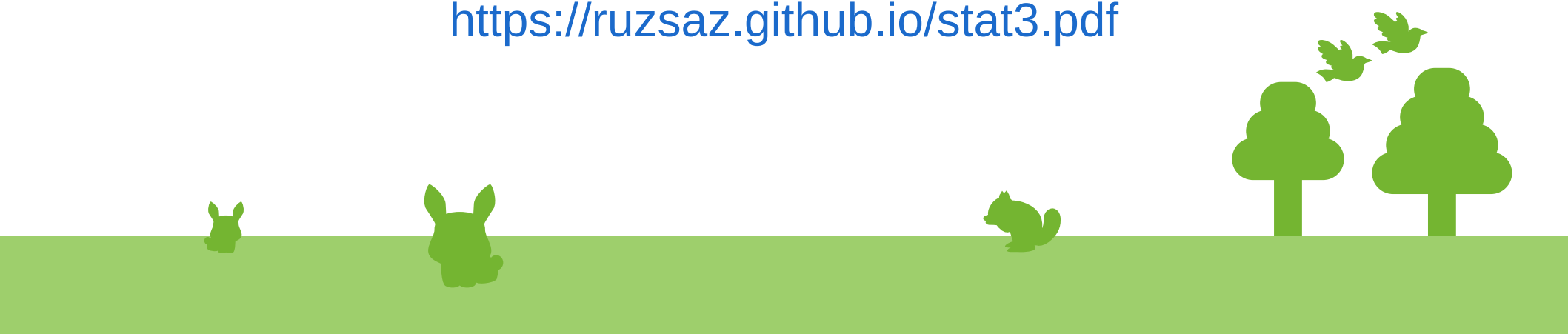
Kvantitatív kutatási módszertanok 3

Ruzsa Zoltán
ruzsa.zoltan@emk.semmelweis.hu

https://ruzsaz.github.io/stat3.pdf



Elméleti HF volt:

- Szükség van egyáltalán hipotézisvizsgálatra?

  Nem elég paraméterbecsléssel megbecsülni a két minta várható értékét, és megnézni, hogy a hibahatárnál jobban különböznek-e?

- Egy hipotézisvizsgálat feladat megoldására két alkalmazható próbát is találtunk. Elvégeztük mindkettőt.

  1) Ha azonos eredményt adtak, biztosabbak lehetünk az eredményben, mintha csak az egyiket végeztük volna el?

  2) Ha ellentétes eredményt kaptunk, akkor mi a helyzet?

1. 250g kenőcs

50 elemű minta, normális eloszlás, ismer szórás, 248.1 az átlag. Baj van a töltéssel?

2. kézmosás covid előtt/után

Előtte: 26/100,
Utána: 32/100.
Nőtt a hajlandóság?

3. sebfertőzés normál / laparoszkópos esetben, kontingencia táblázat

	-	+
N	37453	4421
L	19454	642

4. Vércukor csökkentő gyógyszer

30 fős minta, m = 6.8%, s = 0.5
99% szignifikancia mellett igaz, hogy az átlag 7% alatti?

5. érettségi eredmények normalitása

15-15 pontszám eloszlása tekinthető normálisnak?

6. hallgatói értékelés, 2 értékelési stílus összehasonlítása

A 15 fiú és 9 lány pontozása ugyanabból az eloszlásból vett minta?

7. működik a homeopátiás altató?

A 10 ember elalvási ideje:

	1	2	3	4	5	6	7	8	9	10
placebo	23.2	14	59.2	31.1	47.4	7.8	44.2	21.1	33.3	9.2
altató	31.5	7.8	54.9	33	11.1	23.3	43	16.7	12.6	8.6

8. jobb a Comirnaty?

	tünetm	enyhe	kórház
Sinopharm	87	44	14
Comirnaty	46	19	6

Tematika

- Leíró statisztika (átlag, medián, módusz, szórás...)
- Paraméterbecslés (mennyi?)
- Hipotézisvizsgálat (igaz vagy nem?)
- Regressziószámítás, modellezés (hogy befolyásol?)

Próbák elnevezése

- Student-féle T-próba
- Z-próba (magyarul u-próba)
- χ^2-próba (khi-négyzet próba)
- Wilcoxon rangpróba
- Illeszkedésvizsgálat
- Fisher-féle egzakt teszt
- ...


GUINNESS
GUINNESS


Chemist & Statistician
WILLIAM SEALY
GOSSET
1876-1937
Chief Brewer
Student 't' test

T-próba

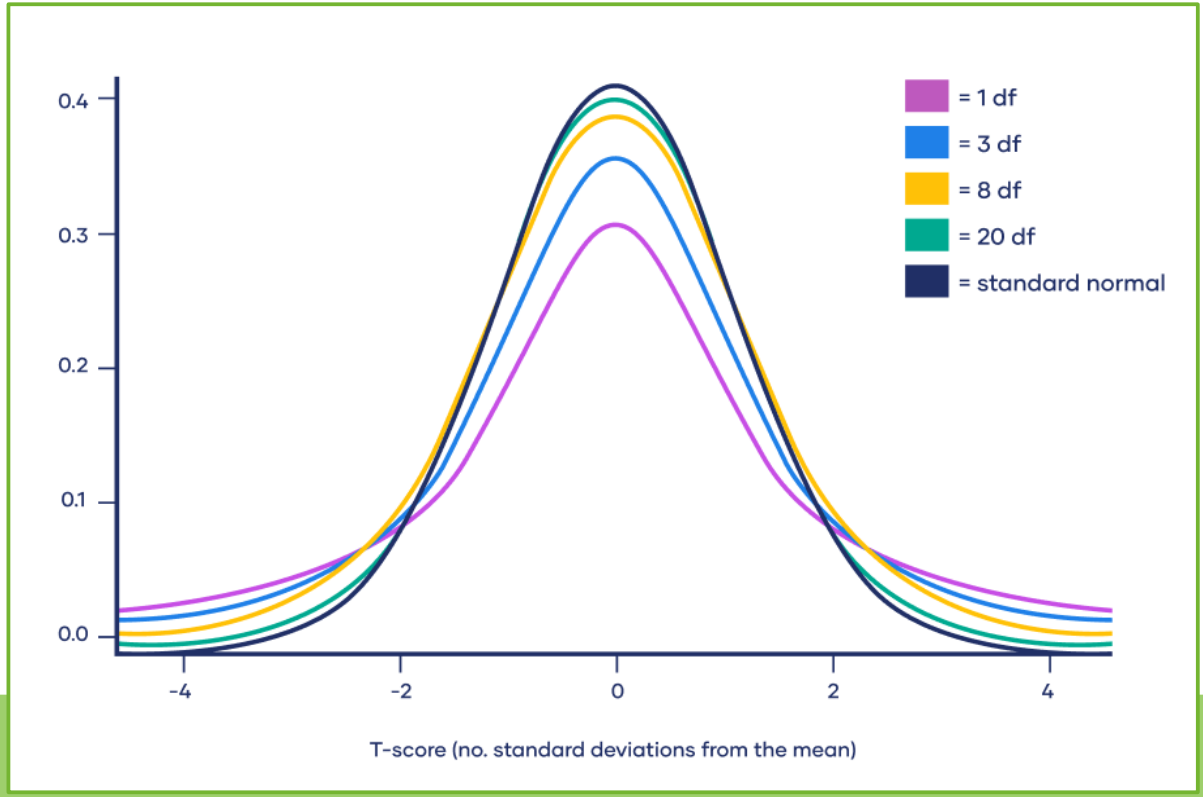
„Minden olyan próba, melyben a próbastatisztika t-eloszlást követ.”

A ν szabadságfokú t-eloszlás sűrűségfüggvénye

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

ahol

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,\mathrm{d}t$$



T-próba („hagyományos”)

Ismeretlen szórású, közel normális eloszlású folytonos valószínűségi változó várható értékére.

Egymintás:
H_0: a minta átlaga m_0

Kétmintás:
H_0: a két minta átlaga azonos

Páros:
H_0: a párok eltérésének átlaga 0

T-próba („hagyományos”)

Ismeretlen szórású, közel normális eloszlású folytonos valószínűségi változó várható értékére.

Egymintás:
H_0: a minta átlaga m_0

Ha a minta elemszáma n, szórása s, átlaga m, akkor a

$$\frac{m-m_0}{s/\sqrt{n}}$$

próbastatisztika eloszlása n-1 szabadságfokú t-eloszlás.

