# ARIMA и сезонная ARIMA

# Буковка I

# Буковка I: план

- Стационарность $ARMA$.
- Определение $ARIMA$.
- Нужно ли переходить к разностям?

# $ARMA$ процесс

**Определение**

$ARMA(p, q)$ процессом с несократимым уравнением

$$y_t = c + \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} + u_t + \alpha_1 u_{t-1} + \ldots + \alpha_q u_{t-q},$$

где $(u_t)$ — белый шум, $\beta_p \neq 0$ и $\alpha_q \neq 0$, называется решение этого уравнения вида $MA(\infty)$ относительно $(u_t)$.

**Определение с лагами**

$ARMA(p, q)$ процессом с уравнением

$$P(L) y_t = c + Q(L) u_t,$$

где $(u_t)$ — белый шум, $P(L)$ степени $p$ и $Q(L)$ степени $q$ несократимы, $P(0) = Q(0) = 1$, называется решение этого уравнения вида $MA(\infty)$ относительно $(u_t)$.

# Нюансы

- Процесс $y_t \sim ARMA(p, q)$ стационарен по определению: $\mathbb{E}(y_t) = \mu_y$, $\mathrm{Var}(y_t) = \gamma_0$, $\mathrm{Cov}(y_t, y_{t-k}) = \gamma_k$.
- В канонической записи $ARMA(p, q)$ процесса $P(L)y_t = c + Q(L)u_t$ у полинома $P(L)$ все корни $|\ell| > 1$.

  Возможны неканонические варианты.
- При оценке $ARMA(p, q)$ процесса методом максимального правдоподобия эти ограничения наложены а-приори.

  Есть упрощённые варианты правдоподобия.

# Что делать с нестационарными процессами?

**Определение**

Случайный процесс $(y_t)$ называется $ARIMA(p, 1, q)$ процессом относительно белого шума $(u_t)$, если $(y_t)$ нестационарен, но $\Delta y_t$ — стационарный $ARMA(p, q)$ процесс относительно белого шума $(u_t)$.

**Определение**

Случайный процесс $(y_t)$ называется $ARIMA(p, 2, q)$ процессом относительно белого шума $(u_t)$, если $(y_t)$ и $(\Delta y_t)$ нестационарны, но $\Delta^2 y_t$ — стационарный $ARMA(p, q)$ процесс относительно белого шума $(u_t)$.

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

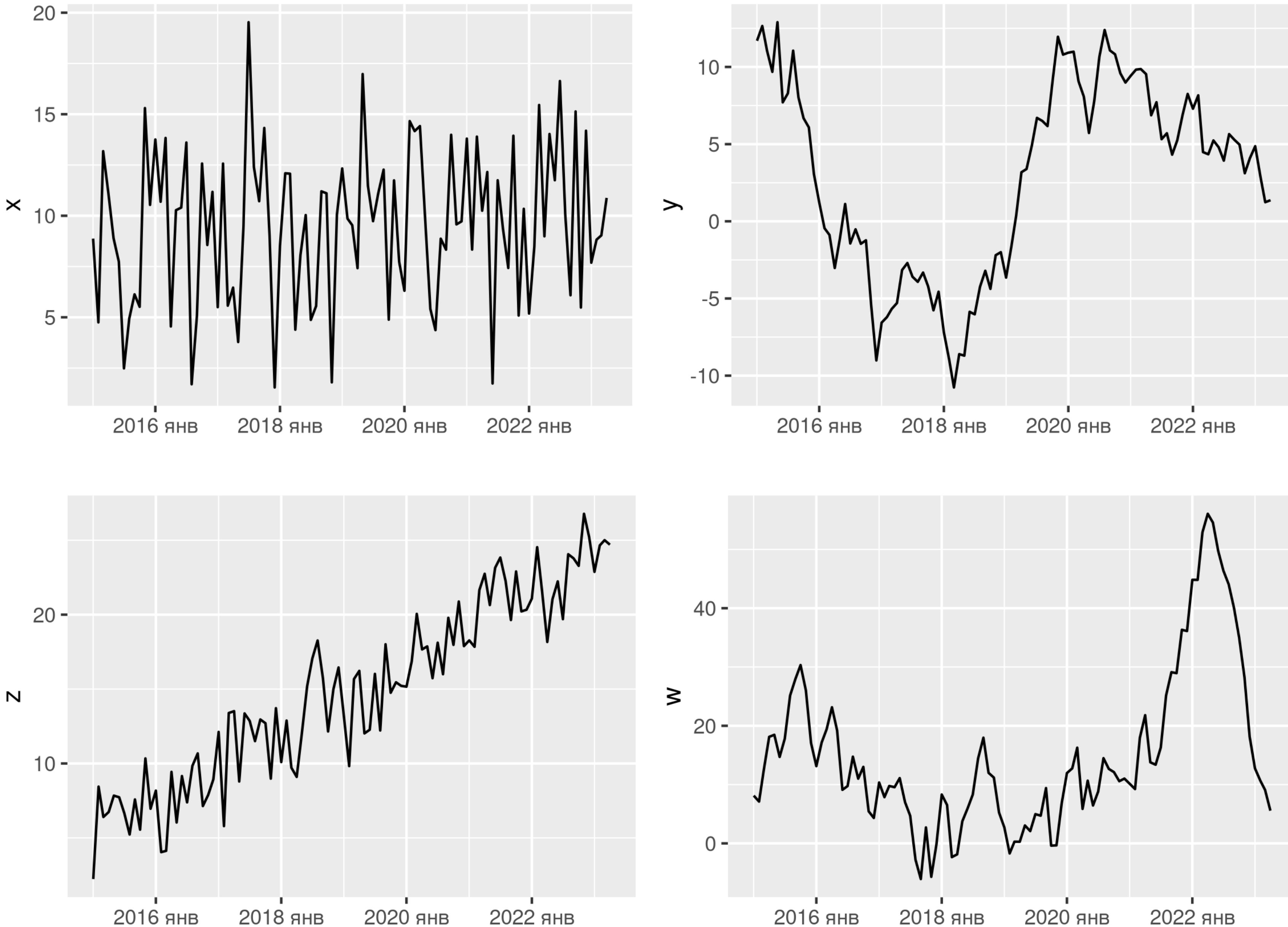
ARIMA — AutoRegressive Integrated Moving Average

