

Временные ряды

Временные ряды

Итак, **временной ряд** — это ряд некоторых числовых величин, измеренных в последовательные моменты времени. Существует принципиальное различие между анализом пространственных выборок (*cross-section*) и временных рядов (*time series*): наблюдения в последних строго упорядочены во времени, в то время как порядок наблюдений в перекрестных выборках можно менять произвольно. Переставлять наблюдения во временных рядах некорректно, так как из-за этого меняется закономерность развития ряда. По той же причине во временных рядах нельзя удалять отдельные наблюдения: если в одномоментной кросс-секционной базе данных какие-то точки являются выбросами или закодированы с ошибкой, то тогда их удаление может улучшить качество подгонки оцениваемой модели, во временных рядах приходится работать со всем временным диапазоном без возможности удалить ухудшающие модель наблюдения (исключениями являются случаи, когда нестандартные наблюдения находятся в начале или конце временного ряда).

Зачем нужно моделировать временные ряды ?

1. Чтобы строить прогнозы

Каким будет уровень инфляции через год?

2. Чтобы исследовать динамические причинно-следственные связи

Если фирма сегодня увеличит расходы на рекламу на миллион рублей, то как это скажется на объеме продаж через месяц? А через полгода?

Примеры (одномерные временные ряды)

Объем торгов на Российской фондовой бирже

Дата	Объем торгов, млн. долл.
2 июня	24,93
3 июня	27,13
4 июня	36,17
5 июня	36,57
6 июня	33,62
9 июня	37,43
10 июня	37,9
11 июня	40,12
13 июня	20,99
16 июня	54,97
17 июня	45,71
18 июня	69,42
19 июня	73,28
20 июня	77,26
23 июня	71,76
24 июня	82,37
25 июня	94,27

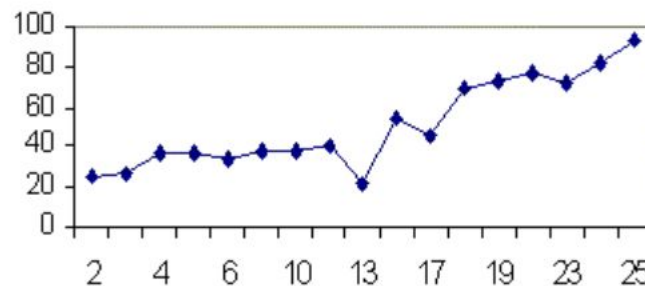
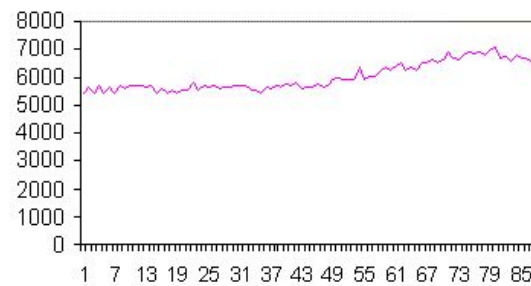


Рис. 5.2. Объем торгов на Российской фондовой бирже

Примеры (одномерные временные ряды)

Индустриальный индекс Доу - Джонса

Номер наблюдения	Индекс	Номер наблюдения	Индекс	Номер наблюдения	Индекс
1	5400	31	5680	61	6390
2	5630	32	5700	62	6500
3	5420	33	5520	63	6260
4	5680	34	5530	64	6310
5	5400	35	5380	65	6280
6	5620	36	5600	66	6500
7	5410	37	5580	67	6510
8	5700	38	5700	68	6590
9	5590	39	5620	69	6490
10	5700	40	5730	70	6610
11	5680	41	5680	71	6880
12	5700	42	5800	72	6680
13	5600	43	5540	73	6590
14	5680	44	5640	74	6760
15	5400	45	5610	75	6870
16	5570	46	5720	76	6820
17	5410	47	5610	77	6900
18	5500	48	5700	78	6790
19	5410	49	5900	79	6980
20	5510	50	5990	80	7070
21	5490	51	5900	81	6660
22	5800	52	5930	82	6720
23	5530	53	5940	83	6550
24	5700	54	6300	84	6810
25	5600	55	5930	85	6670
26	5700	56	6020	86	6680
27	5550	57	6000	87	6500
28	5630	58	6200	88	6490
29	5600	59	6300		
30	5700	60	6270		



Пример (многомерные временные ряды)

В эксперименте проверяются связи между тремя временными рядами: объем грузоперевозок нефти x , цена на нефть h и объем добычи нефти q . Временные ряды заданы в период 21 год по месяцам. Исходные и продифференцированные ряды представлены на рис.17 и рис.18 соответственно.

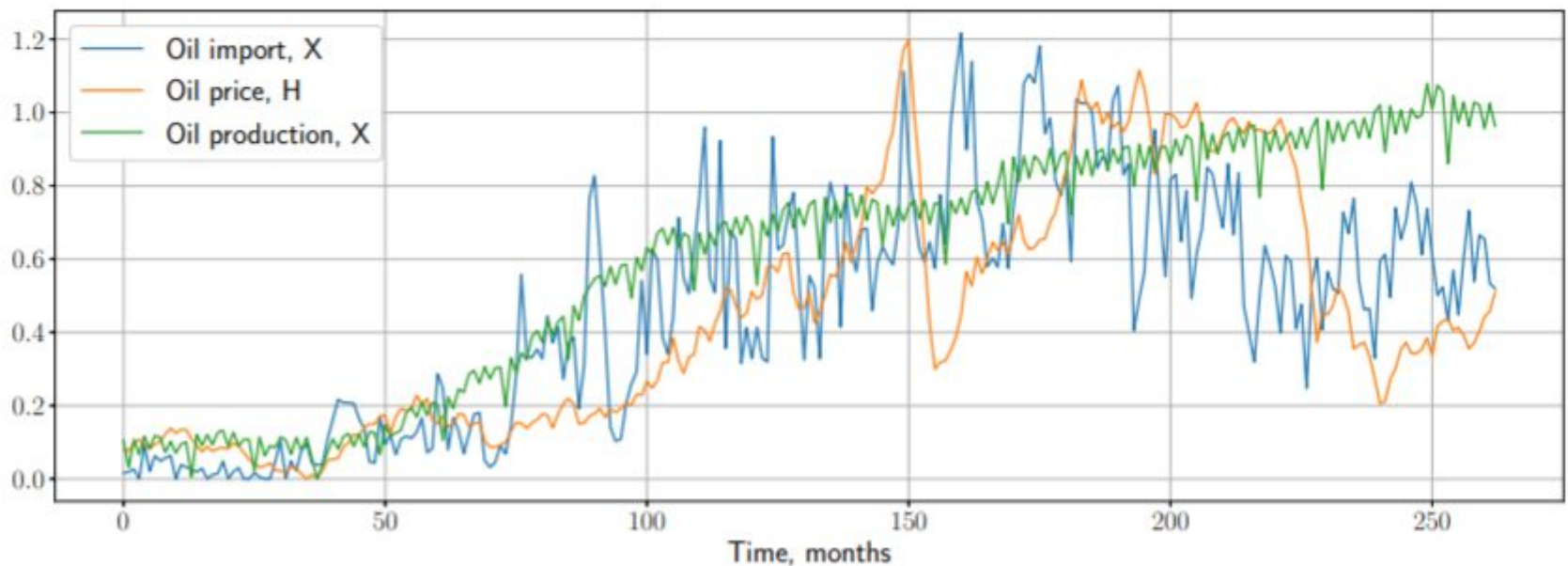


Рис. 17: Исследуемые временные ряды

Алгоритм прогнозирования для временного ряда

1. Строим график временного ряда, графики АСФ, РАСФ. Делаем предварительные выводы о стационарности.
2. Тестируем ряд на стационарность (ADF-тест).
3. Если ряд нестационарный, приводим к стационарному виду (основное преобразование – взятие разностей, разностей второго порядка, сезонных разностей, логарифмирование).
4. Оцениваем модель ARIMA (p, d, q). Можно оценить несколько моделей и выбрать наилучшую по критериям AIC, BIC.
 - Проверяем остатки (WN).
 - Прогнозируем на основе модели п.4.
 - Оцениваем точность прогноза.

Алгоритм прогнозирования для временного ряда

1. Строим график временного ряда, графики ACF, PACF. Делаем предварительные выводы о стационарности.
2. Тестируем ряд на стационарность (ADF-тест).
3. Если ряд нестационарный, приводим к стационарному виду (основное преобразование – взятие разностей, разностей второго порядка, сезонных разностей, логарифмирование).
4. Оцениваем модель ARIMA (p,d,q). Можно оценить несколько моделей и выбрать наилучшую по критериям AIC, BIC.
 - Проверяем остатки (WN).
 - Прогнозируем на основе модели п.4.
 - Оцениваем точность прогноза.

Стационарность временного ряда

Временной ряд называется стационарным в широком смысле (слабо стационарным),

если для всех t :

$$E(Y_t) = \mu < \infty \text{ (const)}$$

$$V(Y_t) = \gamma_0 \text{ (const)}$$

$$\text{Cov}(Y_t, Y_{t-k}) = \gamma_k$$

Стационарность временного ряда

Интуитивно, «стационарность» означает, что поведение ВР в будущем «похоже» на его поведение в прошлом.

Точнее, основные вероятностные характеристики ВР неизменны во времени

Именно «неизменность поведения» во времени позволяет строить прогнозы стационарных временных рядов на основе их предыстории

Основные причины нестационарности

- Наличие тренда;
- Необратимый лаговый полином AR – части (наличие единичного корня);
- Нестационарные остатки.

Проверка стационарности ВР

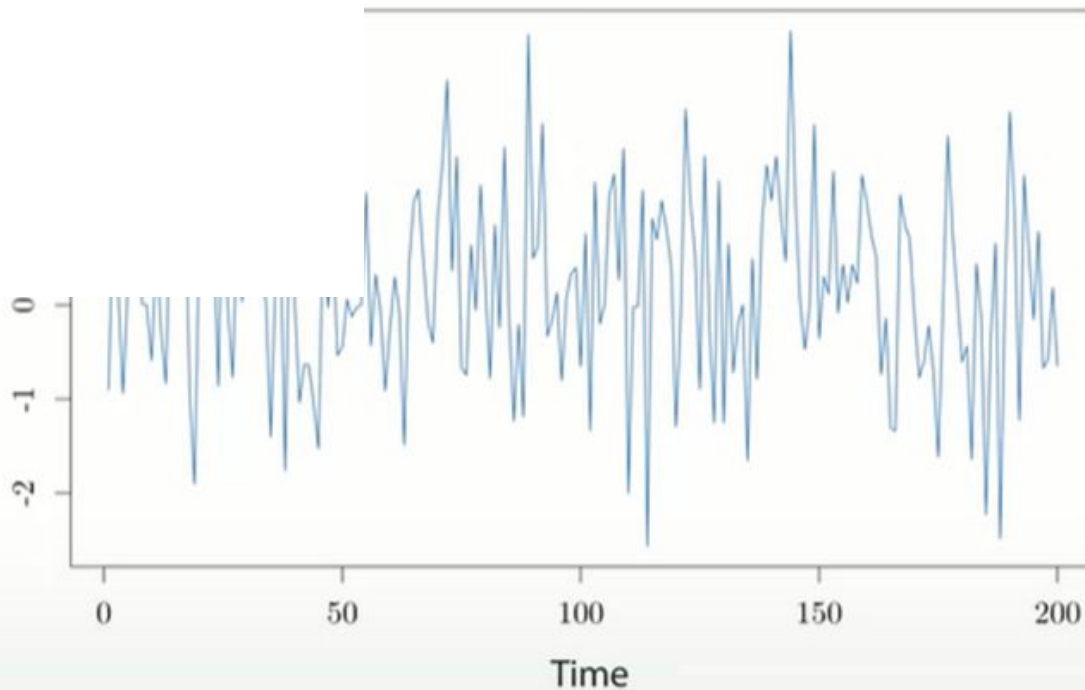
- 1. Анализ графика, АСF, PACF
- 2. Тесты на стационарность (Дики-Фулера, KPSS)

Пример стационарного процесса

является стационарным
(или стационарным),

$$\mu < \infty$$
$$= \gamma_0$$
$$(-k) = \gamma_k$$

Белый шум, $y_t = \varepsilon_t$

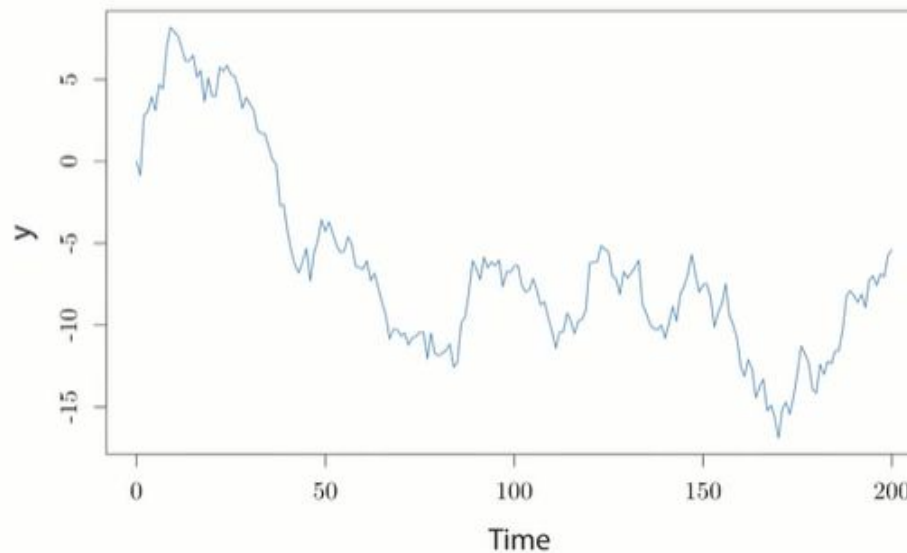


Пример нестационарного процесса

ОНЫМ
НЫМ),

Случайное блуждание (нестационарный процесс!)

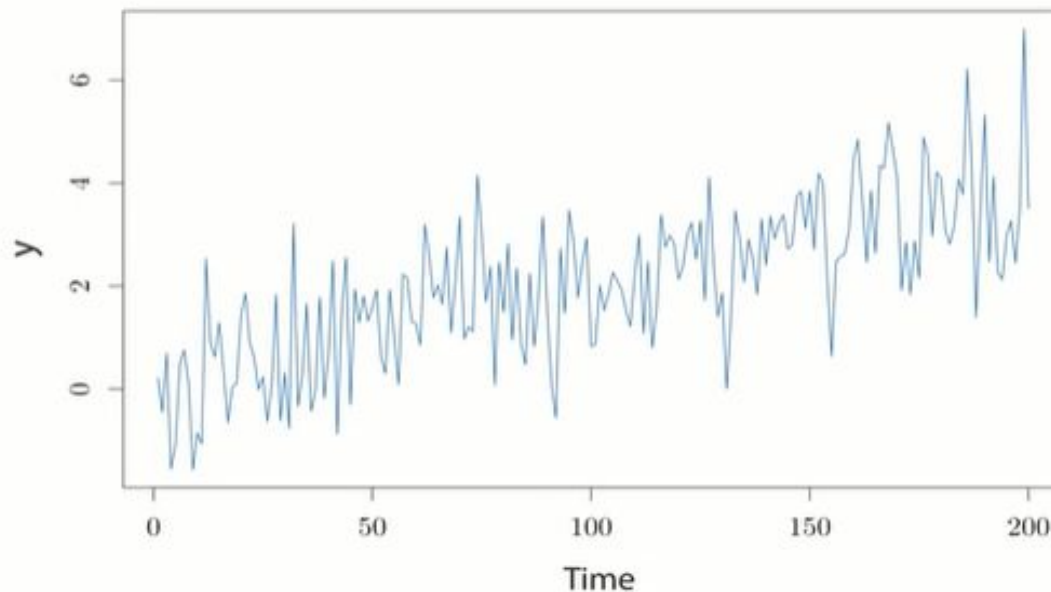
Случайное блуждание, $y_t = y_{t-1} + \varepsilon_t$



Пример нестационарного процесса

Процесс с трендом (нестационарный процесс!)

