Tutorial básico de Regresión

Lucia Fatima Carbajal Falcon Arlette Alashka Carmen Tullume
Gabriel Omar Evaristo Jacinto

2026-01-26

[1] Universidad Nacional Agraria La Molina; [2] Departamento Estadística e Informática

Introducción

a. Importancia de la regresión lineal múltiple:

La regresión lineal múltiple es fundamental porque permite modelar fenómenos complejos donde una variable de interés depende de varios factores a la vez. A diferencia de la regresión simple, esta técnica ofrece una visión más realista al analizar cómo múltiples variables independientes influyen simultáneamente en un resultado, permitiendo aislar el efecto individual de cada una mientras se mantienen las demás constantes.

Su relevancia radica en su doble capacidad: **explicativa y predictiva**. Por un lado, ayuda a identificar qué factores tienen un impacto real y significativo en un problema; por otro, permite construir fórmulas precisas para predecir escenarios futuros, lo que la convierte en una herramienta indispensable para la toma de decisiones basada en datos en cualquier disciplina científica.

b. Objetivos de aprendizaje

- Objetivo General:

  - Capacitar al estudiante en la implementación y validación de un modelo de regresión lineal múltiple utilizando el software estadÍstico R.

- Objetivos Específicos:

  - Comprender la importancia de incluir múltiples variables predictoras para explicar un fenómeno real.
  - Ejecutar los ajustes de modelos lineales mediante la función lm() de R.
  - Interpretar los coeficientes de regresión y su impacto en la variable respuesta.
  - Evaluar la calidad del ajuste mediante el coeficiente de determinación R-cuadrado.
  - Validar los supuestos de normalidad, homecedasticidad e independencia a través de pruebas diagnósticas.
  - Predecir nuevos valores de la variable respuesta utilizando el modelo final optimizado.

Caso (dataset)

En la actualidad educativa, el uso de recursos de inteligencia artificial se ha popularizado entre los alumnos como un medio complementario para el aprendizaje y la ejecución de tareas académicas. Sin embargo, hay dudas respecto al verdadero efecto de estas tecnologías en el desempeño académico. Así, el objetivo de este

tutorial es examinar cómo el uso de la inteligencia artificial se relaciona con el desempeño académico de los estudiantes, el cual se medirá a través de sus calificaciones posteriores, teniendo en cuenta además aspectos importantes como las horas dedicadas al estudio cada día y el rendimiento académico previo, utilizando un modelo de regresión lineal múltiple.

b. Origen del dataset: https://www.kaggle.com/datasets/aminasalamt/students-ai-usage-and-academic-performance

Definición de variables

Variable dependiente

1. **nota_despues**

- **Tipo:** Cuantitativa continua
- **Descripción:** Calificación final del estudiante tras el periodo de uso de herramientas de IA. Es la variable que deseamos predecir o explicar.

Variables independientes

1. **horas_estudio**

- **Tipo:** Cuantitativa continua
- **Descripción:** Tiempo promedio diario (en horas) dedicado al estudio personal.

2. **usa_ia**

- **Tipo:** Cualitativa dicotómica (dummy)
- **Descripción:** Variable que indica si el estudiante utiliza herramientas de inteligencia artificial como apoyo en sus estudios (1 = sí, 0 = no).

3. **nota_antes**

- **Tipo:** Cuantitativa continua
- **Descripción:** Calificación obtenida por el estudiante antes del uso de herramientas de inteligencia artificial.

Exploración de datos

¿Por qué explorar los datos?

Antes de ajustar un modelo de regresión, es fundamental conocer el comportamiento de las variables. La exploración de datos permite identificar valores atípicos, rangos plausibles, niveles de dispersión y posibles problemas que podrían afectar la validez del modelo.

```
# Cargar librerías
library(dplyr)

# Lecutra del dataset
datos = read.csv("students_ai_usage.csv")

# Renombrar columnas para un mejor manejo
datos <- datos |>
  rename(
    edad = age,
    nivel_educativo = education_level,
    horas_estudio = study_hours_per_day,
    usa_ia = uses_ai,
    herramientas_ia = ai_tools_used,
    proposito_ia = purpose_of_ai,
    nota_antes = grades_before_ai,
    nota_despues = grades_after_ai,
    tiempo_pantalla = daily_screen_time_hours
  )

# Convertimos a factor "usa_ia" para que R entienda que "Sí" y "No" son categorías, no solo texto
datos$usa_ia <- factor(datos$usa_ia, levels = c("No", "Yes"), labels = c("No", "Sí"))

# Verificamos los nuevos nombres
head(datos)
```

```
##   edad nivel_educativo horas_estudio usa_ia herramientas_ia proposito_ia
## 1   19         college           1.4     No            None         None
## 2   15          school           3.9     Sí         Copilot     Research
## 3   15          school           1.9     Sí         Copilot     Homework
## 4   15          school           2.8     No            None         None
## 5   19         college           2.7     No            None         None
## 6   16          school           1.4     No            None         None
##   nota_antes nota_despues tiempo_pantalla
## 1         62           62               3
## 2         56           61               2
## 3         75           88               5
## 4         55           55               3
## 5         59           59               3
## 6         58           58               4
```

b. Resumen descriptivo univariado (además de realizar el resumen descriptivo, explicar por qué se hace)

El análisis univariado es el primer paso esencial para entender la distribución, tendencia central y dispersión de cada variable por separado. Nos permite identificar valores atípicos (outliers), errores de digitación o

desequilibrios en las categorías (por ejemplo, si hay muy pocos estudiantes que “No” usan IA), lo cual podría afectar la validez del modelo de regresión posterior.

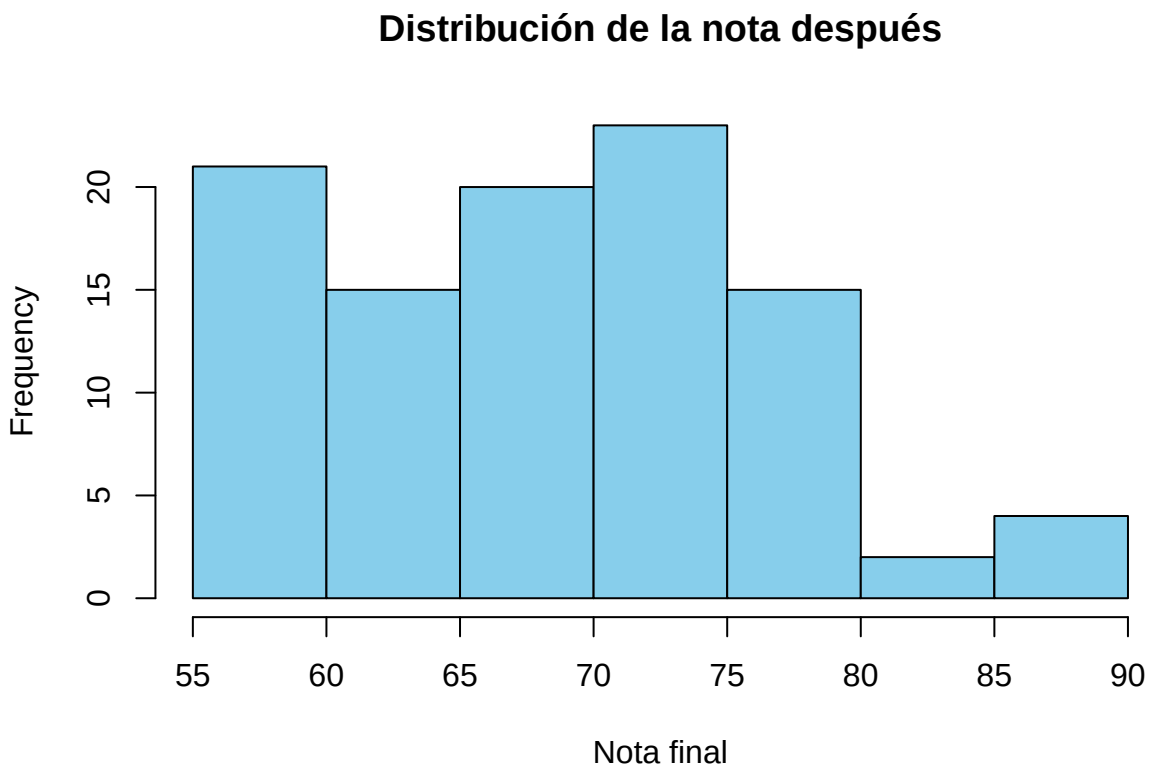
```
summary(datos$nota_despues)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    55.0    61.0    69.0    68.7    74.0    89.0
```

```
sd(datos$nota_despues)
```

```
## [1] 8.140806
```

```
hist(datos$nota_despues,
     col = "skyblue",
     main = "Distribución de la nota después",
       xlab = "Nota final")
```



Variable dependiente: nota_despues

La calificación final de los estudiantes presenta:

Media: 68.7

Mediana: 69

Mínimo – Máximo: 55 a 89

Desviación estándar: 8.14

Esto indica que las notas finales se concentran alrededor de 69 puntos, con una dispersión moderada. Ademas permite ver cómo se distribuyen las calificaciones finales, si están concentradas en un rango específico y si la forma es aproximadamente simétrica.