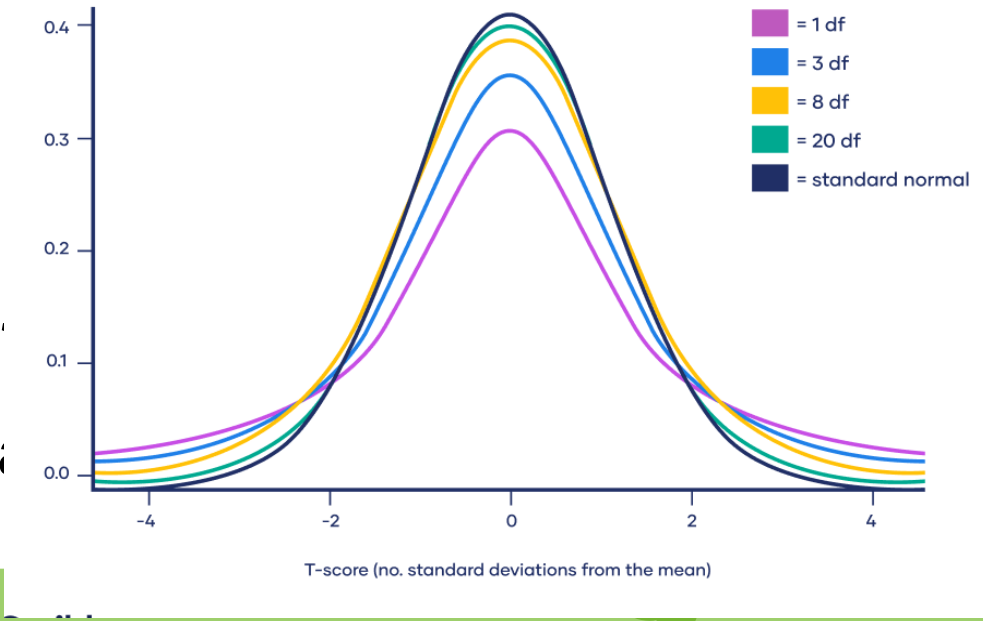
T-próba („hagyományos”)

Ismeretlen szórású, közel normális eloszlású folytonos valószínűségi változó várható értékére.

Egymintás:
H_0: a minta átlaga m_0

Ha a minta elemszáma n, szórása

$$\frac{m-m_0}{s/\sqrt{n}}$$ próbastatisztika eloszlás



T-próba („hagyományos”)

Ismeretlen szórású, közel normális eloszlású folytonos valószínűségi változó várható értékére.

Kétmintás:
H_0: a két minta átlaga azonos

+ feltétel: a két minta elemszáma azonos

Ha a két minta elemszáma n, szórása s_1 és s_2, átlaga m_1 és m_2 akkor a

$$\frac{m_1 - m_2}{\sqrt{\frac{s_1^2 + s_2^2}{2}}\sqrt{\frac{2}{n}}}$$

próbastatisztika eloszlása $2n$-2 szabadságfokú t-eloszlás.

T-próba („hagyományos”)

Ismeretlen szórású, közel normális eloszlású folytonos valószínűségi változó várható értékére.

Páros:
H_0: a párok eltérésének átlaga 0

Egymintás t-próba a párok különbségére

Ha a párok elemszáma n, a különbségek szórása s, átlaga m, akkor a

$$\frac{m}{s/\sqrt{n}}$$

próbastatisztika eloszlása n-1 szabadságfokú t-eloszlás.

Fisher-féle egzakt teszt

	Tejbe tea (valóság)	Teába tej (valóság)
Tejbe tea (tipp)	9	3
Teába tej (tipp)	7	8

H_0: A két esemény független

Fisher-féle egzakt teszt

	Tejbe tea (valóság)	Teába tej (valóság)
Tejbe tea (tipp)	9	3
Teába tej (tipp)	7	8

H_0: A két esemény független

Fisher-féle egzakt teszt

	Tejbe tea (valóság)	Teába tej (valóság)	
Tejbe tea (tipp)	9	3	12
Teába tej (tipp)	7	8	15
	16	11	

H_0: A két esemény független

Fisher-féle egzakt teszt

	Tejbe tea (valóság)	Teába tej (valóság)	
Tejbe tea (tipp)			12
Teába tej (tipp)			15
	16	11	

H_0: A két esemény független

Fisher-féle egzakt teszt

	Tejbe tea (valóság)	Teába tej (valóság)	
Tejbe tea (tipp)	7.11	4.89	12
Teába tej (tipp)	8.89	6.11	15
	16	11	

H_0: A két esemény független

Fisher-féle egzakt teszt

	Tejbe tea (valóság)	Teába tej (valóság)	
Tejbe tea (tipp)	$9_{7.11}$	$3_{4.89}$	12
Teába tej (tipp)	$7_{8.89}$	$8_{6.11}$	15
	16	11	

H_0: A két esemény független

szignifikáns az eltérés?

Fisher-féle egzakt teszt

	Tejbe tea (valóság)	Teába tej (valóság)	
Tejbe tea (tipp)	9 (7.11)	3 (4.89)	12
Teába tej (tipp)	7 (8.89)	8 (6.11)	15
	16	11	

Feladat: írjunk R programot, ami eldönti, hogy elfogadjuk vagy elutasítsuk-e H_0-at!

H_0: A két esemény független

szignifikáns az eltérés?

Működik a gyógyszer?

Egy új, vaspótló gyógyszert vérszegény patkányokon teszteltek. Megmérték a patkányok vérében található vas szintjét a 0. napon és a 48. napon:

	Pat. 1	Pat. 2	Pat. 3	Pat. 4	Pat. 5
0. nap	9	11.4	21.3	79.8	5.8
48. nap	24.2	10.1	24.4	77.7	12.4

A placebót kapó kontrollcsoport:

	Pat. 6	Pat. 7	Pat. 8	Pat. 9	Pat. 10
0. nap	12.2	9.5	86.6	14.1	15.6
48. nap	10	6.6	79.3	11.3	18.1

Működik a gyógyszer?

2 mintás t-próba

Gyógyszeres csoport:

	Pat. 1	Pat. 2	Pat. 3	Pat. 4	Pat. 5
diff.	15.2	-1.3	3.1	-2.1	6.6

Kotrollcsoport:

	Pat. 6	Pat. 7	Pat. 8	Pat. 9	Pat. 10
diff.	-2.2	-2.9	-7.3	-2.8	2.5

H_0: a két minta azonos várható értékű eloszlásból származik.

2 mintás t-próba

Gyógyszeres csoport:

	Pat. 1	Pat. 2	Pat. 3	Pat. 4	Pat. 5
diff.	15.2	-1.3	3.1	-2.1	6.6

$m_1 = 4.3, \quad s_1 = 7.03$

Kotrollcsoport:

	Pat. 6	Pat. 7	Pat. 8	Pat. 9	Pat. 10
diff.	-2.2	-2.9	-7.3	-2.8	2.5

$m_2 = -2.54, \quad s_2 = 3.48$

$$\frac{m_1 - m_2}{\sqrt{\frac{s_1^2 + s_2^2}{2}}\sqrt{\frac{2}{n}}} = 1.94$$

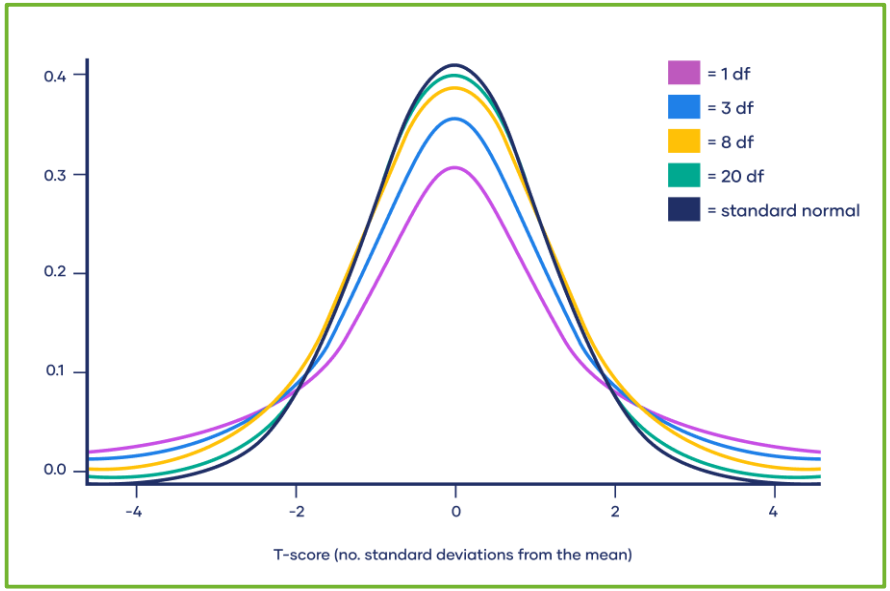
H_0 95%-os elfogadási tartománya a 8 szabadságfokú t-eloszlásból: [−2.306, 2.306]


Hogyan válasszuk ki
a mintaméretet?

