таблицам (см. Приложение 1) определяются критические значения Дарбина-Уотсона d_L и d_U для заданного числа наблюдений *n*, числа независимых переменных *m* и уровня значимости *α*.
- По найденным значениям разбиваем промежуток [0; 4] на 5 отрезков.

Есть положительная автокорреляция остатков (H_0 отклоняется)	Зона неопределен-ности	Автокорреляция остатков отсутствует (нет оснований отклонять H_0)	Зона неопределен-ности	Есть отрицательная автокорреляция остатков (H_0 отклоняется)

0 d_L d_U 2 $4-d_U$ $4-d_L$ 4

- Если расчетное значение попадает в зону неопределенности, то подтверждается существование автокорреляции и гипотезу H_0 отклоняют.

Для проверки *пятой предпосылки* о нормальном распределении остатков используют Q-Q график. Строятся квантили распределения остатков относительно нормального распределения (квантили – это значения, которые делят случаи на ряд групп одинакового размера). Если точки, соответствующие наблюдаемым данным, образуют прямую, проведенную из левого нижнего угла в правый верхний угол, значит, данные распределены приближенно нормально. С другой стороны, если эти точки отклоняются от прямой линии, распределение данных отличается от нормального.

Предпосылка о нормальном распределении остатков позволяет проводить проверку параметров регрессии и корреляции с помощью критериев *t*, *F*. Вместе

с тем оценки регрессии, найденные с применением МНК, обладают хорошими свойствами даже при отсутствии нормального распределения остатков, т.е. при нарушении пятой предпосылки.

3. Описание оборудования и используемых программных комплексов

При выполнении лабораторной работы необходим специализированный компьютерный класс с минимальными системными требованиями компьютеров:

- Процессор – Intel Pentium IV;
- ОЗУ – 500 Mb;
- видеокарта – 64 Mb.
- Требуемое программное обеспечение:
- Операционная система Microsoft Windows;
- Microsoft Excel версии 2007 и выше.

4. Краткое руководство по эксплуатации оборудования

При использовании оборудования необходимо:

- соблюдать общие правила нахождения в учебных лабораториях, работы с компьютером и использования программных средств;
- осмотреть рабочее место, убрать все мешающие работе предметы;
- визуально проверить правильность подключения ПЭВМ к электросети.

5. Задание

Для выданного варианта с помощью пакета Анализ данныхMS Excel построить модель множественной регрессии и проверить ее качество. Для этого необходимо:

1. Выбрать факторы для включения в модель:
    1) рассчитать коэффициенты парной корреляции (с помощью инструмента Корреляция):
        - между зависимой и независимыми переменными,
        - независимых переменных между собой;
    2) исключить незначимые факторы;
    3) исключить мультиколлинеарность;
2. Для выбранных факторов построить линейное уравнение множественной регрессии (с помощью инструмента Регрессия):
    1) определить коэффициенты уравнения множественной регрессии в естественной и стандартизованной форме;
    2) определить коэффициенты эластичности;

3) на основе коэффициентов стандартизованного уравнения и коэффициентов эластичности сравнить степень влияния факторов на эндогенную переменную

3. Проверить качество построенного уравнения (с помощью инструмента Регрессия):
   1) исследовать качество модели на основе коэффициента множественной корреляции, коэффициента детерминации, скорректированного коэффициента детерминации;
   2) проверить значимость уравнения с помощью *F*-критерия Фишера (вместо расчета критического значения *F* использовать столбец Значимость *F*);
   3) проверить значимость параметров уравнения с помощью *t*-критерия Стьюдента(вместо расчета критического значения *t* использовать столбец *p*-значение)
4. Провести анализ остатков на выполнение пяти предпосылок МНК:
   1) проверить выполнение предпосылки о случайном характере остатковс помощью точечной диаграммы зависимости остатков от *y*;
   2) проверить выполнение предпосылки о нулевой средней величине остатков с помощью Графика остатков инструмента Регрессия;
   3) проверить выполнение предпосылки о гомоскедастичности;
   4) проверить выполнение предпосылки об отсутствии автокорреляции остатков с помощью теста Дарбина-Уотсона;
   5) проверить выполнение предпосылки о нормальном распределении остатков с помощью Графика нормальной вероятности инструмента Регрессия.

Варианты заданий

Исследуется производительность труда на предприятиях одной из отраслей.

y – производительность труда на предприятии, руб./чел.
x_1 – средняя заработная плата, руб.
x_2 – доля высококвалифицированных работников, %
x_3 – инвестиции в основные фонды в текущем квартале, тыс.руб.
x_4 – инвестиции в основные фонды в предыдущем квартале, тыс.руб.

Вариант 1

y	x_1	x_2	x_3	x_4
14163,4	14343,9	37,9	3712,08	2896,18
15700,9	15736,6	42,2	2750,49	3299,60
10191,4	15851,6	43,7	1811,41	1520,13

y	x_1	x_2	x_3	x_4
15172,9	26047,8	25,3	2568,23	2017,30
8180,0	19781,5	36,8	1071,07	543,27
7871,5	14780,0	23,4	2297,24	1356,53
13750,9	18226,3	30,6	3811,98	2507,30
21531,9	23431,7	28,1	3609,77	2871,25
18951,8	18575,0	40,8	3796,96	2531,42
14812,4	14061,3	33,2	2801,13	2202,12
9863,4	19419,2	28,5	2927,71	2826,32
9461,9	17065,5	38,7	1768,73	2697,51
9446,1	13195,9	46,2	2630,41	1816,78
13895,7	14720,8	34,0	3592,92	2456,30
9111,3	18256,5	36,7	2991,88	2379,43
10593,1	12189,2	37,2	3346,85	2249,65
16296,4	20172,3	40,3	2918,41	2921,22
10426,8	12648,4	40,9	2755,49	2601,14
6068,3	14139,8	40,0	1094,32	778,60
11125,6	14866,3	42,0	2332,58	2225,41

Вариант 2

y	x_1	x_2	x_3	x_4
15537,2	24985,3	39,2	2719,24	3154,72
20492,3	20866,0	60,0	1000,19	2391,42
25377,1	26352,3	62,0	1643,73	1789,64
16829,4	18454,5	62,2	1839,19	3330,46
23735,4	22073,7	65,8	2037,06	2077,28
31366,0	29346,9	48,5	1486,61	3327,56
16009,4	25767,9	50,7	2075,16	3348,70
21754,7	25919,2	47,5	1845,89	4004,76
20210,3	21953,3	59,2	1518,67	2652,36
50740,4	29218,5	52,2	2077,48	4432,78
5162,7	27258,4	26,4	2048,12	2949,15
13057,4	28558,2	30,1	1905,20	2717,73
11535,7	29302,3	44,6	1387,90	1297,64
49570,9	25787,5	71,1	2018,61	3961,47
16784,6	28079,8	38,6	2587,69	4165,44
12189,6	24380,5	56,4	1651,15	2271,13
8640,0	23793,8	31,2	1867,40	3891,67

Вариант 3

y	x_1	x_2	x_3	x_4
7948,1	6773,9	34,8	1285,63	838,99
11193,7	17663,1	23,3	1816,52	1133,25
9923,6	15657,6	18,1	1523,58	1722,58
6441,2	14918,7	20,6	1539,46	1270,27
8024,9	14334,5	20,2	724,16	728,06
5094,2	9059,6	18,0	1036,19	1314,51
7755,2	16329,4	22,8	1773,48	396,21
6641,0	6668,6	27,6	1245,01	1176,72
9501,4	25650,8	16,5	1020,47	1715,50
11984,9	18391,0	18,9	1937,11	2354,55
6794,2	9919,6	12,8	253,16	1267,62
8462,2	12633,8	23,2	1060,90	1961,94
6572,9	15966,7	19,0	866,50	-314,37
6952,0	11180,3	25,0	1298,73	2320,66
7288,0	12720,7	24,6	1626,90	1119,05
7364,4	19859,3	20,7	931,76	938,37
5742,0	11110,2	32,1	1057,05	49,36
10202,2	19039,9	24,2	1683,96	1386,83
7895,2	14635,7	22,4	1202,91	711,93
6708,9	12762,5	18,8	154,19	-200,88
10847,2	26646,4	21,0	739,32	1512,59
9407,9	17212,6	25,2	1646,26	2049,85
7807,5	12963,8	21,9	1814,23	1227,20
7775,7	14981,8	18,1	1046,08	1566,06
7921,4	14639,8	22,2	1784,37	804,19
6360,7	12833,3	15,7	689,68	830,51
9529,4	16760,3	21,3	1722,09	2505,53
8994,1	20765,1	25,1	756,30	1219,31
6309,3	11609,7	29,7	637,08	527,81
7422,4	13595,5	21,9	1035,73	1363,77

Вариант 4

y	x_1	x_2	x_3	x_4
4602,2	11560,5	23,2	295,65	283,31
7745,3	10372,9	35,2	287,61	289,13
10413,8	12919,9	33,4	261,21	297,93
11773,7	11226,2	34,2	364,04	310,51
4365,0	11727,1	33,2	295,66	242,00

y	x_1	x_2	x_3	x_4
5691,7	11118,0	35,5	269,15	244,29
3795,0	11044,6	19,8	335,76	279,33
5414,8	11779,0	29,5	322,85	269,90
8040,2	12480,7	33,1	315,13	268,38
5730,2	10252,2	25,0	340,29	280,09
6712,6	12587,7	37,6	296,42	230,84
5911,2	9526,3	30,3	353,10	285,15
7329,2	10961,5	25,5	387,98	307,35
7044,2	11703,5	32,9	300,27	262,46
7107,5	10805,3	34,1	237,24	270,81
9348,3	12377,4	36,1	296,18	265,31
7216,9	10787,1	35,4	352,77	275,83
7049,4	10673,8	38,4	288,29	285,30
6661,1	9542,5	31,4	396,93	280,33
10387,3	11788,5	30,7	348,83	292,49
5013,5	9838,4	31,8	374,49	261,34
5433,4	12238,0	33,6	279,90	254,70
4129,6	11888,8	36,1	284,75	237,54
6275,3	12030,8	22,4	340,93	283,29
5426,6	11847,1	22,9	308,16	279,93
5998,8	12298,2	29,9	303,35	265,64
5967,8	9156,8	28,0	352,87	283,18

Вариант 5

y	x_1	x_2	x_3	x_4
7576,3	14239,9	38,1	959,32	711,85
6860,7	14024,0	31,6	636,33	431,74
13413,8	15202,1	33,0	1177,94	911,70
11588,1	18117,9	28,3	1311,09	834,23
12785,8	11918,9	34,0	2286,90	3308,22
19131,5	18692,5	34,1	2864,54	2559,67
15116,4	17185,1	62,8	1668,25	1761,92
13044,9	19210,9	66,5	545,31	476,47
16080,2	18898,9	35,0	760,93	1117,19
20117,6	20344,0	36,5	2266,12	1576,40
9238,6	18141,8	40,9	1218,70	701,55
10899,3	14667,1	38,6	1260,19	558,17
13373,5	15595,1	47,8	1184,23	1765,73
5661,4	13666,4	33,0	1098,98	1608,25

y	x_1	x_2	x_3	x_4
13339,8	20281,1	33,6	1182,13	1354,50
12377,8	18380,2	19,1	1462,09	1677,86
8922,9	17641,2	38,2	935,32	187,81
13192,2	15660,5	31,0	1719,43	2209,61
11817,4	16087,6	46,3	1599,66	2114,28
11612,6	15398,3	26,1	2815,79	3047,27
11377,3	17277,6	22,6	1917,00	1410,80
9771,1	15791,8	28,6	499,24	1176,23

Исследуется динамика цен на первичном рынке жилья.

y – индекс цен на первичном рынке жилья, %

x_1 – среднедушевой доход населения, руб.

x_2 – индекс цен на строительные материалы в текущем году, %

x_3 – индекс цен на строительные материалы в предыдущем году, %

x_4 – индекс цен на строительно-монтажные работы, %

Вариант 6

y	x_1	x_2	x_3	x_4
110,7	14780,6	104,2	107,2	117,3
118,0	16468,8	105,1	107,9	121,9
121,8	16608,1	108,0	112,1	110,8
109,8	28968,7	104,3	107,4	109,9
117,2	21371,5	109,3	107,9	104,2
74,1	15308,7	93,8	100,3	89,5
121,6	19486,5	107,8	112,3	118,9
133,1	25796,8	108,1	111,0	122,5
112,1	19908,9	102,3	108,5	111,5
121,4	14437,9	104,9	109,9	112,3
103,3	20932,7	103,6	109,3	115,8
85,5	18079,4	97,2	103,7	105,7
99,8	13389,8	102,3	107,2	105,3
126,6	15237,2	109,2	110,5	120,5
87,9	19523,0	99,5	106,5	106,2
105,5	12168,9	104,5	105,1	112,2
114,5	21844,9	102,5	113,3	115,1
127,7	12725,8	109,9	113,0	122,8
94,1	14533,3	100,8	105,2	94,4
105,5	12168,9	104,5	105,1	112,2
114,5	21844,9	102,5	113,3	115,1

y	x_1	x_2	x_3	x_4
127,7	12725,8	109,9	113,0	122,8
94,1	14533,3	100,8	105,2	94,4
101,6	15414,0	100,0	110,2	105,6

Вариант 7

y	x_1	x_2	x_3	x_4
106,6	17634,7	105,7	102,1	107,2
110,4	8710,1	110,7	102,7	108,2
100,9	20596,6	101,2	100,8	101,2
97,7	3485,5	101,1	101,6	105,0
97,9	11327,0	98,8	102,0	100,5
119,8	27085,0	111,8	103,2	111,2
106,8	19330,5	107,4	103,7	108,7
104,7	19658,3	102,8	100,1	107,7
103,8	11065,8	104,6	101,8	105,3
130,3	26806,7	111,8	105,7	114,0
105,4	22559,8	106,5	99,6	107,2
108,9	25376,2	103,7	101,2	104,9
105,2	26988,3	105,4	101,9	102,4
113,3	19373,2	107,5	101,2	110,3
103,8	24339,6	102,6	100,9	107,9
101,9	16324,5	107,4	102,3	106,0
109,4	15052,9	105,3	102,4	108,8
107,6	11001,8	101,0	101,0	105,6
103,1	20705,9	102,7	99,1	105,1
103,6	15901,9	106,5	103,6	102,0

Вариант 8

y	x_1	x_2	x_3	x_4
126,2	9352,6	122,6	108,3	112,0
123,2	19708,6	119,7	99,6	113,4
133,3	17801,2	127,1	107,6	121,3
101,8	17098,2	114,0	101,9	112,4
114,0	16542,6	117,6	99,9	109,1
109,4	11524,8	119,0	105,9	114,7
123,2	18440,1	130,4	108,6	111,2
117,4	9251,6	120,7	105,7	114,2
109,8	27307,5	109,9	103,2	114,6
129,1	20401,3	114,2	105,7	121,7

y	x_1	x_2	x_3	x_4
121,3	12342,5	123,2	102,9	115,9
109,9	14924,9	109,8	99,4	116,7
125,9	18095,2	132,5	111,2	105,9
116,0	13542,5	120,4	106,3	123,8
115,2	15007,6	126,5	102,1	115,9
104,9	21797,8	113,8	102,1	109,5
119,5	21018,2	112,1	106,9	112,6
110,0	16829,1	115,2	100,8	108,1
124,0	15047,2	127,3	106,4	104,9

Вариант 9

y	x_1	x_2	x_3	x_4
105,3	17361,4	104,4	107,5	108,3
100,9	19261,8	104,1	108,4	106,8
102,8	16779,0	103,0	109,7	111,0
113,1	20812,9	108,5	111,6	118,7
99,3	17640,4	103,3	101,3	110,0
97,2	20607,1	102,1	101,6	104,3
105,2	10769,4	106,3	106,9	114,6
101,4	20662,5	105,4	105,5	104,5
101,2	18746,5	105,0	105,3	110,8
107,4	22141,8	106,5	107,0	117,2
97,5	16443,5	103,1	99,6	102,2
102,7	16948,8	107,3	107,8	110,4
108,9	13949,5	109,6	111,1	119,0
99,1	16359,4	104,1	104,4	111,8
102,4	20914,3	101,1	105,6	108,8
96,8	17175,2	103,9	104,8	100,9
113,8	19118,6	107,0	106,4	122,0
102,6	14026,2	104,0	107,8	115,7
108,5	17045,6	109,4	107,1	119,2
113,8	19118,6	107,0	106,4	122,0
102,6	14026,2	104,0	107,8	115,7
108,5	17045,6	109,4	107,1	119,2
108,0	18236,8	107,3	108,9	119,4

Вариант 10

y	x_1	x_2	x_3	x_4
113,7	20666,8	104,9	107,2	102,2
110,5	23362,2	103,3	106,9	102,8
113,0	18346,0	105,0	106,4	104,4
104,7	18450,8	103,0	104,5	101,9
104,7	18942,3	102,1	107,1	100,5
112,6	17021,6	103,0	109,0	102,9
110,1	18301,7	104,0	103,3	104,0
101,7	21092,0	102,3	104,4	102,7
113,6	20504,0	104,2	104,6	104,2
97,4	16378,1	102,7	104,3	101,4
117,3	22126,6	105,7	105,5	103,8
110,9	18206,4	103,1	106,3	104,1
114,4	20063,5	104,3	102,5	105,1
100,2	24192,3	100,3	103,2	101,2
111,4	18782,1	103,8	106,3	102,0
110,1	18301,7	104,0	103,3	104,0
101,7	21092,0	102,3	104,4	102,7
113,6	20504,0	104,2	104,6	104,2
97,4	16378,1	102,7	104,3	101,4
117,3	22126,6	105,7	105,5	103,8
110,9	18206,4	103,1	106,3	104,1
114,4	20063,5	104,3	102,5	105,1
100,2	24192,3	100,3	103,2	101,2
111,4	18782,1	103,8	106,3	102,0
106,2	20272,6	103,6	102,2	102,8

6. Методика выполнения заданий

Исследуется взаимосвязь показателей качества жизни населения по выборке для 25 регионов (рис. 4.2).

y – средняя ожидаемая продолжительность жизни при рождении, лет;

x_1 – уровень рождаемости, чел. на 1000 чел. населения;

x_2 – доля населения с денежными доходами ниже величины прожиточного минимума, % от всего населения;

x_3 – среднедушевые доходы населения, у.е.;

x_4 – объем социальных выплат, млрд. у.е.

	A	B	C	D	E	F
1	i	Y	X1	X2	X3	X4
2	1	68,1	10,2	11,2	14,04	6,09
3	2	68,2	10,5	14,0	16,27	6,79
4	3	69,0	11,7	11,9	23,41	4,50
5	4	68,2	11,3	12,0	16,41	4,71
6	5	66,6	8,8	14,3	11,25	5,72
7	6	68,6	11,9	11,0	21,22	4,69
8	7	68,3	11,4	11,3	14,72	6,11
9	8	67,3	9,0	14,3	11,31	6,65
10	9	68,6	11,4	12,6	23,04	5,18
11	10	68,4	12,0	12,5	21,67	5,41
12	11	69,1	11,1	10,5	20,80	5,83
13	12	69,1	12,3	11,2	21,55	4,85
14	13	68,8	12,0	12,5	18,08	5,57
15	14	68,7	12,5	13,0	19,81	5,58
16	15	68,6	11,2	15,1	16,16	6,52
17	16	68,6	12,5	12,8	18,87	5,70
18	17	69,0	12,2	12,2	22,43	5,72
19	18	68,5	10,5	13,9	17,06	6,84
20	19	67,9	10,9	12,9	20,53	5,43
21	20	69,7	13,1	11,8	23,49	6,02
22	21	68,5	10,4	11,6	21,98	5,11
23	22	68,6	11,9	13,1	19,48	5,34
24	23	68,3	12,5	12,1	21,30	4,95
25	24	67,0	8,1	15,2	11,22	7,43
26	25	68,0	10,1	12,3	20,33	6,06

Рис. 4.2. Исходные данные задачи

1. Для исследования целесообразности включения факторов в модель построим корреляционную матрицу. Воспользуемся инструментом Корреляция пакета Анализ данных (закладка Данные). В окне настройки параметров (рис. 4.3) в качестве входного интервала укажем столбцы значений *x* и *y* с заголовками, в качестве выходного интервала – область ячеек 6×6. Нажав OK, получим корреляционную матрицу (рис. 4.4).

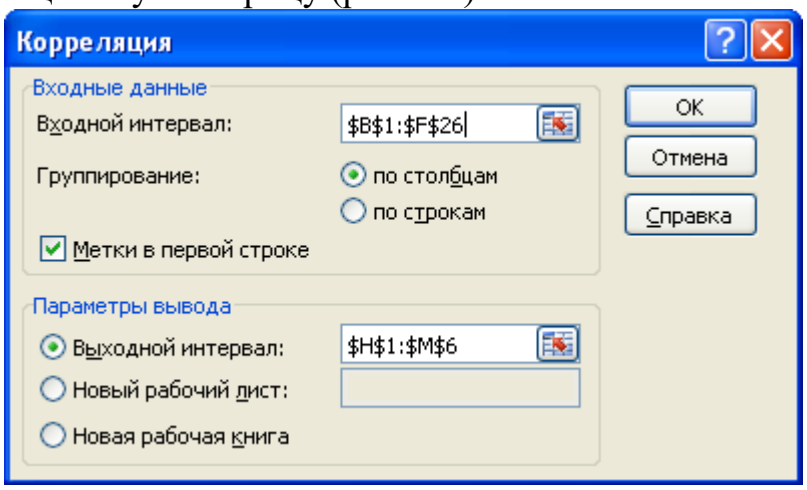


Рис. 4.3. Настройка параметров инструмента Корреляция

	H	I	J	K	L	M
1		*Y*	*X1*	*X2*	*X3*	*X4*
2	Y	1				
3	X1	0,84358	1			
4	X2	-0,56285	-0,52249	1		
5	X3	0,78743	0,75317	-0,57922	1	
6	X4	-0,38186	-0,56991	0,63499	-0,63884	1

Рис. 4.4. Корреляционная матрица

Проанализируем полученную корреляционную матрицу:

1) Между зависимой переменной y и независимыми переменными x_1, x_2, x_3наблюдается корреляционная заметная связь, в то время как между y и x_4 связь не достаточно сильна ($|r_{yx_4}| < 0{,}4$). Следовательно, в модель не стоит включать фактор x_4.

2) Между независимыми переменными x_1иx_3наблюдается высокая нежелательная корреляционная связь ($|r_{x_1x_3}| > 0{,}7$). Для исключения мультиколлинеарности один из этих факторов нужно убрать из модели. По сравнению с x_1 фактор x_3 сильнее связан с x_2 и слабее с y, поэтому в модели следует оставить x_1.

