Modelos lineales generalizados para datos de conteos

La base de datos dada contiene información sobre el número de casos de cáncer de pulmón registrados entre 1968 y 1971 en cuatro ciudades de Dinamarca. En estos casos se registró también la edad de los pacientes. Hay que notar que los casos depende de forma inherente de la población de la ciudad, pues entre más grande la ciudad es mayor el número de casos que se pueden observar. Por esta razón, el estudio debe centrarse en las tasas de incidencia. De este modo, será posible analizar si se puede afirmar que a medida que aumenta la edad, se observa un incremento significativo en la incidencia de cáncer de pulmón.

En el Cuadro 8 se presenta una *muestra aleatoria* de los datos para observar su estructura de acuerdo a las categorias (No son todos los datos proporcionados):

Cuadro 8: Muestra aleatoria de los datos proporcionados

X	Cases	Pop	Age	City
7	13	2879	40-54	Horsens
22	14	631	65-69	Vejle
3	11	710	60-64	Fredericia
14	8	1050	55-59	Kolding
19	5	2520	40-54	Vejle

La visualización de los datos a través de la Figura 13 es la siguiente:

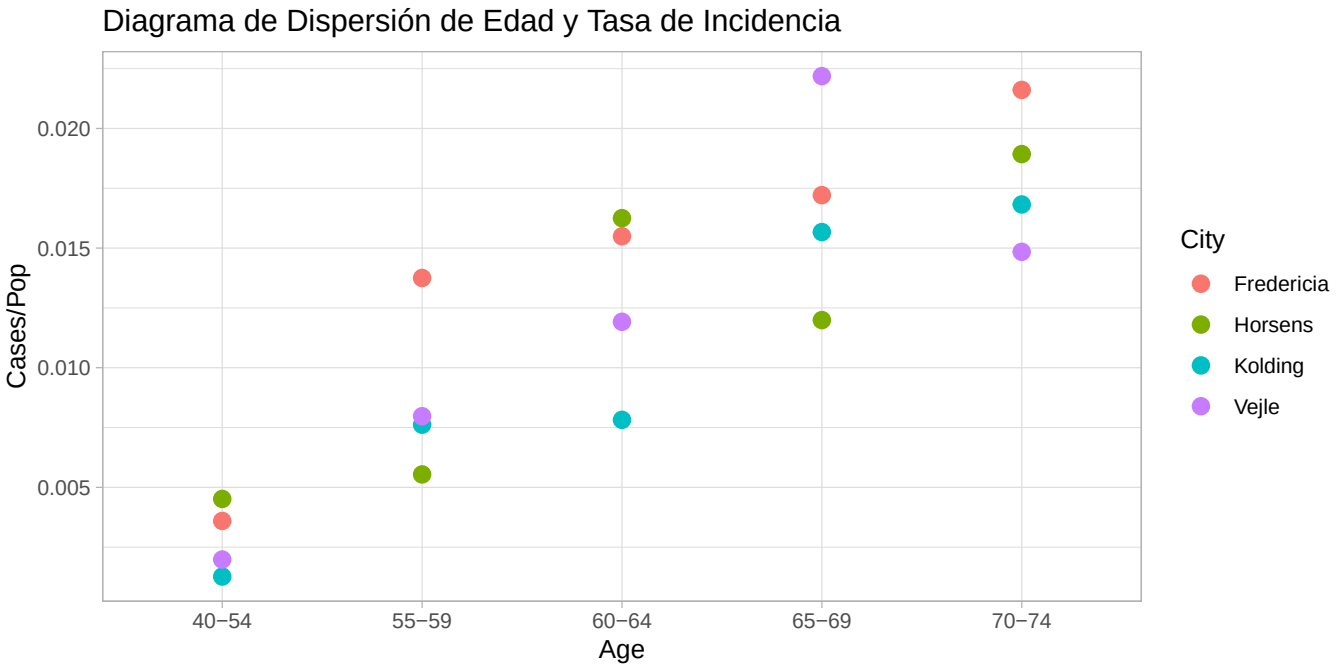

Figura 13: Diagrama de Dispersión

Es importante señalar que en el eje horizontal (*Age*), los grupos de edades se encuentran dispuestos en orden creciente, mientras que en el eje vertical (*Case/Pop*) se representa la tasa de incidencia de cada ciudad. Observamos que la tendencia que reflejan estos datos indica que, a medida que la edad avanza, independientemente de la ciudad, se registra un incremento en la tasa de incidencia.

