„I do see a place where, ultimately, instead of using data only to become more efficient, we will all use data to become more humane.”
Giorgia Lupi

Néhány érdekesség R-rel

Virág Katalin
virag.katalin@med.u-szeged.hu
SzTE Orvosi Fizikai és Orvosi Informatikai Intézet
2019.11.19.

Ábrák fontossága

“Azonos” egyszerű lineáris regressziók

Az *anscombe* adatbázis 4 db *X-Y* adatpárt tartalmaz, melyek statisztikai jellemzői azonosak.

X változók jellemzői: átlagok és szórások

```
library(tidyverse) # tidyverse csomag betöltése

data(anscombe) # anscombe adatbázis betöltése
anscombe %>%
  select(starts_with("x")) %>%
  summarise_all(list(atlag = mean,
                     szoras = sd)) %>%
  kable(digits = 2, align = "c", caption = "*X* változók
jellemzői")
```

x1_atlag	x2_atlag	x3_atlag	x4_atlag	x1_szoras	x2_szoras	x3_szoras	x4_szoras
9	9	9	9	3.32	3.32	3.32	3.32

X változók jellemzői

Y változók jellemzői: átlagok és szórások

```
anscombe %>%
  select(starts_with("y")) %>%
  summarise_all(list(atlag = mean,
                     szoras = sd)) %>%
  kable(digits = 2, align = "c", caption = "*Y* változók
jellemzői")
```

y1_atlag	y2_atlag	y3_atlag	y4_atlag	y1_szoras	y2_szoras	y3_szoras	y4_szoras
7.5	7.5	7.5	7.5	2.03	2.03	2.03	2.03

Y változók jellemzői

Az *X* és *Y* változók közötti korrelációk

```
anscombe %>%
  summarise(x1_y1 = cor(x1, y1),
            x2_y2 = cor(x2, y2),
            x3_y3 = cor(x3, y3),
            x4_y4 = cor(x4, y4)) %>%
  kable(digits = 2, align = "c", caption = "Az *X* és *Y* változók
közötti korrelációk")
```

x1_y1	x2_y2	x3_y3	x4_y4
0.82	0.82	0.82	0.82

Az *X* és *Y* változók közötti korrelációk

Egyszerű lineáris regressziók

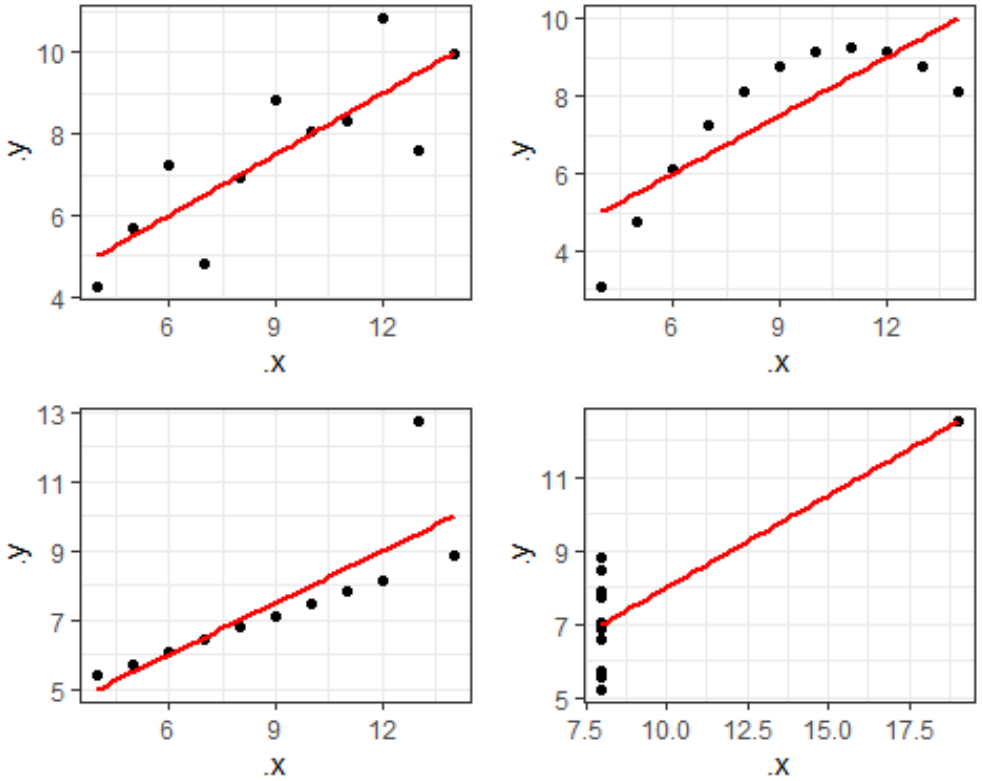
```
x <- anscombe %>%
  select(starts_with("x"))
y <- anscombe %>%
  select(starts_with("y"))
map2(x, y, ~ lm(.y ~ .x)) %>% # Lineáris regressziók
  map_dfr(~ broom::tidy(., conf.int = T), .id = "Magyarázó
változó") %>%
  kable(digits = 2, align = "c", caption = "Egyszerű lineáris
regressziók eredménye")
```

