Выполнила: Пак Елена,
студентка магистратуры факультета социальных наук НИУ ВШЭ
Домашняя работа. «Прикладной анализ временный рядов»

Временным рядом называют случайный процесс с дискретным временем. Реализация временного ряда (X_t) на множестве T={1,2,…200} представлена на рисунке 1.

Цель: построить вероятностную модель временного ряда.

Рисунок 1.

Определение типа временного ряда

Как видно на рисунке 1, ряд не имеет постоянное математическое ожидание для каждого t. Следовательно, он не является стационарным. Для начала, необходимо привести ряд к стационарному виду. Для этого, необходимо определить, относится ли ряд к типу TSP или DSP. Если ряд является типом DSP, то у временного ряда дисперсия зависит от времени, и для остационаривания временного ряда применяют подход Бокса-Дженкинса путем использования разностного оператора. Если ряд принадлежит к группе TSP, то он является нестационарным из-за наличия детерминированного тренда. Разграничив два типа, можно правильно определить случайную составляющую. Для этого, воспользуемся процедурой Доладо-Дженкинса-Сосвилла -Риверо.

Первый шаг процедуры Доладо-Дженкинса-Сосвилла -Риверо состоит в решении проблемы единичного корня для модели с трендом и дрифтом:

$$X_t = \mu + bt + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim WNN(0, \sigma^2) \quad (1)$$

$$\Delta X_t = \mu + bt + \gamma X_{t-1} + \sum_{j=1}^{p-1} \theta_j \Delta X_{t-1} + \varepsilon_t \quad (2)$$

$$\text{Где } \gamma = \sum_{j=1}^{p} \alpha_j - 1$$

Для данных из 200 наблюдений принято начинать проверять модель порядка p = $[\sqrt[3]{200}] = 5$.

Нулевая гипотеза состоит в том, что единичный корень присутствует:

$$H_0: \sum_{j=1}^{p} \alpha_j = 1, \text{ или} \gamma = 0$$

Альтернативная гипотеза состоит в том, что ряд является типом TSP:

$$H_A: \sum_{j=1}^{p} \alpha_j < 1, \text{ или} \gamma < 0$$

Для проверки гипотезы применим статистику Дики-Фуллера:

$$DF_{tr} = \frac{\hat{\gamma}}{S(\hat{\gamma})}$$

Получим следующие результаты расширенного критерия Дики-Фуллера:

```
	Augmented Dickey-Fuller Test

data:  var
Dickey-Fuller = -5.2265, Lag order = 5, p-value = 0.01
alternative hypothesis: stationary
```

Результаты расширенного теста Дики-Фуллера показывают, что $\hat{DF}_{tr} = -5.23$. А истинное значение статистики при нулевой гипотезе на 99% доверительном уровне не меньше, чем -4.04. Оцененное значение статистики Дики-Фуллера меньше, чем критическая точка. Таким образом, на 99% доверительно уровне мы отвергаем нулевую гипотезу и можем сделать вывод, что данный ряд принадлежит к типу TSP, а модель описывается выражением (1) с $\sum_{j=1}^{p} \alpha_j < 1$.

Детрендирование ряда

Так как временной ряд принадлежит к типу TSP, то можно представить его в следующим виде:

$$X_t = f(t) + y_t$$

$f(t)$ – детерминированная часть

y_t – случайная часть с нулевым средним

Далее, проанализируем детерминированную часть. Предположим, что она описывается линейной функцией:

$$f(t, \theta) = \theta_0 + \theta_1 t$$

С помощью оценки МНК оценили θ_0 и θ_1 для модели с трендом и константой. Результаты регрессионного анализа представлены ниже.

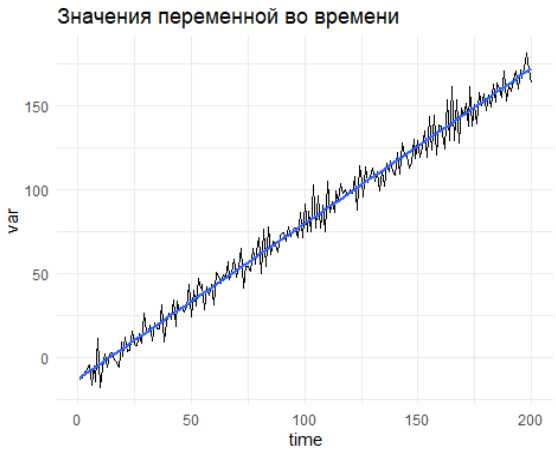
```
Call:
lm(formula = var ~ time, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-14.966  -5.121   0.134   4.468  21.618 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12.920005   1.031082  -12.53   <2e-16 ***
time          0.925643   0.008896  104.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.264 on 198 degrees of freedom
Multiple R-squared:  0.982,	Adjusted R-squared:  0.9819 
F-statistic: 1.083e+04 on 1 and 198 DF,  p-value: < 2.2e-16
```

