

## **Практическое задание № 1. Метод наименьших квадратов. Классическая линейная модель парной регрессии**

### **Задача:**

Вы располагаете файлом данных формата Excel, который содержит: страницу с социально-экономическими показателями, полученными в результате опроса, проведённого Всемирным банком (WCY) в 76 странах мира;

### **Задания:**

1. Выбрать из массива данных требуемые для Вашего исследования данные.
2. Вычислить основные описательные статистики для всех исходных переменных Вашего варианта (включая Y).
3. Вычислить попарные коэффициенты корреляции зависимой переменной с каждой из независимых переменных.
4. Построить попарные графики разброса зависимой переменной со всеми независимыми переменными. Указать на графике линию регрессии. Сделать предположения о степени и характере зависимости.
5. Выбрать один из регрессоров (тот, по которому, как Вам кажется, получится наилучший прогноз) и предположить, какие значения коэффициентов регрессии дадут наименьшую сумму квадратов отклонений.
6. С помощью надстройки «Поиск решения» пакета Excel найти значения коэффициентов, минимизирующие сумму квадратов остатков.

7. Дать интерпретацию найденным оценкам коэффициентов регрессии.

8. Вычислить TSS, RSS, ESS. Вычислить значение коэффициента детерминации  $R^2$ . Сделать предположение о качестве модели.

9. Построить график остатков модели. Вычислить среднее значение остатков.

10. Провести оценку коэффициентов модели парной регрессии с помощью стандартной надстройки Excel «Анализ данных». Убедиться в идентичности полученных результатов.

11. Вычислить прогноз значения зависимой переменной  $\hat{y}_{77}$  для страны 77 «Оз», для которой  $x_{77} = \bar{x}$ , где  $\bar{x}$  - среднее значение данной переменной по Вашей выборке.

## Выполнение практического задания № 1.

1. Выберем из предложенного массива требуемые данные.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1		ВВП на душу населения, в текущих ценах (тыс.долл. СПА)			Интернет-пользователи - Интернет-пользователи на 100 человек					Индекс гравитации демократии						Общие расходы на правоохранение (% от ВВП)			Уровень безработицы			Ожидаемая продолжительность жизни при рождении
2	y				x1		x2				x3		x4			x5						
3	Австралия	67,646	20	242	72,8	79	2,06	1,76	50,9	9,22	406	3,75	312	324	6,13	9,357	4,01	5,225	12,7	30,4	82,537	
4	Азербайджан	7,394	3,4	11,7	80,2	54,2	1,44	1,01	38	3,15	922	7,7	15,7	37	22,9	5,367	39,2	6,048	7,27	93,4	70,896	4
5	Албания	4,248	1,4	0,13	59	54,656	1,04	2,03	44,8	5,67	16,8	7,47	6,71	3,78	22,6	5,589	14,5	13,4	0,44	26,6	77,968	1
6	Алжир	5,584	3,3	39,6	67,4	15,228	7,45	8,9	52,6	3,83	1532	0,72	59,4	76,4	3,15	6,143	30,7	11	0,13	97,1	75,027	4
7	Аргентина	14,357	4,7	67,3	80	55,8	18,7	10	56,5	6,84	551	2,54	88,2	98,2	14,5	5,019	13,8	7,2	7,7	6,27	76,457	0
8	Армения	3,566	3,9	0,94	51,4	37,5	5,35	2,56	41,3	4,09	0	4,68	4,92	2,74	15,8	4,482	14,1	17,3	2,65	7,93	74,886	3
9	Бахрейн	23,063	25	12,7	53,1	88	2,23	2,76	31,4	2,53	41	2,9	13,9	22,9	2,08	4,366	6,97	3,9	0,36	64,1	76,715	3
10	Беларусь	6,722	7	33	46,3	46,91	75,4	59,2	43	3,04	30	2,3	47,7	46,5	27,2	5,008	4,17	0,619	2,88	37,6	71,464	1
11	Бельгия	44,734	13	24,2	67,2	80,72	2,04	2,84	54,2	8,05	0	1,34	432	429	26,5	10,54	3,37	7,65	11,4	11,5	80,984	1
12	Болгария	7,333	7,4	35,1	53,2	51,9	1,56	2,96	51,9	6,72	1	3,34	35,2	35,6	30,6	7,106	8,97	12,38	7,75	16,2	74,322	1
13	Боливия	2,645	1,4	4,41	76,1	35,34	7,1	4,59	63,7	5,84	41,6	3,91	8,78	9,12	4,02	5,557	43,6	3,229	9,22	55	68,743	1
14	Босния и Герцеговина	4,495	7,1	16,1	57	52,78	0,88	1,3	40,7	5,11	0	2,28	10,4	5,98	19,6	9,94	7,77	28	2,46	8,8	76,634	1
15	Бразилия	12,157	2,4	63,4	73,1	48,56	7,82	5,4	44,7	7,12	2061	3,52	302	293	8,69	8,261	20,6	5,483	10,5	11	74,748	1
16	Великобритания	41,295	7,8	197	60,9	87,48	1,63	2,82	55,1	8,21	18	1,78	839	797	25,7	9,411	4,27	7,975	21,7	13,9	80,849	2
17	Венгрия	12,82	5,2	17	69	70,58	3,5	5,67	47,9	6,96	12,9	8,35	114	122	48,6	7,741	4,67	11,07	18,1	3,9	75,313	1
18	Венесуэла	12,772	6,2	53,2	76,4	40,05	14,1	21,1	52,4	5,15	2500	1,31	69,4	95,7	3,06	4,802	14	8,061	0	98,8	74,387	

