

Гетероскедастичность

Общие понятия

Гомоскедастичность

$$\sigma_u^2 = \text{constant}$$

m.e.

$$\sigma_u^2 \neq f(x_{1k}, x_{2k}, \dots, x_{nk})$$

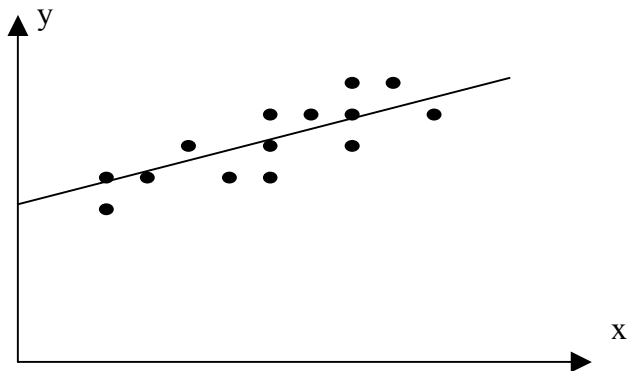
- *Гетероскедастичность* – это нарушение классического предположения о постоянстве дисперсий ошибок, т.е.

$$\sigma_u^2 \neq \text{constant}$$

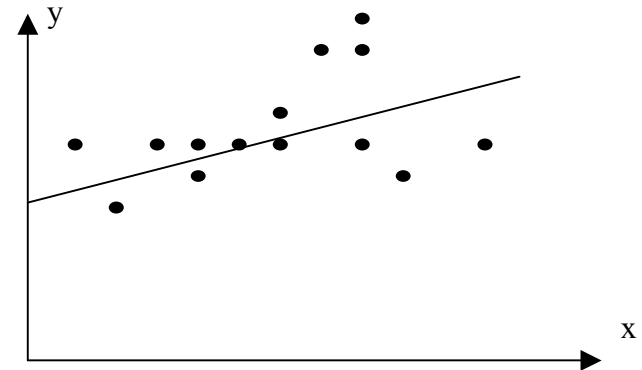
ИЛИ

$$\sigma_u^2 = f(x_{1k}, x_{2k}, \dots, x_{nk})$$

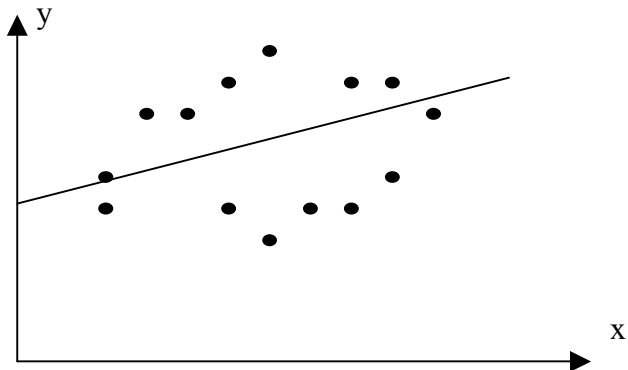
Графическая интерпретация гомо- и гетероскедастичности



