UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

Departamento de Estadística e Informática

Trabajo Grupal - Análisis de Regresión

Tutorial de Regresión

Análisis del Dataset "Insurance"

INTEGRANTES:

Apellidos y Nombres	**Código**
Maguiña Tabory, Leonardo	20241388
Cruzado Flores, Abel	20240000
Montes Yato, Fátima	20220510

GRUPO: 2

DOCENTE A CARGO: Gamboa U., Jesus E.

FECHA DE ENTREGA: Domingo 1 de Febrero del 2026

Lima, 2026

Contents

1. INTRODUCCIÓN

1.1 Importancia de la regresión lineal múltiple

Todo trabajo riguroso requiere de pruebas estadísticas, estas varían de acuerdo al enfoque que se requiera. En tanto, en el mundo real, al observar un fenómeno que quisiéramos explicar este puede depender de una o múltiples variables así como sus observaciones, es así como intentamos comprender aquel fenómeno en base a la ocurrencia de estas observaciones y cómo se comportan. Sin embargo, algo que deberíamos comprender es que no todas las variables pueden estar relacionadas con el fenómeno e incluso aunque encontremos cierta relación, esto no indica la causalidad del fenómeno que queremos explicar; por lo tanto, existen algunas variables más significativas que otras las cuales tienen mayor "peso" en función al fenómeno descrito. Ahora hablemos de una de las técnicas estadísticas fundamentales o centrales dentro del análisis cuantitativo en cualquier disciplina que imaginemos, desde la estadística como tal, economía hasta las ciencias sociales.

En sí, qué es lo que hace a esta técnica tan fundamental para poder comprender los fenómenos u "objetivos" de alguna investigación que querramos elaborar. Anteriormente se comentó acerca de que puede existir cierta relación entre una variable y otra, aunque sin ser causales una de la otra. La regresión lineal múltiple no solo abarca el análisis bivariado del fenómeno, ya que como se mencionó; un fenómeno real difícilmente puede ser explicado por una sola variable, sino múltiples. Entonces, la técnica que desde ahora llamaremos RLM nos permite trascender o ir más allá de las simples relaciones, permitiendo aislar el efecto a una variable objetivo en función a cierta variable mientras las otras se mantienen constantes, controlando también el error. Todo esto se logra mediante un modelo que será definido más adelante. Para el contexto de este documento en formato de tutorial, se espera que finalizado el mismo el lector sea capaz de comprender la importancia de la RLM, siendo esta el primer escalón hacia otras técnicas más avanzadas. Es entonces, de vital importancia explicar los conceptos que acabarcan la misma y cómo se aplican a este ejemplo en concreto utilizando el software RStudio y cómo los resultados e interpretaciones del mismo llegan a ser una gran competencia dentro de cualquier campo en el que nos desarrollemos.

La regresión lineal múltiple es una técnica estadística fundamental que permite modelar la relación entre una variable respuesta cuantitativa y dos o más variables explicativas. A diferencia de la regresión lineal simple, esta metodología considera el efecto conjunto de múltiples factores, lo que permite obtener modelos más realistas y cercanos a situaciones del mundo real.

Esta técnica es ampliamente utilizada en áreas como economía, salud, ingeniería y ciencias sociales, ya que facilita la explicación de fenómenos complejos, la predicción de resultados futuros y la toma de decisiones basadas en datos. El uso del lenguaje R para implementar la regresión lineal múltiple permite realizar análisis reproducibles, verificar supuestos estadísticos y comunicar resultados de manera clara.

1.2 Objetivos de aprendizaje

Objetivo general

- Capacitar al lector para que pueda diseñar, ejecutar, validar e interpretar los modelos de regresión lineal múltiple utilizando el lenguaje R, extrapolando así a problemas reales o de investigaciones requeridas.

