SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE
Fakulta informatiky a informačných technológií

PROGRAMOVANIE PRE DÁTOVÚ VEDU

Projekt 1

Vplyv zdravej výživy na priebeh ochorenia COVID-19

Natália Šeďová
AIS ID: 97062
Tamara Janotková
AIS ID: 110809
Akademický rok: 2021/2022

V dnešnej dobe sa svet snaží vysporiadať s pandémiou ochorenia COVID-19. Vedecká obec prišla už s niekoľkými vakcínami proti tomuto ochoreniu, čo sa prejavilo znížením denného prírastku nakazených v niektorých krajinách. Je však skutočnosťou, že pandémia stále pretrváva a s blížiacim sa chrípkovým obdobím, bude množstvo chorých ľudí len rásť. Ak sa nakazíme, ako aspoň čiastočne zabrániť zlému priebehu ochorenia, ktoré môže v horších prípadoch viesť až k smrti? Je možné svojimi stravovacími návykmi ovplyvniť priebeh ochorenia COVID-19?

V našej štúdii sme sa zamerali na analýzu tabelárnych dát s názvom COVID-19 Healthy Diet Dataset. Dataset obsahuje merania 32 premenných zo 170 krajín sveta. Jednotlivé hodnoty meraní sú vyjadrené v percentách.

Na základe zvolených dát sme stanovili nasledovné hypotézy:

1. Obézni ľudia majú sklon podľahnúť ochoreniu COVID-19.
2. Konzumácia ovocia a zeleniny pomáha ku uzdraveniu z ochorenia COVID-19.

Z hľadiska datasetu, jeho väčšia časť je venovaná stravovacím návykom. V jednotlivých premenných zobrazuje percentuálny podiel konzumovaných potravín v kilogramoch ako napríklad mäso, vajcia, obilniny, mlieko, koreniny a podobne. Pre zvolené hypotézy z tejto časti dát sú kľúčové premenné obezita, ovocie a zelenina.

Zvyšná časť je zameraná na ochorenie COVID-19. Obsahuje potvrdené prípady, úmrtia, množstvo vyliečených pacientov, aktuálne aktívnych pacientov a celkovú populáciu danej krajiny. Hypotézy sa týkajú premenných o percentuálnej úmrtnosti a počte vyliečených a tiež aj o celkovej populácii danej krajiny.

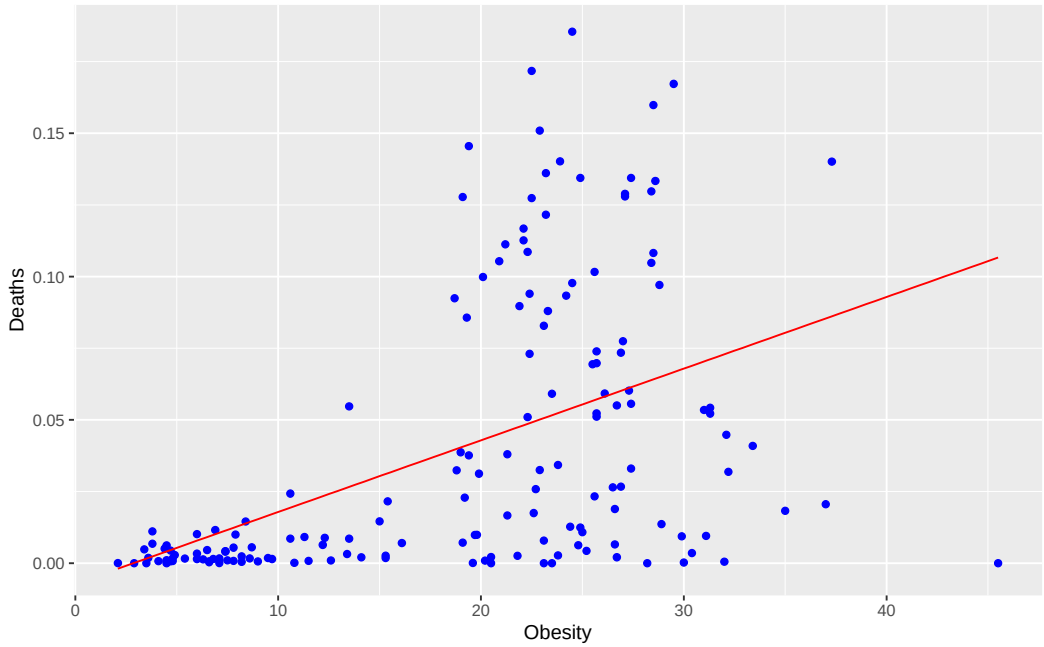
Hypotéza 1: Obézni ľudia majú sklon podľahnúť ochoreniu COVID-19.