Mekkora minta kéne a vaspótló gyógyszerhez?

$m_1 = 4.3,\ s_1 = 7.03$ $m_2 = -2.54,\ s_2 = 3.48$

n	$\frac{m_1 - m_2}{\sqrt{\frac{s_1^2 + s_2^2}{2}}\sqrt{\frac{2}{n}}}$	df	Elfogadási intervallum
4	1.74	6	[-2.45, 2.45]
5	1.94	8	[-2.31, 2.31]
6	2.14	10	[-2.23, 2.23]
7	2.31	12	[-2.18, 2.18]
8	2.47	14	[-2.15, 2.15]
9	2.62	16	[-2.12, 2.12]
10	2.76	18	[-2.10, 2.10]



1. 250g kenőcs

50 elemű minta, normális eloszlás, ismer szórás, 248.1 az átlag. Baj van a töltéssel?

2. kézmosás covid előtt/után

Előtte: 26/100,
Utána: 32/100.
Nőtt a hajlandóság?

3. sebfertőzés normál / laparoszkópos esetben, kontingencia táblázat

	-	+
N	37453	4421
L	19454	642

4. Vércukor csökkentő gyógyszer

30 fős minta, m = 6.8%, s = 0.5
99% szignifikancia mellett igaz, hogy az átlag 7% alatti?

5. érettségi eredmények normalitása

15-15 pontszám eloszlása tekinthető normálisnak?

6. hallgatói értékelés, 2 értékelési stílus összehasonlítása

A 15 fiú és 9 lány pontozása ugyanabból az eloszlásból vett minta?

7. működik a homeopátiás altató?

A 10 ember elalvási ideje:

	1	2	3	4	5	6	7	8	9	10
placebo	23.2	14	59.2	31.1	47.4	7.8	44.2	21.1	33.3	9.2
altató	31.5	7.8	54.9	33	11.1	23.3	43	16.7	12.6	8.6

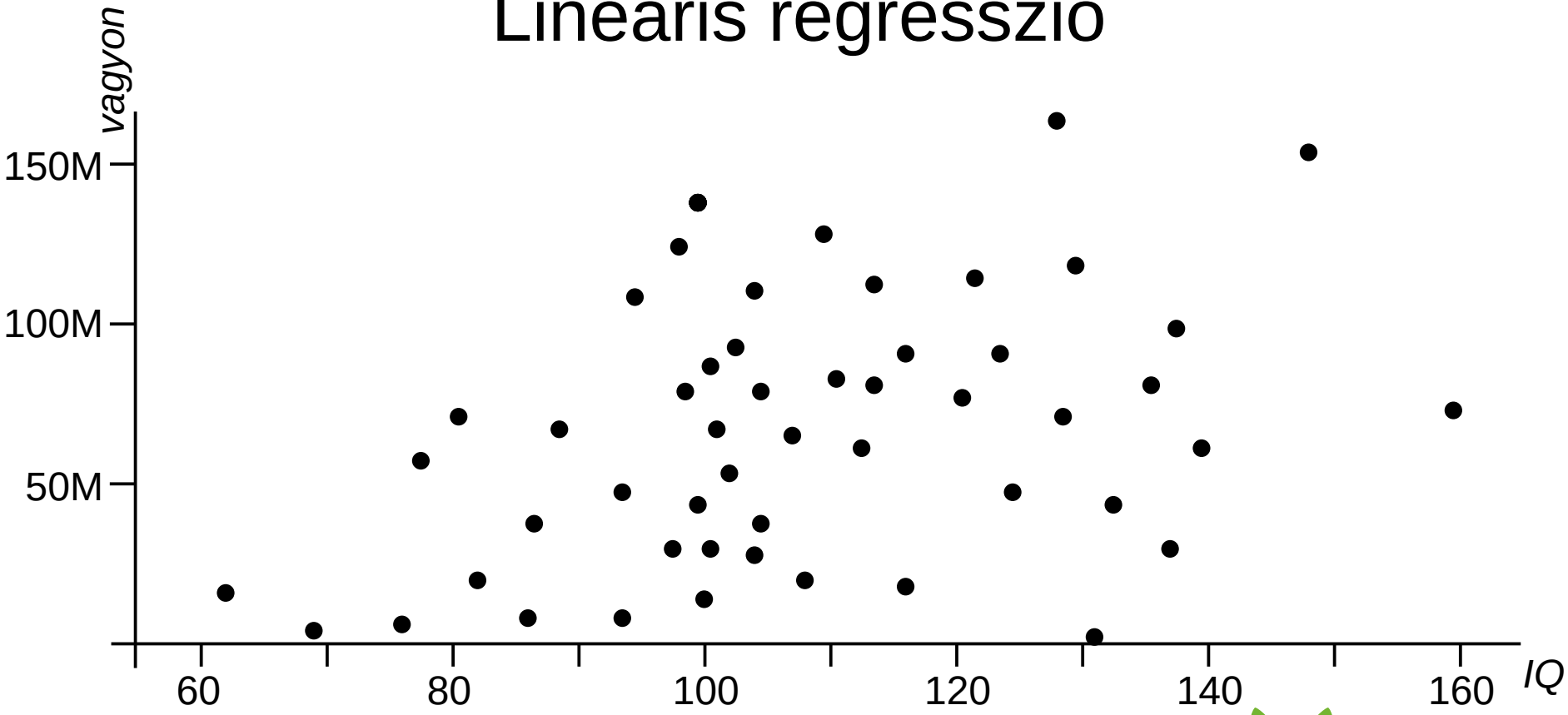
8. jobb a Comirnaty?

	tünetm	enyhe	kórház
Sinopharm	87	44	14
Comirnaty	46	19	6

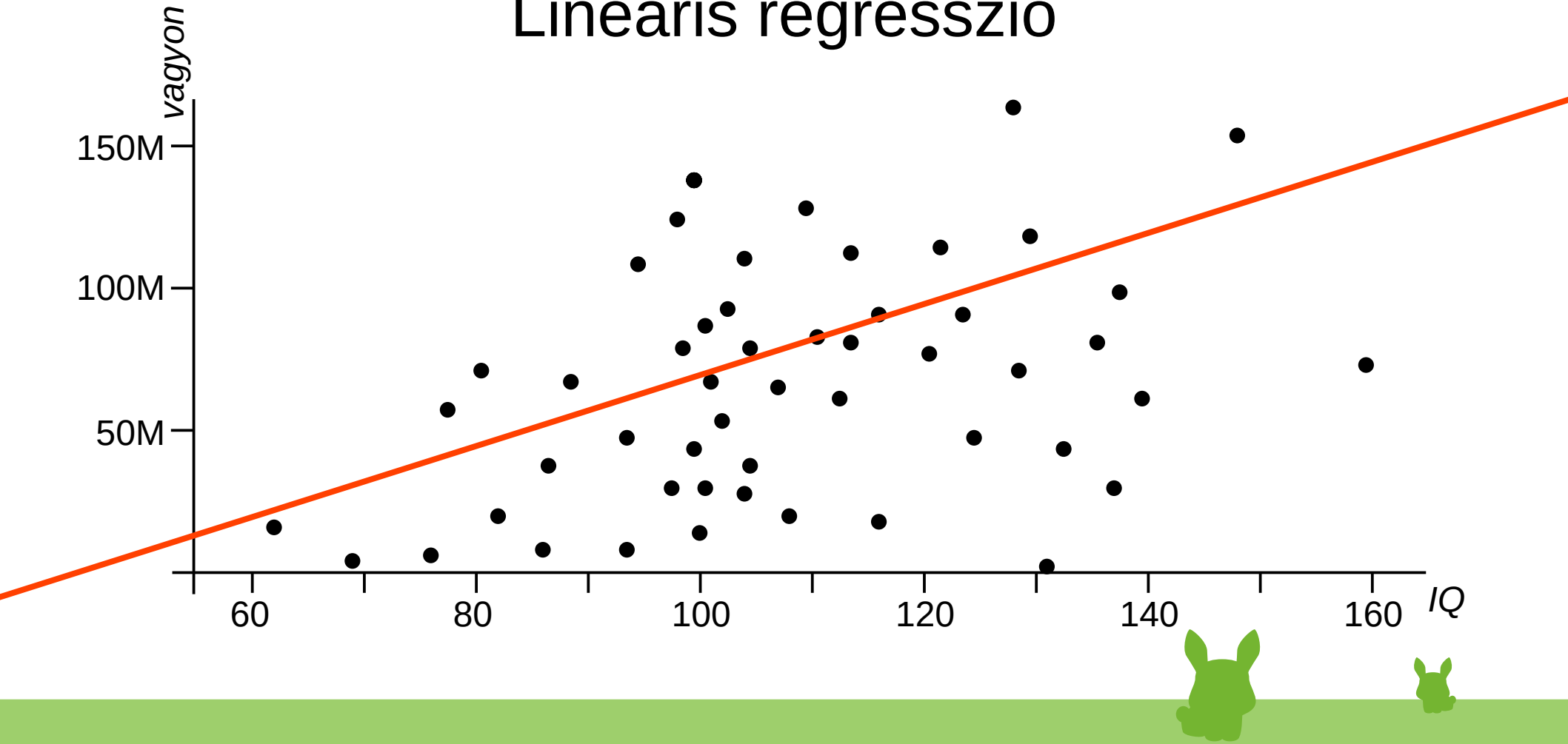
Tematika

- Leíró statisztika (átlag, medián, módusz, szórás...)
- Paraméterbecslés (mennyi?)
- Hipotézisvizsgálat (igaz vagy nem?)
- Regressziószámítás, modellezés (hogy befolyásol?)