# Как выбрать?

$ARIMA(p,0,q)$ или $ARIMA(p,1,q)$ или $ARIMA(p,2,q)$

- Посмотреть на график!
  График стационарного процесса колеблется в полосе постоянной ширины вокруг своего ожидания.
- Оценить все эти модели и выбрать наилучшую по кросс-валидации.
  Затратно по времени!
- Применять $AIC$ нельзя!
  $\ln L(y_1, \ldots, y_n \mid \theta)$ и $\ln L(y_2, \ldots, y_n \mid \theta, y_1)$ и $\ln L(y_3, \ldots, y_n \mid \theta, y_1, y_2)$ несравнимы!
- Есть тесты на единичный корень!
  ADF, KPSS, PP, …

# Выбираем «на глазок»



# Буковка I: итоги

- $ARMA$ подходит только для стационарных рядов.
- Иногда стационарен $\Delta y_t$ или $\Delta^2 y_t$.
- Выбираем между $ARMA$ и $ARIMA$.

# ADF тест

# ADF тест: план

- Предположения теста.
- Алгоритм теста.
- Три вариации теста.

# Зачем нужен ADF тест?

Хотим ответить на вопросы:

- Использовать $ARMA$ модель для $(y_t)$ или для $(\Delta y_t)$?
- Как включать константу в модель?

Название «тест на единичные корни»:

$$\Delta = 1 - L = P(L)$$

Уравнение $1 - \ell = 0$ имеет корень $\ell = 1$.

# ADF тест

| **Расшифровка** |
| --- |
| Augmented Dickey Fuller test |
| Расширенный тест Дики-Фуллера |

Три вариации теста: без константы, с константой, с трендом.

# ADF с константой

$$\Delta y_t = c + \beta y_{t-1} + d_1 \Delta y_{t-1} + \ldots + d_p \Delta y_{t-p} + u_t,$$

$H_0$: $\beta = 0$;

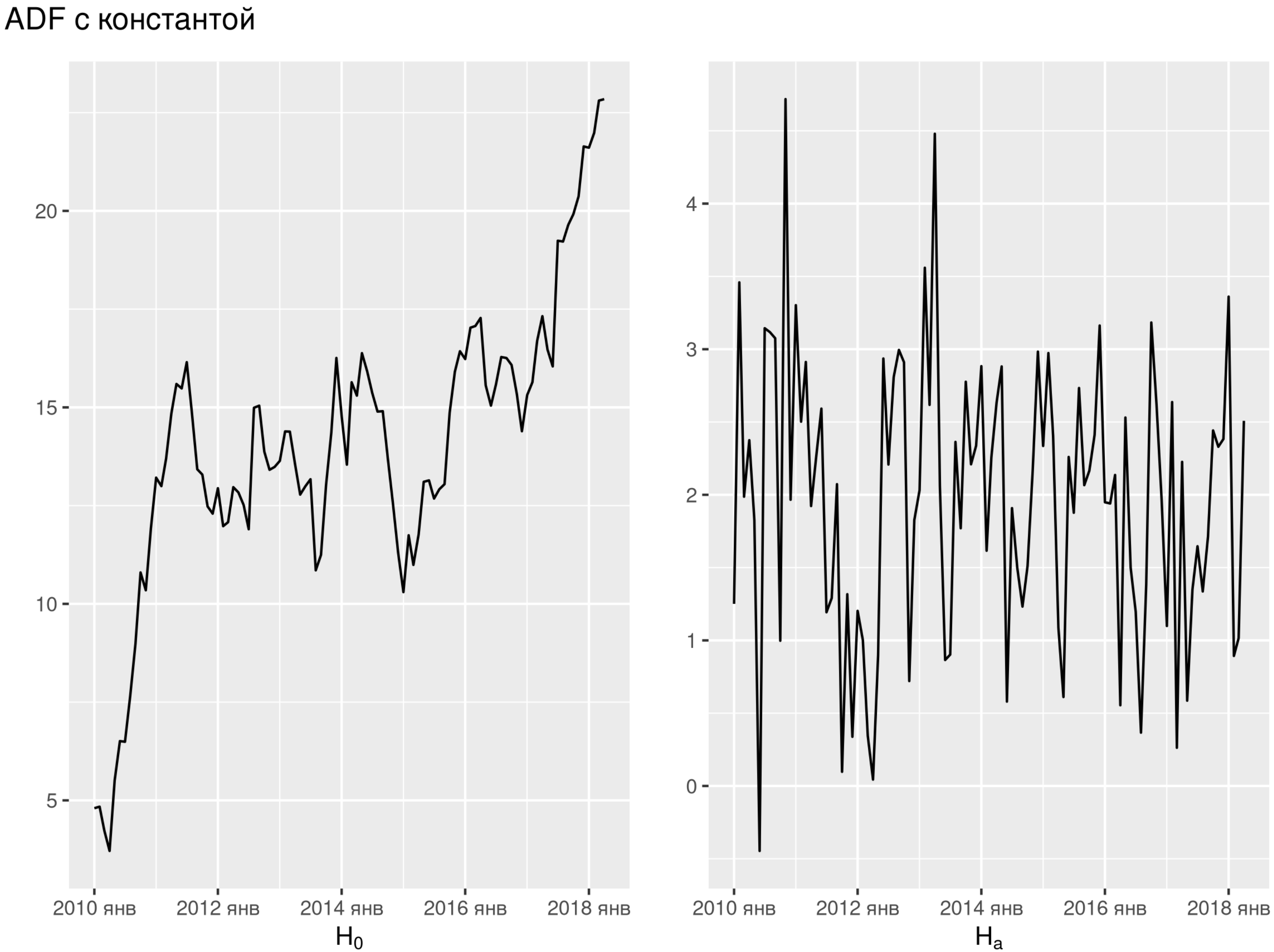
$(\Delta y_t)$ — стационарный $AR(p)$ процесс;

