

**Обобщенная линейная модель
множественной регрессии с
автокоррелированными остатками**

Автокорреляция регрессионных остатков – корреляционная зависимость текущих и предыдущих значений регрессионных остатков.

Линейная модель множественной регрессии

$$Y = X\bar{\beta} + \bar{\varepsilon},$$

для которой нарушено 5 условие Гаусса-Маркова называется обобщенной линейной моделью множественной регрессии (ОЛММР) с автокоррелированными остатками, а именно:

- 1) x_1, \dots, x_k – детерминированные переменные;
 - 2) ранг матрицы X равен " $k+1$ " – среди признаков нет линейно зависимых;
 - 3) $M\varepsilon_i = 0, i = \overline{1, n}$ - нет систематических ошибок в измерении y ;
 - 4) $D\varepsilon_i = M\varepsilon_i^2 = \sigma^2, i = \overline{1, n}$
 - 5) $\text{cov}(\varepsilon_i, \varepsilon_j) = M(\varepsilon_i \cdot \varepsilon_j) \neq 0, i \neq j, i = \overline{1, n} j = \overline{1, n}$
- 4') $\Sigma_\varepsilon = M\overline{\varepsilon\varepsilon^T} = \sigma^2\Sigma_0 \quad (\sigma_i^2 = \sigma_j^2),$ т.е. на диагонали стоят равные

дисперсии.

Виды автокорреляции

Объем продаж
мороженого

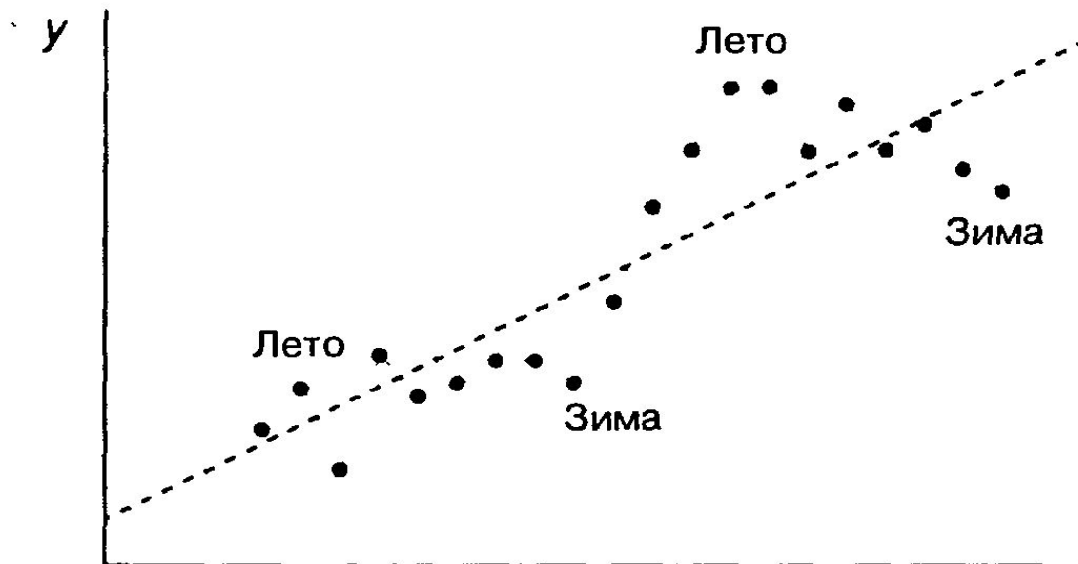


Рис. 1 Положительная автокорреляция

Доход, x

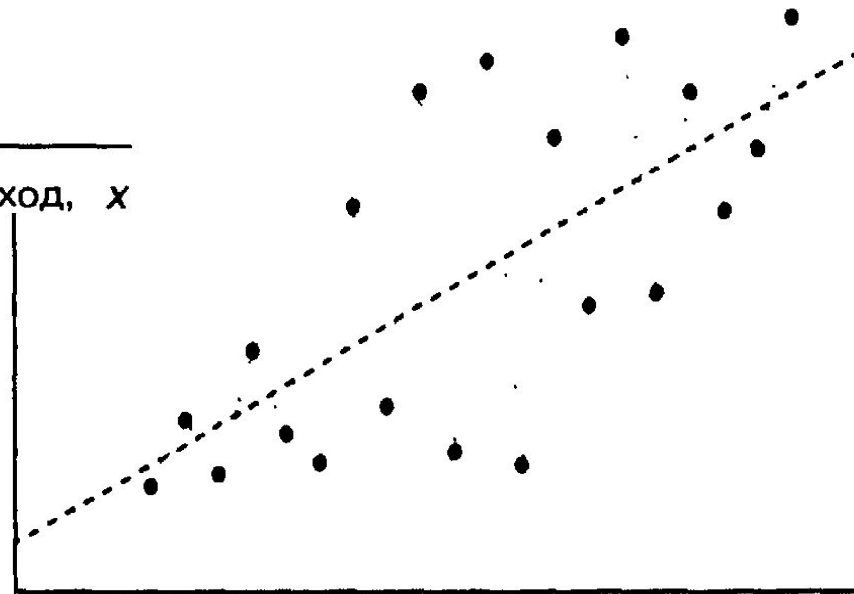


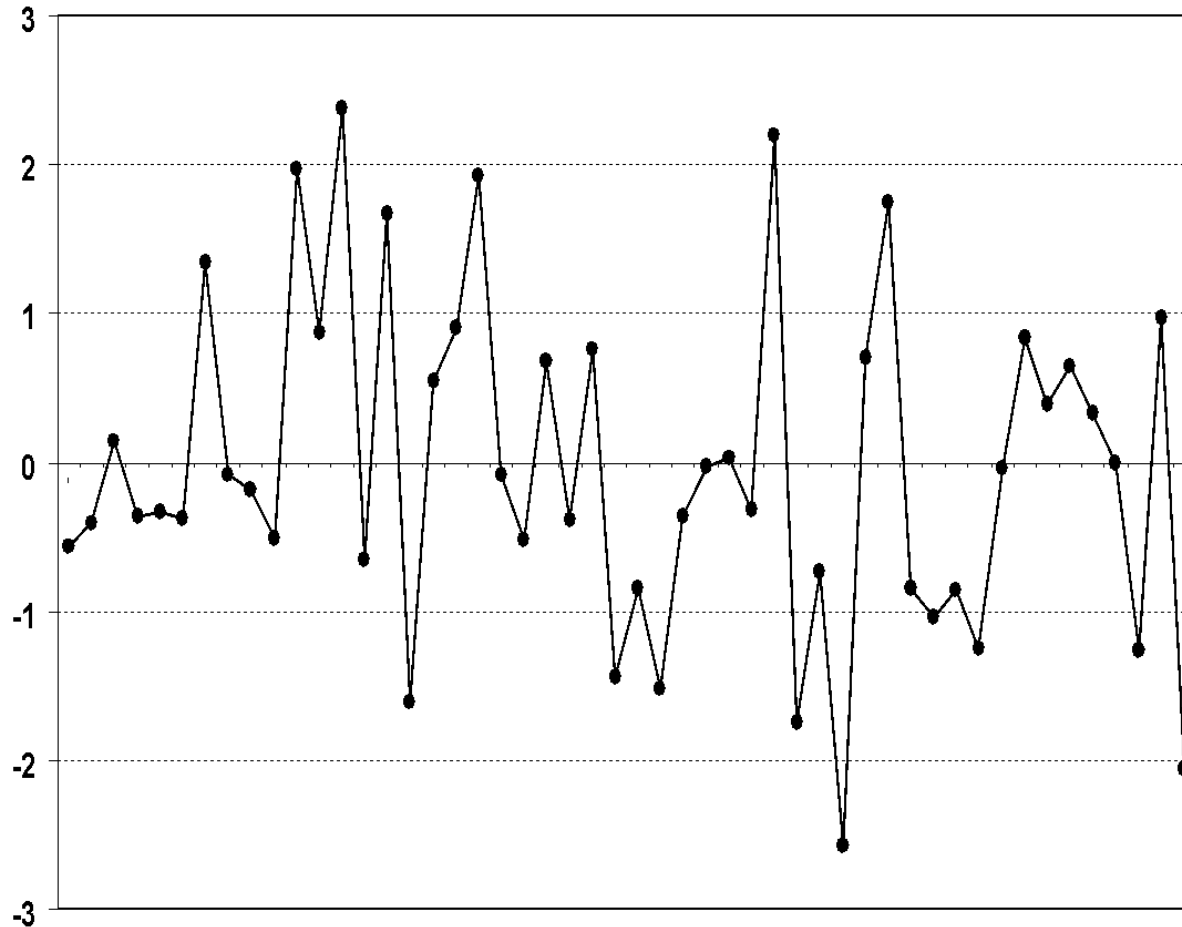
Рис. 2 Отрицательная автокорреляция

Пример влияния автокорреляции на случайную выборку

Рассмотрим выборку из 50 независимых нормально распределенных с нулевым средним значений ε_i .

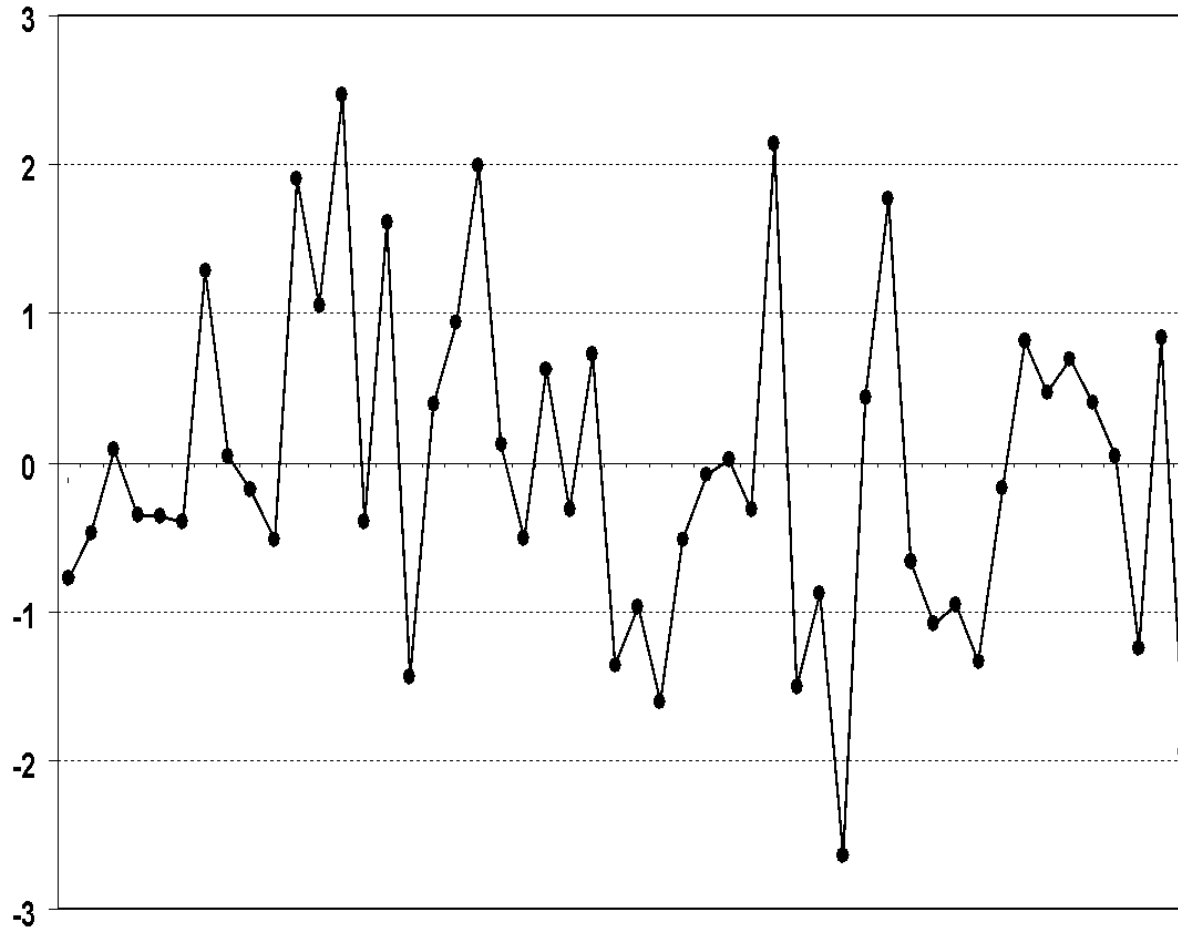
С целью ознакомления с влиянием автокорреляции будем вводить в нее положительную, а затем отрицательную автокорреляцию.

