Práctica 2: Modelado Estadístico de Datos

Alex Herrerías Ramírez

Diciembre 2026

Ejercicio 1

1. (2 puntos) A partir del siguiente código en R se pide completarlo para:
   a) Proporcionar un modelo final.
   b) Realizar un análisis de residuos de dicho modelo final.

a) Selección y desarrollo del modelo

Tenemos que realizar un modelo a partir del conjunto de datos `wine`, el cual contiene los resultados de análisis químicos de vinos. Dado que se nos solicita un "análisis de residuos" y no se nos proporciona una variable objetivo, he decidido desarrollar un modelo de **Regresión Lineal Múltiple**.

Objetivo: Nuestro principal objetivo es conseguir predecir la concentración de **Alcohol** en función del resto de características fisicoquímicas y el tipo de vino.

Estrategia:

1. **Carga y limpieza:** Verificar la estructura de los datos. La variable `Type` debe tratarse como factor.
2. **Modelo Completo:** Ajustar un modelo inicial incluyendo todas las variables predictoras disponibles.
3. **Selección de Variables (Stepwise):** Utilizar el criterio de información de Akaike (AIC) mediante un algoritmo *stepwise* (dirección "both") para eliminar predictores redundantes o no informativos, reduciendo la multicolinealidad y mejorando la parsimonia del modelo.

Código en R propuesto:

```r
#  cargamos las librerias necesarias
library(rattle)
library(MASS)
library(car)
data(wine)

# creamos el modelo lineal usando alcohol como objetivo
modelo_completo <- lm(Alcohol ~ ., data = wine)

# usamos stepAIC para obtener las variables utiles
modelo_final <- stepAIC(modelo_completo, direction = "both", trace =
    0)

summary(modelo_final)

# comprobamos si existe multicolinealidad
vif(modelo_final)
```

Interpretación del Modelo Final

Tras aplicar `stepAIC`, el modelo final es:

$$\text{Alcohol} \sim \text{Type} + \text{Malic} + \text{Ash} + \text{Color} + \text{Hue}$$

El uso de `stepAIC` ha descartado variables como la Alcalinidad, el Magnesio o los Flavonoides.

1. **Variables clave:**
   - **Tipo y Color:** `Type` es el factor determinante; respecto al Tipo 1 (base), los Tipos 2 y 3 reducen significativamente el nivel de alcohol (coef. negativos). El `Color` muestra una relación positiva (+0,12): a mayor intensidad, más alcohol.
   - **Selección AIC:** `Malic` y `Hue` se mantienen en el modelo por criterio AIC. Aunque sus p-valores superan el 0.05, su inclusión mejora la capacidad predictiva global y reduce el error.
2. **Bondad de ajuste:** El modelo presenta un R^2 ajustado de **0.6471**, explicando casi el **65 % de la variabilidad** total. El test F ($p < 2{,}2e^{-16}$) confirma la validez estadística global del ajuste.
3. **Multicolinealidad:** El análisis de VIF descarta correlaciones problemáticas entre variables (todos los valores generalizados están por debajo de 2), lo que garantiza que los coeficientes estimados son estables y fiables.

b) Análisis de Residuos

Para validar el modelo final obtenido, verificaremos los supuestos fundamentales de la Regresión Lineal (Gauss-Markov) mediante diagnóstico visual y contrastes de hipótesis.

Código en R para el diagnóstico:

```r
# configuracion grafica
par(mfrow = c(2, 2))
plot(modelo_final)
par(mfrow = c(1, 1))

# test de normalidad
shapiro_test <- shapiro.test(residuals(modelo_final))
cat("P-valor Shapiro-Wilk:", shapiro_test$p.value, "\n")

# test de homocedasticidad
bp_test <- ncvTest(modelo_final)
cat("P-valor Breusch-Pagan:", bp_test$p, "\n")

# test de independencia
dw_test <- durbinWatsonTest(modelo_final)
cat("P-valor Durbin-Watson:", dw_test$p, "\n")
```

Resultados obtenidos:

1. **Normalidad de los residuos:**
   - *Evidencia visual:* En el gráfico **Normal Q-Q**, los puntos se ajustan casi perfectamente a la línea diagonal discontinua, sin desviaciones severas en las colas.
   - *Evidencia numérica:* El test de **Shapiro-Wilk** arroja un p-valor de **0.634**. Al ser mayor que el nivel de significancia ($\alpha = 0{,}05$), no rechazamos la hipótesis nula. Podemos afirmar que los errores siguen una distribución Normal.
2. **Homocedasticidad (Varianza constante):**
   - *Evidencia visual:* En el gráfico de **Residuals vs Fitted**, la nube de puntos se distribuye de manera aleatoria alrededor de la línea horizontal sin formar patrones de embudo o curvaturas pronunciadas.
   - *Evidencia numérica:* El test de **Breusch-Pagan** devuelve un p-valor de **0.350** ($> 0{,}05$). No hay evidencia estadística para rechazar la homocedasticidad, por lo que la varianza del error es constante.
3. **Independencia:** El test de **Durbin-Watson** presenta un p-valor de **0.454**. Esto indica que no existe autocorrelación significativa entre los residuos.
4. **Puntos influyentes:** En el gráfico **Residuals vs Leverage**, aunque se etiquetan algunas observaciones extremas (como la 72, 116 o 122), ninguna de ellas supera las líneas de la Distancia de Cook (líneas discontinuas rojas en las esquinas, que ni siquiera aparecen en el área visible del gráfico). Esto implica que no hay *outliers* con influencia suficiente para distorsionar los coeficientes del modelo.

Conclusión: El modelo cumple satisfactoriamente con las hipótesis de linealidad, normalidad, homocedasticidad e independencia. Por tanto, el modelo es estadísticamente válido y las inferencias realizadas en el apartado anterior son fiables.

Uso de IA en este ejercicio

He decidido utilizar la IA para que me ayude a evitar la multicolinealidad en el apartado A y para diseñar el análisis de residuos e interpretar los resultados en el apartado B.

- **Prompt utilizado:** " He ejecutado un modelo lineal en R con el dataset 'wine' y he obtenido los siguientes resultados numéricos (Summary, tests de Shapiro-Wilk, Breusch-Pagan) y gráficos de diagnóstico: [Pegar resultados de la consola R]. Basándote en estos datos reales, ¿podrías ayudarme a interpretar si se cumplen las hipótesis del modelo y a redactar las conclusiones técnicas del análisis de residuos?"
- **Integración:** Tras proporcionar a la IA los outputs brutos de la ejecución en R, la herramienta ayudó a confirmar la lectura técnica de los p-valores y los patrones gráficos, facilitando la redacción de las conclusiones sobre la validez de las hipótesis de Gauss-Markov (normalidad y homocedasticidad) presentadas en el ejercicio.

Ejercicio 2

Enunciado: En una tabla 2x2 el valor del estadístico Chi-cuadrado y el del estadístico Chi-Cuadrado de razón de verosimilitudes son asintóticamente equivalentes.

a) Verdadero.

b) Falso.

Respuesta: a) Verdadero.

Justificación

La afirmación es **VERDADERA**.

Para justificarlo, debemos fijarnos en la palabra "asintóticamente". Esto significa que estamos analizando qué ocurre cuando el tamaño de la muestra (n) tiende a infinito (es muy grande).

Tenemos dos formas de medir si dos variables en una tabla de contingencia son independientes:

- El estadístico clásico de Pearson (X^2): compara las diferencias al cuadrado entre lo observado y lo esperado.
- El estadístico de Razón de Verosimilitudes (G^2 o Deviance): utiliza logaritmos de los cocientes entre lo observado y lo esperado.

Matemáticamente, la equivalencia se demuestra utilizando una **Expansión de Taylor**. Si desarrollamos la fórmula del estadístico G^2 mediante una serie de Taylor de segundo orden alrededor del valor 1 (asumiendo que bajo la hipótesis nula lo observado se acerca a lo esperado), los términos logarítmicos se simplifican y la fórmula resultante pasa a ser idéntica a la fórmula del estadístico X^2 de Pearson.

$$G^2 = 2\sum O \ln(O/E) \approx \sum \frac{(O-E)^2}{E} = X^2$$

Por tanto, a medida que aumentamos el tamaño muestral, la diferencia entre el valor de G^2 y X^2 se hace despreciable y ambos convergen a la misma distribución Chi-cuadrado. Sin embargo, es importante notar que para muestras pequeñas esta equivalencia no se cumple tan bien (y de hecho, X^2 suele ser más fiables en esos casos).