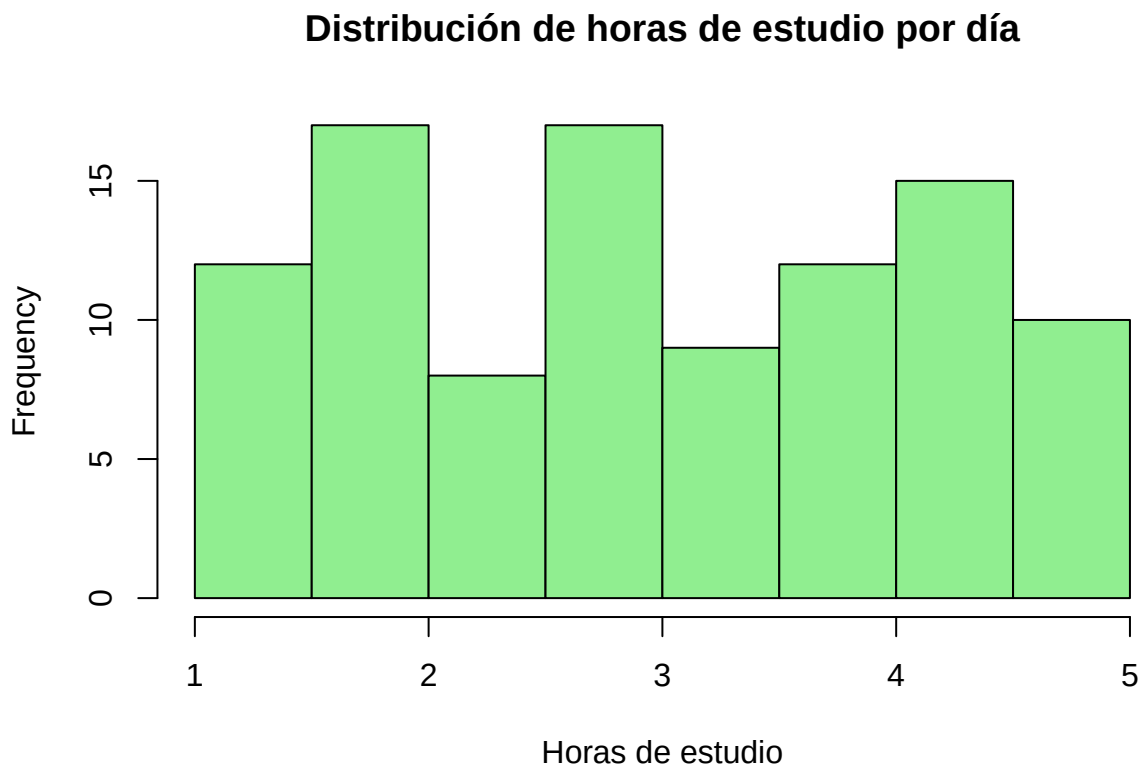
```
summary(datos$horas_estudio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.975   2.800   2.987   4.025   5.000
```

```
sd(datos$horas_estudio)
```

```
## [1] 1.145713
```

```
hist(datos$horas_estudio,
     col = "lightgreen",
     main = "Distribución de horas de estudio por día",
     xlab = "Horas de estudio")
```



Variable independiente: horas_estudio

Las horas de estudio diarias muestran:

Media: 2.99 horas

Mediana: 2.8 horas

Rango: 1 a 5 horas

Desviación estándar: 1.15

Los estudiantes estudian en promedio aproximadamente 3 horas al día. La dispersión es baja a moderada, lo que indica que la mayoría de estudiantes se concentra en un rango similar de tiempo de estudio, sin diferencias excesivas entre ellos.

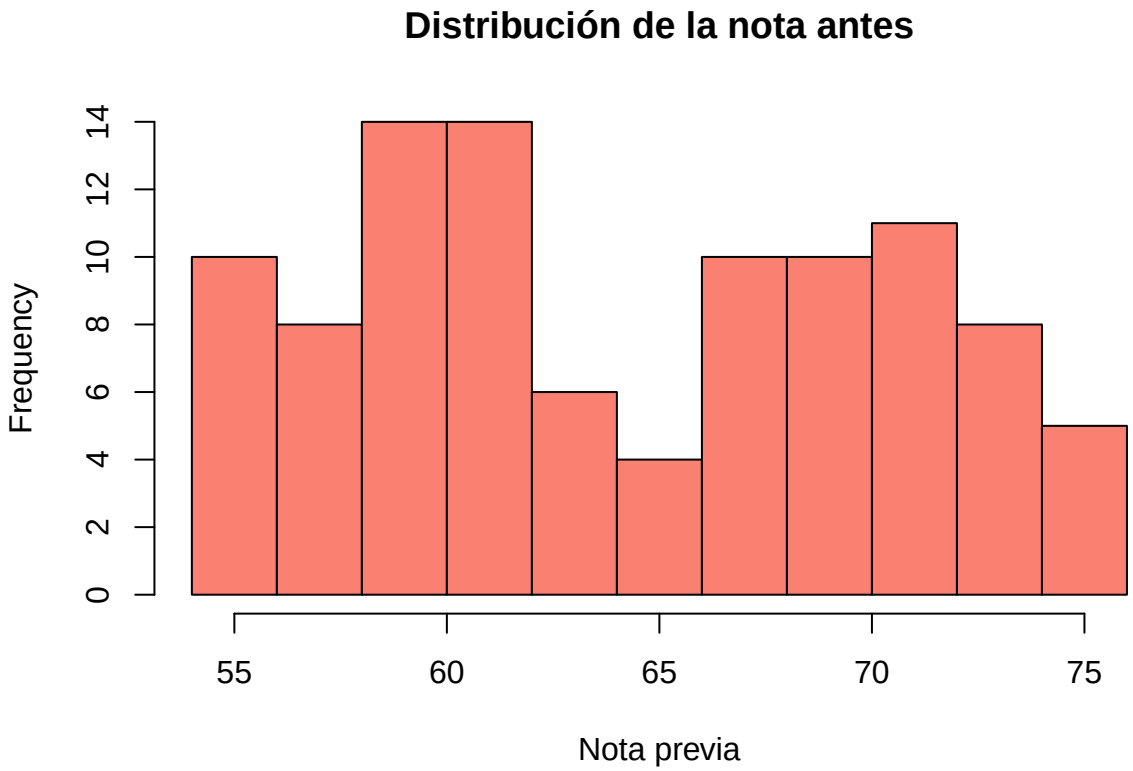
```
summary(datos$nota_antes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   55.00   59.00   63.00   64.77   70.00   75.00
```

```
sd(datos$nota_antes)
```

```
## [1] 6.16909
```

```
hist(datos$nota_antes,
     col = "salmon",
     main = "Distribución de la nota antes",
     xlab = "Nota previa")
```



El rendimiento previo presenta:

Media: 64.77

Mediana: 63

Rango: 55 a 75

Desviación estándar: 6.17

Las calificaciones antes del uso de IA son ligeramente menores que las finales, lo que podría anticipar una mejora general en el desempeño. La variabilidad es moderada, lo que permite que esta variable aporte información relevante al modelo.

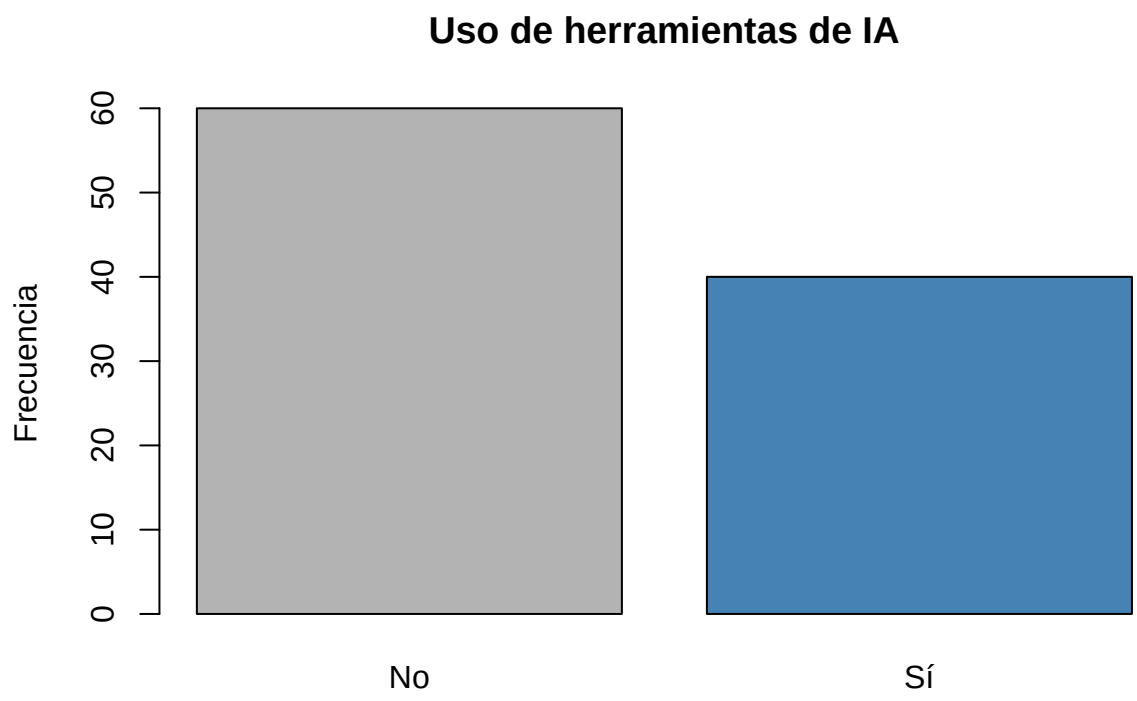
```r
table(datos$usa_ia)
```

```
##
## No Sí
## 60 40
```

```r
prop.table(table(datos$usa_ia))
```

```
##
##  No  Sí
## 0.6 0.4
```

```r
barplot(table(datos$usa_ia),
        col = c("gray70", "steelblue"),
        main = "Uso de herramientas de IA",
        ylab = "Frecuencia")
```



Variable independiente categórica: usa_ia

La distribución del uso de inteligencia artificial es:

No usa IA: 60 estudiantes (60%)

Sí usa IA: 40 estudiantes (40%)