Objetivos específicos

- Ejecutar un análisis exploratorio inicial de los datos requeridos, las variables (sean cuantitativas o cualitativas) necesarias, sus estructuras mediante la visualización, interpretación del mismo (análisis univariado) y la correlación existente entre dos variables (análisis bivariado).
- Formular un modelo de RLM según la notación requerida, interpretando sus distintas componentes.
- Desarrollar las etapas necesarias para la evaluación correcta del modelo.
- Evaluar el ajuste del modelo mediante el Coeficiente de Determinación (R2).
- Verificar el cumplimiento de los supuestos (normalidad, homocedasticidad e independencia de errores).
- Aplicar las medidas necesarias en caso no se verifique el cumplimiento de alguno de los supuestos.
- Predecir valores futuros para nuevas observaciones, comprendiendo los intervalos de confianza de estas predicciones (para la media) y los intervalos de predicción (para casos puntuales).

---

2. CASO DE ESTUDIO (DATASET)

2.1 Descripción del problema

En este tutorial se analiza un conjunto de datos relacionado con los costos de seguros médicos individuales. El objetivo principal es estudiar cómo variables demográficas y de estilo de vida influyen en el monto cobrado por el seguro médico.

El análisis de estos factores permite comprender la variabilidad de los costos y proporciona una base para la predicción del gasto esperado de una persona en función de sus características.

2.2 Origen del dataset

El siguiente dataset titulado "Medical Cost Personal Datasets" ha sido obtenido del siguiente enlace *https://www.kaggle.com/datasets/mirichoi0218/insurance*. Este dataset ha sido extraido de un libro como tal el cual provee distintos datasets para una introducción a Machine Learning utilizando R. Sin embargo, lo utilizaremos para nuestro caso de RLM. Este trata acerca de la previsión de los gastos de seguros médicos en base a distintos factores, los cuales se explicarán a continuación.

2.3 Definición de variables

Variable	Tipo de variable	Observación / Contenido
`age`	Cuantitativa discreta	Edad del beneficiario directo, medida en años completos.
`sex`	Cualitativa nominal	Sexo del beneficiario: Hombre (1) o Mujer (0).
`bmi`	Cuantitativa continua	Índice de masa corporal calculado a partir del peso y la altura (kg/m^2).
`children`	Cuantitativa discreta	Número de hijos o personas a cargo cubiertas por el seguro médico.

Variable	Tipo de variable	Observación / Contenido
`smoker`	Cualitativa nominal	Indica si el asegurado es fumador (1) o no fumador (0).
`region`	Cualitativa nominal	Región de residencia del asegurado en EE.UU.: noreste, sureste, suroeste o noroeste.
`charges`	Cuantitativa continua	Gastos médicos individuales facturados por el seguro médico (variable respuesta).

En el presente estudio, la variable `charges` se considera como la variable respuesta, mientras que las demás variables actúan como variables explicativas dentro del modelo de regresión lineal múltiple.

---

3. EXPLORACIÓN DE DATOS

3.1 Importación del dataset

La importación del dataset permite verificar que los datos se han cargado correctamente y observar las primeras observaciones del conjunto de datos.

```
insurance <- read.csv("data/insurance.csv")
head(insurance)
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

3.2 Resumen descriptivo univariado

El análisis descriptivo univariado permite estudiar cada variable de forma individual, identificando medidas de tendencia central, dispersión y posibles valores atípicos. Este análisis es fundamental para comprender la estructura de los datos antes de realizar inferencias estadísticas.

