Отчет по лабораторным работам.

Выполнил:

Зиновьев Евгений

Гр.09-115(3)

**ЗАДАНИЕ 1.**Заполнение пропущенных значений в таблицах данных.

Задание: Заполнение пропущенных значений в таблицах данных.

Цель: Заменить имеющиеся пропуски на некоторые разумные значения.

Для заполнения пропущенных значений необходимо выполнить следующие шаги:

1. Идентификация недостающих данных.
2. Исследование закономерностей появления отсутствующих значений.
3. Формирование наборов данных, не содержащих пропуски (в результате удаления или замены соответствующих фрагментов).

Атрибутивная информация используемого дата сета:

[http://archive.ics.uci.edu/ml/datasets/StoneFlakes#](http://archive.ics.uci.edu/ml/datasets/StoneFlakes)

Attribute Information:

1. LBI: Length-breadth index of the striking platform
2. RTI: Relative-thickness index of the striking platform
3. WDI: Width-depth index of the striking platform
4. FLA: Flaking angle (the angle between the striking platform and the splitting surface)
5. PSF: platform primery (yes/no, relative frequency)
6. FSF: Platform facetted (yes/no, relative frequency)
7. ZDF1: Dorsal surface totally worked (yes/no, relative frequency)
8. PROZD: Proportion of worked dorsal surface (continuous)

Атрибутивная информация:

1. LBI: Индекс длины и ширины ударной платформы
2. RTI: Индекс относительной толщины ударной платформы
3. WDI: Индекс ширины-глубины ударной платформы
4. FLA: угол отслаивания (угол между ударной платформой и поверхностью расщепления)
5. PSF: пример платформы (да/нет, относительная частота)
6. FSF: Грань платформы (да/нет, относительная частота)
7. ZDF1: Дорсальная поверхность полностью обработана (да/нет, относительная частота)
8. PROZD: Доля обработанной дорсальной поверхности (непрерывная)

Информация о Наборе данных:

Справочная информация: Набор данных относится к самой ранней истории

человечества. Доисторические люди создавали желаемую форму каменного орудия,

ударяя по необработанному камню, таким образом отщепляя чешуйки,

отходы процесса изготовления. Археологи находят не так много инструментов,

но они находят чешуйки. Набор данных посвящен этим хлопьям. Его ряды обозначают

не отдельные хлопья, а целые их запасы. Заданными

характеристиками являются относительные частоты двоичных чисел и средние значения числовых

характеристик, взятые по всем предметам, найденным в инвентаре. Вопрос

с набором данных связано следующее: Отражают ли данные технологический

прогресс за несколько сотен тысяч лет?

Все данные параметры представлены в виде чисел с плавающей точкой. «Таргетная» переменная FLA является предиктором, показывающим насколько человек усовершенствовал свое оружие за промежутки своего существования.

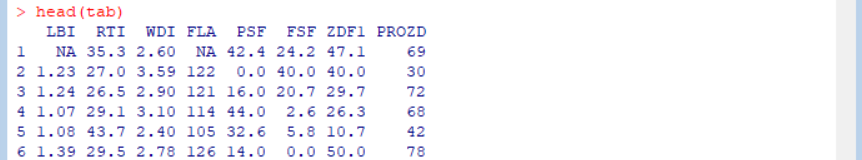
Подключим библиотеку mice, в которой как раз и находятся методы по заполнению пустых значений, а также пакет визуализации VIM.

