Министерство сельского хозяйства Российской Федерации

ФГБОУ ВО Пензенский ГАУ

Г.А. Волкова

ЭКОНОМЕТРИКА (ПРОДВИНУТЫЙ УРОВЕНЬ)

КОМПЬЮТЕРНЫЙ ПРАКТИКУМ

Пенза 2020

Министерство сельского хозяйства Российской Федерации

ФГБОУ ВО Пензенский ГАУ

Г.А. Волкова

ЭКОНОМЕТРИКА (ПРОДВИНУТЫЙ УРОВЕНЬ):

Компьютерный практикум
для студентов, обучающихся по направлению подготовки
38.04.01 Экономика,
квалификация магистр

Пенза 2020

УДК 519.8 (075)
ББК 65.01 (Я7)
В 67

Рецензент – кандидат экономических наук, доцент кафедры «Бухгалтерский учет, анализ и аудит» ФГБОУ ВО ПГАУ Лаврина О.В.

Печатается по решению методической комиссии экономического факультета от 25. 05. 2020 г., протокол № 11.

Волкова, Галина Александровна
В 67 Эконометрика (продвинутый уровень): компьютерный практикум / Г.А. Волкова. – Пенза: РИО ПГАУ, 2020. – 62 с.

СОДЕРЖАНИЕ

Введение

Успешная работа современного экономиста в любой области экономики тесным образом связана с использованием математических методов и средств вычислительной техники. При решении задач из различных областей человеческой деятельности часто приходится использовать методы, основанные на эконометрических моделях.

Зарождение эконометрики является следствием междисциплинарного подхода к изучению экономики. Эта наука возникла в результате взаимодействия и объединения в особый «сплав» трех компонент: экономической теории, статистических и математических методов. Впоследствии к ним присоединилось развитие вычислительной техники как условие развития эконометрики.

В журнале «Эконометрика», основанном в 1933 г. Р. Фришем (1895–1973), он дал следующее определение эконометрики: «Эконометрика – это не то же самое, что экономическая статистика. Она не идентична и тому, что мы называем экономической теорией, хотя значительная часть этой теории носит количественный характер. Эконометрика не является синонимом приложений математики к экономике. Как показывает опыт, каждая из трех отправных точек – статистика, экономическая теория и математика – необходимое, но не достаточное условие для понимания количественных соотношений в современной экономической жизни. Это – единство всех трех составляющих. И это единство образует эконометрику».

Таким образом, эконометрика – это наука, которая дает количественное выражение взаимосвязей экономических явлений и процессов.

Практикум содержит краткие теоретические справки, справочный материал по функциям и надстройкам MS Excel, используемым при решении задач и подробный разбор решения задач в среде MS Excel, сопровождаемый скриншотами и задания для самостоятельного решения. Практикум предназначен для студентов, обучающихся по направлению подготовки 38.04.01 Экономика, квалификация магистр

1 ПАРНАЯ РЕГРЕССИЯ И КОРРЕЛЯЦИЯ

1.1 Теоретическая справка

Парная (простая) линейная регрессия представляет собой модель, где среднее значение зависимой (объясняемой) переменной рассматривается как функция одной независимой (объясняющей) переменной x , т.е. это модель вида:

$$\hat{y}_x = f(x). \quad (1.1)$$

Так же y называют результативным признаком, а x признаком-актором. Знак «^» означает, что между переменными x и y нет строгой функциональной зависимости. Практически в каждом отдельном случае величина y складывается из двух слагаемых:

$$y = \hat{y}_x + \varepsilon \quad (1.2)$$

где y – фактическое значение результативного признака; $\hat{y}_x$ - теоретическое значение результативного признака, найденное исходя из уравнения регрессии; ε – случайная величина, характеризующая отклонения реального значения результативного признака от теоретического, найденного по уравнению регрессии. Случайная величина ε называется также возмущением. Она включает влияние не учтенных в модели факторов, случайных ошибок и особенностей измерения. Ее присутствие в модели порождено тремя источниками: спецификацией модели, выборочным характером исходных данных, особенностями измерения переменных.

Различают *линейные* и *нелинейные* регрессии.

Линейная регрессия: $y = a + b \cdot x + \varepsilon$.

Нелинейные регрессии делятся на два класса: регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам, и регрессии, нелинейные по оцениваемым параметрам. Например:

регрессии, *нелинейные по объясняющим переменным*:

- полиномы разных степеней $y = a + b_1 \cdot x + b_2 \cdot x^2 + ... + b_n \cdot x^n + \varepsilon$;
- равносторонняя гипербола $y = a + \frac{b}{x} + \varepsilon$;

регрессии, *нелинейные по оцениваемым параметрам*:

- степенная $y = a \cdot x^b \cdot \varepsilon$;
- показательная $y = a \cdot b^x \cdot \varepsilon$;
- экспоненциальная $y = e^{a+bx+\varepsilon}$.

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 $\hat{y}_x$ минимальна, т.е.

$$\sum\left(y-\hat{y}_x\right)^2 \to \min \qquad (1.3)$$

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

$$\begin{cases} na + b\sum x = y\sum x \\ a\sum x + b\sum x^2 = \sum xy \end{cases} \qquad (1.4)$$

Можно воспользоваться готовыми формулами, которые вытекают непосредственно из решения этой системы:

$$a = \bar{y} - b\cdot\bar{x}, \qquad b = \frac{\operatorname{cov}(x,y)}{\sigma_x^2}, \qquad (1.5)$$

где $\operatorname{cov}(x,y) = \overline{y\cdot x} - \bar{y}\cdot\bar{x}$ – ковариация признаков *x* и *y* ,

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

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

Ковариация – числовая характеристика совместного распределения двух случайных величин, равная математическому ожиданию произведения отклонений этих случайных величин от их математических ожиданий. Дисперсия – характеристика случайной величины, определяемая как математическое ожидание квадрата отклонения случайной величины от ее математического ожидания. Математическое ожидание – сумма произведений значений случайной величины на соответствующие вероятности.

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

$$r_{xy} = b\cdot\frac{\sigma_x}{\sigma_y} = \frac{\operatorname{cov}(x,y)}{\sigma_x\cdot\sigma_y} \qquad (1.6)$$

и индекс корреляции ρ_{xy} - для нелинейной регрессии $0 \le \rho_{xy} \le 1$

$$\rho_{xy} = \sqrt{1-\frac{\sigma_{ост}^2}{\sigma_y^2}} = \sqrt{1-\frac{\sum\left(y-\hat{y}_x\right)^2}{\sum\left(y-\bar{y}\right)^2}},$$

где $\sigma_y^2 = \sum (y - \bar{y})^2$ – общая дисперсия результативного признака y; $\sigma_{ост}^2 = \sum (y - \hat{y}_x)^2$ – остаточная дисперсия, определяемая исходя из уравнения регрессии $\hat{y}_x = f(x)$.

Оценку качества построенной модели даст коэффициент (индекс) детерминации r_{xy}^2 (для линейной регрессии) либо ρ_{xy}^2 (для нелинейной регрессии), а также средняя ошибка аппроксимации.

Средняя ошибка аппроксимации – среднее отклонение расчетных значений от фактических:

$$\bar{A} = \frac{1}{n} \sum \left| \frac{y - \hat{y}}{y} \right| \cdot 100\% \qquad (1.7)$$

Допустимый придел значений $\bar{A}$ - не более 10%.

Средний коэффициент эластичности $\bar{Э}$ показывает, на сколько процентов в среднем по совокупности изменится результат y от своей средней величины при изменении фактора x на 1% от своего среднего значения:

$$\bar{Э} = f'(x) \frac{\bar{x}}{\bar{y}}. \qquad (1.8)$$

После того как найдено уравнение линейной регрессии, проводится *оценка значимости* как уравнения в целом, так и отдельных его параметров. Проверить значимость уравнения регрессии – значит установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным и достаточно ли включенных в уравнение объясняющих переменных (одной или нескольких) для описания зависимой переменной.

Оценка значимости уравнения регрессии в целом производится на основе *F -критерия Фишера*, которому предшествует дисперсионный анализ. Согласно основной идее дисперсионного анализа, общая сумма квадратов отклонений переменной y от среднего значения y раскладывается на две части – «*объясненную*» и «*необъясненную*»:

$$\sum (y - \bar{y})^2 = \sum (\hat{y}_x - \bar{y})^2 + \sum (y - \hat{y}_x)^2, \qquad (1.9)$$

где $\sum (y - \bar{y})^2$ – общая сумма квадратов отклонений;

$\sum (\hat{y}_x - \bar{y})^2$ – 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.

Схема дисперсионного анализа имеет вид, представленный в таблице 1.1 (n – число наблюдений, m – число параметров при переменной x).

Таблица 1.1 – Схема дисперсионного анализа

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

Определение дисперсии на одну степень свободы приводит дисперсии к сравнимому виду (напомним, что степени свободы – это числа, показывающие количество элементов варьирования, которые могут принимать произвольные значения, не изменяющие заданных характеристик). Сопоставляя факторную и остаточную дисперсии в расчете на одну степень свободы, получим величину F-критерия Фишера:

$$F=\frac{S^2_{факт}}{S^2_{ост}}. \qquad (1.10)$$

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

Для парной линейной регрессии m = 1, поэтому

$$F=\frac{S^2_{факт}}{S^2_{ост}}=\frac{\sum(\hat{y}_x-\bar{y})^2}{\sum(y-\hat{y}_x)^2}\cdot(n-2). \qquad (1.11)$$

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

$$F=\frac{r_{xy}{}^2}{1-r_{xy}{}^2}\cdot(n-2). \qquad (1.12)$$

Для оценки *статистической значимости параметров регрессии и корреляции* рассчитываются *t-критерий Стьюдента* и *доверитель-*

ные интервалы каждого из показателей. Оценка значимости коэффициентов регрессии и корреляции с помощью *t*-критерия Стьюдента проводится путем сопоставления их значений с величиной случайной ошибки:

$$t_b = \frac{b}{m_b}; t_a = \frac{a}{m_a}; t_r = \frac{r_{xy}}{m_r}. \quad (1.13)$$

Стандартные ошибки параметров линейной регрессии и коэффициента корреляции определяются по формулам:

$$m_b = \sqrt{\frac{\sum\left(y-\hat{y}_x\right)^2/\left(n-2\right)}{\sum\left(x-\overline{x}\right)^2}} = \sqrt{\frac{S_{ост}^2}{n\cdot\sigma_x^2}};$$

$$m_a = \sqrt{\frac{\sum\left(y-\hat{y}_x\right)^2}{\left(n-2\right)}\cdot\frac{\sum x^2}{n\sum\left(x-\overline{x}\right)^2}} = \sqrt{S_{ост}^2\frac{\sum x^2}{n^2\cdot\sigma_x^2}}; \quad (1.14)$$

$$m_{r_{xy}} = \sqrt{\frac{1-r_{xy}^{\ 2}}{n-2}}.$$

Сравнивая фактическое и критическое (табличное) значения *t*-статистики ($t_{табл}$ и $t_{факт}$) делаем вывод о значимости параметров регрессии и корреляции. Если $t_{табл} < t_{факт}$, то параметры *a*, *b* и r_{xy} не случайно отличаются от нуля и сформировались под влиянием систематически действующего фактора *x*. Если $t_{табл} > t_{факт}$, то признается случайная природа формирования *a*, *b* или r_{xy}.

Для расчета доверительного интервала определяем *предельную ошибку* Δ для каждого показателя:

$$\Delta_a = t_{табл} m_a, \Delta_b = t_{табл} m_b.$$

Формулы для расчета *доверительных интервалов* имеют следующий вид:

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

$$\gamma_b = b \pm \Delta_b; \gamma_{b_{\min}} = b - \Delta_b; \gamma_{b_{\max}} = b + \Delta_b;$$

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

Связь между *F*-критерием Фишера и *t*-статистикой Стьюдента выражается равенством

$$|t_r| = |t_b| = \sqrt{F}. \quad (1.15)$$

В прогнозных расчетах по уравнению регрессии определяется *предсказываемое индивидуальное значение* y_0 как точечный прогноз

при $x = x_0$, т.е. путем подстановки в линейное уравнение $\hat{y}_x = a + b \cdot x$ соответствующего значения x. Однако точечный прогноз явно нереален, поэтому он дополняется расчетом стандартной ошибки

$$m_{\hat{y}_0} = \sqrt{S_{ост}^2 \left(1 + \frac{1}{n} + \frac{\left(x_p - \overline{x}\right)^2}{\sum \left(x - \overline{x}\right)^2}\right)} = \sqrt{S_{ост}^2 \left(1 + \frac{1}{n} + \frac{\left(x_p - \overline{x}\right)^2}{n \cdot \sigma_x^2}\right)}, \qquad (1.16)$$

где $S_{ост}^2 = \frac{\sum \left(y - \hat{y}_x\right)^2}{n-2}$,

и построением *доверительного интервал* прогнозного значения y_0^*: $\hat{y}_0 - m_{\hat{y}_0} \cdot t_{табл} \le y_0^* \le \hat{y}_x + m_{\hat{y}_x} \cdot t_{табл}$

1.2 Решение типовой задачи в MS Excel с использованием надстройки Анализ данных

С помощью инструмента анализа данных Регрессия можно получить результаты регрессионной статистики, дисперсионного анализа, доверительных интервалов, остатки и графики подбора линии регрессии.

Если на вкладке данные еще нет команды Анализ данных, то необходимо ее установить. В главном меню последовательно выбираем Файл→Параметры→Надстройки в открывшемся диалоговом окне Управление надстройками Microsoft Excel в поле Управление выбираем Надстройки Excel и нажимаем Перейти. В открывшемся диалоговом окне Надстройки устанавливаем «флажок» в строке Пакет анализа (рис. 1.1).

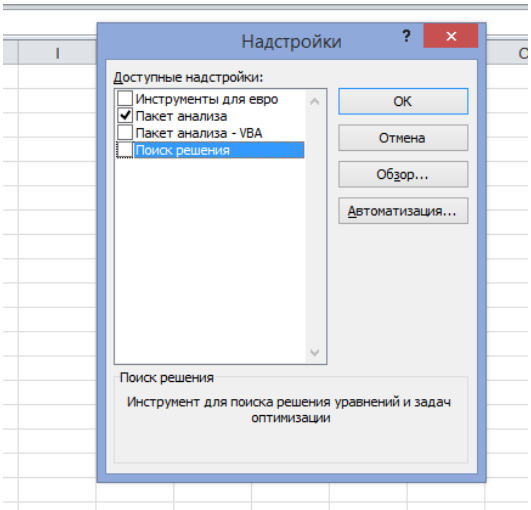


Рисунок 1.1 – Вид окна Надстройки

Пример. В таблице 1.2 приводятся данные по поголовью коров и производству сырого молока в хозяйствах всех категорий РФ за пятнадцать лет.

Таблица 1.2 – Исходные данные

Период	Поголовье коров, тыс. гол.	Производство сырого молока, тыс. т
1	10244,1	31861,2
2	9522,2	31069,9
3	9359,7	31339,0
4	9296,4	31984,2
5	9060,3	32225,7
6	8924,9	32315,1
7	8713,0	31507,8
8	8807,5	31204,3
9	8657,2	31196,8
10	8430,9	29865,2
11	8263,2	29995,2
12	8115,2	29887,5
13	7966,0	29787,2
14	7950,6	30184,5
15	7942,6	30639,7

Требуется:

1. Построить линейное уравнение парной регрессии y (производство молока) по x (поголовье коров).

2. Рассчитать коэффициент корреляции, нескорректированный коэффициент детерминации, скорректированный коэффициент детерминации и среднюю ошибку аппроксимации.

3. Оценить статистическую значимость уравнения регрессии в целом и отдельных параметров регрессии и корреляции с помощью F - критерия Фишера и t -критерия Стьюдента.

4. Выполнить прогноз производства молока при прогнозном значении поголовье коров, составляющем 107% от среднего уровня.

5. Оценить точность прогноза, рассчитав ошибку прогноза и его доверительный интервал.

Решение.

После установки пакета анализа, необходимо сформировать таблицу с исходными данными на листе книги MS Excel.