Пример влияния автокорреляции на случайную выборку



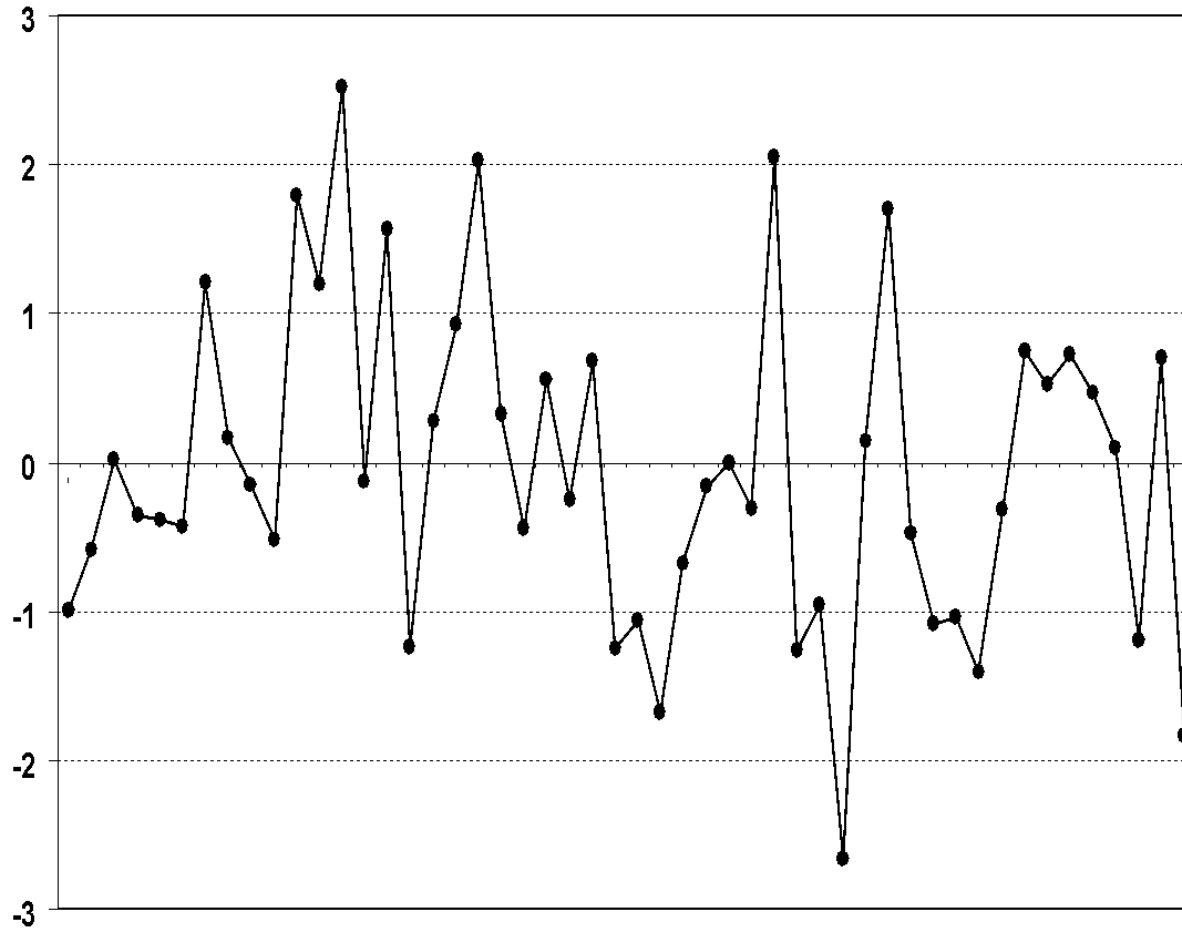
$$u_t = 0.0u_{t-1} + \varepsilon_t$$

Пример влияния автокорреляции на случайную выборку



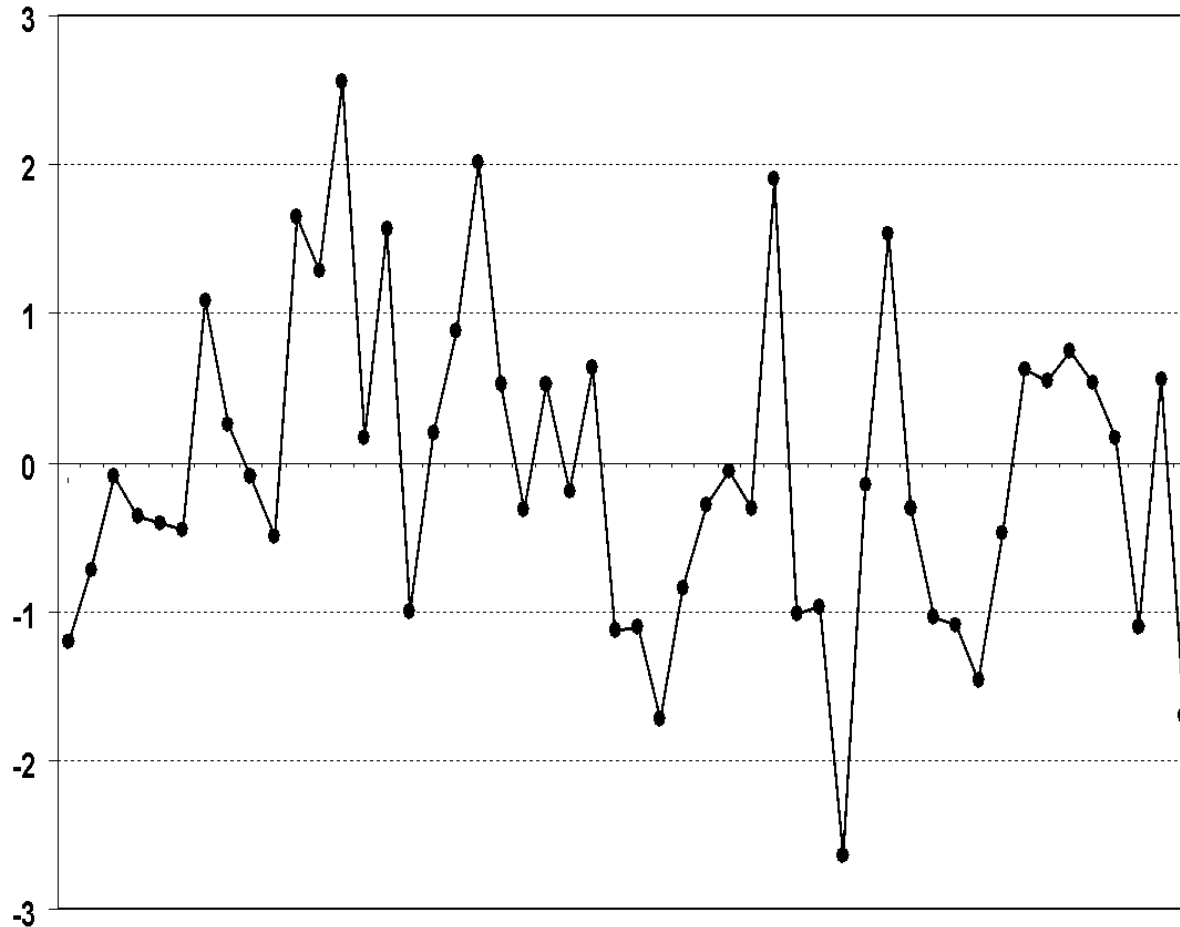
$$u_t = 0.1u_{t-1} + \epsilon_t$$

Пример влияния автокорреляции на случайную выборку



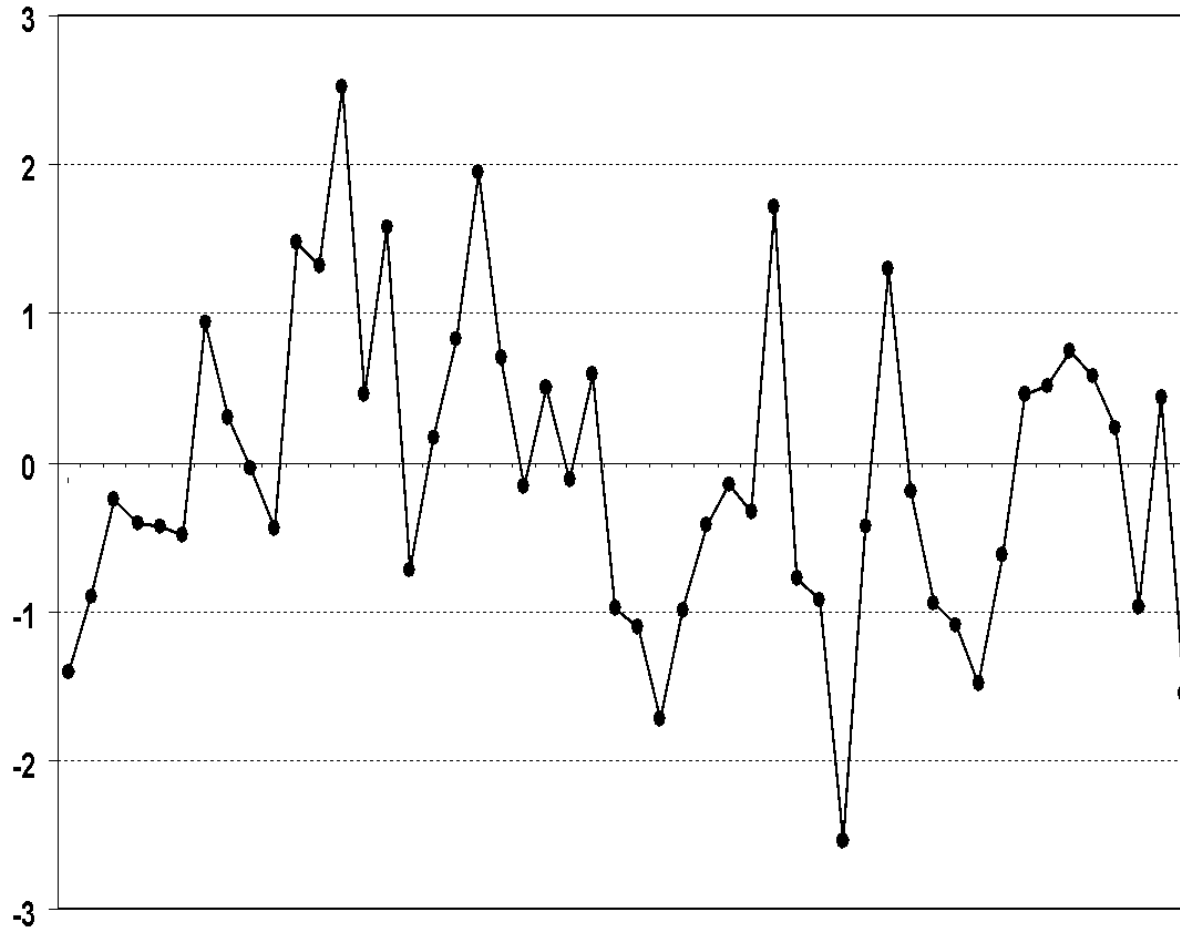
$$u_t = 0.2u_{t-1} + \varepsilon_t$$

Пример влияния автокорреляции на случайную выборку



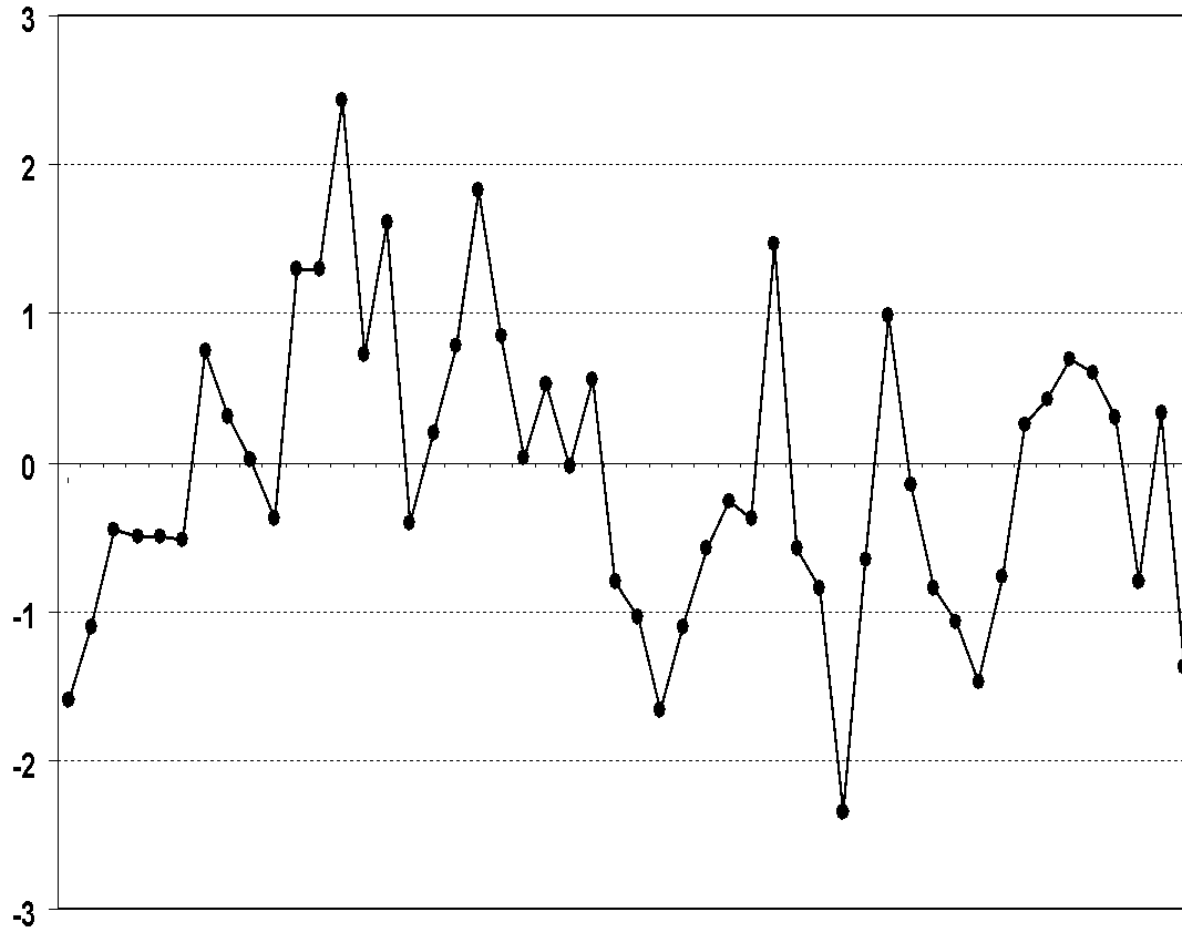
$$u_t = 0.3u_{t-1} + \varepsilon_t$$

Пример влияния автокорреляции на случайную выборку



