Lineární regrese

První video Petra Soukupa https://www.youtube.com/watch?v=r28x1NYeiyA

Jde o vysvětlení tohoto výpočtu v Excelu (česky). Video trvá 55 minut, ale teprve od 4. minuty začínají informace o lineární regresi.

Druhé video Petra Soukupa se stejnými proměnnými - **seminář o lineární regresi v JASP (ale anglicky), se stejnými výpočetními příklady a stejným vysvětlením:** https://www.youtube.com/watch?v=Sxuc1Vvxidk

Lineární regrese je založená na principu Pearsonových korelací. I u lineární regrese pracujeme s kardinálními/spojitými proměnnými, které mají normální rozložení. Mezi nezávisle proměnnými však mohou být i dichotomní proměnné se dvěma možnostmi (např. muži-ženy; dostudoval-nedostudoval; onemocněl-neonemocněl). Tak jako v případě těchto korelací, i u lineární regrese předpokládáme lineární vztah mezi proměnnými, který se dá vyjádřit **přímkou**. V regresním modelu však může být nezávisle proměnných víc.

Lineární regresní analýzu lze použít jen tehdy, když předpokládáme, že mezi proměnnými existují lineární vztahy (tj. přímo úměrné, nebo nepřímo úměrné vztahy).

Výpočet hledá přímku, kterou se proloží data vztahu mezi dvěma proměnnými, a to tak, aby byly všechny body od ní vzdálené co nejméně. Jde o tzv. *metodu nejmenších čtverců*, viz níže.

Rovnice pro tuto přímku (tj. pro regresní analýzu) je: $$\mathbf{Y = b_0 + b_1.x}$$

B_0 je konstanta, tj. průsečík přímky s osou **Y**.

Zbývající vzorec obsahuje informaci, zda přímka jde nahoru, dolů, či je vodorovná. Říká, že s růstem hodnoty X nějak rostou (nebo klesají) i hodnoty na ose „y“.

Kladná hodnota **b_1** znamená, že s růstem hodnoty X roste hodnota Y. Jde o přímou úměrnost (záporná hodnota **b_1** znamená úměrnost nepřímou). Pokud je hodnota **b_1** kladná a vyšší než 1, znamená to, že na ose Y rostou hodnoty ještě o něco rychleji než na ose X. Pokud je tato hodnota kladná, ale nižší než 1, rostou hodnoty na ose Y pomaleji než na ose X, ale pořád jde o přímou úměrnost.

Čím víc jsou body blízké regresní přímce, tím je hodnota lineární regrese vyšší (=bližší hodnotě 1). Přesněji tuto informaci o blízkosti bodů k regresní přímce vyjadřuje **koeficient determinace R^2**, který známe z výpočtu korelací a také se z korelací počítá ($R^2 = r^2 . 100$).

Klademe si otázku, jak velký vliv má jedna (nezávisle) proměnná na druhou (závisle) proměnnou? Například jaký vliv má vzdělání na výdělek? Jedna z proměnných je nezávislá proměnná (=zde vzdělání) a druhá je závislá (zde příjem). Rozhodnutí, která z proměnných je nezávislá a která je závislá, je na nás, kteří tuto úlohu analyzujeme. Zpravidla je nezávislá proměnná ta, u které předpokládáme, že začala respondenty ovlivňovat dříve než proměnná

závislá. Nezávislými proměnnými bývají např. věk, pohlaví, sociodemografické údaje, rodinný výchovný styl apod., závisle proměnnými bývá něco, co podle našeho předpokladu vzniká později než proměnná nezávislá (např. ve vztahu k rodinnému původu jakožto nezávisle proměnné by mohlo být závisle proměnnou dosažené vzdělání).

Tak jako v případě korelací nemůžeme považovat jednu proměnnou za příčinu té druhé, tak ani u regresní analýzy nelze pokládat nezávisle proměnnou za příčinu proměnné závislé. Sledujeme, do jaké míry spolu souvisejí, ale málokdy se jedná o vztah kauzální.

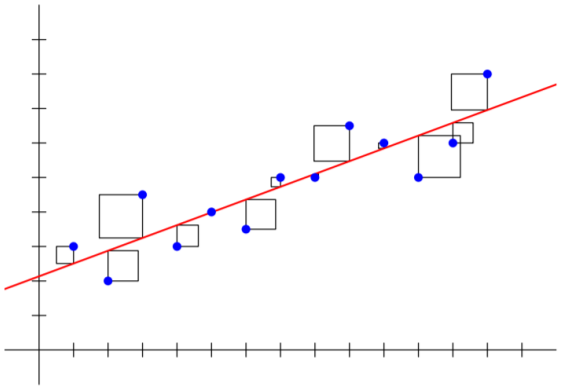
Lineární regresní analýza umožňuje také odhad: Získali jsme data u určité skupiny (populace) o obou proměnných. Pak můžeme i u dalších lidí, kteří se této populaci podobají, a my známe jejich hodnotu proměnné X, dopočítat pravděpodobnostní hodnotu i pro proměnnou Y.