$y_t = y_0 + mt + \sum_{i=1}^{t} (\Delta y_i - \mathbb{E}(\Delta y_i))$;

$H_a$: $\beta < 0$;

$(y_t)$ — стационарный $AR(p+1)$ процесс;

# ADF с константой: $H_0$ и $H_a$



# ADF с константой: алгоритм

Шаг 1. Оцениваем регрессию

$$\widehat{\Delta y_t} = \hat{c} + \hat{\beta} y_{t-1} + \hat{d}_1 \Delta y_{t-1} + \ldots + \hat{d}_p \Delta y_{t-p}.$$

Шаг 2. Считаем по классической формуле $t$-статистику

$$ADF = \frac{\hat{\beta} - 0}{se(\hat{\beta})}.$$

При верной $H_0$ распределение $ADF$-статистики стремится к особому распределению $DF^c$!

Шаг 3. Делаем вывод:

Если $ADF < DF^c$, то $H_0$ отвергается.

# ADF без константы

$$\Delta y_t = \beta y_{t-1} + d_1 \Delta y_{t-1} + \ldots + d_p \Delta y_{t-p} + u_t,$$

$H_0$: $\beta = 0$;

$(\Delta y_t)$ — стационарный $AR(p)$ процесс с $\mathbb{E}(\Delta y_t) = 0$;

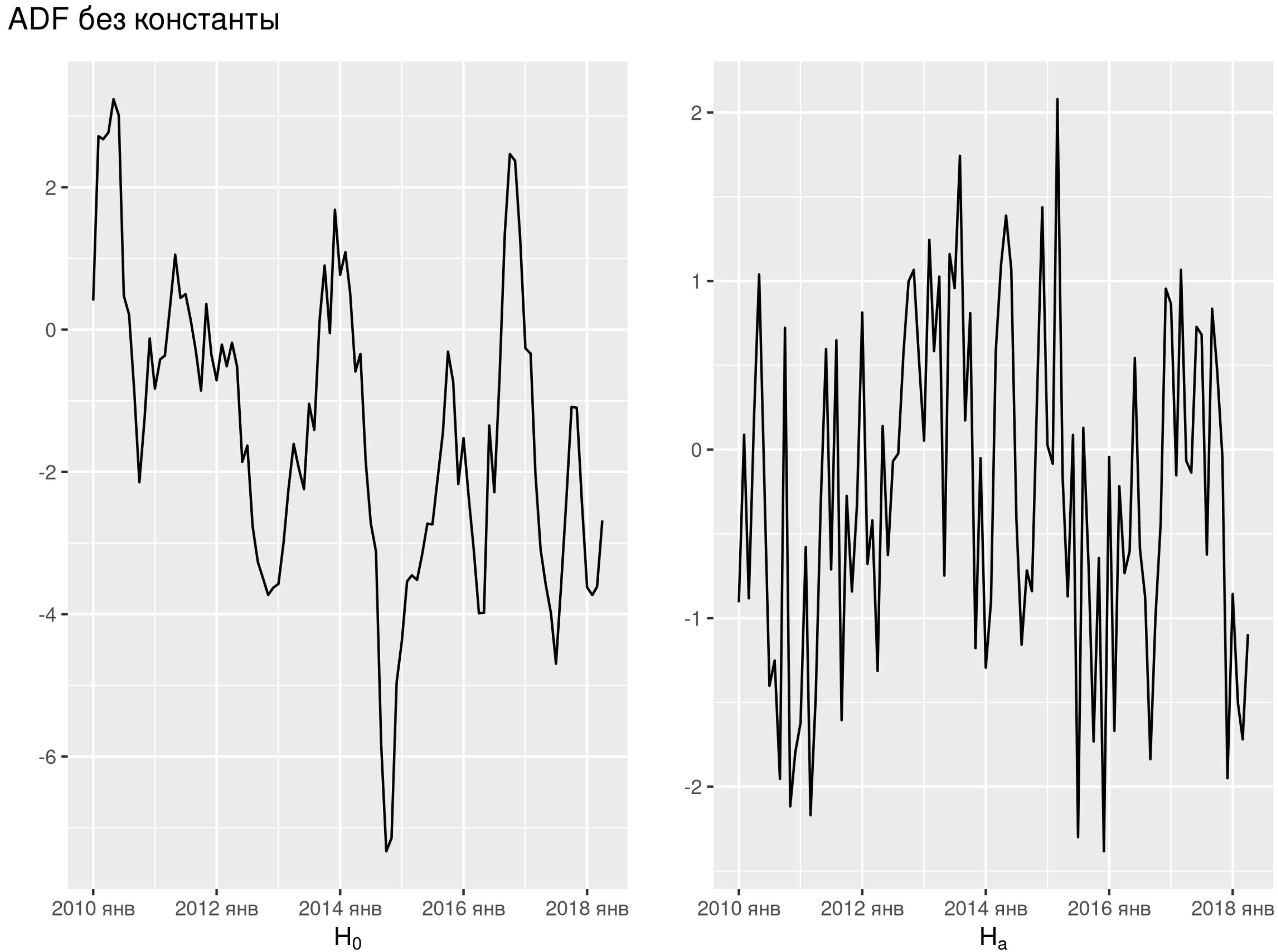
$y_t = y_0 + \sum_{i=1}^t \Delta y_i$;

