Univerzita Karlova v Praze
Přírodovědecká fakulta

DISERTAČNÍ PRÁCE

Jana Rubešová

Statistické metody pro hodnocení predikční validity

Statistical methods for evaluation of predictive validity

Ústav aplikací matematiky a výpočetní techniky

Školitel: doc. RNDr. Karel Zvára, CSc.
Studijní program: Aplikovaná matematika
Studijní obor: Zpracování dat a matematické modelování v přírodních vědách

Praha 2009

Srdečně děkuji svému školiteli, doc. RNDr. Karlu Zvárovi, CSc., za poskytnutí mnoha podnětných konzultací v souvislosti s disertační prací, ale i za trpělivost a ochotu, se kterou mne v průběhu celého studia zasvěcoval do tajů matematické statistiky.

Prohlašuji, že jsem disertační práci vypracovala samostatně a výhradně s použitím citovaných pramenů. Práci ani její podstatnou část jsem nepřeložila k získání jiného nebo stejného akademického titulu.

Jana Rubešová

Obsah

Seznam obrázků

Seznam tabulek

Abstrakt

Název práce: Statistické metody pro hodnocení predikční validity
Autor: Jana Rubešová
Katedra (ústav): Ústav aplikací matematiky a výpočetní techniky
Školitel: doc. RNDr. Karel Zvára, CSc.

Abstrakt: Práce pojednává o statistických metodách a modelech, které lze použít pro hledání a analýzu faktorů ovlivňujících budoucí úspěšnost ve studiu vysoké školy. Jedná se o problémy řešené v celosvětovém měřítku.
Pro zkoumání závislosti konečného průměru známek na dříve známých veličinách jsou využity modely mnohonásobné lineární regrese. Důraz je kladen též na méně běžné postupy pro porovnání síly vlivů jednotlivých regresorů. Model logistické regrese, kdy odhadujeme pravděpodobnost úspěchu ve studiu (dané konkrétním kritériem), je doplněn analýzou ROC křivek.
Analýzy provedené na datovém souboru 1340 studentů Univerzity Karlovy v Praze, Přírodovědecké fakulty, kteří se zapsali ke studiu bakalářských studijních programů v akademickém roce 2003/04 a 2004/05, jsou porovnány s dalšími studiemi zabývajícími se predikční validitou.
Klíčová slova: predikční validita, přijímací zkoušky, korelace, lineární regrese, logistická regrese, ROC křivka.

Abstract

Title: Statistical methods for evaluation of predictive validity
Author: Jana Rubešová
Department: Institute of Applied Mathematics and Information Technologies
Advisor: doc. RNDr. Karel Zvára, CSc.

Abstract: The thesis deals with statistical methods and models suitable to analyze factors that may inluence on future success in graduation. These problems are solved all over the world.
Multiple regression models are used for detection dependence final college grade points average on other known variables like admission exam scores and high school grades. The thesis emphasizes less usual methods for comparing effect power of regressors. Logistic regression model where we estimate probability of success in study are complemented by ROC analysis.
Analyses provided on the data of 1340 students of Charles University, Faculty of Science who matriculated to bahelor study programs in 2003/04 and 2004/05 academic years where compared with other Czech and international publications about predictive validity.
Keywords: predictive validity, admission tests, correlation, linear regression, logistic regression, ROC curve.

Úvod

Tématem disertační práce jsou matematické metody a modely, které je možno použít pro hodnocení predikční validity. Predikční validitou nazýváme schopnost předpovídat nějaké kritérium (v našem případě úspěšnost studia na vysoké škole) na základě dříve provedených testů či jiných známých skutečností.

V první kapitole jsou podrobně popsány matematické modely, které je možno pro hodnocení predikční validity použít. Publikace, zabývající se tímto tematem, jsou většinou úzce zaměřeny na konkrétní metody pro vybraná kritéria úspěšnosti, proto jsem se snažila o komplexnější pojetí problematiky.

Pro demonstraci jednotlivých statistických metod, které doplňují teoretický výklad, jsem použila data o 140 studentech bakalářského studijního oboru Geografie–kartografie, kteří se zapsali ke studiu na Přírodovědecké fakultě v akademickém roce 2003/04. Tato skupina studentů byla pro prezentaci metod zvolena, protože zahrnuje největší homogenní skupinu studentů (všichni konali přijímací zkoušky z matematiky a geografie, jejich studijní plány obsahují až na několik výjimek ve vyšších ročnících tytéž předměty).

Druhá kapitola je zaměřena prakticky. Jsou zde podrobně analyzovány údaje získané ze studijního informačního systému Univerzity Karlovy týkající se studentů bakalářských studijních programů, kteří se zapsali ke studiu na Přírodovědecké fakultě v akademickém roce 2003/04 a 2004/05. Všechny výpočty byly provedeny ve volně šiřitelném statistickém programu R verze 2.9.1. Zjištěné výsledky jsou porovnány se závěry dalších studií českých i zahraničních (zejména amerických) autorů.

Přílohy obsahují popis jednotlivých položek datového souboru, simulace doplňující některé teoretické partie a texty všech skriptů, kde je zaznamenán postup výpočtu. Tyto skripty (včetně dalších potřebných funkcí) jsou součástí přiloženého CD, zájemci mají možnost kterýkoli krok analýzy zopakovat. Datový soubor je na CD uložen jednak ve formátu pro použití v programu R, jednak jako soubor pro MS Excel.

Kapitola 1

Statistické metody a modely

Kritéria hodnocení úspěšnosti studentů dělíme na dvě různé skupiny podle použitého měřítka. Kvantitativním kritériem je nejčastěji celkový průměr všech známek za studium vysoké školy, případně průměry známek za jednotlivé ročníky, počet opravných termínů, atd., kdy používáme metody regresní a korelační analýzy.

Za kvalitativní kritéria úspěšnosti můžeme považovat zjištění, zda student studium vůbec dokončil, zda ho dokončil ve standardní době či později, zda absolvoval s vyznamenáním, zda úspěšně splnil podmínky pro zápis do 2. ročníku, apod. Data v tomto případě analyzujeme nejčastěji metodami logistické regrese.

1.1 Korelační analýza

1.1.1 Pearsonův korelační koeficient

Pearsonův (momentový) korelační koeficient vyjadřuje sílu lineární závislosti mezi dvěma číselnými (spojitými) náhodnými veličinami, kde hodnoty obou veličin zpravidla měříme na jednom subjektu.

Mějme náhodné veličiny X a Y, u nichž předpokládáme dvourozměrné normální rozdělení

$$\mathsf{N}\left(\begin{pmatrix}\mu_X\\ \mu_Y\end{pmatrix}, \begin{pmatrix}\sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y\\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2\end{pmatrix}\right).$$

Populační korelační koeficient ρ_{XY} je

$$\rho_{XY} = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2\sigma_Y^2}} \tag{1.1}$$

kde σ_{XY} je kovariance definovaná

$$\sigma_{XY} = \mathsf{E}\,(X-\mu_X)(Y-\mu_Y)$$

Populační hodnotu korelačního koeficientu (1.1) zpravidla neznáme. Jako odhad ρ_{XY} lze použít výběrový korelační koeficient [36].

$$r_{XY} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{s_{XY}}{s_X \cdot s_Y}$$

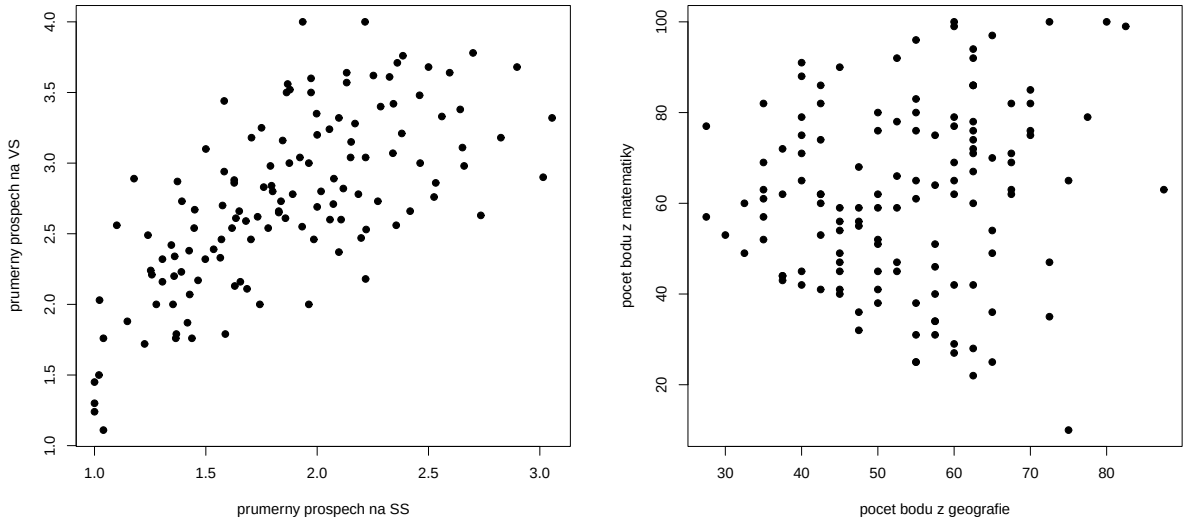
$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1.2}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_X}\right)\left(\frac{y_i - \bar{y}}{s_Y}\right) \tag{1.3}$$

Korelační koeficient nezávisí na volbě měřítka ani jedné z veličin, jak je vidět ze zápisu (1.3), kde jsou násobeny bezrozměrné z-skóry a platí $-1 \leq r_{XY} \leq +1$, kde $|r_{XY}| = 1$ nastává u deterministické lineární závislosti. Korelační koeficient je kladný, když s rostoucími hodnotami jedné veličiny v průměru rostou i hodnoty druhé, a záporný, když s rostoucími hodnotami jedné veličiny hodnoty druhé veličiny spíše klesají.

Chceme-li názorně zobrazit vzájemnou závislost číselných veličin, použijeme bodový graf (scatter plot). Dvojice hodnot (x_i, y_i) vynášíme jako body do kartézské souřadné soustavy.

Příklad: Chceme graficky znázornit závislost průměrného prospěchu na střední a vysoké škole. Omezíme se pouze na 129 studentů, kteří dostali na VŠ alespoň jednu známku. Na vodorovnou osu vynášíme průměr známek na SŠ, na svislou osu dosažený průměr na vysoké škole. Obr. 1.1 (vlevo) ukazuje zřetelnou souvislost mezi oběma veličinami, což odpovídá poměrně vysoké hodnotě vypočteného korelačního koeficientu ($r_{XY} = 0{,}686$).

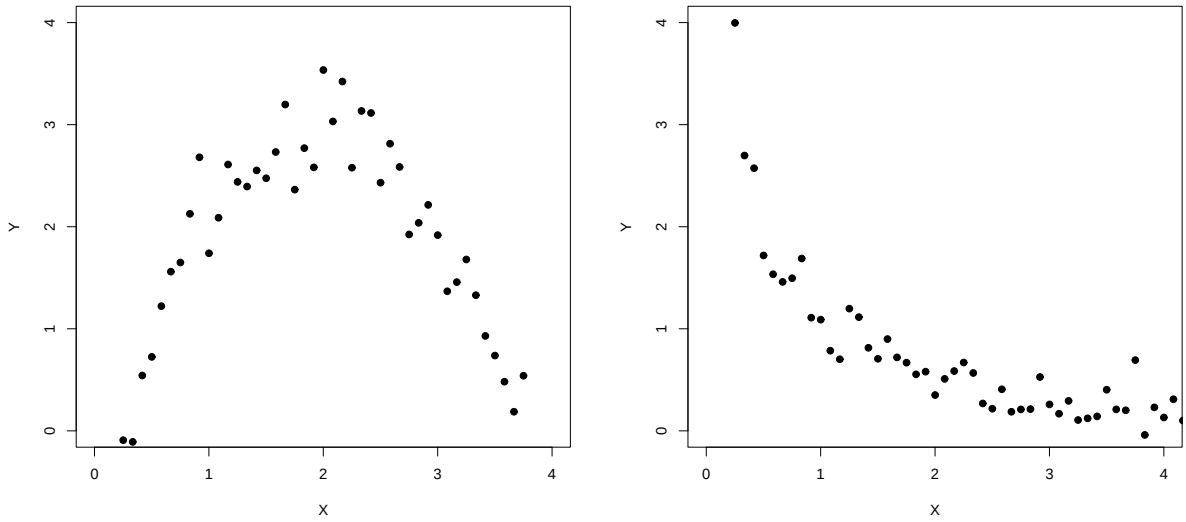


Obrázek 1.1: Souvislost průměrného prospěchu na SŠ a VŠ (vlevo) a počtu bodů z jednotlivých částí přijímacích zkoušek (vpravo)

Zcela jiné je znázornění souvislosti jednotlivých částí přijímacích zkoušek u stejných studentů. Na obr. 1.1 (vpravo) je na vodorovné ose vyznačen počet bodů z geografie a na svislé ose počet bodů dosažených v testu z matematiky. Body reprezentující jednotlivé studenty jsou téměř náhodně rozmístěny po celé ploše grafu, což odpovídá nízké

absolutní hodnotě korelačního koeficientu $r_{XY} = -0{,}129$. Samozřejmě chybí pozorování v levém dolním rohu s malými počty bodů z obou zkoušek, protože takoví studenti nebyli přijati. $\diamondsuit$

Vypočtená hodnota výběrového korelačního koeficientu blízká nule může znamenat nejen nezávislost dvou náhodných veličin, ale i situaci, kdy jejich závislost je jiná než lineární. Na obrázku 1.2 jsou příklady veličin, jejichž závislost není lineární. Situace vlevo svědčí spíše pro kvadratickou závislost, v pravé části obrázku body odpovídají nepřímé úměrnosti tvaru $Y = 1/X$.



Obrázek 1.2: Příklady nelineárních závislostí veličin

1.1.2 Test významnosti

Prokázat lineární závislost dvou veličin je možné využitím vztahu, že pro dvě nezávislé veličiny X, Y platí $\rho_{XY} = 0$. Zamítneme-li nulovost korelačního koeficientu ρ_{XY} při splnění předpokladu dvourozměrného normálního rozdělení, zamítneme zároveň hypotézu o nezávislosti X a Y.

Náhodná veličina T má za platnosti $\mathsf{H}_0 : \varrho_{XY} = 0$ rozdělení t_{n-2}.

$$T = \frac{r_{XY}}{\sqrt{1 - r_{XY}^2}} \sqrt{n-2}\,. \tag{1.4}$$

Příklad: Chceme prokázat, že prospěch na vysoké škole souvisí s prospěchem na střední škole, což je graficky znázorněno na obr. 1.1. Bereme v úvahu 129 studentů, kteří dostali alespoň jednu známku. V testu normality průměrného prospěchu na VŠ je $p = 0{,}398$, u průměrného prospěchu na střední škole (resp. průměru z průměrů na výročních vysvědčeních v 1. až 3. ročníku a pololetním ze 4. ročníku) je $p = 0{,}152$. Vypočteme $r_{XY} = 0{,}686$ a po dosazení do vzorce (1.4) dostaneme

$$T = \frac{0{,}686}{\sqrt{1 - 0{,}686^2}} \sqrt{129-2} = \frac{0{,}686}{\sqrt{1 - 0{,}471}} \sqrt{127} = 10{,}632. \tag{1.5}$$

Protože $|T| > t_{n-2}(0{,}05) = 1{,}98$, zamítáme nezávislost veličin na každé běžně používané hladině ($p < 0{,}00001$).

Při obdobném výpočtu týkající se obou částí přijímací zkoušky z obr. 1.1 (vpravo) vypočteme $T = 1{,}466$, takže závislost obou veličin neprokážeme ($p = 0{,}145$). $\diamondsuit$

1.1.3 Interval spolehlivosti pro populační korelační koeficient

Podobně jako jsme na základě hodnoty výběrového průměru schopni určit interval spolehlivosti pro populační průměr, existují metody pro výpočet přibližného intervalu spolehlivosti populačního korelačního koeficientu. Jednou z nich je využití tzv. Fisherovy z-transformace korelačního koeficientu r, která má na rozdíl od korelačního koeficientu r rozptyl téměř nezávislý na hodnotě ρ, viz [2].

Položíme-li

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r}\,, \tag{1.6}$$

pak

$$\mathsf{E}\, Z \doteq \frac{1}{2} \ln \frac{1+\rho}{1-\rho}\,, \quad \mathsf{var}\, Z \doteq \frac{1}{n}\,.$$

Přesnější výpočet vede k výsledku

$$\mathsf{E}\, Z \doteq \frac{1}{2} \ln \frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}\,, \quad \mathsf{var}\, Z \doteq \frac{1}{n-3}\,,$$

ale v praktických výpočtech se používají aproximace

$$\mathsf{E}\, Z \doteq \frac{1}{2} \ln \frac{1+\rho}{1-\rho}\,, \quad \mathsf{var}\, Z \doteq \frac{1}{n-3}.$$

Transformovaná veličina zpravidla označovaná symbolem Z má tedy za předpokladu, že populační korelační koeficient je roven ρ, přibližně rozdělení

$$\mathsf{N}\Big(\tfrac{1}{2} \ln \tfrac{1+\rho}{1-\rho}, \tfrac{1}{n-3}\Big)$$

Označíme

$$\zeta = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$$

a platí

$$\mathsf{P}\left[\sqrt{n-3}\,|Z-\zeta| < z(\alpha/2)\right] \doteq 1-\alpha.$$

Krajní meze přibližného intervalu spolehlivosti pro ζ jsou

$$Z_L = Z - z(\alpha/2)/\sqrt{n-3}\,,$$
$$Z_U = Z + z(\alpha/2)/\sqrt{n-3}\,.$$

Postupnými úpravami dostaneme:

$$\begin{aligned} \zeta &= \frac{1}{2} \ln \frac{1+\rho}{1-\rho} \\ \mathrm{e}^{2\zeta} &= \frac{1+\rho}{1-\rho} \\ \mathrm{e}^{2\zeta} - \rho\mathrm{e}^{2\zeta} &= 1+\rho \\ \mathrm{e}^{2\zeta} - 1 &= \rho(\mathrm{e}^{2\zeta}+1) \\ \rho &= \frac{\mathrm{e}^{2\zeta}-1}{\mathrm{e}^{2\zeta}+1} \end{aligned} \tag{1.7}$$

Inverzní transformací podle vzorce (1.7) určíme krajní meze přibližného intervalu spolehlivosti pro ρ.

Příklad: Chceme určit interval pro populační korelační koeficient závislosti průměrného prospěchu na vysoké a střední škole. Hladinu α volím 5 %, tedy populační korelační koeficient bude tímto intervalem překryt přibližně s pravděpodobností 95 %. Transformací výběrového korelačního koeficientu $r_{XY} = 0{,}686$ vypočteme

$$Z = \frac{1}{2} \ln \frac{1 + 0{,}686}{1 - 0{,}686} = 0{,}841$$

Meze intervalu spolehlivosti pro ζ jsou

$$Z_L = Z - z(\alpha/2)/\sqrt{129 - 3} = 0{,}666$$

a

$$Z_U = Z + z(\alpha/2)/\sqrt{129 - 3} = 1{,}015.$$

Dosazením vypočtených hodnot Z_L a Z_U do vzorce (1.7) dostaneme přibližný interval spolehlivosti $(0{,}582, 0{,}768)$. Ačkoliv je tento interval pouze přibližný, zjevně neobsahuje 0, což je zcela v souladu s přesným testem (1.5), kde jsme hypotézu o nulovosti populačního korelačního koeficientu zamítli. $\diamondsuit$

1.1.4 Test shody korelačních koeficientů

Použití správné metody pro hodnocení shody korelačních koeficientů závisí na použitých datech, resp. jejich struktuře.

Odpovídající veličiny v nezávislých výběrech

Předpokládejme, že máme dva nezávislé výběry o rozsazích n_1 a n_2. Oba výběry zahrnují stejné nebo odpovídající si veličiny X a Y. Výše uvedenou z-transformaci můžeme využít pro testování hypotézy o shodě populačního korelačního koeficientu ρ_1 mezi veličinami X a Y z prvního výběru s korelačním koeficientem ρ_2 stejných veličin z druhého výběru.

Provedeme Fisherovu z-transformaci výběrových korelačních koeficientů r_1 a r_2:

$$Z_1 = \frac{1}{2} \ln \frac{1 + r_1}{1 - r_1}, \quad Z_2 = \frac{1}{2} \ln \frac{1 + r_2}{1 - r_2}.$$

Protože rozptyl rozdílu veličin $Z_1 - Z_2$ je přibližně roven $1/(n_1 - 3) + 1/(n_2 - 3)$, použijeme testovou statistiku

$$Z^* = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}. \tag{1.8}$$

Nulovou hypotézu o shodě korelačních koeficientů zamítneme na aproximativní hladině významnosti α v případě, že $|Z^*| \geq z(\alpha/2)$ ([36]).

Příklad: Ve skupině studentů chceme zjistit, zda se liší korelační koeficienty závislosti prospěchu na vysoké a střední škole u chlapců a dívek. Prospěch na VŠ je znám

pouze u 67 (n_1) chlapců a 62 (n_2) dívek. Korelační koefiecient ve skupině chlapců je $r_1 = 0{,}643$, ve skupině dívek $r_2 = 0{,}778$. Použitím z-transformace dostaneme:

$$Z_1 = \frac{1}{2} \ln \frac{1 + 0{,}643}{1 - 0{,}643} = 0{,}763$$

a

$$Z_2 = \frac{1}{2} \ln \frac{1 + 0{,}778}{1 - 0{,}778} = 1{,}040.$$

Dosazením do (1.8) vypočteme:

$$Z^* = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{67-3} + \frac{1}{62-3}}} = 1{,}533 < z(0{,}025) = 1{,}96.$$

Tedy jsme na 5% hladině neprokázali rozdíl síly lineární závislosti mezi chlapci a dívkami. $\diamondsuit$

Závislé dvojice veličin v tomtéž výběru

Nepříliš známé a používané postupy se týkají situace, kdy chceme zjistit, zda se korelační koeficient veličin X_1 a Y liší od korelačního koeficientu mezi X_2 a Y. Y, X_1 a X_2 jsou proměnné z výběru z trojrozměrného normálního rozdělení, kde u každého subjektu známe hodnoty všech tří veličin. Korelační matice veličin (Y, X_1, X_2) je

$$\mathbf{P} = \begin{pmatrix} 1 & \rho_{01} & \rho_{02} \\ \rho_{01} & 1 & \rho_{12} \\ \rho_{02} & \rho_{12} & 1 \end{pmatrix}, \tag{1.9}$$

kde index 0 odpovídá veličině Y, 1 je použita pro zkrácené označení X_1 a 2 pro X_2.

Pro test hypotézy $\mathsf{H}_0 : \rho_{01} = \rho_{02}$ je možno použít Hotellingovu statistiku T_1 [14]

$$T_1 = (r_{01} - r_{02}) \cdot \sqrt{\frac{(n-3) \cdot (1 + r_{12})}{2|\mathbf{R}|}}, \tag{1.10}$$

kde výběrové korelační koeficienty r_{ij} odpovídají populačním charakteristikách ρ_{ij} a $|\mathbf{R}|$ je determinant matice výběrových korelačních koeficientů odpovídající korelační matici $\mathbf{P}$.

Statistika T_1 má za platnosti nulové hypotézy rozdělení t_{n-3}. Rozhodování je však podmíněno konkrétní realizací náhodných veličin X_1 a X_2.

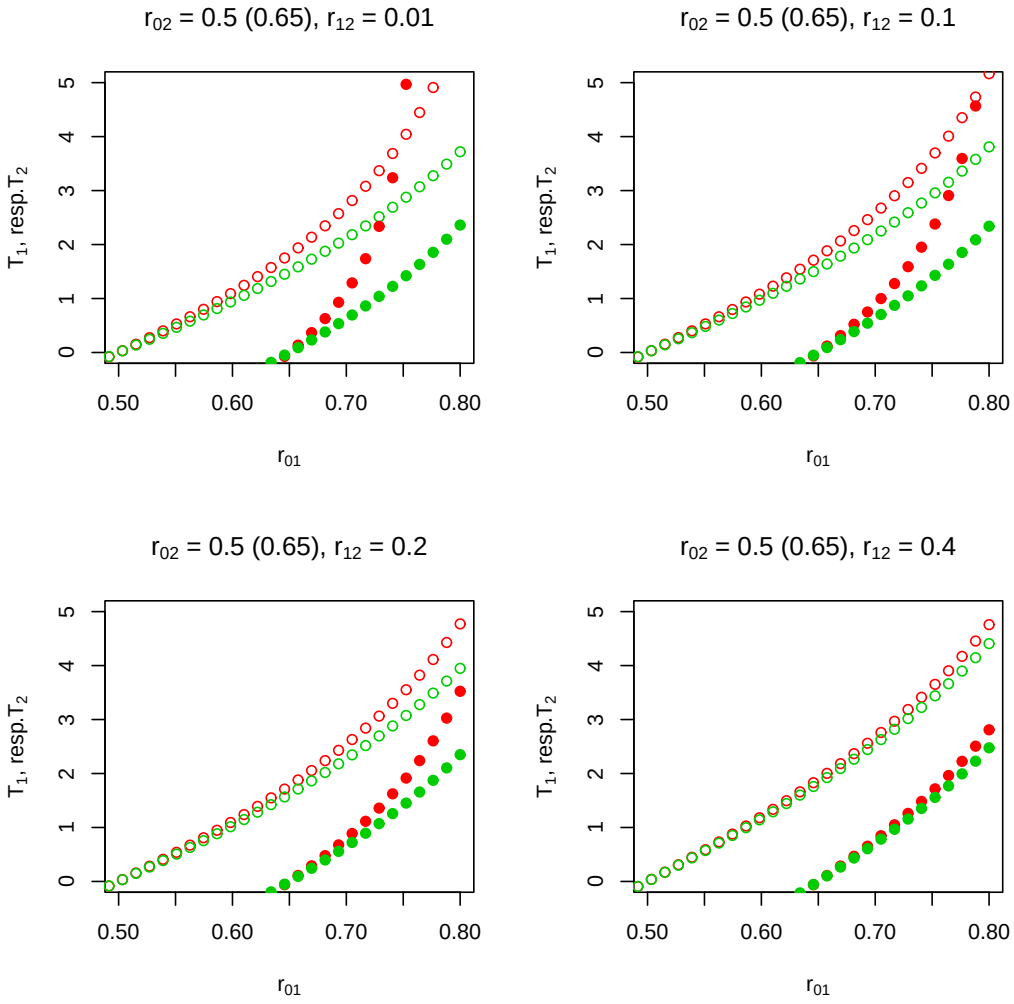
E. J. Williams [34] navrhl modifikaci T_1, kde upravil vzorec (1.10) tak, aby byl zohledněn větší rozptyl rozdílu výběrových korelačních koeficientů způsobený variabilitou veličin X_1, X_2.

$$T_2 = (r_{01} - r_{02}) \cdot \sqrt{\frac{(n-1) \cdot (1 + r_{12})}{2\frac{n-1}{n-3}|\mathbf{R}| + ((r_{01} + r_{02})/2)^2 (1 - r_{12})^3}} \tag{1.11}$$

Statistika T_2 má za platnosti hypotézy asymptoticky rozdělení t_{n-3}. Podmíněnost rozdělení byla odstraněna za cenu přibližného rozdělení.

Neill a Dunn ve svém článku [24] porovnali 11 testů hypotézy $\mathsf{H}_0 : \rho_{01} = \rho_{02}$. Na základě provedených simulací dávají přednost statistice T_2, která má vyšší sílu testu a lépe drží hladinu α než T_1 .

Také Steiger [30] zdůvodňuje preferenci použití statistiky T_2 z (1.11) před T_1 z (1.10) a poukazuje na skutečnost, že i v některých psychologických statistických pracech bylo nevhodně doporučováno T_1. Dále zmiňuje situaci, kdy za předpokladu $\rho_{12} = 0$, $\rho_{01} = \rho_{02} = \sqrt{0{,}5}$ je na základě hodnoty statistiky T_1 téměř vždy zamítnuta nulová hypotéza $\mathsf{H}_0 : \rho_{01} = \rho_{02}$. Ze vzorce (1.10) je zřejmá velká citlivost statistiky T_1 na singularitu varianční matice veličin (Y, X_1, X_2). Když Steiger předpokládá, že X_1 a X_2 jsou nezávislé, singularitu může způsobit jedině skutečnost, že Y je lineární funkcí X_1 a X_2. To znamená, že podmíněné rozdělení Y při daných hodnotách $X_1 = x_1$ a $X_2 = x_2$ je degenerované (s pravděpodobností 1 je Y konstanta). V praxi můžeme podobnou situaci těžko očekávat, ale je zřejmé, že když bude možno predikovat Y pomocí regrese velmi dobře (skoro přesně), je Hotellingův postup nepoužitelný. V příloze B.1 je ukázka simulace, kdy určujeme hodnoty T_1 a T_2 za výše zmíněných předpokladů.



Obrázek 1.3: Porovnání statistik T_1 (červeně) a T_2 (zeleně) při různých hodnotách výběrových korelačních koeficientů a rozsahu výběru $n = 100$

Rozdíly mezi statistikami T_1 a T_2 pro některé konkrétní hodnoty výběrových korelačních koeficientů jsou znázorněny na obrázku 1.3.

Příklad: Chceme porovnat, zda se liší vliv středoškolského průměru a přijímacích

zkoušek na průměr známek na VŠ. Veličinou Y je známkový průměr na VŠ, X_1 průměrný prospěch na SŠ a X_2 je dosažený počet bodů u přijímacích zkoušek, opět u studentů geografie. Korelační koeficienty jsou $r_{01} = 0{,}686$, $r_{02} = -0{,}639$ a $r_{12} = -0{,}426$. Vzhledem k tomu, ze mezi prospěchem a dosaženým počtem bodů u přijímacích zkoušek je záporná korelace, pro porovnání obrátíme znaménka (odpovídalo by změně stupnice známkování, kdy vyšší známka znamená lepší prospěch). Determinant matice výběrových korelačních koeficientů je $|\mathbf{R}| = 0{,}313$.

Dosazením do (1.11) postupně dostaneme

$$\begin{aligned} T_2 &= (0{,}686 - 0{,}639) \cdot \sqrt{\frac{128 \cdot (1 + 0{,}426)}{2\frac{128}{126} \cdot 0,313 + \left(\frac{0{,}686+0{,}639}{2}\right)^2 \cdot (1 - 0{,}426)^3}} \\ &= 0,047 \cdot \sqrt{\frac{182,48}{0,636 + 0,439 \cdot 0,190}} = 0,752 \end{aligned}$$

Vzhledem k tomu, že $|T_2| < t_{126}(0{,}05) = 1{,}98$, nemůžeme zamítnout hypotézu o stejném vlivu přijímacích zkoušek a středoškolského prospěchu ($p = 0{,}453$). Podobné výsledky bychom dostali v tomto případě i ze vzorce (1.10), kde vyjde $T_1 = 0{,}799$. $\diamondsuit$

1.1.5 Odhad korelačního koeficientu z neúplných dat

Zkoumáme závislost dvou veličin X a Y, u kterých předpokládáme dvourozměrné normální rozdělení

$$\mathsf{N}\left(\begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix}, \begin{pmatrix}\sigma_x^2 & \rho\sigma_x\sigma_y\\ \rho\sigma_x\sigma_y & \sigma_y^2\end{pmatrix}\right).$$

Hodnoty X_i známe u všech subjektů, ale hodnoty Y_j pouze u části, tedy máme k dispozici n úplných dvojic (X_i, Y_i) a u dalších $N - n$ subjektů neznáme hodnotu Y_i:

$$\begin{aligned} &x_1, \ldots, x_n, x_{n+1}, \ldots, x_N \\ &y_1, \ldots, y_n. \end{aligned}$$

Maximálně věrohodné odhady parametrů $\mu_x, \mu_y, \sigma_x, \sigma_y$ a ρ nezávisle na sobě odvodili Cohen [9] a Anderson [3].

Při odvození se vychází z toho, že sdruženou hustotu všech X, Y lze zapsat jako součin hustoty X a hustoty podmíněného rozdělení $Y|X = x$, které je však opět normální $\mathsf{N}(\mu_{y.x} + \beta x, \sigma_{y.x})$, kde

$$\beta = \rho\sigma_y/\sigma_x, \tag{1.12}$$

$$\begin{aligned} \mu_{y.x} &= \mu_y - \beta\mu_x, \\ \sigma^2_{y.x} &= \sigma_y^2(1 - \rho^2). \end{aligned} \tag{1.13}$$

Maximalizací logaritmické věrohodnostní funkce odhadneme parametry μ_x a σ_x^2 marginálního rozdělení veličiny X

$$\hat{\mu}_x = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i, \qquad \hat{\sigma}_x^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

a parametry $\beta, \mu_{y.x}, \sigma_{y.x}$ podmíněného rozdělení

$$\hat{\beta} = \frac{\sum_{j=1}^{n}(X_j - \bar{X}^*)(Y_j - \bar{Y}^*)}{\sum_{j=1}^{n}(X_j - \bar{X}^*)^2}$$
$$\hat{\mu}_{y.x} = \bar{Y}^* - \hat{\beta}\bar{X}^*$$
$$\hat{\sigma}^2_{y.x} = \frac{1}{n}\sum_{j=1}^{n}((Y_j - \bar{Y}^*) - \hat{\beta}(X_j - \bar{X}^*)^2,$$

kde $\bar{X}^*$ a $\bar{Y}^*$ jsou aritmetické průměry z úplných dvojic pozorování.

Odhad ρ odvodíme z (1.12) a (1.13) vyloučením $\sigma^2_{y.x}$ a dostaneme vztah

$$\rho = \frac{\beta\sigma_x}{\sqrt{\beta^2\sigma_x^2 + \sigma^2_{y.x}}}.$$

Postupnými úpravami nakonec dostaneme odhad

$$\hat{\rho} = \frac{r_{xy}}{\sqrt{1 - \lambda(1 - r^2_{xy})}}, \tag{1.14}$$

kde r_{xy} je výběrový korelační koeficient spočítaný z úplných dvojic pozorování a

$$\lambda = 1 - \frac{\frac{1}{n}\sum_{j=1}^{n}(X_j - \bar{X}^*)^2}{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}. \tag{1.15}$$

Koeficient λ vlastně měří neshodu mezi odhadem rozptylu X pouze z úplných dvojic pozorování a odhadem rozptylu ze všech pozorování této veličiny. Jsou-li chybějící hodnoty veličiny Y dány omezením na X (například studenti s nízkým počtem bodů u přijímacích zkoušek nejsou přijati), je zpravidla odhad rozptylu X z úplných dvojic (čitatel zlomku v (1.15)) menší než odhad ze všech pozorování (jmenovatel zlomku v (1.15)). Koeficient λ pak vyjde kladný a $|\hat{\rho}| > |r_{xy}|$.

Lze očekávat, že úprava pomocí náhodného λ zvětší rozptyl maximálně věrohodného odhadu $\hat{\rho}$ v porovnání s r_{xy}. Pochopitelně záleží na tom, jakým způsobem se rozhoduje, která pozorování Y nejsou k dispozici. V situaci, kdy chybí hodnoty Y takové, že odpovídající X je menší než pevné x_0, lze dokázat [23], že asymptotický rozptyl odhadu $\hat{\rho}$ je dán výrazem

$$\mathsf{var}\,(\hat{\rho}) \doteq \frac{(1-\rho^2)^2}{N} \cdot \frac{2 - \rho^2\Phi(\frac{x_0-\mu_x}{\sigma_x})}{2(1 - \Phi(\frac{x_0-\mu_x}{\sigma_x}))}. \tag{1.16}$$

Nepříjemnou vlastností tohoto odhadu je skutečnost, že může být velice citlivý na splnění předpokladu o normálním rozdělení.