Pre hypotézu č. 1 sú kľúčové premenné obezita (Obesity), krajina (Country), populácia (Population) a percentuálna úmrtnosť (Deaths). Merania premennej obezita obsahujú hodnoty NA v troch meraniach. V prípade premennej o úmrtnosti bolo zaznamenaných 6 NA hodnôt. Keďže hodnoty NA pre túto analýzu nemajú výpovednú lehotu, po ich odstránení bolo získaných 163 meraní, na ktorých bola ďalšia analýza vykonaná.

Hodnoty premenných obezita a úmrtnosť sú uvedené v dátovom type double, nakoľko ide o percentuálnu mieru uvádzanú v tvare desatinného čísla. V meraniach premennej obezita sa hodnoty pohybujú od minima 2,1% identifikovaného v krajine Vietnam až po maximum 45,5% v krajine Samoa. Priemerná hodnota obezity v meraniach dosahuje 18,7%. Stredná hodnota je 21,3%.

V meraniach premennej úmrtnosti bolo zaznamenané minimum 0 úmrtí v 8 krajinách, zatiaľ čo maximum 0.185% z celkovej populácie danej krajiny bolo zaznamenané v krajine Belgicko. V priemere krajiny dosahujú 0,039611% úmrtnosť a stredná hodnota meraní premennej úmrtnosti je 0,012443%. Medzi krajinami s nulovou úmrtnosťou sa vyskytla aj krajina Samoa, ktorá bola spomenutá ako krajina dosahujúca maximálnu mieru obezity spomedzi ostatných krajín.
Analýza hypotézy spočívala v nájdení lineárneho regresného modelu pre zvolené dáta a jeho následné overenie.

Na zvolených dátach bol identifikovaný lineárny model, ktorý je zobrazený na grafe č.1 v podobe červenej priamky. Tento model vyjadruje vzťah kauzality medzi premennými obezita a úmrtia. Nárast počtu obéznych ľudí spôsobuje nárast úmrtnosti populácie na ochorenie COVID-19. Inými slovami, človek, ktorý je obézny môže mať pridružené aj rôzne iné ochorenia, a tak sa stáva zraniteľnejší voči ochoreniu.



Graf 1 Lineárny model hypotéza č. 1

Z hľadiska korelácie v tomto prípade je taktiež identifikovaný vzťah medzi obezitou a úmrtnosťou, kedy s rastúcou obezitou v populácii, rastie aj úmrtnosť.

Na overenie vhodnosti modelu bol použitý prístup krížnej validácie dát. Celé dáta boli rozdelené na menšie podskupiny, ktoré obsahovali niekoľko percent meraní z celej množiny dát v náhodnom poradí. Našim cieľom bolo ukázať, že ak má platiť hypotéza č. 1, musí platiť aj na náhodne zvolených, menších úsekoch dát.

Zo všetkých podskupín sme vybrali koeficienty pre im prislúchajúci lineárny model, ako aj reziduály. Po vypočítaní štandardnej chyby pre koeficienty β_0 a β_1 sme zistili nasledovné.

