

Métodos Estadísticos de Predicción



Diapositivas de la
Asignatura

Grado en Matemáticas

AUTORES

- Víctor Aceña Gil
- Isaac Martín de Diego

2025-2026



Copyright © 2025 Víctor Aceña Gil, Isaac Martín de Diego. Esta obra está licenciada bajo CC BY-SA 4.0, Creative Commons Atribución-Compartir Igual 4.0 Internacional.

Índice de Diapositivas

Índice de Diapositivas

- Tema 0: Introducción a los Modelos Estadísticos para la Predicción
- Tema 1: Regresión Lineal Simple
- Tema 2: Regresión Lineal Múltiple
- Tema 3: Ingeniería de Características
- Tema 4: Selección de Variables, Regularización y Validación
- Tema 5: Modelos Lineales Generalizados (GLM)

Regresión Lineal Simple

Víctor Aceña - Isaac Martín

DSLab

2025-08-08





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 → formalización → estimación → inferencia → diagnóstico



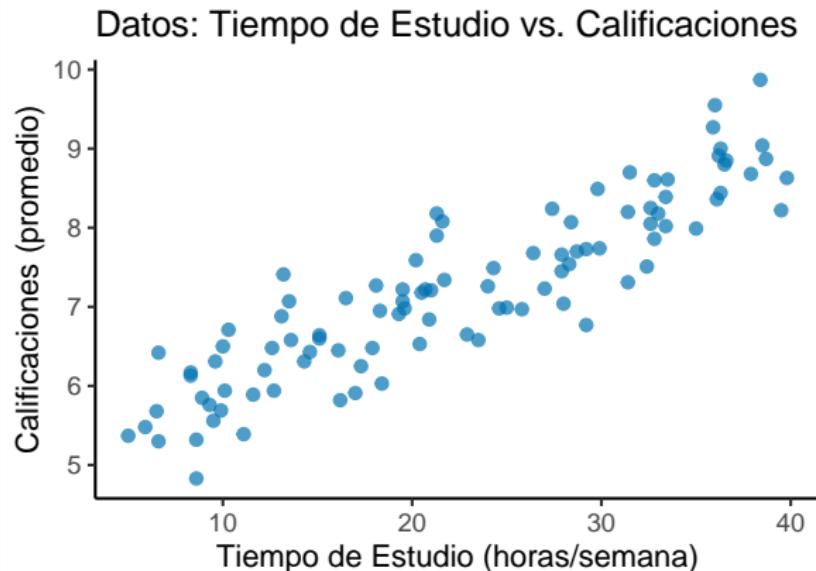
Al finalizar este tema, serás capaz de:

- 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



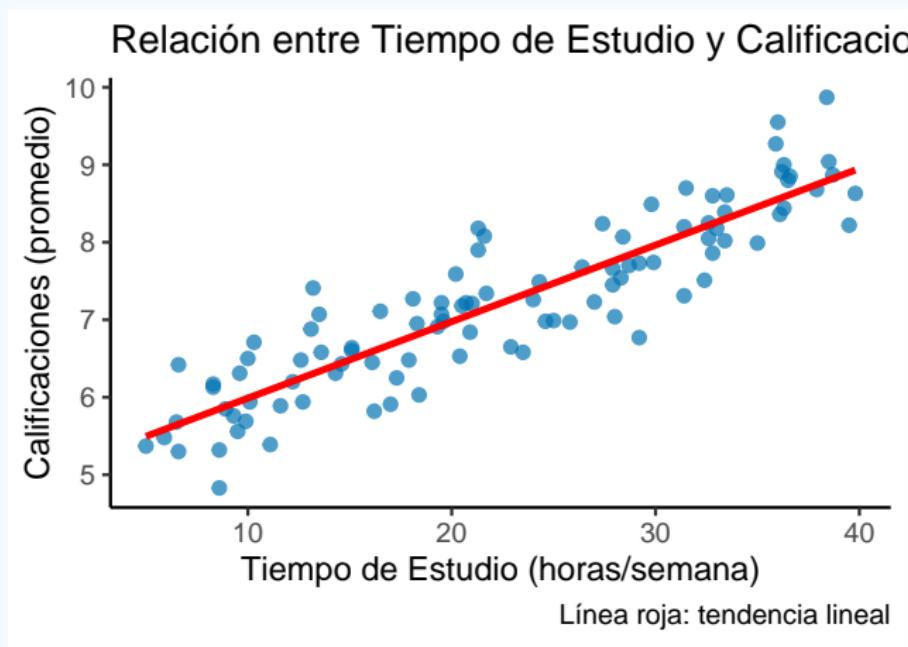
Pregunta de investigación: ¿Influye el tiempo de estudio semanal en las calificaciones finales?

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





Lo que vemos en el gráfico anterior:



Observación clave: Clara tendencia lineal positiva → Justifica un modelo de regresión lineal



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



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

Covarianza muestral:

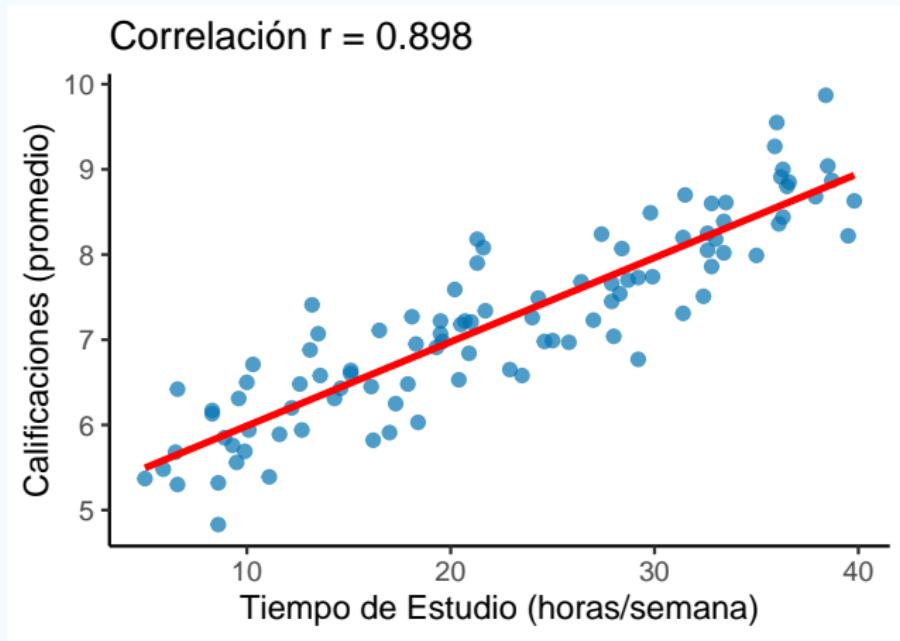
$$\text{Cov}(x, y) = s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

- **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*



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



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

¿Por qué?

Possibles 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



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 : **Intercesto** (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 → Usamos la muestra para estimar el **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



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:** $\text{Var}(\varepsilon_i|X_i) = \sigma^2$ (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.



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.



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):

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:

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



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



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**



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

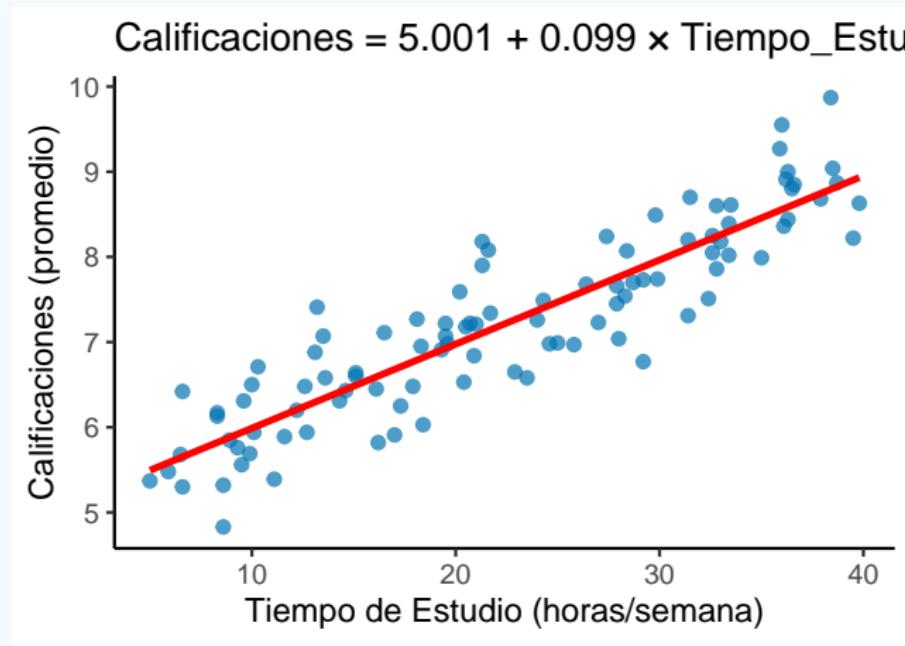
Pendiente ($\hat{\beta}_1$):

- 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.



Interpretación: Por cada hora adicional de estudio, la calificación aumenta en promedio 0.099 puntos.



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

1. Insesgadez:

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

2. Varianza mínima: Entre todos los estimadores lineales insesgados

3. Varianzas conocidas:

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 $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$ es la suma de cuadrados de X

Para el intercepto:

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



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* → Mide la dispersión promedio alrededor de la recta



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:

$$SST = SSR + SSE$$

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



Definición:

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

¿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.



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.



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 ecuación clave:

$$SST = SSR + SSE$$

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 → El modelo es útil



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



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: $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \approx 0$ (muy pequeña)

Consecuencia matemática:

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

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



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
- $F \gg 1 \rightarrow$ Rechazamos $H_0 \rightarrow$ El modelo **sí es útil**



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:

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

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



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



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

Distribución de los estimadores:

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

Estadístico t para la pendiente:

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

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



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 $\rightarrow \beta_1$ 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



Coeficientes:

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

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

Significancia:

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

Error estándar residual: 0.484 → Dispersión típica alrededor de la recta de regresión



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.



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:

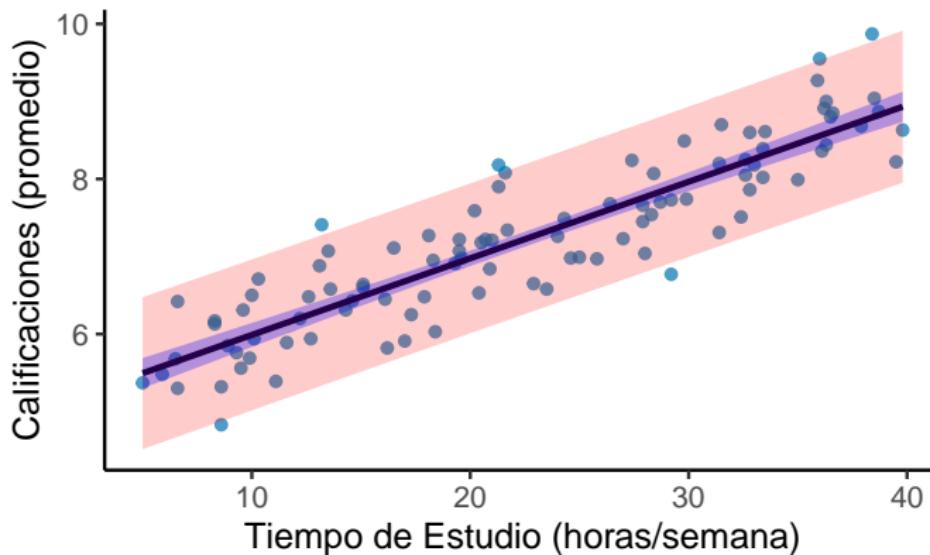
$$\hat{y}_0 \pm t_{\alpha/2, n-2} \cdot \sqrt{\text{MSE} \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)}$$

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



Intervalos de Confianza y Predicción



IC 95% para la media (azul) vs IP 95% para nueva observación (rojo)



¿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** → Conclusiones erróneas
- **Errores estándar incorrectos** → Intervalos de confianza y p-valores inválidos
- **Predicciones poco fiables** → Pérdida de poder predictivo

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



Recordatorio de supuestos:

- 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:** $\text{Var}(\varepsilon_i|X_i) = \sigma^2$ (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



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?

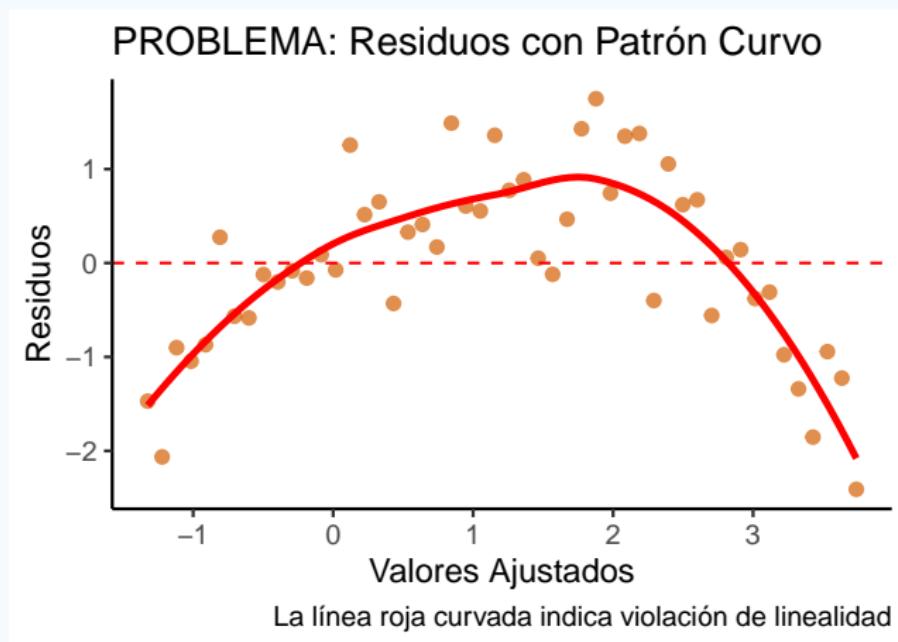
- **Patrón ideal:** Nube aleatoria de puntos centrada en cero
- **Violación:** Patrón curvilíneo (forma de "U" o parábola)

Test de Ramsey RESET:

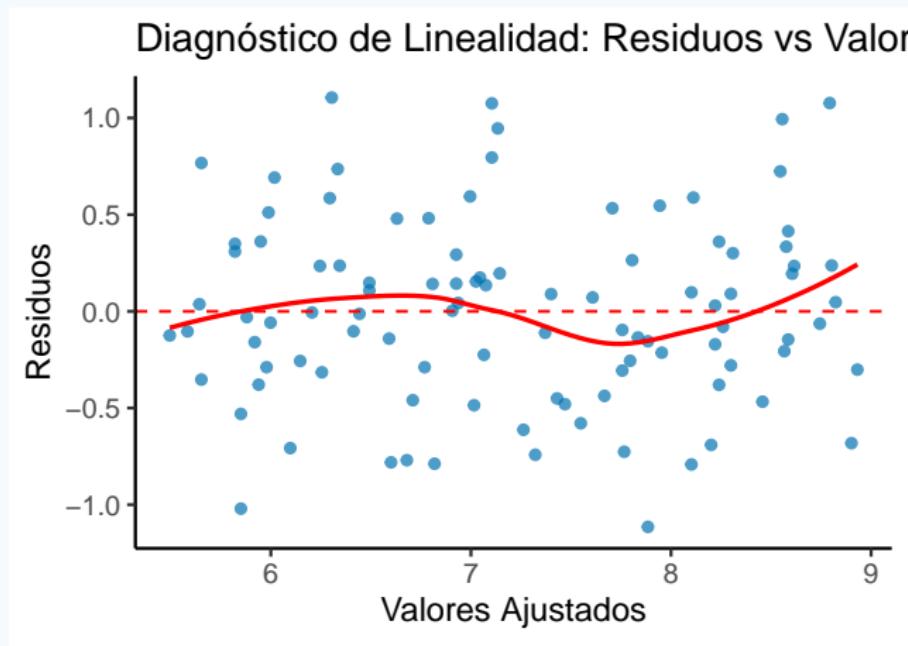
- H_0 : La forma funcional es correcta (lineal)
- H_1 : La forma funcional es incorrecta (no lineal)
- Añade términos $\hat{y}^2, \hat{y}^3, \dots$ al modelo y testa su significancia



Problema: Ajustar un modelo lineal a datos con relación cuadrática



Diagnóstico: Patrón curvo en residuos → **NO linealidad**



Resultados:

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



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?

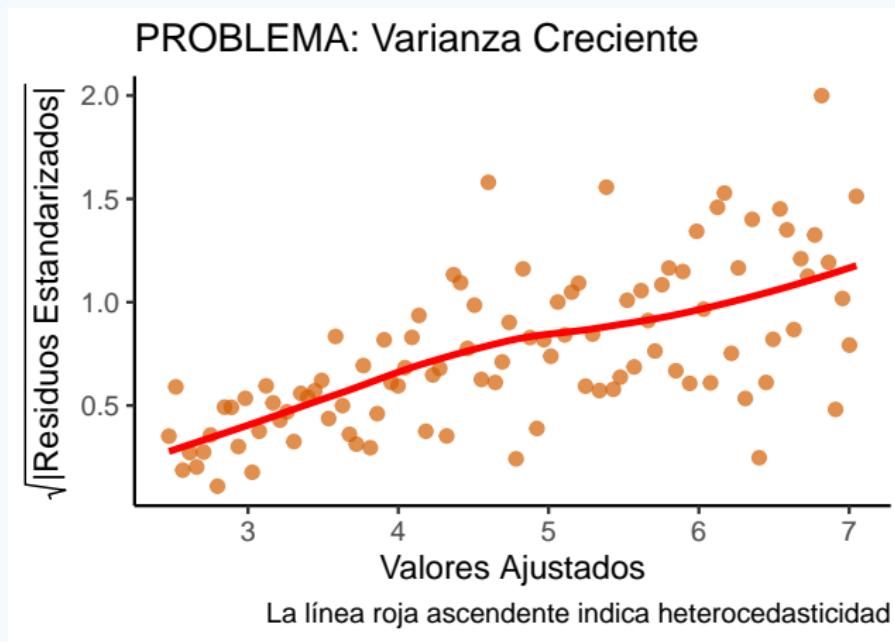
- **Patrón ideal:** Dispersion constante a lo largo del rango
- **Violación:** Forma de “embudo” (dispersion creciente o decreciente)

Tests de Heterocedasticidad:

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



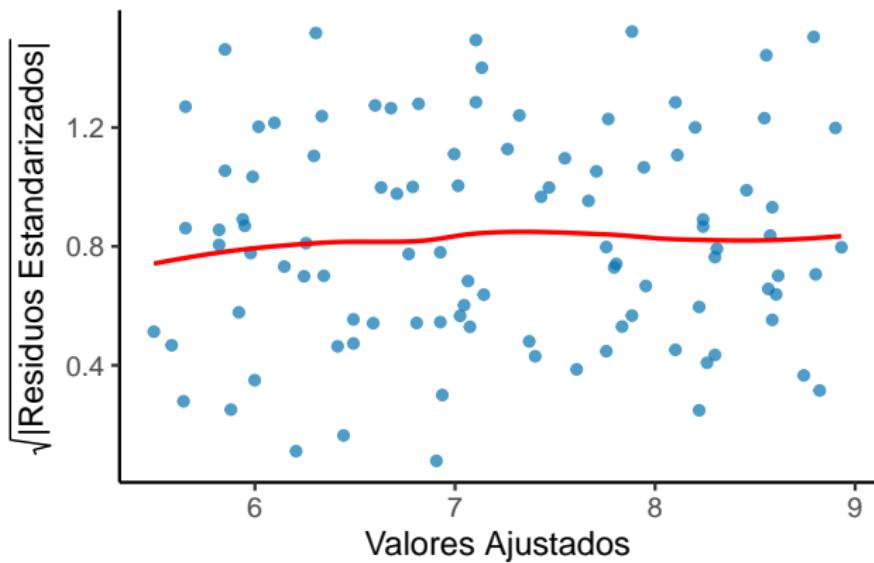
Problema: Varianza de los errores que aumenta con los valores predichos (heterocedasticidad)



Diagnóstico: Tendencia creciente → **Heterocedasticidad** (violación de)



Diagnóstico de Homocedasticidad: Scale-Lo



Resultados:

- **Gráfico:** Línea roja horizontal → Varianza constante
- **Breusch-Pagan:** $LM = 0.02$, $p = 0.889 \rightarrow$ Homocedasticidad



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? - **Q-Q Plot ideal:** Puntos sobre la línea diagonal -

Violación: Desviaciones sistemáticas de la línea (colas pesadas, asimetría)

Tests de Normalidad: - **Shapiro-Wilk:** H_0 : Los residuos siguen distribución normal - **Jarque-Bera:** Basado en asimetría y curtosis -

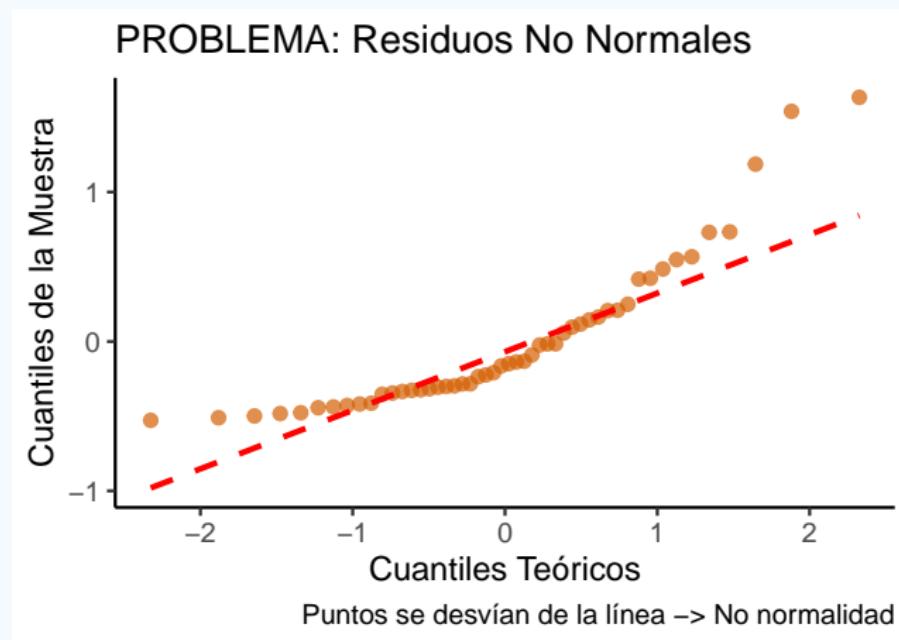
Anderson-Darling: Más sensible en las colas de la distribución

¿Qué buscamos?