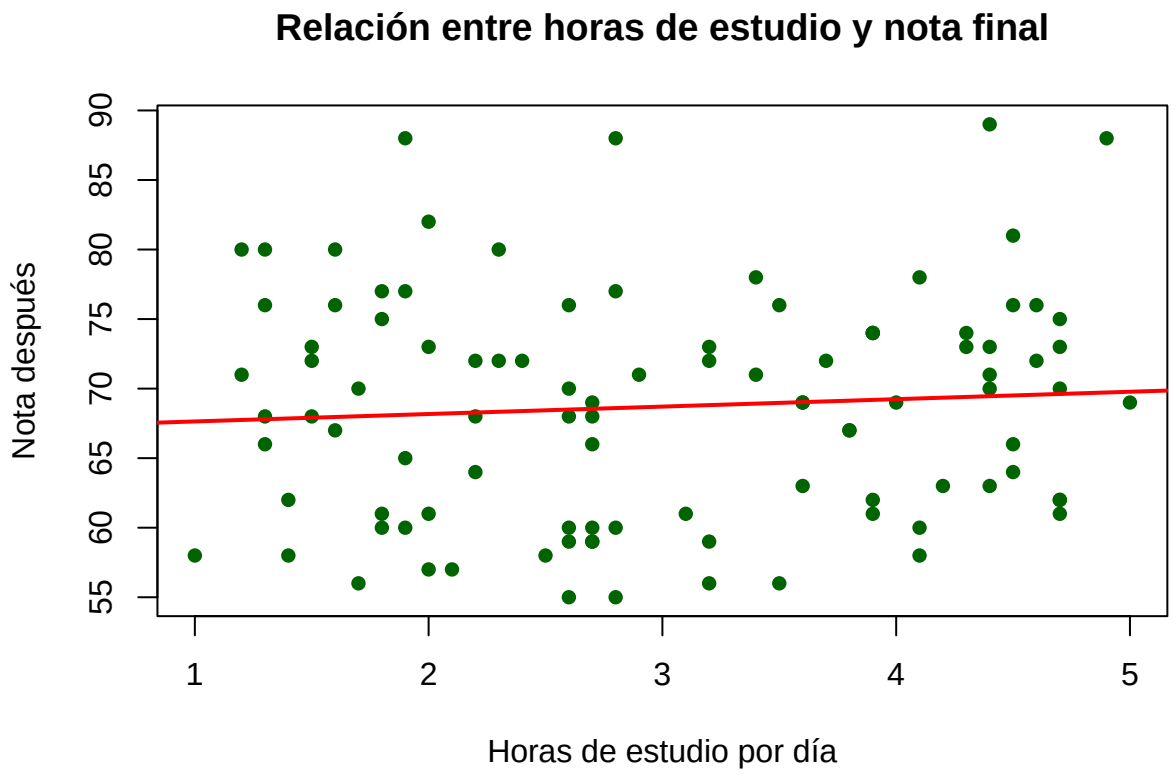
Las proporciones son relativamente equilibradas, lo cual es adecuado para el análisis, ya que ambos grupos tienen tamaños suficientes para comparar sus efectos en el modelo sin generar inestabilidad estadística.

c. Resumen descriptivo bivariado (además de realizar el resumen descriptivo, explicar por qué se hace)

Se realiza para explorar preliminarmente la relación entre la variable respuesta (Y) y cada predictor (X). En regresión lineal, buscamos identificar visualmente si existe una tendencia lineal, la fuerza de la asociación y si la variable cualitativa (uses_ai) desplaza de manera evidente el promedio de las calificaciones. Esto justifica la inclusión de estas variables en el modelo final.

1) Horas de estudio vs Nota despues

```r
plot(datos$horas_estudio, datos$nota_despues,
     col = "darkgreen",
     pch = 16,
     xlab = "Horas de estudio por día",
     ylab = "Nota después",
     main = "Relación entre horas de estudio y nota final")

abline(lm(nota_despues ~ horas_estudio, data = datos), col = "red", lwd = 2)
```



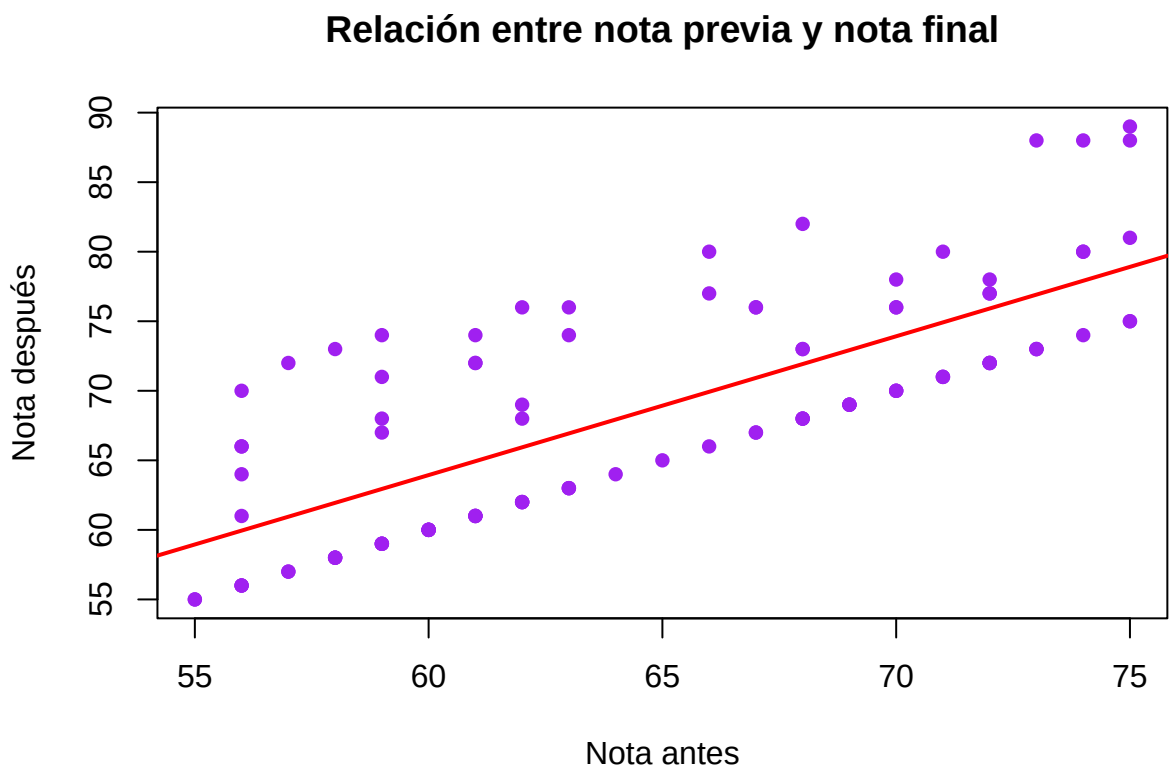
```r
cor(datos$horas_estudio, datos$nota_despues)
```

```
## [1] 0.07516985
```

El diagrama de dispersión muestra la relación entre las horas de estudio diarias y la calificación final obtenida por los estudiantes. Visualmente, los puntos aparecen bastante dispersos y no se observa una tendencia lineal fuerte. La línea de regresión (en rojo) tiene una pendiente levemente positiva, pero muy pequeña. El coeficiente de correlación calculado es 0.075 el cual valor es muy cercano a 0, lo que indica una relación lineal positiva extremadamente débil entre las horas de estudio y la nota final. En términos prácticos, esto sugiere que, en este conjunto de datos, aumentar las horas de estudio no está fuertemente asociado con un incremento notable en la calificación después del uso de IA.

2) Nota antes vs Nota despues

```r
plot(datos$nota_antes, datos$nota_despues,
     col = "purple",
     pch = 16,
     xlab = "Nota antes",
     ylab = "Nota después",
     main = "Relación entre nota previa y nota final")

abline(lm(nota_despues ~ nota_antes, data = datos), col = "red", lwd = 2)
```



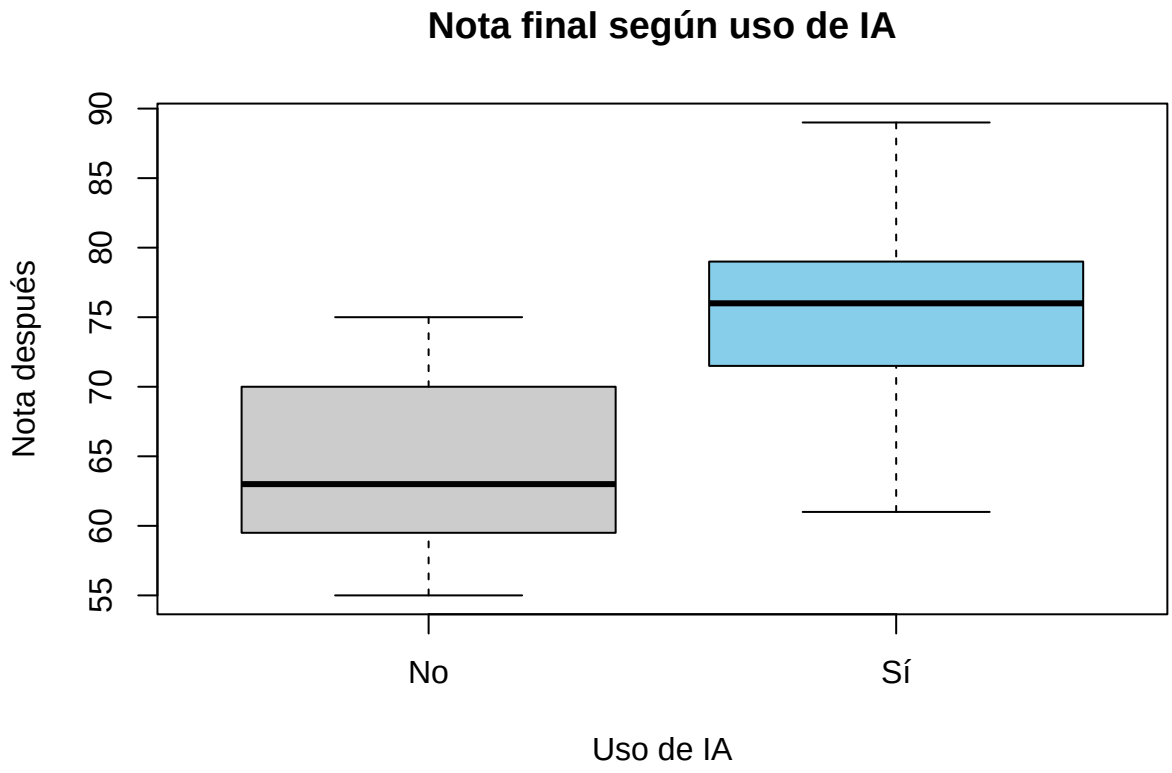
```r
cor(datos$nota_antes, datos$nota_despues)
```

```
## [1] 0.7562679
```

El diagrama de dispersión muestra una tendencia lineal positiva bien definida entre la calificación previa del estudiante y su calificación final. A medida que aumenta la nota antes, también aumenta la nota después. La línea de regresión (en rojo) refleja claramente esta pendiente ascendente. El coeficiente de correlación obtenido es 0.756 el cual indica una correlación lineal positiva fuerte. En términos prácticos, significa que los estudiantes que ya tenían un buen rendimiento académico tienden a mantener ese buen desempeño posteriormente.

3) Uso de IA vs Nota despues

```r
boxplot(nota_despues ~ usa_ia, data = datos,
        col = c("gray80", "skyblue"),
        xlab = "Uso de IA",
        ylab = "Nota después",
        main = "Nota final según uso de IA")
```




```r
tapply(datos$nota_despues, datos$usa_ia, mean)
```

```
##    No    Sí
## 64.40 75.15
```

El diagrama de cajas permite comparar la distribución de las calificaciones finales entre los estudiantes que usan IA y los que no la usan.

