Introducción a los Modelos Estadísticos para la Predicción

Víctor Aceña - Isaac Martín

DSLAB

2025-09-10

DATA SCIENCE LABORATORY

Universidad Rey Juan Carlos

El Modelado de Regresión: Una Herramienta Universal

El modelado de regresión constituye una de las herramientas más potentes y flexibles del arsenal estadístico. Su aplicabilidad abarca un espectro extraordinariamente amplio de disciplinas:

- **Ciencias Físicas**: Física de partículas, ingeniería aeroespacial para modelar sistemas complejos.
- **Ciencias Sociales**: Econometría, psicometría para entender comportamientos.
- **Ciencias de la Salud**: Epidemiología para identificar factores de riesgo.
- **Finanzas**: Para entender mercados y comportamientos económicos.

El Propósito de Este Curso

Este curso tiene como misión construir el **andamiaje conceptual y filosófico** sobre el que se asienta el modelado estadístico moderno.

Objetivos:

- **Contextualizar** la regresión no solo como una técnica, sino como un **marco de pensamiento** indispensable.
- **Explorar en profundidad** el propósito dual de la regresión (predicción vs. inferencia).
- **Desgranar** los componentes axiomáticos hasta el último detalle.
- **Ofrecer una visión panorámica** de la vasta familia de modelos de regresión.

El objetivo es prepararte, **con solidez y sin prisas**, para las inmersiones técnicas posteriores.

Predecir vs. Explicar

El modelado de regresión ofrece un marco para investigar y cuantificar las relaciones entre variables.

Conceptualmente, el modelado estadístico se orienta hacia uno de dos polos (Shmueli, 2010):

- **Predicción**: Enfoque orientado a la **precisión** en la estimación de valores futuros o no observados.
- **Explicación/Inferencia**: Enfoque orientado a la **comprensión** de las relaciones entre las variables.

Ambos paradigmas tienen objetivos, métodos y criterios de evaluación distintos.

Predicción: El Paradigma de la Precisión

Objetivo Principal: La **precisión**. Se busca construir un modelo que pueda estimar con el menor error posible el valor de una variable de interés (la *respuesta*).

Características Clave:

- El modelo puede ser tratado como una **“caja negra”** (*black box*).
- El funcionamiento interno o la interpretabilidad son secundarios.
- Lo importante es que las predicciones sean **consistentemente fiables y robustas** en datos no observados previamente.

Inferencia: El Paradigma de la Comprensión

Objetivo Principal: La **comprensión** e **interpretación**. No solo predecir, sino dilucidar la naturaleza de las interdependencias.

Características Clave:

- Se busca **cuantificar** cómo un cambio en una variable predictora influye en la respuesta.
- La **interpretabilidad del modelo es primordial**.
- El interés reside en la **magnitud, signo e incertidumbre estadística** de los parámetros (errores estándar, intervalos de confianza, p-valores).

Predecir vs. Explicar: Ejemplos

Ejemplo de Predicción
Una entidad financiera quiere predecir la probabilidad de que un cliente incurra en impago. Su principal interés es tener un modelo que clasifique correctamente a los futuros solicitantes como de alto o bajo riesgo para minimizar pérdidas.

Ejemplo de Inferencia
Una epidemióloga investiga los factores de riesgo de una enfermedad cardíaca. Su objetivo es entender y cuantificar la relación: *¿En cuántos mmHg aumenta la presión arterial, en promedio, por cada gramo adicional de sal consumido al día?*

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, 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.

El Primer Paso: Convertirse en un Modelador Eficaz

Comprender la distinción entre predicción e inferencia es fundamental para el desarrollo como estadístico.

- **En la Práctica**: Ambos objetivos a menudo se entrelazan y se complementan mutuamente.
- **Decisión Estratégica**: La elección del enfoque determina el tipo de modelo, las métricas de evaluación y la interpretación de resultados.
- **Contexto del Problema**: El dominio de aplicación y las preguntas de investigación guían esta decisión fundamental.

Anatomía de un Modelo de Regresión

Todo modelo de regresión se construye sobre tres pilares fundamentales (Kutner et al., 2005):

1. La variable de respuesta
2. Las variables predictoras
3. El término de error aleatorio

Estos componentes son los ladrillos con los que edificaremos todo nuestro conocimiento.