Таким образом, включаем в модель факторы x_1 и x_2.

С помощью инструмента Регрессия построим и оценим уравнение линейной множественной регрессии. Для этого в окне параметров инструмента (рис. 4.5) зададим в качестве входного интервала – столбец со значениями y, в качестве выходного интервала – столбцы со значениями факторов x, которые мы ранее выбрали для включения в модель – x_1 и x_2. Для последующего анализа выполнения предпосылок нам также пригодится информация об остатках и график нормальной вероятности. Часть результатов Регрессии, необходимая для построения и оценки качества уравнения, представлена на рис. 4.6.

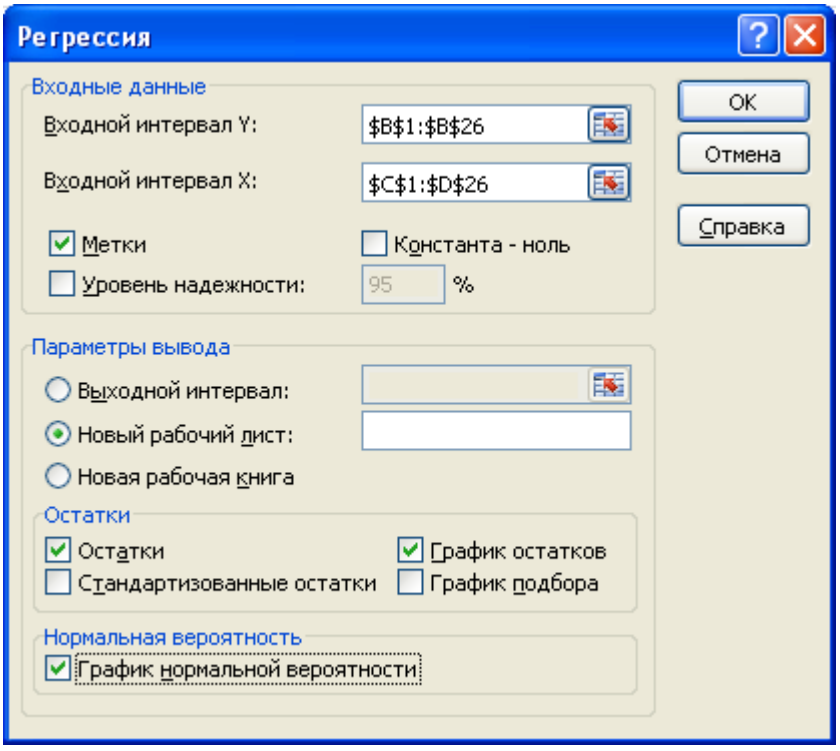


Рис. 4.5. Настройка параметров инструмента Регрессия

	A	B	C	D	E	F	G	H	I
1	ВЫВОД ИТОГОВ								
2									
3	*Регрессионная статистика*								
4	Множественный R	0,855645232							
5	R-квадрат	0,732128764							
6	Нормированный R-квадрат	0,707776833							
7	Стандартная ошибка	0,359476251							
8	Наблюдения	25							
9									
10	Дисперсионный анализ								
11		*df*	*SS*	*MS*	*F*	*Значимость F*			
12	Регрессия	2	7,770061865	3,885030932	30,06450605	0,000000510			
13	Остаток	22	2,842909854	0,129223175					
14	Итого	24	10,61297172						
15									
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
17	Y-пересечение	65,02301895	1,418055429	45,85365114	0,00000000	62,08215199	67,9639	62,0822	67,9639
18	X1	0,400802485	0,068625655	5,840417654	0,00000708	0,258481588	0,54312	0,25848	0,54312
19	X2	-0,087990916	0,067805884	-1,297688494	0,20783669	-0,228611712	0,05263	-0,22861	0,05263

Рис. 4.6. Итоги регрессии (регрессионная статистика и дисперсионный анализ)

2. Построим линейное уравнение множественной регрессии:

1) В ячейках B17:B19 (рис. 4.6) выведены параметры уравнения в естественной форме. Следовательно, наше уравнение имеет вид:

$$\hat{y}_x = 65{,}02 + 0{,}401 \cdot x_1 - 0{,}088 \cdot x_2$$

Параметры найденного уравнения показывают, что:

- при увеличении уровня рождаемости на 1 чел./1000 чел. средняя ожидаемая продолжительность жизни увеличится в среднем на 0,401 лет при неизменном значении других факторов;
- при увеличении доли населения с денежными доходами ниже величины прожиточного минимума на 1% средняя ожидаемая продолжительность жизни уменьшится в среднем на 0,088 лет;
- при нулевом уровне рождаемости и отсутствии населения с денежными доходами ниже величины прожиточного минимума средняя ожидаемая продолжительность жизни составит 65,02 года.

2) Определим параметры стандартизованного уравнения по формулам перехода $\beta_i = b_i \frac{\sigma_{x_i}}{\sigma_y}$. Стандартные отклонения рассчитаем с помощью функции Excel СТАНДОТКЛОН.Г. Результаты вычислений параметров уравнения в стандартизованной форме показаны на рис. 4.7. Уравнение в стандартизованной форме имеет вид:

$$t_y = 0{,}756 \cdot t_{x_1} - 0{,}168 \cdot t_{x_2}.$$

Исходя из коэффициентов уравнения регрессии в стандартизованном виде, x_1 оказывает б*о*льшее влияние на y, чем x_2.

3) Определим коэффициенты эластичности по формуле $\overline{Э}_{yx_j} = b_j \frac{\overline{x}_j}{\overline{y}}$. $\overline{Э}_{yx_1} = 0{,}065$, $\overline{Э}_{yx_2} = -0{,}016$. Исходя из средних коэффициентов эластичности, x_1 оказывает б*о*льшее влияние на y, чем x_2. Это полностью подтверждает выводы, сделанные по стандартизованному уравнению.

B32 f_x =СТАНДОТКЛОН.Г(B2:B26)

	A	B	C	D	E	F
28	Средние	Y	X1	X2		
29	значения	68,4	11,2	12,6		
30						
31	Стандартное	Y	X1	X2		
32	отклонение	0,6516	1,2287	1,2435		
33						
34	Параметры		β_1	β_2		
35	стандартизованного уравнения		0,75583	-0,1679		
36						
37	Коэффициенты		$Э_{yx1}$	$Э_{yx2}$		
38	эластичности		0,06549	-0,0162		

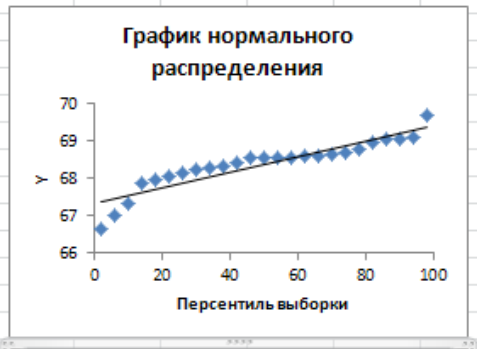
Рис. 4.7. Сравнение степени влияния факторов

3. Проверим качество построенного уравнения. Для этого воспользуемся результатами Регрессии (рис. 4.6).

1) Коэффициент множественной корреляции R = 0,856 говорит о высокой степени тесноты линейной связи средней ожидаемой продолжительности жизни с совокупностью факторов x_1 и x_2. Коэффициент детерминации $R^2 = 0,732$ и скорректированный коэффициент детерминации $\overline{R}^2 = 0,708$ свидетельствуют о высоком качестве регрессионной модели.
2) проверим значимость уравнения с помощью *F*-критерия Фишера:на рис 4.6 расчетное значение $F = 30,06$, при этом Значимость $F = 0,00000051$, что существенно меньше заданного уровня значимости $\alpha = 0,05$. Таким образом, уравнение статистически значимо.
3) проверим значимость параметров уравнения с помощью *t*-критерия Стьюдента: на рис 4.6 расчетные значения *t*-статистики для свободного члена и параметров b_1, b_2 составляют соответственно 45,854, 5,840 и -1,298. При этом *p*-значения для свободного члена и b_1существенно меньше 0,05, следовательно, гипотезу о незначимости этих параметров следует отвергнуть. *p*-значение для $b_2 = 0,2 > 0,05$, т.е. у нас нет оснований отвергать гипотезу о незначимости этого параметра.

4. Проведем анализ остатков на выполнение пяти предпосылок МНК:
   1) Проверим выполнение предпосылки о случайном характере остатков: инструмент Регрессия имеет опцию вывода остатков, результат представлен на рис. 4.8. Построим на основе этих данных точечную диаграмму: по оси абсцисс отложим теоретические значения $\hat{y}_x$, по оси ординат – остатки. Из графика остатков на рис. 4.8 видно, что точки расположены случайным образом, то есть первая предпосылка выполняется.
   2) Проверим выполнение предпосылки о нулевой средней величине остатков: инструмент Регрессия имеет опцию вывода Графиков остатков, результат представлен на рис. 4.9. На графиках видно, что остатки расположены случайным образом, то есть вторая предпосылка выполняется.
   3) Проверим выполнение предпосылки о гомоскедастичности. Наличие гомоскедастичности или гетероскедастичности можно видеть по графику зависимостиостатков от теоретического значения результативного признака $\hat{y}_x$ (рис. 4.8). На графике видно, что дисперсия остатков в целом не меняется при переходе от одного значения $\hat{y}_x$ к другому. Следовательно, третья предпосылка выполняется.

	A	B	C	D	E	F
23	ВЫВОД ОСТАТКА				ВЫВОД ВЕРОЯТНОСТИ	
24						
25	*Наблюдение*	*Предсказанное Y*	*Остатки*		*Персентиль*	*Y*
26	1	68,1261753	-0,050215406		2	66,63889118
27	2	67,99777676	0,225167058		6	67,02356858
28	3	68,65731585	0,38304324		10	67,33789524
29	4	68,49660015	-0,325331342		14	67,88444355
30	5	67,28205069	-0,643159505		18	67,99174002
31	6	68,81157292	-0,258767194		22	68,0759599
32	7	68,5963647	-0,290084532		26	68,17126881
33	8	67,38268449	-0,044789245		30	68,22294382
34	9	68,46668241	0,123323421		34	68,30628017
35	10	68,73400046	-0,328925594		38	68,32765768
36	11	68,5670487	0,542206845		42	68,40507486
37	12	68,96921647	0,105573248		46	68,541426
38	13	68,74142385	0,05938629		50	68,54190711
39	14	68,87382312	-0,194610569		54	68,55280572
40	15	68,18341302	0,408605432		58	68,5773495
41	16	68,90071264	-0,323363141		62	68,59000583
42	17	68,85620214	0,113680352		66	68,59201845
43	18	67,99418822	0,547718884		70	68,64390963
44	19	68,27297193	-0,388528375		74	68,67921255
45	20	69,23089102	0,471413842		78	68,80081014
46	21	68,17785563	0,363570365		82	68,96988249
47	22	68,62726096	0,016648677		86	69,04035909
48	23	68,95986873	-0,632211056		90	69,07478972
49	24	66,92200027	0,101568308		94	69,10925555
50	25	67,97366002	0,018079997		98	69,70230486



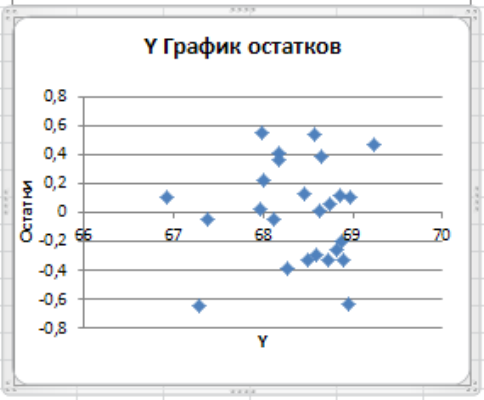


Рис. 4.8. Итоги регрессии (вывод остатка и вывод вероятности)

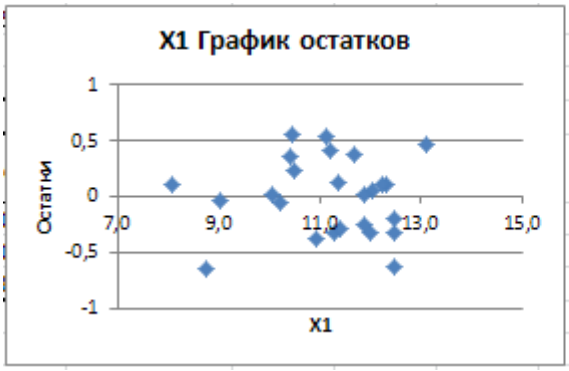


Рис. 4.9. Итоги регрессии (графики остатков)

4) Проверим выполнение предпосылки об отсутствии автокорреляции остатков с помощью теста Дарбина-Уотсона:
   - Выдвигаем гипотезу H_0 об отсутствии автокорреляции;
   - Для расчетов используем значения из столбца «Остатки» результатов Регрессии (рис. 4.8). Занесем их в столбец B нового листа (рис. 4.10);

B2 | =Регрессия!C26

	A	B	C	D	E	F	G
1	i	u_i	u^2_i	$(u_i-u_{i-1})^2$		Проверка наличия автокорреляции в остатках	
2	1	-0,050	0,003	---		Статистика Дарбина-Уотсона DW	2,165
3	2	0,225	0,051	0,076		Число наблюдений n	25
4	3	0,383	0,147	0,025		Число независимых переменных m	2
5	4	-0,325	0,106	0,502		Критическое значение d_L	1,21
6	5	-0,643	0,414	0,101		Критическое значение d_U	1,55
7	6	-0,259	0,067	0,148		Вывод	автокорреляция отсутствует
8	7	-0,290	0,084	0,001			
9	8	-0,045	0,002	0,060			
10	9	0,123	0,015	0,028			
11	10	-0,329	0,108	0,205			
12	11	0,542	0,294	0,759			
13	12	0,106	0,011	0,191			
14	13	0,059	0,004	0,002			
15	14	-0,195	0,038	0,065			
16	15	0,409	0,167	0,364			
17	16	-0,323	0,105	0,536			
18	17	0,114	0,013	0,191			
19	18	0,548	0,300	0,188			
20	19	-0,389	0,151	0,877			
21	20	0,471	0,222	0,740			
22	21	0,364	0,132	0,012			
23	22	0,017	0,000	0,120			
24	23	-0,632	0,400	0,421			
25	24	0,102	0,010	0,538			
26	25	0,018	0,000	0,007			
27	сумма		2,843	6,155			

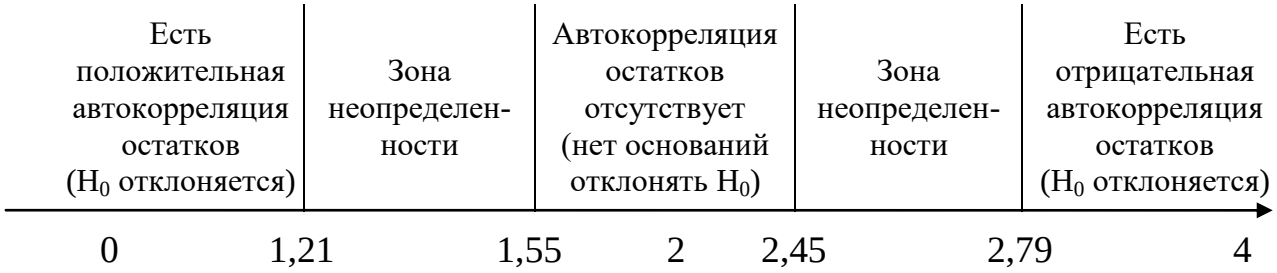
Рис. 4.10. Проверка гипотезы об отсутствии автокорреляции

- В столбце C возведем остатки в квадрат, в столбце D рассчитаем квадраты отклонений текущего значения остатка от предыдущего и просуммируем значения (рис. 4.10);
- Рассчитаем статистику Дарбина-Уотсона по формуле:

$$DW = \frac{\sum_{i=2}^{n}(u_i - u_{i-1})^2}{\sum_{i=1}^{n} u_i^2};$$

- Для числа наблюдений n=25 и числа независимых переменных m=2 и уровня значимости α=0,05 по специальной таблице (см. приложение 1) найдем критические значения d_L=1,21 и d_U=1,55;
- Построим шкалу Дарбина-Уотсона и зададим автоматический вывод результата с помощью функции Excel ЕСЛИ. Используя условное форматирование, выделим ячейку красным цветом в случае наличия автокорреляции, зеленым –

в случае отсутствия, желтым – при попадании значения в зону неопределенности.



5) Проверим выполнение предпосылки о нормальном распределении остатков с помощью Графика нормальной вероятности инструмента Регрессия (рис. 4.8): поскольку все точки расположены близко к прямой, то можно считать, что распределение остатков близко к нормальному и пятая предпосылка выполняется.

Таким образом, все предпосылки МНК выполняются.

7. Контрольные вопросы

1. В каких случаях строится уравнение множественной регрессии? Какова основная цель множественной регрессии?
2. Какой метод применяется для оценки параметров линейного уравнения множественной регрессии? В каких функциях MS Excel он реализован?
3. Каким требованиям должны отвечать факторы, включаемые во множественную регрессию?
4. Что показывает корреляционная матрица? Как построить корреляционную матрицу с помощью MS Excel?
5. Каким образом строится и для чего используется уравнение множественной регрессии в стандартизованном масштабе?
6. Что показывает линейный коэффициент множественной корреляции?
7. В чем особенности скорректированного коэффициента детерминации?
8. Как с помощью инструмента Регрессия можно оценить значимость уравнения регрессии и его параметров?
9. Назовите пять предпосылок МНК. Каким образом можно проверить выполнение предпосылок средствами MS Excel?
10. Для чего используется тест Дарбина-Уотсона? Опишите его алгоритм.

8. Требования к содержанию отчета

Отчет к лабораторной работе предоставляется в электронном виде и должен содержать:

- название и цель работы;
- номер и исходные данные своего варианта;
- описание хода выполнения заданий, в том числе:
  - скриншоты из MS Excel, отображающие заданные в ячейках формулы и функции, использованные для вычислений;
  - скриншоты из MS Excel с полученными при расчетах результатами;
  - анализ и интерпретация полученных результатов;
- выводы по лабораторной работе.

9. Требования к оформлению отчета

- Все рисунки в отчете должны быть подписаны.
- Все скриншоты должны быть читаемыми в масштабе документа 100%. При необходимости используйте обрезку и пропорциональное изменение размера рисунка.
- Все скриншоты таблиц MS Excel должны содержать системное наименование строк (1, 2, 3….) и столбцов (A, B, C, …)

Лабораторная работа №5
Множественный линейный регрессионный анализ в условиях мультиколлинеарности с помощью пакета STATISTICA.

1. Цель и задачи лабораторной работы

Цель работы: изучить возможности пакета STATISTICAдля построения многофакторной линейной регрессии, оценки ее качества и анализа остатков.

Задачи:

– научиться проводить диагностику проблемы мультиколлинеарности;

– приобрести навыки определения факторного состава регрессии с помощью пошаговых алгоритмов;

– приобрести навыки нахождения параметров уравнения многофакторной линейной регрессии в естественной и стандартизованной форме с помощью пакета STATISTICA;

– научиться проверять качество уравнения многофакторной линейной регрессии с помощью пакета STATISTICA;

– научиться проводить анализ остатков на выполнение пяти предпосылок метода наименьших квадратов с помощью пакета STATISTICA.

2. Теоретическая часть

2.1. Механизм возникновения проблемы мультиколлинеарности

Проблема мультиколлинеарности возникает, если между факторами, включенными в регрессионную модель, существует тесная линейная зависимость. Механизм возникновения проблемы:

1. Для нахождения вектора регрессионных коэффициентов используется формула $b=(X^TX)^{-1}X^TY$, поэтому для нахождения вектора b необходимо найти матрицу $(X^TX)^{-1}$.

2. При наличии тесной линейной связи между факторами определитель матрицы X^TXблизок к 0, следовательно, элементы матрицы $(X^TX)^{-1}$–большие числа.

3. При выполнении теста Стьюдента t-статистика рассчитывается по следующей формуле: $t_j = \frac{b_j}{S_{b_j}}$. Так как элементы матрицы $(X^TX)^{-1}$- большие числа, то S_{b_j}–также большие числа, следовательно, t_j– маленькие числа и гипотезу H_0: $\beta_j=0$ нужно принимать. Таким образом, многие факторы в модели будут незначимы, следовательно, результаты, полученные по выборке о влиянии факторов на результативный показатель, нельзя распространять на всю генеральную совокупность, следовательно, полученная модель практически бесполезна.

2.2. Симптомы мультиколлинеарности

- Незначимость большинства регрессионных коэффициентов (по результатам тестов Стьюдента в большинстве случаев нулевую гипотезу принимаем) при значимости уравнения в целом (по результату теста Фишера нулевую гипотезу отвергаем);
- Значительные изменения регрессионных коэффициентов при незначительных изменениях объема выборки или состава факторов, включенных в модель;
- Чрезмерно высокие или противоречащие по знакам экономической теории значения регрессионных коэффициентов.

2.3. Методы диагностики проблемы мультиколлинеарности

Наиболее очевидный способ выявления мультиколлинеарности – это анализ *матрицы парных коэффициентов корреляции* между факторами, включаемыми в модель. Если между какими-либо парами факторов парный коэффициент корреляции по модулю больше 0,7, то в модели будет проблема мультиколлинеарности.

На первый взгляд может показаться, что матрица парных коэффициентов корреляции играет главную роль в отборе факторов. Вместе с тем вследствие взаимодействия факторов парные коэффициенты не могут в полной мере решать вопрос о целесообразности включения в модель того или иного фактора. Эту роль выполняют показатели частной и получастной корреляции, оценивающие в чистом виде тесноту связи фактора с результатом.

Для их расчета находят остатки $e(x_j)$ регрессии $x_j = b_0 + b_1x_1 + \cdots + b_{j-1}x_{j-1} + b_{j+1}x_{j+1} + \cdots + b_mx_m$ и остатки $e(y\backslash x_j)$ регрессии $y_x = b_0 + b_1x_1 + \cdots + b_{j-1}x_{j-1} + b_{j+1}x_{j+1} + \cdots + b_mx_m$.

Коэффициент корреляции, подсчитанный между $e(x_j)$ и $e(y\backslash x_j)$, называется *частным коэффициентом корреляции* между x_j и y. Частный коэффициент корреляции оценивает направление и тесноту линейной связи между x_j и y при исключении влияния остальных факторов на оба показателя. Если частный коэффициент корреляции по модулю большой, то фактор должен присутствовать в модели.

