Centro de Investigación en Matemáticas (CIMAT)
Unidad Monterrey

Computo Estadístico

Tarea 1

Regresion Logistica y Regresion Poisson

Diego Paniagua Molina
diego.paniagua@cimat.mx

29 de Agosto 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:

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

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

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

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:

$$\begin{aligned} x_i &\in \{0, 1\} : \text{Aprueba (1) o no (0) en la primer encuesta.} \\ y_i &\in \{0, 1\} : \text{Aprueba (1) o no (0) en la segunda encuesta.} \end{aligned}$$

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:

$$P_{11} = 1 - P_{10} \tag{10}$$
$$P_{00} = 1 - P_{01} \tag{11}$$
$$P_{10} = 1 - P_{11} \tag{12}$$
$$P_{01} = 1 - P_{00} \tag{13}$$

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-versoimilitud en términos de ellas utilizando los conteos n_{xy} para cada caso, tal que:

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

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

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

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:

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

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:

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

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 estadisticamente 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:

$$\begin{aligned} n_{10} &= 150 : \text{ciudadanos que aprobaron y luego desaprobaron} \\ n_{10} &= 86 : \text{ciudadanos que desaprobaron y luego aprobaron} \end{aligned}$$

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

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

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 critico de la distribución Chi-cuadrada es $\chi^2_{\alpha=0.05} = 3.841$, entonces:

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

Por lo tanto, se rechaza la hipótesis nula a un nivel de significancia del 5% lo que indica que si existe un cambio estadisticamente 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

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

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

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

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

con

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

SOLUCIÓN

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

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{20}$$

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

$$P(Y_i = 1|X_i = x_i) = \frac{P(X_i = x_i|Y_i = 1)P(Y_i = 1)}{P(X = x_i)} \tag{21}$$

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 (29) pasa a tener la siguiente forma:

$$P(Y_i = 1|X_i = x_i) = \frac{f(x_i|Y_i = 1)P(Y_i = 1)}{f(x_i)} \tag{22}$$

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 $\{B_1, B_2, \ldots, B_n\}$ que forman una partición del espacio muestral, entonces para cualquier otro evento A en ese mismo espacio:

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i) \tag{23}$$

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 (22):

$$P(Y_i = 1|X_i = x_i) = \frac{f(x_i|Y_i = 1)P(Y_i = 1)}{f(x_i|Y_i = 0)P(Y_i = 0) + f(x_i|Y_i = 1)P(Y_i = 1)} \tag{24}$$

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

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

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:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu^2)}{2\sigma^2}\right) \tag{25}$$

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 (24) 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:

$$P(Y_i = 1|X_i = x_i) = \frac{1}{Z+1} \tag{26}$$

Recordemos que la función logit se define como:

$$\text{logit}(P) = \log\left(\frac{P}{1-P}\right) \tag{27}$$

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:

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

Problema 3

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

a. Haga una comparación gráfica exploratoria de estos datos.

b. Ajuste un modelo de Poisson apropiado.

c. 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 utilizo el lenguaje de programación `R` para resolver este problema, los códigos utilizados se encuentran al final del documento en el apéndice A.

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 realizo un boxplot:

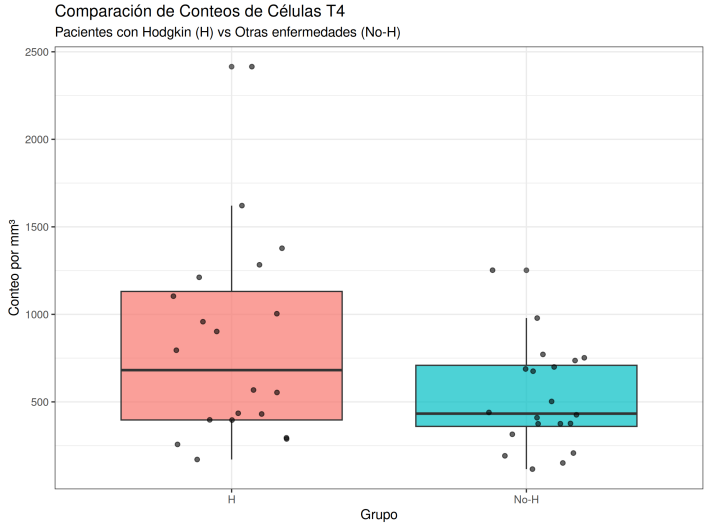

Figure 0.1: Boxplot de conteos de células T4 para el grupo H y No-H.

