VYSOKÁ ŠKOLA EKONOMICKÁ V PRAZE

FAKULTA INFORMATIKY A STATISTIKY

BAKALÁŘSKÁ PRÁCE

2016 Radim Nešpor

VYSOKÁ ŠKOLA EKONOMICKÁ V PRAZE

FAKULTA INFORMATIKY A STATISTIKY

Název bakalářské práce:

Ekonometrická analýza ceny mobilních telefonů

Autor:	**Radim Nešpor**
Katedra:	**Katedra ekonometrie**
Obor:	**Statistika a ekonometrie**
Vedoucí práce:	**Ing. Tomáš Formánek, Ph.D.**

Prohlášení:

Prohlašuji, že jsem bakalářskou práci na téma „Ekonometrická analýza ceny mobilních telefonů“ zpracoval samostatně. Veškerou použitou literaturu a další podkladové materiály uvádím v seznamu použité literatury.

V Praze dne 30.5. 2016 .....

Radim Nešpor

Poděkování:

Rád bych poděkoval Ing. Tomáši Formánkovi, Ph.D. za cenné rady a připomínky při zpracování této práce a za trpělivost, kterou se mnou měl.

Abstrakt

Práce se zabývá ekonometrickou analýzou ceny mobilních telefonů na základě vybraných parametrů. Cílem této práce je zjistit, které z vybraných parametrů mají vliv na výslednou cenu nového mobilního telefonu a jakým způsobem ji ovlivňují. V teoretické části se nejdříve zabývám regresním modelem a určováním statisticky významných proměnných a významnosti modelu jako celku. Dále pak představím různé postupy pro hledání a výběr nejlepšího odhadu modelu. V této práci se zaměřím na tři postupy výběru vysvětlujících proměnných, a to: výběr nejlepší podmnožiny, postupný výběr vpřed a postupný výběr zpět. Nakonec se ještě zmíním o zjišťování multikolinearity a testování heteroskedasticity. V druhé části pak aplikuji jednotlivé postupy výběru vysvětlujících proměnných a pro každý z nich, na základě kritérií stanovených v první části, vyberu nejvhodnější model. Vybrané modely nakonec porovnám mezi sebou a určím ten nejlepší z nich.

Klíčová slova: ekonometrická analýza, regresní model, výběr proměnných, mobilní telefon

Abstract

The work deals with the econometric analysis of the prices of mobile phones based on the selected parameters. The aim of this thesis is to determine, which of the selected parameters affect the price of a new mobile phone and how they affects the price. The theoretical part deals at first with the regression model and determining statistically significant variables and significance of the model as a whole. Then I introduce different procedures for finding and selecting the best estimate for the model. In this thesis I will focus on three procedures for the selection of explanatory variables: best subset selection, forward stepwise selection and backward stepwise selection. Finally I mention the detection of multicollinearity and heteroscedasticity testing. In the second part, I apply different procedures for the selection of explanatory variables and afterwards, based on the criteria defined in the first part, I will select the best model. Finally, I compare these models and choose the best one.

Keywords: econometric analysis, regression model, selection of the variables, mobile phone

Obsah

1 Úvod

Časy, kdy se mobilní telefony používaly pouze k telefonování, jsou dávno pryč a v posledních letech se díky technologickému pokroku stávají stále všestrannějšími produkty. Mobilní telefony jsou v dnešní době naprosto běžnou součástí našeho života. Dalo by se říci, že vlastnit mobilní telefon je dnes téměř nutnost. Na trhu je jich k dispozici nepřeberné množství s velkým počtem technických i netechnických parametrů. Ve své práci se pokusím o ekonometrickou analýzu závislosti ceny mobilních telefonů na vybraných technických parametrech.

Pro tuto práci jsem zvolil devět vysvětlujících proměnných, které významně ovlivňují výslednou cenu telefonu. Vysvětlovanou proměnnou je pak samotná cena mobilních telefonů. Data jsem čerpal z webu www.mobilmania.cz a z internetového obchodu www.czc.cz. Data jsou považována za průřezová data a byla získána v jednom okamžiku, nejedná se o časovou řadu. Pro tuto práci jsem použil 70 mobilních telefonů různých značek a cenových kategorií (jedná se o skromný vzorek z momentální velmi rozsáhlé nabídky). Do výběru jsem zařadil jak telefony známých značek, jako Samsung, či Apple, tak i méně známých jako např. CAT, Zopo a jiné. Vybíral jsem také telefony všech cenových kategorií, aby byl vzorek reprezentativní a zahrnul co nejširší nabídku. Cenový rozsah mého výběru je od 1 700 Kč do 20 000 Kč.

2 Vývoj cen mobilních telefonů

První opravdu mobilní telefon, který spatřil svět, byl Motorola DynaTAC 8000X a bylo to roku 1983. Prototyp telefonu byl představen už roku 1973, ale kvůli nezbytným úpravám a vybudováním sítě byl vypuštěn na trh až o 10 let později. Rozměry telefonu byly 33 × 4,3 × 8,9 cm, vážil necelých 800g a jeho cena byla 3 995 USD (přibližně 100 000 Kč). Baterie vydržela na jedno nabití přibližně 30 min hovoru, popř. 8 hodin pohotovostního stavu. I přes vysokou zaváděcí cenu byl o tuto novinku obrovský zájem. Pro porovnání např. iPhone první generace, představený roku 2007 měl rozměry 115 × 61 × 11,6 mm, vážil 135g, baterie měla vydržet až 8 hodin hovoru, nebo 250 hodin pohotovostního režimu a cena byla 399 USD (přibližně 10 000Kč).

Obrázek 1: Motorola DynaTAC 8000X

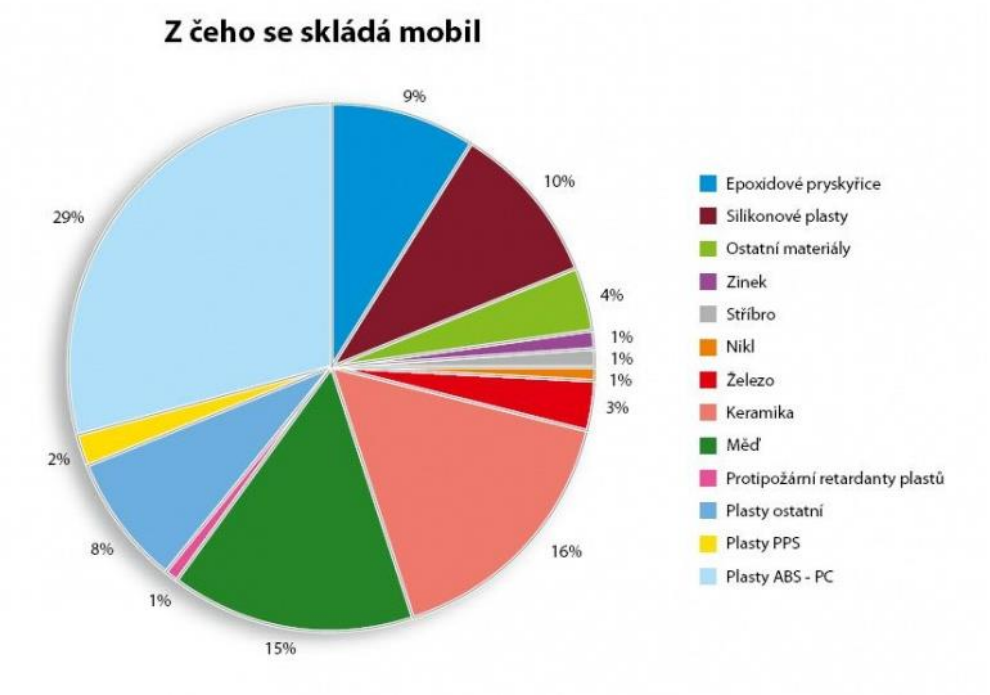
Zdroj: http://cdr.cz/

Roku 1989 Motorola uvedla na trh nový druh mobilních telefonů s odklopným krytem, tzv. flipem. Model MicroTAC už byl o více než polovinu lehčí než model z roku 1983 a základní verze stála 495 USD. Byl to první kapesní telefon a poprvé se začalo uvažovat o mobilních telefonech jako o kapesním zařízení. Dalším revolučním telefonem byla Motorola StarTAC z roku 1996, která vážila pouhých 88g a měla rozměry 95 × 57 × 23 mm, její cena ovšem činila 1 000 USD. Až do této doby uměly všechny telefony pouze volat, v nadstandartních případech i číst SMS zprávy a displeje těchto telefonů se podobaly spíše dnešním kalkulačkám, než mobilům, jak je známe.

Jedním z prvních mobilů, který se už více podobá těm dnešním, byla Nokia 3210. Ta se na trhu objevila roku 1999 a její cena byla, v porovnání s výše zmíněnými, pouze 149,99 liber. Byl to jeden z prvních modelů, který obsahoval skrytou anténu. Nokia prodala 160 milionů kusů tohoto modelu. Můžeme tedy sledovat, že se postupem času cena mobilních telefonů snižuje a ty se tak stávají stále dostupnější i široké veřejnosti. Například v roce 1996

cena nových mobilních telefonů na českém trhu začínala na částkách okolo 10 000 Kč, což bylo více než průměrná hrubá měsíční mzda v té době. Dnes máme možnost zakoupit nový telefon za podstatně nižší částku.

Na cenu mobilních telefonů mají vliv rovněž materiály, z jakých jsou vyrobeny. Jako hlavní materiál, který se používá, je plast. Jak ukazuje následující graf, až 50% mobilu se skládá z plastu, ale podstatnou část tvoří například měď a v nižším množství se nachází i drahé kovy jako stříbro, zlato a palladium. I když se v jednotlivých kusech nachází jen malé množství těchto drahých kovů, v roce 2008 se na výrobu 1,126 miliardy nových telefonů spotřebovalo mimo jiné 280 tun stříbra, 27 tun zlata a 10 tun palladia.



Obrázek 2: Z čeho se skládá mobil

Zdroj: https://www.enviwiki.cz/

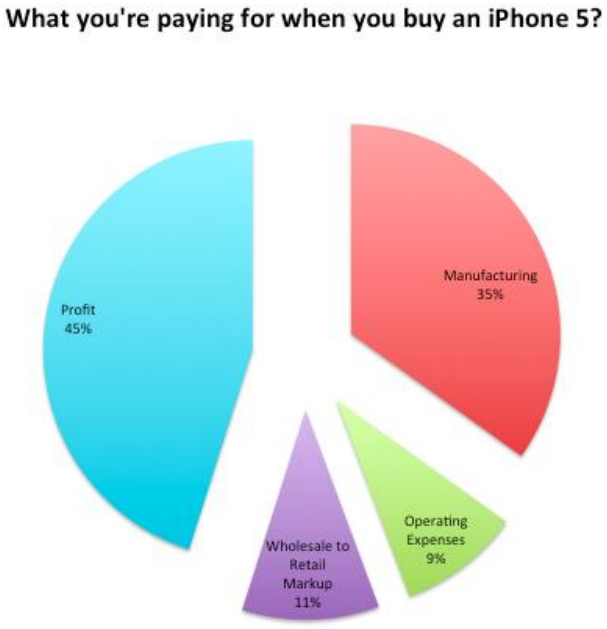
Zajímavým tématem je také výrobce, protože, stejně jako u většiny všech výrobků, za značku se platí. Do roku 2011 byla jednoznačně největším producentem mobilních telefonů Nokia, která ještě v roce 2011 prodala přes 422 milionů kusů, což bylo 23,8% z celého trhu. Od té doby ale pozice Nokie slábla a už v roce 2012 ji na pomyslné první příčce vystřídal Samsung. Poslední roky zůstává největším výrobcem právě Samsung a na druhém místě je Apple. Následující tabulka ukazuje pět největších výrobců a jejich pozici na trhu za roky 2015 a 2014.

Company	2015 Units	2015 Market Share (%)	2014 Units	2014 Market Share (%)
Samsung	320,219.7	22.5	307,596.9	24.7
Apple	225,850.6	15.9	191,425.8	15.4
Huawei	104,094.7	7.3	68,080.7	5.5
Lenovo/Motorola	72,748.2	5.1	81,415.8	6.5
Xiaomi	65,618.6	4.6	56,529.3	4.5
Others	635,368.5	44.6	539,691.3	43.4
Total	1,423,900.3	100.0	1,244,739.8	100.0

Tabulka 2.1: Worldwide Smartphone Sales to End Users by Vendor in 2015 (Thousands of Units)

Zdroj: http://www.gartner.com/

Jakou hodnotu má taková značka, nelze úplně snadno vyčíslit. Například nový iPhone 5 s pamětí 16 GB stál 649,99 USD. Ovšem pouze 226,85 USD stojí jednotlivé komponenty a materiály potřebné k výrobě jednoho kusu. Tato částka je dokonce včetně výrobních nákladů a licenčních poplatků. Dále bychom měli rovněž zahrnout další typy nákladů jako výzkum a vývoj, prodejní, všeobecné a administrativní a v neposlední řadě maloobchodní marže. Tyto náklady jsou odhadnuty následovně: výzkum a vývoj 18,85 USD na kus, prodejní, všeobecné a administrativní náklady 42,33 USD na kus a marže 68,90 USD za kus. Pokud náklady sečteme, obdržíme celkové náklady na výrobu a prodej jednoho iPhonu 5 v hodnotě 356,93 USD. Z toho vyplývá, že profit společnosti Apple je 293,06 USD za kus, což je téměř 50%, jak můžeme vidět v následujícím grafu.



