Tutorial básico de Regresión: Impacto de la IA y el Estudio

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

2026-01-26

Contents

Introducción 1
Caso (dataset) 2
Definición de variables 2
Exploración de datos 3
Resumen descriptivo univariado 4
Resumen descriptivo bivariado 9
Modelamiento 13
Formulación del modelo 13
Ajuste del modelo 13
Interpretación de los coeficientes de regresión 14
Coeficiente de determinación 14
Verificación del supuesto de normalidad de errores 15
Verificación del supuesto de homocedasticidad de errores 20
Verificación del supuesto de independencia de errores 26
Significancia del Modelo de Regresión 29
Inferencia y Predicción del Modelo 32
Conclusión final 35

[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.

• 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
  )
```

```r
# 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
```

Resumen descriptivo univariado

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.

• Variable Dependiente: Nota Final (nota_despues)

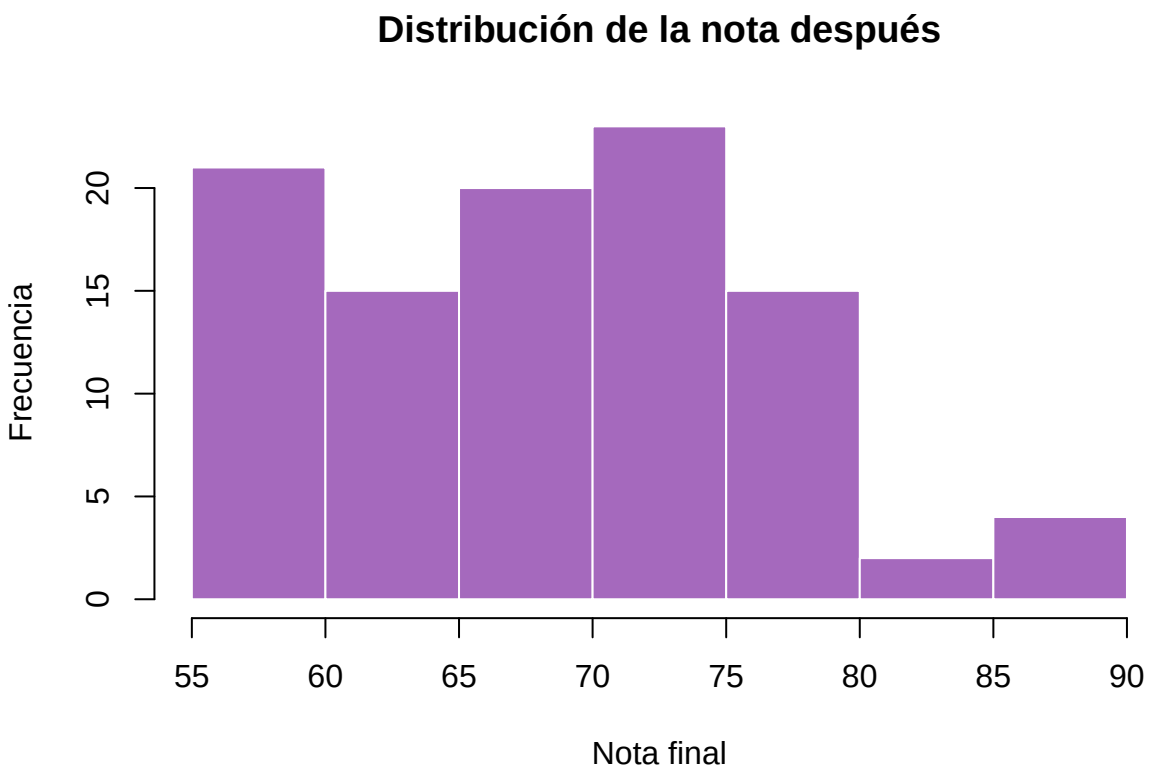
```r
# Para comprender el comportamiento de la variable dependiente, no solo visualizamos su distribución, s

# Estadístico de tendencia central
summary(datos$horas_estudio)
```

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

```r
# Estadístico de dipersión
sd(datos$horas_estudio)
```

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

```r
#Gráfico
hist(datos$nota_despues,
     col = "#A569BD",
     border = "white",
     main = "Distribución de la nota después",
     xlab = "Nota final",
     ylab = "Frecuencia")
```



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.

• Variable Independiente: Horas de Estudio (horas_estudio)

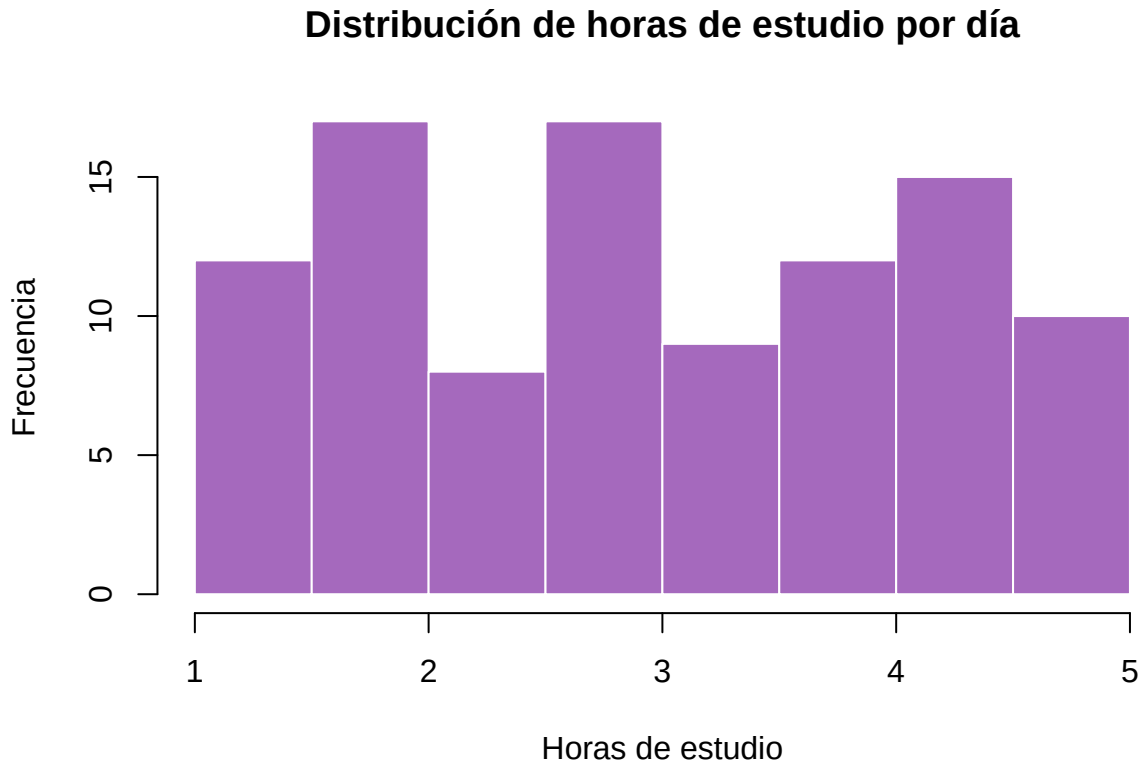
```
# Para comprender el comportamiento de la variable dependiente, no solo visualizamos su distribución, s

# Estadístico de tendencia central
summary(datos$horas_estudio)
```

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

```
# Estadístico de dipersión
sd(datos$horas_estudio)
```

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

```r
# Gráfico
hist(datos$horas_estudio,
     col = "#A569BD",
     border = "white",
     main = "Distribución de horas de estudio por día",
     xlab = "Horas de estudio",
     ylab = "Frecuencia")
```



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.

• Variable Independiente: Nota Anterior (nota_antes)

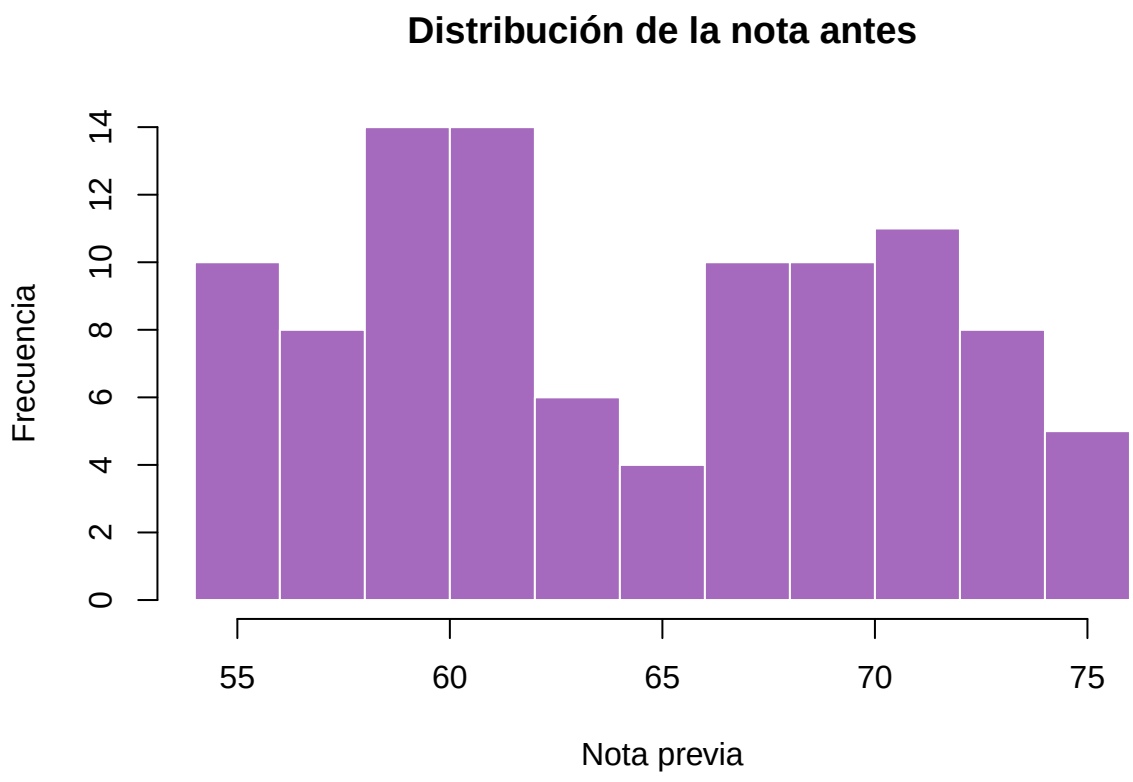
```r
# Para comprender el comportamiento de la variable dependiente, no solo visualizamos su distribución, s

