TAREA 1- SHIPS

María Pallares Diez

2025-04-29

Introducción

El objetivo de esta tarea es modelizar el número de incidentes sufridos por buques de carga debido a daños ocasionados por las olas, en función de diversas características del barco y su periodo de servicio. Para ello se emplea el conjunto de datos `ships.sav`, que recoge información agregada sobre buques de distintos tipos, años de construcción, periodos de operación, duración del servicio (en meses) y número total de incidentes registrados.

Dado que la variable respuesta es un conteo de sucesos y que los barcos presentan tiempos de exposición distintos, el enfoque adecuado es utilizar un **modelo lineal generalizado con distribución Poisson y función de enlace logarítmica**, incorporando un **offset** con el logaritmo de los meses de servicio para ajustar por la exposición.

Presentación de las variables

La variable respuesta del estudio es `damage_incidents`, que representa el número de incidentes por daño registrados. Se trata de una **variable cuantitativa discreta**, ya que cuenta el número de sucesos ocurridos en cada combinación de barco y periodo.

Las variables explicativas son las siguientes:

- `type`: indica el tipo de barco con números del 1 al 5. Es una **variable categórica nominal**.
- `construction`: representa el año de construcción del barco, desde el 60 al 75. Se considera una **variable categórica ordinal**, ya que sigue un orden temporal.
- `operation`: hace referencia al periodo de operación del barco. También es una **variable categórica ordinal**.
- `months_service`: número total de meses en servicio acumulado. Es una **variable cuantitativa continua**, utilizada para calcular la exposición.
- `log_months_service`: logaritmo del número de meses de servicio, que se utilizará como **offset** en el modelo para ajustar por la exposición temporal.

Todas estas variables se incorporarán en el modelo propuesto con el objetivo de explicar la variabilidad observada en el número de incidentes.