Obrázek 3: What you're paying for when you buy an iPhone 5?

Zdroj: http://www.digitaltrends.com/

Porovnáním mobilních telefonů různých značek, například Motorola a Apple, zjistíme, že si za značku Apple připlatíme víc. Porovnejme výše zmíněný iPhone 5 s Moto G od Motoroly. Nový iPhone 5 stál 649,99 USD v základní verzi s 16GB pamětí, zatímco podobná Moto G stála pouhých 179 USD. V prvé řadě je třeba zahrnout součástky, ze kterých se oba

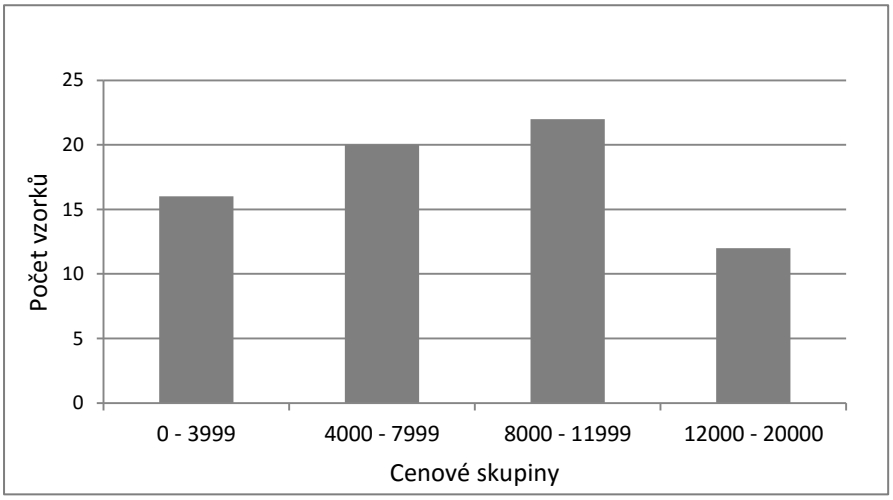
telefony vyrábí. Rozbor těchto komponent prokázal, že celková cena komponent telefonu iPhone 5 je téměř dvakrát vyšší než u Moto G. Při porovnání cen jednotlivých dílů zjistíme, že kromě baterie jsou všechny díly iPhone 5 dražší než u Moto G. Celková cena komponent je tedy 109,75 USD u Moto G a 198,85 USD u iPhonu 5. Tady je poprvé zřetelný vliv značky na cenu. Zatímco maloobchodní cena Moto G je pouze o 73 USD vyšší než suma cen jednotlivých komponent, finální cena iPhonu 5 je vyšší o 450 USD. Jak jsme již vysvětlili, do výsledné ceny musíme zahrnout i další náklady. Ale i kdybychom u Motoroly už další náklady nezapočítávali, byl by profit z jednoho kusu přibližně 41%. Už tento podíl je menší než u Applu a jeho skutečná hodnota bude ještě o poznání menší.

Dalším významným vlivem, který se promítne do výsledné ceny, je také strategie společnosti, postavení společnosti na trhu a způsob, jakým se snaží vydělávat peníze. Apple se snaží vydělat peníze na hardwaru, což je klíčový důvod, proč jsou jeho produkty dražší. Proto má Apple vysokou marži na telefonech i dalších zařízeních, neboť to je hlavní zdroj jeho příjmů. Stejnou strategii volí např. i Samsung. Rozdíl mezi Samsungem a Applem je v tom, že Samsung se zaměřuje na širší spektrum zákazníků a tím pádem se zajímá hlavně o celkové množství prodaných zařízení. Proto vyrábí jak luxusní telefony, tak i telefony cenově dostupnější, a nemá na jednotlivých kusech takovou marži, jako právě Apple. Rozdílnou strategii má např. právě Motorola nebo Google. Google se snaží svými telefony Nexus rozšířit operační systém android mezi co nejvíce uživatelů, protože jeho zisky neplynou hlavně z prodeje jednotlivých zařízení, ale zejména z reklam, vyhledávání a dalších služeb, které poskytuje. Strategie, kterou praktikuje Google, se zaměřuje na dlouhodobou spolupráci s jednotlivými zákazníky a tím pádem i dlouhodobý výnos. Proto se snaží prodávat svá zařízení tak levně, jak je to možné.

3 Faktory ovlivňující cenu mobilních telefonů

3.1 Data

Data jsem získal na internetových serverech www.mobilmania.cz a www.czc.cz 12. 12. 2014. Pro svou analýzu jsem využil 70 mobilních telefonů různých značek a parametrů. Jednotlivé vzorky byly vybrány náhodně tak, aby bylo zahrnuto co nejširší cenové spektrum a aby v každé cenové kategorii byl podobný počet vzorků. Jediným kritériem pro výběr pak bylo, aby se u uvedeného vzorku nacházely všechny potřebné informace. Ačkoliv se jednalo o jediné kritérium, většina telefonů ho nesplňovala. U mnohých chyběla alespoň část informací, které byly nezbytné k mé analýze, a i u některých vzorků, které jsem použil, jsem musel některé údaje dohledávat buď přímo na stránkách výrobce, nebo na internetových fórech k jednotlivým modelům. Obrázek 4 znázorňuje četnosti vzorků v jednotlivých cenových skupinách.



Obrázek 4: Četnosti vzorků v jednotlivých skupinách

Zdroj: autor

3.2 Proměnné

Do regresního modelu jsem zahrnul deset proměnných, z nichž dvě jsou definovány jako tzv. dummy proměnné. To znamená, že nabývají pouze hodnot jedna, pokud je daná skutečnost pravda, a v ostatních případech má hodnotu nula. Ostatní jsou kvantitativní proměnné.

Jednotlivé proměnné jsou:

- cena
- paměť
- RAM

- jemnost displeje
- úhlopříčka
- výdrž
- počet jader
- fotoaparát
- android
- iPhone

V následující tabulce jsou uvedeny základní charakteristiky jednotlivých proměnných. Z těchto charakteristik je pak zřetelné, že do analýzy jsou zahrnuty vzorky z nejrůznějších nejen cenových, ale i výkonnostních kategorií. U dummy proměnných je uveden pouze počet telefonů, které nabývají hodnotu jedna.

Proměnná	Průměr	Medián	Min	Max	Jednotky	Počet
Cena	7684,29	7000	1700	20000	Kč	
Paměť	15,60	16	4	64	GB	
RAM	1581,71	1536	128	3072	MB	
Jemnost displeje	324,97	312	165	534	PPI	
Úhlopříčka	4,84	5	4	6	palce	
Výdrž	2319,89	2200	1150	4000	mAh	
Počet jader	4,23	4	2	8	kusy	
Fotoaparát	10,84	11	2	21	Mpix	
Android						56
iPhone						5

Tabulka 3.1: Základní charakteristiky proměnných

Zdroj: autor

3.2.1 Vysvětlovaná proměnná

Vysvětlovanou proměnnou, jak už bylo zmíněno, je „cena“ jednotlivých telefonů. Ne každý si může dovolit vybírat jen podle vlastností a parametrů a obvykle je hlavní kritérium, podle kterého se nakonec rozhodujeme při koupi nového telefonu právě cena. Cena bude v celé práci vyjádřena v Kč a je aktuální ke konci roku 2014.

3.2.2 Vysvětlující proměnné

Ostatní proměnné jsou vysvětlující neboli nezávislé proměnné, budu se pomocí nich snažit popsat variabilitu vysvětlované proměnné. Většina těchto proměnných je podle mého názoru důležitá pro výkon a celkový dojem z telefonu a s jejich rostoucími hodnotami očekávám, že se bude zvyšovat i cena. V následující části se ke každé z vysvětlujících proměnných vyjádřím podrobněji.

Paměť

Paměť telefonu značí, kolik dat je možno do něho uschovat a pro většinu lidí to bývá jedno z hlavních kritérií. Paměťová karta bývá obvykle jednou z nejdražších komponent.

RAM

Jedná se o operační paměť, na které závisí rychlost celého systému telefonu. Také je na ní závislí multitasking, možnost přepínat mezi aplikacemi a pokračovat v místě, kde se skončilo.

Jemnost displeje

Tato hodnota udává kolik bodů je displej schopen zobrazit na jednom palci. Je to jeden z faktorů vypovídajících o kvalitě displeje, který bývá obdobně jako paměťová karta jednou z nejdražších částí telefonů.

Úhlopříčka

Udává velikost displeje. Obecně platí, že větší displeje jsou dražší.

Výdrž

Výdrž telefonu je udávána v mAh (=miliampere-hour). Značí kapacitu baterie, resp. jak dlouho vydrží telefon v pohotovosti.

Počet jader

Udává počet jader v procesoru. Ačkoliv výkon procesoru mimo jiné nezávisí pouze na jejich počtu, ale také na typu jednotlivých jader, z důvodu velkého množství různých druhů jsem se omezil pouze na jejich počet. Efekt většího počtu je takový, že nám pomůžou udržet v chodu větší počet aplikací, bez výrazného zpomalení telefonu.

Fotoaparát

Udává jaké rozlišení má zadní fotoaparát. Pokud telefon obsahuje i přední, není jeho hodnota do této proměnné zahrnuta.

Android (A)

První dummy proměnná, která má hodnotu jedna v případě, že telefon využívá operační systém android. Pokud tomu tak není, hodnota této proměnné je nula.

iPhone

Druhá dummy proměnná nám říká, zda se jedná o telefon značky Apple, hodnota se rovná jedné pokud je telefon od značky Apple, nebo se hodnota rovná nule pokud tomu tak není. Tuto proměnnou jsem vybral, protože velmi velký vliv na cenu jednotlivých telefonů má mimo jiné právě její značka. Značek je ovšem obrovské množství a zahrnout je všechny by bylo problematické pro řešení. Telefony značky Apple mají ovšem obvykle v porovnání s telefony stejné kategorie nižší hodnoty všech sledovaných technických parametrů. V tabulce 3.2 je právě takové porovnání.

Název	Cena	Paměť	RAM	Jemnost Displeje	Úhlopříčka	Výdrž	Počet Jader	A	Fotoaparát	iPhone
	[Kč]	[GB]	[MB]	[PPI]	[']	[mAh]	[ks]		[Mpx]	
Samsung Galaxy Note 4	18500	32	3072	515	5,7	3220	4	1	16	0
Apple iPhone 6 16GB	18400	16	1024	326	4,7	1810	2	0	8	1

Tabulka 3.2: Porovnání vybraných vzorků

Zdroj: autor

4 Výběr vysvětlujících proměnných regresního modelu – kritéria a metody

4.1 Regresní model

Regresním modelem se snažíme pomocí koeficientů popsat vliv vysvětlujících proměnných na hodnoty vysvětlovaných proměnných. Vícenásobný lineární regresní model vypadá takto:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u, \text{ pro } k = 1, 2, \ldots, j \qquad 4.1$$

kde:

y je vysvětlovaná proměnná,

β_0 je konstanta,

β_j je *j*-tý regresní koeficient *j*-té vysvětlující proměnné x_j

u je náhodná složka a

k představuje počet vysvětlujících proměnných.

Hodnoty koeficientů toho modelu jsou neznámé, a proto se odhadují. Pro takovýto odhad se používá metoda nejmenších čtverců, jejímž cílem je minimalizovat součty čtverců reziduí, resp. odchylek vyrovnaných hodnot závislé na pozorovaných hodnotách. Odhad regresního modelu vypadá následovně:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k, \text{ pro } k = 1, 2, \ldots, j, \qquad 4.2$$

$\hat{y}$ je vyrovnaná hodnota *y*,

$\hat{\beta}_0$ je bodový odhad konstanty β_0,

$\hat{\beta}_j$ je *j*-tý bodový odhad neznámých parametrů β_j.

U takových to regresních modelů se pomocí testů určuje, zda jsou jednotlivé vysvětlující proměnné statisticky významné, či nevýznamné. Existuje pak několik postupů, podle kterých se pak model upravuje, aby jeho konečná verze obsahovala jen proměnné, které jsou statisticky významné, tzn., že zbylé vysvětlující proměnné statisticky významně ovlivňují vysvětlovanou proměnnou *y*. Nesmíme také zapomenout na celkový F-test, který zhodnotí vhodnost modelu jako celku. F-testem testujeme, jestli vysvětlovaná proměnná

závisí na lineární kombinaci vysvětlujících proměnných, které jsme zahrnuli do našeho modelu. Pokud bychom v F-test nezamítli nulovou hypotézu, museli bychom vybrat novou množinu vysvětlujících proměnných. [1] [2]

4.2 Určování statisticky významných a nevýznamných vysvětlujících proměnných

Určit, jaké proměnné jsou pro model statisticky významné a jaké ne, je zásadní pro celou analýzu. Bez ohledu na to, jakým způsobem se finální model sestrojí, určování významnosti proměnných je vždy stejné, používá se k tomu tzv. t-test.

