Socioekonomická analýza západního Švýcarska

Tomáš Spurný

Popis datového souboru

Datový soubor pochází z knihovny datasets. Originálním zdrojem bylo sčítání lidu v roce 1888, jež shrnul Francine Van der Walle. Dle The Office of population research zpracoval i data z sčítání lidu z let 1870, 1910 a 1930, za vzniku mnohem většího souboru, který nejen pokrýval všechny okresy Švýcarska ale i sledoval spoustu dalších proměnných. Bohužel tento soubor nebylo možno najít.

Data pocházejí tedy z období kolem roku 1888, kdy ve Švýcarsku probíhala demografická tranzice. To je jev význačný změnami v reprodukci populace, konkrétně poklesem plodnosti. Pokrývají pouze 47 francouzsky mluvících kantonů/okresů.

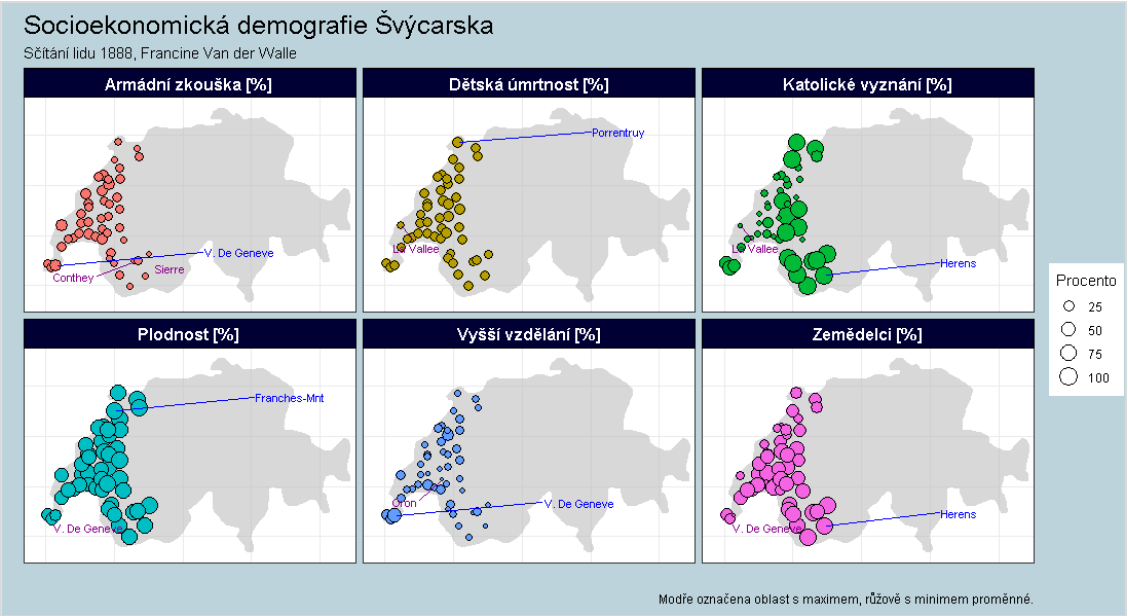
Sledované veličiny jsou (spolu s novými ekvivalenty) jsou níže:

Přehled proměnných datového souboru

Původní označení	Nové označení	Popis proměnné
Fertility	Plodnost	Počet dětí na 100 žen v reprodukčním věku
Agriculture	Zemědělci	Procento zemědělců z povolání z mužské populace
Examination	Armádní zkouška	Procento vojáků s nejlepší známkou ze zkoušky
Education	Vyšší vzdělání	Procento populace se vzděláním nad rámec základní školy
Catholic	Katolické vyznání	Procento populace s katolickým vyznáním, doplněk má protestantské
Infant.Mortality	Dětská úmrtnost	Procento živě narozených dětí, které žili méně než 1 rok
Provincy	Oblast	47 francouzsky mluvících kantonů k roku 1888

K tomu byly dohledány zeměpisné souřadnice pro jednotlivé oblasti[1].

[1] Zde se vyskytl problém, jelikož některé již zanikly nebo se rozpadly/sloučily. V těchto případech byly souřadnice určeny dle současných, nejlépe odpovídajících, místopisných názvů odpovídajících měst.

