PSYb2520
Statistická analýza dat v psychologii II
Přednáška 2

{*Mnohonásobná, vícenásobná*} lineární regrese

Multiple linear regression

Lineární model
Lineárně-regresní model

- Vztahy mezi proměnnými umožňují **predikovat/ modelovat** hodnoty proměnné , která nás zajímá – **závislé proměnné/outcomu/výsledku *Y***
- Má-li **prediktor** *X* hodnotu x_i, jakou má asi hodnotu *Y*?
- Z mnoha možností modelování nejčastěji používáme **lineární model:**

$$Y_i = (b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_kX_{ki}) + e_i$$

Y_i jsou hodnoty závislé pro jedince i – ty modelujeme

$X_{1i} \ldots X_{ki}$ jsou hodnoty prediktorů jedince i – ty známe

$b_0 \ldots b_k$ jsou regr. koeficienty/**parametry** – ty stanovujeme, odhadujeme

e_i je reziduum, chyba, rozdíl mezi predikcí a skutečnou hodnotou Y_i

Účel modelování

- Prozkoumání vztahů mezi proměnnými
  - analyticko-konceptuální využití
  - středem zájmu jsou pak *b*
- Predikce
  - praktické využití
  - středem zájmu jsou predikované/modelované hodnoty a jejich chyba
  - na datech, kde známe hodnoty Y_i odhadneme parametry modelu – cvičná, tréninková data
  - na datech, kde neznáme Y_i, predikujeme se známou přesností

Model

- může odrážet naši kauzální představu o procesu, jímž *X* přímo, nebo nepřímo ovlivňují *Y*
  - data-generation process
- může být nekauzální, čistě asociační, korelační - prediktivní

Statisticky v tom nejsou rozdíly – ty leží v teorii a metodologii

Příklad Long2

- Y: deprese
- X: selfe (self-esteem), duv_r (důvěra k rodičům), duv_v (důvěra k vrstevníkům)

Krok 1 – Specifikace modelu

- Rozhodnutí o tom jaký model použiji - lineární
- Rozhodnutí o tom, jaké prediktory do modelu zahrnu a jaké regresní koeficienty budeme tedy odhadovat
- V jednoduchém modelu odpovídá jednomu prediktoru jeden parametr – regresní koeficient

$$\text{deprese}_i = b_0 + b_1\text{selfe}_i + e_i$$

$$\text{deprese}_i = b_0 + b_1\text{selfe}_i + b_2\text{duv\_r}_i + b_3\text{duv\_v}_i + e_i$$

Krok 1 - Specifikace

„Správnost“ modelu podmíněna

- skutečnou linearitou vztahů
- přítomností všech proměnných ovlivňujících Y
  - LOVE – Left-Out Variable Error

Krok 2 – Odhad parametrů modelu – estimation, fitting

- … odpovídá počítání *a* a *b* v PSY117
- Parametry odhadne počítač
- Odhadne je podle kritéria, které budeme chtít
  - Nejmenší čtverce – ordinary least squares OLS – minimalizuje rozptyl reziduí (sumu kvadr. reziduí)
  - Maximální věrohodnost – maximum likelihood – pro jednoduché modely stejný výsledek jako OLS
  - Mohou být i jiná kritéria

Podívejme se na to

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT deprese
  /METHOD=ENTER selfe.


GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=selfe deprese
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE
  /FITLINE TOTAL=YES.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: selfe=col(source(s), name("selfe"))
  DATA: deprese=col(source(s), name("deprese"))
  GUIDE: axis(dim(1), label("self-esteem"))
  GUIDE: axis(dim(2), label("deprese"))
  GUIDE: text.title(label("Simple Scatter with Fit Line of deprese by
self-esteem"))
  ELEMENT: point(position(selfe*deprese))
END GPL.
```

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT deprese
  /METHOD=ENTER selfe duv_r duv_v
  /PARTIALPLOT ALL.
```

Novinky oproti PSY117

- Regr. koeficienty jsou $\boldsymbol{b_0}$ (průsečík, *a, (constant)*) a $\boldsymbol{b_1}$(směrnice, *b*)
- **Beta** – standardizovaný regresní koeficient.
  - O kolik víc násobku SD proměnné *Y* predikujeme člověku, který má o 1SD proměnné X víc. S jedním prediktorem = *r*.

