

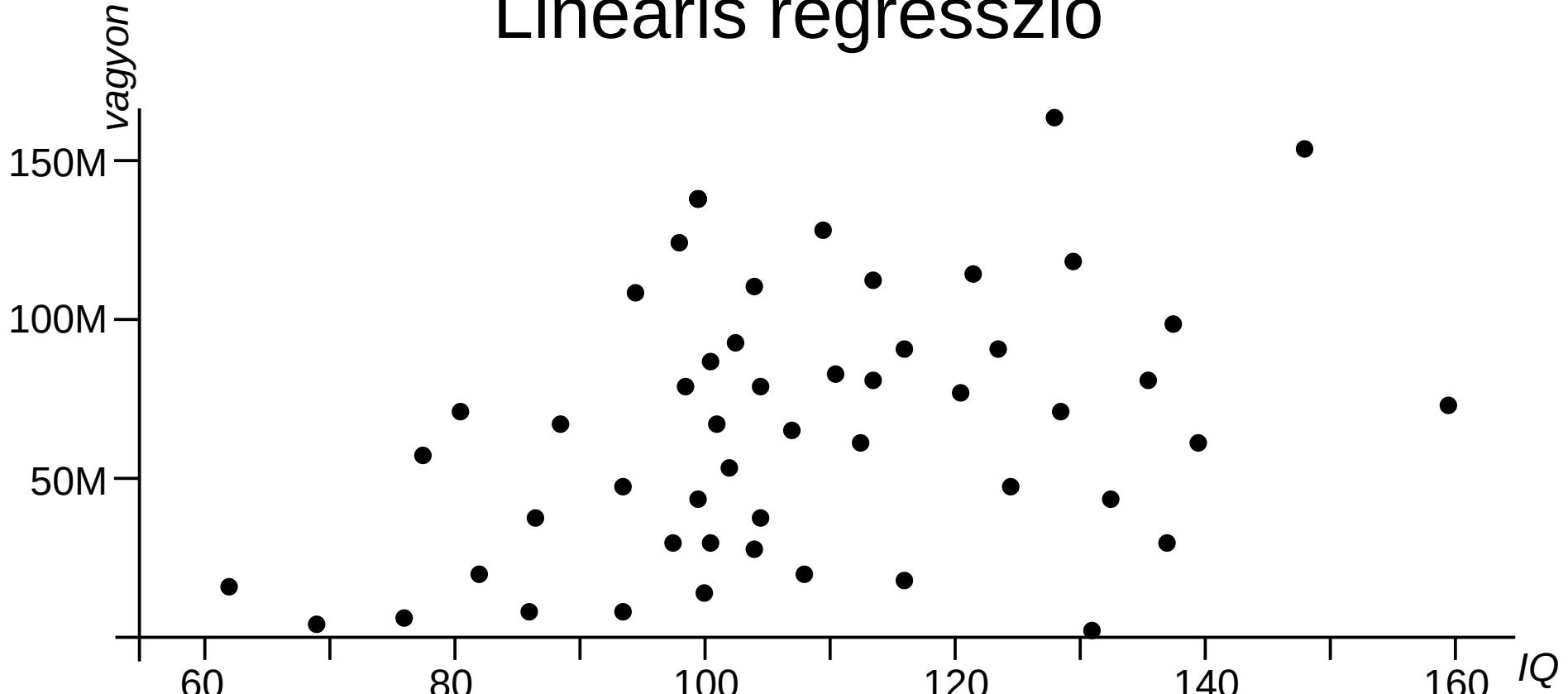
Tematika

- ~~Leíró statisztika (átlag, median, módsz, szórás...)~~
 - Paraméterbecslés (mennyi?)
 - Hipotézisvizsgálat (igaz vagy nem?)
 - Regressziószámítás, modellezés (hogyan befolyásol?)