Nulová hypotéza t-testu říká, že není žádný lineární statisticky významný vztah mezi x_k a y, oproti tomu alternativní hypotéza tvrdí, že vztah mezi x_k a y je, tudíž je proměnná statisticky významná. Můžeme to zapsat následovně:

$$H_0: \beta_k = 0, \text{ pro } k = 1, 2, \ldots, j, \tag{4.3}$$

$$H_a: \beta_k \neq 0, \text{ pro } k = 1, 2, \ldots, j. \tag{4.4}$$

Pokud platí $\beta_k = 0$, pak proměnná x_k není statisticky významná a z modelu ji můžeme vyřadit. Pro určení, zda můžeme nulovou hypotézu odmítnout nebo nikoliv, používáme t-hodnotu. T-hodnota se spočítá jako poměr odhadu koeficientu vysvětlující proměnné a odhadu její střední chyby:

$$t_k = \hat{\beta}_k / se(\hat{\beta}_k). \tag{4.5}$$

T-hodnotu pak porovnáme s tabulkovou kritickou hodnotou Studentova t-rozdělení s hladinou významnosti α a $n - j$ stupni volnosti.

$$t^* \sim t_{\alpha/2,(n-j)} \tag{4.6}$$

[1] WOOLDRIDGE, Jeffrey M. *Introductory econometrics: a modern approach*. 4th ed. Mason, OH: South Western, Cengage Learning, c2009.,s. 71-74. ISBN 0324660545.

[2] JAMES, Gareth, Daniela WITTEN, Trevor HASTIE a Robert TIBSHIRANI.*An introduction to statistical learning: with applications in R*. Springer texts in statistics, 103,s. 71-72.

Pokud je hodnota nižší než kritická hodnota, je proměnná statisticky nevýznamná a můžeme ji vypustit.

$$t \leq t^* \tag{4.7}$$

Je-li vyšší, je proměnná statisticky významná a má smysl ji v modelu ponechat.

$$t > t^* \tag{4.8}$$

Výstupy počítačových programů často zahrnují také tzv. p-hodnotu. Pomocí p-hodnot můžeme rozhodovat od statistické významnosti stejně jako pomocí t-testů, p-hodnotu ovšem porovnáváme rovnou se zvolenou hladinou významnosti. Pokud je p-hodnota nižší, než zvolená hladina významnost α, zamítáme nulovou hypotézu ve prospěch hypotézy alternativní a odmítáme nulovou hypotézu H_0 o nevýznamnosti testované proměnné.

4.3 Celkový F-test

Celkový F-test testuje významnost modelu jako celku, neboli jak dobře vybraná množina vysvětlujících proměnných vysvětluje variabilitu vysvětlované proměnné. Nulová a alternativní hypotéza pro model obsahující *k* proměnných vypadá následovně:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \tag{4.9}$$

$$H_a: H_0 \text{není pravdivá} \tag{4.10}$$

Alternativní hypotéza znamená, že alespoň jeden z β_j není roven nule.

Testovací statistika pro spočítání významnosti modelu jako celku je následující:

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} \sim F_{k,n-k-1} \tag{4.11}$$

Pokud nezamítneme nulovou hypotézu, znamená to, že množina vysvětlujících proměnných je špatně zvolená a nevysvětluje skoro žádnou variabilitu vysvětlované proměnné. Museli bychom najít jinou, lepší skladbu vysvětlujících proměnných.

4.4 Výběr nejlepší podmnožiny (Best subset selection)

Vybírání nejlepší podmnožiny je jedna z metod pro výběr vysvětlujících proměnných. Přestože hlavním nástrojem ekonometrické analýzy jsou teoreticky definované a odůvodněné specifikace modelu, existuje řada pomocných metod pro stanovení přesné specifikace regresního modelu. V této a následující kapitole je podrobněji popíši: první metodou je výběr nejlepší podmnožiny, druhá metoda je postupný výběr.

Abychom provedli výběr nejlepší podmnožiny, musíme provést samostatnou regresi pro všechny možné kombinace prediktorů, resp. vysvětlujících proměnných p. Musíme tedy sestrojit p-počet modelů, které obsahují pouze jeden prediktor, $\binom{p}{2} = \frac{p(p-1)}{2}$-počet modelů, které obsahují přesně dvě proměnné atd. Poté je potřeba prohlédnout všechny získané modely a vybrat z nich ten nejlepší. Problém nastává v exponenciálním růstu možností, při růstu prediktorů. Řešení 2^p možností je při větším počtu proměnných extrémně dlouhé. Postup řešení je pak popsán v následujícím algoritmu:

1. M_0 je tzv. nulový model, model který neobsahuje žádné prediktory. Takový model pouze předpovídá výběrový průměr pro každé měření.
2. Pro $k = 1, 2, \ldots p$:
    a. Sestrojíme $\binom{p}{k}$ modelů, které obsahují přesně k-prediktorů.
    b. Vybereme nejlepší model a pojmenujeme ho M_k. Za nejlepší model považujeme ten, který má nejmenší RSS (=Součet čtverců residuí), nebo největší R^2 (=Koeficient determinace – určuje, jaké procento variability vysvětlované proměnné se podařilo modelem vysvětlit).
3. Vybereme jeden nejlepší model z řady $M_0, \ldots, M_p$ modelů, pomocí jednoho nebo více z následujících kritérií: chyby predikce křížové validace (crossvalidated prediction error), C_p (AIC), BIC, nebo upravený $\bar{R}^2$. K jednotlivým kritériím se vyjádřím později.

Pomocí tohoto algoritmu se snažíme snížit počet možných modelů z 2^p na $p+1$. Musíme ovšem postupovat opatrně, protože s rostoucím počtem proměnných v modelu se monotónně snižuje RSS a zvyšuje R^2. Nízké RSS nebo vysoké R^2 označuje modely, které mají nízkou tréninkovou chybu, my ale chceme modely s nízkou testovou chybou. Proto je potřeba přistoupit k třetímu kroku, protože při posuzování pouze těchto statistik by vždy nejlepší model byl ten, který by obsahoval všechny, resp. co největší počet proměnných.

Ačkoliv je tento postup relativně jednoduchý, trpí výpočetními omezeními. Počet modelů potřebných prověřit rapidně roste s počtem prediktorů. Pokud máme 10 proměnných, resp. $p = 10$, je přibližně 1 000 možností, které se musí prověřit a pro 20 proměnných, $p = 20$ existuje přes jeden milion možností. Pokud pracujeme s modelem, obsahujícím kolem 40 proměnných či více, je tento postup výpočetně nemožný i s použitím extrémně rychlých

moderních počítačů. Existují sice výpočetní usnadnění, která eliminují některé možnosti, ale s rostoucím počtem proměnných jejich efektivnost slábne.

4.5 Postupný výběr (Stepwise Selection)

Kvůli výpočetní složitosti nemůže být výběr nejlepší podmnožiny použit pro velké p. Postupný výběr je alternativa pro výše zmíněný postup. Tato metoda není tak výpočetně náročná při větším počtu proměnných, protože prohledává mnohem omezenější množinu modelů. Je možné ji rozdělit do tří skupin, které se liší postupem hledání.

4.5.1 Postupný výběr – vpřed (Forward Stepwise Selection)

Jedná se o výpočetně efektivní alternativu k výběru nejlepší podmnožiny. Neuvažuje totiž všech 2^p možností, ale mnohem menší množinu modelů. Výběr vpřed začíná stejně jako výběr nejlepší podmnožiny s nulovým modelem, který neobsahuje žádný prediktor. Postupně pak po jedné přidává proměnné do modelu, dokud model neobsahuje všechny proměnné. V každém kroku je přidána jen ta proměnná, která přidává největší dodatečné zlepšení. Tento proces zachycuje následující algoritmus:

1. M_0 je tzv. nulový model, model který neobsahuje žádné prediktory.
2. Pro $k = 0, \ldots, p-1$:
    a. Zvážíme $(p-k)$-počet modelů, ve kterých navýší počet prediktorů v M_k o jeden další prediktor.
    b. Vybereme nejlepší model z $p-k$ modelů a nazveme ho M_{k+1}. Nejlepší je definováno jako model s nejnižším RSS nebo nejvyšší R^2.
3. Vybereme jeden nejlepší model z řady $M_0, \ldots, M_p$ modelů, pomocí jednoho nebo více z následujících kritérií: chyby predikce křížové, C_p (AIC), BIC, nebo upravený $\bar{R}^2$.

Narozdíl od výběru nejlepší podmnožiny, který obsahuje 2^p modelů, tato metoda obsahuje pouze nulový model a $p–k$ modelů pro k-tou iteraci, pro $k = 0, \ldots, p–1$. Počet modelů potřebných k prozkoumání pak zjistíme pomocí následujícího vzorce:

$$1 + \sum_{k=0}^{p-1}(p-k) = 1 + p(p+1)/2. \qquad 4.12$$

Pokud pracujeme s 20 proměnnými, $p = 20$, tak při vybírání nejlepší podmnožiny přímo musíme projít 1 048 576 modelů, zatímco tento postup toto číslo zreguluje na pouhých 211 modelů.

Ačkoliv je výpočetní výhoda této metody zřejmá, není bohužel zaručeno, že vždy najde nejlepší možný model ze všech 2^p možností. Pokud například budeme uvažovat soubor dat se 3 proměnnými a nejlepší model obsahující jednu proměnnou bude obsahovat x_1, ale nejlepší model obsahující dvě proměnné bude obsahovat x_2 a x_3, pak postupný výběr selže. Bude totiž

při vytváření modelu obsahujícího dvě proměnné vycházet z toho, že M_2 musí obsahovat x_1 společně s další proměnnou, protože x_1 je obsažena v M_1. Tato metoda navíc může být použita i v případě, že je počet proměnných p větší než počet pozorování n.

4.5.2 Postupný výběr – zpět (Backward Stepwise Selection)

Obdobně jako výběr vpřed, je tato metoda efektivní alternativou výběru nejlepší podmnožiny. Tato metoda funguje obráceně než výběr vpřed, to znamená, že začíná s modelem, který obsahuje všechny prediktory p a pak postupně po jednom odstraňuje nejméně užitečné prediktory. Detailně je tento postup popsán v následujícím algoritmu:

1. M_p značí model, který obsahuje všechny prediktory p.
2. Pro $k = p, p-1, \ldots, 1$:
    a. Uvažuje všechny k-modely, které obsahují všechny kromě jednoho prediktoru v M_k, tedy celkem $k-1$ prediktorů.
    b. Vybere nejlepší z těchto k-modelů a pojmenuje ho M_{k-1}. Nejlepší je definováno jako model s nejnižším RSS nebo nejvyšší R^2.
3. Vybere jeden nejlepší model z řady $M_0, \ldots, M_p$ modelů, pomocí jednoho nebo více z následujících kritérií: chyby predikce křížové validace, C_p (AIC), BIC, nebo upravený $\bar{R}^2$.

Stejně jako postup vpřed, i zpětný postup hledá pouze $1 + p(p+1)/2$ modelů. Stejně tak bohužel přináší i riziko, že vždy nebude odhalen nejlepší možný model. Na rozdíl od postupu vpřed pak tato metoda navíc vyžaduje, aby byl počet pozorování n větší než počet proměnných p.

4.5.3 Hybridní přístup (Hybrid Approaches)

Všemi třemi již zmíněnými postupy se obvykle dá dosáhnout podobných, ale ne identických modelů. Další alternativou je hybridní verze výběru vpřed a zpět. Proměnné jsou do modelu přidávány postupně analogicky, jako při výběru vpřed, ale po přidání každé proměnné může metoda i jakoukoliv proměnnou z modelu vyřadit, pokud už daná proměnná nemá dostatečný přínos. Tato metoda se více podobá výběru nejlepší podmnožiny, ale přitom si zachovává výpočetní zvýhodnění postupného výběrů.

4.6 Výběr optimálního modelu

Všechny postupy vyústí ve vytvoření množiny modelů, kde každý obsahuje podmnožinu prediktorů p. Abychom mohli tyto metody implementovat, musíme být schopni rozhodnout, který z modelů je nejlepší. Jak již bylo napsáno, model, který bude obsahovat všechny proměnné, bude mít nejmenší RSS a největší R^2. Proto musíme využít jiná kritéria, která nebudou závislá na počtu proměnných obsažených v jednotlivých modelech.

Problém je v tom, že hodnoty RSS a R^2 souvisejí s tzv. tréninkovou chybou (training MSE).

$$MSE = \frac{RSS}{n} \qquad 4.13$$

kde:

MSE = mean squared error = střední čtvercová chyba,

RSS = residual sum of squares = součet čtverců residuí a

n = počet pozorování.

Tato hodnota znázorňuje, jak se liší hodnoty predikované modelem od reálných hodnot. Logicky se tedy snažíme tuto chybu minimalizovat, resp. vybrat model, kde je tato chyba co nejmenší, protože pak jsou naše odhady udělané na základě takovýchto modelů přesnější. Problém ovšem tkví v tom, že tréninková chyba a tudíž i hodnoty RSS a R^2 se vztahují pouze na výběrový soubor, na základě kterého konstruujeme model, ale nijak nevypovídají o tom, jak přesné odhady nám model poskytne, pokud se dostaneme k hodnotám, ke kterým jsme původně neměli přístup. Abychom si tedy udělali představu o tom, jak bude model fungovat, pokud budeme mít přístup k dalším hodnotám, musíme se soustředit na tzv. testovou chybu (test MSE). Tato chyba už vypovídá o tom, jak bude model odhadovat hodnoty u dat, ke kterým jsme neměli přístup při jeho konstruování. Často se bohužel stává, že hodnoty obou chyb se liší, v takovém případě nás zajímá hodnota testová chyba.