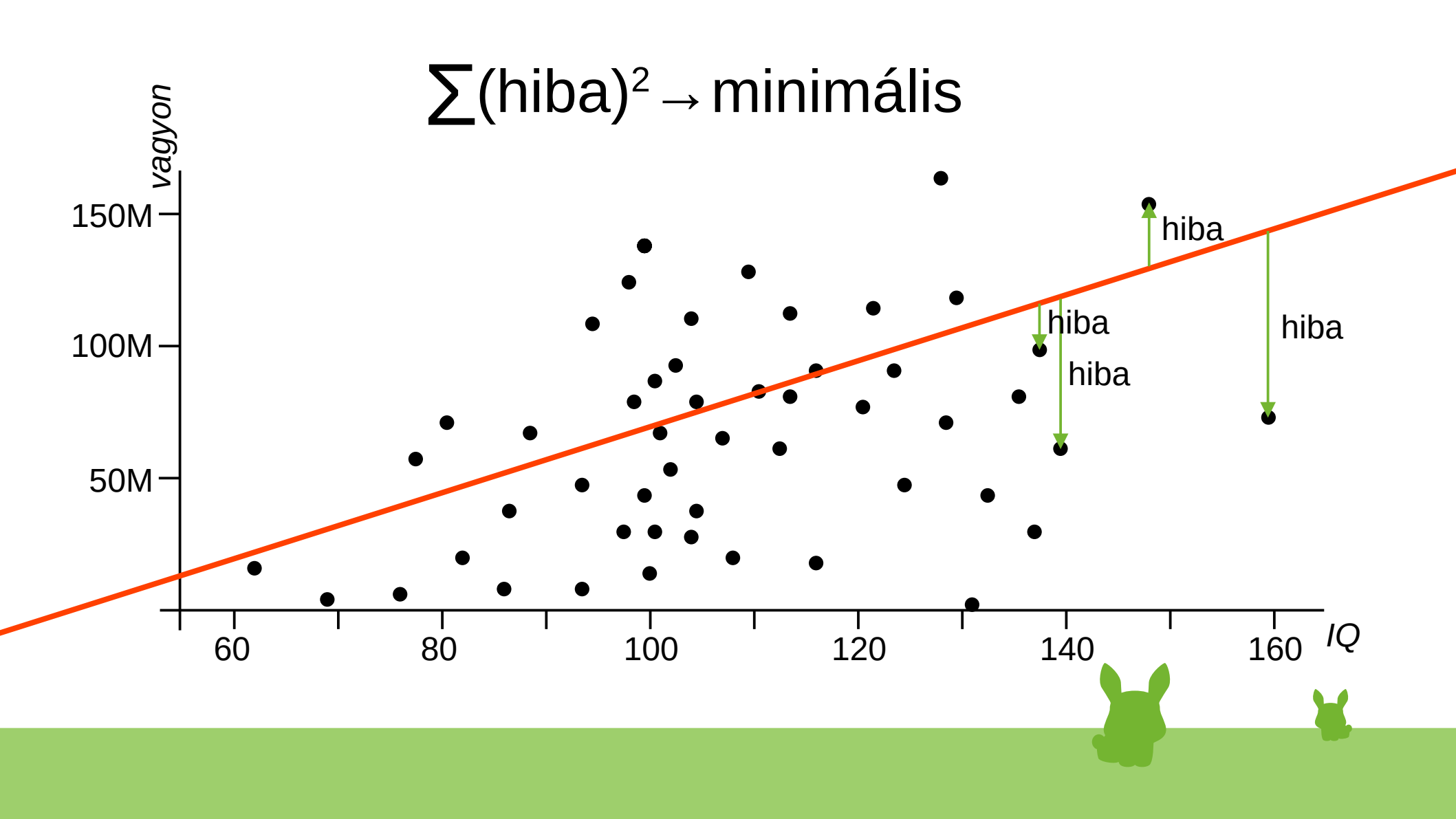
Lineáris regresszió

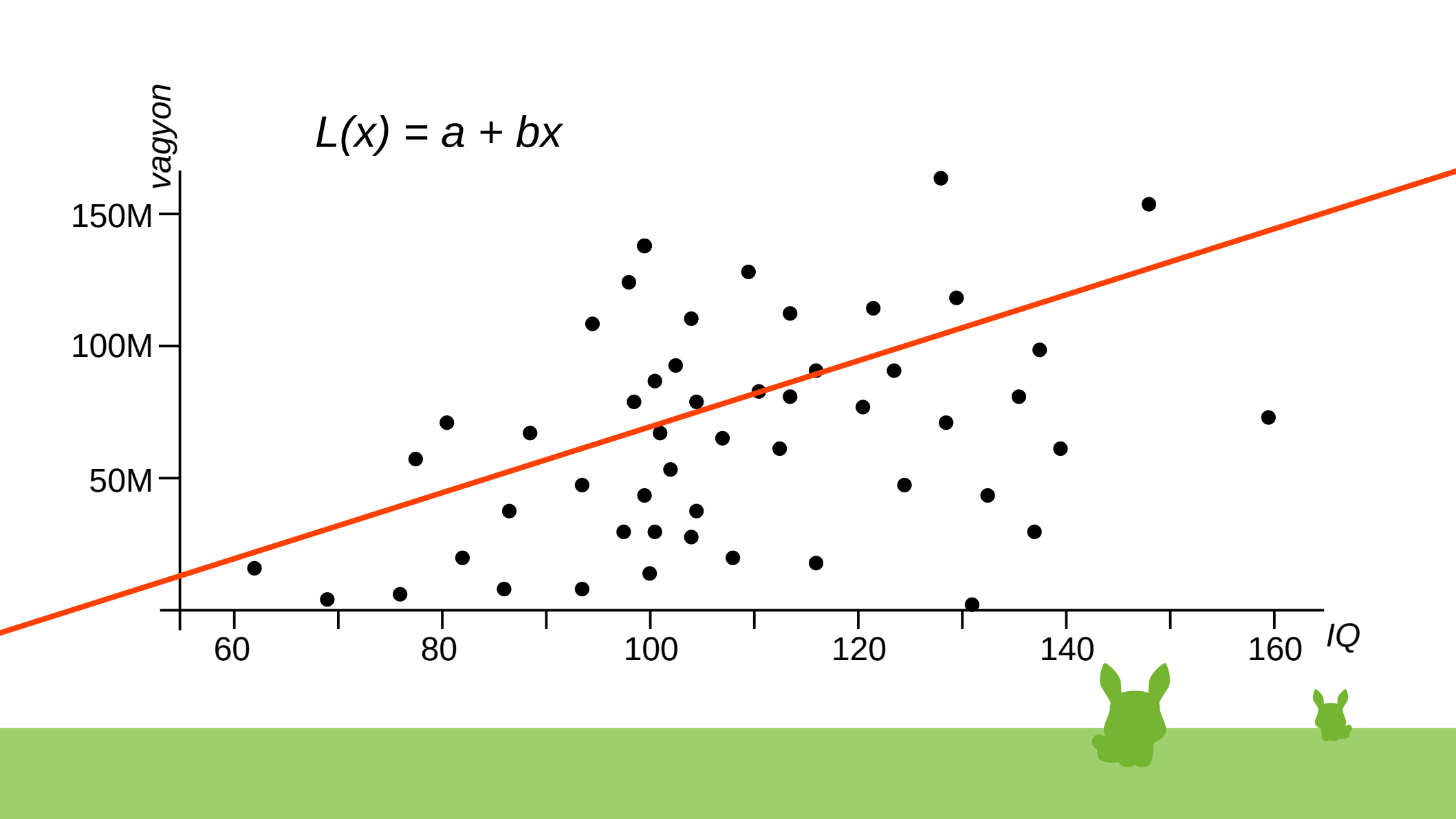


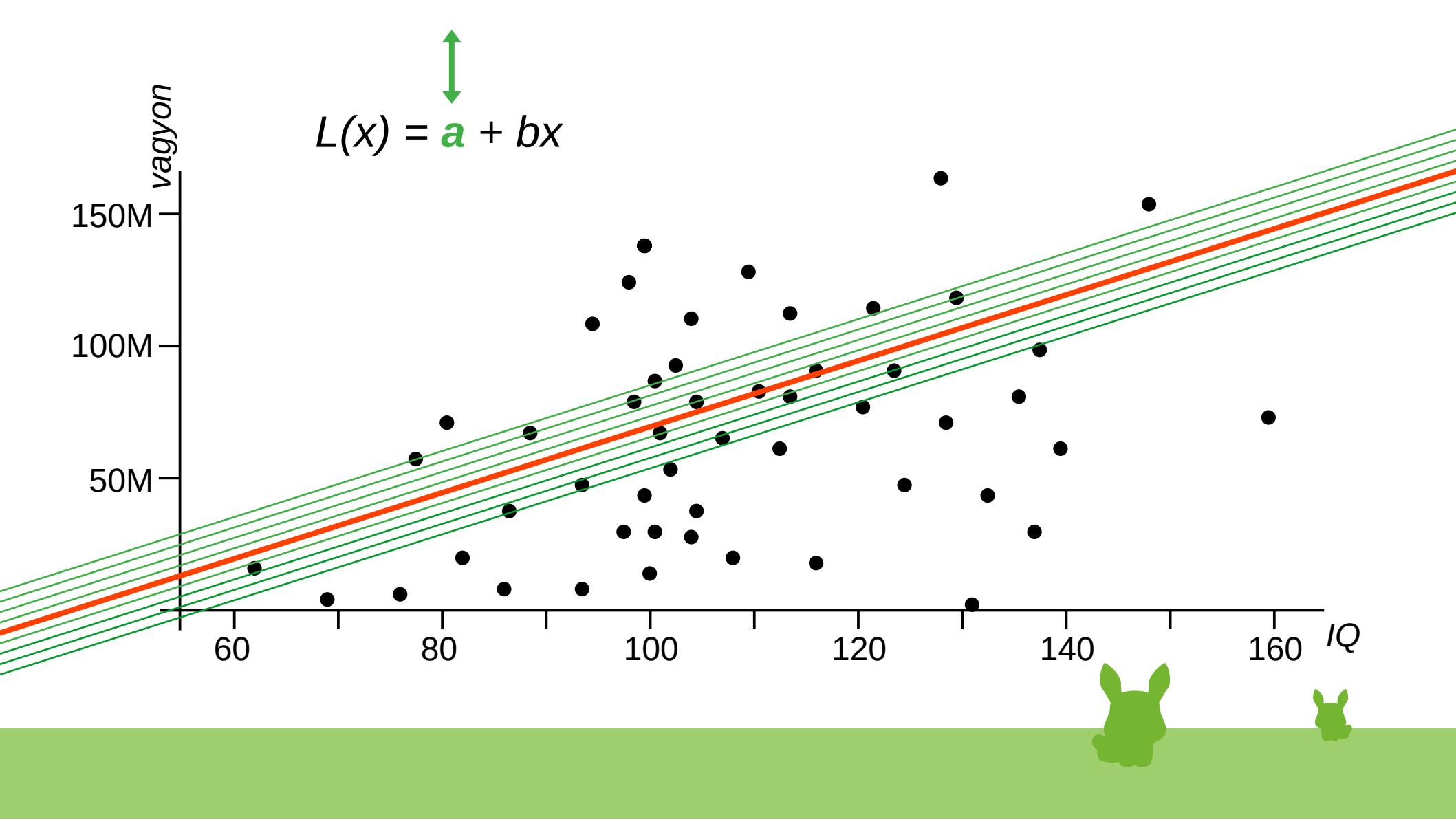
