Отчёт к лабораторной работе

по дисциплине  
«Интеллектуальный анализ данных»

выполнил   
студент гр. ИС/б-18-1-з Демиденко А. А.  
зачётная книжка № 481483  
принял Шумейко И. П.

Лабораторной работа № 2.1  
«Корреляционный и регрессионный анализ данных. Создание набора данных»

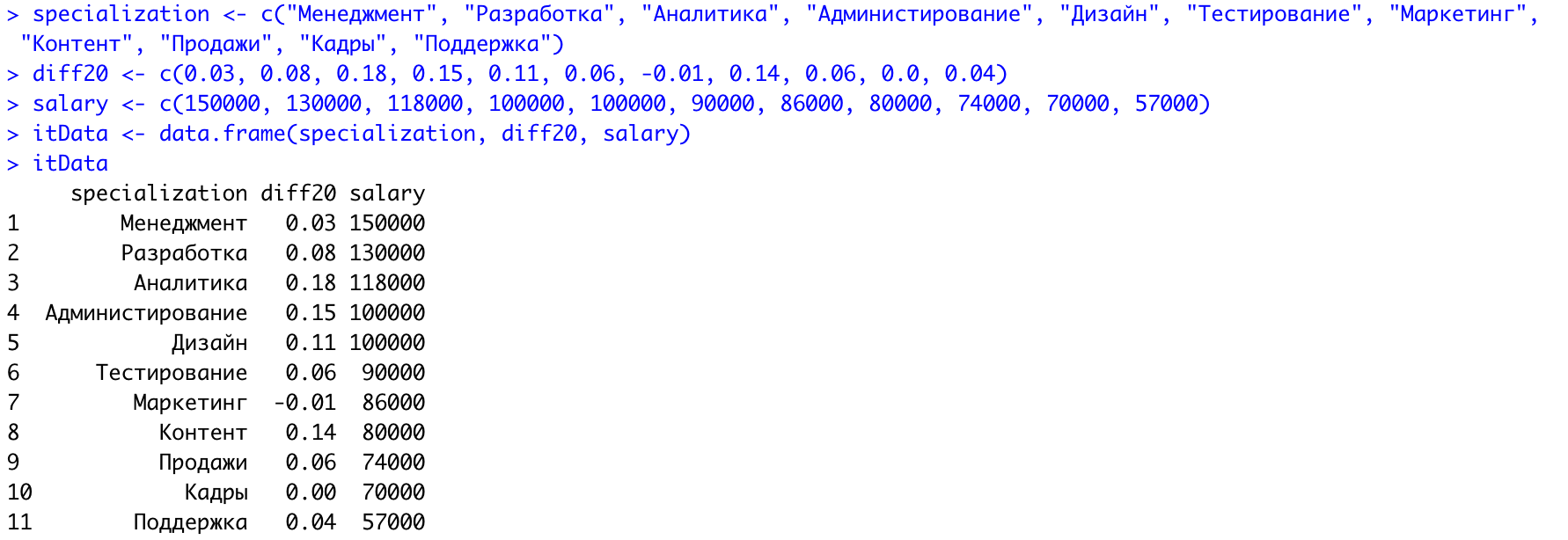
## Цель работы

* исследовать возможности языка R для проведения корреляционного и регрессионного анализа данных;
* создание набора данных для проведения корреляционного и регрессионного анализа данных

## Ход работы

1. Создадим таблицу данных при помощи функции `data.frame()`.

Данная таблица представляет собой данные по изменению зарплат в ИТ в первом полугодии 2021 года по сравнению со вторым полугодием 2020 года:



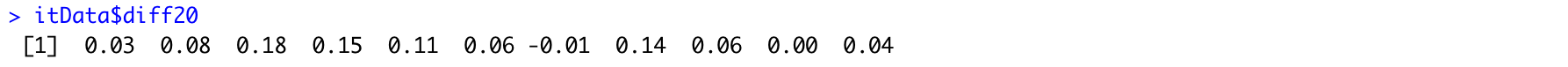
Выведем первый и второй столбцы, отображающие специализацию и изменение заработной платы в процентах:



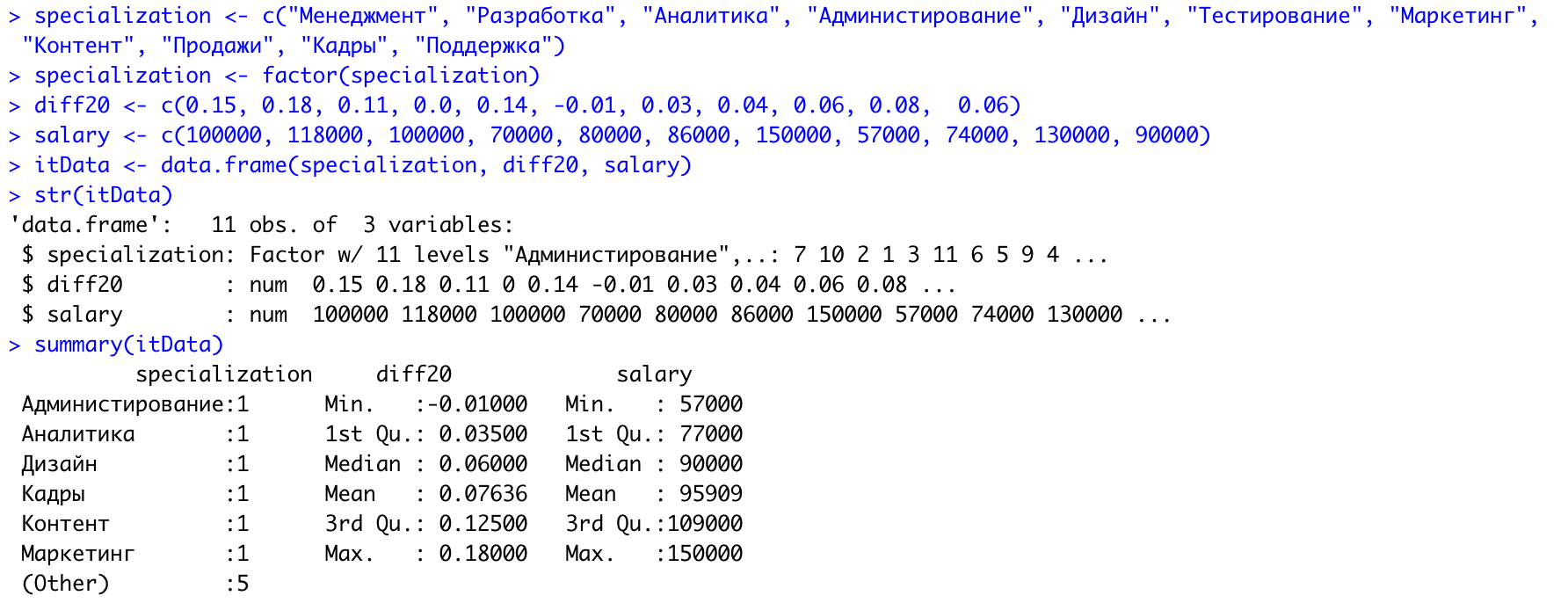
И выведем первый и третий столбцы, отображающие специализацию и медиану зарплаты:



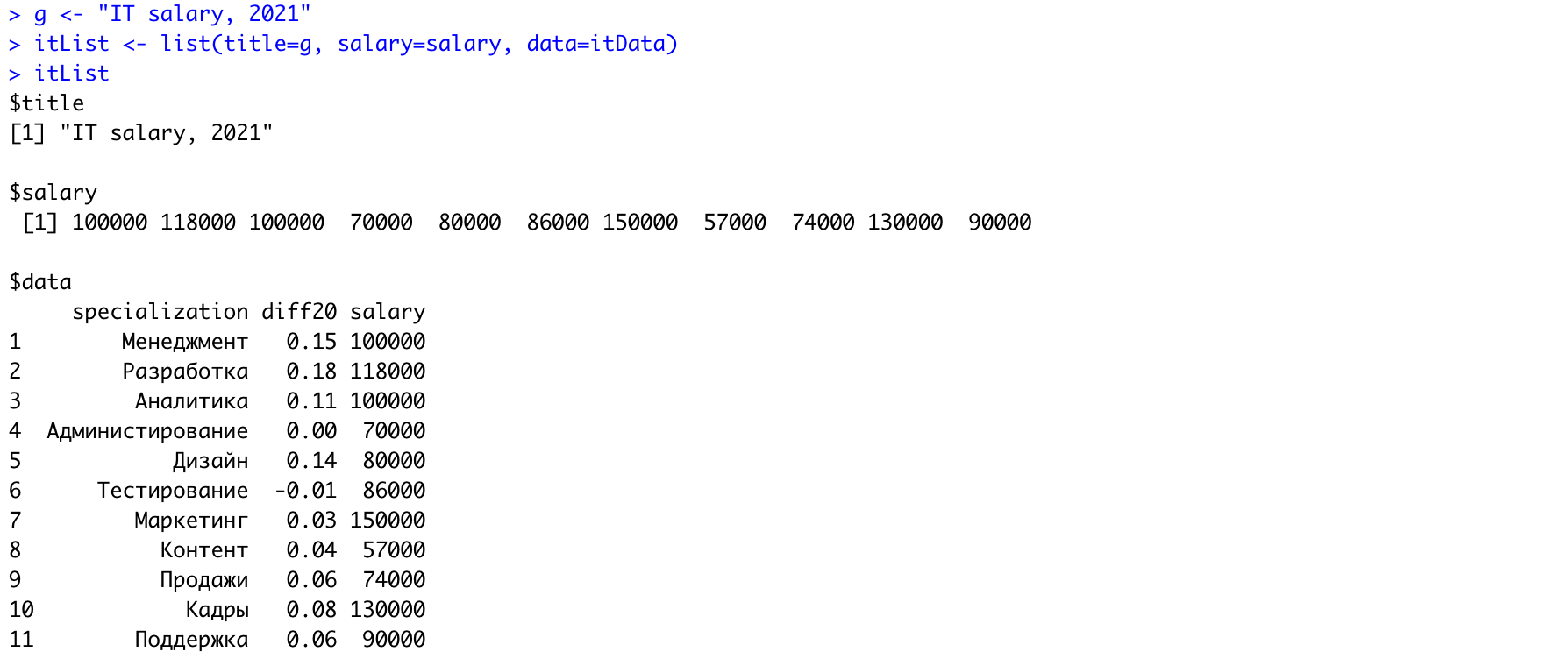
Выведем процент изменения при помощи символа $, для обозначения переменной:



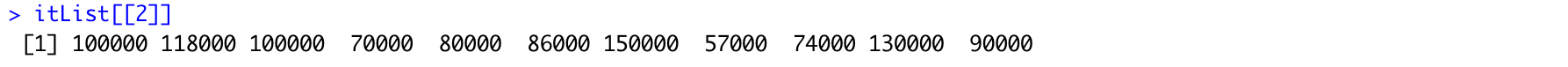
При помощью функции `factor()` сохраним категориальные данные в виде вектора из целых чисел. Далее выведем информацию о таблице данных с помощью функции `str()` и, воспользовавшись функцией `summary()` отобразим сводку по таблице:



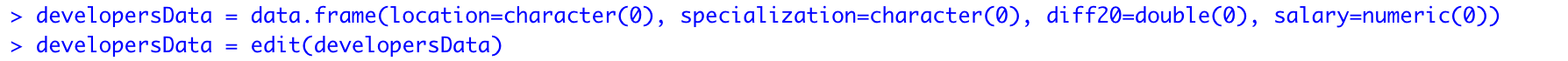
Создадим список, являющийся упорядоченным набором объектов и объединяющий различные объекты под одним именем:

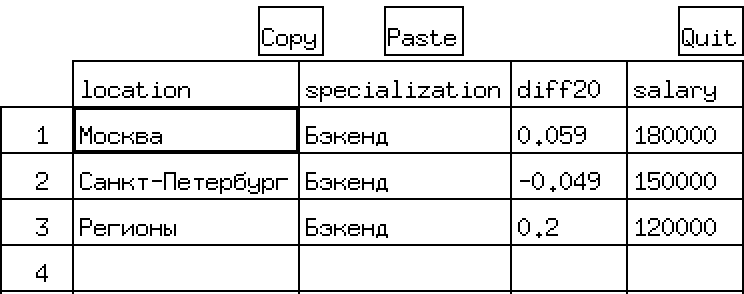


Выведем на экран второй элемент списка:



Произведём ввод данных с клавиатуры с помощью команды `edit()`:





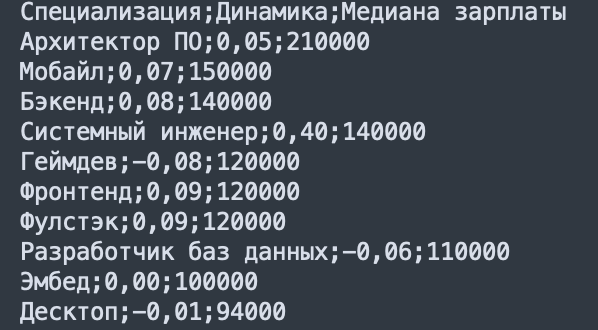
Проверим, что данные сохранились:



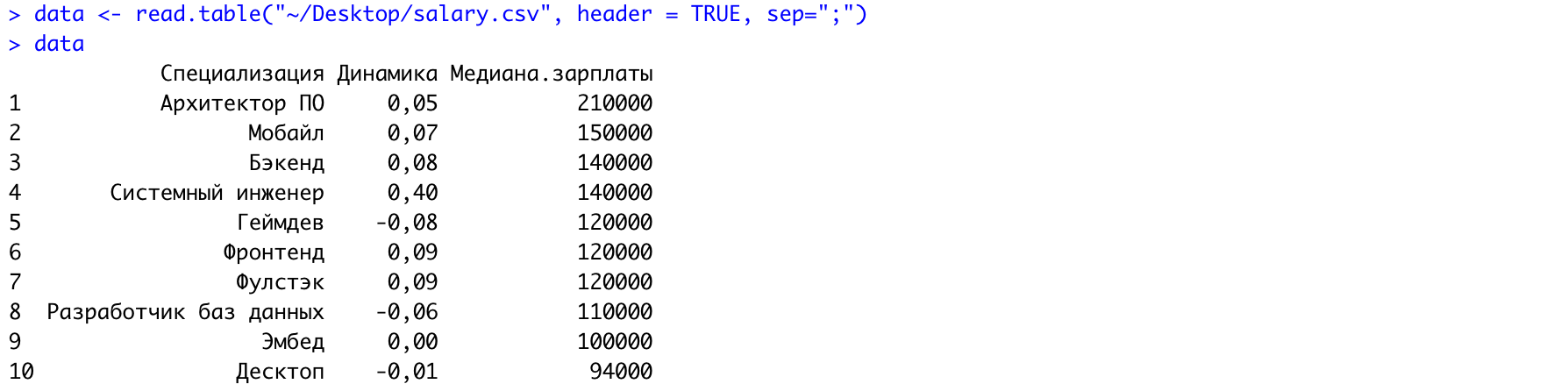
Произведём импорт данных из текстового файла с разделителями.

После формирования таблицы Excel с данными для анализа, сохраним их в формате csv (в формате текстового файла с разделителями).

Данные в файле выглядят следующим образом:

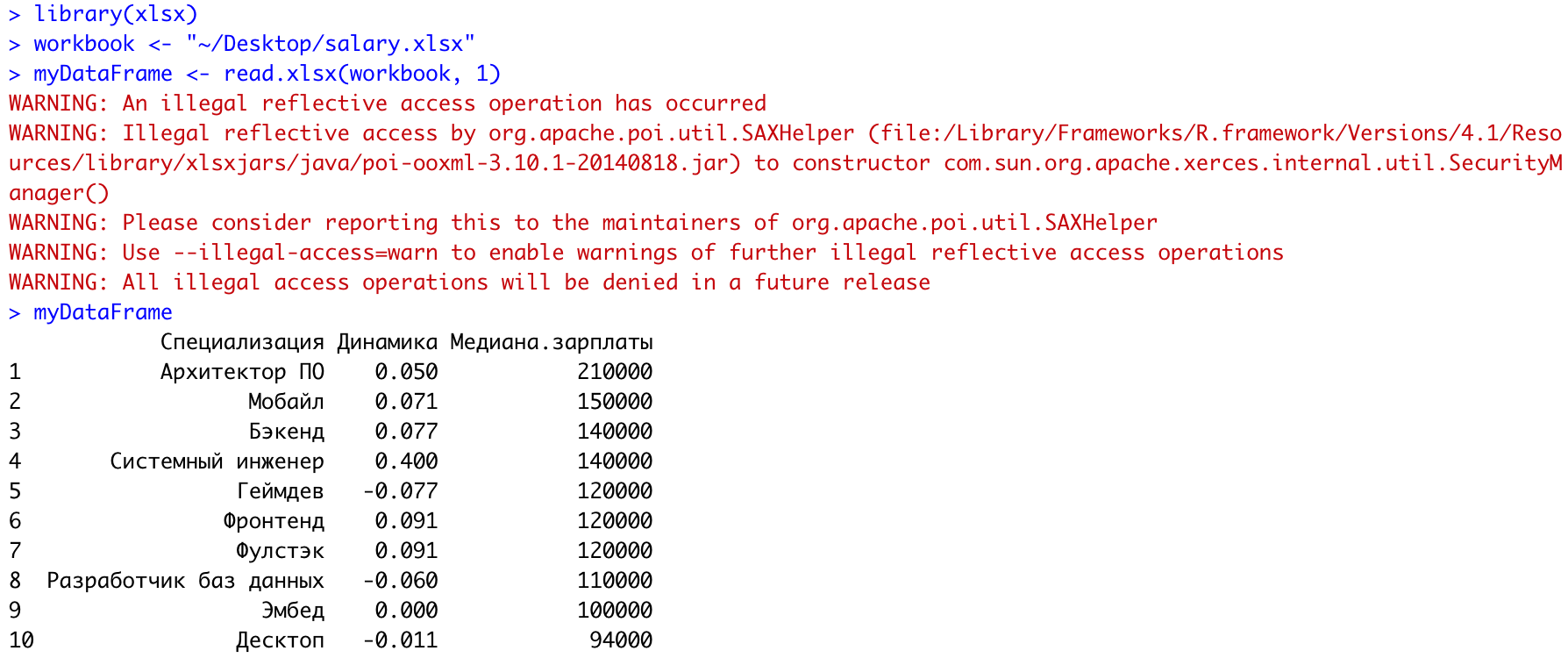


Функция `read.table()` поможет сохранить данные из текстового файла в виде таблицы:



Воспроизведём импорт данных из Excel.

Сохраним в переменную workbook путь к Excel файлу, затем с помощью команды `read.xlsx()` импортируем первый лист книги:



1. Подберём экспериментальные данные для анализа.

Для анализа будут использоваться данные о пассажирах Титаника (источник: <https://www.kaggle.com/c/titanic/data?select=test.csv>).

Файл с данными имеет следующие столбцы:

* PassengerId — идентификатор пассажира;
* Survival — поле, в котором указано спасся человек (1) или нет (0);
* Pclass — содержит социально-экономический статус (высокий, средний, низкий);
* Name — имя пассажира;
* Sex — пол пассажира;
* Age — возраст;
* SibSp — содержит информацию о количестве родственников 2-го порядка (муж, жена, братья, сестры);
* Parch — содержит информацию о количестве родственников на борту 1-го порядка (мать, отец, дети);
* Ticket — номер билета;
* Fare — цена билета;
* Cabin — каюта;
* Embarked — порт посадки (C - Cherbourg, Q - Queenstown, S - Southampton).

1. Ответим на контрольные вопросы.

* Функциональная связь.

Связь, при которой определённому значению факторного признака соответствует одно и только одно значение результативного. Такая связь возможна при условии, что на поведение одного признака (результативного) влияет только второй признак (факторный) и никакие другие.

* Корреляционная связь.

Связь, при которой определённому значению факторного признака может соответствовать множество значений результативного.

* Корреляционный анализ.

Статистический метод изучения взаимосвязи между двумя и более случайными величинами. Суть корреляционного анализа заключается в расчете коэффициентов корреляции. Коэффициенты корреляции могут принимать, как правило, положительные и отрицательные значения. Знак коэффициента корреляции позволяет интерпретировать направление связи, а абсолютное значение – силу связи.

* Корреляционное поле.

Вспомогательное средство анализа выборочных данных. Если даны значения двух признаков x и y (от 1 до n), то при составлении поля точки с координатами (x\_1, y\_l) (х\_n, у\_n) наносят на плоскость. Расположение точек позволяет сделать предварительное заключение о характере и форме зависимости.

* Корреляционный анализ: форма зависимости.

По форме взаимосвязи корреляционные зависимости принято разделять на линейные (см. рисунок 1) и нелинейные (см. рисунок 2).

