Universidad Rey Juan Carlos

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

Métodos Estadísticos de Predicción

Apuntes de la Asignatura

Grado en Matemáticas

AUTORES

- Víctor Aceña Gil
- Isaac Martín de Diego

2025-2026

En el siguiente enlace se puede encontrar la versión html actualizada de los presentes apuntes: https://urjcdslab.github.io/ModelosEstadisticosPrediccion/

Índice

Prefacio

Los modelos estadísticos han emergido como herramientas fundamentales en la era de la información, donde la capacidad de analizar y predecir comportamientos a partir de datos se ha convertido en una habilidad esencial. En este contexto, los modelos para la predicción juegan un papel crucial al permitirnos describir y cuantificar las relaciones entre variables, así como anticipar resultados futuros. Este libro está diseñado para proporcionar una comprensión profunda y práctica de estas técnicas, basándose en el contenido de la asignatura impartida en el **Grado en Matemáticas**.

A lo largo de los capítulos, encontrarás una combinación de teoría rigurosa y aplicaciones prácticas. Se abordarán temas como la regresión lineal simple y múltiple, métodos de selección de variables y regularización, ingeniería de características y modelos generalizados, entre otros. Además, todos los conceptos se ilustrarán con ejemplos en **R**, permitiéndote aplicar lo aprendido a conjuntos de datos reales.

El objetivo de este libro es doble: por un lado, proporcionar herramientas avanzadas para analizar relaciones sujetas a incertidumbre y, por otro, capacitarte para elegir el método más apropiado para resolver problemas de predicción o explicación, analizando la naturaleza de las variables y sus posibles interacciones. Al finalizar, habrás desarrollado una comprensión sólida de los modelos estadísticos y estarás preparado para enfrentar desafíos en el análisis predictivo con confianza y creatividad.

Filosofía pedagógica del volumen

La filosofía que subyace a la obra es un enfoque **"teórico-práctico"** deliberado y sin concesiones. No nos conformamos con una mera aplicación de "recetas" o una guía de funciones de software. Buscamos fomentar una comprensión profunda del *modus operandi* de cada modelo y método. Perseguimos un equilibrio entre la técnica estadística y la estrategia de resolución de problemas, bajo la firme convicción de que la labor práctica se desarrolla con mayor fluidez, creatividad y éxito cuando se cimienta en una comprensión robusta de los principios matemáticos y estadísticos subyacentes, tal y como defiende (Harrell 2015) en su influyente obra.

¿Qué aprenderás con este libro?

Al completar este recorrido, habrás desarrollado habilidades clave para:

- **Modelar** la dependencia entre una variable respuesta y múltiples predictores en conjuntos de datos complejos.
- **Resolver** problemas con iniciativa y creatividad, eligiendo las técnicas estadísticas más adecuadas para cada caso.
- **Evaluar** de forma crítica las ventajas e inconvenientes de diferentes alternativas metodológicas.
- **Implementar** estos modelos utilizando software estadístico profesional como R.
- **Interpretar** correctamente los resultados, proponer mejoras y tomar decisiones basadas en datos.
- **Adquirir** las competencias y la autonomía necesarias para emprender con éxito estudios de posgrado o proyectos profesionales en ciencia de datos.

Agradecemos a los profesores y colegas que han contribuido al desarrollo de esta asignatura y a la elaboración de este libro. Su dedicación y conocimiento han sido fundamentales para la creación de este recurso.

Esperamos que esta guía te resulte útil y enriquecedora.

¡Comenzamos!

! Grado en Matemáticas

Este libro presenta el material de la asignatura de Modelos Estadísticos para la Predicción del Grado en Matemáticas de la Universidad Rey Juan Carlos. Su contenido está fuertemente relacionado con las asignaturas de Estadística Matemática y Minería de Datos.

! Conocimientos previos

Es altamente recomendable que los alumnos que cursen esta materia manejen con soltura los conocimientos adquiridos en las asignaturas de Probabilidad y Estadística Matemática, así como herramientas de cálculo univariante, multivariante y álgebra lineal.

i Sobre los autores

Víctor Aceña Gil es graduado en Matemáticas por la UNED, máster en Tratamiento Estadístico y Computacional de la Información por la UCM y la UPM, doctor en Tecnologías de la Información y las Comunicaciones por la URJC y profesor del departamento de Informática y Estadística de la URJC. Miembro del grupo de investigación de alto rendimiento en Fundamentos y Aplicaciones de la Ciencia de Datos, DSLAB, de la URJC. Pertenece al grupo de innovación docente, DSLAB-TI.
Isaac Martín de Diego es diplomado en Estadística por la Universidad de Valladolid (UVA), licenciado en Ciencias y Técnicas Estadísticas por la Universidad Carlos III de

Madrid (UC3M), doctor en Ingeniería Matemática por la UC3M, catedrático de Ciencias de la Computación e Inteligencia Artificial del departamento de Informática y Estadística de la URJC. Es fundador y coordinador del DSLAB y del DSLAB-TI.

1 Introducción a los modelos de regresión

Este tema inaugural tiene como misión construir el andamiaje conceptual y filosófico sobre el que se asienta el modelado estadístico moderno. A lo largo de estas páginas, contextualizaremos la regresión no solo como una técnica, sino como un marco de pensamiento indispensable en la ciencia de datos y en cualquier disciplina de investigación cuantitativa. Exploraremos en profundidad su propósito dual, desgranaremos sus componentes axiomáticos hasta el último detalle, y ofreceremos una visión panorámica, rica en matices, de la vasta familia de modelos de regresión. El objetivo es preparar al lector, con solidez y sin prisas, para las inmersiones técnicas que seguirán en los capítulos posteriores. Como lectura complementaria que comparte esta filosofía de aprendizaje profundo pero aplicado, recomendamos encarecidamente la obra de (James et al. 2021).

1.1 Predecir vs. explicar

El modelado de regresión constituye una de las herramientas más potentes y flexibles del arsenal estadístico. Ofrece un marco metodológico riguroso para investigar y cuantificar las relaciones entre un conjunto de variables, y su aplicabilidad abarca un espectro extraordinariamente amplio de disciplinas: desde la física de partículas y la ingeniería aeroespacial, donde se usa para modelar sistemas complejos, hasta la econometría, la psicometría, la epidemiología o las finanzas, donde es fundamental para entender mercados y comportamientos.

Aunque en la práctica ambos objetivos a menudo se entrelazan, conceptualmente, el modelado estadístico se orienta hacia uno de dos polos, una dicotomía fundamental articulada brillantemente por (Shmueli 2010): la **predicción** o la **inferencia (explicación)**. Comprender esta distinción es el primer paso para convertirse en un modelador eficaz.

1. **Predicción**: El objetivo principal es la **precisión**. Se busca construir un modelo que pueda estimar con el menor error posible el valor de una variable de interés (la *respuesta*) basándose en la información proporcionada por otras variables (las *predictoras*). En este paradigma, el modelo puede ser tratado como una "caja negra" (*black box*). Su funcionamiento interno o la interpretabilidad de sus componentes son secundarios, siempre y cuando sus predicciones sean consistentemente fiables y robustas en datos no observados previamente.

Ejemplo

Una entidad financiera quiere predecir la probabilidad de que un cliente incurra en impago de un crédito. Utilizan variables como la edad, ingresos, nivel de estudios y historial crediticio. El banco no necesita necesariamente entender la "causa" exacta del impago; su principal interés es tener un modelo que clasifique correctamente a los futuros solicitantes como de alto o bajo riesgo para minimizar pérdidas.

2. **Inferencia**: El foco se desplaza radicalmente hacia la **comprensión** y la **interpretación**. El objetivo no es solo predecir, sino dilucidar la naturaleza de las interdependencias entre las variables. Se busca cuantificar cómo un cambio en una variable predictora influye, ya sea de forma causal o asociativa, en la variable de respuesta. Aquí, la interpretabilidad del modelo es primordial. El interés reside en la magnitud, el signo y, crucialmente, la incertidumbre estadística (expresada mediante errores estándar, intervalos de confianza y p-valores) de los parámetros estimados.

Ejemplo

Una epidemióloga investiga los factores de riesgo de una enfermedad cardíaca. Modela la presión arterial en función de variables como el índice de masa corporal (IMC), el consumo diario de sal y las horas de ejercicio semanales. Su objetivo no es solo predecir la presión arterial de un paciente, sino entender y cuantificar la relación: "¿En cuántos mmHg aumenta la presión arterial, en promedio, por cada gramo adicional de sal consumido al día, manteniendo constantes el IMC y el ejercicio?". La respuesta a esta pregunta tiene implicaciones directas para la salud pública y las recomendaciones dietéticas.

Una relación simbiótica

Aunque conceptualmente distintos, ambos objetivos no son mutuamente excluyentes; a menudo se benefician el uno del otro. Un modelo con una base inferencial sólida, que captura relaciones causales o asociativas verdaderas, suele tener un buen rendimiento predictivo. A la inversa, un modelo que demuestra una alta precisión predictiva en datos nuevos nos da confianza en que las relaciones que ha aprendido no son meras casualidades del conjunto de datos de entrenamiento, sino que probablemente reflejen patrones reales y generalizables. La tensión entre interpretabilidad y precisión es uno de los debates más fascinantes en la ciencia de datos moderna.

1.2 Anatomía de un modelo de regresión: los componentes axiomáticos

Todo modelo de regresión, desde el más simple hasta el más sofisticado, se construye sobre tres pilares fundamentales. Estos componentes, definidos en textos clásicos como el de (Kutner et al. 2005), son los ladrillos con los que edificaremos todo nuestro conocimiento.

1.2.1 La variable de respuesta

También designada como **variable dependiente**, **variable de salida**, **target**, **variable objetivo** o **variable explicada**. Representa el fenómeno o la característica principal cuyo comportamiento se busca modelar, comprender o predecir. La naturaleza de esta variable es, quizás, el factor más determinante a la hora de elegir el tipo de modelo de regresión. Puede ser:

- **Continua**: Una variable que puede tomar cualquier valor dentro de un rango. Ej: temperatura, altura, precio de una acción, concentración de un compuesto químico.
- **Discreta de Conteo**: Una variable que representa un número de eventos. Ej: número de accidentes en una intersección, número de clientes que entran en una tienda, número de mutaciones en un gen.
- **Binaria o Dicotómica**: Una variable con solo dos resultados posibles. Ej: éxito/fracaso, enfermo/sano, compra/no compra, spam/no spam.
- **Categórica**: Una variable que representa grupos o categorías. Si no tiene orden, es **nominal** (ej: tipo de sangre, partido político); si tiene un orden intrínseco, es **ordinal** (ej: nivel de satisfacción "bajo/medio/alto", estadio de una enfermedad "I/II/III/IV").

1.2.2 Las variables predictoras

Conocidas indistintamente como **variables independientes**, **explicativas**, **regresoras**, **covariables** o **características** (*features*). Son las magnitudes, atributos o factores que se postula que influyen o están asociados con el comportamiento de la variable de respuesta. Al igual que la variable de respuesta, pueden ser de diversa naturaleza (continuas, categóricas, etc.). La selección de estas variables es una de las fases más críticas del modelado, requiriendo una combinación de conocimiento del dominio, análisis exploratorio de datos y técnicas estadísticas formales.

1.2.3 El término de error aleatorio

Este componente, a menudo subestimado, es conceptualmente crucial. Simboliza la variabilidad intrínseca de la variable de respuesta que **no es capturada o explicada** por las variables

predictoras incluidas explícitamente en el modelo. El término de error ϵ no es un simple "error" en el sentido de equivocación; es un componente estocástico que amalgama múltiples fuentes de variabilidad:

- **Variables Omitidas**: Ningún modelo es perfecto. Siempre habrá factores que influyen en Y pero que no han sido medidos o incluidos en el modelo (variables latentes).
- **Error de Medición**: Las mediciones de Y (y también de X) pueden no ser perfectamente precisas.
- **Aleatoriedad Intrínseca**: Muchos fenómenos naturales y sociales tienen un componente de variabilidad irreducible. Dos individuos con idénticos valores en todas las variables predictoras pueden, aun así, tener valores distintos en la variable de respuesta.

Formalmente, la relación fundamental de la regresión se expresa como la descomposición de la variable de respuesta en una parte sistemática y una parte aleatoria:

$$Y = \underbrace{f(X_1, \dots, X_k)}_{\text{Componente Sistemática}} + \underbrace{\epsilon}_{\text{Componente Aleatoria}}$$

donde $f(\cdot)$ denota la **componente sistemática** (o determinística) del modelo, que representa el valor esperado de Y para unos valores dados de las X. La función f es lo que intentamos estimar a partir de los datos. Por su parte, ϵ es la **componente aleatoria**, y gran parte del diagnóstico y la inferencia en regresión se basa en verificar los supuestos que hacemos sobre la distribución de este término (ej: que su media es cero, que su varianza es constante, etc.).

Linealidad en los parámetros, no en las variables

Una característica que define a los **modelos de regresión lineal** (y que se extiende a muchos otros tipos de modelos) es que la función $f(\cdot)$ mantiene una relación lineal con respecto a sus **parámetros desconocidos** (los coeficientes beta, β_j). Es crucial enfatizar que esta "linealidad en los parámetros" **no impone una restricción de linealidad en las variables predictoras mismas**.
Por el contrario, es común y metodológicamente válido incorporar transformaciones no lineales de los predictores o interacciones complejas entre ellos para capturar relaciones más sofisticadas. Por ejemplo, el siguiente modelo es un **modelo de regresión lineal**:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 \log(X_2) + \beta_4 (X_1 \cdot X_2) + \epsilon$$

Aunque la relación entre Y y las variables X_1 y X_2 es claramente no lineal (es cuadrática en X_1, logarítmica en X_2 e incluye una interacción), el modelo es **lineal en los parámetros** $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$. La función f es una combinación lineal de estos coeficientes. Esta flexibilidad es una de las razones de la enorme potencia de los modelos lineales.
El siguiente bloque de código en R genera un ejemplo visual. Simulamos datos que siguen una relación cuadrática y luego ajustamos un modelo lineal que incluye un término

cuadrático (X^2). Como se puede observar en la figura, la línea de regresión (azul) captura perfectamente la curvatura de los datos, demostrando que un modelo lineal en sus parámetros puede modelar relaciones no lineales en sus variables.

```r
# Cargar la librería necesaria para la visualización
library(ggplot2)

# 1. Simulación de datos
set.seed(42) # Para reproducibilidad
n <- 100 # Número de observaciones
x <- runif(n, -5, 5)
# La relación verdadera es cuadrática: y = 1.5 + 0.5*x + 0.8*x^2 + error
y <- 1.5 + 0.5 * x + 0.8 * x^2 + rnorm(n, mean = 0, sd = 5)
datos <- data.frame(x, y)

# 2. Ajuste del modelo lineal
# Usamos I(x^2) para indicar que tratamos x^2 como una variable
modelo_cuadratico <- lm(y ~ x + I(x^2), data = datos)

# 3. Visualización con ggplot2
ggplot(datos, aes(x = x, y = y)) +
  geom_point(alpha = 0.6, color = "gray40") + # Puntos de los datos originales
  geom_smooth(method = "lm", formula = y ~ x + I(x^2), se = FALSE, color = "#0072B2", size
  labs(
    title = "Modelo Lineal con Término Cuadrático",
    x = "Variable Predictora (X)",
    y = "Variable de Respuesta (Y)"
  ) +
  theme_classic(base_size = 14)
```

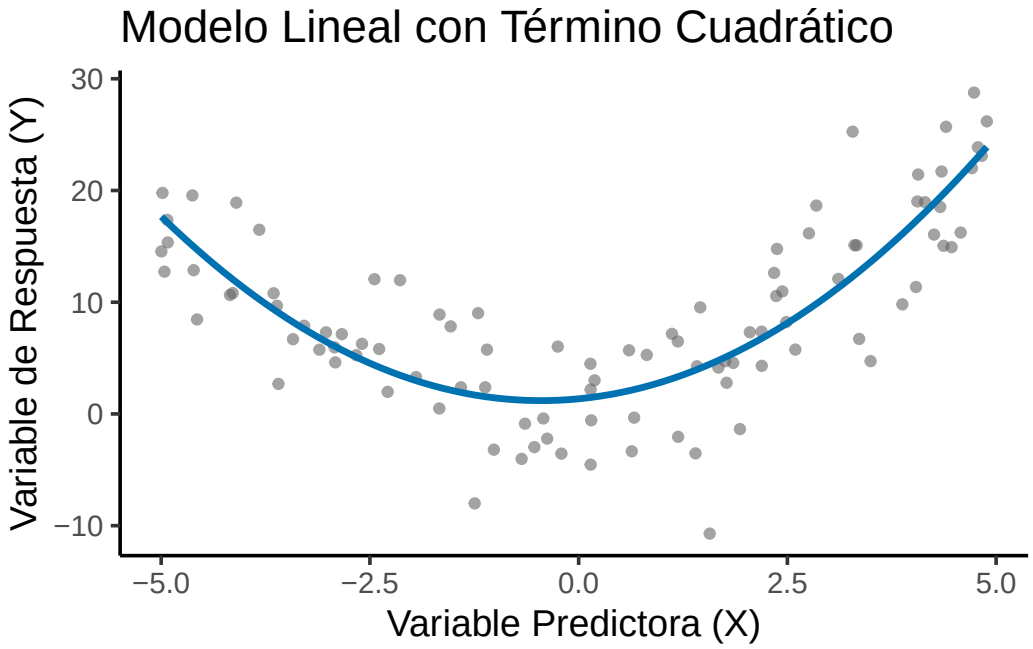


Figure 1.1: Ejemplo de un modelo lineal en los parámetros que captura una relación no lineal (cuadrática) en los datos.

1.3 Un viaje preliminar por el universo de los modelos de regresión

La regresión lineal clásica, que será el objeto de estudio de los primeros capítulos, es el punto de partida y la piedra angular sobre la cual se erige una prolífica y fascinante gama de metodologías estadísticas avanzadas. Este volumen se dedicará a desentrañar con rigor las siguientes extensiones y especializaciones, que permiten al analista abordar una variedad casi infinita de problemas.

1.3.1 Modelos lineales (LMs)

Constituyen el paradigma fundamental, el alfabeto sobre el que se escribe el lenguaje del modelado estadístico. Son mucho más que una simple técnica para ajustar una recta a una nube de puntos; son el laboratorio donde se forjan y se comprenden los conceptos esenciales que nos acompañarán durante todo nuestro viaje. Es aquí donde aprenderemos a:

- **Estimar parámetros** e interpretar su significado en el contexto del problema.

- **Cuantificar la incertidumbre** de nuestras estimaciones mediante errores estándar e intervalos de confianza.
- **Realizar contrastes de hipótesis** para evaluar si la relación entre nuestras variables es estadísticamente significativa o fruto del azar.
- **Diagnosticar la "salud" de un modelo**, examinando si los supuestos sobre los que se construye son razonables para nuestros datos.

En su forma más clásica, el modelo lineal asume que la variable de respuesta (y, por consecuencia, el término de error aleatorio) sigue una distribución Normal o Gaussiana. Esta asunción es la clave que desbloquea todo el elegante aparato de la inferencia estadística, permitiéndonos realizar pruebas exactas y derivar propiedades matemáticas bien conocidas. Técnicas tan ubicuas en la ciencia como el Análisis de la Varianza (ANOVA) o el Análisis de la Covarianza (ANCOVA) no son más que casos particulares de la gran familia de los modelos lineales, un hecho que unifica campos de la estadística que históricamente se estudiaban por separado. Dominar los LMs es, sencillamente, un requisito indispensable.

1.3.2 Modelos lineales generalizados (GLMs)

Si los LMs son el alfabeto, los GLMs son la gramática que nos permite construir frases complejas y con significado en una variedad de contextos mucho más amplia. Introducidos en el influyente y verdaderamente revolucionario trabajo de (Nelder and Wedderburn 1972), los GLMs representan un salto conceptual que expande de forma masiva el universo de problemas que podemos abordar. Suponen una generalización elegante que nos permite escapar de la "tiranía" de la distribución Normal y modelar respuestas con una variedad mucho más amplia de naturalezas y escalas.

Esta flexibilidad se logra mediante la combinación de dos ingeniosos mecanismos que son el corazón de la teoría:

1. **La familia exponencial de distribuciones**: Los GLMs no funcionan con cualquier distribución, sino con aquellas que pertenecen a una "familia" matemática con propiedades muy convenientes: la **familia exponencial**. Este "club" de distribuciones es muy selecto, pero incluye a miembros tan importantes como la Normal, la Poisson (para datos de conteo), la Binomial (para datos de proporciones o binarios), la Gamma (para datos continuos positivos y asimétricos) o la Binomial Negativa. Su estructura matemática común permite desarrollar una teoría unificada para la estimación de parámetros, lo que es un logro teórico de primer orden.
2. **La función de enlace (link function)**: Este es el verdadero golpe de genialidad. El predictor lineal de nuestro modelo, $X\beta$, puede tomar cualquier valor en la recta real, desde $-\infty$ hasta $+\infty$. Sin embargo, la media de nuestra variable de respuesta, $E[Y] = \mu$, a menudo está restringida. Por ejemplo, una probabilidad (μ en un modelo binomial) debe estar entre 0 y 1; un conteo (μ en un modelo de Poisson) debe ser positivo.

La función de enlace, $g(\cdot)$, actúa como un **"traductor"** o un **"puente"** que conecta estos dos mundos. Transforma la media restringida de la respuesta para que pueda ser modelada por el predictor lineal no restringido. La relación fundamental es, por tanto, $g(E[Y]) = g(\mu) = X\beta$.

- Para datos de **conteo** (Poisson), se usa un **enlace logarítmico** ($g(\mu) = \log(\mu)$). Esto garantiza que, al invertir la función para obtener la media ($\mu = \exp(X\beta)$), el resultado será siempre positivo, como debe ser un conteo.
- Para datos **binarios** (Binomial), se usa un **enlace logit** ($g(\mu) = \log(\frac{\mu}{1-\mu})$). Esta función toma una probabilidad μ en el rango (0, 1) y la proyecta sobre toda la recta real, permitiendo que sea modelada por $X\beta$.

Gracias a los GLMs, podemos usar el mismo marco conceptual de la regresión lineal para modelar una gama de fenómenos increíblemente diversa, desde predecir la cantidad de ciclistas en una ciudad (Poisson) hasta la probabilidad de que un paciente responda a un tratamiento (logística).

1.3.3 Modelos de efectos mixtos (Mixed Models)

Su desarrollo responde a la necesidad crítica de analizar datos que exhiben estructuras de dependencia o correlación, como agrupamientos, anidamientos o jerarquías. En datos estándar, asumimos que las observaciones son independientes, pero esta asunción se viola en casos como: * **Medidas repetidas** sobre los mismos sujetos (ej: medir la presión arterial de un paciente cada mes). * **Datos longitudinales** (un tipo de medida repetida a lo largo del tiempo). * **Datos agrupados** (ej: estudiantes anidados dentro de clases, que a su vez están anidadas dentro de colegios). Estos modelos, detallados en obras como la de (Pinheiro and Bates 2000), introducen explícitamente una estructura de correlación en el término de error mediante la incorporación de **efectos aleatorios**, que permiten capturar la variabilidad entre los diferentes grupos o individuos, además de los **efectos fijos** que representan a la población general.

1.3.4 Modelos aditivos generalizados (GAMs)

Representan una extensión natural y altamente flexible de los GLMs que relaja el supuesto de linealidad entre el predictor transformado y las covariables. Los GAMs, cuya implementación moderna se debe en gran parte al trabajo de (Wood 2017), permiten modelar estas relaciones mediante **funciones suaves** no paramétricas (como *splines*), manteniendo al mismo tiempo la estructura aditiva del modelo. La forma general es $g(\mu) = \alpha + f_1(x_1) + f_2(x_2) + ... + f_p(x_p)$, donde las $f_i(\cdot)$ son funciones suaves de los predictores estimadas a partir de los datos. Esto permite capturar patrones no lineales complejos sin necesidad de especificar una forma funcional paramétrica a priori, logrando un equilibrio excepcional entre flexibilidad e interpretabilidad.

R como lenguaje del modelado estadístico

Este compendio no es un texto puramente teórico. Fusiona intrínsecamente la exposición de los conceptos con su aplicación computacional directa a través del lenguaje y entorno estadístico **R**. R se ha consolidado como el estándar de facto en la investigación estadística y la ciencia de datos académica por su potencia, flexibilidad y el inmenso ecosistema de paquetes contribuidos por la comunidad científica. Se presupone en el lector una familiaridad operativa básica con R, y se fomenta activamente el desarrollo de una fluidez progresiva mediante la reproducción, modificación y experimentación con los numerosos ejemplos y fragmentos de código presentados.
La capacidad de ejecutar análisis en R es fundamental para todo el ciclo de vida del modelado:

- La **exploración de datos** y la visualización inicial.
- La **estimación de parámetros** y el ajuste de los modelos.
- El **diagnóstico riguroso** de la adecuación del modelo y la validación de sus supuestos.
- La **producción de gráficos** y tablas de alta calidad para comunicar los resultados.

En R, las herramientas fundamentales para la regresión lineal (`lm()`) y los modelos lineales generalizados (`glm()`) están incluidas en el paquete **`stats`**, que es uno de los **paquetes base** y se carga automáticamente con cada sesión. Por lo tanto, no necesitamos instalarlo ni cargarlo.
A lo largo del libro, extenderemos esta funcionalidad base con paquetes especializados que sí requieren instalación y carga. Entre los más importantes que usaremos se encuentran:

- `mgcv`: La implementación de referencia para GAMs, mantenida por su creador, Simon Wood, y citada en (Wood 2017).
- `lme4` y `nlme`: Los dos paquetes fundamentales para el ajuste de modelos de efectos mixtos, desarrollados por los pioneros en el campo (Pinheiro and Bates 2000; Bates et al. 2015).
- `rms`: Un paquete y una filosofía de trabajo para implementar estrategias de modelado de regresión robustas, como se detalla en la obra de (Harrell 2015).
- `gamair`: Contiene numerosos conjuntos de datos que acompañan al libro de (Wood 2017), ideales para practicar con GAMs.

1.4 Una breve crónica del desarrollo de la regresión

1.4.1 Los orígenes: Galton y la "regresión a la mediocridad"

La gestación de la metodología de regresión se traza hasta las investigaciones pioneras de Sir **Francis Galton**, un polímata de la era victoriana. A finales del siglo XIX, estudiando la herencia de la estatura, Galton recopiló datos de padres e hijos y notó un fenómeno curioso: los padres muy altos tendían a tener hijos altos, pero, en promedio, no tan altos como ellos. Análogamente, los padres muy bajos tenían hijos bajos, pero no tan bajos como ellos. Acuñó el término **"regresión a la mediocridad"** (hoy diríamos "regresión a la media") para describir esta tendencia de las características de la descendencia a "regresar" hacia la media de la población, en lugar de perpetuar los extremos de los progenitores (Galton 1886).

Estudios de Galton sobre estatura

Datos recopilados

- Galton recopiló datos sobre las estaturas de **928 hijos** y sus respectivos **padres**.
- Las medidas fueron expresadas en pulgadas (1 pulgada = 2.54 cm).

- En sus análisis, utilizó el promedio de las estaturas de ambos padres, conocido como **estatura media parental**, para compararlo con la estatura de los hijos.

Principales hallazgos

1. **Relación lineal entre padres e hijos:**
   Galton observó que existe una relación positiva entre la estatura de los padres y la de los hijos. Los padres altos tienden a tener hijos altos, y los padres bajos tienden a tener hijos bajos. Esta relación puede modelarse con una línea recta, lo que inspiró la formulación de la regresión lineal.

2. **Regresión a la media:**

   - Aunque los hijos de padres altos son, en promedio, más altos que el promedio general de la población, también tienden a ser **menos altos que sus padres**.

   - De manera similar, los hijos de padres bajos son más bajos que el promedio general, pero suelen ser **menos bajos que sus padres**.

   - Este fenómeno, que Galton llamó "regresión a la media", ocurre porque las características extremas tienden a suavizarse en la siguiente generación debido a la influencia de múltiples factores genéticos y ambientales.

3. **Ecuación de la recta de regresión:**
   Galton ajustó una recta para describir la relación entre la estatura media parental (X) y la estatura de los hijos (Y):

$$Y = \beta_0 + \beta_1 X$$

   Donde:

   - β_0: Intercepto, representa la estatura promedio de los hijos cuando la estatura parental es promedio.
   - β_1: Pendiente, indica cómo cambia la estatura de los hijos por cada unidad de cambio en la estatura media parental.

Importancia en la Estadística

1. **Regresión lineal:**
   Este estudio introdujo el concepto de **recta de regresión**, que describe cómo varía la media de una variable dependiente en función de una variable independiente.

2. **Correlación:**
   Galton también estudió el grado de relación entre variables, precursor del concepto de **coeficiente de correlación** desarrollado posteriormente por Karl Pearson, un discípulo suyo.

3. **Regresión a la media:**
   El término y la idea detrás de "regresión a la media" surgieron de estos estudios y son hoy fundamentales en estadística y genética.

Ejemplo Gráfico
Galton representó sus datos en gráficos de dispersión, mostrando cómo los puntos (pares de estatura media parental y estatura de los hijos) se agrupan alrededor de la recta de regresión, ilustrando la tendencia general de la relación.

```r
# Cargar los paquetes necesarios
library(ggplot2)
library(HistData)

# Cargar los datos de Galton
data("GaltonFamilies")

# Crear el modelo de regresión lineal para obtener los coeficientes
modelo <- lm(childHeight ~ midparentHeight, data = GaltonFamilies)

# Crear la etiqueta para la ecuación de la recta de forma más limpia
# Usamos sprintf() para un formato más controlado y legible
eq_label <- sprintf("y = %.2f + %.2f * x", coef(modelo)[1], coef(modelo)[2])

# --- Gráfico Mejorado ---
# Usamos un tema más limpio y colores más suaves para una apariencia profesional.
# geom_jitter() es mejor que geom_point() para estos datos, ya que evita la superposición c
ggplot(GaltonFamilies, aes(x = midparentHeight, y = childHeight)) +

  # 1. Puntos de datos: Usamos geom_jitter para visualizar mejor los puntos superpuestos
  #    y añadimos transparencia (alpha) para ver la densidad.
  geom_jitter(alpha = 0.3, color = "gray50", width = 0.1, height = 0.1) +

  # 2. Línea de regresión: En un color azul profesional y más gruesa para que destaque.
  geom_smooth(method = "lm", se = FALSE, color = "#0072B2", size = 1.2) +

  # 3. Anotación: Añadimos la ecuación de la recta de forma elegante,
  #    usando el mismo color que la línea para crear cohesión visual.
  annotate(
    "text",
    x = 66, y = 74, # Posición ajustada para mejor visibilidad
    label = eq_label,
    color = "#0072B2", # Mismo color que la línea
    size = 4.5, # Tamaño de la fuente
    fontface = "italic" # Cursiva para la ecuación
  ) +

  # 4. Títulos y etiquetas: Mejorados para mayor claridad y contexto.
  #    Añadimos un subtítulo y una fuente.
  labs(
    title = "Regresión de la Estatura de Hijos vs. Padres",
    x = "Promedio de Estatura de los Padres (pulgadas)",
    y = "Estatura del Hijo/a (pulgadas)",
    caption = "Fuente: Paquete HistData de R"
  ) +

  # 5. Tema: Usamos un tema limpio y profesional como base.
  theme_classic(base_size = 14)
```

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

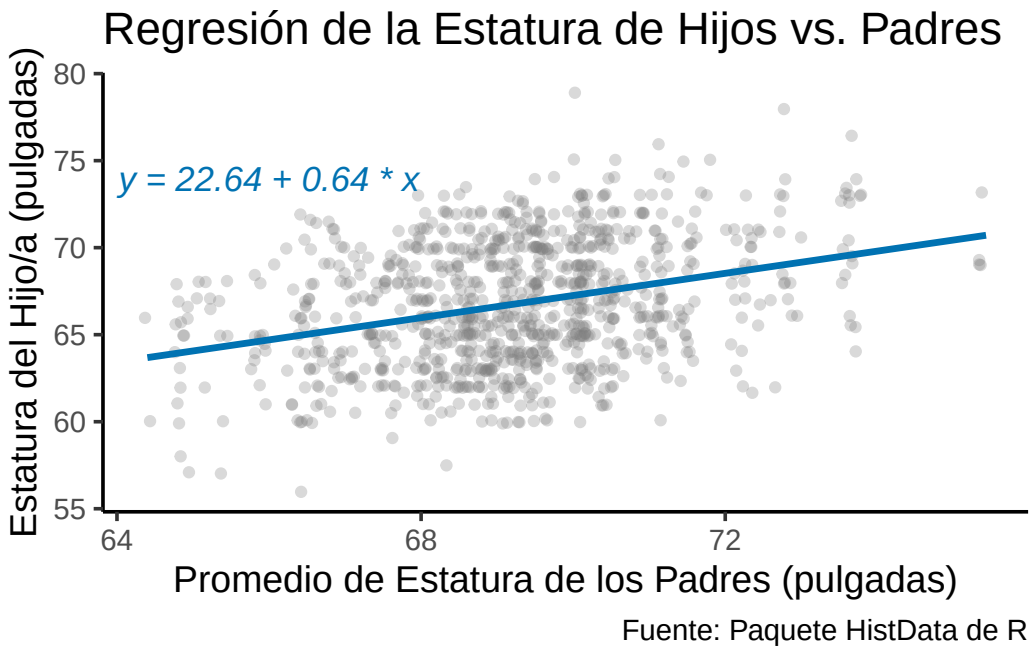


Figure 1.2: Datos históricos del estudio sobre la 'regresión a la media'.

1.4.2 La formalización matemática: Legendre y Gauss

Aunque Galton sentó las bases conceptuales e introdujo el término, la formalización matemática de la estimación de parámetros en modelos lineales se atribuye a dos de los más grandes matemáticos de la historia. **Adrien-Marie Legendre** publicó en 1805 el "Método de los mínimos cuadrados" como un procedimiento numérico para ajustar observaciones astronómicas. Pocos años después, **Carl Friedrich Gauss** no solo publicó que había desarrollado el mismo método de forma independiente años antes, sino que lo dotó de una profundidad teórica mucho mayor, conectándolo con la teoría de la probabilidad y derivándolo bajo el supuesto de errores distribuidos normalmente, convirtiéndolo en la técnica fundamental para la estimación en modelos lineales que sigue siendo hoy.

1.4.3 El desarrollo moderno: la revolución de los GLMs

A lo largo del siglo XX, la regresión experimentó un desarrollo explosivo. Sin embargo, el hito que probablemente más ha influido en la práctica estadística moderna fue la publicación del

artículo sobre **Modelos Lineales Generalizados (GLMs)** por **John Nelder** y **Robert Wedderburn** en 1972 (Nelder and Wedderburn 1972). Esta obra seminal fue revolucionaria porque unificó bajo un mismo paraguas conceptual y computacional diversas clases de modelos que hasta entonces se trataban por separado: la regresión lineal para datos normales, la regresión logística para datos binarios y la regresión de Poisson para datos de conteo. Esto estimuló enormemente el desarrollo de software y la aplicación del modelado estadístico a una nueva y vasta gama de problemas.

1.4.4 La evolución contemporánea

Este legado continúa evolucionando a un ritmo vertiginoso, con la inclusión de modelos jerárquicos y bayesianos, métodos no paramétricos y de *machine learning* como los árboles de regresión, y la adaptación de la regresión al análisis de datos masivos (*big data*). La regresión ha evolucionado desde una observación sobre la herencia biológica hasta convertirse en una de las herramientas más versátiles y poderosas del arsenal analítico moderno.

2 El modelo de regresión lineal simple

La regresión lineal constituye uno de los pilares fundamentales de la modelización estadística. Es, a menudo, el primer y más importante modelo predictivo que se aprende, no solo por su simplicidad e interpretabilidad, sino porque los conceptos que exploraremos aquí son la base sobre la que se construyen técnicas mucho más avanzadas, como el **modelo de regresión lineal múltiple**, los **modelos lineales generalizados (GLM)** o incluso conceptos utilizados en algoritmos de *machine learning* (Draper 1998; Kutner et al. 2005; James et al. 2021).

En este capítulo, daremos el primer y más crucial paso en nuestro viaje por el modelado predictivo: el estudio del **modelo de regresión lineal simple**. Para ello, seguiremos el ciclo de vida completo de un proyecto de modelado: comenzaremos con la exploración visual y cuantitativa de los datos, formalizaremos después nuestras observaciones mediante el lenguaje matemático del modelo y sus supuestos, aprenderemos a estimar sus parámetros, realizaremos inferencias sobre ellos y, finalmente, diagnosticaremos la validez de nuestro modelo (Fox and Weisberg 2018; Harrell 2015).

La comprensión profunda que desarrollaremos aquí es esencial, ya que los principios de estimación, inferencia y diagnóstico que aprenderemos son directamente escalables al **modelo de regresión lineal múltiple**, que exploraremos en el siguiente capítulo.

! Objetivos de aprendizaje

Al finalizar este capítulo, serás capaz de:

1. **Comprender y aplicar** el proceso de modelización estadística para un problema con una única 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 el significado práctico de sus parámetros.
4. **Estimar los coeficientes** del modelo mediante el método de mínimos cuadrados ordinarios (MCO) y entender su derivación matemática y propiedades.
5. **Realizar inferencias sobre los parámetros** del modelo y evaluar su bondad de ajuste mediante el análisis de la varianza y el coeficiente de determinación R^2.
6. **Diagnosticar la adecuación del modelo**, evaluando visual y analíticamente si se cumplen los supuestos del modelo lineal.

2.1 Exploración inicial: visualización y cuantificación de la relación

Antes de sumergirnos en la teoría de la regresión, debemos hacer lo que todo buen analista hace primero: **observar y cuantificar la relación en los datos**. Este paso exploratorio es fundamental para formular hipótesis y justificar la elección de un modelo lineal.

2.1.1 Visualización: el gráfico de dispersión

La herramienta más potente para examinar la relación entre dos variables continuas es el **gráfico de dispersión** (*scatterplot*). Nos permite intuir visualmente la **forma**, la **dirección** y la **fuerza** de la relación. Una inspección visual es siempre el punto de partida.

2.1.2 Cuantificación de la asociación: covarianza y correlación

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

2.1.2.1 Covarianza

La **covarianza** es una medida de la variabilidad conjunta de dos variables aleatorias, X e Y. Nos indica la dirección de la relación lineal. La covarianza muestral, calculada a partir de nuestras observaciones (x_i, y_i), es:

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

El principal inconveniente de la covarianza es que su magnitud depende de las unidades de las variables, lo que la hace difícil de interpretar.

2.1.2.2 Coeficiente de correlación de Pearson

Para solucionar el problema de la escala, estandarizamos la covarianza, dividiéndola por el producto de las desviaciones típicas de cada variable. El resultado es el **coeficiente de correlación de Pearson** (r):

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

Este coeficiente es **adimensional** y siempre varía entre **-1 y 1**, lo que permite una interpretación universal de la fuerza de la asociación *lineal*.

Ejemplo práctico: Horas de estudio vs. Calificaciones

Vamos a plantear un problema que nos acompañará durante todo el capítulo: queremos saber si el tiempo de estudio semanal influye en las calificaciones finales.

```
library(ggplot2)
set.seed(123) # Para reproducibilidad

# Simulación de datos
datos <- data.frame(
  Tiempo_Estudio = round(runif(100, min = 5, max = 40), 1)
)
datos$Calificaciones <- round(5 + 0.1 * datos$Tiempo_Estudio + rnorm(100, mean = 0, sd = 0.

# Visualización
ggplot(datos, aes(x = Tiempo_Estudio, y = Calificaciones)) +
  geom_point(color = "#0072B2", alpha = 0.7) +
  labs(
    title = "Relación entre Tiempo de Estudio y Calificaciones",
    x = "Tiempo de Estudio (horas/semana)",
    y = "Calificaciones (promedio)"
  ) +
  theme_classic(base_size = 14)

# Cuantificación (los objetos se guardan para usarlos en el texto)
covarianza <- cov(datos$Tiempo_Estudio, datos$Calificaciones)
correlacion <- cor(datos$Tiempo_Estudio, datos$Calificaciones)
```

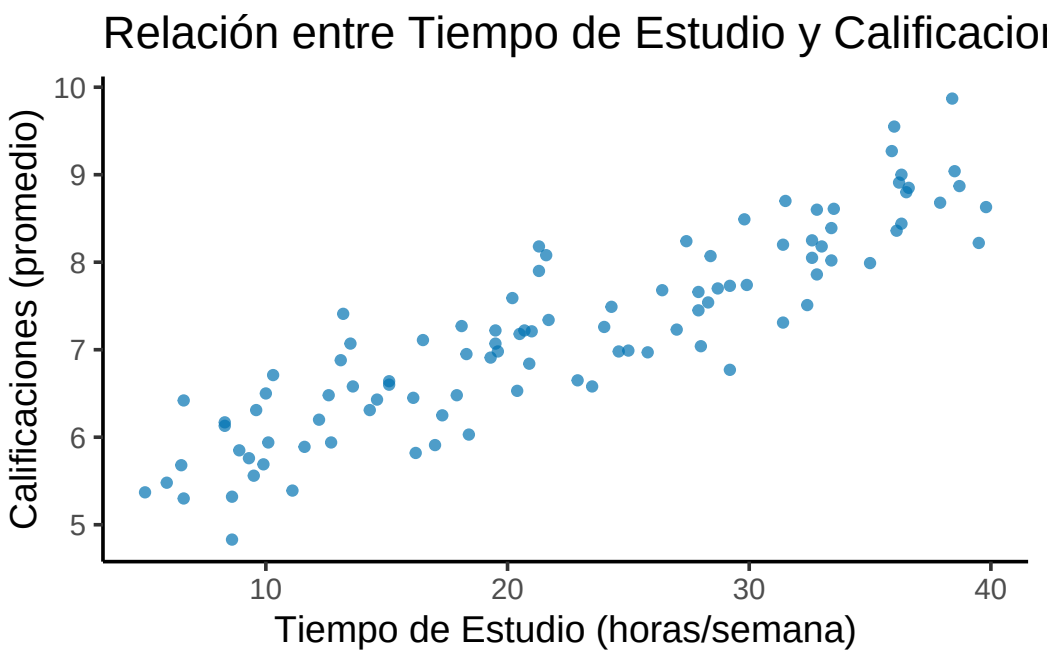


Figure 2.1: Relación entre tiempo de estudio y calificaciones.

El gráfico muestra una clara tendencia lineal positiva. La covarianza toma un valor de 9.82, y el coeficiente de correlación de Pearson es de 0.9. Ambos valores confirman que la asociación lineal es, además de positiva, muy fuerte. Esta evidencia visual y numérica nos da una base sólida para proponer un modelo de regresión lineal.

¡Correlación no implica causalidad!

El haber encontrado una fuerte correlación positiva entre el tiempo de estudio y las calificaciones (0.9) no nos autoriza a concluir que *una cosa causa la otra*. La regresión lineal puede demostrar que las variables se mueven juntas y nos permite predecir una a partir de la otra, pero no explica el porqué de la relación.
Podría existir una tercera variable oculta (p. ej., el interés del alumno en la materia) que influya tanto en las horas de estudio como en las calificaciones. Establecer causalidad requiere un diseño experimental riguroso (asignando aleatoriamente a los estudiantes a diferentes tiempos de estudio), no solo un análisis observacional.

2.2 Formulación teórica del modelo

Una vez que la exploración sugiere una relación lineal, el siguiente paso es formalizarla matemáticamente. Aquí es donde definimos la estructura teórica del modelo y los supuestos bajo los cuales operará.

2.2.1 El modelo poblacional y sus componentes

El **modelo poblacional** postula que la relación verdadera entre la variable respuesta Y y la predictora X sigue una línea recta, aunque contaminada por cierta aleatoriedad. Para cualquier individuo i de la población, esta relación se describe como:

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

En esta ecuación, β_0 y β_1 son los **parámetros poblacionales** (el intercepto y la pendiente verdaderos pero desconocidos), y ε_i es el **error aleatorio**, un componente fundamental que captura todas las fuentes de variabilidad que el modelo no puede explicar por sí solo. Específicamente, este término incluye:

- **Variables omitidas:** Factores que también afectan a las calificaciones (como la calidad del sueño, la motivación del estudiante o su conocimiento previo) y que no están en el modelo.
- **Error de medida:** Pequeñas imprecisiones al medir las variables (p. ej., un estudiante podría reportar 20 horas de estudio cuando en realidad fueron 19.5).
- **Aleatoriedad inherente:** La variabilidad puramente estocástica o impredecible en el comportamiento humano.

Como nunca observamos la población entera, nuestro trabajo consiste en usar una muestra para estimar el **modelo muestral**:

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

Aquí, los "gorros" ($\hat{\cdot}$) denotan **estimaciones** calculadas a partir de la muestra. La diferencia entre el valor real y el predicho, $e_i = y_i - \hat{y}_i$, se conoce como **residuo**.

2.2.2 Los supuestos del modelo lineal clásico (Gauss-Markov)

Para que el puente entre nuestro modelo muestral y la realidad poblacional sea sólido, debemos asumir que los errores teóricos ε_i se comportan de una manera predecible y ordenada. Estos supuestos, conocidos como condiciones de Gauss-Markov (Kutner et al. 2005; Weisberg 2005), son fundamentales para las propiedades óptimas de los estimadores de mínimos cuadrados.

1. **Linealidad**: La relación entre X y el valor esperado de Y es, en promedio, una línea recta: $E[Y_i|X_i] = \beta_0 + \beta_1 X_i$.
2. **Independencia de los errores**: El error de una observación no está correlacionado con el error de ninguna otra: $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ para $i \neq j$.
3. **Homocedasticidad**: La varianza del error es constante (σ^2) para todos los valores de X: $Var(\varepsilon_i|X_i) = \sigma^2$. Esto significa que la dispersión de los datos alrededor de la línea de regresión es la misma a lo largo de todos los valores de la variable predictora. La violación de este supuesto se conoce como **heterocedasticidad**, donde la dispersión de los errores cambia (p. ej., aumenta a medida que X crece).

Cuando el objetivo no es sólo estimar la recta, sino inferir con ella, entonces se asume una hipótesis más: la normalidad de la variable respuesta, o lo que es lo mismo, del error aleatorio:

4. **Normalidad de los errores**: Para la inferencia, se asume que los errores siguen una distribución Normal con media cero y varianza σ^2: $\varepsilon_i \sim N(0, \sigma^2)$.

Estos supuestos son esenciales para garantizar la validez de las estimaciones y conclusiones derivadas del modelo.

2.3 Estimación de los parámetros

Necesitamos un método para encontrar la "mejor" recta de ajuste. El **Método de Mínimos Cuadrados Ordinarios (MCO/OLS)** nos proporciona este criterio.

2.3.1 El criterio de mínimos cuadrados

MCO busca la recta que minimice la **Suma de los Cuadrados del Error (SSE)**, es decir, la suma de las distancias verticales al cuadrado entre los puntos observados y la recta de regresión:

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

2.3.2 Derivación matemática de los estimadores

Para encontrar los valores de β_0 y β_1 que minimizan esta función, recurrimos al cálculo. Tratamos la SSE como una función de dos variables y calculamos sus derivadas parciales, igualándolas a cero para encontrar el mínimo.

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

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

La resolución de este sistema de dos ecuaciones (conocidas como las **ecuaciones normales**) nos proporciona las fórmulas para los estimadores de MCO:

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

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

2.3.2.1 Interpretación práctica de los coeficientes

Una vez estimados, los coeficientes tienen una interpretación muy concreta y útil:

- **Pendiente** ($\hat{\beta}_1$): Representa el cambio promedio estimado en la variable respuesta Y por cada **aumento de una unidad** en la variable predictora X. En nuestro ejemplo, sería el número de puntos que se espera que aumente la calificación final por cada hora adicional de estudio semanal.
- **Intercepto** ($\hat{\beta}_0$): Es el valor promedio estimado de la variable respuesta Y cuando la variable predictora X es igual a cero. La interpretación del intercepto solo tiene sentido práctico si $X = 0$ es un valor plausible y se encuentra dentro del rango de nuestros datos. De lo contrario (como en nuestro ejemplo, donde nadie estudia 0 horas), a menudo se considera simplemente un ancla matemática para la recta de regresión.

> **i** Minimización de SSE
>
> La obtención de los estimadores de mínimos cuadrados para la regresión lineal simple se basa en minimizar la suma de los cuadrados de los residuos (SSE). Este método, desarrollado por Legendre y Gauss a principios del siglo XIX (Galton 1886; Weisberg 2005), es fundamental en la estadística moderna. Aquí está el proceso paso a paso:
> Para minimizar SSE, derivamos parcialmente con respecto a β_0 y β_1 y resolvemos el

sistema de ecuaciones.

1. **Primera derivada con respecto a** β_0:

$$\frac{\partial SSE}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i) \right).$$

`Igualando a cero:`

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0.$$

`Reordenando:`

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i. \tag{1}$$

2. **Primera derivada con respecto a** β_1:

$$\frac{\partial SSE}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i \left(y_i - (\beta_0 + \beta_1 x_i) \right).$$

`Igualando a cero:`

$$\sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0.$$

`Reordenando:`

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \tag{2}$$

Resolución del Sistema de Ecuaciones
El sistema está dado por las ecuaciones (1) y (2):

1. $n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$.
2. $\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$.

Resolviendo para β_0 y β_1:

1. De la primera ecuación, despejamos β_0:

$$\beta_0 = \frac{\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i}{n}. \tag{3}$$

2. Sustituimos β_0 en la segunda ecuación:

$$\frac{\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i}{n} \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i.$$

   Simplificando:

$$\beta_1 \left(\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \right) = \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}.$$

3. Expresamos β_1:

$$\beta_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}}{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}}.$$

   Esta es la fórmula para β_1, que puede reescribirse como:

$$\beta_1 = \frac{\text{Cov}(x, y)}{\text{Var}(x)},$$

   donde $\text{Cov}(x, y)$ y $\text{Var}(x)$ son la covarianza y la varianza muestral de x y y.

4. Finalmente, sustituimos β_1 en la ecuación (3) para obtener β_0:

$$\beta_0 = \bar{y} - \beta_1 \bar{x},$$

   donde $\bar{x}$ y $\bar{y}$ son las medias de x y y.

Bajo los supuestos del modelo, el **Teorema de Gauss-Markov** demuestra que estos estimadores son los **Mejores Estimadores Lineales Insesgados (MELI / BLUE)**.

2.4 Inferencia y bondad de ajuste

Una vez hemos estimado los parámetros del modelo, nuestro trabajo apenas ha comenzado. Ahora debemos pasar de la descripción a la inferencia. Necesitamos un conjunto de herramientas que nos permitan responder a preguntas cruciales: ¿Son nuestros coeficientes estimados, $\widehat{\beta}_0$ y $\widehat{\beta}_1$, meras casualidades de nuestra muestra o reflejan una relación real en la población? ¿Qué

tan bueno es nuestro modelo para explicar la variabilidad de la variable respuesta? Esta sección se dedica a responder estas preguntas.

2.4.1 Propiedades de los estimadores de MCO

Antes de realizar inferencias, es fundamental entender las propiedades teóricas de los estimadores que hemos calculado.

- **Insesgadez**: Los estimadores de MCO son insesgados. Esto significa que si pudiéramos repetir nuestro muestreo muchísimas veces y calcular los estimadores en cada muestra, el promedio de todas nuestras estimaciones de $\hat{\beta}_0$ y $\hat{\beta}_1$ convergería a los verdaderos valores poblacionales β_0 y β_1. Matemáticamente:

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

- **Varianza de los estimadores**: Las fórmulas para la varianza de nuestros estimadores cuantifican su precisión. Una varianza pequeña implica que el estimador es más estable a través de diferentes muestras.

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

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

  Donde σ^2 es la varianza (desconocida) del término de error ε.

- **Teorema de Gauss-Markov**: Este es uno de los resultados más importantes de la teoría de la regresión. Establece que, bajo los supuestos de linealidad, independencia y homocedasticidad (no se requiere normalidad), los estimadores de MCO son los **Mejores Estimadores Lineales Insesgados** (MELI, o BLUE en inglés). Esto significa que, de entre toda la clase de estimadores que son lineales e insesgados, los de MCO son los que tienen la menor varianza posible.

> **i** Propiedades adicionales para las predicciones y para los residuos
>
> - La suma de los residuos es cero:
>
> $$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n}(y_i - \hat{y}_i) = 0$$
>
> - La suma de los valores observados es igual a la suma de los valores ajustados:
>
> $$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$

- La suma de los residuos ponderados por los regresores es cero:

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

- La suma de los residuos ponderados por las predicciones es cero:

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

- La recta de regresión contiene el punto $(\bar{x}, \bar{y})$:

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

Ejemplo

Para los datos de calificaciones y tiempo de estudio, estos son los estimadores de los parámetros del modelo de regresión:

```r
# 1. Ajustamos el modelo lineal
modelo_estudio <- lm(Calificaciones ~ Tiempo_Estudio, data = datos)

# 2. Obtenemos el resumen completo del modelo
summary(modelo_estudio)
```

```
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
```

2.4.2 Estimación de la varianza del error

Las fórmulas de la varianza de los estimadores dependen de σ^2, la varianza del error poblacional, que es desconocida. Por lo tanto, necesitamos estimarla a partir de nuestros datos. Un estimador insesgado de σ^2 es la **Media Cuadrática del Error (MSE)**:

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

Dividimos por $n-2$, los **grados de libertad del error**, porque hemos "gastado" dos grados de libertad de nuestros datos para estimar los dos parámetros, β_0 y β_1. La raíz cuadrada de la MSE, $\hat{\sigma}$, se conoce como el **error estándar de los residuos** y es una medida de la dispersión promedio de los puntos alrededor de la recta de regresión.