```
summary(insurance)
```

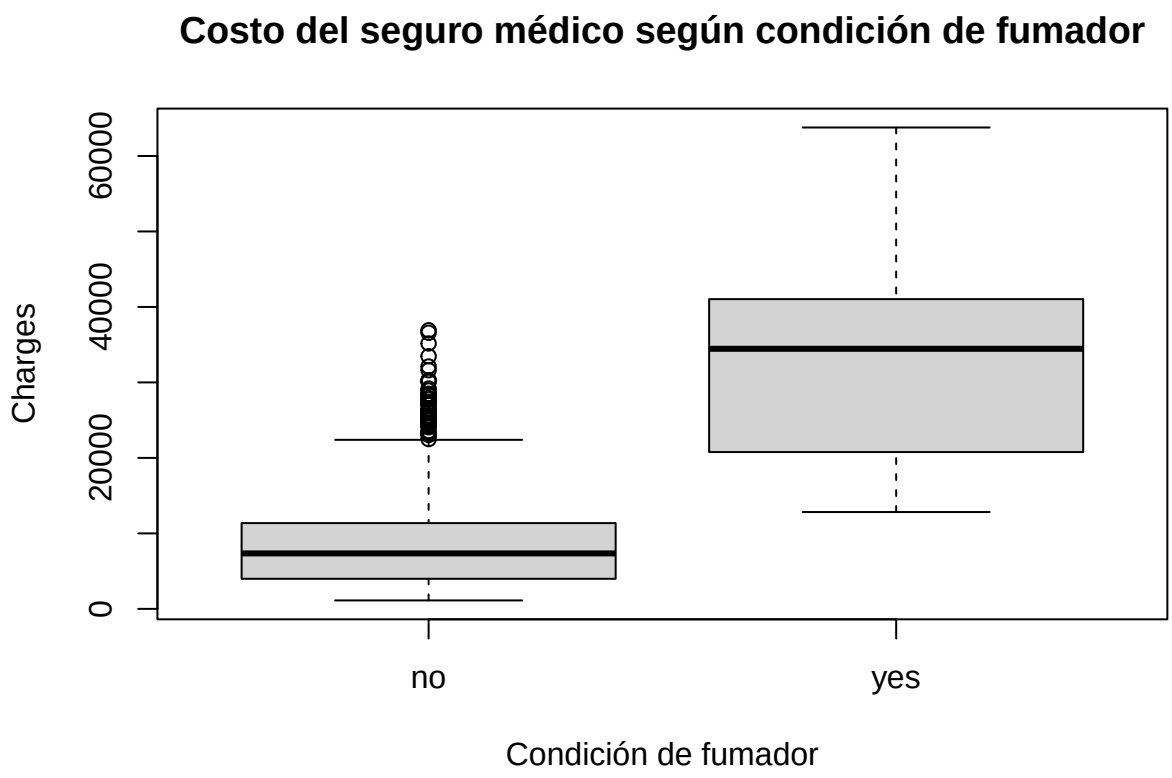
```
##       age            sex                 bmi           children
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000
##  Mean   :39.21                      Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00                      Max.   :53.13   Max.   :5.000
```

```
##     smoker             region            charges
##  Length:1338        Length:1338        Min.   : 1122
##  Class :character   Class :character   1st Qu.: 4740
##  Mode  :character   Mode  :character   Median : 9382
##                                        Mean   :13270
##                                        3rd Qu.:16640
##                                        Max.   :63770
```

3.3 Resumen descriptivo bivariado

El análisis bivariado permite explorar la relación entre la variable respuesta y las variables explicativas, ayudando a identificar patrones relevantes que justifican el uso de un modelo de regresión lineal múltiple.

```
boxplot(charges ~ smoker,
        data = insurance,
        main = "Costo del seguro médico según condición de fumador",
        xlab = "Condición de fumador",
        ylab = "Charges")
```



4.MODELAMIENTO ESTADÍSTICO

4.1 Formulación del modelo

El modelo de regresión lineal múltiple propuesto es:

$$\text{charges} = \beta_0 + \beta_1(\text{age}) + \beta_2(\text{bmi}) + \beta_3(\text{children}) + \beta_4(\text{sex}) + \beta_5(\text{smoker}) + \beta_6(\text{region}) + \varepsilon$$

donde:

- β_0 es el intercepto del modelo.
- $\beta_1, \beta_2, \ldots, \beta_6$ son los coeficientes de regresión asociados a cada variable explicativa.
- ε representa el término de error aleatorio, el cual recoge la variabilidad no explicada por el modelo.

4.2 Ajuste del modelo en R

Antes de ajustar el modelo, es necesario convertir las variables cualitativas a factores para que R las trate adecuadamente dentro del modelo de regresión lineal múltiple. El ajuste del modelo permite estimar los coeficientes de regresión y evaluar la significancia estadística de las variables explicativas incluidas en el análisis.