Рисунок 2.



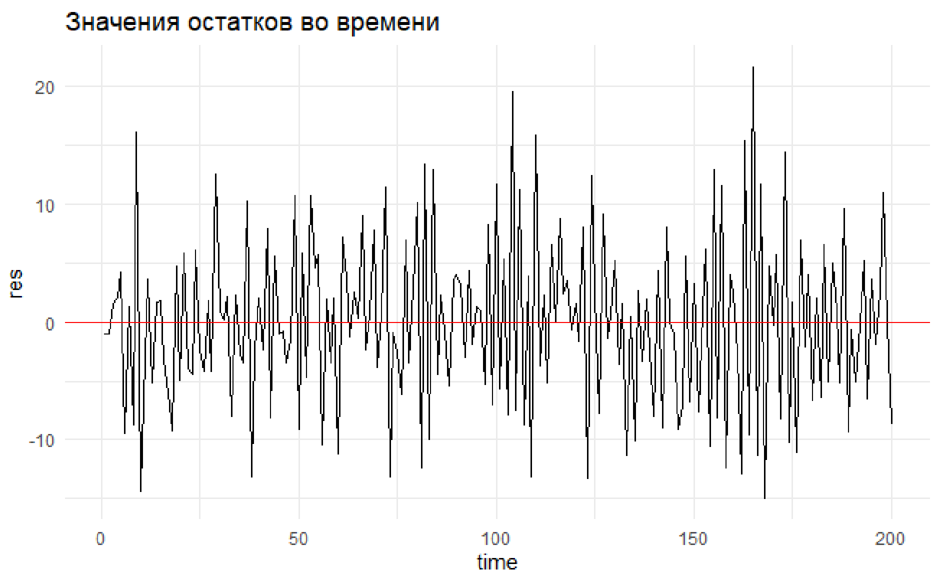
Как видно из модели, R-квадрат равен 0.982, что говорит о хорошем качестве модели. Также на рисунке 2 мы видим, что предсказанные значения вполне хорошо описывают ряд, так как линейный тренд прослеживается на протяжении всего ряда. Константа и коэффициент наклона оказались значимо отличными от нуля на вероятностном уровне 0.001. Таким образом, детерминированную часть запишем следующим образом:

$$\hat{f(t)} =- 12.92 + 0.926t$$

Удалим из первоначального временного ряда оцененную детерминированную часть, чтобы получить случайную часть ряда, которую можно выразить следующим образом:

$$\hat{y(t)} = X_t - 12.92 + 0.926t$$

Рисунок 3.



На рисунке 3, красной линией выделено среднее значение, равное нулю (так как использовали МНК оценку). Мы видим, что дисперсия постоянна во времени. Делаем вывод, что полученный ряд является стационарным.

Идентификация случайной части

Для того, чтобы идентифицировать стационарный ряд, мы можем рассчитать выборочную автокорреляционную ($\rho(\hat{k})$) и выборочную частную автокорреляционную функции ($\Phi(\hat{k})$) и выбрать подходящую линейную модель среди MA(q), AR(p) или ARMA(p,q). Выборочная оценки для АКФ и ЧАКФ являются асимптотически несмещенной.

Для белого шума АКФ равна нулю. А выборочная АКФ распределена асимптотически нормально со средним -1/n и дисперсией 1/n. При больших значениях n среднее становится близкой к нулю. Чтобы оценить равенсто АКФ нулю для белого шума, можно найти доверительный интервал для АКФ для белого шума, используя квантили для нормального распределения.

$$\rho(k) \sim N(\frac{1}{n}, \frac{1}{n})$$

$$P(-\ L < \ \rho(k) < L) = \ 0.95$$

$$L\ = \frac{1.96}{\sqrt[2]{n}} = \frac{1.96}{\sqrt[2]{200}} \approx 0.14$$

Таким образом, значения выборочной АКФ должны находиться в доверительной трубочке [-0.14 ; 0.14], при условии, что истинное значение АКФ равно 0.