Hodnotu této chyby ovšem obvykle neznáme a musíme ji odhadnout, abychom byli schopni vybrat nejlepší možný model z množiny modelů, kterou jsme získali použitím některé z výše zmíněných metod. Existují dva obvyklé přístupy:

1. Můžeme nepřímo odhadnout hodnotu testové chyby provedením úpravy tréninkové chyby. V této metodě využijeme hodnoty ukazatelů: C_p, AIC, BIC, upravený $\bar{R}^2$.
2. Můžeme přímo odhadnout testovou chybu, pomocí buď validačního setu, nebo křížové validace.

4.6.1 Nepřímý odhad - C_p, AIC, BIC a upravený $\bar{R}^2$

Jak bylo řečeno, RSS a R^2 jsou neklesající funkcí počtu proměnných, proto je nelze použít při určování nejlepšího modelu, pokud vybíráme z modelů s různým počtem proměnných, protože by vždy preferovaly model, který by obsahoval všechny proměnné. Existují ale různé způsoby, jak se dají taková kritéria upravit, aby se dala použít i pro porovnání modelů s různým počtem proměnných. Uvádím následující čtyři kritéria:

- C_p,
- „Akaike information criterion“ (AIC),
- „Bayesian information criterion“ (BIC),
- upravený koeficient determinace $\bar{R}^2$.

Pro lineární regresní model obsahující *k*-prediktorů, se C_p, jakožto odhad tréninkové MSE, vypočítá pomocí následující rovnice:

$$C_p = \frac{1}{n}(RSS + 2k\hat{\sigma}^2), \qquad 4.14$$

kde $\hat{\sigma}^2$ je odhad rozptylu chyby ϵ spojenou s každou odezvou v měření. C_p statistika přidává penalizaci v podobě $2k\hat{\sigma}^2$ k tréninkovému RSS, aby tím regulovala to, že tréninková chyba má tendenci podhodnocovat testovou chybu. Penalizace pak roste s počtem proměnných, aby vyrovnala RSS, jež se s rostoucím počtem proměnných snižuje. Pokud je tedy $\hat{\sigma}^2$ nezkreslený odhad σ^2, pak je C_p nezkresleným odhadem tréninkové MSE. Z toho vyplývá, že C_p nabývá nižších hodnot pro modely, které mají jen malou testovou chybu. Pokud se tedy rozhodujeme na základě této charakteristiky, snažíme se vybrat model s nejnižší hodnotou C_p.

AIC kritérium se používá pro velké třídy modelů odhadovaných metodou maximální věrohodnosti. V případě některých modelů je ale hledání maximální věrohodnosti ekvivalentní hledání nejmenších čtverců. V takovýchto případech se AIC spočítá následovně:

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2k\hat{\sigma}^2). \qquad 4.15$$

Z toho plyne, že pro metodu nejmenších čtverců jsou C_p a AIC proporcionální. AIC vždy bude $\hat{\sigma}^2$-krát větší než C_p. Při výběru modelu na základě AIC preferujeme modely s nejnižší hodnotou tohoto kritéria.

BIC je založeno na Bayesovském přístupu ke statistické analýze. Vzorec pro výpočet ale vypadá obdobně jako u C_p a AIC. Pro model nejmenších čtverců s *k*-prediktory vypadá následovně:

$$BIC = \frac{1}{n}(RSS + \log(n)\, k\hat{\sigma}^2). \qquad 4.16$$

Obdobně jako C_p i BIC bude nabývat menších hodnot pro modely s nízkou testovou chybou. V porovnání s C_p má BIC tendenci vybírat modely, které obsahují menší počet proměnných. To je dáno tím, že je při výpočtu BIC v rovnici *log (n)*, který nabývá větších hodnot než 2 pro jakékoliv $n > 7$. BIC proto penalizuje modely obsahující více proměnných silněji než C_p, resp. AIC.[3]

Posledním kritériem je upravený koeficient determinace $\bar{R}^2$. Koeficient determinace R^2 se spočítá pomocí residuálních (RSS) a celkových (TSS) součtů čtverců:

[3] JAMES, Gareth, Daniela WITTEN, Trevor HASTIE a Robert TIBSHIRANI.*An introduction to statistical learning: with applications in R*. Springer texts in statistics, 103,s. 210-213.

$$R^2 = 1 - \frac{RSS}{TSS}. \qquad 4.17$$

Hodnota R^2 udává, jaký podíl variability *y* je vysvětlen pomocí *x*, tato interpretace platí ale pouze v případě přítomnosti úrovňové konstanty. Hodnota blížící se k jedné značí, že jsme pomocí *x* schopni vysvětlit velkou část variability *y*. Pokud se hodnota blíží k nule, znamená to, že regrese vysvětlila jen malý podíl variability *y*. To může značit, mimo jiné, špatný model. Jak víme, tak RSS je neklesající funkcí počtu proměnných. Z toho vyplývá, že R^2 se bude vždy zvyšovat, pokud budeme do model přidávat více proměnných. Je tedy potřeba výpočet upravit tak, aby byl zohledněn (penalizován vysoký) počet proměnných v modelu. Toho docílíme následovně:

$$\bar{R}^2 = 1 - \frac{RSS/(n-k-1)}{TSS/(n-1)}, \qquad 4.18$$

$\bar{R}^2$ je atraktivní právě proto, že penalizuje přidání nových proměnných do modelu. Neplatí tedy, na rozdíl od R^2, že při přidání nových proměnných bude jeho hodnota vždy stoupat. Toho je docíleno zahrnutím počtu proměnných *k* do rovnice. Pokud přidáme do modelu další proměnnou, RSS se sníží, ale zároveň se sníží n–k–1, takže zlomek $\frac{RSS}{(n-d-1)}$ se může s přidáním nové proměnné do regrese jak zvýšit, tak snížit. Takovýto upravený koeficient nám pak je schopný pomoci i při výběru mezi modely s různými počty proměnných. Na rozdíl od ostatních kritérií hledáme při posuzování kvality modelu co možná nejvyšší $\bar{R}^2$.[4]

4.6.2 Přímý odhad – validace a křížová validace

Pomocí validace a křížové validace můžeme testovou chybu odhadnout přímo. Jedná se o alternativu k výše zmíněným postupům. Můžeme pro každý model spočítat chybu validace a chybu křížové validace a pak vybrat model, kde bude odhad chyby nejmenší. Tento postup má výhodu oproti nepřímým odhadům, že poskytuje přímý odhad testovací chyby. Navíc je tento odhad možné provádět i na modelech, kdy je těžké určit počet stupňů volnosti, nebo když je složité odhadnutí chyby rozptylu.

Princip validační metody spočívá v tom, že data, která máme k dispozici, náhodným výběrem rozdělíme na tréninkovou a testovací množinu. Na tréninkovou množinu pak použijeme metodu výběru nejlepší podmnožiny a pak na základě MSE vybereme model

[4] WOOLDRIDGE, Jeffrey M. *Introductory econometrics: a modern approach*. 4th ed. Mason, OH: South Western, Cengage Learning, c2009.,s. 200-201. ISBN 0324660545.

s nejmenší chybou. Nesmíme pak zapomenout, že model, který takto získáme, ještě není výsledný. Musíme nezávisle na těchto výpočtech použít metodu výběru nejlepší podmnožiny na celý datový soubor. Model, který má stejný počet proměnných jako ten, který jsme získali z testovací množiny, je námi hledaný nejlepší model. Oba dva modely se totiž liší nejen v koeficientech u jednotlivých proměnných, ale mohou obsahovat i různé proměnné, proto je důležité vybrat ten model, který jsme získali z celého datového souboru.

Hlavní rozdíl mezi validační metodou a metodou křížové validace spočívá v tom, že při křížové validace nedělíme model pouze na dvě množiny, ale model rozdělíme na větší počet stejných množin. Jednu z nich označíme jako testovací a ostatní jako tréninková. Tato metoda je výpočetně složitá, protože musíme pro každou tréninkovou množinu použít metodu výběru nejlepší podmnožiny. Tímto způsobem se vybere nejlepší model z každé tréninkové množiny a následně se tyto modely otestují na testovací množině a vybere se nejlepší. V dalším kroku se pak stává jiná množina testovací a celý proces se opakuje několikrát, pokaždé s jinou podmnožinou tvořící tréninkovou a testovací množinu. V praxi se obvykle kvůli výpočetní náročnosti používá rozdělení na $k=5$ nebo $k=10$ skupin.

Metoda křížové validace se dá popsat následujícím algoritmem:

1. Nejdříve náhodně rozdělíme všechna pozorování do k skupin, kdy mají všechny skupiny přibližně stejnou velikost.
2. První skupinu označíme jako testovací množinu a ostatní $k-1$ množiny považujeme za tréninkové. Tento postup poté opakujeme k-krát, vždy s jinou skupinou pozorování označenou jako testovací množina.
3. Tento proces má za výsledek k odhadů testových chyb (MSE_1, MSE_2, …, MSE_k) a krossvalidační odhad testové chyby získáme zprůměrováním těchto hodnot:

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i \qquad 4.19$$

4. Nakonec použijeme metodu výběru nejlepší podmnožiny na celý datový soubor a vybereme takový model, který obsahuje stejný počet proměnných, jako obsahuje model, který měl nejmenší odhad krossvalidační testové chyby.

Tyto postupy jsou výpočetně náročné. Proto se dříve dávala přednost odhadování testových chyb nepřímým metodám, ovšem dnes s moderními počítači už není tento problém aktuální a proto je tento přístup velmi lákavý.[5]

4.7 Multikolinearita

4.7.1 Příčiny a důsledky

Multikolinearita je vlastnost jednoho konkrétního výběru, kterou netestujeme, ale zjišťujeme. Multikolinearita značí existenci vztahu lineární závislosti mezi pozorováními vysvětlujících proměnných, to znamená, že dvě nebo více vysvětlujících proměnných spolu úzce souvisí. Existuje perfektní multikolinearita, to znamená, že mezi sloupci matice **X** existuje lineární vztah a pak nemůžeme použít metodu nejmenších čtverců, protože je porušen Gauss-Markův předpoklad o plné hodnosti matice **X**. Pokud mezi sloupci matice **X** existuje lineární závislost, pak jde buď o kolinearitu, pokud se jedná pouze o jeden vztah lineární závislosti. Pokud se jedná o více vztahů lineární závislosti, pak jde o multikolinearitu.

Multikolinearita se často vyskytuje v modelech s časovými řadami, zejména v modelech pracujících s makroekonomickými údaji, které se často vyvíjejí stejným směrem. Často také vzniká mezi proměnnou a její zpožděnou hodnotou, nebo pokud v regresi použijeme proměnnou a její násobek s jinou proměnnou. Výskyt perfektní multikolinearity hrozí, např. když chybně použijeme dummy proměnné a chybně formulujeme model.

V případě zjištění multikolinearity, zůstávají odhady nestranné a vydatné. Důsledky jsou snížená přesnost odhadů regresních koeficientů, velká citlivost odhadové funkce metody nejmenších čtverců na velmi malé změny v matici **X**. Dalším důsledkem jsou velké standardní chyby odhadů a i při vysokém R^2 jsou jednotlivé t-testy statisticky nevýznamné.

4.7.2 Zjišťování a měření

Pokud máme jen dvě vysvětlující proměnné, stačí nám spočítat jejich párový koeficient korelace.

$$r_{x_1x_2} = \frac{cov(x_1x_2)}{s_{x_1}s_{x_2}} \in \langle -1,1 \rangle \qquad 4.20$$

Multikolinearitu považuje za únosnou, pokud není větší než 0,9 nebo větší než koeficient determinace modelu.

U většího počtu proměnných už párové korelační koeficienty nestačí. Pro změření kolinearity potřebujeme spočítat pomocné regrese každé vysvětlující proměnné na všech

[5] JAMES, Gareth, Daniela WITTEN, Trevor HASTIE a Robert TIBSHIRANI.*An introduction to statistical learning: with applications in R*. Springer texts in statistics, 103,s. 181-183,213-214.

ostatních vysvětlujících proměnných, abychom zjistili koeficienty determinace pro každou pomocnou regresi R_i^2, pro i vysvětlujících proměnných. Multikolinearitu považujeme za únosnou, pokud žádný z těchto koeficientů determinace z pomocných regresí není větší než koeficient determinace z hlavní regrese.

$$R_i^2 < R^2 \tag{4.21}$$

Pokud v modelu zjistíme přítomnost multikolinearity a chceme se jí zbavit, můžeme například zvětšit rozsah výběru, kombinovat v modelu průřezová data a časové řady, změnit specifikaci modelu tak, že vynecháme vysvětlující proměnné se statisticky nevýznamnými parametry nebo transformovat pozorování a přejít na diference, nebo použít podíly proměnných. Multikolinearita ovšem jen znemožňuje zjistit vliv jednotlivých vzájemně závislých vysvětlujících proměnných na vysvětlovanou proměnnou, ale stále může určit jejich společný vliv. Pokud tedy chceme model použít pouze pro predikce, tak nám přítomnost multikolinearity nevadí.