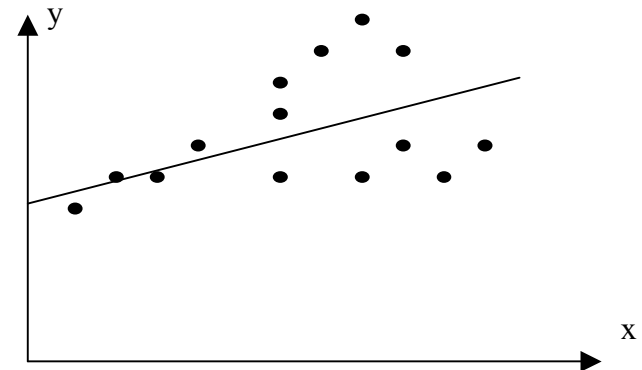
Гомоскедастичность



Гетероскедастичность



Гетероскедастичность

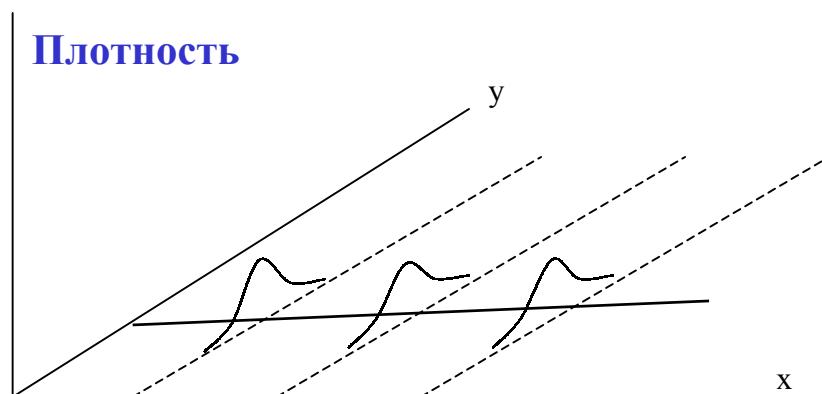


Гетероскедастичность

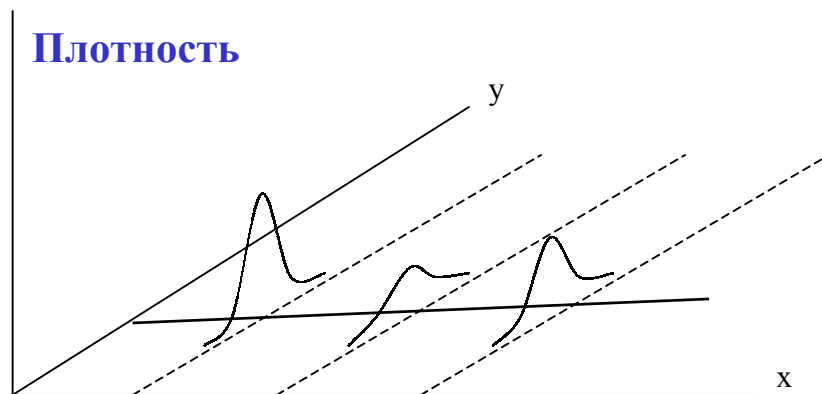
Причины гетероскедастичности

- большие различия между наименьшими и наибольшими значениями наблюдений выборки, взятой в пространственном разрезе (высокая дисперсия значений наблюдений)
- существенное различие в качестве исходных данных внутри выборки

Другая иллюстрация гетероскедастичности



Гомоскедастичность



Гетероскедастичность

- Чистая гетероскедастичность – нарушение предположения о постоянстве дисперсий остатков в корректно специфицированном уравнении регрессии
- Смешанная (нечистая) гетероскедастичность – возникает при неверной спецификации модели в случае не включения в нее существенно влияющих переменных