Uso de IA en este ejercicio

He utilizado la IA para entender la demostración matemática que hay detrás de esta equivalencia teórica, ya que no recordaba exactamente qué herramienta matemática se usaba para igualar las fórmulas.

- **Prompt utilizado:** "Ayúdame a entender por qué el estadístico G2 (Likelihood Ratio) se parece al Chi-cuadrado de Pearson con muestras grandes. ¿Existe alguna demostración matemática sencilla usando Taylor que pruebe que son equivalentes?"
- **Integración:** La IA me explicó que al aproximar el logaritmo mediante Taylor, el término G^2 se transforma en la suma de diferencias cuadráticas típica de Pearson. He usado esa explicación para redactar la justificación en mis propias palabras.

Ejercicio 3

Enunciado: La multicolinealidad perfecta no se puede dar en un modelo de regresión de Poisson.

a) Verdadero.

b) Falso.

Respuesta: b) Falso.

Justificación

La afirmación es **FALSA**.

La multicolinealidad es un problema relacionado exclusivamente con las variables predictoras (la matriz de diseño X), y no tiene nada que ver con la distribución de la variable respuesta (en este caso, Poisson).

Para entenderlo de forma sencilla: la multicolinealidad perfecta ocurre cuando una variable predictora es una combinación lineal exacta de otra (por ejemplo, si tenemos una variable "Temperatura en Celsius" y otra "Temperatura en Fahrenheit"). Son esencialmente la misma información repetida.

Cuando intentamos ajustar un modelo, ya sea una Regresión Lineal clásica o un Modelo Lineal Generalizado (GLM) como la regresión de Poisson, el algoritmo matemático necesita invertir una matriz derivada de nuestros datos (concretamente la matriz de información de Fisher o X^TWX en el algoritmo IRLS).

- Si existe multicolinealidad perfecta, el determinante de esta matriz es 0.
- Una matriz con determinante 0 es singular y **no se puede invertir**.

Por lo tanto, el problema **sí se puede dar** en la regresión de Poisson. De hecho, si intentas ejecutarlo en R, el software detectará que no puede calcular coeficientes únicos para ambas variables y probablemente lanzará un error o automáticamente eliminará una de las variables (generando valores `NA` en los coeficientes) para poder ajustar el modelo. Que el modelo sea Poisson no lo inmuniza contra tener predictores redundantes.

Uso de IA en este ejercicio

He recurrido a la IA para confirmar si existía alguna excepción teórica en los GLM que no recordara, pero la lógica matricial se mantiene.

- **Prompt utilizado:** "¿Afecta la multicolinealidad perfecta a Poisson igual que a la regresión lineal estándar? ¿O el algoritmo de máxima verosimilitud lo evita?"
- **Integración:** La IA me confirmó que la multicolinealidad perfecta hace que la matriz de diseño no sea de rango completo, impidiendo la estimación única de los coeficientes también en Poisson. He usado esa confirmación para explicar que el problema está en la matriz de las X y no en la distribución de la Y.

Ejercicio 4

Enunciado: Se pide analizar el siguiente conjunto de datos con regresión logística e interpretar el resultado.

```
datos <- expand.grid(Sex=c(1,0), Survival=c(1,0))
datos$frecuencia <- c(233,81,109,468)
```

Código y Análisis en R

Teniendo en cuenta que la función `expand.grid` en R genera todas las combinaciones variando primero la columna `Sex` y que el conjunto de datos esta en formato agregado, la extructura interna de datos es:

- `Sex=1, Survival=1` (Frec: 233)
- `Sex=0, Survival=1` (Frec: 81)
- `Sex=1, Survival=0` (Frec: 109)
- `Sex=0, Survival=0` (Frec: 468)