Процесс с трендом, $y_t = 0.02 \cdot t + \varepsilon_t$



Пример нестационарного процесса

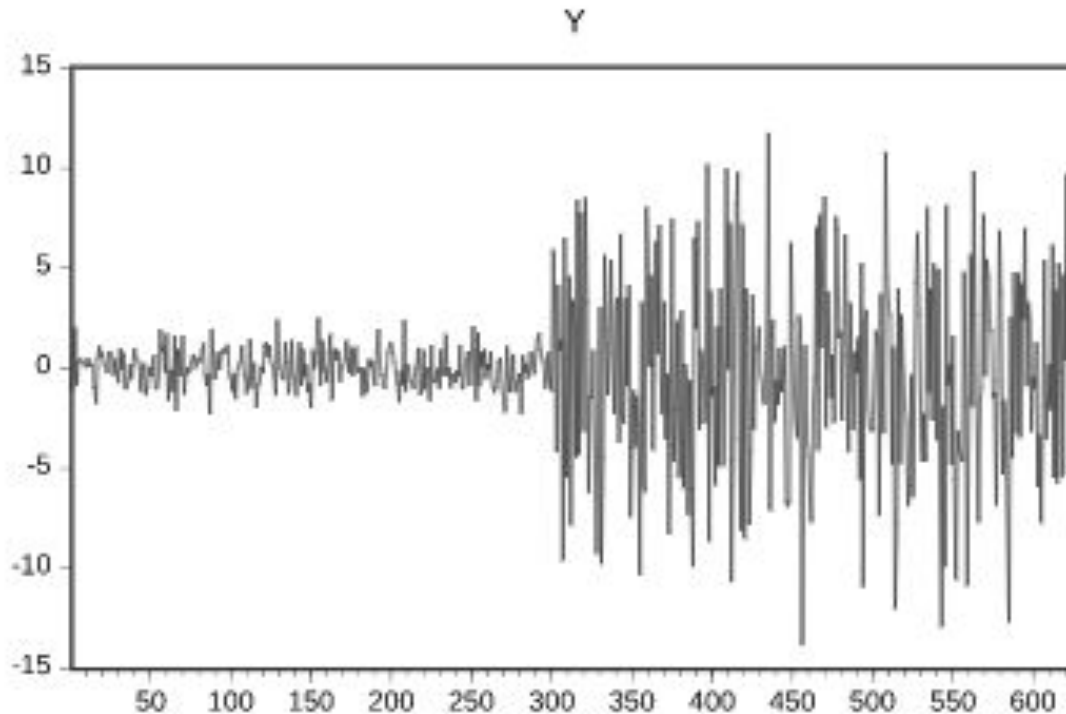


Рис. 12.4. Ряд с непостоянной дисперсией

Пример нестационарного процесса

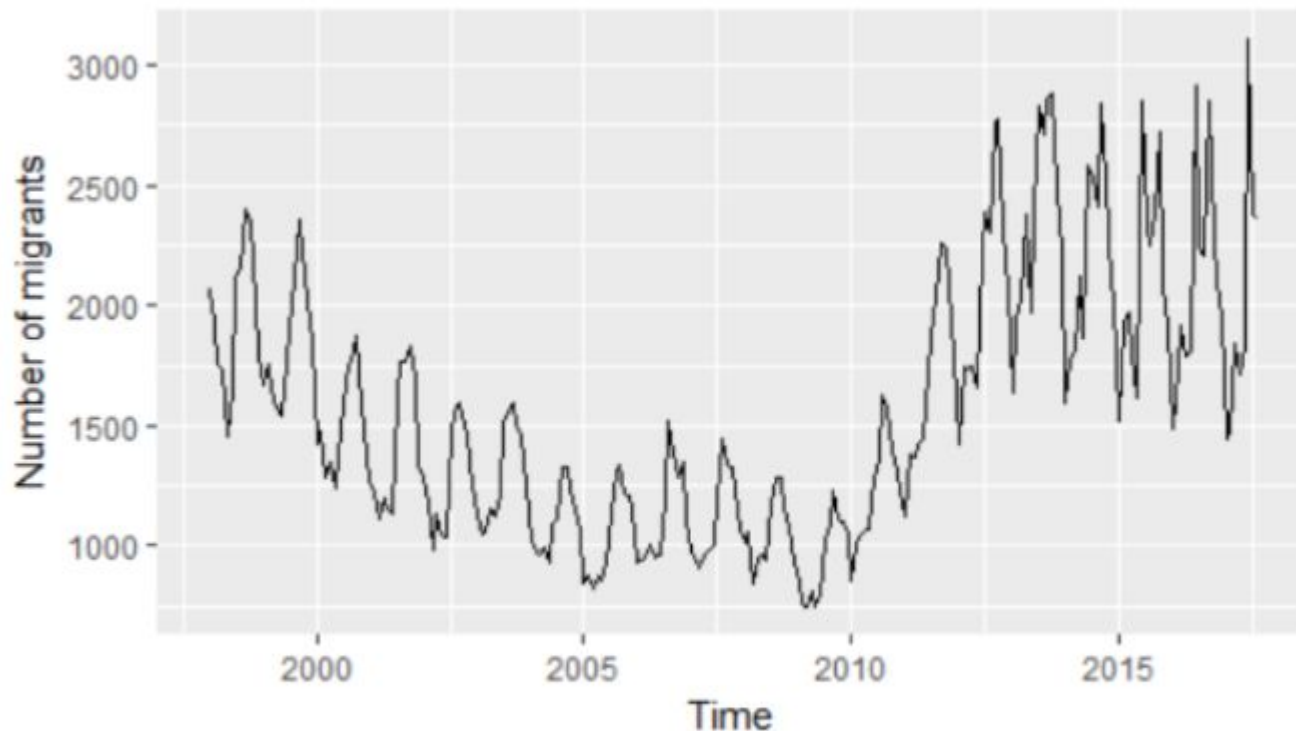


Рисунок 1. Количество мигрантов, ежемесячно уезжавших из Приморского края на постоянное место жительства, 1998-2017гг., чел.

АСФ (автокорреляционная функция)

$$\text{Cov}(Y_t, Y_{t-k}) = \gamma_k$$

Величина γ_τ называется автоковариацией.

Совокупность значений γ_τ в зависимости от τ называется автоковариационной функцией.

При $\tau=0$ получаем дисперсию временного ряда $\text{Var}(Y_t) = \gamma_0$

Автокорреляционная функция (autocorrelation function, ACF):

$$\rho_\tau = \frac{\text{Cov}(Y_t, Y_{t-\tau})}{\text{Var}(Y_t)} = \frac{\gamma_\tau}{\gamma_0}$$

$$\rho_0 = 1, \quad |\rho_\tau| \leq 1$$

АСФ играет важную роль в идентификации временных рядов.

АСФ (пример построения в Excel)

t	Yt	Yt-1	Yt-2	Yt-3	Yt-4	Yt-5	Yt-6	Yt-7	Yt-8
1	6								
2	4,4	6							
3	5	4,4	6						
4	9	5	4,4	6					
5	7,2	9	5	4,4	6				
6	4,8	7,2	9	5	4,4	6			
7	6	4,8	7,2	9	5	4,4	6		
8	10	6	4,8	7,2	9	5	4,4	6	
9	8	10	6	4,8	7,2	9	5	4,4	6
10	5,6	8	10	6	4,8	7,2	9	5	4,4
11	6,4	5,6	8	10	6	4,8	7,2	9	5
12	11	6,4	5,6	8	10	6	4,8	7,2	9
14	9	11	6,4	5,6	8	10	6	4,8	7,2
15	6,6	9	11	6,4	5,6	8	10	6	4,8
16	7	6,6	9	11	6,4	5,6	8	10	6
17	10,8	7	6,6	9	11	6,4	5,6	8	10

Коэф. Корреляции

Cor(Yt,Yt-1)	Cor(Yt,Yt-2)	Cor(Yt,Yt-3)	Cor(Yt,Yt-4)	Cor(Yt,Yt-5)	Cor(Yt,Yt-6)	Cor(Yt,Yt-7)	Cor(Yt,Yt-8)
0,17	-0,57	0,11	0,98	0,12	0,15	0,00	0,97

ACF (пример построения в Excel)

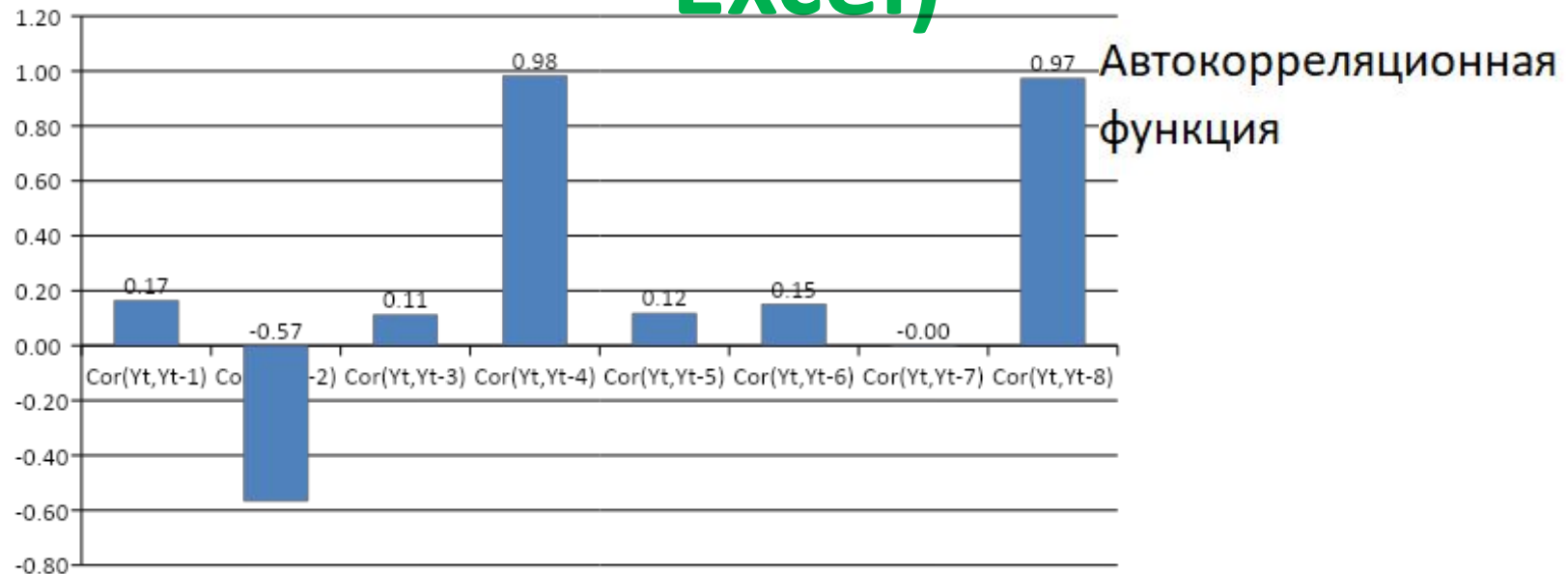
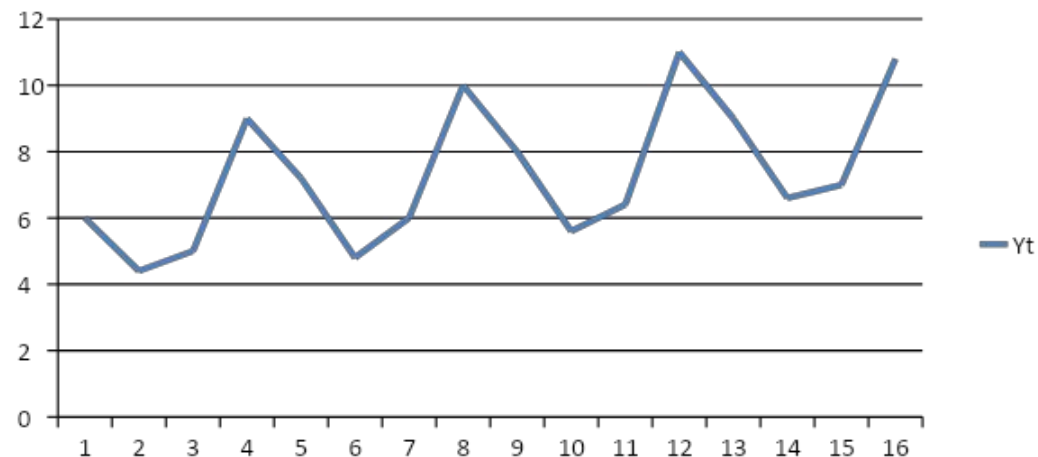


График исходного
временного ряда



РАСФ (частная автокорр. функция)

Получим $\hat{\phi}_k$ из оценки регрессии

$$\hat{y}_t = * + * \cdot y_{t-1} + * \cdot y_{t-2} + \dots + * \cdot y_{t-k+1} + \phi_k y_{t-k} + u_t$$

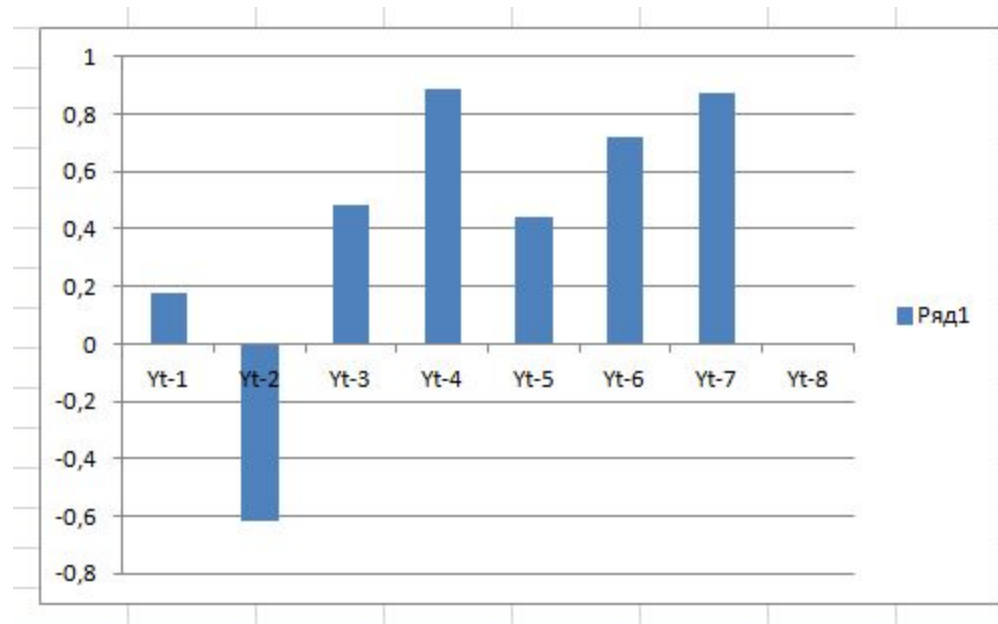
- Для оценки каждого $\hat{\phi}_k$ строится отдельная регрессия
- Из каждой регрессии нужен только последний коэффициент

Значение автокорреляции, полученное из данной регрессии, будет очищенным от влияния промежуточных лагов: этот эффект достигается за счет включения в модель всех промежуточных лагов в качестве регрессоров.