Procederemos a desarrollar modelos destinados a capturar el comportamiento de los datos, con el fin de proporcionar una respuesta a la interrogante planteada: "¿Se puede indicar que a mayor edad existe mayor incidencia de cáncer de pulmón?". No obstante, debido a la imposibilidad de comparar muestras aleatorias en el contexto de datos de conteo, dado que estas dependen del tamaño de la población, resulta necesario realizar un ajuste en la población para adecuarla a estos modelos: El ajuste requerido implica tomar el logaritmo de la población para estandarizarla. Así, empleando una muestra de nuestros datos, se presentarán como lo están en el Cuadro 9:

Cuadro 9: Muestra aleatoria de los datos de conteo

X	Cases	Pop	Age	City	logPop
7	13	2879	40-54	Horsens	7.965
22	14	631	65-69	Vejle	6.447
3	11	710	60-64	Fredericia	6.565
14	8	1050	55-59	Kolding	6.957
19	5	2520	40-54	Vejle	7.832
13	4	3142	40-54	Kolding	8.053

De esta manera, nos enfocaremos en la utilización de Modelos Lineales Generalizados para el análisis de datos de conteo. En una etapa inicial, se evaluaron dos modelos, ambos con distribución Poisson y función liga logarítmica:

- El primer modelo consideraba la totalidad de las variables en nuestra base de datos, incluyendo las dos variables categóricas (Age y City).
- El segundo modelo, con las mismas características de distribución y función liga, excluía la variable categórica City, manteniendo el resto de las variables.

Con el propósito de simplificar el análisis, se optó por utilizar el segundo modelo, fundamentando esta elección en el resultado de una prueba de hipótesis posterior.

$$H_0 : \text{Se utiliza el modelo reducido} \quad vs \quad H_a : \text{Se utiliza el modelo completo}$$

```
## Analysis of Deviance Table
##
## Model 1: Cases ~ Age * City + offset(logPop)
## Model 2: Cases ~ Age + offset(logPop)
##   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
## 1         0          0
## 2        15         17 -15      -17     0.32
```

Dado que estamos trabajando con un nivel de significancia de $\alpha = 0.05$, notemos que

$$0.32 = P - value > \alpha = 0.05$$

por lo que no se rechaza H_0. Esto nos dice que no hay evidencia en contra de poder trabajar con el modelo reducido. Además, se utilizaron los criterios AIC y BIC para determinar qué modelo era mejor.

Cuadro 10: Puntajes de AIC y BIC para ambos modelos

	AIC	BIC	Dispersión_Parametros
Modelo_completo	121.5	141.4	-Inf
Modelo_reducido	108.5	113.4	1.132

Notar que en el Cuadro 10 los valores del modelo reducido son menores. Además, es importante señalar que la dispersión de los parámetros en el modelo reducido es más próxima a uno en comparación con el modelo completo, donde el resultado parece no estar bien definido.

Posteriormente, se tomó la determinación de incorporar un tercer modelo para el análisis de los datos, caracterizado por una distribución Binomial Negativa y una función liga logarítmica. Siguiendo una configuración similar al segundo modelo, se excluyó la variable categórica “City” manteniendo el resto de las variables intactas.