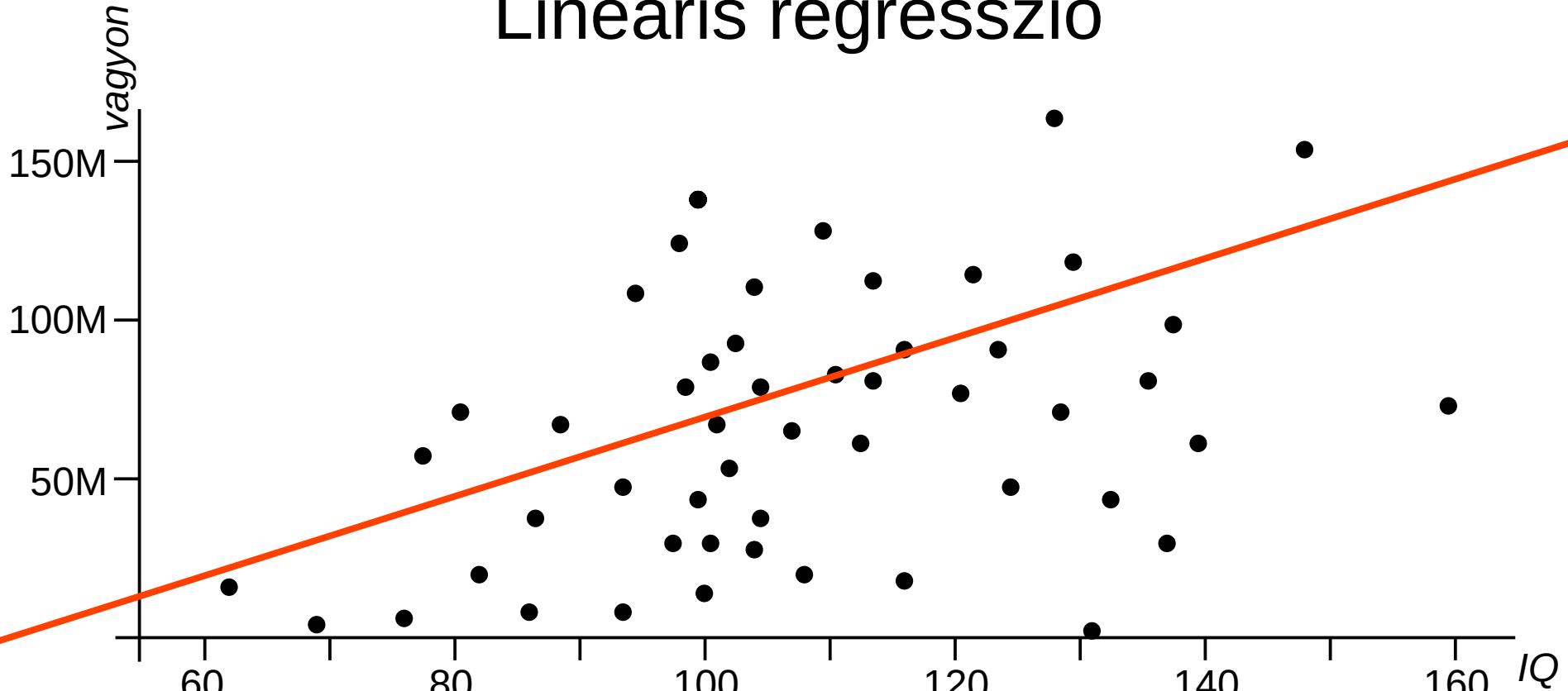
<https://ruzsaz.github.io/stat4.pdf>



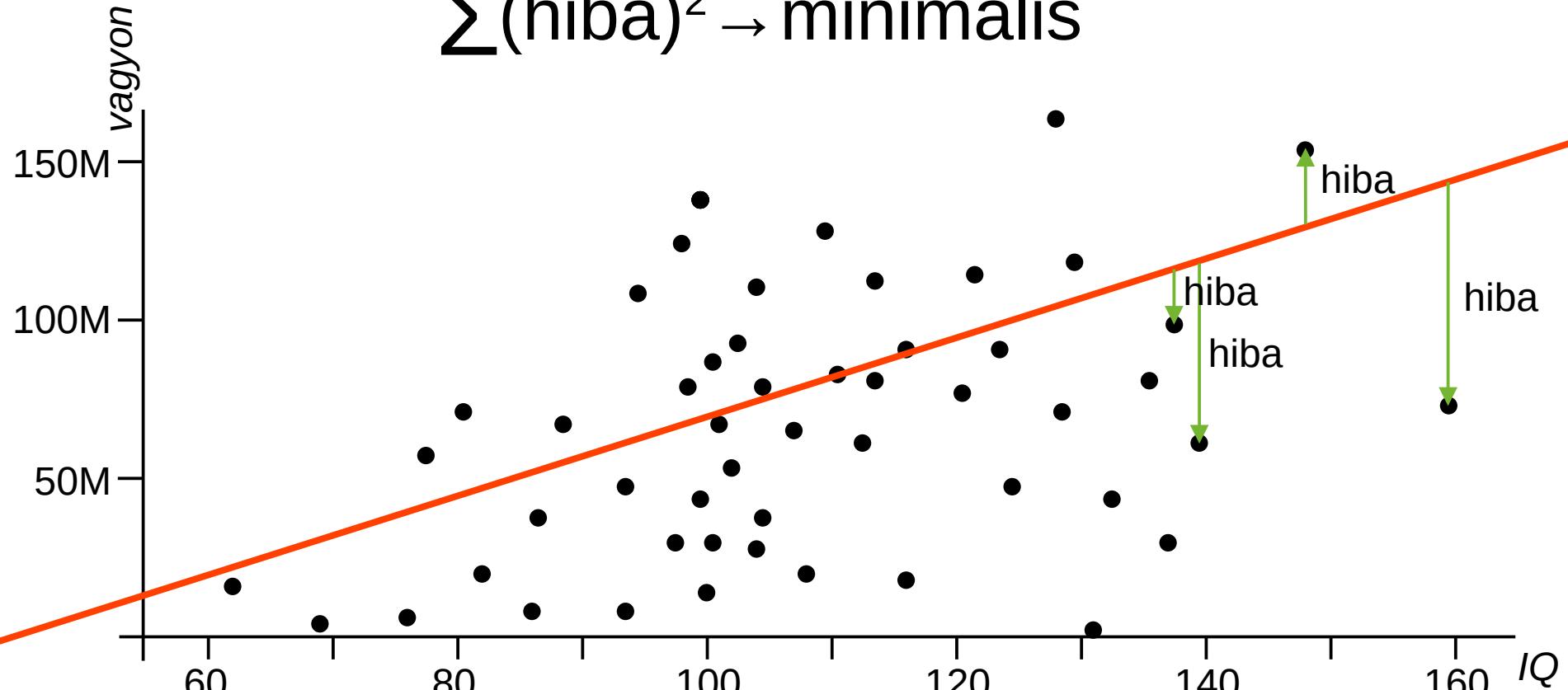
Lineáris regresszió



Lineáris regresszió



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



vagyon

$$L(x) = a + bx$$

150M

100M

50M

60

80

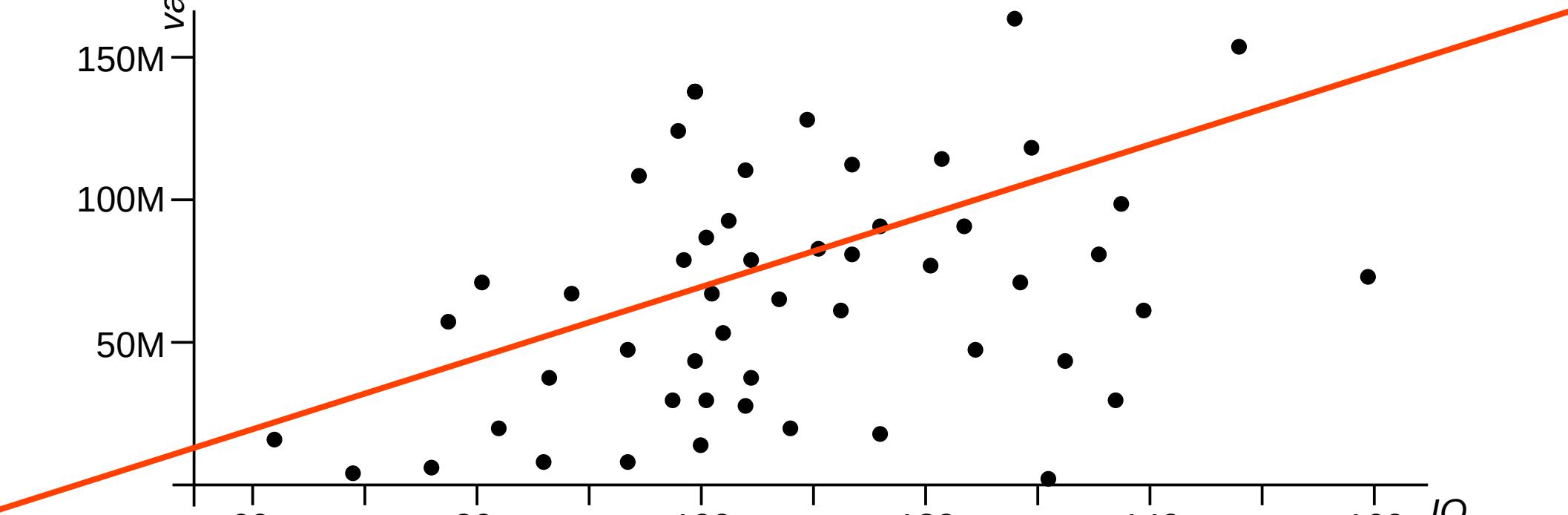
