Regresión Lineal Simple

Víctor Aceña - Isaac Martín

DSLAB

2025-09-10

Fundamento del Modelado Estadístico

La regresión lineal constituye uno de los **pilares fundamentales** de la modelización estadística.

¿Por qué es tan importante?

- Es el **primer modelo predictivo** que se aprende por su simplicidad e interpretabilidad
- Los conceptos aquí desarrollados son la **base para técnicas avanzadas**: regresión múltiple, GLMs, machine learning
- Proporciona el **marco conceptual** para toda la inferencia estadística en modelos lineales

Nuestro enfoque:

Seguiremos el **ciclo completo** de un proyecto de modelado: exploración $\rightarrow$ formalización $\rightarrow$ estimación $\rightarrow$ inferencia $\rightarrow$ diagnóstico

Objetivos de Aprendizaje

1. **Comprender y aplicar** el proceso de modelización estadística para problemas con una variable predictora
2. **Identificar y medir** la correlación lineal entre dos variables como paso previo al modelado
3. **Describir la formulación matemática** del modelo de regresión lineal simple e interpretar sus parámetros
4. **Estimar los coeficientes** mediante mínimos cuadrados ordinarios (MCO) y entender sus propiedades
5. **Realizar inferencias** sobre los parámetros del modelo y evaluar su bondad de ajuste
6. **Diagnosticar la adecuación** del modelo verificando si se cumplen los supuestos

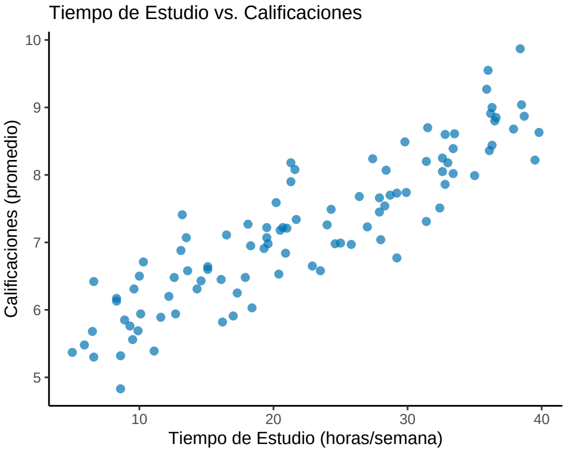
Ejemplo Motivador: Los Datos

Pregunta de investigación: ¿Influye el tiempo de estudio semanal en las calificaciones finales?

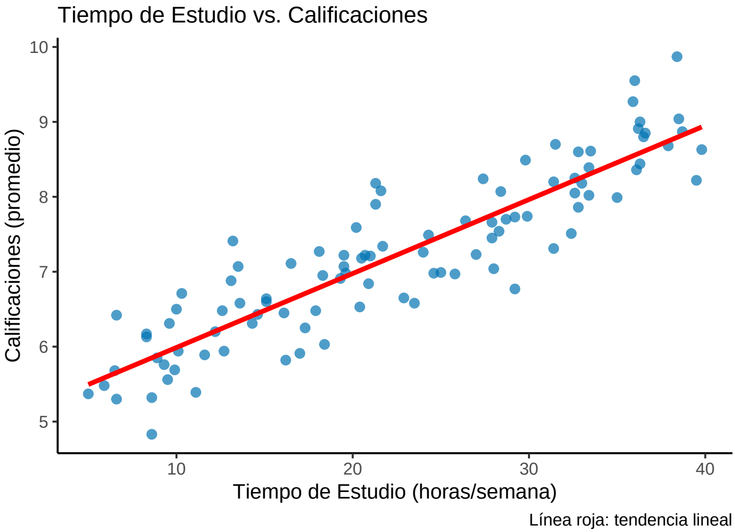
Simulación de datos realista:

- 100 estudiantes universitarios
- Tiempo de estudio: entre 5 y 40 horas/semana
- Calificaciones: escala de 0 a 10 puntos

Pregunta: ¿Qué patrón observas en los datos?



Ejemplo Motivador: Primera Observación



- **Tendencia clara:** A más tiempo de estudio, mejores calificaciones
- **Relación lineal:** Los puntos siguen aproximadamente una línea recta
- **Dirección positiva:** Pendiente ascendente

Conclusión: La tendencia lineal positiva justifica un modelo de regresión lineal.

Paso 1: Exploración Visual - El Gráfico de Dispersión

El **gráfico de dispersión** (*scatterplot*) es la herramienta más potente para examinar la relación entre dos variables continuas.

¿Qué nos permite evaluar?

Características de la relación:

- **Forma**: ¿Es lineal, curva, o sin patrón?
- **Dirección**: ¿Positiva o negativa?
- **Fuerza**: ¿Qué tan estrecha es la relación?
- **Valores atípicos**: ¿Hay observaciones extremas?

Criterios para regresión lineal:

- **Linealidad**: Los puntos siguen una tendencia recta
- **Variabilidad constante**: La dispersión es similar en todo el rango
- **Sin valores atípicos extremos**: No hay puntos que distorsionen la relación

Principio: La visualización SIEMPRE precede a la cuantificación

Paso 2: Cuantificación - Covarianza y Correlación

Una vez que la visualización sugiere una tendencia, necesitamos métricas para cuantificarla.

Covarianza muestral:

$$\text{Cov}(x, y) = s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

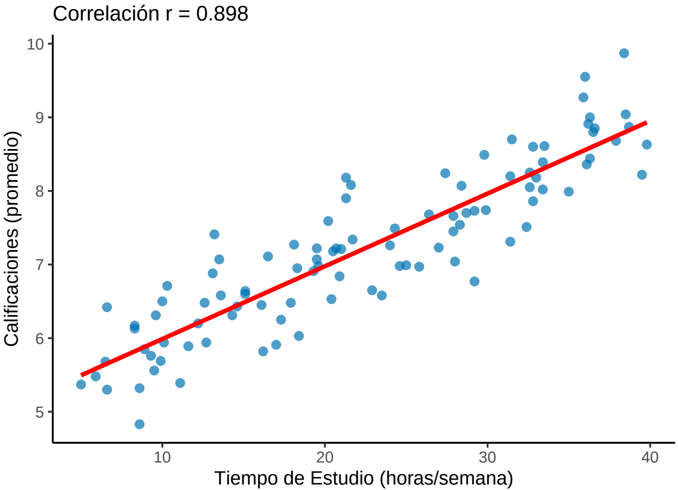
- **Problema**: Su magnitud depende de las unidades de las variables

Coeficiente de correlación de Pearson:

$$r = r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- **Ventajas**: Adimensional, siempre entre -1 y 1
- **Interpretación**: Fuerza de la asociación *lineal*

Cuantificación en Nuestro Ejemplo



Resultados del análisis:

- **Covarianza**: 9.82
  (difícil de interpretar por las unidades)
- **Correlación**: **0.898**
  (asociación lineal muy fuerte y positiva)

La línea roja muestra la **tendencia lineal** de los datos.

Correlación NO Implica Causalidad

Encontrar una **correlación fuerte** (0.898) entre tiempo de estudio y calificaciones **NO** nos autoriza a concluir que *una causa la otra*.

