Домашнее задание по эконометрике для студентов исследовательского потока

Дедлайн: 20 июня 2020 года, 20:20

1 июня 2020 г.

Задание следует выполнять в `R`. Итоговый файл должен представлять из себя скрипт `.R` с кодом и вашими комментариями к нему. Куда отправить работу будет сообщено позднее.

Задание состоит из двух частей, каждая из которых оценивается в 5 баллов. Разбивка баллов внутри частей указана рядом с номером заданий в скобках **жирным** шрифтом. За некоторые задания можно получить бонусы, которыми можно восполнить потерянные баллы в других заданиях. При этом один бонус = 0.5 балла. Максимальная оценка за работу – 10, то есть нельзя получить 11, если всё сделано верно и получено два бонуса. Но мы подумаем, куда и как перенести избыток бонусов.

Выполненные задания должны идти по порядку, и их следует разделять так, чтобы проверяющему было понятно, где заканчивается одно задание и начинается другое. Вы можете использовать любые пакеты и библиотеки, и их подключение должно быть в начале работы. Вы также можете использовать материалы (в том числе и код) из любых открытых источников, однако в месте использования стоит привести ссылку на источник, чтобы избежать подозрения в плагиате. За обнаруженный и доказанный плагиат за всю работу ставится 0. Такая же оценка ставится и при обнаружении списывания, причём всем участникам, даже если можно однозначно определить кто у кого списал.

Перед выполнением заданий не забудьте зафиксировать `seed` для воспроизводимости результатов. Все графики должны быть визуально понятными: не забудьте подписать оси и заголовки. Также к каждому графику должен быть приведён комментарий: графики без пояснений того, какой вывод можно сделать на их основе, не оцениваются. По возможности, старайтесь писать как можно больше выводов и комментариев к тому, что вы делаете и почему.

1 Вперёд и с песней!

В этой части мы будем применять полученные в течение курса эконометрические навыки на практике. Формат этой части достаточно свободный, оцениваются любые разумные действия и выводы.

Загрузите набор данных Country Statistics – UNData, содержащий различные географические, экономические и социальные показатели по разным странам мира. Обратите внимание, что по ссылке для скачивания доступно два файла, и нам требуется файл `country_profile_variables.csv` Для загрузки понадобится регистрация на kaggle.

1. **(0.5)** Сформулируйте исследовательский вопрос. В соответствии с ним выберите непрерывную зависимую переменную. Заметим, что в серьёзных научных работах выбор следует объяснять ссылкой на литературу.

   Например: «Я хочу изучить, как связаны темпы экономического роста и площадь территории страны. Для этого в качестве зависимой переменной я беру темп прироста ВВП».

2. **(0.5)** Выберите и/или создайте объясняющие переменные. Итоговая матрица регрессоров должна включать:

   (a) Не менее одной непрерывной переменной.

   (b) Не менее одной бинарной переменной.

(c) Не менее одной нелинейной переменной (квадрат, логарифм и т.д.)

И по желанию для самых смелых **(+1 бонус, если дальше правильно интерпретируется)**:

(d) Не менее одной переменной взаимодействия.

Поясните логику выбора переменных.

В дальнейших пунктах потребуется дать смысловую интерпретацию оценок коэффициентов при выбранных регрессорах. Заметим, что если в исследовательском вопросе фигурирует одна независимая переменная (как в примере выше), то прочие переменные можно интерпретировать как *контрольные*, то есть позволяющие учесть влияние сторонних факторов, что может быть важно по каким-либо причинам.

Например: «В качестве непрерывных регрессоров я беру площадь территории страны и её квадрат, потому что (здесь идёт объяснение вашего выбора, отсутствующее из-за странности используемого примера)».

3. **(1 + 1 бонус за особо красивые графики)** Проведите визуальный анализ данных:

   (a) На наличие выбросов.

   (b) На наличие пропущенных значений.

   При необходимости обработайте (например, удалите) выбросы. При наличии пропущенных значений удалите их или замените на какое-то значение (например, среднее, медиану и т.д. по регрессору). В любом случае, поясните ваши действия.

   Результатом данного пункта являются воспроизводимые графики и пояснения к ним. Графики должны быть визуально понятными: не забудьте подписать оси и заголовки.

4. **(1)** Задайте спецификацию модели. Проведите тестирование на наличие:

   (a) Мультиколлинеарности.

   (b) Гетероскедастичности.

   (c) Эндогенности.

   Для тестирования наличия каждой проблемы используйте не менее двух статистических тестов. Для проведения тестов используйте готовую реализацию: нужные пакеты и функции в `R` достаточно легко ищутся в поисковике (проверено на собственном опыте!) Для того чтобы показать наличие или отсутствие мультиколлинеарности, можно использовать теоретические знания линейной алгебры (но это не обязательно!)

   Поясните, как найденные проблемы исказят оценки МНК. В зависимости от найденных проблем, выберите метод оценки модели и поясните ваш выбор.

