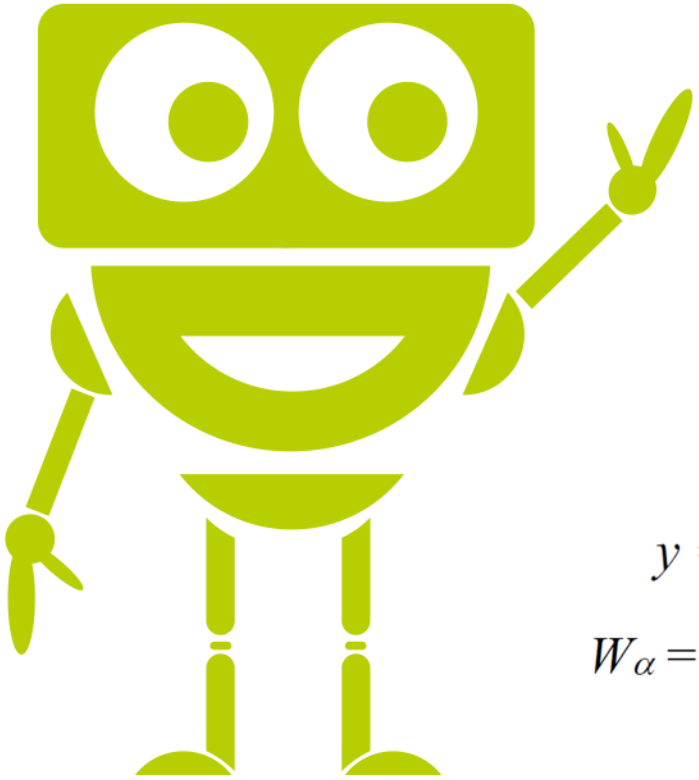
PRÁCE V SPSS

PROGRAM PRO ZPRACOVÁNÍ DAT

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \qquad \delta_t = k_t - 1 \qquad U = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

$$\bar{\Delta} y = \frac{\sum_{t=2}^{T} \Delta y_t}{T-1}$$



$$\sigma = \sqrt{D(X)} \qquad E(X) = \pi$$

$$T = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad W_\alpha = \{F;\ F \geq F_{1\text{-}\alpha}\}$$

STATISTICKY NEKLASICKY

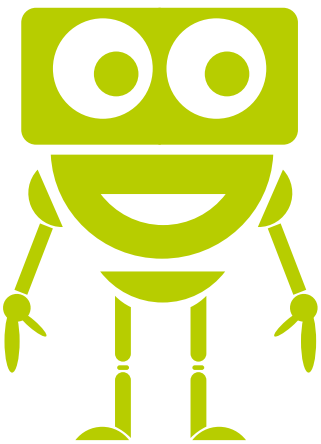
Adriana Řeháčková
www.statistickyneklasicky.cz

Práce v programu **SPSS**

V tomto materiálu si ukážeme základní práci v programu SPSS, která poslouží primárně ke zpracování statistických dat. Naučíme se základní ovládání programu, ukážeme si jak testovat hypotézy, s čím je potřeba pracovat a co je nutné ověřit.

O TOMTO MATERIÁLU:

Autor materiálu: Adriana Řeháčková
Sazba a grafické úpravy: Adriana Řeháčková



STATISTICKY
NEKLASICKY

PRÁCE V SPSS

ZÁKLADY

Data vkládáme uspořádaná ve sloupcích (první sloupec jméno respondenta, druhý věk, třetí počet získaných bodů,…)

Nejjednodušší cesta je data zkopírovat z excelu a vložit do **Data view** (strana co máte otevřenou), případně použít **File – Open –Data** a vybrat požadovaný soubor.

Na data můžeme koukat dvěma způsoby:

Detailní pohled na data umožňuje záložka **Data View** a informace o proměnných najdeme v záložce **Variable View**. Na této záložce je možné proměnné přidávat, mazat nebo měnit jejich pořadí.