Na základě provedení simulací v příloze B.2 se jeví, že odhad není ani tak citlivý na normalitu, ale spíše na symetrii rozdělení. Porušení předpokladu normality resp. symetrie může nastat jednak u veličiny X s kompletními údaji, jednak u veličiny Y, kde část pozorování chybí. V jednotlivých případech dostaneme odlišné výsledky. Jestliže má jedna z veličin například rovnoměrné rozdělení, které je symetrické, nedochází k podstatné deformaci výsledků. Jestliže má však veličina Y, u které neznáme všechny hodnoty, exponenciální rozdělení (případně obecnější gamma rozdělení), které je výrazně asymetrické, výpočet podle vzorce (1.14) hodnotu opraveného korelačního koeficientu

výrazně nadhodnocuje. Při nejistotě o možném rozdělení je tedy třeba dbát zvýšené opatrnosti.

Typickým příkladem použití je situace, kdy známe výsledky všech studentů u přijímacích zkoušek, ale jen u těch, kteří byli přijati, zapsali se a studovali, i průměr známek na vysoké škole. Rozdělení průměru známek na vysoké škole se může na jednotlivých školách či oborech lišit a nebývá vždy blízké normálnímu, což může výpočet znehodnotit.

Příklad: Podle normálního diagramu (i testu normality) můžeme přepokládat normální rozdělení průměrného prospěchu na VŠ, a proto můžeme využít (1.14).

Na základě znalosti korelačního koeficientu mezi průměrným prospěchem na VŠ a počtem bodů dosažených v přijímacím řízení $r_{xy} = -0{,}639$, odhadneme korelační koeficient u všech uchazečů.

Ze vzorce (1.15) vypočteme hodnotu koeficientu $\lambda = 1 - \frac{615{,}87}{1229{,}04} = 0{,}499$. Opravený korelační koeficient pak vychází:

$$\hat{\rho} = \frac{-0{,}639}{\sqrt{1 - 0{,}499(1 - (-0{,}639)^2)}} = -0{,}761. \diamondsuit$$

1.1.6 Spearmanův korelační koeficient

Mnohdy chceme prokázat závislost dvou veličin, ale není splněn předpoklad normality. Pro prokázání monotónní závislosti se využívá Spearmanův korelační koeficient.

Původní pozorování x_i a y_i nahradíme pořadími R_i a Q_i a dále aplikujeme vzorec pro výpočet Pearsonova korelačního koeficientu (1.2) na zjištěná pořadí ve tvaru

$$r_{XY}^{(S)} = \frac{\sum_{i=1}^n R_i Q_i - n\bar{R}\bar{Q}}{\sqrt{(\sum_{i=1}^n R_i^2 - n\bar{R}^2)(\sum_{i=1}^n Q_i^2 - n\bar{Q}^2)}}\,. \tag{1.17}$$

Využitím vztahů

$$\bar{R} = \frac{1}{n}\sum_{i=1}^n R_i = \frac{1}{n}\sum_{i=1}^n i = \frac{n+1}{2} = \bar{Q}\,,$$

$$\sum_{i=1}^n R_i^2 = \sum_{i=1}^n Q_i^2 = \sum_{i=1}^n i^2 = \frac{n(n+1)(2n+1)}{6}\,, \tag{1.18}$$

$$\sum_{i=1}^n (R_i - Q_i)^2 = \sum_{i=1}^n R_i^2 + \sum_{i=1}^n Q_i^2 - 2\sum_{i=1}^n R_i Q_i\,, \tag{1.19}$$

dosazením z (1.18) do (1.19) vyjádříme

$$\sum_{i=1}^n R_i Q_i = \frac{n(n+1)(2n+1)}{6} - \frac{1}{2}\sum_{i=1}^n (R_i - Q_i)^2$$

a vzorec (1.17) můžeme upravit na tvar (pro jednodušší výpočet).

$$r_{XY}^{(S)} = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^n (R_i - Q_i)^2. \tag{1.20}$$

Pro malé hodnoty n se používají speciální tabulky kritických hodnot, pro vyšší n je výpočetní náročnost velká, ale vzhledem k platnosti centrální limitní věty nulovou hypotézu $r_{XY}^{(S)} = 0$ zamítneme, platí-li $r_{XY}^{(S)}\sqrt{n-1} \geq z(\alpha/2)$ (viz např. [2]).

Je důležité si uvědomit, že vzorec (1.20) dává tentýž výsledek jako (1.17) pouze za předpokladu, že všechna pozorování x_i i všechna pozorování y_i jsou navzájem různá. Jinak totiž neplatí vztah (1.18), kde se vlastně jedná o hodnotu součtu druhých mocnin přirozených čísel od 1 do n. V případě shodných pozorování by byl skutečný součet vždy o něco menší.

Jestliže se v datech vyskytuje mnoho shodných hodnot pozorování, doporučuje se pracovat s korigovaným Spearmanovým korelačním koeficientem (viz [15] a [29])

$$r_{XY}^{(S,\mathrm{korig})} = 1 - \frac{6}{n(n^2-1) - T_{x'} - T_{y'}} \sum_{i=1}^{n} (R_i - Q_i)^2 \,, \tag{1.21}$$

kde

$$T_{x'} = \frac{1}{2}\sum (t_{x'}^3 - t_{x'}), \qquad T_{y'} = \frac{1}{2}\sum (t_{y'}^3 - t_{y'})$$

a $t_{x'}$, resp. $t_{y'}$ jsou rozsahy skupin se stejnou hodnotou veličiny X, resp. Y.

Příklad: Chceme-li například prokázat závislost mezi počtem bodů u přijímacích zkoušek a dosaženým průměrem známek na VŠ (ve skupině 89 studentů, kteří studium úspěšně absolvovali) a testy u obou veličin zamítnou normalitu ($p = 0{,}028$, resp. $p = 0{,}015$), můžeme použít Spearmanův korelační koeficient. Jeho hodnota je v tomto případě $r_{XY}^{(S)} = -0{,}5229$. Výpočty podle vzorce (1.20) bez odečtení vlivu shodných pozorování ($-0{,}5225$) i podle (1.21) ($-0{,}5227$) dávají v tomto případě velmi podobné výsledky, protože opakování hodnot je spíše výjimkou. U přijímacích zkoušek se vyskytují devětkrát dvě stejné hodnoty a třikrát tři stejné hodnoty, tedy $T_{x'} = 1/2\,(9\cdot 6 + 3\cdot 24) = 63$. prospěch na VŠ obsahuje šestkrát dvě stejné hodnoty a jednou tři stejné hodnoty, tedy $T_{y'} = 1/2\,(6\cdot 6 + 1\cdot 24) = 30$. $\diamondsuit$

1.2 Lineární regresní model

V následujícím textu nám nepůjde pouze o to, prokázat, že spolu dvě či více veličin souvisí, ale též odhadnout příslušnou závislost. Chceme vysvětlit variabilitu náhodné veličiny Y (závisle proměnná, vysvětlovaná proměnná, odezva) závislostí její střední hodnoty na jedné nebo několika nezávisle proměnných (prediktorech, regresorech).

Mějme náhodný vektor $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ a matici daných čísel $\mathbf{X}_{n\times k}$. Předpokládáme, že se $\mathbf{Y}$ řídí tzv. lineárním modelem

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}\,, \tag{1.22}$$

kde $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)'$ je vektor neznámých parametrů a $\mathbf{e} = (e_1, \ldots, e_n)'$ je vektor nezávislých náhodných veličin, které mají rozdělení $\mathsf{N}(\mathbf{0}, \sigma^2\mathbf{I})$, kdy σ^2 je též neznámý parametr. Vektor $\mathbf{e}$ je možno charakterizovat jako vektor chyb, kterými se zpravidla rozumějí chyby vyplývající z nepřesností při stanovování vektoru Y. Vektor $\mathbf{e}$ se pozorovat nedá. Předpoklad nulové střední hodnoty pro všechna e_i odpovídá tomu, že pozorování Y_i nejsou zatížena systematickými chybami. Diagonální varianční matice (spolu s předpokladem normality) znamená, že chyby jednotlivých měření Y_i jsou nezávislé, stejný rozptyl odpovídá stejné přesností měření. V mnoha případech se nejedná v pravém slova smyslu o chyby měření, ale např. o biologickou variabilitu.

1.2.1 Regresní přímka

Nejjednodušší je situace, kdy odhadujeme střední hodnoty závisle proměnné na základě známých hodnot jediné nezávisle proměnné.

Mějme lineární model $Y_i = \beta_0 + \beta_1 x_i + e_i$, kde $Y_1, Y_2, \ldots, Y_n$ jsou nezávislé náhodné veličiny, o jejichž středních hodnotách předpokládáme:

$$\mathsf{E}\,Y_i = \beta_0 + \beta_1 x_i, \qquad i = 1, \ldots, n\,, \tag{1.23}$$

kde $x_1, \ldots, x_n$ jsou známé konstanty a β_0, β_1 jsou neznámé parametry. Dále předpokládáme, že pro všechna i platí $\mathsf{var}\,Y_i = \sigma^2$.

Předpokládáme, že střední hodnota náhodné veličiny Y_i je lineární funkcí známé hodnoty x_i, a tedy body o souřadnicích $[x_i, \mathsf{E}\,Y_i]$ leží na přímce. Rovnici této přímky neznáme, protože neznáme její parametry v rovnici (1.23).

Střední hodnotu náhodné veličiny Y_i (vektor $\mathsf{E}\,\mathbf{Y}$) odhadujeme pomocí lineární funkce $\hat{Y}_i = b_0 + b_1 x_i$. Rezidui nazýváme rozdíly mezi skutečnou a odhadovanou hodnotou, tj. $U_i = Y_i - \hat{Y}_i$. Reziduální součet čtverců, označovaný RSS, SS_e nebo S_e, udává druhou mocninu normy (délky) rozdílu vektorů $\mathbf{Y}$ a $\hat{\mathbf{Y}}$ a jediným číslem tak vyjadřuje jejich neshodu.

$$RSS = \sum_{i=1}^{n} U_i^2 = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 x_i))^2 \tag{1.24}$$

Odhady b_0, b_1 parametrů β_0, β_1 z rovnice (1.23) určíme *metodou nejmenších čtverců*, kdy hledáme minimum funkce dvou proměnných (b_0 a b_1) z (1.24). Položíme-li obě první parciální derivace rovny nule, řešením soustavy dvou lineárních rovnic o dvou neznámých dostaneme

$$\begin{aligned} b_0 &= \bar{Y} - b_1\bar{x}\,, \\ b_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{aligned} \tag{1.25}$$

kde

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{a} \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Metoda nejmenších čtverců souvisí s předpokladem o normalitě chyb měření e_i. Pokud bude tento předpoklad splněn, odhady metodou maximální věrohodnosti (jejímž konkrétním případem v tomto případě metoda nejmenších čtverců je) budou mít dobré statistické vlastnosti.

Odvození koeficientů b_0 a b_1 můžeme popsat i následujícím způsobem, kdy hledáme řešení rovnice:

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{Y},$$

kde $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$ a $\mathbf{X}$ je regresní matice (matice modelu)

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

Koeficienty b_0, b_1 dostaneme řešením soustavy lineárních rovnic

$$\begin{aligned} \mathbf{X}'\mathbf{X}\mathbf{b} &= \mathbf{X}'\mathbf{Y}, \\ \begin{pmatrix} 1 & \ldots & 1 \\ x_1 & \ldots & x_n \end{pmatrix} \cdot \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \cdot \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} &= \begin{pmatrix} 1 & \ldots & 1 \\ x_1 & \ldots & x_n \end{pmatrix} \cdot \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \\ \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} \cdot \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} &= \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \end{pmatrix}. \end{aligned} \tag{1.26}$$

Koeficient b_1 můžeme interpretovat také následujícím způsobem (viz [35]). Předpokládáme, že všechna pozorování veličiny x se liší od aritmetického průměru $\bar{x}$. Pak můžeme vzorec (1.25) upravit

$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \sum_{i}^{n} \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \frac{Y_i - \bar{Y}}{x_i - \bar{x}} \\ &= \sum_{i}^{n} w_i \,\mathsf{tg}\,\alpha_i\,, \end{aligned}$$

kde váha w_i je

$$w_i = \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \tag{1.27}$$

a tedy součet vah w_i je roven 1 a α_i je úhel, který s vodorovnou osou svírá přímka spojující body $[x_i, Y_i]$ a $[\bar{x}, \bar{Y}]$. Směrnice regresní přímky je tedy váženým průměrem směrnic všech přímek, které procházejí napozorovanými body $[x_i, Y_i]$ a jejich těžištěm $[\bar{x}, \bar{Y}]$. Z (1.27) je vidět, že každý bod má tím větší váhu, čím více je jeho x-ová souřadnice x_i vzdálena od průměru $\bar{x}$. Jestliže by byly hodnoty Y_i pro takovéto x_i zatíženy

hrubými chybami, mohlo by dojít k výraznému zkreslení odhadů regresních koeficientů.

Parametry regresní přímky můžeme vypočítat vždy pro pozorování, kde nejsou všechny hodnoty x_i stejné. Lineární závislost veličiny Y na x je statisticky průkazná (na zvolené hladině α) v případě, že směrnice regresní přímky β_1 je statisticky významně nenulová. Chceme-li testovat $\mathsf{H}_0 : \beta_1 = 0$, vypočteme hodnotu testové statistiky T, která má za platnosti H_0 rozdělení t_{n-2}.

$$T = \frac{b_1}{\mathsf{S.E.}(b_1)} = \frac{b_1}{s}\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,, \tag{1.28}$$

kde

$$s^2 = \frac{RSS}{n-2} = \frac{1}{n-2}\left(\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - b_1\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})\right)\,. \tag{1.29}$$

Testová statistika T ze vzorce (1.28) je tatáž jako T statistika ze vzorce (1.4) pro test nulovosti korelačního koeficientu.

Koeficient determinace

Koeficient determinace ukazuje, jak velký díl variability hodnot závisle proměnné se nám podařilo vysvětlit uvažovanou závislostí při daných hodnotách nezávisle proměnné $x_1, \ldots, x_n$.
Celková variabilita náhodné veličiny Y je

$$SS_T = \sum_{i=1}^{n}(Y_i - \bar{Y})^2\,,$$

variabilita nevysvětlená modelem (reziduální součet čtverců)

$$RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\,,$$

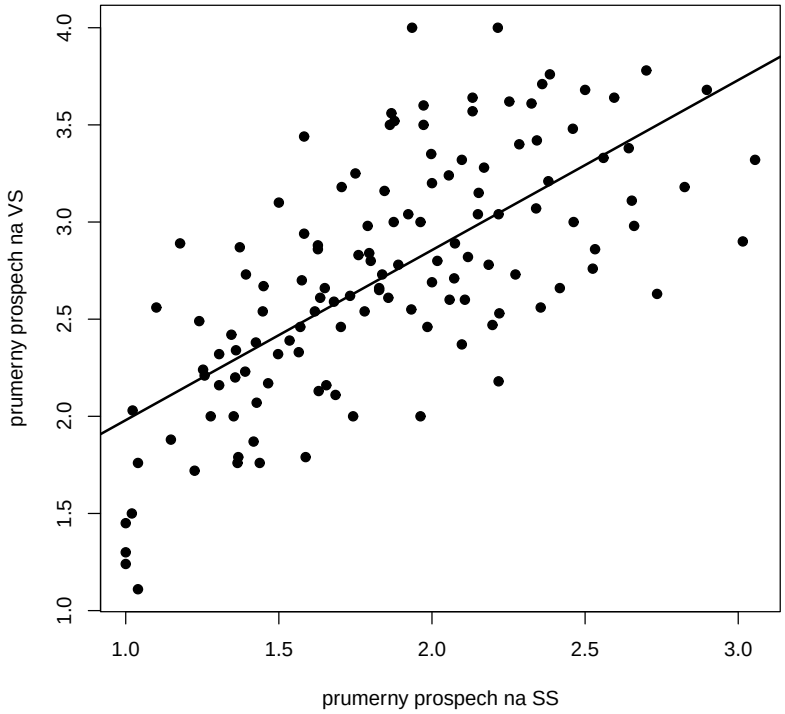
a tedy

$$R^2 = \frac{SS_T - RSS}{SS_T} = 1 - \frac{RSS}{SS_T} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}. \tag{1.30}$$

Vzhledem k (1.30) platí $0 \leq R^2 \leq 1$, kdy $R^2 = 1$ platí pro deterministickou lineární závislost. V případě jediného regresoru je koeficient determinace roven druhé mocnině korelačního koeficientu mezi veličinami Y a $\hat{Y}$, který je vzhledem k lineárnímu vztahu $\hat{Y} = b_0 + b_1 x$ stejný jako korelační koeficient veličin Y a x.

U modelu s více regresory je $R = \sqrt{R^2}$ označován jako výběrový koeficient mnohonásobné korelace.

Příklad: Již dříve jsme prokázali, že prospěch na střední a vysoké škole spolu souvisí, nyní odhadneme parametry modelu závislosti a ověříme předpoklady použití modelu. Do výpočtu je zahrnuto 129 studentů, kteří dostali alespoň jednu známku.



Obrázek 1.4: Regresní přímka modelu závislosti prospěchu na VŠ na předchozím prospěchu na SŠ

Dosazením vypočtených koeficientů b_0, b_1 dostaneme rovnici regresní přímky, která je znázorněna na obrázku 1.4:

$$prumer = 1{,}105 + 0{,}875 \cdot ssprum\,, \tag{1.31}$$

kde *prumer* je průměrný prospěch na vysoké škole a *ssprum* je průměr ze tří průměrů na výročních vysvědčeních a pololetním vysvědčení v posledním ročníku střední školy.

Dosazením do (1.28) vychází T statistika pro test hypotézy $\mathsf{H}_0 : \beta_1 = 0$

$$T = \frac{0{,}875}{0{,}0823} = 10{,}632,$$

což odpovídá stejné hodnotě při testu nulovosti korelačního koeficientu v (1.5).

Ze směrnice regresní přímky v (1.31) je zřejmé, že liší-li se průměr dvou studentů na střední škole o jeden stupeň, očekávaný rozdíl na vysoké škole je 0,875. Například u studentů s průměrným prospěchem 2,00 na střední škole očekáváme na vysoké škole v průměru prospěch 2,855.

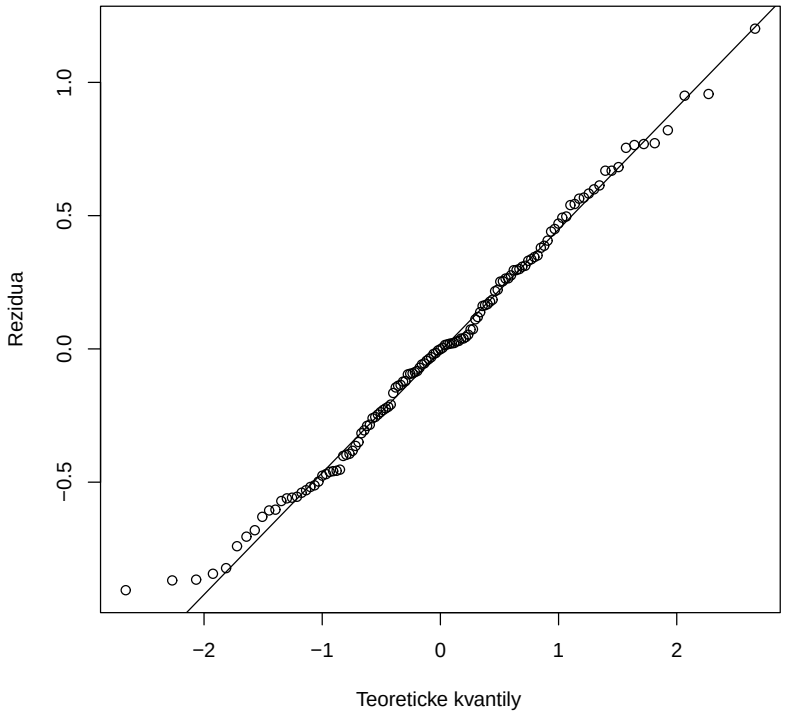
Koeficient determinace modelu vyjde $R^2 = 0{,}471$, což znamená, že 47 % variability prospěchu na vysoké škole jsme vysvětlili závislostí na prospěchu na střední škole.

Normalita reziduí

Pro grafické znázornění se používá tzv. normální diagram (probability plot). Jednotlivá rezidua jsou po uspořádání vynášena na osu y, kde jejich x-ová souřadnice odpovídá

teoretickému kvantilu normovaného normálního rozdělení. Hodnocení normálního diagramu je založeno na představě, že kdyby byl $U_1, \ldots, U_n$ náhodný výběr z rozdělení $\mathsf{N}(\mu, \sigma^2)$, zobrazované body by měly náhodně kolísat kolem přímky $y = \mu + \sigma x$. Pokud body naznačují konvexní závislost, je to známka kladné šikmosti (normální rozdělení má šikmost nulovou), konkávní průběh je důsledkem záporné šikmosti. Esovitý průběh naznačuje jinou špičatost než přepokládáme u normálního rozdělení.

Číselným hodnocením kvality přiblížení bodů k přímce je Shapirův–Wilkův test normality [28], jehož testová statistika je blízká čtverci výběrového korelačního koeficientu mezi hodnotami reziduí a kvantily uspořádaného výběru z $\mathsf{N}(0, 1)$.



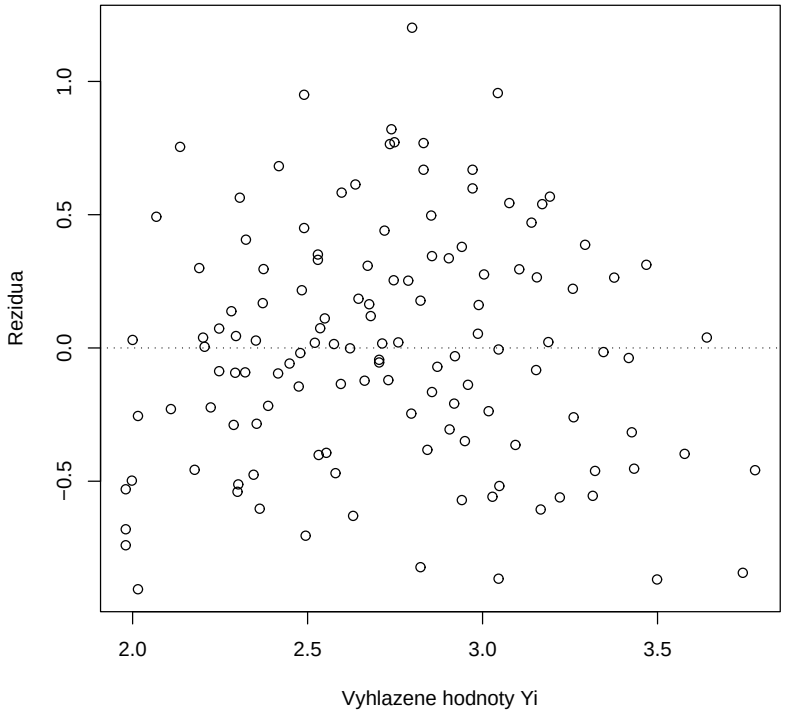
Obrázek 1.5: Normální diagram

Na obrázku 1.5 je znázorněn normální diagram reziduí hodnoceného modelu, kde je patrné, že není problém s normalitou. Odpovídá tomu i vysoká p-hodnota 0,53 Shapirova–Wilkova testu normality.

Homogenita rozptylu (homoskedasticita)

Častým případem porušení předpokladu o konstantním rozptylu je monotonní závislost rozptylu na střední hodnotě veličiny Y. Pro grafické znázornění se používá bodový graf, kde na ose x vynášíme vyhlazené hodnoty $\hat{Y}_i$ a na svislé ose rezidua $Y_i - \hat{Y}_i$ (případně jejich druhé mocniny).

Testová statistika Breuschova–Paganova testu, který testuje homoskedasticitu proti monotónní závislosti na střední hodnotě [6], je blízká testové statistice nulovosti směrnice závislosti čtverců reziduí na vyhlazených hodnotách $\hat{Y}_i$. Na obr. 1.6 jsou body rozmístěny náhodně, což odpovídá nezávislosti rozptylu na střední hodnotě Y_i ($p = 0{,}56$).



Obrázek 1.6: Závislost reziduí na střední hodnotě odhadované veličiny

Boxova–Coxova transformace

V některých případech je při nesplnění předpokladů normálního lineárního modelu nutno využít transformací. Použitím vhodné funkce nezávisle proměnné jako regresoru dostaneme bohatší množinu možných středních hodnot vysvětlované veličiny. Kvalitativně jiná situace nastane při transformaci závisle proměnné, kdy chceme linearizovat závislost, stabilizovat rozptyl, přiblížit vliv náhodné složky normálnímu rozdělení apod.

Boxova–Coxova transformace [5] je pro kladné y definována předpisem

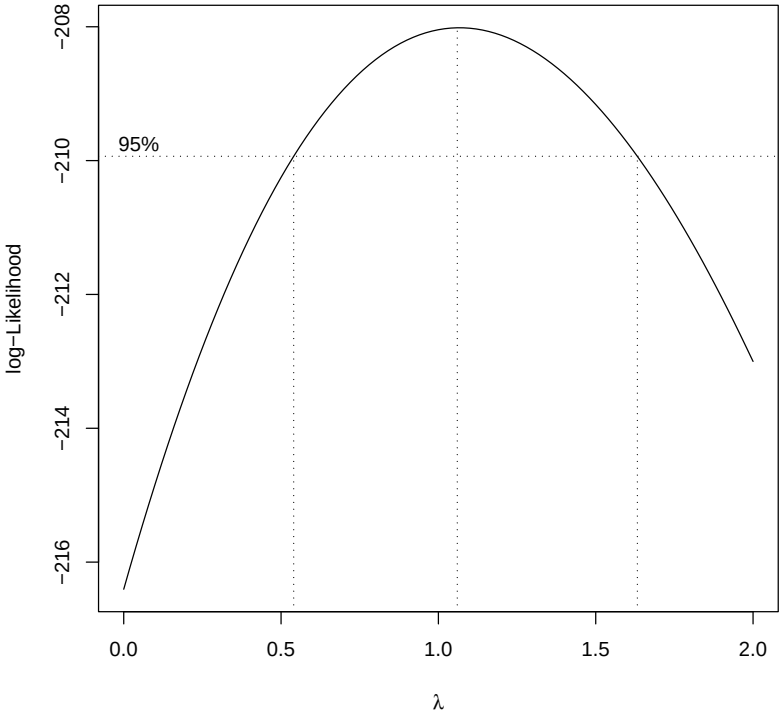
$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & \text{pro } \lambda \neq 0 \\ \ln y & \text{pro } \lambda = 0. \end{cases} \tag{1.32}$$

Běžný lineární model modifikujeme tak, že předpokládáme přibližnou platnost modelu

$$\mathbf{Y}^{(\lambda)} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})\,. \tag{1.33}$$

Pro pevné λ minimalizuje logaritmickou věrohodnostní funkci parametrů $\boldsymbol{\beta}, \sigma^2$ odhad metodou nejmenších čtverců $\mathbf{b}^{(\lambda)}$ v modelu (1.33).

Grafické znázornění závislosti hodnoty logaritmické věrohodnostní funkce na zvolené hodnotě λ je na obrázku 1.7. Vzhledem ke splnění dříve uvedených předpokladů lineárního modelu závislosti průměrného prospěchu nepřekvapuje, že $\lambda = 1$ je takřka uprostřed intervalu spolehlivosti a tedy není třeba provést transformaci vysvětlované veličiny (průměrného prospěchu na vysoké škole).



Obrázek 1.7: Maximálně věrohodný odhad parametru λ Boxovy–Coxovy transformace modelu závislosti prospěchu na VŠ na průměrném prospěchu na SŠ, kde je vyznačen přibližný 95% interval spolehlivosti pro λ

1.2.2 Mnohonásobná lineární regrese – dva regresory

Mějme lineární model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 v_i + e_i\,.$$

Koeficient β_1 lze interpretovat jako střední změnu odhadované veličiny Y při jednotkové změně x a nezměněné hodnotě v, koeficient β_2 jako střední změnu Y při jednotkové změně v a nezměněné hodnotě x. Regresní funkce má vyjádření:

$$\mathsf{E}\, Y_i = \beta_0 + \beta_1 x_i + \beta_2 v_i, \qquad i = 1, \ldots, n\,. \tag{1.34}$$

Odhady regresních koeficientů dostaneme jako v případě jediného regresoru řešením soustavy lineárních rovnic:

$$\mathbf{X'Xb} = \mathbf{X'Y} \tag{1.35}$$

$$\begin{pmatrix} 1 & \ldots & 1 \\ x_1 & \ldots & x_n \\ v_1 & \ldots & v_n \end{pmatrix} \cdot \begin{pmatrix} 1 & x_1 & v_1 \\ \vdots & \vdots & \vdots \\ 1 & x_n & v_n \end{pmatrix} \cdot \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 1 & \ldots & 1 \\ x_1 & \ldots & x_n \\ v_1 & \ldots & v_n \end{pmatrix} \cdot \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

$$\begin{pmatrix} n & \sum x_i & \sum v_i \\ \sum x_i & \sum x_i^2 & \sum x_i v_i \\ \sum v_i & \sum x_i v_i & \sum v_i^2 \end{pmatrix} \cdot \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \\ \sum v_i Y_i \end{pmatrix}$$

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y} \tag{1.36}$$

Reziduální součet čtverců *RSS* a reziduální rozptyl s^2 vypočteme

$$\begin{aligned} RSS &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\ s^2 &= \frac{RSS}{n-3} \end{aligned}$$

Pro test hypotézy $\mathsf{H}_0 : \beta_j = 0, (j = 1, 2)$, určíme

$$T_j = \frac{b_j}{\mathsf{S.E.}(b_j)} = \frac{b_j}{s\sqrt{w_{jj}}}\,, \tag{1.37}$$

kde w_{jj} je odpovídající prvek matice $(\mathbf{X}'\mathbf{X})^{-1}$. Jestliže platí $|T_j| \geq t_{n-3}(\alpha)$, zamítneme H_0; prokázali jsme závislost vysvětlované veličiny na j-tém regresoru.

Mějme model závislosti průměrného prospěchu na vysoké škole na předchozím prospěchu na střední škole a počtu bodů získaných z přijímacích zkoušek. Podle (1.37) prokážeme nenulovost koeficientů u obou regresorů. Dosazením odhadnutých koeficientů dostaneme regresní funkci:

$$prumer = 2{,}725 + 0{,}645 \cdot ssprum - 0{,}0104 \cdot zcel\,.$$

Závislost na měřítku

Sílu vlivu obou regresorů nemůžeme posuzovat na základě vypočtených odhadů regresních koeficientů. Jejich velikost závisí na zvoleném měřítku. Kdybychom například úspěšnost v přijímacím řízení hodnotili počtem procent z maximálně dosažitelných 200 bodů (každou hodnotu bychom vydělili dvěma), vyšla by regresní funkce

$$prumer = 2{,}725 + 0{,}645 \cdot ssprum - 0{,}0207 \cdot zcel/2\,,$$

ze které je patrná dvojnásobná hodnota koeficientu u regresoru *zcel*.

Z původních hodnot regresorů vypočteme bezrozměrné z-skóry, které vyjadřují, o kolik směrodatných odchylek se jednotlivá pozorování liší od průměru. Při použití takto standardizovaných veličin *zssprum* a *zzcel* už můžeme hodnoty koeficientů porovnat. V rovnici regresní funkce

$$prumer = 2{,}730 + 0{,}308 \cdot zssprum - 0{,}258 \cdot zzcel$$

má koeficient u průměrného prospěchu na střední škole větší hodnotu, ale porovnáním absolutních hodnot příslušných korelačních koeficientů neprokážeme rozdíl v síle vlivu obou regresorů (využitím vzorců (1.10), resp. (1.11) z kap. 1.1.4).

Model s interakcemi

V předchozím jsme předpokládali, že vliv obou regresorů je aditivní (citlivost na změnu jednoho regresoru je stejná pro všechny hodnoty druhého regresoru). Přítomnost interakcí v modelu

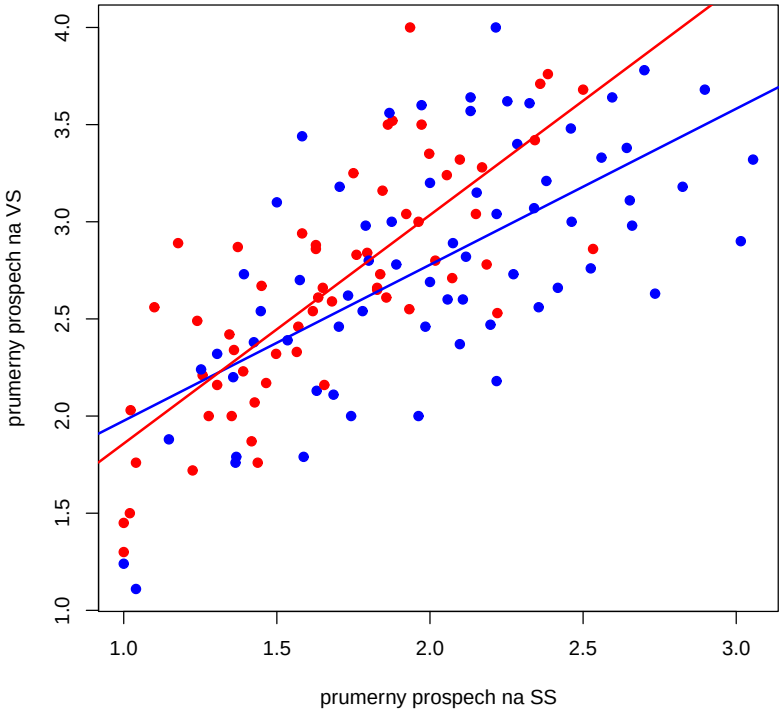
$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 v_i + \beta_3 x_i v_i + e_i$$

prokážeme zamítnutím hypotézy $\mathsf{H}_0 : \beta_3 = 0$.

Příklad: Mějme model závislosti průměrného prospěchu na vysoké škole na předchozím prospěchu na střední škole a pohlaví studenta:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 v_i + \beta_3 x_i v_i + e_i \,, \tag{1.38}$$

kde x je veličina $ssprum$, v je faktor označující pohlaví studenta. Při výpočtu se nahrazuje faktor s I úrovněmi $I - 1$ umělými proměnnými; tady jednou novou proměnnou, která označuje, zda se jedná o dívku.



Obrázek 1.8: Grafické znázornění závislosti průměrného prospěchu na VŠ na pohlaví studentů a předchozím prospěchu na střední škole. Modře jsou označeni chlapci, červeně dívky.

