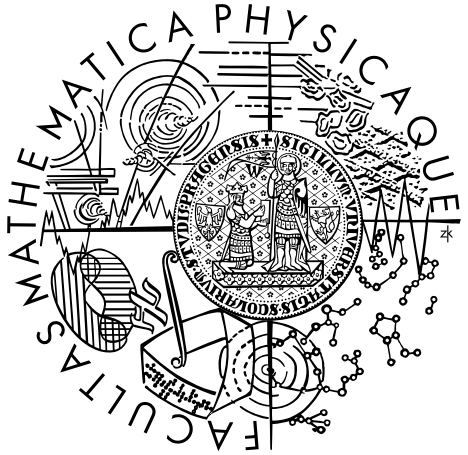
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Filip Šimsa

Kritéria těsnosti regrese dle typu vysvětlované proměnné

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	Mgr. Tomáš Hanzák
Studijní program:	Matematika
Studijní obor:	Obecná Matematika

Praha 2012

Na tomto místě bych rád poděkoval vedoucímu práce za odborné vedení mé bakalářské práce, trpělivost a věcné přípomínky. Neméně si cením velkého množství času a energie, které mi vedoucí práce věnoval.

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V Praze dne 1.8.2012 Filip Šimsa

Název práce: Kritéria těsnosti regrese dle typu vysvětlované proměnné
Autor: Filip Šimsa
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: Mgr. Tomáš Hanzák
E-mail vedoucího: tomas.hanzak@post.cz

Abstrakt: Práce se věnuje popisu modelů lineární, logistické, ordinální a multinomické regrese a interpretaci jejích parametrů. Dále zavádí různé ukazatele kvality modelu a vztahy mezi nimi. Soustředí se zejména na Giniho koeficient a koeficient determinace R^2. První zmíněný je zaveden pomocí modifikace Lorenzovy křivky pro ordinální a spojitou proměnnou a na základě porovnávání odhadnutých pravděpodobností pro proměnnou nominální. Koeficient determinace R^2 je nově definován pro nominální proměnnou, u které je zkoumán jeho vztah k Giniho koeficientu. Za předpokladu normálně rozdělených skóre a chyb modelu je numericky odvozena závislost mezi Giniho koeficientem a koeficientem determinace pro různá spojitá rozdělení vysvětlované proměnné. Teoretické výpočty a definice jsou ilustrovány na dvou sadách reálných dat.

Klíčová slova: Logistická regrese, Giniho koeficient, Lorenzova křivka, koeficient determinace

Title: Regression goodness-of-fit criteria according to dependent variable type
Author: Filip Šimsa
Department: Department of Probability and Mathematical Statistics
Supervisor: Mgr. Tomáš Hanzák
Supervisor's e-mail address: tomas.hanzak@post.cz

Abstract: This work is devoted to the description of linear, logistic, ordinal and multinominal regression models and interpretation of its parameters. Then it introduces a variety of quality indicators of mathematical models and the relations between them. It focuses mainly on the Gini coefficient and the coefficient of determination R^2. The first mentioned is established by modifying the Lorenz curve for ordinal and continuous variables and by comparing the estimated probabilities for nominal variable. The coefficient of determination R^2 is newly defined for the nominal variable and is examined its relationship with Gini coefficient. Assuming normally distributed scores and errors of the model is numerically derived the relation between the Gini coefficient and the coefficient of determiantion for different distribution of continuous dependent variable. Theoretical calculations and definitions are illustrated on two real data sets.

Keywords: Logistic regression, Gini coefficient, Lorenz curve, coefficient of determination.

Obsah

Úvod

V životě se často setkáváme se situacemi, které se opakují s různým výsledkem, a my bychom je rádi byli schopni předpovídat. Ať pracujeme jako bankéř a chceme odhadnout, jací lidé budou schopni splácet hypotéku, či jako doktor snažící se určit, jaký druh lidí je náchylný vůči nově propuklé nákaze nebo si jen chceme vsadit na to, jaká země vyhraje zlato na olympiádě, abychom maximalizovali naší šanci na výhru. Ve všech těchto případech jsme schopni zjistit určité informace. Bankéř bude klást dotazy na výši příjmu, stabilitu zaměstnání a vlastnictví jiných majetků. Doktor zjistí celkový zdravotní stav nakažených jedinců a bude hledat společné rysy. A my si můžeme zjistit výkony jednotlivých sportovců za poslední sezónů. Poté se ovšem vyskytuje otázka, jak moc jsou naměřené faktory důležité.

Právě k odhadu budoucích výsledků a k vyčíslení míry závislosti na zjištěných údajích slouží regresní analýza. Ta se snaží z určitého množství napozorovaných dat odhadnout vliv zjištěných faktorů, tzv. vysvětlujících proměnných na veličinu, kterou chceme predikovat. Ta se nazývá vysvětlovaná proměnná. Nejčastěji se setkáváme se situací, kdy vysvětlovaná proměnná je spojitá a používá se model lineární regrese.

Ovšem v bankovním sektoru se typicky snažíme vysvětlovat chování binární proměnné. V takovém případě nás zajímá pravděpodobnost nastání určitého jevu (klient splatí či nesplatí poskytnutý úvěr). Za těchto okolností se používá tzv. logistická regrese. Vyskytují se ovšem i případy, kdy vysvětlovaná proměnná je vícehodnotová. V závislosti na tom, zda její hodnoty jsou uspořádatelné, či nikoli, zpravidla volíme ordinální či nominální logistickou regresi. Ordinální proměnná také běžně vzniká kategorizací spojité proměnné, proto by se hodil společný ukazatel, který by oba odlišné přístupy dával do souvislosti. Popis a interpretace parametrů těchto rozdílných modelů je jedním z cílů této práce.

Po sestrojení určitého modelu bychom rádi zjistili, zda náš model dobře odpovídá napozorovaným datům, či nikoli. Za tímo účelem používáme různé ukazatele kvality modelu. V lineární regresi je nejběžněji používán koeficient determinace R^2, naproti tomu v regresi logistické se většinou používá Giniho koeficient. Dalším z cílů této práce je rozšíření těchto běžně používaných koeficientů i pro typy vysvětlované proměnné, kde nejsou definovány. K tomu nás motivuje fakt, že jak ordinální, tak nominální logistická regrese je běžnou součástí statistického softwaru (např. program R, SPSS) a pro měření kvality modelu se používají obecné pseudo koeficienty determinace, které nevyužívají specifické rysy těchto modelů. Často se setkáváme s ordinální proměnnou v různých dotazníkových šetřeních, kde jsme např. tázáni na to, jak moc jsme spokojeni s určitým produktem. Nedostatek matematických nástrojů pak může vést k upřednostnění binární logistické proměnné, která ponese jen informaci o spokojenosti, či nespokojenosti. Tento nedostatek se pokusíme alespoň částečně odstranit novými definicemi.

Struktura této práce je následující: v první kapitole si postupně popíšeme modely lineární, binární, multinomické a ordinální regrese a uvedeme interpretaci jejích parametrů. Dále zavedeme běžně používané ukazatele kvality modelu a připomeneme důležité vztahy mezi nimi. Ve stěžejní druhé kapitole nejprve zadefinujeme Giniho koeficient pro ordinální vysvětlovanou proměnnou. Uvedeme několik možných definic a budeme diskutovat jejich výhodnost, případně vztahy

mezi nimi.

V další části definujeme Giniho koeficient a koeficient determinace R^2 pro nominální vysvětlovanou proměnnou. Zavedeme dílčí binární proměnné, které vzniknou z našich dat seskupením všech kategorií vždy až na jednu. Poté oba nově definované koeficienty vyjádříme jako vážený průměr odpovídajících dílčích koeficientů pro tyto binární proměnné. Následně se budeme zabývat otázkou, kdy je R^2 při pevných hodnotách dílčích Giniho koeficientů maximální.

V poslední části druhé kapitoly zavedeme Giniho koeficient obdobně jako jednu z definic pro ordinální proměnnou. Opět použijeme analogii Lorenzovy křivky a vyjádříme Giniho koeficient pomocí pořadí. Poté náhlédneme na jeho vztah ke Spearmanovu korelačnímu koeficientu r_S. V závěru této kapitoly učiníme předpoklady o rozdělení skóre a chyb modelu. Poté pro určitá spojitá rozdělení vysvětlované proměnné budeme hledat závislost mezi Giniho koeficientem a koeficientem determinace R^2.

V třetí kapitole na dvou reálných datech aplikujeme modely ordinální a multinomické regrese. Vypočteme číselné hodnoty zavedených koeficientů a vyčíslíme aproximativní rovnosti. Pokusíme se doporučit využití definic u ordinální proměnné, kde spočteme i Giniho koeficient dle definice nominální, abychom ověřili, že využití informace ordinality má v jeho případě smysl.

Statistické testy jsou prováděny v programu R. K vyčíslení jednotlivých koeficientů a vytvoření grafů byl použit program Wolfram Mathematica 8.0.

1. Různé typy vysvětlované proměnné a kritéria těsnosti modelu

V této kapitole se seznámíme s různými typy vysvětlované proměnné a s rozdílnými modely snažícími se odhadnout jejich hodnotu na základě určitých informací, tzv. vysvětlujících proměnných (regresorů). Dále se budeme zabývat charakteristikami diverzifikace, na jejichž základě bychom rádi rozhodli, do jaké míry se náš model hodí na naše data. Než přejdeme k první podkapitole, podívejme se, jakých různých typů může být libovolná proměnná.

a) *Nominální proměnná*, pro její dvě libovolné hodnoty jsme schopni určit pouze, zda jsou stejné či rozdílné. Například: druh auta (osobní, nákladní, autobus) či obor studia (matematická analýza, matematické struktury, finanční matematika).

b) *Ordinální proměnná*, u jejíž libovolných dvou hodnot lze stanovit jejich pořadí. Například: Oblíbenost určité kapely (velmi oblíbená, oblíbená, neoblíbená, velmi neoblíbená) nebo ratingové hodnocení (AAA, AA, ..., D).

c) *Poměrová proměnná* je ta, pro jejíž dvě libovolné hodnoty můžeme určit, kolikrát je jedna větší než druhá. Tedy jedná se pouze o kladná čísla. Například: Počet členů domácnosti (1,2,..), hmotnost jedince.

d) *Intervalová proměnná* je taková, u níž lze pro každé dvě libovolné hodnoty stanovit, o kolik je jedna větší než druhá. Například: Teplota v Praze ve °C či zůstatek na kreditním kontě.

Poměrové a Intervalové proměnné se společně označují jako *kvantitativní* proměnné a dále je můžeme dělit podle hodnot, kterých mohou nabývat na

- *diskrétní*: nabývá jen konečně či spočetně mnoha hodnot
- *spojitá*: nabývá libovolné hodnoty z určitého intervalu

Nominální, ordinální a diskrétní proměnné se souhrnně nazývají *kategoriální* proměnné a podle počtu hodnot, kterých mohou nabývat, je lze nazvat

- *binární (dichotomická)*: nabývá jen dvou kategorií
- *multinomická (vícekategoriální)*: nabývá více než dvou kategorií

Obecně lze mezi jednotlivými typy proměnných přecházet, ale pouze jednostranně vždy při ztrátě určité informace. K tomu můžeme mít různé důvody, například zjednodušení při dotazníkovém šetření (věk pouze jako celá čísla) či pro rozdělení výběrové populace do několika skupin, které budou mít různý vliv na vysvětlovanou proměnnou. Typické je zejména kategorizování při zachování porovnatelnosti, neboť nám zůstane alespoň informace o pořadí. Přechod od ordinální k nominální proměnné je velmi neobvyklý.

Ilustrujme takový přechod na příkladu:

kvantitativní spojitá	ordinální	nominální
hmotnost jedince	váhové kategorie	jedinec v podváze, nadváze, ostatní

Kde kategorie ostatní zahrnuje jak lidi v normě, tak lidi obézní. Proto není ordinalita dodržena.

Před tím než přejdeme k rozebírání přístupu odhadování vysvětlované proměnné dle jejího typu podívejme se, jak se v modelech reprezentuje kategoriální vysvětlující proměnná x, která je v praxi běžná. V případě, že se jedná o dvouhodnotovou proměnnou (označovanou jako dummy proměnná) stačí ji kódovat čísly 1 a 0 podle toho, zda daný jev nastal, či nenastal a dále s ní zacházet jako s jinými regresory. Ovšem pro vícekategorionální již čiré zakódování nestačí, neboť mezi jednotlivými jevy nemusí být žádná souvislost. Proto ji zakódujeme jako lineární kombinaci dummy proměnných. Toto si ukážeme pro x nabývající čtyř různých hodnot $\{a, b, c, d\}$. Zadefinujeme tři dummy proměnné x_b, x_c, x_d následovně:

x	x_b	x_c	x_d
a	0	0	0
b	1	0	0
c	0	1	0
d	0	0	1

V obecném modelu pak místo βx_i bude $\beta_1 x_{bi} + \beta_2 x_{ci} + \beta_3 x_{di}$. Tento přístup tedy implicitně uvažuje, že nastal jev a a pokud tomu tak není, je kompenzován velikostí příšlušného parametru β_1, β_2 nebo β_3.

1.1 Kvantitativní vysvětlované proměnné

Jedná se o situaci, kdy se na základě regresorů snažíme odhadnout vysvětlovanou proměnnou, která zpravidla nabývá reálných hodnot z nějakého intervalu. V takovémto případě se závislost nejčastěji modeluje pomocí lineárního regresního modelu, který je lineární ve svých parametrech. Popišme si jeho teoretický základ. Nechť jsou dány nezávislé náhodné veličiny $Y_1, ..., Y_n$ a matice nenáhodných čísel $\boldsymbol{X}$ typu $n \times k$, $k < n$. Dále nechť sloupce matice $\mathbf{X}$ jsou nezávislé, tedy $h(\boldsymbol{X}) = k$. Označme $\boldsymbol{Y}' = (Y_1, ..., Y_n)$ a předpokládejme, že platí:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)'$ je vektor neznámých parametrů typu $k \times 1$ a $\boldsymbol{\varepsilon}' = (\varepsilon_1, ..., \varepsilon_n)$ je náhodný vektor typu $n \times 1$. Dále o $\boldsymbol{\varepsilon}$ předpokládáme, že $\mathsf{E}\,\boldsymbol{\varepsilon} = 0$ a $\mathsf{Var}\,\boldsymbol{\varepsilon} = \sigma^2 \boldsymbol{I}$, kde σ^2 je také neznámý parametr. Tedy $\mathsf{cov}\,(\varepsilon_i, \varepsilon_j) = 0$ a $\mathsf{var}\,\varepsilon_j = \sigma^2$, $i, j = 1, ...n, i \neq j$. Za těchto předpokladů platí $\mathsf{E}\,\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}$, $\mathsf{Var}\,\boldsymbol{Y} = \sigma^2 \boldsymbol{I}$.

V matici $\boldsymbol{X}$ se nejčastěji jako první sloupec volí sloupec jedniček, který společně s parametrem β_1 má význam absolutního členu (tzv. intercept). Parametry $\beta_2, ..., \beta_k$ mají přímočarou interpretaci, v případě, že se hodnota i-tého sloupce matice $\boldsymbol{X}$ zvýší o 1, pak se očekávaná hodnota $\boldsymbol{Y}$ změní o β_i.

Zabývejme se odhady parametrů $\boldsymbol{\beta}$, k tomu se používá metoda nejmenších čtverců. Tedy chceme minimalizovat výraz:

$$SSE(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{Y}'\boldsymbol{Y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{Y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}.$$

Před tím, než tento výraz zderivujeme, si uvědomíme, že $h(\boldsymbol{X}'\boldsymbol{X}) = k$, důkaz lze nalézt v [1], a tedy se zřejmě jedná o pozitivně definitní matici. Nyní již k samotným derivacím:

$$SSE'(\boldsymbol{\beta}) = -2\boldsymbol{X}'\boldsymbol{Y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{X}'\boldsymbol{Y} \tag{1.1}$$

$$SSE''(\boldsymbol{\beta}) = 2\boldsymbol{X}'\boldsymbol{X} \tag{1.2}$$

Výraz 1.1 položíme roven nule a dostaneme odhad $\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$, z pozitivní definitnosti matice $\boldsymbol{X}'\boldsymbol{X}$ pak plyne, že se jedná o minimum. Poznamenejme ještě, že tyto odhady se počítají z rovnic $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{Y}$, které se nazývají normální rovnice. Důvod, proč se výpočet provádí pomocí normálních rovnic, je nižší numerická náročnost, která je významná hlavně pro případy s více pozorováními a větším počtem regresorů. Přirozeně zvolíme $\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{b}$ jako odhad náhodné veličiny $\boldsymbol{Y}$. Ten se dá vyjádřit i pomocí symetrické, idempotentní matice $\boldsymbol{H} = \boldsymbol{X}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}$ takto $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$. Snadným výpočtem zjistíme $\mathsf{E}\,\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\mathsf{E}\,\boldsymbol{Y} = \boldsymbol{\beta}$, z čehož plyne, že se jedná o nestranný odhad parametru $\boldsymbol{\beta}$. V knize [5] lze nalézt důkaz Gauss-Markovovy věty, dle které je $\hat{\boldsymbol{Y}}$ dokonce nejlepším lineárním nestranným odhadem vektoru $\boldsymbol{X}\boldsymbol{\beta}$, který je roven $\mathsf{E}\,\boldsymbol{Y}$. Označme $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ a zavedeme další velice důležité veličiny:

Reziduální součet čtverců (residual sum of squares)

$$RSS = SSE(\boldsymbol{b}) = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})'(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{H}\boldsymbol{Y} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{1.3}$$

Úplný součet čtverců (total sum of squares)

$$TSS = (\boldsymbol{Y} - \boldsymbol{e}\overline{Y})'(\boldsymbol{Y} - \boldsymbol{e}\overline{Y}) = \boldsymbol{Y}'\boldsymbol{Y} - n\overline{Y}^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \tag{1.4}$$

a Vysvětlený součet čtverců (explained sum of squares)

$$ESS = (\hat{\boldsymbol{Y}} - \boldsymbol{e}\overline{Y})'(\hat{\boldsymbol{Y}} - \boldsymbol{e}\overline{Y}) = \hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}} - n\overline{Y}^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2. \tag{1.5}$$

V (1.5) jsme využili vztahu $\overline{\hat{Y}} = \overline{Y}$, který plyne z toho, že vektor samých jedniček ($\boldsymbol{e}$) je součástí regresní matice, tedy z normálních rovnic dostáváme: $\boldsymbol{e}'\hat{\boldsymbol{Y}} = \boldsymbol{e}'\boldsymbol{H}\boldsymbol{Y} = \boldsymbol{e}'\boldsymbol{Y}$. Dále uvedeme užitečnou rovnost mezi právě zavedenými veličinami.

$$\begin{aligned} ESS + RSS &= \hat{\boldsymbol{Y}}'\hat{\boldsymbol{Y}} - n\overline{Y}^2 + \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{H}\boldsymbol{Y} \\ &= \boldsymbol{Y}'\boldsymbol{H}'\boldsymbol{H}\boldsymbol{Y} - \boldsymbol{Y}'\boldsymbol{H}\boldsymbol{Y} + \boldsymbol{Y}'\boldsymbol{Y} - n\overline{Y}^2 = \boldsymbol{Y}'\boldsymbol{Y} - n\overline{Y}^2 = TSS. \end{aligned} \tag{1.6}$$

Zvláštní význam má RSS, který měří čtvercovou vzdálenost našeho odhadu od skutečných hodnot. Uveďme ještě, že veličina $S^2 = \frac{RSS}{n-k}$ je nestranným odhadem parametru σ^2.