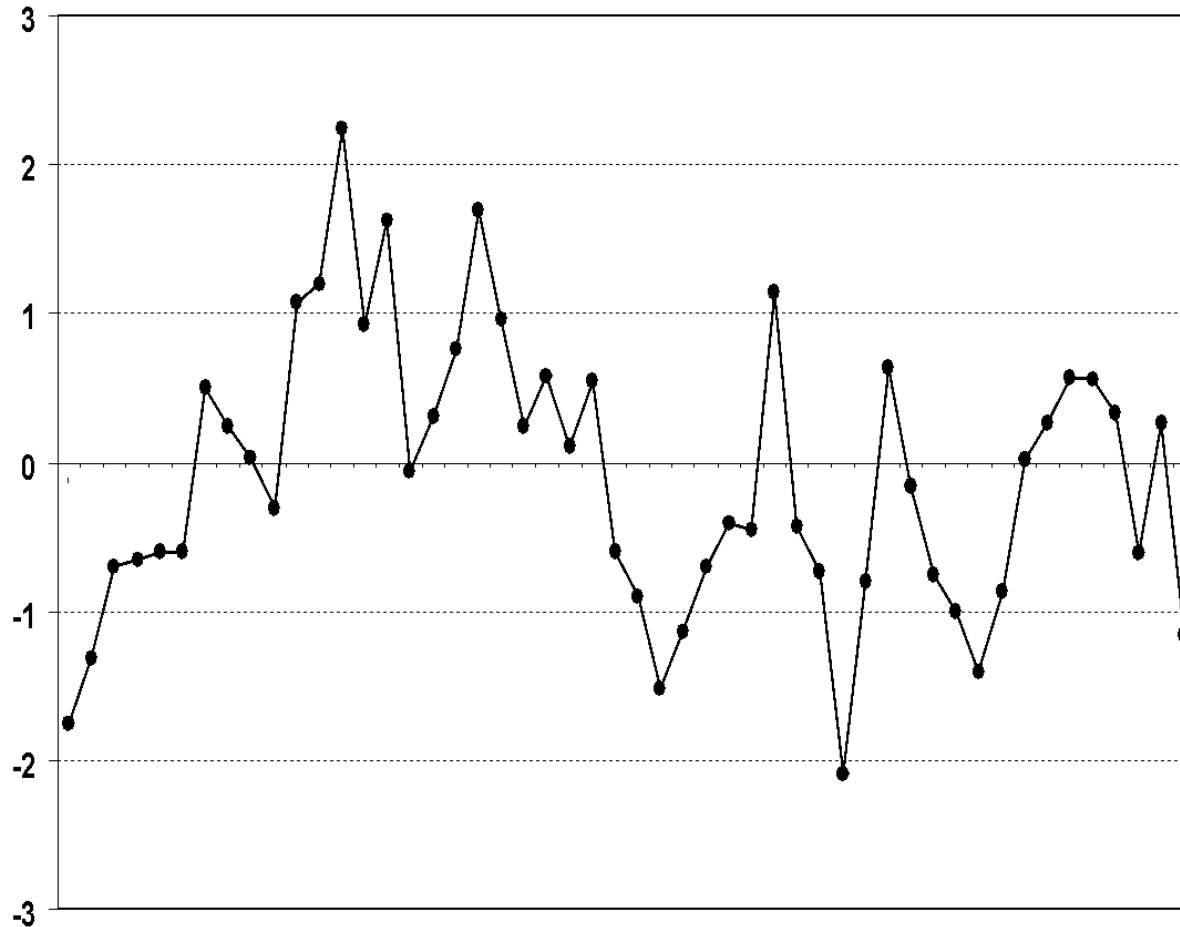
$$u_t = 0.4u_{t-1} + \varepsilon_t$$

Пример влияния автокорреляции на случайную выборку



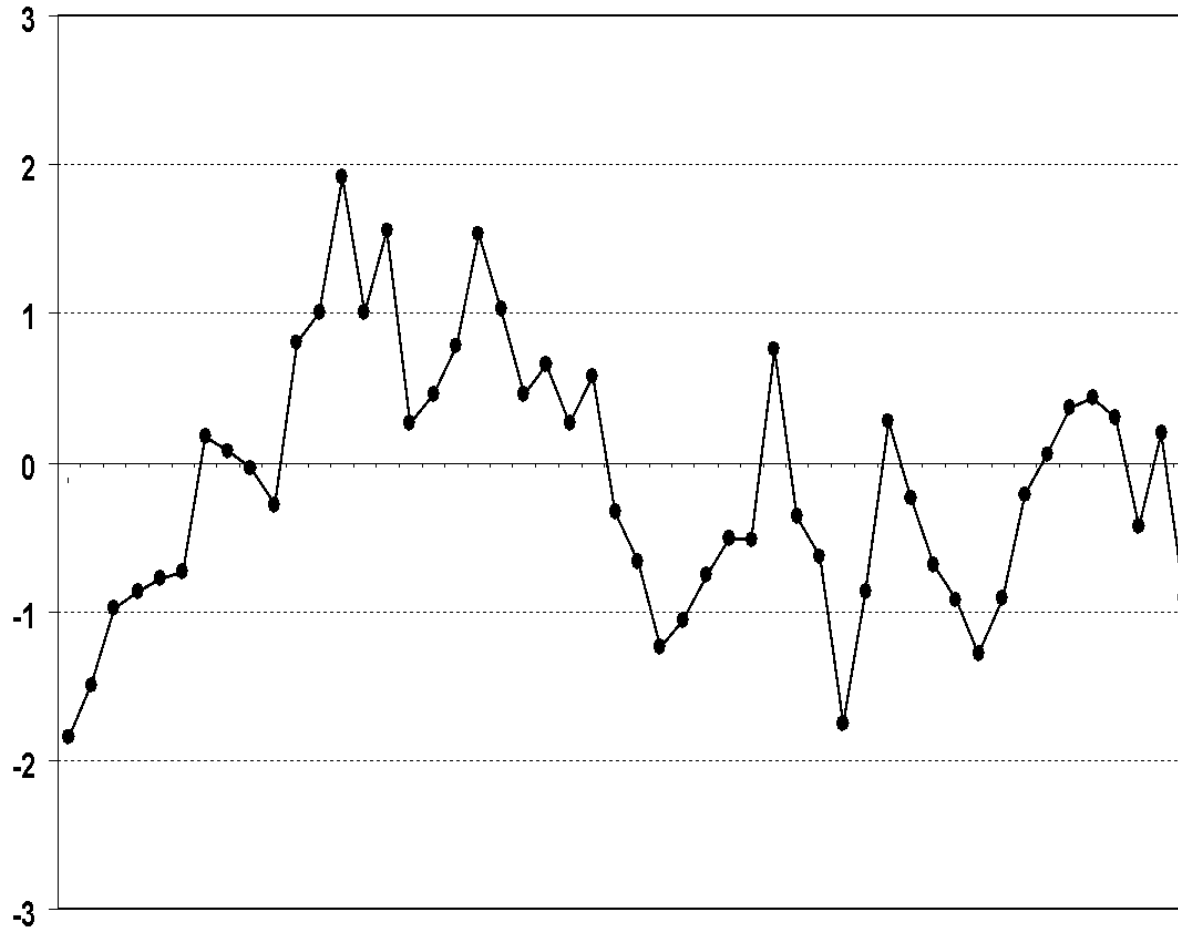
$$u_t = 0.5u_{t-1} + \varepsilon_t$$

Пример влияния автокорреляции на случайную выборку



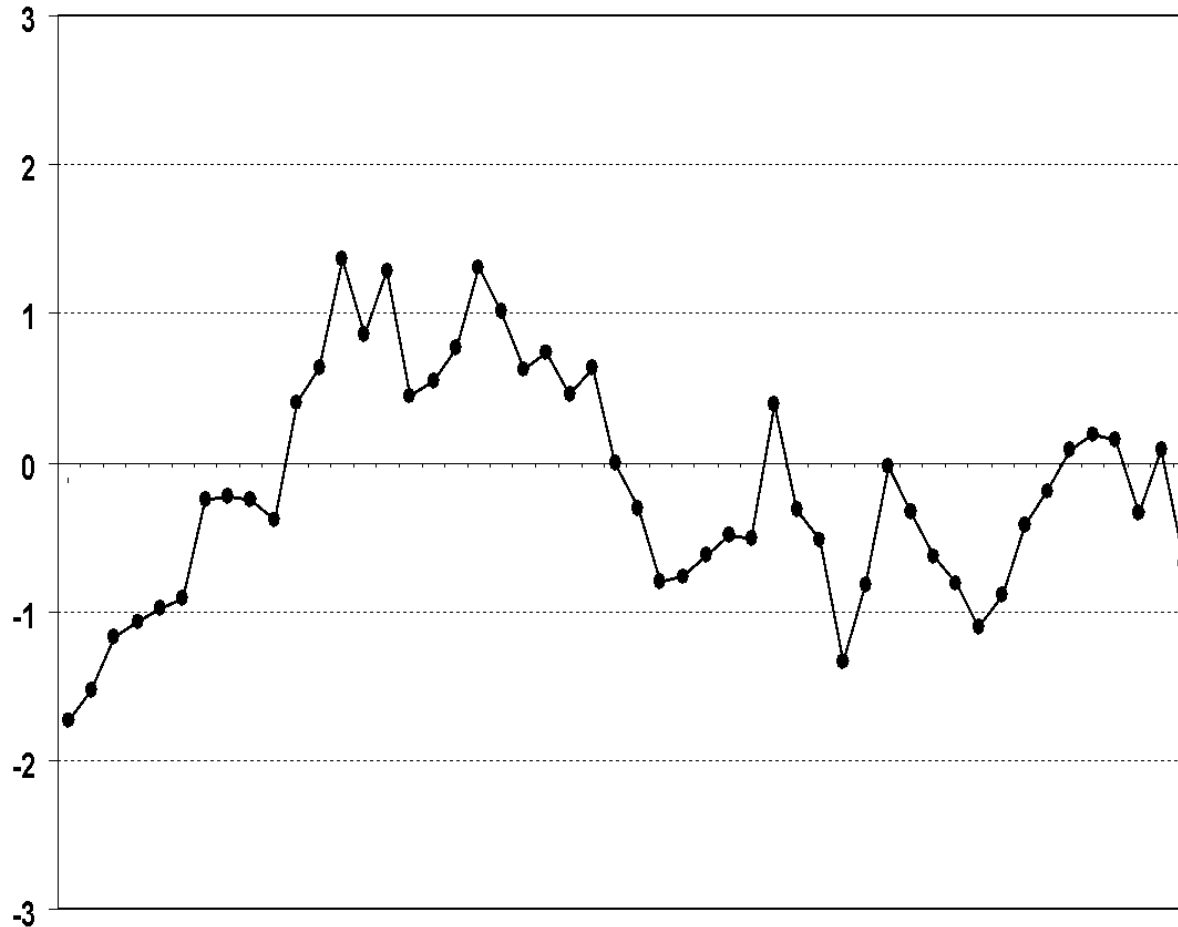
$$u_t = 0.6u_{t-1} + \varepsilon_t$$

Пример влияния автокорреляции на случайную выборку



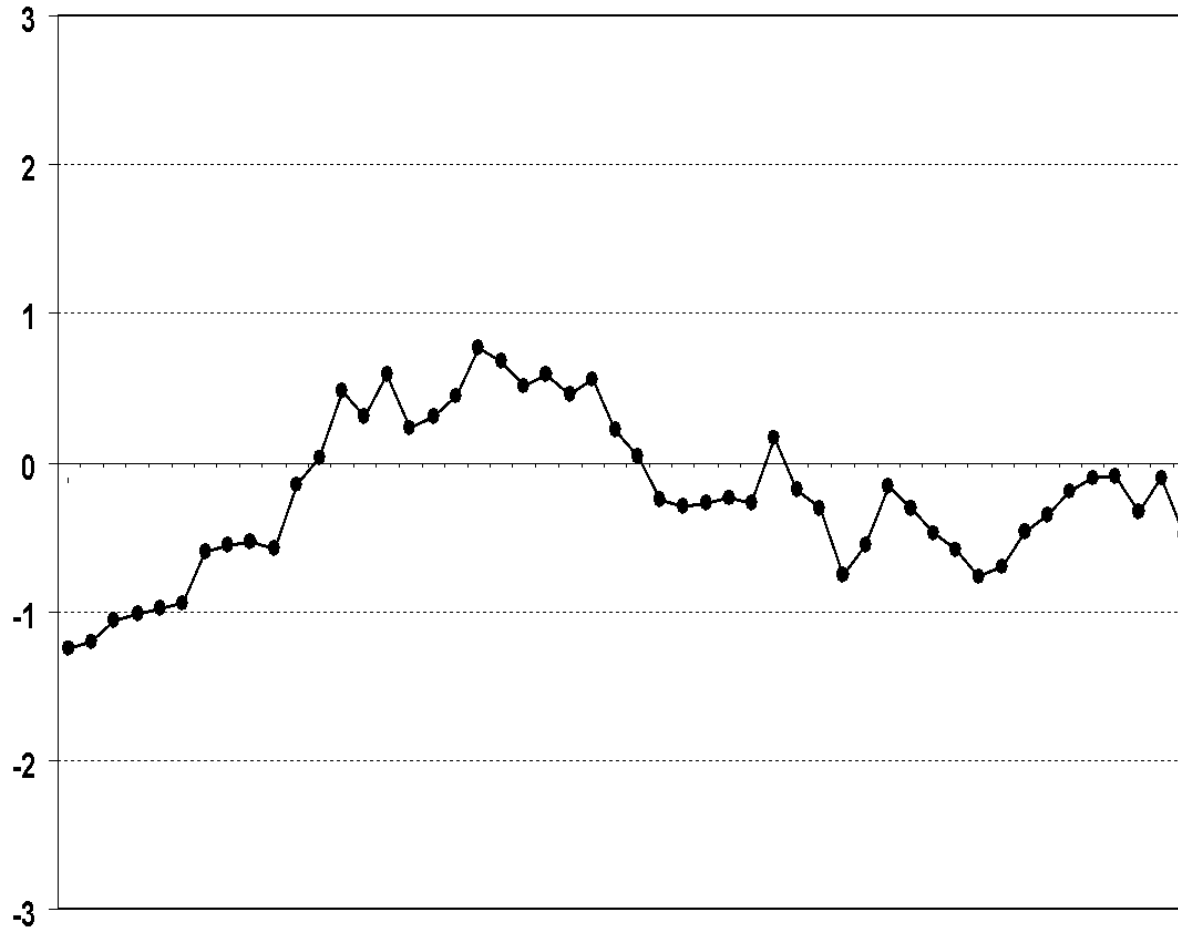
$$u_t = 0.7u_{t-1} + \varepsilon_t$$

Пример влияния автокорреляции на случайную выборку



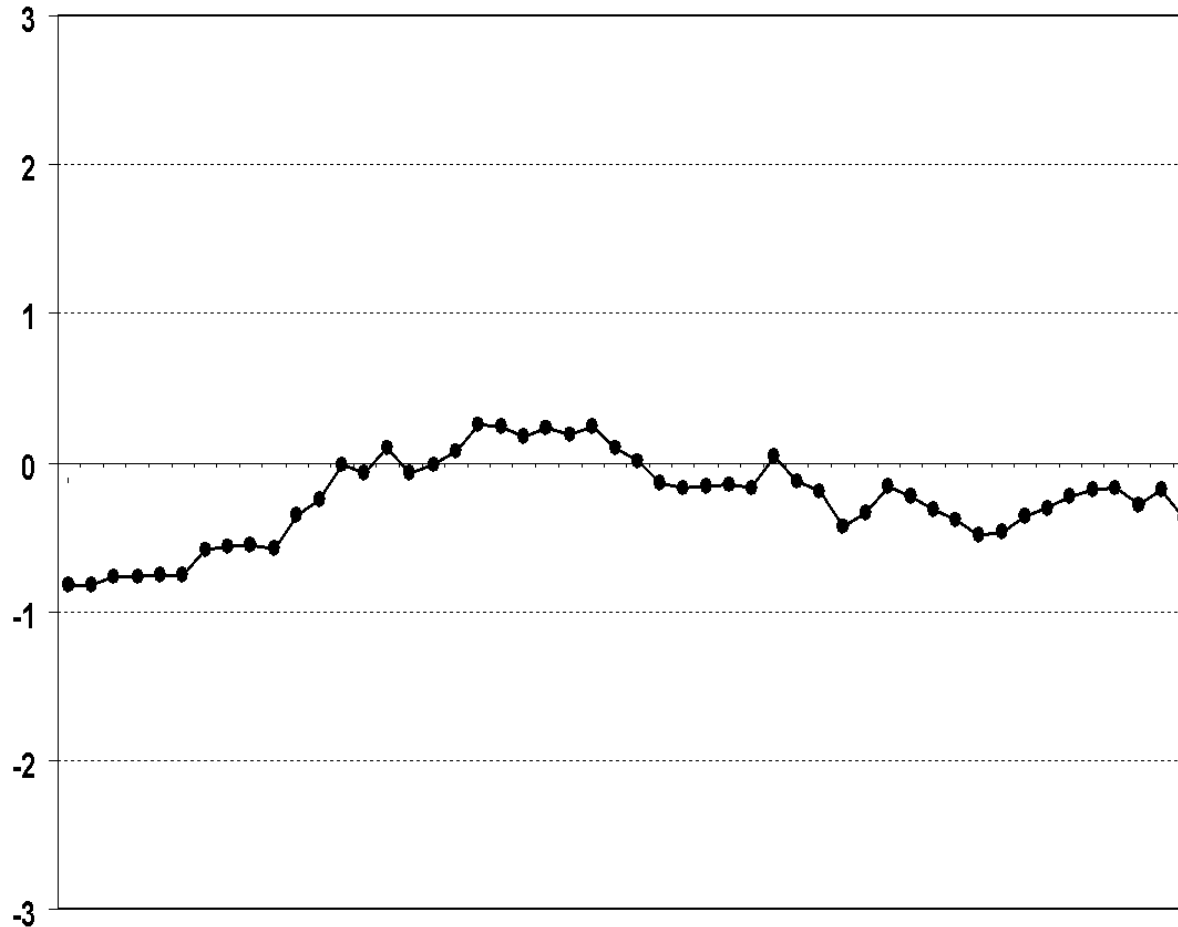
