

Métodos Estadísticos de Predicción

Guía de la
Asignatura

Grado en Matemáticas

AUTORES

- Víctor Aceña Gil
- Isaac Martín de Diego

2025-2026



Material de apoyo para el estudio de Modelos Estadísticos de Predicción.

Índice de Contenidos

Introducción a la asignatura	3
Objetivos de la asignatura	3
Temario de la asignatura	4
Desarrollo de la asignatura	5
Semana 1: Presentación e Introducción a los Modelos de Regresión	5
Semana 2: Regresión Lineal Simple (RLS) - Fundamentos y Estimación	6
Semana 3: RLS - Inferencia y Diagnóstico	6
Semana 4: Regresión Lineal Múltiple (RLM) - Formulación e Interpretación	7
Semana 5: Ingeniería de Características I - Transformaciones	7
Semana 6: Ingeniería de Características II - Interacciones	8
Semana 7: Consolidación y Práctica de Regresión Lineal	9
Semana 8: Evaluación Parcial e Introducción a la Selección de Variables	9
Semana 9: Métodos de Selección de Variables	10
Semana 10: Regularización y Validación de Modelos	10
Semana 11: Modelos Lineales Generalizados (GLM) - Fundamentos	11
Semana 12: GLM - Regresión Logística	11
Semana 13: GLM - Regresión de Poisson	12
Semana 14: Tópicos Avanzados y Repaso	12
Semana 15: Presentación de Proyectos Finales	13
Bibliografía	14
Bibliografía Principal y Recomendada	14
Referencias Adicionales y Tópicos Relacionados	15

Introducción a la asignatura

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**.

El modelado estadístico tiene un propósito dual que exploraremos en profundidad: la **predicción** (la búsqueda de la máxima precisión) y la **inferencia** (la búsqueda de la explicación y la comprensión). Entender esta dicotomía es clave para seleccionar y evaluar los modelos de manera efectiva.

Comenzaremos sentando las bases de la regresión lineal, desde su formulación más simple hasta la complejidad del modelo múltiple. A partir de ahí, aprenderemos a enriquecer nuestros modelos mediante la **ingeniería de características**, a seleccionar las variables más relevantes con técnicas de **selección y regularización**, y finalmente, a extender nuestro alcance más allá de la normalidad con los **Modelos Lineales Generalizados (GLM)**.

Objetivos de la asignatura

- Aprender a formular, estimar e interpretar modelos de regresión, comprendiendo los supuestos teóricos que los sustentan.
- Dominar el diagnóstico de los modelos para verificar su validez y detectar problemas como la multicolinealidad o la heterocedasticidad.
- Saber aplicar técnicas de ingeniería de características, selección de variables y regularización para optimizar el rendimiento y la interpretabilidad de los modelos.
- Extender el conocimiento de los modelos lineales a los GLM para poder modelar diferentes tipos de variables de respuesta (binarias, de conteo, etc.).
- Validar rigurosamente la capacidad predictiva de los modelos para asegurar su generalización a nuevos datos.

Temario de la asignatura

- **Tema 1:** Introducción a los Modelos de Regresión
- **Tema 2:** El Modelo de Regresión Lineal Simple
- **Tema 3:** El Modelo de Regresión Lineal Múltiple
- **Tema 4:** Ingeniería de Características: Transformaciones e Interacciones
- **Tema 5:** Selección de Variables, Regularización y Validación
- **Tema 6:** Modelos de Regresión Generalizada (GLM)
- **Tema 7:** Tópicos Avanzados en Modelado Predictivo

Desarrollo de la asignatura

El curso se desarrolla a lo largo de 15 semanas, con dos sesiones semanales. La metodología busca integrar la teoría y la práctica de forma fluida: los conceptos teóricos se presentarán junto con su implementación en R, fomentando un aprendizaje aplicado en cada clase.