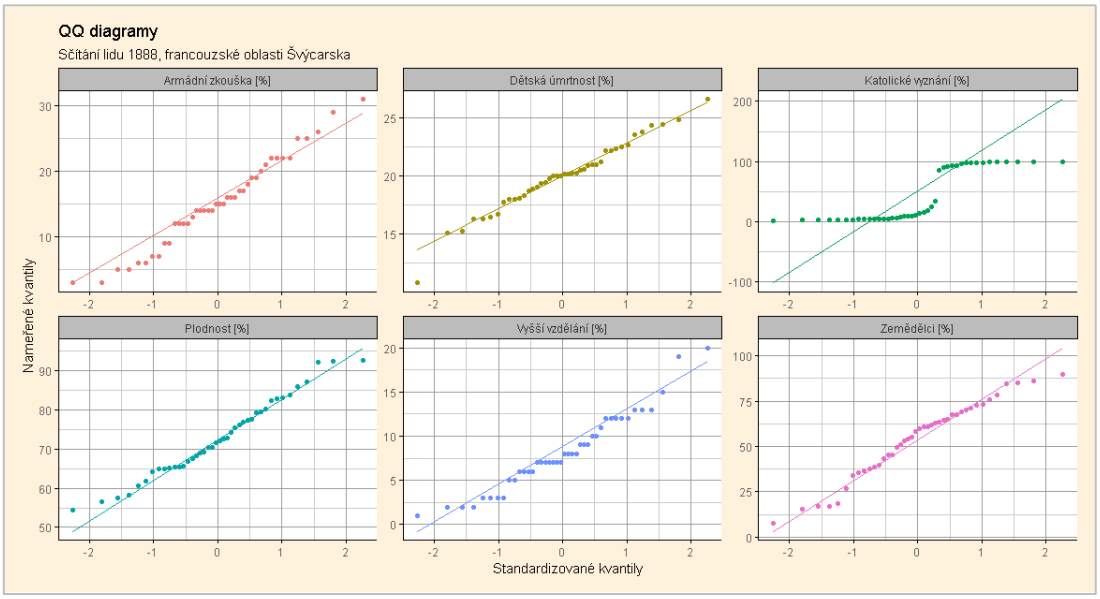


Z map lze vyčíst, že

- protestantské vyznání dominuje u hranic s Francií s relativně ostrým rozhraním.
- Nejvíc vzdělané obyvatelé jsou v okolí Ženevského a Neuchâtelského jezera. Výsledky armádní zkoušky s tím souhlasí.
- Plodnost je o něco málo menší ve východní než západní části sledované oblasti.
- Nejnižší zastoupení zemědělství je při Francouzské hranici.

Normalita dat

Byly vytvořeny QQ grafy pro každou sledovanou proměnnou:



Veličiny lze z hlediska rozpětí hodnot rozdělit na dvě skupiny. Jednu tvoří ty co se vyskytují na celé procentuelní škále (Plodnost, Zemědělci, Katolické vyznání) a druhou které se pohybují v dolních deítkách procent (zbylé tři).

Dle QQ grafů usuzovat na jisté porušení normality a odlehlé hodnoty. Byl proto proveden Shapiro-Wilkův test. Ten v případě katolického vyznání a vyššího vzdělání opravdu zamítl nulovou hypotézu.

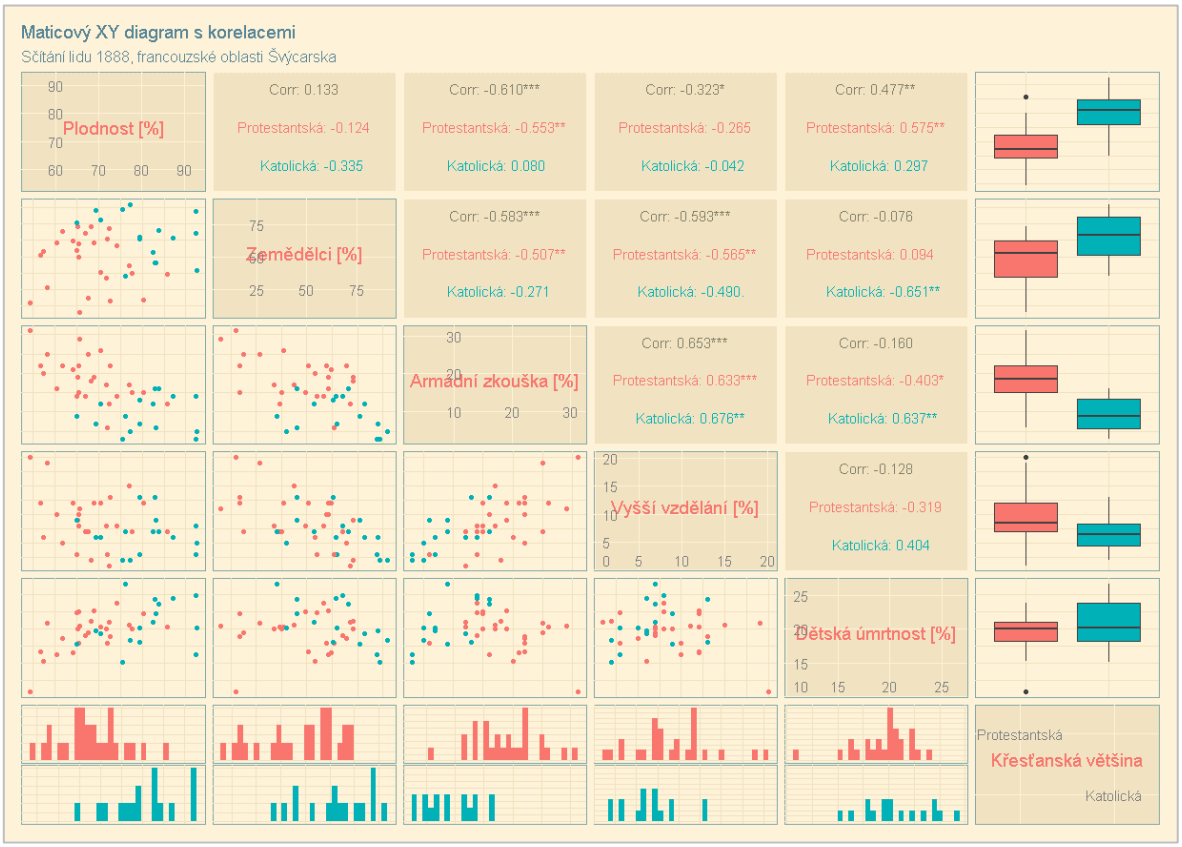
Ověření normality proměnných pomocí Shapiro-Wilkova testu