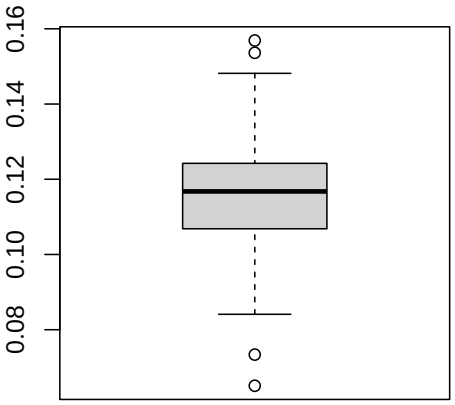
Hodnota štandardnej chyby pre koeficient β_0 bola 0,004943418. Vyjadruje o koľko sa hodnoty β_0 koeficientov v jednotlivých podskupinách líšili od β_0 koeficientu modelu. Ak hodnota koeficientu v modeli bola –0,0071585, hodnoty sa vo veľkej miere menili. Z hľadiska získaných β_0 koeficientov môžeme potvrdiť, že vždy išlo o zápornú hodnotu.

Čo sa týka koeficientu β_1, štandardná chyba bola 0,0003886894. V modeli bol β_1 koeficient rovný 0,0025009. V tomto prípade získanej štandardnej chyby nedochádzalo ku markantným výkyvom ako v prípade koeficientu β_0. Vo všetkých podskupinách bol identifikovaný β_1 koeficient ako kladné číslo, čo znamená, že vždy šlo o lineárny rast. Z toho vyplýva, že vo všetkých podskupinách sa potvrdilo, že sú premenné priamoúmerné- ak rastie jedna, rastie aj druhá. Existujúci vzťah medzi premennými bol potvrdený aj pomocou t-testu, ktorý určil 95% spoľahlivosť.

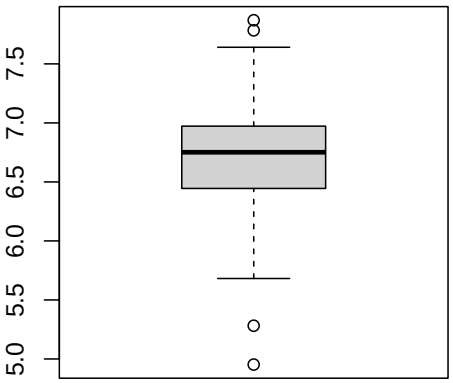
Pri jednotlivých podskupinách boli identifikované reziduály. Ide o hodnoty, ktoré vyjadrujú chybovosť- teda, o koľko sa reálna hodnota líši od hodnoty, ktorá danému meraniu prislúcha podľa identifikovaného modelu. Použili sme výpočet RSE a RSS (zobrazené nižšie) pre určenie chybovosti modelu.

V krabicovom grafe vľavo môžeme pozorovať rozloženie hodnôt reziduálov vo všetkých podskupinách. Body umiestnené mimo lúčov vedúcich zo štvorca predstavujú odľahlé hodnoty („outliers“). Vyjadrujú nekonzistenciu modelu- teda, že medzi danými hodnotami boli aj také, ktoré sa vymykajú z normálu a po ich odstránení by výpočet lepšie sadol na dané dáta. Ďalej si v grafe môžeme všimnúť hrubú čiaru uprostred štvorca predstavujúcu medián. Ak by bola čiara presne uprostred štvorca, išlo by o dáta, ktoré sú v normálnom rozdelení. Rozptyl lúčov určuje celkový rozptyl dát. V tomto prípade, ide o nerovnomerne rozptýlené reziduály a veľký rozptyl, čo predstavuje horší „fit“ modelu na dané dáta.

Krabicový graf napravo sa týka štandardným chybám reziduálov. Podobne ako v predchádzajúcom prípade, sa v grafe nachádzajú odľahlé hodnoty. Takisto, aj tu je rozptyl lúčov pomerne veľký, čo vyjadruje, že chyby reziduálov sú z väčšieho intervalu hodnôt a teda aj reziduály boli niekedy malé, inokedy veľké.



