



**Centro de Investigación en Matemáticas**  
Unidad Monterrey

---

## Cómputo Estadístico

### Tarea 2

---

Gustavo Hernández Angeles

20 de septiembre de 2025

## Índice

<b>1</b>	<b>Ejercicio 1: Desempeño del Primer Ministro</b>	<b>3</b>
1.1	Solución: . . . . .	3
<b>2</b>	<b>Ejercicio 2: Ronquidos y Enfermedad Cardíaca</b>	<b>6</b>
2.1	Solución: . . . . .	6
<b>3</b>	<b>Ejercicio 3: Cangrejos Cacerola</b>	<b>9</b>
3.1	Solución: . . . . .	9
<b>4</b>	<b>Ejercicio 4: Regresión Logística y Teorema de Bayes</b>	<b>11</b>
4.1	Solución: . . . . .	11
<b>5</b>	<b>Ejercicio 5: Curva ROC para Daño Coronario</b>	<b>12</b>
5.1	Solución: . . . . .	12
<b>6</b>	<b>Ejercicio 6: Conteos de Células T4</b>	<b>14</b>
6.1	Solución: . . . . .	14
<b>7</b>	<b>Ejercicio 7: Reclamos de Pólizas de Seguros</b>	<b>19</b>
7.1	Solución: . . . . .	19
<b>8</b>	<b>Ejercicio 8: Estimación por Mínima Ji-Cuadrada</b>	<b>23</b>
8.1	Solución: . . . . .	23
<b>9</b>	<b>Ejercicio 9: Modelo Log-Lineal para el Titanic</b>	<b>25</b>
9.1	Solución: . . . . .	25
<b>10</b>	<b>Ejercicio 10: Ácido Ascórbico y la Gripe Común</b>	<b>27</b>
10.1	Solución: . . . . .	27

### Ejercicio 1: Desempeño del Primer Ministro

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. En los renglones se observa que 944 ciudadanos aprobaban el desempeño del funcionario, mientras que las columnas muestran que, seis meses después, sólo 880 aprueban su desempeño:

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

a) Considere el modelo de regresión logística

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

Escriba la logverosimilitud correspondiente. Muestre explícitamente (i.e. maximizando la logverosimilitud), que el estimador máximo verosímil para  $\beta_1$  es el logaritmo de la tasa de momios de la tabla dada.

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:

#### Inciso a)

Sea  $\pi_x = P(Y = 1|x)$ . El modelo implica dos ecuaciones, una para cada valor de  $x$ :

- Si  $x = 0$  (desaprobaban en la primera encuesta):  $\log \frac{\pi_0}{1-\pi_0} = \beta_0$ .
- Si  $x = 1$  (aprobaban en la primera encuesta):  $\log \frac{\pi_1}{1-\pi_1} = \beta_0 + \beta_1$ .

Los datos de la tabla se pueden ver como el resultado de dos muestras binomiales independientes: una de  $n_1 = 944$  individuos que aprobaron inicialmente ( $x = 1$ ) y otra de  $n_0 = 656$  que desaprobaron ( $x = 0$ ). La variable de respuesta  $Y$  es si aprueban en la segunda encuesta.

Denotemos las celdas de la tabla como  $n_{xy}$ , donde  $x$  es el resultado de la primera encuesta y  $y$  el de la segunda.

- $n_{11} = 794$  (Aprueba  $\rightarrow$  Aprueba)
- $n_{10} = 150$  (Aprueba  $\rightarrow$  Desaprueba)
- $n_{01} = 86$  (Desaprueba  $\rightarrow$  Aprueba)
- $n_{00} = 570$  (Desaprueba  $\rightarrow$  Desaprueba)

La función de log-verosimilitud, agrupando por los valores de  $x$ :

$$\ell(\beta_0, \beta_1) = \ell(\pi_0, \pi_1) = [n_{01} \log(\pi_0) + n_{00} \log(1 - \pi_0)] + [n_{11} \log(\pi_1) + n_{10} \log(1 - \pi_1)]$$

Para maximizar  $\ell$ , podemos maximizar cada parte por separado. El estimador de máxima verosimilitud para  $\pi_0$  (la probabilidad de aprobar en la segunda encuesta, dado que se desaprobaba en la primera) es la proporción muestral:

$$\hat{\pi}_0 = \frac{n_{01}}{n_{01} + n_{00}} = \frac{86}{86 + 570} = \frac{86}{656}$$

El EMV para  $\pi_1$  (la probabilidad de aprobar en la segunda encuesta, dado que se aprobaba en la primera) es:

$$\hat{\pi}_1 = \frac{n_{11}}{n_{11} + n_{10}} = \frac{794}{794 + 150} = \frac{794}{944}$$

Usando la propiedad de invarianza de los EMV, podemos estimar  $\beta_0$  y  $\beta_1$  sustituyendo  $\hat{\pi}_0$  y  $\hat{\pi}_1$  en las ecuaciones del modelo:

$$\hat{\beta}_0 = \log \left( \frac{\hat{\pi}_0}{1 - \hat{\pi}_0} \right)$$

$$\hat{\beta}_0 + \hat{\beta}_1 = \log \left( \frac{\hat{\pi}_1}{1 - \hat{\pi}_1} \right)$$

Despejando  $\hat{\beta}_1$  de la segunda ecuación:

$$\hat{\beta}_1 = \log \left( \frac{\hat{\pi}_1}{1 - \hat{\pi}_1} \right) - \hat{\beta}_0 = \log \left( \frac{\hat{\pi}_1}{1 - \hat{\pi}_1} \right) - \log \left( \frac{\hat{\pi}_0}{1 - \hat{\pi}_0} \right)$$

$$\hat{\beta}_1 = \log \left( \frac{\hat{\pi}_1 / (1 - \hat{\pi}_1)}{\hat{\pi}_0 / (1 - \hat{\pi}_0)} \right)$$

Mostrando así, que el estimador de  $\hat{\beta}_1$  es el logaritmo de la tasa de momios de la tabla dada.

### Inciso b)

La hipótesis a probar es  $H_0 : p_1 = p_2$ , donde  $p_1$  y  $p_2$  son las proporciones poblacionales de aprobación en la primera y segunda encuesta, respectivamente. Los datos no provienen de muestras independientes, sino que son mediciones repetidas sobre los mismos 1600 individuos. Por lo tanto, se trata de datos pareados. La prueba adecuada para comparar proporciones en muestras pareadas es la *prueba de McNemar*.

