Tarea 1B. Introducción a los modelos lineales generalizados

Equipo 1

2025-03-27

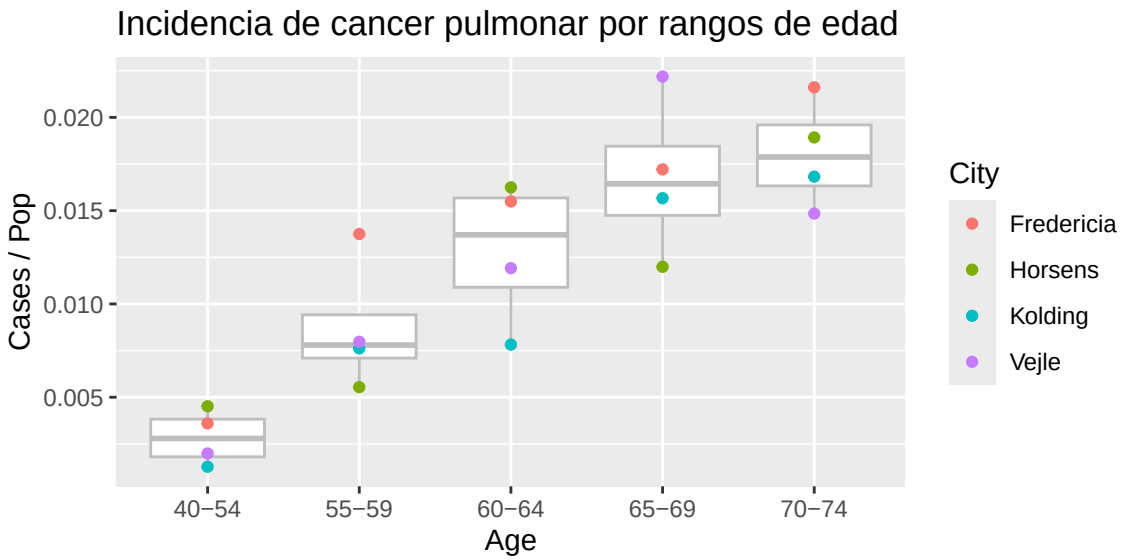
Análisis relacional de la incidencia de cáncer de pulmón respecto a la edad.

Introducción

El objetivo de este estudio es analizar la relación entre la edad y la incidencia de cáncer pulmonar. Para ese análisis se proporciona una base de daos obtenidso entre los años 1968 y 1971 en cuatro ciudades de Dinamarca (Fredericia, Horsens, Kolding y Vejle), en las cuales se registraron las incidencias de cáncer pulmonar en 5 rangos de edad (40-54, 55-59, 60-64, 65-69 y 70-74). Dado que se estudian ciudades distintas, se ha registrado también la población total de cada rango de edad presente en cada una de las ciudades registradas, lo que permite utilizar intensidades (incidencias relativas a las poblaciones totales) para poder comparar las poblaciones. Se espera poder argumentar que la relación entre la edad y la incidencia de cáncer es proporcional, es decir, a mayor edad, mayor incidencia de cáncer de pulmón.

Primer vistazo

Lo primero que hacemos es realizar un scatterplot para darnos una idea general de cómo se encuentran los datos y que relación tienen entre sí. Para ello, comparamos los casos de incidencia (variable `cases`), con la población relativa de cada rango de edad (variable `Age`), diferenciando la ciudad de donde provienen las observaciones (variable `City`). Esto nos dará una idea de cuál es la rlación que sepretende estudiar para cada ciudad.



A partir del gráfico anterior podemos notar una aparente tendencia creciente de la incidencia de cancer pulmonar respecto a la edad, de manera general, también se podría considerar que la varianza es creciente, aunque el rango `60-64` parece que rompe esta tendencia, pero no se puede concluir más allá a partir de la información que tenemos.

Selección del modelo

Dado que estamos trabajando con conteos (las observaciones son entreros no negativos) comenzamos con un modelo *Poisson* con todas las variables disponibles y las interacciones entre estas, empleando la función liga *log*, comunmente usada junto a esta distribución.

Este modelo emplea el uso de variables binarias para las catrgorías de las variables categóricas que tenemos (`Age` y `City`) exeptuando las primeras categorías de cada variable (consideradas valor de referencia) y para las interacciones tentdermos todas plas posibles combinaciones de productos entre estas variables, lo que nos da un total de 19 variables.

