  <!– Header –> <img src="../images/urjc$_l$ogo.png"alt = "$LogoURJC$"$class$ = "$portada-logo$" >

 Escuela Técnica Superior de
 Ingeniería Informática

<!– Contenido principal –> <h1 class="portada-titulo"> Métodos Estadísticos de Predicción </h1>

<h2 class="portada-subtitulo"> Ejercicios de la Asignatura </h2>

 Grado en Matemáticas

 Autores Víctor Aceña Gil Isaac Martín de Diego

2025-2026

<!– Footer –> <img src="https://i.creativecommons.org/l/by-sa/4.0/88x31.png" alt="Licencia Creative Commons" class="portada-licencia-logo">

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

<!– Elementos decorativos –>

Índice ejercicios

Introducción

Este documento recopila una serie de ejercicios prácticos y teóricos diseñados para complementar la asignatura "Modelos Estadísticos de Predicción" del Grado en Matemáticas. El objetivo de esta colección es afianzar los conocimientos adquiridos en cada tema, fomentando tanto la comprensión de los fundamentos teóricos como la habilidad para implementar y diagnosticar modelos en R.

Estructura de los Ejercicios

Los ejercicios están organizados por temas, siguiendo la estructura del curso. Para cada tema, encontrarás una mezcla de:

- **Preguntas Conceptuales:** Diseñadas para reforzar la comprensión de la teoría subyacente.
- **Problemas Prácticos con R:** Enfocados en la aplicación de las técnicas a conjuntos de datos reales o simulados.
- **Ejercicios de Interpretación:** Centrados en la habilidad crítica de interpretar correctamente las salidas de los modelos estadísticos.

Requisitos Previos

Para abordar estos ejercicios, se asume que el estudiante ha estudiado el contenido teórico del tema correspondiente y posee un manejo básico del entorno de programación R y RStudio.

Regresión Lineal Simple

Ejercicio 1: Fundamentos Conceptuales

Basándote en el texto, explica con tus propias palabras por qué un coeficiente de correlación de Pearson (r) alto no es suficiente para modelar una relación y por qué la regresión lineal es un paso más allá. Menciona al menos dos cosas que el modelo de regresión proporciona y que la correlación por sí sola no ofrece.

Ejercicio 2: Interpretación de Coeficientes

Un analista ajusta un modelo para predecir el gasto anual en compras online (`gasto`, en euros) basándose en la edad del cliente (`edad`). El modelo ajustado es:

```
gasto = 1500 + 12 * edad
```

a) ¿Cuál es el gasto predicho para un cliente de 30 años?
b) Interpreta el significado de la pendiente (12) en el contexto específico de este problema.
c) Interpreta el significado del intercepto (1500). ¿Crees que esta interpretación tiene sentido práctico en el mundo real? ¿Por qué?

Ejercicio 3: Aplicación Práctica con R (Ajuste e Inferencia)

Utiliza el conjunto de datos `pressure` de R, que contiene mediciones de temperatura y presión de vapor de mercurio.

a) Ajusta un modelo de regresión lineal simple para predecir la presión (`pressure`) en función de la temperatura (`temperature`). Guarda el modelo en un objeto.
b) Utiliza la función `summary()` sobre el objeto del modelo.
c) Interpreta el valor del **coeficiente de determinación R^2**. ¿Qué porcentaje de la variabilidad de la presión es explicado por la temperatura?
d) Interpreta el **p-valor del estadístico F**. ¿Es el modelo útil en su conjunto?
e) ¿Es el coeficiente de la temperatura estadísticamente significativo a un nivel de $\alpha = 0.05$? Justifica tu respuesta basándote en el p-valor del test t.

Ejercicio 4: Intervalos de Confianza y Predicción

Usando el modelo del ejercicio anterior (`lm(pressure ~ temperature, data = pressure)`):

a) Calcula el **intervalo de confianza al 95%** para la *presión media* esperada cuando la temperatura es de 250 grados.
b) Calcula el **intervalo de predicción al 95%** para la presión de una *única y nueva* medición realizada a 250 grados.
c) ¿Cuál de los dos intervalos es más ancho? Explica la razón teórica de esta diferencia.

