REGRESNÍ ANALÝZA

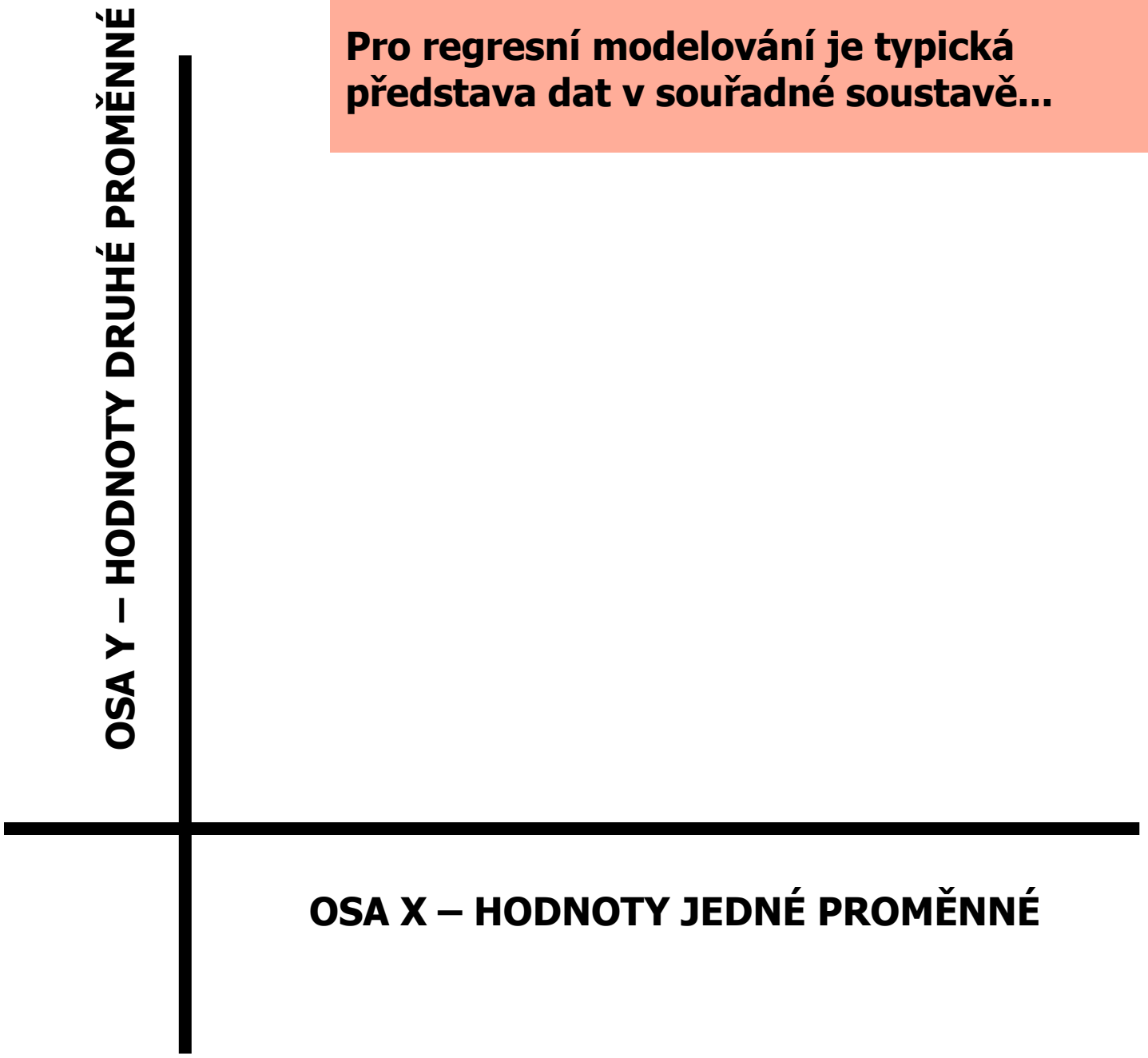
Příklad:

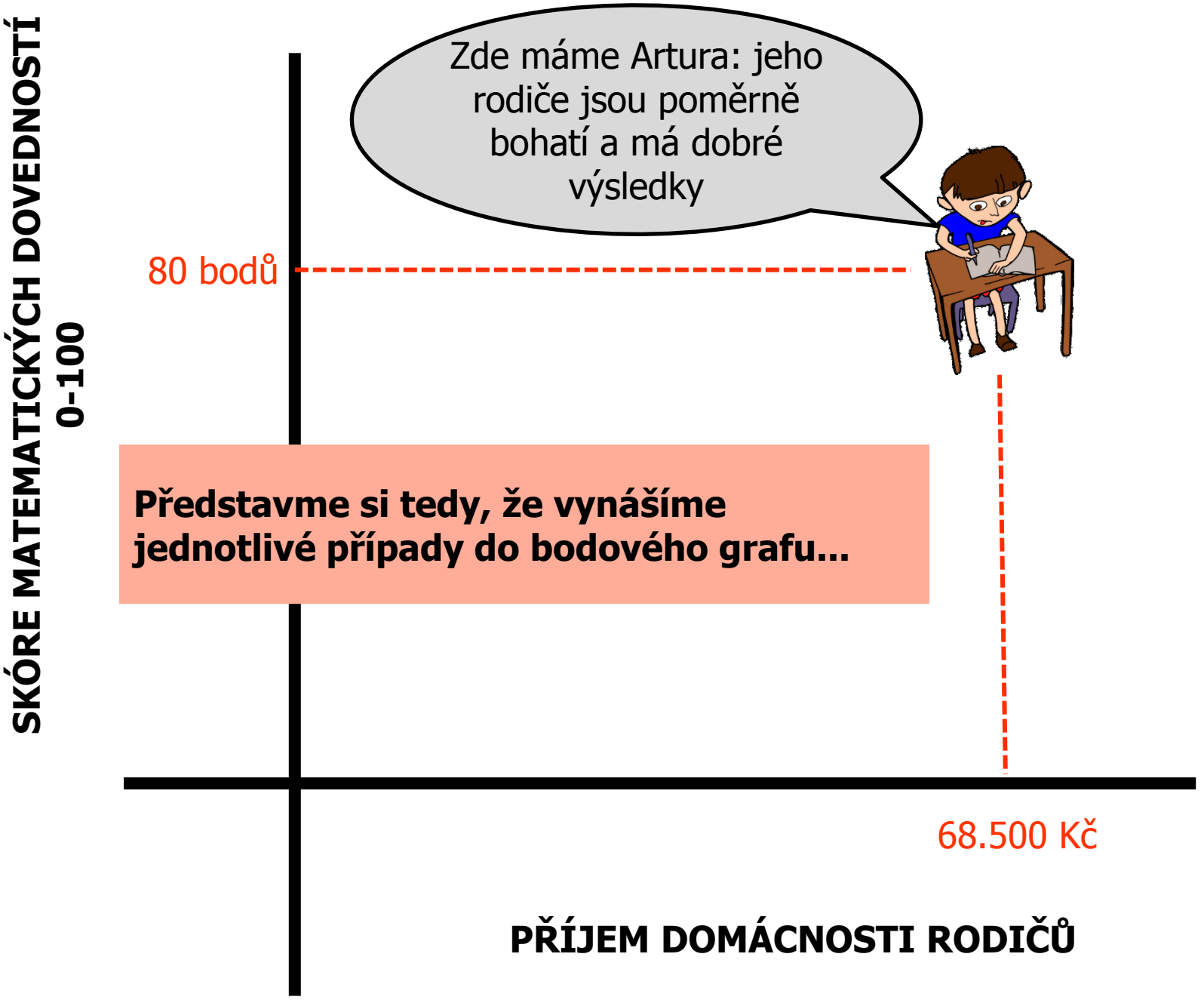
Z dat PISA zjišťuji vztah mezi:

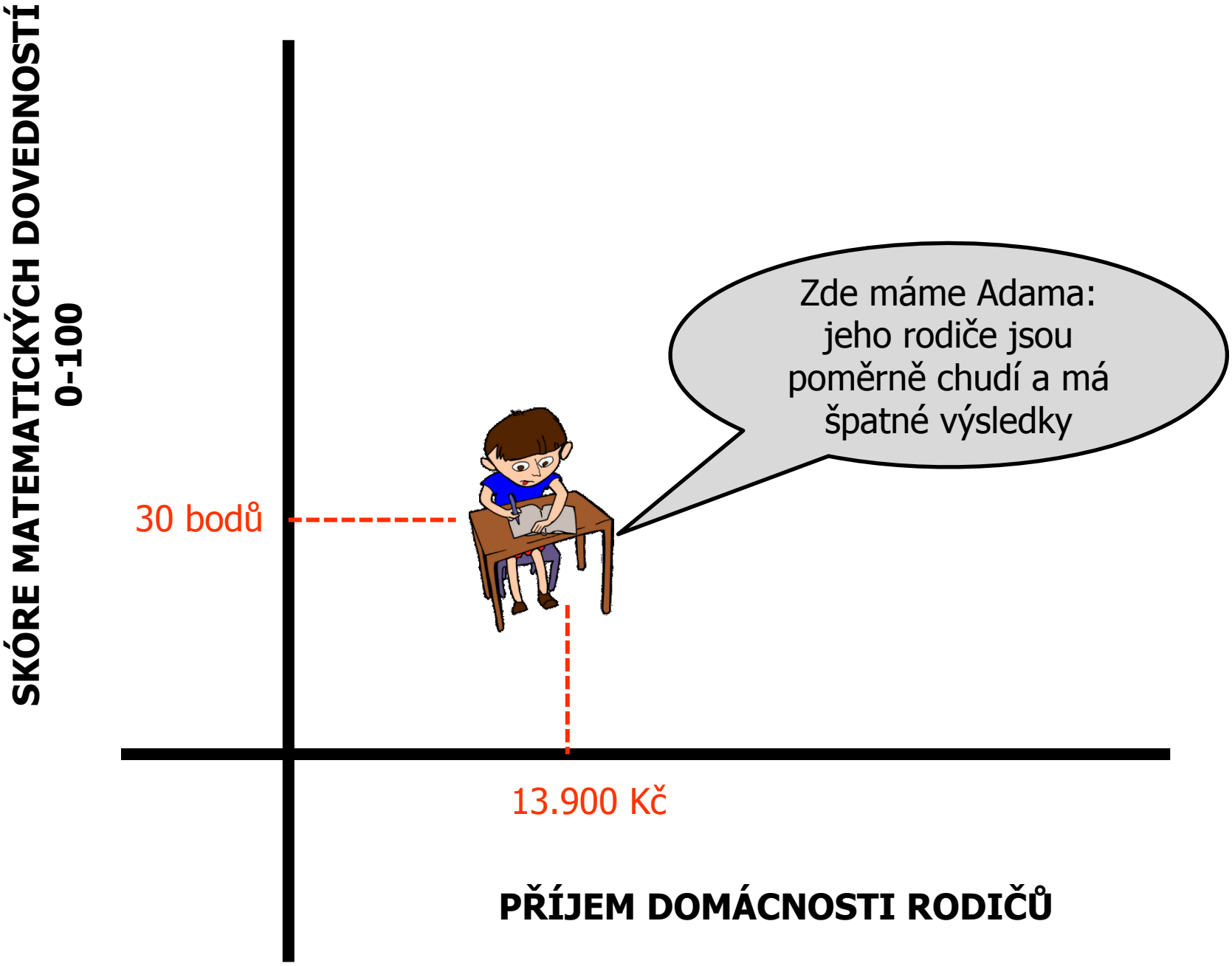
matematickým skóre žáků
a
úrovní příjmů domácnosti v níž žijí