Interakce mezi pohlavím studenta a středoškolským prospěchem jsou v tomto modelu průkazné ($p = 0{,}036$). Směrnice regresních přímek popisujících závislosti jsou statisticky významně odlišné u chlapců a dívek, jak je znázorněno na obrázku 1.8. Dosazením odhadů regresních koeficientů

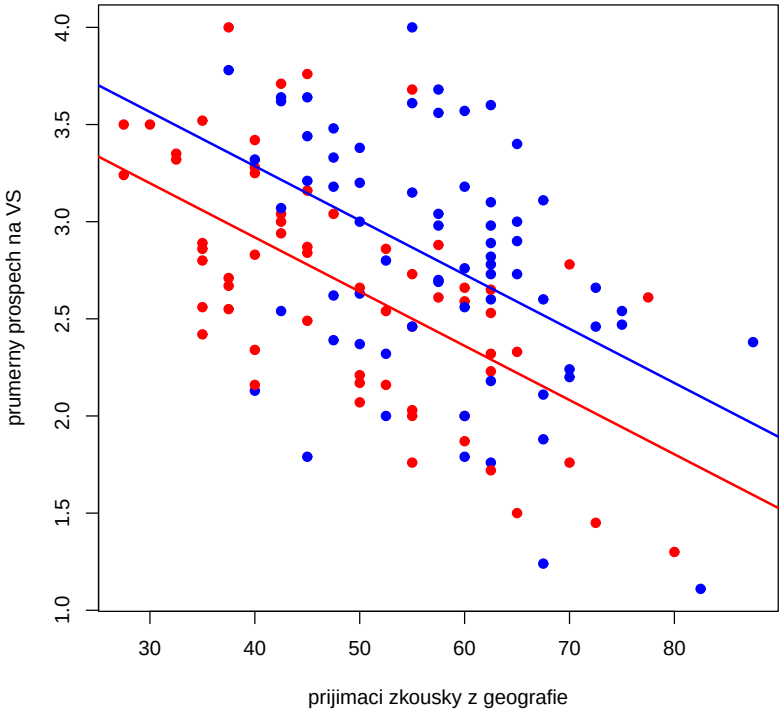
$$b_0 = 1{,}663 \,, \quad b_1 = 0{,}430 \,, \quad b_2 = -0{,}490 \,, \quad b_3 = 0{,}373$$

do (1.38) dostaneme rovnice dvou různoběžných regresních přímek:

$$\begin{aligned} \text{chlapci:} \quad & prumer = \; 1{,}663 + 0{,}430 \cdot ssprum \\ \text{dívky:} \quad & prumer = \; 1{,}663 - 0{,}490 + (0{,}430 + 0{,}373) \cdot ssprum \end{aligned}$$

Mějme jiný model závislosti průměrného prospěchu na vysoké škole, a to na počtu bodů dosažených v přijímacím řízení v testu z geografie. Chceme zjistit, zda má vliv i pohlaví studenta. Neprokážeme vliv interakcí ($p = 0{,}51$), proto můžeme použít jednodušší model bez interakcí, kde je vliv testu z geografie ($p < 0{,}00001$) i pohlaví ($p = 0{,}00025$) statisticky významný. V tomto případě jsou regresní přímky rovnoběžné a prokazatelně různé. Dosazením odhadnutých parametrů dostaneme rovnice:

$$\begin{aligned} \text{chlapci:} \quad & prumer = \; 4{,}769 - 0{,}279 \cdot zem \\ \text{dívky:} \quad & prumer = \; 4{,}769 - 0{,}367 - 0{,}279 \cdot zem \end{aligned}$$



Obrázek 1.9: Grafické znázornění závislosti průměrného prospěchu na VŠ na pohlaví studentů a počtu bodů z přijímací zkoušky z geografie. Modře jsou označeni chlapci, červeně dívky.

Z obr. 1.9 je patrné, že při stejném počtu bodů v testu z geografie, očekáváme u dívek lepší prospěch. ♢

Koeficient parciální korelace

Mějme tři náhodné veličiny Y, Z a X. Připouštíme, že X může působit na Y i Z. Zajímá nás, jaká by byla závislost mezi Y a Z bez vlivu veličiny X.

Koeficient parciální korelace mezi Y a Z při daném X vypočteme

$$\rho_{Y,Z.X} = \frac{\rho_{YZ} - \rho_{XY}\rho_{XZ}}{\sqrt{(1-\rho_{XY}^2)(1-\rho_{XZ}^2)}}.$$

Jedná se vlastně o obyčejný korelační koeficient mezi rezidui dvou modelů (závislosti Y na X a Z na X). Popisuje, jaká závislost mezi Y na Z zůstane, když obě veličiny vysvětlíme co nejlépe pomocí X, neboli vůči X adjustujeme.

Na rozdíl od koeficientu mnohonásobné korelace neplatí žádné nerovnosti mezi obyčejným korelačním koeficientem a koeficientem parciální korelace.

Příklad: Chceme určit hodnotu korelačního koeficientu mezi dosaženým prospěchem na vysoké škole a počtem bodů v přijímací zkoušce z matematiky.

Výběrový korelační koeficient má hodnotu $r_{YZ} = -0{,}510$, po odstranění vlivu středoškolského prospěchu vypočteme parciální korelační koeficient $r_{Y,Z.X} = -0{,}348$. Jeho nižší absolutní hodnota je zřejmě způsobena vysokou korelací počtu bodů z matematiky se středoškolským prospěchem, která je tímto výpočtem eliminována. ♢

Za předpokladu, že vektory $(X_1, Y_1, Z_1)', \ldots (X_n, Y_n, Z_n)'$ jsou výběrem z regulárního normálního rozdělení a platí $\rho_{Y,Z.X} = 0$, má veličina

$$T = \frac{r_{Y,Z.X}}{\sqrt{1 - r^2_{Y,Z.X}}} \sqrt{n-3}$$

rozdělení $\mathtt{t}_{n-3}$ ($r_{Y,Z.X}$ je výběrovým protějškem $\rho_{Y,Z.X}$).

1.2.3 Mnohonásobná lineární regrese – více regresorů

Adjustovaný koeficient determinace

Jak již bylo uvedeno výše, koeficient determinace R^2 ukazuje, jaký díl variability se podařilo vysvětlit závislostí na regresorech modelu. Přidáváme-li do modelu další proměnné, zpravidla se koeficient determinace R^2 zvětší. Pro porovnání modelů s různým počtem regresorů se často používá adjustovaný koeficient determinace, který zohledňuje počet stupňů volnosti pro reziduální složku modelu:

$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-p-1}(1 - R^2)\,, \tag{1.39}$$

kde p je počet regresorů v modelu.

Na rozdíl od R^2 může být R^2_{adj} výjimečně i záporný.

Test podmodelu

Mějme modely

$$Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_j x_{ji} + \beta_{j+1} x_{(j+1)i} + \ldots + \beta_k x_{ki} + e_i\,, \tag{1.40}$$

$$Y_i = \beta'_0 + \beta'_1 x_{1i} + \ldots + \beta'_j x_{ji} + e_i\,. \tag{1.41}$$

Reziduální součet čtverců v modelu (1.40) označíme RSS a v modelu (1.41) RSS_0. Platí-li v lineárním modelu podmodel (zde konkrétně (1.41) je podmodelem (1.40)), pak

$$F_0 = \frac{(RSS_0 - RSS)/(r - r_0)}{RSS/(n-r)} \sim F_{r-r_0,n-r}\,, \tag{1.42}$$

kde r je počet odhadovaných parametrů v modelu a r_0 v podmodelu. V čitateli zlomku (1.42) je, o kolik se v podmodelu zvětší reziduální součet čtverců (dělený rozdílem počtu parametrů), ve jmenovateli je reziduální rozptyl modelu.

Příklad: Mějme model se dvěma regresory (počet bodů z přijímacích zkoušek a středoškolský prospěch), kde vysvětlujeme dosažený prospěch na vysoké škole. Chceme zjistit, zda přidáním informace o pohlaví studenta a době maturity dostaneme významně lepší model.

V testu podmodelu má F-statistika hodnotu $F_0 = 2{,}454$, která je menší než kritická hodnota $F_{2,124}(0{,}05) = 3{,}069$, tedy nezamítáme platnost podmodelu ($p = 0{,}09$). Neprokázali jsme, že bychom zařazením dalších dvou regresorů dostali významně lepší model. Adjustované koeficienty determinace R^2_{adj} modelu 0,6207 a podmodelu 0,6119 jsou podobné. $\diamondsuit$

Testováním podmodelu, který vznikne vyřazením jediného regresoru (a nejedná se o faktor s alespoň třemi úrovněmi), vyjde p-hodnota testu podmodelu stejně jako p-hodnota testu hypotézy nulovosti příslušného regresního koeficientu v modelu, protože platí $F_{1,n-3} = t^2_{n-3}$.

Kroková regrese

Bylo by možné postupně přidávat do modelu jednotlivé veličiny z dostupné množiny možných regresorů a zkoumat, zda dojde ke zlepšení předpovědi. Jedna z možností porovnání různých modelů je pomocí Akaikeho informačního kritéria

$$AIC = -2\ell(\boldsymbol{\theta}) + 2q\,,$$

kde ℓ je logaritmická věrohodnostní funkce a q je počet nezávislých složek $\boldsymbol{\theta}$.

V programu R je k dispozici procedura `step()`, která hledá model s malou hodnotou AIC. Nezaručuje však nejmenší možnou hodnotu AIC, protože neprozkoumá všechny možné množiny regresorů. Algoritmus se v každém kroku pokusí přidat do modelu každou proměnnou, která tam v tuto chvíli není a ubrat každou z těch, které již v modelu jsou. Krok zakončí přidáním (resp. odebráním) takové proměnné, kdy bude hodnota AIC nejmenší. Výpočet algoritmu skončí ve chvíli, kdy žádná jednokroková změna nevede ke zmenšení AIC.

Multikolinearita

Multikolinearitou označujeme situaci, kdy některé regresory, případně skupiny regresorů, jsou mezi sebou skoro závislé. Dokážeme tedy jejich hodnotu z velké části vysvětlit závislostí na dalších regresorech v modelu.

Inflační faktor (Variance Inflation Factor) označovaný VIF_j ukazuje, kolikrát se zhorší rozptyl odhadu b_j v důsledku korelovanosti j-tého regresoru s ostatními regresory. Odmocnina z inflačního faktoru tedy říká, kolikrát je interval spolehlivosti pro zvolené β_j delší.

Příklad: Vysvětlujeme průměrný prospěch na vysoké škole pomocí tří regresorů: přijímací zkouška z geografie (zem), průměrný prospěch na SŠ na konci 3. ročníku ($ss3$) a průměrný prospěch v pololetí 4. ročníku SŠ ($ss4$). Regresní funkce modelu je:

$$prumer = -0{,}018 \cdot zem + 0{,}370 \cdot ss3 + 0{,}382 \cdot ss4 + 2{,}288\,.$$

Regresní koeficienty v tomto modelu jsou statisticky průkazně nenulové ($p < 0{,}00001$, $p = 0{,}0057$, resp. $p = 0{,}0072$), ale oba ukazatele středoškolského prospěchu jsou vzájemně silně korelované ($r = 0{,}87$). Variabilitu $ss3$ (i $ss4$) dokážeme z více než 76 % vysvětlit závislostí na zbývajících dvou regresorech v modelu.

Přidáme-li do modelu se dvěma regresory $(zem, ss3)$ ještě $ss4$, zvýší se sice adjustovaný koeficient determinace R^2_{adj} z 0,566 na 0,587 a dostaneme model, který je statisticky významně lepší, ale zároveň dojde k více než čtyřnásobnému zvýšení rozptylu odhadovaných koeficientů u středoškolské prospěchu ($VIF_1 = 1{,}04; VIF_2 = 4{,}28; VIF_3 = 4{,}20$). Tedy intervaly spolehlivosti pro β_2 a β_3 jsou více než dvojnásobné (funkce `vif()` z knihovny `car` programu R). Kdybychom přidali do modelu ještě další středoškolský regresor $ss2$, vzrostla by hodnota VIF u $ss3$ dokonce na 8,02.

Jako nejvýhodnější se jeví použití jediného středoškolského regresoru, a to vypočteného aritmetického průměru (ze čtyř hodnot uvedených v datovém souboru), kdy dostaneme i nejvyšší $R^2_{\text{adj}} = 0{,}5875$. $\diamondsuit$

1.3 Logistická regrese

Logistickou regresi využíváme v případě, kdy kritérium (úspěšnosti) může nabývat pouze dvou hodnot, zpravidla se volí 1 pro úspěch a 0 pro neúspěch.

Máme nezávislé náhodné veličiny $Y_1, \ldots, Y_n$ s alternativními rozděleními. Jejich střední hodnoty μ_i odpovídají pravděpodobnostem úspěchu (tedy jedničky) a mohou záviset na nějakých nenáhodných veličinách, pro jednoduchost uvažujeme pouze jedinou veličinu x. Pro náhodnou veličinu s alternativním rozdělením je rozptyl roven $\mathsf{var}\, Y_i = \mu_i(1 - \mu_i)$, na rozdíl od normálního lineárního modelu tedy závisí na střední hodnotě této veličiny.

Podíl $\frac{\mu_i}{1-\mu_i}$, který porovnává pravděpodobnost úspěchu a neúspěchu, se nazývá šance (angl. odds) a

$$\eta(\mu) = \ln \frac{\mu}{1-\mu} \tag{1.43}$$

je funkcí označovanou jako logit. Vyjádřením μ ze vztahu (1.43) dostaneme

$$\mu = \frac{\mathrm{e}^{\eta(\mu)}}{1 + \mathrm{e}^{\eta(\mu)}}.$$

Předpokládáme-li dále, že logit pravděpodobnosti je lineární funkcí neznámých parametrů $\boldsymbol{\beta} = (\beta_0, \beta_1)'$

$$\eta_i = \boldsymbol{\beta}' x_i \,,$$

střední hodnota $\mathsf{E}\, Y_i$, resp. pravděpodobnost úspěchu pak je

$$\begin{aligned} \mu_i(\boldsymbol{\beta}) &= \frac{\mathrm{e}^{\boldsymbol{\beta}' x}}{1 + \mathrm{e}^{\boldsymbol{\beta}' x}} \\ &= \frac{1}{\mathrm{e}^{-(\boldsymbol{\beta}' x_i)}(1 + \mathrm{e}^{\boldsymbol{\beta}' x_i})} \\ &= \frac{1}{1 + \mathrm{e}^{-(\boldsymbol{\beta}' x_i)}} \\ &= \frac{1}{1 + \mathrm{e}^{-(\beta_0 + \beta_1 x_i)}} \end{aligned} \tag{1.44}$$

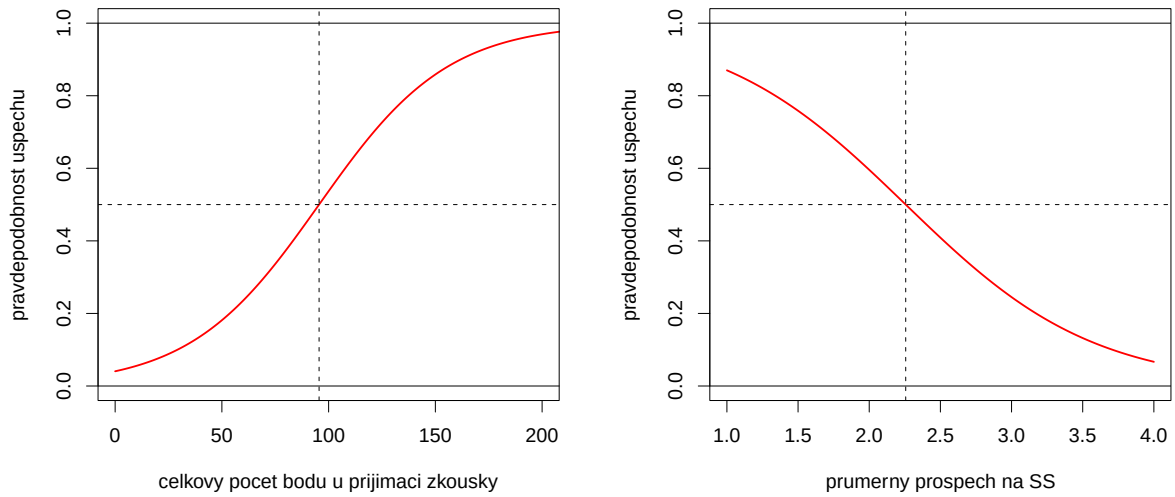
Metodou maximální věrohodnosti iteračními výpočty určíme odhady parametrů β_0 a β_1 z rovnice (1.44), kde parametr β_1 je roven logaritmu poměru šancí úspěchu při $x + 1$ a x. Jestliže pravděpodobnost sledovaného jevu nezávisí na hodnotě x, je poměr

šancí (odds ratio) roven jedné a $\beta_1 = 0$, což odpovídá i významu tohoto parametru v lineární regresi.

Waldův test hypotézy $\mathsf{H}_0 : \beta_1 = 0$ je založen na statistice

$$Z = \frac{b_1}{\sqrt{v_{11}}}, \tag{1.45}$$

která má za platnosti nulové hypotézy asymptoticky rozdělení $\mathsf{N}(0, 1)$ a kde v_{11} je odhad asymptotického rozptylu b_1.



Obrázek 1.10: Graf logistické funkce závislosti úspěšnosti na počtu bodů u přijímacích zkoušek (vlevo), na prospěchu na střední škole (vpravo). Čárkovaně jsou vyznačeny hodnoty, pro které je pravděpodobnost úspěchu 50 %.

Podobnou roli jakou hraje koeficient determinace v lineární regresi má v případě logistické regrese tzv. deviance. Nejbohatší možný model (kde předpokládáme pro každého jedince individuální střední hodnotu) se nazývá saturovaný a hodnota věrohodnostní funkce je maximální (označíme ji $\ell_{\max}$). Každý jiný model je podmodelem saturovaného modelu a jeho přiléhavost můžeme posoudit pomocí deviance

$$D(\mathbf{b}) = 2(\ell_{\max} - \ell(\mathbf{b})). \tag{1.46}$$

Stejně jako v případě reziduálního součtu čtverců platí, že čím je model méně přiléhavý, tím je hodnota deviance větší.

Jiný způsob testování hypotézy $\mathsf{H}_0 : \beta_1 = 0$ je test poměrem věrohodností pomocí deviance modelu a podmodelu (v tomto případě u podmodelu očekáváme pro všechna x_i stejné hodnoty Y_i).

$$\begin{aligned} LR &= 2\left(\ell(b_0, b_1) - \ell(b_0)\right) \\ &= 2\left(\ell_{\max} - \ell(b_0)\right) - 2\left(\ell_{\max} - \ell(b_0, b_1)\right) \\ &= D(b_0) - D(b_0, b_1) \end{aligned}$$

Testová statistika LR má za platnosti podmodelu asymptoticky rozdělení χ^2_1.

Příklad: Ze sledované skupiny studentů 89 úspěšně ukončilo studium a 51 studentů z nejrůznějších důvodů studium neukončilo. Chceme odhadovat úspěšnost studia na základě známého počtu bodů u přijímacích zkoušek. Graf příslušné logistické funkce je na obr. 1.10 vlevo.

Odhad parametru β_1 je $b_1 = 0{,}033$. Testováním hypotézy $\mathsf{H}_0 : \beta_1 = 0$ pomocí Waldova testu dostaneme p-hodnotu 0,000133, porovnáním deviancí modelu a podmodelu vychází podobná p-hodnota 0,000029. Závislost je tedy prokázána. S každým bodem u přijímacích zkoušek roste o 3 % šance, že student bude ve studiu úspěšný. ♢

1.3.1 Logistická regrese – více regresorů

Odhadovat pravděpodobost úspěchu (jedničky) můžeme i na základě znalosti více faktorů. Výpočet odhadů jednotlivých koeficientů je v principu stejný jako v případě jediného regresoru.

Podobně jako v případě jediného regresoru můžeme testovat hypotézu $\mathsf{H}_0 : \beta_i = 0$ dvěma způsoby. Jednak můžeme použít testovou statistiku Z asymptotického (Waldova) testu (1.45), jednak test poměrem věrohodností pomocí deviance modelu a podmodelu (který vznikl vyloučením regresoru u odpovídajícího β_i).

$$LR = 2\left(\ell(\mathbf{b}) - \ell(\mathbf{b}')\right) = 2\left(\ell_{\max} - \ell(\mathbf{b}')\right) - 2\left(\ell_{\max} - \ell(\mathbf{b})\right) = D(\mathbf{b}') - D(\mathbf{b}), \quad (1.47)$$

kde $\mathbf{b}$ jsou odhady parametrů $\boldsymbol{\beta}$ v modelu a $\mathbf{b}'$ v podmodelu.

Testová statistika LR má za platnosti podmodelu asymptoticky rozdělení χ^2_f, kde f je rovno rozdílu počtu nezávislých parametrů v porovnávaných modelech.

Příklad: Chceme zjistit, zda přidáním informace o středoškolském prospěchu k předchozímu modelu (závislosti úspěšnosti studia na počtu bodů u přijímacích zkoušek) tuto předpověď zlepšíme. Waldovým testem hypotézy $\mathsf{H}_0 : \beta_2 = 0$ se dostaneme k p-hodnotě 1,1 %. Použitím testu podmodelu (1.47), který porovnává deviance, vychází obdobná (ale nikoliv stejná) p-hodnota 0,9 %. ♢

Vzhledem k podobě deviance a residuálního součtu čtverců byla snaha pojem koeficientu determinace rozšířit i na logistickou regresi. Zřejmá analogie s lineární regresí, kde

$$R^2 = 1 - \frac{RSS}{\sum (Y_i - \bar{Y})^2}, \quad (1.48)$$

vede k McFaddenovu koeficientu determinace [21]

$$R^2_L = 1 - \frac{D(\mathbf{b})}{D_0},$$

kde $D(\mathbf{b})$ je deviance modelu a D_0 deviance tzv. nulového modelu, kde jsou všechny střední hodnoty $\mu_i = \mathsf{E}\, Y_i$ shodné, tedy v modelu zůstane jen absolutní člen.

Jiná definice koeficientu determinace využívá vyjádření hodnoty logaritmické věrohodnostní funkce normálního lineárního modelu, kdy po dosazení do (1.48) dostaneme

$$R^2 = 1 - \exp\left(\frac{1}{n}\left(D(\mathbf{b}) - D_0\right)\right). \quad (1.49)$$

Deviance saturovaného modelu je rovna nule, takže koeficient z (1.49) nemůže překročit hodnotu

$$R^2_{\max} = 1 - \exp\left(-\frac{1}{n} D_0\right). \tag{1.50}$$

Nagelkerke [22] navrhl upravit definici zobecněného koeficientu determinace z (1.49) na

$$R^2_N = \frac{R^2}{R^2_{\max}} = \frac{1 - \exp\left(\left(D(\mathbf{b}) - D_0\right)/n\right)}{1 - \exp\left(-D_0/n\right)}. \tag{1.51}$$

Příklad: Pro dříve uvedený model závislosti úspěšnosti studia na výsledcích u přijímacích zkoušek vycházejí následující koeficienty determinace:

$$R^2_L = 0{,}095, \quad R^2 = 0{,}117, \quad R^2_{\max} = 0{,}731, \quad R^2_N = 0{,}161.$$

Při zařazení průměrného středoškolského prospěchu mezi regresory dostaneme model, který je statisticky významně lepší, a tomu odpovídají i vyšší hodnoty koeficientů determinace. $R^2_{\max}$ je samozřejmě v obou modelech stejné, protože pro oba modely jsou použita stejná data.

$$R^2_L = 0{,}132, \quad R^2 = 0{,}160, \quad R^2_{\max} = 0{,}731, \quad R^2_N = 0{,}218. \diamondsuit$$

1.3.2 ROC analýza

ROC křivka (Reciever Operating Characteristic) byla poprvé použita během 2. světové války pro analýzu radarových signálů. Po útoku na Pearl Harbor zahájila americká armáda nový výzkum pro zlepšení predikování správné detekce japonského letectva.

V současnosti je ROC analýza využívána v mnoha oborech v lékařství (hodnocení diagnostických testů, radiologie, epidemiologie, apod.), psychologii i sociálních vědách.

Předpokládejme, že máme vstupní veličinu T (kvantitativní) a podle její intenzity máme rozhodnout, zda jsme přijali signál nebo ne. Při překročení zvolené mezní hodnoty t_0 rozhodneme, že signál byl přijat, v opačném případě je naše rozhodnutí, že o signál nešlo. Mohou tedy nastat čtyři různé situace (signál byl nebo nebyl vyslán, byl nebo nebyl přijat), které můžeme přehledně zapsat do tabulky (viz [11]).

predikce	skutečnost	
	signál nevyslán	signál vyslán
signál nepřijat	True Negative	False Negative
signál přijat	False Positive	True Positive

Velmi podobné je statistické rozhodování, kdy chceme rozhodnout, zda platí nulová hypotéza (H_0) nebo hypotéza alternativní (H_1). Jestliže se data dostatečně liší od situace, jakou bychom očekávali za platnosti hypotézy H_0, nulovou hypotézu zamítneme (přikloníme se k hypotéze alternativní). Nulová hypotéza odpovídá neexistenci signálu, její zamítnutí znamená, že jsme signál odhalili.

rozhodnutí	skutečnost	
	platí H_0	platí H_1
nezamítnout H_0	správné rozhodnutí	chyba II. druhu
zamítnout H_0	chyba I. druhu	správné rozhodnutí

Tabulka 1.1: Matice záměn

Predikce	Skutečnost	
	0 (neúspěch)	1 (úspěch)
0 (neúspěch)	TN True Negative	FN False Negative
1 (úspěch)	FP False Positive	TP True Positive

Symbolem TPR (True Pozitive Rate) se značí podmíněná pravděpodobnost přijetí signálu za podmínky, že signál byl opravdu vyslán. V kontextu statistického rozhodování se jedná o sílu testu (doplněk pravděpodobnosti chyby II. druhu do jedničky). Používá se též označení *senzitivita* testu (viz [38]).

Podmíněná pravděpodobnost přijetí signálu za podmínky, že ve skutečnosti signál vyslán nebyl, se označuje FPR (False Positive Rate, Alarm Rate), což ve statistickém rozhodování odpovídá pravděpodobnosti chyby I. druhu. Doplněk FPR do jedničky tzv. *specificita* ukazuje, s jakou pravděpodobností nepřijmeme signál v případě jeho nevyslání.

Existuje podstatný rozdíl mezi modelem statistického rozhodování a modelem přijímání signálu. U statistického rozhodování nulová hypotéza buď platí nebo neplatí, pouze my neznáme skutečnost, takže nemá smysl hovořit o pravděpodobnosti nulové nebo alternativní hypotézy. Pravděpodobnosti chyb obou druhů jsou pravděpodobnostmi počítanými za konkrétních podmínek a nejedná se o podmíněné pravděpodobnosti. V případě přijímání signálu je i vyslání signálu chápáno jako náhodný jev, a proto TPR a FPR jsou pravděpodobnosti podmíněné a mění se v závislosti na volbě mezní hodnoty t_0.

Když použijeme ROC křivku k vyjádření schopnosti logistické regrese předpovídat hodnotu závisle proměnné, za přítomnost signálu označíme skutečnost, že přijatý uchazeč úspěšně vystuduje. Rozhodovací statistikou T bude odhadnutá hodnota logistické funkce pro dané hodnoty nezávisle proměnných.

Podle konkrétního stanovení tzv. *prahové hodnoty* t_0 dostaneme různé klasifikace, neboli roztřídění studentů do skupin úspěšných a neúspěšných podle pravidla

$$T_i \leq t_0 \Rightarrow \widehat{G} = 0 \quad \wedge \quad T_i > t_0 \Rightarrow \widehat{G} = 1, \tag{1.52}$$

kde G je binární indikátor úspěšnosti, kdy 1 znamená úspěšné ukončení studia.

Senzitivita tedy udává schopnost klasifikátoru rozpoznat úspěšnost ve skupině úspěšných $\mathsf{Se} = \mathsf{P}(\widehat{G} = 1 | \, G = 1)$, specificita odpovídá pravděpodobnosti správně klasifikovaného objektu ve skupině neúspěšných $\mathsf{Sp} = \mathsf{P}(\widehat{G} = 0 | \, G = 0)$.

Výsledky klasifikace zapíšeme do tzv. *matice záměn* (confusion matrix) viz tab. 1.1 a empirické odhady specificity $\mathsf{Se}(t_0)$ a senzitivity $\mathsf{Sp}(t_0)$ se pak dají vyjádřit

$$\begin{aligned} \widehat{\mathsf{Se}}(t_0) &= \frac{\mathsf{TP}}{\mathsf{TP}+\mathsf{FN}} = \mathsf{TPR} = 1 - \mathsf{FNR} = 1 - \frac{\mathsf{FN}}{\mathsf{TP}+\mathsf{FN}} \\ \widehat{\mathsf{Sp}}(t_0) &= \frac{\mathsf{TN}}{\mathsf{TN}+\mathsf{FP}} = 1 - \mathsf{FPR} = 1 - \frac{\mathsf{FP}}{\mathsf{TN}+\mathsf{FP}} \, , \end{aligned}$$

kde FNR označuje doplněk senzitivity do jedničky, tzv. False Negative Rate.

Cílem je najít takovou prahovou hodnotu rozhodovacího pravidla, aby byla co nejlépe odlišena přítomnost signálu od jeho nepřítomnosti (resp. odlišeni úspěšní studenti od neúspěšných), neboli nalézt optimální klasifikaci.

Často užívanou mírou úspěšnosti klasifikace je pravděpodobnost správné klasifikace, která se též nazývá celková správnost (overall accuracy), označovaná Acc. Komplementární mírou je celková chyba Err, pravděpodobnost chybné klasifikace.

$$\begin{aligned} \mathsf{Acc}(t_0) &= \mathsf{P}(T \leq t_0 | \, G = 0) \cdot \pi_0 + \mathsf{P}(T > t_0 | \, G = 1) \cdot \pi_1 \\ &= \pi_0 \cdot \mathsf{Sp}(t_0) + \pi_1 \cdot \mathsf{Se}(t_0), \end{aligned}$$

kde π_0 je apriorní pravděpodobnost neúspěchu a π_1 úspěchu.

Empirický odhad Acc se spočítá z matice záměn jako podíl správně zařazených objektů

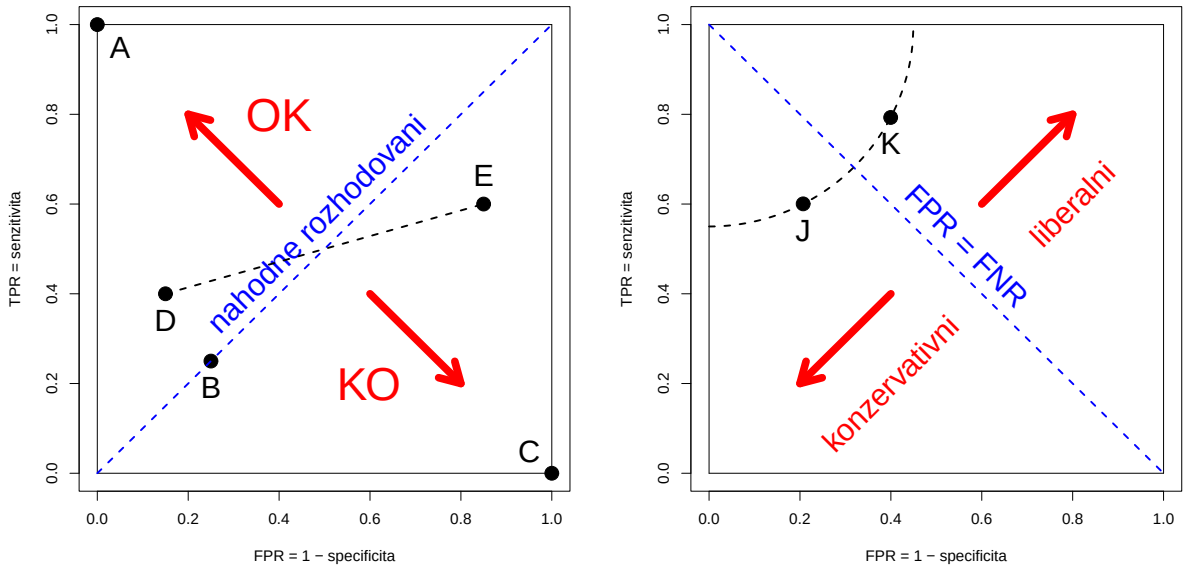
$$\widehat{\mathsf{Acc}}(t_0) = \frac{\mathsf{TP} + \mathsf{TN}}{n},$$

kde n je celkový počet pozorování.

V případě výrazně nerovnoměrného zastoupení skupiny úspěšných a neúspěšných v datech je použití Acc jako míry kvality klasifikátoru nevhodné. V případě malého podílu úspěšných bude klasifikátor, který zařadí všechny objekty mezi neúspěšné, vykazovat vysoké Acc, i když nebude schopen žádného rozlišení.

ROC křivka

ROC křivka znázorňuje závislost pravděpodobnosti TPR (senzitivity) na pravděpodobnosti FPR (doplňku specificity do jedničky). Pro konkrétní hodnotu t_0 se na vodorovné ose vynáší $\mathsf{FPR}(t_0) = 1 - \widehat{\mathsf{Sp}}(t_0)$ a na svislé ose $\mathsf{TPR}(t_0) = \widehat{\mathsf{Se}}(t_0)$.



Obrázek 1.11: ROC prostor

Bod $A[0,1]$ na obr. 1.11 odpovídá nejlepší možné klasifikaci, kde je senzitivita i specificita rovna jedné a všechna pozorování jsou správně zařazena. Na rostoucí diagonále se zobrazují body, kdy klasifikátor rozhoduje náhodně, součet senzitivity a specificity je roven 1. Například pro bod $B[0{,}25, 0{,}25]$ je $\mathsf{Se} = 0{,}25$ a $\mathsf{Sp} = 0{,}75$. Body ležící nad diagonálou odpovídají lepší klasifikaci než je náhodné rozhodování, pro body pod diagonálou poskytuje klasifikace horší než náhodné rozhodování. Invertováním rozhodnutí můžeme v tomto případě vždy vytvořit klasifikaci lepší než náhodnou. Obraz inverzní klasifikace bude středově souměrný podle $[0{,}5, 0{,}5]$, jak je znázorněno umístěním bodů D a E.

Na obr. 1.11 vpravo jsou znázorněny body J a K, které jsou stejně daleko od ideálního rozhodnutí a liší se různým hodnocením chybných rozhodnutí. Na klesající diagonále leží obrazy klasifikací, kde jsou oba typy klasifikačních chyb stejně pravděpodobné ($\mathsf{FPR} = \mathsf{FNR}$) a zároveň $\mathsf{Sp} = \mathsf{Se}$. Pod klesající diagonálou leží body (například J), kde $\mathsf{FPR} < \mathsf{FNR}$, tedy kde je menší chyba zařazení objektu z $\mathcal{G}_0$ do $\mathcal{G}_1$ než zařazení objektu z $\mathcal{G}_1$ do $\mathcal{G}_0$, kde $\mathcal{G}_0$, resp. $\mathcal{G}_1$ je v našem pojetí skupina neúspěšných, resp. úspěšných studentů. Tento typ klasifikátoru se nazývá konzervativní a je mu dávána přednost v případě, že tato chyba (zařazení z $\mathcal{G}_0$ do $\mathcal{G}_1$) má závažnější následky. Bod K reprezentuje tzv. liberální klasifikátor, kde platí $\mathsf{FPR} > \mathsf{FNR}$.

Jak již bylo řečeno výše, není vždy celková přesnost (Acc) tím nejvhodějším kritériem pro hodnocení klasifikace. Jinou možností je požadavek, aby součet senzitivity a specificity byl nejvyšší možný. Tomu odpovídá i minimální součet pravděpodobností chyb $\mathsf{FNR} + \mathsf{FPR}$. Přímky rovnoběžné s rostoucí diagonálou obsahují body, které jsou obrazem klasifikací, kde je konstantní součet specificity a senzitivity, stejně jako součet $\mathsf{FNR} + \mathsf{FPR}$. Čím leží příslušná rovnoběžka výše, tím je součet specificity a senzitivity vyšší.

Máme-li zobrazeny v ROC prostoru všechny klasifikace pro různá t_0, hledáme bod na ROC křivce, který je leží na rovnoběžce s rostoucí diagonálou a zároveň je od diagonály nejvíce vzdálen, protože pro odpovídající t_0 je součet pravěpodobností chyb $\mathsf{FNR} + \mathsf{FPR}$ minimální.

Příklad: Na obrázku 1.12 je zobrazena ROC křivka odpovídající dříve uvedenému modelu logistické regrese závislosti úspěšnosti ve studiu na počtu bodů u přijímacích zkoušek. Modře je vyznačena hodnota t_0 pro maximální součet $\widehat{\mathsf{Se}}+\widehat{\mathsf{Sp}}$, červeně hodnota odpovídající nejvyššímu $\widehat{\mathsf{Acc}}$. V tabulce 1.2 je matice záměn, která odpovídá modelu z obr. 1.12 se zvolenou hraniční mezí $t_0 = 0{,}663$ (modře vyznačenému bodu), kde je dosažen nejnižší součet chyb FNR a FPR.

