Vysoká škola ekonomická v Praze

Fakulta informatiky a statistiky

Studijní program: Bakalářský

Studijní obor: Statistika a ekonometrie

VLIV VYBRANÝCH FAKTŮRŮ NA VELIKOST GRE A PŘIJETÍ NA ZAHRANIČNÍ UNIVERZITU ČI COLLEGE

Semestrální práce

Autor: Lenka Raslová

Akademický rok 2020/2021

Úvod

Každý rok se tisíce zahraničních studentů hlásí na univerzity a college v USA. Jedním z důvodů, proč se studenti hlásí na vysoké školy do zahraniční může být touha získat velmi kvalitní vzdělání anebo si zažít život v USA. Studijní oddělení každý rok obdrží tisíce žádostí o přijetí, ale jen zlomek studentů bývá přijat. Cílem této semestrální práce je vytvořit pomocí vícenásobné regrese model, kterým by se dala odhadnout velikost počtu bodů ze zkoušky GRE, která je nutná pro přijetí na zahraniční školy, v závislosti na různých vlastnostech potencionálního studenta. A dále ověřit, jestli je s vyšším GPA pravděpodobnost možného přijetí vyšší.

1 Teoretický základ

Na základě informacích o zahraničních studentech jsou v semestrální práci předpokládány následující hypotézy:

H1: Student s vyšším GPA získá více bodů ze zkoušky GRE.

H2: U studentů s průměrným GPA se velikost získaných bodů z GRE pro muže a ženy významně liší.

H3: Při vyšším GPA existuje větší pravděpodobnost, že student bude přijat na univerzitu či college.

H4: Socioekonomické prostředí, ze kterého studenti pochází, má významný vliv na přijetí na univerzitu či college.

2 Data

Data, se kterými se pracuje v semestrální práci, pochází z *Kagglu*. *Kaggle* je internetové vědecko-datové prostředí, které umožňuje uživatelům nahrávat a stahovat různé datové sety (Kaggle Inc © 2021). Soubor, se kterým se pracuje v semestrální práci, obsahuje základní údaje z přihlášek zahraničních studentů (např. GPA, GRE, rasa, pohlaví, úroveň univerzity/college na jakou se student hlásí a atd.).

Data jsou průřezového typu, mají velikost 400 pozorování. Datový soubor byl na *Kaggle* nahrán v roce 2019 pro studijní účely. V popisu datového souboru je uvedeno, že data pochází z *Education Department* (nikde však není uvedeno nic konkrétnějšího). Žádné další informace o datovém souboru tedy nejsou zveřejněné. Z tohoto důvodu není známé datum a ani způsob sběru dat. Ovšem dá se předpokládat, že data jsou primárního charakteru. Data byla s největší pravděpodobností získána z přihlášek zahraničních studentů, kteří se na univerzity a college hlásili. Údaje uvedené v souboru se běžně používají pro analýzy faktorů, které ovlivňují přijetí studentů na univerzity a college v USA.

V datovém souboru se nacházejí kvantitativní a kategoriální proměnné. Mezi kvantitativní proměnné patří *GPA* (*Grade Point Average*), což je vážený průměr známek (známky pochází ze 4. stupňové škály, kde 4 je horní hranicí). Další kvantitativní proměnnou je *GRE* (*Graduate Record Exam Scores*), což je počet bodů získaný z této zkoušky. *GRE* má tři části: verbální, analytickou a esej (ETS © 2021). Při přirovnání k českým zkouškám je tato zkouška podobná Scio testu.

Mezi kategoriální proměnné patří *úroveň univerzity* (resp. college), kde jsou rozlišovány 4 stupně (od 1 do 4), 1 značí nejvíce prestižní univerzitu, oproti tomu 4 značí nejhorší možnou úroveň. Další proměnnou je *přijetí* (1 značí přijetí na danou instituci a 0 značí nepřijetí), *pohlaví studenta* (0 pro ženu, 1 pro muže) a *SE* (*socioeconomic status*), kde se rozlišují 3 úrovně socioekonomických statusů zahraničních studentů (kde 1 je nejhorší a 3 nejlepší úroveň socioekonomického statusu). Poslední kategoriální proměnnou je *rasa*, kde 1 značí hispánce, 2 asiaty a 3 afroameričany.

3 Metodologie

3.1 Lineární model

Vícenásobný regresní model odhadneme na základě metody nejmenších čtverců. Závislou proměnnou je počet bodů z GRE a nezávislé proměnné jsou charakteristiky studentů: GPA, úroveň univerzity, socioekonomická situace studenta, rasa studenta a pohlaví studenta a interakce mezi pohlavím a GPA. Model má následující tvar:

$$\begin{aligned} GRE = \beta_0 + \beta_1 GPA\_3.4 + \beta_2 druhá\ nejvyšší\ úroveň\ univerzity \\ + \beta_3\ střední\ úroveň\ univerzity + \beta_4 velmi\ špatná\ úroveň\ univerzity \\ + \beta_5 střední\ socioekonomická\ situace\ studenta \\ + \beta_6 vysoká\ socioekonomická\ situace\ studenta + \beta_7 muž + \beta_8 asiat \\ + \beta_9 afroameričan + \beta_{10} muž * GPA\_3.4 + u. \end{aligned}$$

Referenční skupinou pro tento model je žena hispánka, která dosáhla průměrného GPA (*GPA* bylo vycentrováno hodnotou 3.4), hlásí se na nejprestižnější univerzitu a její socioekonomická situace je na nejhorší možné úrovni.