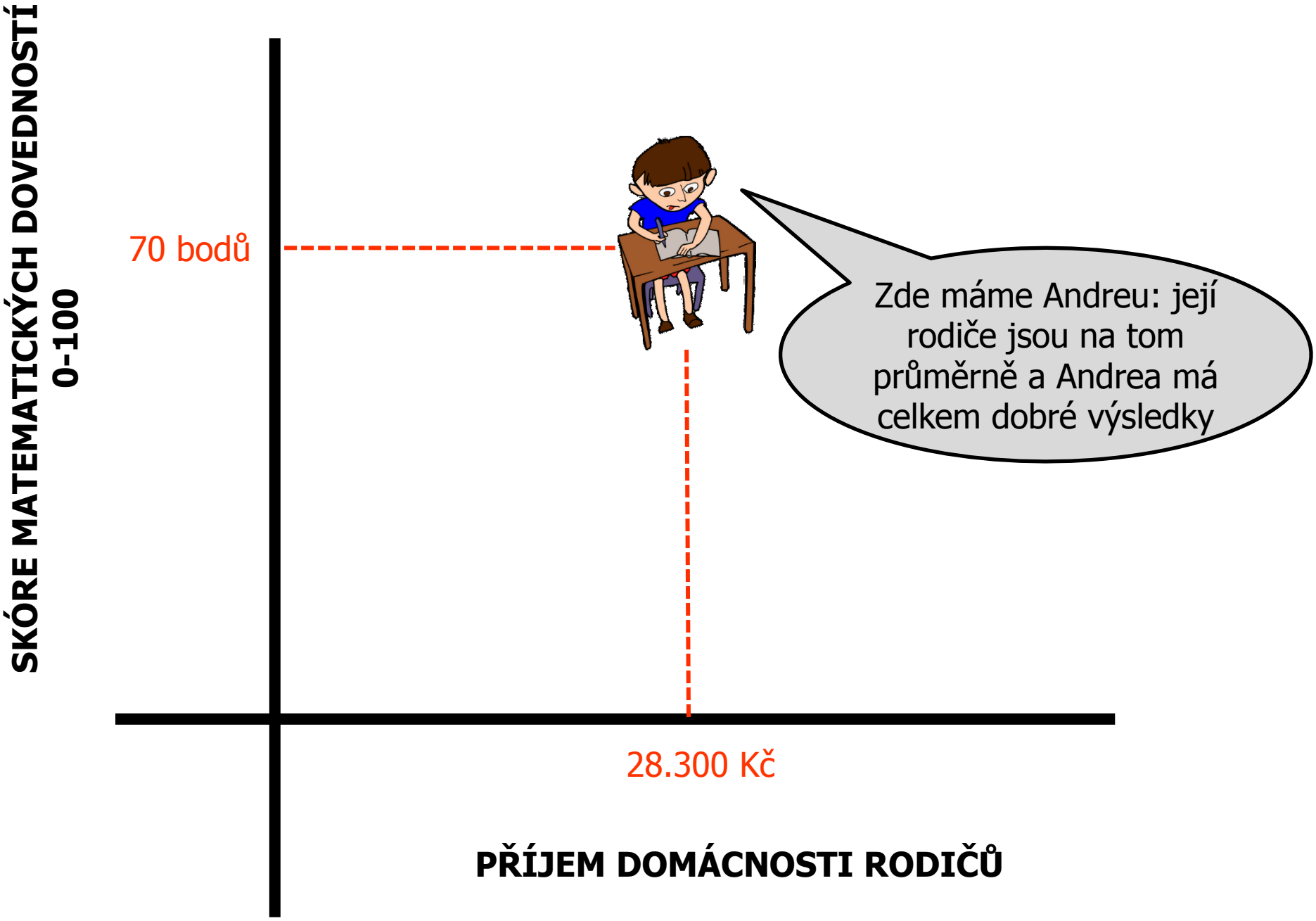
/pro přehlednost je vše v tomto příkladu vymyšleno a souvislost mezi příjmem a výkony žáků není zdaleka tak silná

Pro regresní modelování je typická představa dat v souřadné soustavě...




SKÓRE MATEMATICKÝCH DOVEDNOSTÍ
0-100
Zde máme Artura: jeho rodiče jsou poměrně bohatí a má dobré výsledky
80 bodů
Představme si tedy, že vynášíme jednotlivé případy do bodového grafu...
68.500 Kč
PŘÍJEM DOMÁCNOSTI RODIČŮ


SKÓRE MATEMATICKÝCH DOVEDNOSTÍ
0-100
Zde máme Adama:
jeho rodiče jsou
poměrně chudí a má
špatné výsledky
30 bodů
13.900 Kč
PŘÍJEM DOMÁCNOSTI RODIČŮ


SKÓRE MATEMATICKÝCH DOVEDNOSTÍ
0-100
70 bodů
Zde máme Andreu: její rodiče jsou na tom průměrně a Andrea má celkem dobré výsledky
28.300 Kč
PŘÍJEM DOMÁCNOSTI RODIČŮ

SKÓRE MATEMATICKÝCH DOVEDNOSTÍ
0-100

andrea

artur

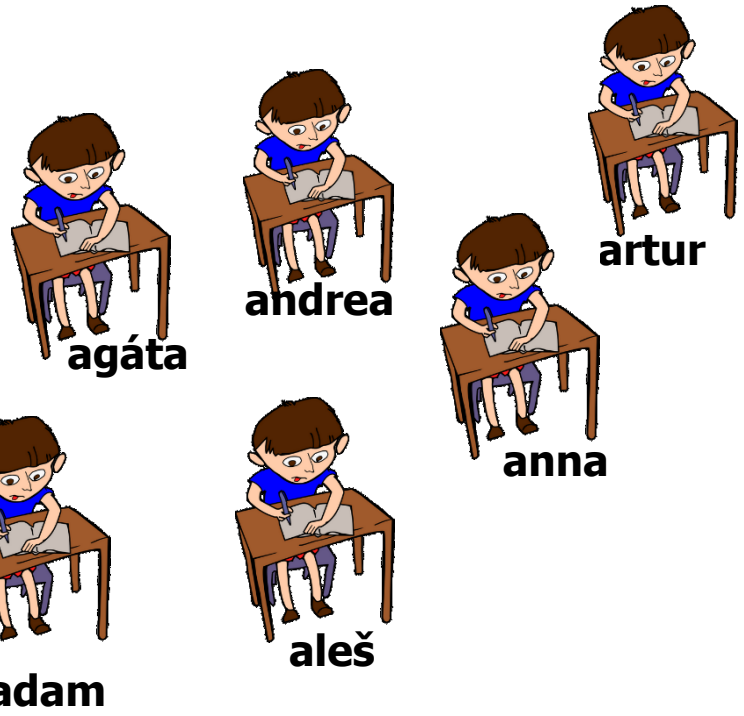
adam

Takto vypadají uspořádaná data za tři žáky...

PŘÍJEM DOMÁCNOSTI RODIČŮ

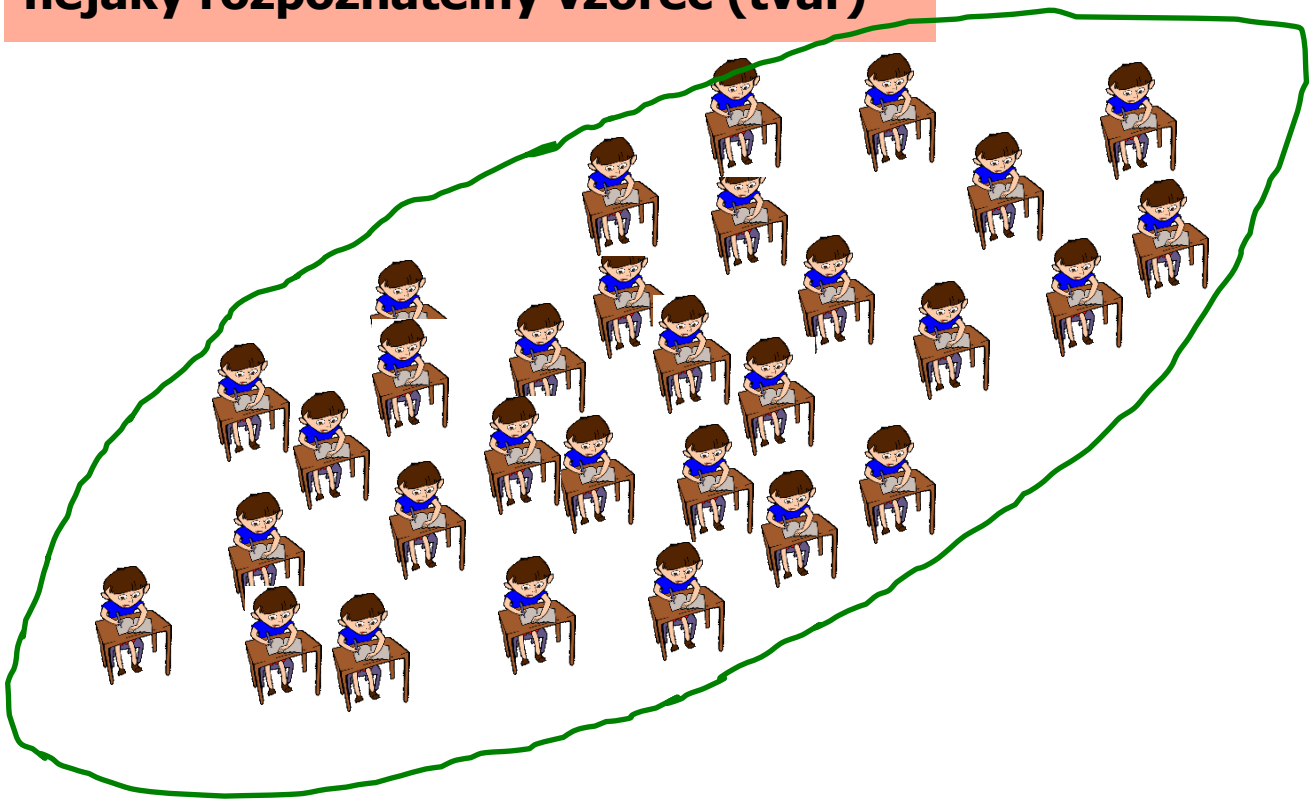
Přidáváme další...