РАСФ (построение в Excel)

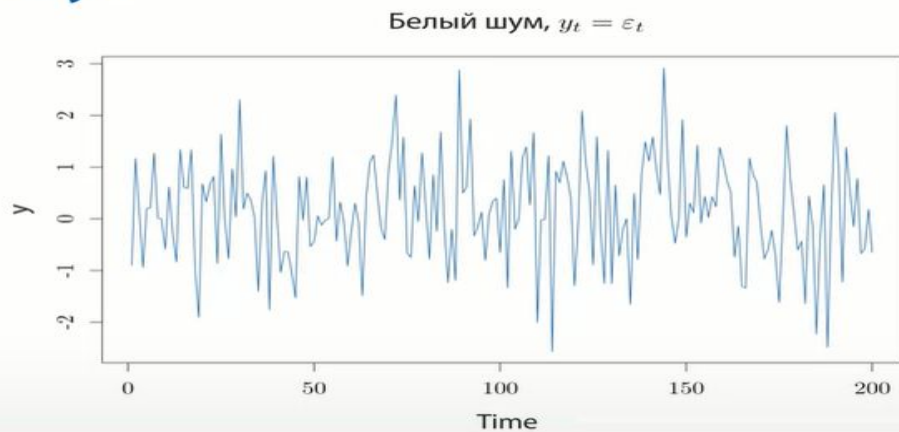
Yt	Yt-1	Yt-2	Yt-3	Yt-4	Yt-5	Yt-6	Yt-7	Yt-8
6								
4,4	6							
5	4,4	6						
9	5	4,4	6					
7,2	9	5	4,4	6				
4,8	7,2	9	5	4,4	6			
6	4,8	7,2	9	5	4,4	6		
10	6	4,8	7,2	9	5	4,4	6	
8	10	6	4,8	7,2	9	5	4,4	6
5,6	8	10	6	4,8	7,2	9	5	4,4
6,4	5,6	8	10	6	4,8	7,2	9	5
11	6,4	5,6	8	10	6	4,8	7,2	9
9	11	6,4	5,6	8	10	6	4,8	7,2
6,6	9	11	6,4	5,6	8	10	6	4,8
7	6,6	9	11	6,4	5,6	8	10	6
10,8	7	6,6	9	11	6,4	5,6	8	10
Коэффициент из регрессионного уравнения (частные коэф-ты корреляции)								
	Yt-1	Yt-2	Yt-3	Yt-4	Yt-5	Yt-6	Yt-7	Yt-8
	0,18	-0,62	0,48	0,89	0,44	0,72	0,87	

РАСФ (построение в Excel)

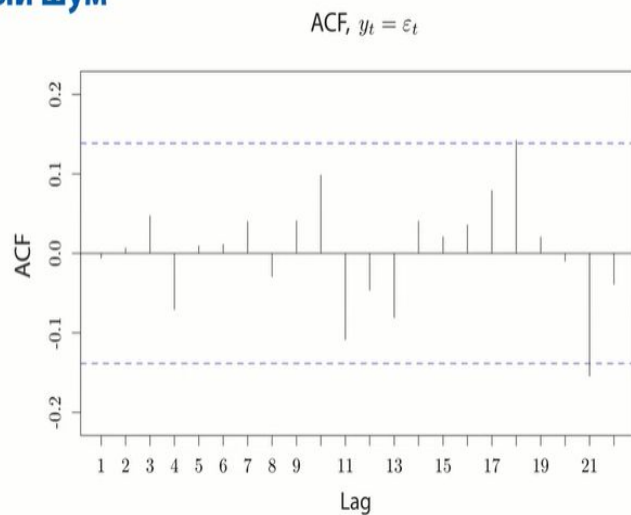


Примеры АСФ, РАСФ

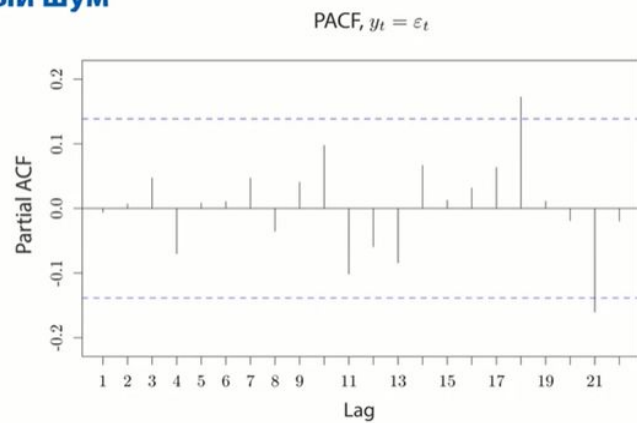
Белый шум



Белый шум

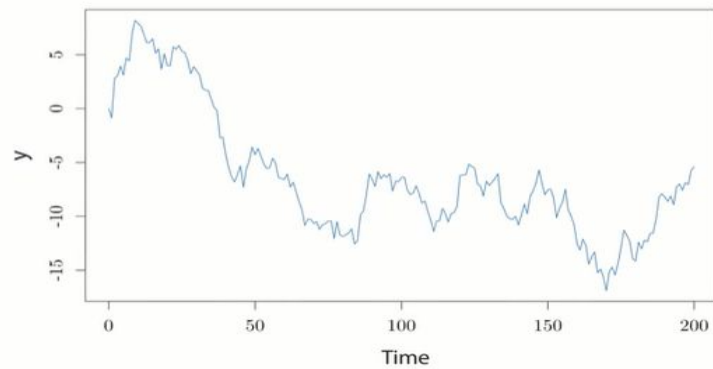


Белый шум



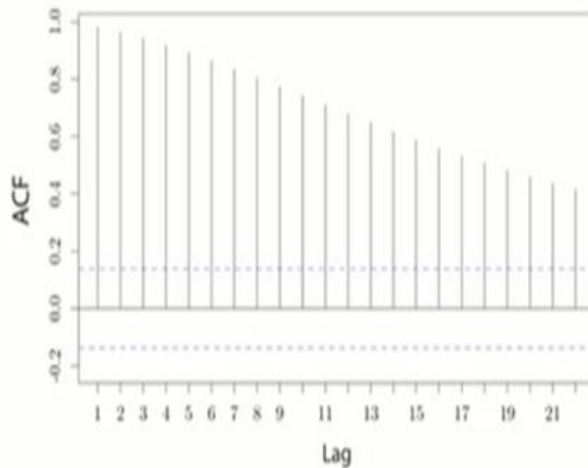
Случайное блуждание (нестационарный процесс!)

Случайное блуждание, $y_t = y_{t-1} + \varepsilon_t$

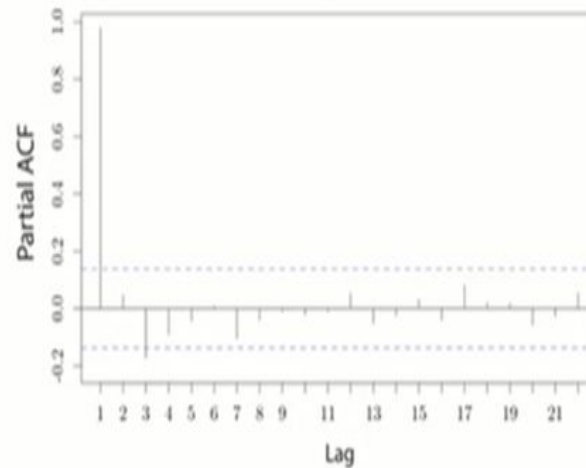


Случайное блуждание (нестационарный процесс!)

ACF, $y_t = y_{t-1} + \varepsilon_t$

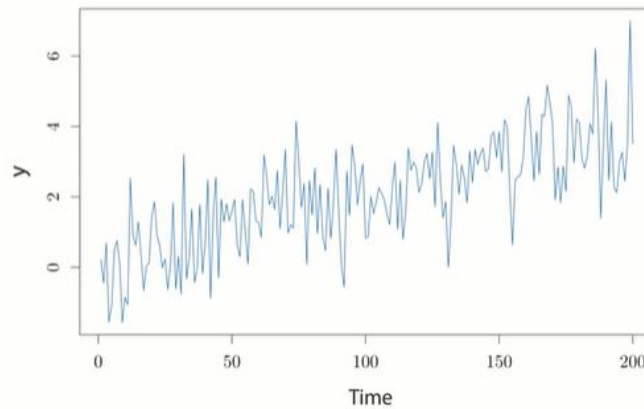


PACF, $y_t = y_{t-1} + \varepsilon_t$



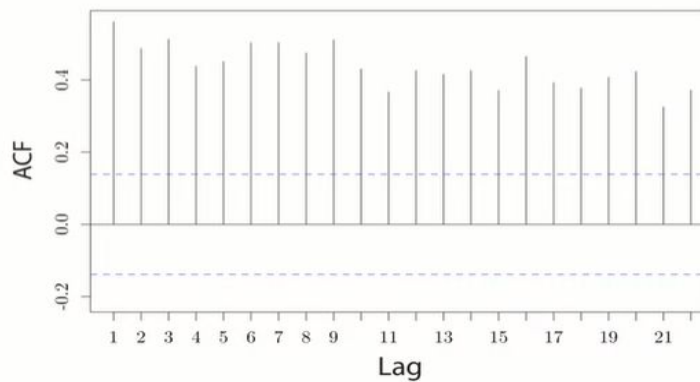
Процесс с трендом (нестационарный процесс!)

Процесс с трендом, $y_t = 0.02 \cdot t + \varepsilon_t$



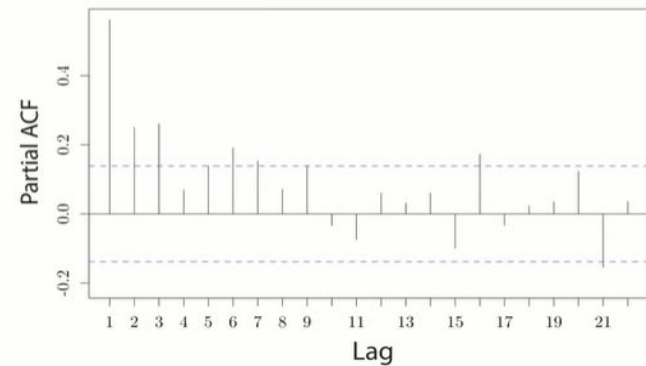
Процесс с трендом (нестационарный процесс!)

ACF, $y_t = 0.02 \cdot t + \varepsilon_t$



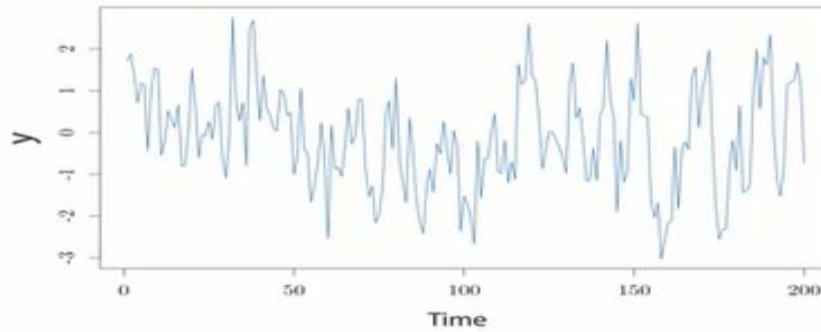
Процесс с трендом (нестационарный процесс!)

PACF, $y_t = 0.02 \cdot t + \varepsilon_t$

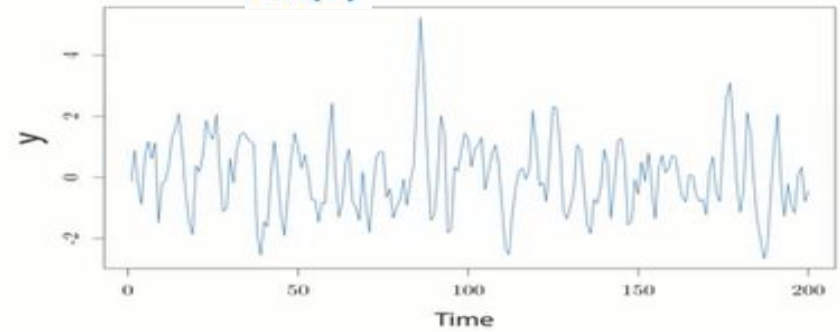


AR(1) и AR(2)

AR(1), $y_t = 0.7y_{t-1} + \varepsilon_t$

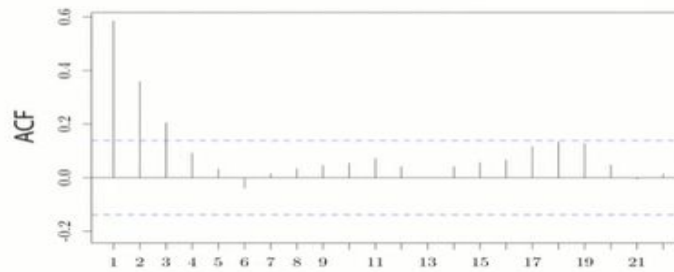


AR(2) $y_t = 0.9y_{t-1} - 0.5y_{t-2} + \varepsilon_t$

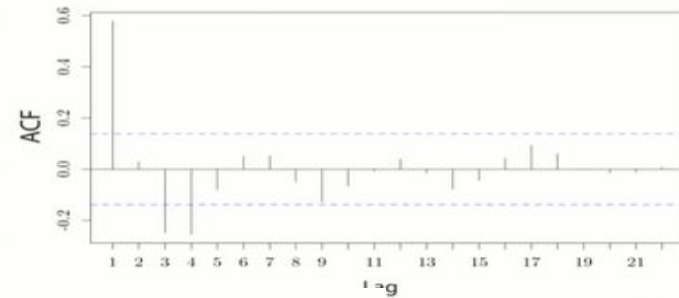


AR(1) и AR(2)

ACF, AR(1), $y_t = 0.7y_{t-1} + \varepsilon_t$

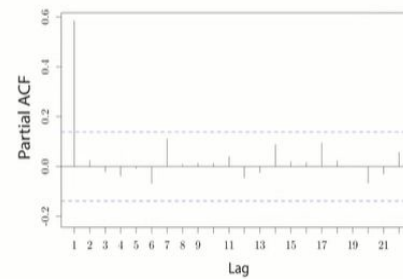


ACF, AR(2), $y_t = 0.9y_{t-1} - 0.5y_{t-2} + \varepsilon_t$

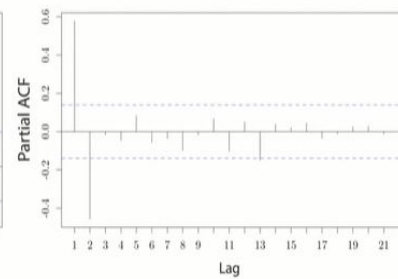


AR(1) и AR(2)

PACF, AR(1), $y_t = 0.7y_{t-1} + \varepsilon_t$

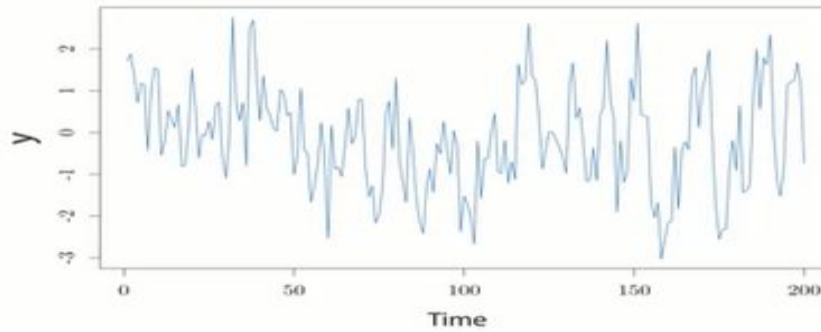


PACF, AR(2), $y_t = 0.9y_{t-1} - 0.5y_{t-2} + \varepsilon_t$

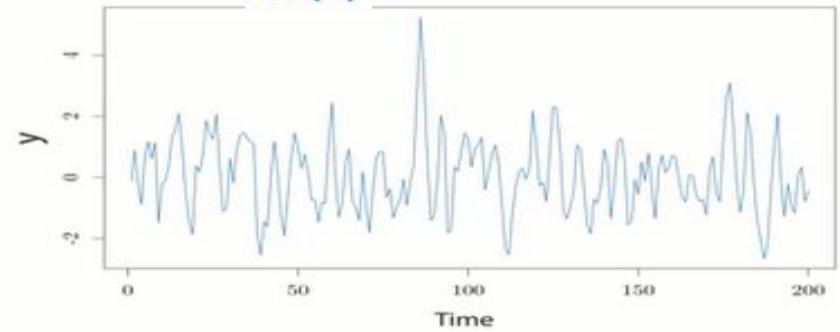


AR(1) и AR(2)