¿Por qué?

Posibles explicaciones alternativas:

- **Variable oculta**: El interés del estudiante influye tanto en las horas de estudio como en las calificaciones
- **Causalidad inversa**: Los estudiantes con mejores calificaciones se motivan a estudiar más
- **Terceras variables**: Calidad del sueño, técnicas de estudio, etc.

La regresión lineal puede:

Demostrar que las variables se mueven juntas
Permitirnos predecir una a partir de la otra
Cuantificar la fuerza de la asociación

NO puede:

Explicar el porqué de la relación
Establecer causalidad sin diseño experimental

El Modelo Poblacional

Una vez confirmada la relación lineal, **formalizamos** matemáticamente nuestra observación.

El modelo poblacional postula que la relación verdadera sigue una línea recta con aleatoriedad:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Componentes:

Parte sistemática:

- β_0: **Intercepto** (parámetro poblacional desconocido)
- β_1: **Pendiente** (parámetro poblacional desconocido)

Parte aleatoria:

- ε_i: **Error aleatorio** que incluye:
  - Variables omitidas
  - Error de medición
  - Aleatoriedad intrínseca

Nunca observamos la población $\rightarrow$ Usamos la muestra para estimar el **modelo muestral**

Del Modelo Poblacional al Modelo Muestral

Modelo poblacional (desconocido):

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Modelo muestral (estimado):

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Terminología clave:

- Los "**gorros**" ($\hat{\cdot}$) indican **estimaciones** calculadas de la muestra
- La diferencia $e_i = y_i - \hat{y}_i$ es el **residuo** (aproximación empírica del error ε_i)
- $\hat{y}_i$ es el **valor predicho** por el modelo

Objetivo: Usar la muestra para encontrar la "mejor" recta de ajuste

Los Supuestos del Modelo

Para que nuestras estimaciones e inferencias sean válidas, asumimos que los errores ε_i se comportan ordenadamente:

1. **Linealidad**: $E[Y_i|X_i] = \beta_0 + \beta_1 X_i$
2. **Independencia**: $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ para $i \neq j$
3. **Homocedasticidad**: $Var(\varepsilon_i|X_i) = \sigma^2$ (varianza constante)
4. **Normalidad** (para inferencia): $\varepsilon_i \sim N(0, \sigma^2)$

Importancia: Estos supuestos garantizan las **propiedades óptimas** de los estimadores de mínimos cuadrados y la **validez** de la inferencia estadística.

El Método de Mínimos Cuadrados Ordinarios (MCO)

Criterio: Encontrar la recta que **minimice** la suma de los cuadrados de los errores.

$$\text{SSE}(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

¿Por qué este criterio?

- Los errores positivos y negativos no se cancelan
- Penaliza más los errores grandes
- Tiene solución analítica única
- Proporciona estimadores con propiedades óptimas

Interpretación geométrica:

Minimizamos la suma de las **distancias verticales al cuadrado** entre los puntos observados y la recta de regresión.

Derivación de las Fórmulas de MCO

Para encontrar β_0 y β_1 que minimizan SSE, usamos cálculo:

Derivadas parciales:

$$\frac{\partial \text{SSE}}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial \text{SSE}}{\partial \beta_1} = -2\sum_{i=1}^{n}x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

Resolviendo el sistema (ecuaciones normales):

Al resolver estas ecuaciones simultáneamente, obtenemos las fórmulas de MCO.

Fórmulas Finales de MCO

Fórmula para la pendiente:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{s_{xy}}{s_{xx}} = \frac{\text{Covarianza muestral}}{\text{Varianza muestral de } X}$$

Fórmula para el intercepto:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Notación:

- $s_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ (suma de productos cruzados)
- $s_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ (suma de cuadrados de X)

Interpretación intuitiva:
La pendiente es la razón entre la covariación de X e Y y la variación de X.

Propiedades de las Predicciones

Las estimaciones MCO generan predicciones ($\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$) con **propiedades matemáticas específicas**:

1. **La recta pasa por el centro de los datos**: $(\bar{x}, \bar{y})$

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y}$$

Demostración: Sumando la ecuación de predicción para todas las observaciones:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

2. **Promedio de predicciones = Promedio observado**:

$$\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \bar{y}$$

Importancia: La recta de regresión siempre pasa por el **punto central** de los datos

Propiedades de los Residuos

Los residuos MCO ($e_i = y_i - \hat{y}_i$) tienen **propiedades fundamentales**:

3. **Suma de residuos = 0**:

$$\sum_{i=1}^{n} e_i = 0$$

4. **Residuos no correlacionados con X**:

$$\sum_{i=1}^{n} x_i e_i = 0$$

5. **Residuos no correlacionados con predicciones**:

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0$$

Implicación: Estas propiedades garantizan que MCO es **insesgado** y **óptimo**

Interpretación de los Coeficientes

Una vez estimados, los coeficientes tienen interpretación concreta y práctica:

Pendiente ($\hat{\beta}_1$):

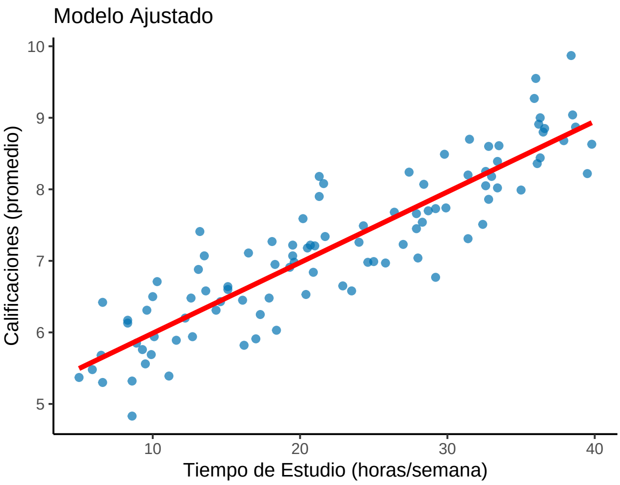
- Representa el **cambio promedio esperado** en Y por cada **aumento de una unidad** en X
- En nuestro ejemplo: puntos que aumenta la calificación por cada hora adicional de estudio

Intercepto ($\hat{\beta}_0$):

- Valor promedio esperado de Y cuando $X = 0$
- Solo tiene sentido práctico si $X = 0$ es plausible y está en el rango de los datos
- A menudo es solo un "ancla matemática" para la recta

Nota importante: La interpretación siempre debe considerar el contexto del problema y la plausibilidad de los valores.

Aplicación: Ajuste del Modelo en Nuestro Ejemplo



Ecuación del modelo:

$\text{Calificaciones} = 5.001 + 0.099 \times \text{Tiempo\_Est}$

Interpretación de coeficientes:

- **Intercepto** ($\hat{\beta}_0 = 5.001$): Calificación esperada con 0 horas de estudio
- **Pendiente** ($\hat{\beta}_1 = 0.099$): Por cada hora adicional de estudio, la calificación aumenta en promedio 0.099 puntos

Conclusión: Relación positiva y significativa entre tiempo de estudio y calificaciones.