Pro odvození testových statistik je třeba dále předpokládat, že rozdělení chyb má normální rozdělení, tj. $\boldsymbol{\varepsilon} \sim \mathsf{N}(0, \sigma^2 \boldsymbol{I})$, a tím pádem $\boldsymbol{Y} \sim \mathsf{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$. Jednotlivé statistiky pro testování hypotéz o významnosti jednotlivých parametrů β_i, či více parametrů najednou lze nalézt v [1].

Za předpokladu normality si ještě odvoďme odhady metodou maximální věrohodnosti, které využijeme v následující části.

$$l(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}\right)$$

$$L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \sigma^{-2}\boldsymbol{X}'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{1.7}$$

$$\frac{\partial L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} SSE(\boldsymbol{\beta}) \tag{1.8}$$

Při položení výrazů (1.7), (1.8) nule dospějeme k odhadům $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{Y}$ a odhad rozptylu $\hat{\sigma}^2 = \frac{1}{n} SSE(\hat{\boldsymbol{\beta}}) = \frac{1}{n} RSS$. Při dosazení těchto odhadů do logaritmické věrohodnostní rovnice dostaneme:

$$L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = -\frac{n}{2}\log\left(2\pi\frac{RSS}{n}\right) - \frac{n}{2RSS}RSS = -\frac{n}{2}(1 + \log 2\pi - \log n + \log(RSS)). \tag{1.9}$$

1.2 Kvalitativní vysvětlované proměnné

V případě, že náhodná veličina nabývá nějakého konečného počtu kategorií, nemá smysl ji modelovat pomocí regresní přímky, neboť odhadnutá přímka bude neomezená a tedy určité hodnoty $\boldsymbol{x}'\boldsymbol{\beta}$ nelze interpretovat. Z tohoto důvodu budeme hledat modely, které budou omezené. Nejčastěji se setkáváme s případy, kdy vysvětlovaná proměnná nabývá právě dvou kategorií (např. klient splatil, či nesplatil úvěr, po nasazení léků pacient zemřel, či nezemřel). Proto se nejdříve budeme věnovat pouze dvouhodnotové vysvětlované proměnné, jež nám poskytne základ k modelování ostatních situací.

1.2.1 Binární logistický model

Nechť máme nezávislé náhodné veličiny $Y_1, ..., Y_n$ s alternativním rozdělením, tj. $\mathsf{P}(Y_i = 1) = p_i$, $\mathsf{P}(Y_i = 0) = 1 - p_i$, $p_i \in (0, 1)$. Odhad střední hodnoty $\mathsf{E}\, Y_i$ je shodný s odhadem pravděpodobnosti p_i, neboť $\mathsf{E}\, Y_i = 1 \cdot p_i + 0 \cdot (1 - p_i) = p_i$. Uvažme dále matici regresorů $\boldsymbol{X}$ typu $n \times (k+1)$, $k+1 < n$, kde první sloupec je tvořen vektorem jedniček. Označme $\boldsymbol{x}_i{}' = (1, x_{i1}, ..., x_{ik})$, $i = 1, ..., n$ vektor

vysvětlujících proměnných odpovídající náhodné veličině Y_i. Naším cílem bude odhadnout $\mathsf{E}\,(Y_i|\boldsymbol{x}_i) = \pi(\boldsymbol{x}_i)$ a k tomu použijeme model logistické regrese:

$$\pi(\boldsymbol{x}_i) = \frac{e^{\boldsymbol{x}_i{}'\boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i{}'\boldsymbol{\beta}}} = \frac{1}{1 + e^{-\boldsymbol{x}_i{}'\boldsymbol{\beta}}} \tag{1.10}$$

kde $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)$ je vektor neznámých parametrů. Veličina $\pi(\boldsymbol{x}_i)$ nabývá hodnot mezi 0 a 1 pro všechny $(k+1)$ rozměrné vektory regresorů, a tudíž splňuje náš požadavek na omezenost. Nalezněme funkci hodnoty $\pi(\boldsymbol{x}_i)$, která bude lineární funkcí vektoru parametrů $\boldsymbol{\beta}$. K tomu si stačí povšimnout, že $\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)} = e^{\boldsymbol{x}_i{}'\boldsymbol{\beta}}$, a zlogaritmováním dostaneme požadovanou funkci

$$g(\boldsymbol{x}_i) = \log(\frac{\pi(\boldsymbol{x}_i)}{1 - \pi(\boldsymbol{x}_i)}) = \boldsymbol{x}_i{}'\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k,$$

která se běžně označuje jako logit. Tato funkce připomíná lineární regresní model, neboť je lineární ve svých parametrech a může nabývat všech reálných hodnot v závislosti na $\boldsymbol{x}_i$.

Podívejme se nyní na interpretaci parametrů β_i, k tomu nám poslouží podíl pravděpodobností $\omega(\boldsymbol{x}_i) = \frac{\mathsf{P}(Y_i=1|\boldsymbol{x}_i)}{\mathsf{P}(Y_i=0|\boldsymbol{x}_i)} = \frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}$, jenž se nazývá šance.

Pokud $\boldsymbol{x}_i = (1, 0, ..., 0)$, pak

$$\omega((1, 0, ..., 0)) = \frac{\frac{e^{\beta_0}}{1+e^{\beta_0}}}{\frac{1}{1+e^{\beta_0}}} = e^{\beta_0}.$$

Tedy koeficient β_0 je roven logaritmu šance neboli logitu pravděpodobnosti $\mathsf{P}\,(Y = 1|\boldsymbol{x}_i = (1, 0, ..., 0))$.

Dále uvažme vektor $\tilde{\boldsymbol{x}}_i = (1, x_{i1}, ..., x_{ij-1}, x_{ij} + 1, x_{ij+1}, ..., x_{ik})$, který vznikl z $\boldsymbol{x}_i$ přičtením jedničky k j-té složcea šance se dá vyjádřit jako

$$\omega(\tilde{\boldsymbol{x}}_i) = \frac{\frac{e^{\tilde{\boldsymbol{x}}_i{}'\boldsymbol{\beta}}}{1+e^{\tilde{\boldsymbol{x}}_i{}'\boldsymbol{\beta}}}}{\frac{1}{1+e^{\tilde{\boldsymbol{x}}_i{}'\boldsymbol{\beta}}}} = e^{\tilde{\boldsymbol{x}}_i{}'\boldsymbol{\beta}} = e^{\beta_j} e^{\boldsymbol{x}_i{}'\boldsymbol{\beta}} = e^{\beta_j}\omega(\boldsymbol{x}_i).$$

Poměr těchto dvou šancí (odds ratio) je $OR_i = \frac{\omega(\tilde{\boldsymbol{x}}_i)}{\omega(\boldsymbol{x}_i)} = e^{\beta_i}$. Tudíž v případě, že se hodnota x_{ij} zvýší o jedničku, se logaritmus šancí změní o β_i.

K odhadu parametrů β_i se zpravidla používá metoda maximální věrohodnosti, lze ovšem použít i metodu nejmenších čtverců. Obě metody jsou odvozeny a porovnány v [4]. Uveďme jen, že dávají velmi podobné výsledky. Jiný možný přístup odhadu koeficientů v případě, kdy máme jen jednu vysvětlující proměnnou a to kategoriálního typu, je založen na kontingenčních tabulkách. Data utřídíme do kontingenční tabulky a určíme poměr šancí pomocí relativních četností nastalých jevů. Hledaný odhad pak bude roven logaritmu tohoto poměru. Ovšem v běžném případě pracujeme s velkým počtem vysvětlujících proměnných různého typu, pak tento přístup není možný.

Uveďme si příklad symbolizující význam OR, nechť Y označuje, zda daná osoba trpí či netrpí rakovinou plic a budeme mít jen jeden regresor x určující, jestli daná osoba kouří, či nekouří. Pokud vyjde $\hat{\beta}_1 = 3$, pak poměr šancí $\hat{OR} = e^3 \doteq 20.09$ odhaduje, že rakovina plic se vyskytne přibližně s dvacetkrát větší pravděpodobností u kuřáků než u nekuřáků. Tato interpretace je pravdivá pouze v případě, že $\frac{1-\mathsf{P}(Y=1|x=0)}{1-\mathsf{P}(Y=1|x=1)} \approx 1$, neboť pak $OR \approx \frac{\mathsf{P}(Y=1|x=1)}{\mathsf{P}(Y=1|x=0)}$.

1.2.2 Multinomický logistický model

V této sekci zobecníme logistickou regresi pro vysvětlovanou proměnnou Y nabývající více než dvou kategorií bez jakéhokoli uspořádání. Kategorie kódujeme čísly $\{0, 1, ..., m\}$ a předpokládáme, že máme k vysvětlujících proměnných. V modelu budeme opět uvažovat intercept, a tedy vektor regresorů bude tvaru $\boldsymbol{x} = (1, x_1, ..., x_k)$. Model binární logistické regrese jsme založili na logitu $g(\boldsymbol{x}) = \log\left(\frac{\mathsf{P}(Y=1|\boldsymbol{x})}{\mathsf{P}(Y=0|\boldsymbol{x})}\right)$, nyní máme kategorií více a přímým rozšířením tedy bude použít m logitů. Je ovšem důležité si rozmyslet vůči jaké kategorii porovnávat. Přímým rozšířením je užít $Y = 0$ jako referenční. Pak i-tý logit bude tvaru:

$$g_i(\boldsymbol{x}) = \log\left(\frac{\mathsf{P}\left(Y = i|\boldsymbol{x}\right)}{\mathsf{P}\left(Y = 0|\boldsymbol{x}\right)}\right) = \boldsymbol{x}'\boldsymbol{\beta}_i = \beta_{i0} + \beta_{i1}x_1 + ... + \beta_{ik}x_k m,$$

kde $\boldsymbol{\beta}_i = (\beta_{i0}, ..., \beta_{ik})'$ je vektor neznámých parametrů, kterých je celkem $(k+1)m$. Nyní odvoďme vzorce pro podmíněné pravděpodobnosti $\mathsf{P}\left(Y = i|\boldsymbol{x}\right)$.

$$\mathsf{P}\left(Y = 0|\boldsymbol{x}\right) = \frac{1}{\frac{1}{\mathsf{P}(Y=0|\boldsymbol{x})}} = \frac{1}{\frac{\mathsf{P}(Y=0|\boldsymbol{x})+\mathsf{P}(Y=1|\boldsymbol{x})+...+\mathsf{P}(Y=m|\boldsymbol{x})}{\mathsf{P}(Y=0|\boldsymbol{x})}} = \frac{1}{1 + e^{g_1(\boldsymbol{x})} + ... + e^{g_m(\boldsymbol{x})}}$$

$$\mathsf{P}\left(Y = i|\boldsymbol{x}\right) = \frac{\frac{\mathsf{P}(Y=i|\boldsymbol{x})}{\mathsf{P}(Y=0|\boldsymbol{x})}}{\frac{1}{\mathsf{P}(Y=0|\boldsymbol{x})}} = \frac{e^{g_i(\boldsymbol{x})}}{1 + e^{g_1(\boldsymbol{x})} + ... + e^{g_m(\boldsymbol{x})}}$$

Opět velmi důležité budou jednotlivé šance, které jsou vztažené k referenční kategorii. Dostaneme tedy m různých šancí porovnávajících, že nastane jev $Y = i$ a ne jev $Y = 0$, vzorcem zapsáno $\omega_i(\boldsymbol{x}) = \frac{\mathsf{P}(Y=i|x)}{\mathsf{P}(Y=0|x)}$, které opět použijeme k interpretaci parametrů β_{ij}.

Pokud $\boldsymbol{x} = (1, 0, ..., 0)$, pak

$$\omega_i((1, 0, ..., 0)) = \frac{\frac{e^{\beta_{i0}}}{1+e^{\beta_{10}}+...+e^{\beta_{k0}}}}{\frac{1}{1+e^{\beta_{10}}+...+e^{\beta_{k0}}}} = e^{\beta_{i0}}$$

z toho vidíme, že β_{i0} je roven logaritmu šance ω_i při vektoru regresorů rovnému $(1, 0, ..., 0)$.

Uvažme vektor $\tilde{\boldsymbol{x}} = (1, x_1, ..., x_{j-1}, x_j + 1, x_{j+1}, ..., x_k)$, který vznikl z $\boldsymbol{x}$ přičtením jedničky k j-té složce (či změnou kategorie z 0 na 1). Pak se šance dá vyjádřit

$$\omega_i(\tilde{\boldsymbol{x}}) = e^{\tilde{\boldsymbol{x}}'\boldsymbol{\beta}_i} = e^{\beta_{ij}}e^{\boldsymbol{x}'\boldsymbol{\beta}_i} = e^{\beta_{ij}}\omega_i(\boldsymbol{x}),$$

a tedy poměr šancí je $OR_i = \frac{\omega_i(\tilde{\boldsymbol{x}})}{\omega_i(\boldsymbol{x})} = e^{\beta_{ij}}$. Tudíž v případě, že hodnota x_j se zvýší o jedničku, pak se logaritmus šancí změní o β_{ij}. Odhady parametrů se zpravidla provádí pomocí metody maximální věrohodnosti. Na závěr zmiňme, že pro $m = 1$ je multinomický model shodný s modelem binárním.

1.2.3 Ordinální logistický model

Pro kategoriální vysvětlovanou proměnnou s přirozeným pořadím můžeme samozřejmě použít multinomický model, který ovšem nebere v potaz pořadí jednotlivých kategorií, a proto nemusí být výhovující. Z tohoto důvodu zde uvedeme

jiné modely, které budou opět založené na logitech. Značení budeme používat stejné jako v předešlých sekcích.

Opět je velmi důležité si rozmyslet, jak jednotlivé logity sestrojit. Tentokráte se nám nabízí možností více. Nejjednodušší z nich je porovnávat sousední kategorie a logity definovat následovně:

$$h_i(x) = \log\left(\frac{\mathsf{P}\left(Y = i|\boldsymbol{x}\right)}{\mathsf{P}\left(Y = i-1|\boldsymbol{x}\right)}\right) = \boldsymbol{x}'\boldsymbol{\beta}_i.$$

Mezi takto sestrojeným modelem a multinomickým modelem je následující vztah

$$\begin{aligned}\log\left(\frac{\mathsf{P}\left(Y = i|\boldsymbol{x}\right)}{\mathsf{P}\left(Y = 0|\boldsymbol{x}\right)}\right) &= \log\left(\frac{\mathsf{P}\left(Y = 1|\boldsymbol{x}\right)}{\mathsf{P}\left(Y = 0|\boldsymbol{x}\right)}\right) + \log\left(\frac{\mathsf{P}\left(Y = 2|\boldsymbol{x}\right)}{\mathsf{P}\left(Y = 1|\boldsymbol{x}\right)}\right) + ...+ \\ &+ \log\left(\frac{\mathsf{P}\left(Y = i|\boldsymbol{x}\right)}{\mathsf{P}\left(Y = i-1|\boldsymbol{x}\right)}\right) = h_1(\boldsymbol{x}) + h_2(\boldsymbol{x}) + ... + h_i(\boldsymbol{x}) \\ &= \boldsymbol{x}'\boldsymbol{\beta}_1 + \boldsymbol{x}'\boldsymbol{\beta}_2 + ... + \boldsymbol{x}'\boldsymbol{\beta}_i = \boldsymbol{x}'(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 + ... + \boldsymbol{\beta}_i)\end{aligned}$$

Vidíme, že pokud vypočteme odhady koeficientů $\boldsymbol{\beta}_i$ v libovolném ze dvou modelů, jsme schopni dopočítat příslušné odhady pro model druhý.

Jinou možností, jak zavést logity a šance, je porovnávat pravděpodobnosti, že Y padne do i-té kategorie, a ne do kategorie nižší. Z čehož dostaneme:

$$q_i(\boldsymbol{x}) = \log\left(\frac{\mathsf{P}\left(Y = i|\boldsymbol{x}\right)}{\mathsf{P}\left(Y < i|\boldsymbol{x}\right)}\right) = \boldsymbol{x}'\boldsymbol{\beta}_i.$$

Výhodou tohoto modelu je možnost odhadovat parametry postupně při uvažování m binárních logistických regresí. Tedy v i-tém kroku lze ztotožnit kategorie kódované $0, 1, ..., i-1$ a jako druhou kategorii uvažovat i.

Nevýhodou obou předchozích modelů je velký počet parametrů, který je stejný jako v případě multinomického modelu. Proto pokud je to možné, tak se upřednostňuje model, který předpokládá stejné poměry šancí mezi všemi úrovněmi. Ten porovnává pravděpodobnosti, že Y padne do kategorie i-té nebo nižší ku pravděpodobnosti, že bude v kategorii vyšší.

$$r_i(\boldsymbol{x}) = \log\left(\frac{\mathsf{P}\left(Y \leq i|\boldsymbol{x}\right)}{\mathsf{P}\left(Y > i|\boldsymbol{x}\right)}\right) = \alpha_i - \boldsymbol{x}'\boldsymbol{\beta}$$

Zde bude $\boldsymbol{x} = (x_1, ..., x_k)$, neboť intercepty uvažujeme zvlášť a různé pro odlišné logity. V tomto modelu odhadujeme jen $m+k$ parametrů za předpokladu, že $\boldsymbol{x}$ má stejný vliv na všech m binárních vysvětlovaných proměnných vzniklých z našich dat tímto způsobem: $Y_j = 0$ pro $Y \leq j$ a s $Y_j = 1$ pro $Y > j$, $j = 0, 1, ..., m-1$.