PARAMETRY

name = pojmenování (nutné bez diakritiky a mezer)
type = *string* (textové proměnné)
numeric (číselné proměnné)
width = počet písmen/číslic maximálně.
decimals = počet desetinných míst
label = popis proměnné.
values = pouze u kvalitativních proměnných; určují, co znamenají kódy (0 = žena, 1 = muž)
missing = zadá se sem to, s čím nemá databáze pracovat.
measure = typ proměnných:
scale (kvantitativní proměnné)
ordinal (kvalitativní proměnná; dá se určit pořadí)
nominal (kvalitativní proměnná; nelze určit pořadí)

UŽITEČNÉ FUNKCE

seřazení veličiny -> DATA -> SORT CASES
- ascending (vzestupně), descending (sestupně)

rozdělení databáze, aby byly pohromadě veličiny dle kritéria -> DATA -> SPLIT FILES

vybrání části databáze -> DATA -> SELECT CASES
- veličiny, které nejsou vybrány, budou proškrtnuty a databáze je nebude brát v potaz. Dá se zde vybrat výběr podle nějaké podmínky (if) například věk > 30 a nebo je i možnost zcela

náhodného výběru, zde si jen vyberete kolik % hodnot chcete nebo i konkrétní velikost vzorku.

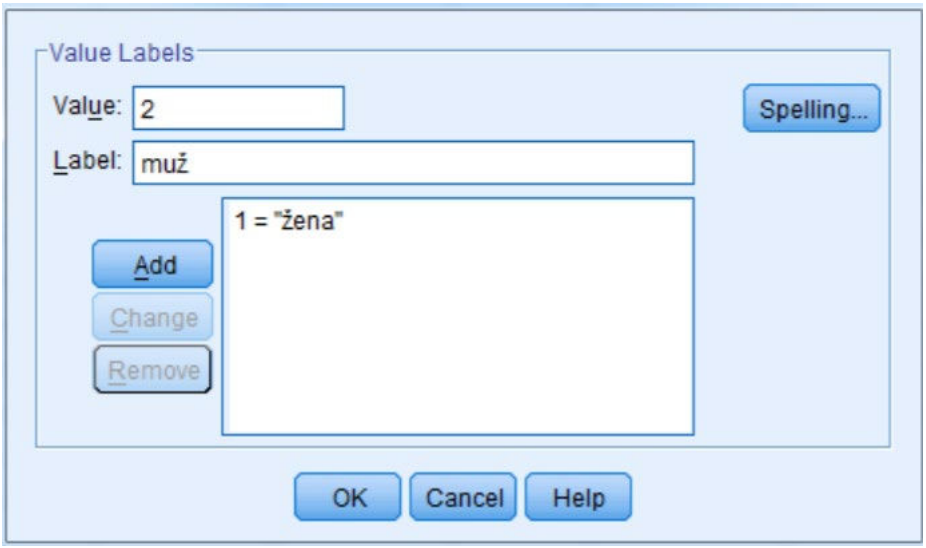
vypočítat novou veličinu -> TRANSFORM -> COMPUTE VARIABLE
- zde se zadá, co s čím se má sečíst, odečíst, apod. a dále, name a label pro novou veličinu

udělat graf -> GRAPHS -> CHART BUILDER
- nejdříve zvolit typ grafu, poté jakou/jaké veličiny chceme v grafu znázornit
- graf se zobrazí v Outputu a po dvojkliku (označení) jej lze upravovat

PŘEKODOVÁNÍ

Doporučuji dávat do SPSS data z Excelu už přepsaná, například když v dotazníku někdo bude zaškrtávat jestli je žena nebo muž, tak pak to rovnou přepsat ve formu žena = 0, muž = 1 (nebo klidně naopak, ať tu není diskriminace 😉). To samé, když někdo bude vybírat „vždycky, občas, nikdy" – potom (vždycky = 1, občas = 2, nikdy = 3).

Když máme taková data v SPSS, doporučuji přiřadit číselným kódům jednotlivých položkek textový popis. To je možné vidět a upravovat v záložce Variable View ve sloupci Values. Po kliknutí do buňky ve sloupci Values a řádku kategoriální proměnné se zobrazí dialog.



CHARAKTERISTIKY

popis proměnné – průměr, medián, apod. -> ANALYZE -> DESCRIPTIVE STATISTICS -> *EXPLORE*
- zde se zadá, pro jakou proměnnou se to počítá (*dependent list*), popřípadě můžeme rozdělit dle např. pohlaví (*factor list*) – to je může udělat i přes *Split files.*