Lineáris regresszió

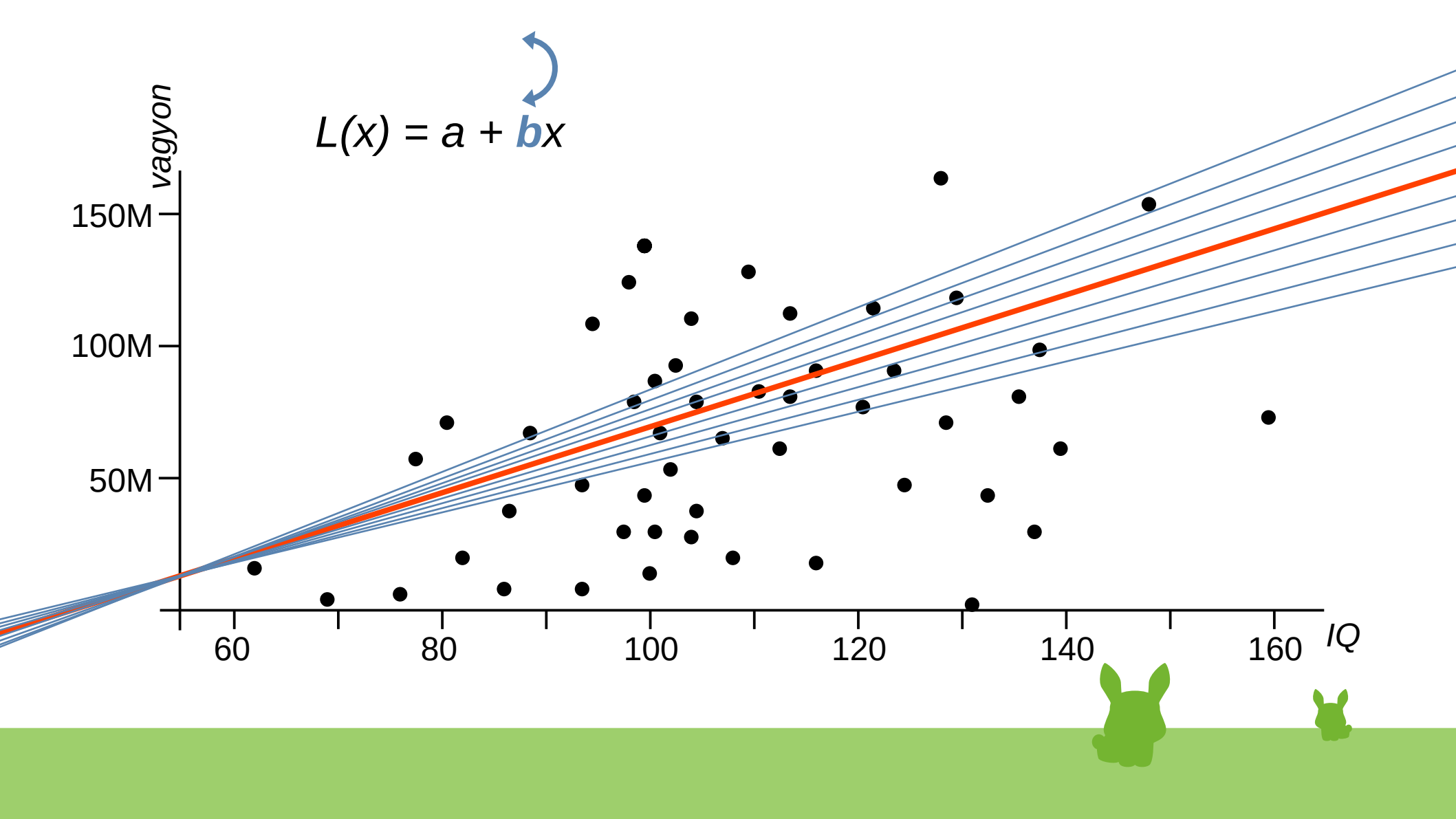


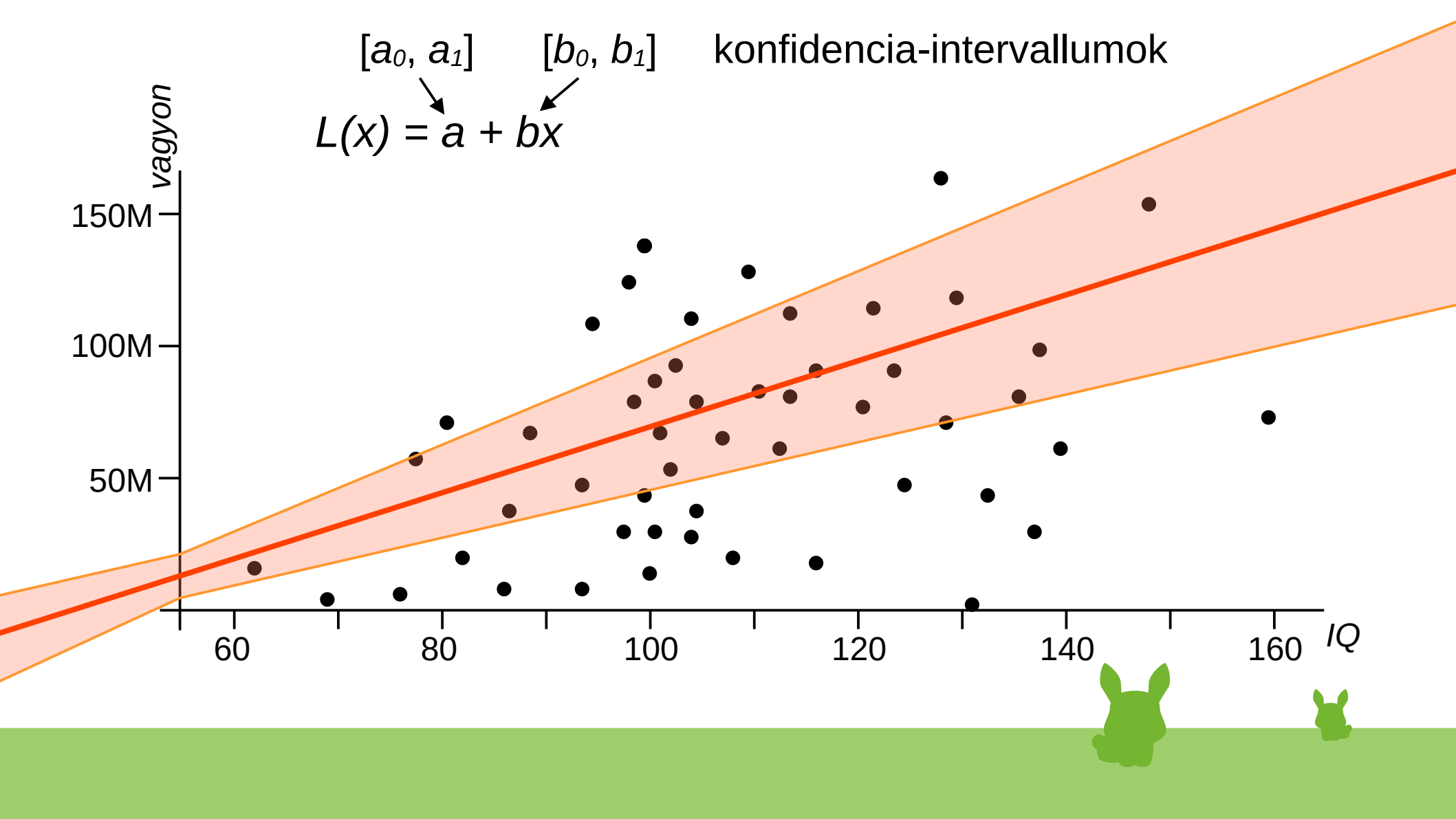
$\sum(\text{hiba})^2 \rightarrow$ minimális