1.3 Existující ukazatelé diverzifikace

Zatím jsme se zabývali různými regresními modely, které se snažily odhadnout vliv jistého počtu vysvětlujících proměnných na vysvětlovanou proměnnou.

Po sestrojení určitého modelu a odhadnutí jeho parametrů nejprve zjistíme, jaké vysvětlovanné proměnné jsou signifikantní a jaké je možné vynechat. Dále je ovšem klíčové zjistit, zda náš model popisuje situaci dostatečně dobře a odpovídá empiricky získaným datům. K tomu nám právě slouží různé diverzifikační ukazatelé, na jejichž základě se můžeme rozhodnout, zda je model dostatečný, či je třeba provést nějaké úpravy nebo dokonce zkusit model jiný.

V lineární regresi je nejpoužívanějším ukazatelem koeficient determinace R^2, který je definován pomocí vzorců (1.3), (1.4) následovně:

$$R^2 = 1 - \frac{RSS}{TSS}. \tag{1.11}$$

Jelikož RSS je reziduální součet čtverců, platí, že čím menší tento výraz je, tím přesnější máme model. Ze vztahu (1.6) již víme, že $R^2 = \frac{ESS}{TSS}$ a tedy R^2 nabývá hodnot od 0 do 1. Kde hodnoty 1 se nabývá, pokud $RSS = 0$ neboli pro naše data platí $Y_i = \hat{Y}_i \ \forall i = 1, ..., n$. Což odpovídá situaci, kdy se náš model shoduje se získanými daty. Naopak $R^2 = 0$ v případě, že $ESS = 0$, tedy $\hat{Y}_i = \overline{Y} \ \forall i = 1, ..., n$. Jedná se o triviální model, kdy znalost hodnot $\boldsymbol{X}$ nepřináší žádnou informaci o veličině $\boldsymbol{Y}$. Maximalizace koeficientu R^2 je stejná úloha jako minimalizace výrazu RSS, který minimalizujeme metodou nejmenších čtverců.

Spočtěme nyní druhou mocninu výběrového korelačního koeficientu $r_{\boldsymbol{Y},\hat{\boldsymbol{Y}}}$ mezi veličinami $\boldsymbol{Y}, \hat{\boldsymbol{Y}}$.

$$\begin{aligned} r^2_{\boldsymbol{Y},\hat{\boldsymbol{Y}}} &= \frac{\left(\sum_{i=1}^n (Y_i - \overline{Y})(\hat{Y}_i - \overline{\hat{Y}})\right)^2}{\sum_{i=1}^n (Y_i - \overline{Y})^2 \sum_{i=1}^n (\hat{Y}_i - \overline{\hat{Y}})^2} = \frac{((\boldsymbol{Y} - \boldsymbol{e}\overline{Y})'(\hat{\boldsymbol{Y}} - \boldsymbol{e}\overline{Y}))^2}{TSS \cdot ESS} \\ &= \frac{(\boldsymbol{Y}'\hat{\boldsymbol{Y}} - \boldsymbol{e}'\hat{\boldsymbol{Y}})^2}{TSS \cdot ESS} = \frac{(\boldsymbol{Y}'\hat{\boldsymbol{Y}} - \boldsymbol{Y}'\boldsymbol{Y} + \boldsymbol{Y}'\boldsymbol{Y} - n\overline{Y})^2}{TSS \cdot ESS} \\ &= \frac{(-RSS + TSS)^2}{TSS \cdot ESS} = \frac{ESS}{TSS} = R^2. \end{aligned} \tag{1.12}$$

Z tohoto vidíme, že při odhadu parametrů $\boldsymbol{\beta}$ metodou nejmenších čtverců se druhá mocnina výběrového korelačního koeficientu shoduje s koeficientem determinace R^2.

Koeficient determinace R^2 má ovšem i některé nevýhody. Jednou z nich je jeho nepoužitelnost v případě nezařazení vektoru samých jedniček do regresní matice, neboť poté se nerovnají průměry hodnot $\boldsymbol{Y}$ a $\hat{\boldsymbol{Y}}$ a tudíž neplatí vztah (1.6). Následkem toho nemůžeme obecně určit obor hodnot, kterých koeficient determinace může nabývat a tím pádem jeho hodnota nelze interpretovat. Další nevýhodou je, že po přidání nových regresorů do modelu se hodnota R^2 může jen zvýšit. Tedy při maximalizaci R^2 bychom dospěli ke složitým modelům s velkým počtem regresorů. Z tohoto důvodu se někdy dává přednost upravenému koeficientu determinace, který je definován jako

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{RSS}{TSS} = 1 - \frac{n-1}{n-k-1}(1 - R^2).$$

Ten se nazývá korigovaný koeficient determinace $\bar{R}^2$ a jeho hodnota může klesnout s přídáním nového regresoru. Protože $\frac{n-1}{n-k-1} \geq 1$ dostáváme, že $\bar{R}^2 \leq R^2$.

V případě binární vysvětlované proměnné se v praxi nejčastěji používá Giniho koeficient. Ten je definován pomocí tzv. Lorenzovy křivky. K jejímu zavedení musíme nejprve učinit určité předpoklady. Nechť máme pozorování y_i, $i = 1, ..., n$ a model, který se snaží predikovat hodnotu y_i na základě regresorů $\boldsymbol{x}_i$. Výstupem modelu budou čísla nazývaná skóre s_i, $i = 1, ..., n$. Čím vyšší je hodnota skóre s_i, tím spíše dle modelu nastane, že $y_i = 1$. K výpočtu skóre lze například použít binární logistický model (1.10). Označme $n_1 = \sum_{i=1}^{n} y_i$ a $n_0 = \sum_{i=1}^{n} 1 - y_i$ a definujme dvě empiricky získané distribuční funkce. První příslušná prvkům s $y_i = 1$ je definována jako

$$F_G(s) = \frac{1}{n_1} \sum_{i=1}^{n} I_{[-\infty,s]}(s_i) y_i$$

a druhá pro prvky $y_i = 0$:

$$F_B(s) = \frac{1}{n_0} \sum_{i=1}^{n} I_{[-\infty,s]}(s_i)(1 - y_i).$$

Nyní definujeme Lorenzovu křivku jako lineární lomenou čáru spojující body $[0, 0]$ a $[F_B(s_i), F_G(s_i)], i = 1, ..., n$. Jedná se o křivku ležící uvnitř jednotkového čtverce s krajními body $[0, 0]$ a $[1, 1]$. Plochu pod Lorenzovou křivkou a mezi stranami čtverce označme AUC a definujme Giniho koeficient následovně:

$$Gini = 1 - 2AUC. \tag{1.13}$$

Obor hodnot $Gini$ je $[-1, 1]$, přičemž hodnota 1 znamená dokonale diverzifikovaná data (po setřízení prvků dle velikosti skóre budou nejprve prvky s $y_i = 0$ a poté prvky s $y_i = 1$), 0 naopak žádnou schopnost diverzifikace. Záporné hodnoty nastávají pro převrácený model.

Ještě uveďme, jak lze spočítat plocha AUC. Její velikost se odvíjí od počtu prohození (skóre odpovídající prvku s hodnotou 1 je menší než skóre odpovídající prvku s hodnotou 0). Každé prohození znamená nárůst plochy AUC o $\frac{1}{n_0 n_1}$, v případě shody dvou skóre odpovídající prvkům s různou hodnotou y bude nárůst poloviční. Z tohoto pozorování dostáváme:

$$AUC = \frac{\sum_{i,y_i=0} \sum_{j,y_j=1} (1 + sgn(s_i - s_j))\frac{1}{2}}{n_0 n_1} \tag{1.14}$$

Jiný koeficient diverzifikace založený na Lorenzově křivce je Kolmogorov Smirnovova statistika měřící maximální vzdálenost Lorenzovy křivky od její diagonály vynásobenou číslem $\sqrt{2}$. K tomu, abychom ji dokázali explicitně vyjádřit, je třeba najít průsečík diagonály (označme jej $[a(s), a(s)]$) a na ní spuštěné kolmice procházející obecným kolbodem $[F_B(s), F_G(s)]$. Ta lze analyticky vyjádřit $y = -x + F_G(s) + F_B(s)$ a tedy $a(s) = \frac{1}{2}(F_B(s) + F_G(s))$. Z tohoto dostáváme:

$$\begin{aligned} KS &= \sqrt{2} \sup_{s \in \mathbb{R}} \sqrt{(F_G(s) - a(s))^2 + (F_B(s) - a(s))^2} \\ &= \sqrt{2} \sup_{s \in \mathbb{R}} \sqrt{(\frac{1}{2}(F_G(s) - F_B(s))^2 + (\frac{1}{2}(F_B(s) - F_G(s))^2} = \sup_{s \in \mathbb{R}} |F_B(s) - F_G(s)| \end{aligned} \tag{1.15}$$

Z definice je vidět, že $KS \in [0,1]$ a pomineme-li případy, kdy $\exists s \in \mathbb{R}$ takové, že: $F_G(s) > F_B(s)$, bude vyšší hodnota KS znamenat lepší schopnost diverzifikace modelu. Mezi KS a $Gini$ není pevný vztah, neboť záleží na tvaru Lorenzovy křivky.

Číselnou charakteristikou úzce spjatou s $Gini$ je C-statistika definována jako pravděpodobnost, že náhodně vybrané skóre s_i příslušející prvku s $y_i = 0$ je menší než náhodně vybrané skóre s_j odpovídající prvku $y_j = 1$. V případě stejného skóre příslušného prvkům s různou hodnotou se započítá $\frac{1}{2}$.

$$\begin{aligned} C &= \mathsf{P}\left(s_k < s_l | y_k = 0, y_l = 1\right) + \frac{1}{2}\mathsf{P}\left(s_k = s_l | y_k = 0, y_l = 1\right) \\ &= \sum_{i,y_i=0} \sum_{j,y_j=1} \frac{(1 + sgn(s_j - s_i))\frac{1}{2}}{n_0 n_1}. \end{aligned} \tag{1.16}$$

Z poslední rovnosti a vzorce pro AUC (2.3) vidíme, že $C = 1 - AUC$ a z toho dostáváme rovnost

$$Gini = 2C - 1 = \sum_{i,y_i=0} \sum_{j,y_j=1} \frac{sgn(s_j - s_i)}{n_0 n_1}. \tag{1.17}$$

Další běžně používaný koeficient je Spearmanův korelační koeficient. Obecně se používá pro náhodný výběr $(X_1, Y_1)', ..., (X_n, Y_n)'$ z dvourozměrného rozdělení, kdy známe jejich pořadí. Označme je $R_1, ..., R_n$ pro veličiny $X_1, ..., X_n$ a obdobně $Q_1, ..., Q_n$ pro $Y_1, ..., Y_n$. V případě shody několika hodnot jim přiřadíme průměrné pořadí. Analyticky lze tedy pořadí zapsat jako $R_i = \frac{1}{2} + \sum_{j=1}^{n} I_{[X_j < X_i]} + \frac{1}{2}\sum_{j=1}^{n} I_{[X_j = X_i]}$. Spearmanův korelační koeficient je definován jako výběrový korelační koeficient dvojic $(R_1, Q_1), ..., (R_n, Q_n)$, tedy:

$$\begin{aligned} r_S &= \frac{\sum_{i=1}^{n}(R_i - \overline{R})(Q_i - \overline{Q})}{\sqrt{\sum_{i=1}^{n}\left(R_i - \overline{R}\right)^2 \sum_{i=1}^{n}(Q_i - \overline{Q})^2}} \\ &= \frac{\sum_{i=1}^{n} R_i Q_i - n\overline{RQ}}{\sqrt{\left(\sum_{i=1}^{n} R_i^2 - n\overline{R}^2\right)\left(\sum_{i=1}^{n} Q_i^2 - n\overline{Q}^2\right)}}. \end{aligned} \tag{1.18}$$

Nás bude zajímat situace, kdy $X_i = s_i$ a $Y_i = y_i$. Pak pořadí prvků bude $Q_i = \frac{n_0+1}{2}$ pro $y_i = 0$ a $Q_i = n_0 + \frac{n_1+1}{2}$ pro $y_i = 1$. Při zavedení transformace $\tilde{Q}_i = (Q_i - \frac{n_0+1}{2})\frac{2}{n}$ se hodnota r_S nezmění a bude platit rovnost $y_i = \tilde{Q}_i$. To nám umožňuje vyjádřit Spearmanův korelační koeficient mezi s_i a y_i (předpokládáme, že skóre jsou navzájem různá, jinak by první rovnost platila jen aproximativně)

$$r_S = \frac{\sum_{i=1}^{n} R_i y_i - n\frac{n+1}{2}\frac{n_1}{n}}{\sqrt{\left(\frac{n(n+1)(2n+1)}{6} - n(\frac{n+1}{2})^2\right)\left(n_1 - n(\frac{n_1}{n})^2\right)}} = \sqrt{3}\frac{2\sum_{i=1}^{n} R_i y_i - n_1(n+1)}{\sqrt{(n+1)(n-1)n_0 n_1}}.$$

V práci [4] je odvozen přibližný vztah $r_S \approx \sqrt{3\overline{y}(1-\overline{y})}Gini$.

Při snaze definovat koeficient determinace R^2 v logistické regresi je dobré si uvědomit, jak lze $R^2 = 1 - \frac{RSS}{TSS}$ interpretovat v lineárním regresním modelu. Uvedeme zde tři různé náhledy

1. *Variabilita vysvětlená modelem*
   TSS udává variabilitu vysvětlované proměnné kolem svého průměru, zatímco RSS je variabilita vysvětlované proměnné, kterou se nepodařilo vysvětlit modelem. Čím více variability se podaří vysvětlit, tím máme lepší model.

2. *Zlepšení kvality predikce vysvětlované proměnné*
   TSS je suma čtvercových chyb při použití triviálního modelu (bez použití vysvětlujících proměnných, pouze za použití interceptu), zatímco RSS je suma čtvercových chyb použitého modelu. Celkově zlomek vyjadřuje zlepšení při použití vysvětlujících proměnných.

3. *Druhá mocnina výběrového korelačního koeficientu*
   využívá rovnosti $R^2 = r^2$, tedy se jedná o druhou mocninu korelace mezi regresory a regresandem.

V logistické regresi se definují koeficienty, které se dají interpretovat stejně jako R^2 podle některého z možných náhledů, a proto se označují jako pseudo R^2. Jejich společnou vlastností je, že nabývají hodnot z intervalu $[0, 1]$ (ne nutně všech) a vyšší číslo znamená lepší model. Označme si $\hat{\pi}_i$ odhad pravěpodobnosti $\mathsf{P}(Y = 1|\boldsymbol{x}_i)$, $l(\boldsymbol{b}) = \prod_{i=1}^{n} \hat{\pi}_i^{Y_i}(1 - \hat{\pi}_i)^{(1-Y_i)}$ je odhad věrohodnostní funkce a $L(\boldsymbol{b}) = \sum_{i=1}^{n} (Y_i \log \hat{\pi}_i + (1 - Y_i) \log(1 - \hat{\pi}_i))$ odhaduje logaritmus věrohodnostní funkce. Pokud v modelu použijeme pouze intercept, příslušné veličiny označíme $l(0)$, $L(0)$.

- *Efronův koeficient determinace* $R_e^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{\pi}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$
  Jedná se o analogickou definici jako v lineární regresi za použití RSS, TSS. V práci [4] lze nalézt, že přibližně platí rovnosti (1.6) a (1.12). Z čehož vidíme, že lze přibližně interpretovat podle všech tří bodů uvedených výše.

- *McFaddenův koeficient determinace* $R_{MF}^2 = 1 - \frac{L(\boldsymbol{b})}{L(0)}$
  Porovnává věrohodnost při použití modelu s použitými regresory oproti triviálnímu modelu. Lze zavést obdobu korigovaného koeficientu korelace vzorcem $\bar{R}_{MF}^2 = 1 - \frac{L(\boldsymbol{b})-k}{L(0)}$, kde k je počet regresorů. V tomto případě může ovšem nabývat i záporných hodnot.

- *Cox and Snellův koeficient determinace* $R_{CS}^2 = 1 - \left(\frac{l(\boldsymbol{b})}{l(0)}\right)^{-\frac{2}{n}}$
  Je založen na skutečnosti, že v lineární regresi lze logaritmická věrohodnostní rovnice po dosazení odhadů zapsat ve tvaru (1.9). A pro triviální model platí: $L(0) = -\frac{n}{2}(1 + \log 2\pi - \log n + \log(TSS))$ a tedy

  $$\log l(\boldsymbol{b}) - \log l(0) = L(\boldsymbol{b}) - L(0) = -\frac{n}{2}(\log RSS - \log TSS)$$
  $$\left(\frac{l(\boldsymbol{b})}{l(0)}\right)^{-\frac{2}{n}} = \frac{RSS}{TSS}$$

  Podíl věrohodností ukazuje zlepšení modelu oproti modelu triviálnímu. Jelikož maximální možná věrohodnost je rovna jedné, nemůže hodnota tohoto koeficientu přesáhnout $(1 - l(0))^{\frac{2}{n}}$. Z tohoto důvodu se zavedl upravený koeficient nazvaný *Naggelkerkův* definovaný $R_N^2 = \frac{R_{CS}^2}{(1-l(0))^{\frac{2}{n}}}$, který může nabývat všech hodnot od nuly do jedné.

Na závěr kapitoly se podívejme na vztahy výše uvedených koeficientů v situaci, kde budeme uvažovat spojitou náhodnou veličinu X, jejíž hodnoty jsou analogií skóre (jedná se o jedno reálné číslo x_i, které v praxi lze vypočítat pomocí regresorů $\boldsymbol{x}_i$ a odhadnutých koeficientů $\boldsymbol{\beta}$ z modelu (1.10) jako $x_i = \boldsymbol{x}_i\boldsymbol{\beta}$) a diskrétní náhodnou veličinu Y s alternativním rozdělením. Dále předpokládejme, že:

$$\mathcal{L}(X|Y=0) = \mathsf{N}(0,1)$$
$$\mathcal{L}(X|Y=1) = \mathsf{N}(D,1)$$

kde $D > 0$, což znamená pozitivní korelaci mezi X a Y. Hustoty těchto náhodných veličin budou $\varphi_G(x) = \varphi(x)$ a $\varphi_B(x) = \varphi(x-D)$. Pro odvození dalších vztahů je předpoklad normality a shodnost rozptylů v podmíněných rozděleních nutný. Naopak volba středních hodnot nijak neubírá na obecnosti. Dále si označme $G = \mathsf{P}(Y=1)$ a $B = \mathsf{P}(Y=0)$ a předpokládejme, že $G > 0$ a $B > 0$. Nyní již můžeme vyjádřit KS dle definice jako

$$KS = \sup_{s\in\mathbb{R}}(\Phi(s) - \Phi(s-D)) = 2\Phi\left(\frac{D}{2}\right) - 1, \tag{1.19}$$

neboť $(\Phi(s) - \Phi(s-D))' = e^{-\frac{s^2}{2}} - e^{-\frac{(s-D)^2}{2}} = 0 \Leftrightarrow s = \frac{D}{2}$ a jedná se o funkci konkávní, tedy jsme opravdu nalezli maximum.