AR(1), $y_t = 0.7y_{t-1} + \varepsilon_t$

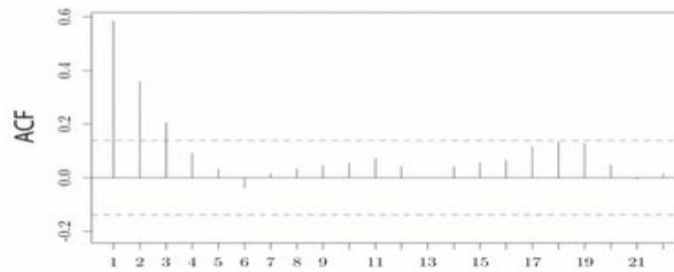


AR(2) $y_t = 0.9y_{t-1} - 0.5y_{t-2} + \varepsilon_t$

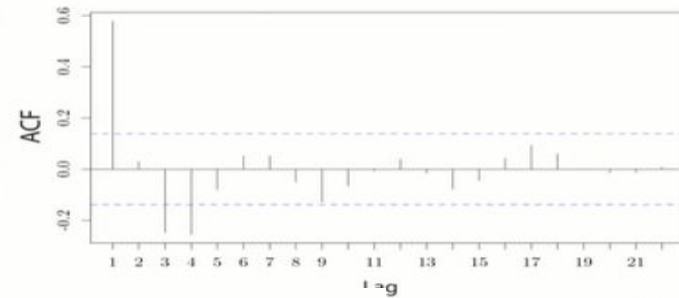


AR(1) и AR(2)

ACF, AR(1), $y_t = 0.7y_{t-1} + \varepsilon_t$

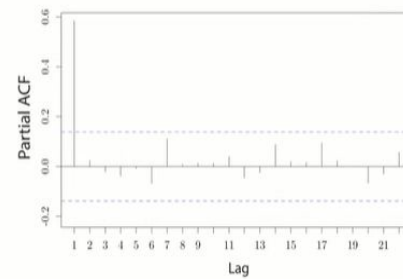


ACF, AR(2), $y_t = 0.9y_{t-1} - 0.5y_{t-2} + \varepsilon_t$

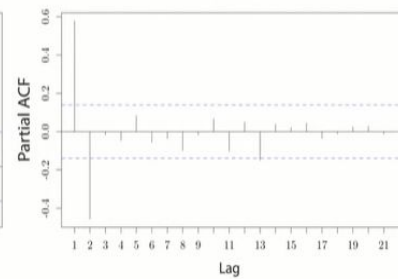


AR(1) и AR(2)

PACF, AR(1), $y_t = 0.7y_{t-1} + \varepsilon_t$



PACF, AR(2), $y_t = 0.9y_{t-1} - 0.5y_{t-2} + \varepsilon_t$



Алгоритм прогнозирования для временного ряда

1. Строим график временного ряда, графики АСФ, РАСФ. Делаем предварительные выводы о стационарности.
2. Тестируем ряд на стационарность (ADF-тест).
3. Если ряд нестационарный, приводим к стационарному виду (основное преобразование – взятие разностей, разностей второго порядка, сезонных разностей, логарифмирование).
4. Оцениваем модель ARIMA (p,d,q). Можно оценить несколько моделей и выбрать наилучшую по критериям AIC, BIC.
 - Проверяем остатки (WN).
 - Прогнозируем на основе модели п.4.
 - Оцениваем точность прогноза.

Расширенный тест Дики-Фуллера

Расширенный тест Дики — Фуллера является одним из основных методов проверки нестационарности ряда. В основе этого теста лежит предположение, что любой ряд можно аппроксимировать с заданной точностью процессом типа $AR(p)$ необходимого порядка:

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \dots + \alpha_p y_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim WN(0; \sigma^2). \quad (12.1)$$

Сначала введем понятие первой и второй разностей значений ряда y_t :

$$\Delta y_t = y_t - y_{t-1};$$

$$\Delta^2 y_t = \Delta y_t - \Delta y_{t-1}.$$

Оригинальный ряд y_t называется рядом в уровнях.

Расширенный тест Дики-Фуллера (ADF-тест)

Вычтем y_{t-1} из обеих частей уравнения (12.1)

$$\Delta y_t = (\alpha_1 - 1)y_{t-1} + \alpha_2 y_{t-2} + \dots + \alpha_p y_{t-p} + \varepsilon_t.$$

Затем вычтем и прибавим к правой части уравнения $\alpha_p y_{t-(p-1)}$:

$$\Delta y_t = (\alpha_1 - 1)y_{t-1} + \alpha_2 y_{t-2} + \dots + (\alpha_{p-1} + \alpha_p)y_{t-(p-1)} - \alpha_p \Delta y_{t-(p-1)} + \varepsilon_t.$$

На следующем шаге добавим и вычтем $(\alpha_{p-1} + \alpha_p)y_{t-(p-2)}$:

$$\begin{aligned} \Delta y_t = & (\alpha_1 - 1)y_{t-1} + \alpha_2 y_{t-2} + \dots + (\alpha_{p-2} + \alpha_{p-1} + \alpha_p)y_{t-(p-2)} - \\ & - (\alpha_{p-1} + \alpha_p)y_{t-(p-2)} - \alpha_p \Delta y_{t-(p-1)} + \varepsilon_t. \end{aligned}$$

Эту процедуру необходимо повторить для каждого лага, и в итоге получится следующее выражение:

$$\Delta y_t = \gamma y_{t-1} + \sum_{i=1}^{p-1} c_i \Delta y_{t-i} + \varepsilon_t$$

где $\gamma = \alpha_1 - 1$, $A_i = - \sum_{j=i+1}^p \alpha_j$.

Расширенный тест Дики-Фуллера

Существует три спецификации расширенного теста Дики — Фуллера.

1. Без константы:

$$\Delta y_t = \gamma y_{t-1} + \sum_{i=1}^{p-1} c_i \Delta y_{t-i} + \varepsilon_t.$$

2. С константой:

$$\Delta y_t = \alpha + \gamma y_{t-1} + \sum_{i=1}^{p-1} c_i \Delta y_{t-i} + \varepsilon_t.$$

3. С трендом и константой:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p-1} c_i \Delta y_{t-i} + \varepsilon_t.$$

Для каждой спецификации нулевая и альтернативная гипотезы выглядят следующим образом:

$H_0: \gamma = 0$, есть UR , процесс нестационарен;

$H_1: \gamma < 0$, нет UR , процесс стационарен.

Расширенный тест Дики-Фуллера

Если в ряде нет ярко выраженной инерции или закономерности, среднее значение ряда примерно равно нулю, то желательно использовать первую спецификацию. Если в ряде присутствуют значимая инерционность и его среднее значение явно не равно нулю, то более корректной является вторая спецификация. Третья спецификация применима, когда в данных присутствует значимый квадратичный тренд. Более того, нет однозначного критерия выбора числа лагов, но существует большое количество эвристических критериев. Например, зачастую используют информационные критерии Акаике (AIC) и Шварца (BIC) для определения наилучшего в ка-

Расширенный тест Дики-Фуллера

ком-то смысле числа лагов. Критерии, соответственно, выглядят следующим образом:

$$AIC = -2\ln\mathcal{L}^* + 2k;$$
$$BIC = -2\ln\mathcal{L}^* + k\ln T,$$

где $\ln\mathcal{L}^*$ — значение логарифма функции правдоподобия оцениваемой модели в точке максимума; k — количество оцениваемых параметров; T — количество наблюдений.

Как можно заметить, формулы для данных критериев состоят из двух частей. Первое слагаемое обозначает штраф за введение дополнительных параметров. Критерий Акаике вводит меньший штраф за добавление параметров, чем критерий Шварца¹. Первое слагаемое, домноженное на -1 , $2\ln\mathcal{L}^*$, обозначает качество подгонки модели (чем выше значение, тем качественнее модель).

Исходя из всего вышесказанного можно сделать вывод о том, что чем меньше значение информационного критерия, тем лучше модель. Однозначно корректного метода выбора того или иного информационного критерия для оценивания качества моделей нет. Обычно при исследовании временных рядов лучше пользоваться критерием *AIC*. Однако применение этих критериев обосновано с теоретической точки зрения лишь для независимых наблюдений, что не выполнено естественным образом для временных рядов, таким образом, применение этих критериев носит эвристический характер. Более того, опираться на результаты этих критериев можно лишь при достаточно большом количестве наблюдений.

Расширенный тест Дики-Фуллера (ADF-тест) в R-studio

Перед началом работы с временными рядами нужно перевести данные в формат `ts`

`#unem` – исходные данные по безработице (1960-1995гг.)

```
unemts = ts(unem)
```

`#ADF test` (расширенный тест Дики-Фуллера)

Устанавливаем пакет `fUnitRoots`

```
library(fUnitRoots)
```

`adfTest(unemts, lags=4)` #спецификация без константы (если в ряде нет ярко выраженной инерции или закономерности, среднее значение ряда примерно равно 0).

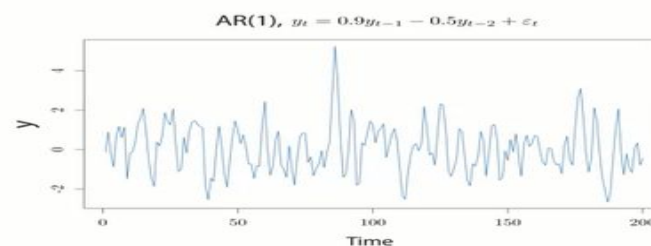
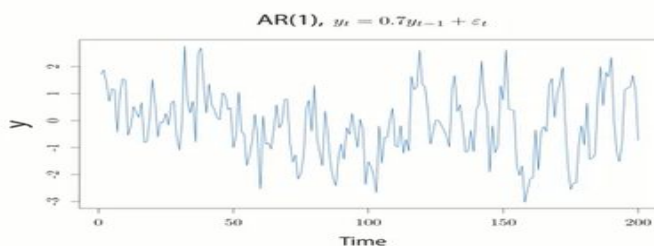
`adfTest(unemts, lags=4, type="c")` #спецификация с константой (если в ряде присутствует значимая инерционность и его среднее значение явно не равно 0).

`adfTest(unemts, lags=4, type="ct")` #спецификация с трендом и константой (если в ряде присутствует значимый линейный или квадратичный тренд).

Расширенный тест Дики-Фуллера (ADF-тест) в R-studio

```
adfTest(unemts,lags=4) #спецификация без константы (если в ряде нет ярко выраженной инерции  
или закономерности, среднее значение ряда примерно равно 0).
```

AR(1) и AR(2)

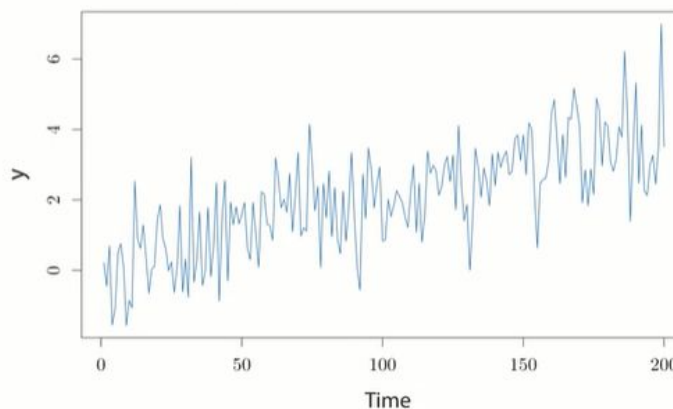


Расширенный тест Дики-Фуллера (ADF-тест) в R-studio

```
adfTest(unemts, lags=4, type="c") # спецификация с константой (если в ряде присутствует значимая инерционность и его среднее значение явно не равно 0).
```

Процесс с трендом (нестационарный процесс!)

Процесс с трендом, $y_t = 0.02 \cdot t + \varepsilon_t$



Расширенный тест Дики-Фуллера (ADF-тест) в R-studio

```
adfTest(unemts,lags=4,type="ct") #спецификация стрендом и константой (если в ряде присутствует значимый линейный или квадратичный тренд).
```

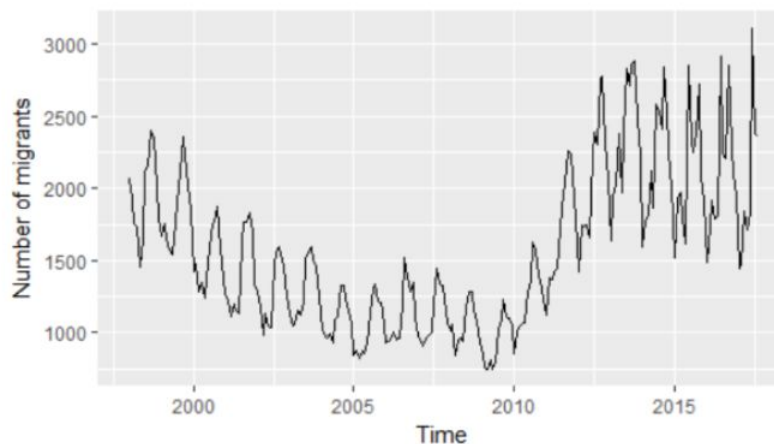


Рисунок 1. Количество мигрантов, ежемесячно уезжавших из Приморского края на постоянное место жительства, 1998-2017гг., чел.

Расширенный тест Дики-Фуллера (ADF-тест) в R-studio

```
Title:  
Augmented Dickey-Fuller Test
```

```
Test Results:  
PARAMETER:  
  Lag Order: 3  
STATISTIC:  
  Dickey-Fuller: -7.9365  
P VALUE:  
  0.01
```

- Вывод: H_0 отвергается, ряд стационарный

Алгоритм прогнозирования для временного ряда

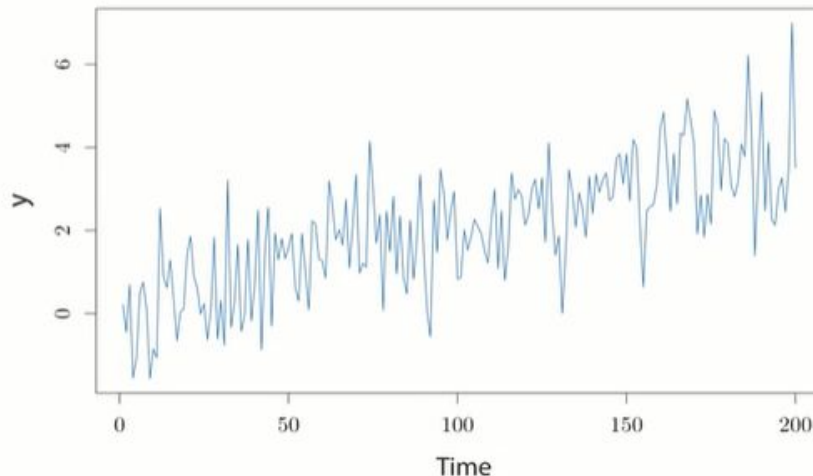
1. Строим график временного ряда, графики АСФ, РАСФ. Делаем предварительные выводы о стационарности.
2. Тестируем ряд на стационарность (ADF-тест).
3. Если ряд нестационарный, приводим к стационарному виду (основное преобразование – взятие разностей, разностей второго порядка, логарифмирование).
4. Оцениваем модель ARIMA (p,d,q). Можно оценить несколько моделей и выбрать наилучшую по критериям AIC, BIC.
 - Проверяем остатки (WN).
 - Прогнозируем на основе модели п.4.
 - Оцениваем точность прогноза.