V této práci budu hodnotit multikolinearitu na základě VIF (Variance inflation factor). VIF značí, o kolik je standardní chyba určité vysvětlující proměnné větší, než jaká by byla, pokud by daná proměnná nekorelovala s ostatními. VIF spočítáme podle následujících kroků:

1. Provedeme metodu nejmenší čtverců, kde použijeme x_i jako vysvětlovanou proměnnou:

$$x_i = \beta_{i+1}x_{i+1} + \beta_{i+2}x_{i+2} + \cdots + \beta_k x_k + c_0 + e, \tag{4.22}$$

   kde c_0 je konstanta a e je náhodná složka.

2. Spočítáme hodnotu VIF pro jednotlivé $\hat{\beta}_i$ pomocí následujícího vzorce:

$$VIF(\hat{\beta}_i) = \frac{1}{1-R_i^2}, \tag{4.23}$$

   kde R_i^2 je koeficient determinace regrese z kroku jedna, kdy jsme jako vysvětlovanou proměnnou použili x_i a jako vysvětlující proměnné všechny ostatní x.

3. Ve třetím kroku analyzujeme hodnotu VIF($\hat{\beta}_i$). Pravidlem bývá, že pokud hodnota VIF($\hat{\beta}_i$) > 10, tak je multikolinearita vysoká.

4.8 Heteroskedasticita

4.8.1 Příčiny a důsledky

Heteroskedasticita je situace, kdy je porušena podmínka konečného a konstantního rozptylu náhodných složek. To znamená, že náhodná složka může mít pro každé pozorování odlišný rozptyl.

$$E(u_i^2) = \sigma_i^2 \neq konst \tag{4.24}$$

Jedna z možných příčin heteroskedasticity je chybná specifikace modelu, kdy jsou v modelu vynechané významné vysvětlující proměnné. Další příčinou může být odhad z průřezových dat se značnou variabilitou ve výběru, nebo kumulace chyb měření proměnných s rostoucí hodnotou vysvětlované proměnné. Heteroskedasticita může vznikat i při odhadování z upravovaných dat místo z původních pozorování, např. ze skupinových průměrů získaných z tříděných dat.

Důsledky heteroskedasticity nemají vliv na nestrannost a konzistenci OLS (Ordinary least squares) estimátoru, ale na vydatnost. Bodové odhady nemají minimální rozptyl a to znamená, že nejsou vydatné a ani asymptoticky vydatné, takže OLS estimátor už není BLUE (Best linear unbiased estimator) a tudíž mohou existovat vydatnější lineární estimátory. Další důsledky souvisí s intervalovými odhady, které jsou nesměrodatné, a se statistickými t-testy a F-testy, které ztrácejí na síle. Existují různé formy heteroskedasticity, pro které existují různé postupy.

4.8.2 Robustní indukce

Jak bylo poznamenáno výše, v případě heteroskedasticity nemůžeme používat t-testy a F-testy, protože t-statistika nemá *t* rozdělení a F-statistika nemá *F* rozdělení. Byly ale vyvinuty asymptoticky platné úpravy vzorců pro OLS standartní chyby, *t* a *F* statistiky, které jsou platné i v přítomnosti heteroskedasticity neznámé formy. Tyto robustní postupy jsou platné, alespoň pro velké vzorky, bez ohledu na to, jestli chyby mají konstantní rozptyl nebo ne.

Hodnotu OLS standardní chyby robustní vůči heteroskedasticitě pro $\hat{\beta}_j$ získáme jako druhou odmocninu následujícího vzorce:

$$\widehat{Var(\hat{\beta}_j)} = \frac{\sum_{i=1}^{n} \hat{r}_{ij}^2 \hat{u}_i^2}{RSS_j^2}, \tag{4.25}$$

kde $\hat{r}_{ij}$ značí *i-té* residuum z regrese x_j na všech ostatních nezávislých proměnných a RSS_j je součet čtverců residuí z této regrese. V případě, že spočítáme vůči heteroskedasticitě

robustní standardní chyby, lze snadno zkonstruovat vůči heteroskedasticitě robustní t-statistiku, pouze při výpočtu použijeme robustní standardní chyby. Obvyklá F-statistika při heteroskedasticitě platná není, její robustní verze nemá žádnou jednoduchou formu, ale může být vypočítána použitím určitých statistických balíčků.

4.8.3 Testování heteroskedasticity

Když víme, že existují postupy, které je možné použít bez ohledu na to, zda víme, jestli je nebo není přítomna heteroskedasticita, nabízí se otázka, zda má vůbec smysl testovat její přítomnost. Existuje několik dobrý důvodů, proč má smysl tyto testy provádět. Zaprvé normální t-statistika má přesné t-rozdělení, zatímco u robustní verze toto platí jen pro velké vzorky. Zadruhé, pokud je přítomna heteroskedasticita, OLS estimátor už není nejlepší nestranný lineární odhad, takže je možné získat lepší odhad než OLS.

Pro testování přítomnosti heteroskedasticity existují různé druhy parametrických a neparametrických testů. Mezi parametrické řadíme např.: Breusch-Paganův test, Whiteův test, Parkův test a další. Mezi neparametrické testy řadíme například: Spearmanův test korelace pořadí a Goldfeld-Qaundtův test.

4.8.4 Breusch-Paganův Test

Jako obvykle začínáme s lineárním modelem. Jako nulovou hypotézu testujeme, že hodnota u^2 nezávisí na vysvětlujících proměnných, předpokládáme, že v modelu je homoskedasticita. Breusch-Paganův test má tři základní kroky:

1. Odhadneme model pomocí metody nejmenších čtverců jako obvykle. Získáme druhou mocninu OLS residuí $\hat{u}^2$.
2. Provedeme regresy:

$$\hat{u}^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \cdots + \delta_k x_k + error \qquad 4.26$$

3. Sestrojíme F-statistiku nebo LM-statistiku a spočítáme p-hodnoty. Pokud je p-hodnota dostatečně malá, to znamená nižší než zvolená hladina významnosti, tak zamítneme nulovou hypotézu o homoskedasticitě.

5 Empirická analýza a vyhodnocení výsledků

V praktické části se zaměřím na hledání nejlepšího odhadu modelu pomocí postupů uvedených ve 3. kapitole. Pomocí každého postupu vyberu model, který bude nejlépe splňovat kritéria (nepřímý odhad – upravený R^2, C_p a BIC a přímý odhad – validace a křížová validace), a z těchto modelů nakonec vyberu model nejlepší. Všechny výpočty byly prováděny v programu R.

Původní model obsahující všechny proměnné vypadá následovně:

$$Cena = \beta_0 + \beta_1 * Pamet + \beta_2 * RAM + \beta_3 * Jemnonst\ Displeje + \beta_4 * Uhlopricka + \beta_5 * Vydrz + \beta_6 * Pocet\ Jader + \beta_7 * Foto + \beta_8 * OS + \beta_9 * iPhone \quad (5.1)$$

Tento model postupně pomocí postupů popsaných v předchozí kapitole upravím a nakonec určím nejlepší možný model.

5.1 Výběr nejlepší podmnožiny (Best subset selection)

Jak jsem uvedl v předchozí kapitole, jedná se o hledání nejlepšího modelu v množině všech možných modelů. Na základě kritérií RSS a R^2 se vybere podmnožina nejlepších modelů pro jednotlivé počty proměnných a z těchto modelů se pomocí upraveného R^2, C_p a BIC vybere jeden nejlepší model. AIC statistiku do hodnocení nezahrnuji, protože je proporcionální s C_p. Tento postup prohledává 2^p modelů, v našem případě je devět vysvětlujících proměnných, $p = 9$, tato metoda tedy postupně prověří 512 modelů.

Následující tabulka znázorňuje, jaké proměnné obsahují jednotlivé modely. Pokud se u proměnné v řádku nachází hvězdička, bude model tuto proměnnou obsahovat. Např. nejlepší model obsahující právě pět vysvětlujících proměnných obsahuje následující proměnné: „RAM“, „Jemnost displeje“, „Vydrz“, „Pocet jader“ a „iPhone“.

	Pamet	RAM	Jemnost.Displeje	Uhlopricka	Vydrz	Pocet.Jader	Foto	Android	iPhone
1 (1)	" "	" "	"*"	" "	" "	" "	" "	" "	" "
2 (1)	" "	"*"	" "	" "	" "	" "	" "	" "	"*"
3 (1)	" "	"*"	"*"	" "	" "	" "	" "	" "	"*"
4 (1)	" "	"*"	"*"	" "	"*"	" "	" "	" "	"*"
5 (1)	" "	"*"	"*"	" "	"*"	"*"	" "	" "	"*"
6 (1)	"*"	"*"	"*"	" "	"*"	"*"	" "	" "	"*"
7 (1)	"*"	"*"	"*"	"*"	"*"	"*"	" "	" "	"*"
8 (1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"	" "	"*"
9 (1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"

Tabulka 5.1: Výběr nejlepší podmnožiny: Proměnné modelů

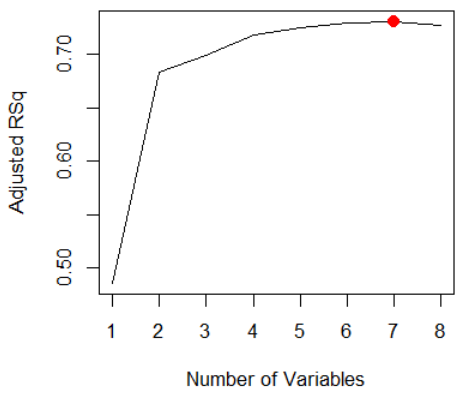
Zdroj: autor

Snížili jsme tedy počet modelů, které je potřeba dále testovat, z původních $2^p = 512$ na podmnožinu obsahující $p+1 = 9+1 = 10$ modelů, devět výše znázorněných modelů pro různý počet proměnných a jeden zahrnující pouze konstantu.

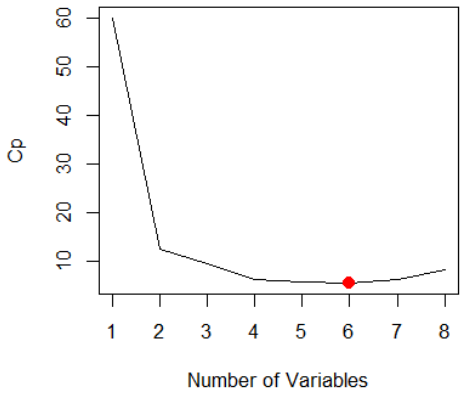
Dalším krokem je vybrat z této podmnožiny modelů jeden nejlepší model. K tomu využiji přímý a nepřímý odhad testovací chyby.

5.1.1 Nepřímý odhad

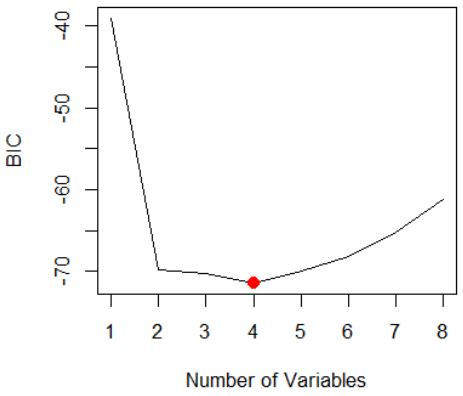
Nejdříve se zaměřím na nepřímý odhad. Model budu vybírat na základě kritérií $\bar{R}^2$, C_p a BIC. V následujících obrázcích 2 - 4 jsou znázorněny nejlepší modely, určené jednotlivými kritérii. Každé ovšem určilo jiný model, podle upraveného koeficientu determinace $\bar{R}^2$ je nejlepší model ten, který obsahuje sedm proměnných. Tento model vysvětluje 73,04% variability vysvětlované proměnné a očekává se u něj nejmenší testovací chyba. C_p označuje jako model s nejnižší testovací chybou model s šesti proměnnými, tento model vysvětluje 72,96% variability, hodnota $C_p = 5{,}43$. Podle Bayesovského informačního kritéria má nejnižší hodnotu testovací chyby model obsahující pouze čtyři proměnné, hodnota BIC = -71,47, ten vysvětluje 71,77% variability vysvětlující proměnné.



Obrázek 5: Výběr nejlepší podmnožiny $\bar{R}^2$



Obrázek 6: Výběr nejlepší podmnožiny - C_p



Obrázek 7: Výběr nejlepší podmnožiny -BIC

Zdroj: autor

Pokud se na obrázky podíváme pozorněji, tak zjistíme, že nejen hodnoty $\bar{R}^2$, ale ani hodnoty C_p se pro všechny tři vybrané modely příliš neliší. Rozdíly hodnot BIC jsou znatelnější, ale jak již bylo zmíněno v minulé kapitole, BIC penalizuje hodnoty obsahující více proměnných silněji, z části tedy můžeme tyto rozdíly přičíst i tomuto faktoru.