SKÓRE MATEMATICKÝCH DOVEDNOSTÍ
0-100



PŘÍJEM DOMÁCNOSTI RODIČŮ

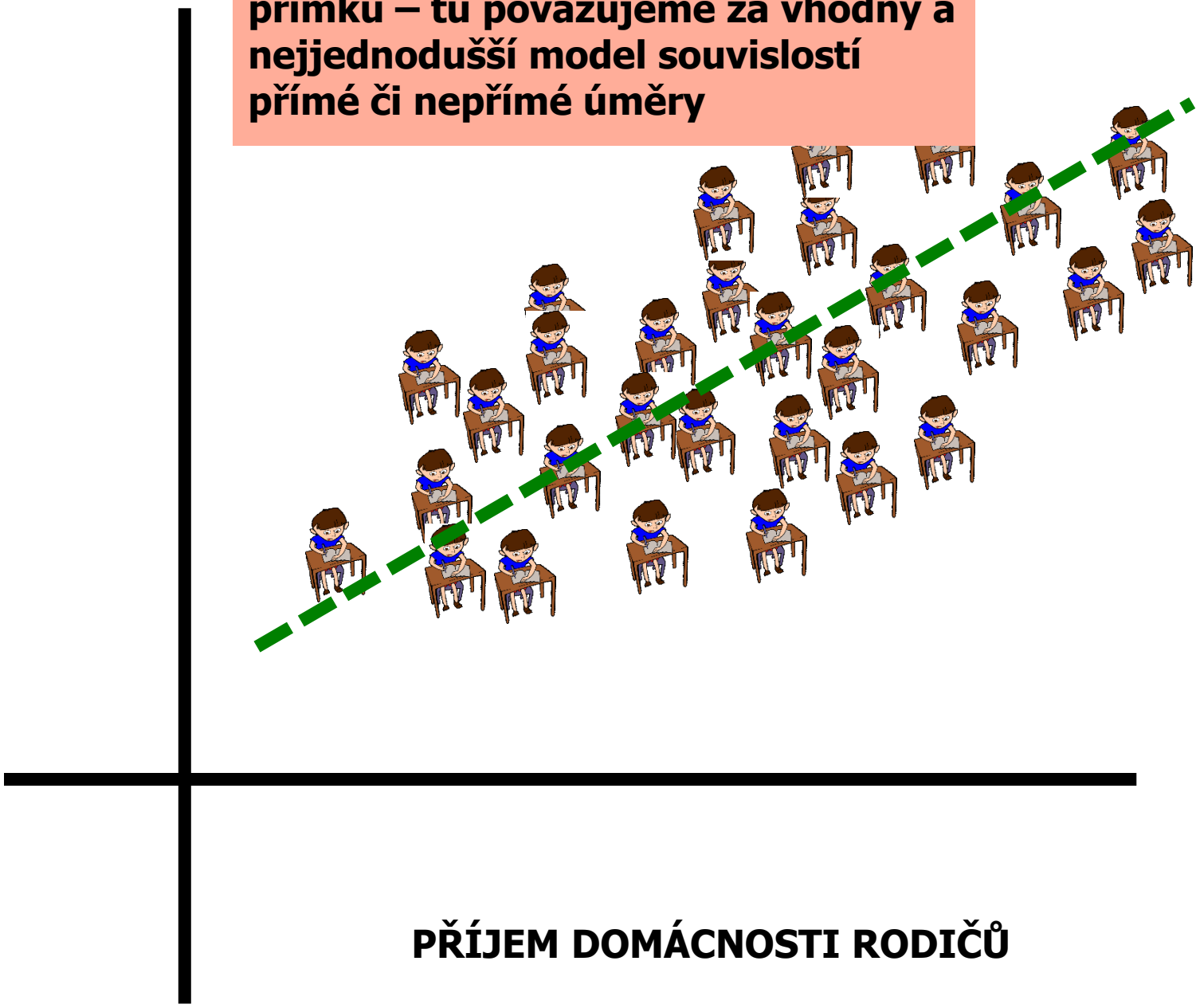
a další...
A sledujeme, jestli v grafu vidíme nějaký rozpoznatelný vzorec (tvar)