- **Patrón ideal:** Puntos siguen la línea diagonal
- **Violación:** Desviaciones sistemáticas de la línea (colas pesadas, asimetría)



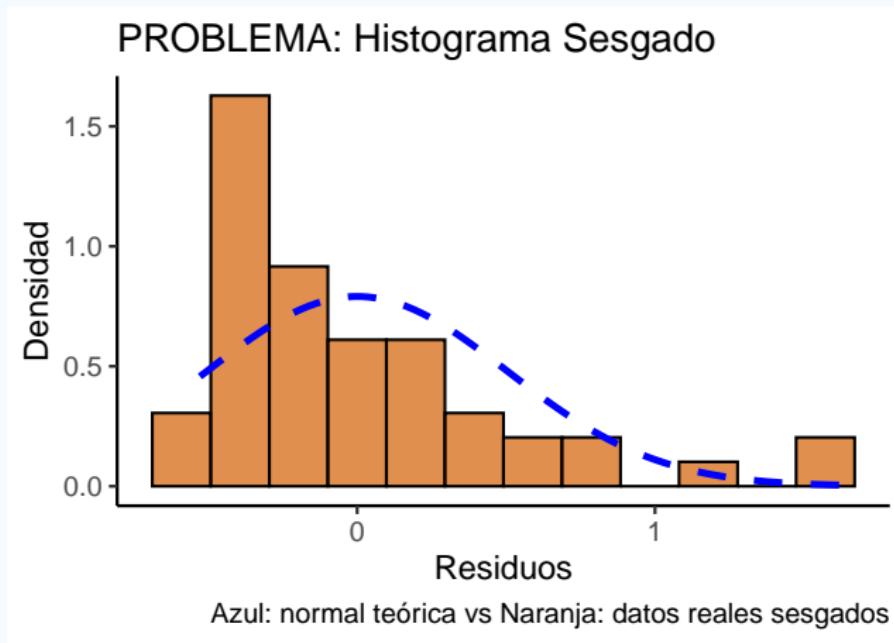
Problema: Errores con distribución asimétrica o con colas pesadas



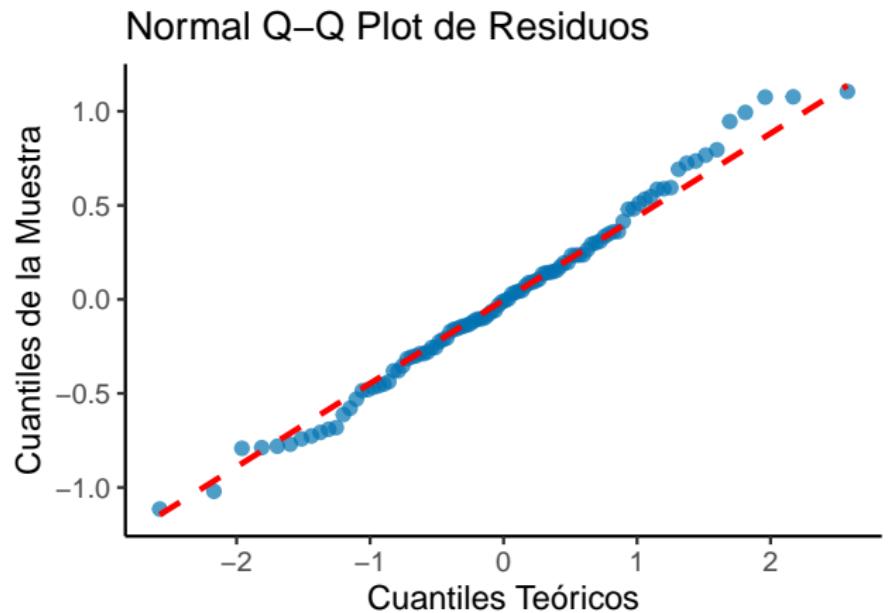
Diagnóstico: Puntos se alejan sistemáticamente de la línea → **NO normalidad**



Problema: Distribución asimétrica de los residuos (histograma sesgado)



Diagnóstico: Distribución sesgada \neq curva normal ($p = 0$) \rightarrow **NO normalidad**



Resultados: - **Gráfico:** Puntos siguen la línea diagonal → Normalidad - **Shapiro-Wilk:** $W = 0.99$, $p = 0.671 \rightarrow$ Normalidad - **Jarque-Bera:** $JB = 0.685$, $p = 0.71 \rightarrow$ Normalidad

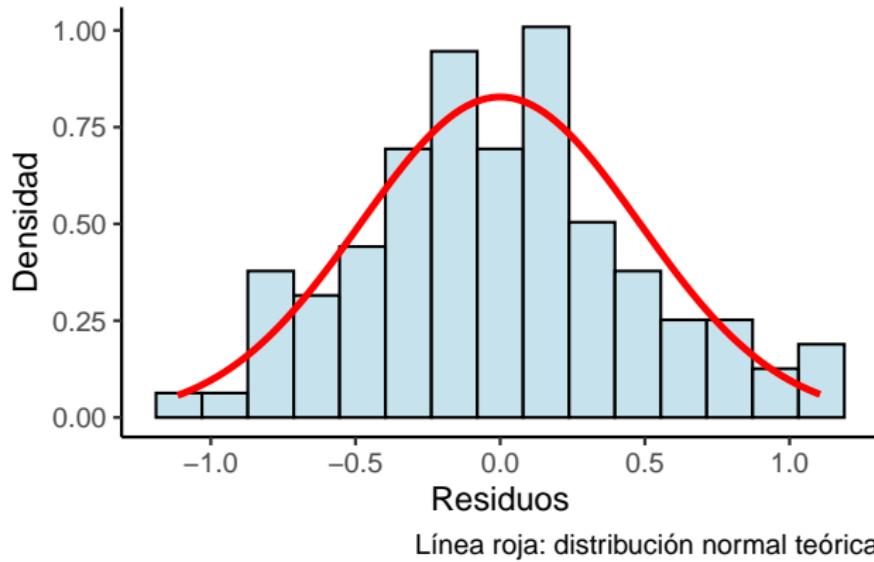


Complemento visual: Histograma de residuos con curva normal superpuesta

¿Qué buscamos?

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

Histograma de Residuos vs. Distribución Normal



Resultado: Distribución simétrica y campaniforme → Normalidad confirmada



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

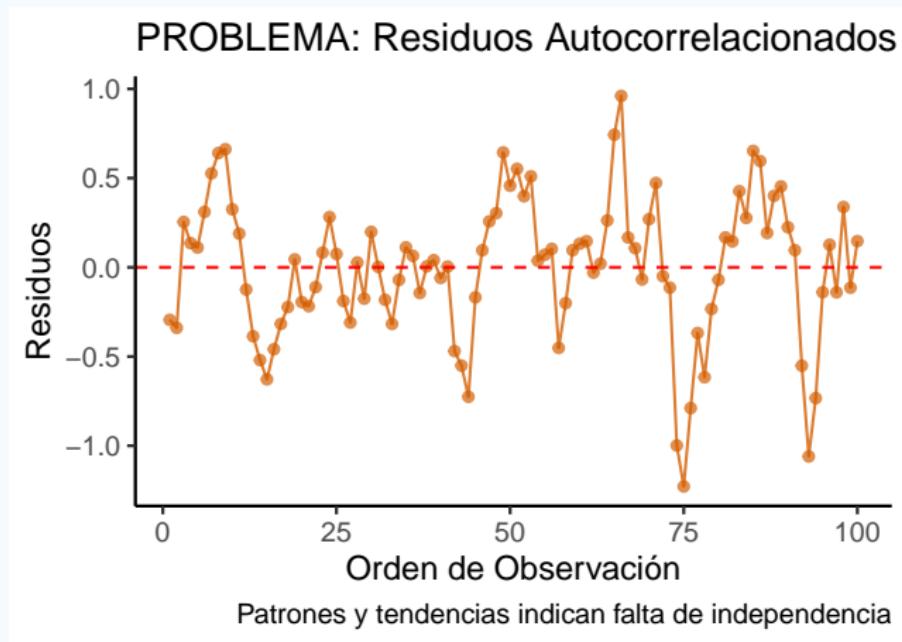
Tests de Autocorrelación:

- **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

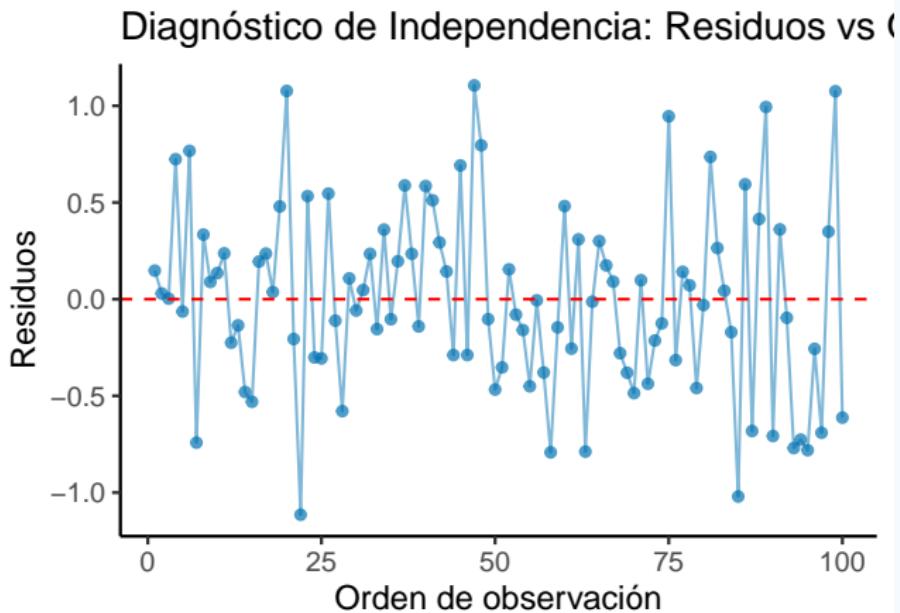




Problema: Residuos con autocorrelación (típico en series temporales)



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



Resultados: - **Gráfico:** Sin patrones temporales → Independencia -

Durbin-Watson: $DW = 2.056, p = 0.61 \rightarrow$ Sin autocorrelación orden 1 -

Breusch-Godfrey: $LM = 0.14, p = 0.932 \rightarrow$ Sin autocorrelación orden 2



Objetivo: Identificar puntos que tienen influencia desproporcionada en el modelo

Métricas Principales:

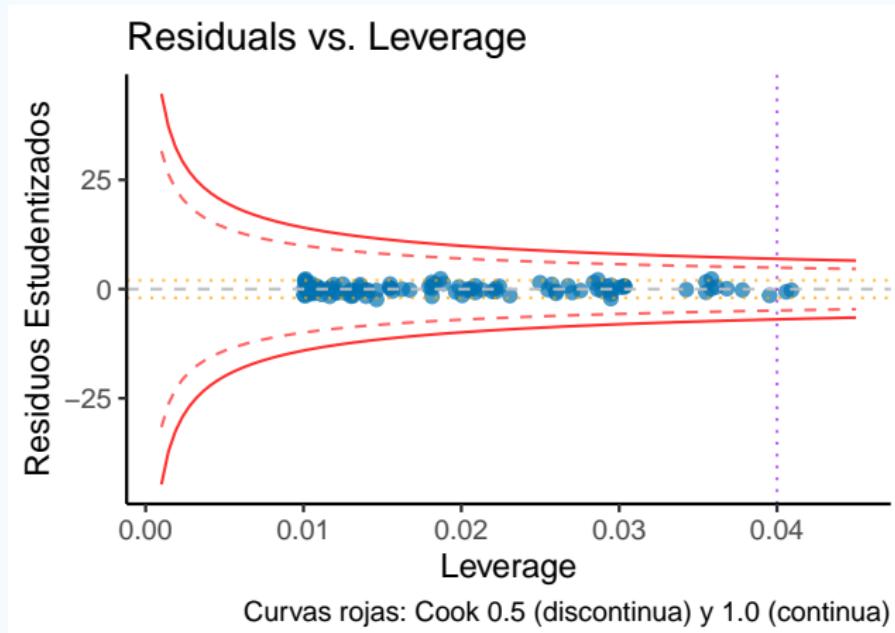
- **Leverage** (h_{ii}): Distancia en el espacio X (valores atípicos en X)
- **Residuos Estudentizados**: Outliers en Y ajustado por su varianza
- **Distancia de Cook** (D_i): Influencia global en los coeficientes

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**



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



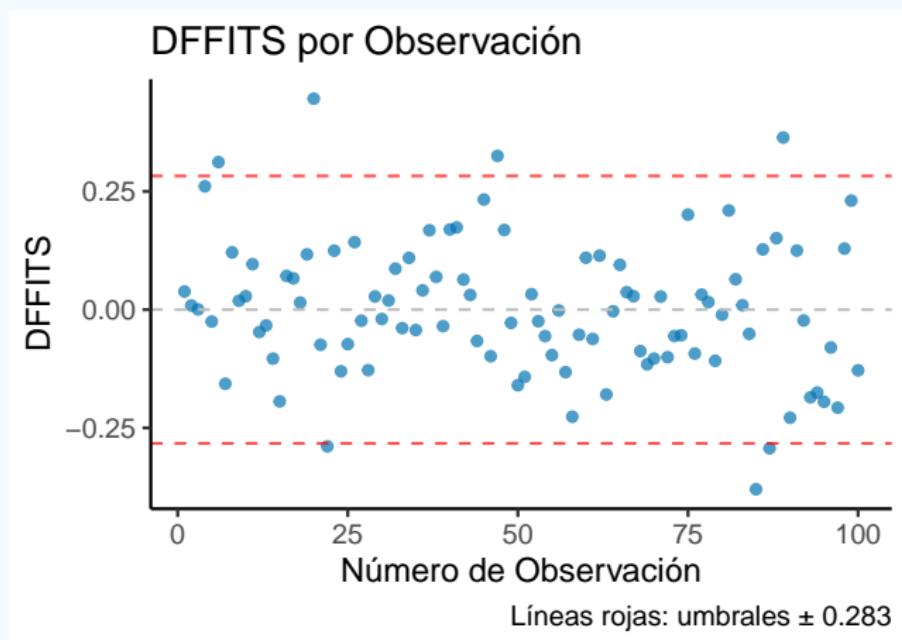
Identificación:

- **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





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 → Investigar casos especiales



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



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



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



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** → Las inferencias son válidas
- Existen observaciones influyentes que requieren atención



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



- 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.

Regresión Lineal Múltiple

Víctor Aceña - Isaac Martín

DSLab

2025-08-17





El modelo de regresión lineal múltiple constituye la **extensión natural y más potente** del modelo simple.

Diferencias clave:

Regresión Simple:

- Una variable respuesta
- Un único predictor
- Relación bivariada

Capacidades únicas:

- **Modelar simultáneamente** el efecto de múltiples variables predictoras
- **Interpretación de coeficientes** en presencia de otros predictores
- **Diagnóstico específico** del modelo múltiple
- Manejo de la **multicolinealidad**

Regresión Múltiple:

- Una variable respuesta
- **Múltiples predictores**
- Relación multivariada



Al finalizar este tema, serás capaz de:

- ① **Formular y estimar** modelos de regresión lineal múltiple, comprendiendo las diferencias clave respecto al caso simple
- ② **Interpretar coeficientes** en el contexto multivariante, entendiendo el concepto de *ceteris paribus*
- ③ **Realizar inferencia estadística** construyendo intervalos de confianza y contrastes de hipótesis
- ④ **Evaluar la calidad del ajuste** usando medidas como R^2 , R^2 ajustado y descomposición ANOVA
- ⑤ **Diagnosticar el modelo múltiple** aplicando técnicas específicas como gráficos CPR
- ⑥ **Identificar y tratar la multicolinealidad** usando el VIF como herramienta de diagnóstico
- ⑦ **Realizar predicciones** distinguiendo entre intervalos de confianza e intervalos de predicción



Para n observaciones y p variables predictoras, el **modelo poblacional** postula:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, \dots, n$$

Componentes:

- Y_i : i -ésima variable respuesta aleatoria
- X_{ij} : i -ésima variable predictora aleatoria del j -ésimo predictor
- ε_i : término de error aleatorio
- $\beta_0, \beta_1, \dots, \beta_p$: coeficientes poblacionales verdaderos pero desconocidos

Características clave:

- Relación **lineal en los parámetros**
- Los errores son **no observables**
- Los parámetros son **constantes poblacionales**



En la práctica, trabajamos con **datos observados** y estimamos el modelo:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}, \quad i = 1, \dots, n$$

Componentes:

- \hat{y}_i : i -ésima predicción
- x_{ij} : i -ésima observación del j -ésimo predictor
- $\hat{\beta}_j$: coeficientes estimados

Interpretación clave de $\hat{\beta}_j$:

El cambio estimado en la media de Y ante un cambio de una unidad en X_j , **manteniendo constantes todas las demás variables predictoras**.

Este principio se conoce como **ceteris paribus** ("lo demás constante")

**Modelo poblacional:**

$$\mathbf{Y} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \tilde{\mathbf{X}} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Nota: $\tilde{\mathbf{X}}$ contiene variables aleatorias (mayúsculas X_{ij})

**Modelo muestral:**

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \quad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$$

Observaciones:

- \mathbf{X} contiene datos observados (minúsculas x_{ij})
- \mathbf{X} y $\tilde{\mathbf{X}}$ son matrices de dimensión $n \times (p + 1)$
- La primera columna de unos corresponde al intercepto β_0



Condiciones de Gauss-Markov:

- ① **Linealidad en los parámetros:** El modelo $E[\mathbf{Y}|\tilde{\mathbf{X}}] = \tilde{\mathbf{X}}\boldsymbol{\beta}$ está bien especificado
- ② **Exogeneidad:** Los errores tienen media cero: $E[\boldsymbol{\varepsilon}|\tilde{\mathbf{X}}] = \mathbf{0}$
- ③ **Homocedasticidad e independencia:** $\text{Var}(\boldsymbol{\varepsilon}|\tilde{\mathbf{X}}) = \sigma^2 \mathbf{I}_n$
- ④ **Ausencia de multicolinealidad perfecta:** \mathbf{X} tiene rango completo ($p + 1$)
- ⑤ **Normalidad (para inferencia):** $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$

Implicación: Estos supuestos garantizan que los estimadores MCO sean **insesgados, consistentes y eficientes**



Principio: Minimizar la discrepancia entre valores observados y predichos

Función objetivo:

$$S(\beta) = \sum_{i=1}^n e_i^2 = \mathbf{e}^T \mathbf{e} = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$

¿Por qué cuadrados?