Nejvyšší $\widehat{\mathsf{Acc}}$ dostaneme pro $t_0 = 0{,}487$, ale z tabulky 1.3 je patrna mnohem nižší hodnota specificity, což znamená výrazně horší rozlišení ve skutečnosti neúspěšných studentů, kdy byla v tomto konkrétním případě více než polovina nesprávně klasifikována jako úspěšní. $\diamondsuit$

Podobně jako je na rovnoběžkách s diagonálou konstatní součet $\mathsf{Se} + \mathsf{Sp}$, na rovnoběžkách, jejichž směrnice je rovna podílu počtu neúspěšných ku úspěšným, je vždy stejná hodnota Acc ([4]). Graficky bychom mohli najít bod maxima Acc jako průsečík ROC křivky s nejvýše položenou tzv. *iso-Acc* přímkou.

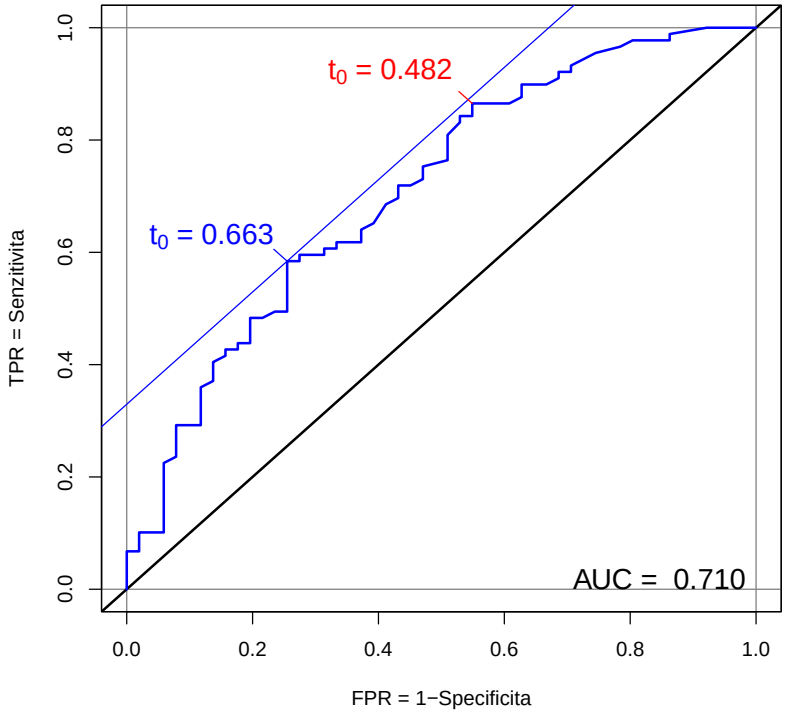
Na ROC křivku je možno pohlížet i obecněji. Nechť $G(t)$ je distribuční funkce rozhodovací statistiky tam, kde není signál ($Y = 0$, neúspěšné studium), $F(t)$ distribuční funkce funkce rozhodovací statistiky za přítomnosti signálu ($Y = 1$, úspěšné studium).

Tabulka 1.2: Matice záměn pro závislost úspěšnosti studia na počtu bodů u přijímacích zkoušek pro $t_0 = 0{,}663$

$t_0 = 0{,}663$			
	skutečnost		
predikce	0 (neúspěch)	1 (úspěch)	celkem
0 (neúspěch)	38	37	75
1 (úspěch)	13	52	65
celkem	51	89	140
$\mathsf{FPR} + \mathsf{FNR} = 0{,}255 + 0{,}416 = 0{,}671$			
$\widehat{\mathsf{Sp}} + \widehat{\mathsf{Se}} = 0{,}745 + 0{,}584 = 1{,}329$			
$\widehat{\mathsf{Acc}} = 90/140 = 0{,}643$			

Tabulka 1.3: Matice záměn pro závislost úspěšnosti studia na počtu bodů u přijímacích zkoušek pro $t_0 = 0{,}487$

$t_0 = 0{,}487$			
	skutečnost		
predikce	0 (neúspěch)	1 (úspěch)	celkem
0 (neúspěch)	23	12	35
1 (úspěch)	28	77	105
celkem	51	89	140
$\mathsf{FPR} + \mathsf{FNR} = 0{,}549 + 0{,}135 = 0{,}684$			
$\widehat{\mathsf{Sp}} + \widehat{\mathsf{Se}} = 0{,}451 + 0{,}865 = 1{,}316$			
$\widehat{\mathsf{Acc}} = 100/140 = 0{,}714$			



Obrázek 1.12: ROC křivka závislosti úspěšnosti studia na počtu bodů u přijímacích zkoušek

Dále předpokládáme, že hustota $g(t)$ odpovídající distribuční funkci $G(t)$ je stejně jako $f(t)$ odpovídající $F(t)$ na celé reálné přímce kladná. Pro dané t_0 v případě nepřítomnosti signálu signálu platí $\mathsf{P}(T > t_0) = 1 - G(t_0)$, v případě jeho existence pak $\mathsf{P}(T > t_0) = 1 - F(t_0)$.

Označíme-li x-ovou souřadnici bodu na ROC křivce jako

$$x = 1 - G(t_0)$$

pak lze odpovídající t_0 vyjádřit pomocí kvantilové funkce G^{-1} (inverzní k distribuční funkci G)

$$t_0 = G^{-1}(1 - x)\,.$$

A y-ová souřadnice téhož bodu bude rovna

$$1 - F(t_0) = 1 - F(G^{-1}(1 - x))\,.$$

ROC křivka je tedy v případě spojitého rozdělení grafem funkce

$$R(x) = 1 - F(G^{-1}(1 - x)), \quad 0 < x < 1. \tag{1.53}$$

AUC

Méně podrobnou informaci než samotná ROC křivka vyjadřuje statistika AUC (Area Under Curve), neboli plocha pod ROC křivkou. Používá se zejména při porovnání dvou nebo více křivek.

Vzhledem k zavedení ROC funkce v (1.53) platí

$$AUC = \int_0^1 R(x)\,\mathrm{d}x \tag{1.54}$$

Označíme jako T_1 rozhodovací statistiku za existence signálu (úspěšné studium), jako T_0 rozhodovací statistiku v případě neexistence signálu (neúspěšné studium). Předpokládáme, že veličina T má spojité rozdělení na celém oboru reálných čísel. Hustotu T_0 budeme značit $g(t)$. Dosazením do (1.54) za $R(x)$ ze vztahu (1.53) postupně dostaneme

$$\begin{aligned} AUC &= \int_0^1 (1 - F(G^{-1}(1-x)))\,\mathrm{d}x \\ &= \int_0^1 \mathsf{P}(T_1 > G^{-1}(1-x))\,\mathrm{d}x. \end{aligned}$$

Po provedení substituce $y = G^{-1}(1-x)$, což je za našich předpokladů prostá a klesající funkce, můžeme poslední integrál postupně upravit na

$$AUC = \int_{-\infty}^{\infty} \mathsf{P}(T_1 > y) g(y)\,\mathrm{d}y \tag{1.55}$$

Chceme dokázat, že $AUC = \mathsf{P}(T_1 > T_0)$. Nyní budeme upravovat pravou stranu dokazované rovnosti. Označíme $h(x, y)$ sdruženou hustotu (T_1, T_0):

$$\begin{aligned} \mathsf{P}(T_1 > T_0) &= \int_{-\infty}^{\infty} \left(\int_y^{\infty} h(x,y)\mathrm{d}x \right) \mathrm{d}y \\ &= \int_{-\infty}^{\infty} \left(\int_y^{\infty} h(x|y)g(y)\mathrm{d}x \right) \mathrm{d}y \\ &= \int_{-\infty}^{\infty} \left(\int_y^{\infty} h(x|y)\mathrm{d}x \right) g(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} \mathsf{P}(T_1 > y) g(y)\,\mathrm{d}y. \end{aligned} \tag{1.56}$$

Ze vztahů (1.55) a (1.56) je zřejmé, že (populační) AUC je rovna pravděpodobnosti, že rozhodovací statistika T je v případě existence signálu (úspěšné studium) větší než stejná statistika bez přítomnosti signálu (neúspěšné studium) - viz [27], [26].

Další interpretace se týká výběrových vlastností AUC. Označíme $Y_{11}, \ldots, Y_{1n_1}$ rozhodovací statistiky za přítomnosti signálu (resp. odhadované pravděpodobnosti úspěšného dokončení studia u úspěšných studentů) a $Y_{01}, \ldots, Y_{0n_0}$ tytéž statistiky bez přítomnosti (odhadované pravděpodobnosti úspěšného dokončení studia u neúspěšných studentů). Máme celkem $n = n_1 + n_0$ hodnot, z nichž můžeme vytvořit celkem $n_1 \cdot n_0$ dvojic, kde každý člen z dvojice patří do jiné skupiny. Pro odhad pravděpodobnosti $\mathsf{P}(Y_1 > Y_0)$ použijeme relativní četnost

$$u = \frac{U}{n_1 \cdot n_0}\,, \tag{1.57}$$

kde

$$U = \#(T_1 > T_0) + \#(T_1 = T_0)/2\,, \tag{1.58}$$

neboli počet dvojic, kdy je rozhodovací statistika T_1 za přítomnosti signálu větší než statistika T_0 bez signálu, zvětšený o polovinu případů, kdy jsou hodnoty obou statistik ve dvojici stejné. Jinými slovy se jedná o počet případů, kdy je pravděpodobnost úspěchu u skutečně úspěšných studentů větší než u neúspěšných zvětšené a polovinu shodných dvojic.

Testová statistika z (1.58) je testová statistika U, kterou navrhli Mann a Whitney v roce 1947 pro testování hypotézy, že distribuční funkce spojitých rozdělení T_1 a T_0 jsou totožné ([20]). V závislosti na zvoleném pořadí výběrů (často si programy určují samy) můžeme místo statistiky U dostat $n_1 \cdot n_0 - U$.

Příklad: Na obrázku 1.12 je zobrazena ROC křivka závislosti úspěšnosti studia na počtu bodů dosažených u přijímacích zkoušek a vyznačena hodnota $AUC = 0{,}710$. Studium úspěšně ukončilo 89 studentů ($n_1 = 89$), neukončilo 51 studentů ($n_0 = 51$). Použitím dvouvýběrového Mannova-Whitneyova testu, který porovnává predikované pravděpodobnosti u dvou skupin studentů (úspěšných a neúspěšných), dostaneme hodnotu statistiky $U = 1315$. S ohledem na pořadí zvolené programem musíme vypočítat doplněk této hodnoty do $n_1 \cdot n_0$. Dosazením do vzorce (1.57) pak vychází

$$u = \frac{n_1 \cdot n_0 - U}{n_1 \cdot n_0} = \frac{89 \cdot 51 - 1315}{89 \cdot 51} = \frac{3224}{4539} = 0{,}710\,,$$

což odpovídá výše uvedené hodnotě AUC. $\diamondsuit$

1.4 Ordinální regrese

Předpokládáme, že veličina Y označující úspěšnost dokáže rozlišovat podrobněji a zařazovat subjekty do více disjunktních kategorií, které jsou vzájemně uspořádané. Příkladem může být rozdělení studentů do tří kategorií, kdy za nejúspěšnější považujeme ty, kteří absolvovali studium ve standardní době studia, ve druhé skupině budou studenti, kteří studium sice úspěšně absolvovali, ale později než ve standardní době a zbytek tvoří ti, kteří studia zanechali nebo doposud studium neukončili.

1.4.1 Model s latentní proměnnou

Zařazení studenta do jedné z J kategorií může být modelováno pomocí latentní proměnné Y^* a dělících bodů $t_0, t_1, \ldots, t_J$, pro které platí $-\infty = t_0 < t_1 \leq \ldots, \leq t_J < t_J = \infty$.

Pro latentní proměnnou Y^* platí lineární model $Y^* = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, kde $\mathbf{X} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')'$ a $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$ je vektor neznámých parametrů s absolutním členem β_0, n označuje počet pozorování. O chybových členech $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ předpokládáme, že jsou nezávislé náhodné veličiny, mají nulovou střední hodnotu, stejné rozptyly a jsou z rozdělení s distribuční funkcí F.

Klasifikaci jednotlivých hodnot Y_i závisle proměnné provedeme následujícím způsobem:

$$\begin{aligned} Y_i &= 1, \quad \text{pokud} \quad Y_i^* < t_1 \\ Y_i &= m, \quad \text{pokud} \quad t_{m-1} \leq Y_i^* < t_m, \quad m = 2, \ldots, J. \end{aligned} \tag{1.59}$$

Ze vztahu (1.59) je možno dále odvodit

$$\mathsf{P}(Y_i \leq m|\,\mathbf{x}_i) = \mathsf{P}(Y_i^* < t_m|\,\mathbf{x}_i) = \mathsf{P}(\varepsilon_i < t_m - \boldsymbol{\beta}'\mathbf{x}_i|\,\mathbf{x}_i) = F(t_m - \boldsymbol{\beta}'\mathbf{x}_i) \tag{1.60}$$

a

$$\mathsf{P}(Y_i = m|\,\mathbf{x}_i) = F(t_m - \boldsymbol{\beta}'\mathbf{x}_i) - F(\theta_{m-1} - \boldsymbol{\beta}'\mathbf{x}_i).$$

V modelu s absolutním členem β_0 a tedy i stejné první složce vektoru $\mathbf{x}_i$ rovné jedničce není řešení jednoznačné. Je nutno stanovit nějakou identifikační podmínku. Program R využívá ve funkci `polr` z knihovny `MASS` identifikační podmínku $\beta_0 = 0$.

Nejčastěji je za funkci F volena distribuční funkce logistického nebo normovaného normálního rozdělení.

Podíl šancí, že veličina Y nabude hodnoty menší nebo rovné m a hodnoty větší než m je při daném $\mathbf{x}$

$$\omega_m(\mathbf{x}) = \frac{\mathsf{P}(Y \leq m|\,\mathbf{x})}{\mathsf{P}(Y > m|\,\mathbf{x})} = \frac{F(t_m - \boldsymbol{\beta}'\mathbf{x})}{1 - F(t_m - \boldsymbol{\beta}'\mathbf{x})}. \tag{1.61}$$

Dosazením distribuční funkce logistického rozdělení (1.44) za funkci F dostaneme

$$\omega_m(\mathbf{x}) = \exp(t_m - \boldsymbol{\beta}'\mathbf{x}),$$

kde je po zlogarimování

$$\ln \omega_m(\mathbf{x}) = t_m - \boldsymbol{\beta}'\mathbf{x}$$

vidět lineární vztah logaritmu šance a $\mathbf{x}$.

Pro dvě různé hodnoty $\mathbf{x}_1$ a $\mathbf{x}_2$ je podíl šancí roven

$$\frac{\omega_m(\mathbf{x}_1)}{\omega_m(\mathbf{x}_2)} = \frac{\exp(t_m - \boldsymbol{\beta}'\mathbf{x}_1)}{\exp(t_m - \boldsymbol{\beta}'\mathbf{x}_2)} = \exp\left(\boldsymbol{\beta}'(\mathbf{x}_2 - \mathbf{x}_1)\right), \tag{1.62}$$

který nezávisí na m. Změna v některé z proměnných x_k má tedy stejný vliv na šanci, že znak nabude hodnoty větší nebo rovné m, tak na šanci, že nabude hodnoty větší nebo rovné l ($l \neq m$). Vždy je tedy třeba zjistit, zda je tento předpoklad rovnoběžnosti regresních křivek oprávněný. Test rovnoběžnosti regresních křivek, který navrhl Brant, je podrobně popsán v [17], uvedenou funkci jsem použila pro své výpočty v programu R.

Jestliže také do (1.60) dosadíme za funkci F distribuční funkci logistického rozdělení, dostaneme předpis pro výpočet odpovídajících pravděpodobností

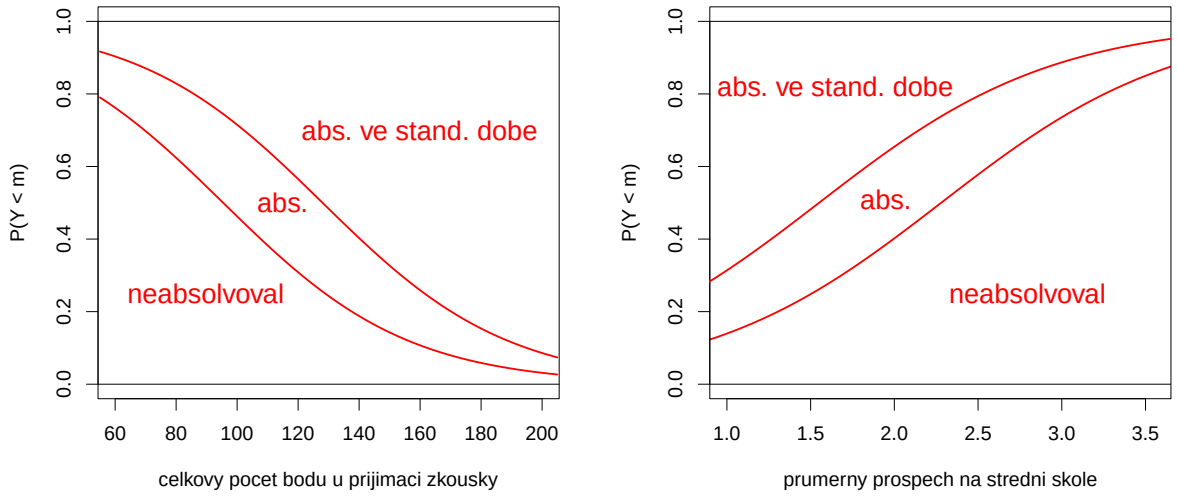
$$\mathsf{P}(Y_i \leq m|\,\mathbf{x}_i) = F(t_m - \boldsymbol{\beta}'\mathbf{x}_i) = \frac{1}{1 + \exp(\boldsymbol{\beta}'\mathbf{x}_i - t_m)}\,. \tag{1.63}$$

Jedná se rovnoběžné logistické funkce, které se liší jen absolutním členem.

V případě jediného regresoru se (1.63) vzhledem k identifikační podmínce ($\beta_0 = 0$) výpočet odhadu koeficientů $\boldsymbol{\beta}$ a $t_1, \ldots, t_{J-1}$ zjednoduší na

$$\mathsf{P}(Y_i \leq m|\,x_i) = F(t_m - \beta_1 x_i) = \frac{1}{1 + \exp(\beta_1 x_i - t_m)}.$$

Příklad: Vypočtené pravděpodobnosti (odpovídající regresním křivkám) můžeme znázornit v grafu, kde pro každou hodnotu nezávislé veličiny postupně vynášíme součty těchto pravděpodobností.



Obrázek 1.13: Grafické znázornění regresních křivek závislosti úspěšnosti studia na počtu bodů u přijímacích zkoušek (vlevo) a průměrného prospěchu na střední škole (vpravo)

Na obrázku 1.13 jsou znázorněny regresní křivky dvou logistických ordinálních modelů, kde je možno křivky považovat za paralelní. V testu rovnoběžnosti regresních křivek pro model vlevo dostaneme p-hodnotu 0,89, vpravo $p = 0{,}88$.

Je však patrné, že v obou případech nikdo nebude zařazen do prostřední skupiny mezi studenty, kteří studium sice úspěšně absolvovali, ale nikoliv ve standardní době studia. Pro každou hodnotu na ose x je totiž buď větší pravděpodobnost nedokončení studia nebo naopak úspěšné ukončení ve standardní době. Je to způsobeno malou hodnotou koeficientu b_1, která odpovídá nepříliš těsné závislosti. Aby pro některé hodnoty na ose x převážila pravděpodobnost zařazení do prostřední kategorie, musely by mít regresní křivky větší sklon.

Pro model závislosti na počtu bodů u přijímacích zkoušek (vlevo) budou všichni, kteří dosáhli nejvýše 111,5 bodů klasifikováni jako neúspěšní. Pro ty, kdo získali alespoň 112 bodů, je naopak největší pravděpodobnost absolvování ve standardní době. Modelu závislosti na průměrném prospěchu na střední škole (vpravo) odpovídá mezní hodnota 1,915. ♢

1.5 Hodnocení shody predikce

Cílem je najít postup, jak hodnotit shodu či neshodu v rozhodování (predikci) dvou nesouvisejících modelů založených na stejných datech.

Ve statistické literatuře je možno najít řadu článků, které se snaží porovnat měření dvou expertů při hodnocení v číselném (spojitém) nebo kvalitativním měřítku. Pro kvalitativní data se používá Cohenův kappa koeficient, v případě spojitého měřítka Linův konkordanční korelační koeficient.

1.5.1 Konkordanční korelační koeficient

Předpokládejme, že pomocí spojitých náhodných veličin X a Y měříme stejnou vlastnost. V případě naprosté shody budou dvojice měření ležet na přímce $y = x$. Lin [19] zavedl statistiku

$$\rho_{\mathrm{c}} = 1 - \frac{\mathsf{E}\,(X-Y)^2}{\mathsf{E}_{\,\mathrm{indep}}(X-Y)^2}, \tag{1.64}$$

kde $\mathsf{E}_{\,\mathrm{indep}}(X-Y)^2$ je střední hodnota vypočtená za předpokladu, že X a Y jsou nezávislé náhodné veličiny, což znamená, že je založena na součinu marginálních rozdělení ke sdruženému rozdělení použitému v čitateli.

Konkordanční korelační koeficient ρ_{c} označovaný zkráceně jako CCC (Concordance Corelation Coefficient) porovnává skutečnou střední hodnotu čtverce rozdílu mezi X a Y se situací, kdy by obě náhodné veličiny byly nezávislé. Nulové hodnoty nabývá pro nezávislé veličiny X a Y, jedné se rovná v případě, že jsou obě veličiny s jednotkovou pravděpodobností totožné.

Souvislost s Pearsonovým korelačním koeficientem je patrna z následujících úprav vzorce (1.64)

$$\begin{aligned}
\rho_{\mathrm{c}} &= 1 - \frac{\mathsf{E}\,((X-\mu_X)-(Y-\mu_Y)+(\mu_X-\mu_Y))^2}{\mathsf{E}_{\,\mathrm{indep}}((X-\mu_X)-(Y-\mu_Y)+(\mu_X-\mu_Y))^2} \\
&= 1 - \frac{\sigma_X^2+\sigma_Y^2+2\cdot\sigma_{XY}+(\mu_X-\mu_Y)^2}{\sigma_X^2+\sigma_Y^2+(\mu_X-\mu_Y)^2} \\
&= \frac{2\sigma_{XY}}{\sigma_X^2+\sigma_Y^2+(\mu_X-\mu_Y)^2}\cdot\frac{\sigma_X\sigma_Y}{\sigma_X\sigma_Y} \qquad (1.65) \\
&= \rho_{XY}\frac{2}{v+1/v+u^2}, \qquad (1.66)
\end{aligned}$$

kde jsme zavedli

$$u = \frac{\mu_X-\mu_Y}{\sqrt{\sigma_X\sigma_Y}}, \qquad v = \frac{\sigma_X}{\sigma_Y}.$$

V případě shodných středních hodnot a stejných rozptylů je $\rho_{\mathrm{c}} = \rho_{XY}$, ve všech ostatních případech bude $\rho_c < \rho_{XY}$. Vztah (1.66) je možno interpretovat tak, že CCC je penalizovaný korelační koeficient, který zohledňuje případnou neshodu středních hodnot nebo rozptylů.

1.5.2 Koeficient kappa

Konkordanční korelační koeficient můžeme zobecnit, když při jeho zavedení použijeme místo kvadratické funkce obecně nějakou nezápornou měřitelnou funkci $w(x,y)$

$$\rho_{\mathrm{cw}} = 1 - \frac{1-\mathsf{E}\,w(X,Y)}{1-\mathsf{E}_{\,\mathrm{indep}}\,w(X,Y)}. \tag{1.67}$$

Uvažujme měření nominálního znaku s hodnotami $1,2,\ldots,k$ dvěma experty X,Y. Zvolme

$$w(i,j) = \begin{cases} 1 & \text{pro } i=j, \\ 0 & \text{pro } i\neq j. \end{cases}$$

Střední hodnota $w(X,Y)$ je pak rovna pravděpodobnosti, že $X = Y$ a platí tedy

$$\mathsf{P}(X = Y) = \sum_{j=1}^{k} \pi_{jj}\,,$$

kde $\sum_{j=1}^{k} \pi_{jj}$ je pravděpodobnost shody.

Vztah (1.67)) pak přejde na definici *kappa koeficientu* ([1])

$$\begin{aligned}
\kappa &= 1 - \frac{1 - \sum_{j=1}^{k} \pi_{jj}}{1 - \sum_{j=1}^{k} \pi_{j+}\pi_{+j}} \\
&= \frac{(1 - \sum_{j=1}^{k} \pi_{j+}\pi_{+j}) - (1 - \sum_{j=1}^{k} \pi_{jj})}{1 - \sum_{j=1}^{k} \pi_{j+}\pi_{+j}} \qquad (1.68) \\
&= \frac{\sum_{j=1}^{k} \pi_{jj} - \sum_{j=1}^{k} \pi_{j+}\pi_{+j}}{1 - \sum_{j=1}^{k} \pi_{j+}\pi_{+j}}\,, \qquad (1.69)
\end{aligned}$$

kde symboly π_{j+}, π_{+j} označují řádkové a sloupcové marginální pravděpodobnosti.

Hodnotíme-li pomocí kappa koeficientu kvalitu predikce, porovnáváme shodu předpovědi se skutečností ve dvou modelech: v hodnoceném modelu a v modelu, v němž se rozhodujeme náhodně. Ve vyjádření (1.68) je význam kappa koeficientu dobře patrný. V čitateli zlomku je rozdíl pravděpodobností chybných rozhodnutí při náhodném rozhodování a rozhodování podle modelu, ve jmenovateli je pravděpodobnost chybného rozhodnutí při náhodném rozhodování. Kappa koeficient ukazuje, o kolik procent je rozhodování na základě modelu lepší než náhodné.

Příklad: Na obrázku 1.13 jsou znázorněny regresní křivky pro ordinální model závislosti úspěšnosti ve studiu na dosaženém počtu bodů u přijímacích zkoušek. Úspěšnost je ordinální veličina, která nabývá tří hodnot: 0 znamená neúspěch (zanechání studia, příp. jeho neukončení do pěti let od začátku studia), 1 označuje úspěšné ukončení studia, ale nikoliv ve standardní době studia a 2 je u studentů, kteří úspěšně ukončili studium ve standardní době studia (3 roky).

Porovnání predikce za platnosti modelu a skutečné úspěšnosti je v tabulce 1.4. Dosazením konkrétních hodnot do vzorce (1.69) dostaneme

$$\widehat{\kappa} = \frac{(35 + 0 + 38)/140 - (71 \cdot 51 + 0 + 69 \cdot 57)/140^2}{1 - (71 \cdot 51 + 0 + 69 \cdot 57)/140^2} = 0{,}221\,,$$

což znamená, že předpověď podle modelu je o 22 % lepší náhodné rozhodování. ♢

Pokud neměříme znak nominální, ale ordinální (s k uspořádanými hodnotami), můžeme zohlednit počet úrovní, o které se jednotlivá rozhodnutí liší (lineárně nebo kvadraticky). Na hlavní diagonále (shoda expertů) je $w(i,i) = w_{ii} = 1$, při nejvyšším možném rozdílu počtu úrovní $k-1$, $w(1,k) = w_{1k} = 0$. V lineárním případě budeme při k úrovních pracovat s funkcí (vahami)

$$w(i,j) = w_{ij} = 1 - \frac{|i-j|}{k-1}, \qquad i,j = 1,\ldots,k\,.$$

Tabulka 1.4: Porovnání predikce a skutečné úspěšnosti ve studiu v závislosti na počtu bodů u přijímacích zkoušek

	skutečnost			
predikce	0	1	2	celkem
0 neúspěch	35	17	19	71
1 úspěch ne ve std. době	0	0	0	0
2 úspěch ve std. době	16	15	38	69
celkem	51	32	57	140

Po dosazení do (1.67) dostaneme tzv. vážený kappa koeficient.

$$\begin{aligned}\kappa_{\mathrm{w}} &= 1 - \frac{1 - \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\pi_{ij}}{1 - \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\pi_{i+}\pi_{+j}} \\ &= \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\pi_{ij} - \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\pi_{i+}\pi_{+j}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}\pi_{i+}\pi_{+j}}\end{aligned} \tag{1.70}$$

Příklad: Pro výše uvedený model využijeme toho, že úspěšnost ve studiu, která nabývá hodnot $0, 1, 2$, můžeme považovat za ordinální veličinu. Vzhledem k počtu kategorií je matici vah:

$$\mathbf{W} = \begin{pmatrix} 1 & 0{,}5 & 0 \\ 0{,}5 & 1 & 0{,}5 \\ 0 & 0{,}5 & 1 \end{pmatrix}$$

a dosazením do (1.70) vychází $\widehat{\kappa_{\mathrm{w}}} = 0{,}272$.

Nízké hodnoty kappa koeficientu (neváženého i váženého) ukazují, že predikce tímto způsobem nedává dobré výsledky. $\diamondsuit$

Kapitola 2

Aplikace metod na reálná data

2.1 Popis dat

Data pro analýzu byla získána ze studijního informačního systému Univezity Karlovy (dále SIS). Již od akademického roku 2002/03 jsou v této databázi (do roku 2007 nebyla pro celou Univezitu Karlovu centralizována) evidovány předměty, které si jednotliví studenti zapisují a zároveň i pedagogové elektronicky vyplňují výsledky zkoušek. Tím je umožněno získat rozsáhlé datové soubory v elektronické podobě pro další zpracování. Samozřejmě není možno data hned použít, bylo potřeba je důsledně prohlédnout, a zkontrolovat, zda některé údaje nejsou zjevně neprávné (například zadaný středoškolský prospěch mimo možné hodnoty). Mnohé nejasnosti byly vyřešeny za ochotné spolupráce referentek studijního oddělení.

Přesný popis jednotlivých veličin s jejich označením a kódováním je uveden v příloze A.

2.1.1 Studium na PřF UK

Na PřF je možno studovat devět bakalářských studijních programů, které se dále člení na obory. Pro větší názornost byly příbuzné studijní programy sloučeny do následujících skupin s charakterizujícím označením, které je používáno v tabulkách a grafech.

označení	skutečné studijní programy	používaná barva
B	Biologie	žlutá (yellow)
B	Speciální chemicko–biologické obory	žlutá (yellow)
C	Chemie	fialová (magenta)
C	Biochemie	fialová (magenta)
C	Klinická a toxikologická analýza	fialová (magenta)
D	Demografie	světle modrá (cyan)
Z	Geografie	modrá (blue)
G	Geologie	zelená (green)
O	Ekologie a ochrana prostředí	oranžová (orange)
U	všichny obory zaměřené na vzdělávání	červená (red)

Přijímací zkoušky sestávaly ze zkoušky z jednoho nebo ze dvou předmětů. Maximum z každého předmětu bylo 100 bodů. Pro jednotné měřítko celkového počtu bodů získaných v přijímacím řízení byl počet bodů v případě jediné zkoušky zdvojnásoben. Přijímací zkoušky se u jednotlivých skupin programů konaly z následujících předmětů:

B biologie + chemie

C chemie

D matematika + geografie / historie (podle oboru)

Z matematika + geografie

G volba dvou předmětů

O biologie + chemie

U předměty podle oboru studia

2.1.2 Zapsaní v akademickém roce 2003/04

V akademickém roce 2003/04 byli studenti v rámci nově zaváděného trojstupňového studia přijímáni poprvé ke studiu bakalářských studijních oborů.

Ke studiu bylo podáno celkem 3376 přihlášek. Mnoho studentů podalo přihlášek několik, velká část uchazečů se ke zkouškám nedostavila (odpovídající více než 20 % přihlášek). V případě, že student podal více přihlášek a různé obory vyžadovaly zkoušku z téhož předmětu, student ji vykonal pouze jednou a výsledky byly zahrnuty do hodnocení všech žádostí o přijetí.

Podmínkou pro přijetí bez přijímacích zkoušek byla účast v celostátním kole olympiády. Na základě této podmínky bylo akceptováno 26 přihlášek, které se týkaly 23 studentů (jeden byl přijat na dva a jeden na tři obory). Z těchto studentů se pouze 12 ke studiu skutečně zapsalo. Nepřijatým studentům oboru Ochrana životního prostředí byla nabídnuta možnost studia geologických oborů, pakliže v přijímacím řízení získali více bodů než byla hranice přijímání na geologické obory. Ke studiu se takto dodatečně zapsalo 22 uchazečů.

Z celkového počtu 1203 přihlášek, které byly na základě přijímacího řízení akceptovány, se ke studiu zapsalo celkem 680 studentů (za toho tři studenti na dva obory).

U dvou dívek nejsou výsledky přijímacích zkoušek zaznamenány v databázi SIS v programu Student a jejich složky jsou již v archivu RUK, proto byly ze studie vyřazeny. Mezi studenty bylo celkem 22 cizinců, z toho 18 ze Slovenska. Ti byli s ohledem na srovnatelné údaje ponecháni, zbývajících 4 cizinci byli vyřazeni. Studenti, kteří se zapsali ke studiu dvou studijních programů, ve dvou případech jednoho ze studií zanechali a jedna studentka druhý obor absolvovala o rok později. Pro lepší vypovídací schopnost byly ponechány údaje o úspěšnějším studiu.

Celkem tedy datový soubor obsahuje informace o 705 studentech.

Celkem u 14 studentů (různých programů) nebyly k dispozici údaje o středoškolském prospěchu, proto nebyli v případě některých výpočtů pro výpočty vzati v úvahu.

V tabulce 2.1 jsou uvedeny počty zapsaných uchazečů v jednotlivých skupinách programů. Šířka sloupců na obrázku 2.1 (vlevo) je úměrná těmto četnostem, barevně jsou rozlišeni studenti podle pohlaví a úspěšnosti ve studiu.

Tabulka 2.2 slouží pro porovnání průměrných ukazatelů podle pohlaví a skupin programů. Jsou patrny velké rozdíly mezi jednotlivými skupinami oborů. Ty mohou být způsobeny různou intenzitou zájmu, odlišnou hranicí minimálního počtu bodů pro

Tabulka 2.1: Počty studentů podle skupin programů, pohlaví a úspěšnosti studia

začátek studia v akademickém roce 2003/04									
skupiny programů	počet studentů								
	celkem			absolventů			abs. ve stand. době		
	muži	ženy	celkem	muži	ženy	celkem	muži	ženy	celkem
B	38	109	147	31	89	120	24	79	103
C	61	88	149	35	60	95	24	44	68
D	15	35	50	6	29	35	3	17	20
G	41	48	89	20	34	54	16	23	29
O	7	11	18	4	10	14	2	7	9
U	34	75	109	13	39	52	4	8	12
Z	72	71	143	43	48	91	32	26	58
celkem	268	437	705	152	309	461	105	204	309
začátek studia v akademickém roce 2004/05									
	muži	ženy	celkem	muži	ženy	celkem	muži	ženy	celkem
B	49	115	164	37	94	131	34	80	114
C	67	113	180	39	74	113	30	62	92
D	22	28	50	13	18	31	6	10	16
G	25	45	70	13	31	44	9	21	30
O	13	24	37	10	21	31	7	18	25
U	36	91	127	11	48	59	8	33	41
Z	42	41	83	20	27	47	12	21	33
celkem	254	457	711	143	313	456	106	245	351

Tabulka 2.2: Aritmetické průměry ukazatelů podle skupin programů a pohlaví

začátek studia v akademickém roce 2003/04									
skupiny programů	body u přijímaček (u těch, kteří je konali)			prospěch na SŠ (kde je znám)			prospěch na VŠ (u absolventů)		
	muži	ženy	celkem	muži	ženy	celkem	muži	ženy	celkem
B	160,3	157,3	158,0	1,68	1,59	1,62	1,90	1,80	1,83
C	141,8	143,0	142,5	1,89	1,76	1,81	2,02	2,08	2,06
D	121,6	120,9	121,1	2,11	1,74	1,85	2,30	2,04	2,09
G	116,9	121,8	119,5	2,17	1,97	2,06	2,23	2,19	2,21
O	159,1	159,6	159,4	1,84	1,60	1,70	2,30	2,20	2,23
U	127,0	118,2	120,9	2,08	1,87	1,94	2,52	2,64	2,61
Z	115,1	116,5	115,8	2,00	1,69	1,85	2,47	2,42	2,45
celkem	136,9	138,1	137,7	1,97	1,74	1,83	2,21	2,14	2,16
začátek studia v akademickém roce 2004/05									
	muži	ženy	celkem	muži	ženy	celkem	muži	ženy	celkem
B	144,2	141,8	142,5	1,72	1,60	1,64	1,80	1,84	1,83
C	132,7	129,1	130,4	1,82	1,74	1,77	1,88	2,19	2,08
D	112,2	117,9	115,4	2,06	1,86	1,95	2,14	2,09	2,11
G	100,1	92,9	95,5	2,33	1,96	2,09	2,42	2,36	2,38
O	124,2	125,3	124,9	1,95	1,68	1,78	2,26	2,12	2,17
U	118,5	117,0	117,4	2,15	1,84	1,93	2,23	2,31	2,30
Z	118,7	124,9	121,8	1,96	1,72	1,84	2,41	2,28	2,33
celkem	124,5	124,8	124,7	1,94	1,75	1,82	2,06	2,12	2,10

přijetí na základě přijímacích zkoušek, obtížností jednotlivých studijních oborů i různou úrovní známkování na konkrétní sekci fakulty, která zajišťuje většinu výuky.