$$u_t = 0.8u_{t-1} + \varepsilon_t$$

Пример влияния автокорреляции на случайную выборку



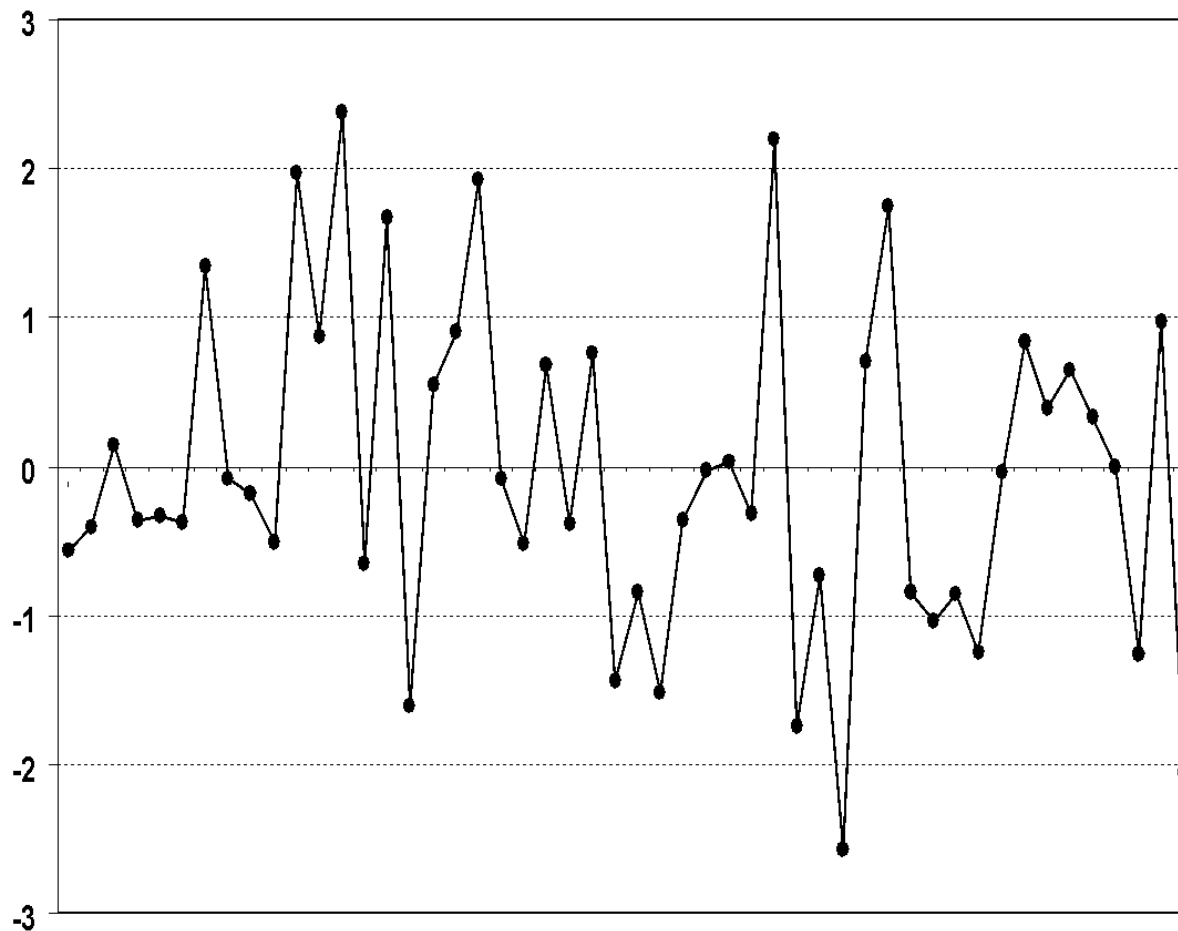
$$u_t = 0.9u_{t-1} + \varepsilon_t$$

Пример влияния автокорреляции на случайную выборку



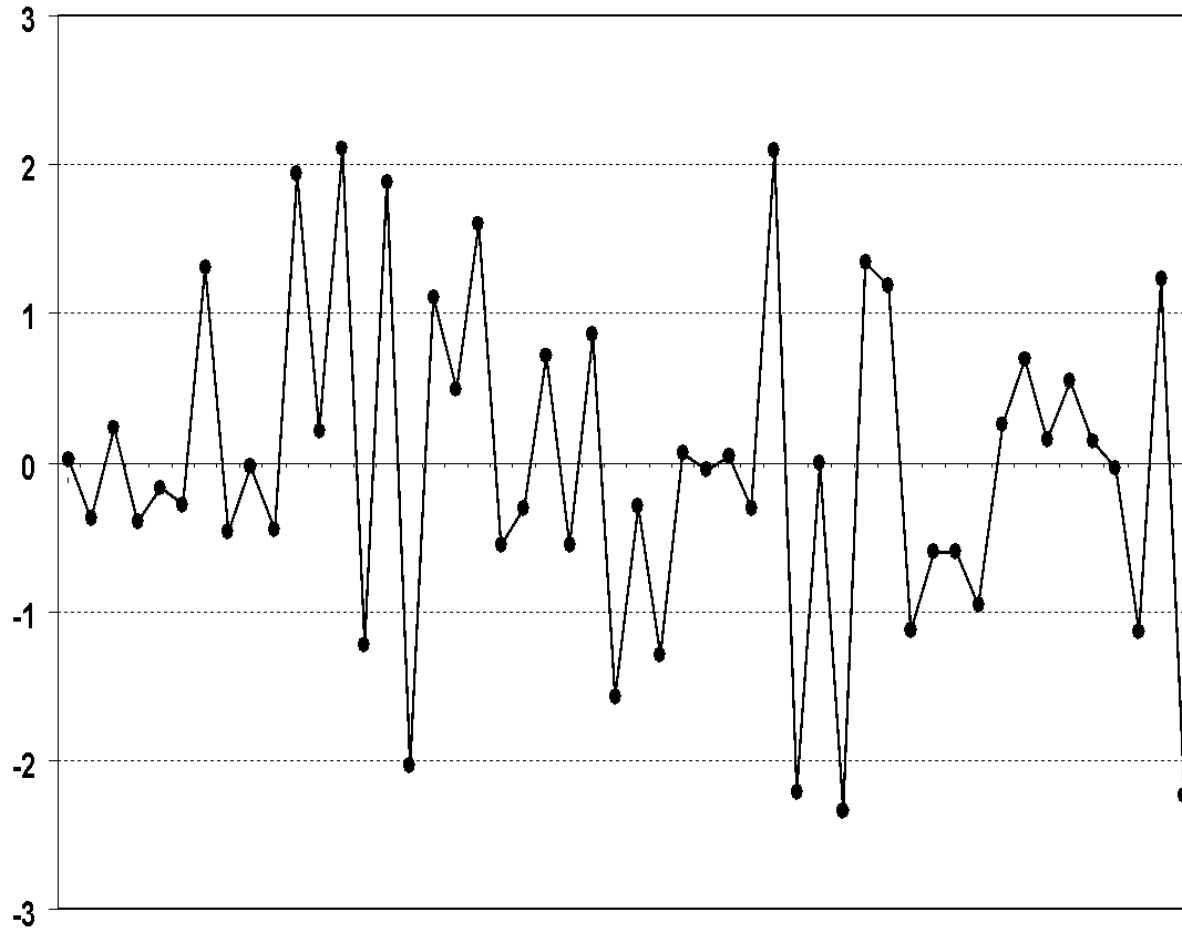
$$u_t = 0.95u_{t-1} + \varepsilon_t$$

Пример влияния автокорреляции на случайную выборку



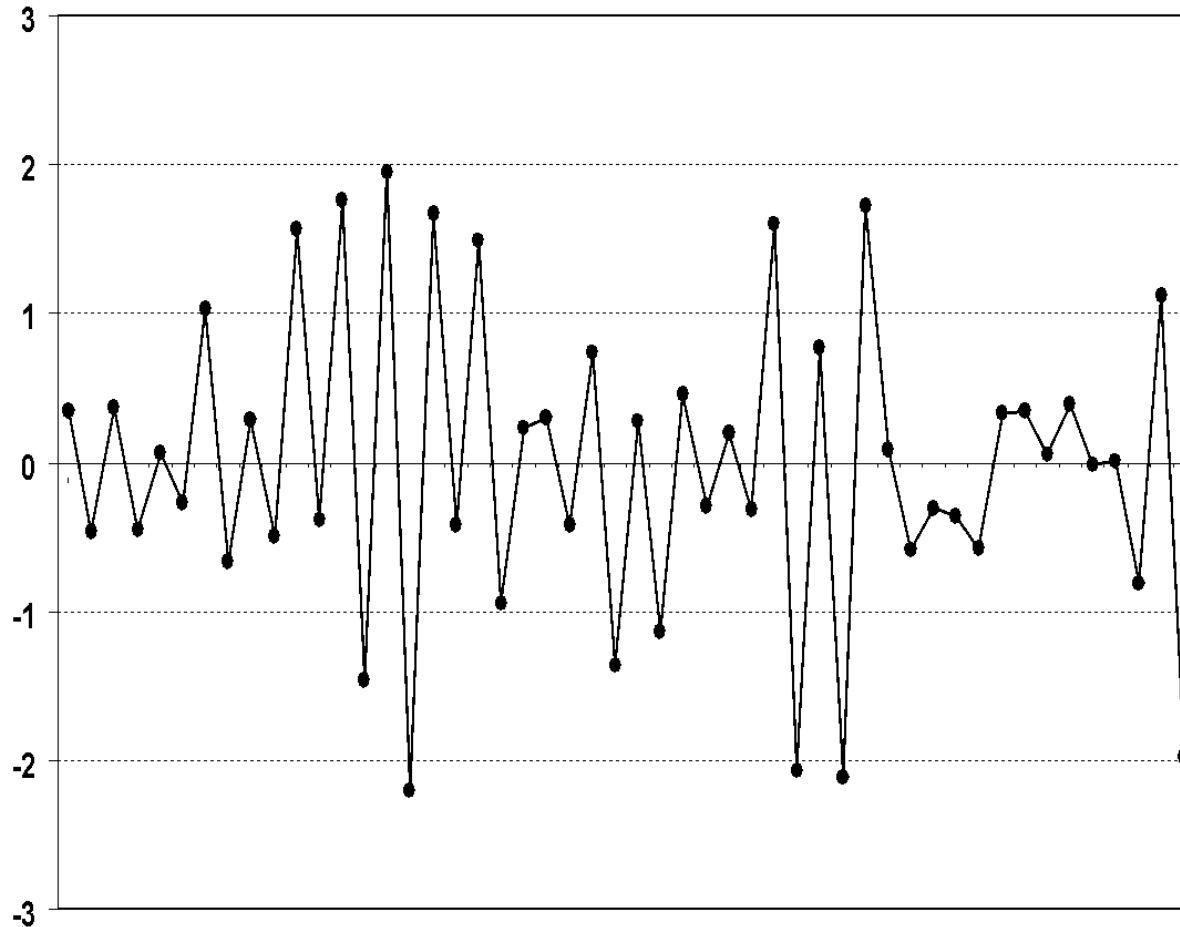
$$u_t = 0.0u_{t-1} + \varepsilon_t$$

Пример влияния автокорреляции на случайную выборку



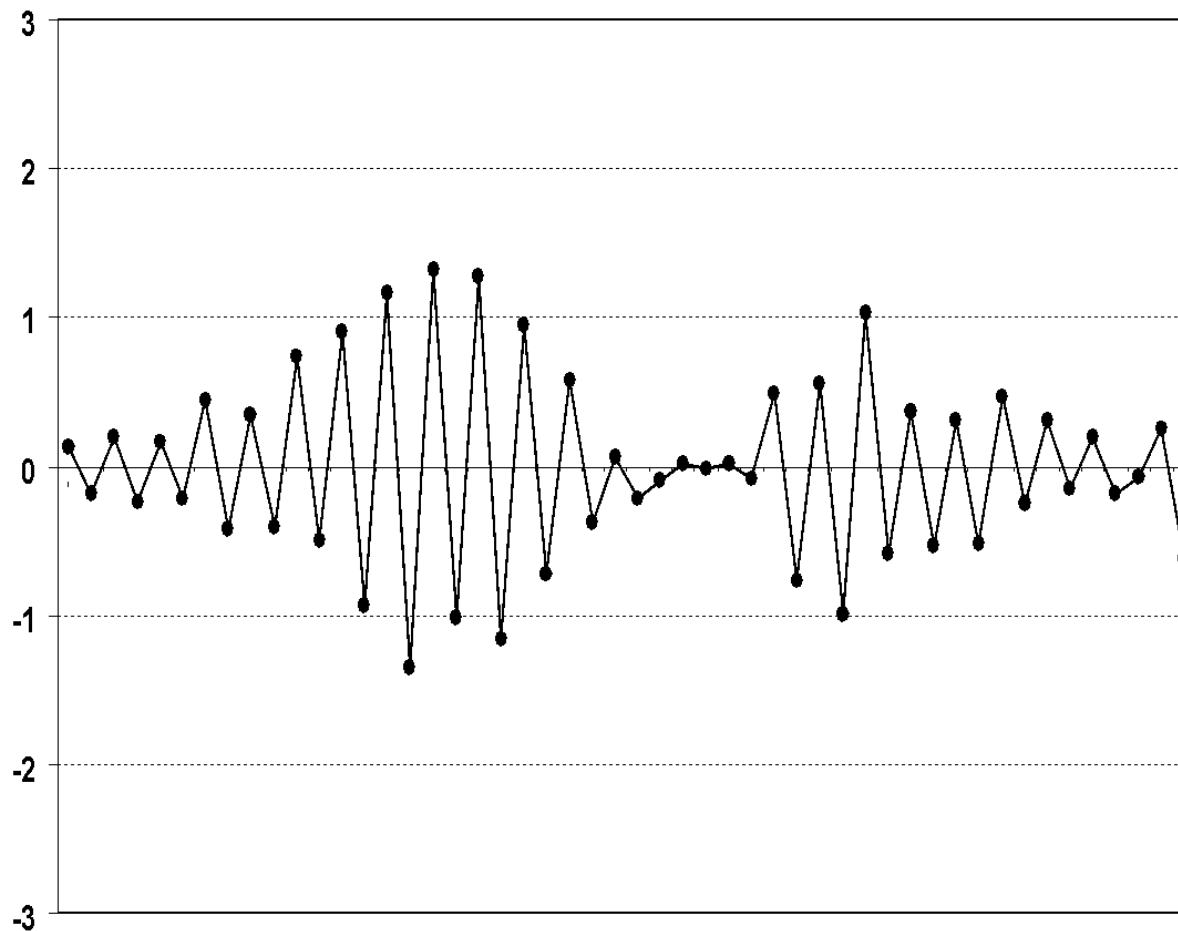
$$u_t = -0.3u_{t-1} + \varepsilon_t$$