Se observa que:

La mediana de las notas finales es claramente mayor en el grupo que sí usa IA.

La mayor parte de las calificaciones del grupo "Sí" se concentra en valores más altos que las del grupo "No".

Aunque existe cierta variabilidad en ambos grupos, la distribución completa del grupo que usa IA está desplazada hacia arriba.

Esto indica que, de manera descriptiva, los estudiantes que utilizan herramientas de inteligencia artificial tienden a obtener mejores calificaciones finales en comparación con aquellos que no las utilizan.

Modelamiento

a. Formulación del modelo

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Dónde:

- Y : nota despues (Calificación final del estudiante)
- X_1 : horas estudio (Horas de estudio diarias)
- X_2 : usa ia (Variable dummy: 1 si usa IA, 0 si no usa)
- X_3 : nota antes (Calificación previa del estudiante)
- β_0 : intercepto del modelo
- $\beta_1, \beta_2, \beta_3$: coeficientes de regresión para cada variable explicativa
- ϵ : error del modelo

b. Ajuste del modelo

Se construye el modelo considerando como variable dependiente la nota final del estudiante y como variables explicativas las horas de estudio, el uso de IA y la nota previa.

```
modelo = lm(nota_despues ~ horas_estudio + usa_ia + nota_antes, data = datos)
modelo |> coef()
```

```
##   (Intercept) horas_estudio      usa_iaSí    nota_antes
##     3.7374711     0.1521088     9.9252391     0.9346624
```

Este comando indica que:

- nota_despues es la variable a explicar,
- horas_estudio, usa_ia y nota_antes actúan como predictores,

el modelo se ajusta usando el método de mínimos cuadrados ordinarios.

c. Interpretación de los coeficientes de regresión

$$\hat{Y} = 3.737 + 0.152X_1 + 9.925X_2 + 0.935X_3$$

- 3.737: Es el valor esperado de la nota_despues cuando las horas de estudio, el uso de IA y la nota anterior son cero. En este contexto académico, funciona principalmente como un ajuste matemático del modelo.
- 0.152: Por cada hora adicional de estudio diario, la nota final aumenta en 0.15 puntos manteniendo constante el uso de IA y su nota previa.
- 9.925: Los estudiantes que sí usan IA obtienen, en promedio, 9.92 puntos más en su calificación final que aquellos que no la usan, manteniendo constantes las horas de estudio y su nota previa.
- 0.935: Por cada punto adicional que el estudiante tenía en su calificación previa, su nota final aumenta en 0.93 puntos manteniendo constante las horas de estudio y el uso de IA.

d. Coeficiente de determinación

El coeficiente de determinación R^2 mide qué proporción de la variabilidad de la variable dependiente (nota_despues) es explicada por el conjunto de variables independientes incluidas en el modelo: horas de estudio, uso de IA y nota previa.

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

```
## [1] 0.9275558
```

Esto indica que aproximadamente el 92.8% de la variación en las calificaciones finales de los estudiantes puede ser explicada por las variables del modelo. Este es un valor muy alto, lo que sugiere que el modelo tiene un gran poder explicativo.

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

```
## [1] 0.9252919
```

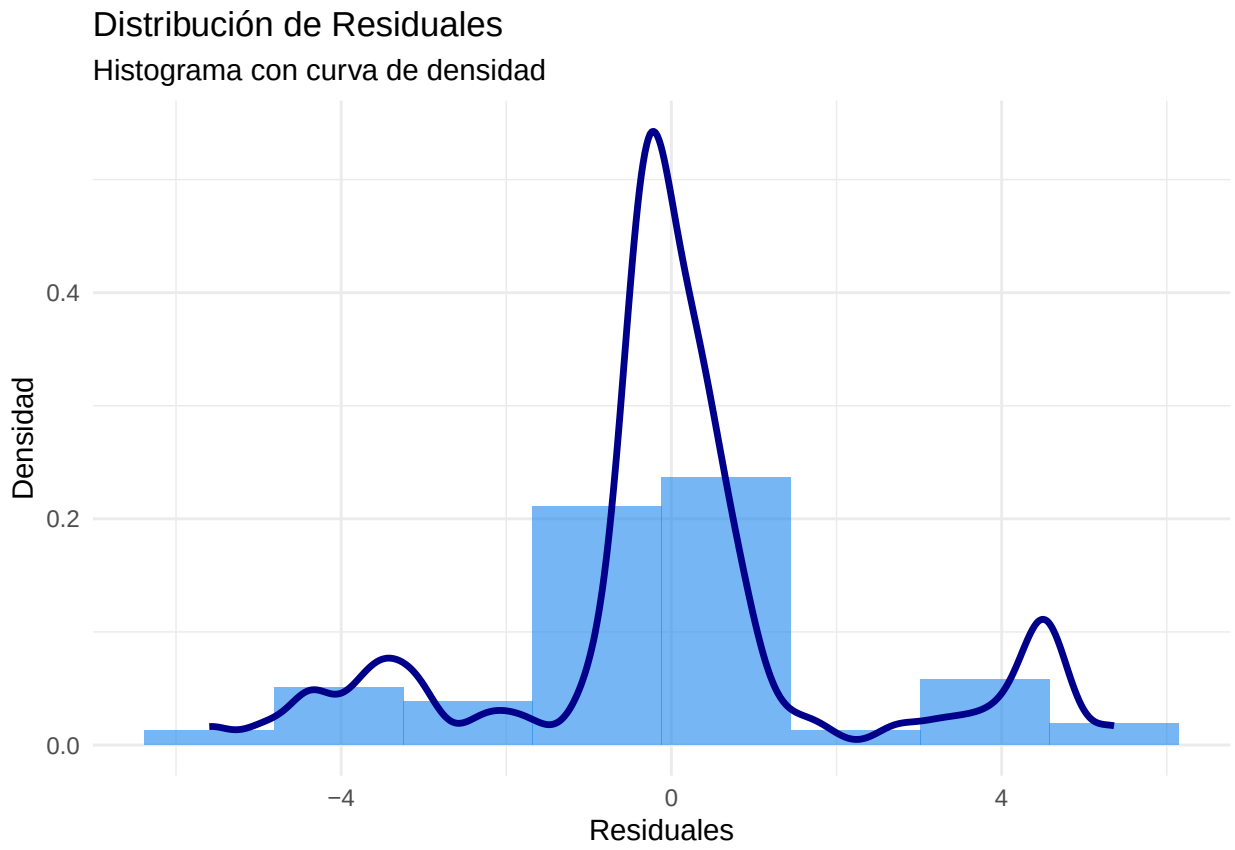
El coeficiente de determinación ajustado fue: $R^2_{ajustado} = 92.5\%$ Este valor es muy cercano al R^2, lo que indica que las variables incluidas son relevantes y que el modelo no está sobreajustado. Es decir, cada variable aporta información útil para explicar el rendimiento académico final.

e. Verificación del supuesto de normalidad de errores

Este supuesto establece que los residuos del modelo de regresión deben seguir aproximadamente una distribución normal. Su verificación es importante porque garantiza la validez de las pruebas de hipótesis y de los intervalos de confianza asociados a los coeficientes del modelo.

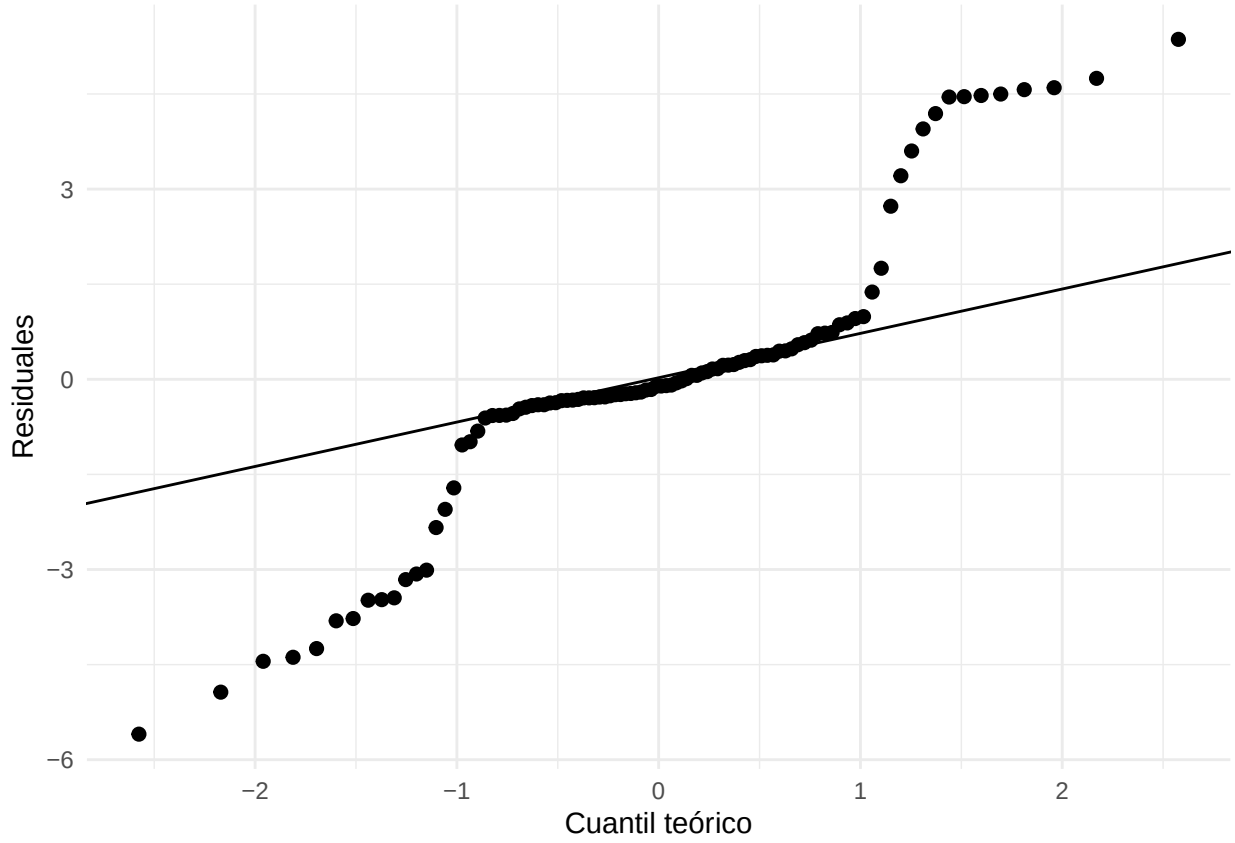
```
library(ggplot2)
# Guardamos los residuales en un objeto
res <- residuals(modelo)

# Gráfico de Normalidad sugerido
```

```r
data.frame(res) |>
  ggplot(aes(x = res)) +
  geom_histogram(aes(y = ..density..),
                 bins = round(1 + 3.3 * log10(nrow(datos))),
                 fill = "dodgerblue2",
                 alpha = 0.6) +
  geom_density(size = 1.2, color = "darkblue") +
  labs(title = "Distribución de Residuales",
       subtitle = "Histograma con curva de densidad",
       x = "Residuales",
       y = "Densidad") +
  theme_minimal()
```



