В таблице 7 приведены данные по территориям региона за 199X год. Число k рассчитывается по формуле

$$k = 100 + 10 \cdot i + j,$$

где i, j – две последние цифры зачетной книжки соответственно. (i = 1, j = 6)

Требуется:

1. Построить поле корреляции.

2. Для характеристики зависимости y от x:

а) построить линейное уравнение парной регрессии y от x;

б) оценить тесноту связи с помощью показателей корреляции и коэффициента детерминации;

в) оценить качество линейного уравнения с помощью средней ошибки аппроксимации;

г) дать оценку силы связи с помощью среднего коэффициента эластичности и бета – коэффициента;

д) оценить статистическую надежность результатов регрессионного моделирования с помощью F – критерия Фишера.

е) оценить статистическую значимость параметров регрессии и корреляции.

3. Проверить результаты, полученные в п. 2 с помощью **ППП *Excel*.**

4. Рассчитать параметры показательной парной регрессии. Проверить результаты с помощью **ППП *Excel***. Оценить статистическую надежность указанной модели с помощью F – критерия Фишера.

5. Обоснованно выбрать лучшую модель и рассчитать по ней прогнозное значение результата, если прогнозное значение фактора увеличится на 5% от среднего уровня. Определить доверительный интервал прогноза при уровне значимости $\gamma = 0{,}05$.

Таблица 1

№ региона	Среднедушевой прожиточный минимум в день, *руб. x*	Среднедневная зарплата, руб., y
1	97	118
2	79	92
3	86	122
4	77	113
5	104	117
6	69	111
7	100	110
8	93	128
9	81	115
10	102	120
11	74	98
12	90	116

Решение

1. Построим поле корреляции, для чего отложим на плоскости в прямоугольной системе координат точки (x_i, y_i) (рис 1.)