Рисунок 1 – Представление данных

Удаляем лишние для нашего варианта столбцы и дополнительные страницы; получаем файл, содержащий 6 столбцов по 76 строк в каждом.

2. Вычислим основные описательные статистики для каждой из 6 имеющихся переменных. Слева указаны стандартные команды русифицированного Excel, в правых – вычисленные значения для переменных  $x_4$  и  $x_5$ .

A	B	C	D	E	F	G	H
	ВВП на одну населения (тыс. долл. США)	Индиаг- популяции на 100 человек	Индекс развития демократии	Общие расходы на здравоохранение	Уровень безработицы	Ожидаемая продолж. жизни при рождении	
1							
2	y	x1	x2	x3	x4	x5	
66	Франция	40,838	81,44	7,88	11,439	9,758	82,359
67	Хорватия	13,236	61,94	6,93	7,797	15,225	77,495
68	Чехия	19,641	73,43	8,19	7,548	6,978	78,775
69	Чили	15,253	61,418	7,54	7,238	6,432	81,956
70	Швейцария	83,209	85,2	9,09	11,589	2,905	83,133
71	Швеция	57,134	93,18	9,73	11,802	7,967	82,347
72	Шри-Ланка	3,351	18,285	5,75	3,21	4	75,049
73	Эквадор	5,702	35,135	5,78	6,479	4,121	76,121
74	Эстония	17,491	78,39	7,61	6,364	10,023	77,012
75	Южная Африка	7,59	41	7,79	8,795	24,875	57,658
76	Южная Корея	24,454	84,07	8,13	7,013	3,225	82,128
77	Ямайка	5,446	33,79	7,39	5,657	13,925	75,82
78	Япония	46,701	79,496	8,08	10,17	4,325	83,684
79							
80	Описательные статистики:						
81	Выборочное среднее	=СРЗНАЧ(В3:В78)				8,8797	76,6774
82	Выборочная дисперсия	=ДИСП.В(В3:В78)				36,309	22,8496
83	Стандартное отклонение	=СТАНДОТКЛОН.В(В3:В78)				6,0257	4,78012
84	Минимальное	=МИН(В3:В78)				0,619	57,658
85	Первый квартиль (Q1)	=КВАРТИЛЬ.ВКЛ(В3:В78;1)				5,454	74,337
86	Медиана (Q2)	=КВАРТИЛЬ.ВКЛ(В3:В78;2)				7,2625	76,265
87	Третий квартиль (Q3)	=КВАРТИЛЬ.ВКЛ(В3:В78;3)				10,756	81,0175
88	Максимальное значение	=МАКС(В3:В78)				31	83,684
89	Коэффициент асимметрии	=СКОС(В3:В78)				1,7607	-0,8855
90	Островершинность	=ЭКСПЕЦС(В3:В78)				3,4434	1,98083
91							

Рисунок 2 – Описательные статистики

3. Вычислим попарные коэффициенты корреляции переменной  $y$  с каждым из  $x$ .

