VYSOKÁ ŠKOLA EKONOMICKÁ V PRAZE

FAKULTA INFORMATIKY A STATISTIKY

BAKALÁŘSKÁ PRÁCE

2015

Adam Zikmund

VYSOKÁ ŠKOLA EKONOMICKÁ V PRAZE

FAKULTA INFORMATIKY A STATISTIKY

Název bakalářské práce:

Ekonometrická analýza úmrtnosti při dopravních nehodách ve vybraných světových zemích

Autor: **Adam Zikmund**

Katedra: **Katedra ekonometrie**

Obor: **Matematické metody v ekonomii**

Vedoucí práce: **Ing. Elena Kuchina**

Prohlášení:

Prohlašuji, že jsem bakalářskou práci na téma „Ekonometrická analýza úmrtnosti při autonehodách ve vybraných světových zemích“ zpracoval samostatně. Veškerou použitou literaturu a další podkladové materiály uvádím v seznamu použité literatury.

V Praze dne 10. května 2015 .....

Adam Zikmund

Poděkování:

Rád bych na tomto místě poděkoval Ing. Eleně Kuchině za formální i věcné rady a připomínky při vedení mé bakalářské práce, které výrazně přispěly ke konečnému výsledku a práci jednoznačně obohatily.

Abstrakt

Název práce: Ekonometrická analýza úmrtnosti při dopravních nehodách ve vybraných světových zemích.

Autor: Adam Zikmund
Katedra: Katedra ekonometrie
Vedoucí práce: Ing. Elena Kuchina

Silniční doprava je podstatnou součástí každodenního života. Na silničních komunikacích ale velmi často dochází k nehodám, které mohou způsobit zranění zúčastněných osob. V nejhorším případě může dojít dokonce k úmrtí. Tato bakalářská práce se zabývá ekonometrickou analýzou nejdůležitějších faktorů, které mají vliv na úmrtnost. Nejdříve seznamuje s použitou metodologií a s proměnnými, které jsou pro daný problém relevantní. V praktické části je představeno několik desítek ekonometrických modelů, pomocí nichž jsou zjišťovány a kvantifikovány vlivy proměnných. V závěru jsou vyhodnoceny a interpretovány výsledky analýzy.

Klíčová slova: úmrtnost při autonehodách, ekonometrická analýza, metoda nejmenších čtverců

Abstract

Title: Econometric Analysis of Mortality in Road Accidents in Selected States of the World
Author: Adam Zikmund
Department: Department of Econometrics
Supervisor: Ing. Elena Kuchina

The road transportation is a fundamental part of everyday life. Of course car accidents can arise on the roads and cause injuries of the participating people. In the worst case the accidents can give rise to the death. This bachelor's thesis concerns with the econometric analysis of the most important factors influencing the mortality rate. At first, it presents the methodology later applied to the data set and variables relevant to the problem. In the application part there are introduced several tens of econometric models which help to discover and quantify influences of the variables. In conclusion the results of analysis are evaluated and interpreted.

Keywords: mortality in road accidents, econometric analysis, ordinary least squares

Obsah

ÚVOD ..... **- 1 -**

1 POUŽITÉ EKONOMETRICKÉ METODY **2**

1.1 REGRESNÍ ANALÝZA A KLASICKÝ LINEÁRNÍ REGRESNÍ MODEL 2

1.2 METODA NEJMENŠÍCH ČTVERCŮ 4

1.2.1 Koeficient determinace 5

1.2.2 Statistická verifikace odhadnutých parametrů 7

1.2.3 Gaussovy-Markovovy předpoklady 8

1.2.4 Heteroskedasticita 9

1.2.5 Multiokolinearita 9

1.3 DUMMY PROMĚNNÉ 10

1.4 INSTRUMENTÁLNÍ PROMĚNNÉ 11

2 DATA A PROMĚNNÉ **13**

2.1 ZDROJE DAT 13

2.2 SPECIFIKA DAT A MOŽNÉ PROBLÉMY 15

2.2.1 Počet úmrtí na silnicích 15

2.2.2 Hrubý národní důchod na obyvatele 16

2.2.3 Maximální povolená rychlost ve městě a mimo město 16

2.2.4 Maximální povolený obsah alkoholu v krvi řidiče 16

2.2.5 Povinnost nosit helmu při jízdě na motocyklu 17

2.2.6 Počet registrovaných vozidel 17

2.2.7 HDI 17

2.2.8 Délka dálniční sítě 17

2.3 VYSVĚTLOVANÁ PROMĚNNÁ 18

2.3.1 Některé hodnoty vysvětlované proměnné 18

2.4 VYSVĚTLUJÍCÍ PROMĚNNÉ A JEJICH OČEKÁVANÝ VLIV 20

2.4.1 Hustota zalidnění 21

2.4.2 Míra urbanizace 21

2.4.3 Hrubý národní důchod 22

2.4.4 HDI 22

2.4.5 Maximální povolené rychlosti 23

2.4.6 Maximální povolené množství obsahu alkoholu v krvi 23

2.4.7 Povinnost nošení helmy na motocyklu 24

2.4.8 Povinnost použití pásů v automobilu na všech místech 24

2.4.9 Počet registrovaných vozidel ... 24
2.4.10 Roční úhrn srážek ... 25
2.4.11 Regiony ... 26
2.4.12 Hustota dálniční sítě ... 26
2.4.13 Doplňující informace a některé statistiky vysvětlujících proměnných ... 27

3 APLIKACE EKONOMETRICKÝCH MODELŮ ... 31

3.1 REGRESNÍ MODELY S JEDNOU VYSVĚTLUJÍCÍ PROMĚNNOU ... 31
3.2 VÍCEROZMĚRNÉ REGRESNÍ MODELY ... 38
3.3 MODELY S INTERAKČNÍMI PROMĚNNÝMI ... 55
3.4 INSTRUMENTÁLNÍ PROMĚNNÉ ... 59

ZÁVĚR ... 63

LITERATURA ... 65

PŘÍLOHY ... 66

Seznam grafů, obrázků a tabulek

Tabulka 2-1, Vybrané statistiky vysvětlované proměnné ... 19
Tabulka 2-2, Vážený průměr míry úmrtnosti dle regionů ... 20
Tabulka 2-3, Zkratky názvů proměnných ... 28
Tabulka 2-4, Některé statistiky vysvětlujících proměnných ... 30
Tabulka 3-1, Regresní modely s jednou vysvětlující proměnnou ... 33
Graf 3-1, Závislost míry úmrtnosti na hustotě zalidnění ... 33
Graf 3-2, Závislost míry úmrtnosti na logaritmech hustoty zalidnění ... 34
Graf 3-3, Závislost míry úmrtnosti na HDI ... 35
Graf 3-4, Závislost míry úmrtnosti na HDI, kvadratická forma ... 35
Graf 3-5, Závislost míry úmrtnosti na proměnné *pasy* ... 36
Graf 3-6, Závislost míry úmrtnosti na ročním úhrnu srážek ... 37
Tabulka 3-2, Modely s regiony s nízkou mírou úmrtnosti ... 39
Tabulka 3-3, Modely s regiony vysokou mírou úmrtnosti ... 39
Tabulka 3-4, Modely s regiony s oběma extrémními hodnotami ... 40
Tabulka 3-5, Výsledky modelu 3.4 ... 41
Tabulka 3-6, Breusch-Paganův test modelu 3.4 ... 42
Tabulka 3-7, Zjišťování multikolinearity v modelu 3.4 ... 42
Tabulka 3-8, Výsledky modelu 3.5 ... 43
Tabulka 3-9, Breusch-Paganův test modelu 3.5 ... 43
Tabulka 3-10, Zjišťování multikolinearity v modelu 3.5 ... 44
Tabulka 3-11, Výsledky modelu 3.6 ... 45
Tabulka 3-12, Breusch-Paganův test modelu 3.6 ... 45
Tabulka 3-13, Zjišťování multikolinearity v modelu 3.6 ... 45

Tabulka 3-14, Výsledky modelu 3.7 46
Tabulka 3-15, Breusch-Paganův test modelu 3.7 47
Tabulka 3-16, Zjišťování multikolinearity v modelu 3.7 47
Tabulka 3-17, Výsledky modelu 3.8 48
Tabulka 3-18, Breusch-Paganův test modelu 3.8 48
Tabulka 3-19, Zjišťování multikolinearity v modelu 3.8 49
Tabulka 3-20, Výsledky modelu 3.9 50
Tabulka 3-21, Breusch-Paganův test modelu 3.9 50
Tabulka 3-22, Zjišťování multikolinearity v modelu 3.8 51
Tabulka 3-23, Výsledky modelu 3.10 52
Tabulka 3-24, Breusch-Paganův test modelu 3.10 52
Tabulka 3-25, Zjišťování multikolinearity v modelu 3.8 53
Tabulka 3-26, Běžné a robustní standardní chyby modelu 3.7 54
Tabulka 3-27, Výsledky modelu 3.12 55
Tabulka 3-28, Výsledky modelu 3.13 56
Tabulka 3-29, Modely s interakčními proměnnými 57
Graf 3-7, Závislost míry úmrtnosti na míře urbanizace v různých regionech 58
Graf 3-8, Závislost míry úmrtnosti na hustotě dálniční cítě v Asii a jinde ve světě 59
Tabulka 3-30, Potenciální instrumentální proměnné a jejich statistiky 60
Tabulka 3-31, Modely s instrumentálními proměnnými 61

Úvod

Již mnoho desítek let zažívá svět rozvoj v oblasti dopravy, ať osobní či nákladní. Výjimkou není ani doprava na silničních komunikacích, která je typická velkým počtem každodenních uživatelů. Mezi její stinné stránky patří mimo jiné poměrně častý výskyt dopravních nehod. Ty jednak komplikují dopravní situaci na dané komunikaci, ale navíc mohou být spojeny s nepříjemnými okolnostmi pro jejich účastníky – zranění, nebo v horším případě úmrtí.

Faktorů ovlivňujících vznik dopravní nehody je mnoho a jistě by nedělalo nikomu potíže vyjmenovat některé z nich, jako například rychlost vozidla, stav vozovky, zkušenosti řidiče, počasí, obsah alkoholu v krvi řidiče atd.

Cílem této práce je provést ekonometrickou analýzu, tedy pokusit se kvantifikovat velikosti vlivu jednotlivých faktorů, přičemž vysvětlovanou proměnnou bude představovat míra úmrtnosti při dopravních nehodách.

První kapitola je zaměřená na metody, které jsou použity v aplikační části práce. Základním stavebním kamenem budou klasický lineární regresní model a metoda nejmenších čtverců, ale dojde i k využití mírně pokročilejšího ekonometrického aparátu.

V druhé kapitole jsou popsána data, a to jejich původ, dále potom specifika pro různé údaje, která by mohla ovlivnit ekonometrickou analýzu (vhodnost či nevhodnost použití té či oné proměnné, odlehlá pozorování atp.). Ústřední roli při sběru dat hrál dokument Světové zdravotnické organizace *Global Status Report on Road Safety* z roku 2013 [5], ale čerpáno bylo i z dalších internetových zdrojů. V druhé části této kapitoly budou představeny proměnné modelu a budou vysloveny hypotézy o jejich očekávaném vlivu na vysvětlovanou proměnnou.

Stěžejní částí této bakalářské práce bude kapitola třetí, kde bude aplikována popsaná metodologie na získaná data. Průběžně s odhadováním jednotlivých modelů bude probíhat ekonomická, statistická a ekonometrická verifikace. Tento průběžný přístup umožňuje, aby další modely reagovaly na vzniklé problémy. Tématu verifikace tedy nebude věnována žádná samostatná kapitola. V každém případě bude představeno a odhadnuto více modelů, nejprve pomocí základních, později pomocí rozvinutějších ekonometrických metod (interakční proměnné, instrumentální proměnné a 2SLS).

Poslední část tvoří závěr, ve kterém bude zhodnocena celá bakalářská práce.

1 Použité ekonometrické metody

1.1 Regresní analýza a klasický lineární regresní model

Regresní analýza je v dnešní době hojně používaná kvantitativní metoda, jejímž úkolem je potvrzení existence vlivu jedné proměnné na druhou (nebo skupiny proměnných na jednu proměnnou a samozřejmě existují i sofistikovanější a složitější možnosti použití). Z matematického hlediska se jedná o nalezení takové funkce, která by co možná nejlépe vystihovala závislost mezi danými veličinami - proměnnými. V této regresní funkci vystupují dva druhy proměnných. Ta proměnná, jejíž hodnoty mají být pomocí regresní funkce vysvětleny, se nazývá vysvětlovaná či závislá proměnná. Naopak proměnné, které mají vliv na vysvětlovanou, jsou proměnné vysvětlující, nezávislé, nebo tzv. regresory. Regresní funkce má tedy obecně následující podobu:

$$y = f(x_0, x_1, \dots, x_k), \tag{1.1}$$

kde y představuje proměnnou vysvětlovanou, a x_0 až x_k jsou proměnné vysvětlující.

Prvním úkolem ekonometra je tedy určit takové vysvětlující proměnné, které podle něj v reálném systému mají vliv na hodnoty vysvětlované. To je samo o sobě úkol obtížný a ve většině případů také neřešitelný. Je nutno mít na paměti, že ekonometrický model (navzdory snaze ekonometra) bude vždy pouze více či méně přesným zjednodušením reality, ve které existují velmi složité vlivy a závislosti mezi různými proměnnými.

Další komplikací je to, že ve většině případů je možné pracovat pouze s omezeným vzorkem populace - tzv. výběrem. I kdyby se tedy podařilo určit řekněme většinu relevantních regresorů, kvantifikace jejich vlivu bude zkreslená (v závislosti na velikosti a vlastnostech výběru). Proto se říká, že regresní funkce (velikost parametrů) se pouze odhaduje. Při použití jiného výběru se může totiž dojít k odlišnému odhadu.

Zpět k problému s odhalením všech relevantních regresorů: v realitě je teoreticky možné, aby hodnota určité veličiny (závislé proměnné) byla naprosto dokonale vysvětlena jinými veličinami (nezávislými proměnnými). Při znalosti hodnot těchto nezávislých proměnných by potom bylo možné se stoprocentní jistotou určit hodnotu závislé. Počet všech relevantních vysvětlujících proměnných by ale byl zřejmě obrovský. Ekonometrie se z logických důvodů vzdává nalezení všech, místo toho ovšem musí zavést proměnnou, která bude vyrovnávat rozdíly mezi skutečnými hodnotami vysvětlované proměnné a hodnotami odpovídajícími regresní funkci. Tento rozdíl se nazývá náhodná složka (někdy také náhodná chyba) a značí se u.

Náhodná složka v sobě zahrnuje několik faktorů. Za prvé jsou to právě výše zmíněné opomenuté proměnné. Dalšími jsou zvolení nesprávné funkční závislosti, chyby měření, agregace dat a náhodné (stochastické) chování subjektů, které tedy nelze postihnout ani teoreticky.

Nyní bude předpokládáno, že ekonometr svůj první úkol splnil a stanovil relevantní proměnné (ne nutně všechny, ale alespoň ty nejvýznamnější, ostatní reprezentuje náhodná složka). Pro jednoduchost bude uvažována lineární forma závislosti u výše uvedené regresní funkce (1.1) a navíc bude zaveena náhodná složka u. Výsledná regresní funkce bude mít potom následující podobu: (zdroj: Hušek [1], s tím rozdílem, že vysvětlující proměnné a parametry značí od 1).

$$y = \beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u. \tag{1.2}$$

Takto specifikovaný model se nazývá klasický lineární regresní model. Lineární je z toho důvodu, že regresní parametry β jsou v lineárním tvaru (nikoli například v exponentu).

První člen rovnice (1.2) se od ostatních odlišuje. Za x_0 (speciální umělou proměnnou) se totiž pro všechna pozorování dosadí jednička, a β_0 je potom tzv. úrovňová konstanta, která vyjadřuje velikost (úroveň) vysvětlované proměnné, jestliže se za všechny ostatní regresory dosadí nula. Z geometrického hlediska je to tedy průsečík regresní funkce s osou y. Umělou proměnnou x_0 již není potřeba dále do regresní funkce zaznamenávat, a tak vzniká nový tvar regresní rovnice (1.3). Případně lze indexem i odlišit jednotlivá pozorování, regresní funkce pro i-té pozorování bude mít tvar (1.4).

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u. \tag{1.3}$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + u_i. \tag{1.4}$$

Parametry β (a tím pádem i náhodná složka u pro jednotlivá pozorování) jsou v reálném systému konkrétní hodnoty. Jak již bylo řečeno, při použití určitého populačního výběru zřejmě nebude možné přesně určit hodnoty parametrů, ale alespoň mohou být co nejpřesněji odhadnuty. Z toho logicky vyplývá, že pro konkrétní pozorování nelze zjistit ani velikost náhodné složky a opět bude pouze odhadována. Tento odhad se nazývá reziduum a je tedy rozdílem hodnoty pozorování od vyrovnané hodnoty odhadnuté regresní funkcí. Označme odhady parametrů jako b_0 až b_k a reziduum jako e. Odhadnutá regresní funkce má obecně tento tvar:

$$y = b_0 + b_1 x_1 + \cdots + b_k x_k + e. \tag{1.5}$$

Takto odhadnutou regresní funkci lze například použít k odhadu hodnoty vysvětlované proměnné, jestliže se za vysvětlující dosadí libovolné hodnoty. Toto použití znázorňuje rovnice (1.6).

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k, \tag{1.6}$$

kde $\hat{y}$ představuje odhadnutou, nebo tzv. vyrovnanou hodnotu. Reziduum v této rovnici nefiguruje, jelikož dané teoretické pozorování je vlastně součástí regresní funkce.

Zpět k prvnímu úkolu ekonometra. Tím bylo nalezení relevantních proměnných. Pokud je tento úkol splněn, k dispozici je tedy k vysvětlujících proměnných. Druhým jeho úkolem bude určení správných funkčních tvarů závislostí. Ve značné části modelů se používá závislost lineární (regresní funkcí je potom přímka, rovina a další její zobecnění ve vyšších

dimenzích). Ne všechny závislosti v realitě ale mají lineární tvar. Někdy se lze setkat například s tím, že se zvyšujícími se hodnotami regresoru hodnoty závislé proměnné nejprve klesají a později rostou (nebo přesně naopak). Pro vysvětlení závislosti potom rozhodně bude vhodnější parabola, což bude znamenat kvadratickou formu závislosti (v regresní funkci se objeví některá proměnná ve formě kvadratické funkce). Často používaná je rovněž logaritmická transformace. Těmito možnostmi specifikace modelu se bude text zabývat ještě později. Důležité ovšem je, že tyto jiné než lineární formy se týkají vysvětlujících proměnných, a nikoli regresních parametrů. Ty se vyskytují stále v lineární podobě, jinak už by se nejednalo o klasický lineární regresní model.

Třetím ekonometrovým úkolem je logicky určení metody, kterou se dospěje k odhadům parametrů. Takových metod existuje více. Významné postavení mezi nimi má metoda nejmenších čtverců, která je jako jediná použita v aplikační části práce, a je jí tedy věnována samostatná následující podkapitola.

1.2 Metoda nejmenších čtverců

Metoda nejmenších čtverců funguje na principu minimalizace součtu druhých mocnin odchylek skutečných hodnot od hodnot odhadnutých pomocí regresní funkce. Jedná se tedy o minimalizaci součtu druhých mocnin (čtverců) reziduí.

Regresní funkci se tedy ekonometr snaží „proložit“ daty tak, aby celkový součet druhých mocnin odchylek pozorování od regresní funkce byl co nejmenší. Přirozeně by existovaly i jiné způsoby. Například by se daly použít nikoli čtverce, ale pouze absolutní hodnoty odchylek. Nebo by se dala minimalizovat nejvyšší odchylka. Jistě musí existovat důvod, proč se používá nejčastěji odhad pomocí metody nejmenších čtverců.

Důvody lze jmenovat hned dva. Za prvé: takto získané odhady mají při splnění určitých předpokladů (tzv. Gaussovy-Markovovy předpoklady, kterým se bude text věnovat později) nejlepší vlastnosti oproti jiným podobně definovaným metodám z toho hlediska, že odhady parametrů jsou nevychýlené, a navíc mají nejnižší rozptyl (tato vlastnost se nazývá vydatnost). Všechny vlastnosti parametrů odhadnutých pomocí metody nejmenších čtverců shrnuje Hušek [1].

Za druhé: tyto odhady lze vypočítat pomocí tzv. odhadové funkce, která je například ve srovnání s metodou absolutních odchylek mnohem jednodušší na výpočet.

Pro ilustraci odvození odhadové funkce MNČ nejdřív vyjádřeme rovnici (1.4) nikoli rozepsanou po jednotlivých členech, nýbrž maticově (1.7).

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + u_i. \tag{1.4}$$

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}, \tag{1.7}$$

kde $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$ a $\boldsymbol{u}$ jsou následující matice a vektory: (zdroj: Hušek [1]).

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix}; \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}; \quad \boldsymbol{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}.$$

Z maticové rovnice (1.7) lze vyjádřit vektor náhodných složek $\boldsymbol{u}$.

$$\boldsymbol{u} = \boldsymbol{y} - \boldsymbol{X\beta}. \tag{1.8}$$

V případě výběrové regrese figuruje místo náhodné složky v rovnici reziduum a místo skutečných parametrů β jejich odhady b.

$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{Xb}. \tag{1.9}$$

Metoda nejmenších čtverců funguje na principu minimalizace součtu čtvercových odchylek. Tato suma se jednoduše vypočítá jako skalární součin vektoru $\boldsymbol{e}$ samotného se sebou, což je to samé, jako následující maticové násobení:

$$\boldsymbol{e}'\boldsymbol{e}. \tag{1.10}$$

Cílem je najít minimum výrazu (1.10). K tomu bude zapotřebí zjistit první derivaci a tu položit rovnu nule. Tento extrém bude jistě minimem, protože si lze snadno představit, že regresní funkci je možné nekonečně oddalovat od pozorování a tím nekonečně navyšovat sumu čtverců reziduí. Jestliže extrém existuje, potom to nutně musí být minimum. Stačí tedy dosadit za vektor $\boldsymbol{e}$ výraz z rovnice (1.9).

$$\boldsymbol{e}'\boldsymbol{e} = (\boldsymbol{y} - \boldsymbol{Xb})'(\boldsymbol{y} - \boldsymbol{Xb}). \tag{1.11}$$

Pokračuje se dalšími úpravami a již zmíněnou derivací. Celý důkaz uvádí například Hušek [1]. Výsledný tvar odhadové funkce regresních parametrů vypadá takto:

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{y}). \tag{1.12}$$

Pro existenci minima je nutné splnění podmínky existence inverzní matice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Toho je dosaženo, jestliže má matice $\boldsymbol{X}$ plnou hodnost.

