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	20241377
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, identificando las variables cuantitativas y cualitativas, su estructura mediante visualizaciones e interpretaciones descriptivas (análisis univariado), así como la correlación existente entre dos variables (análisis bivariado).
- Formular un modelo de Regresión Lineal Múltiple (RLM) según la notación requerida, interpretando adecuadamente sus componentes principales.
- Desarrollar la evaluación del modelo mediante el análisis del ajuste (R^2) y la verificación de supuestos estadísticos fundamentales (normalidad, homocedasticidad e independencia).
- Aplicar las medidas necesarias en caso de que no se verifique el cumplimiento de alguno de los supuestos del modelo.
- Predecir valores futuros para nuevas observaciones, comprendiendo tanto los intervalos de confianza para la media como los intervalos de predicción para casos individuales.

---

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).

Variable	Tipo de variable	Observación / Contenido
`bmi`	Cuantitativa continua	Índice de masa corporal calculado a partir del peso y la altura (kg/m²).
`children`	Cuantitativa discreta	Número de hijos o personas a cargo cubiertas por el seguro médico.
`smoker`	Cualitativa nominal	Indica si el asegurado es fumador (1) o no fumador (0).
`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.

```
library(dplyr)
```

```
##
## Adjuntando el paquete: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
insurance <- read.csv("data/insurance.csv")
insurance <- insurance %>% select(-region)
head(insurance)
```

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

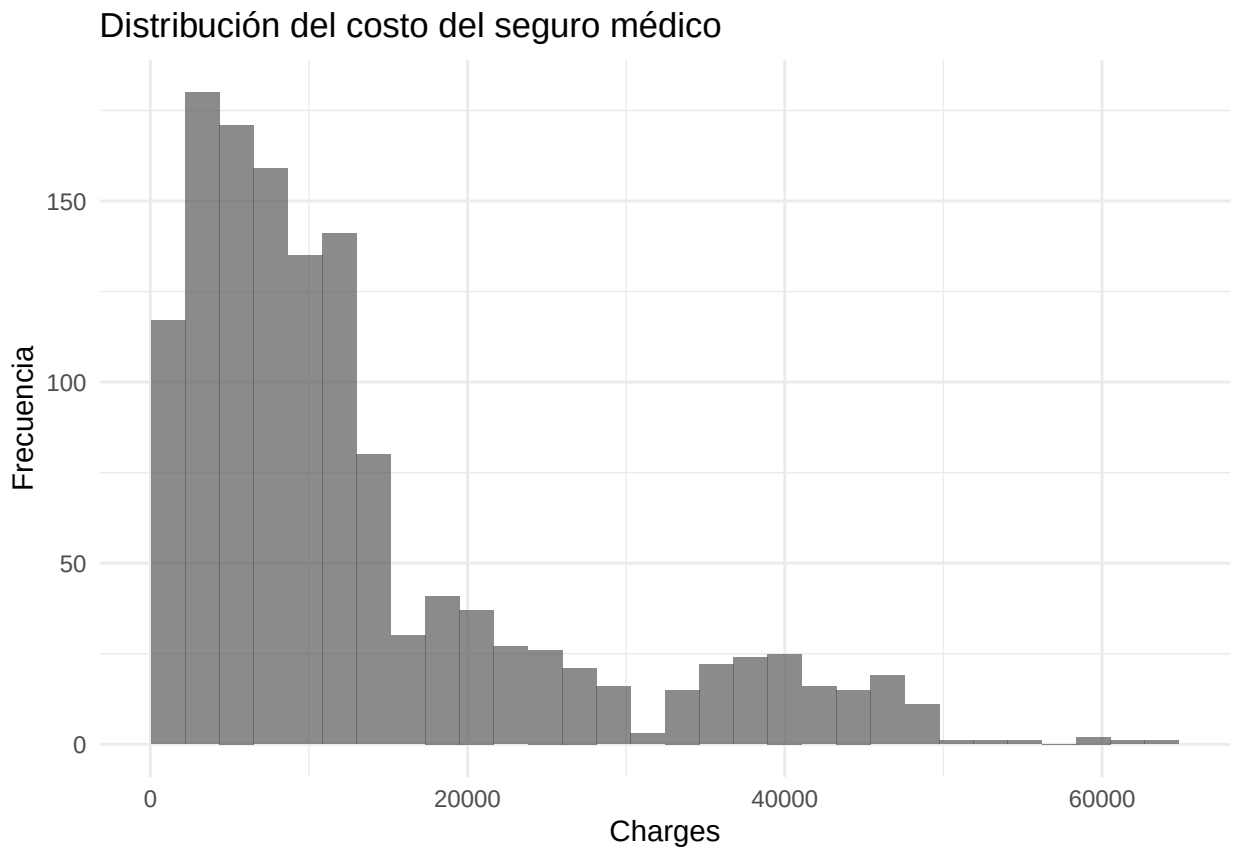
3.2 Resumen descriptivo univariado

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

```
library(ggplot2)
ggplot(insurance, aes(x = charges)) +
  geom_histogram(bins = 30, alpha = 0.7) +
  labs(title = "Distribución del costo del seguro médico",
       x = "Charges",
       y = "Frecuencia") +
  theme_minimal()
```

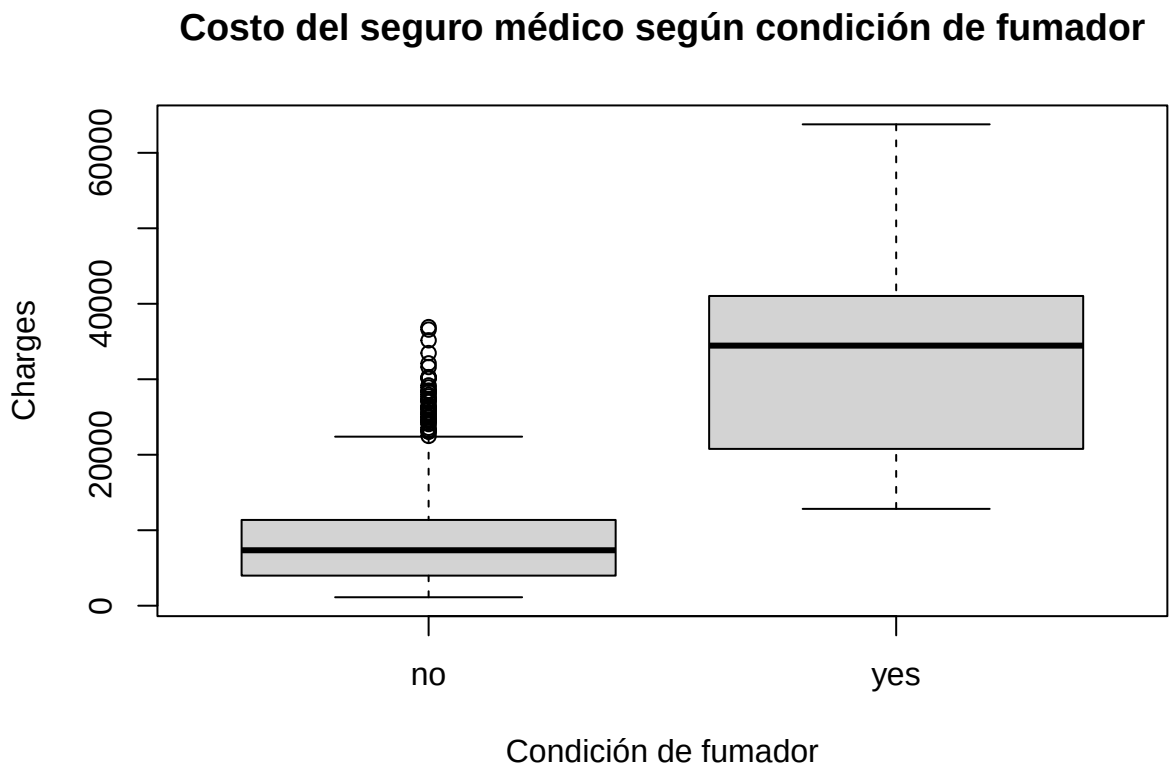


Se observa que la variable `charges` presenta una distribución claramente asimétrica positiva, lo que indica que la mayoría de los asegurados registra costos médicos moderados concentrados en valores bajos e intermedios. Sin embargo, existen algunos casos aislados con gastos extremadamente altos que generan una cola larga hacia la derecha. Este patrón es común en datos de seguros médicos, donde un pequeño grupo de individuos puede incurrir en costos elevados debido a tratamientos complejos o condiciones particulares de salud. Esta asimetría sugiere que, en etapas posteriores del análisis, podría ser necesario considerar transformaciones o ajustes adicionales para mejorar el cumplimiento de supuestos como la normalidad de los errores.