100

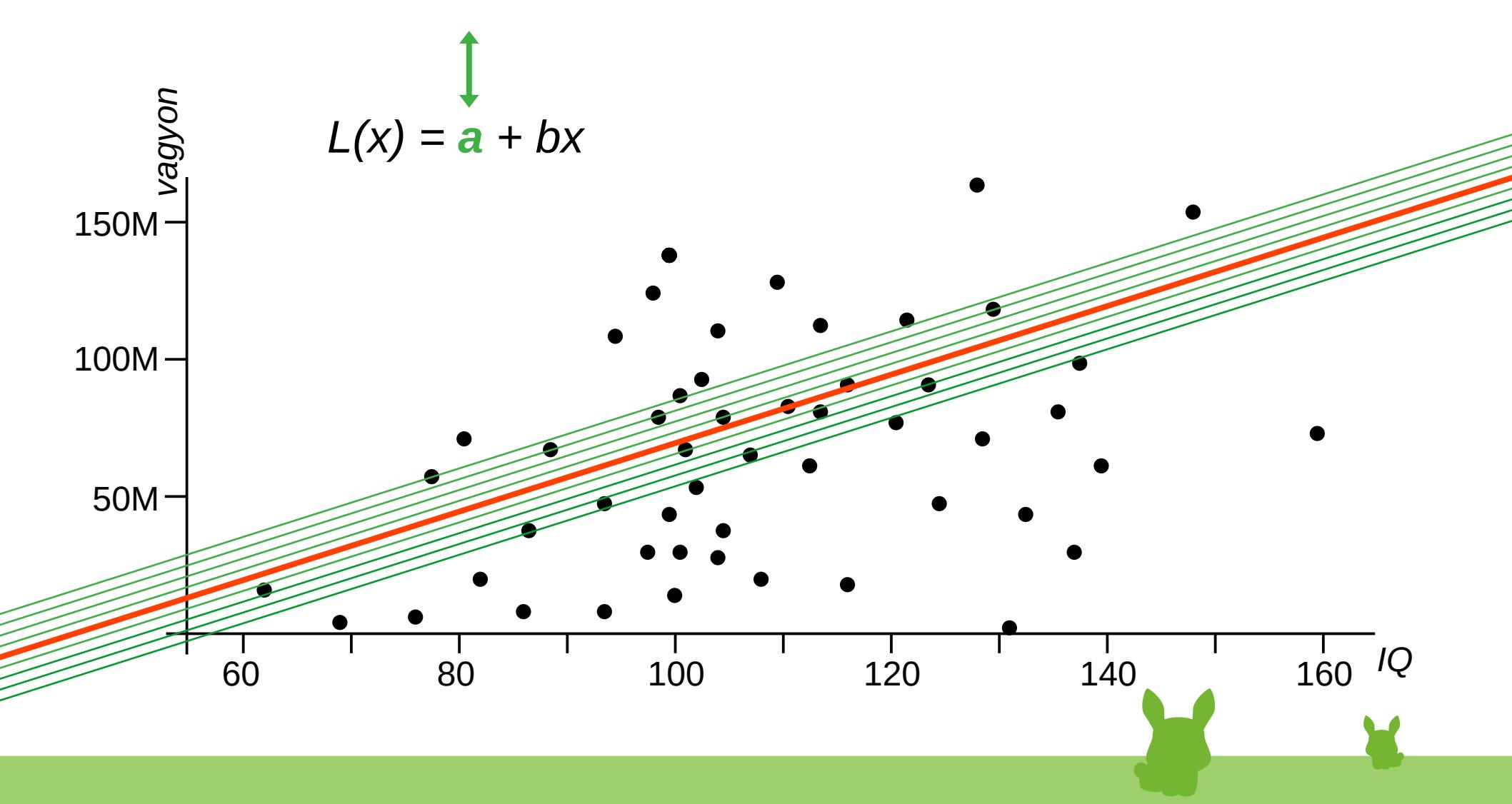
120

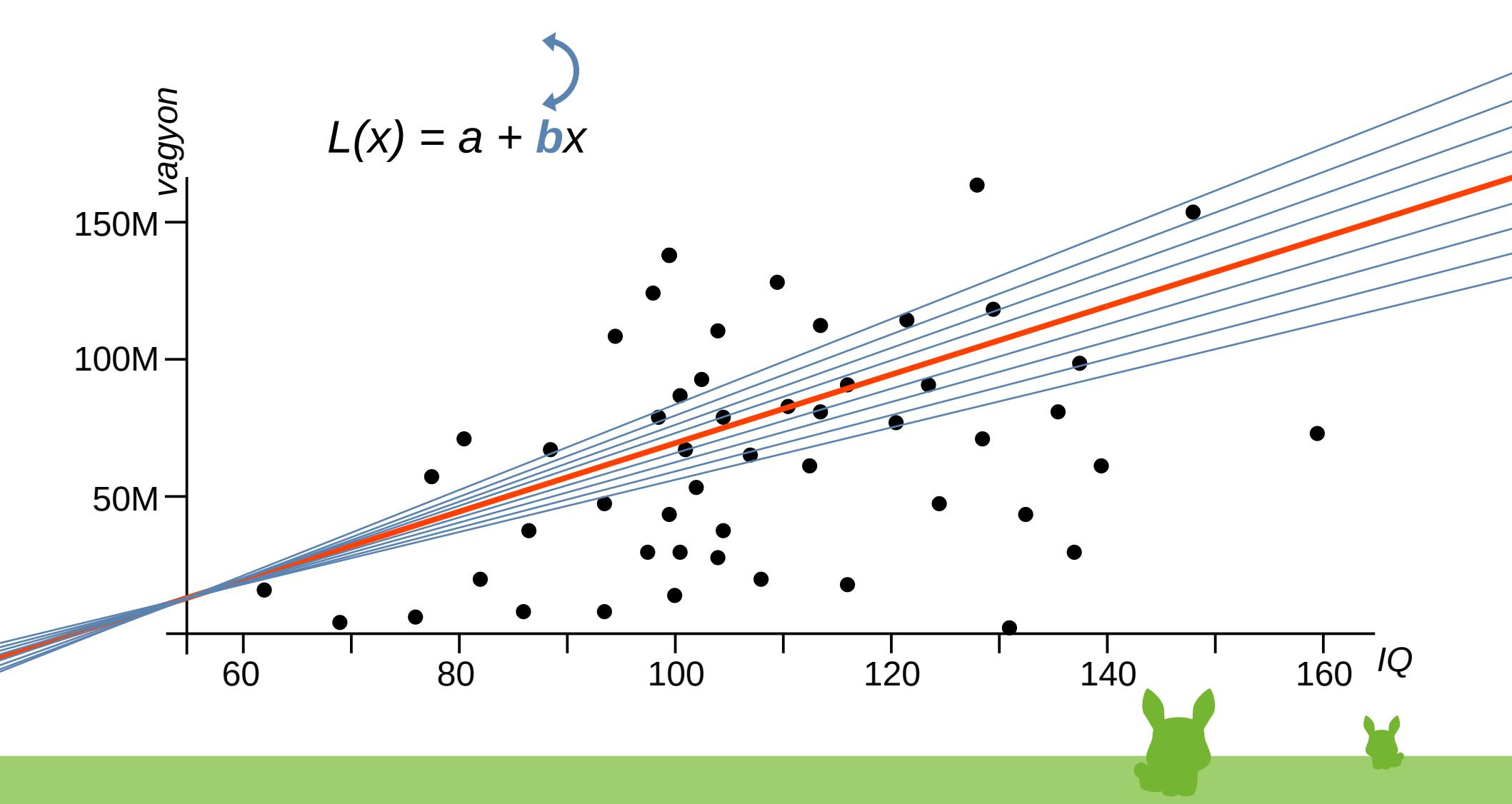
140

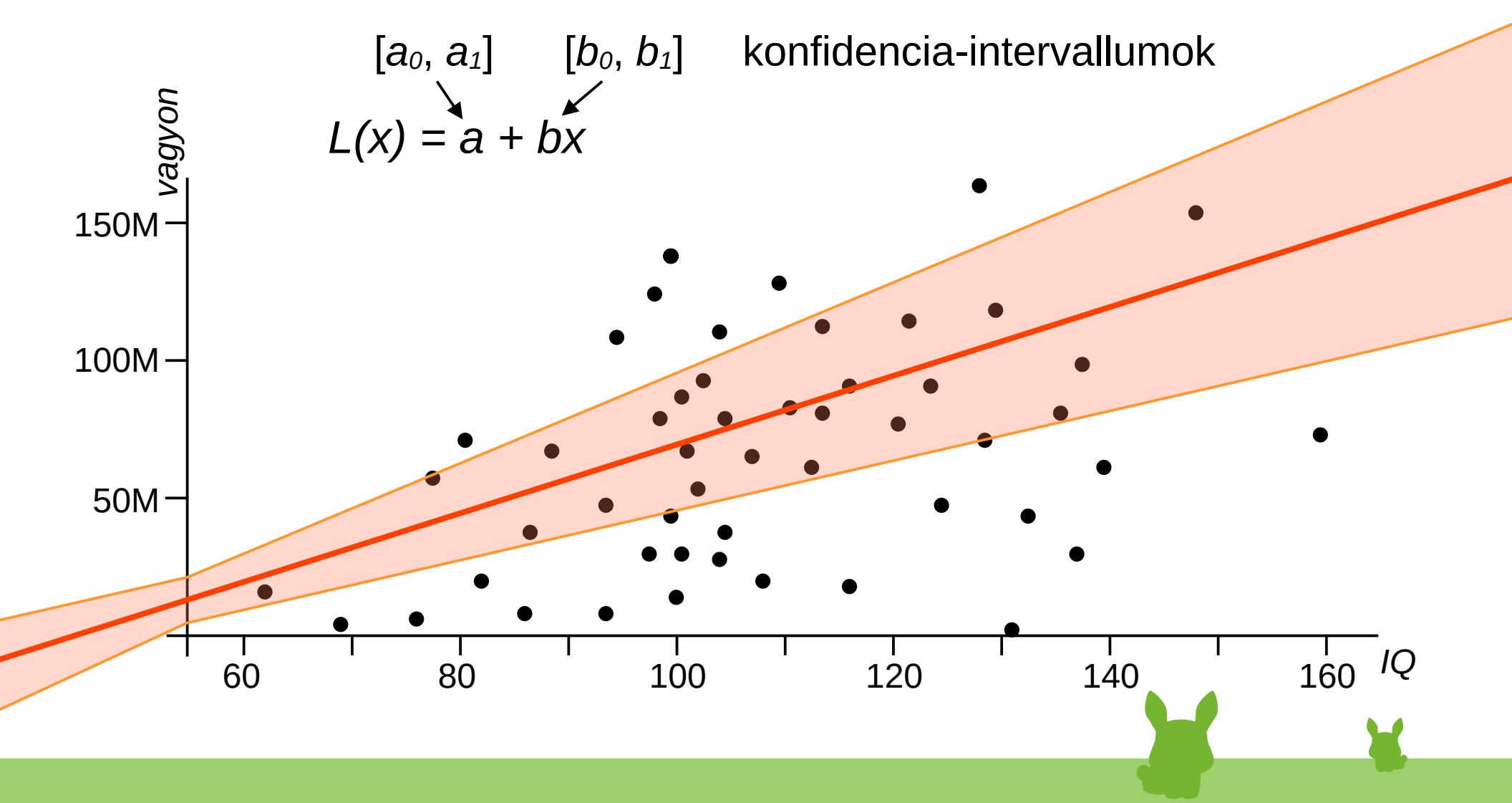
160

IQ



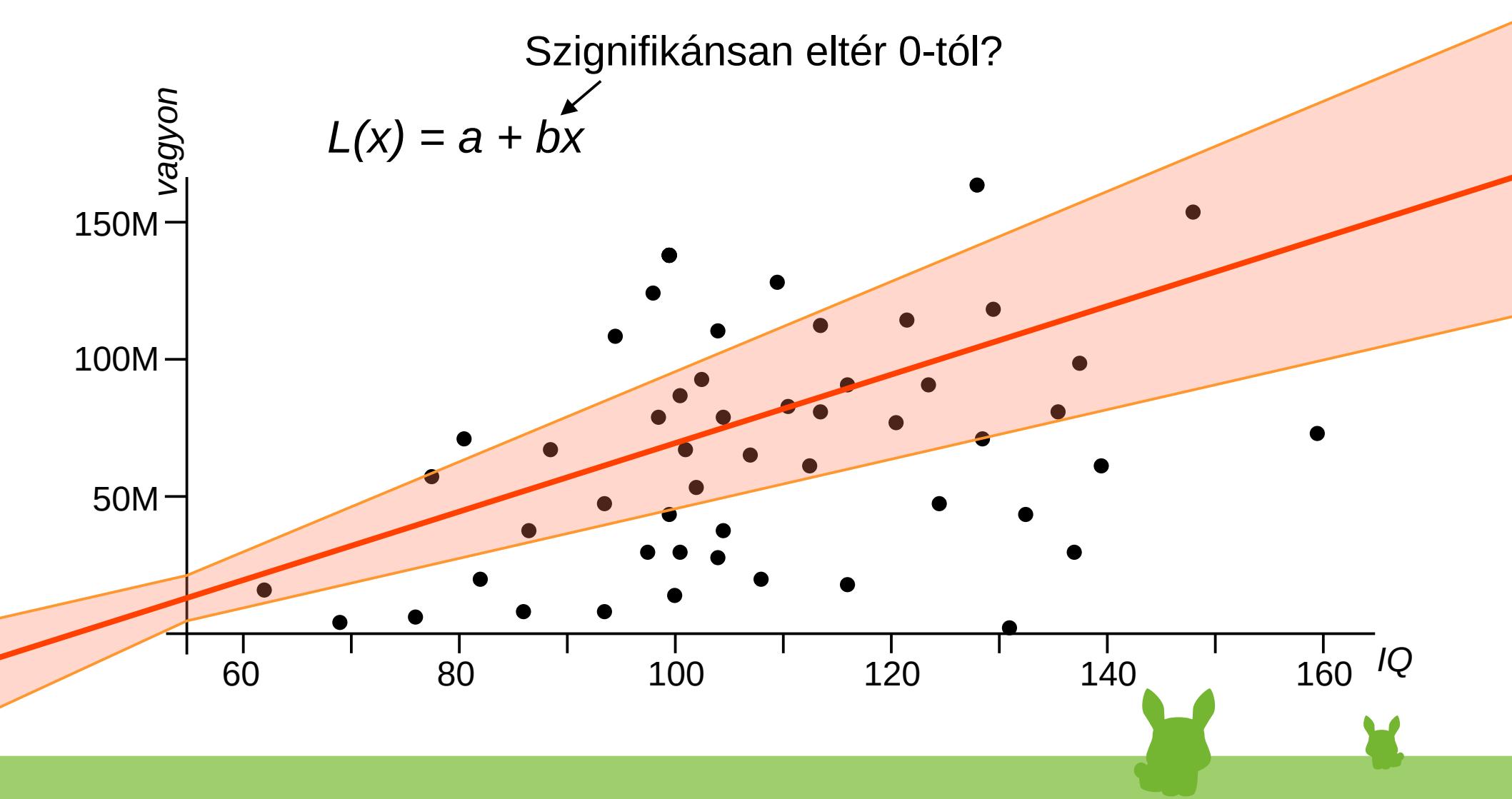




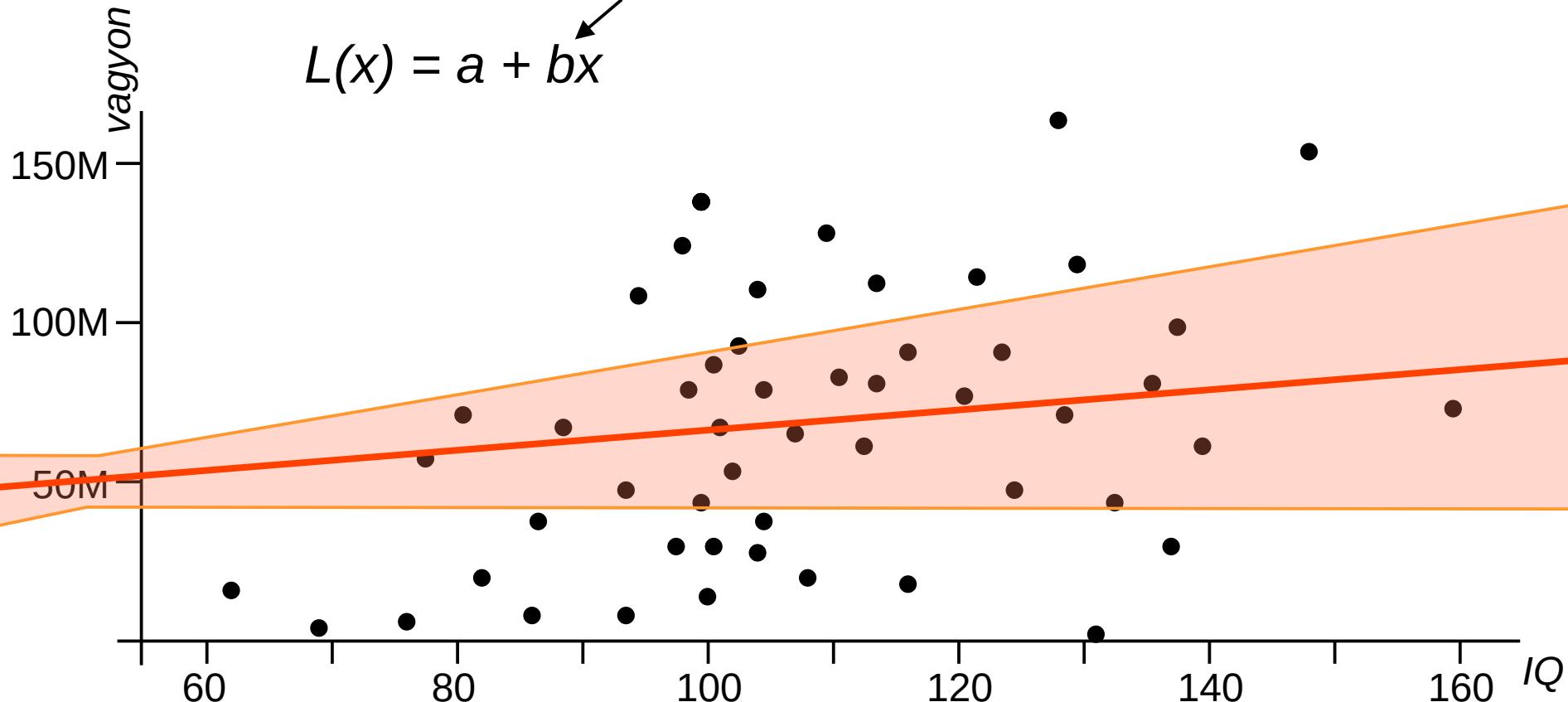


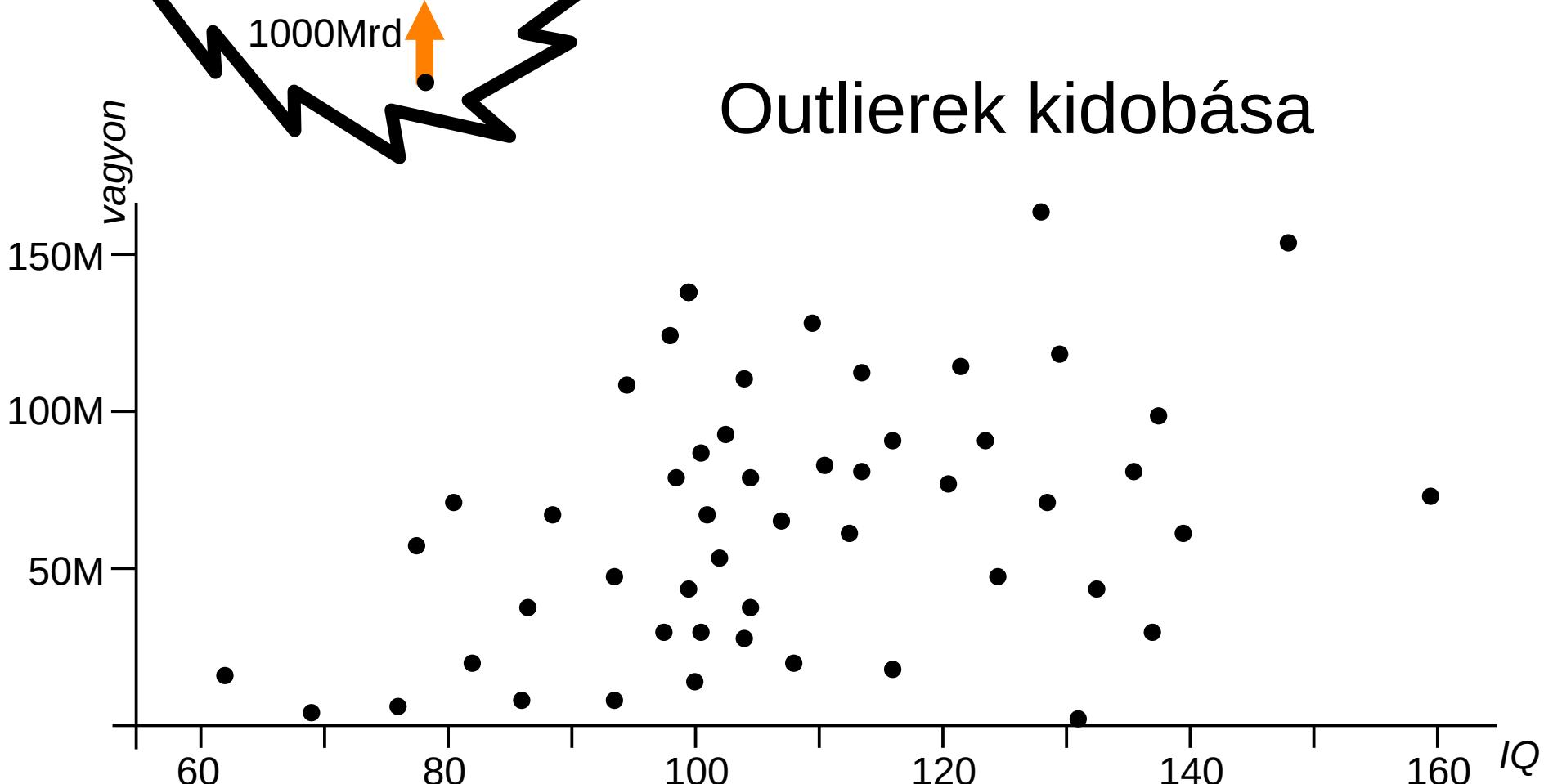
Szignifikánsan eltér 0-tól?

