

CENTRO DE INVESTIGACIÓN EN MATEMÁTICAS (CIMAT)
UNIDAD MONTERREY

Computo Estadístico

Tarea 2

Modelos Lineales Generalizados para Datos Categóricos y de Conteo

Diego Paniagua Molina
diego.paniagua@cimat.mx

21 de Septiembre del 2025



Problema 1

La siguiente tabla muestra los resultados parciales de dos encuestas que forman parte de un estudio para evaluar el desempeño del Primer Ministro del Canadá. Se tomó una muestra aleatoria de 1600 ciudadanos canadienses mayores de edad y en los renglones se observa que 944 ciudadanos aprobaban el desempeño del funcionario, mientras que las columnas muestran que, seis meses después de la primera encuesta, sólo 880 aprueban su desempeño.

Primera encuesta	Segunda encuesta		Total
	Y = 1, Aprueba	Y = 0, Desaprueba	
x = 1, Aprueba	794	150	944
x = 0, Desaprueba	86	570	656
Total	880	720	1600

a) Considere el modelo de regresión logística

$$\log \left(\frac{P(Y_i = 1|x_i)}{1 - P(Y_i = 1|x_i)} \right) = \beta_0 + \beta_1 x_i$$

Escriba la logverosimilitud correspondiente. Muestre explícitamente (i.e. maximizando la logverosimilitud), que el estimador máximo verosimilitud para β_1 es el logaritmo de la tasa de momios de la tabla dada (en general, en regresión logística los estimadores de máxima verosimilitud no tienen una forma explícita, sin embargo, en el presente caso si).

b) Sea p_1 la proporción de ciudadanos que aprueban el desempeño del ministro al tiempo inicial y sea p_2 la proporción correspondiente seis meses después. Considere la hipótesis $H_0 : p_1 = p_2$, ¿Cómo puede hacerse esta prueba?

SOLUCIÓN

a) Para mostrar que el estimador de máxima verosimilitud para β_1 es el logaritmo de la tasa de momios de la tabla dada, comencemos despejando del modelo de regresión logística las probabilidades condicionales del modelo:

$$\begin{aligned}
 \frac{P(Y_i = 1|x_i)}{1 - P(Y_i = 1|x_i)} &= e^{\beta_0 + \beta_1 x_i} \\
 P(Y_i = 1|x_i) &= [1 - P(Y_i = 1|x_i))]e^{\beta_0 + \beta_1 x_i} \\
 P(Y_i = 1|x_i) &= e^{\beta_0 + \beta_1 x_i} - P(Y_i = 1|x_i)e^{\beta_0 + \beta_1 x_i} \\
 P(Y_i = 1|x_i) + P(Y_i = 1|x_i)e^{\beta_0 + \beta_1 x_i} &= e^{\beta_0 + \beta_1 x_i} \\
 P(Y_i = 1|x_i)(1 + e^{\beta_0 + \beta_1 x_i}) &= e^{\beta_0 + \beta_1 x_i} \\
 \Rightarrow P(Y_i = 1|x_i) &= \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \tag{1}
 \end{aligned}$$

Tal que, podemos reescribir la probabilidad de que $Y_i = 1$ de la siguiente forma:

$$P(Y_i = 1|x_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}} \tag{2}$$

Ahora, para calcular la probabilidad de que $Y_i = 0$ podemos obtener el complemento:

$$\begin{aligned}
 P(Y_i = 0|x_i) &= 1 - P(Y_i = 1|x_i) \\
 &= 1 - \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}} \\
 &= \frac{1}{1 + e^{\beta_0 + \beta_1 x_i}}
 \end{aligned} \tag{3}$$

Por otro parte, por cada ciudadano i observamos un par $n_{xy} \equiv (x_i, y_i)$ donde:

$x_i \in \{0, 1\}$: Aprueba (1) o no (0) en la primer encuesta.

$y_i \in \{0, 1\}$: Aprueba (1) o no (0) en la segunda encuesta.

Entonces, siendo la variable de respuesta $Y_i \in \{0, 1\}$ una variable dicotomica, por ende podemos asumir que sigue una distribución condicional Bernoulli con respecto a la covariable x_i , es decir:

$$Y_i|x_i \sim \text{Bernoulli}(p(x_i))$$

Siendo su función de masa de probabilidad:

$$P(Y = y_i|x_i) = p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

Por lo tanto, tenemos que la forma general de la función de verosimilitud para esta variable aleatoria tipo Bernoulli toma la siguiente forma:

$$\begin{aligned}
 L(\beta_0, \beta_1) &= \prod_{i=1}^n P(Y_i = y_i|x_i) \\
 &= \prod_{i=1}^n [P(Y_i = 1|x_i)]^{y_i} [P(Y_i = 0|x_i)]^{1-y_i}
 \end{aligned} \tag{4}$$

Siendo la log-verosimilitud:

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^n \{y_i \log[P(Y_i = 1|x_i)] + (1 - y_i) \log[P(Y_i = 0|x_i)]\} \tag{5}$$

Para simplificar la notación definamos las probabilidades P_{xy} utilizando (2) y (3):

$$P_{11} \equiv P(Y_i = 1|x_i = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1)}} \tag{6}$$

$$P_{01} \equiv P(Y_i = 1|x_i = 0) = \frac{1}{1 + e^{-\beta_0}} \tag{7}$$

$$P_{10} \equiv P(Y_i = 0|x_i = 1) = \frac{1}{1 + e^{\beta_0 + \beta_1}} \tag{8}$$

$$P_{00} \equiv P(Y_i = 0|x_i = 0) = \frac{1}{1 + e^{\beta_0}} \tag{9}$$

Podemos ver que estas expresiones cumplen las siguientes propiedades:

$$P_{11} = 1 - P_{10} \quad (10)$$

$$P_{00} = 1 - P_{01} \quad (11)$$

$$P_{10} = 1 - P_{11} \quad (12)$$

$$P_{01} = 1 - P_{00} \quad (13)$$

Explícitamente tenemos que:

$$\begin{aligned} P_{11} &= 1 - \frac{1}{1 + e^{\beta_0 + \beta_1}} = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} \\ P_{00} &= 1 - \frac{1}{1 + e^{-\beta_0}} = \frac{e^{-\beta_0}}{1 + e^{-\beta_0}} \\ P_{10} &= 1 - \frac{1}{1 + e^{-(\beta_0 + \beta_1)}} = \frac{e^{-(\beta_0 + \beta_1)}}{1 + e^{-(\beta_0 + \beta_1)}} \\ P_{01} &= 1 - \frac{1}{1 + e^{\beta_0}} = \frac{e^{\beta_0}}{1 + e^{\beta_0}} \end{aligned}$$

Entonces, dadas estas cuatro posibles combinaciones de probabilidades, podemos escribir la log-verosimilitud en términos de ellas utilizando los conteos n_{xy} para cada caso, tal que:

$$\ell(\beta_0, \beta_1) = n_{11} \log(P_{11}) + n_{01} \log(P_{01}) + n_{10} \log(P_{10}) + n_{00} \log(P_{00}) \quad (14)$$

Sustituyendo de (6)-(9) y aplicando que $\log(a/b) = \log(a) - \log(b)$ y que $\log(1) = 0$:

$$\begin{aligned} \ell &= n_{11} \log\left(\frac{1}{1 + e^{-(\beta_0 + \beta_1)}}\right) + n_{01} \log\left(\frac{1}{1 + e^{-\beta_0}}\right) + n_{10} \log\left(\frac{1}{1 + e^{\beta_0 + \beta_1}}\right) + n_{00} \log\left(\frac{1}{1 + e^{\beta_0}}\right) \\ &= -n_{11} \log(1 + e^{-(\beta_0 + \beta_1)}) - n_{01} \log(1 + e^{-\beta_0}) - n_{10} \log(1 + e^{\beta_0 + \beta_1}) - n_{00} \log(1 + e^{\beta_0}) \end{aligned}$$

Para maximizar la log-verosimilitud derivamos e igualamos a cero, comenzando con β_0 :

$$\frac{\partial \ell}{\partial \beta_0} = n_{11} \left(\frac{e^{-(\beta_0 + \beta_1)}}{1 + e^{-(\beta_0 + \beta_1)}} \right) + n_{01} \left(\frac{e^{-\beta_0}}{1 + e^{-\beta_0}} \right) - n_{10} \left(\frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} \right) - n_{00} \log\left(\frac{e^{\beta_0}}{1 + e^{\beta_0}}\right) = 0$$

Reescribiendo tenemos en términos de P_{xy} y utilizando de (10)-(13) tenemos que:

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_0} &= n_{11} P_{10} + n_{01} P_{00} - n_{10} P_{11} - n_{00} P_{01} = 0 \\ &= n_{11} P_{10} + n_{01} P_{00} - n_{10} (1 - P_{10}) - n_{00} (1 - P_{00}) = 0 \\ &= n_{11} P_{10} + n_{01} P_{00} - n_{10} + n_{10} P_{10} - n_{00} + n_{00} P_{00} = 0 \\ &= P_{10} (n_{11} + n_{10}) + P_{00} (n_{01} + n_{00}) - n_{10} - n_{00} = 0 \\ &= \left(\frac{1}{1 + e^{\beta_0 + \beta_1}} \right) (n_{11} + n_{10}) + \left(\frac{1}{1 + e^{\beta_0}} \right) (n_{01} + n_{00}) - n_{10} - n_{00} = 0 \end{aligned} \quad (15)$$

Para poder continuar con el desarrollo necesitamos encontrar alguna expresión, ya sea para β_0 o β_1 .

Ahora calculemos la derivada parcial con respecto a β_1 e igualemos a cero para encontrar el máximo:

$$\frac{\partial \ell}{\partial \beta_1} = n_{11} \left(\frac{e^{-(\beta_0 + \beta_1)}}{1 + e^{-(\beta_0 + \beta_1)}} \right) - n_{10} \left(\frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} \right) = 0$$

Reescribiendo nuevamente en términos de P_{xy} tenemos que:

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_1} &= n_{11}P_{10} - n_{10}P_{11} = 0 \\ &= n_{11}P_{10} - n_{10}(1 - P_{10}) = 0 \\ &= n_{11}P_{10} - n_{10} + n_{10}P_{10} = 0 \\ &= P_{10}(n_{11} + n_{10}) - n_{10} = 0 \end{aligned}$$

Despejando P_{10} y sustituyendo su forma explicita:

$$\begin{aligned} \Rightarrow P_{10} &= \frac{n_{10}}{n_{11} + n_{10}} \\ \frac{1}{1 + e^{\beta_0 + \beta_1}} &= \frac{n_{10}}{n_{11} + n_{10}} \\ 1 + e^{\beta_0 + \beta_1} &= \frac{n_{11} + n_{10}}{n_{10}} \\ 1 + e^{\beta_0 + \beta_1} &= \frac{n_{11}}{n_{10}} + 1 \\ e^{\beta_0 + \beta_1} &= \frac{n_{11}}{n_{10}} \\ \beta_0 + \beta_1 &= \log \left(\frac{n_{11}}{n_{10}} \right) \\ \Rightarrow \beta_1 &= \log \left(\frac{n_{11}}{n_{10}} \right) - \beta_0 \end{aligned} \tag{16}$$

Sustituimos esta expresión de β_1 en (15), tal que:

$$\begin{aligned} \left[\frac{1}{1 + e^{\beta_0 + \log(n_{11}/n_{10}) - \beta_0}} \right] (n_{11} + n_{10}) + \left(\frac{1}{1 + e^{\beta_0}} \right) (n_{01} + n_{00}) &= n_{10} + n_{00} \\ \left[\frac{1}{1 + \left(\frac{n_{11}}{n_{10}} \right)} \right] (n_{11} + n_{10}) + \left(\frac{1}{1 + e^{\beta_0}} \right) (n_{01} + n_{00}) &= n_{10} + n_{00} \\ \left(\frac{n_{10}}{n_{11} + n_{10}} \right) (n_{11} + n_{10}) + \left(\frac{1}{1 + e^{\beta_0}} \right) (n_{01} + n_{00}) &= n_{10} + n_{00} \\ n_{10} + \left(\frac{1}{1 + e^{\beta_0}} \right) (n_{01} + n_{00}) &= n_{10} + n_{00} \\ \left(\frac{1}{1 + e^{\beta_0}} \right) (n_{01} + n_{00}) &= n_{00} \end{aligned}$$

Continuando con el desarrollo:

$$\begin{aligned}
 \left(\frac{1}{1 + e^{\beta_0}} \right) &= \frac{n_{00}}{n_{01} + n_{00}} \\
 1 + e^{\beta_0} &= \frac{n_{01}}{n_{00}} + 1 \\
 e^{\beta_0} &= \frac{n_{01}}{n_{00}} \\
 \Rightarrow \beta_0 &= \log \left(\frac{n_{01}}{n_{00}} \right)
 \end{aligned} \tag{17}$$

Por lo tanto, sustituyendo (17) en (16):

$$\beta_1 = \log \left(\frac{n_{11}}{n_{10}} \right) - \log \left(\frac{n_{01}}{n_{00}} \right)$$

Aplicando propiedades de los logaritmos queda demostrado que β_1 es el logaritmo de la tasa de momios de una tabla de contingencia 2×2 :

$$\beta_1 = \log \left(\frac{n_{11}/n_{10}}{n_{01}/n_{00}} \right)$$

b) Se nos pide probar la hipótesis nula:

$$H_0 : p_1 = p_2 \tag{18}$$

Podemos calcular las proporciones p_1 y p_2 directamente de la tabla de contingencia dada, tal que:

$$\begin{aligned}
 p_1 &= \frac{\text{total aprobaron en 1er encuesta}}{\text{total de encuestados}} = \frac{944}{1600} = 0.55 \equiv 55\% \\
 p_2 &= \frac{\text{total aprobaron en 2da encuesta}}{\text{total de encuestados}} = \frac{880}{1600} = 0.59 \equiv 59\%
 \end{aligned}$$

Podemos ver que solo existe una diferencia del 4%, en principio, podríamos pensar que esta diferencia no es estadísticamente significativa, es por ello necesario que determinemos formalmente si la diferencia que observamos es lo suficientemente grande como para no ser atribuible al azar, o si es tan pequeña que podría ser simplemente una fluctuación aleatoria.