3.3 Resumen descriptivo bivariado

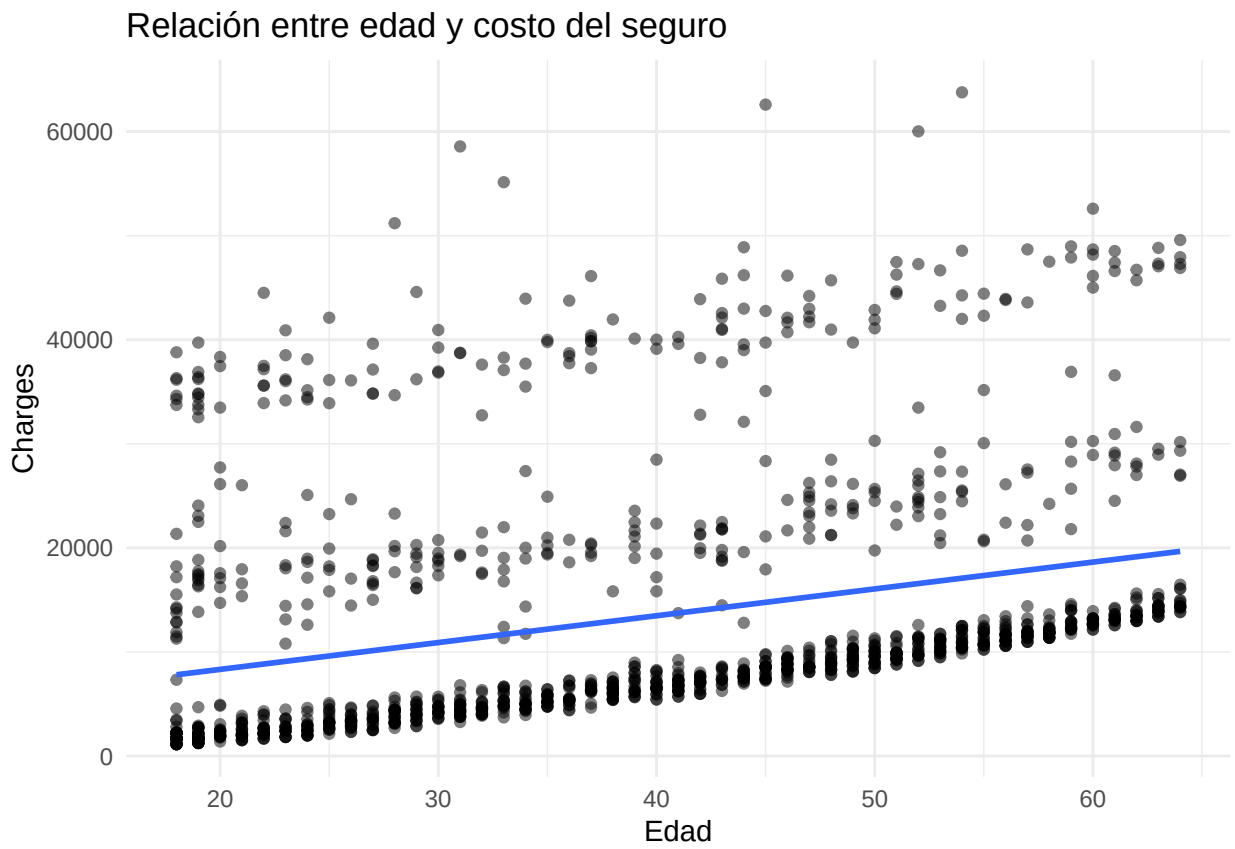
El análisis bivariado permite explorar la relación entre la variable respuesta y las variables explicativas, ayudando a identificar patrones relevantes que justifican el uso de un modelo de regresión lineal múltiple.

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




```
ggplot(insurance, aes(x = age, y = charges)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Relación entre edad y costo del seguro",
       x = "Edad",
       y = "Charges") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



El análisis bivariado mediante el boxplot de `charges` según la condición de fumador permite observar diferencias claras en los costos médicos entre ambos grupos. Se aprecia que los individuos fumadores presentan, en promedio, gastos significativamente más altos que los no fumadores, además de una mayor dispersión en sus valores. Esto sugiere que el hábito de fumar está asociado con un incremento importante en el costo del seguro médico, probablemente debido al mayor riesgo de enfermedades y tratamientos más frecuentes o costosos. Por lo tanto, la variable `smoker` se perfila como un factor explicativo relevante dentro del modelo de regresión lineal múltiple, ya que contribuye a explicar parte de la variabilidad observada en los gastos médicos individuales.

```r
# Correlación de Pearson
cor.test(insurance$age, insurance$charges, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  insurance$age and insurance$charges
## t = 11.453, df = 1336, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2494139 0.3470381
## sample estimates:
##       cor
## 0.2990082
```

La correlación de Pearson permite cuantificar la intensidad y dirección de la relación lineal entre dos variables cuantitativas, en este caso entre la edad (`age`) y los costos médicos (`charges`). El coeficiente obtenido fue

$r = 0.299$, lo que indica una relación lineal positiva pero moderada: a medida que aumenta la edad del asegurado, los gastos del seguro tienden a incrementarse, aunque no de manera extremadamente fuerte. Además, el valor p asociado fue menor que 2.2×10^{-16}, por lo que se rechaza la hipótesis nula de correlación igual a cero, concluyendo que la relación observada es estadísticamente significativa. El intervalo de confianza al 95% $[0.249,\ 0.347]$ confirma que la correlación poblacional es positiva. Este análisis bivariado proporciona evidencia inicial de que la edad podría ser un predictor relevante dentro del modelo de regresión lineal múltiple.

4.MODELAMIENTO ESTADÍSTICO

4.1 Formulación del modelo

El modelo de regresión lineal múltiple propuesto es:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

donde:

- Y : Gastos médicos individuales facturados por el seguro médico (variable respuesta).
- β_0: es el intercepto del modelo.
- $\beta_1, \beta_2, \ldots, \beta_6$: son los coeficientes de regresión asociados a cada variable explicativa.
- X_1 : Edad del beneficiario directo, medida en años completos.
- X_2 : Índice de masa corporal calculado a partir del peso y la altura (kg/m2).
- X_3 : Número de hijos o personas a cargo cubiertas por el seguro médico.
- X_4 : Sexo del beneficiario: Hombre (1) o Mujer (0).
- X_5 : Indica si el asegurado es fumador (1) o no fumador (0).
- 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)

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

summary(modelo)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + sex + smoker, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11837.2  -2916.7   -994.2   1375.3  29565.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12052.46     951.26 -12.670  < 2e-16 ***
## age            257.73      11.90  21.651  < 2e-16 ***
## bmi            322.36      27.42  11.757  < 2e-16 ***
## children       474.41     137.86   3.441 0.000597 ***
## sexmale       -128.64     333.36  -0.386 0.699641
## smokeryes    23823.39     412.52  57.750  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6070 on 1332 degrees of freedom
## Multiple R-squared:  0.7497, Adjusted R-squared:  0.7488
## F-statistic:   798 on 5 and 1332 DF,  p-value: < 2.2e-16
```

4.3 Interpretación de los coeficientes de regresión

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

Intercepto

El intercepto estimado es -11938.5, el cual representa el valor esperado de `charges` cuando todas las variables cuantitativas valen cero y el individuo pertenece a las categorías de referencia (sexo femenino, no fumador). Sin embargo, este valor no necesariamente tiene una interpretación práctica directa.

Edad (`age`)

El coeficiente asociado a `age` es 256.9, lo que indica que por cada año adicional de edad, el costo médico promedio aumenta aproximadamente en 256.9 unidades monetarias, manteniendo constantes las demás variables.

Índice de masa corporal (`bmi`)

El coeficiente estimado para `bmi` es 339.2. Esto sugiere que un aumento de una unidad en el BMI incrementa en promedio los costos médicos en 339.2 unidades, bajo las mismas condiciones.

Número de hijos (`children`)

La variable `children` presenta un coeficiente de 475.5, indicando que cada hijo adicional está asociado con un aumento promedio de 475.5 unidades en los gastos médicos.

Sexo (`sexmale`)

Para la variable categórica `sexmale`, el coeficiente es -131.3, lo que implica que, en promedio, los hombres presentan costos ligeramente menores que las mujeres. No obstante, este efecto no resulta estadísticamente significativo.

Hábito de fumar (`smokeryes`)

El coeficiente para `smokeryes` es 23848.5, mostrando que los fumadores tienen costos médicos considerablemente más altos: en promedio, 23848.5 unidades adicionales en comparación con los no fumadores. Este es el efecto más fuerte dentro del modelo.

El predictor más influyente es smoker, lo que confirma lo observado en el análisis bivariado.En conjunto, estos resultados permiten identificar qué factores influyen con mayor intensidad sobre el costo del seguro médico, destacando principalmente el hábito de fumar como el predictor más importante.