Режимы просмотра книги		Показать					Масштаб	
СУММ								
1								
2	y	x1	x2	x3	x4	x5		
64	Филиппины	2,605	36,235	6,3	4,458	6,975	68,34	
65	Финляндия	47,416	89,88	9,06	9,298	7,742	81,006	
66	Франция	40,838	81,44	7,88	11,439	9,758	82,359	
67	Хорватия	13,236	61,94	6,93	7,797	15,225	77,495	
68	Чехия	19,641	73,43	8,19	7,548	6,978	78,775	
69	Чили	15,253	61,418	7,54	7,238	6,432	81,956	
70	Швейцария	83,209	85,2	9,09	11,589	2,905	83,133	
71	Швеция	57,134	93,18	9,73	11,802	7,967	82,347	
72	Шри-Ланка	3,351	18,285	5,75	3,21	4	75,049	
73	Эквадор	5,702	35,135	5,78	6,479	4,121	76,121	
74	Эстония	17,491	78,39	7,61	6,364	10,023	77,012	
75	Южная Африка	7,59	41	7,79	8,795	24,875	57,658	
76	Южная Корея	24,454	84,07	8,13	7,013	3,225	82,128	
77	Ямайка	5,446	33,79	7,39	5,657	13,925	75,82	
78	Япония	46,701	79,496	8,08	10,17	4,325	83,684	
79								
80								
81		=КОРРЕЛ(\$B3:\$B78;C3:C78)				0,69519		
82								
83								

Рисунок 3 – Расчет коэффициента корреляции

Напомним, что символ «\$», вставленный перед названием столбца, фиксирует его. Коэффициент корреляции принимает значения от -1 до 1 и показывает меру линейной связи между переменными. Так, значение выше 0,9 характерно для сильно зависимых величин, а независимые переменные имеют близкий к 0 коэффициент корреляции.

4. Построим попарные графики разброса. Выберем пункт меню «Вставка» затем «Точечная» и в пункте меню «Выбрать данные», затем «Изменить» указать требуемый диапазон. По построенным диаграммам рассеяния можно сделать вывод, что, например, переменные  $y$  и  $x_1$  имеют прямую зависимость средней силы, а зависимость между  $y$  и  $x_4$  визуально установить затруднительно.