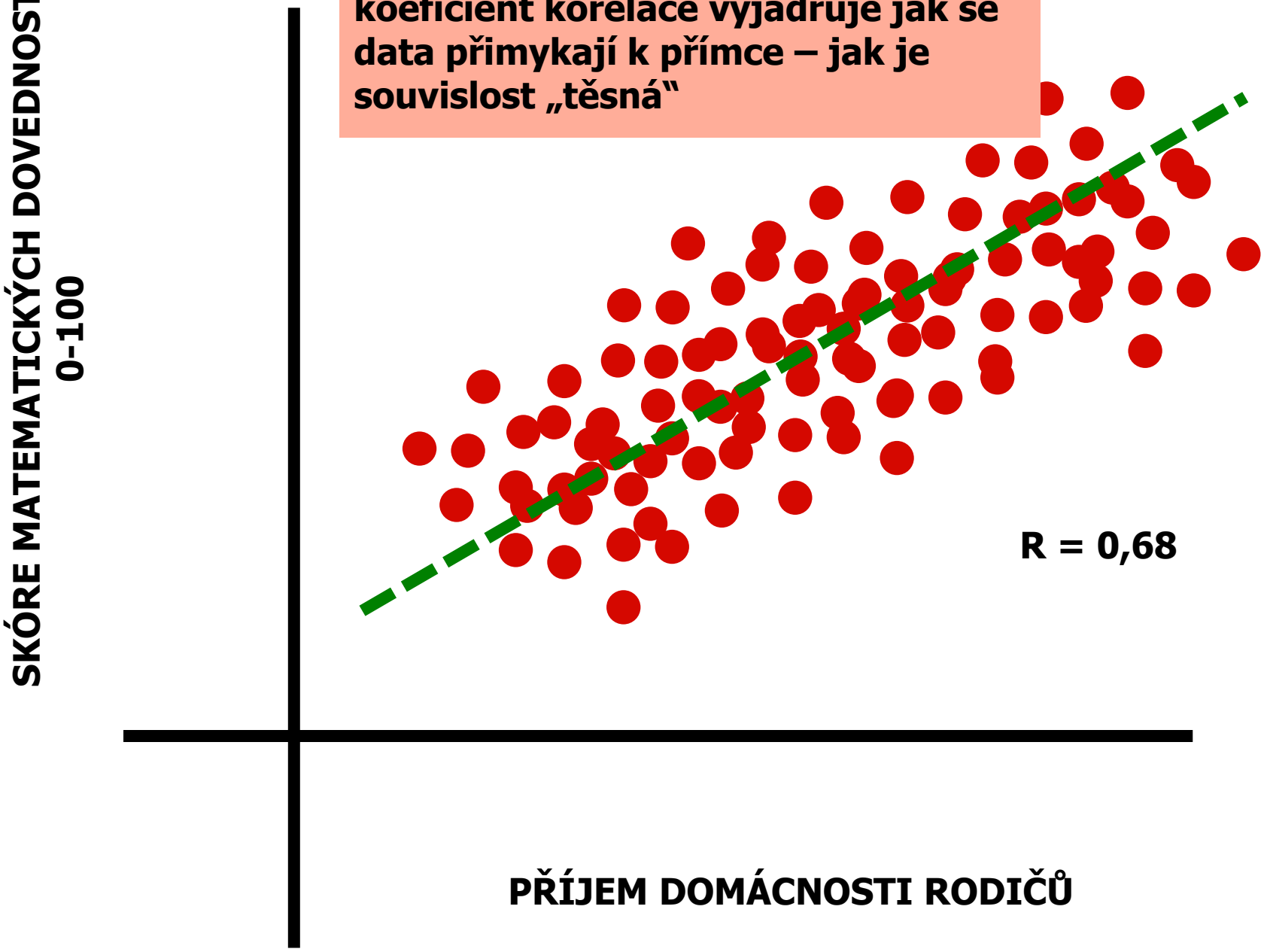
SKÓRE MATEMATICKÝCH DOVEDNOSTÍ 0-100

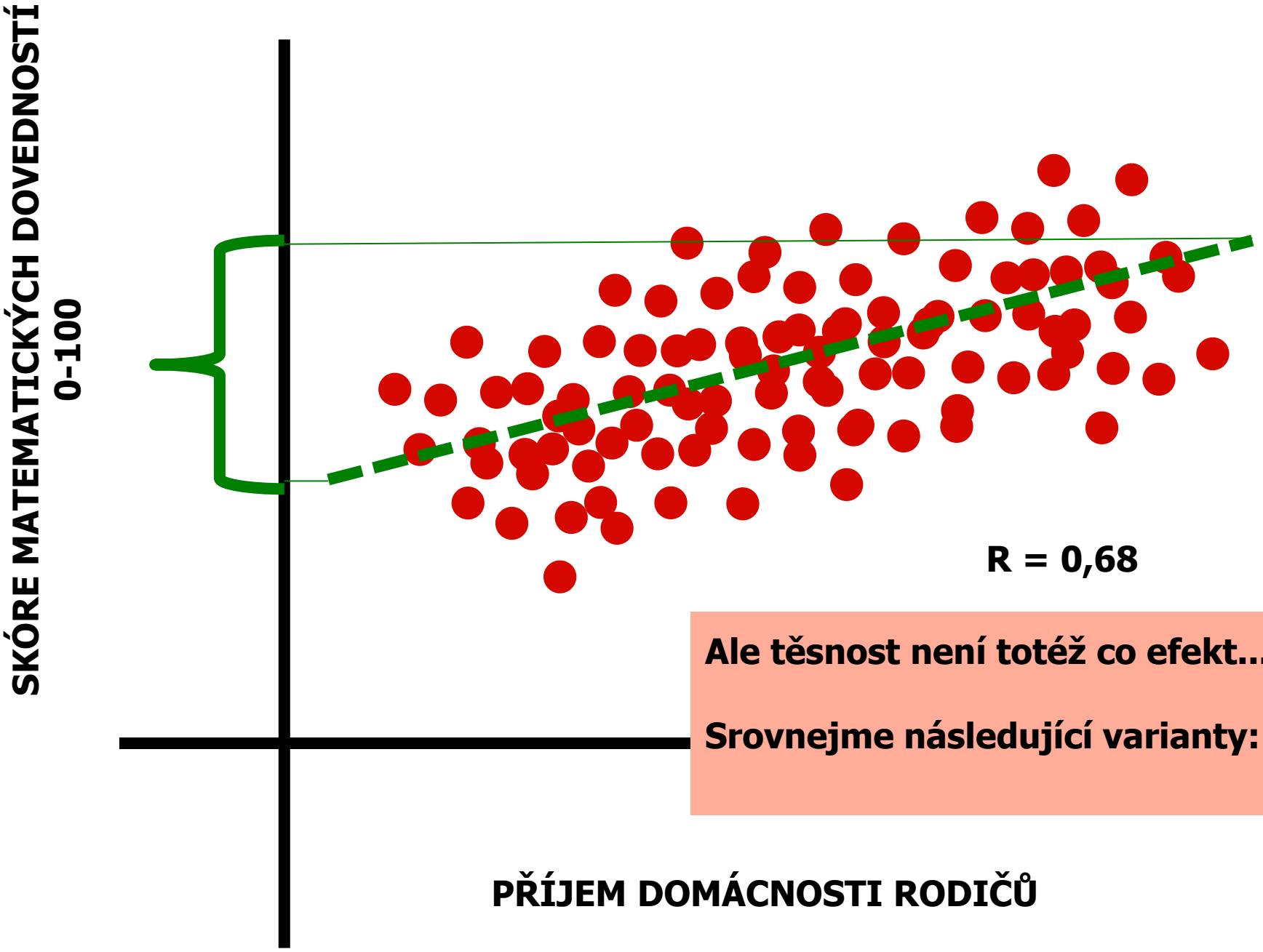
PŘÍJEM DOMÁCNOSTI RODIČŮ

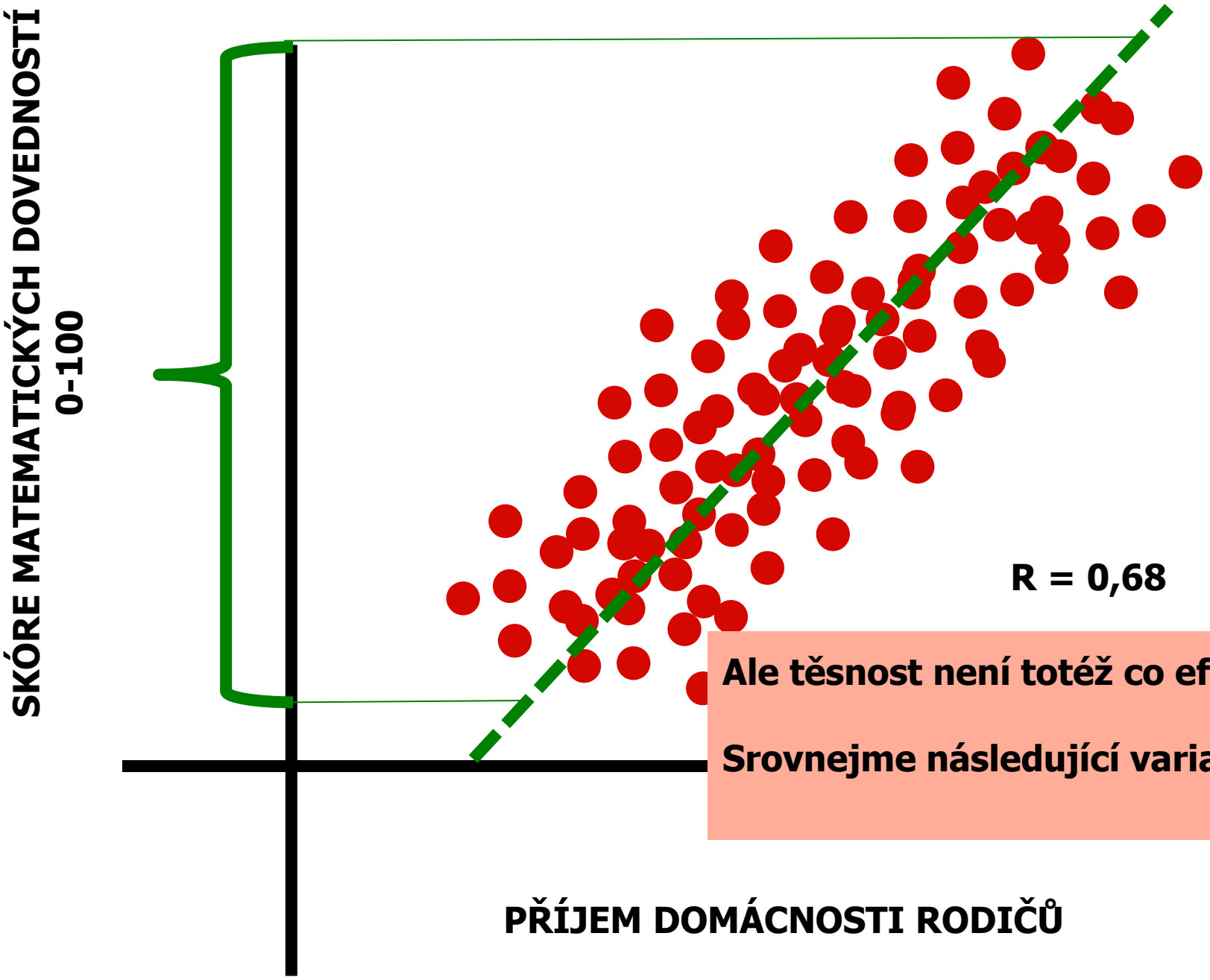
Úkolem regrese je proložit daty přímku – tu považujeme za vhodný a nejjednodušší model souvislostí přímé či nepřímé úměry



Tohle už známe z korelací – koeficient korelace vyjadřuje jak se data přimykají k přímce – jak je souvislost „těsná“

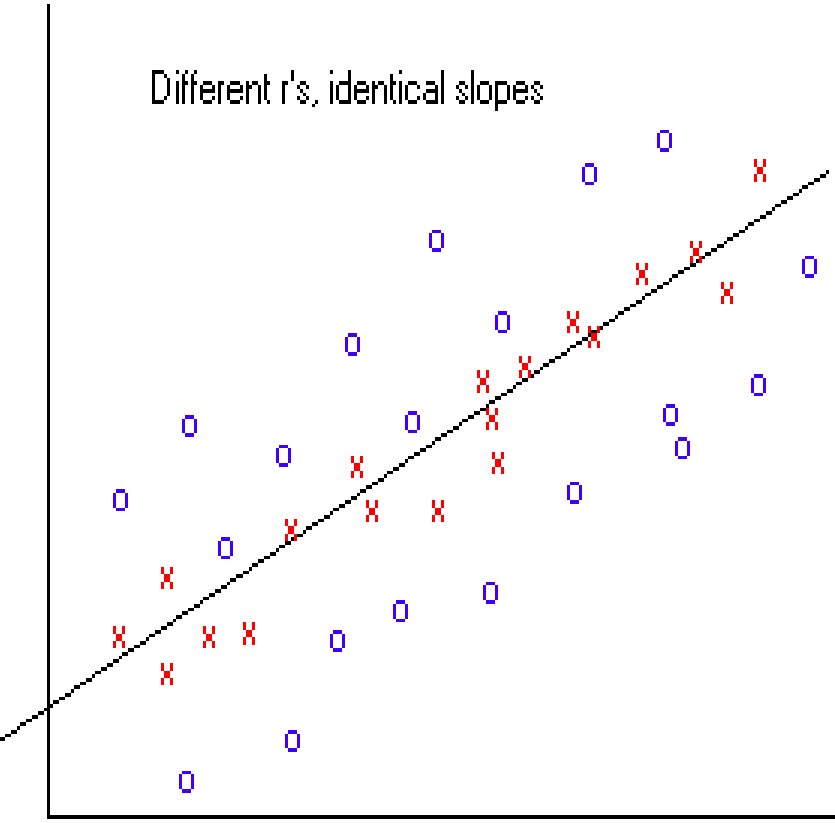


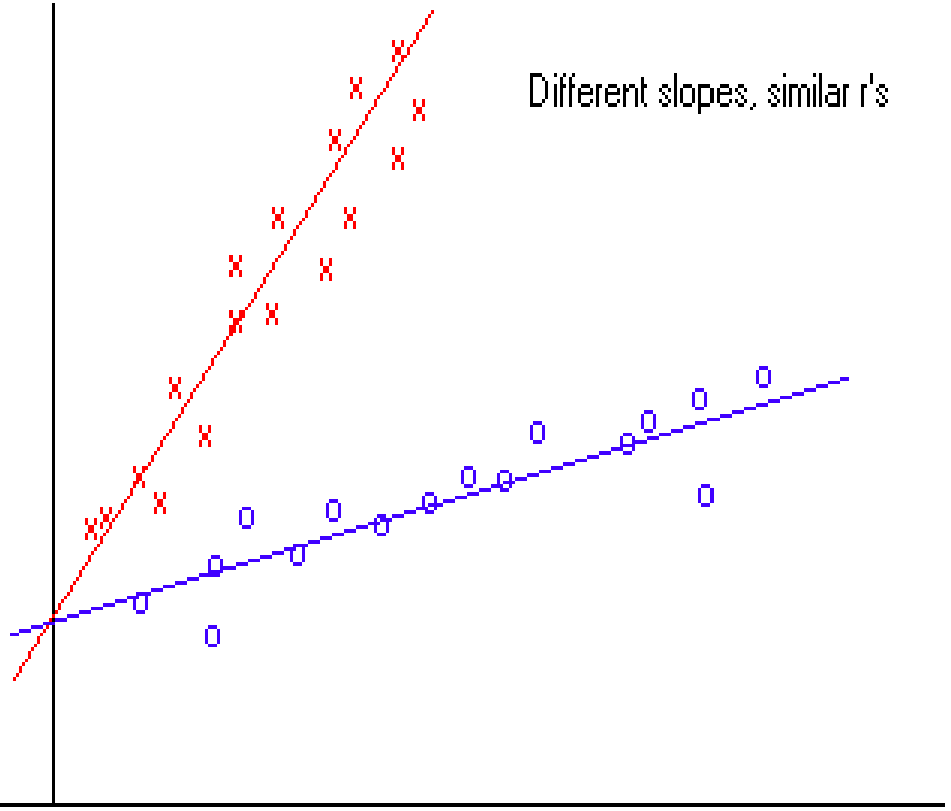

SKÓRE MATEMATICKÝCH DOVEDNOSTÍ
0-100
R = 0,68
Ale těsnost není totéž co efekt…
Srovnejme následující varianty:
PŘÍJEM DOMÁCNOSTI RODIČŮ


SKÓRE MATEMATICKÝCH DOVEDNOSTÍ
0-100
R = 0,68
Ale těsnost není totéž co efekt…
Srovnejme následující varianty:
PŘÍJEM DOMÁCNOSTI RODIČŮ

