Лабораторная работа №6.

Регрессионный анализ

Цели занятия:

приобрести навыки установки и оценки зависимости изучаемой случайной величины Y от одной или нескольких других величин X, и создания прогноза значений Y.

Задание:

1) ***построить график*** исходных данных и попытаться зрительно, приближенно определить характер зависимости;
2) ***выбрать вид функции*** регрессии, которая может описывать связь исходных данных;
3) ***определить численные коэффициенты*** функции регрессии;
4) ***оценить силу*** найденной регрессионной зависимости на основе коэффициента детерминации r^2;
5) ***сделать прогноз*** (при $r^2 \geq 75\%$) или сделать вывод о невозможности прогнозирования с помощью найденной регрессионной зависимости. При этом не рекомендуется использовать модель регрессии для тех значений независимого параметра X, которые не принадлежат интервалу, заданному в исходных данных.

Краткие сведения из теории

Регрессионный и корреляционный анализ позволяет установить и оценить зависимость изучаемой случайной величины Y от одной или нескольких других величин X, и делать прогнозы значений Y. Параметр Y, значение которого нужно предсказывать, является **зависимой** переменной. Параметр X, значения которого нам известны заранее и который влияет на значения Y, называется **независимой** переменной. Например, X – количество внесенных удобрений, Y – снимаемый урожай; X – величина затрат компании на рекламу своего товара, Y – объем продаж этого товара и т.д.

Корреляционная зависимость Y от X – это функциональная зависимость

$$\overline{y}_x = f(x), \qquad (6.1)$$

где $\overline{y}_x$ – среднее арифметическое (**условное среднее**) всех возможных значений параметра Y, которые соответствуют значению X=x. Уравнение (6.1) называется **уравнением регрессии** Y на X, функция f(x) – **регрессией** Y на X**,** а ее график – **линией регрессии** Y на X.

Основная задача регрессионного анализа – установление ***формы*** корреляционной связи, т.е. вида функции регрессии (линейная, квадратичная, показательная и т.д.).

Метод наименьших квадратов позволяет определить коэффициенты уравнения регрессии таким образом, чтобы точки, построенные по исходным

данным (x_i, y_i), лежали как можно ближе к точкам линии регрессии (6.1). Формально это записывается как минимизация суммы квадратов отклонений (ошибок) функции регрессии и исходных точек

$$S = \sum_{i=1}^{n} (y_i^p - y_i)^2 \rightarrow \min,$$

$\boldsymbol{y^p_i}$ – значение, вычисленное по уравнению регрессии; $(\boldsymbol{y^p_i - y_i})$ – отклонение ε (ошибка, остаток) (рис. 6.1); n – количество пар исходных данных.

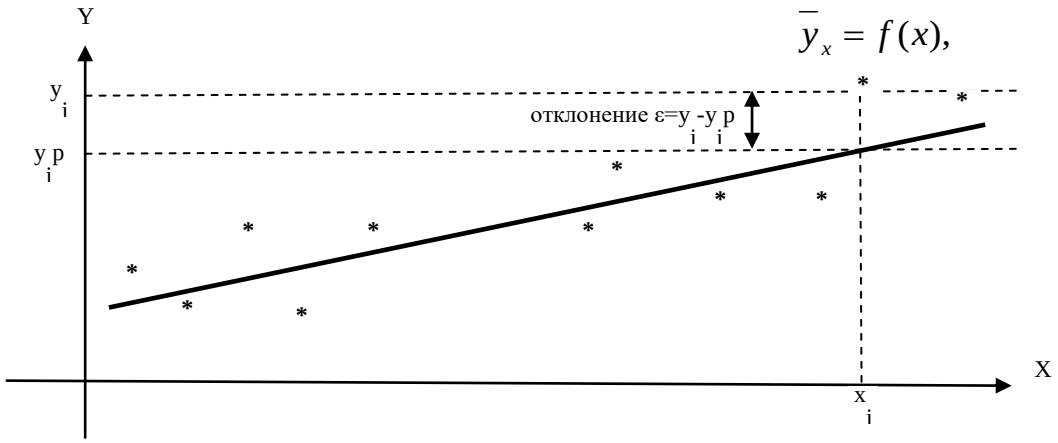


Рис. 6.1. Понятие отклонения ε для случая линейной регрессии

В регрессионном анализе предполагается, что математическое ожидание случайной величины ε равно нулю и ее дисперсия одинакова для всех наблюдаемых значений Y. Отсюда следует, что рассеяние данных возле линии регрессии должно быть одинаково при всех значениях параметра X. В случае, показанном на рис. 6.2 данные распределяются вдоль линии регрессии неравномерно, поэтому метод наименьших квадратов в этом случае неприменим.