```
insurance$sex <- as.factor(insurance$sex)
insurance$smoker <- as.factor(insurance$smoker)
insurance$region <- as.factor(insurance$region)

modelo <- lm(charges ~ age + bmi + children + sex + smoker + region,
             data = insurance)

summary(modelo)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + sex + smoker +
##     region, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
## age                256.9       11.9  21.587  < 2e-16 ***
## bmi                339.2       28.6  11.860  < 2e-16 ***
## children           475.5      137.8   3.451 0.000577 ***
## sexmale           -131.3      332.9  -0.394 0.693348
## smokeryes        23848.5      413.1  57.723  < 2e-16 ***
## regionnorthwest   -353.0      476.3  -0.741 0.458769
## regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
## regionsouthwest   -960.0      477.9  -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
```

4.3 Interpretación de los coeficientes de regresión

Cada coeficiente estimado del modelo de regresión lineal múltiple representa el cambio promedio esperado en el costo del seguro médico ante un incremento de una unidad en la variable explicativa correspondiente, manteniendo constantes las demás variables del modelo.

En el caso de las variables cuantitativas, el coeficiente indica cuánto aumenta o disminuye el valor esperado de `charges` por cada unidad adicional de la variable. Para las variables cualitativas, los coeficientes se interpretan en relación con una categoría de referencia, mostrando la diferencia promedio en el costo del seguro médico respecto a dicha categoría.

4.4 Coeficiente de determinación

El coeficiente de determinación R^2 indica la proporción de la variabilidad total del costo del seguro médico que es explicada por el conjunto de variables explicativas incluidas en el modelo. Un valor alto de R^2 sugiere un buen ajuste del modelo a los datos observados.

```
summary(modelo)$r.squared
```

```
## [1] 0.750913
```

4.5 Verificación del supuesto de normalidad de errores

Primero verificamos mediante un histograma que la distribución de los residuales se parezca a una distribución Normal.