Mediante la utilización nuevamente de los criterios AIC y BIC, se concluyó que el modelo más adecuado para el análisis seguía siendo el segundo modelo, ya que presentaba los valores más bajos de AIC y BIC en comparación con los otros modelos considerados. Véase el Cuadro 11

Cuadro 11: Puntajes de AIC y BIC para los tres modelos

	Distribución	Funcion_Liga	AIC	BIC
Primer Modelo	Poisson	Log	121.5	141.4
Segundo Molelo	Poisson	Log	108.5	113.4
Tercer Modelo	Bin. Neg.	Log	110.5	116.4

Es por este motivo, por el cual procederemos a hacer la verifiación de supuestos para el segundo modelo:

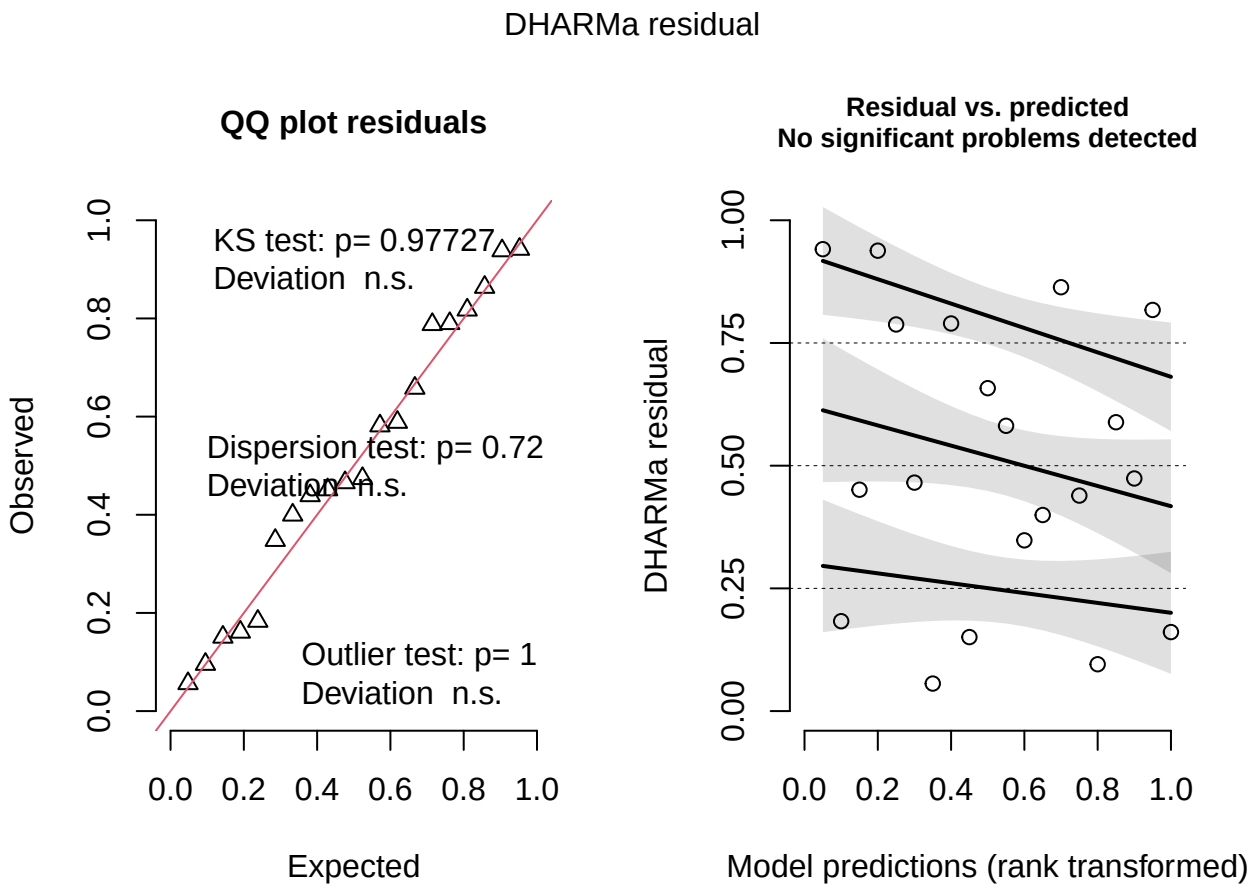

Figura 14: Comprobación de supuestos

Vemos que, en la Figura 14, las gráficas que se encuentran del lado izquierdo, no parece haber problemas con los supuestos, pues:

- **KS test** - NO se rechaza: Se puede asumir una distribución Poisson con los residuales simulados.
- **Dispersion test** - NO se rechaza: Se cumple la estimación de la varianza.
- **Outlier test** - NO se rechaza: No hay que poner atención a un punto fuera de lo común.