Propiedades de los Estimadores de MCO

Bajo los supuestos de Gauss-Markov, los estimadores MCO son **MELI** (Mejores Estimadores Lineales Insesgados):

1. Insesgadez:

$$E[\hat{\beta}_0] = \beta_0 \quad \text{y} \quad E[\hat{\beta}_1] = \beta_1$$

Los estimadores MCO son **correctos en promedio** - no tienen sesgo sistemático.

2. Varianza mínima:

Entre todos los estimadores lineales insesgados, los MCO tienen la menor varianza posible.

3. Varianzas conocidas:

Las varianzas de los estimadores tienen formas matemáticas específicas y conocidas.

Varianzas de los Estimadores de MCO

Para la pendiente:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}$$

donde $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$

Interpretación:

- $Var(\hat{\beta}_1)$ disminuye con mayor dispersión en X (mayor S_{xx})
- Más datos esparcidos $\rightarrow$ mejor estimación de la pendiente

Para el intercepto:

$$Var(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]$$

Interpretación:

- $Var(\hat{\beta}_0)$ aumenta cuando $\bar{x}$ está lejos de cero
- El intercepto se estima mejor cuando $\bar{x}$ está cerca de cero
- Mayor tamaño de muestra (n) reduce la varianza

Estimación de la Varianza del Error

Las fórmulas de varianza dependen de σ^2 (desconocida). La estimamos con:

Media Cuadrática del Error (MSE):

$$\hat{\sigma}^2 = \text{MSE} = \frac{\text{SSE}}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

¿Por qué $n-2$?

- Son los **grados de libertad del error**
- Hemos "gastado" 2 grados de libertad estimando β_0 y β_1

Error estándar de los residuos:

$$\hat{\sigma} = \sqrt{\text{MSE}}$$

También llamado **RMSE** en *machine learning* $\rightarrow$ Mide la dispersión promedio alrededor de la recta

Análisis de la Varianza (ANOVA) para la Regresión

Pregunta clave: ¿Es el modelo útil o la relación observada es casualidad?

Contraste de hipótesis:

- $H_0 : \beta_1 = 0$ (no hay relación lineal)
- $H_1 : \beta_1 \neq 0$ (sí hay relación lineal)

Descomposición de la variabilidad total:

$$\text{SST} = \text{SSR} + \text{SSE}$$

Esta ecuación es fundamental: Toda la variabilidad se divide en **explicada** y **no explicada**

Suma Total de Cuadrados (SST)

Definición:

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

¿Qué mide?

- La **variabilidad total** de Y respecto a su media $\bar{y}$
- Es la varianza muestral de Y multiplicada por $(n-1)$
- Representa **toda la dispersión** que queremos explicar con nuestro modelo

Interpretación:

Si no tuviéramos ningún modelo y solo usáramos $\bar{y}$ para predecir, SST sería el **error total** que cometeríamos.

Suma de Cuadrados de la Regresión (SSR)

Definición:

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

¿Qué mide?

- La variabilidad que **explica** nuestro modelo de regresión
- Es la variabilidad de las predicciones $\hat{y}_i$ respecto a la media $\bar{y}$
- Representa la **señal** que nuestro modelo logra captar

Interpretación:

Mide cuánto **mejor** es nuestro modelo comparado con simplemente usar $\bar{y}$ como predicción.

Suma de Cuadrados del Error (SSE)

Definición:

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

¿Qué mide?

- La variabilidad **no explicada** por nuestro modelo
- Es la suma de los cuadrados de los residuos
- Representa el **ruido** que nuestro modelo no puede captar

Interpretación:

Es exactamente lo que **minimiza el método MCO** para encontrar la mejor recta.

La Descomposición Fundamental de la Variabilidad

La ecuación clave:

$$\text{SST} = \text{SSR} + \text{SSE}$$

Interpretación intuitiva:

En palabras:

- **SST**: "¿Cuánta variabilidad hay que explicar?"
- **SSR**: "¿Cuánta variabilidad explica mi modelo?"
- **SSE**: "¿Cuánta variabilidad queda sin explicar?"

En porcentajes:

- **SST**: 100% de la variabilidad
- **SSR**: % explicado por el modelo
- **SSE**: % no explicado (error)

Consecuencia: Si SSR es grande comparado con SSE $\rightarrow$ El modelo es útil

El Estadístico F

Estadístico F:

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/1}{\text{SSE}/(n-2)}$$

Interpretación:

- **MSR**: Variabilidad explicada por grado de libertad
- **MSE**: Variabilidad no explicada por grado de libertad
- **F**: Ratio entre variabilidad explicada vs no explicada

La Lógica del Contraste F (I): Si No Hay Relación

Si $H_0 : \beta_1 = 0$ fuera cierta (no hay relación lineal):

- El modelo lineal sería **inútil** para explicar Y
- Todas las predicciones $\hat{y}_i$ serían iguales a $\bar{y}$
- Por tanto: $\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \approx 0$ (muy pequeña)

Consecuencia matemática:

$$F = \frac{\text{SSR}/1}{\text{SSE}/(n-2)} \approx \frac{0}{\text{MSE}} \approx 0$$

En palabras: Si no hay relación, F debería ser cercano a **cero**

La Lógica del Contraste F (II): Si Hay Relación

Si $H_1 : \beta_1 \neq 0$ fuera cierta (sí hay relación lineal):

- El modelo **captura** la relación entre X e Y
- Las predicciones $\hat{y}_i$ varían siguiendo el patrón de los datos
- Por tanto: SSR sería **grande** (el modelo explica mucha variabilidad)

Consecuencia matemática:

$$F = \frac{\text{SSR grande}}{\text{MSE}} >> 1$$

Decisión estadística:

- **F $\approx$ 0** $\rightarrow$ No rechazamos H_0 $\rightarrow$ El modelo no es útil
- **F » 1** $\rightarrow$ Rechazamos H_0 $\rightarrow$ El modelo **sí es útil**

Tabla ANOVA: Resumen de la Descomposición

Fuente	df	SS	$MS = SS/df$	Estadístico F
Regresión	1	SSR	MSR	$F = MSR/MSE$
Error	$n-2$	SSE	MSE	
Total	$n-1$	SST		

¿Cómo leer esta tabla?

- **Fila "Regresión"**: Cuantifica lo que el modelo **explica**
- **Fila "Error"**: Cuantifica lo que el modelo **no explica**
- **Fila "Total"**: La variabilidad total que queremos explicar

El estadístico F resume todo:

$$F = \frac{\text{Variabilidad explicada por df}}{\text{Variabilidad no explicada por df}} = \frac{MSR}{MSE}$$

En regresión simple, $F = t^2$ donde t es el estadístico para contrastar $\beta_1 = 0$

Bondad del Ajuste: Coeficiente de Determinación (R^2)

El R^2 cuantifica **qué proporción** de la variabilidad total es explicada por el modelo:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

Interpretación:

- $R^2 = 0$: El modelo no explica nada (tan malo como usar $\bar{y}$)
- $R^2 = 1$: El modelo explica toda la variabilidad (ajuste perfecto)
- $R^2 = 0.7$: El modelo explica el 70% de la variabilidad

En regresión simple: $R^2 = r^2$ (cuadrado de la correlación)

Precaución: Un R^2 alto no garantiza un buen modelo ni implica causalidad

Inferencia sobre los Coeficientes

Para realizar inferencias necesitamos el supuesto de **normalidad** de los errores.