# Estadístico de tendencia central
summary(datos$nota_antes)
```

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

```r
# Estadístico de dipersión
sd(datos$nota_antes)
```

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

```r
# Gráfico
hist(datos$nota_antes,
     col = "#A569BD",
     border = "white",
     main = "Distribución de la nota antes",
     xlab = "Nota previa",
     ylab = "Frecuencia")
```



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.

• Variable Categórica: Uso de IA (usa_ia)

Para esta variable, en lugar de un histograma, lo ideal es mencionar un conteo o gráfico de barras

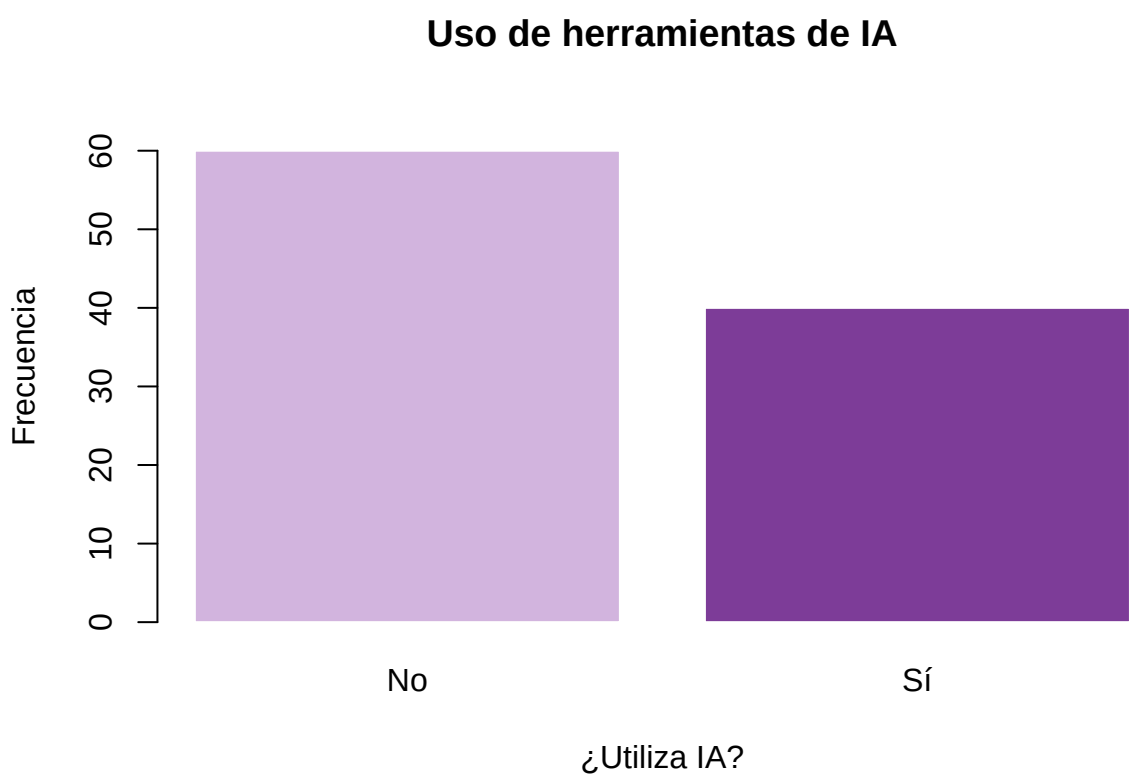
```
# Frecuencia Absoluta
table(datos$usa_ia)
```

```
##
## No Sí
## 60 40
```

```
# Frecuencia Relativa
prop.table(table(datos$usa_ia))
```

```
##
##  No  Sí
## 0.6 0.4
```

```
# Gráfico
barplot(table(datos$usa_ia),
        col = c("#D2B4DE", "#7D3C98"),
        border = "white",
        main = "Uso de herramientas de IA",
        ylab = "Frecuencia",
        xlab = "¿Utiliza IA?",
        ylim = c(0, max(table(datos$usa_ia)) + 5))
```



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.

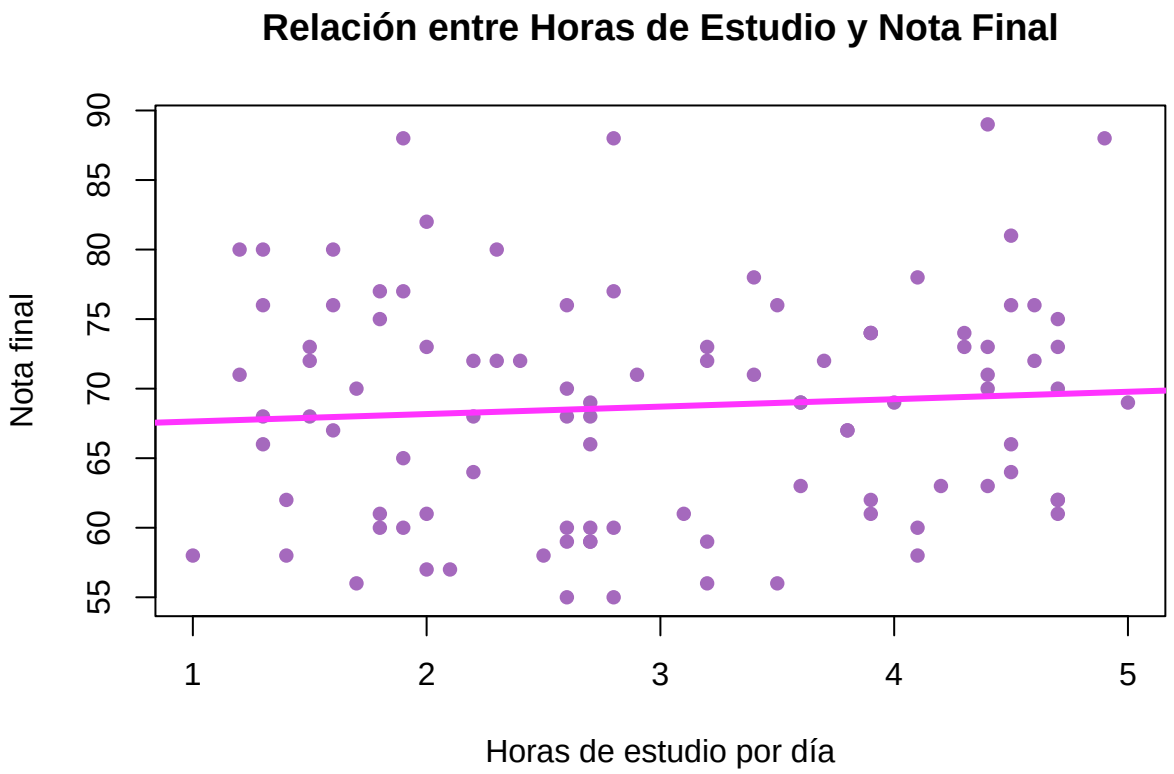
Resumen descriptivo bivariado

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
# Gráfico de dispersión bivariado
plot(datos$horas_estudio, datos$nota_despues,
     col = "#A569BD",
     pch = 16,
     xlab = "Horas de estudio por día",
     ylab = "Nota final",
     main = "Relación entre Horas de Estudio y Nota Final")
```