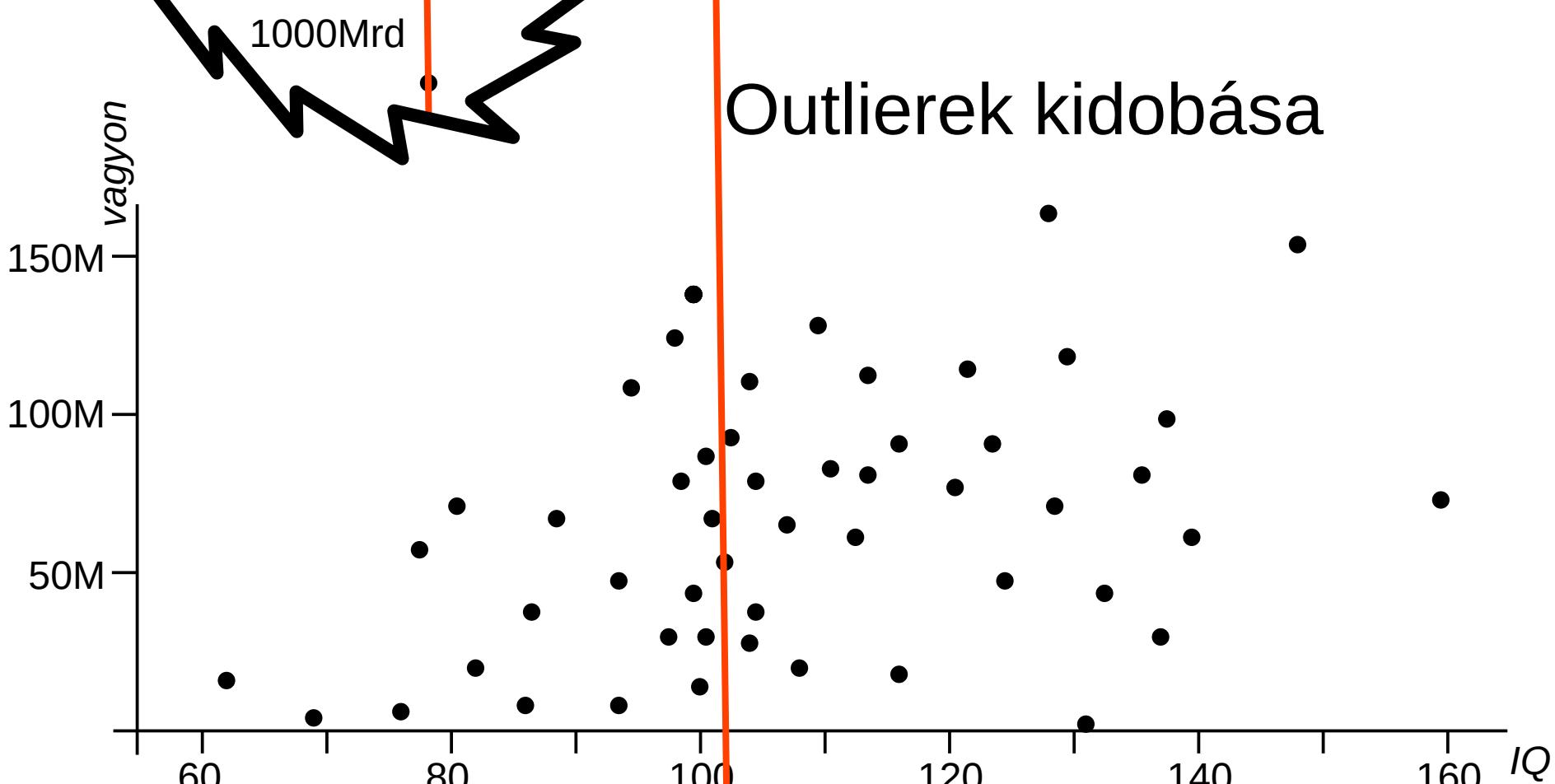
$$L(x) = a + bx$$



Szignifikánsan eltér 0-tól?