Коэффициент корреляции, подсчитанный между $e(x_j)$ и y, называется *получастным коэффициентом корреляции* между x_j и y. Получастный коэффициент корреляции оценивает направление и тесноту линейной связи между x_j и y при исключении влияния остальных факторов на показатель x_j. Получастные коэффициенты корреляции позволяют ответить на вопрос о вкладе каждого фактора в множественный коэффициент корреляции. Квадрат j-го получастного коэффициента корреляции показывает, насколько уменьшится величина коэффициента детерминации, если j-й фактор исключить из модели.

Другой метод – расчет коэффициентов увеличения, или разбухания дисперсии:

$$VIF_j = \frac{1}{1 - R_j^2}$$

где R_j^2– коэффициент детерминации, подсчитанный по уравнению регрессии (в правой части уравнения присутствуют все факторы, кроме *j*-го фактора). Коэффициент VIF_j показывает, во сколько раз возрастает дисперсия *j*-го регрессионного коэффициента по сравнению со случаем отсутствия линейной связи между *j*-м фактором и остальными факторами модели. На практике считается, что если существуют $VIF_j \geq 3$, то в модели будет проблема мультиколлинеарности.

Толерантность переменной определяется как 1 минус квадрат множественной корреляции этой переменной со всеми другими независимыми переменными уравнения регрессии. Чем меньше толерантность переменной тем в большей степени ее вклад в регрессию является избыточным. Если толерантность каких-либо переменных в уравнении регрессии равна (или близка к) нулю, то оценки не могут быть вычислены.

2.4. Борьба с мультиколлинеарностью путем изменения состава факторов

Алгоритм пошагового включения

Шаг 1. Выбирают фактор $x_{(1)}$, наиболее сильно коррелированный с результативным показателем, и стоят модель парной регрессии $y=b_0+b_{(1)}x_{(1)}$.

Шаг 2. Оценивают все варианты регрессионных моделей с двумя факторами: $y = b_0+b_{(1)}x_{(1)}+b_jx_j$, $j=2..k$. Выбирают фактор $x_{(2)}$, обеспечивающий максимальный прирост коэффициента детерминации (фактор с максимальным коэффициентом получастной корреляции).

Шаг 3. Оценивают значимость включения фактора в модель с помощью теста Фишера.

Обозначим:

$R^2_{\text{без } x_{(2)}}$–коэффициент детерминации для модели $y = b_0+b_{(1)}x_{(1)}$,

R^2–коэффициент детерминации для модели $y = b_0+b_{(1)}x_{(1)}+b_{(2)}x_{(2)}$,

$\Delta R^2 = R^2 - R^2_{\text{без } x_{(2)}}$.

Нулевая гипотеза этого теста – H_0: $\Delta R^2 = 0$означает, что включение в модель фактора $x_{(2)}$ в генеральной совокупности не приводит к увеличению объясненной вариации результативной переменной. Тестовая статистика рассчитывается по формуле:

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

Доказано, что в случае справедливости нулевой гипотезы F является случайной величиной, распределенной по закону Фишера с 1, n-m-1 степенями свободы, где m– текущее количество факторов в модели. Поэтому по значению F можно определить вероятность p наблюдать полученное или большее значение F в том случае, если нулевая гипотеза верна. Если p окажется малой (например, если p< 0,05), то нулевую гипотезу нужно отвергать и принимать альтернативную $\Delta R^2 > 0$, т.е. включение в модель фактора $x_{(2)}$и в генеральной совокупности приводит к увеличению объясненной вариации результативной переменной.

Шаг 4. Если не все потенциальные факторы включены в модель на предыдущем шаге, то идем на шаг 2, определяя следующий фактор для включения в модель по критерию максимума прироста коэффициента детерминации, и затем идем на шаг 3 для оценки значимости включения следующего фактора в модель. Если же все потенциальные факторы включены в модель, то идем на шаг 5.

Шаг 5. Анализируем все рассмотренные модели и отбираем из них те, в которых все факторы значимы (во всех тестах Стьюдента нулевая гипотеза отвергалась). Из отобранных моделей выбираем ту, в который коэффициент детерминации максимален.

Алгоритм пошагового исключения

Шаг 1. Строим регрессионную модель, включающую все факторы, имеющиеся в нашем распоряжении.

Шаг 2. Если в текущей регрессионной модели есть факторы, то проверяем факторы на значимость с помощью тестов Стьюдента, выбираем самый незначимый фактор (тот, у которого p в тесте Стьюдента самая большая) и идем на шаг 3, иначе идем на шаг 4.

Шаг 3. Исключаем самый незначимый фактор, строим новую регрессионную модель без исключенного фактора, оцениваем значимость исключения фактора по тесту Фишера (см. алгоритм пошагового включения) и идем на шаг 2.

Шаг 4. Анализируем все рассмотренные модели и отбираем из них те, в которых все факторы значимы (во всех тестах Стьюдента нулевая гипотеза отвергалась). Из отобранных моделей выбираем ту, в который коэффициент детерминации максимален.

3. Описание оборудования и используемых программных комплексов

При выполнении лабораторной работы необходим специализированный компьютерный класс с минимальными системными требованиями компьютеров:

- Процессор – Intel Pentium IV;

- ОЗУ – 500 Mb;
- видеокарта – 64 Mb.
- Требуемое программное обеспечение:
- Операционная система Microsoft Windows;
- Microsoft Excel версии 2007 и выше;
- STATISTICA версии 6 и выше.

4. Краткое руководство по эксплуатации оборудования

При использовании оборудования необходимо:

- соблюдать общие правила нахождения в учебных лабораториях, работы с компьютером и использования программных средств;
- осмотреть рабочее место, убрать все мешающие работе предметы;
- визуально проверить правильность подключения ПЭВМ к электросети.

5. Задание

Для выданного варианта (табл. 1, 2) провести множественный линейный регрессионный анализ с помощью пакета STATISTICA. Для этого необходимо:

5. В программе STATISTICA выполнить регрессионный анализ, включив в модель все факторные переменные;
6. Провести диагностику проблемы мультиколлинеарности следующими способами:
    4) рассчитать матрицу парных коэффициентов корреляции факторов;
    5) рассчитать для каждого фактора R_j^2, $1 - R_j^2$, VIF_j частный и получастный коэффициенты корреляции;
7. Найти самую хорошую регрессионную модель:
    4) реализовав алгоритм пошагового включения;
    5) реализовав алгоритм пошагового исключения;
    6) сверить результаты с автоматической пошаговой регрессией пакета STATISTICA (Forward stepwise и Backward stepwise), в случае расхождений объяснить их причину;
8. Для выбранной модели определить степень влияния экзогенных переменных на эндогенную:
    1) по коэффициентам стандартизованного уравнения;
    2) по коэффициентам частной корреляции;
9. Проверить качество выбранного уравнения:
    1) исследовать качество модели на основе коэффициента множественной корреляции, коэффициента детерминации, скорректированного коэффициента детерминации;

2) проверить значимость уравнения с помощью *F*-критерия Фишера;
3) проверить значимость параметров уравнения с помощью *t*-критерия Стьюдента;

10.Провести анализ остатков на выполнение пяти предпосылок МНК.

Варианты заданий

Исследуется взаимосвязь показателей качества жизни населения по выборке для 50 стран (табл. 1).

Обозначения признаков:

x_1 – численность населения (в тыс. чел.);

x_2 – рождаемость (на 1000 чел.);

x_3 – смертность (на 1000 чел.);

x_4 – младенческая смертность – число детей, умерших в возрасте до 1 г. (на 1000 чел.);

x_5 – среднее число детей в семье;

x_6 – ожидаемая продолжительность жизни мужчины (в годах);

x_7 – ожидаемая продолжительность жизни женщины (в годах);

x_8 – ВВП на душу населения (в долларах США);

x_9 – плотность населения (количество человек на кв. км)

x_{10} – процент городского населения;

x_{11} – процент грамотных;

x_{12} – прирост населения (% в год).

Таблица 1. Значения признаков

n	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
1	17 800	15	8	7,3	1,90	74	80	16 848	2,3	85	100	1,38
2	8 000	12	11	6,7	1,50	73	79	18 396	94	58	99	0,20
3	33 900	20	9	26,6	2,80	68	75	3 408	12	86	95	1,30
4	125 000	35	11	106,0	4,70	53	53	202	800	16	35	2,40
5	10 300	13	11	19,0	1,88	66	76	6 500	50	65	99	0,32
6	10 100	12	11	7,2	1,70	73	79	17 912	329	96	99	0,20
7	156 600	21	9	66,0	2,70	57	67	2 354	18	75	81	1,28
8	10 000	47	18	118,0	6,94	47	50	357	36	15	18	2,81
9	58 400	13	11	7,2	1,83	74	80	15 974	237	89	99	0,20
10	73 100	27	8	46,0	3,33	63	68	230	218	20	88	1,78
11	6 500	40	19	109,0	5,94	43	47	383	231	29	53	1,63
12	81 200	11	11	6,5	1,47	73	79	17 539	227	85	99	0,36
13	5 600	35	6	45,0	4,90	65	70	1 030	46	44	73	2,73
14	5 800	13	6	5,8	1,40	75	80	14 641	5494	94	77	-0,09
15	60 000	29	9	76,4	3,77	60	63	748	57	44	48	1,95

n	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
16	9 100	46	18	85,0	6,68	44	45	573	11	42	73	2,80
17	3 600	14	9	7,4	1,99	73	78	12 170	51	57	98	0,30
18	39 200	11	9	6,9	1,40	74	81	13 047	77	78	95	0,25
19	58 100	11	10	7,6	1,30	74	81	17 500	188	69	97	0,21
20	29 100	14	8	6,8	1,80	74	81	19 904	2,8	77	97	0,70
21	35 600	24	6	28,0	2,47	69	75	1 538	31	70	87	2,00
22	3 300	26	4	11,0	3,10	76	79	2 031	64	47	93	2,30
23	11 100	17	7	10,2	1,90	74	78	1 382	99	74	94	0,95
24	19 500	29	5	25,6	3,51	66	72	2 995	58	43	78	2,30
25	28 600	29	6	50,0	3,83	66	70	1 062	63	46	50	2,12
26	91 800	28	5	35,0	3,20	69	77	3 604	46	73	87	1,90
27	15 400	13	9	6,3	1,58	75	81	17 245	366	89	99	0,58
28	3 524	16	8	8,9	2,03	73	80	14 381	13	84	99	0,57
29	4 300	13	10	6,3	2,00	74	81	17 755	11	75	99	0,40
30	2 800	28	3	22,0	4,50	70	74	14 193	32	81	68	4,80
31	38 600	14	10	13,8	1,94	69	77	4 429	123	62	99	0,30
32	10 500	12	10	9,2	1,50	71	78	9 000	108	34	85	0,36
33	149 200	13	11	27,0	1,83	64	74	6 680	8,8	74	99	0,20
34	18 000	38	6	52,0	6,67	66	70	6 651	7,7	77	62	3,20
35	23 100	24	6	27,7	2,40	67	73	1 000	189	60	99	1,83
36	2 900	16	6	5,7	1,88	73	79	14 990	4456	100	88	1,20
37	260 800	15	9	8,1	2,06	73	79	23 474	26	75	97	0,99
38	59 400	19	6	37,0	2,10	65	72	1 800	115	22	93	1,40
39	62 200	26	6	49,0	3,21	69	73	3 721	79	61	81	2,02
40	51 800	12	13	20,7	1,82	65	75	2 340	87	67	97	0,05
41	69 800	27	7	51,0	3,35	63	68	867	221	43	90	1,92
42	5 100	13	10	5,3	1,80	72	80	15 877	39	60	100	0,30
43	58 000	13	9	6,7	1,80	74	82	18 944	105	73	99	0,47
44	14 000	23	6	14,6	2,50	71	78	2 591	18	85	93	1,70
45	7 000	12	9	6,2	1,60	75	82	22 384	170	62	99	0,70
46	8 800	14	11	5,7	2,10	75	81	16 900	19	84	99	0,52
47	55 200	45	14	110,0	6,81	51	54	122	47	12	24	3,10
48	43 900	34	8	47,1	4,37	62	68	3 128	35	49	76	2,60
49	45 000	16	6	21,7	1,65	68	74	6 627	447	72	96	1,00
50	125 500	11	7	4,4	1,55	76	82	19 860	330	77	99	0,30

Таблица 2. Варианты заданий

Вариант	Номер результативного признака	Номера факторных признаков
1	12	1, 2, 3, 4, 11

Вариант	Номер результативного признака	Номера факторных признаков
2	12	1, 2, 3, 9, 11
3	8	1, 2, 3, 5, 9
4	8	1, 2, 3, 4, 5
5	8	1, 2, 3, 5, 12
6	7	1, 2, 3, 4, 8
7	7	1, 2, 3, 8, 10
8	6	1, 2, 3, 8, 10
9	6	1, 2, 4, 8, 11
10	6	1, 2, 5, 8, 11

6. Методика выполнения заданий

Запустим программу STATISTICA и откроем файл с исходными данными Lab6_Data.sta. В качестве эндогенной переменной будем рассматривать x_6, в качестве экзогенных – x_1, x_2, x_3, x_4, x_8.

1. Выполним регрессионный анализ в пакете STATISTICA, включив в модель все факторные переменные. Выберем в меню Statistics пункт Multiple Regression. В открывшемся окне Multiple Linear Regression (рис. 1) нажмем на кнопку Variables, выделим в левом списке зависимую переменную (x_6), в правом – независимые (x_1, x_2, x_3, x_4, x_8)и нажмем OK.

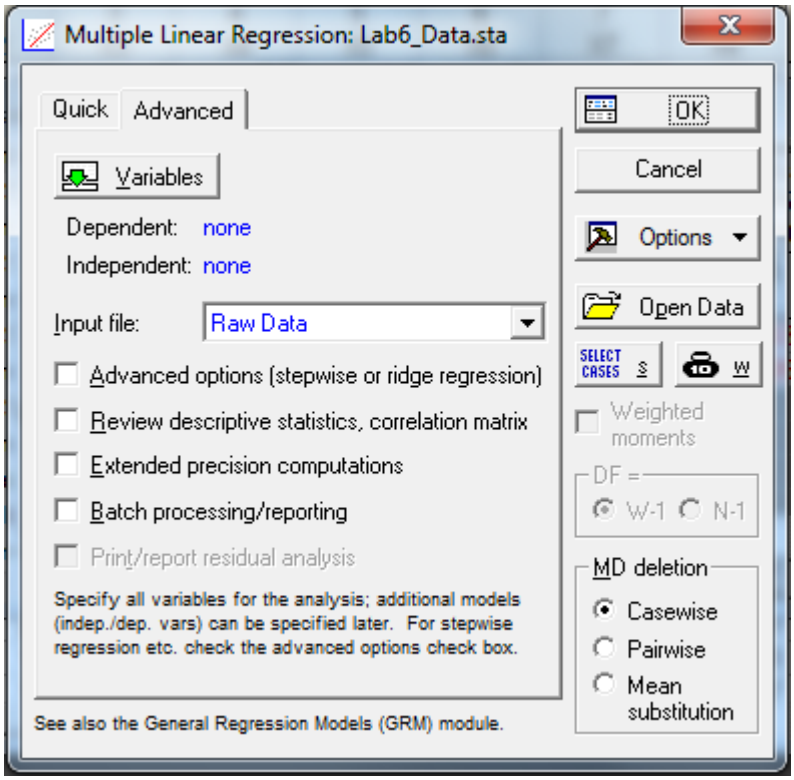


Рис. 1 – Окно Multiple Linear Regression

Результаты множественной регрессии представлены на рис. 2

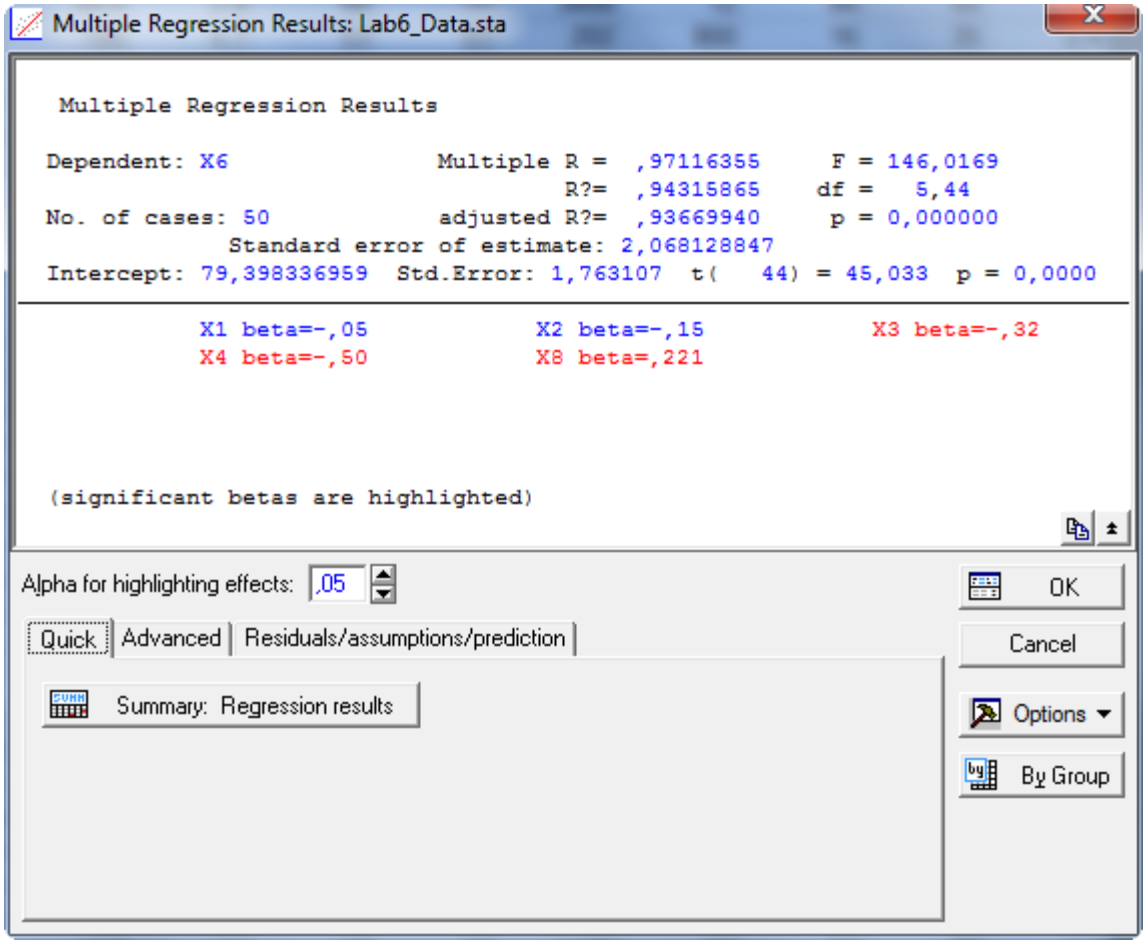


Рис. 2 – Результаты множественной регрессии

На рис. 2 отображена следующая информация:

Dependent: X6 – имя зависимой переменной.

No. of cases: 50 – число наблюдений.

Multiple R= ,97116355 (Множественный R) – коэффициент множественной корреляции между зависимой и независимыми переменными.

R? = ,94315865 (R-квадрат) – коэффициент множественной детерминации (показывает, какую долю дисперсии зависимой переменной объясняет уравнение регрессии).

adjusted R?= ,93669940 (Скорректированный *R*-квадрат) – скорректированный на число степеней свободы коэффициент детерминации

Standard error of estimate: 2,068128847 – Стандартная ошибка оценки, измеряет рассеяние наблюдаемых значений относительно линии регрессии.

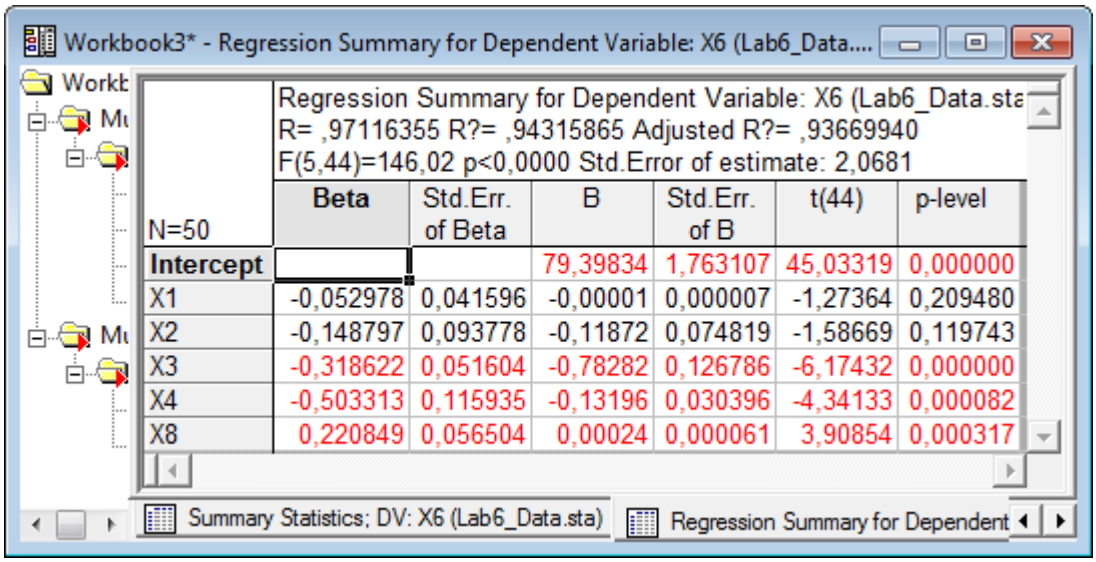
Intercept: 79,398336959 – Свободный член.

Std.Error: 1,763107 – Стандартная ошибка оценки свободного члена.

t(44) = 45,033 p = 0,0000 – *t*-статистика с числом степеней свободы и соответствующее ей *p*-значение, используются для проверки гипотезы о равенстве нулю свободного члена в уравнении регрессии.

F = 146,0169 – F-критерий, **df = 5,44** – степень свободы и **p = 0,0000** – *p*-значение – используются в качестве общего *F*-критерия для проверки гипотезы о зависимости предикторов и отклика.

В поле параметров модели отображаются и выделяются красным статистически значимые коэффициенты регрессии (бета-коэффициенты) для переменных, включенных в анализ. Критерий статистической значимости (выделяемый уровень значимости - альфа) может быть выбран из интервала [0,0001; 0,5). По умолчанию он равен 0,05.