Na základě získaných hodnot jsem jako nejlepší model vybral ten, který obsahuje šest proměnných. Tento model má hodnotu $\bar{R}^2 = 0{,}7296$, to znamená, že model vysvětluje 72,96% variability vysvětlované proměnné cena a jedná se o druhou nejvyšší hodnotu. Hodnota C_p je nejlepší $C_p = 5{,}43$ a hodnota $BIC = -68{,}17$, ta se od nejlepší možné hodnoty -71,47 zásadně neliší. Zvolený model vypadá následovně:

$$Cena = -3335{,}29 + 54{,}22 * Pamet + 1{,}73 * RAM + 11{,}3 * Jemnost\ displeje + 1{,}98 * Vydrz - 319{,}75 * Pocet\ jader + 7375{,}79 * iPhone \quad (5.2)$$

```
Residuals:
    Min      1Q  Median      3Q     Max
-4284.6 -1130.5  -282.9   570.8  6212.8

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -3335.2874  1484.7466  -2.246  0.02819 *
Pamet               54.2217    35.9588   1.508  0.13658
RAM                  1.7256     0.8418   2.050  0.04455 *
Jemnost.Displeje    11.2957     5.3360   2.117  0.03823 *
Vydrz                1.9824     0.7318   2.709  0.00868 **
Pocet.Jader       -319.7527   174.7508  -1.830  0.07202 .
iPhone            7375.7909  1234.6811   5.974 1.18e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2250 on 63 degrees of freedom
Multiple R-squared:  0.7531,    Adjusted R-squared:  0.7296
F-statistic: 32.02 on 6 and 63 DF,  p-value: < 2.2e-16
```

Tabulka 5.2: Nepřímý odhad: Statistiky modelu s šesti proměnnými

Zdroj: autor

V nejlepším modelu vybraném pomocí nepřímého odhadu testovací chyby zůstávají následující proměnné „Pamet“, „RAM“, „Jemnost displeje“, „Vydrz“, „Pocet jader“ a „iPhone“. Hodnoty jednotlivých proměnných pak lze interpretovat následovně: Konstanta má hodnotu -3335,29. Hodnoty proměnných značí, že cena telefonu stoupá o 55,22 Kč za každý GB paměti, o 1,73 Kč za každý MB RAM, o 11,3 Kč za každý PPI bod, o 1,98 Kč za každou mAh a o 7375,95 Kč pokud se jedná o iPhone. Cena mobilu klesá, pokud procesor obsahuje více jader, a to o 319,75 Kč za jedno jádro.

Většina koeficientů značí růst ceny telefonu. Takovýto výsledek se dal očekávat, protože každá z proměnných nějakým způsobem vystihuje kvalitu telefonu, a čím vyšší hodnotu jednotlivé proměnné nabývají, tím kvalitnější by měl telefon být. Zajímavá je hodnota koeficientu u proměnné iPhone, kde jsem sice očekával, že značka bude mít nemalý vliv na finální cenu, ale neočekával jsem, že bude takto výrazný v porovnání s ostatními proměnnými. Pokud by model obsahoval více proměnných reprezentujících jednotlivé značky, mohlo by být zajímavé sledovat a porovnat jejich vliv na výslednou cenu. Jediná proměnná, která se nevyvíjí podle očekávání, je počet jader. Vyšší počet jader v procesoru by měl mít za výsledek vyšší výkon telefonu, proto jsem očekával, že cena telefonu bude růst společně s jejich počtem. Toto tvrzení ovšem není vždy pravdivé a je poměrně běžné, že procesory obsahující pouze jedno jádro mohou být výkonnější než ty, které obsahují dvě resp. čtyři. To by mohl být důvod, proč náš model s vyšším počtem jader cenu snižuje, protože v takových případech klesá i výkon, čili i kvalita. Důvodem také může být to, že v procesoru nejde jen o počet jader, ale také například o frekvenci a na výsledný výkon má vliv více faktorů.

Ještě provedeme test heteroskedasticity a zjistíme, jak je to s přítomností multikolinearity. Multikolinearitu zjistíme pomocí VIF. Bude-li hodnota VIF pro proměnnou větší než 10, naznačuje to silnou multikolinearitu. Pro model obsahující šest proměnných jsou hodnoty VIF pro jednotlivé proměnné znázorněny v následující tabulce:

Pamet	RAM	Jemnost.Displeje	Vydrz	Pocet.Jader	iPhone
2.199955	5.926840	3.303311	2.918757	1.377480	1.398152

Tabulka 5.3: VIF hodnoty pro jednotlivé proměnné modelu s šesti proměnnými

Zdroj: autor

Nejvyšší hodnotu má proměnná „RAM“ VIF[RAM]=5,927. Z toho vyplývá, že nejvíce koreluje s ostatními proměnnými. U ostatních proměnných je hodnota nižší, takže jsme splnili podmínku, že hodnota všech VIF je nižší než 10 a můžeme říci, že se multikolinearita v našem modelu nevyskytuje.

Dále otestuje heteroskedasticitu. Testovat jí budeme pomocí Breusch-Paganova testu, který byl zmíněn ve 4. kapitole. Hodnoty B-P testu pro model s šesti proměnnými jsou následující: BP = 28.063, df = 6, p-value = 9.141e-05. P-hodnota je znatelně menší než jedno-procentní hladina významnosti, takže se v modelu heteroskedasticita vyskytuje. To znamená, že náš model není vydatný, a bylo by možné získat lepší odhad při použití jiné metody než metody nejmenších čtverců. Když víme, že se v modelu vyskytuje heteroskedasticita, můžeme provést testy robustní vůči heteroskedasticitě, v následující tabulce jsou zaznamenány hodnoty standardních chyb robustních vůči heteroskedasticitě a robustních t-poměrů.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-3335.28737	1544.06344	-2.1601	0.0345791	*
Pamet	54.22166	33.70975	1.6085	0.1127292	
RAM	1.72557	1.01985	1.6920	0.0955916	.
Jemnost.Displeje	11.29569	5.64747	2.0001	0.0498011	*
Vydrz	1.98236	0.89352	2.2186	0.0301252	*
Pocet.Jader	-319.75270	178.26121	-1.7937	0.0776553	.
iPhone	7375.79086	1813.65151	4.0668	0.0001349	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

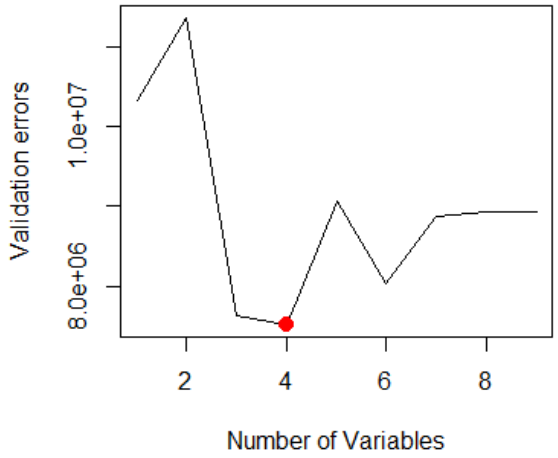
Tabulka 5.4: Standardní chyby a t-testy robustní vůči heteroskedasticitě modelu s šesti proměnnými

Zdroj: autor

Pokud porovnáme hodnoty v tabulce 5.4 s hodnotami v tabulce 5.2, vidíme, že t-testy robustní vůči heteroskedasticitě jsou statisticky významné jen na vyšší hladině významnosti, než tomu bylo u normálních t-testů. Také všechny hodnoty standardních chyb robustních vůči heteroskedasticitě, kromě proměnné „Pamet“, jsou vyšší než normální hodnoty. Celkově z toho můžeme usuzovat, že model je horší, než se původně zdálo.

5.1.2 Přímý odhad

Další možností, jak vybrat nejlepší model, je přímý odhad testovací chyby. Ten získáme pomocí validace, nebo křížové validace. Pomocí validace soubor náhodně rozdělíme na dvě množiny, testovací a tréninkovou a dále postupujeme tak, jak bylo vysvětleno v minulé kapitole. Následující graf a tabulka obsahují hodnoty validačních chyb nejlepších modelů s různým počtem proměnných.



Obrázek 8: Validační chyby

Zdroj: autor

Počet proměnných	1	2	3	4	5	6	7	8	9
Validační chyba	10 320 896	11 350 947	7 628 623	7 526 282	9 055 822	8 031 356	8 882 737	8 923 021	8 925 797

Tabulka 5.5: Validační chyby

Zdroj: autor

Nejlépe dopadl model obsahující pouze čtyři proměnné. Tento model vypadá následovně:

$$Cena = -6451{,}47 + 20.24 * Jemnost\ displeje + 3{,}9 * Vydrz - 456{,}35 * Pocet\ jader + 7548.84 * iPhone \quad (5.3)$$

Jedná se ale o model získaný pouze z tréninkové množiny, musíme tedy ještě najít nejlepší model z celého datového souboru, který bude obsahovat právě čtyři proměnné:

$$Cena = -4192{,}01 + 2{,}17 * RAM + 12.23 * Jemnost\ displeje + 1{,}68 * Vydrz + 8295{,}19 * iPhone \quad (5.4)$$

```
Residuals:
    Min      1Q  Median      3Q     Max
-4645.8 -1163.6  -140.0   577.6  5627.9

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -4192.0111  1463.5675  -2.864  0.00562 **
RAM                  2.1645     0.8162   2.652  0.01005 *
Jemnost.Displeje    12.2295     5.2917   2.311  0.02401 *
Vydrz                1.6751     0.7291   2.297  0.02483 *
iPhone            8295.1870  1186.7289   6.990 1.81e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2299 on 65 degrees of freedom
Multiple R-squared:  0.734,     Adjusted R-squared:  0.7177
F-statistic: 44.85 on 4 and 65 DF,  p-value: < 2.2e-16
```

Tabulka 5.6: Přímý odhad: Statistiky modelu se čtyřmi proměnnými

Zdroj: autor

Modely 5.3 a 5.4 mají různé nejen koeficienty u jednotlivých proměnných, obsahují dokonce i různé proměnné.

Opět ještě provedeme kontrolu multikolinearity a heteroskedasticity. Multikolinearitu rozhodneme podle hodnot VIF, které jsou pro model se čtyřmi proměnnými udány v následující tabulce:

RAM	Jemnost.Displeje	Vydrz	iPhone
5.337251	3.111839	2.775641	1.237257

Tabulka 5.7: VIF hodnoty pro jednotlivé proměnné modelu se čtyřmi proměnnými

Zdroj: autor

Stejně jako v případě modelu se šesti proměnnými má nejvyšší hodnotu proměnná „RAM“, je dokonce nižší, stejně jako hodnoty všech ostatních. Opět jsou všechny hodnoty nižší než hranice 10, takže ani v tomto modelu jsme multikolinearitu nezjistili.

Jako u minulého modelu, opět provedeme B-P test a otestujeme, zda je v modelu přítomná heteroskedasticita. Výsledky B-P testu pro model se čtyřmi proměnnými jsou následující: BP = 23.421, df = 4, p-value = 0.0001043. P-hodnota je sice vyšší než v minulém případě, stále je ale nižší než jedno-procentní hladina významnosti, a tudíž zamítáme nulovou hypotézu o homoskedasticitě. V následující tabulce jsou znázorněny standardní chyby a t-testy robustní vůči heteroskedasticitě:

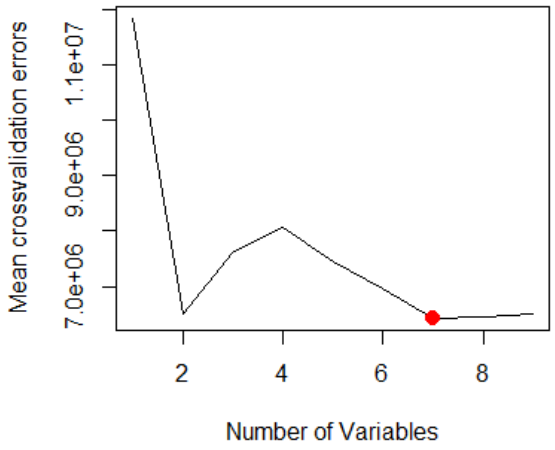
```
                    Estimate  Std. Error t value  Pr(>|t|)
(Intercept)      -4192.01112  1587.62567 -2.6404   0.01036 *
RAM                  2.16447     1.00528  2.1531   0.03503 *
Jemnost.Displeje    12.22948     5.54156  2.2069   0.03086 *
Vydrz                1.67507     0.88608  1.8904   0.06316 .
iPhone            8295.18697  1786.36591  4.6436 1.715e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tabulka 5.8: Standardní chyby a t-testy robustní vůči heteroskedasticitě modelu se čtyřmi proměnnými

Zdroj: autor

Když hodnoty v tabulce 5.8 porovnáme s normálními hodnotami z tabulky 5.6, vidíme, že stejně jako v předchozím modelu hodnoty standardních chyb jsou větší a t-test u proměnné „Vydrz“ je statisticky významný pouze na vyšší hladině významnosti.

Pokud použijeme křížovou validaci, pak při rozdělení datového souboru na sedm množin dostaneme průměrné chyby křížové validace, které jsou zaznamenány v následujícím grafu a tabulce:



Obrázek 9: Průměrné chyby křížové validace

Zdroj: autor

Počet proměnných	1	2	3	4	5	6	7	8	9
Průměrná chyba křížové validace	11 823 402	6 495 493	7 615 828	8 069 193	7 443 951	6 972 735	6 434 524	6 459 427	6 507 310

Tabulka 5.9: Průměrné chyby křížové validace

Zdroj: autor

Jak můžeme vidět, v tomto případě dopadl nejlépe model obsahující právě sedm proměnných. Model vypadá takto:

$$Cena = -6167{,}93 + 48{,}88 * Pamet + 1{,}81 * RAM + 10{,}85 * Jemnost\ displeje + 916{,}52 * Uhlopricka + 1{,}42 * Vydrz - 369{,}22 * Pocet\ jader + 7561{,}87 * iPhone \quad (5.5)$$

