

Adatorientált programozás tanítása kezdőknek

Koren Miklós

CEU Department of Economics and Business

@korenmiklos

Scratch @ Budapest

Magamról

Első hasznos programom

2.

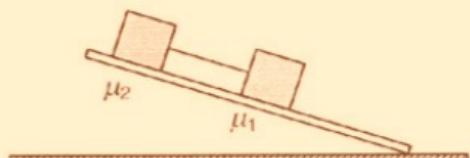
Termoszban lévő 16°C hőmérsékletű vízbe 100°C -ra melegített fémdarabot helyezünk. $36,5^{\circ}\text{C}$ közös hőmérséklet alakul ki.

Mekkora lesz a közös hőmérséklet, ha még további két, az elsővel megegyező tömegű és hőmérsékletű fémdarabot helyezünk a termoszba?

(Kopcsa József)

3.

Vízszintes helyzetű deszkán lévő, két $m = 2\text{ kg}$ tömegű testet elhanyagolható tömegű feszес fonál köt össze. A testek és a deszka közötti tapadási súrlódási együttható különböző, $\mu_1 = 0,2$ és $\mu_2 = 0,5$. A deszka egyik végét lassan emelni kezdjük. Határozzuk meg a fonálerőt e testek közös megcsúszásának határhelyzetében, a megcsúszást megelőző pillanatban!



(Kotek László)

4.

Egy szánkó össztömege 40 kg , a csúszási súrlódási tényezője a hóban $0,08$. A szánkót vízszintes talajon, álló helyzetből indulva 4 másodperc alatt egy állandó nagyságú, vízszintes irányú erővel

Első haszontalan programom



Első beruházásom az adatforradalomba



Programozás oktatása

Két megközelítés

A programozás egyszerű

„Menj előre 10 lépést!”

A programozás bonyolult

„Szimuláld a bayesi poszterior eloszlást GPU-n!”

Hogyan lesz valakiből adattudós?

- 1. Tanuld meg a matekot!**
- 2. Tanuld meg a statisztikát!**
- 3. Fejleszd a geometriai intuiciód!**
- 4. Tanulj meg kezelní egy statisztikai programcsomagot!**
- 5. Ha túl bonyolult dolgot akarsz csinálni, programozd le!**

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}.$$

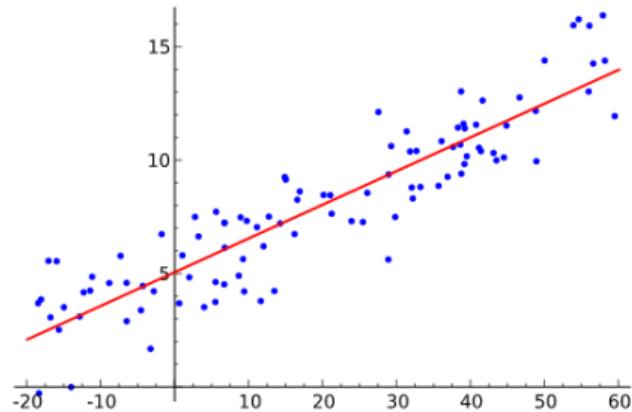
Hogyan lesz valakiből adattudós?

1. Tanuld meg a matekot!
2. **Tanuld meg a statisztikát!**
3. Fejleszd a geometriai intuiciód!
4. Tanulj meg kezelní egy statisztikai programcsomagot!
5. Ha túl bonyolult dolgot akarsz csinálni, programozd le!

$$\hat{\beta} = (X'X)^{-1} X'Y$$

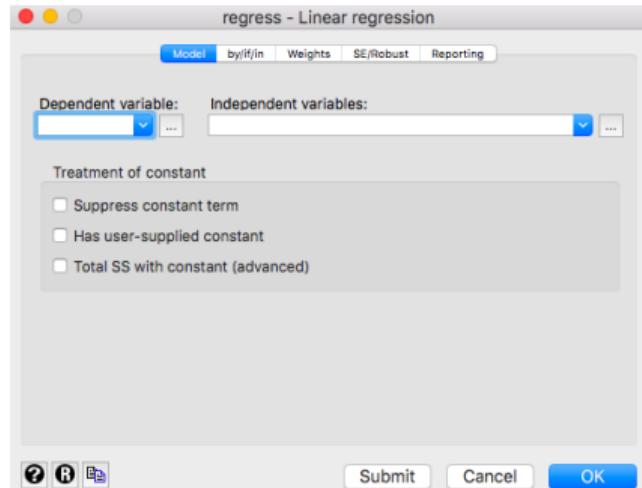
Hogyan lesz valakiből adattudós?