Dado que no debemos "desagregar" las filas manualmente, usamos el argumento `weights` en la función `glm`, indicando que cada fila representa a n individuos.

```r
datos <- expand.grid(Sex=c(1,0), Survival=c(1,0))
datos$frecuencia <- c(233, 81, 109, 468)

# tabla de contingencia
xtabs(frecuencia ~ Sex + Survival, data=datos)

# Ajuste de GLM binomial modelando en funcion de Sex
modelo_logistico <- glm(Survival ~ Sex,
                        weights = frecuencia,
                        family = binomial(link = "logit"),
                        data = datos)

# vemos los resultados del modelo
summary(modelo_logistico)

# calculamos los odds ratios e intervalos de confianza
exp(cbind(OR = coef(modelo_logistico), confint(modelo_logistico)))
```

Interpretación de los Resultados

A partir de la salida del modelo (`summary` y `confint`), obtenemos las siguientes conclusiones estadísticas:

1. **Análisis de los Coeficientes (β):**
   - **Intercepto ($\beta_0 = -1{,}7540$):** Representa el *log-odds* de supervivencia para el grupo de referencia ($Sex = 0$). Su p-valor es extremadamente significativo ($< 2e^{-16}$).

- **Sexo ($\beta_{Sex} = 2{,}5137$):** El coeficiente es positivo, lo que indica que cambiar de $Sex = 0$ a $Sex = 1$ aumenta la probabilidad de supervivencia.

2. **Interpretación en términos de Probabilidad y Odds Ratio:** Para interpretar la magnitud del efecto, exponenciamos los coeficientes:
   - **Odds del grupo base ($e^{-1{,}7540} \approx 0{,}173$):** Para el grupo $Sex = 0$, el odds de sobrevivir es 0.173. Convertido a probabilidad ($P = \frac{Odds}{1+Odds}$), esto significa que el grupo $Sex = 0$ tiene una **tasa de supervivencia de solo el 14.75 %**.
   - **Odds Ratio del Sexo ($e^{2{,}5137} \approx 12{,}35$):** Este es el valor clave. Indica que la oportunidad (odds) de sobrevivir para el grupo $Sex = 1$ es **12.35 veces mayor** que para el grupo $Sex = 0$.
3. **Intervalos de Confianza (95 %):** El intervalo de confianza para el Odds Ratio es [**8.94, 17.23**]. Como el intervalo no contiene el valor 1 (valor nulo para OR), confirmamos con un 95 % de confianza que el efecto del sexo sobre la supervivencia es positivo y estadísticamente significativo.
4. **Bondad de Ajuste:** Observamos una reducción notable en la *Deviance* (desvianza), pasando de una *Null Deviance* de 1156.39 a una *Residual Deviance* de 887.54. Esta caída confirma que la inclusión de la variable `Sex` mejora sustancialmente el ajuste del modelo en comparación con un modelo nulo (solo intercepto).

Conclusión Final: Existe una asociación fortísima entre la variable Sexo y la Supervivencia. Asumiendo el contexto habitual de este dataset (pasajeros del Titanic, donde $Sex = 1$ suelen ser mujeres y $Sex = 0$ hombres), el modelo demuestra que las mujeres tuvieron una probabilidad de supervivencia drásticamente superior, siendo su ventaja en términos de odds más de 12 veces la de los hombres.

Uso de IA en este ejercicio

He utilizado la IA para validar la interpretación probabilística del modelo y contrastar la relación entre los parámetros estimados y la estructura de la tabla de contingencia subyacente.

- **Prompt utilizado:** "En un GLM binomial con enlace logit sobre datos agregados, ¿cómo se vinculan los coeficientes β con el logaritmo de la Odds Ratio de una tabla 2×2? Además, ayúdame a interpretar el descenso de la Deviance como una medida de la varianza explicada."
- **Integración:** La IA facilitó la transición de la interpretación aritmética a la teórica, permitiéndome identificar que el coeficiente β_{Sex} es exactamente el logaritmo natural de la razón de ventajas cruzada (Cross-product Ratio). Asimismo, la asistencia de la IA fue clave para analizar la *Residual Deviance*, permitiéndome concluir que, aunque el factor `Sex` es altamente significativo y reduce la entropía del modelo, la desvianza residual sugiere que existen otros predictores no observados que podrían explicar la variabilidad remanente en la supervivencia.