Magyarázó változó	term	estimate	std.error	statistic	p.value	conf.low	conf.high
x1	(Intercept)	3.0	1.12	2.67	0.03	0.46	5.54
x1	.x	0.5	0.12	4.24	0.00	0.23	0.77
x2	(Intercept)	3.0	1.13	2.67	0.03	0.46	5.55
x2	.x	0.5	0.12	4.24	0.00	0.23	0.77
x3	(Intercept)	3.0	1.12	2.67	0.03	0.46	5.55
x3	.x	0.5	0.12	4.24	0.00	0.23	0.77
x4	(Intercept)	3.0	1.12	2.67	0.03	0.46	5.54
x4	.x	0.5	0.12	4.24	0.00	0.23	0.77

Egyszerű lineáris regressziók eredménye

Ábrák, amikkel kezdeni kellett volna az elemzést

```
abrak <- map2(x, y, ~ ggplot(anscombe, aes(x = .x, y = .y)) +
                      geom_point() +
                      geom_smooth(method = lm, se = FALSE, color =
"red") +
                      theme_bw())
grid.arrange(abrak[[1]], abrak[[2]], abrak[[3]], abrak[[4]])
```


10
8
6
4
.y
6
9
12
.x
10
8
6
4
.y
6
9
12
.x
13
11
9
7
5
.y
6
9
12
.x
11
9
7
5
.y
7.5
10.0
12.5
15.0
17.5
.x

Pontábrák

Regressziós diagnosztika fontossága

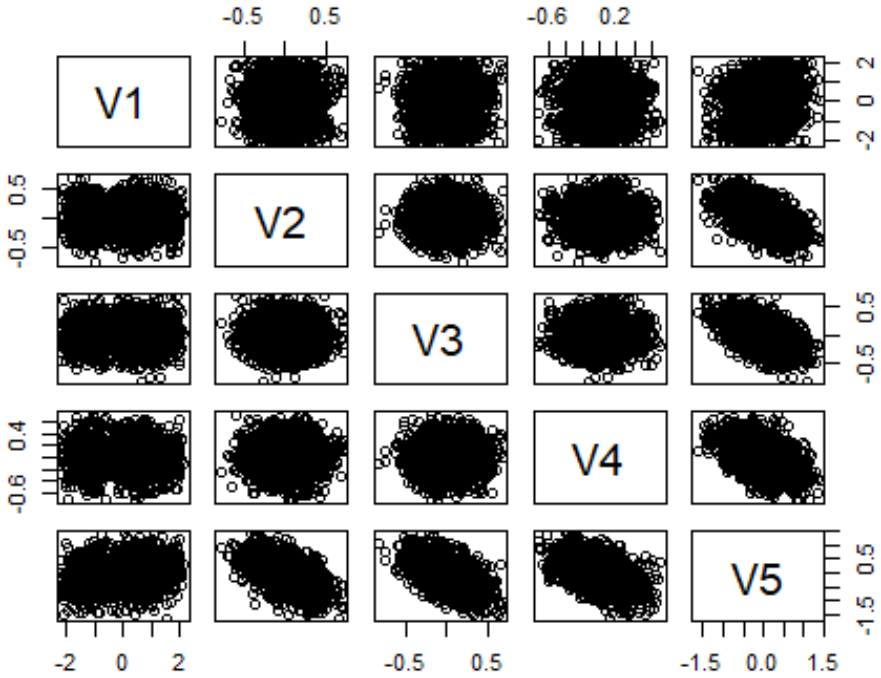
Többszörös lineáris regresszió

Vizsgáljuk meg, hogy a *V1* változó hogyan függ a többi 4 változótól.

```
dat <-
read.table('http://www4.stat.ncsu.edu/~stefanski/NSF_Supported/Hid
den_Images/orly_owl_files/orly_owl_Lin_4p_5_flat.txt', header =
FALSE) # Adatbázis beolvasása
```