1. Tanuld meg a matekot!
2. Tanuld meg a statisztikát!
3. **Fejleszd a geometriai intuiciód!**
4. Tanulj meg kezelní egy statisztikai programcsomagot!
5. Ha túl bonyolult dolgot akarsz csinálni, programozd le!



Hogyan lesz valakiből adattudós?

1. Tanuld meg a mateket!
2. Tanuld meg a statisztikát!
3. Fejleszd a geometriai intuiciód!
4. **Tanulj meg kezelní egy statisztikai programcsomagot!**
5. Ha túl bonyolult dolgot akarsz csinálni, programozd le!

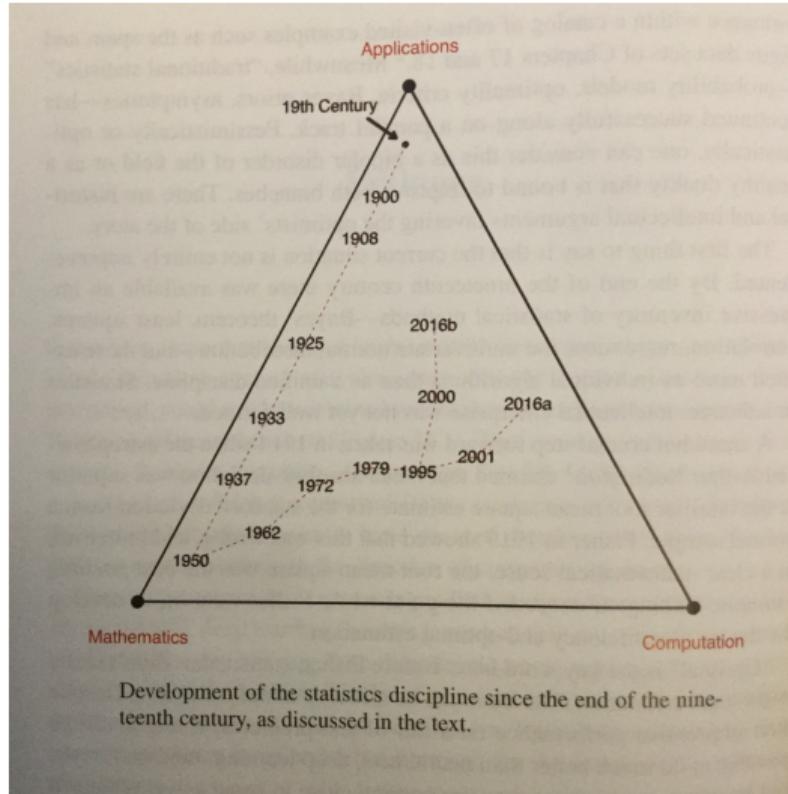


Hogyan lesz valakiből adattudós?

1. Tanuld meg a mateket!
2. Tanuld meg a statisztikát!
3. Fejleszd a geometriai intuiciód!
4. Tanulj meg kezelní egy statisztikai programcsomagot!
5. **Ha túl bonyolult dolgot akarsz csinálni, programozd le!**

```
program weib2
    version 13
    args todo b lnf g H           // H added
    tempvar ti t2
    mleval `t1' = `b', eq(1)
    mleval `t2' = `b', eq(2)
    local t "$ML_y1"
    local d "$ML_y2"
    tempvar p M R
    quietly gen double `p' = exp(`t2')
    quietly gen double `M' = (`t' * exp(-`t1')) ^ `p'
    quietly gen double `R' = ln(`t') - `t1'
    mlsum `lnf' = -`M' + `d' * (`t2' - `t1' + (`p' - 1) * `R')
    if (`todo' == 0 | `lnf' >= .) exit
    tempname d1 d2
    mlvecsnum `lnf' `d1' = `p' * (`M' - `d'), eq(1)
    mlvecsnum `lnf' `d2' = `d' - `R' * `p' * (`M' - `d'), eq(2)
    matrix `g' = (`d1', `d2')
    if (`todo' == 1 | `lnf' >= .) exit          // new from here down
    tempname d11 d12 d22
    mlmatsum `lnf' `d11' = -`p'^2 * `M', eq(1)
    mlmatsum `lnf' `d12' = `p' * (`M' - `d' + `R' * `p' * `M'), eq(1,2)
    mlmatsum `lnf' `d22' = -`p' * `R' * (`R' * `p' * `M' + `M' - `d'), eq(2)
    matrix `H' = (`d11', `d12' \ `d22', `d22')
end
```