2.4.2.1 El error estándar de los residuos y el RMSE

La raíz cuadrada de la MSE, $\hat{\sigma}$, se conoce formalmente como el **error estándar de los residuos** (*Residual Standard Error*). Este valor es nuestra estimación de la desviación estándar del error poblacional, σ, y es una medida de la dispersión promedio de los puntos alrededor de la recta de regresión.

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

En el campo del modelado predictivo y el *machine learning*, esta misma cantidad se conoce como la **Raíz del Error Cuadrático Medio** o **RMSE** (*Root Mean Squared Error*). Aunque la fórmula es idéntica, la interpretación del RMSE se centra en la **evaluación del rendimiento predictivo** del modelo. El RMSE nos dice, en promedio, cuál es la magnitud del error de predicción de nuestro modelo, y tiene la ventaja de estar en las **mismas unidades** que la variable respuesta Y. Por ejemplo, si estamos prediciendo precios de viviendas en euros, un RMSE de 5000 significa que nuestras predicciones se desvían, en promedio, unos 5000 € de los precios reales.

2.4.3 Análisis de la Varianza (ANOVA) para la significancia de la regresión

Una vez hemos estimado los coeficientes, necesitamos una prueba formal para determinar si el modelo en su conjunto es útil. Es decir, ¿la variable predictora X explica una porción de la variabilidad de la variable respuesta Y que sea estadísticamente significativa, o la relación que observamos podría deberse simplemente al azar? El **Análisis de la Varianza (ANOVA)** nos proporciona la herramienta para responder a esta pregunta a través del **contraste F de significancia global**.

Las hipótesis de este contraste son:

- $H_0 : \beta_1 = 0$: La hipótesis nula postula que no existe una relación lineal entre X e Y. El modelo no tiene poder explicativo y no es mejor que usar simplemente la media, $\bar{y}$, como predicción para cualquier valor de x.
- $H_1 : \beta_1 \neq 0$: La hipótesis alternativa sostiene que sí existe una relación lineal significativa.

> Repaso
>
> Es conveniente repasar el tema de *Análisis de la Varianza* estudiado en la asignatura de Inferencia, ya que los conceptos son directamente aplicables aquí.

La idea fundamental del ANOVA es comparar la variabilidad que nuestro modelo *explica* con la variabilidad que *no puede explicar* (el error residual). Para ello, se descompone la variabilidad total de nuestras observaciones (y_i) en dos partes ortogonales.

1. La **Suma Total de Cuadrados (SST)** mide la variabilidad total de los datos alrededor de su media. Es nuestra referencia base de la dispersión total que hay que explicar.

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

2. Esta variabilidad se descompone en:

   - **Suma de Cuadrados de la Regresión (SSR)**: Mide la parte de la variabilidad total que es explicada por nuestro modelo. Cuantifica cuánto se desvían las predicciones del modelo ($\hat{y}_i$) de la media general ($\bar{y}$).

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

   - **Suma de Cuadrados del Error (SSE)**: Mide la variabilidad residual, es decir, la parte que el modelo no puede capturar. Cuantifica la dispersión de los puntos reales (y_i) alrededor de la recta de regresión ($\hat{y}_i$).

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

La descomposición fundamental de la varianza es, por tanto: $\text{SST} = \text{SSR} + \text{SSE}$.

Para poder comparar estas sumas de cuadrados de forma justa, las estandarizamos dividiéndolas por sus respectivos **grados de libertad**, obteniendo así las **Medias Cuadráticas** (MS):

$$\text{MSR} = \frac{\text{SSR}}{1} \qquad \text{MSE} = \frac{\text{SSE}}{n-2}$$

Finalmente, el **estadístico F** se construye como el cociente entre la variabilidad explicada por el modelo y la variabilidad no explicada:

$$F = \frac{\text{MSR}}{\text{MSE}}$$

Intuitivamente, el estadístico F actúa como una **ratio de señal a ruido**. La MSR (la "señal") representa la variabilidad que nuestro modelo captura sistemáticamente, mientras que la MSE (el "ruido") representa la variabilidad aleatoria o residual. Un valor de F grande nos dice que la señal es mucho más fuerte que el ruido, lo que apoya la hipótesis de que la relación que hemos modelado es real y no fruto del azar.

Toda esta información se organiza de forma estándar en la **tabla ANOVA**:

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

Bajo la hipótesis nula ($H_0 : \beta_1 = 0$), el estadístico F sigue una distribución F con 1 y $n-2$ grados de libertad. Si el p-valor asociado a nuestro estadístico F es suficientemente pequeño ($p < \alpha$), rechazamos H_0 y concluimos que nuestro modelo tiene un poder explicativo estadísticamente significativo.

2.4.4 Bondad del ajuste: coeficiente de determinación

El coeficiente de determinación (R^2) es una medida clave que cuantifica qué proporción de la variabilidad total observada en la muestra (y_i) es explicada por la relación lineal con X a través del modelo. Su fórmula se deriva de la descomposición de la varianza:

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

Donde las sumas de cuadrados se calculan a partir de los datos muestrales:

- SST $= \sum_{i=1}^{n}(y_i - \bar{y})^2$: Suma Total de Cuadrados, mide la variabilidad total de las observaciones.
- SSR $= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$: Suma de Cuadrados de la Regresión, mide la variabilidad explicada por el modelo.
- SSE $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$: Suma de Cuadrados del Error, mide la variabilidad no explicada (residual).

Un R^2 cercano a 1 indica que el modelo ajusta bien los datos, mientras que un R^2 cercano a 0 indica un ajuste pobre.

Relación entre R^2 y el coeficiente de correlación

En el caso específico del modelo de regresión lineal simple, existe una relación directa y simple: el coeficiente de determinación R^2 es literalmente el cuadrado del coeficiente de correlación de Pearson (r) entre X e Y.

$$R^2 = (r_{xy})^2$$

Esto refuerza la idea de que ambos miden la fuerza de la asociación *lineal*, aunque R^2 lo hace desde la perspectiva de la varianza explicada por el modelo.

Interpretación de R^2

El coeficiente de determinación, R^2, es una métrica muy popular, pero su interpretación requiere cautela. Un valor alto no garantiza un buen modelo, y un valor bajo no siempre implica un modelo inútil. Es fundamental tener en cuenta las siguientes observaciones:

- R^2 no mide la linealidad de la relación. Un modelo puede tener un R^2 muy alto incluso si la relación subyacente entre las variables X e Y no es lineal. Por ello, un R^2 elevado nunca debe sustituir a un análisis gráfico de los residuos para verificar el supuesto de linealidad.
- R^2 es sensible al rango de la variable predictora X. Si el modelo de regresión es adecuado, la magnitud de R^2 aumentará si aumenta la dispersión de las observaciones x_i (es decir, si S_{xx} crece). Esto se debe a que un mayor rango en X tiende a aumentar la Suma Total de Cuadrados (SST), lo que puede inflar el valor de R^2 sin que la precisión del modelo (medida por la MSE) haya mejorado.
- **Un rango restringido en** X puede producir un R^2 artificialmente bajo. Como consecuencia del punto anterior, si los datos se han recogido en un rango muy estrecho de la variable X, el R^2 puede ser muy pequeño, aunque exista una relación fuerte y significativa entre las variables. Esto podría llevar a la conclusión errónea de que el predictor no es útil.

2.4.5 Inferencia sobre los coeficientes

Además de la prueba F global, podemos realizar inferencias sobre cada parámetro individualmente. Para ello, necesitamos el supuesto de normalidad de los errores.

2.4.5.1 Distribución de los estimadores

Bajo el supuesto de normalidad, se puede demostrar que los estimadores también siguen una distribución Normal:

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

Al estandarizar y reemplazar la desconocida σ^2 por su estimador $\hat{\sigma}^2 = \text{MSE}$, obtenemos un estadístico que sigue una distribución t-Student con $n-2$ grados de libertad:

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

donde $\text{SE}(\hat{\beta}_1) = \sqrt{\frac{\text{MSE}}{S_{xx}}}$ es el **error estándar** del estimador $\hat{\beta}_1$.

2.4.5.2 Contraste de hipótesis para la pendiente

El contraste más común es el de la significancia de la pendiente: * $H_0 : \beta_1 = 0$ * $H_1 : \beta_1 \neq 0$

Bajo H_0, el estadístico de contraste es:

$$t_0 = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)}$$

Rechazamos H_0 si $|t_0| > t_{\alpha/2, n-2}$ o, equivalentemente, si el p-valor asociado es menor que α.

> **i** Relación entre el contraste F y el contraste t
>
> En el contexto de la **regresión lineal simple** (y solo en este caso), el contraste F para la significancia global del modelo es matemáticamente equivalente al contraste t para la significancia del coeficiente β_1. Se puede demostrar que $F = t^2$, y el p-valor de ambos contrastes será idéntico.

2.4.5.3 Intervalo de confianza para la pendiente

A partir de la distribución t, podemos construir un intervalo de confianza al $100(1-\alpha)\%$ para el verdadero valor de la pendiente β_1:

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

Este intervalo nos da un rango de valores plausibles para el efecto de X sobre Y. Si el intervalo no contiene el cero, es equivalente a rechazar la hipótesis nula $H_0 : \beta_1 = 0$.

! Para recordar

En los programas estadísticos se suele proporcionar el p-valor del contraste. Puedes repasar el significado de p-valor proporcionado en la asignatura de Inferencia.

Ejemplo: Interpretación del `summary`

La función `summary()` en R nos proporciona toda esta información.

```
summary(modelo_estudio)
```

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

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

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

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

Interpretación:

- **Coefficients**: El p-valor para `Tiempo_Estudio` (<0.001) es muy pequeño, por lo que rechazamos H_0 y concluimos que la variable es un predictor significativo.

- **R-squared**: El valor de R^2 (0.81) nos indica que el 81% de la variabilidad en las calificaciones es explicada por el tiempo de estudio.
- **F-statistic**: El p-valor del estadístico F

  (98) confirma que el modelo en su conjunto es estadísticamente significativo.

2.5 Predicción de nuevas observaciones

Una vez que hemos ajustado y validado un modelo de regresión, uno de sus propósitos más importantes es utilizarlo para hacer predicciones. Sin embargo, es fundamental distinguir entre dos tipos de predicción:

1. **Estimar la respuesta media** para un valor dado de X. Por ejemplo: "¿Cuál es la calificación *promedio* que esperamos para todos los estudiantes que estudian 25 horas semanales?".
2. **Predecir una respuesta individual** para un valor dado de X. Por ejemplo: "Si un estudiante *concreto* estudia 25 horas semanales, ¿entre qué valores esperamos que se encuentre su calificación?".

Estos dos objetivos, aunque parecidos, responden a preguntas distintas y manejan diferentes fuentes de incertidumbre, lo que da lugar a dos tipos de intervalos.

2.5.1 Intervalo de confianza para la respuesta media

Este intervalo estima el valor esperado de Y para un valor concreto del regresor, x_0. Su objetivo es acotar dónde se encuentra la **línea de regresión poblacional verdadera** para ese punto x_0. La estimación puntual es $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

El intervalo de confianza al $100(1-\alpha)\%$ para la respuesta media $E[Y|X = x_0]$ viene dado por:

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

La anchura de este intervalo depende de dos fuentes de error: la incertidumbre en la estimación de la recta y la distancia del punto x_0 a la media $\bar{x}$. El intervalo es más estrecho cerca del centro de los datos y más ancho en los extremos.

2.5.2 Intervalo de predicción para una respuesta individual

Este intervalo es el que debemos usar cuando queremos predecir el valor para **una única observación futura**, no para la media. Como indicas, este intervalo debe tener en cuenta dos fuentes de variabilidad:

1. La incertidumbre sobre la localización de la verdadera recta de regresión (la misma que en el intervalo de confianza).
2. La variabilidad inherente de una observación individual alrededor de la recta de regresión (el error aleatorio ε_i, cuya varianza estimamos con la MSE).

Por esta razón, el intervalo de predicción **siempre será más ancho** que el intervalo de confianza para la respuesta media. El intervalo de predicción al $100(1-\alpha)\%$ para una observación futura y_0 en el punto x_0 es:

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

La única diferencia matemática es el **"+1"** dentro de la raíz cuadrada, que representa la varianza σ^2 del error de una sola observación.

2.5.2.1 Predicción para la media de *m* observaciones futuras

Si se desea un intervalo de predicción para la media de m futuras observaciones en un valor x_0, la fórmula se modifica ligeramente. Este intervalo será más estrecho que el de una sola observación pero más ancho que el de la respuesta media:

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

Ejemplo práctico: Predicción de calificaciones

Vamos a calcular y visualizar los intervalos para nuestro modelo de estudio. Usaremos la función `predict()` de R, que calcula estos intervalos de forma automática.

```r
# 1. Crear una secuencia de nuevos valores de X para predecir
nuevos_datos <- data.frame(
  Tiempo_Estudio = seq(min(datos$Tiempo_Estudio), max(datos$Tiempo_Estudio), length.out = 1
)

# 2. Calcular el intervalo de confianza para la RESPUESTA MEDIA
conf_interval <- predict(
  modelo_estudio,
  newdata = nuevos_datos,
  interval = "confidence",
  level = 0.95
)

# 3. Calcular el intervalo de predicción para una OBSERVACIÓN INDIVIDUAL
pred_interval <- predict(
  modelo_estudio,
  newdata = nuevos_datos,
  interval = "prediction",
  level = 0.95
)

# 4. Unir todo para graficar con ggplot2
plot_data <- cbind(nuevos_datos, as.data.frame(conf_interval), pred_pred = as.data.frame(pr
colnames(plot_data) <- c("Tiempo_Estudio", "fit_conf", "lwr_conf", "upr_conf", "fit_pred",

# 5. Visualización
ggplot() +
  # Capa 1: Puntos originales del dataframe 'datos'
  geom_point(data = datos, aes(x = Tiempo_Estudio, y = Calificaciones), color = "#0072B2",

  # Capa 2: Línea de regresión del dataframe 'plot_data'
  geom_line(data = plot_data, aes(x = Tiempo_Estudio, y = fit_conf), color = "black", linew

  # Capa 3: Banda de predicción (roja) del dataframe 'plot_data'
  geom_ribbon(data = plot_data, aes(x = Tiempo_Estudio, ymin = lwr_pred, ymax = upr_pred),

  # Capa 4: Banda de confianza (azul) del dataframe 'plot_data'
  geom_ribbon(data = plot_data, aes(x = Tiempo_Estudio, ymin = lwr_conf, ymax = upr_conf),

  # Etiquetas y tema
  labs(
    title = "Intervalos de Confianza y Predicción",
    x = "Tiempo de Estudio (horas/semana)",
    y = "Calificaciones (promedio)",
    caption = "La banda azul (más estrecha) es el IC del 95% para la media.\nLa banda roja
  ) +
  theme_classic(base_size = 14)
```

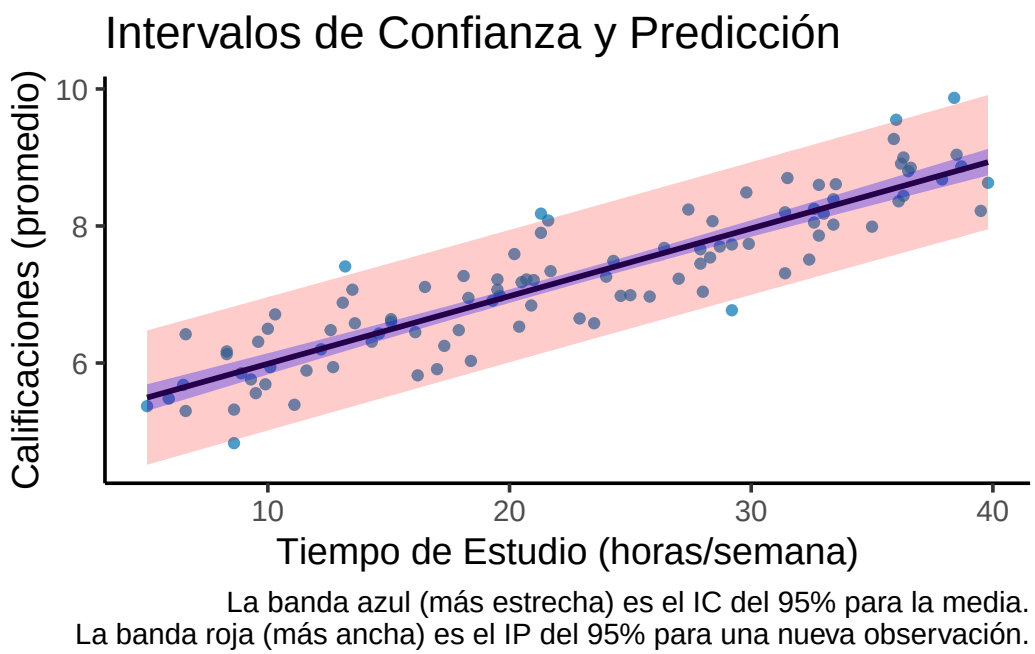


Figure 2.2: Comparación visual del intervalo de confianza (azul, más estrecho) y el intervalo de predicción (rojo, más ancho).

El gráfico muestra claramente que la incertidumbre al predecir una calificación individual es mucho mayor que la incertidumbre al estimar la calificación promedio. Ambas bandas se ensanchan al alejarse del centro de los datos.
Si quisiéramos una predicción para un estudiante que estudia 25 horas:

```
dato_nuevo <- data.frame(Tiempo_Estudio = 25)

# Guardamos la predicción para la media en un objeto
pred_media <- predict(modelo_estudio, newdata = dato_nuevo, interval = "confidence")

# Guardamos la predicción para un individuo en un objeto
pred_indiv <- predict(modelo_estudio, newdata = dato_nuevo, interval = "prediction")
```

Interpretación:

- Con un 95% de confianza, la calificación **promedio** de los estudiantes que estudian 25 horas está entre **7.37** y **7.57**.
- Con un 95% de confianza, la calificación de **un estudiante concreto** que estudia 25 horas estará entre **6.5** y **8.44**.

2.6 Diagnóstico del Modelo

Una vez que hemos ajustado un modelo y evaluado su significancia, el trabajo no ha terminado. Un paso crucial, a menudo subestimado, es el **diagnóstico del modelo** (Fox and Weisberg 2018; Harrell 2015). Este proceso consiste en verificar si se cumplen los supuestos del modelo de regresión lineal clásico. La fiabilidad de nuestras inferencias (los p-valores de los contrastes t y F, y los intervalos de confianza) depende directamente de la validez de estos supuestos.

El diagnóstico se realiza principalmente a través del **análisis de los residuos** del modelo ($e_i = y_i - \widehat{y}_i$). Los residuos son nuestra mejor aproximación empírica de los errores teóricos no observables (ε_i). A continuación, se detalla cómo verificar cada uno de los supuestos clave.

2.6.1 Linealidad

Este supuesto establece que la relación entre la variable predictora X y el valor esperado de la variable respuesta Y es, en promedio, una línea recta: $E[Y|X] = \beta_0 + \beta_1 X$.

Métodos de diagnóstico:

1. **Diagnóstico visual:** El gráfico de **residuos** (e_i) frente a los **valores ajustados** por el modelo ($\widehat{y}_i$) es la herramienta fundamental. La lógica es sencilla pero potente: si el modelo lineal es adecuado, los errores que comete (los residuos) deberían ser completamente aleatorios, sin guardar relación alguna con la magnitud de las predicciones.
2. **Test de Ramsey RESET:** Este test estadístico detecta violaciones de la forma funcional mediante la inclusión de términos no lineales ($\widehat{y}^2, \widehat{y}^3, ...$) en el modelo.
    - **H :** La forma funcional es correcta (lineal)
    - **H :** La forma funcional es incorrecta (no lineal)
    - **Estadístico:** Sigue una distribución F bajo H
    - **Interpretación:** p-valor bajo indica violación de linealidad

En un escenario ideal, el gráfico debería parecer una nube de puntos distribuida horizontalmente y sin estructura aparente, centrada en la línea del cero. Esto nos indica que los errores son, en promedio, nulos para todos los niveles de predicción, cumpliendo así el supuesto de linealidad. La línea roja que R superpone en este gráfico, que suaviza la tendencia de los puntos, debería ser prácticamente plana y pegada al cero, confirmando la ausencia de patrones.

Ejemplo de un modelo válido

Para nuestro `modelo_estudio`, podemos generar específicamente el primer gráfico de diagnóstico, que es el de Residuos vs. Valores Ajustados.

```
# Crear un dataframe con los datos para ggplot2
library(ggplot2)
library(broom)

# Extraer residuos y valores ajustados
datos_diagnostico <- data.frame(
  residuos = residuals(modelo_estudio),
  valores_ajustados = fitted(modelo_estudio)
)

# Gráfico de Residuos vs. Valores Ajustados con ggplot2
ggplot(datos_diagnostico, aes(x = valores_ajustados, y = residuos)) +
  geom_point(color = "#0072B2", alpha = 0.7) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  geom_smooth(method = "loess", se = FALSE, color = "red", linewidth = 0.8, formula = y ~ x
  labs(
    title = "Residuos vs. Valores Ajustados",
    x = "Valores Ajustados",
    y = "Residuos"
  ) +
  theme_classic(base_size = 12)
```

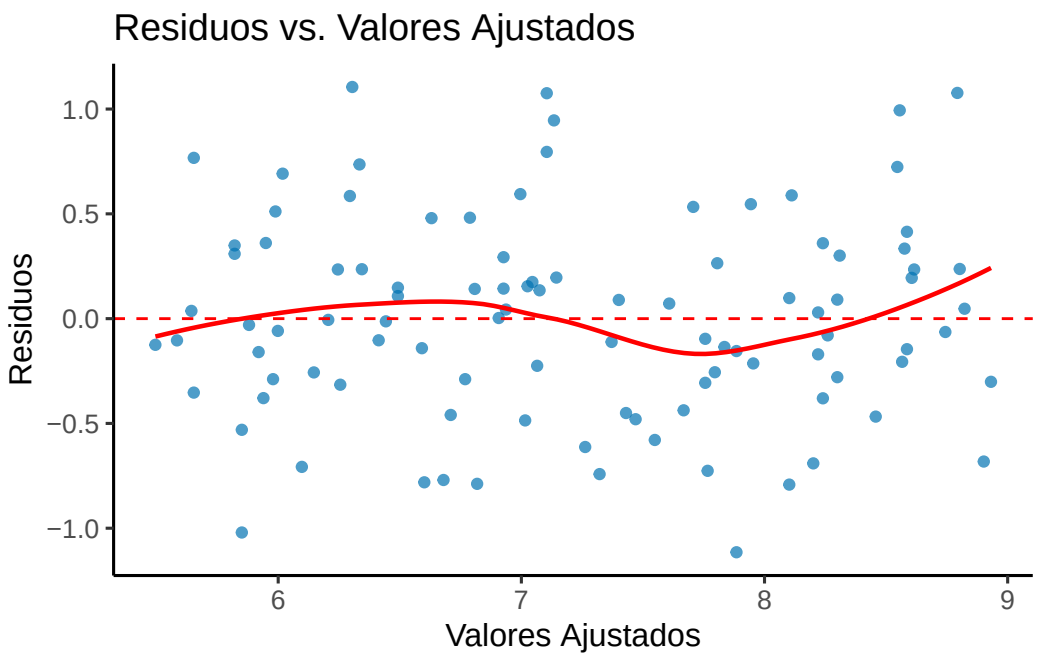


Figure 2.3: Gráfico de Residuos vs. Valores Ajustados para el modelo de estudio. No se observan patrones.

Como se puede observar, los puntos se distribuyen de forma aleatoria alrededor de la línea horizontal en cero. La línea roja, que suaviza la tendencia de los residuos, es prácticamente plana. Esto es un claro indicativo de que el supuesto de linealidad se cumple en nuestro modelo.

Por el contrario, la aparición de un **patrón sistemático** en los residuos es la señal de alarma de que algo anda mal. En lo que respecta al supuesto de **linealidad**, la evidencia más clara de una violación es una **tendencia curvilínea** (como una "U" o una parábola). Este patrón nos dice que el modelo es estructuralmente incapaz de capturar la forma de los datos y, por lo tanto, comete errores predecibles. Por ejemplo, puede subestimar la respuesta en los extremos (generando residuos positivos) y sobreestimarla en el centro (residuos negativos), lo que invalida el modelo lineal.

Contraejemplo: Violación del supuesto de linealidad

Ahora, vamos a simular a propósito unos datos que siguen una relación cuadrática (curva) y ajustaremos incorrectamente un modelo lineal para ver cómo se manifiesta el problema en el gráfico de diagnóstico.

```r
# 1. Simulación de datos no lineales
set.seed(42) # Nueva semilla para este ejemplo
x_no_lineal <- runif(100, 0, 10)
# La relación verdadera es cuadrática (y = 10 - (x-5)^2) más un error
y_no_lineal <- 10 - (x_no_lineal - 5)^2 + rnorm(100, 0, 4)
datos_no_lineal <- data.frame(x = x_no_lineal, y = y_no_lineal)

# 2. Ajuste de un modelo lineal (incorrecto)
modelo_no_lineal <- lm(y ~ x, data = datos_no_lineal)

# 3. Gráfico de Residuos vs. Valores Ajustados con ggplot2
datos_diag_no_lineal <- data.frame(
  residuos = residuals(modelo_no_lineal),
  valores_ajustados = fitted(modelo_no_lineal)
)

ggplot(datos_diag_no_lineal, aes(x = valores_ajustados, y = residuos)) +
  geom_point(color = "#D55E00", alpha = 0.7) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  geom_smooth(method = "loess", se = FALSE, color = "red", linewidth = 0.8, formula = y ~ x
  labs(
    title = "Residuos vs. Valores Ajustados (Violación de Linealidad)",
    x = "Valores Ajustados",
    y = "Residuos"
  ) +
  theme_classic(base_size = 12)
```

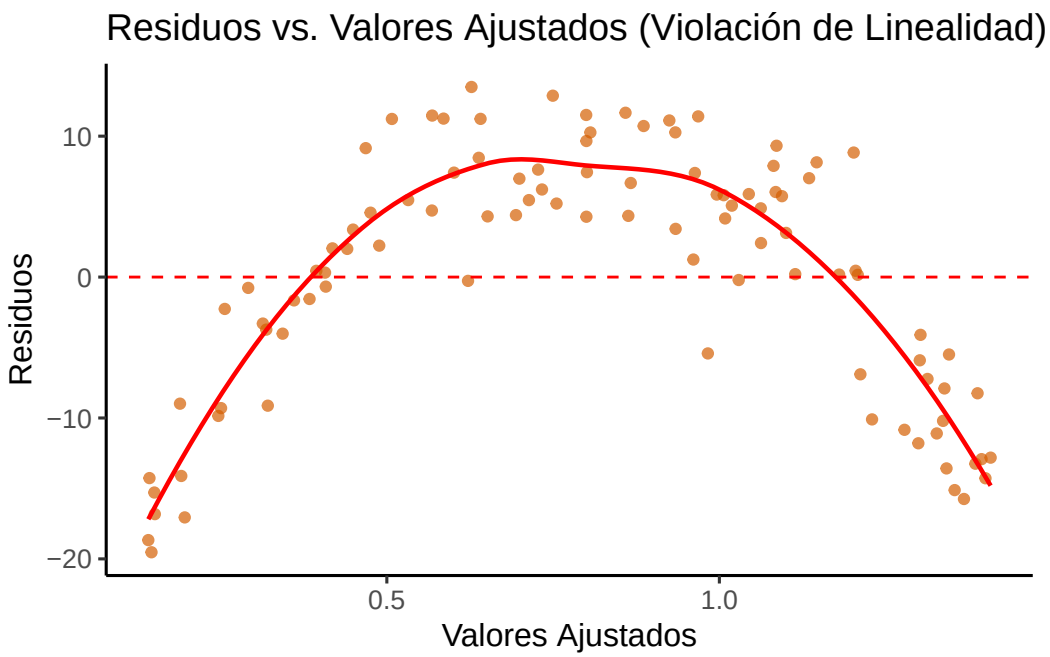


Figure 2.4: Patrón curvo evidente en los residuos, violando el supuesto de linealidad.

El gráfico de diagnóstico es inequívoco. A diferencia del ejemplo anterior, donde los puntos formaban una nube aleatoria, aquí los residuos dibujan un **patrón parabólico** perfecto (una "U" invertida). La línea roja de tendencia, en lugar de ser plana, sigue fielmente esta curva.

Test de Ramsey RESET para Linealidad

Además del diagnóstico visual, podemos usar el test de Ramsey RESET (Regression Equation Specification Error Test) para detectar violaciones de linealidad:

```
suppressPackageStartupMessages(library(lmtest))
# Test RESET para el modelo correcto (horas de estudio)
reset_resultado <- resettest(modelo_estudio, power = 2:3, type = "fitted")
print(reset_resultado)
```

```
    RESET test

data:  modelo_estudio
RESET = 1.0513, df1 = 2, df2 = 96, p-value = 0.3535
```

```r
# Test RESET para el modelo incorrecto (no lineal)
reset_no_lineal <- resettest(modelo_no_lineal, power = 2:3, type = "fitted")
print(reset_no_lineal)
```

```
    RESET test

data:  modelo_no_lineal
RESET = 231.9, df1 = 2, df2 = 96, p-value < 2.2e-16
```

El test de Ramsey RESET confirma nuestro diagnóstico visual:

- Para el modelo correcto: p-valor = 0.3535 → No rechazamos H (forma funcional correcta)
- Para el modelo no lineal: p-valor = 0 → Rechazamos H (forma funcional incorrecta)

2.6.2 Homocedasticidad

El supuesto de homocedasticidad establece que la varianza de los errores del modelo debe ser constante para todos los niveles de la variable predictora. Es decir, la dispersión de los datos alrededor de la línea de regresión es la misma en todo su recorrido ($Var(\varepsilon_i|X_i) = \sigma^2$). La violación de este supuesto se conoce como **heterocedasticidad**, y es un problema común en el modelado.

Métodos de diagnóstico:

1. **Diagnóstico visual:** El gráfico **Scale-Location** muestra la raíz cuadrada de los residuos estandarizados absolutos frente a los valores ajustados. Una línea horizontal indica homocedasticidad.
2. **Test de Breusch-Pagan:** Test clásico para detectar heterocedasticidad.
   - **H :** Homocedasticidad (varianza constante)
   - **H :** Heterocedasticidad
   - **Estadístico:** LM ~ ² bajo H
3. **Test de White:** Versión más robusta que no asume una forma específica de heterocedasticidad.
   - **H :** Homocedasticidad
   - **H :** Heterocedasticidad (forma general)
   - **Ventaja:** No requiere especificar la forma funcional de la heterocedasticidad

¿Por qué es tan importante? Si un modelo es heteroscedástico, los errores estándar de los coeficientes (β_0, β_1) estarán calculados de forma incorrecta. Como consecuencia, los intervalos de confianza y los contrastes de hipótesis (p-valores) no serán fiables, pudiendo llevarnos a conclusiones erróneas sobre la significancia de nuestras variables.

> **i** Sobre los residuos estandarizados
>
> Los **residuos simples** ($e_i = y_i - \hat{y}_i$) no son directamente comparables entre sí porque tienen diferentes varianzas dependiendo de su apalancamiento (*leverage*). Por eso, en los gráficos de diagnóstico se utilizan **residuos estandarizados** o, mejor aún, **residuos estudentizados**, que ponen todos los residuos en una escala común. Esto facilita la identificación de patrones y valores atípicos. La explicación detallada de estos conceptos se encuentra en la sección de identificación de observaciones influyentes.

La heteroscedasticidad se detecta principalmente buscando patrones en la dispersión de los residuos.

- **Gráfico de Residuos vs. Valores Ajustados:** Como en la prueba de linealidad, este gráfico es nuestra primera herramienta. Aquí no buscamos patrones en la media de los residuos (que debe ser cero), sino en su **dispersión**. La señal de alarma inequívoca de heteroscedasticidad es una **forma de embudo o megáfono**, donde la dispersión de los residuos aumenta o disminuye a medida que cambian los valores ajustados.
- **Gráfico Scale-Location:** Este gráfico está diseñado específicamente para detectar heteroscedasticidad. Muestra la raíz cuadrada de los residuos estandarizados en el eje Y (`sqrt(|Standardized residuals|)`) frente a los valores ajustados en el eje X. Al usar la raíz cuadrada, se suaviza la distribución de los residuos, haciendo los patrones de varianza más fáciles de ver. Si la varianza es constante (homocedasticidad), deberíamos ver una nube de puntos aleatoria con una línea de tendencia roja aproximadamente plana. Una pendiente en esta línea roja indica que la varianza cambia con el nivel de la respuesta.
- **Prueba de Breusch-Pagan:** Es el contraste de hipótesis formal. Su lógica es ingeniosa: realiza una regresión auxiliar donde intenta predecir los residuos al cuadrado a partir de las variables predictoras originales. Si las variables predictoras ayudan a explicar la magnitud de los residuos al cuadrado, significa que la varianza del error depende de los predictores, y por tanto, hay heteroscedasticidad.
  - **Hipótesis Nula (H_0):** El modelo es homocedástico.
  - **Decisión:** Un p-valor pequeño (p. ej., < 0.05) es evidencia en contra de la homocedasticidad.

Ejemplo de un modelo válido

Analicemos nuestro `modelo_estudio`. Nos centraremos en el gráfico `Scale-Location` (`which = 3`) y en la prueba de Breusch-Pagan.

```
# Crear datos para el gráfico Scale-Location
datos_scale_loc <- data.frame(
  valores_ajustados = fitted(modelo_estudio),
  residuos_std_sqrt = sqrt(abs(rstandard(modelo_estudio)))
)

# Gráfico Scale-Location con ggplot2
ggplot(datos_scale_loc, aes(x = valores_ajustados, y = residuos_std_sqrt)) +
  geom_point(color = "#0072B2", alpha = 0.7) +
  geom_smooth(method = "loess", se = FALSE, color = "red", linewidth = 0.8, formula = y ~ x
  labs(
    title = "Scale-Location",
    x = "Valores Ajustados",
    y = expression(sqrt("|Residuos Estandarizados|"))
  ) +
  theme_classic(base_size = 12)

# Prueba de Breusch-Pagan
suppressPackageStartupMessages(library(lmtest))
bp_resultado <- bptest(modelo_estudio)
print(bp_resultado)
```

```
    studentized Breusch-Pagan test

data:  modelo_estudio
BP = 0.019638, df = 1, p-value = 0.8886
```

```
# Prueba de White (versión robusta)
white_resultado <- bptest(modelo_estudio, ~ fitted(modelo_estudio) + I(fitted(modelo_estudi
print(white_resultado)
```

```
    studentized Breusch-Pagan test

data:  modelo_estudio
BP = 0.12238, df = 2, p-value = 0.9406
```

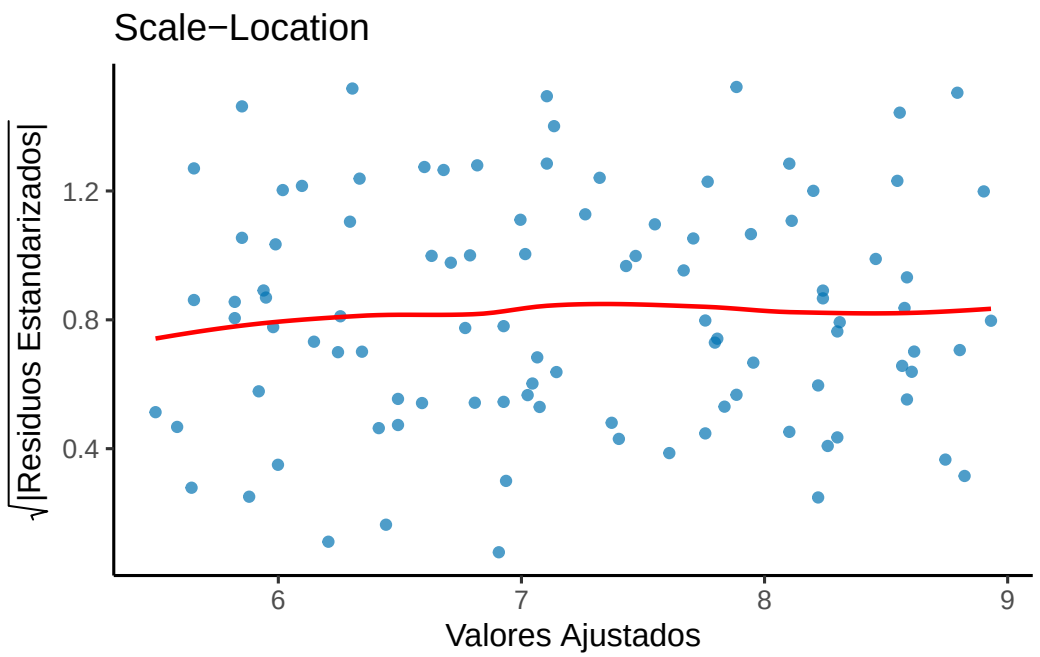


Figure 2.5: Gráfico Scale-Location para el modelo de estudio. La línea de tendencia es casi plana.

El diagnóstico es positivo. En el gráfico `Scale-Location`, la línea roja es casi horizontal, lo que indica que la varianza de los residuos es estable a lo largo de los valores ajustados. Esto se confirma con ambas pruebas:

- **Breusch-Pagan**: p-valor = 0.8886 → No rechazamos H (homocedasticidad)
- **White**: p-valor = 0.9406 → No rechazamos H (varianza constante) **Nuestro modelo cumple el supuesto**.

Contraejemplo: Violación del supuesto de homocedasticidad

Ahora, simularemos datos donde el error aumenta a medida que `x` crece, un caso clásico de heteroscedasticidad.

```r
# 1. Simulación de datos heteroscedásticos
set.seed(101)
x_hetero <- 1:100
y_hetero <- 10 + 2 * x_hetero + rnorm(100, mean = 0, sd = 0.4 * x_hetero)
datos_hetero <- data.frame(x = x_hetero, y = y_hetero)
modelo_hetero <- lm(y ~ x, data = datos_hetero)

# 2. Preparar datos para los gráficos
datos_diag_hetero <- data.frame(
  residuos = residuals(modelo_hetero),
  valores_ajustados = fitted(modelo_hetero),
  residuos_std_sqrt = sqrt(abs(rstandard(modelo_hetero)))
)

# 3. Gráfico de Residuos vs. Valores Ajustados
p1 <- ggplot(datos_diag_hetero, aes(x = valores_ajustados, y = residuos)) +
  geom_point(color = "#D55E00", alpha = 0.7) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  geom_smooth(method = "loess", se = FALSE, color = "red", linewidth = 0.8, formula = y ~ x
  labs(
    title = "Residuos vs. Valores Ajustados",
    x = "Valores Ajustados",
    y = "Residuos"
  ) +
  theme_classic(base_size = 10)

# 4. Gráfico Scale-Location
p2 <- ggplot(datos_diag_hetero, aes(x = valores_ajustados, y = residuos_std_sqrt)) +
  geom_point(color = "#D55E00", alpha = 0.7) +
  geom_smooth(method = "loess", se = FALSE, color = "red", linewidth = 0.8, formula = y ~ x
  labs(
    title = "Scale-Location",
    x = "Valores Ajustados",
    y = expression(sqrt("|Residuos Estandarizados|"))
  ) +
  theme_classic(base_size = 10)

# 5. Mostrar ambos gráficos lado a lado
library(gridExtra)
grid.arrange(p1, p2, ncol = 2)

# 6. Prueba de Breusch-Pagan
suppressPackageStartupMessages(library(lmtest))
test_values <- bptest(modelo_hetero)
```

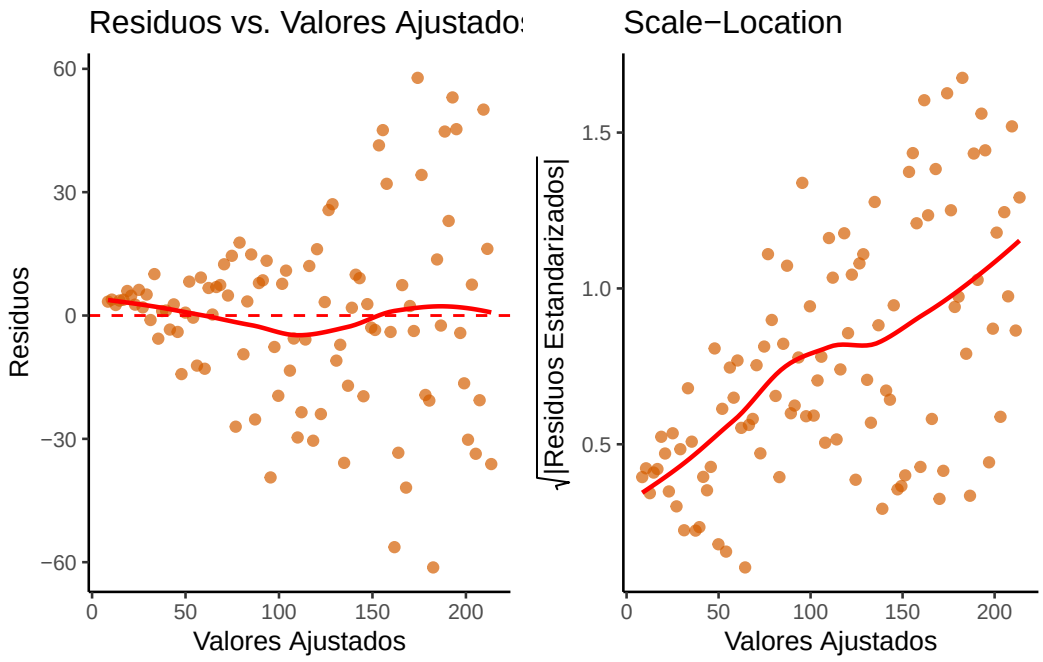


Figure 2.6: Diagnóstico de heteroscedasticidad. Izquierda: Gráfico de Residuos vs. Ajustados (patrón de embudo). Derecha: Gráfico Scale-Location (tendencia ascendente).

Los resultados son un libro de texto sobre la heteroscedasticidad.

- El gráfico de **Residuos vs. Valores Ajustados** (izquierda) tiene una **forma de embudo** inconfundible: la dispersión de los puntos aumenta drásticamente de izquierda a derecha.
- El gráfico **Scale-Location** (derecha) confirma el problema, mostrando una línea roja con una **clara pendiente ascendente**.
- La **prueba de Breusch-Pagan** arroja un **p-valor 7.43e-07**, dándonos una fuerte evidencia estadística para rechazar la hipótesis nula de homocedasticidad.

Este modelo viola claramente el supuesto, y las inferencias basadas en él (como el p-valor del coeficiente de `x`) no serían fiables.

2.6.3 Normalidad de los residuos

Este supuesto postula que los residuos del modelo (ε_i) siguen una distribución normal: $\varepsilon_i \sim N(0, \sigma^2)$. Es especialmente importante para la validez de los intervalos de confianza y los

contrastes de hipótesis cuando el tamaño de la muestra es pequeño.

Métodos de diagnóstico:

1. **Diagnóstico visual:**
   - **Gráfico Q-Q:** Los puntos deben seguir la línea diagonal si hay normalidad
   - **Histograma:** Debe mostrar forma campaniforme y simétrica
2. **Test de Shapiro-Wilk:** Test más potente para muestras pequeñas y medianas (n $<$ 50).
   - **H :** Los residuos siguen distribución normal
   - **H :** Los residuos no siguen distribución normal
   - **Limitación:** Muy sensible en muestras grandes
3. **Test de Jarque-Bera:** Basado en medidas de asimetría y curtosis.
   - **H :** Los residuos son normales (asimetría = 0, curtosis = 3)
   - **H :** Los residuos no son normales
   - **Ventaja:** Menos sensible al tamaño muestral que Shapiro-Wilk

Para evaluar la normalidad disponemos de estas herramientas visuales y analíticas:

- **Gráfico Normal Q-Q (`Normal Q-Q Plot`):** Compara los cuantiles de los residuos estandarizados con los cuantiles de una distribución normal teórica. Los puntos deben caer muy cerca de la línea diagonal de 45 grados.
- **Histograma de los Residuos:** Un simple histograma de los residuos debe mostrar una forma aproximada de campana de Gauss.
- **Prueba de Shapiro-Wilk:** Es uno de los contrastes más potentes para la normalidad.
  - **Hipótesis Nula (H_0):** Los residuos provienen de una distribución normal.
  - **Decisión:** Un p-valor pequeño (< 0.05) sugiere rechazar H_0.

Ejemplo de normalidad válida

Para nuestro `modelo_estudio`, examinamos la normalidad mediante el gráfico Q-Q y la prueba de Shapiro-Wilk.

```r
# Crear datos para los gráficos
residuos <- residuals(modelo_estudio)

# 1. Gráfico Q-Q con ggplot2
datos_qq <- data.frame(residuos = residuos)

p1 <- ggplot(datos_qq, aes(sample = residuos)) +
  geom_qq(color = "#0072B2", alpha = 0.7) +
  geom_qq_line(color = "red", linetype = "dashed") +
  labs(
    title = "Normal Q-Q Plot",
    x = "Cuantiles Teóricos",
    y = "Cuantiles de la Muestra"
  ) +
  theme_classic(base_size = 10)

# 2. Histograma con ggplot2
datos_hist <- data.frame(residuos = residuos)

p2 <- ggplot(datos_hist, aes(x = residuos)) +
  geom_histogram(aes(y = after_stat(density)), bins = 15, fill = "lightblue",
                 color = "black", alpha = 0.7) +
  stat_function(fun = dnorm,
                args = list(mean = mean(residuos), sd = sd(residuos)),
                color = "red", linewidth = 1) +
  labs(
    title = "Histograma de Residuos",
    x = "Residuos",
    y = "Densidad"
  ) +
  theme_classic(base_size = 10)

# 3. Mostrar ambos gráficos lado a lado
library(gridExtra)
grid.arrange(p1, p2, ncol = 2)

# Prueba de Shapiro-Wilk
suppressPackageStartupMessages(library(lmtest))
shapiro_resultado <- shapiro.test(residuals(modelo_estudio))
print(shapiro_resultado)
```

```
    Shapiro-Wilk normality test

data:  residuals(modelo_estudio)
W = 0.99008, p-value = 0.671
```

```r
# Prueba de Jarque-Bera
suppressPackageStartupMessages(library(tseries))
jb_resultado <- jarque.bera.test(residuals(modelo_estudio))
print(jb_resultado)
```

```
    Jarque Bera Test

data:  residuals(modelo_estudio)
X-squared = 0.68515, df = 2, p-value = 0.7099
```

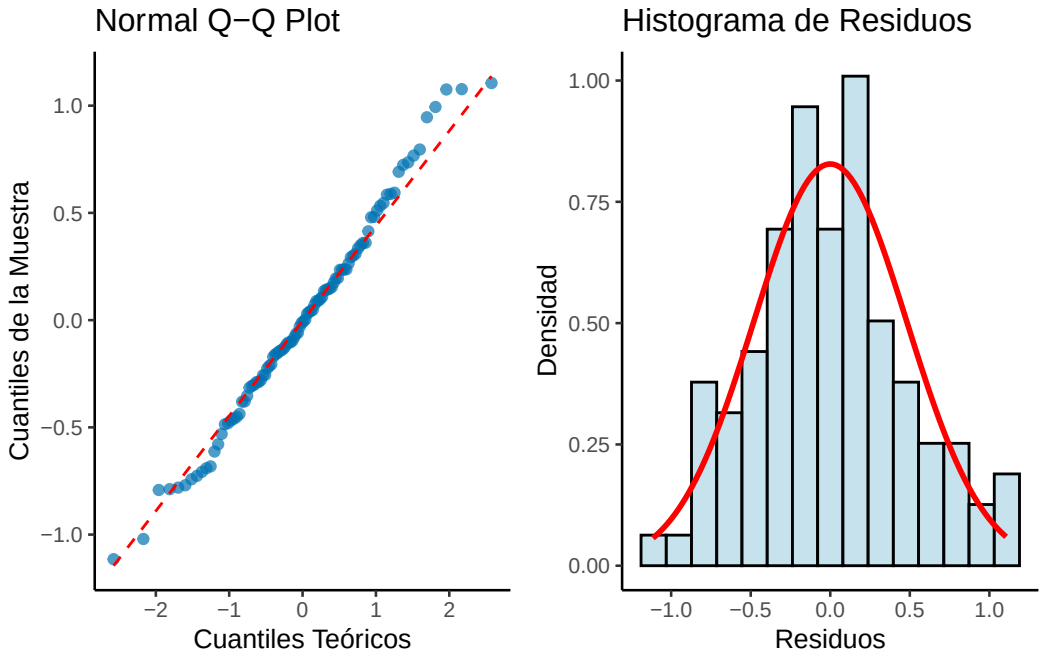


Figure 2.7: Diagnóstico de normalidad del modelo de estudio. Izquierda: Q-Q Plot. Derecha: Histograma de residuos.

El diagnóstico es excelente. En el gráfico Q-Q, los puntos se alinean muy bien con la línea diagonal, indicando normalidad. El histograma muestra una distribución aproximadamente

simétrica que se ajusta bien a la curva normal teórica (línea roja). Ambas pruebas estadísticas confirman la normalidad:

- **Shapiro-Wilk**: p-valor = 0.671 → No rechazamos H (normalidad)
- **Jarque-Bera**: p-valor = 0.7099 → No rechazamos H (normalidad)

Contraejemplo: Violación del supuesto de normalidad

Ahora simularemos datos donde los residuos siguen una distribución asimétrica (distribución exponencial) para mostrar una violación clara del supuesto de normalidad.

```r
# 1. Simulación de datos con errores no normales (exponenciales)
set.seed(456)
x_no_normal <- 1:100
# Errores exponenciales (muy asimétricos) centrados en 0
errores_exp <- rexp(100, rate = 1) - 1  # Restamos 1 para centrar en 0
y_no_normal <- 5 + 2 * x_no_normal + errores_exp * 10
datos_no_normal <- data.frame(x = x_no_normal, y = y_no_normal)
modelo_no_normal <- lm(y ~ x, data = datos_no_normal)

# 2. Crear datos para los gráficos
residuos_no_normal <- residuals(modelo_no_normal)

# 3. Gráfico Q-Q con ggplot2
datos_qq_mal <- data.frame(residuos = residuos_no_normal)

p1_mal <- ggplot(datos_qq_mal, aes(sample = residuos)) +
  geom_qq(color = "#D55E00", alpha = 0.7) +
  geom_qq_line(color = "red", linetype = "dashed") +
  labs(
    title = "Normal Q-Q Plot (Violación)",
    x = "Cuantiles Teóricos",
    y = "Cuantiles de la Muestra"
  ) +
  theme_classic(base_size = 10)

# 4. Histograma con ggplot2
datos_hist_mal <- data.frame(residuos = residuos_no_normal)

p2_mal <- ggplot(datos_hist_mal, aes(x = residuos)) +
  geom_histogram(aes(y = after_stat(density)), bins = 15, fill = "lightcoral",
                 color = "black", alpha = 0.7) +
  stat_function(fun = dnorm,
                args = list(mean = mean(residuos_no_normal), sd = sd(residuos_no_normal)),
                color = "blue", linewidth = 1) +
  labs(
    title = "Histograma de Residuos (Violación)",
    x = "Residuos",
    y = "Densidad"
  ) +
  theme_classic(base_size = 10)

# 5. Mostrar ambos gráficos lado a lado
grid.arrange(p1_mal, p2_mal, ncol = 2)

# Prueba de Shapiro-Wilk
shapiro_resultado <- shapiro.test(residuals(modelo_no_normal))
```

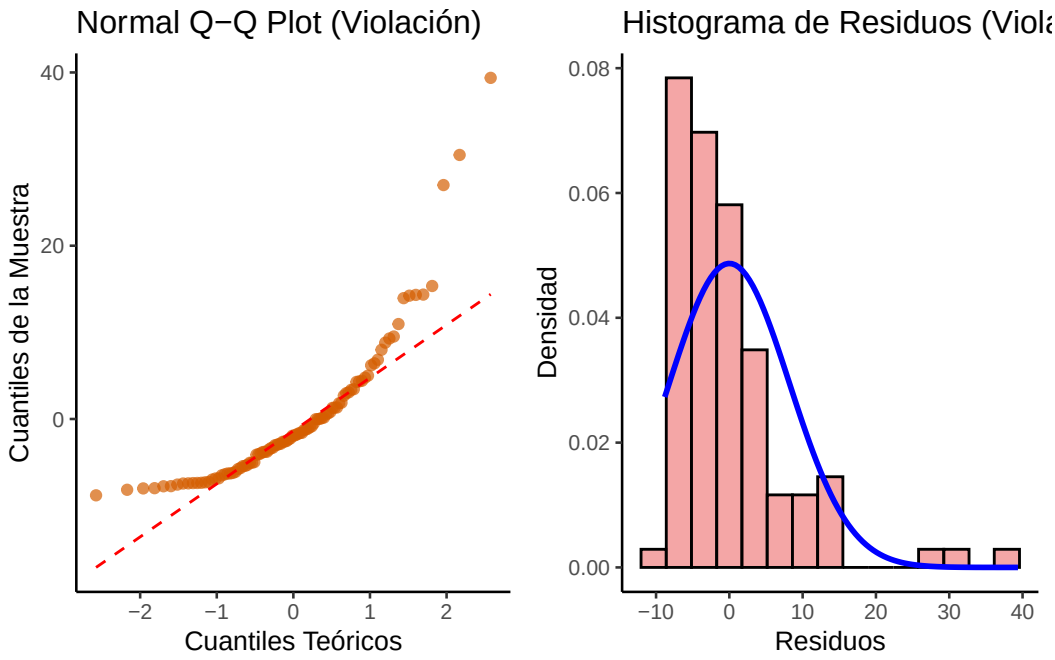


Figure 2.8: Violación del supuesto de normalidad. Izquierda: Q-Q Plot con clara desviación. Derecha: Histograma asimétricamente distribuido.