```
Residuals:
    Min      1Q  Median      3Q     Max
-3789.6 -1074.4  -249.6   513.6  6326.9

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -6167.9301  2995.2526  -2.059   0.0437 *
Pamet               48.8828    36.2397   1.349   0.1823
RAM                  1.8075     0.8440   2.142   0.0362 *
Jemnost.Displeje    10.8493     5.3440   2.030   0.0466 *
Uhlopricka         916.5195   842.0865   1.088   0.2806
Vydrz                1.4173     0.8964   1.581   0.1189
Pocet.Jader       -369.2161   180.3164  -2.048   0.0448 *
iPhone            7561.8698  1244.6740   6.075  8.3e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2247 on 62 degrees of freedom
Multiple R-squared:  0.7577,    Adjusted R-squared:  0.7304
F-statistic:  27.7 on 7 and 62 DF,  p-value: < 2.2e-16
```

Tabulka 5.10: Přímý odhad: Statistiky modelu se sedmi proměnnými

Zdroj: autor

Opět zjistíme, zda se v modelu vyskytuje multikolinearita, a otestujeme heteroskedasticitu. Existenci multikolinearity rozhodneme na základě následující tabulky:

Pamet	RAM	Jemnost.Displeje	Uhlopricka	Vydrz	Pocet.Jader	iPhone
2.241013	5.974399	3.322886	3.137645	4.392368	1.470916	1.425038

Tabulka 5.11: VIF hodnoty pro jednotlivé proměnné modelu se sedmi proměnnými

Zdroj: autor

Stejně jako v minulých případech je nejvyšší hodnota u proměnné „RAM“ a opět žádná není vyšší než 10. Zároveň si můžeme všimnout, že všechny proměnné jsou o něco vyšší, než tomu tak bylo v případě modelu s šesti proměnnými, tabulka 5.3.

Breusch-Paganův test heteroskedasticity dopadl následovně: BP = 25.419, df = 7, p-value = 0.0006392. Opět jsme zaznamenali malé zlepšení oproti předchozím modelům, každopádně i v tomto případě zamítáme nulovou hypotézu o homoskedasticitě. Tabulka s robustními standardními chybami a robustními t-testy vypadá takto:

```
                  Estimate  Std. Error t value  Pr(>|t|)
(Intercept)      -6167.93007  3333.04161 -1.8505   0.06900 .
Pamet               48.88284    33.50196  1.4591   0.14958
RAM                  1.80752     0.99950  1.8084   0.07539 .
Jemnost.Displeje    10.84926     5.42792  1.9988   0.05002 .
Uhlopricka         916.51951   833.39085  1.0997   0.27569
Vydrz                1.41727     0.91732  1.5450   0.12743
Pocet.Jader       -369.21613   193.54806 -1.9076   0.06107 .
iPhone            7561.86981  1667.55063  4.5347 2.698e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tabulka 5.12: Standardní chyby a t-testy robustní vůči heteroskedasticitě modelu se sedmi proměnnými

Zdroj: autor

Při porovnání tabulek 5.12 a 5.10, můžeme opět zaznamenat nárůst většiny standardních chyb a navíc je tu největší rozdíl ve vyhodnocení t-testů. Zatímco při použití normálních t-testů jsme na pěti-procentní hladině významnosti považovali za statisticky významné čtyři t-testy, u robustních t-testů bychom za statisticky významný mohli považovat pouze jeden.

5.2 Postupný výběr vpřed (Forward Stepwise Selection)

Dalším možným postupem při hledání nejlepšího modelu je postupný výběr vpřed. Tato metoda začíná s prázdným modelem a vždy přidá jednu proměnnou, která má největší dodatečné zlepšení, a takto postupuje, dokud do modelu nepřidá všechny proměnné. Tato metoda je oproti výběru nejlepší podmnožiny výpočetně úspornější, protože prohledává mnohem menší počet modelů. Zatímco výběr nejlepší podmnožiny v našem případě prohledával 512 modelů, tato metoda sníží jejich množství na 46. V následující tabulce jsou znázorněny proměnné, které jednotlivé modely obsahují. Např. nejlepší model obsahující právě pět vysvětlujících proměnných obsahuje následující proměnné: „RAM“, „Jemnost displeje“, „Vydrz“, „Pocet jader“ a „iPhone“.

		Pamet	RAM	Jemnost.Displeje	Uhlopricka	Vydrz	Pocet.Jader	Foto	Android	iPhone
1	(1)	" "	" "	"*"	" "	" "	" "	" "	" "	" "
2	(1)	" "	" "	"*"	" "	" "	" "	" "	" "	"*"
3	(1)	" "	"*"	"*"	" "	" "	" "	" "	" "	"*"
4	(1)	" "	"*"	"*"	" "	"*"	" "	" "	" "	"*"
5	(1)	" "	"*"	"*"	" "	"*"	"*"	" "	" "	"*"
6	(1)	"*"	"*"	"*"	" "	"*"	"*"	" "	" "	"*"
7	(1)	"*"	"*"	"*"	"*"	"*"	"*"	" "	" "	"*"
8	(1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"	" "	"*"
9	(1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"

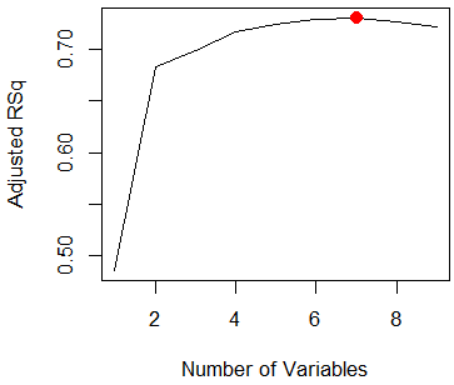
Tabulka 5.13: Postupný výběr - vpřed: Proměnné modelů

Zdroj: autor

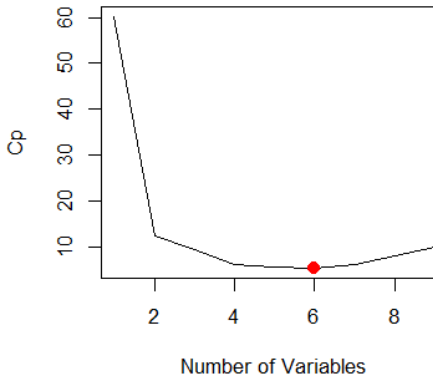
Můžeme si všimnout, že oproti výběru nejlepší podmnožiny se změní proměnné obsahující nejlepší model s dvěma proměnnými. Zatímco touto metodou je jako nejvhodnější proměnný pro model s jednou proměnnou zvolena „Jemnost displeje“. Objevuje se pak ve všech dalších modelech, ale podle metody výběru nejlepší podmnožiny se v případě, kdy model obsahuje právě dvě proměnné, do modelu hodí lépe proměnná „RAM“. Teď musíme z této množiny modelů vybrat ten nejlepší. Toho docílíme pomocí stejných postupů jako při výběru nejlepší podmnožiny.

5.2.1 Nepřímý odhad

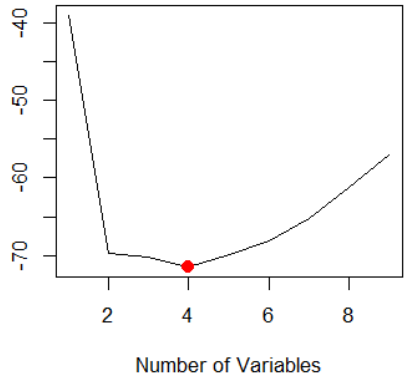
V následujících obrázcích jsou uvedeny hodnoty jednotlivých ukazatelů a vyznačený jejich maxima, resp. minima.



Obrázek 10: Postupný výběr - vpřed - $\bar{R}^2$



Obrázek 11: Postupný výběr - vpřed - C_p



Obrázek 12: Postupný výběr - vpřed - BIC

Zdroj: autor

Všechny tři testy určily nejlepší modely stejné, jako při výběru nejlepší podmnožiny a jejich hodnoty jsou identické. Upravený koeficient determinace určil jako nejlepší model ten, který obsahuje sedm proměnných, podle C_p se jedná o model obsahující šest proměnných a podle BIC model obsahující čtyři proměnné. Jako nejlepší model opět zvolíme model obsahující šest proměnných, jedná se o rovnici 5.2 v kapitole 5.1.1.

5.3 Postupný výběr zpět (Backward Stepwise Selection)

Poslední postup, kterému se zde budu věnovat, je postupný výběr zpět. Tato metoda na rozdíl od minulé, výběru vpřed, postupuje „odzadu“. Začíná s modelem obsahujícím všechny proměnné a postupně je po jedné vyřazuje. Vždy vyřadí tu, která má pro stávající model nejmenší přínos. Hlavní nevýhoda této metody pak spočívá v tom, že jednou vyřazené proměnné už znovu nezvažuje, jako je tomu v případě hybridního přístupu, takže se může stát, že je vyřazena proměnná, která by ovšem v dalších modelech byla významnější, než některá z proměnných, která v něm zůstala. Tato metoda opět prohledává pouze v našem případě 46 různým modelů, výpočetně je tedy opět úspornější, než proces hledání nejlepší podmnožiny. Následující tabulka znázorňuje postup vyřazování jednotlivých proměnných:

		Pamet	RAM	Jemnost.Displeje	Uhlopricka	Vydrz	Pocet.Jader	Foto	Android	iPhone
1	(1)	" "	"*"	" "	" "	" "	" "	" "	" "	" "
2	(1)	" "	"*"	" "	" "	" "	" "	" "	" "	"*"
3	(1)	" "	"*"	"*"	" "	" "	" "	" "	" "	"*"
4	(1)	" "	"*"	"*"	" "	"*"	" "	" "	" "	"*"
5	(1)	" "	"*"	"*"	" "	"*"	"*"	" "	" "	"*"
6	(1)	"*"	"*"	"*"	" "	"*"	"*"	" "	" "	"*"
7	(1)	"*"	"*"	"*"	"*"	"*"	"*"	" "	" "	"*"
8	(1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"	" "	"*"
9	(1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"

Tabulka 5.14: Postupný výběr - zpět: Proměnné modelů

Zdroj: autor

Zde si můžeme všimnout, že na rozdíl od předchozích dvou metod nezůstává jako poslední proměnná v modelu proměnná „Jemnost displeje“, ale proměnné „RAM“. Je to příklad toho, že když metoda v nějaké části postupu jednu proměnnou vyřadí, už ji v další části není možné do modelu vrátit. Je to jediná změna oproti předcházejícím dvěma postupům.

Vyhodnocení jednotlivých kritérií je identické, jako tomu bylo u postupného výběru – vpřed. Jako nejlepší model by byl tedy opět vybrán model obsahující šest proměnných, rovnice 5.2 v kapitole 5.1.1.

5.4 Vyhodnocení výsledků

Jednotlivými postupy jsme nakonec získali tři modely, obsahující různé počty proměnných. U výběru nejlepší podmnožiny jsme použili jednak nepřímý odhad, který jako nejlepší model zvolil model obsahující šest proměnných, rovnice 4.2 na straně 26. Pak jsme použili přímý odhad a validační metodu a metodu křížové validace. Validační metoda vybrala model obsahující čtyři proměnné, rovnice 5.4 na straně 32 a metoda křížové validace zvolila model, který obsahuje sedm proměnných, rovnice 5.5 na straně 34. Dále jsme použili postupné výběry, a to postupy vpřed a zpět. U obou dvou postupů jsme jako nejlepší model označili model obsahující šest proměnných, jako tomu bylo v případě nepřímého odhadu při výběru nejlepší podmnožiny, rovnice 5.2.

U všech modelů jsme pak zjišťovali, zda je přítomna multikolinearita, a testovali jsme heteroskedasticitu. Pro zjišťování multikolinearity jsme použili faktor VIF a zjistili jsme, že ani v jednom případě jeho hodnota nepřesahuje hranici, která by značila vysokou multikolinearitu. Nejvyšší naměřená hodnota faktoru VIF = 5,794 je u modelu obsahujícího sedm proměnných, můžeme tedy konstatovat, že se ani v jednom modelu multikolinearita nevyskytuje. Heteroskedasticitu jsme každého modelu testovali pomocí Breusch-Paganova testu heteroskedasticity. Výsledek Breusch-Paganových testů jednotlivých modelů zjistil, že se heteroskedasticity vyskytuje u všech tří modelů. Znamená to, že odhady sice zůstávají nestranné a konzistentní, ale nejsou vydatné a ani asymptoticky vydatné, takže mohou existovat vydatnější lineární estimátory.

