Лабораторная работа.
Использование электронных таблиц MS Excel при решении задач парного регрессионного анализа

Цель работы – научиться выявлять: 1) факт изменчивости изучаемого явления при определенных, но не всегда четко фиксированных условиях; 2) тенденцию как периодическое изменение признака; 3) закономерность, выраженную в виде корреляционного уравнения (регрессии).

Методы регрессионного анализа рассчитаны, главным образом, на случай устойчивого нормального распределения, в котором изменения от опыта к опыту проявляются лишь в виде независимых испытаний.

1. Линейная модель парной регрессии

Регрессия – функция, позволяющая по величине одного корреллируемого признака определить среднюю величину другого признака.

Выделим основные этапы регрессионного анализа.

Первый этап. Предположение. На этом этапе происходит выбор формы связи между переменными (модель).

Второй этап. Параметризация – происходит оценка значений параметра в выбранной формуле статистической связи. Форма связи (функция) линейная, нелинейная.

Третий этап. Проверка надёжности полученных оценок. На этом этапе осуществляются следующие тесты: *F*-тест (проверка статистической значимости выбранной формы связи), *t*-тест (проверка статистической значимости найденных числовых значений параметра). В результате анализа статистических данных, выбора и построения модели последовательно выполняются все три этапа.

Рассмотрим простейшую модель регрессии – **линейную регрессию**. Линейная регрессия для *x* и *y* сводится к нахождению уравнения вида

$$y_x = a + b \cdot x, \text{ или } y = a + b \cdot x + \varepsilon.$$

Параметры a и b могут быть найдены по готовым формулам: $a = \overline{y} - b \cdot \overline{x}$, $b = \dfrac{\overline{yx} - \overline{y} \cdot \overline{x}}{\overline{x^2} - (\overline{x})^2}$, где $\overline{x} = \dfrac{1}{n}\sum x$, $\overline{y} = \dfrac{1}{n}\sum y$, $\overline{y \cdot x} = \dfrac{1}{n}\sum y \cdot x$, $\overline{x^2} = \dfrac{1}{n}\sum x^2$.

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

Уравнение регрессии всегда дополняется показателем тесноты связи. При использовании линейной регрессии в качестве такого показателя выступает линейный коэффициент корреляции $r_{xy} = b \cdot \dfrac{\sigma(x)}{\sigma(y)} = \dfrac{\overline{yx} - \overline{y} \cdot \overline{x}}{\sigma(x) \cdot \sigma(y)}$.

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 аналитическая 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.

Эмпирическое значение критерия Фишера находят по формуле:

$$F_{эмп} = \frac{r_{xy}^2}{1 - r_{xy}^2} \cdot (n-2)$$

Критическое значение критерия Фишера находят по статистической функции **F.ОБР.ПХ**: $F_{крит}(\alpha;\ k_1;\ k_2)$, где число степеней свободы: $k_1 = 1$ и $k_2 = n - 2$ при уровне значимости α (чаще 0,05). Эмпирическое и критическое значения критерия между собой сравниваются с учетом того, что критерий Фишера правосторонний: если $F_{эмп} < F_{крит}$, то на уровне значимости α признаётся статистическая незначимость уравнения регрессии в целом.

В парной линейной регрессии оценивается значимость не только уравнения в целом, но и отдельных его параметров.

t-распределение Стьюдента применяется для **проверки существенности коэффициента регрессии**. Эмпирическое значение t-критерия Стьюдента: $t_b = t_r = \sqrt{F_{эмп}}$ сравнивается с его критическим значением **СТЬЮДЕНТ.ОБР.2Х** $t_{крит}(\alpha;\ k_2)$, α – уровень значимости, число степеней свободы $k_2 = n - 2$. При этом мы учитываем, что критерий Стьюдента правосторонний: если $t_{эмп} < t_{крит}$, то на уровне значимости α признаётся статистическая незначимость коэффициента b регрессии.

Пример. Изучалась зависимость между массой матерей x_i, измеряемой в начале беременности (кг), и массой новорождённых детёнышей шимпанзе y_i (кг). Найти уравнение линейной регрессии, проверить модель и ее параметры на статистическую значимость и сделать в прогноз для 15 кг.

Решение. Здесь под независимой переменной *x* будем понимать массу матерей, а под зависимой переменной *y* – массу новорожденных детёнышей.