La violación es evidente. En el gráfico Q-Q, los puntos se desvían sistemáticamente de la línea diagonal, especialmente en los extremos, formando una curva característica de distribuciones asimétricas. El histograma muestra una clara asimetría hacia la derecha que no se ajusta a la curva normal teórica (línea azul). La prueba de Shapiro-Wilk arroja un **p-valor muy pequeño** (1.78e-10), rechazando fuertemente la hipótesis nula de normalidad.

2.6.4 Independencia de los residuos

Este supuesto afirma que el error de una observación no está correlacionado con el de ninguna otra: $Cov(\varepsilon_i, \varepsilon_j) = 0$ para $i \neq j$. La violación, conocida como **autocorrelación**, es común en datos de series temporales.

Métodos de diagnóstico:

1. **Diagnóstico visual:** Gráfico de residuos vs orden de observación. No debe mostrar patrones temporales o secuenciales.
2. **Test de Durbin-Watson:** Test clásico para autocorrelación de primer orden.

- **H :** No hay autocorrelación ($\rho = 0$)
- **H :** Hay autocorrelación de orden 1
- **Estadístico:** $DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$
- **Interpretación:** Valores cerca de 2 → no autocorrelación; cerca de 0 → autocorrelación positiva; cerca de 4 → autocorrelación negativa

3. **Test de Breusch-Godfrey (LM):** Generalización del Durbin-Watson para órdenes superiores y regresores retardados.

   - **H :** No hay autocorrelación serial hasta el orden especificado
   - **H :** Hay autocorrelación serial
   - **Ventaja:** Más general y potente que Durbin-Watson

El estadístico de Durbin-Watson varía entre 0 y 4. Un valor cercano a 2 sugiere no autocorrelación. Valores cercanos a 0 indican autocorrelación positiva, y cercanos a 4, autocorrelación negativa.

Ejemplo de independencia válida

Para nuestro `modelo_estudio`, evaluamos la independencia mediante el gráfico de residuos vs orden y la prueba de Durbin-Watson.

```
# Gráfico de residuos vs orden de observación con ggplot2
datos_orden <- data.frame(
  orden = 1:length(residuals(modelo_estudio)),
  residuos = residuals(modelo_estudio)
)

ggplot(datos_orden, aes(x = orden, y = residuos)) +
  geom_point(color = "#0072B2", alpha = 0.7) +
  geom_line(color = "#0072B2", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  labs(
    title = "Residuos vs Orden de Observación",
    x = "Orden de observación",
    y = "Residuos"
  ) +
  theme_classic(base_size = 12)

# Prueba de Durbin-Watson
suppressPackageStartupMessages(library(lmtest))
dw_resultado_valido <- dwtest(modelo_estudio)
print(dw_resultado_valido)
```

```
    Durbin-Watson test

data:  modelo_estudio
DW = 2.0565, p-value = 0.6104
alternative hypothesis: true autocorrelation is greater than 0
```

```r
# Prueba de Breusch-Godfrey (más general)
bg_resultado <- bgtest(modelo_estudio, order = 2)
print(bg_resultado)
```

```
    Breusch-Godfrey test for serial correlation of order up to 2

data:  modelo_estudio
LM test = 0.14002, df = 2, p-value = 0.9324
```

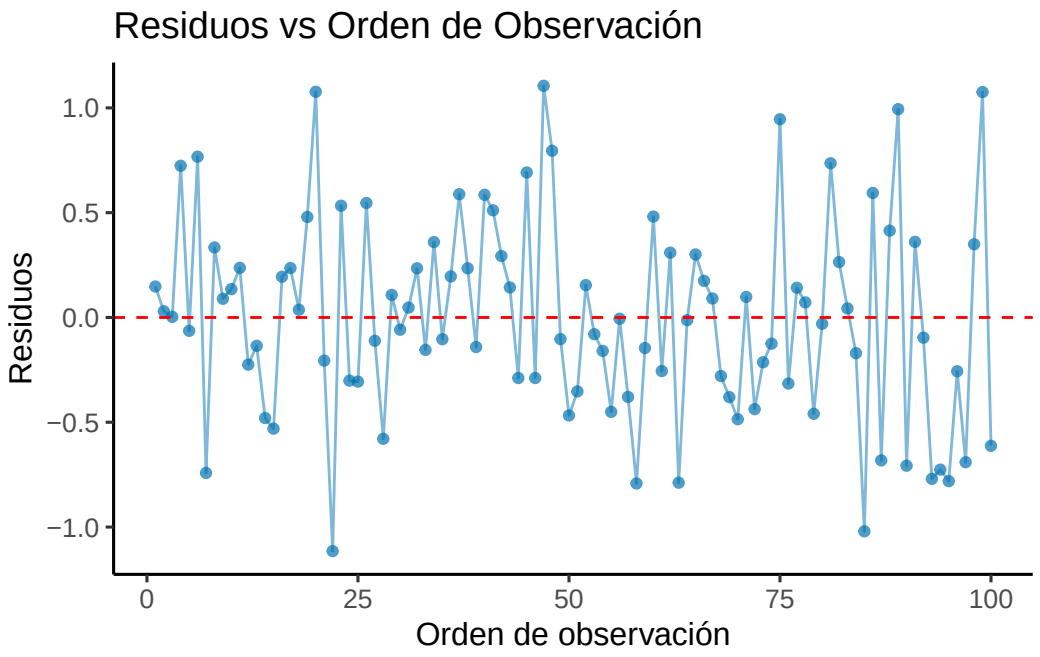


Figure 2.9: Diagnóstico de independencia del modelo de estudio. Los residuos no muestran patrones temporales.

El diagnóstico es satisfactorio. El gráfico de residuos vs orden no muestra ningún patrón sistemático o tendencia temporal. Ambas pruebas estadísticas confirman la independencia:

- **Durbin-Watson**: DW =2.056, p-valor = 0.6104 → No hay autocorrelación de orden 1
- **Breusch-Godfrey**: LM =0.14, p-valor = 0.9324 → No hay autocorrelación de orden 2 fluctúan aleatoriamente alrededor del cero.

La prueba de Durbin-Watson arroja un estadístico de **2.056** (cercano a 2) y un **p-valor de 0.61**, confirmando que no hay evidencia de autocorrelación. **El supuesto de independencia se cumple**.

Contraejemplo: Violación del supuesto de independencia

Simularemos datos con autocorrelación positiva, donde cada residuo está correlacionado con el anterior, violando el supuesto de independencia.

```r
# 1. Simulación de datos con autocorrelación
set.seed(789)
n <- 100
x_autocorr <- 1:n
# Generamos errores autocorrelacionados (AR1 con phi = 0.7)
errores_autocorr <- numeric(n)
errores_autocorr[1] <- rnorm(1)
for(i in 2:n) {
  errores_autocorr[i] <- 0.7 * errores_autocorr[i-1] + rnorm(1, sd = 0.5)
}

y_autocorr <- 10 + 1.5 * x_autocorr + errores_autocorr * 3
datos_autocorr <- data.frame(x = x_autocorr, y = y_autocorr)
modelo_autocorr <- lm(y ~ x, data = datos_autocorr)

# 2. Gráfico de residuos vs orden con ggplot2
datos_orden_autocorr <- data.frame(
  orden = 1:length(residuals(modelo_autocorr)),
  residuos = residuals(modelo_autocorr)
)

ggplot(datos_orden_autocorr, aes(x = orden, y = residuos)) +
  geom_point(color = "#D55E00", alpha = 0.7) +
  geom_line(color = "#D55E00", alpha = 0.5) +
  geom_hline(yintercept = 0, color = "blue", linetype = "dashed") +
  labs(
    title = "Residuos vs Orden de Observación (Violación)",
    x = "Orden de observación",
    y = "Residuos"
  ) +
  theme_classic(base_size = 12)

# 3. Prueba de Durbin-Watson
dw_resultado <- dwtest(modelo_autocorr)
```

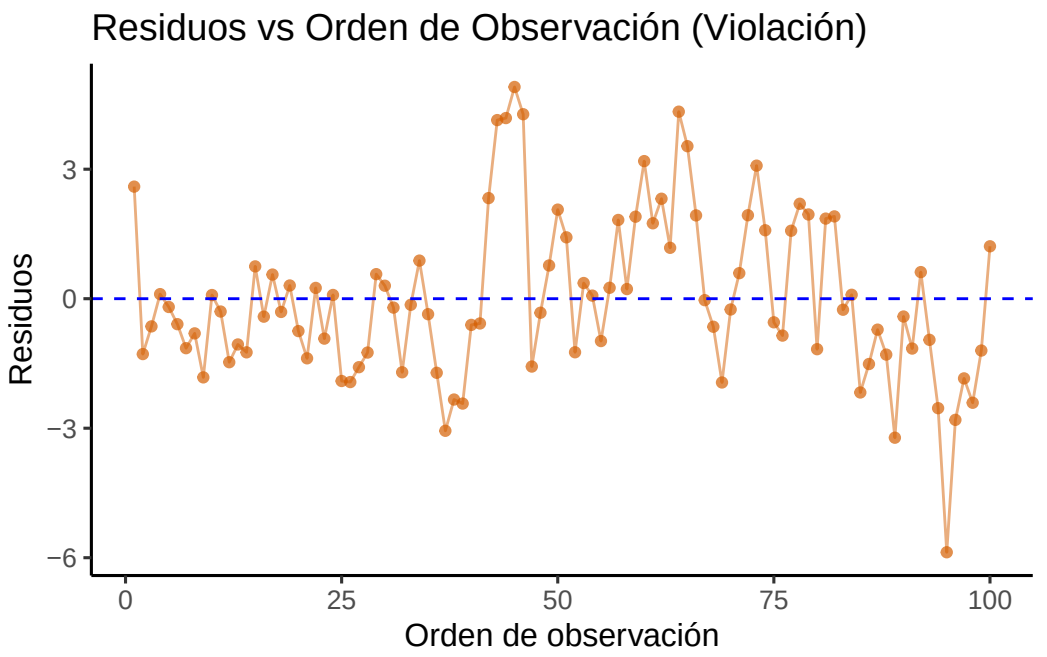


Figure 2.10: Violación del supuesto de independencia. Los residuos muestran un patrón de autocorrelación positiva.

La violación es clara. El gráfico de residuos vs orden muestra un patrón ondulante característico: los residuos tienden a mantenerse del mismo signo durante varias observaciones consecutivas (rachas de valores positivos seguidas de rachas de valores negativos). Esto indica **autocorrelación positiva**. La prueba de Durbin-Watson confirma esto con un estadístico muy por debajo de 2 (DW = **0.74**) y un **p-valor muy pequeño** (**5.01e-11**), rechazando fuertemente la hipótesis nula de independencia.

2.6.5 Media nula de los residuos

Un requisito fundamental del modelo es que la media de los residuos debe ser exactamente cero: $E[e_i] = 0$. Esta propiedad se deriva matemáticamente del método de mínimos cuadrados y su verificación sirve como una comprobación de que nuestros cálculos son correctos.

2.6.6 Identificación de observaciones influyentes y atípicas

Algunos puntos pueden tener una influencia desproporcionada en el modelo. Es crucial identificarlos usando diferentes métricas que evalúan aspectos complementarios de la influencia

(Kutner et al. 2005; Fox and Weisberg 2018). Las métricas desarrolladas por Cook, Belsley, Kuh y Welsch proporcionan herramientas robustas para este diagnóstico.

i Fundamento teórico: de los residuos simples a los estudentizados

Antes de analizar las métricas de influencia, debemos entender por qué no todos los **residuos simples** ($e_i = y_i - \hat{y}_i$) son comparables entre sí. El problema fundamental es que **no tienen la misma varianza**, incluso bajo homocedasticidad.
La varianza teórica del residuo e_i depende del **apalancamiento** (*leverage*) de la observación:

$$\text{Var}(e_i) = \sigma^2(1 - h_{ii})$$

donde el apalancamiento h_{ii} se define como:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

Las observaciones con valores de X más alejados de la media tendrán mayor apalancamiento y, paradójicamente, residuos con **menor varianza**. Por esto, un residuo pequeño en una observación de alto leverage puede ser más preocupante que un residuo grande en el centro de los datos.
Los residuos estandarizados solucionan parcialmente este problema:

$$r_i^* = \frac{e_i}{\sqrt{\text{MSE}(1 - h_{ii})}}$$

Pero **los residuos estudentizados** van un paso más allá, eliminando el sesgo de autoinfluencia:

$$r_i = \frac{e_i}{\sqrt{\text{MSE}_{(-i)}(1 - h_{ii})}}$$

donde $\text{MSE}_{(-i)}$ excluye la observación i del ajuste. Esto evita que un outlier "contamine" su propia evaluación y proporciona una distribución teórica exacta (t de Student con $n - k - 2$ grados de libertad).
¿Por qué son superiores los residuos estudentizados? Por tres razones clave: (1) **eliminan el sesgo de autoinfluencia** al excluir cada observación de su propia evaluación, (2) **evitan la contaminación** que un outlier produce en la MSE global, y (3) **siguen una distribución conocida** (t de Student) que permite umbrales estadísticamente precisos.
En la práctica: $|r_i| > 2$ indica posibles outliers (5% en normalidad) y $|r_i| > 3$ outliers muy probables (<1%).

Las métricas fundamentales de influencia para identificar observaciones problemáticas son:

- **Apalancamiento (Leverage,** h_{ii}**)**: Mide cuán atípico es el valor de la variable predictora X_i de una observación. Un apalancamiento alto significa que el punto tiene el *potencial*

de ser muy influyente. En regresión simple, se calcula como:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

Una regla común es considerar un apalancamiento alto si $h_{ii} > \frac{2(k+1)}{n}$, donde k es el número de predictores (1 en regresión simple).

- **Distancia de Cook (D_i)**: Mide la influencia global de una observación, combinando su apalancamiento y su residuo. Representa cuánto cambian los coeficientes del modelo si la i-ésima observación es eliminada.

$$D_i = \frac{r_i^2}{k+1} \cdot \frac{h_{ii}}{1 - h_{ii}}$$

Se considera que un punto es influyente si su distancia de Cook es grande, por ejemplo, si $D_i > 1$ o $D_i > 4/(n-k-1)$.

- **DFFITS:** Mide cuánto cambia la predicción $\hat{y}_i$ cuando se elimina la i-ésima observación. Es una medida estandarizada que combina el residuo estudentizado y el apalancamiento.

$$\text{DFFITS}_i = r_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

Un punto se considera influyente si $|\text{DFFITS}_i| > 2\sqrt{(k+1)/n}$, donde k es el número de predictores.

Ejemplo: Cálculo y análisis de DFFITS

DFFITS es especialmente útil para evaluar cómo cada observación afecta a su propia predicción. Analicemos esta medida con nuestro `modelo_estudio`.

```r
# Calcular DFFITS y sus componentes
dffits_vals <- dffits(modelo_estudio)
residuos_stud <- rstudent(modelo_estudio)  # Residuos estudentizados
leverage_vals <- hatvalues(modelo_estudio)

# Crear dataframe para análisis
datos_dffits <- data.frame(
  observacion = 1:length(dffits_vals),
  dffits = dffits_vals,
  residuo_stud = residuos_stud,
  leverage = leverage_vals
)

# Umbral de DFFITS
n <- nrow(datos)
k <- 1  # número de predictores
dffits_threshold <- 2 * sqrt((k + 1) / n)

# Gráfico de DFFITS
ggplot(datos_dffits, aes(x = observacion, y = dffits)) +
  geom_point(color = "#0072B2", alpha = 0.7) +
  geom_hline(yintercept = 0, color = "gray", linetype = "dashed") +
  geom_hline(yintercept = c(-dffits_threshold, dffits_threshold),
             color = "red", linetype = "dashed", alpha = 0.7) +
  labs(
    title = "DFFITS por Observación",
    x = "Número de Observación",
    y = "DFFITS",
    caption = paste("Líneas rojas: umbrales ±", round(dffits_threshold, 3))
  ) +
  theme_classic(base_size = 12)

# Análisis cuantitativo
influential_dffits <- which(abs(dffits_vals) > dffits_threshold)
top_indices <- order(abs(dffits_vals), decreasing = TRUE)[1:5]
```

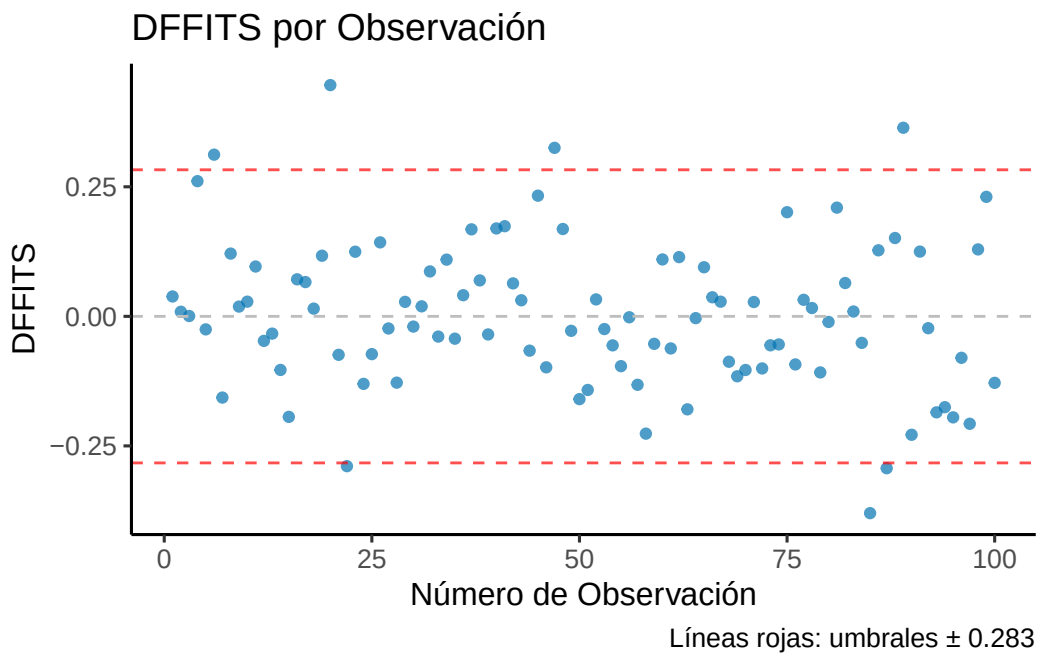


Figure 2.11: Análisis de DFFITS para identificar observaciones que afectan significativamente a sus propias predicciones.

Análisis de resultados:
El umbral de influencia para DFFITS es 0.283. En nuestro modelo, **7 observaciones superan este umbral**: las observaciones 6, 20, 22, 47, 85, 87, 89, lo que las clasifica como influyentes según este criterio.
Las **cinco observaciones con mayor |DFFITS|** son las observaciones 20, 85, 89, 47, 6, con valores de 0.446, -0.38, 0.364, 0.325, 0.312 respectivamente. Lo más notable es que **todas estas cinco observaciones** (20, 85, 89, 47, 6) **superan el umbral de DFFITS**, confirmando su carácter influyente.
Interpretación clave: La observación 20 es el caso más destacado: tiene un DFFITS de 0.446, superando el umbral de 0.283. Esta combinación de residuo y apalancamiento resulta en un DFFITS significativo que indica cambios sustanciales en su predicción.
Conclusión práctica: Tenemos **7 observaciones influyentes** según DFFITS (6, 20, 22, 47, 85, 87, 89) que merecen investigación adicional. Estas observaciones cambian significativamente sus propias predicciones cuando son eliminadas del modelo, sugiriendo que podrían representar casos especiales o errores de medición que deberían ser examinados más detalladamente.

El gráfico `Residuals vs. Leverage` es la herramienta visual más importante para el diagnóstico de influencia, ya que combina en un solo gráfico el **apalancamiento** (eje X) y los **residuos estudentizados** (eje Y), permitiendo identificar simultáneamente observaciones con

alto leverage y outliers. Además, incluye curvas que delimitan regiones de alta **Distancia de Cook**, facilitando la identificación visual de los puntos más problemáticos.

Ejemplo: Gráfico Residuals vs. Leverage

Vamos a analizar el gráfico más importante para el diagnóstico de influencia usando nuestro `modelo_estudio`.

```r
# Crear datos para el gráfico Residuals vs. Leverage
leverage_vals <- hatvalues(modelo_estudio)
residuos_stud <- rstudent(modelo_estudio)  # Residuos estudentizados
cook_dist <- cooks.distance(modelo_estudio)

datos_leverage <- data.frame(
  leverage = leverage_vals,
  residuos_stud = residuos_stud,
  cook = cook_dist,
  observacion = 1:length(leverage_vals)
)

# Calcular umbrales
n <- nrow(datos)
k <- 1
leverage_threshold <- 2 * (k + 1) / n
cook_threshold <- 4 / (n - k - 1)

# Función para crear curvas de Cook
cook_curve <- function(leverage, cook_value, k) {
  sqrt(cook_value * (k + 1) * (1 - leverage) / leverage)
}

# Crear curvas de Cook para diferentes valores
lev_seq <- seq(0.001, max(leverage_vals) * 1.1, length.out = 100)
cook_05 <- data.frame(
  leverage = lev_seq,
  pos = cook_curve(lev_seq, 0.5, k),
  neg = -cook_curve(lev_seq, 0.5, k)
)
cook_1 <- data.frame(
  leverage = lev_seq,
  pos = cook_curve(lev_seq, 1, k),
  neg = -cook_curve(lev_seq, 1, k)
)

# Gráfico Residuals vs. Leverage con ggplot2
ggplot(datos_leverage, aes(x = leverage, y = residuos_stud)) +
  # Curvas de Cook
  geom_line(data = cook_05, aes(x = leverage, y = pos),
            color = "red", linetype = "dashed", alpha = 0.6, inherit.aes = FALSE) +
  geom_line(data = cook_05, aes(x = leverage, y = neg),
            color = "red", linetype = "dashed", alpha = 0.6, inherit.aes = FALSE) +
  geom_line(data = cook_1, aes(x = leverage, y = pos),
            color = "red", linetype = "solid", alpha = 0.8, inherit.aes = FALSE) +
  geom_line(data = cook_1, aes(x = leverage, y = neg),
            color = "red", linetype = "solid", alpha = 0.8, inherit.aes = FALSE) +
  # Puntos de datos
  geom_point(color = "#0072B2", alpha = 0.7, size = 2) +
  # Líneas de referencia
  geom_hline(yintercept = 0, color = "gray", linetype = "dashed") +
  geom_hline(yintercept = c(-2, 2), color = "orange", linetype = "dotted", alpha = 0.7) +
```

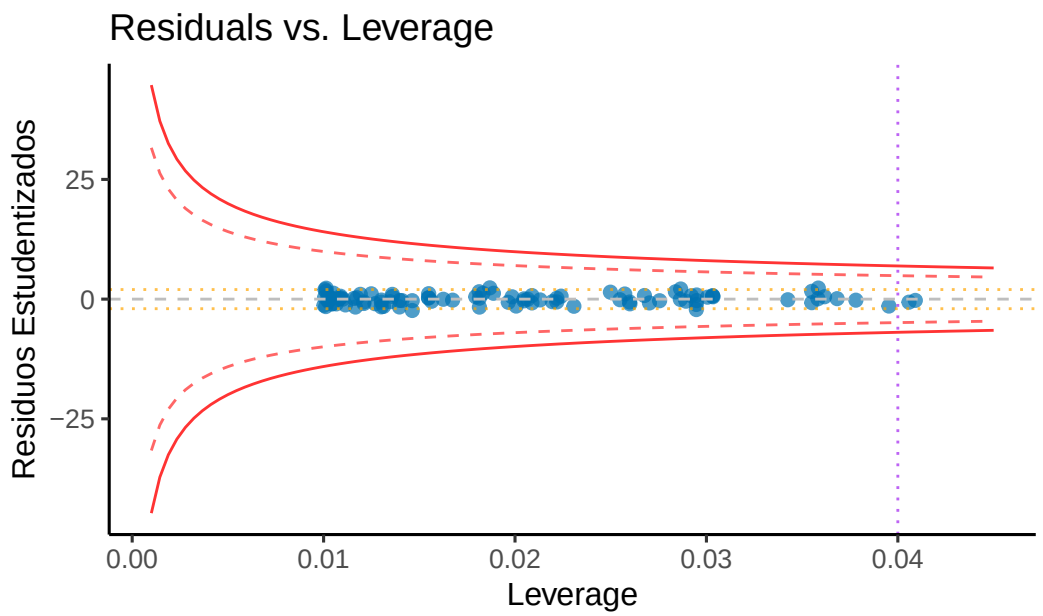


Figure 2.12: Gráfico Residuals vs. Leverage para identificar observaciones influyentes.

Análisis del gráfico:
El gráfico revela varios puntos importantes. Tenemos 6 outliers (residuos estudentizados > 2): las observaciones 20, 22, 47, 85, 89, 99. Además, 2 observaciones superan el umbral de leverage (> 0.04): las observaciones 24, 74.
Interpretación por regiones:

- **Zona derecha (alto leverage)**: Las observaciones 24, 74 superan el umbral de leverage, lo que significa que tienen valores de X atípicos y **alto potencial influyente**
- **Zona izquierda superior/inferior**: Los 6 outliers (20, 22, 47, 85, 89, 99) están distribuidos aquí, con leverage bajo-moderado pero residuos grandes
- **Esquinas críticas**: Afortunadamente vacías (alto leverage + outlier sería muy problemático)

Distancia de Cook: Las curvas rojas muestran que aunque ningún punto supera Cook = 1.0 (línea continua), **varios puntos se acercan a la curva de Cook = 0.5** (línea discontinua), indicando influencia moderada. Las observaciones con alto leverage están en esta zona de influencia moderada.
Conclusión práctica: El modelo presenta una **situación favorable**: aunque tenemos **outliers** (observaciones 20, 22, 47, 85, 89, 99) que son atípicos en Y, y **observaciones**

de alto leverage (observaciones 24, 74) que son atípicos en X, **crucialmente no hay solapamiento entre ambos grupos**. Esto significa que no tenemos la situación más problemática (alto leverage + outlier). Aun así, ambos grupos merecen investigación.

2.6.6.1 Interpretación práctica de las medidas de influencia

Cada medida nos proporciona información complementaria sobre diferentes aspectos de la influencia:

- **Leverage (Apalancamiento):** Identifica observaciones con valores "raros" en las variables predictoras. Alto leverage no es necesariamente problemático, pero indica potencial para ser influyente.
- **Distancia de Cook:** Es la medida más general de influencia. Valores altos indican que eliminar esa observación cambiaría substancialmente los coeficientes del modelo.
- **DFFITS:** Se enfoca específicamente en cómo cambia la predicción de cada punto cuando se elimina esa observación. Es especialmente útil para evaluar el impacto en las predicciones.

En la práctica, una observación es especialmente preocupante si es problemática según **múltiples criterios** a la vez.

Diagnóstico completo del modelo de estudio

A continuación, realizamos todas las verificaciones de diagnóstico para nuestro `modelo_estudio`:

```r
# Preparar todos los datos necesarios para los gráficos
residuos_completo <- residuals(modelo_estudio)
valores_ajustados_completo <- fitted(modelo_estudio)
residuos_std <- rstandard(modelo_estudio)
leverage_vals <- hatvalues(modelo_estudio)

# 1. Gráfico Residuos vs. Valores Ajustados
p1_completo <- ggplot(data.frame(x = valores_ajustados_completo, y = residuos_completo),
                      aes(x = x, y = y)) +
  geom_point(color = "#0072B2", alpha = 0.7) +
  geom_hline(yintercept = 0, color = "red", linetype = "dashed") +
  geom_smooth(method = "loess", se = FALSE, color = "red", linewidth = 0.8, formula = y ~ x
  labs(title = "Residuos vs. Valores Ajustados", x = "Valores Ajustados", y = "Residuos") +
  theme_classic(base_size = 10)

# 2. Gráfico Q-Q Normal
datos_qq_completo <- data.frame(residuos = residuos_std)

p2_completo <- ggplot(datos_qq_completo, aes(sample = residuos)) +
  geom_qq(color = "#0072B2", alpha = 0.7) +
  geom_qq_line(color = "red", linetype = "dashed") +
  labs(title = "Normal Q-Q Plot", x = "Cuantiles Teóricos", y = "Cuantiles de la Muestra") +
  theme_classic(base_size = 10)

# 3. Gráfico Scale-Location
p3_completo <- ggplot(data.frame(x = valores_ajustados_completo,
                                 y = sqrt(abs(residuos_std))),
                      aes(x = x, y = y)) +
  geom_point(color = "#0072B2", alpha = 0.7) +
  geom_smooth(method = "loess", se = FALSE, color = "red", linewidth = 0.8, formula = y ~ x
  labs(title = "Scale-Location", x = "Valores Ajustados",
       y = expression(sqrt("|Residuos Estandarizados|"))) +
  theme_classic(base_size = 10)

# 4. Gráfico Residuos vs. Leverage (con curvas de Cook)
p4_completo <- ggplot(datos_leverage, aes(x = leverage, y = residuos_stud)) +
  # Curvas de Cook
  geom_line(data = cook_05, aes(x = leverage, y = pos),
            color = "red", linetype = "dashed", alpha = 0.6, inherit.aes = FALSE) +
  geom_line(data = cook_05, aes(x = leverage, y = neg),
            color = "red", linetype = "dashed", alpha = 0.6, inherit.aes = FALSE) +
  geom_line(data = cook_1, aes(x = leverage, y = pos),
            color = "red", linetype = "solid", alpha = 0.8, inherit.aes = FALSE) +
  geom_line(data = cook_1, aes(x = leverage, y = neg),
            color = "red", linetype = "solid", alpha = 0.8, inherit.aes = FALSE) +
  # Puntos de datos
  geom_point(color = "#0072B2", alpha = 0.7, size = 1.5) +
  # Líneas de referencia
  geom_hline(yintercept = 0, color = "gray", linetype = "dashed") +
  geom_hline(yintercept = c(-2, 2), color = "orange", linetype = "dotted", alpha = 0.7) +
  geom_vline(xintercept = leverage_threshold, color = "purple", linetype = "dotted", alpha
  labs(title = "Residuals vs. Leverage", x = "Leverage", y = "Residuos Estudentizados") +
  theme_classic(base_size = 10)
```

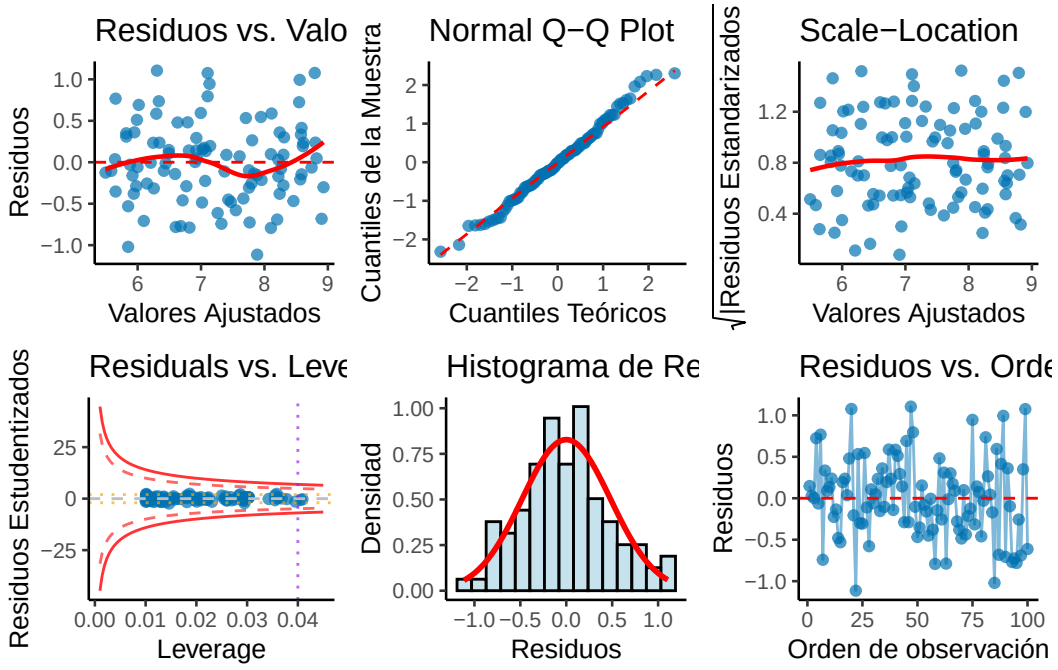


Figure 2.13: Gráficos de diagnóstico completo del modelo de regresión.

Conclusión del diagnóstico:
Nuestro modelo de estudio pasa exitosamente todas las verificaciones:

Linealidad: Los gráficos de residuos vs valores ajustados no muestran patrones sistemáticos, confirmando que la relación lineal es apropiada.

Homocedasticidad: La prueba de Breusch-Pagan arroja un p-valor de 0.8886, que es mayor a 0.05, por lo que no hay evidencia de heterocedasticidad. La varianza de los errores es constante.

Normalidad: La prueba de Shapiro-Wilk presenta un p-valor de 0.671, superior a 0.05, confirmando que los residuos siguen una distribución normal. Esto se corrobora visualmente en el Q-Q plot.

Independencia: El estadístico de Durbin-Watson es 2.056 con un p-valor de 0.61, indicando ausencia de autocorrelación en los residuos.

Media nula: La media de los residuos es 0, prácticamente cero como se esperaría en un modelo bien especificado.

```
Se identificaron las siguientes observaciones que requieren atención:
• Alto apalancamiento (> 0.04 ): 24, 74
• Influyentes según Cook (> 0.041 ): 6, 20, 47, 85, 87, 89
• Influyentes según DFFITS (> 0.283 ): 6, 20, 22, 47, 85, 87, 89
• Posibles outliers (|residuo std| > 2): 20, 22, 47, 85, 89, 99
```

```
Estas observaciones deberían ser investigadas más detalladamente antes de proceder con las
```

Esto confirma que nuestras inferencias estadísticas (p-valores, intervalos de confianza) son válidas y confiables (James et al. 2021; Harrell 2015).

3 El modelo de regresión lineal múltiple

El modelo de regresión lineal múltiple constituye la extensión natural y más potente del modelo simple que estudiamos en el capítulo anterior. Mientras que la regresión simple nos permitía examinar la relación entre una variable respuesta y un único predictor, la regresión múltiple nos capacita para **modelar simultáneamente el efecto de múltiples variables predictoras**, una situación mucho más realista en la mayoría de aplicaciones prácticas (Kutner et al. 2005; James et al. 2021; Fox and Weisberg 2018).

En este capítulo profundizaremos en los aspectos únicos de la regresión múltiple que no están presentes en el caso simple: la **interpretación de coeficientes en presencia de otros predictores**, el **diagnóstico específico** del modelo múltiple, y el problema crucial de la **multicolinealidad**. Estos conceptos son fundamentales para desarrollar modelos predictivos robustos y interpretables (Harrell 2015; Draper 1998).

! Objetivos de aprendizaje

Al finalizar este capítulo, 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* ("manteniendo las demás variables constantes").
3. **Realizar inferencia estadística** construyendo intervalos de confianza y contrastes de hipótesis para los parámetros del modelo múltiple.
4. **Evaluar la calidad del ajuste** usando medidas como R^2, R^2 ajustado y la descomposición ANOVA.
5. **Diagnosticar el modelo múltiple**, aplicando técnicas específicas como gráficos CPR y gráficos de regresión parcial.
6. **Identificar y tratar la multicolinealidad**, comprendiendo sus causas, consecuencias y usando el VIF como herramienta de diagnóstico.
7. **Realizar predicciones** con el modelo ajustado, distinguiendo entre intervalos de confianza e intervalos de predicción.

3.1 Formulación teórica del modelo

El paso de la regresión simple a la múltiple es más que una simple adición de términos; es un salto conceptual. Nos permite construir modelos que reflejan mejor la complejidad del mundo real, donde los resultados raramente dependen de una única causa. Al controlar simultáneamente por varios factores, podemos aislar con mayor precisión el efecto de una variable de interés, reduciendo el riesgo de llegar a conclusiones sesgadas por variables omitidas.

3.1.1 El modelo poblacional

Para n observaciones y p variables predictoras, el **modelo poblacional** postula que la relación verdadera entre la variable respuesta Y y los predictores $X_1, X_2, \dots, X_p$ sigue una relación lineal:

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

Donde Y_i es la i-ésima variable respuesta aleatoria, X_{ij} es la i-ésima variable predictora aleatoria del j-ésimo predictor, y ε_i es el término de error aleatorio. Los parámetros $\beta_0, \beta_1, \dots, \beta_p$ son los coeficientes poblacionales verdaderos pero desconocidos.

3.1.2 El modelo muestral

En la práctica, trabajamos con datos observados y estimamos el modelo usando la muestra disponible:

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

Donde $\hat{y}_i$ es la i-ésima predicción, x_{ij} es la i-ésima observación del j-ésimo predictor, y $\hat{\beta}_j$ son los coeficientes estimados. El coeficiente $\hat{\beta}_j$ representa el cambio estimado en la media de Y ante un cambio de una unidad en el predictor X_j, **manteniendo constantes todas las demás variables predictoras del modelo**. Este principio, conocido como *ceteris paribus* (del latín, "lo demás constante"), es la piedra angular de la interpretación en regresión múltiple.

3.1.3 Notación matricial

La notación matricial es fundamental para el desarrollo teórico y computacional. Nos permite expresar el sistema de n ecuaciones de forma compacta y elegante.

Modelo poblacional:

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

donde:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \tilde{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Donde $\tilde{X}$ contiene variables aleatorias (denotadas con mayúsculas X_{ij}).

Modelo muestral:

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

donde:

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

Donde $\mathbf{X}$ contiene datos observados (denotados con minúsculas x_{ij}).

La matriz $\mathbf{X}$ (datos observados) y $\tilde{X}$ (variables aleatorias), ambas de dimensión $n \times (p+1)$, se denominan **matriz de diseño** y contienen toda la información de los predictores. La primera columna de unos corresponde al término del intercepto β_0.

3.1.4 Supuestos del modelo lineal múltiple

Para que nuestros estimadores tengan propiedades deseables (como ser insesgados y eficientes), el modelo debe cumplir una serie de supuestos sobre el comportamiento del término de error, conocidos como las **condiciones de Gauss-Markov** (Kutner et al. 2005; Weisberg 2005).

1. **Linealidad en los parámetros**: El valor esperado de la respuesta es una función lineal de los parámetros β. El modelo $E[\mathbf{Y}|\tilde{X}] = \tilde{X}\beta$ está bien especificado.

2. **Exogeneidad (media del error nula)**: Los errores tienen una media de cero para cualquier valor de los predictores, $E[\varepsilon|\tilde{X}] = \mathbf{0}$. Esto implica que los predictores no contienen información sobre el término de error.

3. **Homocedasticidad e independencia**: Los errores no están correlacionados entre sí y tienen una varianza constante σ^2 para cualquier valor de los predictores. En notación matricial: $\text{Var}(\varepsilon|\tilde{X}) = \sigma^2\mathbf{I}_n$.

4. **Ausencia de multicolinealidad perfecta**: Ningún predictor es una combinación lineal exacta de los otros. Esto asegura que la matriz $\mathbf{X}$ tiene rango completo $(p+1)$, lo cual es necesario para poder estimar de forma única todos los coeficientes.

5. **Normalidad de los errores (para inferencia)**: Para poder realizar contrastes de hipótesis e intervalos de confianza, se añade el supuesto de que los errores siguen una distribución Normal: $\varepsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$.

3.2 Estimación de los parámetros

Una vez definido el modelo y sus supuestos, el siguiente paso es estimar los parámetros desconocidos del vector β. El método más extendido es el de Mínimos Cuadrados Ordinarios.

3.2.1 El principio de mínimos cuadrados y la función objetivo

La idea de "mejor ajuste" se traduce matemáticamente en minimizar la discrepancia entre los valores observados $\mathbf{y}$ (datos muestrales) y los valores predichos por el modelo, $\hat{\mathbf{y}} = \mathbf{X}\widehat{\beta}$. Esta discrepancia se captura a través de los residuos, $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$.

MCO no minimiza simplemente los residuos (ya que residuos positivos y negativos se cancelarían), sino la **Suma de los Cuadrados de los Residuos** (SCR o *SSR* en inglés). Al elevarlos al cuadrado, nos aseguramos de que todos los errores contribuyan positivamente y, además, penalizamos más fuertemente los errores grandes.

La función objetivo a minimizar, $S(\beta)$, usando datos observados es:

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

3.2.2 Derivación de las ecuaciones normales

Para encontrar el vector $\widehat{\beta}$ que minimiza esta función, utilizamos cálculo diferencial. Primero, expandimos la expresión cuadrática de $S(\beta)$:

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

Un punto clave aquí es notar que $\beta^T\mathbf{X}^T\mathbf{y}$ es un escalar (una matriz 1×1), por lo que es igual a su transpuesta: $\beta^T\mathbf{X}^T\mathbf{y} = \mathbf{y}^T\mathbf{X}\beta$. 1×1). Por lo tanto, es igual a su propia transpuesta: $(\beta^T\mathbf{X}^T\mathbf{y})^T = \mathbf{y}^T\mathbf{X}\beta$. Esto nos permite simplificar la expresión:

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

Ahora, derivamos esta función con respecto al vector β e igualamos el resultado a un vector de ceros para encontrar el mínimo. Usando las reglas de la derivación matricial:

- La derivada de $\mathbf{y}^T\mathbf{y}$ respecto a β es $\mathbf{0}$.
- La derivada de $2\beta^T\mathbf{X}^T\mathbf{y}$ respecto a β es $2\mathbf{X}^T\mathbf{y}$.
- La derivada de la forma cuadrática $\beta^T(\mathbf{X}^T\mathbf{X})\beta$ respecto a β es $2(\mathbf{X}^T\mathbf{X})\beta$.

Aplicando estas reglas, obtenemos el gradiente de la función de pérdida:

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

Igualando a cero y sustituyendo β por el estimador $\widehat{\beta}$ que cumple esta condición:

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

Simplificando, llegamos al célebre sistema de $p + 1$ ecuaciones conocido como las **Ecuaciones Normales**:

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

3.2.3 La solución MCO y la condición de invertibilidad

Para resolver este sistema y despejar $\widehat{\beta}$, necesitamos multiplicar por la inversa de la matriz $(\mathbf{X}^T\mathbf{X})$. Esta inversa existe si y solo si la matriz es invertible, lo que está directamente garantizado por el **supuesto de ausencia de multicolinealidad perfecta**.

Si el rango de la matriz de diseño $\mathbf{X}$ es $p+1$ (sus columnas son linealmente independientes), entonces la matriz $\mathbf{X}^T\mathbf{X}$ (de dimensión $(p+1)\times(p+1)$) será de rango completo, simétrica y definida positiva, y por tanto, invertible.

La solución única para el vector de estimadores MCO es:

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

Esta compacta y poderosa ecuación es la base de la estimación en regresión lineal y es implementada por todo el software estadístico.

Claro, aquí tienes el texto completo en formato Markdown y con las fórmulas clave sin los recuadros.

3.2.4 Propiedades de los estimadores de MCO

Una vez que hemos obtenido la fórmula para calcular nuestros coeficientes, $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, la pregunta fundamental es: ¿qué tan buenos son estos estimadores? La teoría estadística nos proporciona una respuesta contundente a través de sus propiedades en el muestreo, que son la base para toda la inferencia estadística posterior.

El **Teorema de Gauss-Markov** es el resultado central. Afirma que, si se cumplen los supuestos del modelo lineal clásico (1-4), los estimadores de Mínimos Cuadrados Ordinarios son los **Mejores Estimadores Lineales Insesgados (MELI o BLUE)**. Desglosemos esto:

- **Lineal**: $\widehat{\beta}$ es una combinación lineal de la variable respuesta $\mathbf{y}$.
- **Insesgado (Unbiased)**: En promedio, a lo largo de infinitas muestras, nuestro estimador acertará al verdadero valor poblacional β. No tiene un sesgo sistemático. La demostración formal es directa:

$$\begin{aligned} E[\widehat{\beta}] &= E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] \\ &= E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\beta+\varepsilon)] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E[\varepsilon] \\ &= I\beta + \mathbf{0} = \beta \end{aligned}$$

- **Mejor (Best)**: "Mejor" significa que tiene la mínima varianza posible dentro de la clase de todos los estimadores lineales e insesgados. No existe otro estimador de este tipo que sea más preciso. La precisión de nuestros estimadores se captura en su **matriz de varianzas-covarianzas**:

$$\begin{aligned}\text{Var}(\widehat{\beta}) &= \text{Var}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\text{Var}(\mathbf{y})\,[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]^T \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,(\sigma^2\mathbf{I}_n)\,\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\end{aligned}$$

  Por tanto, la matriz que define la incertidumbre de nuestro estimador es:

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

  Los elementos de la diagonal de esta matriz nos dan la varianza de cada coeficiente individual, $Var(\widehat{\beta}_j)$, mientras que los elementos fuera de la diagonal nos dan la covarianza entre pares de coeficientes, $Cov(\widehat{\beta}_j, \widehat{\beta}_k)$.

Finalmente, si añadimos el **supuesto de normalidad de los errores** ($\varepsilon_i \sim N(0, \sigma^2)$), las propiedades del estimador se completan. Dado que $\widehat{\beta}$ es una combinación lineal de $\mathbf{y}$ (que ahora es normal), el propio estimador seguirá una distribución normal:

$$\widehat{\beta} \sim N\big(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\big)$$

Esto implica que cada coeficiente individual también se distribuye normalmente:

$$\widehat{\beta}_j \sim N\big(\beta_j, \sigma^2[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}\big)$$

donde $[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}$ es el j-ésimo elemento de la diagonal de la matriz inversa. Este resultado es la puerta de entrada a la inferencia, permitiéndonos construir intervalos de confianza y realizar contrastes de hipótesis (como los test-t).

3.2.5 Estimación de la varianza del error

La matriz de varianzas-covarianzas de $\widehat{\beta}$ depende de σ^2, la varianza de los errores poblacionales, que es desconocida. Por lo tanto, el siguiente paso lógico es encontrar un buen estimador para ella a partir de nuestros datos.

El punto de partida natural son los residuos del modelo, $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, que son la contraparte muestral de los errores teóricos ε. La suma de los cuadrados de los residuos (SSE) es la base de nuestro estimador:

$$SSE = \mathbf{e}^T\mathbf{e} = \mathbf{y}^T(\mathbf{I}_n - \mathbf{H})^T(\mathbf{I}_n - \mathbf{H})\mathbf{y} = \mathbf{y}^T(\mathbf{I}_n - \mathbf{H})\mathbf{y}$$

(usando las propiedades de simetría e idempotencia de la matriz de proyección $\mathbf{H}$).

Para encontrar un estimador insesgado, calculamos el valor esperado de la SSE. Utilizando el lema $E[\mathbf{z}^T\mathbf{A}\mathbf{z}] = \text{traza}(\mathbf{A}\Sigma) + \mu^T\mathbf{A}\mu$ con $\mathbf{z} = \mathbf{y}$, $\mathbf{A} = \mathbf{I}_n - \mathbf{H}$, $\mu = \mathbf{X}\beta$ y $\Sigma = \sigma^2\mathbf{I}_n$:

$$E[SSE] = \text{traza}[(\mathbf{I}_n - \mathbf{H})\sigma^2\mathbf{I}_n] + \beta^T\mathbf{X}^T(\mathbf{I}_n - \mathbf{H})\mathbf{X}\beta$$

El segundo término se anula porque $(\mathbf{I}_n - \mathbf{H})\mathbf{X} = \mathbf{X} - \mathbf{H}\mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{0}$. Nos queda:

$$\begin{aligned} E[SSE] &= \sigma^2\text{traza}(\mathbf{I}_n - \mathbf{H}) \\ &= \sigma^2(\text{traza}(\mathbf{I}_n) - \text{traza}(\mathbf{H})) \\ &= \sigma^2(n - (p+1)) \end{aligned}$$

El valor esperado de la SSE no es σ^2, sino un múltiplo de ella. Esto nos lleva directamente a un estimador insesgado para σ^2 dividiendo la SSE por sus **grados de libertad**, $n - p - 1$:

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-p-1} = \frac{\mathbf{e}^T\mathbf{e}}{n-p-1}$$

Intuitivamente, perdemos un grado de libertad por cada parámetro que hemos estimado en el modelo (los p coeficientes de las pendientes y el intercepto). La raíz cuadrada de este valor, $\hat{\sigma}$, se conoce como el **Error Estándar de la Regresión** y representa la magnitud de un error de predicción típico.

Con este estimador, podemos calcular el **error estándar de cada coeficiente**, que mide la incertidumbre de nuestra estimación para β_j:

$$\text{se}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}$$

Bajo normalidad, se puede demostrar además que la cantidad $\frac{SSE}{\sigma^2}$ sigue una distribución **Chi-cuadrado** con $n-p-1$ grados de libertad, un resultado clave para la inferencia formal.

Ejemplo: Estimación de un modelo múltiple

Para ilustrar estos conceptos, usemos un ejemplo con datos de precios de viviendas. Supongamos que queremos predecir el precio de una vivienda basándonos en su superficie, número de habitaciones, antigüedad, distancia al centro y si tiene garaje.

```
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
```

Este output nos muestra:

- **Coeficientes estimados** ($\widehat{\beta}$) y sus errores estándar
- **Estadísticos t** y p-valores para cada coeficiente
- **Error estándar residual** ($\widehat{\sigma} = 1.5952 \times 10^4$ euros)
- **R² múltiple** (0.9094) - proporción de varianza explicada
- **Estadístico F global** para contrastar la significancia del modelo

3.3 La interpretación de los coeficientes

Estimar los coeficientes y sus errores estándar es solo la mitad del trabajo. La otra mitad, y a menudo la más importante, es interpretarlos correctamente.

El concepto fundamental en regresión múltiple es el de **ceteris paribus** ("lo demás constante"). Cada coeficiente β_j representa el cambio esperado en Y por un cambio de una unidad en X_j, **manteniendo todas las demás variables predictoras del modelo fijas**. Es el efecto "puro" o "aislado" de X_j sobre Y, después de haber controlado por la influencia de las otras variables incluidas en el modelo. Matemáticamente, es la derivada parcial del valor esperado de Y con respecto a X_j:

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

Esta interpretación es crucialmente diferente de la que se obtiene en una regresión simple. El coeficiente de una regresión simple de Y sobre X_j captura no solo el efecto directo de X_j, sino

también los efectos indirectos de cualquier otra variable omitida que esté correlacionada tanto con Y como con X_j. Por ello, el valor de $\hat{\beta}_j$ en una regresión múltiple casi nunca es igual al de una regresión simple.

La forma más precisa de entender $\hat{\beta}_j$ es a través del concepto de **regresión parcial**. El coeficiente $\hat{\beta}_j$ de la regresión múltiple es idéntico al coeficiente de una regresión simple entre dos conjuntos de residuos:

1. Los residuos de una regresión de **y** sobre todas las demás variables predictoras (excepto X_j).
2. Los residuos de una regresión de $\mathbf{x_j}$ sobre todas las demás variables predictoras.

En otras palabras, $\hat{\beta}_j$ mide la relación entre la parte de Y que no puede ser explicada por las otras variables y la parte de X_j que tampoco puede ser explicada por las otras variables. Es la asociación entre Y y X_j después de haber "limpiado" o "netado" la influencia de todos los demás predictores de ambas. Este concepto se visualiza en los **gráficos de regresión parcial** (o *added-variable plots*), que son una herramienta de diagnóstico fundamental.

Interpretación práctica de los coeficientes

Volviendo a nuestro ejemplo de viviendas, interpretemos cada coeficiente aplicando el principio *ceteris paribus*:

- **Superficie** (1172 €/m²): Por cada metro cuadrado adicional, el precio aumenta en promedio 1172 euros, **manteniendo constantes** el número de habitaciones, antigüedad, distancia al centro y presencia de garaje.
- **Habitaciones** (1.5072×10^4 €): Cada habitación adicional incrementa el precio en 1.5072×10^4 euros en promedio, **controlando por** la superficie y demás variables.
- **Antigüedad** (-745 €/año): Por cada año de antigüedad, el precio disminuye en 745 euros en promedio, **ceteris paribus**.
- **Distancia al centro** (-2028 €/km): Cada kilómetro adicional de distancia reduce el precio en 2028 euros en promedio, **manteniendo todo lo demás constante**.
- **Garaje** (2.5829×10^4 €): Las viviendas con garaje cuestan 2.5829×10^4 euros más que las que no tienen, **en promedio y controlando por las demás variables**.

Punto clave: Estos efectos son diferentes de los que obtendríamos con regresiones simples, ya que aquí hemos "limpiado" la influencia de las otras variables.

La perspectiva geométrica de mínimos cuadrados

La estimación por mínimos cuadrados tiene una interpretación geométrica elegante y potente que nos ayuda a comprender qué está ocurriendo.
Podemos pensar en el vector de observaciones $\mathbf{y}$ como un punto en un espacio de n dimensiones. Las columnas de la matriz de diseño $\mathbf{X}$ generan un subespacio vectorial dentro de $\mathbb{R}^n$, conocido como el **espacio columna** de $\mathbf{X}$, denotado $C(\mathbf{X})$. Este subespacio contiene todas las posibles combinaciones lineales de nuestros predictores.
El método MCO encuentra el vector de valores ajustados $\hat{\mathbf{y}} = \mathbf{X}\widehat{\beta}$ que está "más cerca" de $\mathbf{y}$. Geométricamente, este punto no es otro que la **proyección ortogonal** del vector $\mathbf{y}$ sobre el subespacio $C(\mathbf{X})$.
Esta proyección se realiza a través de una matriz especial llamada **matriz de proyección** o **matriz sombrero** (*hat matrix*), denotada por $\mathbf{H}$:

$$\hat{\mathbf{y}} = \text{Proj}_{C(\mathbf{X})}\,\mathbf{y} = \mathbf{H}\,\mathbf{y}, \qquad \text{donde} \quad \mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

Esta operación induce la **descomposición ortogonal fundamental** del vector de respuesta:

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$$

El hecho de que la proyección sea ortogonal implica que el vector de residuos $\mathbf{e}$ es ortogonal (perpendicular) al vector de valores ajustados $\hat{\mathbf{y}}$ y, de hecho, a todo el subespacio $C(\mathbf{X})$. Esta ortogonalidad, $\hat{\mathbf{y}}^T\mathbf{e} = 0$, es la base del **Teorema de Pitágoras para la regresión**, que permite la descomposición de la variabilidad total en una parte explicada y una no explicada.