Változók páronkénti ábrázolása

```r
pairs(dat)
```



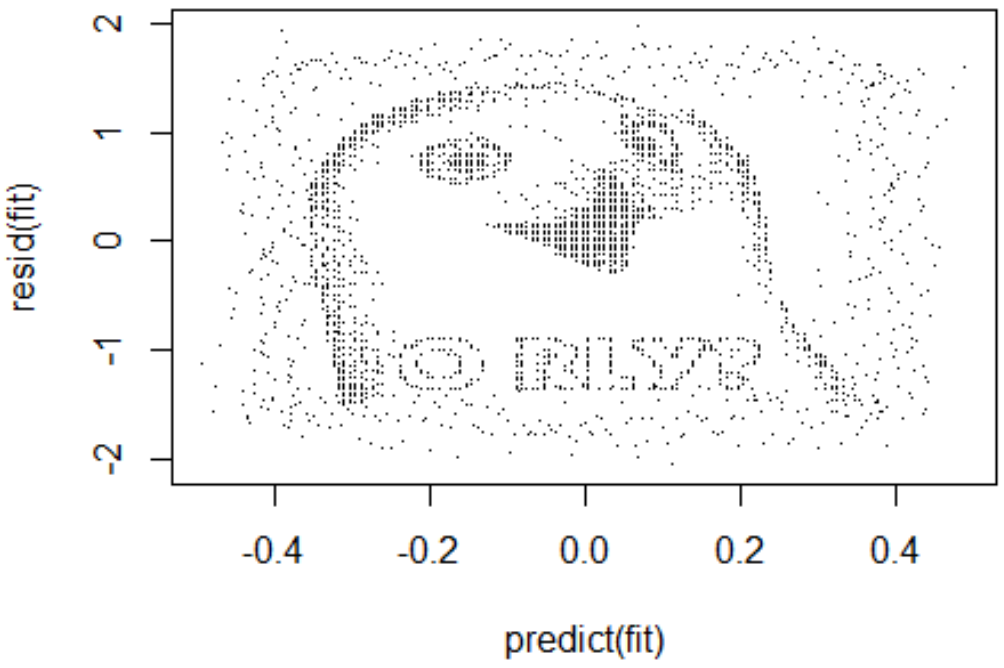
Többszörös lineáris regresszió

```
fit <- lm(V1 ~ . - 1, data = dat)
summary(fit)$coef # Regresszió eredménye

##      Estimate Std. Error   t value     Pr(>|t|)

## V2 0.9856157 0.12798121  7.701253 1.989126e-14

## V3 0.9714707 0.12663829  7.671225 2.500259e-14

## V4 0.8606368 0.11958267  7.197003 8.301184e-13

## V5 0.9266981 0.08328434 11.126919 4.778110e-28
```

A hibatagok ábrázolása a becsült értékek függvényében

```
plot(predict(fit), resid(fit), pch = '.')
```

Tényleg?



A hibatagok ábrázolása a becsült értékek függvényében

A Pearson-féle korrelációs együttható

A Pearson-féle korrelációs együttható mintabeli becslése

```r
library(datasauRus) # Csomag betöltése
dino <- datasaurus_dozen[datasaurus_dozen$dataset == "dino", 2:3]
# Adatbázis beolvasása
cor.test(dino$x, dino$y) # Pearson-féle korrelációs együtthatóra
vonatkozó hipotézisvizsgálat
```

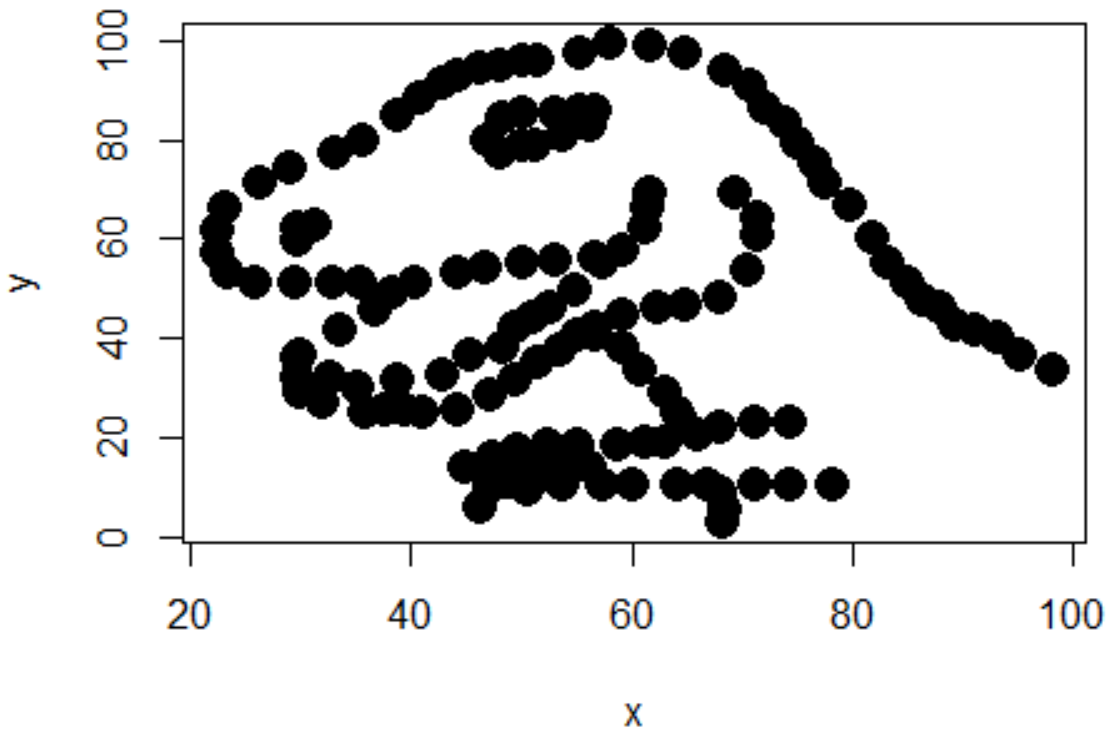
A Pearson-féle korrelációs együttható

```
##
##  Pearson's product-moment correlation
##
## data:  dino$x and dino$y
## t = -0.76443, df = 140, p-value = 0.4459
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.2267905  0.1013316
## sample estimates:
##         cor
## -0.06447185
```