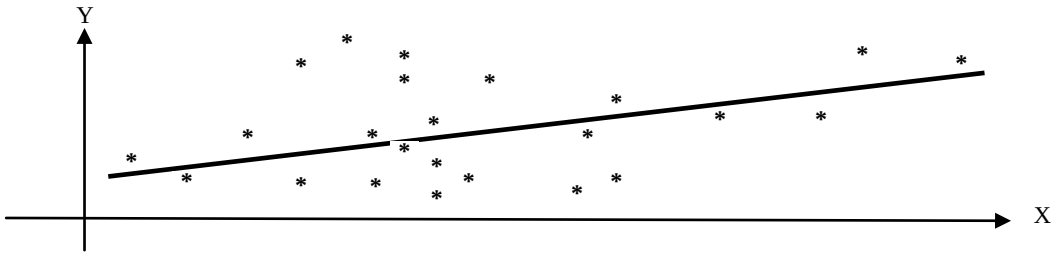


Рис. 6.2. Неравномерное распределение исходных точек вдоль линии регрессии

Основная задача корреляционного анализа – оценка ***тесноты*** (силы) корреляционной связи. Теснота корреляционной зависимости Y от X оценивается по величине рассеяния значений параметра Y вокруг условного среднего $\bar{y}_x$. Большое рассеяние говорит о слабой зависимости Y от X, либо об ее отсутствии и, наоборот, малое рассеяние указывает на наличие достаточно сильной зависимости.

Коэффициент детерминации r^2 показывает, на сколько процентов (r^2*100%) найденная функция регрессии описывает связь между исходными значениями параметров X и Y

$$r^2 = \frac{\sum_{i=1}^{n}(y_i^P - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}, \quad (6.2)$$

где $(y_i^p - \overline{y})^2$ - объяснённая вариация; $(y_i - \overline{y})^2$ - общая вариация (рис. 6.3).

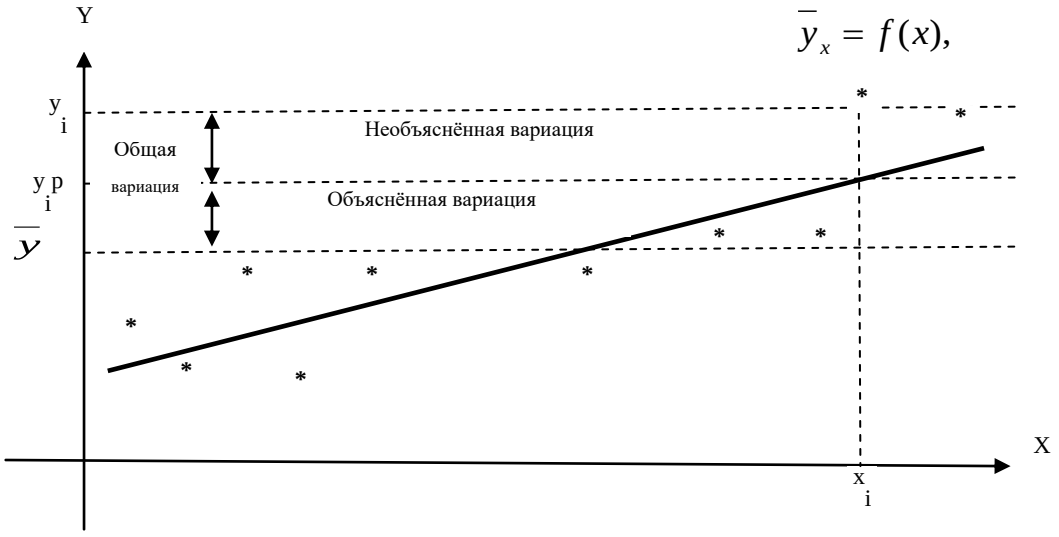


Рис. 6.3. Графическая интерпретация коэффициента детерминации для случая линейной регрессии

Соответственно, величина $(1-r^2)*100\%$ показывает, сколько процентов вариации параметра Y обусловлены факторами, не включенными в регрессионную модель. При высоком ($r^2 \geq 75\%$) значении коэффициента детерминации можно делать прогноз y*=f(x*) для конкретного значения x.

6.2 Методические рекомендации

6.2.1 Линейная регрессия

Коэффициенты **линейной** регрессии $y=a_0+a_1x$ вычисляются по следующим формулам (все суммы берутся по n парам исходных данных):

$$a_1 = \frac{n(\sum_{i=1}^{n} y_i x_i) - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}; \qquad (6.3)$$

$$a_0 = \frac{1}{n}(\sum_{i=1}^{n} y_i - a_1 \sum_{i=1}^{n} x_i).$$

Задача № 6.01

Некоторая фирма занимается поставками различных грузов на короткие расстояния внутри города. Перед менеджером стоит задача оценить стоимость таких услуг, зависящую от затрачиваемого на поставку времени. В качестве наиболее важного фактора, влияющего на время поставки, менеджер выбрал пройденное расстояние. Были собраны исходные данные о десяти поставках (табл. 6.1).