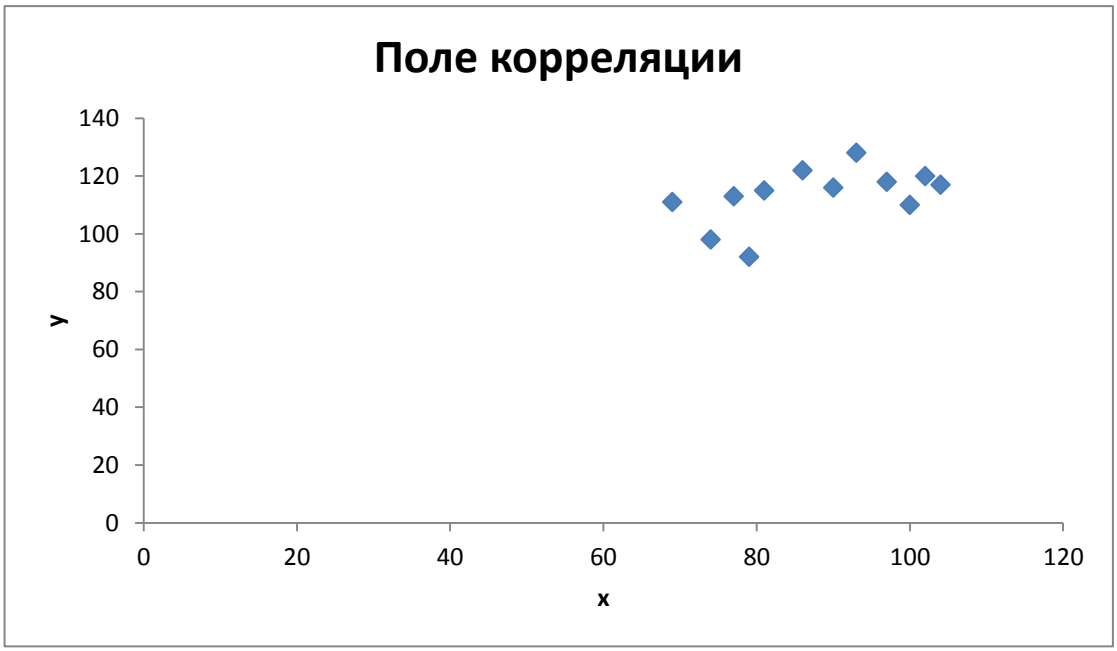

Рисунок 1

2. Для расчета параметров линейной регрессии строим расчетную таблицу 2

Таблица 2

	x	x^2	y	xy	y^2	$(y-\bar{y})^2$	$(x-\bar{x})^2$	$\hat{y}$	$(y-\hat{y})^2$	A(%)
1	97	9409	118	11446	13924	21,78	87,11	117,38	0,38	0,005
2	79	6241	92	7268	8464	455,11	75,11	109,57	308,76	0,191
3	86	7396	122	10492	14884	75,11	2,78	112,61	88,17	0,077
4	77	5929	113	8701	12769	0,11	113,78	108,70	18,46	0,038
5	104	10816	117	12168	13689	13,44	266,78	120,42	11,72	0,029
6	69	4761	111	7659	12321	5,44	348,44	105,23	33,28	0,052
7	100	10000	110	11000	12100	11,11	152,11	118,69	75,46	0,079
8	93	8649	128	11904	16384	215,11	28,44	115,65	152,56	0,096
9	81	6561	115	9315	13225	2,78	44,44	110,44	20,80	0,040
10	102	10404	120	12240	14400	44,44	205,44	119,55	0,20	0,004
11	74	5476	98	7252	9604	235,11	186,78	107,40	88,38	0,096
12	90	8100	116	10440	13456	7,11	5,44	114,35	2,74	0,014
Среднее	87,7	7811,8	113,3	9990,4	12935					6,01%
Сумма	1052	93742	1360	119885	155220	1086,67	1516,67		800,91	0,721
σ	11,24		9,52							
σ^2	126,39		90,56							

2 а) Построим линейное уравнение парной регрессии *y* по *x*. Используя данные таблицы 2, имеем

$$\beta = \frac{\overline{y \cdot x} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2} = \frac{9990{,}4 - 113{,}3 \cdot 87{,}7}{7811{,}8 - 87{,}7^2} = 0{,}434 ,$$

$$\alpha = \bar{y} - \beta\bar{x} = 113{,}3 - 0{,}434 \cdot 87{,}7 = 75{,}28 .$$

Тогда линейное уравнение парной регрессии имеет вид

$$\hat{y} = 75{,}28 + 0{,}434 \cdot x .$$

Оно показывает, что с увеличением среднедушевого прожиточного минимума на 1 руб. средняя зарплата возрастает в среднем на 0,434 руб.

2 б) Тесноту линейной связи оценим с помощью линейного коэффициента парной корреляции

$$r_{xy} = \beta \frac{\sigma_x}{\sigma_y} = 0{,}434 \cdot \frac{11{,}24}{9{,}52} = 0{,}513 .$$

Найдем коэффициент детерминации

$$R^2 = r_{xy}^2 = 0{,}513^2 = 0{,}26 .$$

Это означает, что 26% вариации заработной платы *y* объясняется вариацией фактора *x* – среднедушевого прожиточного минимума.

2 в) Для оценки качества полученной модели найдем среднюю ошибку аппроксимации

$$\overline{A} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\,\% = \frac{1}{n}\sum_{i=1}^{n}|A_i| \cdot 100\,\% = \frac{0{,}721}{12} \cdot 100\,\% = 6{,}01\% .$$

В среднем, расчетные значения отклоняются от фактических на 6,01%. Качество построенной модели оценивается как хорошее, т.к. значение $\overline{A}$ менее 8 %.