Esta prueba se centra en los individuos que cambiaron de opinión entre las dos encuestas, es decir, las celdas discordantes de la tabla:

- $n_{10} = 150$ : Personas que aprobaron en la primera encuesta pero desaprobaron en la segunda.
- $n_{01} = 86$ : Personas que desaprobaron en la primera encuesta pero aprobaron en la segunda.

La hipótesis nula de igualdad de proporciones marginales ( $H_0 : p_1 = p_2$ ) es equivalente a la hipótesis de que la probabilidad de cambiar de opinión en una dirección es igual a la probabilidad de cambiar en la dirección opuesta.

El estadístico de prueba de McNemar se calcula como:

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

Bajo la hipótesis nula, este estadístico sigue una distribución Chi-cuadrada ( $\chi^2$ ) con 1 grado de libertad. Calculamos el valor del estadístico  $\chi^2$  con los datos de la tabla.

$$\chi^2 = \frac{(150 - 86)^2}{150 + 86} = \frac{64^2}{236} = \frac{4096}{236} \approx 17.36$$

Comparando el valor del estadístico con el valor crítico de una distribución  $\chi^2$  con 1 grado de libertad para un nivel de significancia, digamos  $\alpha = 0.05$ , el valor crítico es 3.841. En este caso,  $17.36 > 3.841$ , lo que indica que se rechaza la hipótesis nula  $H_0$  y concluimos que hay una diferencia estadísticamente significativa entre la proporción de aprobación en la primera encuesta y la proporción de aprobación en la segunda.

## Ejercicio 2: Ronquidos y Enfermedad Cardíaca

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

Ajuste un modelo lineal generalizado logit y probit para analizar si existe una relación entre los ronquidos y la posibilidad de tener una enfermedad cardíaca.

### Solución:

El objetivo es modelar la probabilidad de tener una enfermedad cardíaca en función de una puntuación que cuantifica la frecuencia de los ronquidos. Dado que la respuesta es binaria (SI/NO) y los datos están agrupados, se utiliza un modelo lineal generalizado para datos binomiales.

Primero, preparamos los datos:

```
# Datos del problema
ronquido_nivel <- c("Nunca", "Ocasional", "Casi cada noche", "Cada noche")
no_enfermedad <- c(1355, 603, 192, 224)
si_enfermedad <- c(24, 35, 21, 30)
ronquido_score <- c(0, 2, 4, 5)

# Crear matriz de respuesta (éxitos, fracasos)
y <- cbind(si_enfermedad, no_enfermedad)
```

### Modelo Logit

El modelo logit utiliza la función de enlace logit, que modela el logaritmo de los momios de tener la enfermedad. La variable de respuesta se especifica como una matriz de éxitos (casos 'SI') y fracasos (casos 'NO'), y la variable predictora es la puntuación de ronquidos.

```
# Ajustar modelo logit
modelo_logit <- glm(y ~ ronquido_score, family = binomial(link = "logit"))
summary(modelo_logit)

##
## Call:
```

```
## glm(formula = y ~ ronquido_score, family = binomial(link = "logit"))
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.86625    0.16621 -23.261  < 2e-16 ***
## ronquido_score  0.39734    0.05001   7.945 1.94e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 65.9045  on 3  degrees of freedom
## Residual deviance:  2.8089  on 2  degrees of freedom
## AIC: 27.061
##
## Number of Fisher Scoring iterations: 4
```

El coeficiente estimado para la puntuación de ronquidos indica que por cada aumento de una unidad en la escala de ronquidos, los momios de padecer una enfermedad cardíaca aumentan significativamente ( $\exp(0.397) \approx 1.487$  entonces un aumento de 48.7%). El p-valor extremadamente pequeño indica que existe una relación estadísticamente muy significativa entre la frecuencia de los ronquidos y la probabilidad de tener una enfermedad cardíaca.

## Modelo Probit

El modelo probit es una alternativa al modelo logit que, en lugar de basarse en la distribución logística, se fundamenta en la distribución normal estándar. Su función de enlace es la distribución acumulada normal ( $\Phi(p)$ ), lo que en términos más sencillos equivale a modelar el Z-score asociado a la probabilidad de que ocurra el evento. Es decir

$$P(Y = 1|X) = \Phi(X^T\beta)$$

```
# Ajustar modelo probit
modelo_probit <- glm(y ~ ronquido_score, family = binomial(link = "probit"))
summary(modelo_probit)

##
## Call:
## glm(formula = y ~ ronquido_score, family = binomial(link = "probit"))
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.06055    0.07017 -29.367  < 2e-16 ***
## ronquido_score  0.18777    0.02348   7.997 1.28e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##  
## (Dispersion parameter for binomial family taken to be 1)  
##  
##      Null deviance: 65.9045  on 3  degrees of freedom  
## Residual deviance:  1.8716  on 2  degrees of freedom  
## AIC: 26.124  
##  
## Number of Fisher Scoring iterations: 4
```

El coeficiente estimado para la puntuación de ronquidos indica que por cada aumento de una unidad en la escala de ronquidos, el "índice probit" (el valor Z) aumenta significativamente. Al igual que en el modelo logit, el p-valor muy bajo confirma una fuerte evidencia de la relación entre roncar y la enfermedad cardíaca.

## Comparación de modelos

```
# Comparar AIC de ambos modelos  
cat("AIC Modelo Logit:", AIC(modelo_logit), "\n")  
  
## AIC Modelo Logit: 27.06147  
  
cat("AIC Modelo Probit:", AIC(modelo_probit), "\n")  
  
## AIC Modelo Probit: 26.12412
```

El modelo Probit tiene un AIC ligeramente menor (diferencia de  $\sim 0.94$ ), lo que sugiere que ajusta marginalmente mejor a los datos. Sin embargo, esta diferencia es pequeña y ambos modelos son prácticamente equivalentes en términos de ajuste.

## Conclusión

Ambos modelos, logit y probit, coinciden; existe una fuerte evidencia estadística para afirmar que hay una relación positiva y significativa entre la frecuencia de los ronquidos y la probabilidad de tener una enfermedad cardíaca. A medida que aumenta la puntuación en la escala de ronquidos, también aumenta la probabilidad de padecer dicha enfermedad.

### Ejercicio 3: Cangrejos Cacerola

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
1	2	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:

El objetivo es modelar el número de machos concubinos (satélites) de un cangrejo cacerola hembra en función de sus características físicas, utilizando únicamente la muestra de 7 observaciones proporcionada. Dado que la variable respuesta es un conteo, se utiliza un modelo lineal generalizado de tipo **Poisson**.