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)$adj.r.squared
```

```
## [1] 0.748783
```

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.

Esto con la finalidad de tener una conclusión más fuerte, sobre el cumplimiento supuesto de normalidad de errores, cual es que los errores se distribuyen normalmente. supuesto necesario para la validez de las inferencias estadísticas.

$$\epsilon_i \sim N(0, \sigma^2)$$

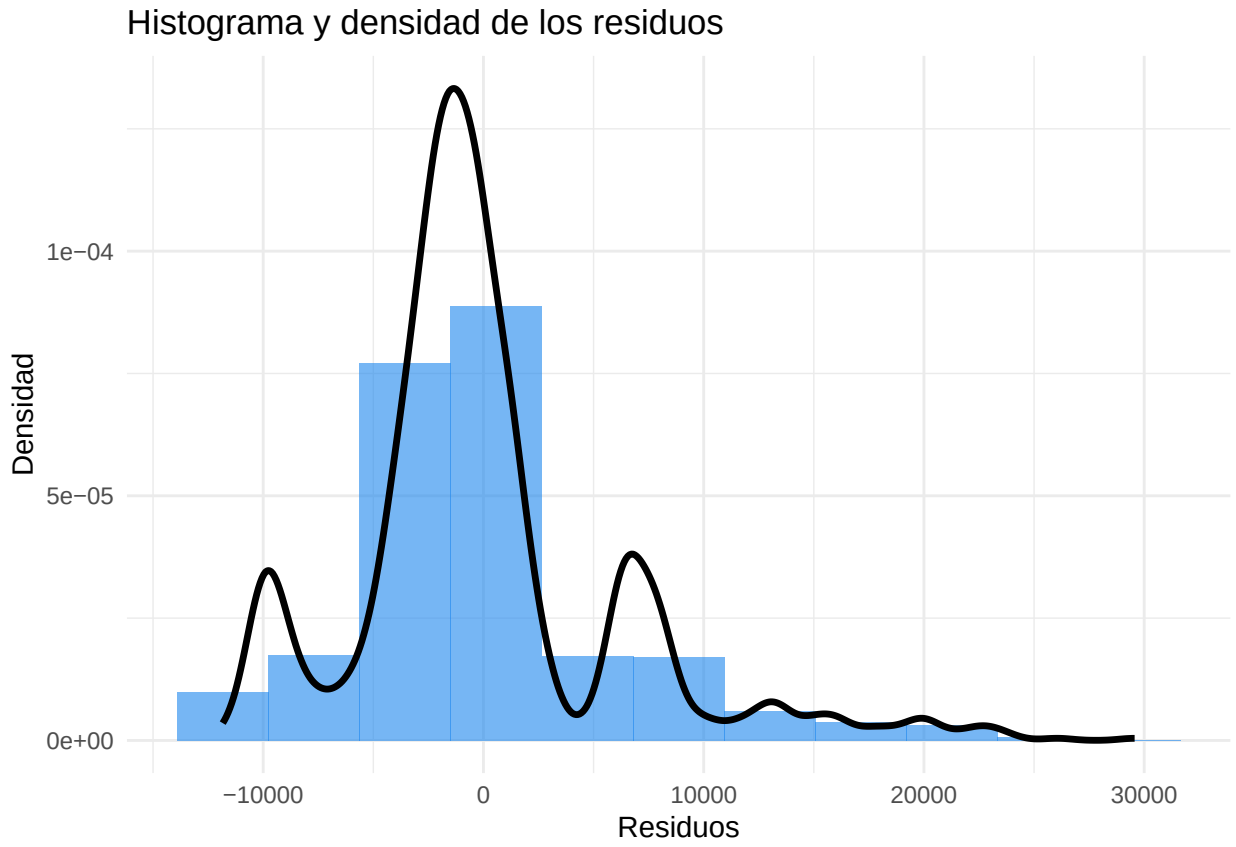
```
library(ggplot2)

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.
```

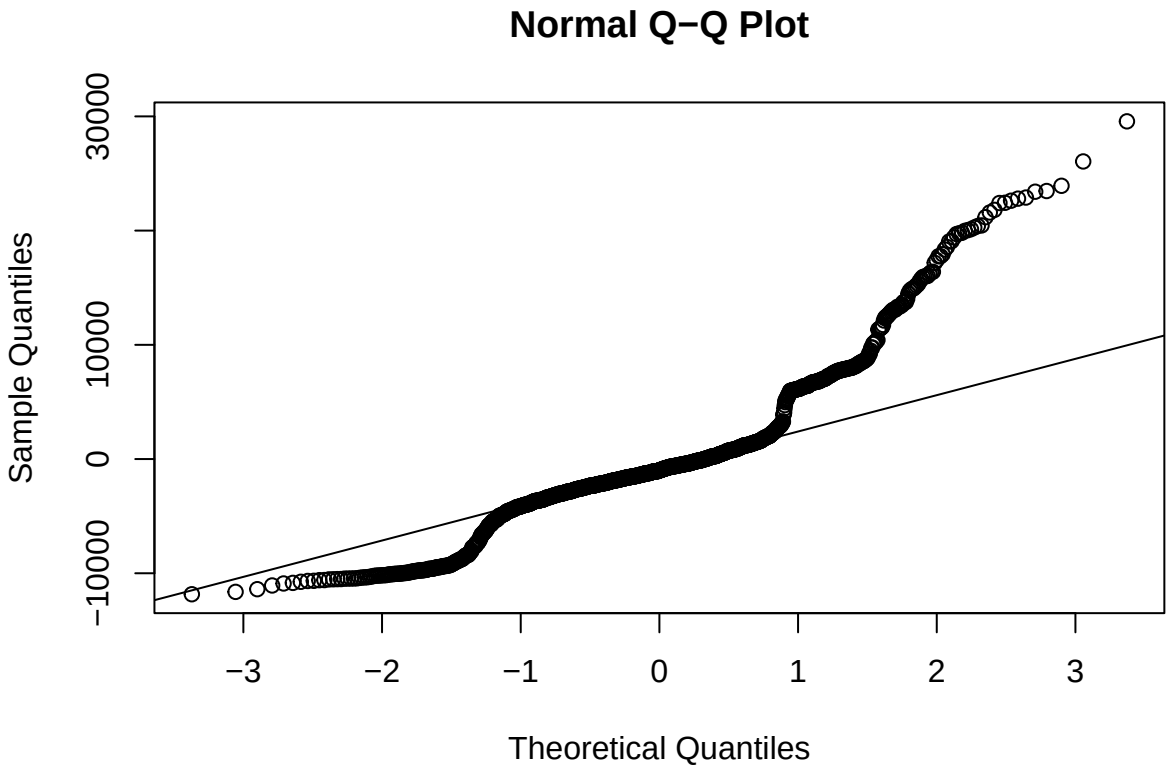
```
## Warning: The dot-dot notation (‘..density..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(density)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



Se aprecia que los errores no se acercan a una distribución normal, en cambio, presenta una marcada asímetria positiva, además la distribución muestra una posible forma leptocúrtica, por la forma apuntalada en el centro, y la cola pesada hacía la derecha.

Usamos ahora un gráfico de probabilidad normal, con la finalidad de detecar outliers, asímetria y curtosis.

```
qqnorm(residuales); qqline(residuales)
```



Se puede observar un fuerte alejamiento de los puntos hacía la esquina superior derecha, lo que podría implicar asímetría y así mismo, el incumplimiento del supuesto de normalidad.

Agostino test Se utiliza para tener pruebas estadísticas de que el coeficiente de asimetría de los residuos es diferente significativamente de 0:

$H_0 : As = 0 \qquad H_1 : As \neq 0 \qquad \alpha = 0.05$

```
library(moments)
```

```
## Warning: package 'moments' was built under R version 4.5.2
```

```
residuales |> agostino.test()
```

```
##
##  D'Agostino skewness test
##
## data:  residuales
## skew = 1.2143, z = 14.5974, p-value < 2.2e-16
## alternative hypothesis: data have a skewness
```

Se rechaza H_0, entonces existe evidencia de que el coeficiente de asímetria es diferente de cero, por lo que existe asímetría

Anscombe test Se utiliza para tener pruebas estadísticas de que el coeficiente de kurtosis de los residuos es diferente significativamente de 3:

$H_0 : K = 3 \qquad H_1 : K \neq 3 \qquad \alpha = 0.05$

```r
library(dplyr)
residuales %>%  anscombe.test
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  .
## kurt = 5.6502, z = 9.3819, p-value < 2.2e-16
## alternative hypothesis: kurtosis is not equal to 3
```

Se rechaza H_0, por lo que existe suficiente evidencia para concluir que el coeficiente de kurtosis es distinto de 3(los errores no son mesocúrticos). Por lo que es una prueba más de que el supuesto de normalidad no se cumple.