Distribución de los estimadores:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right) \qquad \hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]\right)$$

Estadístico t para la pendiente:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\mathsf{SE}(\hat{\beta}_1)} \sim t_{n-2}$$

donde $\mathsf{SE}(\hat{\beta}_1) = \sqrt{\frac{\mathsf{MSE}}{S_{xx}}}$

Contraste para la pendiente:

- $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$
- Estadístico: $t_0 = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)}$
- Decisión: Rechazar H_0 si $|t_0| > t_{\alpha/2,n-2}$

Intervalo de confianza al $(1-\alpha)100\%$ para β_1:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot \text{SE}(\hat{\beta}_1)$$

Interpretación: Si el IC no contiene el cero $\rightarrow$ β_1 es significativo

Resultados del Modelo en Nuestro Ejemplo

```
Call:
lm(formula = Calificaciones ~ Tiempo_Estudio, data = datos)

Residuals:
     Min       1Q   Median       3Q      Max
-1.11465 -0.30262 -0.00942  0.29509  1.10533

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.00118    0.11977   41.76   <2e-16 ***
Tiempo_Estudio  0.09875    0.00488   20.23   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4842 on 98 degrees of freedom
Multiple R-squared:  0.8069,    Adjusted R-squared:  0.8049
F-statistic: 409.5 on 1 and 98 DF,  p-value: < 2.2e-16
```

Interpretación de los Resultados

Coeficientes:

- **Intercepto:** 5.001 $\rightarrow$ Calificación esperada cuando el tiempo de estudio es 0 horas
- **Pendiente:** 0.0987 $\rightarrow$ Por cada hora adicional de estudio, la calificación aumenta en promedio 0.0987 puntos

Bondad de ajuste: R-cuadrado: 0.8069 $\rightarrow$ El modelo explica el **80.7%** de la variabilidad en las calificaciones

Significancia:

- **Coeficientes:** Ambos son altamente significativos ($p < 2e\text{-}16$)
- **Modelo global:** $F = 409.5$ con $p < 2.2e\text{-}16$ $\rightarrow$ El modelo es **estadísticamente útil**

Error estándar residual: 0.484 $\rightarrow$ Dispersión típica alrededor de la recta de regresión

Predicción con el Modelo

Una vez validado, usamos el modelo para **hacer predicciones**. Hay dos tipos:

1. **Intervalo de confianza para la respuesta media:**
   - Pregunta: ¿Cuál es la calificación *promedio* esperada para todos los estudiantes que estudian x_0 horas?
   - Estima dónde se encuentra la **línea de regresión verdadera**
2. **Intervalo de predicción para una respuesta individual:**
   - Pregunta: ¿Entre qué valores esperamos la calificación de *un estudiante específico* que estudia x_0 horas?
   - Considera tanto la incertidumbre del modelo como la variabilidad individual

Diferencia clave: El intervalo de predicción siempre es **más ancho** porque incluye la variabilidad σ^2 del error individual.

Fórmulas para Predicción

Intervalo de confianza para la respuesta media:

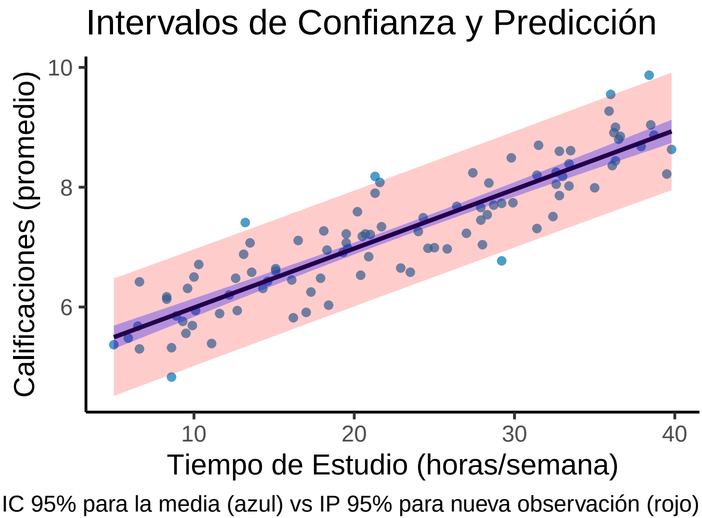
$$\hat{y}_0 \pm t_{\alpha/2,n-2} \cdot \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Intervalo de predicción para respuesta individual:

$$\hat{y}_0 \pm t_{\alpha/2,n-2} \cdot \sqrt{\text{MSE}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

Observaciones importantes:

- Ambos intervalos son más estrechos cerca del **centro** de los datos ($\bar{x}$)
- La diferencia entre ambos es el término **"+1"** que representa σ^2
- Nunca extrapolar más allá del rango de los datos observados



Interpretación gráfica:

- **Banda azul (IC)**: Incertidumbre sobre la **media poblacional**
- **Banda roja (IP)**: Incertidumbre para **nueva observación**
- **Diferencia clave**: IP incluye variabilidad individual (σ^2)
- **Patrón**: Ambos intervalos se estrechan cerca de $\bar{x}$
- **Aplicación**: Usar IC para estimar tendencia, IP para predicciones individuales

El Diagnóstico del Modelo: Fundamento Teórico

¿Por qué es crucial el diagnóstico?

El diagnóstico **NO es opcional**. Las inferencias estadísticas (p-valores, intervalos de confianza, predicciones) solo son válidas si se cumplen los supuestos del modelo.

Consecuencias de ignorar el diagnóstico:

- **Estimadores sesgados** $\rightarrow$ Conclusiones erróneas
- **Errores estándar incorrectos** $\rightarrow$ Intervalos de confianza y p-valores inválidos
- **Predicciones poco fiables** $\rightarrow$ Pérdida de poder predictivo

Filosofía del diagnóstico: Los residuos son la "ventana" hacia los errores verdaderos ε_i

Los Cuatro Pilares del Diagnóstico

Recordatorio de supuestos:

1. **Linealidad**: $E[Y_i|X_i] = \beta_0 + \beta_1 X_i$
2. **Independencia**: Cov$(\varepsilon_i, \varepsilon_j) = 0$ para $i \neq j$
3. **Homocedasticidad**: $Var(\varepsilon_i|X_i) = \sigma^2$ (varianza constante)
4. **Normalidad**: $\varepsilon_i \sim N(0, \sigma^2)$ (para inferencia)

Herramienta fundamental: Análisis de **residuos** $(e_i = y_i - \hat{y}_i)$

Principio clave: Si los supuestos se cumplen, los residuos deben comportarse como **ruido aleatorio** sin patrones sistemáticos

Diagnóstico: Linealidad

Supuesto: $E[Y|X] = \beta_0 + \beta_1 X$ (relación promedio es lineal)

