Домашнее задание 3 Корпусные методы, 1.12.2018

Дата сдачи: 9.12. до 21.00

1. *Линейная регрессия [5 баллов]*

Существуют две противоположные точки зрения на изучение иностранных языков детьми. Первая точка зрения: изучение иностранных языков детьми препятствует полноценному освоению родного языка. Такой точки зрения в своей языковой политике традиционно придерживалась Япония. Вторая точка зрения: изучение иностранных языков детьми не препятствует полноценному освоению родного языка, возможно, даже способствует ему. Многие страны, например, Голландия, в своей языковой политике отражают вторую точку зрения.

Задание:

В файле hw3lr.csv находятся данные по 200 ученикам 10 лет школ одного и того же города. Переменная NofL (number of languages) характеризует детей по тому, сколькими языками они владеют и/или изучают в школе или в объёме не менее школьного. Например, если NofL=2, это может значить, что школьник билингв и владеет двумя языками с рождения, а может значить, что он владеет одним языком с рождения и изучает второй. Переменная Score отражает оценку за экзамен по родному языку. Переменная принимает значения от 0 до 100. Родной язык у всех школьников один и тот же.

- 1 балл: посчитайте в R среднее арифметическое и медиану для переменной NofL и для переменной Score. Приведите 4 числа и формулу, по которой Вы их находили в R.
- 1 балл: изобразите данные и линию регрессии между переменными NofL и Score на графике. Приведите формулы, которые Вы использовали при построении графика и сам график.
- 1 балл: приведите коэффициент корреляции между переменными NofL и Score и формулу, по которой Вы его посчитали.
- 1 балл: проинтерпретируйте результат. Существует ли линейная зависимость между двумя переменными? Если да, то какая: прямая (положительная) или обратная (отрицательная)?
- 1 балл: кратко (5-10 предложений) выскажите своё мнение по поводу языковой политики изучения детьми иностранных языков. Если Ваше мнение подтверждается проанализированными данными, используйте их в Вашей аргументации. Если Ваше мнение расходится с проанализированными в задании данными, объясните, почему их можно проигнорировать. Подсказка: наличие линейной зависимости между двумя переменными не гарантирует наличие причинно-следственной связи между ними.

2. *Метод главных компонент [5 баллов]*

Проанализируйте данные из файла hw3pca.csv методом главных компонент. Данные взяты из книги *Политический атлас современности: Опыт многомерного статистического*

анализа политических систем современных государств. — М.: Изд-во «МГИМО-Университет», 2007. — 272 с. https://www.hse.ru/data/2009/12/15/1230161701/politatlas.pdf

StateshipIndex соответствует индексу государственности стран (стр. 161--163). ThreatIndex соответствует индексу внутренних и внешних угроз (стр. 164--166). InfluenceIndex соответствует индексу потенциала международного влияния (стр. 168--169). DemocracyIndex соответствует индексу институциональных основ демократии (стр. 174--175). Значения округлены.

Задание:

- 1 балл: код, который вводит переменную pcadata, в которой лежат все данные, кроме столбца Country.
- 1 балл: приведите код, позволяющий оценить важность каждой из компонент. Сколько компонент можно оставить для описания вариативности данных? Приведите два критерия, которые помогают ответить на этот вопрос.
- 1 балл: приведите столбчатую диаграмму вклада каждой из компонент в описание вариативности данных.
- 1 балл: приведите код, позволяющий оценить вклад каждой переменной в первую и вторую компоненты. Какие из переменных вносят наибольший вклад в первую компоненту?
- 1 балл: приведите график вклада переменных в первую и вторую компоненты.