Primero, creamos el conjunto de datos en R. Es crucial convertir las variables **Color** y **Spine** a factores para que el modelo las trate como variables categóricas. Además, el peso se convierte de gramos a kilogramos para obtener coeficientes más interpretables.

```
# Crear el data.frame a partir de la muestra de 7 observaciones
crabs <- data.frame(
  Width = c(28.3, 22.5, 26.0, 24.8, 26.0, 23.8, 26.5),
  Color = c(3, 4, 1, 4, 4, 3, 2),
  Spine = c(3, 3, 2, 3, 3, 3, 1),
  Satellite = c(8, 0, 9, 0, 4, 0, 0),
  Weight = c(3050, 1550, 2300, 2100, 2600, 2100, 2350)
)
```

### Ajuste del Modelo Poisson

Ajustamos el modelo de regresión de Poisson. **Nota importante:** Con solo 7 observaciones y múltiples predictores (incluyendo los niveles de las variables categóricas), el modelo tiene muy poca información para generar estimaciones estables.

```
# Ajustar el modelo GLM de tipo Poisson con la muestra
modelo_poisson <- glm(Satellite ~ Color + Spine + Width + Weight,
```

```

                                data = crabs,
                                family = poisson)
summary(modelo_poisson)

##
## Call:
## glm(formula = Satellite ~ Color + Spine + Width + Weight, family = poisson,
##      data = crabs)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.068605  25.825272   0.235  0.81422
## Color       -1.326618   0.506622  -2.619  0.00883 **
## Spine        0.809956   1.186898   0.682  0.49498
## Width       -0.572408   1.504357  -0.381  0.70357
## Weight       0.004572   0.006767   0.676  0.49924
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 37.770  on 6  degrees of freedom
## Residual deviance: 10.152  on 2  degrees of freedom
## AIC: 31.41
##
## Number of Fisher Scoring iterations: 7

```

La interpretación del modelo debe tomarse con cuidado debido al tamaño de la muestra tan pequeña (solo 2 grados de libertad residuales).

Los coeficientes estimados indican cómo cambia el logaritmo del número esperado de concubinos. Para una interpretación más intuitiva, lo convertimos a su forma exponencial.

En este ejercicio, solo el coeficiente de **Color** es estadísticamente significativo ( $p\text{-value} = 0.008 < 0.05$ ). Esto sugiere que el color del cangrejo tiene un impacto relevante en el número esperado de concubinos. Mientras que las otras variables no muestran evidencia estadística suficiente para afirmar que influyen en el número de concubinos.

- **Color:** El valor de este coeficiente es -1.32, lo que sugiere que, manteniendo las demás variables constantes, un aumento en el nivel de color está asociado con una disminución en el número esperado de concubinos. Específicamente, cada unidad de aumento en el nivel de color reduce el número esperado de concubinos en un factor de  $\exp(-1.32) \approx 0.267$ , es decir, una reducción del 73.3 %.

También debemos tomar en cuenta el valor de la devianza residual (10.152) siendo mayor a sus 2 grados de libertad. Esto es un fuerte indicio de sobredispersión, lo que significa que la variabilidad de los datos es mayor a la que el modelo de Poisson asume. Consecuentemente, los errores estándar y los valores  $p$  pueden no ser del todo confiables.

#### Ejercicio 4: Regresión Logística y Teorema de Bayes

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

- $Y_j \sim \text{Bernoulli}(p), i = 1, \dots, n$
- $X_i | \{Y_i = 1\} \sim N(\mu_1, \sigma^2)$
- $X_i | \{Y_i = 0\} \sim N(\mu_0, \sigma^2)$

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

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

#### Solución:

Podemos aplicar directamente el Teorema de Bayes sobre la probabilidad  $P(Y_i = 1 | X_i)$ , tomando en cuenta que al ser  $X$  una variable continua, su “probabilidad” se calcula con su función de densidad de probabilidad  $f_X(X_i | Y_i = y)$ .

$$P(Y_i = 1 | X_i) = \frac{f_X(X_i | Y_i = 1)P(Y_i = 1)}{f_X(X_i | Y_i = 1)P(Y_i = 1) + f_X(X_i | Y_i = 0)P(Y_i = 0)}$$

Determinamos el momio de esta probabilidad para acercarnos a la forma del logit.

$$\begin{aligned} \frac{P(Y_i = 1 | X_i)}{1 - P(Y_i = 1 | X_i)} &= \frac{\frac{f_X(X_i | Y_i = 1)P(Y_i = 1)}{f_X(X_i | Y_i = 1)P(Y_i = 1) + f_X(X_i | Y_i = 0)P(Y_i = 0)}}{\frac{f_X(X_i | Y_i = 0)P(Y_i = 0)}{f_X(X_i | Y_i = 1)P(Y_i = 1) + f_X(X_i | Y_i = 0)P(Y_i = 0)}} \\ &= \frac{f_X(X_i | Y_i = 1)P(Y_i = 1)}{f_X(X_i | Y_i = 0)P(Y_i = 0)} \\ &= \frac{\exp(-(X_i - \mu_1)^2/2\sigma^2)p}{\exp(-(X_i - \mu_0)^2/2\sigma^2)(1-p)} \\ &= \frac{p}{1-p} \exp\left[\frac{-(X_i - \mu_1)^2 + (X_i - \mu_0)^2}{2\sigma^2}\right] \\ &= \frac{p}{1-p} \exp\left[\frac{((\mu_1 - \mu_2)X_i + \mu_0^2 - \mu_1^2)/2\sigma^2}{1}\right] \end{aligned}$$

Ahora aplicamos logaritmo natural para obtener el logit de la probabilidad deseada  $\text{logit}(P(Y_i = 1 | X_i))$ .

$$\begin{aligned} \text{logit}(P(Y_i = 1 | X_i)) &= \log\left(\frac{P(Y_i = 1 | X_i)}{1 - P(Y_i = 1 | X_i)}\right) \\ &= \log\left(\frac{p}{1-p}\right) + \frac{\mu_0^2 - \mu_1^2}{2\sigma^2} + \frac{\mu_1 - \mu_2}{2\sigma^2} X_i \end{aligned}$$

Haciendo  $\alpha = \log\left(\frac{p}{1-p}\right) + \frac{\mu_0^2 - \mu_1^2}{2\sigma^2}$  y  $\beta = \frac{\mu_1 - \mu_2}{2\sigma^2}$ , finalmente obtenemos:

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

### Ejercicio 5: Curva ROC para Daño Coronario

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

#### Solución:

Primero, cargamos los datos de edad y daño coronario. Las curvas ROC grafican las tasas *TPR* vs *FPR* para diferentes umbrales de clasificación de nuestro modelo logístico. Es por ello que esta vez nos interesan las probabilidades de tener daño coronario en función de la edad.os proporcionados

```
edad <- c(
  20, 23, 24, 25, 25, 26, 26, 28, 28, 29, 30, 30, 30, 30, 30, 30,
  32, 32, 33, 33, 34, 34, 34, 34, 34, 35, 35, 36, 36, 36, 37, 37,
  37, 38, 38, 39, 39, 40, 40, 41, 41, 42, 42, 42, 42, 43, 43, 43,
  44, 44, 44, 44, 45, 45, 46, 46, 47, 47, 47, 48, 48, 48, 49, 49,
  49, 50, 50, 51, 52, 52, 53, 53, 54, 55, 55, 55, 56, 56, 56, 57,
  57, 57, 57, 57, 57, 58, 58, 58, 59, 59, 60, 60, 61, 62, 62, 63,
  64, 64, 65, 69)

coro <- c(
  0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,1,0,
  0,1,0,0,0,0,1,0,1,0,0,0,0,0,1,0,0,1,0,0,1,1,0,1,0,1,0,0,1,0,
  1,1,0,0,1,0,1,0,0,1,1,1,1,0,1,1,1,1,1,0,0,1,1,1,1,0,1,1,1,1,
  0,1,1,1,1,1,0,1,1,1)

datos <- data.frame(edad = edad, coro = coro)
modelo_logistico <- glm(coro ~ edad, data = datos, family = binomial)
probabilidades <- predict(modelo_logistico, type = "response")
```

Ahora, usamos la librería **pROC** para construir y graficar la curva ROC. El área bajo la curva (AUC) es una medida de qué tan bien el modelo discrimina entre las clases. Esta librería facilita el cálculo y la visualización de la curva ROC. La función **roc()** toma como entrada las respuestas reales y las probabilidades predichas por el modelo, y devuelve un objeto que contiene toda la información necesaria para graficar la curva ROC y calcular el AUC. En la figura 5.1 se muestra la curva ROC obtenida.

```
library(pROC)
roc_curva <- roc(response = datos$coro, predictor = probabilidades)#f
```

```
plot(roc_curva,
     lwd = 2,           # Grosor de la línea
     print.auc = TRUE, # Imprimir el valor del AUC en el gráfico
     auc.polygon = TRUE, # Rellenar el área bajo la curva
     auc.polygon.col = "#dceafc",
```

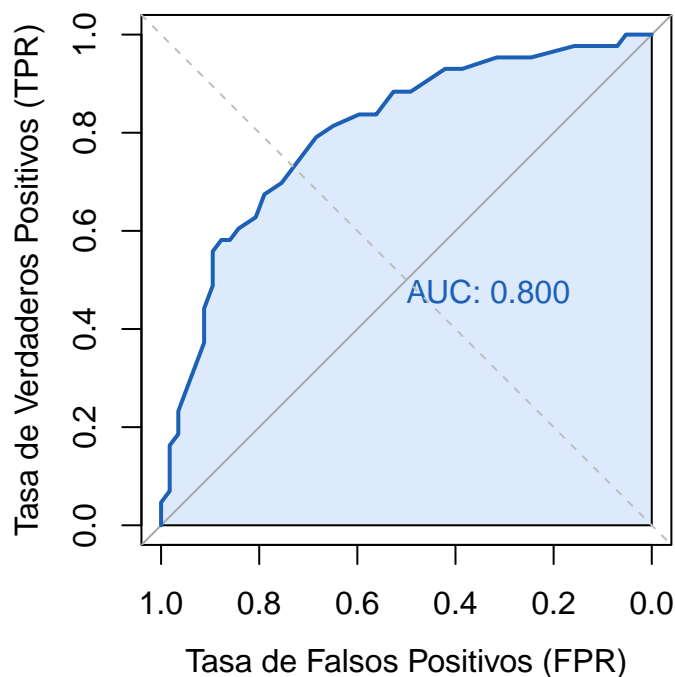


Figura 5.1: Curva ROC del modelo logístico para Daño Coronario por Edad

```
col="#1c61b6",  
xlab="Tasa de Falsos Positivos (FPR)",  
ylab="Tasa de Verdaderos Positivos (TPR)",  
)  
abline(a=0, b=1, lty=2, col="gray")
```

El valor del AUC es aproximadamente 0.8, lo que indica que el modelo tiene una buena capacidad para discriminar entre pacientes con y sin daño coronario basado en la edad. Un AUC de 0.5 indicaría que el modelo no tiene capacidad discriminativa (equivalente a adivinar al azar), mientras que un AUC de 1.0 indicaría una discriminación perfecta.

### Ejercicio 6: Conteos de Células T4

La siguiente tabla muestra conteos de células  $T_4$  por  $mm^3$  en muestras de sangre de 20 pacientes con enfermedad de Hodgkin y 20 pacientes en remisión de otras enfermedades. Se busca determinar 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	795	431	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:

#### Inciso a)

Primero leemos los datos. Estableceremos la variable binaria *has.h* para especificar los pacientes con enfermedad de Hodgkin.

```
library(ggplot2)

h_counts <- c(396, 568, 1212, 171, 554, 1104, 257, 435, 295, 397,
             288, 1004, 431, 795, 1621, 1378, 902, 958, 1283, 2415)
noh_counts <- c(375, 375, 752, 208, 151, 116, 736, 192, 315, 1252,
               675, 700, 440, 771, 688, 426, 410, 979, 377, 503)

df <- data.frame(
  group = factor(c(rep("H", length(h_counts)), rep("No-H", length(noh_counts)))),
  counts = c(h_counts, noh_counts)
)
```

Podemos comparar las distribuciones del conteo de células de los pacientes con y sin la enfermedad de Hodgkin mediante Kernel Density Estimation (KDE) (Figura 6.1) y Boxplot (Figura 6.2).

```
ggplot(data = df) +
  geom_density(aes(x = counts, fill = group), alpha = 0.5) +
  labs(x = "Número de células", y = "Densidad", fill = "Grupos") +
  theme(legend.position="inside", legend.position.inside = c(0.9, 0.9))
```