1. feladat: várható élettartam elemzése

<https://ruzsaz.github.io/eletartam.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      10     Median      30      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
```

95%-os konfidencia intervallum

Élettartam = $67.4 + 0.00025 * \text{gdp}$

T-próba, $H_0: \text{együttható} = 0$

szignifikancia

illeszkedés várható hibája

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

F-próba, $H_0: \text{minden együttható*} = 0$
(*: a konstans tag nincs beleértve)

Lineáris? regresszió

$$L(x) = a + bx$$

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

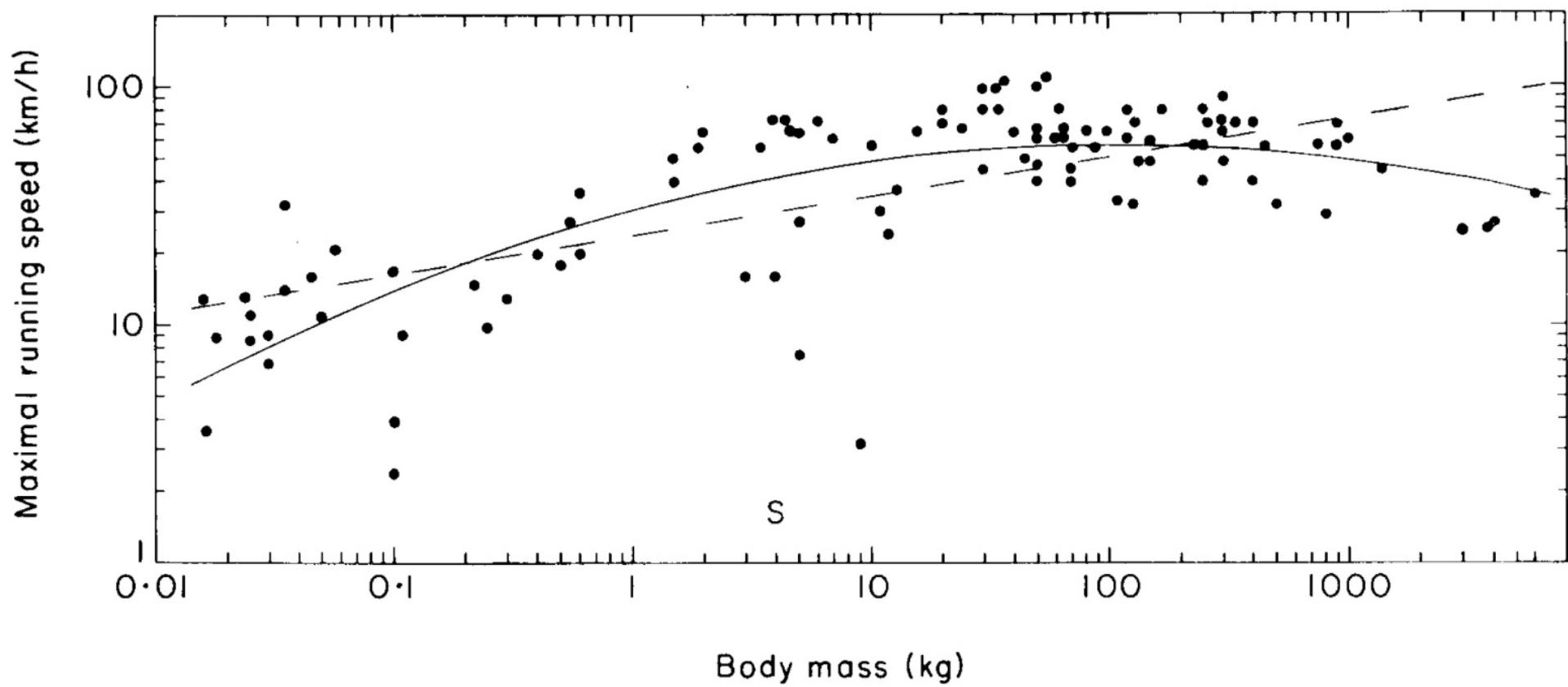
$$Q(x) = a + bx + cx^2 + dx^3$$

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

...



MAXIMAL RUNNING SPEEDS OF MAMMALS



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>

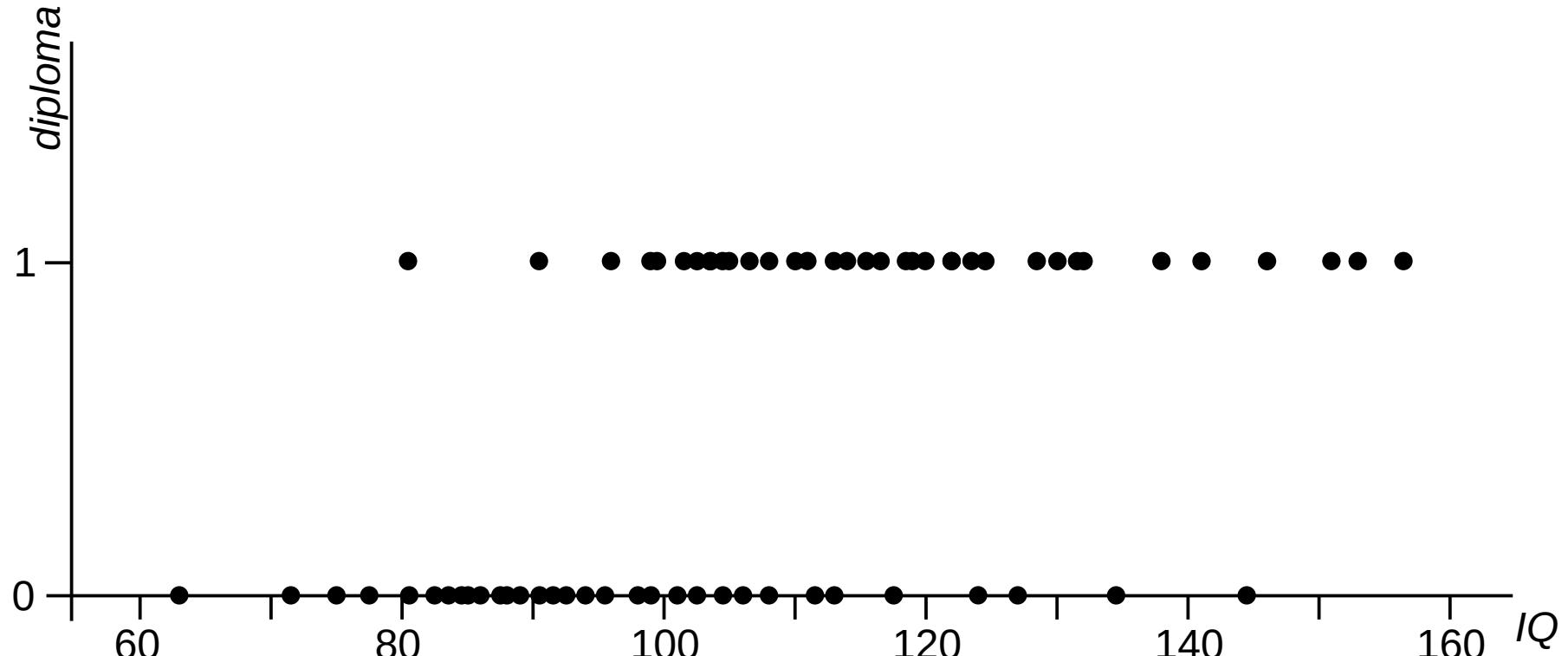
Célváltozó: health (0-100)

Magyarázó változók:

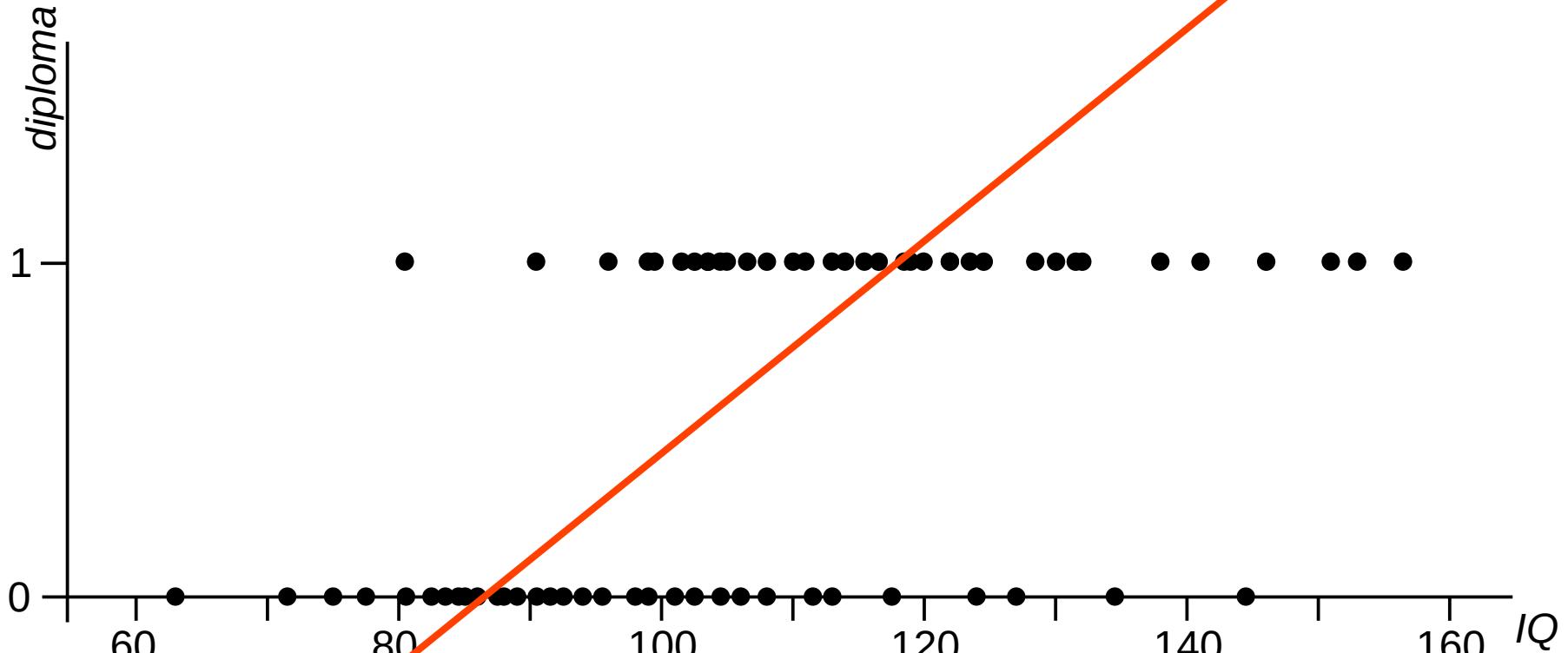
- age, bmi, exercise (0-7), diet (0-100), sleep (óra)
- smoke, sex: (kategória változók: 0, 1)



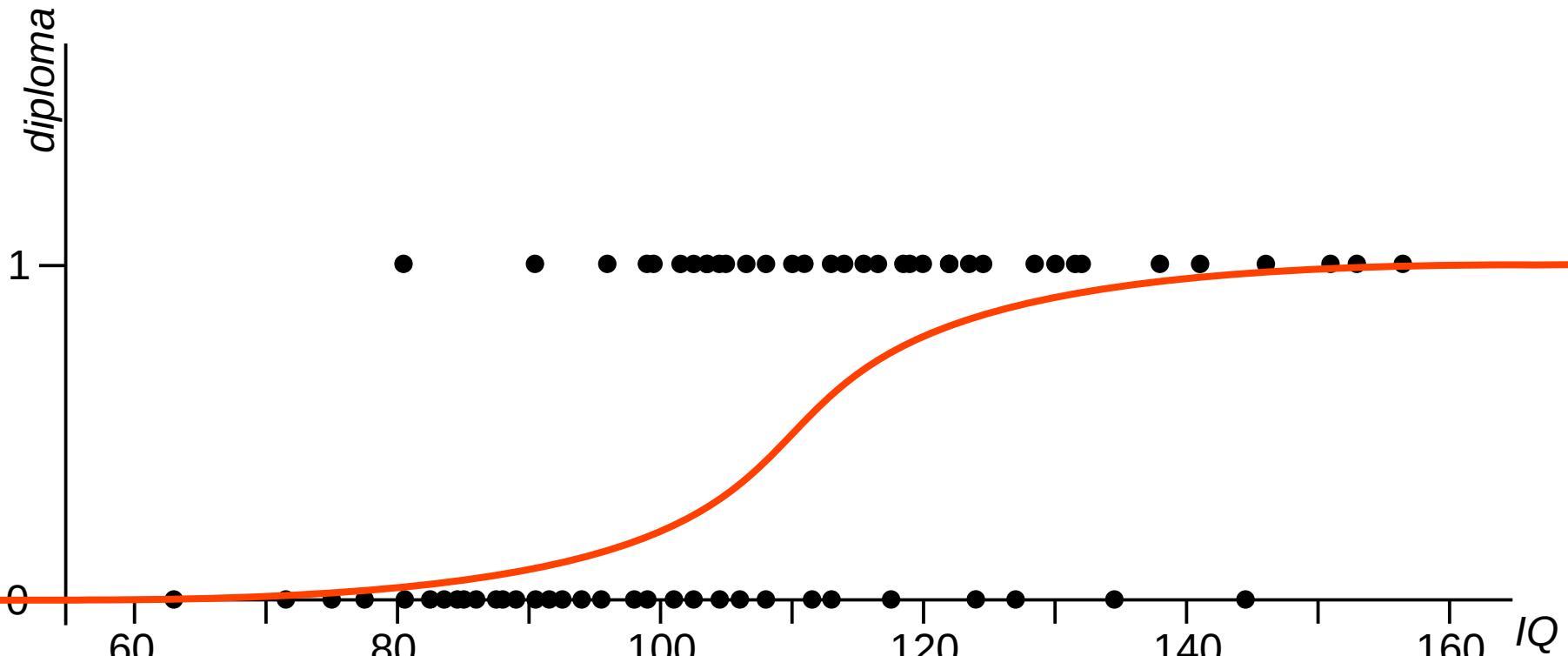
Logisztikus regresszió

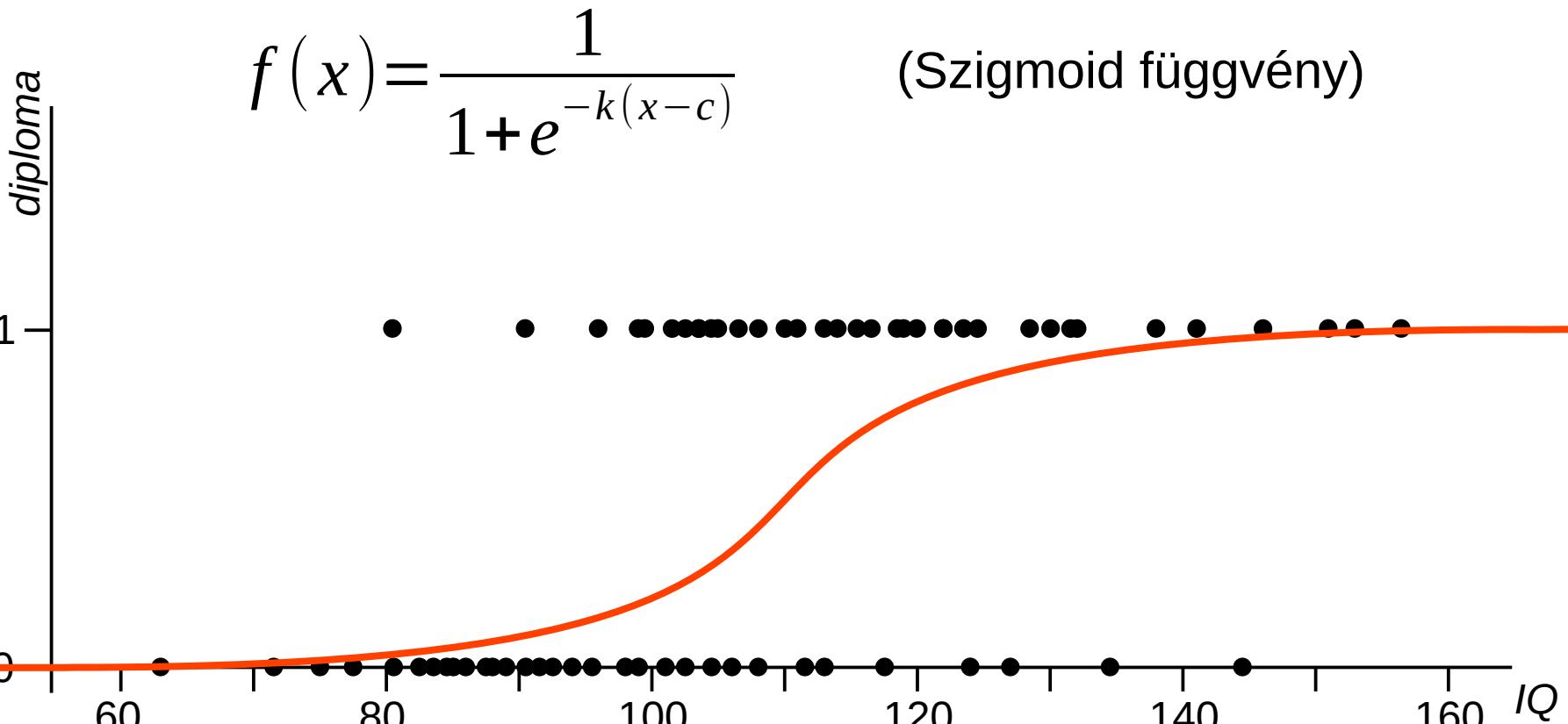


Logisztikus regresszió



Logisztikus regresszió





diploma

$$f(x) = \frac{1}{1 + e^{-k(x - c)}}$$

1

0

60

80

100

120

140

160

IQ



diploma

$$f(x) = \frac{1}{1 + e^{-k(x - c)}}$$

1

60

80

100

120

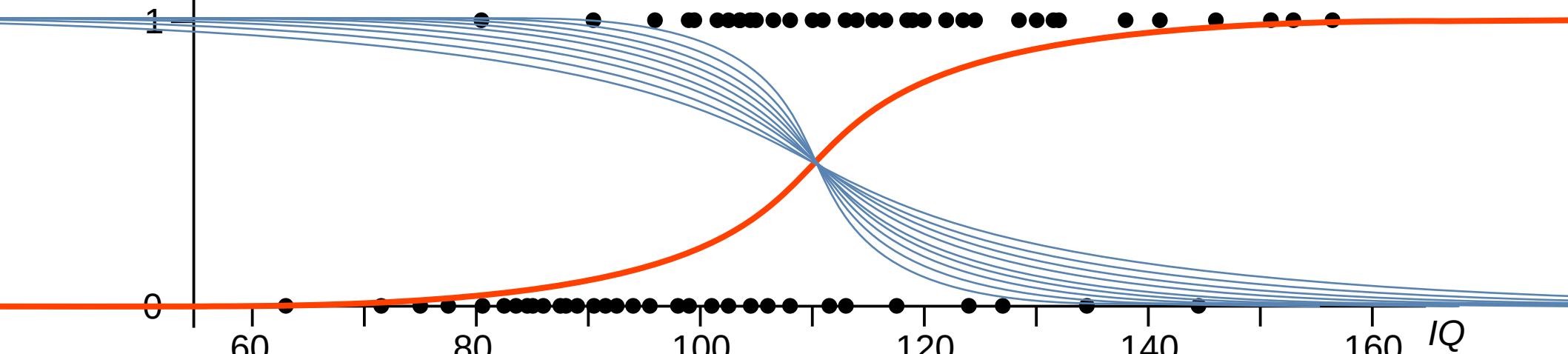
140

160

IQ



$$f(x) = \frac{1}{1 + e^{-k(x - c)}}$$

diploma