Jak bylo zmíněno, jednou z velkých výhod metody nejmenších čtverců je existence jednoduché odhadové funkce, která má právě tvar (1.12). Například metoda absolutních odchylek je komplikovanější z toho důvodu, že některé odchylky jsou kladné a jiné záporné. Jejich součet je vždy nulový. Pro minimalizaci součtu se tedy musí použít absolutní hodnoty, což představuje z výpočetního hlediska určité obtíže.

1.2.1 Koeficient determinace

Jedním z ukazatelů kvality modelu jako celku je koeficient determinace, který se značí R^2. Jeho hodnota udává, jakou část variability vysvětlované proměnné vysvětluje ekonometrický model, jak uvádí Krkošková, Ráčková a Zouhar [2]. Variabilitou je myšlen rozptyl vysvětlované proměnné. Pro odvození je potřeba zavést tři následující pojmy:

SST – celkový součet čtverců (sum of squares total)

SSE – vysvětlený součet čtverců (sum of squares explained)

SSR – reziduální součet čtverců (sum of squares residual)

Při výpočtu jednotlivých sum se pracuje se třemi hodnotami vysvětlované proměnné. Jednak je to její průměrná hodnota $\bar{y}$, a dále pro každé pozorování skutečná hodnota y_i a vyrovnaná (odhadnutá) hodnota $\hat{y}_i$. Jednotlivé sumy čtverců se z uvedených hodnot vysvětlované proměnné vypočítají dle následujících vzorců, které uvádí Wooldridge [4].

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{1.13}$$

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2. \tag{1.14}$$

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{1.15}$$

Mezi jednotlivými sumami platí následující rovnost: (důkaz: Wooldridge [4]).

$$SST = SSR + SSE. \tag{1.16}$$

Pokud budou hodnoty vysvětlované proměnné hodně rozptýlené a nebudou vykazovat silnou závislost na proměnných vysvětlujících, potom lze očekávat, že odchylky skutečných a vyrovnaných hodnot (respektive čtverce odchylek) budou velké. To znamená vysoký součet reziduálních čtverců SSR. Naopak jestliže bude existovat silná závislost, pozorované hodnoty se budou jen nepatrně lišit od hodnot vyrovnaných a to bude hodnotu SSR snižovat. Naopak poroste SSE.

Pomocí regresní analýzy se samozřejmě ekonometr sanží vysvětlit pokud možno co největší část variability, takže bude žádoucí, aby vysvětlený součet čtverců SSE představoval co nejvyšší podíl na celkovém součtu čtverců SST. Tento podíl se nazývá právě koeficient determinace.

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}. \tag{1.17}$$

Jak bylo naznačeno, snahou ekonometra bude vysvětlit co největší procento variability vysvětlované proměnné. Nicméně samoúčelné zvyšování koeficientu determinace také není žádoucí. Hlavním důvodem je to, že přidáním nové proměnné do modelu koeficient determinace nikdy neklesne (nová proměnná vysvětlí alespoň nepatrnou část variability navíc). Z toho plyne, že koeficient determinace se zvýši i po přidání téměř irelevantní proměnné.

Správný analytik by ale měl sledovat nejen koeficient determinace, ale také statistickou významnost jednotlivých regresních koeficientů (to bude probráno v textu dále) a další důležité statistiky. Další možností je používat korigovaný koeficient determinace, který zachycuje navíc počet proměnných. Hodnota korigovaného koeficientu již přidáním nové proměnné klesnout může. Výpočet korigovaného (nebo také adjustovaného) koeficientu determinace uvádí opět Hušek [1].

1.2.2 Statistická verifikace odhadnutých parametrů

Předchozí podkapitola se zabývala koeficientem determinace, podle kterého je posuzována kvalita modelu jako celku. Koeficient determinace ale nic nevypovídá o statistické významnosti jednotlivých odhadů regresních parametrů $\boldsymbol{b}$. Tyto statistické významnosti jsou ale velmi důležité – především u vícerozměrných modelů. Pro tento účel se odhady statisticky testují.

Jak bylo zmíněno výše, při ekonometrické analýze se většinou pracuje s výběrovým souborem, a proto lze pouze odhadnout sílu vlivu regresorů na vysvětlovanou proměnnou (tedy odhadnout velikosti jednotlivých parametrů β, které mají v reálném systému přesnou hodnotu). Při použití jiného výběrového souboru lze dospět k odlišným výsledkům. Odhady b_j jsou tedy náhodné veličiny.

Jak již bylo uvedeno, metoda nejmenších čtverců poskytuje při splnění Gaussových-Markovových předpokladů (které uvádí následující podkapitola) nestranné odhady. Z toho vyplývá, že střední hodnotou náhodné veličiny b_j je skutečná hodnota parametru β_j.

Gaussovy-Markovovy předpoklady tedy stačí k odhadu nestranných parametrů pomocí odhadové funkce metody nejmenších čtverců. Ale pro intervalové odhady je podle Huška [1] zapotřebí ještě splnění dalšího předpokladu, a to že jsou náhodné složky rozděleny normálně.

Předpoklad normality náhodných složek bude ale s vyšší pravděpodobností splněn ve větších výběrových souborech. Jedná se vlastně o aplikaci centrální limitní věty. Jinými slovy, i kdyby se rozdělení vysvětlované proměnné od normálního odlišovalo, v dostatečně velkém souboru se bude rozdělení náhodných složek přibližovat rozdělení normálnímu. V aplikační části této práce se bude pracovat se souborem obsahujícím 136 pozorování. Obecně se považuje za dostatečnou velikost souboru přesahující 30 pozorování.

Jak uvádí Hušek [1], odhady b_j mají za splnění předpokladu normality náhodných složek normální rozdělení se střední hodnotou βj a kovarianční maticí $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$, kde σ^2 představuje konstantní rozptyl náhodných složek. Znovu se ovšem naráží na fakt, že se velikosti náhodných složek (a tím pádem i jejich rozptyl) při výběrové regresi pouze odhadují. Konstantní rozptyl náhodných složek σ^2 není známý, známé jsou však alespoň jeho nestranné odhady. Proto se při statistické verifikaci odhadů b_j používá Studentovo rozdělení. Testovací kritérium, jehož distribuce tedy odpovídá tomuto rozdělení, má následující podobu:

$$t_j = \frac{b_j - \beta_j}{s_{b_j}}, \tag{1.18}$$

kde výraz ve jmenovateli představuje standardní chybu odhadů b_j. Toto testovací kritérium se poměřuje s kvantilem Studentova rozdělení s *n-k-1* stupni volnosti (kvantil dle požadované hladiny významnosti).

Pomocí testovacího kritéria lze testovat, zda se hodnota regresního parametru významně liší od určité zvolené hodnoty. Při ekonometrické analýze je relevantní zjišťovat odlišnost od nuly. Z hlediska interpretace se tedy jedná o to, zda má daná vysvětlující proměnná nějaký statisticky významný vliv na vysvětlovanou (ať už pozitivní, nebo negativní).

Testovací kritérium lze potom psát ve zjednodušeném tvaru:

$$t_j = \frac{b_j}{s_{b_j}}, \tag{1.19}$$

který se označuje jako *t*-poměr a je běžně součástí výstupu ekonometrického softwaru. Nejčastěji se testuje významnost parametrů od nuly na pětiprocentní hladině významnosti, což znamená šanci minimálně 95%, že závislost mezi vysvětlovanou a vysvětlující proměnnou skutečně existuje.

1.2.3 Gaussovy-Markovovy předpoklady

Gaussovy-Markovovy předpoklady jsou taková kritéria, která zajišťují optimální vlastnosti odhadové funkce metody nejmenších čtverců. V literatuře je možné setkat se s různými zápisy Gaussových-Markovových předpokladů. Především někteří autoři explicitně rozepisují druhý bod na dva vlastní předpoklady. V této práci se odkážeme na předpoklady podle Huška [1], které mají následující podobu:

1. $E(\boldsymbol{u}) = \boldsymbol{0}$.
2. $E(\boldsymbol{u}\boldsymbol{u}') = \sigma^2 \boldsymbol{I}_n$.
3. $E(\boldsymbol{X}'\boldsymbol{u}) = \boldsymbol{0}$.
4. Matice $\boldsymbol{X}$ má plnou hodnost *k*.

První požadavek znamená, že střední hodnota náhodných složek je nulová.

Druhý předpoklad v sobě zachycuje dva aspekty, jak uvádí například Hušek [1]. Za prvé: náhodné složky mají mít podle druhého předpokladu konstantní rozptyl, což je požadavek tzv. homoskedasticity (prvky na diagonále kovarianční matice jsou rovny σ^2). Nesplnění by znamenalo heteroskedasticitu, tedy situaci, kdy náhodné složky vykazují různě velké rozptyly v závislosti na hodnotě vysvětlující proměnné. Za druhé: náhodné složky jsou podle tohoto předpokladu nezkorelované (prvky mimo hlavní diagonálu jsou nulové). Opačný případ by znamenal autokorelaci, tedy sériovou závislost náhodných složek. Velikost jedné náhodné složky by potom byla závislá na předchozích hodnotách náhodné složky. Tento

problém je obvyklý při práci s časovými řadami. Druhý předpoklad lze zapsat přehledněji vzorcem (1.20)

$$E(\boldsymbol{uu}') = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix} \tag{1.20}$$

Třetí předpoklad znamená, že hodnoty vysvětlujících proměnných nejsou ovlivněny náhodnou složkou, nebo jinými slovy, že $\boldsymbol{X}$ je nestochastická matice.

Čtvrtý předpoklad je dokonce nezbytný k tomu, aby vůbec existovaly odhady dle odhadové funkce metody nejmenších čtverců. Pokud by matice $\boldsymbol{X}$ neměla plnou hodnost, pak by neexistovala inverzní matice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Tento předpoklad může být porušen například neadekvátní prací s binárními proměnnými. Většina ekonometrických softwarů však dokáže tyto chyby odhalit a nepozorného uživatele ihned upozornit.

1.2.4 Heteroskedasticita

Druhý předpoklad požaduje homoskedasticitu dat. V případě, že se vyskytne v modelu heteroskedasticita, bodové odhady zůstávají nestranné a konzistentní, ale ztrácejí vydatnost i asymptotickou vydatnost (Hušek [1]). Parametry zůstávají nadále nevychýlené (alespoň asymptoticky). Může ale dojít k nepřesným odhadům standardních chyb, což má vliv na statistickou indukci. To může vést k nesprávným závěrům o statistické významnosti jednotlivých parametrů.

Místo běžných chyb, vypočtených pomocí metody nejmenších čtverců, je potřeba použít tzv. metodu robustních standardních chyb, se kterými již lze provést statistickou indukci. Tato metoda bude použita v případě heteroskedasticity v aplikační části práce.

Druhou otázkou je ale samotné testování heteroskedasticity. Ekonometry bylo v minulosti vyvinuto více metod. V této práci bude použit Breusch-Paganův test. Jeho myšlenka spočívá v regresi druhých mocnin reziduí na všech vysvětlujících proměnných. Při homoskedasticitě očekáváme nezávislost na jednotlivých regresorech. Pokud by taková závislost existovala (tedy že velikost reziduí by závisela na úrovni některé vysvětlující proměnné), znamenalo by to heteroskedasticitu. Pro ověření celého modelu se používá buďto *F*-test a nebo *LM* statistika. Druhá jmenovaná metoda bude použita i v této práci. V obou případech se ale testovací statistika porovnává s kvantilem příslušného rozdělení (v prvním případě je to *F*-rozdělení a v druhém chí-kvadrát). Pro zamítnutí nulové hypotézy o homoskedasticitě je nutné, aby *p*-hodnota těchto testů byla dostatečně nízká.

1.2.5 Multiokolinearita

Další typickou komplikací při ekonometrické analýze je multikolinearita, která znamená vzájemnou závislost dvou nebo více vysvětlujících proměnných. Čtvrtý Gaussův-Markovův

předpoklad vyžaduje plnou hodnost matice $\boldsymbol{X}$. Nesplnění by znamenalo nemožnost odhadu regresních parametrů pomocí odhadové funkce metody nejmenších čtverců.

S problémem tohoto typu multikolinearity (tzv. perfektní multikolinearity) se lze setkat například při chybné specifikaci modelu s dummy proměnnými. V jiných případech se však perfektní kolinearita (závislost pouze dvojice proměnných) nebo perfektní multikolinearita (závislost tří a více proměnných) v praxi téměř nevyskytuje.

Komplikací ale nemusí být jen perfektní multikolinearita. I silná (byť ne perfektní) závislost mezi proměnnými může mýt vliv na odhadovou funkci. Různé problémy popisuje Hušek [1]. Především se snižuje přesnost odhadnutých parametrů a odhadová funkce je při silné multikolinearitě také citlivá i na malé změny v matici $\boldsymbol{X}$ (například vynechání, nebo přidání pozorování).

Kolinearita a multikolinearita jsou vlastnosti konkrétního výběru, nikoli základního souboru. Proto se hovoří nikoli o testování multikolinearity, nýbrž o jejím zjišťovaní, či měření v konkrétním výběrovém souboru.

Možností měření existuje více, zde bude popsána metoda, kterou využívá software gretl a která tedy bude používána v aplikační části práce. Tato metoda spočívá v použití pomocné regrese jedné vysvětlující proměnné na všech ostatních. Důležitá je potom hodnota koeficientu determinace z této pomocné regrese R_j^2. Pro každou vysvětlující proměnnou se jeho hodnota dosadí do následujícího výrazu:

$$VIF_j = \frac{1}{1 - R_j^2}. \tag{1.21}$$

Jestliže je hodnota VIF_j vyšší než 10 (hodnota dílčího koeficientu determinace vyšší než 0,9), značí to vysokou závislost j-té proměnné na ostatních, což znamená multikolinearitu.

1.3 Dummy proměnné

Dosud byly proměnné uvažovány jako kvantitativní, tedy zachycující velikost, úroveň, počet, množství. To ale není jediná možnost, kterou regresní analýza nabízí. Při aplikaci jsou běžně používány rovněž proměnné kvalitativní. K zachycení kvalitativních znaků se používají tzv. dummy (rovněž binární, nebo bivalentní) proměnné. Ze své povahy jsou to proměnné umělé – uměle vytvořené.

Tyto proměnné mohou nabývat pouze dvou hodnot, a to nuly a jedničky. Takto konstruovaná proměnná tím pádem dělí soubor na dvě skupiny pozorování, z nichž jedna plní určitou podmínku a druhá nikoliv. Jako příklad je možné uvést pohlaví. Proměnná *žena* potom přiřazuje jedničku všem ženám z výběrového souboru a pochopitelně nulu všem mužům.

V praxi je možné se setkat s dummy proměnnými jak v roli regresorů, tak v roli vysvětlované proměnné. V této práci je ovšem vysvětlovaná proměnná kvantitativní, takže druhou možností práce nebude dále zabývat.

Dummy proměnné nabízejí více možností použití. Prvním z nich je takové, kdy dummy proměnná stojí v regresní rovnici samostatně. Pro ilustraci se bude dále pracovat s proměnnou *žena*, která byla popsána výše.

$$y_i = \beta_0 + \beta_1 \text{žena}_i + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i. \tag{1.22}$$

Model (1.22) potom bude odhadovat jiné hodnoty pro muže a pro ženy. Lze říct tedy, že jsou to vlastně dva modely v jednom. Rozdíl je zachycen v úrovňové konstantě. Pro muže má hodnotu β_0 (za proměnnou žena se dosadí nula), zatímco pro ženy je to součet $\beta_0 + \beta_1$. Takto lze zachytit například rozdíl ve velikosti příjmů, nebo ve výšce, váze, věku atd. Tento rozdíl je však prokazatelný pouze tehdy, je-li koeficient β_1 statisticky významně rozdílný od nuly.

Další možností použití je interakce s jinou proměnnou, jak zachycuje rovnice (1.23).

$$y_i = \beta_0 + \beta_1 \text{žena}_i + \beta_2 x_{2i} + \beta_3 \text{žena}_i x_{2i} + \cdots + \beta_k x_{ki} + u_i. \tag{1.23}$$

Pro ilustraci lze uvažovat příklad s výší platu. Tento model jednak zachycuje rozdíl v úrovňové konstantě, tak jako v předchozím příkladu. Pokud by proměnná x_2 zachycovala například počet let strávených prací, potom by model odhadoval navíc rozdílné navýšení platu za každý odpracovaný rok pro obě pohlaví. Muži dostanou v průměru za každý rok strávený v práci plat vyšší o β_2, ženy o $\beta_2 + \beta_3$. Jestli bude navýšení platu vyšší u žen, či mužů, o tom rozhoduje znaménko u koeficientu β_3.

To jsou tedy dvě základní možnosti použití binárních proměnných. Je potřeba si uvědomit, že pomocí nich lze dělit soubor i na více než dvě skupiny. Příkladem může být dosažené vzdělání, nebo například rozdělení na kontinenty, jak je tomu v aplikační části této práce.

1.4 Instrumentální proměnné

Technika instrumentálních proměnných patří mezi mírně pokročilejší aplikace ekonometrického aparátu. Její použití lze popsat následujícím příkladem. Podrobněji se problematikou zabývá například Wooldridge [4].

Často se v ekonometrické praxi lze setkat s takovým problémem, že v modelu chybí nějaká významná, nepozorovaná proměnná. To může mít za následek vychýlení estimátorů. Tento problém lze řešit více způsoby. Buďto je možné spokojit se s vychýlenými odhady, nebo lze najít vhodnou proxy proměnnou, která bude v modelu zastupovat opomenutou proměnnou. Třetí možností je využít metody instrumentálních proměnných.

Zpět k uvažovanému modelu, kde je opomenuta významná proměnná. To bude mít dva následky. Za prvé: jak již bylo řečeno, odhady regresních parametrů budou vychýlené. Za

druhé: hodnoty nezařazené proměnné budou obsaženy v náhodné složce. Metoda instrumentálních proměnných je vhodná právě k odhalení chybějící proměnné.

Instrumentální proměnná musí splňovat základní dvě kritéria. Jednak nesmí být korelovaná s vysvětlovanou proměnnou *y*. To znamená, že při použití metody nejmenších čtverců by její regresní parametr vyšel statisticky nevýznamný. Jinými slovy, instrumentální proměnná nemá žádný přímý vliv na závislou proměnnou. Druhým kritériem je to, aby byla korelovaná (ať už pozitivně, nebo negativně) s některou nezávislou proměnnou *x*, která je v modelu zahrnuta.

Předchozí odstavec se dá jednoduše shrnout takto: Instrumentální proměnná je taková, která má přímý vliv na některou vysvětlující proměnnou, ale nemá vliv na proměnnou vysvětlovanou.

Vychýlení estimátoru u zařazené proměnné *x* je tedy způsobeno tím, že *x* není zcela exogenní. Tento problém lze vyřešit nalezením vhodné instrumentální proměnné, která ovlivňuje hodnoty *x*. Cílem použití techniky instrumentálních proměnných je tedy získání nevychýleného estimátoru.

2 Data a proměnné

První kapitola představila metody, které budou aplikovány v praktické části práce. Následující kapitola 2 představuje data a proměnné, se kterými bude ekonometrická analýza provedena.

V první části je popsáno, z kterých zdrojů byla data převzata. Následuje podkapitola 2.2, která se věnuje specifikům a nastiňuje možné komplikace, které by mohlo přinést použití určitých proměnných, nebo komplikace spojené s určitými pozorováními.

Třetí a čtvrtá část druhé kapitoly navazuje na specifika dat a budou v ní představeny všechny proměnné. Nejprve se v podkapitole 2.3 seznámíme podrobněji s vysvětlovanou proměnnou, tedy s mírou úmrtnosti na silničních komunikacích, v další podkapitole 2.4 budou popsány všechny vysvětlující proměnné a to bez ohledu na to, jestli se nakonec stanou součástí regresního modelu.

V názvu této práce figuruje slovní spojení „vybrané země“. Hned na začátek by mělo být vysvětleno, co tento pojem znamená. Ideální situace by vypadala tak, že by statistický soubor byl kompletní, tedy že by se do modelů podařilo zahrnout všechny země světa. To se pochopitelně nepovedlo. Důvodem jsou chybějící, nebo nevyhovující údaje u některých zemí. I přes snahu o koordinaci metodiky při sběru dat v jednotlivých zemích jednoduše není možné změnit zaběhlé systémy všude na jeden společný tak, aby si údaje odpovídaly. Nebo jiný důvod: Některé země jsou v určitém ohledu v porovnání s ostatními tak odlišné, že jejich zahrnutí do modelu by přineslo více škody než užitku. O jednotlivých problémech se zmiňuje podkapitola 2.2. Výběrový soubor tedy netvoří všechny země světa, ale pouze 136 z nich, které vyhovovaly všem požadavkům.

2.1 Zdroje dat

Původním úmyslem pro téma této bakalářské práce byla analýza nikoli počtu úmrtí, nýbrž počtu všech nehod. Problém nastal ve chvíli vyhledávání vhodných dat. Nelze říct, že by neexistoval nějaký vhodný zdroj k takové analýze, například byla nalezena zpráva o počtu nehod v jednotlivých okresech České republiky (dále jen ČR). Taková analýza by jistě mohla být zajímavá, ale přeci jen bylo prvním očekáváním analyzovat situaci alespoň v Evropě (případně pouze v zemích Evropské unie), nebo ještě lépe v celém světě.

Je ale zřejmé, že dokument obsahující relevantní údaje o počtu nehod by bylo velmi obtížné, ne-li nemožné vyhledat. Ve spoustě zemí (především v méně vyspělých) se podobná statistika možná vůbec neeviduje, a když, tak údaje by nemusely být zrovna nejpřesnější a důvěryhodné. Je snadné si představit, že účastníci nehod se často snaží celou záležitost ututlat, aby nemuseli čelit následkům vyšetřování nehody policií. Dalším důvodem je vůbec definice pojmu nehoda. Lze každý střet dvou automobilů (či jiných dopravních prostředků)

ihned považovat za dopravní nehodu? Co třeba drobné škrábnutí na parkovišti? O tom se často nedozví ani majitel poškozeného vozu, natožpak aby tato událost vstoupila do nějaké statistiky nehod.

Především z těchto důvodů byl návrh o počtu nehod jako vysvětlované proměnné zamítnut a jako nová vysvětlovaná proměnná byla vybrána míra úmrtnosti při dopravních nehodách. Na rozdíl od předešlého návrhu zjevně existuje jasná hranice mezi tím, kdy se nehoda do statistiky započítá a kdy ne. Ve vyspělých zemích navíc existuje určitá jistota přesností údajů, protože počet úmrtí na silnicích je veřejností i odborníky velmi sledovaný údaj. V ČR přinášejí sdělovací prostředky pravidelné informace o počtu smrtelných dopravních nehod, např. za uplynulý víkend, mnohdy s konkrétními detaily, a nechybí ani úhrnné statistiky za jednotlivé měsíce, srovnání letošních prázdnin s loňskými atd. Z pohledu pečlivosti evidence by mohla naše republika být označena za vzornou zemi. V dalších vyspělých zemích světa lze očekávat podobný přístup. Bohužel není možné vyhnout se menší přesnosti údajů u některých zemí, to je ale problém, se kterým se musí vyrovnávat při své práci snad každý analytik.