3.4 Evaluación del modelo y descomposición de la varianza

Una vez estimado el modelo, el siguiente paso es evaluar su desempeño. ¿Qué tan bien se ajustan nuestras predicciones a los datos reales? Aunque ya vimos la perspectiva geométrica y el Teorema de Pitágoras en regresión simple, es importante revisitar estos conceptos porque en regresión múltiple la interpretación y el cálculo de la descomposición de varianza presenta matices adicionales que debemos entender claramente.

En regresión múltiple, la **Descomposición de la Varianza** o **ANOVA** (*Analysis of Variance*) cobra especial relevancia porque ahora tenemos múltiples variables explicativas y necesitamos evaluar el aporte conjunto de todas ellas, así como su significancia global.

La idea fundamental es que la variabilidad total de la variable respuesta (Y) puede descomponerse en dos partes: una parte que es explicada por nuestro modelo de regresión (ahora con múltiples variables) y otra parte que queda sin explicar, atribuida al error aleatorio.

Partimos de la identidad: $(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) = (\hat{y}_i - \bar{y}) + e_i$.

Elevando al cuadrado y sumando para todas las observaciones (y gracias a la propiedad de ortogonalidad $\hat{\mathbf{y}}^T\mathbf{e} = 0$, que hace que los productos cruzados se anulen), llegamos a la descomposición fundamental de las sumas de cuadrados:

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

Esto se conoce como la **ecuación de ANOVA**:

$$SST = SSR + SSE$$

Donde:

- **SST (Suma de Cuadrados Total)**: Es la variabilidad total de Y. Mide la dispersión de los datos observados alrededor de su media.
- **SSR (Suma de Cuadrados de la Regresión)**: Es la variabilidad **explicada** por el modelo. Mide la dispersión de los valores predichos alrededor de la media.
- **SSE (Suma de Cuadrados del Error)**: Es la variabilidad **no explicada** o residual. Mide la dispersión de los datos observados alrededor de la línea de regresión.

Esta tabla no es solo un resumen; es el motor de las principales herramientas de evaluación e inferencia del modelo.

3.4.1 Coeficiente de determinación múltiple

El **coeficiente de determinación**, R^2, es la medida de ajuste más popular. Responde a la pregunta: *¿Qué proporción de la variabilidad total de Y es explicada por las variables predictoras del modelo?*

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

Propiedades clave:

- Su valor siempre está entre 0 (el modelo no explica nada) y 1 (el modelo explica toda la variabilidad).
- Puede interpretarse como el cuadrado de la correlación entre los valores observados y los valores predichos, $R^2 = \text{corr}^2(\mathbf{y}, \hat{\mathbf{y}})$.
- **Problema**: R^2 **nunca decrece** al añadir una nueva variable predictora al modelo, incluso si esta es completamente irrelevante. Esto lo convierte en una métrica engañosa para comparar modelos con distinto número de predictores.

3.4.2 El coeficiente de determinación ajustado

Para solucionar el problema de R^2, utilizamos el R^2 ajustado, que introduce una penalización por cada variable añadida. Lo hace comparando las varianzas (sumas de cuadrados divididas por sus grados de libertad) en lugar de solo las sumas de cuadrados:

$$R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)} = 1 - \frac{\hat{\sigma}^2}{s^2_Y}$$

Donde s^2_Y es la varianza muestral de Y. El $R^2_{ajustado}$ solo aumentará si la nueva variable mejora el modelo más de lo que se esperaría por puro azar. Es, por tanto, la métrica preferida para comparar la calidad de ajuste de modelos anidados.

¡Excelente observación! Tienes toda la razón. En la sección de "Inferencia" me centré en los contrastes de hipótesis (el test t y el test F) pero omití una parte igualmente importante que sí estaba en tu Tema 1: la **construcción de intervalos de confianza para los parámetros**.

Un contraste de hipótesis te da una respuesta de "sí/no" sobre la significancia, pero un intervalo de confianza te ofrece un rango de valores plausibles para el efecto, lo cual es mucho más informativo.

Aquí tienes una versión revisada y ampliada de la sección **"Inferencia estadística en el modelo múltiple"** que incluye explícitamente los intervalos de confianza, manteniendo el flujo que hemos construido. Puedes reemplazar la sección anterior por esta.

3.5 Inferencia Estadística en el Modelo Múltiple

La estimación nos da los valores de los coeficientes para nuestra muestra, pero la inferencia nos permite usar esos valores para sacar conclusiones sobre los parámetros de la población. ¿Son estos coeficientes "reales" o podrían ser fruto del azar muestral? Para responder, nos basamos en las propiedades distributivas de nuestros estimadores.

3.5.1 Contraste de hipótesis sobre los coeficientes

El **test t** nos permite decidir si una variable predictora X_j tiene una relación estadísticamente significativa con Y, después de controlar por el efecto de todas las demás variables en el modelo.

- **Hipótesis**: La hipótesis nula es que el coeficiente es cero en la población ($H_0 : \beta_j = 0$), lo que implicaría que X_j no tiene un efecto lineal sobre Y una vez que se consideran los otros predictores. La alternativa es que el coeficiente es distinto de cero ($H_1 : \beta_j \neq 0$).

- **Estadístico de contraste**: Construimos el estadístico t, que mide cuántos errores estándar separan nuestro coeficiente estimado del valor nulo (cero).

$$\frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} \sim t_{n-p-1}$$

  Bajo la hipótesis nula, el estadístico que calculamos con nuestra muestra es $t_{obs} = \hat{\beta}_j/\text{se}(\hat{\beta}_j)$.
- **Decisión**: Comparamos el valor observado t_{obs} con la distribución t de Student con $n - p - 1$ grados de libertad. Si el **p-valor** asociado es suficientemente pequeño (normalmente < 0.05), rechazamos la hipótesis nula y concluimos que la variable es un predictor estadísticamente significativo.

3.5.2 Intervalo de confianza para los coeficientes

Mientras que el test t nos da una decisión binaria, el **intervalo de confianza** nos proporciona un **rango de valores plausibles** para el verdadero parámetro poblacional β_j. Es una herramienta de estimación más informativa.

La estructura del intervalo se basa en la distribución t que acabamos de ver:

$$\text{Estimación puntual} \pm (\text{Valor crítico}) \times (\text{Error estándar})$$

Para un nivel de confianza del $100(1-\alpha)\%$, el intervalo para β_j es:

$$\hat{\beta}_j \pm t_{\alpha/2, n-p-1} \cdot \text{se}(\hat{\beta}_j)$$

- **Interpretación**: Tenemos una confianza del $100(1-\alpha)\%$ de que el verdadero valor del parámetro poblacional β_j se encuentra dentro de este rango.
- **Dualidad con el Contraste de Hipótesis**: Existe una relación directa entre el intervalo de confianza y el test t. Si el valor 0 **no está incluido** en el intervalo de confianza del 95% para $\hat{\beta}_j$, es matemáticamente equivalente a rechazar la hipótesis nula $H_0 : \beta_j = 0$ con un nivel de significancia $\alpha = 0.05$. Esto nos da dos formas de llegar a la misma conclusión sobre la significancia de un predictor.

3.5.3 Inferencia sobre la significancia global del modelo

El **test F** evalúa si el modelo en su conjunto tiene poder predictivo. Es decir, contrasta si **al menos uno** de los predictores tiene una relación significativa con Y.

- **Hipótesis**: La hipótesis nula es que todos los coeficientes de las pendientes son simultáneamente cero ($H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$), frente a la alternativa de que al menos uno es distinto de cero ($H_1 :$ Algún $\beta_j \neq 0$).
- **Estadístico de contraste**: El estadístico F se construye a partir de la tabla ANOVA, comparando la varianza explicada por el modelo con la varianza residual, ajustando por sus respectivos grados de libertad.

$$F = \frac{\text{Varianza Explicada}}{\text{Varianza No Explicada}} = \frac{SSR/p}{SSE/(n-p-1)}$$

- **Decisión**: Comparamos el valor del estadístico F con una distribución F de Snedecor con p y $n-p-1$ grados de libertad. Un p-valor pequeño indica que el modelo es globalmente significativo y que, como conjunto, nuestros predictores explican una parte de la variabilidad de Y que no es atribuible al azar.

El test F es una herramienta fundamental, ya que representa el primer paso en la validación de cualquier modelo de regresión múltiple.

Interpretación de las pruebas estadísticas

En nuestro ejemplo de viviendas:
Prueba F global:

- F(5, 194) = 389.45, p < 0.001
- **Conclusión**: El modelo es globalmente significativo. Al menos una variable predictora tiene una relación real con el precio.

Pruebas t individuales (ejemplos):

- **Superficie**: t = 24.78, p < 0.001 → **Significativa**
- **Garaje**: t = 10.99, p < 0.001 → **Significativa**

Intervalos de confianza (95%):

- **Superficie**: [1079, 1265] euros/m²
- No incluye el 0, confirma la significancia estadística

Interpretación práctica: Estamos 95% confiados de que el verdadero efecto de la superficie está entre 1079 y 1265 euros por m², controlando por las demás variables.

3.6 Predicción con el modelo múltiple

Una vez que hemos ajustado y validado nuestro modelo, podemos utilizarlo para uno de sus propósitos más poderosos: hacer predicciones para nuevas observaciones. Es fundamental

distinguir entre dos objetivos de predicción diferentes, ya que cada uno conlleva un nivel de incertidumbre distinto.

Supongamos que tenemos un nuevo conjunto de valores para las variables predictoras, representado por el vector $\mathbf{x}_0^T = [1, x_{01}, x_{02}, \dots, x_{0p}]$. La **predicción puntual** en ambos casos es la misma:

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

Sin embargo, esta estimación puntual está sujeta a error. Para cuantificar esta incertidumbre, construimos dos tipos de intervalos.

3.6.1 Intervalo de confianza para la respuesta media

Este intervalo responde a la pregunta: *¿cuál es el valor **promedio** de Y para todas las observaciones con las características* $\mathbf{x}_0$? Su objetivo es acotar la posición de la verdadera (pero desconocida) superficie de regresión poblacional en el punto $\mathbf{x}_0$.

La incertidumbre aquí proviene únicamente de la estimación de los coeficientes $\hat{\beta}$. La varianza de esta predicción media es:

$$\text{Var}(\hat{y}_0) = \text{Var}(\mathbf{x}_0^T \hat{\beta}) = \mathbf{x}_0^T \text{Var}(\hat{\beta}) \mathbf{x}_0 = \sigma^2 \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0$$

Reemplazando σ^2 por su estimador insesgado $\hat{\sigma}^2$, el **intervalo de confianza al** $100(1-\alpha)\%$ para la respuesta media $E[Y|\mathbf{X} = \mathbf{x}_0]$ es:

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

3.6.2 Intervalo de predicción para una observación individual

Este intervalo responde a una pregunta más ambiciosa: *¿entre qué valores se encontrará la respuesta de una **única y nueva** observación con las características* $\mathbf{x}_0$?

Este intervalo debe considerar **dos fuentes de incertidumbre**:

1. La incertidumbre sobre la localización de la verdadera superficie de regresión (la misma que en el intervalo de confianza).
2. La variabilidad aleatoria inherente a una sola observación, que se desvía de la media poblacional (el error ε_0, cuya varianza es σ^2).

Por esta razón, un intervalo de predicción **siempre será más ancho** que un intervalo de confianza para el mismo nivel de significancia. La varianza del error de predicción es la suma de las dos fuentes de varianza:

$$\text{Var}(y_0 - \hat{y}_0) = \text{Var}(\varepsilon_0) + \text{Var}(\hat{y}_0) = \sigma^2 + \sigma^2 \mathbf{x}_0^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0$$

El **intervalo de predicción al** $100(1-\alpha)\%$ para una observación individual y_0 es:

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

La diferencia clave es el **"+1"** dentro de la raíz cuadrada, que representa la varianza de la nueva observación. Visualmente, tanto la banda de confianza como la de predicción son más estrechas cerca del "centroide" de los datos (la media multivariante de los predictores) y se ensanchan a medida que nos alejamos hacia valores más extremos de los predictores.

3.7 Diagnóstico del modelo múltiple

Al igual que en la regresión simple, una vez ajustado el modelo es fundamental realizar un diagnóstico exhaustivo para verificar que los supuestos del modelo lineal se cumplen. La fiabilidad de todas nuestras inferencias (p-valores e intervalos de confianza) descansa sobre la validez de estos supuestos. El proceso sigue basándose en el **análisis de los residuos**, nuestra ventana a los errores teóricos no observables (Fox and Weisberg 2018; Harrell 2015).

3.7.1 Verificación de los supuestos clásicos

Los supuestos de linealidad, homocedasticidad, normalidad e independencia se verifican con herramientas muy similares a las vistas en el capítulo anterior, por lo que aquí las resumiremos y presentaremos una herramienta adicional para la linealidad.

- **Normalidad**: El **gráfico Q-Q** de los residuos estudentizados y el **test de Shapiro-Wilk** siguen siendo las herramientas principales para comprobar que los errores se distribuyen de forma Normal.
- **Independencia**: Para datos de series temporales, el **gráfico de residuos contra el tiempo** y el **test de Durbin-Watson** se utilizan de la misma manera para detectar autocorrelación.
- **Homocedasticidad (Varianza Constante)**: El **gráfico Scale-Location** y el **test de Breusch-Pagan** siguen siendo los métodos de referencia para detectar heterocedasticidad (patrones de embudo en la dispersión de los residuos).

Para la **linealidad**, el gráfico de **Residuos vs. Valores Ajustados** sigue siendo la primera herramienta a inspeccionar. Una nube de puntos sin patrones alrededor del cero sugiere que la forma funcional del modelo es globalmente correcta. Sin embargo, en el contexto múltiple, este gráfico podría ocultar una relación no lineal con *una variable específica*. Para ello, disponemos de una herramienta más precisa.

3.7.1.1 Gráficos de componente más residuo

El gráfico de Componente más Residuo (CPR o Partial Residual Plots) nos permite visualizar la relación entre la variable respuesta y **un único predictor** X_j, ajustando por el efecto de todos los demás predictores. Para cada predictor X_j, el gráfico muestra:

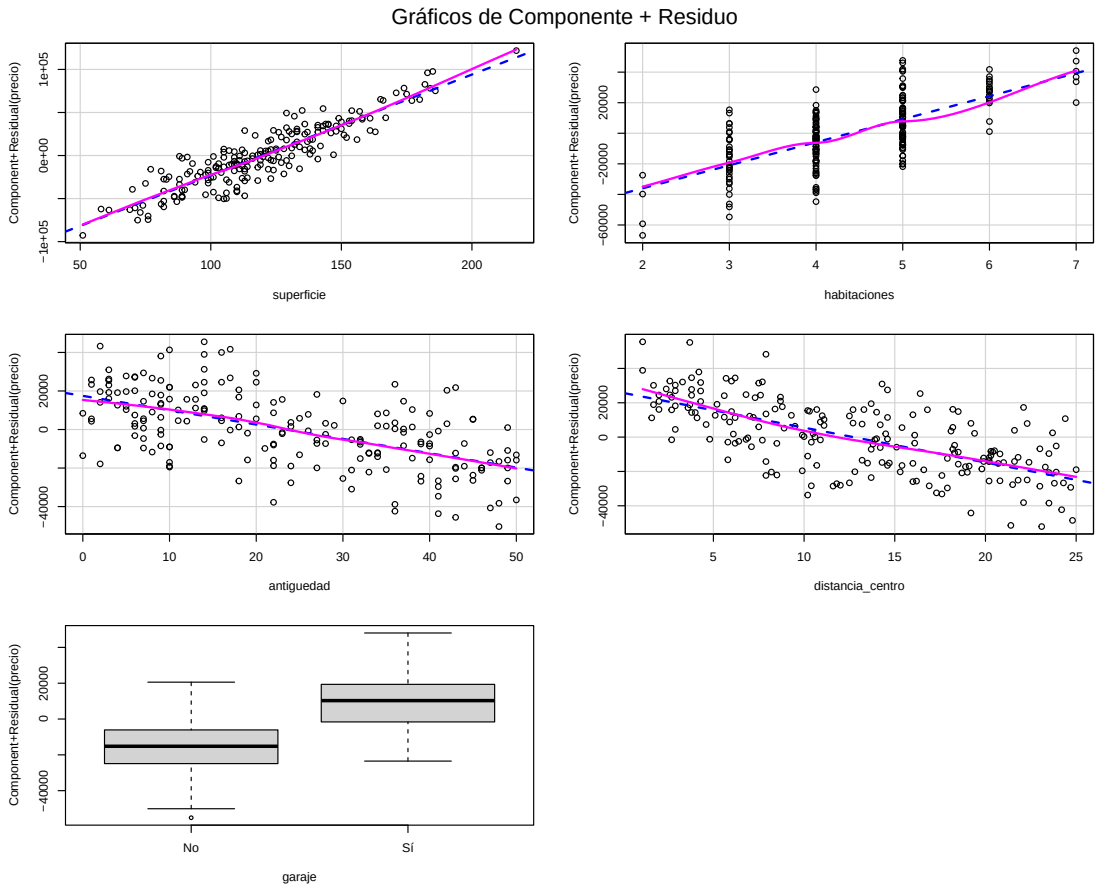
$$\text{Residuo Parcial} = e_i + \hat{\beta}_j x_{ij} \quad \text{vs.} \quad x_{ij}$$

La pendiente de una línea ajustada a estos puntos es exactamente $\hat{\beta}_j$. Si la relación es lineal, los puntos deben seguir la línea de regresión. Una desviación sistemática (una curva) sugiere que la relación con esa variable específica no es lineal y que podría necesitar una transformación (p.ej., logarítmica o cuadrática).

Ejemplo práctico: Gráficos de componente + residuo

Los gráficos CPR nos permiten visualizar si la relación de cada predictor con la respuesta es verdaderamente lineal:

```
# Cargar librerías necesarias
suppressPackageStartupMessages(library(car))

# Gráficos de Componente más Residuo (CPR)
crPlots(modelo, main = "Gráficos de Componente + Residuo")
```



Interpretación:

- **Línea sólida**: Muestra la relación lineal esperada (pendiente = coeficiente de regresión)
- **Línea punteada**: Suavizado no paramétrico de los puntos
- **Coincidencia de líneas**: Sugiere linealidad adecuada
- **Divergencia significativa**: Indica posible no-linealidad que requiere transformación

Si vemos curvas sistemáticas en algún gráfico CPR, es señal de que esa variable podría necesitar una transformación (log, cuadrática, etc.).

3.7.2 Diagnóstico de multicolinealidad

La **multicolinealidad** es un problema que solo existe en la regresión múltiple. Ocurre cuando dos o más variables predictoras están fuertemente correlacionadas entre sí.

3.7.2.1 Consecuencias de la multicolinealidad

La multicolinealidad no viola los supuestos de Gauss-Markov (los estimadores siguen siendo insesgados y eficientes), pero puede arruinar la interpretación práctica de un modelo:

1. **Varianza de los estimadores inflada**: Los errores estándar de los coeficientes de las variables colineales se vuelven muy grandes. Esto dificulta o imposibilita declarar un predictor como estadísticamente significativo, incluso si lo es.
2. **Inestabilidad de los coeficientes**: Pequeños cambios en los datos (o añadir/quitar una variable) pueden provocar cambios drásticos en las estimaciones de los coeficientes, incluso cambiando su signo, lo que hace que la interpretación sea poco fiable.
3. **Contradicciones en los contrastes**: Se puede dar la paradoja de tener un modelo globalmente significativo (test F con p-valor bajo y R^2 alto) pero con ningún predictor individual significativo (tests t con p-valores altos).

3.7.2.2 Detección de la multicolinealidad

1. **Matriz de correlaciones**: Un primer paso es examinar la matriz de correlaciones entre los predictores. Coeficientes de correlación altos (p. ej., > 0.8) son una señal de alerta. Sin embargo, este método no detecta la colinealidad que involucra a tres o más variables.
2. **Factor de Inflación de la Varianza (VIF)**: Es la herramienta de diagnóstico definitiva. Para cada predictor X_j, se calcula su VIF de la siguiente manera:
    1. Se ajusta una regresión lineal de X_j en función de **todas las demás variables predictoras**: $X_j \sim X_1 + \cdots + X_{j-1} + X_{j+1} + \cdots + X_p$.
    2. Se obtiene el R^2 de este modelo auxiliar, que llamamos R_j^2. Este valor nos dice qué proporción de la varianza de X_j es explicada por los otros predictores.
    3. El VIF se calcula como:
    $$VIF_j = \frac{1}{1 - R_j^2}$$
    - **Interpretación**: El VIF nos dice por qué factor se infla la varianza del estimador $\hat{\beta}_j$ debido a su relación con los otros predictores.

> **!** Reglas prácticas para interpretar VIF
>
> - **VIF = 1**: Ausencia de colinealidad (ideal)
> - **VIF > 5**: Valores preocupantes que requieren atención
> - **VIF > 10**: Multicolinealidad seria que debe ser tratada
>
> **Recordar**: El VIF indica por qué factor se multiplica la varianza del coeficiente debido a la multicolinealidad.

Ejemplo: Diagnóstico de multicolinealidad

Caso 1: Sin problemas de multicolinealidad (nuestro modelo actual)

```
# Calculamos VIF para nuestro modelo de viviendas
vif_values <- vif(modelo)
cat("VIF en nuestro modelo:\n")
```

```
VIF en nuestro modelo:
```

```
round(vif_values, 2)
```

```
      superficie     habitaciones       antiguedad distancia_centro
            1.40             1.40             1.01             1.01
          garaje
            1.01
```

Como todos los VIF < 5, no hay problemas de multicolinealidad.

Caso 2: Ejemplo con multicolinealidad problemática

```
# Simulamos un caso con multicolinealidad
set.seed(456)
n <- 100

# Creamos variables altamente correlacionadas
superficie <- runif(n, 50, 200)
habitaciones_sim <- round(superficie/25 + rnorm(n, 0, 0.5))  # Muy correlacionada con super
metros_cuadrados <- superficie + rnorm(n, 0, 5)  # Esencialmente la misma que superficie
precio_sim <- 1000*superficie + 5000*habitaciones_sim + rnorm(n, 0, 10000)

# Modelo con multicolinealidad
modelo_colineal <- lm(precio_sim ~ superficie + habitaciones_sim + metros_cuadrados)

# VIF del modelo problemático
vif_problematico <- vif(modelo_colineal)
cat("\nVIF en modelo con multicolinealidad:\n")
```

```
VIF en modelo con multicolinealidad:
```

```
round(vif_problematico, 2)
```

```
      superficie habitaciones_sim metros_cuadrados
           90.96            10.55            76.57
```

```r
# Matriz de correlaciones para explicar el problema
datos_problema <- data.frame(superficie, habitaciones_sim, metros_cuadrados)
cor_problema <- cor(datos_problema)
cat("\nMatriz de correlaciones:\n")
```

```
Matriz de correlaciones:
```

```r
round(cor_problema, 3)
```

```
                 superficie habitaciones_sim metros_cuadrados
superficie            1.000            0.951            0.993
habitaciones_sim      0.951            1.000            0.942
metros_cuadrados      0.993            0.942            1.000
```

Interpretación: - VIF > 10: Problema serio de multicolinealidad
- La correlación superficie-metros_cuadrados 1 explica el problema

i Soluciones a los problemas de multicolinealidad

Enfrentar la multicolinealidad no siempre significa que debamos alterar el modelo. La solución adecuada depende de la severidad del problema (medido con el VIF) y, sobre todo, del **objetivo de nuestro análisis**.

1. **No hacer nada**: Si el objetivo principal del modelo es la **predicción** y no la interpretación de los coeficientes individuales, la multicolinealidad no es un problema grave. El modelo en su conjunto puede tener un buen poder predictivo, aunque los efectos individuales de las variables colineales sean inestables. Además, si las variables colineales no son las variables de interés de nuestra investigación, podemos ignorar su multicolinealidad.

2. **Eliminar una de las variables correlacionadas**: Esta es la solución más simple y común. Si dos o más variables miden esencialmente el mismo concepto (p. ej., "educación en años" y "nivel educativo alcanzado"), una de ellas es redundante. Se puede eliminar la que tenga menor relevancia teórica o la que, individualmente, tenga una menor correlación con la variable respuesta.

3. **Combinar las variables colineales**: En lugar de eliminar información, podemos combinar las variables colineales en un único predictor compuesto. Por ejemplo,

si tenemos `gasto_en_publicidad_tv` y `gasto_en_publicidad_radio`, podríamos crear una nueva variable `gasto_total_en_medios`. Para casos más complejos, se pueden utilizar técnicas de reducción de dimensionalidad como el **Análisis de Componentes Principales (PCA)** para crear un índice que capture la información conjunta de las variables correlacionadas.

4. **Utilizar métodos de estimación alternativos**: Si no es posible o deseable modificar los predictores, se pueden usar técnicas de regresión penalizada que están diseñadas para manejar la multicolinealidad.
    - **Regresión Ridge**: Es el método por excelencia para este problema. Añade un pequeño sesgo a las estimaciones de los coeficientes para reducir drásticamente su varianza, produciendo un modelo mucho más estable y fiable.
    - **Lasso y Elastic Net**: Son otras técnicas de regresión penalizada que también manejan bien la colinealidad y, además, pueden realizar selección de variables al hacer que algunos coeficientes sean exactamente cero.
5. **Aumentar el tamaño de la muestra**: En algunos casos, la multicolinealidad puede ser un artefacto de una muestra pequeña. Recolectar más datos puede, en ocasiones, reducir la correlación entre los predictores y disminuir la varianza de los coeficientes.

La elección de la estrategia debe ser una decisión meditada, sopesando la simplicidad, la interpretabilidad y la robustez del modelo final.

3.7.3 Identificación de observaciones influyentes

Los conceptos de *outlier* (residuo grande), *leverage* (valor atípico en los predictores) e *influencia* (impacto en el modelo) son los mismos que en regresión simple. Sin embargo, el caso múltiple nos ofrece herramientas de diagnóstico más específicas.

3.7.3.1 DFBETAS: Influencia sobre coeficientes individuales

Mientras que la Distancia de Cook mide la influencia global de una observación sobre *todos* los coeficientes a la vez, los **DFBETAS** miden el impacto que tiene eliminar la observación i sobre **cada coeficiente** β_j individualmente.

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

Un DFBETA grande para el coeficiente β_j indica que la observación i tiene un fuerte poder de atracción sobre la estimación de ese coeficiente en particular. Una regla común es considerar problemáticos los puntos con $|\text{DFBETA}_{j,i}| > \frac{2}{\sqrt{n}}$.

3.7.3.2 Gráficos de regresión parcial

Estos gráficos son una de las herramientas visuales más potentes y elegantes de la regresión múltiple. Un gráfico de regresión parcial para un predictor X_j nos permite ver su relación con la respuesta Y **después de haber eliminado el efecto lineal de todos los demás predictores**.

Se construye de la siguiente forma: 1. Se calculan los residuos de la regresión de Y en función de todos los predictores excepto X_j. Llamemos a estos residuos $e_{Y|X_{-j}}$. 2. Se calculan los residuos de la regresión de X_j en función de todos los demás predictores. Llamemos a estos residuos $e_{X_j|X_{-j}}$. 3. Se grafica $e_{Y|X_{-j}}$ (eje Y) contra $e_{X_j|X_{-j}}$ (eje X).

⚠ Propiedad clave de los gráficos de regresión parcial

La magia de este gráfico: La pendiente de la línea de regresión ajustada a estos puntos es **exactamente** el coeficiente de regresión múltiple $\hat{\beta}_j$.
Esta equivalencia matemática es lo que hace que estos gráficos sean tan poderosos para entender la interpretación de los coeficientes en regresión múltiple.

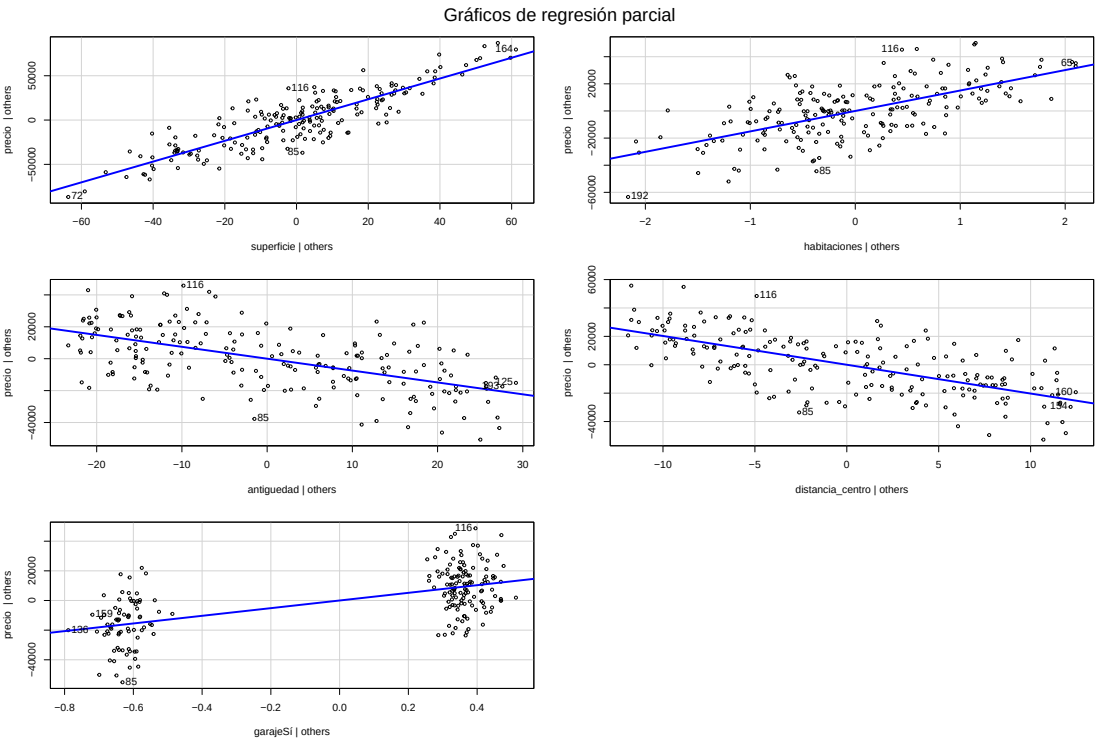
Estos gráficos son útiles para:

- Visualizar la magnitud y significancia del efecto de una variable "ajustada".
- Detectar no-linealidades en la relación parcial de una variable.
- Identificar observaciones que son influyentes para la estimación de un coeficiente específico (puntos con alto leverage o residuos grandes en este gráfico parcial).

💡 Ejemplo: Gráficos de Regresión Parcial

Los gráficos de regresión parcial nos permiten visualizar la relación "limpia" entre cada predictor y la variable respuesta:

```
# Gráficos de regresión parcial para todas las variables
avPlots(modelo, main = "Gráficos de regresión parcial")
```



¿Qué vemos en cada gráfico?

- **Eje X**: Residuos de la regresión de X_j vs. todos los demás predictores
- **Eje Y**: Residuos de la regresión de Y vs. todos los demás predictores (excepto X_j)
- **Pendiente**: Es exactamente el coeficiente $\hat{\beta}_j$ del modelo múltiple
- **Puntos alejados**: Observaciones influyentes para ese coeficiente específico

Interpretación práctica: - Una relación lineal clara confirma la validez del modelo - Puntos muy alejados de la línea pueden ser observaciones influyentes - Patrones curvos sugieren no-linealidad en esa variable específica

4 Ingeniería de características: transformaciones de variables e interacciones

En la práctica del análisis de datos, los datos en su estado bruto raramente están en la forma óptima para el modelado estadístico. La **ingeniería de características** es el proceso fundamental que transforma, combina y crea variables para maximizar la capacidad predictiva y la interpretabilidad de nuestros modelos (Kuhn and Johnson 2019; Zheng and Casari 2018).

Este proceso abarca tres áreas principales que exploraremos en profundidad:

Transformaciones de variables: Las transformaciones matemáticas nos permiten abordar múltiples problemas simultaneamente: linearizar relaciones no lineales, estabilizar la varianza (heterocedasticidad), aproximar la distribución de los errores a la normalidad, y reducir la influencia de valores atípicos. Dominar cuándo y cómo aplicar transformaciones logarítmicas, potenciales, Box-Cox o Yeo-Johnson es fundamental para optimizar nuestros modelos (Box and Cox 1964; Yeo and Johnson 2000).

Tratamiento de variables categóricas: Las variables categóricas requieren estrategias específicas de codificación que preserven la información relevante sin introducir supuestos erróneos. La elección entre codificación ordinal, one-hot encoding, o técnicas más avanzadas puede impactar significativamente el rendimiento del modelo (Potdar, Pardawala, and Pai 2017).

Interacciones entre variables: Las interacciones capturan cómo el efecto de una variable puede cambiar según el nivel de otra variable, revelando patrones que los efectos principales por sí solos no pueden detectar. Comprender los diferentes tipos de interacciones y sus aplicaciones es clave para modelar relaciones complejas en el mundo real (Jaccard and Turrisi 2003).

! Objetivos de aprendizaje

Al finalizar este capítulo, serás capaz de:

1. **Identificar cuándo aplicar transformaciones** específicas según el problema: linearización, heterocedasticidad, normalidad, o atípicos.
2. **Aplicar transformaciones clásicas** (logarítmica, potencial, inversa) y **avanzadas** (Box-Cox, Yeo-Johnson) de manera apropiada.
3. **Interpretar modelos transformados**, comprendiendo cómo cambia el significado de los coeficientes después de la transformación.

4. **Codificar variables categóricas** using ordinal encoding y one-hot encoding según la naturaleza de las categorías.
5. **Crear e interpretar términos de interacción** entre variables continuas, categóricas, y mixtas.
6. **Aplicar ingeniería de características** para crear nuevas variables predictivas mediante combinaciones, ratios y transformaciones.

4.1 Transformaciones de variables: propósitos y aplicaciones

En el análisis de datos y la construcción de modelos estadísticos, los datos en su forma original no siempre están preparados para obtener el máximo rendimiento de nuestros modelos. Las **transformaciones de variables** son herramientas fundamentales que nos permiten modificar la estructura matemática de nuestros datos para abordar problemas específicos y mejorar significativamente el ajuste del modelo (Box and Cox 1964; Carroll and Ruppert 1988).

La clave del éxito en las transformaciones está en diagnosticar correctamente el problema que enfrentamos y seleccionar la transformación apropiada. Cada transformación tiene propósitos específicos y consecuencias interpretativas que debemos comprender profundamente.

4.1.1 Diagnóstico: ¿Cuándo transformar?

El arte de las transformaciones no está en aplicarlas indiscriminadamente, sino en diagnosticar correctamente cuál es el problema que enfrentamos y seleccionar la transformación más apropiada para resolverlo. Un diagnóstico erróneo puede llevarnos a aplicar una transformación innecesaria o, peor aún, contraproducente que distorsione las relaciones reales en los datos.

La práctica común de "probar transformaciones hasta que mejore el ajuste" es metodológicamente peligrosa. Este enfoque de transformación por ensayo y error puede llevarnos a:

- **Sobreajuste**: Optimizar el modelo para los datos específicos que tenemos, perdiendo capacidad de generalización.
- **Pérdida de interpretabilidad**: Aplicar transformaciones complejas sin comprender su significado teórico.
- **Violación de supuestos**: Resolver un problema creando otros nuevos (ej. transformar para normalidad pero introducir heterocedasticidad).
- **Sesgo de selección**: Elegir la transformación que da los "mejores" resultados sin justificación teórica.

El proceso de diagnóstico debe ser **sistemático y basado en evidencia visual y estadística**. No basta con aplicar transformaciones porque "mejoran el R^2"; debemos entender qué problema

específico estamos resolviendo y cómo la transformación aborda ese problema desde una perspectiva teórica sólida.

Un enfoque metodológicamente sólido sigue estos principios:

1. **Diagnóstico previo**: Identificar problemas específicos mediante análisis visual y tests estadísticos antes de decidir transformar.
2. **Justificación teórica**: Cada transformación debe tener una base conceptual sólida. Por ejemplo, usar logaritmos para relaciones multiplicativas o raíz cuadrada para estabilizar varianza Poisson.
3. **Evaluación integral**: No solo considerar el ajuste estadístico, sino también la interpretabilidad, robustez y generalización del modelo transformado.
4. **Validación posterior**: Verificar que la transformación realmente resuelve el problema identificado sin crear nuevos problemas.
5. **Parsimonia**: Preferir la transformación más simple que resuelva efectivamente el problema (principio de Occam aplicado a transformaciones).

i Recordatorio: Diagnóstico de problemas en regresión lineal

Identificación de no linealidad: La no linealidad es uno de los problemas más comunes que enfrentamos en el modelado. *Diagnóstico visual:* gráficos de dispersión (Y vs. X), gráficos de componente + residuo (CPR plots), análisis de residuos vs. valores ajustados. *Diagnóstico estadístico:* Test de Ramsey RESET que evalúa si potencias de los valores ajustados mejoran significativamente el modelo.
Detección de heterocedasticidad: La heterocedasticidad (varianza no constante) viola supuestos fundamentales de la regresión lineal y sesga las inferencias estadísticas. *Diagnóstico visual:* gráfico de residuos vs. valores ajustados (patrón "embudo"), gráfico Scale-Location, residuos vs. variables predictoras individuales. *Diagnóstico estadístico:* Test de Breusch-Pagan (el más utilizado) y Test de White (más general).
Evaluación de normalidad de residuos: Aunque la normalidad no es crítica para estimación de coeficientes, sí es importante para inferencia estadística. *Diagnóstico visual:* histograma de residuos, QQ-plot de residuos. *Diagnóstico estadístico:* Test de Shapiro-Wilk (muestras pequeñas n<50) y Test de Jarque-Bera.
Detección de outliers y observaciones influyentes: Es fundamental distinguir entre outliers (valores extremos en Y) y observaciones influyentes (impacto en coeficientes). *Outliers:* boxplot de variable respuesta, residuos estudentizados, criterios $|t_i| > 2$. *Observaciones influyentes:* análisis de leverage, distancia de Cook, DFBETAS, DFFITS, con criterios específicos como $h_i > 2p/n$ y $D_i > 4/n$. Las observaciones se clasifican en: normales, outliers no influyentes, influyentes sin ser outliers, y outliers influyentes.

4.2 Escalado y normalización: preparando variables para el análisis

Antes de aplicar transformaciones complejas, es fundamental asegurar que nuestras variables estén en escalas comparables. Aunque la regresión lineal ordinaria no requiere estrictamente el escalado de variables para obtener estimadores insesgados, el escalado se vuelve crítico para la interpretación y esencial en métodos avanzados de modelado.

En regresión múltiple, los coeficientes representan el cambio en Y por unidad de cambio en cada variable predictora. Cuando las variables tienen escalas muy diferentes, los coeficientes pierden comparabilidad directa. Una variable medida en miles de euros tendrá coeficientes numéricamente pequeños, mientras que una variable medida en porcentajes tendrá coeficientes grandes, independientemente de su importancia real en el modelo.

Esta disparidad de escalas genera problemas interpretativos fundamentales: comparar la "importancia" relativa de las variables se vuelve imposible basándose únicamente en la magnitud de los coeficientes. El escalado resuelve este problema permitiendo que los coeficientes estandarizados (beta coefficients) representen cambios en desviaciones estándar, facilitando comparaciones directas entre predictores y proporcionando una base sólida para evaluar la importancia relativa de cada variable.

> **i** Escalado en métodos de regularización
>
> **En regresión con regularización (Ridge, Lasso)**, el problema se agrava dramáticamente. Las penalizaciones L1 y L2 afectan **desproporcionadamente** a variables con escalas grandes, llevando a regularización injusta donde variables con unidades grandes son penalizadas más severamente que variables con unidades pequeñas, independientemente de su relevancia predictiva. Esto puede resultar en selección de variables sesgada y estimadores subóptimos. Este tema se desarrollará en profundidad en el siguiente capítulo sobre métodos de regularización.

El escalado no es solo una cuestión técnica, sino una decisión metodológica que afecta directamente la interpretación y validez de nuestros resultados. La elección entre estandarización, normalización min-max, o escalado robusto debe basarse en las características de los datos y los objetivos del análisis, considerando siempre el impacto en la interpretabilidad de los resultados finales.

4.2.1 Estandarización (Z-Score)

La **estandarización** es la técnica de escalado más utilizada en estadística. Transforma cada variable para que tenga media cero y desviación estándar uno, preservando la forma de la distribución original.

$$X_{\text{estandarizado}} = \frac{X - \bar{X}}{\sigma_X}$$

Propiedades de la estandarización:

- **Preserva la forma de la distribución**: Si X era normal, X estandarizado también lo será.
- **Facilita la comparación**: Los coeficientes en regresión múltiple se vuelven comparables.
- **Robusta ante outliers moderados**: La media y desviación estándar son menos sensibles que min-max a valores extremos.

Cuándo usar estandarización:

- Variables con distribuciones aproximadamente normales.
- Cuando necesitamos preservar la información sobre la variabilidad relativa.
- En regresión múltiple para comparar la importancia relativa de las variables.
- Como paso previo a técnicas multivariantes (PCA, análisis discriminante).

4.2.2 Normalización Min-Max

La **normalización Min-Max** escala las variables a un rango específico, típicamente [0,1], preservando las relaciones relativas entre los valores.

$$X_{\text{normalizado}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Propiedades de la normalización Min-Max:

- **Rango acotado**: Todas las variables transformadas tienen el mismo rango [0,1].
- **Preserva relaciones**: Las distancias relativas entre observaciones se mantienen.
- **Interpretación intuitiva**: 0 representa el mínimo observado, 1 el máximo observado.

Cuándo usar normalización Min-Max:

- Cuando necesitamos un rango específico (ej. entradas de redes neuronales).
- Variables con distribuciones uniformes o sin outliers extremos.
- Cuando la interpretación en términos de mínimo-máximo es relevante.
- En algoritmos que requieren entradas en [0,1] (algunos métodos de ensemble).

Limitaciones:

- **Muy sensible a outliers**: Un solo valor extremo puede comprimir toda la distribución.
- **No preserva la normalidad**: Una distribución normal se vuelve uniforme tras Min-Max.

4.2.3 Escalado robusto

Para datos con outliers significativos, el escalado robusto utiliza la mediana y el rango intercuartílico (IQR) en lugar de la media y desviación estándar:

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

Este método es menos sensible a valores extremos y preserva mejor la estructura de los datos en presencia de outliers.

Ejemplo comparativo: Escalado de variables

```r
# Generar datos de ejemplo con diferentes escalas
set.seed(123)
n <- 100

# Variable de ejemplo: Ingresos en miles de euros
ingresos <- rnorm(n, mean = 50, sd = 15)

# Aplicar diferentes transformaciones
ingresos_std <- scale(ingresos)[,1]  # Estandarización
ingresos_norm <- (ingresos - min(ingresos)) / (max(ingresos) - min(ingresos))  # Min-Max
ingresos_robust <- (ingresos - median(ingresos)) / IQR(ingresos)  # Robusto

# Crear tabla comparativa
library(knitr)
tabla_escalado <- data.frame(
  Método = c("Original", "Estandarización", "Min-Max", "Escalado Robusto"),
  Media = round(c(mean(ingresos), mean(ingresos_std), mean(ingresos_norm), median(ingresos_
  Desviación = round(c(sd(ingresos), sd(ingresos_std), sd(ingresos_norm), mad(ingresos_robu
  Mínimo = round(c(min(ingresos), min(ingresos_std), min(ingresos_norm), min(ingresos_robus
  Máximo = round(c(max(ingresos), max(ingresos_std), max(ingresos_norm), max(ingresos_robus
  Rango = round(c(
    max(ingresos) - min(ingresos),
    max(ingresos_std) - min(ingresos_std),
    max(ingresos_norm) - min(ingresos_norm),
    max(ingresos_robust) - min(ingresos_robust)
  ), 3)
)

kable(tabla_escalado, caption = "Comparación de métodos de escalado en variable Ingresos")
```

Table 4.1: Comparación de métodos de escalado en variable Ingresos

Método	Media	Desviación	Mínimo	Máximo	Rango
Original	51.356	13.692	15.362	82.810	67.448
Estandarización	0.000	1.000	-2.629	2.297	4.926
Min-Max	0.534	0.203	0.000	1.000	1.000
Escalado Robusto	0.000	0.750	-2.000	1.793	3.792

Interpretación de los resultados:

- **Original**: Ingresos en miles de euros con gran variabilidad (SD 15)
- **Estandarización**: Media = 0, SD = 1, preservando la forma de la distribución
- **Min-Max**: Valores acotados entre [0,1], comprimiendo toda la variabilidad en este rango
- **Escalado Robusto**: Centrado en la mediana, menos sensible a valores extremos

Cada método transforma los datos de manera diferente según el objetivo: comparabilidad (estandarización), rango específico (min-max), o robustez ante outliers (robusto).

Ejemplo con outliers: Escalado robusto

```
# Datos con outliers
set.seed(456)
datos_normales <- rnorm(95, 10, 2)
outliers <- c(25, 30, 35, 40, 45)
datos_outliers <- c(datos_normales, outliers)

# Aplicar diferentes métodos de escalado
std_clasica <- scale(datos_outliers)[,1]
norm_minmax <- (datos_outliers - min(datos_outliers)) / (max(datos_outliers) - min(datos_ou
escalado_robusto <- (datos_outliers - median(datos_outliers)) / IQR(datos_outliers)

# Crear tabla comparativa
library(knitr)
tabla_outliers <- data.frame(
  Método = c("Original", "Estandarización", "Min-Max", "Escalado Robusto"),
  Media_Mediana = round(c(mean(datos_outliers), mean(std_clasica), mean(norm_minmax), media
  Desviación = round(c(sd(datos_outliers), sd(std_clasica), sd(norm_minmax), mad(escalado_r
  Mínimo = round(c(min(datos_outliers), min(std_clasica), min(norm_minmax), min(escalado_ro
  Máximo = round(c(max(datos_outliers), max(std_clasica), max(norm_minmax), max(escalado_ro
  Q25_Q75 = c(
    paste(round(quantile(datos_outliers, c(0.25, 0.75)), 2), collapse = " - "),
    paste(round(quantile(std_clasica, c(0.25, 0.75)), 2), collapse = " - "),
    paste(round(quantile(norm_minmax, c(0.25, 0.75)), 2), collapse = " - "),
    paste(round(quantile(escalado_robusto, c(0.25, 0.75)), 2), collapse = " - ")
  )
)

kable(tabla_outliers, caption = "Comparación de métodos de escalado con outliers presentes"
```

Table 4.2: Comparación de métodos de escalado con outliers presentes

Método	Media_Mediana	Desviación	Mínimo	Máximo	Q25_Q75
Original	11.447	5.996	5.491	45.000	8.93 - 11.81
Estandarización	0.000	1.000	-0.993	5.595	-0.42 - 0.06
Min-Max	0.151	0.152	0.000	1.000	0.09 - 0.16
Escalado Robusto	0.000	0.778	-1.654	12.108	-0.45 - 0.55

Análisis del impacto de outliers:

- **Datos originales**: Los outliers extienden el rango de ~6-14 a 6-45, distorsionando las medidas centrales
- **Estandarización**: Afectada por outliers en media y desviación estándar, resultando en distribución asimétrica
- **Min-Max**: Extremadamente sensible. Los datos normales quedan comprimidos en un rango muy pequeño (~0.0-0.2)
- **Escalado robusto**: Mantiene mejor las proporciones de la distribución central, minimizando la influencia de valores extremos

Conclusión: El escalado robusto es superior cuando hay outliers, preservando la estructura de la mayoría de observaciones.

4.3 Catálogo de transformaciones según el propósito

Una vez realizado el diagnóstico, debemos seleccionar la transformación más apropiada. Cada transformación tiene propósitos específicos y efectos secundarios que debemos considerar. La clave está en entender no solo qué transformación aplicar, sino por qué esa transformación específica resuelve nuestro problema.

4.3.1 Transformaciones para linearizar relaciones

Muchas relaciones en el mundo real no son lineales, pero pueden linearizarse mediante transformaciones apropiadas. La linearización no solo mejora el ajuste del modelo, sino que también facilita la interpretación y cumple con los supuestos de la regresión lineal.

4.3.1.1 Transformación logarítmica

La transformación logarítmica es probablemente la más utilizada en estadística aplicada debido a su versatilidad y propiedades interpretativas únicas.

Cuándo utilizarla:

- Relaciones exponenciales: Cuando Y crece exponencialmente respecto a X, $Y = ae^{bX}$ se linealiza como $\log(Y) = \log(a) + bX$
- Relaciones multiplicativas: En modelos donde los efectos se combinan multiplicativamente
- Procesos de crecimiento proporcional: Donde la tasa de cambio es proporcional al nivel actual
- Variables con crecimiento acelerado: Ingresos, precios, donde cada unidad adicional tiene impacto decreciente

Patrones de identificación:

- Curva cóncava que se aplana hacia la derecha (rendimientos decrecientes)
- Relación donde duplicar X no duplica Y, sino que el efecto se atenúa
- Heterocedasticidad donde la varianza aumenta con el nivel de Y

Aplicaciones matemáticas:

- $Y' = \log(Y)$: Linealiza relaciones exponenciales en Y
- $X' = \log(X)$: Linealiza relaciones de potencia en X

- $\log(Y) = a + b\log(X)$: Modelo log-log que produce elasticidades constantes

Interpretación especial: En modelos log-lineales y log-log, los coeficientes tienen interpretaciones económicas directas. En el modelo log-lineal $\log(Y) = a + bX$, el coeficiente b representa el cambio porcentual en Y por unidad de cambio en X. En el modelo log-log $\log(Y) = a + b\log(X)$, b es la elasticidad.

Casos típicos: Economía (relaciones ingreso-consumo, funciones de producción), biología (crecimiento poblacional, relaciones alométricas), finanzas (rendimientos de inversión).

4.3.1.2 Transformación de potencia

Las transformaciones de potencia son fundamentales cuando trabajamos con leyes físicas o relaciones alométricas donde esperamos relaciones del tipo $Y = aX^b$.

Identificación y aplicación:

- Relaciones curvilíneas que en escala log-log se vuelven lineales
- Método de linearización: tomar logaritmo de ambas variables $\log(Y) = \log(a) + b\log(X)$

- El exponente b representa la elasticidad o exponente de escalamiento

Ejemplos clásicos: Ley de Stevens en psicofísica, relaciones masa-metabolismo (Ley de Kleiber), economía urbana donde PIB de ciudades escala con población elevada a una potencia.

4.3.2 Transformaciones para estabilizar la varianza

La heterocedasticidad no solo viola supuestos del modelo, sino que también indica que diferentes observaciones tienen diferentes niveles de información. Las transformaciones de varianza estabilizan esta heterogeneidad.

4.3.2.1 Transformación de raíz cuadrada

La transformación $Y' = \sqrt{Y}$ es especialmente útil para datos de conteo donde la varianza es proporcional a la media, característica típica de la distribución de Poisson.

Fundamento teórico: En una distribución de Poisson con parámetro λ, tanto la media como la varianza son iguales a λ. La transformación de raíz cuadrada estabiliza la varianza porque $\text{Var}(\sqrt{Y}) \approx \frac{1}{4}$ (constante).

Cuándo aplicarla:

- Conteos de eventos: número de defectos, llamadas telefónicas, accidentes, ventas por período
- Datos de frecuencia: número de visitas, clicks, transacciones
- Variables discretas con varianza creciente proporcional al nivel

Patrón de diagnóstico: Gráfico de residuos con forma de embudo donde la dispersión aumenta linealmente con la media, y gráfico de varianza vs. media de grupos muestra relación lineal.

Limitaciones: Solo apropiada para valores no negativos, interpretación complicada (unidades en raíz cuadrada), y para conteos con muchos ceros puede requerir $\sqrt{Y + c}$.

4.3.2.2 Transformación logarítmica para heterocedasticidad multiplicativa

Cuando la varianza es proporcional al cuadrado de la media (heterocedasticidad multiplicativa), la transformación logarítmica es la solución natural.

Características típicas:

- Variables monetarias: ingresos, precios, costos donde el error relativo es constante
- Porcentajes de crecimiento: donde el error de medición es proporcional al nivel
- Procesos multiplicativos: donde los errores se acumulan multiplicativamente

Efectos múltiples: La transformación logarítmica frecuentemente resuelve múltiples problemas simultáneamente: lineariza relaciones exponenciales, estabiliza varianza multiplicativa, reduce el impacto de outliers extremos, y aproxima distribuciones asimétricas a la normalidad.

4.3.3 Transformaciones para normalizar residuos y controlar outliers

Algunas transformaciones son especialmente efectivas para aproximar distribuciones a la normalidad y reducir la influencia de valores extremos.

4.3.3.1 Transformación inversa

La transformación inversa $Y' = \frac{1}{Y}$ o $X' = \frac{1}{X}$ es útil para relaciones hiperbólicas y distribuciones con colas pesadas hacia la derecha.