- Наличие гетероскедастичности затрагивает эффективность оценок модели



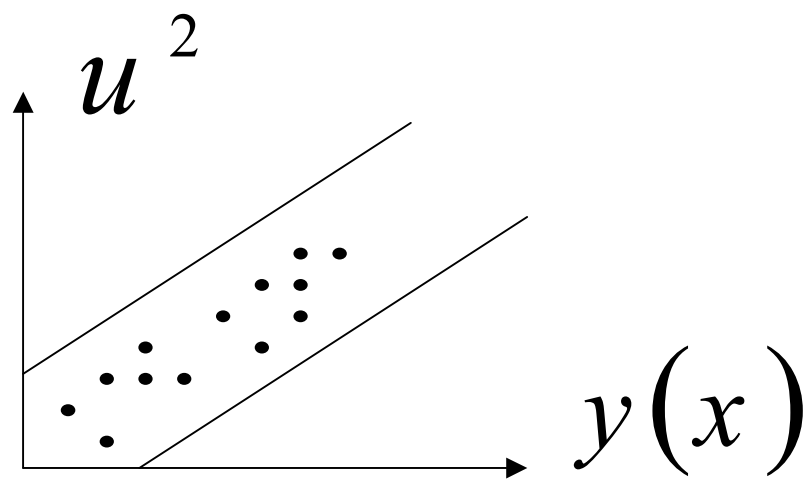
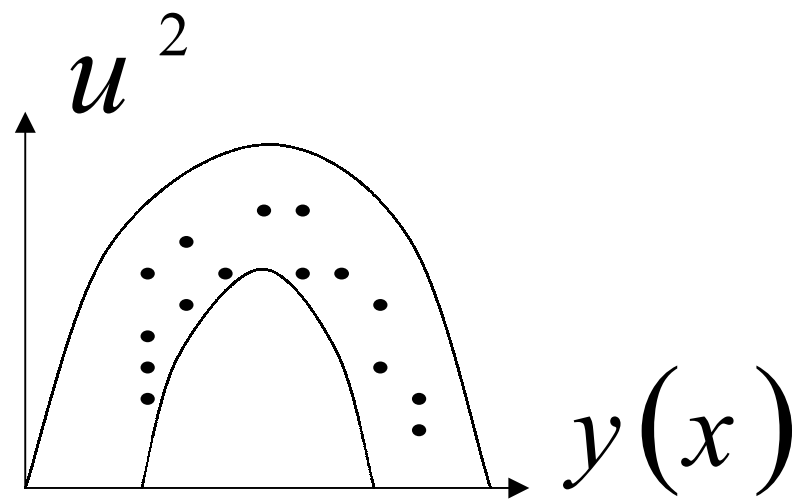
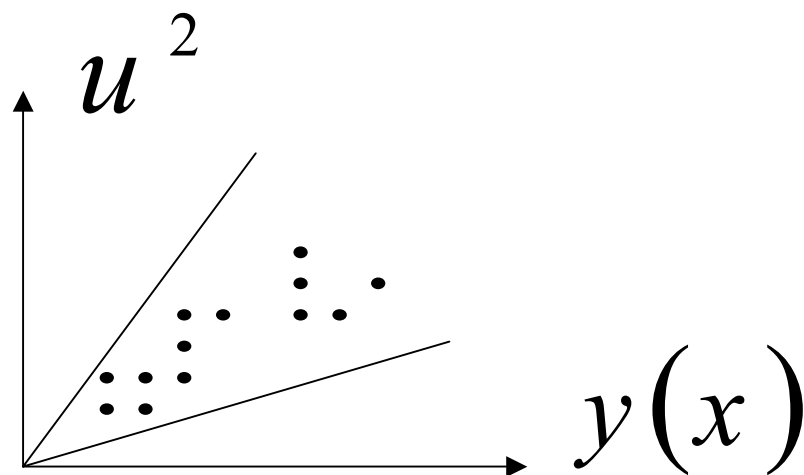
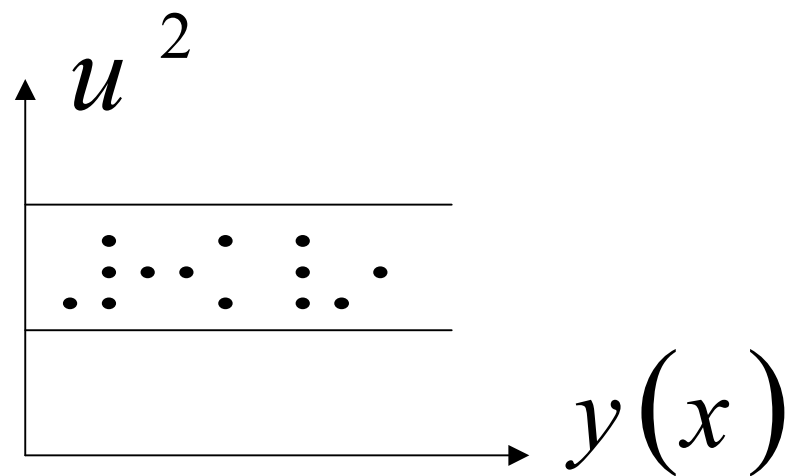
оценка дисперсии их ошибок $\hat{\sigma}_u^2$ не может быть использована для проверки значимости параметров модели и расчета их доверительных интервалов

Методы выявления гетероскедастичности

- анализ содержания проблемы,
- графический анализ,
- тест ранговой корреляции Спирмена,
- μ -критерий,
- параметрический и непараметрический тесты Гольдфельда-Квондта,
- тест Глейсера,
- тест Парка,
- тест Бреуша-Пэйгана,
- тест Уайта и др.

Графический анализ

- Для проведения графического анализа необходимо рассчитать значения квадратов отклонений u_i^2 и затем определить, имеют ли они какую-либо систематичность



μ -критерий

- 1-й шаг.

Все наблюдения зависимой переменной разбиваются на p групп $(r = \overline{1, p})$ в соответствии с уровнем изменения величины y

- 2-й шаг.

Для каждой группы рассчитываются
суммы квадратов отклонений

$$S_r = \sum_{i=1}^{n_r} (y_{ir} - \bar{y}_r)^2$$

- 3-й шаг.

Находится общая сумма квадратов
отклонений по всем группам

$$S = \sum_{r=1}^p S_r$$

- 4-й шаг.