- Los residuos positivos y negativos no se cancelan
- Se penalizan más fuertemente los errores grandes
- Facilita el tratamiento matemático

Resultado: MCO minimiza la **Suma de los Cuadrados de los Residuos** (SSR)



Expandiendo la función objetivo:

$$S(\beta) = \mathbf{y}^T \mathbf{y} - 2\beta^T \mathbf{X}^T \mathbf{y} + \beta^T (\mathbf{X}^T \mathbf{X}) \beta$$

Derivando respecto a β :

$$\frac{\partial S(\beta)}{\partial \beta} = -2\mathbf{X}^T \mathbf{y} + 2(\mathbf{X}^T \mathbf{X})\beta$$

Igualando a cero:

$$-2\mathbf{X}^T \mathbf{y} + 2(\mathbf{X}^T \mathbf{X})\hat{\beta} = \mathbf{0}$$

Ecuaciones Normales:

$$(\mathbf{X}^T \mathbf{X})\hat{\beta} = \mathbf{X}^T \mathbf{y}$$



Solución única:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Condición necesaria: La matriz $(\mathbf{X}^T \mathbf{X})$ debe ser invertible

¿Cuándo es invertible?

- Cuando \mathbf{X} tiene rango completo ($p + 1$)
- Cuando las columnas de \mathbf{X} son linealmente independientes
- Cuando no hay multicolinealidad perfecta

Propiedades de $(\mathbf{X}^T \mathbf{X})$:

- Dimensión: $(p + 1) \times (p + 1)$
- Simétrica
- Definida positiva (si es invertible)



Bajo los supuestos de Gauss-Markov:

- ① **Insesgados:** $E[\hat{\beta}] = \beta$
- ② **Eficientes:** Varianza mínima entre todos los estimadores lineales insesgados
- ③ **Consistentes:** $\hat{\beta} \xrightarrow{p} \beta$ cuando $n \rightarrow \infty$

Matriz de varianza-covarianza:

$$\text{Var}(\hat{\beta}) = \sigma^2(\mathbf{X}^T \mathbf{X})^{-1}$$

Bajo normalidad adicional:

$$\hat{\beta} \sim N\left(\beta, \sigma^2(\mathbf{X}^T \mathbf{X})^{-1}\right)$$



Estimador insesgado de σ^2 :

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n - p - 1} = \frac{\sum_{i=1}^n e_i^2}{n - p - 1} = \frac{\mathbf{e}^T \mathbf{e}}{n - p - 1}$$

Grados de libertad: $n - p - 1$

- n : número de observaciones
- $p + 1$: número de parámetros estimados

Distribución:

$$\frac{(n - p - 1)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-p-1}^2$$

Error estándar de los coeficientes:

$$\hat{\sigma}_{\beta_j} = \hat{\sigma} \sqrt{(\mathbf{X}^T \mathbf{X})_{jj}^{-1}}$$



Datos: Precios de viviendas basados en características

Variables predictoras:

- superficie: Metros cuadrados
- habitaciones: Número de habitaciones
- antiguedad: Años de antigüedad
- distancia_centro: Distancia al centro (km)
- garaje: Presencia de garaje (Sí/No)



Coeficiente de regresión parcial:

$$\beta_j = \frac{\partial E[Y|\tilde{X}]}{\partial X_j}$$

Interpretación: β_j representa el cambio esperado en Y por una unidad de cambio en X_j , **manteniendo todas las demás variables constantes**

Diferencia crucial:

Regresión Simple:

- Efecto **total** (directo + indirecto)
- Puede estar **confundido**
- $\hat{\beta}_j$ captura toda la asociación

Regresión Múltiple:

- Efecto **puro o parcial**
- **Controla** por otras variables
- Interpretación más **causal**

Concepto clave: El coeficiente proviene de una regresión entre residuos



	Estimate	Std. Error
(Intercept)	53750.9705	6666.70624
superficie	1171.7780	47.28087
habitaciones	15072.3104	1303.41715
antiguedad	-744.5896	75.42075
distancia_centro	-2028.2715	164.87756
garajeSí	25829.4317	2349.44285

Interpretación *ceteris paribus*:

- **Superficie** (+1,172 €/m²): Cada m² adicional incrementa el precio
- **Habitaciones** (+15,072 €): Cada habitación adicional aumenta el precio
- **Antigüedad** (-745 €/año): Cada año de antigüedad reduce el precio
- **Distancia centro** (-2,028 €/km): Cada km más lejos del centro reduce el precio
- **Garaje** (+25,829 €): Tener garaje incrementa el precio



Descomposición ANOVA:

$$SST = SSR + SSE$$

Donde:

- **SST** (Sum of Squares Total): $\sum_{i=1}^n (y_i - \bar{y})^2$
- **SSR** (Sum of Squares Regression): $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$
- **SSE** (Sum of Squares Error): $\sum_{i=1}^n (y_i - \hat{y}_i)^2$

Interpretación:

- **SST**: Variabilidad total en los datos
- **SSR**: Variabilidad explicada por el modelo
- **SSE**: Variabilidad no explicada (residual)



R-cuadrado:

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

Interpretación:

- Proporción de la variabilidad en Y explicada por el modelo
- Rango: $0 \leq R^2 \leq 1$
- $R^2 = 0$: El modelo no explica nada
- $R^2 = 1$: El modelo explica toda la variabilidad

Problema: R^2 siempre aumenta al añadir variables (incluso irrelevantes)

En regresión múltiple: R^2 es el cuadrado de la correlación entre y y \hat{y}



R-cuadrado ajustado:

$$R_{\text{adj}}^2 = 1 - \frac{\text{SSE}/(n - p - 1)}{\text{SST}/(n - 1)} = 1 - (1 - R^2) \frac{n - 1}{n - p - 1}$$

Ventajas:

- **Penaliza** la inclusión de variables irrelevantes
- **Puede decrecer** si una variable no aporta información suficiente
- Mejor para **comparar modelos** con diferente número de predictores

Criterio de decisión:

- Si R_{adj}^2 aumenta al añadir una variable → la variable es útil
- Si R_{adj}^2 disminuye → la variable no aporta información suficiente



Hipótesis sobre un coeficiente:

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0$$

Estadístico de contraste:

$$t = \frac{\hat{\beta}_j - 0}{\hat{\sigma}_{\beta_j}} = \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{(\mathbf{X}^T \mathbf{X})_{jj}^{-1}}} \sim t_{n-p-1}$$

Interpretación:

- **Rechazar** H_0 : La variable X_j es estadísticamente significativa
- **No rechazar** H_0 : No hay evidencia de efecto lineal de X_j sobre Y

Valor p: Probabilidad de observar un estadístico t tan extremo o más, bajo H_0



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

$$\hat{\beta}_j \pm t_{\alpha/2, n-p-1} \cdot \hat{\sigma}_{\beta_j}$$

Interpretación:

- Con $(1 - \alpha)\%$ de confianza, el verdadero valor de β_j está en este intervalo
- Si el intervalo **no contiene cero** $\rightarrow \beta_j$ es significativo
- Si el intervalo **contiene cero** $\rightarrow \beta_j$ no es significativo

Relación con el test de hipótesis:

- Intervalo de confianza del 95% \equiv Test de hipótesis con $\alpha = 0.05$
- Si 0 está en el IC del 95% \rightarrow No se rechaza H_0 al 5%



Hipótesis global:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad \text{vs} \quad H_1 : \text{Al menos un } \beta_j \neq 0$$

Estadístico F:

$$F = \frac{\text{SSR}/p}{\text{SSE}/(n - p - 1)} = \frac{R^2/p}{(1 - R^2)/(n - p - 1)} \sim F_{p,n-p-1}$$

Interpretación:

- **Rechazar H_0 :** El modelo es globalmente significativo
- **No rechazar H_0 :** El modelo no explica variabilidad significativa

Relación con R^2 : El test F evalúa si R^2 es significativamente diferente de cero

Ejemplo: Inferencia en el Modelo de Viviendas



Call:

```
lm(formula = precio ~ superficie + habitaciones + antiguedad +
    distancia_centro + garaje, data = viviendas)
```

Residuals:

Min	1Q	Median	3Q	Max
-38847	-11074	867	9898	38486

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	53750.97	6666.71	8.063	7.53e-14 ***
superficie	1171.78	47.28	24.783	< 2e-16 ***
habitaciones	15072.31	1303.42	11.564	< 2e-16 ***
antiguedad	-744.59	75.42	-9.872	< 2e-16 ***
distancia_centro	-2028.27	164.88	-12.302	< 2e-16 ***
garajeSí	25829.43	2349.44	10.994	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15950 on 194 degrees of freedom

Multiple R-squared: 0.9094, Adjusted R-squared: 0.9071

F-statistic: 389.4 on 5 and 194 DF, p-value: < 2.2e-16





Predicción puntual: Para un nuevo vector \mathbf{x}_0 :

$$\hat{y}_0 = \mathbf{x}_0^T \hat{\boldsymbol{\beta}}$$

Dos tipos de intervalos:

Intervalo de Confianza:

- Para la **respuesta media**
 $E[Y|\mathbf{x}_0]$
- Incertidumbre en la estimación
- Más estrecho

Intervalo de Predicción:

- Para una **observación individual** Y_0
- Incertidumbre + variabilidad natural
- Más amplio

Fórmulas: Ambos dependen de $\hat{\sigma}^2$ y de la matriz $(\mathbf{X}^T \mathbf{X})^{-1}$



Intervalo de confianza para la respuesta media:

$$\hat{y}_0 \pm t_{\alpha/2, n-p-1} \cdot \hat{\sigma} \sqrt{\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}$$

Intervalo de predicción para una observación individual:

$$\hat{y}_0 \pm t_{\alpha/2, n-p-1} \cdot \hat{\sigma} \sqrt{1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}$$

Diferencia clave: El “+1” en el intervalo de predicción refleja la variabilidad adicional de una observación individual

Amplitud: Intervalo de predicción > Intervalo de confianza



Una vez ajustado el modelo, es **fundamental realizar un diagnóstico exhaustivo** para verificar que los supuestos se cumplen.

Base del diagnóstico: Análisis de los residuos - nuestra ventana a los errores teóricos no observables

Supuestos a verificar:

1. Normalidad

- Gráfico Q-Q de residuos
- Test de Shapiro-Wilk

2. Independencia

- Residuos vs tiempo
- Test de Durbin-Watson

3. Homocedasticidad

- Gráfico Scale-Location
- Test de Breusch-Pagan

4. Linealidad

- Residuos vs valores ajustados
- **Gráficos CPR** (específicos de múltiple)



Problema: El gráfico residuos vs ajustados puede ocultar una relación no lineal con **una variable específica**

Solución: Gráficos CPR para cada predictor X_j :

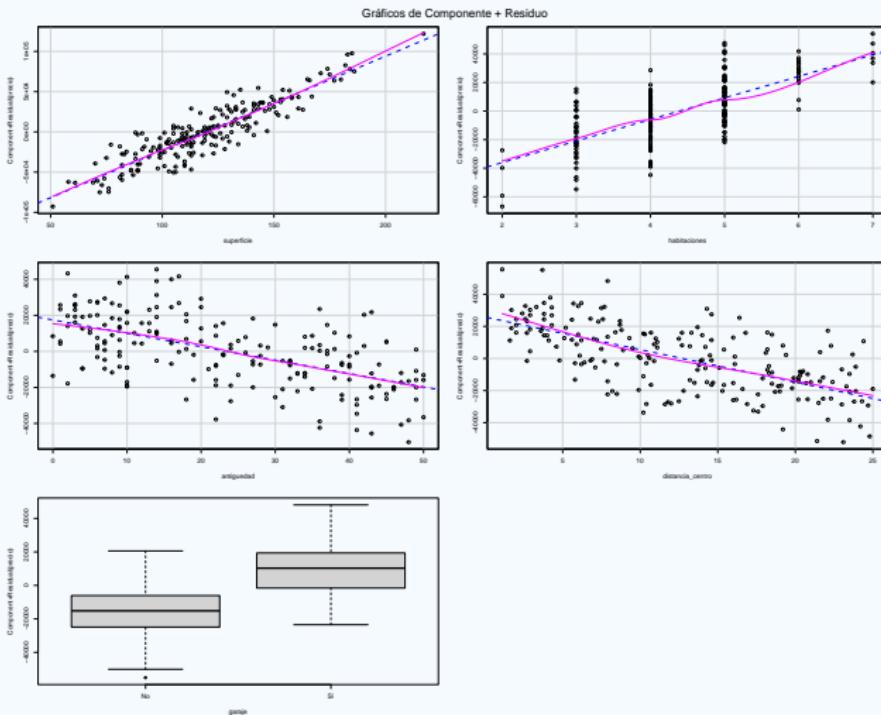
$$\text{Residuo Parcial} = e_i + \hat{\beta}_j x_{ij} \quad \text{vs.} \quad x_{ij}$$

Interpretación:

- **Línea sólida:** Relación lineal esperada (pendiente = $\hat{\beta}_j$)
- **Línea punteada:** Suavizado no paramétrico
- **Coincidencia:** Linealidad adecuada
- **Divergencia:** Posible no-linealidad → necesita transformación

Ventaja: Permite detectar no-linealidades específicas de cada variable

Ejemplo: Diagnóstico con Gráficos CPR





¿Qué observamos en las 5 variables?

Superficie y Habitaciones:

- Líneas sólida y punteada coinciden
- **Conclusión:** Relación lineal adecuada

Antigüedad y Distancia:

- Líneas coinciden bien
- **Conclusión:** Linealidad confirmada

Interpretación general:

- Relaciones lineales apropiadas
- No se necesitan transformaciones

Garaje:

- Separación clara entre grupos (No/Sí)
- **Conclusión:** Efecto categórico apropiado

Clave: Si las líneas divergen significativamente → considerar transformaciones



¿Qué es? Correlación alta entre variables predictoras

Consecuencias:

- ① **Varianza inflada:** Errores estándar muy grandes
- ② **Inestabilidad:** Pequeños cambios en datos → grandes cambios en coeficientes
- ③ **Contradicciones:** Modelo globalmente significativo pero ningún predictor individual significativo

Nota importante: La multicolinealidad **NO viola** los supuestos de Gauss-Markov, pero **arruina la interpretación práctica**

Detección:

- **Matriz de correlaciones:** Correlaciones > 0.8 son señal de alerta
- **VIF:** Herramienta definitiva de diagnóstico



Proceso de cálculo del VIF para X_j :

- ① Regresar X_j sobre **todas las demás variables predictoras**
- ② Obtener el R_j^2 de este modelo auxiliar
- ③ Calcular:

$$VIF_j = \frac{1}{1 - R_j^2}$$

Interpretación: Factor por el cual se infla la varianza de $\hat{\beta}_j$ debido a multicolinealidad

Reglas prácticas:

- **VIF = 1:** Ausencia de colinealidad (ideal)
- **VIF > 5:** Valores preocupantes que requieren atención
- **VIF > 10:** Multicolinealidad seria que debe ser tratada

Ejemplo: Diagnóstico de Multicolinealidad



Caso 1: Sin problemas de multicolinealidad

superficie	habitaciones	antiguedad
1.40	1.40	1.01

distancia_centro	garaje
1.01	1.01

Caso 2: Con multicolinealidad problemática

superficie_sim	habitaciones_sim	metros_cuadrados
86.4	8.5	81.2

Correlación superficie-metros_cuadrados: 0.994



La estrategia depende del objetivo del análisis:

1. No hacer nada

- Si el objetivo es **predicción**
- Si variables colineales no son de interés

2. Eliminar variables

- Quitar la menos relevante teóricamente
- Mantener la más correlacionada con Y

3. Combinar variables

- Crear índices compuestos
- Análisis de Componentes Principales

4. Métodos alternativos

- **Ridge regression:** Reduce varianza añadiendo sesgo
- **Lasso/Elastic Net:** Regresión penalizada

5. Aumentar muestra

- Más datos pueden reducir correlaciones
- No siempre factible



Conceptos básicos (como en regresión simple):

- **Outlier:** Residuo grande
- **Leverage:** Valor atípico en predictores
- **Influencia:** Impacto en el modelo

Herramientas específicas de regresión múltiple:

DFBETAS: Influencia sobre coeficientes individuales

$$\text{DFBETA}_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{\text{se}(\hat{\beta}_{j(-i)})}$$

Criterio: $|ext{DFBETA}_{j,i}| > \frac{2}{\sqrt{n}}$ es problemático

Ventaja: Permite identificar qué observaciones afectan a qué coeficientes específicos



Objetivo: Visualizar la relación entre Y y X_j **después de eliminar el efecto lineal de todos los demás predictores**

Construcción:

- ① Residuos de Y regresado sobre todos los predictores excepto X_j :
 $e_{Y|X_{-j}}$
- ② Residuos de X_j regresado sobre todos los demás predictores: $e_{X_j|X_{-j}}$
- ③ Graficar: $e_{Y|X_{-j}}$ vs $e_{X_j|X_{-j}}$

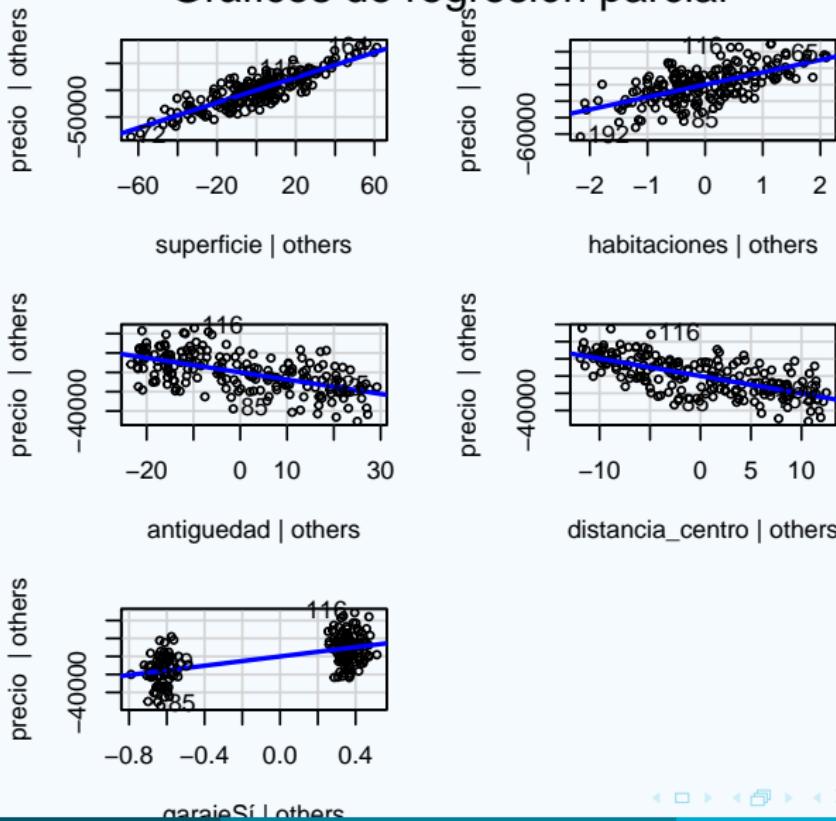
Propiedad mágica: La pendiente de la línea ajustada es **exactamente** $\hat{\beta}_j$

Utilidades:

- Visualizar magnitud y significancia del efecto “ajustado”
- Detectar no-linealidades en relaciones parciales
- Identificar observaciones influyentes para coeficientes específicos



Gráficos de regresión parcial





¿Qué vemos en cada gráfico?

- **Eje X:** Residuos de X_j vs. todos los demás predictores
- **Eje Y:** Residuos de Y vs. todos los demás predictores (excepto X_j)
- **Pendiente:** Es exactamente el coeficiente $\hat{\beta}_j$ del modelo múltiple

Interpretación por variable:

- **Superficie:** Relación lineal clara, pendiente positiva
- **Habitaciones:** Relación positiva, algunos puntos influentes
- **Antigüedad:** Relación negativa evidente
- **Distancia:** Relación negativa clara
- **Garaje:** Separación clara entre grupos (No/Sí)



La regresión múltiple permite:

- ① **Efectos parciales:** Aislar el impacto de cada variable predictora
- ② **Control de confusores:** Reducir sesgos por variables omitidas
- ③ **Mejores predicciones:** Incorporar múltiples fuentes de información
- ④ **Relaciones complejas:** Modelar fenómenos multifactoriales

Aspectos críticos:

- **Interpretación condicional:** Los coeficientes son efectos parciales (*ceteris paribus*)
- **Notación matricial:** Fundamental para la comprensión y computación
- **Supuestos:** Base para las propiedades de los estimadores
- **R^2 ajustado:** Mejor que R^2 para comparar modelos
- **Inferencia:** Tests individuales (t) y global (F)

Próximo paso: Diagnóstico del modelo y tratamiento de problemas específicos

Ingeniería de Características

Víctor Aceña - Isaac Martín

DSLab

2025-08-25





La **ingeniería de características** es el proceso fundamental que transforma y crea variables para maximizar la capacidad predictiva y la interpretabilidad de los modelos.

