Ekonometria: Semestrálny Projekt

Priemerná dĺžka života

Rudolf Nosek, 3DAV

23. apríla 2023

Dáta

Budeme analyzovať dáta o priemernej dĺžke života v 193 krajinách počas obdobia 15-tich rokov. Dáta pochádzajú zo stránky Kaggle(https://www.kaggle.com/datasets/kumarajarshi/l expectancy-who). Zo zoznamu všetkých možných ukazovateľov sme si na modelovanie problému vybrali nasledovné:

- Alcohol - počet litrov vypitého čistého alhoholu na človeka v danom roku. Predpokladáme zápornú závislosť skrz škodlivosti alkoholu
- BMI - priemerné BMI krajiny v danom roku. Predpokladáme zápornú závislosť, pretože vysoké BMI je ukazovateľom obezity.
- GDP - HDP v dolároch krajiny v danom roku. Predpokladáme kladnú závislosť, bohatšie krajiny budú mať lepšie zdravotníctvo.
- Total.expenditure - percento štátneho rozpočtu investovaného do zdravotníctva. Tiež očakávame kladnú závislosť.
- Schooling - priemerný počet rokov strávený v škole. Očakávame kladnú závislosť, lebo v bohatších krajinách budú ľudia dlhšie chodiť do školy.

Následne sme vyhodili všetky riadky, kde chýbali údaje. Zostala nám tabuľka s 2308 riadkami.

Lineárny model

Priemernú dĺžku života sme sa rozhodli modelovať nasledovne:

$$Life.expectancy = \beta_0 + \beta_1 * Alcohol + \beta_2 * BMI + \beta_3 * GDP + \beta_4 * Total.expenditure + \beta_5 * Schooling + \varepsilon$$

Definujeme si maticu X, ktorá bude mať prvý stĺpec jednotky a ostatné stĺpce budú vektory jednotlivých vysvetlujúcich premenných. Potom odhad vektora β je nasledovný:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

kde vektor Y je vektor hodnôt vysvetľovanej premennej. Po odhade koeficientov nám softvér R ukázal aj p-hodnoty pre test signifikancie jednotlivých vysvetľujúcich premenných. Premenná *Total.expenditure* nám vyšla nedostatočne signifikantná, preto ju z modelu odstránime. Nebudeme sa podrobnejšie zaoberať metódou, ako sme prišli k danému záveru, kedže to nie je cieľom projektu. Pre lepšiu interpretáciu koeficientov transformujeme niektoré vysvetlujúce premenné. Výsledný model je teda nasledovný:

$$\begin{gathered} Life.expectancy = \\ \beta_0 + \beta_1 * Alcohol + \beta_2 * (BMI - \overline{BMI}) + \\ \beta_3 * (GDP - \overline{GDP}) + \beta_4 * (Schooling - \overline{Schooling}) + \varepsilon \end{gathered}$$

Odhadnuté koeficienty sú nasledovné: $\hat{\beta}_0 = 70.2102, \hat{\beta}_1 = -0.1928, \hat{\beta}_2 = 0.1079, \hat{\beta}_3 = 0.0001, \hat{\beta}_4 = 1.7418$. Nultý koeficient nám hovorí aká je priemerná dĺžka života v krajine, kde sa nepije alkohol a kde je priemerná BMI, priemerné HDP a priemerný počet rokov sa chodí do školy. Ostatné koeficienty nám hovoria o tom o koľko sa zmení priemerná dĺžka života ak sa daná premenná zvýši o 1.

Ďalej si odhadneme rozptyl náhodných chýb nasledovne:

$$s^2 = \frac{1}{n-k} \sum_{i=1}^{n} \hat{\varepsilon}_i{}^2$$

kde k je počet parametrov(5), n je počet dátových bodov(2308) a $\hat{\varepsilon}$ je odhad reziduí ($\hat{\varepsilon} = Y - X\hat{\beta}$). Odhad rozptylu chýb je 36.03475, čo je časť rozptylu nevysvetlená modelom, ktorú by sme chceli minimalizovať.

Koeficient determinácie

Koeficient determinácie je štatistika, ktorá nám hovorí o tom aký pomer rozptylu závislej premennej náš model vysvetľuje. Počíta sa nasledovne:

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

kde RSS je suma štvorcov reziduí a TSS je suma štvorcov odchyliek závislej premennej od jej priemeru. Existuje aj upravený koeficient determinácie, ktorý berie do úvahy počet vysvetlujúcich premennych.