Téma je tedy jasné – analyzována bude míra úmrtnosti při dopravních nehodách na silničních komunikacích v různých zemích světa. Stěžejní úlohu při výběru dat sehrál dokument Světové zdravotnické organizace (dále jen WHO) s názvem *Global Status Report on Road Safety,* vydaný v roce 2013 [5] a shrnující údaje za rok 2010. Tento dokument od začátku působil velmi spolehlivě a věrohodně. Poměrně detailně analyzuje situaci téměř ve všech zemích světa (13 států se výzkumu nezúčastnilo – Alžírsko, Antigua a Barbuda, Džibutsko, Eritrea, Grenada, Haiti, Libye, Monako, Nauru, Portoriko, Somálsko, Turkmenistán a Tuvalu). V každém státě měl důležitou funkci takzvaný národní koordinátor, který zjišťoval a poskytl data za svou zemi. Věrohodnost údajů tedy do značné míry záleží právě na této osobě a nezbývá než doufat, že národní koordinátoři postupovali při práci korektně, dle vymezené metodiky, a že se vyhnuli chybám nebo nepřesnostem.

Z tohoto dokumentu byla tedy čerpána následující data: počet úmrtí na silnicích, velikost populace, hrubý národní důchod na hlavu, maximální povolená rychlost, a to ve městě a mimo něj, maximální povolený obsah alkoholu v krvi řidiče (pro běžného řidiče, nováčka, nebo řidiče profesionála), dále zda existuje zákonem stanovená povinnost nosit při jízdě na motocyklu helmu, stejně tak zda existuje povinnost být v automobilu připoután (jak vpředu, tak vzadu), a konečně počet registrovaných vozidel.

Další využitý internetový zdroj byly stránky Světové banky [6] a [7]. Odtud pocházejí údaje o ročním úhrnu srážek a míře urbanizace.

Údaje o rozlohách států byly převzaty z databáze The World Factbook americké vládní organizace CIA [8].

Údaje o HDI v roce 2010 pocházejí z dokumentu *Human Development Report 2010* [3].

A konečně posledním internetovým zdrojem byla stránka www.nationmaster.com [9], zde byla zjištěna délka dálniční sítě v jednotlivých zemích.

2.2 Specifika dat a možné problémy

Jak už bylo zmíněno výše, při sběru dat se objevil nejeden problém, či alespoň mírná komplikace, na kterou je potřeba pamatovat při samotném odhadování regresních modelů, protože by mohla nepříznivě ovlivnit odhady parametrů. V následujících odstavcích budou probrány některé údaje z výše uvedeného výčtu a popsány jednotlivé problémy.

2.2.1 Počet úmrtí na silnicích

Tento datový údaj bude později v modelech vysvětlovanou proměnnou (přesněji řečeno bude použita míra úmrtnosti vztažená na 100 000 obyvatel, aby bylo možné srovnávat ji v různě velkých zemích), a proto je potřeba podrobit ho obzvláště pozorné kontrole.

WHO uvádí ve svém dokumentu [5] více údajů. Jednak je to nahlášený počet úmrtí, za druhé bodový odhad, k němuž se přidává ještě interval spolehlivosti (95%). Už při rychlém zhlédnutí dat si čtenář snadno všimne, že bodový odhad a skutečný počet nahlášených úmrtí se v některých případech podstatně liší. Jako příklad může být uvedena hned první země v abecedě, tedy Afghánistán. Nahlášený počet úmrtí činil ve sledovaném roce (2010) 1501 obětí, zatímco bodový odhad byl 6209, což je více než čtyřikrát tolik. Hned na začátku musely být položeny důležité otázky: 1) Který z údajů pro analýzu použít? 2) Nebylo by lepší do modelů zahrnout pouze státy, kde se nahlášený počet neliší od bodového odhadu (například o určité procento)?

Nejprve k druhé položené otázce. Odpověď bude záležet na tom, jakým způsobem se dospělo k bodovým odhadům v jednotlivých zemích. Pro tento účel byly země v dokumentu [5] rozřazeny do čtyř kategorií, v nichž probíhal odhad odlišně, pomocí různých regresních modelů. V každém případě by ale odhad měl především u méně vyspělých zemí odpovídat skutečnému počtu úmrtí více. Podívejme se zpátky na případ Afghánistánu. Rozdíl mezi oběma údaji se zdá propastný, ale důvod je zřejmý. Afghánská evidence úmrtí při autonehodách bude zjevně nekvalitní. Pokud by totiž platil počet nahlášených úmrtí, byl by na tom Afghánistán velmi dobře, figuroval by mezi nejbezpečnějšími zeměmi. To svědčí o tom, že mnohem lepší vypovídací schopnost bude mít bodový odhad, provedený analytiky WHO.

Navíc pokud by z modelu měly být vyřazeny všechny takové země jako v uvedeném příkladě, ubylo by tím ze souboru přesně 78 zemí, a to by znamenalo velkou ztrátu. Proto bude lepší podstoupit riziko spojené s provedeným odhadem a spolehnout se na to, že úředníci WHO pracovali při odhadech korektně.

Tím pádem byla zodpovězena i první položená otázka. Bodový odhad bude použit pro všechny země, a to z důvodu jednotnosti údajů.

2.2.2 Hrubý národní důchod na obyvatele

Tato data převzal dokument WHO [5] ze statistik World Bank a uvádí, že pro zjištění byla použita metodologie Atlas. V ČR se častěji používá pro vyjádření ekonomické vyspělosti země spíš hrubý domácí produkt. Ačkoli metodické vymezení HDP a HND není zcela stejné, interpretace hodnot může být v podstatě totožná (především co se týče pořadí států ve světě a rozdílů mezi nimi).

Bohužel chybějící údaje způsobily, že musely být vyřazeny následující země – Cookovy ostrovy, Korejská lidově demokratická republika, Myanmar a Niue.

2.2.3 Maximální povolená rychlost ve městě a mimo město

Také u maximální rychlosti (ať už ve městě či mimo něj) se objevil problém chybějících dat. Tentokráte ubylo dokonce 14 zemí – Burkina Faso, Burundi, Filipíny, Irák, Jemen, Kamerun, Kapverdy, Libérie, Madagaskar, Salomonovy ostrovy, Salvador, Tanzanie, Tunisko a Vietnam.

Druhá komplikace spočívá v tom, že někdy byla maximální rychlost uvedena jako interval, nikoli jako jedna hodnota. V Evropě jsme zvyklí na to, že maximální povolené rychlosti jsou stanoveny jednotně pro celé území státu, ale to rozhodně nemusí platit i pro jiné části světa. Komplikaci budou představovat především federace, kde jednotlivé země mají pravomoc stanovit maximální rychlosti rozdílně. Vyvstává tedy otázka, jakým způsobem pracovat s daty v situaci, kdy je zadán interval (například 30 – 50 km/h ve městě).

Pro regresní model a metodu nejmenších čtverců je přirozeně potřeba mít pouze jednu hodnotu. Jako reprezentant bude použit aritmetický průměr z krajních hodnot zadaného intervalu. V pozadí stojí předpoklad, že maximální povolené rychlosti v dané zemi jsou rozděleny v tomto intervalu symetricky. Ověřen může být těžko, nicméně nezbývá než doufat, že chyba odhadu pomocí aritmetického průměru nebude příliš významná. Pro výše uvedený interval 30 – 50 km/h tedy do datového souboru vstoupí hodnota 40 km/h.

2.2.4 Maximální povolený obsah alkoholu v krvi řidiče

Tato statistika se v datech objevuje dokonce ve třech variantách: Jednak je to hodnota pro běžné řidiče, za druhé pro začínající řidiče a za třetí pro řidiče profesionály. Údaje jsou na naše poměry uvedeny v celkem nezvyklých jednotkách, a to gram na decilitr (gram alkoholu na decilitr krve). V ČR se běžně používají jednotky promile, což představuje gram na litr. Při ekonometrické analýze bude potřeba pamatovat na tuto odlišnost, nebo později zavést tyto tři proměnné i v jednotkách gram na litr.

Chybějící údaje vyřadily z modelu následujících 19 zemí: Barbados, Bolívii, Gambii, Guatemalu, Guineu-Bissau, Indonésii, Kazachstán, Keňu, Kyrgyzstán, Marshallovy ostrovy, Nepál, Niger, Papuu Novou Guineu, Paraguay, Senegal, Sieru Leone, Svatého Tomáše, Togo a Vanuatu.

Další zajímavost týkající se těchto dat byl způsob, jakým vyhodnotit situaci v islámských zemích, kde je pití alkoholu všeobecně zakázané. Zde totiž údaje rovněž chyběly, ale dotčené země nakonec nebyly vyřazeny. Místo toho byly zadány hodnoty 0, tedy nulová tolerance (jako je například v ČR). Jestliže je požívání alkoholu zakázáno úplně, vyplývá z toho, že nesmí být konzumován ani před jízdou. Otázka je, jak vypadá situace ve skutečnosti.

2.2.5 Povinnost nosit helmu při jízdě na motocyklu

Tato povinnost existuje ve všech zkoumaných zemích kromě dvou výjimek. Helma není povinná v Afghánistánu a na Kiribati. Je tedy otázka, do jaké míry bude tento údaj vhodný jako vysvětlující proměnná, ale tomu se bude věnovat text později.

2.2.6 Počet registrovaných vozidel

Vyřazeny byly další 3 země (opět pro ně nebyly uvedeny údaje). Byli jimi Belize, Lesotho a Uzbekistán.

Další problém způsobilo nečekaně Finsko. Ve Finsku se totiž zvlášť registrují přívěsy a karavany a uvedený údaj tedy obsahuje jak automobily, tak tato přívěsná zařízení. Bylo poměrně překvapující, že k vyhledání nebyl údaj pouze o počtu registrovaných vozidel. Opět nezbývá než doufat, že tato nepřesnost nebude příliš zkreslovat výsledky odhadů regresních parametrů. V nejhorším případě bude později nutné vyřadit z datového souboru i Finsko, prozatím bude ale v datech ponecháno.

2.2.7 HDI

HDI, neboli index lidského rozvoje, byl čerpán přímo z dokumentu Human Development Report 2010 [3]. O přesnosti údajů tedy nemusí být pochybnost. HDI může být považován za poměrně zajímavý protipól k hrubému národnímu důchodu. Zatímco HND vypovídá o vyspělosti ekonomiky z hlediska produkce, HDI v sobě implicitně obsahuje spíše informaci o tom, jakým způsobem jsou použity veřejné finance. U zemí s vysokým HDI lze očekávat například obecně lepší úroveň infrastruktury, zdravotnictví atd., což rozhodně může mít vliv na úmrtnost na silnicích. Při ekonometrické analýze budou vyzkoušeny obě dvě proměnné a prokáže se, která bude pro model vhodnější. Obě najednou do modelu pravděpodobně nevstoupí, a to z důvodu silné korelace.

HDI se bohužel vůbec nesleduje v San Marinu, proto musí být i tato země vypuštěna z datového souboru.

2.2.8 Délka dálniční sítě

Tato statistika byla převzata z internetových stránek www.nationmaster.com [9]. Dálnice je podle tohoto zdroje definovaná jako komunikace, která je ve střední části oddělena fyzickou bariérou tak, aby se zamezilo střetu vozidel jedoucích opačnými směry.

Opět můžeme jen zopakovat, že v různých zemích existují také různé typy dálnic, a proto vyvstává otázka, do jaké míry se tato statistika ukáže být vhodným regresorem. Dalším problémem je to, že délku dálniční sítě bude potřeba pro srovnatelnost mezi státy vztáhnout k jiné proměnné, například k rozloze, nebo počtu obyvatel. Třetí komplikací bude jistě heteroskedasticita, v důsledku toho, že velký počet zemí žádnou dálnici nemá. Mezi nimi se jistě najdou pozorování s vysokou i nízkou mírou úmrtnosti.

2.3 Vysvětlovaná proměnná

Vysvětlovanou proměnnou bude v modelu počet úmrtí při dopravních nehodách připadajících na 100 000 obyvatel dané země. Jedná se tedy o míru úmrtnosti. Bez vztažení na počet obyvatel by celá analýza postrádala téměř smysl, protože je jasné, že čím větší je populace dané země, tím více lidí také při nehodách zemře. Tím není řečeno, že by nebylo možné nelézt dvě země, které by se například trojnásobně lišily co do velikosti populace a zároveň by vykazovaly stejný počet úmrtí v absolutních číslech. Takových dvojic existuje jistě i více, ale přece jen obecně bude spíše platit první tvrzení, a racionální přístup nedovolí srovnávat absolutní počty zemřelých pro státy, které mají desítky či stovky milionů obyvatel s těmi, které jich mají desítky či stovky tisíc.

Možná, že zajímavější otázka ale bude, proč bylo vybráno vztažení zrovna ku 100 000 obyvatel? Stalo se tak kvůli snadnější interpretaci výsledků. Počet úmrtí na 100 000 obyvatel se totiž pohybuje přibližně mezi 3 – 40 jednotkami. To lze považovat za rozumné měřítko, při kterém si lze později snadno představit jednotlivé vlivy nezávislých proměnných. Například lze uvažovat nějaké podobně velké české město (nejblíže sto tisícům má Liberec nebo Olomouc). V případě úmrtnosti 30 lidí na 100 000 obyvatel ročně by potom představa byla taková, že při stejné úmrtnosti by při nehodách zemřelo 30 obyvatel Liberce, či Olomouce za rok.

Pro vytvoření představy o hodnotách pro různé země či regiony slouží následující podkapitola.

2.3.1 Některé hodnoty vysvětlované proměnné

Následující statistiky se týkají pouze výběru 136 vybraných zemí světa, je tedy možné, že vyřazené země zahrnují ještě extrémnější hodnoty, nebo by se při jejich nevyřazení změnily statistiky jako průměr, medián apod. Pro potřeby analýzy v této práci ale bude podstatnější seznámit se s hodnotami za tento výběr.

Nejprve z toho pozitivního konce, tedy od zemí, kde byla sledovaná míra úmrtnosti v roce 2010 nejnižší. Vůbec nejlépe jsou na tom Maledivy se skvělým výsledkem 1,899 (zemřelo 6 lidí, populaci tvoří 315 885 obyvatel). Na druhém místě Island – 2,811 (9 úmrtí na 320 136 obyv.) a na třetím místě Švédsko – 2,964 (278 na 9 379 687). Další v pořadí v první

desítce jsou Západní břeh Jordánu a Pásmo Gazy, Andorra, Spojené království, Malta, Nizozemí, Norsko a Švýcarsko.

Při zběžném prohlédnutí této desítky je možné si všimnout, že nejlépe jsou na tom buďto vyspělé evropské země, nebo ostrovní státy. U ostrovních zemí obecně zřejmě převládá klidnější styl života a navíc jsou k dopravě často do značné míry užívány vodní plochy.

Česká republika obsadila pěkné 32. místo. V roce 2010 zde bylo evidováno 802 obětí dopravních nehod, přičemž počet obyvatel činil 10 492 960, což znamená úmrtnost 7,643. Tato práce není zaměřená na hodnocení dopravní situace, ať už ve světě či u nás, ale přesto následuje krátký komentář k tomuto výsledku, který může být považován za poměrně pozitivní. V žádném případě není situace výrazně horší než v podobných evropských zemích, a naopak některé vyspělejší naše republika dokonce překonala (jako Belgii, Nový Zéland, Spojené státy americké).

Následuje pohled na druhý konec výsledkové listiny. Nejhůře je na tom Dominikánská republika s úmrtností 41,733, za ní Thajsko – 38,066 a Venezuela – 37,236. Následují další země první (nebo spíše poslední) desítky. Írán, Nigérie, JAR, Omán, Uganda, Mauretánie, Guyana. S výjimkou Venezuely a Ománu jsou to všechno země s hrubým národním důchodem pod 10 000 USD na hlavu.

Následující tabulka uvádí vybrané statistiky vysvětlované proměnné za celý datový soubor, tedy za všech 136 zemí.

Aritmetický průměr	14,997
Vážený průměr (dle velikosti populace)	17,979
Medián	14,063
Minimum	1,899
Maximum	41,733
Směrodatná odchylka	8,1403
Variační koeficient	0,54280

Tabulka 2-1, Vybrané statistiky vysvětlované proměnné (zdroj – gretl a vlastní výpočty)

Vážený průměr je vyšší v porovnání s mediánem a průměrem aritmetickým. To ukazuje na to, že vyšší úmrtnost bude obecně ve větších zemích. Vysvětlením by opět mohly být malé ostrovní země, které snižují hodnotu aritmetického průměru i mediánu.

Zajímavé také bude podívat se na hodnoty vysvětlované proměnné podle regionů. Ty jsou uvedeny v následující tabulce 3-2. Jako porovnávací statistika byl zvolen vážený aritmetický průměr. Hodnota v závorce je počet zemí, které spadají do daného regionu.

Rozčlenění do regionů v podstatě odpovídá světadílům, jedině v Americe si zaslouží členění o něco vetší pozornost, jak bude uvedeno dále. A Rusko, které se nachází na dvou světadílech, bylo zařazeno do Evropy, vzhledem k tomu, že kulturně odpovídá spíše tomuto regionu a také hlavní město Moskva a další velká centra se nacházejí v evropské části země.

Evropa (41)	9,633
Severní Amerika (2)	10,972
Střední Amerika (12)	16,136
Jižní Amerika (10)	20,920
Afrika (29)	23,603
Asie (36)	19,096
Austrálie a Oceánie (6)	6,660

Tabulka 2-2, Vážený průměr míry úmrtnosti dle regionů (zdroj – vlastní výpočty)

Výsledky nejsou překvapivé. Nejnižší míry úmrtnosti jsou v Austrálii a Oceánii a v Evropě. Zajímavé je sledovat, jak se úmrtnost na americkém kontinentu zvyšuje při postupu od severu k jihu. Do Severní Ameriky byly zařazeny pouze Spojené státy a Kanada. Mexiko podle charakteru a kultury spadá již do Střední Ameriky. Na severu se tedy podle tohoto členění nacházejí pouze dvě vyspělé země. Ve Střední Americe už je úmrtnost vyšší, ale v tomto regionu se nachází také hodně ostrovních zemí s klidnějším stylem života, jak už bylo zmíněno výše. A konečně na jihu Ameriky registrujeme středně vyspělé země, ve kterých se často můžeme setkat s hektickým stylem života a jihoamerickým temperamentem, ke kterému patří jistá agresivita a netrpělivost. Zde tedy míra úmrtnosti dosahuje nejvyšších hodnot.

Po krátkém představení vysvětlované proměnné a ukázce několika statistik následuje podkapitola popisující proměnné vysvětlující.

2.4 Vysvětlující proměnné a jejich očekávaný vliv

Ze získaných dat bylo tedy vytvořeno celkem 23 nezávislých proměnných. Nyní je možno je všechny představit a především upozornit na velmi důležitý údaj, a to měrné jednotky, které je potřeba znát pro účely interpretace výsledků. Zároveň, jak již nadpis poukazuje, budou vysloveny hypotézy o očekávaném vlivu na proměnnou závislou. V závěru podkapitoly se nachází tabulka, ve které budou uvedeny zkratky, kterými jsou nezávislé proměnné v modelech označeny.

2.4.1 Hustota zalidnění

Hustota zalidnění se měří v počtu obyvatel na kilometr čtvereční a výjimkou nebude ani tato proměnná. Hodnoty byly získány výpočtem z údajů o počtu obyvatel a rozloze. Jaký vliv lze očekávat?

Určitý význam by tato proměnná mít mohla. V řídce osídlených zemích totiž logicky bude i řidší provoz, a tak nastane méně nehod dvou a více vozidel. Pokud dojde k nějakému úmrtí, spíše bude způsobeno rychlou jízdou jednoho vozidla. Naopak v hustě osídlených oblastech se dá čekat obecně více srážek s chodci a také vyšší četnost nehod, kterých se účastní alespoň dva automobily.

Na druhou stranu ve městech mají nehody často charakter spíše srážek v nižší rychlosti, což z městské dopravy naopak činí relativně bezpečnou. Navíc mnoho lidí používá k přepravě ve městech hromadnou dopravu, která bude opět působit protichůdně k prvně zmíněné hypotéze.

V tuto chvíli tedy není možné určit jednoznačný závěr o vlivu hustoty zalidnění, a dokonce ani hypotézu, zda se bude jednat o pozitivní či negativní závislost.

Problém při použití hustoty zalidnění jako nezávislé proměnné budou navíc představovat odlehlá pozorování v podobě městských nebo hustě osídlených malých států. Nadpoloviční většina zemí z datového souboru má hustotu zalidnění do 100 obyv./km^2, zbytek až na pět výjimek se vejde do 700 obyv./km^2. Těchto pět výjimek se ale značně odlišuje. U čtyř se hodnota proměnné pohybuje mezi 1000 až 2000 obyv./km^2, absolutním rekordmanem je ale Singapur s hustotou zalidnění 7340 obyv./km^2! Zařazení tohoto odlehlého pozorování by mohlo mít za následek vychýlení regresních parametrů. Ke zvážení bude při použití této proměnné vyřazení Singapuru, nebo dokonce i zbylé čtveřice odlehlých pozorování.

K tomu se přidává další komplikace, a tou je fakt, že hustota zalidnění nemusí mít úplně dobrou vypovídací schopnost z toho důvodu, že v ní není zahrnuta informace o tom, jakým způsobem je dané území osídleno. Pokud se totiž na území daného státu nachází velká vodní plocha, nebo neprostupná území, jako hory, husté lesy, či pouště, hustota zalidnění bude poměrně nízká. Pokud by se ale odečetly tyto plochy, výsledný údaj by byl směrodatnější.

Jak je vidět, komplikací s touto proměnnou je mnoho. Proto se jeví praktičtější použít místo ní proměnnou následující – tedy míru urbanizace.

2.4.2 Míra urbanizace

Tato proměnná je svým způsobem jistý protipól k předchozí hustotě zalidnění. Její výhodou je, že nebude obsahovat odlehlá pozorování, a to z toho důvodu, že se měří v procentech (hodnoty jsou zaokrouhleny na celá procenta).