Hodnoty jednotlivých kritérií jsou u všech tří modelů velmi podobné. $\bar{R}^2$ má nejvyšší hodnotu model se sedmi proměnnými, ale jen o 0,08 oproti tomu s šesti, hodnota C_p je nejlepší pro model obsahující šest proměnných, a jen u ukazatele BIC je větší rozdíl mezi modelem se čtyřmi proměnnými a ostatními dvěma modely. Takový výsledek můžeme ale z části poukázat na to, že BIC silněji penalizuje modely obsahující více proměnných, má tedy tendenci vybírat spíše modely s menším počtem proměnných. I v případě obou validačních metod, nejsou hodnoty toho modelu extrémně rozdílné oproti nejlepší variantě. U validační metody je model s šesti proměnnými třetí nejlepší model v pořadí a model se sedmi čtvrtý. V případě metody křížové validace je model s šesti proměnnými pátý nejlepší model a model se čtyřmi až předposlední. Na základě těchto testů provedených v této kapitole jsem jako

nejlepší z těchto tří modelů vybral ten, který obsahuje právě šest proměnných, rovnice 5.2 strana 30. Model vysvětluje 72,96% variability vysvětlované proměnné a obsahuje vysvětlující proměnné „Pamet“, „RAM“, „Jemnost displeje“, „Vydrz“, „Pocet jader“ a „iPhone“.

6 Závěr

V této práci jsem chtěl zjistit, jaké z vybraných technických parametrů ovlivňují cenu mobilních telefonů a jak. Postupně jsem se dostal k očekávaným výsledkům v tom smyslu, že všechny parametry krom jedné výjimky mají pozitivní vliv na růst výsledné ceny, neboli že s růstem jednotlivých parametrů se zvedá i cena. Toto vysvětluji tím, že každý z vybraných parametrů by měl mít pozitivní vliv na kvalitu telefonu, takže s rostoucí kvalitou se bude zvyšovat i cena. Jedinou zmiňovanou výjimkou byla proměnná označující počet jader v procesoru daného telefonu, kde platí, že pokud se počet jader v procesoru zvyšoval, výsledná cena se snižovala. Toto by mohl mít za následek fakt, že procesory obsahující více jader nejsou nutně kvalitnější a výkonnější než ty, které jich obsahují menší počet.

Zajímavé je také zaměření se na jedinou proměnnou zabývající se značkou mobilního telefonu. Očekávaný výsledek byl znovu takový, že pokud bude telefon značky Apple, resp. bude se jednat o iPhone, bude cena růst. Telefon se stejnými vlastnostmi od značky Apple bude tedy dražší než srovnatelné přístroje ostatních značek, ale neočekával jsem, že rozdíl bude až tak výrazný.

Na všechny odhady má samozřejmě také vliv přítomnost heteroskedasticity, takže výsledky musíme brát s jistou rezervou. Na vytvoření základní představy jsou ale tyto odhady dostačující.

Zajímavé by určitě bylo zkoumat tuto problematiku na větším vzorku a zahrnout do testování i více značek a dalších proměnných, které jsem v tomto relativně jednoduchém modelu nemohl použít. Jedná se např. o proměnné jako rozlišení a typ displeje a typ procesoru, které jsem bohužel nebyl schopný do mého modelu zahrnout kvůli jejich složitému zaznamenání. Nejen tyto parametry totiž jistě mají nemalý vliv na konečnou cenu telefonu a mohly by výsledný odhad podstatně zlepšit.

7 Zdroje

4EK214 Úvod do ekonometrie [online]. [cit. 2016-05-10]. Dostupné z: https://sites.google.com/site/ekonometrievse/4ek214

CZC.cz [online]. [cit. 2014-12-12]. Dostupné z: https://www.czc.cz/

Enviwiki.cz [online]. [cit. 2016-05-27]. Dostupné z: https://www.enviwiki.cz/wiki/Modernizace_GSM_s%C3%ADt%C4%9B_a_recyklace_mobiln%C3%ADch_telefon%C5%AF

Finanace.cz [online]. [cit. 2016-05-27]. Dostupné z: http://www.finance.cz/makrodata-eu/trh-prace/statistiky/mzda/

Fundamentals of Statistics: Variance Inflation Factor [online]. [cit. 2016-05-10]. Dostupné z: http://www.statistics4u.info/fundstat_eng/ee_mlr_vif.html

Gartner, Inc. [online]. [cit. 2016-05-27]. Dostupné z: http://www.gartner.com/technology/home.jsp

How much does an iPhone cost to make? [online]. [cit. 2016-05-27]. Dostupné z: http://www.alphr.com/features/388273/how-much-does-an-iphone-cost-to-make

IASTAT - INTERAKTIVNÍ UČEBNICE STATISTIKY [online]. [cit. 2016-05-10]. Dostupné z: http://iastat.vse.cz/

IDnes.cz: Mobil [online]. 2006 [cit. 2016-05-27]. Dostupné z: http://mobil.idnes.cz/pamatujete-10-let-stare-mobily-tezke-a-neforemne-krabice-co-nic-neumi-1f8-/mob_tech.aspx?c=A060630_014450_mob_tech_jm

IDnes.cz: Mobil [online]. 2012 [cit. 2015-05-10]. Dostupné z: http://mobil.idnes.cz/valka-procesor-jednojadro-dvoujadro-ctyrjadro-nabidne-stejny-vykon-1gh-/mob_tech.aspx?c=A120311_160032_mob_tech_vok

JAMES, Gareth, Daniela WITTEN, Trevor HASTIE a Robert TIBSHIRANI. *An introduction to statistical learning: with applications in R*. New York: Springer, 2013. Springer texts in statistics, 103. ISBN 1461471370.

Jan Zouhar, materiály k výuce [online]. [cit. 2016-05-10]. Dostupné z: http://nb.vse.cz/~zouharj/econCZ.html

Jana Sekničková, Výuka [online]. [cit. 2016-05-10]. Dostupné z: http://jana.seknicka.eu/vyuka/

Mobilmania.cz [online]. [cit. 2014-12-12]. Dostupné z: http://www.mobilmania.cz/

SHERMAN, Joshua. *Spendy but indispensable: Breaking down the full $650 cost of the iPhone 5* [online]. 2013 [cit. 2016-05-27]. Dostupné z: http://www.digitaltrends.com/mobile/iphone-cost-what-apple-is-paying/#:fC6rODmM7mVo5A

Tretikura.cz [online]. 2009 [cit. 2016-05-27]. Dostupné z: http://www.tretiruka.cz/news/mobilni-telefony-se-take-stanou-odpadem-zajimava-cisla-a-suroviny-v-mobilnich-telefonech/

VOŘÍŠEK, Lukáš. *Okno do historie mobilních telefonů: 12 perliček, které měnily mobilní svět* [online]. 2013 [cit. 2016-05-27]. Dostupné z: http://cdr.cz/clanek/okno-do-historie-12-telefonu-ktere-menily-mobilni-svet

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

8 Seznam obrázků

9 Seznam tabulek

10 Příloha

Kompletní seznam všech dat použitých pro výpočty v této práci.

Zdroj: www.mobilmania.cz a www.czc.cz, 12.12. 2014

	Název	Cena	Paměť	RAM	Jemnost Displeje	Úhlopříčka	Výdrž	Počet Jader	Fotoaparát	Android	iPhone
1	Apple iPhone 6 Plus 16GB	20000	16	1024	401	5,5	2915	2	8	0	1
2	Samsung Galaxy Note 4	18500	32	3072	515	5,7	3220	4	16	1	0
3	Apple iPhone 6 16GB	18400	16	1024	326	4,7	1810	2	8	0	1
4	Sony Xperia Z3 Dual	17200	16	3072	424	5,2	3100	4	21	1	0
5	Xiaomi Mi4 64GB	15000	64	3072	441	5	3080	4	13	1	0
6	Samsung Galaxy Note 3	14900	32	3072	386	5,7	3200	4	13	1	0
7	Lenovo Vibe Z2 Pro	14000	32	3072	490	6	4000	4	16	1	0
8	Samsung Galaxy Alpha	12800	32	2048	312	4,7	1860	8	12	1	0
9	Samsung Galaxy S5	12400	16	2048	432	5,1	2800	4	16	1	0
10	CAT S50	12400	8	2048	312	4,7	2630	4	8	1	0
11	Sony Xperia Z3 Compact	12000	16	2048	319	4,6	2600	4	21	1	0
12	Apple iPhone 5 16GB	12000	16	1024	326	4	1420	2	8	0	1
13	LG G3 32GB	11100	32	3072	534	5,5	3000	4	13	1	0
14	Nokia Lumia 1520	11000	32	2048	367	6	3400	4	20	0	0
15	Alcatel One Touch HERO 2	10500	16	2048	367	6	3100	8	13	1	0
16	Sony Xperia Z2	10000	16	3072	424	5,2	3200	4	21	1	0
17	Nokia Lumia 930	10000	32	2048	441	5	2420	4	20	0	0
18	LG Nexus 5	10000	16	2048	445	4,95	2300	4	8	1	0
19	Xiaomi Mi4 16GB	9900	16	3072	441	5	3080	4	13	1	0
20	Meizu MX4 32GB	9500	32	2048	418	5,36	3100	8	21	1	0

21	Sony Xperia Z1	9500	16	2048	441	5	3000	4	21	1	0
22	Samsung Galaxy S4	9300	16	2048	441	5	2600	4	13	1	0
23	Lenovo Vibe X2	9000	32	2048	441	5	2300	8	13	1	0
24	HTC One mini 2	9000	16	1024	326	4,5	2110	4	13	1	0
25	Zopo ZP990+	8800	32	2048	367	6	3000	8	14	1	0
26	CAT B15Q	8500	4	1024	233	4	2000	4	5	1	0
27	Meizu MX4 16GB	8500	16	2048	418	5,36	3100	8	21	1	0
28	Huawei Ascend P7	8500	16	2048	441	5	2500	4	13	1	0
29	HTC One	8400	32	2048	469	4,7	2300	4	4	1	0
30	Lenovo Vibe Z2	8400	32	2048	267	5,5	3000	4	13	1	0
31	Nokia Lumia 830	8300	16	1024	294	5	2200	4	10	0	0
32	Sony Xperia Z1 Compact	8300	16	2048	342	4,3	2300	4	21	1	0
33	Motorola Moto X	8200	16	2048	312	4,7	2200	2	10	1	0
34	Zopo ZP998	8000	16	2048	401	5,5	2400	8	14	1	0
35	Lenovo P70	7000	16	2048	294	5	4000	8	13	1	0
36	Apple iPhone 4S 8GB	7000	8	512	330	3,5	1432	2	8	0	1
37	Nokia Lumia 925	7000	16	1024	332	4,5	2000	2	9	0	0
38	Zopo ZP980+	6900	16	2048	441	5	2000	8	14	1	0
39	Alcatel One Touch Idol X+	6500	32	2048	441	5	2000	8	13	1	0
40	HTC Desire 601	6200	8	1024	245	4,7	2400	4	8	1	0
41	Nokia Lumia 735	6200	8	1024	312	4,7	2220	4	7	0	0
42	Sony Xperia T3	6000	8	1024	277	5,3	2500	4	8	1	0
43	Samsung Galaxy S4 mini	6000	8	1536	256	4,3	1900	2	8	1	0
44	Lenovo S850	5600	16	1024	294	5	2000	4	13	1	0
45	Xiaomi Redmi NOTE	5100	8	2048	267	5,5	3200	8	13	1	0
46	Apple iPhone 3G 16GB	4700	16	128	165	3,5	1150	2	2	0	1
47	Samsung Galaxy S3 Neo	4700	16	1536	306	4,8	2100	4	8	1	0
48	BenQ F5	4600	16	2048	294	5	2520	4	13	1	0
49	HTC Desire 516 Dual SIM	4500	4	1024	220	5	1950	4	5	1	0

50	Gigabyte Gsmart MIKA M3	4500	8	1024	294	5	1900	4	13	1	0
51	Alcatel One Touch Idol Alpha	4500	16	2048	312	4,7	2000	4	13	1	0
52	myPhone CUBE	4200	16	1024	220	5	1800	4	8	1	0
53	Sony Xperia E3	4200	4	1024	218	4,5	2330	4	5	1	0
54	Samsung Galaxy Grand Neo Duos	4000	8	1024	187	5	2100	4	5	1	0
55	Gigabyte Gsmart AKTA A4	3900	8	1024	220	5	1800	4	13	1	0
56	Sony Xperia M	3700	4	1024	245	4	1750	2	5	1	0
57	BenQ T3	3600	4	1024	245	4,5	1900	4	8	1	0
58	Gigabyte Gsmart Simba SX1	3500	4	1024	294	5	1900	2	13	1	0
59	Myphone Next-S	3200	4	1024	245	4,5	1800	4	8	1	0
60	Nokia Lumia 635	3200	8	512	218	4,5	1830	4	5	0	0
61	Microsoft Lumia 535 Dual SIM	3200	8	1024	220	5	1905	4	5	0	0
62	Sencor Element P501	3100	4	1024	196	5	2000	4	5	1	0
63	Samsung Galaxy S Duos 2	2900	4	768	233	4	1500	2	5	1	0
64	Samsung Galaxy Trend Plus	2700	4	768	233	4	1500	2	5	1	0
65	Nokia Lumia 630 Dual SIM	2500	8	512	218	4,5	1830	4	5	0	0
66	Lenovo A328	2500	4	1024	218	4,5	2000	4	5	1	0
67	Alcatel Onetouch POP D3	2300	4	512	233	4	1400	2	5	1	0
68	Nokia Lumia 530	2200	4	512	245	4	1430	4	5	0	0
69	Lenovo A319	2000	4	512	233	4	1500	2	5	1	0
70	GoCleaver Quantum 400	1700	4	512	233	4	1600	4	2	1	0