Regresní analýza

Použití

- TESTOVÁNÍ TEORIÍ !!!

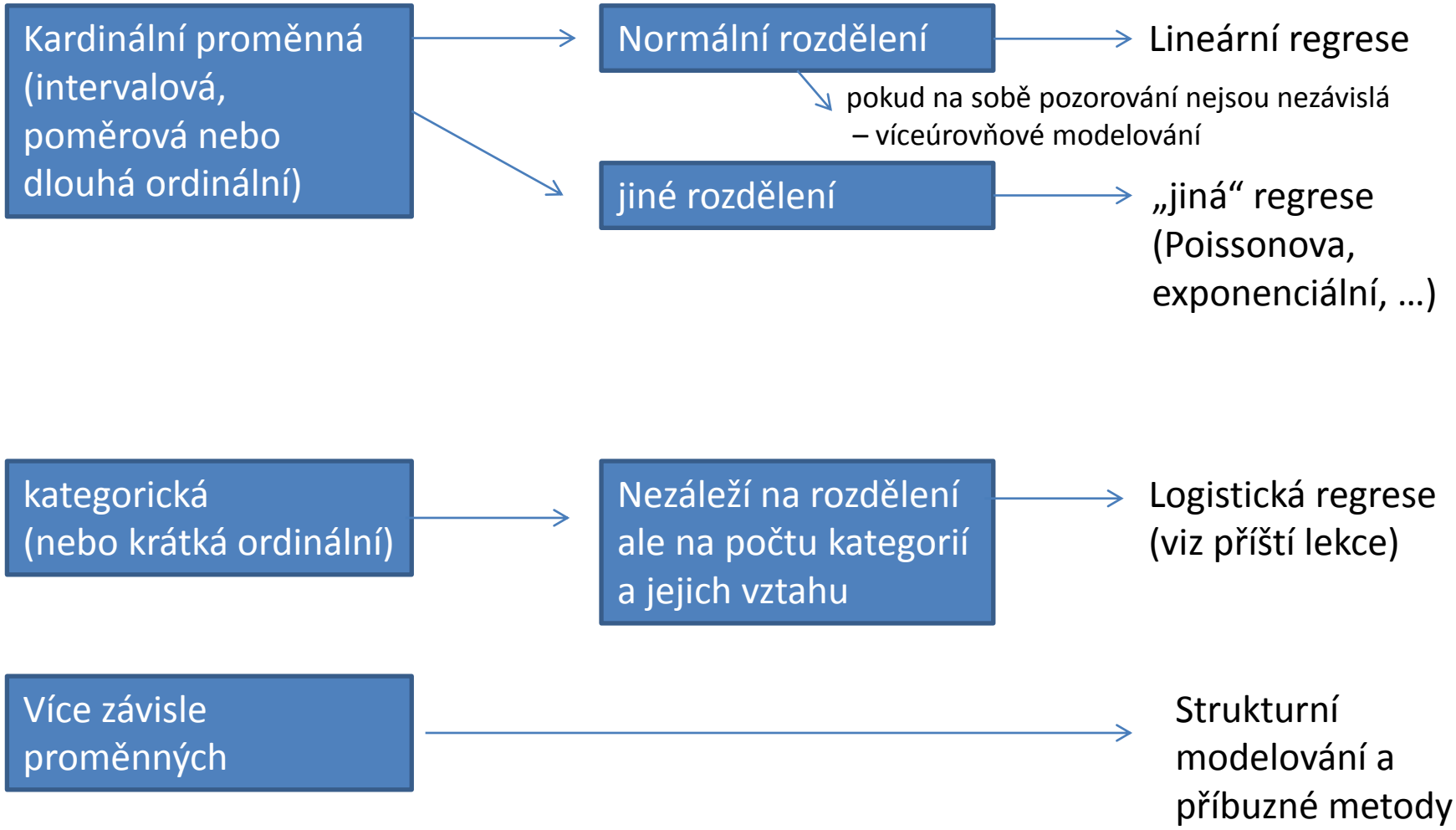
- Zjištění vlivu nezávisle proměnné na závisle proměnnou
  - Při kontrole dalších možných faktorů
  - (predikce: jakou hodnotu bude mít závisle proměnná při určité kombinaci nezávisle proměnných)