Proměnná	Původní data		Odstranění odlehlých hodnot	
	SW statistika	p-hodnota	SW statistika	p-hodnota
Armádní zkouška [%]	0,970	0,256	0,977	0,552
Dětská úmrtnost [%]	0,978	0,498	0,978	0,579
Katolické vyznání [%]	0,746	0,000	kategorizace	kategorizace
Plodnost [%]	0,973	0,345	0,975	0,495
Vyšší vzdělání [%]	0,748	0,000	0,953	0,084
Zemědelci [%]	0,966	0,193	0,964	0,201

Kvůli nenormalitě byly u proměnné Vyšší vzdělání určeny odlehlé body (celkem 5 - oblasti Lausanne, Neuchatel, Rive Droite, Rive Gauche, V. de Geneve). Proměnná Katolické vyznání byla kategorizována, dle toho, jestli je v oblasti katolická nebo protestantská většina.

Vztahy mezi veličinami

K posouzení vztahů byl vytvořen následující graf, který zahrnuje i výběrové korelační koeficienty.



Je vidět, že křesťanská většina významně ovlivňuje hodnotu ostatních proměnných – oblasti s katolickou většinou mají vyšší podíl zemědělců a plodnost, naopak menší frakci lidí s vyšším vzděláním nebo skvělým výsledkem armádní zkoušky.

Z bodových grafů jsou znát korelační vztahy, které jsou však nanejvýš střední (do 0,7). Nejvyšší korelaci mají oblasti v procentu lidí s vyšším vzděláním a lidí se skvělou armádní zkouškou. Nejmenší korelaci má dětská úmrtnost se zemědělstvím.

Regresní modely

Jakožto důležitý demografický ukazatel, byla modelována **plodnost**.

Jednorozměrné lineární regrese

Nejprve se přistoupilo k jednoduchým lineárním modelům. Výsledky za všechny shrnuje následující tabulka.

Jednorozměrné modely pro plodnost ve francouzských oblastech Švýcarska

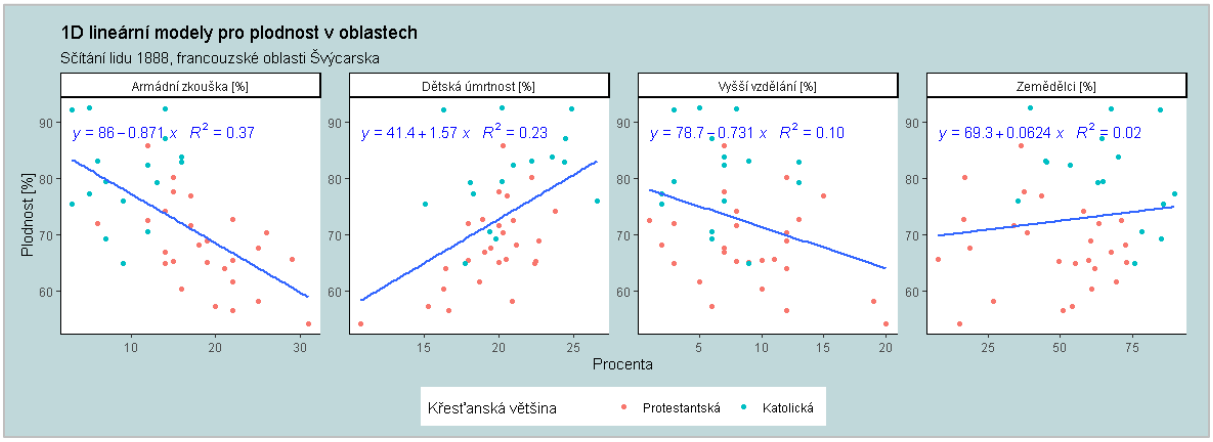
Vysvětlující proměnná	Model				Rezidua			Predikce
	Úsek	Směrnice	F-test p-hodnota	R^2	SW test p-hodnota	T-test p-hodnota	DW statistika	MAPE [%]
Zemědělci [%]	69,3	0,062	0,4027	0,018	0,3563	1	1,77	11,3
Armádní zkouška [%]	86,0	-0,871	0,0000	**0,372**	0,3516	1	2,08	**9,1**
Vyšší vzdělání [%]	78,7	-0,731	0,0370	0,104	0,0908	1	1,76	10,9
Dětská úmrtnost [%]	41,4	1,568	0,0014	0,228	0,1270	1	1,78	9,8
Křesťanská většina (ref: Protestantská)	67,9	12,654	8,7E-06	**0,394**	0,8748	1	1,99	**8,6**