N=50	Beta	Std.Err. of Beta	B	Std.Err. of B	t(44)	p-level
Intercept			79,39834	1,763107	45,03319	0,000000
X1	-0,052978	0,041596	-0,00001	0,000007	-1,27364	0,209480
X2	-0,148797	0,093778	-0,11872	0,074819	-1,58669	0,119743
X3	-0,318622	0,051604	-0,78282	0,126786	-6,17432	0,000000
X4	-0,503313	0,115935	-0,13196	0,030396	-4,34133	0,000082
X8	0,220849	0,056504	0,00024	0,000061	3,90854	0,000317

Рис. 3 – Таблица результатов множественной регрессии

Для отображения более подробных результатов нажмем кнопку Summary: Regression results. Эта опция строит таблицы результатов со стандартизованными (Beta) и нестандартизованными (B) регрессионными коэффициентами, их стандартными ошибками (Std.Err.), *t*-статистиками и уровнями значимости (p-level) (рис. 3).

2. Построим матрицу парных коэффициентов корреляции. Для этого вернемся к окну Multiple Linear Regression, где в разделе Variables уже заданы переменные, поставим галочку около пункта Review descriptive statistics, correlation matrix и нажмем OK, появится окно Review Descriptive Statistics (рис. 4). Нажмем на кнопку Correlations для просмотра корреляционной матрицы (рис. 5).

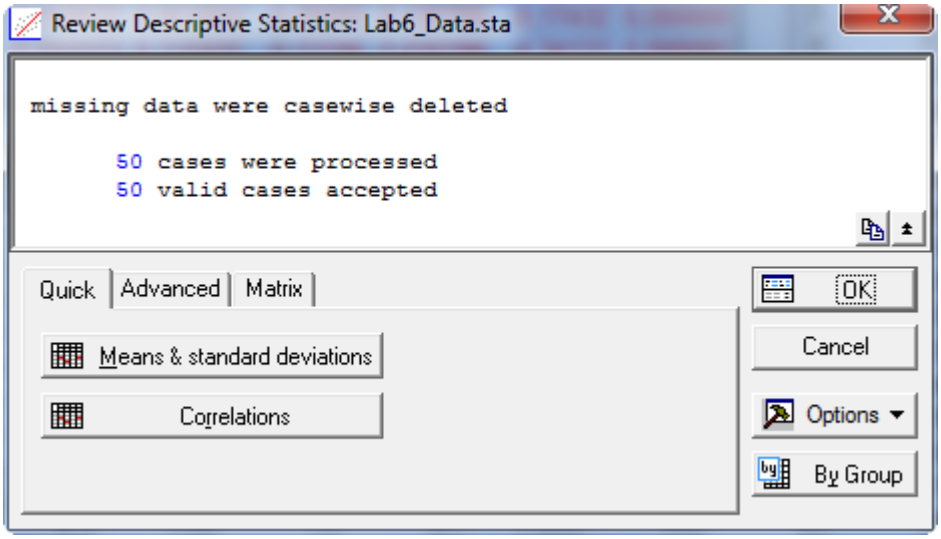
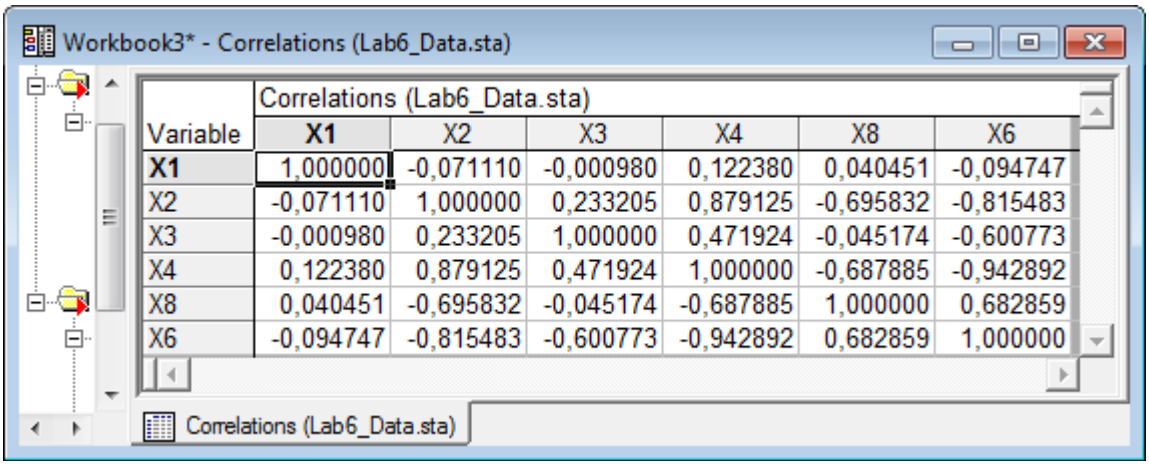


Рис. 4 – Окно Review Descriptive Statistics



Variable	X1	X2	X3	X4	X8	X6
X1	1,000000	-0,071110	-0,000980	0,122380	0,040451	-0,094747
X2	-0,071110	1,000000	0,233205	0,879125	-0,695832	-0,815483
X3	-0,000980	0,233205	1,000000	0,471924	-0,045174	-0,600773
X4	0,122380	0,879125	0,471924	1,000000	-0,687885	-0,942892
X8	0,040451	-0,695832	-0,045174	-0,687885	1,000000	0,682859
X6	-0,094747	-0,815483	-0,600773	-0,942892	0,682859	1,000000

Рис. 5 – Корреляционная матрица

3. Выведем для всех факторов значения R_j^2, $1-R_j^2$, частный и получастный коэффициенты корреляции. Для этого вернемся к окну Multiple Linear Regression, уберем галочку около пункта Review descriptive statistics, correlation matrix и нажмем OK. В появившемся окне результатов множественной регрессии перейдем к вкладке Advanced и нажмем на кнопку Partial Correlations. Результат представлен на рис. 6: коэффициенты частной и получастной корреляции выведены соответственно в столбцах Partial Cor. и Semipart Cor., R_j^2 – в столбце R-square, $1-R_j^2$– Tolerance.

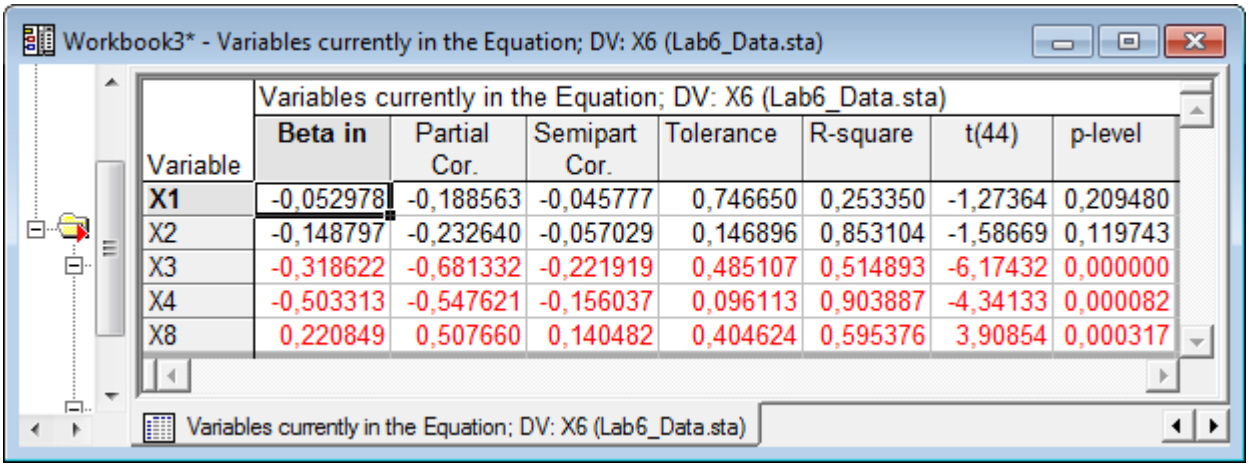


Рис. 6 – Таблица результатов Частной корреляции

4. Для отбора факторов, которые целесообразно включать в модель, воспользуемся алгоритмом пошагового включения:

1) С помощью выведенной ранее корреляционной матрицы (рис. 5), выберем фактор, наиболее сильно коррелированный с результативным показателем x_6, и построим модель парной регрессии.

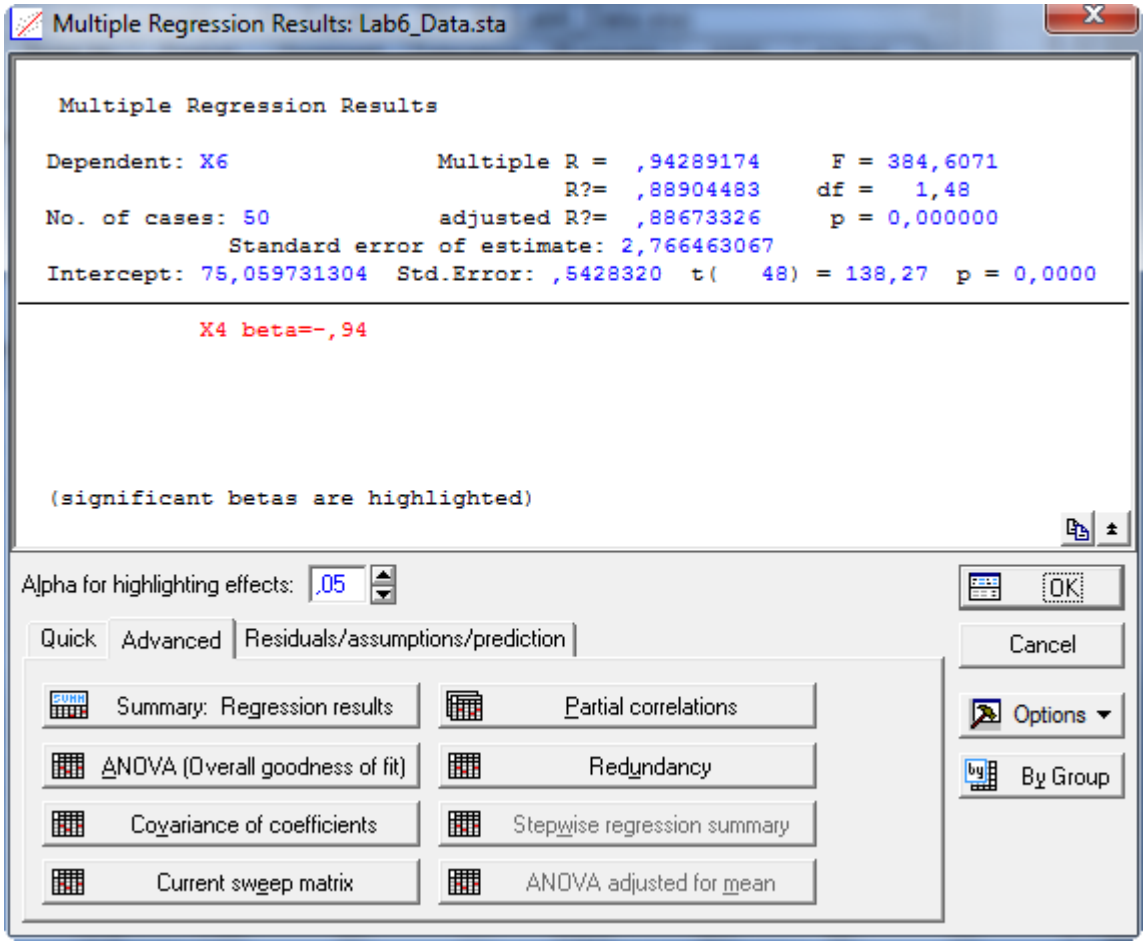


Рис. 7 – Результаты построения однофакторной модели.

Наибольшее значение коэффициента корреляции x_6имеет с фактором x_4 ($r_{x_6x_4} = -0{,}943$), для них и будем строить уравнение линейной регрессии. Вернемся к окну Multiple Linear Regression, в списке независимых переменных раздела Variables оставим только x_4. Нажмем OK. Результаты множественной регрессии представлены на рис. 7. Выведем более подробную информацию о модели, нажав на кнопку Summary: Regression results (рис. 8).

Regression Summary for Dependent Variable: X6 (Lab6_Data.sta)
R= ,94289174 R?= ,88904483 Adjusted R?= ,88673326
F(1,48)=384,61 p<0,0000 Std.Error of estimate: 2,7665

N=50	Beta	Std.Err. of Beta	B	Std.Err. of B	t(48)	p-level
Intercept			75,05973	0,542832	138,2743	0,000000
X4	-0,942892	0,048079	-0,24721	0,012605	-19,6114	0,000000

Рис. 8 – Подробная информация об однофакторной модели

Тогда уравнение в нестандартизованном виде: $y_x = 75{,}06 - 0{,}25x_4$

2) Оценим все варианты регрессионных моделей с двумя факторами:

$$y_x = b_0 + b_4x_4 + b_mx_m, m = 1,2,3,8.$$

Возвращаемся к окну MultipleLinearRegression, добавляем еще один независимый фактор. Результаты для моделей с двумя факторами представлены на рис. 9-12, а также сведены в таблице 3.

Regression Summary for Dependent Variable: X6 (Lab6_Data.sta)
R= ,94312115 R?= ,88947751 Adjusted R?= ,88477442
F(2,47)=189,13 p<0,0000 Std.Error of estimate: 2,7903

N=50	Beta	Std.Err. of Beta	B	Std.Err. of B	t(47)	p-level
Intercept			74,93296	0,622178	120,4366	0,000000
X1	0,020958	0,048860	0,00000	0,000008	0,4290	0,669919
X4	-0,945457	0,048860	-0,24788	0,012810	-19,3503	0,000000

Рис. 9 – Результаты регрессионного анализа для y, x_1, x_4

Regression Summary for Dependent Variable: X6 (Lab6_Data.sta)
R= ,94331315 R?= ,88983970 Adjusted R?= ,88515203
F(2,47)=189,83 p<0,0000 Std.Error of estimate: 2,7857

N=50	Beta	Std.Err. of Beta	B	Std.Err. of B	t(47)	p-level
Intercept			74,46713	1,155111	64,46751	0,000000
X2	0,059157	0,101582	0,04720	0,081046	0,58235	0,563112
X4	-0,994898	0,101582	-0,26084	0,026633	-9,79403	0,000000

Рис. 10 – Результаты регрессионного анализа для y, x_2, x_4

Regression Summary for Dependent Variable: X6 (Lab6_Data.sta)
R= ,95930874 R?= ,92027326 Adjusted R?= ,91688064
F(2,47)=271,26 p<0,0000 Std.Error of estimate: 2,3699

N=50	Beta	Std.Err. of Beta	B	Std.Err. of B	t(47)	p-level
Intercept			78,70225	0,967959	81,3075	0,000000
X3	-0,200440	0,046716	-0,49246	0,114775	-4,2906	0,000088
X4	-0,848299	0,046716	-0,22241	0,012248	-18,1588	0,000000

Рис. 11 – Результаты регрессионного анализа для y, x_3, x_4

Regression Summary for Dependent Variable: X6 (Lab6_Data.sta)
R= ,94407229 R?= ,89127250 Adjusted R?= ,88664579
F(2,47)=192,64 p<0,0000 Std.Error of estimate: 2,7675

N=50	Beta	Std.Err. of Beta	B	Std.Err. of B	t(47)	p-level
Intercept			74,09908	1,119481	66,1906	0,000000
X4	-0,898160	0,066266	-0,23548	0,017374	-13,5538	0,000000
X8	0,065027	0,066266	0,00007	0,000071	0,9813	0,331466

Рис. 12 – Результаты регрессионного анализа для y, x_4, x_8

Таблица 3. Характеристики двухфакторных моделей

Включае-мый фактор	Уравнение регрессии	Коэффициент детерминации	Прирост коэффициента детерминации
----	$y_x = 75{,}06 - 0{,}25x_4$	0,88904483	----
x_1	$y_x = 74{,}9 - 0{,}00x_1 - 0{,}25x_4$	0,88947751	0,00043268
x_2	$y_x = 74{,}5 - 0{,}05x_2 - 0{,}26x_4$	0,88983970	0,00079487
x_3	$y_x = 78{,}7 - 0{,}49x_3 - 0{,}22x_4$	0,92027326	0,03122843
x_8	$y_x = 74{,}1 - 0{,}24x_4 + 0{,}00007x_8$	0,89127250	0,00222767

Выбираем фактор, обеспечивающий максимальный прирост коэффициента детерминации – x_3, тогда двухфакторное уравнение регрессии будет иметь вид:

$$y_x = 78{,}7 - 0{,}49x_3 - 0{,}22x_4$$

Оценим значимость включения фактора в модель с помощью теста Фишера: коэффициент детерминации для однофакторной модели: $R^2_{\text{без } x_3} = 0{,}88904483$, коэффициент детерминации для двухфакторной модели: $R^2 = 0{,}92027326$, $\Delta R^2 = R^2 - R^2_{\text{без } x_3} = 0{,}03122843$

Выдвигаем гипотезу о том, что включение в модель фактораX_3в генеральной совокупности не приводит к увеличению объясненной вариации результативной переменной –H_0: $\Delta R^2 = 0$.Рассчитываем тестовую статистику:

$$F = \frac{\Delta R^2 (n - m - 1)}{1 - R^2} = \frac{0{,}03122843(50 - 2 - 1)}{1 - 0{,}92027326} = 18{,}40$$

В случае справедливости нулевой гипотезы F является случайной величиной, распределенной по закону Фишера с 1, n-m-1=47 степенями свободы. Табличное значение можно высчитать с помощью вероятностного калькулятора. Для этого необходимо выбрать в меню Statistics / Probability Calculator / Distribution. В поле Distribution – выбираем критерий (в данном случае критерий Фишера – F(Fisher)), в поле p – 0,95, в полях df – число степеней свободы:df1 – 1, df2 – 47. При нажатии на кнопку Compute в поле F появится табличное значение $F_{табл}$ = 4,047 (рис. 13).

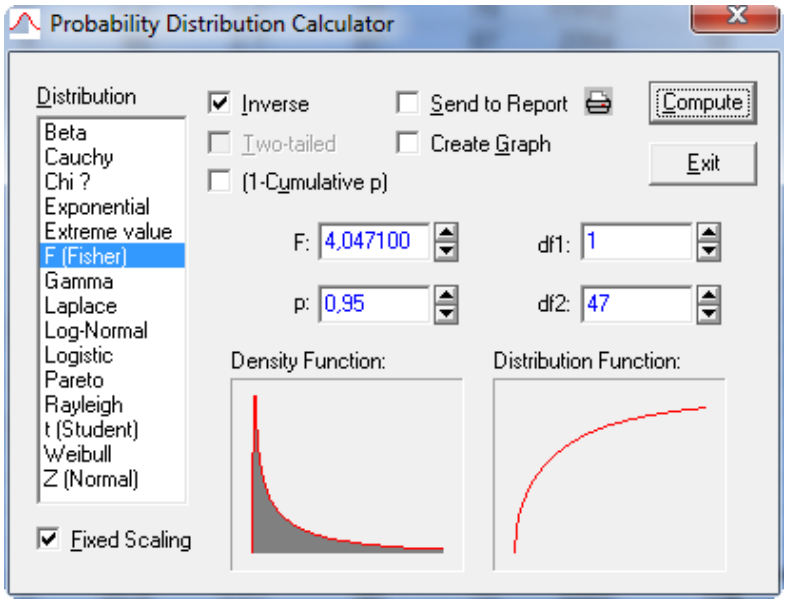


Рис. 13 – Вероятностный калькулятор

Поскольку $F_{табл}$<F, следовательно, нулевую гипотезу следует отвергнуть и принять альтернативную – $\Delta R^2 > 0$, т.е. включение в модель фактораx_3ив генеральной совокупности приводит к увеличению объясненной вариации результативной переменной.

3) Аналогично оценим все варианты регрессионных моделей с тремя факторами: $y_x = b_0 + b_3x_3 + b_4x_4 + b_mx_m, m = 1,2,8$. Результаты для моделей с тремя факторами представлены в таблице 4.

Таблица 4. Характеристики двухфакторных моделей

Вклю-чаемый фактор	Уравнение регрессии	Коэффици-ент детер-минации	Прирост коэффициента детерминации
----	$y_x = 78,7 - 0,49x_3 - 0,22x_4$	0,92027326	----
x_1	$y_x = 78,6 + 0,00x_1 - 0,49x_3 - 0,22x_4$	0,92035353	0,00008027
x_2	$y_x = 80,5 - 0,1x_2 - 0,56x_3 + 0,19x_4$	0,92313229	0,00285903
x_8	$y_x = 76,96 - 0,68x_3 - 0,17x_4 + 0,0002x_8$	0,93938683	0,01911357

Выбираем фактор, обеспечивающий максимальный прирост коэффициента детерминации – x_8, тогда трехфакторное уравнение регрессии будет иметь вид:

$$y_x = 76{,}96 - 0{,}68x_3 - 0{,}17x_4 + 0{,}0002x_8$$

Оценим значимость включения фактора в модель с помощью теста Фишера:

$$\Delta R^2 = 0{,}93938683 - 0{,}92027326 = 0{,}01911357$$

$$F = \frac{0{,}01911357(50 - 3 - 1)}{1 - 0{,}93938683} = 14{,}51$$

$$F_{табл} = 4{,}052.$$

Поскольку $F_{табл} < F$, следовательно, нулевую гипотезу следует отвергнуть и принять альтернативную – $\Delta R^2 > 0$, т.е. включение в модель фактора x_8 оправдано.

4) Оценим все варианты регрессионных моделей с четырьмя факторами: $y_x = b_0 + b_3x_3 + b_4x_4 + b_8x_8 + b_mx_m, m = 1{,}2$. Результаты для моделей с тремя факторами приведены в таблице 5.

Таблица 5. Характеристики двухфакторных моделей

Включаемый фактор	Уравнение регрессии	Коэффициент детерминации	Прирост коэффициента детерминации
----	$y_x = 76{,}96 - 0{,}68x_3 - 0{,}17x_4 + 0{,}0002x_8$	0,93938683	----
x_1	$y_x = 77{,}07 + 0{,}00x_1 - 0{,}69x_3 - 0{,}17x_4 - 0{,}0002x_8$	0,93990629	0,00051946
x_2	$y_x = 78{,}37 - 0{,}07x_2 - 0{,}73x_3 - 0{,}15x_4 + 0{,}0002x_8$	0,94106307	0,00167624

Выбираем фактор, обеспечивающий максимальный прирост коэффициента детерминации – x_2, тогда трехфакторное уравнение регрессии будет иметь вид:

$$y_x = 79{,}12 + 0{,}11x_2 - 0{,}77x_3 - 0{,}14x_4 - 0{,}0002x_8$$

Оценим значимость включения фактора в модель с помощью теста Фишера:

$$\Delta R^2 = 0{,}94106307 - 0{,}93938683 = 0{,}00167624$$

$$F = \frac{0{,}00167624(50 - 4 - 1)}{1 - 0{,}94106307} = 1{,}28$$

$$F_{табл} = 4{,}057.$$

Поскольку $F_{табл} > F$, следовательно, нулевую гипотезу следует принять, т.е. включение в модель фактора x_2 незначительно увеличивает долю объясненной вариации результативной переменной.