Таблица 6.1

Расстояние, км.	3,5	2,4	4,9	4,2	3,0	1,3	1,0	3,0	1,5	4,1
Время, мин.	16	13	19	18	12	11	8	14	9	16

Построить график исходных данных, определить по нему характер зависимости между расстоянием и затраченным временем, проанализировать применимость метода наименьших квадратов, построить уравнение регрессии, проанализировать силу регрессионной связи и сделать прогноз времени поездки на 2 км.

Решение

На рис. 6.4 построены исходные данные по десяти поездкам.

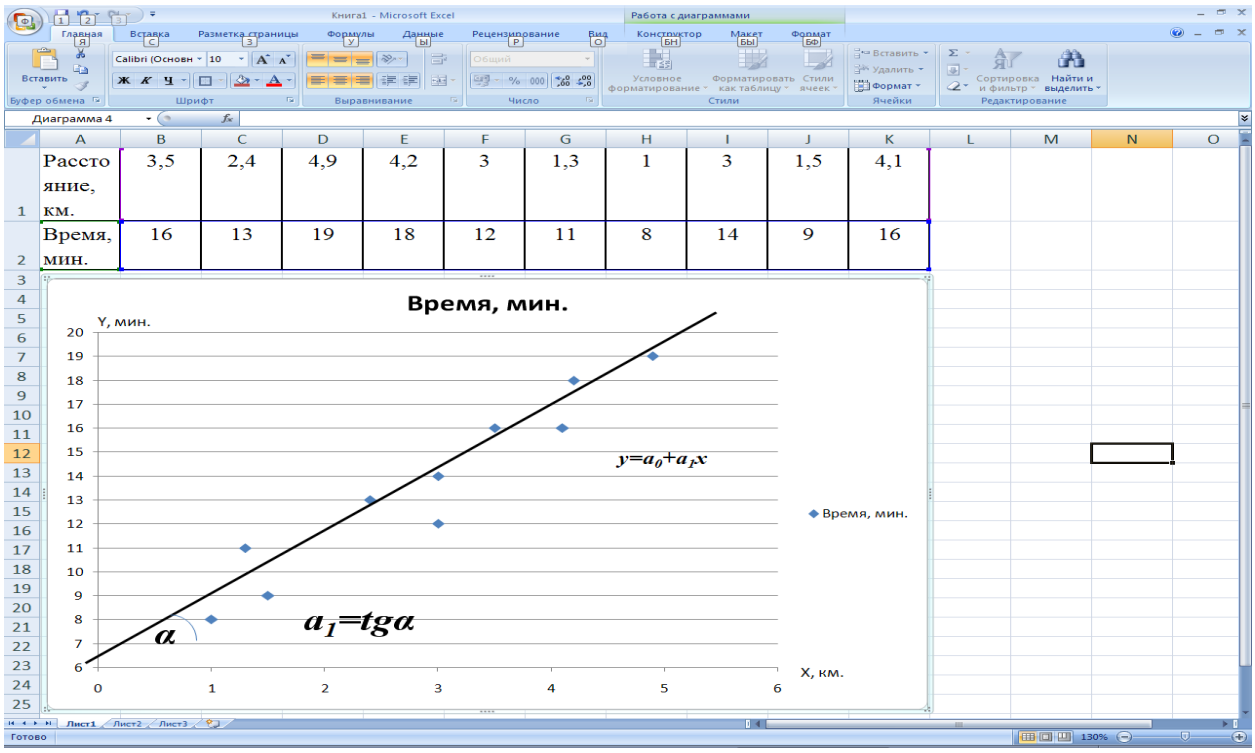
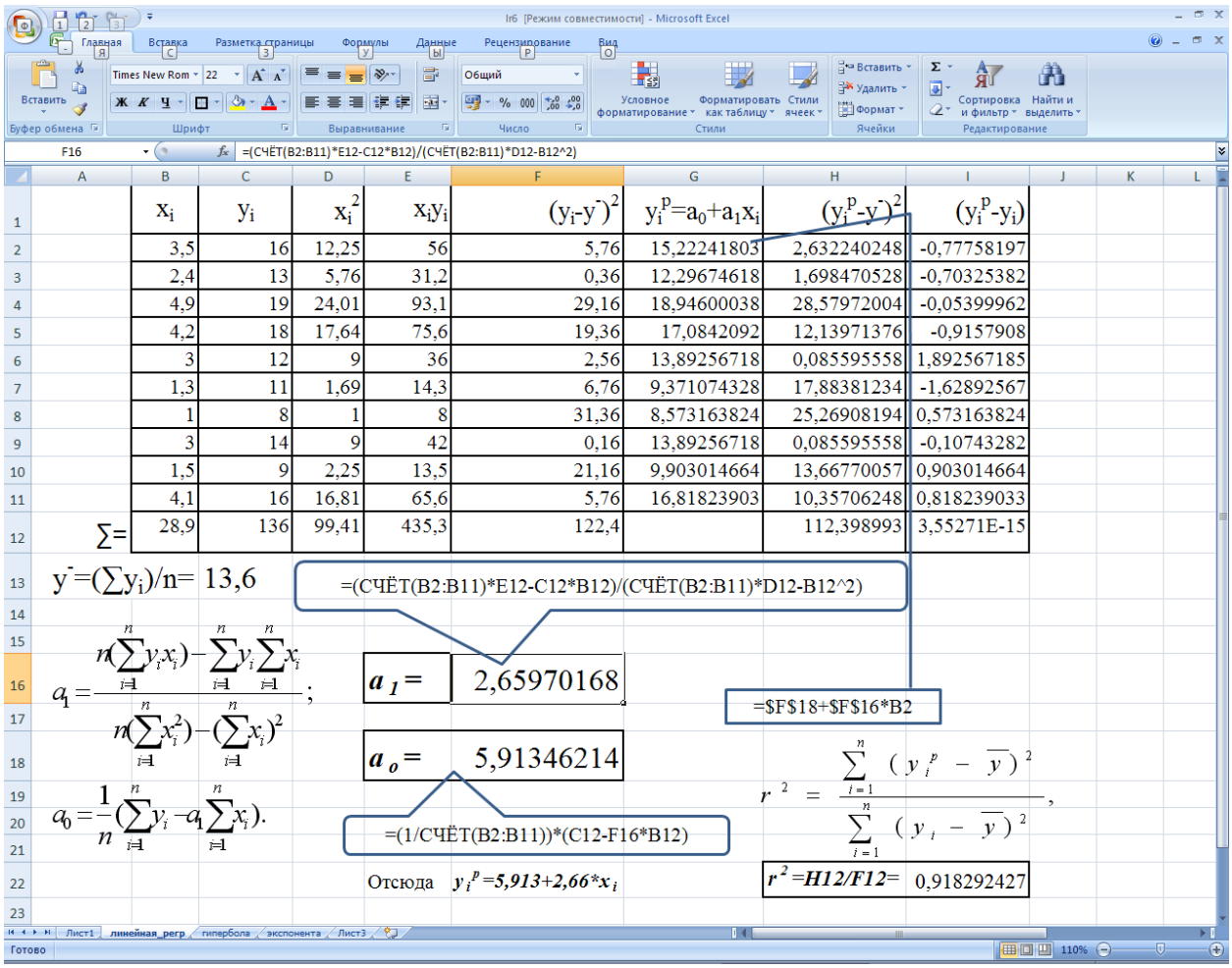