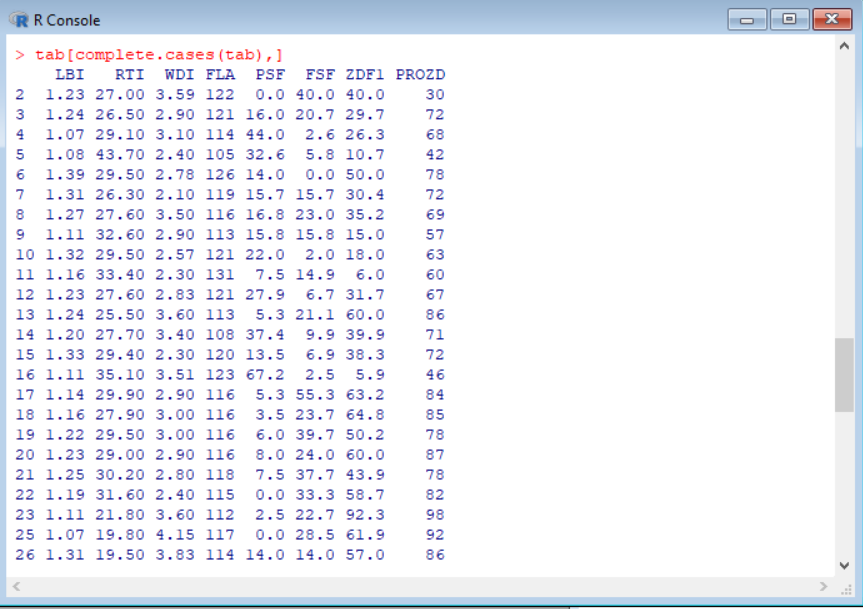
*library(mice)*

*library(VIM)*

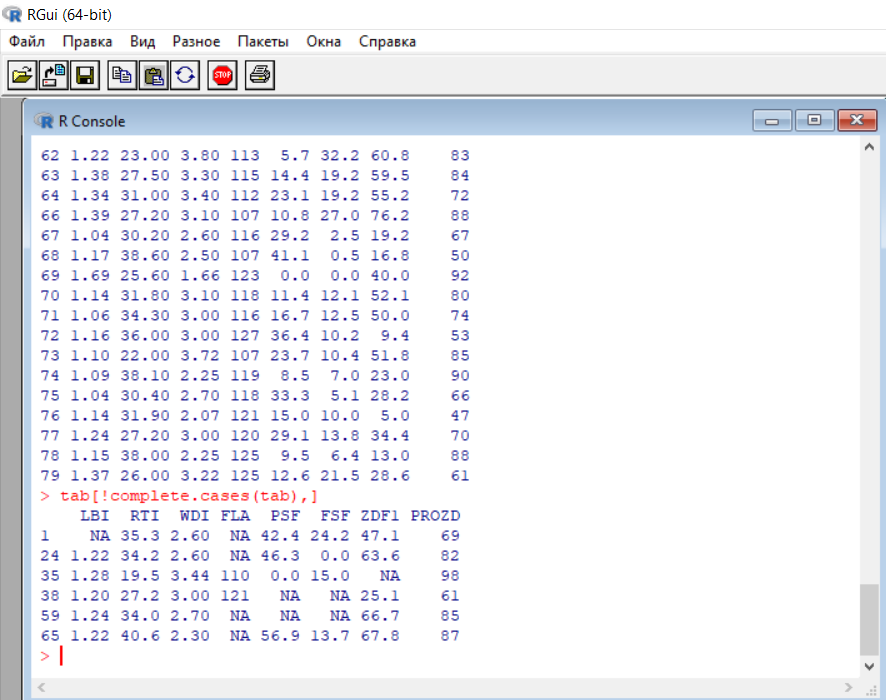
Прочитаем файл и просмотрим данные:

*tab<-read.table(file="C:\\Users\\pc\\Downloads\\StoneFlakes.dat", header = TRUE, sep=",")*

Полностью заполненные строки:



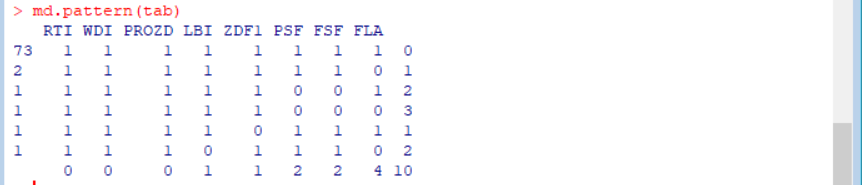
Идентифицируем пустые значения в датасете:

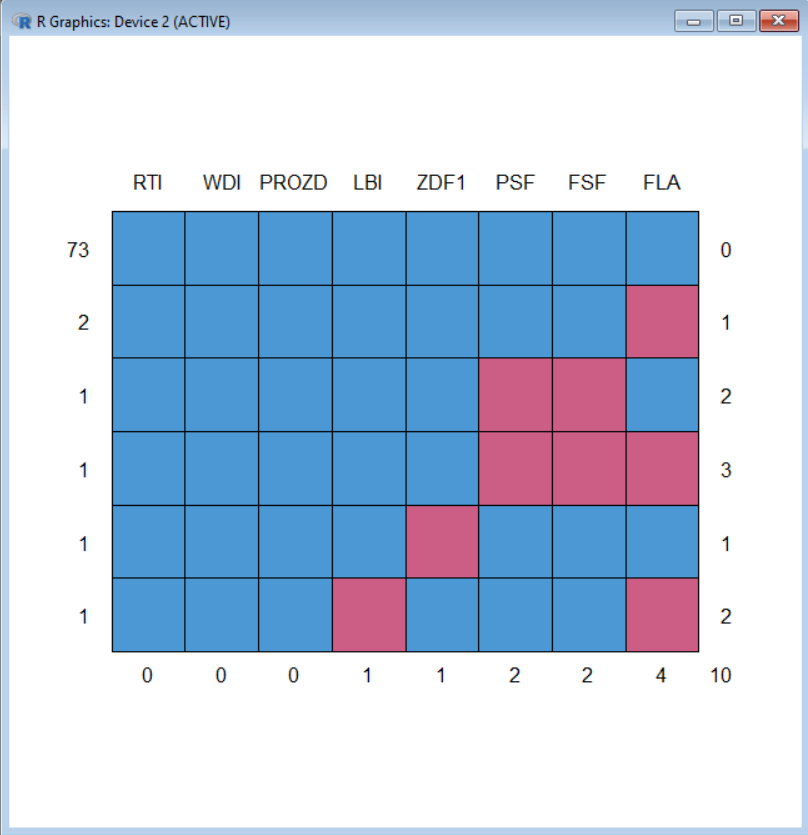


Посчитаем общее количество пропущенных значений:

**

Выведем матрицу, показывающую при каких комбинациях параметров пропущены значения, и соответствующую ей бинарный рисунок:

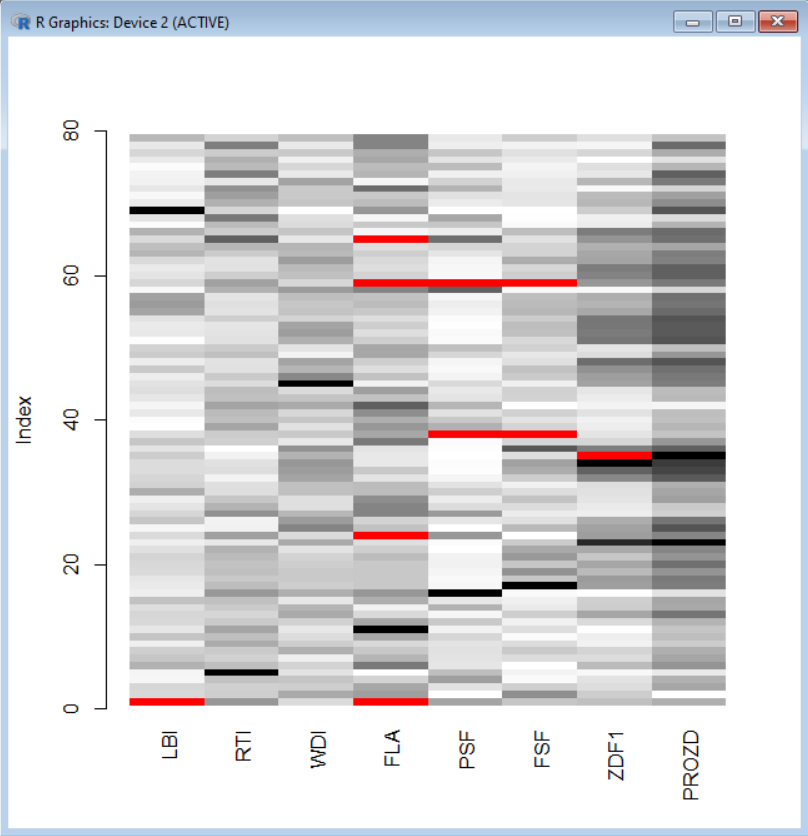




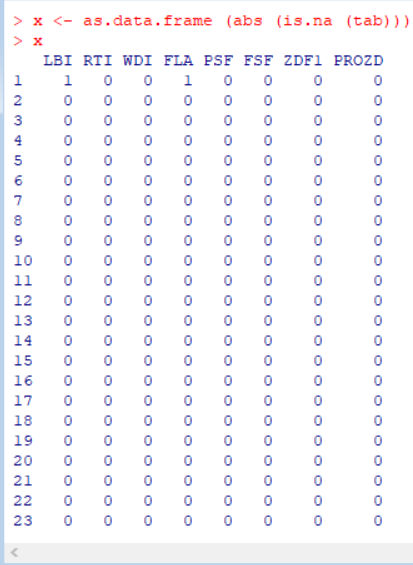
Значения 0 в теле приведенной таблицы соответствуют недостающим значениям, причем в первой строке пропусков нет, а последующие строки упорядочены по числу их появления. Первый столбец указывает число случаев в каждом образце данных, а последний столбец – число переменных с отсутствующими значениями в каждой строке.