La Variable de Respuesta (Dependiente)

Representa el fenómeno cuyo comportamiento se busca modelar, comprender o predecir. Su naturaleza determina el tipo de modelo a elegir.

- **Continua**: Cualquier valor en un rango (temperatura, precio).
- **Discreta de Conteo**: Número de eventos (nº de accidentes, nº de clientes).
- **Binaria o Dicotómica**: Dos resultados posibles (éxito/fracaso, enfermo/sano).
- **Categórica**: Grupos o categorías.
    - **Nominal** (sin orden): tipo de sangre, partido político.
    - **Ordinal** (con orden): nivel de satisfacción "bajo/medio/alto".

La Variable de Respuesta: Ejemplos Detallados

La naturaleza de la variable de respuesta es el factor más determinante para elegir el tipo de modelo:

- **Continuas**: Temperatura ambiente, altura de una persona, precio de una acción, concentración de un compuesto químico.
- **Discretas de Conteo**: Número de accidentes en una intersección, número de clientes que entran en una tienda, número de mutaciones en un gen.
- **Binarias**: Éxito/fracaso en un tratamiento, enfermo/sano, compra/no compra, correo spam/no spam.
- **Categóricas**:
  - **Nominales**: Tipo de sangre (A, B, AB, O), partido político preferido.
  - **Ordinales**: Estadio de una enfermedad (I/II/III/IV), nivel educativo.

Las Variables Predictoras

También llamadas independientes, explicativas, regresoras, covariables o *features*.

Son las magnitudes, atributos o factores que se postula que influyen o están asociados con el comportamiento de la variable de respuesta.

- Pueden ser de diversa naturaleza (continuas, categóricas, etc.).
- Su selección es una fase crítica del modelado que requiere:
  - Conocimiento del dominio.
  - Análisis exploratorio de datos.
  - Técnicas estadísticas formales.

La Selección de Variables: Una Fase Crítica

La selección de variables predictoras es una de las fases más críticas del modelado estadístico:

Requiere una Combinación de:

- **Conocimiento del Dominio**: Comprensión profunda del fenómeno que se está modelando.
- **Análisis Exploratorio**: Visualización y exploración inicial de los datos para identificar patrones.
- **Técnicas Estadísticas Formales**: Métodos como selección hacia adelante, hacia atrás, o criterios de información.

Consideraciones:

- Las variables pueden ser de diversa naturaleza (continuas, categóricas, etc.).
- No todas las variables disponibles deben incluirse en el modelo.

El Término de Error Aleatorio (ϵ)

Este componente, a menudo subestimado, es conceptualmente crucial. Simboliza la variabilidad de la respuesta **no capturada** por los predictores.

No es un simple "error" en el sentido de equivocación, sino un componente estocástico que amalgama:

- **Variables Omitidas**: Factores que influyen en Y pero no han sido medidos o incluidos.
- **Error de Medición**: Imprecisiones en la medición de las variables.
- **Aleatoriedad Intrínseca**: Variabilidad irreducible inherente a muchos fenómenos.

Los Supuestos sobre el Error Aleatorio

El término de error ϵ es la clave del diagnóstico en regresión. Gran parte de la inferencia se basa en verificar los supuestos sobre su distribución:

Supuestos Fundamentales:

- **Media cero**: $E[\epsilon] = 0$
- **Varianza constante** (homocedasticidad): $Var[\epsilon] = \sigma^2$
- **Independencia**: Los errores no están correlacionados
- **Normalidad**: $\epsilon \sim N(0, \sigma^2)$ (para inferencia exacta)

Dos individuos con idénticos valores en las variables predictoras pueden tener valores distintos en la respuesta debido a este componente irreducible.

La Ecuación Fundamental de la Regresión

La relación se expresa como la descomposición de la variable de respuesta en una parte sistemática y una parte aleatoria:

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

- $f(\cdot)$ es la **componente sistemática**, que representa el valor esperado de Y para unos valores dados de las X. Es lo que intentamos estimar.
- ϵ es la **componente aleatoria**. El diagnóstico en regresión se basa en verificar los supuestos sobre su distribución.