Métodos de Diagnóstico:

- **Gráfico:** Residuos vs Valores Ajustados
- **Test estadístico:** Test de Ramsey RESET (Regression Equation Specification Error Test)

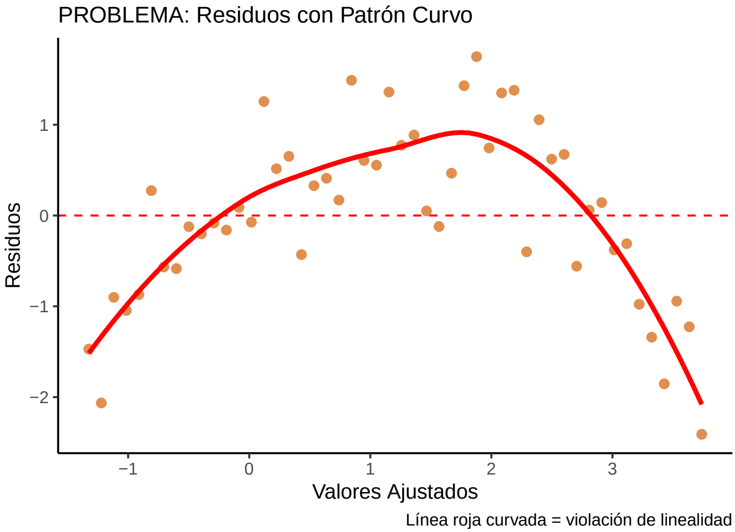
¿Qué buscamos?

- **Patrón ideal:** Nube aleatoria de puntos centrada en cero
- **Violación:** Patrón curvilíneo (forma de "U" o parábola)

Test de Ramsey RESET:

- H_0: La forma funcional es correcta (lineal)
- H_1: La forma funcional es incorrecta (no lineal)
- Añade términos $\hat{y}^2, \hat{y}^3, \ldots$ al modelo y testa su significancia

Violación del supuesto: NO Linealidad



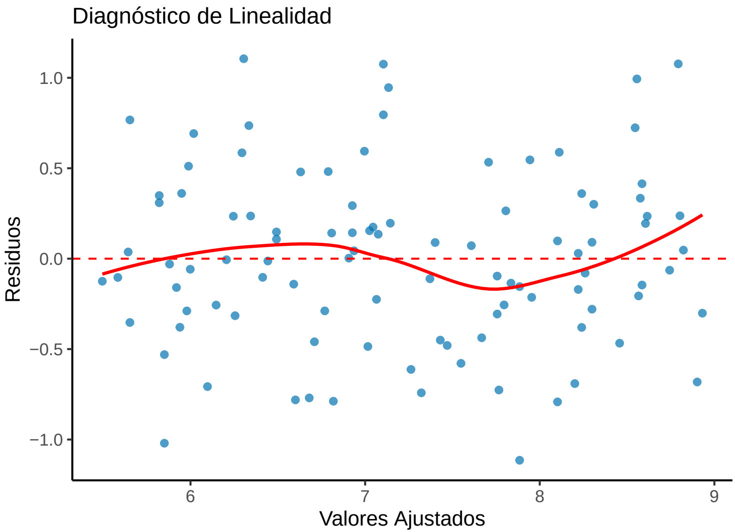
Problema detectado:
Ajustar un modelo **lineal** a datos con relación **cuadrática**

¿Qué observamos?

- **Patrón curvo** en los residuos
- Los residuos no están aleatoriamente dispersos
- La línea suavizada (roja) no es horizontal

Diagnóstico:
Patrón sistemático en residuos → **Violación de linealidad**

Verificación de linealidad



¿Qué evaluar?
Si la relación es verdaderamente lineal

Resultados:

- **Gráfico:** Línea roja prácticamente plana → Linealidad
- **Test RESET:** F = 1.051, p = 0.353 → Forma funcional correcta

Diagnóstico: Homocedasticidad

Supuesto: $Var(\varepsilon_i|X_i) = \sigma^2$ (varianza constante)

Métodos de Diagnóstico:

- **Gráficos:** Scale-Location, Residuos vs Valores Ajustados
- **Tests estadísticos:** Test de Breusch-Pagan, Test de Goldfeld-Quandt, Test de White

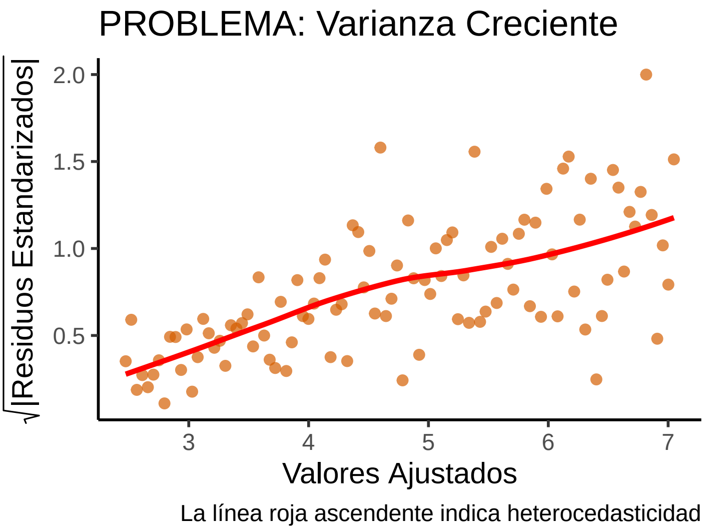
¿Qué buscamos?

- **Patrón ideal:** Dispersión constante a lo largo del rango
- **Violación:** Forma de "embudo" (dispersión creciente o decreciente)

Tests de Heterocedasticidad:

- **Breusch-Pagan:** H_0: Homocedasticidad, H_1: Heterocedasticidad
- **White:** Versión robusta que no asume forma específica de heterocedasticidad

Violación del Supuesto: NO Homocedasticidad



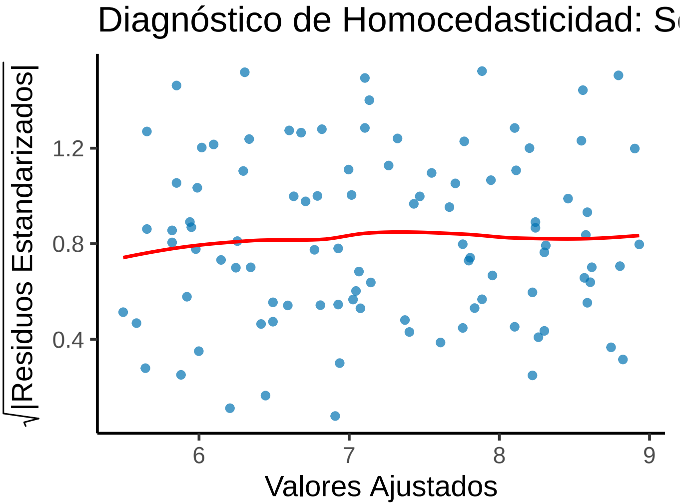
Problema: Varianza de los errores que aumenta con los valores predichos (heterocedasticidad)

¿Qué observamos?