četnost proměnných -> ANALYZE -> DESCRIPTIVE STATISTICS -> *FREQUENCIES*
- zde se zadá, pro jaké proměnné se má četnost počítat
- lze tu zjistit *modus – nejpočetnější znak* – a jeho hodnotu

srovnání proměnných ->ANALYZE->DESCRIPTIVE STATISTICS -> *DESCRIPTIVES*
- jednoduše popíše proměnnou; počet, průměr, min, max a směrodatná odchylka, můžeme si vybrat co všechno chceme zobrazit v možnostech **options**.
- vhodné pro porovnávání např. mezi muži a ženami

Mean = průměr
Median = prostřední znak
Variance = rozptyl (s^2)
Std. Deviation = směrodatná odchylka (s)
Range = variační rozpětí (R)
Minimum
Maximum

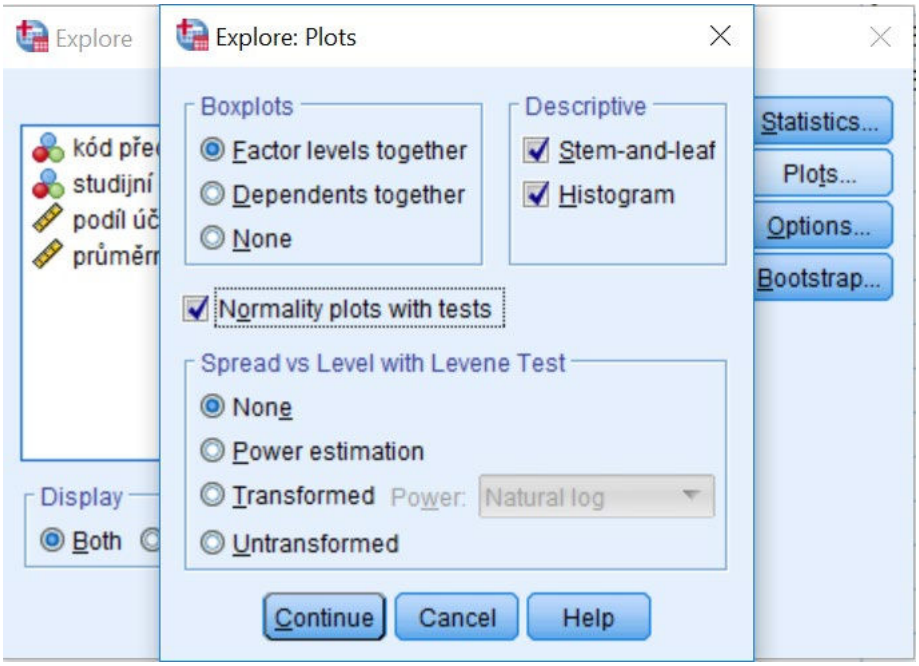
Variační koeficient – musí se dopočítat; = (směrodatná odchylka / průměr) * 100
- když je zadán rozptyl, musí se odmocnit, protože rozptyl je s^2 a směr. odch. je s.

TESTOVÁNÍ NORMALITY

Testování normality je nezbytný předpoklad pro výběr vhodného typu testu (parametrické = mají normální rozdělení, neparametrické = nemají normální rozdělení).

Postup je následující: **-> ANALYZE -> DESCRIPTIVE STATISTICS -> *EXPLORE***
A do pole **DEPENDENT LIST** vybereš proměnnou, kterou chceš testovat a následně do **FACTOR LIST** můžeš zvolit zde ji chceš podle něčeho třídit (například vyberu v dependent list výsledky z testů a ve factor list pohlaví, tím získám testy za každé pohlaví zvlášť.

U normality platí, že pokud je **sig (p-hodnota) menší než 0,05**, tak data **nemají normální rozdělení** (zamítáme H0, která hovoří o normalitě) a pokud je **větší než 0,05**



potom data **mají normální rozdělení**. Rovněž si zobrazíme histogram, kde se můžeme i vizuálně vidět, zda se data chovají podle Gaussovy křivky (mají normální rozdělení) nebo ne.

