

Moderná Aplikovaná Regresia 1

Lukáš Lafférs

KM FPV UMB
www.lukaslaoffers.com

O čom bude tento kurz

Úlohou tohoto kurzu je sprístupniť študentom lineárne regresné modely. Dôraz je kladený na praktické aplikácie v štatistickom programe R. Počas kurzu budú študenti analyzovať desiatky datasetov, na ktorých budú demonštrovať naučené techniky. Od študentov sa očakáva znalosť základov pravdepodobnosti, štatistiky a lineárnej algebry.

www.lukaslauffers.com/mar

13 týždňov (17. Sept - 14. Dec, týždne 38-50)

- Utorok 10:40 - 12:05 Počítačové cvičenia so Samuelom Hudecom (samuel.hudec@umb.sk)
- Streda 08:15 - 11:20 F237 Prednáška + cvičenia
- [4ECTS] - KMA FPV/2d-mxx-047 Moderná aplikovaná regresia 1
- [3ECTS] - KMA FPV/2d-mxx-547 Moderná aplikovaná regresia 1 - cvičenie

Hodnotenie

30% 10 Domácich úloh (každá 3%)

20% Písomná skúška

10% Projekt

40% Ústna skúška

Odporúčaná literatúra

- Povinná

F Julian Faraway: Linear Models with R, second edition, CRC press
(web accompanying the book
<http://www.maths.bath.ac.uk/~jjf23/LMR/>).

- Doplnujúca

C Crawley, Michael J. "Statistical modelling."The R Book, Second Edition (2007): 388-448.

B Berk, Richard A. Statistical learning from a regression perspective. Springer Science & Business Media, 2008.

G Greene, William H. Econometric analysis (International edition)."(2000).

H Wickham, Hadley. ggplot2: elegant graphics for data analysis. Springer Science & Business Media, 2009.

CH Claeskens, Gerda, and Nils Lid Hjort. Model selection and model averaging. Vol. 330. Cambridge: Cambridge University Press, 2008.

FHT Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. Springer, Berlin: Springer series in statistics, 2001.
<http://statweb.stanford.edu/~tibs/ElemStatLearn/>

WHT James, G., Witten D., Hastie T., Tibshirani R. An introduction to statistical learning with applications in R, Springer, 2013.

Prehľad kurzu [odkazy] (1)

- Week1** Pravidlá, úvod, motivácia, základy exploratívnej analýzy dát a vizualizácie, sumárne štatistiky. [F:ch1, B:ch1,2, H]
- Week2** Odhadovanie parametrov lineárneho modelu, estimátor, geometrická inuícia, príklady, QR dekompozícia, Gauss-Markovova veta s dôkazom. [F:ch2]
- Week3** Štatistická inferencia - základy testovania hypotéz, test pomocou pomeru vierohodností, F-distribúcia, rozklad súm štvorcov, testovanie hypotéz, permutačný test, konfidenčný interval, bootstrap [F:ch3, G:ch4,5]
- Week4** Predikcia vs. Vysvetlenie - konfidenčné intervaly pre predikciu a vysvetľovanie, príklady, koncept kauzality, nadizajnované experimenty vs pozorované dáta, matching, kvalitatívna podpora kauzality, interpretácia. [F:ch4,5]
- Week5** Diagnostika - predpokady o náhodných chybách: konštantná variancia, normalita, korelované chyby, nezvyčajné pozorovania: pákové pozorovania, outliery, Cookova štatistika, štruktúrna časť modelu: parciálna regresia. [F:ch6]

Prehľad kurzu [odkazy] (2)

- Week6** Kolinearita - variance inflation factor, praktické problémy, chyby v prediktoroch, zmena škály, Metóda zovšeobecnených najmenších štvorcov, testovanie nesprávneho fitu, Robustná regresia [F:ch7,8]
- Week7** Transformácia, transformácia odozvy, Box-Cox-ova transformácia aj s odvodením, transformácia prediktorov, "hokejková" regresia, polynómy [F:ch9]
- Week8** Výber modelu - spätná eliminácia, kritériálne metódy - AIC, BIC - konzistentnosť vs. efektívnosť, príklady. [F:ch10, CH:ch2,3,4]
- Week9** Zmenšovanie - metóda hlavných komponentov, parciálne najmenšie štvorce, Ridge regresia, LASSO, krížová validácia. [F:ch11, FHT:ch3.4, WHT]
- Week10** Kompletný príklad z oblasti poisťovníctva - uvodná analýza dát, sumárne štatistiky, kompletný model a diagnostika, analýza citlivosti s diskusiou. Chýbajúce pozorovania, MCAR (missing completely at random), MAR (missing at random), MNAR (missing not at random), imputácia [F:ch12,13]

Prehľad kurzu [odkazy] (3)

- Week11** Kategorické premenné - dvojlevelový faktor, príklad, faktory a kvalitatívne prediktory, interakčné členy, viaclevelové faktory, ANOVA, kódovanie kvalitatívnych prediktorov. [F:ch14]
- Week12** Jednofaktorový model, diagnostika, párové porovnávania, false discovery rate. [F:ch15]
- Week13** Prezentácie projektov, kompletný príklad, diskusia, zopakovanie.