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

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(modelo)
## W = 0.8994, p-value < 2.2e-16
```

Prueba de normalidad (Shapiro–Wilk)

Para evaluar el supuesto de normalidad de los residuos del modelo, se aplicó la prueba de Shapiro–Wilk, la cual contrasta las siguientes hipótesis:

- **Hipótesis nula** H_0: los residuos siguen una distribución Normal.
- **Hipótesis alternativa** H_1: los residuos no siguen una distribución Normal.

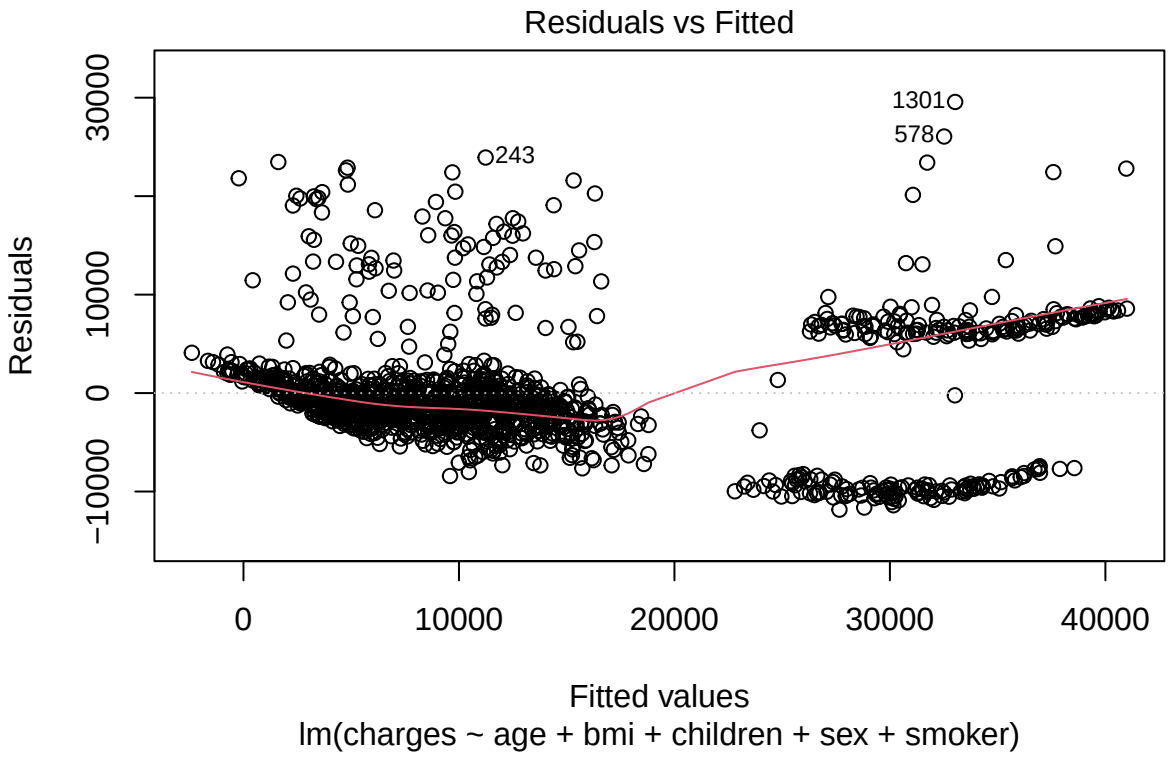
El resultado obtenido fue $W = 0.89894$ con un valor-p extremadamente pequeño ($p < 2.2 \times 10^{-16}$).
Dado que el valor-p es menor que el nivel de significancia usual ($\alpha = 0.05$), se **rechaza la hipótesis nula**, concluyendo que existe evidencia estadística de que los residuos no cumplen completamente el supuesto de normalidad. Esto es consistente con la asimetría observada en la variable `charges`, por lo que podría considerarse una transformación logarítmica para mejorar el ajuste.

4.6 Verificación del supuesto de homocedasticidad

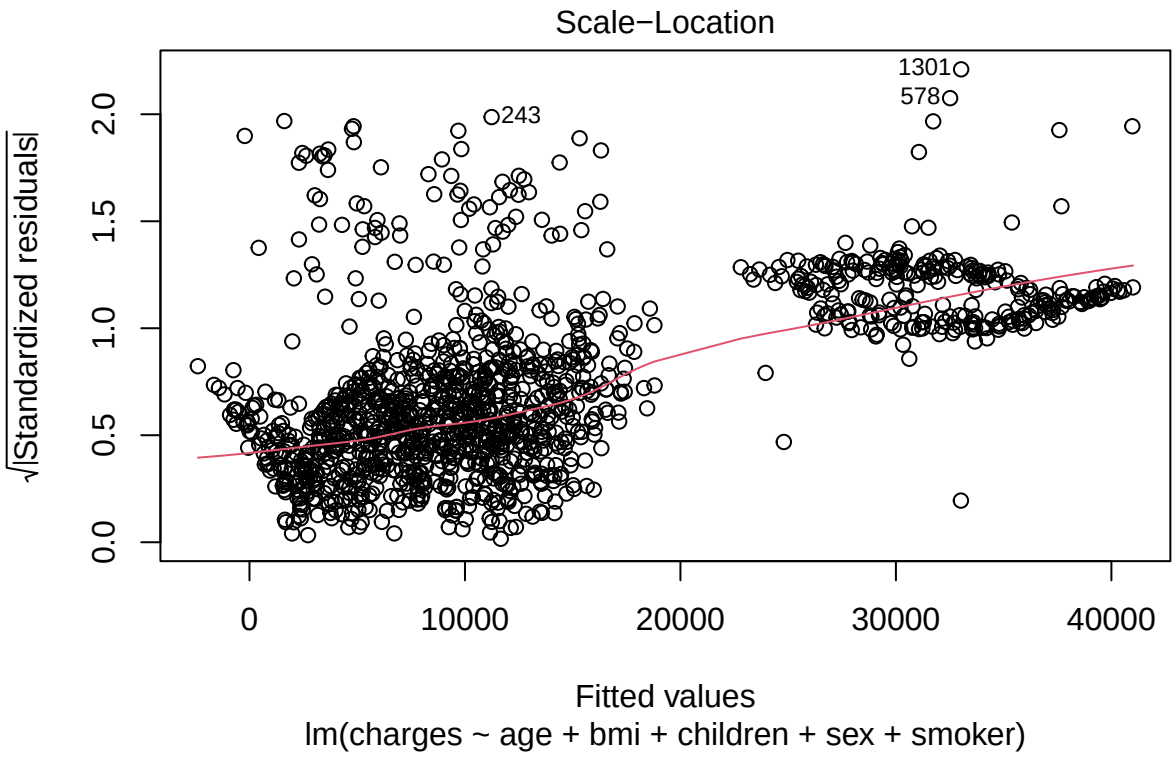
La homocedasticidad implica que la varianza de los errores del modelo es constante para todos los valores de las variables explicativas. Este supuesto es fundamental para que las estimaciones de los coeficientes y las pruebas de hipótesis sean válidas.

Primero, elaboramos dos gráficos, uno de los puntos de los residuales vs los valores ajustados y residuales estandarizados vs los valores ajustados , esto para comprobar que los residuales tienen una distribución aleatoria, lo que sería un indicio del cumplimiento del supuesto de homocedasticidad.

```r
modelo |> plot(which=1)
```




```
modelo |> plot(which=3)
```



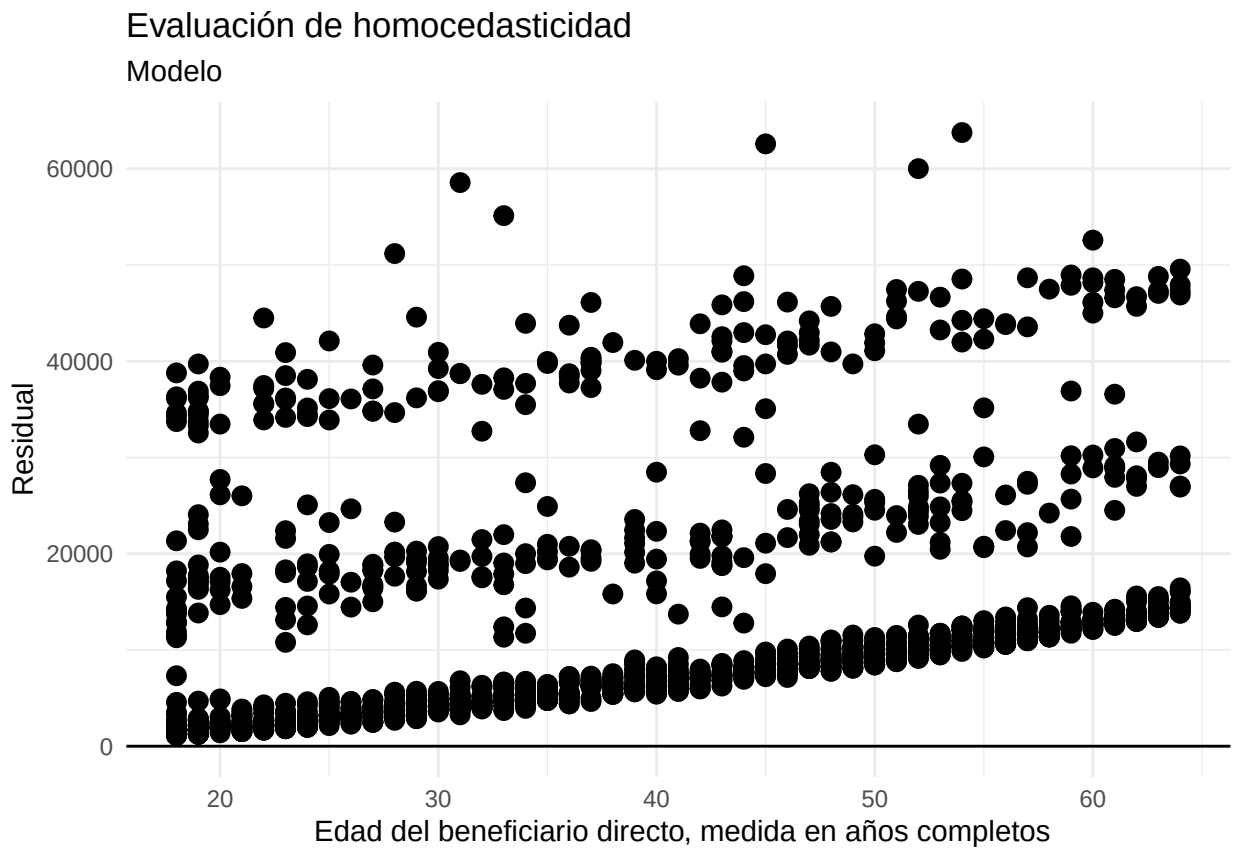
En Ambos gráficos se observa como que la linea que debería de "envolver" los puntos divergen cuando los valores ajustados disminuyen, debido a que los puntos no están distribuidos homogeneamente. Es un indicio de heterocedasticidad.

Con la finalidad de obtener más indisios de homocedasticidad, se hará un diagrama de puntos de manera individual de cada variable explicativa vs variable respuesta.