Рис. 6.4. График исходных данных задачи

Помимо расстояния на время поставки влияют пробки на дорогах, время суток, дорожные работы, погода, квалификация водителя, вид транспорта. Построенные точки не находятся точно на линии, что обусловлено описанными выше факторами. Но эти точки собраны вокруг прямой линии, поэтому можно предположить линейную связь между параметрами. Все исходные точки равномерно распределены вдоль предполагаемой прямой линии, что позволяет применить метод наименьших квадратов.



	x_i	y_i	x_i^2	$x_i y_i$	$(y_i-\bar{y})^2$	$y_i^p=a_0+a_1x_i$	$(y_i^p-\bar{y})^2$	$(y_i^p-y_i)$
	3,5	16	12,25	56	5,76	15,22241803	2,632240248	-0,77758197
	2,4	13	5,76	31,2	0,36	12,29674618	1,698470528	-0,70325382
	4,9	19	24,01	93,1	29,16	18,94600038	28,57972004	-0,05399962
	4,2	18	17,64	75,6	19,36	17,0842092	12,13971376	-0,9157908
	3	12	9	36	2,56	13,89256718	0,085595558	1,892567185
	1,3	11	1,69	14,3	6,76	9,371074328	17,88381234	-1,62892567
	1	8	1	8	31,36	8,573163824	25,26908194	0,573163824
	3	14	9	42	0,16	13,89256718	0,085595558	-0,10743282
	1,5	9	2,25	13,5	21,16	9,903014664	13,66770057	0,903014664
	4,1	16	16,81	65,6	5,76	16,81823903	10,35706248	0,818239033
∑=	28,9	136	99,41	435,3	122,4		112,398993	3,55271E-15

$\bar{y}=(\sum y_i)/n = 13{,}6$

$$a_1 = \frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2};$$

a_1 = 2,65970168

$$a_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - a_1 \sum_{i=1}^{n} x_i\right).$$

a_0 = 5,91346214

$$r^2 = \frac{\sum_{i=1}^{n} (y_i^p - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2},$$

r^2 =H12/F12= 0,918292427

Рис. 6.5

Наклон линии регрессии $a_1=2{,}66$ мин. на км. – это количество минут, приходящееся на 1 км. Координата $a_0=5{,}913$ мин. – это время, которое не зависит от пройденного расстояния, а обуславливается всеми остальными возможными факторами, явно не учтёнными при анализе.