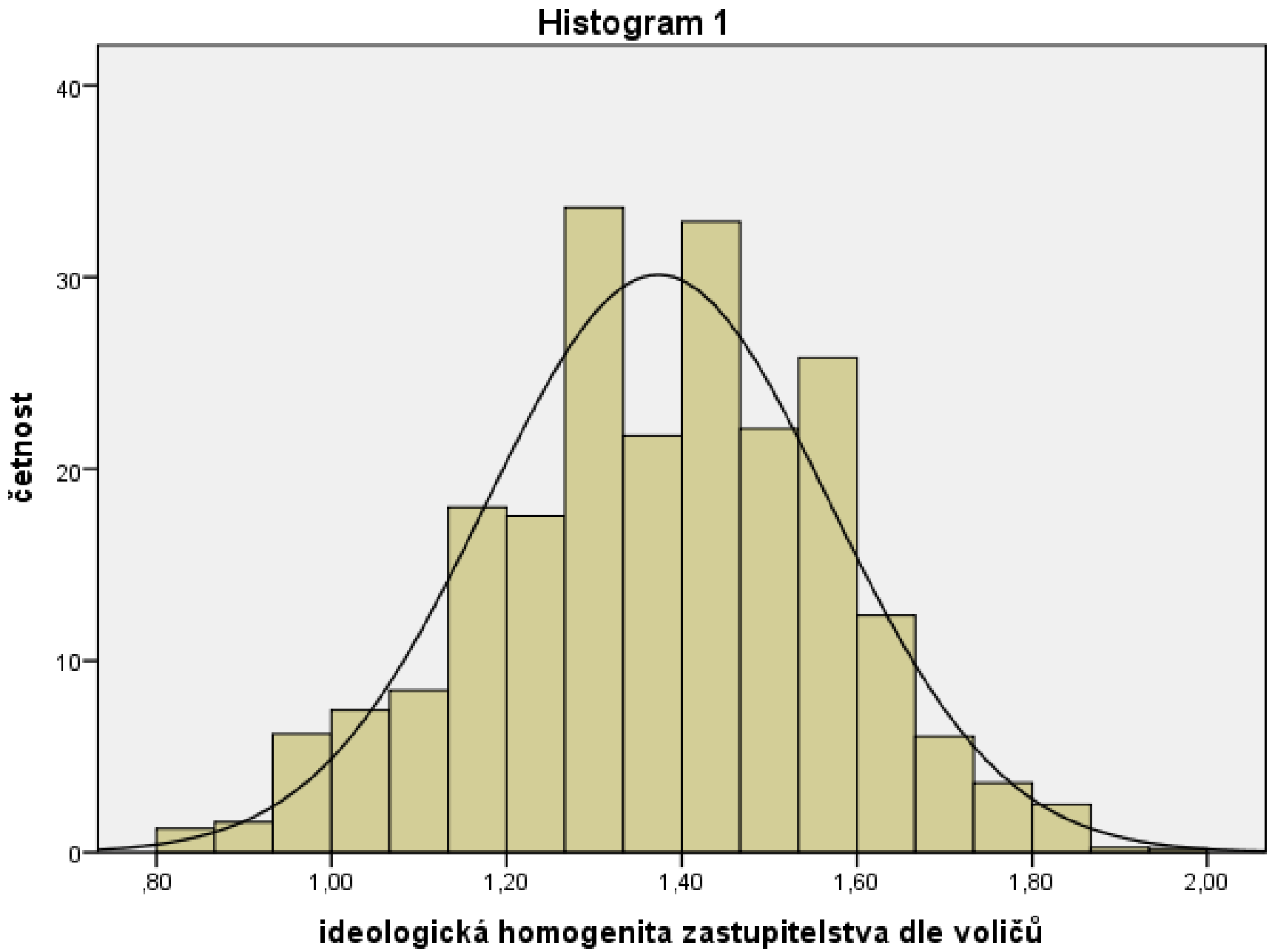
Podmínky

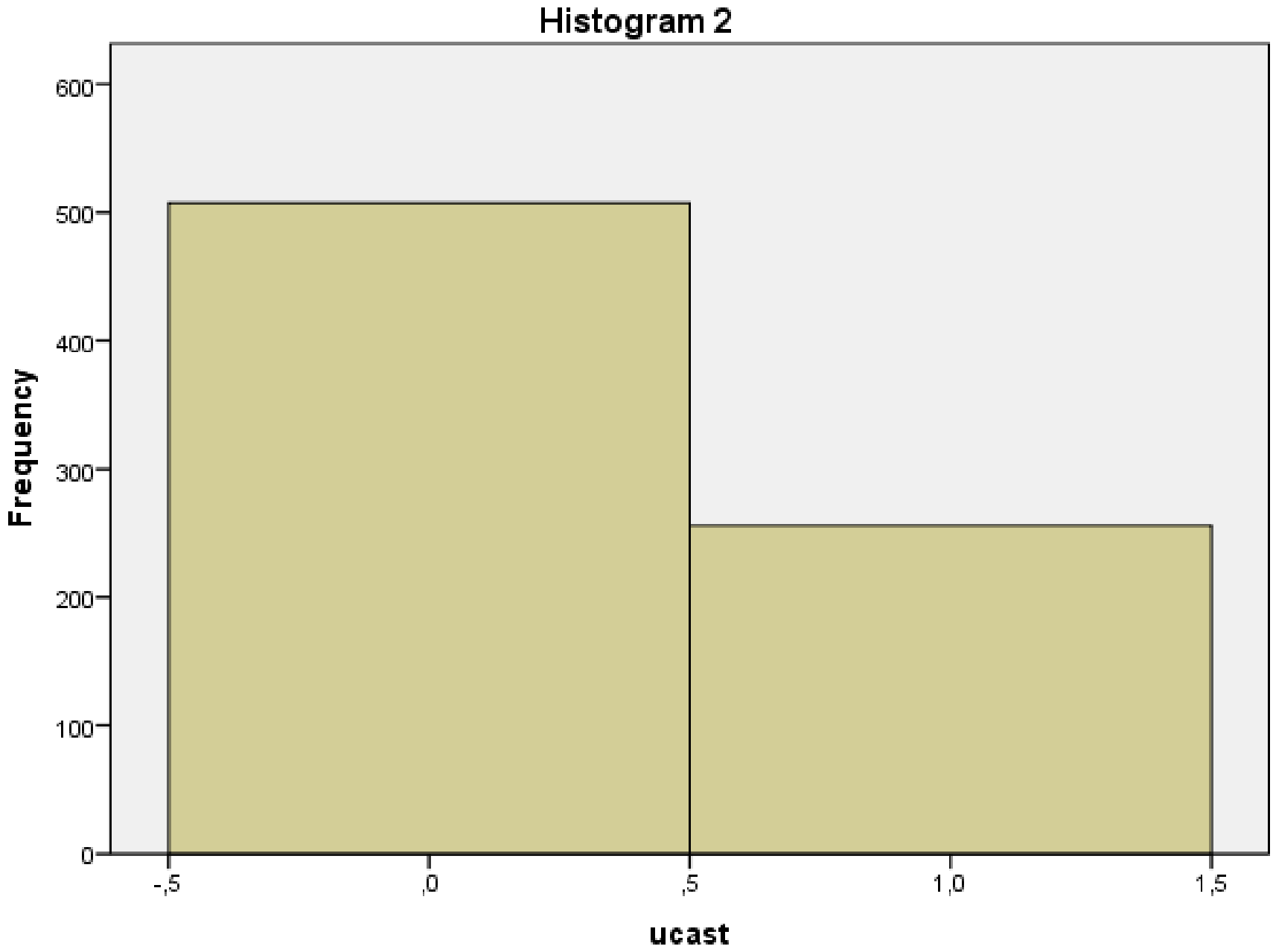
- Jedna závisle proměnná
  + Jedna nebo více nezávisle proměnných
- Normálně rozdělená základní proměnná
  - Rozdělení a typ nezávisle proměnné může být jakékoli
  + několik dalších různě důležitých podmínek
    - Nezávislost pozorování
    - Předpoklad lineárního vztahu

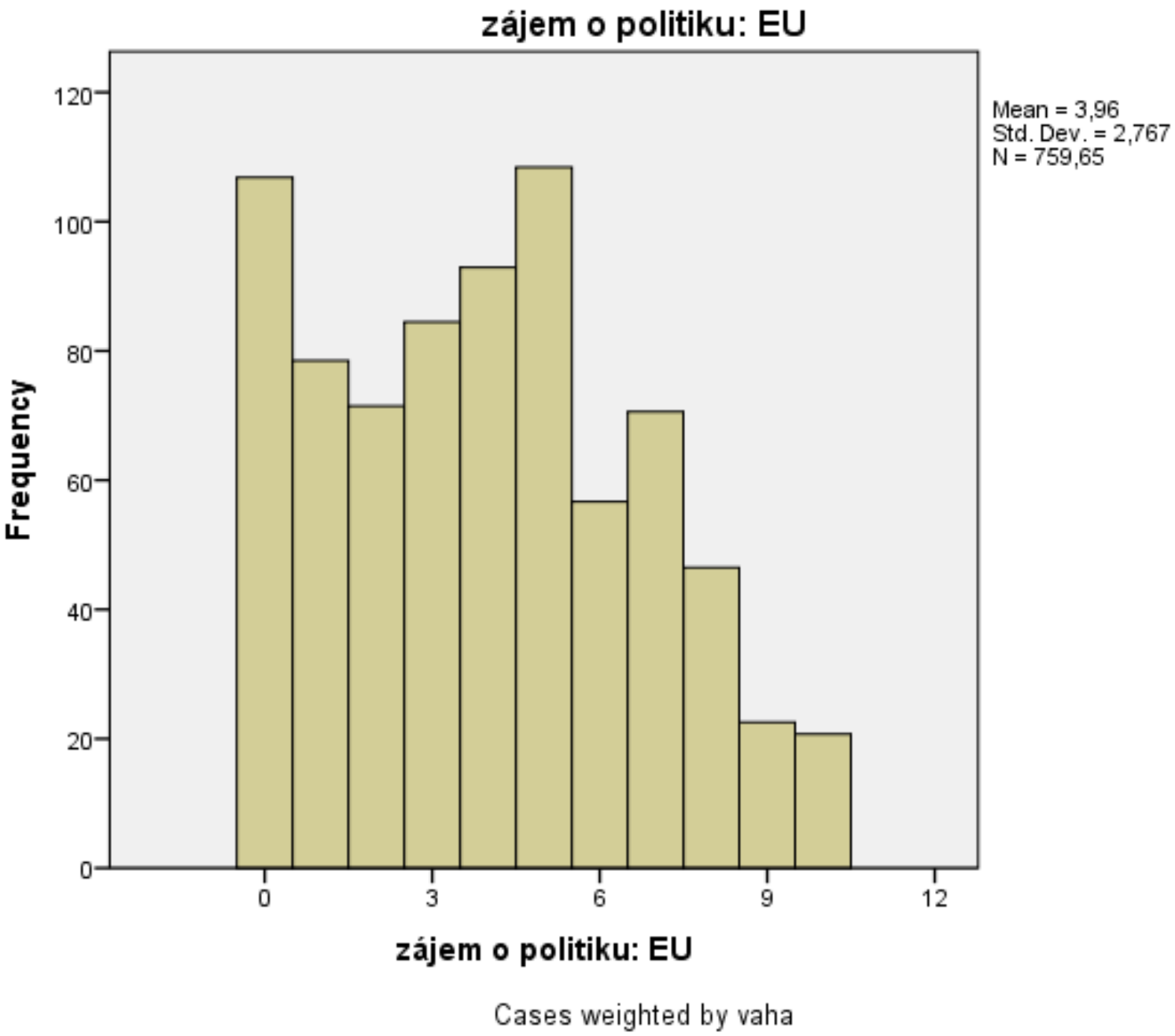
    - Nezávislost nezávisle proměnných mezi sebou
    - Homogenní rozptyl reziduí

Rozhodovací strom




Histogram 1
40
30
20
10
0
četnost
,80
1,00
1,20
1,40
1,60
1,80
2,00
ideologická homogenita zastupitelstva dle voličů
Cases weighted by vaha


Histogram 2
Frequency
600
500
400
300
200
100
0
-,5
,0
,5
1,0
1,5
ucast
Cases weighted by vaha


zájem o politiku: EU
Mean = 3,96
Std. Dev. = 2,767
N = 759,65
Frequency
120
100
80
60
40
20
0
0
3
6
9
12
zájem o politiku: EU
Cases weighted by vaha

Co regrese dělá

- Odhad parametrů přímky (při 1 nezávisle proměnné), roviny (při 2) či nadroviny (při více)
- Parametry: sklon (pro každou proměnnou) a konstanta (jedna pro celý model)
- Parametry popisují vztah mezi nezávisle a závisle proměnnou
- Hodnota závisle proměnné = konstanta + sklon*hodnota nezávisle proměnné
- $y = a + b*x$
- $y = a + b1*x + b2*x + b3*x + \ldots$

Postup

- Nadefinování modelu pomocí hypotéz vycházejících z teorie
- Sestavení datasetu obsahujícího závisle a nezávisle proměnné dle specifikace
- Zkontrolování normality závisle proměnné
- Zkontrolování vlastností nezávisle proměnných

Normalita závisle proměnné

- Jinakost rozdělení
  - ovlivňuje především hodnoty signifikance
  - Zkresluje odhady parametrů
- Prvně vizuální zhodnocení pomocí histogramu
- Testy
  - K-S a S-W
  - Ve velkých souborech lze brát s rezervou
  - Šikmost a strmost není větší než 3