2.1.3 Zapsaní v akademickém roce 2004/05

Ke studiu bakalářských studijních oborů bylo podáno celkem 3572 přihlášek. Poprvé bylo možno studovat obor Molekulární biologie a biochemie mikroorganismů (součást skupiny programů B). Mnoho studentů podalo přihlášek několik, velká část uchazečů se ke zkouškám nedostavila (odpovídající 24 % přihlášek). V případě, že student podal více přihlášek a různé obory vyžadovaly stejný předmět přijímací zkoušky, student zkoušku z daného předmětu vykonal pouze jednou a výsledky byly zahrnuty do hodnocení všech žádostí o přijetí.

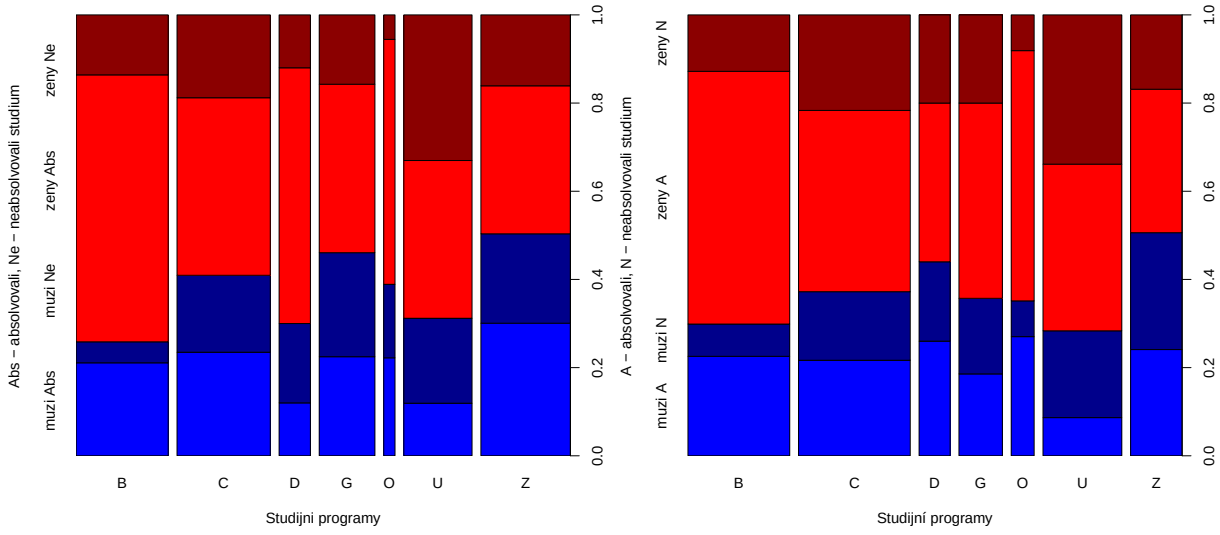
Podmínky pro přijetí bez přijímacích zkoušek byly stejné jako v předcházejících letech. Na základě účasti v celostátním kolem olympiády (biologické, chemické, matematické apod.) bylo akceptováno 34 přihlášek, které se týkaly 31 studentů (10 dívek a

21 chlapců; dva chlapci a jedna dívka byli přijati na dva obory). 22 z nich se ke studiu skutečně zapsalo (15 chlapců a 7 dívek, jeden student na dva obory). Podobně jako v minulém roce bylo dodatečně přijato 37 studentů na geologické obory (9 chlapců a 28 dívek), z nich se ke studiu zapsalo 32 (8 chlapců a 24 dívek).

Z 1276 přihlášek, které byly na základě přijímacího řízení akceptovány, se ke studiu zapsalo celkem 671 studentů (za toho 9 studentů na dva obory). Mezi studenty bylo celkem 37 cizinců, z toho 34 ze Slovenska. Ti byli s ohledem na srovnatelné údaje ponecháni, na rozdíl od zbývajících 3 cizinců. Dále bylo vyřazeno celkem 7 studentů, u nichž nebyly v databázi zaznamenány výsledky přijímacích zkoušek. Většinou se jednalo o přestupy z jiných oborů. Z 9 studentů, kteří se zapsali ke studiu dvou studijních oborů, pouze jediná studentka dokončila obě studia a jedna obou studií zanechala. Náhodně bylo vybráno jedno ze studií. U ostatních byly ponechány údaje o úspěšnějším (absolvovaném) studiu.

Celkem tedy datový soubor obsahuje informace o 711 studentech. U 30 studentů (různých programů) nebyly k dispozici údaje o středoškolském prospěchu, proto nebyli v případě některých výpočtech vzati v úvahu.

Podíly studentů jednotlivých skupin oborů, kteří absolvovali nebo neabsolvali studium, jsou přehledně znázorněny na obrázku 2.1 (vpravo), jejich absolutní počty v tabulce 2.1, průměrné ukazatele podle pohlaví a skupin programů v tabulce 2.2.



Obrázek 2.1: Studenti PřF UK podle pohlaví, jednotlivých skupin oborů a úspěšnosti ve studiu zapsaní v roce 2003 (vlevo) a v roce 2004 (vpravo). Modře jsou vyznačeni chlapci, červeně dívky, tmavším odstínem ti, kteří studium neabsolvovali.

2.2 Predikce číselných kritérií úspěšnosti

2.2.1 Průměrný prospěch na VŠ

Chceme předpovídat konečný průměrný prospěch na základě znalosti prospěchu na střední škole a počtu bodů dosažených v přijímacím řízení. Datový soubor omezíme

na studenty, u nichž je znám středoškolský prospěch, výsledky přijímacích zkoušek a studium úspěšně absolvovali.

Ve výsledku funkce `summary()` pro lineární model je vždy nejdříve model znovu schematicky zapsán a dále jsou uvedeny kvartily reziduí. Odhady jednotlivých regresních koeficientů jsou doplněny testem jejich nulovosti, přehledně označené symbolem pro hladinu významnosti ($***$ 0.001, $**$ 0.01, $*$ 0.05, $\cdot$ 0.1). Kromě hodnoty koeficientu determinace je uvedena i hodnota adjustovaného koeficientu determinace.

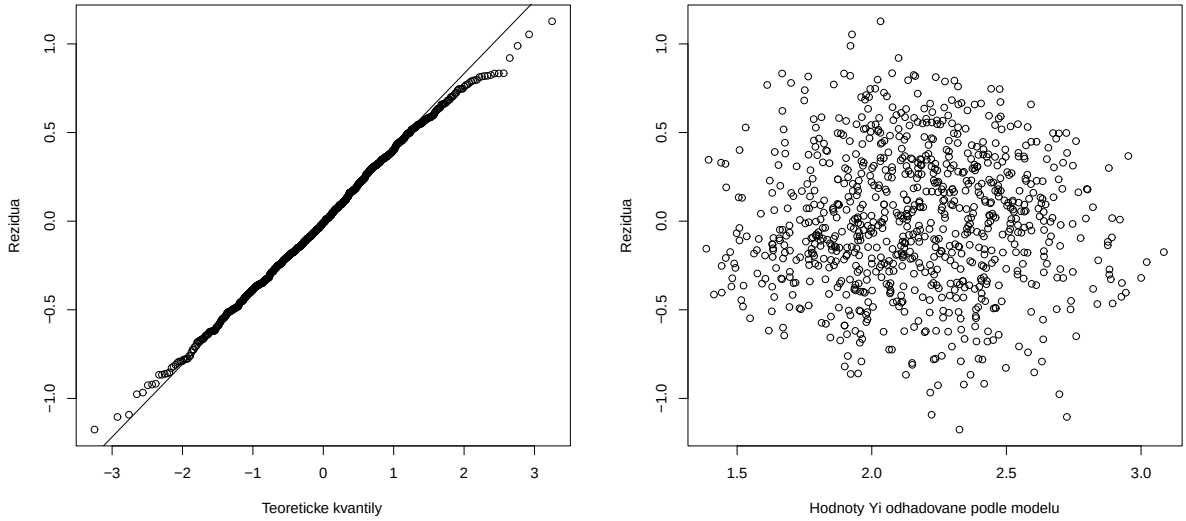
Z údajů o 867 studentech určíme následující odhady koeficientů regresní funkce:

```
a<-lm(prumer~zcel+ssprum, subset=(JeSS=="ano"&Prijeti!="79"&Abs=="1"))
summary(a)

Call:
lm(formula = prumer ~ zcel + ssprum, subset = (JeSS == "ano" &
    Prijeti != "79" & Abs == "1"))

Residuals:
      Min        1Q    Median        3Q       Max
-1.175752 -0.264099 -0.009366  0.289036  1.127443

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.5252240  0.1099229   22.97   <2e-16 ***
zcel        -0.0081774  0.0005496  -14.88   <2e-16 ***
ssprum       0.4209454  0.0333249   12.63   <2e-16 ***

Residual standard error: 0.3896 on 864 degrees of freedom
Multiple R-squared: 0.4118,     Adjusted R-squared: 0.4104
F-statistic: 302.4 on 2 and 864 DF,  p-value: < 2.2e-16
```

Vliv obou regresorů je průkazný na jakékoli stanovené hladině.

Je však nutno ověřit předpoklady lineárního modelu. Obr. 2.2 demonstruje splnění předpokladu normality residuí (vlevo) a nezávislosti rozptylu reziduí na odhadované střední hodnotě (vpravo). Potvrzují to i následující výpočty (pro funkci `bptest` je třeba načíst knihovnu `lmtest`).

```
a<-lm(prumer~zcel+ssprum, subset=(JeSS=="ano"&Prijeti!="79"&Abs=="1"))
shapiro.test(resid(a))

        Shapiro-Wilk normality test
data:  resid(a)
W = 0.9975, p-value = 0.2164

library(lmtest)
bptest(a,~fitted(a))
        studentized Breusch-Pagan test
data:  a
BP = 1.3617, df = 1, p-value = 0.2432
```



Obrázek 2.2: Normální diagram (vlevo) a závislost reziduí na odhadované střední hodnotě (vpravo) modelu závislosti konečného průměru známek na vysoké škole na průměrném prospěchu na střední škole a počtu bodů dosažených v přijímacím řízení.

V tomto modelu

$$prumer = -0{,}0082 \cdot zcel + 0{,}421 \cdot ssprum + 2{,}525$$

je koeficient determinace $R^2 = 0{,}412$, podařilo se nám vysvětlit více než 40 % variability konečného průměru závislostí na dvou regresorech.

Normováním veličin můžeme vliv obou regresorů lépe porovnat. V zápise pomocí standardizovaných veličin

$$prumer = -0{,}212 \cdot zzcel + 0{,}180 \cdot zssprum + 2{,}160$$

je zřejmá vyšší absolutní hodnota koeficientu u přijímacích zkoušek.

Sílu vlivu obou regresorů můžeme porovnat pomocí vzorce (1.11), kde použitím funkce `paired.r` z knihovny `psych` neprokážeme rozdíl ($p = 0{,}19$).

```
library(psych)
attach(vse[JeSS=="ano"&Prijeti!="79"&Abs=="1",])
(r01=abs(cor(prumer,zcel)))
[1] 0.5505713
(r02=abs(cor(prumer,ssprum)))
[1] 0.5109236
(r12=abs(cor(zcel,ssprum)))
[1] 0.3724115
(n=length(prumer))
[1] 867

paired.r(r01,r02,r12,n)
$test
```

```
[1] "test of difference between two correlated  correlations"
$t
[1] 1.312037
$p
[1] 0.1898559
```

Jestliže standardizujeme i vysvětlovanou veličinu, dostaneme model bez absolutního členu a vypočtené odhady koeficientů se označují jako tzv. betaváhy [10], které se využívají k porovnání modelů s odpovídajícími regresory, ale u jiných datových souborů.

$$zprumer = -0,418 \cdot zzcel + 0,355 \cdot zssprum$$

```
attach(vse[JeSS=="ano"&Prijeti!="79"&Abs=="1",])
summary(lm(scale(prumer)~scale(zcel)+scale(ssprum)))

Coefficients:
                Estimate Std. Error  t value Pr(>|t|)
(Intercept)    4.660e-16  2.608e-02 1.79e-14        1
scale(zcel)   -4.183e-01  2.812e-02   -14.88   <2e-16 ***
scale(ssprum)  3.551e-01  2.812e-02    12.63   <2e-16 ***

Residual standard error: 0.7679 on 864 degrees of freedom
Multiple R-squared: 0.4118,     Adjusted R-squared: 0.4104
F-statistic: 302.4 on 2 and 864 DF,  p-value: < 2.2e-16
```

Pokusíme se přidat do modelu další známé veličiny. Přidáním informace, zda student maturoval v roce přijímacích zkoušek (*Maturdrive*), nedostaneme statisticky významně lepší model, jak se můžeme přesvědčit například testem podmodelu.

```
a<-lm(prumer~zcel+ssprum, subset=(JeSS=="ano"&Prijeti!="79"&Abs=="1"))
a2<-lm(prumer~zcel+ssprum+Maturdrive,
 subset=(JeSS=="ano"&Prijeti!="79"&Abs=="1"))
anova(a,a2)
Analysis of Variance Table

Model 1: prumer ~ zcel + ssprum
Model 2: prumer ~ zcel + ssprum + Maturdrive
  Res.Df     RSS  Df Sum of Sq      F Pr(>F)
1    864 131.173
2    863 131.147   1     0.026 0.1739 0.6768
```

Zopakujeme-li test podmodelu po zařazení pohlaví studenta *Pohlavi* nebo faktoru státní příslušnosti *Stat* (Česká republika, Slovensko) do původního modelu se dvěma regresory, ani v jednom případě nedostaneme statisticky významně lepší model ($p = 0,72$, resp. $p = 0,36$).

Prokážeme však rozdíl mezi oběma roky začátku studia (významnost faktoru *Rok*).

```
a2<-lm(prumer~zcel+ssprum+Rok,
 subset=(JeSS=="ano"&Prijeti!="79"&Abs=="1"))
```

```
summary(a2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.6566711  0.1133456  23.439  < 2e-16 ***
zcel        -0.0085864  0.0005532 -15.522  < 2e-16 ***
ssprum       0.4079569  0.0331581  12.303  < 2e-16 ***
Rok2004     -0.1113704  0.0266583  -4.178 3.24e-05 ***


Residual standard error: 0.386 on 863 degrees of freedom
Multiple R-squared: 0.4234,     Adjusted R-squared: 0.4214
F-statistic: 211.3 on 3 and 863 DF,  p-value: < 2.2e-16
```

Koeficient $-0{,}111$ u faktoru Rok2004 znamená, že u studentů, kteří se zapsali ke studiu v akademickém roce 2004/05, očekáváme v průměru o 0,111 lepší prospěchový průměr při stejné úspěšnosti v přijímacím řízení a stejném průměru známek na střední škole.

Statistická významnost faktoru *Rok* může být způsobena změnami hranice počtu bodů dosažených v přijímacím řízení nutných pro přijetí. Významně se liší průměrné počty bodů z přijímacího řízení v obou letech ($p < 0{,}0001$) na rozdíl od průměrného prospěchu na střední škole nebo na vysoké škole, kde dvouvýběrovým testem rozdíl neprokážeme ($p = 0{,}39$, resp. $p = 0{,}12$).

```
t.test(zcel~Rok, var.equal=TRUE,
 subset=(JeSS=="ano"&Prijeti!="79"&Abs=="1"))

        Two Sample t-test
data:  zcel by Rok
t = 4.5661, df = 865, p-value = 5.689e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4.540085 11.385657
sample estimates:
mean in group 2003 mean in group 2004
          137.7247           129.7618
```

Podobně zlepšíme předpověď rozlišením skupin studentů podle studijních programů, tedy přidáním prediktoru *Prog* do modelu. Funkce `Anova()` (z knihovny `car`) ukazuje statistickou významnost jednotlivých regresorů modelu. Vždy je uveden součet čtverců, o které by se zvýšil reziduální součet čtverců *RSS*, kdybychom daný regresor z modelu vyřadili. Obdobný výstup funkce `anova()` závisí na zvoleném pořadí regresorů. Součet čtverců v každém řádku udává v tomto případě rozdíl od modelu, který obsahuje všechny výše uvedeneé regresory.

```
library(car)
Anova(a3)
Anova Table (Type II tests)
```

```
Response: prumer
           Sum Sq  Df F value    Pr(>F)
zcel       21.795   1 168.300 < 2.2e-16 ***
ssprum     26.244   1 202.654 < 2.2e-16 ***
Rok         1.992   1  15.383  9.48e-05 ***
Prog       17.589   6  22.637 < 2.2e-16 ***
Residuals 110.984 857
```

```
Analysis of Variance Table

Response: prumer
           Df  Sum Sq Mean Sq F value    Pr(>F)
zcel        1  67.595  67.595 521.963 < 2.2e-16 ***
ssprum      1  24.224  24.224 187.054 < 2.2e-16 ***
Rok         1   2.600   2.600  20.079  8.44e-06 ***
Prog        6  17.589   2.932  22.637 < 2.2e-16 ***
Residuals 857 110.984   0.130
```

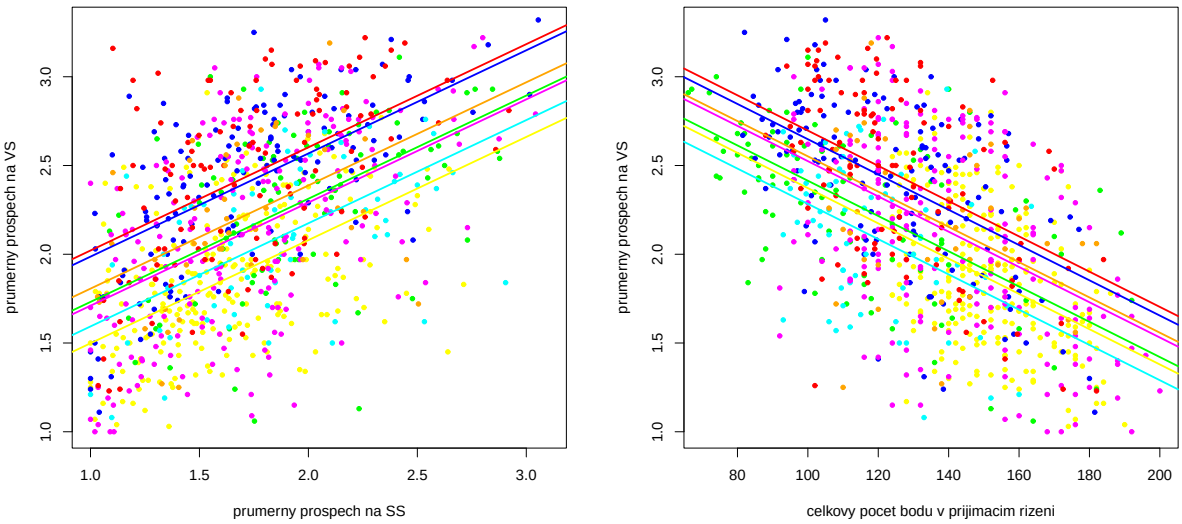
U funkce `summary()` mají koeficienty u jednotlivých skupin programů následující význam. Program B je zvolen jako základní úroveň a například koeficient 0,14 u ProgC (resp. *Prog* = C) znamená, že u studentů chemických programů očekáváme o 0,14 horší konečný průměr než u studentů biologických programů za předpokladu nezměněných hodnot ostatních regresorů.

```
a3<-lm(prumer~zcel+ssprum+Rok+Prog,
 subset=(JeSS=="ano"&Prijeti!="79"&Abs=="1"))
summary(a3)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.3604768  0.1234482  19.121  < 2e-16 ***
zcel        -0.0077331  0.0005961 -12.973  < 2e-16 ***
ssprum       0.4433974  0.0311469  14.236  < 2e-16 ***
Rok2004     -0.1002566  0.0255619  -3.922 9.48e-05 ***
ProgC        0.1406662  0.0358151   3.928 9.27e-05 ***
ProgD       -0.1250282  0.0537084  -2.328 0.020149 *
ProgG       -0.0216145  0.0490417  -0.441 0.659514
ProgO        0.2126292  0.0598387   3.553 0.000401 ***
ProgU        0.3366708  0.0444389   7.576 9.23e-14 ***
ProgZ        0.2772809  0.0423469   6.548 1.00e-10 ***

Residual standard error: 0.3599 on 857 degrees of freedom
Multiple R-squared: 0.5023,     Adjusted R-squared: 0.4971
F-statistic:  96.1 on 9 and 857 DF,  p-value: < 2.2e-16
```

Grafické znázornění je v případě více regresorů obtížné. Jen pro ilustraci jsou na obrázku 2.3 znázorněny závislosti ve dvou různých modelech. Levá část odpovídá modelu závislosti průměrného prospěchu na vysoké škole na prospěchu na střední škole se

zařazením vlivu studijního programu. Vpravo je znázorněna také závislost průměrného prospěchu na vysoké škole, ale tentokrát na počtu bodů dosažených v přijímacím řízení a studijním programu. Je patrné, že při stejném průměrném prospěchu na SŠ, očekáváme u studentů geografie (modře) a učitelství (červeně) horší známkové průměry, což může (ale nemusí) být způsobeno větší obtížností takto zaměřeného studia.



Obrázek 2.3: Závislost konečného průměru známek na vysoké škole na průměrném prospěchu na střední škole (vlevo) a na počtu bodů dosažených v přijímacím řízení (vpravo) podle jednotlivých skupin studijních programů.

Konečný model závislosti

Pro odhad průměru známek na VŠ je tedy vhodné využít model se čtyřmi regresory *ssprum*, *zcel*, *Rok* a *Prog*, který obsahuje celkem 10 parametrů a kde dokážeme vysvětlit více než 50 % kolísání konečného průměru závislostí na uvedených regresorech. Ověřili jsme, zda nejsou porušeny předpoklady normálního lineárního modelu:

```
a3<-lm(prumer~zcel+ssprum+Rok+Prog,
 subset=(JeSS=="ano"&Prijeti!="79"&Abs=="1"))
shapiro.test(resid(a3))

        Shapiro-Wilk normality test
data:  resid(a3)
W = 0.9983, p-value = 0.5864

bptest(a3,~fitted(a3))

        studentized Breusch-Pagan test
data:  a3
BP = 1.362, df = 1, p-value = 0.2432
```

Funkce `vif()` z knihovny `car` vypočte hodnoty inflačního faktoru *VIF* (označené *GVIF* s ohledem na možné zobecnění) pro všechny odhadované regresní koeficienty.

Ve třetím sloupci jsou uvedeny odmocniny z *VIF*, u faktoru *Prog* se sedmi úrovněmi, které nahrazujeme šesti umělými proměnnými, je *VIF* umocněno na 1/12. Z nízkých hodnot *VIF* je vidět, že každý regresor se chová skoro nezávisle na ostatních regresorech v modelu.

```
library(car)
vif(a3)
           GVIF Df GVIF^(1/2Df)
zcel   1.601031  1     1.265319
ssprum 1.189017  1     1.090421
Rok    1.092849  1     1.045394
Prog   1.439410  6     1.030818
```

Kroková regrese

Využitím funkce `step()` je možno nechat program mechanicky navrhnout model krokovou regresí. Vždy musíme zvolit počáteční (lower) a nejbohatší možný model (upper). Je možno se omezit pouze na přidávání nebo odebírání regresorů z modelu.

V tomto případě začneme modelem pouze s absolutním členem a do nejbohatšího modelu zařadíme všechny známé veličiny (bez vzájemných interakcí). Ve výpisu provádění jednotlivých kroků jsou změny seřazeny vzestupně podle hodnoty AIC. V každém řádku je uvedeno, jak by se přidáním konkrétního regresoru do modelu nebo jeho odebráním z modelu (v případě, že tam již je) změnil součet čtverců (resp. reziduální součet čtverců) a kolik by byla hodnota AIC.

```
attach(vse[JeSS=="ano"&Prijeti!="79"&Abs=="1",])
aS<-step(lm(prumer~1),scope=list(lower=~1,
 upper=~Prog+zcel+ssprum+ss1+ss2+ss3+ss4+
 Maturdrive+Pohlavi+Prijeti+Stat+Rok))
Start:  AIC=-1175.3
prumer ~ 1

             Df Sum of Sq      RSS      AIC
+ zcel        1     67.60   155.40 -1486.42
+ ssprum      1     58.21   164.78 -1435.58
+ ss3         1     54.99   168.00 -1418.79
+ ss4         1     54.46   168.53 -1416.07
+ ss2         1     49.51   173.48 -1390.98
+ ss1         1     42.98   180.01 -1358.95
+ Prog        6     39.12   183.87 -1330.55
+ Maturdrive  1      1.94   221.05 -1180.88
+ Prijeti     1      0.93   222.07 -1176.91
+ Pohlavi     1      0.68   222.32 -1175.93
+ Rok         1      0.61   222.38 -1175.67
<none>                      222.99 -1175.30
+ Stat        1      0.20   222.79 -1174.07

Step:  AIC=-1486.42
prumer ~ zcel
```

```
             Df Sum of Sq      RSS      AIC
+ ssprum      1     24.22   131.17 -1631.35
+ ss3         1     22.40   133.00 -1619.37
+ ss4         1     21.70   133.70 -1614.83
+ ss2         1     20.19   135.21 -1605.08
+ ss1         1     18.13   137.26 -1592.00
+ Prog        6     14.45   140.94 -1559.07
+ Rok         1      4.27   151.13 -1508.59
+ Maturdrive  1      1.14   154.25 -1490.83
+ Prijeti     1      0.87   154.53 -1489.28
+ Pohlavi     1      0.66   154.74 -1488.12
<none>                      155.40 -1486.42
+ Stat        1      0.08   155.32 -1484.85
- zcel        1     67.60   222.99 -1175.30

Step:  AIC=-1631.35
prumer ~ zcel + ssprum

             Df Sum of Sq      RSS      AIC
+ Prog        6     18.20   112.98 -1748.83
+ Rok         1      2.60   128.57 -1646.71
+ Prijeti     1      0.70   130.48 -1633.97
<none>                      131.17 -1631.35
+ ss1         1      0.18   130.99 -1630.56
+ ss4         1      0.15   131.02 -1630.33
+ Stat        1      0.13   131.05 -1630.17
+ ss3         1      0.11   131.06 -1630.08
+ ss2         1      0.08   131.09 -1629.89
+ Maturdrive  1      0.03   131.15 -1629.52
+ Pohlavi     1      0.02   131.15 -1629.47
- ssprum      1     24.22   155.40 -1486.42
- zcel        1     33.61   164.78 -1435.58

Step:  AIC=-1748.83
prumer ~ zcel + ssprum + Prog

             Df Sum of Sq      RSS      AIC
+ Rok         1      1.99   110.98 -1762.25
+ ss4         1      0.31   112.67 -1749.18
+ ss1         1      0.27   112.71 -1748.88
<none>                      112.98 -1748.83
+ Stat        1      0.25   112.73 -1748.74
+ Maturdrive  1      0.25   112.73 -1748.74
+ Pohlavi     1      0.20   112.77 -1748.39
+ ss2         1      0.07   112.91 -1747.34
+ ss3         1      0.05   112.93 -1747.19
+ Prijeti     1 0.0002455   112.98 -1746.83
```

```
- Prog        6     18.20   131.17 -1631.35
- zcel        1     19.87   132.85 -1610.34
- ssprum      1     27.97   140.94 -1559.07

Step:  AIC=-1762.25
prumer ~ zcel + ssprum + Prog + Rok

             Df Sum of Sq      RSS      AIC
+ Maturdrive  1      0.38   110.60 -1763.23
+ Stat        1      0.29   110.69 -1762.55
<none>                      110.98 -1762.25
+ ss1         1      0.24   110.75 -1762.11
+ ss4         1      0.24   110.75 -1762.10
+ Pohlavi     1      0.17   110.81 -1761.59
+ ss3         1      0.03   110.95 -1760.51
+ ss2         1      0.03   110.95 -1760.50
+ Prijeti     1  0.004033   110.98 -1760.29
- Rok         1      1.99   112.98 -1748.83
- Prog        6     17.59   128.57 -1646.71
- zcel        1     21.80   132.78 -1608.80
- ssprum      1     26.24   137.23 -1580.22

Step:  AIC=-1763.23
prumer ~ zcel + ssprum + Prog + Rok + Maturdrive

             Df Sum of Sq      RSS      AIC
<none>                      110.60 -1763.23
+ Stat        1      0.23   110.38 -1763.00
+ ss1         1      0.20   110.40 -1762.81
+ Pohlavi     1      0.19   110.42 -1762.71
+ ss4         1      0.18   110.42 -1762.65
- Maturdrive  1      0.38   110.98 -1762.25
+ ss3         1      0.04   110.56 -1761.53
+ ss2         1      0.03   110.57 -1761.45
+ Prijeti     1      0.01   110.59 -1761.30
- Rok         1      2.12   112.73 -1748.74
- Prog        6     17.89   128.50 -1645.21
- zcel        1     22.14   132.74 -1607.03
- ssprum      1     24.21   134.81 -1593.61
```

Výpočet skončí, když žádná změna nepřinese nižší AIC než $-1763{,}23$. Do modelu byly zařazeny regresory v tomto pořadí *zcel*, *ssprum*, *Prog*, *Rok* a *Maturdrive*. V tomto případě jsme dostali krokovou regresí podobný model jako předchozím postupem.

Vzhledem k tomu, že se jedná o zcela mechanický postup, který nebere v úvahu logické souvislosti, je vhodné krokovou regresi používat například jako pomůcku, máme-li k dispozici velké množství možných regresorů a nemáme dobrou představu o jejich smyslu.

2.2.2 Průměrný prospěch v 1. ročníku VŠ

Velmi často používaným kritériem úspěšnosti (zejména v USA) je průměrný prospěch na konci 1. ročníku VŠ. Do výpočtů zahrneme záznamy o 1126 studentech, u nichž známe středoškolský prospěch i výsledky přijímacích zkoušek a kteří splnili podmínky pro zápis do 2. ročníku. Použijeme-li stejné regresory jako pro přepověď průměrného prospěchu za celé studium, prokážeme jejich statistickou významnost i v tomto modelu.

```
a<-lm(pr1~zcel+ssprum+Prog+Rok,
 subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
summary(a)

Call:
lm(formula = pr1 ~ zcel + ssprum + Prog + Rok, subset = (JeSS ==
    "ano" & Prijeti != "79" & Splnil1 == "1"))

Residuals:
     Min       1Q   Median       3Q      Max
-1.37854 -0.31458  0.02439  0.31097  1.38318

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.8818991  0.1424885  20.225  < 2e-16 ***
zcel        -0.0100323  0.0007007 -14.317  < 2e-16 ***
ssprum       0.5593509  0.0341999  16.355  < 2e-16 ***
ProgC       -0.1192052  0.0424448  -2.808  0.00506 **
ProgD       -0.3504528  0.0637541  -5.497 4.79e-08 ***
ProgG       -0.4881855  0.0583402  -8.368  < 2e-16 ***
ProgO        0.0635554  0.0736049   0.863  0.38807
ProgU        0.2391283  0.0502833   4.756 2.24e-06 ***
ProgZ        0.1080364  0.0503510   2.146  0.03211 *
Rok2004     -0.1241488  0.0294825  -4.211 2.75e-05 ***

Residual standard error: 0.474 on 1116 degrees of freedom
Multiple R-squared: 0.4808,     Adjusted R-squared: 0.4766
F-statistic: 114.8 on 9 and 1116 DF,  p-value: < 2.2e-16
```

Koeficient determinace pro tento lineární model je jen nepatrně nižší ($R^2 = 0{,}48$) než při odhadu konečného prospěchu.

```
shapiro.test(resid(a))

        Shapiro-Wilk normality test
data:  resid(a)
W = 0.9982, p-value = 0.2965

bptest(a,~fitted(a))

        studentized Breusch-Pagan test
```

```
data:  a
BP = 2.1923, df = 1, p-value = 0.1387
```

Splnění předpokladů použití lineárního modelu můžeme taktéž předpokládat.

Přidáním prediktoru *Maturdrive* nebo *Stat* nedostaneme významně lepší model. Kdybychom rozhodovali na hladině významnosti 10 %, zřejmě bychom zařadili *Pohlavi* jako pátý regresor do modelu.

```
a2<-lm(pr1~zcel+ssprum+Prog+Rok+Pohlavi,
 subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
summary(a2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.38911 -0.30940  0.01836  0.31461  1.37567

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.797943   0.150267  18.620  < 2e-16 ***
zcel        -0.009904   0.000704 -14.068  < 2e-16 ***
ssprum       0.574007   0.035186  16.314  < 2e-16 ***
ProgC       -0.114663   0.042486  -2.699  0.00706 **
ProgD       -0.345102   0.063770  -5.412 7.64e-08 ***
ProgG       -0.482077   0.058392  -8.256 4.23e-16 ***
ProgO        0.067708   0.073576   0.920  0.35764
ProgU        0.240871   0.050247   4.794 1.86e-06 ***
ProgZ        0.121854   0.050924   2.393  0.01688 *
Rok2004     -0.122846   0.029465  -4.169 3.29e-05 ***
Pohlavi2     0.053954   0.030916   1.745  0.08123 .

Residual standard error: 0.4736 on 1115 degrees of freedom
Multiple R-squared: 0.4822,     Adjusted R-squared: 0.4776
F-statistic: 103.8 on 10 and 1115 DF,  p-value: < 2.2e-16
```

Koeficient 0,054 u Pohlavi2 odpovídá tomu, že u žen ($Pohlavi = 2$) bychom očekávali horší průměr na konci 1. ročníku než u mužů při stejných hodnotách ostatních regresorů.