2 г) Для оценки силы связи признаков *y* и *x* найдем средний коэффициент эластичности

$$Э_{\hat{y}x} = f'(x)\frac{\bar{x}}{\bar{y}} = \frac{\beta\bar{x}}{\alpha + \beta\bar{x}} = \frac{0{,}434 \cdot 87{,}7}{75{,}28 + 0{,}434 \cdot 87{,}7} = 0{,}336 .$$

Т.о., в среднем на 0,336% по совокупности изменится среднедневная зарплата от своей средней величины при изменении среднедушевого прожиточного минимума в день одного трудоспособного на 1%.

Бета–коэффициент

$$\beta_{yx} = \beta \frac{\sigma_x}{\sigma_y} = 0{,}434 \cdot \frac{11{,}24}{9{,}52} = 0{,}513$$

показывает, что среднее квадратическое отклонение среднедневной зарплаты изменится в среднем на 51,3% от своего значения при изменении прожиточного минимума в день одного трудоспособного на величину его среднего квадратического отклонения.

2 д) Для оценки статистической надежности результатов используем F – критерий Фишера.

Выдвигаем нулевую гипотезу H_o о статистической незначимости полученного линейного уравнения.

Рассчитаем фактическое значение F – критерия при заданном уровне значимости $\gamma = 0{,}05$

$$F_{факт} = \frac{r_{xy}^2}{1 - r_{xy}^2}(n-2) = \frac{0{,}513^{\,2}}{1 - 0{,}513^{\,2}}(12-2) = 3{,}57 \ .$$

Сравнивая табличное $F_{табл}$=4,96 и фактическое $F_{факт} = 3{,}57$ значения, отмечаем, что

$$F_{факт} < F_{табл},$$

что указывает на необходимость принять выдвинутую гипотезу H_o.

2 е) Оценку статистической значимости параметров регрессии проведем с помощью t – статистики Стьюдента и путем расчета доверительного интервала для каждого из показателей.

Выдвигаем гипотезу H_0 о статистически незначимом отличии показателей регрессии от нуля: $\alpha = \beta = r_{xy} = 0$.

Табличное значение t – статистики $t_{табл}$ для числа степеней свободы

$$df = n - 2 = 12 - 2 = 10$$

при заданном уровне значимости $\gamma = 0{,}05$ составляет 2,23.

Определим величину случайных ошибок

$$m_\alpha = \frac{\sigma_{ост} \cdot \sqrt{\sum_{i=1}^{n} x^2}}{n\sigma_x} = \frac{8{,}95 \cdot \sqrt{93742}}{12 \cdot 11{,}24} = 20{,}311 ,$$

$$m_\beta = \frac{\sigma_{ост}}{\sigma_x \sqrt{n}} = \frac{8{,}95}{11{,}24\sqrt{12}} = 0{,}23 ,$$

$$m_{r_{xy}} = \sqrt{\frac{1 - r_{xy}^2}{n-2}} = \sqrt{\frac{1-0{,}26}{12-2}} = 0{,}271 .$$

Найдем соответствующие фактические значения t – критерия Стьюдента

$$t_\beta = \frac{\beta}{m_\beta} = \frac{0{,}434}{0{,}230} = 1{,}889 , \; t_\alpha = \frac{\alpha}{m_\alpha} = \frac{75{,}280}{20{,}311} = 3{,}706 ,$$

$$t_r = \frac{r_{xy}}{m_{r_{xy}}} = \frac{0{,}513}{0{,}271} = 1{,}889 .$$

Фактические значения t – статистики не превосходят табличное значение $t_{табл}$= 2,23

$$t_\beta = 1{,}889 < t_{табл} , \quad t_r = 1{,}889 < t_{табл}$$

поэтому гипотеза H_0 о статистически незначимом отличии показателей регрессии от нуля отклоняется, т.е. параметры статистически не значимы.

Для расчета доверительных интервалов для параметров α и β определим их предельные ошибки

$$\Delta_\alpha = t_{табл} m_\alpha = 2{,}23 \cdot 20{,}311 = 45{,}255 ,$$

$$\Delta_\beta = t_{табл} m_\beta = 2{,}23 \cdot 0{,}230 = 0{,}512 .$$

Доверительные интервалы

для параметра α: (30,025; 120,535),
для параметра β: (-0,078; 0,946).