```
library(broom)
```

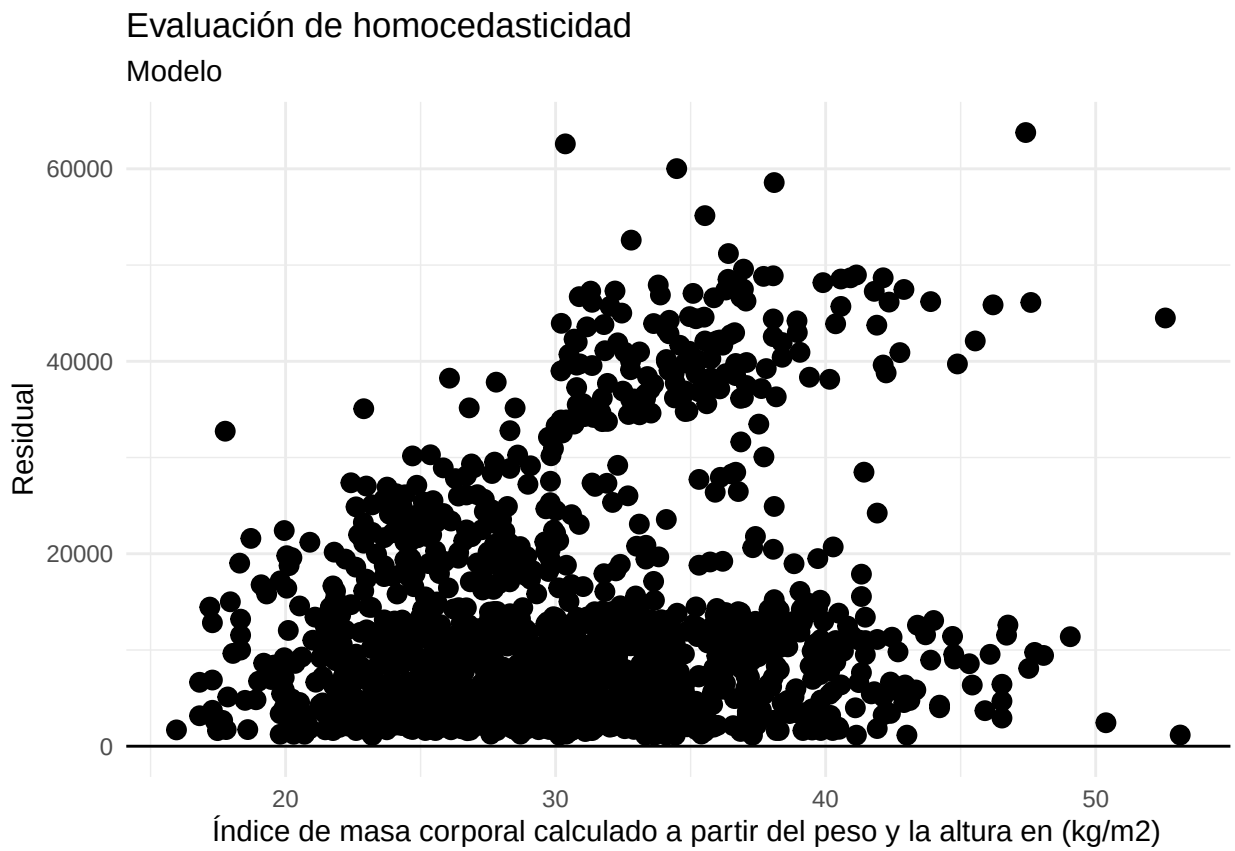
```
## Warning: package 'broom' was built under R version 4.5.2
```

```
modelo |> augment() |>
ggplot(aes(x=age,y=charges))+
geom_point(size = 3) +
geom_hline(yintercept=0)+
labs(x = " Edad del beneficiario directo, medida en años completos",
y = "Residual",
title = "Evaluación de homocedasticidad",
subtitle = "Modelo")+
theme_minimal()
```



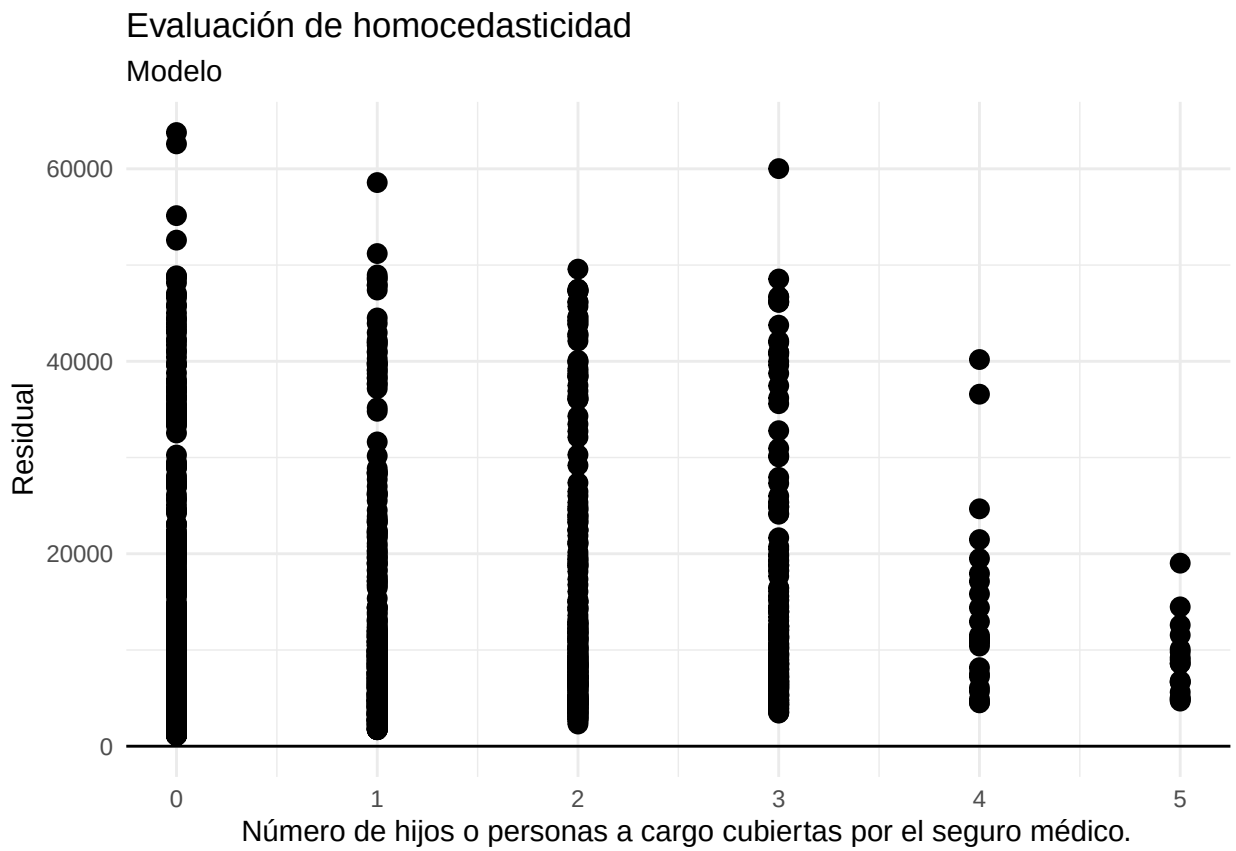
Los residuales tienden a ser más dispersos conforme la edad del beneficiario aumenta.

```
modelo |> augment() |>
ggplot(aes(x=bmi,y=charges))+
geom_point(size = 3) +
geom_hline(yintercept=0)+
labs(x = "  Índice de masa corporal calculado a partir del peso y la altura en (kg/m2)",
y = "Residual",
title = "Evaluación de homocedasticidad",
subtitle = "Modelo")+
theme_minimal()
```



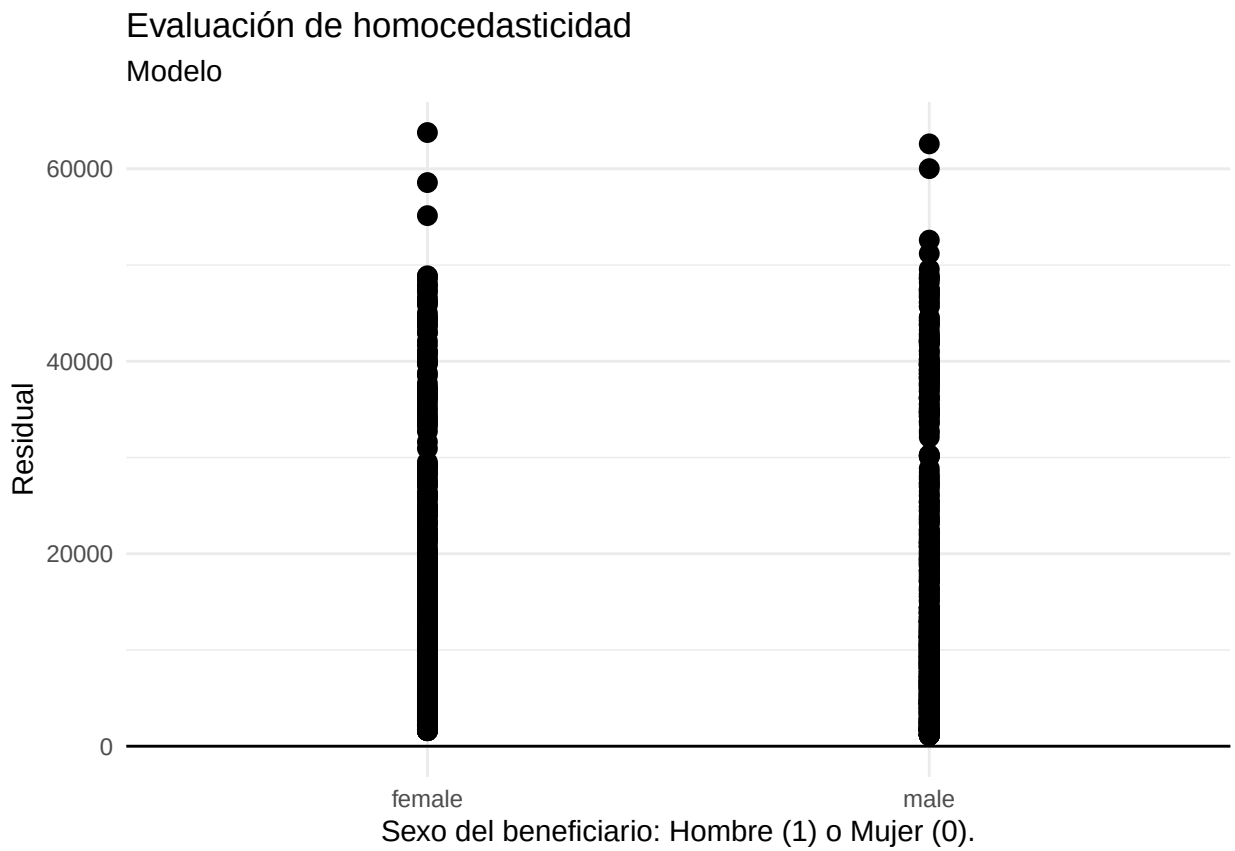
Los residuales presentan una mayor dispersión cuando el índice de masa corporal es alto.