Pro odvození C-statistiky vyjdeme z definice a využijeme znalosti, že lineární kombinace dvou nezávislých náhodných veličin s normálním rozdělením bude mít opět normální rozdělení. Označme $X_0 \sim \mathcal{L}(X|Y=0)$ a $X_1 \sim \mathcal{L}(X|Y=1)$, pak C-statistika je:

$$C = \mathsf{P}(X_0 < X_1) = P\left(\frac{X_0 - X_1 + D}{\sqrt{2}} < \frac{D}{\sqrt{2}}\right) = \Phi\left(\frac{D}{\sqrt{2}}\right).$$

Pro vyjádření Giniho koeficientu použijeme jeho vztah k C-statistice a dostáváme

$$Gini = 2C - 1 = 2\Phi\left(\frac{D}{\sqrt{2}}\right) - 1. \tag{1.20}$$

Nyní k určení vztahu KS a $Gini$ stačí z jedné z rovnic (1.19) nebo (1.20) vyjádřit D a dosadit do druhé. Tedy například $D = \sqrt{2}\Phi^{-1}\left(\frac{Gini+1}{2}\right)$ a

$$KS = 2\Phi\left(\frac{1}{\sqrt{2}}\Phi^{-1}\left(\frac{Gini+1}{2}\right)\right) - 1.$$

Navíc platí přibližná rovnost $KS \approx \frac{\sqrt{2}}{2}Gini$, která je poměrně přesná pro hodnoty $Gini < 0,6$.

Pro zkoumání vzájemného vztahu koeficientu determinace a Giniho koeficientu se nejvíce hodí R^2 podle Efronovy definice, neboť platí $R^2 \approx r^2$, kde $r = \mathsf{corr}(Y, \hat{Y})$. Ovšem v naší situaci je třeba si ujasnit, jak se vypočítá $\pi(x)$, RSS, TSS. Podrobný postup lze nalézt v [4], pro představu se dá postupovat následovně

1. Pomocí Bayesovy věty bychom vypočetli $\pi(x) = \mathsf{P}(Y=1|X=x)$.

2. Analogii $RSS = \sum_{i,y_i=0} \pi(x)^2 + \sum_{i,y_i=1} (1-\pi(x))^2$ zavedli jako

$$RSS = \int B\pi(x)^2 \varphi_B(x)\mathrm{d}x + \int G(1-\pi(x))^2 \varphi_G(x)\mathrm{d}x.$$

3. Úplný součet čtverců, pak dostaneme jako speciální případ, kdy $\pi(x) = \mathsf{P}(Y=1) = G$ a platí tedy $TSS = BG^2 + GB^2 = GB = G(1-G)$.

Po úpravách a výpočtech dostaneme:

$$\pi(x) = \frac{G}{G + B\exp\left(\frac{D^2}{2} - xD\right)}$$
$$R^2 = 1 - \frac{1}{G}\int_{-\infty}^{\infty} \pi(x)\varphi(x)\mathrm{d}x.$$

Nyní stačí použít rovnosti $D = \sqrt{2}\Phi^{-1}\left(\frac{Gini+1}{2}\right)$ a dostaneme vztah mezi R^2 a $Gini$. Jedná se o velmi komplikovaný vzorec, a proto využijeme aproximativních vztahů $R^2 \approx r^2$ a $r_S \approx \sqrt{3G(1-G)}Gini$ odvozených dříve. Nyní by nás zajímal vztah mezi r_S a r. V knize [1] nalezneme, že v případě, kdy $(X_1, Y_1), ..., (X_n, Y_n)$ je výběr z regulárního dvourozměrného normálního rozdělení, platí přibližná rovnost $r \approx 2\sin\left(\frac{\pi}{6}r_S\right)$. S využitím Taylorova vzorce dostaneme

$$r \approx 2\sin\left(\frac{\pi}{6}r_S\right) = \frac{\pi}{3}r_S + o(r_S^2) \approx r_S. \tag{1.21}$$

V práci [4] byla tato přibližná rovnost ukázána i na naší situaci. Odtud vidíme, že pro R^2 a $Gini$ platí přibližný vztah

$$R^2 \approx 3G(1-G)Gini^2. \tag{1.22}$$

Poněkud překvapující může být, že vztah významně závisí na $G = \mathsf{P}(Y=0)$ a tedy pro nízkou či vysokou pravděpodobnost úspěchu můžeme dostat velice rozdílné hodnoty koeficientů R^2 a $Gini$.

2. Giniho koeficient a koeficient determinace

Jak jsme již uvedli, hlavní ukazatel vhodnosti modelu v lineární regresi je koeficient determinace R^2. Podobně v logistické binární regresi se nejběžněji používá Giniho koeficient ($Gini$). Z tohoto důvodu bychom rádi rozšířili jejich možné definice i pro případy, kde nejsou definovány.

2.1 Gini pro ordinální vysvětlovanou proměnnou

V této části uvedeme tři možné definice Giniho koeficientu pro ordinální vysvětlovanou proměnnou Y a ukážeme jeho základní vlastnosti. Nejprve začneme s definicí, kde budou hrát kličovou roli pozorované kategorie Y. V druhé pak jejich pořadí, což nám umožní nalézt vztah ke Spearmanovu korelačnímu koeficientu. Třetí bude založená na rozšířené definici C-statistiky. Vlastnosti jednotlivých $Gini$ pro první dvě možné definice budou velice obdobné, proto výklad při druhé z nich zkrátíme.

V modelu budeme předpokládat kódování jednotlivých kategorií čísly od 0 do m. Dále nechť máme nezávislá pozorování y_i, $i = 1, ..., n$ a každá kategorie je v našem výběru zahrnuta alespoň jednou, symbolicky zapsáno $\forall i \in \{0, 1, ..., m\} \exists j \in \{1, ..., n\} : y_j = i$. Uvažujme matematický model, který se snaží predikovat hodnotu ordinální proměnné y_i na základě nespecifikovaných regresorů. Výstupem modelu je reálné číslo s_i nazývané skóre.

2.1.1 Gini pomocí hodnot y_i

Stejně jako pro $Gini$ u binární proměnné nejprve sestrojíme obdobu Lorenzovy křivky a na jejím základě definujeme příslušný koeficient. Na rozdíl od Giniho pro binární vysvětlovanou proměnnou definujeme pouze jednu distribuční funkci. Označme n_i, $i = 0, 1, ..., m$ počet pozorování, kdy $y_j = i$, vzorcem zapsáno $n_i = \sum_{j=1}^{n} I_{[y_j=i]}$. Dále označme $N = \sum_{i=1}^{m} i \cdot n_i = \sum_{i=1}^{n} y_i$ a $N_k = \sum_{i=0}^{k} n_i$. Nyní definujme funkci F předpisem:

$$F(x) = \frac{1}{N} \sum_{i=1}^{n} I_{(-\infty,x]}(s_i) y_i, \ x \in \mathbb{R}.$$

Povšimněme si nejprve základních vlastností funkce F. Pro $x > \max_{i \in \{1,...,n\}} s_i$ je hodnota $F(x) = \frac{1}{N} \sum_{i=1}^{n} y_i = 1$ a skoky (body, kde funkce F není spojitá) mohou nastat v bodech s_i, $i = 1, ..., n$.

K tomu, abychom mohli sestrojit analogii Lorenzovy křivky, je potřeba zadefinovat ještě jednu pomocnou funkci předpisem:

$$G(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(s_i). \tag{2.1}$$

Spojením bodů $[0,0]$ a $[G(s_i), F(s_i)]$, kde $i = 1, ..., n$ dostáváme klíčovou křivku (nazvěme ji L-křivka) ležící uvnitř jednotkového čtverce. Větší přimykavost L-křivky ke stěnám čtverce, stejně jako u Lorenzovy křivky, znamená lepší diverzifikační sílu modelu. Ovšem v našem případě pro dokonalý (resp. převrácený) model (tj. model, kde po uspořádání hodnot y_i dle velikosti skóre s_i budou seřazeny od nejmenší po největší (resp. od největší po nejmenší)) bude mít L-křivka právě $m-1$ zlomů a to v bodech $s_{N_k}, k = 0, ..., m-1$. Tuto křivku označme L_{id} (resp. L_{-id}). Plochu mezi touto křivkou a hranami jednotkového čtverce označme K.

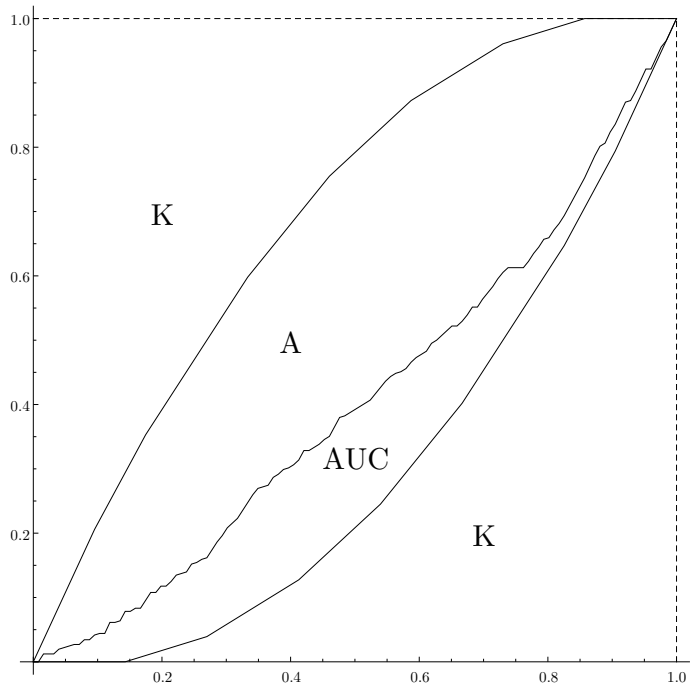
Velikost této plochy je:

$$K = \frac{1}{nN}\left(\sum_{i=0}^{m-1}(n - N_i - \frac{n_{i+1}}{2})n_{i+1}(i+1)\right) = \frac{M}{nN},$$

kde $M = \sum_{i=0}^{m-1}(n - N_i - \frac{n_{i+1}}{2})n_{i+1}(i+1)$. Dále označme plochu mezi L-křivkou a L_{id} (resp. L_{-id}) jako AUC (resp. A). Pak zřejmě platí $AUC + A + 2K = 1$. A proto definujme Giniho koeficient následujícím způsobem:

$$Gini = 1 - \frac{2AUC}{1-2K}. \tag{2.2}$$

Z definice vidíme, že $Gini \in [-1, 1]$, přičemž hodnota 1 znamená dokonalou diverzifikační schopnost modelu, naopak 0 žádnou.



Obrázek 2.1: L-křikvy

Nyní si ukážeme, jak lze spočítat plocha AUC bez použití funkce F. K tomu si stačí všimnout, že v případě záměny hodnot skóre s_i a s_j, příslušné prvkům y_i a y_j $(y_i < y_j)$, se AUC změní o hodnotu $\frac{|y_j - y_i|}{N \cdot n}$. Klesne (resp. stoupne)

v případě, kdy po záměně bude $s_i < s_j$ (resp. $s_i > s_j$). Pro $i < j$ definujme $P_{i,j} = \sum_{l,y_l=j} \sum_{k,y_k=i} \frac{1+\mathrm{sgn}(s_k - s_l)}{2}, i, j \in 1, ..., m$. Číslo $P_{i,j}$ udává počet skóre odpovídající prvkům s hodnotou i vyšších než skóre odpovídající prvkům s hodnotou j. Pak plochu AUC spočteme následovně

$$AUC = \sum_{j=1}^{m} \sum_{i=0}^{j-1} P_{i,j} \frac{j-i}{Nn}. \tag{2.3}$$

Dále odvoďme shodnost definic Giniho koeficientů v případě binární proměnné y_i. Pro $m = 1$ vyjde $K = \frac{n_1}{2n}$ a tedy $\frac{1}{1-2K} = \frac{n}{n_0}$, $AUC = P_{0,1}\frac{1}{n_1 n}$ a z toho dostáváme $Gini = 1 - \frac{2P_{0,1}}{n_0 n_1}$. Zbývá si uvědomit, že $\frac{P_{0,1}}{n_0 n_1}$ odpovídá ploše pod grafem Lorenzovy křivky. Což je jistě pravda, neboť se jedná o shodný výraz jako (1.14). Tedy naše definice je opravdu konzistentní s definicí Giniho koeficientu pro binární vysvětlovanou proměnnou.

Další vlastností Giniho koeficientu je možnost jeho vyjádření pomocí dílčích Giniho koeficientů mezi všemi dvojicemi hodnot $(i, j), i < j, i,j \in \{1, ..., m\}$, které označme $Gini_{i,j}$. Pro představu si z uspořádané n-tice skóre $(s_1 < s_2 < ... < s_n)$ vytvoříme uspořádanou $n_i + n_j$-tici skóre odpovídající prvkům s hodnotami i a j $(\tilde{s}_1 < \tilde{s}_2 < ... < \tilde{s}_{n_i+n_j})$ a odpovídající binární vysvětlovanou proměnnou $\tilde{Y}$. S využitím konzistence definic lze vyjádřit $Gini_{i,j} = 1 - 2AUC_{i,j} = 1 - 2\frac{P_{i,j}}{n_i n_j}$, tudíž $P_{i,j} = n_i n_j \frac{1-Gini_{i,j}}{2}$. Tento vztah dosadíme do (2.3) a dostáváme

$$\begin{aligned}
AUC &= \sum_{j=1}^{m} \sum_{i=0}^{j-1} n_i n_j \frac{1 - Gini_{i,j}}{2} \frac{j-i}{Nn} \\
Gini &= 1 - \frac{2AUC}{1-2K} = 1 - \frac{1}{nN - 2M} \sum_{j=1}^{m} \sum_{i=0}^{j-1} n_i n_j (1 - Gini_{i,j})(j-i) \\
&= \frac{1}{nN - 2M} \sum_{j=1}^{m} \sum_{i=0}^{j-1} n_i n_j Gini_{i,j}(j-i).
\end{aligned}$$

Povedlo se nám vyjádřit Giniho koeficient jako vážený průměr dílčích $Gini$ a vidíme, že větší význam na hodnotu Giniho koeficientu mají především dílčí $Gini$ pro vzdálené a více zastoupené kategorie.

Dále se inspirujeme v práci [4] a pokusíme se odvodit podobný vzorec pro výpočet Giniho koeficientu založený na pořadí. Seřaďme hodnoty y_i, $i = 1, ..., n$ podle velikosti skóre od nejmenší po největší a přiřaďme prvkům jejich pořadí v tomto uspořádání. Pořadí prvku y_i označme R_i. Tedy jestliže je pro i-tý prvek s_i nejmenší (resp. největší) bude $R_i = 1$ (resp. $R_i = n$). Pokud se některá skóre shodují, přiřadíme všem příslušným prvkům průměrné pořadí. V následujících vzorcích bude S_n značit symetrickou grupu permutací na n-prvkové množině. Nyní definujeme

$$S = \sum_{i=1}^{n} R_i y_i, \tag{2.4}$$

$$S_M = \max_{\pi \in S_n} \sum_{i=1}^{n} i y_{\pi(i)} = \sum_{i=0}^{m-1} \sum_{j=N_i+1}^{N_{i+1}} j(i+1) = \frac{1}{2}N + \frac{1}{2}\sum_{i=0}^{m-1}(i+1)n_{i+1}(n_{i+1}+2N_i), \tag{2.5}$$

$$\begin{aligned} S_m &= \min_{\pi \in S_n} \sum_{i=1}^{n} i y_{\pi(i)} = \sum_{i=0}^{m-1} \sum_{j=n-N_{m-i-1}+1}^{n-N_{m-i-1}} (m-i)j, \\ &= nN + \frac{1}{2}\sum_{i=0}^{m-1}(i+1)n_{i+1}(n_i + 1 - 2N_{i+1}), \end{aligned} \tag{2.6}$$

$$S_0 = \frac{S_M + S_m}{2} = \frac{n+1}{2}N. \tag{2.7}$$

Hodnota S vyjadřuje diverzifikační schopnost modelu, čím je S vyšší, tím přesnější model máme. Maximální (resp. minimální) možnou hodnotou S je S_M (resp. S_m), což odpovídá situaci, kdy jsou všechna pozorování setřízena od nejmenší hodnoty po největší (resp. od největší po nejmenší). V obou případech jsou data dokonale diverzifikována, a proto pro $S = S_M$ bude $Gini = 1$ a pro $S = S_m$ bude $Gini = -1$, neboť se jedná o převrácený model.

Klíčové pro vyjádření $Gini$ pomocí výše zmíněných veličin je si uvědomit rovnost $S_M - S = \sum_{j=1}^{m}\sum_{i=0}^{j-1} P_{i,j}(j-i)$. Ta se dá dokázat například indukcí podle počtu prohození $\sum_{j=1}^{m}\sum_{i=0}^{j-1} P_{i,j}$. Nyní již stačí úpravami odvodit vztah $S_M - S_m = (1-2K)Nn$ a dostáváme:

$$Gini = 1 - \frac{2AUC}{1-2K} = 1 - 2\frac{\sum_{j=1}^{m}\sum_{i=0}^{j-1} P_{i,j}(j-i)}{(1-2K)Nn} = 1 - 2\frac{S_M - S}{S_M - S_m} = \frac{S - S_0}{S_M - S_0}.$$

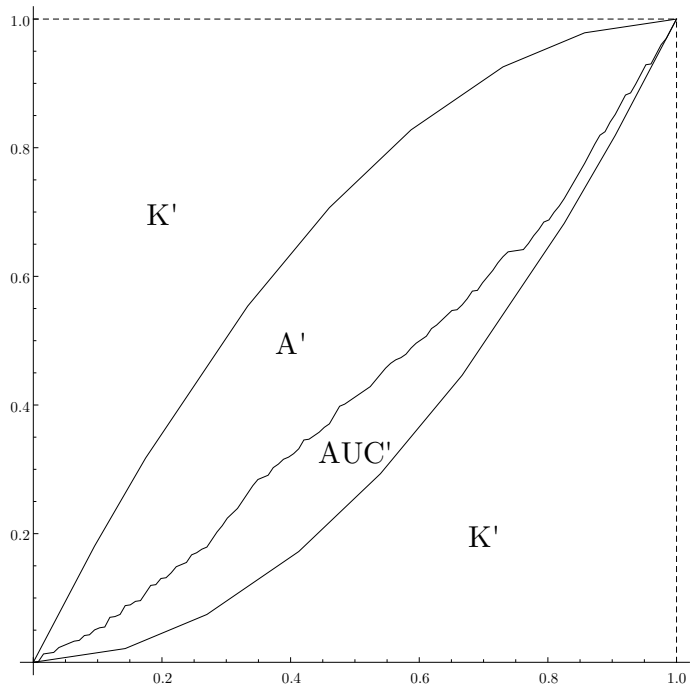
Vidíme, že se jedná o rozšíření vztahu pro ordinální vysvětlovanou proměnnou.