Interpretace regresních koeficientů

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k + e$$

- $\boldsymbol{B_i}$; $\boldsymbol{b_i}$ vyjadřuje **nárůst** *Y'* při nárůstu X_i o jednu jednotku; v jednotkách *Y*, při kontrole všech ostatních prediktorů (≈semiparciální korelace); jedinečný přínos
  - K porovnání síly prediktoru v různých skupinách, vzorcích
- $\boldsymbol{\beta_i}$; $\boldsymbol{b_i^*}$; ***BETA*** vyjadřuje **nárůst** *Y'* při nárůstu X_i o 1; jsou-li X_i i *Y* standardizovány, při kontrole všech ostatních prediktorů (≈semiparciální korelace); jedinečný přínos
  - k porovnání prediktorů mezi sebou v rámci jednoho modelu
  - k porovnání různě operacionalizovaného prediktoru v různých modelech
  - ukazatel velikosti účinku
- $\boldsymbol{b_0}$ – obtížně interpretovatelný průsečík … leda by prediktory byly **centrované**
- V různých modelech nemusí být vliv prediktoru stejný

Statistická kontrola

- Co bylo na předchozím slajdu komplikace, je vlastně velmi užitečné
- Dozvídame se efekt prediktoru očištěný o vliv ostatních prediktorů
- Doplňuje designové způsoby kontroly intervenujících
- Není samospásná, zvyšuje nároky na N

Predikované hodnoty

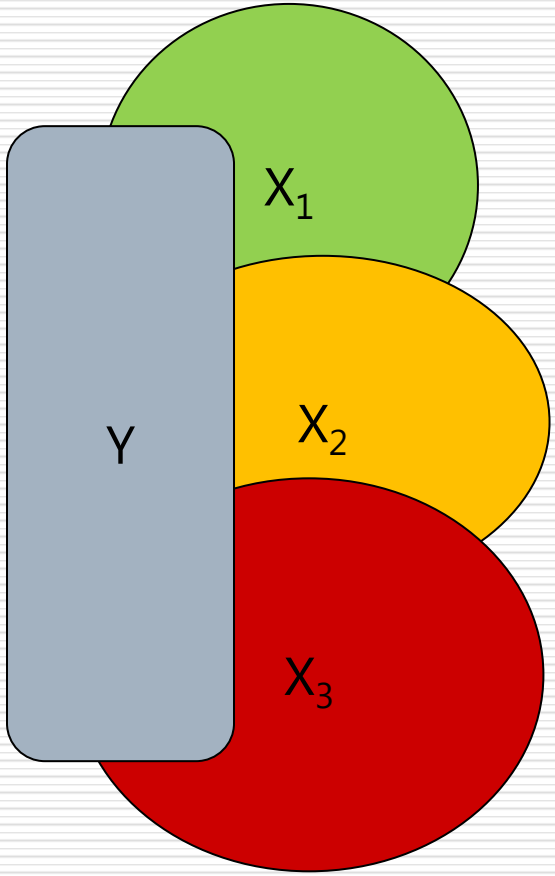
- Dosazení hodnot prediktorů do regresní rovnice – modelu
- Někdy používáme k tvorbě grafů

Krok 3 – Posouzení shody modelu s daty
Rezidua a jejich rozložení

- Měřítkem jsou rezidua – jejich „průměrná velikost“ – rozptyl
  - Samotná SD reziduí nás zajímá při predikci
  - $R^2 = (SS_{Total} - SS_{reg})/SS_{Total} \approx s^2_{res}/s^2_Y$
  - R^2 je podíl rozptylu Y vysvětlený prediktory
  - $R = r_{YY'} = r_{Y(b1X1+b2X2+ \ldots bkXk)}$
- Lze si představit i jiná měřítka
- Obvykle R^2 konstatujeme, ale nemáme na něj specifické nároky, tj. nemusí být větší než …

Rozptyl vysvětlený modelem a jednotlivými prediktory

- Část rozptylu *Y* vysvětleného dohromady všemi prediktory
- Predikční síla sady prediktorů
- Ukazatel velikosti účinku
- *R*: Mnohonásobná (mutiple) korelace
- Vždy nadhodnocuje >> při replikaci vychází nižší R^2



Krok 4 – Zvážení možných zdrojů zkreslení

Jsou případy, které model predikuje zvlášť špatně?

- Outlieři – mohou zvyšovat i snižovat *b* (jako u *r*)
- **Rezidua** – případy s vysokými r. regrese predikovala nejhůř, standardizovaná, studentizovaná ±3

Nemají některé případy příliš velký vliv na výsledky regrese?

