25.01.20 / Консултация / R /

I Описателна статистика

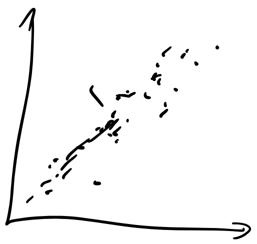
II Тестове

III Разпределения

IV Линейна регресия → ако го направиш, минаваш

IV Лин. регресия

- ще имаме графики



$y = a + bx + c$

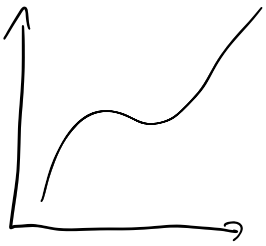
↳ също са лин. регресии

корелационен анализ

cor ∈ [−1, 1] → pearson, spearman, Kendal ②
↓ ефективността намалява, ↑ стабилността се увеличава

|r| = [0, 1]
↙ няма корелация ↘ детер.

да определим по графики ф-ята y



→ $y = ax^3 + bx^2 + cx + d$

cor(X, Y, method = ①)

X, Z, m ⟷ Y

с корелацията взимаме връзката и избираме най-силната променлива

Линейна регресия

model = lm(Y ~ X) → връща a и b
(Y — зависима; X — зависи от)

summary(model) → таблица: Est., SE, z-value, p-value

↳ определяме дали a и b са значими → при [|t| > 2 ; pvalue < 0,05]

от Est. взимаме a и b

с най-голямо |t value| → най-значим

Да определим дали модел е добър или не — Adjusted R^2 (→ идва от summary)

добър е когато Adj. $R^2 \to 1$

Може и с Multiple R^2

Трябва да проверим дали изпълнява условията — подмножество остатъци с грешки

- Хомоскедастичност на остатъците

r = residuals(model)

$r = Y_i - \hat{Y}$

реална, прогнозирана (predict(model))

Условия за модел да е правилен:

1. Homoscedasicity
   - през графика r  → по ялата графика трябва да са
   - през тест — bptest(model) → трябва p-value > 0,05

Кое да изберем — има const вариация

ако я има графика — plot($\hat{Y}$, r)

2. Да няма автокорелация

lmtest::dwtest(model)

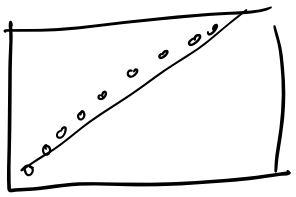
Хипотеза: липсва автокорелация — p-value > 0.05

H_0

3. $r \sim N(0, \sigma)$

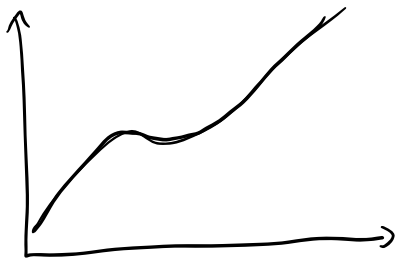
а) тест - shapiro.test(r)

б) qqnorm(r); qqline(r)



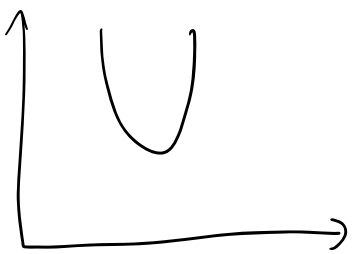
→ когато опашките са близо до линията

По-сложни връзки



$y = ax^3 + \dots + nx + b$

$y \sim I(x^3) + I(x^2) + x$



$y \sim I(x^2) + x$

IV Тестове

1, 2 променливи - t.test, wilcoxon

> 3 - ANOVA, ~~Kruskal~~ Kruskal (!!)

```
y = [   ]
x = [   ]
z = [   ]
```

1. Дали са $N(\ )$ → shapiro.test(y) (z)

ако всички са p-value > 0.05

=> ANOVA

H_0 - локациите са равни/различни

ANOVA: aov(value ~ group, DF) => p_value — от него казваме дали са различни или не
>0.05 са еднакви (зависи от H_0)

↳ първо трябва DF = data.frame(group = c(1, 2, 3),
value = c(x, y, z))

ако има разлика от shapiro => ползваме Kruskal
т.е
поне едно p-value < 0.05

III Разпределения

1. Дискретни разпределения

Bi, Po, Ge
Ber

Ber — ползваме за "нещо дали ще стане или не"
— параметър p

p / q / r / d binom (n, p)
size = 1, p)

Bi — n-на брой независими опита

Примерна задача
n = 100 (100 заведени щети)
p = 0,1 (плащаме ги в 10% от случ)
→ затова е p = → тогава е d

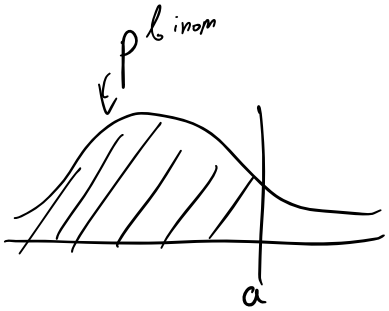
$\mathbb{P}(k>3) = ?$

pbinom (3, 100, 0,1) => $\mathbb{P}(x \leq 3)$
k n p(0,1)

=> трябва 1 - pbinom(3, 100, 0,1)
или
pbinom(3, 100, 0,1), low_tail = TRUE

Ако имаме $P(k \geq 2, k < 4)$ → dbinom с 2 и 3

→ $P(k < 4) - P(k < 2)$



qbinom → знаем вероятността, търсим a

rbinom → симулация

dbinom → в точка

- Геометрично - до първи успех
- Поасоново - времеви интервал
  ↳ колко събития във времеви интервал

2. Непрекъснати

- равномерно - хвърляне на зар
  гадост делегата
  тотото
- експоненциално - време между две събития
- нормално разпределение - природата

~~2 разредени~~

I Описателна статистика

↙ непрекъснат

	параметрична	непараметрична
локация	mean	median
дисперсия	sd var	IQR → 3-ти quartile − 1-ви quartile mad

да покажем разпред. hist()
qqnorm

фискретно → mode
категорийни с малки брой
↳ barplot()

Как се правят ф-ции

$$x^2 + x^4 + x^6 + \dots + x^{16}$$

```
foo <- function (x) {
  s <- 0
  for(i in 1:8){
     s=s+ x^(2*i)
  }
  return(s)   //може s
}
```

```
x= [      ]
```

```
mean = function (x) {
  return ( sum(x)/len(x));
}
```

дисперсия, стандартно отклонение