Обработка и анализ экспериментальных данных средствами MS Excel

Вопросы:

I. Статистическая обработка данных
II. Реализация типовых задач в ППП Excel

I. Статистическая обработка данных

Различают два вида зависимостей между явлениями и процессами:

- функциональную,
- статистическую.

Статистическая зависимость существует, когда каждому значению одной переменной соответствует множество возможных значений другой переменной;

т.е. каждому значению одной переменной соответствует определенное распределение другой переменной.

Пример статистической связи: зависимость веса человека от роста.

Статистическая зависимость называется корреляционной, если каждому значению одной переменной соответствует определенное математическое ожидание (среднее значение) другой.

В регрессионном анализе рассматривается односторонняя зависимость случайной переменной *y* от одной (или нескольких) неслучайной независимой переменной *x*.

Односторонняя статистическая зависимость выражается с помощью функции, которая называется регрессией.

Виды регрессии

1. В зависимости от количества переменных, включенных в уравнение регрессии, различают простую (парную) и множественную регрессии.

   Простая (парная) регрессия- это регрессия между двумя переменными.

   Множественная регрессия- это регрессия между зависимой переменной y и несколькими независимыми (объясняющими) переменными $x_1, x_2 \ldots x_m$

2. Относительно формы зависимости различают:

- линейную регрессию, выражаемую линейной функцией;
- нелинейную регрессию, выражаемую нелинейной функцией.

3. В зависимости от характера регрессии различают:

положительную и отрицательную регрессию.

Положительная регрессия имеет место, если с увеличением или уменьшением значений объясняющей переменной, значения зависимой переменной также соответственно увеличиваются или уменьшаются.

В случае отрицательной регрессии с увеличением или уменьшением значений объясняющей переменной, значения зависимой переменной соответственно уменьшаются или увеличиваются.

Понятие положительной или отрицательной регрессии имеют смысл только для простой регрессии.

4. Относительно типа соединения явлений различают:

- непосредственную регрессию (зависимая и объясняющая переменные связаны непосредственно друг с другом);
- косвенную регрессию (объясняющая переменная действует через какую то третью или ряд других переменных на зависимую переменную);
- ложную регрессию (возникает при формальном подходе к исследуемым явлениям, без уяснения того, какие причины обуславливают данную связь).

ПАРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

Теоретическая линейная регрессионная модель представляется в виде:

$Y= a+b*x+\varepsilon$, где

a, b- параметры (коэффициенты) уравнения (регрессии)

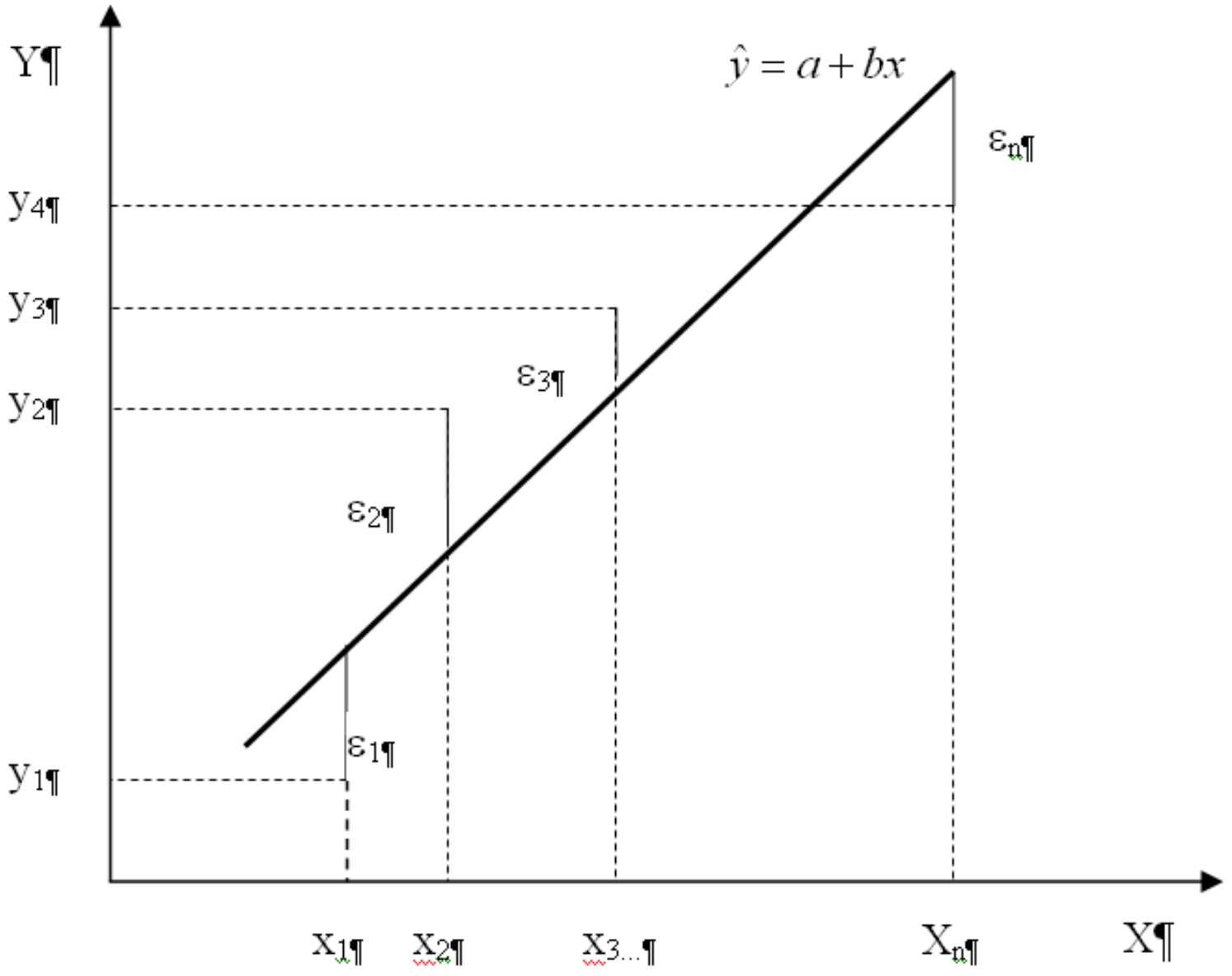
ε- случайная переменная, характеризующая отклонение реального значения зависимой переменной от теоретического, найденного по уравнению регрессии.

