Расчетно-графическая работа №1.
Парная линейная регрессия.

Исходные данные.

Дана выборка: себестоимость Y (руб.) одного экземпляра книги в зависимости от тиража X (тыс.шт.)

X	0,8	1,9	2,7	4,1	5,2	7,5	8,3	9,8
Y	87	62	59	45	40	35	20	17

Задания.

1. Выполняя непосредственные вычисления, найдите эмпирическое уравнение линейной регрессии Y от X.
2. Выполните задание из пункта 1, используя встроенные функции Microsoft Excel. Сравните полученные результаты.
3. Постройте корреляционное поле выборки и график эмпирического уравнения регрессии.
4. Найдите общую и регрессионную суммы квадратов отклонений, рассчитайте по ним коэффициент детерминации и проверьте значимость модели по критерию Фишера при уровне значимости $\alpha = 0.05$.
5. Найдите стандартную ошибку регрессии и 95%-ные доверительные интервалы для коэффициентов регрессии.
6. Найдите все величины из предыдущих пунктов при помощи инструмента **Регрессия** раздела **Анализ данных.**

Выполнение.

1. Теоретическое уравнение линейной регрессии имеет вид

$$Y = \alpha + \beta X + \varepsilon.$$

Найти эмпирическое уравнение линейной регрессии – это значит получить оценки a и b для параметров α и β по данным выборки. Эмпирическое уравнение будет иметь вид

$$y_i^* = a + bx_i.$$

Замечание. *Так как нам доступна лишь некоторая выборка из генеральной совокупности, то точные значения параметров* α *и* β *определить невозможно, и мы можем найти только оценки для* α *и* β*, которые, естественно, будут отличаться от истинных значений параметров.*

Для отыскания a и b воспользуемся формулами

$$a = \overline{y} - b\overline{x}, \qquad b = \frac{Cov(x,y)}{D_x}.$$

Соответствующие расчеты представлены на рисунках 1 и 2.

	A	B	C	D	E	F	G	H	I
1	№ набл.	X	Y	X^2	Y^2	XY		Характеристики	
2	1	0,8	87	0,64	7569	69,6		$X_{ср}$	5,0375
3	2	1,9	62	3,61	3844	117,8		$Y_{ср}$	45,625
4	3	2,7	59	7,29	3481	159,3		D_x	9,194844
5	4	4,1	45	16,81	2025	184,5		D_y	472,4844
6	5	5,2	40	27,04	1600	208		S_x	3,0323
7	6	7,5	35	56,25	1225	262,5		S_y	21,73671
8	7	8,3	20	68,89	400	166		cov(x,y)	-63,0484
9	8	9,8	17	96,04	289	166,6		r_{xy}	-0,95655
10	Сумма	40,3	365	276,57	20433	1334,3			
11									
12				Коэфф. Регрессии					
13				a	80,166805				
14				b	-6,856934				

Рис. 1. Вычисление эмпирических коэффициентов регрессии

	A	B	C	D	E	F	G	H	I
1	№ набл.	X	Y	X^2	Y^2	XY		Характеристики	
2	1	0,8	87	=B2^2	=C2^2	=B2*C2		$X_{ср}$	=B10/A9
3	2	1,9	62	=B3^2	=C3^2	=B3*C3		$Y_{ср}$	=C10/A9
4	3	2,7	59	=B4^2	=C4^2	=B4*C4		D_x	=D10/A9-I2^2
5	4	4,1	45	=B5^2	=C5^2	=B5*C5		D_y	=E10/A9-I3^2
6	5	5,2	40	=B6^2	=C6^2	=B6*C6		S_x	=КОРЕНЬ(I4)
7	6	7,5	35	=B7^2	=C7^2	=B7*C7		S_y	=КОРЕНЬ(I5)
8	7	8,3	20	=B8^2	=C8^2	=B8*C8		cov(x,y)	=F10/A9-I2*I3
9	8	9,8	17	=B9^2	=C9^2	=B9*C9		r_{xy}	=I8/(I6*I7)
10	Сумма	=СУММ(B2:B9)	=СУММ(C2:C9)	=СУММ(D2:D9)	=СУММ(E2:E9)	=СУММ(F2:F9)			
11									
12				Коэфф. Регрессии					
13				a	=I3-E14*I2				
14				b	=I8/I4				

Рис. 2. Формулы для вычисления эмпирических коэффициентов регрессии

Таким образом, эмпирическое уравнение линейной регрессии для данной выборки будет иметь вид

$$y_i^* = 80{,}17 - 6{,}86x_i.$$

2. Воспользуемся встроенными функциями НАКЛОН и ОТРЕЗОК, которые возвращают значения коэффициентов a и b, соответственно. Результаты вычислений и формулы представлены на рисунках 3 и 4.