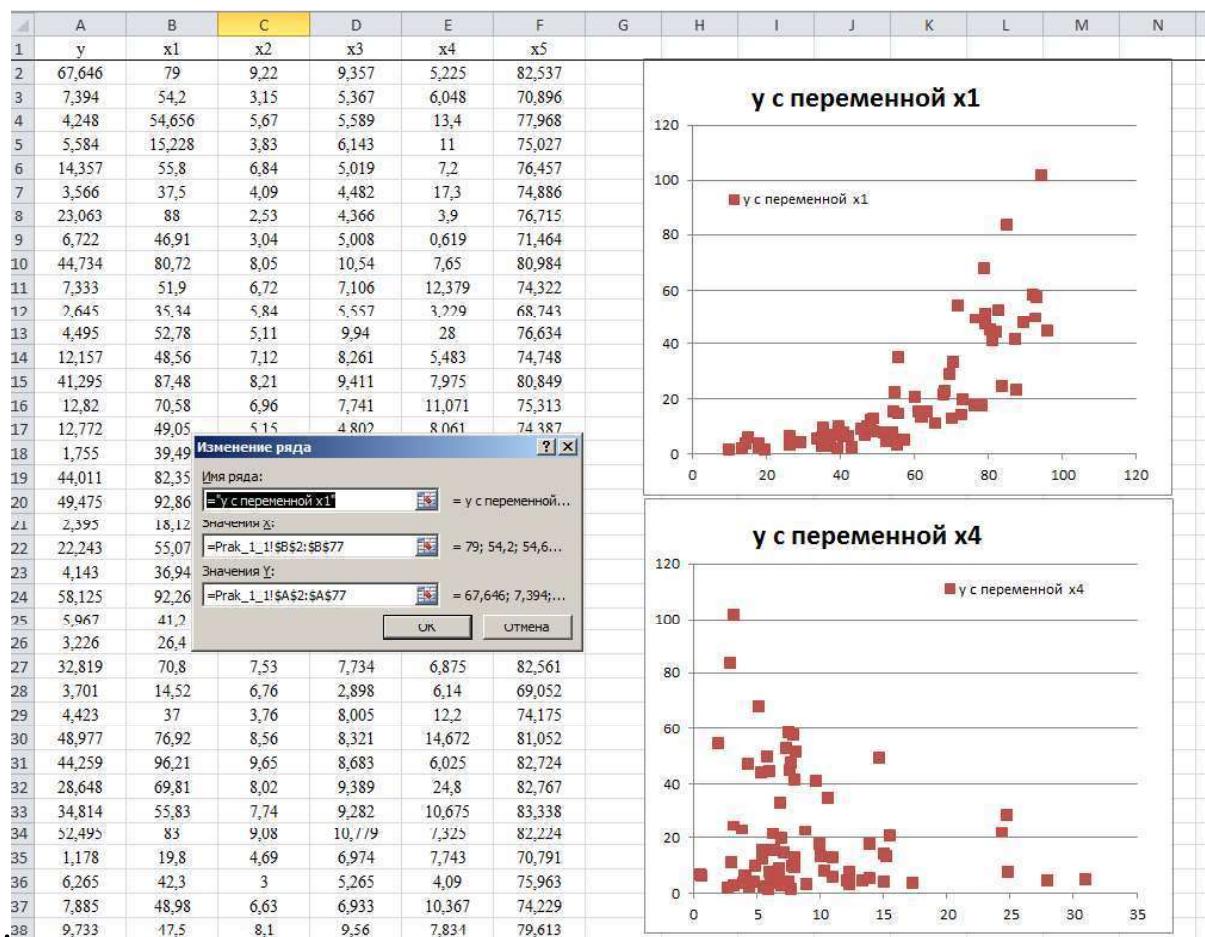


Рисунок 4 – Попарные графики разброса

5. Далее нам предстоит вычислить такие значения  $\hat{\Theta}_0, \hat{\Theta}_1$ , которые минимизируют сумму квадратов отклонений наблюдаемой зависимой переменной от предсказанных значений этой величины. Таким образом, пусть столбцы А и В содержат истинные значения зависимой и независимой переменных. Ячейки К2 и L2 будут содержать требуемые значения  $\hat{\Theta}_0, \hat{\Theta}_1$ , а пока поместим в них произвольные начальные числа (например,  $\Theta_0 = 2$  и  $\Theta_1 = 3$ ). Столбец С содержит предсказания переменной  $y$ , полученные с помощью (пока неоптимальной) модели  $y = \Theta_0 + \Theta_1 \cdot x$ .

	A	B	C	D	E	F	G	H	I	J	K	L
1	Наблюдаемо у	Наблюдаемо х	Предсказанное у		Остатки	Квадраты остатков	у-mean(y)					
2	67,646	79	=\\$K\$2+B2*\$L\$2								θ_0	θ_1
3	7,394	54,2									2	3
4	4,248	54,656										
5	5,584	15,228										
6	14,357	55,8										
7	3,566	37,5										

Рисунок 5 – Нахождение предсказанных значений

Поскольку нам предстоит вычислить  $\hat{y}$  для каждой страны, не забудем зафиксировать ячейки K2 и L2, поставив на соответствующих местах в формуле знак \$ (или воспользовавшись горячей клавишей F4).

Разница между истинным и предсказанным значениями объясняемой переменной, т.е. между столбцами А и С, называется остатками (ошибками, невязками) модели.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Наблюдаемо у	Наблюдаемо х	Предсказанное у		Остатки	Квадраты остатков	у-mean(y)							
2	67,646	79	239	-171,354							θ_0	θ_1		
3	7,394	54,2	164,6	-157,206							2	3		
4	4,248	54,656	165,968	-161,72										
5	5,584	15,228	47,684	-42,1										
6	14,357	55,8	169,4	=A6-C6										
7	3,566	37,5	114,5	-110,934										
8	23,063	88	266	-242,937										
9	6,722	46,91	142,73	-136,008										

Рисунок 6 – Нахождение остатков модели

Возводя остатки модели в квадрат, и суммируя по всем наблюдениям (т.е. по всем 76 странам), получаем, что сумма квадратов остатков модели  $y=2+3x$  равна 1937052. Заметим, что, изменяя начальные значения (2 и 3) параметров модели, сумма квадратов остатков изменяется. Далее нам предстоит подобрать такие