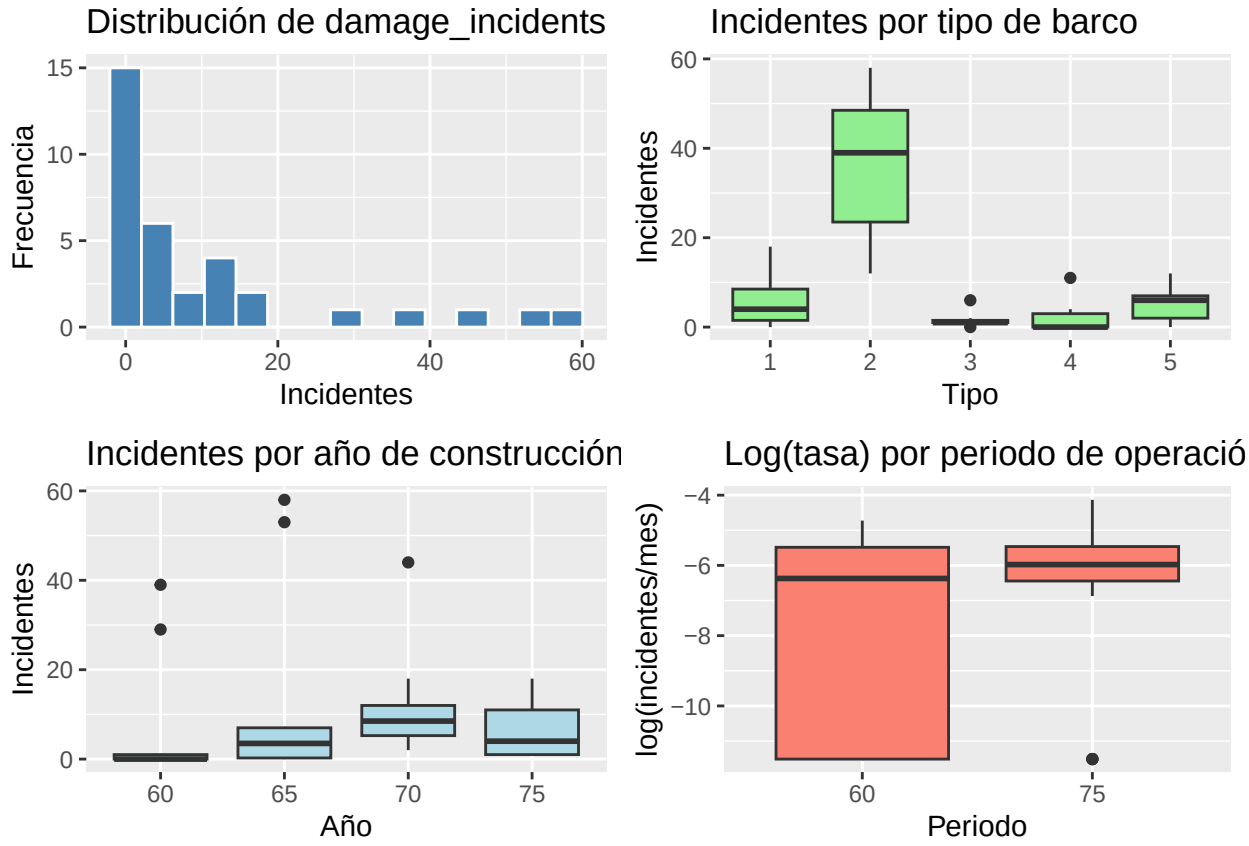
Variables numéricas

Variable	Mín	Q1	Mediana	Media	Q3	Máx	NA's
damage_incidents	0	1.00	4.00	10.47	11.75	58.00	6
months_service	0.0	175.8	782.0	4089.3	2078.5	44882.0	0

Variable	Mín	Q1	Mediana	Media	Q3	Máx	NA's
log_months_service	3.81	5.91	6.99	7.05	7.71	10.71	6

Variables categóricas

Variable	Categorías (Frecuencia)
type	1–5 (8 casos cada uno)
construction	60, 65, 70, 75 (10 casos cada uno)
operation	60, 75 (20 casos cada uno)



A partir de los cuatro gráficos exploratorios, se pueden destacar las siguientes observaciones:

- **Distribución de `damage_incidents`**: La mayoría de los barcos presenta pocos o ningún incidente. Se observa una distribución asimétrica con valores extremos (outliers) que alcanzan hasta 58 incidentes. Esto sugiere una fuerte dispersión y la presencia de pocos barcos con muchos eventos.
- **Incidentes por tipo de barco (`type`)**: Los barcos del tipo 2 tienen una mediana mucho más alta de incidentes en comparación con los otros tipos, y muestran una variabilidad considerable. El resto de los tipos presentan distribuciones mucho más concentradas y con menor número de incidentes.
- **Incidentes por año de construcción (`construction`)**: Se observa un aumento progresivo en la mediana y dispersión de los incidentes a medida que avanza el año de construcción. Los barcos construidos en 60 presentan valores bajos, mientras que los de 75 presentan más incidentes, en promedio.

- **Log(tasa de incidentes) por periodo de operación (`operation`)**: Los barcos operando en el año 75 muestran una mayor log(tasa) de incidentes que los del año 60. Esto sugiere que el periodo de operación podría estar asociado a un mayor riesgo ajustado por el tiempo de exposición.

Modelización como Regresión de Poisson

La variable respuesta `damage_incidents` representa un **conteo de eventos discretos no negativos**, por lo que se asume que sigue una **distribución de Poisson**. Esta distribución es adecuada para modelar el número de ocurrencias de un evento en un intervalo de tiempo o espacio, bajo el supuesto de independencia entre eventos y baja probabilidad individual.

- **Distribución asumida**:
  $Y_i \sim \text{Poisson}(\mu_i)$, donde Y_i es el número de incidentes para el barco i.
- **Parámetro de interés**:
  μ_i, que representa el **número esperado de incidentes** para el barco i. Este parámetro depende de las características del barco (tipo, año de construcción y periodo de operación), así como del tiempo de servicio (exposición).