Переменную ε называют возмущением.

Задача линейного регрессионного анализа состоит в том, чтобы по имеющимся статистическим данным (x_i,y_i) для переменных X и Y получить наилучшие оценки неизвестных параметров α и β.

Согласно методу наименьших квадратов неизвестные параметры a и b выбираются таким образом, чтобы сумма квадратов отклонений эмпирических значений y_i от значений, найденных по уравнению регрессии, была минимальной.

$$S = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \rightarrow \min$$


Y¶
$\hat{y} = a + bx$
εn¶
y4¶
y3¶
ε3¶
y2¶
ε2¶
ε1¶
y1¶
x1¶
x2¶
x3...¶
Xn¶
X¶

Параметр b называется коэффициентом регрессии. Его величина показывает, на сколько единиц в среднем изменяется переменная y при увеличении переменной x на одну единицу.

Формально a- значение результата y при x=0. Если фактор x не имеет и не может иметь нулевого значения, то такая трактовка коэффициента a не имеет смысла.

Коэффициент корреляции

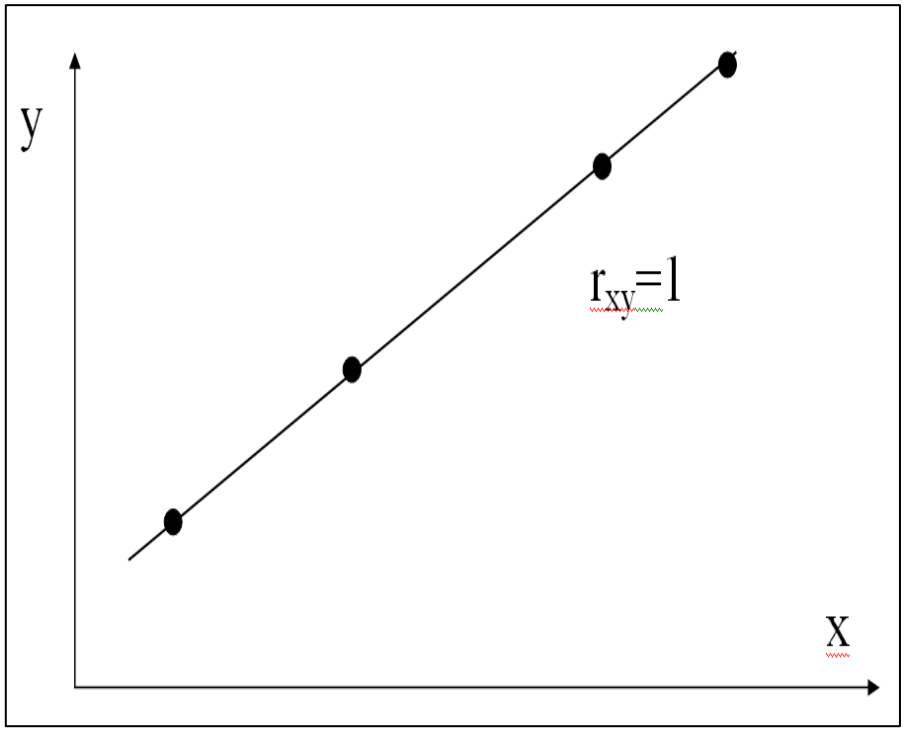
Для оценки тесноты корреляционной зависимости (функциональная зависимость между значением одной переменной и условным математическим ожиданием другой) в случае линейной регрессии используют линейный коэффициент корреляции.

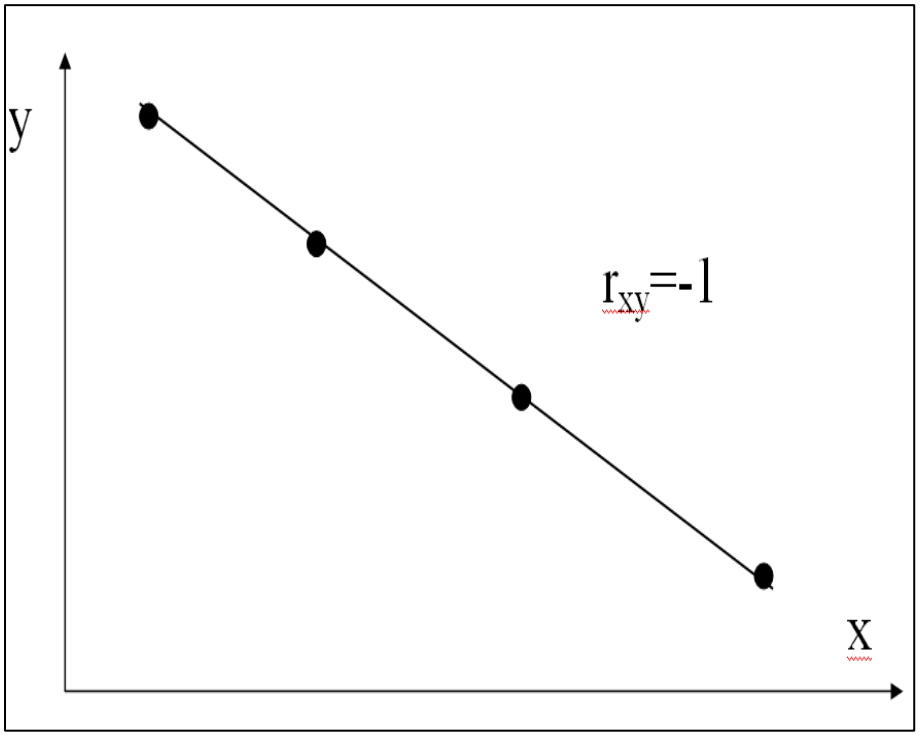
Если $r_{xy}>0$, то корреляционная связь между переменными называется прямой; если $r_{xy}<0$- обратной.