$H_a$: $\beta < 0$;

$(y_t)$ — стационарный $AR(p+1)$ процесс с $\mathbb{E}(y_t) = 0$;

В алгоритме будет регрессия без константы и другое распределение $DF^0$.

# ADF без константы: $H_0$ и $H_a$



# ADF с трендом

$$\Delta y_t = c + gt + \beta y_{t-1} + d_1 \Delta y_{t-1} + \ldots + d_p \Delta y_{t-p} + u_t,$$

$H_0$: $\beta = 0$;

$\Delta y_t = k_1 + k_2 t + x_t$;

$(x_t)$ — стационарный $AR(p)$ процесс с $\mathbb{E}(x_t) = 0$;

$y_t = y_0 + m_1 t + m_2 t^2 + \sum_{i=1}^{t} x_i$;

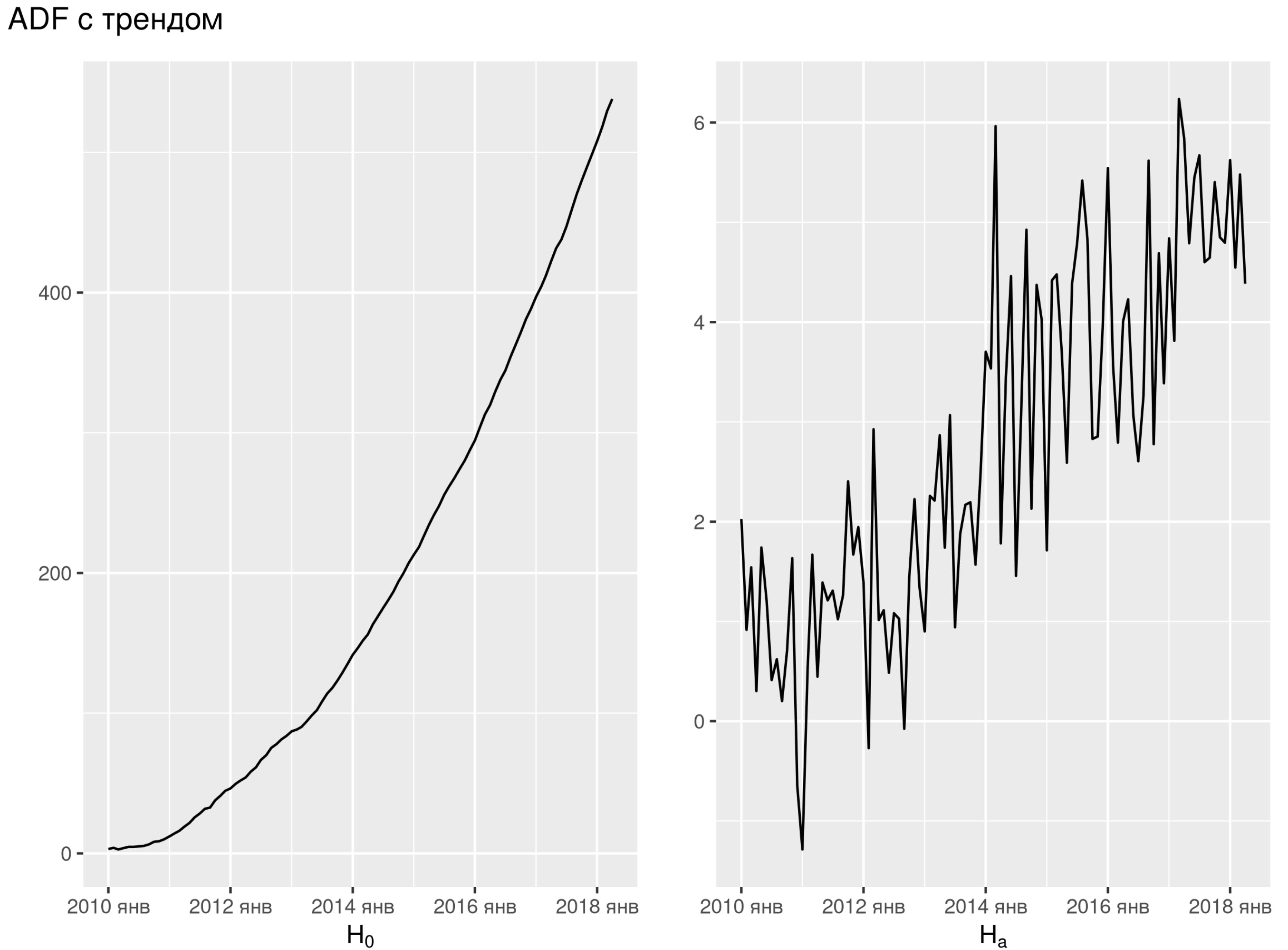
$H_a$: $\beta < 0$;