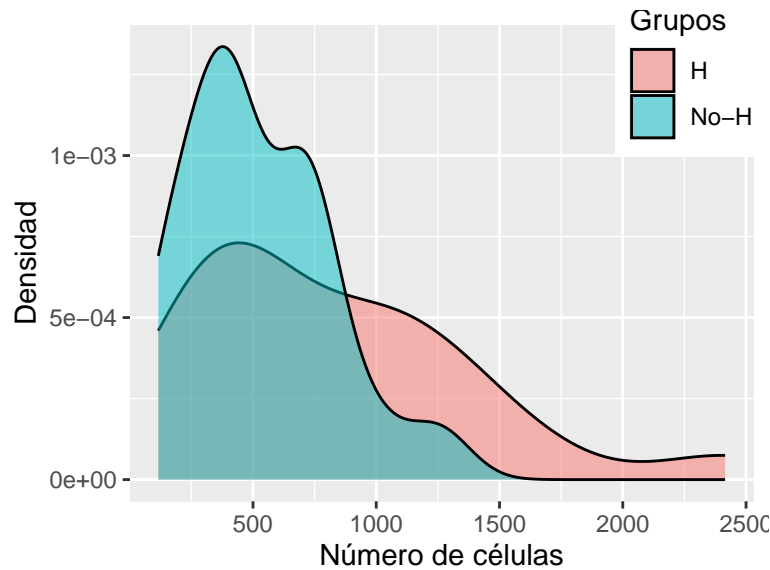


Figura 6.1: Gráfico de densidad de la distribución del conteo de células  $T_4$  por  $mm^3$  para pacientes con la enfermedad de Hodgkin (color rojo) y sin ella (color azul)

```
boxplot(counts ~ group, data=df, ylab="Número de células", xlab="Grupos")
```

El gráfico de densidad muestra una distribución más sesgada "hacia la derecha" para los pacientes con la enfermedad de Hodgkin que se traduce en una mayor varianza y una media muestral mayor respecto al grupo de pacientes sin la enfermedad.

El gráfico de caja muestra los cuartiles y la mediana de las distribuciones. La comparación visual sugiere también una mayor mediana en el conteo de las células  $T_4$ .

### Inciso b)

En este inciso codificamos la variable binaria para que posea valores 0 y 1, en donde 1 representa a los pacientes con la enfermedad de Hodgkin.

```
df <- data.frame(
  has_h = factor(c(rep(1, length(h_counts)), rep(0, length(noh_counts)))),
  counts = c(h_counts, noh_counts)
)
poisson_model <- glm(counts ~ has_h, data = df, family = poisson)
summary(poisson_model)

##
## Call:
## glm(formula = counts ~ has_h, family = poisson, data = df)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
```

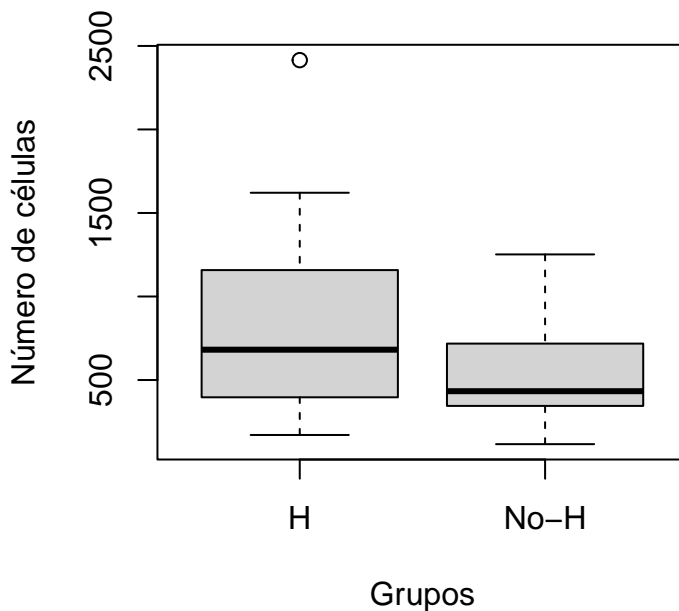


Figura 6.2: Gráfico de caja de la distribución del conteo de células  $T_4$  por  $mm^3$  para pacientes con la enfermedad de Hodgkin y sin ella.

```
## (Intercept) 6.257763 0.009787 639.4 <2e-16 ***
## has_h1      0.455436 0.012511 36.4 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
## Null deviance: 11325  on 39  degrees of freedom
## Residual deviance: 9965  on 38  degrees of freedom
## AIC: 10294
##
## Number of Fisher Scoring iterations: 5
```

El modelo de regresión de Poisson ajustado se especifica de la siguiente manera:

$$\ln(\mathbb{E}[\text{conteo}]) = \beta_0 + \beta_1 \cdot \text{has\_h}$$

Los resultados del modelo, proporcionados en el `summary` de R, nos dan las estimaciones de los coeficientes:

- **Coefficiente del Intercepto** ( $\beta_0$ ): 6.257763
- **Coefficiente para has\_h1** ( $\beta_1$ ): 0.455436

Estos coeficientes tienen la siguiente interpretación:

- **El Intercepto:** El valor de 6.257763 es el logaritmo del conteo promedio de células  $T_4$  para el grupo de pacientes sin la enfermedad de Hodgkin (cuando `has_h` = 0). Para obtener el conteo promedio en la escala original, se calcula  $e^{6.257763} \approx 522.2$ .
- **El Coeficiente `has_h1`:** El valor de 0.455436 representa el **cambio** en el logaritmo del conteo promedio de células cuando se pasa del grupo sin la enfermedad de Hodgkin (`has_h` = 0) al grupo con la enfermedad (`has_h` = 1). Este valor indica un incremento en el logaritmo de la media.

Para una interpretación más intuitiva, se puede calcular la *Razón de Tasa de Incidencia (RTI)*, la cual se obtiene exponentiando el coeficiente:  $e^{0.455436} \approx 1.577$ . Esto significa que el conteo promedio de células  $T_4$  en los pacientes con la enfermedad de Hodgkin es aproximadamente 1.577 veces mayor que en los pacientes sin la enfermedad.

**Dispersión del Modelo:** Es crucial notar que el modelo presenta sobredispersión. Esto se evidencia al comparar la devianza residual (9965) con los grados de libertad residuales (38). El valor de la devianza es mucho mayor que los grados de libertad, lo que indica que la varianza de los datos es significativamente mayor que la media. Aunque el modelo de Poisson muestra una diferencia significativa, los errores estándar pueden estar subestimados, lo que infla el valor del estadístico  $z$  y, por lo tanto, reduce el  $p$ -valor de manera artificial.

#### Inciso c)

Para completar el inciso, se utiliza la propiedad de normalidad asintótica de los estimadores de máxima verosimilitud para construir un intervalo de confianza del 90 % para la diferencia en las medias. En el modelo de Poisson, esta diferencia se modela a través del coeficiente  $\beta_1$  que representa el logaritmo de la razón de las medias.