```
modelo |> augment() |>
ggplot(aes(x=children,y=charges))+
geom_point(size = 3) +
geom_hline(yintercept=0)+
labs(x = "  Número de hijos o personas a cargo cubiertas por el seguro médico.",
y = "Residual",
title = "Evaluación de homocedasticidad",
subtitle = "Modelo")+
theme_minimal()
```



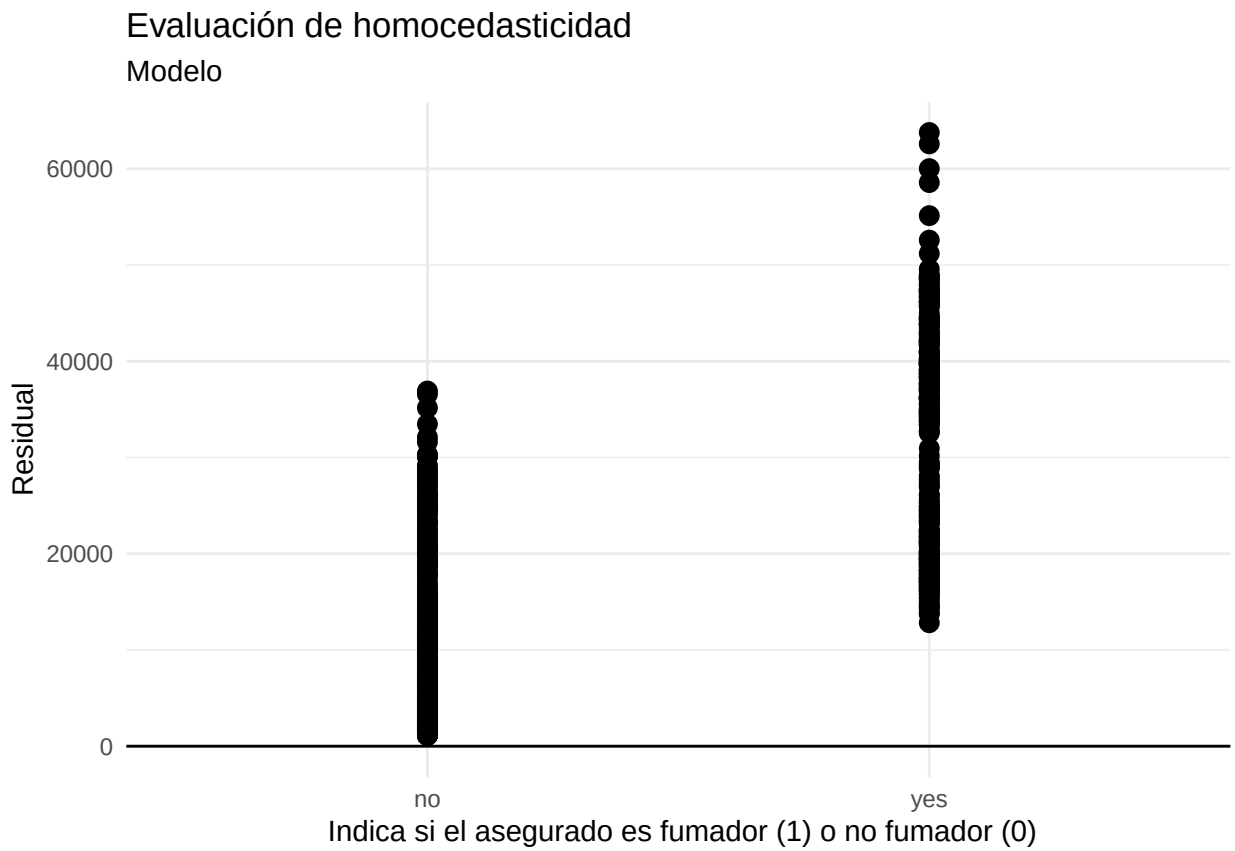
Los residuales presentan una leve disminución de dispersión conforme se tienen más hijos o personas a cargo cubiertas por el seguro médico, hasta que llega a 4, donde la dispersión baja más a comparación de los anteriores, y sigue bajando cuando la variable toma el valor de 5.

```
modelo |> augment() |>
ggplot(aes(x=sex,y=charges))+
geom_point(size = 3) +
geom_hline(yintercept=0)+
labs(x = "Sexo del beneficiario: Hombre (1) o Mujer (0).",
y = "Residual",
title = "Evaluación de homocedasticidad",
subtitle = "Modelo")+
theme_minimal()
```



Los residuales tienden a tener la misma dispersión tanto para hombres como para mujeres.

```
modelo |> augment() |>
ggplot(aes(x=smoker,y=charges))+
geom_point(size = 3) +
geom_hline(yintercept=0)+
labs(x = " Indica si el asegurado es fumador (1) o no fumador (0)",
y = "Residual",
title = "Evaluación de homocedasticidad",
subtitle = "Modelo")+
theme_minimal()
```



Los residuales son más dispersos para los que fuman.

Como se ha encontrado patrones residuales cuando se compara la variable respuesta con ciertas variables explicativas , es una evidencia más de que haya heterocedasticidad.

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 4.5.2
```

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

```
## Warning: package 'zoo' was built under R version 4.5.2
```

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

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

```
bptest(modelo)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  modelo
## BP = 118.02, df = 5, p-value < 2.2e-16
```

Para evaluar el supuesto de **homocedasticidad** (varianza constante de los errores), se aplicó la prueba estadística de **Breusch–Pagan**. En este caso, se plantean las siguientes hipótesis:

- **Hipótesis nula** H_0: la varianza de los errores es constante (existe homocedasticidad).
- **Hipótesis alternativa** H_1: la varianza de los errores no es constante (existe heterocedasticidad).

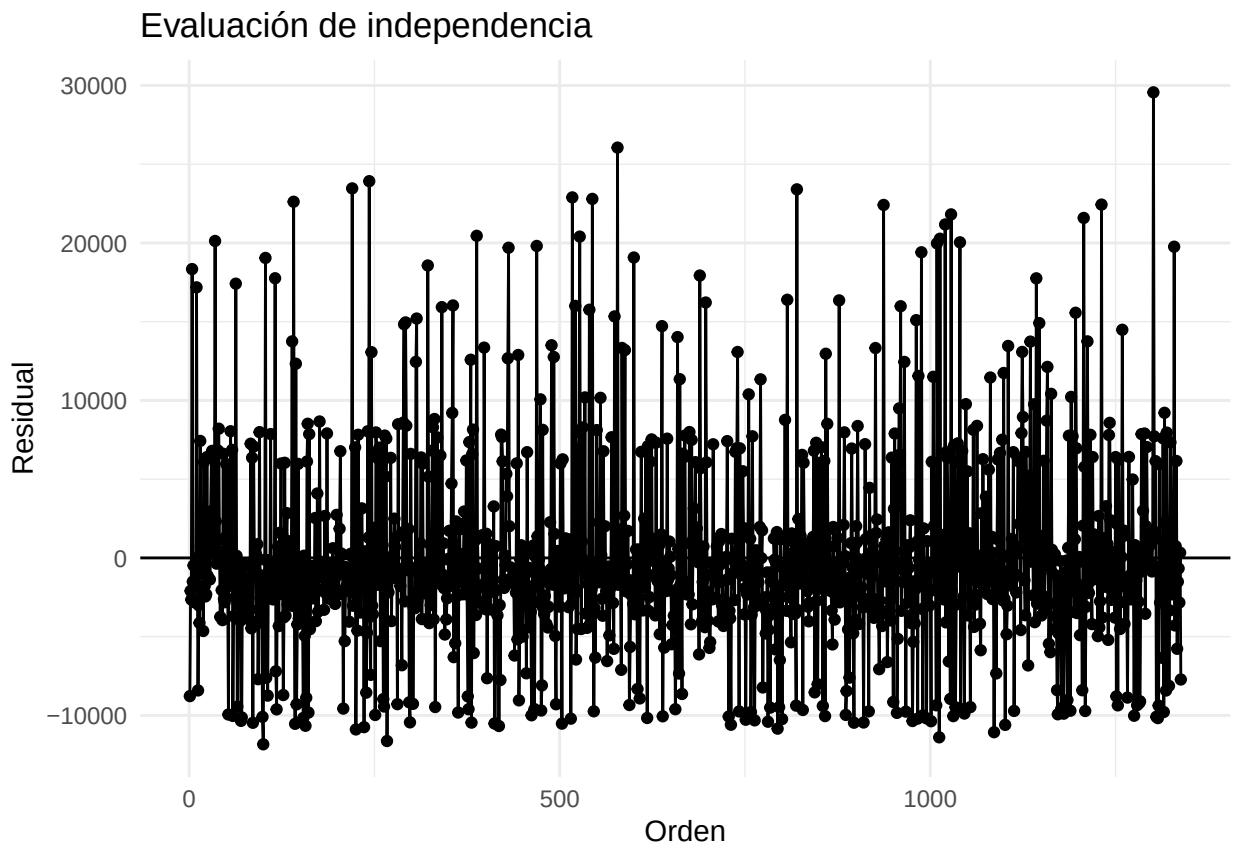
El resultado obtenido fue $BP = 121.74$ con un valor p extremadamente pequeño ($p < 2.2 \times 10^{-16}$). Dado que este valor p es menor que el nivel de significancia usual ($\alpha = 0.05$), se **rechaza la hipótesis nula**. Por lo tanto, existe evidencia estadística de que los residuos del modelo presentan **heterocedasticidad**, es decir, la variabilidad de los errores no se mantiene constante a lo largo de los valores ajustados. Esto es común en datos como el dataset *Insurance*, donde los costos médicos pueden mostrar gran dispersión en ciertos grupos, como se ha observado con los gráficos anteriores, por lo que podría considerarse una transformación (como el logaritmo) o el uso de métodos robustos para mejorar el ajuste del modelo.