Na rozdíl od hustoty zalidnění má větší vypovídací schopnost o charakteru zalidnění, protože odpovídá tomu, jaké procento obyvatel žije ve městech. Míra urbanizace se sice

nevypočítává ve všech zemích dle stejné metodiky, ale to by vlastně nemuselo až tak vadit. Možná je to naopak prospěšné, protože v různých částech světa nemusí pojem město znamenat totéž. U nás se považuje za hranici mezi vesnicí a městem přibližně 3000 obyvatel (kritérií je pochopitelně více), ale jinde ve světě jistě existuje určité místo, kde bydlí mnohem více lidí pohromadě, ale charakter tohoto sídla nebude odpovídat městu.

Co se týče vlivu urbanizace na míru úmrtnosti, může být jen zopakováno to, co bylo řečeno u hustoty zalidnění. Ve městech se jezdí zpravidla pomaleji, značná část obyvatelstva používá k přepravě městskou hromadnou dopravu, která je relativně bezpečná, a případné srážky automobilů většinou končí odřenou karoserií či promáčknutými plechy. Naopak větší nebezpečí může hrozit chodcům a cyklistům (především v méně kultivovaných částech světa, kde je dopravní situace ve městech chaotická).

I přes poslední zmíněný fakt je očekávání takové, že vyšší míra urbanizace povede k nižšímu počtu úmrtí na silnicích.

2.4.3 Hrubý národní důchod

Hrubý národní důchod jistě bude mít vliv na vysvětlovanou proměnnou, ale na druhou stranu musíme přiznat, že je to velmi obecný údaj. Tato závislost potom může být zhodnocena tak, že ve vyspělejších zemích umírá méně lidí na silničních komunikacích. Co z toho ale vyplývá pro chudší regiony? K vyřešení problému stačí zbohatnout? Ale jakým způsobem to dokázat a co konkrétně způsobuje vysokou míru úmrtnosti, to zůstává nezodpovězeno. Za hrubým domácím důchodem se jednoduše skrývá značné množství dalších vlivů a faktorů, které není vždy snadné odlišit.

Navíc lze očekávat heteroskedasticitu. Mezi chudšími státy bude existovat určitě větší rozptyl, než mezi vyspělými. Mezi státy s nízkým HND se totiž vyskytují jak malé ostrovní státy, které mají často velmi nízkou úmrtnost, tak vnitrozemské státy s podobným důchodem, ale rozvinutější dopravní sítí. Zde bude úmrtnost naopak vysoká.

Měrné jednotky HND jsou dolary na obyvatele. Koeficient beta tedy bude vycházet zřejmě velmi blízký nule, ale nemusí to vůbec znamenat statistickou nevýznamnost. Při interpretaci se bude komentovat výsledek tak, jak by se změnila úmrtnost, kdyby se HND zvýšilo třeba o 1000 USD/obyv., nikoli pouze o jeden dolar. Koeficient tedy bude pro interpretaci užitečné vynásobit tisícem.

2.4.4 HDI

Podobně jako předchozí proměnná je Human Development Index (HDI), neboli Index lidského rozvoje, velmi obecným ukazatelem. Přece jen by ale mohl být o něco relevantnější. Existují totiž státy, které mají poměrně vysokou produktivitu práce (například díky bohatství nerostných surovin atd.), ale z tohoto bohatství těží pouze omezená skupina osob. Ta výrazně ovlivňuje výši HDP (potažmo HND), ale mezi běžným obyvatelstvem může panovat chudoba.

Jednoduše řečeno, HND nevypovídá nic o tom, jakým způsobem je produkt spotřebován, jakým způsobem jsou použity veřejné finance, jak se investuje do školství, zdravotnictví, do infrastruktury atd. Naproti tomu HDI tyto informace částečně zachycuje. Vypočítává se jako geometrický průměr tří dílčích indexů, a to indexu očekávané délky života, indexu vzdělání a indexu hrubého národního produktu. Především první dva dílčí indexy v sobě zachycují výše zmíněné informace.

U států s vyšším HDI lze očekávat tedy nižší úmrtnost, vycházející z předpokladu lépe fungujících veřejných služeb, kvalitnějších komunikací, vyššího vzdělání obyvatel.

Index lidského rozvoje je bezrozměrné číslo v intervalu od 0 do 1. Vyplývá to z toho, že dílčí indexy jsou také čísla v tomto intervalu, geometrický průměr tedy nemůže tento interval opustit.

2.4.5 Maximální povolené rychlosti

Rychlost vozidla před srážkou je jednoznačně podstatná pro to, zda nehoda vůbec nastane, nebo ne. Zároveň také spoluurčuje, jestli dojde pouze ke zranění zúčastněných osob, nebo k jejich úmrtí.

Tento předpoklad nemusí být nijak dokazován, ale o něco obtížnější už bude prokázat, že stejnou roli hraje také maximální povolená rychlost, stanovená legislativou dané země. Problém spočívá v tom, zda je tato rychlost také dodržovaná řidiči a případně vynucovaná státem (formou pokut, odebrání řidičského oprávnění a dalšími sankcemi).

Na výběr bude dokonce ze dvou proměnných, a to maximální rychlost ve městě a mimo město. Obě jsou vedeny ve stejných jednotkách, a to v kilometrech za hodinu. Očekávaný vliv lze tedy formulovat tak, že v zemích s nižšími povolenými rychlostmi bude i nižší míra úmrtnosti na silnicích.

2.4.6 Maximální povolené množství obsahu alkoholu v krvi

Obsah alkoholu v krvi řidiče je často první věc, která lidi zajímá, když se dozví o nějaké nehodě (ať už má fatální následky nebo ne). V ČR je navíc situace zajímavější tím, že je požívání alkoholu před jízdou zcela zakázáno, a tak se často vedou debaty o tom, jestli nulová tolerance není příliš přísná a podobně. Tato bakalářská práce by mohla přinést na podobné otázky odpovědi.

Je až překvapující, v kolika zemích je alkohol před jízdou povolen, a možná ještě překvapivější, v jakém množství! Nejvyšší povolená hranice je 0,08 g/dl (v datovém souboru jsou použity právě tyto jednotky), což představuje 0,8 promile alkoholu v krvi! To znamená, že řidič si může dát před jízdou dvě až tři piva, a přesto mu nic nebrání vyjet (záleží jistě na dalších faktorech, především na celkovém čase konzumace, váze řidiče a mnoha dalších). Pro nás Čechy je tento fakt poměrně zarážející.

Stejně jako u předchozí proměnné i pro povolené množství alkoholu bude k dispozici při specifikaci modelů více variant, tentokrát dokonce tři. Jsou jimi hodnoty pro běžnou populaci, pro nováčky (vymezení dle každé země jinak) a pro profesionální řidiče (řidiče MHD, kamionů, autobusů, taxikáře atd.).

U všech variant lze očekávat pozitivní vztah, tedy čím vyšší povolená hranice obsahu alkoholu v krvi, tím vyšší bude úmrtnost. Jestli se ale podaří očekávanou závislost potvrdit, to se zjistí až později při tvorbě modelů v aplikační části práce.

2.4.7 Povinnost nošení helmy na motocyklu

Tato proměnná je proměnnou binární, nabývá pouze hodnoty 1 v případě, že je v dané zemi helma pro motocyklisty povinná, a hodnoty 0 v opačném případě. Počet zemí, kde je povoleno jezdit bez helmy, bude možná překvapivý, existují totiž pouze dvě. Afghánistán a Kiribati (alespoň z vybraných 136 zemí).

Tento nízký počet může zapříčinit, že výsledky analýzy nebudou přesné, neboli že bude estimátor vychýlený. Obecně lze očekávat, že nošení helmy zvýší bezpečnost, a tedy i sníží úmrtnost. Situace odpovídá tomuto tvrzení v Afghánistánu, kde je úmrtnost poměrně vysoká, ale druhé pozorování – Kiribati - má naopak úmrtnost velmi nízkou. Při zařazení této proměnné do modelu koeficient beta vyjde s opačným znaménkem, než dle očekávané hypotézy.

Proto tato proměnná do modelu určitě nevstoupí samotná, možná že bude možné ji použít v interakci s jinou. Ale zřejmě ani tento přístup nemusí projít ekonomickou verifikací.

2.4.8 Povinnost použití pásů v automobilu na všech místech

I tato proměnná je binární, a přiřazení hodnot odpovídá předchozí proměnné. Jednička tedy bude znamenat povinnost být při jízdě automobilem připoután (a to na všech místech), nula znamená, že tato povinnost neexistuje.

Očekávaný vliv je samozřejmě snížení úmrtnosti při povinnosti být připoután, ale jak bylo možné se přesvědčit u předchozí proměnné, analýzu může zkomplikovat řada dalších faktorů. Jak to tedy vypadá s počtem zemí bez zavedení této povinnosti? Z tohoto hlediska jich je ve výběru více, přesně 40. To poskytuje větší šanci na přesnost regresního koeficientu beta pro tuto proměnnou.

2.4.9 Počet registrovaných vozidel

Dalším regresorem je počet registrovaných vozidel, přesněji řečeno je to počet registrovaných vozidel na 1000 obyvatel dané země, opět ze zřejmého důvodu, a to aby bylo možno srovnávat země mezi sebou. Hodnoty této proměnné se pohybují od jednotek (většinou chudé africké země) až po stovky, přičemž rekordman Malta má dokonce 1022 vozidel na 1000 obyvatel. Na druhém místě je Finsko, o kterém byla zmínka v kapitole o datech. Ve Finsku se totiž zvlášť registrují přívěsná zařízení, a to „míru automobilizace“

uměle navyšuje. Na druhou stranu Finsko nevybočuje extrémním způsobem, takže budeme doufat, že toto pozorování nevychýlí odhady.

První hypotéza o očekávaném vlivu této proměnné na míru úmrtnosti by logicky vypadala tak, že čím více lidí bude jezdit osobními dopravními prostředky, tím více úmrtí na silnicích nastane. Pokud naopak připadá několik jednotek vozidel na 1000 obyvatel, ostatní budou zřejmě k dopravě používat bicykly, vlastní nohy, zvířata atd. Ale při pohledu na datový soubor si lze uvědomit, že vysoký počet registrovaných vozidel je běžný ve vyspělých zemích, kde je zároveň nízká úmrtnost, a naopak nejnižší v již zmíněných afrických zemích, kde je úmrtnost zase vysoká.

To vůbec neodpovídá původní hypotéze. Že by celá situace nakonec dopadla úplně jinak? Možná, že naopak vysoký počet aut na obyvatele svědčí jednak o vyspělosti země, což může být spojeno například s kvalitou a bezpečností komunikací, jednak o řidičské gramotnosti obyvatel, takže nedochází k nehodám způsobeným řidiči laiky. Zrovna tak v zemích s velkým množstvím automobilů a motocyklů bude spíše fungovat kvalitnější systém výuky a testování začínajících řidičů. V Evropě jsou občané v autoškolách podrobeni výuce v řádech několika desítek hodin a zároveň se učí teorii. Všechny dovednosti potom prokazují u závěrečných zkoušek. Takový systém si lze těžko představit v rozvojových zemích.

To všechno svědčí o tom, že výsledný vliv bude zřejmě úplně opačný, než dle původní hypotézy.

Jiný problém při použití tohoto regresoru však může spočívat ve vysoké korelaci s HND a HDI. Dále by se mohla objevit i heteroskedasticita, protože mezi zeměmi s nízkým počtem registrovaných vozidel se vyskytuje jak vysoká, tak nízká úmrtnost (již poněkolikáté lze poukázat na ostrovní země, například z Oceánie, nebo některé středoamerické ostrovy, a na druhé straně na rozvojové země pevninské).

2.4.10 Roční úhrn srážek

Dopravní nehody jsou často ovlivněny také počasím. Nepřizpůsobení stylu jízdy především u nezkušených řidičů často vede k nehodě, ačkoli nemusí zároveň nutně dojít k porušení pravidel silničního provozu nebo k předchozímu požití alkoholu.

Mezi nejčastější faktory patří déšť, sníh, silný vítr, námraza. Všechny není možné kvantifikovat, ale bude nutné spokojit se alespoň s prvními dvěma údaji. Roční úhrn srážek vypovídá jak o dešťových, tak o sněhových srážkách, a je proto ideální statistikou.

Měří se v milimetrech vodního sloupce, který by vytvořily srážky kapalného původu. Sníh má totiž větší objem než déšť, pro jednotnost údajů se tedy nedá použít pouze velikost sloupce. S tím ale nebude nutné se zatěžovat, tento technický problém za nás řeší meteorologové, kteří vědí, jakým způsobem velikost vodního sloupce určit.

Pro představu možno také uvést, že jeden milimetr srážkové vody by na ploše jednoho metru čtverečního znamenal objem přesně jednoho litru. Údaj v milimetrech si tedy lze představit jako počet litrů vody, které spadnou na jeden metr čtvereční.

Predikce o vlivu tohoto regresoru opět nebudou překvapivé. Více srážek by mělo znamenat větší pravděpodobnost nehody, tím pádem větší pravděpodobnost úmrtí.

2.4.11 Regiony

Ačkoli má tato podkapitola název regiony, nejedná se o nějakou zvláštní proměnnou. Ve skutečnosti jde pouze o 7 binárních proměnných, které zaznamenávají regionální příslušnost dané země. O tomto tématu se již zmiňuje předešlá kapitola, takže stačí připomenout, že regiony jsou Evropa, Severní Amerika, Střední Amerika, Jižní Amerika, Afrika, Asie a Austrálie a Oceánie. Rusko bylo zařazeno mezi evropské země, ačkoli většinou své rozlohy se nachází na kontinentu asijském.

Zařazení těchto binárních proměnných může být poměrně užitečné. Jednak umožní analyzovat situaci pro jednotlivé kontinenty odděleně, ale i při použití celého datového souboru se mohou některé regionální proměnné aplikovat zejména v interakci s jinou proměnnou. Mezi jednotlivými kontinenty totiž jistě budou existovat rozdíly, ať již ve výši úmrtnosti, nebo v síle vlivu jiných proměnných. Interakční proměnné pomohou odhalit v pořadí druhé jmenované rozdíly.

2.4.12 Hustota dálniční sítě

Na míru úmrtnosti má jistě vliv i kvalita silničních komunikací. To je další těžko kvantifikovatelný údaj. I kdyby se podařilo nějak kvalitu zjistit, nemuselo by to nic znamenat, protože ve všech zemích nejsou stejné další podmínky (počasí atd.). To by znamenalo, že v některých zemích stačí pro stejnou bezpečnost silničního provozu použít pro stavbu komunikací například méně kvalitní materiály. Naopak v zemích s náročnějšími podmínkami bude větší důraz na kvalitu, tak aby materiál odolal nepříznivým vlivům podnebí.

Proto byl vybrán alespoň jeden reprezentant, který v sobě výše popsané odráží. Je to počet kilometrů dálnic. Dálnice jsou totiž poměrně bezpečné vzhledem k tomu, kolik osob a v jaké rychlosti obslouží. Samozřejmě pokud se stane ve vysoké rychlosti nehoda, může mít velmi rychle fatální následky, ale alespoň je na dálnici zamezeno střetu protijedoucích vozidel díky bariérám ve střední části.

Opět by asi nemělo smysl použít tuto nezávislou proměnnou v absolutních číslech, tedy v hodnotách udaných v kilometrech komunikace. Bude nutné tedy tato čísla vztáhnout k jinému údaji. Nabízí se hned několik možností.

Za prvé lze délku dálniční sítě poměřovat vzhledem k rozloze země. Potom se mluví o tzv. hustotě dálniční sítě. Jednotkami této proměnné jsou kilometry dálnice na 1000km^2 plochy. (taková jednotka by potom vlastně byla 1/1000km). Tato varianta vztažení asi

napadne každého jako první, protože o hustotě dálniční (nebo třeba železniční) sítě se běžně diskutuje. Ale nezajímavé by nemusely být ani další možnosti.

Tou druhou je vztáhnout kilometry dálnic k počtu obyvatel. Lze provést malou, skoro až filozofickou úvahu. Dálnice neslouží ploše dané země, nýbrž lidem. Jestliže bude země velmi hustě osídlená, potom se relativně vysoká hustota dálniční sítě může zdát dostatečná v porovnání s jinými zeměmi, ale ve skutečnosti může být tato síť příliš řídká na obsluhu všech obyvatel. Délku dálnice v kilometrech potom vztahujeme na 100000 obyvatel.

Zajde-li se v předchozí úvaze ještě o kus dále, může vyvstat argument o tom, že autem se dá cestovat s různou efektivitou. Bohužel i v Praze je často možné vidět řidiče, jak osamoceně míří za volantem do práce. V ideálním případě by ale mohli 3 ze 4 takových řidičů nechat svoje vozy doma a cestovat všichni společně. (Při takovém způsobu dojíždění se samozřejmě vyskytne mnoho jiných problémů, příklad tedy slouží pouze pro představu). Předchozí příklad lze aplikovat na danou problematiku a dojít k názoru, že pokud dálnice souží lidem, pak je ale také důležité, jestli jezdí do práce (a jinam) jednotlivě či ve větší skupině. V druhém případě budou dálnice prázdnější. Touto úvahou může být pozměněna předešlá myšlenka, a tedy lze tvrdit, že dálnice neslouží lidem, ale dopravním prostředkům.

Potom můžeme délku dálniční sítě vztáhnout nikoli na počet obyvatel, nýbrž na počet registrovaných vozidel (to sice neodpovídá zcela přesně předešlé úvaze, určité množství registrovaných aut nemusí být k přepravě vůbec používáno). Nicméně vyzkoušet jako vysvětlující proměnnou lze i tuto třetí variantu vztažení délky dálniční sítě.

Jednotkami potom budou kilometry dálnice na 100 000 registrovaných vozidel.

2.4.13 Doplňující informace a některé statistiky vysvětlujících proměnných

Podkapitola 2.4 se zabývala popisem vysvětlujících proměnných, se kterými bude prováděna ekonometrická analýza. Při tvorbě modelů ale bude zřejmě praktické, aby jednotlivé regresory měly krátké a zjednodušené názvy tak, aby rovnice byly přehlednější. Zkratky všech proměnných jsou uvedeny v tabulce 2-3.

V předchozích odstavcích byly probírány vlastnosti nezávislých proměnných. Většinou se ale jednalo spíše o kvalitativní vlastnosti. Proto následující tabulka 2-4 obsahuje navíc některé statistiky, tedy znaky kvantitativní. Konkrétně se jedná o minimum a maximum, aritmetický průměr, vážený aritmetický průměr (vahami je velikost populace dané země), medián a směrodatnou odchylku. Všechny jednotky odpovídají výše uvedeným. V této tabulce nebudou uvedeny binární proměnné popisující region.

Proměnná:	Zkratka:
Hustota zalidnění	*hust_zal*
Míra urbanizace	*urban*
Hrubý národní důchod	*gni*
Human development index	*hdi*
Maximální povolená rychlost ve městě	*rychlost_mesto*
Maximální povolená rychlost mimo město	*rychlost_mimo*
Maximální povolený obsah alkoholu v krvi řidiče - běžní řidiči	*alk_norm*
Maximální povolený obsah alkoholu v krvi řidiče - nováčci	*alk_nov*
Maximální povolený obsah alkoholu v krvi řidiče – profesionální řidiči	*alk_prof*
Povinnost nosit helmu při jízdě na motocyklu	*helma*
Povinnost být připoután pásy na všech místech v automobilu	*pasy*
Počet registrovaných vozidel	*reg_voz*
Roční úhrn srážek	*srazky*
Evropa	*evropa*
Severní Amerika	*sev_amer*
Střední Amerika	*stred_amer*
Jižní Amerika	*jiz_amer*
Afrika	*afrika*
Asie	*asie*
Austrálie a Oceánie	*australie*
Délka dálniční sítě na počet obyvatel	*daln_obyv*
Délka dálniční sítě na rozlohu	*daln_rozl*
Délka dálniční sítě na počet registrovaných vozidel	*daln_reg_voz*

Tabulka 2-3, Zkratky názvů proměnných

Proměnná	Maximum	Minimum	Aritmetický průměr	Vážený průměr	Medián	Směrodatná odchylka
hust_zal	7339,71 (Singapur)	1,76 (Mongolsko)	196,96	198,24	77,99	661,32
urban	100 (Singapur)	9 (Trinidad a Tobago)	60,06	52,89	63,00	22,86
gni	86390 (Norsko)	180 (Dem. republika Kongo)	14870,51	10148,63	6530	18395,33
hdi	0,938 (Norsko)	0,140 (Zimbabwe)	0,68	0,64	0,71	0,17
rychlost_mesto	100 (Omán)	20 (Rovníková Guiena)	53,54	53,44	50	13,34
rychlost_mimo	120 (Kolumbie, Mozambik)	25 (Svatá Lucie)	74,80	65,24	80	19,08
alk_norm	0,08 (38 zemí)	0,00 (18 zemí)	0,047	0,033	0,05	0,027
alk_nov	0,08 (35 zemí)	0,00 (30 zemí)	0,040	0,027	0,05	0,03
alk_prof	0,08 (36 zemí)	0,00 (34 zemí)	0,038	0,026	0,037	0,031
helma	1 (134 zemí)	0 (Kiribati a Afghánistán)	0,985	0,995	1	0,12
pasy	1 (96 zemí)	0 (40 zemí)	0,706	0,815	1	0,456
reg_voz	1021,61 (Malta)	1,09 (Středoafr. republika)	305,79	240,50	212,15	266,67
srazky	3240 (Kolumbie)	51 (Egypt)	1120,59	959,91	972	740,24
daln_obyv	85,68 (Bhútán)	0 (58 zemí)	6,36	5,65	1,11	11,06
daln_rozl	232,32 (Singapur)	0 (58 zemí)	8,42	6,32	0,42	22,90

Proměnná	Maximum	Minimum	Aritmetický průměr	Vážený průměr	Medián	Směrodatná odchylka
daln_reg_voz	1079,52 (Bhútán)	0 (58 zemí)	22,37	20,17	6,34	93,10

Tabulka 2-4, Některé statistiky vysvětlujících proměnných (zdroj - gretl a vlastní výpočty)

3 Aplikace ekonometrických modelů

V předchozí kapitole byla provedena analýza získaných dat, vysvětlované proměnné i regresorů. Byly uvedeny některé statistiky jak vysvětlujících tak vysvětlované proměnné a upozorněno bylo na možné problémy, které by se mohly vyskytnout při samotné specifikaci ekonometrických modelů. Nyní tedy lze přistoupit k aplikaci ekonometrických metod.

3.1 Regresní modely s jednou vysvětlující proměnnou

Hned na úvod bude užitečné prozkoumat vlivy regresorů samostatně, tedy vyzkoušet závislosti pomocí jednoduchého regresního modelu. Získán bude díky tomu přehled o výběrovém souboru, zjistí se, které proměnné budou vykazovat silný vliv, a tím pádem s nimi bude možné pracovat i při specifikaci složitějších modelů; naopak u některých proměnných se možná dojde k závěru, že na vysvětlovanou proměnnou žádný přímý vliv nemají. To ovšem nemusí znamenat nic negativního. Statistický důkaz o nezávislosti je často stejně důležitý, jako prokázání vlivu. Anebo taková proměnná může později posloužit jako instrumentální.