Ejercicio 5: Supuestos del Modelo

Enumera los cuatro supuestos del modelo de regresión lineal clásico (también conocidos como supuestos de Gauss-Markov) y explica brevemente la importancia de cada uno.

Ejercicio 6: Diagnóstico de Linealidad y Homocedasticidad

Para el modelo del ejercicio 3:

a) Genera y muestra el gráfico de **Residuos vs. Valores Ajustados**. Basándote en este gráfico, ¿se cumple el supuesto de **linealidad**? Explica en qué te basas.
b) Genera y muestra el gráfico **Scale-Location**. Basándote en este gráfico, ¿se cumple el supuesto de **homocedasticidad**? Describe el patrón que indicaría un problema de heterocedasticidad.

Ejercicio 7: Diagnóstico de Normalidad

Para el modelo del ejercicio 3:

a) Genera un gráfico **Normal Q-Q** de los residuos. ¿Parecen seguir los residuos una distribución normal?
b) Realiza un **test de Shapiro-Wilk** sobre los residuos del modelo. ¿Qué concluyes a partir del p-valor?

Ejercicio 8: Descomposición de la Varianza (ANOVA)

Explica qué representan la **Suma de Cuadrados Total (SST)**, la **Suma de Cuadrados de la Regresión (SSR)** y la **Suma de Cuadrados del Error (SSE)**. ¿Cuál es la ecuación fundamental que las relaciona?

Ejercicio 9: Observaciones Influyentes

Basado en la teoría de los apuntes:

a) Explica la diferencia entre un residuo simple (e_i), un residuo estandarizado y un residuo estudentizado. ¿Por qué se prefieren los estudentizados para el diagnóstico?
b) ¿Qué mide el **leverage** (h_{ii})? ¿Y la **distancia de Cook** (D_i)? ¿Puede una observación tener un leverage alto y no ser influyente?

Ejercicio 10: Relación entre Pruebas de Hipótesis

En el contexto **exclusivo** de la regresión lineal simple, ¿qué relación matemática existe entre el estadístico **F** del test ANOVA y el estadístico **t** del test para la pendiente β_1? ¿Qué implica esto para sus respectivos p-valores?

Regresión Lineal Múltiple

Ejercicio 1: Conceptual (Interpretación *Ceteris Paribus*)

Un analista ajusta dos modelos para predecir el consumo de un coche (`mpg`):

1. `lm(mpg ~ wt)` obtiene un coeficiente para `wt` de -5.3.
2. `lm(mpg ~ wt + hp)` obtiene un coeficiente para `wt` de -3.8.

Explica detalladamente por qué el coeficiente para la variable `wt` (peso) cambia al añadir la variable `hp` (caballos de fuerza). ¿Cuál de los dos coeficientes representa el efecto "puro" o "aislado" del peso? Fundamenta tu respuesta en el principio de ***ceteris paribus***.

Ejercicio 2: Práctico (Ajuste e Interpretación de un Modelo Múltiple)

Usa el conjunto de datos `iris` de R. Queremos modelar la anchura del pétalo (`Petal.Width`) en función de la longitud del pétalo (`Petal.Length`) y la anchura del sépalo (`Sepal.Width`).

a) Ajusta un modelo de regresión lineal múltiple: `lm(Petal.Width ~ Petal.Length + Sepal.Width, data = iris)`.
b) Interpreta el coeficiente estimado para `Petal.Length`.
c) Interpreta el coeficiente estimado para `Sepal.Width`.
d) Interpreta el intercepto del modelo. ¿Tiene un significado práctico en este contexto biológico?

Ejercicio 3: Conceptual (R^2 vs. R^2 Ajustado)

Cuando pasamos de un modelo simple a uno múltiple, introducimos el **R^2 ajustado** como medida de bondad de ajuste.

a) ¿Cuál es el principal problema de usar el R^2 tradicional para comparar modelos con diferente número de predictores?
b) ¿Cómo soluciona el R^2 ajustado este problema? Explica qué "penalización" introduce en su fórmula.