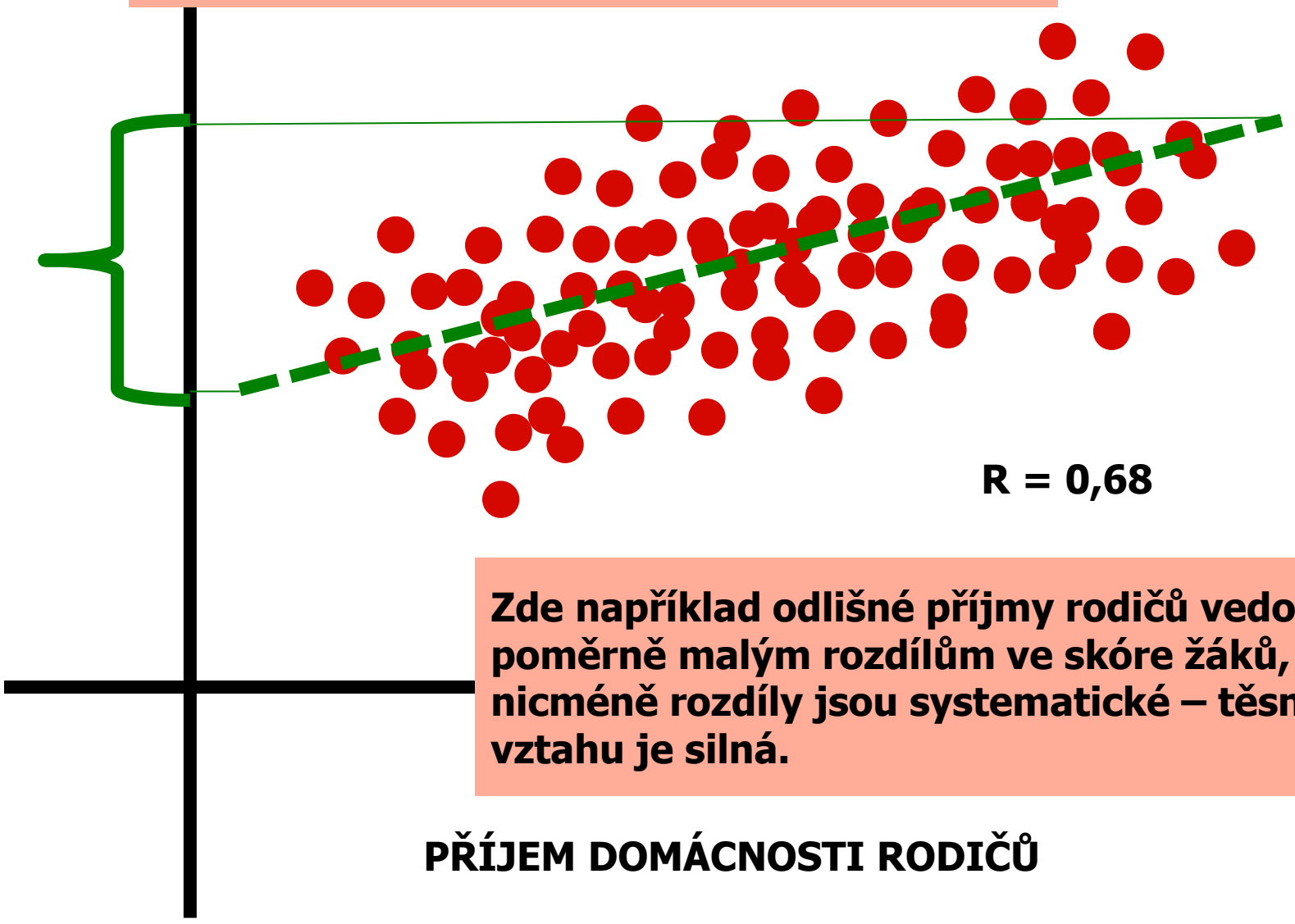
TĚSNOST = jak často je změna v příjmu rodičů asociována se změnou skóre žáka

EFEKT = jak moc se mění skóre žáka v závislosti na příjmu rodičů

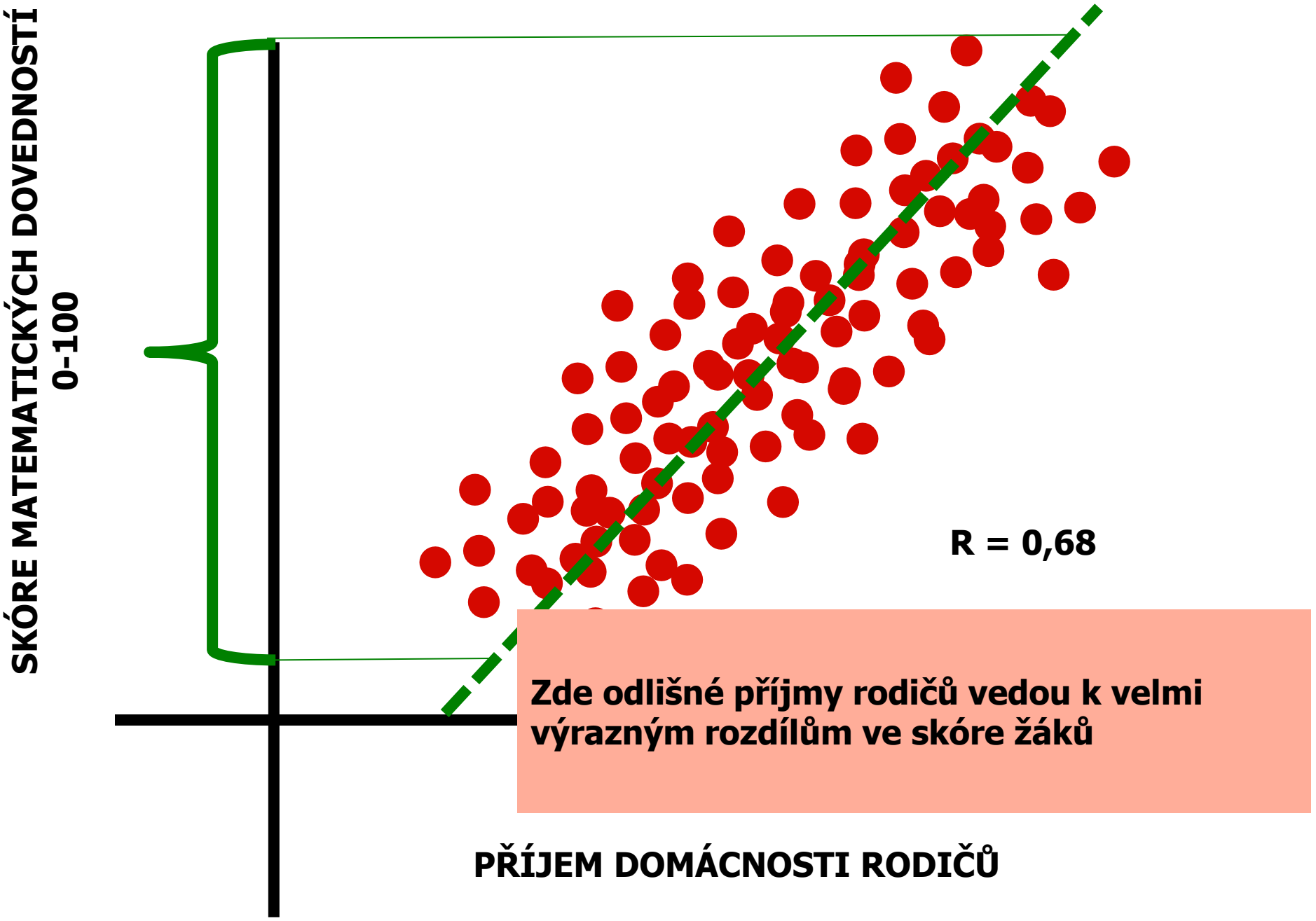




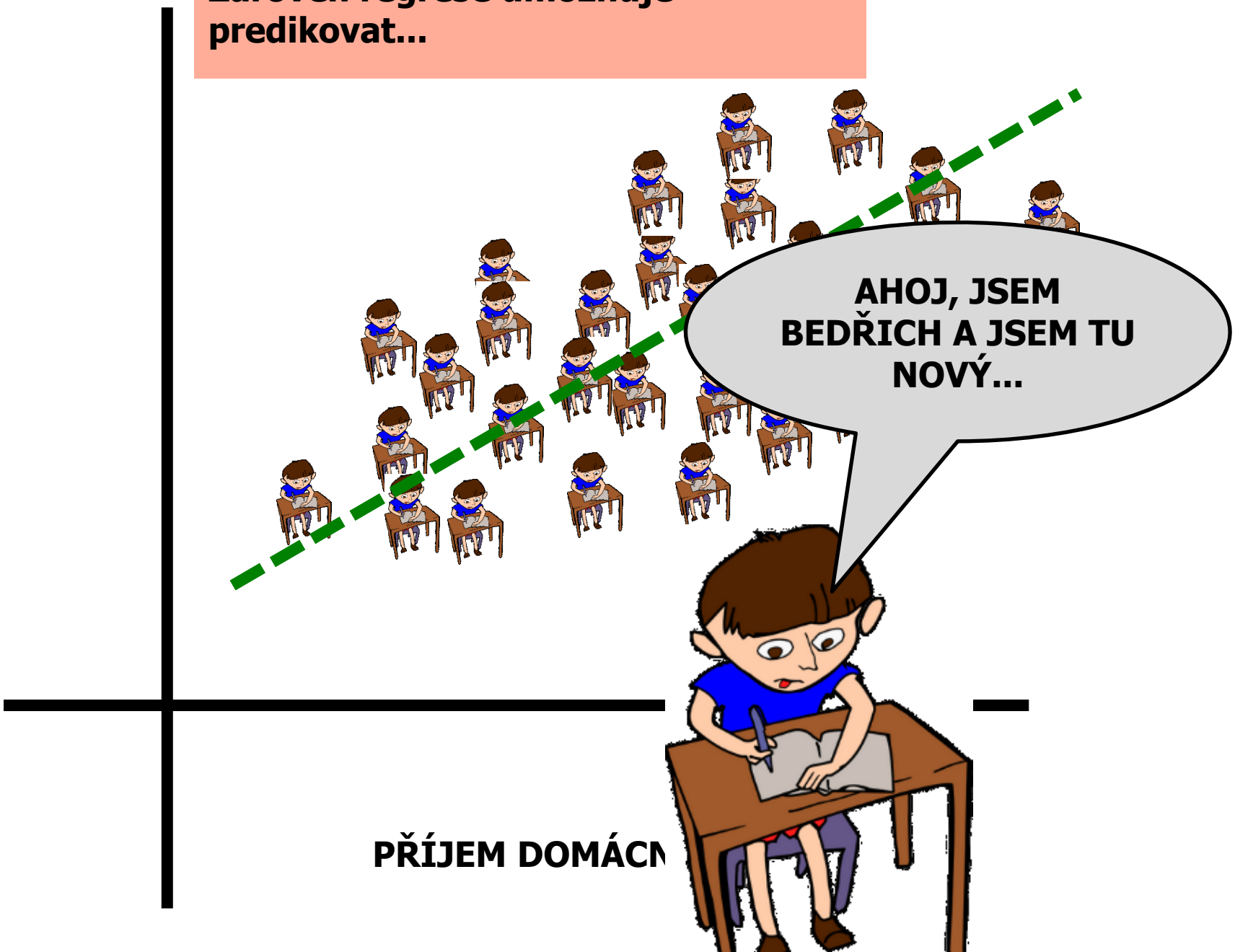
REGRESE nám tedy na rozdíl od korelace prozradí, jak velký vliv má příjem rodičů na skóre žáků