Všechny modely jsou významné, až na model se **zemědělci**, kdy se nezamítla nulová hypotéza a postačuje tedy model konstanty. Tudíž i jen u zemědělců nebyl dílčí t-test směrnice významný. Nejrychleji roste závislost na dětské úmrtnosti – čím větší zastoupení dětské úmrtnosti v oblasti, tím vyšší zdejší porodnost. Je také jediná kladná a má nejmenší úsek. S ostatními proměnnými plodnost klesá – nejvíc u skvělé armádní zkoušky.

Nejvíce variability je schopen vysvětlit model **s křesťanskou většinou** a model s armádní zkouškou. Nejmenší střední absolutní procentuální chybu predikce má také křesťanská příslušnost.

U každého modelu vyšly **předpoklady u reziduí v pořádku** – hypotézy o nulovosti střední hodnoty, o normalitě nebyly zamítnuty na 5% hladině spolehlivosti. Durbin – Watsonovy statistiky se též drží v intervalu od 1,4 do 2,6 a rezidua jsou tedy i navzájem nezávislá. Jejich homoskedasticita byla splněna, i když ne ideálně (z diagnostických grafů).

Následující diagramy zachycují proložení regresních přímek daty.



Vícerozměrné lineární regrese

Metoda Enter – 2D modely s Křesťanským vyznáním

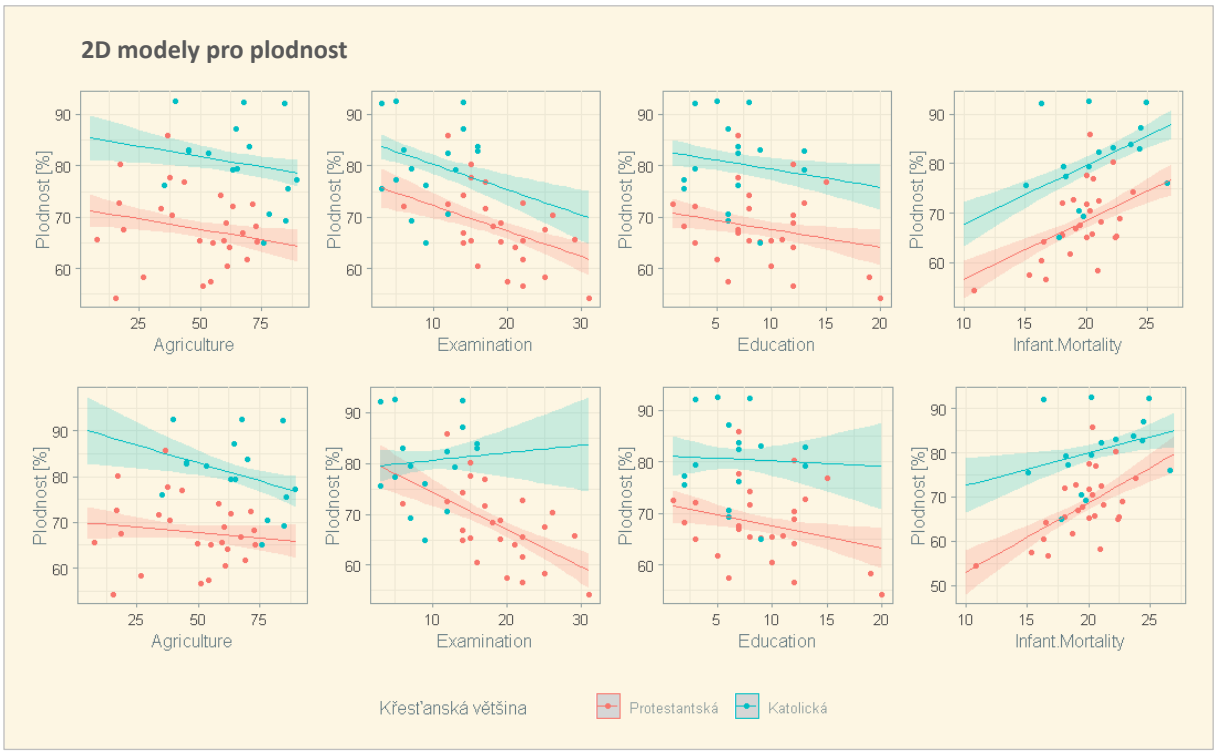
Jelikož kategoriální proměna křesťanského vyznání byla velice významnou veličinou v 1D regresích a jelikož bodových diagramů může existovat podezření na interakce, byly provedeny dvourozměrné modely se zbylými prediktory.