Del resumen de los resultados del modelo de regresión de Poisson, se extraen los siguientes valores para el coeficiente  $\beta_1$  (`has_h1`):

- Estimación ( $\hat{\beta}_1$ ): 0.455436
- Error Estándar ( $SE(\hat{\beta}_1)$ ): 0.012511

El valor crítico para un intervalo de confianza del 90 % en una distribución normal estándar es  $Z_{0.95} = 1.645$ . El intervalo de confianza se calcula usando la fórmula de normalidad asintótica:

$$IC_{90\%}(\beta_1) = \hat{\beta}_1 \pm Z_{0.95} \cdot SE(\hat{\beta}_1)$$

Sustituyendo los valores:

$$\begin{aligned} IC_{90\%}(\beta_1) &= 0.455436 \pm 1.645 \cdot (0.012511) \\ &= 0.455436 \pm 0.02058 \\ &= [0.434856, 0.476016] \end{aligned}$$

Este intervalo corresponde al logaritmo natural de la razón de las medias,  $\ln(\lambda_H/\lambda_{No-H})$ . Para obtener el intervalo en la escala original, se aplica la función exponencial a los límites del intervalo anterior:

$$IC_{90\%}(\lambda_H/\lambda_{No-H}) = [e^{0.434856}, e^{0.476016}] = [1.545, 1.609]$$

El intervalo de confianza del 90 % para la razón de las medias de los conteos de células  $T_4$  es  $[1.545, 1.609]$ . Dado que este intervalo no contiene el valor 1, se concluye que existe evidencia estadística de que las medias de los conteos entre los dos grupos son significativamente

diferentes. Específicamente, se estima con un 90 % de confianza que la media de conteo para los pacientes con la enfermedad de Hodgkin es entre 1.545 y 1.609 veces mayor que para los pacientes sin la enfermedad.

### Ejercicio 7: Reclamos de Pólizas de Seguros

Los datos de la tabla son números,  $n$ , de pólizas de seguros y los correspondientes números,  $y$ , de reclamos. Las variables son CAR (clase de carro), EDAD (edad del titular) y DIST (distrito).

- 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) así como sus interacciones.
- Basados en resultados previos, se decidió que ninguna interacción era importante y que CAR y EDAD podían ser 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

### Solución:

#### Inciso a)

Primero, creamos el data frame con los datos proporcionados y calculamos la tasa de reclamos  $y/n$  para cada categoría. Luego, graficamos estas tasas contra las diferentes variables para visualizar los efectos principales.

```
seguros$tasa <- seguros$y / seguros$n # Tasa de reclamos
```

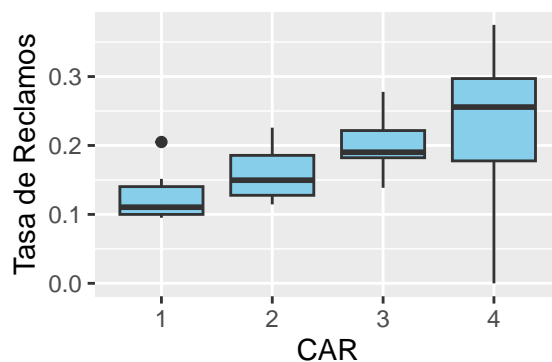


Figura 7.1: Tasa de reclamos vs Clase de Carro (CAR)

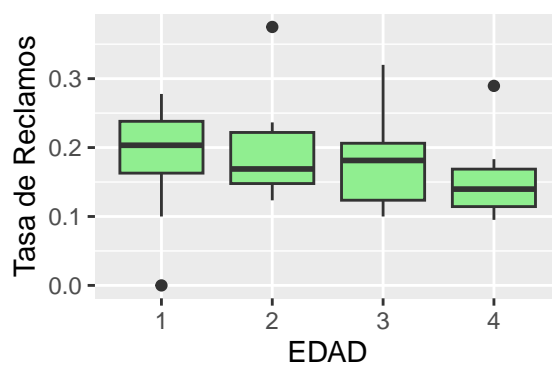


Figura 7.2: Tasa de reclamos vs Grupo de Edad (EDAD)

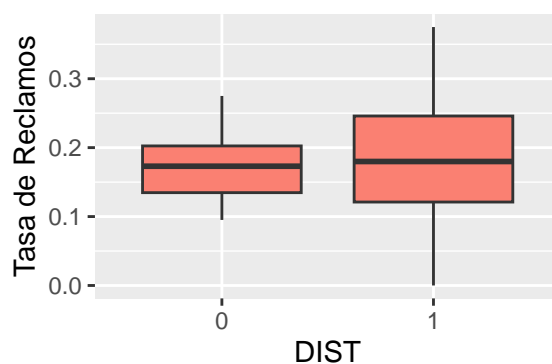


Figura 7.3: Tasa de reclamos vs Grupo de Distrito (DIST)