Como estamos analizando un conteo, utilizamos un ajuste con una **regresión de Poisson**:

- **Variable respuesta**: Y_i, número de incidentes de daño sufridos por los barcos para $i = 1, \ldots, 34$ combinaciones barco-período.

$$Y_i \sim \text{Poisson}(\mu_i), \quad i = 1, \ldots, 34$$

- La **componente sistemática** está formada por 3 variables explicativas categóricas del estudio: `type`, `construction` y `operation`.
- Se dispone también de una covariable de exposición cuantitativa (`months_service`), que se incorpora como **offset** mediante su logaritmo.
- Utilizamos la transformación **logarítmica** como **función de enlace** (*link*) para relacionar el predictor lineal con la media de la respuesta.

$$\boxed{g(\mu_i) = \log(\mu_i)}$$

La forma funcional del modelo es la siguiente:

$$\log(\mu_i) = \beta_0 + \beta_1 \cdot \text{type}_i + \beta_2 \cdot \text{construction}_i + \beta_3 \cdot \text{operation}_i + \log(\text{months\_service}_i), \quad i = 1, \ldots, 34$$

El término $\log(\text{months\_service}_i)$ se incluye como **offset** para ajustar por el tiempo de exposición de cada barco. El modelo resultante permite interpretar los efectos de las covariables sobre la **tasa de incidentes por unidad de tiempo de servicio**.

Resolución del modelo inicial y evaluación del método

Se ajustó un modelo lineal generalizado (GLM) con distribución de Poisson, función de enlace logarítmica y offset correspondiente al logaritmo del número de meses en servicio (`log_months_service`), lo que permite modelar tasas de incidentes por unidad de tiempo. Las variables explicativas son `type`, `construction` y `operation`, todas categóricas. Tras eliminar 6 observaciones con valores perdidos, se dispuso de 34 casos válidos.

Resultados del ajuste

El modelo revela efectos significativos en los niveles 2 y 3 de `type`, indicando tasas de incidentes más bajas respecto al grupo de referencia (tipo 1). Para `construction`, los años 65 y 70 presentan coeficientes positivos altamente significativos, lo que implica tasas considerablemente más elevadas. También se observa un efecto relevante en `operation`, con una mayor tasa de incidentes en los barcos operando en 1975.

El valor de la devianza residual es 38.7 con 25 grados de libertad, en comparación con la devianza nula de 146.3 con 33 grados de libertad. Esto implica que el modelo explica aproximadamente un 73.6% de la variabilidad inicial. El test global basado en la devianza residual ofrece un p-valor de 0.039, lo que sugiere un ajuste adecuado sin evidencias claras de falta de ajuste.

Análisis gráfico

Los gráficos diagnósticos muestran resultados consistentes con un modelo bien ajustado. Los residuos de devianza frente a los valores ajustados no presentan patrones evidentes. El histograma de residuos indica ligera asimetría sin valores extremos destacables, mientras que el gráfico Q-Q refleja una leve desviación en las colas, esperable en modelos de Poisson. En conjunto, los resultados gráficos refuerzan la validez del modelo ajustado.

Evaluación de la sobredispersión

Se calcula el cociente entre la suma de residuos de Pearson al cuadrado y los grados de libertad del modelo, obteniendo un valor de 1.31, ligeramente superior a 1 pero sin indicar un problema grave. Un test formal de sobredispersión basado en un estadístico tipo z confirma que no hay evidencia significativa de sobredispersión ($p = 0.178$).

Intentando mejorar el modelo inicial

Dado que todas las variables son categóricas, se planteó evaluar interacciones de primer orden que podrían reflejar efectos combinados entre factores. Se consideraron las siguientes:

- **type:construction** – para analizar si el impacto del tipo de barco cambia según su año de construcción.
- **type:operation** – para detectar si el efecto del tipo varía entre los dos periodos de operación.
- **construction:operation** – para identificar posibles interacciones entre antigüedad y condiciones de uso.