Dado que los datos son pareados, es decir, tenemos dos observaciones para los $n = 1600$ ciudadanos encuestados y estos datos son dependientes entre si, la prueba adecuada para comparar las proporciones es la prueba de McNemar. Esta prueba se enfoca unicamente en los individuos que cambiaron de opinión entre las dos encuestas, es decir, se centra en las celdas discordantes de la tabla de contingencia:

$n_{10} = 150$: ciudadanos que aprobaron y luego desaprobaron

$n_{01} = 86$: ciudadanos que desaprobaron y luego aprobaron

El estadístico de la prueba de McNemar se calcula con la siguiente fórmula y sigue una distribución χ^2 con 1 grado de libertad:

$$\chi^2 = \frac{(n_{10} - n_{01})^2}{n_{10} + n_{01}} \quad (19)$$

Si el valor de χ^2 es mayor que el valor crítico (o si el p -valor es menor que α), rechazamos la hipótesis nula H_0 y concluimos que hubo un cambio significativo en la proporción de aprobación. Pero si χ^2 es menor, no tenemos evidencia suficiente para rechazar H_0 .

Sustituyendo los valores de las celdas discordantes en (19):

$$\chi^2 = \frac{(150 - 86)^2}{150 + 86} = \frac{(64)^2}{236} = \frac{4096}{236} \approx 17.3559$$

Para un nivel de significancia de $\alpha = 0.05$ y 1 grado de libertad el valor crítico de la distribución Chi-cuadrada es $\chi^2_{\alpha=0.05} = 3.841$, entonces:

$$\chi^2 > \chi^2_{\alpha=0.05}$$

Por lo tanto, se rechaza la hipótesis nula a un nivel de significancia del 5% lo que indica que si existe un cambio estadísticamente significativo entre las proporciones de ciudadanos que aprueban el desempeño del ministro en la primer encuesta y los que la aprobaron en la segunda encuesta.

Problema 2

Se tiene la siguiente tabla donde se eligen varios niveles de ronquidos y se ponen en relación con una enfermedad cardíaca. Se toman como puntuaciones relativas de ronquidos los valores $\{0, 2, 4, 5\}$.

Ronquido	Enfermedad Cardíaca		Proporción de SI
	SI	NO	
Nunca	24	1355	0.017
Ocasional	35	603	0.055
Casi cada noche	21	192	0.099
Cada noche	30	224	0.118

Ajuste un modelo lineal generalizado **logit** y **probit** (investigar sobre el link **probit**) para analizar si existe una relación entre los ronquidos y la posibilidad de tener una enfermedad cardíaca.

SOLUCIÓN

Se utilizo el lenguaje de programación **R** para resolver este problema, el código utilizado se encuentran al final del documento en el apéndice **A.1**

Primero recordemos que el propósito de una función de enlace en un GLM busca modelar la relación entre uno o más predictores X y una variable de respuesta Y . La función de enlace g , es el puente que conecta el predictor lineal, $\eta = \beta_0 + \beta_1 X_1 + \dots$, con el valor esperado de la variable de respuesta $\mu = E(Y)$. Su forma fundamental es:

$$g(\mu) = \eta \quad (20)$$

En este problema, la respuesta es binaria (Sí/No enfermedad), por lo que su valor esperado es una probabilidad, sea $\pi = P(Y = 1)$, que siempre está entre 0 y 1. Sin embargo, el predictor lineal η puede tomar cualquier valor real de $-\infty$ a $+\infty$. La función de enlace se encarga de mapear el rango $(0, 1)$ de la probabilidad al rango $(-\infty, +\infty)$ del predictor lineal. El **enlace logit**, utiliza la función logaritmo de las “odds” (razón de momios) que habíamos visto antes:

$$g(\pi) = \ln \left(\frac{\pi}{1 - \pi} \right) \quad (21)$$

Esta función es la inversa de la función de distribución acumulada, CDF, de la distribución logística estándar. Por otra parte el **enlace probit** es muy similar conceptualmente, pero utiliza una distribución diferente como base, esta es la función cuantil:

$$g(\pi) = \Phi^{-1}(\pi) \quad (22)$$

Donde la función cuantil Φ^{-1} es la inversa de la CDF de la distribución normal estándar, $N(0, 1)$. En otras palabras, la función probit toma una probabilidad π y te devuelve el puntaje Z (z-score) correspondiente a esa probabilidad.

Entonces, para resolver el problema primero cargamos los datos manualmente en un data frame y se procedió a ajustar los modelos logit y probit obteniendo los siguientes resultados respectivamente:

Table 0.1: Resumen del ajuste del modelo Logit

	Estimate	Std. Error	z value	Pr(> z)
Intercept	-3.86625	0.16621	-23.261	$< 2 \times 10^{-16}***$
score	0.39734	0.05001	7.945	$1.94 \times 10^{-15}***$

Signif. codes: ***0.001, **0.01, *0.05, .0.1

Null deviance: 65.9045 on 3 d.f.

Residual deviance: 2.8089 on 2 d.f.

AIC: 27.061

Fisher Scoring iterations: 4

Analizando estos resultados, el resumen del modelo logit nos dice que la variable **score** es un predictor muy importante, puesto que:

- **Coefficiente de score:** El valor estimado es 0.39734. Al ser positivo, indica que a medida que aumenta la puntuación de ronquido, aumentan los log-odds (y por lo tanto la probabilidad) de tener una enfermedad cardíaca.
- **Significancia Estadística:** El p-valor es 1.94×10^{-15} , un número extremadamente pequeño (prácticamente cero). Esto, junto con el código de significancia ***, nos permite concluir con mucha confianza que la relación entre el ronquido y la enfermedad cardíaca no es una casualidad.
- **Ajuste del Modelo:** La **Residual deviance** (2.81) es drásticamente menor que la **Null deviance** (65.90), lo que confirma que incluir la variable **score** mejora drásticamente el ajuste del modelo.
- **Odds Ratio:** Podemos calcular el Odds Ratio (OR) a partir del coeficiente $e^{0.39734} \approx 1.488$. Esto significa que por cada punto que aumenta la escala de ronquido, los odds de tener una enfermedad cardíaca aumentan en aproximadamente un 48.8%.

Table 0.2: Resumen del ajuste del modelo Probit

	Estimate	Std. Error	z value	Pr(> z)
Intercept	-2.06055	0.07017	-29.367	$< 2 \times 10^{-16}***$
score	0.18777	0.02348	7.997	$1.28 \times 10^{-15}***$

Signif. codes: ***0.001, **0.01, *0.05, .0.1

Null deviance: 65.9045 on 3 d.f.

Residual deviance: 1.8716 on 2 d.f.

AIC: 26.124

Fisher Scoring iterations: 4

Por otra parte, el modelo probit cuenta exactamente la misma historia, solo que con una escala diferente:

- **Coefficiente de score:** El valor estimado es 0.18777. De nuevo, es positivo, lo que confirma que a medida que aumenta la puntuación de ronquido, aumenta la probabilidad de enfermedad cardíaca. En este caso, el coeficiente representa el cambio en el puntaje Z asociado a dicha probabilidad.
- **Significancia Estadística:** El p-valor es 1.28×10^{-15} y al igual que en el modelo logit, es extremadamente pequeño ***. La conclusión es idéntica: la variable **score** es un predictor altamente significativo.
- **Ajuste del Modelo:** Al igual que antes, la **Residual deviance** (1.87) es mucho menor que la **Null deviance** (65.90), indicando un buen ajuste.

Por lo tanto, ambos modelos llegan a la misma conclusión y es que existe una relación positiva y estadísticamente muy significativa entre la intensidad del ronquido y la probabilidad de padecer una enfermedad cardíaca.

Para comparar formalmente ambos modelos, observamos el AIC (Akaike Information Criterion). Un AIC más bajo indica un mejor equilibrio entre ajuste y simplicidad, aquí tenemos:

- AIC Logit: 27.061
- AIC Probit: 26.124

El modelo probit tiene un AIC ligeramente más bajo, lo que sugiere que se ajusta marginalmente mejor a estos datos específicos. Sin embargo, la diferencia es muy pequeña. En la práctica, ambos modelos son excelentes para este problema y la elección entre uno u otro a menudo se reduce a la preferencia por la interpretación.

Finalmente, se realizó una gráfica para visualizar los resultados del ajuste, mostrando las proporciones observadas de la enfermedad (puntos negros) junto con las curvas de probabilidad predichas por los modelos logit (línea continua) y probit (línea discontinua):

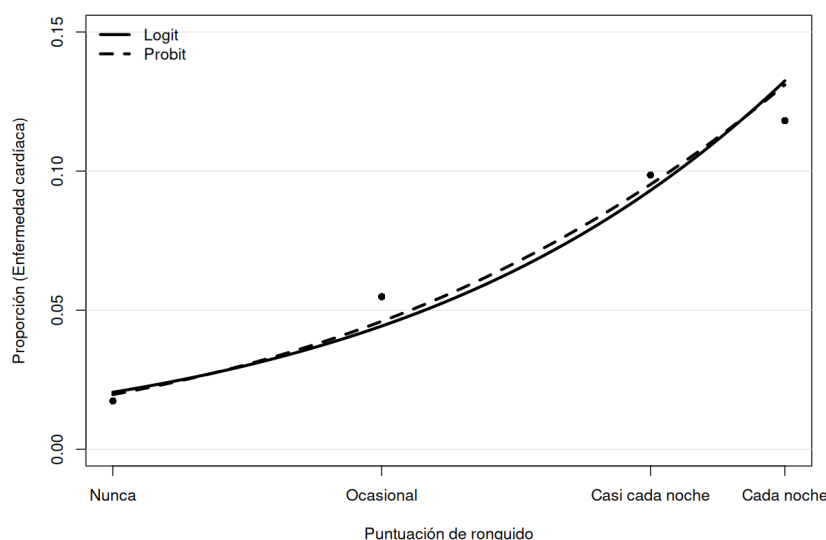


Figure 0.1: Curvas de probabilidad ajustadas (logit y probit).

La figura confirma la relación positiva y monótona entre la puntuación de ronquido y la probabilidad de padecer una enfermedad cardíaca, evidenciada por la clara tendencia ascendente tanto de los datos como de las curvas. Asimismo, se observa un excelente ajuste de ambos modelos. La superposición casi perfecta de las dos curvas demuestra visualmente la conclusión del análisis: los modelos logit y probit ofrecen resultados prácticamente idénticos y son igualmente efectivos para describir la relación en este conjunto de datos.

Problema 3

Entre los cangrejos cacerola se sabe que cada hembra tiene un macho en su nido, pero puede tener más machos concubinos.

Se considera que la variable respuesta es el número de concubinos y las variables explicativas son: color, estado de la espina central, peso y anchura del caparazón.

color	Spine	Width	Satellite	Weight
3	3	28.3	8	3050
4	3	22.5	0	1550
2	1	26.0	9	2300
4	3	24.8	0	2100
4	3	26.0	4	2600
3	3	23.8	0	2100
2	1	26.5	0	2350

Realizar e interpretar los resultados de ajustar un modelo lineal generalizado tipo **poisson**.

SOLUCIÓN

Se utilizo el lenguaje de programación R para resolver este problema, el código utilizado se encuentran al final del documento en el apéndice A.2

Tras ajustar un GLM de tipo Poisson para predecir el número de machos satélite el proceso de ajuste presentó problemas críticos de convergencia e inestabilidad. Los resultados numéricos del ajuste se muestran a continuación:

Predictor	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.40×10^3	1.97×10^6	0.001	0.999
color3	-1.42×10^2	2.12×10^5	-0.001	0.999
color4	-1.47×10^2	2.01×10^5	-0.001	0.999
Spine3	NA	NA	NA	NA
Width	-9.68×10^1	1.33×10^5	-0.001	0.999
Weight	4.87×10^{-1}	6.70×10^2	0.001	0.999

Nota: 1 coeficiente (Spine3) no definido por singularidades.

Parámetro de dispersión para la familia Poisson tomado como 1.

Devianza Nula: 37.77 con 6 g.l.

Devianza Residual: 6.69×10^{-10} con 2 g.l.

AIC: 21.258

Iteraciones de Fisher Scoring: 25

Analizando estos resultados, el resumen del modelo poisson nos muestra múltiples indicadores de que el modelo ha fallado, puesto que:

- Singularidad y Coeficientes NA:** La nota al pie de la tabla y la fila para **Spine3** indican una singularidad. Esto se debe a una multicolinealidad perfecta en la muestra de datos (todos los cangrejos con **Spine=1** también tienen **color=2**), lo que impide al modelo estimar el efecto de **Spine3** de forma independiente.
- Estimaciones y Errores Estándar:** Los valores en las columnas **Estimate** y **Std. Error** son extremadamente grandes. Por ejemplo, el intercepto tiene un error estándar de 1.97×10^6 .

Errores estándar de esta magnitud nos indican que las estimaciones de los coeficientes no tienen ninguna precisión y son completamente inestables.

3. **Significancia Estadística:** Como consecuencia de los enormes errores estándar, todos los valores en la columna `z value` son casi cero y los p -valores son efectivamente 1. Esto confirma que ninguna variable puede ser considerada estadísticamente significativa, no porque no tengan un efecto, sino porque el modelo fue incapaz de estimarlo.
4. **Devianza Residual:** La devianza residual es prácticamente cero, 6.69×10^{-10} , lo que es un síntoma claro de un sobreajuste (overfitting) severo. El modelo se ha ajustado de manera casi perfecta a los 7 puntos de datos, perdiendo toda capacidad de generalización.

Por lo tanto, el modelo es estadísticamente inválido. La causa principal es el tamaño de muestra extremadamente pequeño, $n=7$, en relación con el número de predictores, lo que provoca multicolinealidad e inestabilidad numérica. Por lo tanto, no es posible extraer ninguna conclusión fiable sobre la relación entre las variables predictoras y el número de machos satélite a partir de este análisis. Adicionalmente, el alto número de iteraciones de Fisher (25) y las advertencias de R sobre la falta de convergencia confirman el fallo del ajuste.