Identificación matemática:

- Relación hiperbólica: $Y = \frac{a}{X} + b$ se lineariza como $Y = a \cdot \frac{1}{X} + b$
- Asíntota horizontal: la relación se aproxima a un valor límite cuando X aumenta

Aplicaciones específicas: Tiempo hasta el evento (con asíntota natural), tasas de decaimiento, relaciones dosis-respuesta en farmacología, curvas de demanda con elasticidad precio variable.

Efecto en outliers: La transformación inversa invierte la escala, comprimiendo fuertemente los valores grandes y expandiendo los pequeños. Útil para reducir influencia de outliers extremos, pero debe usarse con precaución ya que amplifica errores en valores pequeños.

Consideraciones prácticas: Solo aplicable a valores estrictamente positivos (o negativos), los coeficientes representan el impacto en la escala inversa, y requiere tratamiento especial para valores cercanos a cero.

4.4 Transformación de Box-Cox

La transformación de Box-Cox es un método que optimiza automáticamente el parámetro de transformación para maximizar la normalidad y homocedasticidad de los residuos (Box and Cox 1964). En lugar de elegir manualmente entre transformación logarítmica, raíz cuadrada o inversa, Box-Cox encuentra el valor λ (lambda) que mejor normaliza los datos.

4.4.1 Definición matemática

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

Los valores especiales de λ corresponden a transformaciones clásicas:

- $\lambda = \mathbf{1}$: Sin transformación (identidad)
- $\lambda = \mathbf{0.5}$: Transformación de raíz cuadrada

- $\lambda = \mathbf{0}$: Transformación logarítmica
- $\lambda = \mathbf{-1}$: Transformación inversa

4.4.2 Propósito y ventajas

Para qué sirve Box-Cox:

- Encuentra automáticamente la transformación óptima sin prueba y error
- Maximiza la verosimilitud del modelo, mejorando simultáneamente normalidad y homocedasticidad
- Proporciona un método objetivo para seleccionar la transformación apropiada
- Incluye intervalos de confianza para λ, permitiendo evaluar la incertidumbre de la estimación

Procedimiento de aplicación:

1. Se ajusta el modelo original y se calculan los residuos
2. Se evalúa la función de verosimilitud para diferentes valores de λ
3. Se selecciona el λ que maximiza la verosimilitud
4. Se aplica la transformación con el λ óptimo encontrado

4.4.3 Limitaciones importantes

Restricción de dominio: Box-Cox requiere que todos los valores de Y sean estrictamente positivos. Esta es su limitación más importante, ya que muchos conjuntos de datos reales incluyen ceros o valores negativos.

Aplicación tradicional: Se aplica principalmente a la variable dependiente Y, no a las variables predictoras. Aunque técnicamente es posible aplicarla a X, la interpretación se complica considerablemente.

Interpretación compleja: Cuando λ no corresponde a valores "simples" (como 0, 0.5, o 1), la interpretación de los coeficientes se vuelve difícil. Por ejemplo, si $\lambda = 0.37$, ¿cómo interpretar un coeficiente en la escala transformada?

Dependencia del modelo: El λ óptimo depende del modelo específico (predictores incluidos), por lo que cambiar el modelo puede requerir recalcular la transformación.

i Extensión: Transformación de Yeo-Johnson

La transformación de Yeo-Johnson (Yeo and Johnson 2000) fue desarrollada específicamente para superar la limitación principal de Box-Cox: la restricción a valores positivos. **Ventajas de Yeo-Johnson sobre Box-Cox:**

- **Sin restricción de dominio**: Acepta cualquier valor real, incluyendo negativos y cero
- **Preserva el signo**: Los valores negativos permanecen negativos tras la transformación
- **Continuidad**: La transformación es continua en Y = 0, evitando discontinuidades
- **Casos especiales familiares**: Incluye como casos especiales todas las transformaciones comunes

Cuándo usar cada una:

- **Box-Cox**: Para datos estrictamente positivos, especialmente cuando se busca comparabilidad con literatura existente
- **Yeo-Johnson**: Cuando los datos incluyen valores negativos o cero, o cuando se necesita mayor flexibilidad

La elección entre ambas depende fundamentalmente de las características de sus datos y los objetivos del análisis.

4.5 Tratamiento de variables categóricas

Las **variables categóricas** son fundamentales en el modelado estadístico, pero requieren una preparación especial antes de ser utilizadas en algoritmos que esperan entradas numéricas (Potdar, Pardawala, and Pai 2017). La elección del método de codificación puede impactar significativamente tanto la interpretabilidad como el rendimiento del modelo.

4.5.1 Principios de codificación categórica

¿Por qué codificar? La mayoría de algoritmos de machine learning y modelos estadísticos requieren entradas numéricas. Las variables categóricas deben transformarse preservando su información semántica sin introducir supuestos erróneos sobre relaciones entre categorías.

Criterios de selección del método:

- **Naturaleza de la variable**: ¿Existe orden inherente entre categorías?
- **Número de categorías**: Variables con muchas categorías requieren consideraciones especiales
- **Interpretabilidad**: ¿Qué método facilita la interpretación de resultados?
- **Eficiencia computacional**: Balance entre precisión y complejidad

4.5.2 Codificación One-Hot (variables nominales)

El **One-Hot Encoding** transforma variables categóricas nominales en un conjunto de variables binarias (0/1), donde cada nueva variable representa la presencia o ausencia de una categoría específica. Esta técnica es fundamental cuando trabajamos con variables categóricas que no tienen orden inherente, como color, género, región geográfica, o tipo de producto.

La transformación convierte una variable categórica con k categorías en k nuevas columnas binarias (o *k-1* para evitar colinealidad). Cada fila tendrá exactamente un "1" en la columna correspondiente a su categoría y "0" en todas las demás.

> ⚠ Dummy Variable Trap
>
> Cuando se crean todas las columnas (k para k categorías), una puede expresarse como combinación lineal de las demás, causando colinealidad perfecta en modelos lineales. **Solución:** Eliminar una categoría de referencia (usar k-1 columnas).

¿Por qué es necesario? La mayoría de algoritmos de machine learning requieren entradas numéricas y no pueden procesar directamente texto categórico. Más importante aún, el One-Hot Encoding no impone orden artificial entre categorías, tratando cada una como completamente independiente.

Ejemplo práctico: Consideremos una variable "Color" con valores [Rojo, Verde, Azul]. La codificación One-Hot creará tres columnas:

ID	Color	Color_Rojo	Color_Verde	Color_Azul
1	Rojo	1	0	0
2	Verde	0	1	0
3	Azul	0	0	1

ID	Color	Color_Rojo	Color_Verde	Color_Azul
4	Rojo	1	0	0

Cada observación queda representada por un vector binario que identifica unívocamente su categoría sin asumir relaciones ordinales entre colores.

Interpretación en regresión: En un modelo de regresión lineal, cada variable binaria creada tendrá su propio coeficiente que representa la diferencia en la variable respuesta entre esa categoría específica y la categoría de referencia (la omitida). Por ejemplo, si omitimos "Azul", el coeficiente de "Color_Rojo" indicará cuánto mayor (o menor) es el valor esperado de Y cuando el color es Rojo comparado con cuando es Azul.

Implementación práctica

```r
# Crear datos de ejemplo
suppressPackageStartupMessages(library(caret))
datos <- data.frame(
  ID = 1:5,
  Color = c("Rojo", "Verde", "Azul", "Rojo", "Verde")
)
```

Método 1: usando model.matrix (incluye todas las categorías)

```r
one_hot_completo <- model.matrix(~ Color - 1, data = datos)
one_hot_completo
```

```
  ColorAzul ColorRojo ColorVerde
1         0         1          0
2         0         0          1
3         1         0          0
4         0         1          0
5         0         0          1
attr(,"assign")
[1] 1 1 1
attr(,"contrasts")
attr(,"contrasts")$Color
[1] "contr.treatment"
```

Método 2: eliminando categoría de referencia (evita colinealidad)

```r
one_hot_referencia <- model.matrix(~ Color, data = datos)[, -1]
one_hot_referencia
```

```
  ColorRojo ColorVerde
1         1          0
2         0          1
3         0          0
4         1          0
5         0          1
```

Método 3: usando `caret::dummyVars` con `fullRank` para evitar colinealidad

```r
dummy_vars <- dummyVars(~ Color, data = datos, fullRank = TRUE)
one_hot_caret <- predict(dummy_vars, newdata = datos)
one_hot_caret
```

```
  ColorRojo ColorVerde
1         1          0
2         0          1
3         0          0
4         1          0
5         0          1
```

Ventajas y desventajas

Ventajas del One-Hot Encoding:

- **No asume orden**: Trata cada categoría como independiente
- **Interpretabilidad**: Cada coeficiente representa el efecto específico de esa categoría
- **Compatibilidad**: Funciona con todos los algoritmos numéricos

Desventajas:

- **Dimensionalidad**: Crea k columnas para k categorías (o k-1 con categoría de referencia)
- **Dispersión**: Matrices resultantes son muy dispersas (muchos ceros)
- **Colinealidad**: Riesgo de "dummy variable trap" sin categoría de referencia

4.5.3 Codificación Ordinal (variables ordinales)

La **codificación ordinal** transforma variables categóricas ordinales en números enteros que preservan el orden jerárquico natural de las categorías. Esta técnica es fundamental cuando trabajamos with variables categóricas que tienen un orden inherente y significativo, como nivel educativo, satisfacción del cliente, grado de severidad, o calificaciones de crédito.

La transformación asigna números enteros consecutivos que reflejan la jerarquía natural de las categorías, preservando tanto la información categórica como el orden relativo entre ellas.

> ⚠ Cuándo usar codificación ordinal
>
> La codificación ordinal es apropiada cuando las categorías tienen un orden natural claro y este orden es relevante para el fenómeno que estamos modelando. El modelo puede aprovechar esta información ordinal para capturar tendencias o patrones relacionados con la jerarquía de las categorías.

¿Por qué preservar el orden? Los algoritmos de machine learning pueden aprovechar la información ordinal para identificar tendencias y patrones que se perderían con one-hot encoding. Cuando el orden es significativo, la codificación ordinal es más eficiente y puede mejorar el rendimiento del modelo.

Ejemplo práctico: Consideremos una variable "Satisfacción" con valores ordenados [Muy Insatisfecho, Insatisfecho, Neutral, Satisfecho, Muy Satisfecho]. La codificación ordinal asignará:

ID	Satisfacción	Satisfacción_Codificada
1	Muy Insatisfecho	1
2	Insatisfecho	2
3	Neutral	3
4	Satisfecho	4
5	Muy Satisfecho	5

Cada observación queda representada por un número entero que preserva el orden jerárquico de las categorías originales.

Interpretación en regresión: En un modelo de regresión lineal, el coeficiente de la variable ordinal codificada representa el cambio promedio en la variable respuesta por cada incremento de una unidad en el nivel ordinal. Por ejemplo, si el coeficiente es 2.5, esto significa que pasar del nivel 1 al 2, o del 3 al 4, se asocia en promedio con un aumento de 2.5 unidades en la variable respuesta, asumiendo intervalos uniformes entre niveles.

Implementación práctica

```r
# Crear datos de ejemplo
datos <- data.frame(
  ID = 1:5,
  Satisfaccion = c("Muy Insatisfecho", "Insatisfecho", "Neutral", "Satisfecho", "Muy Satisf
)

# Convertir en factor ordenado
datos$Satisfaccion_factor <- factor(datos$Satisfaccion,
                                    levels = c("Muy Insatisfecho", "Insatisfecho", "Neutra
                                    ordered = TRUE)

# Codificación ordinal manual
datos$Satisfaccion_codificada <- as.numeric(datos$Satisfaccion_factor)

# Verificar la codificación
datos
```

```
  ID     Satisfaccion Satisfaccion_factor Satisfaccion_codificada
1  1 Muy Insatisfecho    Muy Insatisfecho                       1
2  2     Insatisfecho        Insatisfecho                       2
3  3          Neutral             Neutral                       3
4  4       Satisfecho          Satisfecho                       4
5  5   Muy Satisfecho      Muy Satisfecho                       5
```

Ejemplo de uso en regresión:

```r
# Simular variable respuesta correlacionada con el orden
set.seed(123)
datos$Puntuacion <- c(2, 4, 6, 8, 10) + rnorm(5, mean = 0, sd = 0.5)

# Modelo de regresión
modelo <- lm(Puntuacion ~ Satisfaccion_codificada, data = datos)
summary(modelo)
```

```

Call:
lm(formula = Puntuacion ~ Satisfaccion_codificada, data = datos)

Residuals:
      1       2       3       4       5
```

```
-0.2090 -0.1279  0.6826 -0.1455 -0.2002

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               -0.1552     0.4640  -0.335 0.759968
Satisfaccion_codificada    2.0840     0.1399  14.896 0.000657 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4424 on 3 degrees of freedom
Multiple R-squared:  0.9867,    Adjusted R-squared:  0.9822
F-statistic: 221.9 on 1 and 3 DF,  p-value: 0.0006565
```

Ventajas y desventajas

Ventajas de la codificación ordinal:

- **Preserva la jerarquía**: Mantiene el orden natural entre categorías
- **Eficiencia dimensional**: Una sola columna independiente del número de categorías
- **Interpretabilidad**: Coeficientes representan cambios por unidad de nivel ordinal
- **Eficiencia computacional**: Menor uso de memoria y procesamiento

Desventajas:

- **Supuesto de intervalos uniformes**: Asume que las diferencias entre niveles consecutivos son iguales
- **Riesgo con variables no-ordinales**: Puede imponer orden artificial en variables nominales
- **Pérdida de flexibilidad**: No puede capturar relaciones no-lineales entre niveles ordinales

4.5.4 Comparación directa: Ordinal vs One-Hot Encoding

La elección entre codificación ordinal y one-hot encoding depende fundamentalmente de la naturaleza de la variable categórica y los objetivos del análisis. Una decisión incorrecta puede llevar a interpretaciones erróneas y modelos subóptimos.

Característica	Codificación Ordinal	One-Hot Encoding
Preserva el orden	Sí, refleja la jerarquía entre categorías	No, trata cada categoría como independiente
Dimensionalidad	Una sola columna	k columnas (o k-1)

Característica	Codificación Ordinal	One-Hot Encoding
Adecuado para	Variables con orden natural (educación, satisfacción)	Variables nominales (color, género, región)
Interpretación	Cambio por unidad de nivel	Diferencia vs. categoría de referencia
Eficiencia computacional	Alta (menos parámetros)	Menor (más parámetros)
Riesgo principal	Orden artificial en variables nominales	Dimensionalidad excesiva

4.6 Interacciones entre variables

Las **interacciones entre variables** representan uno de los conceptos más poderosos y subestimados en el modelado estadístico. Mientras que los efectos principales capturan el impacto promedio de cada variable por separado, las interacciones revelan cómo el efecto de una variable cambia según el nivel de otra variable. Este fenómeno es omnipresente en el mundo real: el efecto del precio en las ventas depende del nivel de publicidad, el impacto de la experiencia en el salario varía según la educación, o la efectividad de un tratamiento médico puede diferir entre grupos demográficos.

! Principios para feature engineering efectivo

1. **Justificación teórica**: Cada nueva variable debe tener sentido conceptual en el dominio
2. **Validación rigurosa**: Evaluar el poder predictivo real en datos no vistos
3. **Simplicidad primero**: Preferir transformaciones simples e interpretables
4. **Documentación exhaustiva**: Registrar el proceso de creación y la lógica detrás de cada feature
5. **Monitoreo continuo**: Verificar que las relaciones se mantienen en producción

Ignorar las interacciones relevantes puede llevar a conclusiones erróneas y pérdida significativa de poder predictivo. Por otro lado, incluir interacciones irrelevantes incrementa la complejidad del modelo sin beneficios, violando el principio de parsimonia. La clave está en identificar, interpretar y validar interacciones de manera sistemática y teoricamente fundamentada.

4.6.1 Interacciones entre variables continuas

El caso más directo es la interacción entre dos variables continuas:

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

Interpretación del coeficiente de interacción (β_3):

- Si $\beta_3 > 0$: El efecto de X_1 se amplifica cuando X_2 aumenta
- Si \$ _3 < 0: El efecto de X_1 se atenúa cuando X_2 aumenta
- Si $\beta_3 = 0$: No hay interacción (efectos aditivos)

Ejemplo: Interacción precio-publicidad en ventas

```r
# Simulación: efecto de precio y publicidad en ventas
# con interacción (mayor publicidad reduce sensibilidad al precio)
set.seed(789)
n <- 200
precio <- runif(n, 50, 150)  # precio en euros
publicidad <- runif(n, 0, 10)  # gasto en publicidad (miles €)

# Efecto principal negativo del precio, positivo de publicidad
# Interacción: mayor publicidad reduce sensibilidad negativa al precio
ventas <- 1000 - 5*precio + 50*publicidad + 0.8*precio*publicidad/10 +
          rnorm(n, 0, 50)

datos_inter <- data.frame(precio, publicidad, ventas)

# Modelo con interacción
modelo_interaccion <- lm(ventas ~ precio * publicidad, data = datos_inter)
summary(modelo_interaccion)
```

```
Call:
lm(formula = ventas ~ precio * publicidad, data = datos_inter)

Residuals:
     Min       1Q   Median       3Q      Max
-158.256  -35.577    0.296   37.460  118.605

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       968.37048   27.53011  35.175   <2e-16 ***
precio             -4.74324    0.26860 -17.659   <2e-16 ***
publicidad         55.77675    4.48175  12.445   <2e-16 ***
```

```
precio:publicidad   0.03470    0.04432   0.783    0.435
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50.87 on 196 degrees of freedom
Multiple R-squared:  0.9518,    Adjusted R-squared:  0.9511
F-statistic:  1291 on 3 and 196 DF,  p-value: < 2.2e-16
```

Interpretación numérica de los coeficientes:

```
# Extraer coeficientes y p-valores para interpretación
coef_int <- coef(modelo_interaccion)
summary_model <- summary(modelo_interaccion)
p_valores <- summary_model$coefficients[, "Pr(>|t|)"]

# Crear tabla de interpretación de efectos
tabla_efectos <- data.frame(
  Nivel_Publicidad = c("0 (sin publicidad)", "5 (media)", "10 (alta)"),
  Efecto_Precio = c(
    round(coef_int[2], 2),
    round(coef_int[2] + 5*coef_int[4], 2),
    round(coef_int[2] + 10*coef_int[4], 2)
  )
)

kable(tabla_efectos,
      caption = "Efecto del precio según el nivel de publicidad (hipotético)",
      col.names = c("Nivel de Publicidad", "Efecto del Precio (€/unidad)"))
```

Table 4.6: Efecto del precio según el nivel de publicidad (hipotético)

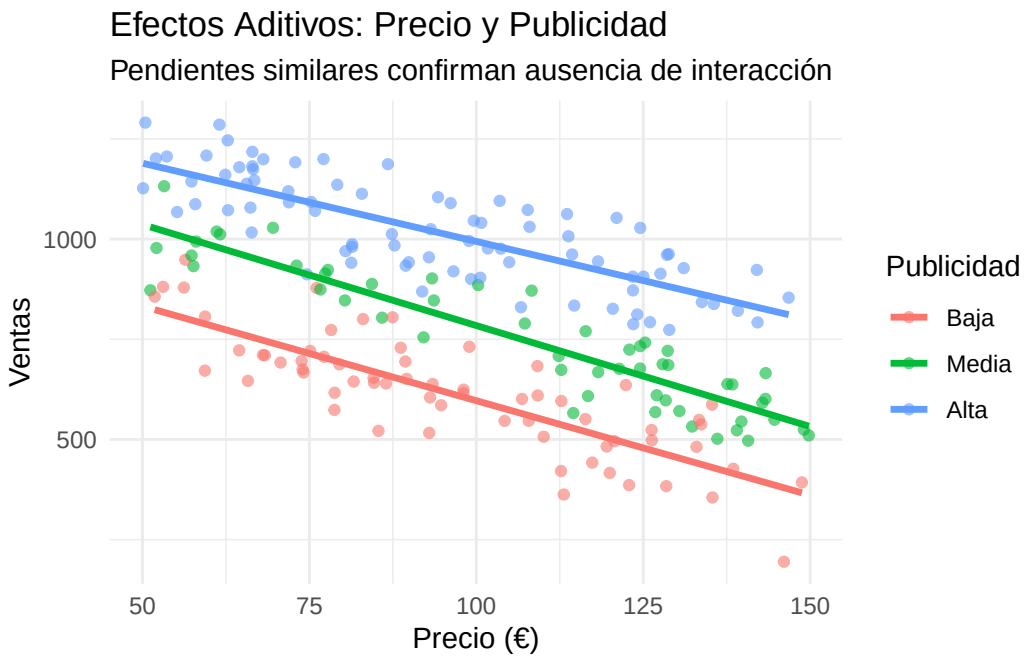
Nivel de Publicidad	Efecto del Precio (€/unidad)
0 (sin publicidad)	-4.74
5 (media)	-4.57
10 (alta)	-4.40

Evidencia estadística: La interacción tiene un coeficiente de **0.035** con un p-valor de **0.435**. Dado que $p > 0.05$, **no hay evidencia estadística de interacción** entre precio y publicidad.
Interpretación del gráfico: El primer gráfico (scatter plot con líneas de tendencia por grupos de publicidad) muestra **líneas prácticamente paralelas**, confirmando la ausencia de interacción. Aunque visualmente las pendientes parecen ligeramente diferentes, esta variación está dentro del rango esperado por el ruido aleatorio.
Implicaciones prácticas:

- El efecto del precio sobre las ventas es **constante** (-€5 por unidad) independientemente del nivel de publicidad
- Los efectos son **aditivos**: cada €1000 adicional en publicidad aumenta las ventas base en ~50 unidades, sin modificar la sensibilidad al precio
- **Modelo recomendado:** `ventas ~ precio + publicidad` (sin término de interacción)

Lección metodológica: Este caso demuestra la importancia de **confiar en la evidencia estadística formal** sobre las impresiones visuales cuando hay variabilidad muestral significativa.
Visualización para verificar ausencia de interacción:

```
# Visualización de la interacción
library(ggplot2)
# Crear grupos de publicidad para visualizar
datos_inter$pub_grupo <- cut(datos_inter$publicidad,
                             breaks = 3,
                             labels = c("Baja", "Media", "Alta"))

ggplot(datos_inter, aes(x = precio, y = ventas, color = pub_grupo)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", formula = y ~ x, se = FALSE, linewidth = 1.2) +
  labs(title = "Efectos Aditivos: Precio y Publicidad",
       subtitle = "Pendientes similares confirman ausencia de interacción",
       x = "Precio (€)", y = "Ventas", color = "Publicidad") +
  theme_minimal()
```



Interpretación crítica: Visual vs Estadística
Observando el gráfico, las líneas **parecen tener pendientes diferentes**, lo que visualmente sugeriría la presencia de interacción. Sin embargo, el análisis estadístico formal nos indica que esta diferencia **no es estadísticamente significativa** ($p = 0.435$).
¿Por qué esta aparente contradicción?

- **Variabilidad aleatoria**: Las diferencias observadas pueden deberse al ruido aleatorio en los datos
- **Tamaño de muestra**: Puede no ser suficiente para detectar una interacción débil si realmente existe
- **Poder estadístico**: El test puede no tener suficiente poder para detectar efectos pequeños
- **Agrupación artificial**: Los grupos de publicidad se crearon artificialmente para visualización, no reflejan la variable continua real

Decisión metodológica correcta:

- **Confiar en la estadística formal**: El p-valor > 0.05 indica no significancia
- **Modelo parsimonioso**: Eliminar el término de interacción no significativo
- **Interpretación conservadora**: Los efectos son aditivos hasta que se demuestre lo contrario

Lección crucial:

Este ejemplo demuestra por qué **la inspección visual nunca debe ser el único criterio** para decidir sobre la inclusión de términos de interacción. La estadística inferencial formal debe prevalecer sobre las impresiones visuales, especialmente cuando hay incertidumbre debido a la variabilidad muestral.
Modelo final recomendado: `ventas ~ precio + publicidad` (sin interacción)

4.6.2 Interacciones entre variables categóricas

Cuando ambas variables son categóricas, las interacciones representan efectos específicos de combinaciones de categorías que no pueden explicarse por los efectos principales por separado.

Para dos variables categóricas A (con niveles i) y B (con niveles j), el modelo incluye:

$$Y = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon$$

Donde $(\alpha\beta)_{ij}$ representa la interacción específica entre el nivel i de A y el nivel j de B.

La interacción $(\alpha\beta)_{ij}$ indica cuánto la combinación específica (i,j) se desvía del efecto que esperaríamos si solo sumáramos los efectos principales $\alpha_i + \beta_j$.

Ejemplo: Interacción género-departamento en salarios

```r
# Simulación: salarios por género y departamento con interacción
# (brecha salarial varía según departamento)
set.seed(456)
n <- 300

# Variables categóricas
genero <- sample(c("Masculino", "Femenino"), n, replace = TRUE)
departamento <- sample(c("Ventas", "IT", "RRHH"), n, replace = TRUE)

# Efectos principales y de interacción simulados
efecto_base <- 40000  # salario base
efecto_masculino <- ifelse(genero == "Masculino", 2000, 0)
efecto_it <- ifelse(departamento == "IT", 8000, 0)
efecto_rrhh <- ifelse(departamento == "RRHH", 3000, 0)

# Interacción: brecha de género mayor en IT
interaccion <- ifelse(genero == "Masculino" & departamento == "IT", 4000, 0)

salario <- efecto_base + efecto_masculino + efecto_it + efecto_rrhh +
           interaccion + rnorm(n, 0, 3000)

datos_cat <- data.frame(genero, departamento, salario)

# Modelo con interacción
modelo_cat <- lm(salario ~ genero * departamento, data = datos_cat)
summary(modelo_cat)
```

```
Call:
lm(formula = salario ~ genero * departamento, data = datos_cat)

Residuals:
    Min      1Q  Median      3Q     Max
-8037.9 -1930.7    48.4  1968.1  7486.9

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      47627.4      461.2 103.270  < 2e-
16 ***
generoMasculino                   6166.3      593.4  10.391  < 2e-
```

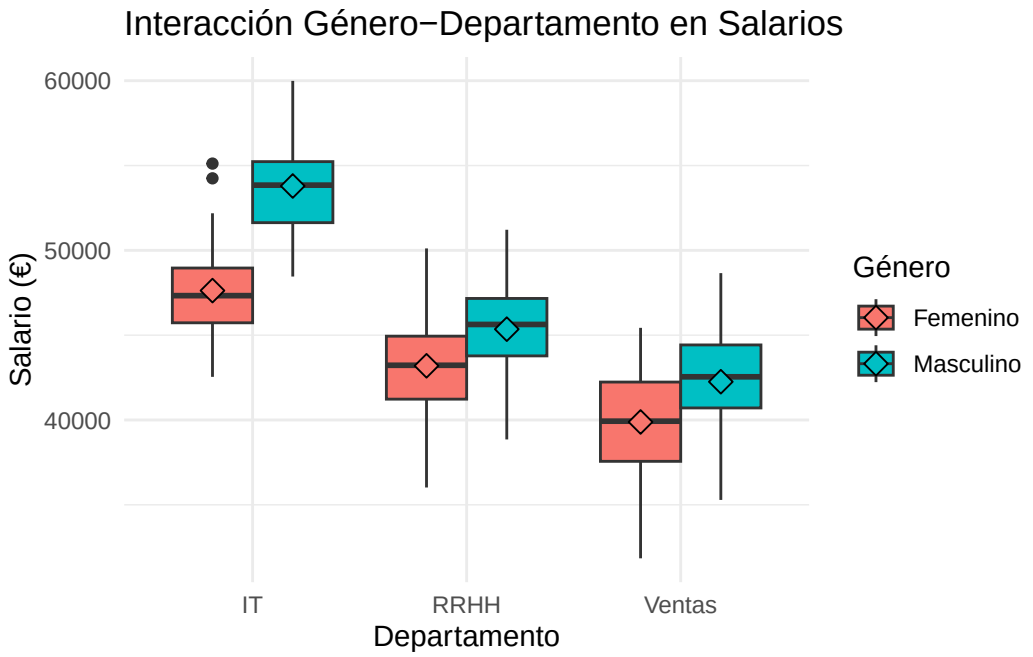
```
16 ***
departamentoRRHH                    -4433.3      633.9  -6.994 1.80e-
11 ***
departamentoVentas                  -7744.7      597.4 -12.964  < 2e-
16 ***
generoMasculino:departamentoRRHH    -4007.2      857.1  -4.675 4.48e-
06 ***
generoMasculino:departamentoVentas  -3793.8      814.4  -4.659 4.83e-
06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2917 on 294 degrees of freedom
Multiple R-squared:  0.7372,    Adjusted R-squared:  0.7327
F-statistic: 164.9 on 5 and 294 DF,  p-value: < 2.2e-16
```

```r
# Medias por grupo para interpretar la interacción
medias_grupo <- aggregate(salario ~ genero + departamento, data = datos_cat, FUN = mean)
medias_grupo <- medias_grupo[order(medias_grupo$departamento, medias_grupo$genero), ]
kable(medias_grupo, caption = "Salario promedio por género y departamento")
```

Table 4.7: Salario promedio por género y departamento

genero	departamento	salario
Femenino	IT	47627.44
Masculino	IT	53793.71
Femenino	RRHH	43194.15
Masculino	RRHH	45353.23
Femenino	Ventas	39882.76
Masculino	Ventas	42255.23

```r
# Visualización de la interacción
library(ggplot2)
ggplot(datos_cat, aes(x = departamento, y = salario, fill = genero)) +
  geom_boxplot(position = "dodge") +
  stat_summary(fun = mean, geom = "point", shape = 23, size = 3,
               position = position_dodge(0.75)) +
  labs(title = "Interacción Género-Departamento en Salarios",
       x = "Departamento", y = "Salario (€)", fill = "Género") +
  theme_minimal()
```



Evidencia visual clara: El segundo gráfico (boxplots por género y departamento) revela **patrones de interacción marcados** que se manifiestan de forma diferente en cada departamento.

Patrones específicos por departamento:

- **IT:** La brecha de género es **máxima** (~€8,000). Los hombres tienen salarios significativamente superiores y mayor variabilidad salarial
- **RRHH:** Brecha **moderada** (~€3,000) con distribuciones más similares entre géneros
- **Ventas:** **Menor brecha de género** (~€1,500), con salarios más homogéneos entre grupos

Interpretación de la interacción: Las **líneas no paralelas** en el patrón de medias confirman que el efecto del género sobre el salario **varía significativamente según el departamento**. Esto sugiere:

- Diferencias en **culturas departamentales** respecto a equidad salarial
- **Estructuras de compensación** variables entre departamentos
- Posibles diferencias en **poder de negociación** o **demanda de talento**

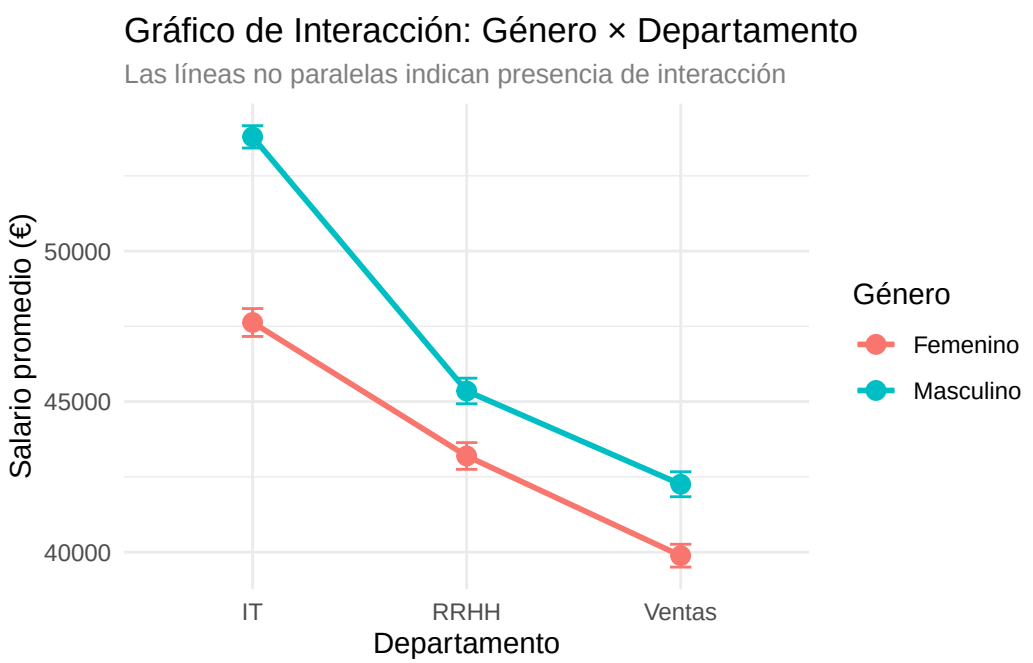
Implicaciones organizacionales: La interacción indica que las políticas salariales no son uniformes y que intervenciones de equidad deberían ser **diferenciadas por departamento**.
Gráfico de interacción clásico:

```
# Calcular medias y errores estándar por grupo para el gráfico
suppressPackageStartupMessages(library(dplyr))
medias_se <- datos_cat %>%
  group_by(genero, departamento) %>%
  summarise(
    media = mean(salario),
    se = sd(salario) / sqrt(n()),
    .groups = "drop"
  )

# Mostrar los datos calculados
kable(medias_se, caption = "Medias y errores estándar por grupo")
```

Table 4.8: Medias y errores estándar por grupo

genero	departamento	media	se
Femenino	IT	47627.44	463.4703
Femenino	RRHH	43194.15	443.9292
Femenino	Ventas	39882.76	378.5932
Masculino	IT	53793.71	371.1923
Masculino	RRHH	45353.23	425.9146
Masculino	Ventas	42255.23	414.4745

```r
# Gráfico de interacción estilo clásico
ggplot(medias_se, aes(x = departamento, y = media, color = genero, group = genero)) +
  geom_point(size = 3) +
  geom_line(linewidth = 1) +
  geom_errorbar(aes(ymin = media - se, ymax = media + se), width = 0.1) +
  labs(title = "Gráfico de Interacción: Género × Departamento",
       subtitle = "Las líneas no paralelas indican presencia de interacción",
       x = "Departamento", y = "Salario promedio (€)", color = "Género") +
  theme_minimal() +
  theme(plot.subtitle = element_text(size = 10, color = "gray50"))
```



Interpretación del gráfico: Las líneas **no son paralelas**, confirmando la presencia de interacción significativa. La brecha salarial de género varía considerablemente: mayor en IT, moderada en RRHH, y menor en Ventas.

4.6.3 Interacciones mixtas (continua × categórica)

Las interacciones mixtas son especialmente útiles para modelar cómo el efecto de una variable continua varía entre grupos categóricos. Esto es fundamental cuando sospechamos que la relación funcional cambia según el contexto definido por la variable categórica.

Formulación matemática:

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 X \cdot D + \varepsilon$$

Donde D es una variable dummy (0/1) que representa la variable categórica.

Interpretación geométrica: La interacción permite que cada grupo categórico tenga:

- **Intercepto diferente**: β_0 (grupo de referencia) vs. $\beta_0 + \beta_2$ (otro grupo)
- **Pendiente diferente**: β_1 (grupo de referencia) vs. $\beta_1 + \beta_3$ (otro grupo)

Ejemplo: Interacción experiencia-género en salarios

```
# Simulación: efecto de experiencia en salario varía según género
set.seed(789)
n <- 250

experiencia <- runif(n, 0, 20)  # años de experiencia
genero <- sample(c("Femenino", "Masculino"), n, replace = TRUE)

# Efecto diferencial: pendiente de experiencia menor para mujeres
salario_base <- 35000
efecto_experiencia_hombres <- 2000  # €2000 por año para hombres
efecto_experiencia_mujeres <- 1200  # €1200 por año para mujeres (brecha creciente)
efecto_genero_base <- ifelse(genero == "Masculino", 3000, 0)

# Crear variable dummy para interacción
dummy_masculino <- ifelse(genero == "Masculino", 1, 0)

# Salario con interacción
salario <- salario_base +
           efecto_experiencia_mujeres * experiencia +  # pendiente base (mujeres)
           efecto_genero_base * dummy_masculino +      # diferencia intercepto
           (efecto_experiencia_hombres - efecto_experiencia_mujeres) * experiencia * dummy_
           rnorm(n, 0, 4000)

datos_mixta <- data.frame(experiencia, genero, salario, dummy_masculino)

# Modelo con interacción
modelo_mixta <- lm(salario ~ experiencia * genero, data = datos_mixta)
summary(modelo_mixta)
```

```
Call:
lm(formula = salario ~ experiencia * genero, data = datos_mixta)

Residuals:
     Min       1Q   Median       3Q      Max
-12340.2  -2742.7    162.8   2586.6   9293.8

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                35650.27     765.04  46.599   <2e-16 ***
experiencia                 1123.03      69.74  16.103   <2e-16 ***
generoMasculino             2613.33    1038.08   2.517   0.0125 *
experiencia:generoMasculino  905.46      92.44   9.795   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3995 on 246 degrees of freedom
Multiple R-squared:  0.8889,    Adjusted R-squared:  0.8876
F-statistic: 656.2 on 3 and 246 DF,  p-value: < 2.2e-16
```

Interpretación detallada de los coeficientes:

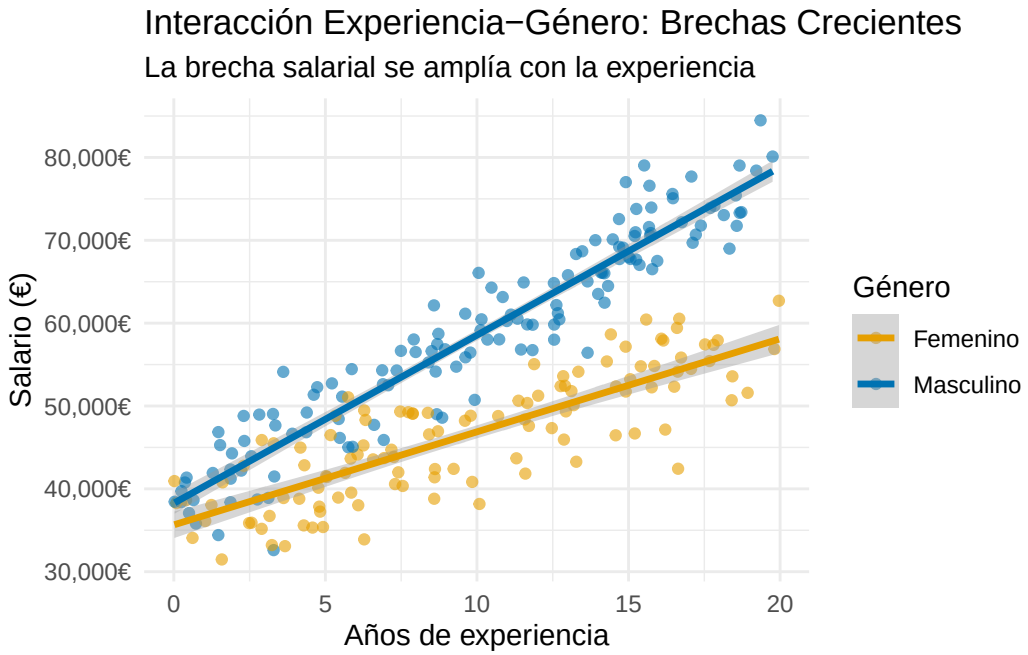
```r
# Extraer coeficientes para interpretación
coef_mixta <- coef(modelo_mixta)

# Crear tabla de interpretación por género
tabla_genero <- data.frame(
  Parámetro = c("Intercepto (salario inicial)", "Pendiente (€ por año experiencia)"),
  Mujeres = c(
    paste0("€", format(round(coef_mixta[1], 0), big.mark = ",")),
    paste0("€", round(coef_mixta[2], 0))
  ),
  Hombres = c(
    paste0("€", format(round(coef_mixta[1] + coef_mixta[3], 0), big.mark = ",")),
    paste0("€", round(coef_mixta[2] + coef_mixta[4], 0))
  ),
  Diferencia = c(
    paste0("€", format(round(coef_mixta[3], 0), big.mark = ",")),
    paste0("€", round(coef_mixta[4], 0), " adicionales")
  )
)

kable(tabla_genero,
      caption = "Comparación de parámetros del modelo por género")
```

Table 4.9: Comparación de parámetros del modelo por género

Parámetro	Mujeres	Hombres	Diferencia
Intercepto (salario inicial)	€35,650	€38,264	€2,613
Pendiente (€ por año experiencia)	€1123	€2028	€905 adicionales

Implicaciones de la interacción:
El coeficiente de interacción (**905**) indica que cada año adicional de experiencia aumenta el salario masculino en **€905 más** que el salario femenino. Esto crea una **brecha creciente**: inicialmente la diferencia es de €2,613, pero después de 20 años de experiencia, la brecha total alcanza €20,722.
Visualización de la divergencia salarial:

```
# Visualización de líneas de regresión por grupo
ggplot(datos_mixta, aes(x = experiencia, y = salario, color = genero)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", formula = y ~ x, se = TRUE, linewidth = 1.2) +
  labs(title = "Interacción Experiencia-Género: Brechas Crecientes",
       subtitle = "La brecha salarial se amplía con la experiencia",
       x = "Años de experiencia", y = "Salario (€)", color = "Género") +
  theme_minimal() +
  scale_color_manual(values = c("Femenino" = "#E69F00", "Masculino" = "#0072B2")) +
  scale_y_continuous(labels = scales::comma_format(suffix = "€"))
```



Evidencia de brechas crecientes: El tercer gráfico (scatter plot con líneas de regresión) demuestra claramente el patrón de **interacción experiencia-género** mediante **líneas divergentes** con pendientes notablemente diferentes.
Interpretación cuantitativa de la divergencia:

- **Punto de inicio (0 años):** Brecha inicial de €2,613 a favor de los hombres

- **Pendientes diferenciadas:** Los hombres ganan €2028 adicionales por año vs. €1123 para las mujeres
- **Brecha acumulativa:** Cada año adicional de experiencia **amplía la brecha** en €905 adicionales

Implicaciones del patrón de interacción: La **divergencia progresiva** visible en las líneas revela que:

- El **retorno a la experiencia** es sistemáticamente **mayor para hombres** que para mujeres
- A los 20 años de experiencia, la brecha total alcanza €20,722 (inicial + acumulativa)
- Este patrón sugiere **barreras estructurales** que impiden que las mujeres capitalicen plenamente su experiencia

Significancia social: Esta interacción documenta un fenómeno preocupante donde la **inequidad salarial se agrava con el tiempo**, indicando que las brechas de género no son meramente diferencias de entrada sino **desventajas acumulativas** a lo largo de la carrera profesional.

4.6.4 Identificación y detección de interacciones

La detección sistemática de interacciones requiere combinar justificación teórica, exploración visual y validación estadística. No debemos buscar interacciones aleatoriamente, sino guiados por el conocimiento del dominio y patrones observables en los datos.

La justificación teórica previa es fundamental: la teoría del dominio debe sugerir dónde pueden existir interacciones. Por ejemplo, en economía esperamos efectos precio-publicidad o educación-experiencia; en medicina son comunes las interacciones dosis-edad o tratamiento-comorbilidad; en marketing encontramos interacciones producto-canal o temporada-promoción.

La exploración visual sistemática complementa la teoría con evidencia empírica. Para variables continuas utilizamos gráficos de dispersión coloreados por grupos categóricos; para variables categóricas empleamos gráficos de interacción (interaction plots); para interacciones mixtas analizamos líneas de regresión por grupo.

Los tests estadísticos formales proporcionan validación objetiva: el test F para interacciones compara modelos con y sin términos de interacción, el test de significancia individual evalúa coeficientes específicos mediante t-test, y los criterios de información (AIC/BIC) guían la selección entre modelos alternativos.

Estrategia de modelado jerárquico

Principio de jerarquía: Si incluimos una interacción A×B, siempre debemos incluir los efectos principales A y B, incluso si no son significativos individualmente. Esto preserva la interpretabilidad y evita sesgos en los coeficientes de interacción.
Proceso de construcción del modelo:

1. **Modelo base**: Solo efectos principales
2. **Modelo con interacciones**: Agregar términos de interacción teoricamente justificados
3. **Comparación**: Test F para evaluar mejora significativa
4. **Selección**: Usar criterios estadísticos y de parsimonia
5. **Validación**: Verificar supuestos y estabilidad en datos de prueba

4.6.5 Consideraciones prácticas y limitaciones

Las interacciones incrementan exponencialmente la complejidad interpretativa del modelo. Un modelo con k efectos principales puede tener hasta $k(k-1)/2$ interacciones de segundo orden, y el número crece exponencialmente con interacciones de orden superior. Esta explosión combinatorial hace que incluso modelos aparentemente simples se vuelvan rápidamente inmanejables desde el punto de vista interpretativo. Como reglas prácticas para la complejidad, recomendamos limitar a máximo 2-3 interacciones de segundo orden en modelos explicativos, evitar interacciones de tercer orden salvo justificación teórica muy sólida, y priorizar interacciones con efectos grandes sobre mera significancia estadística.

Advertencia sobre interpretación

Cuidado con la interpretación automática. Las interacciones en escalas transformadas tienen significados diferentes que en escalas originales. Siempre verificar la interpretación en el contexto de la transformación aplicada y considerar la retransformación para comunicación con audiencias no técnicas.

Un problema adicional surge cuando las interacciones crean multicolinealidad severa, especialmente cuando las variables principales están correlacionadas, se incluyen múltiples interacciones con variables comunes, o se usan variables categóricas con muchos niveles. Esta multicolinealidad puede hacer que los coeficientes individuales sean inestables y difíciles de interpretar, incluso cuando el modelo en conjunto funcione bien predictivamente. Las estrategias de mitigación incluyen el centrado de variables continuas para reducir la correlación entre X y X×Z, la selección cuidadosa de interacciones sin incluir todas las combinaciones posibles, y el uso de métodos de regularización como Ridge o Lasso cuando hay múltiples interacciones.

La situación se complica aún más cuando las variables están transformadas (logarítmica, Box-

Cox). En estos casos, la interpretación de las interacciones adquiere significados completamente diferentes que en escalas originales. Por ejemplo, en un modelo como $\log(Y) = \beta_0 + \beta_1 \log(X_1) + \beta_2 X_2 + \beta_3 \log(X_1) \cdot X_2 + \varepsilon$, el coeficiente β_3 representa cómo cambia la elasticidad de Y respecto a X cuando X aumenta en una unidad, lo que requiere una interpretación mucho más sofisticada que una interacción en escalas lineales.

! Principios para el uso de interacciones

1. **Justificación teórica primero**: No buscar interacciones sin base conceptual
2. **Principio de jerarquía**: Mantener efectos principales cuando se incluyen interacciones
3. **Parsimonia**: Preferir modelos simples que expliquen bien sobre modelos complejos
4. **Validación robusta**: Verificar estabilidad en múltiples contextos
5. **Interpretación cuidadosa**: Asegurar comprensión completa antes de conclusiones
6. **Comunicación efectiva**: Usar visualizaciones para explicar efectos complejos

Las interacciones son herramientas poderosas que pueden revelar patrones importantes ocultos en los efectos principales. Sin embargo, su uso requiere disciplina metodológica, justificación teórica sólida, y validación rigurosa para evitar conclusiones espurias y modelos sobreajustados.

4.7 Ingeniería de características avanzada: combinaciones, ratios y transformaciones

Más allá de las transformaciones individuales e interacciones, la ingeniería de características avanzada implica crear nuevas variables predictivas mediante combinaciones matemáticas, ratios y transformaciones compuestas que capturen relaciones complejas no evidentes en las variables originales (Kuhn and Johnson 2019). Esta aproximación es fundamental cuando las variables individuales contienen información parcial que, al combinarse, revelan patrones predictivos más potentes.

4.7.1 Combinaciones lineales y no lineales

Las combinaciones lineales crean nuevas variables mediante sumas ponderadas de variables existentes, útiles especialmente cuando trabajamos con variables que miden aspectos relacionados del mismo fenómeno pero con diferentes escalas o unidades.

Ejemplos de combinaciones lineales efectivas:

- **Índices compuestos**: Combinan múltiples indicadores en un score único que captura un constructo multidimensional. Por ejemplo, un índice de riesgo cardiovascular podría definirse como `índice_salud = 0.4×presión_arterial_normalizada + 0.3×colesterol_normalizada + 0.3×IMC_normalizado`. Los pesos (0.4, 0.3, 0.3) reflejan la importancia relativa establecida por evidencia médica, creando una métrica integrada que es más informativa que cualquier indicador individual. Este tipo de índices son especialmente valiosos en dominios donde múltiples factores contribuyen conjuntamente al outcome de interés.
- **Scores balanceados**: Representan equilibrios o trade-offs entre dimensiones competitivas. Un ejemplo típico es `balance_trabajo_vida = horas_trabajo / (tiempo_personal + tiempo_familia + tiempo_descanso)`. Esta métrica captura no solo la intensidad laboral, sino también su contexto relativo dentro del estilo de vida completo. Valores altos indican desbalance hacia el trabajo, mientras que valores cercanos a 1.0 sugieren equilibrio saludable. Los scores balanceados son fundamentales para capturar dinámicas de compensación que no son evidentes en variables absolutas.
- **Factores sintéticos**: Cuando múltiples variables correlacionadas miden aspectos del mismo constructo subyacente, pueden condensarse en un factor común que preserve la información esencial eliminando redundancia. Por ejemplo, si tenemos variables `ingresos`, `educación`, y `prestigio_ocupacional` (todas correlacionadas), podemos crear un factor `estatus_socioeconomico` que capture la varianza común. Esto es especialmente útil cuando la colinealidad entre predictores compromete la estabilidad del modelo, pero cada variable aporta información valiosa.

Las combinaciones no lineales van más allá de las sumas ponderadas para capturar interacciones multiplicativas, sinergias y compensaciones entre variables mediante productos, cocientes, potencias y funciones más complejas:

- **Productos de eficiencia**: Capturan sinergias multiplicativas donde el rendimiento depende de la combinación simultánea de múltiples factores. Por ejemplo, `rendimiento_efectivo = capacidad_instalada × utilización_porcentual × factor_calidad`. Esta métrica reconoce que el rendimiento real no es aditivo: tener alta capacidad pero baja utilización, o alta utilización con problemas de calidad, resulta en rendimiento subóptimo. Los productos de eficiencia son esenciales en contextos operacionales donde el desempeño emerge de la coordinación entre recursos.
- **Ratios de rendimiento ajustado por riesgo**: Normalizan beneficios por su costo o riesgo asociado, creando métricas comparables entre contextos diferentes. Un ejemplo financiero sería `eficiencia_ajustada = (rendimiento_esperado - tasa_libre_riesgo) / (volatilidad + costos_transaccion)`. Esta formulación reconoce que los rendimientos absolutos son engañosos sin considerar el riesgo asumido y los costos incurridos. Los ratios ajustados son cruciales para decisiones de optimización donde debemos comparar alternativas con perfiles de riesgo-retorno heterogéneos.

- **Funciones de utilidad**: Capturan percepciones subjetivas o valores no lineales mediante transformaciones que reflejan preferencias reales. Por ejemplo, `valor_percibido = √(calidad_producto) × precio ·` reconoce que la utilidad del consumidor tiene rendimientos decrecientes tanto en calidad como en ahorro de precio. La raíz cuadrada de la calidad refleja que mejoras incrementales tienen menor impacto en niveles altos, mientras que el exponente negativo del precio captura la sensibilidad decreciente a cambios de precio en productos caros.

4.7.2 Ratios y proporciones como features

Los ratios son especialmente poderosos porque normalizan automáticamente las diferencias de escala y pueden revelar relaciones proporcionales fundamentales que permanecen ocultas en variables absolutas. A diferencia de las medidas absolutas, los ratios capturan relaciones estructurales que son invariantes bajo cambios de escala y contexto, lo que los convierte en herramientas fundamentales para crear features robustos y comparables.

La potencia de los ratios radica en su capacidad para transformar información absoluta en información relativa. Por ejemplo, una empresa con €1M en ventas y €100K en marketing tiene un ratio ventas/marketing de 10, igual que una empresa con €10M en ventas y €1M en marketing. Esta normalización automática permite comparaciones directas y elimina sesgos de tamaño que podrían distorsionar el análisis.

Categorización detallada de ratios efectivos:

1. **Ratios de eficiencia**: Miden qué tan efectivamente se convierten los inputs en outputs, revelando productividad y optimización operacional. Ejemplos fundamentales incluyen:

   - `ROI_marketing = (ventas_generadas - gasto_marketing) / gasto_marketing`: Captura el retorno neto por euro invertido
   - `eficiencia_produccion = unidades_producidas / (horas_trabajo + costo_materiales)`: Normaliza productividad por recursos consumidos
   - `conversion_rate = ventas_completadas / visitantes_web`: Revela la efectividad del funnel de conversión

   Estos ratios son especialmente valiosos porque eliminan el efecto escala y permiten comparar unidades de diferentes tamaños en términos de eficiencia pura.

2. **Ratios de riesgo ajustado**: Normalizan retornos o beneficios por la incertidumbre o costo asociado, proporcionando métricas de valor ajustado por riesgo:

   - `sharpe_ratio = (rendimiento_promedio - tasa_libre_riesgo) / volatilidad`: Mide retorno por unidad de riesgo asumido
   - `stability_score = beneficio_promedio / desviacion_estandar_beneficios`: Indica consistencia en el desempeño

- `risk_adjusted_growth = crecimiento_promedio / max_drawdown`: Captura crecimiento sostenible

Estos ratios son cruciales en análisis financiero y gestión de riesgos, donde los valores absolutos pueden ser engañosos sin considerar la variabilidad subyacente.