El problema:

- Los datos raramente están en forma óptima
- Relaciones no lineales ocultas
- Variables categóricas sin procesar
- Efectos de interacción ignorados

La solución:

- Transformaciones matemáticas precisas
- Codificación inteligente de categorías
- Creación de interacciones significativas
- Combinaciones y ratios informativos

Principio clave: “Los datos y la preparación de características determinan el límite superior del rendimiento; los modelos solo se aproximan a ese límite” - *Andrew Ng*



- ① **Identificar cuándo aplicar transformaciones** específicas según el problema detectado
- ② **Aplicar transformaciones clásicas y avanzadas** (logarítmica, Box-Cox, Yeo-Johnson) apropiadamente
- ③ **Interpretar modelos transformados** comprendiendo cómo cambian los coeficientes
- ④ **Codificar variables categóricas** usando ordinal encoding y one-hot encoding según su naturaleza
- ⑤ **Crear e interpretar interacciones** entre variables continuas, categóricas y mixtas
- ⑥ **Aplicar ingeniería avanzada** mediante combinaciones, ratios y transformaciones compuestas



- **Principio Clave:** Diagnosticar antes de transformar.
- **Riesgos de una Mala Práctica:**
 - **Sobreajuste:** Se pierde capacidad de generalización.
 - **Pérdida de Interpretabilidad:** Se aplican transformaciones sin base teórica.
 - **Violación de Supuestos:** Solucionar un problema creando otro.
 - **Sesgo de Selección:** Elegir por “mejor resultado” sin justificación.

Regla de oro: No transformes sin un diagnóstico previo. Cada cambio debe estar justificado.

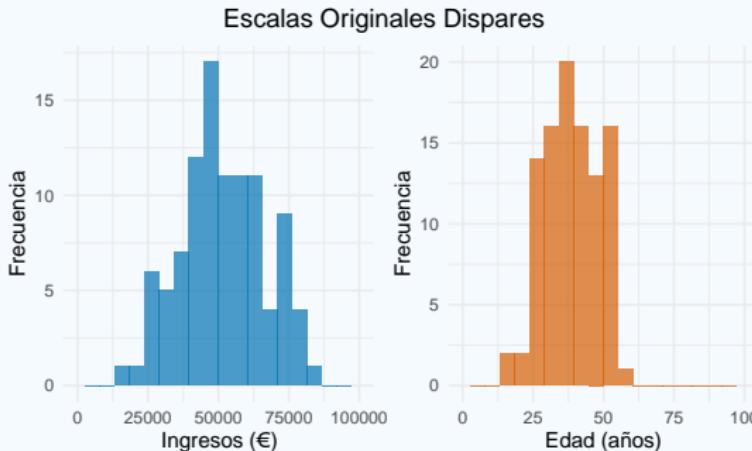


- **Clave:** Proceso sistemático basado en evidencia, no solo en métricas de ajuste (como el R^2).
- **Principios de Actuación:**
 - ① **Diagnóstico Previo:** Análisis visual y estadístico.
 - ② **Justificación Teórica:** Base conceptual para cada transformación.
 - ③ **Evaluación Integral:** Medir ajuste, interpretabilidad y robustez.
 - ④ **Validación Posterior:** Verificar la solución sin crear nuevos problemas.
 - ⑤ **Parsimonia:** Preferir siempre la solución más simple.



¿Por qué escalar?

- **Comparabilidad:** Coeficientes en misma escala
- **Regularización:** Penalización justa en Ridge/Lasso
- **Convergencia:** Optimización más eficiente
- **Interpretación:** Efectos estandarizados





$$X_{std} = \frac{X - \bar{X}}{\sigma_X}$$

Es la técnica más común. Transforma los datos para que tengan una **media de 0** y una **desviación estándar de 1**, pero **preserva la forma** de la distribución original.

Propiedades Clave

- **Preserva la forma:** Una distribución normal seguirá siendo normal.
- **Comparación fácil:** Permite evaluar el peso de variables con distintas unidades.
- **Robustez moderada:** Es menos sensible a *outliers* que la normalización Min-Max.

Aplicaciones Comunes

- En **regresión**, para comparar la importancia de los coeficientes.
- Como paso previo a **PCA** o análisis discriminante.
- Cuando las variables tienen distribuciones aproximadamente simétricas.



$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Esta técnica escala los datos a un rango fijo, comúnmente **[0, 1]**, donde 0 es el mínimo valor observado y 1 es el máximo.

- **Cuándo Usarlo:**

- Algoritmos que requieren entradas en un rango específico, como las **redes neuronales**.
- Cuando la interpretación en términos de mínimo y máximo es útil para el problema.
- En datos con distribuciones uniformes o sin *outliers* extremos.

- **Limitación Principal:**

- Es **muy sensible a outliers**. Un solo valor extremo puede comprimir el resto de los datos en un rango muy pequeño, perdiendo información sobre su variabilidad.



$$X_{robust} = \frac{X - \text{mediana}(X)}{\text{IQR}(X)}$$

Diseñado específicamente para datos con *outliers*. Utiliza la **mediana** y el **rango intercuartílico (IQR)**, que son medidas estadísticas no afectadas por valores extremos.

- **Principio Clave:**

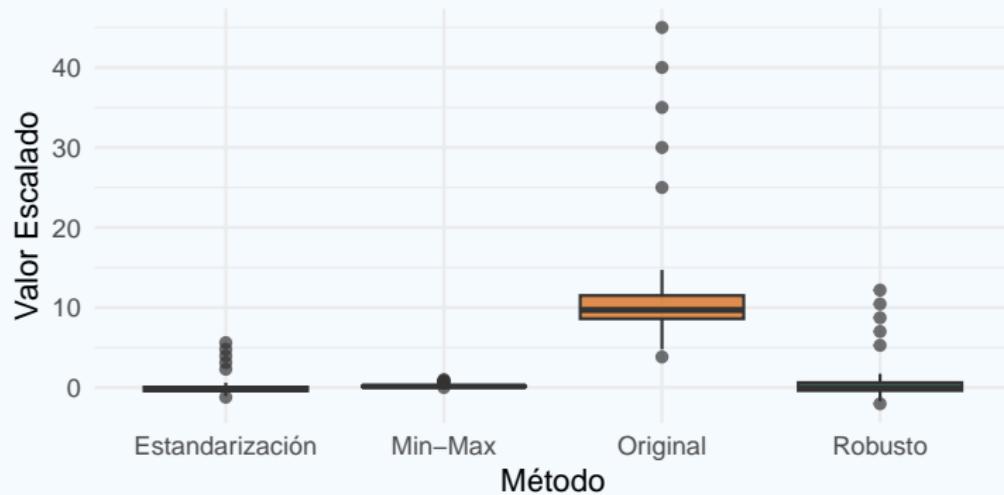
- Al no usar la media ni la desviación estándar, los *outliers* no distorsionan el resultado del escalado.

- **Cuándo Usarlo:**

- Es la opción preferida cuando se sabe o se sospecha que los **datos contienen *outliers*** significativos.
- Cuando se quiere preservar la estructura de la mayor parte de los datos sin la influencia de los valores extremos.



Impacto de Diferentes Métodos de Escalado Datos con outliers (puntos extremos en original)



Observación: El escalado robusto mantiene mejor la estructura central ante outliers



Una vez realizado el diagnóstico, debemos seleccionar la transformación más apropiada. La clave no está en *qué* transformación aplicar, sino en entender *por qué* esa transformación específica resuelve nuestro problema.

Exploraremos tres familias de transformaciones:

- ① Para **linearizar relaciones** no lineales.
- ② Para **estabilizar la varianza** (heterocedasticidad).
- ③ Para **normalizar residuos** y controlar *outliers*.



Es la transformación más versátil, ideal para linearizar relaciones de crecimiento exponencial o donde los efectos son multiplicativos.

Cuándo Usarla

- Relaciones **exponentiales** o **multiplicativas**.
- Procesos de **crecimiento proporcional**.
- Variables con **rendimientos decrecientes** (ingresos, precios).
- **Casos Típicos:** Economía, biología, finanzas.

Diagnóstico y Aplicación

- **Diagnóstico:** Curva cóncava que se aplana o varianza que aumenta con Y.
- **Aplicación:** $\log(Y) \sim X$, $Y \sim \log(X)$ o $\log(Y) \sim \log(X)$.
- **Interpretación:** Los coeficientes se leen como cambios porcentuales o elasticidades.



Fundamental para relaciones curvilíneas que siguen una **ley de potencia** del tipo $Y = a * X^b$.

- **Diagnóstico:** La relación entre las variables se vuelve lineal al graficarla en una **escala log-log**.
- **Aplicación:** Se toman **logaritmos en ambas variables** para linearizar el modelo: $\log(Y) = \log(a) + b * \log(X)$.
- **Interpretación:** El exponente b representa la **elasticidad** o el exponente de escalamiento.
- **Ejemplos Clásicos:** Ley de Kleiber (relación masa-metabolismo), economía urbana (población-PIB).



Especialmente útil para **datos de conteo** (típicamente de una distribución de Poisson), donde la varianza es proporcional a la media.

Cuándo Aplicarla

- **Conteos de eventos:** número de defectos, llamadas, ventas por período.
- **Datos de frecuencia:** visitas, clics, transacciones.
- Para conteos con muchos ceros puede requerir $\text{sqrt}(Y + 0.5)$.

Diagnóstico y Limitaciones

- **Diagnóstico:** Gráfico de residuos en forma de embudo donde la dispersión crece linealmente con la media.
- **Limitaciones:** La interpretación es menos directa y solo es apropiada para valores no negativos.



Es la solución natural cuando la **varianza es proporcional al cuadrado de la media**, lo que se conoce como heterocedasticidad multiplicativa.

- **Cuándo Aplicarla:**

- **Variables monetarias** (ingresos, precios, costos), donde el error relativo tiende a ser constante.
- Porcentajes de crecimiento o procesos donde los **errores se acumulan multiplicativamente**.

- **La Gran Ventaja (Efectos Múltiples):**

- Con frecuencia, esta única transformación resuelve varios problemas a la vez: **lineariza la relación, estabiliza la varianza, normaliza la distribución y reduce el impacto de outliers**.



Útil para relaciones **hiperbólicas** (del tipo $Y = 1/X$) y para distribuciones con colas muy pesadas a la derecha.

Cuándo Usarla

- Relaciones que se aproximan a una **asíntota horizontal**.
- **Tasas de decaimiento** o relaciones dosis-respuesta en farmacología.
- **Tiempo hasta un evento**.

Efecto y Precauciones

- **Efecto en Outliers:** Comprime fuertemente los valores grandes y expande los pequeños.
- **Precaución:** Amplifica errores en valores pequeños y requiere tratamiento especial para datos cercanos a cero.
- Solo aplicable a valores **no nulos**.



Es un método que **optimiza automáticamente** el parámetro de transformación λ (lambda) para maximizar la normalidad y homocedasticidad de los residuos. En lugar de elegir manualmente, Box-Cox encuentra el valor λ que mejor normaliza los datos.

Definición Matemática:

$$Y(\lambda) = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log(Y), & \lambda = 0 \end{cases}$$

Casos Especiales de λ :

- $\lambda = 1$: Sin transformación (identidad).
- $\lambda = 0.5$: Transformación de raíz cuadrada.
- $\lambda = 0$: Transformación logarítmica.
- $\lambda = -1$: Transformación inversa.



Propósito y Ventajas

- Encuentra la transformación **óptima** sin necesidad de prueba y error.
- Maximiza la verosimilitud del modelo, mejorando simultáneamente la **normalidad** y la **homocedasticidad**.
- Proporciona un método **objetivo** para seleccionar la transformación más apropiada.

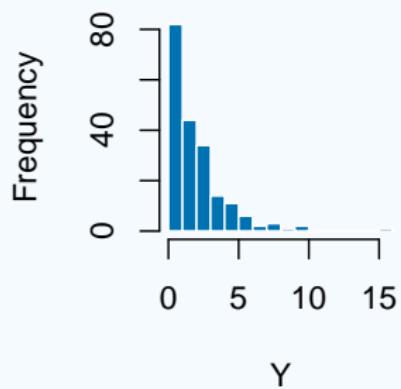
Limitaciones Importantes

- **Requiere que los datos (Y) sean estrictamente positivos.** Esta es su principal restricción.
- La **interpretación** es compleja si λ no es un valor simple (0, 0.5, 1).
- El λ óptimo **depende del modelo** específico, por lo que puede cambiar si se modifican los predictores.

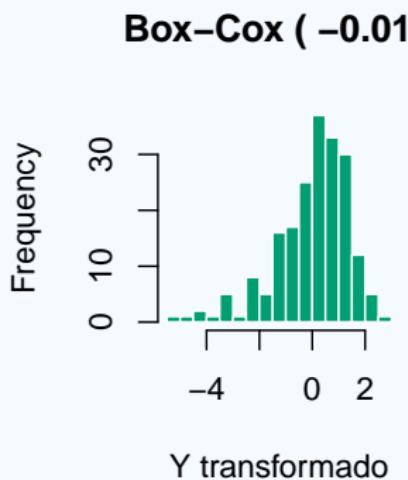


A continuación, se visualiza cómo la transformación Box-Cox corrige la asimetría de una distribución original.

Original (Sesgado)



Box-Cox (-0.01)





Fue desarrollada para superar la principal limitación de Box-Cox: **acepta cualquier valor real (positivo, negativo o cero).**

Cuándo Usar Box-Cox

- Cuando los datos son **estrictamente positivos.**
- Si se busca comparabilidad con literatura existente que la utilice.

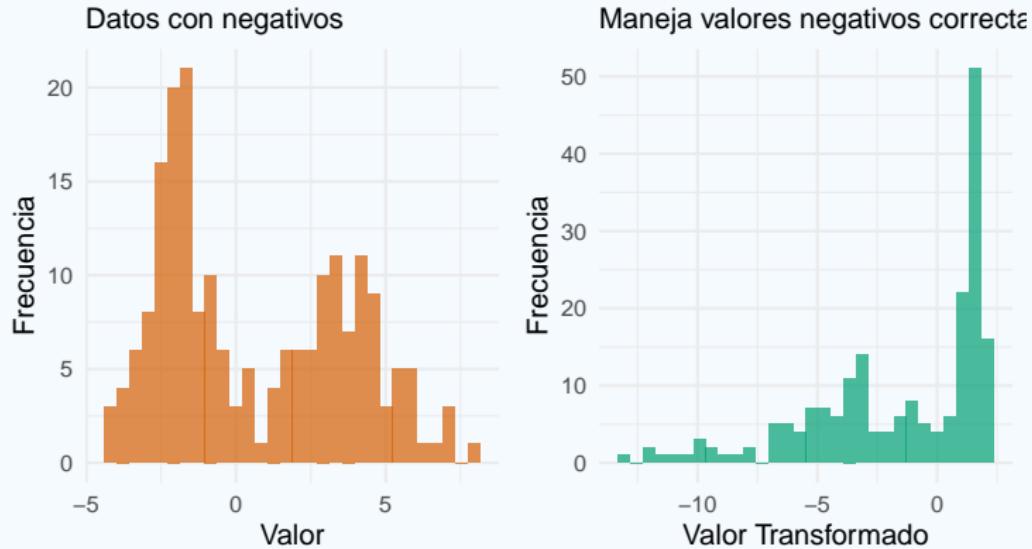
Cuándo Usar Yeo-Johnson

- Cuando los datos incluyen **valores negativos o cero.**
- Si se necesita mayor **flexibilidad** y no hay restricciones de dominio.



Este ejemplo muestra cómo Yeo-Johnson maneja un conjunto de datos que incluye valores negativos, algo que Box-Cox no podría hacer.

Original vs. Transformación Yeo–Johnson





La mayoría de los algoritmos estadísticos requieren entradas numéricas. El objetivo de la codificación es transformar las categorías de texto en números, **preservando su información semántica sin introducir supuestos erróneos**.

Criterios para Seleccionar el Método de Codificación:

- **Naturaleza de la variable:** ¿Existe un orden inherente entre las categorías? Esta es la pregunta más importante.
 - **Nominal:** Sin orden (ej. color, país).
 - **Ordinal:** Con orden (ej. nivel educativo, satisfacción).
- **Número de categorías:** Variables con muchas categorías (“alta cardinalidad”) pueden requerir técnicas especiales.
- **Interpretabilidad del modelo:** ¿Qué método facilita una explicación clara de los resultados?



Transforma una variable con **k categorías sin orden** (ej. color, región) en **k-1 variables binarias** (0/1), también conocidas como *variables dummy*.

La “Dummy Variable Trap”:

- **Problema:** Usar k columnas (una para cada categoría) crea colinealidad perfecta en modelos lineales, ya que una columna es una combinación lineal de las otras.
- **Solución:** Siempre se omite una categoría, que se convierte en la **categoría de referencia** del modelo.

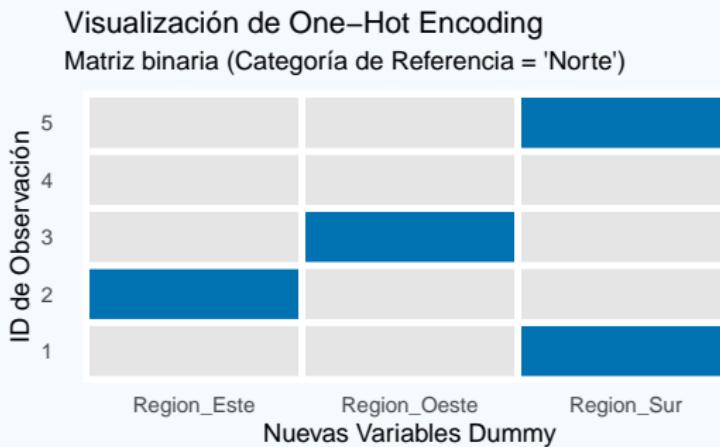
Ventajas

- No asume un orden entre categorías.
- Cada nueva variable tiene un coeficiente directamente interpretable (la diferencia con la categoría de referencia).

Desventajas

- Aumenta mucho la dimensionalidad si hay muchas categorías.
- Genera matrices de datos con muchos ceros (*dispersas*).

Imaginemos una variable Region con 4 categorías. Al codificarla, elegimos “Norte” como categoría de referencia.



En un modelo de regresión, el coeficiente de Region_Sur se interpretaría como la diferencia promedio en la variable respuesta al estar en la región “Sur” **en comparación con la región “Norte”**.



Asigna números enteros consecutivos a las categorías, **respetando su orden o jerarquía natural**.

Ejemplo Práctico:

- **Variable:** Nivel_Satisfaccion con categorías ["Bajo", "Medio", "Alto"].
- **Codificación:** Se asignan los valores [1, 2, 3].

Ventajas

- Preserva la información jerárquica de la variable.
- Es muy eficiente: crea una sola columna, sin importar el número de categorías.

Desventajas

- **Supone que la “distancia” entre niveles es uniforme** (asume que el cambio de “Bajo” a “Medio” es igual que de “Medio” a “Alto”).
- Puede imponer un orden artificial si se aplica por error a una variable nominal.



La elección incorrecta del método puede llevar a modelos con menor rendimiento e interpretaciones erróneas.

Usar Codificación Ordinal

- **Cuándo:** Para variables con un **orden natural y significativo** (ej. nivel educativo, satisfacción del cliente, grado de severidad).
- **Resultado:** Una sola columna numérica (1, 2, 3, ...).
- **Riesgo Principal:** Asumir que los intervalos entre categorías son iguales.

Usar One-Hot Encoding

- **Cuándo:** Para variables **nominales**, sin un orden inherente (ej. color, género, país).
- **Resultado:** $k-1$ columnas binarias (0 o 1).
- **Riesgo Principal:** Aumento excesivo del número de variables si la cardinalidad es alta.



Mientras los efectos principales miden el impacto promedio de una variable, las **interacciones** revelan cómo el efecto de una variable **cambia según el nivel de otra**.

Modelo con Interacción:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \times X_2 + \varepsilon$$

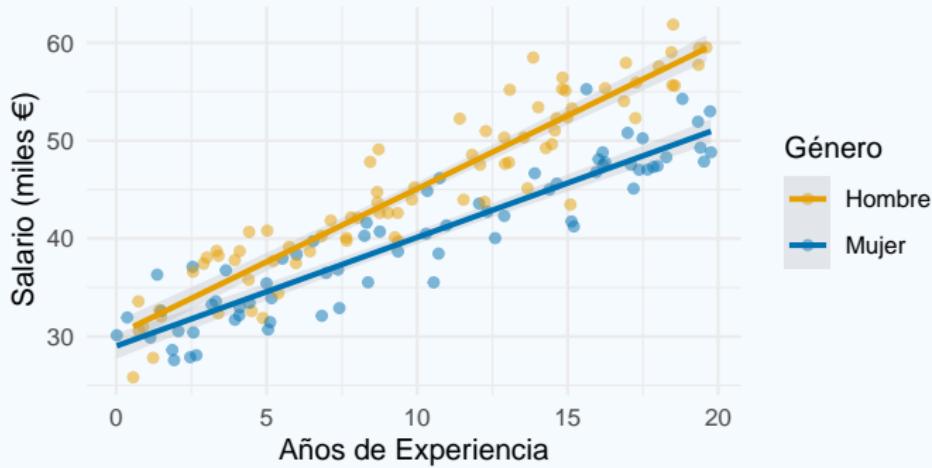
Interpretación de β_3 :

- Si $\beta_3 > 0$: Los efectos se **potencian** (sinergia).
- Si $\beta_3 < 0$: Los efectos se **atenúan** (compensación).