Вычисляется параметр $\mathcal{W}$

$$\mathcal{W} = \frac{\prod_{r=1}^p \left(\frac{S_r}{n_r} \right)^{n_r/2}}{\left(\frac{S}{n} \right)^{n/2}}$$

n_r – число наблюдений r -й группы

n – общее число наблюдений

- 5-й шаг.

Рассчитывается значение критерия μ

$$\mu = -2 \ln w$$

(приблизительно соответствует критерию χ^2
при числе степеней свободы $\nu = p - 1$,
когда дисперсия всех наблюдений
однородна)

Если

$$\mu \geq \chi^2$$

при заданном уровне значимости α , то
имеет место гетероскедастичность

Пример

- Разобьем наблюдения на пять групп по пять наблюдений в каждой группе

Группа 1	Группа 2	Группа 3	Группа 4	Группа 5
1,82	3,71	3,24	2,83	1,49
2,19	3,07	2,12	3,03	2,69
4,23	1,75	3,95	3,08	2,29
1,66	4,78	2,28	3,49	1,92
1,84	3,35	0,47	2,00	4,22

- Найдем средние значения наблюдений каждой группы

$$\bar{y}_1 = 2,384$$

$$\bar{y}_2 = 3,332$$

$$\bar{y}_3 = 2,412$$

$$\bar{y}_4 = 2,886$$

$$\bar{y}_5 = 2,522$$

• Найдем сумму квадратов отклонений индивидуальных значений каждой группы от своего среднего значения

$$S_1 = \sum_{i=1}^5 (y_{i1} - \bar{y}_1)^2 = 4,57708$$

$$S_2 = \sum_{i=1}^5 (y_{i2} - \bar{y}_2)^2 = 4,81128$$

$$S_3 = \sum_{i=1}^5 (y_{i3} - \bar{y}_3)^2 = 6,92508$$

$$S_4 = \sum_{i=1}^5 (y_{i4} - \bar{y}_4)^2 = 1,21132$$

$$S_5 = \sum_{i=1}^5 (y_{i5} - \bar{y}_5)^2 = 4,39268$$

- Рассчитаем общую сумму квадратов отклонений по пяти группам

$$S = \sum_{r=1}^5 S_r = 21,91744$$

• Определим параметр W

$$W = \frac{\prod_{r=1}^p \left(\frac{S_r}{n_r} \right)^{n_r/2}}{\left(\frac{S}{n} \right)^{n/2}} =$$

$$= \frac{\left(\frac{4,58}{5} \right)^{5/2} \cdot \left(\frac{4,81}{5} \right)^{5/2} \cdot \left(\frac{6,93}{5} \right)^{5/2} \cdot \left(\frac{1,21}{5} \right)^{5/2} \cdot \left(\frac{4,39}{5} \right)^{5/2}}{\left(\frac{21,92}{25} \right)^{25/2}} = 0,178$$

- Найдем критерий μ

$$\mu = -2 \ln w = -2 \cdot (-1.726) = 3,542$$

Табличное значение критерия $\chi^2 = 9,49$

для числа степеней свободы

$$\nu = p - 1 = 5 - 1 = 4$$

и уровня значимости $\alpha = 0,05$

Так как $\mu < \chi^2$, то делаем заключение об

отсутствии гетероскедастичности для

исследуемого набора данных

Тест ранговой корреляции Спирмена

- 1-й шаг.

На основе построенного уравнения
регрессии рассчитываем величины
остатков u_i

- 2-й шаг.

Ранжируем по возрастанию (убыванию)

абсолютные значения остатков $|u_i|$ и

значения независимой переменной x_j

Рассчитываем коэффициент ранговой
корреляции Спирмена

$$r_s = 1 - 6 \cdot \left[\frac{\sum_{i=1}^n d_i^2}{n(n^2 - 1)} \right]$$

d_i – разность между рангами, которые
приписываются двум характеристикам

$(x_j \text{ и } u)$ i -го объекта,

n – количество ранжируемых объектов.

- 3-й шаг.

Проверяем значимость коэффициента ранговой корреляции по критерию Стьюдента. Рассчитываем статистику Стьюдента

$$t = \frac{r_S \sqrt{n-2}}{\sqrt{1-r_S^2}}$$

Если $t_{расч.} \geq t_{табл.}$ для числа степеней свободы

$\nu = n - 2$ и уровня значимости α , то принимается

гипотеза о наличии гетероскедастичности остатков.