С вероятностью

$$p = 1 - \gamma = 1 - 0{,}05 = 0{,}95$$

можно утверждать, что параметр β, принимает нулевое значение, т.е. является статистически незначимым.

3. Проверим результаты, полученные в п. 2 с помощью **ППП *Excel***.

Параметры парной регрессии вида $y = \alpha + \beta x$ определяет встроенная статистическая функция **ЛИНЕЙН**. Порядок вычисления следующий:

1) введем исходные данные, содержащий анализируемые данные;

2) выделим область пустых ячеек 5x2 (5 строк, 2 столбца) для вывода результатов регрессионной статистики;

3) выберем **Мастер функций**

4) в окне Категория (рис. 2) выберем **Статистические**, в окне Функция – **ЛИНЕЙН**. Щелкнем по кнопке **ОК** (в результате появится диалоговое окно ввода аргументов функции **ЛИНЕЙН** (рис. 3));

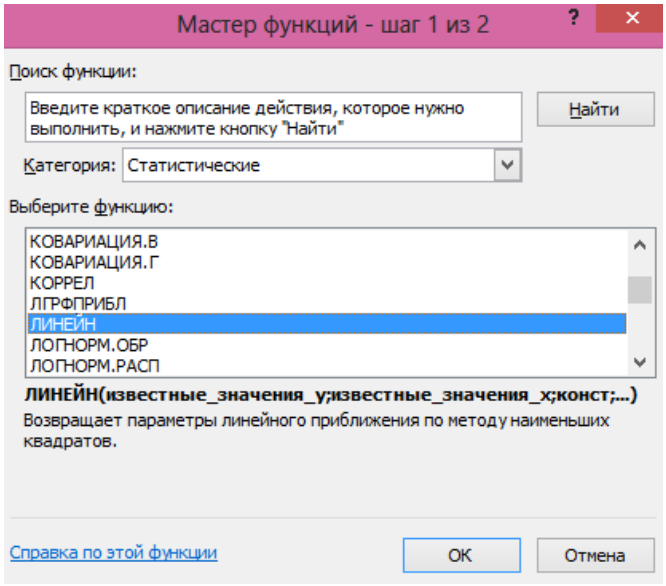

Рисунок 2. Диалоговое окно «Мастер функций»

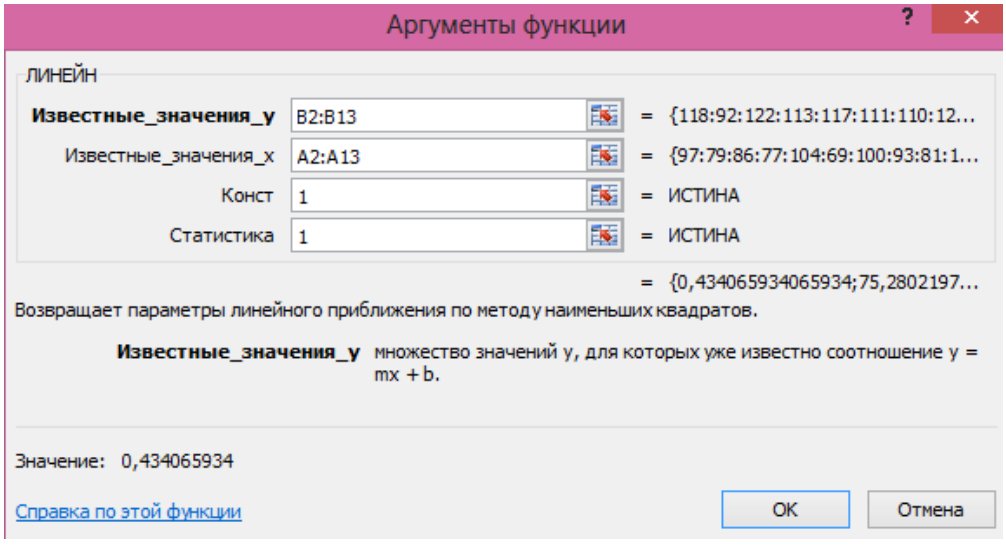

Рисунок 3. Диалоговое окно ввода аргументов функции **ЛИНЕЙН**

5) заполним аргументы функции (рис. 3):