$y_t = m_1 + m_2 t + x_t$;

$(x_t)$ — стационарный $AR(p+1)$ процесс с $\mathbb{E}(x_t) = 0$;

В алгоритме будет регрессия с константой и трендом и другое распределение $DF^{ct}$.

# ADF с трендом: $H_0$ и $H_a$



# $ADF$ тест: итоги

- Применим для принятия решения о переходе к $\Delta y_t$.
- Есть три варианта теста с разными предпосылками.

# KPSS тест

# KPSS тест: план

- Долгосрочная дисперсия.
- Предпосылки теста.
- Две вариации теста.

# Зачем нужен KPSS тест?

Хотим ответить на вопросы:

- Использовать $ARMA$ модель для $(y_t)$ или для $(\Delta y_t)$?
- Как включать константу в модель?

# KPSS тест

| **Расшифровка** |
|---|
| Kwiatkowski–Phillips–Schmidt–Shin test<br><br>Тест Квятковского-Филлипса-Шмидта-Шина |

Две вариации теста: с константой, с трендом.

# Долгосрочная дисперсия

## Определение

Для стационарного процесса $(y_t)$ величина $\lambda^2$ называется долгосрочной дисперсией, если

$$\operatorname{Var}(\bar{y}) = \frac{\lambda^2}{T} + o(1/T)$$

или

$$\lim_{T\to\infty} T \operatorname{Var}(\bar{y}) = \lambda^2,$$

где $\bar{y} = (y_1 + \ldots + y_T)/T$.

## Мотивация

Для независимых наблюдений с одинаковой дисперсией

$$\operatorname{Var}(\bar{y}) = \frac{\sigma^2}{T}, \text{ где } \sigma^2 = \operatorname{Var}(y_i).$$

# KPSS с константой

$$y_t = c + rw_t + x_t,$$

$H_0$: $rw_t = 0$;

$(x_t)$ — стационарный процесс с $\mathbb{E}(x_t) = 0$;

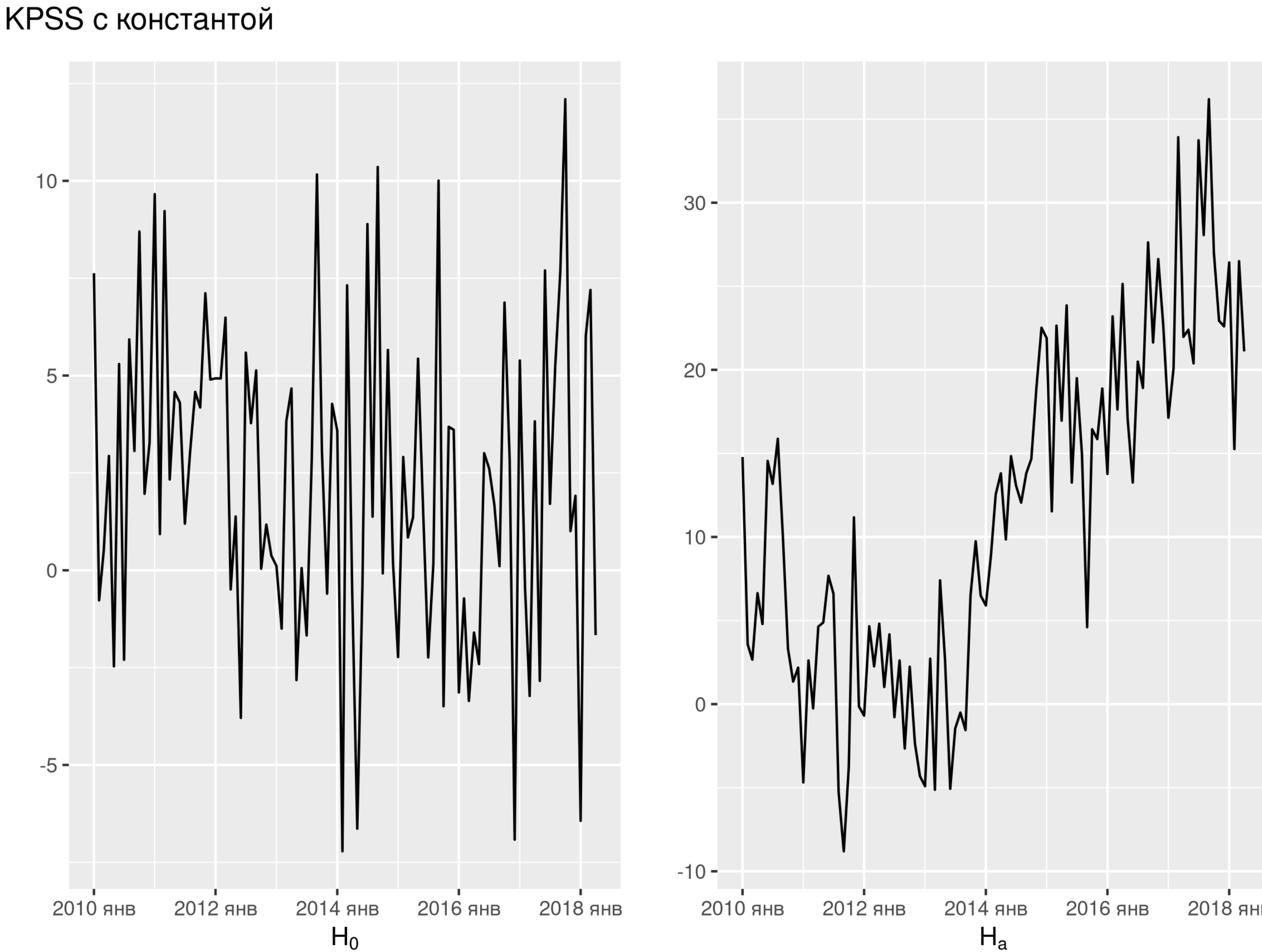
$H_a$: $rw_t = rw_{t-1} + u_t$;