Při práci s jednoduchými regresními modely bude navíc užitečné analyzovat vztah pohledem na grafické zobrazení závislosti. Díky němu se může lineární závislost ukázat jako ne zcela vhodná. Ke zvážení potom bude použití například kvadratické funkce. Za druhé může být z dat zřejmé, že bude lepší použít logaritmickou transformaci. Ta je vhodná, pokud v datech existují malé rozdíly pro nízké hodnoty, intervaly se postupně rozšiřují a mezi vysokými hodnotami jsou rozdíly největší. Za třetí, pohledem na graf lze odhalit například heteroskedasticitu.

Není třeba vypisovat každý model zvlášť, pro rychlost a přehlednost jsou výsledky všech jednoduchých regresních modelů uvedeny v tabulce 3-1. Tabulka obsahuje pro každý model odhad parametru β, jeho p-hodnotu a koeficient determinace modelu (obyčejný i korigovaný). Pro ilustraci může být uveden příklad jednoduchého regresního modelu s proměnnou hustota zalidnění.

$$mira_umrtnosti_i = \beta_0 + \beta_1 hust_zal_i + u_i. \tag{3.1}$$

U některých proměnných se zřejmě objeví i model s kvadratickou formou závislosti, nebo logaritmickou transformací, jak bylo popsáno výše.

$$mira_umrtnosti_i = \beta_0 + \beta_1 gni_i + u_i. \tag{3.2}$$

$$mira_umrtnosti_i = \beta_0 + \beta_1 \log(gni)_i + u_i. \tag{3.3}$$

Vešechny modely třetí kapitoly jsou odhadovány pro kompletní soubor 136 pozorování, tedy $i = 1,2,\dots,136$.

Proměnná	Odhad parametru	*p*-hodnota	Koeficient determinace	Korigovaný koeficient determinace
hust_zal	-0,00226	0,0315	0,034040	0,026832
log(hust_zal)	-1,55478	0,0015	0,072928	0,066009
urban	-0,11460	0,0001	0,104312	0,097628
gni	-0,00024	<0,0001	0,292482	0,287202
log(gni)	-2,89543	<0,0001	0,271998	0,266565
hdi	-25,82150	<0,0001	0,295676	0,290420
hdi hdi^2	71,83150 -80,24100	0,0005 <0,0001	0,405087	0,396141
rychlost_mesto	0,12260	0,0186	0,040642	0,033482
rychlost_mimo	-0,05533	0,1310	0,016934	0,009598
alk_norm	11,34140	0,6658	0,001396	-0,006056
alk_nov	47,38130	0,0421	0,030473	0,023238
alk_prof	18,24010	0,4253	0,004750	-0,002677
helma	2,12969	0,7149	0,000999	-0,006456
pasy	-5,72160	0,0001	0,103325	0,096634
reg_voz	-0,01621	<0,0001	0,283912	0,278568
log(reg_voz)	-2,25707	<0,0001	0,169877	0,163682
srazky	0,00011	0,9089	0,000098	-0,007364
evropa	-9,42181	<0,0001	0,284196	0,278854
sev_amer	-5,99301	0,3031	0,007912	0,000508
stred_amer	2,06902	0,4025	0,005236	-0,002188
jiz_amer	6,70310	0,0117	0,046533	0,039418
afrika	7,89438	<0,0001	0,158950	0,152673
asie	2,15634	0,1738	0,013759	0,006399
australie	-7,02514	0,0383	0,031641	0,024414
daln_obyv	-0,23357	0,0002	0,101539	0,094834

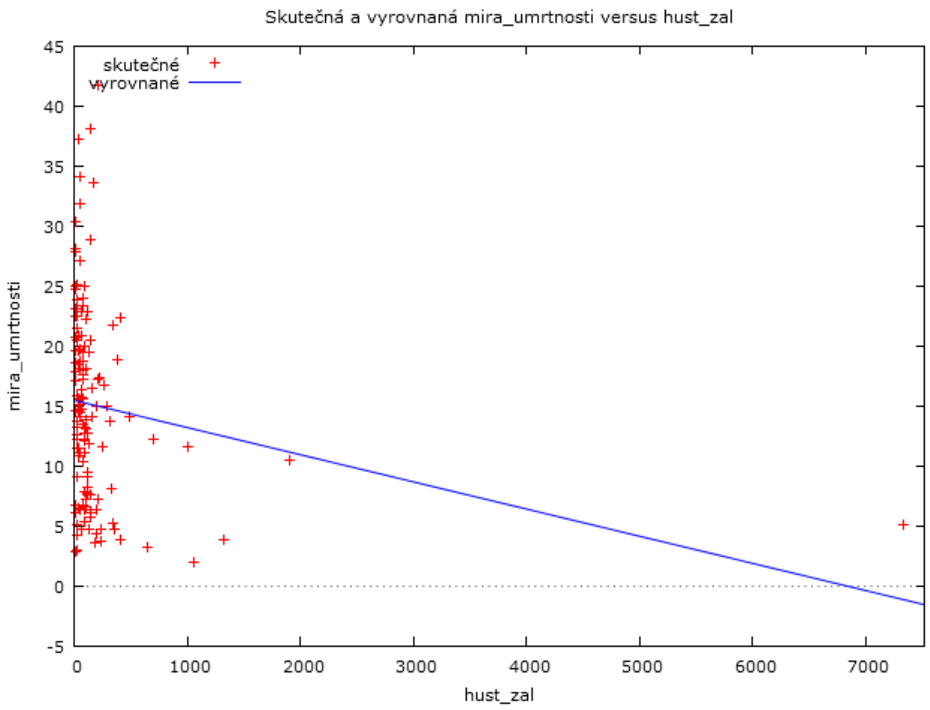
Proměnná	Odhad parametru	*p*-hodnota	Koeficient determinace	Korigovaný koeficient determinace
log(daln_obyv)	-2,46233	0,0003	0,161412	0,150378
daln_rozl	-0,10464	0,0005	0,087280	0,080468
log(daln_rozl)	-1,95721	<0,0001	0,185707	0,174993
daln_reg_voz	-0,00545	0,4695	0,003912	-0,003522
log(daln_reg_voz)	-0,98247	0,2747	0,015675	0,002724

Tabulka 3-1, Regresní modely s jednou vysvětlující proměnnou (zdroj - výpočty v softwaru gretl)

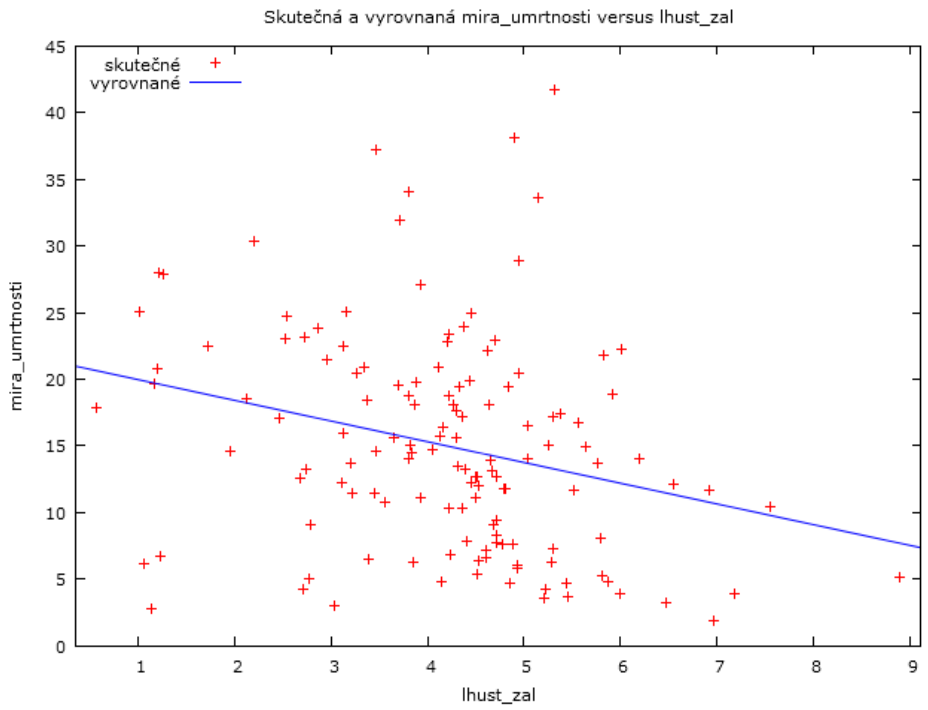
V následujících odstavcích jsou uvedeny poznámky k některým zajímavým výsledkům předchozích modelů.

Hustota zalidnění vykazuje ideální znaky pro logaritmickou transformaci. Data jsou mnohem hustší pro nízké hodnoty, ale pro vyšší existují naopak velmi odlehlá pozorování (odlehlá z hlediska hodnot vysvětlující proměnné). Při použití logaritmu se výsledky zlepšily, a to jak z hlediska statistické významnosti regresního parametru, tak koeficientu determinace.

Následující grafy zobrazují závislost jak proměnné hustota zalidnění, tak její logaritmické transformace (pomocí přirozeného logaritmu).



Graf 3-1, Závislost míry úmrtnosti na hustotě zalidnění (zdroj – graf ze softwaru gretl)

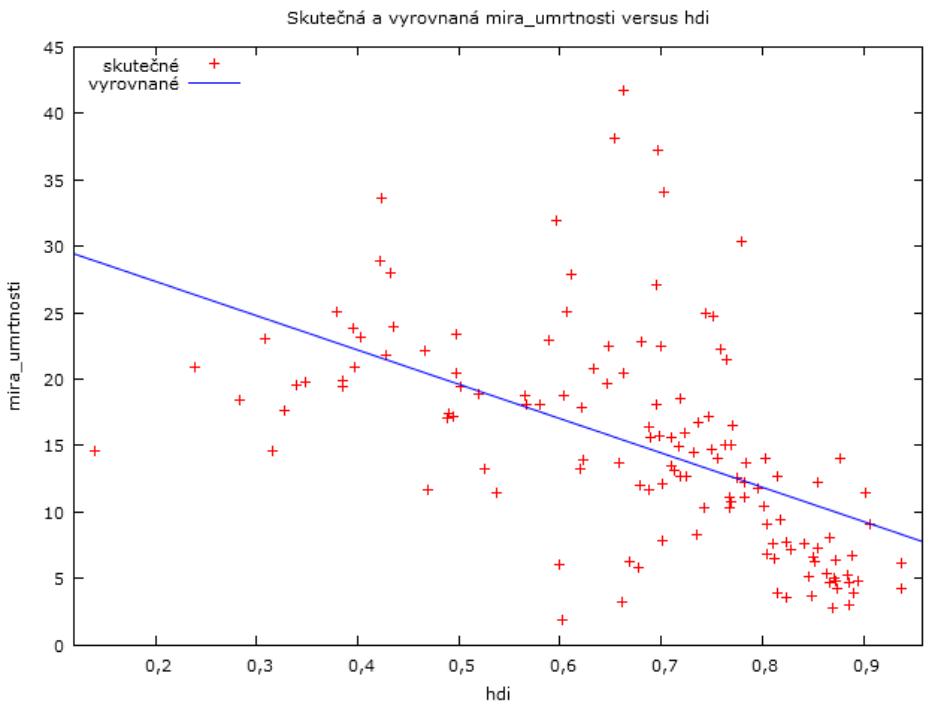


Graf 3-2, Závislost míry úmrtnosti na logaritmech hustoty zalidnění (zdroj – graf ze softwaru gretl)

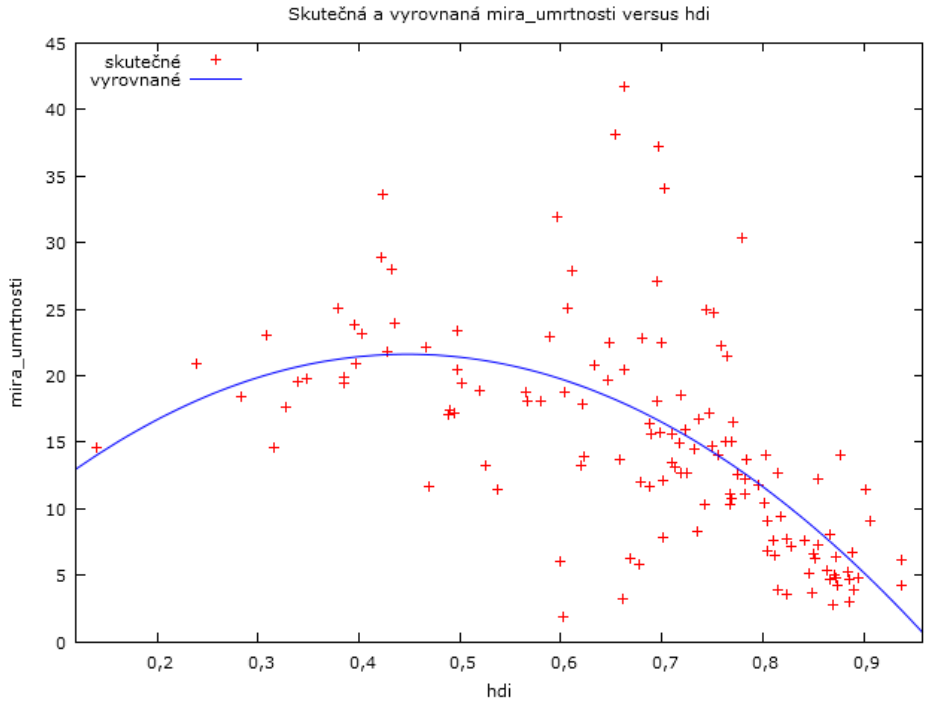
Poměrně zajímavý výsledek poskytuje odhad modelu pro hrubý národní důchod (gni). Použití jednoduchého regresního modelu nebo logaritmické transformace totiž vykazuje přibližně podobnou kvalitu (viz *p*-hodnota a koeficient determinace). Při odhadování složitějších modelů mohou být vyzkoušeny obě varianty. Přece jen o něco lepší výsledky by ale mohl dávat model s logaritmovanými daty, jelikož u jednoduchého modelu je na první pohled zřejmá heteroskedasticita.

Velmi potěšující jsou výsledky u regresoru HDI. Už model s lineární formou závislosti vykazuje koeficient determinace 0,296 (korigovaný 0,290). Při použití kvadratické formy se ale ještě výrazně zvýší na 0,405 (korigovaný 0,396). Parabola tedy vysvětluje závislost v datech více. Nejvyšší úmrtnost přisuzuje zemím s hodnotou HDI 0,4476. Pro to by ale mělo být nalezeno vhodné vysvětlení. Zřejmě to bude znamenat, že ve vyspělých státech je úmrtnost relativně nižší z důvodů kvalitní infrastruktury, zdravotnictví (vliv záchranné služby po potenciální nehodě), nutnost absolvování autoškoly a zkoušek pro získání řidičského oprávnění. Země s nejnižším HDI budou naopak také velmi chudé, to znamená, že lidé budou obecně vlastnit méně dopravních prostředků. Nekvalitní infrastruktura neumožní jízdu ve vysoké rychlosti atp. V zemích s hodnotou HDI okolo zmiňovaných středních hodnot se tyto dva principy kombinují a více lidí si pořídí auto, které používá na ne příliš kvalitních komunikacích. Po nehodě má vliv hůře fungující systém záchranné služby, což přispěje k vyšší úmrtnosti.

Obě dvě varianty, tedy jak použití neupravených, tak logaritmovaných dat zachycují následující grafy 3-3 a 3-4.



Graf 3-3, Závislost míry úmrtnosti na HDI (zdroj – graf ze softwaru gretl)



Graf 3-4, Závislost míry úmrtnosti na HDI, kvadratická forma (zdroj – graf ze softwaru gretl)

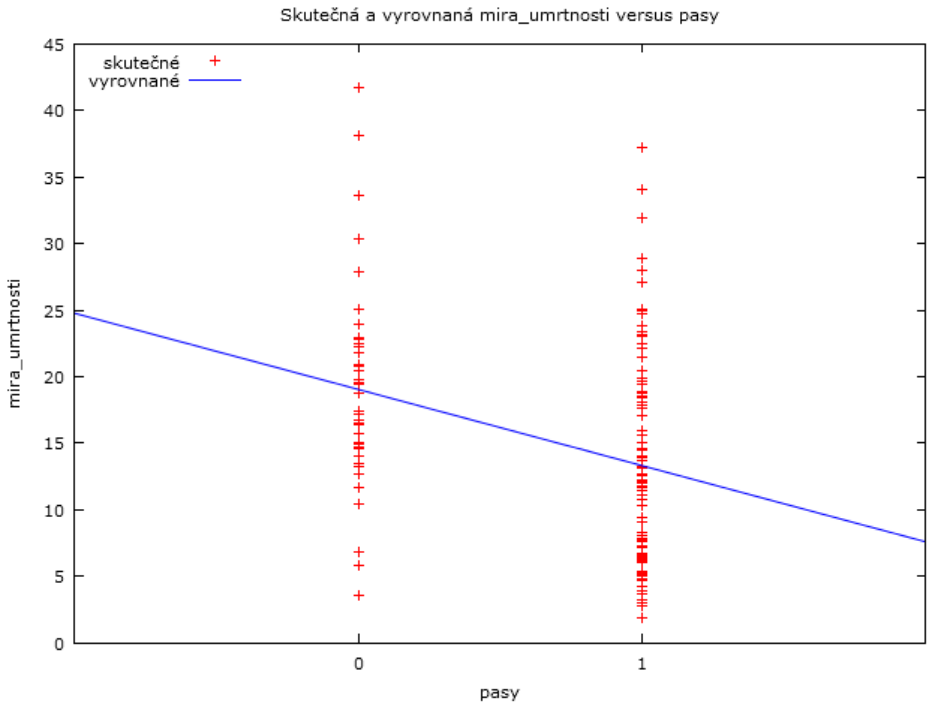
Zklamáním je naopak vliv maximální povolené rychlosti. Oba dva modely mají nízkou vysvětlovací schopnost. Rychlost mimo město dokonce vykazuje úplně opačnou závislost, než jaká byla předpokládaná (tedy že vyšší rychlost bude mít za následek vyšší úmrtnost). Důvod je zřejmě takový, že mnoho fatálních nehod se stane při rychlostech, které převyšují povolené hranice. Dobrým regresorem by tedy možná byla spíše taková proměnná, která by

například udávala poměr času, po který průměrný řidič jede vyšší rychlostí než povolenou. Nebo ještě lépe, o kolik kilometrů v hodině v průměru je tato rychlost překročena. Je evidentní, že taková data bude nemožné vyhledat.

V podstatě katastrofálně dopadla analýza pro maximální povolený obsah alkoholu v krvi řidiče. Ze tří variant (běžný řidič, nováček a profesionál) pouze u nováčků se podařilo prokázat statistickou významnost parametru beta alespoň na pětiprocentní hladině významnosti. Ani tento model ale nemá příliš vysokou kvalitu. Mohou být zopakovány závěry z minulého odstavce. Jeden aspekt jsou maximální povolené hodnoty, realita a dodržování těchto předpisů je věc druhá. Anebo je možné, že alkohol v malém množství skutečně nemá vliv na vznik nehody.

Další problematickou proměnnou je povinnost nošení přilby na motocyklu. Na toto téma už byla učiněna poznámka v druhé kapitole. Může tedy být zopakováno, že povinnost existuje ve všech zemích kromě Afghánistánu a Kiribati. Jelikož v Kiribati je úmrtnost poměrně nízká, výsledky tohoto modelu vůbec neodpovídají původní hypotéze. Data vykazují opačnou závislost. Navíc *p*-hodnota a koeficient determinace potvrzují nízkou kvalitu modelu.

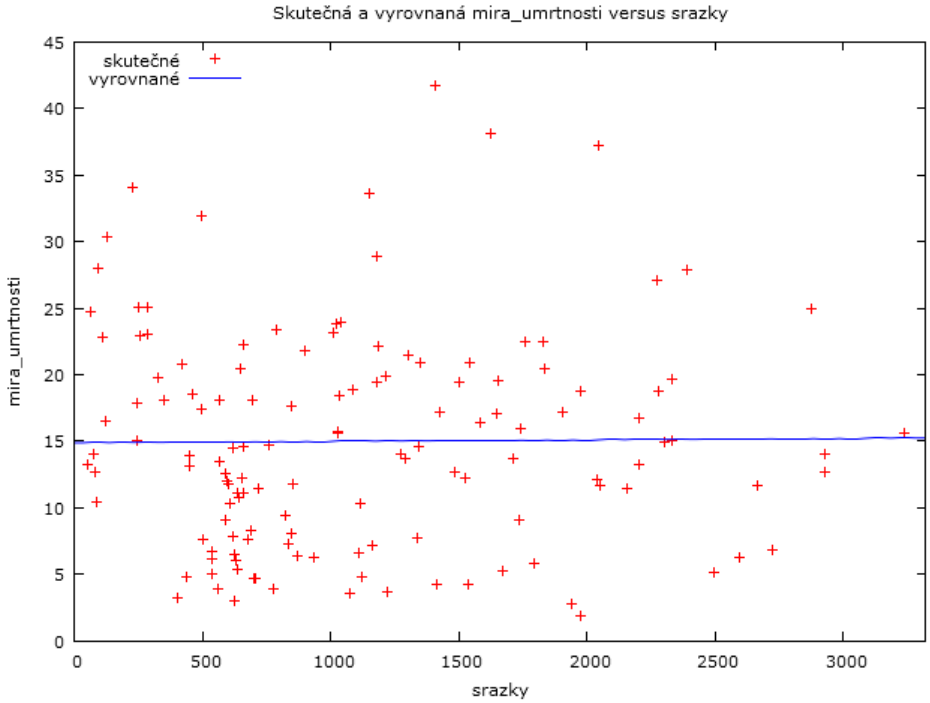
Povinnost být při jízdě připoután bezpečnostními pásy je proměnná, která se na první pohled podobá předchozí. Na rozdíl od helmy ale model alespoň určitou vysvětlovací schopnost má. V zemích, kde je tato povinnost zavedena, je možné očekávat nižší úmrtnost v průměru o 5,72 lidí ze sta tisíc obyvatel. Graficky je tato závislost znázorněna na grafu 3-5.