Ejercicio 4: Interpretación de Salidas de R

Te presentan el siguiente resumen de un modelo que predice el prestigio de una ocupación (`prestige`) en función de los ingresos (`income`) y el nivel educativo (`education`).

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.0647    4.2750   -1.419   0.1595
income       0.0013    0.0003    4.524   1.9e-05 ***
education    4.1832    0.3887   10.762   < 2e-16 ***

Multiple R-squared:  0.79,  Adjusted R-squared:  0.785
F-statistic: 185.6 on 2 and 99 DF,  p-value: < 2.2e-16
```

a) ¿Es el modelo globalmente significativo? ¿En qué te basas?
b) ¿Son los predictores `income` y `education` individualmente significativos, después de controlar por el efecto del otro? Justifica tu respuesta.
c) Explica la diferencia conceptual entre lo que evalúa el **test F global** y lo que evalúan los **tests t individuales** en este modelo.

Ejercicio 5: Conceptual (Multicolinealidad)

Describe con tus propias palabras qué es la **multicolinealidad**. Menciona tres consecuencias negativas que puede tener la multicolinealidad severa en un modelo de regresión y si afecta más a la **predicción** o a la **inferencia**.

Ejercicio 6: Práctico (Diagnóstico de Multicolinealidad)

Usa el dataset `mtcars`. Ajusta un modelo para predecir el consumo (`mpg`) usando como predictores el número de cilindros (`cyl`), la cilindrada (`disp`), los caballos de fuerza (`hp`) y el peso (`wt`).

a) Observa el `summary()` del modelo. ¿Hay alguna variable que, a pesar de tener una alta correlación simple con `mpg`, no resulte significativa en el modelo múltiple?
b) Carga la librería `car` y calcula el **Factor de Inflación de la Varianza (VIF)** para cada predictor.
c) Basándote en los valores del VIF, ¿qué variables presentan un problema de multicolinealidad? ¿Cuál es tu recomendación para simplificar el modelo?

Ejercicio 7: Teórico (Notación Matricial)

a) Escribe la fórmula del estimador de Mínimos Cuadrados Ordinarios ($\hat{\beta}$) en notación matricial.
b) ¿Qué supuesto fundamental del modelo de regresión múltiple garantiza que la matriz ($\mathbf{X}^T\mathbf{X}$) sea invertible?

Ejercicio 8: Práctico (Gráficos de Regresión Parcial)

Usa el dataset `Prestige` de la librería `car`.

a) Ajusta el modelo `lm(prestige ~ income + education + women, data = Prestige)`.
b) Genera los gráficos de regresión parcial (o "added-variable plots") para este modelo usando la función `avPlots(tu_modelo)`.
c) Explica qué representa el gráfico para la variable `education`. ¿Qué significan los ejes X e Y de ese gráfico específico? ¿A qué corresponde la pendiente de la línea en ese gráfico?

Ejercicio 9: Inferencia (F-test vs. t-tests)

Describe un escenario hipotético en el que el **test F global** de un modelo de regresión múltiple sea altamente significativo ($p < 0.001$), pero **ninguno de los tests t individuales** para los coeficientes sea significativo. ¿Cuál es la causa estadística más probable de este fenómeno?

Ejercicio 10: Práctico (Comparación de Modelos Anidados)

Usa el dataset `swiss`.

a) Ajusta un **modelo reducido** para predecir `Fertility` usando solo `Agriculture` y `Education`.
b) Ajusta un **modelo completo** que, además de las variables anteriores, incluya `Catholic` y `Infant.Mortality`.
c) Utiliza la función `anova()` para comparar formalmente los dos modelos. ¿Aportan las variables `Catholic` y `Infant.Mortality` una mejora estadísticamente significativa al modelo? Interpreta el p-valor del test F resultante.

Ingeniería de Características

Ejercicio 1: Conceptual (Diagnóstico antes de Transformar)

El texto desaconseja fuertemente el enfoque de "ensayo y error" al aplicar transformaciones. Explica con tus propias palabras por qué la práctica de probar transformaciones hasta que mejore el R^2 es metodológicamente peligrosa. Menciona al menos tres de los riesgos específicos discutidos en los apuntes.

Ejercicio 2: Práctico (Escalado de Variables)