Ejercicio 5

Enunciado: Se pide proporcionar un ejemplo en R con una única variable explicativa donde el análisis discriminante cuadrático mejore los resultados claramente del análisis discriminante lineal.

Fundamentación del Ejemplo

El Análisis Discriminante Lineal (LDA) asume que las clases tienen distintas medias pero **igual matriz de covarianza** (homocedasticidad). En una dimensión, esto implica que busca separar las clases mediante un único punto de corte. Por el contrario, el Análisis Discriminante Cuadrático (QDA) permite que cada clase tenga su propia varianza.

Para demostrar la superioridad de QDA, simulamos el caso de **medias idénticas y varianzas distintas** (datos concéntricos):

- **Clase 0:** Datos concentrados en torno al 0 (Media 0, SD 1).
- **Clase 1:** Datos muy dispersos en torno al 0 (Media 0, SD 10).

Dado que las medias coinciden, el LDA no encontrará un punto de separación efectivo. El QDA, sin embargo, podrá definir un intervalo central para la Clase 0 y asignar los extremos a la Clase 1.

Código en R

```r
library(MASS)
set.seed(123)

# 500 observaciones por clase
n <- 500

# clase 0: distribución estrecha
x1 <- rnorm(n, mean = 0, sd = 1)
g1 <- rep(0, n)

# clase 1: distribución ancha
x2 <- rnorm(n, mean = 0, sd = 10)
g2 <- rep(1, n)

# unimos los datos
x <- c(x1, x2)
grupo <- factor(c(g1, g2))
datos <- data.frame(x, grupo)

# ajustamoslos modelos
# LDA
modelo_lda <- lda(grupo ~ x, data = datos)

# QDA
modelo_qda <- qda(grupo ~ x, data = datos)

```

```r
# predicciones
pred_lda <- predict(modelo_lda)$class
pred_qda <- predict(modelo_qda)$class

# calculamos accuracy
acc_lda <- mean(pred_lda == datos$grupo)
acc_qda <- mean(pred_qda == datos$grupo)

cat("Precisión LDA:", round(acc_lda * 100, 2), "%\n")
cat("Precisión QDA:", round(acc_qda * 100, 2), "%\n")
```

Interpretación de los Resultados

Tras ejecutar la simulación con la semilla fijada, los resultados numéricos confirman drásticamente la teoría:

- **Precisión LDA: 50.20 % (Ineficacia total).** El modelo lineal ha obtenido un resultado estadísticamente equivalente a lanzar una moneda al aire (azar). **Causa:** Como ambas clases tienen la media en 0 ($\bar{x}_0 \approx \bar{x}_1 \approx 0$), el LDA es incapaz de encontrar un punto de corte único que separe los grupos. Al no poder modelar la diferencia de dispersión, asigna las clases casi aleatoriamente o asigna todo a una clase, fallando estrepitosamente.
- **Precisión QDA: 89.20 % (Mejora sustancial).** El modelo cuadrático mejora el rendimiento en casi 40 puntos porcentuales. **Causa:** Al permitir varianzas distintas, el QDA detecta que la Clase 0 es .^estrecha"($SD = 1$) y la Clase 1 es .^ancha"($SD = 10$). En lugar de un punto de corte, el QDA genera una frontera de decisión cuadrática (un intervalo). El modelo aprende correctamente la regla: *"Si el dato está muy cerca de 0, es Clase 0; si se aleja hacia los extremos, es Clase 1"*.

Conclusión: Este ejercicio demuestra empíricamente que cuando la información discriminante reside en la **varianza** y no en la media (como en distribuciones concéntricas), el LDA es inservible y el uso de QDA se vuelve imprescindible.

Uso de IA en este ejercicio

He utilizado la IA para diseñar el escenario de simulación. Sabía que debían tener varianzas distintas, pero quería confirmar cómo configurar los parámetros para maximizar la diferencia de resultados.

- **Prompt utilizado:** "Dame un ejemplo en R con una sola variable X donde QDA tenga un accuracy mucho mayor que LDA. ¿Qué distribución deben seguir los datos?"
- **Integración:** La IA sugirió el escenario de "medias iguales, varianzas desiguales"($N(0,1)$ vs $N(0,10)$). Explicó que en este caso el LDA colapsa porque los centros de gravedad coinciden. He implementado ese código y añadido la interpretación sobre el ïntervalo"vs "punto de corte".