Un ejemplo clásico es cómo la relación entre experiencia y salario no es la misma para todos los grupos. La interacción permite que la pendiente de la experiencia sea diferente para cada género, como se ve en el gráfico.

Interacción Experiencia × Género
Pendientes diferentes = Interacción presente





Este tipo de interacción ocurre cuando el efecto de una variable continua sobre el resultado depende del valor de otra variable continua.

- **Concepto:** El efecto de una variable se **amplifica** o **atenúa** a medida que otra variable cambia.
- **Modelo:** $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 \times X_2$. El coeficiente β_3 captura la interacción.

Sintaxis en R:

```
modelo <- lm(ventas ~ precio * publicidad, data = datos)
# Interpretación: El efecto del precio sobre las ventas varía
# según el nivel de inversión en publicidad.
```



Ocurre cuando el efecto de pertenecer a una categoría depende de la pertenencia a otra categoría.

- **Concepto:** Existe un **efecto específico para una combinación de categorías** que no se puede explicar sumando los efectos individuales.
- **Ejemplo:** El efecto del género en el salario puede ser diferente en cada departamento de una empresa.

Sintaxis en R:

```
modelo <- lm(salario ~ genero * departamento, data = datos)
# Interpretación: La brecha salarial de género es diferente
# en cada departamento.
```



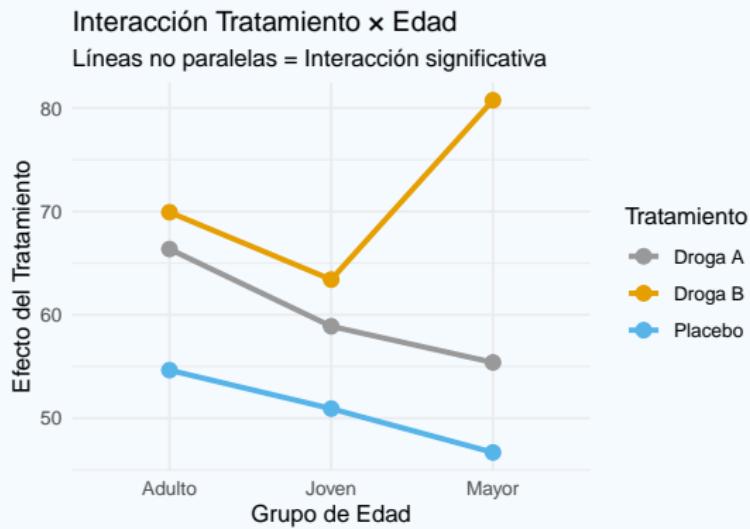
Este es uno de los tipos más comunes e intuitivos. Permite que la relación entre una variable continua y el resultado sea diferente para distintos grupos.

- **Concepto:** La **pendiente** de la variable continua es diferente para cada nivel de la variable categórica.
- **Ejemplo:** La relación entre los años de experiencia y el salario puede tener una pendiente más pronunciada para un grupo que para otro.

Sintaxis en R:

```
modelo <- lm(rendimiento ~ horas_estudio * metodo, data = datos)
# Interpretación: La efectividad de las horas de estudio
# (la pendiente)
# sobre el rendimiento varía según el método de estudio
# utilizado.
```

Los *interaction plots* son la mejor herramienta para entender interacciones entre variables categóricas. **Las líneas no paralelas son una señal visual clara de una posible interacción.**



La Droga B es especialmente efectiva en el grupo de “Mayor”, un efecto que no se podría ver analizando las variables por separado.



No debemos buscar interacciones al azar. El enfoque correcto combina la teoría con la evidencia de los datos.

¿Cómo Detectarlas?

- **Justificación Teórica:** El conocimiento del dominio es la guía principal para saber dónde buscar.
- **Exploración Visual:** Usar gráficos de dispersión por grupos o *interaction plots*.
- **Tests Estadísticos:** Utilizar el Test F para comparar formalmente un modelo con y sin la interacción.

El Principio de Jerarquía

- **Regla de Oro:** Si incluyes una interacción $A \times B$, **siempre debes incluir los efectos principales A y B por separado.**
- **Justificación:** Preserva la interpretabilidad del modelo y evita sesgos en los coeficientes.



Las interacciones son poderosas, pero su uso requiere cuidado para no complicar el modelo innecesariamente.

- **Riesgo de Complejidad:**

- El número de interacciones posibles crece exponencialmente.
- **Recomendación:** Limitar el modelo a las 2 o 3 interacciones más importantes y con justificación teórica.

- **Riesgo de Multicolinealidad:**

- Las interacciones pueden hacer que los coeficientes sean inestables.
- **Mitigación:** Centrar las variables continuas antes de crear el término de interacción.

- **Interpretación con Transformaciones:**

- **Advertencia:** Una interacción en una escala log no significa lo mismo que en una escala lineal y requiere una interpretación mucho más cuidadosa.



Consiste en crear nuevas variables mediante **combinaciones, ratios y transformaciones compuestas**. El objetivo es capturar relaciones complejas que no son evidentes en las variables originales.

A menudo, las variables individuales contienen información parcial. Al combinarlas de forma inteligente, podemos revelar **patrones predictivos mucho más potentes**.

Técnicas Clave que Exploraremos:

- **Combinaciones:** Creación de nuevas variables a partir de sumas o productos de las existentes.
- **Ratios y Proporciones:** Normalización de variables para revelar relaciones estructurales.
- **Manejo de Colinealidad:** Estrategias para condensar información de predictores correlacionados.



Combinaciones Lineales

- **Concepto:** Sumas ponderadas de variables que miden aspectos de un mismo fenómeno.
- **Ejemplo (Índice Compuesto):** Un índice de riesgo cardiovascular. $\text{Índice_Riesgo} = 0.4 * \text{Presión} + 0.3 * \text{Colesterol} + 0.3 * \text{IMC}$
- **Aplicación:** Crear un *score* único a partir de múltiples indicadores para capturar un constructo multidimensional.

Combinaciones No Lineales

- **Concepto:** Productos, cocientes o funciones complejas que capturan sinergias o efectos multiplicativos.
- **Ejemplo (Producto de Eficiencia):** Rendimiento = Capacidad \times Utilización \times Calidad
- **Aplicación:** Modelar efectos donde el resultado depende de la combinación simultánea de varios factores.



Los ratios son muy potentes porque **normalizan automáticamente las diferencias de escala** y revelan relaciones estructurales que las variables absolutas ocultan.

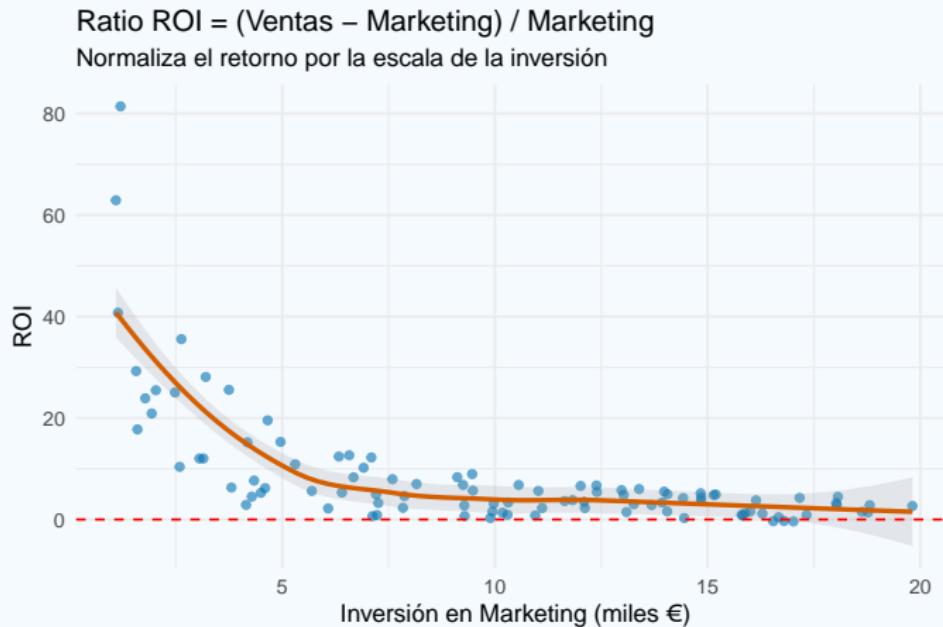
Ventajas Principales:

- **Normalización Automática:** Permiten comparar entidades de diferentes tamaños (ej. una startup vs. una multinacional).
- **Interpretación Intuitiva:** Tienen significados claros y directos (ej. Deuda / Patrimonio).
- **Robustez:** Suelen ser menos sensibles a valores atípicos (*outliers*).

Ejemplo Práctico: Retorno de Inversión (ROI)



Un ratio clásico en negocio es el ROI, que mide la eficiencia de una inversión normalizando el beneficio obtenido por el coste de la misma.





El Problema: Simplemente eliminar variables correlacionadas es una mala práctica, ya que **se pierde información predictiva valiosa**.

La Solución: Condensar la información redundante en nuevas variables, **preservando la información única** de cada predictor original.

Estrategias Principales

- **Componentes Principales (PCA):** Extrae la máxima varianza común en componentes ortogonales. Su principal desventaja es que **pierde interpretabilidad directa**.
- **Ratios Informativos:** Capturan la relación estructural entre dos variables (ej. Deuda / Patrimonio). **Preservan la interpretabilidad económica**.
- **Índices Ponderados:** Crean un score único a partir de varias variables, usando pesos teóricos o empíricos.



A medida que aumenta la complejidad de la técnica, generalmente se gana poder predictivo pero se pierde facilidad de interpretación.

Técnicas Más Simples

- **Estandarización:** La pérdida de interpretación es **mínima** (solo cambia la escala).
- **Transformación Logarítmica:** La pérdida es **baja**, ya que se puede interpretar como cambios porcentuales.

Técnicas Más Potentes

- **Box-Cox (con λ complejo):** La escala transformada no es intuitiva, dificultando la interpretación.
- **PCA (Componentes Principales):** La pérdida es **muy alta**, ya que los componentes son constructos abstractos.
- **Interacciones Múltiples:** La interpretación se complica al haber **efectos condicionales**.

- **Si buscas EXPLICAR → Prioriza la Interpretabilidad.**
- **Si buscas PREDECIR → Prioriza el Rendimiento.**



① Transformar sin Diagnóstico Previo

- Aplicar `log()` "por si acaso" en lugar de identificar un problema específico que lo justifique.

② Ignorar el Dominio del Problema

- Usar transformaciones que no tienen sentido teórico (ej. `log(edad)`).

③ Caer en la “Dummy Variable Trap”

- Incluir k columnas para k categorías en lugar de $k-1$ (categoría de referencia).

④ Usar Interacciones sin Efectos Principales

- Modelar $Y \sim A:B$ en lugar de la forma correcta y jerárquica $Y \sim A + B + A:B$.

⑤ Sobreingeniería (*Over-engineering*)

- Crear cientos de *features* automáticamente sin una justificación clara y validada para cada uno.



Un proceso sistemático garantiza resultados robustos y reproducibles.

① Análisis Exploratorio

- Visualizar distribuciones, patrones y *outliers*.

② Diagnóstico de Problemas

- Buscar no linealidad, heterocedasticidad, asimetría, etc., en los datos.

③ Selección de Transformaciones

- Elegir la técnica adecuada para el problema diagnosticado.

④ Aplicación y Validación

- Transformar las variables y verificar si el problema original se ha resuelto.

⑤ Comparación de Modelos

- Usar métricas (R^2 , RMSE, AIC) y validación cruzada para evaluar la mejora.

⑥ Interpretación Final

- Traducir los resultados del modelo final al contexto del negocio o la investigación.



1 Documentación Rigurosa

- Registrar cada transformación y su justificación para mantener la trazabilidad.

2 Validación en Datos Nuevos

- Guardar los parámetros de transformación del set de entrenamiento (ej. media, sd, lambda) y aplicarlos al de test.

3 Evitar *Data Leakage*

- Calcular todos los parámetros de las transformaciones **únicamente con los datos de entrenamiento**.

4 Considerar el Contexto

- Asegurarse de que las nuevas variables creadas son interpretables para los *stakeholders*.

5 Parsimonia

- Preferir siempre el modelo más simple que funcione adecuadamente. No añadir complejidad por mejoras marginales.



¿Por Qué es Fundamental?

- **Resuelve problemas** específicos de los datos (no linealidad, etc.).
- **Mejora el rendimiento** del modelo significativamente.
- **Revela relaciones ocultas** mediante interacciones y combinaciones.
- **Adapta los datos** a los requisitos de los algoritmos.

Próximo tema: Métodos de regularización para manejar la complejidad agregada.

Principios Clave

- **Diagnóstico** antes que transformación.
- **Justificación** teórica o empírica.
- **Validación** rigurosa.
- **Balance** entre complejidad e interpretabilidad.
- **Documentación** exhaustiva.

Selección de Variables, Regularización y Validación

Víctor Aceña - Isaac Martín

DSLAB

2025-09-02





En modelos de regresión con gran número de variables predictoras enfrentamos el desafío crítico de identificar qué variables son realmente relevantes

Los problemas principales:

- Inclusión de demasiadas variables → **sobreajuste**, pérdida de **interpretabilidad**, complejidad innecesaria
- Exclusión de variables importantes → **modelos subóptimos**
- Con p variables explicativas: 2^p modelos diferentes posibles
- Exploración exhaustiva **computacionalmente inviable** cuando p es grande

El objetivo del tema: Seleccionar el subconjunto óptimo de variables predictoras y validar la calidad del modelo resultante



Seis enfoques sistemáticos:

① Filtrado basado en información básica

- Eliminación preliminar de variables irrelevantes
- Criterios: variabilidad, correlación, VIF

② Criterios de bondad de ajuste

- Métricas para comparar modelos: AIC, BIC, Cp de Mallows

③ Métodos de selección exhaustiva

- Evaluación sistemática: Best Subset Selection

④ Métodos automáticos paso a paso

- Selección iterativa: forward, backward, stepwise

⑤ Métodos basados en regularización

- Penalización de complejidad: Ridge, Lasso, Elastic Net

⑥ Validación del modelo

- División train/test y validación cruzada



Etapas del proceso sistemático:

- ① Definición del problema y variables de interés
- ② Recogida de datos (fiabilidad, validez, ética, control de sesgos)
- ③ Análisis Exploratorio de Datos (EDA)
- ④ Ajuste del modelo inicial
- ⑤ Evaluación del modelo (R^2 , ANOVA, significancia)
- ⑥ Diagnóstico del modelo (residuos, observaciones atípicas)
- ⑦ Reducción de variables ← Enfoque principal del tema
- ⑧ Validación del modelo ← Enfoque principal del tema

Este tema se centra en las etapas 7 y 8



Clasificación según el diseño:

- **Experimentos controlados:** Manipulación deliberada de variables independientes
- **Estudios observacionales exploratorios:** Sin intervención, registro natural
 - Transversales (un momento temporal)
 - Longitudinales (seguimiento temporal)
- **Estudios observacionales confirmatorios:** Testear hipótesis específicas
- **Encuestas y cuestionarios:** Datos estructurados sobre actitudes/comportamientos
- **Experimentos naturales:** Fenómenos naturales como intervención
- **Estudios de simulación:** Modelos matemáticos/computacionales
- **Datos secundarios:** Bases de datos existentes



Objetivo: Filtrado preliminar antes de métodos sofisticados

Criterios de eliminación básicos:

① Variabilidad de las variables predictoras

$$\text{Var}(X_j) = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 < \epsilon$$

(típicamente $\epsilon = 0.01$)

② Correlación con la variable respuesta

$$r_{X_j,Y} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Umbral mínimo: $|r_{X_j,Y}| > \delta$ (ej: $\delta = 0.1$)



Detectar y eliminar variables redundantes:

- ③ **Multicolinealidad extrema** Si $|r_{X_j, X_k}| > 0.95 \rightarrow$ eliminar una variable
- ④ **Factor de Inflación de la Varianza (VIF)**

$$VIF_j = \frac{1}{1 - R_j^2}$$

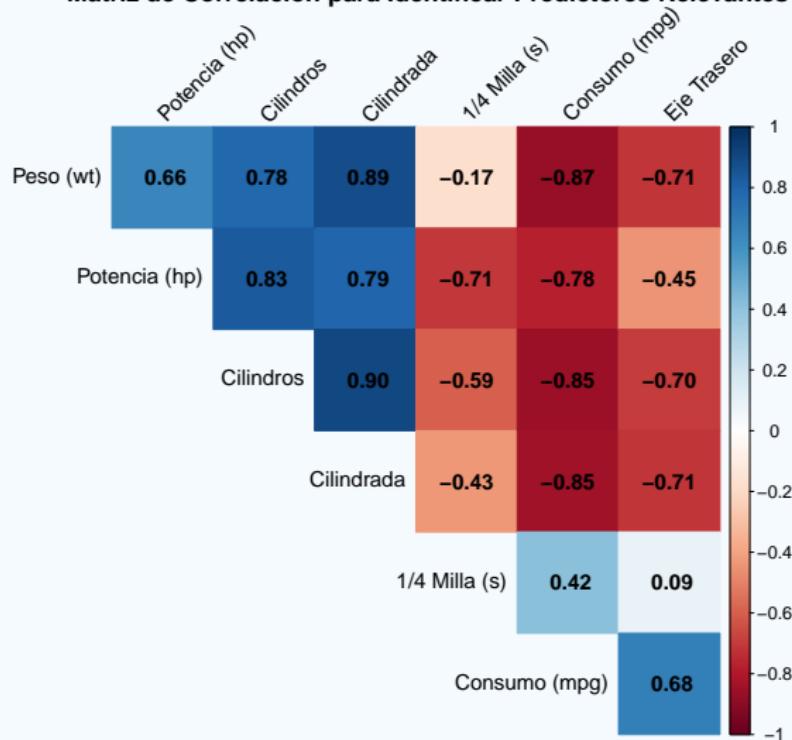
Valores $VIF_j > 10$ indican multicolinealidad problemática

Estrategia: Eliminar variables con mayor VIF iterativamente hasta que todos los VIF sean aceptables

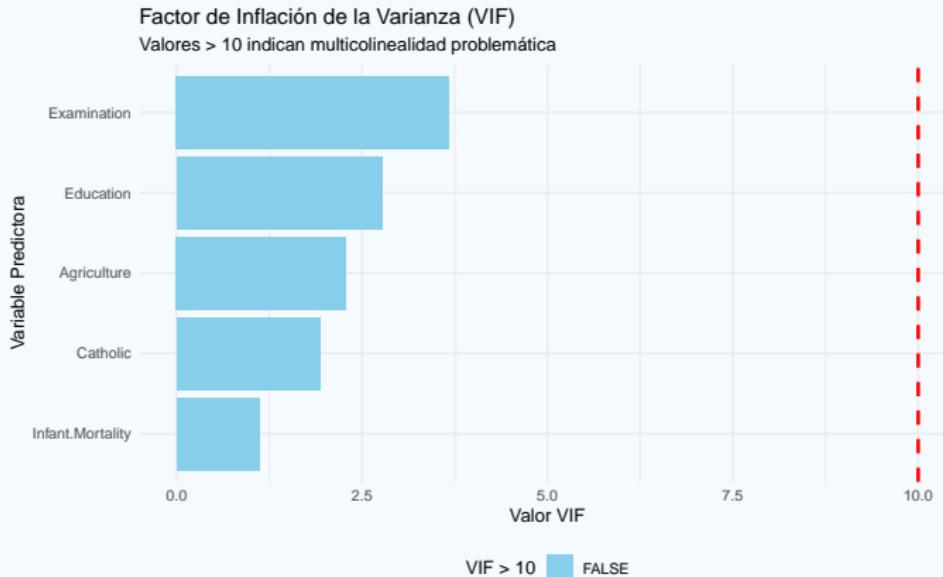
Visualización: Matriz de Correlación



Matriz de Correlación para Identificar Predictores Relevantes



Visualización: Diagnóstico de Multicolinealidad





Problema: Equilibrar capacidad explicativa vs complejidad del modelo

Subajuste vs Sobreajuste: - Muy pocas variables → subajuste (underfitting) - Demasiadas variables → sobreajuste (overfitting)

Tres criterios principales:

- ① Criterio de Información de Akaike (AIC)
- ② Criterio de Información Bayesiano (BIC)
- ③ Estadístico Cp de Mallows

Estrategia: Seleccionar el modelo que minimice el criterio elegido



Fundamento: Teoría de la información de Hirotugu Akaike

Objetivo: Estimar la pérdida de información del modelo

Fórmula:

$$AIC = n \ln \left(\frac{SSE}{n} \right) + 2(p + 1)$$

Componentes:

- $n \ln(SSE/n)$: **Bondad de ajuste** (relacionado con log-verosimilitud)
- $2(p + 1)$: **Penalización por complejidad** (aumenta 2 unidades por parámetro)

Interpretación:

- Menor AIC = mejor modelo
- Asintóticamente eficiente
- **Orientado a predicción**



