



ปัจจัยที่มีผลต่อค่าใช้จ่ายรักษาพยาบาล (Medical Charges)



กíมาและความสำคัญ

ในปัจจุบันค่ารักษาพยาบาลเป็นภาระสำคัญของประชาชนและครอบครัว เนื่องจากโรคภัยไข้เจ็บเกิดขึ้นได้ตลอดเวลา ค่าใช้จ่ายที่สูงอาจส่งผลต่อคุณภาพชีวิต และการวางแผนการเงิน การศึกษาปัจจัยที่ส่งผลต่อค่าใช้จ่ายดังกล่าว จึงมีความสำคัญทั้งในมุมมองของบุคคลทั่วไปและในเชิงนโยบาย เช่น การกำหนดเบี้ยประกันสุขภาพ

สรุปผลที่ได้จากการใช้สต็อก

1. ตัวแปรที่ศึกษา

- ตัวแปรตาม (Dependent Variable, Y) = ค่ารักษาพยาบาล
- ตัวแปรอิสระ (Independent Variable, X) = เช่น อายุ, จำนวนวันนอนโรงพยาบาล, รายได้

2. สมการทดแทนที่ได้ (Regression Equation)

$$Y = \beta_0 + \beta_1 X$$

3. การทดสอบนัยสำคัญ (Hypothesis Testing)

- $H_0 : \beta_0 = 0$ (ตัวแปรอิสระไม่มีผล)
- $H_1 : \beta_1$ ไม่เท่ากับ 0 (ตัวแปรอิสระมีผล)

4. ค่าการอธิบายความแปรปรวน (R^2)

- เช่น ถ้า $R^2 = 0.68$ → แสดงว่า ตัวแปรอิสระที่เลือกสามารถอธิบายความผันแปรของค่ารักษาพยาบาลได้ 68%

5. การประเมินคุณภาพของโมเดล (Model Evaluation)

- ค่า Residuals กระจายแบบสุ่ม → แสดงว่าแบบจำลองมีความเหมาะสม
- ค่า RMSE (Root Mean Square Error) อยู่ในระดับต่ำ → แบบจำลองมีความแม่นยำในการพยากรณ์
- ตรวจสอบ Assumption ของ SLR → พบว่าเป็นไปตามข้อสมมติส่วนใหญ่

จากการวิเคราะห์ด้วยส้นประสิกหรือสันนิษฐานพบว่า ปัจจัยที่นำมาศึกษามีความสัมพันธ์กับค่ารักษาพยาบาลในระดับที่แตกต่างกัน โดยค่า r ที่ได้จะก้อนถึงกิจกรรมและความแน่นแฟ้นของความสัมพันธ์ระหว่างปัจจัยอิสระกับค่ารักษาพยาบาล

ปัจจัยบางตัวมีความสัมพันธ์เชิงบวก หมายความว่าค่ารักษาพยาบาลมีแนวโน้มเพิ่มขึ้นตามการเปลี่ยนแปลงของปัจจัยบัน្ត ในขณะที่บางปัจจัยอาจมีความสัมพันธ์เชิงลบ หรือไม่มีความสัมพันธ์อย่างมีนัยสำคัญ

การวิเคราะห์โดยใช้ The Simple Linear Regression Model แสดงให้เห็นว่าปัจจัยที่นำมาศึกษามีความสัมพันธ์กับค่ารักษาพยาบาล และสามารถนำมาร่างสมการเชิงเส้นเพื่อใช้ในการพยากรณ์ได้ อย่างไรก็ตาม การตีความผลการวิเคราะห์ต้องอยู่ภายใต้เงื่อนไขและข้อสมมติของแบบจำลอง ได้แก่

1. ความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามต้องเป็นเส้นตรง (Linearity)

2. ความคลาดเคลื่อนต้องมีการกระจายตัวแบบคงที่ (Homoscedasticity)

3. ความคลาดเคลื่อนต้องมีการแจกแจงใกล้เคียงปกติ (Normality of errors)

4. ความคลาดเคลื่อนไม่ควรมีความสัมพันธ์กันเอง (Independence of errors)

ภายใต้ข้อสมมติดังกล่าวแบบจำลองทดแทนเชิงเส้นแบบง่ายเชิงสารภาพใช้เพื่อปรับยาและประเมินปัจจัยที่มีผลต่อค่าใช้จ่ายรักษาพยาบาลได้อย่างมีประสิทธิภาพ และช่วยให้การคาดการณ์มีความถูกต้องมากขึ้น

วัตถุประสงค์

- เพื่อศึกษาความสัมพันธ์ระหว่างอายุ (BMI) การสูบบุหรี่ และภูมิภาคที่อยู่อาศัย กับค่าใช้จ่ายรักษาพยาบาล
- เพื่อสร้างโมเดลทางสถิติในการพยากรณ์ค่าใช้จ่ายรักษาพยาบาล