Problema 4

Suponga $(x_1, y_1), \dots, (x_n, y_n)$ observaciones independientes de variables aleatorias definidas como sigue:

$$\begin{aligned} Y_i &\sim \text{Bernoulli}(p), \quad i = 1, \dots, n \\ X_i | \{Y_i = 1\} &\sim N(\mu_1, \sigma^2) \\ X_i | \{Y_i = 0\} &\sim N(\mu_0, \sigma^2) \end{aligned}$$

Usando el Teorema de Bayes, muestre que $P(Y_i = 1|X_i)$ satisface el modelo de regresión logística, esto es

$$\text{logit}(P(Y_i = 1|X_i)) = \alpha + \beta X_i$$

con

$$\beta = \frac{\mu_1 - \mu_0}{\sigma^2}.$$

SOLUCIÓN

Comencemos recordando la forma del teorema de Bayes para dos eventos A y B :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (23)$$

Para este caso, dadas las variables aleatorias tenemos que el teorema de Bayes toma la siguiente forma:

$$P(Y_i = 1|X_i = x_i) = \frac{P(X_i = x_i|Y_i = 1)P(Y_i = 1)}{P(X_i = x_i)} \quad (24)$$

Esto debido a que nos interesa encontrar $P(Y_i = 1|X_i)$. Ya que la variable Y_i es tipo Bernoulli sabemos que es una variable discreta y por ende podemos trabajar con sus probabilidades, pero por otro lado, la variable X_i sigue una distribución normal, es decir, es una variable continua por lo que no hablamos de su probabilidad en un punto exacto, sino que podemos trabajar con su función de densidad de probabilidad (pdf), tal que, la ecuación (33) pasa a tener la siguiente forma:

$$P(Y_i = 1|X_i = x_i) = \frac{f(x_i|Y_i = 1)P(Y_i = 1)}{f(x_i)} \quad (25)$$

Ya que no sabemos que distribución sigue la variable aleatoria X por si sola, podemos expandir el denominador mediante el teorema de probabilidad total que establece lo siguiente; dado un conjunto de eventos $\{B_1, B_2, \dots, B_n\}$ que forman una partición del espacio muestral, entonces para cualquier otro evento A en ese mismo espacio:

$$P(A) = \sum_{i=1}^n P(A|B_i)P(B_i) \quad (26)$$

Entonces, para este caso tenemos que:

$$\begin{aligned} f(x_i) &= P(X_i = x_i|Y_i = 0)P(Y_i = 0) + P(X_i = x_i|Y_i = 1)P(Y_i = 1) \\ &= f(x_i|Y_i = 0)P(Y_i = 0) + f(x_i|Y_i = 1)P(Y_i = 1) \end{aligned}$$

Sustituyendo en (25):

$$P(Y_i = 1|X_i = x_i) = \frac{f(x_i|Y_i = 1)P(Y_i = 1)}{f(x_i|Y_i = 0)P(Y_i = 0) + f(x_i|Y_i = 1)P(Y_i = 1)} \quad (27)$$

Ahora podemos sustituir los valores de las distribuciones conocidas, recordando que $Y_i \sim \text{Bernoulli}(p)$:

$$P(Y_i = 1) = p, \quad P(Y_i = 0) = 1 - p$$

Mientras que para $X_i|\{Y_i = 1\} \sim N(\mu_1, \sigma^2)$ y $X_i|\{Y_i = 0\} \sim N(\mu_0, \sigma^2)$, tenemos que la formula para la densidad de la distribución normal es:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu^2)}{2\sigma^2}\right) \quad (28)$$

Tal que:

$$f(x_i|Y_i = 1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu_1^2)}{2\sigma^2}\right), \quad f(x_i|Y_i = 0) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu_0^2)}{2\sigma^2}\right)$$

Entonces, sustituyendo en (27) y trabajando con la expresión:

$$\begin{aligned} P(Y_i = 1|X_i = x_i) &= \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu_1)^2}{2\sigma^2}\right) p}{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu_0)^2}{2\sigma^2}\right) (1 - p) + \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu_1)^2}{2\sigma^2}\right) p} \\ &= \frac{\exp\left(-\frac{(x_i - \mu_1)^2}{2\sigma^2}\right) p}{\exp\left(-\frac{(x_i - \mu_0)^2}{2\sigma^2}\right) (1 - p) + \exp\left(-\frac{(x_i - \mu_1)^2}{2\sigma^2}\right) p} \\ &= \frac{1}{\frac{\exp\left(-\frac{(x_i - \mu_0)^2}{2\sigma^2}\right) (1 - p)}{\exp\left(-\frac{(x_i - \mu_1)^2}{2\sigma^2}\right) p} + 1} \\ &= \frac{1}{\exp\left(-\frac{(x_i - \mu_0)^2}{2\sigma^2} + \frac{(x_i - \mu_1)^2}{2\sigma^2}\right) \frac{(1 - p)}{p} + 1} \\ &= \frac{1}{\exp\left(\frac{-x_i^2 + 2x_i\mu_0 - \mu_0^2}{2\sigma^2} + \frac{x_i^2 - 2x_i\mu_1 + \mu_1^2}{2\sigma^2}\right) \frac{(1 - p)}{p} + 1} \\ &= \frac{1}{\exp\left(\frac{2x_i\mu_0 - 2x_i\mu_1 - \mu_0^2 + \mu_1^2}{2\sigma^2}\right) \frac{(1 - p)}{p} + 1} \\ &= \frac{1}{\exp\left(\frac{2x_i\mu_0 - 2x_i\mu_1 - \mu_0^2 + \mu_1^2}{2\sigma^2}\right) \frac{(1 - p)}{p} + 1} \\ &= \frac{1}{\exp\left(\frac{(2\mu_0 - 2\mu_1)x_i}{2\sigma^2} + \frac{(\mu_1^2 - \mu_0^2)}{2\sigma^2}\right) \frac{(1 - p)}{p} + 1} \end{aligned}$$

Para simplificar la notación, sea $Z = \exp\left(\frac{(2\mu_0 - 2\mu_1)x_i + (\mu_1^2 - \mu_0^2)}{2\sigma^2}\right) \frac{(1-p)}{p}$, tal que:

$$P(Y_i = 1|X_i = x_i) = \frac{1}{Z + 1} \quad (29)$$

Recordemos que la función logit se define como:

$$\text{logit}(P) = \log\left(\frac{P}{1-P}\right) \quad (30)$$

Entonces, buscamos la misma forma del argumento, para nuestro caso primero:

$$1 - P = 1 - \frac{1}{1 + Z} = \frac{Z}{1 + Z}$$

Tal que, el argumento completo:

$$\frac{P}{1-P} = \frac{\frac{1}{1+Z}}{\frac{Z}{1+Z}} = \frac{1}{Z}$$

Para obtener el logit aplicamos el logaritmo:

$$\text{logit}(P(Y_i = 1|X_i = x_i)) = \log\left(\frac{1}{Z}\right) = \log(Z^{-1}) = -\log(Z)$$

Expandimos Z para trabajar con el logaritmo:

$$\begin{aligned} \text{logit}(P(Y_i = 1|X_i = x_i)) &= -\log\left[\exp\left(\frac{(2\mu_0 - 2\mu_1)x_i}{2\sigma^2} + \frac{(\mu_1^2 - \mu_0^2)}{2\sigma^2}\right) \frac{1-p}{p}\right] \\ &= -\left\{\log\left[\exp\left(\frac{(2\mu_0 - 2\mu_1)x_i}{2\sigma^2} + \frac{(\mu_1^2 - \mu_0^2)}{2\sigma^2}\right)\right] + \log\left(\frac{1-p}{p}\right)\right\} \\ &= -\frac{(2\mu_0 - 2\mu_1)x_i}{2\sigma^2} - \frac{\mu_1^2 - \mu_0^2}{2\sigma^2} - \log\left(\frac{1-p}{p}\right) \\ &= \log\left(\frac{p}{1-p}\right) + \frac{2(\mu_1 - \mu_0)x_i}{2\sigma^2} + \frac{\mu_0^2 - \mu_1^2}{2\sigma^2} \\ &= \underbrace{\left[\log\left(\frac{p}{1-p}\right) + \frac{\mu_0^2 - \mu_1^2}{2\sigma^2}\right]}_{\alpha} + \underbrace{\left(\frac{\mu_1 - \mu_0}{\sigma^2}\right)}_{\beta} x_i \end{aligned}$$

Por lo tanto, queda demostrado que se satisface:

$$\boxed{\text{logit}(P(Y_i = 1|X_i = x_i)) = \alpha + \beta x_i}$$

Problema 5

Cuando usamos un modelo de regresión logística para clasificación, tenemos que definir el umbral, p , a partir del cual declaramos un “positivo”.

Las curvas ROC grafican las tasas TPR vs FPR para diferentes umbrales p .

$$TPR = \text{True Positive Rate} = \frac{TP}{P} = \text{“sensitividad”}$$

$$FPR = \text{False Positive Rate} = \frac{FP}{N} = 1 - \text{“especificidad”}$$

Decisión	Observados	
	1	0
1	TP	FP
0	FN	TN
Total	P	N

- La gráfica de TPR vs FPR puede interpretarse como una gráfica de “poder” vs “error tipo I”.
- Idealmente, una regla de decisión estaría en el punto $(0, 1)$.
- El área bajo la curva, AUC , puede verse, es la probabilidad de que un individuo de los positivos, tomado al azar, tenga un riesgo estimado mayor que un individuo de los negativos, tomado al azar.
- El estadístico J de Youden, es una medida que, con un sólo número, trata de capturar el desempeño de una prueba de diagnóstico. Es la máxima distancia vertical, entre la diagonal y la curva ROC, o equivalentemente

$$J = \text{sensitividad} - (1 - \text{especificidad})$$

Construyan la curva ROC para el problema de daño coronario y su relación con la edad visto en la clase 3 del curso.

```

1 # Hosmer, D.W. & Lemeshow, S.(1989) Applied logistic regression. Wiley
2 # Edad y Coronaria (daño significativo en coronaria)
3
4 edad <- c(
5   20, 23, 24, 25, 25, 26, 26, 28, 28, 29, 30, 30, 30, 30, 30, 30, 32, 32, 33, 33,
6   34, 34, 34, 34, 34, 35, 35, 36, 36, 36, 37, 37, 37, 38, 38, 39, 39, 40, 40, 41,
7   41, 42, 42, 42, 42, 43, 43, 43, 44, 44, 44, 44, 45, 45, 46, 46, 47, 47, 47, 48,
8   48, 48, 49, 49, 49, 50, 50, 51, 52, 52, 53, 53, 54, 55, 55, 55, 56, 56, 56, 57,
9   57, 57, 57, 57, 57, 58, 58, 58, 59, 59, 60, 60, 61, 62, 62, 63, 64, 64, 65, 69
10 )
11
12 coro <- c(
13   0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,1,0,0,1,0,0,1,0,1,0,1,0,
14   0,0,0,1,0,0,1,0,0,1,1,0,1,0,1,0,0,1,1,0,1,0,0,1,1,1,1,0,1,1,1,1,0,0,1,1,1,1,0,
15   1,1,1,1,0,1,1,1,1,1,1,0,1,1,1
16 )

```

Listing 1: Datos de Edad y Daño Coronario

SOLUCIÓN

Se utilizó el lenguaje de programación *R* para resolver este problema, el código utilizado se encuentran al final del documento en el apéndice **A.3**

Buscamos construir y analizar la curva ROC para un modelo que prediga la probabilidad de daño coronario significativo a partir de la edad de un paciente, básicamente la curva ROC nos servirá como herramienta gráfica para evaluar el rendimiento de un clasificador binario a medida que se varía el umbral de discriminación.

Para este problema se nos pide ajustar un modelo de regresión logística binomial para predecir la probabilidad de daño coronario, $y = 1$, a partir de la edad. El modelo tiene la forma:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot \text{edad} \quad (31)$$

donde $p = P(y = 1 | \text{edad})$. Para la construcción de la curva ROC se generaron las probabilidades predichas \hat{p} para cada observación. Se evaluó un conjunto de umbrales de decisión t sobre estas probabilidades. Para cada umbral, se clasificó una observación como positiva si $\hat{p} \geq t$ y se calcularon los correspondientes valores de TPR y FPR para construir la curva. El área bajo la curva (AUC) se calculó numéricamente utilizando la regla del trapecio. La curva ROC obtenida se muestra a continuación:

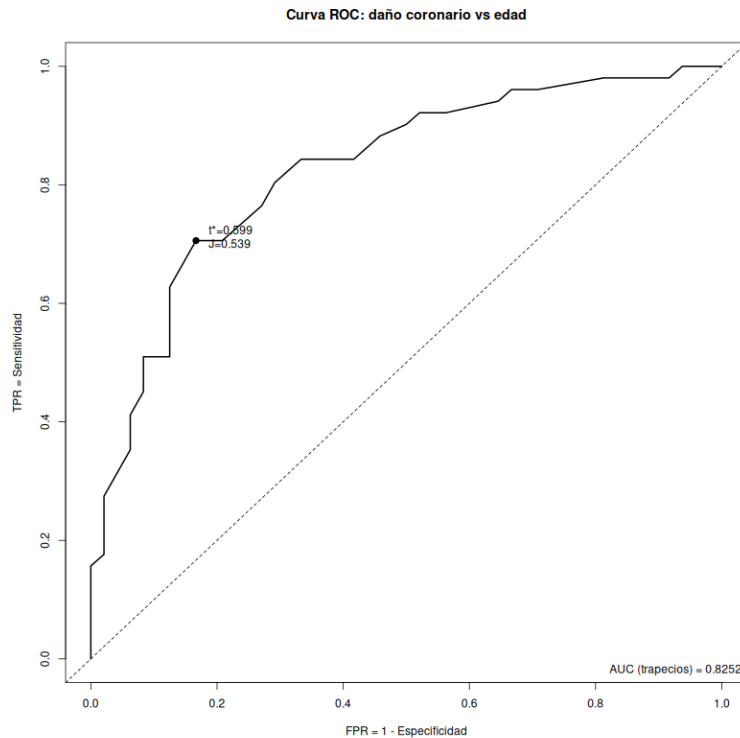


Figure 0.2: Curva ROC para el modelo de daño coronario vs. edad. Se muestra el punto óptimo según el índice J de Youden y el valor del AUC.

Podemos observar que el modelo de regresión logística demostró una buena capacidad de discriminación. El área bajo la curva calculada fue de **AUC = 0.8252**, lo que nos indica un buen rendimiento del clasificador (considerando que un valor de 0.5 corresponde al azar y 1.0 a una clasificación perfecta).

Para determinar el umbral óptimo t^* se maximizó el estadístico J de Youden y se obtuvo que $t^* = 0.5992$, este umbral representa el punto de corte en la probabilidad predicha que ofrece el mejor equilibrio entre sensibilidad y especificidad. Las métricas de rendimiento en este punto óptimo son:

- **Índice J de Youden máximo:** $J_{max} = 0.5392$
- **Sensitividad (TPR):** 0.7059
- **Especificidad:** 0.8333
- **Tasa de Falsos Positivos (FPR):** 0.1667

Al aplicar el umbral óptimo $t^* = 0.5992$ a los datos, se obtuvo la siguiente matriz de confusión:

Table 0.4: Matriz de confusión en el umbral óptimo

Decisión	Observados	
	1 (Positivo)	0 (Negativo)
1 (Positivo)	36 (TP)	8 (FP)
0 (Negativo)	15 (FN)	40 (TN)

A partir de lo obtenido en la matriz de confusión, podemos derivar las siguientes métricas globales:

- **Accuracy:** $\frac{36 + 40}{36 + 8 + 15 + 40} = \frac{76}{99} \approx 0.7677$
- **Precision:** $\frac{36}{36 + 8} = \frac{36}{44} \approx 0.8182$

Por lo tanto, podemos decir que el modelo de regresión logística basado en la edad es una herramienta útil para discriminar entre pacientes con y sin daño coronario, como lo demuestra un AUC de 0.8252. El umbral de probabilidad óptimo de 0.5992 permite clasificar a los pacientes con una sensibilidad del 70.6% y una especificidad del 83.3%. Esto significa que el modelo, con este punto de corte, identifica correctamente al 70.6% de los enfermos, mientras que clasifica correctamente al 83.3% de los sanos. La precisión del 81.8% indica que, de todos los pacientes clasificados como positivos, la gran mayoría lo son realmente.