2.1.2 Gini pomocí pořadí y_i

Jinou možností jak zavést Giniho koeficient je pomocí pořadí náhodných veličin y_i, nikoli pomocí jejích hodnot. Proto označme Q_i pořadí prvku y_i v uspořádané n-tici prvků y_j dle velikosti, tj. $Q_i = \frac{1}{2} + \sum_{j=1}^{n}\left(I_{(-\infty,y_i)}(y_j) + \frac{1}{2}I_{[y_i=y_j]}\right) = \sum_{j=0}^{y_i-1} n_j + \frac{1}{2}(n_{y_i}+1)$. Dále označme $U_i = \sum_{j=0}^{i-1} n_j + \frac{1}{2}(n_i+1)$, $i = 1, ..., m$ a $N' = \frac{n(n+1)}{2}$. Distribuční funkci F' definujeme následujícím způsobem:

$$F'(x) = \frac{1}{N'}\sum_{i=1}^{n} I_{(-\infty,x]}(s_i)Q_i, \ x \in \mathbb{R}.$$

Pomocnou funkci G zadefinujeme stejně jako v (2.1) a L-křivku dostaneme spojením bodů $[0,0]$ a $[G(s_i), F(s_i)]$, $i = 1, ..., n$. L-křivky pro ideální a převrácený model označme L'_{id} a L'_{-id}. Značení pro další klíčové veličiny zůstává obdobné, viz obrázek (modelováno na stejná data jako dříve).



Obrázek 2.2: L'-křikvy

Plocha K' lze spočítat pomocí veličin U_i jako součet obsahů lichoběžníků$K' = \frac{1}{2N'n}\sum_{i=0}^{m}\left(n_i^2 U_i + 2n_i\sum_{j=0}^{i-1} U_j n_j\right) = \frac{M'}{N'n}$. A obměněný Giniho koeficient definujeme jako

$$Gini' = 1 - \frac{2AUC'}{1-2K'}.$$

Klíčovou plochu AUC' lze opět vyjádřit pomocí počtu všech prohození $P_{i,j}$ jako $AUC' = \sum_{j=1}^{m}\sum_{i=0}^{j-1} P_{i,j}\frac{U_j - U_i}{N'n}$. Ukažme, že i tato definice je konzistentní s definicí Giniho koeficientu pro binární vysvětlovanou proměnnou. K tomu stačí spočítat, že pro $m = 1$ je $1 - 2K' = \frac{n_0 n_1}{2N'}$ a $AUC' = \frac{P_{0,1}}{2N'}$ a tedy $Gini' = 1 - 2\frac{P_{0,1}}{n_0 n_1}$. Již dříve jsme si rozmysleli, že výraz $\frac{P_{0,1}}{n_0 n_1}$ odpovídá ploše pod Lorenzovou křivkou, tím je konzistence dokázána.

Díky této vlastnosti, lze $Gini'$ vyjádřit pomocí dílčích Giniho koeficientů mezi všemi dvojicemi hodnot (i, j), $i < j$, $i, j \in \{1, ..., m\}$. Počet prohození mezi prvky nabývající hodnot i a j je $P_{i,j} = n_i n_j \frac{1 - Gini_{i,j}}{2}$ a dostaneme:

$$\begin{aligned}
AUC &= \frac{1}{N'n}\sum_{j=1}^{m}\sum_{i=0}^{j-1} n_i n_j \frac{1 - Gini_{i,j}}{2}(U_j - U_i) \\
Gini' &= 1 - \frac{1}{N'n - 2M'}\sum_{j=1}^{m}\sum_{i=0}^{j-1} n_i n_j (1 - Gini_{i,j})(U_j - U_i) \\
&= \frac{1}{N'n - 2M'}\sum_{j=1}^{m}\sum_{i=0}^{j-1} n_i n_j Gini_{i,j}(U_j - U_i).
\end{aligned}$$

Opět se jedná o velmi podobný vzorec jako v první definici, kde hlavní vliv na Giniho koeficient mají dílčí $Gini_{i,j}$ pro vzdálené a více zastoupené kategorie. Vý-

hodou $Gini'$ je jeho neměnnost na vytovoření nové, méně početné kategorie, či sloučení málo početných kategorií, neboť hodnoty U_i se zásadně nezmění.

Pro vyjádření $Gini'$ pomocí pořadí je opět třeba modifikovat definice veličin (2.4), (2.5), (2.6), (2.7). Symbolem R_i budeme opět označovat pořadí skóre s_i. Zavedeme

$$S' = \sum_{i=1}^{n} R_i Q_i, \tag{2.8}$$

$$S'_M = \max_{\pi \in S_n} \sum_{i=1}^{n} iQ_{\pi(i)} = \sum_{j=0}^{m} \sum_{i=N_{j-1}+1}^{N_j} iU_j = \sum_{j=0}^{m} n_j U_j^2, \tag{2.9}$$

$$S'_m = \min_{\pi \in S_n} \sum_{i=1}^{n} iQ_{\pi(i)} = \sum_{j=0}^{m} \sum_{i=n-N_{m-j-1}}^{n-N_{m-j}+1} iU_{m-j} = \sum_{j=0}^{m} n_j U_j (n - U_j + 1), \tag{2.10}$$

$$S'_0 = \frac{S'_M + S'_m}{2} = \frac{n+1}{2} \sum_{j=0}^{m} n_j U_j = \frac{n+1}{2} N'. \tag{2.11}$$

Pro dokázání rovnosti $Gini' = \frac{S'-S'_0}{S'_M-S'_0}$ bychom postupovali zcela analogicky. Nejdříve algebraickými úpravami dokázali vztah $S'_M - S'_m = (1-2K')N'n$ a poté indukcí podle počtu všech prohození rovnost $S'_M - S' = \sum_{j=1}^{m} \sum_{i=0}^{j-1} P_{i,j}(U_j - U_i)$. Poté již snadno dostaneme:

$$\begin{aligned} Gini' = 1 - \frac{2AUC'}{1-2K'} = \frac{S' - S'_0}{S'_M - S'_0} = 1 - 2\frac{S'_M - S'}{S'_M - S'_m} \\ = \frac{\sum_{i=1}^{n} R_i Q_i - \frac{n(n+1)^2}{2}}{S'_M - S'_0}. \end{aligned} \tag{2.12}$$

Zavedení Giniho koeficientu přes pořadí veličin y_i umožňuje nalézt funkční závislost mezi $Gini'$ a r_S. Pro vyjádření r_S spočteme $\overline{Q} = \frac{1}{n}\sum_{i=0}^{m} n_i U_i = \frac{N'}{n} = \frac{n+1}{2}$ a proto $n\overline{Q}^2 = S'_0$. Dále $\sum_{i=1}^{n} Q_i^2 = \sum_{i=0}^{m} n_i U_i^2 = S'_M$ což umožňuje Spearmanův korelační koeficient vyjádřit jako:

$$r_S = \frac{\sum_{i=1}^{n} R_i Q_i - n\overline{RQ}}{\sqrt{\left(\sum_{i=1}^{n} R_i^2 - n\overline{R}^2\right)\left(\sum_{i=1}^{n} Q_i^2 - n\overline{Q}^2\right)}} = \frac{\sum_{i=1}^{n} R_i Q_i - \frac{n(n+1)^2}{2}}{\sqrt{\left(\sum_{i=1}^{n} R_i^2 - n\overline{R}^2\right)(S'_M - S'_0)}}.$$

Tedy dostáváme:

$$r_S = \sqrt{\frac{S'_M - S'_0}{\sum_{i=1}^{n} R_i^2 - n\overline{R}^2}} Gini' = \sqrt{\frac{12(S'_M - S'_0)}{n(n+1)(n-1)}} Gini', \tag{2.13}$$

kde druhá rovnost platí pouze, pokud jsou všechna skóre navzájem různá.

Další vlastností $Gini'$ (resp. $Gini$) je jeho invariantnost na lineární transformace pořadí Q_i (hodnoty y_i). Ta lze ukázat nejlépe ze vzorce (2.12). Pokud zavedeme

$\tilde{Q}_i = aQ_i + b$, pak $\tilde{S} = \sum_{i=1}^{n} R_i\tilde{Q}_i = b\sum_{i=1}^{n} R_iQ_i + aS$. Stejná transformace bude i u ostatních veličin, z čehož vyjde:

$$\frac{\tilde{S} - \tilde{S}_0}{\tilde{S_M} - \tilde{S}_0} = \frac{b\sum_{i=1}^{n} R_iQ_i + aS - b\sum_{i=1}^{n} R_iQ_i - aS_M}{aS_M + b\sum_{i=1}^{n} R_iQ_i - aS_0 - b\sum_{i=1}^{n} R_iQ_i} = \frac{S - S_0}{S_M - S_m}.$$

Díky této invariantnosti budou $Gini$ a $Gini'$ shodné v případě, že všechny kategorie budou stejně zastoupené. Pak totiž $Q_i = in_0 + \frac{n_0+1}{2}$ a stačí zvolit transformaci $\tilde{Q}_i = (Q_i - \frac{n_0+1}{2})\frac{1}{n_0} = y_i$.

2.1.3 Gini pomocí C-statistiky

Jelikož pro binární vysvětlovanou proměnnou Giniho koeficient úzce souvisí s C-statistikou, rozšíříme její definici i pro ordinální vysvětlovanou proměnnou. C-statistiku (C_m) definujeme jako pravděpodobnost, že skóre náhodně vybraného prvku s hodnotou i bude menší než skóre náhodně vybraného prvku s hodnotou j, kde $i < j$. V případě shody skóre započítáme jednu polovinu. Zapsáno vzorcem

$$\begin{aligned}
C_m &= \mathsf{P}\left(s_k < s_l | y_k < y_l\right) + \frac{1}{2}\mathsf{P}\left(s_k = s_l | y_k < y_l\right) \\
&= \frac{\sum_{j=1}^{m}\sum_{i=0}^{j-1} \mathsf{P}\left(s_k < s_l | y_k = i, y_l = j\right)\mathsf{P}\left(y_k = i, y_l = j\right)}{\mathsf{P}\left(y_k < y_l\right)} \\
&+ \frac{1}{2}\frac{\sum_{j=1}^{m}\sum_{i=0}^{j-1} \mathsf{P}\left(s_k = s_l y_k = i, y_l = j\right)\mathsf{P}\left(y_k = i, y_l = j\right)}{\mathsf{P}\left(y_k < y_l\right)} \\
&= \frac{\sum_{j=1}^{m}\sum_{i=0}^{j-1} \left(\mathsf{P}\left(s_k < s_l | y_k = i, y_l = j\right) + \frac{1}{2}\mathsf{P}\left(s_k = s_l | y_k = i, y_l = j\right)\right) n_i n_j}{\sum_{j=1}^{m}\sum_{i=0}^{j-1} n_i n_j} \\
&= \frac{\sum_{j=1}^{m}\sum_{i=0}^{j-1} C_{i,j} n_i n_j}{\sum_{j=1}^{m}\sum_{i=0}^{j-1} n_i n_j}
\end{aligned}$$

Kde $C_{i,j}$ jsme označili C-statistiku mezi prvky s hodnotami i a j. Fakt, že jak $Gini$ ($Gini'$), tak C_m se dají vyjádřit pomocí dílčích charakteristik, by mohl vést k přesvědčení o vzájemném pevném vztahu. Ten ovšem existovat nemůže, což ilustrujeme na následujícím příkladě:

Nechť máme dva modely, kde setřízené hodnoty dle skóre jsou $3, 0, 1, 2$ pro první model a $0, 3, 2, 1$ pro druhý model. Spočteme $Gini_1 = -\frac{1}{5}$, $Gini'_1 = -\frac{1}{5}$, $C_1 = \frac{1}{2}$ a $Gini_2 = \frac{1}{5}$, $Gini'_2 = \frac{1}{5}$, $C_2 = \frac{1}{2}$ z čehož plyne, že $Gini$ ani $Gini'$ nemůže být funkcí C-statistiky.

Obdobně po setřízení hodnot dle skóre jiných modelů dostaneme $0, 1, 1, 2, 2, 0$ a $1, 1, 0, 2, 0, 2$ a charakteristiky diverzifikace vyjdou $Gini_1 = \frac{1}{4}$, $Gini'_1 = \frac{1}{4}$, $C_1 = \frac{7}{12}$ a $Gini_2 = \frac{1}{4}$, $Gini'_2 = \frac{1}{4}$, $C_2 = \frac{2}{3}$. A proto C-statistika nemůže být funkcí $Gini$ ani $Gini'$.Tato skutečnost je důsledkem toho, že C-statistika (oproti $Gini$ a $Gini'$) hodnotí pouze, zda jsou odhadnuté prvky setřízené dle skóre správně či špatně, nezávisle na rozdílu jejich hodnot.

Z toho vidíme, že odlišný Giniho koeficient lze definovat na základě C-statistiky následovně:

$$Gini_C = 2C - 1 = \frac{\sum_{j=1}^{m}\sum_{i=0}^{j-1}(2C_{i,j}-1)n_i n_j}{\sum_{j=1}^{m}\sum_{i=0}^{j-1} n_i n_j} = \frac{\sum_{j=1}^{m}\sum_{i=0}^{j-1} Gini_{i,j} n_i n_j}{\sum_{j=1}^{m}\sum_{i=0}^{j-1} n_i n_j}.$$

Pro binární proměnnou je i tato definice shodná s původní definicí Giniho koeficientu, jak je odvozeno v (1.17).

2.2 Gini a koeficient determinace pro nominální vysvětlovanou proměnnou

V této části zadefinujeme koeficient determinace R^2 a Giniho koeficient pro nominální vysvětlovanou proměnnou Y. Stejně jako výše budeme předpokládat kódování jednotlivých kategorií čísly $\{0, 1, ..., m\}$. Nechť máme nezávislá pozorování y_i, $i = 1, ..., n$ a každá kategorie je v našem výběru zahrnuta alespoň jednou. Výstupem matematického modelu nebude jedno číslo, ale vektor odhadnutých pravděpodobností, že Y patří do dané kategorie. Tento vektor označme $\hat{Y}$. Tedy pro regresory odpovídající našim pozorováním dostaneme $\hat{y}_i = (\hat{p_i}^1, ..., \hat{p_i}^m)$, kde čísla $\hat{p_i}^j$, $j = 1, ..., m$ značí odhadnutou pravděpodobnost, že i-tý prvek patří do j-té skupiny.

Opět označme počet prvků v j-té kategorii, číslem $n_j, j = 0, ..., m$. Giniho koeficient definujeme pouze na základě spočtených $\hat{y}_i$ a jejích složek. Nechť prvek $y_i = j, j \in \{0, ..., m\}$, pak j-tou složku odhadu $\hat{y}_i$ porovnáme s j-tou složkou všech ostatních odhadů pozorování, které nepatří do kategorie j. Pokud bude $\hat{p_i}^j > \hat{p_k}^j$, $k \neq j$ započítáme $\frac{1}{n-n_{y_i}}$, v případě opačné nerovnosti daný výraz odečteme. Při rovnosti $\hat{p_i}^j = \hat{p_k}^j$, $k \neq j$ nepřičteme ani neodečteme nic. Nyní již k samotné definici:

$$Gini_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1, y_i \neq y_j}^{n} \frac{\mathrm{sgn}(\hat{p_i}^{y_i} - \hat{p_j}^{y_i})}{n - n_{y_i}}$$

Takto definovaný $Gini_n$ nabývá hodnot v intervalu $[-1, 1]$. Záporné hodnoty znamenají převrácený model v tom smyslu, že čím menší je $\hat{p_i}^j$ tím větší je pravděpodobnost, že $y_i = j$. V případě ideálního modelu, kde $\hat{p_i}^{y_i} = 1$, $i = 1, ..., n$, bude $Gini = 1$. Nyní se zaměřme na případ, kdy $m = 1$. Označme $s_i = \hat{p_i}^1$, pak podle definice bude:

$$\begin{aligned} Gini_n &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j, y_i=0, y_j=1}\frac{\mathrm{sgn}(s_j - s_i)}{n_1} + \sum_{j, y_i=1, y_j=0}\frac{\mathrm{sgn}(s_i - s_j)}{n_0}\right) \\ &= \frac{1}{n}\sum_{i, y_i=0}\sum_{j, y_j=1}\frac{n_0 + n_1}{n_1 n_0}\mathrm{sgn}(s_j - s_i) = \sum_{i, y_i=0}\sum_{j, y_j=1}\frac{\mathrm{sgn}(s_j - s_i)}{n_0 n_1}. \end{aligned}$$

Z čehož podle (1.17) bude pro $m = 1$ $Gini$ pro nominální a ordinální vysvětlovanou proměnnou stejný, pokud jako skóre modelu zvolíme hodnoty $\hat{p_i}^1$. Pro $m > 1$

již shodnost definic nemá smysl diskutovat, neboť není jasné, jak zvolit skóre na základě odhadu $\hat{y}_i = (\hat{p}_i^1, ..., \hat{p}_i^m)$.