En total tenemos 19 variables por lo que el componente lineal constara de 20 paámentros β's. Luego, como no estamos considerando las observaciones como tal sino las intensidades y empleamos la función liga *log*, ya que usaremos un término `offset` nuestro modelo se ve así:

$$\eta_1(x) = \log\left(\frac{\mu}{p}\right) = \beta_0 + \sum_{i=1}^{19} \beta_i x_i$$

Donde η es nuestro componente lineal, x es un vector con todas nuestra variables, $\log(p)$ es nuestra variables `offset` y v_i son nuestras variables.

Al tener demasiadas variabels consideramos reducir el modelo para facilitar el trabajo, decidimo eliminar las interacciones entre las variables `Age` y `City` para retirar 12 variables de nuestro primer modelo. El modelo 2 queda de la siguiente manera:

$$\eta_2(x) = \log\left(\frac{\mu}{p}\right) = \beta_0 + \sum_{i=1}^{7} \beta_i x_i$$

Para ver si el modelo anidado ha mejorado, empleamos pruebas de hipótesis simultaneas que nos permiten comparar ambos modelos, como el segundo modelo está anidado, se puede considerar que el modelo reducido es como el primero pero considerando algunos β's iguales a cero.
Realizamos la prueba usando la función `anova()` de `R base`, la cual compara nuestros modelos anidados.
Basandonos en el `p-value` $= 0.4785 > \alpha = 0.05$ (nuestra significancia), podemos considerar plausible que el modelo 2 sea más adecuado que el modelo 1.

Considerando que aún tenemos bastantes variables en nuestro modelo reducido,realizamos un tercer modelo considerando únicamente la variable `Age`, con el fin de simplificar el modelo descartando la variable `City` el nuevo modelo quedaría de la siguiente manera:

$$\eta_3(x) = \log\left(\frac{\mu}{p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Nuevamente empleamos pruebas de hipótesis usando la función `anova()` de `R base`, la cual compara nuestros modelos anidados.

Comparando los `AIC`'s y `BIC`'s de los modelos anteriores:

	Modelo 1	**Modelo 2**	**Modelo 3**
AIC	121.4730	109.0705	108.4512
BIC	141.3876	117.0363	113.4299

Podemos ver que el modelo 3 tiene menor `AIC` y `BIC`. Basandonos en la prueba de comparación de modelos cuyo `p-value` $= 0.1459 > \alpha = 0.05$ junto a la tabla anterior, es plausible considerar que el modelo 3 es el mejor. Esto implica que , según lo observado, la variable `City` no está afectando de manera considerable la incidencia de cáncer pulmonar. Volviendo a la gráfica, como se mencionó en el análisis de la misma, el comportamiento en las 4 ciudades parece ser el mismo.

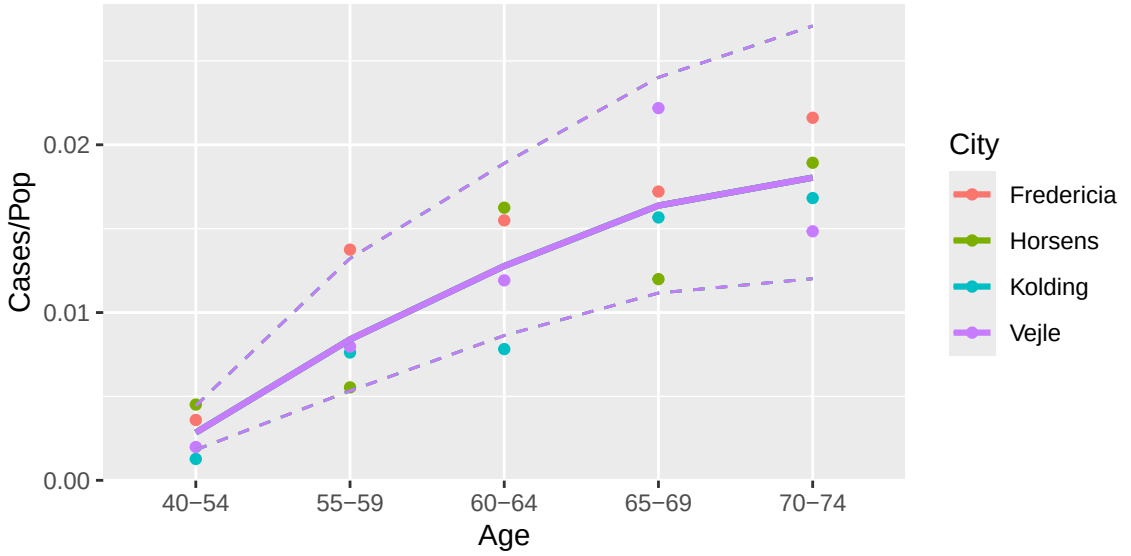
Con base a todo lo anterior optamos por pobar con un modelo binomial negativo con la misma función liga y suando únicamente la variables `Age` como en el modelo 3, para ello se emplea la función `glm.nb()` del paquete `MASS`, para luego comparar los modelos, esta vez veremos el `AIC` junto al `BIC`, y usaremos la estimación del paŕametro de dispersión, el cual debe ser cercano a 1.