สมมติฐาน

สมมติฐานหลักของการวิจัย

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$

$$H_1 : \beta_i \neq 0$$

สมมติฐานย่อยของการวิจัย

$$H_{01} : \beta_1 = 0$$

$H_{11} : \beta_1 \neq 0 \rightarrow$ อายุ (Age) ไม่มี/มีอิทธิพลต่อค่าใช้จ่ายในการรักษาพยาบาล

$$H_{02} : \beta_2 = 0$$

$H_{12} : \beta_2 \neq 0 \rightarrow$ ดัชนีมวลกาย (BMI) ไม่มี/มีอิทธิพลต่อค่าใช้จ่ายในการรักษาพยาบาล

$$H_{03} : \beta_3 = 0$$

$H_{13} : \beta_3 \neq 0 \rightarrow$ การสูบบุหรี่ (Smoker) ไม่มี/มีอิทธิพลต่อค่าใช้จ่ายในการรักษาพยาบาล

$$H_{04} : \beta_4 = \beta_5 = \beta_6 = 0$$

$H_{14} : \text{อย่างน้อยหนึ่ง } \beta \text{ ของ Region} \neq 0 \rightarrow$ ภูมิภาคมี/ไม่มีอิทธิพลต่อค่าใช้จ่ายในการรักษาพยาบาล

$$H_{05} : \beta_7 = 0$$

$H_{15} : \beta_7 \neq 0 \rightarrow$ การรักษาพยาบาลในโรงพยาบาล

$$H_{06} : \text{charges} \sim \text{Normal}$$

$H_{16} : \text{charges} \neq \text{Normal} \rightarrow$ ความไม่สมมาตรของค่าใช้จ่ายในการรักษาพยาบาล

$$H_{07} : \text{charges} \sim \text{Homoscedastic}$$

$H_{17} : \text{charges} \neq \text{Homoscedastic} \rightarrow$ ความไม่เท่ากันของความแปรปรวนของค่าใช้จ่ายในการรักษาพยาบาล

สมมติฐานของแบบจำลอง

$$H_0 : \text{error} \sim \text{Normal}$$

$H_1 : \text{error} \sim \text{Non-Normal} \rightarrow$ ความไม่สมมาตรของค่าใช้จ่ายในการรักษาพยาบาล

$$H_0 : \text{residuals} \sim \text{Normal}$$

$H_1 : \text{residuals} \sim \text{Non-Normal} \rightarrow$ ความไม่สมมาตรของค่าใช้จ่ายในการรักษาพยาบาล

ประโยชน์ที่จะได้รับ

เข้าใจปัจจัยที่มีอิทธิพลต่อค่ารักษาพยาบาล ซึ่งจะช่วยให้ผู้มีส่วนเกี่ยวข้องสามารถวางแผนด้านสุขภาพและการเงินได้ดีขึ้น และสนับสนุนการตัดสินใจเชิงนโยบายของรัฐและหน่วยงานด้านสุขภาพเป็นข้อมูลเชิงวิเคราะห์สำหรับพัฒนาแบบจำลองการพยากรณ์ที่ช่วยคาดการณ์ค่าใช้จ่ายได้อย่างแม่นยำ

บรรณานุกรม

MOSAPABDEL-GHANY.(2568).Medical Insurance Cost Dataset.สืบค้าเมื่อ 6 ตุลาคม 2568, จาก <https://www.kaggle.com/datasets/mosapabdelghany/medical-insurance-cost-dataset>