En la Figura 0.1 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 mas grande lo que implica una mayor dispersión de los datos, es decir, mas variabilidad entre el conteo de células T4 de los pacientes mientras que el grupo No-H tiene una caja mas compacta indicando lo opuesto. El grupo H tiene valores muy

altos que aparecen como outliers (≈ 2400), el grupo No-H tambien tiene algunos outliers altos pero no tan extremos (≈ 1250), esto podría indicarnos que la distribución del grupo H podria 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:

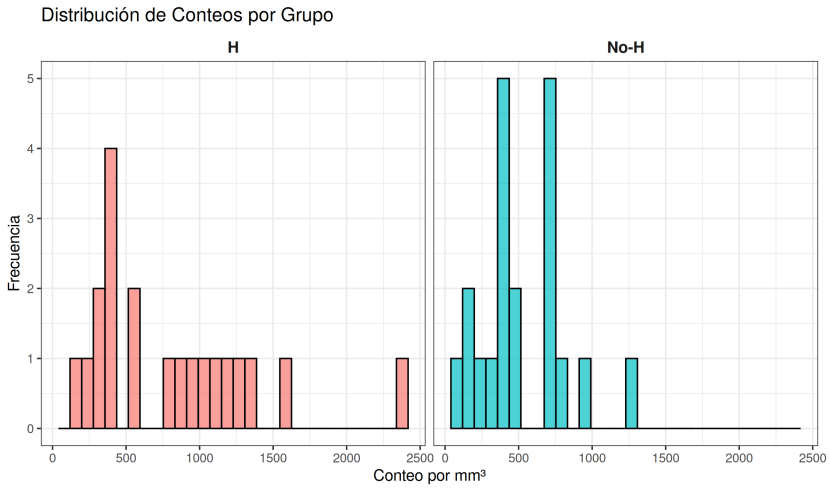

Figure 0.2: Histogramas de frecuencia de los conteos de células T4 para el grupo H y No-H.

En la Figura 0.2 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:

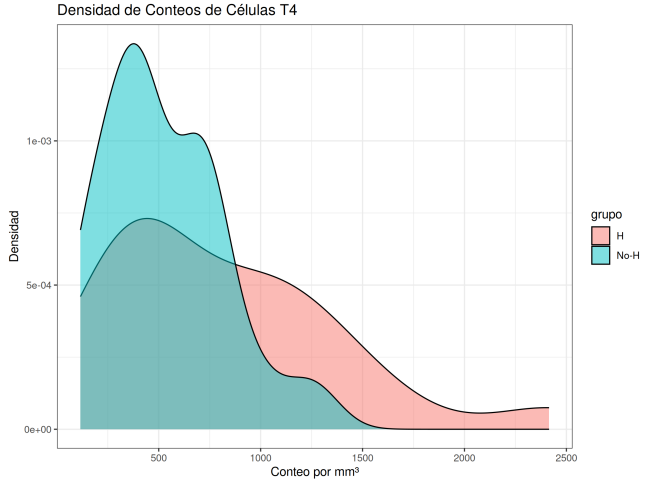

Figure 0.3: Gráfico de densidad de conteos de células T4 para el grupo H y No-H.

La Figura 0.3 muestra que los conteos de celulas 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:

$$\log(\mu_i) = \beta_0 + \beta_1 x_i \tag{28}$$

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:

$$\beta_0 = \log(\mu_{\text{H}}) \quad \text{y} \quad \beta_1 = \log\left(\frac{\mu_{\text{No-H}}}{\mu_{\text{H}}}\right) \tag{29}$$

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

Table 0.1: 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 $\widehat{\beta}_0 = 6.713$. Esto corresponde al grupo H. Por lo tanto, despejando de la ecuación (29), tenemos que la media esperada en este grupo es:

$$\widehat{\mu}_{\text{H}} = e^{6.713} \approx 824.8$$

Mientras que el efecto del grupo No-H es $\widehat{\beta}_1 = -0.455$. Despejando de la ecuación (29), esto implica que manteniendo todo lo demás constante, la razón de medias entre No-H y H es:

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

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 desviancia 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 $\widehat{\beta}$ 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 0.2: 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 $\widehat{\beta}_1 = -0.4554$ 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:

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

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.