Dáta

- sú to najdôležitejšie čo máme
- Akým spôsobom boli dáta zbierané?
- Experiment ako zlatý štandard
- Chýbajúce pozorovania? Prečo?
- Čo ktoré premenné znamenajú?
- O čo nám ide? Predikcia vs. porozumenie.
- Chybme zaznamenané dáta?
- Podozrivo veľké/malé hodnoty?

- sú to najdôležitejšie čo máme
 - Akým spôsobom boli dáta zbierané?
 - Experiment ako zlatý štandard
 - Chýbajúce pozorovania? Prečo?
 - Čo ktoré premenné znamenajú?
 - O čo nám ide? Predikcia vs. porozumenie.
 - Chybme zaznamenané dáta?
 - Podozrivo veľké/malé hodnoty?
-
- komunikácia s klientom, poznať jeho potreby
 - jednoduchý vs. zložitý model

Vizualizácia je extrémne dôležitá.

Vizualizácia je extrémne dôležitá.



No. of Meetings



0



1



2



3



5



7

#1 FREE DAILY NEWSPAPER IN NYC



328,296



metro

326,889



amNY

Source: For month March-April average net circulation.
Circulation's Statement for period ended 3/31/2011

METRO NEW YORK

is the No. 1 circulated free daily newspaper in New York City reaching an unduplicated audience of young professionals in the big apple.

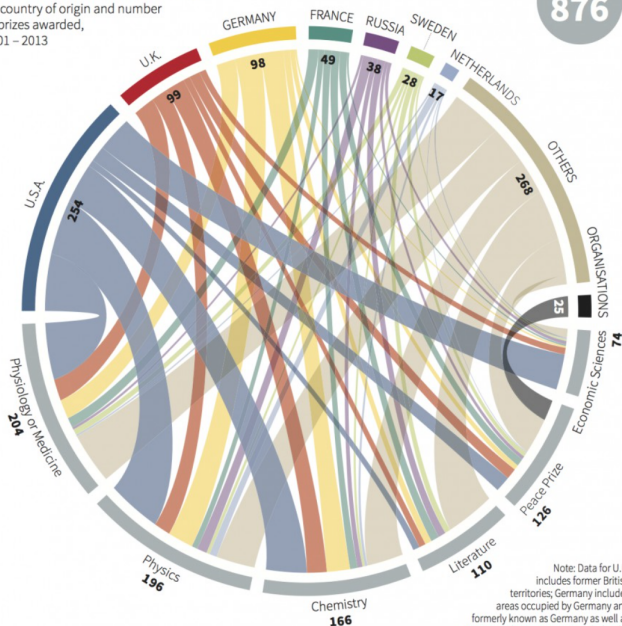
metro

THE PUB. IN NEW YORK CITY

NOBEL LAUREATES

By country of origin and number of prizes awarded, 1901 – 2013

TOTAL
876



Note: Data for U.K. includes former British territories; Germany includes areas occupied by Germany and formerly known as Germany as well as Bavaria and Prussia; Russia includes the former Russian Empire and U.S.S.R.

How do the Poor Spend Their Day?

QUICK
VIEW:



RADIAL

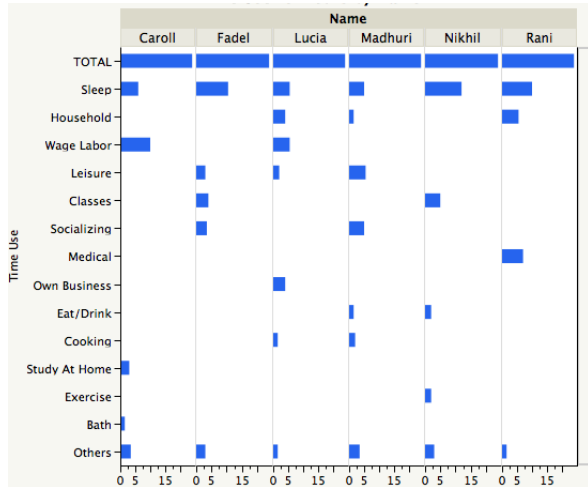


LINEAR

SELECT:

COMPARE DAILY ACTIVITIES

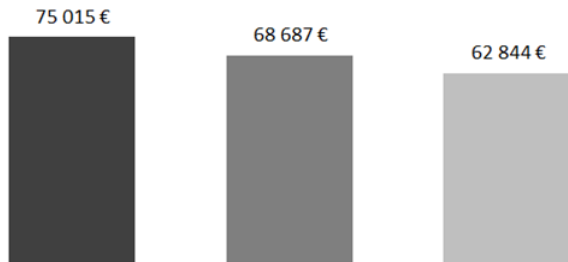




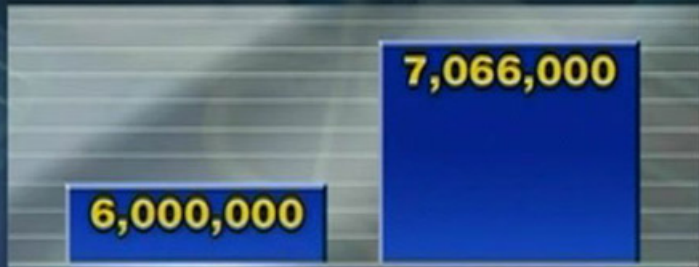
Graf Partners Group



Skutočnosť



OBAMACARE ENROLLMENT



AS OF
MARCH 27

MARCH 31
GOAL

SOURCE: HHS

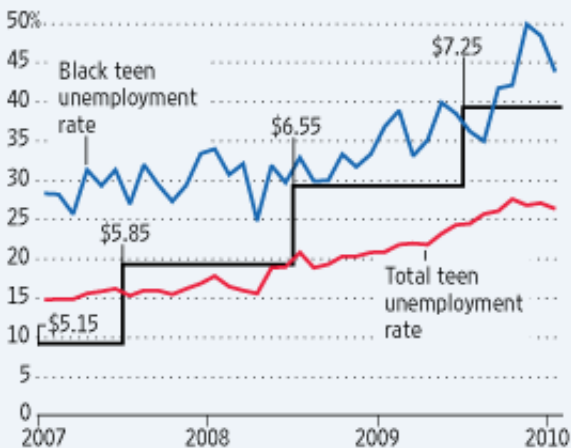


mediamatters.

