

numpy - библиотека за „scientific computing“, слична је matplotlib

`import numpy as np`

matplotlib - библиотека за *plotting*

`import matplotlib.pyplot as plt`

pandas - библиотека за манипулацију подацима

`import pandas as pd`

Главни објект је **DataFrame** који служи за рачунање табеларних података; може се дефинисати као матрица редова и колона и многи други додатних функционисања

- **диференцијална (функционална) зависност** - одржавање вр. зависне пром. и независ. пром. x (објасни) - тако је описан однос две променљиве

- **статистика зависности** - из разматрања, без ит. две променљиве није савршено јасно

- према x пронађемо линију која ће најбоље нашим подацима да одговара: $\hat{y} = ax + b$

x_i - вредности независне променљиве за свако мерење i

y_i - вредности зависне променљиве за свако мерење i

\hat{y}_i - предвиђене вредности зависне променљиве за свако мерење i

y_i и \hat{y}_i разликује се одређеном у неким мерењима, а циљ је да ту разлику смањимо минимизује се грешка функције $\epsilon_i = y_i - \hat{y}_i$

- један од начина јесте **метод најмањих квадрата OLS** - минимизује суму квадрата грешака **SSE** (говори колико добро линија одговара подацима)

- **линеарност** - претпоставља да важи некако однос између x и y линеаран. Можемо да сада исцртамо ВИЗУЕЛНО одређено графика, а други начин за исцртавање је коришћење статистичких метода (користимо Т-тест)

- **Т-тест** испитује да ли је линеаран однос између x и y - тј. претпоставља да НЕ ПОСТОЈИ да $a=0$ и $y=ax+b$

• Т-вредности није честита (далеко од нуле) - објаснило би значење и кажемо да је однос између x и y **линеаран**

• Т-вредности није честита (као нуле) - не можемо да објаснимо хипотезу

ако је вр. чести не знамо да је нелинеаран, већ не знамо да ли је **већи линеаран или нелинеаран**

- уз помоћ Т-вредности, линеарности можемо проверити и користити **Р-вредности** - то је вероватноћа да ћемо из t -дистрибуције изабрати конкретну Т-вредност или неку још мање вероватну

- ако је Р-вредност мала, Т-вредности није честита и **већи** је **линеаран**

Р-вредности је мала $\leq 0,05 \rightarrow 95\%$ смо сигурни да је **већи линеаран**

- више не можемо користити графика и као нешто проверили линеарност већ користимо **т-тест** 1

- Екуј бр. са којим радимо садржи податаке о 280 кућа у Францу. Користимо линеарни модел $y = ax + b$ како бисмо пронашли бр. кућа (немамо податаке о ценама свих кућа, већ само о одр. бр. кућа)
- други податаке су смо имали - доб. бисмо модел са др. бр. за најбољ и пресек
- тако ће се променити параметри модела када бисмо проширили скуп података, колико је иф. модел добар - да бисмо не ова одговорили - првенствено сигурнош кућ зовемо **интервал поверења** које је распон у коме ће се наћи права вистинити параметри (прв бр. би добили када бисмо имали све податке)
- интервал поверења креира се за неки проценат поузданости, гласимо смо код р-вистинити узимамо праг од 95%. - штако повећемо неможе са 95% вистинити одређено измеђ којих бр. се налази најбољ
- `model.conf-int() @ model.summary()`
- до сада смо интервал поверења израдили за параметре најбољ и одсечак, али можели смо дефинисати и за САМЕ ПРЕДИКЦИЈЕ
- измеђ којих бр. се налази y_i за дамо x_i ? (који распон у коме куће са њим $y_{\text{отп}}?$)
- `model.get-prediction-with-const().summary-frame()`

L.I.N.E - 4 предпоставке (могуће је да неке ипак нису биле сигурне за је референца модела погодан за даљи провек)

- 1) **линеарност** - важи када је однос измеђ X и Y линеаран
тест - графички преглед да ли је близу праве $x=y$ како би предикције биле вистините
 ВЕЛИКА одступања - модел није адекватан
 МАЛА одступања - модел је адекватан
- 2) **независност грешака** - важи када су грешке статистички независне
 Грешке се нег резидуумама не из осцилација
 резидуали - проценат грешака из узорка података. Разлика измеђ прве и проценатне бр. $e = y - \hat{y}$
 \rightarrow испитано штако што ЦРТАМО ГРАФИК РЕЗИДУАЛА
 Резидуали имају очигледан шаблон - предпоставка **нарушена**
 нема очигледан шаблон - предпоставка **није нарушена**

Дуринт-Замсон статистички тест

$1.5 \leq d \leq 2$ нема аутокорекулације \rightarrow ПРЕТРАЖИВАЊЕ
 < 1.5 и/или > 2 и/или \rightarrow НЕ ВАЖИ

- 3) **нормалност грешака**
 - средња бр. 0 за грешке
 - грешке су нормално дистрибуиране **Андерсон-Дарлинг** статистички тест - **гласимо**
 - р-вр. $>$ = праг \rightarrow дистрибуција НОРМАЛНА
 - р-вр. $<$ = праг \rightarrow дистрибуција НИЈЕ НОРМАЛНА

- ор. бр 0 али нису нормално дистрибуирани \rightarrow ПРЕТП. НЕ ВАЖИ
- али ако је узорак податка велик, занемарљивим нису нормално дистрибуирани \rightarrow асимпт. важе

4) једнака варијанса - важи када тачијане фреквенце око регресионе праве имају једнаку варијансу (ако видимо са дефинице резидуала) - када неможемо да кажемо

Голдберг-Куант исити - бр. p - бр. q

- p бр \geq q - ВАРИЈАНСА ЈЕ ЈЕДНАКА
- p бр $<$ q - ВАРИЈАНСА НИЈЕ ЈЕДНАКА

ФИШЕСТРУКА

+ још једна исити.

5) независни садржане колинеарности између 2 или више независних променљивих садржане колинеарности исити када су променљиве x_1 и x_2 повезане линеарном фнкцијом:

$$x_2 = ax_1 + b$$

колинеарност доказано преко МАТРИЦЕ КОРРЕЛАЦИЈЕ

- $\text{свои} = \text{има}$ велич. r припадајућим r

- ИНТЕРПОЛАЦИЈА - ПРЕМЕШАЊА ПОДАТАКА