**Vargas Carrillo Raul Paulino**

**Selección de Variables**

Obtener la información utilizando la función stepAIC de la librería textitMASS de R, deberá adjuntar tanto el código utilizado como los resultados obtenidos.

Utilizando el conjunto de datos personales de gastos médicos (insurance.csv).

charges: Costos médicos individuales facturados por el seguro de salud.

age: Edad del beneficiario principal.

sex: Género del contratante del seguro, femenino, masculino.

bmi: Índice de masa corporal, proporcionando una comprensión de los pesos corporales que son relativamente altos o bajos en relación con la altura, índice objetivo de peso corporal (kg / m2) utilizando la proporción de altura a peso, idealmente 18.5 a 24.9.

children: Número de hijos cubiertos por el seguro de salud / Número de dependientes.

smoker: Fumador.

region: Área residencial del beneficiario en los EE. UU., noreste, sureste, suroeste, noroeste.

Complete las siguientes actividades:

1. **Estime el modelos de regresión lineal empleando como variable dependiente el gasto médicos individual (charges) y el resto de variable como variables independientes.**

> modinsurance <- lm(charges ~ ., insurance)

> summary(modinsurance)

Call:

lm(formula = charges ~ ., data = insurance)

Residuals:

Min 1Q Median 3Q

-11304.9 -2848.1 -982.1 1393.9

Max

29992.8

Coefficients:

Estimate Std. Error

(Intercept) -11938.5 987.8

age 256.9 11.9

sexmale -131.3 332.9

bmi 339.2 28.6

children 475.5 137.8

smokeryes 23848.5 413.1

regionnorthwest -353.0 476.3

regionsoutheast -1035.0 478.7

regionsouthwest -960.0 477.9

t value Pr(>|t|)

(Intercept) -12.086 < 2e-16 \*\*\*

age 21.587 < 2e-16 \*\*\*

sexmale -0.394 0.693348

bmi 11.860 < 2e-16 \*\*\*

children 3.451 0.000577 \*\*\*

smokeryes 57.723 < 2e-16 \*\*\*

regionnorthwest -0.741 0.458769

regionsoutheast -2.162 0.030782 \*

regionsouthwest -2.009 0.044765 \*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05

‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom

Multiple R-squared: 0.7509, Adjusted R-squared: 0.7494

F-statistic: 500.8 on 8 and 1329 DF, p-value: < 2.2e-16

1. **Obtenga el mejor modelo utilizando el método de selección de variables hacia adelante, empleando como criterio el AIC.**

> insurancefor <- stepAIC(modinsurance, direction = "forward", trace = T)

Start: AIC=23316.43

charges ~ age + sex + bmi + children + smoker + region

1. **Obtenga el mejor modelo utilizando el método de selección de variables hacia atrás, empleando como criterio el AIC.**

> insuranceback <- stepAIC(modinsurance, direction = "backward", trace = T)

Start: AIC=23316.43

charges ~ age + sex + bmi + children + smoker + region

Df Sum of Sq RSS AIC

- sex 1 5.7164e+06 4.8845e+10 23315

<none> 4.8840e+10 23316

- region 3 2.3343e+08 4.9073e+10 23317

- children 1 4.3755e+08 4.9277e+10 23326

- bmi 1 5.1692e+09 5.4009e+10 23449

- age 1 1.7124e+10 6.5964e+10 23717

- smoker 1 1.2245e+11 1.7129e+11 24993

Step: AIC=23314.58

charges ~ age + bmi + children + smoker + region

Df Sum of Sq RSS AIC

<none> 4.8845e+10 23315

- region 3 2.3320e+08 4.9078e+10 23315

- children 1 4.3596e+08 4.9281e+10 23325

- bmi 1 5.1645e+09 5.4010e+10 23447

- age 1 1.7151e+10 6.5996e+10 23715

- smoker 1 1.2301e+11 1.7186e+11 24996

**4. Obtenga el mejor modelo utilizando el método de selección de variables por pasos, empleando como criterio el AIC.**

> insuranceboth <- stepAIC(modinsurance, direction = "both", trace = T)

Start: AIC=23316.43

charges ~ age + sex + bmi + children + smoker + region

Df Sum of Sq RSS AIC

- sex 1 5.7164e+06 4.8845e+10 23315

<none> 4.8840e+10 23316

- region 3 2.3343e+08 4.9073e+10 23317

- children 1 4.3755e+08 4.9277e+10 23326

- bmi 1 5.1692e+09 5.4009e+10 23449

- age 1 1.7124e+10 6.5964e+10 23717

- smoker 1 1.2245e+11 1.7129e+11 24993

Step: AIC=23314.58

charges ~ age + bmi + children + smoker + region

Df Sum of Sq RSS AIC

<none> 4.8845e+10 23315

- region 3 2.3320e+08 4.9078e+10 23315

+ sex 1 5.7164e+06 4.8840e+10 23316

- children 1 4.3596e+08 4.9281e+10 23325

- bmi 1 5.1645e+09 5.4010e+10 23447

- age 1 1.7151e+10 6.5996e+10 23715

- smoker 1 1.2301e+11 1.7186e+11 24996

1. **Con los resultados obtenidos de los métodos de selección, determine cual es el mejor modelo.**

> AIC(insurancefor)

[1] 27115.51

> AIC(insuranceback)

[1] 27113.66

> AIC(insuranceboth)

[1] 27113.66

Tanto el modelo de selección de variables hacia atrás como el de selección de variables por pasos son el mejor modelo, ambos con el mismo AIC