Dvourozměrné modely pro plodnost ve francouzských oblastech Švýcarska
(Prvním regresorem je Křesťanská většina, modely jsou s i bez interakcí)

Druhý regresor (i s operátorem)	Modely						Rezidua			Pred.
	Úsek	Směrnice Katolíci	Směrnice 2. regresor	Interakce	F-test p-hodnota	R^2	SW test p-hodnota	T-test p-hodnota	DW stat.	MAPE [%]
+ Zemědělci	71,7	14,16	-0,081	NA	2,63E-5	0,418	0,961	1	1,95	8,3
* Zemědělci	70,1	20,78	-0,046	-0,111	8,50E-5	0,427	0,898	1	1,98	8,1
+ Armádní zkouška	77,2	8,03	-0,495	NA	5,75E-6	0,461	0,893	1	2,12	8,3
* Armádní zkouška	81,6	-2,49	-0,732	0,880	6,07E-6	0,504	0,694	1	2,10	7,9
+ Vyšší vzdělání	71,1	11,75	-0,351	NA	2,82E-5	0,416	0,782	1	1,98	8,6
* Vyšší vzdělání	71,8	9,42	-0,426	0,323	1,11E-4	0,419	0,739	1	2,00	8,4
+ Dětská úmrtnost	44,9	11,11	1,183	NA	6,69E-7	**0,518**	0,194	1	2,26	7,1
* Dětská úmrtnost	37,3	28,02	1,573	-0,838	1,94E-6	**0,533**	0,494	1	2,32	**6,9**

V tabulce jsou jak modely s interakcí tak bez ní v tandemu. Směrnice interakce se týká dvojice Katolická křesťanská většina a druhý regresor. Zamítnutí dílčího t-testu na 5% hladině, a tedy statistická významnost, je zachyceno podtržením koeficientu.

Oproti jednorozměrným modelům mají všechny tyto modely vyšší indexy determinace. Všechny jsou významné oproti modelu konstanty. Nicméně u většiny, dílčí t-testy zamítly významnost koeficientů, krom úseku. Takže jen jediný model, s Dětskou úmrtností bez interakce, má všechny koeficienty významné.



Na proložených grafech lze pozorovat, že modely bez interakce mají regresní přímky rovnoběžné (čtveřice horních grafů), což vychází z teorie. Dle všech modelů mají katolické oblasti vyšší plodnost jak protestantské.

V porovnání s grafy 1D regresí, se u zemědělství změnil celý trend přímek – dle tohoto modelu mají oblasti s více zemědělci menší porodnost a protestantské oblasti ji mají vždy menší jak katolické. Nicméně už v jednorozměrný model nebyl významný.

Metoda Enter – plný model bez interakcí

Byl uvažován nový model, kde prediktory plodnosti v oblasti jsou všechny proměnné. Interakce se nebraly v úvahu. Výsledek je zachycen v následující tabulce:

Výsledky plného modelu bez interakcí

Člen modelu	Odhad	Sm. chyba	T-statistiky	VIF koeficient	p-hodnota
Úsek	69.59	10.3318	6.7357	-	1e-06
Zemědělci [%]	-0.148	0.0657	-2.2493	1.92	0.031
Armádní zkouška [%]	-0.635	0.2529	-2.5111	3,09	0.017
Vyšší vzdělání [%]	-0.158	0.3417	-0.4639	2,24	**0.646**
Dětská úmrtnost [%]	0.956	0.3529	2.7099	1,14	0.010
Křesťanská většina : Katolická	7.821	2.9108	2.6869	2,05	0.011

F-test plného modelu

Střední chyba	6.395	Stupně volnosti	5 a 36	F-statistika	12.5
Adjustované R^2	0.5838	R^2	0.6345	p-hodnota	4.54e-7
				AIC	282.58

Analýza reziduí plného modelu

SW test 1D normality (p-hodnota)	0.77	T-test s H_0: m=0 (p-hodnota)	1
Durbin-Watsonova statistika	2.092		

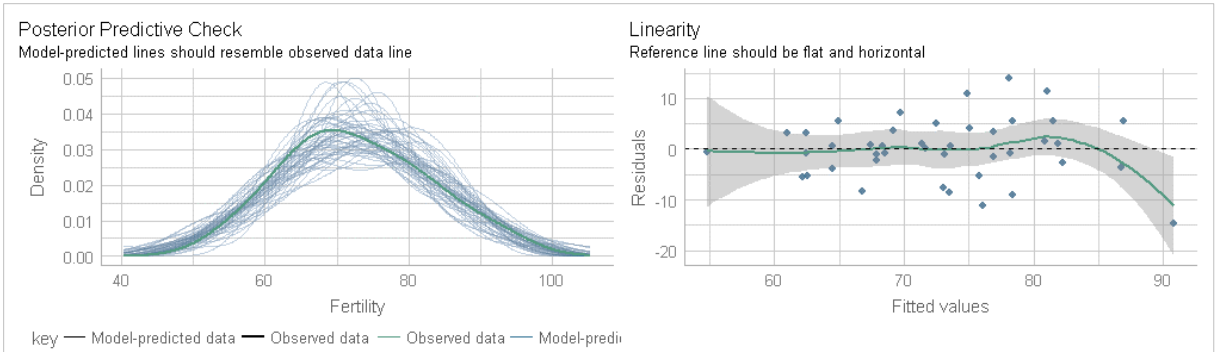
Plný model bez interakcí je lepší než model konstanty (p-hodnota F-testu) a vysvětluje 64% variability plodnosti mezi oblastmi Švýcarska. Všechny dílčí t-testy zamítly nevýznamnost koeficientů, kromě toho pro Vyšší vzdělání. Pokud by se pokračovalo metodou Enter, vyloučila by se z modelu. Pokračování je uskutečněno však Stepwise metodou.