En el análisis gráfico de los residuales, se observa una distribución aproximadamente simétrica. Sin embargo, la curva de densidad muestra una forma leptocúrtica (apuntamiento pronunciado), lo que sugiere que los errores están muy concentrados alrededor de la media.

```r
data.frame(res) |>
ggplot(aes(sample=res))+
stat_qq(size = 2) +
stat_qq_line(distribution = stats::qnorm)+
labs(x = "Cuantil teórico", y = "Residuales")+
theme_minimal()
```



Al observar el gráfico Q-Q, se aprecia que la mayoría de los residuales se alinean con la diagonal teórica, especialmente en el sector central. No obstante, en los extremos se observa unalejamiento de los puntos, lo cual es coherente con la leptocurtosis detectada en el histograma.

- Coeficiente de asimetría

$H_0 : As = O$

$H_1 : As \neq 0$

$\alpha = 0.5$

```
library(moments)
res |> agostino.test()
```

```
##
##  D'Agostino skewness test
##
## data:  res
## skew = 0.15699, z = 0.67883, p-value = 0.4972
## alternative hypothesis: data have a skewness
```

Dado que el p-valor (0.4972) es notablemente mayor al nivel de significancia $\alpha = 0.05$, no se rechaza la hipótesis nula (H_0). Esto significa que no existe evidencia suficiente para afirmar que los errores son asimétricos. En términos estadísticos, los residuales presentan una simetría aceptable, lo cual es un indicio favorable para el cumplimiento del supuesto de normalidad.

- Curtosis

$H_0 : k = 3$

$H_1 : k \neq 3$

$\alpha = 0.05$

```
res |> anscombe.test()
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  res
## kurt = 3.7859, z = 1.6453, p-value = 0.09991
## alternative hypothesis: kurtosis is not equal to 3
```

Observamos que el p-valor (0.09991) es mayor que el nivel de significancia. Por lo tanto, no se rechaza la hipótesis nula (los datos son mesocúrticos).

En conclusión, tras realizar las pruebas de D'Agostino y Anscombe-Glynn, se obtuvieron p-valores de 0.4972 y 0.0999 respectivamente. Dado que ambos son superiores al nivel de significancia $\alpha = 0.05$, no se rechaza la hipótesis nula de simetría ni la de mesocurtosis. Por lo tanto, a pesar de las irregularidades visuales observadas en los gráficos, contamos con evidencia estadística suficiente para asumir el cumplimiento del supuesto de normalidad en los residuales del modelo.

- Prueba de normalidad

H_0 : los errores siguen una distribucion normal

H_1 : los errores no siguen una distribucion normal

```
res |> shapiro.test()
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.89347, p-value = 7.057e-07
```

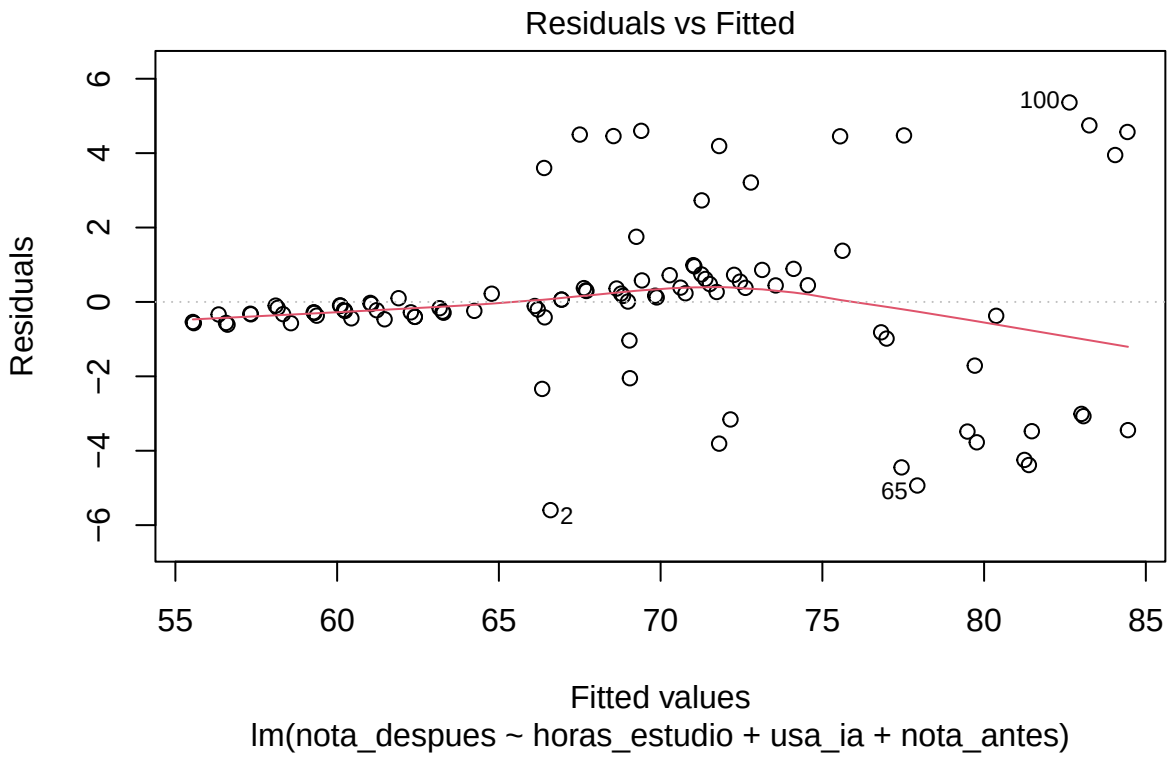
"Al realizar la prueba de Shapiro-Wilk, se obtuvo un p-valor de 7.057×10^{-7}. Al ser menor al nivel de significancia $\alpha = 0.05$, se rechaza la hipótesis nula, concluyendo que los residuales no siguen una distribución normal. Esta falta de normalidad, impulsada por la marcada leptocurtosis observada en el histograma, sugiere que las inferencias del modelo deben tomarse con cautela.

Al evaluar el supuesto de normalidad, se observa una discrepancia entre las pruebas. Por un lado, las pruebas de D'Agostino ($p = 0.4972$) y Anscombe-Glynn ($p = 0.0999$) sugieren que los residuales mantienen niveles aceptables de simetría y curtosis. Sin embargo, la prueba de Shapiro-Wilk arroja un p-valor de 7.057×10^{-7}, rechazando la hipótesis nula de normalidad global. Esta sensibilidad de Shapiro-Wilk confirma que existen desviaciones en las colas de la distribución, como se aprecia en la forma de 'S' del gráfico Q-Q."

f. Verificación del supuesto de homocedasticidad de errores

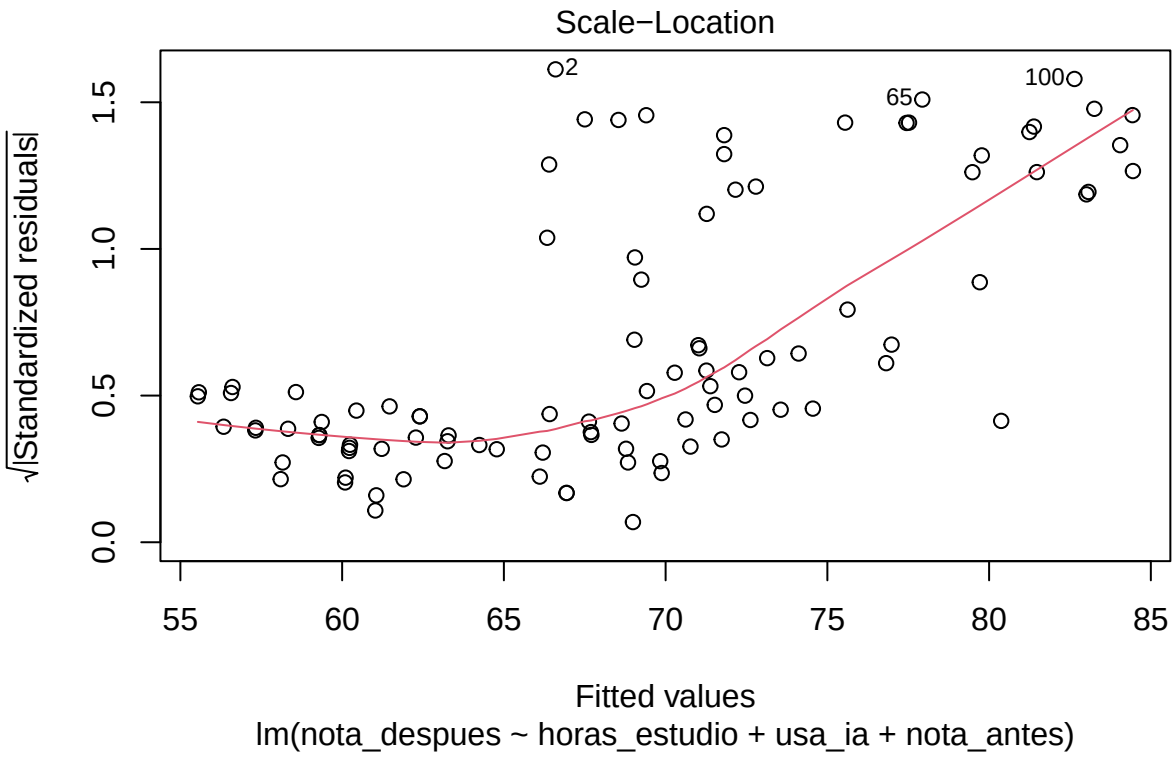
Este supuesto indica que la varianza de los residuos debe ser constante para todos los valores de la variable explicativa. Su cumplimiento asegura que las estimaciones de los coeficientes sean eficientes y que las inferencias estadísticas sean confiables.