- **Línea roja ascendente**: Claro patrón de aumento de varianza
- **Dispersión creciente**: Los residuos se dispersan más para valores altos
- **Violación**: Supuesto de varianza constante **NO** se cumple

Diagnóstico: Tendencia creciente → **Heterocedasticidad**

Verficación de homocedasticidad



Tests estadísticos:

Breusch-Pagan:

- LM = 0.02
- p = 0.889
- **Conclusión**: Homocedasticidad

White:

- LM = 0.122
- p = 0.941
- **Conclusión**: Varianza constante

Interpretación gráfica:

- Línea roja horizontal → Varianza constante
- Sin patrones sistemáticos → Supuesto cumplido

Diagnóstico: Normalidad

Supuesto: $\varepsilon_i \sim N(0, \sigma^2)$ (errores normalmente distribuidos)

Métodos de Diagnóstico:

Gráficos:

- Normal Q-Q Plot
- Histograma de residuos

Tests estadísticos:

- Test de Shapiro-Wilk
- Test de Jarque-Bera
- Test de Anderson-Darling

¿Qué buscamos en Q-Q Plot?

- **Ideal:** Puntos sobre la línea diagonal
- **Problema:** Desviaciones sistemáticas

Detalles de los Tests:

Shapiro-Wilk:

- H_0: Los residuos siguen distribución normal
- Más potente para muestras pequeñas (n < 50)

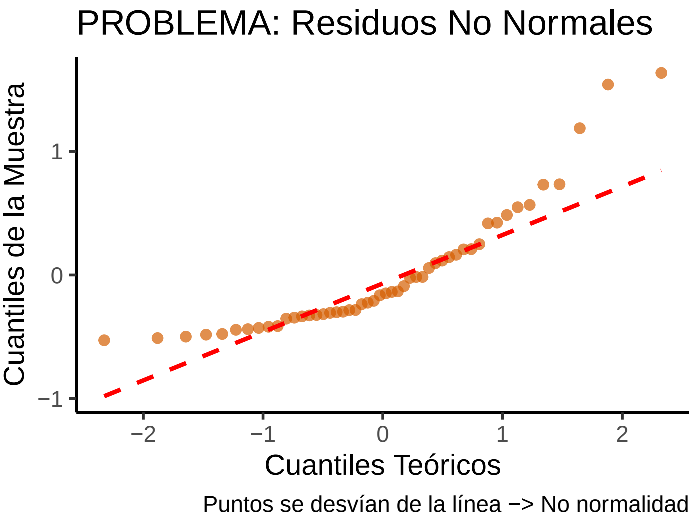
Jarque-Bera:

- Basado en asimetría y curtosis
- Asintóticamente válido para muestras grandes

Anderson-Darling:

- Más sensible en las colas de la distribución
- Mejor detección de desviaciones extremas

Violación del Supuesto: NO Normalidad



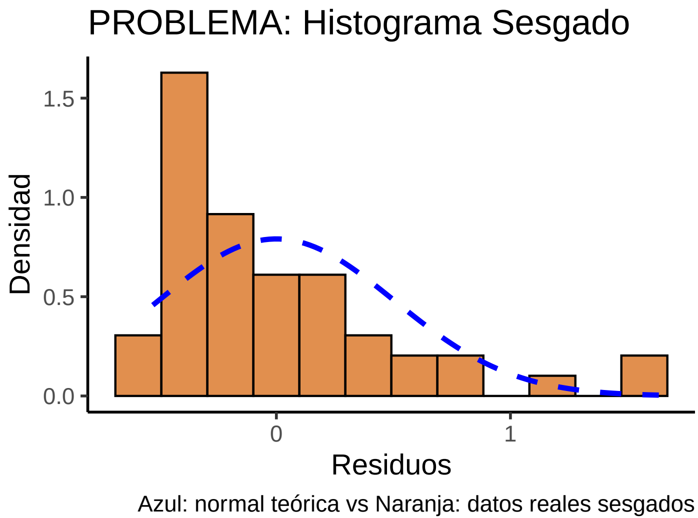
Problema: Errores con distribución asimétrica o con colas pesadas

¿Qué observamos?

- **Desviación sistemática**: Los puntos no siguen la línea diagonal
- **Curvatura**: Patrón curvo indica distribución sesgada
- **Violación**: Supuesto de normalidad **NO** se cumple

Diagnóstico: Puntos se alejan sistemáticamente de la línea → **NO normalidad**

Violación del Supuesto: Histograma NO Normal



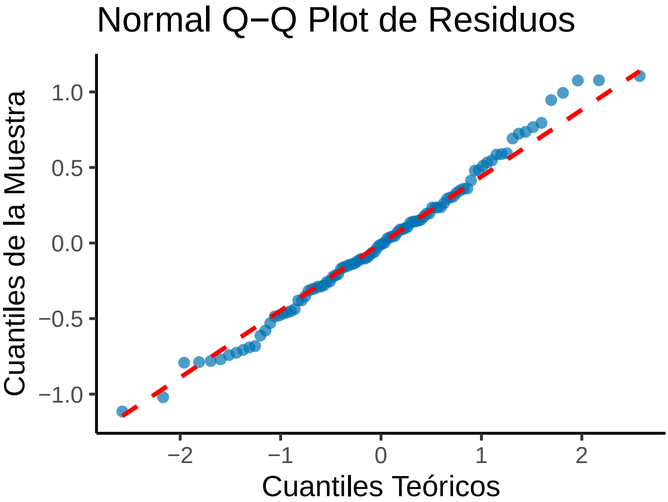
Problema: Distribución asimétrica de los residuos (histograma sesgado)

¿Qué observamos?

- **Asimetría**: Distribución sesgada hacia la derecha
- **Desajuste**: Curva normal (azul) no coincide con histograma
- **Test Shapiro-Wilk**: $p = 0 < 0.05$

Diagnóstico: Distribución sesgada curva normal $\rightarrow$ **NO normalidad**

Verificación de normalidad



Tests estadísticos:

Shapiro-Wilk:

- W = 0.99
- p = 0.671
- **Conclusión**: Normalidad

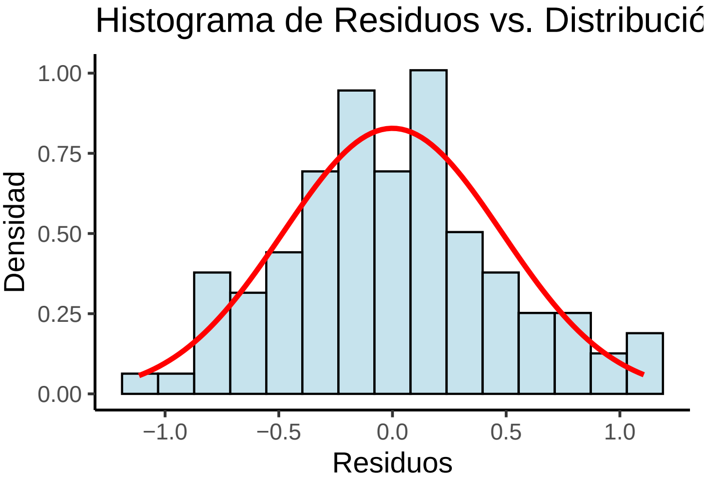
Jarque-Bera:

- JB = 0.685
- p = 0.71
- **Conclusión**: Normalidad

Interpretación gráfica:

- Puntos siguen la línea diagonal → Normalidad cumplida

Diagnóstico: Normalidad (Histograma)



Línea roja: distribución normal teórica

Complemento visual: Histograma de residuos con curva normal superpuesta

¿Qué buscamos?