Utiliza el dataset `iris` de R y céntrate en las cuatro variables predictoras continuas (`Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width`).

a) Calcula la media y la desviación estándar de estas cuatro variables en su escala original. ¿Son sus escalas directamente comparables?
b) Crea un nuevo data frame donde hayas aplicado la **estandarización Z-Score** a estas cuatro variables. Verifica que las nuevas variables tienen una media cercana a 0 y una desviación estándar de 1.
c) ¿Por qué este paso de escalado es crucial antes de aplicar métodos de regularización como Ridge o Lasso, tal y como se menciona en el texto?

Ejercicio 3: Conceptual (Elección del Método de Escalado)

Describe un escenario hipotético para cada uno de los siguientes casos, explicando por qué el método de escalado elegido sería el más apropiado:

a) Un escenario donde la **estandarización Z-Score** es preferible.
b) Un escenario donde la **normalización Min-Max** es preferible.
c) Un escenario donde el **escalado robusto** (usando mediana y IQR) es necesario.

Ejercicio 4: Práctico (Transformación para Linealizar)

En el tema anterior vimos que la relación en el dataset `cars` (entre `speed` y `dist`) no era perfectamente lineal.

a) Ajusta el modelo `lm(dist ~ speed, data = cars)` y genera el gráfico de residuos vs. valores ajustados para confirmar visualmente la no linealidad (patrón curvo).
b) Los apuntes sugieren que la transformación logarítmica es útil para relaciones con "rendimientos decrecientes". Propón y aplica una transformación (ej. sobre el predictor, la respuesta, o ambos) para intentar linealizar la relación. Por ejemplo, ajusta `lm(log(dist) ~ speed, data = cars)`.
c) Genera de nuevo el gráfico de residuos vs. valores ajustados para el nuevo modelo. Compara ambos diagnósticos. ¿Ha mejorado la linealidad?

Ejercicio 5: Práctico (Transformación de Box-Cox)

Usa el dataset `Boston` de la librería `MASS`. La variable respuesta `medv` (valor mediano de la vivienda) es estrictamente positiva y tiene cierta asimetría.

a) Carga la librería `MASS` y utiliza la función `boxcox()` para encontrar el valor de λ óptimo para la variable `medv` en un modelo simple frente a `lstat`. La fórmula sería `boxcox(medv ~ lstat, data = Boston)`.
b) Observando el gráfico que se genera, ¿a qué valor "simple" (como -1, 0, 0.5, 1) se aproxima el λ óptimo?
c) Basándote en este resultado, ¿cuál de las transformaciones clásicas (logarítmica, raíz cuadrada, inversa, etc.) sería la más recomendable para la variable `medv`?

Ejercicio 6: Conceptual (Codificación de Variables Categóricas)

Explica la diferencia fundamental entre la **Codificación Ordinal** y la **Codificación One-Hot**. Para cada una de las siguientes variables, indica qué método de codificación usarías y justifica tu elección:

- `mes`: ("Enero", "Febrero", "Marzo", ...)
- `nivel_riesgo`: ("Bajo", "Medio", "Alto", "Crítico")
- `pais_origen`: ("España", "Francia", "Alemania", "Italia")

Ejercicio 7: Práctico (Interacción entre Variables Continuas)

Usa el dataset `mtcars` para investigar si el efecto del peso de un coche (`wt`) sobre su consumo (`mpg`) depende de su potencia (`hp`).

a) Ajusta un modelo que incluya un término de interacción entre `wt` y `hp`. Escribe la fórmula en R.
b) Observa el `summary()` del modelo. ¿Es el término de interacción (`wt:hp`) estadísticamente significativo a un nivel de $\alpha = 0.05$?
c) Basándote en el signo del coeficiente de la interacción, ¿cómo cambia el efecto del peso sobre el consumo a medida que aumenta la potencia? (Es decir, ¿el efecto negativo del peso se hace más fuerte o más débil en los coches más potentes?).

Ejercicio 8: Interpretación de una Interacción (Continua x Categórica)

Un investigador modela el salario (`salario`, en euros) en función de los años de experiencia (`experiencia`) y si el empleado tiene o no un máster (`master`, con "No" como categoría de referencia). El modelo ajustado es:

```
salario = 30000 + 1200*experiencia + 8000*masterSi + 300*experiencia:masterSi
```

a) Escribe la ecuación de regresión específica para los empleados que **no tienen** un máster.
b) Escribe la ecuación de regresión específica para los empleados que **sí tienen** un máster.
c) Interpreta el coeficiente de la interacción (300). ¿Qué nos dice sobre el retorno económico de la experiencia para ambos grupos?

Ejercicio 9: Conceptual (Principio de Jerarquía)

Explica el **principio de jerarquía** en el contexto de los modelos de regresión con interacciones. Si un modelo incluye el término de interacción `A:B`, ¿por qué es una buena práctica incluir siempre los efectos principales `A` y `B`, incluso si sus tests t individuales no son significativos?

Ejercicio 10: Conceptual (Ingeniería de Características Avanzada)

Los apuntes discuten la creación de nuevas variables mediante **ratios** y **combinaciones**. Para cada uno de los siguientes escenarios, propón una nueva variable (feature) que podrías crear y explica qué relación podría capturar mejor que las variables originales por sí solas.

a) Para predecir la rentabilidad de una tienda, tienes las variables `ventas_totales` y `numero_de_empleados`.
b) Para predecir el riesgo de impago de un solicitante de préstamo, tienes las variables `ingresos_anuales` y `deuda_total`.

Selección de variables, Regularización y Validación

Ejercicio 1: Conceptual (Sobreajuste vs. Subajuste)

Explica con tus propias palabras qué es el **sobreajuste (overfitting)** y el **subajuste (underfitting)**. Describe los síntomas de cada uno comparando el error de entrenamiento con el error de validación (o de test), y menciona la solución principal para cada problema.

Ejercicio 2: Práctico (Filtrado Básico)

Imagina que recibes un nuevo conjunto de datos con 50 predictores para un modelo de regresión. Antes de aplicar métodos computacionalmente costosos, decides hacer un filtrado inicial. Describe los **cuatro criterios básicos** que aplicarías para descartar variables de forma preliminar, según lo explicado en los apuntes.

Ejercicio 3: Conceptual (AIC vs. BIC)

Tanto el AIC como el BIC son criterios para comparar modelos, pero se basan en filosofías distintas y tienen penalizaciones diferentes.

a) Escribe la fórmula de la penalización por complejidad para el AIC y para el BIC.
b) ¿Cuál de los dos criterios tenderá a seleccionar modelos más simples (más parsimoniosos)? ¿Por qué?
c) Si tu objetivo principal es la **precisión predictiva**, ¿cuál de los dos criterios es generalmente preferido?

Ejercicio 4: Práctico (Best Subset y Criterios de Información)

Usa el conjunto de datos `mtcars` y la librería `leaps`.

a) Utiliza la función `regsubsets()` para realizar una selección del mejor subconjunto (`best subset selection`) para predecir `mpg` usando el resto de variables.

b) Obtén el `summary()` de los resultados. ¿Qué modelo (cuántas variables) es el mejor según el criterio **Cp de Mallows**?
c) ¿Y cuál es el mejor modelo según el **R^2 ajustado**?
d) ¿Coinciden ambos criterios en el número de variables del modelo óptimo?

Ejercicio 5: Conceptual (Métodos Stepwise)

Los métodos automáticos paso a paso (forward, backward, stepwise) son computacionalmente eficientes, pero el texto advierte sobre su uso. Menciona y explica brevemente **tres de las principales limitaciones o problemas** de estos métodos.

Ejercicio 6: Práctico (Selección Backward Stepwise)

Utiliza el conjunto de datos `swiss` para predecir `Fertility`.

a) Ajusta el modelo completo: `modelo_completo <- lm(Fertility ~ ., data = swiss)`.
b) Utiliza la función `step()` para realizar una selección **regresiva (backward)** basada en el criterio AIC.
c) Reporta la fórmula del modelo final que selecciona el algoritmo y su valor de AIC.