No se incluyen transformaciones adicionales ni se modifica el offset, ya que no hay variables cuantitativas susceptibles de ello. El modelo extendido con todas las interacciones es comparado con el modelo inicial mediante análisis de devianza, AIC y validación cruzada.

Selección del modelo más adecuado en términos de ajuste

Se compararon tres modelos con distribución Poisson y offset logarítmico:

- **Modelo nulo**: solo intercepto y offset.
- **Modelo inicial**: efectos aditivos de `type`, `construction` y `operation`.
- **Modelo con interacciones**: todas las interacciones de primer orden entre las tres variables.

Modelo	Df	Deviance	AIC
Modelo nulo	33	146.33	246.19
Modelo inicial	25	38.70	**154.56**
Modelo con interacciones	7	6.86	158.72

Contraste de devianza

Comparación 1: Modelo nulo vs inicial

- **H0**: el modelo nulo se ajusta igual de bien.
- **H1**: el modelo inicial mejora significativamente.
- **Estadístico**: D = 146.33 - 38.70 = 107.63 con gl = 8

- **p-valor**: $< 0.0001 \rightarrow$ Se rechaza H0.

Comparación 2: Modelo inicial vs con interacciones

- **H0**: el modelo inicial se ajusta igual de bien.
- **H1**: el modelo con interacciones mejora significativamente.
- **Estadístico**: D = 38.70 - 6.86 = 31.84 con gl = 18

- **p-valor**: 0.02297 $\rightarrow$ Se rechaza H0.

Comparación mediante AIC

El modelo inicial presenta el AIC más bajo, por lo que resulta más parsimonioso. Aunque el modelo con interacciones ajusta mejor, el incremento en complejidad no compensa la ligera mejora.

Proporción de devianza explicada

- Modelo inicial: $1 - \frac{38.70}{146.33} \approx 73.6\%$
- Modelo con interacciones: $1 - \frac{6.86}{146.33} \approx 95.3\%$

El modelo con interacciones explica más variabilidad, pero a costa de un ajuste posiblemente sobredimensionado.

Comparación mediante validación cruzada

Se empleó validación cruzada tipo LOOCV para estimar la capacidad predictiva de los modelos. Se calcularon los errores medios de predicción:

Modelo	Error LOOCV
Modelo nulo	219.21
Modelo inicial	**28.15**
Modelo con interacciones	2.03 $\times$ 10^7

El modelo inicial obtiene el menor error de predicción. El modelo con interacciones muestra un error extremadamente alto, reflejo de un sobreajuste evidente.

Conclusión: el modelo inicial es el más adecuado, pues combina buen ajuste, capacidad predictiva y parsimonia.

Interpretación de los coeficientes del mejor modelo

El modelo final seleccionado es el **modelo inicial**, que incluye las variables categóricas `type`, `construction` y `operation`. El modelo es un GLM con distribución Poisson y enlace logarítmico, e incorpora como offset el logaritmo del número de meses en servicio (`log_months_service`), para modelar tasas de incidentes.

El modelo estimado es:

$$\log(\mu_i) = \log(\text{months}_i) + \beta_0 + \beta_{\text{type}_i} + \beta_{\text{construction}_i} + \beta_{\text{operation}_i}$$

donde μ_i es el número esperado de incidentes para la observación i, ajustado por su tiempo de exposición.

En un modelo de regresión Poisson con enlace logarítmico, los coeficientes estimados representan el **logaritmo de la razón de tasas** entre niveles de una variable. Para interpretar estos coeficientes en términos porcentuales, se aplica la transformación:

$$\text{Cambio porcentual} = \left(e^{\beta} - 1\right) \times 100$$

- Si el resultado es positivo, indica un **aumento porcentual** en la tasa de incidentes respecto al grupo de referencia.
- Si es negativo, indica una **disminución porcentual**.