Další postup

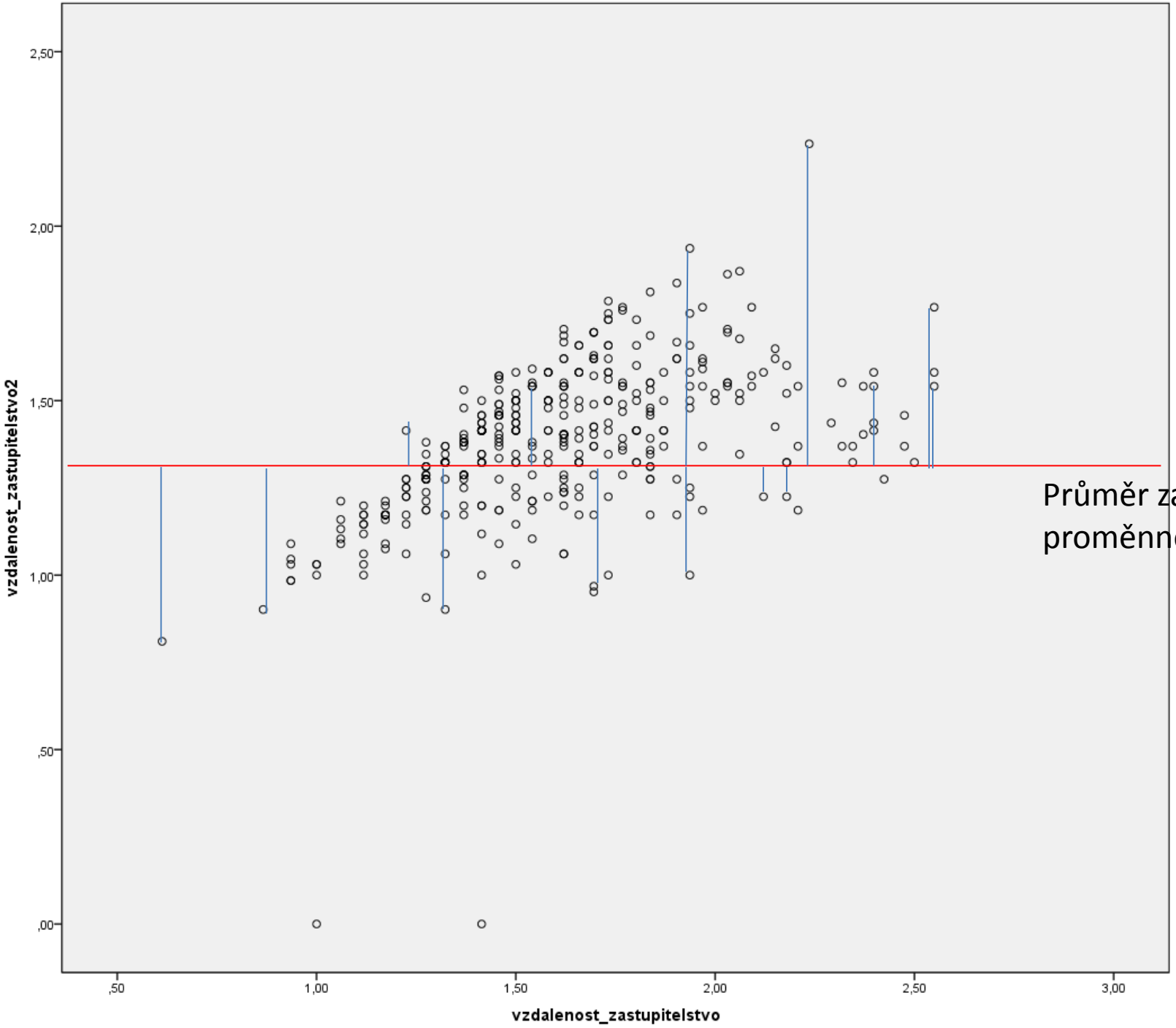
- Pokud je závisle proměnná v pořádku
  - Rekódování nezavisle proměnných
  - Kontrola multikolinearity nezávisle proměnných
    - Nezávisle proměnné by mezi sebou neměly příliš souviset
    - První kontrola pomocí korelačního koeficientu
    - Další kontrola přímo v modelu
  - Výpočet

Co nám výpočet poskytne?

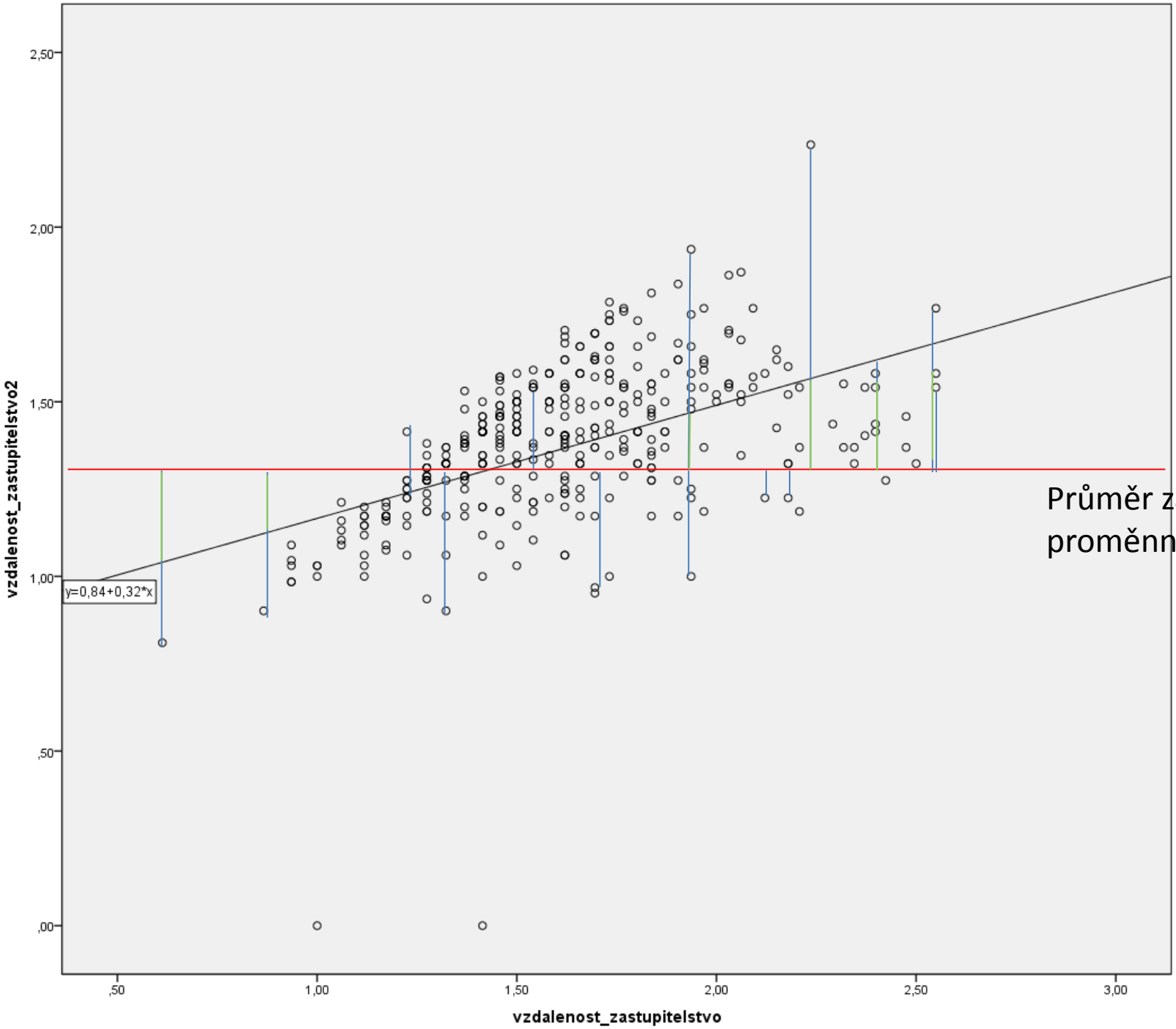
- R-square (česky index determinace)
  - Ukazuje jak dobře model sedí na data
- Parametry
  - Unstandardized beta (nestandardizovaný beta koeficient)
  - Constant (konstanta)
- Hodnoty signifikance

Co je to R-square?

- Ukazuje, kolik procent variability závisle proměnné je vysvětleno přidáním nezávisle proměnných
- Původní variabilita je vypočtena jako suma kvadratických odchylek mezi průměrem a jednotlivými hodnotami závisle proměnné
- „nová“ variabilita je vypočtena jako suma odchylek od regresní přímky/roviny
- Rozdíl mezi původní a novou variabilitou vydělený původní variabilitou = R-square
- Čím víc proměnných, tím nižší R-square
  - Řešeno pomocí adjusted R-squre