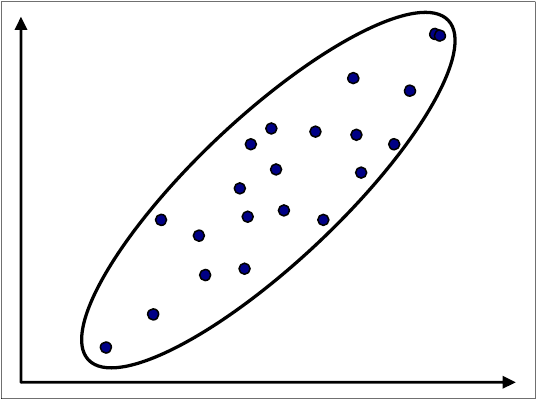
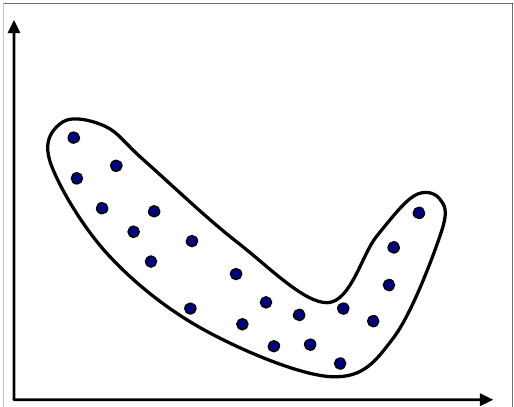


Рисунок 1 – Линейная статистическая связь

При линейной зависимости огибающая корреляционного поля близка к эллипсу. Линейная взаимосвязь двух случайных величин состоит в том, что при увеличении одной случайной величины другая случайная величина имеет тенденцию возрастать (или убывать) по линейному закону.

  
Рисунок 2 – Нелинейная статистическая связь

* Корреляционный анализ: направленность взаимосвязи.

Направленность взаимосвязи бывает отрицательная и положительная. Направленность отрицательная, если увеличение значения одного признака приводит к уменьшению значения второго, положительная, если увеличение значения одного признака приводит к увеличению значения второго.

* Корреляционный анализ: теснота (сила) взаимосвязи.

Для количественного описания связи используются коэффициенты корреляции. Они характеризуют силу (или тесноту) связи, а также ее направление. Любой коэффициент корреляции по абсолютному значению может находиться в диапазоне от 0 до 1, и чем он ближе к 1, тем сильнее (теснее) связь между двумя показателями.

* Коэффициент корреляции. Его свойства.

Величина коэффициента корреляции отражает силы связи. При оценке силы связи коэффициентов корреляции используется условная шкала (см. таблицу 1):

|  |  |
| --- | --- |
| Значение | Интерпретация |
| | r | = 0 | связь отсутствует |
| | r | < 0.3 | слабая связь |
| 0.3 < | r | < 0.7 | умеренная связь |
| | r | > 0.7 | сильная связь |
| | r | = 1 | связь функциональная |

Таблица 1 – шкала для оценки силы корреляции

При r > 0 связь оценивается, как прямая, при r < 0 – обратная.

* Регрессионный анализ.

Если расчёт корреляции характеризует силу связи между двумя переменными, то регрессионный анализ служит для определения вида этой связи и дает возможность для прогнозирования значения одной (зависимой) переменной отталкиваясь от значения другой (независимой) переменной.

* Этапы регрессионного анализа.

1. Формулировка задачи. На этом этапе формируются предварительные гипотезы о зависимости исследуемых явлений.
2. Определение зависимых и независимых (объясняющих) переменных.
3. Сбор статистических данных. Данные должны быть собраны для каждой из переменных, включенных в регрессионную модель (гипотеза).
4. Формулировка гипотезы о форме связи (простая или множественная, линейная или нелинейная).
5. Определение функции регрессии (заключается в расчете численных значений параметров уравнения регрессии).
6. Оценка точности регрессионного анализа.
7. Интерпретация полученных результатов. Полученные результаты регрессионного анализа сравниваются с предварительными гипотезами. Оценивается корректность и правдоподобие полученных результатов.
8. Предсказание неизвестных значений зависимой переменной.

**Выводы**

В ходе лабораторной работы были изучена возможности языка R для работы с данными, способы их ввода в систему; изучена особенный тип данных – факторы. Полученные базовые теоретические знания о корреляционном и регрессионном анализе, собраны данные для дальнейшего анализа.