	Modelo 1	Modelo 2	Modelo 3	Modelo 4
AIC	121.4730	109.0704725	108.451248	110.451499
BIC	141.3876	117.0363307	113.429909	116.425893
$\hat{\phi}$	-Inf	0.9664595	1.131886	1.131818

El parámetro de dispersión le los modelos 3 y 4 es precticamente el mismo, pero el modelo 3 tiene tanto `AIIC` como `BIC` menores al modelo 4, por lo que consideramos al modelo 3 el más adecuado para el análisis.

Intervalos de confianza

Procedemos a calcular intervalos de confianza simultáneos de las tasas de incidencia para cada uno de nuestros grupos de edad.



Con base en el gráfico anterior, dada la cerciente varianza en la incidencia con respecto a la edad, solo podemos indicar que a mayor edad existe mayor incidencia de cáncer pulmonar entre las edades 40 y 59, más allá no podemos asegurar nada.

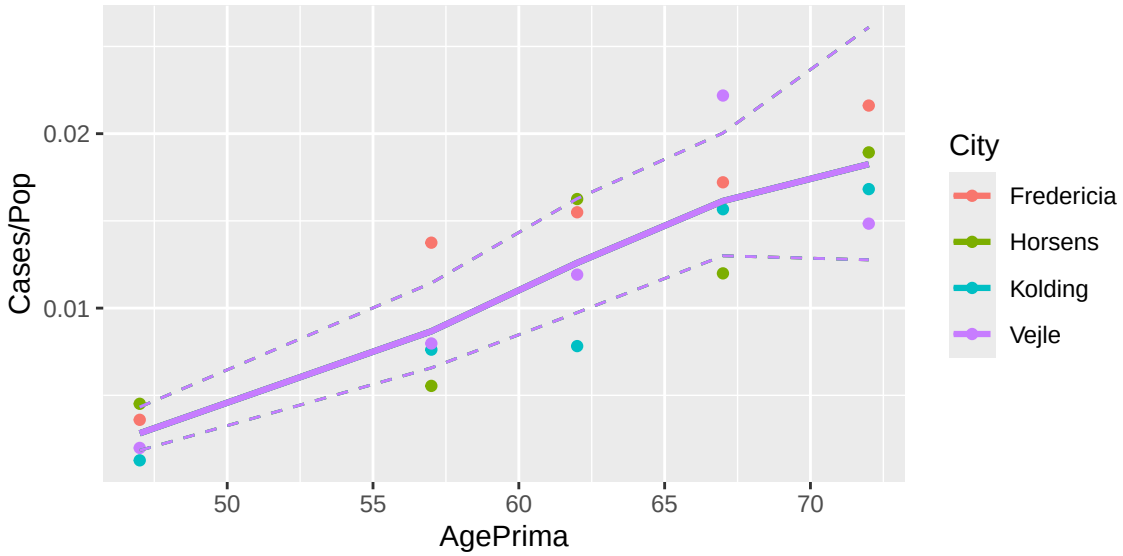
Edad continua

Utilizar la variable `Age` como categórica dificulta la interpretación de los resultados por lo que optamos por transformarla en una variable continua usando como referencia el valor medio de cada rango (`AgePrima`), ajustamos 4 nuevos modelos usando la distribución *Poisson* y la *Binoial Negativa*, empleando `AgePrima` y $\texttt{AgePrima}^2$.

	Modelo continuo 1	Modelo continuo 2	Modelo continuo 3	Modelo continuo 4
AIC	107.921925	104.509536	109.889748	106.509783
BIC	109.913389	107.496733	112.876945	110.492712
$\hat{\phi}$	1.247165	1.002152	1.179734	1.002091

La tabla anterior compara el `AIC`, el `BIC` y la estimación del parámetro de dispersión $\hat{\phi}$ de los 4 modelos nuevos. Basándonos en dicha tabla, parece que el modelo más adecuado para el análisis puede ser el modelo continuo 2, el cual emplea una distribución *Poisson* y un polinomio de segundo grado de la variable `AgePrima` como se muestra a continuación:

$$\eta(x) = \log\left(\frac{\mu}{p}\right) = \beta_0 + \beta_1 x + \beta_2 x^2$$



Con base en el nuevo modelo, podemos indicar que a mayor edad hay mayor incidencia en cancer de pulmón al menos entre los 40 y los 67 años, saliendo de ese rango no podemos aseurar lo mismo.