TESTOVÁNÍ STATISTICKÝCH HYPOTÉZ

- máme testy parametrické a neparametrické

PARAMETRICKÉ TESTY

- podmínkou je normální rozdělení
- jednovýběrové, dvouvýběrové a vícevýběrové testy (1 / 2 / více výběrových souborů)

JEDNOVÝBĚROVÉ TESTY

-> ANALYZE -> COMPARE MEANS -> *ONE SAMPLE T-TEST*
- vloží se proměnná z databáze, kterou chceme porovnávat, testovat ji a do **Test value** zadáme hodnotu, s kterou chceme porovnávat.
- **vyjede tabulka, ze které zjistíme-> t= testové kritérium; sig.(2 tailed)= p**

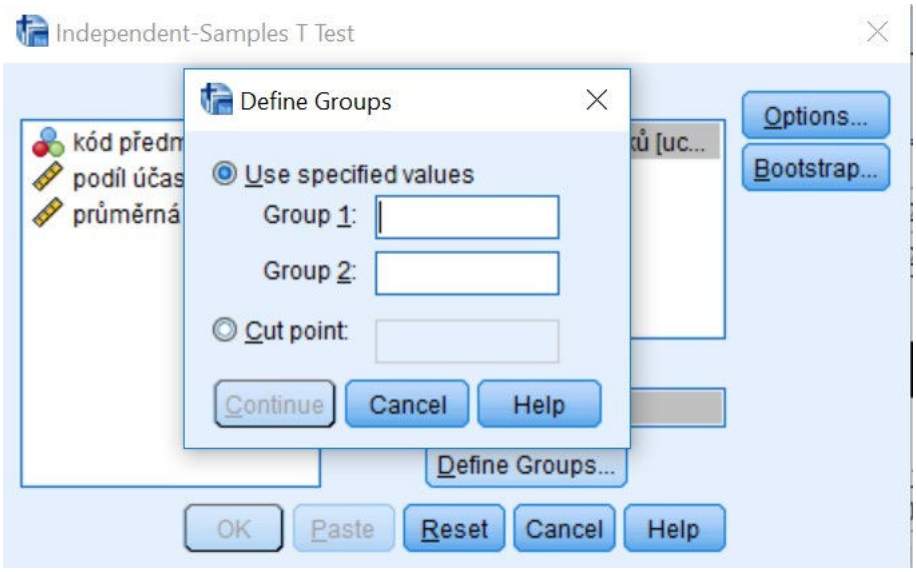
1) formulace nulové H_0 a alternativní hypotézy H_1
H_0 = shoda; předpoklad je shodný se skutečností / H_1 = alternativa; skutečnost se s předpokladem neshoduje
2) volba hladiny významnosti α, většinou 0,05
3) Výsledek sig (2 tailed (p-hodnota))nám řekne, zda H0 zamítáme nebo nezamítáme, je-li p-hodnota větší než hladina významnosti, potom zamítáme H0, jestliže není větší, tak nezamítáme H0.
4) Zamítáme H0 – existuje statisticky významný rozdíl mezi předpokladem a skutečností.

DVOUVÝBĚROVÉ TESTY

- dělí se na závislé výběry a nezávislé výběry

- závislé – když jsou dva sledované parametry na sobě závislé (1 student -> 2 testy)
  - **párový t-test -> ANALYZE -> COMPARE MEANS -> *PAIRED–SAMPLES T-TEST***
  - vloží se obě porovnávané proměnné, tak aby tvořily pár. Musí jich tedy být stejný počet.

- nezávislé – používá se, když dva parametry na sebe nemají vliv (porovnání pohlaví)
  -> ANALYZE -> COMPARE MEANS -> *INDEPENDENT SAMPLES*
  - napíše se testová proměnná -> to, co zjišťujeme (test variable) a proměnná, podle které se porovnává -> to, podle čeho se proměnné dělí (grouping variable), pak se definují kódy (co se skrývá pod 1, co pod 2)



Leveneho test pro zjištění variability (rozptylu)

-> zjistí se, jestli mají hodnoty stejné nebo odlišné rozptyly.

- test nám H_0 zamítne nebo nezamítne a to má vliv na konečné řešení (ukazuje, který údaj z tab. použijeme)
  - -> když se nezamítne, použije se první řádek „assumed“
  - -> když se zamítne, použije se druhý řádek „not assumed“

- poté se použije dvouvýběrový test (postup zápisu znovu jako u jednovýběrového), ale všechna data, která potřebujeme, jsou již obsažena v tabulce po výpočtu Leveneho testu Když se zjišťuje jen shoda rozptylu, použije se pouze výsledek Leveneho testu.

Leveneho test

1. formulace nulové H_0 a alternativní hypotézy H_1

   $H_0 : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 \neq \sigma_2^2$

2. volba hladiny významnosti α

   $\alpha = 0{,}05$

3. Výpočet v SPSS
4. Rozhodnutí o zamítnutí nebo nezamítnutí nulové hypotézy -> $p > \alpha$ -> H_0 (nezamítnutí)
5. interpretace statistického rozhodnutí

   Neexistuje statisticky významný rozdíl v kolísání hodnot v průměrné době dojížďky mezikluky a holkami.