	A	B	C	D	E	F
1	№ набл.	X	Y		Коэфф. регрессии	
2	1	0,8	87		a	80,167
3	2	1,9	62		b	-6,857
4	3	2,7	59			
5	4	4,1	45			
6	5	5,2	40			
7	6	7,5	35			
8	7	8,3	20			
9	8	9,8	17			
10						

Рис. 3. Вычисление эмпирических коэффициентов регрессии с помощью встроенных функций

	A	B	C	D	E	F
1	№ набл.	X	Y		Коэфф. регрессии	
2	1	0,8	87		a	=ОТРЕЗОК(C2:C9;B2:B9)
3	2	1,9	62		b	=НАКЛОН(C2:C9;B2:B9)
4	3	2,7	59			
5	4	4,1	45			
6	5	5,2	40			
7	6	7,5	35			
8	7	8,3	20			
9	8	9,8	17			
10						

Рис. 4. Формулы для вычисления эмпирических коэффициентов регрессии с помощью встроенных функций

3. Для того, чтобы построить график эмпирического уравнения регрессии, необходимо найти значения y_i^* для всех значений x_i из выборки. Расчеты представлены на рисунках 5 и 6.

	A	B	C	D
1	№	X	Y	Y*
2	1	0.80	87.00	74.68
3	2	1.90	62.00	67.14
4	3	2.70	59.00	61.65
5	4	4.10	45.00	52.05
6	5	5.20	40.00	44.51
7	6	7.50	35.00	28.74
8	7	8.30	20.00	23.25
9	8	9.80	17.00	12.97
10				
11				
12	Коэфф. регрессии			
13	a	80.17		
14	b	-6.86		
15				

Рис. 5. Вычисление значений эмпирической функции

	A	B	C	D
1	№	X	Y	Y*
2	1	0.8	87	=B$13+B$14*B2
3	2	1.9	62	=B$13+B$14*B3
4	3	2.7	59	=B$13+B$14*B4
5	4	4.1	45	=B$13+B$14*B5
6	5	5.2	40	=B$13+B$14*B6
7	6	7.5	35	=B$13+B$14*B7
8	7	8.3	20	=B$13+B$14*B8
9	8	9.8	17	=B$13+B$14*B9
10				
11				
12		Коэфф. регрессии		
13	a	=ОТРЕЗОК(C2:C9;B2:B9)		
14	b	=НАКЛОН(C2:C9;B2:B9)		

Рис. 6. Формулы для вычисления значений эмпирической функции

Далее, на точечной диаграмме строим два ряда данных: (X,Y) и (X,Y*).

Ряд (X,Y) дает поле корреляции, а ряд (X,Y*) график эмпирической функции регрессии.