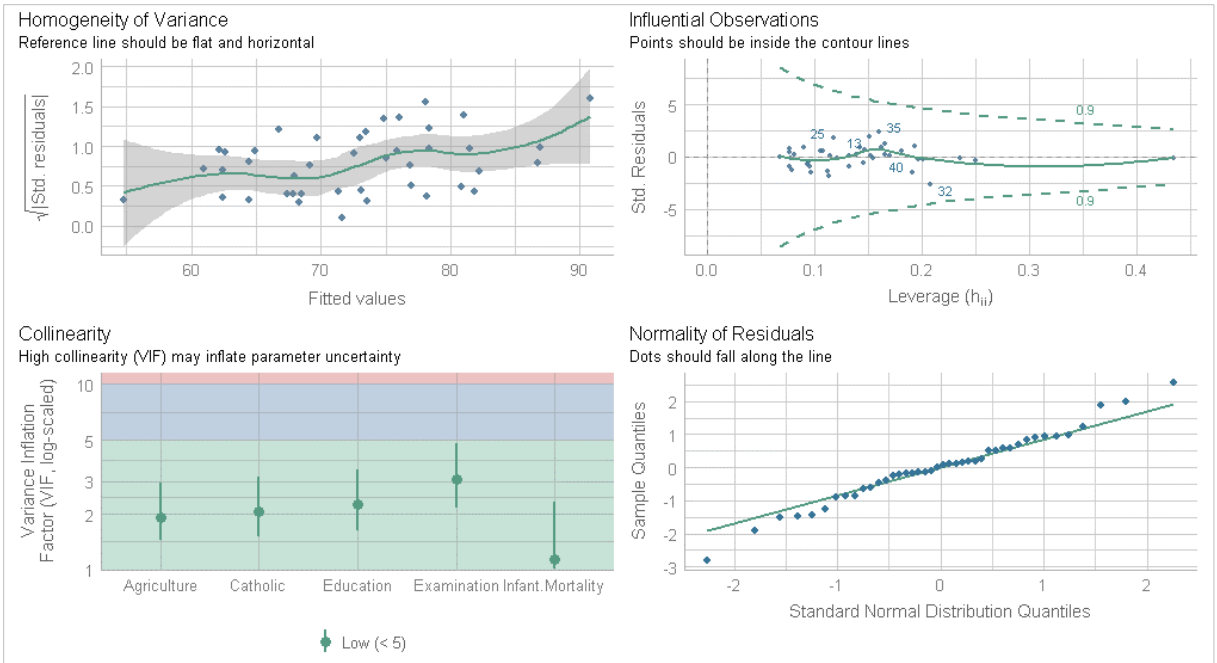
Akaikovo kritérium má rovno 282,58 a předčil tak také kterýkoli jednorozměrný model, kterým AIC nekleslo pod 295.

Koeficienty VIF svědčí při nejhorším (3,09) jen o mírné korelaci. Jelikož jsou z intervalu od 1 do 5, kdy 1 znamená žádnou korelaci mezi regresory, není multikolinearita závažná.

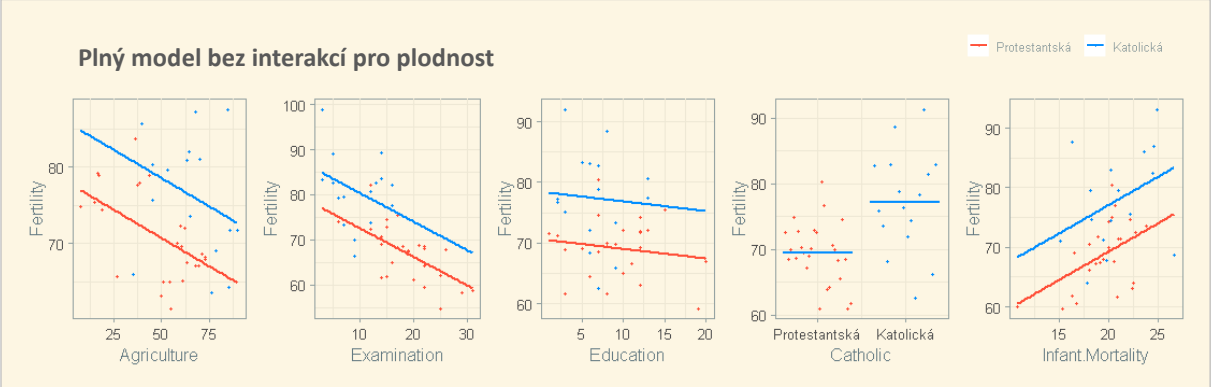
Co je zajímavé, že porodnost je výrazně vyšší v oblastech s větší dětskou úmrností a katolickou křesťanskou většinou.