$rw_0 = 0$;

$(x_t)$ — стационарный процесс с $\mathbb{E}(x_t) = 0$;

$(u_t)$ — белый шум, независимый с $(x_t)$.

# KPSS с константой: $H_0$ и $H_a$



# KPSS с константой: алгоритм

Шаг 1. Оцениваем регрессию на константу

$$\widehat{y_t} = \hat{c}.$$

Шаг 2. Считаем $KPSS$ статистику

$$KPSS = \frac{\sum_{t=1}^{T} S_t^2}{T^2 \hat{\lambda}^2},$$

где $S_t$ — накопленная сумма остатков, $S_t = \hat{u}_1 + \ldots + \hat{u}_t$,

а $\hat{\lambda}^2$ — состоятельная оценка долгосрочной дисперсии.

При верной $H_0$ распределение $KPSS$-статистики стремится к особому распределению $KPSS^c$!

Шаг 3. Делаем вывод:

Если $KPSS > KPSS^c$, то $H_0$ отвергается.

# KPSS с трендом

$$y_t = c + bt + rw_t + x_t,$$

$H_0$: $rw_t = 0$;

$(x_t)$ — стационарный процесс с $\mathbb{E}(x_t) = 0$;

$H_a$: $rw_t = rw_{t-1} + u_t$;

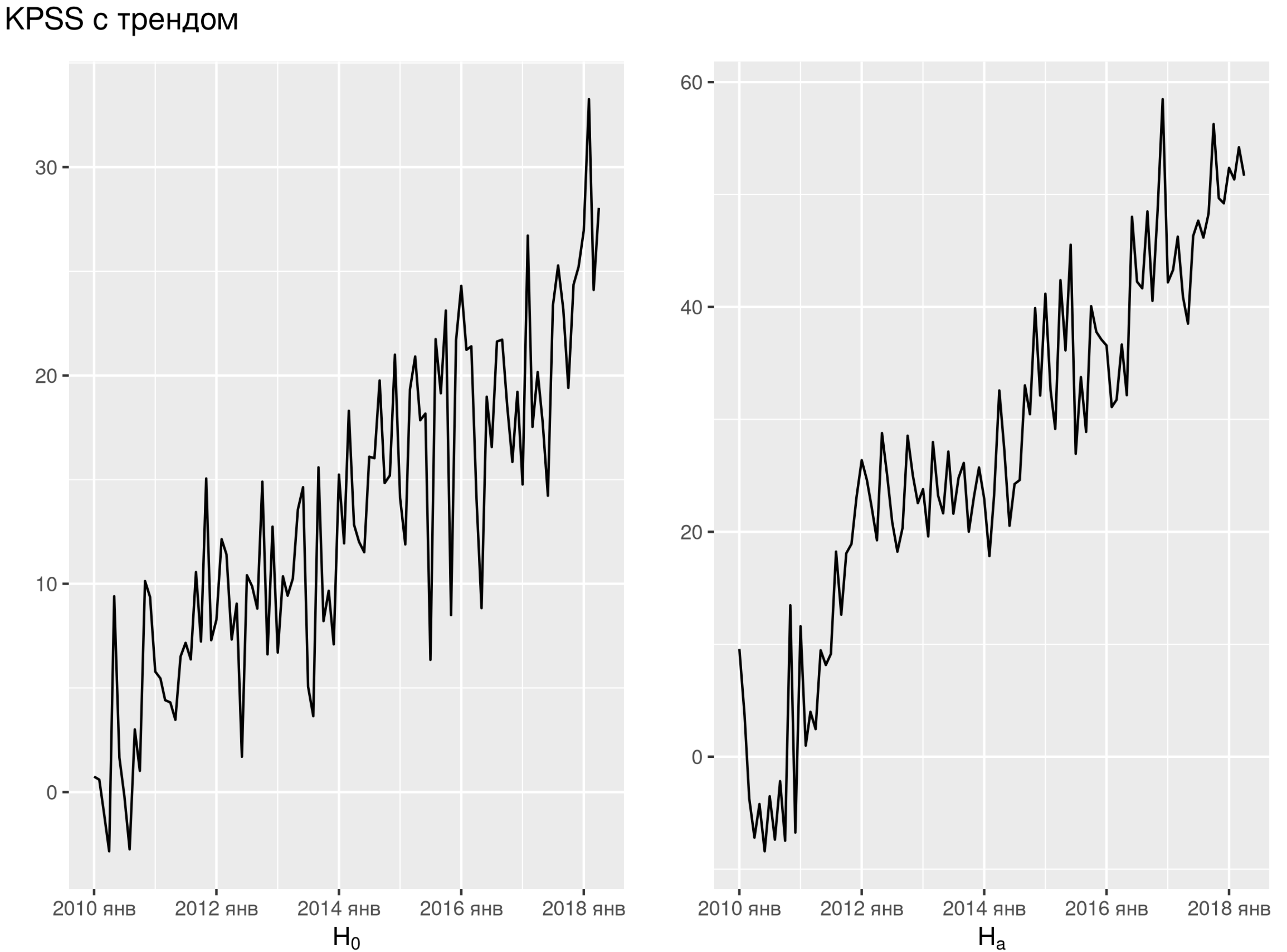
$rw_0 = 0$;