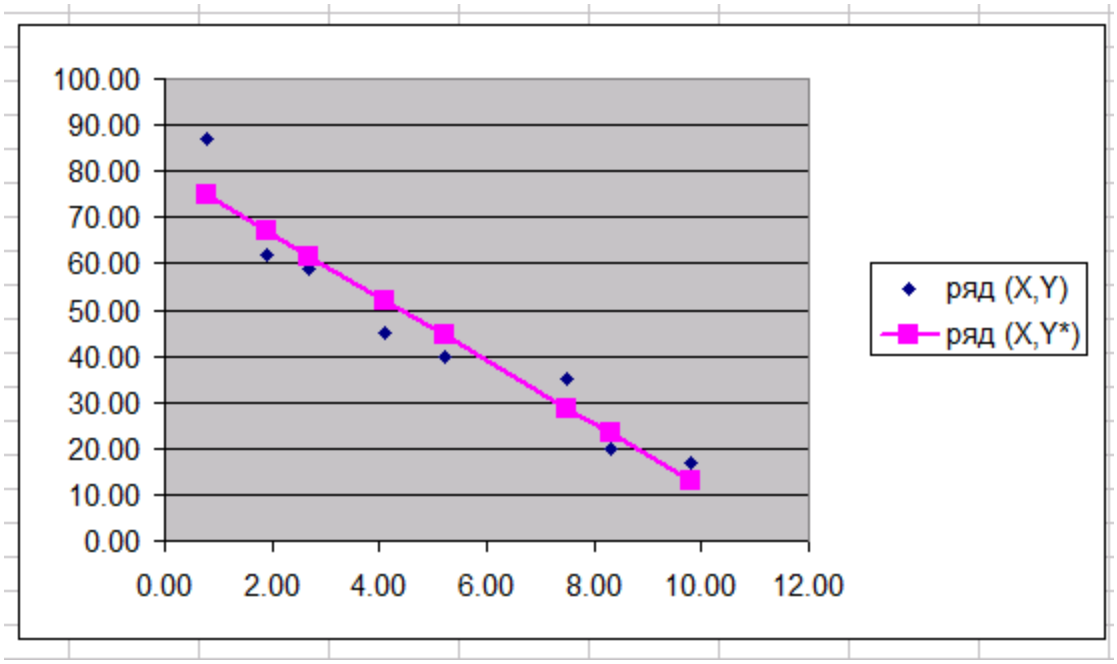

Рис. 7. Поле корреляции и график эмпирической функции регрессии

4. Объясненная (регрессионная) и общая суммы вычисляются по формулам

$$\sum_{рег} = \sum \left(y_i^* - y_{ср} \right)^2, \qquad \sum_{общ} = \sum \left(y_i - y_{ср} \right)^2.$$

Коэффициент детерминации R^2 – это отношение объясненной суммы квадратов к общей, т.е.

$$R^2 = \frac{\sum_{\text{рег}}}{\sum_{\text{общ}}}.$$

Проверка значимости линейной регрессионной модели сводится к проверке нулевой гипотезы $H_0 : \beta = 0$ (модель незначима) при альтернативной гипотезе $H_1 : \beta \neq 0$ (модель значима).

Чтобы проверить H_0 по выборке объема n, необходимо вычислить наблюдаемое значение критерия

$$F_{\text{набл}} = \frac{R^2(n-2)}{1-R^2}.$$

Если H_0 верна, то статистика $F_{\text{набл}}$ имеет распределение Фишера со степенями свободы 1 и $(n-2)$. При помощи функции FРАСПОБР по заданному уровню значимости α и степеням свободы 1 и $(n-2)$ определяем критическую точку $F_{\text{кр}} = F(\alpha, 1, n-2) = \text{FРАСПОБР}(\alpha, 1, n-2)$. Если $F_{\text{набл}} > F_{\text{кр}}$, то нулевая гипотеза отвергается и регрессия считается значимой. В противном случае нет оснований для того чтобы отвергнуть нулевую гипотезу, поэтому полученное уравнение регрессии считается незначимым.

Расчетный лист для вычисления сумм квадратов отклонений, коэффициента детерминации и проверки значимости регрессии представлен на рисунках 8 и 9.

	A	B	C	D	E	F	G
1	**№ набл.**	**X**	**Y**	**Y***	**$(Y-Y_{ср})^2$**	**$(Y^*-Y_{ср})^2$**	
2	1	0,8	87	74,681	1711,891	844,266	
3	2	1,9	62	67,139	268,141	462,836	
4	3	2,7	59	61,653	178,891	256,899	
5	4	4,1	45	52,053	0,391	41,324	
6	5	5,2	40	44,511	31,641	1,242	
7	6	7,5	35	28,740	112,891	285,110	
8	7	8,3	20	23,254	656,641	500,450	
9	8	9,8	17	12,969	819,391	1066,424	
10	**Сумма**				3779,875	3458,552	
11							
12							
13	**Коэфф. регрессии**			R^2	0,915		
14	**a**	80,16681		$F_{\text{набл}}$	64,581		
15	**b**	-6,85693		$F_{\text{кр}}$	5,987		
16							
17	**$Y_{ср}$**	45,625					

Рис. 8. Проверка значимости модели по критерию Фишера (значения)