Предполагается, что 20-30 шагов достаточно, чтобы понять, является ли процесс белым шумом или нет.

Для процесса MA(q) АКФ(k) не равна нулю для значений $k \in [1; q]$, но равна нулю для значений больше q. Выборочная АКФ для процесса MA(q) для всех значениях $k\ >\ q$ при больших значениях n имеет распределение:

$$\rho(k) \sim N(0, \frac{1}{n})$$

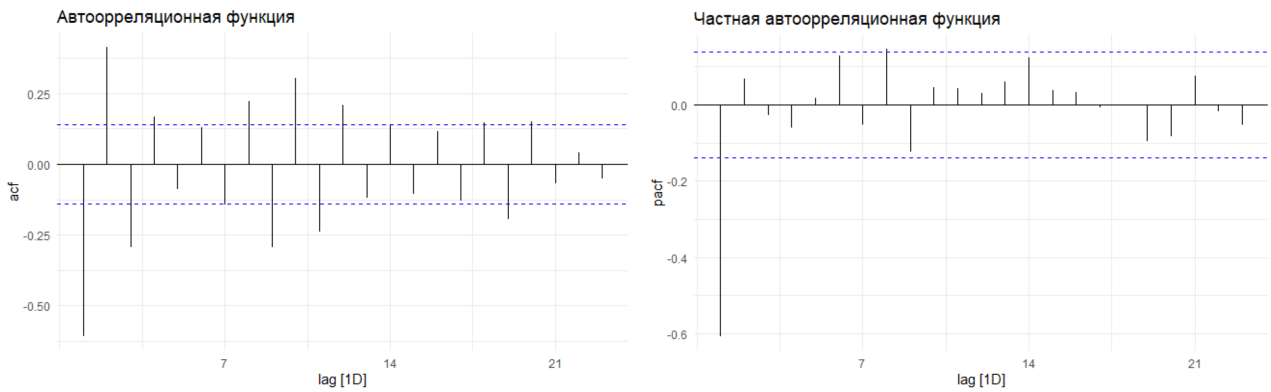
Для процесса AR(p) мы не предполагаем, что АКФ равна нулю, но при больших расстояниях k значение АКФ очень маленькое. ЧАКФ не равна нулю для значений $k \in [1; p]$, но равна нулю для значений больше p. А выборочная ЧАКФ для процесса AR(p) для всех значениях $k\ >\ p$ при больших значениях n имеет распределение:

$$\Phi(k) \sim N(0, \frac{1}{n})$$

Если АКФ и ЧАКФ не равны нулю на больших расстояниях, то нужно рассмотреть модель ARMA(p,q).

Для детрендированного ряда получили значения для выборочных АКФ и ЧАКФ (рисунок 4).

Рисунок 4.



Значения для выборочной АКФ			Значения для выборочной ЧАКФ		
	lag	acf		lag	pacf
1	1D	-0.608	1	1D	-0.608
2	2D	0.412	2	2D	0.0680
3	3D	-0.293	3	3D	-0.0291
4	4D	0.166	4	4D	-0.0607
5	5D	-0.0895	5	5D	0.0160
6	6D	0.128	6	6D	0.129
7	7D	-0.145	7	7D	-0.0535
8	8D	0.220	8	8D	0.146
9	9D	-0.294	9	9D	-0.123
10	10D	0.302	10	10D	0.0458
11	11D	-0.238	11	11D	0.0421
12	12D	0.209	12	12D	0.0295
13	13D	-0.120	13	13D	0.0589
14	14D	0.141	14	14D	0.124
15	15D	-0.107	15	15D	0.0375
16	16D	0.115	16	16D	0.0315
17	17D	-0.131	17	17D	-0.00725
18	18D	0.148	18	18D	-0.00279
19	19D	-0.194	19	19D	-0.0957
20	20D	0.150	20	20D	-0.0829
21	21D	-0.0700	21	21D	0.0762
22	22D	0.0390	22	22D	-0.0176
23	23D	-0.0505	23	23D	-0.0527

Сделаем выводы на основе полученных значений. Выборочная АКФ долго не заходит в доверительную трубочку, в то время как выборочная ЧАКФ заходит в трубочку на втором шаге. На основе это, можем предположить, что представленный ряд является процессом AR(1).