Ejercicio 7: Conceptual (Ridge vs. Lasso)

La regresión Ridge y Lasso son dos métodos de regularización muy populares, pero tienen un efecto fundamentalmente diferente sobre los coeficientes del modelo.

a) ¿Qué tipo de penalización utiliza cada método (L_1 o L_2)?
b) ¿Cuál de los dos métodos puede realizar selección de variables (es decir, anular coeficientes por completo)?
c) Describe un escenario en el que preferirías usar Ridge sobre Lasso.

Ejercicio 8: Práctico (Regresión Lasso)

Utiliza el paquete `glmnet` y el conjunto de datos `mtcars` para predecir `mpg`.

a) Prepara los datos: crea una matriz `x` para los predictores y un vector `y` para la respuesta.
b) Utiliza la función `cv.glmnet()` para realizar una validación cruzada y encontrar el valor de `lambda` óptimo para una regresión **Lasso** (`alpha = 1`).
c) Extrae y muestra los coeficientes del modelo Lasso ajustado con el `lambda.min`.
d) ¿Qué variables ha eliminado el modelo (coeficientes iguales a cero)?

Ejercicio 9: Conceptual (Validación)

Explica la diferencia entre la estrategia de validación **Train/Test Split simple** y la **Validación Cruzada k-fold**. ¿Cuál es la principal ventaja de la validación cruzada sobre la división simple? ¿En qué situación (tamaño del dataset) recomendarías usar cada una?

Ejercicio 10: Práctico (Validación Cruzada)

Imagina que has ajustado dos modelos para predecir `mpg` en el dataset `mtcars`: 1. Un modelo simple: `mpg ~ wt + hp` 2. Un modelo complejo: `mpg ~ .` (todas las variables)

Utilizando la librería `caret` y la función `train()`, como se muestra en el `callout-tip` "La maldición del sobreajuste", configura y ejecuta una **validación cruzada de 10 particiones** para estimar el **RMSE** de ambos modelos. ¿Cuál de los dos modelos generaliza mejor a nuevos datos según esta estimación?

Modelos de Regresión Generalizada

Ejercicio 1: Conceptual (Fundamentos de GLM)

Explica los **tres componentes clave** que definen a cualquier Modelo Lineal Generalizado (GLM) y describe brevemente la función de cada uno.

Ejercicio 2: Conceptual (Función de Enlace)

¿Cuál es el propósito fundamental de la **función de enlace** en un GLM? ¿Por qué la regresión lineal clásica es considerada un caso particular de un GLM? (Pista: piensa en su función de enlace).

Ejercicio 3: Práctico (Ajuste de un Modelo Logístico)

Usa el conjunto de datos `mtcars` de R. La variable `am` indica si la transmisión de un coche es automática (0) o manual (1).

a) Ajusta un modelo de regresión logística para predecir la probabilidad de que una transmisión sea manual (`am`) en función del peso del coche (`wt`) y los caballos de fuerza (`hp`).
b) Utiliza la función `summary()` para examinar el modelo. ¿Qué variables parecen ser significativas?
c) Obtén los coeficientes del modelo. ¿Cómo interpretarías el signo del coeficiente para la variable `wt`?

Ejercicio 4: Interpretación (Odds Ratios)

Basado en el modelo del ejercicio anterior:

a) Calcula el **Odds Ratio (OR)** para el coeficiente de la variable `hp`.
b) Interpreta este Odds Ratio en el contexto del problema. Específicamente, ¿cómo cambian las "odds" (la razón de probabilidad) de tener una transmisión manual por cada caballo de fuerza adicional, manteniendo el peso constante?

Ejercicio 5: Práctico (Validación del Modelo Logístico)

Continuando con el modelo logístico de `mtcars`:

a) Genera las predicciones de probabilidad del modelo para los datos.
b) Convierte estas probabilidades en clases ("0" o "1") usando un umbral de decisión de 0.5.
c) Crea la **matriz de confusión** comparando las predicciones con los valores reales.
d) Calcula la **precisión (accuracy)** global del modelo.