5) Проанализируем рассмотренные модели:

$$y_x = 75{,}06 - 0{,}25x_4$$

$$y_x = 78{,}7 - 0{,}49x_3 - 0{,}22x_4$$
$$y_x = 76{,}96 - 0{,}68x_3 - 0{,}17x_4 + 0{,}0002x_8$$

Все факторы значимы (во всех тестах Стьюдента нулевая гипотеза отвергалась). Выбираем модель, в который коэффициент детерминации максимален (0,93938683): $y_x = 76{,}96 - 0{,}68x_3 - 0{,}17x_4 + 0{,}0002x_8$.

5. Проверим правильность наших рассуждений с помощью автоматической пошаговой регрессии в пакете Statistica:

В стартовой панели Множественной регрессии (рис. 1), щелкнем на кнопке Variables и выберем переменную x_6 в качестве зависимой переменной, а переменные x_1, x_2, x_3, x_4, x_8 в качестве независимых. Отметим галочкой пункт Advanced options (stepwise or ridge regression) и нажмем OK. Откроется окно Model Definition – определение модели (рис. 14)

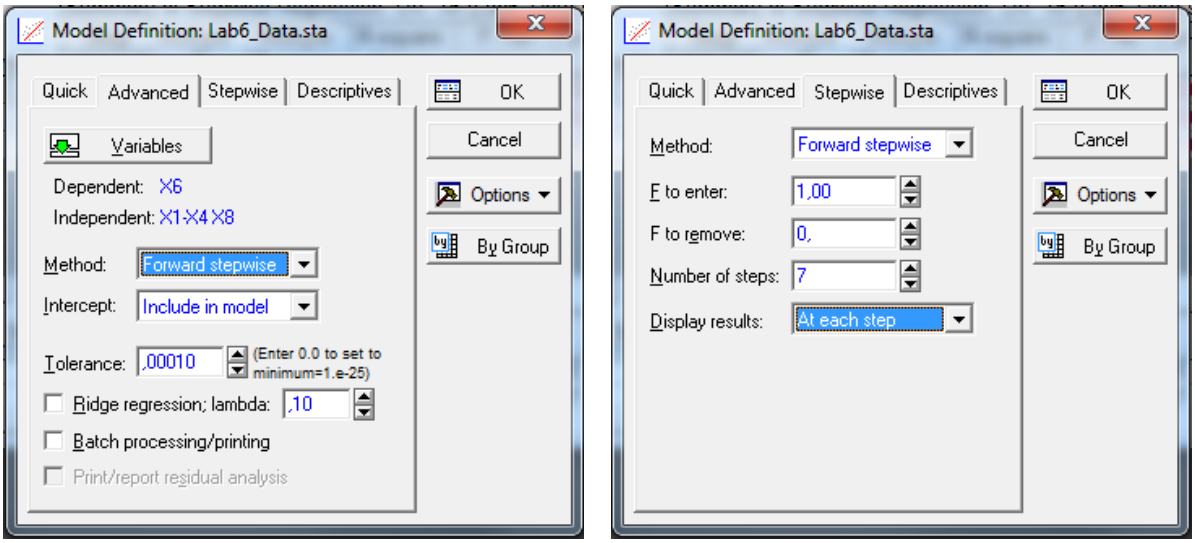


Рис. 14 – Диалоговое окно Model Definition

На выбор предлагаются следующие процедуры для проведения анализа данных: Standard (стандартная), Forward stepwise (пошаговая с включением) или Backward stepwise (пошаговая с исключением). Выберем процедуру с пошаговым включением.

Далее перейдем к вкладке Stepwise. Здесь помимо выбора метода, можно поменять следующие настройки:

F to enter (F-включить) – определяет насколько значимым должен быть вклад переменной в регрессию, чтобы она была добавлена в уравнение. По умолчанию для метода пошаговой с включением установлено значение 1,00.

F to remove (F-исключить) – определяет насколько «незначимым» должен быть вклад переменной в регрессию, чтобы она могла быть исключена из

регрессионного уравнения. По умолчанию для метода пошаговой с включением установлено значение 0.

В общем случае чем большие значения заданы в этих полях, тем больше переменных будет включено в модель, при этом Ftoremovевсегда должно быть меньше, чем Ftoenter.

Number of steps (число шагов) – максимальное число шагов, которое процедура оценивания выполнит до вывода окончательных результатов пошагового регрессионного анализа.

Display results (Отображение результатов) – можно выбрать для просмотра только окончательных (итоговых) результатов пошагового регрессионного анализа (Summary only) или результатов каждого шага анализа (At each step).

Оставим опции F и количество шагов без изменения, установим переключатель на Summary only и нажмем OK. Окно результатов пошаговой регрессии (рис. 15) будет отличаться от результатов простой множественной регрессии (рис. 2) наличием информации о шаге, на котором было получено окончательное решение.

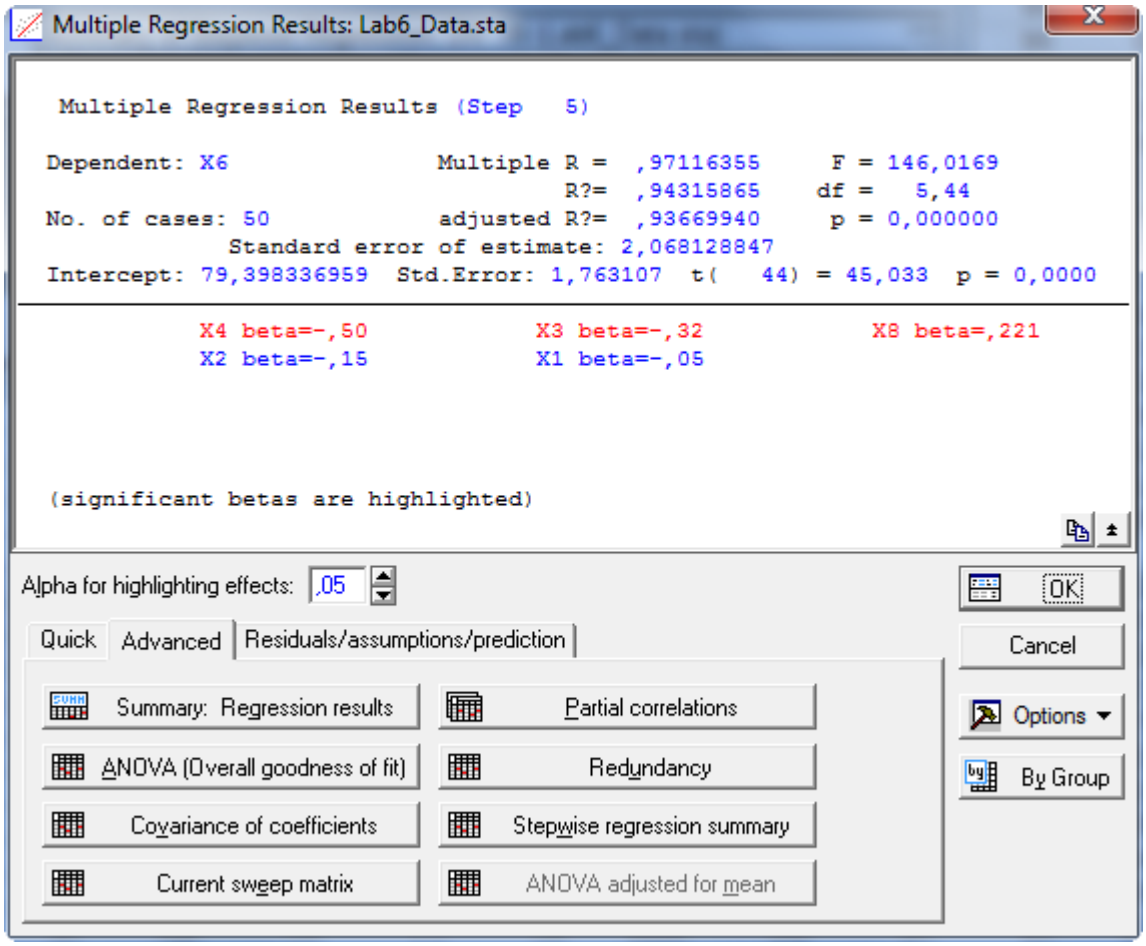
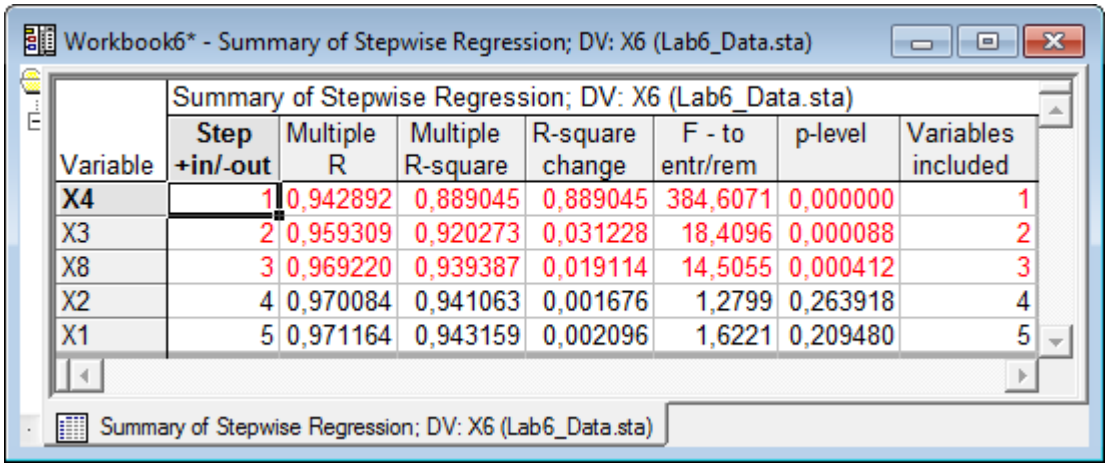


Рис. 15 – Результаты пошаговой множественной регрессии

В нашем случае решение было получено на пятом шаге (Step 5), при этом в модель были включены все 5 фактором. Подробную историю включения факторов и их характеристики можно увидеть, нажав на кнопку Stepwise regression summary на закладке Advanced (рис. 16).



Variable	Step +in/-out	Multiple R	Multiple R-square	R-square change	F - to entr/rem	p-level	Variables included
X4	1	0,942892	0,889045	0,889045	384,6071	0,000000	1
X3	2	0,959309	0,920273	0,031228	18,4096	0,000088	2
X8	3	0,969220	0,939387	0,019114	14,5055	0,000412	3
X2	4	0,970084	0,941063	0,001676	1,2799	0,263918	4
X1	5	0,971164	0,943159	0,002096	1,6221	0,209480	5

Рис. 16 – Подробные результаты пошаговой множественной регрессии

На рис. 6 факторы расположены в порядке включения их в модель. Также по столбцам приведена следующая информация:

Multiple R – коэффициент множественной корреляции после включения соответствующего фактора,

Multiple R-square – коэффициент множественной детерминации после включения соответствующего фактора,

R-square change – приращение коэффициента множественной детерминации на каждом шаге,

F-to ent/rem – F-критерий для включения/исключения соответствующего фактора,

p-level – уровень значимости для F-статистики.

Как видно из рис. 15-16, результаты автоматической пошаговой регрессии не совпадают с полученными нами ранее. Это объясняется тем, что в модель включались переменные с незначимыми F (p-level для x_1 и x_2 превышает 0,05).

Повысим порог F для включения факторов в модель. Для этого вернемся к диалоговому окну Model Definition, изменим значения F to enter на 4,10, F to remove на 4,00 и снова запустим пошаговую регрессию. Результаты представлены на рис. 17. На этот раз в модель были включены только 3 значимых фактора и результат совпал с полученным при расчетах «вручную».

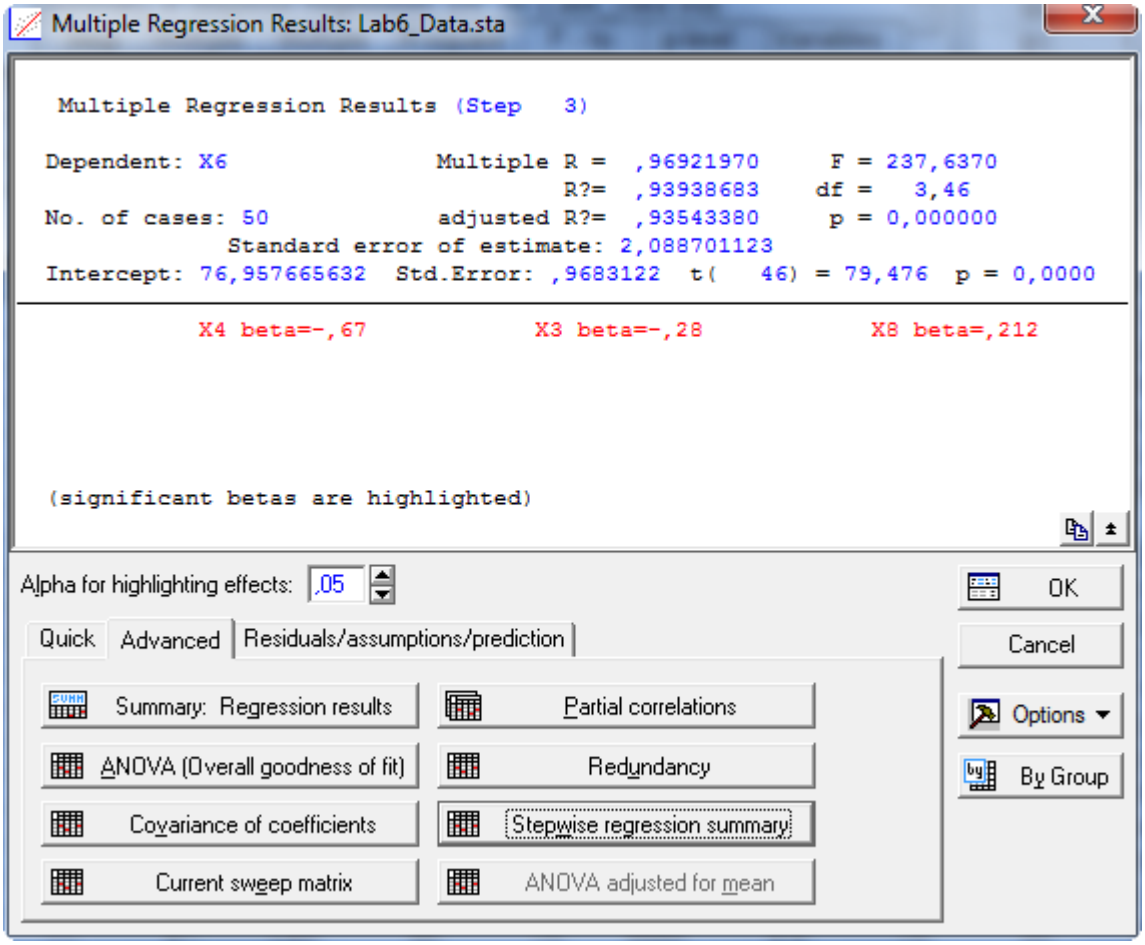


Рис. 17 – Результаты пошаговой множественной регрессии

6. Для анализа качества построенной модели можно использовать информацию из окна результатов множественной регрессии (рис. 2), таблицы результатов множественной регрессии (рис. 3), также можно в проводнике слева от таблицы выбрать пункт Summary Statistics (рис. 18).

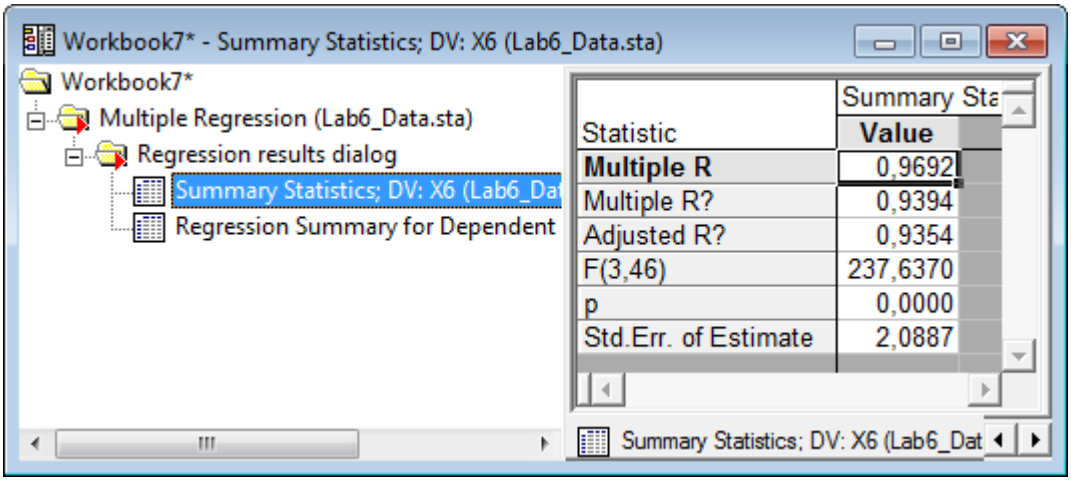


Рис. 18 – Summary Statistics

7. Для анализа остатков в окне результатов множественной регрессии (рис. 2) нужно перейти на последнюю вкладку – Residual/assumptions/prediction и нажать на кнопку Perform residual analysis. Откроется окно Residual Analysis (рис. 19)

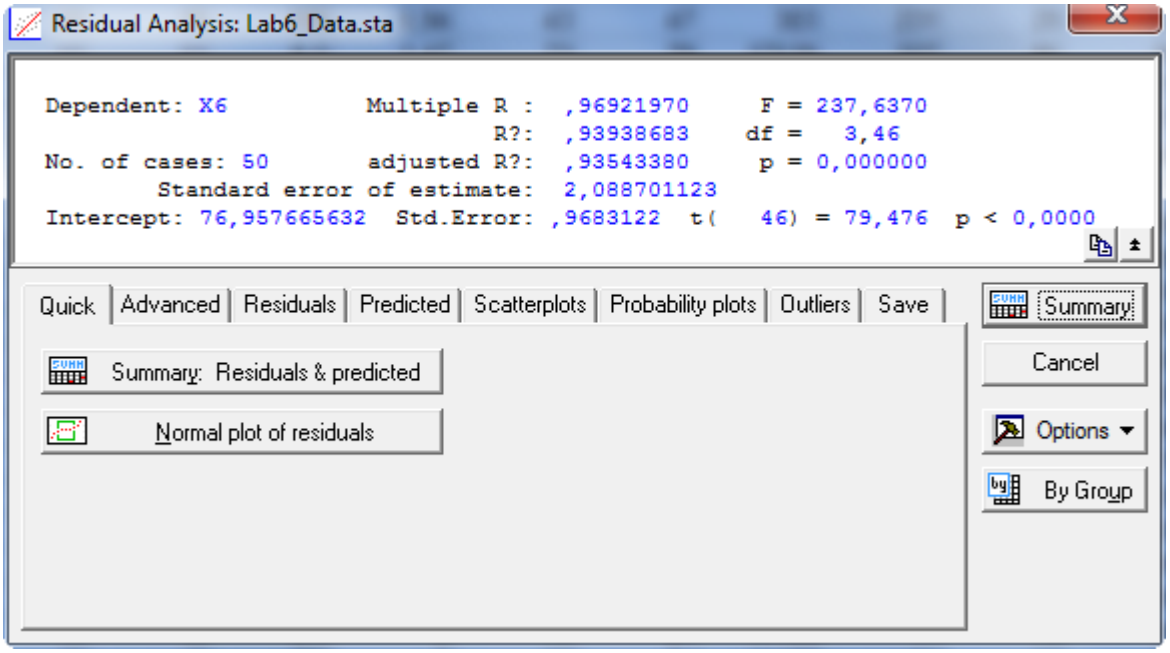


Рис. 19 – Окно Residual Analysis

1) Для построения графика зависимости остатков от значений эндогенной переменной нужно перейти к вкладке Scatterplots и нажать на кнопку Observed vs. Residuals (результат – правый нижний график на рис. 20).
2) Для построения графиков зависимости остатков от значений экзогенных переменных нужно перейти к вкладке Residuals, нажать на кнопку Residuals vs. Independent var., выделить нужную переменную и нажать OK (результаты – верхние и левый нижний график на рис. 20).
3) Для Определения значения статистики Дарбина-Уотсона нужно перейти к вкладке Advanced и нажать на кнопку Durbin-Watson statistic (результат на рис. 21).
4) Для анализа закона распределения остатков можно построить диаграмму остатков (вкладка Residuals, кнопка Histogram of Residuals) и нормальный график остатков (вкладка Probability plots, кнопка Normal plot of residuals). Полученные графики, представленные на рис. 22, наглядно показывают, что распределение остатков близко к нормальному.

