Slovenská technická univerzita v Bratislave

Fakulta informatiky a informačných technológií

Vplyv vzdelania na príjem

Dokumentácia

Peter Smreček

AIS ID: 103130

E-mail: xsmrecek@stuba.sk

Martin Schön

AIS ID: 103121

E-mail: xschon@stuba.sk

Predmet: Programovanie pre dátovú vedu

Deň a čas cvičenia: Streda 17:00

Semester: ZS 21/22

Ročník: 3.

Úvod

V súčasnej dobe sa zvyšuje platové ohodnotenie tradičných, remeselných, prác. Mnoho ľudí, preto zvažuje, či pokračovať v štúdiu, alebo ísť do praxe. Hlavnou motiváciou mladých ľudí pri voľbe budúcej kariéry je často práve platové ohodnotenie. Preto sme sa rozhodli skúmať závislosť výšky platu od rôznych atribútov. Všeobecným predpokladom je, že inteligencia (bežne meraná pomocou výšky IQ) súvisí so vzdelaním. Vysoké školy strácajú svoju prestíž, v súčasnej dobe sa zdá, že titul môže mať každý a jeho vlastníctvo sa už nemusí automaticky spájať s výhodnejšími platovými podmienkami zamestnanca. Súčasná doba sa čím ďalej tým viac javí orientovaná na praktiktické znalosti a skúsenosti z praxe.

Napriek tomuto všetkému ale verejným predpokladom ostáva, že dlhšie vzdelanie by malo znamenať lepšiu možnosť zárobku. Taktiež je očakávané, že výška IQ súvisí s dĺžkou vzdelania a teda tým nepriamo vplýva aj na výsledné peniaze, ktoré je človek schopný svojou činnosťou zarobiť. V tomto článku budeme teda pomocou štatistických metód skúmať, či platia tieto predpoklady alebo či sa už naozaj viac oplatí pracovať v sektore s nižším požadovaným vzdelaním. Rozhodli sme sa skúmať túto tématiku, pretože nás ako študentov zaujímajú naše vyhliadky do budúcna a radi by sme konštruktívne vedeli zhodnotiť, aký úžitok nám ďalšie roky vzdelávania vedia priniesť. Za vhodné považujeme uviesť, že pri našom prieskume využívame jazyk R.

Obsah dát

Dátaset

Pre potreby našej práce sme si zvolili dátaset obsahujúci výšky platov a rôzne iné údaje o osobách, ktoré ich poberajú. Set je verejne dostupný, prípadný záujemca ho vie nájsť na nasledovnej adrese: https://www.kaggle.com/saadsikander/wages-as-per-education.

Základné deskriptívne štatistiky

Súbor dát obsahuje 17 rôznych stĺpcov - 17 rôznych atribútov o jednotlivých záznamoch - ľuďoch, ktorí sú reprezentovaní v dátasete. Atribútmi sú nasledovné informácie: (informácie významné pre naše potreby sú znázornené **hrubým** textom):

1. **wage** - mesačný príjem osoby v amerických dolároch
2. **hours** - priemerný počet odrobených hodín týždenne
3. **IQ** - počet bodov inteligenčného kvocientu
4. KWW - hodnotenie znalosti pracovného sveta (vyjadrené v porovnávacom bodovaní)
5. **educ** - roky venované vzdelávaniu
6. exper - roky pracovných skúseností
7. tenure - roky strávené u súčasného zamestnávateľa
8. age - vek v rokoch
9. married - hodnota 1, ak je osoba ženatý/ vydatá, inak 0
10. black - hodnota 1, ak je osoba čiernej pleti, inak 0
11. south - hodnota 1, ak osoba žije na juhu, 0 na severe
12. urban - hodnota 1, ak osoba žije v metropolitnej štatistickej oblasti (oblasti s najmenej jednou urbanizovanou oblasťou s minimálnym počtom obyvateľov 50 000)
13. sibs - počet súrodencov
14. brthord - znázorňuje koľké dieťa svojich rodičov je dotyčná osoba
15. meduc - počet rokov vzdelávania matky
16. feduc - počet rokov vzdelávania otca

17. lwage - prirodzený logaritmus výšky mesačného platu

Ako je vidieť, máme množstvo naozaj zaujímavých vlastností. Tento dátaset by bol teda výborný aj na hlbšiu analýzu vplyvu faktorov na platy zamestnancov. Nás ale aktuálne zaujíma faktor inteligencie a vzdelania - a tomu sa budeme aj venovať.

Počet riadkov (teda jednotlivých záznamov) tohto dátasetu je 935. Teda tento set je pomerne rozsiahly a mal by byť dostatočne veľký pre overenie nami skúmaných vlastností.

Číselný rozbor určitých atribútov tabuľky

Wage
Najmenšia hodnota nachádzajúca sa v stĺpci wage je 115($). Naopak, najvyššou hodnotou platu je 3078. Toto sú ale okrajové hodnoty zárobkov - dokazujú to aj hodnoty mediánu (905) a priemeru (957.9) platov.