e) (Bonus) Utiliza el paquete `pROC` para calcular y visualizar la **curva ROC** y obtener el valor del **AUC**. ¿Qué tan buena es la capacidad discriminativa del modelo?

Ejercicio 6: Conceptual (Regresión de Poisson)

a) ¿Qué tipo de variable respuesta está diseñada para modelar la regresión de Poisson?
b) ¿Cuál es el supuesto fundamental de la distribución de Poisson respecto a la relación entre la media y la varianza?
c) ¿Cómo se llama el problema que surge cuando este supuesto se viola y la varianza es mayor que la media?

Ejercicio 7: Práctico (Ajuste de un Modelo de Poisson)

El dataset `discoveries` de R es una serie temporal que cuenta el número de "grandes inventos" por año.

a) Crea un gráfico de la serie temporal. ¿Parece la media del conteo constante a lo largo del tiempo?
b) Ajusta un modelo de regresión de Poisson simple donde `discoveries` es la respuesta y el tiempo (`time(discoveries)`) es el predictor.
c) Interpreta el coeficiente del tiempo. (Pista: recuerda exponenciarlo para obtener el Incidence Rate Ratio - IRR).

Ejercicio 8: Diagnóstico (Sobredispersión)

a) Para el modelo de Poisson del ejercicio anterior, calcula el **estadístico de dispersión ($\hat{\phi}$)**. (Pista: $\hat{\phi} = \frac{\sum r_i^2}{n-p}$, donde los r_i son los residuos Pearson).
b) Basándote en el valor de $\hat{\phi}$, ¿hay evidencia de sobredispersión?
c) Si encuentras sobredispersión, ¿cuál es el modelo alternativo que proponen los apuntes? ¿Qué ventaja teórica ofrece este modelo alternativo?

Ejercicio 9: Conceptual (Deviance)

La **deviance** es la medida principal de bondad de ajuste en los GLM. Explica conceptualmente qué mide. ¿Cómo se utiliza la diferencia en deviance entre dos modelos anidados para decidir cuál es mejor?

Ejercicio 10: Elección del Modelo Adecuado

Para cada uno de los siguientes escenarios, indica qué tipo de GLM (Logístico, Poisson, Binomial Negativo, Gamma...) sería el más apropiado y por qué.

a) Quieres modelar el **tiempo (en minutos)** que tarda un cliente en resolver una consulta en un centro de atención telefónica. El tiempo es siempre positivo y muchos valores se agrupan en tiempos cortos, con una cola larga de tiempos muy largos.
b) Quieres predecir la **presencia o ausencia** de una especie de planta en diferentes parcelas de un bosque.
c) Quieres modelar el **número de visitas** que cada usuario hace a una página web en un mes. Observas que la varianza del número de visitas es mucho mayor que la media.

Ejercicios Avanzados

Ejercicio 1: Derivación de Estimadores

Considera el modelo de regresión lineal simple $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. Partiendo de la función objetivo de Mínimos Cuadrados Ordinarios (MCO), $S(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$, realiza la derivación matemática completa para obtener las expresiones de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$. Muestra todos los pasos, desde el cálculo de las derivadas parciales hasta la resolución de las ecuaciones normales.

Ejercicio 2: El Impacto de la Multicolinealidad

En un modelo de regresión múltiple con dos predictores estandarizados (X_1, X_2), la varianza del estimador $\hat{\beta}_1$ viene dada por $\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{n(1-r_{12}^2)}$, donde r_{12} es la correlación entre X_1 y X_2.

a) Explica matemáticamente qué le ocurre a la varianza de $\hat{\beta}_1$ cuando la correlación entre los predictores (r_{12}) se aproxima a 1 (multicolinealidad perfecta).
b) Relaciona esta fórmula con la del Factor de Inflación de la Varianza (VIF). ¿Cómo demuestra esta expresión que la multicolinealidad "infla" la varianza de los estimadores de los coeficientes?

Ejercicio 3: Interpretación de Coeficientes en Modelos Transformados