V modelu předpokládáme, že jsou splněny základní předpoklady pro vícenásobnou lineární regresi. Tedy že platí Gauss-Markovovy předpoklady.

3.2 Pravděpodobnostní model

Pro zodpovězení třetí hypotézy je potřeba využít logistickou regresi. Pro účely semestrální práce je využit logit a probit, obecné charakteristiky těchto modelů podrobně popisuje Wooldridge, 2016. Model má následující tvar:

$$\begin{aligned} přijetí = \beta_0 + \beta_1 GPA\_3.4 + \beta_2 GRE\_588 + \beta_3 druhá\ nejvyšší\ úroveň\ univerzity \\ + \beta_4\ střední\ úroveň\ univerzity + \beta_5 velmi\ špatná\ úroveň\ univerzity \\ + \beta_6 střední\ socioekonomická\ situace\ studenta \\ + \beta_7 vysoká\ socioekonomická\ situace\ studenta + \beta_8 muž + \beta_9 asiat \\ + \beta_{10} afroameričan + u. \end{aligned}$$

Proměnné *GPA* a *GRE* byly vycentrovány svými průměrnými hodnotami. *GRE* se zaokrouhlilo na celá čísla a u *GPA* došlo k zaokrouhlení na jedno desetinné místo. Tedy *GPA* bylo centrováno hodnotou 3.4 a *GRE* 588. Referenční skupinou pro tento druhý model je žena hispánka, která dosáhla průměrného GPA a GRE, hlásí se na nejprestižnější univerzitu a její socioekonomická situace je na nejhorší možné úrovni.

4 Výsledné modely

4.1 Lineární model

Model byl testován na přítomnost heteroskedasticity prostřednictvím Breusch-Paganova testu (p = 0.614). Na 5% hladině významnosti nedošlo k prokázání heteroskedasticity. V modelu je přítomna homoskedasticita. Z tohoto důvodu není nutné model upravit na robustní verzi.

Dále bylo v modelu testováno, jestli se v něm nevyskytuje multikolinearita. Řešeno to bylo prostřednictvím faktoru zvyšujícího rozptyl (VIF). V případě, že hodnota je větší jak 10, tak máme problém s multikolinearitou (Wooldridge, 2016). Na základě dat v tabulce 4.1 je patrné, že v modelu není problém s multikolinearitou.

Tabulka 4.1: VIF

	VIF
GPA_3.4	2.13
úroveň univerzity	1.05
socioekonomická situace studenta	1.03
rasa	1.05
pohlaví	1.01
*GPA_3.4*muž*	2.10

V následující tabulce 4.2 jsou uvedeny koeficienty, směrodatné chyby. Na základě výše uvedených hodnot můžeme říct, že na 1% hladině je statisticky významná *konstanta* a *GPA_3.4*. Upravený koeficient determinace má hodnotu 0,15. Tímto modelem se podařilo vysvětlit 15 % variability.

Tabulka 4.2: Vícenásobný lineární regresní model

	Regresní koeficienty ($\boldsymbol{\beta}$)	**Směrodatná chyba**
konstanta	609.69***	17.92
GPA_3.4	128.21***	20.41
druhá nejvyšší úroveň univerzity	-2.87	16.32
střední úroveň univerzity	-32.22	16.83
velmi špatná úroveň univerzity	-24.75	18.96
střední socioekonomická situace studenta	5.01	13.06
vysoká socioekonomická situace studenta	-12.60	13.24
muž	0.81	13.05
asiat	6.68	13.00
afroameričan	-19.00	10.70
*GPA_3.4*muž*	-17.23	28.14

*Poznámka: *** p <0.01*

4.2 Pravděpodobnostní modely

V modelu je závislou proměnnou binární proměnná *přijat*, která nabývá hodnot 1 (v případě, kdy je student přijat) a 0 (v případě, kdy student přijat není). Z důvodu binární závislé proměnné je pro odhad regresních koeficientů využita logistická regrese (neboli logit) a probit. V tabulce 4.3 jsou uvedené hodnoty regresních koeficientů, směrodatných chyb a dále hodnoty průměrných mezních efektů.