Las figuras 7.1, 7.2 y 7.3 muestran las tasas de reclamos en función de las variables CAR, EDAD y DIST, respectivamente. Se observa que la tasa de reclamos varía significativamente con la clase del carro (CAR) y la edad del titular (EDAD), mientras que la variación con el distrito (DIST) parece menos pronunciada. En particular, los carros de clase 4 y los titulares en el grupo de edad 1 muestran tasas de reclamos más altas, siguiendo una tendencia lineal en ambos casos.

```

modelo_inter_2do_orden <- glm(
  cbind(y, n - y) ~ (as.factor(CAR) + as.factor(EDAD) + as.factor(DIST))^2,
  data = seguros,
  family = binomial)

summary(modelo_inter_2do_orden)

##
## Call:
## glm(formula = cbind(y, n - y) ~ (as.factor(CAR) + as.factor(EDAD) +
##   as.factor(DIST))^2, family = binomial, data = seguros)
##
## Coefficients:
##
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.38483    0.13770  -10.057  < 2e-16 ***
## as.factor(CAR)2         0.02125    0.17520    0.121  0.903480
## as.factor(CAR)3        -0.05980    0.21380   -0.280  0.779711
## as.factor(CAR)4         0.28533    0.37617    0.759  0.448135
## as.factor(EDAD)2        -0.43724    0.18875   -2.317  0.020531 *
## as.factor(EDAD)3        -0.75621    0.19787   -3.822  0.000132 ***
## as.factor(EDAD)4        -0.86369    0.14909   -5.793  6.91e-09 ***
## as.factor(DIST)1        -0.15682    0.33540   -0.468  0.640092
## as.factor(CAR)2:as.factor(EDAD)2  0.12035    0.23275    0.517  0.605112
## as.factor(CAR)3:as.factor(EDAD)2  0.58884    0.26904    2.189  0.028623 *
## as.factor(CAR)4:as.factor(EDAD)2  0.40244    0.43647    0.922  0.356506
## as.factor(CAR)2:as.factor(EDAD)3  0.22430    0.23766    0.944  0.345265
## as.factor(CAR)3:as.factor(EDAD)3  0.78262    0.27158    2.882  0.003956 **
## as.factor(CAR)4:as.factor(EDAD)3  0.35526    0.43381    0.819  0.412829
## as.factor(CAR)2:as.factor(EDAD)4  0.17951    0.18712    0.959  0.337404
## as.factor(CAR)3:as.factor(EDAD)4  0.48243    0.22577    2.137  0.032615 *
## as.factor(CAR)4:as.factor(EDAD)4  0.33333    0.38819    0.859  0.390522
## as.factor(CAR)2:as.factor(DIST)1  0.09790    0.18935    0.517  0.605148
## as.factor(CAR)3:as.factor(DIST)1  0.16050    0.20506    0.783  0.433805
## as.factor(CAR)4:as.factor(DIST)1  0.57313    0.24923    2.300  0.021469 *
## as.factor(EDAD)2:as.factor(DIST)1 -0.07595    0.37813   -0.201  0.840811
## as.factor(EDAD)3:as.factor(DIST)1  0.32584    0.35287    0.923  0.355799
## as.factor(EDAD)4:as.factor(DIST)1  0.31717    0.31831    0.996  0.319048
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 244.3272  on 31  degrees of freedom
## Residual deviance:   6.4246  on   9  degrees of freedom
## AIC: 214.35
##
## Number of Fisher Scoring iterations: 4

```

En el modelo ajustado se utilizan todas las variables como categóricas, incluyendo las interacciones de segundo orden entre ellas. Podemos observar que solo 4 de los más de 15 coeficientes de segundo orden son estadísticamente significativos. Esto sugiere que las interacciones entre las variables no aportan información relevante para explicar la variabilidad en la tasa de reclamos. Además, los coeficientes de las variables CAR y EDAD muestran una tendencia lineal, lo que indica que estas variables podrían ser tratadas como continuas en lugar de categóricas.

### Inciso c)

```
modelo_sin_interacciones <- glm(cbind(y, n - y) ~ CAR + EDAD + as.factor(DIST),
                                data = seguros,
                                family = binomial)
summary(modelo_sin_interacciones)

##
## Call:
## glm(formula = cbind(y, n - y) ~ CAR + EDAD + as.factor(DIST),
##      family = binomial, data = seguros)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.66749    0.08748 -19.061  < 2e-16 ***
## CAR             0.23168    0.02266  10.225  < 2e-16 ***
## EDAD          -0.20967    0.02040 -10.278  < 2e-16 ***
## as.factor(DIST)1 0.25891    0.06420   4.033  5.5e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 244.327  on 31  degrees of freedom
## Residual deviance:  30.086  on 28  degrees of freedom
## AIC: 200.01
##
## Number of Fisher Scoring iterations: 4
```

Los coeficientes del nuevo modelo son todos estadísticamente significativos, lo que indica que cada una de las variables tiene un efecto relevante en la probabilidad de reclamo. En particular, los coeficientes de CAR y vivir en el distrito 1 son positivos, lo que sugiere que a medida que aumenta la clase del carro y si el titular vive en el distrito 1, la probabilidad de reclamo también aumenta. Por otro lado, el coeficiente negativo de EDAD indica que a medida que aumenta la edad del titular, la probabilidad de reclamo disminuye.

También es importante destacar que la medida AIC del nuevo modelo (200) es menor en comparación con el modelo anterior (214), lo que sugiere que este modelo es más parsimonioso. Esto nos indica que, aunque el modelo anterior ajustaba mejor a los datos, el modelo reducido logra un buen desempeño con menos parámetros, lo que es preferible en términos de simplicidad y generalización.

### Ejercicio 8: Estimación por Mínima Ji-Cuadrada

El método de la mínima ji-cuadrada consiste en estimar  $\theta$  mediante el valor que minimice el estadístico de Pearson:

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

Considere una población muy grande con tres categorías A, B y C. Se toman 3 muestras de tamaños  $n_1, n_2, n_3$  y se registra:

- 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  $\pi_1, \pi_2$  y  $\pi_3$  usando el método de la mínima ji-cuadrada, suponiendo  $n_1 = 100, y_1 = 22, n_2 = 150, y_2 = 52, n_3 = 200, y_3 = 77$ . Es decir, minimice:

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

con la restricción  $\pi_3 = 1 - \pi_1 - \pi_2$ .

### Solución:

Como el problema sugiere, utilizaremos la utilidad `nlminb()` de R para minimizar la función objetivo con la restricción dada. Primero, definimos la función objetivo que representa el estadístico de Pearson a minimizar.

```
minima_chi_cuadrada <- function(params) {
  pi1 <- params[1]
  pi2 <- params[2]
  pi3 <- 1 - pi1 - pi2 # restricción

  # penalizamos prob invalidas
  if (pi1 <= 0 || pi1 >= 1 || pi2 <= 0 || pi2 >= 1 || pi3 <= 0 || pi3 >= 1) {
    return(Inf)
  }

  chi1 <- (y1 - n1 * pi1)^2 / (n1 * pi1) +
    ((n1-y1) - n1*(1-pi1))^2 / (n1 * (1 - pi1))
  chi2 <- (y2 - n2 * pi2)^2 / (n2 * pi2) +
    ((n2-y2) - n2*(1-pi2))^2 / (n2 * (1 - pi2))
  chi3 <- (y3 - n3 * pi3)^2 / (n3 * pi3) +
    ((n3-y3) - n3*(1-pi3))^2 / (n3 * (1 - pi3))

  return(chi1 + chi2 + chi3)
}
```

Para utilizar `nlminb()`, necesitamos proporcionar un punto inicial para las probabilidades  $\pi_1$  y  $\pi_2$ . Un punto inicial razonable podría ser  $(y_1/n_1, y_2/n_2)$ , que son las proporciones observadas en las muestras.

```
iniciales <- c(y1/n1, y2/n2)

resultado <- nlminb(
  start = iniciales,
  objective = minima_chi_cuadrada,
  lower = c(0.0001, 0.0001),
  upper = c(0.9999, 0.9999)
)
```