Для расчёта необходимых сумм и произведений составим вспомогательную таблицу.

Масса матерей x_i	Масса детёнышей y_i	$x_i \cdot y_i$	x_i^2	y_i^2
10	0,7	7	100	0,49
10	0,7	7	100	0,49

10,1	0,65	6,565	102,01	0,4225
10,2	0,61	6,222	104,04	0,3721
10,8	0,73	7,884	116,64	0,5329
11	0,65	7,15	121	0,4225
11,1	0,65	7,215	123,21	0,4225
11,3	0,75	8,475	127,69	0,5625
11,3	0,7	7,91	127,69	0,49
11,4	0,7	7,98	129,96	0,49
11,8	0,69	8,142	139,24	0,4761
12	0,72	8,64	144	0,5184
12	0,6	7,2	144	0,36
12,1	0,75	9,075	146,41	0,5625
12,3	0,63	7,749	151,29	0,3969
13	0,8	10,4	169	0,64
13,4	0,78	10,452	179,56	0,6084
13,5	0,7	9,45	182,25	0,49
14,5	0,7	10,15	210,25	0,49
15,6	0,85	13,26	243,36	0,7225
Сумма 237,4	14,06	167,919	2861,6	9,9598
Среднее 11,87	0,703	8,39595	143,08	0,49799

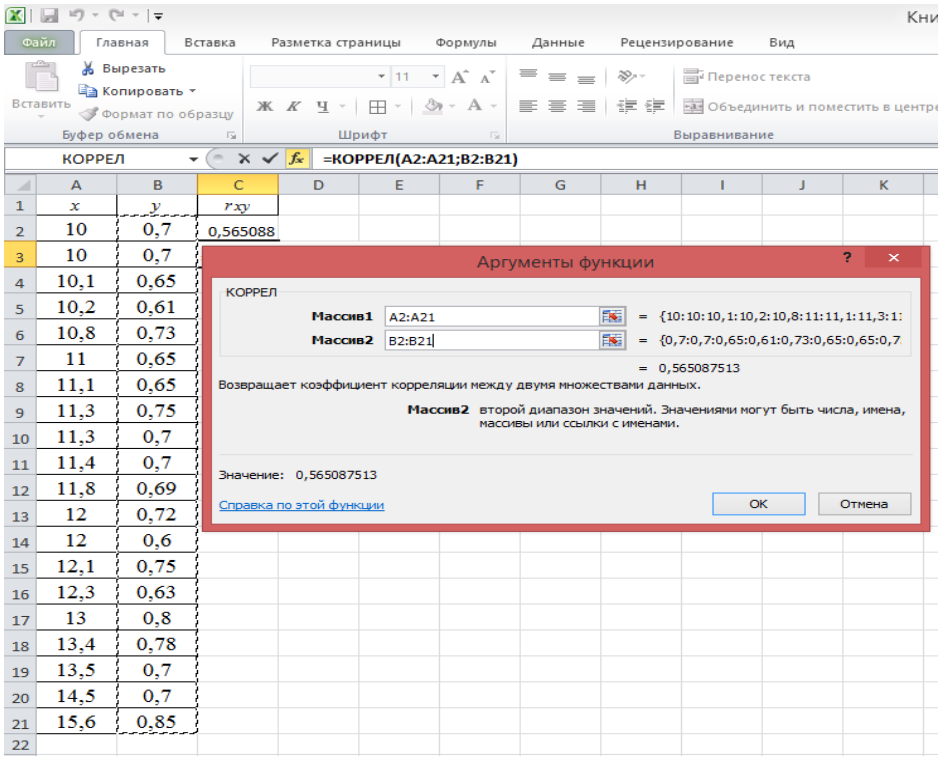
Определим параметры линейной регрессии, используя формулу и ранее найденное значение коэффициента корреляции.

$$b = \frac{\overline{yx} - \bar{y} \cdot \bar{x}}{\overline{x^2} - (\bar{x})^2} = \frac{8{,}39595 - 0{,}703 \cdot 11{,}87}{143{,}08 - 11{,}87^2} = 0{,}0235.$$

$$a = \bar{y} - b \cdot \bar{x} = 0{,}703 - 0{,}0235 \cdot 11{,}87 = 0{,}424.$$

Следовательно, фактическое уравнение регрессии массы детёнышей шимпанзе по значениям массы их матерей имеет вид $\hat{y}_x = 0{,}024 \cdot x + 0{,}424$, то есть при увеличении массы матери на 1 кг у детёныша ожидается увеличение массы на 0,024 кг.