значения в ячейках K2 и L2, чтобы значение в ячейке E79 оказалось наименьшим.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Наблюдаемое у	Наблюдаемое х	Предсказанное у	Остатки	Квадраты остатков	у-mean(y)							
2	67,646	79	239	-171,354	29362,19								
3	7,394	54,2	164,6	-157,206	24713,73								
4	4,248	54,656	165,968	-161,72	26153,36								
68	15,253	61,418	186,254	-171,001	29241,34								
69	83,209	85,2	257,6	-174,391	30412,22								
70	57,134	93,18	281,54	-224,406	50358,05								
71	3,351	18,285	56,855	-53,504	2862,678								
72	5,702	35,135	107,405	-101,703	10343,5								
73	17,491	78,39	237,17	-219,679	48258,86								
74	7,59	41	125	-117,41	13785,11								
75	24,454	84,07	254,21	-229,756	52787,82								
76	5,446	33,79	103,37	-97,924	9589,11								
77	46,701	79,496	240,488	-193,787	37553,4								
78													
79						=СУММ(E2:E77)							
80						СУММ(число1; [число2]; ...)							
81													

Рисунок 7 – Квадраты остатков

6. Поскольку «на глазок» это сделать проблематично (а попробуйте!), воспользуемся встроенной надстройкой для поиска оптимальных значений.

Если данная надстройка не активирована, следует выбрать пункт меню «файл – параметры – надстройки», и внизу страницы выбрать

кнопку. 

В появившемся окне выбираем требуемые надстройки «Пакет анализа», «Поиск решения».

В результате в пункте меню «Данные» появится раздел «Поиск решения».

Напомним, что, изменяя значения в ячейках K2 и L2, нам требуется добиться минимально возможного значения в ячейке E72 (т.е. суммы квадратов остатков).