Свойства коэффициента корреляции

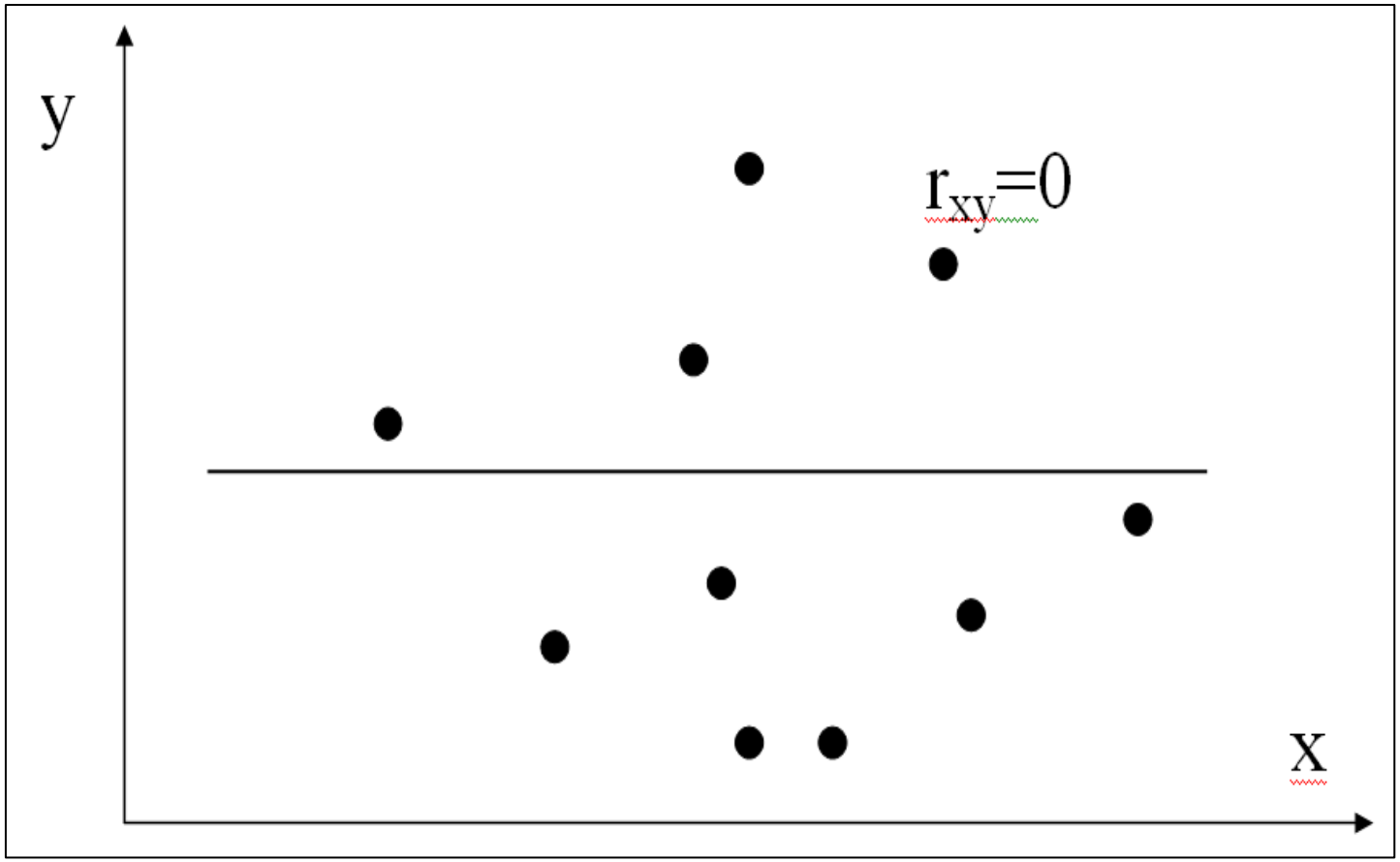
1. Коэффициент корреляции принимает значения на отрезке [-1;1].
   Чем ближе $|r_{xy}|$ к 1,тем теснее связь.
2. При $r_{xy}=\pm 1$ корреляционная связь представляет линейную функциональную зависимость.
   При этом все наблюдаемые значения расположены на прямой линии.