Ejercicio 6

Enunciado: ¿Se puede decir que en los datos del Titanic se da el fenómeno de la confusión?

Respuesta: Sí, rotundamente.

Justificación y Código en R

El fenómeno de confusión ocurre cuando una tercera variable distorsiona la asociación real entre una exposición y un desenlace. En el Titanic, el **Sexo** actúa como un fuerte factor de confusión en la relación entre **Clase** y **Supervivencia**.

Para demostrarlo, observamos primero la distribución de sexos y posteriormente comparamos los coeficientes de dos modelos:

1. **Modelo Crudo:** Supervivencia $\sim$ Clase (ignora el sexo).
2. **Modelo Ajustado:** Supervivencia $\sim$ Clase + Sexo (controla el efecto del sexo).

Si el coeficiente de la clase cambia sustancialmente (regla general $> 10\,\%$) al añadir el sexo, confirmamos la confusión.

```r
# carga de datos
data("Titanic")
df <- as.data.frame(Titanic)

# visualizamos la desproporción de sexos
xtabs(Freq ~ Class + Sex, data = df)

# modelo 1  estimación cruda
modelo_crudo <- glm(Survived ~ Class,
                    weights = Freq,
                    data = df,
                    family = binomial)

# modelo 2: estimación ajustada por sexo
modelo_ajustado <- glm(Survived ~ Class + Sex,
                       weights = Freq,
                       data = df,
                       family = binomial)

# comparamos OR
print(exp(coef(modelo_crudo)))
print(exp(coef(modelo_ajustado)))

# calculo del cambio porcentual para crew
beta_crudo <- coef(modelo_crudo)["ClassCrew"]
beta_ajustado <- coef(modelo_ajustado)["ClassCrew"]
cambio <- abs((beta_crudo - beta_ajustado) / beta_ajustado) * 100

cat("\nCambio porcentual en el coeficiente de 'Crew':", round(cambio,
    2), "%\n")
```

Interpretación de los Resultados

La ejecución del código proporciona evidencias claras del fenómeno:

1. **El mecanismo de la confusión (Tabla cruzada):** Observamos un desbalance extremo en la clase *Crew*: hay **862 hombres** frente a solo **23 mujeres**. Dado que las mujeres tenían prioridad en los botes, la tripulación parece tener una mortalidad altísima simplemente porque está compuesta casi exclusivamente por hombres.
2. **Comparación de Odds Ratios (OR):**
   - **Modelo Crudo:** El OR para *ClassCrew* es **0.189**. Esto sugería que la probabilidad de sobrevivir de la tripulación era ínfima en comparación con la 1ª Clase.
   - **Modelo Ajustado:** Al controlar por sexo, el OR para *ClassCrew* sube a **0.414**. Aunque sigue siendo bajo, es más del doble que en el modelo crudo. Esto indica que "ser tripulante"no era tan letal como parecía; lo letal era "ser hombre".
3. **Cambio en los coeficientes ($\Delta\beta$):** El cambio porcentual en el coeficiente de la tripulación entre el modelo crudo y el ajustado es del **88.96 %**.

   Conclusión: Dado que el cambio es drásticamente superior al umbral estándar del 10 %, confirmamos estadísticamente que **existe confusión**. El modelo crudo estaba sesgado: subestimaba la supervivencia de la tripulación al atribuirle a la "Clase" un riesgo que en realidad pertenecía al "Sexo".

Uso de IA en este ejercicio

He utilizado la IA para confirmar cuál es la combinación de variables más evidente para enseñar el concepto de confusión en este dataset específico.

- **Prompt utilizado:** "En el dataset Titanic de R, ¿cuál es el mejor ejemplo para demostrar 'confounding' (confusión)? ¿Es la relación Clase->Supervivencia confundida por Sexo? Dame el código para comparar los OR crudos vs ajustados."
- **Integración:** La IA me sugirió centrarme en la discrepancia de la clase "Crew", ya que es el grupo con mayor desbalance de género. He seguido esa línea argumental para el código, comparando el modelo con y sin la variable de ajuste.