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

Náš model má $R^2 = 0.6176763$ a $\overline{R^2} = 0.6170122$, čo znamená, že nás model vysvetľuje približne 62% rozptylu závislej premennej.

Test hypotézy o významnosti regresie

Ďalej môžeme testovať náš model oproti tzv. úbohému modelu:

$$Life.expectancy = \beta_0$$

Naše hypotézy sú:

$$H_0 : \beta_1 = 0 \wedge \beta_2 = 0 \wedge \beta_3 = 0 \wedge \beta_4 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0 \vee \beta_2 \neq 0 \vee \beta_3 \neq 0 \vee \beta_4 \neq 0$$

Zapísané maticovo:

$$H_0 : R\beta = r \quad vs. \quad H_1 : R\beta \neq r$$

$$R = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} r = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Testovú štatistiku vypočítame nasledovne:

$$F = \frac{[R\hat{\beta} - r]^T [R(X^T X)^{-1} R^T]^{-1} [R\hat{\beta} - r]}{qs^2}$$

kde q je počet testovaných podmienok (v našom prípade 4). Pre náš model $F = 930.1728$. Tuto štatistiku keď porovnáme s 95%-ným kvantylom Fisherovho rozdelenia so stupňami voľnosti q a $n - k$ tak zistíme, že naša štatistika je väčšia a preto **zamietame** nulovú hypotézu.

Chow breakpoint test

Jedna z nespomínaných premenných v pôvodnom datasete je premennea *Status*, ktorá rozdeľuje dáta podľa toho či je krajina vyspelá alebo rozvojová. Formálne si vytvoríme dva lineárne modeli, jeden pre vyspelé a druhý pre rozvojové krajiny. Chceme testovať či sa ich koeficienty rovnajú. Dva spomínané modely spojíme do jedného:

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0_{n_1 * k} \\ 0_{n_2 * k} & X_2 \end{bmatrix} \begin{bmatrix} \beta_{(1)} \\ \beta_{(2)} \end{bmatrix} + \varepsilon$$

kde matice s indexom 1 patria modelu pre vyspelé krajiny a s indexom 2 pre rozvojové. Potom test vypadá nasledovne:

$$H_0 : \beta_{(1)} = \beta_{(2)} \quad vs. \quad \beta_{(1)} \neq \beta_{(2)}$$

maticovo:

$$H_0 : R\beta = r \quad vs. \quad H_1 : R\beta \neq r$$

$$R = \begin{bmatrix} I_k & -I_k \end{bmatrix}, \beta = \begin{bmatrix} \beta_{(1)} \\ \beta_{(2)} \end{bmatrix}, r = 0_k$$

k je počet parametrov pôvodného modelu. Použijeme rovnakú testovú štatistiku ako v predchádzajúcej úlohe len tento krát ju porovnáme s 95%-tným kvantilom Fisherovho rozdelenia so stupňami voľnosti k a $n - 2k$. Testová štatistika (45.46596) je väčšia ako spomínaný kvantil (2.217991) preto nulovú hypotézu **zamietame**.

Záver

V prvom rade sme vytvorili model a interpretovali jeho koeficienty. Očakávali sme zápornú závislosť medzi BMI a závislou premennou ale model nám tvrdí opak, náš

odhad prečo to tak je, pretože vo väčšine krajinách stále nie je hlavný problém obezita ale naopak podvýživa. Ďalej nás prekvapila nízka signifikancia premennej *Total.expenditure*. Ukázali sme si ako zistiť rozptyl nevysvetlený modelom a následne ako vypočítať pomer rozptylu vysvetlený modelom (koeficient determinácie). Testovali sme význam regresie, pomocou ktorého sme sa ujistili, že regresia má v našom prípade význam. V posledom rade sme chceli vedieť či sa koeficienty líšia pre vyspelé krajiny od rozvojových. Výsledok testu ukázal, že koeficienty sú naozaj rozdielne, čo znamená, že ak zvýšíme nejakú vysvelujúcu premennú o 1, tak sa v rozvojových krajinách inak zmení priemerná dĺžka života ako vo vyspelých.