$r_{xy}=\pm 1$

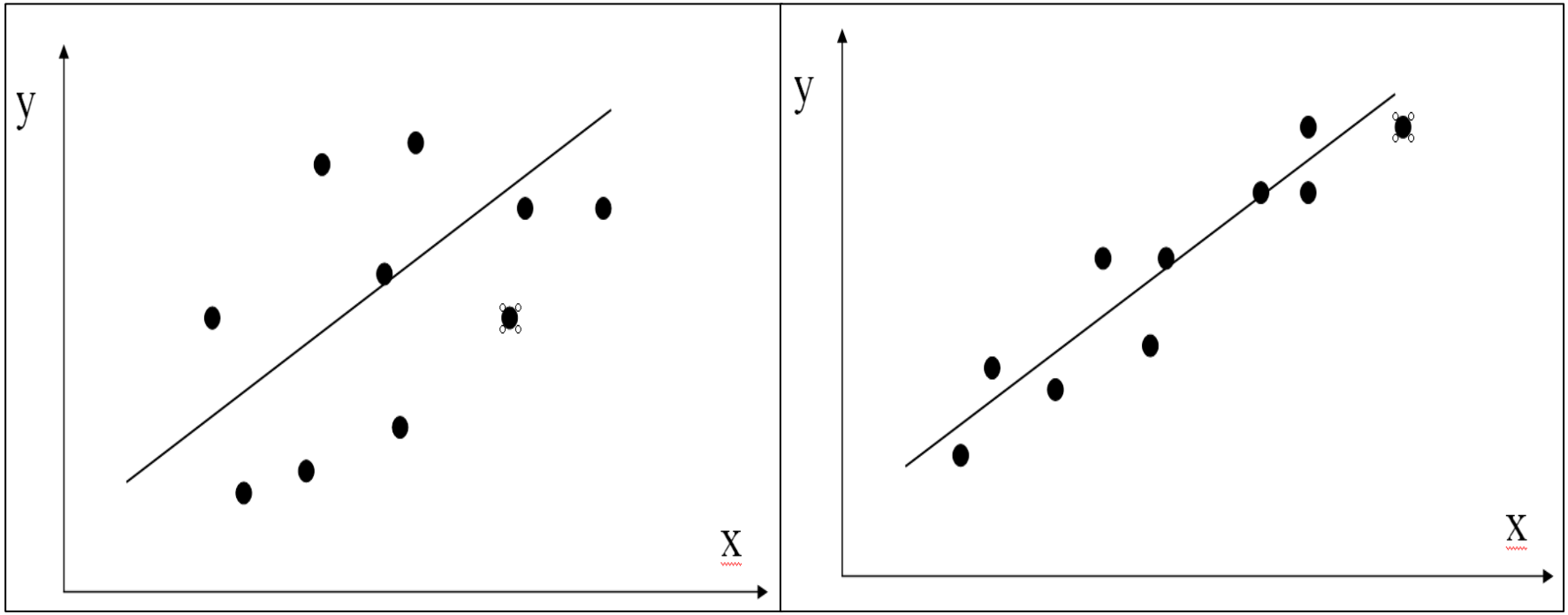




3. При $r_{xy}=0$ линейная корреляционная связь отсутствует. При этом линия регрессии параллельна оси х.



Рассмотрим две корреляционные зависимости представленные графически



а) слабая корреляция

б) тесная корреляция

Оценка тесноты связи на основе шкалы Чеддока

Теснота связи	Значение коэффициента корреляции при наличии	
	Прямой связи	Обратной связи
Слабая	0,1 - 0,3	(-0,1) – (-0,3)
Умеренная	0,3 – 0,5	(-0,3) – (-0,5)
Заметная	0,5 – 0,7	(-0,5) – (-0,7)
Высокая	0,7 – 0,9	(-0,7) – (-0,9)
Весьма высокая	0,9 – 0,99	(-0,9) – (-0,99)

Оценка значимости уравнения регрессии

Проверить значимость уравнения регрессии- значит установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным, и достаточно ли включенных в уравнение объясняющих переменных для описания зависимой переменной.

Проверка значимости уравнения регрессии проводится на основе дисперсионного анализа.

Центральное место в нем занимает разложение общей суммы квадратов отклонений переменной y от среднего значения на две части- «объясненную» и «необъясненную»

Определение F критерия Фишера

Для этого проводится проверка статистической значимости коэффициента детерминации R^2 на основе F-критерия Фишера:

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

где n – число наблюдений;

m – число факторов в уравнении регрессии.

Если в уравнении регрессии свободный член $a_0 = 0$, то числитель n-m-1 следует увеличить на 1, т.е. он будет равен n-m.

- Если $F_{факт.} > F_{табл.}$, то объясненная дисперсия существенно больше остаточной дисперсии, а следовательно уравнение регрессии достаточно качественно отражает динамику изменения зависимой переменной.

Если $F_{факт.} < F_{табл.}$, то объясненная дисперсия соизмерима с дисперсией, вызванной случайными факторами.

Это дает основание считать, что совокупное влияние объясняющих переменных модели несущественно, а общее качество модели невысоко, т.е. признается статистическая незначимость, ненадежность уравнения регрессии.

Одной из наиболее эффективных оценок адекватности регрессионной модели, мерой качества уравнения регрессии является коэффициент детерминации

$$R^2$$

Величина R^2 показывает, какая часть (доля) вариации зависимой переменной обусловлена вариацией объясняющей переменной.

R^2 принимает значения от 0 до 1 включительно.

Если R^2=0, то вариация зависимой переменной полностью обусловлена воздействием неучтенных в модели переменных.

В случае парной линейной регрессии коэффициент детерминации равен квадрату коэффициента корреляции, т.е.

$$R^2 = r_{xy}^2$$

Оценка существенности параметров линейной регрессии и корреляции.

Оценка значимости коэффициента B проводится на основе t- статистики.

Если $|t_b| > t_{табл.}$, то коэффициент регрессии b значим на уровне α.

Если $|t_b| < t_{табл.}$, то признается случайная природа формирования коэффициента b.

II. Реализация типовых задач на компьютере ППП Excel

1. Встроенная статистическая функция **ЛИНЕЙН** определяет параметры линейной регрессии.

   Порядок вычисления:

- введите исходные данные
- выделите область пустых ячеек 5*2 (5 строк, 2 столбца)
- вызовите **Мастер функций**, выберите категорию **Статистические**, функцию **ЛИНЕЙН**, щелкните **ОК**

- Заполните аргументы функции:

  Известные значения y – диапазон, содержащий данные результативного признака;

  Известные значения x – диапазон, содержащий данные независимого признака;