IN SECOND PLACE WITH \$26.5 MIL, WHILE "MUPP (DOW FUT 16,325.0

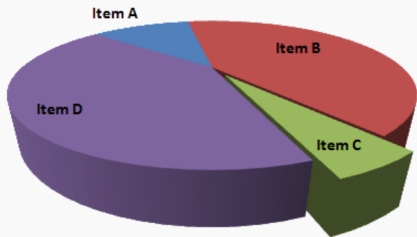
The Young and the Jobless

Teen unemployment rates and minimum wage, 2007-2010

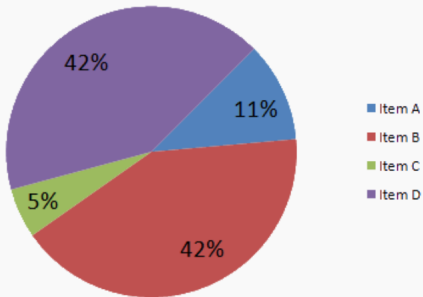


Source: Bureau of Labor Statistics

Misleading pie chart



Regular pie chart



Home Sales



2000



2001



Yiewsley ▼

Now 17°C

1am 16°C

HOME

ELECTION 2017

FOOTBALL

SPORT

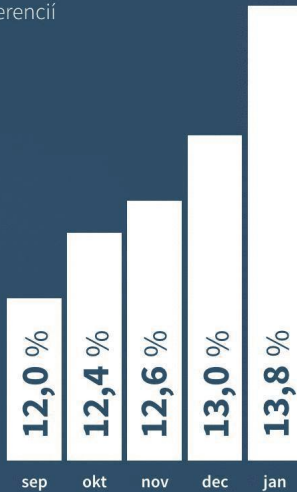
TV & SHOWBIZ

NEWS

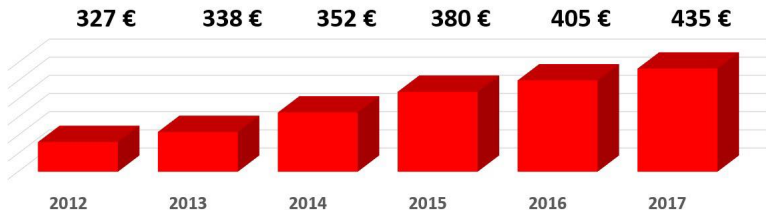
LIVING



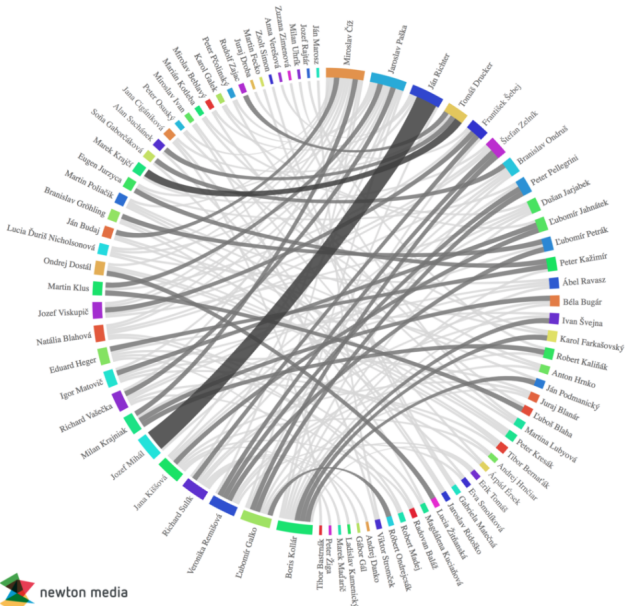
agentúra Polis - prieskum preferencií

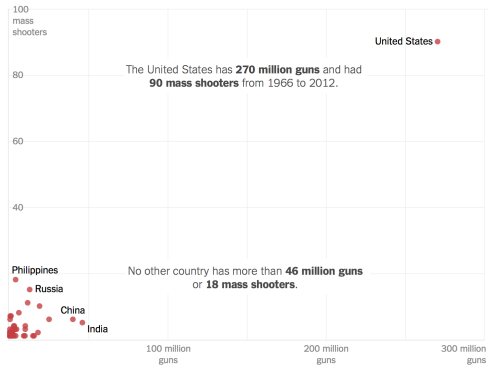


Vývoj výšky minimálnej mzdy od roku 2012



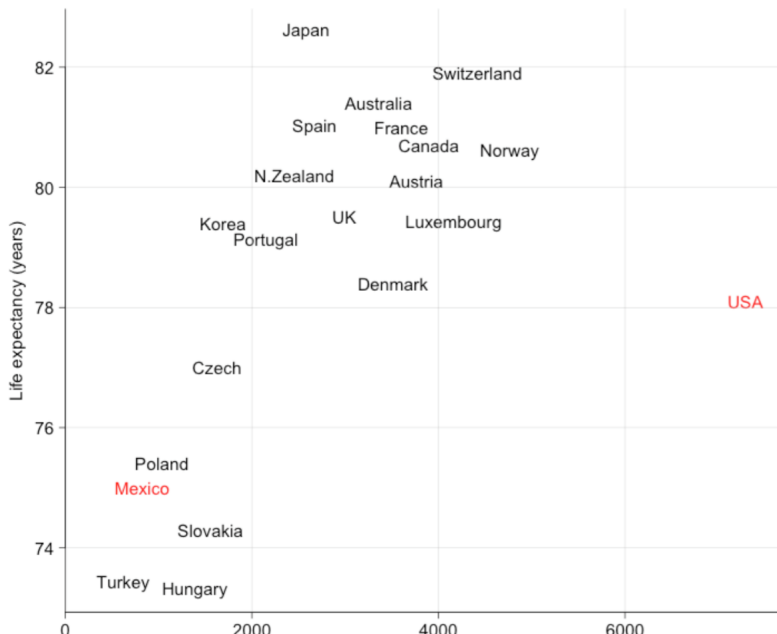
Dvojice politikov, ktoré sa stretli v diskusiách v roku 2017. Zdroj – Newton Media





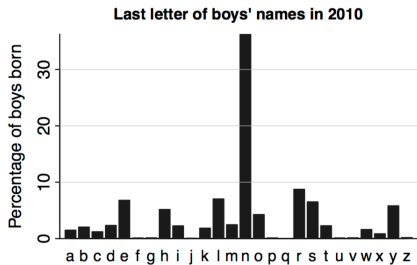
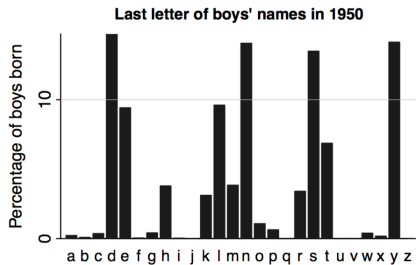
The New York Times | Source: Adam Lankford, The University of Alabama (shooters); Small Arms Survey (guns). Note: Includes countries with more than 10 million people and at least one mass public shooting with four or more victims.