Graf 2 Residual sum of squares (RSS)



Graf 3 Residual standard error (RSE)

Zo zistených skutočností bol potvrdený vzťah medzi obezitou a úmrtnosťou a tiež bola potvrdená hypotéza, avšak pre ďalšie analýzy by bolo vhodnejšie zvoliť iný model pre zvolené dáta ako lineárny.

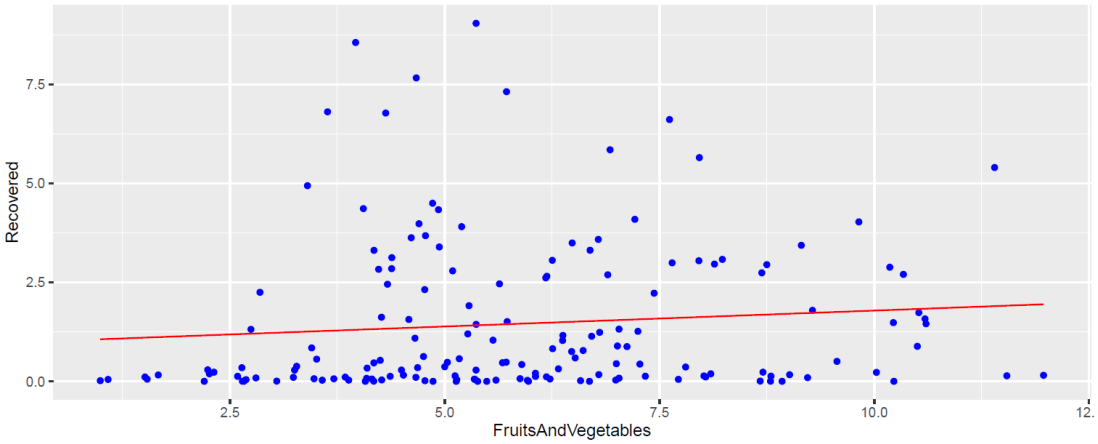
Hypotéza 2: Konzumácia ovocia a zeleniny pomáha ku uzdraveniu z ochorenia COVID-19.

Pre hypotézu č. 2 sú kľúčové premenné krajina (Country), ovocie (Fruits – Excluding Wine), zelenina (Vegetables), populácia (Population) a uzdravení (Recovered). Vzhľadom na našu hypotézu, potrebujeme zjednotiť stĺpce ovocie a zelenina. Dané stĺpce si zjednotíme jednoduchým spôsobom. Spočítame ich a vydelíme dvomi a dáme ich do stĺpca ovocie a zelenina (FruitsAndVegetables). Ďalej si musíme z daných stĺpcov odstrániť hodnoty NA, nakoľko tieto hodnoty pre nás nemajú výpovednú lehotu. Po odstránení daných hodnôt nám zostalo 164 meraní, na ktorých bola vykonaná analýza.

Hodnoty premenných ovocie a zelenia a uzdravení sú uvedené v percentuálnej miere, ktorá je uvádzaná v desatinných číslach, a preto tieto stĺpce majú dátový typ double. V meraniach premennej ovocie a zelenina sa hodnoty pohybujú od minima 0,99% identifikovaného v krajine Čad až po maximum 11,97% v krajine Dominika. Priemerná hodnota spotreby ovocia a zeleniny v meraniach dosahuje 5,84%. Stredná hodnota je 5,26%.

V meraniach premennej uzdravení bolo zaznamenané minimum 0 úmrtí v 4 krajinách, zatiaľ čo maximum 9,04 % z celkovej populácie danej krajiny bolo zaznamenané v krajine Čierna Hora. V priemere krajiny dosahujú 1,45% uzdravenie a stredná hodnota meraní premennej uzdravenia je 0,48%.