Pro ordinální vysvětlovanou proměnnou jsme odvodili vztah pro výpočet Giniho koeficientu pomocí dílčích Giniho koeficientů mezi všemi dvojicemi hodnot $(k, l), k < l, k, l \in \{1, ..., m\}$. Tento přístup pro nominální vysvětlovanou proměnnou ztrácí smysl, neboť nemusí platit ekvivalence $\text{sgn}(\hat{p}_i^{y_i} - \hat{p}_j^{y_i}) = 0 \Leftrightarrow \text{sgn}(\hat{p}_j^{y_j} - \hat{p}_i^{y_j}) = 1$. Proto uvážíme pouze $m + 1$ Giniho koeficientů ($Gini_k$) a to mezi veličinami $y_i^k = I_{[y_i=k]}$ a $s_i^k = \hat{p}_i^k$ s četnostmi jevů $n_1^k = \sum_{i=1}^n y_i^k = n_k$ a $n_0^k = n - n_k$. Jedná se o binární proměnnou y_i^k, která se rovná jedné v případě, že její hodnota padne do k-té kategorie a skóre rovné odhadu této pravděpodobnosti. Proto bude:

$$Gini_k = \sum_{i,y_i^k=0} \sum_{j,y_j^k=1} \frac{\text{sgn}(s_j^k - s_i^k)}{n_0^k n_1^k} = \sum_{i,y_i \neq k} \sum_{j,y_j=k} \frac{\text{sgn}(\hat{p}_j^k - \hat{p}_i^k)}{(n - n_k)n_k} \tag{2.14}$$

A celkový $Gini_n$ vyjádříme pomocí dílčích $Gini_k$ následovně:

$$\begin{aligned} Gini_n &= \frac{1}{n} \sum_{k=0}^m \sum_{i,y_i=k} \sum_{j,y_j \neq k} \frac{\text{sgn}(\hat{p}_i^k - \hat{p}_j^k)}{n - n_k} = \frac{1}{n} \sum_{k=0}^m n_k \sum_{i,y_i \neq k} \sum_{j,y_j=k} \frac{\text{sgn}(\hat{p}_j^k - \hat{p}_i^k)}{(n - n_k)n_k} \\ &= \frac{1}{n} \sum_{k=0}^m n_k Gini_k. \end{aligned} \tag{2.15}$$

Jedná se tedy o jakýsi vážený průměr jednotlivých $Gini_k$, kde váhy jsou četnosti jevů $y_i = k$.

Pro ordinální vysvětlovanou proměnnou je s $Gini_n$ úzce spojena C-statistika, tu lze velmi obdobně definovat i pro nominální vysvětlovanou proměnnou. A to jako pravděpodobnost, že k-tá složka odhadu náhodně vybraného prvku s hodnotou různou od k bude menší než k-tá složka odhadu náhodně vybraného prvku s hodnotou k. V případě shody započítáme $\frac{1}{2}$. Symbolicky zapsáno:

$$\begin{aligned} C &= \mathsf{P}\left(\hat{p}_i{}^{y_j} < \hat{p}_j{}^{y_j} | y_i \neq y_j\right) + \frac{1}{2}\mathsf{P}\left(\hat{p}_i{}^{y_j} = \hat{p}_j{}^{y_j} | y_i \neq y_j\right) \\ &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1, y_i \neq y_j}^n \frac{1 + \text{sgn}(\hat{p}_i{}^{y_i} - \hat{p}_j{}^{y_i})}{2(n - n_{y_i})} \end{aligned}$$

Z tohoto vzorce vidíme, že platí vztah $Gini_n = 2C - 1$.

Při snaze definovat koeficient determinace pro nominální vysvětlovanou proměnnou se inspirujeme v logistické regresi, kde jsme uvedli některé možné definice. McFaddenův a Cox and Snellův koeficient determinace se dají zavést zcela analogicky. Naproti tomu Efronův koeficient je třeba rozumně modifikovat, neboť nyní nemáme odhad Y dán jedním číslem, nýbrž máme odhady jednotlivých pravděpodobností. Proto RSS bude součet rozdílů čtverců složek odhadu od indikátoru jevu, že Y patří do dané kategorie. Tedy

$$RSS = \sum_{i=1}^n \sum_{k=0}^m (\hat{p}_i{}^k - I_{[y_i=k]})^2 = \sum_{i=1}^n \sum_{k=0}^m (\hat{p}_i{}^k)^2 - 2 \sum_{i=1}^n \sum_{k=0}^m I_{[y_i=k]} \hat{p}_i{}^k + n.$$

Při použití nominálního modelu bez vysvětlujících proměnných budou jednotlivé odhady pravděpodobností kategorie Y_i dány následovně: $\hat{p_i}^k = \frac{n_k}{n}$, $k = 1, ..., m$. A proto úplný a vysvětlený součet čtverců jsou

$$TSS = \sum_{i=1}^{n}\sum_{k=0}^{m}(I_{[y_i=k]} - \frac{n_k}{n})^2 = n - 2\sum_{i=1}^{n}\sum_{k=0}^{m}\frac{n_k}{n}I_{[y_i=k]} + \sum_{k=0}^{m}\frac{n_k^2}{n},$$
$$ESS = \sum_{i=1}^{n}\sum_{k=0}^{m}(\hat{p_i}^k - \frac{n_k}{n})^2 = \sum_{i=1}^{n}\sum_{k=0}^{m}(\hat{p_i}^k)^2 - 2\sum_{i=1}^{n}\sum_{k=0}^{m}\frac{n_k}{n}\hat{p_i}^k + \sum_{k=0}^{m}\frac{n_k^2}{n}.$$

Rozšířený Efronův koeficient determinace definujeme pomocí zavedených veličin jako

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}\sum_{k=0}^{m}(\hat{p_i}^k - I_{[y_i=k]})^2}{\sum_{i=1}^{n}\sum_{k=0}^{m}(I_{[y_i=k]} - \frac{n_k}{n})^2}.$$

S použitím rovností výše dostáváme

$$ESS + RSS - TSS = 2\left(\sum_{i=1}^{n}\sum_{k=0}^{m}\hat{p_i}^k(\hat{p_i}^k - I_{[y_i=k]}) + \sum_{k=0}^{m}\frac{n_k}{n}\sum_{i=1}^{n}(I_{[y_i=k]} - \hat{p_i}^k)\right).$$

Přičemž při použití logitového modelu a odhadu koeficientů pomocí metody maximální věrohodnosti platí: $\sum_{i=1}^{n}(I_{[y_i=k]} - \hat{p_i}^k) = 0$. Dále máme $\mathsf{E}\, I_{[Y=k]} = \mathsf{P}(Y = k)$ a tedy první sčítanec je ve střední hodnotě nulový. Z toho plyne, že při velkém počtu pozorování bude blízký nule. Celkově dostáváme přibližnou rovnost $TSS \approx RSS + ESS$.

Pro $m = 1$ bude definice shodná s definicí Efronova koeficientu determinace, neboť $(\hat{p_i}^0 - I_{[y_i=0]})^2 = (\hat{p_i}^1 - I_{[y_i=1]})^2$ a $(\frac{n_0}{n} - I_{[y_i=0]})^2 = (\frac{n_1}{n} - I_{[y_i=1]})^2$ a tedy

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{p_i}^1 - I_{[y_i=1]})^2}{\sum_{i=1}^{n}(\frac{n_1}{n} - I_{[y_i=1]})^2} = 1 - \frac{\sum_{i=1}^{n}(\hat{p_i}^1 - y_i)^2}{\sum_{i=1}^{n}(y_i - \frac{n_1}{n})^2} = R_e^2.$$

Stejně jako pro Giniho koeficient vyjádříme koeficient determinace pomocí $m + 1$ dílčích R_k^2 opět mezi binární proměnnou $I_{[y_i=k]}$ a $\hat{p}_i^k = \mathsf{P}(Y_i = k)$. Tedy $RSS_k = \sum_{i=1}^{n}(\hat{p_i}^k - I_{[y_i=k]})^2$, $TSS_k = \sum_{i=1}^{n}(\hat{p_i}^k - \frac{n_k}{n})^2$ a $R_k^2 = 1 - \frac{RSS_k}{TSS_k}$. Celkový součet čtverců zjednodušíme a vyjádříme reziudální součet čtverců v závislosti na R_k^2 a TSS_k. Ještě označme $p_k = \frac{n_k}{n}$:

$$\begin{aligned} TSS_k &= n_k(1 - \frac{n_k}{n})^2 + (n - n_k)(0 - \frac{n_k}{n})^2 = n_k - 2\frac{n_k^2}{n} + \frac{n_k^3}{n^2} + \frac{n_k^2}{n} - \frac{n_k^3}{n^2} \\ &= n_k - \frac{n_k^2}{n} = np_k(1 - p_k) \\ RSS_k &= (1 - R_k^2)TSS_k = np_k(1 - p_k)(1 - R_k^2) \end{aligned}$$

Po dosazení těchto vztahů do definice R^2 a snadných algebraických úpravách dospějeme k vyjádření:

$$R^2 = 1 - \frac{\sum_{k=0}^{m}RSS_k}{\sum_{k=0}^{m}TSS_k} = 1 - \frac{\sum_{k=0}^{m}np_k(1 - p_k)(1 - R_k^2)}{\sum_{k=0}^{m}np_k(1 - p_k)} = \frac{\sum_{k=0}^{m}p_k(1 - p_k)R_k^2}{\sum_{k=0}^{m}p_k(1 - p_k)}. \tag{2.16}$$

Zrekapitulujme si výsledky, ke kterým jsme dospěli. Podařilo se nám vyjádřit $Gini_n$ i R^2 jako vážené lineární kombinace dílčích koeficientů pro binární proměnné. Vztah (1.22), který v našem značení má tvar $R_k^2 \approx 3p_k(1-p_k)Gini_k^2$, nám tyto dílčí koeficienty dává do souvislosti. Použijeme-li tuto aproximaci $m+1$ krát, dostaneme:

$$R^2 \approx 3\frac{\sum_{k=0}^{m} p_k^2(1-p_k)^2 Gini_k^2}{\sum_{k=0}^{m} p_k(1-p_k)}. \tag{2.17}$$

Bez dodatečných předpokladů další úpravy neprovedeme, podívejme se ale na případ, kdy budou všechny kategorie stejně zastoupené, tj. $n_0 = n_k$, $k = 1, ..., m$. Pak $n = n_0(m+1)$ a $p_0 = \frac{1}{m+1}$ a vyjádření z (2.17) upravíme do tvaru:

$$R^2 \approx 3\frac{m}{(m+1)^3}\sum_{k=0}^{m} Gini_k^2. \tag{2.18}$$

Pokud budou hodnoty všech dílčích Giniho koeficientů přibližně stejné, lze výraz z (2.18) upravit na tvar:

$$R^2 \approx 3\frac{m}{(m+1)^2}Gini.$$

Z kterého vidíme, že při rostoucím počtu kategorií R^2 klesá, pokud dílčí Giniho koeficienty jsou přibližně stejné.

Pokusme se ještě určit případy, kdy bude R^2 maximální při pevných $Gini_k$. Nejprve se, ale budeme zabývat úlohou ve zjednodušeném tvaru, tj. budeme řešit:

$$\begin{gathered}\text{maximalizovat } \frac{\sum_{k=0}^{m} p_k^2(1-p_k)^2}{\sum_{k=0}^{m} p_k(1-p_k)} = f(p_0, ..., p_m) \\ \text{za podmínek } 0 \leq p_i < 1 \ \forall i = 0, 1, ..., m, \\ \sum_{i=0}^{m} p_i = 1\end{gathered}$$

Spočteme parciální derivaci funkce f podle i-té proměnné:

$$\partial_{p_i} f(p_0, ..., p_m) = \frac{(1-2p_i)\left(2p_i(1-p_i)\sum_{k=0}^{m} p_k(1-p_k) - \sum_{k=0}^{m} p_k^2(1-p_k)^2\right)}{\left(\sum_{k=0}^{m} p_k(1-p_k)\right)^2}.$$

Ta se rovná nule, buď když $p_i = \frac{1}{2}$ nebo $2p_i(1-p_i) = f(p_0, ..., p_m)$. Z čehož dostáváme, že $2p_i(1-p_i) = 2p_j(1-p_j)$. Jelikož součet všech p_i se rovná jedné, platí buď, že:

1. $m = 1$ a $p_0 = p_1 = \frac{1}{2}$
2. jedno $p_i = \frac{1}{2}$ a pro ostatní platí $p_i = p_j$ nebo $p_i = 1 - p_j$. Z podmínek omezenosti však druhá rovnost nemá smysl, a proto $p_k = p_j$ a jejich součet se rovná $\frac{1}{2}$. Dosazením tohoto bodu do rovnosti $2p_j(1-p_j) = f(p_0, ..., p_m)$ zjistíme, že rovnost není splněna pro žádné m přirozené.

Zkoumejme nyní body na hranici množiny M vyhraničené našemi podmínkami. Pro $p_i = 1$ platí, že ostatní $p_j = 0$, a tedy funkce f není definována, ale pro $p_i = 0$ se nám funkce f zjednodušší a bude stejného tvaru jen v dimenzi o jedna menší. Při opětovném výpočtu derivace dostaneme stejnou úlohu, o které již víme, že má jen jediné možné řešení.

Jelikož množina M není kompakt musíme ještě určit limity v bodech nespojitosti. Ty ovšem vyjdou nulové, a proto se maxima nabývá v bodech, kdy $p_i = p_j = \frac{1}{2}$ a funkční hodnota je $f(0, ..., \frac{1}{2}, ..., \frac{1}{2}, ..., 0) = \frac{1}{4}$. S využitím tohoto výsledku si již snadno rozmyslíme, že výraz (2.17) bude svého maxima nabývat při $p_i = p_j = \frac{1}{2}$ a $p_k = 0$, kde i, j jsou takové indexy, pro které platí $Gini_i^2 \geq Gini_j^2 \geq Gini_k^2$, $k \in \{0, ..., m\}, k \neq j, k \neq i$. V našem případě ovšem rovnost $p_k = 0$ nemá smysl, neboť předpokládáme $n_k > 0$. Ovšem ze spojitosti plyne, že maximální hodnoty se bude nabývat pro případy blížící se uvedenému limitnímu. Proto R^2 při pevném počtu kategorií vysvětlované proměnné bude maximální, když dvě marginální četnosti s nejvyšším $Gini_k$ budou zhruba stejné a ostatní budou zastoupeny jen jedním jevem.

2.3 Vztah R^2 a Gini pro spojitou vysvětlovanou proměnnou

V této části zadefinujeme Giniho koeficient pro spojitou vysvětlovanou proměnnou. Nechť Y je spojitá náhodná veličina a X je náhodná veličina mající roli skóre. V praxi se může spočítat například za použití odhadnutých parametrů $\boldsymbol{\beta}$ jako $x = \boldsymbol{x}'\boldsymbol{\beta}$. Nechť máme nezávislá pozorování $(x_1, y_1), ..., (x_n, y_n)$. Pro zavedení L-křivky se můžeme inspirovat u ordinální vysvětlované proměnné. Jelikož budeme chtít využít informace o jejích hodnotách, nabízí se nám dvě možné definice. Pokud bychom chtěli použít přímo hodnoty y_i není jasné, jak pracovat se zápornými hodnotami. Proto nejpřirozenější přístup bude pomocí pořadí Q_i pozorování y_i. Označme $N = \sum_{i=1}^{n} Q_i = \frac{n(n+1)}{2}$ a zcela analogicky jako pro ordinální vysvětlovanou proměnnou definujeme empirickou distribuční funkci F a pomocnou funkci G předpisem:

$$F(x) = \frac{1}{N} \sum_{i=1}^{n} I_{(-\infty,x]}(x_i) Q_i, \ x \in \mathbb{R}$$

$$G(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(x_i).$$

Analogii L-křivky dostaneme spojením bodů $[0, 0]$ a $[G(x_i), F(x_i)]$, kde $i = 1, ..., n$. Opět uvážíme křivky L_{id} (resp. L_{-id}) jako L-křivku pro data setřízená modelem od nejmenší po největší (resp. od největší po nejmenší). A použijeme stejné značení tj. K bude plocha mezi L_{id} a hranami jednotkového čtverce a AUC plocha mezi L-křivkou a L_{id}. Oproti ordinální proměnné můžeme předpokládat, že všechny hodnoty y_i, $i = 1, ..., n$ jsou navzájem různé, protože se jedná o výběr ze spojitého rozdělení. Velikost plochy K můžeme obecně vyjádřit:

$$K = \frac{1}{Nn} \left(\sum_{k=1}^{n} \left(\frac{k}{2} + k(n-k) \right) \right) = \frac{2n+1}{6n}.$$

Samotná definice Gini je:

$$Gini = 1 - \frac{2AUC}{1-2K}.$$

Vidíme, že $Gini \in [-1, 1]$. K tomu, abychom mohli vyjádřit Gini za pomoci pořadí R_i prvku y_i podle velikosti skóre x_i použijeme opět veličiny $S = \sum_{i=1}^{n} R_i Q_i$ a její maximální (resp. minimální) možné hodnoty S_M (resp. S_m) a průměr jejích hodnot S_0, které se dají upravit následovně:

$$\begin{aligned} S_M &= \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6} \\ S_m &= \sum_{i=1}^{n} i(n+1-i) = \frac{n(n+1)(n+2)}{6} \\ S_0 &= \frac{S_M + S_m}{2} = n\left(\frac{n+1}{2}\right)^2. \end{aligned}$$

Výpočtem zjistíme, že $S_M - S_m = \frac{n(n^2-1)}{6} = (1-2K)Nn$. Stačí proto ověřit rovnost $AUC \cdot Nn = S_M - S$. Výpočet plochy pod L-křivkou lze spočítat jako součet pravoúhlých lichoběžníků s délkami rovnoběžných hran $1 - \frac{i-1}{n}$ a $1 - \frac{i}{n}$ a výškou $\frac{Q_j}{Nn}$, kde $j = R_i$ a $i = 1, ..., n$. Tedy

$$\begin{aligned} AUC + K &= \sum_{i=1}^{n} \left(\frac{Q_i}{2Nn} + \frac{\sum_{j=1}^{n} I_{[R_j=i]} Q_j}{nN}(n-i) \right) \\ &= \frac{1}{2n} + 1 - \sum_{i=1}^{n} \frac{i \sum_{j=1}^{n} I_{[R_j=i]} Q_j}{Nn} = 1 + \frac{1}{2n} - \frac{1}{Nn} \sum_{j=1}^{n} R_j Q_j \\ AUC &= \frac{1}{Nn} \left(\frac{2Nn+N}{3} - \sum_{i=1}^{n} R_i Q_i \right) = \frac{1}{Nn} \left(S_M - S \right). \end{aligned}$$

Z tohoto dostáváme $Gini = 1 - \frac{2AUC}{1-2K} = 1 - 2\frac{S_M - S}{S_M - S_m} = \frac{S - S_0}{S_M - S_0}$. A jelikož $S_M = \sum_{i=1}^{n} R_i^2 = \sum_{i=1}^{n} Q_i^2$ a $S_0 = n\overline{RQ}$ vidíme, že

$$Gini = \frac{S - S_0}{S_M - S_0} = \frac{\sum_{i=1}^{n} R_i Q_i - n\overline{RQ}}{\sum_{i=1}^{n} Q_i^2 - n\overline{Q}^2} = r_S.$$

Při zavedení Giniho koeficientu pomocí pořadí hodnot y_i se definice shoduje se Spearmanovým korelačním koeficientem.

