НЕТОЛОГИЯ
групп

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ #1. РАЗБОР КЕЙСОВ

Максим Чикуров

Data Scientist и руководитель команды аналитики

Работал в компаниях Citibank, BNP Paribas, Barclays Bank, Teradata

maxim.chikurov@gmail.com

О ЧЕМ ПОГОВОРИМ И ЧТО СДЕЛАЕМ

План занятия

- Выводы домашнего задания
- Несколько слов о t-распределении
- Формат CSV
- Gretl
- Регрессионный анализ
- Практика

ВЫВОДЫ ДОМАШНЕГО ЗАДАНИЯ

ВЫВОДЫ ДОМАШНЕГО ЗАДАНИЯ

На основе результатов решения задачи мы сделали выводы о том, что:

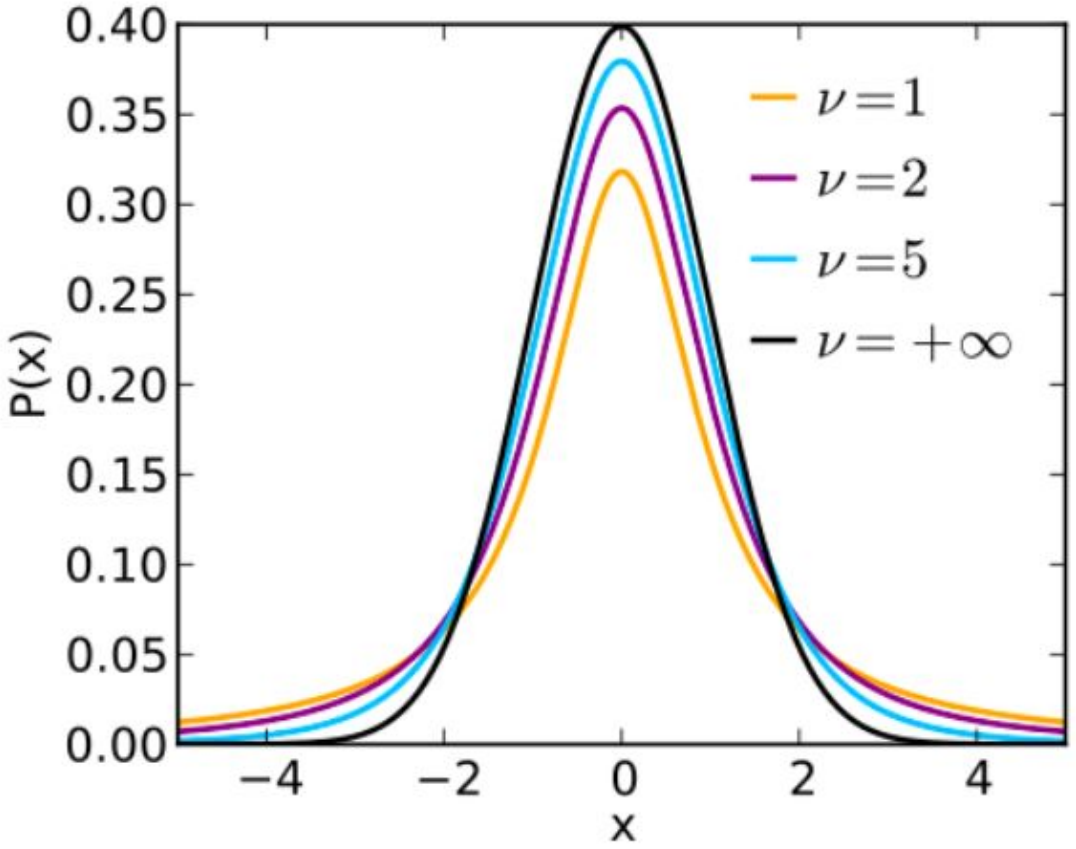
- в рамках групп (по количеству комнат) распределение цен близко к нормальному
- Коэффициент корреляции цен от площади в рамках групп ниже общего.
- Чем выше дисперсия тем ниже коэффициент корреляции

Несколько слов о t-распределении

t-распределение (Стьюдента)

Распределение Стьюдента может быть использовано для:

- оценки статистического значения разницы между двумя выборочными средними
- построения доверительных интервалов разницы между двумя доверительными средними
- оценки того, насколько вероятно, что истинное среднее находится в каком-либо заданном диапазоне



T-РАСПРЕДЕЛЕНИЕ

Для n значений выборки из непрерывного распределения с известным средним.