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



En la primera imagen, se observa que a medida que los valores ajustados aumentan (hacia la derecha en el eje X), la dispersión de los puntos se vuelve mucho mayor. Los puntos no forman una "banda horizontal" uniforme, sino que parecen abrirse en forma de embudo o abanico.

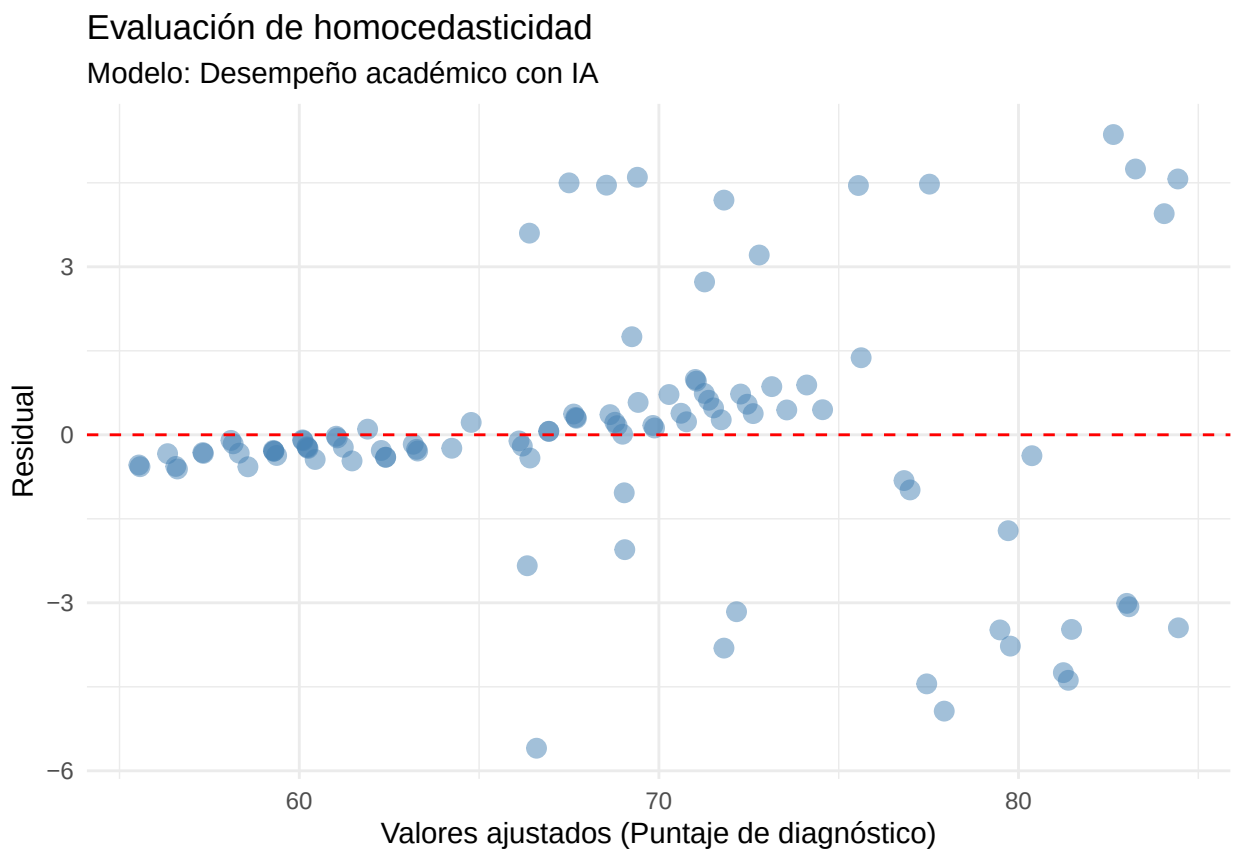
```
modelo |> plot(which=3)
```



En la segunda imagen, la línea roja tiene una tendencia ascendente marcada. Esto confirma que la magnitud de los residuos (la variabilidad) está creciendo junto con los valores predichos.

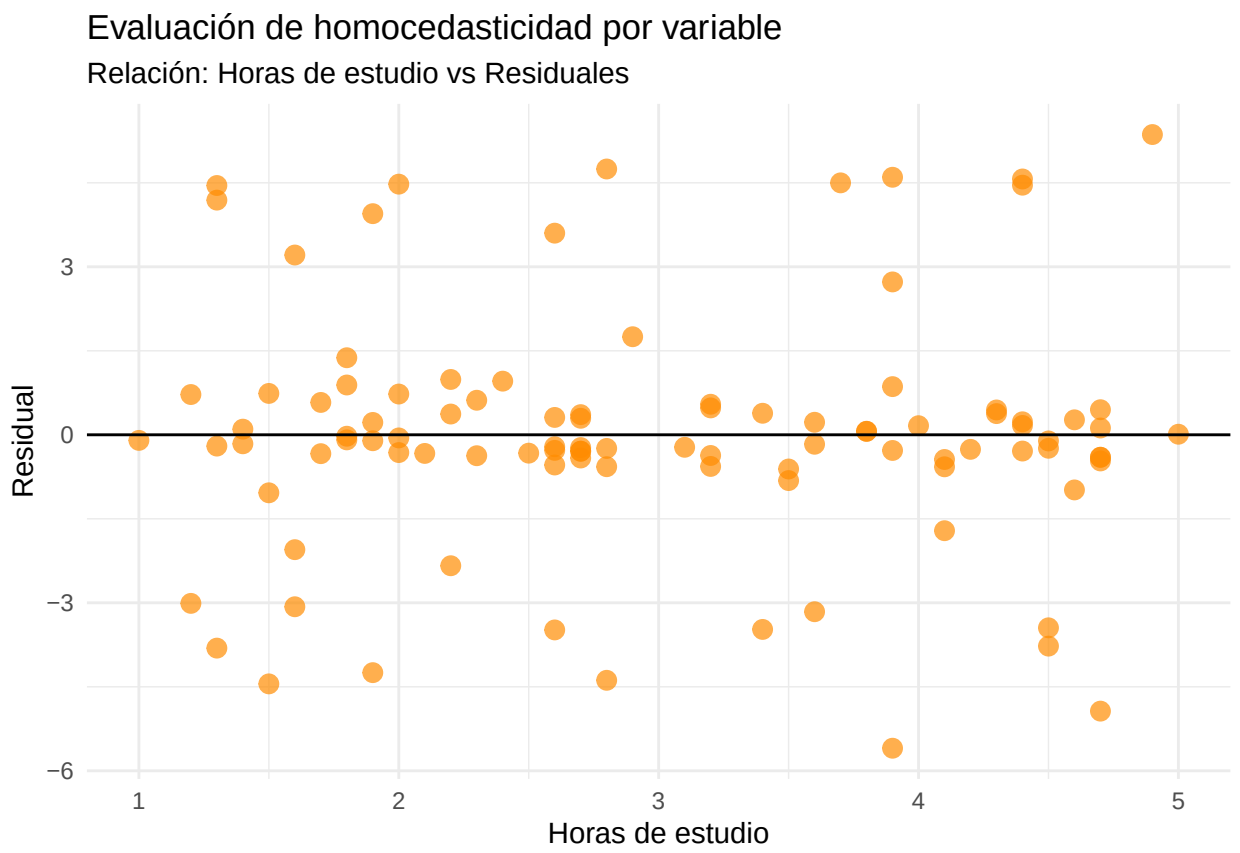
Ambos gráficos sugieren la presencia de heterocedasticidad (varianza no constante). Esto indica que el modelo es más preciso para predecir notas bajas que para predecir notas altas.

```
library(broom)
# Usamos augment() para obtener los valores ajustados (.fitted) y los residuales (.resid)
modelo |>
  augment() |>
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point(size = 3, color = "steelblue", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Valores ajustados (Puntaje de diagnóstico)",
       y = "Residual",
       title = "Evaluación de homocedasticidad",
       subtitle = "Modelo: Desempeño académico con IA") +
  theme_minimal()
```



Existe heterocedasticidad. Visualmente, el gráfico tiene forma de "abanico" o "embudo". Esto nos dice que, aunque el modelo capta la tendencia general, hay factores adicionales que afectan a los alumnos destacados que no estamos logrando explicar solo con las horas de estudio o el uso de IA.

```
# Usamos el modelo y la variable 'horas_estudio'
modelo |> augment() |>
  ggplot(aes(x = horas_estudio, y = .resid)) +
  geom_point(size = 3, color = "darkorange", alpha = 0.7) +
  geom_hline(yintercept = 0, color = "black", linetype = "solid") +
  labs(x = "Horas de estudio",
       y = "Residual",
       title = "Evaluación de homocedasticidad por variable",
       subtitle = "Relación: Horas de estudio vs Residuales") +
  theme_minimal()
```



Aunque el modelo identifica una tendencia, la heterocedasticidad detectada en la variable 'horas de estudio' sugiere que el impacto del estudio en la nota final es muy variable entre los alumnos destacados. Esto indica que nuestro modelo es útil, pero sus predicciones deben tomarse con cautela en el rango de alto rendimiento académico.