Вычислим коэффициент детерминации (6.2):

$$r^2 = 112{,}4/122{,}4 = 0{,}918 \text{ или } 91{,}8\ \%.$$

Таким образом, линейная модель объясняет 91,8 % вариации времени доставки, а не объясняет 8,2 %.

Поскольку r^2 имеет достаточно высокое значение и расстояние 2 км., для которого нужно сделать прогноз, находится в пределах диапазона исходных данных (табл. 6.1), то возможно использовать полученное уравнение регрессии для прогнозирования:

$$y^*(2 \text{ км.}) = 5{,}913 + 2{,}66 * 2 = 11{,}23 \text{ мин.}$$

Приблизительным, но простым и наглядным способом проверки удовлетворительности регрессионной модели является сравнение отклонений функции от расчетных значений, т.е. $(y_i^p - y_i)$. Если регрессионная модель близка к реальной зависимости, то отклонения будут носить случайный характер и их сумма будет близка к нулю. В примере $\sum(y_i^p - y_i) = 3{,}55E\text{-}15$.

6.2.2 Нелинейная регрессия

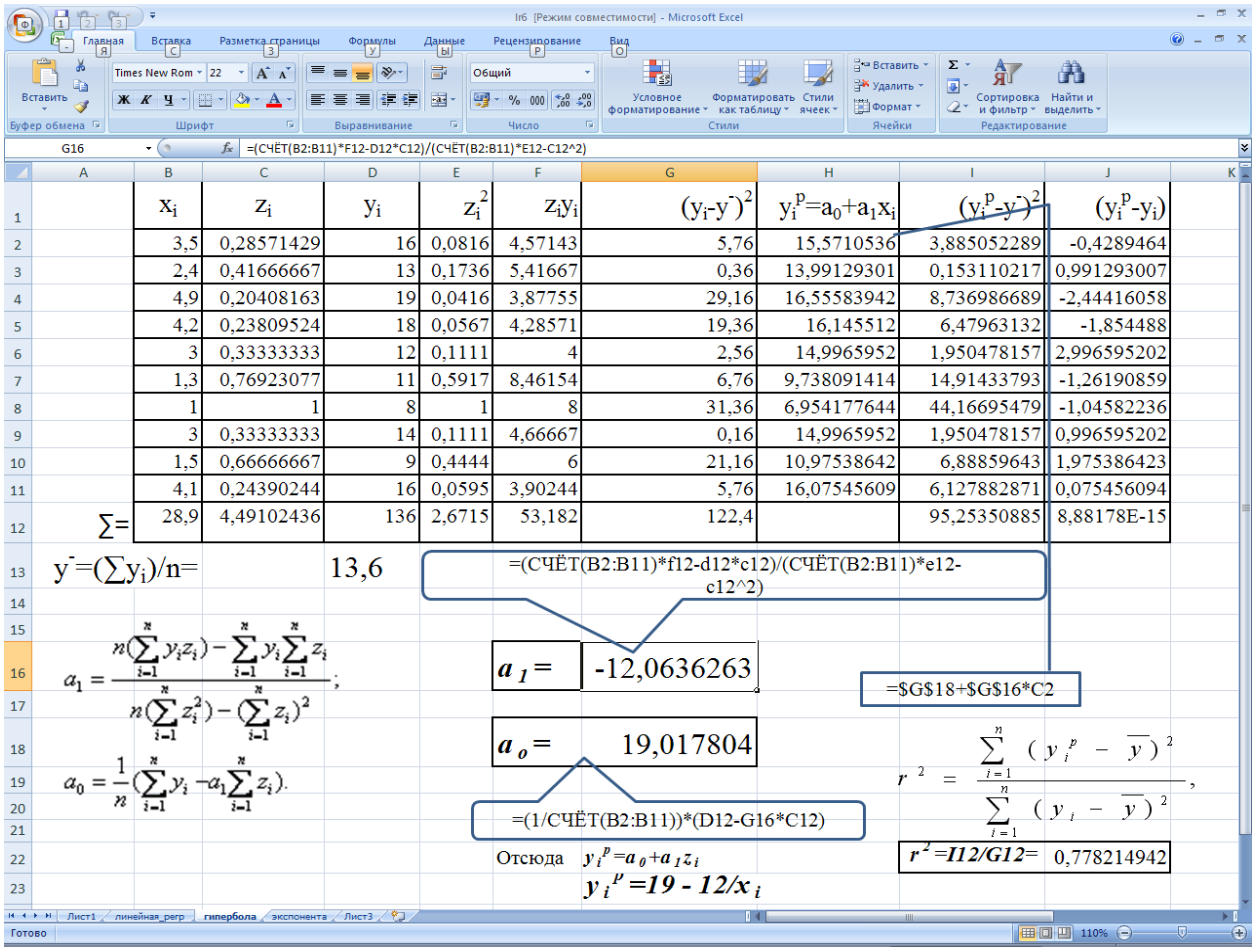
Наиболее простые случаи – гипербола, экспонента и парабола. При нахождении коэффициентов гиперболы и экспоненты используют приём приведения нелинейной регрессионной зависимости к линейному виду. Это позволяет использовать для вычисления коэффициентов функций регрессии формулы (6.3).