6) в левой верхней ячейке выделенной области появится первый элемент итоговой таблицы. Чтобы раскрыть всю таблицу. Нажмите клавишу <F2>, а затем – на комбинацию клавиш <CTRL>+<SHIFT>+<ENTER>.

Дополнительная регрессионная статистика будет выводиться в порядке, указанном в следующей схеме (табл. 4)

Таблица 4

Значение коэффициента β	Значение коэффициента α
Среднее квадратическое отклонение β	Среднее квадратическое отклонение α
Коэффициент детерминации R^2	Среднеквадратическое отклонение y
F – статистика	Число степеней свободы
Регрессионная сумма квадратов	Остаточная сумма квадратов

Для данных рассматриваемого примера результат вычисления функции **ЛИНЕЙН** представлен на рис. 4

	A	B	C	D	E
1	x	y		0,434066	75,28022
2	97	118		0,229798	20,31058
3	79	92		0,262969	8,949338
4	86	122		3,567958	10
5	77	113		285,7601	800,9066
6	104	117			
7	69	111			
8	100	110			
9	93	128			
10	81	115			
11	102	120			
12	74	98			
13	90	116			
14					

Рисунок 4. Результат вычисления функции **ЛИНЕЙН**

Замечание

С помощью инструмента анализа данных Регрессия, помимо результатов регрессионной статистики, дисперсионного анализа и доверительных интервалов, можно получить остатки и графики подбора линии регрессии, остатков и нормальной вероятности. Порядок действий следующий:

1) проверим доступ к пакету анализа. (рис. 5);

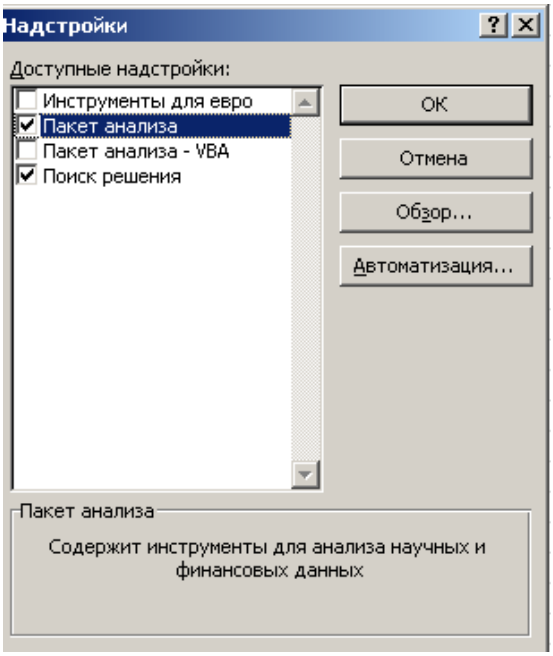

Рисунок 5. Подключение надстройки **Пакет анализа**

2) в главном меню выберем **Регрессия** (рис. 6). Щелкните по кнопке **ОК**;

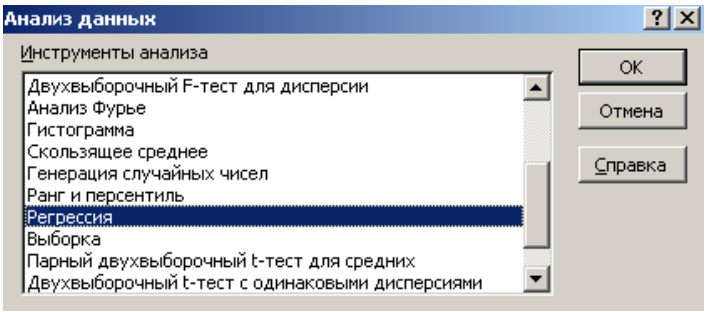

Рисунок 6. Диалоговое окно **Анализ данных**

3) заполним диалоговое окно ввода данных и параметров вывода (рис. 7):