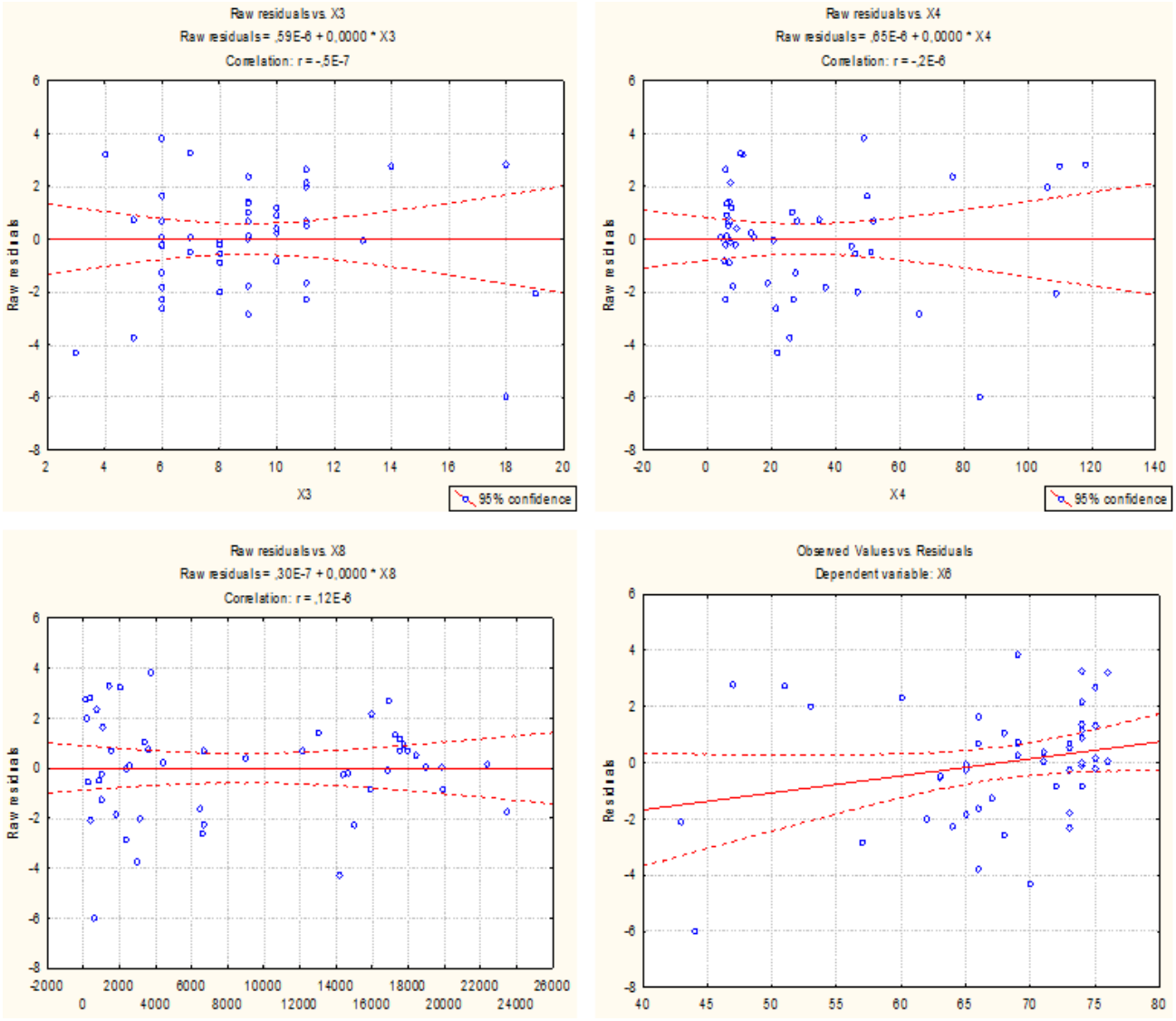


Рис. 20 – Графики зависимости остатков от переменных x_3, x_4, x_8, x_6

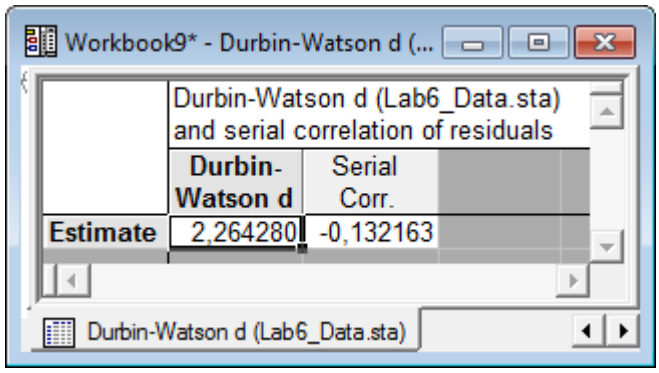


Рис. 21 – Результат расчета статистики Дарбина-Уотсона

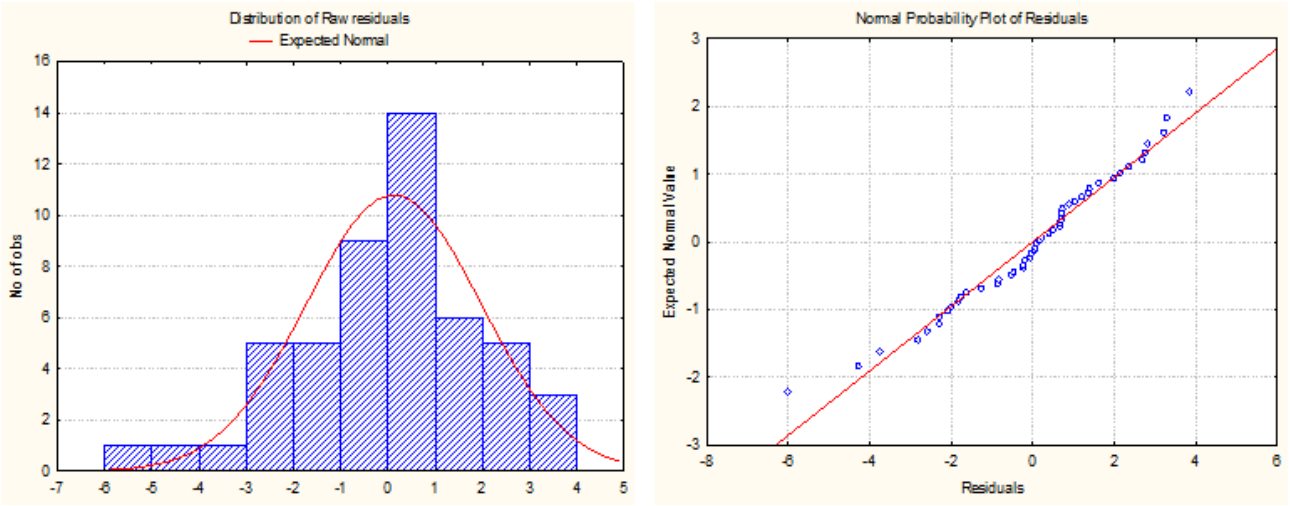


Рис. 22 – Диаграмма остатков и нормальный график остатков (Q-Q график)

7. Контрольные вопросы

1. Что такое мультиколлинеарность и каковы ее последствия?
2. Каковы основные способы выявления мультиколлинеарности?
3. В чем отличие парных, частных и множественных коэффициентов корреляции?
4. Опишите алгоритм пошагового включения.
5. Опишите алгоритм пошагового исключения.
6. Какой критерий используется при принятии решения о включении (исключении) фактора в модель в пошаговых алгоритмах?
7. В чем удобство вероятностного калькулятора пакета STATISTICA?
8. Как с помощью пакета STATISTICA определить параметры уравнения регрессии и оценить их значимость?
9. Как с помощью пакета STATISTICA оценить значимость уравнения регрессии?
10. Какие возможности для анализа остатков предоставляет пакет STATISTICA?

Лабораторная работа №6.

Моделирование одномерных временных рядов с помощью пакета MSExcel.

1. Цель и задачи лабораторной работы

Цель работы: изучить возможности пакета MS Excel для построения мультипликативных и аддитивных моделей временных рядов.

Задачи:

– научиться выявлять структуру временных рядов с помощью автокорреляционной функции;

– приобрести навыки моделирования сезонных и циклических колебаний методом скользящей средней;

– приобрести навыки моделирования сезонных и циклических колебаний постоянной амплитуды с помощью фиктивных переменных;

– приобрести навыки прогнозирования уровней временного ряда.

2. Теоретическая часть

2.1. Основные элементы временного ряда

Временной ряд – это 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;
- факторы, формирующие циклические колебания ряда;
- случайные факторы.

Соответственно случаев фактический уровень временного ряда в большинстве можно представить как сумму или произведение трендовой, циклической и случайной компонент.

Модель, в которой временной ряд представлен как сумма перечисленных компонент, называется ***аддитивной моделью временного ряда***.

Модель, в которой временной ряд представлен как произведение перечисленных компонент, называется ***мультипликативной моделью временного ряда***.

Основная задача эконометрического исследования отдельного временного ряда – выявление и придание количественного выражения каждой из перечисленных выше компонент с тем, чтобы использовать полученную информацию для прогнозирования будущих значений ряда или при построении моделей взаимосвязи двух или более временных рядов.

2.2. Автокорреляция уровней временного ряда

При наличии во временном ряде тенденции и циклических колебаний значения каждого последующего уровня ряда зависят от предыдущих. Корреляционную зависимость между последовательными уровнями временного ряда называют ***автокорреляцией уровней ряда***.

Количественно ее можно измерить с помощью линейного коэффициента корреляции между уровнями исходного временного ряда y_t и уровнями этого ряда, сдвинутыми на несколько шагов во времени $y_{t-\tau}$.

Число периодов, по которым рассчитывается коэффициент автокорреляции, называют лагом. С увеличением лага число пар значений, по которым рассчитывается коэффициент автокорреляции, уменьшается. Максимальный лаг должен быть не больше (n/4).

Коэффициент автокорреляции уровней ряда первого порядка, измеряющий зависимость между соседними уровнями ряда y_t и y_{t-1}, т.е. при лаге 1, рассчитывается по формуле:

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

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

Аналогично определяются коэффициенты автокорреляции второго и более высоких порядков. Так, коэффициент автокорреляции второго порядка характеризует тесноту связи между уровнями y_t и y_{t-2} и определяется по формуле:

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

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

Коэффициент автокорреляции характеризует тесноту только линейной связи текущего и анализируемого уровней ряда. Поэтому по нему можно судить о наличии линейной (или близкой к линейной) тенденции. Для некоторых временных рядов, имеющих сильную нелинейную тенденцию, коэффициент автокорреляции уровней исходного ряда может приближаться к нулю.

Последовательность коэффициентов автокорреляции уровней первого, второго и т. д. порядков называют ***автокорреляционной функцией*** временного ряда. График зависимости ее значений от величины лага называется ***коррелограммой***.

При помощи анализа автокорреляционной функции и коррелограммы можно выявить структуру ряда.

Анализ структуры ряда можно проводить следующим образом:

- если наиболее высоким оказался коэффициент автокорреляции первого порядка, исследуемый ряд содержит только тенденцию;
- если наиболее высоким оказался коэффициент автокорреляции порядка τ, ряд содержит циклические колебания с периодичностью в τ моментов времени;
- если ни один из коэффициентов автокорреляции не является значимым, можно сделать одно из предположений относительно структуры ряда: 1) ряд не содержит тенденции и циклических колебаний, а включает только случайную компоненту; 2) ряд содержит сильную нелинейную тенденцию.

2.3. Моделирование тенденции временного ряда

Одним из наиболее распространенных способов моделирования тенденции временного ряда является построение аналитической функции, характеризующей зависимость уровней ряда от времени (тренда). Этот способ называют аналитическим выравниванием временного ряда. Для построения трендов чаще всего применяются следующие функции: линейный тренд, гипербола, экспоненциальный тренд, степенная функция, парабола второго и более высоких порядков. Параметры каждого из них можно определить обычным МНК, используя в качестве независимой переменной время t = 1, 2, …, n, а в качестве зависимой переменной – фактические уровни временного ряда y_t. Для нелинейных трендов предварительно проводят стандартную процедуру их линеаризации.

2.4. Моделирование сезонных и циклических колебаний методом скользящей средней

Существует несколько подходов к анализу структуры временных рядов, содержащих циклические колебания. Простейший подход – расчет значений сезонной компоненты методов скользящей средней и построение аддитивной или мультипликативной модели временного ряда.

Общий вид аддитивной модели: $Y = T + S + E$.

Общий вид мультипликативной модели: $Y = T \cdot S \cdot E$.

Здесь Y – уровни временного ряда, T – трендовая компонента, S – сезонная компонента, E – случайная компонента.

Выбор одной из двух моделей осуществляется на основе анализа структуры сезонных колебаний. Если амплитуда колебаний приблизительно постоянна, строят аддитивную модель временного ряда, в которой значения сезонной компоненты полагаются постоянными для различных циклов. Если амплитуда сезонных колебаний возрастает или уменьшается, строят

мультипликативную модель временного ряда, которая ставит уровни ряда в зависимость от значений сезонной компоненты.

Построение аддитивной и мультиплиакативной моделей сводится к расчету значений *T*, *S*, *E* для каждого уровня ряда.

Основные этапы построения модели:

1. Выравнивание исходного ряда методом скользящей средней.
2. Расчет значений сезонной компоненты *S*.
3. Устранение сезонной компоненты из исходных уровней ряда и получение выровненных данных $(T + E)$ в аддитивной или $(T \cdot E)$ в мультипликативной модели.
4. Аналитическое выравнивание уровней $(T + E)$ или $(T \cdot E)$ и расчет значений *T* с использованием полученного уравнения тренда.
5. Расчет полученных по модели значений $(T + S)$ или $(T \cdot S)$.
6. Расчет абсолютных и/или относительных ошибок

Если полученные значения ошибок не содержат автокорреляции, ими можно заменить исходные уровни ряда и в дальнейшем использовать временной ряд ошибок *E* для анализа взаимосвязи исходного ряда и других временных рядов.

2.5. Применение фиктивных переменных для моделирования сезонных колебаний

Помимо метода скользящей средней используются и другие подходы. Один из них – построение модели регрессии с включением фактора времени и фиктивных переменных. Количество фиктивных переменных в такой модели должно быть на единицу меньше числа периодов времени внутри одного цикла колебаний. Каждая фиктивная переменная отражает циклическую компоненту временного ряда для какого-либо одного периода. Она равна единице для данного периода и нулю для всех остальных периодов.

Пусть имеется временной ряд, содержащий циклические колебания периодичностью *k*. Модель регрессии с фиктивными переменными для этого ряда будет иметь вид:

$$y_t = a + b \cdot t + c_1 x_1 + \cdots + c_j x_j + \cdots + c_{k-1} x_{k-1} + \varepsilon_t$$

где $x_j = \begin{cases} 1 \text{ для каждого } j \text{ внутри каждого цикла,} \\ 0 \text{ во всех остальных случаях.} \end{cases}$

Фиктивные переменные позволяют дифференцировать величину свободного члена уравнения регрессии для каждого периода времени. Она составит:

для 1 периода ($a + c_1$);

для 2 периода ($a + c_2$);

…

для *k*-1 периода ($a + c_{k-1}$);

для *k* периода *a*.

Параметр *b* в этой модели характеризует среднее абсолютное изменение уровней ряда под воздействием тенденции. В сущности, эта модель – аналог аддитивной модели временного ряда, поскольку фактический уровень временного ряда есть сумма трендовой, сезонной и случайной компонент.

Основной недостаток модели с фиктивными переменными для описания сезонных и циклических колебаний – наличие большого количества переменных. При небольшом объеме выборки число степеней свободы невелико, что снижает вероятность получения статистически значимых оценок параметров уравнения регрессии.

3. Описание оборудования и используемых программных комплексов

При выполнении лабораторной работы необходим специализированный компьютерный класс с минимальными системными требованиями компьютеров:

- Процессор – Intel Pentium IV;
- ОЗУ – 500 Mb;
- видеокарта – 64 Mb.
- Требуемое программное обеспечение:
- Операционная система Microsoft Windows;
- MicrosoftExcel версии 2007 и выше.

4. Краткое руководство по эксплуатации оборудования

При использовании оборудования необходимо:

- соблюдать общие правила нахождения в учебных лабораториях, работы с компьютером и использования программных средств;
- осмотреть рабочее место, убрать все мешающие работе предметы;
- визуально проверить правильность подключения ПЭВМ к электросети.

5. Задание

Для исходных данных согласно выданному варианту (табл. 1, 2) построить модель временного ряда:

1. Построить и проанализировать автокорреляционную функцию и коррелограмму.
2. Нанести исходные данные на график и сделать предположение о наличии и виде сезонных колебаний.
3. Если можно предположить наличие сезонных колебаний, построить аддитивную или мультипликативную модель временного ряда (для построения модели использовать данные **без последнего уровня ряда**).
4. Осуществить прогноз уровня временного ряда, который не был включен в построение модели, и сравнить его с фактическим значением.

5. Построить модель регрессии с фиктивными переменными для временного ряда с постоянной амплитудой.

6. Сравнить данную модель с полученной в п.3.

Таблица 1. Исходные данные по временному ряду 1.

t		Варианты исходных данных								
	1	2	3	4	5	6	7	8	9	10
1	6,992	1258,0	54,29	54,21	22,38	4,91	6,17	32,85	19,94	5,688
2	3,763	1246,1	53,67	81,25	17,40	2,12	4,10	27,21	15,45	5,516
3	2,217	1190,5	50,46	82,06	16,57	3,64	7,07	29,37	13,81	3,964
4	5,604	1181,5	52,37	73,69	22,76	9,81	10,56	37,90	18,38	2,362
5	9,496	1225,0	56,50	96,26	20,18	5,34	6,36	30,01	22,25	6,283
6	8,818	1253,2	56,82	103,87	17,57	5,29	10,01	31,70	18,22	8,446
7	1,308	1204,3	52,81	92,07	26,79	11,92	12,49	38,40	15,99	8,920
8	4,496	1180,4	55,67	109,69	23,37	7,76	7,73	35,89	20,58	6,023
9	10,157	1183,4	58,63	124,30	19,78	6,61	12,36	33,79	23,73	4,000
10	7,932	1211,7	57,87	109,77	28,37	11,73	12,43	39,48	21,26	9,493
11	4,545	1240,2	57,94	124,13	27,90	11,43	10,40	35,22	17,06	12,921
12	3,587	1207,7	59,32	145,22	21,57	9,87	16,28	34,41	21,38	11,563
13	7,775	1168,2	61,58	132,19	28,90	14,87	15,94	42,09	24,03	10,836
14	9,799	1199,6	63,45	136,61	31,06	12,81	10,95	39,92	21,13	10,683
15	6,384	1217,2	59,83	161,60	24,75	10,23	16,88	36,66	18,65	12,199
16	3,907	1221,9	61,35	153,26	27,32	15,66	17,28	43,95	23,50	16,408
17	9,465	1186,6	65,06	152,23	36,34	16,48	14,07	45,22	25,79	15,685
18	11,908	1164,2	65,65	178,81	28,02	13,10	19,58	37,86	25,06	13,847
19	9,331	1202,4	64,86	175,47	27,93	17,11	19,34	45,50	20,98	12,787
20	6,648	1206,7	62,61	169,05	38,36	20,05	17,47	46,39	22,95	14,448
21	10,036	1199,9	66,77	195,16	31,99	16,61	22,55	42,25	28,29	21,080
22	13,226	1168,6	69,04	196,34	29,73	17,32	20,11	45,26	24,68	20,981
23	12,621	1137,6	64,79	186,07	39,07	22,25	19,91	51,18	21,53	18,421

Таблица 2. Исходные данные по временному ряду 2.

t	Варианты исходных данных									
	1	2	3	4	5	6	7	8	9	10
1	68,292	8,18	48,55	9,84	22,89	11,91	9,84	12,82	12,64	50,10
2	89,076	6,06	58,24	12,81	27,60	17,70	12,86	29,08	12,28	12,63
3	102,017	5,28	46,77	14,64	30,91	20,97	14,65	16,56	9,68	48,53
4	88,504	9,64	50,92	12,67	27,73	17,90	12,59	6,56	10,16	69,09
5	64,361	13,99	82,07	9,19	23,99	16,60	9,19	17,27	11,63	43,33
6	77,037	10,72	52,71	10,98	23,65	23,44	11,08	27,23	15,89	38,24

t	Варианты исходных данных									
	1	2	3	4	5	6	7	8	9	10
7	96,434	7,24	52,03	13,70	32,51	25,79	13,72	18,64	14,43	49,57
8	77,876	14,23	53,21	11,04	26,10	22,20	10,90	14,87	12,59	58,25
9	64,536	20,17	55,55	9,12	20,30	19,38	9,09	21,62	12,40	51,99
10	64,723	10,01	70,18	9,12	21,18	29,04	9,27	28,80	15,12	44,89
11	79,589	11,26	53,65	11,17	29,52	30,82	11,20	21,40	18,38	52,15
12	74,090	25,52	60,24	10,36	24,35	25,35	10,18	17,56	18,95	56,69
13	56,683	31,43	57,50	7,90	21,21	26,51	7,86	25,77	15,23	51,86
14	57,517	13,71	65,03	7,98	21,53	35,73	8,17	31,96	14,17	44,46
15	70,706	16,94	67,15	9,77	22,11	40,51	9,82	28,44	15,44	49,84
16	61,300	27,65	60,35	8,43	23,73	30,47	8,24	24,29	19,95	57,88
17	45,685	28,31	66,08	6,25	18,40	29,79	6,20	28,38	21,54	61,68
18	53,541	18,05	62,41	7,29	17,44	41,32	7,51	34,26	17,31	56,78
19	64,606	21,23	70,69	8,75	20,91	41,09	8,82	35,19	17,17	48,97
20	57,617	35,28	68,62	7,76	22,17	32,06	7,54	31,18	18,79	57,23
21	41,199	30,00	61,83	5,51	19,25	36,24	5,43	34,50	25,75	60,55
22	42,506	21,53	67,87	5,64	15,12	47,50	5,85	40,62	23,62	51,28
23	57,354	23,72	67,87	7,56	19,55	48,39	7,65	35,51	19,16	54,62
24	50,512	44,88	72,90	6,60	19,78	34,68	6,39	33,38	16,57	63,08
25	34,677	34,49	73,44	4,49	16,82	37,84	4,40	37,61	22,08	58,10
26	35,078	21,23	70,84	4,49	16,12	51,22	4,68	40,58	26,09	55,34
27	44,831	35,98	78,69	5,68	17,84	54,43	5,77	43,49	24,65	61,02
28	44,111	59,69	71,56	5,51	20,22	38,20	5,32	40,50	22,97	63,45
29	26,916	33,16	81,97	3,32	15,61	50,40	3,23	40,83	20,22	59,59
30	27,989	24,44	79,66	3,39	13,70	60,52	3,55	45,97	23,25	59,18

6. Методика выполнения заданий

6.1. Построение аддитивной модели.

Имеются ежемесячные исходные данные о заработной плате в Республике Таджикистан за 2004-2005 гг. (табл. 3).

Таблица 3. Исходные данные для построения аддитивной модели

t	y_t	t	y_t
1	54,05	11	71,77
2	53,74	12	81,36
3	60,37	13	76,92
4	55,00	14	77,61
5	55,96	15	87,77

6	61,20	16	83,61
7	57,25	17	84,56
8	58,11	18	91,10
9	73,73	19	87,18
10	75,18	20	87,03

Построим модель временного ряда по 19 значениям.

1. Построим автокорреляционную функцию для выбранного временного ряда. Для этого можно воспользоваться функцией КОРРЕЛ, задавая в качестве параметров «Массив» уровни одного и того же временного ряда со сдвигом. Так для нахождения r_1 нужно задать в качестве массива 1 – y_t при t = 1, 2, …, 18; в качестве массива 2 – y_t при t = 2, 3,…,19.