Найдем коэффициент корреляции, используя статистическую функцию **КОРРЕЛ**.



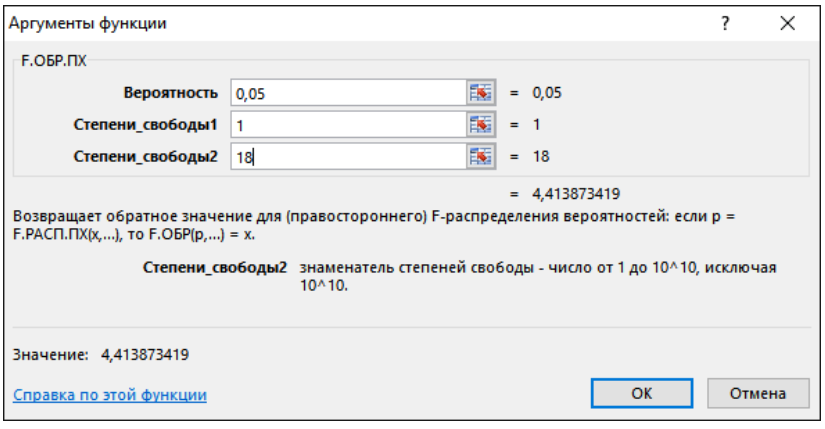
x	y
10	0,7
10	0,7
10,1	0,65
10,2	0,61
10,8	0,73
11	0,65
11,1	0,65
11,3	0,75
11,3	0,7
11,4	0,7
11,8	0,69
12	0,72
12	0,6
12,1	0,75
12,3	0,63
13	0,8
13,4	0,78
13,5	0,7
14,5	0,7
15,6	0,85

В свою очередь найдём квадрат коэффициента корреляции (коэффициент детерминации): $R^2 = (r_{xy})^2 = 0{,}565^2 = 0{,}319$. Коэффициент детерминации показывает, что вариация массы новорождённых детёнышей на 31,9% обусловлена изменчивостью массы матерей.

Оценим качество уравнения регрессии в целом с помощью F-критерия Фишера, найдём эмпирическое значение критерия $F_{эмп}$ по формуле:

$$F_{эмп} = \frac{R^2}{1-R^2} \cdot (n-2) = \frac{0{,}319}{1-0{,}319} \cdot 18 = 8{,}43.$$

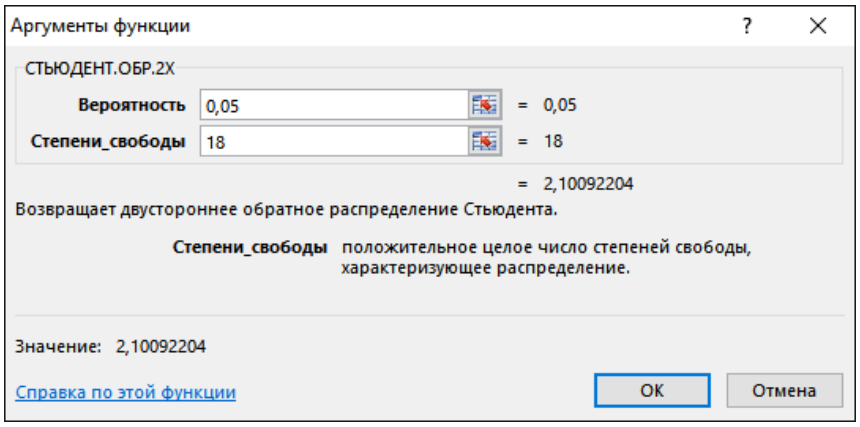
Найдем критическое значение критерия Фишера, используя статистическую функцию **F.ОБР.ПХ**, оно равно $F_{крит}$ (0,05; 1; 18) = 4,41.



Таким образом, $F_{эмп} > F_{крит}$, так как $8{,}43 > 4{,}41$, и на уровне значимости 0,05 признаётся статистическая значимость уравнения в целом.