Как привести ВР к стационарному виду?

- Очень часто стационарность возникает из-за присутствия в ВР трендовой составляющей

Процесс с трендом
(нестационарный процесс!)

Процесс с трендом, $y_t = 0.02 \cdot t + \varepsilon_t$



Как избавиться от трендовой составляющей?

Заменить исходные данные первыми (вторыми, сезонными) разностями

Сначала введем понятие первой и второй разностей значений ряда y_t :

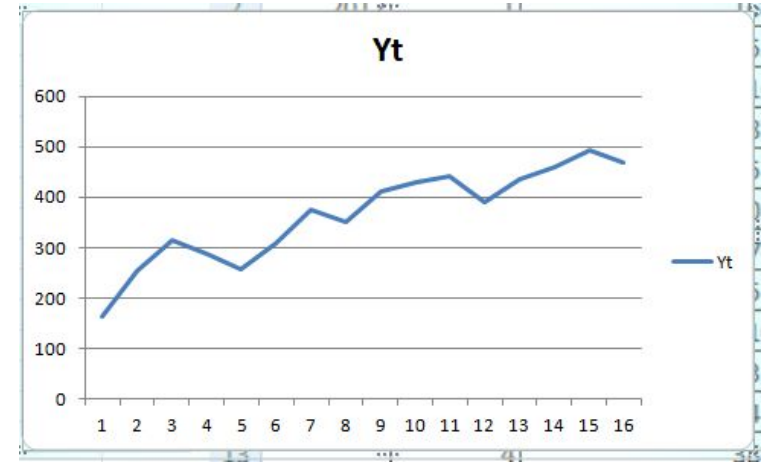
$$\Delta y_t = y_t - y_{t-1};$$

$$\Delta^2 y_t = \Delta y_t - \Delta y_{t-1}.$$

Оригинальный ряд y_t называется рядом в уровнях.

Построение ряда в первых разностях (Excel)

Y_t	$\Delta Y_t = Y_t - Y_{t-1}$
165	
253	88
316	63
287	-29
257	-30
308	51
376	68
351	-25
410	59
431	21
443	12
389	-54
436	47
459	23
492	33
470	-22



Построение ряда в первых разностях (R-studio)

- Пакет tseries

```
diff(df_ts, 1)
```

Иногда, например, в случае квадратичного тренда, для приведения ряда к стационарному необходимо взять вторые разности.

Логарифмирование уровней ВР

- Делаем, если дисперсия уровней ВР непостоянна (сначала логарифмируем, потом, если нужно, берем разности)

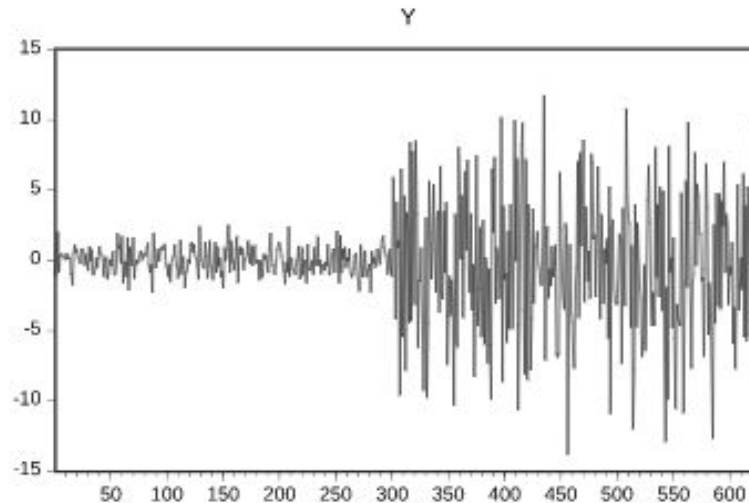


Рис. 12.4. Ряд с непостоянной дисперсией

Логарифмирование уровней ВР (Excel)

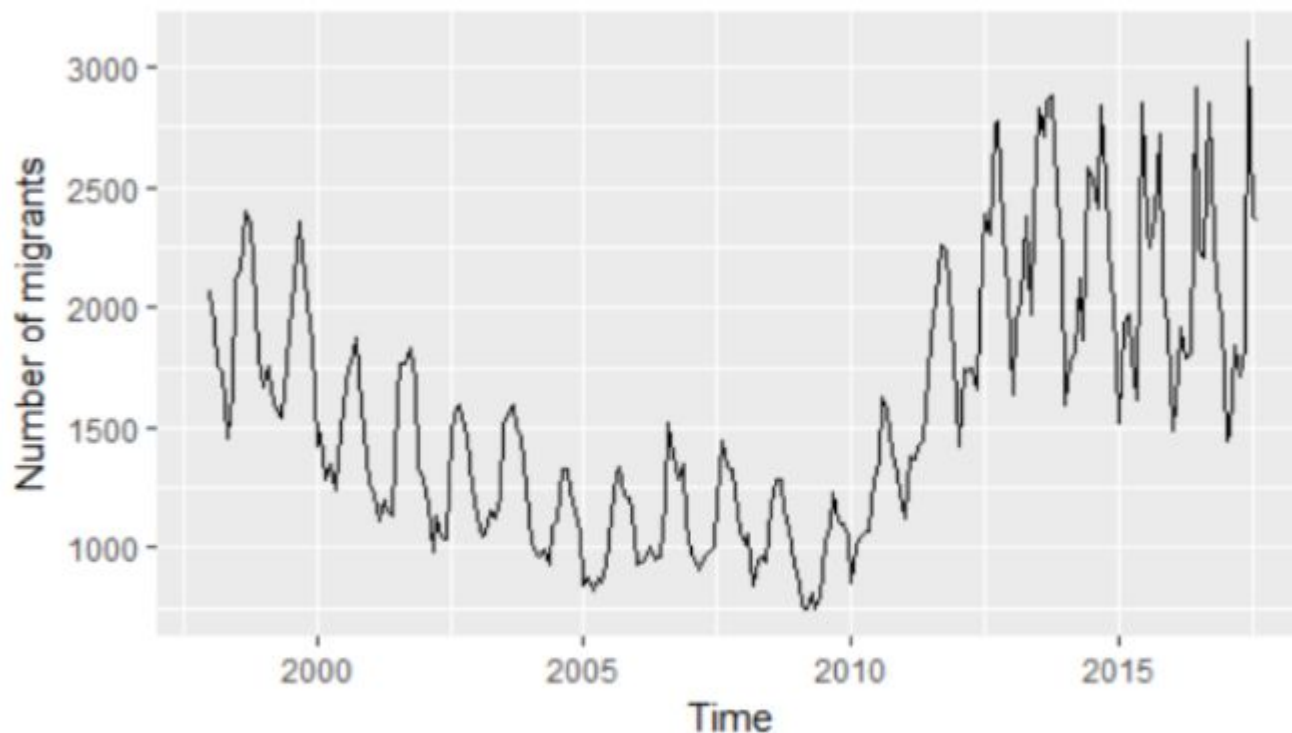
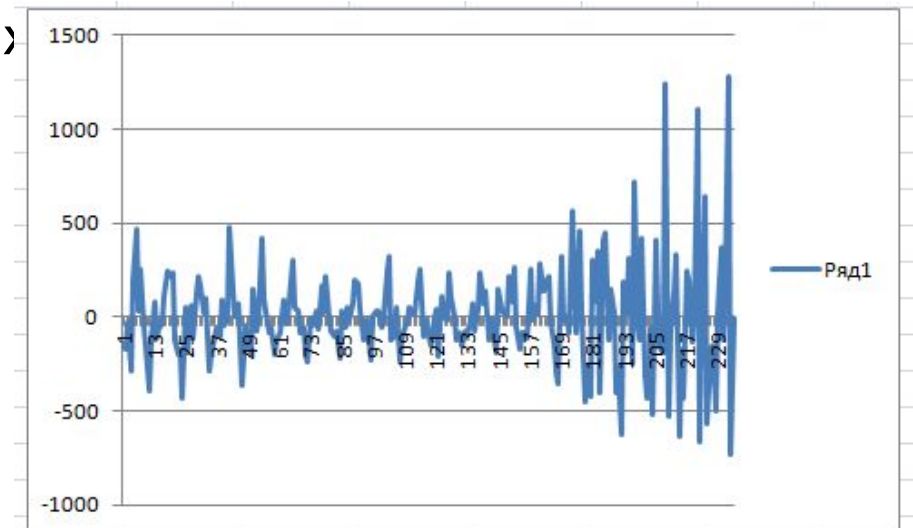


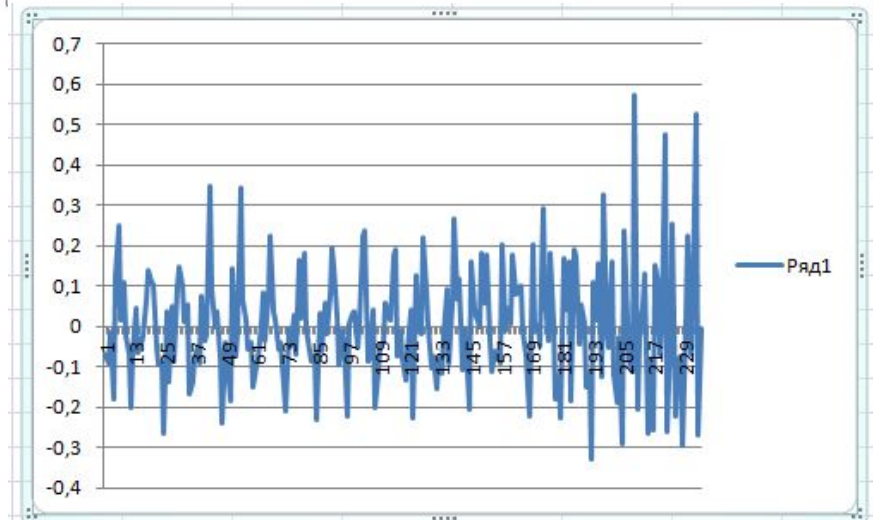
Рисунок 1. Количество мигрантов, ежемесячно уезжавших из Приморского края на постоянное место жительства, 1998-2017гг., чел.

Логарифмирование уровней ВР (Excel)

- ВР после взятия обычных
- разностей



- ВР после
- логарифмирования
- и взятия разностей



Алгоритм прогнозирования для временного ряда

1. Строим график временного ряда, графики АСФ, РАСФ. Делаем предварительные выводы о стационарности.
2. Тестируем ряд на стационарность (ADF-тест).
3. Если ряд нестационарный, приводим к стационарному виду (основное преобразование – взятие разностей, разностей второго порядка, логарифмирование).
4. Оцениваем модель ARIMA (p,d,q). Можно оценить несколько моделей и выбрать наилучшую по критериям AIC, BIC.
 - Проверяем остатки (WN).
 - Прогнозируем на основе модели п.4.
 - Оцениваем точность прогноза.

Важные примеры одномерных случайных процессов

0. Белый шум
1. $AR(1)$ — процесс авторегрессии первого порядка
2. $AR(p)$ — процесс авторегрессии порядка p
3. $MA(1)$ — процесс скользящего среднего первого порядка
4. $MA(q)$ — процесс скользящего среднего порядка q
5. $ARMA(p,q)$ — процесс авторегрессии со скользящим средним в остатках
6. RW — случайное блуждание
7. $ARIMA(p,d,q)$ — интегрированный порядка k процесс авторегрессии со скользящим средним в остатках

Важные примеры одномерных случайных процессов

Почему мы рассматриваем именно эти модели?

Потому что они неплохо описывают динамику многих экономических переменных: инфляции, валютных курсов, доходностей активов...

С помощью ARMA (p,q) можно компактно и очень точно описать любой стационарный процесс.

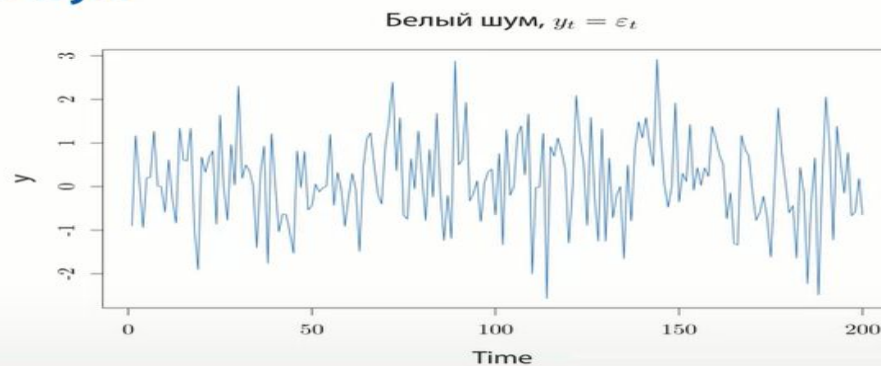
Белый шум (WN-white noise)

Процессом белого шума мы будем называть последовательность независимых и одинаково распределенных случайных величин с нулевым математическим ожиданием и дисперсией σ^2 :

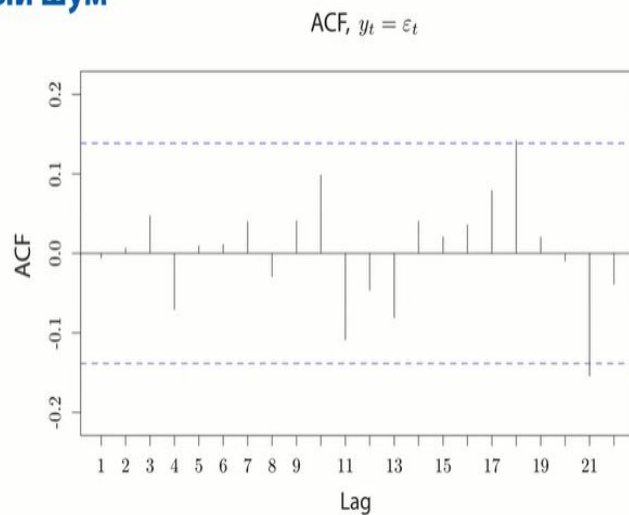
$$\varepsilon_1, \varepsilon_2, \dots, \varepsilon_{t-1}, \varepsilon_t, \dots$$

Белый шум

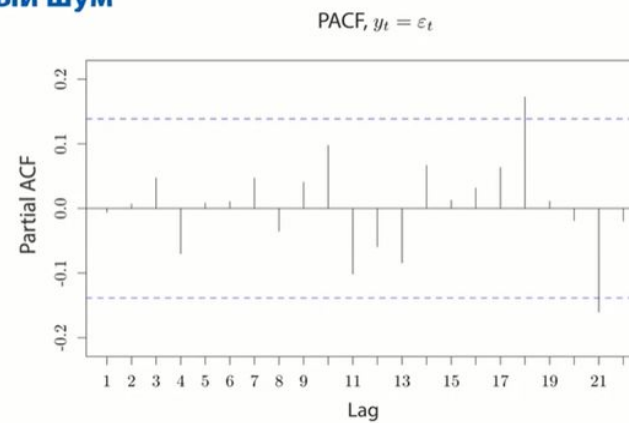
Белый шум



Белый шум



Белый шум



AR – процессы

- AR(p) – процесс авторегрессии порядка p

$$y_t = \delta + \theta_1 y_{t-1} + \dots + \theta_p y_{t-p} + \varepsilon_t$$

$$(\xi_t \sim \text{WN}(0; \sigma^2))$$

AR (1) – процесс авторегрессии первого порядка

$$y_t = \delta + \theta y_{t-1} + \varepsilon_t, |\theta| < 1 \quad (\xi_t \sim \text{WN}(0; \sigma^2))$$