Máme-li víc nezávisle proměnných, potřebujeme, aby mezi nimi nebyl příliš silný korelační vztah. Kolinearita (tj. prokorelovanost) by měla být mezi nimi zhruba nižší než 0,6.[1]

Při lineární regresi vycházíme **z metody nejmenších čtverců**:

Tato metoda se snaží najít takovou regresní přímku, aby od ní byly všechny body (=všechna měření) co nejméně vzdálené. Matematicky jde o součet druhých mocnin těchto vzdáleností (proto čtverců, viz obrázek).

Metoda nejmenších čtverců



V prvním videu dále najdete popis způsobu, jak lze zadat výpočet lineární regresní analýzy v Excelu.

Zajímavé jsou zde i informace o rozložení hodnot ***t***, tedy výsledků ***t*-testu**. V rozmezí hodnot t = -2 až +2 leží celkově 95 % naměřených hodnot. Na videu vyšla hodnota t = 15, což znamená extrémně vysokou hodnotu. V souvislosti s tím vyšla i extrémní hodnota p (se 37 nulami za desetinnou čárkou). Zjednodušeně to znamená, že mezi proměnnými existuje reálný, velmi těsný lineární vztah mezi oběma proměnnými (signifikantní hodnota t znamená nutnost zamítnout nulovou hypotézu H_0, že mezi oběma proměnnými neexistuje signifikantní vztah. Musíme přijmout alternativní hypotézu H_1 ve znění, že mezi oběma proměnnými tento vztah existuje).

[1] Na přesné určení přípustné míry kolinearity existují určité testy (např. VIP apod.) s určitou metodikou, které lze najít v učebnici Rabušice, Soukupa a Mareše (2018): Statistická analýza sociálněvědních dat (prostřednictvím SPSS). Brno, MUNI.

Hodnoty týkající se intervalu spolehlivosti/confidence interval (95 %) vypovídají o tom, že v dané lokalitě (nejen v rámci souboru, z nějž výpočet pochází) se prodává odhadem 95 % bytů za cenu za metr čtvereční v rozmezí, které je v tabulce vymezeno tímto intervalem (v JASP je to zkratka „CI“). Jedná se o odhad pro zbývající podobnou populaci.

VÝPOČET LINEÁRNÍ REGRESE V JASP

Pro počítání v JASP je zapotřebí podívat se nejprve na vztahy mezi oběma proměnnými v sekci „**Regression**“ **=>** „**Correlate**“. Zde zadáváme do okénka „variable“ proměnné tak, abychom dostali graf se správně řazenými proměnnými „X“ (vodorovně) a „Y“ (svisle).

V tomto výpočtu si zároveň zvolíme graf: „scatter plot“, „statistics“, „confidence interval“ (tj. interval spolehlivosti, viz výše).

V grafu kontrolujeme, zda v datech nemáme nějaké odlehlé (=vlivné) proměnné, které by výsledky extrémně zkreslily. Jde-li o jednu nebo několik takovýchto proměnných, bývá v rámci regresní analýzy někdy vhodné je vynechat (za předpokladu, že máme těchto dat dostatek).

Přes horní levý roh s obrázkem trychtýře („filter“) je možné vyčlenit některá data, která jsou extrémní.

Výpočet jednoduché lineární regrese v JASP:

„**Regression**“ **=>** „**Classical linear regression**“. Do kolonky „Dependent variable“ zadáváme jednu závisle proměnnou. Do „Covariates“ zadáváme nezávisle proměnné (**spojité/kardinální**, anebo **dichotomní** jen s dvěma možnými hodnotami – ty však předtím musíme předefinovat na spojité: zvolit „oranžové pravítko“ v rohu popisku proměnných).

Výpočtem získáváme:

- **R,** což je korelace mezi reálnými hodnotami závisle proměnné a jejich očekávanými hodnotami, tedy takovými hodnotami, které by ležely na regresní přímce. V případě jedné závislé a jedné nezávislé proměnné jde o hodnotu Pearsonova korelačního koeficientu.
- Dále **R^2**, což je velikost vysvětleného rozptylu závisle proměnné nezávisle proměnnou (R^2 vynásobený 100 udává velikost vysvětleného rozptylu v procentech, jde o koeficient determinace, viz výše).
- **Adjusted R^2** je upravená hodnota R^2 o chybu, která vzniká při zařazování dalších závisle proměnných. Platí totiž, že když přidáme jakoukoli další závisle proměnnou, hodnota „R^2“ se vždycky automaticky zvýší. Hodnota **adjusted R^2** bývá proto nižší a pro naši intepretaci přesnější, a proto ji uvádíme a interpretujeme i v našich závěrečných pracích.
- Dále **velikost sklonu** (angl. *slope* => tj. **o kolik se zvýší hodnota Y, pokud se hodnota X zvýší o jednu jednotku**). *V Soukupově modelu to znamená: o kolik se zvýší hodnota bytu, když se jeho velikost zvýší i 1 metr čtvereční. Je to cca 14 tis.*