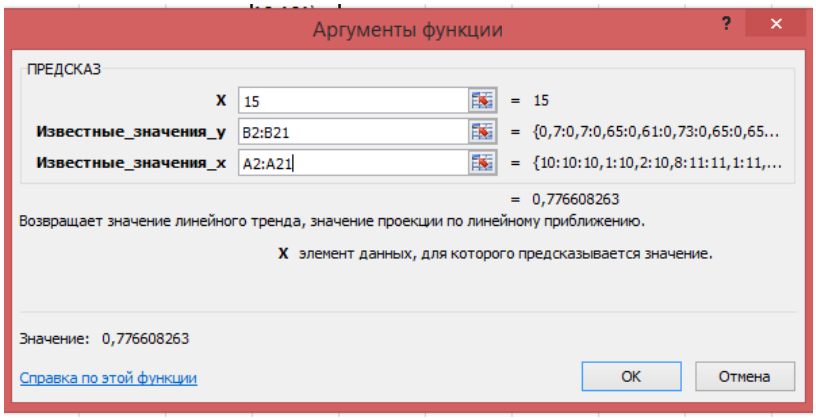
Для оценки статистической значимости коэффициентов регрессии и корреляции рассчитаем t-критерий Стьюдента: $t_b = \sqrt{F_{эмп}} = \sqrt{8{,}43} = 2{,}90$. Для уровня значимости

α = 0,05 найдём критическое значение критерия Стьюдента, используя статистическую функцию **СТЬЮДЕНТ.ОБР.2Х**, оно равно $t_{крит}$ (0,05; 18) = 2,10.



Таким образом, $t_b \geq t_{крит}$, так как 21,34 > 2,1, и на уровне значимости 0,05 делаем вывод о статистической значимости показателя, стоящего перед x.

Для выполнения прогноза воспользуемся функцией **ПРЕДСКАЗ** (категория статистические).



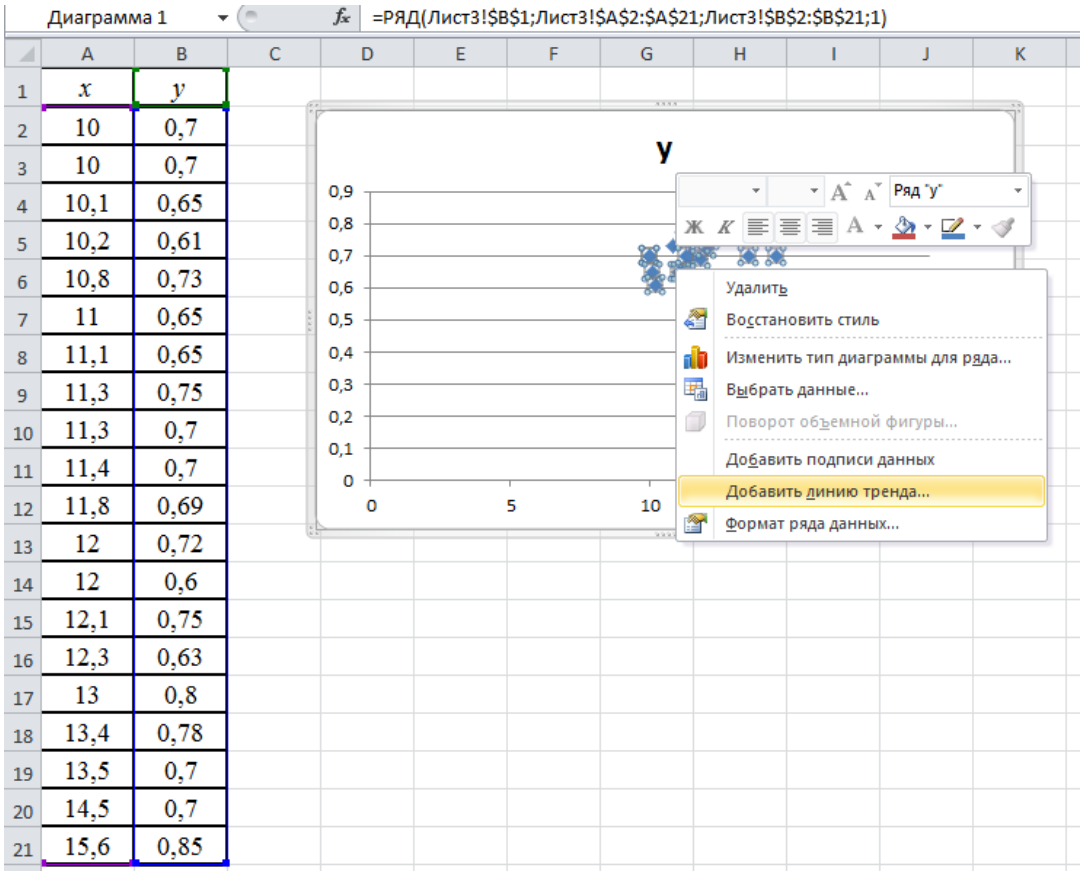
В силу того, что значение коэффициента корреляции, уравнение регрессии и параметр при x статистически значимы, по найденному уравнению регрессии можем делать статистические прогнозы, так, если масса самки шимпанзе равна 15 кг, то ожидаемая масса новорождённого детёныша будет равна 0,78 кг (можно легко проверить, подставив в найденную модель прогнозное значение, $\hat{y}_x = 0{,}024 \cdot 15 + 0{,}424 = 0{,}784$ кг).