Докажем, что:

$$\mu = \frac{\delta}{1-\theta},$$

$$\gamma_0 = \frac{\sigma^2}{1-\theta^2},$$

$$\gamma_k = \theta^k \frac{\sigma^2}{1-\theta^2}$$

-
- $|\theta| < 1$ - условие стационарности процесса AR(1)

AR(1) – вычисление математического ожидания

$$y_t = \delta + \theta y_{t-1} + \varepsilon_t, |\theta| < 1$$

$$E y_t = E(\delta + \theta y_{t-1} + \varepsilon_t)$$

$$E y_t = \delta + \theta E(y_{t-1}) + E(\varepsilon_t)$$

$$\mu = \delta + \theta \mu + 0$$

$$\mu = \frac{\delta}{1 - \theta}$$

AR(1) – вычисление дисперсии

$$y_t = \delta + \theta y_{t-1} + \varepsilon_t, |\theta| < 1$$

$$V(y_t) = V(\delta + \theta y_{t-1} + \varepsilon_t)$$

$$V(y_t) = \theta^2 V(y_{t-1}) + V(\varepsilon_t)$$

$$\gamma_0 = \theta^2 \gamma_0 + \sigma^2$$

$$\gamma_0 = \frac{\sigma^2}{1 - \theta^2}$$

AR(1) – вычисление автоковариаций

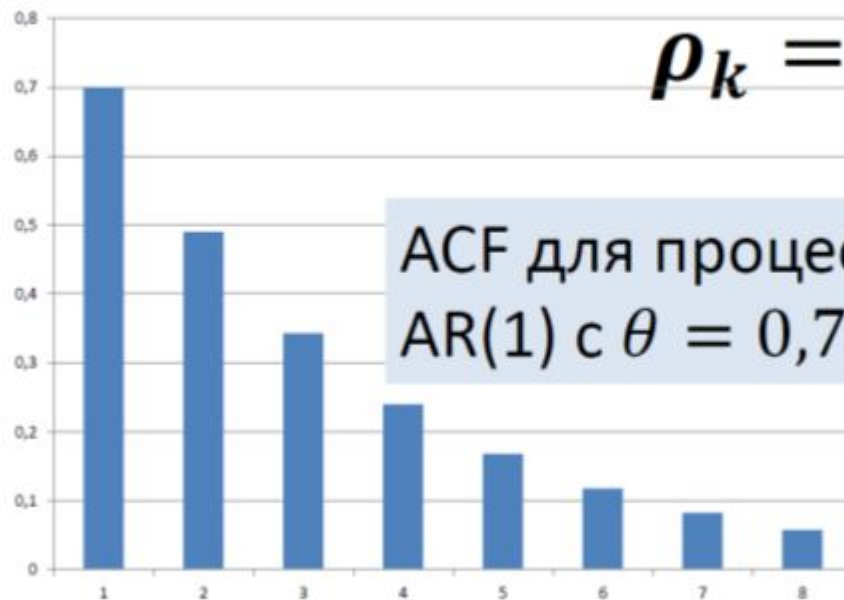
$$\begin{aligned}\gamma_1 &= Cov(y_t, y_{t-1}) = Cov(\delta + \theta y_{t-1} + \varepsilon_t, y_{t-1}) = \\ &= Cov(\theta y_{t-1} + \varepsilon_t, y_{t-1}) = \\ &= \theta * Cov(y_{t-1}, y_{t-1}) + Cov(\varepsilon_t, y_{t-1}) = \\ &= \theta * \gamma_0 + 0 = \theta \frac{\sigma^2}{1 - \theta^2}\end{aligned}$$

Аналогично: $\gamma_k = Cov(y_t, y_{t-k}) = \theta^k \frac{\sigma^2}{1 - \theta^2}$

AR(1)- процесс авторегрессии первого порядка

$$y_t = \delta + \theta y_{t-1} + \varepsilon_t, |\theta| < 1$$

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \theta^k$$



AR(1) — процесс авторегрессии 1-го порядка



Пример временного ряда порожденного AR(1) процессом с $\theta = 0,7$

Пример (AR(1))

Для временного ряда AR(1) $y_t = 2 + 0,5y_{t-1} + \xi_t$, ($\xi_t \sim WN(0; \sigma^2)$), найти математическое ожидание μ , автокорреляционную функцию ρ_k , частную автокорреляционную функцию ϕ_k .

$$\mu = \frac{\delta}{1-\theta},$$

$$\gamma_0 = \frac{\sigma^2}{1-\theta^2},$$

$$\gamma_k = \theta^k \frac{\sigma^2}{1-\theta^2}$$

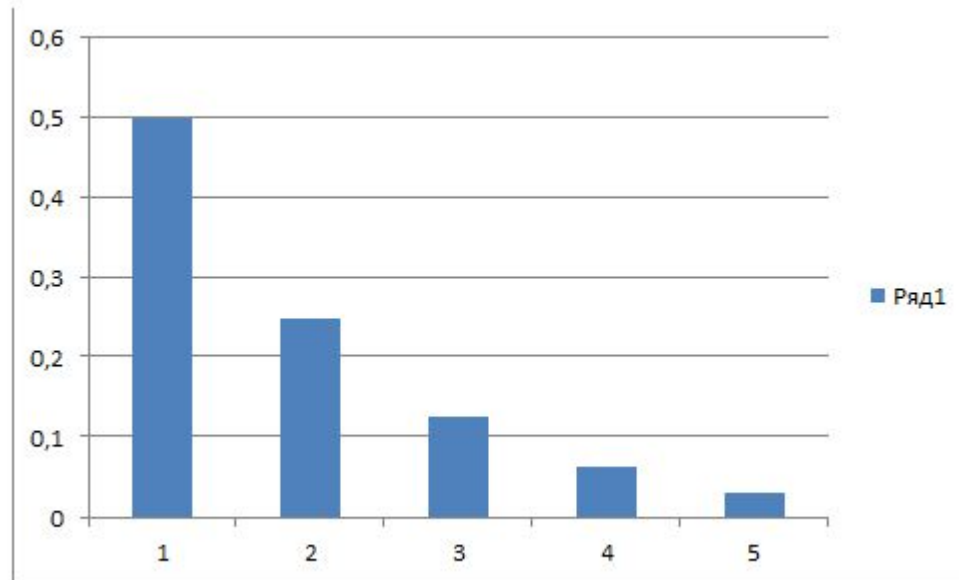
$$\rho_k = \frac{\gamma_k}{\gamma_0} = \theta^k$$

- $\mu = 2/0,5 = 4$ $\rho_k = 0,5^k$
- $\gamma_0 = \sigma^2/0,75$ $(\rho_1 = 0,5; \rho_2 = 0,25; \rho_3 = 0,125)$
- $\gamma_k = 0,5^k * (\sigma^2/0,75)$ $\psi_1 = 0,5, \psi_2 = \psi_3 = \dots 0$

Пример (AR(1)) - АСФ

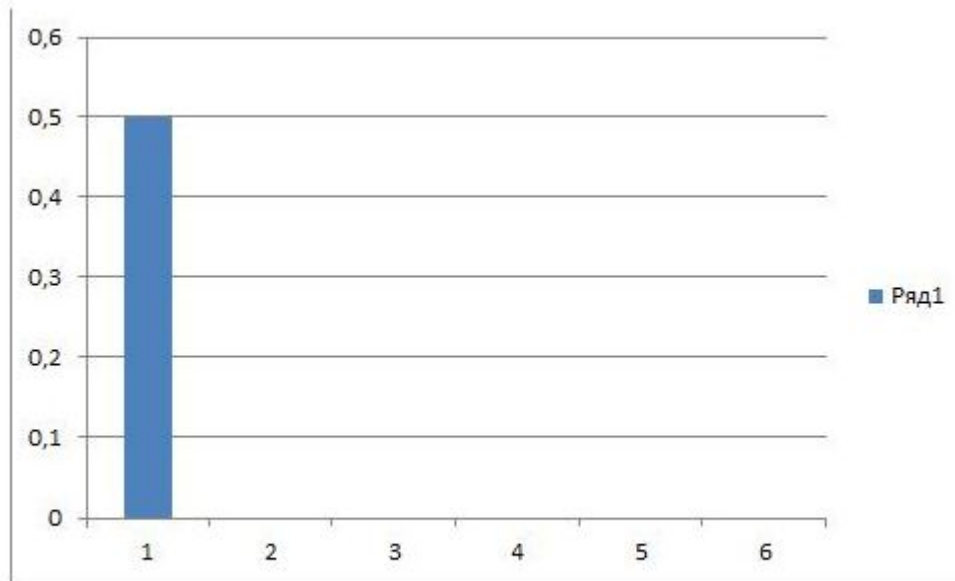
$$\rho_k = 0,5^k$$

$$(\rho_1 = 0,5; \rho_2 = 0,25; \rho_3 = 0,125)$$



Пример (AR(1)) - РАСФ

$$\psi_1=0,5, \psi_2=\psi_3=\dots 0$$



AR(p) – процесс авторегрессии порядка p

$$y_t = \delta + \theta_1 y_{t-1} + \dots + \theta_p y_{t-p} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0; \sigma^2).$$

Запись с использованием лагового оператора:

$$y_t = \delta + \theta_1 L y_t + \dots + \theta_p L^p y_t + \varepsilon_t$$

$$(1 - \theta_1 L - \dots - \theta_p L^p) y_t = \delta + \varepsilon_t$$

AR(p) – процесс авторегрессии порядка p

$$(1 - \theta_1 L - \dots - \theta_p L^p) y_t = \delta + \varepsilon_t$$

Характеристическое уравнение:

$$1 - \theta_1 z - \dots - \theta_p z^p = 0$$

Условие стационарности: все корни
характеристического уравнения по
модулю больше единицы $|z_i| > 1$

AR(p) – процесс авторегрессии порядка p

- Если характеристическое уравнение имеет корень $|z_i| \leq 1$, то говорят, что полином необратим, т.е. соответствующий ему временной ряд нестационарен.
- Полином называется необратимым, если он не может быть разложен в произведение двух непостоянных полиномов (непостоянный полином – если он не равен константе)
- Лаговые полиномы, корень которых по модулю меньше единицы, практически не встречаются на практике, т.к. такие процессы носят взрывной характер.
- На практике часто встречается случай, когда $z = 1$.

AR(p) – процесс авторегрессии порядка p

- На практике часто встречается случай, когда $z = 1$. Такая проблема получила название UR (unit root problem).
- Вспомним гипотезу для расширенного теста Дики-Фулера:

$H_0: \gamma = 0$, есть UR, процесс нестационарен;

$H_1: \gamma < 0$, нет UR, процесс стационарен.

- Т.е. для проверки ряда на стационарность необходимо составить уравнение для AR-процесса и проанализировать значения его корней

Пример

Проверим на стационарность процесс

$$y_t = 1 + 1,3y_{t-1} - 0,4y_{t-2} + \xi_t, \quad (\xi_t \sim WN(0; \sigma^2)).$$

Это процесс AR(2).

$$y_t = 1 + 1,3Ly_t - 0,4L^2y_t + \xi_t$$

$$y_t - 1,3Ly_t + 0,4L^2y_t = 1 + \xi_t$$

$$(1 - 1,3L + 0,4L^2)y_t = 1 + \xi_t$$

$$(1 - 1,3z + 0,4z^2)y_t = 1 + \xi_t$$

Решим характеристическое уравнение

$$1 - 1,3z + 0,4z^2 = 0$$

$$D = 1,3^2 - 4 \cdot 0,4 = 0,09$$

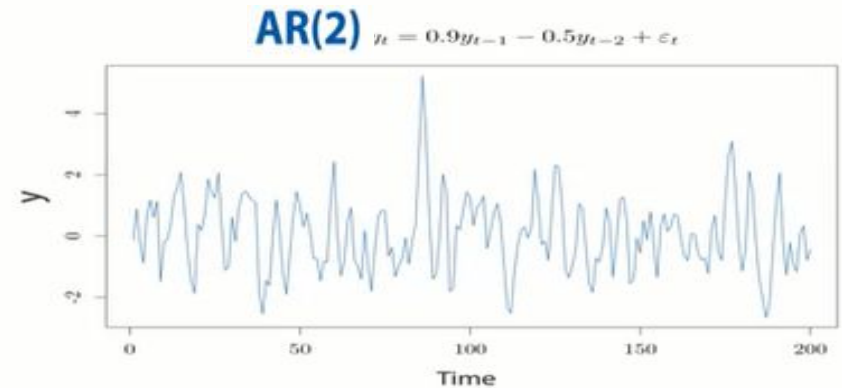
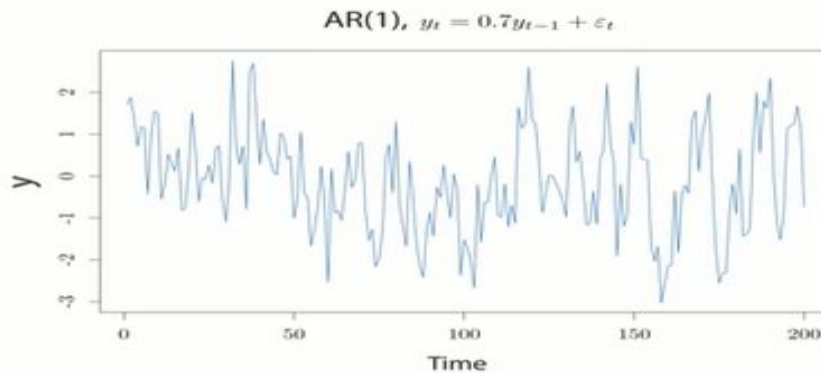
$$Z_1 = (1,3 - 0,3) / (0,8) = 1,25$$

$$Z_2 = (1,3 + 0,3) / (0,8) = 2$$

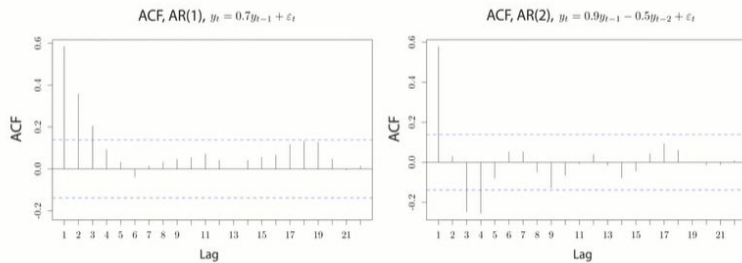
Процесс стационарен

Графики AR(1), AR(2)

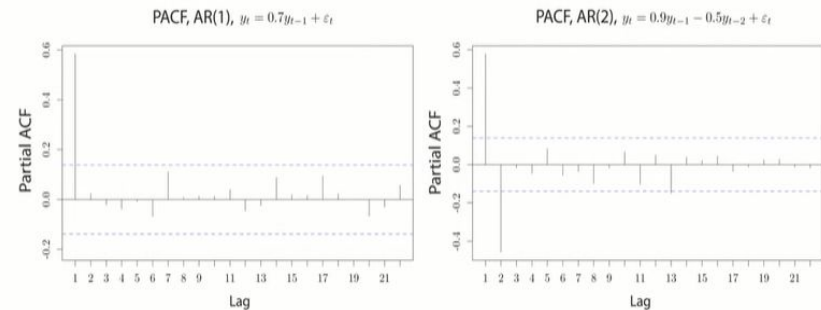
AR(1) и AR(2)