H_0: La varianza de los errores son constantes

H_1: La varianza de los errores no son constantes

$\alpha = 0.05$

```
library(car)
```

```
modelo |> ncvTest()
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 44.40423, Df = 1, p = 2.6711e-11
```

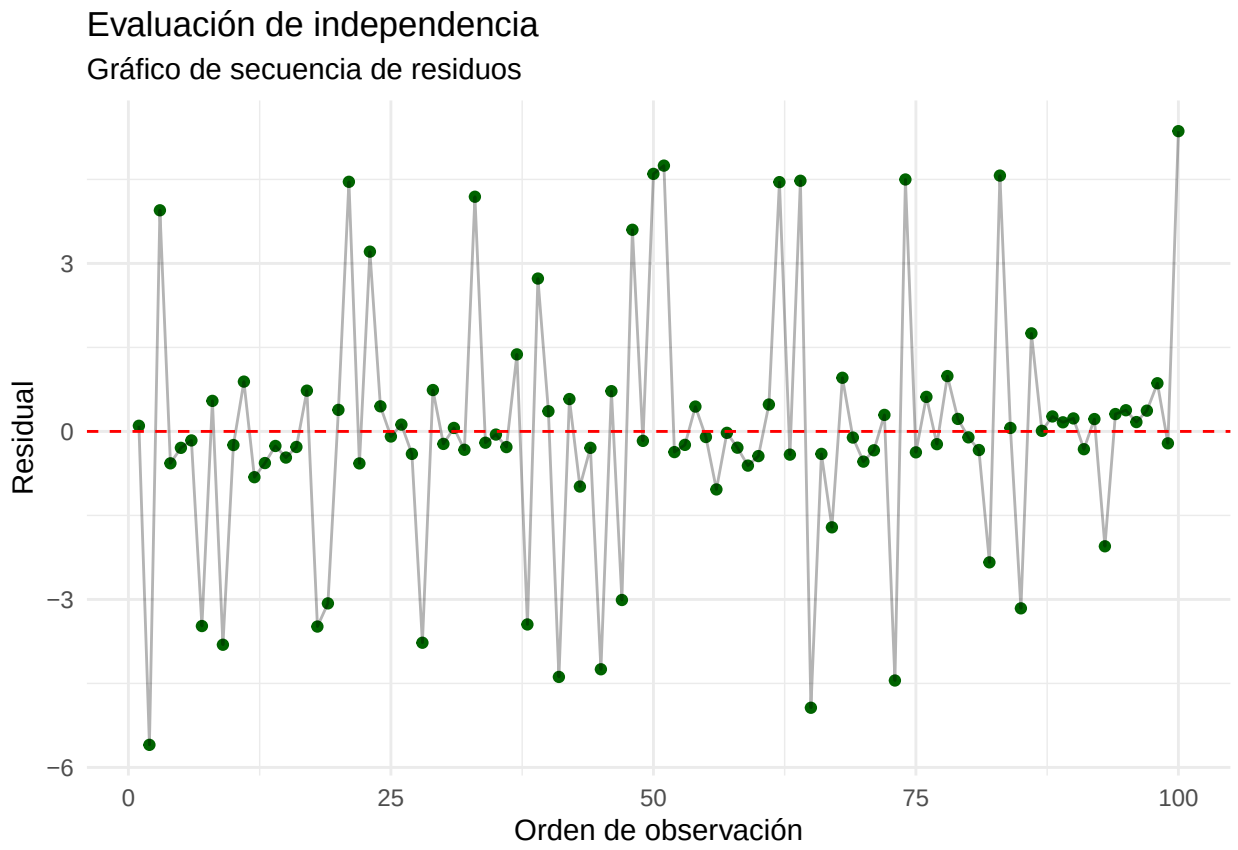
Se rechaza la hipotesis nula, por lo tanto no se verifica el supuesto de homogeneidad de varianza de los errores

g. Verificación del supuesto de independencia de errores

Este supuesto establece que los errores del modelo no deben estar correlacionados entre sí. Su verificación es fundamental para evitar sesgos en la estimación de los coeficientes y en las pruebas estadísticas.

- Usamos el gráfico de los residuales en orden

```
# Usamos los residuales de tu modelo (res) y el total de datos
data.frame(res) |>
  ggplot(aes(x = 1:nrow(datos), y = res)) +
  geom_point(size = 1.5, color = "darkgreen") +
  geom_line(alpha = 0.3) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(x = "Orden de observación",
       y = "Residual",
       title = "Evaluación de independencia",
       subtitle = "Gráfico de secuencia de residuos") +
  theme_minimal()
```



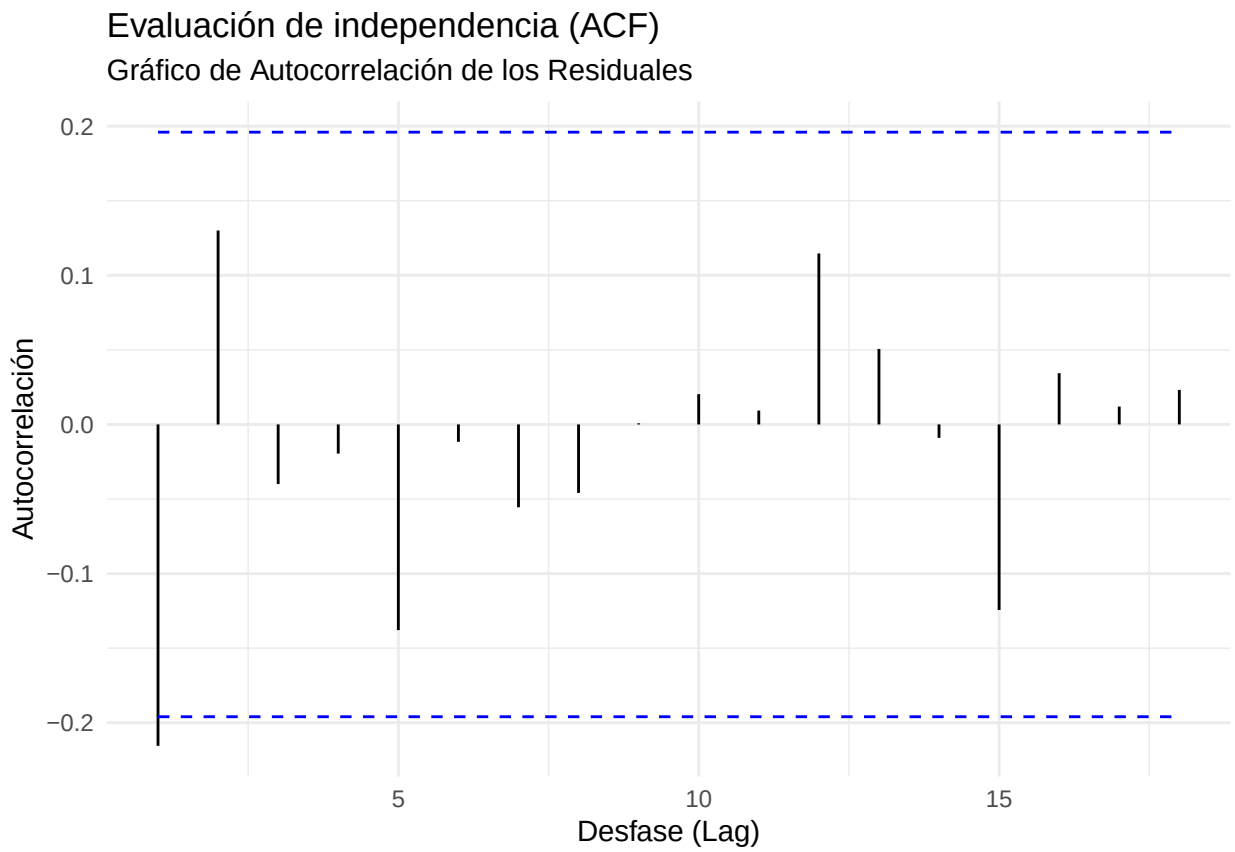
El gráfico de secuencia de residuos no muestra patrones sistemáticos, lo que indicaría que las observaciones son independientes entre sí.

- Usamos el correlograma

```
library(ggfortify)
```

```
# Usamos los residuales de tu modelo
res |>
  TSA::acf(lag = 18, plot = F) |>
  autoplot() +
```

```
  labs(x = "Desfase (Lag)",
       y = "Autocorrelación",
       title = "Evaluación de independencia (ACF)",
       subtitle = "Gráfico de Autocorrelación de los Residuales") +
  theme_minimal()
```



Para verificar la independencia de los errores, se generó un correlograma (ACF). Se observa que casi la totalidad de las barras de autocorrelación se encuentran dentro de las bandas de confianza (líneas azules), a excepción de una ligera desviación en el primer desfase. Esto sugiere que no existe una dependencia lineal significativa entre los residuos consecutivos, cumpliéndose de manera razonable el supuesto de independencia necesario para la validez del modelo.

- Usamos prueba de hipótesis para verificar la independencia con Durbin Watson

H_0: Los errores son independientes

H_1: Los errores no son independientes

$\alpha = 0.05$

```
library(car)
modelo |>
  durbinWatsonTest(alternative = "two.sided",
                   max.lag = 10,
                   reps = 1e5)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1   -0.2155056389      2.370510 0.06274
##    2    0.1300585592      1.613377 0.06314
##    3   -0.0399744913      1.919083 0.86530
##    4   -0.0195435120      1.877249 0.76196
##    5   -0.1378895450      2.113701 0.33642
##    6   -0.0116585487      1.860884 0.87386
##    7   -0.0555569733      1.923065 0.81350
##    8   -0.0459027999      1.894279 0.84510
##    9    0.0007482973      1.770341 0.73944
##   10    0.0203025993      1.730896 0.66634
##  Alternative hypothesis: rho[lag] != 0
```

Para validar el supuesto de independencia, se aplicó la prueba de Durbin-Watson extendida hasta 10 retardos. Los resultados muestran que para todos los niveles de desfase, los p-valores son superiores al nivel de significancia $\alpha = 0.05$ (p-valor lag 1 = 0.062). Al no rechazar la hipótesis nula, se concluye que no existe autocorrelación significativa en los residuales. Este hallazgo es consistente con el gráfico ACF, confirmando que el modelo cumple con el supuesto de independencia necesario para la inferencia estadística.

Aunque se detectó falta de normalidad y heterocedasticidad, lo cual sugeriría una transformación de variables, se ha decidido proceder con los datos originales para mantener la interpretabilidad pedagógica del tutorial. No obstante, se incluye una Nota de Precaución indicando que las inferencias (pruebas t y F) podrían presentar ligeras desviaciones.

i) Prueba de hipótesis global

Una vez ajustado el modelo, es necesario verificar si este tiene capacidad predictiva global. La Prueba de Hipótesis Global utiliza el estadístico F para determinar si existe una relación lineal entre la variable respuesta y el conjunto de variables explicativas.

$H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$: Ninguna de las variables explicativas influye linealmente en la nota después.

H_1 : Al menos un $\beta_j \neq 0$: Al menos una de las variables tiene un efecto significativo sobre la nota después.