Hours
Čo sa týka priemerne odpracovaných hodín týždenne, najmenší počet dosahuje hodnotu 20, zatiaľ čo najviac pracujúci človek odpracoval 80 hodín týždenne. Opäť ale vidíme, že priemer a medián sú oveľa očakávanejšie - s veľkosťou 40 hodín v mediáne, respektíve 43.93 priemerne.

IQ
Ako je očakávané, priemer IQ sa hýbe okolo 100 - konkrétne sa jedná o 101.3 IQ bodov priemeru, alebo 102 bodov, ak rozprávame o mediáne. IQ je rozdelené po celom spektre - s najmenšou hodnotou počtu 50 a najvyššou 145.

Educ
Počet rokov vzdelania je spodne ohraničený hodnotou 9, zatiaľ čo najdlhšie študujúci človek nášho dátasetu venoval vzdelaniu presne dvakrát takú dlhú dobu - teda 18 rokov. Medián dĺžky vzdelania sa nachádza na hodnote 12, zatiaľčo priemer je trochu vyšší - 13.47 rokov.

Age
Hoci priamo nevyužívame atribút veku v našich výskumoch, je dobré si povedať, že sa jedná o skupinu ľudí medzi 28. a 38. rokom života. Priemerný vek v skupine je niečo málo nad 33 rokov.

Black
Porovnávali sme aj podiel ľudí čiernej ľudí. V dátasete je táto hodnota 12.83%, pričom toto percento v celej USA je 12.1%. Odchýlka je malá - vidíme teda, že dáta vyzerajú byť dostatočne kvalitné a aj táto hodnota nám pomáha veriť tomu, že zber dát bol riadne vykonaný.

Pridávanie stĺpcov

V dátasete máme stĺpec označujúci mesačný zárobok osoby (wage) a stĺpec s priemerným počtom odrobených hodín za týždeň (hours). Z týchto 2 veličín sme si vytvorili novú veličinu, hodinovú mzdu. Vypočítali sme ju ako podiel mesačného zárobku a štvornásobku priemeru odrobených hodín (stĺpec hours) týždenne. Hypotézy budeme formulovať v spojitosti s touto novou veličinou.

Z čistej mesačnej mzdy bez počtu odrobených hodín nedokážeme spoľahlivo určiť finančné ohodnotenie osoby. Mesačný zárobok ľudí je najlepšie smerodajný iba v kontexte odrobených hodín. To, ako je niekto platovo ohodnotený vidíme najmä z hodinovej mzdy a preto ju používame v našich hypotézach.

Chýbajúce a vychýlené hodnoty

V pôvodnom dátasete boli chýbajúce hodnoty reprezentované bodkami, pre jednoduchšiu prácu sme bodky nahradili NA. Chýbajúcich hodnôt je 355 a aj to len v stĺpcoch brthord, meduc a feduc. Hoci existuje viacero techník, ako sa zbaviť chýbajúcich hodnôt, ako je napríklad ich odstránenie, či nahradenie mediánom, priemerom, alebo kNN, rozhodli sme sa ich ignorovať. Tieto chýbajúce hodnoty by nemali ovplyvňovať naše hypotézy, keďže sa ich netýkajú. Preto tieto dáta nenahrádzame.

Vychýlené hodnoty možno vidieť pri stĺpcoch wages, hours, IQ a KWW. Keďže sa jedná o skutočné dáta skutočných ľudí, nebudeme ich odstraňovať, ani nahrádzať hraničnými hodnotami rozdelenia. Predsa len, ide o hodnoty ktoré v skutočnom svete môžu existovať. Predpokladáme, že nejde o chyby v dátach, ale len o bežné anomálie v ľudskej spoločnosti.

Hypotézy a ich overenie

Pre potreby nášho projektu sme si sformulovali 3 hypotézy. Overujeme ich lineárnym regresným modelom.

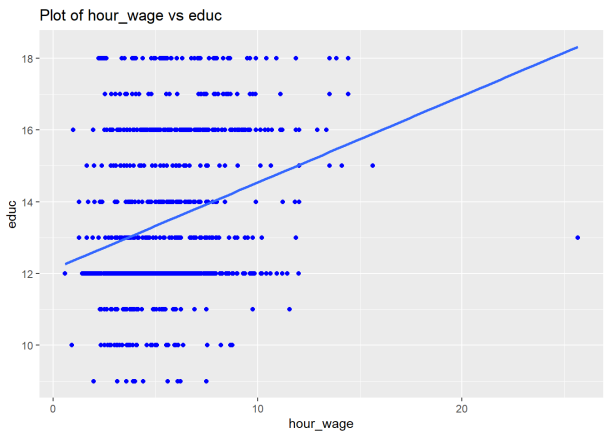
Regresia je vyjadrená vzťahom $Y \approx \beta_0 + \beta_1 X$. Dôležitejším pre nás je koeficient β_1, keďže určuje sklon priamky. Zisťujeme či má priamka sklon, alebo nie. Pre overenie hypotéz je nutné určiť hodnoty koeficientov.