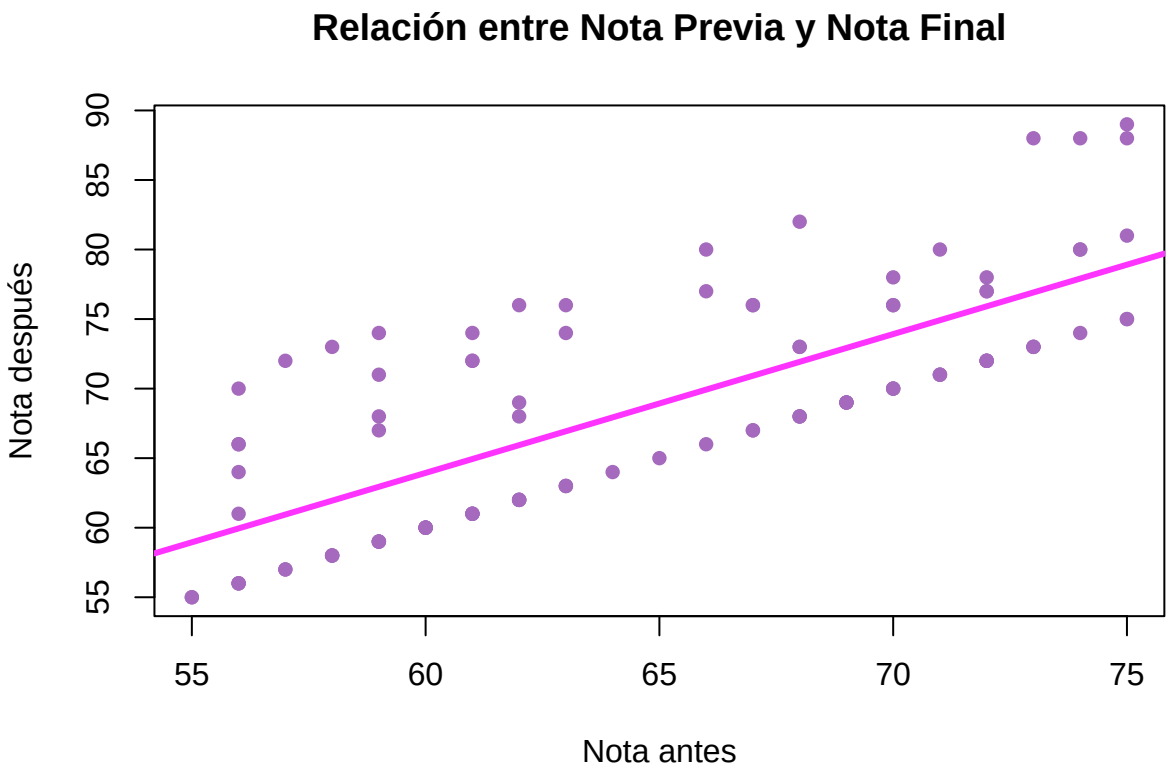
```r
# Línea de regresión lineal simple
abline(lm(nota_despues ~ horas_estudio, data = datos),
       col = "#FF33FF",
       lwd = 3)
```



- **Los Puntos (•):** Representan a cada estudiante.
- **La Línea de Tendencia:** Esta línea es una vista previa de nuestra regresión. Si sube de izquierda a derecha, indica una **correlación positiva**: a más horas de estudio, mayor nota esperada.
- **Dispersión:** Observar qué tan cerca están los puntos de la línea nos da una idea visual de qué tan fuerte es la relación antes de calcular el R^2.

2) Nota antes vs Nota despues

```r
# Gráfico de dispersión: Nota antes vs Nota después
plot(datos$nota_antes, datos$nota_despues,
     col = "#A569BD",
     pch = 16,
     xlab = "Nota antes",
     ylab = "Nota después",
     main = "Relación entre Nota Previa y Nota Final")

# Línea de tendencia
abline(lm(nota_despues ~ nota_antes, data = datos),
       col = "#FF33FF",
       lwd = 3)
```




```
# Cálculo de la correlación
cor(datos$nota_antes, datos$nota_despues)
```

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

¿Por qué incluimos el `cor()` aquí?

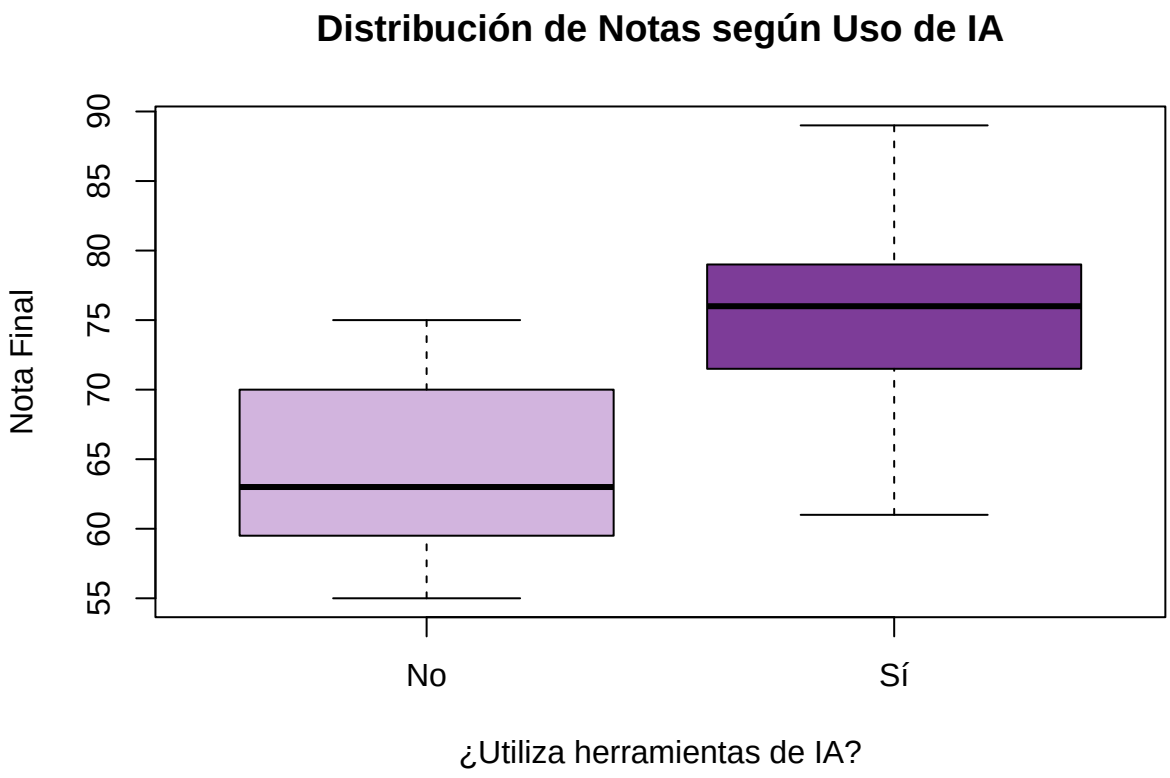
Como este es un **tutorial pedagógico**, es vital explicar que el número que sale al final (`cor`) es la traducción matemática de lo que vemos en el gráfico:

- **Interpretación del Gráfico:** "Al ver que los puntos morados siguen de cerca la línea rosada, visualizamos que existe una relación lineal fuerte".
- **Interpretación del `cor()`:** "El coeficiente de correlación de Pearson nos da un valor entre -1 y 1. Cuanto más cerca esté de 1, más fuerte es la relación positiva entre la base académica del alumno y su resultado final".

3) Uso de IA vs Nota despues

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

```r
# Boxplot comparativo: Nota vs Uso de IA
boxplot(nota_despues ~ usa_ia, data = datos,
        col = c("#D2B4DE", "#7D3C98"), # Lila para 'No', Morado para 'Sí'
        xlab = "¿Utiliza herramientas de IA?",
        ylab = "Nota Final",
        main = "Distribución de Notas según Uso de IA")
```




```r
# Cálculo de promedios por grupo
tapply(datos$nota_despues, datos$usa_ia, mean)
```

```
##    No    Sí
## 64.40 75.15
```

- **Comparación de Medianas:** La línea negra dentro de cada caja nos indica el valor central de las notas para cada grupo.
- **Dispersión:** El tamaño de las cajas moradas nos permite ver qué grupo tiene notas más variadas.

Si observamos un desplazamiento hacia arriba en el grupo "Sí", existe una relación positiva preliminar entre el uso de tecnología y el rendimiento. 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.

Síntesis del Análisis Bivariado

Tras explorar la relación entre nuestra variable dependiente (Nota Final) y cada uno de los predictores, podemos concluir lo siguiente:

- Fuerza Académica: Existe una correlación positiva y marcada entre la Nota Anterior y el resultado final, lo que sugiere que la base previa es el predictor más sólido.
- Impacto del Esfuerzo: El número de Horas de Estudio también muestra una tendencia ascendente, aunque con una dispersión que justifica un análisis más profundo mediante el modelo.
- Efecto de la Tecnología: El análisis comparativo (boxplot) revela que el grupo que Sí usa IA mantiene una mediana de notas superior al grupo que no la utiliza.

Conclusión: Estos hallazgos visuales y numéricos (proporcionados por cor y tapply) nos dan la "luz verde" para proceder con la Regresión Lineal Múltiple. Al haber confirmado que todas las variables tienen una relación lógica con la nota, nuestro modelo final buscará cuantificar exactamente cuánto aporta cada factor al éxito del estudiante.

Modelamiento

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

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.

```
# Creamos el modelo: Variable Respuesta ~ Predictores
modelo = lm(nota_despues ~ horas_estudio + usa_ia + nota_antes, data = datos)

# Vemos solo los Coeficientes (los Betas)
modelo |> coef()
```

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

¿Qué nos dice el comando `coef()`?

Al ejecutar `modelo |> coef()`, R nos devuelve unos números llamados **Coeficientes** (β). ¡Es lo que buscábamos!.