- **Patrón ideal:** Distribución simétrica y campaniforme
- **Violación:** Asimetría marcada o múltiples modas

Resultado: Distribución simétrica y campaniforme → Normalidad confirmada

Diagnóstico: Independencia

Supuesto: $Cov(\varepsilon_i, \varepsilon_j) = 0$ para $i \neq j$ (errores independientes)

Métodos de Diagnóstico:

- **Gráfico:** Residuos vs Orden de observación
- **Tests estadísticos:** Test de Durbin-Watson, Test de Breusch-Godfrey (LM), Ljung-Box

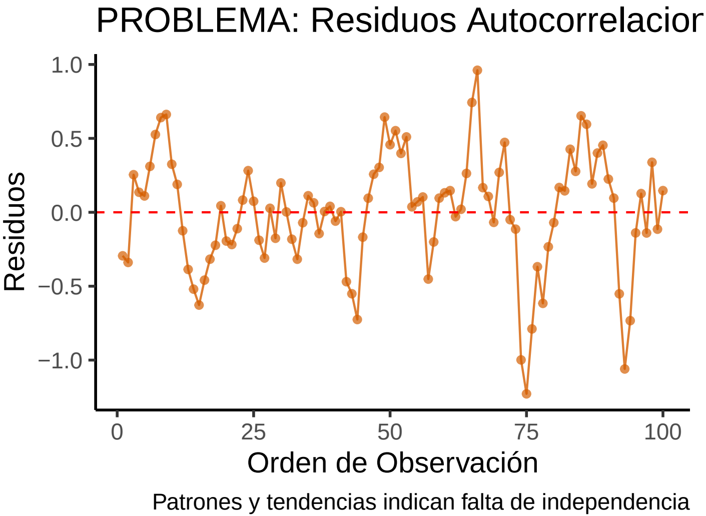
¿Qué buscamos?

- **Patrón ideal:** Residuos sin patrones temporales o secuenciales
- **Violación:** Tendencias, ciclos, o correlaciones entre residuos consecutivos

Tests de Autocorrelación:

- **Durbin-Watson:** H_0: No hay autocorrelación de primer orden ($\rho = 0$)
- **Breusch-Godfrey:** Generaliza DW para órdenes superiores y regresores retardados
- **Ljung-Box:** Testa autocorrelación conjunta en múltiples retardos

Violación del Supuesto: NO Independencia



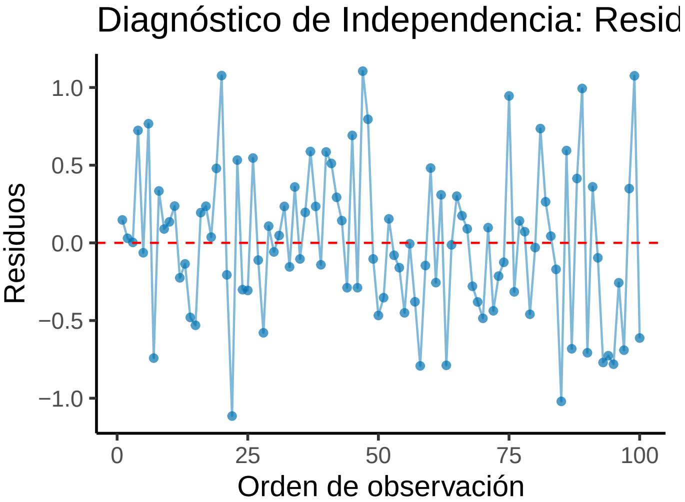
Problema: Residuos con autocorrelación (típico en series temporales)

¿Qué observamos?

- **Patrones sistemáticos**: Los residuos muestran tendencias claras
- **Conexiones**: Residuos consecutivos están correlacionados
- **Violación**: Supuesto de independencia **NO** se cumple

Diagnóstico: Patrones sistemáticos y tendencias → **NO independencia**

Verificación de Independencia



Tests estadísticos:

Durbin-Watson:

- DW = 2.056
- p = 0.61
- **Conclusión**: Sin autocorrelación orden 1

Breusch-Godfrey:

- LM = 0.14
- p = 0.932
- **Conclusión**: Sin autocorrelación orden 2

Interpretación gráfica:

- Sin patrones temporales → Independencia cumplida

Diagnóstico 5: Observaciones Influyentes

Objetivo: Identificar puntos que tienen influencia desproporcionada en el modelo

Métricas Principales:

- **Leverage** (h_{ii}): Distancia en el espacio X (valores atípicos en X)
- **Residuos Estudentizados**: Outliers en Y ajustado por su varianza
- **Distancia de Cook** (D_i): Influencia global en los coeficientes

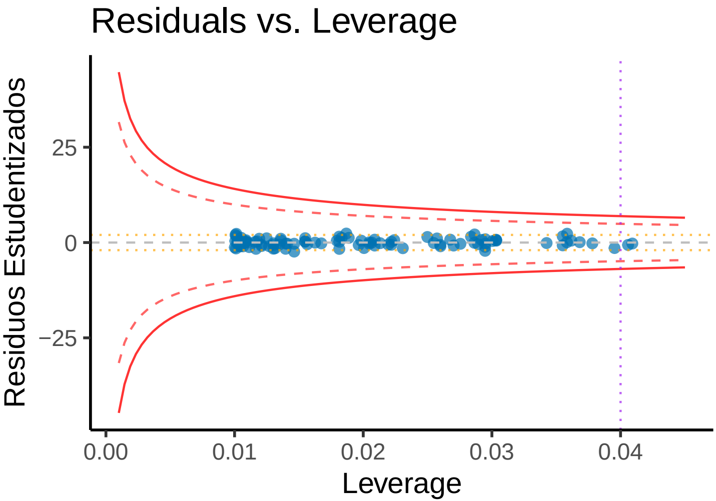
Umbrales de Referencia:

- **Leverage:** $h_{ii} > \frac{2(k+1)}{n}$ (k = número de predictores)
- **Cook:** $D_i > \frac{4}{n-k-1}$ (regla conservadora)
- **Residuos:** $|t_i| > 2$ (fuera de 2 desviaciones estándar)

Combinaciones Problemáticas:

- Alto leverage + alto residuo = **Muy influyente**
- Alto leverage + bajo residuo = **Punto de anclaje** (puede ser bueno)
- Bajo leverage + alto residuo = **Outlier sin influencia**

Diagnóstico: Observaciones Influyentes



Curvas rojas: Cook 0.5 (discontinua) y 1.0 (continua)

Resultados:

- **Leverage máximo:** 0.041 (umbral: 0.04)
- **Cook máximo:** 0.095 (umbral: 0.041)
- **Outliers ($|t| > 2$):** 6 observaciones
- **Conclusión:** Revisar observaciones: 24, 74, 6, 20, 47, 85, 87, 89

Interpretación por zonas:

- **Derecha:** Alto leverage → Potencial influencia
- **Arriba/Abajo:** Outliers → Residuos grandes
- **Esquinas:** ¡Vacías!