- **Vlivné případy** – případy, které nejvíc ovlivňují parametry
  - Co se stane s parametry regrese, když případ odstraníme?
  - DFBeta – změna *b*, když se případ odstraní; standardizovaná DFbeta > 1
  - DFFit – rozdíl mezi predikovanou hodnotou a predikovanou hodnotou bez případu (adjustovanou)
  - Cookova vzdálenost > 1
  - Leverage > $2(k+1)/n$, kde k = počet prediktorů, n= velikost vzorku

- Případy s vysokými rezidui či vlivné případy **NEODSTRAŇUJEME**
  - …leda by šlo o zjevnou chybu v datech či vzorku
  - …leda by nám šlo výhradně o zpřesnění predikce (nikoli o testy hypotéz)

Krok 5 - Zobecnění ze vzorku na populaci

1. Testy signifikance

- Testy jednotlivých regresních koeficientů.
  - Testují H_0: b_k=0. (t=b/SE$_b$, t-rozložení s df=N-k-1)
- Test H_0: R^2 = 0 (ANOVA)
- Předpoklady
  - Linearita vztahů
  - Nezávislost reziduí …. Případů
  - Homoskedascita
  - Normalita reziduí
  - Žádné další proměnné nekorelují se závislou
  - Absence výrazné multikolinearity

Krok 5 - Zobecnění ze vzorku na populaci

2. Krosvalidace – R2

- Kolik rozptylu bychom vysvětlili v populaci?
  - Méně – overfitting
  - Korekce R2 (adjusted R2)
- Kolik rozptylu bychom stejným modelem vysvětlili v jiném náhodném vzorku?
  - Vzorec 9.15
  - Půlením dat – na náhodné půlce data odhadneme, na druhé zjišťujeme shodu modelu s daty.

Síla testu a velikost vzorku v MLR

Přibývá nový faktor síly testu: **množství prediktorů**
2 efekty – 2 síly: Síla detekovat R^2, síla detekovat *b*.

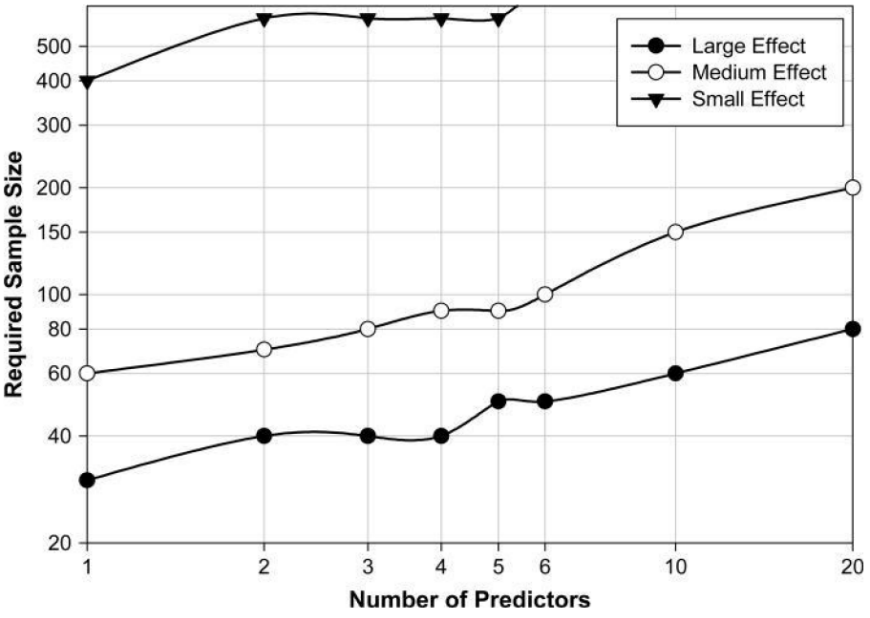


TABLE 5 Minimum R^2 That Can Be Found Statistically Significant with a Power of .80 for Varying Numbers of Independent Variables and Sample Sizes

	Significance Level (α) = .01 *No. of Independent Variables*				*Significance Level (α) = .05* *No. of Independent Variables*			
Sample Size	**2**	**5**	**10**	**20**	**2**	**5**	**10**	**20**
20	45	56	71	NA	39	48	64	NA
50	23	29	36	49	19	23	29	42
100	13	16	20	26	10	12	15	21
250	5	7	8	11	4	5	6	8
500	3	3	4	6	3	4	5	9
1,000	1	2	2	3	1	1	2	2

Note: Values represent percentage of variance explained.
NA = not applicable.