- **Počátek/konstantu** (angl. ***intercept*** => **hodnota závisle proměnné Y, pokud je hodnota nezávisle proměnné X rovna nule**). *V Soukupově modelu vychází trochu nesmysl, protože (neexistující) byt o rozloze 0 m^2 by stál mínus cca 29 tisíc, tedy při nákupu takovéhoto bytu by kupující tyto peníze dostal… V tomto počátku však můžeme posunout měřítko a začít jej interpretovat např. od nejmenšího z bytů, které jsou v nabídce. Pokud by to byl byt s rozlohou např. 30 m^2, byla by jeho hodnota podle regresní rovnice: Y = 29 tis. + 14 tis. × 30 m^2, tj. 449 tis.*

  *Jinou možností je **centrování**, kdy se velikost každého bytu odečte od průměrné velikosti bytu. Centrování používáme vždy, když do modelu zadáváme větší počet nezávisle proměnných. Centrování v JASP prostřednictvím černého „plus" v pravém horním rohu, zadá název proměnné, klikne na obrázek „ruky" a „pravítka" (že jde o spojitou proměnnou) a „column". V datech se objeví nový sloupeček s názvem „**f**$_x$csize", JASP ví, že bude tento sloupeček počítat. Do horního okénka nad daty se zadává centrovaná proměnná SIZE, z horních příkazů se vybere znaménko „mínus" (bude se něco odečítat) a dopíše se hodnota průměrné velikosti bytu, tj. 64.653. Na liště níže zmáčkne „compute column". Pokud zadáme do regresního výpočtu tuto novou proměnnou, změní se pouze intercept/konstanta, která vyjadřuje hodnotu průměrně velkého bytu. Pokud máme více nezávislých proměnných, musí se upravit všechny.*

- **Parciální *t*-test** (a hodnota p) odpovídá na otázku, zda zjištění vztah mezi proměnnými můžeme zobecnit na celou populaci podobných jedinců. Zajímá nás až hodnota v posledním řádku (*t* = 15), která je extrémně vysoká (viz výše). Pokud podle hodnoty „p" vyjde významně, znamená to, že cena bytu je statisticky významně ovlivněna jeho velikostí a že tedy mezi nimi existuje statisticky významný vztah. (Pokud hodnotu „sklonu" vydělíme hodnotou „standard error", získáváme hodnotu *t*-testu.)

V první tabulce JASP automaticky počítá dva modely: jeden pro nulovou hypotézu H_0 a druhý pro alternativní hypotézu H_1. Modelu H_0 se říká nulový nebo prázdný model. Zajímá nás až hodnota ve druhém řádku: hodnota „R" je v případě jediné nezávisle proměnné výpočet Pearsonova korelačního koeficientu (viz též výše).

Tabulka Correlate je stěžejní a údaje v ní uvádíme ve svých závěrečných pracích. Zajímá nás v ní **řádek s H_1**.

Podobně jako u Excelu i JASP nás může zajímat interval spolehlivosti (viz výše). Níže uvedený popis je vlastně stejný jako u Excelu: Abychom mohli získat odhad hodnot i pro zbytek podobné populace, zvolíme v nabídce „**Confidence interval**". Jde o hodnoty 95 % CI (lower, upper) na konci nové tabulky, tj. dolní a horní hranice intervalu spolehlivosti, ve které se pohybuje 95 % celkové populace, která se svými charakteristikami podobá populaci měřené. V případě Soukupových dat jde hrubý odhad ceny ostatních bytů ve sledované lokalitě, které nejsou na prodej, a proto se nedostaly do jeho analýzy. 95% interval spolehlivosti odpovídá hodnotě „velkosti sklonu", k níž přičteme (a odečteme) dvojnásobek hodnoty „standard error". Interval spolehlivosti je vlastně přesnější, reálnější hodnotou pro běžné využití statistických výsledků než ta konkrétní hodnota velikosti sklonu. Podobně je tomu také v případě intervalu spolehlivosti ceny za byt o průměrné velikosti (viz řádek výše).

Poslední interpretovanou hodnotnou je **standardizovaný koeficient** („standardized"). Pokud máme pouze jedinou nezávislou proměnnou, rovná se hodnotě korelačního koeficientu **R**.