Пример влияния автокорреляции на случайную выборку



$$u_t = -0.6u_{t-1} + \varepsilon_t$$

Пример влияния автокорреляции на случайную выборку



$$u_t = -0.9u_{t-1} + \varepsilon_t$$

Автокорреляция первого порядка

$$\varepsilon_i = \rho \varepsilon_{i-1} + \delta_i,$$

ρ - коэффициент автокорреляции первого порядка ($|\rho| \leq 1$).

δ_i - случайные величины, удовлетворяющие

условиям
$$M[\delta_i] = 0, M[\delta_i, \delta_j] = \begin{cases} \sigma_\delta^2, & \text{при } i = j; \\ 0, & \text{при } i \neq j \end{cases}$$

.

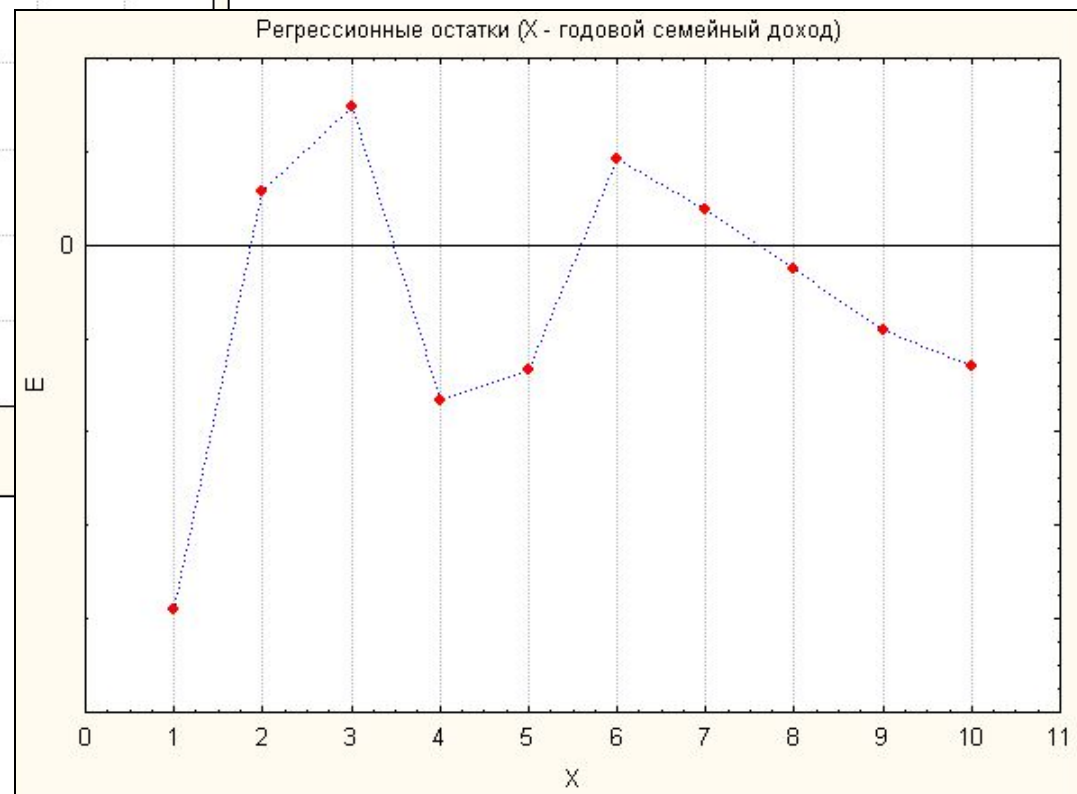
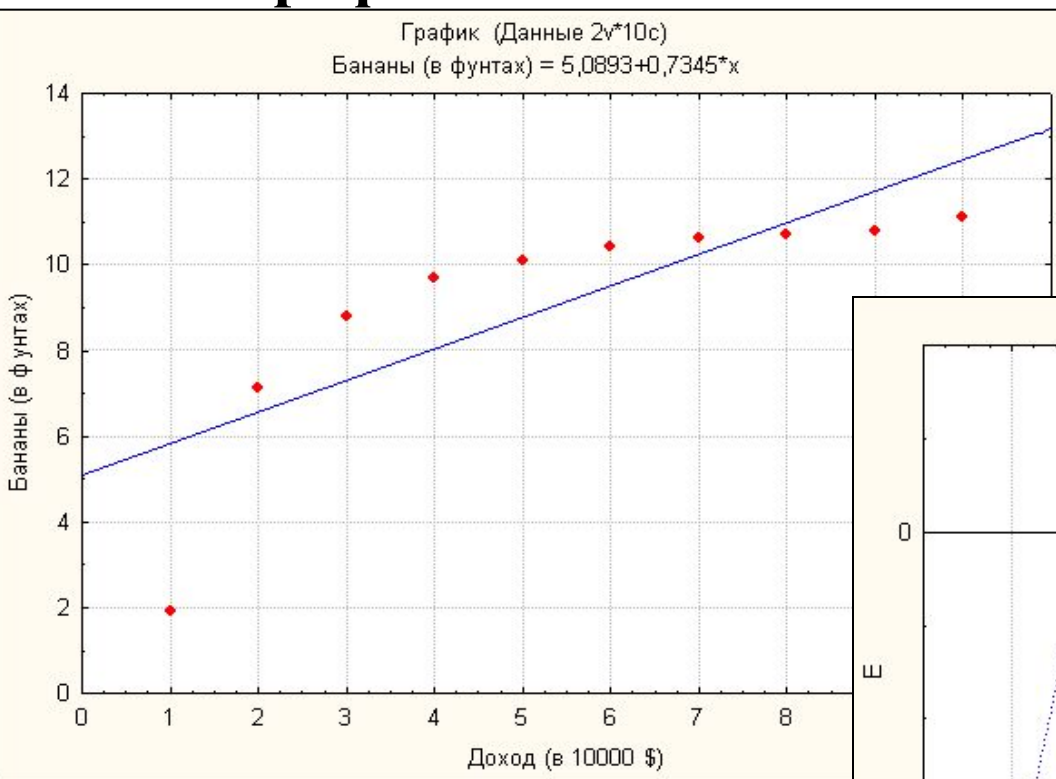
Автокорреляция первого порядка

$$\Sigma_{\varepsilon} = \frac{\sigma_0^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix},$$

$$\Sigma_0^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho & 0 & \dots & 0 \\ -\rho & 1+\rho^2 & -\rho & \dots & 0 \\ 0 & -\rho & 1-\rho^2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

Обнаружение автокорреляции

1. Графический анализ статистической информации



Обнаружение автокорреляции

2. Статистика Дарбина-Уотсона

$H_0: \rho = 0$ (нет явления автокорреляции)

$H_1: \rho \neq 0$ (есть явление автокорреляции)