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.

Se creará un gráfico de secuencia de residuos en busca de patrones con la finalidad de verificar con evidencias el supuesto.

```
data.frame(residuales) |>
ggplot(aes(x=1:nrow(insurance),y=residuales))+
geom_point(size = 1.5) +
geom_line()+
geom_hline(yintercept=0)+
labs(x = "Orden", y = "Residual", title = "Evaluación de independencia") +
theme_minimal()
```



Si bien podríamos observar a simple vista que no se cumple ningún patron, la cantidad de datos impide el claro analisis de este gráfico.

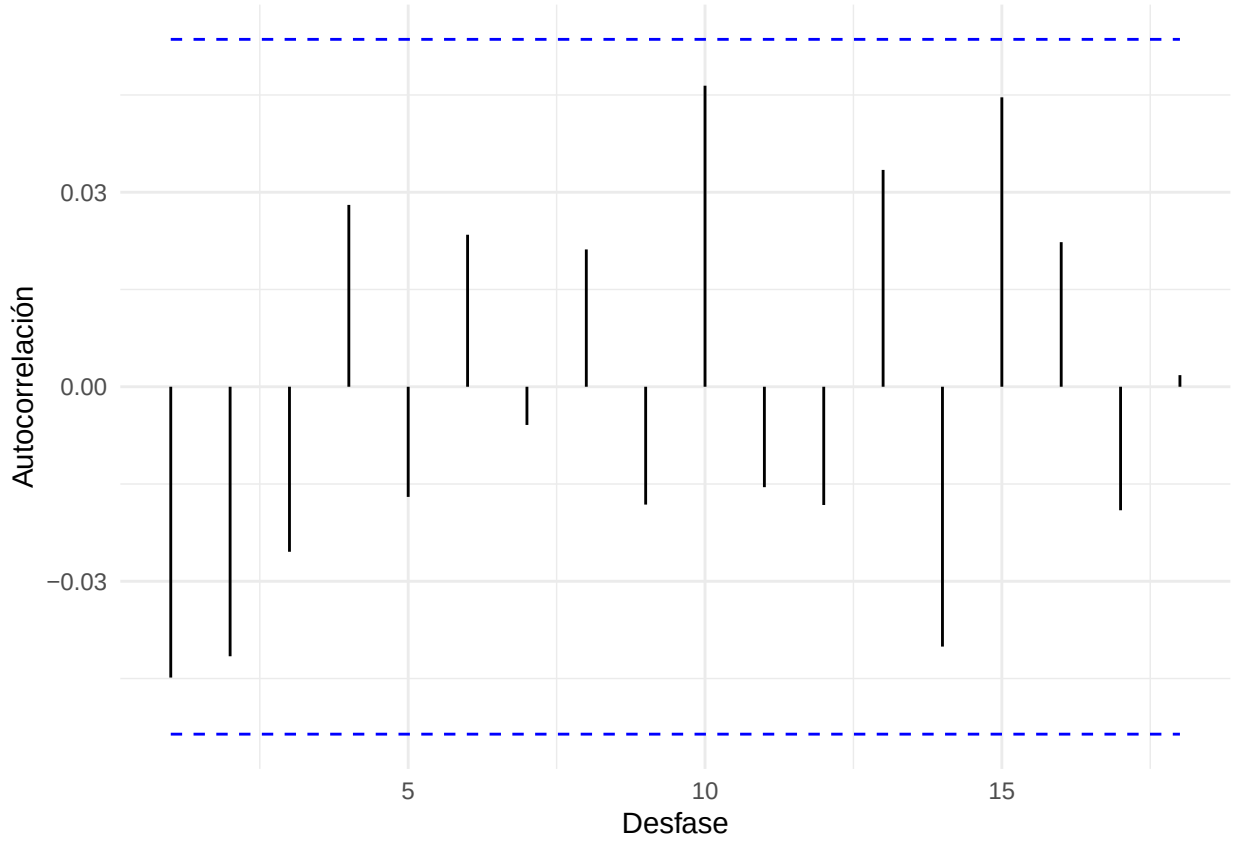
Para solucionar el problema de la interpretación del gráfico anterior, haremos un correlograma con la finalidad de verificar si existe o no independencia de errores.

```
library(ggfortify)
```

```
## Warning: package 'ggfortify' was built under R version 4.5.2
```

```
library(pacman)
p_load(TSA)
residuales |>
TSA::acf(lag = 18, plot=F) |>
autoplot() +
labs(x = "Desfase",y = "Autocorrelación") +
theme_minimal()
```

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## i The deprecated feature was likely used in the ggfortify package.
##   Please report the issue at <https://github.com/sinhrks/ggfortify/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Debido a que todas las autocorrelaciones están dentro de los límites azuales de confianza, decimos que son estadisticamente iguales a cero, lo que evidenciaría independencia de errores.

```
library(lmtest)
dwtest(modelo)
```

```
##
##  Durbin-Watson test
##
## data:  modelo
## DW = 2.0869, p-value = 0.944
## alternative hypothesis: true autocorrelation is greater than 0
```

Para evaluar el supuesto de **independencia de los errores** (ausencia de autocorrelación), se aplicó la prueba estadística de **Durbin–Watson**, la cual es utilizada para detectar correlación serial en los residuos del modelo.

Las hipótesis planteadas son:

- **Hipótesis nula** H_0: no existe autocorrelación en los errores (los residuos son independientes).
- **Hipótesis alternativa** H_1: existe autocorrelación positiva en los errores.

El resultado obtenido fue $DW = 2.0884$, un valor muy cercano a 2, lo cual sugiere ausencia de autocorrelación en los residuos. Además, el valor p asociado fue $p = 0.9472$, que es considerablemente mayor que el nivel de significancia usual ($\alpha = 0.05$).

Por lo tanto, **no se rechaza la hipótesis nula**, concluyendo que no existe evidencia estadística de autocorrelación positiva en los errores. En consecuencia, el supuesto de **independencia de los residuales** se cumple adecuadamente en este modelo de regresión lineal múltiple.

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`.

```
modelo_log <- lm(log(charges) ~ age + bmi + children + sex + smoker ,
                 data = insurance)

summary(modelo_log)
```

```
##
## Call:
## lm(formula = log(charges) ~ age + bmi + children + sex + smoker,
##     data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.08241 -0.20315 -0.05185  0.07057  2.11173
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.0121103  0.0701685  99.932  < 2e-16 ***
## age          0.0347158  0.0008781  39.536  < 2e-16 ***
## bmi          0.0109087  0.0020225   5.394 8.16e-08 ***
## children     0.1017275  0.0101688  10.004  < 2e-16 ***
## sexmale     -0.0750088  0.0245899  -3.050  0.00233 **
## smokeryes    1.5502366  0.0304293  50.946  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4477 on 1332 degrees of freedom
## Multiple R-squared:  0.7638, Adjusted R-squared:  0.7629
## F-statistic: 861.5 on 5 and 1332 DF,  p-value: < 2.2e-16
```

Debido a que en el modelo inicial se evidenciaron desviaciones en los supuestos, se aplicó una transformación logarítmica a la variable respuesta, ajustando el modelo:

$$\log(charges) \sim age + bmi + children + sex + smoker$$

Los resultados muestran que la mayoría de variables explicativas continúan siendo estadísticamente significativas. En particular, la edad ($p < 2e-16$), el índice de masa corporal ($p = 2.42 \times 10^{-10}$), el número de hijos ($p < 2e-16$) y el hábito de fumar ($p < 2e-16$) presentan un efecto importante sobre los costos médicos. El coeficiente asociado a `smokeryes` es el más alto (1.5543), indicando que ser fumador incrementa considerablemente el costo esperado del seguro, incluso en escala logarítmica. Además, el modelo transformado alcanza un $R^2 = 0.7679$, lo que significa que aproximadamente el 76.8% de la variabilidad en $\log(charges)$ es explicada por las variables incluidas. En conjunto, esta transformación mejora la estabilidad de los errores y contribuye a un ajuste más adecuado del modelo.

formulación del modelo con la transformación logarítmica

$$log(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

obtenemos sus coeficientes para tener la ecuación estimada

```
modelo_log %>% coef()
```

```
## (Intercept)         age         bmi    children     sexmale   smokeryes
##  7.01211034  0.03471578  0.01090867  0.10172752 -0.07500875  1.55023665
```

$$log(y) = 7.03 + 0.035 X_1 + 0.013 X_2 + 0.102 X_3 - 0.075 X_4 + 1.554 X_5$$

- 0.035: Por cada año adicional de edad, la mediana de los costos médicos aumenta en 3.52%, manteniendo constantes las demás variables.