A statisztika fejlődése (Efron és Hastie, 2016)



Pedig a programozás egyszerű

Pedig a programozás egyszerű

1. Programozás számok nélkül.
2. Világos célok.
3. Azonnali visszacsatolás.

Programozás számok nélkül

- ▶ Az „adatmesterség” (data carpentry) az adatokkal való bánás mestersége.
- ▶ Az „adattudomány” (data science) előszobája.

Világos célok

- ▶ web scraping
- ▶ adattisztítás

Web scraping

1. Felderítés
2. Letöltés
3. Adatkinyerés
4. Adatmentés

Egy egyszerű statikus honlap

```
In [18]: KOZBESZ_URL = 'http://ceumicrodata.github.io/regi.kozbeszerzes.hu/static/Kearchiv/index.html'
```

```
In [12]: response = requests.get(KOZBESZ_URL)
# en ezt python2-ben irom, te tegyel zarojelet!
print response.text
```

```
<html><head><title>Elérzések számkönyök</title><meta http-equiv="Content-Type" content="text/html; charset=iso-8859-2"></head><body text=black link=black alink=blue vlink=black><h2><b><center>KOZBESZERZÉSI ÖRTÉNETTÖRTÉNÉK</b><br>A KÖZBESZERZÉSEK TANCSÁNAK HIVATALOS LAPJA</h2><font size=3><br>
<br><a href="04087/index.html">X. Övfüolyam 87. szám - 2004. július 30.</a>
<br><a href="04086/index.html">X. Övfüolyam 86. szám - 2004. július 28.</a>
<br><a href="04085/index.html">X. Övfüolyam 85. szám - 2004. július 26.</a>
<br><a href="04084/index.html">X. Övfüolyam 84. szám - 2004. július 23.</a>
<br><a href="04083/index.html">X. Övfüolyam 83. szám - 2004. július 21.</a>
<br><a href="04082/index.html">X. Övfüolyam 82. szám - 2004. július 19.</a>
<br><a href="04081/index.html">X. Övfüolyam 81. szám - 2004. július 16.</a>
```

Találjuk meg benne a struktúrát!

```
In [10]: # helper function
def all_links(html_page):
    """
    This is some useful HTML magic. For more see http://lxml.de/tutorial.html
    """
    parsed = etree.HTML(html_page)
    return [html_element.attrib['href'] for html_element in parsed.xpath("//a")]

# testing
print all_links(response.text)

['04087/index.html', '04086/index.html', '04085/index.html', '04084/index.html', '04083/index.htm
l', '04082/index.html', '04081/index.html', '04080/index.html', '04079/index.html', '04078/index.h
tml', '04077/index.html', '04076/index.html', '04075/index.html', '04074/index.html', '04073/inde
x.html', '04072/index.html', '04071/index.html', '04070/index.html', '04069/index.html', '04068/in
dex.html', '04067/index.html', '04066/index.html', '04065/index.html', '04064/index.html', '04063/
index.html', '04062/index.html', '04061/index.html', '04060/index.html', '04059/index.html', '0405
8/index.html', '04057/index.html', '04056/index.html', '04055/index.html', '04054/index.html', '04
053/index.html', '04052/index.html', '04051/index.html', '04050/index.html', '04049/index.html',
```

Adattisztítási példa: Címfeldolgozás

- ▶ 2,5 millió címhely (1,7 millió székhely, 900 ezer telephely)
- ▶ Egy cím anatómiája: 1075 Budapest, Károly körút 9.