Fundamento: Estadística bayesiana (Gideon Schwarz)

Objetivo: Encontrar el modelo más probable de ser el “verdadero”

Fórmula:

$$BIC = n \ln\left(\frac{SSE}{n}\right) + (p + 1) \ln(n)$$

Diferencia clave con AIC:

- Penalización: $(p + 1) \ln(n)$ en lugar de $2(p + 1)$
- Más restrictivo cuando $n > 7$ (ya que $\ln(n) > 2$)

Características:

- **Consistencia:** Si el modelo verdadero está entre candidatos, $P(\text{selección}) \rightarrow 1$
- **Orientado a explicación**
- Favorece modelos más simples (parsimonia)



Fundamento: Error cuadrático medio de predicción

Objetivo: Modelo con bajo sesgo y baja varianza

Fórmula:

$$C_p = \frac{SSE_p}{MSE_{full}} - n + 2(p + 1)$$

donde MSE_{full} es el error cuadrático medio del modelo completo

Interpretación:

- **Modelo bien especificado:** $C_p \approx p + 1$
- $C_p > p + 1$: modelo sesgado (variable importante omitida)
- $C_p \leq p + 1$: buen ajuste

Estrategia: Buscar modelos donde $C_p \approx p + 1$, elegir el menor entre ellos



Si el objetivo principal es la predicción:

- **AIC** es la opción preferida
- Diseñado para minimizar error de predicción
- Penalización más moderada
- Ideal en contextos de pronóstico

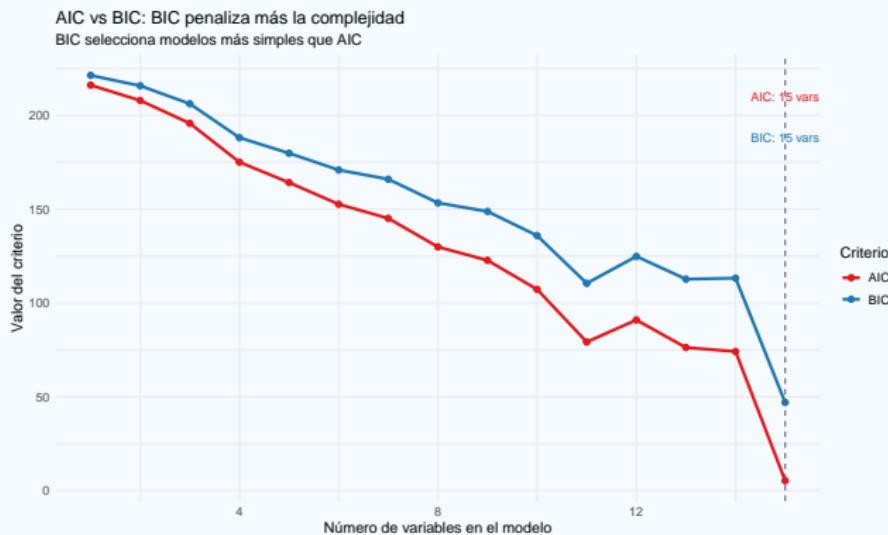
Si el objetivo es la explicación/inferencia:

- **BIC** es la elección más sólida
- Identifica el modelo más parsimonioso
- Penalización más fuerte contra sobreajuste
- Propiedad de consistencia en muestras grandes

Para análisis exploratorio:

- **Cp de Mallows** es especialmente valioso
- Compromiso explícito entre sesgo y varianza
- Visualización clara del “codo” de complejidad óptima

Visualización: Criterios de Información (AIC vs BIC)

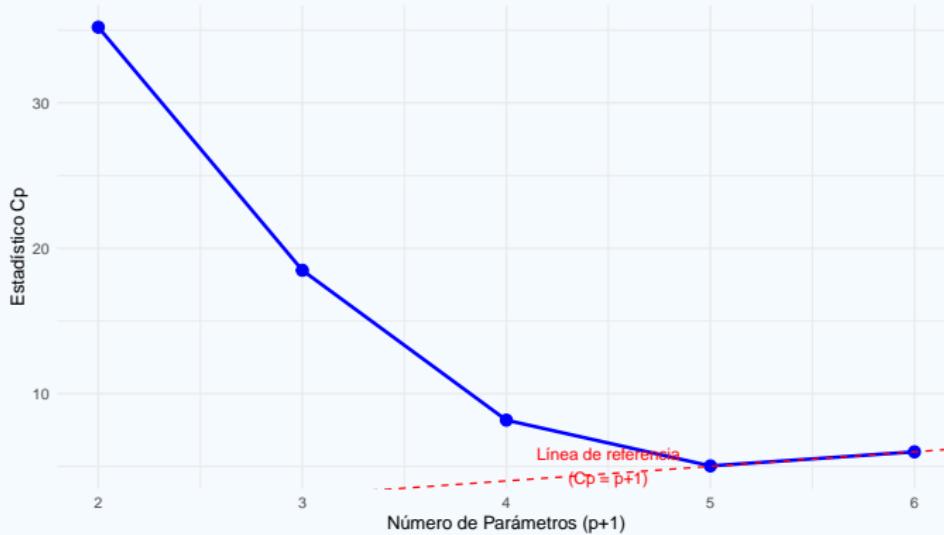


Visualización: Selección con Cp de Mallows



Gráfico Cp de Mallows

Buscamos modelos donde Cp es pequeño y cercano al número de parámetros ($p+1$)





Best Subset Selection: Evalúa **todos** los subconjuntos posibles

Proceso:

- Para $k = 1, 2, \dots, p$ variables
- Construir todos los modelos posibles con k variables
- Seleccionar el mejor modelo de cada tamaño según criterio elegido

Ventajas:

- Garantiza encontrar el modelo óptimo según el criterio
- Evaluación completa de todas las combinaciones
- Estándar para comparar otros métodos

Limitaciones:

- Complejidad computacional: 2^p modelos posibles
- Impracticable para $p > 15 - 20$
- Puede seleccionar modelos sobreajustados sin validación cruzada



Principio: Construir modelo iterativamente, añadiendo o quitando predictores uno a uno

Forward Selection:

- ① Comenzar con modelo nulo (solo intercepto)
- ② Añadir variable que más mejora el criterio
- ③ Repetir hasta que ninguna variable mejore significativamente
- ④ **Problema:** No puede eliminar variables una vez incluidas

Backward Elimination:

- ① Comenzar con modelo completo (todas las variables)
- ② Eliminar variable menos significativa
- ③ Repetir hasta que todas las variables sean significativas
- ④ **Problema:** Requiere $n > p$



Stepwise Regression:

- ① Combina forward + backward
- ② Puede añadir y eliminar variables
- ③ **Problema:** Solo encuentra óptimo local

Limitaciones importantes de métodos automáticos:

- **Inestabilidad:** Pequeños cambios en datos pueden alterar el modelo
- **In validez de p-valores:** Múltiples comparaciones sesgan la inferencia
- **Óptimo local:** No garantizan la mejor combinación
- **Inflación del error tipo I:** Sin corrección para comparaciones múltiples

Uso recomendado: Como herramientas exploratorias, no para inferencia final



Principio: Introducir penalización en la función de ajuste del modelo

Objetivos:

- Controlar el sobreajuste reduciendo complejidad
- Forzar selección de subconjunto más parsimonioso
- Mejorar estabilidad y precisión del modelo

Tres métodos principales:

- **Ridge Regression:** Penalización $L_2 = \lambda \sum \beta_j^2$
- **Lasso:** Penalización $L_1 = \lambda \sum |\beta_j|$
- **Elastic Net:** Combina $L_1 + L_2$

Ventaja clave: Control automático del balance sesgo-varianza



Fundamento: Penalización L_2 en la estimación de coeficientes

Formulación:

$$SSE_{ridge} = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^p \beta_j^2$$

Estimación:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

Interpretación del parámetro λ :

- $\lambda = 0$: equivalente a regresión lineal tradicional (OLS)
- λ aumenta: coeficientes se reducen en magnitud
- λ muy grande: coeficientes se acercan a cero

Propiedades:

- Manejo de multicolinealidad
- Menor varianza en predicciones
- **No realiza selección de variables** (no anula coeficientes)





Fundamento: Penalización L_1 que permite eliminación de variables

Formulación:

$$SSE_{\text{Lasso}} = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^p |\beta_j|$$

Diferencia clave con Ridge:

- Ridge reduce magnitud de coeficientes
- **Lasso puede eliminar variables por completo** (coeficientes = 0)

Interpretación del parámetro λ :

- $\lambda = 0$: regresión lineal tradicional
- λ aumenta: más coeficientes $\rightarrow 0$
- λ muy grande: elimina demasiadas variables

Propiedades:

- Selección automática de variables
- Manejo de multicolinealidad
- Simplicidad e interpretabilidad
- Reduce sobreajuste



Fundamento: Combinación de penalizaciones Ridge (L_2) y Lasso (L_1)

Formulación:

$$SSE_{\text{Elastic Net}} = \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \left[\alpha \sum_{j=1}^p |\beta_j| + (1 - \alpha) \sum_{j=1}^p \beta_j^2 \right]$$

Parámetro α controla la mezcla:

- $\alpha = 1 \rightarrow$ Comportamiento como Lasso
- $\alpha = 0 \rightarrow$ Comportamiento como Ridge
- $0 < \alpha < 1 \rightarrow$ Combinación de ambos métodos

Ventajas principales:

- Manejo superior de multicolinealidad
- Selección de variables más estable
- Evita selección arbitraria cuando hay grupos correlacionados



Ridge Regression:

- Todas las variables aportan información
- Fuerte multicolinealidad presente
- Objetivo: reducir varianza sin eliminar variables

Lasso:

- Muchas variables irrelevantes esperadas
- Selección sparse deseable
- Interpretabilidad prioritaria

Elastic Net:

- Variables correlacionadas en grupos
- Balance entre selección y estabilidad
- Rendimiento predictivo como objetivo principal

Estrategia práctica: Optimizar α mediante validación cruzada junto con λ



El problema: ¿Cómo elegir el valor óptimo de lambda?

La solución: Validación cruzada

Proceso:

- ① Definir secuencia de valores lambda candidatos
- ② Para cada lambda, calcular error de validación cruzada
- ③ Seleccionar lambda que minimiza el error

Dos criterios principales:

- **lambda_min:** Valor que minimiza el error de CV
- **lambda_1SE:** Valor más grande cuyo error está dentro de 1 error estándar del mínimo

Regla 1-SE: Preferir modelo más simple (mayor lambda) si su error es comparable al mínimo



Partición inicial (paso obligatorio):

Antes de cualquier análisis, dividir datos originales en:

- ① **Datos de modelado (80%)**: Para todo el proceso de construcción
- ② **Conjunto de prueba final (20%)**: Guardado para evaluación final

Dentro de los datos de modelado, tres estrategias principales:

- ① **División Train/Test simple**
- ② **Validación cruzada k-fold**
- ③ **Leave-One-Out Cross-Validation (LOOCV)**

Cada estrategia tiene sus ventajas e inconvenientes según el contexto del problema



Concepto: División única de los datos de modelado

Proceso:

- **Conjunto de entrenamiento (70-80%):** Ajustar el modelo
- **Conjunto de test (20-30%):** Evaluar rendimiento

Ventajas:

- Computacionalmente muy eficiente
- Fácil de implementar y entender
- Apropiado para datasets grandes

Desventajas:

- **Alta variabilidad:** Resultados dependen de la división específica
- Puede ser optimista o pesimista según qué observaciones caigan en test
- Menos datos disponibles para entrenamiento

Cuándo usar: Datasets grandes ($n > 1000$), recursos limitados, evaluación rápida



Concepto: Múltiples evaluaciones para obtener estimación más estable

Proceso de k-fold CV:

- ① Dividir datos en k particiones de tamaño similar
- ② Para cada partición $i = 1, 2, \dots, k$:
 - Usar partición i como conjunto de test
 - Usar las $k - 1$ particiones restantes como entrenamiento
 - Calcular métrica de error
- ③ **Error de CV:** $CV_{(k)} = \frac{1}{k} \sum_{i=1}^k \text{Error}_i$

Valores típicos: $k = 5$ o $k = 10$

Ventajas:

- Estimación más estable y menos sesgada
- Todos los datos se usan para entrenamiento y test
- Reduce variabilidad de la estimación

#Esquema del Proceso de Validación Cruzada ($k=5$)



Concepto: Caso extremo donde $k = n$ (número de observaciones)

Proceso:

- Para cada observación i :
 - Entrenar modelo con $n - 1$ observaciones
 - Predecir la observación i excluida
 - Calcular error de predicción
- **Error LOOCV:** $CV_{(n)} = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{1 - h_{ii}} \right)^2$

Ventajas:

- Estimación prácticamente insesgada
- Determinística (no depende de divisiones aleatorias)
- Máximo uso de datos para entrenamiento

Desventajas:

- Computacionalmente costoso
- Alta varianza en la estimación



Train/Test Split:

- Dataset grande ($n > 1000$)
- Recursos computacionales limitados
- Necesidad de evaluación rápida
- Primera aproximación al problema

Validación Cruzada k-fold:

- Dataset de tamaño moderado ($100 < n < 1000$)
- Balance entre eficiencia y precisión
- Estimación robusta del rendimiento
- **Más recomendado en general**

LOOCV:

- Dataset pequeño ($n < 100$)
- Necesidad de estimación menos sesgada
- Recursos computacionales abundantes
- Regresión lineal (fórmula rápida disponible)



Una vez obtenidas las predicciones, necesitamos “calificar” el modelo

Raíz del Error Cuadrático Medio (RMSE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Características del RMSE:

- Penaliza desproporcionadamente errores grandes
- Sensible a valores atípicos
- Mismas unidades que la variable respuesta
- Interpretación: “desviación típica de los residuos”



Error Absoluto Medio:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Características del MAE:

- Trata todos los errores proporcionalmente
- Más robusto frente a valores atípicos
- Interpretación directa: “error promedio en valor absoluto”

¿Cuándo usar cada métrica?

- **RMSE:** Cuando errores grandes son especialmente problemáticos
- **MAE:** Cuando se prefiere robustez frente a valores atípicos
- **Ambas:** Para análisis completo del rendimiento predictivo



La comparación clave: Error en entrenamiento vs Error en validación

Sobreajuste (Overfitting):

- **Síntoma:** Error entrenamiento bajo + Error validación mucho más alto
- **Causa:** Modelo memoriza ruido específico de los datos de entrenamiento
- **Solución:** Simplificar modelo, usar regularización, más datos

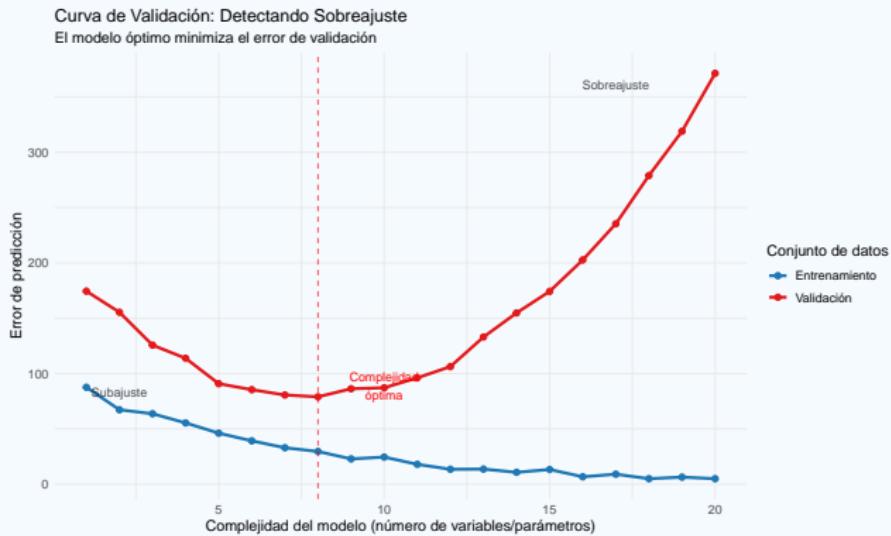
Subajuste (Underfitting):

- **Síntoma:** Error entrenamiento alto + Error validación alto y similar
- **Causa:** Modelo demasiado simple, no captura estructura subyacente
- **Solución:** Aumentar complejidad, añadir variables, términos de interacción

Modelo bien calibrado:

- **Síntoma:** Error entrenamiento y validación similares y bajos
- **Interpretación:** Buen equilibrio entre sesgo y varianza

Error de Entrenamiento vs Validación





Después de todo el proceso de modelado:

- ① Filtrado de variables
- ② Selección del mejor método
- ③ Optimización de hiperparámetros
- ④ Validación cruzada para elegir modelo final

El paso final: Evaluar el modelo seleccionado en el conjunto de prueba final

¿Por qué es necesario?

- La validación cruzada se usó para **tomar decisiones** sobre el modelo
- Existe riesgo de sobreajuste al proceso de validación mismo
- Necesitamos una evaluación completamente independiente

Interpretación:

- Error similar a validación cruzada → modelo robusto
- Error mucho mayor → posible sobreajuste al proceso de modelado



Los métodos stepwise (forward, backward, stepwise) requieren precaución especial

Problemas fundamentales:

- ① **In validez de p-valores:** Los p-valores y errores estándar están sesgados
- ② **Inestabilidad:** Pequeños cambios en datos pueden cambiar radicalmente el modelo
- ③ **Óptimo local:** No garantizan encontrar la mejor combinación de variables
- ④ **Inflación del error tipo I:** Múltiples comparaciones sin corrección

Uso recomendado:

- Como herramientas **exploratorias** únicamente
- Generar modelos candidatos para evaluación posterior
- Siempre validar con técnicas robustas
- No reportar p-valores del modelo final como definitivos



Flujo de trabajo recomendado:

- ① **Partición inicial:** Separar conjunto de prueba final (20%)
- ② **En datos de modelado (80%):**
 - Filtrado básico de variables
 - Aplicar métodos de selección (exhaustivos, stepwise, regularización)
 - Comparar modelos con validación cruzada
 - Seleccionar modelo final
- ③ **Evaluación final:** Probar modelo seleccionado en conjunto de prueba
- ④ **Reportar:** Error de validación cruzada Y error en conjunto de prueba

Criterios de decisión:

- Número de variables vs tamaño de muestra → método de selección
- Objetivo (predicción vs explicación) → criterio de información
- Multicolinealidad → regularización vs selección clásica



Antes del modelado:

- EDA completo para entender los datos
- Conocimiento del dominio para variables importantes
- Objetivo claro: ¿predicción o explicación?
- Relación entre tamaño muestral y número de variables

Durante la selección:

- Usar validación cruzada para todos los hiperparámetros
- Comparar múltiples métodos de selección
- No guiarse solo por métricas: considerar interpretabilidad
- Documentar todas las decisiones tomadas

Después de la selección:

- Diagnóstico completo de residuos del modelo final
- Análisis de sensibilidad a observaciones influyentes
- Intervalos de confianza para coeficientes importantes
- Validación en el conjunto de prueba final

Lo aprendido en este tema:

- ① **Filtrado inicial:** Elimina problemas básicos de forma eficiente
- ② **Criterios de información:** Guían comparación objetiva de modelos
- ③ **Métodos exhaustivos:** Garantizan óptimo pero son computacionalmente costosos
- ④ **Regularización:** Controla sobreajuste y realiza selección automáticamente
- ⑤ **Validación:** Indispensable para evaluar capacidad de generalización

Recomendaciones principales:

- **Combinar métodos:** Ningún método es perfecto en todas las situaciones
- **Validar siempre:** Con datos que el modelo no ha visto
- **Preferir simplicidad:** Cuando el rendimiento es comparable
- **Incorporar conocimiento del dominio:** Los datos no lo dicen todo

El mejor modelo es aquel que resuelve el problema con la mayor simplicidad.

Modelos Lineales Generalizados (GLM)

Víctor Aceña - Isaac Martín

DSLAB

2025-09-04





Punto de Partida: La Regresión Lineal

- Es una herramienta potente para modelar una variable dependiente **continua**.
- Sin embargo, sus supuestos (normalidad, homocedasticidad) no siempre se cumplen.

El Desafío: Datos No Normales

- ¿Cómo modelamos una variable de respuesta **binaria** (ej. enfermo/sano)?
- ¿O datos de **conteo** (ej. nº de accidentes en una intersección)?

La Solución: Modelos Lineales Generalizados (GLM)

- Son una **extensión** de la regresión lineal que permite modelar respuestas con distribuciones como la Binomial o la de Poisson, utilizando **funciones de enlace** para mayor flexibilidad.



Los **Modelos Lineales Generalizados (GLM)** son una extensión de los modelos de regresión lineal que permiten manejar una mayor variedad de tipos de datos y relaciones entre variables.

Mientras que la regresión lineal clásica asume que la variable dependiente (Y) es continua y sigue una distribución Normal, los GLM permiten que Y sea:

- **Binaria:** Éxito/Fracaso, Sí/No (ej. Regresión Logística).
- **De Conteo:** N° de eventos (ej. Regresión de Poisson).
- **Continua y Positiva** con sesgo (ej. tiempos, costos).