	A	B	C	D	E	F
1	№ набл.	X	Y	Y*	(Y-$Y_{ср}$)^2	(Y*-$Y_{ср}$)^2
2	1	0,8	87	=B$14+B$15*B2	=(C2-B$17)^2	=(D2-B$17)^2
3	2	1,9	62	=B$14+B$15*B3	=(C3-B$17)^2	=(D3-B$17)^2
4	3	2,7	59	=B$14+B$15*B4	=(C4-B$17)^2	=(D4-B$17)^2
5	4	4,1	45	=B$14+B$15*B5	=(C5-B$17)^2	=(D5-B$17)^2
6	5	5,2	40	=B$14+B$15*B6	=(C6-B$17)^2	=(D6-B$17)^2
7	6	7,5	35	=B$14+B$15*B7	=(C7-B$17)^2	=(D7-B$17)^2
8	7	8,3	20	=B$14+B$15*B8	=(C8-B$17)^2	=(D8-B$17)^2
9	8	9,8	17	=B$14+B$15*B9	=(C9-B$17)^2	=(D9-B$17)^2
10	Сумма				=СУММ(E2:E9)	=СУММ(F2:F9)
11						
12						
13	Коэфф. регрессии			R^2	=F10/E10	
14	a	=ОТРЕЗОК(C2:C9;B2:B9)		$F_{набл}$	=E13*(A9-2)/(1-E13)	
15	b	=НАКЛОН(C2:C9;B2:B9)		$F_{кр}$	=FРАСПОБР(0,05;1;A9-2)	
16						
17	$Y_{ср}$	=СРЗНАЧ(C2:C9)				
18						

Рис. 9. Проверка значимости модели по критерию Фишера (формулы)

Суммы квадратов отклонений $\Sigma_{общ}$ и $\Sigma_{рег}$ вычислены в ячейках E10 и F10 соответственно.

Поскольку $F_{набл}=64,581>F_{кр}=5,987$, то коэффициент детерминации статистически значим, а следовательно статистически значимо и уравнение регрессии.

5. Для выполнения расчетов в этом задании понадобятся следующие формулы.

Стандартная ошибка регрессии

$$S = \sqrt{\frac{\sum_{ост}}{n-2}} = \sqrt{\frac{\sum\left(y_i - y_i^*\right)^2}{n-2}}.$$

Стандартная ошибка оценки b

$$m_b = \sqrt{\frac{S^2}{\sum\left(x_i - \bar{x}\right)^2}}.$$

Доверительный интервал для коэффициента регрессии β при заданном уровне значимости α

$$b \pm t(\alpha, n-2) \cdot m_b,$$

где $t\left(\alpha, n-2\right)$ – критическое значение t-критерия Стьюдента при уровне значимости α и числе степеней свободы $\left(n-2\right)$.

Стандартная ошибка параметра a

$$m_a = \sqrt{S^2 \cdot \frac{\sum x_i^2}{n \cdot \sum \left(x_i - \bar{x}\right)^2}}.$$

Доверительный интервал для свободного члена уравнения регрессии при заданном уровне значимости α определяется по формуле:

$$a \pm t(\alpha, n-2) \cdot m_a.$$

Необходимые расчеты представлены на рисунках 10 и 11.

	A	B	C	D	E	F	G	H	I
1	№	X	Y	X^2	$(X-X_{ср})^2$	Y*	$(Y-Y_{ср})^2$	$(Y^*-Y_{ср})^2$	$(Y-Y^*)^2$
2	1	0,80	87,00	0,64	17,96	74,68	1711,89	844,27	151,75
3	2	1,90	62,00	3,61	9,84	67,14	268,14	462,84	26,41
4	3	2,70	59,00	7,29	5,46	61,65	178,89	256,90	7,04
5	4	4,10	45,00	16,81	0,88	52,05	0,39	41,32	49,75
6	5	5,20	40,00	27,04	0,03	44,51	31,64	1,24	20,35
7	6	7,50	35,00	56,25	6,06	28,74	112,89	285,11	39,19
8	7	8,30	20,00	68,89	10,64	23,25	656,64	500,45	10,59
9	8	9,80	17,00	96,04	22,68	12,97	819,39	1066,42	16,25
10	Сумма			276,57	73,56		3779,88	3458,55	321,32
11									
12	Коэфф. регрессии					$X_{ср}$	5,04		
13	a	80,17				$Y_{ср}$	45,63		
14	b	-6,86							
15						S	7,32		
16						m_b	0,85		
17						m_a	5,02		
18						t(0,05;n-2	2,45		
19									
20					Доверительные интервалы				
21					-8,94	<b<	-4,77		
22					67,89	<a<	92,44		