Semana 1: Presentación e Introducción a los Modelos de Regresión

- **Contenidos:** Presentación de la asignatura, guía docente y sistema de evaluación. Se abordarán los conceptos fundamentales del modelado estadístico: el propósito dual de **predecir vs. explicar** y la anatomía de un modelo de regresión.
- **Objetivos de Aprendizaje:**
 - Diferenciar entre un objetivo de predicción y uno de inferencia.
 - Identificar los componentes clave de un modelo de regresión (respuesta, predictores, error).
 - Comprender el ecosistema de modelos de regresión y el rol de R en el curso.
- **Materiales Utilizados:**
 - Apuntes (Tema 1), Diapositivas (“Introducción a los Modelos Estadísticos”), `lab0_introduccion.html`.
- **Actividades Planificadas:**
 - Discusión sobre los objetivos de la asignatura.
 - Sesión práctica introductoria con `lab0_introduccion.html` para configurar el entorno de R y RStudio.
- **Trabajo Personal Recomendado:**
 - Lectura del Tema 1 de los apuntes.
 - Asegurar que el entorno de R/RStudio está correctamente instalado y funcionando.

Semana 2: Regresión Lineal Simple (RLS) - Fundamentos y Estimación

- **Contenidos:** Formulación del modelo RLS, supuestos teóricos y estimación de parámetros mediante **Mínimos Cuadrados Ordinarios (MCO)**. Interpretación de coeficientes y coeficiente de determinación R^2 .
- **Objetivos de Aprendizaje:**
 - Formular matemáticamente el modelo de RLS.
 - Estimar los coeficientes β_0 y β_1 usando MCO.
 - Interpretar el significado práctico de la pendiente, el intercepto y R^2 .
- **Materiales Utilizados:**
 - Apuntes (Secciones 2.1 a 2.3), Diapositivas (“Regresión Lineal Simple”), `lab1_regresion_simple.html`.
 - Ejercicios: Boletín de RLS (Ejercicios 1, 2, 3).
- **Actividades Planificadas:**
 - Ajuste de un primer modelo `lm()` en R.
 - Análisis e interpretación de la salida de la función `summary()`.
- **Trabajo Personal Recomendado:**
 - Lectura de las secciones 2.1, 2.2 y 2.3 de los apuntes.
 - Realización de los **ejercicios 1, 2 y 3** de RLS.

Semana 3: RLS - Inferencia y Diagnóstico

- **Contenidos:** Inferencia sobre los coeficientes (test-t), ANOVA para la significancia global (test F) y diagnóstico completo de los supuestos a través del análisis de residuos.
- **Objetivos de Aprendizaje:**
 - Realizar e interpretar contrastes de hipótesis para los coeficientes.
 - Evaluar la bondad de ajuste de un modelo mediante ANOVA y R^2 .
 - Diagnosticar la validez de los supuestos del modelo analizando los residuos.
- **Materiales Utilizados:**
 - Apuntes (Secciones 2.4 a 2.6), Diapositivas (“Regresión Lineal Simple”), `lab1_regresion_simple.html`.
 - Ejercicios: Boletín de RLS (Ejercicios 4 al 10).
- **Actividades Planificadas:**

- Taller práctico de diagnóstico de modelos, analizando gráficamente los residuos (linealidad, homocedasticidad, normalidad).
- Uso de tests estadísticos (Shapiro-Wilk, Breusch-Pagan) para confirmar el diagnóstico visual.
- **Trabajo Personal Recomendado:**
 - Completar el laboratorio `lab1_regresion_simple.html`.
 - Resolver los **ejercicios 4 al 10** de RLS.

Semana 4: Regresión Lineal Múltiple (RLM) - Formulación e Interpretación

- **Contenidos:** Extensión al caso multivariante. Interpretación de coeficientes bajo el principio *ceteris paribus*. Coeficiente de determinación ajustado (R_{adj}^2). Diagnóstico e identificación de la **multicolinealidad**.
- **Objetivos de Aprendizaje:**
 - Interpretar correctamente los coeficientes de una RLM como efectos parciales.
 - Diagnosticar la presencia de multicolinealidad utilizando el Factor de Inflación de la Varianza (VIF).
- **Materiales Utilizados:**
 - Apuntes (Tema 3), Diapositivas (“Regresión Lineal Múltiple”), `lab2_regresion_multiple.html`.
 - Ejercicios: Boletín de RLM (Ejercicios 1 al 6).
- **Actividades Planificadas:**
 - Ajuste de un modelo de RLM en R.
 - Taller práctico de diagnóstico de multicolinealidad, calculando e interpretando los valores VIF.
- **Trabajo Personal Recomendado:**
 - Lectura completa del Tema 3 de los apuntes.
 - Resolver los **ejercicios 1 al 6** de RLM.