A Código en R

A continuación, se presentan los scripts de R utilizados para la carga de datos, el análisis exploratorio y el ajuste de los modelos estadísticos discutidos en este reporte.

```r
# ==============================================================================
# Script 1: Carga y Limpieza de Datos - Conteos de Celulas T4
# Autor: Diego Paniagua Molina
# Fecha: 2025-08-29
# ==============================================================================

# Instalar paquetes necesarios ------------------------------------------------
if (!require("dplyr")) install.packages("dplyr")
if (!require("readr")) install.packages("readr")
if (!require("here")) install.packages("here")

# Cargar paquetes -------------------------------------------------------------
library(dplyr)
library(readr)
library(here)

# Crear los datos manualmente -------------------------------------------------
datos <- tibble(
  grupo = rep(c("H", "No-H"), each = 20),
  conteo = c(
    # Grupo H (Hodgkin) - Primera fila
    396, 568, 1212, 171, 554, 1104, 257, 435, 295, 397,
    # Grupo H (Hodgkin) - Segunda fila
    288, 1004, 431, 795, 1621, 1378, 902, 958, 1283, 2415,
    # Grupo No-H (No Hodgkin) - Primera fila
    375, 375, 752, 208, 151, 116, 736, 192, 315, 1252,
    # Grupo No-H (No Hodgkin) - Segunda fila
    675, 700, 440, 771, 688, 426, 410, 979, 377, 503
  )
)

# Verificar estructura de los datos -------------------------------------------
cat("Estructura de los datos:\n")
glimpse(datos)

cat("\nResumen por grupos:\n")
datos %>%
  group_by(grupo) %>%
  summarise(
    n = n(),
    media = mean(conteo),
    mediana = median(conteo),
    sd = sd(conteo),
    min = min(conteo),
    max = max(conteo)
  ) %>%
  print()

# Guardar datos en formato CSV ------------------------------------------------
dir.create(here("data"), recursive = TRUE, showWarnings = FALSE)
write_csv(datos, here("data", "raw", "datos_pacientes.csv"))

# Mensaje de confirmacion -----------------------------------------------------
cat("\nDatos creados y guardados en: data/raw/datos_pacientes.csv\n")
cat("Total de observaciones:", nrow(datos), "\n")
cat("Grupos:", unique(datos$grupo), "\n")
```