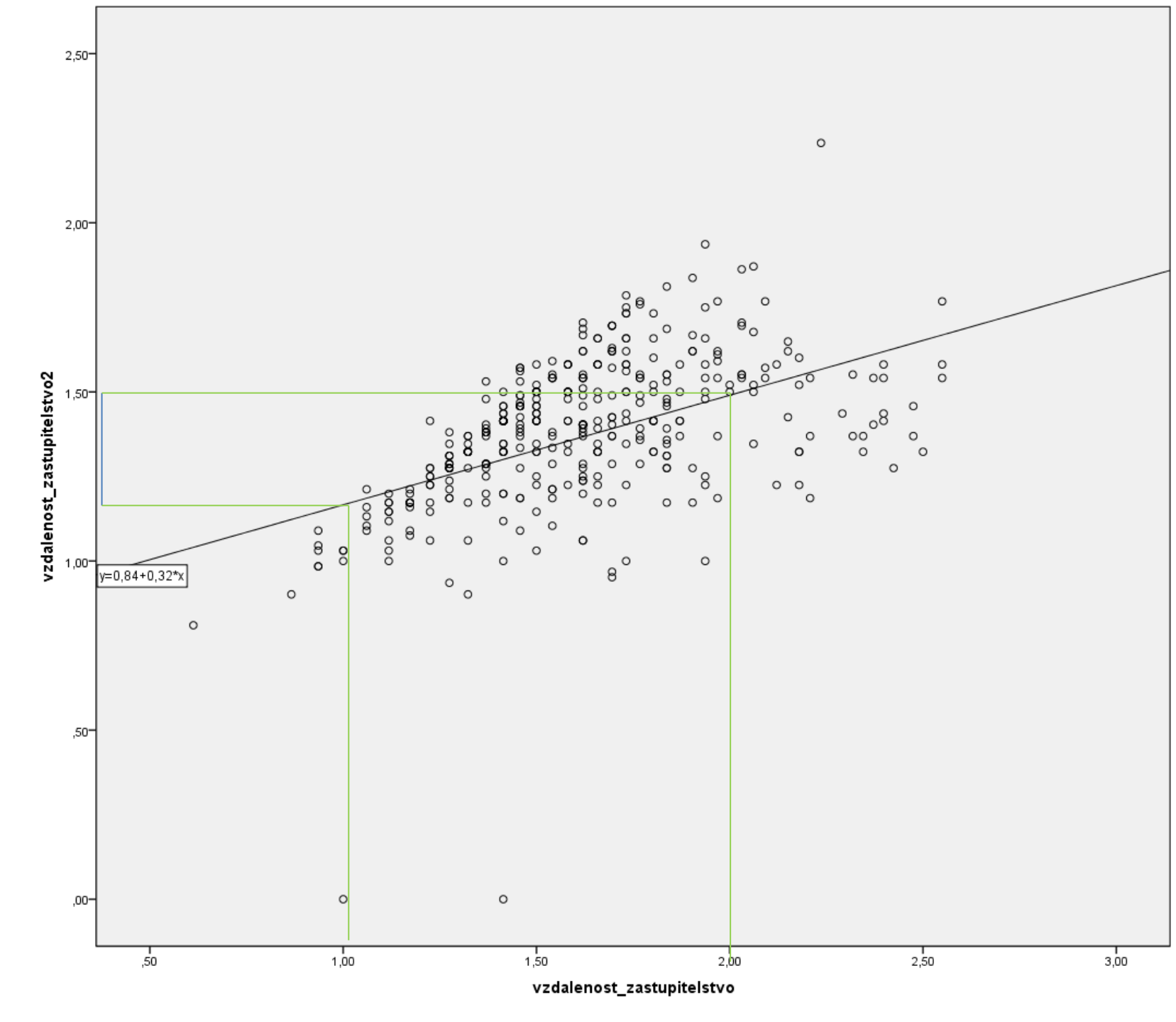
Průměr závisle proměnné



Průměr závisle proměnné

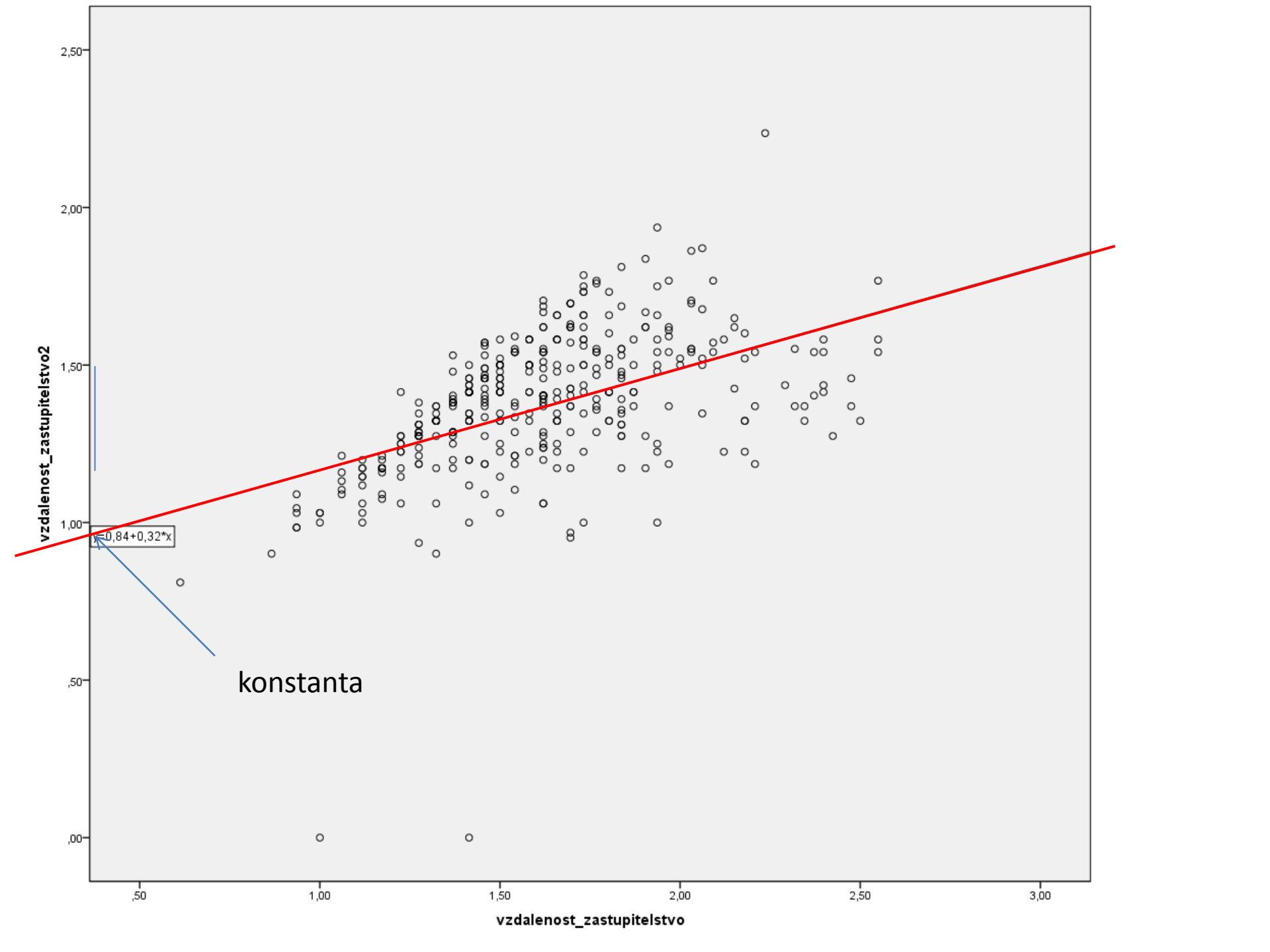
Nestandardizovaný Beta koeficient

- „o kolik se změní hodnota závisle proměnné, pokud se hodnota nezávisle proměnné změní o jednotku“
- Různé proměnné se mohou změnit o různý počet jednotek
  - Pro srovnání síly proměnných v modelu – standardizovaný koeficient beta (jakou změnu v počtu směrodatných odchylek závisle proměnné způsobí změna o směrodatnou odchylku nezávisle proměnné)


2,50
2,00
1,50
1,00
,50
,00
vzdalenost_zastupitelstvo2
y=0,84+0,32*x
,50
1,00
1,50
2,00
2,50
3,00
vzdalenost_zastupitelstvo

Konstanta

- Jaká je očekávaná hodnota nezávisle proměnné, pokud jsou hodnoty všech nezávisle proměnných 0
- Pro smysluplnou interpretaci je často potřeba rekódovat proměnné
  - Každý má nějaký věk, pohlaví, výšku, váhu, …


2,50
2,00
1,50
1,00
,50
,00
vzdalenost_zastupitelstvo2
=0,84+0,32*x
konstanta
,50
1,00
1,50
2,00
2,50
3,00
vzdalenost_zastupitelstvo

Následná kontrola

- Outlieři
- Homogenita rozptylu reziduí (homoskedascita)
- multikolinearita

Příklad

- Téma: Vnímání ideologické homogenity zastupitelstva
- Popis problému:
  - v zastupitelstvu zasedá 8 stran
  - Voliči mezi respondenty hodnotili pozici stran na škále levice pravice
  - Někteří voliči si myslí, že strany v zastupitelstvu reprezentují odlišné ideologické pozice, někteří si myslí, že strany jsou nerozlišitelné
- Otázka: Co způsobuje rozdílné vnímání ideologických rozdílů mezi stranami?