Linealidad en los Parámetros

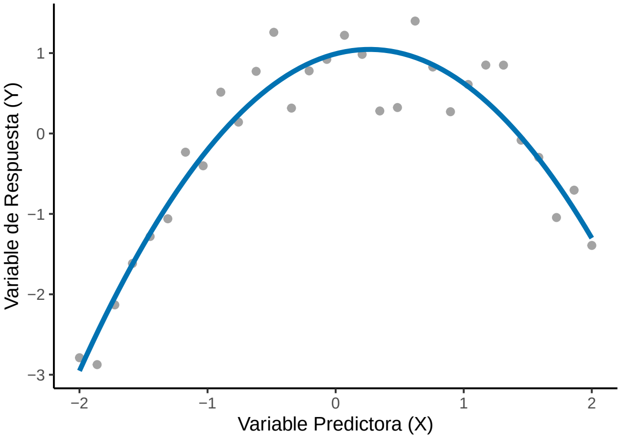
Una característica clave de los modelos de regresión lineal es que son **lineales en los parámetros (β_j)**, no necesariamente en las variables.

Este modelo es **lineal**, aunque la relación con las variables no lo sea:

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

La función f es una combinación lineal de los coeficientes β. Esta flexibilidad es una de las razones de la enorme potencia de los modelos lineales.

Ejemplo: Modelo Lineal con Término Cuadrático



$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

¿Es "lineal" este modelo?

- **SÍ**: Lineal en los parámetros $\beta_0, \beta_1, \beta_2$
- **NO**: No lineal en la variable X

La "linealidad" se refiere a los coeficientes, no a la forma de la curva.

Una parábola que encaja perfectamente en el marco de regresión lineal clásica.

Universo de los Modelos de Regresión

La regresión lineal clásica es el punto de partida para una gama prolífica de metodologías avanzadas:

- Modelos Lineales (LMs)
- Modelos Lineales Generalizados (GLMs)
- Modelos de Efectos Mixtos (Mixed Models)
- Modelos Aditivos Generalizados (GAMs)

Modelos Lineales (LMs)

Constituyen el paradigma fundamental. Es el laboratorio donde se forjan los conceptos esenciales:

- **Estimar parámetros** e interpretar su significado.
- **Cuantificar la incertidumbre** (errores estándar, intervalos de confianza).
- **Realizar contrastes de hipótesis** para evaluar la significancia estadística.
- **Diagnosticar la "salud" de un modelo** examinando sus supuestos.

Asumen que la variable de respuesta sigue una distribución Normal.

Modelos Lineales: La Base Unificadora

Los Modelos Lineales no son solo una técnica más, sino el fundamento que unifica toda la estadística clásica:

- **ANOVA (Análisis de la Varianza)**: Es un caso particular de los LMs cuando todas las variables predictoras son categóricas.
- **ANCOVA (Análisis de la Covarianza)**: Combina variables categóricas (factores) con variables continuas (covariables).
- **Unificación Histórica**: Técnicas que se estudiaban por separado ahora se entienden como manifestaciones del mismo principio matemático.

Esta perspectiva unificadora revolucionó la enseñanza y comprensión de la estadística.

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 contextos mucho más amplios.

Representan un salto conceptual que expande masivamente el universo de problemas que podemos abordar (Nelder & Wedderburn, 1972).

Permiten escapar de la "tiranía" de la distribución Normal para modelar respuestas con otras naturalezas y escalas.

Se logra mediante dos mecanismos:

1. La **familia exponencial de distribuciones**.
2. La **función de enlace (link function)**.

GLMs: Una Revolución Conceptual

Los GLMs representan uno de los avances más significativos en la estadística del siglo XX:

- **Unificación**: Por primera vez, se unificaron bajo un mismo marco conceptual diversas clases de modelos que antes se trataban por separado.
- **Flexibilidad**: Permitieron abordar una gama masivamente amplia de problemas que antes requerían técnicas especializadas.
- **Impacto**: Estimularon enormemente el desarrollo de software estadístico y la aplicación del modelado a nuevos dominios.

Gracias a los GLMs, podemos usar el mismo marco conceptual para modelar desde la cantidad de ciclistas en una ciudad (Poisson) hasta la probabilidad de respuesta a un tratamiento (logística).

GLMs: La Familia Exponencial de Distribuciones

Los GLMs funcionan con distribuciones que pertenecen a la **familia exponencial**, un “club” con propiedades matemáticas convenientes que permiten una teoría unificada.

- **Miembros Notables**: Normal, Poisson (conteo), Binomial (proporciones/binarios), Gamma (positivos asimétricos), Binomial Negativa.
- **Estructura Común**: Su forma matemática compartida es la clave que permite unificar la estimación de parámetros para todos estos modelos.