ผลลัพธ์

```
> #Linear regression
> names(insurance_clean)
[1] "age"          "bmi"          "smoker_yes"
[4] "region_southeast" "region_southwest" "region_northwest"
[7] "charges"      >
> with(insurance_clean, cor(charges, age, method = "pearson", use = "two.sided"))
[1] 0.8319583
> with(insurance_clean, cor(charges, bmi, method = "pearson", use = "two.sided"))
[1] 0.1533936
> with(insurance_clean, cor(charges, smoker_yes, method = "pearson", use = "two.sided"))
[1] 0.4396389
> with(insurance_clean, cor.test(charges, age, method = "pearson"))
[1] 0.8319583
Pearson's product-moment correlation
data: charges and age
t = 21.099, df = 198, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7856748 0.8702492
sample estimates:
cor
0.8319583
> with(insurance_clean, cor.test(charges, bmi, method = "pearson"))
Pearson's product-moment correlation
data: charges and bmi
t = 2.1843, df = 198, p-value = 0.83012
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.01497123 0.28684699
sample estimates:
cor
0.1533936
> with(insurance_clean, cor.test(charges, smoker_yes, method = "pearson"))
Pearson's product-moment correlation
data: charges and smoker_yes
t = 0.8978, df = 198, p-value = 0.39812
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3264436 0.5451292
sample estimates:
cor
0.4396389
> cor(charges, age, method = "pearson")
Pearson's product-moment correlation
data: charges and age
t = 2.1843, df = 198, p-value = 0.83012
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.01497123 0.28684699
sample estimates:
cor
0.1533936
> cor(charges, bmi, method = "pearson")
Pearson's product-moment correlation
data: charges and bmi
t = 2.1843, df = 198, p-value = 0.83012
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.01497123 0.28684699
sample estimates:
cor
0.1533936
> cor(charges, smoker_yes, method = "pearson")
Pearson's product-moment correlation
data: charges and smoker_yes
t = 0.8978, df = 198, p-value = 0.39812
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3264436 0.5451292
sample estimates:
cor
0.4396389
> cor(charges, region_southeast, method = "pearson")
Pearson's product-moment correlation
data: charges and region_southeast
t = 0.8978, df = 198, p-value = 0.39812
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3264436 0.5451292
sample estimates:
cor
0.4396389
> cor(charges, region_southwest, method = "pearson")
Pearson's product-moment correlation
data: charges and region_southwest
t = 0.8978, df = 198, p-value = 0.39812
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3264436 0.5451292
sample estimates:
cor
0.4396389
> cor(charges, region_northwest, method = "pearson")
Pearson's product-moment correlation
data: charges and region_northwest
t = 0.8978, df = 198, p-value = 0.39812
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3264436 0.5451292
sample estimates:
cor
0.4396389
> cor(charges, log_charges, method = "pearson")
Pearson's product-moment correlation
data: charges and log_charges
t = 0.8978, df = 198, p-value = 0.39812
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3264436 0.5451292
sample estimates:
cor
0.4396389
> cor(charges, region_southeast, region_southwest, method = "pearson")
Pearson's product-moment correlation
data: charges and region_southeast
t = 0.8978, df = 198, p-value = 0.39812
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3264436 0.5451292
sample estimates:
cor
0.4396389
> cor(charges, region_southeast, region_northwest, method = "pearson")
Pearson's product-moment correlation
data: charges and region_northwest
t = 0.8978, df = 198, p-value = 0.39812
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3264436 0.5451292
sample estimates:
cor
0.4396389
> cor(charges, region_southwest, region_northwest, method = "pearson")
Pearson's product-moment correlation
data: charges and region_northwest
t = 0.8978, df = 198, p-value = 0.39812
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3264436 0.5451292
sample estimates:
cor
0.4396389
> #E ~ Normal
> nulldata charges
> insurance_clean$log_charges <- log(insurance_clean$charges)
> shapiro.test(insurance_clean$log_charges)
Shapiro-Wilk normality test
data: insurance_clean$log_charges
W = 0.98326, p-value = 4.07e-10
> qnorm(insurance_clean$log_charges)
> qqline(insurance_clean$log_charges, col = "red")
> hist(insurance_clean$log_charges,
+ main = "Histogram of log(charges)",
+ xlab = "log(charges)",
+ col = "lightgray",
+ border = "white")
> #VIF
> vif(model)
> #Independent
> lmtest::dwtest(model)
Durbin-Watson test
data: model
DW = 1.9411, p-value = 0.3409
alternative hypothesis: true autocorrelation is greater than 0
> #Equality
> lmtest::htest(model)
studentized Breusch-Pagan test
data: model
BP = 4.4711, df = 6, p-value = 0.6132
> car::ncvTest(model)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.168573, Df = 1, p = 0.68138
> #Anderson-Darling normality test
Anderson-Darling normality test
data: residuals(model)
A = 0.67075, p-value = 0.07873
```

ผลการตรวจสอบสมมติฐานทั้งหมดแสดงว่า แบบจำลองดัดแปลงเชิงเส้นนี้ ผ่านการทดสอบเกือบทุกข้อ ยกเว้นตัวแปรตาม (charges) ที่ไม่เป็นปกติในระดับสากล แต่การแยกแขวงของ Residuals มีลักษณะไขกลับปีกตัวอ่อน multicollinearity, autocorrelation หรือ heteroscedasticity จึงสามารถสรุปได้ว่า แบบจำลองนี้มีความเหมาะสม เชื่อถือได้ และสามารถใช้อธิบายความแปรปรวนของค่าใช้จ่ายในการรักษาพยาบาล (charges) โดยเฉพาะตัวแปร age และ smoker_yes ที่มีอิทธิพลมากที่สุดต่อค่าใช้จ่าย