Teorie

- Prostorové modely volebního chování
- Občané jsou snáze schopni rozlišit strany blízké jejich pozici a hůře strany jim vzdálené
- Více středových stran – voliči umisťující se do středu spektra by měli vnímat strany v zastupitelstvu jako různorodější než levicoví či pravicoví voliči
- Voliči s vyšším zájmem o politiku by měli mít lepší informace o stranách a být lépe schopni rozlišit ideologické rozdíly mezi nimi
- Role vzdělání a politické znalosti

	Průměr
KSČM	1.54
ČSSD	3.44
SZ	5.22
KDU-ČSL	5.28
Žít Brno	5.42
ANO	5.83
TOP09	7.58
ODS	7.73

Hypotézy

- H1: Pravicoví a levicoví voliči vnímají strany jako méně rozdílné než středoví voliči
- H2: S rostoucím zájmem o politiku roste vnímána rozdílnost mezi stranami.
- H3: s vyšší vzděláním roste vnímaná rozdílnost mezi stranami
- H4: voliči s nějakou politickou znalostí vnímají strany jako více rozdílné než voliči bez znalosti
- H5: voliči umisťující se dále od všech stran vnímají strany jako podobnější

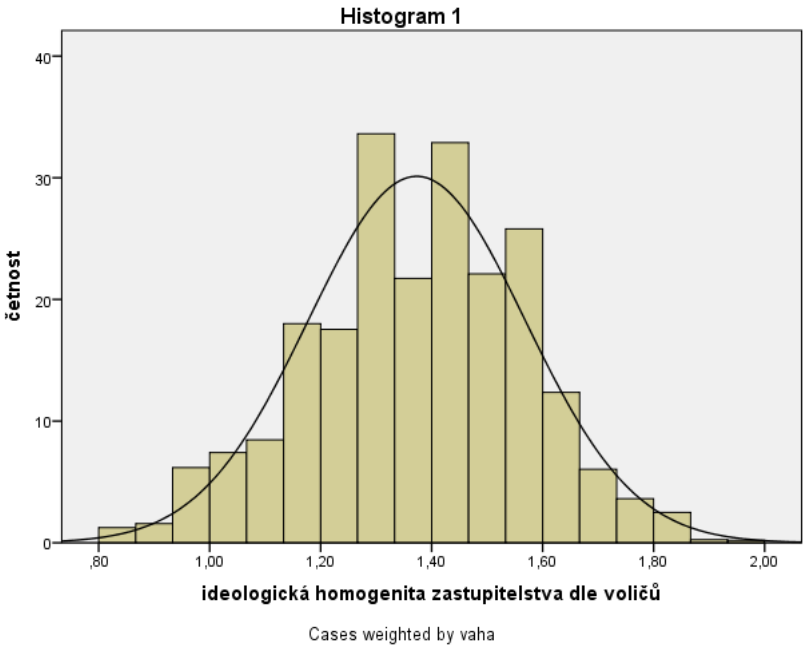
- H0 proměnná nemá vliv

Proměnné

- Závisle proměnná: Homogenita zastupitelstva
  - Spočítána jako odmocnina sumy odchylek pozice jednotlivých stran na škále levice pravice od průměrné pozice stran dle respondta
  - COMPUTE zastupitelstvo_lp=(q9_1+q9_2+ q9_3+ q9_4+ q9_5+ q9_6 + q9_8 + q9_7) /8.
  - COMPUTE ideolog_homog= sqrt((abs(q9_1- zastupitelstvo_lp)+ abs(q9_2- zastupitelstvo_lp)+ abs(q9_3- zastupitelstvo_lp)+ abs(q9_4- zastupitelstvo_lp)+ abs(q9_5- zastupitelstvo_lp)+ abs(q9_6- zastupitelstvo_lp)+ abs(q9_8- zastupitelstvo_lp)+ abs(q9_7- zastupitelstvo_lp)) /8).

Test normality závisle proměnné

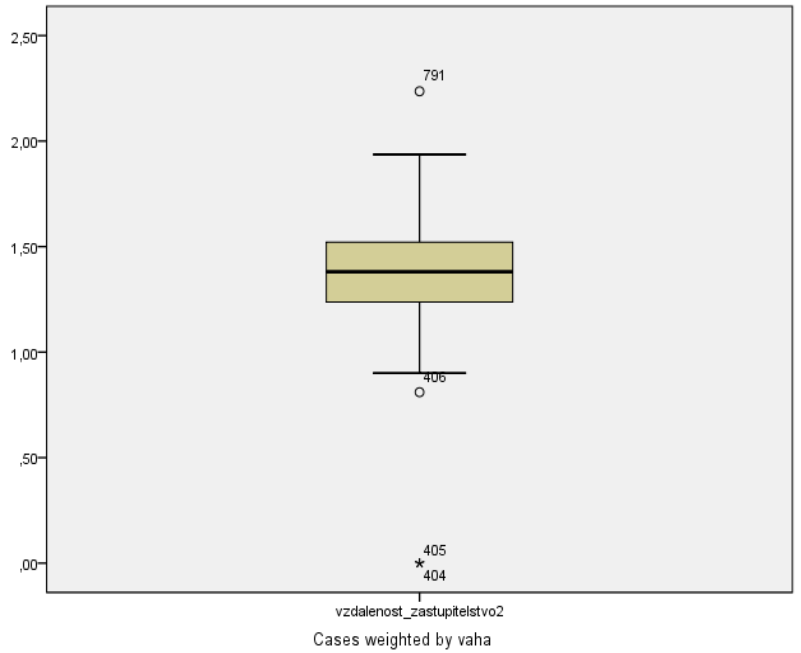
- Histogram
  - Analyze- descriptive stat- frequncies – plots
- Kolmogorův-Smirnovův test
  - Analyze – descriptive stat – explore – plots

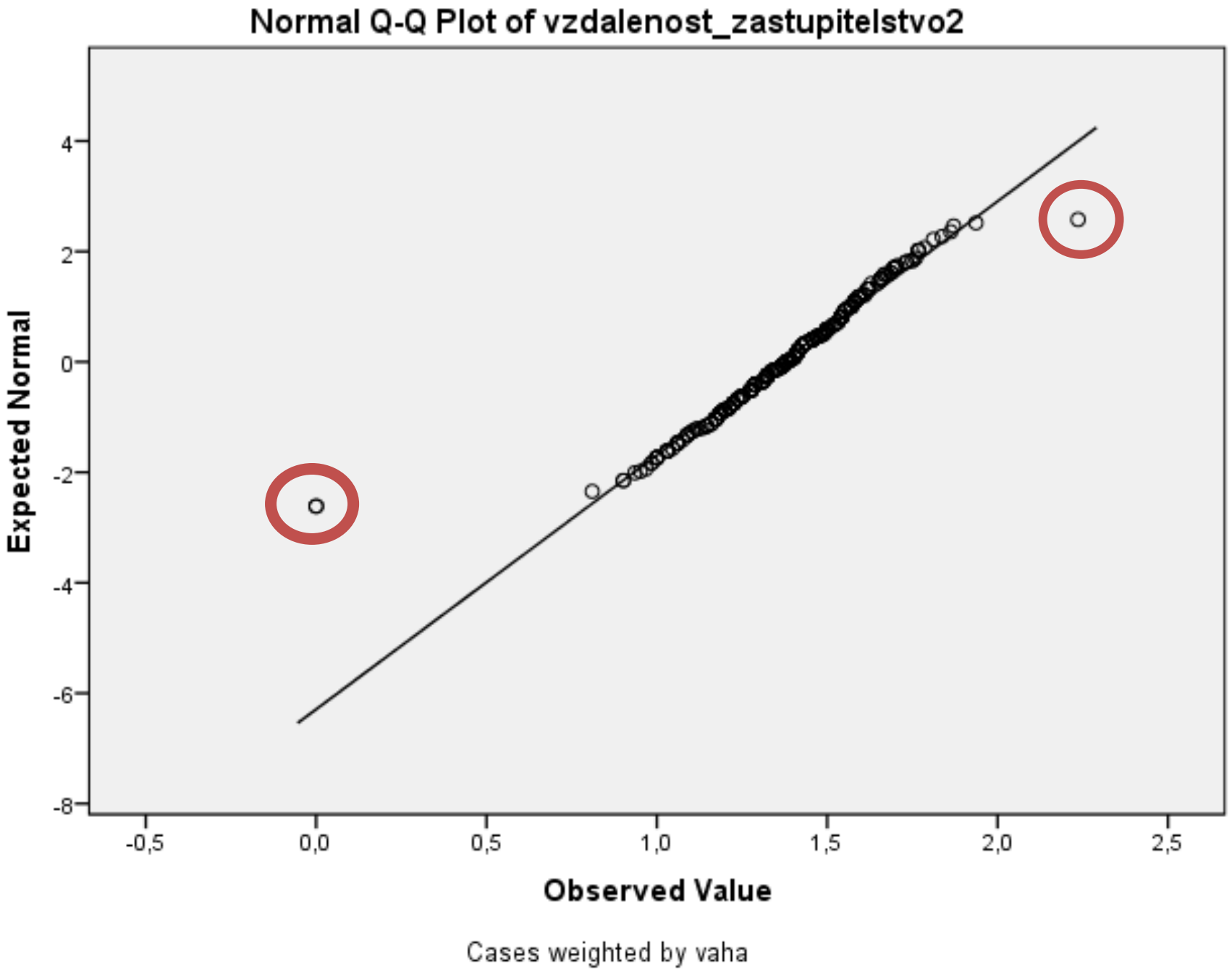


Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
vzdalenost_zastupitelstvo 2	,059	223	,060

a. Lilliefors Significance Correction




Normal Q-Q Plot of vzdalenost_zastupitelstvo2
Expected Normal
4
2
0
-2
-4
-6
-8
-0,5
0,0
0,5
1,0
1,5
2,0
2,5
Observed Value
Cases weighted by vaha

Po odebrání outlierů

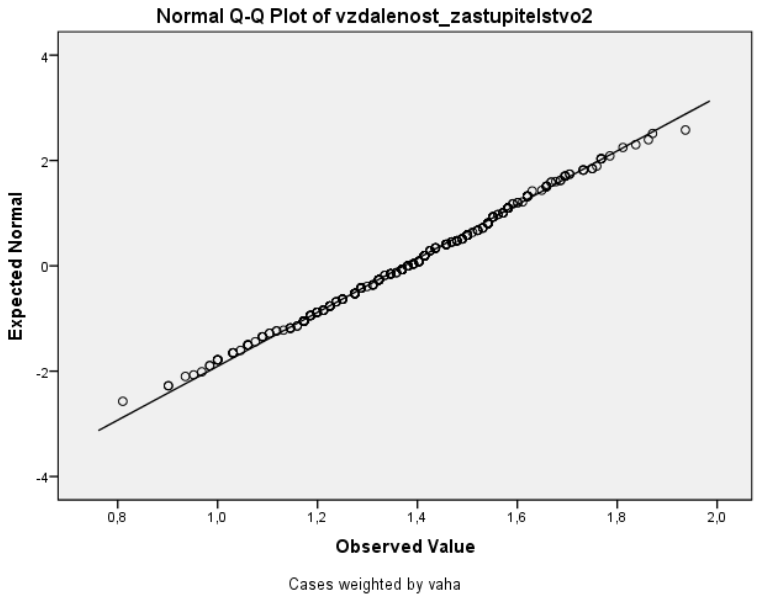
- Data – select cases – use filter variable – „filtr“

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
vzdalenost_zastupitelstvo 2	,040	222	,200*

*. This is a lower bound of the true significance.