На вкладке Данные открыть Анализ данных, в открывшемся диалоговом окне выбрать инструмент анализа: Регрессия (рис. 1.2).

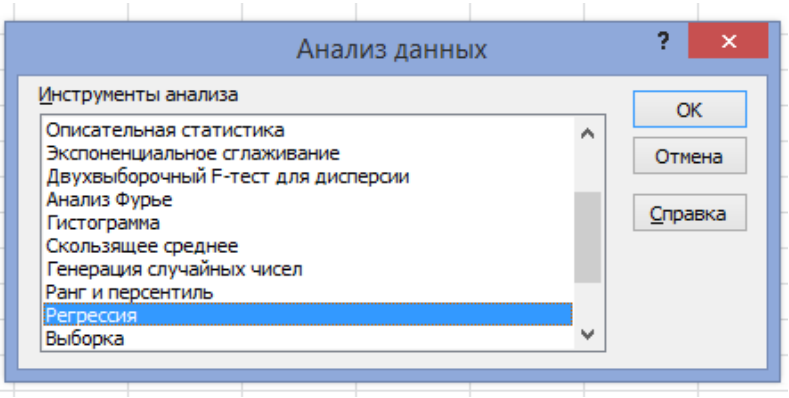


Рисунок 1.2 – Вид окна Анализ данных

Заполнить диалоговое окно Регрессия занеся в него данные из таблицы, определить параметры (рис. 1.3).

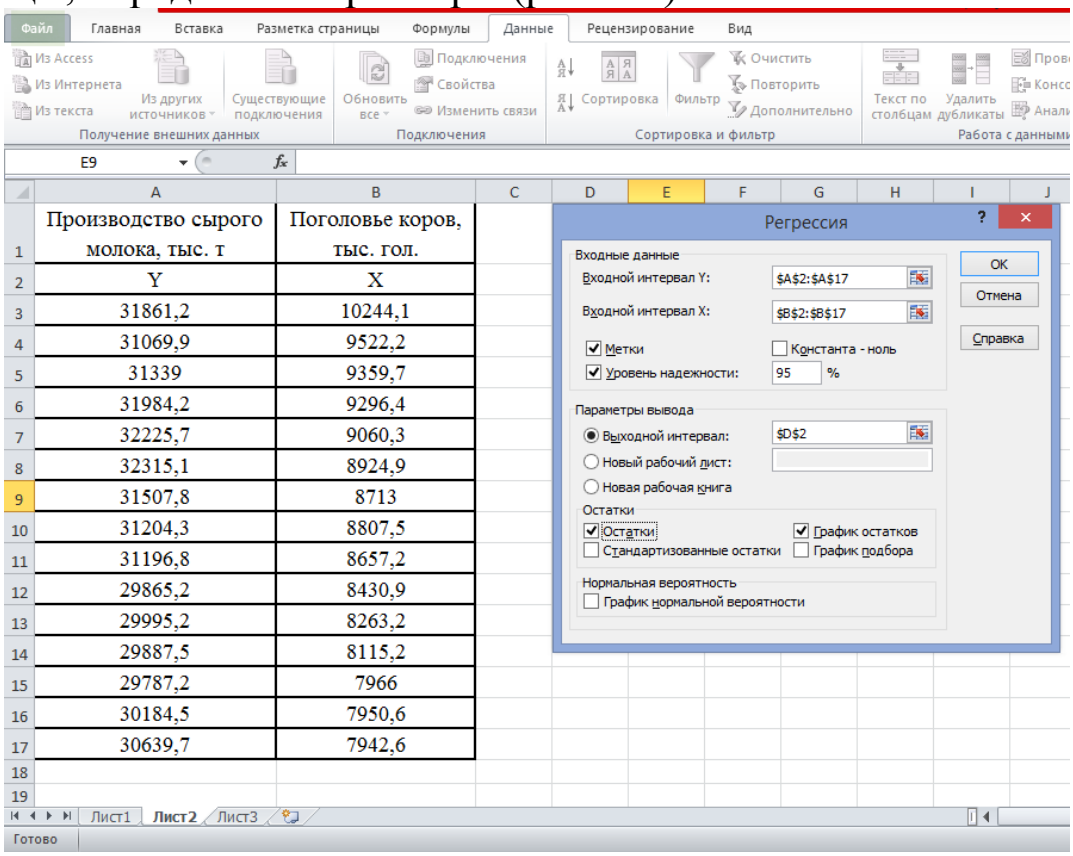


Рисунок 1.3 – Ввод данных

В диалоговом окне Регрессия:

Входной интервал Y – диапазон, содержащий данные результативного признака;

Входной интервал X – диапазон, содержащий данные признака-фактора;

Метки – «флажок», который указывает, содержи ли первая строка названия столбцов;

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 указать произвольное имя нового листа (или не указывать, тогда результаты выводятся на вновь созданный лист).

Результаты расчетов будут представлены в виде таблиц (рис. 1.4 и 1.5).

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,729485003
R-квадрат	0,532148369
Нормированный R-квадрат	0,496159782
Стандартная ошибка	634,9024276
Наблюдения	15

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	5960489,34	5960489,34	14,78658691	0,002025371
Остаток	13	5240314,204	403101,0926		
Итого	14	11200803,54			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	22483,83417	2221,828968	10,11951617	1,56728E-07	17683,86451	27283,80383	17683,86451	27283,80383
X	0,973730189	0,253223883	3,845333134	0,002025371	0,426673248	1,520787129	0,426673248	1,520787129

Рисунок 1.4 – Результаты расчетов

ВЫВОД ОСТАТКА

Наблюдение	*Предсказанное Y*	*Остатки*
1	32458,8236	-597,6235967
2	31755,88777	-685,9877734
3	31597,65662	-258,6566178
4	31536,0195	448,1805032
5	31306,1218	919,5782007
6	31174,27873	1140,821268
7	30967,9453	539,8546953
8	31059,96281	144,3371925
9	30913,61116	283,1888398
10	30693,25602	-828,0560185
11	30529,96147	-534,7614658
12	30385,8494	-498,3493979
13	30240,56885	-453,3688537
14	30225,57341	-41,0734088
15	30217,78357	421,9164327

Рисунок 1.5 – Вывод остатка

Выписываем, округляя до 4 знаков после запятой:

1. Уравнение регрессии:

$\hat{y}_x = 22483{,}83 + 0{,}9737x.$

Что свидетельствует о том, что при увеличении поголовья коров в хозяйствах всех категорий на 1 тыс. голов производство молока увеличивается на 0,9737 тыс. т.

2. Коэффициент корреляции:

$r_{xy} = 0{,}7295.$

Коэффициент выявляет высокую корреляцию между фактором и результативным признаком.

Нескорректированный коэффициент детерминации:

$R^2_{xy} = 0{,}5321.$

В данном случае он показывает, что на 53,21 % вариация производства молока обусловлена вариацией поголовья коров. Можно сделать вывод, что данный признак является существенным, но не объясненная доля так же является значительной и объясняется не включенными в модель признаками.

Скорректированный коэффициент детерминации:

$\hat{R}^2_{xy} = 0{,}4962.$

Он дает такую оценку тесноты связи, которая не зависит от числа факторов в модели и поэтому может сравниваться по моделям с разным числом факторов. Это позволит оценить улучшение модели в случае включения в нее дополнительных факторов.

С помощью средней ошибки аппроксимации можно оценить качество уравнений. На основании данных регрессионного анализа представленного на рисунках 1.4 и 1.5 можно составить новую таблицу, изображенную на рисунке 1.6. В столбце С произведено вычисление относительной ошибки аппроксимации по формуле:

$A_i = \left|\frac{(y-\hat{y}_x)}{y}\right| * 100.$

Средняя ошибка аппроксимации рассчитана по формуле:

$\bar{A} = \frac{1}{n} * \sum A_i = \frac{25{,}036}{15} = 1{,}67.$

Качество построенной модели оценивается как хорошее, так как значение средней ошибки аппроксимации не превышает 8 – 10 %.

C3 | =ABS(B3/A3*100)

	A	B	C	D
1	Производство сырого молока, тыс. т	Остатки	A	
2	Y			
3	31861,2	-597,624	1,8757	
4	31069,9	-685,988	2,2079	
5	31339	-258,657	0,8254	
6	31984,2	448,1805	1,4013	
7	32225,7	919,5782	2,8536	
8	32315,1	1140,821	3,5303	
9	31507,8	539,8547	1,7134	
10	31204,3	144,3372	0,4626	
11	31196,8	283,1888	0,9077	
12	29865,2	-828,056	2,7726	
13	29995,2	-534,761	1,7828	
14	29887,5	-498,349	1,6674	
15	29787,2	-453,369	1,522	
16	30184,5	-41,0734	0,1361	
17	30639,7	421,9164	1,377	
18	Итого		25,036	
19	Среднее значение		1,6691	

Рисунок 1.6 – Вычисление средней ошибки аппроксимации

3. Фактическое значение *F*-критерия Фишера:

F =14,7866.

Табличное значение критерия при уровне значимости α = 0,05 и $k_1 = m = 1$, $k_2 = n – m –1 = 15 – 1 –1 = 13$:

$F_{табл} = F(0,05; 1; 13) = 4,67$ (приложение 1).

Так как $F_{табл} < F_{набл}$, то с вероятностью $(1 – \alpha) = 0,95$ делаем заключение о значимости уравнения регрессии и коэффициента детерминации.

Стандартные ошибки для параметров регрессии:

$m_a = 2221,829$, $m_b = 0,2532$.

Они показывают, какое значение данной характеристики сформировалось под влиянием случайных факторов. Эти значения используются для расчета t-критерия Стьюдента.

Фактические значения t-критерия Стьюдента:

$t_a = 10,1195$, $t_b = 3,8453$.

Табличное значение критерия при уровне значимости α = 0,05 и $k = n – 2 = 15 – 2 = 13$: $t_{табл} = 2,1604$.

Так как $t_a > t_{табл}$ и $t_b > t_{табл}$, можно сделать вывод о существенности данных параметров, которые формируются под воздействием неслучайных причин. Статистически значимыми являются a и b.

На это же указывает показатель вероятности случайных значений параметров регрессии (P-Значение) если $\alpha_{факт}$ меньше принятого нами уровня (обычно 0,1; 0,05 или 0,01; это соответствует 10 %; 5 %

или 1 % вероятности), делают вывод о неслучайной природе данного значения параметра, т.е. статистически значим и надежен.

P-значение (a) = 0,00000016 < 0,01 < 0,05,

P-значение (b) = 0,002 < 0,01 < 0,05,

следовательно, коэффициенты (a, b) значимы при 1 %-ном уровне, а тем более при 5 %-ном уровне значимости. Таким образом, коэффициенты регрессии значимы и модель адекватна исходным данным.

4. Полученное уравнение регрессии позволяет использовать его для прогноза.

$\hat{y}_x$ = 22483,83+0,9737x.

Среднее значение $\bar{x}$ = 8750,25. Прогнозное значение поголовья коров, составляющее 107% от среднего уровня, имеет значение 9362,77 тыс. голов. Подставив прогнозное значение в уравнение регрессии получим прогнозное значение производства молока:

$\hat{y}_x = 22483{,}83 + 0{,}9737 * 9362{,}77 = 31600{,}36.$

5. Остаточная дисперсия на одну степень свободы:

$S^2_{\text{ост}} = 403101{,}0926.$

Корень квадратный из остаточной дисперсии (стандартная ошибка):

$S_{\text{ост}} = 634{,}9024,$

показывает на сколько велика ошибка предсказания значений переменной y на основании значений x.

Доверительные интервалы:

17683,8645 ≤ a* ≤ 27283,8038,

0,4267 ≤ b* ≤ 1,5208.

Анализ верхней и нижней границ доверительных интервалов приводит к выводу о том, что с вероятностью p =1 – α = 0,95 параметры a и b, находясь в указанных границах, не принимают нулевых значений, т.е. являются статистически значимыми и существенно отличны от нуля.

Ошибка прогноза рассчитанного в пункте 4 примера составит:

$y_{m0} = \sqrt{S^2_{\text{ост}} + (1 + \frac{1}{n} + \frac{(x_0 - \bar{x})\text{^}2}{n * \sigma_x^2})} = \sqrt{5240314{,}2036 + (1 + \frac{1}{15} + \frac{(9362{,}77 - 8750{,}25)^2}{15 * 419095{,}8758})},$

$y_{m0} = 2289{,}1735.$

При этом $\sigma_x^2 = \overline{x^2} - \bar{x}^2$.

Предельная ошибка прогноза, которая в 95 % случаев не будет превышена, составит:

$\Delta_{\hat{y}_0} = t_{\text{табл}} \cdot y_{m0} = 2{,}1604 \cdot 2289{,}1735 = 4945{,}5304.$

Доверительный интервал прогноза:

$\gamma_{\hat{y}_0} = \hat{y}_0 \pm \Delta_{\hat{y}_0} = 31600{,}36 \pm 4945{,}5304,$

$26654{,}8296 \leq y_0^* \leq 36545{,}8904.$

Выполненный прогноз производства молока является надежным (p=1 – a = 1 – 0,05 = 0,95) и находится в пределах от 26654,8296 тыс. т до 36545,8904 тыс. т.

1.3 Использование функции ЛИНЕЙН в MS Excel

Функция ЛИНЕЙН рассчитывает статистику для ряда с применением метода наименьших квадратов, чтобы вычислить прямую линию, которая наилучшим образом аппроксимирует имеющиеся данные и затем возвращает массив, который описывает полученную прямую.

Уравнение прямой имеет следующий вид:

y = a + b x

Синтаксис функции ЛИНЕЙН:

ЛИНЕЙН(известные_значения_y; [известные_значения_x]; [конст]; [статистика])

Аргументы функции ЛИНЕЙН:

Известные_значения_y. Обязательный аргумент. Множество значений y, которые уже известны для соотношения y = a + b x.

Если массив известные_значения_y имеет один столбец, то каждый столбец массива известные_значения_x интерпретируется как отдельная переменная.

Если массив известные_значения_y имеет одну строку, то каждая строка массива известные_значения_x интерпретируется как отдельная переменная.

Известные_значения_x. Необязательный аргумент. Множество значений x, которые уже известны для соотношения y = a + b x

Массив известные_значения_x может содержать одно или несколько множеств переменных. Если используется только одна переменная, то массивы известные_значения_y и известные_значения_x могут иметь любую форму — при условии, что они имеют одинаковую размерность. Если используется более одной переменной, то известные_значения_y должны быть вектором (т. е. интервалом высотой в одну строку или шириной в один столбец).

Если массив известные_значения_x опущен, то предполагается, что это массив {1;2;3;...}, имеющий такой же размер, что и массив известные_значения_y.

Конст. Необязательный аргумент. Логическое значение, которое указывает, требуется ли, чтобы константа b была равна 0.

Если аргумент конст имеет значение ИСТИНА, то константа b вычисляется обычным образом.

Если аргумент конст имеет значение ЛОЖЬ, то значение b полагается равным 0 и значения m подбираются таким образом, чтобы выполнялось соотношение y = bx.

Статистика. Необязательный аргумент. Логическое значение, которое указывает, требуется ли возвратить дополнительную регрессионную статистику.