Rezidua krom výše uvedených testů hodnotí i následující diagnostické grafy. Model nepodhodnocuje ani nenahodnocuje. Homogenita rozptylu není splněna dokonale.





Na výsledcích je vidět pokles plodnosti s prvními třemi proměnnými a vzestup s křesťanskou většinou katolíků a dětskou smrtností. Vzhledem k oboru hodnot je smrtnost s armádní zkouškou nejvýznamnější. Křivky u zemědělství a vyššího vzdělání jsou téměř konstantní.



Stepwise metody a nejvhodnější model dle AIC

V dalším postupu se využilo stepwise metod. Jak dopředné tak zpětné postupy od plného modelu (ať už s interakcemi nebo bez) skončili **na modelu bez interakcí se všemi prediktory krom vyššího vzdělání**.

Výsledky modelu

Člen modelu	Odhad	Sm. chyba	T-statistiky	VIF	p-hodnota
Úsek	68,16	9,7555	6,9869	-	0,000
Zemědělci [%]	-0,14	0,0594	-2,2798	1.605	0,028
Armádní zkouška [%]	-0,70	0,2127	-3,2769	2.231	0,002
Dětská úmrtnost [%]	0,98	0,3439	2,8636	1.104	0,007
Křesťanská většina : Katolická	7,38	2,7245	2,7101	1.837	0,010

F-test plného modelu

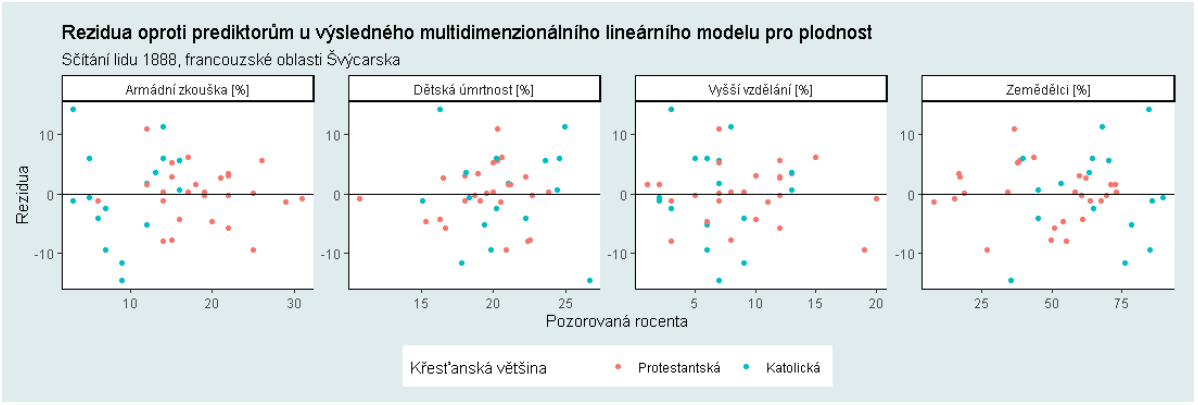
Střední chyba	6.327	Stupně volnosti	4 a 37	F-statistika	15.91
Adjustované R^2	0.5926	R^2	0.6324	p-hodnota	1.159e-07
				AIC	159.64