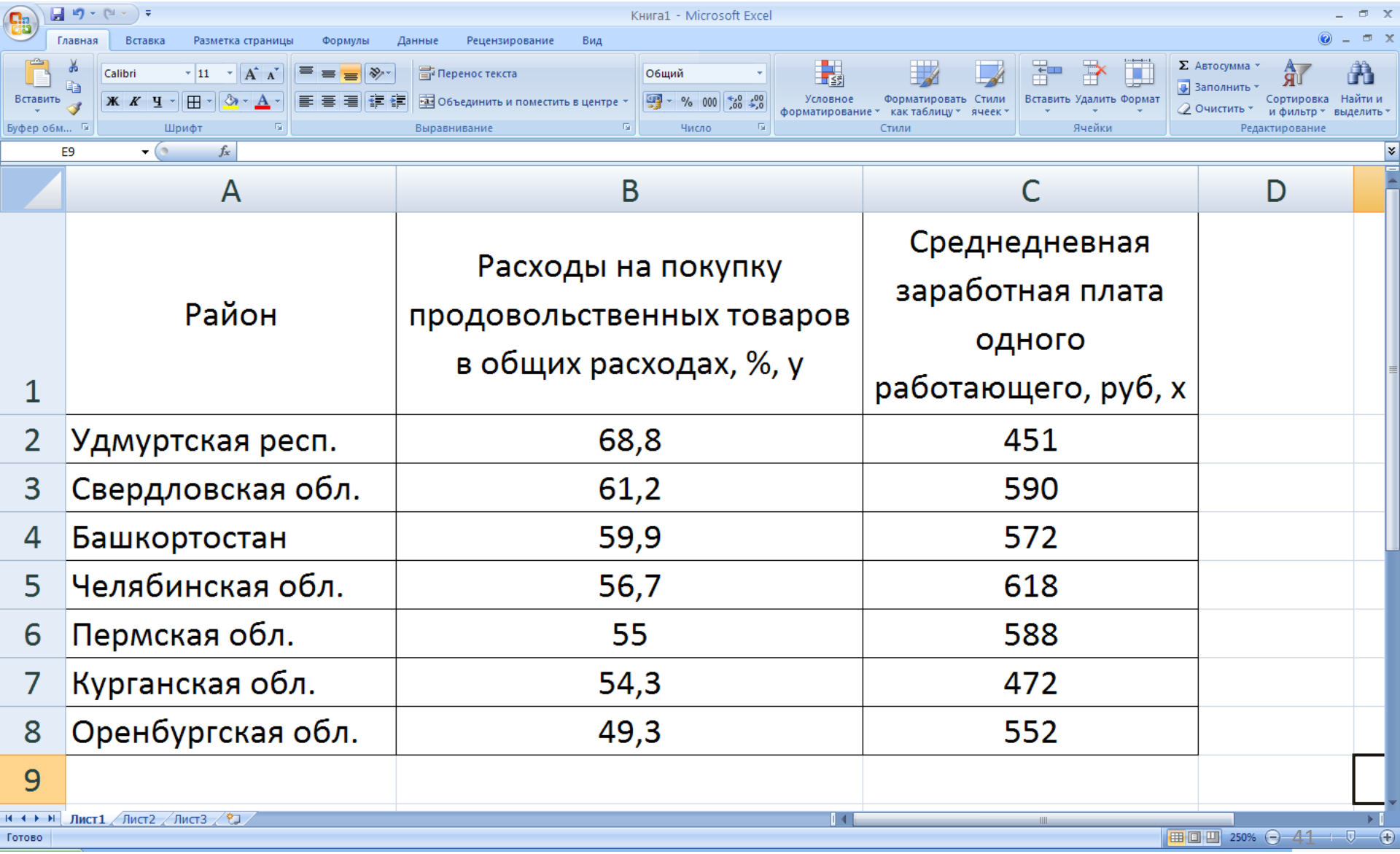
  Константа – если константа равна единице, предполагается наличие свободного члена уравнения, если *Константа* равна 0, то свободный член равен 0;

  Статистика – равна 1, если нужна вся информация по регрессионному анализу.

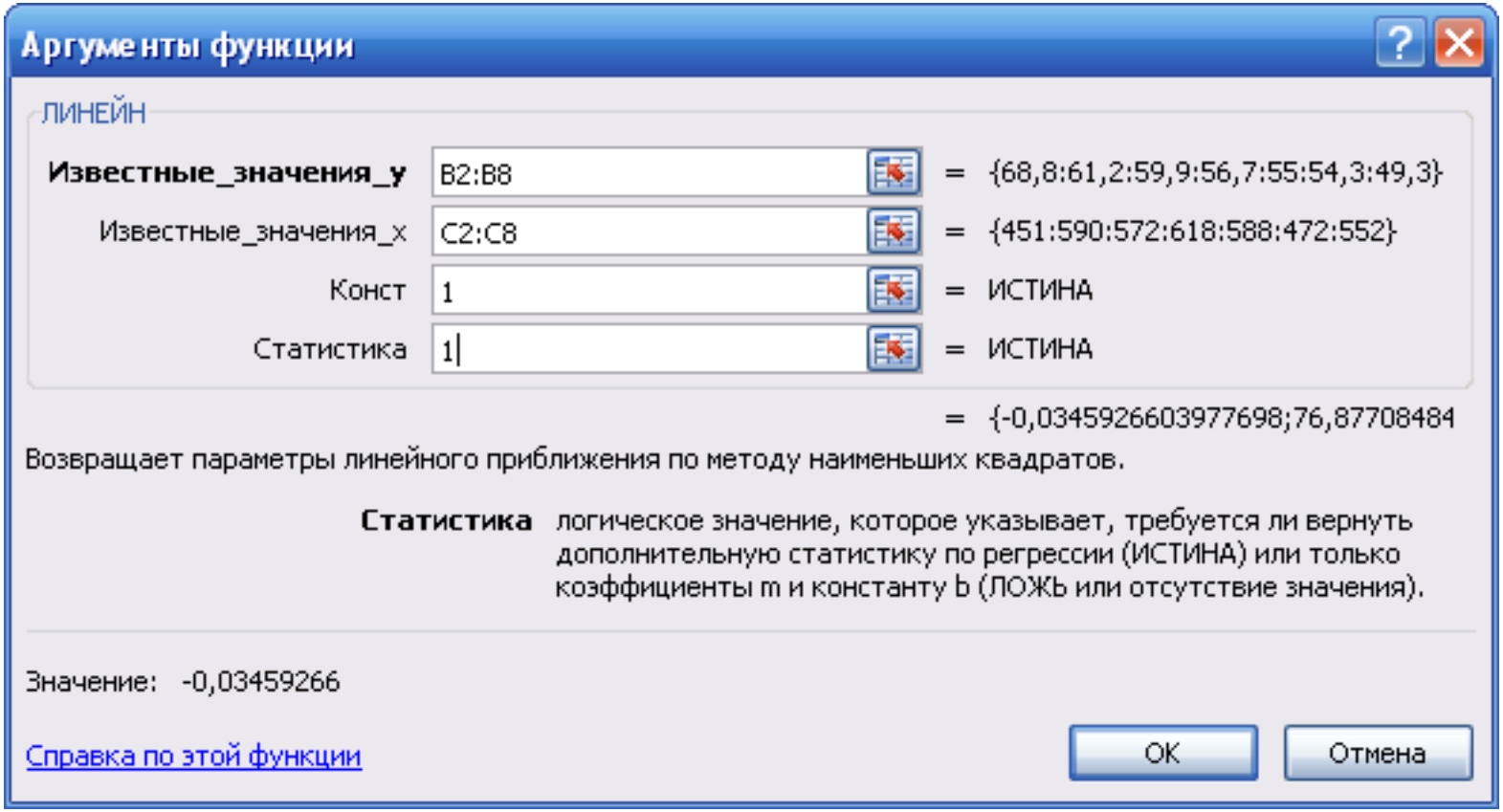
  щелкните ОК.