Гипербола

При нахождении гиперболы $y=a_0+a_1/x$ вводят новую переменную $z=1/x$, тогда уравнение гиперболы принимает линейный вид $y=a_0+a_1z$. После этого используют формулы (6.3) для нахождения линейной функции, но вместо значений x_i используется $z_i=1/x_i$.

$$a_1 = \frac{n(\sum_{i=1}^{n} y_i z_i) - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} z_i}{n(\sum_{i=1}^{n} z_i^2) - (\sum_{i=1}^{n} z_i)^2}; \qquad (6.4)$$

$$a_0 = \frac{1}{n}(\sum_{i=1}^{n} y_i - a_1 \sum_{i=1}^{n} z_i).$$



	A	B	C	D	E	F	G	H	I	J
1		x_i	z_i	y_i	z_i^2	z_iy_i	$(y_i-\bar{y})^2$	$y_i^p=a_0+a_1x_i$	$(y_i^p-\bar{y})^2$	$(y_i^p-y_i)$
2		3,5	0,28571429	16	0,0816	4,57143	5,76	15,5710536	3,885052289	-0,4289464
3		2,4	0,41666667	13	0,1736	5,41667	0,36	13,99129301	0,153110217	0,991293007
4		4,9	0,20408163	19	0,0416	3,87755	29,16	16,55583942	8,736986689	-2,44416058
5		4,2	0,23809524	18	0,0567	4,28571	19,36	16,145512	6,47963132	-1,854488
6		3	0,33333333	12	0,1111	4	2,56	14,9965952	1,950478157	2,996595202
7		1,3	0,76923077	11	0,5917	8,46154	6,76	9,738091414	14,91433793	-1,26190859
8		1	1	8	1	8	31,36	6,954177644	44,16695479	-1,04582236
9		3	0,33333333	14	0,1111	4,66667	0,16	14,9965952	1,950478157	0,996595202
10		1,5	0,66666667	9	0,4444	6	21,16	10,97538642	6,88859643	1,975386423
11		4,1	0,24390244	16	0,0595	3,90244	5,76	16,07545609	6,127882871	0,075456094
12	∑=	28,9	4,49102436	136	2,6715	53,182	122,4		95,25350885	8,88178E-15
13	$\bar{y}=(\sum y_i)/n=$			13,6						

$$a_1 = \frac{n(\sum_{i=1}^{n} y_i z_i) - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} z_i}{n(\sum_{i=1}^{n} z_i^2) - (\sum_{i=1}^{n} z_i)^2};$$

$$a_0 = \frac{1}{n}(\sum_{i=1}^{n} y_i - a_1 \sum_{i=1}^{n} z_i).$$

$$r^2 = \frac{\sum_{i=1}^{n} (y_i^p - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2},$$

Рис. 6.6

Поскольку коэффициент детерминации r^2 имеет достаточно высокое значение (≈ 78 %)и расстояние 2 км., для которого нужно сделать прогноз, находится в пределах диапазона исходных данных (табл. 6.1), то возможно использовать полученное уравнение регрессии для прогнозирования:

$$y^*(2 \text{ км.}) = 19\text{-}12/2 = 13 \text{ мин.}$$

Экспонента

Для приведения к линейному виду экспоненты $y = a_0 e^{a_1 x}$ проведём логарифмирование