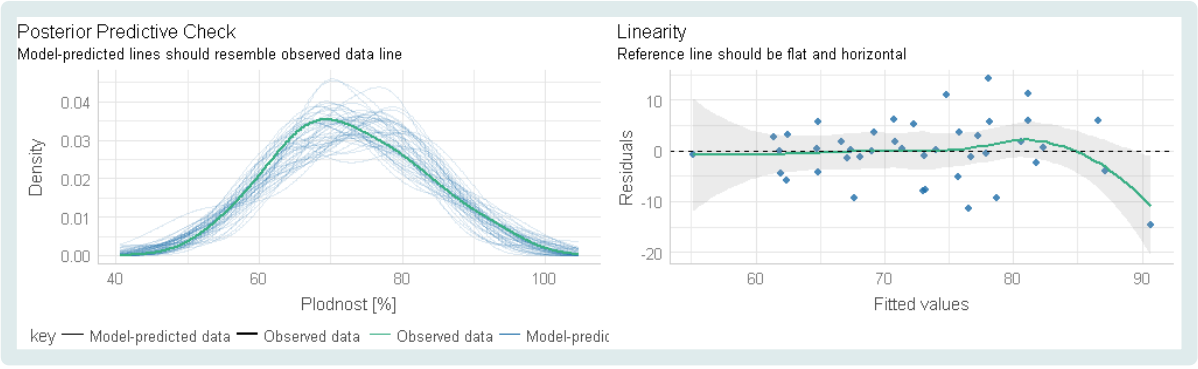
Analýza reziduí plného modelu

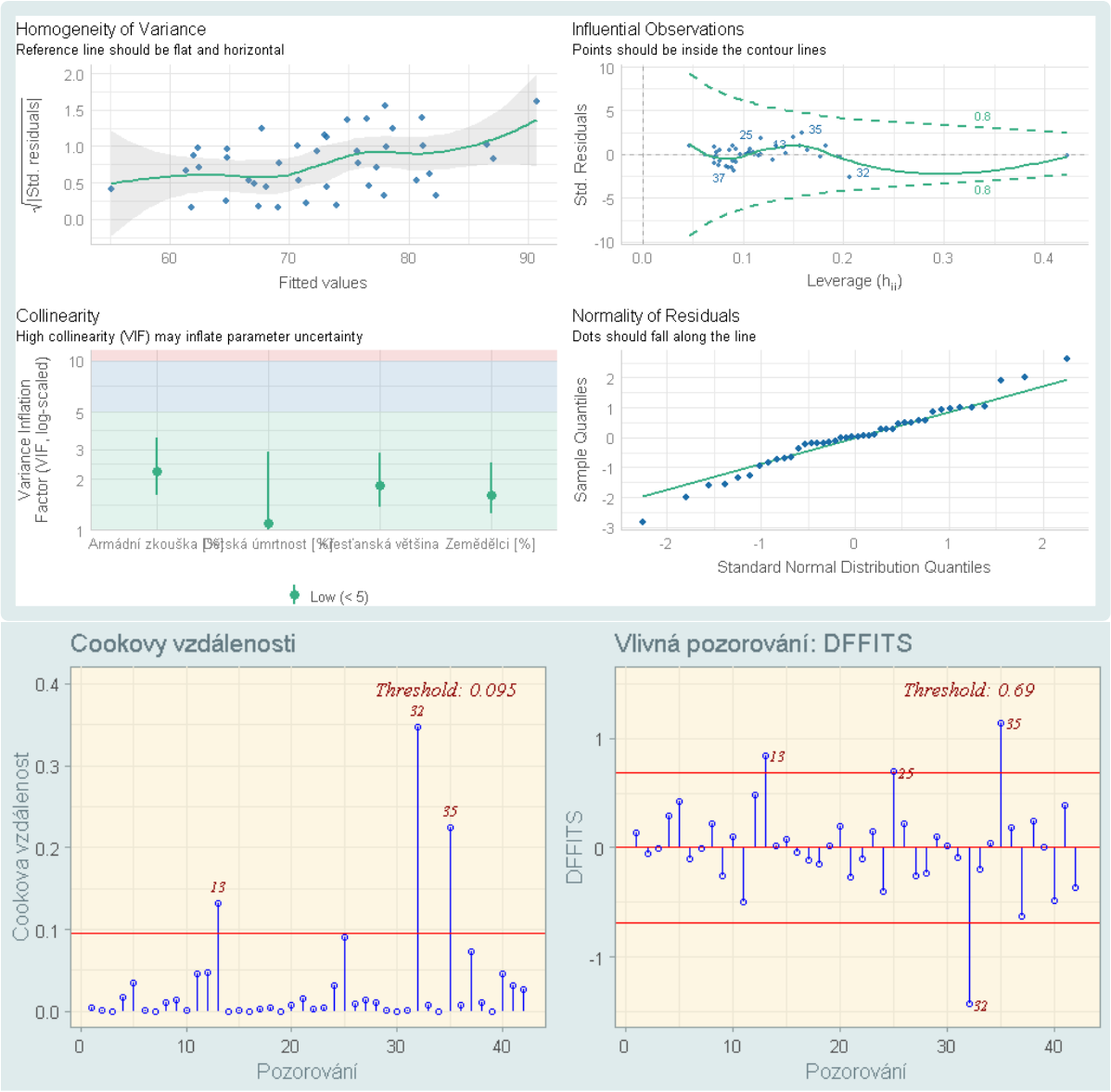
SW test 1D normality (p-hodnota)	0.7087	T-test s H_0: m=0 (p-hodnota)	1
Durbin-Watsonova statistika	2.142		

Hodnota Akaikova informačního kritéria je u tohoto modelu minimální. Variabilitu plodnosti vysvětluje model z 63%, tj. téměř stejně jako plný model a znatelně více než zkoumané 1D a 2D modely. Trend a míra závislostí se příliš nezměnil oproti plnému modelu.

Vícerozměrné ověření normality reziduí následuje v grafech níže.







Předpoklady linearity a normality reziduí lineárního modelu nejsou závažně porušeny. Homoskedasticita není extrémní. V modelu existují vlivná pozorování a to č. 13 (oblast Glane), 25 (Moutier), 32 (Porrentruy), 35 (Sierre) a 37 (St. Mauric). Ty byly opakovaně určeny ať už pomocí Cookových vzdáleností, deleční DFFITS nebo pákových bodů. Bylo vyzkoušeno je odebrat ze souboru, ale modely se významně nezměnily a nakonec stejně obsahovali vlivná pozorování.