Problema 6

La siguiente tabla muestra conteos de células T_4 por mm^3 en muestras de sangre de 20 pacientes (en remisión) con enfermedad de Hodgkin, así como conteos en 20 pacientes en remisión de otras enfermedades. Una cuestión de interés es si existen diferencias en las distribuciones de conteos en ambos grupos.

H	396	568	1212	171	554	1104	257	435	295	397
No-H	375	375	752	208	151	116	736	192	315	1252
H	288	1004	431	795	1621	1378	902	958	1283	2415
No-H	675	700	440	771	688	426	410	979	377	503

- Haga una comparación gráfica exploratoria de estos datos.
- Ajuste un modelo de Poisson apropiado.
- Usando la normalidad asintótica de los estimadores de máxima verosimilitud, dé un intervalo del 90% de confianza para la diferencia en medias. ¿Hay evidencia de diferencias en los dos grupos en cuanto a las medias de los conteos?.

SOLUCIÓN

Se utilizó el lenguaje de programación R para resolver este problema, el código utilizado se encuentra al final del documento en el apéndice A.4

a) Primero se cargaron manualmente los datos y se guardaron en un archivo .csv para trabajar con ellos de mejor manera. Antes de trabajar con un modelo estadístico siempre es bueno realizar una visualización gráfica de los datos de ser posible, para este caso primero se realizó un boxplot:

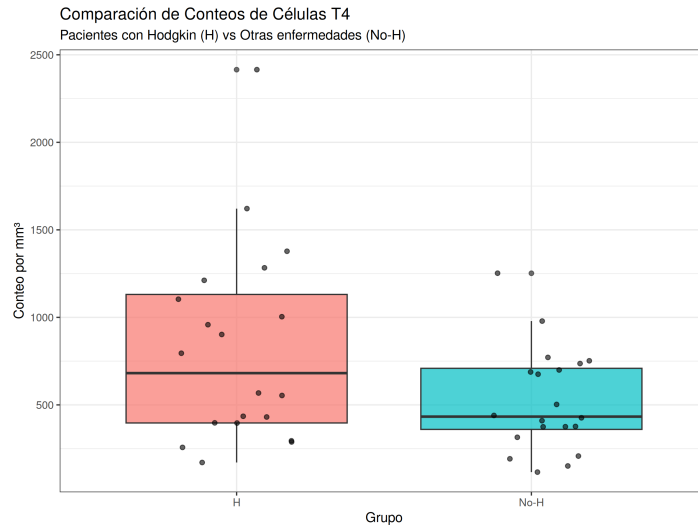


Figure 0.3: Boxplot de conteos de células T4 para el grupo H y No-H.

En la Figura 0.3 podemos observar que la mediana del grupo H es mayor en comparación con la del grupo No-H, esto nos sugiere que los pacientes con Hodgkin tienen conteos más altos de células T4 que los otros. También podemos ver que la caja del grupo H es más grande lo que implica una mayor dispersión de los datos, es decir, más variabilidad entre el conteo de células T4 de los pacientes mientras que el grupo No-H tiene una caja más compacta indicando lo opuesto. El grupo H tiene valores muy

altos que aparecen como outliers (≈ 2400), el grupo No-H también tiene algunos outliers altos pero no tan extremos (≈ 1250), esto podría indicarnos que la distribución del grupo H podría tener una cola derecha larga. En general el grupo H tiene mayor mediana y mayor dispersión, siendo su distribución más heterogénea y con valores extremos grandes, mientras que el grupo No-H tiene menor mediana, menos dispersión y valores más concentrados.

Para complementar esta descripción visual también se realizaron histogramas de frecuencia:

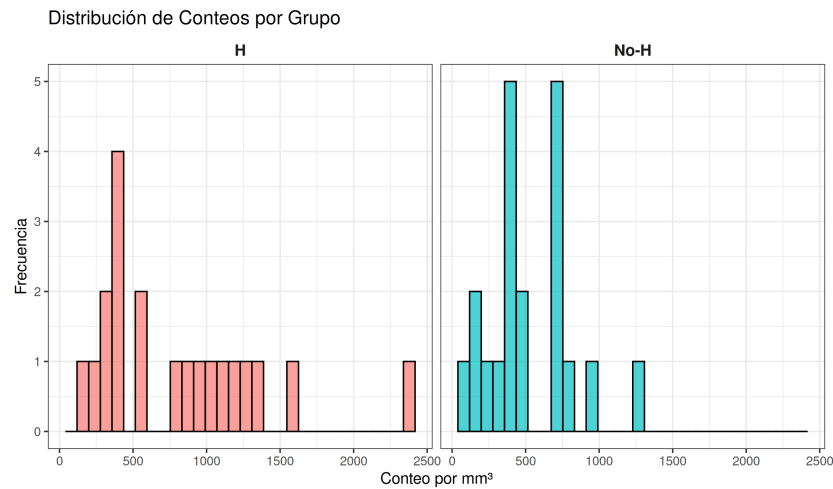


Figure 0.4: Histogramas de frecuencia de los conteos de células T4 para el grupo H y No-H.

En la Figura 0.4 podemos observar que la cola larga hacia la derecha que se había sospechado en el boxplot, así mismo podemos ver que los pacientes del grupo H tienden a tener conteos de células T4 mas altos en promedio, mientras que los pacientes del grupo No-H presenta conteos mas bajos y mas consistentes entre si ya que la mayoría de conteos se encuentran agrupados en valores intermedios.

Por ultimo se realizo un gráfico de densidad:

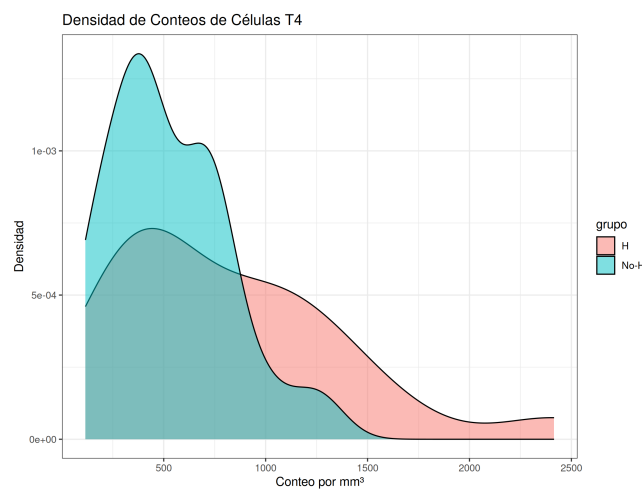


Figure 0.5: Gráfico de densidad de conteos de células T4 para el grupo H y No-H.

La Figura 0.5 muestra que los conteos de células T4 en pacientes del grupo H son más heterogéneos, con una cola derecha larga y valores muy altos, mientras que en pacientes del grupo No-H la distribución es más concentrada en torno a valores moderados y con poca presencia de extremos.

Por lo tanto, en primera instancia estas comparaciones gráficas parecen indicarnos que si existen diferencias en las distribuciones de conteos en ambos grupos.

b) En lugar de comparar solo medias con un test clásico, usamos un Modelo Lineal Generalizado (GLM) con respuesta Poisson, que es natural para modelar datos de conteos. Los supuestos del modelo Poisson son los siguientes:

1. La variable respuesta Y_i (conteo de células del paciente i) sigue una distribución Poisson:

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

donde $\mu_i = E(Y_i)$ es la media (también varianza bajo el modelo básico).

2. Los Y_i son condicionalmente independientes.
3. La media μ_i depende de covariables (aquí, el grupo H vs No-H) a través de una función de enlace.

En los GLM con respuesta Poisson, el enlace canónico es el logaritmo:

$$\log(\mu_i) = \beta_0 + \beta_1 x_i \quad (32)$$

Donde:

- $x_0 = 0$ si el paciente está en el grupo H,
- $x_1 = 1$ si el paciente está en el grupo No-H.

Buscamos ajustar este modelo bajo los supuestos mencionados para estimar los parámetros β_0 y β_1 , siendo cada uno:

$$\beta_0 = \log(\mu_H) \quad \text{y} \quad \beta_1 = \log\left(\frac{\mu_{\text{No-H}}}{\mu_H}\right) \quad (33)$$

Entonces, ajustando el modelo mediante la función `glm()` se obtuvo el summary" siguiente:

Table 0.5: Resumen del ajuste del modelo Poisson				
	Estimate	Std. Error	z value	Pr(> z)
Intercept	6.713199	0.007793	861.4	$< 2 \times 10^{-16}***$
grupoNo-H	-0.455436	0.012511	-36.4	$< 2 \times 10^{-16}***$

Signif. codes: ***0.001, **0.01, *0.05, .0.1

Null deviance: 11325 on 39 d.f.

Residual deviance: 9965 on 38 d.f.

AIC: 10294

Fisher Scoring iterations: 5

Podemos observar que el intercepto es $\hat{\beta}_0 = 6.713$. Esto corresponde al grupo H. Por lo tanto, despejando de la ecuación (33), tenemos que la media esperada en este grupo es:

$$\hat{\mu}_H = e^{6.713} \approx 824.8$$

Mientras que el efecto del grupo No-H es $\hat{\beta}_1 = -0.455$. Despejando de la ecuación (33), esto implica que manteniendo todo lo demás constante, la razón de medias entre No-H y H es:

$$\frac{\hat{\mu}_{\text{No-H}}}{\hat{\mu}_{\text{H}}} = e^{-0.455} \approx 0.634$$

Es decir, los pacientes No-H presentan, en promedio, un conteo de células T4 aproximadamente un 36.6% menor que los pacientes del grupo H.

De acuerdo a los supuestos mencionados, el modelo de Poisson asume que la varianza de Y_i es igual a su media, $\text{Var}(Y_i) = \mu_i$. Sin embargo, en nuestros datos observamos que la desviación residual (9965) es mucho mayor que los grados de libertad residuales (38) y que los conteos de T4 muestran gran heterogeneidad y colas largas en el grupo H, lo que sugiere que $\text{Var}(Y_i) \gg E(Y_i)$. Por lo tanto, esto indica una sobre-dispersión, es decir, el modelo Poisson subestima la variabilidad real de los datos.

Para corregir la inferencia, ajustaremos un modelo quasi-Poisson, que conserva la misma forma funcional pero permite que:

$$\text{Var}(Y_i) = \phi \mu_i$$

donde $\phi > 1$ es un parámetro de dispersión que captura el exceso de variabilidad. De esta manera, los estimadores $\hat{\beta}$ permanecen iguales, pero sus errores estándar se corrigen multiplicándose por $\sqrt{\phi}$, proporcionando intervalos de confianza y pruebas más realistas.

Ahora, ajustando el modelo quasi-Poisson mediante la función `glm()` se obtuvo el summary” siguiente:

Table 0.6: Resumen del ajuste del modelo quasi-Poisson

	Estimate	Std. Error	t value	Pr(> t)
Intercept	6.7132	0.1297	51.750	$< 2 \times 10^{-16}***$
grupoNo-H	-0.4554	0.2082	-2.187	0.035*

Signif. codes: ***0.001, **0.01, *0.05, .0.1

Dispersion parameter: $\phi = 277.0613$

Null deviance: 11325 on 39 d.f.

Residual deviance: 9965 on 38 d.f.

AIC: NA

Fisher Scoring iterations: 5

Podemos ver que tras ajustar por sobredispersión se obtuvo un $\phi = 277.06$ mediante el modelo quasi-Poisson, la diferencia asociada al grupo No-H persistió como estadísticamente significativa con $p = 0.035$.

c) Una vez ajustado el modelo quasi-Poisson, que corrige la sobredispersión de los datos, procedemos a evaluar si existe evidencia estadística de una diferencia en las medias de los conteos celulares entre los dos grupos. Para ello, utilizamos la normalidad asintótica de los estimadores de máxima verosimilitud para construir un intervalo de confianza para el parámetro β_1 , que representa el efecto del grupo No-H en la escala logarítmica. Del resumen del modelo quasi-Poisson, obtuvimos que $\hat{\beta}_1 = -0.4554$ y un error estándar: 0.2082.

Para construir un intervalo de confianza del 90%, usamos el valor crítico correspondiente a un nivel de significancia de $\alpha = 0.10$, es decir, $z_{0.95} \approx 1.645$ proveniente de una distribución normal estándar.

El intervalo se calcula como:

$$\begin{aligned}\hat{\beta}_1 \pm z_{0.95} \cdot SE(\hat{\beta}_1) &= -0.4554 \pm 1.645 \cdot 0.2082 \\ &= -0.4554 \pm 0.3425 \\ &= [-0.7979, -0.1129]\end{aligned}$$

El intervalo de confianza del 90% para β_1 es $[-0.798, -0.113]$. Dado que este intervalo está completamente por debajo de cero y no lo incluye, existe evidencia estadística significativa (con un nivel de confianza del 90%) de que la media de conteos en el grupo No-H es diferente a la del grupo H.

Para facilitar la interpretación, podemos transformar este resultado a la escala original (razón de medias) exponenciando los límites del intervalo:

$$IC_{90\%} \left(\frac{\mu_{\text{No-H}}}{\mu_{\text{H}}} \right) = [e^{-0.7979}, e^{-0.1129}] \approx [0.450, 0.893]$$

$$\Rightarrow IC_{90\%} \left(\frac{\mu_{\text{No-H}}}{\mu_{\text{H}}} \right) \approx [0.450, 0.893]$$

Este intervalo para la razón de medias no contiene el valor 1, lo que confirma nuestra conclusión. Específicamente, podemos afirmar con un 90% de confianza que la media de conteos de células T4 en los pacientes del grupo No-H es entre un 45% y un 89.3% de la media de los pacientes del grupo H.

Por lo tanto, los pacientes del grupo No-H tienen, en promedio, un conteo celular significativamente menor.

Problema 7

Los datos de la tabla en la siguiente hoja son números, n , de pólizas de seguros y los correspondientes números, y , de reclamos (esto es, número de accidentes en los que se pidió el amparo de la póliza). La variable CAR es una codificación de varias clases de carros, $EDAD$ es la edad del titular de la póliza y $DIST$ es el distrito donde vive el titular.