AR(1) и AR(2)



AR(1) и AR(2)



МА(1) – процесс скользящего среднего 1-го порядка

$$y_t = \delta + \varepsilon_t + \alpha \varepsilon_{t-1}$$

$$\mu = \delta,$$

$$\gamma_0 = (1 + \alpha^2)\sigma^2,$$

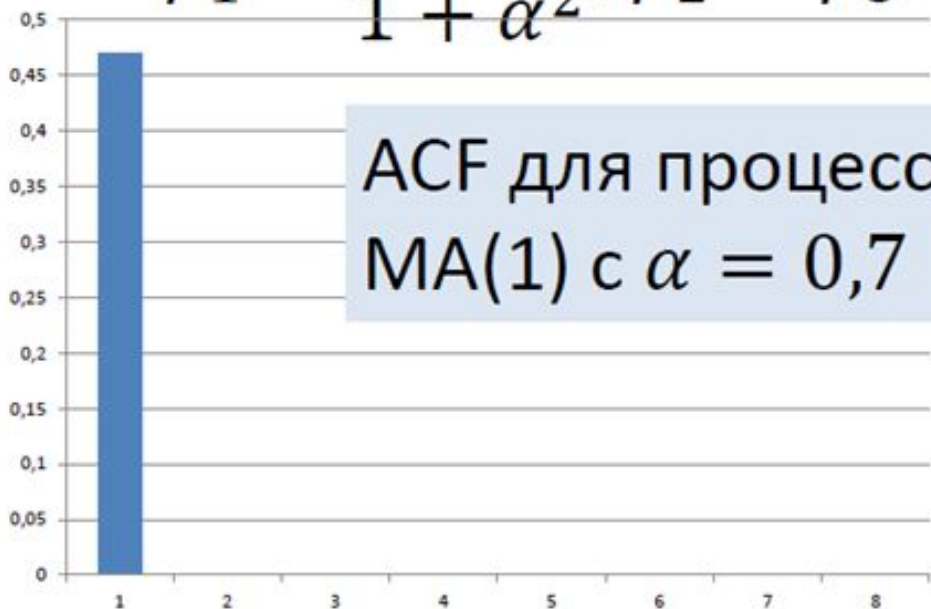
$$\gamma_1 = \alpha\sigma^2,$$

$$\gamma_2 = \gamma_3 = \dots = 0$$

МА(1) – процесс скользящего среднего 1-го порядка

$$y_t = \delta + \varepsilon_t + \alpha \varepsilon_{t-1}$$

$$\rho_1 = \frac{\alpha}{1 + \alpha^2}, \rho_2 = \rho_3 = \dots = 0$$



МА(q) – процесс скользящего среднего порядка q

$$y_t = \delta + \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \dots + \alpha_q \varepsilon_{t-q}$$

Процесс стационарен при любых значениях $\alpha_1, \dots, \alpha_q$.

Пример процесса МА(1) приведен на рис. 12.2.

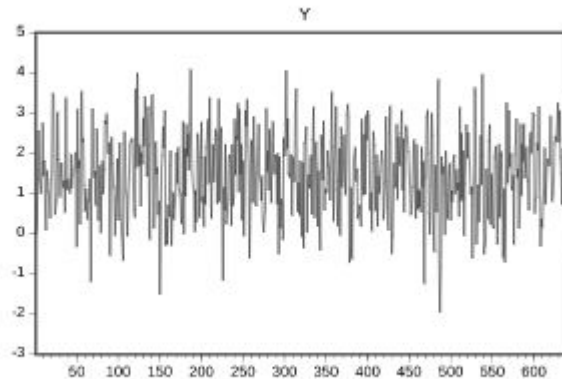


Рис. 12.2. Процесс МА(1): $y_t = 2 + \varepsilon_t + 0,3\varepsilon_{t-1}$, $\varepsilon_t \sim i.i.d.: N(0; 1)$

Процесс ARMA (p,q)

Естественное обобщение подхода к моделированию: объединение AR и MA процессов в одну модель ARMA

$$y_t = \delta + \theta_1 y_{t-1} + \dots + \theta_p y_{t-p} + \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \dots + \alpha_q \varepsilon_{t-q}$$

Процесс ARMA (p,q)

Естественное обобщение подхода к моделированию: объединение AR и MA процессов в одну модель ARMA

$$y_t = \delta + \theta_1 y_{t-1} + \dots + \theta_p y_{t-p} + \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \dots + \alpha_q \varepsilon_{t-q}$$

С помощью ARMA (p,q) можно компактно и очень точно описать любой стационарный процесс.

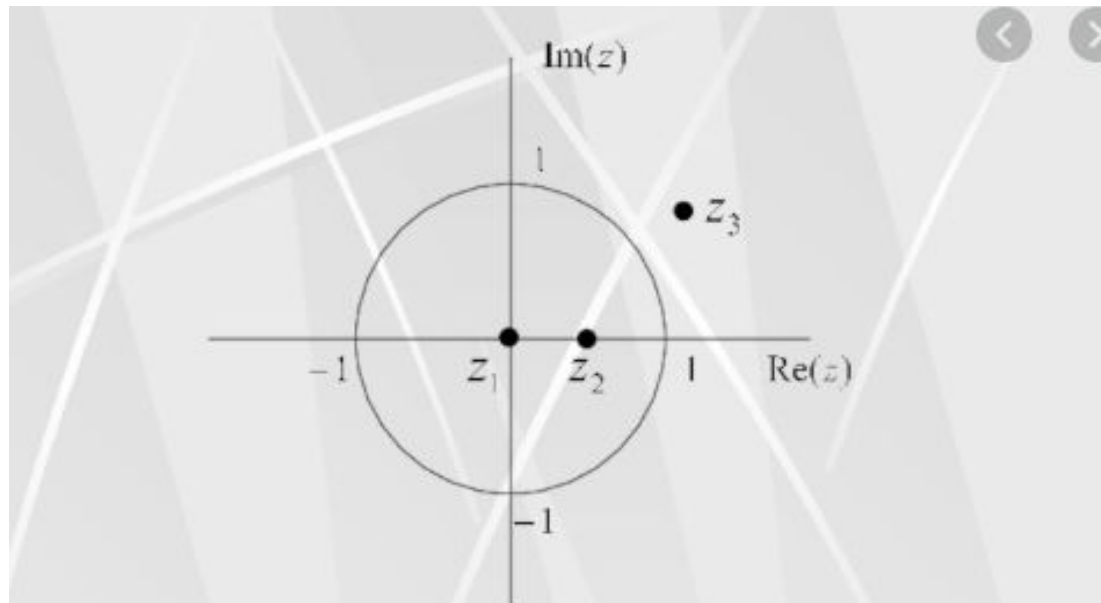
Пример

Проверим на стационарность процесс
 $y_t = 1,2 + 0,7y_{t-1} - 0,4y_{t-2} + \xi_t - 0,2 \xi_{t-1}$, ($\xi_t \sim WN(0; \sigma^2)$).

- Это процесс ARMA(2,1)
- $y_t - 0,7Ly_t + 0,4L^2y_t = 1,2 + \xi_t - 0,2 \xi_{t-1}$
- $(1 - 0,7z + 0,4z^2) y_t = 1,2 + \xi_t - 0,2 \xi_{t-1}$
- $1 - 0,7z + 0,4z^2 = 0$
- $D = -1,11$
- $Z_1 = (0,7 - i*1,054)/0,8 = 0,875 - 1,3175*i$
- $Z_1 = (0,7 + i*1,054)/0,8 = 0,875 + 1,3175*i$
- *Корни находятся за пределами единичного круга, поэтому процесс стационарен*

$$Z_1 = (0,7 - i * 1,054) / 0,8 = 0,875 - 1,3175 * i$$

$$Z_1 = (0,7 + i * 1,054) / 0,8 = 0,875 + 1,3175 * i$$



На втором рисунке показаны полюсы передаточной функции $Z_2(z)$ второго фильтра. Полюсов тоже три, два из них z_1, z_2 лежат внутри единичного круга $|z| < 1$, а один полюс z_3 вне единичного круга $|z| \geq 1$. Значит, второй фильтр неустойчив.

Случайное блуждание (Random walk, RW)

$$y_t = y_{t-1} + \varepsilon_t$$

Стационарен ли этот процесс?

$$(1 - L)y_t = \varepsilon_t$$

Характеристическое уравнение:

$$1 - z = 0$$

$z = 1$. **Единичный корень.** Процесс не стационарен

Случайное блуждание (Random walk, RW)



Пример RW: Обменный курс доллара США к немецкой марке (ежедневные данные, 2 января 1980 — 21 мая 1987)

Случайное блуждание (Random walk, RW)

Более общий случай — случайное блуждание с дрейфом:

$$y_t = \delta + y_{t-1} + \varepsilon_t$$

Как устроена первая разность для этого процесса?

Она стационарна:

$$y_t - y_{t-1} = \delta + \varepsilon_t$$
$$\Delta y_t = \delta + \varepsilon_t$$

Случайное блуждание (Random walk, RW)

Процесс случайного блуждания не стационарен, однако его первая разность является стационарным процессом.

На практике многие финансовые или макроэкономические переменные также не стационарны сами по себе но стационарны в разностях.

Порядок интегрированности ВР

- Если процесс стационарен, то он называется процессом, интегрированным нулевого порядка. Обозначение $I(0)$
- Если процесс не стационарен, а его первые разности стационарны, то он называется процессом, интегрированным 1-го порядка. $I(1)$

Порядок интегрированности ВР

- Если процесс y_t не стационарен, его первые разности Δy_t тоже не стационарны, но его вторые разности
$$\Delta^2 y_t = (\Delta y_t - \Delta y_{t-1})$$
стационарны, то процесс называется интегрированным 2-го порядка. $I(2)$
- Аналогично определяется процесс интегрированный k -го порядка. $I(k)$
- На практике почти не встречается процессов интегрированных выше второго порядка

Процесс ARIMA (p,d,q)

Если

- процесс является интегрированным k -го порядка,
- и его разность k -го порядка описывается процессом ARMA(p,q),

то исходный процесс называется интегрированным процессом авторегрессии со скользящим средним в остатках ARIMA(p,d,q)
(d=k)

Процесс ARIMA (p,d,q)

Пример: $y_t = 0,9y_{t-1} + 0,1y_{t-2} + \varepsilon_t$

$p = ?, k = ?, q = ?$

$$y_t - 0,9y_{t-1} - 0,1y_{t-2} = \varepsilon_t$$

$$(1 - 0,9L - 0,1L^2)y_t = \varepsilon_t$$

$$(1 - L)(1 + 0,1L)y_t = \varepsilon_t$$

Процесс содержит единичный корень, то есть нестационарен.

Рассмотрим первые разности процесса

Процесс ARIMA (p,d,q)

Пример: $y_t = 0,9y_{t-1} + 0,1y_{t-2} + \varepsilon_t$

Рассмотрим первые разности процесса:

$$y_t - y_{t-1} = 0,9y_{t-1} - y_{t-1} + 0,1y_{t-2} + \varepsilon_t$$
$$\Delta y_t = -0,1\Delta y_{t-1} + \varepsilon_t$$

Первые разности описываются
стационарным процессом AR(1) или, что
тоже самое ARMA(1,0).

Значит исходный процесс описывается
моделью ARIMA(1,1,0)

Прогнозирование в рамках модели $ARIMA(p,d,q)$

Главная цель построения $ARIMA$ моделей — прогнозирование будущих значений экономических переменных.

- Хорошая новость: $ARIMA$ модели при своей относительной простоте часто справляются с задачей прогнозирования лучше, чем сложные структурные модели.
- Плохая новость: При этом у $ARIMA$ моделей может не быть содержательной экономической интерпретации.

Как подобрать для временного ряда подходящую ARIMA – модель?

- Обычно смотрят на график приведенного к стационарному виду ВР и графики ACF, PACF
- Строят несколько моделей и из их выбирают наилучшую по критерию AIC
- В R-studio можно использовать функцию **autoarima**, она сама выберет модель с наименьшим значением AIC

Пример в R-studio

```
library(wooldridge)
library(timeseries)
library(foreign)
df = phillips# загружаем данные из пакета wooldridge
df
cinf=df$cinf
cinf_ts=ts(cinf)#переводим переменную cinf в формат временных рядов
plot.ts(cinf_ts)# строим график временного ряда
library(funitRoots)
adfTest(cinf_ts,lags=3,type="c")# проводим ADF-тест на стационарность

auto.arima(cinf_ts)#строим подходящую ARIMA-модель
```

```
> auto.arima(cinf_ts)
Series: cinf_ts
ARIMA(0,0,2) with zero mean

Coefficients:
      ma1      ma2
    0.1774 -0.6738
s.e.  0.1042  0.1000

sigma^2 estimated as 3.73: log likelihood=-114.01
AIC=234.02   AICc=234.49   BIC=240.04
```

Запишем соответствующее регрессионное уравнение

$$Y_t = \varepsilon_t + 0,1774 * \varepsilon_{t-1} - 0,6738 * \varepsilon_{t-2}$$

Пример в R-studio

```
library(timeseries)
library(forecast)
data=read.csv("GDP.csv",header=TRUE, sep=";",dec=".")
gdp=data$GDP[21:81]
gdp
plot.ts(gdp)
gdp_ts=ts(gdp)
acf(gdp)
pacf(gdp)
```

```
adfTest(data$GDP[21:81],lag=3,type="ct")
```

```
model=auto.arima(gdp)
model
```

Series: gdp
ARIMA(2,1,2) with drift

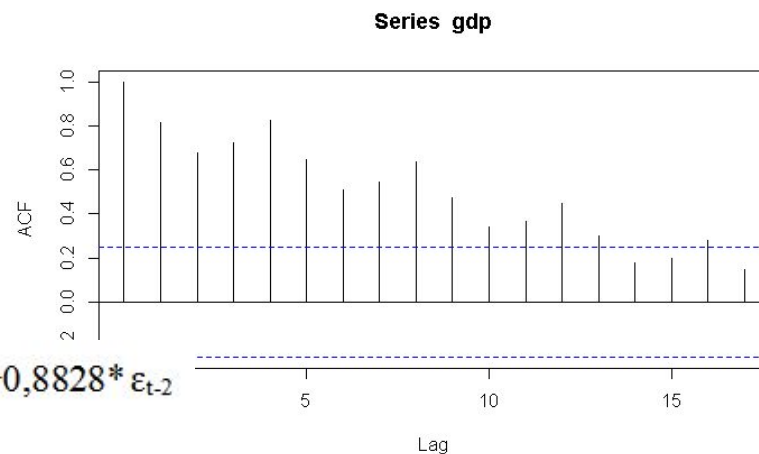
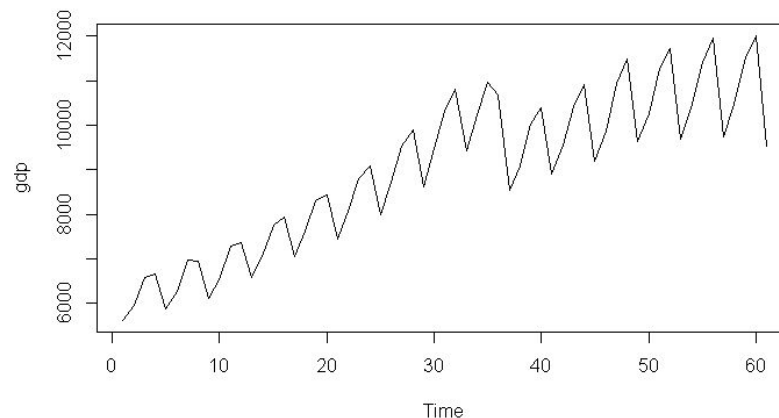
Coefficients:

	ar1	ar2	ma1	ma2	drift
	-0.0019	-0.9708	-0.5253	0.8828	77.5129
s.e.	0.0352	0.0243	0.0877	0.0701	37.7642

sigma^2 estimated as 196173: log likelihood=-451.3
AIC=914.6 AICC=916.18 BIC=927.16

