Домашнее задание:
Временные ряды в R

21 июня 2022 г.

Общая информация

В этом задании вам предстоит попрактиковаться в анализе временных рядов, используя функционал R. Для удобства вспомним последовательность установки инструментов с первого занятия:

1. На сайте [RStudio] выбрать RStudio Desktop –> Download.
2. Установить R по [ссылке], которая отобразится в описании шага Install R. Рекомендуется установить последнюю версию: R 4.2.0 (2022-04-22) – "Vigorous Calisthenics".
3. Установить RStudio для вашей системы.

Все задания нужно выполнять внутри скрипта `.R`. Весь код должен быть написан на R, а выводы оформлены в тексте скрипта в виде комментариев. Для удобства проверяющих постарайтесь разделить файл на секции при помощи комментариев `# ==== Section name ====`

Не забудьте объявить рабочей директорией ту же папку, в которой находится скрипт.

Для некоторых заданий ниже приведены подсказки, однако использовать код из них не обязательно. Все тесты нужно проводить на вашем любимом уровне значимости.

Во всех заданиях предполагается, что в ряде отсутствует сезонность (даже если на самом деле это не так).

Дедлайн

28 июня 2022, 23:59 МСК

Оценивание

По умолчанию за полностью выполненную работу ставится 10 баллов, из которых вычитаются штрафы за мелкие (-0.1 – 0.2) и грубые (-0.5) ошибки. Субъективное или корректно объяснённое действие не считается ошибкой (например, визуальный выбор лагов или определение типа преобразования Бокса-Кокса). Но ошибкой является неправильное применение или интерпретация функций или статистических тестов. За каждое невыполненное задание вычитается 0.7 балла.

Чем можно пользоваться

Можно использовать любой код с практических занятий, а также любой код из открытых источников с указанием ссылки на источник.

Задания

1. Файл `fdi.xlsx` содержит данные по чистым иностранным инвестициям, скачанные из базы [Всемирного Банка]. Прочитайте данные и сохраните их в какую-нибудь переменную.

   Hint

   Для чтения файлов `.xlsx` можно использовать пакет `readxl`.

2. Переведите данные из «горизонтального» в «вертикальный» формат, который затем можно будет использовать для построения временного ряда.

   Hint

   Могут пригодиться функции `gather()` и `arrange()` из пакета `dplyr`.

3. Выведите ряд и убедитесь, что несколько первых значений в нём пропущены. Так как функция `ts()` умеет работать только с регулярными рядами, от пропущенных значений необходимо избавиться. Удалите пропущенные значения.

4. При помощи функции `ts()` переведите данные в формат временного ряда.

5. Если на разных участках ряда наблюдается разная дисперсия наблюдений (например, в начале ряд колеблется слабо, а в конце сильно), то на этапе предварительной обработки для сглаживания дисперсии часто применяется преобразование Бокса-Кокса

   $$y_t := \begin{cases} \log y_t, & \text{если } \lambda = 0, \\ \dfrac{y_t^\lambda - 1}{\lambda}, & \text{если } \lambda \neq 0, \end{cases}$$

   где λ – гиперпараметр, подбираемый эвристически. Если вы считаете, что дисперсия ряда нестабильна, выберите какую-нибудь λ и примените преобразование Бокса-Кокса для стабилизации дисперсии.

   Hint

   Может пригодиться функция `BoxCox()` из пакета `forecast`.

6. Разделите выборку на обучающую и тестовую. На тестовую выборку оставьте последние пять лет.

7. Постройте графики ряда, ACF и PACF на обучающей выборке. Примерно оцените, какие значения следует брать для параметров p и q. Попробуйте визуально определить наличие детерминированного тренда в ряде.

8. Для статистического тестирования наличия стохастического тренда можно использовать тест **ADF (Augmented Dickey-Fuller)**, который проверяет гипотезу о том, что ряд содержит стохастический тренд:

   $$\begin{cases} H_0 : \text{Ряд содержит стохастический тренд}, \\ H_1 : \text{Ряд не содержит стохастический тренд} \end{cases}$$

   Для проведения теста ADF используйте функцию `ur.df(series, lags = ..., type = ...)` из пакета `urca`. Параметр `lags` этой функции – это количество лагов переменной, которые будут использоваться для проверки гипотезы (имеет смысл установить этот параметр чуть больше, чем p, выбранный выше). Параметр `type` отвечает за то, какую спецификацию модели требуется проверить:

   - `type = none` подразумевает, что в модели нет ни константы, ни детерминированного тренда, то есть модель имеет вид $y_t = y_{t-1} + \dots$
   - `type = drift` подразумевает, что в модели есть константа, то есть модель имеет вид $y_t = C + y_{t-1} + \dots$

- `type = trend` подразумевает, что в модели есть детерминированный тренд, то есть модель имеет вид $y_t = C + bt + y_{t-1} + \dots$

Распределение статистики ADF зависит от наличия в модели константы и детерминированного тренда. Выберите нужную спецификацию модели и проведите ADF-тест на обучающей выборке. Для удобного вывода результатов используйте команду `summary(ur.df(series, lags = ..., type = ...))`. Вне зависимости от спецификации, нам всегда нужно первое (левое) значение в строчке «Value of test-statistic is:» и первая строчка (tau3) критических значений.

Определите, есть ли в данных стохастических тренд.

Если тренд обнаружен, используйте команду `diff(...)` для взятия первых разностей. Проведите ADF-тест для ряда из разниц (обратите внимание, что отсутствие константы в ряде разностей подразумевает наличие константы в исходном ряде, а наличие константы в ряде разностей подразумевает наличие детерминированного тренда в исходном ряде, поэтому будьте внимательны при выборе `type`). Если детрендирование не помогло, используйте команду `diff(..., 2)` для взятия вторых разностей. Повторите тест.

Определите порядок разностей d, который позволяет избавиться от тренда (он вполне может оказаться равен 0).

9. Используя графики ACF и PACF для детрендированных данных обучающей выборки, визуально определите параметры p и q для модели ARIMA(p, d, q).

10. В цикле переберите все возможные комбинации параметров из списков $[p-1,\ p,\ p+1]$ и $[q-1,\ q,\ q+1]$, где p и q – значения из предыдущего пункта, и обучите модель с каждой комбинацией (всего 9 моделей). Для каждой модели рассчитайте AIC на обучающей выборке. Выберите модель с лучшим значением AIC.

Hint

Значения AIC удобно хранить в структуре `matrix`. Значение AIC на обучающей выборке для модели `model` можно получить при помощи команды `model$aic`.

11. Оцените модель ARIMA с выбранными параметрами на обучающей выборке. Проведите тесты на коррелированность и нормальность остатков. Определите, можно ли доверять доверительным интервалам прогнозов. Обозначим эту модель как M1.

12. Оцените автоматическую модель `auto.arima()` (не забудьте обозначить отсутствие сезонности). Проведите тесты на коррелированность и нормальность остатков. Определите, можно ли доверять доверительным интервалам прогнозов этой модели. Обозначим эту модель как M2.

13. Рассчитайте BIC для M1 и M2 на обучающей выборке. Определите лучшую модель с точки зрения качества подгонки.

14. Постройте прогнозы на $h = 1, 2, \dots, 5$ периодов вперёд. Оцените качество прогнозирования M1 и M2 на тестовой выборке при помощи какой-нибудь метрики. Определите лучшую модель с точки зрения качества предсказаний.