Чтобы увидеть ответ полностью, нажмите F2, затем <Ctrl> + <Shift> + <Enter>.

Пример



	A	B	C	D
1	Район	Расходы на покупку продовольственных товаров в общих расходах, %, y	Среднедневная заработная плата одного работающего, руб, x	
2	Удмуртская респ.	68,8	451	
3	Свердловская обл.	61,2	590	
4	Башкортостан	59,9	572	
5	Челябинская обл.	56,7	618	
6	Пермская обл.	55	588	
7	Курганская обл.	54,3	472	
8	Оренбургская обл.	49,3	552	
9				


Аргументы функции
ЛИНЕЙН
Известные_значения_y
B2:B8
= {68,8:61,2:59,9:56,7:55:54,3:49,3}
Известные_значения_x
C2:C8
= {451:590:572:618:588:472:552}
Конст
1
= ИСТИНА
Статистика
1
= ИСТИНА
= {-0,0345926603977698;76,87708484
Возвращает параметры линейного приближения по методу наименьших квадратов.
Статистика логическое значение, которое указывает, требуется ли вернуть дополнительную статистику по регрессии (ИСТИНА) или только коэффициенты m и константу b (ЛОЖЬ или отсутствие значения).
Значение: -0,03459266
Справка по этой функции
OK
Отмена

Ответ

-0,03459	76,8771
0,04097	22,6202
0,12479	6,35151
0,71292	5
28,7603	201,708

коэффициент b	коэффициент a
Среднеквадратическое отклонение b	Среднеквадратическое отклонение a
R^2	Среднеквадратическое отклонение y
F - статистика	Число степеней свободы
Регрессионная сумма квадратов	Остаточная сумма квадратов

Уравнение регрессии:

$\hat{y}$=76,87771-0,03459x.

С увеличением среднедневной заработной платы на 1 руб. доля расходов на покупку продовольственных товаров снижается в среднем на 0,035 процента.

Для вычисления параметров экспоненциальной кривой

$$y = \alpha \cdot \beta^{x}$$

применяется встроенная статистическая функция **ЛГРФПРИБЛ**.

Порядок вычисления аналогичен функции **ЛИНЕЙН**.

2. Инструмент анализа данных Регрессия

- Проверьте доступ к пакету анализа (**Сервис/Надстройки**, установить галочку около **Пакет анализа**) (или Главное меню, **Параметры Excel**, **Надстройки**)
- Выполните **Сервис/Анализ данных/Регрессия**, **ОК**. (или вкладка **Данные**, **Анализ данных/ Регрессия**, **ОК**)

- Заполните диалоговое окно:

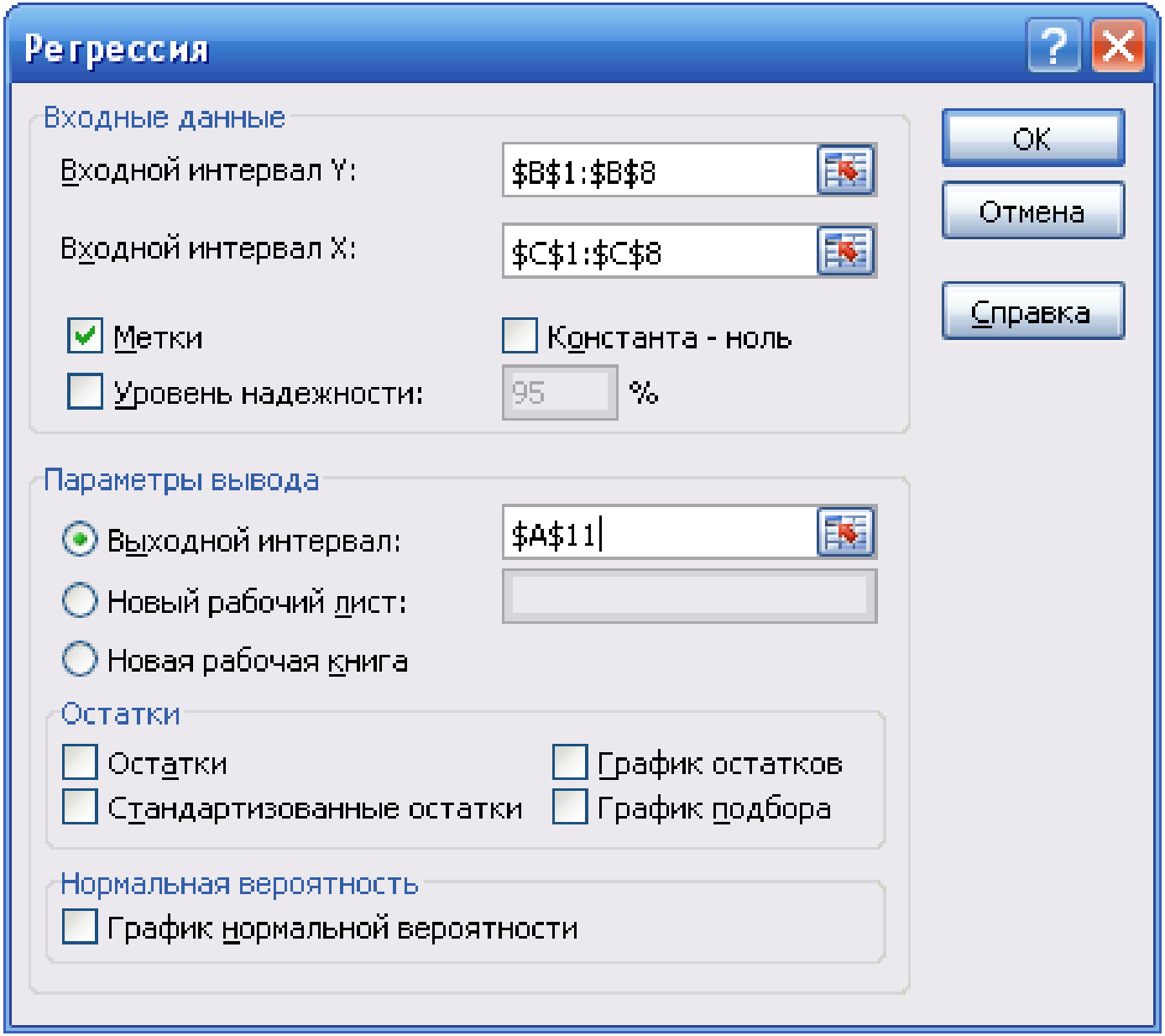
  Входной интервал y – диапазон, содержащий данные результативного признака;