Hodnoty inflačních faktorů u modelů `a` i `a2` jsou také nízké jako při předpovědi konečného průměru.

```
a<-lm(pr1~zcel+ssprum+Prog+Rok,
 subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
vif(a)
           GVIF Df GVIF^(1/2Df)
zcel   1.677395  1     1.295143
ssprum 1.202625  1     1.096643
Prog   1.508003  6     1.034825
Rok    1.088823  1     1.043467

a2<-lm(pr1~zcel+ssprum+Prog+Rok+Pohlavi,
```

```
 subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
vif(a2)
            GVIF Df GVIF^(1/2Df)
zcel    1.696033  1     1.302318
ssprum  1.275272  1     1.129280
Prog    1.552203  6     1.037319
Rok     1.089522  1     1.043802
Pohlavi 1.088965  1     1.043535
```

Pro porovnání vlivu prospěchu na střední škole a přijímacích zkoušek použijeme model s normovanými veličinami a dostaneme následující regresní funkci:

$$pr1 = -0{,}236 \cdot zzcel + 0{,}240 \cdot zssprum + 2{,}470$$

Vzhledem k velmi podobným hodnotám odhadnutých regresních koeficientů nepřekvapí vysoká p-hodnota testu shody odpovídajících korelačních koeficientů.

```
(r01=abs(cor(pr1,zcel)))
[1] 0.5041193
(r02=abs(cor(pr1,ssprum)))
[1] 0.5080482
(r12=abs(cor(zcel,ssprum)))
[1] 0.3926376
(n=length(pr1)) # pocet pozorovani
[1] 1126

paired.r(r01,r02,r12,n)
$test
[1] "test of difference between two correlated  correlations"
$t
[1] -0.1463792
$p
[1] 0.8836483
```

Využitím krokové regrese bychom dostali model s následujícími prediktory:

```
summary(aS)
lm(formula = pr1 ~ ssprum+Prog+zcel+Rok+ss4+Pohlavi+Stat)
```

Současné zařazení dvou středoškolských regresorů, i když je jejich vliv v modelu `aS` statisticky průkazný, zvýší rozptyl odhadů regresních koeficientů.

```
vif(aS)
            GVIF Df GVIF^(1/2Df)
ssprum  7.772933  1     2.787998
Prog    1.587751  6     1.039278
zcel    1.710428  1     1.307833
Rok     1.090932  1     1.044477
ss4     7.760711  1     2.785805
Pohlavi 1.114428  1     1.055665
Stat    1.042811  1     1.021181
```

2.3 Predikce kvalitativních kritérií úspěšnosti

2.3.1 Úspěšné ukončení studia

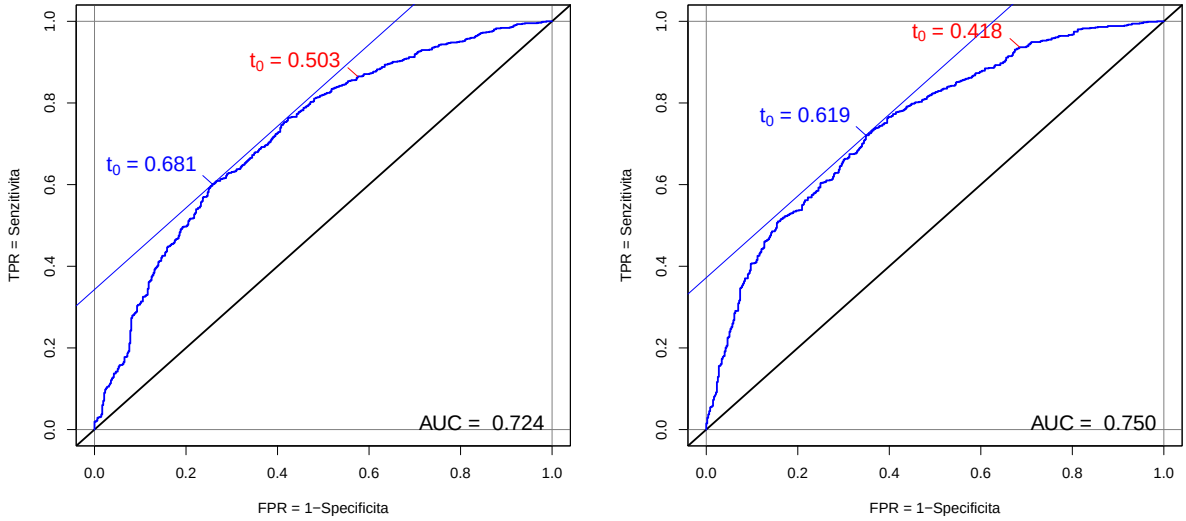
Do modelu logistické regrese, kdy odhadujeme pravděpodobnost úspěšného absolvování studia, můžeme zahrnout data o 1340 studentech, kteří konali přijímací zkoušky a je znám jejich středoškolský prospěch. Z tohoto počtu 867 studentů úspěšně absolvovalo studium a 473 studentů studia zanechalo, případně dosud studium neukončilo (studuje či studium přerušilo).

Podobně jako při předpovědi průměrného prospěchu pomocí lineární regrese prokážeme statistickou významnost středoškolského prospěchu *ssprum*, úspěšnosti v přijímacím řízení *zcel* a příslušnosti ke skupině programů *Prog*, akademický rok začátku studia *Rok* však byl shledán nevýznamným, a proto ho z modelu vyřadíme.

```
a=glm(Abs~zcel+ssprum+Prog+Rok,family =binomial,
 subset=(JeSS=="ano"&Prijeti!="79"))
summary(a)

Call:
glm(formula = Abs ~ zcel + ssprum + Prog + Rok, family = binomial,
    subset = (JeSS == "ano" & Prijeti != "79"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2593  -1.0793   0.5957   0.8922   1.8169

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.351643   0.618464   2.185 0.028853 *
zcel         0.014082   0.003156   4.463  8.1e-06 ***
ssprum      -1.176784   0.145877  -8.067  7.2e-16 ***
ProgC       -0.575629   0.202989  -2.836 0.004572 **
ProgD        0.018438   0.291073   0.063 0.949491
ProgG        0.079780   0.269371   0.296 0.767098
ProgO        0.394755   0.396493   0.996 0.319437
ProgU       -0.845594   0.226211  -3.738 0.000185 ***
ProgZ       -0.332316   0.237673  -1.398 0.162052
Rok2004      0.015010   0.128199   0.117 0.906793

    Null deviance: 1740.1  on 1339  degrees of freedom
Residual deviance: 1543.5  on 1330  degrees of freedom
AIC: 1563.5

Number of Fisher Scoring iterations: 4
```

ROC křivka odpovídající modelu závislosti na uvedených třech prediktorech je na obrázku 2.4 (vlevo). Pro kreslení ROC křivek jsem použila modifikaci funkce `ROC()` z knihovny `Epi`, její znění je v příloze C.2.4.

Přehled klasifikací na základě odhadu pravděpodobnosti úspěchu je v následující tabulce. Při volbě mezní hodnoty $t_0 = 0{,}681$ dostaneme největší součet senzitivity a



Obrázek 2.4: ROC křivka modelu s prediktory *zcel*, *ssprum* a *Prog* (vlevo) a modelu obsahujícího navíc prediktory *Pohlavi*, *Maturdrive* (vpravo).

specificity. Odpovídají hodnota kappa koeficientu znamená, že rozhodováním základě modelu jsme odstranili 31 % nesprávných klasifikací v porovnání s náhodným rozhodováním ([8]). Pro odhad kappa koeficientu jsem použila funkci `Kappa()` z knihovny `vcd`.

Počet správně klasifikovaných studentů je nejvyšší (a to 951) celkem pro pět různých hodnot t_0. V tabulce jsou zjitěné četnosti pro $t_0 = 0{,}503$, kdy je zřejmá velmi nízká hodnota specificity. Ve skupině neúspěšných bychom více než polovinu studentů klasifikovali nesprávně.

Podobně jako můžeme zjistit takové t_0, kde je maximální hodnota Acc, je možné určit i t_0, kdy bude kappa koeficient největší. U tohoto modelu je maximální $\widehat{\kappa} = 0{,}341$ pro $t_0 = 0{,}552$ (viz skripty v příloze).

$t_0 = 0{,}681$				$t_0 = 0{,}503$			
	skutečnost				skutečnost		
predikce	neúspěch	úspěch	celkem	predikce	neúspěch	úspěch	celkem
neúspěch	351	346	697	neúspěch	201	117	318
úspěch	122	521	643	úspěch	272	750	1022
celkem	473	867	1340	celkem	473	867	1340
$\mathsf{FPR} + \mathsf{FNR} = 0{,}258 + 0{,}399 = 0{,}657$ $\widehat{\mathsf{Sp}} + \widehat{\mathsf{Se}} = 0{,}742 + 0{,}601 = 1{,}343$ $\widehat{\mathsf{Acc}} = 872/1340 = 0{,}651$ $\widehat{\kappa} = 0{,}310$				$\mathsf{FPR} + \mathsf{FNR} = 0{,}575 + 0{,}135 = 0{,}710$ $\widehat{\mathsf{Sp}} + \widehat{\mathsf{Se}} = 0{,}425 + 0{,}865 = 1{,}290$ $\widehat{\mathsf{Acc}} = 951/1340 = 0{,}710$ $\widehat{\kappa} = 0{,}313$			

Pokusíme se zjistit, zda přidáním dalších prediktorů do modelu předpověď úspěšnosti zlepšíme. Faktory označující pohlaví studenta *Pohlavi* a dobu maturity *Matur-*

drive mají na předpověď úspěchu statisticky významný vliv, jejich zařazením dostaneme lepší model.

```
a2=glm(Abs~zcel+ssprum+Prog, family =binomial,
 subset=(JeSS=="ano"&Prijeti!="79"))
a3=glm(Abs~zcel+ssprum+Prog+Maturdrive+Pohlavi,
 family =binomial, subset=(JeSS=="ano"&Prijeti!="79"))

anova(a2,a3,test="Chisq")
Analysis of Deviance Table

Model 1: Abs2 ~ zcel2 + ssprum + Prog3
Model 2: Abs2 ~ zcel2 + ssprum + Prog3 + Maturdrive + Pohlavi
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      1331    1543.48
2      1329    1486.19    2    57.29 3.632e-13

summary(a3)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.552597   0.666979  -0.829 0.407384
zcel          0.017081   0.003217   5.310 1.10e-07 ***
ssprum       -0.926668   0.152518  -6.076 1.23e-09 ***
ProgC        -0.503073   0.207439  -2.425 0.015302 *
ProgD         0.251485   0.301338   0.835 0.403965
ProgG         0.317821   0.277171   1.147 0.251521
ProgO         0.668137   0.409874   1.630 0.103079
ProgU        -0.834081   0.229594  -3.633 0.000280 ***
ProgZ        -0.184524   0.242269  -0.762 0.446269
Maturdrivene  1.008691   0.147776   6.826 8.74e-12 ***
Pohlavi2      0.343589   0.135949   2.527 0.011493 *


    Null deviance: 1740.1  on 1339  degrees of freedom
Residual deviance: 1486.2  on 1329  degrees of freedom
AIC: 1508.2
```

Odhadnuté koeficienty jsou v souladu s představou, že úspěšnost by měla růst v závislosti na větším počtu bodů u přijímacích zkoušek a lepším prospěchu na střední škole. U studentů, kteří maturovali dříve než v roce přijímacích zkoušek, kladné znaménko koeficientu odpovídá vyšší pravděpodobnosti úspěchu u „čerstvých“ maturantů, podobně jako u dívek v porovnání s chlapci.

ROC křivka tohoto modelu s pěti regresory je na obrázku 2.4 (vpravo). Přidáním dalších dvou regresorů do modelu jsme zvýšili plochu pod ROC křivkou na 75 %. Tedy pravděpodobnost, že náhodně vybraný úspěšný student bude mít v tomto modelu vyšší odhad pravděpodobnosti úspěchu než náhodně zvolený neúspěšný student, je 0,75. Vzroste i hodnota kappa koeficientu, při klasifikaci ($t_0 = 0{,}619$) s minimálním součtem pravděpodobností obou chyb vychází $\widehat{\kappa} = 0{,}359$, odpovídající četnosti jsou v levé části níže uvedené tabulky. Maximální počet správných rozhodnutí (961) nastane pro čtyři

různé hodnoty t_0, pro $t_0 = 0{,}418$ v pravé části tabulky je vidět opět velmi nízká schopnost správného rozlišení ve skutečnosti neúspěšných studentů, kdy je méně než jedna třetina klasifikována správně. Maximální hodnotu kappa koeficientu $\widehat{\kappa} = 0{,}364$ bychom dosáhli při volbě $t_0 = 0{,}592$.

$t_0 = 0{,}619$				$t_0 = 0{,}418$			
	skutečnost				skutečnost		
predikce	neúspěch	úspěch	celkem	predikce	neúspěch	úspěch	celkem
neúspěch	308	242	550	neúspěch	149	55	204
úspěch	165	625	790	úspěch	324	812	1136
celkem	473	867	1340	celkem	473	867	1340
$\mathsf{FPR} + \mathsf{FNR} = 0{,}349 + 0{,}279 = 0{,}628$				$\mathsf{FPR} + \mathsf{FNR} = 0{,}687 + 0{,}063 = 0{,}750$			
$\widehat{\mathsf{Sp}} + \widehat{\mathsf{Se}} = 0{,}651 + 0{,}721 = 1{,}372$				$\widehat{\mathsf{Sp}} + \widehat{\mathsf{Se}} = 0{,}313 + 0{,}937 = 1{,}250$			
$\widehat{\mathsf{Acc}} = 933/1340 = 0{,}696$				$\widehat{\mathsf{Acc}} = 961/1340 = 0{,}717$			
$\widehat{\kappa} = 0{,}359$				$\widehat{\kappa} = 0{,}289$			

Také poměrně nízké hodnoty koeficietů determinace $R^2_L = 0{,}146$, $R^2_N = 0{,}237$ ukazují na to, že model logistické regrese není moc přiléhavý. Sice jsme prokázali statistickou významnost regresorů, ale ani pomocí modelu s pěti regresory nedostaneme příliš kvalitní předpověď.

2.3.2 Úspěšné absolvování 1. ročníku

Za kritérium úspěšnosti můžeme zvolit i splnění podmínek pro zápis do 2. ročníku. Podmínky 1. ročníku splnilo 1126 studentů a nesplnilo 214 studentů.

V modelu logistické regrese prokážeme statistickou významnost všech tří regresorů *ssprum*, *zcel* a *Prog*.

```
library(car)
Anova(a)
Anova Table (Type II tests)
Response: Splnil1
       LR Chisq Df Pr(>Chisq)
zcel    13.2476  1  0.0002729 ***
ssprum   9.1233  1  0.0025237 **
Prog    23.3005  6  0.0007018 ***
```

Při zařazení dalších možných regresorů *Pohlavi*, *Stat* nebo *Rok* nedostaneme lepší model. Jiná situace nastane u prediktoru *Maturdrive*:

```
anova(a,a2,test="Chisq")
Analysis of Deviance Table
Model 1: Splnil1 ~ zcel + ssprum + Prog
Model 2: Splnil1 ~ zcel + ssprum + Prog + Maturdrive
  Resid. Df Resid. Dev   Df Deviance P(>|Chi|)
1      1331    1097.05
2      1330    1073.42    1    23.62 1.171e-06
```

Pro odhad pravděpodobnosti úspěchu použijeme tedy model se čtyřmi regresory *zcel*, *ssprum*, *Prog* a *Maturdrive*.

```
a2=glm(Splnil1~zcel+ssprum+Prog+Maturdrive, family =binomial,
 subset=(JeSS=="ano"&Prijeti!="79"))
summary(a2)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7030   0.2953   0.4739   0.6420   1.1966

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.477151   0.802077   0.595 0.551914
zcel          0.015491   0.003996   3.876 0.000106 ***
ssprum       -0.349582   0.176659  -1.979 0.047832 *
ProgC        -1.008431   0.306866  -3.286 0.001015 **
ProgD        -0.420240   0.407537  -1.031 0.302461
ProgG        -0.528318   0.374803  -1.410 0.158661
ProgO         0.136944   0.592106   0.231 0.817096
ProgU        -1.139982   0.325537  -3.502 0.000462 ***
ProgZ        -0.521803   0.347765  -1.500 0.133499
Maturdrivene  0.838030   0.169569   4.942 7.73e-07 ***
```

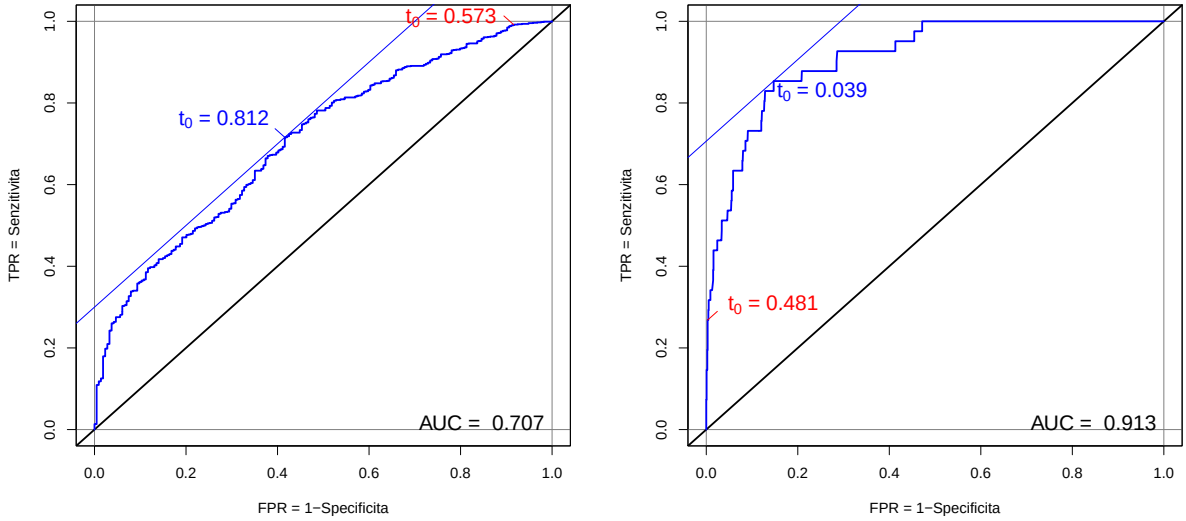
Znaménka regresních koeficientů jsou stejná jako u předpovědi úspěšnosti celého studia. Pravděpodobnost úspěchu roste s vyšším počtem bodů u přijímacích zkoušek, lepším prospěchem na střední škole a je vyšší u maturujících v roce přijímacích zkoušek.

Odpovídající ROC křivka je na obrázku 2.5, kde je vyznačena hodnota $AUC = 0{,}707$. Z níže uvedené tabulky je zcela zřejmé, že když jsou skupiny úspěšných a neúspěšných velmi početně nevyvážené jako v tomto případě, hledisko nejvyššího počtu správně klasifikovaných objektů Acc může být zavádějící.

Hodnota kappa koeficientu je výrazně horší než u předpovědi absolvování celého studia. Maximální kappa koeficient $\widehat{\kappa} = 0{,}235$ bychom dostali při zvolení $t_0 = 0{,}780$.

Nízké hodnoty koeficietů determinace $R^2_L = 0{,}088$, $R^2_N = 0{,}127$ ukazují, že model skutečnost vysvětluje nedostatečně.

$t_0 = 0{,}812$				$t_0 = 0{,}573$			
	skutečnost				skutečnost		
predikce	neúspěch	úspěch	celkem	predikce	neúspěch	úspěch	celkem
neúspěch	125	320	445	neúspěch	18	9	27
úspěch	89	806	895	úspěch	196	1117	1313
celkem	214	1126	1340	celkem	214	1126	1340
$\mathsf{FPR} + \mathsf{FNR} = 0{,}416 + 0{,}284 = 0{,}700$				$\mathsf{FPR} + \mathsf{FNR} = 0{,}916 + 0{,}008 = 0{,}924$			
$\widehat{\mathsf{Sp}} + \widehat{\mathsf{Se}} = 0{,}584 + 0{,}716 = 1{,}300$				$\widehat{\mathsf{Sp}} + \widehat{\mathsf{Se}} = 0{,}084 + 0{,}992 = 1{,}076$			
$\widehat{\mathsf{Acc}} = 931/1340 = 0{,}695$				$\widehat{\mathsf{Acc}} = 1135/1340 = 0{,}847$			
$\widehat{\kappa} = 0{,}209$				$\widehat{\kappa} = 0{,}118$			



Obrázek 2.5: ROC křivka modelu splnění podmínek 1. ročníku s prediktory *zcel*, *ssprum*, *Prog* a *Maturdrive* (vlevo) a modelu absolvování studia s vyznamenáním s prediktory *ssprum*, *Prog* a *Prijeti* (vpravo).

2.3.3 Absolvování studia s vyznamenáním

Jiným kritériem úspěšnosti může být absolvování studia s vyznamenáním. Z celkového počtu 1416 studentů absolvovalo s vyznamenáním pouze 45 studentů. U 41 z nich známe prospěch na střední škole, 14 absolventů s vyznamenáním bylo přijato bez přijímacích zkoušek (8 chlapců a 6 dívek), 27 na základě přijímacích zkoušek (10 chlapců a 17 dívek). Do výpočtu zahrneme záznamy o 1372 studentech, u kterých známe předchozí středoškolský prospěch.

Skutečnosti, že značná část absolventů s vyznamenáním byla přijata bez přijímacích zkoušek, odpovídá i významnost faktoru *Prijeti* v modelu, kde dalšími regresory jsou *ssprum* a *Prog*. ROC křivka tohoto modelu je na obrázku 2.5 vpravo.

V následující tabulce jsou uvedeny četnosti odpovídající zvoleným mezním hodnotám pravděpodobnosti pro maximalizaci součtu senzitivity a specificity (vlevo) a maximalizaci Acc, kdy je součet pravděpodobností obou chyb mnohem vyšší. Maximální kappa koeficient $\widehat{\kappa} = 0{,}433$ dostaneme pro $t_0 = 0{,}179$. Z uvedeného je zřejmé, že při maximalizaci různých kritérií hodnocení klasifikace, dostaneme odlišné výsledky.

Hodnoty koeficietů determinace $R^2_L = 0{,}352$, $R^2_N = 0{,}383$ vycházejí vyšší než u předchozích modelů, což může odpovídat i vysoké hodnotě $AUC = 0{,}913$.

$t_0 = 0{,}039$				$t_0 = 0{,}481$			
	skutečnost				skutečnost		
predikce	neúspěch	úspěch	celkem	predikce	neúspěch	úspěch	celkem
neúspěch	1135	6	1141	neúspěch	1327	30	1357
úspěch	196	35	231	úspěch	4	11	15
celkem	1331	41	1372	celkem	1331	41	1372
$\mathsf{FPR} + \mathsf{FNR} = 0{,}146 + 0{,}147 = 0{,}293$				$\mathsf{FPR} + \mathsf{FNR} = 0{,}732 + 0{,}003 = 0{,}735$			
$\widehat{\mathsf{Sp}} + \widehat{\mathsf{Se}} = 0{,}854 + 0{,}853 = 1{,}707$				$\widehat{\mathsf{Sp}} + \widehat{\mathsf{Se}} = 0{,}268 + 0{,}997 = 1{,}265$			
$\widehat{\mathsf{Acc}} = 1170/1372 = 0{,}853$				$\widehat{\mathsf{Acc}} = 1338/1372 = 0{,}975$			
$\widehat{\kappa} = 0{,}218$				$\widehat{\kappa} = 0{,}383$			

2.3.4 Ordinální regrese

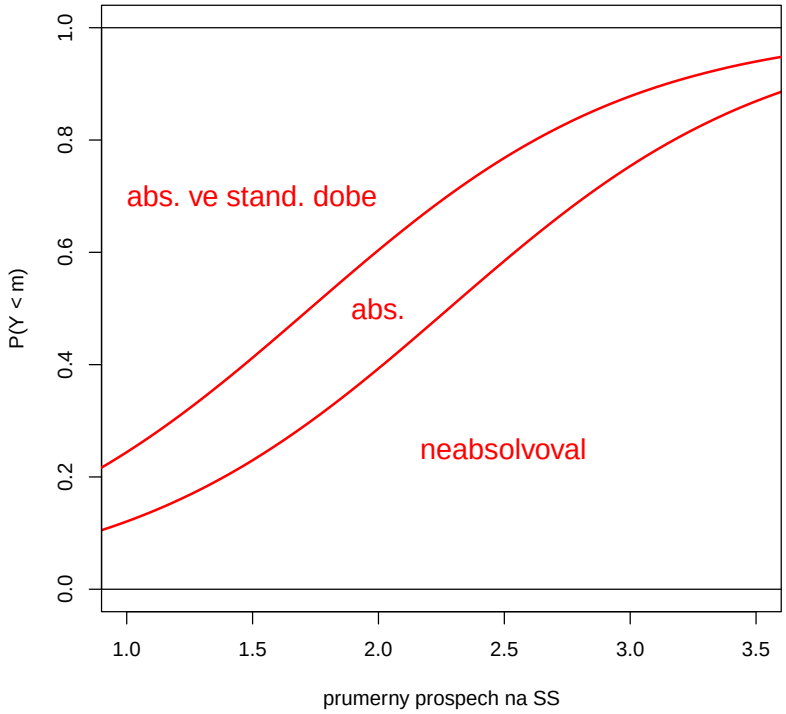
Využijeme informace, zda student úspěšně dokočil studium ve standardní době bakalářského studia 3 roky nebo později. U studentů, kteří konali přijímací zkoušky a je znám jejich prospěch na střední škole budeme předpovídat úspěšnost v podrobnějším ordinálním měřítku na základě ordinálního modelu s jedním regresorem *ssprum*.

Predikcí pomocí funkce `polr` z knihovny `MASS` dostaneme model, jehož kontingenční tabulka četností skutečných hodnot a předpovědí podle modelu následuje:

predikce	skutečnost			
	neúspěch	úspěch (ne ve std. době)	úspěch (ve std. době)	celkem
neúspěch	243	101	115	459
úspěch (ne ve std. době)	0	0	0	0
úspěch (ve std. době)	230	148	503	881
celkem	473	249	618	1340

Z obrázku 2.6 regresních křivek je zřejmé, že pro žádnou hodnotu středoškolského prospěchu není pravděpodobnost zařazení do prostřední kategorie největší a odpovídající řádek tabulky obsahuje nulové četnosti. I s ohledem na nízkou hodnotu kappa koeficientu $\widehat{\kappa} = 0{,}230$ (s lineárními vahami $\widehat{\kappa} = 0{,}275$) je patrné, že předpověď využitím tohoto modelu nedává příliš dobré výsledky.

Přidáním dalších regresorů *zcel* a *Maturdrive* dostaneme průkazně lepší model, kde není zamítnut předpoklad rovnoběžnosti regresních křivek. V níže uvedené tabulce se přesvědčíme, že ani v tomto případě není žádný student klasifikován jako absolvent, ale nikoliv ve standardní době.



Obrázek 2.6: Regresní křivky modelu ordinální regrese závislosti doby absolvování studia na středoškolském prospěchu

predikce	skutečnost			
	neúspěch	úspěch (ne ve std. době)	úspěch (ve std. době)	celkem
neúspěch	274	109	119	502
úspěch (ne ve std. době)	0	0	0	0
úspěch (ve std. době)	199	140	499	838
celkem	473	249	618	1340

Hodnota kappa koeficientu je o něco vyšší $\widehat{\kappa} = 0{,}270$ (s lineárními vahami $\widehat{\kappa} = 0{,}321$) než u modelu s jediným regresorem.

Faktor *Prog* by byl po přidání do modelu shledán jako statisticky významný, ale vzhledem k porušení předpokladu rovnoběžnosti regresních křivek by bylo nutno použít jiný model ordinální regrese.

Tabulka 2.3: Korelační koeficienty mezi prospěchem na vysoké škole, počtem bodů z přijímacích zkoušek a předchozím prospěchem na střední škole u studentů PřF UK

prospěch na vysoké škole	přijímací zkoušky	prospěch na střední škole
1. ročník VŠ	-0,503	0,478
2. ročník VŠ	-0,515	0,436
3. ročník VŠ	-0,449	0,477
VŠ celkem	-0,544	0,452

2.4 Porovnání s dalšími studiemi

2.4.1 Čeští autoři

Škaloudová [33] ve své disertační práci analyzovala data o studentech učitelství, kteří se zapsali ke studiu na Pedagogické fakultě UK v roce 1996/97 a 1997/98. Na základě známého středoškolského prospěchu, výsledků přijímacího řízení, speciálního dotazníku pro studenty učitelství a inteligenčního testu T-S-I se snažila předpovědět úspěšnost studia na VŠ. Jako kritérium stanovila koeficient studijní úspěšnosti, který v sobě zahrnoval jak průměrný prospěch na VŠ, tak počet zkoušek a opravných termínů. Škaloudová zjistila, že středoškolské prediktory přispívají k rozlišení skupiny úspěšných a neúspěšných studentů stejně dobře jako přijímací zkoušky, u některých oborů studia dokonce ještě lépe. V průběhu studia se vliv prospěchu na střední škole snižuje pomaleji než vliv výsledku přijímacího řízení. U inteligenčního testu se neprokázal jeho vliv na úspěšnost studia.

Tabulka 2.3 ukazuje poměrně vyrovnané korelační koeficienty u 961 studentů PřF UK, kteří konali přijímací zkoušky, je znám jejich prospěch na SŠ i průměrný prospěch v jednotlivých ročnících VŠ. Je možné, že v případě pětiletého magisterského studia by se projevil snižující se vliv přijímacích zkoušek prospěch na VŠ, jak naznačuje nižší hodnota korelačního koeficientu u 3. ročníku.

V článku [37] jsou shrnuty výsledky analýzy dat získaných na souboru uchazečů přijatých ke studiu na MFF UK v letech 1993–1997. Pomocí logistické regrese bylo hodnoceno, zda studenti úspěšně absolvovali první dva roky studia na fakultě. Bylo prokázáno, že ve studijních programech fyzika (F), matematika (M) a informatika (I) byla významným faktorem skutečnost, zda byl student přijat bez přijímacích zkoušek (na základě velmi dobrého prospěchu nebo účasti v celostátním kole olympiády), ve studijním programu učitelství (U) a F byl prokázán vliv středoškolského prospěchu a ve studijním programu I a M úspěšnost ovlivnil i počet bodů dosažených v přijímacím řízení, jestliže student přijímací zkoušky konal. Nebyl shledán rozdíl mezi chlapci a dívkami.

V případě předpovídání úspěšného ukončení studia (logistickou regresí) byly na rozdíl od MFF na PřF prokázány jako významné faktory kromě středoškolského prospěchu a výsledků přijímacích zkoušek též pohlaví a skutečnost, zda student maturoval dříve než v roce přijímacích zkoušek. Zřejmě se takto projevily velké rozdíly, které jsou na PřF v podílu těch, kteří ve studiu vytrvali a úspěšně ho absolvovali (muži 56 %, ženy 70 %; maturující dříve 45 %, v roce zkoušek 71 %). Výpočet prokázal významný vliv

příslušnosti studenta ke skupině oborů, nebyl shledán rozdíl mezi studenty s ohledem na začátek studia. Na základě uvedených pěti prediktorů bylo dosaženo sedmdesátiprocentního podílu správných předpovědí při celkovém počtu 1340 studentů, kteří konali přijímací zkoušky a byl znám jejich prospěch na SŠ.

Höschl a Kožený [13] analyzovali údaje o studentech 3. lékařské fakulty UK, kteří byli přijati ke studiu v roce 1992/93 a 1993/94 a dokončili šestý semestr studia. Při stanovení celkového průměru na VŠ za šest semestrů jako kritéria prokázali statisticky významný vliv výsledku přijímací zkoušky z fyziky, středoškolských známek z fyziky, motivace ke studiu medicíny hodnocené komisí a osobnostního dotazníku. Závislostí na těchto prediktorech se podařilo vysvětlit 32 % kolísání dosaženého známkového průměru. Na PřF UK bylo v modelu lineární regrese se dvěma prediktory (celkový počet bodů u přijímacích zkoušek a průměrný prospěchu na střední škole) dosaženo vyššího koeficentu determinace, a to $R^2 = 41\,\%$.

Studie [32] analyzuje data týkající se 673 studentů, kteří započali magisterské studium všeobecného lékařství nebo stomatologie na 1. lékařské fakultě UK v letech 1994–1997. Autoři použili logistickou regresi pro předpověď úspěšného ukončení studia. Model s nejnižší hodnotou Akaikeho informačního kritéria AIC zahrnoval pět regresorů (počet bodů z přijímací zkoušky z fyziky a z chemie, průměrná známka na střední škole z biologie a chemie a rozdíl mezi prospěchem v prvním a posledním ročníku střední školy). Plocha pod ROC křivku u tohoto modelu vyšla $AUC = 0{,}72$. Ačkoliv v přijímacím řízení nebyl zohledněn prospěch na střední škole, tyto výsledky u přijatých studentů lépe predikovali úspěšnost ($AUC = 0{,}66$) než přijímací zkoušky ($AUC = 0{,}64$). Dosažené hodnoty jsou srovnatelné s PřF UK, kdy pro model předpovědi úspěchu pomocí logistické regrese na základě pěti obdobných prediktorů byla zjištěna hodnota $AUC = 0{,}75$. Podobně i na PřF UK nebyl v přijímacím řízení vzat v úvahu prospěch na střední škole.

Do výzkumu predikční validity testu Obecných studijních předpokladů (OSP) společnosti Scio, s.r.o. [31] byly zahrnuty údaje Fakulty sociálních studií MU Brno, Fakulty zdravotně sociální Ostravské univerzity a VŠCHT v Praze. Bylo zjištěno, že korelace testu OSP s průměrem známek na VŠ se postupně snižuje až na nevýznamnou souvislost se závěrečným prospěchem. Lépe než jednotlivé oddíly testu OSP (verbální, analytický a kvantitativní) koreluje celkový výsledek testu OSP.

V analýze [31] je sledovaná VŠCHT svým přírodovědným zaměřením PřF UK nejpodobnější. Velikosti datových souborů jsou též porovnatelné. V tabulce 2.4 jsou uvedeny korelační koeficienty s dosaženým prospěchovým průměrem v 1. ročníku studia VŠ. Vzhledem k tomu, že ze studie není zřejmé, zda jsou u VŠCHT započteni všichni studenti s alespoň jednou známkou nebo pouze ti, kteří splnili podmínky pro zápis do vyššího ročníku, uvádím u PřF obě hodnoty.

Porovnáním korelačních koeficientů s využitím testové statistiky z (1.8) prokážeme, že testy z předmětů na PřF lépe predikují očekávaný průměr v 1. ročníku než na VŠCHT testy OSP. Pro slabou korelaci průměru na VŠ s prospěchem na SŠ v případě VŠCHT jsou jako možná příčina uvedeny rozdíly ve způsobech hodnocení na jednotlivých školách. Toto by však projevilo i ve větším vzorku z PřF. Důvodem je spíše fakt, že větší rozdíly jsou mezi jednotlivými typy středních škol, kde na PřF studuje naprostá většina absolventů gymnázií a pouze malé procento tvoří absolventi středních odborných škol.

Tabulka 2.4: Korelační koeficienty mezi prospěchem v 1. ročníku vysoké školy, počtem bodů z přijímacích zkoušek a předchozím prospěchem na střední škole u studentů VŠCHT a PřF UK

prospěch v 1. r. VŠ	počet studentů	prospěch na SŠ	přijímací zkoušky
VŠCHT	410		-0,381
VŠCHT	357	0,075	
PřF UK - 2003/04	622	0,507	-0,556
Splnili 1. r.	570	0,518	-0,545
PřF UK - 2004/05	621	0,525	-0,509
Splnili 1. r.	556	0,498	-0,479

Z názvů některých příspěvků a abstraktů by se dalo soudit, že se zabývají predikční validitou. Podrobnějším studiem se někdy zjistí, že zřejmě došlo k nedorozumnění, jako například v [7]. Autoři z Masarykovy univerzity v Brně se snaží například logistickou regresí odhadovat úspěšnost absolvování magisterského studia na základě průměrného prospěchu na vysoké škole. Ve svém přípěvku spíše porovnávají známé ukazatele u skupiny úspěšných a neúspěšných studentů, než že by se zabývali predikcí.

2.4.2 Zahraniční autoři

V USA mají standardizované přijímací testy na vysoké školy dlouhou tradici. Předchůdce současných testů studijních předpokladů byl poprvé použit pro osm tisíc uchazečů již v roce 1926 [39].

Program SAT je sponzorován neziskovou organizací College Board. Každoroční testování milionů uchazečů je organizováno Educational Testing Service (ETC), v současnosti největší organizací USA zabývající se testováním, která byla založena již v roce 1947. Do roku 2005 byly používány dva druhy testů. SAT I: Reasoning Test měl část matematickou a verbální a SAT II: Subject Tests, kde byly hodnoceny dosažené znalosti v konkrétních oborech. Predikční validita testů SAT I byla studována mnohem více než SAT II.