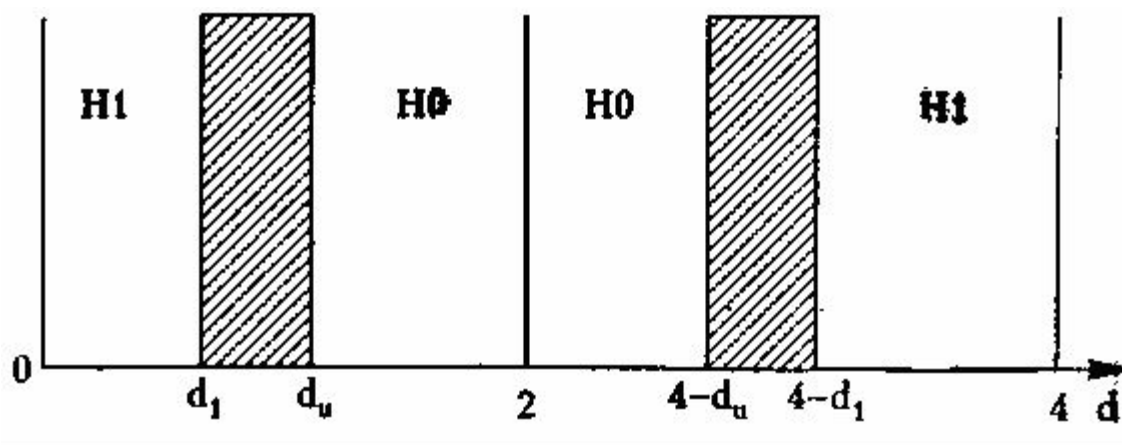
$$DW = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^{n-1} e_i^2} .$$

Так как

$$\sum_{i=2}^n (e_i - e_{i-1})^2 = \sum_{i=2}^n (e_i^2 - 2e_i e_{i-1} + e_{i-1}^2) = \sum_{i=2}^n e_i^2 - 2 \sum_{i=2}^n e_i e_{i-1} + \sum_{i=2}^n e_{i-1}^2 \approx 2 \sum_{i=2}^n e_i^2 - 2 \sum_{i=2}^n e_i e_{i-1}$$

$$DW = \frac{2(\sum e_i^2 - \sum e_i e_{i-1})}{\sum e_i^2} = 2(1 - r_{e_i e_{i-1}})$$

1. $d_6 < DW < 4 - d_6$, то гипотеза об отсутствии автокорреляции принимается.
2. $d_n < DW < d_6$ или $4 - d_6 < DW < 4 - d_n$, область неопределенности критерия (вопрос об отвержении или принятии гипотезы остается открытым).
3. $0 < DW < d_n$, то принимается альтернативная гипотеза о положительной автокорреляции.
4. $4 - d_n < DW < 4$, то принимается альтернативная гипотеза об отрицательной автокорреляции.



Пример исследования регрессионных остатков на автокорреляции

На основе наблюдений по десяти семьям требуется исследовать зависимость между ежегодным потреблением бананов и годовым семейным доходом.

Результативный признак:

Y – Потребление бананов в год (в фунтах)

Факторный признак:

X – Семейный доход (в 10000 \$)

Исходные данные

	1	2
	Бананы (в фунтах)	Доход (в 10000 \$)
1	1,93	1
2	7,13	2
3	8,78	3
4	9,69	4
5	10,09	5
6	10,42	6
7	10,62	7
8	10,71	8
9	10,79	9
10	11,13	10

Оценка функции регрессии

Regression Summary for Dependent Variable: Бананы (в фунт)						
R= ,79547676 R?= ,63278328 Adjusted R?= ,58688119						
F(1,8)=13,786 p<,00593 Std.Error of estimate: 1,7968						
N=10	Beta	Std.Err. of Beta	B	Std.Err. of B	t(8)	p-level
Intercept			5,089333	1,227445	4,146283	0,003225
Доход (в 10000 \$)	0,795477	0,214248	0,734485	0,197821	3,712883	0,005932

$$\hat{y} = 5,089 + 0,734 \cdot x$$

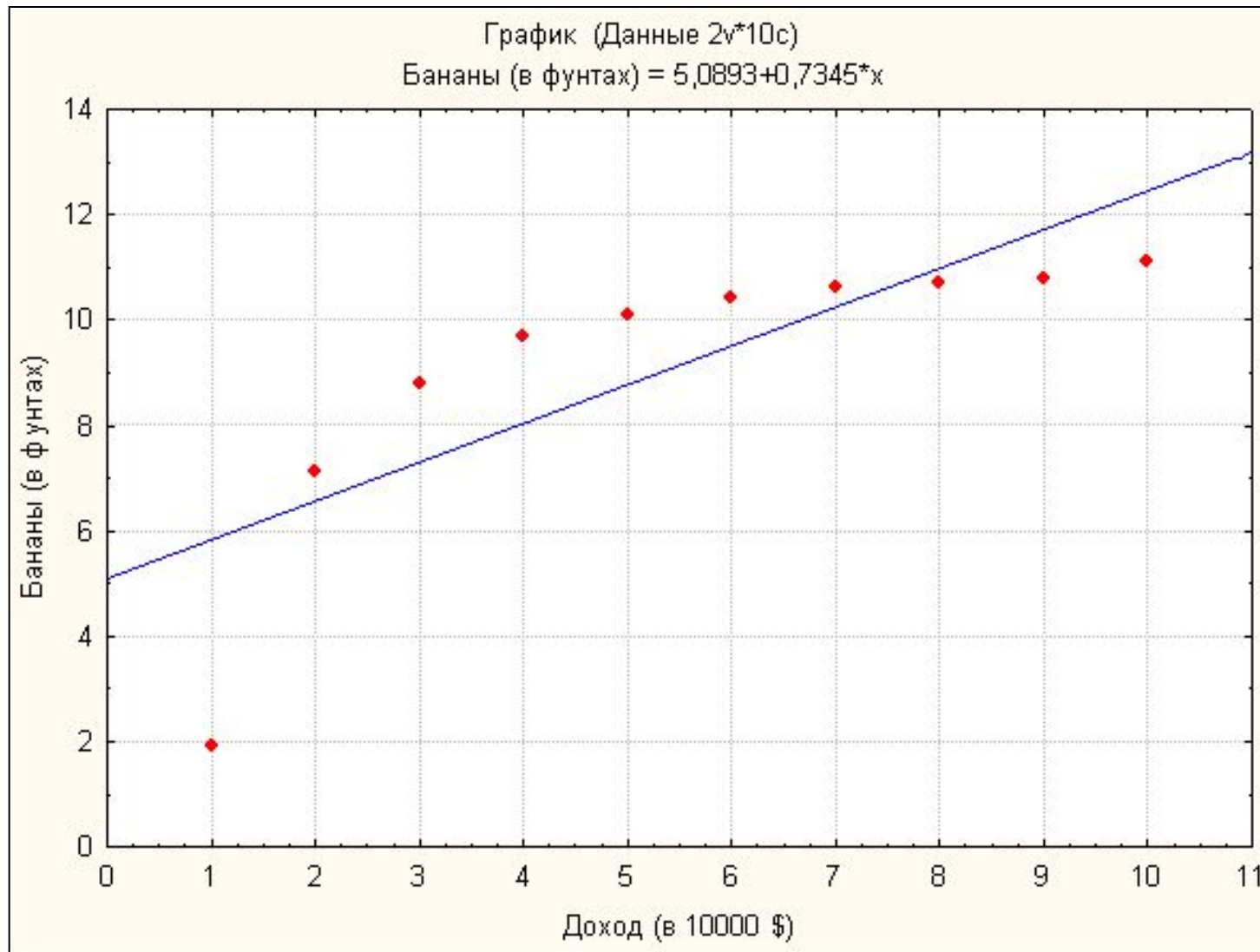
(1,23) (0.198)

$$\hat{R}^2 = 0,637$$

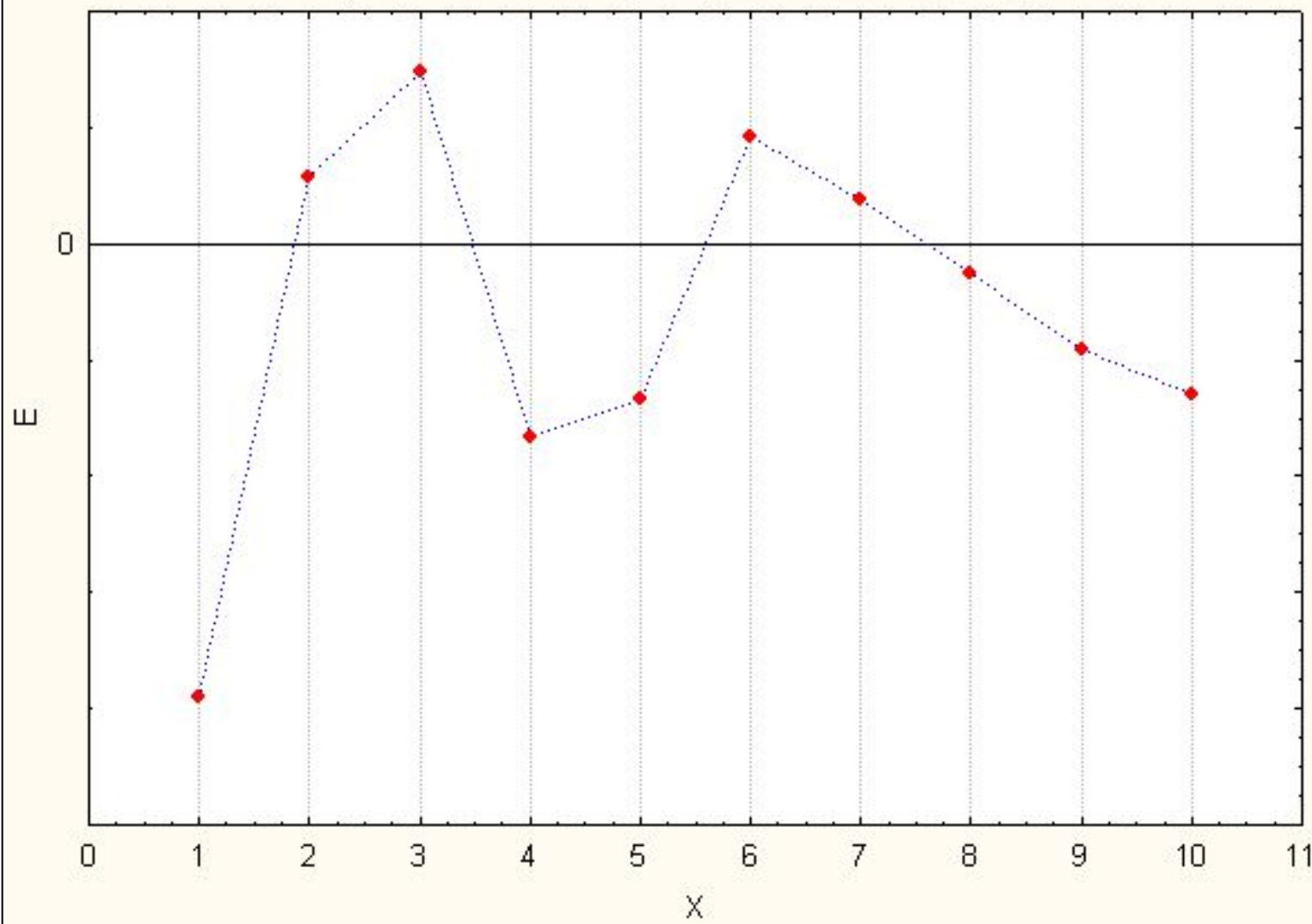
$$F_{набл} = 13,786$$

$$F_{кр} (0,05;1;8) = 5,317$$

График разброса наблюдённых значений относительно линии регрессии



Регрессионные остатки (X - годовой семейный доход)



Вычисление статистики Дарбина-Уотсона

$$D = \frac{\sum_{i=2}^N (e_i - e_{i-1})^2}{\sum_{i=1}^N e_i^2}$$

	Durbin-Watson d	Serial Corr.
Estimate	0,866133	0,257442

$$D_{набл} = 0,866$$

$$D_L = 1,08$$

$$D_U = 1,36$$

Рекламная пауза

Контрольная работа

- У группы P06-201 – 30 марта
- У группы P06-203-204 – 25 марта
- У группы P06-205 – 6 апреля
- У группы P06-206 – 30 марта

Дополнительные тесты

Тест серий (Runs test [Geary test])

Серия – последовательность подряд идущих регрессионных остатков одного знака (даже единичной длины). Длина серии – количество элементов серии. Количество серий является случайной величиной со своим математическим ожиданием (средним числом серий) и дисперсией.