Semana 5: Ingeniería de Características I - Transformaciones

- **Contenidos:** Técnicas para mejorar el modelo: transformaciones de variables para linealizar relaciones y estabilizar la varianza (logarítmica, Box-Cox). Codificación de variables categóricas (One-Hot vs. Ordinal).
- **Objetivos de Aprendizaje:**

- Saber cuándo y cómo aplicar transformaciones para corregir violaciones de supuestos.
 - Codificar correctamente variables categóricas según su naturaleza.
- **Materiales Utilizados:**
 - Apuntes (Secciones 4.1 a 4.5), Diapositivas (“Ingeniería de Características”), `lab3_ingenieria_caracteristicas.html`.
 - Ejercicios: Boletín de Ing. de Características (Ejercicios 1 al 6).
 - **Actividades Planificadas:**
 - Práctica en R sobre diagnóstico de no-linealidad y aplicación de transformaciones.
 - Creación de variables `*dummy`- para predictores categóricos.
 - **Trabajo Personal Recomendado:**
 - Lectura de las secciones sobre transformaciones en los apuntes (4.1-4.5).
 - Resolver los **ejercicios 1 al 6** (Tema 4).

Semana 6: Ingeniería de Características II - Interacciones

- **Contenidos:** Modelado de relaciones complejas mediante la inclusión, visualización e interpretación de **términos de interacción** entre variables continuas y categóricas.
- **Objetivos de Aprendizaje:**
 - Identificar la necesidad de incluir un término de interacción.
 - Ajustar, visualizar e interpretar un modelo con interacciones.
- **Materiales Utilizados:**
 - Apuntes (Sección 4.6), Diapositivas (“Ingeniería de Características”), `lab3_ingenieria_caracteristicas.html`.
 - Ejercicios: Boletín de Ing. de Características (Ejercicios 7 al 10).
- **Actividades Planificadas:**
 - Ajuste de modelos con interacciones en R.
 - Visualización de interacciones mediante gráficos de efectos para interpretar cómo cambia la pendiente.
- **Trabajo Personal Recomendado:**
 - Completar el laboratorio `lab3_ingenieria_caracteristicas.html`.
 - Resolver los **ejercicios 7 al 10** (Tema 4).

Semana 7: Consolidación y Práctica de Regresión Lineal

- **Contenidos:** Repaso y aplicación conjunta de RLS, RLM e ingeniería de características.
- **Objetivos de Aprendizaje:**
 - Integrar todos los conceptos de modelado lineal en un único flujo de trabajo.
 - Afrontar un problema de modelado de principio a fin: desde el análisis exploratorio hasta el diagnóstico final.
- **Materiales Utilizados:**
 - Todos los materiales de los Temas 2, 3 y 4. Ejercicios de repaso de los boletines correspondientes.
- **Actividades Planificadas:**
 - Resolución de un caso de estudio completo en clase.
 - Sesión de tutoría y resolución de dudas para el Trabajo 1.
- **Trabajo Personal Recomendado:**
 - Finalizar la preparación y entrega del Trabajo 1.

Semana 8: Evaluación Parcial e Introducción a la Selección de Variables

- **Contenidos:** Defensa oral del Trabajo 1. Introducción a los criterios de información (AIC, BIC, C_p de Mallows) para la comparación de modelos.
- **Objetivos de Aprendizaje:**
 - Defender y justificar las decisiones tomadas en un proyecto de modelado.
 - Comprender la base teórica del compromiso sesgo-varianza y cómo los criterios de información lo cuantifican.
- **Materiales Utilizados:**
 - Apuntes (Sección 5.3), Diapositivas (“Selección de Variables, Regularización y Validación”).
 - Ejercicios: Boletín de Selección de Variables (Ejercicio 3).
- **Actividades Planificadas:**
 - Presentaciones orales del Trabajo 1.
 - Clase teórica sobre criterios de bondad de ajuste.
- **Trabajo Personal Recomendado:**
 - Comenzar la lectura del Tema 5 y reflexionar sobre el **ejercicio conceptual 3**.