- **El Intercepto:** Es el punto de partida. Si todas las variables fueran cero, esta sería la nota base.
- **Los Betas (β_i):** Cada número te dice cuánto aumenta la nota final por cada unidad que sube esa variable.
  - *Por ejemplo:* Si el coeficiente de `horas_estudio` es 0.5, significa que por cada hora extra de estudio, ¡tu nota sube medio punto!.

Tip: Fíjate bien en el signo de los coeficientes. Si es positivo (+), la variable ayuda a subir la nota; si es negativo (−), la variable la disminuye.

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.

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.

• El coeficiente de determinación ajustado fue: $R^2_{ajustado}$: Nos indica si el modelo es realmente bueno o si solo se está amontonando variables

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

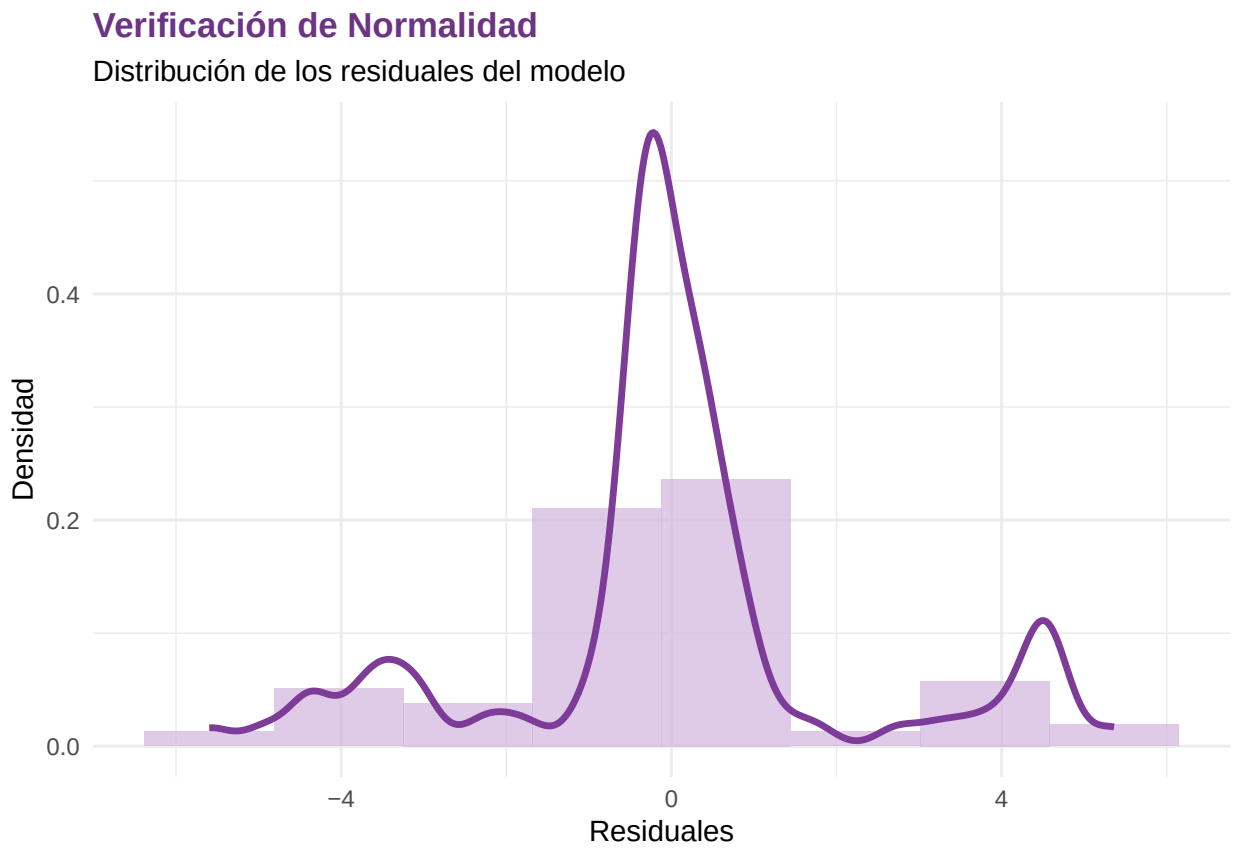
```
## [1] 0.9252919
```

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.

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.

```r
# Activamos la librería
library(ggplot2)

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

# Gráfico de Normalidad
data.frame(res) |>
  ggplot(aes(x = res)) +
  # Histograma en lila claro (coincide con los grupos 'No')
  geom_histogram(aes(y = ..density..),
                 bins = round(1 + 3.3 * log10(nrow(datos))),
                 fill = "#D2B4DE",
                 alpha = 0.7) +
  # Curva de densidad en morado fuerte (coincide con los grupos 'Sí')
  geom_density(linewidth = 1.2, color = "#7D3C98") +
  labs(title = "Verificación de Normalidad",
       subtitle = "Distribución de los residuales del modelo",
       x = "Residuales",
       y = "Densidad") +
  theme_minimal() +
  theme(plot.title = element_text(color = "#6C3483", face = "bold"))
```



¿Cómo entendemos este gráfico?

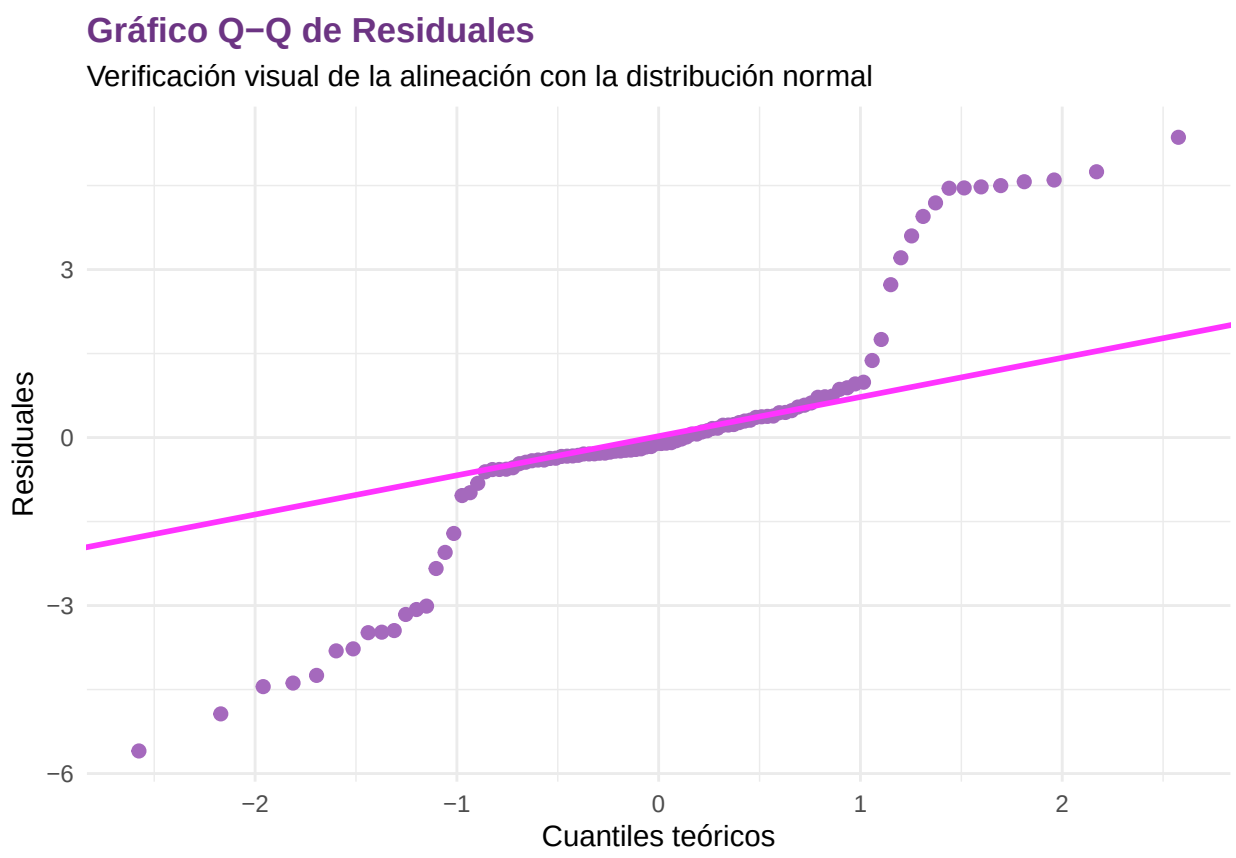
- **La forma de campana:** "Buscamos que el histograma lila siga la forma de la curva morada intensa. Si se parecen a una campana simétrica, nuestro supuesto de normalidad va por buen camino".
- **El centro en cero:** "Los residuales deben estar centrados mayoritariamente en el 0, lo que indica que el modelo no se equivoca sistemáticamente hacia un lado".

Para el conjunto de datos que estamos trabajando, observamos 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.