A teraz nějaké dobré

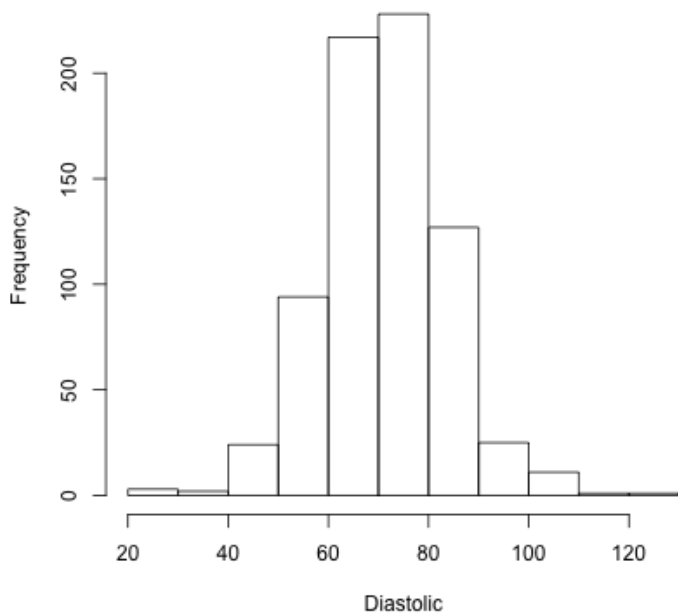




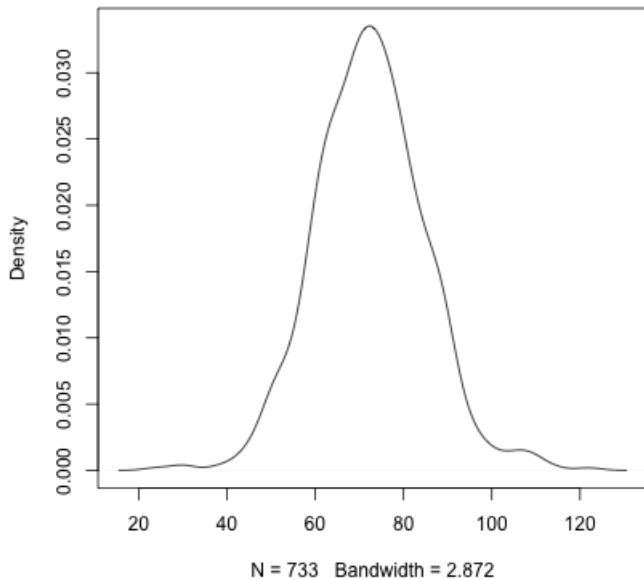
Čo tento?



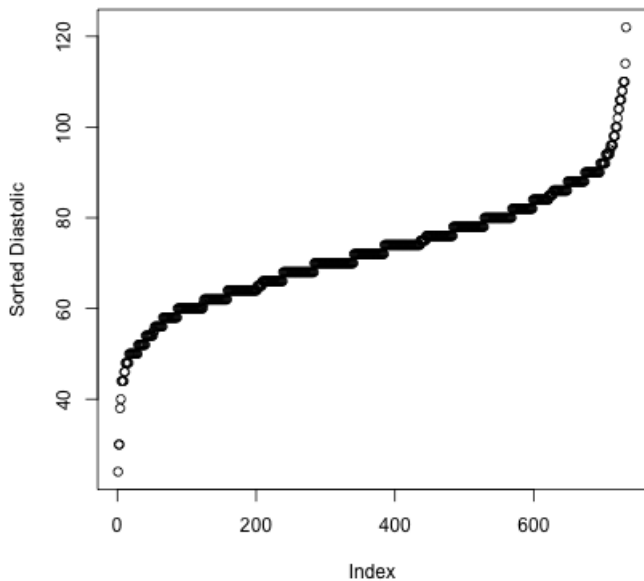
Histogram



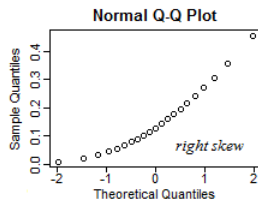
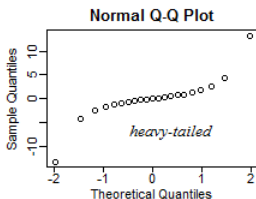
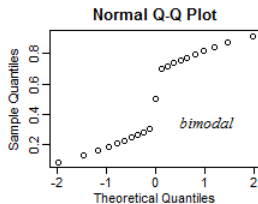
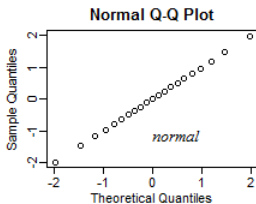
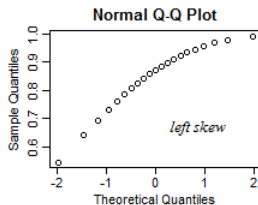
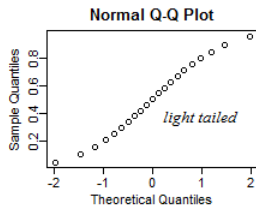
Jadrová hustota