VÍCEVÝBĚROVÉ TESTY – ANALÝZA ROZPTYLU

- předpokládá se normální rozdělení a homogenita rozptylů (Levene test by měl potvrdit H_0)
- porovnává se *m* průměrů, kde $m > 2$
- provede se test pro ANOVU -> **ANALYZE -> COMPARE MEANS -> *ONE-WAY ANOVA***
- vloží se proměnná, která se bude testovat (Dependent list) a proměnná, podle které se kritéria dělí (Factor)
- následně se klikne na možnosti (Options) a zde se zaškrtnou popisné charakteristiky (Descriptive) a test shody rozptylů (Homogenity of variance test)
- zobrazí se tři tabulky -> potřebujeme jen druhou a třetí
- v první jsou popisné charakteristiky pro jednotlivé skupiny (Descriptives)
- druhá zobrazuje test shody rozptylů, což je test nulový hypotézy (H_0) u Leveneho testu
    - zde zjistíme testové kritérium L a p-hodnotu, podle které H_0 ne/zamítneme
- ve třetí je test nulové hypotézy (H_0) průměrů
    - zjistíme testové kritérium F a p-hodnotu, podle které H_0 ne/zamítneme

- pokud nulovou hypotézu zamítneme, znamená to, že mezi danou trojicí (čtveřicí, pěticí, atd.) neexistuje shoda, musí tedy následovat podrobnější vyhodnocení

NEPARAMETRICKÉ TESTY

- není nutná znalost tvaru rozdělení zkoumané veličiny, data nemusí mít normální rozdělení.
- použitelnost pro znaky kvantitativní i kvalitativní (ordinální data)
- charakteristická je výpočetní jednoduchost, avšak je zde menší síla testů

DVOUVÝBĚROVÝ SOUBOR

- *nezávislé výběry – Mann-Whitneyův U test*

    - neparametrickou obdobou (ekvivalentem) dvouvýběrového t-testu pro dva nezávislé výběry

- testuje se hypotéza, že dva nezávislé výběry o rozsazích *m* a *n* pocházejí ze stejného základního souboru (z populací se stejným mediánem)
-**ANALYZE->NONPARAMETRIC TESTS -> LEGACY DIALOGS -> 2 INDEPENDENT SAMPLES** můžeš si vybrat, který test chceš provést a výsledná tabulka nabídne p hodnotu a testové kritérium *Z.*

- *závislé výběry – Znaménkový test nebo Wilcoxonův test*
  - neparametrická obdoba párového t-testu pro dva závislé výběry
  - ověřujeme, zda se dva závislé výběry významně liší svou polohou
  - znaménkový test má menší sílu

  ANALYZE->NONPARAMETRIC TESTS -> LEGACY DIALOGS -> 2 RELATED SAMPLES může se vybrat, který test chceme provést a výsledná tabulka nabídne p hodnotu a testové kritérium *Z.*

VÍCEVÝBĚROVÝ SOUBOR
- použije se *Kruskal-Wallisův test* (jde o neparametrický ekvivalent analýzy rozptylu)
- test nulové hypotézy, že *m* nezávislých výběrů s rozsahy n_1, n_2, … n_m pochází z téhož rozdělení

ANALYZE->NONPARAMETRIC TESTS -> INDEPENDENT SAMPLES
ANALYZE->NONPARAMETRIC TESTS -> LEGACY DIALOGS -> K INDEPENDENT SAMPLES může se vybrat, který test chceme provést a výsledná tabulka nabídne p hodnotu a testové kritérium *Z.*

Korelační analýza

Z menu **Analyze** vybereme **Correlate** a následně **Bivariate**. Do pole **Variables** přeneseme obě proměnné (absenci, věk). V nabídce Correlation Coeffiecients označíme **Pearson** (musí být splněný předpoklad normality, ten ale už víme jak ověříme).
Z výsledků nás potom zajímá hodnota **Pearsonova koeficientu**.
Další věc, kterou musíme posoudit, je statistická významnost (uvedená jako Sig. 2 tailed). Ta nám určuje, jak moc se můžeme na získaný výsledek spolehnout. Signifikance (významnost) by neměla překročit standardní p-hodnotu 0,05.