Konstanta jako model

- *M*: všem predikujeme stejnou hodnotu *c*
- $Y' = c$, $Y = c + e$
- Deviance = $\Sigma(Y_i - c)^2$
- Deviance je nejnižší, když $c = m_Y$
- Deviance = $\Sigma(Y_i - m_Y)^2$
- $s^2_{res} = \Sigma(Y_i - m_Y)^2 / (N-1)$ … tedy s^2_Y
- $s^2_{reg} = 0$ a tedy i $R^2 = 0$
- Nulový model

Možnosti práce s modely

- Odhadneme model, který jsme plánovali.
- Odhadneme řadu modelů, s postupně se rozšiřující sadou prediktorů
- - hierarchická regrese, po blocích
- Necháme nějaký algoritmus vybrat nejlepší sadu prediktorů z dostupných

- Modely srovnáváme podle R2, všímáme si i toho jak se proměňují b a jejich se

Hierarchická lineární regrese

- Bloková, se sadami (sets) prediktorů
- Prediktory vkládáme po skupinách (popř. jednotlivě) v teoreticky zdůvodněném pořadí
- Teoreticky zdůvodněné pořadí umožňuje rozdělit rozptyl Y na smysluplné části (variance partitioning)
  - Změna pořadí prediktorů změní velikost těch částí
- Zajímá nás schopnost sady prediktorů vylepšit model
  - Srovnání různých oblastí vlivu na zkoumaný jev
  - Zkoumání inkrementální validity

Obvyklá řazení bloků

- Dle času, kauzální priority
  - Př. od dispozičním k situačním…
- Od známých k neznámým vlivům
  - kontrola intervenujících proměnných
  - Minimalizace chyby 1. typu
- Podle výzkumné relevance
  - Od ústředních po „co kdyby“; maximalizace síly

Obvyklý postup regresní analýzy

- Na základě teoretických rozvah stanovíme různé modely, jejichž srovnání je potenciálně zajímavé
- Nejjednodušší srovnání je u hierarchických modelů, kdy je jeden model plně vnořen do následujícího – to umožňuje testovat inkrement (nárůst) R^2
- Až v druhé řadě se zabýváme jednotlivými regresními koeficienty v modelu, který je nejúplnější/nejlepší

Diagnostika 2: Kolinearita

- Když 2 prediktory vysvětlují tutéž část variability závislé, jeden z nich je téměř zbytečný
- Komplikuje porovnávání síly preditorů
- Snižuje přesnost odhadu parametrů (=zvyšuje jejich SE)
- V extrému (když lze jeden prediktor přesně vypočítat z ostatních) regresi úplně znemožňuje

- Korelace nad 0,9
- **Tolerance (= 1/VIF) cca pod 0,1**
- (VIF (= 1/tolerance) cca nad 10)

I při korelacích kolem 0,5 komplikuje interpretaci!!