Kalifornská univerzita (UC), sdružující osm kampusů (US Berkeley, UC Davis, US Irvine, UC Los Angeles, UC Riverside, UC San Diego, UC Santa Barbara a UC Santa Cruz), vyžadovala od uchazečů jak testy SAT I (nebo ACT), tak i SAT II. Ve studii [12] je porovnáván vliv SAT I, SAT II a prospěchu na střední škole na predikci průměrného prospěchu v 1. ročníku studia vysoké školy. Analýzou údajů o téměř 80 000 studentech (nováčcích) zapsaných ke studiu v letech 1996–1999 bylo zjištěno, že při znalosti výsledků testů SAT II, je přínos informace o SAT I minimální. V tabulce 2.5 jsou zjištěné koeficienty determinace jednotlivých modelů porovnány s údaji PřF UK, kam jsou však zařazeni nejen studenti poprvé zapsaní ke studiu vysoké školy (v důsledku obtížného zjištění této informace).

Od roku 2006 je obsah testů upraven. SAT Reasoning Test nově obsahuje kromě částí SAT-CR (critical reading), SAT-M (mathematics) i SAT-W (writing). Predikční

Tabulka 2.5: Porovnání koeficientů determinace při předpovědi prospěchu v 1. ročníku VŠ u vybraných modelů

prediktory	**koeficient determinace** R^2				
University of California (údaje z [12])					
	1996	1997	1998	1999	1996–1999
prospěch na SŠ (HSGPA)	0,170	0,167	0,147	0,129	0,154
SAT I	0,138	0,108	0,122	0,142	0,133
SAT II	0,164	0,144	0,156	0,164	0,160
SAT I + SAT II	0,167	0,144	0,156	0,168	0,162
HSGPA + SAT I	0,219	0,201	0,192	0,204	0,208
HSGPA + SAT II	0,230	0,217	0,211	0,215	0,222
HSGPA + SAT I + SAT II	0,232	0,217	0,211	0,219	0,223
PřF UK (1243 studentů s alespoň jednou známkou)					
	2003	2004			2003–2004
prospěch na SŠ	0,257	0,276			0,265
přijímací zkoušky	0,309	0,259			0,274
prospěch na SŠ + přijímací zkoušky	0,397	0,389			0,386
PřF UK (1126 studentů, kteří splnili podmínky 1. ročníku)					
prospěch na SŠ	0,268	0,248			0,258
přijímací zkoušky	0,297	0,230			0,254
prospěch na SŠ + přijímací zkoušky	0,394	0,352			0,367

Tabulka 2.6: Porovnání korelačních koeficientů, resp. koeficientů mnohonásobné korelace průměrného prospěchu v 1. ročníku VŠ a uvedených prediktorů

prediktory	*korelační koeficient (koef. mnohonásobé korelace)*
data z [16] (údaje ze 110 univerzit USA)	
prospěch na SŠ (HSGPA)	0,36
SAT-M	0,26
SAT-CR	0,20
SAT-W	0,33
SAT-M + SAT-CR + SAT-W	0,35
HSGPA + SAT-M + SAT-CR + SAT-W	0,46
PřF UK (1243 studentů s alespoň jednou známkou)	
prospěch na SŠ	0,515
přijímací zkoušky	0,523
prospěch na SŠ + přijímací zkoušky	0,621
PřF UK (1126 studentů, kteří splnili podmínky 1. ročníku)	
prospěch na SŠ	0,504
přijímací zkoušky	0,508
prospěch na SŠ + přijímací zkoušky	0,606

validita upravených testů SAT je analyzována například v [16] za základě údajů týkajících se více než 150 tisíců studentů ze 110 amerických univerzit, kteří poprvé studovali 1. ročník vysoké školy v akademickém roce 2006/07. V USA je prospěch v 1. ročníku u poprvé zapsaných studentů nejčastěji používaným kritériem hodnocení úspěšnosti. Kurzy, které mají studenti v 1. ročníku, jsou si více podobné než v dalších letech studia a průměr v 1. ročníku vysoce koreluje s celkovým dosaženým průměrem. Důležitým faktorem je i skutečnost, že data pro další analýzy jsou dostupná již po prvním roce studia. Zjištěné korelační koeficienty (resp. koeficienty mnohonásobné korelace) jsou porovnány s výsledky analýzy dat PřF UK v tabulce 2.6.

V meta-analýze [18] jsou shrnuty údaje z velkého množství nezávislých studií predikční validity testů GRE (Graduate Record Examination). General test měří verbální, kvatitativní a analytické schopnosti, zatímco Subject Tests měří dosažené znalosti v konkrétní oblasti (biochemie a molekulární biologie, biologie, chemie, informatika, literatura, matematika, fyzika a psychologie).

Pro predikci jak konečného prospěchu, tak prospěchu v 1. ročníku byly Subject tests shledány prokazatelně lepšími prediktory než jednotlivé složky testu GRE nebo předchozí průměrný prospěch. Zjištěné korelační koeficienty jsou velmi blízké hodnotám dosaženým na PřF UK.

Analýza predikční validity testů ACT [40] prokázala podobný vliv testů ACT a prospěchu na střední škole na předpověď průměru známek v 1. ročníku VŠ, na roz-

díl od předpovědi konečného průměrného prospěchu, kde má větší váhu středoškolský prospěch.

ROC analýza byla použita pro předpověď úspěšného začátku studia na Universitě v Udine (Itálie) u více než tří tisíc poprvé zapsaných studentů v letech 1992/93 až 1997/98 ([25]). V modelu, kdy byla za kritérium úspěšnosti zvolena skutečnost, že student absolvoval minimálně čtyři zkoušky v průběhu prvního akademického roku, byl prokázán statisticky významný vliv známek na střední škole, typu střední školy (studenti specializovaných středních škol byli úspěšnější), kategorie pracovního zařazení otce a též byly shledány rozdíly mezi fakultami univerzity. Bylo dosaženo plochy pod ROC křivkou $AUC = 0{,}793$.

Závěr

V dnešní době, kdy začíná studovat vysokou školu více než polovina mladých lidí, se problematika přijímacího řízení dostává do popředí zájmu nejen v České republice, ale prakticky ve všech hospodářsky vyspělých zemích.

Situace v České republice je komplikována možností podávání libovolného počtu přihlášek ke studiu na vysoké škole, kdy v posledních letech podávají uchazeči v průměru více než dvě přihlášky. Mnozí jsou přijati vícekrát a teprve po ukončení přijímacího řízení se rozhodují, kterou školu či obor budou studovat. Vysoké školy mají tedy pouze přibližnou představu o počtu přijatých uchazečů, kteří se ke studiu skutečně zapíší. Snahou je stanovení takových podmínek přijetí ke studiu, aby byli vybráni uchazeči s dobrými předpoklady pro úspěšné absolvování studia.

Metodika hodnocení predikční validity není jednoduchá. Ač je toto téma zmiňováno na mnoha konferencích, neznám dostatečně podrobný materiál, který by mohl sloužit těm, kteří nemají mnoho zkušeností s aplikacemi matematické statistiky. Ve své práci jsem se proto pokusila o ucelený přehled statistických metod a modelů, které je možno pro hodnocení predikce úspěšnosti studia použít, ať jsou kritériem úspěšnosti stanoveny číselné nebo kvalitativní veličiny.

Na teoretickou část navazuje aplikace vysvětlených metod na konkrétní data z Přírodovědecké fakulty Univerzity Karlovy. Pro všechny výpočty jsem zvolila volně šiřitelný programu R, dostupný na: `http://cran.at.r-project.org/`. Vždy zdůvodňuji volbu daného modelu i omezení na data, která je do něj možno zahrnout. Uvedené výstupy z programu jsou doplněny podrobným vysvětlením, abych umožnila potenciálnímu zájemci snadnější vniknutí do problematiky. Pouze mechanické používání procedur jakéhokoli programu může vést k nesprávným závěrům.

Analýzou dat o studentech PřF UK jsem zjistila, že výsledky přijímacích zkoušek spolu s předchozím prospěchem na střední škole poměrně dobře predikují číselná kritéria úspěšnosti (např. prospěchový průměr). Ve srovnání s ostatními studiemi je dosaženo vyšších podílů variability, kterou je možno vysvětlit závislostí na uvedených prediktorech. Do rozsáhlých studií z USA jsou však zahrnuty údaje z vysokých škol s nejrůznějším zaměřením a PřF UK je z tohoto hlediska spíše homogenní skupinou.

Předpověď kvalitativních kritérií úspěšnosti dává na PřF i jinde méně uspokojivé výsledky. Domnívám se, že to souvisí se skutečností, že je velmi obtížné měřit například motivaci ke studiu. Ta je významným faktorem při rozhodování studentů, zda studium splňuje jejich představy a budou se tedy snažit ho úspěšně dokončit.

Porovnání se zahraničními studiemi není rozhodně vyčerpávající, spíše jsem chtěla naznačit, že podobné problémy jsou řešeny v mnoha částech světa. Zavádění změn v přijímacím řízení je vždy obtížné, protože jejich důsledky je možno analyzovat až se značným časovým odstupem. Mezitím však může dojít k dalším změnám ve společnosti. Například v České repulice se v důsledku nepříznivého demografického vývoje zmenšuje

základna potenciálních uchazečů o studium na vysokých školách, která je jen částečně doplňována uchazeči ze zahraničí (zejména ze Slovenska).

Prokázaná souvislost úspěšnosti studia na vysoké škole s předchozím prospěchem na střední škole může být pomůckou při změně koncepce přijímacího řízení. U oborů, kde zájem uchazečů výrazně nepřevyšuje kapacitu, by mohl být dobrý prospěch na střední škole a úspěšně složená maturita jedinou podmínkou pro přijetí ke studiu. Tento postup je již v současné době aplikován na některých technicky zaměřených vysokých školách.

Příloha A

Seznam a popis veličin v databázi údajů o studentech

název veličiny	vysvětlení
`ss1`	průměr na konci 1. ročníku střední školy
`ss2`	průměr na konci 2. ročníku střední školy
`ss3`	průměr na konci 3. ročníku střední školy
`ss4`	průměr v pololetí 4. ročníku střední školy
	(u víceletých gymnázií odpovídajícího ročníku)
`ssprum`	aritmetický průměr čtyř uvedených průměrů známek
`pr1`	průměrný prospěch za 1. ročník studia na VŠ
`prumer`	průměrný prospěch za celé studium VŠ
`biol`	počet bodů z přijímací zkoušky z biologie
`chemie`	počet bodů z přijímací zkoušky z chemie
`fyzika`	počet bodů z přijímací zkoušky z fyziky
`matem`	počet bodů z přijímací zkoušky z matematiky
`dejep`	počet bodů z přijímací zkoušky z historie
`zem`	počet bodů z přijímací zkoušky z geografie
`zcel`	součet bodů za přijímací zkoušky
	(v případě jediné zkoušky je počet bodů zdvojnásoben)
`prijeti`	číselné označení způsobu přijetí studenta
	10 – přijetí na základě přijímací zkoušky
	79 – bez přijímacích zkoušek
	85 – dodatečně
`Prijeti`	přijetí jako faktor
`pohlavi`	označení pohlaví studenta
	(*1* – žena, *2* – muž)
`Pohlavi`	pohlaví studenta jako faktor

název veličiny	vysvětlení
`Prog`	označení skupiny studijních programů
	B – Biologie, Speciální chemicko–biologické obory
	C – Chemie, Biochemie, Klinická a tox. analýza
	D – Demografie
	Z – Geografie
	G – Geologie
	O – Ekologie a ochrana prostředí
	U – všechny obory zaměřené na vzdělávání
`Maturdrive`	faktor označující, kdy student maturoval
	ano - maturoval dříve než v roce přijímacích zkoušek
	ne - maturoval v roce přijímacích zkoušek
`Stat`	státní příslušnost studenta (*203* – ČR, *703* – Slovensko)
`JeSS`	faktor označující, zda je znám prospěch na střední škole
	ano - je znám, *ne* – není znám
`rok`	rok zápisu na fakultu (*2003*, *2004*)
`Rok`	rok zápisu na fakultu jako faktor
`abs`	absolvování studia (*0* – ne, *1* – ano)
`Abs`	absolvování studia jako faktor
`abs3`	absolvování studia se třemi úrovněni
	0 - dosud neabsolvoval (zanechal, ještě studuje, má přerušeno apod.)
	1 - absolvoval, ale ne ve standardní době studia
	2 - absolvoval ve standardní době studia
`Abs3`	absolvování studia se třemi úrovněni jako faktor
`splnil1`	splnění podmínek 1. ročníku (*0* – ne, *1* – ano)
`Splnil1`	splnění podmínek 1. ročníku jako faktor
`vysl`	celkový výsledek studia (*4* – s vyznamenáním, *5* – absolvoval)
`Vysl`	výsledek studia jako faktor
`vyzn`	absolvování s vyznamenáním (*0* – ne, *1* – ano)
`Vyzn`	absolvování s vyznamenáním jako faktor

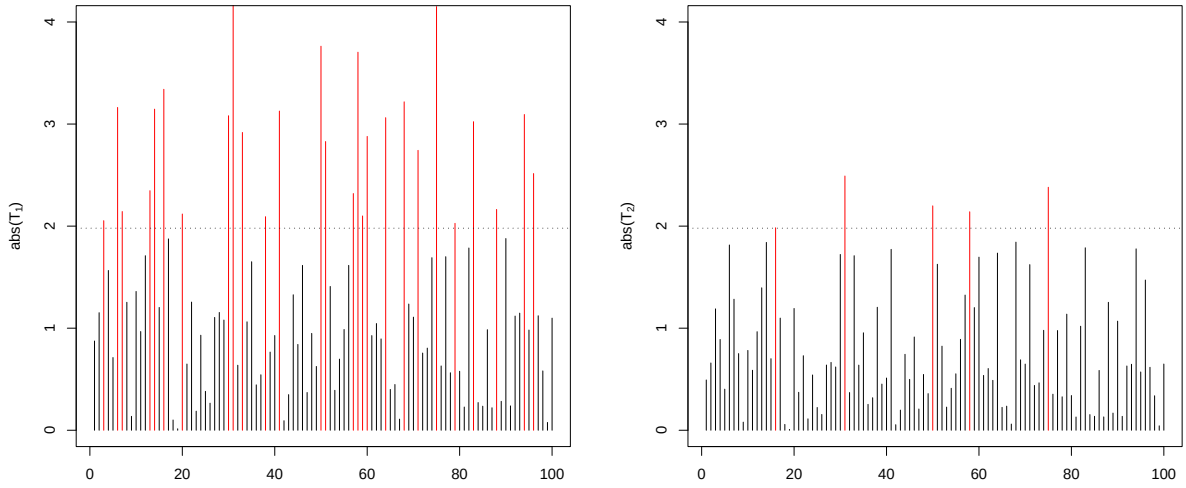
Příloha B

Simulace

B.1 Porovnání statistik T_1 a T_2

Simulujeme situaci popsanou v kap. 1.1.4, kdy je korelační matice veličin (Y, X_1, X_2) „téměř“ singulární. Předpokládáme nezávislost veličin X_1 a X_2 (obě s rozdělením $\mathsf{N}(0,1)$), veličinu Y položíme rovnu součtu $X_1 + X_2$ opraveným o náhodnou chybu (s rozdělením $\mathsf{N}(0,1/4)$).

Na obrázku B.1 jsou znázorněny absolutní hodnoty vypočtených statistik T_1 (vlevo) a T_2 (vpravo) určené v opakovaných výběrech o rozsahu 1000 hodnot. Korelační koeficienty r_{01} a r_{02} vycházejí přibližně 0,67, absolutní hodnota jejich rozdílu maximálně do 0,07. Červeně jsou zvýrazněny hodnoty statistik T_1, resp. T_2 překračující kritickou hodnotu $t_{n-3}(0{,}05)$, která vyznačena vodorovnou čarou.



Obrázek B.1: Porovnání absolutních hodnot statistik T_1 (vlevo) a T_2 (vpravo)

```
# Porovnani statistik T1 a T2
n=1000  # rozsah vyberu
B=100       # pocet opakovani
vysl = matrix(NA,B,2)
```

```
    # matice (Bx2), kam ukladam vypoctene statistiky
for (i in 1:B){
    # nageneruju X1 a X2
    x1=rnorm(n);x2=rnorm(n)
    # y je "skoro" soucet x1 a x2
    y=x1+x2+rnorm(n)/2
    # matice korelacnich koeficientu
    R=matrix(cor(cbind(y,x1,x2)),nrow = 3, ncol=3, byrow=TRUE)
    r01= R[1,2]
    r02= R[1,3]
    r12= R[2,3]
    T1=(r01-r02)*sqrt((n-3)*(1+r12)/2/det(R))
    T2=(r01-r02)*sqrt((n-1)*(1+r12)/(2*(n-1)/(n-3)*det(R)+
    ((r01+r02)/2)^2*(1-r12)^3))
    # ulozim vypoctene absolutni hodnoty do matice
    vysl[i,1]<-abs(T1)
    vysl[i,2]<-abs(T2)
    }
# kresleni T1
plot(vysl[,1],type="h", ylim=c(0,4),xlab="",
ylab=expression(paste("abs(",T[1],")")),
col=ifelse(vysl[,1]>qt(0.975,n-3),"red","black"))
abline(h=qt(0.975,n-3),col="black",lty=3)
points((1:100),vysl[,4]),pch=18,col="blue")
# kresleni T2
plot(vysl[,2],type="h", ylim=c(0,4),xlab="",
ylab=expression(paste("abs(",T[2],")")),
col=ifelse(vysl[,2]>qt(0.975,n-3),"red","black"))
abline(h=qt(0.975,n-3),col="black",lty=3)
```

B.2 Odhad korelačního koeficientu z neúplných dat

Simulujeme reálnou situaci, kdy hodnoty X_i známe u všech subjektů, ale hodnoty Y_j pouze u části, tedy máme k dispozici n úplných dvojic (X_i, Y_i) a u dalších $N-n$ subjektů neznáme hodnotu Y_i. Například u všech studentů, kteří se zúčastnili přijímacího řízení, známe počet získaných bodů, ale pouze u těch, kteří byli přijati a ke studiu se zapsali, je k dispozici jejich prospěch na vysoké škole.

Vygenerujeme data, která mají dvourozměrné normální rozdělení s předem zvoleným korelačním koeficientem. Podíl „chybějících“ pozorování veličiny Y je označen $p = (N - n)/N$. Dále budeme opakovaně (B je počet opakování) provádět následující kroky:

1. vybereme N dvojic,
2. zvolíme „chybějící“ pozorování,
3. spočítáme tři výběrové korelační koeficienty veličin X a Y z těchto údajů:

   `cely vyber` – ze všech vybraných dvojic rozsahu N,

   `uplne dvojice` – z úplných dvojic, kde část hodnot veličiny Y „chybí“,

`Cohenova oprava` – koef. z úplných dvojic upravený podle vzorce (1.14).

Pak určíme aritmetické průměry vypočtených korelačních koeficientů v B opakováních včetně směrodatných odchylek, které jsou součástí výpisu funkce `simuluj()`. Předchází jim zápis hodnot zjištěných v celé množině vygenerovaných dat, resp. zadaných jako parametry při volání funkce:

`rho` – korelační koeficient ze všech vygenerovaných dvojic,

`sd klas.` – odmocnina z asymptotického rozptylu korelačního koeficientu,

`sd Cohen` – odmocnina z rozptylu upraveného korelačního koeficientu vypočteného ze vzorce (1.16),

`N` – rozsah výběru,

`pocet opakovani` – počet opakování výpočtu,

`podil vyrazenych` – desetinné číslo udávající, jaký podíl hodnot Y_i „zapomeneme“ (u nejmenších X_i).

B.2.1 Zdrojový text funkce `simuluj()`

```
simuluj = function(
xx,     #   "populace", ze ktere budeme simulovat, hodnoty x
yy,     #   "populace", ze ktere budeme simulovat, hodnoty y
p = 0.25,   #   kvantil x (jaký dil yy k nejmensim xx vyradime)
N = 100,    #   rozsah vyberu
B = 1000)   #   pocet simulovanych vyberu
{
x0 = quantile(xx,p=p)
rho = cor(xx,yy)
varHatRho = (1-rho^2)^2/N
    # klasicky asymptoticky rozptyl korel. koef.
    # sd klas. je odmocnina z nej
phi = pnorm(x0,mean(xx),sd(xx))
    # hodnota distribucni funkce
varHatRhoCohen = varHatRho*(2-rho^2*phi)/2/(1-phi)
    # rozptyl upraveneho korelacniho koeficientu
stat = matrix(NA,B,3)
    # matice (Bx3), kam postupně ukládám
for (iSim in 1:B){
    # vypocty pro jednotlive simulovane "vybery"
    indexy = sample(length(xx),N,replace=FALSE)
      # z celkoveho poctu vybiram N potenc. indexu (bez vraceni)
    x = xx[indexy]; y = yy[indexy]
      # beru vzdy dvojice, ktere k sobe opravdu patrily
    stat[iSim,1] = cor(x,y)
      # kor. koef. ze vsech vybranych dvojic pozorovani
    PLATNE = (x>=x0)
      # ponecham jen ty, ktere jsou vetsi nez dana mez
      # (urceno podilem vyrazenych p)
```

```
    stat[iSim,2] = rXY = cor(x[PLATNE],y[PLATNE])
      # kor. koef. uplnych dvojic pozorovani
    meanX = mean(x); meanXstar = mean(x[PLATNE])
    lambda = 1-crossprod(x[PLATNE]-meanXstar)/sum(PLATNE)/
        (crossprod(x-meanX)/N)
      # pozn.: crossprod je soucet ctvercu
    stat[iSim,3] = rXY/sqrt(1-lambda*(1-rXY^2))
      # opraveny korelacni koeficient
    }
cat("skutecnost\n","rho              ",rho,"\n",
                   "sd klas.         ",sqrt(varHatRho),"\n",
                   "sd Cohen         ",sqrt(varHatRhoCohen),"\n",
            "N                ",N,"\n",
            "pocet opakovani  ",B,"\n",
            "podil vyrazenych ",p,"\n")
# vypocty prumernych hodnot z B opakovani vyberu
prumery=apply(stat,2,mean)
sd=apply(stat,2,sd)
prehled = rbind(rxy=prumery,sd=sd)
colnames(prehled) = c("cely vyber","uplne dvojice","Cohenova oprava")
return(prehled)
}
```

B.2.2 Výpočet

```
library(mvtnorm)
   # knihovna potrebna ke generovani dvourozmerneho normalniho rozdeleni
rxy=0.4
  # zvolena hodnota korelacniho koeficientu
zz = rmvnorm(1000000,mean=c(0,0),sigma=matrix(c(1,rxy,rxy,1),2,2))
  # generovani a ulozeni dat do matice o dvou sloupcich
# nutno nacist simuluj.R
simuluj(zz[,1], zz[,2], p=0.5, N=500, B=1000)
  # provedeni vypoctu
```

Provedením výše uvedených kroků dostaneme:

```
skutecnost
 rho               0.3984279
 sd klas.          0.03762208
 sd Cohen          0.05215948
 N                 500
 pocet opakovani   1000
 podil vyrazenych  0.5
    cely vyber uplne dvojice Cohenova oprava
rxy 0.39753359    0.24976695      0.39091883
sd  0.03647269    0.06263292      0.08830136
```

Je patrno, že v případě, kdy mají obě veličiny normální rozdělení, přináší Cohenův postup velmi dobré výsledky. Hodnota průměrného opraveného korelačního koeficientu (0,3909) se téměř neliší od průměru korelačních koeficientů vypočtených ze všech hodnot výběru (0,3975), má však větší rozptyl. Při opakovaném volání funkce `simuluj()` bychom dostali podobné průměrné hodnoty `rxy` a `sd`, kdy korelační koeficient spočtený pouze z úplných dvojic bude vždy samozřejmě menší.

V následujícím textu budou uvedeny příklady, kdy jedna z veličin nemá normální rozdělení.

Jedna z veličin má rovnoměrné rozdělení

Při daném X je rozdělení Y normální a veličina X má rozdělení $\mathsf{R}(0,1)$.

```
n=100000
X = runif(n)
Y = rnorm(n, 0, 0.25) + X/3
simuluj(X, Y, p=0.5, N=500, B=1000)
skutecnost
 rho               0.3605179
 sd klas.          0.03890878
 sd Cohen          0.05395316
 N                 500
 pocet opakovani   1000
 podil vyrazenych  0.5
    cely vyber uplne dvojice Cohenova oprava
rxy 0.36256292    0.19059998       0.3561059
sd  0.03637859    0.05962995       0.1019133
```

Z opakovaných výpočtů je zřejmé, že Cohenův odhad nedává zavádějící výsledky, rozptyl opravených korelačních koeficientů je o něco vyšší než za předpokladu normálního rozdělení veličiny X.

Záměna pořadí X a Y při volání funkce způsobí, že předpokládáme „nenormální“ rozdělení u veličiny, kde část pozorování chybí.

```
skutecnost
 rho               0.3605179
 sd klas.          0.03890878
 sd Cohen          0.05418051
 N                 500
 pocet opakovani   1000
 podil vyrazenych  0.5
    cely vyber uplne dvojice Cohenova oprava
rxy 0.36030165    0.22107345      0.35073492
sd  0.03804571    0.05603108      0.08165293
```

Opakováním zjistíme, že Cohenova oprava dává nepatrně menší korelační koeficient než odpovídá skutečnosti.

Jedna z veličin má exponenciální rozdělení

Rozdělení jedné z veličin nyní zvolíme tak, aby nebylo symetrické, tedy například exponenciální. Při daném X je opět rozdělení Y normální a veličina X má rozdělení exp(4).

```
X = rexp(n,4)
Y = X/3 + rnorm(n, 0, 0.25)
simuluj(X, Y, p=0.5, N=500, B=1000)
skutecnost
 rho               0.3155536
 sd klas.          0.04026827
 sd Cohen          0.05063154
 N                 500
 pocet opakovani   1000
 podil vyrazenych  0.5
    cely vyber uplne dvojice Cohenova oprava
rxy 0.31477126    0.31492919      0.31562178
sd  0.04256915    0.06186846      0.05992804
```

Při porušení předpokladu normality (i symetrie) u veličiny, kde známe všechna pozorování, opět nedojde k vychýlení odhadů.

Zcela jinak je tomu u veličiny, kde některá pozorování chybějí. Výpočtem podle vzorce (1.14) dostaneme mnohem vyšší hodnotu opraveného korelačního koeficientu:

```
simuluj(Y, X, p=0.5, N=500, B=1000)
skutecnost
 rho               0.3155536
 sd klas.          0.04026827
 sd Cohen          0.0560012
 N                 500
 pocet opakovani   1000
 podil vyrazenych  0.5
    cely vyber uplne dvojice Cohenova oprava
rxy 0.31384558    0.27023004      0.40906851
sd  0.04271354    0.07184615      0.09627057
```

Příloha C

Skripty použité pro výpočty v programu R

S ohledem na možné problémy se správným zobrazováním českých znaků v programu R na různých počítačích, se kterými jsem se opakovaně setkala, raději je v níže uvedených skriptech nepoužívám. Doufám, že i tak budou dostatečně srozumitelné.

C.1 Predikce číselných kritérií úspěšnosti

C.1.1 Průměrný prospěch na VŠ

```
### Predpoved konecneho prumerneho prospechu na vysoke skole
# do vypoctu jsou zahrnuti studenti, u kterych znam prospech na SS,
# delali prijimaci zkousky a studium uspesne absolvovali
attach(vse)
#
# linearni model se dvema regresory
a<-lm(prumer~zcel+ssprum, subset=(JeSS=="ano"&Prijeti!="79"&Abs=="1"))
summary(a)
# test normality rezidui a normalni diagram
shapiro.test(resid(a))
qqnorm(resid(a),main="",
xlab="Teoreticke kvantily",ylab="Rezidua")
qqline(resid(a))
# test nezavislosti residui na odhadovane stredni hodnote
library(lmtest)     # potrebna knihovna pro funkci bptest()
bptest(a,~fitted(a))
plot(fitted(a),resid(a),
xlab="Hodnoty Yi odhadovane podle modelu",
ylab="Rezidua")
#
# pro dalsi vypocty musim pouzit omezeni na data jinym zpusobem
attach(vse[JeSS=="ano"&Prijeti!="79"&Abs=="1",])
# vypocet koeficientu pro normovane prediktory
summary(lm(prumer~scale(zcel)+scale(ssprum)))
#
```

```
# porovnani sily vlivu obou regresoru
library(psych)  # potrebna knihovna pro funkci paired.r()
# vypocty korelacnich koeficientu
(r01=abs(cor(prumer,zcel)))
(r02=abs(cor(prumer,ssprum)))
(r12=abs(cor(zcel,ssprum)))
(n=length(prumer))  # pocet pozorovani
paired.r(r01,r02,r12,n)
#
# normovanim i vysvetlovane veliciny dostaneme tzv. betavahy
summary(lm(scale(prumer)~scale(zcel)+scale(ssprum)))
#
# data budeme pro analyzu vybirat jinak
detach(vse[JeSS=="ano"&Prijeti!="79"&Abs=="1",])
#
# pokusy o pridaní dalsiho regresoru do modelu
a<-lm(prumer~zcel+ssprum,
 subset=(JeSS=="ano"&Prijeti!="79"&Abs=="1"))
a2<-lm(prumer~zcel+ssprum+Maturdrive,
 subset=(JeSS=="ano"&Prijeti!="79"&Abs=="1"))
# test podmodelu
anova(a,a2)
# neprokazeme vyznamnost Maturdrive
#
# pokus o pridani vlivu pohlavi
a2<-lm(prumer~zcel+ssprum+Pohlavi,
 subset=(JeSS=="ano"&Prijeti!="79"&Abs=="1"))
# test podmodelu
anova(a,a2)
# neprokazeme vyznamnost Pohlavi
#
# pokus o pridani vlivu statni prislusnosti (Ceska republika, Slovensko)
a2<-lm(prumer~zcel+ssprum+Stat,
 subset=(JeSS=="ano"&Prijeti!="79"&Abs=="1"))
# test podmodelu
anova(a,a2)
# neprokazeme vyznamnost statni prislusnosti
#
# pokus o pridani vlivu zacatku studia
a2<-lm(prumer~zcel+ssprum+Rok,subset=(JeSS=="ano"&Prijeti!="79"&Abs=="1"))
# test podmodelu
anova(a,a2)
# prokazeme statisticky vyznamny vliv rok zacatku studia
summary(a2)
#
# porovnani prumeru velicin u dvou skupin studentu podle zacatku studia
t.test(zcel~Rok, var.equal=TRUE,
 subset=(JeSS=="ano"&Prijeti!="79"&Abs=="1"))
```

```
t.test(ssprum~Rok, var.equal=TRUE,
 subset=(JeSS=="ano"&Prijeti!="79"&Abs=="1"))
t.test(prumer~Rok, var.equal=TRUE,
 subset=(JeSS=="ano"&Prijeti!="79"&Abs=="1"))
#
#
a3<-lm(prumer~zcel+ssprum+Rok+Prog,
 subset=(JeSS=="ano"&Prijeti!="79"&Abs=="1"))
# test podmodelu
anova(a2,a3)
# knihovna potrebna pro Anova
library(car)
Anova(a3)
anova(a3)
summary(a3)
# overeni predpokladu modelu se ctyrmi regresory
# test normality rezidui a normalni diagram
shapiro.test(resid(a3))
qqnorm(resid(a3));qqline(resid(a3))
# test nezavislosti residui na odhadovane stredni hodnote
bptest(a3,~fitted(a3))
plot(fitted(a3),resid(a3))
#
# vypocet hodnot inflacniho faktoru
library(car)    # potrebna knihovna
vif(a3)
#
# krokova regrese
# omezeni na data
attach(vse[JeSS=="ano"&Prijeti!="79"&Abs=="1",])
aS<-step(lm(prumer~1),scope=list(lower=~1,
 upper=~Prog+zcel+ssprum+ss1+ss2+ss3+ss4+
 Maturdrive+Pohlavi+Prijeti+Stat+Rok))
summary(aS)
detach(vse[JeSS=="ano"&Prijeti!="79"&Abs=="1",])
```

C.1.2 Průměrný prospěch v 1. ročníku VŠ

```
### Predpoved prumerneho prospechu v 1. rocniku vysoke skoly
# do vypoctu jsou zahrnuti studenti, u kterych znam prospech na SS,
# delali prijimaci zkousky a splnili podminky pro zapis do 2. rocniku
attach(vse)
#
# linearni model se ctyrmi regresory
a<-lm(pr1~zcel+ssprum+Prog+Rok,
subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
summary(a)
# test normality rezidui a normalni diagram
```

```
shapiro.test(resid(a))
qqnorm(resid(a));qqline(resid(a))
# test nezavislosti residui na odhadovane stredni hodnote
library(lmtest)     # potrebna knihovna pro funkci bptest()
bptest(a,~fitted(a))
plot(fitted(a),resid(a))
#
# pokus o pridani dalsich regresoru
# doba maturity Maturdrive
a2<-lm(pr1~zcel+ssprum+Prog+Rok+Maturdrive,
 subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
summary(a2)
# test podmodelu
anova(a,a2)
# neprokazeme vyznamnost Maturdrive
#
# vliv statni prislusnosti
a2<-lm(pr1~zcel+ssprum+Prog+Rok+Stat,
 subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
summary(a2)
# test podmodelu
anova(a,a2)
# neprokazeme vyznamnost faktoru Stat
#
# vliv pohlavi studentu Pohlavi
a2<-lm(pr1~zcel+ssprum+Prog+Rok+Pohlavi,
 subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
summary(a2)
# test podmodelu
anova(a,a2)
# neprokazeme vyznamnost Pohlavi na 5% hladine, ale na 10% ano
Anova(a2)
# test normality rezidui a normalni diagram
shapiro.test(resid(a2))
qqnorm(resid(a2));qqline(resid(a2))
# test nezavislosti residui na odhadovane stredni hodnote
bptest(a2,~fitted(a2))
plot(fitted(a2),resid(a2))
# vypocet hodnot inflacniho faktoru
library(car)    # potrebna knihovna
a<-lm(pr1~zcel+ssprum+Prog+Rok,
subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
vif(a)
a2<-lm(pr1~zcel+ssprum+Prog+Rok+Pohlavi,
 subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
vif(a2)
#
#
```

```
# pro dalsi vypocty musim pouzit omezeni na data jinym zpusobem
attach(vse[JeSS=="ano"&Prijeti!="79"&Splnil1=="1",])
# vypocet koeficientu pro normovane prediktory
summary(lm(pr1~scale(zcel)+scale(ssprum)))
#
# porovnani sily vlivu prospechu na SS a prijimacich zkousek
library(psych)  # potrebna knihovna pro funkci paired.r()
# vypocty korelacnich koeficientu
(r01=abs(cor(pr1,zcel)))
(r02=abs(cor(pr1,ssprum)))
(r12=abs(cor(zcel,ssprum)))
(n=length(pr1))     # pocet pozorovani
paired.r(r01,r02,r12,n)
detach(vse[JeSS=="ano"&Prijeti!="79"&Splnil1=="1",])
#
# porovnani prumeru velicin u dvou skupin studentu podle zacatku studia
t.test(zcel~Rok, var.equal=TRUE,
 subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
t.test(ssprum~Rok, var.equal=TRUE,
 subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
t.test(pr1~Rok, var.equal=TRUE,
 subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
# lisi se pouze prumerny pocet bodu u prijimacich zkousek
#
# krokova regrese
# omezeni na data
attach(vse[JeSS=="ano"&Prijeti!="79"&Splnil1=="1",])
aS<-step(lm(pr1~1),scope=list(lower=~1,
 upper=~Prog+zcel+ssprum+ss1+ss2+ss3+ss4+
 Maturdrive+Pohlavi+Prijeti+Stat+Rok))
summary(aS)
# vypocet inflacniho faktoru
vif(aS)
detach(vse[JeSS=="ano"&Prijeti!="79"&Abs=="1",])
```