El marco teórico unificador de los GLM es que la distribución de la variable dependiente siempre pertenece a la **familia exponencial**.



Todo Modelo Lineal Generalizado se define por la interacción de tres componentes clave:

① Componente Aleatorio:

- Qué es: La **distribución de probabilidad** que se asume para la variable dependiente (Y).
- Proviene de la **familia exponencial**.

② Componente Sistemático:

- Qué es: La **combinación lineal** de las variables predictoras (X), que forma el **predictor lineal** (η).

③ Función de Enlace (*Link Function*):

- Qué es: El **ponte matemático** que conecta el predictor lineal (η) con la media de la variable respuesta (μ).



1. Componente Aleatorio (La Distribución)

Define el tipo de datos que estamos modelando. A diferencia de la regresión lineal (solo Normal), en los GLM podemos usar otras distribuciones:

- **Distribución Binomial:** Para variables categóricas binarias (0/1, éxito/fracaso).
- **Distribución de Poisson:** Para datos de conteo (número de eventos).
- **Distribución Gamma:** Para datos continuos y positivos (como costos o tiempos).

2. Componente Sistemático (El Predictor Lineal)

Describe cómo las variables independientes se combinan linealmente. Su forma es idéntica a la de la regresión lineal:

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Donde η es el predictor lineal y β son los coeficientes del modelo.





La función de enlace (g) conecta el predictor lineal (η) con la media de la variable dependiente (μ). Es la clave de la flexibilidad del modelo.

Relación Fundamental:

$$g(\mu) = \eta$$

Esta función transforma la media de Y para que la relación con los predictores se vuelva lineal.

Ejemplos Matemáticos: * **Logística (Logit):** Para Regresión Logística.

$$g(\mu) = \log\left(\frac{\mu}{1 - \mu}\right)$$

* **Logarítmica:** Para Regresión de Poisson.

$$g(\mu) = \log(\mu)$$

* **Identidad:** Para Regresión Lineal estándar (el GLM más simple).

$$g(\mu) = \mu$$



Regresión Lineal Clásica

- **Distribución:** Normal.
- **Tipo de Respuesta:** Continua.
- **Relación:** Lineal y directa.
- **Función de Enlace:** Identidad.

Modelos Lineales Generalizados

- **Distribución:** Familia Exponencial.
- **Tipo de Respuesta:** Flexible (binaria, conteo...).
- **Relación:** Transformada por una función de enlace.
- **Función de Enlace:** Flexible (Logit, Log...).

Ventajas Principales de los GLM:

- **Flexibilidad:** Permiten modelar muchos más tipos de variables dependientes.
- **Interpretación Coherente:** Los coeficientes siguen siendo interpretables de forma rigurosa.
- **Evaluación Robusta:** Se pueden usar las mismas herramientas de evaluación (AIC, BIC, tests de hipótesis).



La estimación en GLM representa un cambio fundamental respecto a la regresión lineal clásica.

- **Regresión Lineal:** Utiliza **Mínimos Cuadrados Ordinarios (MCO)**, un método que funciona bajo supuestos de normalidad y homocedasticidad.
- **GLM:** Necesita métodos más sofisticados debido a las distribuciones no normales y las funciones de enlace no lineales.

El enfoque para los GLM se basa en un principio unificador:

- **El Principio:** La **Máxima Verosimilitud (MLE)**, que proporciona un marco teórico coherente para toda la familia exponencial.
- **El Algoritmo:** **IRLS** (*Iteratively Reweighted Least Squares*), el método computacional para encontrar la solución de máxima verosimilitud.



A diferencia de Mínimos Cuadrados, el método de **Máxima Verosimilitud** se emplea para estimar los parámetros en un GLM.

¿Por qué es necesario?

- Las distribuciones de la familia exponencial no siempre tienen una relación lineal directa con los predictores.
- La varianza de la respuesta a menudo depende de su media ($Var(Y) = V(\mu)$), violando el supuesto de homocedasticidad que requiere MCO.

Principio Fundamental de MLE: Consiste en encontrar los valores de los parámetros β que hacen **más probable** observar los datos que tenemos.

De acuerdo, tienes razón. Es mejor separar las dos funciones para explicarlas con más claridad.

Aquí tienes el contenido dividido en dos diapositivas.



Para aplicar el principio de Máxima Verosimilitud, primero definimos la **función de verosimilitud**.

- **Definición:** Es la probabilidad conjunta de observar la totalidad de nuestra muestra (y_1, \dots, y_n) dado un conjunto de parámetros β .
- **Fórmula:**

$$L(\beta) = \prod_{i=1}^n f(y_i; \theta_i, \phi)$$

donde $f(y_i; \theta_i, \phi)$ es la función de densidad o masa de probabilidad de cada observación.

- **El Problema Práctico:** Maximizar una función que es un producto de muchos términos es computacionalmente complejo y puede ser numéricamente inestable.



Para solucionar el problema de los productos, en la práctica se trabaja con el **logaritmo** de la verosimilitud.

- **Definición:** La función de log-verosimilitud es simplemente el logaritmo de la función de verosimilitud.

- **Fórmula:**

$$\ell(\beta) = \log(L(\beta)) = \sum_{i=1}^n \log f(y_i; \theta_i, \phi)$$

- **Ventaja Clave:** El logaritmo convierte los **productos en sumas**, lo que simplifica enormemente los cálculos matemáticos y numéricos necesarios para encontrar el máximo de la función.



La clave de los GLM es que todas sus distribuciones (Binomial, Poisson, Gamma...) pertenecen a la **familia exponencial**.

Esto significa que todas se pueden escribir con una **forma matemática unificada**:

$$f(y; \theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

donde θ es el **parámetro natural** y ϕ es el **parámetro de dispersión**.

Propiedades Derivadas de esta Forma: Esta estructura unificada permite derivar propiedades generales para todos los GLM de forma elegante:

- **Esperanza (Media):** $E(Y) = \mu = b'(\theta)$.
- **Varianza:** $\text{Var}(Y) = a(\phi)b''(\theta) = a(\phi)V(\mu)$, donde $V(\mu)$ es la **función de varianza** que caracteriza la relación media-varianza de cada distribución.



Estas son las distribuciones más frecuentes en la práctica.

- **Normal**

- **Uso Típico:** Datos continuos simétricos (es el GLM equivalente a la regresión lineal).
- **Función de Varianza:** $V(\mu) = 1$ (varianza constante).
- **Enlace Canónico:** Identidad ($g(\mu) = \mu$).

- **Binomial**

- **Uso Típico:** Proporciones, datos binarios (éxito/fracaso).
- **Función de Varianza:** $V(\mu) = \mu(1 - \mu)$.
- **Enlace Canónico:** Logit ($g(\mu) = \log(\frac{\mu}{1-\mu})$).

- **Poisson**

- **Uso Típico:** Conteos de eventos.
- **Función de Varianza:** $V(\mu) = \mu$.
- **Enlace Canónico:** Log ($g(\mu) = \log(\mu)$).



Para datos continuos que son estrictamente positivos y tienen sesgo a la derecha.

- **Gamma**

- **Uso Típico:** Tiempos, costos, o cualquier dato continuo positivo y asimétrico.
- **Función de Varianza:** $V(\mu) = \mu^2$.
- **Enlace Canónico:** Inverso ($g(\mu) = 1/\mu$).

- **Inversa Gaussiana**

- **Uso Típico:** Tiempos hasta un evento, o datos con una asimetría aún más pronunciada que la Gamma.
- **Función de Varianza:** $V(\mu) = \mu^3$.
- **Enlace Canónico:** Inverso al cuadrado ($g(\mu) = 1/\mu^2$).



Para entender y comparar los diferentes GLM, dos conceptos derivados de la familia exponencial son fundamentales:

① La Función de Varianza: $V(\mu)$

- **Definición:** Es la “firma” de cada distribución, ya que define la relación teórica entre la media (μ) y la varianza.
- **Implicación Práctica:** Determina la **heterocedasticidad inherente** de los datos (ej. $V(\mu) = \mu$ en Poisson) y, por tanto, influye directamente en los **pesos** que el algoritmo IRLS asigna a cada observación durante la estimación.

② El Enlace Canónico: $g(\mu)$

- **Definición:** Es la función de enlace que surge de forma “natural” de la estructura matemática de cada distribución.
- **Implicación Práctica:** Aunque en la práctica se pueden probar otros enlaces, el canónico suele garantizar las mejores propiedades estadísticas y una estimación computacionalmente más eficiente.



La relación entre media y varianza es fundamental en los GLM. La función de varianza $V(\mu)$ determina la **heterocedasticidad inherente** de cada distribución e influye en la estimación del modelo.

- **Distribución Binomial:** $V(\mu) = \mu(1 - \mu)$. La varianza es máxima cuando la probabilidad $\mu = 0.5$.
- **Distribución de Poisson:** $V(\mu) = \mu$. La varianza aumenta linealmente con la media.
- **Distribución Gamma:** $V(\mu) = \mu^2$. La varianza aumenta cuadráticamente con la media.

Implicación Práctica: Esta función influye directamente en los pesos del algoritmo IRLS. En regresión logística, por ejemplo, las observaciones con probabilidades cercanas a 0.5 tienen mayor varianza y, por tanto, reciben **menor peso** en la estimación.



Las ecuaciones de máxima verosimilitud de los GLM no tienen una solución matemática directa como en la regresión lineal. Por ello, se necesita un **algoritmo iterativo** para encontrar los coeficientes.

El método estándar es **IRLS** (*Iteratively Reweighted Least Squares*), que es una aplicación del método de Newton-Raphson.

¿Cómo funciona conceptualmente? El algoritmo aproxima el problema no lineal del GLM a una **serie de regresiones lineales ponderadas** que se resuelven de forma sucesiva.

- ① Se empieza con una estimación inicial de los coeficientes β .
- ② En cada paso, se calculan unos **pesos** (w_i) para cada observación. Estos pesos reflejan la “fiabilidad” de cada punto según el modelo actual.
- ③ Se resuelve una **regresión por mínimos cuadrados ponderada** para obtener una nueva y mejor estimación de β .
- ④ Se repite el proceso hasta que las estimaciones de β se estabilizan (convergen).



Los estimadores MLE poseen propiedades **asintóticas** (se cumplen cuando $n \rightarrow \infty$) muy deseables, que los validan como el método de estimación preferido.

1. Consistencia: A medida que aumenta el tamaño de la muestra, los estimadores convergen al valor verdadero del parámetro.

$$\hat{\beta} \xrightarrow{p} \beta \text{ cuando } n \rightarrow \infty.$$

2. Normalidad Asintótica: Para muestras grandes, la distribución de los estimadores se aproxima a una Normal multivariada.

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}^{-1}(\beta))$$

Permite construir **intervalos de confianza** y realizar **tests de hipótesis**.

3. Eficiencia: Los estimadores MLE alcanzan la cota de Cramér-Rao, lo que significa que tienen la **menor varianza asintótica posible** entre todos los estimadores insesgados.



La **matriz de información** ($I(\beta)$) es un concepto clave que cuantifica la certeza de las estimaciones de nuestros coeficientes.

Interpretación Intuitiva:

- Mide la “**curvatura**” de la función de verosimilitud en su punto máximo.
- Una función muy “puntiaguda” (alta curvatura) significa **muchísima información** y, por tanto, estimaciones más precisas y fiables.
- Una función más plana (baja curvatura) significa **poca información** y mayor incertidumbre.

Cálculo en la Práctica (GLM):

Aunque existen definiciones teóricas basadas en las segundas derivadas de la log-verosimilitud, el algoritmo IRLS nos proporciona una aproximación computacionalmente eficiente y directa:

$$I(\hat{\beta}) \approx \mathbf{X}^T \mathbf{W} \mathbf{X}$$



Los **errores estándar** de los coeficientes individuales se obtienen a partir de la matriz de información.

Proceso de Cálculo:

- ① **Matriz de Covarianza:** Se calcula como la **inversa** de la matriz de información.
- ② **Errores Estándar:** Se obtienen como las raíces cuadradas de los elementos diagonales de esa matriz de covarianza.

Fórmula:

$$SE(\hat{\beta}_j) = \sqrt{[\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})]_{jj}}$$



Estos errores estándar son fundamentales para la inferencia estadística.

Aplicaciones Principales:

- **Intervalos de Confianza:**

$$\hat{\beta}_j \pm z_{\alpha/2} \cdot \text{SE}(\hat{\beta}_j)$$

- **Estadísticos de Prueba (de Wald):**

$$z_j = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)}$$

- **Evaluación de la Precisión** de nuestras estimaciones.

Advertencia Importante:

- Estos errores estándar **son válidos bajo los supuestos del modelo GLM**.
- Violaciones serias de estos supuestos (como **sobredispersión** en modelos de Poisson) pueden hacer que sean inadecuados.



La **deviance** es la medida principal de ajuste en GLM; es una generalización de la suma de cuadrados residuales.

Idea Fundamental: Mide la discrepancia entre la verosimilitud de nuestro **modelo propuesto** y la de un **modelo saturado** (un modelo teóricamente perfecto que ajusta cada dato).

$$D = 2 \sum_{i=1}^n [\ell(\text{modelo saturado}) - \ell(\text{modelo propuesto})]$$

Interpretación práctica:

- Deviance = 0: Modelo perfecto que ajusta exactamente todos los datos observados
- Deviance baja: Buen ajuste del modelo a los datos
- Deviance alta: Mal ajuste del modelo, sugiere que el modelo no captura adecuadamente los patrones en los datos



La clave para interpretar la deviance es **comparar** la de tu modelo con la de un modelo base.

① El Punto de Partida: La Deviance Nula

- **¿Qué es?** La deviance de un modelo simple, solo con intercepto (que ignora todos tus predictores).
- **Analogía:** Es el “**error máximo**” que sirve como punto de comparación.

② El Resultado: La Deviance Residual

- **¿Qué es?** La deviance de tu modelo final, con todos los predictores incluidos.
- **Analogía:** Es el “**error restante**” después de que tus predictores han hecho su trabajo.

La **reducción de la deviance** (Deviance Nula – Deviance Residual) representa la **mejora en el ajuste** que se debe a tus predictores.





Es la herramienta principal para **comparar modelos anidados**, basándose en el principio de parsimonia.

Estadístico de Prueba: La diferencia en las deviances sigue una distribución **chi-cuadrado**:

$$LRT = D_{\text{reducido}} - D_{\text{completo}} \sim \chi^2_{df}$$

(donde df es la diferencia en el número de parámetros).

Regla de Decisión:

- H_0 : El modelo reducido es suficiente.
- Si el **p-valor** $< \alpha$, se rechaza H_0 . La mejora del modelo completo es **estadísticamente significativa**.



Objetivo: Determinar si añadir la variable `disp` (cilindrada) a un modelo que ya contiene `wt` (peso) mejora significativamente la predicción.

Salida del Test en R

Analysis of Deviance Table

Model 1: `am ~ wt`

Model 2: `am ~ wt + disp`

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	30	19.176			
2	29	17.785	1	1.3913	0.2382



Interpretación del Resultado

El test compara la **Deviance** de ambos modelos. La hipótesis nula (H_0) es que el modelo reducido es suficiente.

- La diferencia en deviance es de **1.391** con **1** grado de libertad.
- El p-valor asociado es **0.238**.

Dado que el p-valor es mayor que 0.05, **no rechazamos la hipótesis nula**.

Conclusión: Añadir disp no aporta una mejora significativa. Nos quedamos con el **modelo reducido** por parsimonia.



La diagnosis de GLMs es el proceso de evaluar los supuestos del modelo y detectar problemas que puedan afectar la validez de las inferencias.

¿Por qué es diferente de la Regresión Lineal?

- A diferencia de la regresión lineal, los residuos ordinarios no son suficientes.
- Los GLM requieren herramientas especializadas debido a la **heterocedasticidad inherente** y a las diferentes distribuciones subyacentes (Poisson, Binomial, etc.).

Nuestro Enfoque: Abordaremos el diagnóstico respondiendo a tres preguntas clave, utilizando diferentes tipos de **residuos** para obtener las respuestas.



Los residuos “crudos” ($y_i - \hat{\mu}_i$) no son homocedásticos, por lo que se utilizan versiones estandarizadas. Los más importantes son:

- **Residuos Pearson:**

- Son un análogo directo a los residuos estandarizados en regresión lineal. Estandarizan el residuo crudo dividiendo por la desviación estándar predicha por el modelo.

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

- **Residuos Deviance:**

- Son los **más recomendados** para la inspección visual en gráficos diagnósticos.
- Su distribución se aproxima mejor a la normalidad y su varianza es más estable que la de otros residuos.

$$d_i = \text{sign}(y_i - \hat{\mu}_i) \sqrt{2[l_i(y_i) - l_i(\hat{\mu}_i)]}$$



Se evalúa si la estructura básica del modelo, $g(\mu) = X\beta$, es adecuada para los datos.

Herramientas de Diagnóstico:

- **Gráfico de Residuos vs. Valores Ajustados:**
 - Es la herramienta fundamental. Se grafican los residuos (idealmente, deviance) contra el predictor lineal ($\hat{\eta}_i$).
 - **Un patrón curvilíneo** es una señal clara de que la forma funcional o la función de enlace son incorrectas.
- **Gráficos de Residuos Parciales:** Evalúan si la relación es apropiada para **cada predictor individualmente**.
- **Test de Especificación de Enlace (Linktest):** Es una prueba formal. Si el predictor lineal al cuadrado ($\hat{\eta}^2$) resulta significativo al añadirlo al modelo, es evidencia de que la función de enlace está mal especificada.



Se evalúa si la elección de la familia de distribución (Poisson, Binomial, etc.) fue acertada, empezando por la relación entre la media y la varianza.

Sobredispersión: Ocurre cuando la varianza real de los datos es **mayor** que la media, violando el supuesto de modelos como el de Poisson ($Var(Y) = \mu$).

- **Consecuencia:** Invalida las inferencias (errores estándar demasiado pequeños, p-valores incorrectos).
- **Detección:** Se calcula el **Estadístico de dispersión ($\hat{\phi}$)**. Si $\hat{\phi}$ es **significativamente mayor que 1**, hay sobredispersión.

$$\hat{\phi} = \frac{\sum r_i^2}{n - p}$$

- **Solución:** Cambiar a un modelo más flexible. El caso clásico es pasar de **Poisson** a **Binomial Negativo**, ya que este último incluye un parámetro (α) para modelar la variabilidad extra: $Var(Y) = \mu + \alpha\mu^2$.



Un segundo aspecto para verificar la distribución es la forma general de los errores.

Herramienta: El Gráfico Q-Q de residuos deviance.

- **Concepto:** Aunque los errores de un GLM no son estrictamente normales, los **residuos deviance** sí deberían tener una distribución aproximadamente normal si el modelo está bien especificado.
- **Interpretación:**
 - Se grafican los cuantiles de los residuos deviance contra los cuantiles teóricos de una distribución normal.
 - Los puntos deberían seguir de cerca la línea diagonal.
 - **Desviaciones sistemáticas** de la línea pueden indicar que la familia de distribución asumida (Poisson, Binomial, etc.) es incorrecta.



Finalmente, buscamos identificar puntos individuales que tienen una influencia desproporcionada en los coeficientes del modelo.

Herramientas Matemáticas Clave:

- **Leverage Generalizado (h_i):** Mide el potencial de una observación para ser influyente debido a su posición en el espacio de los predictores.
- **Distancia de Cook para GLMs (D_i):** Mide la influencia global de una observación en *todos* los coeficientes.

$$D_i = \frac{r_i^2 h_i}{p(1 - h_i)^2}$$

- **DFBETAS:** Mide la influencia de una observación en *cada coeficiente individual*.

Estrategia:

- La herramienta visual principal es el **gráfico de residuos vs. leverage**.
- La estrategia ante estas observaciones no es eliminarlas automáticamente, sino **investigarlas** para entender su naturaleza.





La **regresión logística** es la herramienta fundamental de los GLM para modelar la probabilidad de ocurrencia de un **evento binario**.

Objetivo:

Modelar una variable dependiente que solo toma dos valores:

- Éxito / Fracaso
- Sí / No
- Enfermo / Sano

Pilares del Tema:

- ① **Fundamentos:** La función sigmoide y el enlace Logit.
- ② **Estimación:** Máxima Verosimilitud (MLE) y el algoritmo IRLS.
- ③ **Interpretación:** El concepto clave de los **Odds Ratios**.
- ④ **Evaluación:** Bondad de ajuste (Deviance, Pseudo R²) y validación (Matriz de Confusión, Curva ROC).



El Problema: La regresión lineal puede predecir valores fuera del rango [0,1], lo cual no tiene sentido para modelar una probabilidad.

La Solución: La función logística (o sísmoide) transforma cualquier valor real del predictor lineal (η) en una probabilidad entre 0 y 1.