   Например: «Я провёл (названия тестов) и выявил наличие в данных проблемы эндогенности. Таким образом, если я буду оценивать модель при помощи МНК, оценки коэффициентов будут несостоятельными. В данном случае для устранения проблемы разумно использовать 2МНК».

5. **(1)** Оцените модель:

   (a) При помощи МНК.

   (b) Выбранным вами методом.

   Прокомментируйте результаты: чем отличаются оценки вашего метода от оценок МНК? Насколько сильно проявляется влияние найденных проблем? Выберите уровень значимости, который вам больше нравится, и прокомментируйте значимость коэффициентов на этом уровне. Прокомментируйте адекватность регрессии в целом. Проинтерпретируйте полученные оценки коэффициентов банальным образом *(при увеличении X_1 на единицу, Y увеличивается на* 0.5*)* и по смыслу *(X_1 положительно влияет на Y, что можно объяснить слеующим образом: (объяснение)*. Если это возможно, дайте ответ на исследовательский вопрос.

6. **(1)** Теперь попробуем отвлечься от эконометрических задач и попробовать себя в роли machine learner'а. Допустим, что мы заинтересованы не в получении качественного ответа на некоторый исследовательский вопрос, а в достижении наибольшего качества предсказания зависимой переменной. В такой постановке, вообще говоря, нас не интересует, какие проблемы представлены в данных: нам важно построить некоторую базовую модель, а затем предложить другую спецификацию модели, лучшую по качеству предсказания.

   Задача ставится следующим образом. Предположим, что у нас есть зависимая переменная из пункта 1 и регрессоры из пункта 2, и мы провели визуальный анализ и модификацию данных из пункта 3. Базовой будем считать модель:

   $$Y = X\beta + u,$$

   оцениваемую при помощи МНК.

   Поделите данные на обучающую и тестовую выборку в соотношении 8:2. Оцените базовую модель на обучающей выборке и получите прогноз на тестовой выборке. Рассчитайте среднеквадратичную ошибку прогноза (MSE).

   Выберите и оцените три других спецификации модели. Разрешаются любые модификации: добавление или отброс переменных, взятие функций от регрессоров и проч., использование различных методов оценки. Рассчитайте среднеквадратичную ошибку прогнозов этих моделей. Задание считается выполненным, когда найдена такая спецификация модели, среднеквадратичная ошибка прогноза которой меньше, чем у МНК. В зависимости от того, насколько удалось снизить ошибку прогноза, могут быть выставлены бонусные баллы **(до +2 бонусов)**.

2 Табалуга и река времени

В этой части мы будем работать с временными рядами. Так как анализ реальных временных рядов требует значительной подготовки, мы будем использовать искусственно созданные ряды, которые сами и сгенерируем.

Сейчас в `R` существует два распространённых стиля работы с временными рядами:

- Хорошо устоявшийся пакет `forecast` для работы с небольшим количеством рядов.
- Новый пакет `fable` с кучей модных плюшек для работы с сотнями и тысячами рядов.

Вы можете работать в рамках любого подхода.

1. **(1)** Сгенерируйте временные ряды, задающиеся следующими уравнениями:

   $$y_t = 0.8y_{t-1} + \varepsilon_t$$
   $$y_t = 0.1y_{t-1} + 0.2y_{t-2} + 0.3y_{t-3} + \varepsilon_t$$
   $$y_t = \varepsilon_t + 1.2\varepsilon_{t-1} + 2\varepsilon_{t-2},$$

   каждый из которых состоит из 120 наблюдений.

   (a) Выпишите спецификацию моделей, задаваемых этими уравнениями. Например, AR(5).

   (b) Постройте графики полученных временных рядов.

   (c) Используя графики и/или полученные знания, прокомментируйте, имеют ли данные уравнения стационарные решения.

2. **(1)** На основе предыдущего пункта, сгенерируйте временные ряды, уравнения которых специфицируется как ARIMA(0, 1, 2), ARIMA(0, 0, 0), ARIMA(3, 0, 0). Постройте графики и прокомментируйте, имеют ли соответствующие уравнения стационарные решения.

3. **(0.5)** Вспомните уравнение случайного блуждания. Сгенерируйте соответствующий временной ряд и постройте его график. Имеет ли это уравнение стационарные решения?

4. **(0.5)** Из созданных выше рядов выберите ряд, задаваемый моделью AR(1), уравнение которого имеет стационарные решения. Постройте автокорреляционную и частную автокорреляционную функции для этого ряда. Сравните их с ACF и PACF случайного блуждания. Прокомментируйте результаты.

5. **(2)**

   (a) Сгенерируйте ряд из 120 наблюдений, задаваемый моделью ARIMA(2, 0, 3).

   (b) Разделите ряд на обучающую выборку из 100 наблюдений и тестовую выборку из 20 наблюдений.

   (c) Оцените модель ARIMA(2, 0, 3) на обучающей выборке.

   (d) Постройте прогноз на 20 периодов вперёд для этой модели, используя 95% доверительные интервалы.

   (e) Постройте полученные прогнозные значения и тестовую выборку на одном графике. Визуально оцените качество прогноза.