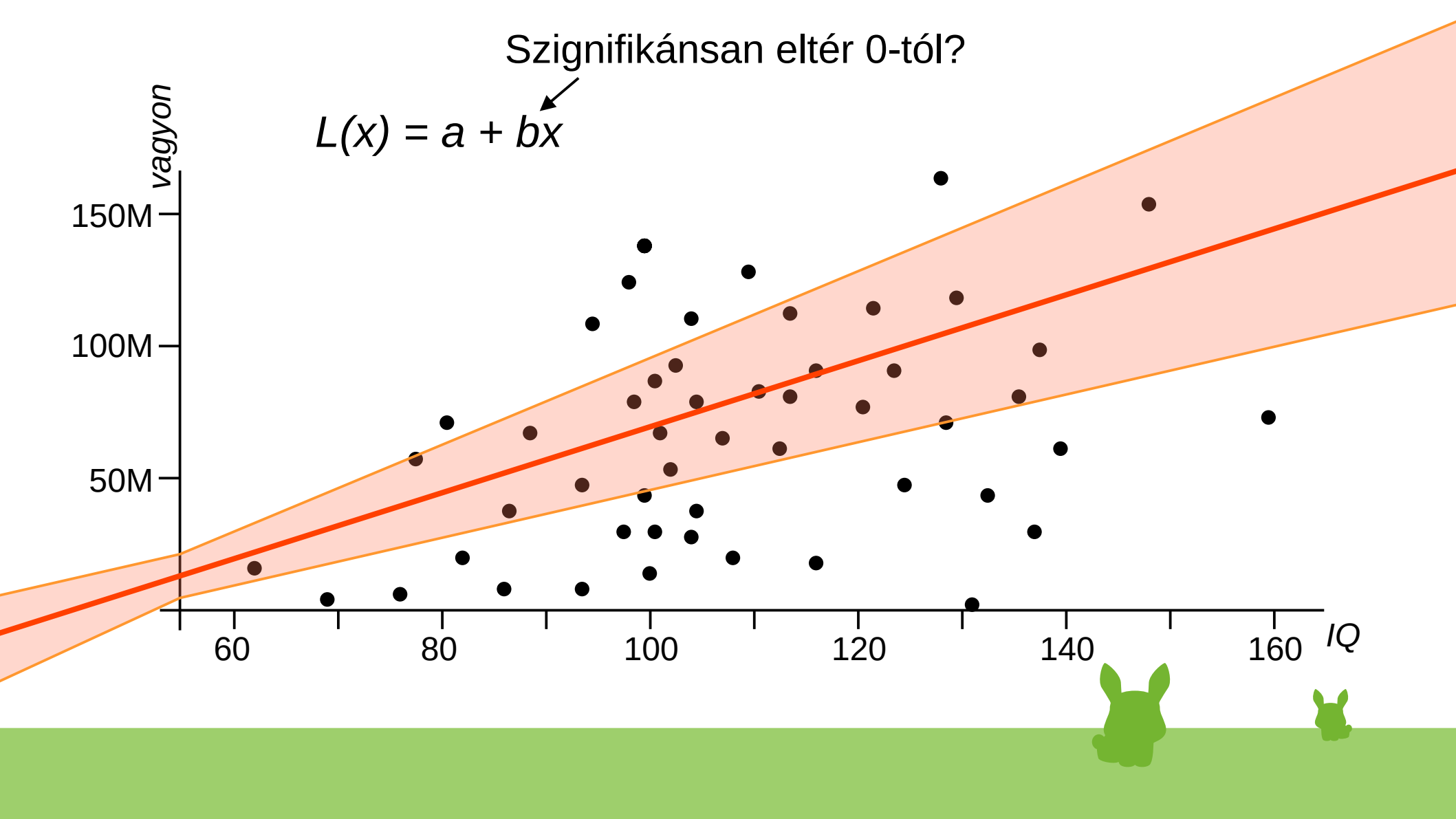


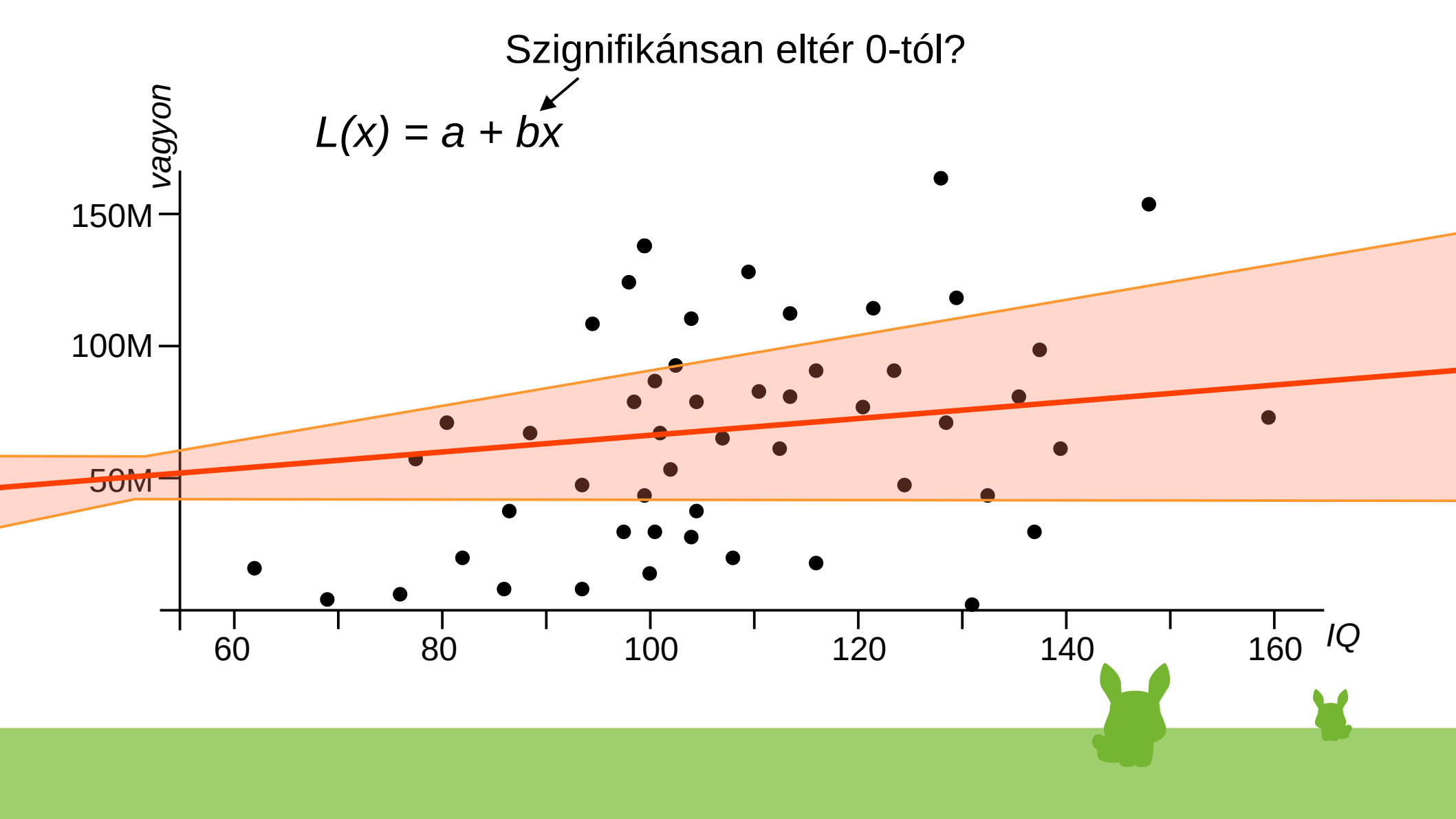

$L(x) = a + bx$
vagyon
150M
100M
50M
60
80
100
120
140
160
IQ


$L(x) = a + bx$
vagyon
150M
100M
50M
60
80
100
120
140
160
IQ

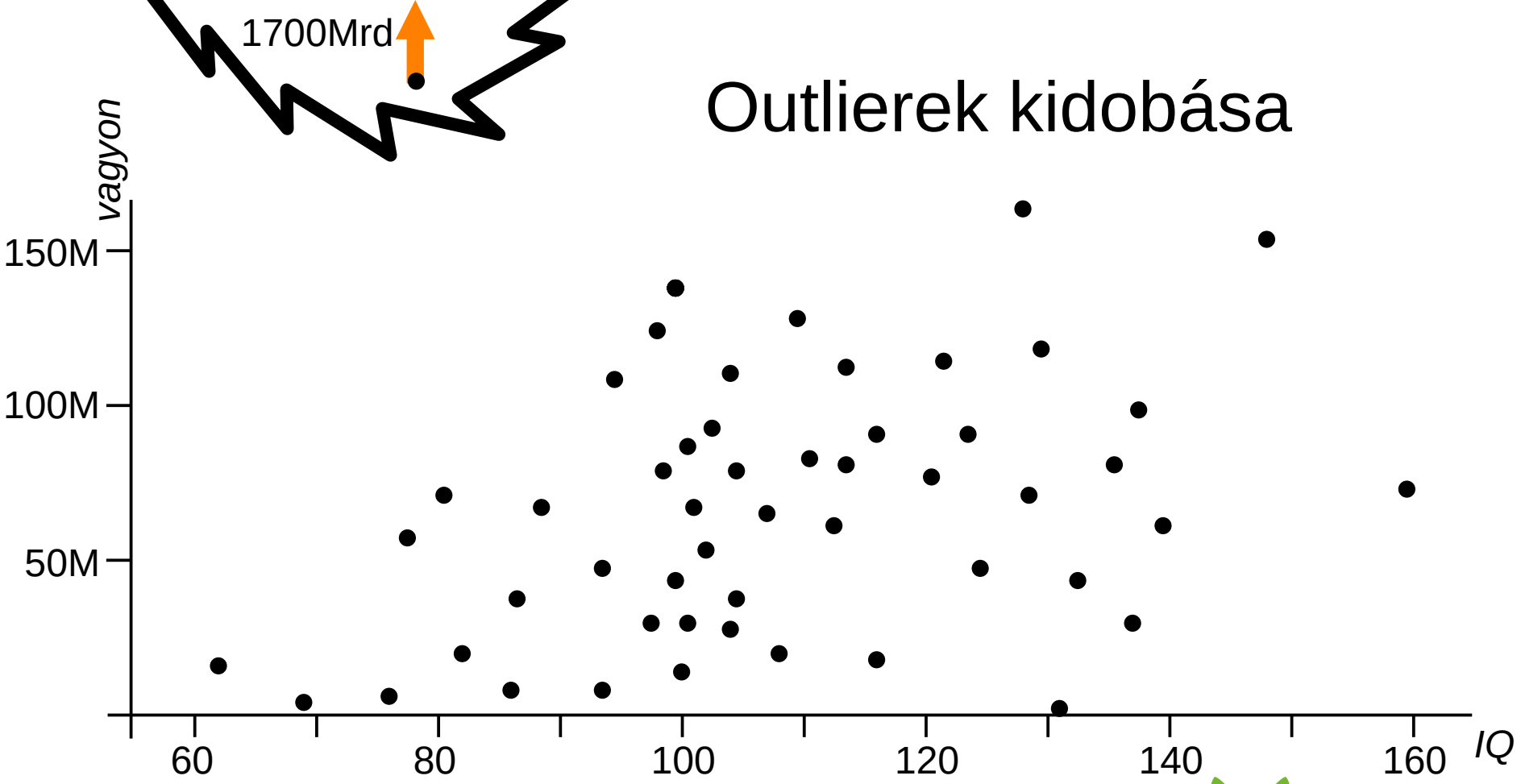

$L(x) = a + bx$
vagyon
150M
100M
50M
60
80
100
120
140
160
IQ


[a_0, a_1]
[b_0, b_1]
konfidencia-intervallumok
$L(x) = a + bx$
vagyon
150M
100M
50M
60
80
100
120
140
160
IQ


Szignifikánsan eltér 0-tól?
L(x) = a + bx
vagyon
150M
100M
50M
60
80
100
120
140
160
IQ