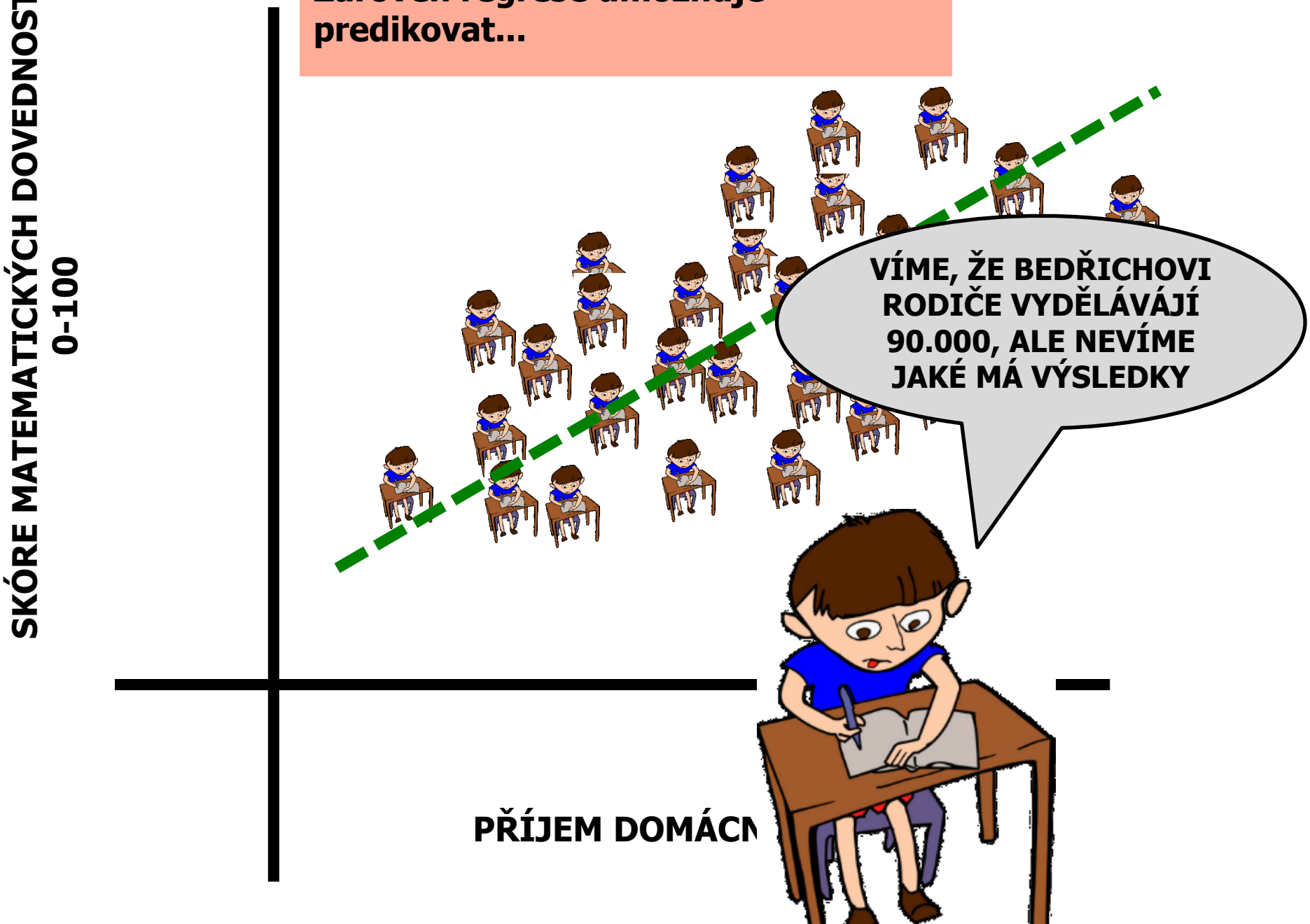
Zde například odlišné příjmy rodičů vedou k poměrně malým rozdílům ve skóre žáků, nicméně rozdíly jsou systematické – těsnost vztahu je silná.


SKÓRE MATEMATICKÝCH DOVEDNOSTÍ
0-100
R = 0,68
Zde odlišné příjmy rodičů vedou k velmi výrazným rozdílům ve skóre žáků
PŘÍJEM DOMÁCNOSTI RODIČŮ

Zároveň regrese umožňuje predikovat...



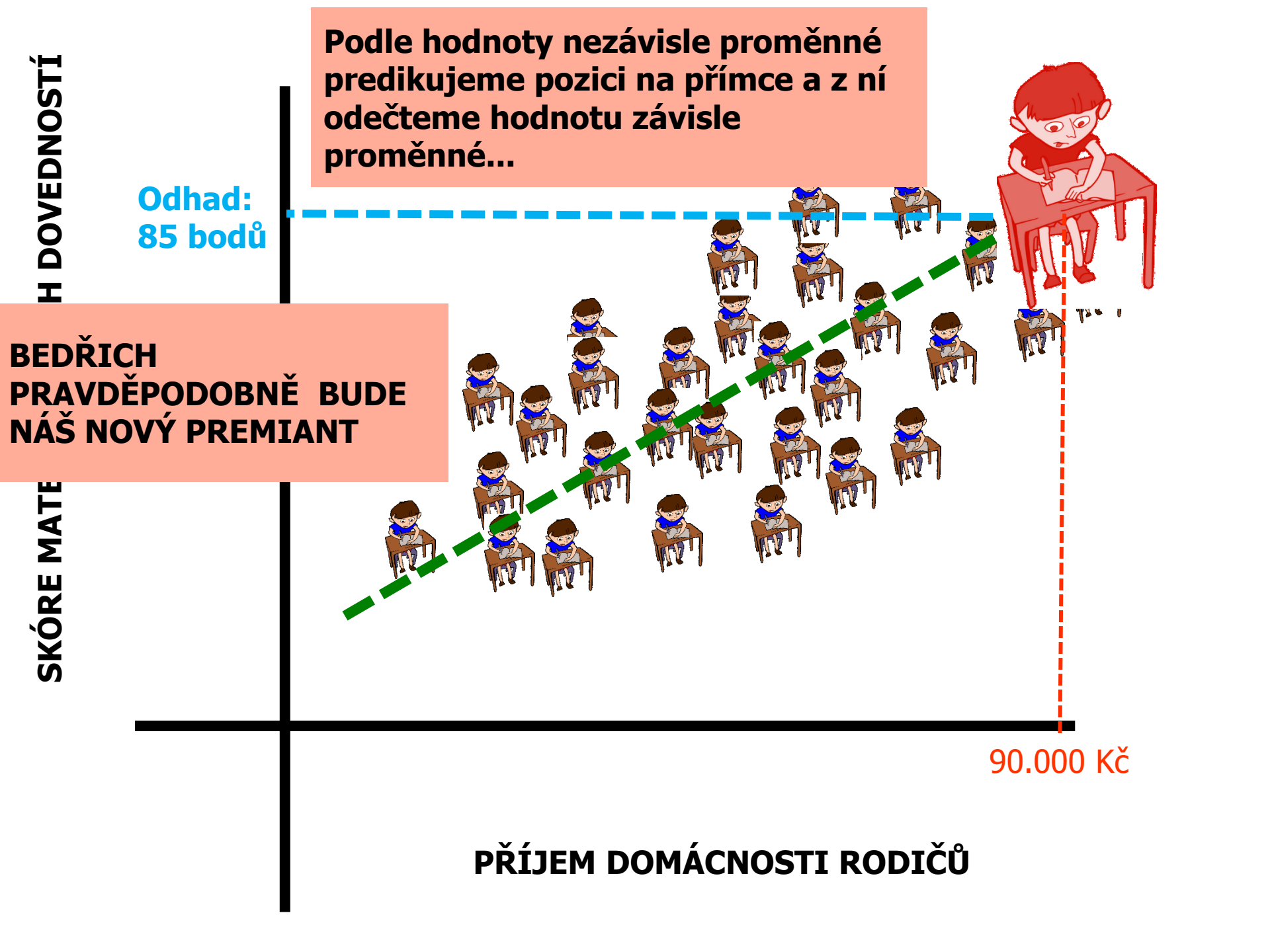
Zároveň regrese umožňuje predikovat...



Podle hodnoty nezávisle proměnné predikujeme pozici na přímce a z ní odečteme hodnotu závisle proměnné…

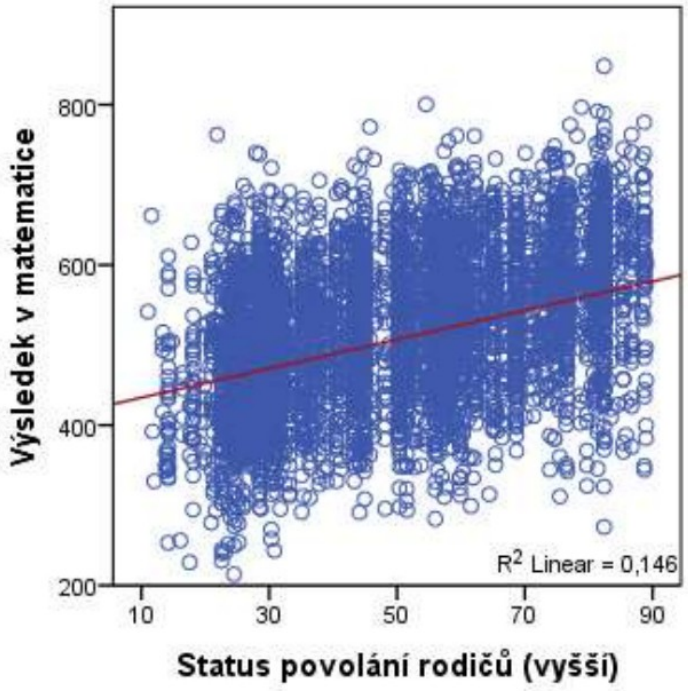
Odhad: 85 bodů

BEDŘICH PRAVDĚPODOBNĚ BUDE NÁŠ NOVÝ PREMIANT



VAROVÁNÍ K PŘÍKLADU:
Omluvte prosím stigmatizující a potenciálně stereotypizující charakter přechozího příkladu – děti chudých se neučí o tolik hůře jako děti bohatých

Graf č. 4 **Vztah úrovně matematické způsobilosti a společenského postavení rodin žáků**