Tabulka 4.3: Porovnání regresních koeficientů a průměrných mezních efektů

	Regresní koeficienty ($\boldsymbol{\beta}$)		**Průměrné mezní efekty (AME)**	
	logit	probit	logit	probit
GRE_588	0.002**	0.001**	0.001*	0.001*
	(0.001)	(0.001)	(0.001)	(0.001)
GPA_3.4	0.814**	0.489**	0.157*	0.157*
	(0.336)	(0.200)	(0.067)	(0.063)
druhá nejvyšší úroveň univerzity	-0.712**	-0.435**	-0.130*	-0.133*
	(0.321)	(0.197)	(0.055)	(0.056)
střední úroveň univerzity	-1.361***	-0.822***	-0.238***	-0.242***
	(0.350)	(0.211)	(0.052)	(0.053)
velmi špatná úroveň univerzity	-1.581***	-0.949***	-0.248***	-0.254***
	(0.423)	(0.248)	(0.049)	(0.050)
střední socioekonomická situace studenta	-0.135	-0.079	-0.026	-0.025
	(0.276)	(0.166)	(0.053)	(0.053)
vysoká socioekonomická situace studenta	-0.258	-0.161	-0.049	-0.051
	(0.285)	(0.170)	(0.053)	(0.053)
muž	-0.192	-0.108	-0.037	-0.035
	(0.230)	(0.138)	(0.044)	(0.044)
asiat	-0.486*	-0.298*	-0.091	-0.094
	(0.283)	(0.168)	(0.051)	(0.051)
afroameričan	-0.313	-0.186	-0.059	-0.059
	(0.230)	(0.165)	(0.051)	(0.051)
konstanta	0.566	0.339		
	(0.365)	(0.222)		

Poznámka: v závorkách jsou uvedené směrodatné chyby, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$

5 Hypotézy

5.1 Hypotéza 1

H1: Student s vyšším GPA získaná více bodů ze zkoušky GRE.

Tato hypotéza vychází z úvahy, že studenti, kteří mají lepší průměr známek, získají i vyšší počet bodů ze zkoušky GRE. Opora pro tuto hypotézu se nachází v tabulce 4.2. Koeficient u *GPA_3.4* je kladný. Můžeme tedy říct, že pokud student zvýší své GPA, tak u zkoušky GRE získá větší počet bodů.

5.2 Hypotéza 2

H2: U studentů s průměrným GPA se velikost získaných bodů z GRE pro muže a ženy významně liší.

Tato hypotéza je založena na úvaze, že muži dosahují lepších výsledků u logických testů (zkouška GRE je založena na podobném principu jako logické testy). Referenční kategorií v modelu je žena. Na základě výsledků v tabulce 4.2 můžeme říct, že rozdíl v počtu bodů ze zkoušky GRE mezi ženou a mužem s průměrným GPA se na 5% hladině významnosti se významně neliší ($p = 0.27$).

5.3 Hypotéza 3

H3: Při vyšším GPA existuje větší pravděpodobnost, že student bude přijat na univerzitu či college.

Tato hypotéza byla postavena na podobné úvaze jako *H1*. Tedy, že studenti s lepším průměrem mají větší šanci být přijati na univerzitu či college. Z tabulky 4.3 je patrné, že modely binární volby pro tuto hypotézu poskytují oporu. Můžeme říct, že se zvyšujícím se GPA se zvyšuje pravděpodobnost přijetí. Na 5% hladině významnosti je výše GPA významná.

5.4 Hypotéza 4

H4: Socioekonomické prostředí, ze kterého studenti pochází, má významný vliv na přijetí na univerzitu či college.

Tato hypotéze byla postavena na základě úvahy, že socioekonomické prostředí studentů má signifikantní vliv na přijetí. Pro tuto hypotéze v datech není podpora na žádné rozumné hladině významnosti (p = 0.66). Tento fakt není překvapivý, protože zahraniční studenti, kteří se hlásí na univerzitu či college do USA, nejsou (ani nemohou) být penalizování za prostředí ze kterého pocházejí.

Závěr

Cílem semestrální práce bylo prozkoumat faktory které ovlivňují počet získaných bodů ze zkoušky GRE a prozkoumat faktory, které ovlivňují pravděpodobnost přijetí zahraničních studentů na univerzitu či college v USA. Na základě dat byla nalezena opora pro $H1$ a $H3$. Nepodařilo se najít oporu pro $H2\ a\ H4$.

Zdroje

ETS. The GRE Tests. In: *ETS Home* [online]. © 2021 [cit. 28.05.2021]. Dostupné z: https://www.ets.org/gre

Hindls, R., Arltová, M., Hronová, S., Malá, I., Marek, L., Pecáková, I., & Řezanková, H. (2018). *Statistika v ekonomii.* Professional Publishing,

Kaggle Inc. Admission | Kaggle. In: *Your Machine Learning and Data Science Community* [online]. © 2021 [cit. 28.05.2021]. Dostupné z: https://www.kaggle.com/eswarchandt/admission

Raslová, L. (2021). *Predikce cen diamantů* [semestrální práce]. VŠE.

Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach.* Nelson Education.