$$\ln y = \ln(a_0 e^{a_1 x})$$

$$\ln y = \ln a_0 + \ln(e^{a_1 x})$$

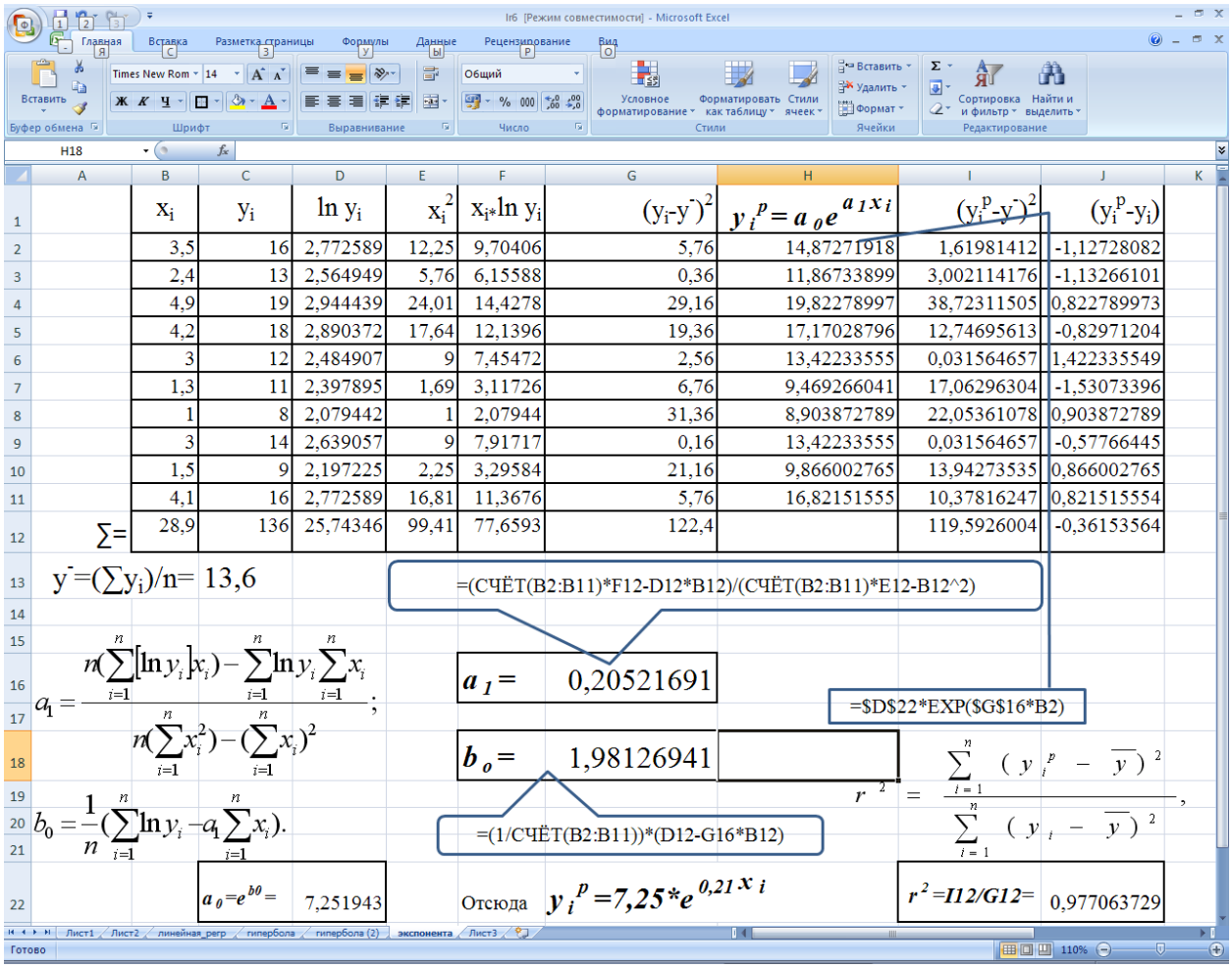
$$\ln y = \ln a_0 + a_1 x.$$

Введём переменную $b_0 = \ln a_0$, тогда $\ln y = b_0 + a_1 x$. Отсюда следует, что можно применять формулы (6.3), в которых вместо значений y_i нужно использовать $\ln y_i$:

$$a_1 = \frac{n(\sum_{i=1}^{n}[\ln y_i]x_i) - \sum_{i=1}^{n}\ln y_i \sum_{i=1}^{n} x_i}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}; \qquad (6.5)$$

$$b_0 = \frac{1}{n}(\sum_{i=1}^{n}\ln y_i - a_1 \sum_{i=1}^{n} x_i).$$

Откуда $a_0 = e^{b_0}$.