3. **Ratios temporales**: Capturan dinámicas y tendencias mediante comparaciones entre períodos, revelando momentum y patrones estacionales:

   - `momentum_growth = crecimiento_último_trimestre / crecimiento_promedio_histórico`: Identifica aceleración o desaceleración
   - `estacionalidad = ventas_período_actual / media_móvil_12_meses`: Captura variaciones cíclicas
   - `trend_strength = (valor_actual - valor_hace_12_meses) / volatilidad_histórica`: Mide significancia de cambios

   Estos ratios son especialmente útiles en análisis de series temporales donde necesitamos distinguir entre variación normal y cambios estructurales significativos.

4. **Ratios de composición**: Revelan la estructura interna de agregados mediante proporciones parte-todo, fundamentales para análisis de portafolios y segmentación:

   - `concentracion_cliente = ventas_top3_clientes / ventas_totales`: Mide dependencia y riesgo de concentración
   - `diversificacion_producto = 1 - suma(proportion_i²)`: Índice de Herfindahl para medir dispersión
   - `market_share = ventas_empresa / ventas_mercado_total`: Posición relativa competitiva

   Los ratios de composición son esenciales para gestión de riesgos y análisis estratégico, revelando vulnerabilidades y fortalezas estructurales.

Ventajas metodológicas profundizadas:

- **Normalización automática**: Los ratios eliminan efectos de escala absoluta, haciendo comparables entidades de diferentes tamaños. Una startup con €10K en ventas y €2K en marketing tiene el mismo ratio ventas/marketing (5.0) que una multinacional con €100M y €20M respectivamente, permitiendo benchmarking directo de eficiencia.
- **Interpretación intuitiva**: Los ratios tienen significados naturales que facilitan la comunicación con stakeholders. Un ratio deuda/patrimonio de 0.3 es inmediatamente comprensible como "30 céntimos de deuda por cada euro de patrimonio", mientras que valores absolutos requieren más contexto.

- **Robustez ante outliers**: Los ratios suelen ser menos sensibles a valores extremos que las variables absolutas. Si una empresa tiene ventas anómalamente altas pero también marketing proporcionalmente alto, el ratio ventas/marketing permanece estable, mientras que ambas variables individuales serían outliers.
- **Invarianza bajo transformaciones**: Los ratios mantienen sus relaciones bajo cambios de unidades o inflación. El ratio precio/ingresos de una acción es el mismo si se mide en euros o dólares, proporcionando estabilidad interpretativa a lo largo del tiempo y contextos.

Consideraciones para construcción robusta de ratios:

Los ratios requieren cuidado especial en su construcción para evitar interpretaciones erróneas o inestabilidad numérica. Es fundamental evitar denominadores cercanos a cero, considerar transformaciones logarítmicas para ratios con rangos amplios, y validar que el ratio tenga significado conceptual en el dominio de aplicación.

4.7.3 Tratamiento de variables colineales mediante feature engineering

Cuando enfrentamos multicolinealidad entre predictores informativos, la ingeniería de características ofrece alternativas más sofisticadas que simplemente eliminar variables o usar interacciones sin efectos principales. Este escenario es común en la práctica: tenemos múltiples variables que aportan información valiosa individualmente, pero están suficientemente correlacionadas como para crear problemas de estabilidad e interpretación en el modelo.

El enfoque tradicional de "eliminar variables correlacionadas" es problemático porque puede resultar en pérdida significativa de información predictiva. Si `variable_A` y `variable_B` tienen correlación $r = 0.75$, ambas comparten 56% de varianza, pero cada una retiene 44% de información única. Eliminar cualquiera de ellas descarta información potencialmente valiosa que podría mejorar el poder predictivo del modelo.

La ingeniería de características para colinealidad busca condensar la información redundante mientras preserva la información única, creando nuevas variables que capturen la esencia predictiva de las variables originales sin los problemas de multicolinealidad. Esta aproximación es especialmente valiosa cuando la correlación entre variables tiene significado teórico: por ejemplo, diferentes medidas de solvencia financiera que capturan aspectos relacionados pero distintos del riesgo crediticio.

Estrategias avanzadas de condensación de información:

1. **Componentes principales (PCA)**: Extraen direcciones de máxima varianza común, creando variables ortogonales que preservan la mayor cantidad de información con la menor dimensionalidad:

```r
# Ejemplo: Variables financieras correlacionadas
pc_financiero <- prcomp(~ ingresos + patrimonio + crédito + liquidez, scale = TRUE)
score_financiero <- pc_financiero$x[,1]  # Primer componente (mayor varianza)
score_diversificacion <- pc_financiero$x[,2]  # Segundo componente (varianza residual)
```

Ventajas del PCA: Elimina completamente la multicolinealidad, preserva máxima varianza, proporciona interpretación de "factores latentes". Desventajas: Pérdida de interpretabilidad directa, todos los componentes dependen de todas las variables originales, sensible a outliers.

2. **Ratios informativos**: Crean cocientes que preservan la información relativa más relevante, eliminando efectos de escala común:

```r
# Ratios que capturan relaciones estructurales fundamentales
ratio_debt_income <- deuda_total / ingresos_anuales  # Capacidad de endeudamiento
ratio_assets_equity <- activos / patrimonio_neto     # Apalancamiento
ratio_liquidity <- activos_liquidos / pasivos_corrientes  # Solvencia a corto plazo
```

Ventajas de los ratios: Mantienen interpretabilidad económica directa, eliminan efectos de escala, capturan relaciones estructurales clave. Aplicabilidad: Especialmente efectivos cuando las variables correlacionadas miden aspectos del mismo fenómeno subyacente (ej. diferentes medidas de tamaño empresarial).

3. **Índices ponderados**: Combinan variables usando pesos derivados de conocimiento teórico o empírico, creando métricas compuestas más robustas que sus componentes individuales:

```r
# Índice de solvencia con pesos basados en evidencia empírica
indice_solvencia <- 0.4 * (ingresos/gastos) + 0.3 * (activos/deudas) + 0.3 * score_credi

# Índice de crecimiento balanceado
indice_crecimiento <- 0.5 * crecimiento_ventas + 0.3 * crecimiento_beneficios + 0.2 * cr
```

Determinación de pesos: Pueden derivarse de análisis factorial confirmatorio, regresión ridge, conocimiento experto, o optimización empírica. Los pesos deben justificarse teóricamente y validarse en datos independientes.

4. **Diferencias y cambios relativos**: Capturan dinámicas temporales y patrones de co-movimiento que revelan información única no presente en niveles absolutos:

```r
# Dinámicas de crecimiento relativo
crecimiento_relativo <- (valor_actual - valor_anterior) / valor_anterior
aceleracion <- (crecimiento_t - crecimiento_t_1) / crecimiento_t_1

# Medidas de estabilidad y volatilidad
```

```
volatilidad <- sd(ultimos_12_meses) / mean(ultimos_12_meses)
consistencia <- 1 / (1 + cv(ultimos_periodos))  # Coeficiente de variación invertido
```

Aplicabilidad temporal: Especialmente útiles para series temporales donde variables están correlacionadas en niveles pero divergen en tasas de cambio, revelando dinámicas diferenciales ocultas en análisis de niveles.

Criterios de selección de estrategia

- **PCA**: Cuando la interpretabilidad no es crítica y maximizar la retención de varianza es prioritario
- **Ratios**: Cuando existe significado teórico claro en las relaciones proporcionales entre variables
- **Índices ponderados**: Cuando hay conocimiento previo sobre la importancia relativa de cada componente
- **Cambios relativos**: Cuando las dinámicas temporales son más informativas que los niveles absolutos

La ingeniería de características es tanto arte como ciencia: requiere creatividad para identificar combinaciones útiles, pero también rigor metodológico para validar que las nuevas variables realmente aportan valor predictivo estable y generalizable.

5 Selección de variables, regularización y validación

En los modelos de regresión, especialmente cuando se trabaja con conjuntos de datos que incluyen un gran número de variables predictoras, es común enfrentarse al desafío ntficar qué variables son realmente relevantes para explicar la variable respuesta. La inclusión de demasiadas variables en un modelo puede llevar a problemas como el sobreajuste, pérdida de interpretabilidad y complejidad innecesaria, mientras que la exclusión de variables importantes puede resultar en modelos subóptimos.

Este tema aborda uno de los aspectos más críticos en la construcción de modelos de regresión: cómo seleccionar el subconjunto óptimo de variables predictoras y cómo validar la calidad del modelo resultante. Una vez realizado el análisis exploratorio y el ajuste inicial del modelo, surge la necesidad crítica de optimizar la selección de variables. Cuando se dispone de p variables explicativas, es posible construir hasta 2^p modelos diferentes considerando todas las combinaciones posibles. Sin embargo, explorar de manera exhaustiva todos estos modelos puede ser computacionalmente inviable cuando p es grande.

Para superar este desafío, en este tema nos enfocaremos en cinco enfoques principales:

1. **Filtrado basado en información básica**: Eliminación preliminar de variables irrelevantes mediante criterios básicos (variabilidad, correlación, VIF)
2. **Criterios de bondad de ajuste**: Métricas para comparar modelos con diferente número de variables (AIC, BIC, Cp de Mallows)
3. **Métodos de selección exhaustiva**: Evaluación sistemática de todas las combinaciones posibles (Best Subset Selection)
4. **Métodos automáticos paso a paso**: Selección iterativa mediante algoritmos forward, backward y stepwise
5. **Métodos basados en regularización**: Técnicas que penalizan la complejidad del modelo (Ridge, Lasso, Elastic Net)
6. **Validación del modelo**: Evaluación rigurosa de la capacidad predictiva mediante división train/test y validación cruzada

Cada enfoque tiene sus propias ventajas y limitaciones, siendo apropiados para diferentes situaciones según el tamaño del dataset, el número de variables y los objetivos del análisis. El objetivo es presentar las técnicas más relevantes para la selección de variables y regularización, entender sus fundamentos teóricos, y aplicarlas a casos prácticos, culminando con métodos robustos de validación que aseguren la calidad y generalización del modelo final.

5.1 Proceso completo de construcción y optimización del modelo

La construcción de un modelo de regresión múltiple es un proceso sistemático que busca explicar la relación entre una variable respuesta (Y) y múltiples variables predictoras ($X_1, X_2, \ldots, X_k$). Este proceso consta de varias etapas clave (Kutner et al. 2005), que en este tema nos enfocaremos particularmente en las etapas de reducción de variables y validación:

1. **Definición del problema y variables de interés:**
    - Identificar claramente el objetivo del análisis, ya sea realizar predicciones, evaluar relaciones o controlar por efectos de variables confusoras.
    - Seleccionar las variables predictoras potenciales en función de su relevancia teórica, conocimiento previo o exploración inicial de los datos.
2. **Recogida de datos:**

- La calidad de los datos recogidos influye directamente en la validez de los resultados y conclusiones obtenidas. El proceso de recogida de datos consiste en recopilar información de manera organizada y sistemática para responder a las preguntas de investigación planteadas. Dependiendo del diseño del estudio y los objetivos del análisis, se pueden emplear diferentes tipos de experimentos o métodos de recogida de datos.
- Debemos asegurar las siguientes características sobre los datos.
    - **Fiabilidad:** Asegurar que los datos sean consistentes y puedan reproducirse bajo condiciones similares.
    - **Validez:** Garantizar que los datos recojan realmente la información necesaria para responder a las preguntas de investigación.
    - **Ética:** Asegurar la privacidad y el consentimiento informado de los participantes.
    - **Control de Sesgos:** Diseñar el estudio de manera que se minimicen los sesgos que puedan distorsionar los resultados.

> **Tipos de experimentos**
>
> La elección del tipo de experimento o método de recogida de datos dependerá de la naturaleza del problema a investigar, los recursos disponibles y las limitaciones del estudio. Una correcta planificación y ejecución de esta etapa sienta las bases para un análisis robusto y confiable.

1. **Experimentos controlados:**

   - Los experimentos controlados son diseñados de manera que los investigadores manipulan deliberadamente una o más variables independientes (llamadas factores o variables controladas) para observar su efecto en la variable dependiente.
   - Incluyen la aleatorización de sujetos entre grupos (por ejemplo, grupos de control y tratamiento) para minimizar sesgos y asegurar comparabilidad.
   - En muchas ocasiones la información suplementaria no se puede incorporar en el diseño del experimento. A esas variables, no controladas, se les suel llamar covariables.
   - **Ejemplo:** Un estudio clínico donde se prueba un nuevo medicamento y se compara su efecto con un placebo.

2. **Estudios observacionales exploratorios:**

   - En este enfoque, los datos se recogen sin intervenir ni manipular las condiciones. Los investigadores observan y registran los fenómenos tal como ocurren en la naturaleza.
   - Pueden clasificarse en:
     - **Estudios transversales:** Los datos se recogen en un único punto temporal.
     - **Estudios longitudinales:** Los datos se recogen durante un periodo para analizar cambios a lo largo del tiempo.
   - **Ejemplo:** Investigar los hábitos alimenticios y su asociación con enfermedades cardiovasculares en una población.

3. **Estudios observacionales confirmatorios:**

   - En este enfoque, los datos se recogen para testear (confirmar o no) hipótesis derivadas de estudios previos o de ideas que pueden tener los investigadores.
   - En este contexto, las variables que aparecen involucradas en la hipótesis que se quiere confirmar se denominan variables primarias, y las variables explicativas que se sabe inluyen en la respuesta se llaman variables de control (en Epidemiología nos referimos a ellas como factores de riesgo)
   - **Ejemplo:** Un equipo de investigadores, basándose en estudios previos, plantea la hipótesis de que existe una relación positiva entre el hábito de fumar (variable explicativa principal) y la incidencia de cáncer de pulmón (variable respuesta). Para confirmar esta hipótesis, realizan un estudio observacional en el que recopilan datos de una población durante un periodo determinado. Dado que no es ético inducir a las personas a fumar para realizar un experimento controlado, este estudio se realiza de forma observacional. Los datos se analizan para evaluar la asociación entre las variables, permitiendo confirmar (o refutar)

la hipótesis planteada con un diseño adecuado y controlando los posibles factores de confusión.

4. **Encuestas y cuestionarios:**
    - Las encuestas son una técnica común para recoger datos de manera estructurada sobre actitudes, opiniones, comportamientos o características demográficas.
    - Pueden aplicarse en formato presencial, en línea, por teléfono o mediante correo.
    - **Ejemplo:** Una encuesta para medir el grado de satisfacción de los clientes con un servicio.
5. **Experimentos naturales:**
    - Se producen cuando un fenómeno natural o social actúa como una intervención en un entorno sin que los investigadores tengan control sobre el experimento.
    - Este tipo de estudio aprovecha eventos únicos para analizar sus impactos.
    - **Ejemplo:** Estudiar los efectos económicos de una nueva política fiscal aplicada en una región específica.
6. **Estudios de simulación:**
    - Los datos se generan a través de modelos matemáticos o computacionales que representan un sistema real o hipotético.
    - Este método se usa cuando es difícil o costoso realizar experimentos reales.
    - **Ejemplo:** Simular el comportamiento de un mercado financiero bajo diferentes escenarios económicos.
7. **Recogida de datos secundarios:**
    - En lugar de recoger datos nuevos, se utilizan datos ya existentes recopilados por terceros, como censos, registros administrativos o bases de datos públicas.
    - Aunque es eficiente en tiempo y costos, el investigador tiene menor control sobre la calidad y las características de los datos.
    - **Ejemplo:** Analizar datos de encuestas nacionales para estudiar tendencias sociales.

3. **Análisis Exploratorio de Datos (EDA):**
    - Inspeccionar los datos mediante análisis descriptivo y visual para identificar posibles problemas como valores atípicos, datos faltantes y multicolinealidad.
    - Escalar o transformar las variables si es necesario, especialmente si están en diferentes escalas o presentan distribuciones no lineales.
4. **Ajuste del modelo:**
    - Especificar el modelo de regresión múltiple en su forma general:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon,$$

donde ε representa los errores aleatorios.

- Estimar los coeficientes del modelo $(\beta_0, \beta_1, \ldots, \beta_p)$ utilizando el método de mínimos cuadrados, que minimiza la suma de los errores al cuadrado.

5. **Evaluación del modelo:**

   - Analizar el ajuste general del modelo utilizando métricas como R^2 y R^2 ajustado, que miden la proporción de la variabilidad explicada.
   - Examinar la tabla ANOVA para evaluar la significancia global del modelo.
   - Realizar pruebas de hipótesis para los coeficientes individuales, verificando si las variables predictoras tienen un efecto significativo en la variable respuesta.

6. **Diagnóstico del modelo:**

   - Examinar los residuos para evaluar supuestos como la linealidad, homocedasticidad, normalidad de los errores y ausencia de autocorrelación.
   - Identificar observaciones atípicas, leverage y puntos de influencia utilizando herramientas como la distancia de Cook, DFBETAS y DFFITS.

7. **Reducción de variables:**

   - En análisis de regresión, especialmente cuando se trabaja con conjuntos de datos de alta dimensionalidad, es común enfrentar situaciones en las que el número de variables explicativas es muy grande. Esto puede llevar a problemas como el sobreajuste, dificultades en la interpretación del modelo y una mayor complejidad computacional. Por ello, reducir el número de variables explicativas, sin perder información relevante, se convierte en un paso crucial para construir modelos más eficientes y robustos.

8. **Validación del modelo:**

   - Evaluar el rendimientodel modelo con datos de validación o mediante técnicas como validación cruzada para garantizar su capacidad predictiva en nuevos conjuntos de datos.

5.2 Filtrado basado en información básica

Antes de aplicar métodos sofisticados de selección de variables, es fundamental realizar un filtrado preliminar basado en información básica. Este primer paso consiste en identificar y descartar variables que claramente no aportan información relevante al modelo, reduciendo significativamente el espacio de búsqueda y mejorando la eficiencia de los métodos posteriores (James et al. 2013).

Los criterios principales para este filtrado incluyen:

1. Variabilidad de las variables predictoras

Variables con varianza muy baja o constantes proporcionan poca información discriminatoria. Se descartan variables donde:

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

para algún umbral pequeño ϵ (típicamente $\epsilon = 0.01$).

2. Correlación con la variable respuesta

Variables con correlación muy baja con Y pueden ser candidatas a eliminación. Se calcula:

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

y típicamente se establece un umbral mínimo $|r_{X_j,Y}| > \delta$ (ej: $\delta = 0.1$).

3. Multicolinealidad extrema

Variables altamente correlacionadas entre sí pueden ser redundantes. Se calcula:

$$r_{X_j,X_k} = \frac{\text{Cov}(X_j, X_k)}{\sqrt{\text{Var}(X_j)\text{Var}(X_k)}}$$

Si $|r_{X_j,X_k}| > 0.95$, se considera eliminar una de las dos variables.

4. Factor de Inflación de la Varianza (VIF)

Para detectar multicolinealidad más compleja se calcula:

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

donde R_j^2 es el coeficiente de determinación de la regresión de X_j sobre las demás variables predictoras. Valores $VIF_j > 10$ indican multicolinealidad problemática.

Ejemplo de filtrado inicial

En este ejemplo aplicamos el proceso completo de filtrado basado en información a un conjunto de datos simulado con diferentes características.

```r
# Configuración y generación de datos
set.seed(123)
n <- 100
p <- 15

# Generar datos con diferentes características
X <- matrix(rnorm(n * p), n, p)
colnames(X) <- paste0("X", 1:p)

# Variable constante (sin variabilidad)
X[, 1] <- 5

# Variable con muy baja variabilidad
X[, 2] <- 5 + rnorm(n, 0, 0.01)

# Variables moderadamente correlacionadas
X[, 4] <- X[, 3] + rnorm(n, 0, 0.5)
X[, 5] <- 0.7 * X[, 3] + rnorm(n, 0, 0.6)

# Variable respuesta con coeficientes conocidos
beta <- c(0, 0, 2, 1.5, 1.2, -1, 0.8, rep(0, 8))
y <- X %*% beta + rnorm(n)

datos <- data.frame(y = y, X)

suppressPackageStartupMessages(library(car))

# 1. Análisis de variabilidad
varianzas <- apply(X, 2, var)
vars_baja_var <- which(varianzas < 0.01)

# 2. Filtrar por correlación con Y
X_filtrada <- if(length(vars_baja_var) > 0) X[, -vars_baja_var] else X
correlaciones <- cor(X_filtrada, y)
vars_baja_corr_idx <- which(abs(correlaciones) < 0.1)
vars_baja_corr <- if(length(vars_baja_corr_idx) > 0) {
  as.numeric(gsub("X", "", colnames(X_filtrada)[vars_baja_corr_idx]))
} else {
  c()
}

# 3. Identificar correlaciones altas entre predictores
cor_matrix <- cor(X_filtrada)
high_cor <- which(abs(cor_matrix) > 0.8 & abs(cor_matrix) < 1, arr.ind = TRUE)

# 4. Calcular VIF para variables restantes
vars_eliminar <- unique(c(vars_baja_var, vars_baja_corr))
datos_final <- if(length(vars_eliminar) > 0) {
  datos[, -(vars_eliminar + 1)] # +1 porque datos incluye y en primera columna
} else {
  datos
}
```

Resultados del filtrado basado en información:
1. Variables eliminadas por baja variabilidad: X 1, X 2
2. Variables eliminadas por baja correlación con Y: X 8, X 9, X 10, X 11, X 12, X 13, X 14
3. Correlaciones altas entre predictores ($|r| > 0.8$): X4 y X3 : 0.846 , X3 y X4 : 0.846
4. Variables restantes: 6 de 15 originales
5. Factores VIF de variables finales: X3 : 4.87 , X4 : 3.56 , X5 : 2.26 , X6 : 1.06 , X7 : 1.03 , X15 : 1.07
Este proceso de filtrado redujo el conjunto original de 15 variables a 6 variables, eliminando efectivamente las variables con problemas de variabilidad y correlación identificados.

El proceso de filtrado se implementa secuencialmente: (1) eliminar variables constantes o con varianza cercana a cero, (2) eliminar variables con correlación muy baja con la variable respuesta, (3) identificar grupos de variables multicolineales y retener solo la más relevante de cada grupo, y (4) calcular VIF y eliminar variables con valores muy altos. Este filtrado inicial típicamente reduce el conjunto de variables candidatas, facilitando significativamente los pasos posteriores de selección.

Es importante considerar que este filtrado no es definitivo, ya que variables eliminadas en esta etapa pueden ser importantes en combinaciones específicas. Además, está basado en relaciones lineales y puede omitir relaciones no lineales importantes. Por tanto, requiere validación posterior del modelo resultante y los umbrales deben ajustarse según el dominio de aplicación específico.

¡Claro! El contenido que tienes es excelente y muy completo. Para hacerlo menos esquemático, lo he reescrito en un formato más narrativo, conectando las ideas en párrafos fluidos. La idea es transformar las listas y tablas en una explicación discursiva, como si lo estuvieras contando en una clase, lo que se adapta mejor al formato de un libro.

Aquí tienes la propuesta:

5.3 Criterios de Bondad de Ajuste

Una vez completado el filtrado preliminar de variables, nos enfrentamos a una de las tareas más importantes del modelado: seleccionar la combinación óptima de predictores. El objetivo es encontrar un equilibrio delicado. Un modelo con muy pocas variables puede ser demasiado simple y no capturar la relación real (**subajuste** o *underfitting*), mientras que un modelo con demasiadas variables puede ajustarse al ruido de la muestra y no generalizar bien a nuevos datos (**sobreajuste** u *overfitting*).

Para navegar este compromiso, utilizamos criterios de información que cuantifican la calidad de un modelo, equilibrando su capacidad explicativa con su complejidad. Estos nos permiten

comparar modelos con diferente número de predictores de forma rigurosa y objetiva. Los tres criterios más influyentes en la estadística clásica son el Criterio de Información de Akaike (AIC), el Criterio de Información Bayesiano (BIC) y el estadístico Cp de Mallows.

5.3.1 Criterio de Información de Akaike

El **Criterio de Información de Akaike (AIC)**, desarrollado por Hirotugu Akaike, es una métrica fundamentada en la teoría de la información (James et al. 2013). Su propósito es estimar la pérdida de información que ocurre cuando usamos un modelo para representar la realidad. El modelo que minimice esta pérdida de información será considerado el mejor.

La fórmula del AIC para un modelo de regresión lineal es:

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

En esta ecuación, n es el tamaño de la muestra, SSE es la Suma de Cuadrados del Error y p es el número de variables predictoras. La fórmula equilibra dos fuerzas opuestas:

1. **Bondad de ajuste**: El primer término, $n \ln(SSE/n)$, está directamente relacionado con la función de log-verosimilitud del modelo. Disminuye a medida que el modelo se ajusta mejor a los datos (es decir, a medida que el SSE se reduce).
2. **Penalización por complejidad**: El segundo término, $2(p+1)$, actúa como un castigo. Aumenta en 2 unidades por cada parámetro adicional que se incluye en el modelo (p pendientes + 1 intercepto).

En la práctica, calculamos el AIC para varios modelos candidatos y **seleccionamos aquel con el valor de AIC más bajo**. Este criterio es asintóticamente eficiente, lo que significa que, con muestras suficientemente grandes, tiende a seleccionar el modelo que minimiza el error de predicción esperado en nuevos datos.

5.3.2 Criterio de Información Bayesiano

El **Criterio de Información Bayesiano (BIC)**, propuesto por Gideon Schwarz, es un competidor directo del AIC, pero con fundamentos en la estadística bayesiana (Hastie et al. 2009). Mientras que el AIC busca el mejor modelo para la predicción, el BIC está diseñado para encontrar el **modelo más probable de ser el "verdadero"** generador de los datos.

Su fórmula es muy similar a la del AIC, pero la penalización por complejidad es diferente y más severa:

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

La diferencia clave reside en el término de penalización. En lugar de $2(p+1)$, el BIC utiliza $(p+1)\ln(n)$. Dado que el logaritmo natural de n es mayor que 2 para cualquier muestra con más de 7 observaciones ($e^2 \approx 7.4$), la penalización del BIC es casi siempre más fuerte que la del AIC. Esta penalización más estricta le confiere al BIC una tendencia hacia la **parsimonia**, favoreciendo modelos más simples. Una de sus propiedades teóricas más importantes es la **consistencia**: si el modelo verdadero se encuentra entre los candidatos, la probabilidad de que el BIC lo seleccione tiende a 1 a medida que el tamaño de la muestra crece.

5.3.3 Estadístico Cp de Mallows

A diferencia del AIC y el BIC, el **estadístico Cp de Mallows** no se basa en la teoría de la información ni en la estadística bayesiana, sino que aborda directamente el **error cuadrático medio de predicción** del modelo (James et al. 2013). Su objetivo es encontrar un modelo que tenga un bajo sesgo y una baja varianza.

La fórmula para el estadístico Cp es:

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

Aquí, SSE_p es la suma de cuadrados del error del modelo candidato con p variables, y MSE_{full} es el error cuadrático medio del **modelo completo** (el que incluye todas las variables predictoras disponibles), que se utiliza como una estimación insesgada de la varianza del error poblacional, σ^2.

La interpretación del Cp es particularmente intuitiva. Si un modelo está bien especificado (es decir, no incluye un sesgo significativo), se espera que su valor de C_p sea cercano al número de parámetros, $p+1$.

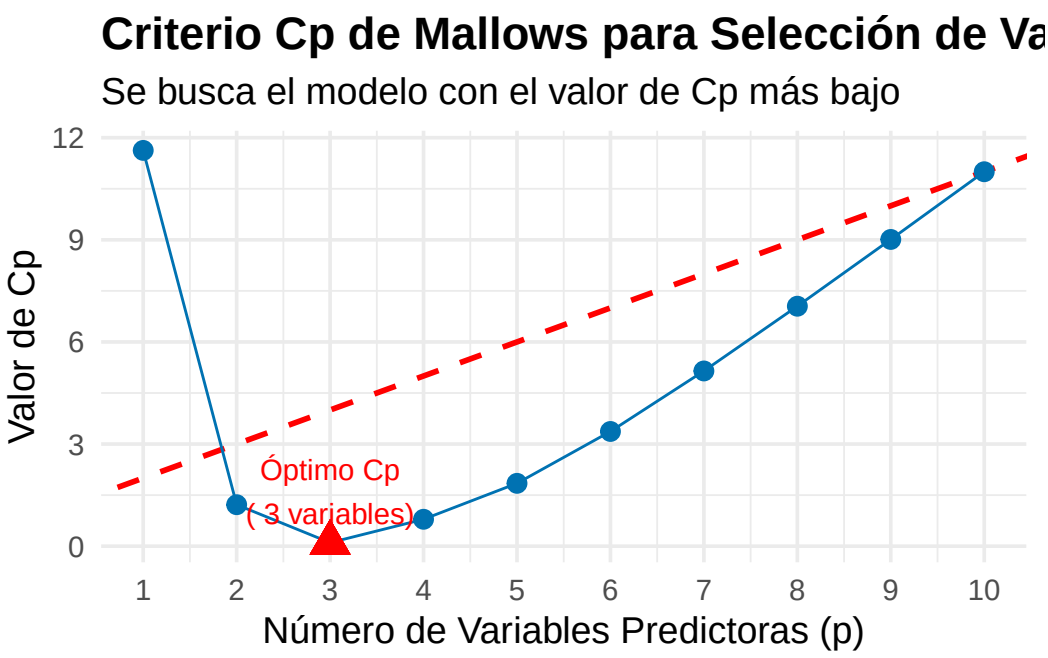
Por lo tanto, la estrategia de selección consiste en **elegir el modelo que tenga el valor de Cp más bajo**. Este modelo representa el mejor equilibrio entre el sesgo y la varianza. Generalmente, observaremos dos cosas en un gráfico de Cp vs. p:

- Los modelos con pocas variables y C_p **muy por encima de la línea** $p+1$ sufren de un sesgo elevado (subajuste).
- El modelo con el C_p **más bajo** es el preferido. Normalmente, este valor mínimo también estará cerca de la línea $p+1$, confirmando su buen ajuste.

> Visualización del Cp de Mallows
>
> Una de las aplicaciones más útiles del Cp de Mallows es su visualización gráfica para identificar el modelo óptimo. El siguiente ejemplo muestra cómo crear un gráfico de Cp que facilita la interpretación y selección del mejor modelo.

```r
# Cargar librerías necesarias
suppressPackageStartupMessages({
  library(leaps)     # Para best subset selection
  library(ggplot2)   # Para gráficos
  library(dplyr)     # Para manipulación de datos
})

# Usar el dataset mtcars para el ejemplo
data(mtcars)

# 1. Realizar best subset selection
best_subset <- regsubsets(mpg ~ ., data = mtcars, nvmax = 10)
subset_summary <- summary(best_subset)

# 2. Extraer información relevante para el gráfico
plot_data <- data.frame(
  n_variables = 1:length(subset_summary$cp),
  Cp = subset_summary$cp
)

# 3. Identificar el mejor modelo según el criterio Cp
#    La regla es simple: escoger el modelo con el menor valor de Cp.
mejor_cp_idx <- which.min(subset_summary$cp)
mejor_cp_valor <- subset_summary$cp[mejor_cp_idx]

# 4. Crear el gráfico del Cp de Mallows
ggplot(plot_data, aes(x = n_variables, y = Cp)) +
  # Línea de referencia ideal (Cp = p+1)
  geom_abline(intercept = 1, slope = 1, color = "red", linetype = "dashed", linewidth = 1)
  # Puntos y línea de los valores Cp de los modelos
  geom_point(color = "#0072B2", size = 3) +
  geom_line(color = "#0072B2") +
  # Resaltar el mejor punto (el que tiene el Cp mínimo)
  geom_point(aes(x = mejor_cp_idx, y = mejor_cp_valor), color = "red", size = 5, shape = 17
  # Etiqueta para el mejor punto
  annotate("text", x = mejor_cp_idx, y = mejor_cp_valor + 1.5,
           label = paste("Óptimo Cp\n(", mejor_cp_idx, "variables)"), color = "red", size =
  labs(
    title = "Criterio Cp de Mallows para Selección de Variables (mtcars)",
    subtitle = "Se busca el modelo con el valor de Cp más bajo",
    x = "Número de Variables Predictoras (p)",
    y = "Valor de Cp"
  ) +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(face = "bold")) +
  scale_x_continuous(breaks = 1:10)
```

```
Warning in geom_point(aes(x = mejor_cp_idx, y = mejor_cp_valor), color = "red", : All aesth
i Please consider using `annotate()` or provide this layer with data containing
  a single row.
```




```
# 5. Mostrar las variables del modelo seleccionado
variables_mejor_modelo_cp <- names(which(subset_summary$which[mejor_cp_idx, -1]))
```

El gráfico resultante nos permite diagnosticar visualmente la calidad de los modelos candidatos. La línea discontinua roja representa la referencia para un modelo sin sesgo ($C_p = p + 1$).
La estrategia de selección es clara: **identificar el modelo que minimice el estadístico Cp**. Este punto representa el mejor equilibrio teórico entre el sesgo del modelo (subajuste) y su varianza (sobreajuste). La línea roja nos ayuda a confirmar visualmente que el modelo elegido, además de ser el de menor Cp, tiene un sesgo bajo.
Como se observa, el estadístico Cp disminuye drásticamente al pasar de uno a dos predictores. El modelo con **3 variables** es el seleccionado como óptimo porque alcanza el valor de Cp más bajo de todos los candidatos, con un valor de **0.1**.
El análisis sugiere que el modelo más parsimonioso y con el mejor rendimiento predictivo se compone de las siguientes variables: **wt, qsec, am**. Añadir más predictores más allá de este punto óptimo no mejora el modelo; de hecho, el valor de Cp comienza a aumentar, lo que indica que estamos añadiendo una complejidad innecesaria y empezando a sobreajustar los datos.

5.3.4 ¿Cuándo Usar Cada Criterio?

La existencia de varios criterios plantea una pregunta natural: ¿cuál debemos usar? La respuesta depende en gran medida del objetivo final de nuestro análisis.

Si el **objetivo principal es la predicción**, el **AIC** suele ser la opción preferida. Su diseño para minimizar el error de predicción lo hace ideal en contextos de pronóstico, donde el rendimiento en datos nuevos es lo más importante. Su penalización más moderada permite incluir variables que, aunque no sean "verdaderas" en un sentido causal, ayudan a mejorar la precisión de las predicciones.

Por otro lado, si el **objetivo es la explicación o la inferencia** —es decir, identificar el modelo más parsimonioso que probablemente representa el verdadero proceso generador de los datos—, el **BIC** es la elección más sólida. Su penalización más fuerte protege de forma más robusta contra el sobreajuste y, en muestras grandes, su propiedad de consistencia le da una base teórica más fuerte para la selección del "modelo verdadero".

El **Cp de Mallows** es especialmente valioso en un contexto más exploratorio, cuando queremos entender explícitamente el compromiso entre el sesgo y la varianza. Al graficar C_p frente a $p + 1$ para diferentes subconjuntos de modelos, podemos visualizar claramente el punto en el que añadir más variables deja de reducir el sesgo y solo empieza a inflar la varianza, ofreciendo una visión muy clara del "codo" de complejidad óptima.

Es común que estos criterios no coincidan en su selección. Cuando esto ocurre, no debe verse como un fracaso, sino como una indicación de que no existe un único modelo "mejor" de forma inequívoca. En tales casos, el juicio del analista es clave, y se pueden usar herramientas adicionales como la **validación cruzada** (*cross-validation*) para comparar el rendimiento predictivo de los modelos finalistas y tomar una decisión informada.

> **!** El Principio de Parsimonia en la Selección de Modelos
>
> El **principio de parsimonia**, también conocido como la "navaja de Occam", es un concepto fundamental que subyace a todos los criterios de bondad de ajuste. Este principio establece que, entre modelos que explican igualmente bien un fenómeno, **se debe preferir el más simple**.

5.4 Métodos de selección exhaustiva

Los métodos de selección exhaustiva son un enfoque fundamental en la búsqueda de un subconjunto óptimo de variables predictoras en modelos de regresión. Este enfoque evalúa de manera sistemática diferentes combinaciones de variables para identificar cuál de ellas proporciona el mejor ajuste al modelo en función de un criterio predefinido, como el coeficiente de determinación ajustado (R^2 ajustado) o criterios de información como AIC o BIC.

A diferencia de los métodos automáticos, los métodos de selección exhaustiva no dependen de un proceso iterativo de adición o eliminación de variables. En cambio, buscan exhaustivamente (o mediante aproximaciones computacionalmente más eficientes) entre todas las posibles combinaciones de variables, lo que garantiza un análisis completo de las interacciones y relevancias potenciales.

El método más conocido dentro de este enfoque es la **selección del mejor subconjunto (Best Subset Selection)**, que evalúa todos los subconjuntos posibles de variables y selecciona el mejor para cada tamaño específico. Es el enfoque más completo pero también el más exigente computacionalmente. Para un conjunto de p variables predictoras, este método construye todos los modelos posibles que incluyen k variables, donde $k = 1, 2, ..., p$, seleccionando el mejor modelo de cada tamaño según el criterio elegido.

Ejemplo de selección exhaustiva

```
# Ejemplo de Best Subset Selection maximizando R² ajustado
suppressPackageStartupMessages(library(leaps))

# Usando el dataset mtcars
data(mtcars)

# Realizar best subset selection
best_subset <- regsubsets(mpg ~ ., data = mtcars, nvmax = 10)

# Obtener estadísticas del mejor modelo según R² ajustado
subset_summary <- summary(best_subset)
mejor_modelo_idx <- which.max(subset_summary$adjr2)
mejor_r2_adj <- max(subset_summary$adjr2)
total_variables <- ncol(mtcars) - 1  # Excluir variable respuesta
variables_seleccionadas <- mejor_modelo_idx
```

El método de selección exhaustiva aplicado al conjunto de datos `mtcars` identifica el modelo óptimo que maximiza el R^2 ajustado. Este modelo alcanza un **R^2 ajustado de 0.8375**, utilizando **5 variables** del total de **10 variables** predictoras disponibles. Esta selección representa un equilibrio óptimo entre la capacidad explicativa del modelo y la penalización por complejidad, demostrando cómo la evaluación exhaustiva puede identificar el subconjunto de variables que mejor explica la variabilidad en el consumo de combustible.

Esta aproximación presenta importantes **ventajas**: garantiza encontrar el mejor subconjunto según el criterio elegido (optimalidad garantizada), examina todas las posibles combinaciones de variables ofreciendo una evaluación completa, y proporciona un estándar sólido para comparar otros métodos de selección. Sin embargo, también tiene **limitaciones** significativas:

la complejidad computacional crece exponencialmente ya que con p variables se generan 2^p modelos posibles, lo que hace que sea impracticable para p grande (típicamente $p > 15 - 20$). Además, sin una validación cruzada adecuada, puede seleccionar modelos sobreajustados.

Estos métodos son especialmente útiles cuando el número de predictores no es demasiado grande, ya que el esfuerzo computacional crece exponencialmente con el número de variables. Aunque el costo computacional puede ser elevado en datasets amplios, los métodos de selección exhaustiva proporcionan una referencia sólida y transparente para evaluar qué variables son fundamentales en el modelo, siendo particularmente valiosos en estudios donde la interpretabilidad y la certeza sobre la selección de variables son prioritarias.

¡Perfecto! El texto que tienes es una excelente introducción. Ahora vamos a expandir cada uno de esos puntos para darles la profundidad teórica y práctica que necesitan en el libro, explicando el algoritmo de cada método, sus criterios de decisión y sus ventajas y limitaciones.

Aquí tienes una propuesta para desarrollar esa sección.

5.5 Métodos automáticos paso a paso

Los métodos automáticos de selección de variables, a menudo llamados métodos secuenciales o por pasos (*stepwise*), son algoritmos diseñados para explorar el espacio de posibles modelos de una manera computacionalmente eficiente. A diferencia del método de mejores subconjuntos (*best subset selection*), que evalúa todos los modelos posibles, estos enfoques siguen un camino restringido, añadiendo o quitando predictores de uno en uno.

El principio clave es construir un modelo de forma iterativa, tomando en cada paso una decisión "localmente óptima" basada en un criterio estadístico. Los criterios más comunes son el p-valor de un predictor, o el cambio que este produce en un indicador global como el **AIC**, el **BIC** o el R^2 **ajustado**.

5.5.1 Selección progresiva (Forward Selection)

Esta estrategia es la más intuitiva: parte de la nada y construye el modelo pieza por pieza, añadiendo en cada paso el predictor que aporta la mayor mejora.

El Algoritmo

1. **Inicio**: Se comienza con el **modelo nulo**, que solo contiene el intercepto ($Y \sim 1$).
2. **Primer Paso**: Se ajustan p modelos de regresión simple, uno para cada una de las p variables predictoras disponibles. Se elige la variable que mejor explica la respuesta (la que tiene el p-valor más bajo en su test t, o la que produce el AIC/BIC más bajo). Esta variable se convierte en el primer predictor del modelo.

3. **Pasos Siguientes**: Se ajustan $p-1$ nuevos modelos, cada uno de los cuales contiene la(s) variable(s) ya seleccionada(s) más una de las variables restantes. De nuevo, se selecciona y se añade la variable que produce la mayor mejora en el criterio elegido.
4. **Finalización**: El proceso se repite y se detiene cuando ninguna de las variables restantes mejora el modelo de forma significativa al ser añadida (por ejemplo, ninguna tiene un p-valor por debajo de un umbral predefinido, o el AIC/BIC del modelo deja de disminuir).

Ventajas y Limitaciones

- **Ventaja**: Es computacionalmente muy eficiente. Puede aplicarse en situaciones con un número de predictores muy grande, incluso cuando hay más predictores que observaciones ($p > n$).
- **Limitación**: Su principal debilidad es su **"miopía"**. Una variable seleccionada en una etapa temprana se queda en el modelo para siempre. Sin embargo, es posible que esa variable se vuelva redundante una vez que se añadan otros predictores. El método *forward* no puede rectificar decisiones pasadas, por lo que no garantiza encontrar el mejor modelo posible.

5.5.2 Eliminación regresiva (Backward Elimination)

Esta estrategia adopta el enfoque opuesto: empieza con todo y va eliminando lo que no es útil, como un escultor que retira el mármol sobrante.

El Algoritmo

1. **Inicio**: Se comienza con el **modelo completo**, que incluye todas las p variables predictoras disponibles ($Y \sim X_1 + X_2 + \cdots + X_p$).
2. **Primer Paso**: Se ajusta el modelo completo y se examina la significancia de cada predictor. Se identifica la variable **menos significativa**, es decir, aquella con el p-valor más alto en su test t (o la que, al ser eliminada, produce la menor disminución en la calidad del modelo según AIC/BIC).
3. **Pasos Siguientes**: Si el p-valor de esa variable supera un umbral de permanencia (p. ej., $\alpha_{out} = 0.10$), se elimina del modelo. A continuación, se vuelve a ajustar el modelo con las $p-1$ variables restantes.
4. **Finalización**: El proceso de identificar y eliminar la variable menos significativa se repite hasta que todas las variables que quedan en el modelo son estadísticamente significativas (es decir, todas tienen un p-valor por debajo del umbral de permanencia).

Ventajas y Limitaciones

- **Ventaja**: Generalmente se considera superior al método *forward* porque empieza evaluando el efecto de cada variable en presencia de todas las demás. Esto proporciona un contexto inicial más completo.

- **Limitación**: No se puede utilizar si el número de predictores es mayor que el número de observaciones ($p > n$), ya que es imposible ajustar el modelo completo inicial. Además, al igual que el método *forward*, una vez que una variable es eliminada, no puede volver a entrar, lo que podría llevar a eliminar por error una variable que es importante en combinación con un subconjunto más pequeño de predictores.

5.5.3 Selección paso a paso (Stepwise Regression)

Este método es un híbrido que intenta combinar lo mejor de las dos estrategias anteriores, permitiendo un proceso de "prueba y error" más flexible.

El Algoritmo

La selección *stepwise* es esencialmente una selección *forward* con un añadido crucial: en cada paso, después de añadir una nueva variable, se realiza una **verificación hacia atrás** para comprobar si alguna de las variables que ya estaban en el modelo se ha vuelto redundante.

1. **Paso Adelante (Forward)**: Al igual que en la selección progresiva, se añade la variable que más mejora el modelo.
2. **Paso Atrás (Backward)**: Después de añadir esa variable, se examinan **todas las variables ya incluidas** en el modelo. Si alguna de ellas ha perdido su significancia (su p-valor ha aumentado por encima de un umbral de eliminación), se elimina.
3. **Repetición**: El proceso continúa, alternando pasos hacia adelante y hacia atrás, hasta que se alcanza un punto de equilibrio en el que ninguna variable puede ser añadida ni eliminada según los umbrales establecidos.

Ventajas y Limitaciones

- **Ventaja**: Es más robusto que los métodos *forward* o *backward* puros, ya que puede corregir decisiones anteriores. Una variable que fue importante al principio puede ser eliminada más tarde si otra la hace redundante.
- **Limitación**: A pesar de su flexibilidad, sigue siendo un algoritmo "codicioso" (*greedy*) que no explora todo el espacio de modelos. Por tanto, tampoco garantiza encontrar el mejor modelo global.

> ⚠ Advertencia sobre los métodos automáticos
>
> Aunque estos métodos son herramientas útiles para un primer cribado de variables, deben usarse con extrema cautela. Su naturaleza automática puede llevar a conclusiones erróneas si no se supervisan con criterio.
>
> 1. **No garantizan el mejor modelo**: Al seguir un camino fijo, pueden pasar por alto el subconjunto de variables verdaderamente óptimo.
> 2. **Invalidez de los p-valores**: Los p-valores, errores estándar e intervalos de confianza

del modelo final están **sesgados** y son excesivamente optimistas. El proceso de selección ha "elegido a los ganadores" de antemano, y la teoría de la inferencia estándar no se aplica a un modelo que ha sido seleccionado de esta manera.
3. **Inestabilidad**: Los resultados pueden ser muy sensibles a pequeñas variaciones en los datos. Añadir o quitar unas pocas observaciones puede cambiar drásticamente el modelo seleccionado.

Por estas razones, los métodos automáticos deben considerarse como **herramientas exploratorias** para generar modelos candidatos, no como un procedimiento definitivo. La selección final siempre debe estar guiada por el conocimiento del dominio, la teoría subyacente y un diagnóstico riguroso.

5.6 Métodos basados en regularización

En los modelos de regresión, especialmente cuando se trabaja con un gran número de variables predictoras o con datos multicolineales, los métodos tradicionales de selección de variables pueden resultar ineficaces o inestables. En estos casos, los métodos basados en regularización surgen como una alternativa poderosa que no solo selecciona variables, sino que también mejora la estabilidad y la precisión del modelo.

La regularización consiste en introducir una penalización en la función de ajuste del modelo, lo que tiene dos efectos principales: controlar el sobreajuste al reducir la complejidad del modelo y forzar la selección de un subconjunto más parsimonioso de predictores. Estas penalizaciones ajustan los coeficientes de las variables predictoras, favoreciendo soluciones más simples y robustas (James et al. 2013).

Entre los métodos de regularización más destacados se encuentran:

- **Ridge Regression:** Aplica una penalización proporcional al cuadrado de los coeficientes, lo que permite manejar problemas de multicolinealidad pero no conduce a la eliminación completa de variables.
- **Lasso (Least Absolute Shrinkage and Selection Operator):** Introduce una penalización basada en el valor absoluto de los coeficientes, lo que no solo reduce su magnitud, sino que también puede anularlos completamente, realizando una selección automática de variables.
- **Elastic Net:** Combina las penalizaciones de Ridge y Lasso, ofreciendo mayor flexibilidad en situaciones donde hay una gran correlación entre los predictores.

Estos métodos son especialmente útiles en problemas donde el número de variables predictoras excede el número de observaciones, o cuando se desea un modelo más interpretable. En esta sección, exploraremos en detalle los fundamentos teóricos, la implementación práctica y las

aplicaciones de cada uno de estos métodos, destacando sus ventajas en escenarios complejos y desafiantes.

5.6.1 Ridge regression

La regresión Ridge introduce una penalización en la estimación de los coeficientes de regresión, lo que ayuda a reducir la varianza del modelo y mejora su capacidad predictiva en presencia de datos altamente correlacionados o con muchas variables (Marquardt and Snee 1975). El modelo de regresión Ridge es una extensión de la regresión lineal estándar. Con datos observados, escribimos:

$$\mathbf{y} = \mathbf{X}\,\beta + \varepsilon$$

donde:

- $\mathbf{y}$ es el vector de respuesta observado de dimensión $n \times 1$.
- $\mathbf{X}$ es la matriz de diseño observada de dimensión $n \times (p+1)$ (la primera columna suele ser de unos para el intercepto).
- $\beta = (\beta_0, \beta_1, ..., \beta_p)^T$ es el vector de coeficientes.
- ε es el vector de errores.

En mínimos cuadrados ordinarios (OLS), los coeficientes se estiman minimizando la suma de los errores al cuadrado:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \|\mathbf{y} - \mathbf{X}\,\beta\|^2.$$

Sin embargo, cuando hay multicolinealidad, la matriz $\mathbf{X}^T\mathbf{X}$ puede ser casi singular, generando coeficientes inestables. Para evitar esto, la regresión Ridge añade un **término de penalización** λ, de la siguiente manera (sin penalizar el intercepto β_0):

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

Este término adicional, es un **término de penalización** ($L_2 = \sum \beta_j^2$) impone una restricción sobre los coeficientes, evitando que tomen valores excesivamente grandes. La estimación de β en Ridge se obtiene resolviendo:

$$\widehat{\beta}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\,\mathbf{P})^{-1}\mathbf{X}^T\mathbf{y}.$$

donde $\mathbf{P}$ es diagonal con $P_{11} = 0$ (no penalizamos el intercepto) y $P_{jj} = 1$ para $j = 2, \ldots, p+1$, y $\lambda \geq 0$ controla la cantidad de penalización aplicada. (Cuando no hay intercepto o se reparametriza, a menudo se escribe con I para simplificar.)

Interpretación del parámetro λ

- Si $\lambda = 0$, el modelo Ridge es equivalente a la regresión lineal tradicional (OLS).
- A medida que λ aumenta, los coeficientes β_j (pendientes) se reducen en magnitud, lo que ayuda a controlar la varianza del modelo y a prevenir el sobreajuste.
- Si λ es demasiado grande, los coeficientes se acercan a cero y el modelo puede perder interpretabilidad.

La elección óptima de λ se determina generalmente mediante **validación cruzada**.

Aviso

Los detalles de la validación cruzada son tratados en la asignatura de Minería de Datos.

Propiedades Clave

- **Manejo de la multicolinealidad:** La regularización reduce la sensibilidad del modelo cuando los predictores están altamente correlacionados.
- **Menor varianza en las predicciones:** El modelo Ridge tiende a ser más estable en comparación con OLS, lo que mejora la capacidad de generalización en conjuntos de datos nuevos.
- **No realiza selección de variables:** A diferencia de Lasso, Ridge **no anula coeficientes**, sino que reduce su magnitud. Esto es útil cuando se sospecha que todas las variables tienen algún grado de importancia en el modelo.

Ejemplo

```r
# Cargar librerías
suppressPackageStartupMessages(library(glmnet))

# Datos simulados
set.seed(123)
X <- matrix(rnorm(100 * 10), 100, 10)  # 100 observaciones, 10 predictores
Y <- X %*% rnorm(10) + rnorm(100)  # Variable de respuesta con ruido

# Ajustar modelo Ridge
modelo_ridge <- glmnet(X, Y, alpha = 0)  # alpha = 0 indica regresión Ridge

# Seleccionar lambda óptimo con validación cruzada
cv_ridge <- cv.glmnet(X, Y, alpha = 0)
lambda_optimo <- cv_ridge$lambda.min  # Mejor valor de lambda

print(lambda_optimo)
```

```
[1] 0.2583753
```

```r
# Ajustar modelo final con lambda óptimo
modelo_ridge_final <- glmnet(X, Y, alpha = 0, lambda = lambda_optimo)

modelo_ridge_final
```

```

Call:  glmnet(x = X, y = Y, alpha = 0, lambda = lambda_optimo)

  Df  %Dev Lambda
1 10 93.55 0.2584
```

```r
# Comparación modelo clásico

modelo_lm <- lm(Y~X)

# Mostrar coeficientes
output=cbind(round(coef(modelo_ridge_final),3),
             round(coef(modelo_lm),3))

colnames(output)=c("RIDGE","OLS")

output
```

```
11 x 2 sparse Matrix of class "dgCMatrix"
              RIDGE    OLS
(Intercept)  0.118  0.132
V1          -0.874 -0.995
V2          -1.019 -1.131
V3           0.040  0.039
V4           0.002  0.001
V5          -2.500 -2.703
V6           1.001  1.104
V7           0.247  0.274
V8           2.125  2.244
V9           0.635  0.658
V10         -0.390 -0.427
```

La regresión Ridge es una técnica poderosa para mejorar la estabilidad de los modelos de regresión en presencia de multicolinealidad. A diferencia de OLS, que puede generar coeficientes inestables, Ridge introduce una penalización que reduce la magnitud de los coeficientes, evitando valores extremos. Aunque Ridge no realiza selección de variables, su capacidad para reducir la varianza y mejorar la capacidad predictiva lo convierte en una herramienta esencial en el análisis de datos modernos.

En la siguiente sección, exploraremos la **regresión Lasso**, que extiende este concepto permitiendo la eliminación de variables irrelevantes del modelo.

5.6.2 Regresión Lasso

Cuando se tiene un conjunto de predictores con posibles redundancias o ruido, Lasso permite identificar cuáles son las variables más relevantes para el modelo, lo que facilita la interpretación y reduce la complejidad del análisis.

Al igual que en Ridge, el modelo de regresión Lasso se define sobre datos observados mediante la minimización (Ranstam and Cook 2018):

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

donde el **término de penalización** ($L_1 = \sum |\beta_j|$) no penaliza el intercepto β_0 y hace que algunos coeficientes de pendiente se reduzcan exactamente a **cero**, eliminando variables del modelo.