На диаграмме, полученной при помощи функции matrixplot(), числовые данные масштабируются к интервалу [0, 1] и представлены уровнями шкалы яркости, причем более темные цветами показаны большие значения. По умолчанию, недостающие значения представлены красным цветом. График является интерактивным: нажатие на столбец отсортирует матрицу по этой переменной:

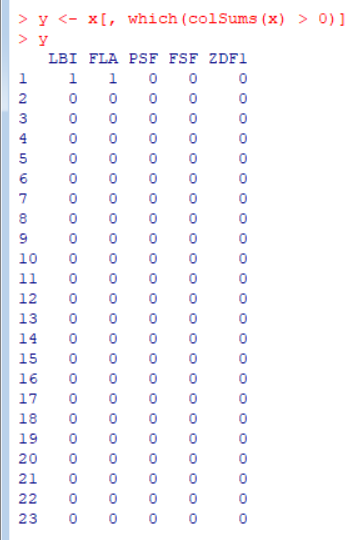
*> matrixplot(tab)*



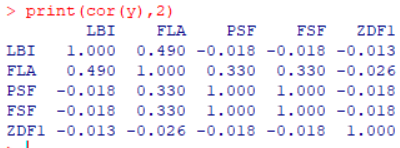
Наложим на исходный датасет булеву маску наличия пустых значений и переведем ложные значения в нули, а истинные в единицы:



Оставим только те столбцы, в которых есть пустые значения:

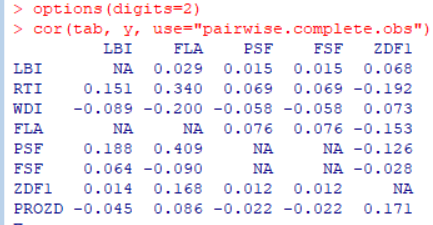


Построим корреляционную матрицу пропусков, которая показывает, как взаимосвязаны частоты пропусков в одних столбцах от пропусков в других:



Как видно из матрицы: без вида платформы нельзя понять ее грань (FSF, PSF).

Далее построим корреляцию параметров между собой:

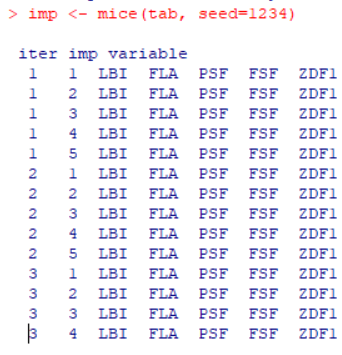


На использовании корреляционных отношений основан популярный метод заполнения пропусков с использованием регрессионных моделей. Другой подход, называемый "множественное присвоение" (multiple imputation), основан на многократно повторяющемся моделировании данных с использованием методов Монте-Карло.

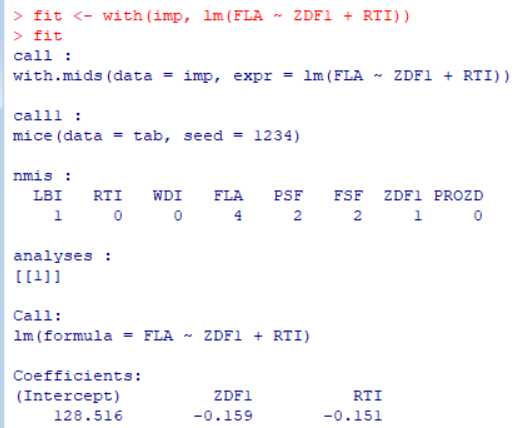
Функция mice() из одноименного пакета реализует метод Гиббса (Gibbs sampler), который представляет собой способ формирования выборок (x1; …; xn) из заданных распределений p(x) m-мерных переменных путем многократного извлечения имитируемых значений. По умолчанию, каждая переменная, содержащая недостающие величины, моделируется в зависимости от всех других переменных в наборе данных. Полученные уравнения предсказания используются, чтобы приписать наиболее вероятные значения для пропущенных данных. Процесс повторяется, пока не будет достигнут определенный критерий сходимости. Для каждой переменной пользователь может выбрать форму модели предсказания, называемую элементарным методом присвоения.