Zdroj: PISA 2012

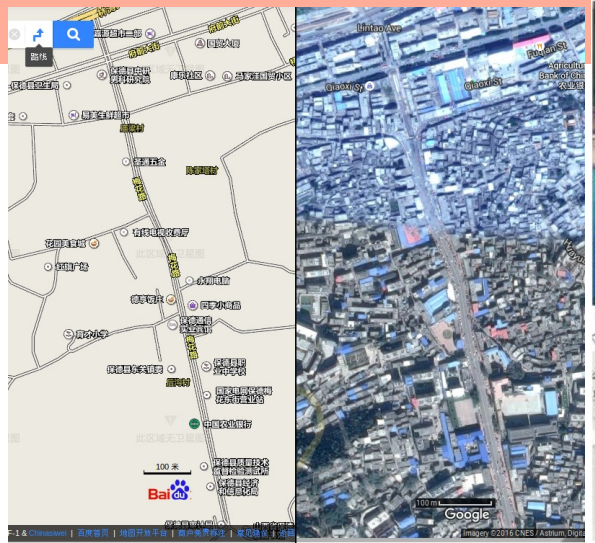
Výsledky z reálných dat PISA 2012

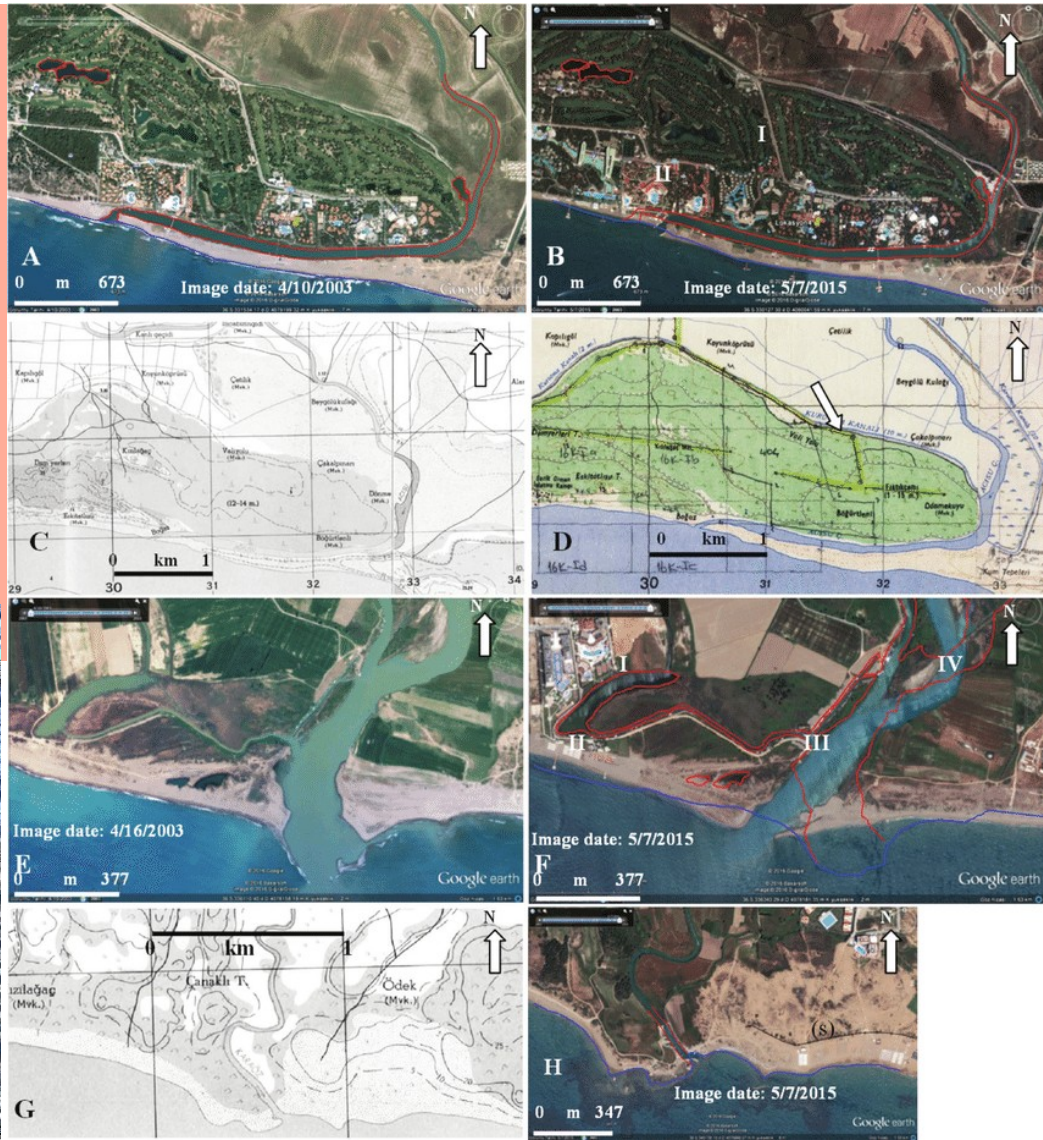
https://www.csicr.cz/Csicr/media/Prilohy/PDF_el._publikace/Mezinárodní%20šetření/PISA_2012_SA.pdf

S regresní analýzou opouštíme oblast popisných statistik a vstupujeme na pole modelovaní.

Model je vždy abstrakce - odhlížíme od detailů ve prospěch zachycení důležitého vzorce

Srov.
Letecký snímek vs. mapa

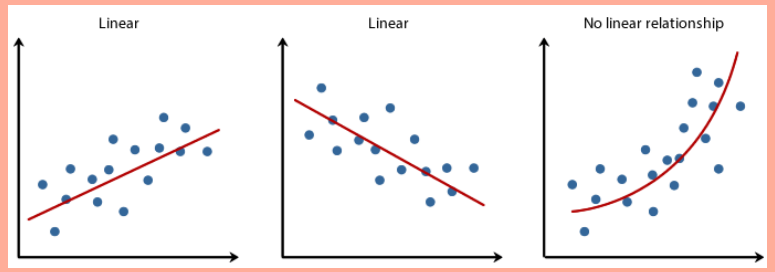




U MODELŮ LZE OBECNĚ UVAŽOVAT O DVOU ASPEKTECH:

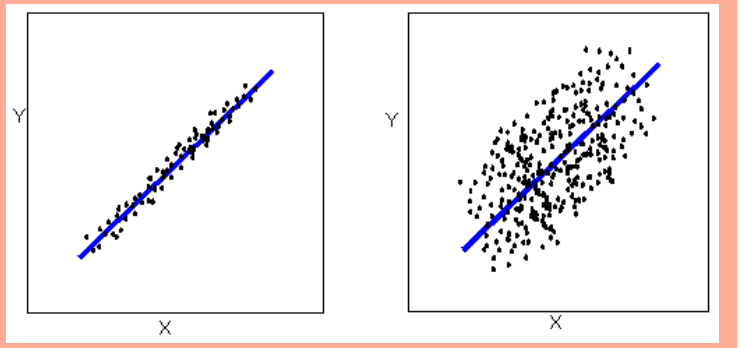
PARAMETRY MODELU: co model říká

- jaký tvar má souvislost,
- jaký je sklon přímky...



Regresní rovnice a její členy

VHODNOST MODELU: jak dobře model reprezentuje data, jak "sedí„



Míry vhodnosti modelu
(nový význam R^2 – zde
jako míra vhodnosti modelu,
podobně také ANOVA – analýza rozptylu)

Cíl lineární regrese

Sumarizovat vztah mezi dvěma proměnnými ve formě **rovnice přímky** (neboť předpokládáme lineární, tj. přímkový vztah) prostřednictvím výpočtu **regresního koeficientu:**

$$\mathbf{y = a + b*x}\ (Y = b_0 + b_1 * X_1)$$

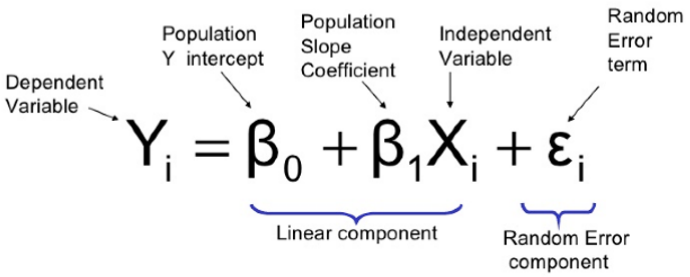
y …. hodnota závisle proměnné, tu chceme predikovat (*outcome*)

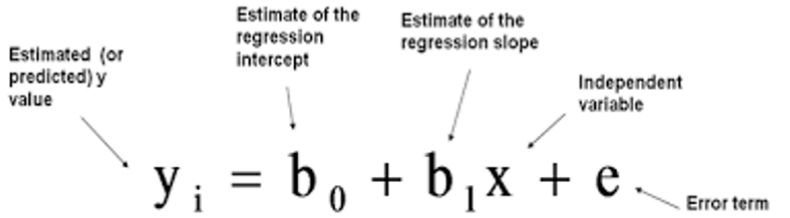
a …. parametr, který říká, v jakém bodě přímka protíná vertikální osu Y (hodnota Y, když X = 0),

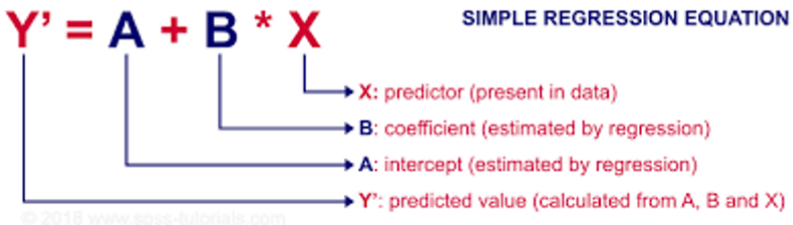
b …. regresní koeficient -- určuje směr přímky, (*predictor*)

x …. hodnota nezávisle proměnné, slouží k predikci hodnoty *y*

S regresní rovnicí se můžeme setkat v různých formách:







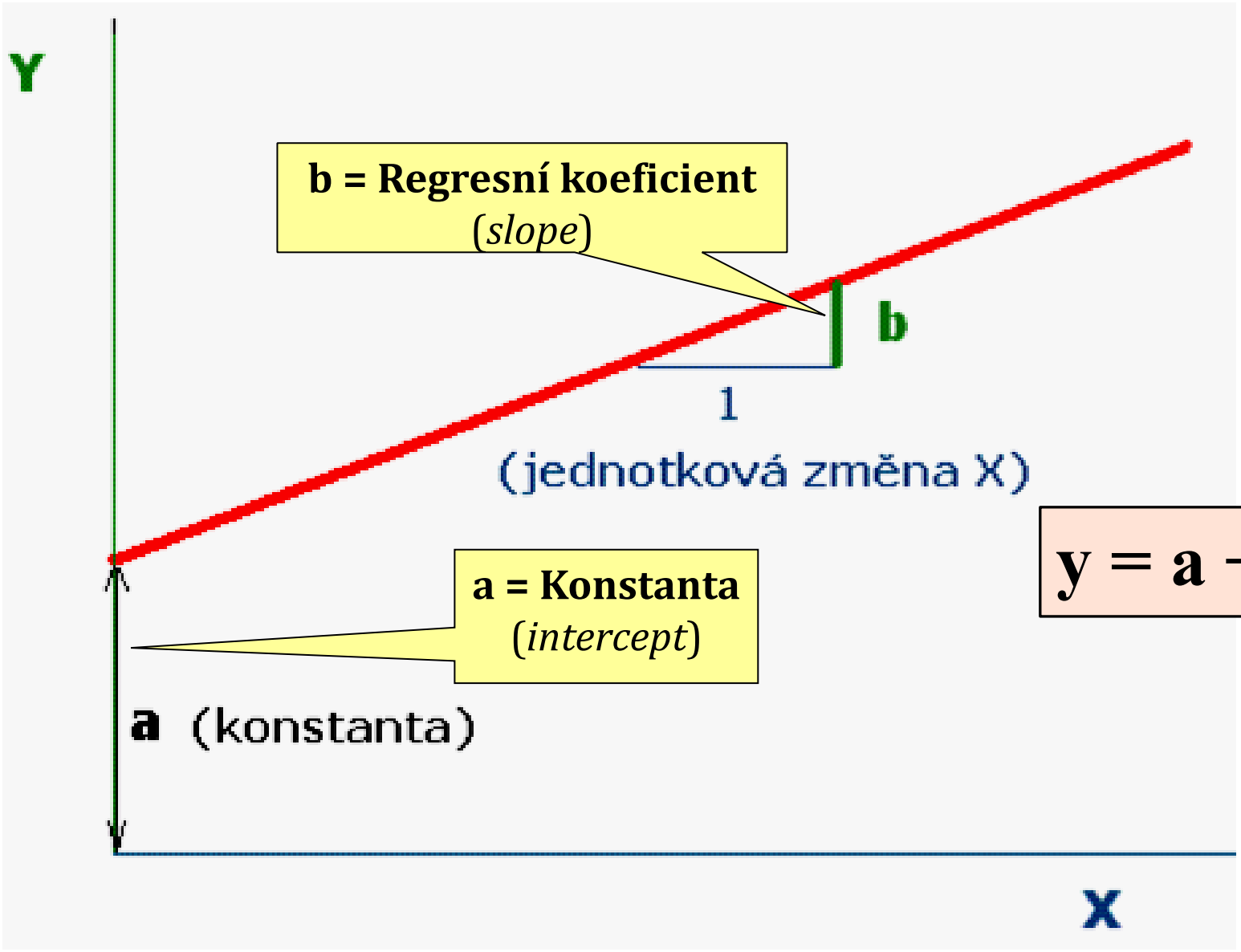
nejčastěji

y = a + b*x

Vždy ale musí obsahovat dva členy (které se však taky nazývají různě):

Něco, co určuje, **kde přímka protíná osu** Y – to je nutné pro její umístění v rovině – názvy: *konstanta, posunutí, intercept*

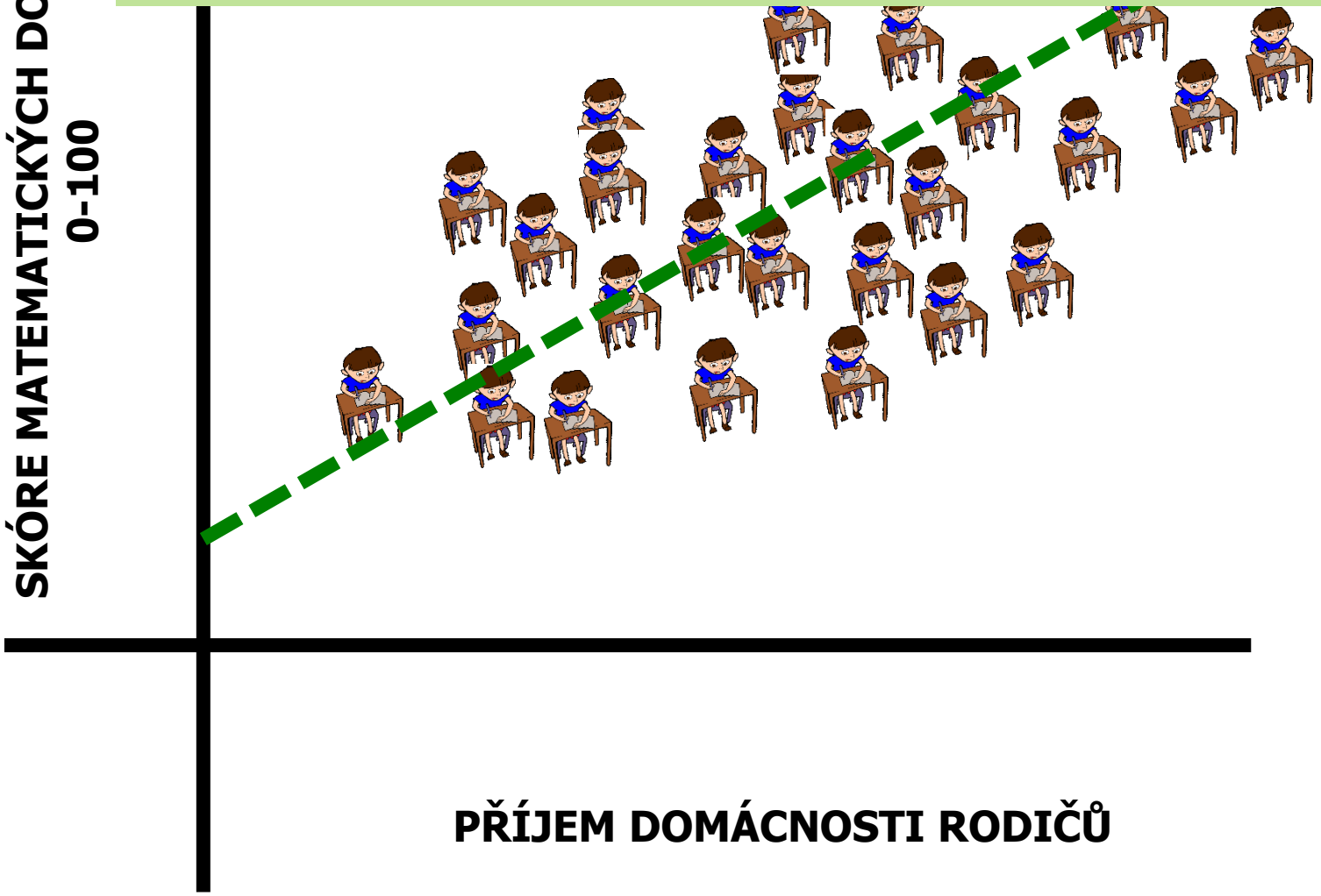
Něco, co určuje **sklon přímky**: *směrnice, koeficient, slope*

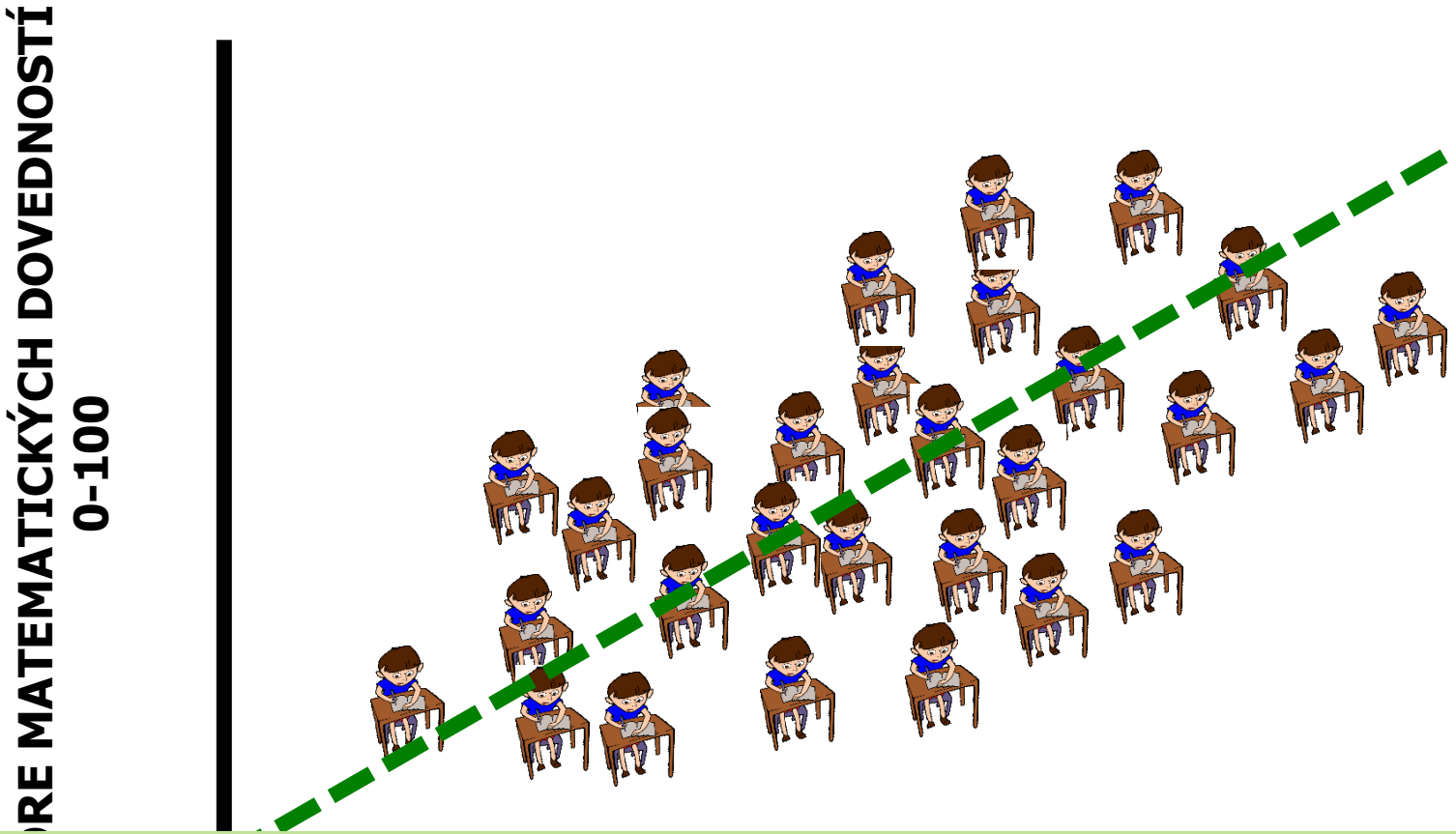

Y
b = Regresní koeficient
(slope)
b
1
(jednotková změna X)
a = Konstanta
(intercept)
a (konstanta)
X

$$y = a + b * x$$

Takže například:

skóre = 15 + 0,005*příjem rodičů





Takže například:

Jestliže se příjem domácnosti zvýší o korunu, zvýší se skóre žáka o 0,005 bodu

… jestliže se zvýší o 1000 Kč, zvýší se skóre o 5 bodů

Podmínky pro užití lineární regresní analýzy:

1. Vztah mezi analyzovanými proměnnými musí být lineární,
2. závisle proměnná Y je měřena na intervalové úrovni a nezávisle proměnná X je buď intervalová, nebo dichotomická,
3. obě proměnné by měly být přibližně normálně rozloženy
4. V datech není přítomna heteroskedasticita