4. Modelos lineales generalizados para datos de conteos

Tarea-examen 1: Aprendizaje estadístico supervisado

Carlos Iván Canto Varela 315649888

La siguiente gráfica de dispersión provee una idea básica de los datos proporcionados.

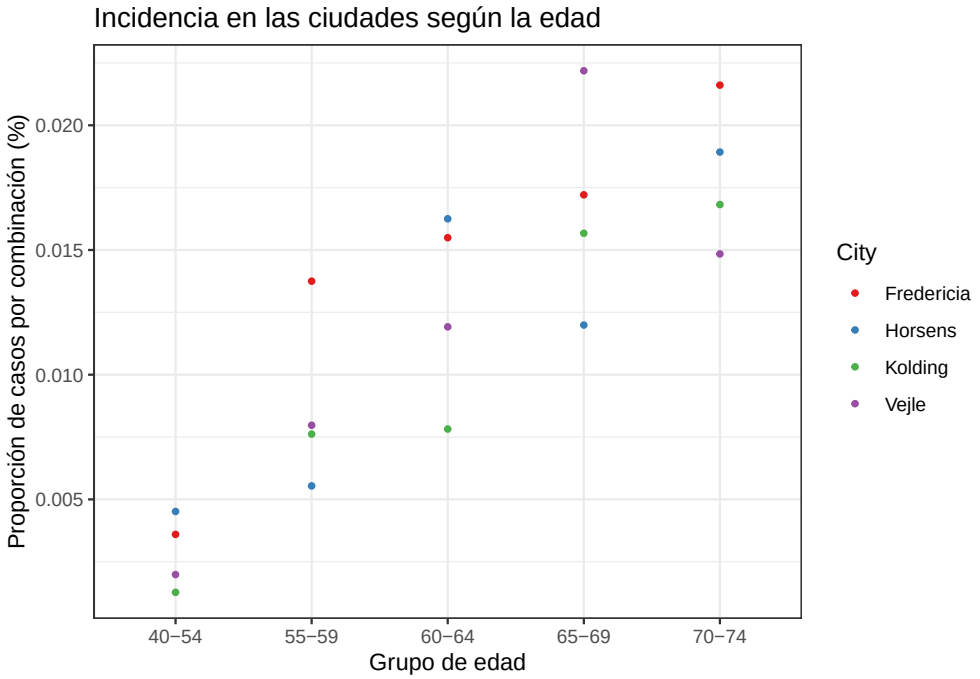

Figura 1: Diagrama de dispersión para los datos

Puede verse una clarísima tendencia al aumento de casos conforme al grupo de edad para todas las ciudades. Adicionalmente, pareciera que la ciudad de Fredericia tiene una proporción mayor relativa a las demás ciudades.

Como primer modelo propuesto se tomará una distribución Poisson con liga logarítimica y las interacciones entre las variables categóricas. El término *offset* es necesario en el modelo pues se usarán las tasas de incidencia.

También se considerará un modelo con sólo la edad como variable.

```
## 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      0.000
## 2        15     16.978 -15  -16.978   0.3202
```

Con un valor p mayor a una significancia de 0.05, se puede asentir que no hay evidencia en favor a la diferencia entre los coeficientes añadidos y el cero. Por otra parte, los criterios de desempeño

Puntajes por versión del modelo		
Modelo	**Puntaje AIC**	**Puntaje BIC**
Original	121.473	141.3876
Reducido	**108.4512**	**113.4299**

indican que el mejor modelo es el reducido. Adicionalmente, se vio que el comportamiento de las incidencias por ciudad es visualmente similar entre sí; por lo que existe un argumento más a favor de la reducción.

Por otra parte, un ajuste con distribución binomial negativa también puede ser significante.

```
## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace =
## control$trace > : iteration limit reached

## Warning in theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace =
## control$trace > : iteration limit reached


##
## Call:
## glm.nb(formula = Cases ~ Age + offset(logPop), data = Data, link = "log",
##     init.theta = 152367.3428)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8558  -0.6424  -0.1428   0.6526   1.5468
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.8623     0.1741 -33.675  < 2e-16 ***
## Age55-59      1.0823     0.2481   4.362 1.29e-05 ***
## Age60-64      1.5017     0.2314   6.489 8.67e-11 ***
## Age65-69      1.7503     0.2292   7.637 2.23e-14 ***
## Age70-74      1.8472     0.2352   7.855 4.01e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(152367.3) family taken to be 1)
##
##     Null deviance: 115.425  on 19  degrees of freedom
## Residual deviance:  16.977  on 15  degrees of freedom
## AIC: 110.45
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  152367
##           Std. Err.:  5232732
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -98.451
```

Nótese la advertencia sobre las iteraciones del cálculo para theta en el modelo binomial negativo. Esto podría ser indicador que los datos no están lo suficientemente dispersos en relatividad a la media, por lo que el modelo reducido con distribución Poisson será más apropiado para el caso en mano. Además, el Poisson mantuvo mejores coeficientes AIC y BIC.