2. Построение линии регрессии на корреляционном поле

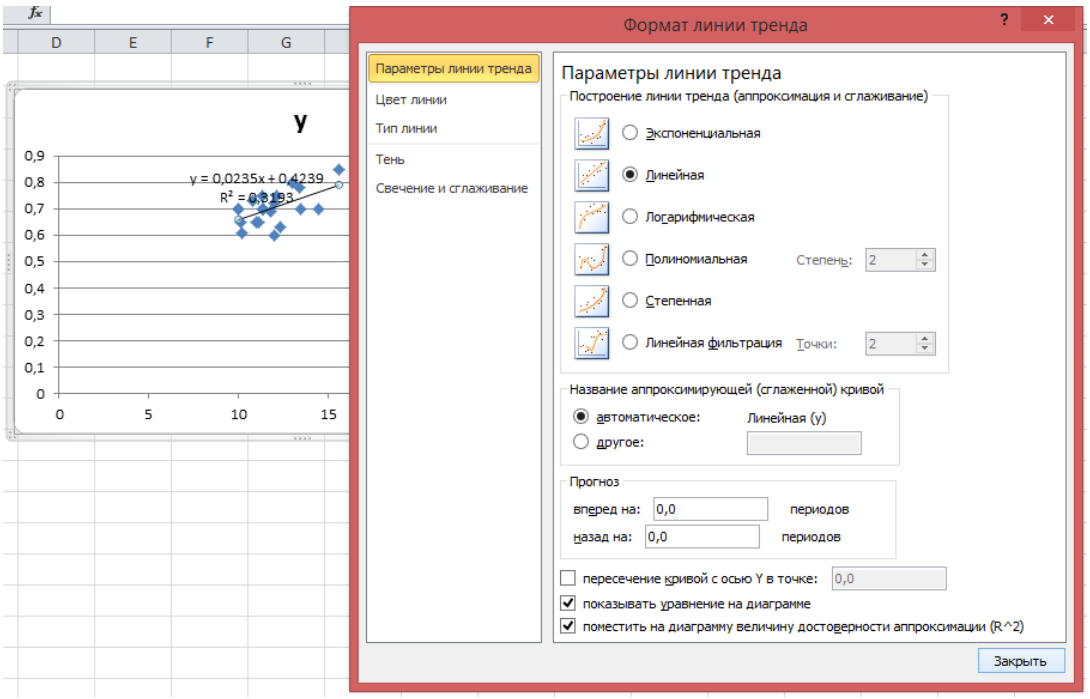
Корреляционное поле – это совокупность (набор) точек с координатами (x_i; y_i). Для его построения можно использовать **Мастер диаграмм**, тип диаграммы **Точечная.**

Для построения линии регрессии направьте курсор мыши на любую точку диаграммы, нажмите правую кнопку мыши, выберите в меню **Добавить линию тренда.** Далее выберите тип линии (**линейная**, **экспоненциальная**,

логарифмическая, **полиномиальная** или **степенная**) и поставьте метку **Показать уравнение на диаграмме**.