- Calcule la tasa de reclamos, y/n , para cada categoría y grafique estas tasas contra las diferentes variables para tener una idea de los efectos principales.
- Use regresión logística para estimar los efectos principales (cada variable tratada como categórica y modelada usando variables indicadoras) así como sus interacciones.
- Basados en los resultados del inciso anterior, los autores del artículo donde aparecieron estos datos, decidieron que ninguna interacción era importante y que podían considerar que CAR y $EDAD$ fuesen tratadas como variables continuas. Ajuste un modelo incorporando estas observaciones y compárelo con el obtenido en (b). ¿Cuáles son las conclusiones?.

CAR	EDAD	DIST = 0		DIST = 1	
		y	n	y	n
1	1	65	317	2	20
1	2	65	476	5	33
1	3	52	486	4	40
1	4	310	3259	36	316
2	1	98	486	7	31
2	2	159	1004	10	81
2	3	175	1355	22	122
2	4	877	7660	102	724
3	1	41	223	5	18
3	2	117	539	7	39
3	3	137	697	16	68
3	4	477	3442	63	344
4	1	11	40	0	3
4	2	35	148	6	16
4	3	39	214	8	25
4	4	167	1019	33	114

Table 0.7: Tabla de Pólizas de Seguros y Reclamos.

SOLUCIÓN

Se utilizó el lenguaje de programación R para resolver este problema, el código utilizado se encuentran al final del documento en el apéndice A.5

Buscamos modelar la tasa de reclamos utilizando un modelo de regresión logística, considerando como variables explicativas el tipo de carro (CAR), el grupo de edad del titular ($EDAD$) y el distrito de residencia ($DIST$). Para ello debemos evaluar dos modelos, uno completo con todas las interacciones posibles y un modelo simplificado con solo efectos principales, para determinar cuál de ellos explica mejor la variabilidad en los datos de forma más parsimoniosa.

a) Primero, calculamos las tasas de reclamo y/n para cada combinación de las variables categóricas y se generaron gráficos de cajas para visualizar la relación entre la tasa de reclamos y cada una de las variables predictoras como se muestra a continuación:

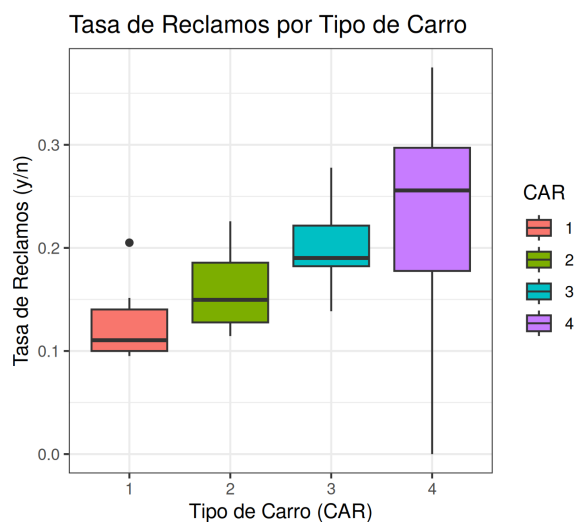


Figure 0.6: Tasa de Reclamos por tipo de carro.

En la Figura 0.6 se observa una tendencia positiva pues a medida que aumenta la categoría del carro, la mediana y la dispersión de la tasa de reclamos también tienden a aumentar, esto nos sugiere que los carros de categorías superiores están asociados a un mayor riesgo.

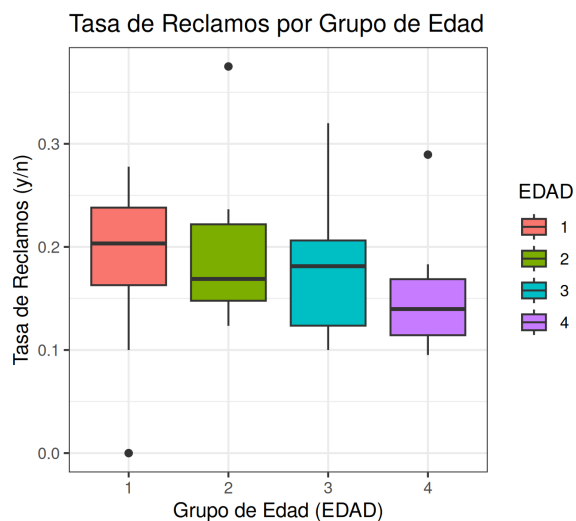


Figure 0.7: Tasa de Reclamos por Grupo de Edad.

En la Figura 0.7 el gráfico muestra una clara tendencia decreciente. Los grupos de edad más jóvenes (especialmente el grupo 1) presentan las tasas de reclamo más altas. La tasa disminuye consistentemente a medida que aumenta la edad, indicando que la edad es un factor protector.

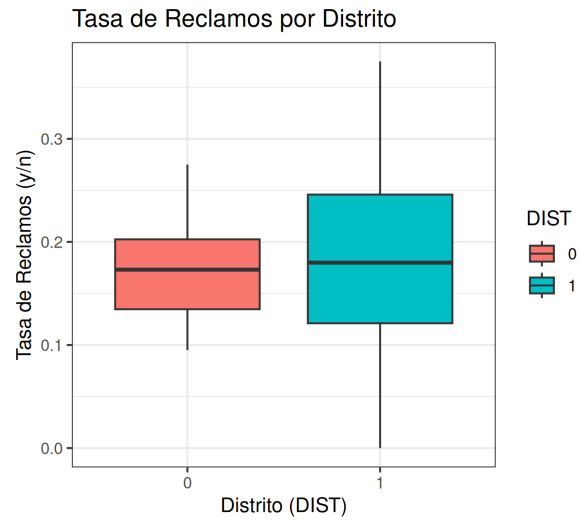


Figure 0.8: Tasa de Reclamos por Distrito

Por ultimo en la Figura 0.8 se aprecia una diferencia notable entre los dos distritos. El distrito 0 presenta tasas de reclamo con una mediana más baja en comparación con el distrito 1, lo que indica que la ubicación geográfica es un predictor potencialmente importante.

b) Se ajustó un primer GLM binomial con función de enlace logística. Este modelo incluyó todos los efectos principales y todas las interacciones de segundo y tercer orden entre *CAR*, *EDAD* y *DIST*, tratadas como variables categóricas. El resumen del ajuste se presenta a continuación:

Table 0.8: Resumen del modelo Logístico completo con interacciones

Predictor	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.3550	0.1391	-9.740	$< 2 \times 10^{-16}***$
CAR2	-0.0210	0.1793	-0.117	0.906760
CAR3	-0.1354	0.2219	-0.610	0.541753
CAR4	0.3856	0.3805	1.014	0.310756
EDAD2	-0.4892	0.1928	-2.537	0.011174*
EDAD3	-0.7668	0.2022	-3.792	0.000149***
EDAD4	-0.8976	0.1514	-5.929	$3.04 \times 10^{-9}***$
DIST1	-0.8422	0.7582	-1.111	0.266686
CAR2:EDAD2	0.1948	0.2396	0.813	0.416346
CAR3:EDAD2	0.6968	0.2792	2.495	0.012586*
CAR4:EDAD2	0.2865	0.4472	0.641	0.521707
CAR2:EDAD3	0.2343	0.2454	0.955	0.339707
CAR3:EDAD3	0.8492	0.2826	3.005	0.002654**
CAR4:EDAD3	0.2349	0.4446	0.528	0.597177
CAR2:EDAD4	0.2280	0.1923	1.185	0.235856
CAR3:EDAD4	0.5609	0.2350	2.387	0.017001*
CAR4:EDAD4	0.2374	0.3943	0.602	0.547094
CAR2:DIST1	0.9861	0.8788	1.122	0.261808
CAR3:DIST1	1.3771	0.9390	1.467	0.142493
CAR4:DIST1	-21.3479	37437.8581	-0.001	0.999545
EDAD2:DIST1	0.9636	0.9102	1.059	0.289733
EDAD3:DIST1	0.7668	0.9350	0.820	0.412179
EDAD4:DIST1	1.0436	0.7809	1.336	0.181439
CAR2:EDAD2:DIST1	-1.3972	1.0711	-1.304	0.192095
CAR3:EDAD2:DIST1	-1.7355	1.1490	-1.510	0.130932
CAR4:EDAD2:DIST1	21.8877	37437.8581	0.001	0.999534
CAR2:EDAD3:DIST1	-0.5163	1.0647	-0.485	0.627725
CAR3:EDAD3:DIST1	-1.0724	1.1278	-0.951	0.341658
CAR4:EDAD3:DIST1	22.1708	37437.8581	0.001	0.999527
CAR2:EDAD4:DIST1	-0.9497	0.9054	-1.049	0.294206
CAR3:EDAD4:DIST1	-1.2466	0.9688	-1.287	0.198168
CAR4:EDAD4:DIST1	21.8782	37437.8581	0.001	0.999534

Signif. codes: ***0.001, **0.01, *0.05

Null deviance: 244.33 on 31 d.f.

Residual deviance: 5.26×10^{-10} on 0 d.f.

AIC: 225.92

Podemos observar que la mayoría de los términos de interacción no son estadísticamente significativos, $\text{Pr}(> |z|) > 0.05$. Además, algunos coeficientes presentan errores estándar extremadamente grandes, por ejemplo *CAR4:DIST1*), lo que sugiere problemas de inestabilidad o cuasi-separación en el modelo. Este resultado indica que el modelo está sobreajustado y es innecesariamente complejo.

c) De acuerdo a los resultados anteriores, se ajustó un segundo modelo más parsimonioso. En este caso, las variables *CAR* y *EDAD* se trataron como continuas (numéricas) y solo se incluyeron los efectos principales, sin interacciones. El resumen del ajuste se muestra a continuación:

Table 0.9: Resumen del modelo Logístico simplificado de efectos principales

Predictor	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.66749	0.08748	-19.061	$< 2 \times 10^{-16}***$
CAR	0.23168	0.02266	10.225	$< 2 \times 10^{-16}***$
EDAD	-0.20967	0.02040	-10.278	$< 2 \times 10^{-16}***$
DIST1	0.25891	0.06420	4.033	$5.5 \times 10^{-5}***$

Signif. codes: ***0.001

Null deviance: 244.327 on 31 d.f.

Residual deviance: 30.086 on 28 d.f.

AIC: 200.01

En este modelo, todos los predictores son altamente significativos, por cada incremento en una unidad en la categoría del carro, el log-odds de un reclamo aumenta en 0.232, por cada incremento en una unidad en el grupo de edad, el log-odds de un reclamo disminuye en 0.210 y el log-odds de un reclamo en el distrito 1 es 0.259 mayor que en el distrito 0 (el nivel de referencia).

Para comparar formalmente ambos modelos, se utilizó un test de razón de verosimilitud (ANOVA). La hipótesis nula de esta prueba es que el modelo simple es suficiente para describir los datos, los resultados obtenidos se muestran a continuación:

Table 0.10: Comparación de modelos mediante Test de Razón de Verosimilitud (ANOVA)

Model	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
Simple	28	30.086			
Completo	0	0.000	28	30.086	0.3591

Podemos ver que el p -valor de la prueba es de 0.3591, el cual es considerablemente mayor que el nivel de significancia de 0.05, por lo tanto, no se rechaza la hipótesis nula, esto nos indica que los términos de interacción y la complejización del modelo completo no aportan una mejora estadísticamente significativa al ajuste. Adicionalmente, el Criterio de Información de Akaike (AIC) del modelo simple (AIC = 200.01) es sustancialmente menor que el del modelo completo (AIC = 225.92), lo que favorece fuertemente al modelo más simple.

Por lo tanto, el análisis demuestra que el modelo simplificado de efectos principales es superior al modelo completo con interacciones. Es más fácil de interpretar, más estable y tiene un mejor rendimiento según el criterio AIC, sin una pérdida significativa en la capacidad de ajuste, como lo demuestra el test ANOVA. Las conclusiones del estudio son que la probabilidad de reclamo de un seguro aumenta significativamente con la categoría del carro (*CAR*) y para los residentes del distrito 1, mientras que disminuye a medida que aumenta el grupo de edad (*EDAD*) del titular.

Problema 8

A lo largo del curso hemos enfatizado el uso del método de Máxima Verosimilitud para todo lo relacionado con estimación. Consideremos ahora una alternativa: El método de la Mínima Ji-Cuadrada. Suponga que las celdas de una multinomial están parametrizadas en términos de un vector $\theta = (\theta_1, \dots, \theta_s)^T$. El método de la mínima ji-cuadrada (ver Agresti, pág. 611) consiste en estimar θ mediante aquel valor que minimice el estadístico de Pearson

$$\chi^2 = \sum \frac{(\text{obs} - \text{esp})^2}{\text{esp}} = \sum_{j=1}^K \frac{(y_j - n\pi_j(\theta))^2}{n\pi_j(\theta)}$$

Considere el siguiente problema. Suponga una población muy grande de objetos que pueden clasificarse en tres categorías, A, B y C. Para estimar las proporciones π_1, π_2 y π_3 correspondientes a cada una de esas categorías, se efectuó un estudio; se obtuvieron tres muestras de tamaños n_1, n_2 y n_3 tomadas de la población global, sin embargo, en vez de registrar la frecuencia observada de A's, B's y C's de cada muestra, lo que se hizo fue anotar:

- Número de A's en la muestra de tamaño $n_1 = y_1$
- Número de B's en la muestra de tamaño $n_2 = y_2$
- Número de C's en la muestra de tamaño $n_3 = y_3$

Estime π_1, π_2 y π_3 usando el método de la mínima ji-cuadrada; suponga que $n_1 = 100, y_1 = 22, n_2 = 150, y_2 = 52, n_3 = 200, y_3 = 77$. Esto es, encuentre π_1, π_2 y π_3 que minimicen

$$\frac{(y_1 - n_1\pi_1)^2}{n_1\pi_1} + \frac{[(n_1 - y_1) - n_1(1 - \pi_1)]^2}{n_1(1 - \pi_1)} + \dots + \frac{(y_3 - n_3\pi_3)^2}{n_3\pi_3} + \frac{[(n_3 - y_3) - n_3(1 - \pi_3)]^2}{n_3(1 - \pi_3)}$$

con la restricción $\pi_3 = 1 - \pi_1 - \pi_2$ (sugerimos usar directamente `nlminb` de R).