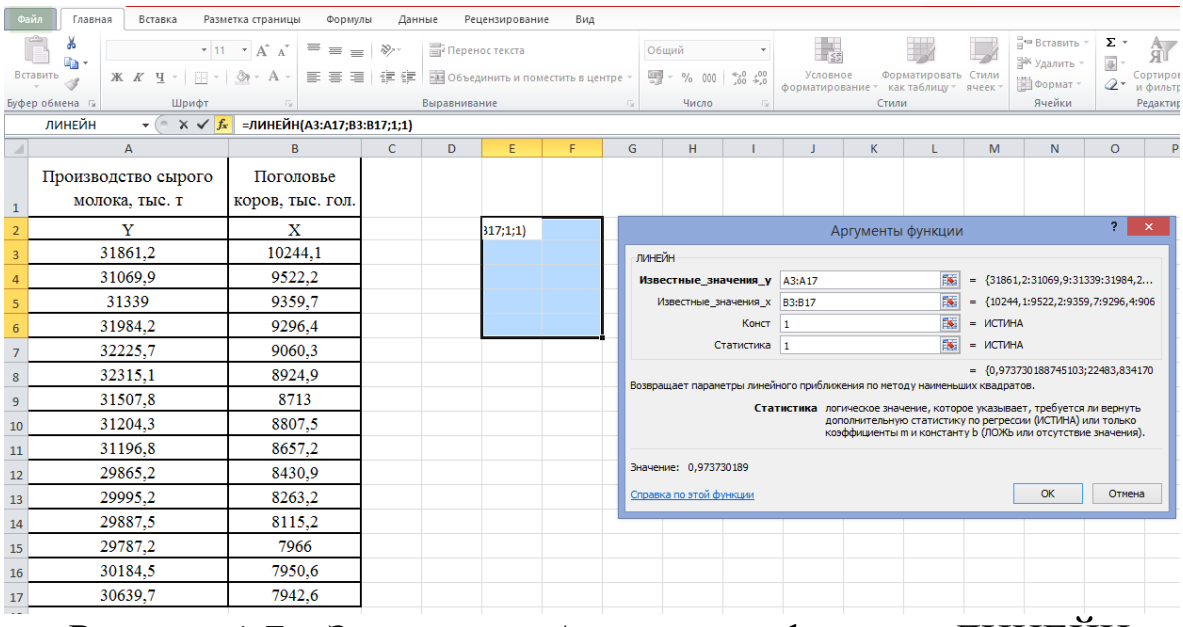
Если значение аргумента Статистика истинно, функция ЛИНЕЙН возвращает дополнительную статистику по регрессии.

Если аргумент статистика имеет значение ЛОЖЬ или опущен, функция ЛИНЕЙН возвращает только коэффициенты b и постоянную a.

Пример. Требуется построить регрессионную модель с использованием исходных данных, представленных в таблице 1.2.

На рабочем листе создается набор данных, т.е. вводится содержимое выборки наблюдений за поведением объекта. Это таблица, содержащая значения эндогенной и экзогенных переменных. Между строками данных не должно быть пустых ячеек.

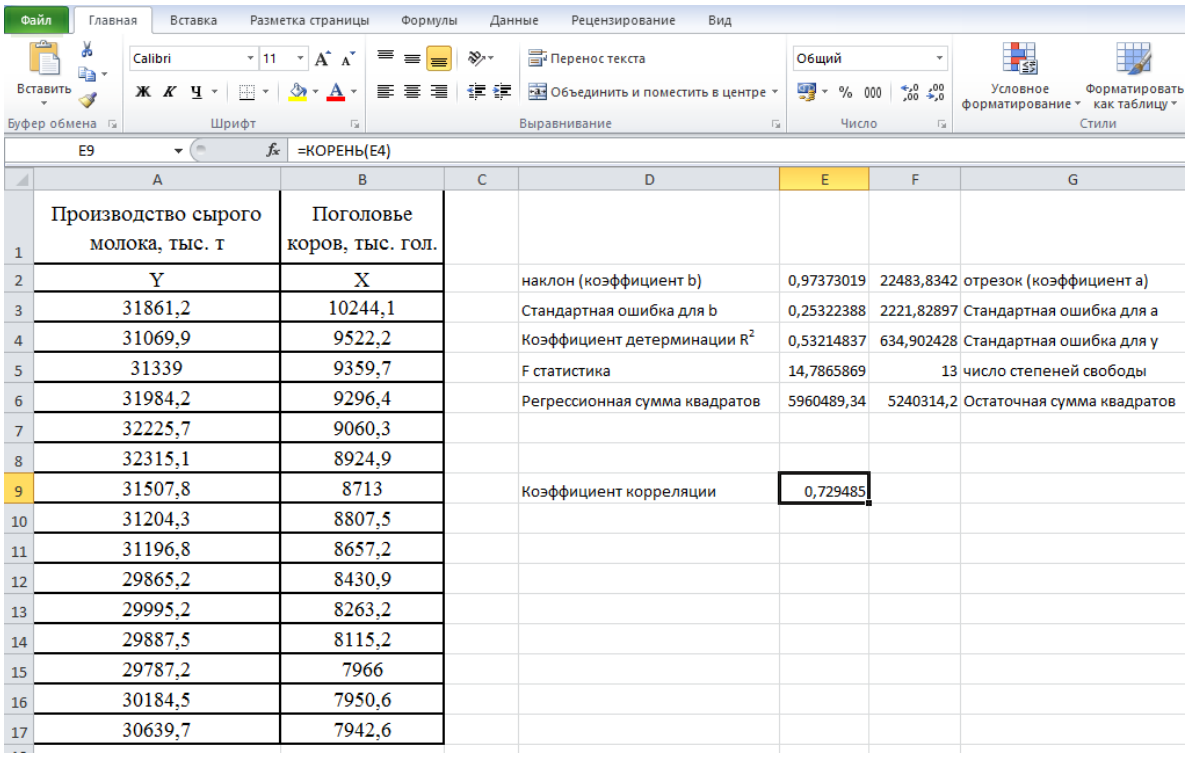
Если необходимо отобразить не только наклон и отрезок, но и дополнительные статистики, выделяется диапазон на один столбец больше, чем столбцов с переменными x, и высотой 5 строк (рис. 1.7). В примере одна переменная x, поэтому выделяется диапазон E2:F6 (2 столбца по 5 строк). Необходимо дополнительно ввести значения двух констант: «Конст» и «Статистика». С помощью первой константы отмечается присутствует (1) или нет (0) в спецификации модели свободный параметр a. Второй константе "Статистика" присваивается значение 1, так как нужно получить полную информацию, т.е. вывести дополнительные статистики. Ввод формулы осуществляется одновременным нажатием клавиш Ctrl+Shift+Enter. Результат должен соответствовать рисунку 1.8.



	A	B
1	Производство сырого молока, тыс. т	Поголовье коров, тыс. гол.
2	Y	X
3	31861,2	10244,1
4	31069,9	9522,2
5	31339	9359,7
6	31984,2	9296,4
7	32225,7	9060,3
8	32315,1	8924,9
9	31507,8	8713
10	31204,3	8807,5
11	31196,8	8657,2
12	29865,2	8430,9
13	29995,2	8263,2
14	29887,5	8115,2
15	29787,2	7966
16	30184,5	7950,6
17	30639,7	7942,6

Рисунок 1.7 – Заполнение Аргументов функции ЛИНЕЙН

На этом этапе наиболее частая ошибка – несинхронное нажатие клавиш Ctrl+Shift+Enter. Ошибка обнаруживается следующим образом: вместо ожидаемых десяти значений выводится одно. Для исправления необходимо вернуться в строку формул и повторить запуск (Ctrl+Shift+Enter) синхронно.



	A	B	C	D	E	F	G
1	Производство сырого молока, тыс. т	Поголовье коров, тыс. гол.					
2	Y	X		наклон (коэффициент b)	0,97373019	22483,8342	отрезок (коэффициент a)
3	31861,2	10244,1		Стандартная ошибка для b	0,25322388	2221,82897	Стандартная ошибка для a
4	31069,9	9522,2		Коэффициент детерминации R^2	0,53214837	634,902428	Стандартная ошибка для y
5	31339	9359,7		F статистика	14,7865869	13	число степеней свободы
6	31984,2	9296,4		Регрессионная сумма квадратов	5960489,34	5240314,2	Остаточная сумма квадратов
7	32225,7	9060,3					
8	32315,1	8924,9					
9	31507,8	8713		Коэффициент корреляции	0,729485		
10	31204,3	8807,5					
11	31196,8	8657,2					
12	29865,2	8430,9					
13	29995,2	8263,2					
14	29887,5	8115,2					
15	29787,2	7966					
16	30184,5	7950,6					
17	30639,7	7942,6					

Рисунок 1.8 – Результаты вычислений

Для вычисления коэффициента корреляции достаточно извлечь квадратный корень из значения коэффициента детерминации.

Полученные данные позволяют:

1. составить уравнение регрессии:

$\hat{y}_x = 22483{,}83 + 0{,}9737x$,

2. рассчитать коэффициент корреляции:

$r_{xy} = 0{,}7295$,

3. нескорректированный коэффициент детерминации:

$R_{xy}^2 = 0{,}5321$,

4. фактическое значение F -критерия Фишера:

F = 14,7866,

5. стандартные ошибки для параметров регрессии:

m_a = 2221,829, m_b = 0,2532,

Эти значения используются для расчета t-критерия Стьюдента.

$t_a = \frac{a}{m_a} = \frac{22483{,}83}{2221{,}829} = 10{,}119$,

$t_b = \frac{b}{m_b} = \frac{0{,}9737}{0{,}2532} = 3{,}845$,

6. корень квадратный из остаточной дисперсии (стандартная ошибка):

$S_{\text{ост}} = 634{,}9024$.

1.4 Линейная и нелинейная регрессия в MS Excel

В Excel имеется еще более быстрый и удобный способ построить график линейной регрессии и основных видов нелинейных регрессий. Это можно сделать следующим образом:

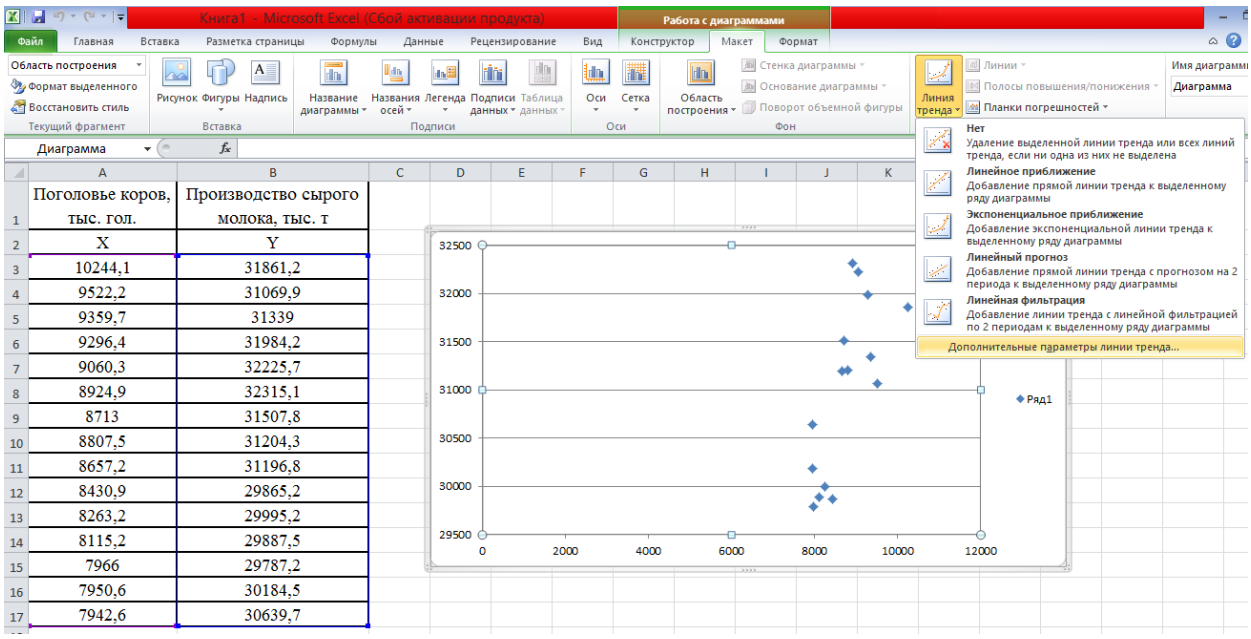
1) выделить столбцы с данными X и Y (они должны располагаться именно в таком порядке);

2) вызвать Мастер диаграмм и выбрать в группе Тип – Точечная, нажать Готово;

3) не сбрасывая выделения с диаграммы, выбрать появившейся пункт основного меню Работа с диаграммами / Макет, в котором следует выбрать пункт Линия тренда / Дополнительные параметры линии тренда (рис. 1.9);

4) в появившемся диалоговом окне Линия тренда во вкладке Тип выбрать один из предлагаемых типов;

5) во вкладке Параметры можно активизировать переключатель Показывать уравнение на диаграмме, что позволит увидеть уравнение линейной регрессии.



Поголовье коров, тыс. гол.	Производство сырого молока, тыс. т
X	Y
10244,1	31861,2
9522,2	31069,9
9359,7	31339
9296,4	31984,2
9060,3	32225,7
8924,9	32315,1
8713	31507,8
8807,5	31204,3
8657,2	31196,8
8430,9	29865,2
8263,2	29995,2
8115,2	29887,5
7966	29787,2
7950,6	30184,5
7942,6	30639,7

Рисунок 1.9 – Вид окна Линия тренда

В этой же вкладке можно активизировать переключатель Поместить на диаграмму величину достоверности аппроксимации (R^2). Это значение коэффициента детерминации (квадрат коэффициента корреляции), и оно показывает, насколько хорошо рассчитанное уравнение описывает экспериментальную зависимость. Если R^2 близок к единице, то теоретическое уравнение регрессии хорошо описывает экспериментальную зависимость (теория хорошо согласуется с экспериментом), а если R^2 близок к нулю, то данное уравнение не пригодно для описания экспериментальной зависимости (теория не согласуется с экспериментом) (рис. 1.10).

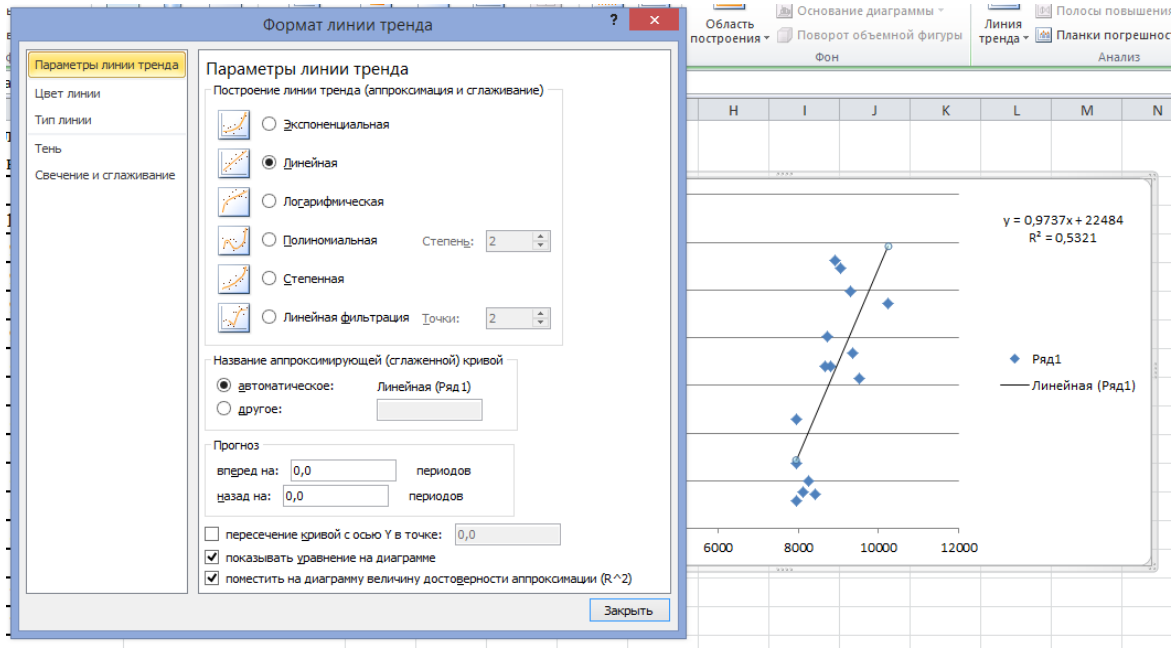


Рисунок 1.10 – Выбор типа уравнения регрессии

В результате выполнения описанных действий получится диаграмма с графиком регрессии и ее уравнение.