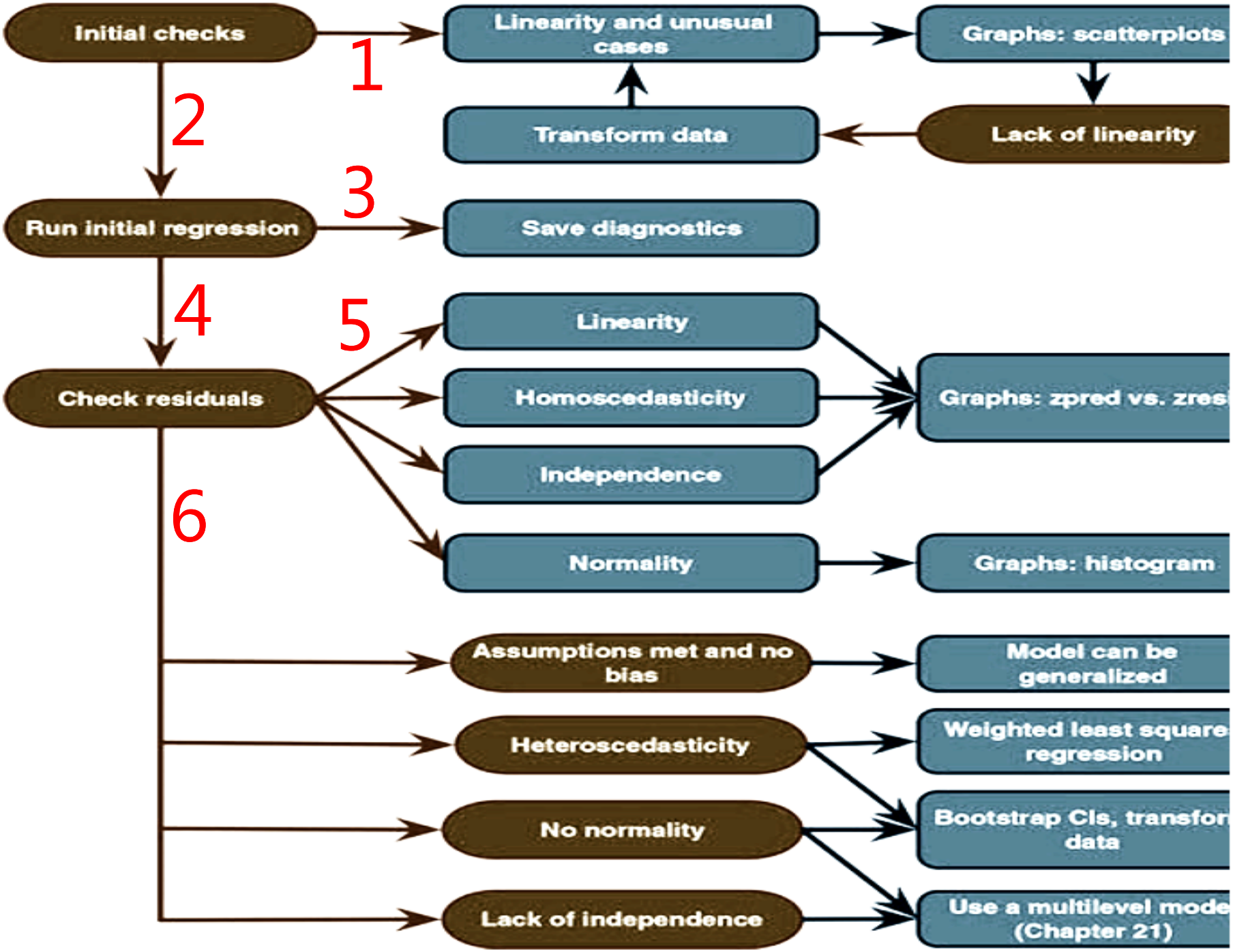
Kolinearita – poznámky a podrobnosti k VIF

- VIF – Variance Inflation Factor – Faktor zvětšení rozptylu.
  - Rozptyl o který se tady jedná, je rozptyl regresních koeficientů (tedy jejich SE^2). Důsledkem kolinearity roste – rostou tedy i CI – klesá přesnost odhadu. VIF udává míru nárůstu SE^2
- $VIF_{X_i} = \frac{1}{(1-R^2)}$, kde R^2 je z regresního modelu, v němž je X_i predikován všemi ostatními prediktory
  - VIF tak udává, jak blízko je prediktor stavu, kdy bychom ho mohli považovat za **lineární kombinaci** ostatních prediktorů.
  - Lineární kombinace je termín z lineární algebry. Když si uvědomíme, že sloupeček hodnot jedné proměnné(prediktoru) je vektor, pak se můžeme ptát, zda prvky tohoto vektoru (=hodnoty) můžeme vypočítat váženým součtem ostatních vektorů(prediktorů). Více viz třeba https://matematika.cz/linearni-kombinace-vektoru.
- Proč se nestačí podívat jen na korelační matici? Protože jeden prediktor může být téměř perfektně predikován nejen jedním dalším prediktorem (= vysoká r v korelační matici), ale i kombinací několika prediktorů – a to v korelační matici vidět není.

Reportování MLR

Základ:

- Popisné statistiky Y a X_i s korelační maticí všech
- Ujištění o naplnění předpokladů
- Popis shody modelu s daty – R^2 , p (někdy i s F-testem)
- Přehled regresních koeficientů, b, β s jejich SE, popř. s intervaly spolehlivosti, nebo p
- Limity, např. možný dopad nedokonalého naplnění předpokladů, vlivných případů apod.


Initial checks
1
Linearity and unusual cases
Graphs: scatterplots
2
Transform data
Lack of linearity
Run initial regression
3
Save diagnostics
4
5
Linearity
Check residuals
Homoscedasticity
Graphs: zpred vs. zres
Independence
6
Normality
Graphs: histogram
Assumptions met and no bias
Model can be generalized
Heteroscedasticity
Weighted least square regression
No normality
Bootstrap CIs, transfor data
Lack of independence
Use a multilevel mode (Chapter 21)