$(x_t)$ — стационарный процесс с $\mathbb{E}(x_t) = 0$;

$(u_t)$ — белый шум, независимый с $(x_t)$.

В алгоритме будет регрессия с константой и трендом и другое распределение $KPSS^{ct}$.

# $KPSS$ с трендом: $H_0$ и $H_a$



# Устоявшаяся терминология:

$$A. y_t = a + bt + x_t;$$

$(y_t)$ — стационарный вокруг тренда (trend stationary).

$(x_t)$ — стационарный процесс с $\mathbb{E}(x_t) = 0$.

Рецепт: оценим регрессию $a + bt$ с $ARMA$ ошибками для $(y_t)$.

$$B. y_t = a + \sum_{i=1}^{t} x_i \text{ или } y_t = a + bt + \sum_{i=1}^{t} x_i$$

$(x_t)$ — стационарный процесс с $\mathbb{E}(x_t) = 0$.

$(y_t)$ — стационарный в разностях (difference stationary).

Рецепт: оценим $ARMA$ для $(\Delta y_t)$.

Оба $(y_t)$ нестационарны!

# $KPSS$ тест: итоги

- Применим для принятия решения о переходе к $\Delta y_t$.
- Есть два варианта теста с разными предпосылками.

# Сезонная ARIMA

# Сезонная ARIMA: план

- ARMA должна быть экономной!
- Сезонные полиномы.
- Нужно ли переходить к сезонным разностям?

# Сезонность и $ARIMA$

С помощью $ARMA$ и $ARIMA$ моделей можно моделировать сезонность!

Только дорого!

$$MA(12): y_t = c + u_t + a_1 u_{t-1} + a_2 u_{t-2} + \ldots + a_{12} u_{t-12}.$$

$$ARIMA(12, 1, 0): \Delta y_t = c + u_t + b_1 \Delta y_{t-1} + \ldots + b_{12} \Delta y_{t-12}.$$

# ARMA должна быть экономной!

Сосредоточимся на коэффициентах сильнее отличных от нуля!

**Определение**

Если стационарную $ARMA$ модель для $y_t$ можно записать с меньшим числом параметров в виде

$$P_{non}(L)P_{seas}(L^{12})y_t = c + Q_{non}(L)Q_{seas}(L^{12})u_t,$$

где степени у лаговых полиномов равны $\deg P_{non} = p$, $\deg P_{seas} = P$, $\deg Q_{non} = q$, $\deg Q_{seas} = Q$, то она также называется $SARMA(p,q)(P,Q)[12]$.

# Примеры

- $SARMA(1, 0)(0, 2)[12]$

$$(1 - b_1 L)y_t = c + (1 + d_1 L^{12} + d_2 L^{24})u_t;$$

- $SARMA(0, 2)(1, 0)[12]$

$$(1 - f_1 L^{12})y_t = c + (1 + a_1 L + a_2 L^2)u_t;$$

- $SARMA(1, 2)(2, 1)[12]$

$$(1 - f_1 L^{12} - f_2 L^{24})(1 - b_1 L^1)y_t = c + (1 + a_1 L + a_2 L^2)(1 + d_1 L^{12})u_t$$

# SARIMA

По аналогии с разностью $\Delta y_t = y_t - y_{t-1}$ можно рассмотреть сезонную разность $\Delta_{12} y_t = y_t - y_{t-12}$.