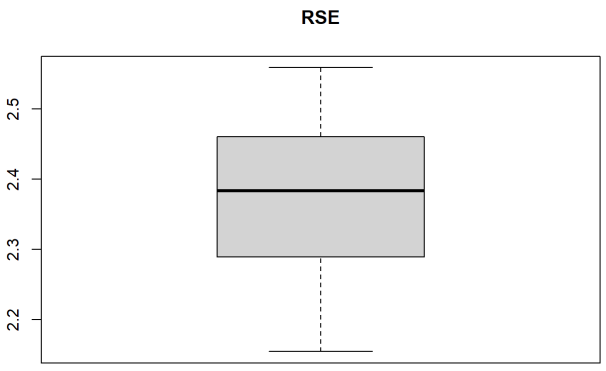
Hypotéza 1

Existuje vzťah medzi výškou hodinovej mzdy a dĺžky vzdelania?
H0: Vzťah medzi výškou hodinovej mzdy a dĺžky vzdelania **neexistuje**.
H1: Vzťah medzi výškou hodinovej mzdy a dĺžky vzdelania **existuje**.

Na základe regresie nad celou množinou dát môžeme vidieť, že priamka má výrazne stúpajúci sklon, čo značí, že existuje vzťah medzi výškou hodinovej mzdy a dĺžkou vzdelania. Zároveň ale vidíme, že dáta sú rozptýlené.

Krížovou validáciou overíme stabilitu modelu. Z pôvodnej množiny dát vytvoríme 100





podmnožín o veľkosti 50% pôvodnej množiny.
Následne sme vypočítali smerodajné odchýlky koeficientov. Pre β_0 nám vyšla smerodajná odchýlka 0.4577959 a pre β_1 0.03469156. Priemerná β_0 v 100 podmnožinách bola 1.49729. β_0 v celom modeli bola 1.467454. Priemerná β_1 v 100

podmnožinách nám vyšla 0.3030743. β_1 v celom modeli bola 0.3053812. Ako vidíme, hodnoty v podmnožinách sú oproti celému modelu približne zhodné, čo značí, že model je stabilný. RSE pre celý model nadobúdal hodnotu 2.378. Z boxplotu RSE pre podmnožiny vidíme, že priemerná RSE je približne rovnaká ako pre podmnožiny, tak aj pre celú množinu.

Výsledky sme overili aj t-testom. Z toho jasne vyplýva, že zamietame H0 v prospech H1. Platí teda H1: Vzťah medzi výškou hodinovej mzdy a dĺžky vzdelania **existuje**.

Hypotéza 2

Existuje vzťah medzi výškou hodinovej mzdy a IQ?
H0: Vzťah medzi výškou hodinovej mzdy a IQ **neexistuje**.
H1: Vzťah medzi výškou hodinovej mzdy a IQ **existuje**.

Analogicky k Hypotéze 1 sme overili aj Hypotézu 2. Postupovali sme rovnako ako pri overovaní Hypotézy 1. Pre zachovanie dobrej čitateľnosti dokumentu postup neuvádzame, je obsiahnutý v priloženom R notebooku.

Na základe krížovej validácie ktorú sme urobili vidíme, že model je stabilný. Z toho jasne vyplýva, že zamietame H0 v prospech H1. Platí teda H1: Vzťah medzi výškou hodinovej mzdy a IQ existuje. Výsledky sme overili t-testom.

Hypotéza 3

Existuje vzťah medzi IQ a dĺžkou vzdelania?
H0: Vzťah medzi IQ a dĺžkou vzdelania **neexistuje**.
H1: Vzťah medzi IQ a dĺžkou vzdelania **existuje**.

Analogicky k Hypotézam 1 a 2 sme overili aj Hypotézu 3. Postupovali sme rovnako ako pri overovaní prvých dvoch Hypotéz. Pre zachovanie dobrej čitateľnosti dokumentu postup neuvádzame, je obsiahnutý v priloženom R notebooku.

Na základe krížovej validácie ktorú sme urobili vidíme, že model je stabilný. Z toho jasne vyplýva, že zamietame H0 v prospech H1. Platí teda H1: Vzťah medzi IQ a dĺžkou vzdelania existuje. Výsledky sme overili t-testom.

Záver

Z našej analýzy dát vyplýva, že i v dnešnej dobe stále existuje štatisticky významné prepojenie medzi dĺžkou štúdia jednotlivca, jeho IQ a jeho finančným ohodnotením. Rovnako tak sme preukázali, že aj IQ zvyčajne súvisí s dĺžkou štúdia.

Lineárnym regresným modelom sme overili 3 hypotézy. Dáta neboli ideálne, jednalo sa prevažne o diskrétne veličiny. Dátaset ale iné veličiny neposkytoval. Lineárnou regresiou sme dokázali vzťahy medzi veličinami, určili RSS a RSE aj rozdiely medzi koeficientami β_0 a β_1. Keďže sa jednalo o prevažne kategorické dáta s malým počtom unikátnych hodnôt, výsledky regresií nemusia na 100% odrážať realitu.

Prílohy

R notebook s riešením - Schon_Smrecek_Project_R.Rmd
HTML súbor s výstupom notebooku - Schon_Smrecek_Project_R.html
Dátaset - Wages.csv