Jak již bylo zmíněno, pro spojitou vysvětlovannou proměnnou se nejčastěji používá lineární regrese a jako ukazatel kvality modelu koeficient determinace R^2. V případě, že jeho hodnota je dosti nízká (kolem 10%) považuje se model za nedostatečný. Ovšem pokud bychom jako hlavní diverzifikační koeficient vzali $Gini$, vzniká otázka, jaké hodnoty koeficientu determinace R^2 mu budou odpovídat v závislosti na rozdělení vysvětlované proměnné. Asi lze obecně očekávat, že po zlepšení modelu a zvýšení hodnoty jednoho z koeficientů, vzrostou i hodnoty druhého. Otázkou je o kolik a na čem nárůst závisí.

K tomu, abychom tento vztah mohli popsat učiníme některé předpoklady. O rozdělení veličiny X budeme předpokládat, že má normální rozdělení s nulovou střední hodnotou a jednotkovým rozptylem, tj. $X \sim \mathsf{N}(0,1)$. Zdůvodnění realističnosti těchto předpokladů lze nálezt v [4]. Zmiňme jen, že je podstatné, aby v modelu měly všechny regresory podobnou váhu a byly na sobě nezávislé.

Další předpoklad učiníme o veličině charakterizující nepřesnost naší predikce. Označme ji ε a opět předpokládáme, že $\varepsilon \sim \mathsf{N}(0,1)$. Posledním z předpokladů je nezávislost ε a X.

Zvolme si spojitou distribuční funkci F, která bude distribuční funkcí vysvětlované proměnné. Dále parametr $r \in [0,1]$, který bude vyjadřovat jak dobrou predikcí je X. Zadefinujeme pomocnou náhodnou veličinu $Z = rX + \sqrt{1-r^2}\varepsilon$. Z věty o konvoluci plyne, že Z bude mít normální rozdělení a snadným výpočtem zjistíme, že $\mathsf{E}\, Z = 0$, $\mathsf{var}\, Z = 1$, tedy $Z \sim N(0,1)$.

Nyní položme $Y = F^{-1}(\Phi(Z))$ a pro distribuční funkci Y, platí:

$$F_Y(y) = \mathsf{P}\left(F^{-1}(\Phi(Z)) < y\right) = \mathsf{P}\left(\Phi(Z) < F(y)\right) = \mathsf{P}\left(Z < \Phi^{-1}(y)\right) = F(y).$$

Z odvozeného vztahu dostáváme shodnost distribučních funkcí $F(x) = F_Y(x)$. Odhad náhodné veličiny Y pomocí X určíme jako podmíněnou střední hodnotu:

$$\hat{Y} = \mathsf{E}\left[Y|X\right] = \mathsf{E}\left[F_Y^{-1}(\Phi(Z))|X\right]. \tag{2.19}$$

Pro zjednodušení této podmíněné střední hodnoty využijeme nezávislosti X a ε a dostaneme integrál:

$$\begin{aligned}\mathsf{E}\left[F_Y^{-1}(\Phi(rX + \sqrt{1-r^2}\varepsilon))|X = x\right] &= \mathsf{E}_{\,\varepsilon}(F_Y^{-1}(\Phi(rx + \sqrt{1-r^2}\varepsilon)) \\ &= \int_{-\infty}^{\infty} F_Y^{-1}(\Phi(rx + \sqrt{1-r^2}e))\varphi(e)\,\mathrm{d}e.\end{aligned}$$

Pro výpočet tohoto integrálu je potřeba znalosti distribuční funkce F_Y. Jelikož integrand obsahuje distribuční funkci normálního rozdělení, pro většinu typů rozdělení veličiny Y bude třeba tento integrál počítat numericky.

Než se pustíme do výpočtů odhadů veličiny Y na základě znalosti její distribuční funkce, podívejme se na vztah *Gini* mezi veličinami X a Z a mezi X a Y. Veličinu Y jsme vyjádřili pomocí Z jako $Y = F^{-1}(\Phi(Z)) = f(Z)$. Distribuční funkce jsou rostoucí a tudíž i F^{-1} je rostoucí, složení dvou rostoucích funkcí bude opět funkce rostoucí. Když budeme mít náhodný výběr z rozdělení Z a vypočteme odpovídající veličiny Y, bude uspořádání prvků podle velikosti stejné. Důsledkem toho bude Spearmanův korelační koeficient (resp. *Gini*) stejný mezi X a Z a mezi X a Y.

Veličinu Z jsme definovali specifickým způsobem nejen kvůli tvaru jejího rozdělení, ale také abychom znali její korelaci s veličinou X.

$$\mathsf{corr}\,(X,Z) = \mathsf{corr}\left(X, rX + \sqrt{1-r^2}\epsilon\right) = r\mathsf{corr}\,(X,X) + \sqrt{1-r^2}\mathsf{corr}\,(X,\epsilon) = r,$$

V poslední rovnosti jsme využili nezávislost veličin X a ϵ. S využití vztahu (1.21) dostaneme přibližnou rovnost mezi korelačním koeficientem r a r_S. Celkově tedy

$$\mathsf{corr}\,(X,Z) = r \approx r_S(X,Z) = r_S(X,Y) = Gini.$$

Nyní budeme chtít dostat vztah mezi $Gini$ a R^2. K tomu musíme určit odhad veličiny Y. Ten bude závislý na rozdělení Y a její distribuční funkci F_Y. Zvlášť si rozebereme případ, kdy $Y \sim \mathsf{N}(0,1)$, neboť ten je specifický. Tedy $Y = \Phi^{-1}(\Phi(Z)) = Z$ a odhad z (2.19) bude:

$$\hat{Y} = \mathsf{E}\,[Z|X] = rX.$$

Předpokládejme, že máme náhodný výběr $(X_1, Y_1), ..., (X_n, Y_n)$. Víme, že $Y_i = rX_i + \sqrt{1-r^2}\epsilon_i$ a odhad $\hat{Y}_i = rX_i$. Spočteme reziduální a celkový součet čtverců.

$$\begin{aligned} RSS &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (\sqrt{1-r^2}\epsilon_i)^2 = (1-r^2)\sum_{i=1}^{n} \epsilon_i^2 \\ TSS &= \sum_{i=1}^{n} (Y_i - \overline{Y}_i)^2 = \sum_{i=1}^{n} (rX_i - \overline{X}) + \sqrt{1-r^2}(\epsilon_i - \overline{\epsilon}))^2 \\ &= r^2 \sum_{i=1}^{n} (X_i - \overline{X})^2 + 2r\sqrt{1-r^2}\sum_{i=1}^{n} (X_i - \overline{X})(\epsilon_i - \overline{\epsilon}) + (1-r^2)\sum_{i=1}^{n} (\epsilon_i - \overline{\epsilon})^2 \end{aligned} \tag{2.20}$$

Z vlastností normálního rozdělení a konzistence odhadů $\overline{X}$ a $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ plyne, že $\frac{1}{n}RSS \longrightarrow (1-r^2), s.j.$ Dále se podívejme na součet z prostředního členu výrazu (2.20)

$$\sum_{i=1}^{n} (X_i - \overline{X})(\epsilon_i - \overline{\epsilon}) = \sum_{i=1}^{n} X_i\epsilon_i - n\overline{X}\overline{\epsilon}.$$

Využitím nezávislosti X_i a ϵ_i dostaneme $\mathsf{E}\,X_i\epsilon_i = \mathsf{E}\,X_i\mathsf{E}\,\epsilon_i = 0$. A z Kolmogorovova silného zákonu velkých čísel, viz [1], platí, že $\frac{1}{n}\sum_{i=1}^{n} X_i\epsilon_i \longrightarrow 0, s.j.$ Celkově jsme odvodili, že

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\frac{1}{n}RSS}{\frac{1}{n}TSS} \longrightarrow r^2, s.j.$$

Pro dostatečně velká n tedy platí aproximativní vztah $R^2 \approx r^2 \approx Gini^2$.

Pro jiné než normální rozdělení tento způsob odvození nebude možný, neboť funkce f nebude identita. Pokusíme se ovšem numericky vyčíslit jednotlivé odhady a koeficient determinace R^2, který následně porovnáme s r^2. K tomu zvolíme následující postup: Nagenerujeme náhodné hodnoty $\epsilon_1, ..., \epsilon_m$ a $x_1, ..., x_n$ z $\mathsf{N}(0,1)$. Spočteme možné hodnoty Z jako $z_{i,k} = rx_i + \sqrt{1-r^2}\epsilon_k$, z kterých pomocí funkcí F_Y^{-1} a Φ určíme $y_{i,k} = F_Y^{-1}(\Phi(z_{i,k}))$. Vypočteme odhady $\hat{y}_i = \frac{1}{m}\sum_{k=1}^{m} y_{i,k}$, $i = 1, ..., n$ a průměr všech hodnot $\overline{y} = \frac{1}{mn}\sum_{k=1}^{m}\sum_{i=1}^{n} y_{i,k}$. Dále spočteme reziduální a celkový součet čtverců následovně:

$$\begin{aligned} RSS &= \frac{1}{mn}\sum_{k=1}^{m}\sum_{i=1}^{n} (y_{i,k} - \hat{y}_i)^2, \\ TSS &= \frac{1}{mn}\sum_{k=1}^{m}\sum_{i=1}^{n} (y_{i,k} - \overline{y})^2. \end{aligned}$$

Poté již snadno určíme $R^2 = 1 - \frac{RSS}{TSS}$ při pevném r.

Celkově jsme tuto metodu aplikovali na šest různých rozdělení a pro ověření i na rozdělení normální. Pro některá rozdělení závislá na parametru vyjde pro různé hodnoty parametru R^2 stejně. Jedná se o případy, kdy lze mezi těmito rozděleními přecházet pomocí lineární transformace. Pro názornost si to předveďme na příkladu, kdy $Y \sim \mathsf{N}(\mu, \sigma^2)$. Víme, že $W = \frac{Y-\mu}{\sigma} \sim \mathsf{N}(0,1)$ a $Y = \mu + \sigma W$. Pak $\hat{Y} = \mathsf{E}[Y|X] = \mu + \sigma \mathsf{E}[W|X]$, a tedy

$$RSS = \frac{1}{mn} \sum_{k=1}^{m} \sum_{i=1}^{n} (\mu + \sigma w_{i,k} - (\mu + \sigma \hat{w}_i))^2 = \frac{\sigma^2}{mn} \sum_{k=1}^{m} \sum_{i=1}^{n} (w_{i,k} - \hat{w}_i)^2.$$

Analogicky bychom upravili TSS a po zkrácení σ^2 je rovnost dokázána.

Pro výpočet jsme zvolili $m = 250$ a $n = 1000$. Předpokládali jsme, že pro nízké hodnoty r bude platit aproximativní vztah $R^2 \approx cGini^2$, a proto jsme hledali konstantu $c \approx \frac{R^2}{Gini^2}$. Nejdříve uveďme první sadu rozdělení, u kterých nám vyšlo, že platí aproximace $R^2 \approx Gini^2$. V tabulce uvedeme různé hodnoty r ($\approx Gini$) a konstantu c získanou jako průměr ze tří výpočtů. Pro všechna z uvedených rozdělení je hodnota c nezávislá na parametrech. Označení jednotlivých parametrů je stejné jako v knize [1].

r	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\mathsf{N}(\mu, \sigma^2)$	1.031	1.030	1.028	1.026	1.022	1.019	1.014	1.010	1.005
$\mathsf{R}(a, b)$	0.954	0.955	0.955	0.956	0.959	0.963	0.968	0.976	0.987
$\mathsf{Lgc}(a, b)$[1]	1.045	1.043	1.039	1.034	1.028	1.022	1.016	1.009	1.004
$\mathsf{Ex}(\lambda)$	0.919	0.923	0.927	0.933	0.939	0.946	0.955	0.966	0.980

Další zkoumaná rozdělení byla lognormální $\mathsf{LN}(a, b)$, paretovo $\mathsf{Par}(a, b)$ a beta rozdělení $\mathsf{Beta}(a, b)$. Každé z nich je závislé alespoň na jednom parametru. Proto ty, na kterých je vztah $Gini$ a R^2 závislý označíme hvězdičkou. Uvedeme tři tabulky, kde v prvním řádku budou různé hodnoty parametru (resp. parametrů) a v příslušných buňkách přibližná hodnota c počítána jako průměr z hodnot podílu $c_r = \frac{R^2}{r^2}$ pro $r \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$.

$\mathsf{LN}(a, b*)$	0.1	0.5	1	2	3
c	0.990	0.917	0.662	0.197	0.0872

U Lognormálního rozdělení pro vysoké hodnoty parametru b se c_r nezvyšuje monotónně, ale nejprve se zvyšujícím se r^2 klesají a poté skokovitě vzrostou.

$\mathsf{Par}(a*, b)$	3.1	4	5	6	10
c	0.594	0.680	0.731	0.762	0.813

U paretova rozdělení jsme volili nejmenší $a = 3.1$, aby byl definován alespoň třetí moment. Ve všech zkoušených případech hodnoty c_r rostly monotónně.

Toto rozdělení je závislé na obou parametrech. Z předchozích příkladů se zdá, že s rostoucí šikmostí hodnota c_r klesá. Proto jsme jednou zvolili hodnotu $a = b$,

[1] Lgc(a,b) značí logistické rozdělení s parametry a,b

$\mathsf{Beta}(a*, b*)$	$(4,2)$	$(\frac{1}{20}, \frac{1}{20})$	$(\frac{1}{200}, \frac{1}{2000})$
c	0.954	0.629	0.349

neboť poté vyjde šikmost nulová. Dále si uvědomíme, že pro případ $a = b = 1$ se z beta rozdělení stane rozdělení normální. Výsledky ukazují, že přinejmenším záleží i na jiných faktorech než jen na šikmosti. Pří hodnotách parametrů $(a, b) = (\frac{1}{200}, \frac{1}{2000})$ se jedná o rozdělení velmi podobné alternativnímu s $p = 0.9$. Při použití aproximace (1.22) vyjde, že hodnota $c \approx 0.27$. Vidíme tedy, že výsledky si navzájem neodporují.

Shrňme si některá rozdělení do tabulky, kde uvedeme, šikmost (α_3), špičatost α_4 a odhadnutou hodnotu c.

	α_3	α_4	c
$\mathsf{N}(\mu, \sigma^2)$	0	3	1
$\mathsf{R}(a, b)$	0	$\frac{9}{5}$	0.956
$\mathsf{Ex}(\lambda)$	2	9	0.680
$\mathsf{Par}(4, b)$	7.071	nedefinováno	0.956
$\mathsf{Par}(5, b)$	4.648	73.80	0.731
$\mathsf{Par}(6, b)$	3.810	38.67	0.762
$\mathsf{LN}(a, 0.1)$	0.302	3.162	0.990
$\mathsf{LN}(a, 1)$	6.185	113.9	0.662
$\mathsf{LN}(a, 2)$	414.4	$9.2 \cdot 10^6$	0.197
$\mathsf{Beta}(4, 2)$	-0.467	2.625	0.954
$\mathsf{Beta}(\frac{1}{20}, \frac{1}{20})$	0	1.065	0.629
$\mathsf{Beta}(\frac{1}{200}, \frac{1}{2000})$	-2.846	9.111	0.349

3. Ilustrace

3.1 Data "Cheese"

Data "Cheese" obsahují výsledky experimentu testování chuti čtyř různých sýrů, která jsou na internetu volně ke stažení (cheese-tasting experiment(McCullagh and Nelder, 1989)). Náhodně vybraní jedinci byli požádáni ochutnat jeden ze čtyř různých sýrů a ohodnotit jejich chuť od 0=silná nechuť až k 8=excelentní. Získané údaje lze shrnout do tabulky, kde v prvním sloupci je uveden typ sýru a V prním řádku hodnocení. V jednotlivých buňkách jsou udány počty lidí, kteří daný sýr ohodnotili příšlušným číslem.

Cheese	0	1	2	3	4	5	6	7	8
A	0	0	1	7	8	8	19	8	1
B	6	9	12	11	7	6	1	0	0
C	1	1	6	8	23	7	5	1	0
D	0	0	0	1	3	7	14	16	11

Při zběžném pohledu můžeme odhadnout přibližné pořadí oblíbenosti jednotlivých sýrů, nejoblíbenější se zdá být D následován A,C,B. Sestrojíme model snažící se predikovat chuť sýru podle jeho typu. Tedy vysvětlovaná proměnná Y bude devíti úrovňová a vysvětlující proměnnou X zakódujeme do tří dummy proměnných X_B, X_C, X_D, kde $X_B = 1$ pro typ B a $X_B = 0$ pro jiný typ. Analogicky definujeme X_C a X_D.

Pro fitování použijeme ordinální model proportional odds. Testem zjistíme, že předpoklad shodnosti sloupcových koeficientů není v rozporu s daty. Pro kompletnost připomeňme, že logity jsou definovány tak, aby při kladném koeficientu β_i a zvýšení hodnoty x_i byla zvýšená pravděpodobnost, že Y_i padne do vyšší kategorie, tj.:

$$\text{logit}(\mathsf{P}\left(Y_i \leq k|\boldsymbol{x}_i\right)) = \alpha_k - x_{i,B}\beta_B - x_{i,C}\beta_C - x_{i,D}\beta_D.$$

Pro výpočet ordinálních Giniho koeficientů by nám stačilo znát pouze odhady koeficientů β. Ovšem pro zajímavost vypočteme i Giniho koeficient pro nominální vysvětlovanou proměnnou za pomoci odhadnutých pravděpodobností stejným modelem. Proto uvedeme i odhady koeficientů α.

$\hat{\beta}_B$	$\hat{\beta}_C$	$\hat{\beta}_D$	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$	$\hat{\alpha}_5$	$\hat{\alpha}_6$	$\hat{\alpha}_7$
-3.35	-1.71	1.61	-5.47	-4.41	-3.31	-2.24	-0.91	0.044	1.55	3.11

Skóre spočítáme z odhadnutých koeficientů $\hat{\beta}_i$, tedy $s_i = \boldsymbol{x}_i{}'\hat{\boldsymbol{\beta}} = x_{i,B}\hat{\beta}_B + x_{i,C}\hat{\beta}_C + x_{i,D}\hat{\beta}_D$. Vidíme že v celém souboru se budou vyskytovat jen čtyři různé hodnoty skóre seřazené podle očekávání: $s_B < s_C < s_A < s_D$, kde s_i značí skóre sýru typu i. Nyní spočteme Giniho koeficienty podle čtyř možných definic.