SOLUCIÓN

Se utilizó el lenguaje de programación R para resolver este problema, el código utilizado se encuentran al final del documento en el apéndice A.6

De acuerdo a la definición dada del estadístico χ^2 de Pearson, tenemos que para la categoría A, la contribución al estadístico es:

$$\chi_1^2 = \frac{(y_1 - n_1\pi_1)^2}{n_1\pi_1} + \frac{((n_1 - y_1) - n_1(1 - \pi_1))^2}{n_1(1 - \pi_1)} \quad (34)$$

Simplificando el segundo numerador llegamos a que:

$$\begin{aligned} \chi_1^2 &= \frac{(y_1 - n_1\pi_1)^2}{n_1\pi_1} + \frac{(y_1 - n_1\pi_1)^2}{n_1(1 - \pi_1)} \\ &= (y_1 - n_1\pi_1)^2 \left[\frac{1}{n_1\pi_1} + \frac{1}{n_1(1 - \pi_1)} \right] \\ &= (y_1 - n_1\pi_1)^2 \left[\frac{1 - \pi_1 + \pi_1}{n_1\pi_1(1 - \pi_1)} \right] \\ &= \frac{(y_1 - n_1\pi_1)^2}{n_1\pi_1(1 - \pi_1)} \end{aligned}$$

Generalizando para las tres muestras, la función objetivo completa $Q(\boldsymbol{\pi})$ a minimizar sería:

$$Q(\boldsymbol{\pi}) = \frac{(y_1 - n_1\pi_1)^2}{n_1\pi_1(1 - \pi_1)} + \frac{(y_2 - n_2\pi_2)^2}{n_2\pi_2(1 - \pi_2)} + \frac{(y_3 - n_3\pi_3)^2}{n_3\pi_3(1 - \pi_3)} \quad (35)$$

Las proporciones deben sumar 1, siendo la restricción:

$$\pi_1 + \pi_2 + \pi_3 = 1 \implies \pi_3 = 1 - \pi_1 - \pi_2$$

Sustituyendo π_3 en Q , obtenemos una función de dos variables:

$$Q(\pi_1, \pi_2) = \frac{(y_1 - n_1\pi_1)^2}{n_1\pi_1(1 - \pi_1)} + \frac{(y_2 - n_2\pi_2)^2}{n_2\pi_2(1 - \pi_2)} + \frac{(y_3 - n_3(1 - \pi_1 - \pi_2))^2}{n_3(1 - \pi_1 - \pi_2)(\pi_1 + \pi_2)} \quad (36)$$

Esta función es compleja para minimizar analíticamente, por lo que se utilizó optimización numérica mediante la función `nlminb` en R como se sugiere. El optimizador convergió exitosamente, arrojando los siguientes parámetros:

$$\hat{\pi}_1 = 0.2443 \quad \text{y} \quad \hat{\pi}_2 = 0.3665 \quad (37)$$

El valor mínimo del estadístico χ^2 fue de 1.0539. Calculamos $\hat{\pi}_3$ a partir de la restricción:

$$\hat{\pi}_3 = 1 - \hat{\pi}_1 - \hat{\pi}_2 = 1 - 0.2443 - 0.3665 = 0.3892 \quad (38)$$

Por lo tanto, las proporciones poblacionales estimadas por el método de la Mínima Ji-Cuadrada son:

- **Proporción de A ($\hat{\pi}_1$):** 24.43%
- **Proporción de B ($\hat{\pi}_2$):** 36.65%
- **Proporción de C ($\hat{\pi}_3$):** 38.92%

Problema 9

Se toman los datos relacionados con el hundimiento del Titanic en abril de 1912. El resultado se puede expresar en una tabla de dimensión 4.

Las variables son **Class** de los pasajeros (1, 2, 3, Tripulación), **Sex** de los pasajero (Male, Female), **Age** de los pasajeros (Child, Adult), y **Survived** si los pasajeros sobrevivieron o no (No, Yes). Usar librería en R “titanic” y los datos se encuentran en la variable “Titanic”.

Considerar entonces un modelo log-lineal para analizar los posibles efectos:

- **Class**: Hay más pasajeros en algunas clases que en otras.
- **Sex**: Hay más pasajeros en un sexo que en otro.
- **Age**: Hay más pasajeros en un grupo de edad que en otro.
- **Survived**: Hay más pasajeros o vivos o muertos que la alternativa.
- **Class × Sex**: **Class** y **Sex** no son independientes.
- **Class × Age**: **Class** y **Age** no son independientes.
- **Class × Survived**: **Class** y **Survived** no son independientes.
- **Sex × Age**: **Sex** y **Age** no son independientes.
- **Sex × Survived**: **Sex** y **Survived** no son independientes.
- **Age × Survived**: **Age** y **Survived** no son independientes.
- **Class × Sex × Age**, **Class × Sex × Survived**, **Class × Age × Survived**, **Sex × Age × Survived**: hay interacción triple entre las variables.
- **Class × Sex × Age × Survived**: hay interacción cuádruple entre las variables.

SOLUCIÓN

Se utilizo el lenguaje de programación R para resolver este problema, el código utilizado se encuentran al final del documento en el apéndice A.7

Buscamos identificar las asociaciones e interacciones significativas entre estas variables para entender los patrones de supervivencia en el desastre del Titanic. Primero se comenzó con el ajuste de un “modelo saturado”, que incluye todos los efectos principales y todas las interacciones posibles hasta el cuarto orden (**Class × Sex × Age × Survived**). Este modelo se ajusta perfectamente a los datos, como lo indica una devianza residual de prácticamente cero, pero es inherentemente complejo.

Para obtener un modelo más parsimonioso, se utilizó una estrategia de eliminación hacia atrás, comenzando por evaluar la significancia del término de interacción de orden más alto. Se utilizó la función `drop1()` para realizar un análisis de devianza, cuyos resultados se muestran a continuación:

Table 0.11: Análisis de devianza para el modelo saturado

Término a eliminar	Df	Devianza	AIC	Pr(>Chi)
<none>		4.46e-10	191.4	
Class:Sex:Age:Survived	3	4.24e-10	185.4	1.000

Podemos observar que el p -valor asociado a la interacción de cuarto orden (**Class:Sex:Age:Survived**) es de 1.0, lo que indica una falta total de significancia estadística. Esto nos sugiere que cualquier relación entre tres de las variables no cambia a través de los niveles de la cuarta variable, por lo tanto, este término puede ser eliminado del modelo sin una pérdida significativa de ajuste.

Con base en este resultado, se ajustó un segundo modelo que incluye únicamente interacciones de hasta tres vías (**Model 1**). Para confirmar formalmente que este modelo más simple es suficiente, se comparó con el modelo saturado (**Model 2**) mediante un test de ANOVA:

Table 0.12: Comparación ANOVA de modelos anidados

Modelo	Resid. Df	Resid. Dev	Df	Devianza	Pr(>Chi)
1 (3-vías)	3	4.24e-10			
2 (Saturado)	0	4.46e-10	3	-2.25e-11	

Esta comparación confirma que el modelo saturado no ofrece una mejora significativa sobre el modelo con interacciones de tres vías. Por lo tanto seleccionamos el modelo de interacciones de tres vías como el modelo final por ser el más parsimonioso que describe adecuadamente las relaciones en los datos.

Los coeficientes del modelo final se presentan a continuación:

Table 0.13: Coeficientes del modelo de interacciones de tres vías

Coeficiente	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.292e+01	5.750e+04	0.000	0.999682
SexFemale	-5.206e+00	1.326e+00	-3.925	8.68e-05 ***
Class3rd:SexFemale	4.483e+00	1.293e+00	3.468	0.000525 ***
SexFemale:SurvivedYes	3.596e+00	7.478e-01	4.809	1.52e-06 ***
Class3rd:SexFemale:AgeAdult	-2.569e+00	1.183e+00	-2.171	0.029895 *
Class3rd:SexFemale:SurvivedYes	-2.800e+00	5.687e-01	-4.923	8.52e-07 ***

Nota: Solo se muestran los términos más relevantes y significativos para la discusión.

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

De la tabla de coeficientes, se extraen las siguientes conclusiones clave:

- **Interacción Sex:Survived (***)**: El término **SexFemale:SurvivedYes** es altamente significativo. Esto confirma la conocida política de “mujeres y niños primero”, indicando que la probabilidad de supervivencia estaba fuertemente asociada con ser mujer.
- **Interacción Class:Sex (***)**: La interacción **Class3rd:SexFemale** es también significativa. Esto sugiere que la distribución de hombres y mujeres no era homogénea a través de las clases, la proporción de mujeres en tercera clase era diferente a la de primera clase (categoría de referencia).
- **Interacción Class:Sex:Survived (***)**: El término **Class3rd:SexFemale:SurvivedYes** también es altamente significativo. Esto implica que la ventaja de supervivencia para las mujeres (observada en la interacción **Sex:Survived**) **no era la misma en todas las clases**. Específicamente, la ventaja de supervivencia de ser mujer era significativamente diferente (en este caso, menor) en la tercera clase en comparación con la primera clase.
- **Interacción Class:Sex:Age (*)**: El término **Class3rd:SexFemale:AgeAdult** es significativo, lo que indica que la combinación de ser una mujer adulta en tercera clase ocurría con una frecuencia distinta a la que se esperaría si estos factores fueran independientes.

Por lo tanto, el análisis log-lineal no solo confirma los efectos principales conocidos (que la clase, el sexo y la edad influyeron en la supervivencia), sino que también cuantifica las complejas interacciones entre ellos. El hallazgo más importante es que los factores no actuaron de forma aislada pues el privilegio de la clase social moduló la ventaja de supervivencia otorgada a las mujeres, pintando un cuadro más matizado de la dinámica social a bordo del Titanic..

Problema 10

Se ha realizado un análisis sobre el valor terapéutico del ácido ascórbico (vitamina C) en relación a su efecto sobre la gripe común. Se tiene una tabla 2×2 con los recuentos correspondientes para una muestra de 279 personas:

	Gripe	No Gripe	Totales
Placebo	31	109	140
Ácido Ascórbico	17	122	139
Totales	48	231	279

Aplicar un modelo lineal para determinar si existe evidencia suficiente para asegurar que el ácido ascórbico ayuda a tener menos gripe.

SOLUCIÓN

Se utilizo el lenguaje de programación R para resolver este problema, el código utilizado se encuentran al final del documento en el apéndice A.8

Para evaluar la relación entre el tipo de tratamiento y la probabilidad de contraer gripe, se ajustó un GLM con una familia binomial y una función de enlace logística. Sabemos que tiene la forma:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 \quad (39)$$

Para este problema particular π es la probabilidad de contraer gripe, β_0 es el log-odds de contraer gripe para el grupo de referencia (placebo), X_1 sera la variable indicadora que toma el valor 1 si el tratamiento es ácido ascórbico y 0 si es placebo, por ultimo β_1 representa el cambio en el log-odds de contraer gripe al pasar del grupo placebo al grupo de ácido ascórbico. Buscamos estimar los coeficientes β_0 y β_1 y evaluar si β_1 es estadísticamente diferente de cero. Al ajustar el modelo de regresión logística a los datos, se obtuvieron los siguientes resultados :

Table 0.14: Resumen del ajuste del modelo para el tratamiento

	Estimate	Std. Error	z value	Pr(> z)
Intercept	-1.2574	0.2035	-6.177	$6.53 \times 10^{-10}***$
TratamientoAcidoAscorbico	-0.7134	0.3293	-2.166	0.0303*

Signif. codes: ***0.001, **0.01, *0.05, .0.1

Null deviance: 4.8717 on 1 d.f.

Residual deviance: 7.55×10^{-15} on 0 d.f.

AIC: 13.578

Fisher Scoring iterations: 3

Observamos que se obtuvo un $\hat{\beta}_0 = -1.2574$ y $\hat{\beta}_1 = -0.7134$ este valor indica que el log-odds de contraer gripe para el grupo que tomó ácido ascórbico es 0.7134 unidades menor que el del grupo Placebo. El p -valor asociado al coeficiente del tratamiento con ácido ascórbico es **0.0303**. Dado que este valor es menor que el nivel de significancia estándar, $\alpha = 0.05$, se rechaza la hipótesis nula de que el coeficiente es igual a cero, es decir $H_0 : \beta_1 = 0$. Esto implica que el tipo de tratamiento tiene un efecto estadísticamente significativo en la probabilidad de contraer gripe.

Para completar esta interpretación, se calculo el Odds Ratio (OR):

$$\text{OR} = e^{\hat{\beta}_1} = e^{-0.7134} \approx 0.49 \quad (40)$$

Un Odds Ratio de **0.49** nos dice que las odds (chances) de contraer gripe en el grupo que recibió ácido ascórbico son aproximadamente el 49% de las *odds* del grupo que recibió el placebo. Dicho de otra manera, el tratamiento con ácido ascórbico reduce las odds de contraer gripe en un 51%.

Por lo tanto, el análisis del GLM nos proporciona **evidencia estadística suficiente** para asegurar que el ácido ascórbico tiene un efecto protector contra la gripe común en la población estudiada. El efecto es estadísticamente significativo, con $p = 0.0303$, y un Odds Ratio de 0.49 indica que el tratamiento reduce a casi la mitad las “chances” de enfermarse en comparación con el placebo.

A Código en R

A continuación, se presentan los scripts de R utilizados para cada uno de los problemas resueltos en este reporte.