> **Mini-Diccionario: Mesocúrtica:** Normal y equilibrada (Lo que buscamos). **Leptocúrtica:** Muy punteaguda (Errores muy concentrados). **Platicúrtica:** Muy plana (Errores muy dispersos).

```
# Activación de paquete
library(ggplot2)

# Gráfico de Probabilidad Normal (Q-Q Plot)
data.frame(res) |>
  ggplot(aes(sample = res)) +
  stat_qq(size = 2, color = "#A569BD") +
  stat_qq_line(distribution = stats::qnorm, color = "#FF33FF", linewidth = 1) +
  labs(title = "Gráfico Q-Q de Residuales",
```

```
      subtitle = "Verificación visual de la alineación con la distribución normal",
      x = "Cuantiles teóricos",
      y = "Residuales") +
 theme_minimal() +
 theme(plot.title = element_text(color = "#6C3483", face = "bold"))
```



¿Qué estamos viendo en este gráfico? Imagina que la línea magenta es una **pista de aterrizaje** y los puntos morados son aviones:

- **¿Qué es lo ideal?** Queremos que todos los puntos morados estén **encima o muy cerca** de la línea magenta. Si se alinean como soldados, podemos asegurar que nuestros residuales siguen una distribución normal.
- **¿Qué pasa si se alejan en las puntas?** Si los puntos se curvan hacia arriba o hacia abajo al final de la línea (en las "colas"), significa que tenemos valores extremos o que nuestra campana es más ancha o flaca de lo normal.

¿Qué nos dicen las "puntas" del gráfico?

¡Ojo con los extremos!

- Puntas hacia afuera (Efecto "S"): Si los puntos se alejan de la línea en las esquinas formando una especie de "S", significa que tienes colas pesadas (distribución Leptocúrtica). ¡Hay más valores extremos de lo esperado!
- Puntas que caen o suben rápido: Si los puntos se curvan alejándose mucho, podrías tener una distribución Platicúrtica (más plana que la normal).
- Puntos rebeldes: Un solo puntito muy alejado en la punta es un Outlier. Ese estudiante tuvo un comportamiento tan raro que el modelo no pudo predecirlo bien.

En nuestro conjunto de datos, podemos apreciar que la mayoría de los residuales se alinean con la diagonal, especialmente en el sector central. No obstante, en los extremos se observa un alejamiento de los puntos, lo cual es coherente con la leptocurtosis detectada en el histograma.

A veces los ojos nos engañan, por eso usamos la estadística para confirmar si nuestra campana morada es realmente normal. Vamos a medir qué tan "chueca" está (Asimetría) y qué tan "puntiaguda" resulta (Curtosis).

Subsección A: Análisis de la Asimetría

$H_0 : As = O$

$H_1 : As \neq 0$

$\alpha = 0.5$

```
# Activamos la librería
library(moments)

## Realizamos el test de D'Agostino para asimetría
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
```

Interpretación de los resultados:

Dado que el p-valor 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.

¿Qué significa ese p-valor tan alto?. En estadística, un p-valor grande es como una "bandera blanca". Nos dice: "No hay pruebas para decir que algo anda mal". Como el 0.49 es mucho mayor que el 0.05 que pusimos de límite, aceptamos que nuestra campana está bien equilibrada.

Subsección B: Curtosis

$H_0 : k = 3$

$H_1 : k \neq 3$

$\alpha = 0.05$

```
# Prueba de Anscombe-Glynn para Curtosis
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
```

Interpretación de los resultados:

Observamos que el p-valor (0.09991) es mayor que nuestro nivel de significancia $\alpha = 0.05$. Por lo tanto, **no se rechaza la hipótesis nula**; esto indica que los datos son **mesocúrticos**.

Conclusión

Tras realizar las pruebas de **D'Agostino** (asimetría) y **Anscombe-Glynn** (curtosis), obtuvimos p-valores de 0.4972 y 0.0999 respectivamente. Como ambos son mayores a 0.05, no rechazamos la simetría ni la mesocurtosis.

En palabras simples: Aunque en los gráficos veas algunas "arrugas" o puntos fuera de lugar, estadísticamente tus errores se comportan como una campana normal.

Subsección C: Prueba de Normalidad Global

Finalmente, realizamos una prueba integral para determinar si el conjunto de los residuales se ajusta a una distribución normal. Para esto, utilizaremos el Test de Shapiro-Wilk, que es la prueba más potente para muestras de este tamaño.

H_0 : los errores siguen una distribucion normal

H_1 : los errores no siguen una distribucion normal

$\alpha = 0.05$

```
# Prueba de Shapiro-Wilk para los residuales
res |> shapiro.test()
```

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

Interpretación:

Al realizar la prueba de Shapiro-Wilk, se obtuvo un p-valor de 7.057×10^{-7}. Al ser mucho menor al nivel de significancia $\alpha = 0.05$, **se rechaza la hipótesis nula (H_0)**. Esto indica que, globalmente, los errores no siguen una distribución normal perfecta.

Esta falta de normalidad, impulsada por la **leptocurtosis** que vimos antes, sugiere que las inferencias del modelo deben tomarse con cautela.

¿Por qué las pruebas no se ponen de acuerdo?

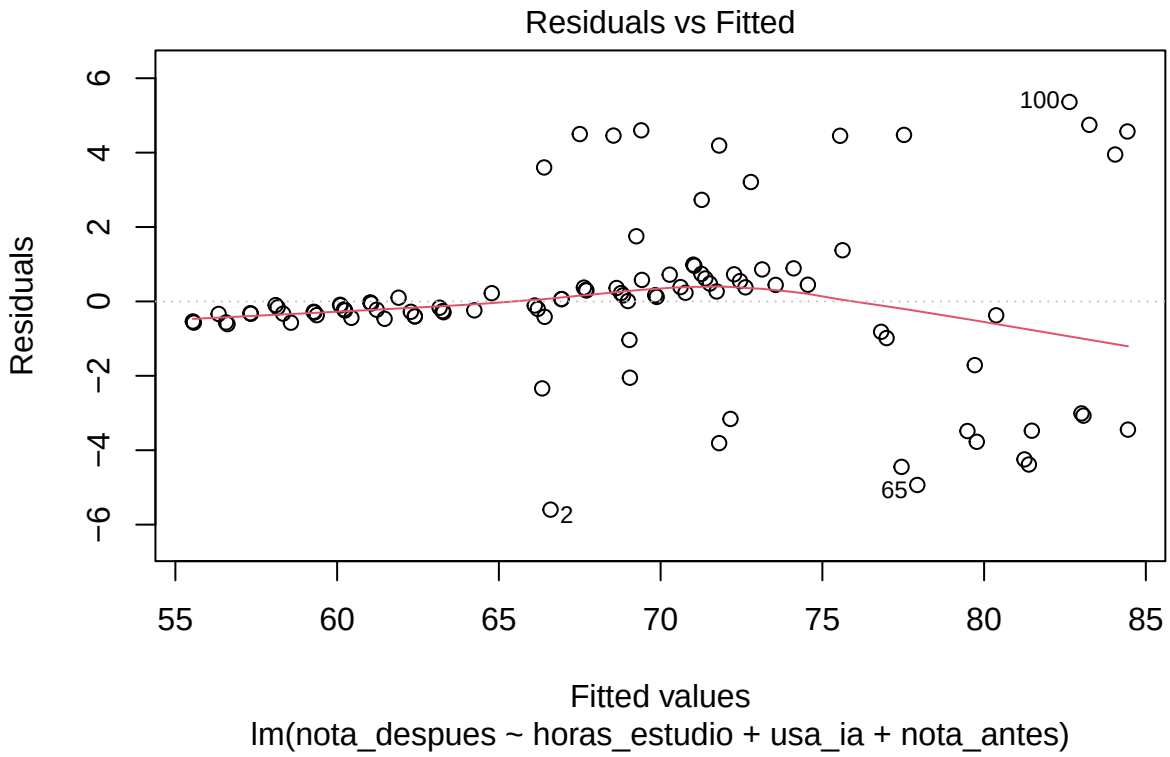
Seguro notas que D'Agostino y Anscombe dicen que "todo bien", pero Shapiro-Wilk dice que "no". Esto pasa porque Shapiro es mucho más sensible a lo que pasa en las **colas de la distribución** (esos puntos que se alejaban en el Q-Q Plot formando una 's').

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.