Далее, найдем коэффициенты для модели AR(1). Также, посмотрим на модель AR(2), чтобы сравнить с выбранной моделью.

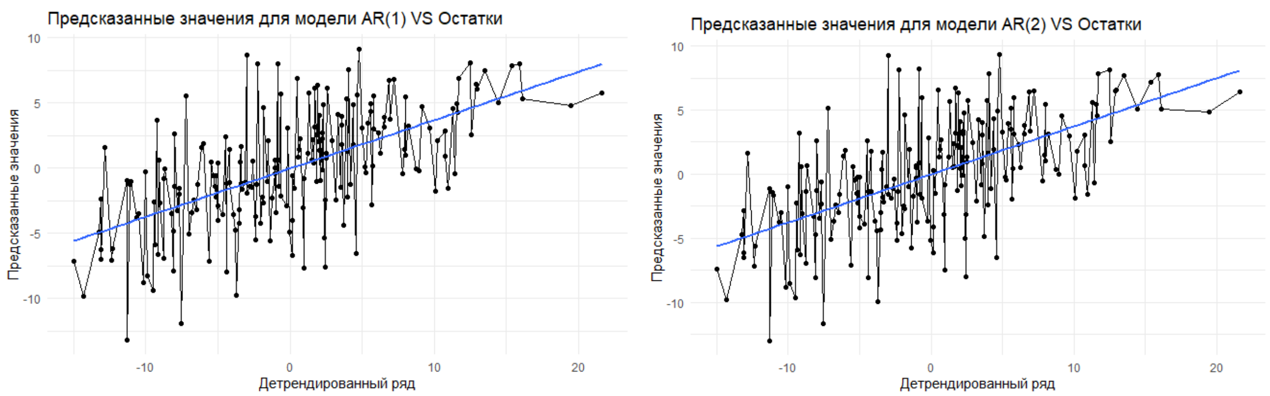
Если считать коэффициенты с помощью системы уравнений Юла-Уокера, то в качестве оценок коэффициентов мы бы взяли значения ЧАКФ. Таким образом, для AR(1) $\hat{\alpha}_1 = -\ 0.608$. Модель процесса AR(1) описывается следующим образом:

$$y_t = \alpha_1 y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim WNN(0, \sigma^2)$$

Подставив значение коэффициента, получим построенную модель для случайной составляющей ряда:

$$\hat{y}_t = -\ 0.608\hat{y}_{t-1} + \hat{\varepsilon}_t$$

Рисунок 5.



На рисунке 5 нарисованы графики для предсказанных значений случайной составляющей в модели AR(1) (слева) и AR(2) (справа) и детрендированный ряд. Как мы видим, они мало чем отличаются. Для первой модели дисперсия для остатков $\hat{\varepsilon}_t$ равна 33.26, а для второй - 33.27. Дисперсия для $\hat{\varepsilon}_t$ также включена в оценки качества моделей AIC, BIC. Данные информационные критерии показывают, что наиболее подходящей является модель AR(1).

Значения AIC для моделей AR(1), AR(2), AR(3)

```
  p AIC-Exact AIC-Approx
1 1  702.4575  -87.97258
2 2  703.6370  -86.05380
3 8  706.3423  -86.96146
```

Значения BIC для моделей AR(1), AR(2), AR(3)

```
  p BIC-Exact BIC-Approx
1 1  709.0541  -78.07763
2 2  713.5319  -72.86053
3 3  718.7501  -68.20755
```