A.1 Problema 2

```

1 # =====
2 # Problema 2
3 # Script: Ronquido y enfermedad cardíaca
4 # Autor: Diego Paniagua Molina
5 # Fecha: 2025-09-18
6 # =====
7
8 # Carpetas -----
9 dir_report <- file.path("report")
10 dir_models <- file.path("report", "models")
11 dir_figs <- file.path("report", "figures")
12
13 for (d in c(dir_report, dir_models, dir_figs)) {
14   if (!dir.exists(d)) dir.create(d, recursive = TRUE, showWarnings = FALSE)
15 }
16
17 # Datos -----
18 snore <- c("Nunca", "Ocasional", "Casi cada noche", "Cada noche")
19 score <- c(0, 2, 4, 5) # puntuaciones relativas
20 SI <- c(24, 35, 21, 30)
21 NO <- c(1355, 603, 192, 224)
22
23 df <- data.frame(
24   snore = factor(snore, levels = snore), # convierte a factor y preserva orden
25   score = score,
26   SI = SI,
27   NO = NO
28 )
29 df$tot <- df$SI + df$NO
30 df$prop <- df$SI / df$tot
31
32 cat("\n=== Datos ===\n")
33 print(df)
34
35 # Modelos GLM: logit y probit -----
36 fit_logit <- glm(cbind(SI, NO) ~ score, data = df,
37   family = binomial(link = "logit"))
38 fit_probit <- glm(cbind(SI, NO) ~ score, data = df,
39   family = binomial(link = "probit"))
40
41 cat("\n=== Resumen LOGIT ===\n")
42 print(summary(fit_logit))
43 cat("\n=== Resumen PROBIT ===\n")
44 print(summary(fit_probit))
45
46 # Guardar modelos -----
47 saveRDS(fit_logit, file.path(dir_models, "snoring_logit.rds"))
48 saveRDS(fit_probit, file.path(dir_models, "snoring_probit.rds"))
49
50 # Figura: observados vs curvas logit/probit -----
51 png(file.path(dir_figs, "snoring_glm_curves.png"),
52   width = 1200, height = 800, res = 140)
53
54 op <- par(mar=c(4.2, 4.2, 1, 1))
55 plot(df$score, df$prop, pch = 16,
56   ylim = c(0, max(df$prop, 0.15)),
57   xlab = "Puntuación de ronquido",
58   ylab = "Proporción (Enfermedad cardíaca)",

```

```

59     xaxt = "n")
60 axis(1, at = df$score, labels = df$snore)
61 abline(h = pretty(c(0, max(df$prop, 0.15))), col = "gray90", lty = 1)
62
63 grid_sc <- data.frame(score = seq(0, 5, length.out = 301))
64 lines(grid_sc$score, predict(fit_logit, newdata = grid_sc, type = "response"), lwd = 3)
65 lines(grid_sc$score, predict(fit_probit, newdata = grid_sc, type = "response"), lwd = 3, lty
    = 2)
66
67 legend("topleft", bty = "n", lwd = 3, lty = c(1,2),
68       legend = c("Logit", "Probit"))
69 par(op)
70 dev.off()
71
72 # Mensaje de confirmacion -----
73 cat("\nArchivos generados:\n")
74 cat(" - Modelos: report/models/snoring_logit.rds, snoring_probit.rds\n")
75 cat(" - Figura  : report/figures/snoring_glm_curves.png\n\n")

```

Listing 2: Script: Ronquido y enfermedad cardiaca

A.2 Problema 3

```

1  # =====
2  # Problema 3
3  # Script: Cangrejos cacerola
4  # Autor: Diego Paniagua Molina
5  # Fecha: 2025-09-20
6  # =====
7
8
9  # Cargar librerias -----
10 library(here)
11 library(dplyr)
12
13 # Datos -----
14 crabs_df <- data.frame(
15   color = c(3, 4, 2, 4, 4, 3, 2),
16   Spine = c(3, 3, 1, 3, 3, 3, 1),
17   Width = c(28.3, 22.5, 26.0, 24.8, 26.0, 23.8, 26.5),
18   Satellite = c(8, 0, 9, 0, 4, 0, 0),
19   Weight = c(3050, 1550, 2300, 2100, 2600, 2100, 2350)
20 )
21
22 # Preparacion de datos -----
23 # Las variables 'color' y 'Spine' son categóricas. Las convertimos a factores.
24 crabs_df <- crabs_df %>%
25   mutate(
26     color = as.factor(color),
27     Spine = as.factor(Spine)
28   )
29
30 print("Estructura de los datos:")
31 glimpse(crabs_df)
32
33 # Ajuste GLM Poisson -----
34 poisson_model <- glm(Satellite ~ color + Spine + Width + Weight,
35   data = crabs_df,
36   family = poisson(link = "log"))
37
38 # Guardar resultados -----
39 model_summary <- summary(poisson_model)
40
41 print("Resumen del Modelo de Poisson:")
42 print(model_summary)

```

Listing 3: Script: Cangrejos cacerola

A.3 Problema 5

```

1
2 # =====
3 # Problema 5
4 # Script: Edad y daño coronario
5 # Autor: Diego Paniagua Molina
6 # Fecha: 2025-09-19
7 # =====
8
9 # Datos -----
10 edad <- c(
11   20, 23, 24, 25, 25, 26, 26, 28, 28, 29, 30, 30, 30, 30, 30, 30, 32, 32, 33, 33,
12   34, 34, 34, 34, 34, 35, 35, 36, 36, 37, 37, 37, 38, 38, 39, 39, 40, 40, 41, 41,
13   42, 42, 42, 42, 43, 43, 43, 44, 44, 44, 44, 45, 45, 46, 46, 47, 47, 47, 48, 48,
14   48, 49, 49, 49, 50, 50, 51, 52, 52, 53, 53, 54, 55, 55, 55, 56, 56, 56, 57, 57,
15   57, 57, 57, 58, 58, 58, 59, 59, 60, 60, 61, 62, 62, 63, 64, 64, 65, 69)
16
17 coro <- c(
18   0,0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,1,0,0,0,1,0,1,0,
19   0,0,0,1,0,0,1,0,1,1,0,1,0,1,1,0,1,0,0,1,0,1,1,1,0,1,1,1,1,1,1,0,
20   0,1,1,1,1,0,1,1,1,1,1,0,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
21
22 stopifnot(length(edad) == length(coro))
23 y <- as.integer(coro) # 0/1
24 n <- length(y)
25 P <- sum(y == 1L) # Positivos observados
26 N <- sum(y == 0L) # Negativos observados
27
28
29 # Modelo logístico -----
30 mod <- glm(y ~ edad, family = binomial)
31 phat <- as.numeric(predict(mod, type = "response"))
32
33 # Curva ROC (barrido de umbrales) -----
34 thr_unique <- sort(unique(phat))
35 thr <- c(Inf, rev(thr_unique), -Inf) # asegura extremos (0,0) y (1,1)
36
37 TPR <- FPR <- SPEC <- SENS <- J <- numeric(length(thr))
38
39 for (i in seq_along(thr)) {
40   t <- thr[i]
41   yhat <- as.integer(phat >= t)
42   TP <- sum(yhat == 1L & y == 1L)
43   FP <- sum(yhat == 1L & y == 0L)
44   TN <- sum(yhat == 0L & y == 0L)
45   FN <- sum(yhat == 0L & y == 1L)
46
47   SENS[i] <- if (P > 0) TP / P else NA_real_ # TPR
48   SPEC[i] <- if (N > 0) TN / N else NA_real_ # Specificity
49   TPR[i] <- SENS[i]
50   FPR[i] <- 1 - SPEC[i]
51
52   J[i] <- SENS[i] - (1 - SPEC[i]) # = TPR - FPR = sens + espec - 1
53 }
54
55 # Ordenamos por FPR ascendente (eje x) para AUC con trapecios
56 ord <- order(FPR, TPR)
57 FPRo <- FPR[ord]
58 TPRo <- TPR[ord]
59
60 # AUC (regla del trapecio) -----
61 AUC <- sum(diff(FPRo) * (head(TPRo, -1) + tail(TPRo, -1)) / 2)
62
63 # Umbral óptimo por Youden J -----
64 Jmax <- max(J, na.rm = TRUE)
65 cand <- which(J == Jmax)
66

```

```

67 # Elegimos el candidato con menor FPR; si empata, mayor TPR
68 if (length(cand) > 1L) {
69   sub <- cand[order(FPR[cand], -TPR[cand])]
70   idx_opt <- sub[1L]
71 } else {
72   idx_opt <- cand
73 }
74 t_opt <- thr[idx_opt]
75 tpr_opt <- TPR[idx_opt]
76 fpr_opt <- FPR[idx_opt]
77 spec_opt <- 1 - fpr_opt
78 sens_opt <- tpr_opt
79
80 # Matriz de confusión en t_opt y métricas -----
81 yhat_opt <- as.integer(phat >= t_opt)
82 TP <- sum(yhat_opt == 1L & y == 1L)
83 FP <- sum(yhat_opt == 1L & y == 0L)
84 TN <- sum(yhat_opt == 0L & y == 0L)
85 FN <- sum(yhat_opt == 0L & y == 1L)
86
87 acc <- (TP + TN) / (P + N)
88 prec <- if ((TP + FP) > 0) TP / (TP + FP) else NA_real_
89
90 # Gráfica ROC -----
91 png("results/figures/roc_coro.png", width = 800, height = 800)
92 plot(FPRo, TPRo, type = "l", lwd = 2,
93      xlab = "FPR = 1 - Especificidad",
94      ylab = "TPR = Sensitividad",
95      main = "Curva ROC: daño coronario vs edad")
96 abline(0, 1, lty = 2) # diagonal
97 points(fpr_opt, tpr_opt, pch = 19, cex = 1.2)
98 text(fpr_opt, tpr_opt,
99      labels = sprintf(" t*=%.3f\n J=%.3f", t_opt, Jmax),
100     pos = 4)
101 legend("bottomright",
102       legend = sprintf("AUC (trapecios) = %.4f", AUC),
103       bty = "n")
104 dev.off()
105
106 cat("Imagen guardada en: report/figures/roc_coro.png\n")
107
108 # Resumen de resultados -----
109 cat("\n===== RESUMEN =====\n")
110 cat(sprintf("AUC (trapecios): %.6f\n", AUC))
111 cat(sprintf("Umbral óptimo (Youden J): t* = %.6f\n", t_opt))
112 cat(sprintf("Sensitividad (TPR): %.4f\n", sens_opt))
113 cat(sprintf("Especificidad: %.4f\n", spec_opt))
114 cat(sprintf("FPR: %.4f\n", fpr_opt))
115 cat(sprintf("J = TPR - FPR: %.4f\n", Jmax))
116
117 cat("\nMatriz de confusión en t*:\n")
118 tab <- matrix(c(TP, FP, FN, TN), nrow = 2, byrow = TRUE,
119             dimnames = list("Decisión" = c("1", "0"),
120                             "Observado" = c("1", "0")))
121 print(tab)
122 cat(sprintf("\nAccuracy: %.4f", acc))
123 cat(sprintf("\nPrecision: %.4f", prec))

```

Listing 4: Script: Edad y daño coronario

A.4 Problema 6

```

1 # =====
2 # Problema 6
3 # Script 1: Carga y Limpieza de Datos - Conteos de Celulas T4
4 # Autor: Diego Paniagua Molina
5 # Fecha: 2025-08-29
6 # =====
7
8 # Instalar paquetes necesarios -----
9 if (!require("dplyr")) install.packages("dplyr")
10 if (!require("readr")) install.packages("readr")
11 if (!require("here")) install.packages("here")
12
13 # Cargar paquetes -----
14 library(dplyr)
15 library(readr)
16 library(here)
17
18 # Crear los datos manualmente -----
19 datos <- tibble(
20   grupo = rep(c("H", "No-H"), each = 20),
21   conteo = c(
22     # Grupo H (Hodgkin) - Primera fila
23     396, 568, 1212, 171, 554, 1104, 257, 435, 295, 397,
24     # Grupo H (Hodgkin) - Segunda fila
25     288, 1004, 431, 795, 1621, 1378, 902, 958, 1283, 2415,
26     # Grupo No-H (No Hodgkin) - Primera fila
27     375, 375, 752, 208, 151, 116, 736, 192, 315, 1252,
28     # Grupo No-H (No Hodgkin) - Segunda fila
29     675, 700, 440, 771, 688, 426, 410, 979, 377, 503
30   )
31 )
32
33 # Verificar estructura de los datos -----
34 cat("Estructura de los datos:\n")
35 glimpse(datos)
36
37 cat("\nResumen por grupos:\n")
38 datos %>%
39   group_by(grupo) %>%
40   summarise(
41     n = n(),
42     media = mean(conteo),
43     mediana = median(conteo),
44     sd = sd(conteo),
45     min = min(conteo),
46     max = max(conteo)
47   ) %>%
48   print()
49
50 # Guardar datos en formato CSV -----
51 dir.create(here("data"), recursive = TRUE, showWarnings = FALSE)
52 write_csv(datos, here("data", "raw", "datos_pacientes.csv"))
53
54 # Mensaje de confirmacion -----
55 cat("\nDatos creados y guardados en: data/raw/datos_pacientes.csv\n")
56 cat("Total de observaciones:", nrow(datos), "\n")
57 cat("Grupos:", unique(datos$grupo), "\n")

```

Listing 5: Script 1: Carga y Limpieza de Datos - Conteos de Células T4.

```

1 # =====
2 # Problema 6
3 # Script 2: Analisis Exploratorio - Comparacion Grafica (Inciso a)
4 # =====
5
6 # Instalar paquetes necesarios -----
7 if (!require("dplyr")) install.packages("dplyr")
8 if (!require("ggplot2")) install.packages("ggplot2")
9 if (!require("here")) install.packages("here")
10
11 # Cargar paquetes -----
12 library(dplyr)
13 library(ggplot2)
14 library(here)
15
16 # Cargar datos -----
17 datos <- read_csv(here("data", "raw", "datos_pacientes.csv"), show_col_types = FALSE)
18
19 # 1. Boxplot comparativo con puntos -----
20 boxplot <- ggplot(datos, aes(x = grupo, y = conteo, fill = grupo)) +
21   geom_boxplot(alpha = 0.7) +
22   geom_point(position = position_jitter(width = 0.2), alpha = 0.6) +
23   labs(title = "Comparacion de Conteos de Celulas T4",
24        subtitle = "Pacientes con Hodgkin (H) vs Otras enfermedades (No-H)",
25        x = "Grupo",
26        y = "Conteo por mm³") +
27   theme_bw() +
28   theme(legend.position = "none")
29
30 # 2. Histogramas comparativos -----
31 histogramas <- ggplot(datos, aes(x = conteo, fill = grupo)) +
32   geom_histogram(alpha = 0.7, position = "identity", color = "black") +
33   facet_wrap(~ grupo, nrow = 1) +
34   labs(title = "Distribucion de Conteos por Grupo",
35        x = "Conteo por mm³",
36        y = "Frecuencia") +
37   theme_bw() +
38   theme(aspect.ratio = 1,
39        legend.position = "none",
40        strip.background = element_blank(),
41        strip.text = element_text(face = "bold", size = 11))
42
43 # 3. Grafico de densidad -----
44 densidad <- ggplot(datos, aes(x = conteo, fill = grupo)) +
45   geom_density(alpha = 0.5) +
46   labs(title = "Densidad de Conteos de Celulas T4",
47        x = "Conteo por mm³",
48        y = "Densidad") +
49   theme_bw()
50
51 # Guardar graficos -----
52 dir.create(here("results", "figures"), recursive = TRUE, showWarnings = FALSE)
53 ggsave(here("results", "figures", "boxplot_comparativo.png"),
54        plot = boxplot, width = 8, height = 6, dpi = 300)
55 ggsave(here("results", "figures", "histogramas_comparativos.png"),
56        plot = histogramas, width = 8, height = 6, dpi = 300)
57 ggsave(here("results", "figures", "densidad_comparativa.png"),
58        plot = densidad, width = 8, height = 6, dpi = 300)
59
60 # Mostrar graficos -----
61 print(boxplot)
62 print(histogramas)
63 print(densidad)
64
65 cat("\n Graficos guardados en la carpeta: results/figures/\n")