Donde finalmente obtuvimos los siguientes resultados:

```
pi1_hat <- resultado$par[1]
pi2_hat <- resultado$par[2]
pi3_hat <- 1 - pi1_hat - pi2_hat

cat("Valor de Chi-Cuadrada Min:", round(resultado$objective, 4), "\n") #ℓ

## Valor de Chi-Cuadrada Min: 0.5123

cat("Pi 1 estimado:", round(pi1_hat, 4), "\n")

## Pi 1 estimado: 0.2395

cat("Pi 2 estimado:", round(pi2_hat, 4), "\n")

## Pi 2 estimado: 0.3629

cat("Pi 3 estimado:", round(pi3_hat, 4), "\n")

## Pi 3 estimado: 0.3976
```

Por lo tanto, las estimaciones obtenidas mediante el método de la mínima ji-cuadrada son:

- $\hat{\pi}_1 \approx 0.24$
- $\hat{\pi}_2 \approx 0.36$
- $\hat{\pi}_3 \approx 0.40$

### Ejercicio 9: Modelo Log-Lineal para el Titanic

Se analizan los datos del hundimiento del Titanic. Las variables son Class (1, 2, 3, Tripulación), Sex (Male, Female), Age (Child, Adult), y Survived (No, Yes). En R, usar la librería `titanic` y los datos del objeto `Titanic`. Considerar un modelo log-lineal para analizar los siguientes efectos y sus interacciones:

- Class: Hay más pasajeros en algunas clases que en otras.
- Sex: Hay más pasajeros de un sexo que de otro.
- Age: Hay más pasajeros de un grupo de edad que de otro.
- Survived: Hay más pasajeros que sobrevivieron que los que no.
- Class  $\times$  Sex: Class y Sex no son independientes.
- Class  $\times$  Age: Class y Age no son independientes.
- Class  $\times$  Survived: Class y Survived no son independientes.
- Sex  $\times$  Age: Sex y Age no son independientes.
- Sex  $\times$  Survived: Sex y Survived no son independientes.
- Age  $\times$  Survived: Age y Survived no son independientes.
- Interacciones triples y cuádruples.

**Solución:**

Modelo	$\chi^2$	gl	p-valor
Class	5901.080	28	0.000
Sex	5513.260	30	0.000
Age	3647.270	30	0.000
Survived	6508.320	30	0.000
Class x Sex	3358.040	24	0.000
Class x Age	2164.370	24	0.000
Class x Survived	4354.000	24	0.000
Sex x Age	2140.470	28	0.000
Sex x Survived	3836.730	28	0.000
Age x Survived	2668.090	28	0.000
Class x Sex x Age	844.370	16	0.000
Class x Sex x Survived	1832.180	16	0.000
Class x Age x Survived	1303.150	16	0.000
Sex x Age x Survived	1266.570	24	0.000
Independencia	1637.450	25	0.000
Interacciones 2do Orden	109.650	13	0.000
Interacciones 3er Orden	0.000	3	1.000

Cuadro 9.1: Análisis de Bondad de Ajuste para Modelos Log-Lineales del Titanic.

En la tabla 9.1 se resumen los resultados de los modelos ajustados, incluyendo el estadístico de Chi-cuadrado de Pearson, los grados de libertad y el p-valor asociado para cada modelo.

Los resultados indican que todos los modelos, excepto el de interacciones de tercer orden, tienen p-valores nulos, lo que sugiere que estos modelos no se ajustan bien a los datos. Todos los modelos con interacciones de segundo orden para abajo son rechazados categóricamente (incluso los que consideran las interacciones de tercer orden aisladas). Ninguno de ellos, por sí solo, es capaz de explicar la compleja estructura de dependencias en los datos.

No obstante, el modelo con interacciones de tercer orden tiene un p-valor de 1, lo que indica un excelente ajuste a los datos, lo que sugiere que este modelo captura adecuadamente las relaciones entre las variables. Por lo tanto, se concluye que el modelo con interacciones de tercer orden es el más adecuado para describir los datos del Titanic entre los modelos considerados.

### Ejercicio 10: Ácido Ascórbico y la Gripe Común

Se realizó un análisis sobre el valor terapéutico del ácido ascórbico (vitamina C) en relación a su efecto sobre la gripe común. Se tiene una tabla  $2 \times 2$  con los recuentos 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:

Utilizaremos la regresión logística para modelar la probabilidad de contraer gripe en función del tratamiento recibido (Placebo o Ácido Ascórbico).

```
##
## Call:
## glm(formula = cbind(gripe, no_gripe) ~ tratamiento, family = binomial(link = "logit"),
##      data = datos)
##
## Coefficients:
##
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.2574     0.2035  -6.177 6.53e-10 ***
## tratamientoAcido Ascorbico  -0.7134     0.3293  -2.166  0.0303 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 4.8717e+00  on 1  degrees of freedom
## Residual deviance: 7.5495e-15  on 0  degrees of freedom
## AIC: 13.578
##
## Number of Fisher Scoring iterations: 3
```

El modelo resultante confirma que existe una asociación estadísticamente significativa entre el tratamiento y la incidencia de gripe, con un p-valor de 0.0303. Los coeficientes revelan información acerca del efecto del ácido ascórbico en comparación con el placebo.

- **Intercepto:** -1.2574. Este valor representa el logaritmo de las probabilidades de contraer gripe en el grupo de referencia, el cual es Placebo. La probabilidad correspondiente es:

$$\frac{e^{-1.2574}}{1 + e^{-1.2574}} \approx 0.222 \quad (\text{o } 22.2\%)$$

- **Tratamiento (Ácido Ascórbico):** -0.7134. Este coeficiente indica que el log-odds de contraer gripe en el grupo de ácido ascórbico es 0.7134 unidades **menor** que en el grupo de placebo. Para hallar la probabilidad de este grupo, primero calculamos su log-odds total:

- *Log-odds total:*  $-1.2574 + (-0.7134) = -1.9708$ .
- *Probabilidad:* La probabilidad de gripe para el grupo con ácido ascórbico es:

$$\frac{e^{-1.9708}}{1 + e^{-1.9708}} \approx 0.122 \quad (\text{o } 12.2\%)$$

Una forma más directa de interpretar el coeficiente es con el odds ratio:  $e^{-0.7134} \approx 0.49$ . Esto significa que las probabilidades de contraer gripe tomando ácido ascórbico son del 49 % de las probabilidades del grupo placebo (una reducción del 51 % en las probabilidades). Por lo tanto, hay evidencia estadística que sugiere que el ácido ascórbico tiene un efecto protector contra la gripe.