Análisis de Puntos Influyentes

Identificación:

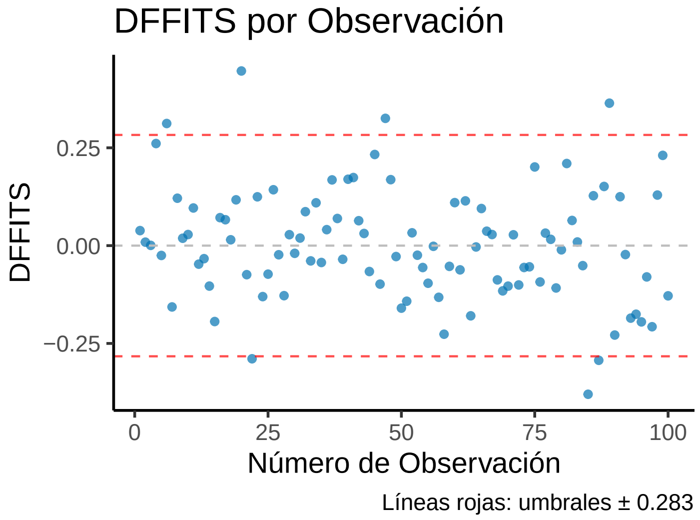
- **Outliers:** observaciones 20, 22, 47, 85, 89, 99
- **Alto leverage:** observaciones 24, 74

Interpretación por regiones:

- **Zona derecha:** Alto leverage (X atípicos) → Potencial influyente
- **Zona izquierda:** Outliers (Y atípicos) → Residuos grandes
- **Esquinas críticas:** ¡Vacías! (Situación favorable)
- **Distancia de Cook:** Influencia moderada (< 1.0)

Conclusión: No hay solapamiento leverage + outlier → Situación manejable

Diagnóstico DFFITS



DFFITS: Evalúa cómo cada observación afecta a su propia predicción

Resultados cuantitativos:

- **Umbral:** 0.283
- **Influyentes:** 7 observaciones (6, 20, 22, 47, 85, 87, 89)
- **Top 5 |DFFITS|:** obs. 20, 85, 89, 47, 6
- **Valores:** 0.446, -0.38, 0.364, 0.325, 0.312 **Interpretación:**
- **Obs. 20:** DFFITS = 0.446 (más influyente)
- **Conclusión:** 7 observaciones cambían significativamente sus propias predicciones

Análisis de Resultados DFFITS

DFFITS: Evalúa cómo cada observación afecta a su propia predicción

Resultados cuantitativos:

- **Umbral de influencia:** 0.283
- **Observaciones influyentes:** 7 observaciones (6, 20, 22, 47, 85, 87, 89)
- **Top 5 |DFFITS|:** observaciones 20, 85, 89, 47, 6
- **Valores:** 0.446, -0.38, 0.364, 0.325, 0.312

Interpretación:

- **Observación 20:** DFFITS $= 0.446$ (la más influyente)
- **Conclusión:** 7 observaciones cambian significativamente sus propias predicciones $\rightarrow$ Investigar casos especiales

Diagnóstico Completo: Supuestos Básicos

Ejemplo: Modelo `horas_estudio ~ nota_examen` (n=100)

1. **LINEALIDAD:** [OK] CUMPLIDO
   - **Gráfico:** Línea loess prácticamente plana en Residuos vs Ajustados
   - **Test RESET:** $F = 1.051$, $p = 0.353 \rightarrow$ Forma funcional correcta
2. **HOMOCEDASTICIDAD:** [OK] CUMPLIDO
   - **Scale-Location:** Línea horizontal, dispersión constante
   - **Breusch-Pagan:** $LM = 0.02$, $p = 0.889 \rightarrow$ Varianza constante
   - **White:** $LM = 0.122$, $p = 0.941 \rightarrow$ Confirmado

Diagnóstico Completo: Supuestos Distribucionales

Ejemplo: Modelo `horas_estudio ~ nota_examen` (n=100)

3. **NORMALIDAD:** [OK] CUMPLIDO
   - **Q-Q Plot:** Puntos siguen línea diagonal perfectamente
   - **Shapiro-Wilk:** W = 0.99, p = 0.671 $\rightarrow$ Normalidad confirmada
   - **Jarque-Bera:** JB = 0.685, p = 0.71 $\rightarrow$ Distribución normal
4. **INDEPENDENCIA:** [OK] CUMPLIDO
   - **Residuos vs Orden:** Sin patrones temporales o secuenciales
   - **Durbin-Watson:** DW = 2.056, p = 0.61 $\rightarrow$ Sin autocorrelación
   - **Breusch-Godfrey:** LM = 0.14, p = 0.932 $\rightarrow$ Independencia confirmada

Diagnóstico Completo: Observaciones Influyentes

Ejemplo: Modelo `horas_estudio ~ nota_examen` (n=100)

DETECCIÓN DE PUNTOS PROBLEMÁTICOS:

- **Outliers:** 6 observaciones con $|t| > 2$
- **Alto Leverage:** 2 observaciones de alta palanca
- **DFFITS influyentes:** 7 observaciones que cambian sus predicciones
- **Cook influyentes:** 6 observaciones con alta influencia global

EVALUACIÓN DE RIESGO:

- **Situación:** [OK] Favorable - Sin solapamiento crítico leverage + outlier
- **Acción:** Revisar 9 observaciones específicas

Veredicto Final del Diagnóstico

Ejemplo: Modelo `horas_estudio ~ nota_examen` (n=100)

Interpretación completa:

- Por cada **hora adicional** de estudio, la calificación aumenta en promedio **0.099 puntos**
- El modelo explica el **80.7%** de la variabilidad en las calificaciones
- La relación es **altamente significativa** ($p < 0.001$)
- Todos los **supuestos se cumplen** $\rightarrow$ Las inferencias son válidas
- Existen obsrevaciones influyentes que requieren atención

Conclusiones y Próximos Pasos

Lo que hemos aprendido:

Proceso completo de modelado: exploración → formalización → estimación → inferencia → diagnóstico

Interpretación de coeficientes y medidas de bondad de ajuste

Validación mediante diagnóstico de supuestos

Limitaciones de la correlación vs. causalidad

Próximo tema: Regresión Lineal Múltiple

- Múltiples variables predictoras
- Control de variables confusas
- Interacciones entre predictores
- Selección de variables

La regresión simple es el fundamento → Todos estos conceptos escalan directamente

Referencias

- Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). Wiley.
- Fox, J., & Weisberg, S. (2018). *An R companion to applied regression* (3rd ed.). Sage.
- Harrell Jr, F. E. (2015). *Regression modeling strategies* (2nd ed.). Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2nd ed.). Springer.
- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill/Irwin.