Graf 3-5, Závislost míry úmrtnosti na proměnné *pasy* (zdroj – graf ze softwaru gretl)

Stejně tak faktor roční úhrn srážek překvapil spíše nepříjemně navzdory očekáváním. Model téměř žádnou závislost neprokazuje (*p*-hodnota 0,91). S ironií lze tvrdit, že na

desetiprocentní hladině významnosti prokazuje nezávislost úmrtnosti na úhrnu srážek. Z hlediska ekonomické interpretace by to mohlo znamenat, že jednotlivé země zahrnují nepřízeň počasí do svých pravidel silničního provozu, také řidiči se v náročnějších podmínkách naučí jezdit bezpečně, a tak nakonec nepřízeň počasí nemá vliv na úmrtnost. Závislost je vyobrazena na grafu 3-6.



Graf 3-6, Závislost míry úmrtnosti na ročním úhrnu srážek (zdroj – graf ze softwaru gretl)

Při jednoduché regresi s regionálními proměnnými nemůžou být očekávány žádné objevné výsledky. Model pouze prozradí, jestli se průměr zemí daného regionu liší od průměru ostatních. Logicky tedy statisticky významně vyjdou odhady parametrů u regionů, které se výrazněji odlišují. V našem případě to bude Evropa (která má úmrtnost výrazně nižší než zbytek světa) a Afrika (naopak s vyšší úmrtností).

Další statisticky významnou proměnnou je délka dálniční sítě na 100 tisíc obyvatel. Potvrdil se očekávaný směr závislosti – tedy že více kilometrů dálnic bude znamenat méně úmrtí. Na data by bylo vhodné aplikovat logaritmickou transformaci, problém spočívá ale v tom, že 58 zemí má délku dálniční sítě 0km. Logaritmus pro tato pozorování tedy neexistuje. Model lze sestavit pouze pro 78 zbývajících zemí, které mají alespoň několik kilometrů dálnice. Koeficient determinace je u omezeného modelu vyšší, nicméně to není překvapivé (vyřazeno bylo 58 zemí se stejnou délkou dálnice, avšak různými hodnotami úmrtnosti). Navíc je žádoucí, aby výsledný model zahrnoval všechny země. Model s logaritmy je uveden tedy pouze pro zajímavost.

Rovněž tak může být interpretováno použití logaritmů u dvou dalších variant, ve kterých je délka dálniční sítě k dispozici (vztažení na rozlohu a 100 tisíc registrovaných vozidel). Obecně nejlepší výsledky poskytuje první model, tedy vztažení hodnot na obyvatele.

3.2 Vícerozměrné regresní modely

Po úvodní podkapitole aplikační části práce, kde byla analyzována závislost pouze na jednotlivých nezávislých proměnných odděleně, se naskýtá možnost použít více proměnných najednou. Bude se tedy jednat o vícerozměrné regresní modely.

U nich už existuje mnohem více možností aplikace regresní analýzy, lze očekávat vyšší hodnoty koeficientu determinace (tedy vyšší vysvětlenou variabilitu závislé proměnné).

Na začátek je možné analyzovat například vliv binárních proměnných, které mají z hlediska interpretace vztah k úrovňové konstantě. Významnou skupinou binárních proměnných jsou proměnné regionů. Statisticky významné regresní koeficienty se dají očekávat pouze u těch regionů, které se od zbytku výběru liší (v úrovni míry úmrtnosti). Možností, jak přistoupit k analýze, je více. Buďto lze zařadit do výběru regiony s nejnižší úmrtností, nebo naopak s nejvyšší. Případně lze aplikovat třetí možnost, zařadit nejextrémnější hodnoty z obou konců, které se ale musí významně lišit od zbytku datového souboru (kde bude míra úmrtnosti přibližně průměrná).

Při této práci může být nápomocná tabulka 3-1 z předchozí podkapitoly. Uvádí výsledky jednoduché jednorozměrné regrese regionálních proměnných. Za povšimnutí stojí především ty regiony, u nichž je parametr β významně různý od nuly. Regiony s nejnižší úmrtností jsou Evropa a Austrálie a Oceánie, naopak s nejvyšší Afrika a Jižní Amerika. Regresní parametr u zbylých tří proměnných nebyl v jednoduchém modelu významný. Nicméně je možné zkusit postupně zařazovat do modelu proměnné z jednoho či druhého konce (případně z obou) a sledovat, jak se bude měnit vysvětlující schopnost modelu a hodnoty jednotlivých parametrů.

Tento postup znázorňují tabulky 3-2 až 3-4, které mají stejnou formu jako tabulka 3-1 z předchozí podkapitoly. Obsahují tedy jednotlivé odhady parametrů, jejich *p*-hodnoty a obě varianty koeficientu determinace. Pro všechny parametry platí stejná interpretace. Hodnota parametru říká, o kolik se v průměru liší země daného regionu od těch, jejichž regionální proměnná není v modelu zahrnuta. Je to tedy odchylka od úrovňové konstanty.

Proměnné	Odhady parametrů	*p*-hodnoty	Koeficient determinace	Korigovaný koeficient determinace
evropa	-9,42181	<0,0001	0,284196	0,278854
evropa *australie*	-10,06600 -10,19980	<0,0001 0,0004	0,349566	0,339785

Proměnné	Odhady parametrů	*p*-hodnoty	Koeficient determinace	Korigovaný koeficient determinace
evropa	-10,28190	<0,0001	0,369730	0,355405
australie	-10,41570	0,0002		
sev_amer	-9,60532	0,0419		

Tabulka 3-2, Modely s regiony s nízkou mírou úmrtnosti (zdroj – výpočty v softwaru gretl)

V tabulce 3-2 jsou uvedeny výsledky modelu, do kterého byly postupně přidávány další regiony (počínaje od nejnižší míry úmrtnosti). Přidáním Severní Ameriky se koeficient determinace (ani adjustovaný) už příliš nezvýšil a samotný regresní koeficient je statisticky významný, ale pouze na pětiprocentní hladině významnosti. Přidávat další region (konkrétně by to byla Střední Amerika) tedy již nebude mít velký smysl. Ta má totiž míru úmrtnosti v porovnání s předchozími třemi o poznání vyšší, a tak lze očekávat, že by její regresní parametr nebyl významný.

Následuje podobná analýza, ale jak již bylo avizováno, z druhého konce – tedy od regionů s nejvyšší mírou úmrtnosti.

Proměnné	Odhady parametrů	*p*-hodnoty	Koeficient determinace	Korigovaný koeficient determinace
afrika	7,89438	<0,0001	0,158950	0,152673
afrika	8,70815	<0,0001	0,235782	0,224290
jiz_amer	8,70736	0,0004		
afrika	11,11760	<0,0001	0,342445	0,327500
jiz_amer	11,11680	<0,0001		
asie	6,49212	<0,0001		
afrika	12,78120	<0,0001	0,419504	0,401779
jiz_amer	12,78040	<0,0001		
asie	8,15572	<0,0001		
stred_amer	8,45664	<0,0001		

Tabulka 3-3, Modely s regiony vysokou mírou úmrtnosti (zdroj – výpočty v softwaru gretl)

Tento druhý postup vygeneroval poměrně kvalitní modely, alespoň při porovnání s prvním. Jakmile se do modelu přidávají další a další proměnné, statistická významnost

jednotlivých parametrů roste a koeficient determinace i korigovaný koeficient determinac se každým krokem zvyšují. Všechny regresní parametry, dokonce i u modelu se čtyřmi nejhoršími regiony (s hlediska úmrtnosti), jsou významné, a to dokonce více než na jednoprocentní hladině.

A konečně zmiňme třetí způsob přidávání proměnných. Začneme od modelu, který obsahuje oba dva nejextrémnější regiony. Postupně se budou přidávat další proměnné a opět se bude sledovat, jaký vliv to bude mít na výsledný model.

Proměnné	Odhady parametrů	*p*-hodnoty	Koeficient determinace	Korigovaný koeficient determinace
evropa	-7,94078	<0,0001	0,337208	0,327242
afrika	4,85165	0,0014		
evropa	-8,74824	<0,0001	0,385311	0,371340
australie	-8,88203	0,0016		
afrika	4,04420	0,0064		
evropa	-7,07455	<0,0001	0,368209	0,353850
jiz_amer	5,71708	0,0121		
afrika	5,71788	0,0002		
evropa	-7,93956	<0,0001	0,407241	0,389142
australie	-8,07335	0,0039		
jiz_amer	4,85208	0,0294		
afrika	4,85288	0,0014		

Tabulka 3-4, Modely s regiony s oběma extrémními hodnotami (zdroj – výpočty v softwaru gretl)

Při aplikaci třetího principu mají výsledné modely rovněž poměrně vysoké koeficienty determinace (v obou variantách), statistická významnost parametrů je ale nižší v porovnání s předchozími dvěma přístupy. Zajímavé bylo také pořadí, v jakém se proměnné přidávaly. Úvodní model obsahuje pouze Evropu a Afriku, tedy dva nejextrémnější regiony co do výše míry úmrtnosti na silničních komunikacích. Je otázka, z jakého konce přidat další proměnnou. Vyzkoušeny byly obě varianty. O něco lépe na tom byla varianta s přidáním Austrálie a Oceánie jakožto třetí proměnné. Při vstupu všech čtyř proměnných do modelu je vidět další snížení statistické významnosti téměř všech parametrů, které by pokračovalo i po přidání dalších proměnných ať z té či oné strany pomyslné výsledkové listiny mezi regiony (tyto modely už ani není potřeba uvádět).

Je na čase shrnout první krok této podkapitoly, analýzu pouze s regionálními binárními proměnnými. Analýza byla provedena třemi různými způsoby, přičemž nejlepší výsledky vykazuje model, jehož vysvětlující proměnné jsou Afrika, Jižní Amerika, Asie a Střední Amerika. Koeficienty takto konstruovaného modelu jsou statisticky významné (Jižní Amerika na pětiprocentní, ostatní regiony dokonce na jednoprocentní hladině významnosti) a model vysvětluje téměř 42% variability míry úmrtnosti. Další vícerozměrné modely tedy mohou navazovat na uvedený model.

Pro plynulou návaznost na předchozí regionální proměnné by bylo dobré pokračovat v binárních proměnných. Ty už jsou k dispozici pouze dvě: pásy a helma. O problémech s proměnnou helma už bylo napsáno dost a do modelu se zařazovat nebude. Další proměnnou, která tedy vstoupí do modelu, bude povinnost být v automobilu připoután na všech místech.

K výsledkům následujících modelů, které budou vždy vyobrazeny formou stejné tabulky, budou připojeny také výsledky testů heteroskedasticity. Konkrétně to bude Breusch-Paganův test. Výstup výsledků Breusch-Paganova testu bude obsahovat vždy *p*-hodnoty koeficientů δ_j (ty prozradí, která proměnná případnou heteroskedasticitu způsobuje) a také *p*-hodnotu testu jako celku (dle LM statistiky).

Dále bude pro každý model zjišťována multikolinearita a to pomocí metody popsané v aplikační části práce. Pro připomenutí je třeba poznamenat, že jestliže hodnota VIF_j překročí hodnotu 10, bude to znamenat silnou multikolinearitu.

Do modelu s regionálními proměnnými vstoupí tedy nová proměnná – *pasy*. Výsledky odhadů, test heteroskedasticity a měření multikolinearity se nacházejí v tabulkách 3-5 až 3-7.

$$mira_umrtnosti_i = \beta_0 + \beta_1 afrika_i + \beta_2 jiz_amer_i + \beta_3 asie_i + \beta_4 stred_amer_i + \beta_5 pasy_i + u_i. \tag{3.4}$$

Proměnné	Odhady parametrů	*p*-hodnoty	Koeficient determinace	Kor. koeficient determinace
β_0	11,10690	<0,00001	0,441954	0,420490
afrika	11,81180	<0,00001		
jiz_amer	12,72680	<0,00001		
asie	7,01583	<0,00001		
stred_amer	7,23568	0,00063		
pasy	-2,91840	0,02382		

Tabulka 3-5, Výsledky modelu 3.4 (zdroj – výpočet v softwaru gretl)

Proměnné	*p*-hodnoty	*p*-hodnota testu
δ_0	0,2683	0,001529
afrika	0,4544	
jiz_amer	0,1448	
asie	0,0106	
stred_amer	0,1089	
pasy	0,7685	

Tabulka 3-6, Breusch-Paganův test modelu 3.4 (zdroj – výpočet v softwaru gretl)

Proměnné	VIF_j
afrika	1,359
jiz_amer	1,116
asie	1,447
stred_amer	1,216
pasy	1,197

Tabulka 3-7, Zjišťování multikolinearity v modelu 3.4 (zdroj – software gretl)

Proměnná pásy se i mezi regionálními ukázala jako statisticky významná, byť jen na pětiprocentní hladině. Výsledky tohoto modelu lze velmi snadno interpretovat, což je způsobeno bivalentní povahou proměnných. Můžeme tedy klást otázky typu: „Jaká je očekávaná úmrtnost v africké zemi, kde je povinné používání pásů?“ Odpověď lze vypočítat pouhým sčítáním a odčítáním jednotlivých koeficientů. Výsledek by v tomto konkrétním případě byl 11,1069 (úrovňová konstanta) + 11,8118 (afrika) – 2,9184 (pasy) = 20,0003. Očekávaná úmrtnost je tedy 20 lidí ze 100 tisíc obyvatel.

Nulová hypotéza o homoskedasticitě byla na jednoprocentní hladině významnosti zamítnuta. V modelu se tedy vyskytuje heteroskedasticita. Dle *p*-hodnot jednotlivých parametrů jí nejvíce způsobuje proměnná *asie*. Zatím se bude pokračovat v odhadech nových modelů, ale průběžně budou sledovány výsledky testů heteroskedasticity. Na konci podkapitoly bude potřeba se s heteroskedasticitou vyrovnat.

Multikolinearita nebyla zjištěna, všechny hodnoty jsou dostatečně nízké.

Tím by byly všechny binární proměnné vyčerpány a nezbývá než vyzkoušet zařazení ostatních proměnných do modelu. Jako první například hustotu zalidnění. Jak model s netransformovanými daty, tak logaritmy jsou kvalitní, ale přece jen o něco lepší výsledky byly pozorovány u modelu s logaritmickou transformací. Ten má tedy tvar (3.5).

$$mira_umrtnosti_i = \beta_0 + \beta_1 afrika_i + \beta_2 jiz_amer_i + \beta_3 asie_i + \beta_4 stred_amer_i + \beta_5 pasy_i + \beta_6 \log(hust_zal)_i + u_i. \quad (3.5)$$

Proměnné	Odhady parametrů	*p*-hodnoty	Koeficient determinace	Korigovaný koeficient determinace
β_0	16,31850	<0,00001	0,477135	0,452816
afrika	10,95530	<0,00001		
jiz_amer	11,05630	<0,00001		
asie	7,66382	<0,00001		
stred_amer	7,62781	0,00023		
pasy	-3,08888	0,01411		
log(hust_zal)	-1,19001	0,00382		

Tabulka 3-8, Výsledky modelu 3.5 (zdroj – výpočet v softwaru gretl)

Proměnné	*p*-hodnoty	*p*-hodnota testu
δ_0	0,4846	0,011699
afrika	0,4873	
jiz_amer	0,1278	
asie	0,0890	
stred_amer	0,1138	
pasy	0,5663	
log(hust_zal)	0,8379	

Tabulka 3-9, Breusch-Paganův test modelu 3.5 (zdroj – výpočet v softwaru gretl)

Proměnné	VIF_j
afrika	1,412
jiz_amer	1,198
asie	1,482
stred_amer	1,222
pasy	1,200
log(hust_zal)	1,214

Tabulka 3-10, Zjišťování multikolinearity v modelu 3.5 (zdroj – software gretl)

Koeficient determinace (i adjustovaný) se oproti modelu bez hustoty zalidnění zvýšil, všechny parametry jsou významné. Pásy jako jediné nedosáhly jednoprocentní hladiny významnosti, nicméně se jí velmi blíží. Koeficient u hustoty zalidnění se bude komentovat tímto způsobem: Jestliže se hodnota hustoty zalidnění zvýší o jedno procento, potom lze očekávat snížení míry úmrtnosti o 0,0119 člověka ze sto tisíc obyvatel země. Nebo přesněji - jestliže se hustota zalidnění *e*-krát zvýší (*e* je přibližně 2,718), lze očekávat o 1,19 úmrtí na sto tisíc obyvatel méně.

Breusch-Paganův test dopadl lépe, ale stále zamítáme na pětiprocentní hladině hypotézu o homoskedasticitě. Kromě *asie* je heteroskedasticita přítomná zřejmě kvůli proměnným *jiz_amer* a *stred_amer*.

Multikolinearita se v modelu opět neobjevila.

Jiná možnost je vstup proměnné gni (tedy hrubého národního důchodu) namísto předchozí hustoty zalidnění. Opět byla ze dvou možností (tedy logaritmovaných a nelogaritmovaných dat) vybrána pouze jedna, ale tentokrát naopak varianta netransformovaná, jak ukazuje rovnice 3.6. Výsledky odhadnutého modelu jsou vyobrazeny v tabulce 3-11.

$$mira_umrtnosti_i = \beta_0 + \beta_1 afrika_i + \beta_2 jiz_amer_i + \beta_3 asie_i + \beta_4 stred_amer_i + \beta_5 pasy_i + \beta_6 \mathrm{gni}_i + u_i. \quad (3.6)$$

Proměnné	Odhady parametrů	*p*-hodnoty	Koeficient determinace	Korigovaný koeficient determinace
β_0	14,32410	<0,00001	0,499207	0,475915
afrika	8,84537	<0,00001		
jiz_amer	10,34990	<0,00001		
asie	5,33666	0,00033		
stred_amer	4,96893	0,01683		
pasy	-2,81119	0,02215		
gni	-0,00012	0,00019		

Tabulka 3-11, Výsledky modelu 3.6 (zdroj – výpočet v softwaru gretl)

Proměnné	*p*-hodnoty	*p*-hodnota testu
δ_0	0,2930	0,000387
afrika	0,5294	
jiz_amer	0,1308	
asie	0,0249	
stred_amer	0,1159	
pasy	0,7109	
gni	0,6063	

Tabulka 3-12, Breusch-Paganův test modelu 3.6 (zdroj – výpočet v softwaru gretl)

Proměnné	VIF_j
afrika	1,751
jiz_amer	1,218
asie	1,592
stred_amer	1,326
pasy	1,198
gni	1,377

Tabulka 3-13, Zjišťování multikolinearity v modelu 3.6 (zdroj – software gretl)

Model 3.6 má opět o něco vyšší koeficient determinace (zároveň i korigovaný koeficient) a jednotlivé parametry jsou statisticky významné. Zajímavější než popis *p*-hodnot bude spíše interpretace odhadu parametru β_6. Ukazatel hrubého národního důchodu se vykazuje v hodnotách USD na obyvatele. Rozdíl jednoho dolaru je tedy naprosto zanedbatelný. Významnější rozdíly v bohatství států tedy budou, pokud se hodnoty budou lišit o stovky, spíše tisíce či desetitisíce dolarů. U státu, který má HND o 1000 USD na hlavu vyšší než jiný se očekává úmrtnost nižší v průměru o 0,1238 člověka ze sto tisíc obyvatel. K ušetření jednoho lidského života je tedy zapotřebí HND vyšší přibližně o 8077 USD.

Jiný model obsahuje místo gni proměnnou HDI. Výše bylo zmíněno, že u HDI se hodí k vysvětlení vlivu parabola. Tento fakt se potvrdil i při vícerozměrné regresi. Výsledky jsou opět k nahlédnutí v tabulkách 3-14 až 3-16.

$$mira_umrtnosti_i = \beta_0 + \beta_1 afrika_i + \beta_2 jiz_amer_i + \beta_3 asie_i + \beta_4 stred_amer_i + \beta_5 pasy_i + \beta_6 \mathrm{hdi}_i + \beta_7 hdi_i^2 + u_i. \tag{3.7}$$

Proměnné	Odhady parametrů	*p*-hodnoty	Koeficient determinace	Korigovaný koeficient determinace
β_0	0,16895	0,97974	0,520960	0,494762
afrika	7,66015	0,00063		
jiz_amer	9,57589	0,00001		
asie	4,20440	0,00597		
stred_amer	4,52168	0,02676		
pasy	-2,46130	0,04390		
hdi	63,03340	0,00248		
hdi^2	-61,39680	0,00023		

Tabulka 3-14, Výsledky modelu 3.7 (zdroj – výpočet v softwaru gretl)

Proměnné	*p*-hodnoty	*p*-hodnota testu
δ_0	0,1131	0,000040
afrika	0,3753	
jiz_amer	0,3282	
asie	0,0430	
stred_amer	0,2762	
pasy	0,6580	
hdi	0,0539	
hdi^2	0,0622	

Tabulka 3-15, Breusch-Paganův test modelu 3.7 (zdroj – výpočet v softwaru gretl)

Proměnné	VIF_j
afrika	3,256
jiz_amer	1,175
asie	1,679
stred_amer	1,258
pasy	1,233
hdi	2,691

Tabulka 3-16, Zjišťování multikolinearity v modelu 3.7 (zdroj – software gretl)

Tento model (3.7) jako první překročil hranici 0,5 koeficientu determinace. S nadsázkou lze tvrdit, že se podařilo, aby model více variability vysvětloval, než nevysvětloval. Je ale zřejmé, že úkolem ekonometrické analýzy není nekonečná honba za co možná nejvyšším koeficientem determinace, ale v úvahu se musí brát více kritérií (jakými jsou například již probíraná statistická významnost jednotlivých regresních parametrů, adjustovaný koeficient determinace, výskyt heteroskedasticity a ekonomická verifikace vztahu mezi proměnnými).

Co se týče heteroskedasticity, ta se u předchozího modelu rozhodně vyskytuje. Je to způsobeno tím, že od začátku je přítomna proměnná *asie*, ale nově vstupující *hdi* také vykazuje heteroskedasticitu.

I při měření multikolinearity se vyskytly o něco vyšší hodnoty, nicméně stále se nejedná o silnou multikolinearitu.

Důležitá bude také interpretace koeficientů u proměnné *hdi* a hdi^2. Vrchol paraboly se nachází v bodě 0,5133 (výpočet pomocí derivace regresní funkce dle proměnné hdi) a je to

její maximum. Nejvyšší úmrtnost se očekává u zemí s hodnotou ukazatele HDI okolo této hodnoty. V podkapitole o vysvětlujících proměnných bylo uvedeno, že aritmetický průměr hodnot HDI u 136 zemí ve výběrovém souboru je přibližně 0,68. Analýza tedy přisuzuje vyšší úmrtnost spíše podprůměrným zemím z hlediska *hdi*.