	x	y
2	10	0,7
3	10	0,7
4	10,1	0,65
5	10,2	0,61
6	10,8	0,73
7	11	0,65
8	11,1	0,65
9	11,3	0,75
10	11,3	0,7
11	11,4	0,7
12	11,8	0,69
13	12	0,72
14	12	0,6
15	12,1	0,75
16	12,3	0,63
17	13	0,8
18	13,4	0,78
19	13,5	0,7
20	14,5	0,7
21	15,6	0,85



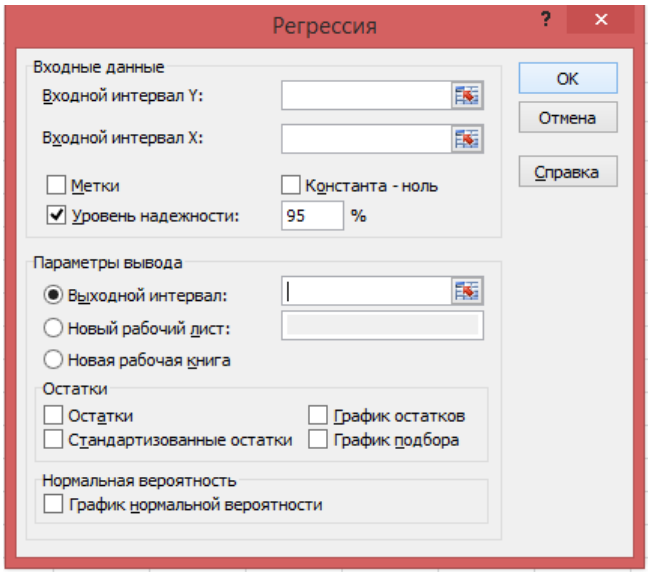
3. Использование Пакета анализа в регрессионном анализе

Для нахождения уравнения линейной регрессии MS Excel используется процедура инструмента анализа данных **Регрессия**. С помощью **Регрессия** можно получить: результаты регрессионной статистики, результаты дисперсионного анализа,

доверительные интервалы, остатки и графики подбора линии регрессии, остатки и нормальную вероятность.

Замечание. Если статистические данные по каждой переменной представлены в виде строк, то таблицу данных необходимо транспонировать (строки сделать столбцами). Для этого необходимо выделить таблицу данных, нажать правую кнопку мыши, выбрать **Копировать**, далее курсор мыши переместить на любую пустую ячейку рабочего листа MS Excel и вновь нажать правую кнопку мыши, выбрать **Специальная вставка / Специальная вставка / Транспонировать.**

Для реализации процедуры инструмента анализа данных необходимо выполнить: **Данные / Анализ данных / Инструменты анализа / Регрессия.**



В появившемся диалоговом окне указать:

Входной интервал Y – адреса ячеек, содержащих выборочные значения переменной *Y*;