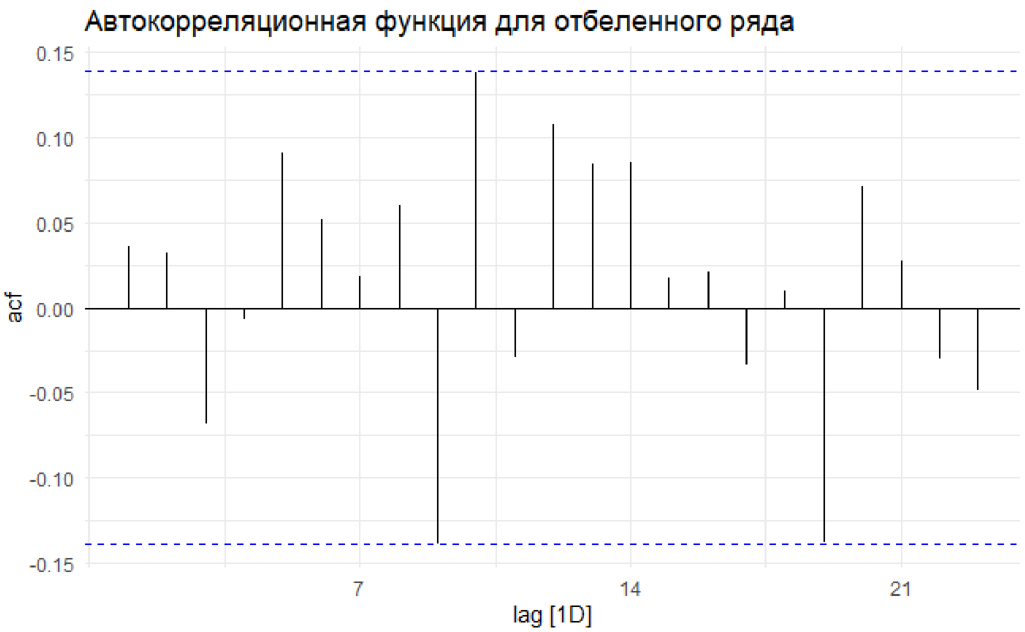
Проверка на белошумность остатков

После процедуры отбеливания получили остатки $\hat{\varepsilon}_t$. Для того, чтобы понять, подходит ли нам выбранная модель, необходимо сделать проверку на белошумность остатков:

$$\varepsilon_t \sim WNN(0, \sigma^2)$$

Во-первых, остатки должны быть некоррелированы и независимы. Для остатков была оценена АКФ (рисунок 6). Как видно на рисунке, значение выборочной АКФ не выходит за доверительную трубочку.

Рисунок 6.



Дополнительно, воспользуемся формальным тестом Люнга-Бокса. В качестве нулевой гипотезы проверяется совместное равенство нулю всех значений автокорреляции. Альтернативной является гипотеза, что хотя бы одно значение автокорреляции не равно нулю.

$$H_0\colon\ \rho_\varepsilon(1) = \rho_\varepsilon(2) = \ldots = \rho_\varepsilon(m) = 0$$

$$H_A\colon\ \sum_{j=1}^{m} \rho_j^2 > 0$$

Результаты теста Бокса-Люнга представлены внизу. Значение статистики оказалось не значимо. Таким образом, не отвергается нулевая гипотеза, на основании чего мы можем сделать вывод о том, что остатки некоррелируемы. Таким образом, мы выбрали подходящую модель для описания стационарного ряда.

```
	Box-Ljung test

data:  res2
X-squared = 0.26179, df = 1, p-value = 0.6089
```

Прогноз на один шаг

В среднеквадратическом смысле наиболее оптимальным в качестве прогноза последующего значения случайного процесса является условное математическое ожидание, при условии того, что мы знаем предыдущие значения реализации процесса. Так как для модели, которую мы определили AR(1) каждое значение зависит только от предыдущего значения на расстоянии 1, то предиктором будет условным математическим ожиданием от этой последней величины.

$$\varphi^*(y) = E(X) = E\left(y_{t-1}\right) = E\left(y_{t-1}\right) = E\left(y_{t-1}\right) + E\left(y_{t-1}\right) = \alpha_1 y_{t-1} + 0 = \alpha_1 y_{t-1} =$$

$$= \alpha_1 y_{200} = \ -0.608 * (-8.6785914) \approx 5.277$$

Мы сделали прогноз для случайной величины. Тогда для прогноза для значения изначального временного ряда прибавим детерминированную часть.

$$\hat{X}_t = f(t) + y_t = -12.92 + 0.926t + y_t$$

$$\hat{X}_{201} = -12.92 + 0.926t + y_t = -12.92 + 0.926 * 201 + y_{201} = 173.206 + 5.277 = 178.483$$

Получили прогноз временного ряда для $t = 201$: 178.483