Чтобы рассчитать коэффициент автокорреляции таким способом, например, для лагов 1, 2,…, 6 нужно 6 раз использовать функцию КОРРРЕЛ, задавая различные диапазоны. Либо можно воспользоваться функцией СМЕЩ(ссылка,смещ_по_строкам,смещ_по_столбцам,[высота],[ширина]) при указании диапазона значений, после чего копировать формулу в смежные ячейки.

Параметры функции ***СМЕЩ***:

Ссылка – ссылка (на ячейку или на диапазон смежных ячеек), от которой вычисляется смещение.

Смещ_по_строкам – количество строк, которые требуется отсчитать вверх (при отрицательном значении) или вниз (при положительном значении), чтобы левая верхняя ячейка результата ссылалась на нужную ячейку.

Смещ_по_столбцам – количество столбцов, которые требуется отсчитать влево (при отрицательном значении) или вправо (при положительном значении), чтобы левая верхняя ячейка результата ссылалась на нужную ячейку.

Высота – число строк возвращаемой ссылки.

Ширина – число столбцов возвращаемой ссылки.

Для нашего примера укажем в качестве первого массива КОРРЕЛ функцию СМЕЩ с аргументами:

Ссылка: весь диапазон значений в виде абсолютной ссылки;

Смещ_по_строкам: величина лага;

Смещ_по_столбцам: 0;

Высота: разность между числом наблюдений и величиной лага;

Ширина: 1.

Для второго массива КОРРЕЛ функция СМЕЩ будет иметь такие же аргументы кроме Смещ_по_строкам, здесь укажем значение 0.

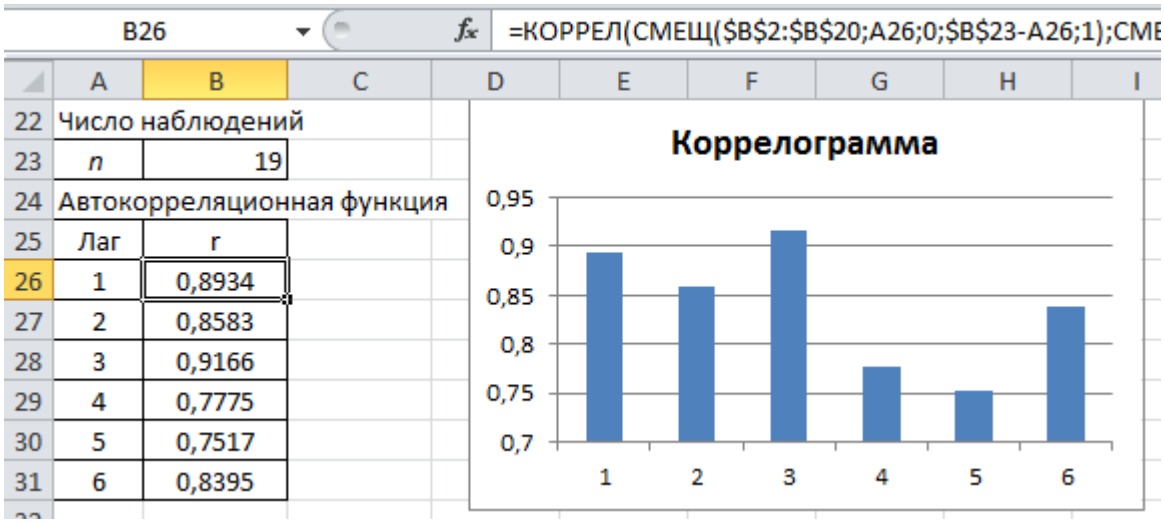


Рис. 1 – Автокорреляционная функция и коррелограмма временного ряда 1

2. Построим коррелограмму, для этого найденные значения автокорреляционной функции представим в виде столбиковой диаграммы (рис. 1).

Из автокорреляционной функции и коррелограммы на рис. 1 видно, что наибольшее значение достигается при лаге равном 3. Значит, временной ряд содержит циклические колебания с периодом 3.

3. На основе данных, приведенных в таблице 1, построим график зависимости уровня ряда от времени (рис. 2).

На графике видно, что данный временной ряд содержит циклические колебания с приблизительно равной амплитудой. Поэтому целесообразно строить аддитивную модель.

Рис. 2 – Исходный временной ряд 1

4. Проведем выравнивание уровней ряда методом скользящей средней. Найдем средние значения уровней ряда последовательно за каждые три периода

со сдвигом на один момент времени (столбец C рис. 3). Так как число моментов в одном периоде равно трем (нечетное число), то значения соответствуют фактическим моментам времени, т.е. нет необходимости искать центрированную скользящую среднюю. Полученные таким образом выровненные значения уже не содержат сезонной компоненты и соответствует фактическим моментам времени.

	A	B	C	D	E	F	G	H	I	J	K
1	t	y_t	Скользящая средняя за 3 периода	Оценка сезонной компоненты	S_t	T+E=Y−S	T	T+S	E	E^2	$(y_t-y_{ср})^2$
2	1	54,05	---	---	-1,1070	55,157	51,269	50,1616	3,8884	15,1196	282,8062
3	2	53,74	56,05	-2,31	-3,0079	56,748	53,453	50,4448	3,2952	10,8585	293,3287
4	3	60,37	56,37	4,00	4,1149	56,255	55,637	59,7516	0,6184	0,3824	110,1837
5	4	55,00	57,11	-2,11	-1,1070	56,107	57,821	56,7138	-1,7138	2,9370	251,7567
6	5	55,96	57,39	-1,43	-3,0079	58,968	60,005	56,9969	-1,0369	1,0752	222,2139
7	6	61,20	58,14	3,06	4,1149	57,085	62,189	66,3038	-5,1038	26,0485	93,44784
8	7	57,25	58,85	-1,60	-1,1070	58,357	64,373	63,2659	-6,0159	36,1915	185,4184
9	8	58,11	63,03	-4,92	-3,0079	61,118	66,557	63,5491	-5,4391	29,5839	162,737
10	9	73,73	69,01	4,72	4,1149	69,615	68,741	72,8559	0,8741	0,7640	8,197673
11	10	75,18	73,56	1,62	-1,1070	76,287	70,925	69,8181	5,3619	28,7499	18,60333
12	11	71,77	76,10	-4,33	-3,0079	74,778	73,109	70,1013	1,6687	2,7847	0,815694
13	12	81,36	76,68	4,68	4,1149	77,245	75,293	79,4081	1,9519	3,8099	110,1064
14	13	76,92	78,63	-1,71	-1,1070	78,027	77,477	76,3703	0,5497	0,3022	36,64072
15	14	77,61	80,77	-3,16	-3,0079	80,618	79,661	76,6534	0,9566	0,9150	45,47018
16	15	87,77	83,00	4,77	4,1149	83,655	81,845	85,9603	1,8097	3,2751	285,7167
17	16	83,61	85,31	-1,70	-1,1070	84,717	84,029	82,9224	0,6876	0,4727	162,3881
18	17	84,56	86,42	-1,86	-3,0079	87,568	86,213	83,2056	1,3544	1,8344	187,5026
19	18	91,10	87,61	3,49	4,1149	86,985	88,398	92,5124	-1,4124	1,9950	409,3807
20	19	87,18	---	---	-1,1070	88,287	90,582	89,4746	-2,2946	5,2652	266,1191

Рис. 3 – Расчет параметров аддтитивной модели

5. Найдем оценки сезонной компоненты как разность между фактическими уровнями ряда и скользящими средними (столбец D рис. 3).

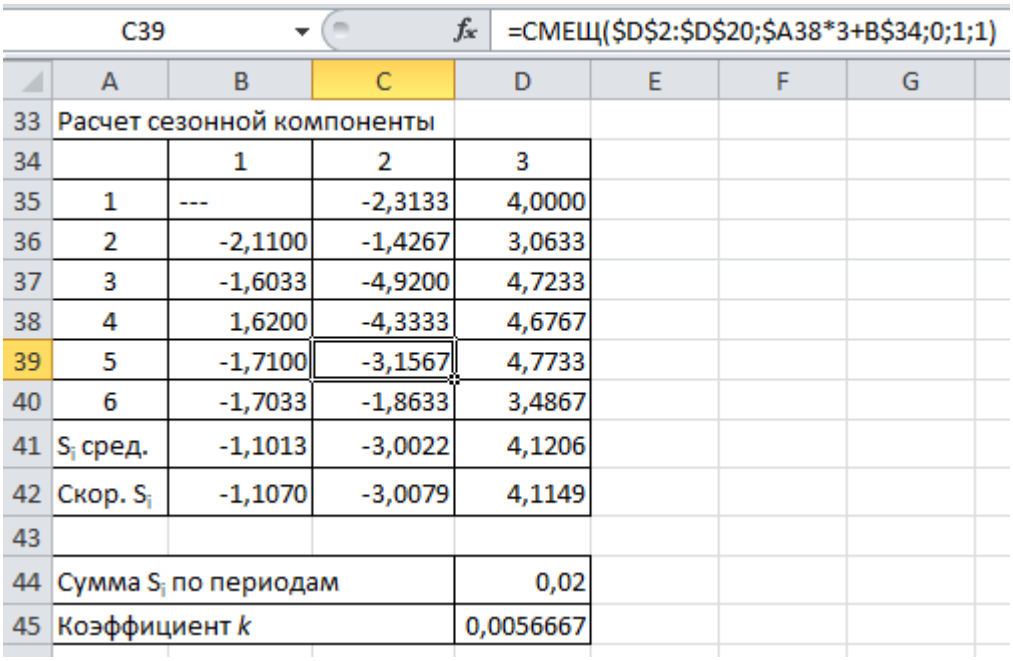
6. Используем найденные оценки для расчета значений сезонной компоненты S_t.

1) Чтобы перенести значения из основной таблицы, снова воспользуемся функцией СМЕЩ (рис. 4).

2) Найдем средние оценки сезонной компоненты S_i за каждый период. В моделях с сезонной компонентой обычно предполагается, что сезонные воздействия за период взаимопоглощаются. Для аддитивной модели это значит, что сумма значений сезонной компоненты по всем кварталам должна быть равна нулю.

Для данной модели имеем: -1,1-3+4,12=0,02.

3) Определим корректирующий коэффициент: k=0,02/3=0,006. Рассчитаем скорректированные значения сезонной компоненты как разность между ее средней оценкой и корректирующим коэффициентом k: $S_i = \overline{S}_i - k$, где i=1,2,3



	A	B	C	D	E	F	G
33	Расчет сезонной компоненты						
34		1	2	3			
35	1	---	-2,3133	4,0000			
36	2	-2,1100	-1,4267	3,0633			
37	3	-1,6033	-4,9200	4,7233			
38	4	1,6200	-4,3333	4,6767			
39	5	-1,7100	-3,1567	4,7733			
40	6	-1,7033	-1,8633	3,4867			
41	S_i сред.	-1,1013	-3,0022	4,1206			
42	Скор. S_i	-1,1070	-3,0079	4,1149			
43							
44	Сумма S_i по периодам			0,02			
45	Коэффициент *k*			0,0056667			

Рис. 4 – Расчет сезонной компоненты для аддитивной модели

Проверим условие равенства нулю суммы значений сезонной компоненты:

-1,107-3,008+4,115=0.

4) Таким образом, получены следующие значения сезонной компоненты:

1 период: S_1=-1,107

2 период: S_2=-3,008

3 период: S_3=-4,115

Используя абсоолютные ссылки, перенесем полученные значения для соответствующих периодов в основную таблицу (столбец Е рис. 3).

7. Элиминируем влияние сезонной компоненты, вычитая ее значения из каждого уровня исходного ВР. Получим: $T+E=Y-S$ (столбец F рис. 3). Эти значения рассчитываются для каждого момента времени и содержат только тенденцию и случайную компоненту.

8. Определим компоненту T данной модели. Для этого проведем аналитическое выравнивание ряда $(T+E)$ с помощью линейного тренда. Воспользуемся встроенной функцией ЛИНЕЙН (рис. 5). Получаем уравнение следующего вида:

Y=2,184t+49,085

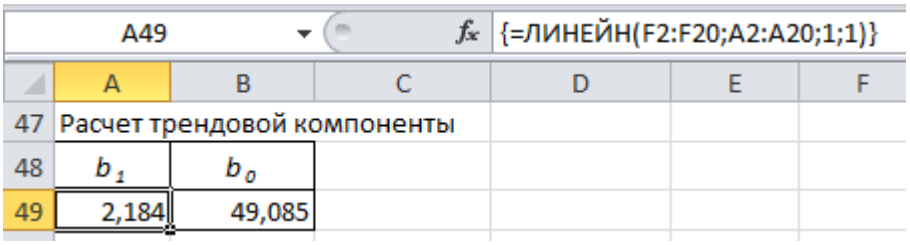


Рис. 5 – Расчет трендовой составляющей временного ряда

Подставим в это уравнение значения t=1,…,19, найдем уровни T для каждого момента времени (столбец G рис. 3).

9. Рассчитаем оцененные значения с учетом циклической компоненты как T+S (столбец H рис. 3) и нанесем их на график (рис.6).

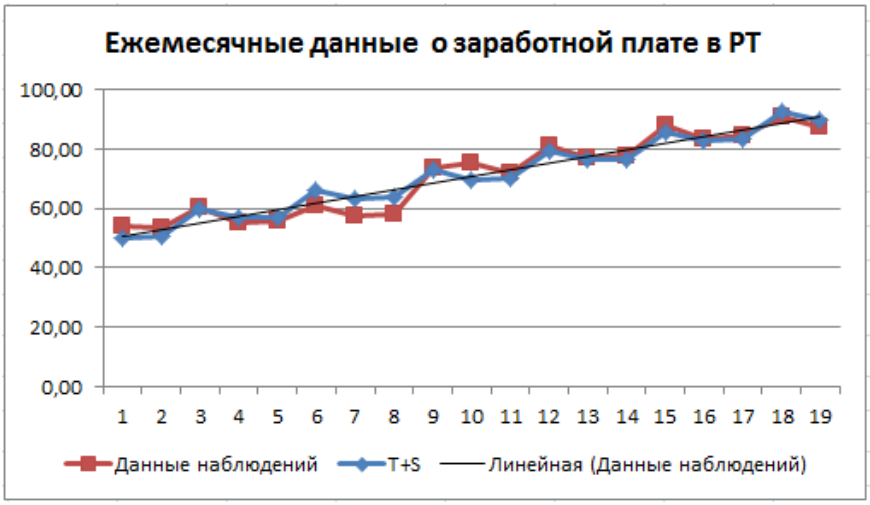


Рис. 6 – Трендовая и сезонная составляющие временного ряда 1

10. Проведем расчет ошибок по формуле E=Y-(T+S). Это абсолютная ошибка.

Численные значения абсолютных ошибок приведены в столбце I рис. 3. Для получения относительной ошибки надо возвести абсолютную ошибку в квадрат (столбец J рис. 3), найти сумму квадратов отклонения от его среднего значения (столбец K рис. 3) и найти их отношение.

Получаем:

$$C_{\text{ост}} = \sum E^2 = 179{,}708$$

$$\sum (y_t - \overline{y})^2 = 3132{,}834$$

Относительная ошибка: (172,365/3132,834)·100%=5,5%

Тогда полученная модель объясняет 94,5% общей вариации уровней временного ряда (рис. 7).

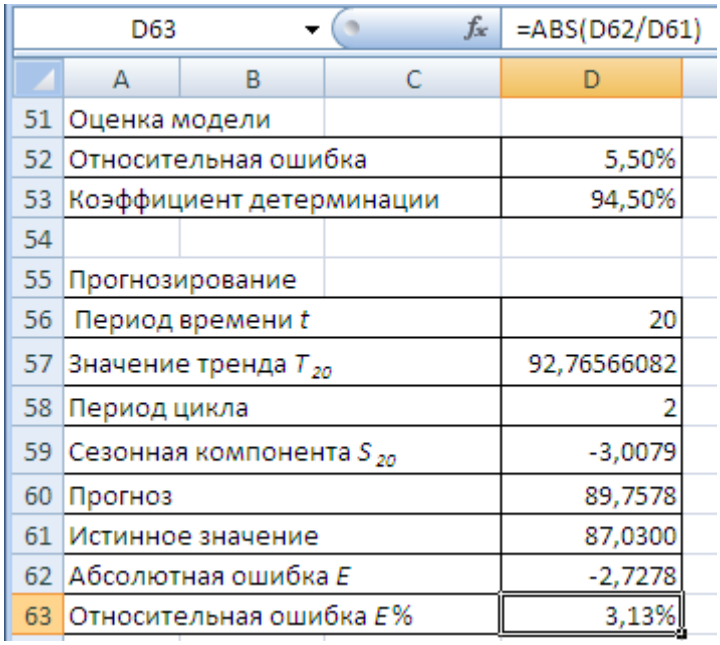


Рис. 7 – Оценка модели и прогнозирование уровня временного ряда 1

11. Найдем прогнозное значение уровня временного ряда для момента времени 20 и сравним его с истинным значением y_{20} = 87,03 (рис. 7):

1) Рассчитаем прогнозное значение трендовой компоненты для момента времени t=20, получим T=2,184·20+49,085=92,766.

2) Рассчитаем прогнозное значение циклической компоненты. Номер периода определяется как остаток от деления t на количество периодов в цикле. Для этого можно воспользоваться функцией ***ОСТАТ*** (она возвращает остаток от деления аргумента «число» на значение аргумента «делитель»).

Для нашего примера S_{20}=S_2=-3,008.

3) Найдем значения уровней ряда, полученные по аддитивной модели. Для этого прибавим к уровню T значение сезонной компоненты:

T+S=92,766-3,008=89,758.

4) Сравним прогнозное значение для t=20 с фактическим: y_t=87,03. Тогда E=-2,728 или 3,13%, т.е. модель дает достаточно хороший прогноз.

6.2. Построение мультипликативной модели.

Имеются данные об авиаперевозках пассажиров поквартально с 1949 по 1956 г. (в тысячах пассажиров).

Таблица 4. Исходные данные для построения мультипликативной модели

t	y_t	t	y_t
1	362	15	681
2	385	16	557
3	432	17	628

4	341	18	707
5	382	19	773
6	409	20	592
7	498	21	557
8	387	22	725
9	473	23	854
10	513	24	661
11	582	25	742
12	474	26	854
13	544	27	1023
14	582	28	789

Построим модель временного ряда по 27 значениям.

1. Аналогично предыдущему примеру построим автокорреляционную функцию и коррелограмму для выбранного временного ряда (рис. 8).

30	Число наблюдений	
31	*n*	27
32	Автокорреляционная функция	
33	Лаг	r
34	1	0,8098
35	2	0,6790
36	3	0,7280
37	4	0,9487
38	5	0,7128
39	6	0,5364
40	7	0,6532

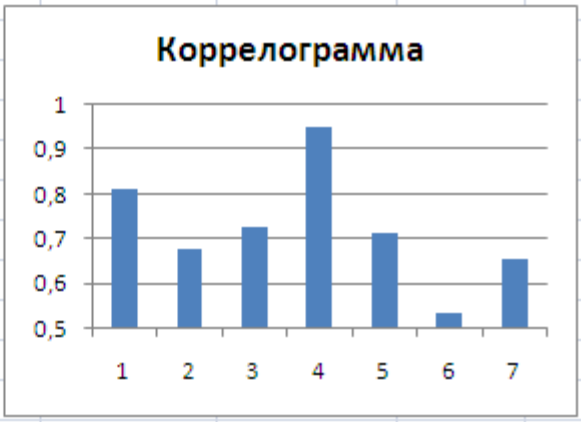


Рис. 8 – Автокорреляционная функция и коррелограмма временного ряда 2

На рис. 8 видно, что наибольший коэффициент автокорреляции достигается при лаге 4. Значит, временной ряд содержит циклические колебания с периодом 4.

2. На основе данных, приведенных в таблице 4, построим график зависимости уровня ряда от времени (рис. 9).

Рис. 9 – Исходный временной ряд 2

На графике рис. 9 видно, что данный временной ряд содержит циклические колебания с возрастающей амплитудой. Поэтому целесообразно строить мультипликативную модель.

3. Проведем выравнивание уровней ряда методом скользящей средней. Найдем средние значения уровней ряда последовательно за каждые четыре периода со сдвигом на один момент времени (столбец C рис. 10). Так как число моментов в одном периоде равно четырем (четное число), то значения не соответствуют фактическим моментам времени, следовательно, необходимо рассчитать центрированную скользящую среднюю. В столбце D рис. 10 центрированная скользящая средняя рассчитывается как среднее арифметическое двух соседних значений скользящей средней из столбца C. Полученные таким образом выровненные значения не содержат сезонной компоненты и соответствует фактическим моментам времени.

4. Найдем оценки сезонной компоненты как частное от деления фактических уровней ряда на центрированную скользящую среднюю (столбец E рис. 10).