Fórmula de la Probabilidad:

$$P(Y = 1|X) = \frac{1}{1 + e^{-\eta}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots)}}$$

La curva en forma de "S" asegura que las predicciones se aplanen hacia 0 y 1 en los extremos.



Para poder usar un modelo lineal ($X\beta$), necesitamos transformar la probabilidad (que está en la escala [0,1]) a una escala que vaya de $-\infty$ a $+\infty$.

La Herramienta: Esto se logra con la **función de enlace logit**.

Definición del Logit: El logit de una probabilidad p es el logaritmo de los *odds* (razón de probabilidades):

$$\text{logit}(p) = \log \left(\frac{p}{1-p} \right)$$

El Modelo Logístico Linealizado: Esta transformación nos permite expresar el modelo de forma lineal:

$$\log \left(\frac{p}{1-p} \right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$



La estimación de los coeficientes β se basa en MLE para la distribución Binomial.

Función de Verosimilitud ($L(\beta)$): Para un resultado binario $y_i \in \{0, 1\}$, la verosimilitud se define como:

$$L(\beta) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}$$

Función de Log-Verosimilitud ($\ell(\beta)$): Maximizamos el logaritmo de la función anterior:

$$\ell(\beta) = \sum_{i=1}^n \left[y_i \mathbf{x}_i^T \beta - \log(1 + e^{\mathbf{x}_i^T \beta}) \right]$$

Ecuaciones de Puntuación (Score Equations): Para encontrar el máximo, se deriva la log-verosimilitud respecto a cada β_j y se iguala a cero:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^n x_{ij} (y_i - p_i) = 0$$



Como las ecuaciones de verosimilitud no tienen solución analítica cerrada, se utiliza el algoritmo IRLS. Para la regresión logística, los componentes específicos son:

Pesos (w_i): El peso de cada observación es la varianza de una distribución Bernoulli, que es máxima cuando la probabilidad es 0.5.

$$w_i = p_i(1 - p_i)$$

Variable Dependiente Ajustada (z_i): Es la versión linealizada de la respuesta en cada iteración.

$$z_i = \eta_i^{(t)} + \frac{y_i - p_i^{(t)}}{w_i^{(t)}}$$

Estos componentes se utilizan en cada paso de la actualización de los coeficientes: $\beta^{(t+1)} = (\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(t)} \mathbf{z}^{(t)}$.



Podemos verificar que los coeficientes encontrados por `glm()` son, en efecto, los que maximizan la función de log-verosimilitud. A continuación se muestra la salida del código R:

Salida del Análisis:

Coeficientes estimados por `glm()`:

(Intercept)	glu	bmi
-8.16560411	0.03433555	0.08585474

Log-verosimilitud en el óptimo: -101.4057

Iteraciones necesarias: 4

¿Convergió? TRUE

Conclusión: La salida confirma que el algoritmo convergió en 4 iteraciones para encontrar los coeficientes que maximizan la verosimilitud del modelo.



Los coeficientes β están en la escala del logit, por lo que no son directamente interpretables en términos de probabilidad. Para interpretarlos, primero necesitamos entender el concepto de **odds**.

Definición de Odds:

- El **odds** es la razón entre la probabilidad de que un evento ocurra y la de que no ocurra.

$$\text{odds} = \frac{p}{1 - p}$$

- El modelo logístico es, en esencia, un modelo lineal para el **logaritmo de los odds**: $\log(\text{odds}) = X\beta$.

Ejemplo:

- Si la probabilidad de éxito $p = 0.8$.
- El odds es $\frac{0.8}{0.2} = 4$.
- **Interpretación:** El evento es **4 veces más probable** que ocurra a que no ocurra.



El **Odds Ratio (OR)** es la herramienta principal para interpretar los coeficientes de una regresión logística.

- **Concepto:** Mide el cambio **multiplicativo** en los *odds* por cada incremento de una unidad en un predictor X_j , manteniendo el resto de variables constantes.
- **Cálculo:** Se obtiene exponenciando el coeficiente:

$$\text{OR} = e^{\beta_j}$$

- **Interpretación:**
 - **OR > 1:** El odds aumenta (el evento se vuelve más probable).
 - **OR < 1:** El odds disminuye (el evento se vuelve menos probable).
 - **OR = 1:** No hay efecto.

Ejemplo: Si $\beta_{BMI} = 0.08$, entonces $OR = e^{0.08} \approx 1.083$. Por cada unidad que aumenta el BMI, el odds de tener diabetes se multiplica por 1.083 (es decir, aumenta un **8.3%**).



El R^2 tradicional no es aplicable en este contexto. La bondad de ajuste en regresión logística se evalúa con métricas basadas en la verosimilitud.

1. Deviance

- Compara la log-verosimilitud de nuestro modelo con la de un modelo saturado. La fórmula específica para la distribución binomial es:

$$D = 2 \sum_{i=1}^n \left[y_i \log \left(\frac{y_i}{\hat{p}_i} \right) + (1 - y_i) \log \left(\frac{1 - y_i}{1 - \hat{p}_i} \right) \right]$$

2. Pseudo R²

- Son análogos al R^2 que miden la mejora en la verosimilitud del modelo en comparación con un modelo nulo (solo con intercepto).
- No representan la “proporción de varianza explicada”, sino la mejora en el ajuste del modelo.



Existen varias formulaciones para el Pseudo R². Las más comunes son:

- **McFadden's R²:** Es el más utilizado.

$$R_{\text{McFadden}}^2 = 1 - \frac{\ell_{\text{modelo}}}{\ell_{\text{nulo}}}$$

- Valores entre 0.2 y 0.4 se consideran indicativos de un buen ajuste.
- **Cox-Snell R²:**

$$R_{\text{Cox-Snell}}^2 = 1 - \left(\frac{L_{\text{nulo}}}{L_{\text{modelo}}} \right)^{2/n}$$

- **Nagelkerke R²:** Es una corrección del Cox-Snell para que su valor máximo sea 1, haciéndolo más comparable al R² tradicional.

$$R_{\text{Nagelkerke}}^2 = \frac{R_{\text{Cox-Snell}}^2}{1 - (L_{\text{nulo}})^{2/n}}$$



La validación de un modelo logístico se centra en su **capacidad de clasificación**.

La Matriz de Confusión

- Es la herramienta fundamental. Compara las clases predichas por el modelo con las clases reales.
- **Proceso:** Se convierten las probabilidades predichas (\hat{p}_i) en clases (“Sí” / “No”) usando un **umbral de decisión** (típicamente 0.5).

Esto genera cuatro posibles resultados:

- **Verdaderos Positivos (VP):** Predijo “Sí” y era “Sí”.
- **Falsos Positivos (FP):** Predijo “Sí” pero era “No” (Error Tipo I).
- **Verdaderos Negativos (VN):** Predijo “No” y era “No”.
- **Falsos Negativos (FN):** Predijo “No” pero era “Sí” (Error Tipo II).



A partir de la matriz de confusión, se calculan las métricas de rendimiento clave:

- **Precisión (Accuracy):**

- $\frac{VP+VN}{Total}$
- Proporción total de predicciones correctas. Cuidado: puede ser engañosa en datasets desbalanceados.

- **Sensibilidad (Recall o Tasa de VP):**

- $\frac{VP}{VP+FN}$
- De todos los positivos reales, ¿qué proporción clasificamos correctamente? Mide la capacidad para identificar los casos positivos.

- **Especificidad:**

- $\frac{VN}{VN+FP}$
- De todos los negativos reales, ¿qué proporción clasificamos correctamente? Mide la capacidad para identificar los casos negativos.



La Curva ROC (*Receiver Operating Characteristic*)

- Es una evaluación global del rendimiento del modelo, **independiente del umbral de decisión**.
- Grafica la **Sensibilidad** (Tasa de VP) en el eje Y frente a **1 - Especificidad** (Tasa de FP) en el eje X para todos los umbrales posibles.

AUC (Área Bajo la Curva ROC)

- Cuantifica la capacidad discriminativa del modelo en un solo número (de 0.5 a 1.0).
 - **AUC = 1.0:** Clasificador perfecto.
 - **AUC = 0.5:** Clasificador inútil (equivalente al azar).
 - **Típicamente, AUC > 0.8** se considera una buena discriminación.

Sí, es una idea excelente. Integrar los ejemplos prácticos es fundamental para conectar la teoría con la aplicación en R.

Ajustamos un modelo para predecir la diabetes en el dataset Pima.tr. Los resultados clave de la validación del modelo son los siguientes:

Resultados Numéricos

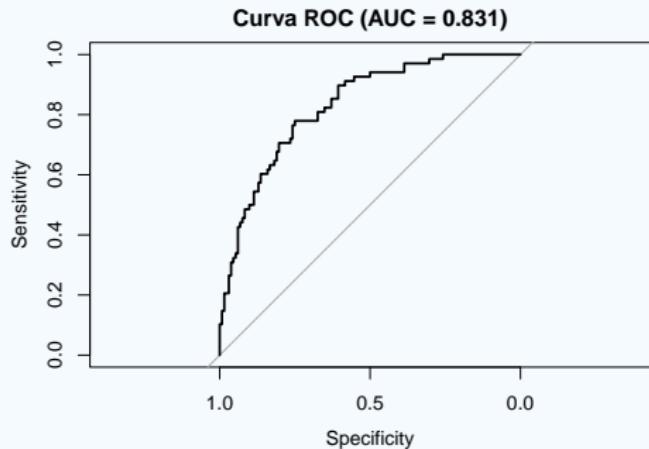
Matriz de Confusión:

		Actual	
		Predicted	No Yes
Predicted	No	114	29
	Yes	18	39

Métricas Clave:

- **Exactitud:** 0.785
- **AUC:** 0.831

Evaluación Visual: Curva ROC





Es la técnica de GLM utilizada para modelar **datos de conteo**: una variable que representa el número de veces que ocurre un evento en un intervalo.

- **Tipo de Variable:** La respuesta toma valores enteros no negativos ($0, 1, 2, \dots$) y se asume que sigue una **distribución de Poisson**.
- **Función de Probabilidad de Poisson:**

$$P(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!}$$

donde λ es la **tasa media de ocurrencia** del evento.



El objetivo del modelo es explicar la relación entre la **tasa de ocurrencia de los eventos** (λ) y un conjunto de variables predictoras X .

- **Forma Funcional:** Se utiliza una **función de enlace logarítmica** para asegurar que la tasa λ sea siempre positiva.

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- **Tasa Esperada:** El modelo puede expresarse en términos de la tasa esperada de eventos como:

$$\lambda = e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}$$



Para que el modelo sea adecuado, se deben cumplir ciertos supuestos:

- **Independencia de los eventos.**
- **Equidispersión:** El supuesto fundamental de la distribución de Poisson es que la **media es igual a la varianza**:

$$E(Y) = \text{Var}(Y) = \lambda$$

Limitaciones Comunes (Violación de Supuestos):

- **Sobredispersión:** Ocurre cuando la varianza es mayor que la media ($\text{Var}(Y) > E(Y)$). La solución es usar una **Regresión Binomial Negativa**.
- **Exceso de Ceros:** Si hay más ceros en los datos de los que predice el modelo. La solución es usar modelos **ZIP** (*Zero-Inflated Poisson*).



Los coeficientes β están en la escala logarítmica de la tasa, por lo que para una interpretación práctica, los exponenciamos.

- **Incidence Rate Ratio (IRR):**

$$\text{IRR} = e^{\beta_j}$$

- **Interpretación:** El IRR es un **factor multiplicativo** que nos dice cuánto cambia la tasa de eventos esperada por cada incremento de una unidad en el predictor X_j .
 - **IRR > 1:** La tasa de eventos aumenta. Un IRR de 1.25 es un aumento del 25%.
 - **IRR < 1:** La tasa de eventos disminuye. Un IRR de 0.80 es una disminución del 20%.
 - **IRR = 1:** No hay efecto.



La estimación se adapta a la distribución de Poisson con enlace logarítmico.

- **Función de Verosimilitud ($L(\beta)$):**

$$L(\beta) = \prod_{i=1}^n \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$$

donde $\lambda_i = e^{\mathbf{x}_i^T \beta}$.

- **Función de Log-Verosimilitud ($\ell(\beta)$):**

$$\ell(\beta) = \sum_{i=1}^n \left[y_i \mathbf{x}_i^T \beta - e^{\mathbf{x}_i^T \beta} - \log(y_i!) \right]$$

- **Ecuaciones de Puntuación:** La solución de máxima verosimilitud se encuentra cuando:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^n x_{ij}(y_i - \lambda_i) = 0$$



Objetivo: Ajustamos un modelo para predecir el número de accidentes en función del tráfico y la visibilidad.

Salida de Coeficientes (summary)

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	9.607e-04	2.316e-03	0.415	0.678
trafico	9.999e-03	1.360e-06	< 2e-16	***
visibilidad	-2.000e-01	1.012e-04	< 2e-16	***

Métricas Globales del Modelo

- **Null deviance:** 1.68e+08 (con 99 g.l.)
- **Residual deviance:** 89.3 (con 97 g.l.)
- **AIC:** 1220.4



Significancia de los Predictores

Basado en los p-valores ($\text{Pr}(>|z|)$) de la diapositiva anterior:

- Tanto trafico como visibilidad son predictores **altamente significativos** (sus p-valores son prácticamente cero).

Interpretación de los Coeficientes (vía IRR)

Para interpretar el efecto práctico, exponenciamos los coeficientes ($\text{IRR} = e^\beta$):

- **IRR (tráfico):** $e^{0.01} \approx 1.01$.
 - Un aumento de 1 unidad en trafico **incrementa** la tasa de accidentes esperada en un **1%**.
- **IRR (visibilidad):** $e^{-0.20} \approx 0.82$.
 - Un aumento de 1 km en visibilidad **reduce** la tasa de accidentes esperada en un **18%**.



Como no hay solución analítica cerrada, se utiliza el algoritmo IRLS con componentes específicos para Poisson.

- **Pesos (w_i):** El peso de cada observación es simplemente la tasa esperada.

$$w_i = \lambda_i$$

- **Variable Dependiente Ajustada (z_i):**

$$z_i = \log(\lambda_i^{(t)}) + \frac{y_i - \lambda_i^{(t)}}{\lambda_i^{(t)}}$$

- **Actualización de parámetros:**

$$\beta^{(t+1)} = (\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(t)} \mathbf{z}^{(t)}$$



La estimación MLE en el modelo de Poisson tiene características particulares:

- ① **Equidispersión:** El modelo asume que la varianza aumenta linealmente con la media ($E(Y_i) = \text{Var}(Y_i) = \lambda_i$).
- ② **Convergencia Rápida:** Generalmente requiere menos iteraciones que la regresión logística.
- ③ **Estabilidad Numérica:** El enlace logarítmico garantiza automáticamente que las tasas estimadas λ_i sean siempre positivas.
- ④ **Interpretación Multiplicativa:** Los coeficientes se interpretan naturalmente como efectos multiplicativos sobre la tasa.



La métrica teórica fundamental sigue siendo la **deviance**, pero en la práctica, la prueba de bondad de ajuste más importante es la **evaluación de la sobredispersión**.

- **Herramienta de Diagnóstico:** El **estadístico de dispersión ($\hat{\phi}$)** se convierte en la medida de facto del ajuste.

$$\hat{\phi} = \frac{X_{\text{Pearson}}^2}{n - p}$$

- **Interpretación:**

- Si $\hat{\phi} \approx 1$: El supuesto de equidispersión se cumple y el ajuste es adecuado.
- Si $\hat{\phi} \gg 1$: Hay **sobredispersión**. El modelo no se ajusta bien a la variabilidad de los datos, y se debe considerar una **Regresión Binomial Negativa**.



La validación se enfoca en la **capacidad de predicción**: ¿qué tan cerca están los conteos predichos de los conteos reales?

- **Proceso:** Se ajusta el modelo en un conjunto de **entrenamiento** y se evalúa su rendimiento en un conjunto de **prueba**.
- **Métricas de Validación Principales:**
 - **Raíz del Error Cuadrático Medio (RMSE):** Mide la desviación estándar de los residuos. Penaliza más los errores grandes.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum (y_i - \hat{\mu}_i)^2}$$

- **Error Absoluto Medio (MAE):** Mide la magnitud promedio de los errores. Es menos sensible a outliers.

$$\text{MAE} = \frac{1}{n} \sum |y_i - \hat{\mu}_i|$$

- **Herramienta Visual: Gráfico de valores predichos vs. valores reales.** En un buen modelo, los puntos deben agruparse cerca de la línea diagonal $y = x$.



Este segundo ejemplo se centra en verificar dos aspectos clave de la estimación: la **convergencia del algoritmo** y el supuesto de **equidispersión**.

Comprobamos si se cumple el supuesto clave de Poisson ($\text{media} \approx \text{varianza}$).

Verificación de Equidispersión:

Media observada: 0.15

Varianza observada: 0.149

Razón varianza/media: 0.993

- **Conclusión:** La razón es **muy cercana a 1**, por lo que **se cumple el supuesto**. No hay evidencia de sobredispersión y el modelo de Poisson es adecuado.



Más allá de la regresión logística y de Poisson, existen otros GLM para manejar situaciones más complejas.

Estos modelos son especialmente útiles cuando los datos presentan características como:

- **Sobredispersión:** La varianza es mayor de lo esperado.
- **Sesgo:** La distribución de los datos es asimétrica.
- **Restricciones en el dominio:** La variable respuesta solo puede tomar valores positivos.

Exploraremos los tres modelos más importantes para estos casos:

- **Regresión Binomial Negativa**
- **Regresión Gamma**
- **Regresión Inversa Gaussiana**



El Problema: Sobredispersión

- Ocurre en datos de conteo cuando la **varianza es mayor que la media**, violando el supuesto clave de la regresión de Poisson ($Var(Y) = \mu$).
- **Causas comunes:** Heterogeneidad no modelada, dependencia entre eventos o exceso de ceros.
- **Consecuencia:** La regresión de Poisson subestima los errores estándar, llevando a conclusiones incorrectas sobre la significancia de los predictores.

La Solución: El Modelo Binomial Negativo

- Es una extensión del modelo de Poisson que introduce un **parámetro de dispersión (α)** para permitir que la varianza sea mayor que la media:

$$Var(Y) = \mu + \alpha\mu^2$$

- Si $\alpha = 0$, el modelo se reduce a la regresión de Poisson.



La forma funcional del modelo es la misma que la de Poisson (con enlace logarítmico), pero debemos interpretar el nuevo parámetro de dispersión y comparar ambos modelos.

Interpretación del Parámetro de Dispersión (θ)

- El software (como la función `glm.nb` en R) estima un parámetro θ , donde $\alpha = 1/\theta$.
- **Valores altos de θ** (ej. > 100): Poca sobredispersión. El modelo es similar a Poisson.
- **Valores bajos de θ** (ej. < 10): Mucha sobredispersión. El modelo Binomial Negativo es claramente más apropiado.

Comparación de Modelos (Poisson vs. Binomial Negativa)

- Se utiliza el **Criterio de Información de Akaike (AIC)**.
- Si el **AIC de la Binomial Negativa es menor** que el AIC de Poisson, debemos preferir el modelo Binomial Negativo.



Cuando la variable dependiente (Y) es **continua**, pero **no sigue una distribución normal**, la regresión lineal clásica no es adecuada.

Este escenario es común en variables que son:

- **Estrictamente positivas** (ej. costos, tiempos).
- Tienen una distribución **sesgada a la derecha**.

Los GLM nos ofrecen alternativas como la **Regresión Gamma** y la **Regresión Inversa Gaussiana**.



Regresión Gamma

- **Uso Típico:** Para variables continuas **positivas y con sesgo a la derecha** (tiempos, costos, reclamos de seguros).
- **Varianza:** Aumenta proporcionalmente al **cuadrado de la media** ($V(\mu) = \mu^2$).
- **Enlace Común:** Logarítmico ($\log(\mu) = X\beta$).

Regresión Inversa Gaussiana

- **Uso Típico:** Para tiempos de respuesta o variables con un **sesgo aún más pronunciado** que la Gamma.
- **Varianza:** Aumenta proporcionalmente al **cubo de la media** ($V(\mu) = \mu^3$).
- **Enlace Común:** Inverso al cuadrado ($1/\mu^2 = X\beta$).



La elección del modelo depende casi exclusivamente de la naturaleza de tu variable respuesta (Y).

- **¿Es Y binaria (0/1, Éxito/Fracaso)?** Usa **Regresión Logística**.
- **¿Es Y un conteo de eventos (nº de accidentes, nº de clientes)?**

Empieza con una **Regresión de Poisson**.

- **Importante:** Después, comprueba si hay **sobredispersión**. Si la hay ($\hat{\phi} > 1.5$), cambia a una **Regresión Binomial Negativa**.
- **¿Es Y continua, positiva y con sesgo a la derecha (tiempos, costos)?** Usa **Regresión Gamma**. Es una excelente alternativa a transformar la variable con logaritmos y usar un modelo lineal.
- **¿Es Y un tiempo hasta un evento con una asimetría muy pronunciada?** Considera una **Regresión Inversa Gaussiana**.