На рисунке 1.10 представлена линейная регрессия, на графике отражено:

уравнение регрессии:

$\hat{y}_x = 22483{,}83 + 0{,}9737x,$

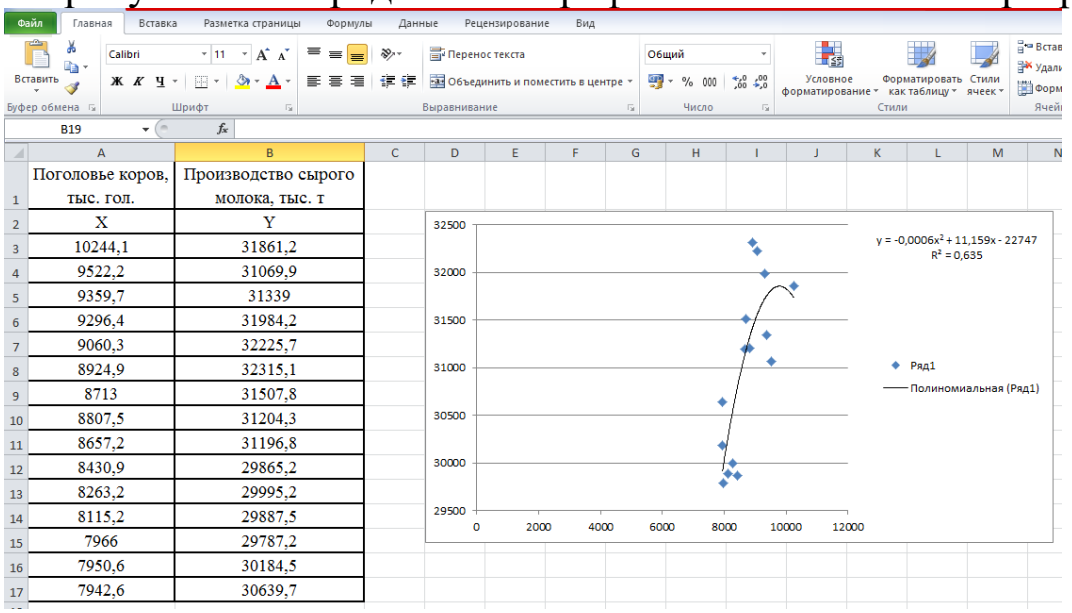
нескорректированный коэффициент детерминации:

$R^2_{xy} = 0{,}5321.$

Можно рассчитать коэффициент корреляции, извлекая квадратный корень из коэффициента детерминации:

$r_{xy} = 0{,}7295.$

На рисунке 1.11 представлен график полиномиальной регрессии.



	A	B
1	Поголовье коров, тыс. гол.	Производство сырого молока, тыс. т
2	X	Y
3	10244,1	31861,2
4	9522,2	31069,9
5	9359,7	31339
6	9296,4	31984,2
7	9060,3	32225,7
8	8924,9	32315,1
9	8713	31507,8
10	8807,5	31204,3
11	8657,2	31196,8
12	8430,9	29865,2
13	8263,2	29995,2
14	8115,2	29887,5
15	7966	29787,2
16	7950,6	30184,5
17	7942,6	30639,7

Рисунок 1.11 – Построение полиномиального уравнения регрессии и его графического изображения

На графике так же отражено:

уравнение регрессии:

$\hat{y}_x = -0{,}0006x^2 + 11{,}159x - 22747,$

нескорректированный коэффициент детерминации:

$R^2_{xy} = 0{,}635.$

Можно рассчитать коэффициент корреляции, извлекая квадратный корень из коэффициента детерминации:

$r_{xy} = 0{,}7969.$

Задачи для самостоятельного решения

Задача 1.

В таблице 1.3 приводятся данные по производству молока в хозяйствах всех категорий и урожайности кормовых корнеплодов (включая сахарную свеклу на корм) за семнадцать лет.

Требуется:

1. Построить линейное уравнение парной регрессии y (производство молока в хозяйствах всех категорий) по x (урожайность кормовых корнеплодов (включая сахарную свеклу на корм)).

2. Рассчитать коэффициент корреляции, нескорректированный коэффициент детерминации, скорректированный коэффициент детерминации и среднюю ошибку аппроксимации.

3. Оценить статистическую значимость уравнения регрессии в целом и отдельных параметров регрессии и корреляции с помощью F - критерия Фишера и t -критерия Стьюдента.

4. Выполнить прогноз производства молока при прогнозном значении урожайности кормовых корнеплодов, составляющем 105 % от среднего уровня.

5. Оценить точность прогноза, рассчитав ошибку прогноза и его доверительный интервал.

Задача 2.

В таблице 1.4 приводятся данные по производству молока в расчете на 100 га сельскохозяйственных угодий и энергообеспеченности сельскохозяйственных организаций (энергетические мощности в расчете на 100 га посевной площади) за восемнадцать лет.

Требуется:

1. Построить линейное уравнение парной регрессии y (производство молока в расчете на 100 га сельскохозяйственных угодий) по x (энергообеспеченность сельскохозяйственных организаций).

2. Рассчитать коэффициент корреляции, нескорректированный коэффициент детерминации, скорректированный коэффициент детерминации и среднюю ошибку аппроксимации.

3. Оценить статистическую значимость уравнения регрессии в целом и отдельных параметров регрессии и корреляции с помощью F - критерия Фишера и t -критерия Стьюдента.

4. Выполнить прогноз производства молока в расчете на 100 га сельскохозяйственных угодий при прогнозном значении энергообес-

печенности сельскохозяйственных организаций, составляющем 106 % от среднего уровня.

5. Оценить точность прогноза, рассчитав ошибку прогноза и его доверительный интервал.

Задача № 3

В таблице 1.5 приводятся данные по производству молока в расчете на 100 га сельскохозяйственных угодий и урожайности кормовых корнеплодов (включая сахарную свеклу на корм) за восемнадцать лет.

Требуется:

1. Построить линейное уравнение парной регрессии y (производство молока в расчете на 100 га сельскохозяйственных угодий) по x (урожайность кормовых корнеплодов (включая сахарную свеклу на корм)).

2. Рассчитать коэффициент корреляции, нескорректированный коэффициент детерминации, скорректированный коэффициент детерминации и среднюю ошибку аппроксимации.

3. Оценить статистическую значимость уравнения регрессии в целом и отдельных параметров регрессии и корреляции с помощью F - критерия Фишера и t -критерия Стьюдента.

4. Выполнить прогноз производства молока в расчете на 100 га сельскохозяйственных угодий при прогнозном значении урожайности кормовых корнеплодов, составляющем 108 % от среднего уровня.

5. Оценить точность прогноза, рассчитав ошибку прогноза и его доверительный интервал.

Задача № 4

В таблице 1.6 приводятся данные по средним потребительским ценам на молоко и сыры сычужные за двадцать лет.

Требуется:

1. Построить линейное уравнение парной регрессии y (средние потребительские цены на сыры сычужные твердые и мягкие) по x (средние потребительские цены на молоко питьевое цельное).

2. Рассчитать коэффициент корреляции, нескорректированный коэффициент детерминации, скорректированный коэффициент детерминации и среднюю ошибку аппроксимации.

3. Оценить статистическую значимость уравнения регрессии в целом и отдельных параметров регрессии и корреляции с помощью F - критерия Фишера и t -критерия Стьюдента.

4. Выполнить прогноз средней потребительской цены на сыры сычужные твердые и мягкие при прогнозном значении средней потребительской цены на молоко, составляющем 103 % от среднего уровня.

5. Оценить точность прогноза, рассчитав ошибку прогноза и его доверительный интервал.

Задача № 5

В таблице 1.7 приводятся данные по производству молока в расчете на 100 га сельскохозяйственных угодий и энерговооруженности труда в сельскохозяйственных организациях (энергетические мощности в расчете на 1 работника) за восемнадцать лет.

Требуется:

1. Построить линейное уравнение парной регрессии y (производство молока в расчете на 100 га сельскохозяйственных угодий) по x (энерговооруженность труда в сельскохозяйственных организациях (энергетические мощности в расчете на 1 работника)).

2. Рассчитать коэффициент корреляции, нескорректированный коэффициент детерминации, скорректированный коэффициент детерминации и среднюю ошибку аппроксимации.

3. Оценить статистическую значимость уравнения регрессии в целом и отдельных параметров регрессии и корреляции с помощью F - критерия Фишера и t -критерия Стьюдента.

4. Выполнить прогноз производства молока в расчете на 100 га сельскохозяйственных угодий при прогнозном значении энерговооруженности труда в сельскохозяйственных организациях, составляющем 109 % от среднего уровня.

5. Оценить точность прогноза, рассчитав ошибку прогноза и его доверительный интервал.

Задача 6.

В таблице 1.3 приводятся данные по производству молока в хозяйствах всех категорий и урожайности кормовых корнеплодов (включая сахарную свеклу на корм) за семнадцать лет.

Требуется построить линейное уравнение парной регрессии y (производство молока в хозяйствах всех категорий) по x (урожай-

ность кормовых корнеплодов (включая сахарную свеклу на корм)) и рассчитать дополнительные параметры, используя функцию ЛИНЕЙН в MS Excel. Оценить статистическую значимость уравнения регрессии в целом и отдельных параметров регрессии и корреляции с помощью F - критерия Фишера и t -критерия Стьюдента.

Задача 7.

В таблице 1.4 приводятся данные по производству молока в расчете на 100 га сельскохозяйственных угодий и энергообеспеченности сельскохозяйственных организаций (энергетические мощности в расчете на 100 га посевной площади) за восемнадцать лет.

Требуется построить линейное уравнение парной регрессии y (производство молока в расчете на 100 га сельскохозяйственных угодий) по x (энергообеспеченность сельскохозяйственных организаций) и рассчитать дополнительные параметры, используя функцию ЛИНЕЙН в MS Excel. Оценить статистическую значимость уравнения регрессии в целом и отдельных параметров регрессии и корреляции с помощью F - критерия Фишера и t -критерия Стьюдента.

Задача 8.

В таблице 1.5 приводятся данные по производству молока в расчете на 100 га сельскохозяйственных угодий и урожайности кормовых корнеплодов (включая сахарную свеклу на корм) за восемнадцать лет.

Требуется построить линейное уравнение парной регрессии y (производство молока в расчете на 100 га сельскохозяйственных угодий) по x (урожайность кормовых корнеплодов (включая сахарную свеклу на корм)) и рассчитать дополнительные параметры, используя функцию ЛИНЕЙН в MS Excel. Оценить статистическую значимость уравнения регрессии в целом и отдельных параметров регрессии и корреляции с помощью F - критерия Фишера и t -критерия Стьюдента.

Задача 9.

В таблице 1.6 приводятся данные по средним потребительским ценам на молоко и сыры сычужные за двадцать лет.

Требуется построить линейное уравнение парной регрессии y (средние потребительские цены на сыры сычужные твердые и мягкие) по x (средние потребительские цены на молоко питьевое цель-

ное) и рассчитать дополнительные параметры, используя функцию ЛИНЕЙН в MS Excel. Оценить статистическую значимость уравнения регрессии в целом и отдельных параметров регрессии и корреляции с помощью F - критерия Фишера и t -критерия Стьюдента.

Задача 10.

В таблице 1.7 приводятся данные по производству молока в расчете на 100 га сельскохозяйственных угодий и энерговооруженности труда в сельскохозяйственных организациях (энергетические мощности в расчете на 1 работника) за восемнадцать лет.

Требуется построить линейное уравнение парной регрессии y (производство молока в расчете на 100 га сельскохозяйственных угодий) по x (энерговооруженность труда в сельскохозяйственных организациях (энергетические мощности в расчете на 1 работника)) и рассчитать дополнительные параметры, используя функцию ЛИНЕЙН в MS Excel. Оценить статистическую значимость уравнения регрессии в целом и отдельных параметров регрессии и корреляции с помощью F - критерия Фишера и t -критерия Стьюдента.

Задача 11.

В таблице 1.3 приводятся данные по производству молока в хозяйствах всех категорий и урожайности кормовых корнеплодов (включая сахарную свеклу на корм) за семнадцать лет.

Требуется построить точечный график. Подобрать уравнение регрессии y (производство молока в хозяйствах всех категорий) по x (урожайность кормовых корнеплодов (включая сахарную свеклу на корм)) лучшим образом аппроксимирующим фактические данные, рассчитать для него значение коэффициента корреляции.

Задача 12.

В таблице 1.4 приводятся данные по производству молока в расчете на 100 га сельскохозяйственных угодий и энергообеспеченности сельскохозяйственных организаций (энергетические мощности в расчете на 100 га посевной площади) за восемнадцать лет.

Требуется построить точечный график. Подобрать уравнение регрессии y (производство молока в расчете на 100 га сельскохозяйственных угодий) по x (энергообеспеченность сельскохозяйственных организаций) лучшим образом аппроксимирующим фактические данные, рассчитать для него значение коэффициента корреляции.

Задача 13.

В таблице 1.5 приводятся данные по производству молока в расчете на 100 га сельскохозяйственных угодий и урожайности кормовых корнеплодов (включая сахарную свеклу на корм) за восемнадцать лет.

Требуется построить точечный график. Подобрать уравнение регрессии y (производство молока в расчете на 100 га сельскохозяйственных угодий) по x (урожайность кормовых корнеплодов (включая сахарную свеклу на корм)) лучшим образом аппроксимирующим фактические данные, рассчитать для него значение коэффициента корреляции.

Задача 14.

В таблице 1.6 приводятся данные по средним потребительским ценам на молоко и сыры сычужные за двадцать лет.

Требуется построить точечный график. Подобрать уравнение регрессии y (средние потребительские цены на сыры сычужные твердые и мягкие) по x (средние потребительские цены на молоко питьевое цельное) лучшим образом аппроксимирующим фактические данные, рассчитать для него значение коэффициента корреляции.

Задача 15.

В таблице 1.7 приводятся данные по производству молока в расчете на 100 га сельскохозяйственных угодий и энерговооруженности труда в сельскохозяйственных организациях (энергетические мощности в расчете на 1 работника) за восемнадцать лет.

Требуется построить точечный график. Подобрать уравнение регрессии y (производство молока в расчете на 100 га сельскохозяйственных угодий) по x (энерговооруженность труда в сельскохозяйственных организациях (энергетические мощности в расчете на 1 работника)) лучшим образом аппроксимирующим фактические данные, рассчитать для него значение коэффициента корреляции.