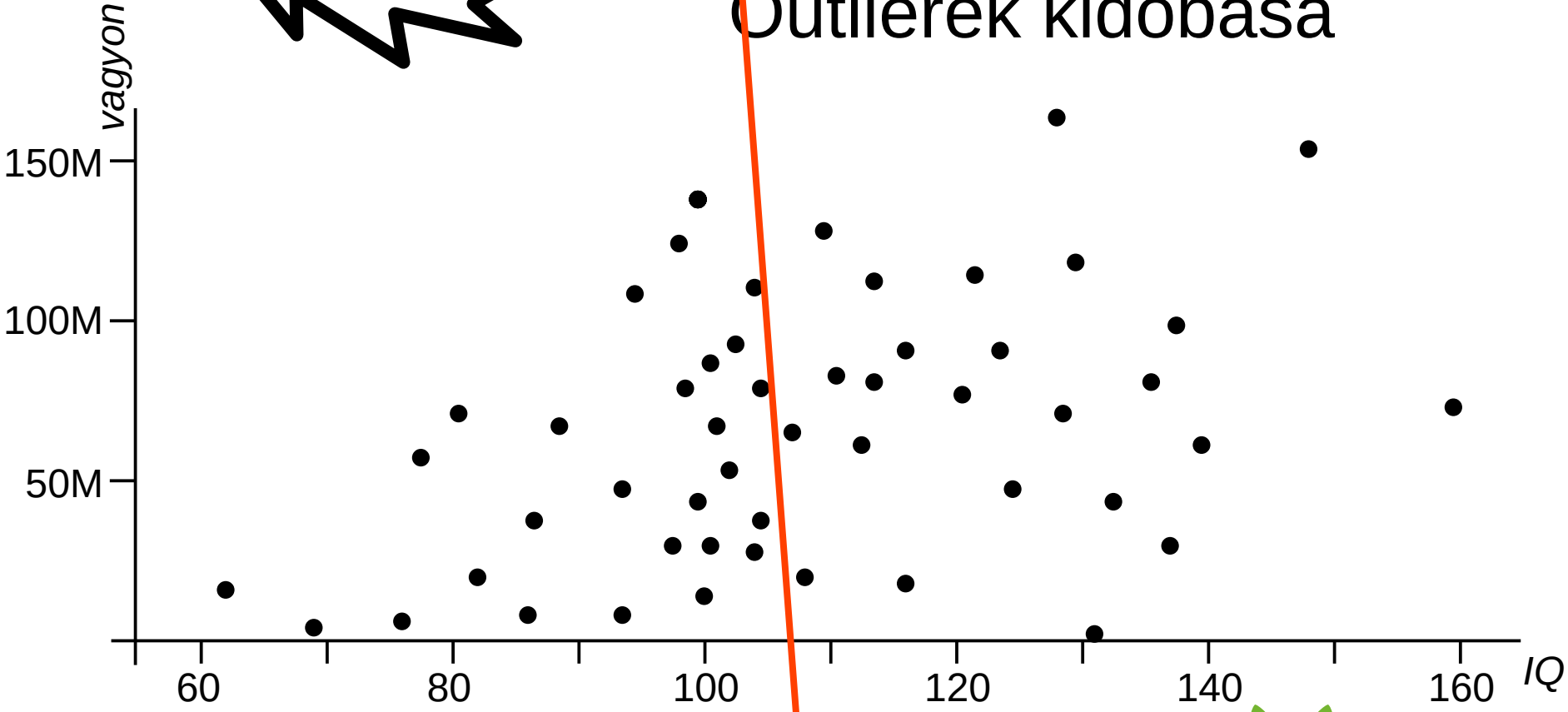

Szignifikánsan eltér 0-tól?
L(x) = a + bx
vagyon
150M
100M
50M
60
80
100
120
140
160
IQ

Outlierek kidobása



1700Mrd

Outlierek kidobása



1. feladat: várható élettartam elemzése

https://ruzsaz.github.io/elettartam.csv

1) Készítsünk lineáris modellt, rajzot: élettartam ~ gdp

```
Make a linear regression to explain life_expectancy with gdp
from the loaded data frame.
Show the confidence intervals.
Draw the data points.
Draw the 95% confidence band.
```

1. feladat: várható élettartam elemzése

```
Call:
lm(formula = life_expectancy ~ gdp, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-19.562  -4.541   1.462   5.222  15.649 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.746e+01  3.027e-01  222.88   <2e-16 ***
gdp         2.594e-04  1.213e-05   21.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.909 on 764 degrees of freedom
Multiple R-squared:  0.3746,    Adjusted R-squared:  0.3737 
F-statistic: 457.5 on 1 and 764 DF,  p-value: < 2.2e-16
                    2.5 %       97.5 %
(Intercept) 6.686499e+01 6.805334e+01
gdp         2.356418e-04 2.832639e-04
```

1. feladat: várható élettartam elemzése

```
Call:
lm(formula = life_expectancy ~ gdp, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-19.562  -4.541   1.462   5.222  15.649

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.746e+01  3.027e-01  222.88   <2e-16 ***
gdp         2.594e-04  1.213e-05   21.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.909 on 764 degrees of freedom
Multiple R-squared:  0.3746,    Adjusted R-squared:  0.3737
F-statistic: 457.5 on 1 and 764 DF,  p-value: < 2.2e-16
                    2.5 %       97.5 %
(Intercept) 6.686499e+01 6.805334e+01
gdp         2.356418e-04 2.832639e-04
```

Élettartam = 67.4 + 0.00025 * gdp

T-próba, H_0: együttható = 0

szignifikancia

illeszkedés várható hibája

modell fittsége: az élettartam szórásának ennyied részét magyarázza a regresszió (0-1)

95%-os konfidencia intervallum

F-próba, H_0: minden együttható* = 0

(*: a konstans tag nincs beleértve)

Nemlineáris regresszió?

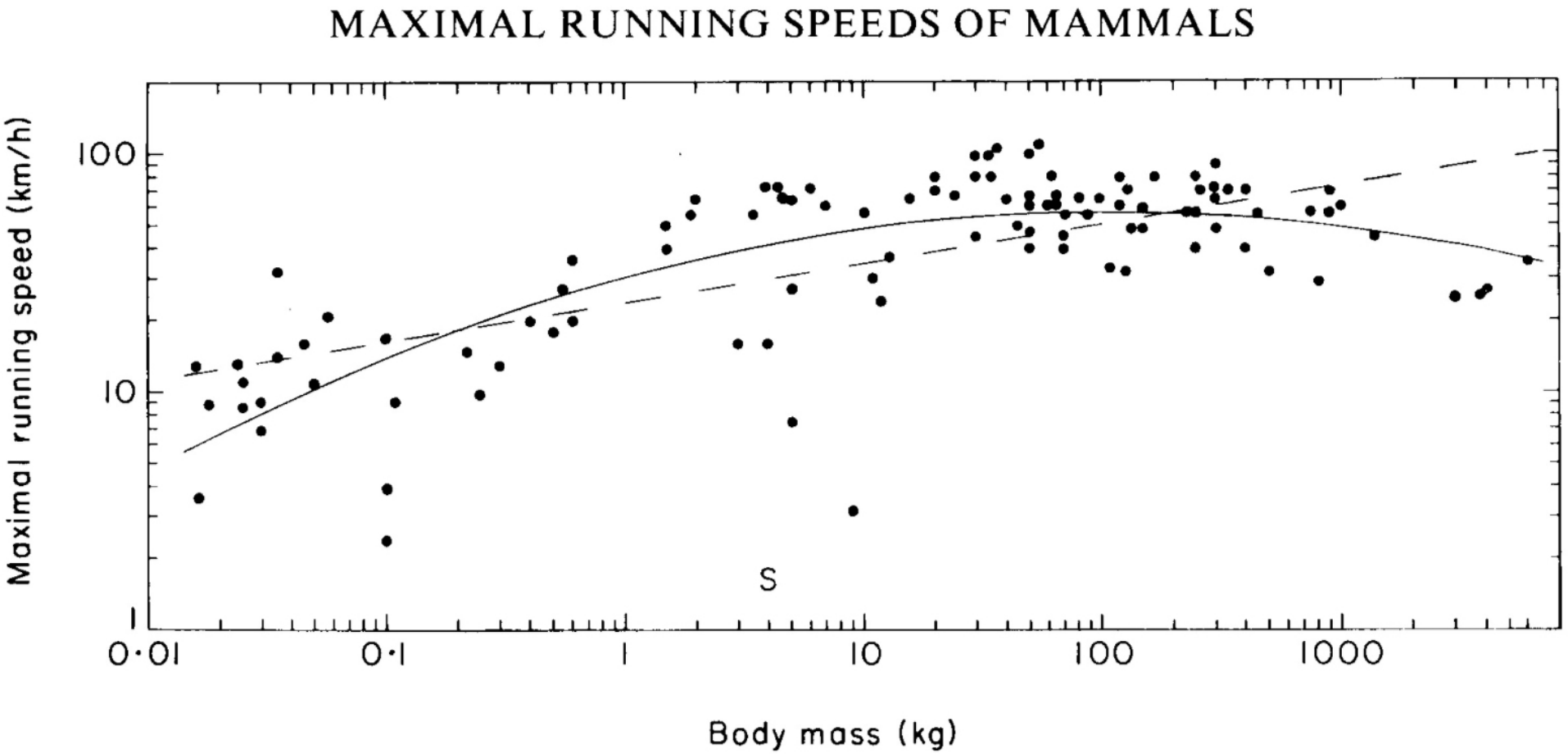
$L(x) = a + bx$

$S(x) = a + bx + cx^2$

$Q(x) = a + bx + cx^2 + dx^3$

$Ln(x) = a + b\ln(x)$

...


MAXIMAL RUNNING SPEEDS OF MAMMALS
Maximal running speed (km/h)
100
10
1
0·01
0·1
1
10
100
1000
S
Body mass (kg)

1. feladat: várható élettartam elemzése

https://ruzsaz.github.io/elettartam.csv

1) Készítsünk lineáris modellt, rajzot: élettartam ~ gdp
2) Négyzetes modell? Más függvény?

1. feladat: várható élettartam elemzése

https://ruzsaz.github.io/elettartam.csv

1) Készítsünk lineáris modellt, rajzot: élettartam ~ gdp
2) Négyzetes modell? Más függvény?
3) Nézzük meg a többi lehetséges magyarázó változót is.

1. feladat: várható élettartam elemzése

https://ruzsaz.github.io/elettartam.csv

1) Készítsünk lineáris modellt, rajzot: élettartam ~ gdp
2) Négyzetes modell? Más függvény?
3) Nézzük meg a többi lehetséges magyarázó változót is.
4) Mely magyarázó változókkal lesz a modell a legjobb?

   (R: step eljárás automatikusan megkeresi)

2. feladat: egészséges életmód elemzése

https://ruzsaz.github.io/egeszseg.csv

Célváltozó: health (0-100)

Magyarázó változók:

- age, bmi, exercise (0-7), diet (0-100), sleep (óra)
- smoke, sex: (kategória változók: 0, 1)