Následuje model (3.8):

$$mira_umrtnosti_i = \beta_0 + \beta_1 afrika_i + \beta_2 jiz_amer_i + \beta_3 asie_i + \beta_4 stred_amer_i + \beta_5 pasy_i + \beta_6 \mathrm{srazky}_i + u_i. \tag{3.8}$$

Proměnné	Odhady parametrů	*p*-hodnoty	Koeficient determinace	Korigovaný koeficient determinace
β_0	12,70070	<0,00001	0,463154	0,438184
afrika	12,01270	<0,00001		
jiz_amer	14,58640	<0,00001		
asie	7,29619	<0,00001		
stred_amer	9,08683	0,00006		
pasy	-2,88593	0,02325		
srazky	-0,00182	0,02569		

Tabulka 3-17, Výsledky modelu 3.8 (zdroj – výpočet v softwaru gretl)

Proměnné	*p*-hodnoty	*p*-hodnota testu
δ_0	0,1450	0,000788
afrika	0,5035	
jiz_amer	0,0498	
asie	0,0126	
stred_amer	0,0736	
pasy	0,6263	
srazky	0,4064	

Tabulka 3-18, Breusch-Paganův test modelu 3.8 (zdroj – výpočet v softwaru gretl)

Proměnné	VIF_j
afrika	1,364
jiz_amer	1,285
asie	1,458
stred_amer	1,414
pasy	1,198
srazky	1,294

Tabulka 3-19, Zjišťování multikolinearity v modelu 3.8 (zdroj – software gretl)

Tento model obsahuje proměnnou srážky. Výsledek je poměrně překvapivý, jelikož při jednorozměrné analýze se srážky ukázaly jako naprosto nevyhovující regresor. Přesto u tohoto vícerozměrného modelu je parametr statisticky významný. Překvapující je navíc směr závislosti. Podle odhadů modelu (3.8) s přibývajícími srážkami ubývá počet úmrtí při autonehodách, což neodpovídá původní hypotéze. Tato situace bude zřejmě způsobena určitou souhrou více proměnných. Například je možná určitá korelace mezi srážkami a jinou vysvětlující proměnnou, což přidává regresnímu koeficientu u proměnné srážky na významnosti.

V závěru této podkapitoly budou uvedeny ještě dva modely, oba ještě s vyšším počtem proměnných.

$$mira_umrtnosti_i = \beta_0 + \beta_1 afrika_i + \beta_2 jiz_amer_i + \beta_3 asie_i + \beta_4 stred_amer_i + \beta_5 pasy_i + \beta_6 \mathrm{hdi}_i + \beta_7 hdi_i^2 + \beta_8 \mathrm{srazky}_i + u_i. \tag{3.9}$$

Proměnné	Odhady parametrů	*p*-hodnoty	Koeficient determinace	Korigovaný koeficient determinace
β_0	2,77362	0,67207	0,548531	0,520092
afrika	7,43313	0,00067		
jiz_amer	11,55870	<0,00001		
asie	4,36096	0,00352		
stred_amer	6,52579	0,00227		
pasy	-2,36828	0,04671		
hdi	62,56290	0,00208		
hdi^2	-62,08180	0,00013		
srazky	-0,00208	0,00617		

Tabulka 3-20, Výsledky modelu 3.9 (zdroj – výpočet v softwaru gretl)

Proměnné	*p*-hodnoty	*p*-hodnota testu
δ_0	0,1439	0,000041
afrika	0,3504	
jiz_amer	0,1681	
asie	0,0399	
stred_amer	0,2362	
pasy	0,5870	
hdi	0,0745	
hdi^2	0,0930	
srazky	0,6996	

Tabulka 3-21, Breusch-Paganův test modelu 3.9 (zdroj – výpočet v softwaru gretl)

Proměnné	VIF_j
afrika	3,261
jiz_amer	1,326
asie	1,682
stred_amer	1,439
pasy	1,234
hdi	2,718
srazky	1,308

Tabulka 3-22, Zjišťování multikolinearity v modelu 3.8 (zdroj – software gretl)

Model (3.9) obsahuje kromě regionů proměnné pásy, hdi (a to v kvadratickém tvaru závislosti) a již diskutované srážky. Ačkoli počet proměnných je poměrně vysoký, jednotlivé koeficienty si stále udržují statistickou významnost.

Nicméně s rostoucím počtem proměnných se prohlubuje problém heteroskedasticity. Způsobuje to opět přítomnost proměnných *asie* a *hdi*.

Zjiťování multikolinearity opět neodhalilo významný problém, je ale možné si všimnout jisté korelace mezi proměnnými *hdi* a *afrika*, což není rozhodně překvapující.

Jestliže vstoupí do tohoto modelu ještě navíc proměnná hustota zalidnění (v logaritmované podobě), vznikne tím model (3.10). Výsledky jsou opět přehledně uvedeny v tabulce 3-23, výsledky Breusch-Paganova testu se nachází v tabulce 3-24 a měření multikolinearity v tabulce 3-25.

$$\begin{gathered} mira_umrtnosti_i = \beta_0 + \beta_1 afrika_i + \beta_2 jiz_amer_i + \beta_3 asie_i + \\ \beta_4 stred_amer_i + \beta_5 pasy_i + \beta_6 hdi_i + \beta_7 hdi_i^2 + \beta_8 srazky_i + \\ \beta_9 log\ (hust_zal)_i + u_i. \end{gathered} \tag{3.10}$$

Proměnné	Odhady parametrů	*p*-hodnoty	Koeficient determinace	Korigovaný koeficient determinace
β_0	5,84588	0,36910	0,572898	0,542391
afrika	7,04975	0,00097		
jiz_amer	9,89262	0,00002		
asie	4,96489	0,00083		
stred_amer	6,58425	0,00163		
pasy	-2,56554	0,02796		
hdi	64,47730	0,00119		
hdi^2	-62,76760	0,00008		
srazky	-0,00175	0,01926		
log(hust_zal)	-1,00676	0,00832		

Tabulka 3-23, Výsledky modelu 3.10 (zdroj – výpočet v softwaru gretl)

Proměnné	*p*-hodnoty	*p*-hodnota testu
δ_0	0,1307	0,000057
afrika	0,3309	
jiz_amer	0,1208	
asie	0,1125	
stred_amer	0,2167	
pasy	0,5513	
hdi	0,0812	
hdi^2	0,0932	
srazky	0,7724	
log(hust_zal)	0,5372	

Tabulka 3-24, Breusch-Paganův test modelu 3.10 (zdroj – výpočet v softwaru gretl)

Proměnné	VIF_j
afrika	3,277
jiz_amer	1,441
asie	1,729
stred_amer	1,440
pasy	1,239
hdi	2,739
srazky	1,344
log(hust_zal)	1,255

Tabulka 3-25, Zjišťování multikolinearity v modelu 3.8 (zdroj – software gretl)

Na závěr podkapitoly s vícerozměrnými modely by bylo vhodné výsledky posledního modelu komentovat. Celkem čtyři regiony (Afrika, Jižní Amerika, Asie a Střední Amerika) vykazují statisticky významné odlišnosti ve výši úmrtnosti oproti zbylým třem (Evropa, Austrálie a Oceánie a Severní Amerika). Pro odhad úmrtnosti se musí tedy čtyřem zmíněným regionům k úrovňové konstantě přičíst navíc regresní parametr u proměnné právě daného regionu.

Pokud je v zemi zavedena povinnost používání bezpečnostních pásů (a to na všech sedadlech v automobilu), ušetří se tím více než 2,5 lidských životů ze sto tisíc obyvatel země. To odpovídá očekávanému vlivu této proměnné.

Závislost úmrtnosti na výši ukazatele HDI má kvadratickou formu. Nejvyšší úmrtnost přisuzuje model zemím s HDI na hodnotě 0,5136. Tento fakt můžeme komentovat tak, že nejvyšší úmrtnost je způsobena například nekvalitní infrastrukturou a službami. Naopak země s vyšším HDI mají úmrtnost nízkou. V ukazateli se totiž odráží bohatství země (což může mít vliv právě na kvalitu komunikací) a délku školní docházky. Vyšší vzdělání zřejmě přispívá k chápání vyšší hodnoty života. Vzdělaní lidé budou spíše tíhnout k bezpečnější jízdě, jelikož nechtějí způsobit újmu sobě, ani nikomu jinému.

Nejtěžší na interpretaci bude proměnná srážky. Podle odhadnutého modelu by vyšší roční úhrn srážek měl znamenat nižší úmrtnost. Koeficient se vztahuje ke změně o jeden milimetr srážek za rok. Při změně o 1000 milimetrů by model odhadoval úmrtnost nižší o 1,7 lidských životů (očekává se jeden ušetřený život při zvýšení úhrnu srážek o 570 mm ročně). Tento výsledek naprosto neodpovídá původní hypotéze o tom, že v náročnějších podmínkách (sníh, déšť) nastane více autonehod. Jak bylo zmíněno výše, touto proměnnou bude potřeba zabývat se ještě později.

A konečně hustota zalidnění. Zvýší-li se hustota zalidnění o jedno procento, očekává se přibližně o 0,01 úmrtí na sto tisíc obyvatel méně.

U několika předešlých modelů byla testována homoskedasticita pomocí Breusch-Paganova testu. U všech byla zamítnuta hypotéza o homoskedasticitě a to především díky proměnným *asie*, *stred_amer*, *jiz_amer* a *hdi*. Heteroskedasticita nemá vliv na nestrannost estimátorů. Má ale vliv na vydatnost odhady standardních chyb. Proto bude užitečné vybrat alespoň jeden z předchozích modelů a provést odhady s pomocí metody robustních standardních chyb. Vyberme například model (3.7), kde se nachází všechny jmenované proměnné.

$$mira_umrtnosti_i = \beta_0 + \beta_1 afrika_i + \beta_2 jiz_amer_i + \beta_3 asie_i + \beta_4 stred_amer_i + \beta_5 pasy_i + \beta_6 \mathrm{hdi}_i + \beta_7 hdi_i^2 + u_i. \quad (3.7)$$

Tabulka 3-26 uvádí výsledky odhadů pomocí metody nejmenších čtverců, a to s běžnými standardními chybami, ale také s robustními odhady.

Proměnné	Odhady parametrů	Běžné standardní chyby	*p*-hodnoty	Robustní standardní chyby	*p*-hodnoty
β_0	0,16895	6,63830	0,97974	5,01536	0,9732
afrika	7,66015	2,18639	0,00063	2,16473	0,0006
jiz_amer	9,57589	2,12378	<0,00001	2,34453	<0,0001
asie	4,20440	1,50374	0,00597	1,64450	0,0117
stred_amer	4,52168	2,01779	0,02676	2,03586	0,0281
pasy	-2,46130	1,20937	0,04390	1,34397	0,0694
hdi	63,03340	20,41750	0,00248	15,27130	<0,0001
hdi^2	-61,39680	16,17330	0,00023	12,30600	<0,0001

Tabulka 3-26, Běžné a robustní standardní chyby modelu 3.7 (zdroj – výpočty v softwaru gretl gretl)

Použití robustních standardních chyb mělo vliv na statistickou významnost jednotlivých parametrů, ale hodnoty odhadů samotných parametrů se nemění. Paradoxně největší byl vliv na proměnnou *pasy*, která již není významná ani na pětiprocentní hladině významnosti. U regionálních proměnných se standardní chyby oproti běžným zvýšily (dle očekávání). Naopak je překvapivé, že u proměnná *hdi* svoji statistickou významnost zvýšila.

Tímto modelem je zakončena podkapitola 3.2, která obsahovala několik vícerozměrných modelů. Tím ale ekonometrická analýza rozhodně nekončí. V další podkapitole se přidá nástroj interakčních proměnných.

3.3 Modely s interakčními proměnnými

Interakční proměnné umožní použít násobek dvou (či více) proměnných. Prvním možným použitím je takové, kdy alespoň jedna z proměnných je binárního charakteru. V našem případě to bude znamenat, že například v různých kontinentech jsou odlišné závislosti určitých proměnných. Regresní koeficient se potom musí interpretovat společně s koeficientem dané proměnné použité bez interakce. Tento postup je názorný na následujícím příkladu.

Analyzován bude vliv hustoty zalidnění. V předešlém textu již bylo řečeno, že pro vysvětlení závislosti se budou hodit logaritmovaná data. Na tom se nic měnit nebude. Uvažujme nyní jednorozměrný model:

$$mira_umrtnosti_i = \beta_0 + \beta_1 log\ (hust_zal)_i + u_i. \tag{3.11}$$

S odhady parametrů b_0 = 21,519 a b_1 = -1,55478. To znamená, že se očekává nižší míra úmrtnosti v zemích s vyšším zalidněním. Model ale nevypovídá nic o rozdílu mezi jednotlivými kontinenty. K tomuto účelu mohou posloužit právě interakční proměnné.

Není například možné, aby existovala jiná závislost na hustotě zalidnění v Evropě oproti zbytku světa? Takovou hypotézu lze ověřit pomocí ekonometrického modelu.

$$mira_umrtnosti_i = \beta_0 + \beta_1 log\ (hust_zal)_i + \beta_2 evropa * log\ (hust_zal)_i + u_i. \tag{3.12}$$

Proměnné	Odhady parametrů	*p*-hodnoty	Koeficient determinace	Korigovaný koeficient determinace
β_0	21,46220	<0,00001	0,302533	0,292045
log(hust_zal)	-0,94114	0,02928		
*evropa*log(hust_zal)*	-1,88384	<0,00001		

Tabulka 3-27, Výsledky modelu 3.12 (zdroj – výpočet v softwaru gretl)

V Evropě tedy očividně existuje jiná síla závislosti na hustotě zalidnění než v ostatních regionech. Se zvyšujícím se zalidněním klesá úmrtnost rychleji. Jestliže se zalidnění zvýší o jedno procento, v evropské zemi klesne úmrtnost přibližně o 0,02285 člověka ze sto tidíc obyvatel, zatímco v ostatních zemích pouze o 0,00941.

Opět bude možné sledovat závislost i z druhé strany (tedy těch regionů, kde síla vlivu zalidnění je naopak nižší).

$$mira_umrtnosti_i = \beta_0 + \beta_1 log\ (hust_zal)_i + \beta_2 afrika * log\ (hust_zal)_i + \beta_3 jiz_amer * \log(hust_{zal})_i + \beta_4 asie * \log(hust_{zal})_i + \beta_5 stred_amer * log\ (hust_zal)_i + u_i. \quad (3.13)$$

Proměnné	Odhady parametrů	*p*-hodnoty	Koeficient determinace	Korigovaný koeficient determinace
β_0	20,07210	<0,00001	0,410919	0,388262
log(hust_zal)	-2,56897	<0,00001		
*afrika*log(hust_zal)*	2,80324	<0,00001		
*jiz_amer*log(hust_zal)*	2,95111	0,00009		
*asie*log(hust_zal)*	1,67041	<0,00001		
*stred_amer*log(hust_zal)*	1,96092	0,00002		

Tabulka 3-28, Výsledky modelu 3.13 (zdroj – výpočet v softwaru gretl)

Výsledky předchozího modelu jsou pozoruhodné. Podle odhadovaných parametrů má totiž v Africe a jižní Americe závislost dokonce opačný směr! Tedy zvyšující se hustota zalidnění přispívá k vyšší úmrtnosti (byť jen nepatrně). Vysvětlení je nasnadě. V méně vyspělých světadílech totiž jistě nebude fungovat podobný systém, který známe ze západního prostředí. Tedy že každé velké město má dobře fungující systém hromadné dopravy. V rozvojových regionech se dá očekávat ve velkých městech spíše chaos, který může často skončit nehodou (a třeba i úmrtím účastníků).

Interpretace koeficientů bude stejná, jako v předchozím příkladě. Pro Evropu, Severní Ameriku a Austrálii a Oceánii platí hodnota parametru β_1, u ostatních regionů je nutné tento parametr sečíst s parametrem příslušné interakční proměnné

Další modely, kde součástí interakční proměnné je proměnná regionální, budou uvedeny již pouze v přehledné tabulce 3-29.

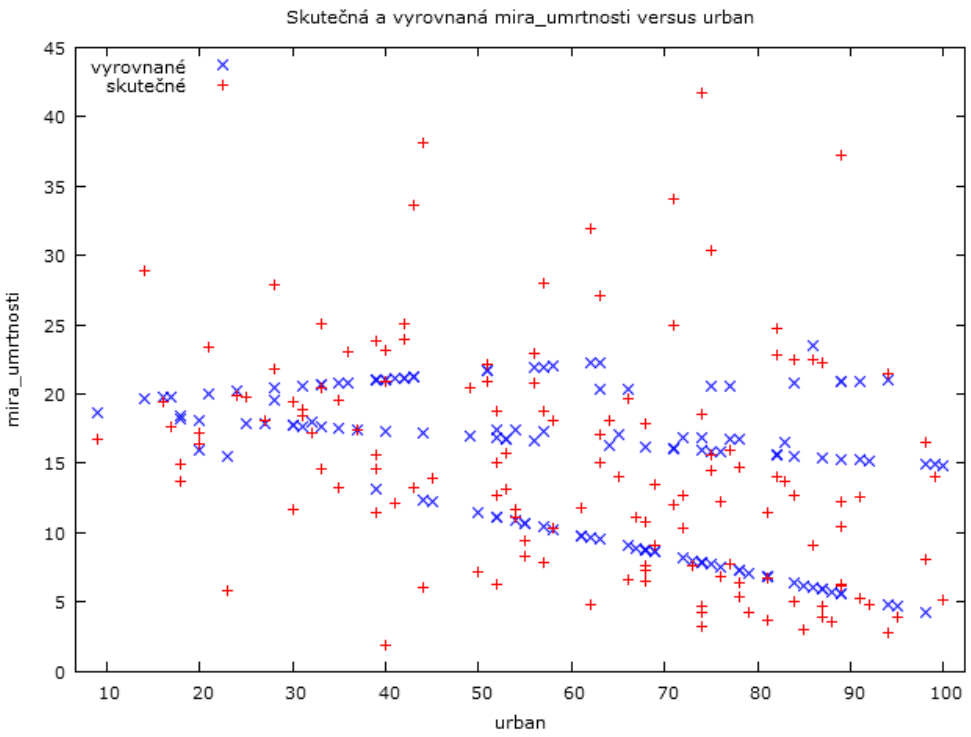
Proměnné	Odhady parametrů	*p*-hodnoty	Koeficient determinace	Korigovaný koeficient determinace
β_0	18,91260	<0,0001	0,436166	0,414480
urban	-0,14955	<0,0001		
*afrika*urban*	0,20296	<0,0001		
*jiz_amer*urban*	0,17155	<0,0001		
*asie*urban*	0,10896	<0,0001		
*stred_amer*urban*	0,12119	0,0003		
β_0	19,03560	<0,0001	0,387893	0,369202
pasy	-10,16610	<0,0001		
*jiz_amer*pasy*	11,60130	<0,0001		
*afrika*pasy*	12,22890	<0,0001		
*asie*pasy*	6,01933	0,0007		
β_0	17,03750	<0,00001	0,201843	0,189840
daln_obyv	-0,42058	<0,00001		
*asie*daln_obyv*	0,41885	<0,00001		

Tabulka 3-29, Modely s interakčními proměnnými (zdroj – výpočet v softwaru gretl)

První model v tabulce, který popisuje vliv urbanizace, koresponduje s modelem, kde byla vysvětlující proměnná hustota zalidnění. Na podobnost mezi těmito dvěma přístupy už bylo upozorňováno, takže výsledky by neměly být překvapující.

Koeficient β_0 představuje takovou úmrtnost, pokud by byla urbanizace nulová (bez rozdílu kontinentů). Proměnná urbanizace je číslo od 0 do 100, přičemž jsou uvažovány pouze diskrétní celočíselné hodnoty. Koeficient u proměnné urbanizace má vztah ke všem regionům, které nejsou uvedeny v interakci (je to tedy Evropa, Severní Amerika a Austrálie a Oceánie). Na těchto třech kontinentech zvyšující se urbanizace znamená nižší úmrtnost. A vliv je celkem znatelný. Pokud se urbanizace zvýší například o 10 bodů, odhaduje se o 1,5 úmrtí na sto tisíc obyvatel méně.

Koeficienty u všech interakčních proměnných jsou kladné a to znamená, že vliv urbanizace je slabší – ba dokonce v Africe a Jižní Americe je vliv opačný (byť nepříliš silný).

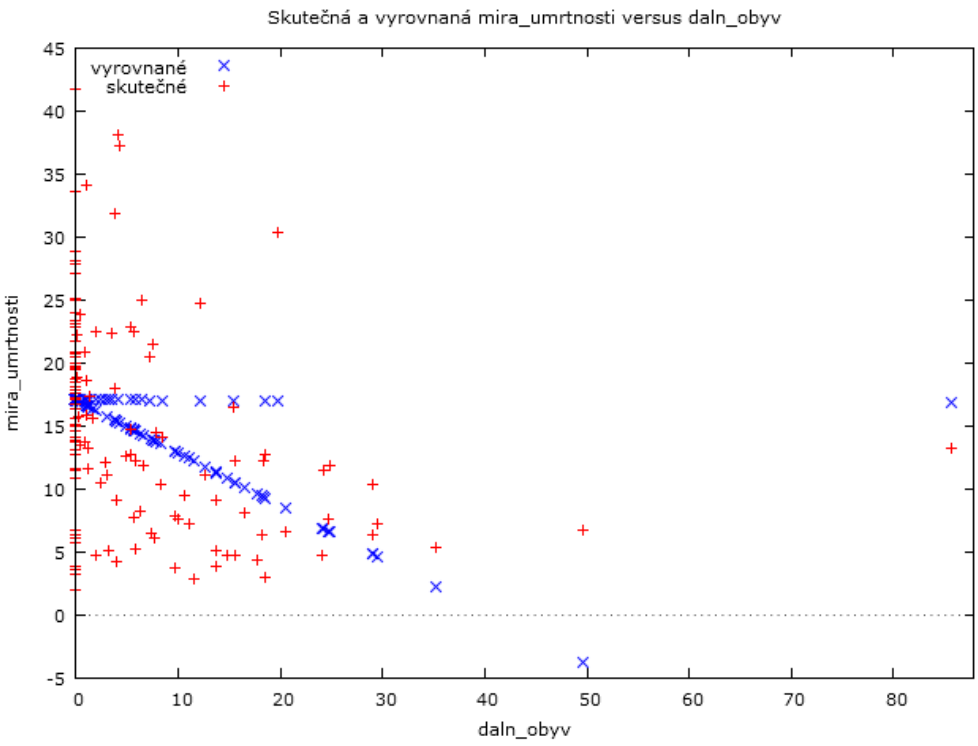


Graf 3-7, Závislost míry úmrtnosti na míře urbanizace v různých regionech (zdroj – gretl)

Celou situaci je možné analyzovat podle grafu 3-7. V modelu jsou státy rozděleny do pěti skupin – čtyři regiony odděleně a zbylé tři tvoří skupinu pátou. Těchto pět skupin je reprezentováno vlastními regresními přímkami (spojnice modrých vyrovnaných hodnot). Ta nejvíce klesající patří Evropě, Severní Americe a Austrálii. Je patrné, že s rostoucí urbanizací klesá v této skupině úmrtnost nejrychleji. Při postupu výše se dospěje nejprve na Asii a vzápětí na Střední Ameriku. Zde je ještě závislost mírně negativní. Naopak rostoucí přímky reprezentují Jižní Ameriku a Afriku.