Если $t_{расч.} < t_{табл.}$, то считается, что в модели

имеет место гомоскедастичность

Пример

Независимая переменная x_1 — численность населения

Проранжируем по возрастанию численность населения и величины остатков модели

Население, тыс. чел.	Ранг	Модули остатков	Ранг	d_i	d_i^2
940,8	1	0,01667	1	0	0
1071,8	2	0,63702	16	-14	196
1172,3	3	0,79428	18	-15	225
1192,5	4	0,69994	17	-13	169
1211,2	5	0,10498	5	0	0
1255,1	6	1,80037	25	-19	361
1288,6	7	0,21139	9	-2	4
1332,1	8	0,81888	19	-11	121
1333,4	9	1,65011	24	-15	225
1384,3	10	1,18054	20	-10	100
1465,6	11	0,09936	3	8	64
1467,9	12	0,56415	15	-3	9
1491,1	13	0,42480	12	1	1
1498,4	14	0,11667	6	8	64
1723,9	15	1,57790	23	-8	64
1862,2	16	0,08118	2	14	196
1880,4	17	0,37616	11	6	36
2059,2	18	0,44007	13	5	25
2566,8	19	0,16261	7	12	144
2580,2	20	0,20826	8	12	144
2743	21	0,10447	4	17	289
2750,6	22	0,53837	14	8	64
3055,1	23	0,25425	10	13	169
3811,2	24	1,36929	21	3	9
5125,4	25	1,38299	22	3	9
Σ					2688

Коэффициент ранговой корреляции

$$r_s = 1 - 6 \cdot \left[\frac{2688}{25(25^2 - 1)} \right] = -0,0338$$

Статистика Стьюдента

$$t = \frac{-0,0338\sqrt{25-2}}{\sqrt{1-(-0,0338)^2}} = -17,9251$$

$$t_{\text{табл.}} = 1,714$$

$$\nu = 23, \alpha = 0,05$$

Вывод: имеет место гетероскедастичность

•Тест Парка

В основе теста лежит предположение:
*дисперсия остатков изменяется
пропорционально значению некоторого
фактора*

$$Var(u_i) = \sigma^2 Z_i^2$$

σ^2 – дисперсия гомоскедастической ошибки,
 Z – фактор пропорциональности.

3 вопроса на засыпку:

1. Имеют ли место очевидные ошибки спецификации модели?

2. Насколько часто предмет исследований «страдает» гетероскедастичностью?

3. Дает ли график отклонений какое-либо свидетельство наличия гетероскедастичности?

- 1-й шаг.

Построение уравнения регрессии и
расчет отклонений $u_i \quad (i = \overline{1, n})$

$$u_i = y_i - \hat{a}_0 - \hat{a}_1 x_1 - \dots - \hat{a}_m x_m$$

- 2-й шаг.

Использование значений отклонений для получения новой зависимой переменной (используется логарифм квадрата отклонений).

Построение вторичного уравнения регрессии

$$\ln(u_i^2) = \alpha_0 + \alpha_1 \ln Z_i + \varepsilon_i$$

- 3-й шаг.

Проверка значимости коэффициента регрессии при по критерию Стьюдента.

- Если коэффициент существенно отличается от нуля, то имеет место гетероскедастичность отклонений по отношению к независимой переменной Z ;
- в противном случае, гетероскедастичности, относящейся к данному конкретному Z , скорее всего, не наблюдается

Пример

С целью выбора места строительства нового круглосуточно работающего кафе необходимо определить зависимость его дохода (y) от трех факторов: количества прямых конкурентов, находящихся в радиусе двух километров от кафе (x_1), количества людей, живущих в радиусе трех километров от кафе (x_2) и среднегодового дохода семьи (x_3) живущей в данной окрестности.

Наблюдение	Доход кафе, грн.	Число конкурентов, ед.	Численность проживающих, чел.	Доход семьи, грн.
1	107919	3	65044	13240
2	118866	5	101376	22554
3	98579	7	124989	16916
4	122015	2	55249	20967
5	152827	3	73775	19576
6	91259	5	48484	15039
7	123550	8	138809	21857
8	160931	2	50244	26435
9	98496	6	104300	24024
10	108052	2	37852	14987
11	144788	3	66921	30902
12	164571	4	166332	31573
13	105564	3	61951	19001
14	102568	5	100441	20058
15	103324	2	39462	16194
16	127030	5	139900	21384
17	166755	6	171740	18800
18	125343	6	149894	15289
19	121886	3	57386	16702
20	134594	6	185105	19093
21	152937	3	114520	26502
22	109622	3	52933	18760
23	149884	5	203500	33242
24	98388	4	39334	14988
25	140791	3	95120	18505
26	101260	3	49200	16839
27	139517	4	113566	28915
28	115236	9	194125	19033
29	136749	7	233844	19200
30	105067	7	83416	22833
31	136872	6	183953	14409
32	117146	3	60457	20307
33	163538	2	65065	20111

1-й шаг.