$\alpha = 0.05$

```
library(broom)
# Usamos las variables: nota_despues (Y), horas_estudio, usa_ia y nota_antes (X's)

X = model.matrix(nota_despues ~ horas_estudio + usa_ia + nota_antes ,data=datos)
anova = aov(nota_despues ~ X, datos) |> tidy()
anova
```

```
## # A tibble: 2 x 6
##   term         df sumsq  meansq statistic   p.value
##   <chr>     <dbl> <dbl>   <dbl>     <dbl>     <dbl>
## 1 X             3 6086. 2029.        410.  1.45e-54
## 2 Residuals    96  475.    4.95       NA  NA
```

Se observa que el p-valor es mucho menor que el alpha, por lo tanto, se rechaza H0, entonces podemos concluir que al menos una de las variables tienen un efecto significativo sobre la variable respuesta (nota_despues).

"A pesar de la significancia global observada en esta prueba F, recordamos que las inferencias deben tomarse con cautela debido a la falta de normalidad en los errores"

j). Pruebas de hipótesis individuales

Esta prueba evalúa si cada variable independiente, de manera individual, tiene un efecto significativo sobre la nota final, manteniendo las demás constantes.

Una vez que sabemos que el modelo es útil globalmente, el siguiente paso es analizar cada variable por separado. ¿Realmente cada una aporta al modelo? Para esto, realizamos una Prueba t.

Definimos el modelo

```
modelo2 = lm(nota_despues ~ horas_estudio + usa_ia + nota_antes, datos)
modelo2 |> tidy()
```

```
## # A tibble: 4 x 5
##   term          estimate std.error statistic  p.value
##   <chr>            <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)      3.74     2.37       1.58  1.17e- 1
## 2 horas_estudio    0.152    0.200      0.762 4.48e- 1
## 3 usa_iaSí         9.93     0.459     21.6   1.16e-38
## 4 nota_antes       0.935    0.0370    25.3   3.11e-44
```

Paso 1: Formular las Hipótesis

- Variable: nota_antes

¿Es posible que la nota final aumente exactamente en la misma proporción (1 a 1) que la nota inicial?

$H_0 : \beta_{\text{nota\_antes}} = 1$

$H_1 : \beta_{\text{nota\_antes}} \neq 1$

Paso 2: Calculamos el estadístico de prueba

$$t_{calc} = \frac{0.935 - 1}{0.037}$$

```
(0.935-1)/0.037
```

```
## [1] -1.756757
```

Paso 3: Calculamos el t-critico, al ser una prueba bilateral, definimos t-crítico para un 95% de confianza (alpha = 0.05) para los 2 lados.

```
qt(0.025, 96)
```

```
## [1] -1.984984
```

```
qt(0.975, 96)
```

```
## [1] 1.984984
```

Al comparar el t-crítico con el t-calculado, no rechazamos H_0. Esto significa que sí se puede afirmar que existe una relación de 1 a 1 entre la nota inicial y la final, confirmando que la base académica previa se mantiene de forma exacta en el resultado.

- Variable: horas_estudio

¿Se puede afirmar que por cada hora adicional de estudio, la nota final aumenta en 1.5 puntos, manteniendo lo demás constante?

Paso 1: Formular las Hipótesis

$H_0 : \beta_{\text{horas\_estudio}} = 1.5$

$H_1 : \beta_{\text{horas\_estudio}} \neq 1.5$

Paso 2: Calculamos el estadístico de prueba

$$t_{calc} = \frac{0.152 - 1.5}{0.200}$$

```
(0.152-1.5)/0.200
```

```
## [1] -6.74
```

Paso 3: Calculamos el t-critico, al ser una prueba bilateral, definimos t-crítico para un 95% de confianza (alpha = 0.05) para los 2 lados.

```
qt(0.025, 96)
```

```
## [1] -1.984984
```

```
qt(0.975, 96)
```

```
## [1] 1.984984
```

Al comparar el t-crítico con el t-calculado, rechazamos H_0.Esto significa que no se puede afirmar que cada hora adicional de estudio incremente la nota exactamente en 1.5 puntos.

- Variable: usa_ia

Un informe previo indica que el uso de IA mejora las notas en 5 puntos. ¿Los datos actuales permiten afirmar que el impacto es distinto a esos 5 puntos?

Paso 1: Formular las Hipótesis

$H_0 : \beta_{\text{usa\_ia}} = 5$

$H_1 : \beta_{\text{usa\_ia}} \neq 5$

Paso 2: Calculamos el estadístico de prueba

$$t_{calc} = \frac{-9.925 - 5}{0.459}$$

```
(-9.925-5)/0.459
```

```
## [1] -32.51634
```

Paso 3: Calculamos el t-critico, al ser una prueba bilateral, definimos t-crítico para un 95% de confianza (alpha = 0.05) para los 2 lados.

```r
qt(0.025, 96)
```

```
## [1] -1.984984
```

```r
qt(0.975, 96)
```

```
## [1] 1.984984
```

Al comparar el t-crítico con el t-calculado, rechazamos H_0. Por lo tanto podemos concluir que existe evidencia estadística suficiente para afirmar que el uso de IA tiene un efecto diferente a 5 puntos.

En definitiva, el éxito del estudiante depende de conservar su base académica, pero se ve potenciado principalmente por el uso de herramientas tecnológicas, el cual supera en efectividad al incremento de horas de estudio tradicional.

Podemos concluir que o todas las variables se comportan como predecimos: mientras que la base académica es una constante predecible (1 a 1), la IA y las horas de estudio requieren un análisis más profundo al alejarse de los estándares teóricos de 5 y 1.5 puntos respectivamente

k. Estimación de una media (puntual e intervalar)

El objetivo es estimar la nota final promedio esperada de un grupo de estudiantes que presentan características específicas en las variables explicativas del modelo. A diferencia de la predicción individual, aquí se busca inferir el valor medio poblacional de la variable respuesta condicionado a ciertos valores de las variables independientes. Esta estimación se realiza utilizando el modelo de regresión lineal múltiple previamente ajustado, manteniendo constantes los coeficientes estimados y modificando únicamente los valores de las variables explicativas.

Paso 1: Definición del perfil promedio a evaluar

Para la estimación de la media, se consideran los siguientes valores:

- Horas de estudio: 3 horas diarias
- Uso de IA: Sí
- Nota previa: 65 puntos

Estos valores no son arbitrarios, ya que:

- 3 horas de estudio corresponde aproximadamente al promedio observado en el dataset.
- 65 puntos es un valor cercano a la media de la variable nota_antes.
- El uso de IA es una categoría relevante y ampliamente representada en la muestra.

k.1 Estimación puntual de la media

Paso 2: Construcción del nuevo conjunto de datos (media poblacional)

Construimos un nuevo data.frame que representa el perfil promedio definido anteriormente.

```r
nuevo_promedio <- data.frame(
  horas_estudio = 3,
  usa_ia = "Sí",
  nota_antes = 65
)
```

A partir del modelo ajustado, se obtiene la estimación puntual de la nota final promedio correspondiente a este perfil.

```
predict(modelo, newdata = nuevo_promedio)
```

```
##        1
## 74.87209
```

La estimación puntual indica que la nota final promedio esperada para estudiantes que estudian 3 horas diarias, utilizan herramientas de inteligencia artificial y tenían una nota previa de 65 es 74.87 puntos.

k.2 Estimación intervalar de la media (Intervalo de Confianza)

Para cuantificar la incertidumbre asociada a la estimación de la media, construimos un intervalo de confianza al 95%.

```
predict(modelo, newdata = nuevo_promedio,
        interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 74.87209 74.16967 75.57452
```

Con un 95% de confianza, la nota final promedio de los estudiantes que estudian 3 horas diarias, utilizan herramientas de inteligencia artificial y tenían una nota previa de 65 puntos se encuentra entre 74.17 y 75.57 puntos.

"Aunque el modelo presenta un alto poder explicativo, se ha detectado la presencia de heterocedasticidad y desviaciones de la normalidad en los errores. Por tal motivo, el intervalo de confianza para la media debe interpretarse con cautela, ya que la varianza no constante puede afectar la precisión de la inferencia estadística. No obstante, la estimación se mantiene con fines descriptivos y pedagógicos."

l. Predicción de un nuevo valor (puntual e intervalar)

En esta sección se realiza la predicción de la nota final de un estudiante individual, considerando características específicas. A diferencia de la pregunta anterior, aquí se incorpora no solo la incertidumbre del modelo, sino también la variabilidad individual, por lo que se utiliza un intervalo de predicción.

Paso 1: Selección y justificación del perfil del estudiante

Consideremos un estudiante con las siguientes características:

- Horas de estudio: 4 horas diarias
- Uso de IA: Sí
- Nota previa: 70 puntos

Estos valores son coherentes porque:

- Se encuentran dentro del rango observado en el dataset.
- Representan un estudiante con desempeño previo relativamente alto.
- Permiten analizar el comportamiento del modelo en un nivel superior al promedio.

l.1 Predicción puntual

Paso 2: Construcción del nuevo conjunto de datos (estudiante individual)

Se ingresa un nuevo data.frame que representa al estudiante individual a predecir.

```
nuevo_estudiante <- data.frame(
  horas_estudio = 4,
  usa_ia = "Sí",
  nota_antes = 70
)
```

Paso 3: Predicción puntual del nuevo valor

Aqui obtenemos la predicción puntual de la nota final del estudiante utilizando el modelo ajustado.

```
predict(modelo, newdata = nuevo_estudiante)
```

```
##        1
## 79.69751
```

Utilizando el modelo de regresión lineal múltiple ajustado, podemos predecir que la nota final esperada para un estudiante que estudia 4 horas diarias, utiliza herramientas de inteligencia artificial y tenía una nota previa de 70 puntos es aproximadamente 79.70 puntos.

l.2 Predicción intervalar (Intervalo de Predicción)

Para incorporar la variabilidad individual y la incertidumbre del modelo, se construye un intervalo de predicción al 95%.

```
predict(modelo, newdata = nuevo_estudiante,
        interval = "prediction", level = 0.95)
```

```
##        fit      lwr      upr
## 1 79.69751 75.19517 84.19985
```

Con un 95% de confianza, la nota final de un estudiante que estudia 4 horas diarias, utiliza herramientas de inteligencia artificial y tenía una nota previa de 70 puntos se encontrará entre 75.20 y 84.20 puntos.

Este intervalo de predicción incorpora tanto la incertidumbre del modelo como la variabilidad individual inherente al desempeño académico de cada estudiante.

“Debido a la presencia de heterocedasticidad en los errores del modelo, el intervalo de predicción puede no reflejar con total precisión la verdadera dispersión de los valores individuales. En consecuencia, la predicción debe interpretarse como una aproximación razonable y no como un valor exacto.”