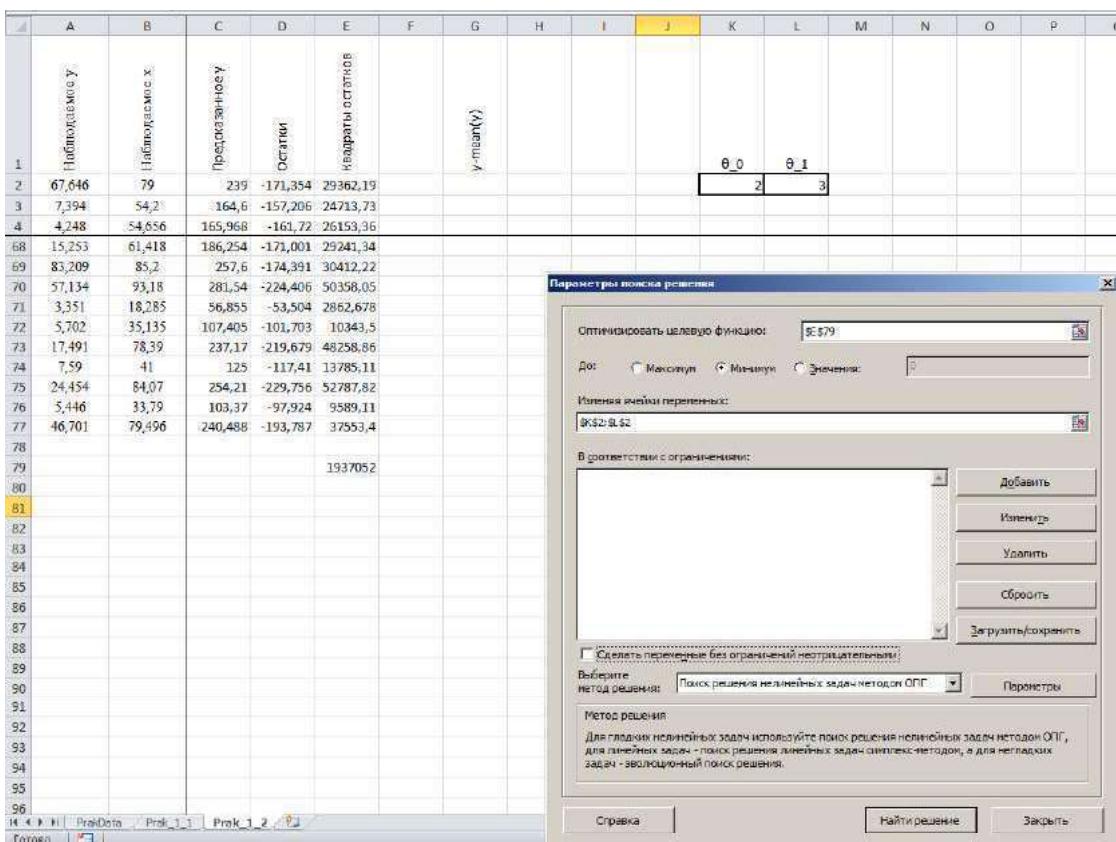


Рисунок 8 – Поиск решения модели

Не забудьте разрешить коэффициентам модели быть отрицательными.

7. Возможно, минимальное значение не будет найдено за один шаг – рекомендуем повторить данную процедуру с изменёнными начальными условиями. В результате получим, что минимальное значение суммы квадратов равно 13027 достигается при значениях коэффициентов  $\Theta_0 = -20,92$  и  $\Theta_1 = 0,7329$ . Эти значения называются «оценками коэффициентов» и обозначаются  $\hat{\Theta}_0$  и  $\hat{\Theta}_1$  соответственно. Значение  $\hat{\Theta}_1$  показывает, на сколько изменится значение зависимой переменной, если значение независимой переменной увеличить на единицу. Значение  $\hat{\Theta}_0$  показывает значение зависимой переменной при  $x=0$ , однако часто свободный член интерпретировать не принято.

8. Обратим внимание, что найденное минимальное значение суммы квадратов остатков в различных учебниках может обозначаться как RSS (Residuals Sum of Squares), так и ESS (Errors Sum of Squares). Это совершенно не принципиально в дальнейшем, поэтому для определённости остановимся на обозначении RSS.

Общая сумма квадратов TSS (Total Sum of Squares) находится как  $TSS = \sum(y - \bar{y})^2$ , где суммирование ведётся по всем наблюдениям.

Рисунок 9 – Расчет RSS, ESS, TSS

Выражение  $\Sigma(\hat{y}_i - \bar{y})^2$ , где  $\hat{y}_i = \hat{\Theta}_0 + \hat{\Theta}_1 \cdot x_i$ , а  $\bar{y}$  - среднее значение переменной  $y$ , обозначим как ESS. Проверим выполнение равенства  $TSS = RSS + ESS$ .

Вычислим коэффициент детерминации:  $R^2 = 1 - \frac{RSS}{TSS}$ .

	A	B	C	D	E	F	G	H	I	J	K	L
1	Наблюдаемое у	Наблюдаемое х	Предсказанное у	Остатки	Квадраты остатков		(y - mean(y))^2	(y^ - mean(y))^2				
2	67,646	79	36,97723	30,668766	940,5732		2239,576	277,4021			$\theta_0$	$\theta_1$
3	7,394	54,2	18,8001	-11,4061	130,0991		167,1291	2,3157			-20,9257	0,732949
4	4,248	54,656	19,13432	-14,886324	221,6027		258,3684	1,410198				
74	7,59	41	9,125173	-1,5351735	2,356758		162,0998	125,3654				1937052
75	24,454	84,07	40,69329	-16,239285	263,7144		17,07473	414,9957				
76	5,446	33,79	3,840611	1,6053885	2,577272		221,2907	271,631				
77	46,701	79,496	37,34078	9,3602234	87,61378		695,86	289,6441				
78												
79				-0,01	13027,59		34188,56	21160,48				
80	20,3218421				RSS		TSS	ESS				
81												
82												
83			R^2=		=1 - E79/G79		0,618949					
84												
85												

Рисунок 10 – Вычисление коэффициента детерминации

Коэффициент детерминации может быть проинтерпретирован как доля вариации зависимой переменной, которая может быть объяснена изменением независимой переменной. Иными словами, в рассматриваемой модели выбранный регрессор более чем на шестьдесят процентов объясняет изменение переменной  $y$ .