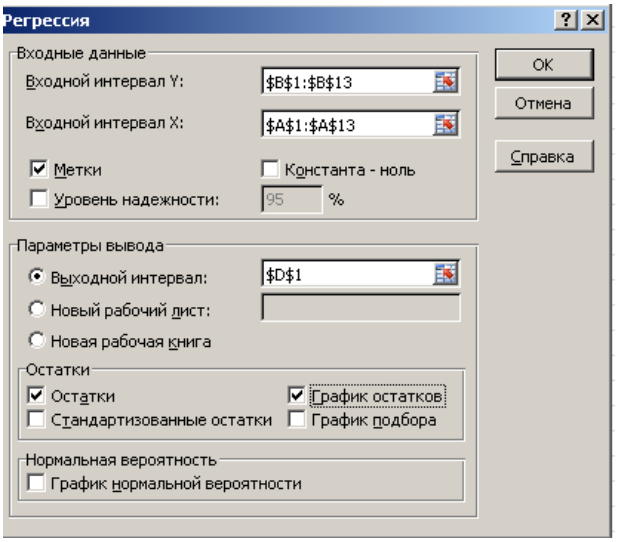

Рисунок 7. Диалоговое окно ввода параметров инструмента **Регрессия**

Результаты регрессионного анализа для данных рассматриваемой задачи представлены на рис. 8

	A	B	C	D	E	F	G	H	I	J	K	L
1	x	y		ВЫВОД ИТОГОВ								
2	97	118										
3	79	92		*Регрессионная статистика*								
4	86	122		Множественный R	0,512805414							
5	77	113		R-квадрат	0,262969393							
6	104	117		Нормированный R-квадрат	0,189266332							
7	69	111		Стандартная ошибка	8,949338486							
8	100	110		Наблюдения	12							
9	93	128										
10	81	115		Дисперсионный анализ								
11	102	120			*df*	*SS*	*MS*	*F*	*Значимость F*			
12	74	98		Регрессия	1	285,7600733	285,7600733	3,567957561	0,088221122			
13	90	116		Остаток	10	800,9065934	80,09065934					
14				Итого	11	1086,666667						
15												
16					*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*ижние 95,0%*	*Верхние 95,0%*
17				Y-пересечение	75,28021978	20,31058105	3,706453281	0,00406501	30,02542503	120,5350145	30,02543	120,535015
18				x	0,434065934	0,229797798	1,888903798	0,08822112	-0,077955469	0,946087337	-0,07796	0,94608734
19												
20												
21												
22				ВЫВОД ОСТАТКА								
23												
24				*Наблюдение*		*Предсказанное*	*Остатки*					
25					1	117,3846154	0,615384615					
26					2	109,5714286	-17,57142857					
27					3	112,6098901	9,39010989					
28					4	108,7032967	4,296703297					
29					5	120,4230769	-3,423076923					
30					6	105,2307692	5,769230769					
31					7	118,6868132	-8,686813187					
32					8	115,6483516	12,35164835					
33					9	110,4395604	4,56043956					
34					10	119,5549451	0,445054945					
35					11	107,4010989	-9,401098901					
36					12	114,3461538	1,653846154					

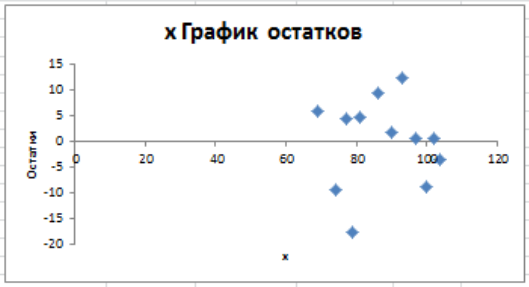

Рисунок 8. Результаты применения инструмента **Регрессия**

Сравнивая полученные вручную и с помощью **ППП *Excel***. данные, убеждаемся в правильности выполненных действий.