Druhý model v tabulce analyzuje vlivy proměnné pásy. Jestliže povinnost neexistuje, odhaduje se úmrtnost 19,04. Pokud je povinnost zavedena, síla vlivu se liší dle různých regionů. V Evropě, Severní a Střední Americe a v Austrálii se snižuje úmrtnost o 10,17, což je znatelný rozdíl. Na zbylých třech kontinentech je situace bohužel opět o něco horší. V Asii tato povinnost zachrání alespoň 4,15 lidských životů ze 100 tisíc obyvatel, ale v Jižní Americe a Africe se opět zdá, že pásy žádnou roli nehrají.

Třetí model je zaměřený na délku dálniční sítě na obyvatele. Jednotky této proměnné jsou kilometry dálnice na sto tisíc obyvatel. Tentokrát se nejvýznamněji odlišovala Asie. Zatímco ve zbylých částech světa více dálnic znamená nižší úmrtnost, v Asii je vliv téměř zanedbatelný. Vizuální znázornění poskytuje graf 3-8. Vyrovnané hodnoty zemí z asijského kontinentu skutečně nevykazují závislost na délce dálniční sítě na obyvatele.



Graf 3-8, Závislost míry úmrtnosti na hustotě dálniční cítě v Asii a jinde ve světě (zdroj – gretl)

3.4 Instrumentální proměnné

Závěrečná podkapitola praktické části práce si klade za cíl aplikovat na datový soubor metodu instrumentálních proměnných. To znamená, že bude nutné vycházet od počátku z jiných předpokladů, protože od instrumentálních proměnných se požadují naprosto odlišné vlastnosti, než od obvyklých vysvětlujících proměnných. V první řadě musí být korelace mezi instrumentem a vysvětlovanou proměnnou velmi blízká nule. Prvním úkolem tedy bude vybrat takové proměnné, které například při jednoduché regresi nevykázaly statisticky významný vliv na míru úmrtnosti. Nápomocná může být opět tabulka 3-1.

Bylo vybráno sedm proměnných, jejichž *p*-hodnota přesáhla 0,3, Jsou to tedy takové regresory, které nejsou významné ani na třicetiprocentní hladině významnosti. Všechny jsou uvedeny v následující tabulce 3-30, která obsahuje jednak zmíněné p-hodnoty z jednoduchých regresních modelů a navíc také korelační koeficienty s vysvětlovanou proměnnou, tedy mírou úmrtnosti.

	alk_norm	*alk_prof*	*helma*	*srazky*	*sev_amer*	*stred_amer*	*daln_reg_voz*
p-hodnota	0,6658	0,4253	0,7149	0,9089	0,3031	0,4025	0,4695
korelační koeficient	0,0374	0,0689	0,0316	0,0099	-0,0889	0,0724	-0,0625

Tabulka 3-30, Potenciální instrumentální proměnné a jejich statistiky (zdroj – gretl)

Jak si lze všimnout, čím vyšší byla *p*-hodnota při jednoduché regresi, tím nižší je korelační koeficient s vysvětlovanou proměnnou.

Těchto sedm proměnných se tedy zdá být vhodnými kandidáty na funkci instrumentální proměnné. Tři z nich jsou ale proměnné binární, což nemusí úplně vhodné. Navíc jedna z nich je proměnná *helma*, která nabývá nuly pouze pro Afghánistán a Kiribati. To je i důvod vysoké *p*-hodnoty (a nízkého korelačního koeficientu). Tato proměnná tedy jistě jako instrument použita nebude. Nejvhodněji se jeví proměnná *srazky*.

To byl ale pouze první krok, po němž musí následovat ještě další dva. Druhým úkolem bude nalezení takové vysvětlující proměnné, která má jednak statisticky významný vliv na míru úmrtnosti, a zároveň je korelována s některou z předchozích kandidátů na instrument. Třetím úkolem bude nakonec ekonomická verifikace, což se může zdát jako nepodstatný krok, ale ve skutečnosti je velmi důležitý. Jistě nemá smysl nalézt model s instrumentální proměnnou, který nebude mít logickou ekonomickou interpretaci.

Při druhém kroku mohou být opět nápomocny koeficienty korelace. Není nutné vypisovat zde celou korelační matici, ale spíše okomentovat vybrané výsledky.

Proměnná *alk_norm* má vysoké koeficienty korelace se zbylými dvěma proměnnými týkajícími se alkoholu. To není překvapivé, ale ani vhodné pro vytvoření modelu s instrumentální proměnnou. Dále má poměrně vysoké koeficienty s proměnnou *srazky* a *stred_amer*. To jsou ovšem další potenciální instrumenty. Žádný zajímavý model s instrumentem *alk_nor*m tedy nevznikne.

Podobné výsledky korelačních koeficientů má regresor *alk_prof*. Korelační koeficient s proměnnou *evropa*, která má významný vliv na míru úmrtnosti, byl -0,2357. To není příliš vysoká korelace, nicméně použít tuto dvojici by rozhodně šlo.

Takovýmto způsobem lze odhalit více potenciálních dvojic vhodných pro konstrukci modelu s instrumentální proměnnou. Obdobných dvojic bylo nalezeno celkem sedm a byla provedena analýza. Výsledky shrnuje následující tabulka. V prvním řádku jsou vždy uvedeny výsledky jednoduchého regresního modelu s jedinou vysvětlující proměnnou (*evropa*), v druhém potom model obsahující instrumentální proměnnou. V tabulce jsou uvedeny údaje o velikostech odhadů regresních parametrů, dále potom *p*-hodnota odhadu parametru β_1 a koeficient determinace. Sledujme především, jaký vliv bude mít použití instrumentu na velikost odhadu parametru β_1.

	β_0	β_1	p-hodnota (β_1)	koeficient determinace
evropa	17,83730	-9,42181	<0,0001	0,284196
evropa, instrumentováno *alk_prof*	16,55520	-5,16907	0,3644	0,284196
rychlost_mesto	8,43316	0,12260	0,0186	0,040642
rychlost_mesto, instrumentováno *srazky*	16,04860	-0,01965	0,9093	0,040642
alk_nov	13,08180	47,38130	0,0421	0,030473
alk_nov, instrumentováno *srazky*	14,68650	7,67813	0,9083	0,030473
evropa	17,83730	-9,42181	<0,0001	0,284196
evropa, instrumentováno *srazky*	15,19630	-0,66156	0,9069	0,284196
jiz_amer	14,50400	6,70310	0,0117	0,046533
jiz_amer, instrumentováno *srazky*	14,92240	1,01339	0,9081	0,046533
daln_obyv	16,48310	-0,23357	0,0002	0,101539
daln_obyv, instrumentováno *sev_amer*	16,22630	-0,19321	0,2782	0,101539
rychlost_mesto	8,43316	0,12260	0,0186	0,040642
rychlost_mesto, instrumentováno *sev_amer*	22,79360	-0,14563	0,4355	0,040642

Tabulka 3-31, Modely s instrumentálními proměnnými (zdroj – výpočty v softwaru gretl)

Zajímavé výsledky jsou například u modelů s regionálními proměnnými, kde instrumentem byla proměnná *srazky*. Vliv samotných regionálních proměnných se zdá podle jednoduchých modelů poměrně značný, ale jak je vidět, velkou roli můžou hrát také srážky. Ty sice mají nízkou korelaci s vysvětlovanou proměnnou, ale zřejmě se jejich úroveň na jednotlivých kontinentech liší. Tím pádem jejich přítomnost v modelu ve funkci instrumentu oslabuje vliv regionálních proměnných.

Podkapitolou s instrumentálními proměnnými je tedy završena celá aplikační část této práce. Analyzováno bylo mnoho modelů, ať už jednorozměrných, nebo složitějších. Bylo vyzkoušeno použití interakčních a instrumentálních proměnných. Rovněž byl analyzován problém heteroskedasticity a k vyřešení byla aplikována metoda robustních chyb.

Komentáře k celkovým výsledkům budou uvedeny v závěru.

Závěr

Bakalářská práce se zabývala problematikou úmrtnosti při dopravních nehodách ve vybraných zemích světa. Celkem obsahoval datový soubor 136 zemí a k analýze byly použity různé ekonometrické metody.

První kapitola seznámila s ekonometrickým aparátem, který byl na datový soubor aplikován. Všechny modely v této práci byly odhadnuty metodou nejmenších čtverců, přičemž byly využity různé možnosti aplikace regresní analýzy – logaritmická transformace, kvadratická forma závislosti, technika binárních, interakčních či instrumentálních proměnných.

Ve druhé kapitole byly popsány všechny zdroje dat, přičemž nejvíce údajů pochází z dokumentu WHO [5] analyzujícího bezpečnostní situaci na silnicích po celém světě. Dále se tato kapitola zabývala proměnnými (jak závislou, tak nezávislými), byly nastíněny očekávané vlivy jednotlivých vysvětlujících proměnných na míru úmrtnosti při autonehodách a také byla vyřčena varování před potenciálními problémy při vstupu konkrétních proměnných do ekonometrického modelu.

Závěrečná, třetí kapitola byla stěžejní částí práce, ve které byly na získaná data aplikovány popsané ekonometrické metody. Analyzováno bylo několik desítek modelů, přičemž některé sloužily pro prvotní orientaci (jednoduché regresní modely) pro pozdější práci s vícerozměrnými modely a také s lehce pokročilejším aparátem. Výsledky analýzy by šlo shrnout asi takto:

Ekonometrická analýza prokázala především velké rozdíly mezi zeměmi z různých kontinentů. Relativně dobře jsou na tom řidiči v Evropě, Severní Americe a Austrálii a Oceánii, naopak nejhůře je na tom Afrika. Značné odlišnosti byly zjištěny nejen v úrovni míry úmrtnosti, ale také v síle vlivu určitých proměnných. Například míra urbanizace na třech relativně nejbezpečnějších kontinentech přispívá ke snížení úmrtnosti, vliv je ale bohužel velmi slabý (spíše neprokazatelný) ve zbylých částech světa. Podobně tomu je s délkou dálniční sítě, která rovněž snižuje míru úmrtnosti, nicméně v Asii tento vliv nelze ověřit.

Přes tyto odlišnosti v datovém souboru se vliv určitých proměnných podařilo prokázat. Především se jedná o míru urbanizace, hrubý národní důchod, ukazatel HDI, povinnost být v automobilu připoután, počet registrovaných vozidel a délku dálniční sítě. Všechny tyto proměnné prokazují negativní závislost, tedy že se zvyšujícími se hodnotami klesá očekávaná míra úmrtnosti. U ukazatele HDI poskytuje ještě lepší výsledky model s kvadratickou formou závislosti.

Naopak některé proměnné neprokázaly vliv na míru úmrtnosti, ačkoli jsou veřejností považovány za významné. Jsou jimi maximální povolená rychlost mimo město, maximální povolené množství alkoholu v krvi řidiče (a to u běžných řidičů a profesionálů) a roční úhrn

srážek. Neprokazatelnost vlivu především dvou prvně jmenovaných proměnných na míru úmrtnosti byla poměrně překvapivým výsledkem.

Ačkoli byla snaha zahrnout do datového souboru co možná nejvíce relevantních proměnných, jistě nebyly vyčerpány všechny možnosti, a tak je možné, že některá opomenutá proměnná by případně výsledky celé analýzy pozměnila nebo alespoň obohatila. Tohoto úkolu se může chopit další zájemce, kterého analýza míry úmrtnosti na silničních komunikacích zaujala.

Literatura

[1] HUŠEK, Roman. *Ekonometrická analýza*. Vyd. 1. Praha: Oeconomica, 2007, 367 s. ISBN 978-80-245-1300-3.

[2] KRKOŠKOVÁ, Šárka, Adéla RÁČKOVÁ a Jan ZOUHAR. *Základy ekonometrie v příkladech*. 2., přeprac. vyd. Praha: Oeconomica, 2010, 276 s. ISBN 9788024517087.

[3] SANDO, Wil a Wil SANDO. *Rethinking Human Development*: The UNDP Human Development Report 2010. *Development* [online]. 2010, vol. 53, issue 1, s. 138-139 [cit. 2015-03-14]. DOI: 10.2172/1054929. Dostupný z WWW: http://hdr.undp.org/sites/default/files/reports/270/hdr_2010_en_complete_reprint.pdf

[4] WOOLDRIDGE, Jeffrey M. *Introductory econometrics: a modern approach*. 4th ed. Mason, OH: South Western, Cengage Learning, c2009, xx, 865 p. ISBN 0324660545.

[5] WORLD HEALTH ORGANIZATION. *Global status report on road safety 2013: supporting a decade of action*. Geneva: World Health Organization, [2013], 316 s. ISBN 978-92-4-069062-2. Dostupný z: http://www.who.int/violence_injury_prevention/road_safety_status/2013/en/

Internetové zdroje:

[6] THE WORLD BANK GROUP. Average precipitation in depth (mm per year). *data.worldbank.org* [online databáze]. ©2015 [cit. 2015-03-10]. Dostupné z: http://data.worldbank.org/indicator/AG.LND.PRCP.MM

[7] THE WORLD BANK GROUP. Urban population (% of Total). *data.worldbank.org* [online databáze]. ©2015 [cit. 2015-03-08]. Dostupné z: http://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS

[8] CENTRAL INTELLIGENCE AGENCY. The World Factbook. *www.cia.gov* [online databáze]. ©2015 [cit. 2015-03-09]. Dostupné z: https://www.cia.gov/library/publications/resources/the-world-factbook/rankorder/2147rank.html

[9] NATIONMASTER.COM. Motorway lenght. *nationmaster.com* [online databáze]. ©2003-2015 [cit. 2015-03-15]. Dostupné z:

http://www.nationmaster.com/country-info/stats/Transport/Road/Motorway-length:

Přílohy

Model 3.4

	Koeficient	*Směr. chyba*	*t-podíl*	*p-hodnota*	
const	11,1069	1,46874	7,5621	<0,00001	***
afrika	11,8118	1,51247	7,8096	<0,00001	***
jiz_amer	12,7268	2,15044	5,9182	<0,00001	***
asie	7,01583	1,44874	4,8427	<0,00001	***
stred_amer	7,23568	2,06612	3,5021	0,00063	***
pasy	-2,9184	1,27614	-2,2869	0,02382	**

Střední hodnota závisle proměnné	14,99687	Sm. odchylka závisle proměnné	8,140346
Součet čtverců reziduí	4992,174	Sm. chyba regrese	6,196881
Koeficient determinace	0,441954	Adjustovaný koeficient determinace	0,420490
F(5, 130)	20,59112	P-hodnota(F)	4,19e-15
Logaritmus věrohodnosti	-437,9777	Akaikovo kritérium	887,9554
Schwarzovo kritérium	905,4314	Hannan-Quinnovo kritétium	895,0572

Model 3.5

	Koeficient	*Směr. chyba*	*t-podíl*	*p-hodnota*	
const	16,3185	2,27291	7,1796	<0,00001	***
afrika	10,9553	1,49816	7,3125	<0,00001	***
jiz_amer	11,0563	2,16517	5,1064	<0,00001	***
asie	7,66382	1,42484	5,3787	<0,00001	***
stred_amer	7,62781	2,01208	3,7910	0,00023	***
pasy	-3,08888	1,24139	-2,4882	0,01411	**
lhust_zal	-1,19001	0,403918	-2,9462	0,00382	***

Střední hodnota závisle proměnné	14,99687	Sm. odchylka závisle proměnné	8,140346
Součet čtverců reziduí	4677,449	Sm. chyba regrese	6,021569
Koeficient determinace	0,477135	Adjustovaný koeficient determinace	0,452816
F(6, 129)	19,61960	P-hodnota(F)	3,52e-16
Logaritmus věrohodnosti	-433,5497	Akaikovo kritérium	881,0993
Schwarzovo kritérium	901,4879	Hannan-Quinnovo kritétium	889,3847

Model 3.6

	Koeficient	*Směr. chyba*	*t-podíl*	*p-hodnota*	
const	14,3241	1,62872	8,7947	<0,00001	***
afrika	8,84537	1,63262	5,4179	<0,00001	***
jiz_amer	10,3499	2,13663	4,8441	<0,00001	***
asie	5,33666	1,44544	3,6921	0,00033	***
stred_amer	4,96893	2,05158	2,4220	0,01683	**
pasy	-2,81119	1,21391	-2,3158	0,02215	**
gni	-0,000123808	3,22391e-05	-3,8403	0,00019	***

Střední hodnota závisle proměnné	14,99687	Sm. odchylka závisle proměnné	8,140346
Součet čtverců reziduí	4479,995	Sm. chyba regrese	5,893101
Koeficient determinace	0,499207	Adjustovaný koeficient determinace	0,475915
F(6, 129)	21,43193	P-hodnota(F)	2,38e-17
Logaritmus věrohodnosti	-430,6168	Akaikovo kritérium	875,2335
Schwarzovo kritérium	895,6221	Hannan-Quinnovo kritétium	883,5189

Model 3.7

	Koeficient	*Směr. chyba*	*t-podíl*	*p-hodnota*	
const	0,16895	6,6383	0,0255	0,97974	
afrika	7,66015	2,18639	3,5036	0,00063	***
jiz_amer	9,57589	2,12378	4,5089	0,00001	***
asie	4,2044	1,50374	2,7960	0,00597	***
stred_amer	4,52168	2,01779	2,2409	0,02676	**
pasy	-2,4613	1,20937	-2,0352	0,04390	**
hdi	63,0334	20,4175	3,0872	0,00248	***
hdi2	-61,3968	16,1733	-3,7962	0,00023	***

Střední hodnota závisle proměnné	14,99687	Sm. odchylka závisle proměnné	8,140346
Součet čtverců reziduí	4285,400	Sm. chyba regrese	5,786163
Koeficient determinace	0,520960	Adjustovaný koeficient determinace	0,494762
F(7, 128)	19,88586	P-hodnota(F)	7,49e-18
Logaritmus věrohodnosti	-427,5970	Akaikovo kritérium	871,1940
Schwarzovo kritérium	894,4953	Hannan-Quinnovo kritétium	880,6630

Model 3.8

	Koeficient	*Směr. chyba*	*t-podíl*	*p-hodnota*	
const	12,7007	1,60936	7,8918	<0,00001	***
afrika	12,0127	1,49186	8,0522	<0,00001	***
jiz_amer	14,5864	2,27201	6,4200	<0,00001	***
asie	7,29619	1,43185	5,0956	<0,00001	***
stred_amer	9,08683	2,19345	4,1427	0,00006	***
pasy	-2,88593	1,25659	-2,2966	0,02325	**
srazky	-0,00181502	0,000804161	-2,2570	0,02569	**

Střední hodnota závisle proměnné	14,99687	Sm. odchylka závisle proměnné	8,140346
Součet čtverců reziduí	4802,523	Sm. chyba regrese	6,101546
Koeficient determinace	0,463154	Adjustovaný koeficient determinace	0,438184
F(6, 129)	18,54870	P-hodnota(F)	1,82e-15
Logaritmus věrohodnosti	-435,3441	Akaikovo kritérium	884,6882
Schwarzovo kritérium	905,0768	Hannan-Quinnovo kritétium	892,9736

Model 3.9

	Koeficient	*Směr. chyba*	*t-podíl*	*p-hodnota*	
const	2,77362	6,53701	0,4243	0,67207	
afrika	7,43313	2,13244	3,4857	0,00067	***
jiz_amer	11,5587	2,18889	5,2806	<0,00001	***
asie	4,36096	1,46664	2,9734	0,00352	***
stred_amer	6,52579	2,09409	3,1163	0,00227	***
pasy	-2,36828	1,17914	-2,0085	0,04671	**
hdi	62,5629	19,8998	3,1439	0,00208	***
hdi2	-62,0818	15,7646	-3,9381	0,00013	***
srazky	-0,00208072	0,000747134	-2,7849	0,00617	***

Střední hodnota závisle proměnné	14,99687	Sm. odchylka závisle proměnné	8,140346
Součet čtverců reziduí	4038,755	Sm. chyba regrese	5,639257
Koeficient determinace	0,548531	Adjustovaný koeficient determinace	0,520092
F(8, 127)	19,28799	P-hodnota(F)	9,41e-19
Logaritmus věrohodnosti	-423,5661	Akaikovo kritérium	865,1323
Schwarzovo kritérium	891,3462	Hannan-Quinnovo kritétium	875,7849

Model 3.10

	Koeficient	*Směr. chyba*	*t-podíl*	*p-hodnota*	
const	5,84588	6,48537	0,9014	0,36910	
afrika	7,04975	2,08721	3,3776	0,00097	***
jiz_amer	9,89262	2,22592	4,4443	0,00002	***
asie	4,96489	1,44976	3,4246	0,00083	***
stred_amer	6,58425	2,04498	3,2197	0,00163	***
pasy	-2,56554	1,15377	-2,2236	0,02796	**
hdi	64,4773	19,4451	3,3159	0,00119	***
hdi2	-62,7676	15,3961	-4,0769	0,00008	***
srazky	-0,00175377	0,000739691	-2,3709	0,01926	**
lhust_zal	-1,00676	0,375496	-2,6811	0,00832	***

Střední hodnota závisle proměnné	14,99687	Sm. odchylka závisle proměnné	8,140346
Součet čtverců reziduí	3820,772	Sm. chyba regrese	5,506686
Koeficient determinace	0,572898	Adjustovaný koeficient determinace	0,542391
F(9, 126)	18,77905	P-hodnota(F)	1,51e-19
Logaritmus věrohodnosti	-419,7932	Akaikovo kritérium	859,5865
Schwarzovo kritérium	888,7130	Hannan-Quinnovo kritétium	871,4228

Model 3.13

	Koeficient	*Směr. chyba*	*t-podíl*	*p-hodnota*	
const	22,0253	1,90066	11,5882	<0,00001	***
lhust_zal	-2,76075	0,479167	-5,7616	<0,00001	***
afr_lhust_zal	2,51656	0,389268	6,4649	<0,00001	***
asie_lhust_zal	1,49636	0,312575	4,7872	<0,00001	***
st_am_lhust_zal	1,74348	0,463644	3,7604	0,00025	***

Střední hodnota závisle proměnné	14,99687	Sm. odchylka závisle proměnné	8,140346
Součet čtverců reziduí	5929,964	Sm. chyba regrese	6,728068
Koeficient determinace	0,337124	Adjustovaný koeficient determinace	0,316883
F(4, 131)	16,65589	P-hodnota(F)	4,65e-11
Logaritmus věrohodnosti	-449,6837	Akaikovo kritérium	909,3674
Schwarzovo kritérium	923,9307	Hannan-Quinnovo kritétium	915,2856