	A	B	C	D	E	F	G	H	I	J
1		x_i	y_i	$\ln y_i$	x_i^2	$x_i \ln y_i$	$(y_i-\bar{y})^2$	$y_i^p = a_0 e^{a_1 x_i}$	$(y_i^p-\bar{y})^2$	$(y_i^p-y_i)$
2		3,5	16	2,772589	12,25	9,70406	5,76	14,87271918	1,61981412	-1,12728082
3		2,4	13	2,564949	5,76	6,15588	0,36	11,86733899	3,002114176	-1,13266101
4		4,9	19	2,944439	24,01	14,4278	29,16	19,82278997	38,72311505	0,822789973
5		4,2	18	2,890372	17,64	12,1396	19,36	17,17028796	12,74695613	-0,82971204
6		3	12	2,484907	9	7,45472	2,56	13,42233555	0,031564657	1,422335549
7		1,3	11	2,397895	1,69	3,11726	6,76	9,469266041	17,06296304	-1,53073396
8		1	8	2,079442	1	2,07944	31,36	8,903872789	22,05361078	0,903872789
9		3	14	2,639057	9	7,91717	0,16	13,42233555	0,031564657	-0,57766445
10		1,5	9	2,197225	2,25	3,29584	21,16	9,866002765	13,94273535	0,866002765
11		4,1	16	2,772589	16,81	11,3676	5,76	16,82151555	10,37816247	0,821515554
12	$\sum$=	28,9	136	25,74346	99,41	77,6593	122,4		119,5926004	-0,36153564

13 $\bar{y}=(\sum y_i)/n=$ 13,6

$$a_1 = \frac{n\left(\sum_{i=1}^{n}[\ln y_i]x_i\right) - \sum_{i=1}^{n}\ln y_i \sum_{i=1}^{n} x_i}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2};$$

$a_1 =$ 0,20521691

$$b_0 = \frac{1}{n}\left(\sum_{i=1}^{n}\ln y_i - a_1 \sum_{i=1}^{n} x_i\right).$$

$b_0 =$ 1,98126941

$a_0 = e^{b0} =$ 7,251943

Отсюда $y_i^p = 7{,}25 * e^{0{,}21 x_i}$

$$r^2 = \frac{\sum_{i=1}^{n}(y_i^p - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

r^2=I12/G12= 0,977063729

Рис. 6.7

Поскольку коэффициент детерминации r^2 имеет достаточно высокое значение (≈ 98 %)и расстояние 2 км., для которого нужно сделать прогноз, находится в пределах диапазона исходных данных (табл. 6.1), то возможно использовать полученное уравнение регрессии для прогнозирования:

$$y_i^p = 7{,}25 * e^{0{,}21 X_i} = y^*(2 \text{ км.}) = 7{,}25 * e^{0{,}21 * 2} = 11{,}19 \text{ мин.}$$

Таким образом, для рассмотренного примера приемлемыми являются все три рассмотренные регрессионные модели. При этом, для рассмотренного примера наилучший коэффициент детерминации $r^2 \approx 98$ % получен при использовании экспоненциальной модели.

Варианты заданий

1. Построить регрессионные модели (линейную, гиперболу, экспоненту) для данных, связывающих расходы на потребление ***y*** с душевым доходом ***x***.
2. Найти коэффициенты детерминации и определить наиболее адекватную модель.
3. Сделать прогноз функции y при прогнозном значении фактора ***x***, превышающем среднее $\bar{x}$ на 20%.

Вариант 1		Вариант 2		Вариант 3		Вариант 4		Вариант 5	
x	y	x	y	x	y	x	y	x	y
200	17	200	114	200	14	200	113	209	109
250	27	250	123	250	20	250	124	251	121
300	39	300	132	300	25	300	132	304	134
350	40	350	143	350	28	350	148	359	149
400	44	400	152	400	88	400	152	402	152
450	68	450	161	450	99	450	164	459	159
500	122	500	169	500	102	500	168	535	165
550	121	550	171	550	222	550	222	559	222
600	288	600	277	600	333	600	179	602	345
650	189	650	266	650	175	650	185	657	180

Вариант 6		Вариант 7		Вариант 8		Вариант 9		Вариант 10	
x	y	x	y	x	y	x	y	x	y
207	107	211	105	204	112	199	109	202	5
242	112	253	124	259	123	259	129	240	12
315	129	308	135	345	135	308	132	308	18
408	152	354	144	364	145	348	140	352	44
460	160	404	154	414	151	445	153	440	77
541	163	456	157	458	160	462	162	450	88
558	171	538	163	529	270	505	269	532	99
605	233	545	171	550	280	550	371	560	11
650	323	604	179	620	387	600	481	640	321
679	184	555	185	698	390	650	494	659	345

Вариант 11		Вариант 12		Вариант 13		Вариант 14		Вариант 15	
x	y	x	y	x	y	x	y	x	y
205	9	203	11	204	14	211	20	200	17
252	12	250	14	250	22	259	24	260	26
300	18	309	18	337	37	338	35	302	31
360	22	370	22	352	42	354	40	350	33
400	27	409	27	449	49	453	54	412	44
440	33	450	29	469	55	462	70	451	66
505	44	510	77	532	77	504	87	520	98
550	46	559	88	560	165	559	90	552	111
602	88	607	111	649	170	662	155	607	150
670	99	658	122	680	181	682	179	653	184