Рис. 10. Расчет доверительных интервалов

	A	B	C	D	E	F	G	H	I
1	№	X	Y	X^2	$(X-X_{ср})^2$	Y*	$(Y-Y_{ср})^2$	$(Y^*-Y_{ср})^2$	$(Y-Y^*)^2$
2	1	0,8	87	=B2^2	=(B2-G$12)^2	=B$13+B$14*B2	=(C2-G$13)^2	=(F2-G$13)^2	=(C2-F2)^2
3	2	1,9	62	=B3^2	=(B3-G$12)^2	=B$13+B$14*B3	=(C3-G$13)^2	=(F3-G$13)^2	=(C3-F3)^2
4	3	2,7	59	=B4^2	=(B4-G$12)^2	=B$13+B$14*B4	=(C4-G$13)^2	=(F4-G$13)^2	=(C4-F4)^2
5	4	4,1	45	=B5^2	=(B5-G$12)^2	=B$13+B$14*B5	=(C5-G$13)^2	=(F5-G$13)^2	=(C5-F5)^2
6	5	5,2	40	=B6^2	=(B6-G$12)^2	=B$13+B$14*B6	=(C6-G$13)^2	=(F6-G$13)^2	=(C6-F6)^2
7	6	7,5	35	=B7^2	=(B7-G$12)^2	=B$13+B$14*B7	=(C7-G$13)^2	=(F7-G$13)^2	=(C7-F7)^2
8	7	8,3	20	=B8^2	=(B8-G$12)^2	=B$13+B$14*B8	=(C8-G$13)^2	=(F8-G$13)^2	=(C8-F8)^2
9	8	9,8	17	=B9^2	=(B9-G$12)^2	=B$13+B$14*B9	=(C9-G$13)^2	=(F9-G$13)^2	=(C9-F9)^2
10	Сумма			=СУММ(D2:D9)	=СУММ(E2:E9)		=СУММ(G2:G9)	=СУММ(H2:H9)	=СУММ(I2:I9)
11									
12	Коэфф. регрессии					$X_{ср}$	=СРЗНАЧ(B2:B9)		
13	a	=ОТРЕЗОК(C2:C9;B2:B9)				$Y_{ср}$	=СРЗНАЧ(C2:C9)		
14	b	=НАКЛОН(C2:C9;B2:B9)							
15						S	=КОРЕНЬ(I10/(A9-2))		
16						m_b	=КОРЕНЬ(G15^2/E10)		
17						m_a	=КОРЕНЬ((G15^2)*D10/(A9*E10)		
18						t(0,05;n-2)	=СТЬЮДРАСПОБР(0,05;A9-2)		
19									
20					Доверительные интервалы				
21					=B14-G18*G16	<b<	=B14+G18*G16		
22					=B13-G18*G17	<a<	=B13+G18*G17		
23									

Рис. 11. Формулы для расчета доверительных интервалов

Подчеркнем, что доверительные интервалы найдены для уровня значимости α=0,05. То есть это такие интервалы, что *коэффициенты теоретического уравнения регрессии принадлежат им с вероятностью 95%.*

6. Величины, вычисленные в предыдущих пунктах, могут быть получены автоматически при помощи встроенного инструмента Microsoft Excel. Для этого необходимо в разделе **Анализ данных** выбрать инструмент **Регрессия.** В результате появляется диалоговое окно, которое нужно заполнить так, как показано на рисунке 12.