```
library(ggplot2)
```

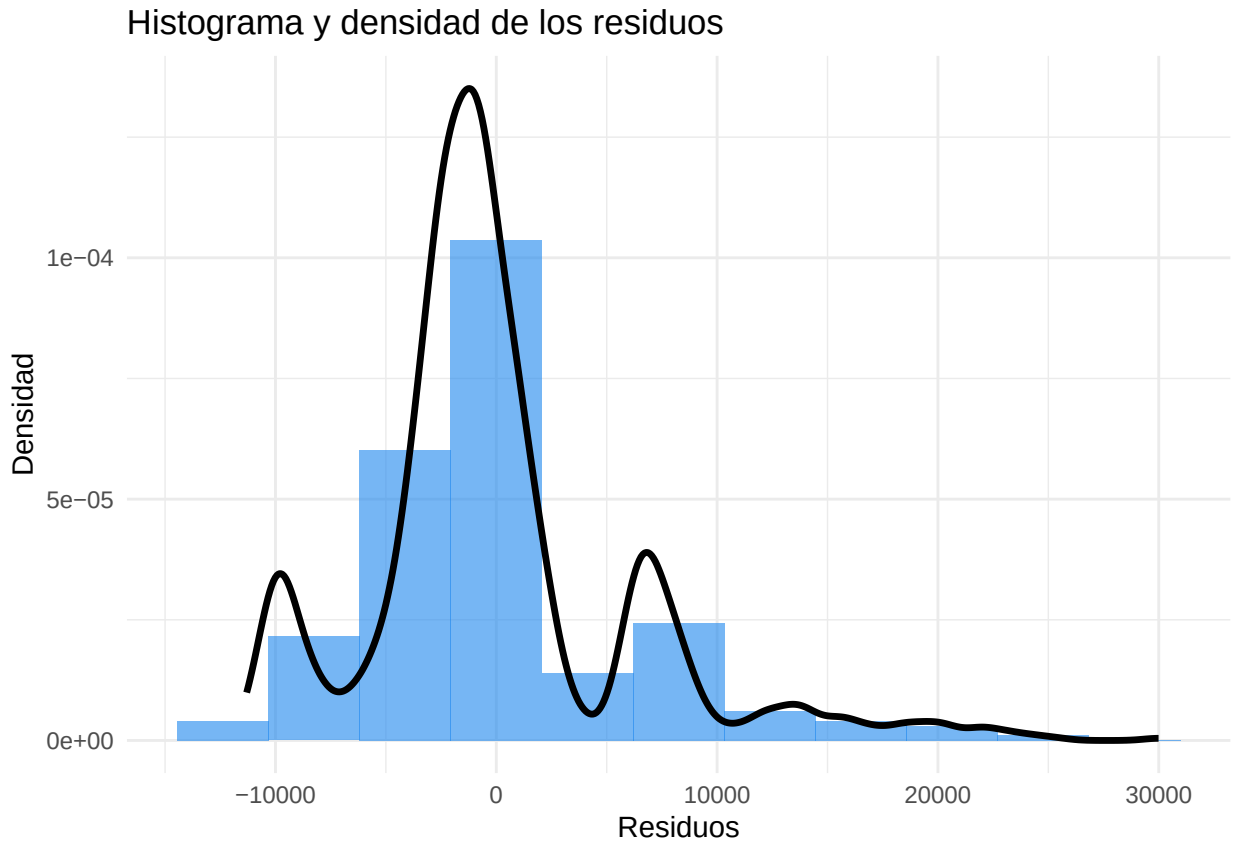
```
## Warning: package 'ggplot2' was built under R version 4.4.2
```

```
residuales <- residuals(modelo)

ggplot(data.frame(residuales), aes(x = residuales)) +
  geom_histogram(aes(y = ..density..),
                 bins = round(1 + 3.3 * log10(nrow(insurance))),
                 fill = "dodgerblue2",
                 alpha = 0.6) +
  geom_density(size = 1.2) +
  labs(title = "Histograma y densidad de los residuos",
       x = "Residuos",
       y = "Densidad") +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Esta prueba permite evaluar si los residuos del modelo de regresión siguen una distribución normal, supuesto necesario para la validez de las inferencias estadísticas.

```
shapiro.test(residuals(modelo))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modelo)
## W = 0.89894, p-value < 2.2e-16
```

4.6 Verificación del supuesto de homocedasticidad

La homocedasticidad implica que la varianza de los errores del modelo es constante para todos los valores de las variables explicativas. Este supuesto es fundamental para que las estimaciones de los coeficientes y las pruebas de hipótesis sean válidas.

```r
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.4.3
```

```
## Cargando paquete requerido: zoo
```

```
## Warning: package 'zoo' was built under R version 4.4.3
```

```
##
## Adjuntando el paquete: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
bptest(modelo)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo
## BP = 121.74, df = 8, p-value < 2.2e-16
```

4.7 Verificación del supuesto de independencia de errores

El supuesto de independencia establece que los errores del modelo no están correlacionados entre sí. Dado que los datos corresponden a observaciones individuales y no a una serie temporal, se asume que este supuesto se cumple.

```r
# CODIGOOOOOOOO
```

4.8 Transformación de datos (si aplica)

En caso de que alguno de los supuestos del modelo no se cumpla, puede considerarse una transformación de la variable respuesta con el objetivo de mejorar el ajuste del modelo y la validez de las inferencias estadísticas. Una transformación común es el uso del logaritmo natural de la variable `charges`.

```r
# CODIGOOOOOOOO
```