También, gracias a la Figura 14, la gráfica del lado derecho, parece no haber problema con la linealidad.

En virtud de que nuestro modelo satisface los supuestos establecidos, estamos en condiciones de proceder con el análisis de las tasas de incidencia en cada grupo de edad. Para llevar a cabo este análisis, haremos uso de intervalos de confianza simultáneos con un nivel de confianza del 95 %. Estos intervalos se presentan gráficamente en el siguiente diagrama de dispersión.

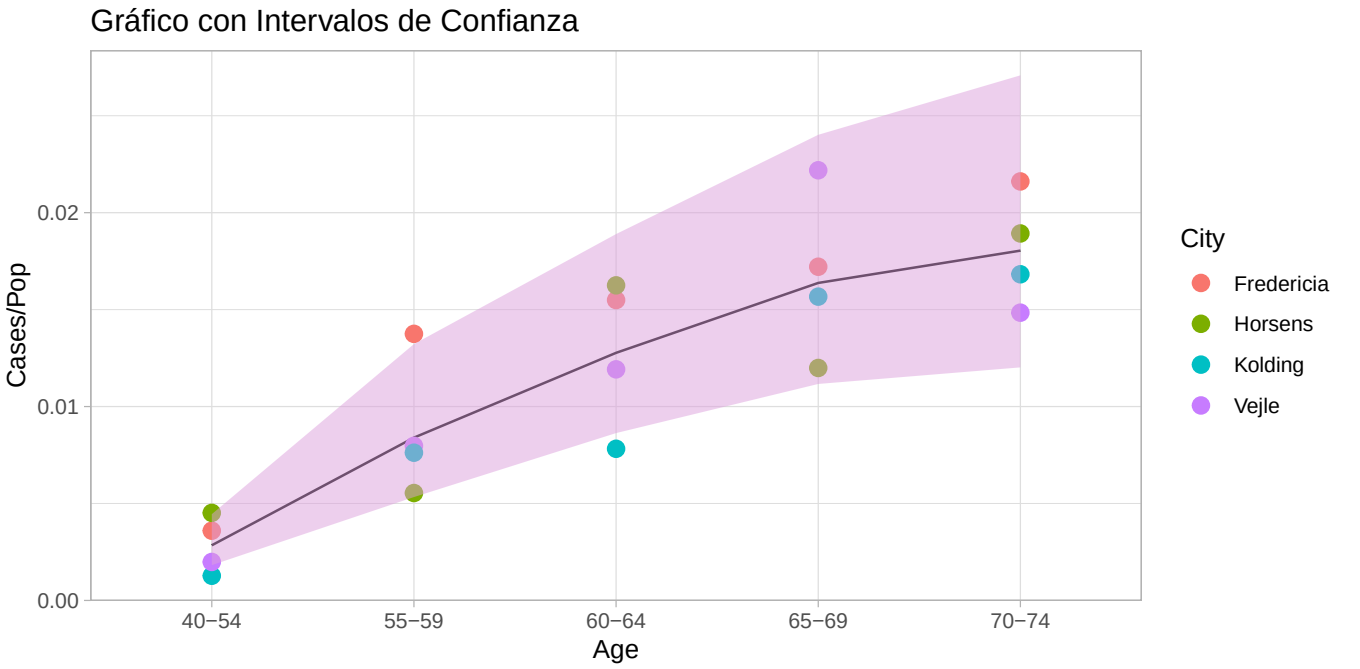

Figura 15: Diagrama de dispersión con intervalos de confianza

A partir de este punto, se puede observar en la Figura 15 que sí existe un aumento en la incidencia de cáncer de pulmón a medida que aumenta la edad. Es evidente que al comparar los grupos de edad iniciales con los últimos, el aumento de los intervalo en la tasa de incidencia de cáncer de pulmón se vuelve más notable.