Correlations

		MAKS	MICA
MAKS	Pearson Correlation	1	,313**
	Sig. (2-tailed)		,003
	N	89	89
MICA	Pearson Correlation	,313**	1
	Sig. (2-tailed)	,003	
	N	89	89

**. Correlation is significant at the 0.01 level (2-tailed).

Pořadová korelace

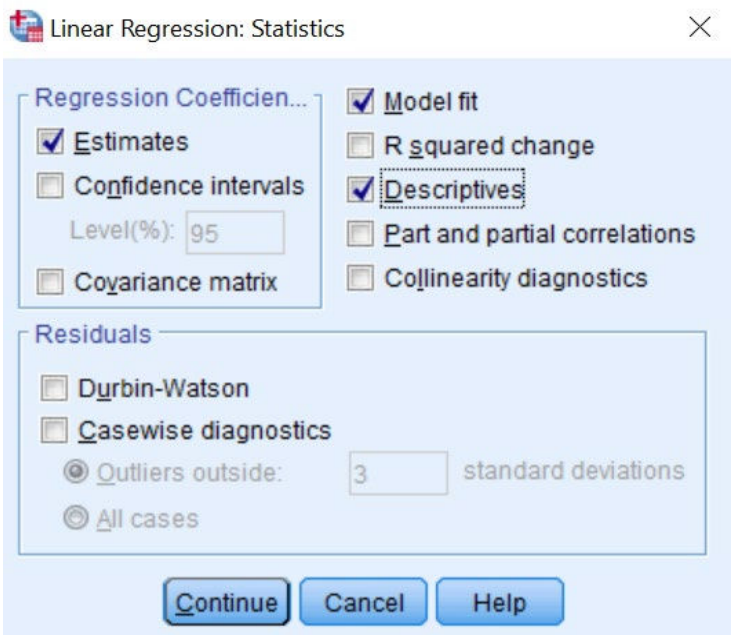
Používá se tehdy, když máme daná pořadí. Například pořadí podle vzhledu a pořadí podle oblíbenosti.
Analyze – Correlate – Bivariate. Do pole **Variables** vložíme obě proměnné (vybavenost, výdaje). V nabídce Correlation Coeficients zaškrtneme **Spearman**.

Zajímá nás hodnota **Spearmanova koeficientu** a další věc, kterou musíme opět posoudit, je statistická významnost (uvedená jako Sig. 2 tailed). Ta nám určuje, jak moc se můžeme na získaný výsledek spolehnout. Signifikance (významnost) by neměla překročit standardní p-hodnotu 0,05.

Korelační matice

Korelační matici získáme, když dáme **Analyze – reggresion – Linear** a zde si rozklineme **Statistics**…a nyní již stačí zaškrnout **Descriptives**

.



Lineární regrese

Máme jednu závislou a jednu nebo více nezávislých proměnných.
Můžeme se podívat na graf: **Graphs – Legacy Dialogs – Scatter/Dot.** Zpravidla platí, že x je nezávislá proměnná a y je závislá proměnná.

Pokud chceme použít pokročilejší metody regresní analýzy, nabídku vyvoláme z menu: **Analyze – Regression – Linear**. V dialogovém okně lineární regrese zvolíme závislou – **Dependent** (letos) a nezávislou – **Independent(s)** (loni) proměnnou, potvrdíme OK.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,648[a]	,420	,420	59,966

Vyjede na nás podobná tabulka z které můžeme pomocí korelačního koeficientu vyčíst sílu a směr závislosti, dále máme koeficient determinace, který nám řekne kolik % variability se daným modelem podařilo vysvětlit a upravený koeficient determinace, ten slouží po porovnání více modelů.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-20,828	3,306		-6,300	,000
	vek v letech	2,147	,075	,335	28,458	,000

a. Dependent Variable: prijmy domacnosti ($)

Následně můžeme vidět hodnoty konstanty a regresního koeficientu, vedle následně máme hodnotu významnosti, podle které rozhodneme jak statisticky významný regresní koeficient a konstanta jsou.

Jiné regresní funkce

Opět stačí jít přes **Analyze – Regression** a zde **Curve Estimation** potom do pole **Dependent** vybrat závislou proměnnou a do pole **Independent** nezávislou proměnnou.
Zároveň si zde můžete vybrat jaké modely (funkce) má program vytvořit, následně pak zvolíme ten nejvhodnější model, zpravidla podle **upraveného koeficientu determinace**.