GLMs: La Función de Enlace - El “Traductor”

El verdadero golpe de genialidad: La función de enlace $g(\cdot)$ actúa como un **“traductor”** o **“puente”** entre dos mundos:

- **El predictor lineal** $X\beta$: Puede tomar cualquier valor real ($-\infty$ a $+\infty$)
- **La media de la respuesta** $\mu = E[Y]$: A menudo está **restringida**

La Relación Fundamental: $g(E[Y]) = g(\mu) = X\beta$

Ejemplos Clave:

- **Enlace Logarítmico** (Poisson): $g(\mu) = \log(\mu) \to \mu = \exp(X\beta)$ (siempre positivo)
- **Enlace Logit** (Binomial): $g(\mu) = \log(\frac{\mu}{1-\mu})$ (proyecta probabilidades al rango completo)

Modelos de Efectos Mixtos

Su desarrollo responde a la **necesidad crítica** de analizar datos que exhiben estructuras de dependencia o correlación.

Violación del Supuesto de Independencia:

- **Medidas repetidas**: Medir la presión arterial de un paciente cada mes.
- **Datos longitudinales**: Un tipo especial de medida repetida a lo largo del tiempo.
- **Datos agrupados**: Estudiantes anidados dentro de clases, clases dentro de colegios.

La Solución: Introducir **efectos aleatorios** para capturar la variabilidad específica entre grupos/individuos, además de los **efectos fijos** que representan a la población general.

GAMs: Flexibilidad sin Perder Interpretabilidad

Los **Modelos Aditivos Generalizados** representan una extensión natural y altamente flexible de los GLMs.

Innovación Clave: Relajan el supuesto de linealidad entre el predictor transformado y las covariables.

Metodología:

- Modelan relaciones mediante **funciones suaves** no paramétricas (*splines*).
- Mantienen la estructura aditiva: $g(\mu) = \alpha + f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p)$
- Las funciones $f_i(\cdot)$ se **estiman a partir de los datos**.

Ventaja: Capturan patrones no lineales complejos sin especificar una forma funcional a priori, logrando un equilibrio excepcional entre **flexibilidad** e **interpretabilidad**.

R y el Ecosistema de Paquetes

Este curso fusiona la teoría con la aplicación computacional directa a través de **R**.

¿Por qué R?

- **Estándar de facto** en la investigación estadística y ciencia de datos académica.
- **Potencia y flexibilidad** incomparables.
- **Inmenso ecosistema** de paquetes contribuidos por la comunidad científica.

Capacidades Fundamentales: Exploración de datos, estimación de parámetros, diagnóstico riguroso, producción de gráficos de alta calidad.

Paquetes Especializados para el Modelado

Funciones Base (paquete `stats`):

- `lm()` para regresión lineal
- `glm()` para modelos lineales generalizados
- Se cargan automáticamente (no requieren instalación)

Paquetes Especializados:

- **`mgcv`**: Implementación de referencia para GAMs (Simon Wood)
- **`lme4` y `nlme`**: Modelos de efectos mixtos
- **`rms`**: Estrategias robustas de modelado
- **`gamair`**: Conjuntos de datos para practicar con GAMs

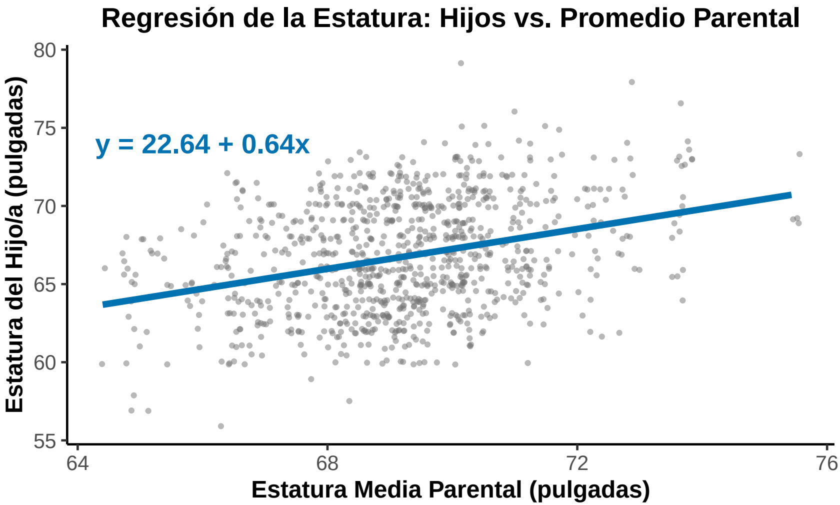
Breve Crónica de la Regresión: Orígenes

La gestación de la regresión se traza hasta Sir **Francis Galton** (S. XIX).