  Входной интервал x – диапазон, содержащий данные независимого признака;

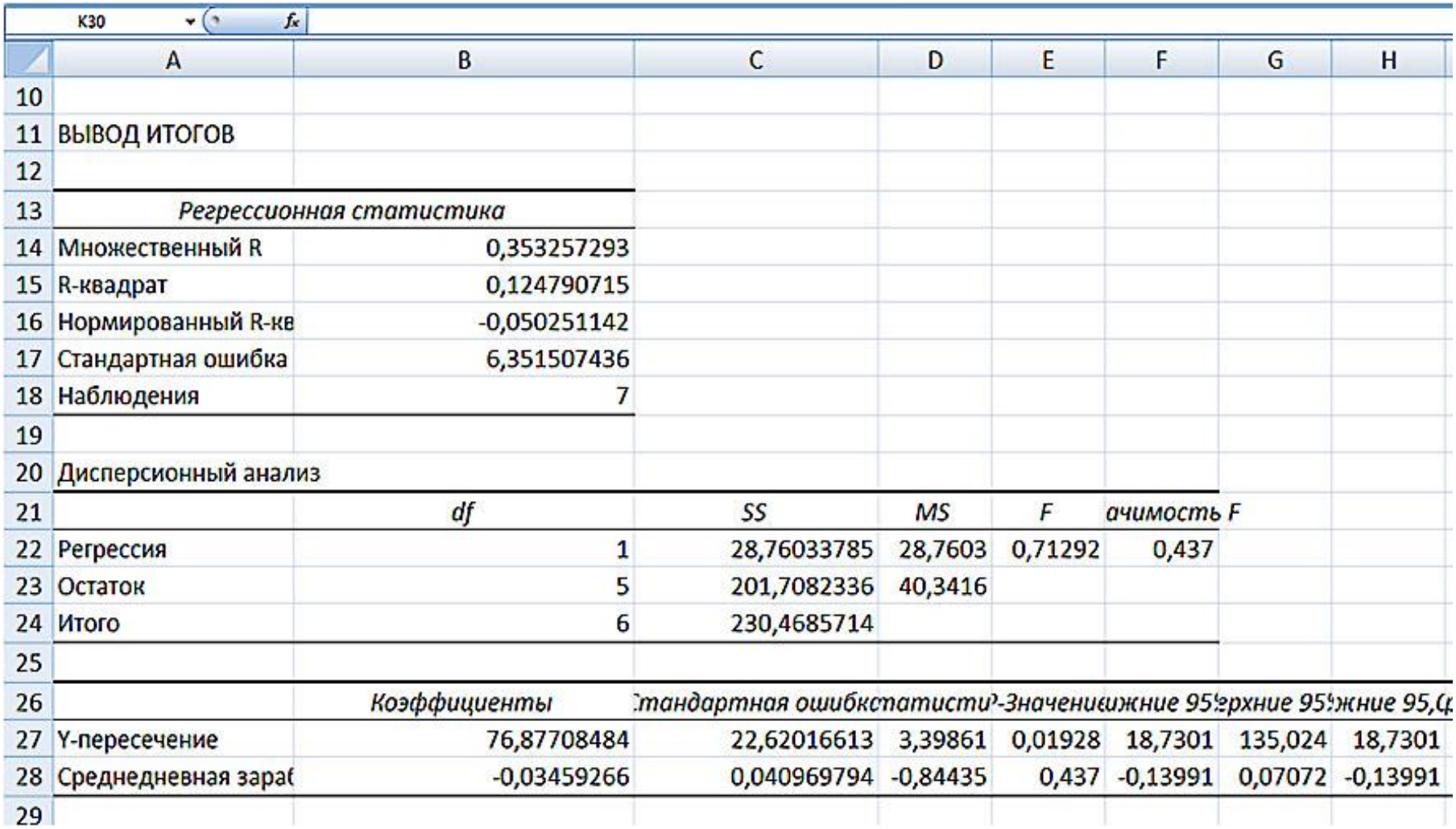
  Метки – указывает, содержит ли первая строка названия столбцов или нет;

  Константа – если константа равна единице, предполагается наличие свободного члена уравнения, если *Константа* равна 0, то свободный член равен 0;

  Выходной интервал – укажите левую верхнюю ячейку ответа.