Considera un modelo de regresión **log-log**: $\log(Y_i) = \beta_0 + \beta_1 \log(X_i) + \varepsilon_i$. Demuestra matemáticamente que el coeficiente β_1 puede interpretarse como una **elasticidad**, es decir, el cambio porcentual en Y ante un cambio del 1% en X. (Pista: utiliza la derivada de $\log(Y)$ con respecto a $\log(X)$).

Ejercicio 4: Fundamentos de la Regularización

Explica desde una perspectiva geométrica por qué la regularización **Lasso (penalización L1)** es capaz de reducir los coeficientes exactamente a cero, realizando así selección de variables, mientras que la regularización **Ridge (penalización L2)** solo puede encoger los coeficientes

hacia cero sin anularlos por completo. Apoya tu explicación con un dibujo o descripción de las "regiones de restricción" de ambos métodos en un espacio de dos coeficientes (β_1, β_2).

Ejercicio 5: La Familia Exponencial y los GLM

La teoría de los Modelos Lineales Generalizados (GLM) se basa en que distribuciones como la Normal, Binomial o Poisson pertenecen a la **familia exponencial**. La forma canónica de esta familia establece una relación directa entre la media y la varianza a través de la **función de varianza** $V(\mu)$. Explica cuál es la función de varianza para un modelo de **Poisson** y para un modelo **Binomial**. ¿Qué implicaciones tiene la forma de $V(\mu)$ en cada caso sobre el comportamiento de los datos y los supuestos del modelo?

Ejercicio 6: El Problema de la Inferencia en Métodos Stepwise

Los apuntes advierten que los p-valores de un modelo final obtenido mediante selección por pasos (stepwise) están **sesgados y son excesivamente optimistas**. Explica el razonamiento estadístico detrás de esta advertencia. ¿Por qué el proceso iterativo de "buscar y seleccionar" la variable más significativa en cada paso invalida los supuestos teóricos del test t estándar?

Ejercicio 7: Propiedades de los Estimadores MCO

El **Teorema de Gauss-Markov** establece que, bajo ciertos supuestos, los estimadores de Mínimos Cuadrados Ordinarios (MCO) son **MELI (Mejores Estimadores Lineales Insesgados)**. Demuestra la propiedad de **insesgadez** para el estimador $\hat{\boldsymbol{\beta}}$ en notación matricial. Es decir, demuestra que $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$. Muestra todos los pasos y menciona qué supuestos del modelo estás utilizando en cada paso.

Ejercicio 8: Intervalos de Confianza vs. Predicción

La fórmula para el intervalo de predicción para una nueva observación en regresión lineal simple es:

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

Explica el origen y el significado de cada uno de los **tres términos** que se encuentran dentro del paréntesis bajo la raíz cuadrada. ¿Qué fuente de incertidumbre representa cada término y por qué la suma de los tres es necesaria para un intervalo de predicción?

Ejercicio 9: Estimación por Máxima Verosimilitud

Para un modelo de regresión logística, la función de log-verosimilitud es:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

donde $p_i = \frac{1}{1+e^{-\mathbf{x}_i^T \boldsymbol{\beta}}}$. Deriva la **ecuación de puntuación (score equation)** para un coeficiente β_j (es decir, calcula $\frac{\partial \ell}{\partial \beta_j}$) y demuestra que se iguala a cero cuando $\sum_{i=1}^{n} x_{ij}(y_i - p_i) = 0$. Interpreta el significado de esta condición final.

Ejercicio 10: El Coeficiente de Regresión Parcial

El texto afirma que el coeficiente $\hat{\beta}_j$ de una regresión múltiple puede entenderse como el coeficiente de una regresión simple entre dos conjuntos de residuos. Explica con detalle este concepto de **regresión parcial**. ¿Qué se está "parcializando" o "eliminando" de la variable respuesta Y y del predictor X_j antes de calcular su relación? ¿Por qué este concepto es fundamental para entender la interpretación *ceteris paribus*?