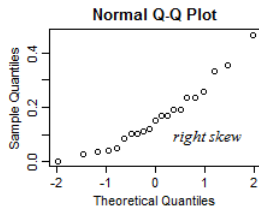
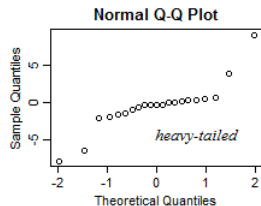
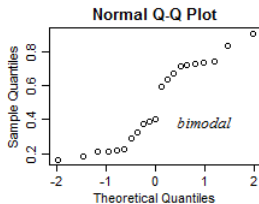
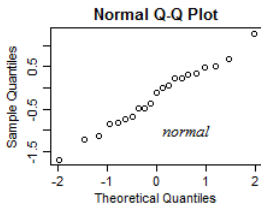
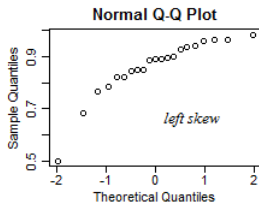
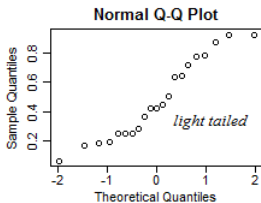
Graf usporiadaných dát



Priemerný QQ-plot

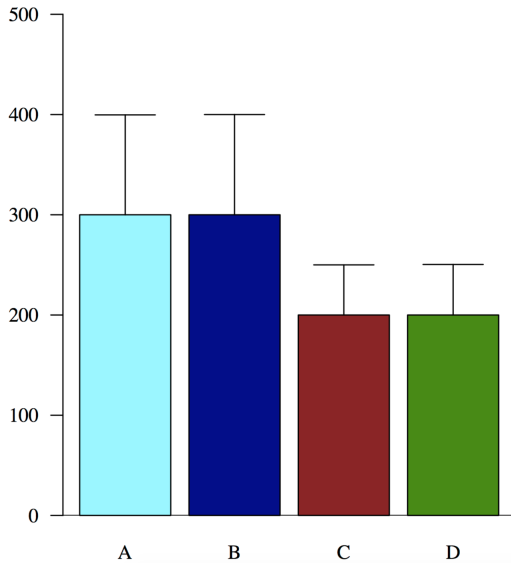


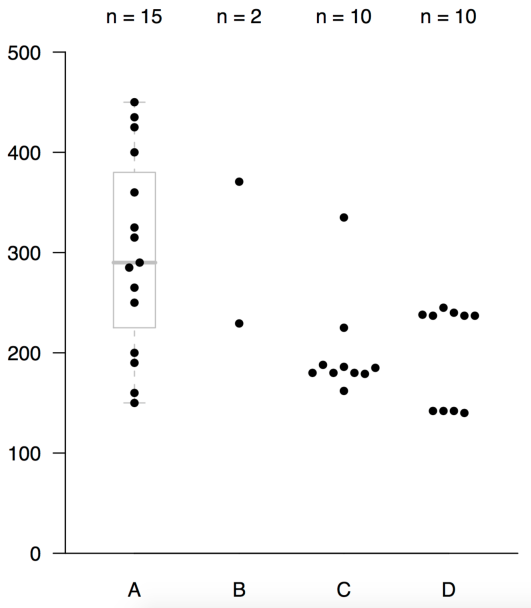
QQ-plot pre konkrétnu dátovú vzorku



Zakladné princípy vizualizácie dát

- Sumárne štatistiky zavádzajú
- Je užitočné zobrazit' surové dáta.
- Dátový bod jedna machuľa atramentu
- d'alšie machule pridávame len ak nám pomôžu pochopiť niečo nové
- Moment prekvapenia
- Graf by mal byť samovysvetľujúci,
- Farby lepšie vyniknú na tmavšom podklade ako na čisto bielom.
- Farby sa zväčša používajú na vysvetľovanú premennú, nie prediktory.
- Individuálnym výstupom rozumieme na základe porovnávania podobných objektov.





Príklad

Námorníci potrebujú vedieť kde sú.