Таблица 1.3 – Исходные данные к задачам № 1, 6, 11

Показатели	2001	2002	2003	2004	2005	2006	2007	2008
Производство молока в в хозяйствах всех категорий, млн. т	32,9	33,5	33,3	31,9	31,1	31,3	32,0	32,2
Урожайность кормовых корнеплодов (включая сахарную свеклу на корм), ц на 1 га	209	199	228	230	238	258	258	265

Продолжение таблицы 1.3 – Исходные данные к задачам № 1, 6, 11

Показатели	2009	2010	2011	2012	2013	2014	2015	2016	2017
Производство молока в в хозяйствах всех категорий, млн. т	32,3	31,5	31,2	31,2	29,8	30,0	29,9	29,8	30,2
Урожайность кормовых корнеплодов (включая сахарную свеклу на корм), ц на 1 га	267	209	189	275	249	273	253	267	255

Таблица 1.4 – Исходные данные к задачам № 2, 7, 12

Показатели	2000	2001	2002	2003	2004	2005	2006	2007	2008
Энергообеспеченность сельскохозяйственных организаций (энергетические мощности в расчете на 100 га посевной площади)	329	321	304	303	283	270	254	243	234
Производство молока в расчете на 100 га сельскохозяйственных угодий, тонн	9,4	9,8	10,4	10,2	9,7	9,9	10,2	11,0	11,4

Продолжение таблицы 1.4 – Исходные данные к задачам № 2, 7, 12

Показатели	2009	2010	2011	2012	2013	2014	2015	2016	2017
Энергообеспеченность сельскохозяйственных организаций (энергетические мощности в расчете на 100 га посевной площади)	227	227	212	211	201	201	197	200	198
Производство молока в расчете на 100 га сельскохозяйственных угодий, тонн	11,8	11,7	11,9	12,3	11,9	12,3	12,0	13,0	13,6

Таблица 1.5 – Исходные данные к задачам № 3, 8, 13

Показатели	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
Производство молока в расчете на 100 га сельскохозяйственных угодий, тонн	9,4	9,8	10,4	10,2	9,7	9,9	10,2	11,0
Урожайность кормовых корнеплодов (включая сахарную свеклу на корм), ц на 1 га	209	199	228	230	238	258	258	265

Продолжение таблицы 1.5 – Исходные данные к задачам № 3, 8, 13

Показатели	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**
Производство молока в расчете на 100 га сельскохозяйственных угодий, тонн	11,8	11,7	11,9	12,3	11,9	12,3	12,0	13,0	13,6
Урожайность кормовых корнеплодов (включая сахарную свеклу на корм), ц на 1 га	267	209	189	275	249	273	253	267	255

Таблица 1.6 – Исходные данные к задачам № 4, 9, 14

Показатели	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Средние потребительские цены на молоко питьевое цельное за 1 л (на конец года, рублей за кг, в масштабе цен соответствующих лет)	9,70	11,37	11,96	13,48	15,52	17,35	18,76	25,39	28,09	26,75
Средние потребительские цены на сыры сычужные твердые и мягкие (на конец года, рублей за кг, в масштабе цен соответствующих лет)	85,17	103,06	102,67	111,95	122,30	138,72	144,26	233,93	212,92	213,11

Продолжение таблицы 1.6 – Исходные данные к задачам № 4, 9, 14

Показатели	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Средние потребительские цены на молоко питьевое цельное за 1 л (на конец года, рублей за кг, в масштабе цен соответствующих лет)	31,99	32,52	33,88	38,64	43,81	47,61	51,44	53,45	54,04	57,70
Средние потребительские цены на сыры сычужные твердые и мягкие (на конец года, рублей за кг, в масштабе цен соответствующих лет)	263,20	273,43	272,57	326,89	388,81	418,61	461,71	478,88	502,55	552,03

Таблица 1.7 – Исходные данные к задачам № 5, 10, 15

Показатели	2000	2001	2002	2003	2004	2005	2006	2007	2008
Энерговооруженность труда в сельскохозяйственных организациях (энергетические мощности в расчете на 1 работника)	51	54	54	55	57	59	60	63	63
Производство молока в расчете на 100 га сельскохозяйственных угодий, тонн	9,4	9,8	10,4	10,2	9,7	9,9	10,2	11,0	11,4

Продолжение таблицы 1.7 – Исходные данные к задачам № 5, 10, 15

Показатели	2009	2010	2011	2012	2013	2014	2015	2016	2017
Энерговооруженность труда в сельскохозяйственных организациях (энергетические мощности в расчете на 1 работника)	61	67	69	70	72	75	74	77	75
Производство молока в расчете на 100 га сельскохозяйственных угодий, тонн	11,8	11,7	11,9	12,3	11,9	12,3	12,0	13,0	13,6

2.МНОЖЕСТВЕННАЯ РЕГРЕССИЯ И КОРРЕЛЯЦИЯ

2.1.Теоретическая справка

Множественная регрессия – это уравнение связи с несколькими независимыми переменными:

$$y = f\left(x_1, x_2, ..., x_m\right) + \varepsilon,$$

где y – зависимая переменная (результативный признак); $x_1, x_2, ..., x_m$ –независимые переменные (признаки-факторы).

Для построения уравнения множественной регрессии чаще используются следующие функции:

-линейная – $y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_m \cdot x_m + \varepsilon$;

-степенная – $y = a \cdot x_1^{b_1} \cdot x_2^{b_2} \cdot ... \cdot x_m^{b_m} \cdot \varepsilon$;

-экспонента – $y = e^{a + b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_m \cdot x_m + \varepsilon}$;

-гипербола – $y = \dfrac{1}{a + b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_m \cdot x_m + \varepsilon}$

Можно использовать и другие функции, приводимые к линейному виду.

Для оценки параметров уравнения множественной регрессий применяют *метод наименьших квадратов* (МНК). Для линейных уравнений

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_m \cdot x_m + \varepsilon \qquad (2.1)$$

строится следующая система нормальных уравнений, решение которой позволяет получить оценки параметров регрессии:

$$\begin{cases} \sum y = na + b_1 \sum x_1 + b_2 \sum x_2 + ... + b_m \sum x_m, \\ \sum yx_1 = a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 + ... + b_m \sum x_m x_1, \\ ..... \\ \sum yx_m = a \sum x_m + b_1 \sum x_1 x_m + b_2 \sum x_2 x_m + b_m \sum x_m^2, \end{cases} \qquad (2.2)$$

Для двухфакторной модели данная система будет иметь вид:

$$\begin{cases} na + b_1 \sum x_1 + b_2 \sum x_2 = \sum y, \\ a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 = \sum yx_1 \\ a \sum x_2 + b_2 \sum x_2^2 + b_1 \sum x_1 x_2 = \sum yx_2 \end{cases} \qquad (2.3)$$

Так же можно воспользоваться готовыми формулами, которые являются следствием из этой системы:

$$b_1 = \frac{\sigma_y}{\sigma_{x_1}} \cdot \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{1 - r^2_{x_1x_2}};$$

$$b_2 = \frac{\sigma_y}{\sigma_{x_2}} \cdot \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{1 - r^2_{x_1x_2}}; \qquad (2.4)$$

$$a = \overline{y} - b_1\overline{x}_1 - b_2\overline{x}_2.$$

В линейной множественной регрессии параметры при *x* называются *коэффициентами «чистой» регрессии*. Они характеризуют среднее изменение результата с изменением соответствующего фактора на единицу при неизмененном значении других факторов, закрепленных на среднем уровне.

Метод наименьших квадратов применим и к уравнению множественной регрессии в *стандартизированном масштабе*:

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + ... + \beta_m t_{x_m} + \varepsilon, \qquad (2.5)$$

где $t_y, t_{x_1}, ..., t_{x_m}$ – *стандартизированные 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: $\sigma_{t_y} = \sigma_{t_{x_i}} = 1$; β_i - *стандартизированные коэффициенты регрессии*.

В силу того, что все переменные заданы как центрированные и нормированные, стандартизованные коэффициенты регрессии β_i можно сравнивать 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 «чистой» регрессии, которые несравнимы между собой.

Применяя МНК к уравнению множественной регрессии в стандартизированном масштабе, получим систему нормальных уравнений вида

$$\begin{cases} r_{yx_1} = \beta_1 + \beta_2 r_{x_1x_2} + \beta_3 r_{x_1x_3} + ... + \beta_m r_{x_1x_m}, \\ r_{yx_2} = \beta_2 + \beta_1 r_{x_1x_2} + \beta_3 r_{x_1x_3} + ... + \beta_m r_{x_1x_m}, \\ ..... \\ r_{yx_m} = \beta_1 r_{x_1x_m} + \beta_2 r_{x_2x_m} + \beta_3 r_{x_3x_m} + ... + \beta_m, \end{cases} \qquad (2.6)$$

где $r_{yx_i}, r_{x_ix_j}$ – коэффициенты парной и межфакторной корреляции.

Коэффициенты «чистой» регрессии b_i связаны со стандартизованными коэффициентами регрессии β_i следующим образом:

$$b_i = \beta_i \frac{\sigma_y}{\sigma_{x_i}} \left(\beta_i = b_i \frac{\sigma_{x_i}}{\sigma_y} \right). \qquad (2.7)$$

Поэтому можно переходить от уравнения регрессии в стандартизованном масштабе (3.5) к уравнению регрессии в натуральном масштабе переменных (3.1), при этом параметр *a* определяется как $a = \overline{y} - b_1\overline{x}_1 - b_2\overline{x}_2 - ... - b_m\overline{x}_m$.

Рассмотренный смысл стандартизованных коэффициентов регрессии позволяет их использовать при отсеве факторов – из модели исключаются факторы с наименьшим значением β_i.

Средние коэффициенты эластичности для линейной регрессии рассчитываются по формуле:

$$\overline{Э}_{yx_j} = b_j \frac{\overline{x}_j}{\overline{y}}, \qquad (2.8)$$

которые показывают на сколько процентов в среднем изменится результат, при изменении соответствующего фактора на 1%. Средние показатели эластичности можно сравнивать друг с другом и соответственно ранжировать факторы по силе их воздействия на результат.

Тесноту совместного влияния факторов на результат оценивает *индекс множественной корреляции:*

$$R_{yx_1x_2...x_m} = \sqrt{1 - \frac{\sigma^2_{y_{ост}}}{\sigma^2_y}}. \qquad (2.9)$$

Значение индекса множественной корреляции лежит в пределах от 0 до 1 и должно быть больше или равно максимальному парному индексу корреляции:

$$R_{yx_1x_2...x_m} \geq r_{yx_i} (i = \overline{1,m}). \qquad (2.10)$$

При линейной зависимости *коэффициент множественной корреляции* можно определить через матрицы парных коэффициентов корреляции:

$$R_{yx_1x_2...x_m} = \sqrt{1 - \frac{\Delta r}{\Delta r_{11}}}, \qquad (2.11)$$

где

$$\Delta r = \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} & \dots & r_{yx_m} \\ r_{yx_1} & 1 & r_{x_1x_2} & \dots & r_{x_1x_m} \\ r_{yx_2} & r_{x_2x_1} & 1 & \dots & r_{x_2x_m} \\ \dots & \dots & \dots & \dots & \dots \\ r_{yx_m} & r_{x_mx_1} & r_{x_mx_2} & \dots & 1 \end{vmatrix}$$

– определитель матрицы парных коэффициентов корреляции;

$$\Delta r_{11} = \begin{vmatrix} 1 & r_{x_1x_2} & \dots & r_{x_1x_m} \\ r_{x_2x_1} & 1 & \dots & r_{x_2x_m} \\ \dots & \dots & \dots & \dots \\ r_{x_mx_1} & r_{x_mx_2} & \dots & 1 \end{vmatrix}$$

– определитель матрицы межфакторной корреляции.

Так же при линейной зависимости признаков формула коэффициента множественной корреляции может быть также представлена следующим выражением:

$$R_{yx_1x_2\dots x_m} = \sqrt{\sum \beta_i \cdot r_{yx_i}}, \qquad (2.12)$$

где β_i – стандартизованные коэффициенты регрессии; r_{yx_i} – парные коэффициенты корреляции результата с каждым фактором.

Качество построенной модели в целом оценивает коэффициент (индекс) детерминации. *Коэффициент множественной детерминации* рассчитывается как квадрат индекса множественной корреляции $R^2_{yx_1x_2\dots x_m}$

Для того чтобы не допустить преувеличения тесноты связи, применяется *скорректированный индекс множественной детерминации*, который содержит поправку на число степеней свободы и рассчитывается по формуле

$$\hat{R}^2 = 1 - (1 - R^2)\frac{(n-1)}{(n-m-1)}, \qquad (2.13)$$

где n – число наблюдений, m – число факторов. При небольшом числе наблюдений нескорректированная величина коэффициента множественной детерминации R^2 имеет тенденцию переоценивать долю вариации результативного признака, связанную с влиянием факторов, включенных в регрессионную модель.

Частные коэффициенты (или индексы) корреляции, измеряющие влияние на *y* фактора x_i , при элиминировании (исключении влияния) других факторов, можно определить по формуле

$$r_{yx_ix_1x_2\ldots x_{i-1}x_{i+1}\ldots x_m}=\sqrt{1-\frac{1-R_{yx_1x_2\ldots x_i\ldots x_m}}{1-R^2_{yx_ix_1x_2\ldots x_{i-1}x_{i+1}\ldots x_m}}}, \qquad (2.14)$$

или по рекуррентной формуле:

$$r_{yx_ix_1x_2\ldots x_{i-1}x_{i+1}\ldots x_m}=\frac{r_{yx_ix_1x_2\ldots x_{i-1}x_{i+1}\ldots x_{m-1}}-r_{yx_mx_1x_2\ldots x_{m-1}}\cdot r_{yx_ix_mx_1x_2\ldots x_{i-1}x_{i+1}\ldots x_{m-1}}}{\sqrt{\left(1-r^2_{yx_mx_1x_m\ldots x_{m-1}}\right)\left(1-r^2_{yx_ix_mx_1x_2\ldots x_{i-1}x_{i+1}\ldots x_{m-1}}\right)}} \qquad (2.15)$$

Рассчитанные по рекуррентной формуле частные коэффициенты корреляции изменяются в пределах от –1 до +1, а по формулам через множественные коэффициенты детерминации – от 0 до 1. Сравнение их друг с другом позволяет ранжировать факторы по тесноте их связи с результатом. Частные коэффициенты корреляции дают меру тесноты связи каждого фактора с результатом в чистом виде.

При двух факторах формулы (2.14) и (2.15) примут вид:

$$r_{yx_1x_2}=\sqrt{1-\frac{1-R^2_{yx_1x_2}}{1-r^2_{yx_2}}};r_{yx_2x_1}=\sqrt{1-\frac{1-R^2_{yx_1x_2}}{1-r^2_{yx_1}}}.$$

$$r_{yx_1x_2}=\frac{r_{yx_1}-r_{yx_2}\cdot r_{x_1x_2}}{\sqrt{\left(1-r^2_{x_2}\right)\left(1-r^2_{x_1x_2}\right)}};r_{yx_2x_1}=\frac{r_{yx_2}-r_{yx_1}\cdot r_{x_1x_2}}{\sqrt{\left(1-r^2_{x_1}\right)\left(1-r^2_{x_1x_2}\right)}}.$$

Значимость уравнения множественной регрессии в целом оценивается с помощью *F*-критерия Фишера:

$$F=\frac{R^2}{1-R^2}\cdot\frac{n-m-1}{m}. \qquad (2.16)$$

Частный F-критерий оценивает статистическую значимость присутствия каждого из факторов в уравнении. В общем виде для фактора *x* частный *F*-критерий определится как

$$F_{x_i}=\frac{R^2_{yx_1\ldots x_i\ldots x_m}-R^2_{yx_1\ldots x_{i-1}x_{i+1}\ldots x_m}}{1-R^2_{yx_1\ldots x_i\ldots x_m}}\cdot\frac{n-m-1}{m}. \qquad (2.17)$$

Фактическое значение *F*-критерия Фишера сравнивается с табличным значением $F_{табл}(\alpha,k_1,k_2)$ при уровне значимости α и степенях свободы $k_1=1$ и $k_2=n-m-1$. При этом, если фактическое значение F_{x_i} - критерия больше табличного, то дополнительное включение фактора x_i в модель статистически оправданно и коэффициент чистой регрессии b_i при факторе x_i статистически значим. Если же фактическое

значение F_{x_i} меньше табличного, то дополнительное включение в модель фактора x_i не увеличивает существенно долю объясненной вариации признака y , следовательно, нецелесообразно его

включение в модель; коэффициент регрессии при данном факторе в этом случае статистически незначим.

Оценка значимости коэффициентов чистой регрессии проводится по *t*-статистике Стьюдента. В этом случае, как и в парной регрессии, для каждого фактора используется формула:

$$t_{b_i} = \frac{b_i}{m_{b_i}} \tag{2.18}$$

Для уравнения множественной регрессии (3.1) средняя квадратическая ошибка коэффициента регрессии может быть определена по формуле:

$$m_{b_i} = \frac{\sigma_y \cdot \sqrt{1 - R^2_{yx_1 \ldots x_m}}}{\sigma_{x_i} \cdot \sqrt{1 - R^2_{yx_i x_1 \ldots x_m}}} \cdot \frac{1}{\sqrt{n - m - 1}}, \tag{2.19}$$

где $R^2_{yx_i x_1 \ldots x_m}$ – коэффициент детерминации для зависимости фактора x_i со всеми другими факторами уравнения множественной регрессии. Для двухфакторной модели (m = 2) имеем:

$$m_{b_1} = \frac{\sigma_y \cdot \sqrt{1 - R^2_{yx_1x_2}}}{\sigma_{x_1} \cdot \sqrt{1 - r^2_{x_1x_2}}} \cdot \frac{1}{\sqrt{n-3}};$$

$$m_{b_2} = \frac{\sigma_y \cdot \sqrt{1 - R^2_{yx_1x_2}}}{\sigma_{x_1} \cdot \sqrt{1 - r^2_{x_1x_2}}} \cdot \frac{1}{\sqrt{n-3}}; \tag{2.20), (2.21}$$

Существует связь между *t*-статистикой Стьюдента и частным *F*-критерием Фишера:

$$\left|t_{b_i}\right| = \sqrt{F_{x_i}} \tag{2.22}$$

Уравнения множественной регрессии могут включать в качестве независимых переменных качественные признаки (например, профессия, пол, образование, климатические условия, отдельные регионы и т.д.). Чтобы ввести такие переменные в регрессионную модель, их необходимо упорядочить и присвоить им те или иные значения, т.е. качественные переменные преобразовать в количественные.