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	t	y_t	Скользящая средняя за 4 периода	Центрированная скользящая средняя	Оценка сезонной компоненты	S_t	TxE=Y/S	T	TxS	E	E'	$(E')^2$	$(y_t-y_{cp})^2$
2	1	362	---	---	---	0,9423	384,174	336,086	316,6875	1,1431	45,31	2053,2	48465,2
3	2	385	380,00	---	---	1,0274	374,718	354,707	364,4401	1,0564	20,56	422,7	38867,4
4	3	432	385,00	382,50	1,13	1,1502	375,588	373,329	429,4019	1,0061	2,60	6,7	22544,5
5	4	341	391,00	388,00	0,88	0,8801	387,463	391,950	344,9486	0,9886	-3,95	15,6	58152,4
6	5	382	407,50	399,25	0,96	0,9423	405,399	410,571	386,8734	0,9874	-4,87	23,8	40059,3
7	6	409	419,00	413,25	0,99	1,0274	398,077	429,193	440,9690	0,9275	-31,97	1022,0	29980,3
8	7	498	441,75	430,38	1,16	1,1502	432,969	447,814	515,0745	0,9669	-17,07	291,5	7080,9
9	8	387	467,75	454,75	0,85	0,8801	439,731	466,435	410,5016	0,9427	-23,50	552,3	38082,8
10	9	473	488,75	478,25	0,99	0,9423	501,974	485,056	457,0593	1,0349	15,94	254,1	11913,3
11	10	513	510,50	499,63	1,03	1,0274	499,300	503,678	517,4979	0,9913	-4,50	20,2	4781,5
12	11	582	528,25	519,38	1,12	1,1502	506,000	522,299	600,7472	0,9688	-18,75	351,5	0,0
13	12	474	545,50	536,88	0,88	0,8801	538,586	540,920	476,0547	0,9957	-2,05	4,2	11696,0
14	13	544	570,25	557,88	0,98	0,9423	577,323	559,542	527,2452	1,0318	16,75	280,7	1455,3
15	14	582	591,00	580,63	1,00	1,0274	566,457	578,163	594,0268	0,9798	-12,03	144,6	0,0
16	15	681	612,00	601,50	1,13	1,1502	592,072	596,784	686,4198	0,9921	-5,42	29,4	9771,7
17	16	557	643,25	627,63	0,89	0,8801	632,895	615,405	541,6078	1,0284	15,39	236,9	632,4
18	17	628	666,25	654,75	0,96	0,9423	666,468	634,027	597,4311	1,0512	30,57	934,5	2102,4
19	18	707	675,00	670,63	1,05	1,0274	688,119	652,648	670,5556	1,0543	36,44	1328,2	15588,0
20	19	773	657,25	666,13	1,16	1,1502	672,058	671,269	772,0924	1,0012	0,91	0,8	36424,4
21	20	592	661,75	659,50	0,90	0,8801	672,664	689,890	607,1609	0,9750	-15,16	229,9	97,1
22	21	557	682,00	671,88	0,83	0,9423	591,119	708,512	667,6170	0,8343	-110,62	12236,1	632,4
23	22	725	699,25	690,63	1,05	1,0274	705,638	727,133	747,0845	0,9704	-22,08	487,7	20406,7
24	23	854	745,50	722,38	1,18	1,1502	742,481	745,754	857,7650	0,9956	-3,77	14,2	73903,4
25	24	661	777,75	761,63	0,87	0,8801	751,066	764,376	672,7139	0,9826	-11,71	137,2	6217,6
26	25	742	820,00	798,88	0,93	0,9423	787,451	782,997	737,8028	1,0057	4,20	17,6	25552,6
27	26	854	---	---	---	1,0274	831,193	801,618	823,6134	1,0369	30,39	923,3	73903,4
28	27	1023	---	---	---	1,1502	889,412	820,239	943,4376	1,0843	79,56	6330,2	194350,4

Рис. 10 – Расчет параметров мультипликативной модели

5. Используем найденные оценки для расчета значений сезонной компоненты S_t (рис. 11).

1) Найдем средние оценки сезонной компоненты S_i за каждый период. Взаимопогашаемось сезонных воздействий в мультипликативной модели выражается в том, что сумма значений сезонной компоненты по всем периодам должна быть равна числу периодов в цикле, т.е. 4. Для данной модели имеем: 0,94+1,025+1,147+0,878=3,99.

2) Определим корректирующий коэффициент: k=4/3,99=1,003.

Определим скорректированные значения сезонной компоненты, умножив ее средние оценки на корректирующий коэффициент k:

$S_i = \overline{S}_i \cdot k$, где i=1,2,3,4.

Проверим условие равенства 4 суммы значений сезонной компоненты: 0,942+1,027+1,15+0,88=4.

3) Таким образом, получены следующие значения сезонной компоненты:

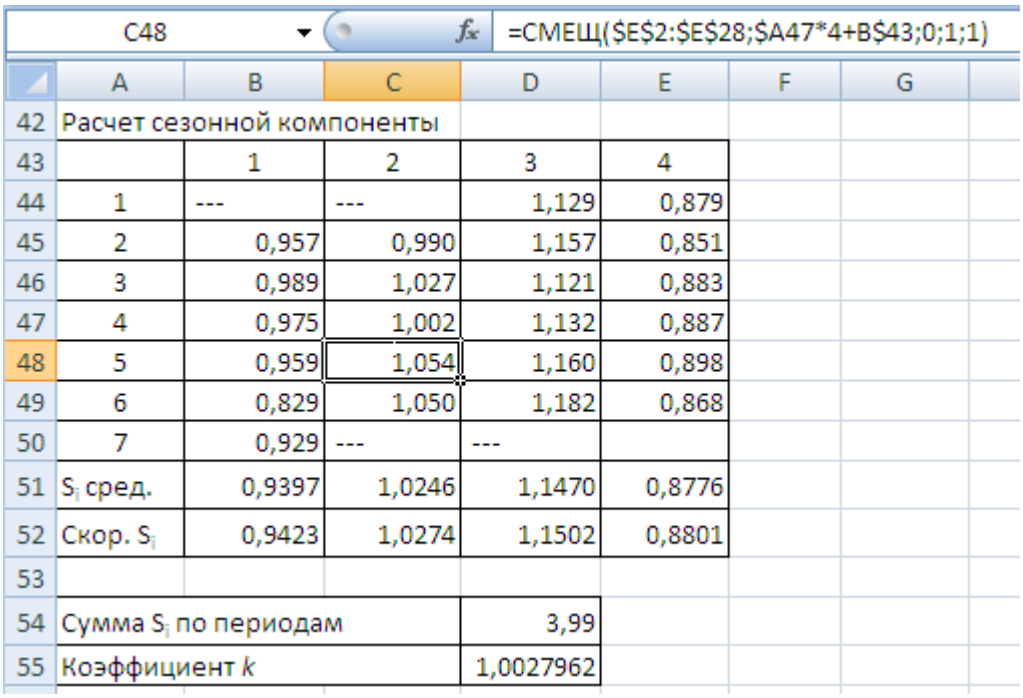
1 период: S_1=0,942

2 период: S_2=1,027

3 период: S_3=1,15

4 период: S_4=0,88

Занесем полученные значения в основную таблицу (столбец F рис. 10).



C48 | =СМЕЩ($E$2:$E$28;$A47*4+B$43;0;1;1)

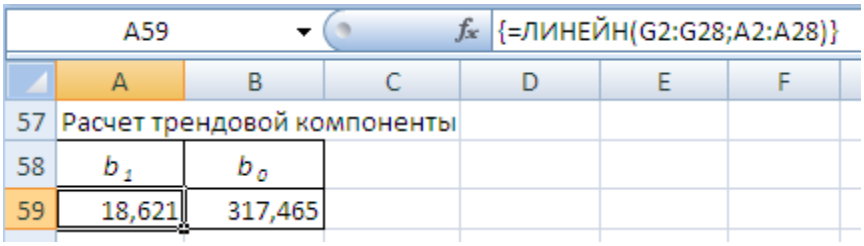
	A	B	C	D	E	F	G
42	Расчет сезонной компоненты						
43		1	2	3	4		
44	1	---	---	1,129	0,879		
45	2	0,957	0,990	1,157	0,851		
46	3	0,989	1,027	1,121	0,883		
47	4	0,975	1,002	1,132	0,887		
48	5	0,959	1,054	1,160	0,898		
49	6	0,829	1,050	1,182	0,868		
50	7	0,929	---	---			
51	S_i сред.	0,9397	1,0246	1,1470	0,8776		
52	Скор. S_i	0,9423	1,0274	1,1502	0,8801		
53							
54	Сумма S_i по периодам			3,99			
55	Коэффициент k			1,0027962			

Рис. 11 – Расчет сезонной компоненты для мультипликативной модели

6. Элиминируем влияние сезонной компоненты, поделив на нее значения уровней исходного временного ряда. Получим: $T \cdot E = Y/S$ (столбец G рис. 10). Эти значения рассчитываются для каждого момента времени и содержат только тенденцию и случайную компоненту.

7. Определим компоненту T данной модели. Для этого проведем аналитическое выравнивание ряда $T \cdot E$ с помощью линейного тренда. Воспользуемся встроенной функцией ЛИНЕЙН (рис. 12). Получаем уравнение следующего вида:

Y=18,621t+317,465



A59 | {=ЛИНЕЙН(G2:G28;A2:A28)}

	A	B	C	D	E	F
57	Расчет трендовой компоненты					
58	b_1	b_0				
59	18,621	317,465				

Рис. 12 – Расчет трендовой составляющей временного ряда

Подставим в это уравнение значения t=1,…,27, найдем уровни T для каждого момента времени (столбец H рис. 10).

8. Рассчитаем оцененные значения с учетом циклической компоненты как $T \cdot S$ (столбец I рис. 10) и нанесем их на график (рис. 13).

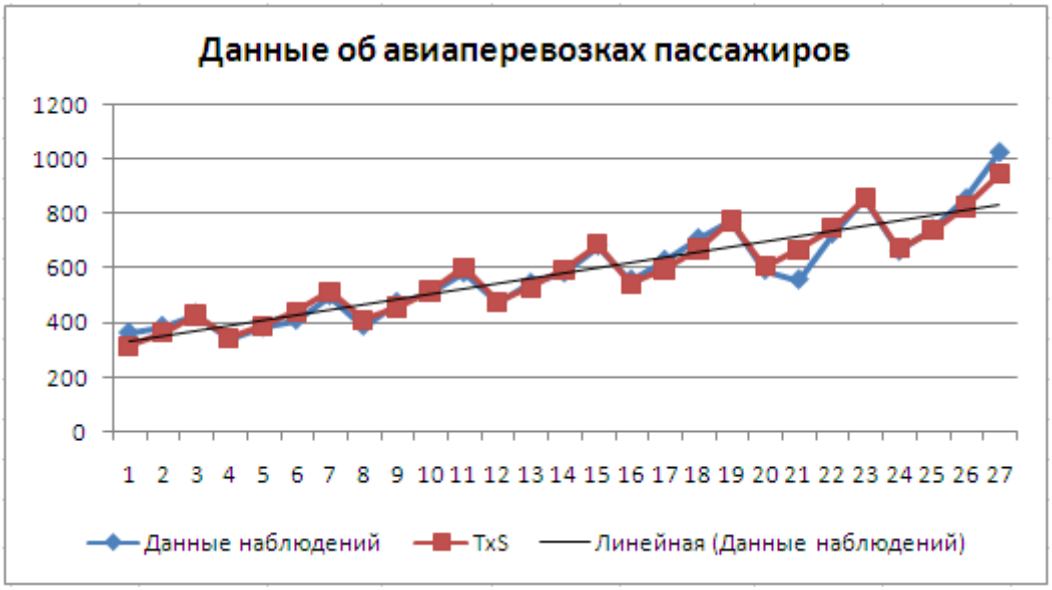


Рис. 13– Трендовая и сезонная составляющие временного ряда 2

9. Проведем расчет ошибок по формуле $E=Y/(T \cdot S)$. Численные значения абсолютных ошибок приведены в столбце J рис. 10. Абсолютные ошибки в мультипликативной модели определяются как $E'=Y-(T \cdot S)$ (столбец K рис. 10). Для получения относительной ошибки надо возвести абсолютную ошибку в квадрат (столбец L рис. 10), найти сумму квадратов отклонения от его среднего значения (столбец M рис. 10) и найти их отношение.

Получаем: $C_{\text{ост}} = \sum(E')^2 = 28349{,}268; \sum(y_t - \overline{y})^2 = 772661{,}407$

Относительная ошибка: $(28349{,}268/772661{,}407) \cdot 100\% = 5{,}5\%$

Тогда полученная модель объясняет 96,331% общей вариации уровней временного ряда (рис. 14).

10. Найдем прогнозное значение уровня временного ряда для момента времени 28 и сравним его с истинным значением $y_{28} = 789$ (рис. 14):

1) Трендовая компонента: $18{,}621 \cdot 28 + 317{,}465 = 838{,}8607$.

2) Циклическая компонента: остаток от деления $t=28$ на количество периодов в цикле (4) равен 0, т.е. мы имеем дело с последним (четвертым) периодом цикла. Тогда $S_{28}=S_4=0{,}8801$.

3) Прогнозное значение: $T \cdot S = 838{,}8607 \cdot 0{,}8801 = 738{,}267$.

4) Сравним прогнозное значение для $t=28$ с фактическим: $y_t=789$. Тогда $E=50{,}733$ или 6,43%, т.е. модель дает достаточно хороший прогноз.

D62 | =СУММ(L2:L28)/СУММ(M2:M28)

	A	B	C	D	E	F	G
61	Оценка модели						
62	Относительная ошибка			3,67%			
63	Коэффициент детерминации			96,33%			
64							
65	Прогнозирование						
66	Период времени t			28			
67	Значение тренда T_{28}			838,86072			
68	Период цикла			0			
69	Сезонная компонента S_{28}			0,8801			
70	Прогноз			738,2670			
71	Истинное значение			789,0000			
72	Абсолютная ошибка E			50,7330			
73	Относительная ошибка E%			6,43%			

Рис. 14 – Оценка модели и прогнозирование уровня временного ряда 2

6.3. Построение модели регрессии с фиктивными переменными.

Для построения модели регрессии с фиктивными переменными используем временной ряд из п. 6.1 (табл.1).

Количество фиктивных переменных должно быть на единицу меньше числа моментов времени одного цикла. Тогда модель для данной задачи должна включать три независимых переменных: две фиктивные переменные и фактор времени. Каждая фиктивная переменная отражает сезонную компоненту временного ряда для какого-либо одного периода. В данной задаче общий вид модели следующий:

$$y_t = a + b \cdot t + c_1 x_1 + c_2 x_2 + \varepsilon_t$$

где $x_1 = \begin{cases} 1 \text{ для первого периода,} \\ 0 \text{ для остальных периодов.} \end{cases}$ $x_2 = \begin{cases} 1 \text{ для второго периода,} \\ 0 \text{ для остальных периодов.} \end{cases}$

Уравнение тренда для каждого периода будет иметь следующий вид:

- для 1-ого периода: $y_t = a + b \cdot t + c_1 x_1 + \varepsilon_t$
- для 2-ого периода: $y_t = a + b \cdot t + c_2 x_2 + \varepsilon_t$
- для 3-его периода: $y_t = a + b \cdot t + \varepsilon_t$

Величину свободного члена для каждого периода времени составит:

- для 1 периода ($a + c_1$);
- для 2 периода ($a + c_2$);
- для 3 периода a.

Составим матрицу исходных данных (рис. 15).

Оценим параметры регрессии обычным МНК, используя инструмент Регрессия в надстройке Пакет Анализа. Результаты приведены на рис. 16.

	A	B	C	D
1	t	x_1	x_2	y_t
2	1	1	0	54,05
3	2	0	1	53,74
4	3	0	0	60,37
5	4	1	0	55,00
6	5	0	1	55,96
7	6	0	0	61,20
8	7	1	0	57,25
9	8	0	1	58,11
10	9	0	0	73,73
11	10	1	0	75,18
12	11	0	1	71,77
13	12	0	0	81,36
14	13	1	0	76,92
15	14	0	1	77,61
16	15	0	0	87,77
17	16	1	0	83,61
18	17	0	1	84,56
19	18	0	0	91,10
20	19	1	0	87,18

Рис. 15 – Матрица исходных данных

	A	B	C	D	E	F	G
1	ВЫВОД ИТОГОВ						
2							
3	*Регрессионная статистика*						
4	Множественный R	0,972168					
5	R-квадрат	0,94511					
6	Нормированный R-квадрат	0,934132					
7	Стандартная ошибка	3,385857					
8	Наблюдения	19					
9							
10	Дисперсионный анализ						
11		*df*	*SS*	*MS*	*F*	*Значимость F*	
12	Регрессия	3	2960,87323	986,9577432	86,091727	0,0000000011	
13	Остаток	15	171,960381	11,4640254			
14	Итого	18	3132,833611				
15							
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
17	Y-пересечение	52,97	2,03464664	26,03400461	0,000000	48,63325336	57,30675
18	t	2,185873	0,142192615	15,37261988	0,000000	1,882796633	2,488949
19	x1	-4,94444	1,885057956	-2,622966805	0,019200	-8,96235035	-0,92654
20	x2	-6,77746	1,959989933	-3,457905677	0,003514	-10,95507995	-2,59984

Рис. 16 – Результаты регрессионного анализа для временного ряда 1

Тогда модель временного ряда будет иметь вид:

$$y_t = 52{,}97 + 2{,}186 \cdot t - 4{,}944 \cdot x_1 - 6{,}777 \cdot x_2$$

Проанализируем результаты, представленные на рис. 16.

Параметр a = 52,074 – это сумма начального уровня ряда и сезонной компоненты в третьем периоде. Сезонные колебания в первом и во втором периоде приводят к снижению этой величины, о чем свидетельствуют отрицательные значения оценок параметров при переменных x_1, x_2. Положительная величина параметра b=2,271 при переменной времени свидетельствует о наличии возрастающей тенденции в уровнях ряда.

Влияние трендовой и сезонных компонент в каждом периоде статистически значимо (P-значение < 0,05).

Коэффициент детерминации в модели с фиктивными переменными R^2 = 0,947 превосходит полученный для аддитивной модели (0,945). Следовательно, модель регрессии с фиктивными переменными описывает динамику данного временного ряда лучше, чем аддитивная модель.

7. Контрольные вопросы

1. В чем отличие временных рядов от пространственных данных?
2. Какие факторы могут воздействовать на формирование уровней временного ряда?
3. В каких случаях строится аддитивная модель, в каких мультипликативная?
4. Дайте определение автокорреляции уровней временного ряда.
5. Каким образом можно выявить наличие циклических колебаний в уровнях временного ряда?
6. Перечислите этапы построения аддитивной и мультипликативной модели временного ряда.
7. Каким образом производится оценка сезонной компоненты в аддитивной и мультипликативной моделях? В чем отличие?
8. Объясните назначение функции MS Excel СМЕЩ, опишите ее параметры.
9. Перечислите основные этапы метода фиктивных переменных для моделирования сезонных колебаний.
10. Назовите достоинства и недостатки метода фиктивных переменных по сравнению с методом скользящей средней.

Приложение 1.

Значения статистик Дарбина-Уотсона при 5%-ном уровне значимости

n	m=1		m=2		m=3		m=4		m=5	
	d_L	d_U	d_L	d_U	d_L	d_L	d_U	d_L	d_U	d_L
6	0,610	1,400	-	-	-	-	-	-	-	-
7	0,700	1,356	0,467	1,896	-	-	-	-	-	-
8	0,763	1,332	0,559	1,777	0,368	2,287	-	-	-	-
9	0,824	1,320	0,629	1,699	0,455	2,128	-	-	-	-
10	0,879	1,320	0,697	1,641	0,525	2,016	-	-	-	-
11	0,927	1,324	0,658	1,604	0,595	1,928	-	-	-	-
12	0,971	1,331	0,812	1,579	0,658	1,864	-	-	-	-
13	1,010	1,340	0,861	1,562	0,715	1,816	-	-	-	-
14	1,045	1,350	0,905	1,551	0,767	1,779	-	-	-	-
15	1,077	1,361	0,946	1,543	0,814	1,750	0,685	1,977	0,562	2,220
16	1,106	1,371	0,982	1,539	0,857	1,728	0,734	1,935	0,615	2,157
17	1,133	1,381	1,015	1,536	0,897	1,710	0,779	1,900	0,664	2,104
18	1,158	1,391	1,046	1,535	0,933	1,696	0,820	1,872	0,710	2,060
19	1,180	1,401	1,074	1,536	0,967	1,685	0,859	1,849	0,752	2,023
20	1,201	1,411	1,100	1,537	0,998	1,676	0,984	1,828	0,792	1,991
21	1,222	1,420	1,125	1,538	1,026	1,669	0,927	1,812	0,829	1,964
22	1,239	1,429	1,147	1,541	1,053	1,664	0,958	1,797	0,863	1,940
23	1,257	1,437	1,168	1,543	1,078	1,660	0,986	1,785	0,895	1,920
24	1,273	1,446	1,188	1,546	1,101	1,656	1,013	1,775	0,925	1,902
25	1,288	1,454	1,206	1,550	1,123	1,654	1,038	1,767	0,953	1,886
26	1,302	1,461	1,224	1,553	1,143	1,652	1,062	1,759	0,979	1,873
27	1,316	1,469	1,240	1,556	1,162	1,651	1,084	1,753	1,004	1,861
28	1,328	1,476	1,255	1,560	1,181	1,650	1,104	1,747	1,028	1,850
29	1,341	1,483	1,270	1,563	1,198	1,650	1,124	1,743	1,050	1,841
30	1,352	1,489	1,284	1,567	1,214	1,650	1,143	1,739	1,071	1,833
31	1,363	1,496	1,297	1,570	1,229	1,650	1,160	1,735	1,090	1,825
32	1,373	1,502	1,309	1,574	1,244	1,650	1,177	1,732	1,109	1,819
33	1,383	1,508	1,321	1,577	1,258	1,651	1,193	1,730	1,127	1,813
34	1,393	1,514	1,333	1,580	1,271	1,652	1,028	1,728	1,144	1,808
35	1,402	1,519	1,343	1,584	1,283	1,653	1,222	1,726	1,160	1,803
36	1,411	1,525	1,354	1,587	1,295	1,654	1,236	1,724	1,175	1,799
37	1,419	1,530	1,364	1,590	1,307	1,655	1,249	1,723	1,190	1,795
38	1,427	1,535	1,373	1,594	1,318	1,656	1,261	1,722	1,204	1,792
39	1,435	1,540	1,382	1,597	1,328	1,658	1,273	1,722	1,218	1,789

Приложение 1 (продолжение)

n	*m*=1		*m*=2		*m*=3		*m*=4		*m*=5	
	d_L	d_U	d_L	d_U	d_L	d_L	d_U	d_L	d_U	d_L
40	1,442	1,544	1,391	1,600	1,338	1,659	1,285	1,721	1,230	1,786
45	1,475	1,566	1,430	1,615	1,383	1,666	1,336	1,720	1,287	1,776
50	1,503	1,585	1,462	1,628	1,421	1,674	1,378	1,721	1,335	1,771
55	1,528	1,601	1,490	1,641	1,452	1,681	1,414	1,724	1,374	1,768
60	1,549	1,616	1,514	1,652	1,480	1,689	1,444	1,727	1,408	1,767
65	1,567	1,629	1,536	1,662	1,503	1,696	1,471	1,731	1,438	1,767
70	1,583	1,641	1,554	1,672	1,525	1,703	1,494	1,735	1,464	1,768
75	1,598	1,652	1,571	1,680	1,543	1,709	1,515	1,739	1,487	1,770
80	1,611	1,662	1,586	1,688	1,560	1,715	1,534	1,743	1,507	1,772
85	1,624	1,671	1,600	1,696	1,575	1,721	1,550	1,747	1,525	1,774
90	1,635	1,679	1,612	1,703	1,589	1,726	1,566	1,751	1,542	1,776
95	1,645	1,687	1,623	1,709	1,602	1,732	1,579	1,755	1,557	1,778
100	1,654	1,694	1,634	1,715	1,613	1,736	1,592	1,758	1,571	1,780
150	1,720	1,746	1,706	1,760	1,693	1,774	1,679	1,788	1,665	1,802
200	1,758	1,778	1,748	1,789	1,738	1,799	1,728	1,810	1,718	1,820