Hypotézu sme analyzovali prostredníctvom lineárneho regresného modelu pre zvolené dáta, a jeho následným overením. Na zvolených dátach bol identifikovaný lineárny model, ktorý je zobrazený na grafe 4 v podobe červenej priamky. Na grafe 4 môžeme vidieť závislosť premenných Recovered a FruitsAndVegetables. Na základe grafu môžeme povedať, že ovocie a zelenina zvyšujú možnosť uzdravenia, aj keď to nie je až také signifikantné.



Graf 4 Lineárny model hypotéza č. 2

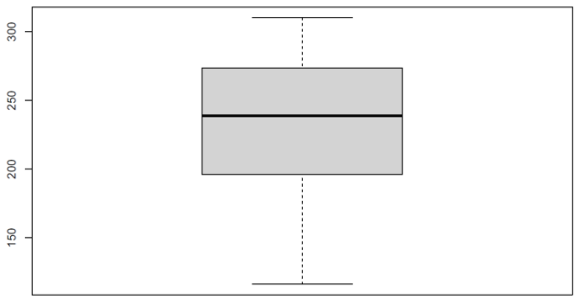
Na overenie vhodnosti modelu bol použitý prístup krížnej validácie dát. Zo všetkých podskupín sme vybrali reziduály a aj koeficienty pre im prislúchajúci lineárny model. Po vypočítaní štandardnej chyby pre koeficienty β_0 a β_1 sme zistili nasledovné.

Hodnota štandardnej chyby pre koeficient β_0 bola 0,4191856. Hodnota koeficientu β_0 je 0,9827556. Z hľadiska získaných β_0 koeficientov môžeme vidieť, že všetky hodnoty sú kladné.

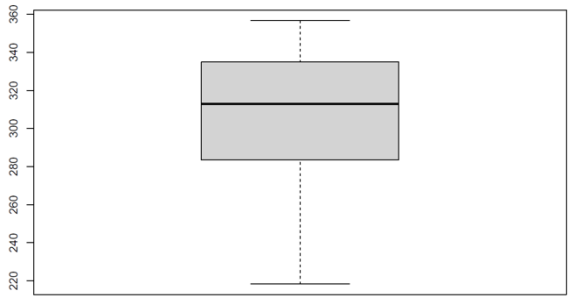
Hodnota štandardnej chyby pre koeficient β_1 bola 0,06845252. Hodnota koeficientu β_1 je 0,05775505. Zo získaných β_1 koeficientov môžeme vidieť, že sa tam nachádzajú aj záporné hodnoty a teda nemôžeme jednoznačne potvrdiť, že sa jedná o lineárny rast. Avšak, podiel záporných hodnôt je veľmi malý. Existujúci vzťah medzi premennými bol potvrdený aj pomocou t-testu, ktorý určil 95% spoľahlivosť.

Grafy 5 a 6, ktoré sú zobrazené nižšie, sú krabicové grafy s hrubou čiarou takmer uprostred. Táto čiara predstavuje medián nameraných dát. Môžeme vidieť, že ani na jednom grafe nie je čiara uprostred štvorca a teda nejedná sa o normálne rozdelenie. Rozptyl lúčov je pomerne veľký, a môžeme povedať, že chyby reziduálov sú z väčšieho intervalu hodnôt. Tieto zistenia značia, že daný model nie je úplne idálny pre zobrazenie dát.

V grafe (graf 5) môžeme pozorovať rozloženie hodnôt reziduálov vo všetkých podskupinách. Krabicový graf napravo (graf 6) sa týka štandardným chybám reziduálov.



Graf 5 Residual sum of squares (RSS)



Graf 6 Residual sum of squares (RSE)

Zo zistených skutočností bol potvrdený vzťah medzi konzumáciou zeleniny a ovocia a počtu uzdravených ľudí. Hypotéza bola potvrdená, avšak pre ďalšie analýzy je vhodnejšie zvoliť iný model pre dané dáta.