Такого вида сконструированные переменные принято в эконометрике называть *фиктивными переменными.* Например, включать в

модель фактор «пол» в виде фиктивной переменной можно в следующем виде:

$$z = \begin{cases} 1 - мужскойпол \\ 0 - женскийпол \end{cases} \quad (2.23)$$

Коэффициент регрессии при фиктивной переменной интерпретируется как среднее изменение зависимой переменной при переходе от одной категории (женский пол) к другой (мужской пол) при неизменных значениях остальных параметров.

2.2 Решение типовой задачи в MS Excel с использованием надстройки Анализ данных

Пример. По 20 предприятиям региона изучается зависимость выработки продукции на одного работника *y* (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%). Исходные данные приведены в таблице 2.1

Таблица 2.1 – Исходные данные

Номер предприятия	Y	X_1	X_2	Номер предприятия	Y	X_1	X_2
1	7,0	3,9	10,0	11	9,0	6,0	21,0
2	7,0	3,9	14,0	12	11,0	6,4	22,0
3	7,0	3,7	15,0	13	9,0	6,8	22,0
4	7,0	4,0	16,0	14	11,0	7,2	25,0
5	7,0	3,8	17,0	15	12,0	8,0	28,0
6	7,0	4,8	19,0	16	12,0	8,2	29,0
7	8,0	5,4	19,0	17	12,0	8,1	30,0
8	8,0	4,4	20,0	18	12,	8,5	31,0
9	8,0	5,3	20,0	19	14,0	9,6	32,0
10	10,0	6,8	20,0	20	14,0	9,0	36,0

Требуется:

1. Построить линейную модель множественной регрессии. На основе коэффициентов регрессии и средних коэффициентов эластичности ранжировать факторы по степени их влияния на результат.

2. Найти коэффициенты парной и множественной корреляции.

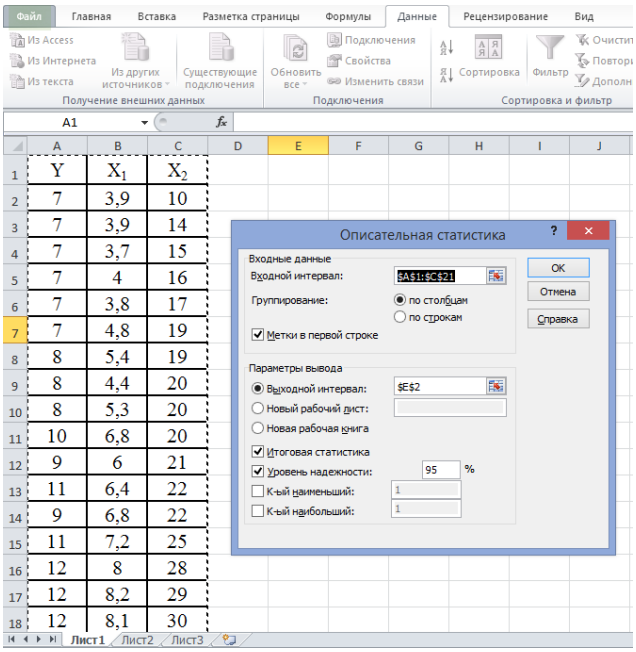
3. Найти скорректированный коэффициент множественной детерминации. Сравнить его с нескорректированным (общим) коэффициентом детерминации.

4. С помощью F -критерия Фишера оценить статистическую надежность уравнения регрессии и коэффициента детерминации $R^2_{yx_1x_2}$

5. С помощью t -критерия оценить статистическую значимость коэффициентов чистой регрессии.

Решение. Ввести исходные данные в таблицу в MS Excel.

На вкладке Данные выбрать Анализ данных, в открывшемся диалоговом окне выбрать инструмент анализа: Описательная статистика (рис. 2.1).



	A	B	C
1	Y	X_1	X_2
2	7	3,9	10
3	7	3,9	14
4	7	3,7	15
5	7	4	16
6	7	3,8	17
7	7	4,8	19
8	8	5,4	19
9	8	4,4	20
10	8	5,3	20
11	10	6,8	20
12	9	6	21
13	11	6,4	22
14	9	6,8	22
15	11	7,2	25
16	12	8	28
17	12	8,2	29
18	12	8,1	30

Рисунок 2.1 – Вид окна Описательная статистика

Заполнить диалоговое окно Описательная статистика.

Входной интервал – исходные данные;

Группировка – по строкам или по столбцам – указывается с учетом формирования таблицы с исходными данными (в примере таблица сформирована с расположением исходных данных по столбцам);

Метки в первой строке – отмечается в случае, если в диапазон Входной интервал входит название столбцов;

Выходной интервал – достаточно указать верхнюю левую ячейку будущего диапазона, либо *Новый рабочий лист*;

Если необходимо указать дополнительные параметры: Итоговая статистика, Уровень надежности и т.д.

Нажать ОК.

Результаты вычислений представлены на рисунке 2.2.

=F8/F4

	E	F	G	H	I	J
	Среднее	9,6	Среднее	6,19	Среднее	22,3
	Стандартная ошибка	0,549641031	Стандартная ошибка	0,433522901	Стандартная ошибка	1,523672848
	Медиана	9	Медиана	6,2	Медиана	20,5
	Мода	7	Мода	3,9	Мода	20
	Стандартное отклонение	2,458069418	Стандартное отклонение	1,938773351	Стандартное отклонение	6,814072127
	Дисперсия выборки	6,042105263	Дисперсия выборки	3,758842105	Дисперсия выборки	46,43157895
	Эксцесс	-1,196054269	Эксцесс	-1,331425706	Эксцесс	-0,53652906
	Асимметричность	0,445095914	Асимметричность	0,188100846	Асимметричность	0,327800798
	Интервал	7	Интервал	5,9	Интервал	26
	Минимум	7	Минимум	3,7	Минимум	10
	Максимум	14	Максимум	9,6	Максимум	36
	Сумма	192	Сумма	123,8	Сумма	446
	Счет	20	Счет	20	Счет	20
	Уровень надежности(95,0%)	1,1504119	Уровень надежности(95,0%)	0,907373859	Уровень надежности(95,0%)	3,189083922
	Коэффициент вариации	0,256048898		0,313210557		0,305563772

Рисунок 2.2 – Результаты вычислений Итоговая статистика

Сравнивая значения средних квадратических отклонений (в таблице – Стандартная ошибка) и средних величин (в таблице – Среднее) определить коэффициент вариации.

Значение коэффициента вариации до 0,35 (или 35%) свидетельствует о том, что совокупность однородна, и для ее изучения могут использоваться метод наименьших квадратов и вероятностные методы оценки статистических гипотез.

Значения линейных коэффициентов парной корреляции определяет тесноту попарно связанных переменных, использованных в уравнении множественной регрессии.

На вкладке Данные выбрать Анализ данных, в открывшемся диалоговом окне выбрать инструмент анализа: Корреляция (рис. 2.3).

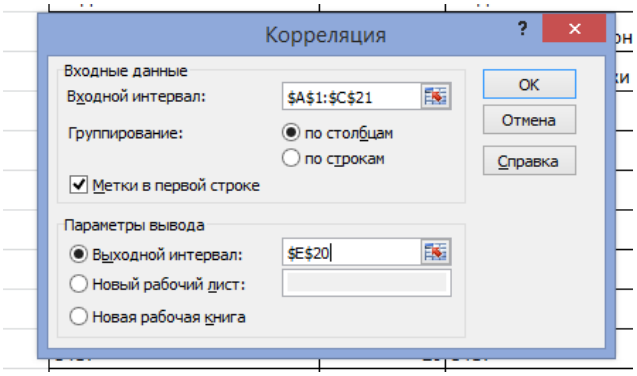


Рисунок 2.3 – Вид окна Корреляция

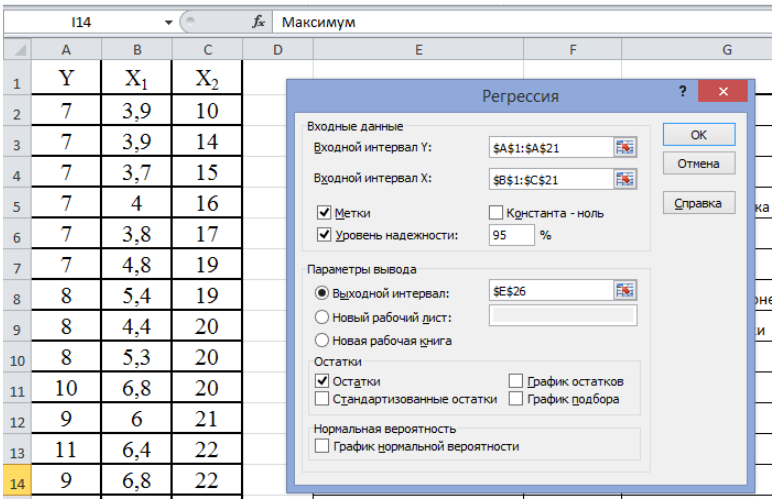
Заполнить диалоговое окно Корреляция. Нажать ОК, сформируется матрица коэффициентов парной корреляции (рис. 2.4).

	Y	*X1*	*X2*
Y	1		
X1	0,969881436	1	
X2	0,940800036	0,942838898	1

Рисунок 2.4 – Матрица коэффициентов парной корреляции

Значения коэффициентов парной корреляции указывают на весьма тесную связь результативного признака у как с фактором x_1, так и с фактором x_2 (r_{yx1}= 0,9699 и r_{yx2}= 0,9408). Но вместе с тем межфакторная связь r_{x1x2}= 0,9428 весьма тесная и превышает тесноту связи x_2 с у. Поэтому для улучшения модели можно исключить из нее фактор x_2, чтобы избежать искажения результатов в связи с высокой межфакторной зависимостью.

Для вычисления параметров линейного уравнения множественной регрессии на вкладке Данные выбрать Анализ данных, в открывшемся диалоговом окне выбрать инструмент анализа: Регрессия (рис. 2.5).



	A	B	C
1	Y	X_1	X_2
2	7	3,9	10
3	7	3,9	14
4	7	3,7	15
5	7	4	16
6	7	3,8	17
7	7	4,8	19
8	8	5,4	19
9	8	4,4	20
10	8	5,3	20
11	10	6,8	20
12	9	6	21
13	11	6,4	22
14	9	6,8	22

Рисунок 2.5 – Вид окна регрессия

Заполнить диалоговое окно Регрессия. Обратите внимание, что входной интервал разделен на две части:

Входной интервал Y – диапазон данных результативного признака;

Входной интервал X – диапазон данных всех факторов независимых признаков.

Нажать ОК, сформируются таблицы с параметрами линейного уравнения множественной регрессии (рис. 2.6).

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,973101182
R-квадрат	0,94692591
Нормированный R-квадрат	0,9406819
Стандартная ошибка	0,598670364
Наблюдения	20

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	108,7070945	54,35354726	151,6534774	1,45045E-11
Остаток	17	6,092905478	0,358406205		
Итого	19	114,8			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	1,83530694	0,471064997	3,896080054	0,001161531	0,841446671	2,829167209	0,841446671	2,829167209
X1	0,945947723	0,212576487	4,449917001	0,00035148	0,497450539	1,394444906	0,497450539	1,394444906
X2	0,085617787	0,060483309	1,415560577	0,174963664	-0,04199084	0,213226414	-0,04199084	0,213226414

Рисунок 2.6 – Результаты применения инструмента Регрессия

Выписываем, округляя до 4 знаков после запятой:

уравнение регрессии:

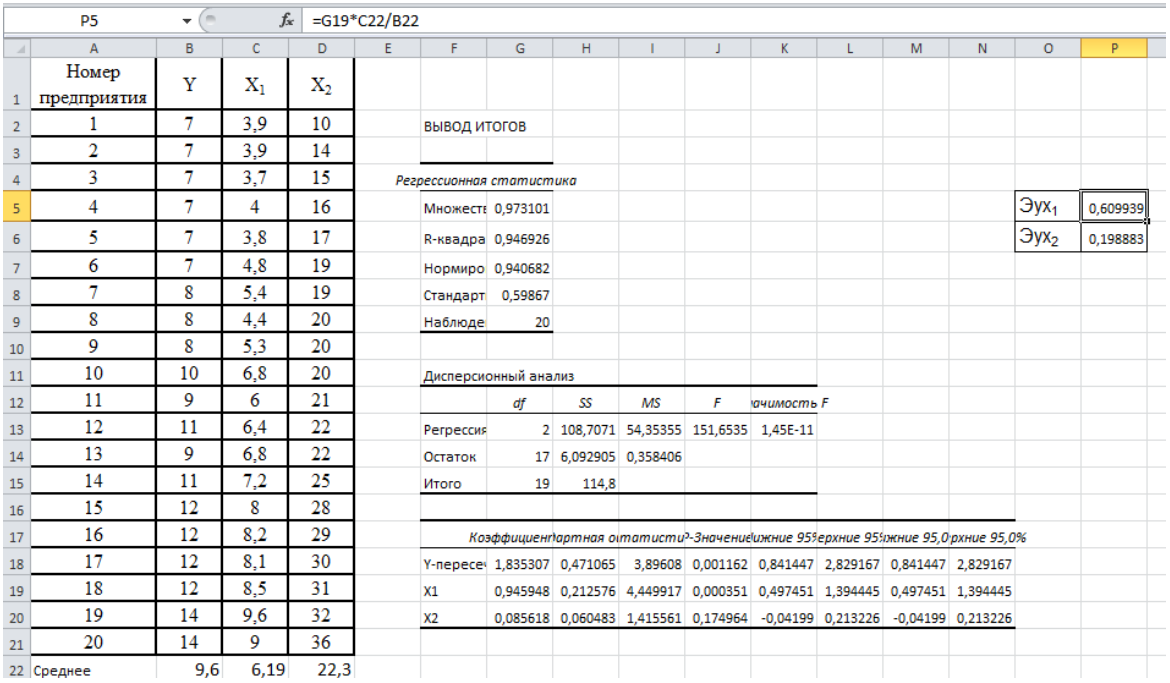
$\hat{y}_x = 1{,}8353 + 0{,}9459x_1 + 0{,}0856x_2$.

Коэффициенты регрессии свидетельствуют о том, что при увеличении ввода в действие новых основных фондов (X_1) на 1% от стоимости фондов на конец года выработка продукции на одного работника (Y) увеличится на 0,9459 тыс. руб., при постоянных значениях остальных условий. А при увеличении удельного веса рабочих высокой квалификации в общей численности рабочих (X_2) на 1% выработка продукции на одного работника (Y) увеличится на 0,0856 тыс. руб., при постоянных значениях остальных условий.

Средний коэффициент эластичности показывает, на сколько процентов изменится результат от значения своей средней $\bar{y}$ при изменении значения фактора на 1 % от своей средней $\bar{x_j}$ и при фиксированном воздействии на результат всех прочих факторов. Средний коэффициент эластичности ($\bar{Э}$) для линейной регрессии рассчитывается как относительное изменение у на единицу относительного изменения х:

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

Воспользуемся функцией СРЕДНЕЕ для расчета средних значений результативного признака и факторов, что позволит определить значения средних коэффициентов эластичности (рис. 2.7).