Среднее выборки и **дисперсия** определены как:

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n},$$
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

t-значение рассчитывается как:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

для t-распределения с n - 1 степенями свободы.

Формат CSV

Применение формата CSV

CSV (от англ. Comma-Separated Values — значения, разделённые запятыми) — текстовый формат, предназначенный для представления табличных данных.

- Каждая строка файла — это одна строка таблицы
- Разделителем (англ. delimiter) значений колонок является символ запятой (,)
- Значения, содержащие зарезервированные символы (двойная кавычка, запятая, точка с запятой, новая строка) обрамляются двойными кавычками (")

Gretl

Инструмент Gretl

GNU Regression, Econometrics and Time-series Library: Библиотека для регрессий, эконометрики и временных рядов) — прикладной программный пакет для эконометрического моделирования.

Эконометрика — наука, изучающая количественные и качественные экономические взаимосвязи с помощью математических и статистических методов и моделей.

http://gretl.sourceforge.net/

Регрессионный анализ

Регрессионный анализ

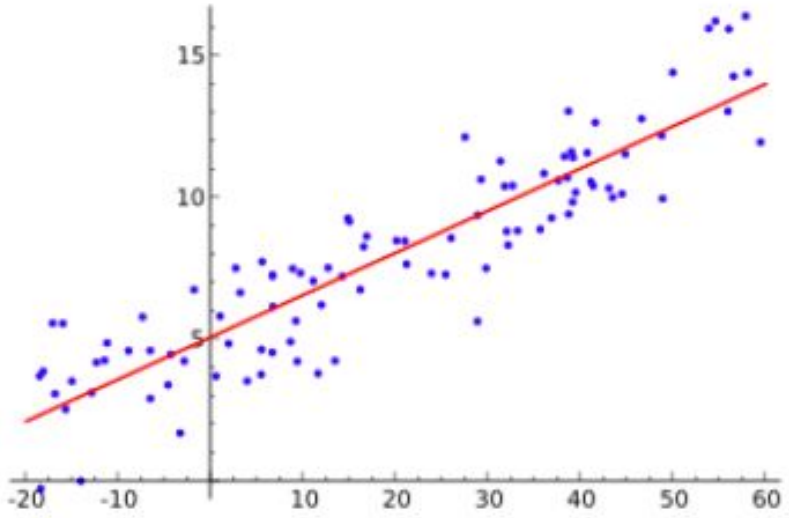
Статистический метод исследования влияния одной или нескольких независимых переменных **x** на зависимую переменную **y**.

Модель линейной регрессии:

$$f(x, b) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

Простейшая линейная регрессия:

$$y_t = a + b x_t + \varepsilon_t$$



Словарь терминов

- ***Коэффициент детерминации (R-квадрат)*** — это доля дисперсии зависимой переменной, объясняемая рассматриваемой моделью зависимости, то есть объясняющими переменными.

- ***Метод наименьших квадратов (МНК, Ordinary Least Squares, OLS)*** — математический метод, применяемый для решения различных задач, основанный на минимизации суммы квадратов отклонений некоторых функций от искомых переменных.

ПРАКТИКА

Датасет №1: оборот по отделам

Датасет содержит агрегированные данные о:

- количестве открытых интересов по отделам на конец месяца
- сумме заключенных сделок в следующем месяце

https://goo.gl/2mHi9Y

Задача: понять есть ли связь в количестве открытых интересов и сумме заключений

Датасет №2: трафик в интернет-магазине

Столбец №1 – clientID (идентификатор клиента).

Далее: данные по транзакциям, количеству сеансов, виды трафика по устройствам, по каналам, и данные о промежуточных конверсиях (значения 1 и 0). Строки с одним clientID могут повторяться, если любой другой параметр отличен от других строк с этим clientID.

goo.gl/tUkN6j

Датасет №2: трафик в интернет-магазине

Вопросы:

- как соотносятся промежуточные конверсии с транзакциями
- как влияют устройства на промежуточные конверсии и транзакции
- как влияют источники трафика на промежуточные конверсии и на транзакции
- есть ли какая-то связь между количеством сеансов, длительностью сеансов и промежуточными конверсиями и на транзакциями
- можно ли спрогнозировать конверсии и транзакции, исходя из предыдущих данных
- какие характеристики присущи клиентам, которые вообще никогда не закажут

СПАСИБО ЗА ВНИМАНИЕ