Semana 9: Métodos de Selección de Variables

- **Contenidos:** Estudio de los métodos automáticos (*stepwise*) y exhaustivos (*best subset*), analizando sus ventajas y, sobre todo, sus limitaciones prácticas.
- **Objetivos de Aprendizaje:**
 - Aplicar algoritmos de selección de variables en R.
 - Analizar críticamente los resultados de los métodos automáticos, entendiendo su inestabilidad.
- **Materiales Utilizados:**
 - Apuntes (Secciones 5.4, 5.5), Diapositivas (“Selección de Variables, Regularización y Validación”), `lab4_seleccion_variables.html`.
 - Ejercicios: Boletín de Selección de Variables (Ejercicios 4, 5, 6).
- **Actividades Planificadas:**
 - Práctica en R con las funciones `regsubsets()` y `step()`.
 - Discusión sobre la invalidez de los p-valores en modelos seleccionados por métodos *stepwise*.
- **Trabajo Personal Recomendado:**
 - Lectura de las secciones 5.4 y 5.5 de los apuntes.
 - Resolver los **ejercicios 4, 5 y 6**.

Semana 10: Regularización y Validación de Modelos

- **Contenidos:** Técnicas de regularización (Ridge, Lasso, Elastic Net) y fundamentos de la validación de modelos, con especial foco en la **validación cruzada** (Cross-Validation).
- **Objetivos de Aprendizaje:**
 - Comprender la diferencia fundamental entre la penalización L1 (Lasso) y L2 (Ridge).
 - Aplicar la validación cruzada para seleccionar hiperparámetros y evaluar el rendimiento predictivo del modelo.
- **Materiales Utilizados:**
 - Apuntes (Secciones 5.6, 5.7), Diapositivas (“Selección de Variables, Regularización y Validación”), `lab4_seleccion_variables.html`.
 - Ejercicios: Boletín de Selección de Variables (Ejercicios 7 al 10).
- **Actividades Planificadas:**
 - Ajuste de modelos Lasso y Ridge con `glmnet` en R.
 - Uso de `cv.glmnet` para encontrar el valor óptimo de λ .

- **Trabajo Personal Recomendado:**
 - Completar el laboratorio `lab4_seleccion_variables.html`.
 - Resolver los **ejercicios 7 al 10**.

Semana 11: Modelos Lineales Generalizados (GLM) - Fundamentos

- **Contenidos:** Introducción a los componentes clave que extienden el modelo lineal: la **familia exponencial**, la **función de enlace** y la estimación por Máxima Verosimilitud (MLE).
- **Objetivos de Aprendizaje:**
 - Identificar los tres componentes que definen cualquier GLM.
 - Entender por qué se necesita MLE en lugar de MCO para estimar los GLM.
- **Materiales Utilizados:**
 - Apuntes (Sección 6.1, 6.2), Diapositivas (“Modelos Lineales Generalizados”), `lab5_modelos_generalizados.html`.
 - Ejercicios: Boletín de GLM (Ejercicios 1, 2, 9).
- **Actividades Planificadas:**
 - Clase teórica vinculando la regresión lineal como un caso particular de los GLM.
 - Análisis de diferentes tipos de variables respuesta para identificar el GLM apropiado.
- **Trabajo Personal Recomendado:**
 - Lectura de las secciones 6.1 a 6.4 de los apuntes y resolución de los **ejercicios conceptuales 1, 2 y 9**.

Semana 12: GLM - Regresión Logística

- **Contenidos:** Estudio del modelo para respuestas binarias, incluyendo la interpretación de **Odds Ratios (OR)** y las métricas de validación específicas como la **curva ROC** y el **AUC**.
- **Objetivos de Aprendizaje:**
 - Ajustar e interpretar un modelo de regresión logística.
 - Calcular e interpretar los Odds Ratios.
 - Validar un clasificador binario mediante la matriz de confusión y la curva ROC.
- **Materiales Utilizados:**

- Apuntes (Sección 6.5), Diapositivas (“Modelos Lineales Generalizados”), `lab5_modelos_generalizados.html`.
- Ejercicios: Boletín de GLM (Ejercicios 3, 4, 5).
- **Actividades Planificadas:**
 - Ajuste de un modelo `glm()` con `family = binomial`.
 - Taller práctico de cálculo de métricas de clasificación y visualización de la curva ROC.
- **Trabajo Personal Recomendado:**
 - Resolver los **ejercicios 3, 4 y 5** sobre regresión logística.