$$f(x) = \frac{1}{1 + e^{-k(x - c)}}$$

$[k_0, k_1]$ $[c_0, c_1]$

konfidencia-intervallumok

1

0

60

80

100

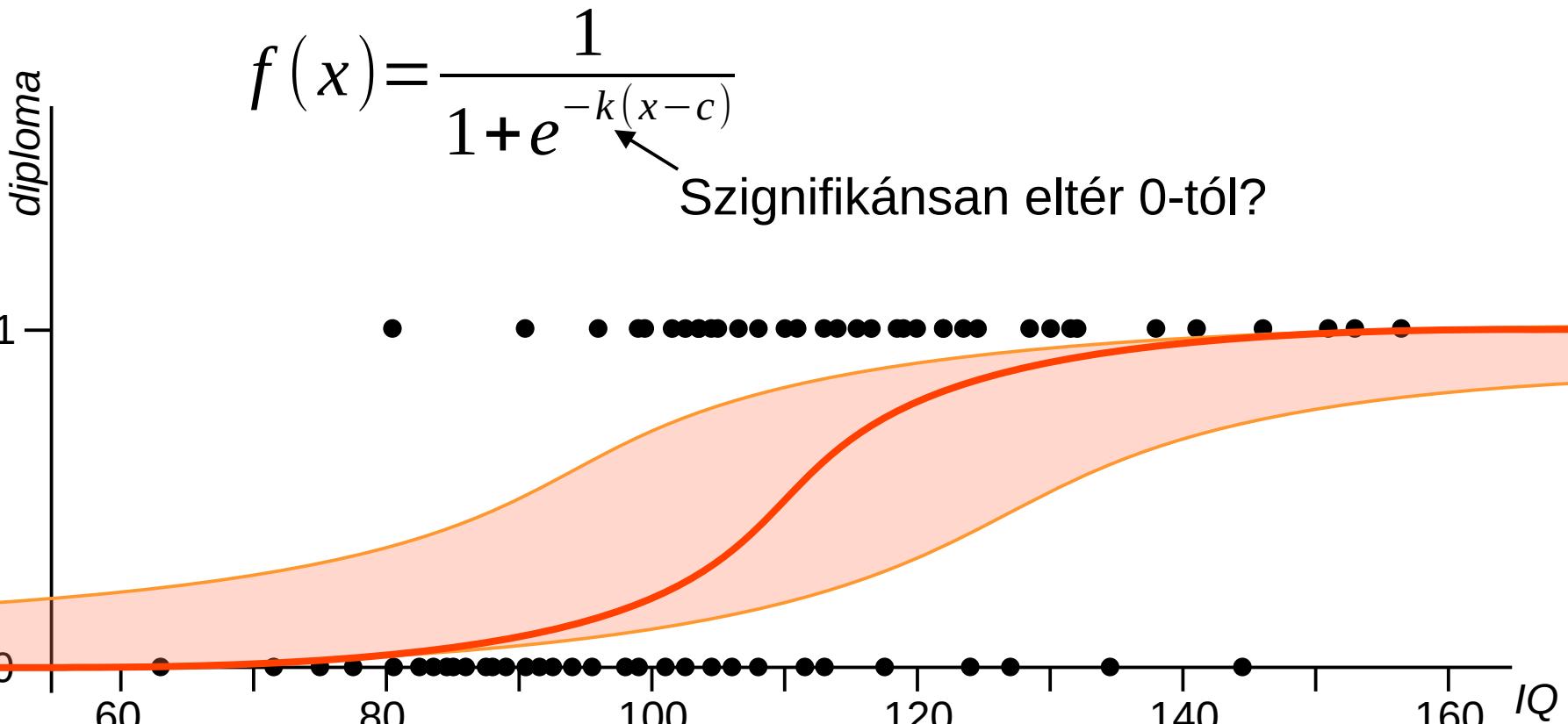
120

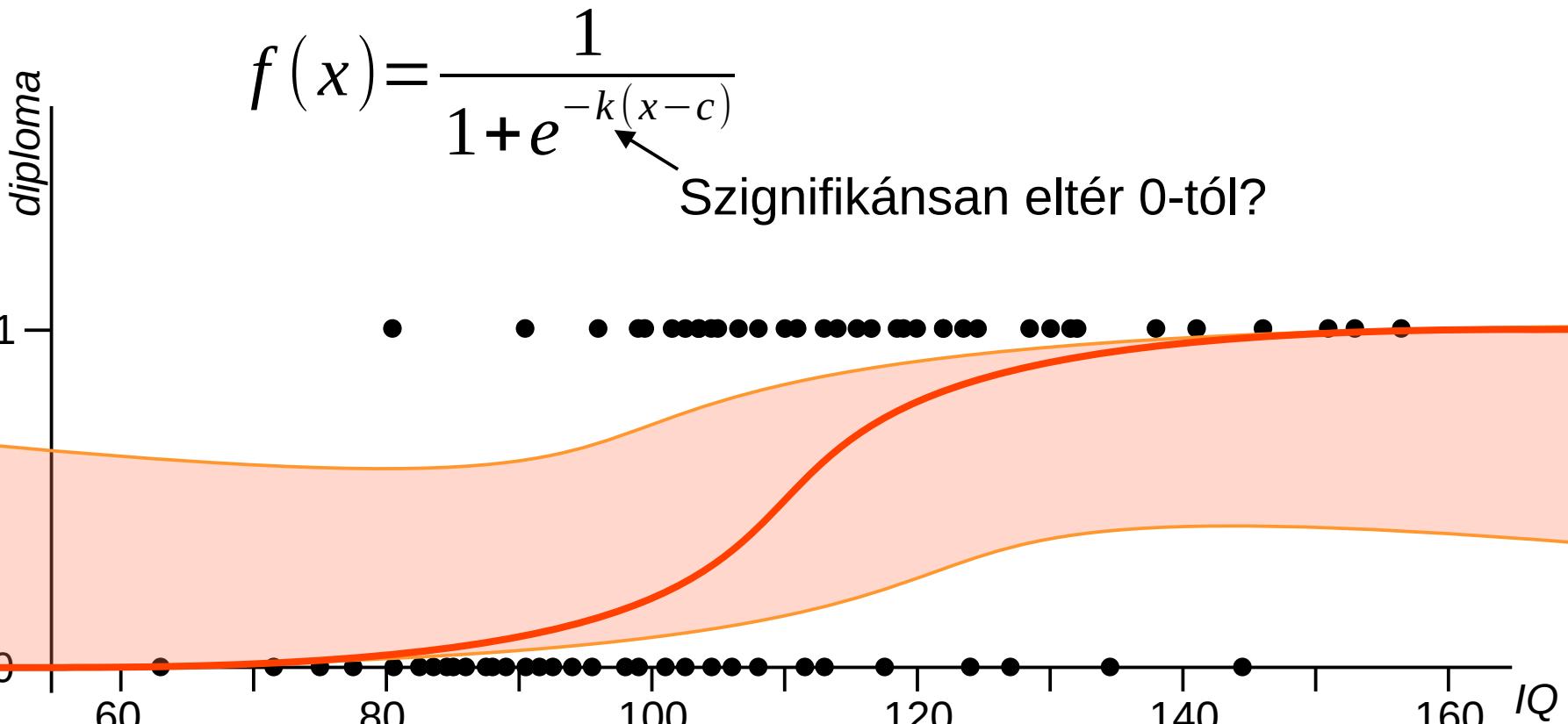
140

160

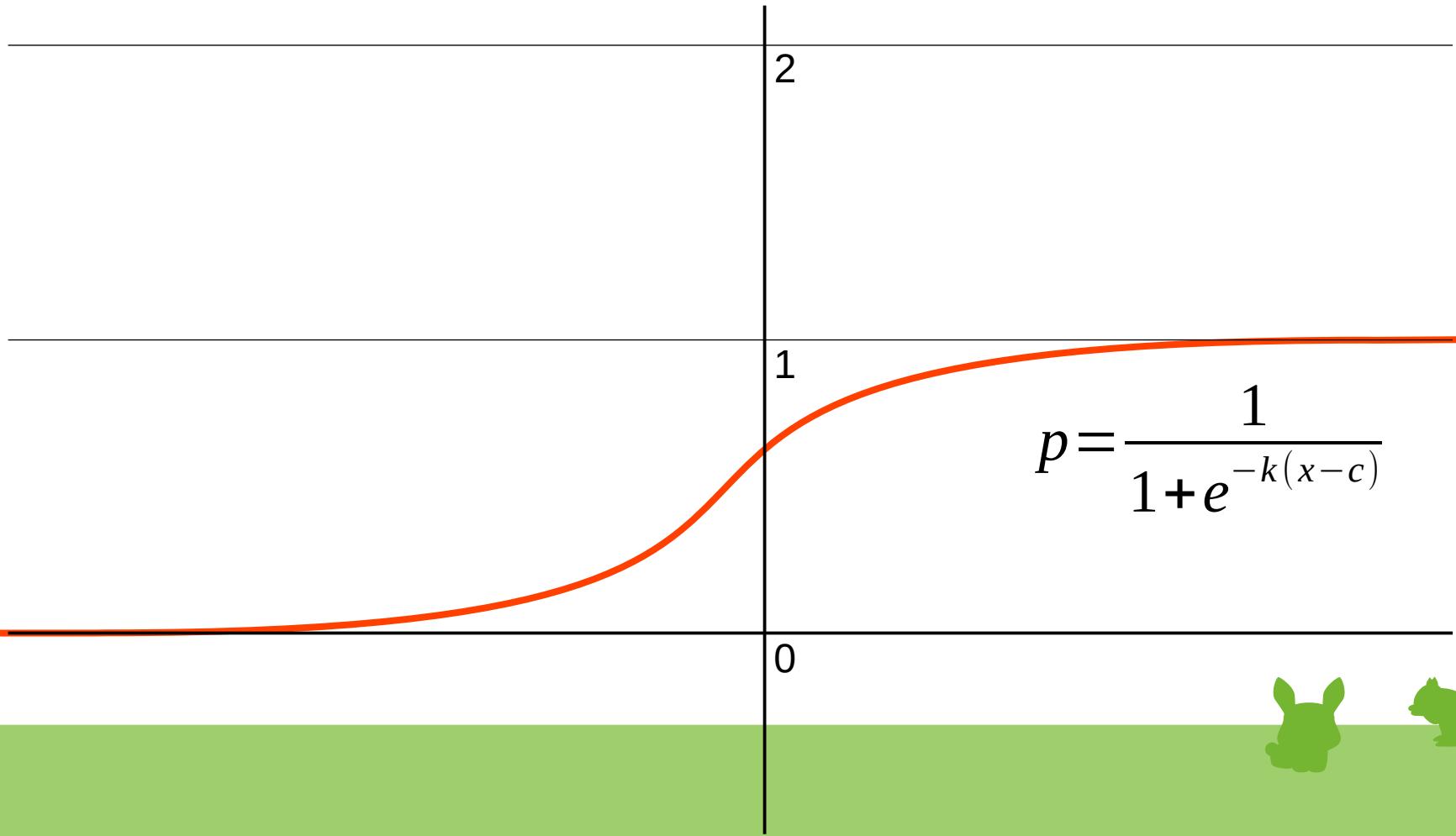
IQ







Mitől logisztikus?



Mitől logisztikus?

$$\frac{p}{1-p} = e^{k(x-c)}$$

$$p = \frac{1}{1 + e^{-k(x-c)}}$$

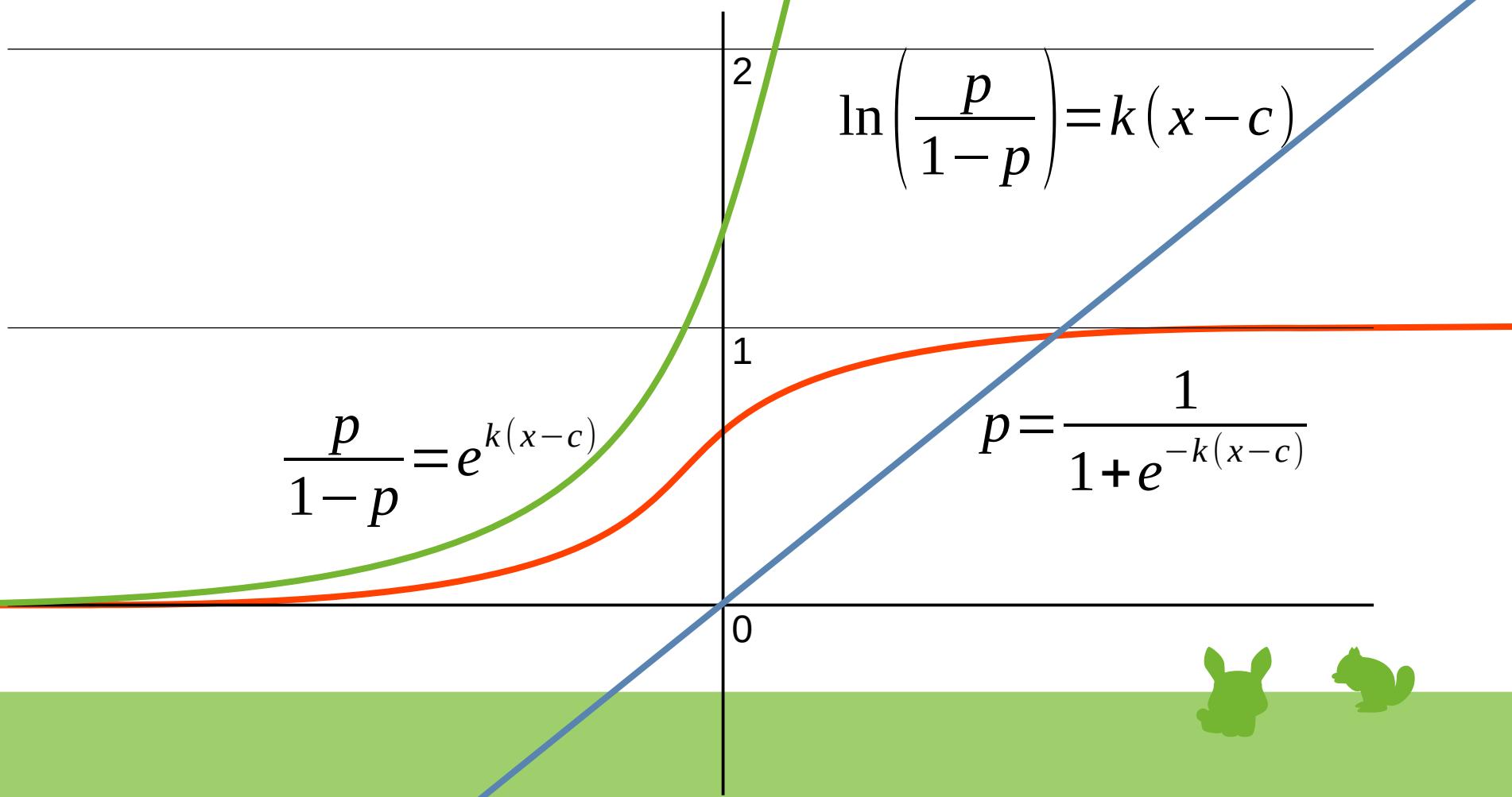
2

1

0



Mitől logisztikus?



Mitől logisztikus?

Logisztikus regresszió p -re

II

Lineáris regresszió $\ln\left(\frac{p}{1-p}\right)$ -re

$$\frac{p}{1-p} = e^{k(x-c)}$$

$$\ln\left(\frac{p}{1-p}\right) = k(x-c)$$

$$p = \frac{1}{1+e^{-k(x-c)}}$$

0

1

2



Logit-függvény

Logisztikus regresszió p -re

II

Lineáris regresszió $\ln\left(\frac{p}{1-p}\right)$ -re

$$\frac{p}{1-p} = e^{k(x-c)}$$

$$\ln\left(\frac{p}{1-p}\right) = k(x-c)$$

$$p = \frac{1}{1+e^{-k(x-c)}}$$

0

1

2



Odds

Logisztikus regresszió p -re

II

Lineáris regresszió $\ln\left(\frac{p}{1-p}\right)$ -re

$$\frac{p}{1-p} = e^{k(x-c)}$$

$$\ln\left(\frac{p}{1-p}\right) = k(x-c)$$

$$p = \frac{1}{1+e^{-k(x-c)}}$$

2

1

0



Valószínűség?

- p
- Kedvező / összes
- $p = 1/5 = 0.2$
 $p = 3/5 = 0.6$
- $[0, 1]$

Odds?

- $p/(1-p)$
- Kedvező : kedvezőtlen
- „1 a 4-hez” = $1 : 4 = 0.25$
„3 a 2-höz” = $3 : 2 = 1.5$
(3x nyer valaki, 2x veszít)
- $[0, \infty]$







ODDS & RETURNS

Fractional	Decimal	Payout*	Fractional	Decimal	Payout*
1/5	1.20	£2.40	7/2	4.50	£9.00
2/5	1.40	£2.80	4/1	5.00	£10.00
1/2	1.50	£3.00	9/2	5.50	£11.00