P5 =G19*C22/B22

Номер предприятия	Y	X_1	X_2
1	7	3,9	10
2	7	3,9	14
3	7	3,7	15
4	7	4	16
5	7	3,8	17
6	7	4,8	19
7	8	5,4	19
8	8	4,4	20
9	8	5,3	20
10	10	6,8	20
11	9	6	21
12	11	6,4	22
13	9	6,8	22
14	11	7,2	25
15	12	8	28
16	12	8,2	29
17	12	8,1	30
18	12	8,5	31
19	14	9,6	32
20	14	9	36
Среднее	9,6	6,19	22,3

ВЫВОД ИТОГОВ

Регрессионная статистика

Множеств	0,973101
R-квадра	0,946926
Нормиро	0,940682
Стандарт	0,59867
Наблюде	20

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*ачимость F*
Регрессия	2	108,7071	54,35355	151,6535	1,45E-11
Остаток	17	6,092905	0,358406		
Итого	19	114,8			

Коэффициенгартная ошатисти²-Значение ижние 95% рхние 95% жние 95,0% рхние 95,0%								
Y-пересе	1,835307	0,471065	3,89608	0,001162	0,841447	2,829167	0,841447	2,829167
X1	0,945948	0,212576	4,449917	0,000351	0,497451	1,394445	0,497451	1,394445
X2	0,085618	0,060483	1,415561	0,174964	-0,04199	0,213226	-0,04199	0,213226

$Эух_1$	0,609939
$Эух_2$	0,198883

Рисунок 2.7 – Расчет частных коэффициентов эластичности

$$\overline{Э}_{yx_1} = 0{,}6099,$$

$$\overline{Э}_{yx_2} = 0{,}1989.$$

Значения коэффициентов эластичности свидетельствуют о более сильном влиянии на результат фактора x_1.

Множественный коэффициент корреляции (Множественный R) имеет значение 0,9731 и характеризует тесноту линейной корреляционной связи между одной случайной величиной и некоторым множеством случайных величин.

Нескорректированный коэффициент детерминации (R-квадрат):

$R^2_{yx1x2} = 0{,}9469$ оценивает долю вариации результата за счет представленных в уравнении факторов в общей вариации результата. В примере эта доля составляет 94,7%.

Скорректированный коэффициент детерминации (Нормированный R-квадрат):

$R^2_{yx1x2} = 0{,}9407$ определяет тесноту связи с учетом степеней свободы общей и остаточной дисперсий. Он дает такую оценку тесноты связи, которая не зависит от количества факторов в модели и следовательно может сравниваться по разным моделям с разным количеством факторов.

Оба показателя указывают на весьма высокую детерминированность результата y в модели факторами x_1 и x_2.

Фактическое значение F -критерия Фишера:

$F = 151{,}6535$.

Вероятность случайного получения такого значения равна 0,000 (Значимость F – 0,000), что не превышает допустимый уровень 5%. Следовательно, полученное значение не случайно, т.е. подтверждается статистическая значимость всего уравнения и показателя тесноты связи (Множественный R).

Остаточная дисперсия на одну степень свободы:

$S^2_{\text{ост}} = 0{,}3584.$

Корень квадратный из остаточной дисперсии (стандартная ошибка):

$S_{\text{ост}} = 0{,}5987.$

Стандартные ошибки для параметров регрессии:

m_a = 0,4711, m_{b1} = 0,2126, m_{b2} = 0,0605.

Они показывают, какое значение данной характеристики сформировалось под влиянием случайных факторов. Эти значения используются для расчета t-статистики Стьюдента.

Фактические значения *t*-статистики Стьюдента:

t_a = 3,90, t_{b1} = 4,45, t_{b2} = 1,42.

Если фактическое значение t-статистики Стьюдента больше табличного при (n-2, α=0,05)=2,1009 (прил. 2), то можно сделать вывод существенности данного показателя. В данном случае статистически значимыми являются b_0 и b_1, а величина b_2 сформировалась под воздействием случайных причин, поэтому фактор x_2 можно исключить как несущественно влияющий, неинформативный.

Доверительные интервалы:

$0{,}8414 \leq b_0^* \leq 2{,}8291$

$0{,}4974 \leq b_1^* \leq 1{,}3944$

$-0{,}0420 \leq b_2^* \leq 0{,}2132$

Одновременно с построением доверительных интервалов обычно производится проверка статистической значимости параметров уравнения регрессии. При этом выдвигаются нулевые гипотезы о равенстве нулю действительных значений параметров. Имеет место следующее правило: если рассчитанные границы доверительного интервала имеют разные знаки, т.е. интервал включает в себя нуль, то соответствующий параметр уравнения регрессии незначим. Напротив, если доверительный интервал не содержит нуля, то соответствующий параметр статистически значим.

Исходя из этого, можно сделать вывод о том, что статистически значимыми являются b_0 и b_1, а величина b_2 имеет случайную природу.

Удаляя из модели переменную x_2, переходим к исследованию однофакторной модели с одной переменной x_1. Подробно подобное исследование представлено в разделе 1.2.

2.3 Использование функции ЛИНЕЙН в MS Excel

Подробное описание и синтаксис функции ЛИНЕЙН рассмотрены в разделе 1.3. Остановимся на особенностях применения функции для множественной регрессии.

Уравнение множественной регрессии имеет следующий вид:

$$y = a + \sum_{j=1}^{n} b_j x_j$$

где n – количество факторов включенных в модель.

Если необходимо отобразить не только наклон и отрезок, но и дополнительные статистики, выделяется диапазон на один столбец больше, чем столбцов с переменными x, и высотой 5 строк. На рисунке 2.7 выделен диапазон E8:G12 (3 столбца по 5 строк), поскольку в примере две переменные x_1 и x_2. Необходимо дополнительно ввести значения двух констант: «Конст» и «Статистика». С помощью первой константы отмечается, присутствует (1) или нет (0) в спецификации модели свободный параметр a. Второй константе «Статистика» присваивается значение 1. Это означает, что нужно получить полную информацию, т.е. вывести дополнительные статистики. После ввода формулы нажатием Ctrl+Shift+Enter, результат должен соответствовать рисунку 2.8.

	A	B	C	D	E		F	G	H	I
1	X_1	X_2	Y							
2	3,9	10	7							
3	3,9	14	7							
4	3,7	15	7							
5	4	16	7							
6	3,8	17	7							
7	4,8	19	7		**Столбец F /**	**Столбец G**				
8	5,4	19	8		Наклон X2 (b_2) /	Наклон X1 (b_1)	0,0856178	0,94594772	1,8353069	Отрезок (a)
9	4,4	20	8		Стандартная ошибка для X2 /	для X1	0,0604833	0,21257649	0,471065	Стандартная ошибка для a
10	5,3	20	8		Коэффициент детерминации R2 /	Стандартная ошибка для y	0,9469259	0,59867036	#Н/Д	
11	6,8	20	10		F статистика /	Число степеней свободы	151,65348	17	#Н/Д	
12	6	21	9		Регрессионная сумма квадратов /	Остаточная сумма квадратов	108,70709	6,09290548	#Н/Д	
13	6,4	22	11							
14	6,8	22	9							
15	7,2	25	11							
16	8	28	12							
17	8,2	29	12							
18	8,1	30	12							
19	8,5	31	12							

Рисунок 2.8 – Результат использования функции ЛИНЕЙН

Необходимо учесть, что, не смотря на то, что в таблице с исходными данными факторы располагаются по порядку, в первом столбце переменная x_1, во втором x_2. Результаты расчета в первых двух строках выходного интервала E8:G12 представлены следующим образом – в столбце E для переменной x_2, в столбце F для переменной x_1.

Полученные данные позволяют:

составить уравнение регрессии:

$\hat{y}_x = 1{,}8353 + 0{,}9459x_1 + 0{,}0856x_2$,

рассчитать множественный коэффициент корреляции. Для вычисления коэффициента корреляции достаточно извлечь квадратный корень из значения коэффициента детерминации:

$\sqrt{0{,}9469} = 0{,}9731$,

оценить нескорректированный коэффициент детерминации:

$R^2_{yx1x2} = 0{,}9469$,

оценить фактическое значение F -критерия Фишера:

F = 151,653,

стандартные ошибки для параметров регрессии:

$m_a = 0{,}4711$; $m_{b1} = 0{,}2126$; $m_{b2} = 0{,}0605$.

Эти значения используются для расчета t-статистики Стьюдента.

$t_a = \frac{a}{m_a} = \frac{1{,}8353}{0{,}4711} = 3{,}8961$,

$t_{b1} = \frac{b_1}{m_{b1}} = \frac{0{,}9459}{0{,}2126} = 4{,}4499$,

$t_{b2} = \frac{b_2}{m_{b2}} = \frac{0{,}0856}{0{,}0605} = 1{,}4156$.

Корень квадратный из остаточной дисперсии (стандартная ошибка):

$S_{ост} = 0{,}5987$.

2 ГЕТЕРОСКЕДАСТИЧНОСТЬ

3.1 Теоретическая справка

Гетероскедастичность – понятие, используемое в прикладной статистике (чаще всего – в эконометрике), означающее неоднородность наблюдений, выражающуюся в неодинаковой (непостоянной) дисперсии случайной ошибки регрессионной модели. Гетероскедастичность противоположна гомоскедастичности, означающей однородность наблюдений, то есть постоянство дисперсии случайных ошибок модели.

Наличие гетероскедастичности случайных ошибок приводит к неэффективности оценок, полученных с помощью метода наименьших квадратов. Кроме того, в этом случае оказывается смещённой и несостоятельной классическая оценка ковариационной матрицы МНК-оценок параметров. Следовательно, статистические выводы о качестве полученных оценок могут быть неадекватными. В связи с этим тестирование моделей на гетероскедастичность является одной из необходимых процедур при построении регрессионных моделей.

Разработан ряд методов и тестов для обнаружения гетероскедастичности. К ним относятся: тест ранговой корреляции Спирмена, тест Гольдфельда-Квандта, тест Уайта.

Для обнаружения гетероскедастичности для рассматриваемого в этом разделе примера будем использовать тест Гольдфельта-Квандта.

Тест Гольдфельда-Квандта применяется, если случайные остатки предполагаются нормально распределенными величинами и объем наблюдений достаточно большой.

Процедура проверки следующая:

1. Все наблюдения упорядочивают по мере возрастания какой-либо независимой переменной, которая, как предполагается, оказывает влияние на изменение дисперсии случайных остатков;

2. Упорядоченную совокупность делят на три группы, причем первая и последняя должны быть равного объема, с числом наблюдений, больших, чем число параметров модели регрессии. Пусть в первую и третью группы отобрано по k наблюдений, тогда во второй группе (n – 2k) наблюдений;

3. По первой и третьей группам находят параметры уравнений регрессии той же структуры, что и исходное уравнение регрессии, и остаточные суммы квадратов по каждой модели;

4. Используя данные об остаточных суммах квадратов моделей первой и третьей групп, рассчитывают фактическое значение F-критерия Фишера по формуле:

$$F = \frac{S_3/(m-p-1)}{S_1/(m-p-1)} = \frac{S_3}{S_1}.$$

Если расчетное значение $F > F_{табл}$, где

$F_{табл} = (\alpha = 0{,}05;\ k_1 = k_2 = m - p - 1)$, то признается наличие гетероскедастичности, в обратном случае – наличие гомоскедастичности.

3.2 Тест Гольдфельда-Квандта в MS Excel

Все 20 наблюдений представленных в таблице 2.1 необходимо упорядочить по возрастанию переменной X_1. Для этого нужно выбрать пункты *Данные* → *Сортировка*.

В диалоговом окне «Сортировка» отметить «Сортировать по»: Столбец – X1; Сортировка – Значения; Порядок – По возрастанию (рис.3.1). Нажать ОК.

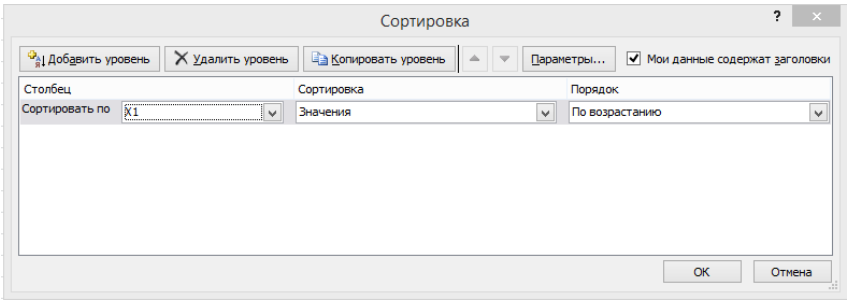


Рисунок 3.1 – Вид окна Сортировка

Вся отсортированная выборка делится на три приблизительно равные подвыборки с размерностью: $m_1 = 7,\ m_2 = 6,\ m_3 = 7$

	A	B	C	D
1	Y	X_1	X_2	
2	7	3,7	15	
3	7	3,8	17	
4	7	3,9	10	
5	7	3,9	14	
6	7	4	16	
7	8	4,4	20	
8	7	4,8	19	
9	8	5,3	20	
10	8	5,4	19	
11	9	6	21	
12	11	6,4	22	
13	10	6,8	20	
14	9	6,8	22	
15	11	7,2	25	
16	12	8	28	
17	12	8,1	30	
18	12	8,2	29	
19	12	8,5	31	
20	14	9	36	
21	14	9,6	32	

Рисунок 3.2 – Деление на подвыборки по X_1

Строится регрессия для первых 7 наблюдений (рис. 3.3) и для последних 7 наблюдений (рис.3.4).

ВЫВОД ИТОГОВ						
Регрессионная статистика						
Множественный R	0,548090142					
R-квадрат	0,300402804					
Нормированный R-квадрат	-0,049395794					
Стандартная ошибка	0,387186886					
Наблюдения	7					
Дисперсионный анализ						
	df	*SS*	*MS*	*F*	*Значимость F*	
Регрессия	2	0,257488118	0,128744059	0,858787901	0,489436237	
Остаток	4	0,599654739	0,149913685			
Итого	6	0,857142857				
	Коэффициенты	*ндартная ошиб*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	6,051318267	1,699398676	3,560858529	0,023567893	1,333031133	10,7696054
X1	0,037664783	0,52465726	0,071789311	0,946215748	-1,419017298	1,494346865
X2	0,059165097	0,061353981	0,964323689	0,389492172	-0,111180863	0,229511057

Рисунок 3.3 – Результат регрессии для первой подвыборки

ВЫВОД ИТОГОВ					
Регрессионная статистика					
Множественный R	0,950026562				
R-квадрат	0,902550469				
Нормированный R-квадрат	0,837584116				
Стандартная ошибка	0,416225432				
Наблюдения	6				
Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	4,813602503	2,406801252	13,89258312	0,03042073
Остаток	3	0,51973083	0,17324361		
Итого	5	5,333333333			
	Коэффициенты	*ндартная ошиб*	*t-статистика*	*P-Значение*	*Нижние 95%*
Y-пересечение	-0,856193404	2,595510546	-0,329874754	0,763186184	-9,116266352
7,2	1,099941846	0,415585231	2,646729874	0,077209582	-0,222635837
25	0,132258868	0,091365762	1,447575833	0,24355137	-0,158507762

Рисунок 3.4 – Результат регрессии для последней подвыборки

Определяется F – статистика:

$$F = \frac{S_3/(m-p-1)}{S_1/(m-p-1)} = \frac{S_3}{S_1},$$

$$F = \frac{0{,}51973083}{0{,}599654739} = 0{,}8667,$$

$$F_{\text{табл}} = (\alpha = 0{,}05;\ k_1 = k_2 = m - p - 1 = 4) = 6{,}39,$$

$F < F_{\text{табл}}$ – гетероскедастичность отсутствует.

Следует отметить, что переменная X_1 гомоскедастична, но это не значит, что по всем остальным переменным модель может не быть гетероскедастичной. Поэтому необходима дальнейшая проверка по остальным переменным.

Теперь все 20 наблюдений необходимо сортировать по возрастанию переменной X_2. После этого применить тест по вновь образованным подвыборкам.

Результат регрессии для первой и последней подвыборок представлены на рисунках 3.5 и 3.6.