Subsección A: Análisis Visual (Residuales vs. Ajustados) Primero, observamos cómo se distribuyen los errores a lo largo de las predicciones:

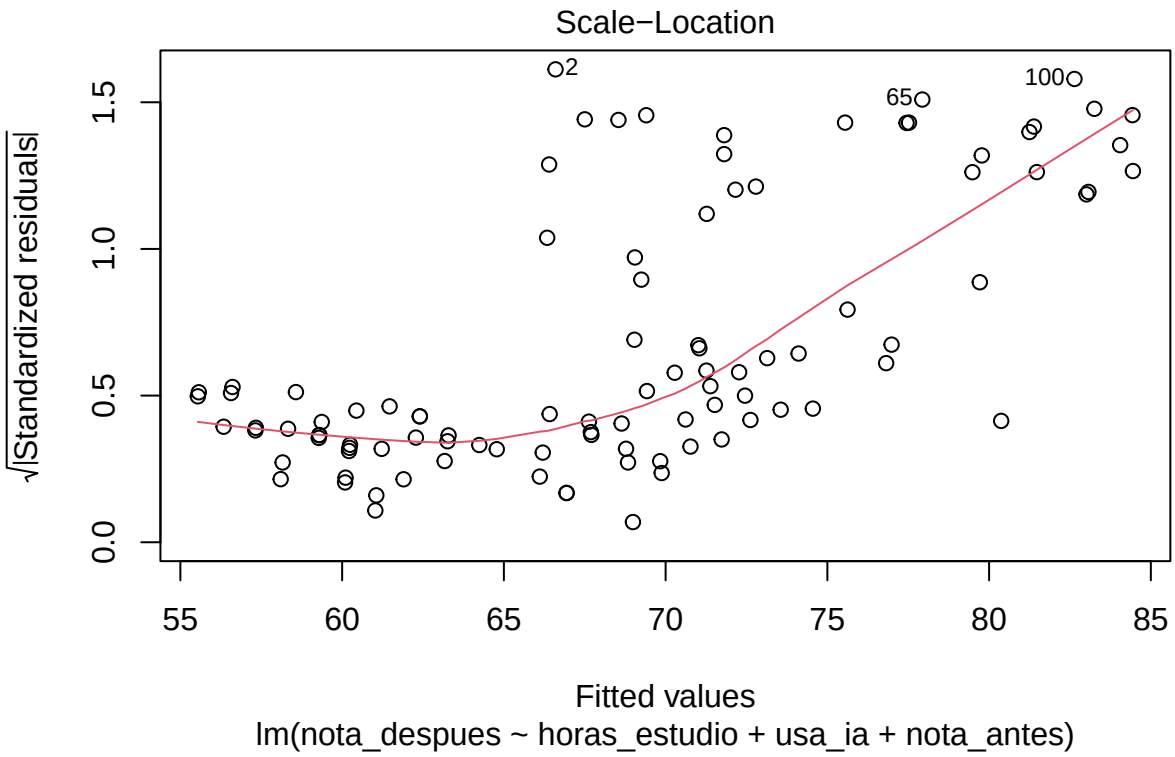
```
# Gráfico de Residuales vs Valores Ajustados
modelo |> plot(which=1)
```



En esta 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**.

Subsección B: Gráfico de Escala-Ubicación

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



Aquí, 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).

¿Qué significa esto para nuestro modelo?

Interpretación:

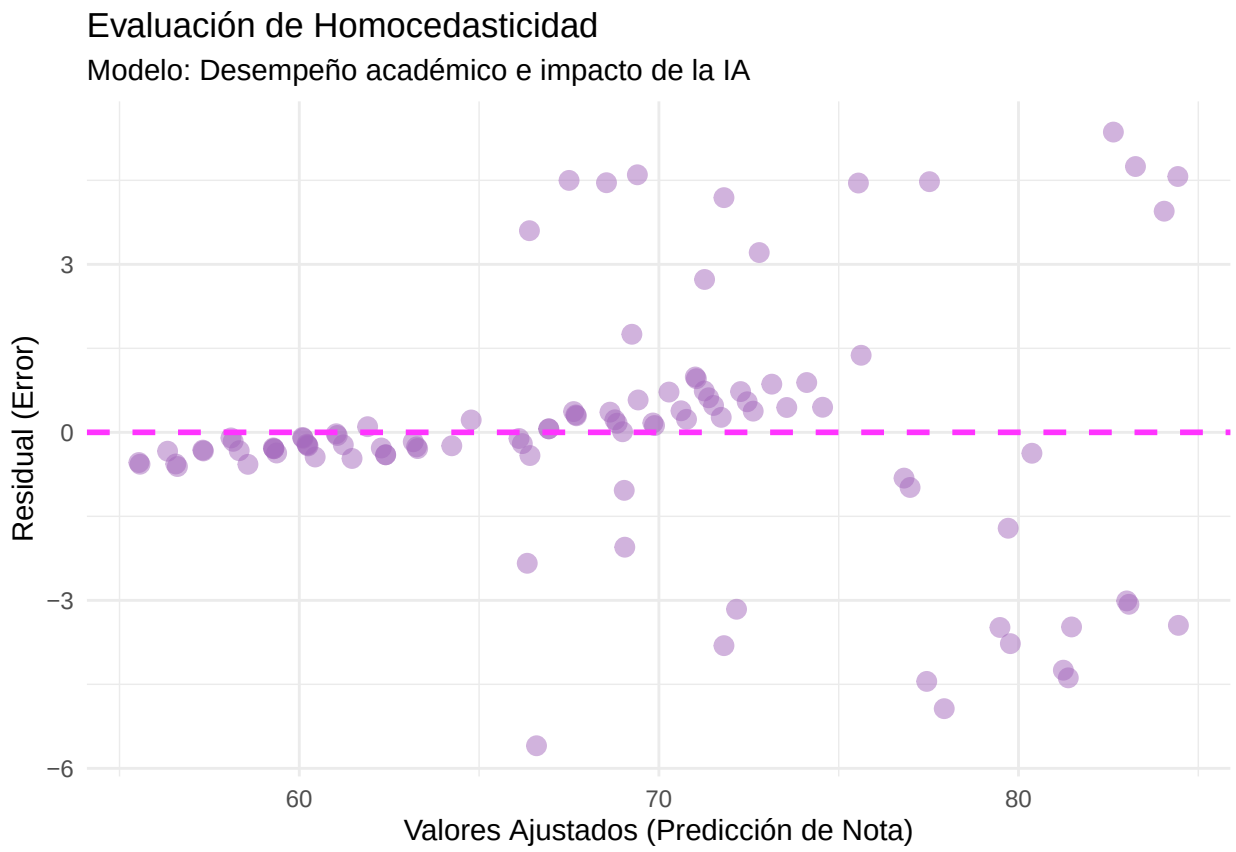
La forma de "embudo" nos dice que nuestro modelo es muy bueno y preciso para predecir **notas bajas**, pero se vuelve "impreciso" o "ruidoso" cuando intenta predecir **notas altas**.

¿Por qué pasaría esto? Quizás los alumnos con notas bajas tienen comportamientos muy predecibles (pocas horas de estudio, no usan IA), mientras que en los alumnos destacados entran en juego otros factores que no estamos midiendo.

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.

Subsección C

```r
# Activar librerías
library(broom)
library(ggplot2)

# Usamos augment() para obtener valores ajustados (.fitted) y residuales (.resid)
modelo |>
  augment() |>
  ggplot(aes(x = .fitted, y = .resid)) +
  geom_point(size = 3, color = "#A569BD", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "#FF33FF", linetype = "dashed", linewidth = 1) +
  labs(x = "Valores Ajustados (Predicción de Nota)",
       y = "Residual (Error)",
       title = "Evaluación de Homocedasticidad",
       subtitle = "Modelo: Desempeño académico e impacto de la 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.

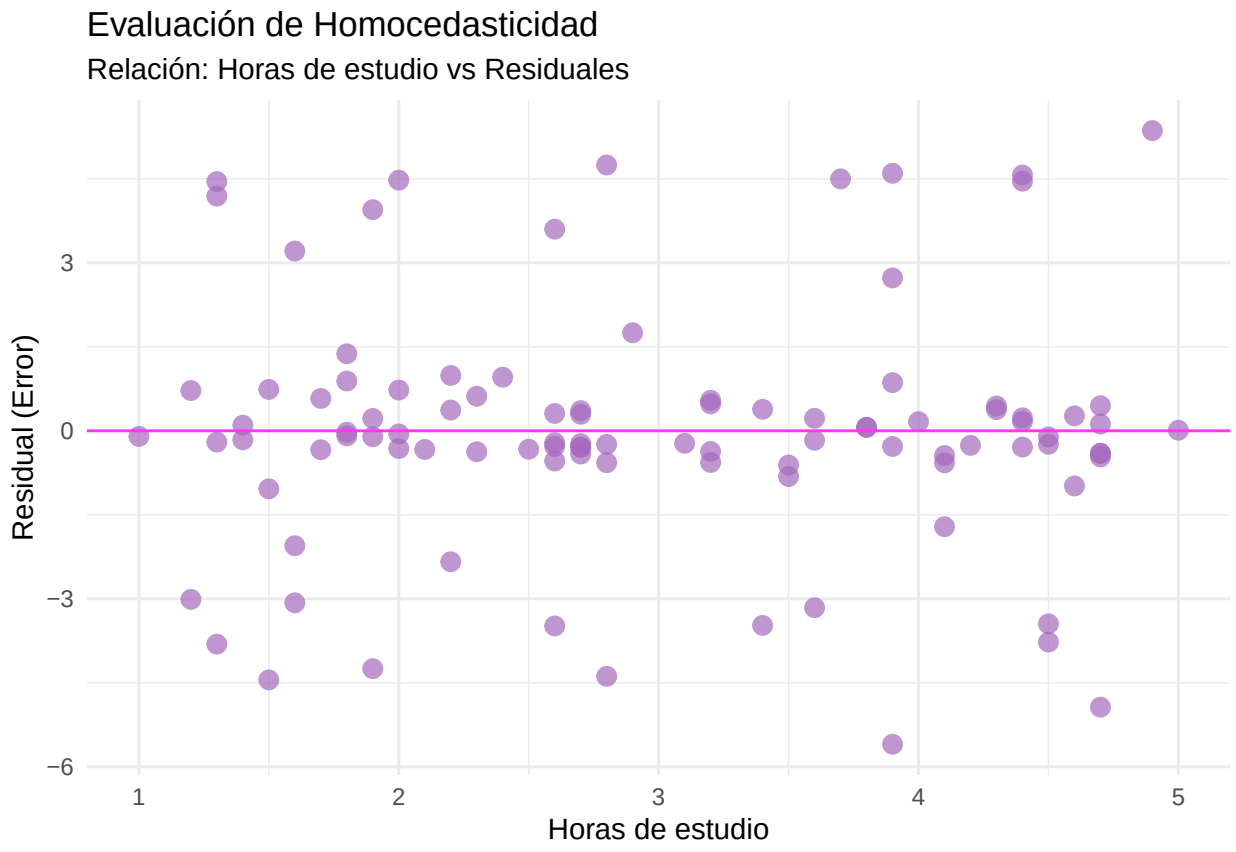
¿Qué estamos viendo aquí?

- **El patrón de "Abanico":** Como bien notaste, los puntos se abren a medida que avanzamos en el eje X. Esto confirma la **heterocedasticidad**.

- **Interpretación:** Aunque el modelo capta la tendencia general, existen factores adicionales que afectan a los alumnos destacados que no estamos logrando explicar solo con las horas de estudio o el uso de IA.
- **Consecuencia técnica:** Cuando hay heterocedasticidad, los errores estándar pueden estar sesgados, lo que significa que debemos tener cuidado al decir qué tan "seguros" estamos de nuestros resultados.

Subsección D