Строим уравнение множественной регрессии

$$\hat{y} = 102188 - 9074x_1 + 0,355x_2 + 1,288x_3$$

Показатели значимости модели и ее параметров:

$$F = 15,65$$

$$t_2 = 4,88$$

$$t_3 = 2,37$$

$$t_1 = -4,42$$

$$\bar{R}^2 = 0,579$$

Рассчитаем
остатки

Наблюдаемое значение y	Предсказанное значение y	Остатки
107919	115087,8	-7168,786
118866	121821,6	-2955,611
98579	104785,9	-6206,934
122015	130640,6	-8625,642
152827	126345,3	26481,71
91259	93383,01	-2124,01
123550	106977,3	16572,75
160931	135908,4	25022,58
98496	115677,7	-17181,75
108052	116768,1	-8716,113
144788	138502,5	6285,495
164571	165550,4	-979,4206
105564	121411,1	-15847,08
102568	118275,1	-15707,1
103324	118893,8	-15569,77
127030	133977,9	-6947,881
166755	132868	33887,02
125343	120597,7	4745,348
121886	116830,9	5055,142
134594	137985,5	-3391,515
152937	149717,1	3219,912
109622	117902,3	-8280,264
149884	171808,1	-21924,14
98388	99146,41	-758,4093
140791	132536,2	8254,777
101260	114104	-12844
139517	143412,4	-3895,421
115236	113884,4	1351,572
136749	146335,2	-9586,22
105067	97662,51	7404,49
136872	131543,9	5328,148
117146	122563,4	-5417,357
163538	133019,5	30518,48

2-й шаг.

Преобразуем остатки – рассчитаем логарифмы их квадратов.

Фактор пропорциональности – размер рынка

$$Z = x_2$$

u_i	-7168,786	u_i^2	51391485,68	$\ln x_{2i}$	17,75498	$\ln(u_i^2)$	11,08282
	-2955,611		8735636,01		15,98292		11,52659
	-6206,934		38526023,5		17,46684		11,73598
	-8625,642		74401693,06		18,12499		10,91961
	26481,71		701280919		20,36842		11,20878
	-2124,01		4511420,475		15,32212		10,78899
	16572,75		274655904,8		19,43103		11,84085
	25022,58		626129613,6		20,25507		10,82465
	-17181,75		295212383,8		19,50321		11,55503
	-8716,113		75970631,17		18,14586		10,54144
	6285,495		39507443,21		17,492		11,11127
	-979,4206		959264,7604		13,77392		12,02174
	-15847,08		251129881,2		19,34148		11,0341
	-15707,1		246712934,4		19,32374		11,51733
	-15569,77		242417729,1		19,30617		10,58309
	-6947,881		48273043,93		17,69238		11,84868
	33887,02		1148330143		20,86157		12,05374
	4745,348		22518328,46		16,92984		11,91768
	5055,142		25554458,54		17,05632		10,95756
	-3391,515		11502375,9		16,25806		12,12868
	3219,912		10367831,76		16,15422		11,6485
	-8280,264		68562769,46		18,04326		10,87678
	-21924,14		480668027,2		19,99069		12,22342
	-758,4093		575184,6448		13,26245		10,57984
	8254,777		68141348,59		18,03709		11,46289
	-12844		164968331,8		18,92126		10,80365
	-3895,421		15174303,91		16,53511		11,64014
	1351,572		1826745,66		14,41805		12,17626
	-9586,22		91895622,88		18,33616		12,36241
	7404,49		54826477,02		17,81968		11,3316
	5328,148		28389161,01		17,16152		12,12244
	-5417,357		29347757,64		17,19473		11,00969
	30518,48		931377408		20,65218		11,08314

Уравнение регрессии

$$\ln(u^2) = 21,05 - 0,286 \ln x_2$$

Параметры качества уравнения

$$F = 0,209 \quad t = -0,457 \quad \bar{R}^2 = 0,0067$$

$$t_{\text{табл.}} = 2,042$$

Вывод: нулевая гипотеза о
гомоскедастичности остатков принимается