ВЫВОД ИТОГОВ					
Регрессионная статистика					
Множественный R	0,846238				
R-квадрат	0,716119				
Нормированный R-квадрат	0,526865				
Стандартная ошибка	0,280813				
Наблюдения	6				
Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	0,596766007	0,298383003	3,783908046	0,151253005
Остаток	3	0,236567327	0,078855776		
Итого	5	0,833333333			
	Коэффициен	*тандартная ошибк*	*-статистика*	*P-Значение*	*Нижние 95%*
Y-пересечение	5,544159	1,024296447	5,412651044	0,012367778	2,28439079
3,9	0,68958	0,331336659	2,081206961	0,128853906	-0,364880965
10	-0,07918	0,109108664	-0,725717579	0,520502174	-0,426414539

Рисунок 3.5 – Результат регрессии для первой подвыборки

ВЫВОД ИТОГОВ					
Регрессионная статистика					
Множественный R	0,950027				
R-квадрат	0,90255				
Нормированный R-квадрат	0,837584				
Стандартная ошибка	0,416225				
Наблюдения	6				
Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	4,813602503	2,406801252	13,89258312	0,03042073
Остаток	3	0,51973083	0,17324361		
Итого	5	5,333333333			
	Коэффициен	*тандартная ошибк*	*-статистика*	*P-Значение*	*Нижние 95%*
Y-пересечение	-0,85619	2,595510546	-0,329874754	0,763186184	-9,116266352
7,2	1,099942	0,415585231	2,646729874	0,077209582	-0,222635837

Рисунок 3.6 – Результат регрессии для последней подвыборки

Определить F – статистику:

$$F = \frac{S_3/(m-p-1)}{S_1/(m-p-1)} = \frac{S_3}{S_1},$$

$$F = \frac{0{,}51973083}{0{,}236567327} = 2{,}1969,$$

$$F_{\text{табл}} = (\alpha = 0{,}05;\ k_1 = k_2 = m - p - 1 = 4) = 6{,}39,$$

$F < F_{\text{табл}}$ – гетероскедастичность отсутствует.

Таким образом, в результате проверки выяснилось, что по всем переменным модель является гомоскедастичной.

Задачи для самостоятельного решения.

Задача 1.

По данным, представленным в табл. 2.2, исследуется зависимость между величиной накладных расходов 36 строительных организаций Y (млн. руб.) и следующими тремя факторами:

x_1 – объемом выполненных работ, млн. руб.

x_2 – численностью рабочих, чел.

x_3 – фондом зарплаты, млн. руб.

Таблица 2.2 – Исходные данные

№	Y	X_1	X_2	X_3
10	3,40	15,00	670,00	5,76
11	4,10	14,70	622,00	6,10
12	4,10	13,30	566,00	6,06
13	3,10	14,60	518,00	4,92
14	2,80	11,70	510,00	4,13
15	2,10	10,60	452,00	4,38
16	2,50	10,00	447,00	4,16
17	2,00	9,00	497,00	4,32
18	2,40	9,50	428,00	4,02
19	2,30	7,00	381,00	3,32
20	2,40	9,10	385,00	3,62
21	2,50	6,80	412,00	3,46
22	2,20	5,50	293,00	2,14
23	1,60	5,10	284,00	2,24
24	3,40	12,20	514,00	3,96
25	2,70	11,00	407,00	3,34
26	3,20	9,30	577,00	3,68
27	2,90	5,90	265,00	2,12
28	4,80	25,90	977,00	10,65
29	3,70	23,50	724,00	6,81
30	4,40	19,80	983,00	9,24
31	3,70	18,80	828,00	8,86
32	4,80	19,10	766,00	7,35
33	3,70	18,80	615,00	5,29
34	3,60	17,40	583,00	5,83
35	4,00	14,10	591,00	6,27
36	3,80	13,80	593,00	5,40

Требуется. Построить уравнение множественной линейной регрессии с использованием пакета Анализ данных с полным набором факторов. Отобрать информативные факторы в модель.

Построить уравнение множественной регрессии только со значимыми факторами. Оценить качество полученного уравнения регрессии. Оценить статистическую значимость уравнения регрессии, используя F-критерий Фишера (а =0,05) и статистическую значимость параметров регрессии, используя t-статистику Стьюдента.

Задача 2.

По данным, представленным в табл. 2.3, исследуется зависимость между ценой автомобиля по 15 объявлениям о продаже Y (усл. ден. ед.) и следующими тремя факторами:

x_1 – пробег, тыс. км.

x_2 – срок эксплуатации, лет.

x_3 – объем двигателя, л.

Таблица 2.3 – Исходные данные

№ автомобиля	Y	X_1	X_2	X_3
1	12500	130	12	2,3
2	13700	120	10	1,9
3	9200	300	15	1,8
4	11400	180	13	2,1
5	15800	150	14	2,6
6	12300	80	8	1,7
7	16300	170	10	2,4
8	10200	210	11	1,9
9	11000	250	7	1,9
10	12700	150	9	1,7
11	15000	90	4	2,2
12	10500	230	13	2,4
13	17200	120	8	2,3
14	16000	110	9	2,5
15	17100	120	6	2,6

Требуется. Построить уравнение множественной линейной регрессии с использованием пакета Анализ данных с полным набором факторов. Отобрать информативные факторы в модель.

Построить уравнение множественной регрессии только со значимыми факторами. Оценить качество полученного уравнения регрессии.

Оценить статистическую значимость уравнения регрессии, используя F-критерий Фишера (а =0,05) и статистическую значимость параметров регрессии, используя t-статистику Стьюдента.

Задача 3.

По данным, представленным в табл. 2.4, исследуется зависимость между прибылью по 13 банкам Y (млн. руб.) и следующими четырьмя факторами:

x_1 – собственный капитал, млн. руб.

x_2 – привлеченные средства, млн. руб.

x_3 – депозитная ставка, % годовых

X_4 – кредитная ставка, % годовых

Таблица 2.4 – Исходные данные

№ банка	Y	X_1	X_2	X_3	X_4
1	115	4428	3278	12,5	17,7
2	80	3756	5696	11,7	18,2
3	97	2970	2210	11,2	19,1
4	92	6231	5823	9,7	15,2
5	129	3960	4569	13,5	18,5
6	223	7354	2896	10,8	18,6
7	251	4662	3526	12,1	15,7
8	267	4760	2259	11,7	16,6
9	137	4569	4596	13,7	17,3
10	163	5274	3271	12,5	19,3
11	225	5418	4596	12,8	17,8
12	278	5359	3256	11,2	14,5
13	367	8254	5189	10,4	13,7

Требуется. Построить уравнение множественной линейной регрессии с использованием пакета Анализ данных с полным набором факторов. Отобрать информативные факторы в модель.

Построить уравнение множественной регрессии только со значимыми факторами. Оценить качество полученного уравнения регрессии. Оценить статистическую значимость уравнения регрессии, используя F-критерий Фишера (а =0,05) и статистическую значимость параметров регрессии, используя t-статистику Стьюдента.

Задача 4.

По данным, представленным в табл. 2.5, исследуется зависимость между объемом реализованной продукции кондитерского предприятия за 12 месяцев Y (тыс. руб.) и следующими четырьмя факторами:

x_1 – затраты на телерекламу, тыс. руб.

x_2 – затраты на радиорекламу, тыс. руб.

x_3 – затраты на газетную рекламу, тыс. руб.

X_4 – затраты на наружную рекламу, тыс. руб.

Таблица 2.5 – Исходные данные

Месяц	Y	X1	X2	X3	X4
1	14 050	240	42	42	34
2	16 310	263	47	44	36
3	15 632	241	55	45	35
4	15 126	276	47	42	32
5	13 972	236	49	47	25
6	15 753	272	44	45	39
7	16 661	276	57	55	45
8	15 584	260	46	47	36
9	15 326	280	40	35	34
10	14 077	248	38	38	29
11	15 528	289	49	45	25
12	1 755	258	56	52	26

Требуется. Построить уравнение множественной линейной регрессии с использованием пакета Анализ данных с полным набором факторов. Отобрать информативные факторы в модель.

Построить уравнение множественной регрессии только со значимыми факторами. Оценить качество полученного уравнения регрессии. Оценить статистическую значимость уравнения регрессии, используя F-критерий Фишера (а =0,05) и статистическую значимость параметров регрессии, используя t-статистику Стьюдента.

Задача 5.

По данным, представленным в табл. 2.2, исследуется зависимость между величиной накладных расходов 36 строительных организаций Y (млн. руб.) и следующими тремя факторами:

x_1 – объемом выполненных работ, млн. руб.

x_2 – численностью рабочих, чел.

x_3 – фондом зарплаты, млн. руб.

Требуется. Построить уравнение множественной линейной регрессии с использованием функции ЛИНЕЙН с дополнительной статистикой. Оценить качество полученного уравнения регрессии. Оценить статистическую значимость уравнения регрессии, используя F-критерий Фишера (а =0,05) и статистическую значимость параметров регрессии, используя t-статистику Стьюдента.

Провести исследование на обнаружение гетероскедастичности с использованием теста Гольдфельта-Квандта.

Задача 6.

По данным, представленным в табл. 2.3, исследуется зависимость между ценой автомобиля по 15 объявлениям о продаже Y (усл. ден. ед.) и следующими тремя факторами:

x_1 – пробег, тыс. км.

x_2 – срок эксплуатации, лет.

x_3 – объем двигателя, л.

Требуется. Построить уравнение множественной линейной регрессии с использованием функции ЛИНЕЙН с дополнительной статистикой. Оценить качество полученного уравнения регрессии. Оценить статистическую значимость уравнения регрессии, используя F-критерий Фишера (а =0,05) и статистическую значимость параметров регрессии, используя t-статистику Стьюдента.

Провести исследование на обнаружение гетероскедастичности с использованием теста Гольдфельта-Квандта.

Задача 7.

По данным, представленным в табл. 2.4, исследуется зависимость между прибылью по 13 банкам Y (млн. руб.) и следующими четырьмя факторами:

x_1 – собственный капитал, млн. руб.

x_2 – привлеченные средства, млн. руб.

x_3 – депозитная ставка, % годовых

X_4 – кредитная ставка, % годовых

Требуется. Построить уравнение множественной линейной регрессии с использованием функции ЛИНЕЙН с дополнительной статистикой. Оценить качество полученного уравнения регрессии. Оценить статистическую значимость уравнения регрессии, используя F-критерий Фишера (а =0,05) и статистическую значимость параметров регрессии, используя t-статистику Стьюдента.

Провести исследование на обнаружение гетероскедастичности с использованием теста Гольдфельта-Квандта.

Задача 8.

По данным, представленным в табл. 2.5, исследуется зависимость между объемом реализованной продукции кондитерского предприятия за 12 месяцев Y (тыс. руб.) и следующими четырьмя факторами:

x_1 – затраты на телерекламу, тыс. руб.

x_2 – затраты на радиорекламу, тыс. руб.

x_3 – затраты на газетную рекламу, тыс. руб.

X_4 – затраты на наружную рекламу, тыс. руб.

Требуется. Построить уравнение множественной линейной регрессии с использованием функции ЛИНЕЙН с дополнительной статистикой. Оценить качество полученного уравнения регрессии. Оценить статистическую значимость уравнения регрессии, используя F-критерий Фишера (а =0,05) и статистическую значимость параметров регрессии, используя t-статистику Стьюдента.

Провести исследование на обнаружение гетероскедастичности с использованием теста Гольдфельта-Квандта.

Список литературы

1. Вуколов Э.А. Основы статистического анализа. Практикум по статистическим методам и исследованию операций с использованием пакетов STATISTICA и EXCEL. – М.: Форум, 2008. – 464 с.
2. Демидова, О.А. Эконометрика: учебник и практикум для вузов / О.А. Демидова, Д.И. Малахов. – Москва: Издательство Юрайт, 2020. – 334с.
3. Евсеев, Е.А. Эконометрика: учебное пособие для вузов / Е.А. Евсеев, В.М. Буре. – 2-е изд., испр. и доп. – Москва: Издательство Юрайт, 2020. – 186с.
4. Елисеева И.И. Практикум по эконометрике: Учебное пособие. – М.: Финансы и статистика, 2005 – 192 с.
5. Кремер, Н.Ш. Эконометрика: учебник и практикум для вузов / Н.Ш. Кремер, Б.А. Путко; под редакцией Н.Ш. Кремера. – 4-е изд., испр. и доп. – Москва: Издательство Юрайт, 2020. – 308с.
6. Официальная статистика: Территориальный орган Федеральной службы государственной статистики по Пензенской области [Электронный ресурс] – Режим доступа: https://pnz.gks.ru/ofstatistics (Дата обращения: 03.05.2020)
7. Просветов Г.И. Эконометрика: задачи и решения. – М.: Альфа-пресс, 2008. – 192 с.
8. Эффективность экономики России: Федеральная служба государственной статистики [Электронный ресурс] – Режим доступа: https://www.gks.ru/folder/11186 (Дата обращения: 10.04.2020)

ПРИЛОЖЕНИЯ

Приложение 1

Таблица значений *F* –критерия Фишера при уровне значимости α = 0,05

k_2 \ k_1	1	2	3	4	5	6	8	12	24	∞
1	161,45	199,50	215,72	224,57	230,17	233,97	238,89	243,91	249,04	254,32
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,83	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,76	1,48
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	1,95	1,74	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
70	3,98	3,13	2,74	2,50	2,35	2,23	2,07	1,89	1,67	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,88	1,65	1,31
90	3,95	3,10	2,71	2,47	2,32	2,20	2,04	1,86	1,64	1,28
100	3,94	3,09	2,70	2,46	2,30	2,19	2,03	1,85	1,63	1,26
150	3,90	3,06	2,66	2,43	2,27	2,16	2,00	1,82	1,59	1,18
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,80	1,57	1,14
300	3,87	3,03	2,64	2,41	2,25	2,13	1,97	1,79	1,55	1,10
400	3,86	3,02	2,63	2,40	2,24	2,12	1,96	1,78	1,54	1,07
500	3,86	3,01	2,62	2,39	2,23	2,11	1,96	1,77	1,54	1,06
1000	3,85	3,00	2,61	2,38	2,22	2,10	1,95	1,76	1,53	1,03
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1

Приложение 2

Критические значения t-критерия Стьюдента при уровне значимости 0,01, 0,05, 0,1 (двухсторонний)

Число степеней свободы d.f.	α			Число степеней свободы d.f.	α		
	00,10	0,05	0,01		00,10	0,05	0,01
1	6,3138	12,706	63,657	18	1,7341	2,1009	2,8784
2	2,9200	4,3027	9,9248	19	1,7291	2,0930	2,8609
3	2,3534	3,1825	5,8409	20	1,7247	2,0860	2,8453
4	2,1318	2,7764	4,5041	21	1,7207	2,0796	2,8314
5	2,0150	2,5706	4,0321	22	1,7171	2,0739	2,8188
6	1,9432	2,4469	3,7074	23	1,7139	2,0687	2,8073
7	1,8946	2,3646	3,4995	24	1,7109	2,0639	2,7969
8	1,8595	2,3060	3,3554	25	1,7081	2,0595	2,7874
9	1,8331	2,2622	3,2498	26	1,7056	2,0555	2,7787
10	1,8125	2,2281	3,1693	27	1,7033	2,0518	2,7707
11	1,7959	2,2010	3,1058	28	1,7011	2,0484	2,7633
12	1,7823	2,1788	3,0545	29	1,6991	2,0452	2,7564
13	1,7709	2,1604	3,0123	30	1,6973	2,0423	2,7500
14	1,7613	2,1448	2,9768	40	1,6839	2,0211	2,7045
15	1,7530	2,1315	2,9467	60	1,6707	2,0003	2,6603
16	1,7459	2,1199	2,9208	120	1,6577	1,9799	2,6174
17	1,7396	2,1098	2,8982	∞	1,6449	1,9600	2,5758

Галина Александровна Волкова

ЭКОНОМЕТРИКА (ПРОДВИНУТЫЙ УРОВЕНЬ)

Компьютерный практикум
для студентов, обучающихся по направлению подготовки
38.04.01 Экономика,
квалификация магистр

Компьютерная верстка *Г.А. Волковой*

Подписано в печать
Бумага SvetoCopy
Тираж экз.

Формат 60×84 1/16
Усл. печ. л. 3,6
Заказ №

РИО ПГАУ
440014, г. Пенза, ул. Ботаническая, 30