Pontábra

A Pearson-féle korrelációs együttható becslése -0.06, pedig egy egész dinoszaurusz bújik meg az adatok között!

```
plot(dino, pch = 16, cex = 2) # Adatok ábrázolása
```



Dino

Simpson-paradoxon

UCBAdmissions adatbázis

A *UCBAdmissions* adatbázis a Berkeley egyetem 6 legnagyobb intézetére vonatkozó felvételi adatokat tartalmaz nemek szerint csoportosítva, 1973-ból.

```
d <- data.frame(UCBAdmissions) # Adatbázis betöltése
# ?UCBAdmissions # Adatbázis leírása
```

Felvételi eredmények

Felvettek száma nemek szerint:

```
(gyak.tablazat <- with(d, xtabs(Freq ~ Admit + Gender)))

##            Gender

## Admit      Male Female

##   Admitted 1198    557

##   Rejected 1493   1278
```

Felvételi eredmények

Felvettek aránya nemek szerint:

```
prop.table(gyak.tablazat, margin = 1)

##            Gender
## Admit            Male    Female
##   Admitted 0.6826211 0.3173789
##   Rejected 0.5387947 0.4612053
```

Felvételi eredmények

Úgy tűnik, hogy kétszer annyi fiút vettek fel, mint lányt.

Valóban?

Nézzük meg intézetenként:

```
d %>%
  group_by(Dept, Gender) %>%
  mutate(Prop = Freq / sum(Freq)) %>%
  filter(Admit == "Admitted") %>%
  kable(digits = 2, align = "c", caption = "Felvételi arányok
nemenként és intézetenként")
```

Admit	Gender	Dept	Freq	Prop
Admitted	Male	A	512	0.62
Admitted	Female	A	89	0.82
Admitted	Male	B	353	0.63
Admitted	Female	B	17	0.68
Admitted	Male	C	120	0.37
Admitted	Female	C	202	0.34
Admitted	Male	D	138	0.33
Admitted	Female	D	131	0.35
Admitted	Male	E	53	0.28
Admitted	Female	E	94	0.24
Admitted	Male	F	22	0.06
Admitted	Female	F	24	0.07

Felvételi arányok nemenként és intézetenként

Felvételi eredmények

Intézetenként már eltűnik vagy megfordul a fiúk és a lányok közötti különbség.

Lehetséges ok

A lányok közül kevesebben jelentkeztek olyan helyekre, ahová könnyű bekerülni, mint például az *A* intézet, ahová a hallgatók 64%-át felvették:

```
d %>%
  group_by(Gender, Dept) %>%
  summarize(Total = sum(Freq)) %>%
  mutate(Prop = Total / sum(Total)) %>%
  arrange(Gender) %>%
  kable(digits = 2, align = "c", caption = "Jelentkezési arányok
nemenként és intézetenként")
```

Gender	Dept	Total	Prop
Male	A	825	0.31
Male	B	560	0.21
Male	C	325	0.12
Male	D	417	0.15
Male	E	191	0.07
Male	F	373	0.14
Female	A	108	0.06
Female	B	25	0.01
Female	C	593	0.32
Female	D	375	0.20
Female	E	393	0.21
Female	F	341	0.19

Jelentkezési arányok nemenként és intézetenként

„Maybe stories are just data with a soul.”
Brené Brown