Запишем соответствующее регрессионное уравнение

$$\Delta Y_t = 77,8129 - 0,0019 * \Delta Y_{t-1} - 0,9708 * \Delta Y_{t-2} + \varepsilon_t - 0,5253 * \varepsilon_{t-1} + 0,8828 * \varepsilon_{t-2}$$



Пример в R-studio

```
data<- read.csv("SP500.csv")
data_ts=ts(data)
data
plot.ts(data_ts)
acf(data)
adfTest(data_ts,lags=4,type="c")

model=auto.arima(data)
model
```

```
Series: data
ARIMA(0,1,0) with drift
```

```
Coefficients:
```

```
drift
```

```
0.9711
```

```
s.e. 0.4313
```

```
sigma^2 estimated as 159.2: log likelihood=-3380.23
AIC=6764.46 AICc=6764.47 BIC=6773.96
```

Пример в R-studio (сезонная ARIMA)

```
library(forecast)
```

```
library(timeSeries)
```

```
empl_manuf <- read.csv("empl_manuf.csv")
```

```
empl <- ts(empl_manuf, start = c(1992, 1), freq = 12)
```

```
empl
```

```
plot.ts(empl)
```

```
adf.test(empl)
```

```
model = auto.arima(empl)
```

```
> summary(model)
```

```
Series: empl
```

```
ARIMA(1,2,2)(1,0,2)[12]
```

```
Coefficients:
```

	ar1	ma1	ma2	sar1	sma1	sma2
	-0.7115	-0.2177	-0.3398	0.8829	-0.7542	0.0538
s.e.	0.1551	0.1836	0.1662	0.1358	0.1615	0.1593

```
sigma^2 estimated as 0.6336: log likelihood=-168.16
```

```
AIC=350.32 AICc=351.15 BIC=371.01
```

Алгоритм прогнозирования для временного ряда

1. Строим график временного ряда, графики АСФ, РАСФ. Делаем предварительные выводы о стационарности.
2. Тестируем ряд на стационарность (ADF-тест).
3. Если ряд нестационарный, приводим к стационарному виду (основное преобразование – взятие разностей, разностей второго порядка, логарифмирование).
4. Оцениваем модель ARIMA (p,d,q). Можно оценить несколько моделей и выбрать наилучшую по критериям AIC, BIC.
5. Проверяем остатки (WN).
 - Прогнозируем на основе модели п.4.
 - Оцениваем точность прогноза.

Проверка наличия автокорреляции в остатках

Рассматривая последовательность остатков как временной ряд, можно построить график их зависимости от времени. В соответствии с предпосылками МНК остатки ε_t должны быть случайными (рис. 6.1 а). Однако при моделировании временных рядов нередко встречается ситуация, когда остатки содержат тенденцию (рис. 6.1 б) и в) или циклические колебания (рис. 6.1 г)). Это свидетельствует о том, что каждое следующее значение остатков зависит от предшествующих. В этом случае говорят о наличии автокорреляции остатков.

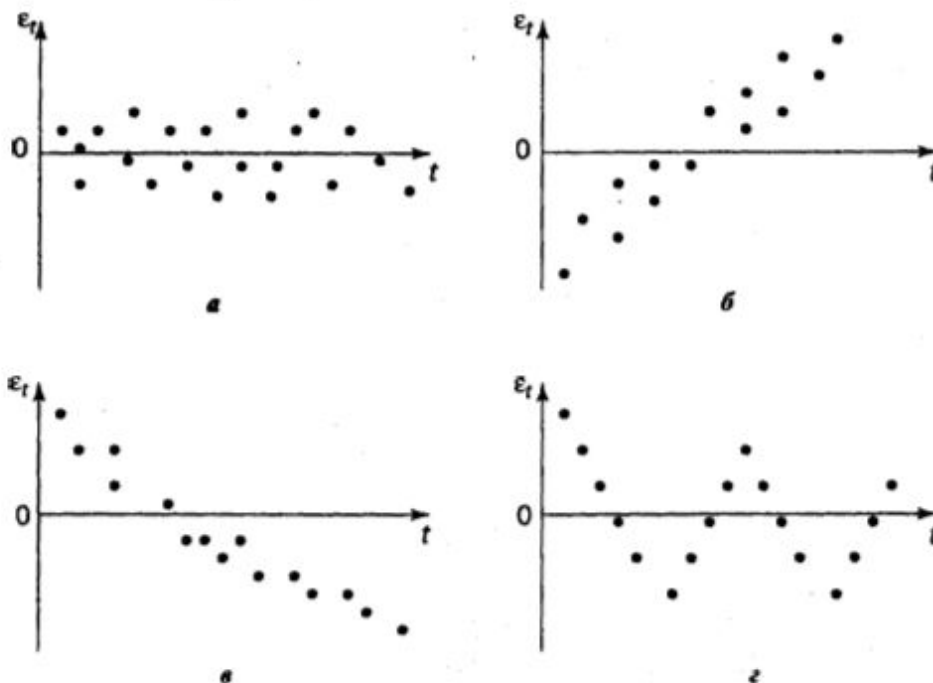


Рис. 6.1. Модели зависимости остатков от времени

а — случайные остатки; б — возрастающая тенденция в остатках;
в — убывающая тенденция в остатках;
г — циклические колебания в остатках

Проверка наличия автокорреляции в остатках

Причины автокорреляции ошибок следующие¹.

1. Систематические ошибки при измерении переменных. Ошибки измерения могут изменяться каким-то определенным образом во времени. Например, при развитии технологий многие переменные могут наблюдаться со значительно меньшими погрешностями, чем раньше (например, количество автомобилей в Москве).

2. Пропущенные переменные. Если в модели пропущены переменные, то они естественным образом попадают в остатки модели, а поскольку обычно пропущенные регрессоры имеют значимую автокорреляцию, то и в остатках также появляется автокорреляция.

3. Неправильная форма модели. Если подобрана неправильная спецификация модели (например, переменные не преобразованы необходимым образом), то это также может быть причиной возникновения автокорреляции.

Проверка наличия автокорреляции в остатках

13.3.3. Тест Бройша — Годфри

Данный тест не имеет таких ограничений, как тест Дарбина — Уотсона. Более того, используя данный тест, можно проверять гипотезу о наличии автокорреляции любых порядков. Нулевая гипотеза данного теста — автокорреляции в остатках нет, альтернативная — автокорреляция есть.

Так же как и для предыдущего теста, на первом шаге оценивается оригинальная модель и рассчитывается для нее ряд остатков. Затем оценивается дополнительная модель: $e_t = \beta_0 + \beta_1 e_{t-1} + \beta_2 e_{t-2} + \dots + \beta_p e_{t-p} + v_t$. На третьем шаге проверяется нулевая гипотеза о том, что модель неадекватна, против альтернативной, что модель адекватна:

$$H_0: \beta_0 = \beta_1 = \dots = \beta_p = 0;$$

$$H_1: \sum_{i=1}^p \beta_i^2 > 0.$$

Проверка наличия автокорреляции в остатках

$t = 1$

Расчетная статистика: $T \cdot R^2 \sim F(p, T - p - 1)$. Коэффициент R^2 берется из модели для остатков. Стоит также отметить, что можно получить представление о порядке автокорреляции, если проанализировать значимость отдельных коэффициентов в дополнительной модели. Также зачастую в дополнительную модель добавляют регрессанты из основной модели, чтобы нивелировать эффект от пропущенных переменных. Обычно при выборе количества лагов в модели следуют такому же правилу, как и при оценивании обычных моделей: двигаются от более общих моделей к частным (можно также основывать выбор на информационных критериях). Также необязательно автокорреляция может адекватно описываться процессом $AR(p)$ конечного порядка, однако на практике это ограничение обычно не создает проблем.

Пример

Протестируем остатки модели на автокорреляцию. Построим коррелограмму остатков (рис. 13.32).

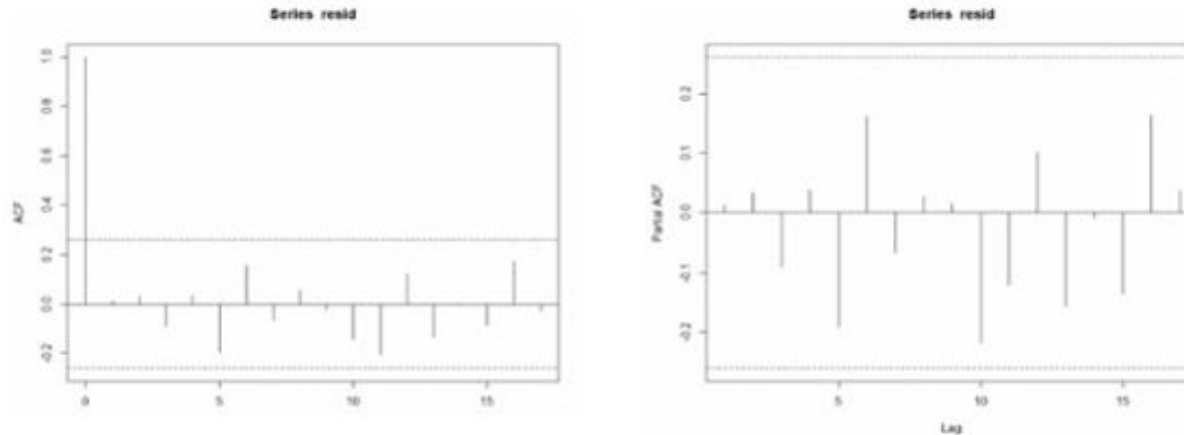


Рис. 13.32. Коррелограмма остатков модели динамики ВВП

Исходя из анализа коррелограммы можно заключить, что остатки похожи на белый шум. Проведем тест Бройша — Пагана — Годфри на выявление автокорреляции второго порядка:

```
|| bgtest(resid ~ 1, order = 2, type = "F")
```

Получим

Breusch-Godfrey test for serial correlation of order up to 2

data: resid - 1

LM test = 0.034575, df1 = 2, df2 = 53, p-value = 0.966

Можно заключить, что автокорреляции нет (ситуация будет аналогична и для автокорреляции первого порядка).

Прогнозирование

```
library(forecast)
library(timeseries)
library(fUnitRoots)
empl_manuf <- read.csv("empl_manuf.csv")
empl_manuf
empl <- ts(empl_manuf, start = c(1992, 1), freq = 12)
empl
plot.ts(empl)
adfTest(empl, lags=3, type="ct")

model=auto.arima(empl)
summary(model)
forecast(model, h=12, level=95)
autoplot(forecast(empl, h=12))
```

Оценка точности прогноза

Используя такой метод построения прогнозов, можно получать достаточно качественные точечные прогнозы для исследуемого временного ряда. Однако возникает релевантный вопрос: как определить качество прогноза? Обычно для ответа на этот вопрос пользуются таким понятием, как «критерий качества прогноза». Существует несколько популярных таких критериев. Первый из них — *MAPE* (*mean absolute percentage error*):

$$100\% \frac{1}{n} \sum_{i=1}^n \left| \frac{y_{t+i} - \hat{y}_{t+i}}{y_{t+i}} \right|,$$

где y_{t+i} — реальное значение ряда на момент времени $(t + i)$; \hat{y}_{t+i} — точечный прогноз.

Другой мерой качества прогноза является *MAE* (*mean absolute error*):

$$\frac{1}{n} \sum_{i=1}^n |y_{t+i} - \hat{y}_{t+i}|.$$

Оценка точности прогноза

Также часто используют *RMSE* (*root mean square error*):

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (y_{t+i} - \hat{y}_{t+i})^2}.$$

Критерий *MAPE* является относительной мерой прогноза, поэтому он нечувствителен к изменению масштаба переменных, более того, он не столь чувствителен к выбросам. Критерий *RMSE* наиболее чувствителен к выбросам. Для оценки качества прогноза всю выборку необходимо разделить на две части: обучающую и тестовую. По обучающей выборке оценивается модель, а по тестовой рассчитывается качество прогноза. Обычно все наблюдения, за исключением последних двух-трех, добавляют в обучающую выборку, а оставшиеся оставляют для тестовой. Для моделей типа *ARIMA*(p, d, q) обычно рекомендуется строить краткосрочные прогнозы, т.е. обычно не более чем на три шага вперед, так как после этого качество прогноза зачастую очень сильно падает.

Пример

#Создаем обучающую выборку

```
empl_0 = ts(head(empl,(nrow(empl)-3)))#убираем последние 3 наблюдения  
empl_0
```

#Создаем тестовую выборку

```
empl_test=ts(tail(empl,3))  
empl_test
```

#Строим модель по обучающей выборке

```
model=auto.arima(empl_0)
```

#Делаем прогноз

```
fore_1 = forecast(model,h=3)  
fore_1
```

Оцениваем точность прогноза

```
empl_test=c(100.04,99.63,99.21)  
accuracy(fore_1,empl_test)
```


Сравнение прогнозных моделей

- Иногда необходимо сравнить прогноз, полученный по сложной модели, с прогнозом, полученным по простой.
- Серьезный вопрос: нужно ли тратить время на построение сложной модели, если простая модель обеспечивает сравнимую прогнозную точность?
- В чем состоят эти простые полезные альтернативы, которые могут играть роль эталона (benchmark)?

Сравнение прогнозных моделей

Прогноз по среднему:

```
meanf(x, h = 12)
```

Наивный прогноз

```
naive(x, h = 12)  
rwf(x, h = 12)
```

Сезонный наивный прогноз

```
snaive(x, h=12)
```

Прогноз по модели со сдвигом

```
rwf(x, drift = TRUE, h=12)
```

#Сезонный наивный прогноз

```
foresnaive<-snaive(empl_0, h=3)
```

```
foresnaive
```

```
accuracy(foresnaive, empl_test)
```

#Прогноз по модели со сдвигом

```
forerwf<-rwf(empl_test, drift=TRUE, h=3)
```

```
forerwf
```

```
accuracy(forerwf, empl_test)
```

Сравнение прогнозных моделей

Метод среднего

- Прогноз по всем будущим значениям равен среднему по историческим данным $\{y_1, \dots, y_T\}$.
- Прогнозы:

$$\hat{y}_{T+h|T} = \bar{y} = \frac{1}{T} \sum_{i=1}^T y_i$$

- Это прогноз с помощью модели WN с константой:
 $y_t = \mu + \nu_t$

Наивный метод

- Прогноз равен последнему наблюдаемому значению.
- Прогноз:

$$\hat{y}_{T+h|T} = y_T$$

- Это прогноз по модели RW: $y_t = y_{t-1} + \nu_t$

Сезонный наивный метод

- Прогноз равен прошлому значению того же сезона.
- Прогноз:

$$\hat{y}_{T+h|T} = y_{T+h-km},$$

где m - период сезонности и k равно $\lfloor (h-1)/m \rfloor + 1$, где $\lfloor \cdot \rfloor$ означает целую часть

- Это прогноз на основе сезонной модели случайного блуждания: $y_t = y_{t-m} + \nu_t$

Метод сдвига (Drift method)

- Прогноз равен сумме последнего доступного значения и среднего изменения:
- Прогнозы:

$$\begin{aligned} y_{T+h,T} &= y_T + \frac{h}{T-1} \sum_{t=2}^T (y_t - y_{t-1}) = \\ &= y_T + \frac{h}{T-1} (y_T - y_1) \end{aligned}$$

- Эквивалентно продолжению прямой линии, соединяющей первое и последнее наблюдения.
- Это прогноз в соответствии с моделью RW со сдвигом:
 $y_t = \mu + y_{t-1} + \nu_t$