3/5	1.60	£3.20	5/1	6.00	£12.00
4/5	1.80	£3.60	6/1	7.00	£14.00
1/1	2.00	£4.00	7/1	8.00	£16.00
6/5	2.20	£4.40	8/1	9.00	£18.00
7/5	2.40	£4.80	9/1	10.00	£20.00
6/4	2.50	£5.00	10/1	11.00	£22.00
8/5	2.60	£5.20	15/1	16.00	£32.00
9/5	2.80	£5.60	20/1	22.00	£44.00
2/1	3.00	£6.00	25/1	26.00	£52.00
5/2	3.50	£7.00	30/1	31.00	£62.00
3/1	4.00	£8.00	50/1	51.00	£102.00

* Payouts based on a £2 stake

MOBILE // WINS SPORTS

Valószínűség = p odds = $p/(1-p)$



$$p = 0.67 \quad \text{odds} = 0.67/0.33 = 2:1$$

$$p = 0.5 \quad \text{odds} = 0.5/0.5 = 1:1$$

$$p = 0.33 \quad \text{odds} = 0.33/0.67 = 1:2$$

$$p = 0.25 \quad \text{odds} = 0.25/0.75 = 1:3$$



ODDS & RETURNS

Fractional	Decimal	Payout*	Fractional	Decimal	Payout*
1/5	1.20	£2.40	7/2	4.50	£9.00
2/5	1.40	£2.80	4/1	5.00	£10.00
1/2	1.50	£3.00	9/2	5.50	£11.00
3/5	1.60	£3.20	5/1	6.00	£12.00
4/5	1.80	£3.60	6/1	7.00	£14.00
1/1	2.00	£4.00	7/1	8.00	£16.00
6/5	2.20	£4.40	8/1	9.00	£18.00
7/5	2.40	£4.80	9/1	10.00	£20.00
6/4	2.50	£5.00	10/1	11.00	£22.00
8/5	2.60	£5.20	15/1	16.00	£32.00
9/5	2.80	£5.60	20/1	22.00	£44.00
2/1	3.00	£6.00	25/1	26.00	£52.00
5/2	3.50	£7.00	30/1	31.00	£62.00
3/1	4.00	£8.00	50/1	51.00	£102.00

* Payouts based on a £2 stake

MOBILE // WINS SPORTS

3. feladat: szívbetegség előrejelzése

<https://ruzsaz.github.io/sziv.csv>

- Célváltozó: chd – 10 koronaér-betegség 10 éven belül
(0: nincs, 1: van)
- Magyarázó változó: cholesterol – koleszterin szint

Végezzük el a logisztikus regressziót, rajzoljuk le az eredményt!