Регрессия
Входные данные
Входной интервал Y:
$B$1:$B$8
Входной интервал X:
$C$1:$C$8
Метки
Константа - ноль
Уровень надежности:
95
%
OK
Отмена
Справка
Параметры вывода
Выходной интервал:
$A$11
Новый рабочий лист:
Новая рабочая книга
Остатки
Остатки
График остатков
Стандартизованные остатки
График подбора
Нормальная вероятность
График нормальной вероятности

Результат расчета:



	A	B	C	D	E	F	G	H
10								
11	ВЫВОД ИТОГОВ							
12								
13	*Регрессионная статистика*							
14	Множественный R	0,353257293						
15	R-квадрат	0,124790715						
16	Нормированный R-кв	-0,050251142						
17	Стандартная ошибка	6,351507436						
18	Наблюдения	7						
19								
20	Дисперсионный анализ							
21		*df*	*SS*	*MS*	*F*	*ачимость F*		
22	Регрессия	1	28,76033785	28,7603	0,71292	0,437		
23	Остаток	5	201,7082336	40,3416				
24	Итого	6	230,4685714					
25								
26		*Коэффициенты*	*:тандартная ошибкстатисти?*		*-Значениижние 95%*	*рхние 95%*	*жние 95,0%*	
27	Y-пересечение	76,87708484	22,62016613	3,39861	0,01928	18,7301	135,024	18,7301
28	Среднедневная зара	-0,03459266	0,040969794	-0,84435	0,437	-0,13991	0,07072	-0,13991
29								

Уравнение регрессии:

$\hat{y}$=76,87771-0,03459x.

С увеличением среднедневной заработной платы на 1 руб. доля расходов на покупку продовольственных товаров снижается в среднем на 0,035 процента.

Линейный коэффициент парной корреляции r=0,35

Связь умеренная, обратная.

Коэффициент детерминации:

$$r_{xy}^2 = (-0{,}35)^2 = 0{,}127.$$

Вариация результата на 12,7% объясняется вариацией фактора х.

F- критерий $F_{факт}$=0,7

Значимость F=0,437

Значимость F больше 0,05, что указывает на необходимость принять гипотезу H_0 о случайной природе выявленной зависимости и статистической незначимости параметров уравнения и показателя тесноты связи.

Получим уравнение $\hat{y}=exp(5.934)\cdot x^{-0.298}$

Индекс корреляции $\rho_{xy}=0,34$;

Коэффициент детерминации $D=0,12$.

Характеристики модели указывают, что она несколько хуже линейной функции описывает взаимосвязь.

Множественная регрессия и корреляция

Уравнение множественной регрессии может быть представлено в виде:

$$Y=f(X)+\varepsilon$$

где $X= (x_1,x_2,\ldots x_p)$- вектор объясняющих переменных,

ε – случайная ошибка (возмущение).

Основная цель множественной регрессии-

построить модель с бóльшим числом факторов,

определив при этом влияние каждого из них в отдельности,

а также совокупное их воздействие на моделируемый показатель.

Построение уравнения множественной регрессии начинается с решения вопроса о спецификации модели.

К проблемам спецификации относят два типа задач:

1) отбор объясняющих переменных,
2) выбор формы уравнения регрессии.

1). Отбор объясняющих переменных

Факторы, включаемые во множественную регрессию должны отвечать следующим требованиям:

a) должны быть количественно измеримы,
b) не должны быть интеркоррелированы и тем более находиться в точной функциональной связи.

Если величина парного коэффициента корреляции между объясняющими переменными x_i и x_j $r_{x^i x j} \geq 0.7$, то переменные x_i и x_j считаются явно коллинеарными.

В этом случае факторы x_i и x_j дублируют друг друга и один из них рекомендуется исключить из регрессии.

Предпочтение отдается фактору, который при достаточно тесной связи с результатом имеет наименьшую тесноту связи с другими факторами.

При построении уравнения множественной регрессии используют различные методы, позволяющие выполнить отбор наиболее существенных факторов модели:

- метод исключения (отсев факторов из полного набора);
- метод включения (дополнительное введение фактора);
- шаговой регрессионный анализ (исключение ранее введенного фактора)

2) Выбор формы уравнения регрессии

Правильный выбор вида экономической модели является отправной точкой для качественного ее анализа.

Для построения «хорошей» модели и сравнения ее с другими возможными моделями необходимо учитывать следующие свойства:

1) Простота. Модель должна быть максимально простой. Поэтому из двух моделей, приблизительно одинаково отражающих реальность, предпочтение отдается модели, содержащей меньшее число объясняющих переменных.

2) Единственность. Для любого набора статистических данных определяемые коэффициенты должны вычисляться однозначно.
3) Максимальное соответствие. Уравнение тем лучше, чем большую часть разброса зависимой переменной оно может объяснить.

4) Согласованность с теорией. Модель должна опираться на теоретический фундамент.
5) Прогнозные качества. Модель может быть признана качественной, если полученные на ее основе прогнозы подтверждаются реальностью.

Виды уравнений множественной регрессии

Наиболее простой из моделей множественной регрессии является линейная множественная регрессия вида:

$$\hat{y}=a+b_1x_1+b_2x_2+\ldots+b_px_p$$

Параметры $b_1,b_2\ldots b_p$ называются коэффициентами «чистой» регрессии.

Они характеризуют среднее изменение результата с изменением соответствующего фактора на единицу при неизменном значении других факторов, закрепленных на среднем уровне.

Пример:

Зависимость прибыли предприятия y (млн. руб.) от величины оборотных средств x_1 (млн. руб.) и стоимости основных фондов x_2 (млн. руб.) характеризуется уравнением:

$$\hat{y}=0,66x_1+0.21x_2$$

Экономическая сущность коэффициентов при x_1 и x_2:

Увеличение оборотных средств на 1 млн. руб. при той же стоимости основных фондов ведет к росту прибыли на 660 тыс. руб.,

а увеличение основных фондов на 1 млн. руб. ведет к росту прибыли на 210 тыс. руб. при той же величине оборотных средств.