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

```r
# ==============================================================================
# Script 2: Analisis Exploratorio - Comparacion Grafica (Inciso a)
# ==============================================================================

# Instalar paquetes necesarios -------------------------------------------------
if (!require("dplyr")) install.packages("dplyr")
if (!require("ggplot2")) install.packages("ggplot2")
if (!require("here")) install.packages("here")

# Cargar paquetes --------------------------------------------------------------
library(dplyr)
library(ggplot2)
library(here)

# Cargar datos -----------------------------------------------------------------
datos <- read_csv(here("data", "raw", "datos_pacientes.csv"), show_col_types = FALSE)

# 1. Boxplot comparativo con puntos --------------------------------------------
boxplot <- ggplot(datos, aes(x = grupo, y = conteo, fill = grupo)) +
  geom_boxplot(alpha = 0.7) +
  geom_point(position = position_jitter(width = 0.2), alpha = 0.6) +
  labs(title = "Comparacion de Conteos de Celulas T4",
       subtitle = "Pacientes con Hodgkin (H) vs Otras enfermedades (No-H)",
       x = "Grupo",
       y = "Conteo por mm³") +
  theme_bw() +
  theme(legend.position = "none")

# 2. Histogramas comparativos --------------------------------------------------
histogramas <- ggplot(datos, aes(x = conteo, fill = grupo)) +
  geom_histogram(alpha = 0.7, position = "identity", color = "black") +
  facet_wrap(~ grupo, nrow = 1) +
  labs(title = "Distribucion de Conteos por Grupo",
       x = "Conteo por mm³",
       y = "Frecuencia") +
  theme_bw() +
  theme(aspect.ratio = 1,
        legend.position = "none",
        strip.background = element_blank(),
        strip.text = element_text(face = "bold", size = 11))

# 3. Grafico de densidad -------------------------------------------------------
densidad <- ggplot(datos, aes(x = conteo, fill = grupo)) +
  geom_density(alpha = 0.5) +
  labs(title = "Densidad de Conteos de Celulas T4",
       x = "Conteo por mm³",
       y = "Densidad") +
  theme_bw()

# Guardar graficos -------------------------------------------------------------
dir.create(here("results", "figures"), recursive = TRUE, showWarnings = FALSE)
ggsave(here("results", "figures", "boxplot_comparativo.png"),
       plot = boxplot, width = 8, height = 6, dpi = 300)
ggsave(here("results", "figures", "histogramas_comparativos.png"),
       plot = histogramas, width = 8, height = 6, dpi = 300)
ggsave(here("results", "figures", "densidad_comparativa.png"),
       plot = densidad, width = 8, height = 6, dpi = 300)

# Mostrar graficos -------------------------------------------------------------
print(boxplot)
print(histogramas)
print(densidad)

cat("\n Graficos guardados en la carpeta: results/figures/\n")
```

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

```r
# ============================================================================
# Script 3: Modelo de Poisson e Inferencia (Inciso b)
# ============================================================================

# Instalar paquetes necesarios ------------------------------------------------
if (!require("dplyr")) install.packages("dplyr")
if (!require("readr")) install.packages("readr")
if (!require("here")) install.packages("here")

# Cargar paquetes -------------------------------------------------------------
library(dplyr)
library(readr)
library(here)

# Cargar datos ----------------------------------------------------------------
datos <- read_csv(here("data", "raw", "datos_pacientes.csv"), show_col_types = FALSE)

# INCISO b: Ajustar modelo de Poisson -----------------------------------------
cat("Ajustando modelo de Poisson...\n")

# Modelo: log(conteo) = beta_0 + beta_1*grupo
modelo_poisson <- glm(conteo ~ grupo,
                      family = poisson(link = "log"),
                      data = datos)

# Resumen del modelo ----------------------------------------------------------
cat("\nResumen del modelo de Poisson:\n")
resumen_modelo <- summary(modelo_poisson)
print(resumen_modelo)

# Verificar sobredispersion ---------------------------------------------------
cat("\nVerificacion de sobredispersion:\n")
deviance_val <- resumen_modelo$deviance
df_residual <- resumen_modelo$df.residual
ratio_sobredispersion <- deviance_val / df_residual

cat("Deviance:", deviance_val, "\n")
cat("Grados de libertad:", df_residual, "\n")
cat("Ratio Deviance/df:", ratio_sobredispersion, "\n")

# Guardar modelo --------------------------------------------------------------
dir.create(here("results", "models"), recursive = TRUE, showWarnings = FALSE)
saveRDS(modelo_poisson, here("results", "models", "modelo_poisson.rds"))
cat("\n Modelo guardado en: results/models/modelo_poisson.rds\n")

# Existe sobredispersion, ajustar modelo quasi-Poisson ------------------------
cat("\nAjustando modelo quasi-Poisson...\n")

modelo_quasi <- glm(conteo ~ grupo,
                    family = quasipoisson(link = "log"),
                    data = datos)

# Resumen del modelo ----------------------------------------------------------
cat("\nResumen del modelo quasi-Poisson:\n")
resumen_quasi <- summary(modelo_quasi)
print(resumen_quasi)

# Dispersion (phi) ------------------------------------------------------------
phi <- resumen_quasi$dispersion
cat("\nEstimador de dispersion (phi):", round(phi, 3), "\n")

# Guardar modelo --------------------------------------------------------------
saveRDS(modelo_quasi, here("results", "models", "modelo_quasipoisson.rds"))
cat("\n Modelo guardado en: results/models/modelo_quasipoisson.rds\n")
```

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