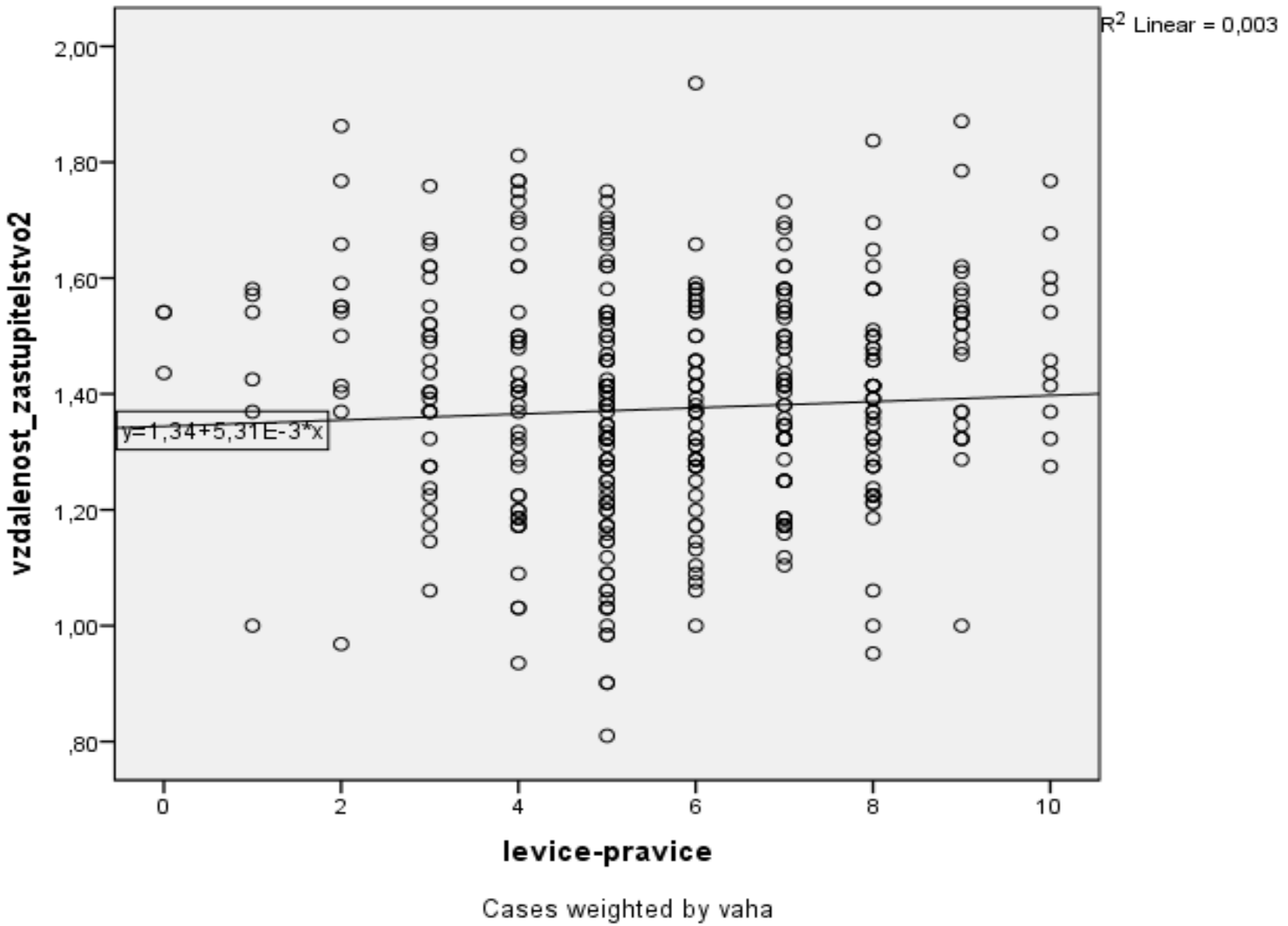
a. Lilliefors Significance Correction



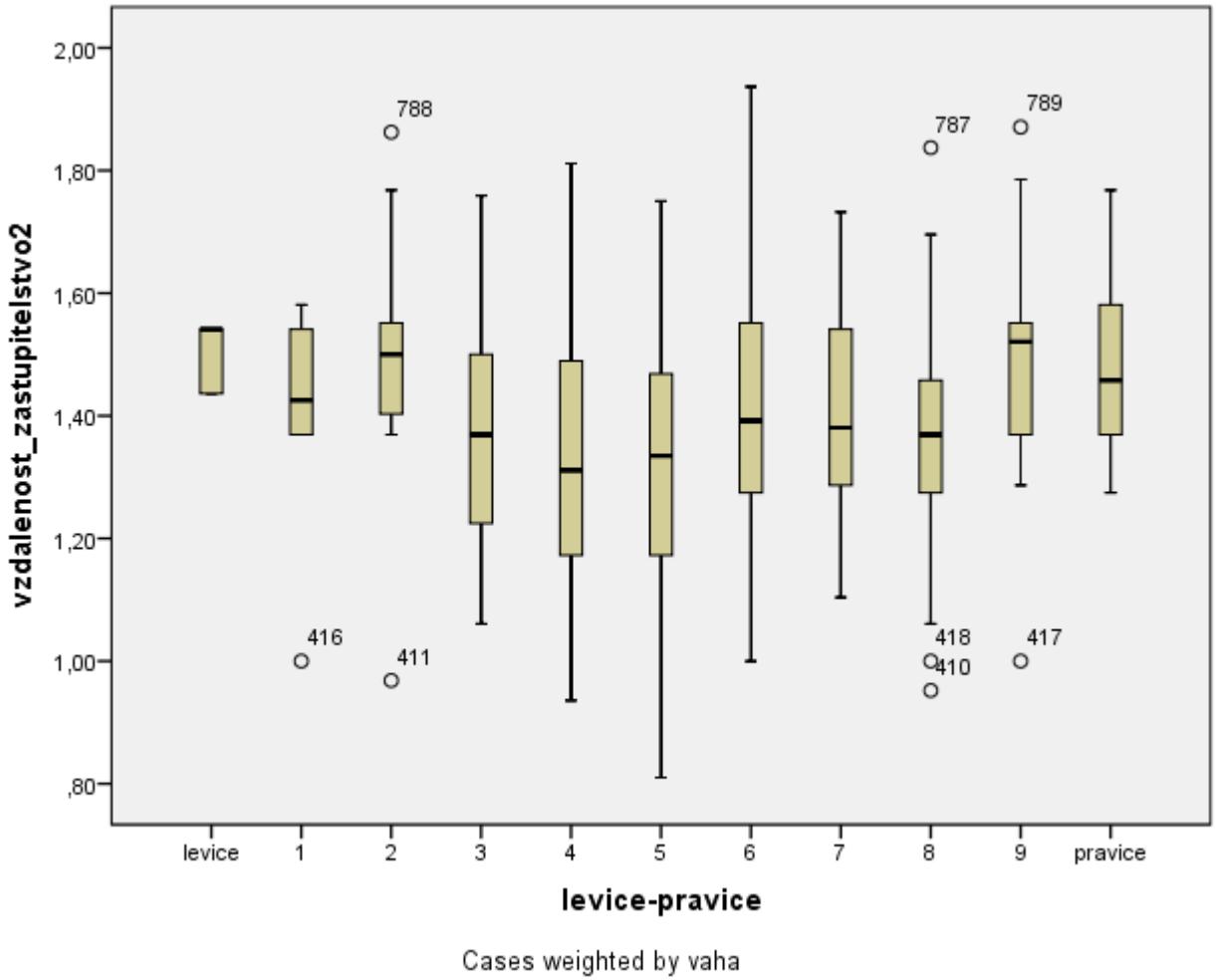
Nezávisle proměnné

- Zájem: součet proměnných ptajících se na zájem (od 0 do 50)
- Responsivita: součet rekódovaných proměnných o plnění slibů (od 0 do 11)
- Levice- pravice: škála od 0 do 10
- Znalost: zná jméno primátora =1, nezná=0
- Vzdlání: kategorická proměnná rekódovaná na dummy proměnné
  - VŠ vzdělání referenční kategorií