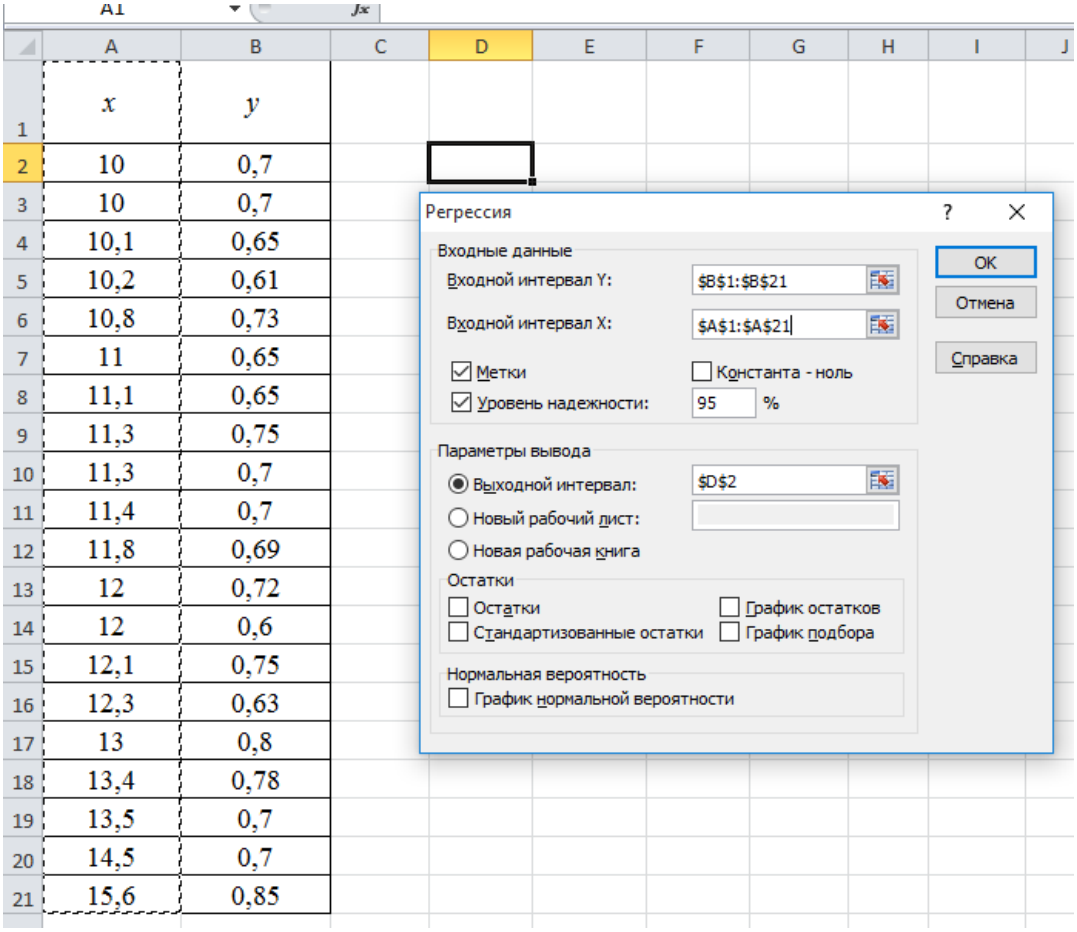
Входной интервал X – адреса ячеек, содержащих выборочные значения переменной *X*;

Метки – включается, если учитываются заголовки столбцов данных;

Уровень надежности – по умолчанию равен 95%;

Выходной интервал – указывается, куда выводятся результаты вычислений.

Далее нажать кнопку **ОК**.



	A	B
1	*x*	*y*
2	10	0,7
3	10	0,7
4	10,1	0,65
5	10,2	0,61
6	10,8	0,73
7	11	0,65
8	11,1	0,65
9	11,3	0,75
10	11,3	0,7
11	11,4	0,7
12	11,8	0,69
13	12	0,72
14	12	0,6
15	12,1	0,75
16	12,3	0,63
17	13	0,8
18	13,4	0,78
19	13,5	0,7
20	14,5	0,7
21	15,6	0,85

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,565087513
R-квадрат	0,319323898
Нормированный R-кв	0,281508559
Стандартная ошибка	0,05347519
Наблюдения	20

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	0,02414727	0,02415	8,444	0,009424365
Остаток	18	0,05147273	0,00286		
Итого	19	0,07562			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
y	0,423853007	0,09680326	4,3785	4E-04	0,220476902	0,627229112
x	0,023517017	0,00809283	2,90591	0,009	0,006514608	0,040519426

Здесь была рассмотрена наиболее простая форма связи между двумя признаками (переменными *X* и *Y*) – линейная. Между тем зависимости между признаками могут принимать самые разнообразные формы.

Задачи для самостоятельного решения

Задача 1. Имеются результаты измерений: образцы некоторого сплава были изготовлены при различных температурах *X*, после чего была измерена прочность каждого образца *Y*.

X	6,7	6,9	7,2	7,3	8,4	8,8	9,1	9,8	10,6	10,7	11,1	11,8	12,1	12,4
Y	2,8	2,2	3	3,5	3,2	3,7	4	4,8	6	5,4	5,2	5,4	6	9

По выборке необходимо построить парную линейную регрессию.

Задача 2. На основе данных по группе хозяйств о среднегодовой численности работников (*X*, чел.) и о стоимости валовой продукции (*Y*, тыс. руб.)

X	96	58	135	153	108	105	76	119	118	149	99
Y	4603	4053	9665	5146	4850	7132	6257	7435	7560	4110	2988

Необходимо построить уравнение линейной регрессии. Вычислить прогноз валового производства при значении среднегодового количества работников, составляющем 115% от среднего уровня.