```
# prompt: use logistic regression to explain "chd" from "cholesterol"

# Fit the logistic regression model
model <- glm(chd ~ cholesterol, data = df, family = "binomial")

# Summarize the model
summary(model)

# Calculate odds ratios and confidence intervals
exp(cbind(OR = coef(model), confint(model)))
```

```
Call:
glm(formula = chd ~ cholesterol, family = "binomial", data = df)

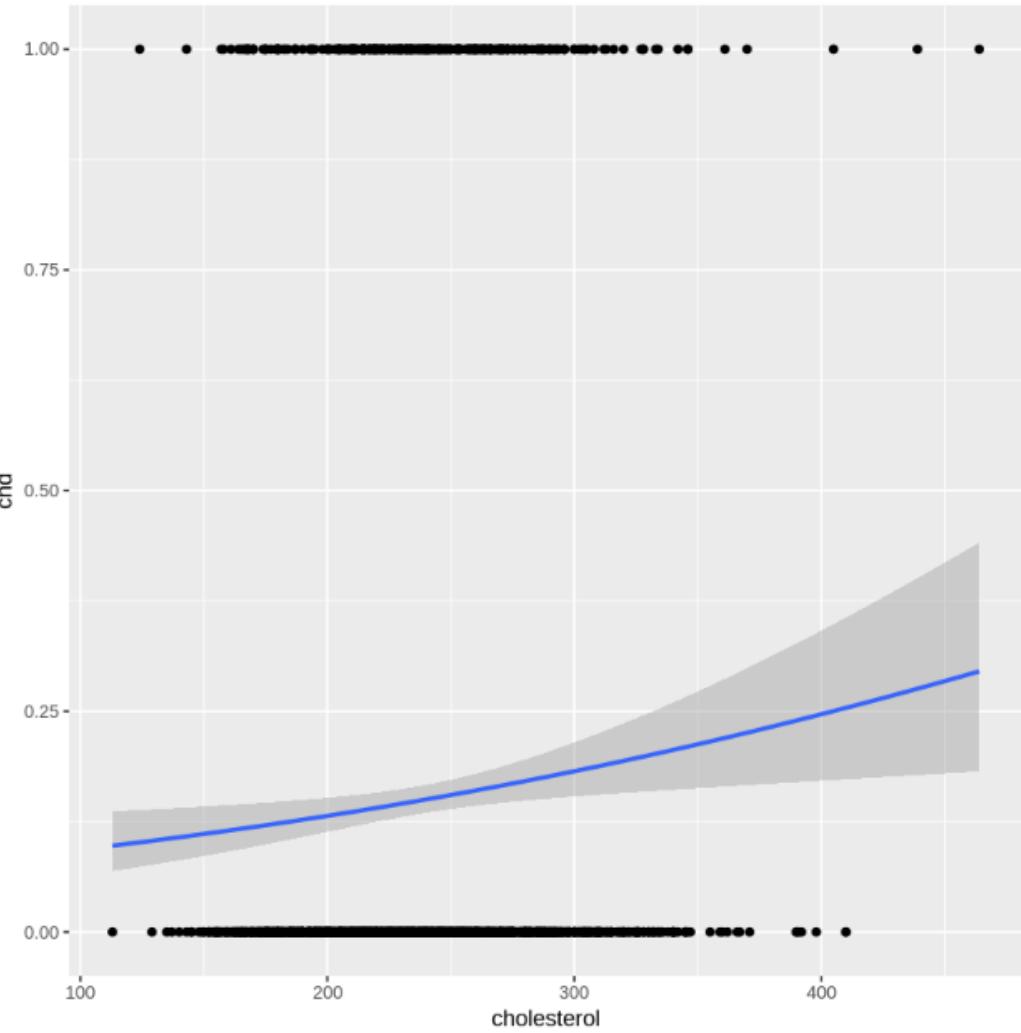
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.658310   0.348271 -7.633  2.3e-14 ***
cholesterol  0.003854   0.001420   2.713  0.00666 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1687.4 on 1999 degrees of freedom
Residual deviance: 1680.1 on 1998 degrees of freedom
AIC: 1684.1

Number of Fisher Scoring iterations: 4
Waiting for profiling to be done...
```

	A matrix: 2 × 3 of type dbl		
	OR	2.5 %	97.5 %
(Intercept)	0.07006656	0.03531783	0.1384529
cholesterol	1.00386139	1.00105543	1.0066495



```
# prompt: use logistic regression to explain "chd" from "cholesterol"

# Fit the logistic regression model
model <- glm(chd ~ cholesterol, data = df, family = "binomial")

# Summarize the model
summary(model)

# Calculate odds ratios and confidence intervals
exp(cbind(OR = coef(model), confint(model)))
```

```
Call:
glm(formula = chd ~ cholesterol, family = "binomial", data = df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.658310  0.348271 -7.633 2.3e-14 ***
cholesterol   0.003854  0.001420   2.713 0.00666 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1687.4 on 1999 degrees of freedom
Residual deviance: 1680.1 on 1998 degrees of freedom
AIC: 1684.1

Number of Fisher Scoring iterations: 4
Waiting for profiling to be done...
```

```
A matrix: 2 × 3 of type dbl
      OR    2.5 %   97.5 %
(Intercept) 0.07006656  0.03531783  0.1384529
cholesterol 1.00386139 1.00105543  1.0066495
```

1-el nő a cholesterol, 0.003854-el nő a log odds

Z-teszt a szignifikanciára

Model illeszkedése a null-modelhez képest

Akaike Information Criterion
(minél kisebb, annál több információt tartalmaz a model)

1-el nő a cholesterol, 1.0038 szorosára változik az odds



Háromszorosára növeli az elhízás kockázatát a házasság a férfiaknál, de a feleségükknél nem

Dóka Boglárka | 2025.03.13. 13:40



Egy friss lengyel kutatás szerint a házasság jelentősen növeli az elhízás esélyét a férfiaknál, míg a nőknél nem figyelhető meg hasonló hatás. A kutatás rámutat arra is, hogy az életkor előrehaladtával mindenkor nem esetében nő az elhízás kockázata.

A kutatásban 2405, átlagosan 50 éves résztvevő egészségügyi adatait elemezték.

Az eredmények szerint a házas férfiak 3,2-szer nagyobb eséllyel válnak elhízottá, mint a nőtlen férfiak.

Ezzel szemben a házasság nem növelte a nőknél az elhízás kockázatát.

A lengyel kutatás azt is kimutatta, hogy az életkor előrehaladtával mindenkor nemnél nő a túlsúly és az elhízás kockázata. A férfiaknál évente 3 százalékkal nő a túlsúly, míg az elhízás kockázata 4 százalékkal emelkedik.

A nőknél ezek az arányok még magasabbak: a túlsúly esélye évente 4 százalékkal, az elhízásé pedig 6 százalékkal növekszik.

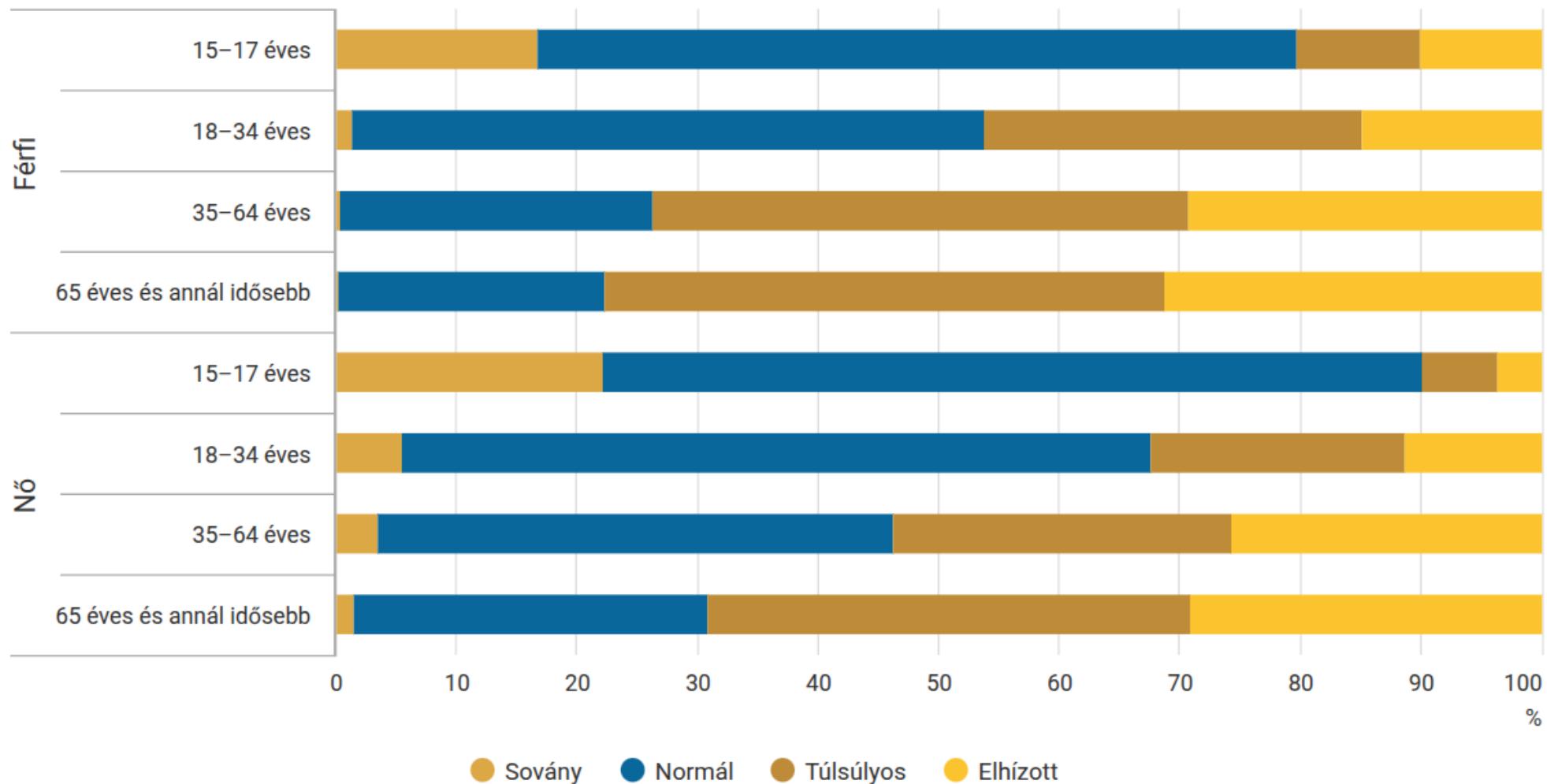
Érdekesség, hogy bizonyos tényezők kizárolag a nőknél növelték az elhízás kockázatát. A depresszió megduplázta, míg az egészségügyi ismeretek hiánya 43 százalékkal növelte az elhízás esélyét. Emellett a kisebb településeken élő nők körében is gyakoribb volt az elhízás, míg ezek a tényezők a férfiakat nem érintették hasonló mértékben.

<https://hirado.hu/extra/eltmod/cikk/2025/03/13/háromszorosára-noveli-az-elhízás-kockázatát-a-házasság-kockázatát-a-férfiaknál-de-a-feleségüknel-nem>

Hivatkozás:

[https://www.thelancet.com/journals/lancet/article/PII-S0140-6736\(25\)00397-6/fulltext](https://www.thelancet.com/journals/lancet/article/PII-S0140-6736(25)00397-6/fulltext)

Tápláltság nem és korcsoport szerint, 2019





Háromszorosára növeli az elhízás kockázatát a házasság a férfiaknál, de a feleségüknél nem

Dóka Boglárka | 2025.03.13. 13:40



Egy friss lengyel kutatás szerint a házasság jelentősen növeli az elhízás esélyét a férfiaknál, míg a nőknél nem figyelhető meg hasonló hatás. A kutatás rámutat arra is, hogy az életkor előrehaladtával mindenkorban nő az elhízás kockázata.

A kutatásban 2405, átlagosan 50 éves résztvevő egészségügyi adatait elemezték.

Az eredmények szerint a házas férfiak 3,2-szer nagyobb eséllyel válnak elhízottá, mint a nőtlen férfiak.

Ezzel szemben a házasság nem növelte a nőknél az elhízás kockázatát.

A lengyel kutatás azt is kimutatta, hogy az életkor előrehaladtával mindenkorban nő a túlsúly és az elhízás kockázata. A férfiaknál évente 3 százalékkal nő a túlsúly, míg az elhízás kockázata 4 százalékkal emelkedik.

A nőknél ezek az arányok még magasabbak: a túlsúly esélye évente 4 százalékkal, az elhízásé pedig 6 százalékkal növekszik.

Érdekkesség, hogy bizonyos tényezők kizárolag a nőknél növelték az elhízás kockázatát. A depresszió megduplázta, míg az egészségügyi ismeretek hiánya 43 százalékkal növelte az elhízás esélyét. Emellett a kisebb településeken élő nők körében is gyakoribb volt az elhízás, míg ezek a tényezők a férfiakat nem érintették hasonló mértékben.

Férfi:

17 éves: 10%

+50 év: $*(1.04)^{50} = 71\%$

+házas: $*3.2 = 227.2\%$

Nő:

17 éves: 4%

+50 év: $*(1.06)^{50} = 73.6\%$

+depi: $*2 = 147.2\%$

+buta: $*1.43 = 210.5\%$

4. feladat: szívbetegség előrejelzése

<https://ruzsaz.github.io/sziv.csv>

- Célváltozó: chd – 10 koronaér-betegség (0: nincs, 1: van)
- Magyarázó változók
 - Katagória:
male, education, smoker, cigarettes, hypertension, diabetes
 - Számszerű:
age, cigarettes, cholesterol, bloodpressure, bmi, heartrate
- Végezzük el az elemzést!