4. Построению показательной модели

$$y = \alpha \cdot \beta^{x} \tag{2}$$

предшествует процедура линеаризации переменных.

Прологарифмируем обе части уравнения (2), получим

$$\ln y = \ln \alpha + x \cdot \ln \beta \,. \tag{3}$$

Введем обозначения

$$Y = \ln y \,,\; C = \ln \alpha \,,\; B = \ln \beta \,.$$

Тогда уравнение (3) запишется в виде

$$Y = C + B \cdot x \,. \tag{4}$$

Параметры полученной линейной модели (4) рассчитываем аналогично тому, как это было сделано выше. Используем данные расчетной таблицы 5

Таблица 5

	x	x2	y	xy	y2	$y-\bar{y}$	$x-\bar{x}$	$(y-\bar{y})^2$	$(x-\bar{x})^2$	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$	A(%)
1	97	9409	4,771	462,76	22,76	4,67	9,333	21,778	87,11	117,19	0,81	0,66	0,007
2	79	6241	4,522	357,22	20,45	-21,33	-8,667	455,111	75,11	109,08	-17,08	291,60	0,186
3	86	7396	4,804	413,15	23,08	8,67	-1,667	75,111	2,78	112,16	9,84	96,78	0,081
4	77	5929	4,727	364,01	22,35	-0,33	-10,667	0,111	113,78	108,21	4,79	22,94	0,042
5	104	10816	4,762	495,27	22,68	3,67	16,333	13,444	266,78	120,50	-3,50	12,28	0,030
6	69	4761	4,710	324,96	22,18	-2,33	-18,667	5,444	348,44	104,81	6,19	38,26	0,056
7	100	10000	4,700	470,05	22,09	-3,33	12,333	11,111	152,11	118,60	-8,60	73,94	0,078
8	93	8649	4,852	451,24	23,54	14,67	5,333	215,111	28,44	115,34	12,66	160,39	0,099
9	81	6561	4,745	384,34	22,51	1,67	-6,667	2,778	44,44	109,95	5,05	25,51	0,044
10	102	10404	4,787	488,32	22,92	6,67	14,333	44,444	205,44	119,55	0,45	0,20	0,004
11	74	5476	4,585	339,29	21,02	-15,33	-13,667	235,111	186,78	106,92	-8,92	79,64	0,091
12	90	8100	4,754	427,82	22,60	2,67	2,333	7,111	5,44	113,96	2,04	4,14	0,018
Среднее	87,67	7811,83	4,73	414,87	22,35								6,12%
Сумма	1052,00	93742,00	56,72	4978,42	268,18			1086,67	1516,67			806,34	0,735
σ	11,24		0,09										
Дисперсия	126,39		0,01										

Построим линейное уравнение парной регрессии Y по x. Используя данные таблицы 5, имеем

$$B = \frac{\overline{Y \cdot x} - \bar{Y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2} = \frac{414{,}87 - 4{,}73 \cdot 87{,}67}{7811{,}83 - 87{,}67^2} = 0{,}004 ,$$

$$C = \bar{Y} - B \cdot \bar{x} = 4{,}73 - 0{,}004 \cdot 87{,}67 = 4{,}377 .$$

Получим линейное уравнение регрессии

$$Y = 4{,}377 + 0{,}004 \cdot x .$$

Тесноту полученной линейной модели характеризует линейный коэффициент парной корреляции

$$r_{xy} = B \frac{\sigma_x}{\sigma_Y} = 0{,}004 \cdot \frac{11{,}24}{0{,}09} = 0{,}510 .$$

Коэффициент детерминации при этом равен σ^2

$$R^2 = r_{xy}^2 = 0{,}26 .$$

Это означает, что почти 59% вариации фактора Y объясняется вариацией фактора x.