```

Listing 6: Script 2: Analisis Exploratorio - Comparacion Grafica (Inciso a).

```

1 # =====
2 # Problema 6
3 # Script 3: Modelo de Poisson e Inferencia (Inciso b)
4 # =====
5
6 # Instalar paquetes necesarios -----
7 if (!require("dplyr")) install.packages("dplyr")
8 if (!require("readr")) install.packages("readr")
9 if (!require("here")) install.packages("here")
10
11 # Cargar paquetes -----
12 library(dplyr)
13 library(readr)
14 library(here)
15
16 # Cargar datos -----
17 datos <- read_csv(here("data", "raw", "datos_pacientes.csv"), show_col_types = FALSE)
18
19 # INCISO b: Ajustar modelo de Poisson -----
20 cat("\nAjustando modelo de Poisson...\n")
21
22 # Modelo: log(conteo) = beta_0 + beta_1*grupo
23 modelo_poisson <- glm(conteo ~ grupo,
24                       family = poisson(link = "log"),
25                       data = datos)
26
27 # Resumen del modelo -----
28 cat("\nResumen del modelo de Poisson:\n")
29 resumen_modelo <- summary(modelo_poisson)
30 print(resumen_modelo)
31
32 # Verificar sobredispersión -----
33 cat("\nVerificación de sobredispersión:\n")
34 deviance_val <- resumen_modelo$deviance
35 df_residual <- resumen_modelo$df.residual
36 ratio_sobredispersión <- deviance_val / df_residual
37
38 cat("Deviance:", deviance_val, "\n")
39 cat("Grados de libertad:", df_residual, "\n")
40 cat("Ratio Deviance/df:", ratio_sobredispersión, "\n")
41
42 # Guardar modelo -----
43 dir.create(here("results", "models"), recursive = TRUE, showWarnings = FALSE)
44 saveRDS(modelo_poisson, here("results", "models", "modelo_poisson.rds"))
45 cat("\nModelo guardado en: results/models/modelo_poisson.rds\n")
46
47 # Existe sobredispersión, ajustar modelo quasi-Poisson -----
48 cat("\nAjustando modelo quasi-Poisson...\n")
49
50 modelo_quasi <- glm(conteo ~ grupo,
51                    family = quasipoisson(link = "log"),
52                    data = datos)
53
54 # Resumen del modelo -----
55 cat("\nResumen del modelo quasi-Poisson:\n")
56 resumen_quasi <- summary(modelo_quasi)
57 print(resumen_quasi)
58
59 # Dispersión (phi) -----
60 phi <- resumen_quasi$dispersion
61 cat("\nEstimador de dispersión (phi):", round(phi, 3), "\n")
62
63 # Guardar modelo -----
64 saveRDS(modelo_quasi, here("results", "models", "modelo_quasipoisson.rds"))
65 cat("\nModelo guardado en: results/models/modelo_quasipoisson.rds\n")

```

Listing 7: Script 3: Modelo de Poisson e Inferencia (Inciso b).

A.5 Problema 7

```

1  # =====
2  # Problema 7
3  # Script: Reclamos polizas de seguro
4  # Autor: Diego Paniagua Molina
5  # Fecha: 2025-09-20
6  # =====
7
8
9  # Librerias -----
10 library(here)
11 library(dplyr)
12 library(tidyr)
13 library(ggplot2)
14
15 # Dataset -----
16 insurance_wide_df <- tibble::tribble(
17   ~CAR, ~EDAD, ~y_dist0, ~n_dist0, ~y_dist1, ~n_dist1,
18   1, 1, 65, 317, 2, 20,
19   1, 2, 65, 476, 5, 33,
20   1, 3, 52, 486, 4, 40,
21   1, 4, 310, 3259, 36, 316,
22   2, 1, 98, 486, 7, 31,
23   2, 2, 159, 1004, 10, 81,
24   2, 3, 175, 1355, 22, 122,
25   2, 4, 877, 7660, 102, 724,
26   3, 1, 41, 223, 5, 18,
27   3, 2, 117, 539, 7, 39,
28   3, 3, 137, 697, 16, 68,
29   3, 4, 477, 3442, 63, 344,
30   4, 1, 11, 40, 0, 3,
31   4, 2, 35, 148, 6, 16,
32   4, 3, 39, 214, 8, 25,
33   4, 4, 167, 1019, 33, 114
34 )
35
36 # Convertimos a formato largo
37 insurance_df <- insurance_wide_df %>%
38   pivot_longer(
39     cols = c(y_dist0, n_dist0, y_dist1, n_dist1),
40     names_to = c(".value", "DIST"),
41     names_pattern = "([yn])_dist(.)"
42   )
43
44 # Inciso (a): EDA -----
45 eda_df <- insurance_df %>%
46   mutate(
47     rate = y / n,
48     CAR = as.factor(CAR),
49     EDAD = as.factor(EDAD),
50     DIST = as.factor(DIST)
51   )
52
53 # Gráfico: Tasa de reclamos vs. Tipo de Carro (CAR)
54 plot_car <- ggplot(eda_df, aes(x = CAR, y = rate, fill = CAR)) +
55   geom_boxplot() +
56   labs(
57     title = "Tasa de Reclamos por Tipo de Carro",
58     x = "Tipo de Carro (CAR)",
59     y = "Tasa de Reclamos (y/n)"
60   ) +
61   theme_bw()
62 ggsave(here("results", "figures", "rate_vs_car.png"), plot_car)
63 print(plot_car)
64
65 # Gráfico: Tasa de reclamos vs. Grupo de Edad (EDAD)
66 plot_edad <- ggplot(eda_df, aes(x = EDAD, y = rate, fill = EDAD)) +

```

```

67 geom_boxplot() +
68 labs(
69   title = "Tasa de Reclamos por Grupo de Edad",
70   x = "Grupo de Edad (EDAD)",
71   y = "Tasa de Reclamos (y/n)"
72 ) +
73 theme_bw()
74 ggsave(here("results", "figures", "rate_vs_edad.png"), plot_edad)
75 print(plot_edad)
76
77 # Gráfico: Tasa de reclamos vs. Distrito (DIST)
78 plot_dist <- ggplot(eda_df, aes(x = DIST, y = rate, fill = DIST)) +
79   geom_boxplot() +
80   labs(
81     title = "Tasa de Reclamos por Distrito",
82     x = "Distrito (DIST)",
83     y = "Tasa de Reclamos (y/n)"
84   ) +
85   theme_bw()
86 ggsave(here("results", "figures", "rate_vs_dist.png"), plot_dist)
87 print(plot_dist)
88
89 cat("\n Inciso a) completado. Gráficos guardados en 'results/figures/'\n")
90
91
92 # Inciso (b): Ajuste modelo logístico completo -----
93 model_b <- glm(
94   cbind(y, n - y) ~ CAR * EDAD * DIST,
95   data = eda_df,
96   family = binomial(link = "logit")
97 )
98
99 # Guardamos el resumen y el modelo
100 summary_b <- summary(model_b)
101 print("--- Resumen del Modelo Completo (Inciso b) ---")
102 print(summary_b)
103 capture.output(summary_b, file = here("results", "summary_model_b.txt"))
104 saveRDS(model_b, file = here("results", "models", "model_b.rds")) # << CORREGIDO
105
106 cat("\n Inciso b) Completado. Resultados guardados. \n")
107
108
109 # Inciso (c): Ajuste modelo simplificado -----
110 simplified_df <- insurance_df %>%
111   mutate(
112     CAR = as.numeric(CAR),
113     EDAD = as.numeric(EDAD),
114     DIST = as.factor(DIST)
115   )
116
117 model_c <- glm(
118   cbind(y, n - y) ~ CAR + EDAD + DIST,
119   data = simplified_df,
120   family = binomial(link = "logit")
121 )
122
123 # Guardamos el resumen y el modelo
124 summary_c <- summary(model_c)
125 print("--- Resumen del Modelo Simplificado (Inciso c) ---")
126 print(summary_c)
127 capture.output(summary_c, file = here("results", "summary_model_c.txt"))
128 saveRDS(model_c, file = here("results", "models", "model_c.rds")) # << CORREGIDO
129
130 cat("\n Inciso c) completado. Resultados guardados. \n")
131
132
133 # Comparacion de modelos -----
134 model_comparison <- anova(model_c, model_b, test = "Chisq")

```

```
135  
136 print("--- Comparación de Modelos (ANOVA)")  
137 print(model_comparison)  
138 capture.output(model_comparison, file = here("results", "model_comparison_anova.txt"))
```

Listing 8: Script: Reclamo polizas de seguros.

A.6 Problema 8

```

1  # =====
2  # Problema 8
3  # Script: Estimacion mediante mínima ji-cuadrada
4  # Autor: Diego Paniagua Molina
5  # Fecha: 2025-09-19
6  # =====
7
8
9  # Datos -----
10 n1 <- 100
11 y1 <- 22
12
13 n2 <- 150
14 y2 <- 52
15
16 n3 <- 200
17 y3 <- 77
18
19 # Función Objetivo: Chi-Cuadrada Q(p1, p2) -----
20 chi_sq_function <- function(params) {
21   pi1 <- params[1]
22   pi2 <- params[2]
23
24   # Evitamos divisiones por cero o valores fuera de dominio [0,1]
25   if (pi1 <= 0 || pi1 >= 1 || pi2 <= 0 || pi2 >= 1 || (pi1 + pi2) >= 1) {
26     return(Inf) # Retorna un valor infinito si los parámetros son inválidos
27   }
28
29   # Término 1 para la muestra 1
30   term1 <- (y1 - n1 * pi1)^2 / (n1 * pi1 * (1 - pi1))
31
32   # Término 2 para la muestra 2
33   term2 <- (y2 - n2 * pi2)^2 / (n2 * pi2 * (1 - pi2))
34
35   # Término 3 para la muestra 3, usando la restricción pi3 = 1 - pi1 - pi2
36   pi3 <- 1 - pi1 - pi2
37   term3 <- (y3 - n3 * pi3)^2 / (n3 * pi3 * (1 - pi3))
38
39   # Suma total
40   return(term1 + term2 + term3)
41 }
42
43 # Valores iniciales para el optimizador -----
44 start_params <- c(y1/n1, y2/n2)
45 cat("Valores iniciales (ingenuos):", "\n")
46 print(start_params)
47
48 # Ejecutar nlminb -----
49 optimizer_result <- nlminb(
50   start = start_params,
51   objective = chi_sq_function,
52   lower = c(1e-9, 1e-9), # Límite inferior cercano a 0
53   upper = c(1 - 1e-9, 1 - 1e-9) # Límite superior cercano a 1
54 )
55
56 # Mostrar los resultados -----
57 cat("\n--- Resultados de la Optimización ---\n")
58 print(optimizer_result)

```

Listing 9: Script: Estimacion mediante mínima ji-cuadrada.

A.7 Problema 9

```

1  # =====
2  # Problema 9
3  # Script: Supervivencia Titanic
4  # Autor: Diego Paniagua Molina
5  # Fecha: 2025-09-20
6  # =====
7
8
9  # Librerias -----
10 library(here)
11 library(dplyr)
12 library(titanic)
13
14 # Cargar dataset -----
15 titanic_df <- as.data.frame(Titanic)
16
17 # Verificamos la estructura de los datos
18 print("Estructura de los datos en formato largo:")
19 glimpse(titanic_df)
20
21
22 # Ajuste modelos log-lineales -----
23 model_saturated <- glm(Freq ~ Class * Sex * Age * Survived,
24                        family = poisson(link = "log"),
25                        data = titanic_df)
26
27 # Guardamos el resumen y el objeto del modelo
28 summary_saturated <- summary(model_saturated)
29 print("--- Resumen del Modelo Saturado ---")
30 print(summary_saturated)
31 capture.output(summary_saturated, file = here("results", "summary_saturated_model.txt"))
32 saveRDS(model_saturated, file = here("results", "models", "model_saturated.rds"))
33
34
35 # Seleccion mejor modelo (backward) -----
36 print("--- Análisis de Devianza (drop1) del Modelo Saturado ---")
37 backward_selection <- drop1(model_saturated, test = "Chisq")
38 print(backward_selection)
39 capture.output(backward_selection, file = here("results", "backward_selection_analysis.txt")
40                )
41
42 # Modelo 2: Modelo con interacciones de 3 vías
43 model_3way <- glm(Freq ~ (Class + Sex + Age + Survived)^3,
44                  family = poisson(link = "log"),
45                  data = titanic_df)
46
47 # Guardamos los resultados de este modelo
48 summary_3way <- summary(model_3way)
49 print("--- Resumen del Modelo con Interacciones de 3 Vías ---")
50 print(summary_3way)
51 capture.output(summary_3way, file = here("results", "summary_3way_model.txt"))
52 saveRDS(model_3way, file = here("results", "models", "model_3way.rds"))
53
54 # Comparacion modelos -----
55 model_comparison <- anova(model_3way, model_saturated, test = "Chisq")
56 print("--- Comparación ANOVA: Modelo 3-vías vs. Modelo Saturado ---")
57 print(model_comparison)
58 capture.output(model_comparison, file = here("results", "model_comparison_anova.txt"))

```

Listing 10: Script: Supervivencia Titanic.

A.8 Problema 10

```

1  # =====
2  # Problema 10
3  # Script: Placebo vs acido ascórbico y gripe
4  # Autor: Diego Paniagua Molina
5  # Fecha: 2025-09-19
6  # =====
7
8
9  # Datos -----
10 datos <- data.frame(
11   Tratamiento = factor(c("Placebo", "AcidoAscorbico")),
12   Gripe = c(31, 17),
13   NoGripe = c(109, 122)
14 )
15
16 # Ajustar el GLM -----
17 datos$Tratamiento <- relevel(datos$Tratamiento, ref = "Placebo")
18
19 modelo_logistico <- glm(cbind(Gripe, NoGripe) ~ Tratamiento,
20   data = datos,
21   family = binomial(link = "logit"))
22
23 # Resumen del modelo -----
24 resultados_modelo <- summary(modelo_logistico)
25
26 print("\nResultados del Modelo Logístico:")
27 print(resultados_modelo)
28
29 # Odds Ratio -----
30 coef_tratamiento <- coef(modelo_logistico)["TratamientoAcidoAscorbico"]
31 odds_ratio <- exp(coef_tratamiento)
32
33 print("\nCálculo del Odds Ratio:")
34 cat("Odds Ratio (Ácido Ascórbico vs. Placebo):", odds_ratio, "\n")
35 cat("Interpretación: Las 'chances' (odds) de contraer gripe en el grupo que tomó Ácido
36   Ascórbico son",
37   round(odds_ratio, 3), "veces las chances del grupo Placebo.\n")

```

Listing 11: Script: Placebo vs acido ascórbico y gripe.