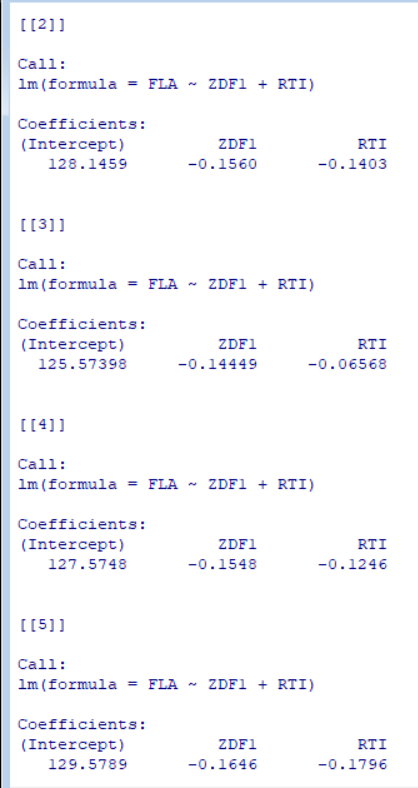
Процесс заполнения пропусков реализуется в несколько этапов:

1. функция mice() обрабатывает tab – матрицу или структуру данных, содержащую недостающие значения, и возвращает объект imp, содержащий несколько полных наборов данных (например, m = 5), отличающихся между собой значениями пропусков, заполненных в ходе процедуры Гиббса, но с включением некоторого случайного компонента:

**

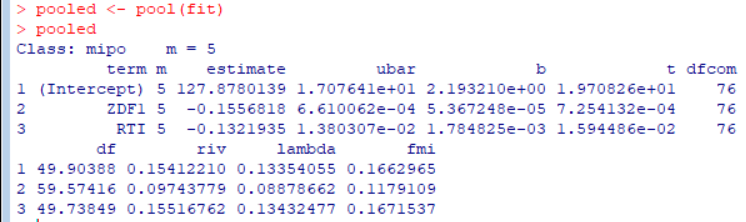
2. функция with() используется, чтобы применить статистическую модель analysis к каждому набору данных из imp; это может быть линейная lm() или обобщенная линейная glm() модели, аддитивная модель gam() или любая иная, задаваемая пользователем. Применим линейную модель зависимости искомой переменной от оставшихся параметров. Выберем в качестве независимых параметров ZDF1 + RTI (0.167 и 0.339 соответственно), так как FLA довольно сильно коррелирует между ними, но при этом они сами не сильно связаны с другими параметрами (PSF, например, коррелирует со всеми переменными, вследствие чего может вносить мелкий вклад в модель линейной функции, поэтому ее не включаем):

**



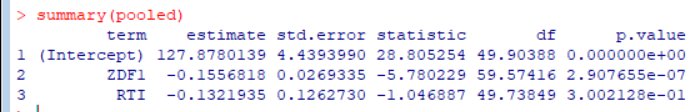
Таким образом, в данном отчете показаны количество пропущенных значений напротив каждого столбца, а также коэффициенты корреляций для пяти построенных наборов данных. Как видим, зависимость показателя FLA от 2 выбранных параметров не сильно меняется от одного набора к другому и приблизительно находится в схожей отрицательной зависимости в районе -0.14, а сами коэффициенты корреляции снизились относительно матрицы корреляций с пропущенными значениями.

3. функция pool() комбинирует результаты и объединяет их в заключительную модель, в которой стандартные ошибки и p-значения сбалансированы относительно неопределенности, порожденной пропусками:



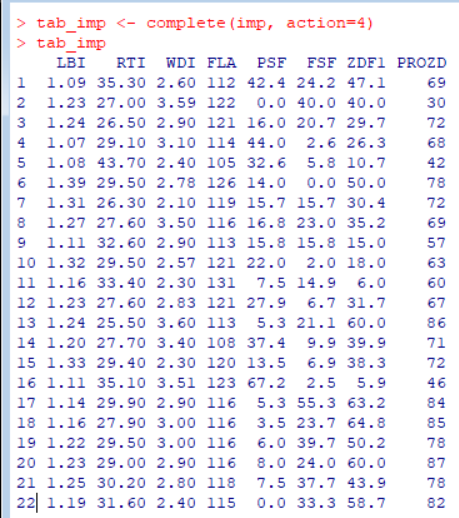
Как и было сказано в предыдущем пункте, их средний коэффициент корреляции равен -0.14 (estimate) и коэффициент Стьюдента (сильно меньше 0.05) говорит о значимости статистической зависимости.

4. наконец, функция complete() создает набор данных с заполненными пропусками:



Как видно, из статистического вывода зависимость искомого параметра от данных параметров статистических значима (р<0.05 для обоих параметров). Получили довольно мелкие значения стандартных ошибок (0.02 и 0.12) при заполнении пустот.

Сохраним и применим 4-ый набор данных на наш датасет:

**

Как видим, пустые данные пропали, посчитаем теперь количество пропущенных значений:



**Вывод:** были получены навыки по выявлению пропущенных значений в датасете, а также их устранение с помощью библиотеки mice().