Средняя ошибка линейной аппроксимации составляет

$$\bar{A} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100\% = \frac{1}{n} \sum_{i=1}^{n} |A_i| \cdot 100\% = \frac{0{,}735}{12}\% = 6{,}12\% .$$

Проведя потенцирование уравнения (5), получим искомую нелинейную (показательную) модель

$$\hat{y} = 79{,}614 \cdot 1{,}004^{x}$$

Результаты вычисления параметров показательной кривой можно проверить с помощью **ППП *Excel***, для чего используем встроенную статистическую функцию **ЛГРФПРИБЛ**. Порядок вычисления аналогичен применению функции **ЛИНЕЙН**.

В результате применения функции **ЛГРФПРИБЛ** дополнительная регрессионная статистика будет выводиться в порядке, указанном выше (табл. 4), причем в первой строке таблицы (рис. 9) функция **ЛГРФПРИБЛ** возвращает коэффициенты показательной модели (2), остальные параметры соответствуют линейной модели (4) (рис. 9).

	A	B	C	D	E
1	x	y		1,003993	79,61407
2	97	118		0,002127	0,18797
3	79	92		0,259916	0,082824
4	86	122		3,511975	10
5	77	113		0,024092	0,068598
6	104	117			
7	69	111			
8	100	110			
9	93	128			
10	81	115			
11	102	120			
12	74	98			
13	90	116			

Рисунок 9. Результат вычисления функции **ЛГРФПРИБЛ**

Для расчета индекса корреляции ρ_{xy} нелинейной регрессии воспользуемся вспомогательной таблицей 6

Таблица 6

	x	y	$\hat{y}$	$(y-\hat{y})^2$	$(y-\bar{y})^2$	$(x-\bar{x})^2$
1	97	118	117,19	0,66	21,778	9,333
2	79	92	109,08	291,60	455,111	-8,667
3	86	122	112,16	96,78	75,111	-1,667
4	77	113	108,21	22,94	0,111	-10,667
5	104	117	120,50	12,28	13,444	16,333
6	69	111	104,81	38,26	5,444	-18,667
7	100	110	118,60	73,94	11,111	12,333
8	93	128	115,34	160,39	215,111	5,333
9	81	115	109,95	25,51	2,778	-6,667
10	102	120	119,55	0,20	44,444	14,333
11	74	98	106,92	79,64	235,111	-13,667
12	90	116	113,96	4,14	7,111	2,333
Среднее	87,67	113,3				
Сумма	1052	93742		806,34	1086,67	1516,667

$$\rho_{xy} = \sqrt{1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \sqrt{1 - \frac{806{,}34}{1086{,}67}} = 0{,}51 .$$

Найдем коэффициент детерминации

$$R^2 = \rho_{xy}^2 = 0{,}51^2 = 0{,}26 .$$

Это означает, что 26% вариации заработной платы y объясняется вариацией фактора x – среднедушевого прожиточного минимума.

Рассчитаем фактическое значение F – критерия при заданном уровне значимости $\gamma = 0{,}05$

$$F_{факт} = \frac{\rho_{xy}^2}{1 - \rho_{xy}^2}(n-2) = \frac{0{,}26}{1 - 0{,}26}(12-2) = 3{,}51 .$$

Сравнивая табличное $F_{табл}$=4,96 и фактическое $F_{факт} = 3{,}51$ значения, отмечаем, что

$$F_{факт} < F_{табл},$$

что указывает на необходимость принять гипотезу H_o о статистически незначимых параметрах уравнения (6).

5. Так как коэффициенты детерминации, соответствующие линейной и показательной моделям практически равны (около 26% вариации заработной платы y объясняется вариацией фактора x – среднедушевого прожиточного минимума в обеих моделях), то нет весомых оснований отдать предпочтение какой либо модели.

Обе модели не значимы в целом, поэтому прогноз мы не можем строить ни по одной модели.