При **отрицательной автокорреляции** количество серий будет велико, а при **положительной** – слишком мало.

Тест серий

Определяем:

K - количество серий

N^+ - количество неотрицательных остатков (суммарная длина положительных серий)

N^- - количество отрицательных остатков (суммарная длина отрицательных серий)

N - всего объем выборки, $N = N^+ + N^-$

Выдвигаем нулевую гипотезу H_0 : нет автокорреляции первого порядка. H_1 : есть автокорреляция первого порядка.

Для проверки гипотезы используется статистика, которая в условиях справедливости H_0 имеет асимптотическое нормальное распределение:

$$U = \frac{K - \frac{2 \cdot N^+ \cdot N^- + N}{N}}{\sqrt{\frac{2 \cdot N^+ \cdot N^- \cdot (2 \cdot N^+ \cdot N^- - N)}{N^2 \cdot (N - 1)}}} \in N(0;1)$$

Для уровня
значимости
0,05

Анализируем U

$U < -1.96$

положительная
автокорреляция

$-1.96 < U < 1.96$

нет автокорреляции

$U > 1.96$

отрицательная автокорреляция

Тест серий (пример реализации)

Microsoft Excel - practice 3 auto

Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

Arial Cyr 10 Ж К У

H24 fx

	A	B	C	D	E	F
1	Year	M-объем импорта	VNP	d	e	
2	1960	23,2	506	0	15,6642	
3	1961	23,1	523,3	0	13,3891	
4	1962	25,2	563,8	0	10,3972	
5	1963	26,4	594,7	0	7,7123	6
6	1964	28,4	635,7	0	4,5575	
7	1965	32	688,1	0	1,5694	
8	1966	37,7	753	0	-0,8902	
9	1967	40,6	796,3	0	-3,4342	
10	1968	47,7	868,5	0	-5,4117	
11	1969	52,9	935,5	0	-8,6353	
12	1970	58,5	982,4	8	-8,9319	
13	1971	64	1063,4	0	-13,6158	
14	1972	75,9	1171,1	0	-15,2565	
15	1973	94,4	1306,6	0	-13,7924	
16	1974	131,9	1424,9	0	0,0000	1
17	1975	126,9	1528,1	2	-9,1408	
18	1976	155,4	1702,2	0	-2,5298	
19	1977	185,8	1899,5	0	3,0643	
20	1978	217,5	2127,6	0	6,0861	3
21	1979	260,9	2368,5	0	19,1986	
22						
23						
24		Всего	N		20,0000	
25		число серий	K		5	
26			N+		10	
27			N-		10	
28		mat o			1	
29		disp			4,736842	
30						
31		U			-2,75681	
32						

! Число серий K=5
значительно меньше
среднего числа серий =11

$$\frac{2 \cdot N^+ \cdot N^- + N}{N}$$

$$\frac{2 \cdot N^+ \cdot N^- \cdot (2 \cdot N^+ \cdot N^- - N)}{N^2 \cdot (N - 1)}$$

U < -1.96 –
положительная автокорреляция

Дополнительные тесты

Общий тест Бройша-Годфри

Этот тест следует использовать для определения максимального порядка лага в авторегрессионной схеме остатков.

Исходная модель имеет вид:
$$Y_i = \beta_0 + \sum_{j=1}^k \beta_j X_{i,j} + \varepsilon_i$$

Относительно регрессионных остатков предполагается, что
$$\varepsilon_i = \rho_1 \varepsilon_{i-1} + \rho_2 \varepsilon_{i-2} + \dots + \rho_p \varepsilon_{i-p} + \delta_i, \delta_i \sim iid N(0; \sigma^2).$$

Нас будет интересовать проверка гипотезы об одновременном равенстве нулю всех ρ_1, \dots, ρ_p - тогда автокорреляция отсутствует

Общий тест Бройша-Годфри

Шаг 1 Находим МНК-оценки исходной модели, оценки регрессионных остатков и Q_{ost1}



Шаг 2 Оцениваем регрессию
$$e_i = \beta_0 + \sum_{j=1}^k \beta_j \cdot X_{i,j} + \sum_{l=1}^p \beta_l \cdot e_{i,l} + \delta_i, \quad \delta_i \sim iid N(0; \sigma^2)$$

Находим оценку R^2 и Q_{ost2}

Тест множителей Лагранжа
(LM-тест)

Делаем
ВЫВОДЫ

Шаг 3а Используем статистику

$$F = \frac{Q_{ost1} - Q_{ost2} / p}{Q_{ost2} / (n - (k + 1) - p)} \quad p \cdot F \sim \chi_p^2$$

Шаг 3б Используем статистику

$$(n - p) \cdot \hat{R}^2 \sim \chi_p^2$$



Методы устранения автокорреляции

Процедура Кохрейна-Оркатта

1. МНК оцениваются коэффициенты $\bar{b}^{(1)}$ регрессионной модели $Y = X\bar{\beta} + \bar{\varepsilon}$;

2. Рассчитываются регрессионные остатки

первой итерации: $e_i^{(1)} = y_i - \hat{y}_i^{(1)}$, где

$$\hat{y}_i^{(1)} = b_0^{(1)} + b_1^{(1)}x_{i1} + \dots + b_k^{(1)}x_{ik};$$

Методы устранения автокорреляции

Процедура Кохрейна-Оркатта

3. Первое приближение $r^{(1)}$ оценки неизвестного параметра ρ определяется с помощью МНК-оценки коэффициента регрессии ρ в модели $e_i^{(1)} = \rho e_{i-1}^{(1)} + \delta_i^{(1)}$

$$\hat{\rho}^{(1)} = \left(\begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_{n-1} \end{pmatrix}^T \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_{n-1} \end{pmatrix} \right)^{-1} \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_{n-1} \end{pmatrix}^T \begin{pmatrix} e_2 \\ e_3 \\ \dots \\ e_n \end{pmatrix} = \frac{\sum_{i=2}^n e_i \cdot e_{i-1}}{\sum_{i=1}^{n-1} e_i^2}$$

Методы устранения автокорреляции

Процедура Кохрейна-Оркатта

4. Вычисляются ОМНК-оценки $\bar{b}_{\text{ОМНК}}(r^{(1)}) = \bar{b}^{(2)}$

$\bar{b}_{\text{ОМНК}} = (X^T \Sigma_0^{-1} X)^{-1} (X^T \Sigma_0^{-1} Y)$, с матрицей

$$\Sigma_0 = \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n-1} \\ \rho & 1 & \rho & \dots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{n-3} \\ \dots & \dots & \dots & \dots & \dots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots & 1 \end{pmatrix},$$

Σ_0

в которой вместо ρ подставлены $r^{(1)}$

Методы устранения автокорреляции

Процедура Кохрейна-Оркатта

5. Рассчитываются регрессионные остатки второй итерации: $e_i^{(2)} = y_i - \hat{y}_i^{(2)}$, где $\hat{y}_i^{(2)} = b_0^{(2)} + b_1^{(2)} x_{i1} + \dots + b_k^{(2)} x_{ik}$ и т.д.

Процесс продолжается до тех пор, пока не будет достигнута требуемая точность (пока r не стабилизируются), а именно пока разность между предыдущей и последующей оценками параметра ρ не станет меньше любого наперед заданного числа.

Методы устранения автокорреляции

Процедура Кохрейна-Оркатта

1) находим оценку модели регрессии: $\hat{Y} = X\hat{\beta}_{МНК}$;

2) используя построенную оценку модели регрессии, находим оценку вектора регрессионных остатков:

$$e \equiv \varepsilon = Y - \hat{Y} = Y - X\hat{\beta}_{МНК}$$

3) находим оценку неизвестного параметра ρ из оценки модели с автокоррелированными остатками:

$$e_i = \rho e_{i-1} + \delta_i, \quad i = 2, 3, \dots, n$$

$$\hat{\rho}^{(1)} = \left(\begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_{n-1} \end{pmatrix}^T \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_{n-1} \end{pmatrix} \right)^{-1} \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_{n-1} \end{pmatrix}^T \begin{pmatrix} e_2 \\ e_3 \\ \dots \\ e_n \end{pmatrix} = \frac{\sum_{i=2}^n e_i \cdot e_{i-1}}{\sum_{i=1}^{n-1} e_i^2}$$

4) Определяем $\hat{\Sigma}_0^{(1)} = \begin{pmatrix} 1 & \hat{\rho}^{(1)} & \hat{\rho}^{(1)2} & \dots & \hat{\rho}^{(1)n-1} \\ \hat{\rho}^{(1)} & 1 & \hat{\rho}^{(1)} & \dots & \hat{\rho}^{(1)n-2} \\ \dots & \dots & \dots & \dots & \dots \\ \hat{\rho}^{(1)n-1} & \hat{\rho}^{(1)n-2} & \dots & \dots & 1 \end{pmatrix}$


5) Находим оценку $\hat{\beta}_{омик} = (X^T (\hat{\Sigma}_0^{(1)})^{-1} X)^{-1} X^T (\hat{\Sigma}_0^{(1)})^{-1} Y$.

Используя вектор $\hat{\beta}_{омик}$, находим оценку $e^{(1)} = Y - \hat{Y}^{(1)} = Y - X\hat{\beta}_{омик}$

Далее рассчитывается второе приближение и т.д., пока два соседних приближения ρ не совпадут с определенной степенью точности.

Точечный прогноз для условного среднего и индивидуальных значений результативного признака

$$\hat{y}_{n+1} = \bar{X}_{n+1}^T \hat{b}_{\text{ОМНК}} + \hat{\rho} e_n$$

- 
- При работе с пространственной статистической информацией, наличие автокоррелированных регрессионных остатков, как правило, обусловлено неправильной спецификацией модели.
 - Поэтому в некоторых практических задачах методом устранения автокорреляции является изменение спецификации (вида функции) регрессионной модели

Изменение спецификации модели как метод устранения автокорреляции

Построим функцию регрессии в следующей форме

$$\tilde{y} = b_0 + \frac{b_1}{x}$$

Regression Summary for Dependent Variable: Бананы (в фунта R= ,99943601 R?= ,99887233 Adjusted R?= ,99873138 F(1,8)=7086,3 p<,00000 Std.Error of estimate: ,09957						
N=10	Beta	Std.Err. of Beta	B	Std.Err. of B	t(8)	p-level
Intercept			12,0805	0,047124	256,3530	0,000000
Доход (в 10000 \$)	-0,999436	0,011873	-10,0768	0,119705	-84,1801	0,000000

$$\hat{y} = 12,08 - 10,08 \cdot z$$

(0,05) (0.12)

$$\hat{R}^2 = 0,998$$

$$z_i = \frac{1}{x_i}$$