- Graphs – legacy dialogs – Scatter/dot

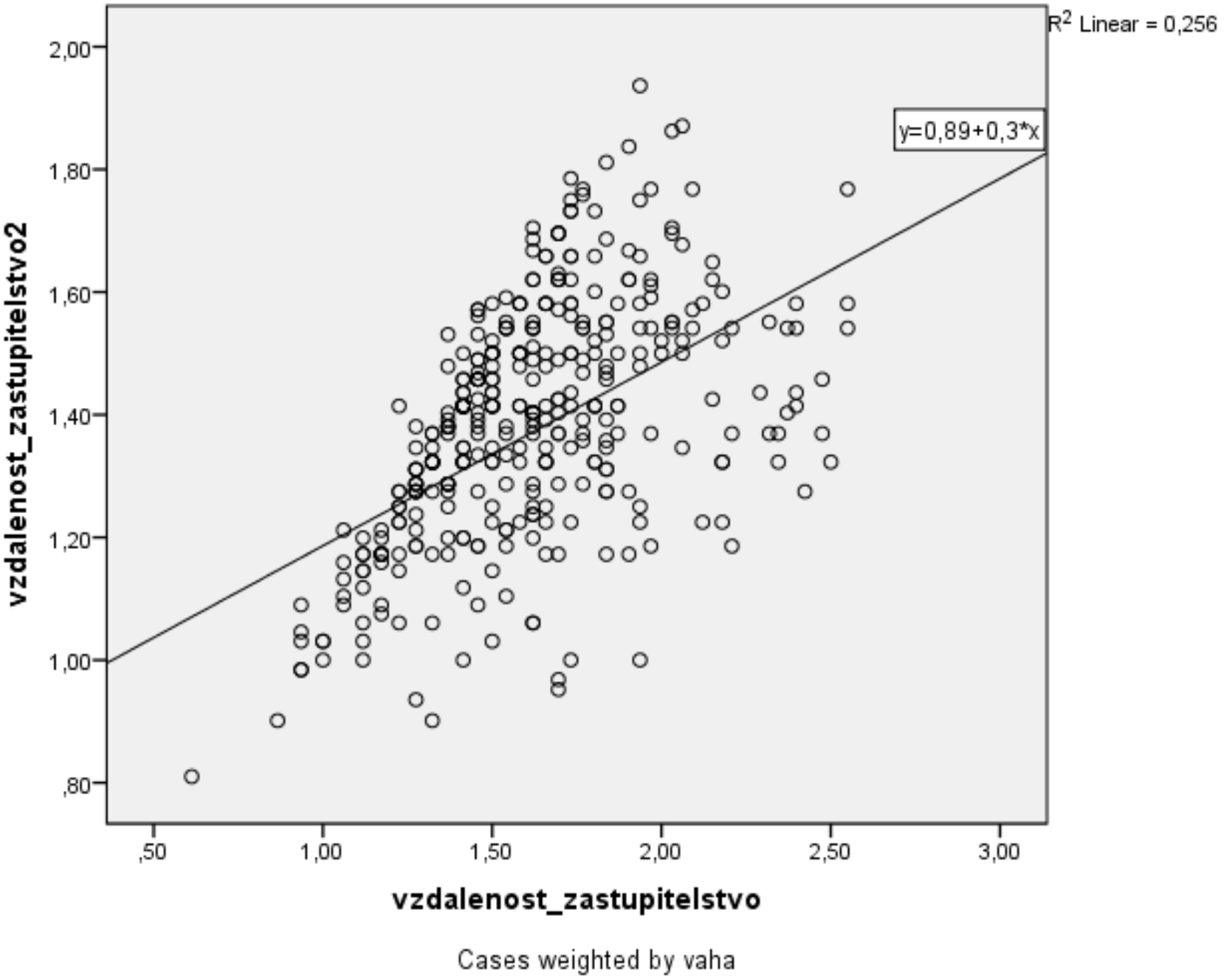


Kontrola linearity



Rekódování levice-pravice

- Vytvoření 2 dummy proměnných
- 0-3 = levice
- 7-70 =pravice
- Středoví voliči jsou referenční kategorií
- Efekty levice a pravice se interpretují jako rozdíl oproti středu
- (totéž platí pro vzdělání)


R2 Linear = 0,256
y=0,89+0,3*x
vzdalenost_zastupitelstvo2
2,00
1,80
1,60
1,40
1,20
1,00
,80
,50
1,00
1,50
2,00
2,50
3,00
vzdalenost_zastupitelstvo
Cases weighted by vaha

Kontrola multikolinearity

- Analyze – correlate - bivariate

Correlations

		vzdalenost_zastupitelstvo2	vzdalenost_zastupitelstvo	responzivita	zajem2	vek	levice-pravice
vzdalenost_zastupitelstvo2	Pearson Correlation	1	,506**	,141*	,046	,053	,057
	Sig. (2-tailed)		,000	,038	,499	,439	,406
	N	222	219	217	222	217	219
vzdalenost_zastupitelstvo	Pearson Correlation	,506**	1	,243**	,119	,145*	,078
	Sig. (2-tailed)	,000		,000	,078	,034	,251
	N	219	219	214	219	214	219
responzivita	Pearson Correlation	,141*	,243**	1	-,062	,099	-,067
	Sig. (2-tailed)	,038	,000		,360	,149	,328
	N	217	214	217	217	212	214
zajem2	Pearson Correlation	,046	,119	-,062	1	,108	,086
	Sig. (2-tailed)	,499	,078	,360		,112	,205
	N	222	219	217	222	217	219
vek	Pearson Correlation	,053	,145*	,099	,108	1	-,329**
	Sig. (2-tailed)	,439	,034	,149	,112		,000
	N	217	214	212	217	217	214
levice-pravice	Pearson Correlation	,057	,078	-,067	,086	-,329**	1
	Sig. (2-tailed)	,406	,251	,328	,205	,000	
	N	219	219	214	219	214	219

Naklikání modelu

- Analyze – regression – linear

- Dependent: ideolog_homog
- Independent: responzivita, ucen, maturita, znalost, koalice, levice, pravice, vzdalenost

- Statistics: colinearity dignostics, casewise diagnistocs

- Plots: Y:*ZRESID, X:*ZPRED, produce all partial plots

OK

Interpretace R^2 a adj. R^2

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,564[a]	,318	,295	,18268

a. Predictors: (Constant), vzdalenost_zastupitelstvo, znalost, ss, zajem, pravice, ucen, levice

b. Dependent Variable: ideolog_homog

- Model vysvětluje 32 % variability závisle proměnné

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3,296	7	,471	14,107	,000[b]
	Residual	7,068	212	,033		
	Total	10,364	219			

a. Dependent Variable: ideolog_homog

b. Predictors: (Constant), vzdalenost_zastupitelstvo, znalost, ss, zajem, pravice, ucen, levice

- Model je statisticky významný (tj. můžeme jeho výstupy zobecnit na populaci)

Interpretace R^2 a adj. R^2

- **neukazuje,** nakolik jsou výsledky platné v celém souboru,
- **neukazuje**, pro jaké procento voličů vztah platí
- ukazuje jak moc model vysvětluje rozptyl v závisle proměnné.
- Jak dobře model popisuje realitu (zaznamenanouv datech)
- Když je model nesignifikantní (tj. žádná z proměnných nepřispívá k vysvětlení rozptylu), tak použité proměnné nejsou vhodné,
  - nikoli, že k analýze proměnné není regrese použitelná
    - To závisí na naplnění předpokladů

Interpretace konstanty

- Nesmyslná, protože pokud volič má vzdálenost od stran 0, pak logicky musel sebe a všechny strany umístit na stejné místo na škále levice pravice
- Proto proměnnou vzdálenost rekódujeme
  - Odečteme 0,61
- V novém modelu je konstantu možné interpretovat:
- hodnota závisle proměnné očekávaná pro středové voliče s vysokoškolským vzděláním, kteří si myslí, že rada plní všechny své sliby a kteří se cítí relativně blízko všem stranám=0,794

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Col
		B	Std. Error	Beta			Tole
1	(Constant)	,794	,078		10,121	,000	
	levice	-,105	,039	-,197	-2,683	,008	
	pravice	-,123	,035	-,261	-3,557	,000	
	zajem	-,001	,001	-,039	-,662	,509	
	ucen	-,111	,035	-,225	-3,153	,002	
	ss	-,053	,030	-,120	-1,806	,072	
	znalost	-,010	,027	-,022	-,365	,715	
	vzdalenost_zastupitelstvo	,438	,048	,667	9,104	,000	

a. Dependent Variable: ideolog_homog

Interpretace nestandardizovaného beta koeficientu

- 2 situace
- Dummy proměnné x kardinální proměnné

- Interpretace efektu dummy proměnné:
  - nestandardizovaný koeficient ukazuje rozdíl dané kategorie oproti referenční kategorii
- Interpretace efektu kardinální proměnné
  - Při změně nezávisle proměnné o jednotku se hodnota závisle proměnné změní o hodnotu nestandardizovaného koeficient

Interpretace efektu dummy proměnné

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Col
		B	Std. Error	Beta			Tole
1	(Constant)	,794	,078		10,121	,000	
	levice	-,105	,039	-,197	-2,683	,008	
	pravice	-,123	,035	-,261	-3,557	,000	
	zajem	-,001	,001	-,039	-,662	,509	
	ucen	-,111	,035	-,225	-3,153	,002	
	ss	-,053	,030	-,120	-1,806	,072	
	znalost	-,010	,027	-,022	-,365	,715	
	vzdalenost_zastupitelstvo	,438	,048	,667	9,104	,000	

a. Dependent Variable: ideolog_homog

Interpretace efektu dummy proměnné

- Levicový volič vnímá strany v zastupitelstvu jako ideologicky navzájem bližší než středový volič a to o 0,1 bodu
- Nebo též
- Rozdíl mezi levicovým a středovým voličem ve vnímání ideologických rozdílů mezi stranami je 0,1. Levicový volič vnímá strany jako ideologicky bližší.