Como un segundo enfoque, se consideró el punto medio definido entre cada categoria de edad para así considerarla como una variable continua.

Con la incorporación de la nueva covariable denominada "*Ageprima*", se llevaron a cabo ajustes de modelos que emplearon distribuciones Poisson o Binomial Negativa, ambos con ligas logarítmicas. Asimismo, se consideró la inclusión o exclusión del aporte de esta misma covariable al cuadrado, es decir, "$Ageprima^2$".

En el Cuadro 12 se muestran los valores de los índices AIC y BIC obtenidos para cada modelo.

Cuadro 12: Puntajes de AIC y BIC para los modelos ajustados

	Distribución	Liga	Ageprima_cuadrada	AIC	BIC
Modelo 1	Poisson	Log	NO	107.9	109.9
Modelo 2	Poisson	Log	SI	104.5	107.5
Modelo 3	Bin. Neg.	Log	NO	109.9	112.9
Modelo 4	Bin. Neg.	Log	SI	106.5	110.5

Notamos que el segundo ajuste es el que cuenta con los menores índices, por lo tanto se seguirá el análisis únicamente con este.

En la Figura 16 se tiene el análisis de supuestos, con la cual se puede concluir que

- **KS test** - NO se rechaza: Se puede asumir una distribución Poisson con los residuales simulados.
- **Dispersion test** - NO se rechaza: Se cumple la estimación de la varianza.
- **Outlier test** - NO se rechaza: No hay que poner atención a un punto fuera de lo común.

Además de que tampoco parece haber problema con la linealidad. Por lo tanto, parece plausible continuar con este modelo.