Tests para verificar cumplimiento de los supuestos:

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.

```
summary(modelo)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + sex + smoker, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11837.2  -2916.7   -994.2   1375.3  29565.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12052.46     951.26 -12.670  < 2e-16 ***
## age            257.73      11.90  21.651  < 2e-16 ***
## bmi            322.36      27.42  11.757  < 2e-16 ***
## children       474.41     137.86   3.441 0.000597 ***
## sexmale       -128.64     333.36  -0.386 0.699641
## smokeryes    23823.39     412.52  57.750  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6070 on 1332 degrees of freedom
## Multiple R-squared:  0.7497, Adjusted R-squared:  0.7488
## F-statistic:   798 on 5 and 1332 DF,  p-value: < 2.2e-16
```

```
anova(modelo)
```

```
## Analysis of Variance Table
##
## Response: charges
##             Df     Sum Sq    Mean Sq  F value    Pr(>F)
## age          1 1.7530e+10 1.7530e+10  475.827 < 2.2e-16 ***
## bmi          1 5.4464e+09 5.4464e+09  147.834 < 2.2e-16 ***
## children     1 5.7152e+08 5.7152e+08   15.513 8.619e-05 ***
## sex          1 5.8245e+08 5.8245e+08   15.810 7.381e-05 ***
## smoker       1 1.2287e+11 1.2287e+11 3335.109 < 2.2e-16 ***
## Residuals 1332 4.9073e+10 3.6842e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para evaluar la significancia global del modelo de regresión lineal múltiple, se realizó un análisis de varianza (ANOVA). Los resultados indican que el modelo es estadísticamente significativo en conjunto, ya que el estadístico F obtenido es $F = 500.8$ con un valor-p menor a 2.2×10^{-16}. Esto confirma que al menos una de las variables explicativas incluidas contribuye de manera importante a explicar la variabilidad en los costos médicos (`charges`).

Además, la tabla ANOVA muestra que variables como la edad (`age`), el índice de masa corporal (`bmi`) y especialmente el hábito de fumar (`smoker`) presentan efectos altamente significativos ($p < 2.2 \times 10^{-16}$), siendo `smoker` el factor con mayor influencia sobre los costos del seguro. También `children` resulta significativa ($p = 8.446 \times 10^{-5}$).

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.

```
summary(modelo)$coefficients
```

```
##                Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) -12052.4620  951.26043 -12.6699919 8.099035e-35
## age            257.7350   11.90389  21.6513323 2.593679e-89
## bmi            322.3642   27.41860  11.7571362 1.954711e-30
## children       474.4111  137.85580   3.4413577 5.967197e-04
## sexmale       -128.6399  333.36051  -0.3858881 6.996412e-01
## smokeryes    23823.3925  412.52338  57.7504052 0.000000e+00
```

Las pruebas de hipótesis individuales permiten evaluar si cada variable explicativa tiene un efecto significativo sobre el costo del seguro médico (`charges`), manteniendo constantes las demás variables del modelo. Para cada coeficiente se contrasta la hipótesis nula $H_0 : \beta_j = 0$, es decir, que la variable no tiene influencia estadísticamente significativa.

Los resultados muestran que la variable `age` es altamente significativa ($p < 2.2 \times 10^{-16}$), indicando que, en promedio, por cada año adicional de edad el costo del seguro aumenta en aproximadamente **256.86 unidades monetarias**, manteniendo constantes las demás variables. Asimismo, el índice de masa corporal `bmi` también es significativo ($p < 2.2 \times 10^{-16}$), lo que sugiere que un incremento de una unidad en el BMI incrementa el gasto esperado en alrededor de **339.19 unidades**.

La variable `children` presenta un efecto positivo y significativo ($p = 0.000577$), indicando que tener más hijos dependientes se asocia con un aumento promedio de **475.50 unidades** en el costo del seguro.

Por otro lado, la variable `sexmale` no resulta significativa ($p = 0.693$), lo que implica que no existe evidencia suficiente para afirmar que el costo del seguro difiera entre hombres y mujeres cuando se controlan las demás variables.

La variable con mayor impacto es `smokeryes`, que es extremadamente significativa ($p < 2.2 \times 10^{-16}$). El coeficiente estimado indica que ser fumador incrementa el costo del seguro en aproximadamente **23848.53 unidades**, convirtiéndose en el predictor más importante del modelo.

En conclusión, las variables más relevantes para explicar el costo del seguro médico son la edad, el BMI y especialmente el hábito de fumar, mientras que el sexo no muestra un efecto estadísticamente significativo dentro del modelo.

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.

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


modelo_log %>%  predict(newdata = nuevo, interval = "confidence") %>%  exp()
```

```
##        fit      lwr      upr
## 1 6868.017 6583.028 7165.343
```

El intervalo de confianza obtenido permite estimar el **costo médico promedio esperado** para un individuo con las características especificadas (edad 40 años, BMI 28, 2 hijos, hombre, no fumador). El modelo predice un gasto medio aproximado de **8103.318 unidades monetarias**. Además, con un nivel de confianza del 95%, se concluye que el verdadero costo promedio para personas con este perfil se encuentra entre **7528.774** y **8677.862**. Este intervalo refleja únicamente la incertidumbre asociada a la estimación de la media poblacional, por lo que es más estrecho que un intervalo de predicción individual.

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.

```
modelo_log %>%  predict(newdata = nuevo, interval = "prediction") %>%  exp()
```

```
##        fit      lwr      upr
## 1 6868.017 2850.599 16547.28
```

El intervalo de predicción obtenido permite estimar el rango en el que podría ubicarse el **costo médico de un caso individual específico** con las características dadas (edad 40 años, BMI 28, 2 hijos, hombre,

no fumador). El modelo predice un gasto aproximado de **8103.318 unidades monetarias**, pero debido a la variabilidad natural entre individuos, se espera que el costo real de una persona con este perfil pueda encontrarse, con un 95% de confianza, entre **-4226.81** y **19612.34**. Este intervalo es considerablemente más amplio que el intervalo de confianza para la media, ya que incorpora no solo la incertidumbre en la estimación del promedio, sino también la dispersión propia de los valores individuales. El límite inferior negativo no representa un costo real posible, sino una consecuencia matemática de la amplitud del intervalo y la variabilidad presente en los datos.

5. REPLICABILIDAD DEL TUTORIAL

La replicabilidad de este tutorial con un conjunto de datos diferente depende mucho del formato, estructura y variables que se desea analizar. Como primeras condiciones tenemos que la variable que necesitamos estimar tiene que ser cuantitativa continua o discreta pero con valores relativamente altos, por lo tanto, las variables explicativas al formular el modelo tienen que ser cuantitativas de igual manera (si no lo son, se harán de forma binaria) sean cualitativas o no.

1. Como primer paso leeremos el dataset de interés, siendo la lectura de estos conocimiento previo con funciones tales como read.csv, read.csv2 o read_excel (del paquete readxl).

```
  datos = read.csv("mi_dataset.csv")
  library(readxl)
  datos = read_excel("mi_dataset.xlsx")
```

2. Luego se realizará un análisis descriptivo del dataset, esto nos permitirá obtener diferentes estadísticas, así como podemos realizar gráficos que nos permitirán analizar las variables de interés y cómo se comportan. Se hará un análisis exploratorio en general. Para nuestro caso se realizaron de la siguiente manera. Resumen descriptivo del dataset importado

```
summary(datos)
```

```
Boxplot básico de las variables que creamos necesarias
```

```
boxplot(var1 ~ var2,
        data = datos,
        main = "Título del boxplot comparando var1 con var2",
        xlab = "Nombre de var2",
        ylab = "Var1")
```

3. Prosiguiendo con el tutorial, pasamos a la formulación del modelo de RLM. Como se había mencionado antes, la variable objetivo preferiblemente debe ser cuantitativa continua y si es discreta esta debe ser de valores altos. Luego, decidiremos qué variables podrían afectar a esta variable objetivo, teniendo así las variables independientes de nuestro modelo.

   Formulamos el modelo y su conversión a factor de ser necesaria

   Cada apartado del mismo han sido explicados en el documento

```
modelo = lm(vartarget \~ var 1 + var 2 + var 3 + var 4, datos)
```

```
Luego podemos ver los coeficientes del modelo ajustado e
interpretarlos
```

```
coef(modelo)
```

4. Como una continuación del anterior apartado, obtenemos el coeficiente de determinación R^2, el cual nos indicará en qué proporción la variabilidad total de nuestra variable objetivo es explicada por las variables explicativas valga la redundancia elegidas del dataset. Si obtenemos un valor alto de este coeficiente, significa que se dio un buen ajuste del modelo con las observaciones del dataset.

   Obtenemos el coeficiente de determinación

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

5. Como ya se explicó anteriormente, se necesitan verificar diferentes supuestos acerca de los errores o residuales dentro del modelo, el primero será el de normalidad de errores.

   Para este primer supuesto se pueden verificar visualmente como con pruebas estadísticas. Primeramente, es necesario extraer los residuales de nuestro modelo.

```
residuales = residuals(modelo)
```

```
Luego, podemos elaborar un histograma y densidad de estos residuales
con el paquete ggplot2 siendo el código completamente adaptable al
usuario del ejemplo ya hecho en el apartado del documento
correspondiente.
```

```
library(ggplot2)
library(dplyr)
residuales |>
  ggplot() + 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 residuales",
       x = "Residuales",
       y = "Densidad") +
  theme_minimal()
```

```
Lo que se trata de encontrar es que este histograma sea lo más
parecido a una distribución normal.

Otra verificación gráfica puede hacerse con el gráfico de quantil vs
quantil.
```

```
qqnorm(residuales); qqline(residuales)
```

```
La verificación de este gráfico debe ser interpretar las
observaciones tanto en la parte inicial como final de la línea que
observamos, se busca que las observaciones sean lo más cercanas a la
línea definida.
```

```
Por último, para este supuesto, para su robustes, podemos
verificarlo mediante la prueba estadística de Shapiro-Wilk podemos
determinar la veracidad de nuestra hipótesis nula o alterna
detallada en el documento buscamos un pvalor alto para su
verificación, por el contrario, si obtenemos uno menor que el nivel
de significancia, se rechazará la hipótesis nula.
```

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

6. El segundo supuesto de Homocedasticidad: Indica que la varianza de los errores del modelo es constante.
   Primero cargamos el paquete necesario

```r
library(lmtest)
```

```
De la misma manera que para la hipótesis nula y alterna del supuesto
de normalidad, buscamos valores altos del pvalor para su
verificación.
```

```r
bptest(modelo)
```

7. Y por último el de independencia de errores: El cual establece que los errores del modelo no están correlacionados entre sí.

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.