Hibás irányítószám

1052 Budapest, Kossuth Lajos tér 3.

Részleges utcanév

1055 Budapest, **Kossuth** tér 3.

Kétértelműség

2700 Cegléd, Kossuth **Lajos** utca 5.

Kétértelműség

2700 Cegléd, Kossuth **Lajos** utca 5.

2700 Cegléd, Kossuth **Ferenc** utca 5.

Elírás

1151 Budapest, **Kosut** utca 7.

Feldolgozhatatlan mezők

6600 Szentes, Ipari park hrsz. 3967/3.

Feldolgozhatatlan mezők

6600 Szentes, Ipari park hrsz. 3967/3.

1093 Budapest, (Pólus Irodaház), Lónyai utca 15.

Interaktív adattisztító alkalmazások

Trifacta

Sample 1 - First 500KB 11 Columns 2,980 Rows 2 Data Types Grid Columns: ✓ All Rows: ✓ All Transformed - 2 Columns Transformed - 2,979 Rows Filter in grid

	Source	Preview										
ABC	ID	ABC	column3	ABC	column4	ABC	column1	ABC	column5	ABC	column6	ABC
2,839 Categories	258 Categories	101 Categories	1 Category	2,475 Categories	2,475 Categories	2,475 Categories	2,959 Categories					
1 customer_id	first_name	last_name		SSN	credit_card	address						
2 "4abe6b808c96e647239677f2a9f247fd'	Julian	"Russell"	"	"451-59-0366"	"4516009576471550"	"2166-Cedar-Lane						
3 "d8983c31f0c1031ca2837f42852fbf24'	Nathan	"Davis"	"	"308-61-6226"	"4407614812304060"	"7475-Madison-St						
4 "f8ebc49d5c03b7e2f934019fc10e9d4'	Elijah	"Wright"	"	"593-19-3579"	"5584472636741872"	"48533-2nd-Street						
5 "7d9c5c49ad12e8233c558dd88ed3c143'	Cole	"Thomas"	"	"177-74-6463"	"4257017589440200"	"2366-Linden-Str						
6 "334ae2126c83dffbe28bd8a13d4ae50b'	Andrew	"Green"	"	"557-30-0305"	"5477064333168580"	"4252-10th-ST,-E						
7 "af67b3a6f43ff02dfedb81ee94cd0bf5'	Adam	"Howard"	"	"076-69-2166"	"5409820014340117"	"278-Orange-Street						
8 "12f35d87b7c6e54aec5593e4c19b9824'	Andrew	"Price"	"	"457-96-9416"	"4087559818775316"	"953-Prospect-ST						
9 "c98623c793c911e68e9c4b7502429983'	Erin	"Barnes"	"	""	""	"370-Dogwood-Drive						
10 "4435b2c7c3712c154bc9d76427ba72f'	Daniel	"Perez"	"	"329-36-9209"	"5290545373364620"	"6268-Fairway-Dr						
11 "30a359c8d57c68de61eb5be6128d8d37'	Blake	"Bell"	"	"071-17-4141"	"4477504255299526"	"7709-Holly-Drive						

SUGGESTIONS

Extract on: `^`

ABC	column4	ABC	column1
last_name			
"Russell"	"		
"Davis"	"		
Affects 1 column, 2979 rows	Creates 1 column		

Countpattern on: `^`

ABC	column4	#	column1
last_name	0		
"Russell"	2		
"Davis"	2		
Affects 1 column, 2979 rows	Creates 1 column		

Extractlist on: `'{any}+` delimiter: `^`

ABC	column4	ABC	column1
last_name	["last_name"]		
"Russell"	["", "Russell", ""]		
"Davis"	["", "Davis", ""]		
Affects 1 column, all rows	Creates 1 column		

Review Pairs
90.2K pairs (from 163.3K total registrations) [Setup how pairs are found](#)