V 18.storočí je celkovo nedostatok GPS satelitov

Thomas Mayer chce určiť polohu lode podľa polohy krátera na mesiaci.

$$\text{arc} = \beta + \alpha \sin \text{ang} + \gamma \cos \text{ang},$$

- β, α, γ - neznáme parametre
- arc , $\sin \text{ang}$ a $\cos \text{ang}$ sú pozorované veličiny

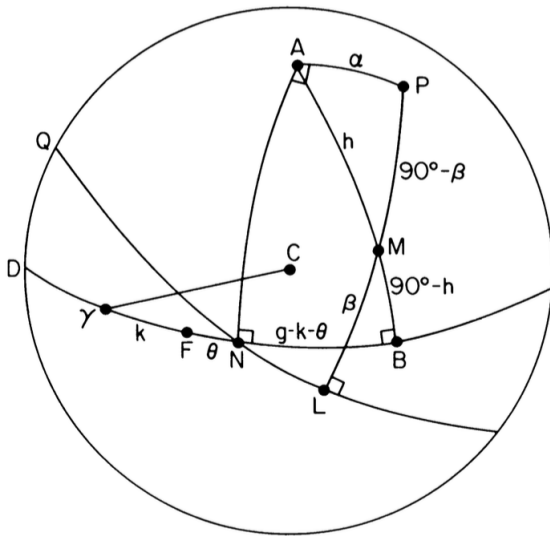


Figure 1.2. The moon. M , The crater Manilius; NL , the moon's equator; P , the moon's equatorial pole; NB , the circumference parallel to the plane of the ecliptic; A , the pole of NB ; $C\gamma$, the direction of the equinox from the moon's center C ; F , the node of the moon's orbit and the plane of the ecliptic. See text for more details.

27 rovnic o 3 neznámých

Eq. no.	Equation	Group
1	$\beta - 13^{\circ}10' = +0.8836\alpha - 0.4682\alpha \sin \theta$	I
2	$\beta - 13^{\circ}8' = +0.9996\alpha - 0.0282\alpha \sin \theta$	I
3	$\beta - 13^{\circ}12' = +0.9899\alpha + 0.1421\alpha \sin \theta$	I
4	$\beta - 14^{\circ}15' = +0.2221\alpha + 0.9750\alpha \sin \theta$	III
5	$\beta - 14^{\circ}42' = +0.0006\alpha + 1.0000\alpha \sin \theta$	III
6	$\beta - 13^{\circ}1' = +0.9308\alpha - 0.3654\alpha \sin \theta$	I
7	$\beta - 14^{\circ}31' = +0.0602\alpha + 0.9982\alpha \sin \theta$	III
8	$\beta - 14^{\circ}57' = -0.1570\alpha + 0.9876\alpha \sin \theta$	II
9	$\beta - 13^{\circ}5' = +0.9097\alpha - 0.4152\alpha \sin \theta$	I
10	$\beta - 13^{\circ}2' = +1.0000\alpha + 0.0055\alpha \sin \theta$	I
11	$\beta - 13^{\circ}12' = +0.9689\alpha + 0.2476\alpha \sin \theta$	I
12	$\beta - 13^{\circ}11' = +0.8878\alpha + 0.4602\alpha \sin \theta$	I
13	$\beta - 13^{\circ}34' = +0.7549\alpha + 0.6558\alpha \sin \theta$	III
14	$\beta - 13^{\circ}53' = +0.5755\alpha + 0.8178\alpha \sin \theta$	III
15	$\beta - 13^{\circ}58' = +0.3608\alpha + 0.9326\alpha \sin \theta$	III
16	$\beta - 14^{\circ}14' = +0.1302\alpha + 0.9915\alpha \sin \theta$	III
17	$\beta - 14^{\circ}56' = -0.1068\alpha + 0.9943\alpha \sin \theta$	III
18	$\beta - 14^{\circ}47' = -0.3363\alpha + 0.9418\alpha \sin \theta$	II
19	$\beta - 15^{\circ}56' = -0.8560\alpha + 0.5170\alpha \sin \theta$	II
20	$\beta - 13^{\circ}29' = +0.8002\alpha + 0.5997\alpha \sin \theta$	III
21	$\beta - 15^{\circ}55' = -0.9952\alpha - 0.0982\alpha \sin \theta$	II
22	$\beta - 15^{\circ}39' = -0.8409\alpha + 0.5412\alpha \sin \theta$	II
23	$\beta - 16^{\circ}9' = -0.9429\alpha + 0.3330\alpha \sin \theta$	II
24	$\beta - 16^{\circ}22' = -0.9768\alpha + 0.2141\alpha \sin \theta$	II
25	$\beta - 15^{\circ}38' = -0.6262\alpha - 0.7797\alpha \sin \theta$	II
26	$\beta - 14^{\circ}54' = -0.4091\alpha - 0.9125\alpha \sin \theta$	II
27	$\beta - 13^{\circ}7' = +0.9284\alpha - 0.3716\alpha \sin \theta$	I

27 rovníc o 3 neznámých

Rozdelil si 27 pozorovaní do 3 rovnako veľkých skupín podľa toho, ako veľmi rôzne boli hodnoty $\sin \alpha$.