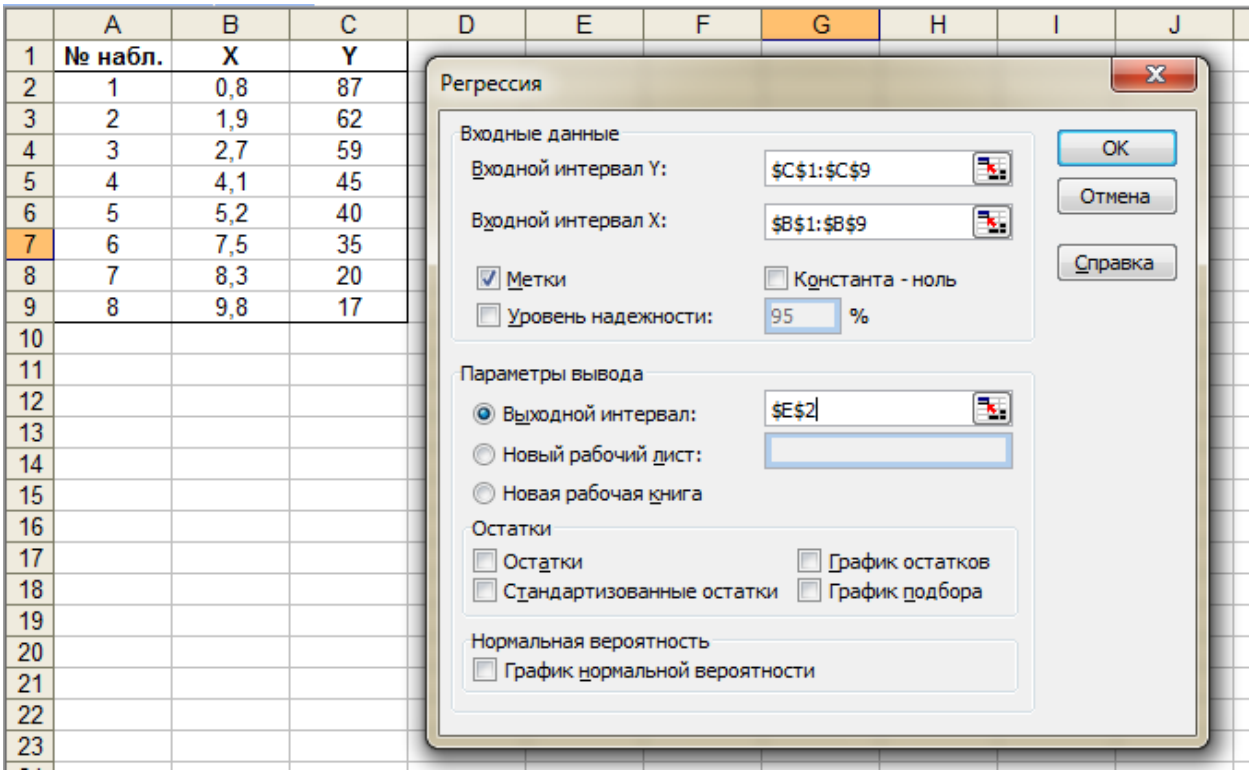

Рис. 12. Заполнение диалогового окна Регрессия

Замечание. Обратите внимание, что интервалы **$C$1:$C$9** и **$B$1:$B$9** содержат заголовки столбцов, а не только числовые данные (для этого ставим флажок **Метки**). Это удобно при отображении результатов проводимого анализа.

После нажатия на **Ок** на экран будут выведены результаты регрессионного анализа (рис. 13).

	A	B	C	D	E	F	G	H	I	J	K
1	№ набл.	X	Y								
2	1	0,8	87		ВЫВОД ИТОГОВ						
3	2	1,9	62								
4	3	2,7	59		*Регрессионная статистика*						
5	4	4,1	45		Множеств	0,956552					
6	5	5,2	40		R-квадрат	0,914991					
7	6	7,5	35		Нормиров	0,900823					
8	7	8,3	20		Стандартн	7,31805					
9	8	9,8	17		Наблюден	8					
10											
11					Дисперсионный анализ						
12						*df*	*SS*	*MS*	*F*	*ачимость F*	
13					Регрессия	1	3458,552	3458,552	64,58081	0,000198	
14					Остаток	6	321,3232	53,55386			
15					Итого	7	3779,875				
16											
17					*Коэффициенты*	*Дартная о*	*статисти*	*P-Значени*	*ижние 95%*	*ерхние 95%*	
18					Y-пересеч	80,16681	5,016904	15,97934	3,81E-06	67,89088	92,44273
19					X	-6,85693	0,853254	-8,03622	0,000198	-8,94477	-4,7691
20											

Рис. 13. Результаты регрессионного анализа

Поскольку терминология, используемая в Microsoft Excel, несколько отличается от той, которая принята в российской научной литературе, то укажем соответствие между введенными ранее величинами и значениями на рисунке 13.

Величина	**Ячейка**
Коэффициент корреляции r_{xy}	**F5**
Коэффициент детерминации R^2	**F6**
Стандартная ошибка регрессии S	**F8**
Регрессионная сумма $\Sigma_{рег}$	**G13**
Остаточная сумма $\Sigma_{ост}$	**G14**
Общая сумма $\Sigma_{общ}$	**G15**
Наблюдаемое значение F-статистики $F_{набл}$	**I13**
Коэффициент регрессии a	**F18**
Коэффициент регрессии b	**F19**
Стандартная ошибка m_a	**G18**
Стандартная ошибка m_b	**G19**
Доверительный интервал для α	**J18-K18**
Доверительный интервал для β	**J19-K19**