4.9 Prueba de hipótesis global

La prueba de hipótesis global evalúa si el modelo de regresión lineal múltiple es estadísticamente significativo en su conjunto. Esta prueba contrasta la hipótesis nula de que todos los coeficientes de regresión, excepto el intercepto, son iguales a cero.

```r
# CODIGOOOOOOOO
summary(modelo)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + sex + smoker +
##     region, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
## age                256.9       11.9  21.587  < 2e-16 ***
## bmi                339.2       28.6  11.860  < 2e-16 ***
## children           475.5      137.8   3.451 0.000577 ***
## sexmale           -131.3      332.9  -0.394 0.693348
## smokeryes        23848.5      413.1  57.723  < 2e-16 ***
## regionnorthwest   -353.0      476.3  -0.741 0.458769
## regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
## regionsouthwest   -960.0      477.9  -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
```

4.10 Pruebas de hipótesis individuales

Las pruebas de hipótesis individuales permiten evaluar la significancia estadística de cada uno de los coeficientes de regresión del modelo. Estas pruebas contrastan la hipótesis nula de que el coeficiente asociado a una variable explicativa es igual a cero, manteniendo constantes las demás variables.

```r
# CODIGOOOOOOOO
summary(modelo)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + sex + smoker +
##     region, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
## age                256.9       11.9  21.587  < 2e-16 ***
## bmi                339.2       28.6  11.860  < 2e-16 ***
## children           475.5      137.8   3.451 0.000577 ***
## sexmale           -131.3      332.9  -0.394 0.693348
## smokeryes        23848.5      413.1  57.723  < 2e-16 ***
```

```
## regionnorthwest   -353.0      476.3  -0.741 0.458769
## regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
## regionsouthwest   -960.0      477.9  -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
```

4.11 Estimación de una media (puntual e intervalar)

La estimación de una media permite obtener el valor promedio esperado del costo del seguro médico para un individuo con características específicas. Además, se puede construir un intervalo de confianza que refleje la incertidumbre asociada a dicha estimación.

```r
nuevo <- data.frame(
  age = 40,
  bmi = 28,
  children = 2,
  sex = "male",
  smoker = "no",
  region = "southwest"
)

predict(modelo, nuevo, interval = "confidence")
```

```
##        fit      lwr     upr
## 1 7692.768 6887.325 8498.21
```

4.12 Predicción de un nuevo valor (puntual e intervalar)

La predicción de un nuevo valor permite estimar el costo del seguro médico para un individuo con características específicas, considerando tanto la incertidumbre del modelo como la variabilidad individual. Por esta razón, el intervalo de predicción suele ser más amplio que el intervalo de confianza de la media.

```r
predict(modelo, nuevo, interval = "prediction")
```

```
##        fit       lwr      upr
## 1 7692.768 -4226.809 19612.34
```

5. REPLICABILIDAD DEL TUTORIAL

Para replicar este tutorial con un conjunto de datos distinto, el estudiante debe mantener la estructura general del documento y modificar únicamente los siguientes elementos:

- El archivo de entrada en la función `read.csv()`, reemplazando el nombre del archivo por el nuevo conjunto de datos.
- La definición de la variable respuesta y las variables explicativas en la fórmula del modelo dentro de la función `lm()`.

El resto del código, incluyendo la exploración de datos, la verificación de supuestos, las pruebas de hipótesis y los procedimientos de estimación y predicción, puede mantenerse sin cambios, lo que garantiza la reproducibilidad del análisis.

6. CONCLUSIONES

En el presente tutorial se aplicó la regresión lineal múltiple para analizar los factores que influyen en el costo del seguro médico. A través del uso del lenguaje R, se realizó un análisis exploratorio de los datos, se ajustó un modelo de regresión, se verificaron los principales supuestos estadísticos y se llevaron a cabo pruebas de hipótesis, estimaciones y predicciones.

El desarrollo de este tutorial permite al estudiante comprender la utilidad de la regresión lineal múltiple como herramienta para el análisis y la predicción en contextos reales, así como replicar el procedimiento utilizando otros conjuntos de datos, manteniendo una estructura clara y reproducible.