Semana 13: GLM - Regresión de Poisson

- **Contenidos:** Modelado de datos de conteo, interpretación de **Incidence Rate Ratios (IRR)**, y diagnóstico y solución del problema de la **sobredispersión** mediante la Regresión Binomial Negativa.
- **Objetivos de Aprendizaje:**
 - Ajustar e interpretar un modelo de regresión de Poisson.
 - Diagnosticar la sobredispersión y saber cuándo usar un modelo Binomial Negativo.
- **Materiales Utilizados:**
 - Apuntes (Sección 6.6), Diapositivas (“Modelos Lineales Generalizados”), `lab5_modelos_generalizados.html`.
 - Ejercicios: Boletín de GLM (Ejercicios 6, 7, 8, 10).
- **Actividades Planificadas:**
 - Ajuste de un modelo `glm()` con `family = poisson`.
 - Cálculo del estadístico de dispersión y comparación con un modelo `glm.nb()`.
- **Trabajo Personal Recomendado:**
 - Completar el laboratorio `lab5_modelos_generalizados.html`.
 - Resolver los **ejercicios 6, 7, 8 y 10**.

Semana 14: Tópicos Avanzados y Repaso

- **Contenidos:** Breve introducción a extensiones de los modelos vistos, como los **Modelos Aditivos Generalizados (GAMs)**. Repaso general de toda la materia.
- **Objetivos de Aprendizaje:**

- Reconocer las situaciones en las que un GAM puede ser una alternativa útil.
 - Consolidar el conocimiento global de la asignatura y resolver dudas.
- **Materiales Utilizados:**
 - Todos los materiales del curso, incluyendo los Ejercicios Avanzados.
 - **Actividades Planificadas:**
 - Sesión de repaso general y consultas para el examen final.
 - Tutorías para el Trabajo Colectivo 2.
 - **Trabajo Personal Recomendado:**
 - Estudio para el examen final y desarrollo del trabajo.
 - Intentar resolver los **Ejercicios Avanzados**.

Semana 15: Presentación de Proyectos Finales

- **Contenidos:** Exposición y defensa del Trabajo Colectivo 2.
- **Objetivos de Aprendizaje:**
 - Aplicar las técnicas de la asignatura a un problema de modelado real.
 - Comunicar de forma efectiva los resultados de un análisis estadístico.
- **Materiales Utilizados:**
 - Proyectos desarrollados por los estudiantes.
- **Actividades Planificadas:**
 - Presentaciones orales de los trabajos finales por parte de los grupos.
- **Trabajo Personal Recomendado:**
 - N/A.

Bibliografía

A continuación, se presenta la bibliografía utilizada y recomendada para el curso, dividida en textos principales y referencias complementarias.

Bibliografía Principal y Recomendada

- Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.
- Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B*, 26(2), 211-243.
- Draper, N. R. (1998). *Applied regression analysis*. McGraw-Hill.
- Fox, J., & Weisberg, S. (2018). *An R companion to applied regression*. Sage publications.
- Harrell, F. E. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer.
- Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.
- Jaccard, J., & Turrisi, R. (2003). *Interaction effects in multiple regression*. Sage.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R*. Springer.
- Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: a practical approach for predictive models*. CRC Press.
- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models*. McGraw-Hill.
- McCullagh, P. (2019). *Generalized linear models*. Routledge.
- Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A*, 135(3), 370-384.
- Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer.
- Shmueli, G. (2010). To Explain or to Predict?. *Statistical Science*, 25(3), 289-310.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA.
- Weisberg, S. (2005). *Applied linear regression*. Wiley.
- Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

- Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: principles and techniques for data scientists*. O'Reilly Media.

Referencias Adicionales y Tópicos Relacionados

- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.
- Bellman, R. E. (1978). *Artificial intelligence: can computers think?*.
- Flach, P. (2012). *Machine learning: the art and science of algorithms that make sense of data*. Cambridge University Press.
- Kelleher, J. D., & Tierney, B. (2018). *Data science*. MIT Press.
- Kelleher, J. D., Mac Namee, B., & D'arcy, A. (2020). *Fundamentals of machine learning for predictive data analytics*. MIT press.
- Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.
- Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.
- Russell, S. J. (2010). *Artificial intelligence a modern approach*. Pearson Education, Inc.
- Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, 1, 29-39.