Estudiando la herencia de la estatura, notó que los padres muy altos tendían a tener hijos que, en promedio, no eran tan altos como ellos (y viceversa).

Acuñó el término **“regresión a la mediocridad”** (hoy “regresión a la media”) para describir esta tendencia de las características a “regresar” hacia la media de la población.

Los Estudios de Galton: Los Datos Históricos



Los Estudios de Galton: Hallazgos e Importancia

Datos y Hallazgos Principales

- Galton recopiló datos de **928 hijos** y sus padres.
- Observó una **relación lineal**: padres altos tenían hijos altos (y viceversa).
- **"Regresión a la mediocridad"**: Las estaturas extremas de los padres no se perpetuaban completamente.
- Los hijos tendían a "regresar" hacia la **media poblacional**.

Importancia Histórica

- **Regresión Lineal**: Introdujo el concepto de la **recta de regresión** para modelar relaciones.
- **Correlación**: Precursor del **coeficiente de correlación** (desarrollado por Karl Pearson).
- **Terminología**: Acuñó el término **"regresión"** que usamos hoy.
- **Método Estadístico**: Fundó las bases del análisis de regresión moderno.

Breve Crónica: La Formalización Matemática

Aunque Galton sentó las bases conceptuales, la formalización matemática se debe a dos gigantes:

Adrien-Marie Legendre

- En 1805 publicó el **"Método de los mínimos cuadrados"**.
- Lo concibió como un procedimiento numérico para ajustar observaciones astronómicas.

Carl Friedrich Gauss

- Desarrolló el método de forma independiente.
- Lo dotó de una profunda base teórica, conectándolo con la **teoría de la probabilidad**.
- Lo derivó bajo el supuesto de **errores normales**, convirtiéndolo en la técnica fundamental que es hoy.

Breve Crónica: El Desarrollo Moderno

El siglo XX fue testigo de un desarrollo explosivo, con dos hitos clave:

La Revolución de los GLMs (1972)

- **John Nelder** y **Robert Wedderburn** publicaron su trabajo sobre Modelos Lineales Generalizados.
- Unificaron la regresión lineal, logística y de Poisson bajo un mismo marco conceptual y computacional.
- Esto estimuló enormemente la aplicación del modelado a una nueva y vasta gama de problemas.

La Evolución Contemporánea

- El legado continúa evolucionando a un ritmo vertiginoso.
- Inclusión de modelos jerárquicos y bayesianos.
- Integración con métodos de *machine learning* (ej: árboles de regresión).
- Adaptación al análisis de datos masivos (*big data*).

De la Herencia Biológica al Big Data

La regresión ha evolucionado de manera extraordinaria desde sus orígenes:

- **Origen Modesto**: Una simple observación sobre la herencia de la estatura por Sir Francis Galton.
- **Desarrollo Matemático**: La formalización rigurosa por Legendre y Gauss en el siglo XIX.
- **Revolución Conceptual**: Los GLMs unificaron múltiples técnicas bajo un marco común.
- **Era Contemporánea**: Adaptación a *machine learning*, métodos bayesianos y *big data*.

La regresión se ha convertido en una de las herramientas más versátiles y poderosas del arsenal analítico moderno, presente en prácticamente todas las disciplinas cuantitativas.

Preparándonos para el Viaje Técnico

Lo que hemos construido:

- **Marco conceptual sólido**: Predicción vs. inferencia
- **Fundamentos axiomáticos**: Los tres pilares de la regresión
- **Perspectiva histórica**: De Galton al *big data*
- **Visión panorámica**: El universo de modelos disponibles

Lo que sigue:

Con este **andamiaje conceptual y filosófico**, estamos preparados para las inmersiones técnicas que seguirán. Cada concepto avanzado se construirá sobre estos cimientos sólidos.

Referencias

- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill/Irwin.
- Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3), 370-384.
- Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289-310.