## 9. Обратим внимание на остатки модели.

По нижеприведенному рисунку видно, что имеются как положительные остатки (т.е. истинное значение больше предсказанного), так и отрицательные. Вычислим среднее значение остатков и убедимся, что с точностью до вычислительной ошибки оно равно нулю. То есть остатков «вверх» в сумме ровно столько же, сколько остатков «вниз».

Оцениваемая модель с подставленными коэффициентами имеет вид  $y = -20,92 + 0,73 \cdot x$ . Видим, что  $\hat{\Theta}_0 = -20,82$  - это значение переменной

$y$  в случае, если  $x=0$ , а на величину  $\hat{\Theta}_1=0,73$  изменится значение переменной  $y$  при изменении переменной  $x$  на одну единицу.

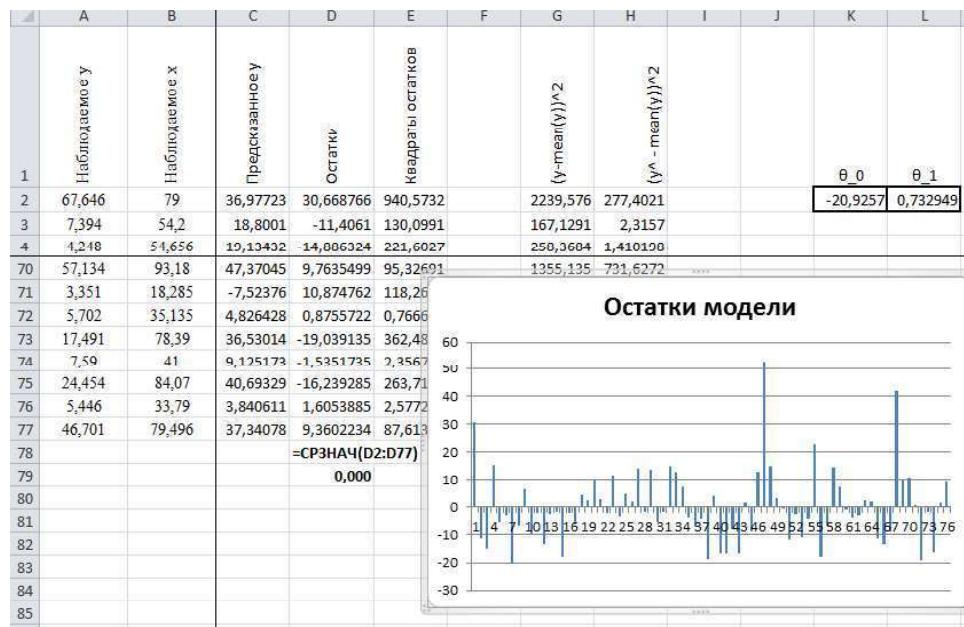


Рисунок 11 – Остатки модели

Напомним, что эти значения получены нами как результат минимизации функционала  $\Sigma(y - (\Theta_0 + \Theta_1 \cdot x)) \rightarrow \min_{\Theta_0, \Theta_1}$ .

10. Полученные выше результаты можно получить с помощью, встроенной в пакет Excel надстройки «Анализ данных». (Если надстройка не установлена, её следует установить в соответствии с приведённой в пункте 6 инструкцией.) Выбираем пункт меню «Данные» → «Анализ данных» и выбираем пункт меню «Регрессия».

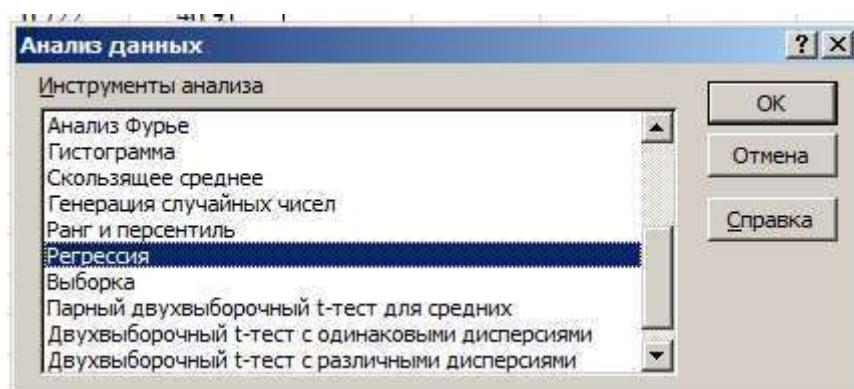


Рисунок 12 – Запуск пакета «Анализ данных»

Выбираем диапазоны зависимой и независимой переменных, и получаем таблицу результатов, которую стоит прокомментировать поподробнее.