C.2 Predikce kvalitativních kritérií úspěšnosti

C.2.1 Úspěšné ukončení studia

```
### Predpoved prumerneho prospechu v 1. rocniku vysoke skoly
# do vypoctu jsou zahrnuti studenti, u kterych znam prospech na SS,
# delali prijimaci zkousky a splnili podminky pro zapis do 2. rocniku
attach(vse)
#
# linearni model se ctyrmi regresory
a<-lm(pr1~zcel+ssprum+Prog+Rok,
subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
summary(a)
```

```
# test normality rezidui a normalni diagram
shapiro.test(resid(a))
qqnorm(resid(a));qqline(resid(a))
# test nezavislosti residui na odhadovane stredni hodnote
library(lmtest)     # potrebna knihovna pro funkci bptest()
bptest(a,~fitted(a))
plot(fitted(a),resid(a))
#
# pokus o pridani dalsich regresoru
# doba maturity Maturdrive
a2<-lm(pr1~zcel+ssprum+Prog+Rok+Maturdrive,
 subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
summary(a2)
# test podmodelu
anova(a,a2)
# neprokazeme vyznamnost Maturdrive
#
# vliv statni prislusnosti
a2<-lm(pr1~zcel+ssprum+Prog+Rok+Stat,
 subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
summary(a2)
# test podmodelu
anova(a,a2)
# neprokazeme vyznamnost faktoru Stat
#
# vliv pohlavi studentu Pohlavi
a2<-lm(pr1~zcel+ssprum+Prog+Rok+Pohlavi,
 subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
summary(a2)
# test podmodelu
anova(a,a2)
# neprokazeme vyznamnost Pohlavi na 5% hladine, ale na 10% ano
Anova(a2)
# test normality rezidui a normalni diagram
shapiro.test(resid(a2))
qqnorm(resid(a2));qqline(resid(a2))
# test nezavislosti residui na odhadovane stredni hodnote
bptest(a2,~fitted(a2))
plot(fitted(a2),resid(a2))
# vypocet hodnot inflacniho faktoru
library(car)    # potrebna knihovna
a<-lm(pr1~zcel+ssprum+Prog+Rok,
subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
vif(a)
a2<-lm(pr1~zcel+ssprum+Prog+Rok+Pohlavi,
 subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
vif(a2)
#
```

```
#
# pro dalsi vypocty musim pouzit omezeni na data jinym zpusobem
attach(vse[JeSS=="ano"&Prijeti!="79"&Splnil1=="1",])
# vypocet koeficientu pro normovane prediktory
summary(lm(pr1~scale(zcel)+scale(ssprum)))
#
# porovnani sily vlivu prospechu na SS a prijimacich zkousek
library(psych)  # potrebna knihovna pro funkci paired.r()
# vypocty korelacnich koeficientu
(r01=abs(cor(pr1,zcel)))
(r02=abs(cor(pr1,ssprum)))
(r12=abs(cor(zcel,ssprum)))
(n=length(pr1))     # pocet pozorovani
paired.r(r01,r02,r12,n)
detach(vse[JeSS=="ano"&Prijeti!="79"&Splnil1=="1",])
#
# porovnani prumeru velicin u dvou skupin studentu podle zacatku studia
t.test(zcel~Rok, var.equal=TRUE,
 subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
t.test(ssprum~Rok, var.equal=TRUE,
 subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
t.test(pr1~Rok, var.equal=TRUE,
 subset=(JeSS=="ano"&Prijeti!="79"&Splnil1=="1"))
# lisi se pouze prumerny pocet bodu u prijimacich zkousek
#
# krokova regrese
# omezeni na data
attach(vse[JeSS=="ano"&Prijeti!="79"&Splnil1=="1",])
aS<-step(lm(pr1~1),scope=list(lower=~1,
 upper=~Prog+zcel+ssprum+ss1+ss2+ss3+ss4+
 Maturdrive+Pohlavi+Prijeti+Stat+Rok))
summary(aS)
# vypocet inflacniho faktoru
vif(aS)
detach(vse[JeSS=="ano"&Prijeti!="79"&Abs=="1",])
```

C.2.2 Úspěšné absolvování 1. ročníku

```
### Predpoved splneni podminek pro zapis do 2. rocniku
# model logisticke regrese
# do vypoctu jsou zahrnuti studenti, u kterych znam prospech na SS,
# a delali prijimaci zkousky
attach(vse)
a=glm(Splnil1~zcel+ssprum+Prog,family =binomial,
 subset=(JeSS=="ano"&Prijeti!="79"))
summary(a)
library(car)
Anova(a)
```

```
#
# pokus o zarazeni veliciny Pohlavi
a2=glm(Splnil1~zcel+ssprum+Prog+Pohlavi,family =binomial,
 subset=(JeSS=="ano"&Prijeti!="79"))
summary(a2)
# test podmodelu
anova(a,a2,test="Chisq")
#
# pokus o zarazeni veliciny Rok
a2=glm(Splnil1~zcel+ssprum+Prog+Rok,family =binomial,
 subset=(JeSS=="ano"&Prijeti!="79"))
# test podmodelu
anova(a,a2,test="Chisq")
#
# pokus o zarazeni veliciny Stat
a2=glm(Splnil1~zcel+ssprum+Prog+Stat,family =binomial,
 subset=(JeSS=="ano"&Prijeti!="79"))
# test podmodelu
anova(a,a2,test="Chisq")
#
## pokus o zarazeni veliciny Pohlavi
a2=glm(Splnil1~zcel+ssprum+Prog+Pohlavi,family =binomial,
 subset=(JeSS=="ano"&Prijeti!="79"))
# test podmodelu
anova(a,a2,test="Chisq")
#
## pokus o zarazeni veliciny Maturdrive
a2=glm(Splnil1~zcel+ssprum+Prog+Maturdrive,family =binomial,
 subset=(JeSS=="ano"&Prijeti!="79"))
# test podmodelu
anova(a,a2,test="Chisq")
summary(a)
summary(a2)
#
# ROC krivka
# je treba nacist funkci ROCmodif.R
# musim omezit mnozinu, odkud vybiram
attach(vse[JeSS=="ano"&Prijeti!="79",])
#
roc<-ROC(predict(a2,type="r"),Splnil1,MI=FALSE,PV=FALSE)
# chci ziskat matici zamen pro zobrazene t0
# do nove promenne ulozim soucty senzitivity a specificity
soucet = apply(roc$res[,c("sens","spec")],1,sum)
(kdeMax = which.max(soucet))
# maximum je v 444. radku
# pro t0 = 0.8115
#
# vypisu odpovidajici matici zamen
```

```
addmargins(matrix(roc$tabulky[444,],2,2,byrow=FALSE))
# ulozeni prislusne kontingencni tabulky
tab=(matrix(roc$tabulky[444,],2,2,byrow=FALSE))
library(vcd)    # knihovna potrebna pro funkci Kappa
# vypocet kappa koeficientu
Kappa(tab)
#
# hledani maximalniho Acc
(maximum = max(apply(roc$tabulky[,c(1,4)],1,sum)))
# maximalni soucet spravne zarazenych je vyjde 1135
# hledani, u kterych to nastane zaznamu
which(apply(roc$tabulky[,c(1,4)],1,sum)==maximum)
# v jedinem pripade t0 je 0.57315 v radku 28
# najdu odpovidajici matici zamen
addmargins(matrix(roc$tabulky[28,],2,2,byrow=FALSE))
# ulozeni tabulky
tab=(matrix(roc$tabulky[28,],2,2,byrow=FALSE))
# vypocet kappa koeficientu
Kappa(tab)
#
#
# prikreslim do obrazku bod s maximalnim Acc
# t0= 0.57315
ROC(predict(a2,type="r"),Splnil1,MI=FALSE,PV=FALSE)
# ulozim hodnoty do pomocne promenne
pind=roc$tabulky[28,]
angle=135
dist=0.03
pomx=1-pind[1]/(pind[1]+pind[2])
pomy=pind[4]/(pind[4]+pind[3])
# pomocna cara
lines(c(pomx, pomx+dist*cos(pi*angle/180)),
 c(pomy, pomy+dist*sin(pi*angle/180)), col="red",lty=1)
# umisteni popisku
text (pomx-0.13,pomy-0.05, expression(paste(t[0]," = 0.573 ")),
 pos = 3, offset = 1.0, cex = 1.4, col = "red")
#
#
# pokus najit, pro jakou hodnotu t0 je kappa nejvetsi
# zjisteni, kolik mam hodnot
pocet<-dim(roc$tabulky)[1]
# priprava vektoru, kam budeme ukladat
kappavec=seq(0,0,length=pocet) #1332
for (i in 1:pocet)
    {
        tab=(matrix(roc$tabulky[i,],2,2,byrow=FALSE))
    # vypocet kappa koeficientu
    (kappavec[i]<-Kappa(tab)$Unweighted[1])
```

```
    }
# konec cyklu if
# hledam maximalni kappa
(maximum=max(kappavec))
# maximalni kappa je 0.2346319
# a v jakem je radku
which((kappavec)==maximum)
# radek 320
dimnames(roc$tabulky)
# odpovidajici hodnota t0 je 0.78040
addmargins(matrix(roc$tabulky[320,],2,2,byrow=FALSE))
# kontrola
tab=(matrix(roc$tabulky[320,],2,2,byrow=FALSE))
Kappa(tab)
#
#
# vypocet koeficientu determinace, resp. deviance modelu
# urceni deviance modelu
D1<-deviance(a2)
# nulova deviance
D0<-a2$null.deviance
n<-length(residuals(a2))
# Mc.Faddenuv koeficient determinace
(R2.McFadden<-1-D1/D0)
# jina definice koeficientu determinace
(R2<-1-exp((D1-D0)/n))
# maximalni koeficient determinace
(R2.max<-1-exp(-D0/n))
# Nagelkerkova modifikace koeficientu determinace
(R2.Nagel<-R2/R2.max)
#
detach(vse[JeSS=="ano"&Prijeti!="79",])
```

C.2.3 Absolvování studia s vyznamenáním

```
### Predpoved absolvovani s vyznamenanim
# model logisticke regrese
# do vypoctu jsou zahrnuti studenti, u kterych je znam prospech na SS
attach(vse)
a=glm(vyzn~ssprum+Prog+Prijeti,family =binomial,
 subset=(JeSS=="ano"))
summary(a)
Anova(a)
#
# ROC krivka
# je treba nacist funkci ROCmodif.R
# musim omezit mnozinu, odkud vybiram
attach(vse[JeSS=="ano",])
```

```
#
roc<-ROC(predict(a,type="r"),vyzn,MI=FALSE,PV=FALSE)
# funkci ROCmodif.R jsme dale upravila,
# oznaceni bodu psala mimo rozsah
# dale doplnim do obrazku rucne
# do nove promenne ulozim soucty senzitivity a specificity
soucet = apply(roc$res[,c("sens","spec")],1,sum)
(kdeMax = which.max(soucet))
# maximum je v 962. radku
# pro t0 = 0.03859
#
# vypisu odpovidajici matici zamen
addmargins(matrix(roc$tabulky[962,],2,2,byrow=FALSE))
# ulozeni prislusne kontingencni tabulky
tab=(matrix(roc$tabulky[962,],2,2,byrow=FALSE))
library(vcd)    # knihovna potrebna pro funkci Kappa
# vypocet kappa koeficientu
Kappa(tab)
#
# hledani maximalniho Acc
(maximum = max(apply(roc$tabulky[,c(1,4)],1,sum)))
# maximalni soucet spravne zarazenych vyjde 1338
# hledani, u kterych to nastane zaznamu
which(apply(roc$tabulky[,c(1,4)],1,sum)==maximum)
# v jedinem pripade t0 je 0.48095 v radku 1137
# najdu odpovidajici matici zamen
addmargins(matrix(roc$tabulky[1137,],2,2,byrow=FALSE))
# ulozeni tabulky
tab=(matrix(roc$tabulky[1137,],2,2,byrow=FALSE))
# vypocet kappa koeficientu
Kappa(tab)
#
# prikreslim do obrazku modre oznaceni bodu
# (funkce psala mimo rozsah)
ROC(predict(a,type="r"),vyzn,MI=FALSE,PV=FALSE)
text (0.26,0.76, expression(paste(t[0]," = 0.039 ")),
 pos = 3, offset = 1.0, cex = 1.4, col = "blue")
#
# prikreslim do obrazku cervene bod s maximalnim Acc
# t0= 0.48095
# ulozim hodnoty do pomocne promenne
pind=roc$tabulky[1137,]
angle=45
dist=0.03
pomx=1-pind[1]/(pind[1]+pind[2])
pomy=pind[4]/(pind[4]+pind[3])
# pomocna cara
lines(c(pomx, pomx+dist*cos(pi*angle/180)),
```

```
 c(pomy, pomy+dist*sin(pi*angle/180)), col="red" ,lty=1)
# umisteni popisku
text (pomx+0.15,pomy-0.03, expression(paste(t[0]," = 0.481 ")),
 pos = 3, offset = 1.0, cex = 1.4, col = "red")
#
#
# pokus najit, pro jakou hodnotu t0 je kappa nejvetsi
# zjisteni, kolik mam hodnot
pocet<-dim(roc$tabulky)[1]
# priprava vektoru, kam budeme ukladat
kappavec=seq(0,0,length=pocet) #1148
for (i in 1:pocet)
    {
        tab=(matrix(roc$tabulky[i,],2,2,byrow=FALSE))
    # vypocet kappa koeficientu
    (kappavec[i]<-Kappa(tab)$Unweighted[1])
    }
# konec cyklu if
# hledam maximalni kappa
(maximum=max(kappavec))
# maximalni kappa je 0.4334941
# a v jakem je radku
which((kappavec)==maximum)
# radek 1117
dimnames(roc$tabulky)
# odpovidajici hodnota t0 je 0.17912552
addmargins(matrix(roc$tabulky[1117,],2,2,byrow=FALSE))
# kontrola
tab=(matrix(roc$tabulky[1117,],2,2,byrow=FALSE))
Kappa(tab)
#
#
# vypocet koeficientu determinace, resp. deviance modelu
# urceni deviance modelu
D1<-deviance(a)
# nulova deviance
D0<-a$null.deviance
n<-length(residuals(a))
# Mc.Faddenuv koeficient determinace
(R2.McFadden<-1-D1/D0)
# jina definice koeficientu determinace
(R2<-1-exp((D1-D0)/n))
# maximalni koeficient determinace
(R2.max<-1-exp(-D0/n))
# Nagelkerkova modifikace koeficientu determinace
(R2.Nagel<-R2/R2.max)
#
detach(vse[JeSS=="ano",])
```

C.2.4 Modifikovaná funkce ROC()

```
steplines <-
function( x,
          y,
       left = TRUE,
      right = !left,
      order = TRUE,
        ... )
{
# A function to plot step-functions
#
  # Get the logic right if right is supplied...
  left <- !right   # ... right!
  n    <- length( x )
  if( any( order ) ) ord <- order(x) else ord <- 1:n
  dbl <- rep( 1:n, rep( 2, n) )
  xv  <- c( !left, rep( T, 2*(n-1) ),  left)
  yv  <- c(  left, rep( T, 2*(n-1) ), !left)
  lines( x[ord[dbl[xv]]],
         y[ord[dbl[yv]]], ... )
}
interp <-
function ( target, fv, res )
{
# Linear interpolaton of the values in the N by 2 matrix res,
# to the target value target on the N-vector fv.
# Used for placing tickmarks on the ROC-curves.
#
where <- which( fv>target )[1] - 1:0
  int <- fv[where]
   wt <- ( int[2] - target ) / diff( int )
wt[2] <- 1-wt
t( res[where,] ) %*% wt
}
ROC.tic <-
function ( tt,
          txt = formatC(tt,digits=2,format="f"),
         dist = 0.02,
        angle = +135,
          col = "black",
          cex = 1.0,
           fv,
          res )
{
# Function for drawing tickmarks on a ROC-curve
#
for (i in 1:length(tt))
    {
```

```
    pnt <- interp ( tt[i], fv, res )
    x <- 1-pnt[2]
    y <- pnt[1]
    lines( c( x, x+dist*cos(pi*angle/180) ),
           c( y, y+dist*sin(pi*angle/180) ), col=col )
    text( x+dist*cos(pi*angle/180),
          y+dist*sin(pi*angle/180), txt[i], col=col,
          adj=c( as.numeric(abs(angle)>=90),
                 as.numeric(    angle <= 0)), cex=cex )
    }
}
ROC <-
function ( test = NULL,
           stat = NULL,
           form = NULL,
           plot = c( "sp", "ROC" ),
             PS = is.null(test),      # Curves on probability scale
             PV = TRUE,               # sn, sp, PV printed at "optimality" point
             MX = TRUE,               # tick at "optimality" point
             MI = TRUE,               # Model fit printed
            AUC = TRUE,               # Area under the curve printed
           grid = seq(0,100,100),     # Background grid, upraveno
      col.grid = "black",        # puvodne bylo gray( 0.9 ),
           cuts = NULL,
            lwd = 2,
           data = parent.frame(),
            ... )
{
# First all the computations
#
# Name of the response
  rnam <- if ( !missing( test ) )
             deparse( substitute( test ) ) else
             "lr.eta"
# Fit the model and get the info for the two possible types of input
  if( is.null( form ) )
    {
    if( is.null( stat ) | is.null( test ) )
      stop( "Either 'test' AND 'stat' OR 'formula' must be supplied!" )
    lr <- glm( stat ~ test, family=binomial )#, data=data )
    resp <- stat
    Model.inf <- paste("Model: ", deparse( substitute( stat ) ), "~",
                                  deparse( substitute( test ) ) )
    }
  else
    {
    lr <- glm( form, family=binomial )#, data=data )
    resp <- eval( parse( text=deparse( form[[2]] ) ) )
```

```
    Model.inf <- paste("Model: ",paste(paste(form)[c(2,1,3)], collapse=" "))
    }
# Form the empirical distribution function for test for each of
# the two categories of resp.

# First a table of the test (continuous variable) vs. the response
  m  <- as.matrix( base:::table( switch( PS+1, test, lr$fit ), resp ) )
# What values do they refer to
  fv <-     sort( unique( switch( PS+1, test, lr$fit ) ) )
# How may different values of the test variable do we have?
  nr <- dim( m )[1]
# Number in the two outcome categories
  a <- apply( m, 2, sum )
# Add a the sum for each value of test
  m <- addmargins( m, 2 )
# The calculate the empirical distribution functions:
  m <- apply( m[nr:1,], 2, cumsum )[nr:1,]
##  MODIFIKACE  - ulozeni spocitanych cetnosti:
  tabulky = cbind(a[1]-m[,1],m[,1], a[2]-m[,2],m[,2])
##  dostaneme cetnosti v poradi TN, FP, FN, TP
##  konec MODIFIKACE
# Then the relevant measures are computed.
  sn <- c( m[,2] / a[2], 0 )
  sp <- c( (a[1]-m[,1]) / a[1], 1 )
  pvp <- c( m[,2] / m[,3], 1 )
  pvn <- (a[1] - m[,1]) / ( sum(a) - m[,3] )
  pvn <- c( pvn, rev( pvn )[1] )
  res <- data.frame( cbind( sn, sp, pvp, pvn, c(NA,fv) ) )
  auc <- sum( ( res[-1,1] + res[-nr,1] ) / 2 * diff( res[,2] ) )
  names( res ) <- c( "sens", "spec", "PV+", "PV-", rnam )

# Plot of sens, spec, PV+, PV-:
if ( any( !is.na( match( c( "SP", "SNSP", "SPV" ), toupper( plot ) ) ) ) )
{
# First for probability scale
if ( PS ) {
       plot( 0:1, 0:1,
             xlim=0:1, xlab="Cutpoint for predicted probability",
             ylim=0:1, ylab=" ",
             type="n" )
       if( is.numeric( grid ) ) abline( h=grid/100, v=grid/100, col=col.grid )
       box()
       for ( j in 4:1 ){
       steplines( fv, res[,j], lty=1, lwd=lwd, col=gray((j+1)/7)) }
       text( 0, 1.01, "Sensitivity", cex=0.7, adj=c(0,0), font=2 )
       text( 1, 1.01, "Specificity", cex=0.7, adj=c(1,0), font=2 )
       text( 0,  a[2]/sum(a)-0.01, "PV+", cex=0.7, adj=c(0,1), font=2 )
       text( 0 + strwidth( "PV+", cex=0.7 ),  a[2]/sum(a)-0.01,
```

```
            paste( " (= ", a[2],"/", sum(a), " =",
                   formatC( 100*a[2]/sum(a), digits=3 ),
                   "%)", sep=""),
            adj=c(0,1), cex=0.7 )
      text( 1, 1-a[2]/sum(a)-0.01, "PV-", cex=0.7, adj=c(1,1), font=2 )
          }
# then for test-variable scale
else {
      xl <- range( test )
      plot( xl, 0:1,
            xlim=xl,
            xlab=paste( deparse( substitute( test ) ), "(quantiles)" ),
            ylim=0:1,         ylab=" ",
            type="n" )
      if( is.numeric( grid ) )
        abline( h=grid/100, v=quantile( test, grid/100 ), col=col.grid )
      box()
      for ( j in 4:1 ){
      steplines( fv, res[,j], lty=1, lwd=lwd, col=gray((j+1)/7))}
      text( xl[1], 1.01, "Sensitivity", cex=0.7, adj=c(0,0), font=2 )
      text( xl[2], 1.01, "Specificity", cex=0.7, adj=c(1,0), font=2 )
      text( xl[1],  a[2]/sum(a)-0.01, "PV+", cex=0.7, adj=c(0,1), font=2 )
      text( xl[1] + strwidth( "PV+", cex=0.7 ),  a[2]/sum(a)-0.01,
            paste( " (= ", a[2],"/", sum(a), " =",
                   formatC( 100*a[2]/sum(a), digits=3 ),
                   "%)", sep=""),
            adj=c(0,1), cex=0.7 )
      text( xl[2], 1-a[2]/sum(a)-0.01, "PV-", cex=0.7, adj=c(1,1), font=2 )
      }
}
# Plot of ROC-curve:
if ( any( !is.na( match( "ROC", toupper( plot ) ) ) ) )
{
      plot( 1-res[,2], res[,1],
            xlim=0:1, xlab="FPR = 1-Specificita",
            ylim=0:1, ylab=  "TPR = Senzitivita",
            type="n", ...)
      if( is.numeric( grid ) ) abline( h=grid/100, v=grid/100, lwd=1,
       lty =1,     # zmena typu kresleni car
       col=gray( 0.5 ) )
      abline( 0, 1, col="black",lwd=2 )
      box()
      lines( 1-res[,2], res[,1], lwd=lwd, col="blue")

  # Tickmarks on the ROC-curve
      if ( !is.null(cuts) )
      {
      ROC.tic( cuts,
```

```
              txt=formatC( cuts, digits=2, format="f" ),
              fv=fv, res=res, dist=0.03, cex=0.7)
     }

# Plot of optimality point
     if (MX)
     {
     mx <- max( res[,1]+res[,2] )
     mhv <- which( (res[,1]+res[,2])==mx )[1]
     mxf <- fv[mhv]
     abline( mx-1, 1, col="blue",lwd=1 )
     ROC.tic( mxf,
  # bud necham vypsat zjistenou hodnotu nebo
  # rozumne zformatuju konkretne vypoctenou
         txt=paste( "t0 =", formatC( mxf, format="f", digits=3 ) ),
    #    txt=expression(paste( t[0], " = 0.663")),
      fv=fv, res=res, dist=0.03, cex=1.4, angle=135,col="blue" )
     }

# Model information
     if (MI)
     {
     crn <- par()$usr
     text(0.95*crn[2]+0.05*crn[1], 0.07, Model.inf,
          adj=c(1,0.5),cex=1.0)
     cf <- summary(lr)$coef[,1:2]
     nf <- dimnames(cf)[[1]]

     text(0.95*crn[2]+0.05*crn[1], 0.10,
          paste("Variable\ \ \ \ \ est.\ \ \ \ \ (s.e.) \ \ \n",
                paste(rbind(nf,
                            rep("\ \ \ \ ",length(nf)),
                            formatC(cf[,1],digits=3,format="f"),
                            rep("\ \ \ (",length(nf)),
                            formatC(cf[,2],digits=3,format="f"),
                            rep(")",length(nf)),
                            rep("\n",length(nf))),
                      collapse=""),
                collapse=""),
          adj=c(1,0), cex=0.7 )
     }

# Print the area under the curve
     if (AUC)
     {
     crn <- par()$usr
     text( 0.95*crn[2]+0.05*crn[1], 0.00,
           paste( "AUC = ",
```

```
                    formatC( auc, format="f", digits=3, width=5 ) ),
              adj=c(1,0), cex=1.4 )
       }

  # Predictive values at maximum
       if (PV)
       {
       if(!MX) { mx <- max(res[,1]+res[,2])
                 mhv <- which((res[,1]+res[,2])==mx)
                 mxf <- fv[mhv]
               }
       ROC.tic(mxf, fv=fv, res=res,
                    txt= paste( "Sens: ",
                                formatC(100*res[mhv,1],digits=1,format="f"),
                           "%\n", "Spec: ",
                                formatC(100*res[mhv,2],digits=1,format="f"),
                           "%\n", "PV+: ",
                                formatC(100*res[mhv,3],digits=1,format="f"),
                           "%\n", "PV-: ",
                                formatC(100*res[mhv,4],digits=1,format="f"),
                           "%", sep="" ),
                    dist=0.1, cex=0.7, angle=-45 )
       }
}
invisible( list( res=res, AUC=auc, lr=lr , tabulky=tabulky) )
}
```

C.3 Ordinální regrese

```
### Predpoved ordinalni veliciny Abs3
#
# potrebna knihovna pro funkci polr
library(MASS)
attach(vse)
# musim omezit mnozinu, odkud vybiram
attach(vse[JeSS=="ano"&Prijeti!="79",])
a<-polr(Abs3~ssprum,Hess=TRUE)
summary(a)
# test paralelnich krivek
# nutno nacist funkci Brant.R
mcezav<-model.matrix(a)
testBrant(abs3,mcezav)
# p-hodnota tesne nad 5%
#
# ulozeni predikovanych hodnot do tabulky
tab=predict(a,type="class")
### vypocet kappa koeficientu
# potrebna knihovna pro funkci Kappa
```

```
library(vcd)
Kappa(xtabs(~tab+Abs3))
# tabulka cetnosti
addmargins(xtabs(~tab+Abs3))
#
# pridani dalsiho regresoru do modelu
a2<-polr(Abs3~ssprum+Maturdrive,Hess=TRUE)
summary(a2)
# test podmodelu
1-pchisq(deviance(a)-deviance(a2),1)
# test paralelnich krivek
mcezav<-model.matrix(a2)
testBrant(abs3,mcezav)
# paralelni krivky OK p=11%
#
# ulozeni predikovanych hodnot do tabulky
tab=predict(a2,type="class")
### vypocet kappa koeficientu
Kappa(xtabs(~tab+Abs3))
# tabulka cetnosti
addmargins(xtabs(~tab+Abs3))
#
#
# pridani dalsich regresoru do modelu
a3<-polr(Abs3~ssprum+Maturdrive+zcel,Hess=TRUE)
summary(a3)
# test podmodelu
1-pchisq(deviance(a2)-deviance(a3),1)
# test paralelnich krivek
mcezav<-model.matrix(a3)
testBrant(abs3,mcezav)
# p-hodnota = 9.5%
#
# ulozeni predikovanych hodnot do tabulky
tab=predict(a3,type="class")
### vypocet kappa koeficientu
Kappa(xtabs(~tab+Abs3))
#
#
# pridani faktoru Prog do modelu
a4<-polr(Abs3~ssprum+Maturdrive+zcel+Prog,Hess=TRUE)
summary(a4)
# test podmodelu (vyssi pocet stupnu volnosti)
1-pchisq(deviance(a3)-deviance(a4),6)
# test paralelnich krivek
mcezav<-model.matrix(a4)
testBrant(abs3,mcezav)
# zamitnuty paralelni krivky, vyradit Prog
```

```
#
# pridani Pohlavi do modelu
a4<-polr(Abs3~ssprum+Maturdrive+zcel+Pohlavi,Hess=TRUE)
summary(a4)
# test podmodelu
1-pchisq(deviance(a3)-deviance(a4),1)
# p-hodnota je 8%
#
# pridani Rok do modelu
a4<-polr(Abs3~ssprum+Maturdrive+zcel+Rok,Hess=TRUE)
summary(a4)
# test podmodelu
1-pchisq(deviance(a3)-deviance(a4),1)
# p-hodnota je 1%, Rok ma vliv
# test paralelnich krivek
mcezav<-model.matrix(a4)
testBrant(abs3,mcezav)
# zamitnuty paralelni krivky
#
# pridani Stat do modelu
a4<-polr(Abs3~ssprum+Maturdrive+zcel+Stat,Hess=TRUE)
summary(a4)
# test podmodelu
1-pchisq(deviance(a3)-deviance(a4),1)
# p-hodnota je 22%, Stat nema vliv
#
#
detach(vse[JeSS=="ano"&Prijeti!="79",])
```

Literatura

[1] Agresti, A. *Categorial Data analysis*. 2nd ed. New York: Wiley, 2002. ISBN 0-471-360093-7.

[2] Anděl, J. *Statistické metody*. 2. vyd. Praha: MATFYZPRESS, 2003. ISBN 80-85863-27-8.

[3] Anderson, T., W. Maximum likelihood estimates for a multivariate normal distribution when some observation are missing. *Journal of the American Statistical Association*. 1957, Vol. 52, No. 278, s. 200–203.

[4] Betinec, M. Použití ROC křivek pro hodnocení klasifikátorů. In *ROBUST 2006*. Jednota českých matematiků a fyziků. 2006. s. 25–34.

[5] Box, G., E., P. – Cox, D., R. An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B.* 1964, Vol. 26, s. 211–252.

[6] Breusch, T., S. – Pagan, A., R. A simple test for heteroscedasticity and random coefficient variation. *Econometrica*. 1979, Vol. 47, No. 5, s. 1287–1294.

[7] Budíková, M. – Mikoláš, Š. Analýza úspěšnosti magisterského studia na Přírodovědecké fakultě Masarykovy univerzity v Brně. In *5th International Conference Aplimat*. Bratislava: Katedra matematiky SJF STU Bratislava, 2006. s. 293–299. ISBN 80-967305-4-1.

[8] Byčkovský, P., Zvára K. *Konstrukce a analýza testů pro přijímací řízení*. Universita Karlova v Praze, Pedagogická fakulta. 2007. ISBN 978-80-7290-331-3.

[9] Cohen, A., C. Jr. Restriction and selection in samples from bivariate normal distributions. *Journal of the American Statistical Association*. 1955, Vol. 50, No. 271, s. 884–893.

[10] Cohen, B., H. *Explaining Psychological Statistics*. 2nd ed. New York – Chichester – Weinheim – Brisbane – Singapore – Toronto: Wiley, 2001. ISBN 0-471-34582-2.

[11] Fawcett, T. ROC Graphs: Notes and practical considerations for researches. *Technical report*. 2004, Palo Alto, USA: HP Laboratories. 38 s.

[12] Geiser, S. – Studley, R. Predictive validity and differential impact of the SAT I and SAT II at the University of California. *Educational Assessment*. 2002, Vol. 8, No. 1, s. 1–26.

[13] Höschl, C. – Kožený, J. Predicting academic performance of medical students: the first three years. *The American Journal of Psychiatry*. 1997, Vol. 154, Num. 6, s. 87–92.

[14] Hotelling, H. The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *Annals of Mathematical Statistics.* 1940, Vol. 11, s. 271–283.

[15] Kendall, M., G. *Rank Correlation Methods*, 3rd ed. Griffin, London, 1962.

[16] Kobrin, J. L. et al. Validity of the SAT for predicting first-year college grade point average *College Board Research Report.* New York: The College Board. No. 2008-5, 10 s.

[17] Kubíček, V. *Multinomická a ordinální regrese.* Diplomová práce. Univerzita Karlova v Praze. Matematicko–fyzikální fakulta. 2005. 79 s. Vedoucí diplomové práce doc. RNDr. Karel Zvára, CSc.

[18] Kuncel, N., R. – Hezlett, S., A. – Ones, D., S. A Comprehensive meta-analysis of the predictive validity of the graduate record examination: Implications for graduate student selection and performance. *Psychological Bulletin*, 2001. Vol. 127, No. 1, s. 162–181.

[19] Lin, L., I. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 1989. Vol. 45, s. 255–268. ISSN 0006-341X.

[20] Mann, H., B. – Whitney, D., R. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 1947. Vol. 18, No. 1, s. 50–60.

[21] Menard, S. Coefficient of determination for multiple logistic regression analysis. *Journal of the American Statistical Association*, 2000. Vol. 54, s. 17–24.

[22] Nagelkerke, N., J., D. A note on a general definition of the coefficient of determination. *Biometrika*, 1991. Vol. 78, s. 691–692.

[23] Nagy, S. *Odhad korelačního koeficientu, když je jedna složka cenzorovaná.* Bakalářská práce. Univerzita Karlova v Praze. Matematicko–fyzikální fakulta. 2009. Vedoucí bakalářské práce doc. RNDr. Karel Zvára, CSc.

[24] Neill, J., J. – Dunn, O., J. Equality of dependent correlation coefficient. *Biometrics*, 1975. Vol. 31, No. 2, s. 531–543.

[25] Pagani, L. – Segheiri, Ch. Predictive validity of high school grade and other characteristics on students' university careers using ROC Analysis. In *Metodološki zvezki*, Ljubljana, 2003, s. 197–204. ISSN 1318-1726.

[26] Pepe, M., S. Receiver operating characteristic methodology. *Journal of the American Statistical Association.* 2000, Vol. 95, No. 449, s. 308–311. ISSN 0162-1459.

[27] Qin, J. – Zhang, B. Using logistic regresion procedures for estimating receiver operating characteristic curves. *Biometrika*, 2003, Vol. 90, No. 3, s. 585–596. ISSN 1464-3510.

[28] Royston, J., P. Algorithm AS 181: The W test for Normality. *Applied Statistics.* 1982, Vol. 31, No. 2, s. 176–180.

[29] Sachs, L. *Angewandte Statistik*, 4. Auflage. Springer–Verlag. Berlin – Heidelberg – New York. 1974.

[30] Steiger, J., H. Test for comparing elements of a correlation matrix. *Psychological Bulletin*, 1980. Vol. 87, No. 2, s. 245–251.

[31] Synek, J. – Otřísal, V. *Predikční validita testu OSP - výsledky analýzy.* 2008, Scio, s.r.o. [cit. 11. 4. 2009]. Dostupné na internetu:
`http://www.testovani.cz/1_download/nsz/Predikcni_validita_OSP.pdf`

[32] Štuka, Č. – Šimeček, P. Studium souvislostí mezi úspěšností studia medicíny, známkami na střední škole a výsledky přijímacích zkoušek. In *Sborník Medsoft 2006*, MEDSOFT 2006.

[33] Škaloudová, A. *Predikce úspěšnosti ve studiu učitelství.* Disertační práce. Univerzita Karlova v Praze. Pedagogická fakulta. 2003. 132 s.

[34] Williams, E., J. The comparison of regression variables. *Journal of the Royal Statistical Society.* Serie B (Methodological), 1959. Vol. 21, No. 2, s. 396–399.

[35] Zvára, K. *Regrese.* Praha: MATFYZPRESS, 2008. ISBN 978-80-7378-041-8.

[36] Zvára, K. *Biostatistika.* 2. vyd. Praha: Karolinum, 2003. ISBN 80-246-0739-5.

[37] Zvára, K. – Anděl, J. Souvislost výsledků přijímacího řízení s úspěšností studia na MFF. *Pokroky matematiky, fyziky a astronomie*, 2001, roč. 46, č. 6, s. 304–312. ISSN 0032-2423.

[38] Zvárová, J. *Biostatistika pro biomedicínské obory I.* Praha: Karolinum, 2007. ISBN 978-80-7184-786-1.

[39] Zwick, R. Higher education admissions testing. 4th ed. In *Educational measurement.* ACE/Praeger series on higher education. 2006. s. 647–679. ISBN 0-275-98125-8.

[40] ACT: The relative predictive validity of ACT scores and high school grades in making college admission decisions. *College Success.* 2008. 4 s.