Interpretace efektu kardinální proměnné

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Col
		B	Std. Error	Beta			Tole
1	(Constant)	,794	,078		10,121	,000	
	levice	-,105	,039	-,197	-2,683	,008	
	pravice	-,123	,035	-,261	-3,557	,000	
	zajem	-,001	,001	-,039	-,662	,509	
	ucen	-,111	,035	-,225	-3,153	,002	
	ss	-,053	,030	-,120	-1,806	,072	
	znalost	-,010	,027	-,022	-,365	,715	
	vzdalenost_zastupitelstvo	,438	,048	,667	9,104	,000	

a. Dependent Variable: ideolog_homog

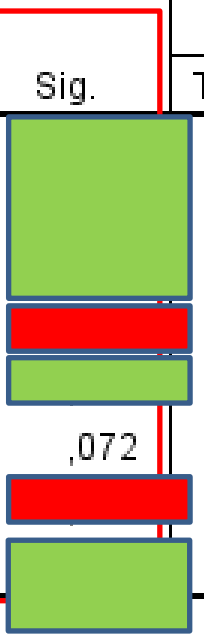
Interpretace efektu kardinální proměnné

- Pokud je volič A o 1 bod dále od všech stran než volič B, měl by volič A považovat strany o 0,44 ideologicky rozdílnější než volič B
- Nebo též
- Pokud se vzdálenost od stran zvýší o 1bod, pak vnímaná ideologická různost stran vzroste o 0,44 bodu
- Lze násobit
  - Pokud se vzdálenost od stran zvýší o 10 bodů, pak vnímaná ideologická různost stran vzroste o 4,4 bodu

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Col
		B	Std. Error	Beta			Tole
1	(Constant)	,794	,078		10,121		
	levice	-,105	,039	-,197	-2,683		
	pravice	-,123	,035	-,261	-3,557		
	zajem	-,001	,001	-,039	-,662		
	ucen	-,111	,035	-,225	-3,153		
	ss	-,053	,030	-,120	-1,806	,072	
	znalost	-,010	,027	-,022	-,365		
	vzdalenost_zastupitelstvo	,438	,048	,667	9,104		

a. Dependent Variable: ideolog_homog



Hodnocení signifikance

- Zobecňování výsledků na populaci
- Obvyklá hranice sig. < 0,05
- Potom považujeme efekt za signifikantní na hladině významnosti 95 %
- Nic nám nebrání zvolit si jinou hladinu významnosti (např. 90%, 99% nebo 99,99%)
- S nižší hladinou roste riziko, že budeme za platný považovat i efekt který v populaci neplatí
- S vyšší hladinou vyšší riziko že budeme za neplatný považovat i efekt, který v populaci platí

Honocení multikolinearity

- VIF
- Arbitární hranice: 5
- A zároveň podobné hodnoty v dimenzích

- Proměnné levice a pravice
  - V pořádku, neboť se jedná o dummy proměnné vytvořené z jedené kategorické proměnné

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	,794	,078		10,121	,000		
	levice	-,105	,039	-,197	-2,683	,008	,599	
	pravice	-,123	,035	-,261	-3,557	,000	,596	
	zajem	-,001	,001	-,039	-,662	,509	,915	1,093
	ucen	-,111	,035	-,225	-3,153	,002	,630	1,587
	ss	-,053	,030	-,120	-1,806	,072	,725	1,380
	znalost	-,010	,027	-,022	-,365	,715	,899	1,112
	vzdalenost_zastupitelstvo	,438	,048	,667	9,104	,000	,599	

a. Dependent Variable: ideolog_homog

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions							
				(Constant)	levice	pravice	zajem	ucen	ss	znalost	vzdalenost_zastupitelstvo
1	1	4,807	1,000	,00	,01	,01	,01	,01	,01	,01	,00
	2	1,265	1,950	,00	,12	,07	,00	,14	,03	,00	,00
	3	,799	2,452	,00	,11	,11	,00	,11	,23	,00	,00
	4	,439	3,308	,00	,40	,38	,01	,06	,02	,09	,00
	5	,400	3,468	,00	,08	,02	,00	,21	,23	,45	,00
	6	,194	4,975	,01	,01	,11	,23	,30	,34	,41	,01
	7	,083	7,607	,07	,03	,08	,75	,17	,11	,00	,07
	8	,013	19,248	,92	,25	,21	,01	,01	,04	,03	,92

a. Dependent Variable: ideolog_homog

Outlieři

- Pro případy č. 499 a 500 očekáváme, že budou považovat strany za relativně různorodé, ale tito respondeti si myslí, že jsou všechny stejné pozici
- Můžeme vyřadit a zjistit, co to udělá s výsledky

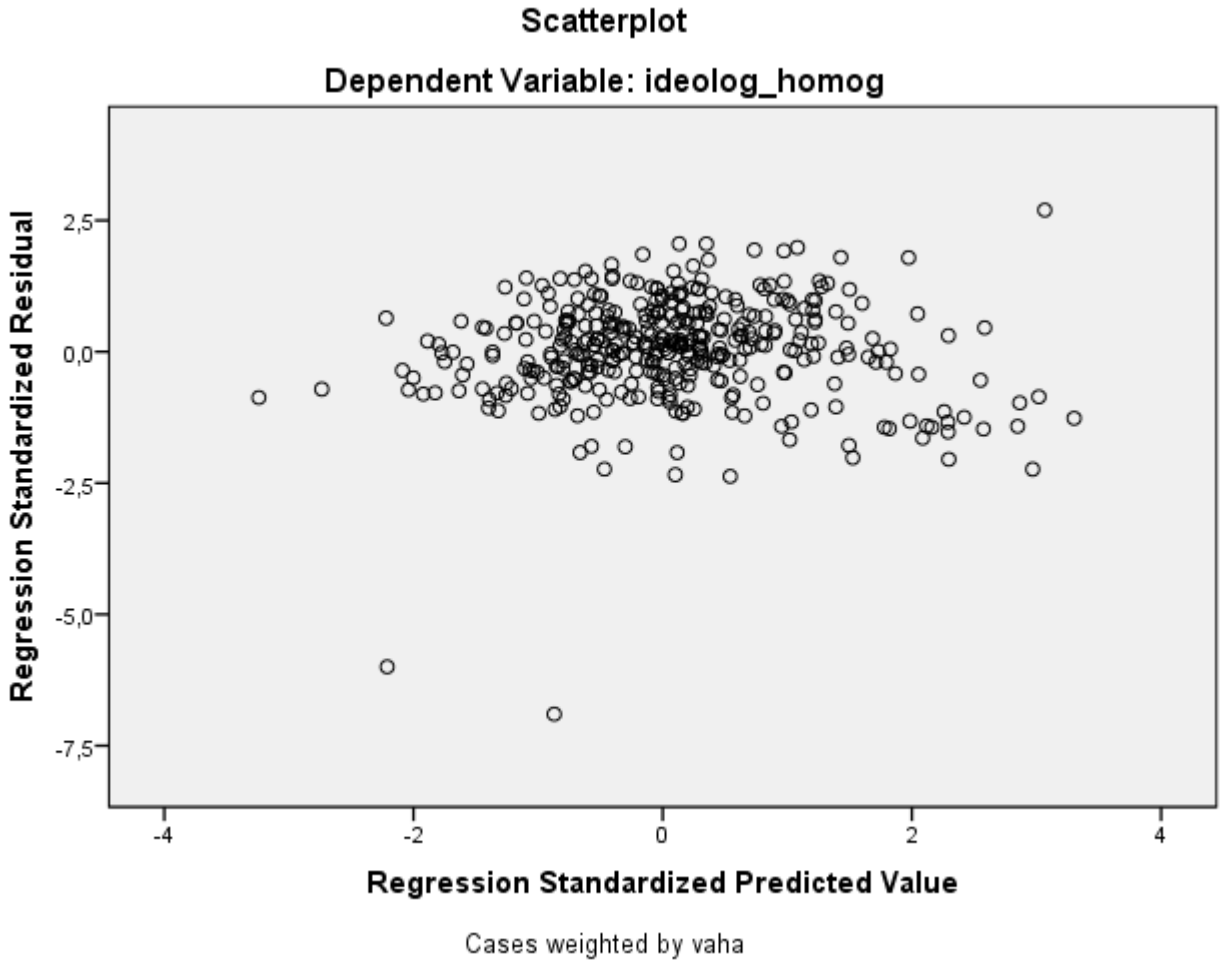
Casewise Diagnostics[a]

Case Number	Std. Residual	ideolog_homog	Predicted Value	Residual
499	-5,998	,00	1,0957	-1,09574
500	-6,899	,00	1,2603	-1,26030

a. Dependent Variable: ideolog_homog

Homoskedascita

- V reziduích by neměl být žádný zřetelný vzorec



heteroskedascita

- Příklad situace kdy homoskedascita není v pořádku