La diferencia clave con **Ridge Regressión**, visto anteriormente, es que Ridge reduce la magnitud de los coeficientes pero no los anula, mientras que **Lasso puede eliminar variables por completo**.

Interpretación del parámetro λ

- Si $\lambda = 0$, el modelo es equivalente a la regresión lineal tradicional (OLS).
- A medida que λ aumenta, más coeficientes de pendiente se reducen a cero, lo que equivale a realizar **selección de variables**.
- Si λ es demasiado grande, se eliminan demasiadas variables, lo que puede resultar en un modelo subóptimo.

Al igual que en el método *Ridge*, la selección óptima de λ se realiza generalmente mediante **validación cruzada**.

Propiedades Clave

- **Selección de variables automática:** Lasso no solo regulariza, sino que también selecciona las variables más importantes eliminando aquellas menos relevantes.
- **Manejo de la multicolinealidad:** Puede mejorar la interpretación del modelo cuando hay muchas variables correlacionadas.
- **Simplicidad y interpretabilidad:** Un modelo con menos variables es más fácil de interpretar y aplicar en la práctica.
- **Reduce el sobreajuste:** La penalización L_1 evita que el modelo se ajuste demasiado a los datos de entrenamiento, mejorando su capacidad predictiva en datos nuevos.

Ejemplo

```r
# Ajustar modelo Lasso
modelo_lasso <- glmnet(X, Y, alpha = 1)  # alpha = 1 indica regresión Lasso

# Seleccionar lambda óptimo con validación cruzada
cv_lasso <- cv.glmnet(X, Y, alpha = 1)
lambda_optimo <- cv_lasso$lambda.min  # Mejor valor de lambda

print(lambda_optimo)
```

```
[1] 0.03260326
```

```r
# Ajustar modelo final con lambda óptimo
modelo_lasso_final <- glmnet(X, Y, alpha = 1, lambda = lambda_optimo)


# Mostrar coeficientes
output=cbind(round(coef(modelo_lasso_final),3),output)

colnames(output)=c("LASSO","RIDGE","OLS")

output
```

```
11 x 3 sparse Matrix of class "dgCMatrix"
             LASSO  RIDGE    OLS
(Intercept)  0.131  0.118  0.132
V1          -0.950 -0.874 -0.995
V2          -1.078 -1.019 -1.131
V3           0.006  0.040  0.039
V4           .      0.002  0.001
V5          -2.652 -2.500 -2.703
V6           1.058  1.001  1.104
V7           0.235  0.247  0.274
V8           2.213  2.125  2.244
V9           0.629  0.635  0.658
V10         -0.392 -0.390 -0.427
```

Consideraciones Importantes

La regresión Lasso es una poderosa técnica de regularización que no solo mejora la estabilidad del modelo en presencia de muchas variables predictoras, sino que también realiza una selección

automática de las más relevantes. Su capacidad para reducir coeficientes a cero la convierte en una herramienta esencial en el análisis de datos de alta dimensión.

- **Lasso puede eliminar demasiadas variables si** λ es demasiado grande, lo que puede llevar a la pérdida de información importante.
- **No maneja bien grupos de predictores altamente correlacionados**, ya que selecciona solo uno de ellos y elimina los demás.
- **Elastic Net**, que combina Ridge y Lasso, puede ser una mejor opción cuando hay **multicolinealidad fuerte** en los datos.

En la siguiente sección, exploraremos **Elastic Net**, una técnica híbrida que combina las ventajas de Ridge y Lasso para mejorar la selección de variables en presencia de predictores altamente correlacionados.

5.6.3 Elastic Net

La regresión **Elastic Net** es una técnica de regularización que combina las propiedades de **Ridge** y **Lasso**, abordando algunas de sus limitaciones individuales (Zou and Hastie 2005). Mientras que Ridge es útil para manejar la multicolinealidad sin eliminar variables y Lasso selecciona un subconjunto de predictores, Elastic Net equilibra ambos enfoques permitiendo la selección de variables en presencia de alta correlación entre los predictores.

Este método es particularmente efectivo cuando el número de predictores es grande y existe **multicolinealidad**, ya que permite controlar simultáneamente la **reducción de la magnitud de los coeficientes** y la **eliminación de variables irrelevantes**.

Elastic Net introduce una penalización que combina los términos de Ridge (L_2) y Lasso (L_1), sobre datos observados:

$$SSE_{\text{Elastic Net}} = \|\mathbf{y} - \mathbf{X}\,\beta\|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$$

donde:

- λ_1 (asociado a Lasso) controla la cantidad de coeficientes que se reducen a **cero**.
- λ_2 (asociado a Ridge) controla la **reducción de magnitud** de los coeficientes sin anularlos.
- α es un parámetro adicional que pondera la combinación entre Lasso y Ridge, con:
  - $\alpha = 1 \rightarrow$ Elastic Net se comporta como Lasso.
  - $\alpha = 0 \rightarrow$ Elastic Net se comporta como Ridge.
  - $0 < \alpha < 1 \rightarrow$ Elastic Net combina ambos métodos.

La estimación de los coeficientes en Elastic Net se obtiene resolviendo (habitualmente sin penalizar el intercepto):

$$\widehat{\beta}_{\text{EN}} = \arg\min_{\beta} \left(\|\mathbf{y} - \mathbf{X}\,\beta\|^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right) \right)$$

Propiedades Clave

- **Manejo de la Multicolinealidad:** A diferencia de Lasso, que selecciona solo una de las variables correlacionadas y elimina las demás, Elastic Net distribuye la penalización entre todas las variables correlacionadas, evitando una selección arbitraria.
- **Selección de variables más estable:** La combinación de Lasso y Ridge permite una selección más robusta, manteniendo información relevante del modelo sin eliminar predictores clave.
- **Mejora del rendimiento predictivo:** Al utilizar validación cruzada para seleccionar los hiperparámetros λ_1, λ_2 y α, se optimiza la capacidad del modelo para generalizar a nuevos datos.

Ejemplo

```r
# Ajustar modelo Elastic Net
modelo_elastic_net <- glmnet(X, Y, alpha = 0.5)  # Alpha = 0.5 (50% Ridge, 50% Lasso)

# Seleccionar lambda óptimo con validación cruzada
cv_elastic_net <- cv.glmnet(X, Y, alpha = 0.5)
lambda_optimo <- cv_elastic_net$lambda.min  # Mejor valor de lambda

print(lambda_optimo)
```

```
[1] 0.0213522
```

```r
# Ajustar modelo final con lambda óptimo
modelo_elastic_final <- glmnet(X, Y, alpha = 0.5, lambda = lambda_optimo)

# Mostrar coeficientes
output=cbind(round(coef(modelo_elastic_final),3),output)

colnames(output)=c("ELASTIC","LASSO","RIDGE","OLS")

output
```

```
11 x 4 sparse Matrix of class "dgCMatrix"
            ELASTIC  LASSO  RIDGE    OLS
(Intercept)   0.131  0.131  0.118  0.132
V1           -0.975 -0.950 -0.874 -0.995
V2           -1.108 -1.078 -1.019 -1.131
V3            0.028  0.006  0.040  0.039
V4            .      .      0.002  0.001
V5           -2.677 -2.652 -2.500 -2.703
V6            1.084  1.058  1.001  1.104
V7            0.260  0.235  0.247  0.274
V8            2.229  2.213  2.125  2.244
V9            0.647  0.629  0.635  0.658
V10          -0.414 -0.392 -0.390 -0.427
```

Para determinar el mejor valor de α, se usa **validación cruzada** probando distintos valores entre 0 y 1. Algunas estrategias comunes incluyen:

- **Si hay muchas variables irrelevantes**, se recomienda α cercano a 1 (Lasso).
- **Si hay fuerte multicolinealidad**, se recomienda α cercano a 0 (Ridge).
- **Si se desea un balance entre selección y estabilidad**, se suele usar $\alpha = 0.5$.

La regresión Elastic Net combina lo mejor de Ridge y Lasso, ofreciendo un método de regularización robusto para modelos con muchas variables predictoras y posible multicolinealidad. Su capacidad para seleccionar variables sin eliminar información clave lo convierte en una opción ideal para modelos complejos y de alta dimensionalidad.

5.6.4 Comparación de los métodos de Regularización

Método	Penalización	Efecto sobre los coeficientes
OLS	Ninguna	Sin restricción, puede haber multicolinealidad
Ridge	L_2	Reduce la magnitud de los coeficientes, pero no los anula
Lasso	L_1	Puede anular coeficientes, permitiendo selección de variables
Elastic Net	$L_1 + L_2$	Combinación de Ridge y Lasso

Lasso es especialmente útil cuando se sospecha que muchas variables son irrelevantes, mientras que Ridge es preferido cuando se espera que todas las variables aporten información al modelo.

Elastic Net es ideal cuando hay **muchas variables correlacionadas** y se desea un modelo **estable y parsimonioso**.

- Elastic Net mejora la estabilidad del modelo en comparación con Lasso, especialmente cuando hay variables predictoras altamente correlacionadas.
- Es más flexible que Ridge y Lasso individualmente, permitiendo un ajuste más fino a distintos tipos de problemas.
- Requiere la selección de hiperparámetros (λ y α), por lo que debe usarse validación cruzada para encontrar la combinación óptima.

5.7 Validación del Modelo

Hemos ajustado un modelo, interpretado sus coeficientes y evaluado su significancia estadística. Pero, ¿cómo podemos estar seguros de que funcionará bien en el futuro, con datos que nunca ha visto? Esta es la pregunta fundamental que la **validación del modelo** busca responder.

Imagina que estás preparando un examen. Si solo memorizas las respuestas de los exámenes de años anteriores (tus **datos de entrenamiento**), puede que saques una nota perfecta en ellos. Sin embargo, cuando te enfrentes al examen real con preguntas nuevas (los **datos de prueba**), es probable que tu rendimiento sea decepcionante. Esto, en esencia, es el **sobreajuste** (*overfitting*): un modelo que se aprende los datos de entrenamiento "de memoria", incluyendo su ruido y peculiaridades, pero que pierde su capacidad de **generalizar** a nuevas observaciones.

La validación es el proceso de simular este "examen final" para obtener una estimación honesta del rendimiento predictivo de nuestro modelo en el mundo real (James et al. 2013). Se compone de dos elementos clave: las **estrategias de validación**, que nos dicen cómo simular el examen, y las **métricas de evaluación**, que nos dicen cómo calificarlo.

5.7.1 Estrategias de Validación

Para evaluar la capacidad de generalización, necesitamos probar el modelo en datos que no se usaron para entrenarlo. Las siguientes estrategias nos permiten hacer precisamente eso.

⚠ El primer paso no negociable: La partición inicial

Antes de escribir una sola línea de código para ajustar un modelo, seleccionar variables o ejecutar una validación cruzada, el procedimiento **siempre** debe comenzar con una única acción:
Dividir el conjunto de datos completo en dos partes y guardar una de ellas bajo llave.

1. **Datos de modelado (p. ej., 80% del total):** Este es el conjunto de datos con el que trabajarás. Lo usarás para todas tus tareas de construcción y evaluación de modelos: entrenar, comparar diferentes conjuntos de variables, y ejecutar estrategias como la validación cruzada.
2. **Conjunto de prueba Final (p. ej., 20% restante):** Este conjunto de datos debe ser guardado y **no ser utilizado bajo ninguna circunstancia** durante el proceso de modelado. Es tu "examen final sorpresa", tu única oportunidad de obtener una estimación verdaderamente honesta y no sesgada del rendimiento del modelo que has seleccionado como el campeón definitivo.

La validación cruzada y la división simple train/test que veremos a continuación son técnicas que se aplican **dentro** de los "datos de modelado".

5.7.1.1 El Conjunto de entrenamiento y test (Train/Test Split)

La estrategia más directa es tomar nuestros "Datos de Modelado" (el 80% inicial) y volver a dividirlos, creando un único "examen" para el proceso de construcción del modelo (Hastie et al. 2009):

1. **Conjunto de entrenamiento (*Training Set*)**: Usualmente, el 70-80% de los datos. El modelo se construye y se ajusta usando únicamente esta porción. Es aquí donde el modelo "aprende".

2. **Conjunto de test (*Test Set*)**: El 20-30% restante. Estos datos se mantienen "ocultos" durante el entrenamiento. Una vez que el modelo está finalizado, lo usamos para predecir la variable respuesta en este conjunto. La comparación entre las predicciones ($\hat{y}$) y los valores reales (y) nos da una medida no sesgada de su rendimiento.

Aunque es simple y computacionalmente barata, esta técnica tiene una debilidad importante: los resultados pueden depender mucho de la división aleatoria específica que se haya hecho. Si por mala suerte en el conjunto de prueba caen observaciones muy atípicas, nuestra evaluación del modelo será excesivamente pesimista. Si caen puntos muy fáciles de predecir, será demasiado optimista. Esta alta variabilidad es un problema, especialmente con muestras de datos pequeñas.

5.7.1.2 Validación cruzada (Cross-Validation)

La **validación cruzada** es la solución a la variabilidad de la división simple y se aplica, de nuevo, **sobre el conjunto total de "datos de modelado"**. En lugar de hacer un único "examen final", la validación cruzada promedia los resultados de múltiples mini-exámenes, proporcionando una estimación del error mucho más estable y fiable (James et al. 2013).

El método más común es la **validación cruzada de k-particiones (*k-fold cross-validation*)**. Su nombre describe el proceso: los datos se dividen en k particiones y se "cruzan" los roles de entrenamiento y validación.

El resultado final de este procedimiento es el **error de validación cruzada**, que se calcula promediando los errores obtenidos en cada una de las k particiones. Esto nos da una única métrica de rendimiento para el modelo.

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} \text{Métrica}_i$$

donde Métrica_i es la métrica de error (como RMSE o MAE) calculada en la i-ésima iteración. La elección de k suele ser 5 o 10, ya que se ha demostrado que estos valores ofrecen un buen equilibrio entre el sesgo y la varianza de la estimación del error.

> **i** Procedimiento de k-particiones
>
> 1. **División**: Dividir aleatoriamente los datos en k **particiones** de tamaño similar.
> 2. **Iteración**: Para cada partición $i = 1, 2, ..., k$:
>
>   - Usar la partición i como conjunto de test.
>   - Usar las restantes $k - 1$ particiones como conjunto de entrenamiento.
>   - Ajustar el modelo y calcular las métricas de desempeño en el conjunto de test.
>
> 3. **Promedio**: Calcular el promedio de las métricas a través de las k iteraciones.

Un caso extremo de este método es la **validación cruzada "dejando uno fuera" (LOOCV)**, donde k es igual al número de observaciones, n. En cada iteración, se entrena el modelo con n-1 datos y se prueba en el único punto restante. Aunque es computacionalmente muy costoso, en regresión lineal existe una afortunada fórmula matemática que nos permite calcular el error LOOCV con la misma rapidez que un solo ajuste:

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_{ii}} \right)^2$$

donde h_{ii} es el apalancamiento (*leverage*) de la i-ésima observación.

! Guía para seleccionar estrategia de validación

Usa Train/Test cuando:

- El dataset es grande (n > 1000)
- Los recursos computacionales son limitados

- Se requiere una evaluación rápida

Usa validación cruzada k-fold cuando:

- El dataset es de tamaño moderado (n < 1000)
- Se requiere una estimación más estable del desempeño
- Se dispone de recursos computacionales adecuados

Usa LOOCV cuando:

- El dataset es pequeño (n < 100)
- Se requiere la estimación menos sesgada posible
- El tiempo computacional no es una restricción crítica

5.7.2 Métricas de rendimiento

Una vez que usamos una estrategia de validación para generar predicciones sobre datos no vistos, necesitamos una "nota" para cuantificar qué tan buenos fueron esos pronósticos. Aquí es donde entran las métricas de error.

5.7.2.1 Raíz del error cuadrático medio

La métrica más utilizada es la **raíz del error cuadrático medio (RMSE)**. Es como una desviación típica de los residuos, y nos da una idea de la magnitud promedio de los errores de predicción.

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

La característica clave del RMSE es que, al elevar los errores al cuadrado, **penaliza de forma desproporcionada los errores grandes**. Un solo error de predicción de 10 unidades contribuye al RMSE mucho más que 10 errores de 1 unidad. Esto lo hace muy sensible a valores atípicos. Su gran ventaja es que se expresa en las mismas unidades que la variable respuesta, facilitando su interpretación.

5.7.2.2 Error absoluto medio

Una alternativa popular es el **error absoluto medio (MAE)**, que simplemente promedia el valor absoluto de los errores.

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

A diferencia del RMSE, el MAE no eleva los errores al cuadrado, por lo que **trata todos los errores de forma proporcional a su magnitud**. Un error de 10 unidades es simplemente el doble de malo que un error de 5. Esto hace que el MAE sea **más robusto frente a valores atípicos** y, para muchos, más fácil de interpretar como "el error promedio" que cometemos en nuestras predicciones.

En resumen, la validación nos obliga a confrontar nuestro modelo con la realidad de datos nuevos. Usando una estrategia robusta como la **validación cruzada** para calcular una métrica interpretable como el **RMSE** o el **MAE**, podemos obtener una estimación fiable de su rendimiento predictivo y construir modelos en los que realmente podamos confiar. Claro, aquí tienes ese contenido reorganizado y resumido dentro de un recuadro (`callout`), ideal para destacar esta idea clave en tu libro.

> **!** Interpretando el Error
>
> La comparación entre el error del modelo en los datos que ha visto (entrenamiento) y en datos que no ha visto (validación) es la herramienta de diagnóstico más importante para entender el ajuste del modelo.

La regla general es que el error de entrenamiento siempre será más bajo (más optimista) que el de test. La clave está en analizar la **diferencia** entre ambos.

Sobreajuste (Overfitting)

- **Síntoma**: Error de entrenamiento **bajo** + Error de test **mucho más alto**.
- **Diagnóstico**: El modelo ha "memorizado" el ruido de los datos de entrenamiento y no es capaz de generalizar a nuevos datos.
- **Solución**: Simplificar el modelo (usar menos variables, aplicar regularización como Ridge o Lasso).

Subajuste (Underfitting)

- **Síntoma**: Error de entrenamiento **alto** + Error de test **alto y similar**.
- **Diagnóstico**: El modelo es demasiado simple y no tiene la capacidad de capturar la estructura subyacente de los datos.
- **Solución**: Aumentar la complejidad del modelo (añadir más variables, incluir interacciones o términos no lineales).

La maldición del sobreajuste

Para ilustrar por qué la validación es indispensable, realizaremos un experimento controlado. Crearemos un conjunto de datos donde **conocemos la verdad**: sabemos exactamente qué variables influyen en la respuesta y cuáles son puro ruido. Luego, compararemos dos modelos:

1. **Modelo Completo**: Un modelo que incluye todas las variables disponibles, tanto las útiles como las de ruido.
2. **Modelo Correcto**: Un modelo que incluye únicamente las variables que realmente tienen un efecto sobre la respuesta.

El objetivo es ver cuál de los dos modelos predice mejor en datos "no vistos", utilizando la validación cruzada para simular este escenario.

1. Simulación de Datos

Creamos un dataset con 100 observaciones. La variable `y` dependerá de `X1`, `X2` y `X3`. Las variables `X4` a `X10` no tendrán ninguna relación real con `y`; serán **predictores de ruido**.

```r
# Cargar la librería 'caret', que simplifica enormemente el proceso de validación
suppressPackageStartupMessages(library(caret))

# Para reproducibilidad
set.seed(42)

# Crear datos de ejemplo
n <- 100
# 3 predictores verdaderos y 7 de ruido
X <- matrix(rnorm(n * 10), n, 10)
colnames(X) <- paste0("X", 1:10)

# La respuesta 'y' depende SOLO de X1, X2 y X3
beta_true <- c(2.5, -1.5, 3, 0, 0, 0, 0, 0, 0, 0)
y <- X %*% beta_true + rnorm(n, sd = 2)

# Combinar en un data frame
datos <- data.frame(y = y, X)

indices_modelado <- createDataPartition(datos$y, p = 0.8, list = FALSE)
datos_modelado <- datos[indices_modelado, ]
datos_prueba_final <- datos[-indices_modelado, ]
```

2. Ajuste y Evaluación de Modelos con Validación Cruzada
Usaremos la función `train()` del paquete `caret`, que es una herramienta increíblemente potente para ajustar y validar modelos. Configuraremos una **validación cruzada de 10 particiones (10-fold CV)** para estimar el error de predicción (RMSE) de nuestros dos modelos.

```r
# Configurar el método de validación cruzada de 10 particiones
control_cv <- trainControl(method = "cv", number = 10)

# Ajustar y validar el MODELO COMPLETO (incluye predictores de ruido)
modelo_completo <- train(
  y ~ .,
  data = datos_modelado,
  method = "lm",
  trControl = control_cv
)

# Ajustar y validar el MODELO CORRECTO (solo los predictores relevantes)
modelo_correcto <- train(
  y ~ X1 + X2 + X3,
  data = datos_modelado,
  method = "lm",
  trControl = control_cv
)

# Comparar los resultados de la validación cruzada de ambos modelos
resultados_cv <- resamples(list(COMPLETO = modelo_completo, CORRECTO = modelo_correcto))
resumen_cv <- summary(resultados_cv)
resumen_cv
```

```
Call:
summary.resamples(object = resultados_cv)

Models: COMPLETO, CORRECTO
Number of resamples: 10

MAE
              Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
COMPLETO 0.8638906 1.284897 1.585880 1.685953 1.895840 2.783321    0
CORRECTO 0.7192667 1.279226 1.602489 1.687326 2.283342 2.477780    0

RMSE
              Min.  1st Qu.   Median     Mean  3rd Qu.     Max. NA's
COMPLETO 1.0722692 1.592663 2.105678 2.116903 2.523848 3.119017    0
CORRECTO 0.9052294 1.474694 1.934304 2.020464 2.480041 3.124041    0
```

```
Rsquared
              Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
COMPLETO 0.5917775 0.7793610 0.8556182 0.8189137 0.8882997 0.9505874    0
CORRECTO 0.5019737 0.7988599 0.8513481 0.8017851 0.9048191 0.9385882    0
```

```r
# Extraer métricas RMSE para uso en el texto
rmse_correcto <- round(resumen_cv$statistics$RMSE["CORRECTO", "Mean"], 3)
rmse_completo <- round(resumen_cv$statistics$RMSE["COMPLETO", "Mean"], 3)
```

3. Análisis de Resultados
Al comparar el **RMSE** promedio obtenido en la validación cruzada, la conclusión es clara: el **Modelo Correcto** (2.02) es consistentemente **mejor** (menor error) que el **Modelo Completo** (2.117).
El **Modelo Completo** sufre de **sobreajuste**. Al incluir las 7 variables de ruido, se esfuerza por encontrar patrones en datos puramente aleatorios. "Aprende" estas relaciones falsas en los datos de entrenamiento, pero falla al predecir en datos nuevos. El **Modelo Correcto**, al ser más parsimonioso, captura la estructura fundamental y generaliza mejor.
Este ejemplo demuestra la lección más importante del modelado: **un buen ajuste en los datos de entrenamiento no garantiza un buen rendimiento predictivo**. La validación es el único método fiable para estimar la verdadera calidad de un modelo.
4. El Veredicto Final en el Conjunto de Prueba
La validación cruzada nos ha servido como un juez imparcial para comparar nuestros modelos candidatos y seleccionar el `Modelo Correcto` como el claro ganador. Ahora, para obtener una estimación final y no sesgada de su rendimiento en el mundo real, tomamos ese modelo elegido y lo enfrentamos a los `datos_prueba_final`, el conjunto de datos que ha permanecido intacto durante todo el proceso.

```r
# Usamos el modelo ganador (modelo_correcto) para predecir sobre los datos de prueba
predicciones_finales <- predict(modelo_correcto, newdata = datos_prueba_final)

# Calculamos el RMSE final comparando las predicciones con los valores reales de prueba
rmse_final <- RMSE(predicciones_finales, datos_prueba_final$y)
```

Al evaluar nuestro modelo final, obtenemos un RMSE en el conjunto de prueba de **1.966**. Este valor es nuestra estimación más honesta del error de predicción que podemos esperar de nuestro modelo al enfrentarse a nuevos datos. Es crucial compararlo con el error que estimamos durante la validación cruzada (2.02). El hecho de que ambos valores sean muy similares confirma que nuestro proceso de validación fue robusto y que no hemos sobreajustado el modelo al conjunto de datos de modelado. Este RMSE final es el que reportaríamos como la medida definitiva del rendimiento predictivo de nuestro modelo.

6 Modelos de regresión generalizada

Hasta ahora hemos estuadiado la regresión lineal como una herramienta poderosa para modelar la relación entre una variable dependiente continua y un conjunto de variables independientes. Sin embargo, en muchos contextos del mundo real, las suposiciones de la regresión lineal tradicional no son adecuadas. ¿Qué sucede si la variable dependiente es binaria, como en un diagnóstico médico (enfermo/sano)? ¿O si estás modelando el número de accidentes en una intersección o la cantidad de compras realizadas por un cliente?

Para abordar estos desafíos, se utilizan los llamados **Modelos Lineales Generalizados (GLM)**. Esta clase de modelos amplía la regresión lineal al permitir que la variable dependiente tenga distribuciones diferentes a la normal, como la binomial o la de Poisson. Además, los GLM utilizan funciones de enlace que transforman la relación entre la variable dependiente y los predictores, permitiendo una mayor flexibilidad en el modelado.

Algunos de los modelos más comunes dentro de los GLM son:

- Regresión Logística: Ideal para variables dependientes binarias (sí/no, éxito/fracaso).
- Regresión de Poisson: Utilizada para modelar datos de conteo (número de eventos).
- Regresión Binomial Negativa: Una extensión de la regresión de Poisson para datos de conteo con sobredispersión.
- Modelos de Gamma y Inverso Gaussiano: Utilizados para modelar variables continuas positivas y sesgadas, como tiempos de espera o costos.

En este tema, exploraremos cómo utilizar estos modelos para resolver problemas del mundo real, interpretar sus resultados y evaluar su ajuste.

6.1 Introducción a los GLM

6.1.1 ¿Qué son los modelos lineales generalizados?

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 (Nelder and Wedderburn 1972). Mientras que la regresión lineal tradicional asume que la variable dependiente es continua y sigue una distribución normal, los GLM permiten trabajar con variables dependientes que:

- Son **binarias** (como éxito/fracaso o sí/no).
- Representan **conteos** de eventos (número de llamadas, accidentes, etc.).
- Son **continuas positivas** y no siguen una distribución normal (como tiempos o costos).

Los GLM proporcionan una estructura flexible para modelar la relación entre una o más variables independientes y una variable dependiente que sigue alguna distribución de la **familia exponencial** (binomial, Poisson, gamma, entre otras).

6.1.2 Componentes de un modelo lineal generalizado

Un GLM se define por tres componentes clave:

1. **Componente Aleatorio:**

   Este componente describe la distribución de la variable dependiente. En la regresión lineal, la variable dependiente sigue una distribución normal. En los GLM, puede seguir otras distribuciones de la **familia exponencial**, como:

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

2. **Componente Sistemático:**
   Este componente describe cómo las variables independientes se combinan linealmente en el modelo. Se define como:

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

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

3. **Función de Enlace:**
   La función de enlace conecta el componente sistemático con la media de la variable dependiente. Mientras que en la regresión lineal la relación es directa ($y = \eta$), en los GLM se utiliza una función de enlace $g(\mu)$ para transformar la media μ y ajustar diferentes tipos de datos.

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

Ejemplos de funciones de enlace:

- **Logística (Logit):** Para la regresión logística, que modela la probabilidad de un evento.

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

- **Logarítmica:** Para la regresión de Poisson, que modela tasas de eventos.

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

- **Identidad:** Para la regresión lineal estándar.

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

i Aplicaciones

Los GLM se utilizan en una amplia variedad de disciplinas para resolver problemas del mundo real:
Regresión Logística (para variables binarias):

- **Medicina:** Predicción de la presencia o ausencia de una enfermedad basada en factores de riesgo.
- **Marketing:** Determinación de la probabilidad de que un cliente compre un producto.
- **Finanzas:** Evaluación de la probabilidad de incumplimiento de pago de un préstamo.

Regresión de Poisson (para datos de conteo):

- **Transporte:** Modelado del número de accidentes en una carretera en un período de tiempo.
- **Ecología:** Conteo de especies en un área determinada.
- **Telecomunicaciones:** Número de llamadas recibidas por un centro de atención.

Regresión Binomial Negativa (para conteos con sobredispersión):

- **Salud Pública:** Modelado del número de visitas al médico o incidentes de una enfermedad en una población.

Modelos Gamma (para variables continuas positivas):

- **Seguros:** Estimación de los costos de reclamos de seguros.
- **Ingeniería:** Modelado de tiempos de falla en procesos industriales.

6.1.3 Diferencias clave entre la regresión lineal y los GLM

Característica	Regresión Lineal	Modelos Lineales Generalizados (GLM)
Distribución de la variable dependiente	Normal	Familia exponencial (binomial, Poisson, gamma, etc.)
Tipo de variable dependiente	Continua	Binaria, de conteo, continua positiva
Relación entre las variables	Lineal directa	Relación transformada mediante una función de enlace
Función de Enlace	Identidad ($g(\mu) = \mu$)	Logit, logarítmica, inversa, etc.

Las ventajas principales de los GLM son:

- **Flexibilidad:** Los GLM permiten modelar diferentes tipos de variables dependientes, lo que amplía significativamente el rango de problemas que se pueden abordar.
- **Interpretación Coherente:** Aunque se utilizan funciones de enlace, los coeficientes de los GLM pueden interpretarse de manera similar a los modelos lineales, proporcionando información sobre el impacto de cada variable independiente.
- **Evaluación Estadística Robusta:** Los GLM permiten la realización de pruebas de hipótesis, la construcción de intervalos de confianza y la evaluación de la bondad del ajuste mediante medidas ya conocidas como el AIC o el BIC.

Los **Modelos Lineales Generalizados** amplían el alcance de la regresión lineal clásica, proporcionando herramientas para modelar una amplia variedad de tipos de datos, desde variables binarias hasta datos de conteo y variables continuas no normales. A través del uso de funciones de enlace y distribuciones flexibles, los GLM permiten resolver problemas complejos del mundo real en campos tan diversos como la medicina, el marketing, la ingeniería y las ciencias sociales.

En las próximas secciones, exploraremos en detalle cómo aplicar estos modelos específicos, como la **regresión logística** y la **regresión de Poisson**, y cómo interpretar sus resultados en diferentes contextos.

6.2 Estimación de parámetros en GLM

La estimación de parámetros en los **Modelos Lineales Generalizados** representa un aspecto fundamental que diferencia estos modelos de la regresión lineal clásica. Mientras que en la regresión lineal utilizamos mínimos cuadrados ordinarios para obtener estimadores con

propiedades óptimas, en los GLM necesitamos métodos más sofisticados debido a la naturaleza no normal de las distribuciones involucradas y las funciones de enlace no lineales.

La estimación en GLM se basa en el **principio de máxima verosimilitud**, que proporciona un marco teórico unificado para todos los modelos de la familia exponencial. Este enfoque no solo garantiza propiedades estadísticas deseables de los estimadores, sino que también permite el desarrollo de algoritmos computacionales eficientes para encontrar las soluciones.

En esta sección exploraremos los fundamentos teóricos de la estimación por máxima verosimilitud, el algoritmo iterativo IRLS (Iteratively Reweighted Least Squares) que implementan los software estadísticos, y los problemas prácticos que pueden surgir durante el proceso de estimación. Comprender estos aspectos es crucial para interpretar correctamente los resultados y diagnosticar posibles problemas en el ajuste de los modelos.

6.2.1 Método de máxima verosimilitud

A diferencia de la regresión lineal que utiliza el método de **mínimos cuadrados**, los GLM emplean el **método de máxima verosimilitud** para estimar los parámetros del modelo. Este cambio metodológico es necesario debido a que las distribuciones de la familia exponencial no siempre tienen una relación lineal directa con los predictores, y además porque la varianza de la variable respuesta depende de su media, violando el supuesto de homocedasticidad que requieren los mínimos cuadrados.

El principio de máxima verosimilitud consiste en encontrar los valores de los parámetros β que hacen más probable observar los datos que tenemos. Para una muestra de n observaciones independientes $y_1, y_2, \dots, y_n$, la **función de verosimilitud** se define como la probabilidad conjunta de observar estos datos dado un conjunto de parámetros:

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

donde $f(y_i; \theta_i, \phi)$ es la función de densidad (o masa) de probabilidad de la observación i. En la práctica, es más conveniente trabajar con el logaritmo de esta función, conocida como **log-verosimilitud**:

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

La ventaja de usar el logaritmo es que convierte productos en sumas, simplificando considerablemente los cálculos matemáticos y numéricos.

La clave para entender los GLM radica en reconocer que todas las distribuciones que podemos usar (binomial, Poisson, gamma, etc.) pertenecen a la **familia exponencial**. Estas distribuciones pueden expresarse en 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** o canónico que está relacionado directamente con la media de la distribución, ϕ es el **parámetro de dispersión** que controla la variabilidad, y $b(\theta)$, $a(\phi)$ y $c(y,\phi)$ son funciones específicas de cada distribución que determinan sus propiedades particulares.

Esta forma unificada tiene propiedades matemáticas muy convenientes que hacen que los GLM sean tanto elegantes teóricamente como computacionalmente eficientes. La esperanza de Y se obtiene como $E(Y) = \mu = b'(\theta)$ (la derivada de b respecto a θ), y la varianza como $\text{Var}(Y) = a(\phi)b''(\theta) = a(\phi)V(\mu)$, donde $V(\mu)$ es la **función de varianza** que caracteriza cómo la varianza depende de la media en cada tipo de distribución.

La Familia Exponencial: Un Vistazo General

La elegancia de los GLM reside en que muchas distribuciones aparentemente distintas comparten una estructura matemática común. Esto permite una teoría unificada. Aquí están los miembros más importantes:

Distribución	Uso Típico	Función de Varianza $V(\mu)$	Enlace Canónico $g(\mu)$
Normal	Datos continuos simétricos	1	Identidad: μ
Binomial	Proporciones, datos binarios (éxito/fracaso)	$\mu(1-\mu)$	Logit: $\log(\frac{\mu}{1-\mu})$
Poisson	Conteos de eventos	μ	Log: $\log(\mu)$
Gamma	Datos continuos positivos y asimétricos (tiempos, costos)	μ^2	Inverso: $1/\mu$
Inversa Gaussiana	Tiempos hasta un evento, datos muy asimétricos	μ^3	Inverso al cuadrado: $1/\mu^2$

La **función de varianza** $V(\mu)$ es la "firma" de cada distribución, ya que define la relación teórica entre la media y la varianza de la respuesta. El **enlace canónico** es la función de enlace que surge de forma natural de la estructura matemática de la distribución, aunque en la práctica se pueden usar otros enlaces.

Esta relación entre media y varianza es fundamental para entender las diferencias entre los diversos GLM. En la regresión lineal clásica, la varianza es constante ($V(\mu) = 1$), pero en otros GLM la función de varianza toma formas específicas:

- **Distribución binomial**: $V(\mu) = \mu(1 - \mu)$ - la varianza es máxima cuando $\mu = 0.5$ y mínima en los extremos.
- **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.

Estas funciones de varianza no solo determinan la heterocedasticidad inherente de cada distribución, sino que también influyen directamente en los pesos del algoritmo IRLS y en la precisión de las estimaciones. Por ejemplo, en regresión logística, las observaciones con probabilidades cercanas a 0.5 tienen mayor varianza y, por tanto, menor peso en la estimación, mientras que en regresión de Poisson, las observaciones con conteos más altos contribuyen con mayor peso al ajuste del modelo.

6.2.1.1 Algoritmo de Newton-Raphson (IRLS)

Para encontrar los valores de β que maximizan la log-verosimilitud, los GLM utilizan un algoritmo iterativo conocido como **Iteratively Reweighted Least Squares (IRLS)**, que es una implementación especializada del método de Newton-Raphson. La necesidad de un algoritmo iterativo surge porque, a diferencia de la regresión lineal donde existe una solución analítica cerrada, en los GLM las ecuaciones de verosimilitud no tienen solución directa debido a la presencia de funciones no lineales.

El algoritmo IRLS se basa en la idea de que podemos aproximar la función de enlace y la varianza de la distribución en torno a un valor central (la media) y luego aplicar mínimos cuadrados de manera iterativa para ajustar los parámetros del modelo. Los pasos básicos del algoritmo son:

1. **Inicialización:** Establecer valores iniciales para los parámetros $\beta^{(0)}$.
2. **Iteración t:**
   - Calcular el predictor lineal: $\eta_i^{(t)} = \mathbf{x}_i^T \beta^{(t)}$.
   - Calcular la media estimada: $\mu_i^{(t)} = g^{-1}(\eta_i^{(t)})$.
   - Calcular los pesos: $w_i^{(t)} = \frac{1}{\text{Var}(\mu_i^{(t)})} \left(\frac{d\mu_i}{d\eta_i}\right)^2$.
   - Calcular la variable dependiente ajustada: $z_i^{(t)} = \eta_i^{(t)} + (y_i - \mu_i^{(t)}) \frac{d\eta_i}{d\mu_i}$.
3. **Actualización:** Actualizar los parámetros del modelo:

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

4. **Convergencia:** Repetir el proceso hasta que la diferencia entre iteraciones sucesivas sea menor que un umbral predefinido.

Ejemplo: Convergencia del algoritmo IRLS

```r
# Ejemplo de seguimiento de la convergencia en regresión logística
library(MASS)
data(Pima.tr)

# Función para mostrar el proceso iterativo
mostrar_convergencia <- function() {
  # Ajustar modelo con seguimiento de iteraciones
  modelo <- glm(type ~ glu + bmi, data = Pima.tr, family = binomial,
                control = glm.control(trace = TRUE, maxit = 10))

  cat("Número de iteraciones necesarias:", modelo$iter, "\n")
  cat("¿Convergió?", modelo$converged, "\n")
  cat("Log-likelihood final:", logLik(modelo), "\n")

  return(modelo)
}

# Ejecutar y mostrar convergencia
modelo_ejemplo <- mostrar_convergencia()
```

```
Deviance = 199.36 Iterations - 1
Deviance = 198.4772 Iterations - 2
Deviance = 198.4704 Iterations - 3
Deviance = 198.4704 Iterations - 4
Número de iteraciones necesarias: 4
¿Convergió? TRUE
Log-likelihood final: -99.23522
```

6.2.1.2 Propiedades de los estimadores de máxima verosimilitud

Los estimadores de máxima verosimilitud en GLM poseen propiedades estadísticas muy atractivas que los convierten en la elección preferida para la estimación de parámetros. Estas propiedades son **asintóticas**, lo que significa que se cumplen cuando el tamaño de la muestra tiende a infinito, pero en la práctica proporcionan una excelente aproximación para muestras moderadamente grandes.

1. Consistencia:

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

La **consistencia** garantiza que a medida que aumentamos el tamaño de la muestra, nuestros estimadores se acercan cada vez más al valor verdadero de los parámetros. Esto significa que con suficientes datos, los estimadores de máxima verosimilitud convergerán al valor real de β, eliminando el sesgo de estimación. Esta propiedad es fundamental porque nos asegura que no estamos introduciendo errores sistemáticos en nuestras estimaciones.

2. Normalidad asintótica:

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

La **normalidad asintótica** establece que la distribución de los estimadores, apropiadamente escalada, se aproxima a una distribución normal multivariada cuando el tamaño de la muestra es grande. Esta propiedad es crucial porque:

- Permite construir **intervalos de confianza** para los parámetros usando la distribución normal
- Facilita la realización de **pruebas de hipótesis** sobre los coeficientes
- Proporciona la base teórica para los **estadísticos de Wald** utilizados en las pruebas de significancia

La matriz de covarianza asintótica $\mathbf{I}^{-1}(\beta)$ nos permite calcular los errores estándar de nuestras estimaciones, que son esenciales para la inferencia estadística.

3. Eficiencia:

Los estimadores MV alcanzan la **cota de Cramér-Rao**, siendo asintóticamente eficientes. Esto significa que:

- Entre todos los estimadores insesgados posibles, los de máxima verosimilitud tienen la **menor varianza asintótica**
- No existe otro método de estimación que, bajo las mismas condiciones, produzca estimadores con menor incertidumbre
- Utilizan la información disponible en los datos de manera **óptima**

En términos prácticos, esta eficiencia se traduce en intervalos de confianza más estrechos y pruebas de hipótesis más poderosas comparado con otros métodos de estimación.

6.2.1.3 Matriz de información y errores estándar

La implementación práctica de estas propiedades teóricas requiere el cálculo de la **matriz de información**, que cuantifica la cantidad de información que contienen los datos sobre los parámetros del modelo.

La **matriz de información de Fisher** se define teóricamente como:

$$\mathbf{I}(\beta) = E\left[-\frac{\partial^2 \ell}{\partial\beta\partial\beta^T}\right]$$

Sin embargo, en la práctica utilizamos la **matriz de información observada**, que se calcula directamente de nuestros datos:

$$\mathbf{I}(\hat{\beta}) = -\frac{\partial^2 \ell}{\partial\beta\partial\beta^T}\bigg|_{\beta=\hat{\beta}}$$

Esta matriz representa las segundas derivadas de la log-verosimilitud evaluadas en nuestras estimaciones. Intuitivamente, mide qué tan "puntiaguda" es la función de verosimilitud alrededor del máximo: una función más puntiaguda indica mayor información y, por tanto, menor incertidumbre en la estimación.

Para los GLM, el algoritmo IRLS proporciona una aproximación computacionalmente eficiente:

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

donde $\mathbf{W}$ es la matriz diagonal de pesos calculada en la última iteración del algoritmo. Esta aproximación es exacta para la distribución normal y muy buena para otras distribuciones de la familia exponencial.

Los **errores estándar** de los coeficientes individuales se obtienen como las raíces cuadradas de los elementos diagonales de la matriz de covarianza:

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

Estos errores estándar son fundamentales para:

- **Intervalos de confianza:** $\hat{\beta}_j \pm z_{\alpha/2} \cdot \text{SE}(\hat{\beta}_j)$
- **Estadísticos de prueba:** $z_j = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)}$
- **Evaluación de la precisión** de nuestras estimaciones

Es importante recordar que estos errores estándar son válidos bajo los supuestos del modelo GLM y que violaciones serias de estos supuestos (como sobredispersión en modelos de Poisson) pueden hacer que sean inadecuados.

6.3 Bondad de ajuste en GLMs

Los **Modelos Lineales Generalizados** requieren métodos específicos para evaluar la calidad del ajuste que van más allá de las métricas tradicionales de la regresión lineal. Mientras que en la regresión lineal clásica utilizamos el coeficiente de determinación (R^2) y la suma de cuadrados residuales como medidas principales de bondad de ajuste, en los GLMs estas métricas no son apropiadas debido a las diferentes distribuciones subyacentes y las funciones de enlace no lineales.

La evaluación de la bondad de ajuste en GLMs se basa fundamentalmente en conceptos de **verosimilitud** y **deviance**, que proporcionan una base teórica sólida para comparar modelos y evaluar su calidad de ajuste. Esta aproximacion basada en la verosimilitud es coherente con el método de estimación utilizado en estos modelos.

6.3.1 La deviance como medida de bondad de ajuste

La **deviance** (o **desviación**) es la medida principal de bondad de ajuste en los **Modelos Lineales Generalizados**. Conceptualmente, representa una generalización de la suma de cuadrados residuales de la regresión lineal para distribuciones no normales, pero su interpretación y cálculo son fundamentalmente diferentes.

La deviance se basa en el principio de **máxima verosimilitud** y mide qué tan bien el modelo propuesto se ajusta a los datos comparado con el mejor ajuste posible. Para entender este concepto, es importante distinguir entre dos tipos de modelos:

1. **Modelo Saturado**: Un modelo hipotético que tiene tantos parámetros como observaciones, por lo que puede predecir perfectamente cada valor observado. Este modelo representa el "ajuste perfecto" teórico.
2. **Modelo Propuesto**: El modelo que estamos evaluando, con un número limitado de parámetros basado en nuestras variables predictoras.

La deviance mide la **diferencia en log-verosimilitud** entre estos dos modelos:

$$D = 2 \sum_{i=1}^{n} [\ell(y_i; y_i) - \ell(y_i; \hat{\mu}_i)]$$

donde:

- $\ell(y_i; y_i)$ = Log-verosimilitud del modelo saturado para la observación i.
- $\ell(y_i; \hat{\mu}_i)$ = Log-verosimilitud del modelo propuesto para la observación i.
- El factor 2 se incluye para que la deviance siga aproximadamente una distribución chi-cuadrado bajo ciertas condiciones.

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

Comparación relativa: La deviance es más útil para **comparar modelos** que para evaluación absoluta. Un modelo con menor deviance indica mejor ajuste, pero el valor absoluto depende del tamaño de la muestra y la naturaleza de los datos.

Deviance residual vs. deviance nula:

- **Deviance nula**: Deviance del modelo que solo incluye el intercepto (sin predictores)
- **Deviance residual**: Deviance del modelo con todos los predictores incluidos
- La diferencia entre ambas indica cuánto mejora el modelo al incluir las variables predictoras

6.3.2 Test de la razón de verosimilitudes

El **test de la razón de verosimilitudes** es la herramienta principal para comparar modelos anidados en GLMs y para evaluar la significancia global del modelo. Se basa en el principio de que si un modelo más complejo no mejora significativamente el ajuste, debemos preferir el modelo más simple por parsimonia.

Cuando comparamos dos modelos anidados (donde uno es un caso especial del otro), la diferencia en sus deviances sigue aproximadamente una distribución chi-cuadrado:

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

donde df es la diferencia en grados de libertad (número de parámetros) entre los modelos.

Interpretación del test:

- **Hipótesis nula (H_0)**: El modelo reducido es adecuado (los parámetros adicionales no son necesarios)
- **Hipótesis alternativa (H_1)**: El modelo completo es significativamente mejor
- **Decisión**: Si p-valor $< \alpha$ (típicamente 0.05), rechazamos H_0 y preferimos el modelo completo

Aplicaciones principales:

1. **Significancia global del modelo:** Comparar el modelo completo con el modelo nulo (solo intercepto) para determinar si las variables predictoras aportan información significativa.
2. **Selección de variables:** Evaluar si la inclusión o exclusión de variables específicas mejora significativamente el ajuste del modelo.
3. **Comparación de especificaciones:** Decidir entre diferentes formas funcionales o distribuciones para el mismo conjunto de datos.