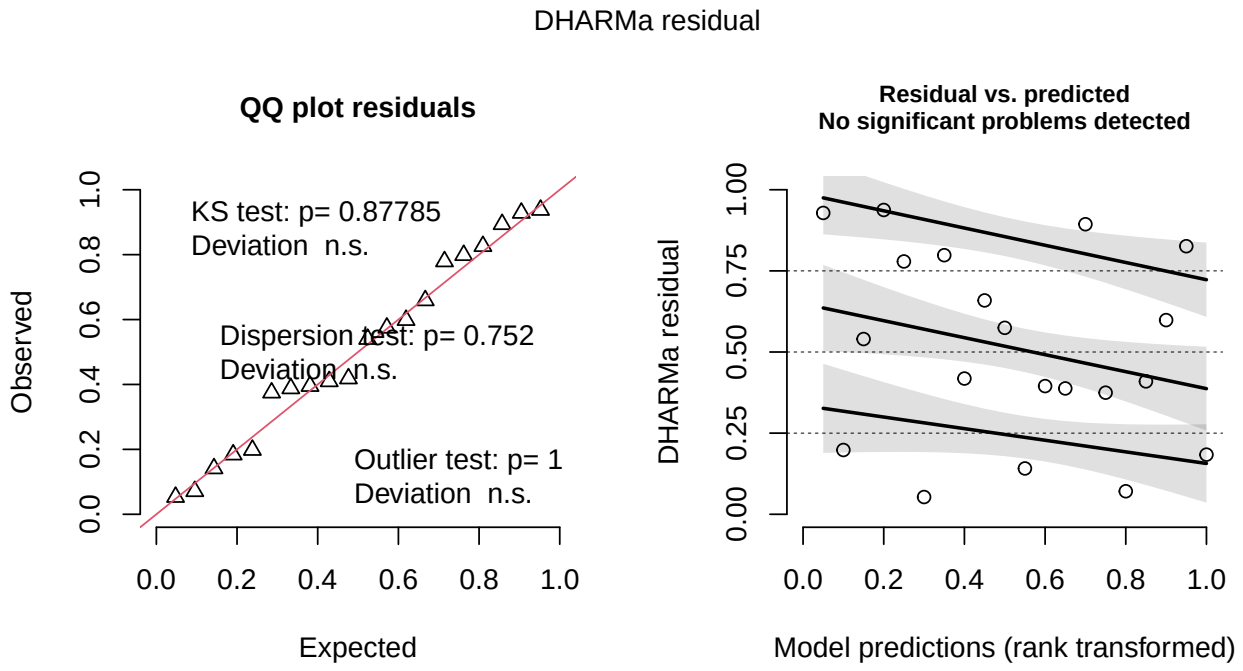

Figura 16: Comprobación de supuestos

Finalmente, en la Figura 17 se muestra la curva ajustada con sus respectivos intervalos de confianza.

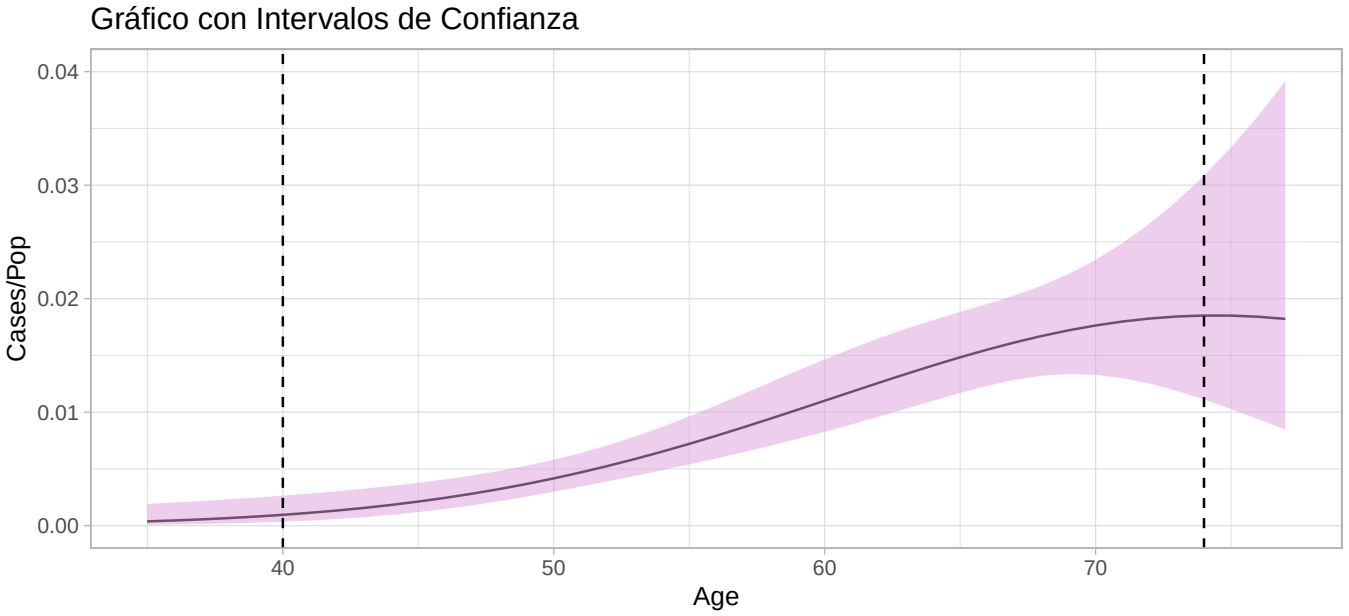

Figura 17: Comportamiento del modelo ajustado, con los intervalos de confianza sombreados

De esta manera, a partir de la observación del gráfico, es evidente el patrón ascendente del modelo. Específicamente, en el intervalo de edades $[40, 74]$, que es de importancia en nuestro estudio y se encuentra resaltado entre línea punteadas, se cumple que a medida que aumenta la edad, la incidencia de cáncer de pulmón también se incrementa.