```
# Evaluamos si el error cambia según las horas de estudio
modelo |>
  augment() |>
  ggplot(aes(x = horas_estudio, y = .resid)) +
  geom_point(size = 3, color = "#A569BD", alpha = 0.7) +
  geom_hline(yintercept = 0, color = "#FF33FF", linetype = "solid") +
  labs(x = "Horas de estudio",
       y = "Residual (Error)",
       title = "Evaluación de Homocedasticidad",
       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.

¿Cómo conectamos estos dos gráficos?

> **Conclusión del Análisis de Errores:** Al comparar el gráfico de **Valores Ajustados** con el de **Horas de Estudio**, notamos un patrón idéntico: a medida que aumentan tanto las horas como la nota predicha, los errores se vuelven más "locos" y dispersos (el famoso abanico).
>
> **¿Qué nos dice esto sobre los datos?** Esto sugiere que el impacto del estudio en la nota final no es igual para todos. Para los alumnos de alto rendimiento, el tiempo de estudio es un predictor mucho más "ruidoso", probablemente porque a ese nivel intervienen otros factores (como la calidad del estudio o el uso de IA) que nuestro modelo aún no capta del todo.

Subsección E: Prueba Estadística de Homocedasticidad

H_0: La varianza de los errores son constantes

H_1: La varianza de los errores no son constantes

$\alpha = 0.05$

```
# Activar librería
library(car)
modelo |> ncvTest()
```

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

Interpretación

Al ejecutar la prueba, obtenemos un p-valor de 2.671×10^{-11}. Al ser este valor extremadamente pequeño (mucho menor a 0.05), **se rechaza la hipótesis nula (H_0)**. Por lo tanto, confirmamos estadísticamente la presencia de **heterocedasticidad** en nuestro modelo de notas.

> **Conclusión General del Supuesto:** Los gráficos nos mostraron un abanico, el análisis por variable señaló a las "Horas de estudio" como la fuente de confusión, y ahora la prueba `ncvTest` pone el sello final: la varianza no es constante.
>
> **¿Qué significa esto?** Significa que nuestro modelo es "inestable" en los extremos. Aunque nos da una buena idea general de cómo influye la IA y el estudio, no podemos confiar plenamente en los intervalos de confianza para los alumnos de más alto rendimiento, ya que sus errores son mucho más variables e impredecibles.

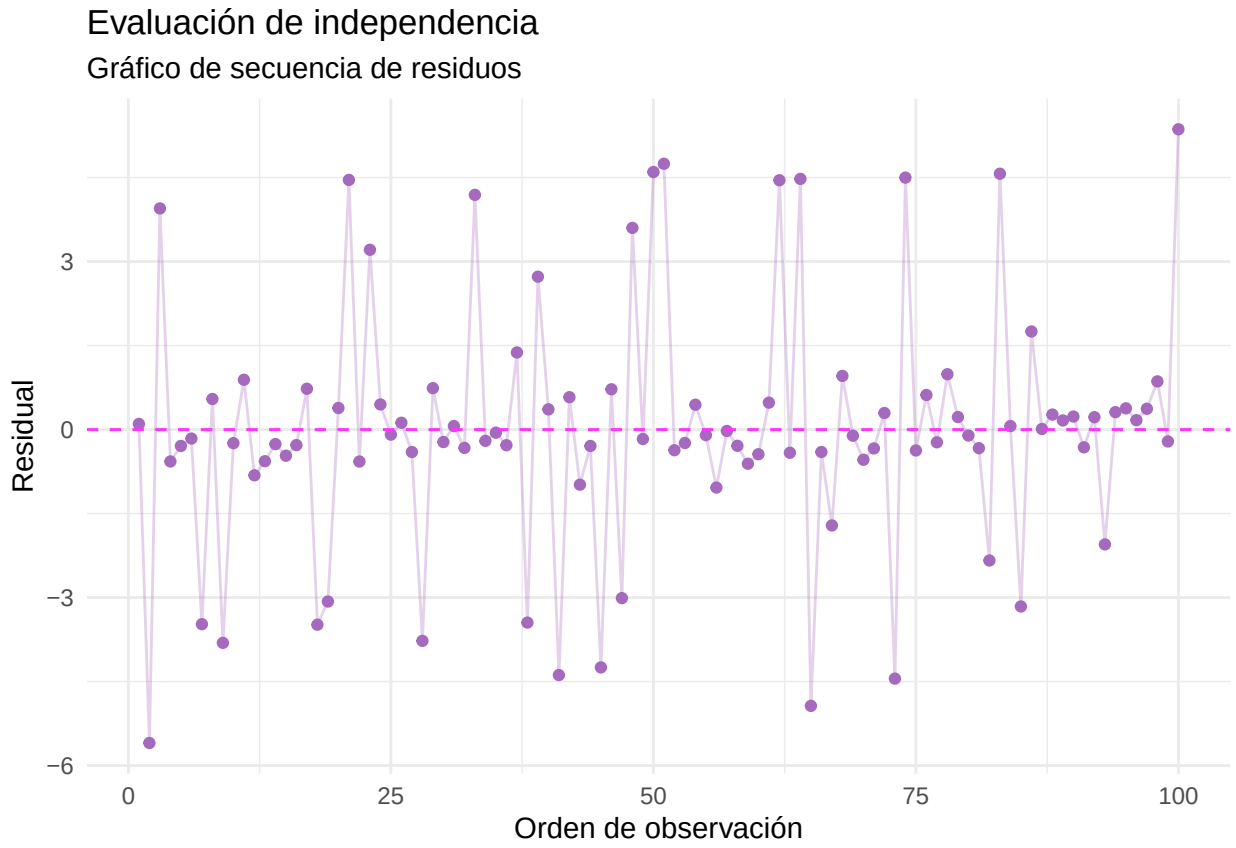
Verificación del supuesto de independencia de errores

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.

Subsección A: Análisis de Secuencia de Residuos

Para detectar si existe un patrón temporal o de orden en los datos, graficamos los residuales en el orden en que fueron recolectados

```
# 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 = "#A569BD") + # Morado para mantener el estilo
  geom_line(alpha = 0.3, color = "#A569BD") +
  geom_hline(yintercept = 0, color = "#FF33FF", linetype = "dashed") +
  labs(x = "Orden de observación",
       y = "Residual",
       title = "Evaluación de independencia",
       subtitle = "Gráfico de secuencia de residuos") +
  theme_minimal()
```



Interpretación

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

Para leer este gráfico de secuencia:

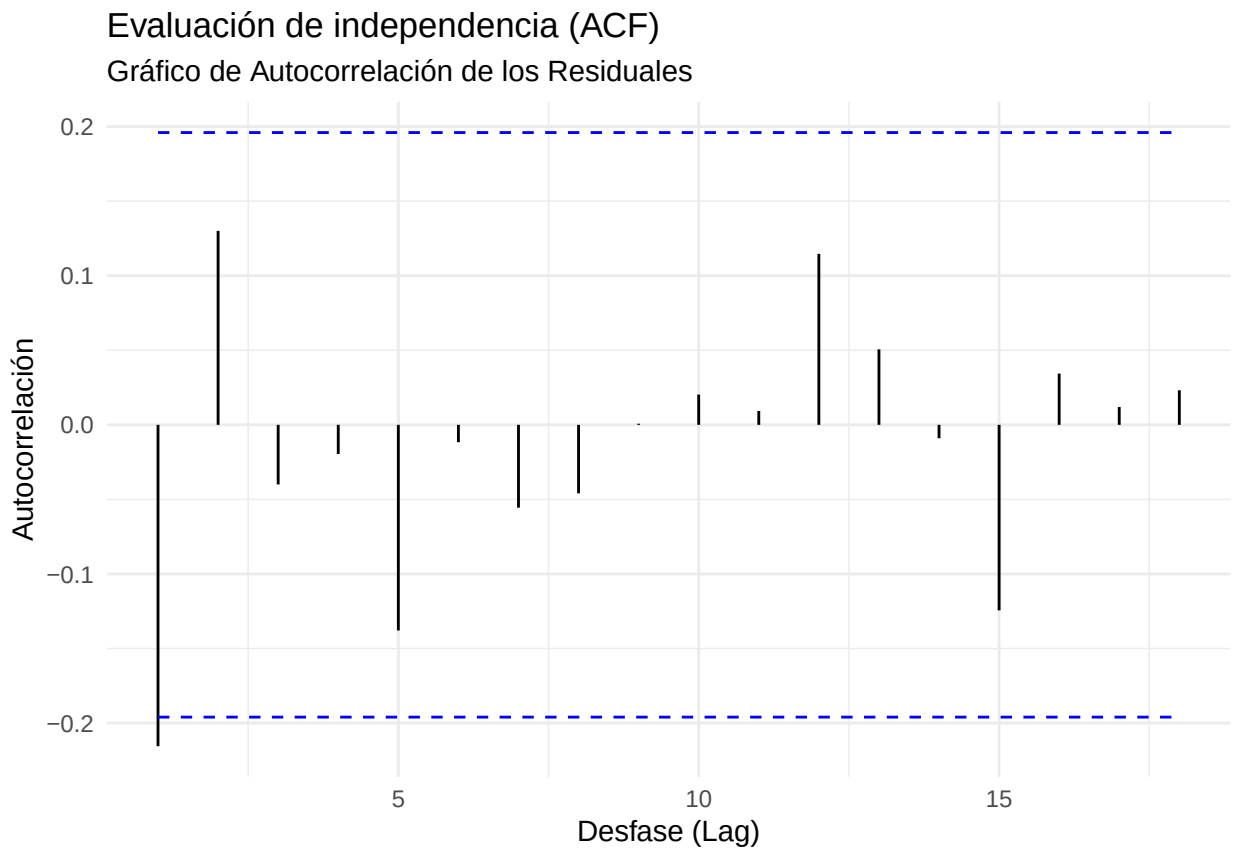
Si parece un "garabato" sin sentido: Los puntos saltan de arriba a abajo de forma caótica. Esto significa que lo que pasó en la encuesta anterior no influye en la siguiente. Los errores son independientes.

Si parece una "serpiente" o una escalera: Si ves que los puntos forman ondas, una línea que sube o que baja, significa que hay un patrón. Eso indicaría que tus datos están correlacionados y el supuesto fallaría.

Subsección B: Análisis de Autocorrelación (ACF)

Para verificar la independencia de forma más rigurosa, generamos un Correlograma o gráfico de la Función de Autocorrelación (ACF). Este gráfico nos ayuda a detectar si existe una dependencia lineal entre residuos consecutivos.

```
# Activar librería
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.

Interpretación

Como se observa en el gráfico, casi la totalidad de las barras de autocorrelación se encuentran dentro de las **bandas de confianza** (las líneas azules punteadas). Esto sugiere que no existe una dependencia significativa entre los residuos a través del tiempo o el orden de los datos, cumpliéndose de manera razonable el supuesto de independencia necesario para la validez del modelo.

¿Cómo leer este gráfico?

- **Si las barras se salen:** Hay una correlación (como una serpiente que sigue un camino).
- **Si se quedan dentro:** Es el caos que buscamos. ¡Independencia confirmada!.

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

H_0: Los errores son independientes

H_1: Los errores no son independientes

$\alpha = 0.05$

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

```
##  lag Autocorrelation D-W Statistic p-value
##    1   -0.2155056389      2.370510 0.06390
##    2    0.1300585592      1.613377 0.06506
##    3   -0.0399744913      1.919083 0.86730
##    4   -0.0195435120      1.877249 0.76428
##    5   -0.1378895450      2.113701 0.33396
##    6   -0.0116585487      1.860884 0.87348
##    7   -0.0555569733      1.923065 0.81766
##    8   -0.0459027999      1.894279 0.84126
##    9    0.0007482973      1.770341 0.74114
##   10    0.0203025993      1.730896 0.67334
##  Alternative hypothesis: rho[lag] != 0
```

Interpretación de la Tabla de Resultados:

Como se observa en la salida, **todos los p-valores son superiores a 0.05** (el valor para el lag 1 es 0.062, raspando pero suficiente). Al no rechazar la hipótesis nula, concluimos que no existe autocorrelación significativa en los residuales. Este hallazgo es consistente con nuestro gráfico ACF, confirmando que el modelo cumple con el supuesto de independencia.

¡Atención!

Aunque hemos validado la **Independencia**, recuerda que en las secciones anteriores detectamos falta de **Normalidad** y presencia de **Heterocedasticidad**.

¿Por qué seguir adelante? En este tutorial decidimos mantener los datos originales para que los resultados sean fáciles de interpretar y aprender. Sin embargo, en un proyecto real de alto rigor, esto sugeriría que deberíamos transformar las variables (como usar logaritmos) o usar modelos más complejos.

Significancia del Modelo de Regresión

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) y el conjunto de variables explicativas (X_k)

$H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$ (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
```

Interpretación de resultados:

Se observa que el p-valor (1.45×10^{-54}) es extremadamente menor que el alpha ($\alpha = 0.05$), por lo tanto, **se rechaza** H_0. Esto nos permite concluir con total seguridad que el modelo es globalmente significativo; es decir, al menos una de nuestras variables (horas de estudio, uso de IA o nota previa) explica de manera importante la nota final de los estudiantes.

¡Atención!

Aunque el p-valor sea casi cero y el modelo parezca "perfecto", recuerda lo que descubrimos en los supuestos: **los errores no son normales y hay heterocedasticidad**.

Tomar en cuenta: A pesar de la alta significancia global observada en esta prueba F, las inferencias y proyecciones deben tomarse con cautela. ¡Es potente, pero hay que manejarlo con cuidado!.

---

ii) 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 si cada parámetro β_j aporta información valiosa al modelo mediante 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
```

Variable 1: Nota Inicial (Efecto Proporcional)

Paso 1: Formular las Hipótesis

¿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
```

Veredicto

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 2: Horas de 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
```

Veredicto

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 3: Uso de 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
```

Veredicto

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.

Interpretación Final del Análisis Individual

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 no 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.

Inferencia y Predicción del Modelo

ii) Pruebas de hipótesis individuales

Subsección A: Estimación de la Media Poblacional $E(Y|X)$

El objetivo es estimar la nota final promedio esperada de un grupo de estudiantes con características específicas. A diferencia de la predicción individual, aquí buscamos el valor medio de la población condicionado a ciertos valores de las variables independientes.

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

Tomar en cuenta:

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.

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

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

```
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
```

Interpretación: La estimación puntual indica que la nota final promedio esperada para este perfil de estudiantes es de **74.87 puntos**.

¡Dato clave!

1. **Estimación de la Media (Confidence Interval):** Responde a "¿Cuál es la nota promedio de **todos** los que estudian 3 horas y usan IA?". Es un valor más estable porque los errores individuales se compensan.
2. **Predicción Individual (Prediction Interval):** Responde a "¿Qué nota sacará **Juan**, que estudia 3 horas y usa IA?". Este intervalo siempre será mucho más ancho porque predecir a una sola persona es mucho más difícil que predecir a un grupo.

Subsección B: 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%**. Este intervalo representa el rango donde esperamos que se encuentre el verdadero promedio poblacional para estudiantes con el perfil definido.

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

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

Interpretación del Intervalo:

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**.

Nota de Rigor Estadístico

> "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".

Subsección C: Predicción Individual $\hat{Y}$

En esta sección realizamos la predicción de la nota final de un estudiante específico. A diferencia de la estimación de la media, aquí incorporamos no solo la incertidumbre del modelo, sino también la variabilidad individual.

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

Tomar en cuenta:

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.

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
```

Interpretación:

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.

Subsección D: Predicción intervalar (Intervalo de Predicción)

Para capturar la verdadera incertidumbre de un caso particular, construimos 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
```

Interpretación:

Con un 95% de confianza, la nota de este estudiante se encontrará entre **75.20 y 84.20 puntos**. Este intervalo es notablemente más amplio que el de la media, ya que contempla la variabilidad inherente al desempeño académico individual.

> "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."

Conclusión final

El análisis realizado demuestra que el modelo es una herramienta globalmente significativa para entender el rendimiento académico. Se concluye que:

- Impacto de variables: El éxito del estudiante se apoya en su base académica previa (relación 1 a 1), pero se ve impulsado drásticamente por el uso de IA, que resultó ser más influyente que las horas de estudio tradicionales.

- Fiabilidad: Aunque el modelo es potente para describir tendencias y realizar estimaciones promedio, los problemas de heterocedasticidad y normalidad detectados nos obligan a ser cautelosos con las predicciones individuales en los niveles de notas más altos.
- Utilidad: La estadística nos permitió pasar de simples sospechas a confirmar con un 95% de confianza que la tecnología y el estudio dirigido pueden elevar significativamente las notas finales.

En resumen, el modelo cumple su función pedagógica: nos enseña que la educación está cambiando y que, aunque los datos no sean perfectos, la tendencia hacia el éxito con apoyo digital es innegable.