Ejemplo: Comparando Modelos con el Test de Razón de Verosimilitudes

Supongamos que queremos determinar si añadir la variable `disp` (cilindrada) a un modelo de regresión logística que ya contiene `wt` (peso) mejora significativamente la predicción de si un coche tiene transmisión automática (`am`).

```
# Usaremos el dataset mtcars
data(mtcars)

# Modelo Reducido: solo contiene 'wt'
modelo_reducido <- glm(am ~ wt, data = mtcars, family = binomial)

# Modelo Completo: contiene 'wt' y 'disp'
modelo_completo <- glm(am ~ wt + disp, data = mtcars, family = binomial)

# Realizamos el Test de Razón de Verosimilitudes (LRT)
# En R, esto se hace con la función anova() especificando el test
lrt_resultado <- anova(modelo_reducido, modelo_completo, test = "LRT")

# Mostramos los resultados
print(lrt_resultado)
```

```
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 (es decir,
$beta\_disp = 0$).

- La diferencia en deviance es de 1.391 con 1 grado de libertad (el parámetro adicional).
- El p-valor asociado es **0.238**.

Dado que el p-valor es mayor que 0.05, **no rechazamos la hipótesis nula**. Esto significa que, una vez que tenemos en cuenta el peso del coche (`wt`), añadir la cilindrada (`disp`) no aporta una mejora estadísticamente significativa al modelo. Nos quedaríamos con el **modelo reducido** por el principio de parsimonia.

6.4 Diagnosis de GLMs

La diagnosis de GLMs implica evaluar los supuestos del modelo y detectar problemas potenciales que puedan afectar la validez de las inferencias. A diferencia de la regresión lineal, donde los residuos ordinarios proporcionan información diagnóstica directa, los GLMs requieren herramientas especializadas debido a la heterocedasticidad inherente y las diferentes distribuciones subyacentes.

En lugar de una simple lista de comprobación, abordaremos el diagnóstico respondiendo a tres preguntas clave que un analista se haría, utilizando diferentes tipos de **residuos** para obtener las respuestas.

6.4.1 Tipos de Residuos en GLMs

La elección del tipo de residuo apropiado es crucial. Los residuos "crudos" ($y_i - \hat{\mu}_i$) no son homocedásticos, por lo que se utilizan versiones estandarizadas. Los más importantes son:

- **Residuos Pearson:** Estandarizan el residuo crudo dividiendo por la desviación estándar predicha por el modelo. Son un análogo directo a los residuos estandarizados en regresión lineal.
$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$
- **Residuos Estudentizados:** Son una mejora de los residuos Pearson que también tienen en cuenta el leverage (h_i) de cada observación. Son más fiables para la detección de outliers.
$$r_{S_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)(1 - h_i)}}$$
- **Residuos deviance:** Son los más recomendados para la inspección visual en gráficos diagnósticos. Su construcción, basada en la contribución de cada punto a la deviance

total, les confiere propiedades muy deseables: su distribución se aproxima mejor a la normalidad y su varianza es más estable que la de otros residuos.

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

6.4.2 ¿La forma del modelo es correcta? (Linealidad y Enlace)

Esta primera pregunta evalúa si la estructura básica del modelo, $g(\mu) = X\beta$, es adecuada para los datos. Para ello, utilizamos varias herramientas de diagnóstico:

- El **gráfico de residuos vs. valores ajustados** es la herramienta fundamental. Se grafican los residuos (idealmente, deviance) contra los valores predichos en la escala del predictor lineal ($\hat{\eta}_i$). Si la forma del modelo es correcta, no deberíamos ver ningún patrón sistemático. Una tendencia curvilínea es una señal clara de que la forma funcional o la función de enlace son incorrectas.
- Los **gráficos de residuos parciales** son esenciales para evaluar si la función de enlace es apropiada para **cada predictor individualmente**. Un patrón no lineal en este gráfico sugiere que la relación de esa variable específica con la respuesta no es la que asume el modelo.
- El **test de especificación de enlace** (como el *Linktest* de Pregibon) ofrece una prueba formal. La idea es ajustar un segundo modelo que incluye el predictor lineal al cuadrado ($\hat{\eta}^2$) como una variable adicional. Si este término cuadrático resulta significativo, es una fuerte evidencia de que la función de enlace está mal especificada.

Si estos diagnósticos revelan problemas, las estrategias de corrección incluyen aplicar **transformaciones** a los predictores (ej. logaritmo, términos polinómicos), **incluir términos de interacción** para capturar relaciones no aditivas, o directamente **cambiar la función de enlace** a una que se ajuste mejor a los datos.

6.4.3 ¿La distribución que elegimos es la correcta? (Varianza y Normalidad)

Esta pregunta evalúa si la elección de la familia de distribución (Poisson, Binomial, etc.) fue acertada, lo que implica verificar la relación entre la media y la varianza, así como la forma general de los errores.

Un primer aspecto clave al verificar la distribución es la **sobredispersión**. Imagina que tu modelo es como una regla estricta sobre el comportamiento de los datos. El modelo de Poisson, por ejemplo, impone que la varianza de los conteos debe ser igual a su media ($Var(Y) = \mu$). La sobredispersión ocurre cuando tus datos reales son más "desordenados" o variables de lo que esta regla permite.

Para detectar este problema de forma objetiva, calculamos el **Estadístico de dispersión ($\hat{\phi}$)**, que compara la varianza observada (a través de los residuos Pearson) con la esperada:

$$\hat{\phi} = \frac{X^2_{\text{Pearson}}}{n-p} = \frac{\sum r_i^2}{n-p}$$

La interpretación es directa:

- **Si** $\hat{\phi} \approx 1$: ¡Perfecto! La dispersión de los datos es la que el modelo esperaba.
- **Si** $\hat{\phi} > 1$: Tienes **sobredispersión**. El modelo está subestimando la variabilidad real de los datos, lo que invalida las inferencias (errores estándar, p-valores).

La estrategia para corregirlo no es forzar los datos, sino cambiar a un modelo más flexible. El caso clásico es pasar del modelo de **Poisson** al modelo **Binomial Negativo**. Este último funciona porque su fórmula para la varianza incluye un parámetro de dispersión adicional (α) que le permite modelar esa variabilidad extra que el modelo de Poisson no puede capturar:

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

Este término adicional se ajusta a la variabilidad de los datos, proporcionando estimaciones y errores estándar mucho más fiables.

Un segundo aspecto es la forma general de la distribución, que se evalúa con el **Gráfico Q-Q de residuos deviance**. 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. Desviaciones sistemáticas de la línea diagonal en el gráfico Q-Q pueden indicar que la distribución asumida para los datos es incorrecta.

6.4.4 ¿Hay observaciones que distorsionan el modelo? (Atípicos e Influyentes)

Finalmente, buscamos identificar puntos individuales que tienen una influencia desproporcionada en el modelo. Las herramientas matemáticas para ello son generalizaciones de las vistas en regresión lineal:

- **Leverage generalizado (h_i):** Mide el potencial de una observación para ser influyente debido a su posición en el espacio de los predictores.

  $$h_i = w_i \mathbf{x}_i^T (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{x}_i$$

  donde w_i y $\mathbf{W}$ son el peso y la matriz de pesos de la última iteración del algoritmo IRLS.

- **Distancia de Cook para GLMs (D_i):** Mide la influencia global de una observación en *todos* los coeficientes. Utiliza el residuo Pearson (r_i).

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

- **DFBETAS:** Mide la influencia de la observación i en *cada coeficiente individual* β_j. Es útil para ver si un punto influyente está afectando a una variable de interés particular.

$$\text{DFBETA}_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{\text{SE}(\hat{\beta}_{j(-i)})} \approx \frac{(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}_{jj} x_{ij}(y_i - \hat{\mu}_i)}{\text{SE}(\hat{\beta}_j)\sqrt{1-h_i}}$$

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

6.5 Regresión logística

La **regresión logística** es una herramienta fundamental para modelar la probabilidad de eventos binarios en una variedad de contextos, desde la medicina hasta la economía y el marketing (Hosmer Jr, Lemeshow, and Sturdivant 2013). La correcta interpretación de los coeficientes mediante **odds ratios**, así como la evaluación del ajuste del modelo mediante curvas **ROC** y matrices de confusión, son esenciales para extraer conclusiones válidas de los datos.

6.5.1 Fundamentos de la regresión logística

La **regresión logística** es una técnica estadística utilizada para modelar la probabilidad de ocurrencia de un evento binario, es decir, cuando la variable dependiente toma solo dos posibles valores (por ejemplo, **éxito/fracaso**, **sí/no**, **enfermo/sano**). A diferencia de la regresión lineal, que modela una relación lineal entre variables, la regresión logística utiliza una **función logística** para asegurar que las predicciones estén en el rango [0,1], lo cual es necesario para interpretar los resultados como probabilidades.

La función Logística (Sigmoide)

La función logística transforma cualquier valor real en un valor comprendido entre 0 y 1. La forma matemática de la función logística es:

$$P(Y = 1|X_1, \ldots, X_p) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}}$$

Donde:

- $P(Y = 1|X_1, \ldots, X_p)$ es la probabilidad de que el evento ocurra.
- β_0 es el intercepto y $\beta_1, \beta_2, \ldots, \beta_p$ son los coeficientes asociados a las variables independientes $X_1, X_2, \ldots, X_p$.

La **curva sigmoide** que representa esta función tiene forma de "S", lo que refleja que para valores muy pequeños o muy grandes del predictor, la probabilidad se aplana hacia 0 o 1, respectivamente.

Función de enlace Logit

En la regresión logística, la relación entre el predictor lineal y la probabilidad se establece mediante la **función de enlace logit**. El logit de una probabilidad p se define como:

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

Esta transformación convierte una probabilidad en una escala que va de $-\infty$ a $+\infty$, lo que permite ajustar un modelo lineal a los datos. El modelo logístico puede expresarse como:

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

6.5.2 Estimación por máxima verosimilitud en regresión logística

La estimación de los parámetros en regresión logística se basa en el método de máxima verosimilitud, adaptado específicamente para la distribución binomial con función de enlace logit.

Para una muestra de n observaciones independientes, donde $y_i \in \{0, 1\}$ representa el resultado binario para la observación i, la **función de verosimilitud** se define como:

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

donde $p_i = P(Y_i = 1|\mathbf{x}_i) = \frac{1}{1+e^{-\mathbf{x}_i^T\beta}}$ es la probabilidad estimada para la observación i.

La **log-verosimilitud** correspondiente es:

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \log(p_i) + (1-y_i)\log(1-p_i)]$$

Sustituyendo la expresión de p_i y simplificando:

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

Para encontrar los valores de β que maximizan la log-verosimilitud, derivamos respecto a cada parámetro:

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

donde x_{ij} es el valor de la variable j para la observación i.

Esta ecuación tiene una interpretación intuitiva: los estimadores de máxima verosimilitud se obtienen cuando la suma de los residuos ponderados por cada variable predictora es igual a cero.

6.5.2.1 Implementación del algoritmo IRLS

Dado que las ecuaciones de verosimilitud no tienen solución analítica cerrada, se utiliza el algoritmo IRLS. Para regresión logística, los elementos específicos son:

Pesos:

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

Variable dependiente ajustada:

$$z_i = \mathbf{x}_i^T \beta^{(t)} + \frac{y_i - p_i^{(t)}}{p_i^{(t)}(1 - p_i^{(t)})}$$

Actualización de parámetros:

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

Ejemplo: Estimación paso a paso en regresión logística

```r
# Demostración del proceso de estimación por máxima verosimilitud
library(MASS)
data(Pima.tr)

# Función para calcular log-verosimilitud manualmente
log_likelihood_logistic <- function(beta, X, y) {
  eta <- X %*% beta
  p <- 1 / (1 + exp(-eta))
  # Evitar problemas numéricos
  p <- pmax(pmin(p, 1-1e-15), 1e-15)
  sum(y * log(p) + (1-y) * log(1-p))
}

# Preparar datos
X <- model.matrix(type ~ glu + bmi, data = Pima.tr)
y <- as.numeric(Pima.tr$type == "Yes")

# Ajuste con glm para comparación
modelo_glm <- glm(type ~ glu + bmi, data = Pima.tr, family = binomial)
beta_glm <- coef(modelo_glm)

cat("=== ESTIMACIÓN POR MÁXIMA VEROSIMILITUD ===\n")
```

```
=== ESTIMACIÓN POR MÁXIMA VEROSIMILITUD ===
```

```r
cat("Coeficientes estimados por glm():\n")
```

```
Coeficientes estimados por glm():
```

```r
print(beta_glm)
```

```
(Intercept)         glu         bmi
-8.21610630  0.03571601  0.09001639
```

```r
# Verificar que estos coeficientes maximizan la log-verosimilitud
ll_optimo <- log_likelihood_logistic(beta_glm, X, y)
cat("\nLog-verosimilitud en el óptimo:", round(ll_optimo, 4), "\n")
```

```

Log-verosimilitud en el óptimo: -99.2352
```

```r
# Comparar con valores ligeramente diferentes
beta_test <- beta_glm + c(0.1, 0, 0)
ll_test <- log_likelihood_logistic(beta_test, X, y)
cat("Log-verosimilitud con perturbación:", round(ll_test, 4), "\n")
```

```
Log-verosimilitud con perturbación: -99.3995
```

```r
cat("Diferencia:", round(ll_optimo - ll_test, 4), "\n")
```

```
Diferencia: 0.1643
```

```r
# Mostrar información de convergencia
cat("\nInformación del algoritmo IRLS:\n")
```

```

Información del algoritmo IRLS:
```

```r
cat("Iteraciones necesarias:", modelo_glm$iter, "\n")
```

```
Iteraciones necesarias: 4
```

```r
cat("¿Convergió?", modelo_glm$converged, "\n")
```

```
¿Convergió? TRUE
```

```r
# Matriz de información y errores estándar
vcov_matrix <- vcov(modelo_glm)
cat("\nErrores estándar de los coeficientes:\n")
```

```

Errores estándar de los coeficientes:
```

```r
print(sqrt(diag(vcov_matrix)))
```

```
(Intercept)         glu         bmi
1.346965130 0.006311023 0.031268458
```

6.5.3 Interpretación de coeficientes y odds ratios

Uno de los aspectos más importantes de la regresión logística es la interpretación de los coeficientes. Dado que los coeficientes están en la escala del logit, su interpretación directa no es tan intuitiva como en la regresión lineal. Sin embargo, podemos interpretarlos utilizando **odds** y **odds ratios**.

El **odds** o razón de probabilidades de que ocurra un evento es el cociente entre la probabilidad de que ocurra el evento y la probabilidad de que no ocurra:

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

Por ejemplo, si la probabilidad de éxito es 0.8, el odds sería:

$$\text{odds} = \frac{0.8}{1-0.8} = 4$$

Esto significa que el evento es **4 veces más probable** que no ocurra.

El **odds ratio (OR)** mide el cambio en los odds cuando una variable independiente aumenta en una unidad. Se calcula como el exponencial del coeficiente de la regresión logística:

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

Interpretación de OR:

- Si $\mathbf{OR > 1}$, el evento es más probable a medida que aumenta la variable independiente.
- Si $\mathbf{OR < 1}$, el evento es menos probable a medida que aumenta la variable independiente.
- Si $\mathbf{OR = 1}$, no hay efecto.

Ejemplo

Supongamos que ajustamos un modelo de regresión logística para predecir la probabilidad de tener diabetes en función del índice de masa corporal (BMI). El coeficiente asociado a **BMI** es 0.08.

$$\text{OR} = e^{0.08} \approx 1.083$$

Esto significa que por cada incremento de 1 unidad en el BMI, la **odds** de tener diabetes aumentan en un **8.3%**.

6.5.4 Bondad de ajuste del modelo logístico

La evaluación de la bondad de ajuste en regresión logística presenta desafíos únicos debido a la naturaleza binaria de la variable dependiente. A diferencia de la regresión lineal, donde el R^2 tradicional proporciona una medida directa de la varianza explicada, en la regresión logística necesitamos adoptar enfoques alternativos que se basen en la verosimilitud del modelo y que sean apropiados para datos categóricos.

La bondad de ajuste en regresión logística se evalúa principalmente a través de dos enfoques complementarios: la **deviance** y los **pseudo R^2**. Ambos métodos nos permiten cuantificar qué tan bien nuestro modelo captura los patrones subyacentes en los datos binarios.

La **deviance** en regresión logística se calcula utilizando la distribución binomial subyacente. Para cada observación, comparamos la probabilidad que predice nuestro modelo con la "probabilidad perfecta" que asignaría un modelo saturado. Matemáticamente, esto se expresa como:

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

donde $y_i \in \{0,1\}$ representa el resultado observado y $\hat{p}_i$ es la probabilidad estimada por nuestro modelo. Es importante notar que cuando $y_i = 0$ o $y_i = 1$, algunos términos en esta expresión se anulan automáticamente, lo que refleja la naturaleza discreta de los datos binarios.

La interpretación de la deviance sigue principios similares a los discutidos anteriormente, pero es crucial recordar que ahora estamos tratando con datos binarios. La deviance nos ayuda a entender qué tan bien nuestro modelo logístico se ajusta a los datos de respuesta binaria en comparación con un modelo que simplemente predice la media.

Los **pseudo R^2** son medidas alternativas que intentan capturar la proporción de variabilidad explicada por el modelo, similar al R^2 en regresión lineal, pero adaptadas a la naturaleza de los datos binarios y la verosimilitud del modelo. Estas medidas son útiles para evaluar y comparar modelos, aunque su interpretación no es tan directa como el R^2 tradicional.

El **McFadden's R^2** es quizás el más utilizado y se define como:

$$R^2_{\text{McFadden}} = 1 - \frac{\log L_{\text{modelo}}}{\log L_{\text{modelo nulo}}}$$

Este pseudo R^2 compara la log-verosimilitud de nuestro modelo con la de un modelo que solo incluye el intercepto. Los valores típicamente oscilan entre 0 y 1, aunque raramente alcanzan valores tan altos como el R^2 en regresión lineal. En contextos aplicados, valores de McFadden entre 0.2 y 0.4 se consideran indicativos de un buen ajuste.

El **Nagelkerke R²** representa una versión normalizada que garantiza que el valor máximo sea 1:

$$R^2_{\text{Nagelkerke}} = \frac{1 - \left(\frac{L_{\text{nulo}}}{L_{\text{modelo}}}\right)^{2/n}}{1 - (L_{\text{nulo}})^{2/n}}$$

Finalmente, el **Cox-Snell R²** se define como:

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

Aunque este último tiene la limitación de que su valor máximo teórico es menor que 1, razón por la cual Nagelkerke propuso su corrección.

Es crucial entender que estos pseudo R^2 no deben interpretarse exactamente como el R^2 tradicional. Mientras que en regresión lineal el R^2 representa la proporción de varianza explicada, en regresión logística estos índices reflejan más bien la mejora en la verosimilitud que aporta nuestro modelo comparado con el modelo nulo. Sin embargo, proporcionan una herramienta valiosa para evaluar y comparar diferentes especificaciones de modelo.

La evaluación integral de la bondad de ajuste en regresión logística requiere considerar tanto la deviance como los pseudo R^2 en conjunto, complementando esta información con pruebas de significancia global del modelo mediante tests de razón de verosimilitudes, que permiten determinar si la inclusión de las variables predictoras mejora significativamente el ajuste comparado con el modelo nulo.

6.5.5 Validación del modelo logístico

Una vez evaluada la bondad de ajuste del modelo logístico, el siguiente paso fundamental es validar su capacidad predictiva y su rendimiento en la clasificación de nuevas observaciones. La validación en regresión logística presenta características particulares debido a la naturaleza categórica de la variable dependiente, lo que requiere métricas y enfoques específicos que van más allá de las medidas tradicionales de error de predicción.

La validación del modelo logístico se centra en dos aspectos complementarios: la **capacidad discriminativa** del modelo (qué tan bien puede distinguir entre las dos clases) y la **precisión de clasificación** (qué proporción de predicciones son correctas). Estas evaluaciones se realizan típicamente mediante la construcción de una **matriz de confusión** y el análisis de curvas **ROC**.

La **matriz de confusión** constituye la herramienta fundamental para evaluar el rendimiento de clasificación en regresión logística. Esta matriz organiza las predicciones del modelo en una tabla de contingencia 2×2 que compara los resultados predichos con los valores reales

observados. Para construir esta matriz, primero debemos convertir las probabilidades estimadas por el modelo en predicciones de clase mediante un umbral de decisión, típicamente 0.5.

La clasificación de cada observación resulta en una de cuatro categorías:

- **Verdaderos Positivos (VP):** Predijo positivo y es positivo.
- **Falsos Positivos (FP):** Predijo positivo pero es negativo.
- **Verdaderos Negativos (VN):** Predijo negativo y es negativo.
- **Falsos Negativos (FN):** Predijo negativo pero es positivo.

A partir de esta clasificación, podemos calcular métricas fundamentales de rendimiento:

- **Precisión (Accuracy):** $\frac{VP+VN}{\text{Total}}$ - representa la proporción total de predicciones correctas
- **Sensibilidad:** $\frac{VP}{VP+FN}$ - mide la capacidad del modelo para identificar correctamente los casos positivos

- **Especificidad:** $\frac{VN}{VN+FP}$ - evalúa la capacidad para identificar correctamente los casos negativos

Es importante reconocer que estas métricas pueden verse influenciadas por el umbral de decisión elegido y por el balance de clases en los datos. Un modelo puede tener alta precisión global pero pobre capacidad para detectar la clase minoritaria, especialmente en datasets desbalanceados.

La **Curva ROC (Receiver Operating Characteristic)** proporciona una evaluación más comprehensiva del rendimiento del modelo al examinar la relación entre la sensibilidad y la especificidad a través de todos los posibles umbrales de decisión. Esta curva grafica la tasa de verdaderos positivos contra la tasa de falsos positivos (1 - especificidad) para cada umbral posible.

Un modelo perfecto produciría una curva ROC que pasaría por la esquina superior izquierda del gráfico (100% sensibilidad, 0% falsos positivos), mientras que un modelo sin capacidad discriminativa produciría una línea diagonal de 45 grados. La **AUC (Área Bajo la Curva ROC)** cuantifica esta capacidad discriminativa en un solo número que varía entre 0.5 (sin capacidad discriminativa) y 1.0 (discriminación perfecta).

La interpretación de la AUC sigue convenciones establecidas:

- **0.9 - 1.0:** Excelente discriminación
- **0.8 - 0.9:** Buena discriminación

- **0.7 - 0.8:** Discriminación aceptable
- **0.6 - 0.7:** Discriminación pobre
- **0.5 - 0.6:** Sin capacidad discriminativa útil

La validación efectiva del modelo logístico requiere considerar tanto las métricas puntuales derivadas de la matriz de confusión como la capacidad discriminativa global capturada por la curva ROC y la AUC. Además, es crucial evaluar estas métricas tanto en los datos de entrenamiento como en conjuntos de validación independientes para detectar posibles problemas de sobreajuste y asegurar que el modelo generalizará adecuadamente a nuevos datos.

Ejemplo

Vamos a aplicar la regresión logística en R utilizando el conjunto de datos `Pima.tr` del paquete `MASS`, que contiene información sobre mujeres pima y si tienen o no diabetes.

```r
# Cargar la librería y el conjunto de datos
library(MASS)
data(Pima.tr)

# Ajustar el modelo de regresión logística
modelo_logistico <- glm(type ~ npreg + glu + bmi, data = Pima.tr, family = binomial)

# Resumen del modelo
summary(modelo_logistico)
```

```
Call:
glm(formula = type ~ npreg + glu + bmi, family = binomial, data = Pima.tr)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.718723   1.411080  -6.179 6.46e-10 ***
npreg        0.149213   0.051833   2.879  0.00399 **
glu          0.033879   0.006327   5.355 8.55e-08 ***
bmi          0.094817   0.032405   2.926  0.00343 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 256.41  on 199  degrees of freedom
Residual deviance: 189.89  on 196  degrees of freedom
AIC: 197.89

Number of Fisher Scoring iterations: 5
```

```r
# Predicciones de probabilidad
predicciones_prob <- predict(modelo_logistico, type = "response")

# Clasificación con un umbral de 0.5
predicciones_clase <- ifelse(predicciones_prob > 0.5, "Yes", "No")

# Crear matriz de confusión
tabla_confusion <- table(Predicted = predicciones_clase, Actual = Pima.tr$type)
print(tabla_confusion)
```

```
         Actual
Predicted  No Yes
      No  114  29
      Yes  18  39
```

```r
# Calcular precisión
accuracy <- sum(diag(tabla_confusion)) / sum(tabla_confusion)
print(paste("Precisión:", round(accuracy, 3)))
```

```
[1] "Precisión: 0.765"
```

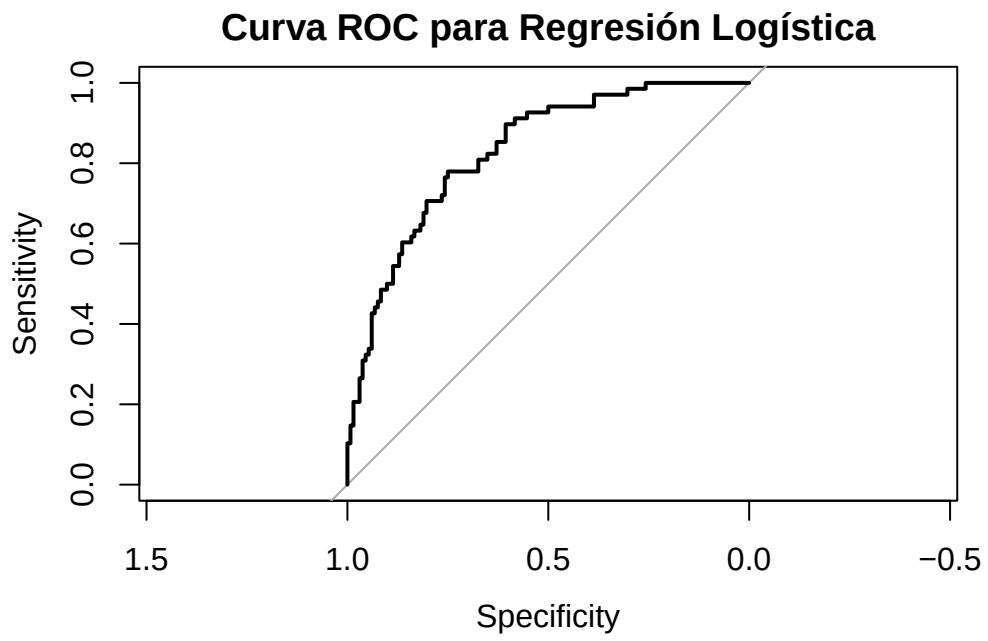
```r
# Cargar librería para curvas ROC
library(pROC)
```

```
Type 'citation("pROC")' for a citation.


Attaching package: 'pROC'

The following objects are masked from 'package:stats':

    cov, smooth, var
```

```r
# Curva ROC
roc_obj <- roc(Pima.tr$type, predicciones_prob)
```

```
Setting levels: control = No, case = Yes

Setting direction: controls < cases
```

```r
plot(roc_obj, main = "Curva ROC para Regresión Logística")
```




```r
# Calcular AUC
auc_valor <- auc(roc_obj)
print(paste("AUC:", round(auc_valor, 3)))
```

```
[1] "AUC: 0.831"
```

6.6 Regresión de Poisson

La **regresión de Poisson** es una técnica estadística utilizada para modelar **datos de conteo**, es decir, situaciones en las que la variable dependiente representa el número de veces que ocurre un evento en un período de tiempo o espacio específico (Coxe, West, and Aiken 2009). Este tipo de modelo es adecuado cuando la variable dependiente toma valores enteros no negativos $(0, 1, 2, ...)$ y sigue una distribución de **Poisson**.

La distribución de Poisson describe la probabilidad de que ocurra un número determinado de eventos en un intervalo fijo, dado que estos eventos ocurren de forma independiente y a una tasa constante.

La **función de probabilidad** de la distribución de Poisson es:

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

Donde:

- Y es la variable aleatoria que representa el número de eventos.
- λ es la **tasa media de ocurrencia** de los eventos (esperanza de Y).
- y es el número de eventos observados ($y = 0, 1, 2, ...$).

6.6.1 Modelo de regresión de Poisson

En la **regresión de Poisson**, el objetivo es modelar la relación entre la **tasa de ocurrencia de los eventos** (λ) y un conjunto de variables predictoras $X_1, X_2, \dots, X_p$.

La **forma funcional** del modelo de Poisson es:

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Donde:

- $\log(\lambda)$ es la **función de enlace logarítmica** que asegura que la tasa λ sea siempre positiva.
- $\beta_0, \beta_1, \dots, \beta_p$ son los coeficientes del modelo que describen la influencia de cada predictor sobre la tasa de eventos.

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

$$\lambda = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p}$$

6.6.2 Supuestos y limitaciones de la regresión de Poisson

Tal y como ocurre en el modelo de regresión lineal, para que la regresión de Poisson sea adecuada, se deben cumplir ciertos **supuestos**:

- **Independencia de los eventos:** Los eventos deben ocurrir de manera independiente unos de otros.
- **Distribución de Poisson de la variable dependiente:** La variable de respuesta debe seguir una distribución de Poisson, donde la **media** y la **varianza** son iguales:

$$E(Y) = Var(Y) = \lambda$$

- **No sobredispersión:** Uno de los problemas comunes en los datos de conteo es la **sobredispersión**, que ocurre cuando la varianza de los datos es mayor que la media ($Var(Y) > E(Y)$). La presencia de sobredispersión indica que el modelo de Poisson puede no ser adecuado, y puede ser necesario considerar modelos alternativos como la **regresión binomial negativa**.
- **No exceso de ceros:** Si hay demasiados ceros en los datos (por ejemplo, en el número de accidentes en diferentes localidades donde muchas tienen cero accidentes), puede ser necesario utilizar modelos de **Poisson inflados en ceros (ZIP)** (Lambert 1992).

6.6.3 Interpretación de los resultados

La interpretación de los coeficientes en la regresión de Poisson difiere de la regresión lineal debido al uso de la función de enlace logarítmica.

Los coeficientes β representan el **logaritmo de la tasa** de eventos asociados con un cambio en la variable independiente. Para interpretar en términos de la tasa de ocurrencia, se utiliza el **exponencial de los coeficientes**:

$$e^{\beta_i}$$

Esto representa el **factor de cambio multiplicativo** en la tasa de eventos por cada unidad adicional en la variable X_i.

! Interpretando el coeficiente de Poisson: Incidence Rate Ratio (IRR)

En la regresión de Poisson, el coeficiente β_j está en la escala logarítmica de la tasa. Para una interpretación práctica, lo exponenciamos para obtener el **Incidence Rate Ratio (IRR)**:

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

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.

- **Si IRR > 1**: Un incremento en X_j se asocia con un **aumento** en la tasa de eventos. Un IRR de 1.25 significa que por cada unidad que aumenta X_j, la tasa de eventos esperada se multiplica por 1.25 (es decir, aumenta un 25%).
- **Si IRR < 1**: Un incremento en X_j se asocia con una **disminución** en la tasa de eventos. Un IRR de 0.80 significa que por cada unidad que aumenta X_j, la tasa de eventos esperada se multiplica por 0.80 (es decir, disminuye un 20%).
- **Si IRR $= 1$**: La variable X_j no tiene efecto sobre la tasa de eventos.

Ejemplo

Vamos a utilizar R para ajustar un modelo de regresión de Poisson. Supongamos que tenemos datos sobre el **número de accidentes de tráfico** en diferentes intersecciones de una ciudad, junto con variables como el volumen de tráfico y la visibilidad.

```
# Simulación de datos para el número de accidentes
set.seed(456)  # Seed diferente para evitar duplicación
n <- 100  # Número de observaciones

# Variables predictoras
trafico_nuevo <- rnorm(n, mean = 1000, sd = 300)  # Volumen de tráfico en vehículos por día
visibilidad_nueva <- rnorm(n, mean = 5, sd = 2)   # Visibilidad en kilómetros

# Generar la tasa de accidentes (lambda) usando un modelo logarítmico
lambda_nuevo <- exp(0.01 * trafico_nuevo - 0.2 * visibilidad_nueva)

# Generar el número de accidentes como una variable de Poisson
accidentes_nuevo <- rpois(n, lambda = lambda_nuevo)

# Crear el data frame
datos_ejemplo <- data.frame(accidentes = accidentes_nuevo, trafico = trafico_nuevo, visibil
head(datos_ejemplo)
```

```
  accidentes   trafico visibilidad
1        140  596.9436    5.236303
2      37153 1186.5327    6.739805
3      92909 1240.2624    4.816128
4        117  583.3323    5.137798
5       1793  785.6929    1.635146
6       1935  902.7817    7.233911
```

```
# Ajustar el modelo de regresión de Poisson
modelo_ejemplo <- glm(accidentes ~ trafico + visibilidad, data = datos_ejemplo, family = po

# Resumen del modelo
summary(modelo_ejemplo)
```

```


Call:
glm(formula = accidentes ~ trafico + visibilidad, family = poisson,
```

```
    data = datos_ejemplo)

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  7350.127   <2e-16 ***
visibilidad -2.000e-01  1.012e-04 -1976.897   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1.6856e+08  on 99  degrees of freedom
Residual deviance: 8.9300e+01  on 97  degrees of freedom
AIC: 1220.4

Number of Fisher Scoring iterations: 3
```

El coeficiente asociado a `trafico` indica cómo el volumen de tráfico afecta la tasa de accidentes.
El coeficiente asociado a `visibilidad` muestra cómo la visibilidad afecta la frecuencia de accidentes.

6.6.4 Estimación por máxima verosimilitud en regresión de Poisson

La estimación de parámetros en regresión de Poisson utiliza también el método de máxima verosimilitud, pero adaptado específicamente para la distribución de Poisson con función de enlace logarítmica. Este enfoque garantiza que las estimaciones aprovechen de manera óptima la información contenida en los datos de conteo.

Para una muestra de n observaciones independientes, donde y_i representa el conteo de eventos para la observación i, la **función de verosimilitud** se define como:

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

donde $\lambda_i = e^{\mathbf{x}_i^T \beta}$ es la tasa esperada para la observación i.

La **log-verosimilitud** correspondiente es:

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \log(\lambda_i) - \lambda_i - \log(y_i!)]$$

Sustituyendo $\lambda_i = e^{\mathbf{x}_i^T \beta}$:

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

Para encontrar los valores de β que maximizan la log-verosimilitud, derivamos respecto a cada parámetro:

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

Esta ecuación indica que los estimadores de máxima verosimilitud se obtienen cuando la suma de los residuos (observado menos esperado) ponderados por cada variable predictora es cero.

6.6.4.1 Implementación del algoritmo IRLS

Dado que las ecuaciones de verosimilitud no tienen solución analítica cerrada, se utiliza el algoritmo IRLS adaptado para regresión de Poisson. Los elementos específicos son:

Pesos:

$$w_i = \lambda_i$$

Variable dependiente ajustada:

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

Actualización de parámetros:

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

6.6.4.2 Propiedades específicas de la estimación Poisson

La estimación por máxima verosimilitud en regresión de Poisson tiene características particulares que la distinguen de otros GLMs:

1. **Equidispersión**: El modelo asume que $E(Y_i) = \text{Var}(Y_i) = \lambda_i$, lo que significa que la varianza aumenta linealmente con la media.
2. **Convergencia**: Generalmente requiere menos iteraciones que la regresión logística debido a la naturaleza más estable de la función de enlace logarítmica.

3. **Estabilidad numérica**: La función de enlace logarítmica garantiza automáticamente que $\lambda_i > 0$, evitando problemas de valores negativos en las tasas estimadas.

4. **Interpretación multiplicativa**: Los coeficientes se interpretan como efectos multiplicativos sobre la tasa, lo que es natural para datos de conteo.

Ejemplo: Estimación paso a paso en regresión de Poisson

```
# Demostración del proceso de estimación por máxima verosimilitud
# Usar los datos simulados anteriores
set.seed(123)
n <- 100
trafico <- rnorm(n, mean = 1000, sd = 300)
visibilidad <- rnorm(n, mean = 5, sd = 2)
lambda <- exp(-2 + 0.001*trafico - 0.2*visibilidad)
accidentes <- rpois(n, lambda)
datos_accidentes <- data.frame(accidentes, trafico, visibilidad)

# Función para calcular log-verosimilitud manualmente
log_likelihood_poisson <- function(beta, X, y) {
  eta <- X %*% beta
  lambda <- exp(eta)
  sum(y * log(lambda) - lambda - lgamma(y + 1))
}

# Preparar datos
X <- model.matrix(accidentes ~ trafico + visibilidad, data = datos_accidentes)
y <- datos_accidentes$accidentes

# Ajuste con glm
modelo_glm_pois <- glm(accidentes ~ trafico + visibilidad,
                       data = datos_accidentes, family = poisson)
beta_glm_pois <- coef(modelo_glm_pois)

cat("=== ESTIMACIÓN POR MÁXIMA VEROSIMILITUD - POISSON ===\n")
```

```
=== ESTIMACIÓN POR MÁXIMA VEROSIMILITUD - POISSON ===
```

```
cat("Coeficientes estimados por glm():\n")
```

```
Coeficientes estimados por glm():
```

```r
print(beta_glm_pois)
```

```
(Intercept)     trafico visibilidad
-5.75588458  0.00284373  0.13082078
```

```r
# Verificar optimización
ll_optimo_pois <- log_likelihood_poisson(beta_glm_pois, X, y)
cat("\nLog-verosimilitud en el óptimo:", round(ll_optimo_pois, 4), "\n")
```

```
Log-verosimilitud en el óptimo: -39.65
```

```r
# Interpretación multiplicativa
cat("\nInterpretación multiplicativa (exp(coeficientes)):\n")
```

```
Interpretación multiplicativa (exp(coeficientes)):
```

```r
exp_coefs <- exp(beta_glm_pois)
print(exp_coefs)
```

```
(Intercept)     trafico visibilidad
0.003164106 1.002847777 1.139763495
```

```r
cat("\nInterpretación:\n")
```

```
Interpretación:
```

```r
cat("- Intercepto: Tasa base =", round(exp_coefs[1], 4), "accidentes\n")
```

```
- Intercepto: Tasa base = 0.0032 accidentes
```

```r
cat("- Tráfico: Por cada vehículo adicional, la tasa se multiplica por", round(exp_coefs[2]
```

```
- Tráfico: Por cada vehículo adicional, la tasa se multiplica por 1.002848
```

```r
cat("- Visibilidad: Por cada km adicional de visibilidad, la tasa se multiplica por", round
```

```
- Visibilidad: Por cada km adicional de visibilidad, la tasa se multiplica por 1.1398
```

```
# Información de convergencia
cat("\nInformación del algoritmo IRLS:\n")
```

```
Información del algoritmo IRLS:
```

```
cat("Iteraciones necesarias:", modelo_glm_pois$iter, "\n")
```

```
Iteraciones necesarias: 6
```

```
cat("¿Convergió?", modelo_glm_pois$converged, "\n")
```

```
¿Convergió? TRUE
```

```
# Verificar supuesto de equidispersión
media_y <- mean(y)
var_y <- var(y)
cat("\nVerificación de equidispersión:\n")
```

```
Verificación de equidispersión:
```

```
cat("Media observada:", round(media_y, 3), "\n")
```

```
Media observada: 0.15
```

```
cat("Varianza observada:", round(var_y, 3), "\n")
```

```
Varianza observada: 0.149
```

```
cat("Razón varianza/media:", round(var_y/media_y, 3), "\n")
```

```
Razón varianza/media: 0.993
```

6.6.5 Bondad de ajuste en la regresión de Poisson

Al igual que en la regresión logística, la bondad de ajuste en los modelos de Poisson se aleja del R^2 tradicional y se centra en medidas basadas en la verosimilitud. El objetivo es cuantificar si el modelo captura adecuadamente la estructura de los datos de conteo.

La **deviance** sigue siendo la métrica fundamental. Para la regresión de Poisson, se calcula comparando la log-verosimilitud del modelo ajustado con la de un modelo saturado (donde $\widehat{\mu}_i = y_i$). La fórmula específica es:

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

Donde el término $y_i \log(y_i)$ se considera cero si $y_i = 0$. Al igual que en otros GLMs, la deviance es clave para comparar modelos anidados mediante el **Test de la Razón de Verosimilitudes (LRT)**.

Sin embargo, para los modelos de Poisson, la prueba de bondad de ajuste más importante en la práctica es la evaluación de la **sobredispersión**. Un buen ajuste del modelo de Poisson implica que se cumple el supuesto de equidispersión ($Var(Y) = \mu$). Por lo tanto, el **estadístico de dispersión** ($\widehat{\phi}$) se convierte en una medida de facto de la bondad de ajuste:

$$\widehat{\phi} = \frac{X^2_{\text{Pearson}}}{n - p}$$

Un valor de $\widehat{\phi}$ cercano a 1 indica que el supuesto de la distribución de Poisson se cumple y que el ajuste del modelo es adecuado. Si $\widehat{\phi}$ es significativamente mayor que 1, el modelo no se ajusta bien a la variabilidad de los datos, y este es el principal indicador para buscar alternativas como la regresión binomial negativa.

Aunque los **pseudo R^2** (como el de McFadden) pueden calcularse, son menos utilizados e informativos en el contexto de la regresión de Poisson en comparación con el análisis de la dispersión.

6.6.6 Validación del modelo de Poisson

A diferencia de la regresión logística, donde la validación se centra en la capacidad de *clasificación*, la validación de un modelo de Poisson se enfoca en su **capacidad de predicción**: ¿qué tan cerca están los conteos predichos por el modelo de los conteos reales observados?

El proceso de validación suele implicar la división de los datos en un conjunto de **entrenamiento (train)** y uno de **prueba (test)**. El modelo se ajusta con los datos de entrenamiento y su rendimiento predictivo se evalúa sobre los datos de prueba, que el modelo no ha visto antes. Esto nos da una estimación honesta de cómo generalizará el modelo a nuevos datos.

Las métricas de validación principales para modelos de conteo son:

- **Raíz del Error Cuadrático Medio (RMSE):** Es una de las métricas más comunes y mide la desviación estándar de los residuos. Penaliza más los errores grandes.

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

- **Error Absoluto Medio (MAE):** Mide la magnitud promedio de los errores, siendo menos sensible a valores atípicos que el RMSE.

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

Ambas métricas se expresan en las mismas unidades que la variable de respuesta (por ejemplo, "accidentes", "compras"), lo que facilita su interpretación. Un modelo con valores de RMSE y MAE más bajos en el conjunto de prueba se considera que tiene un mejor rendimiento predictivo.

Una herramienta visual clave para la validación es el **gráfico de valores predichos vs. valores reales** en el conjunto de prueba. En un modelo con buena capacidad predictiva, los puntos deberían agruparse cerca de la línea diagonal $y = x$, indicando que las predicciones ($\hat{\mu}_i$) son cercanas a los valores observados (y_i).

6.7 Otros GLMs

La **regresión binomial negativa** y los **modelos basados en distribuciones como Gamma e Inversa Gaussiana** amplían la capacidad de los **Modelos Lineales Generalizados (GLM)** para adaptarse a una amplia variedad de situaciones del mundo real. Estos modelos son especialmente útiles cuando los datos presentan características como sobredispersión, sesgo o restricciones en el dominio (por ejemplo, solo valores positivos). La elección adecuada del modelo y la función de enlace garantiza predicciones precisas y válidas, contribuyendo a la toma de decisiones informadas en campos como la salud, la ingeniería y la economía.

6.7.1 Regresión binomial negativa

Tal y como hemos visto en apartados anteriores, la **sobredispersión** ocurre cuando la varianza de los datos de conteo es **mayor que la media**, lo cual viola uno de los supuestos clave de la regresión de Poisson, que asume que la media y la varianza son iguales ($E(Y) = Var(Y) = \lambda$). La sobredispersión puede surgir por varias razones:

- **Heterogeneidad no modelada:** Existen factores que afectan la variable dependiente pero no han sido incluidos en el modelo.

- **Dependencia entre eventos:** Los eventos no ocurren de forma independiente.
- **Exceso de ceros:** Hay más ceros en los datos de los que predice la distribución de Poisson.

Cuando la sobredispersión está presente, la regresión de Poisson subestima los errores estándar, lo que puede llevar a conclusiones incorrectas sobre la significancia de los predictores.

La **regresión binomial negativa** es una extensión de la regresión de Poisson que introduce un parámetro adicional para manejar la sobredispersión. Este modelo permite que la varianza sea mayor que la media:

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

Donde α es el parámetro de dispersión. Si $\alpha = 0$, el modelo se reduce a la regresión de Poisson.

La forma funcional del modelo binomial negativo es similar al de Poisson:

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Pero la varianza ahora incluye el término adicional α para capturar la sobredispersión.

Ejemplo

```r
# Instalar y cargar la librería MASS que contiene la función glm.nb
library(MASS)

# Usar los datos de accidentes del ejemplo anterior
# Ajuste de un modelo binomial negativo
modelo_binom_neg <- glm.nb(accidentes ~ trafico + visibilidad, data = datos_accidentes)
```

```
Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace =
control$trace > : iteration limit reached
Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace =
control$trace > : iteration limit reached
```

```r
# Resumen del modelo
summary(modelo_binom_neg)
```

```
Call:
glm.nb(formula = accidentes ~ trafico + visibilidad, data = datos_accidentes,
```

```
    init.theta = 2158.536301, link = log)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.7560791  1.5223127  -3.781 0.000156 ***
trafico      0.0028439  0.0009831   2.893 0.003818 **
visibilidad  0.1308306  0.1402185   0.933 0.350795
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(2158.536) family taken to be 1)

    Null deviance: 59.679  on 99  degrees of freedom
Residual deviance: 50.680  on 97  degrees of freedom
AIC: 87.301

Number of Fisher Scoring iterations: 1

              Theta:  2159
          Std. Err.:  49427
Warning while fitting theta: iteration limit reached

 2 x log-likelihood:  -79.301
```

```
# Comparar la dispersión con el modelo de Poisson
cat("Dispersión en Poisson (deviance/df.residual):", round(modelo_glm_pois$deviance / model
```

```
Dispersión en Poisson (deviance/df.residual): 0.523
```

```
cat("Parámetro theta en Binomial Negativa:", round(modelo_binom_neg$theta, 3), "\n")
```

```
Parámetro theta en Binomial Negativa: 2158.536
```

```
# Comparar AIC
cat("AIC Poisson:", round(AIC(modelo_glm_pois), 2), "\n")
```

```
AIC Poisson: 85.3
```

```
cat("AIC Binomial Negativa:", round(AIC(modelo_binom_neg), 2), "\n")
```

```
AIC Binomial Negativa: 87.3
```

Interpretación del parámetro θ en binomial negativa:

- El parámetro θ controla el grado de sobredispersión en el modelo
- **Valores altos de** θ (ej: $\theta > 100$): Poca sobredispersión, el modelo se aproxima a Poisson
- **Valores bajos de** θ (ej: $\theta < 10$): Mucha sobredispersión, diferencia significativa respecto a Poisson
- **Regla práctica**: Si θ es pequeño, confirma que había sobredispersión y que el modelo binomial negativa es más apropiado que Poisson

Comparación de modelos: - Si AIC de binomial negativa < AIC de Poisson → preferir binomial negativa - El modelo binomial negativa corrige la subestimación de errores estándar que ocurre en Poisson con sobredispersión

6.7.2 Modelos para variables continuas no normales

Existen situaciones en las que la variable dependiente es **continua**, pero **no sigue una distribución normal**. En estos casos, los **Modelos Lineales Generalizados (GLM)** permiten utilizar distribuciones alternativas como **Gamma** o **Inversa Gaussiana**, junto con funciones de enlace específicas.

6.7.2.1 Regresión gamma para datos positivos y sesgados

La **regresión Gamma** es adecuada para modelar variables continuas que son **positivas** y tienen una distribución **sesgada a la derecha**. Ejemplos típicos incluyen tiempos de espera, costos médicos o duración de procesos.

- La distribución Gamma asume que la variable dependiente es continua y positiva.
- La varianza de la variable dependiente aumenta proporcionalmente al cuadrado de la media.

Función de Enlace Común:

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

Ejemplo

```r
# Simulación de costos médicos
set.seed(123)
n <- 100
ingresos <- rnorm(n, mean = 50000, sd = 10000)
edad <- rnorm(n, mean = 45, sd = 10)
costos <- rgamma(n, shape = 2, rate = 0.00005 * ingresos + 0.01 * edad)

# Ajuste del modelo Gamma
modelo_gamma <- glm(costos ~ ingresos + edad, family = Gamma(link = "log"))

# Resumen del modelo
summary(modelo_gamma)
```

```
Call:
glm(formula = costos ~ ingresos + edad, family = Gamma(link = "log"))

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.440e-01  5.294e-01   0.650   0.5173
ingresos    -1.807e-05  7.804e-06  -2.316   0.0227 *
edad         3.584e-03  7.366e-03   0.487   0.6277
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.5010938)

    Null deviance: 60.771  on 99  degrees of freedom
Residual deviance: 58.345  on 97  degrees of freedom
AIC: 105.47

Number of Fisher Scoring iterations: 5
```

- Los coeficientes muestran cómo los ingresos y la edad afectan los costos médicos esperados.
- El enlace logarítmico asegura que las predicciones sean siempre positivas.

6.7.2.2 Regresión inversa gaussiana

La **regresión Inversa Gaussiana** es útil para modelar tiempos de respuesta o variables donde la varianza disminuye rápidamente a medida que la media aumenta. Este modelo se aplica en campos como la ingeniería, donde se analizan tiempos hasta fallas de sistemas.

Función de Enlace Común:

$$\frac{1}{\mu^2} = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Ejemplo

```
# Instalar y cargar la librería correcta
library(statmod)

# Simulación de datos
set.seed(123)
n <- 100
carga_trabajo <- rnorm(n, mean = 50, sd = 10)

# Generar tiempos hasta el fallo usando la distribución inversa gaussiana
# Aseguramos que los valores de carga_trabajo sean positivos para evitar problemas numéric
carga_trabajo[carga_trabajo <= 0] <- 1
tiempo_fallo <- rinvgauss(n, mean = 100 / carga_trabajo, dispersion = 1)

# Ajuste del modelo Inversa Gaussiana con enlace logarítmico
modelo_inversa_gauss <- glm(tiempo_fallo ~ carga_trabajo, family = inverse.gaussian(link =
```

```
Warning in sqrt(eta): NaNs produced

Warning: step size truncated due to divergence
```

```
# Resumen del modelo
summary(modelo_inversa_gauss)
```

```


Call:
glm(formula = tiempo_fallo ~ carga_trabajo, family = inverse.gaussian(link = "1/mu^2"),
    start = c(0.01, 0.01))

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)   -0.106143   0.462230  -0.230    0.819
carga_trabajo  0.008261   0.009623   0.859    0.393

(Dispersion parameter for inverse.gaussian family taken to be 1.348442)

    Null deviance: 94.085  on 99  degrees of freedom
Residual deviance: 93.171  on 98  degrees of freedom
AIC: 294.2

Number of Fisher Scoring iterations: 5
```

⚠ ¿Qué GLM debo usar?

La elección del modelo correcto depende casi exclusivamente de la naturaleza de tu variable respuesta (Y). Aquí tienes una guía rápida:

- **¿Tu variable respuesta es binaria (Sí/No, 0/1, Éxito/Fracaso)?**
  - Usa **Regresión Logística**.
- **¿Tu variable respuesta es un conteo de eventos (nº de accidentes, nº de clientes, nº de fallos)?**
  - Empieza con una **Regresión de Poisson**.
  - **Importante:** Después de ajustar el modelo, comprueba si hay **sobredispersión**.
  - Si la hay (estadístico de dispersión $\hat{\phi} > 1.5$ o la teoría lo sugiere), cambia a una **Regresión Binomial Negativa**.
- **¿Tu variable respuesta es continua y estrictamente positiva, con una distribución asimétrica hacia la derecha (ej. tiempos, costos, reclamaciones de seguros)?**
  - Usa **Regresión Gamma**. Es una excelente alternativa a transformar la variable con logaritmos y usar un modelo lineal.
- **¿Tu variable respuesta es un tiempo hasta un evento y tiene una asimetría muy pronunciada?**
  - Considera una **Regresión Inversa Gaussiana**.

7 Conclusiones

A lo largo de seis capítulos hemos presentado el material de la asignatura de **Modelos Estadísticos para la Predicción** del Grado en **Matemáticas**.

En este capítulo final, te invitamos a reflexionar sobre las principales lecciones aprendidas durante el curso y a destacar cómo el rigor matemático que caracteriza vuestra formación se convierte en una herramienta indispensable para dominar el modelado estadístico. Además, animamos a los estudiantes a continuar explorando y ampliando sus conocimientos en cursos posteriores, consolidando así una base sólida para afrontar los desafíos de la ciencia de datos moderna desde una perspectiva analítica y fundamentada.

Al finalizar este recorrido, queda patente que el modelado estadístico es mucho más que una colección de técnicas; es un marco de pensamiento estructurado para comprender la incertidumbre y extraer conocimiento a partir de los datos. Hemos transitado desde los axiomas teóricos de la regresión hasta su aplicación computacional, equipando a los futuros matemáticos con las herramientas para construir, validar e interpretar modelos robustos.

7.1 Resumen de los aprendizajes

A lo largo de este manual, hemos construido un conocimiento progresivo sobre el modelado predictivo, cubriendo los siguientes pilares:

1. **Fundamentos del modelado lineal: simple y múltiple**: Hemos partido de la formulación teórica del **modelo lineal**, estableciendo sus componentes axiomáticos y los supuestos de Gauss-Markov que garantizan las propiedades óptimas de los estimadores de **Mínimos Cuadrados Ordinarios (MCO)**. Se ha hecho hincapié en la transición del modelo simple al múltiple, destacando el principio de *ceteris paribus* para la interpretación de coeficientes, el diagnóstico de la **multicolinealidad** mediante el VIF y la evaluación de la bondad de ajuste a través de la descomposición ANOVA y el R^2 ajustado.
2. **Ingeniería de características y flexibilidad del modelo**: Exploramos cómo superar las limitaciones de un modelo estrictamente lineal. Aprendimos a diagnosticar y corregir violaciones de los supuestos mediante **transformaciones de variables** (logarítmica, Box-Cox), a incorporar predictores no numéricos a través de la codificación de **variables categóricas**, y, fundamentalmente, a modelar relaciones complejas mediante la inclusión

de **términos de interacción**, entendiendo cómo el efecto de un predictor puede depender del valor de otro.

3. **Selección de variables, regularización y validación**: Abordamos el crucial dilema sesgo-varianza y la necesidad de construir modelos parsimoniosos que generalicen bien a datos no observados. Se presentaron los **criterios de información (AIC, BIC)** y los **métodos por pasos (stepwise)** como herramientas para comparar y seleccionar modelos. Profundizamos en los métodos de **regularización (Ridge, Lasso y Elastic Net)**, que ofrecen una alternativa moderna y robusta para manejar la multicolinealidad y realizar selección de variables de forma simultánea, especialmente en contextos de alta dimensionalidad. Finalmente, se consolidó la importancia de la **validación cruzada** como el estándar para una evaluación honesta del rendimiento predictivo del modelo.
4. **Modelos lineales generalizados (GLM)**: Expandimos el marco de la regresión más allá de la asunción de normalidad en la variable respuesta. A través de la introducción de la **familia exponencial** de distribuciones y las **funciones de enlace**, entendimos cómo adaptar el modelo lineal a diferentes tipos de datos. Nos centramos en dos de los GLM más importantes: la **Regresión Logística** para modelar resultados binarios y la **Regresión de Poisson** para datos de conteo, aprendiendo a interpretar sus coeficientes en términos de odds ratios y tasas de eventos, respectivamente.

7.2 Reflexiones finales

El estudio de los **Modelos Estadísticos para la Predicción** dota al matemático de un puente entre la teoría abstracta y la resolución de problemas del mundo real. A lo largo de este curso, hemos visto cómo conceptos rigurosos —espacios vectoriales en la geometría de MCO, optimización en la estimación de parámetros, y teoría de la probabilidad en la inferencia— se materializan en herramientas prácticas para la toma de decisiones bajo incertidumbre.

Hemos aprendido que construir un modelo no es un acto mecánico, sino un proceso iterativo de **diagnóstico, crítica y refinamiento**. La capacidad para evaluar la validez de los supuestos, interpretar los resultados con cautela y comunicar tanto las fortalezas como las limitaciones de un modelo es lo que distingue a un analista competente. La **interpretabilidad** y la **validación rigurosa** no son meros pasos finales, sino el núcleo de una práctica estadística honesta y efectiva.

En un mundo saturado de datos, la habilidad para construir modelos que no solo predicen, sino que también explican y ofrecen certidumbre cuantificable, es más valiosa que nunca.

7.3 Proyección futura: El valor del rigor matemático

Las competencias adquiridas en esta asignatura son la culminación de vuestra carrera, el punto donde el álgebra lineal, el cálculo y la optimización se convierten en el motor de la modelización estadística aplicada. Vuestra formación matemática os proporciona una ventaja fundamental: la capacidad de ir más allá de la aplicación mecánica de un algoritmo para comprender en profundidad los supuestos que lo sustentan, la geometría de su funcionamiento y la incertidumbre inherente a sus conclusiones.

Conceptos como la regularización y la validación cruzada son el lenguaje compartido con el **Aprendizaje Automático**. Mientras que el Aprendizaje Automático a menudo se centra en la potencia predictiva de algoritmos complejos, este curso os ha proporcionado la "gramática" estadística para construir modelos interpretables, diagnosticar su validez y cuantificar la fiabilidad de sus resultados. Esta base teórica es indispensable para aplicar, y en un futuro desarrollar, cualquier técnica de modelado de forma rigurosa y responsable.

Esta habilidad para analizar críticamente los modelos es precisamente lo que el mercado y el mundo académico demandan. Os posiciona de manera ideal para roles avanzados como **Científico de Datos**, **Analista Cuantitativo ('Quant')** en el sector financiero, o **Bioestadístico**, así como para continuar vuestra formación con **estudios de postgrado (Máster o Doctorado)** donde la investigación y el desarrollo de nuevos métodos es primordial.

En definitiva, habéis adquirido un conjunto de herramientas analíticas que os permitirá traducir problemas complejos en modelos manejables y basados en evidencia.

¡Mucha suerte en vuestra trayectoria profesional!

Bibliografía

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using Lme4." *Journal of Statistical Software* 67 (1): 1–48.

Box, George EP, and David R Cox. 1964. "An Analysis of Transformations." *Journal of the Royal Statistical Society: Series B (Methodological)* 26 (2): 211–43.

Carroll, Raymond J, and David Ruppert. 1988. "Transformation and Weighting in Regression." *Monographs on Statistics and Applied Probability*.

Coxe, Stefany, Stephen G West, and Leona S Aiken. 2009. "The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives." *Journal of Personality Assessment* 91 (2): 121–36.

Draper, NR. 1998. *Applied Regression Analysis*. McGraw-Hill. Inc.

Fox, John, and Sanford Weisberg. 2018. *An r Companion to Applied Regression*. Sage publications.

Galton, Francis. 1886. "Regression Towards Mediocrity in Hereditary Stature." *The Journal of the Anthropological Institute of Great Britain and Ireland* 15: 246–63.

Harrell, Frank E., Jr. 2015. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Second. Springer.

Hastie, Trevor, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Vol. 2. Springer.

Hosmer Jr, David W, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied Logistic Regression*. John Wiley & Sons.

Jaccard, James, and Robert Turrisi. 2003. *Interaction Effects in Multiple Regression*. Sage.

James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. 2013. *An Introduction to Statistical Learning*. Vol. 112. Springer.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning with Applications in r*. Second. Springer.

Kuhn, Max, and Kjell Johnson. 2019. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press.

Kutner, Michael H, Christopher J Nachtsheim, John Neter, and William Li. 2005. *Applied Linear Statistical Models*. McGraw-hill.

Lambert, Diane. 1992. "Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing." *Technometrics* 34 (1): 1–14.

Marquardt, Donald W, and Ronald D Snee. 1975. "Ridge Regression in Practice." *The American Statistician* 29 (1): 3–20.

Nelder, John Ashworth, and Robert WM Wedderburn. 1972. "Generalized Linear Models." *Journal of the Royal Statistical Society Series A: Statistics in Society* 135 (3): 370–84.

Pinheiro, José C., and Douglas M. Bates. 2000. *Mixed-Effects Models in s and s-PLUS*. New York: Springer.

Potdar, Kedar, Taher S Pardawala, and Chinmay D Pai. 2017. "A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers." *International Journal of Computer Applications* 175 (4): 7–9.

Ranstam, Jonas, and Jonathan A Cook. 2018. "LASSO Regression." *Journal of British Surgery* 105 (10): 1348–48.

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25 (3): 289–310.

Weisberg, S. 2005. "Applied Linear Regression." Wiley.

Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with r*. Second. Chapman; Hall/CRC.

Yeo, In-Kwon, and Richard A Johnson. 2000. "A New Family of Power Transformations to Improve Normality or Symmetry." *Biometrika* 87 (4): 954–59.

Zheng, Alice, and Amanda Casari. 2018. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 67 (2): 301–20.