V každej skupine sčítal všetky pozorovania a potom riešil sústavu 3 lineárnych rovníc o 3 neznámých.

Legendre 1805 - metóda najmenších štvorcov.

Namiesto arbitrárneho zgrupovania pozorovaní do rôznych skupín Legendre explicitne zaviedol chybový člen:

$$\text{arc}_i = \beta + \alpha \sin \text{ang}_i + \gamma \cos \text{ang}_i + \epsilon_i, \quad \text{pre } i = 1, \dots, 27$$

a navrhol zvoliť parametre β, α, γ , tak aby boli minimalizované štvorce odchýliek $\sum_{i=1}^{27} \epsilon_i^2$.

To prečo je toto dobrý nápad sa dozvieme neskôr.

Prečo sa regresia volá regresia

Vysokí rodičia majú vysoké deti, nízki rodičia majú nízke deti.

Prečo sa regresia volá regresia

Vysokí rodičia majú vysoké deti, nízki rodičia majú nízke deti.

V priemere.

Prečo sa regresia volá regresia

Vysokí rodičia majú vysoké deti, nízki rodičia majú nízke deti.

V priemere.

Otázkou teraz je, či najvyšší rodičia budú mať najvyššie deti a najnižší rodičia budú mať najnižšie deti.

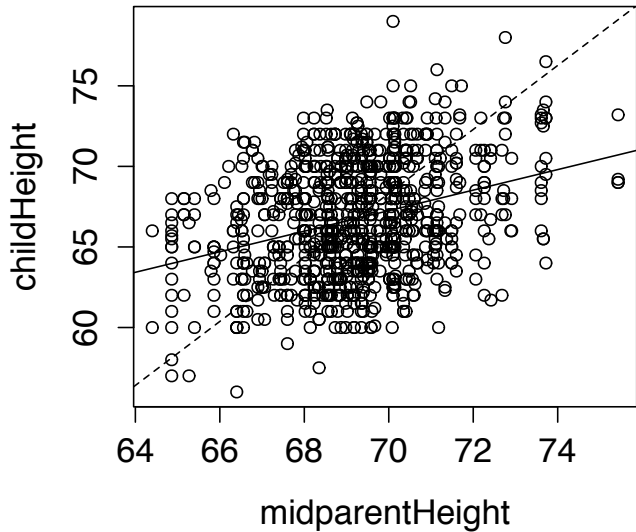
Prečo sa regresia volá regresia

Vysokí rodičia majú vysoké deti, nízki rodičia majú nízke deti.

V priemere.

Otázkou teraz je, či najvyšší rodičia budú mať najvyššie deti a najnižší rodičia budú mať najnižšie deti.

Odpoveďou je, že nie tak celkom.



Prečo?

Možno by sme si boli mysleli, že rodičia, ktorí sú o jednu štandardnú odchýlku vyšší ako priemer rodičov majú aj dieťa, ktoré je o jednu štandardnú odchýlku vyššie ako priemerné dieťa.

Toto by bolo len ak by $r = \text{cov}(x, y) = 1$

$$\frac{y - \bar{y}}{SD_y} = r \frac{x - \bar{x}}{SD_x}.$$

Avšak existujú aj iné faktory, ktoré vplývajú na výšku dieťaťa.

Tento efekt je na veľa miestach

- športovec, ktorému sa extrémne darilo má ďalšiu sezónu len výbornú
- firmy (v konkurenčnom prostredí), ktoré zaznamenali obrovský zisk tak v ďalšom roku len veľký zisk
- hodnotiť školy len podľa toho ako veľmi sa zlepšili nie je dobrý nápad
- po zavedení rýchlostných kamier preceníme ich efekt
- ...

Výnosnosť firiem v konkurenčnom prostredí je v priemere konštantná.

Teda ak sa firme darí veľmi dobre, môžeme očakávať, že v budúcnosti sa jej môže dariť o kus bližšie k priemeru.

Keď na toto poukázal Secrist, tak Hotelling na to reagoval slovami:

“proving the multiplication table by arranging elephants in rows and columns, and then doing the same for numerous other kinds of animals”

Ďakujem za pozornosť.