**Определение**

Если ряд $z_t = \Delta^d \Delta_{12}^D y_t$ описывается стационарной моделью $SARMA(p,q)(P,Q)[12]$, то говорят, что $y_t$ описывается моделью $SARIMA(p,d,q)(P,D,Q)[12]$.

$d$ — количество взятий обычной разности $\Delta = 1 - L$;

$D$ — количество взятий сезонной разности $\Delta_{12} = 1 - L^{12}$;

$y_t \sim SARIMA(0,0,2)(1,1,2)[12]$ означает, что

$\Delta_{12} y_t \sim SARMA(0,2)(1,2)[12]$

# Как выбрать?

$SARIMA(p, 0, q)(P, 0, Q)$ или $SARIMA(p, 0, q)(P, 1, Q)[12]$?

- Посмотреть на график!
  Слишком выраженная сезонность — повод перейти к $\Delta_{12} y_t$.
- Оценить все эти модели и выбрать наилучшую по кросс-валидации.
  Затратно по времени!
- Применять $AIC$ нельзя!
  Условная и безусловная функции правдоподобия содержат разное число слагаемых.
- Есть тесты на единичный корень!
  И эмпирические правила…

# STL разложение и сила сезонности

Шаг 1. Находим $STL$ разложение ряда $(y_t)$.

$$y_t = trend_t + seas_t + remainder_t$$

Шаг 2. Рассчитываем силу сезонности.

$$F_{seas} = \max\left\{1 - \frac{\mathrm{sVar}(remainder)}{\mathrm{sVar}(seas + remainder)}, 0\right\}.$$

Шаг 3. Если сила сезонности выше порога, то переходим к $\Delta_{12} y_t = y_t - y_{t-12}$.

# Сезонная ARIMA: итоги

- Сезонная ARIMA экономит параметры.
- Сила сезонности из STL разложения используется для решения о необходимости сезонной разности $\Delta_{12} y_t$.

# Алгоритм Хандакара-Хиндмана

# Алгоритм Хандакара-Хиндмана: план

- Три шага алгоритма.
- Нюансы и рекомендации.

# Как всё это собрать в кучу?

Шаг 1 (для сезонных рядов). Сколько раз надо брать $\Delta_{12}$?

Шаг 2. Сколько раз надо брать $\Delta$?

Шаг 3. Какую стационарную $SARMA$ модель оценивать после взятия разностей?

## Шаг 1. Сколько раз надо брать $\Delta_{12}$?

Ни разу, раз или два раза.

- Находим $STL$ разложение ряда.
- Если сила сезонности меньше пороговой, то работаем с исходным рядом $y_t$.
- Если сила больше пороговой, то переходим к сезонной разности и после нового $STL$ разложения сравниваем силу сезонности с пороговой ещё раз.
- Если сила сезонности снова больше пороговой, то работаем с $\Delta_{12}^2 y_t$, иначе работаем с $\Delta_{12} y_t$.

Есть альтернатива в виде теста Канова-Хансена (Canova-Hansen).

# Шаг 2. Сколько раз надо брать $\Delta$?

Ни разу, раз или два раза.

- Применяем $KPSS$ тест с константой к исходному ряду.
- Если $H_0$ не отвергается, то работаем с рядом $y_t$.
- Если $H_0$ отвергается, то проводим $KPSS$ тест для разности $\Delta y_t$.
- Если у повторного $KPSS$ теста $H_0$ отвергается, то работаем с $\Delta^2 y_t$ иначе работаем с $\Delta y_t$.

Есть альтернатива в виде $ADF$ теста.

# Шаг 3. Выбор $SARMA$ модели для преобразованного ряда

- Оцениваем большое количество экономных $SARMA$ моделей.
  $\Delta^d \Delta_{12}^D y_t \sim SARMA(p, q)(P, Q)[12]$, $p + q \leq 5$, $P + Q \leq 5$
- Выбираем наилучшую модель по штрафному критерию Акаике:
  $AIC = 2K - 2 \ln L$, где $K$ — общее число параметров, $\ln L$ — логарифм правдоподобия.

Есть альтернатива в виде перебора с помощью кросс-валидации.

# Методология Бокса-Дженкинсона

- Идентификация подходящей модели.
  Графический анализ, тесты. Выбор количества сезонных и обычных разностей.
- Оценивание подходящей модели.
  Оценивание параметров $SARMA$ модели для преобразованного ряда.
- Статистическая проверка модели.
  Визуализация остатков. Тесты на остатки модели.

Алгоритм Хандакара-Хиндмана — практическая реализация методологии.

# Нюансы алгоритма

- Очень много опций…

Возможны отличия реализаций в софте.

- Обратите внимание на включение константы

$P(L)y_t = c + Q(L)u_t$ или $P(L)(y_t - \mu) = Q(L)u_t$

- Требует много времени.

Не стоит использовать кросс-валидацию.

- Суммирует опыт десятилетий.

Не забудьте им воспользоваться!

# Алгоритм Хандакара-Хиндмана: итоги

- Шаг 1. Решение о переходе к сезонным разностям.
- Шаг 2. Решение о переходе к разностям.
- Шаг 3. Оценка множества $SARMA$ моделей с выбором по $AIC$.
- Обязательно попробуйте алгоритм!