[Filter](#) Duplicates Uniques Remove Responses

Responses	address	customerID	email	firstName
You	375 Dwight Ter, Hampton, VA 23668	1017970	ccoxph@who.int	Kathleen
Tamr	375 Dwight Terrace, Hampton, VA 23668	289929		Kitty
	305 Glendale Trl, Los Angeles, CA 90030	1003298	jgraham6t@bigca	Jac
	305 Glendale Trail, Los Angeles, CA 90030	275257		Johnny
	93 Barnett Plaza, Fresno, CA 93786	1021662		Ben
		293621	bevansgx@cdc.go	Benjamin
		1068275		Karen
		340234	kjones66@wooth	
	10 Ruskin Park, Lansing, MI 48919	3419		Tina

OpenRefine

11285 rows

Show as: rows records Show: 5 10 25 50 rows Extensions: Zemanta ▾ Freebase ▾ RDF ▾ CK

« first < previous 1 - 50 next >

		<input checked="" type="checkbox"/> Capital or Reven	Directorate	Transaction Num	Date	Service Area	Expenses Type	Amount	Supp
1.	Revenue	Community Wellbeing & Social Care	5105695746	05.04.2013	Youth & Community	Operational Equipment		120	REDACTE PERSON/
2.	Revenue	Community Wellbeing & Social Care	5105695746	05.04.2013	Youth & Community	Operational Equipment		80	REDACTE PERSON/
3.	Revenue	Community Wellbeing & Social Care	5105698650	24.04.2013	Leaseholds by LA	Accommodation Costs - Leaseholder Payments	edit	695.89	REDACTE PERSON/
4.	Revenue	Community Wellbeing & Social Care	5105698650	24.04.2013	Leaseholds by LA	Accommodation Costs - Leaseholder Payments		695.89	REDACTE PERSON/
5.	Revenue	Community Wellbeing & Social Care	5105698650	24.04.2013	Leaseholds by LA	Accommodation Costs - Leaseholder Payments		695.89	REDACTE PERSON/
6.	Revenue	Community Wellbeing & Social Care	5105698650	24.04.2013	Leaseholds by LA	Accommodation Costs - Leaseholder Payments		695.89	REDACTE PERSON/
7.	Revenue	Community Wellbeing & Social Care	5105698650	24.04.2013	Leaseholds by LA	Accommodation Costs - Leaseholder Payments		695.89	REDACTE PERSON/
8.	Revenue	Chief Executive, Schools & Learning	5105698316	19.04.2013	L&A Commissioned Activity	Bought in Prof Services - Curriculum (Schools)		250	REDACTE PERSON/
9.	Revenue	Chief Executive, Schools & Learning	5105698318	19.04.2013	L&A Commissioned Activity	Bought in Prof Services - Curriculum (Schools)		710	REDACTE PERSON/
10.	Revenue	Economy & Environment	5105695879	05.04.2013	IW Biological Record	General Materials		220.2	REDACTE PERSON/
11.	Revenue	Chief Executive, Schools & Learning	5105696514	12.04.2013	Adult Services Training	Training and Conferences		150	REDACTE PERSON/
12.	Revenue	Community Wellbeing & Social Care	5105695832	10.04.2013	Short Breaks	Payments to Voluntary and Other Associations		1,260.00	REDACTE PERSON/
13.	Capital	Resources	5105696504	12.04.2013	Capital Receipts	External Design and Supervision Fees		400	REDACTE PERSON/
14.	Capital	Resources	5105696505	12.04.2013	Capital Receipts	External Design and Supervision Fees		1,350.00	REDACTE PERSON/
15.	Revenue	Economy & Environment	5105696707	12.04.2013	Schools Reorganisation	Security of Buildings		300	REDACTE PERSON/
16.	Revenue	Economy & Environment	5105696707	12.04.2013	Schools Reorganisation	Security of Buildings		300	REDACTE PERSON/

Azonnali visszacsatolás

- ▶ interaktív IDE
- ▶ csoportmunka, pair programming
- ▶ játékos feladatok

Együtt programozni jó



Sikerélmény



Összefoglaló

Összefoglaló

Programozni jó!

- ▶ A programozásnak legyen kézzel fogható célja!
- ▶ Kezdjük minél korábban!
- ▶ Programozzunk akár számok nélkül!

Adatozás a CEU-n

- ▶ MSc in Business Analytics
- ▶ Data Analysis for Business and Policy
- ▶ Data @ CEU

Kapcsolat

- ▶ economics.ceu.edu
- ▶ twitter.com/korenmiklos