Variable	Estimación	p-valor	Interpretación
(Intercept)	-6.406	< 0.0001	Tasa base de incidentes por mes en barcos tipo 1, construidos antes de 1960 y operando en los años 60.
`type2`	-0.543	0.0022	Los barcos tipo 2 tienen una tasa de incidentes un 42% menor que los tipo 1.
`type3`	-0.687	0.0367	Los barcos tipo 3 tienen una tasa de incidentes un 50% menor que los tipo 1.
`type4`	-0.076	0.7938	Los barcos tipo 4 no presentan diferencias significativas frente a los tipo 1.
`type5`	0.326	0.1675	Los barcos tipo 5 tienen una tasa un 39% mayor que los tipo 1, pero no es significativa.
`construction65`	0.697	< 0.0001	Los barcos construidos en 1965 tienen una tasa de incidentes aproximadamente el doble que los construidos antes de 1960.
`construction70`	0.818	< 0.0001	Los barcos de 1970 tienen una tasa 2.3 veces mayor que los construidos antes de 1960.
`construction75`	0.453	0.0518	Los barcos de 1975 tienen una tasa un 57% mayor que los anteriores a 1960 (resultado marginalmente significativo).
`operation75`	0.384	0.0012	Los barcos operando en los años 70 tienen una tasa de incidentes un 47% mayor que los que operaban en los años 60.

- Los tipos de barco 2 y 3 presentan **una reducción significativa en la tasa de incidentes** respecto al tipo 1.
- La **fecha de construcción** tiene un efecto positivo claro: **barcos más nuevos presentan tasas de incidentes más altas**, lo que puede deberse a su uso más intensivo o condiciones operativas distintas.
- La **variable `operation` también muestra un efecto positivo significativo**, indicando un mayor riesgo de incidentes en los años 70.

En conjunto, el modelo permite identificar **perfiles de mayor o menor riesgo** en función del tipo de barco, su antigüedad y su periodo de operación, lo que puede ser útil para la toma de decisiones en mantenimiento y prevención.

Predicciones

Una vez validado el modelo inicial como el más adecuado en términos de ajuste y capacidad predictiva, procedemos a realizar **predicciones del número esperado de incidentes** para nuevas combinaciones de características de los barcos. Este paso permite aplicar el modelo ajustado a situaciones futuras o no observadas, facilitando la toma de decisiones preventivas.

Recordemos que, en un modelo de regresión de Poisson con función de enlace logarítmica y offset, la predicción del valor esperado de la variable respuesta μ_i se obtiene mediante:

$$\hat{\mu}_i = \exp\left(\hat{\beta}_0 + \hat{\beta}_{\text{type}_i} + \hat{\beta}_{\text{construction}_i} + \hat{\beta}_{\text{operation}_i}\right) \cdot \text{months\_service}_i$$

El término de offset $\log(\text{months\_service}_i)$ permite ajustar la predicción por el tiempo de exposición, es decir, por el número de meses que ha estado en servicio el barco.De este modo, el modelo estima el número esperado de incidentes considerando tanto las características del barco como su exposición.

Predicciones realizadas

Se evalúan dos perfiles distintos de barcos, ambos presentes en los datos de entrenamiento y compatibles con la estructura del modelo:

- **Barco A**: tipo 1, construido en 1960, operando en el periodo 1960–74, con 10.000 meses de servicio.
- **Barco B**: tipo 4, construido en 1970, operando en el periodo 1975–79, con 2.000 meses de servicio.

El modelo ajustado estima los siguientes valores esperados de incidentes para cada uno:

- **Barco A**: se predicen aproximadamente **16,52 incidentes**.
- **Barco B**: se predicen aproximadamente **10,20 incidentes**.

Interpretación

Ambas predicciones reflejan el número esperado de incidentes de daño sufridos por cada barco, teniendo en cuenta sus características estructurales y operativas, así como su tiempo total de actividad (exposición).

La predicción es coherente con el marco teórico del modelo de Poisson: los efectos de las variables explicativas modifican la tasa de incidentes por unidad de tiempo, y el offset multiplica esta tasa por el tiempo total para obtener el número total esperado.

Estas estimaciones son útiles para prever riesgos, planificar tareas de mantenimiento y priorizar medidas de seguridad en función del perfil del barco.