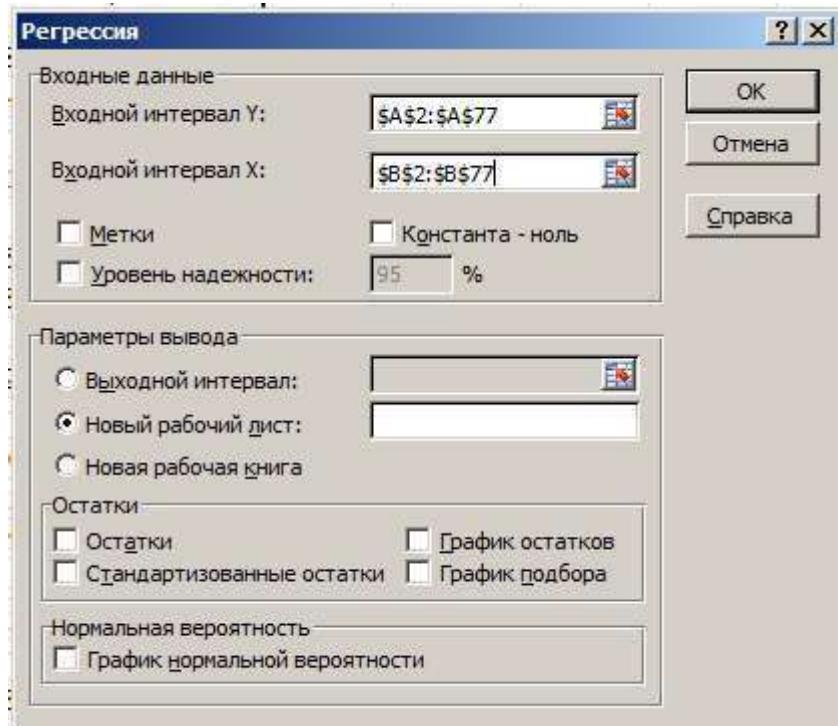


Рисунок 13 – Вычисление регрессии с помощью пакета «Анализ данных»

	A	B	C	D	E	F	G
1	ВЫВОД ИТОГОВ						
2							
3	<i>Регрессионная статистика</i>						
4	Множественны	0,786732989					
5	R-квадрат	0,618948797					
6	Нормированны	0,613799456					
7	Стандартная оц	13,2683278					
8	Наблюдения	76					
9							
10	<i>Дисперсионный анализ</i>						
11	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Значимость F</i>		
12	Регрессия	1	21160,96603	21160,96603	120,1996229	3,65563E-17	
13	Остаток	74	13027,59067	176,0485226			
14	Итого	75	34188,5567				
15							
16	<i>Коэффициенты линейной регрессии</i>	<i>Статистика</i>	<i>P-значение</i>	<i>Нижние 95%</i>	<i>Верхние 95%</i>		
17	Y-пересечение	-20,9263266	4,058486104	-5,156190279	2,03116E-06	-29,01303669	-12,83961651
18	Переменная X	0,732957359	0,066853963	10,96355886	3,65563E-17	0,599747929	0,866166788
19							

Рисунок 14 – Результаты парной регрессии в пакете «Анализ данных»

Напомним, что, например,  $2,03116E-06 = 2,03116 \cdot 10^{-6}$   
 $2,03116 \cdot 10^{-6} = 2,03116 \cdot 0,000001 = 0,00000203116$  и с точностью до тысячных  
округляется как 0,000.

В ячейках B17-D18 содержатся найденные нами ранее оценки коэффициентов, при которых сумма квадратов остатков минимальна. И равна значению в ячейке C13. Ячейка C14 содержит значение TSS, а коэффициент детерминации  $R^2$  содержится в ячейке B5.

С остальными указанными в таблице и пока не знакомыми значениями (и не только с ними!) нам предстоит разобраться в следующих разделах.

В пункте 11 предполагается подставить среднее значение переменной x в полученное уравнение регрессии.