$Gini$	$Gini'$	$Gini_C$	$Gini_n$
0.7184	0.7153	0.5781	0.4674

K vysvětlení vyšších hodnot $Gini$ a $Gini'$ oproti $Gini_C$ se podíváme na tabulku dílčích Giniho koeficientů. Z té je vidět, že velmi vysoké hodnoty se nachází zejména u vzdálených kategorií, které mají v případě prvních dvou koeficientů vyšší váhu. Nízká hodnota $Gini_n$ je způsobena velkým počtem kategorií, mezi kterými nebereme v potaz ordinalitu. Tedy porovnáváme vždy příslušné pravděpodobnosti pro hodnoty navzájem různé, nikoli uspořádané dle velikosti.

$Gini$	0	1	2	3	4	5	6	7
1	-0.0429							
2	0.233	0.274						
3	0.492	0.522	0.304					
4	0.725	0.756	0.520	0.162				
5	0.719	0.739	0.586	0.332	0.218			
6	0.952	0.959	0.885	0.670	0.620	0.312		
7	0.994	0.996	0.966	0.834	0.800	0.536	0.312	
8	1.00	1.00	0.996	0.938	0.904	0.705	0.571	0.280

Tabulka 3.1: Tabulka dílčích Giniho koeficientů

Zkusme nyní různě kategorizovat vysvětlovanou proměnnou Y. Budeme vždy shlukovat sousední kategorie, mezi kterými je logický vztah. Tuto novou proměnnou označíme $Y_{i_1,\dots,i_m}$, kde m je nový počet kategorií a čísla i_j značí počet shluknutých kategorií nově označených jako j. Tedy například trojhodnotovou proměnnou, která vznikne sloučením kategorií 0,1,2,3 na 0, 4,5,6 na 1 a 7,8 na 2, označíme $Y_{4,3,2}$. Podívejme se na číselné hodnoty koeficientů pro různé kategorizace.

	$Y_{2,1,1,1,1,1,2}$	$Y_{3,2,2,2}$	$Y_{1,4,4}$	$Y_{4,1,4}$	$Y_{4,2,3}$	$Y_{3,3,3}$	$Y_{4,5}$	$Y_{5,4}$
$Gini$	0.7172	0.7198	0.7020	0.6907	0.7081	0.7372	0.6774	0.702
$Gini'$	0.7151	0.7176	0.7019	0.6927	0.7095	0.7370	0.6774	0.702
$Gini_C$	0.5853	0.6408	0.6891	0.6489	0.6535	0.6937	0.6774	0.702
$Gini_n$	0.4667	0.5304	0.6606	0.6475	0.5874	0.5799	0.6774	0.702

Tabulka 3.2: Tabulka dílčích Giniho koeficientů

První věcí, kterou zmíníme jsou velmi podobné hodnoty $Gini$ a $Gini'$. Pro libovolná seskupení se liší nejvíce o necelé tři tisíciny. Také vidíme, že se jejich hodnoty při změnách kategorií příliš neliší, neboť obecně lepší diverzifikace mezi skupinami je kompenzována ztrátou vzdálenějších kategorií, mezi kterými model rozlišuje lépe a jsou hodnoceny více. A tedy při sníženém počtu kategorií mohou i klesnout. Naopak $Gini_C$ s klesajícím počtem kategorií roste. To, že se hodnoty koeficientů mohou při nižším počtu kategorií i snižovat se zdá býti jako jejich nevýhoda, neboť při sníženém počtu kategorií nepožadujeme tak přesnou predikci vysvětlovanné proměnné. Ovšem musíme mít na paměti, že naše data vznikla seskupením dat předchozích, a tím pádem již logická vzdálenost mezi kategoriemi nemusí býti tak dobrá. Například přechodem z původní vysvětlovanné proměnné Y na $Y_{4,1,4}$ budeme hodnotit stejně odpovědi 0 a 3, mezi kterými už je poměrně velký rozdíl.

Také hodnota $Gini_n$ zpravidla roste a je nižší než ostatní koeficienty. Z toho je vidět, že informace o uspořádání je cenná a je vhodné ji využít.

Na závěr se zabývejme myšlenkou, která definice $Gini$ pro ordinální proměnnou je ideální. Mezi $Gini$ a $Gini'$ je velmi drobný rozdíl, lze spíše doporučit $Gini'$, jehož hodnota se mění málo při sloučení, či odebrání málo početných kategorií. $Gini_C$ v sobě skrývá informaci o celkovém počtu dvojic, kde uspořádání hodnot odpovídá uspořádání skóre. Proto lze doporučit pracovat vždy s dvojicí $Gini'$ (či $Gini$) a $Gini_C$, neboť nám poskytne jednak informaci o celkovém uspořádání, tak informaci o uspořádání vzdálenějších kategorií.

3.2 Data "Program choice"

Na webové stránce http://www.ats.ucla.edu/stat/R/dae/mlogit.htm je volně stáhnutelný dokument obsahující informace o 200 studentech, kteří nastoupili na střední školu a vybírali si mezi programy general, vocational a academic. Koncept učení těchto tří odvětví je rozdílný u general se studenti zdokonalují ve všech základních předmětech. Vocational program se zaměřuje na konkrétní odborné činnosti, v podstatě mohou být velmi rozdílné od kadeřnictví až po různé montérské či zednické práce. Academic se soustředí na předměty jako matematika, fyzika či programování. Pro pro představu jejich preferencí se podívejme na marginální četnosti zvolených skupin.

n_a	n_v	n_g	n
105	50	45	200

Tabulka 3.3: Relativních četností proměnné *prog*

V uvedených datech se vyskytuje celkem dvanáct proměnných. My si vybereme jen některé z nich, u kterých předpokládáme, že budou mít větší vliv na uvedená rozhodnutí. Proměnné, které použijeme při vytváření modelu, shrňme do tabulky:

názeV	popis	typ proměnné
prog	zvolený program	3 kategoriální nominální
ses	ekonomická situace studenta	3 kategoriální ordinální
schtyp	škola veřejná či soukromá	2 kategoriální nominální
score	vznikne sečtením skóre z testů z *read*, *math*, *write*, *socst*	spojitá

Tabulka 3.4: Proměnné zahrnuté v modelu

K odhadu zvoleného programu použijeme model multinomické regrese. Tedy vysvětlovaná proměnná bude *prog* a jako referenční skupinu zvolíme academic. Vysvětlující jsou uvedené v tabulce výše, jen zmiňme, že proměnnou *ses* zakódujeme pomocí dvou dummy proměnných. Všechny regresní koeficienty vysvětlujících proměnných vyšly významně na pětiprocentní hladině významnosti. Pro přehlednost si uveďme jednotlivé odhady regresních koeficientů vysvětlujících proměnných. Rozeberme nyní jednotlivé odhady. V obou případech vyšší skóre z testů znamená zvýšenou pravděpodobnost volby programu academic. Podobně tomu tak je v případě typu školy, kde soukromá škola naznačuje volbu academic. Odhadnuté koeficienty u ses_{low} a ses_{middle} ukazují, že hůře situovaní studenti dávají přednost academic před general. Ovšem při porovnávání vocation a academic vidíme, že zde není monotonie dodržena, neboť nejvyšší je odhad koeficientu u regresoru ses_{middle}.

Dále pro všechna pozorování určíme odhadnuté pravděpodobnosti, které použijeme pro výpočet Giniho koeficientu a koeficientu determinace, jejichž hodnoty jsou shrnuty v tabulce společně s $Gini_a$, $Gini_g$, $Gini_v$, což jsou Giniho koeficienty

typ programu		odhadnutý koeficient
general	intercept	2.594
	score	-0.016
	ses =low	0.808
	ses =middle	0.581
	schtyp =public	0.560
vocation	intercept	6.372
	score	-0.037
	ses =low	0.133
	ses =middle	1.152
	schtyp =public	1.849

Tabulka 3.5: Tabulka odhadnutých koeficientů v modelu

pro binární proměnné typu $Y_a = I_{[prog=academic]}$ a skórem $s_i^a = \hat{\mathsf{P}}(Y = academic)$. Také si povšimněme, že opravdu platí přibližná rovnost $RSS + ESS \approx TSS$.

$Gini$	$Gini_a$	$Gini_g$	$Gini_v$
0.519	0.576	0.248	0.643

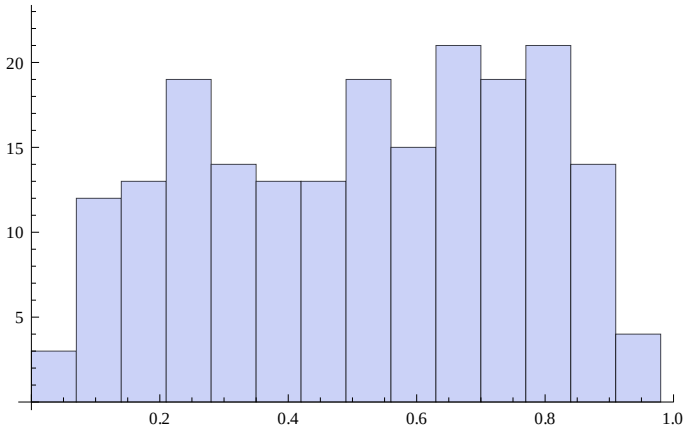
R^2	R^2_a	R^2_g	R^2_v	RSS	ESS	$RSS + ESS$	TSS
0.192	0.252	0.0350	0.259	98.78	22.92	121.70	122.25

Vidíme, že poměrně nízká hodnota $Gini_g$ (resp.R^2_g) zásadně nesníží celkový $Gini$ (resp. R^2_g), neboť studentů, kteří zvolili program general, je nejméně a tím pádem má nejmenší vliv. Poměrně velká hodnota $Gini$ naznačuje, že zpravidla odhadnutá pravděpodobnost $Y = i$ pro prvky v i-té kategorii je vyšší než pro prvky z jiné kategorie. Ovšem nedává nám žádnou informaci o koncentraci jednotlivých odhadů pravděpodobností. Při pohledu na histogram odhadů pravděpodobností vidíme velké rozdíly. Odhady volby academic jsou poměrně rovnoměrně rozdělené, oproti tomu u vocation je velká frekvence nízkých hodnot a pro general se nejvíce odhadů vyskytuje v intervalu od 0.2 do 0.3 a prakticky nepřekročí 0.4. Z toho lze zdůvodnit nižší hodnotu R^2_g, neboť rozdíl $(\hat{p}^i_g - I_{[prog_i=g]})^2$ bude velmi podobný rozdílu $(\frac{n_g}{n} - I_{[prog_i=g]})^2$. Ukazatel $Gini_g$ nám dává informaci o tom, že výchylky od průměru jsou správné jen v něco málo více než polovině případů. Ještě si spočtěme aproximace R^2 pomocí Giniho koeficientů. Spočtené hodnoty jsou velmi podobné hodnotám přesným, žádná z nich se od skutečné neliší více než tři setiny.

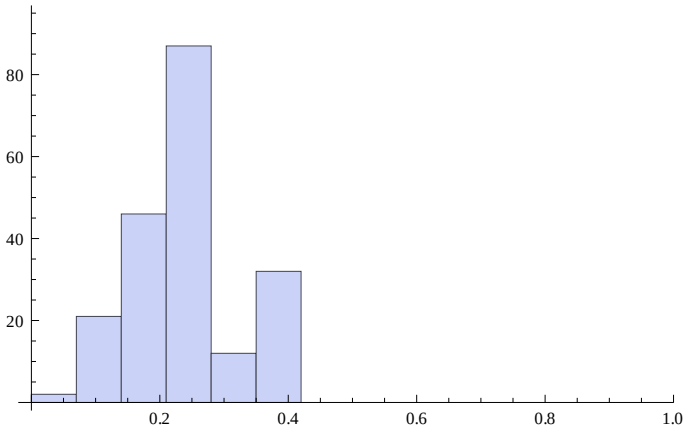
Na tomto příkladu jsme ilustrovali, že pro podrobnější analýzu modelu je dobré spočítat i dílčí koeficienty $Gini$ a R^2, které nám dají představu o problematičnosti určení jednotlivých kategorií.

$3(1-p_a)p_a Gini_a^2$	$3(1-p_g)p_g Gini_g^2$	$3(1-p_v)p_v Gini_v^2$	$3\frac{\sum_{k=a,v,g} p_k^2(1-p_k)^2 Gini_k^2}{\sum_{k=a,v,g} p_k(1-p_k)}$
0.248	0.0321	0.233	0.182

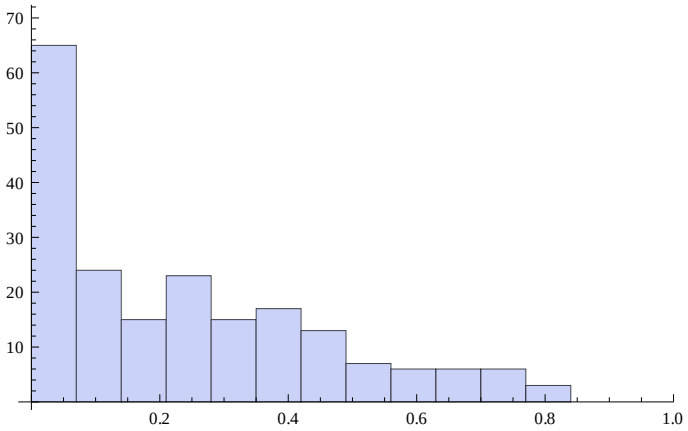
Tabulka 3.6: Tabulka aproximativních vztahů pro R^2



Obrázek 3.1: histogram pravděpodobností *prog*=academic



Obrázek 3.2: histogram pravděpodobností *prog*=general



Obrázek 3.3: histogram pravděpodobností *prog*=vocation

Závěr

V této práci jsme se nejdříve věnovali různým druhům regrese. Byl popsán lineární, logistický, ordinální a multinomický model. Nastínili jsme interpretaci odhadnutých koeficientů a zavedli pojem šance a poměr šancí.

Byly zavedeny běžně používané koeficienty hodnotící kvalitu modelu, či jeho diverzifikační schopnost. Mezi nimi koeficient determinace R^2 a jeho upravená verze tzv. korigovaný $\bar{R}^2$ v případě lineární regrese. v logistické regresi byla zadefinována Lorenzova křivka a z ní vycházející Giniho koeficient a Kolmogorov-Smirnovova statistika. Zabývali jsme se vztahem Giniho koeficientu k jiným používaným charakteristikám diverzifikace jako je C-statistika či Spearmanův korelační koeficient r_S. Právě za pomoci posledně jmenovaného lze za předpokladu normálně rozděleného skóre vyjádřit aproximativní vztah mezi $Gini$ a R^2, který je silně závislý na parametru $G = \mathsf{P}(Y = 1)$ a platí $R^2 \approx 3G(1-G)Gini^2$. Tedy při pevné hodnotě Giniho koeficientu R^2 klesá s klesající pravděpodobností G.

Stěžejní druhá kapitola obsahuje tři různé definice Giniho koeficientu pro ordinální vysvětlovanou proměnnou. Dvě z nich jsou založeny na modifikaci Lorenzovy křivky a berou v potaz rozdíly mezi naměřenými hodnotami. Třetí z nich využívá rozšířenou definici C-statistiky a kvantifikuje poměr správně seřazených skóre podle hodnot vysvětlované proměnné. Všechny tři definice jsou konzistentní s definicí Giniho koeficientu pro binární proměnnou. Na základě teoretických poznatků a ilustrativního příkladu bychom doporučili používání zejména $Gini'$ a $Gini_C$, kde jejich srovnáním získáme informaci o diverzifikaci především vzdálenějších kategorií. Důvodem upřednostnění $Gini'$ před $Gini$ je jeho nízká variabilita při vytvoření či odebrání nové málo početné kategorie. Další jeho výhodou je návaznost na definici Giniho koeficientu pro spojitou vysvětlovanou proměnnou. Koeficient determinace pro tento typ proměnné zaveden nebyl, jelikož odhad vysvětlované proměnné je dán odhadem jednotlivých pravděpodobností a tedy ztrácí přírozenou ordinalitu.

Pro nominální vysvětlovanou proměnnou byl definován Giniho koeficient a koeficient determinace R^2 na základě odhadnutých pravděpodobností padnutí do dané kategorie. Teoreticky byl odvozen a na příkladu ilustrován přibližný vztah $TSS \approx RSS + ESS$. Byly zavedeny binární proměnné, které vzniknou z našich dat seskupením všech kategorií až na jednu. Poté oba nově definované koeficienty vyjádřeny jako vážený průměr odpovídajících koeficientů pro binární proměnné. Bylo zjištěno, že R^2 bude maximální, když dvě kategorie s maximálními dílčími Giniho koeficienty budou každá obsahovat zhruba polovinu všech případů a ostatní kategorie budou zastoupeny jen jedním pozorováním.

Pro spojitou vysvětlovanou proměnnou byl zaveden $Gini$ opět pomocí modifikované Lorenzovy křivky a bylo zjištěno, že při zvolené definici vyjde shodně jako Spearmannův korelační koeficient. Za předpokladu normálně rozdělených skóre a nepřesnosti predikce byl sestrojen model, pomocí kterého jsme zkoumali vztah R^2 a $Gini^2 = r_S^2$. Pro srovnání s výsledky práce [4] jedno ze zkoumaných rozdělení bylo beta rozdělení s koeficienty nastavenými tak, aby bylo velmi blízké alternativnímu. Získané výsledky si navzájem neodporovali.

V ilustrativní části byly na dvou příkladech aplikovány modely multinomické a ordinální regrese a spočteny hodnoty zadefinovaných koeficientů.

Seznam použité literatury

[1] Anděl, Jiří.
Základy matematické statistiky. 2. vydání. Praha: MatfyzPress, 2011.
ISBN 80-7378-001-1.

[2] Cipra, Tomáš.
Finanční Ekonometrie. 1. vydání. Praha: Ekopress, 2008.
ISBN 978-80-86929-43-9.

[3] Hosmer, D. W. a Lemeshow, S.
Applied Logistic Regression. John Wiley Sons, Inc., 2000.
ISBN 0-471-35632-8.

[4] Ondrušková, Markéta.
Odhadování a kritéria těsnosti modelu logistické regrese, Bakalářská práce, MFF UK, 2011.

[5] Zvára, Karel.
Regrese. 1. vydání. Praha:MatfyzPress, 2008.
ISBN 978-80-7378-041-8.

[6] *What are pseudo R-squareds?*[online] UCLA: Academic Technology Services, Statistical Consulting Group. Poslední změna 20.10.2011 [cit. 1.8.2012].
Dostupné z
http://www.ats.ucla.edu/stat/mult_pkg/faq/general/psuedo_rsquareds.htm