Dicho esto, los intervalos de confianza simultáneos al 95 % según la edad son

$$40-54:[0.0018,0.0044],\quad 55-59:[0.0053,0.0132],\quad 60-64:[0.0086,0.0189],$$
$$65-69:[0.0112,0.024]\quad \text{y}\quad 70-74:[0.012,0.0271];$$

y se ilustran a continuación:

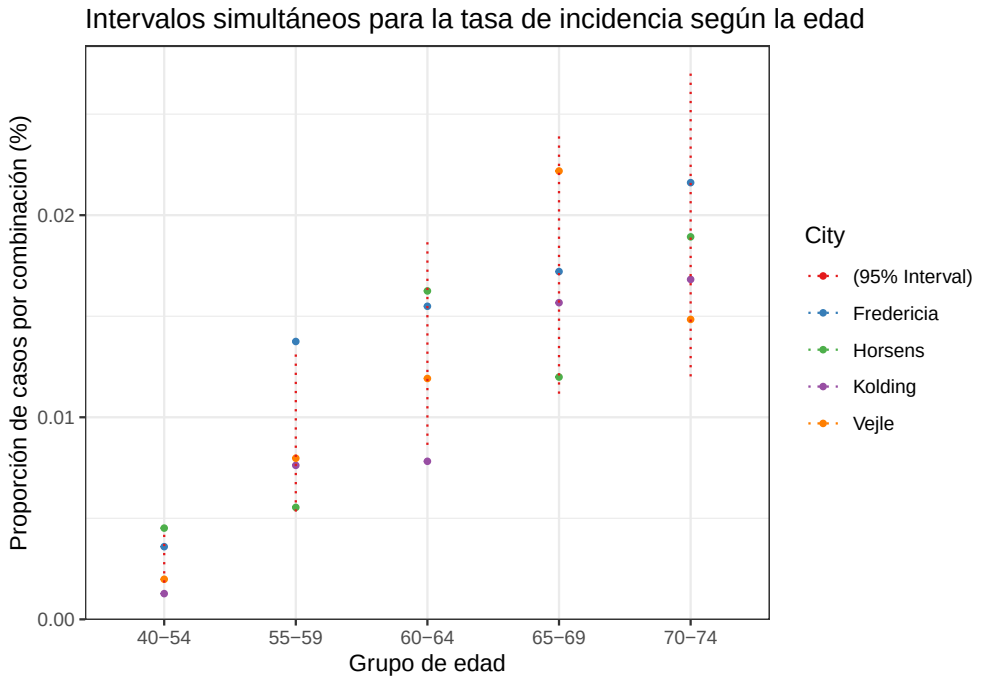

Figura 2: Diagrama de dispersión con intervalos de confianza simultáneos

La otra alternativa es adaptar un modelo donde la edad sea variable continua mediante aproximaciones con los puntos medios de cada intervalo.

Después de ajustar los cuatro modelos, el que parece más adecuado de utilizar es el distribuido Poisson con las variables Edad y Edad cuadrada continuas. Este no sólo cumple los supuestos requeridos, sino que también tuvo los menores puntajes AIC y BIC.

Dicho esto, una pregunta de interés es si a mayor edad existe mayor incidencia de cáncer de pulmón. De forma analítica, esto se puede verificar por la función del modelo ajustado:

$$\begin{aligned}
&\ln(\mu)=\beta_0+\beta_1\text{Edad}+\beta_2\text{Edad}^2+\ln(\text{Pop})\\
\iff\ &L\equiv\ln\left(\frac{\mu}{\text{Pop}}\right)=\beta_0+\beta_1\text{Edad}+\beta_2\text{Edad}^2\\
\iff\ &\frac{d}{d\text{Edad}}L=\beta_1+\beta_2\text{Edad}.
\end{aligned}$$

Para interpretar esto, se usó la prueba de hipótesis para

$$H_a: \quad \beta_2 \geq 0 \quad \text{vs.} \quad H_0: \quad \beta_2 < 0$$

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: glm(formula = Cases ~ Ageprima + Ageprima2 + offset(logPop),
##     family = poisson(link = "log"), data = Data)
##
## Linear Hypotheses:
##         Estimate Std. Error z value Pr(<z)
## 1 >= 0 -0.002502   0.001081  -2.314 0.0103 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

donde el valor p menor a una significancia de 0.05 implica que la derivada será una recta con pendiente negativa. Para las edades de interés se sigue que esta es positiva, por lo que la función del modelo es creciente. En otras palabras, se puede decir que a mayor edad hay una mayor incidencia de cáncer de pulmón.

Véase la gráfica siguiente para una ilustración visual del fenómeno:

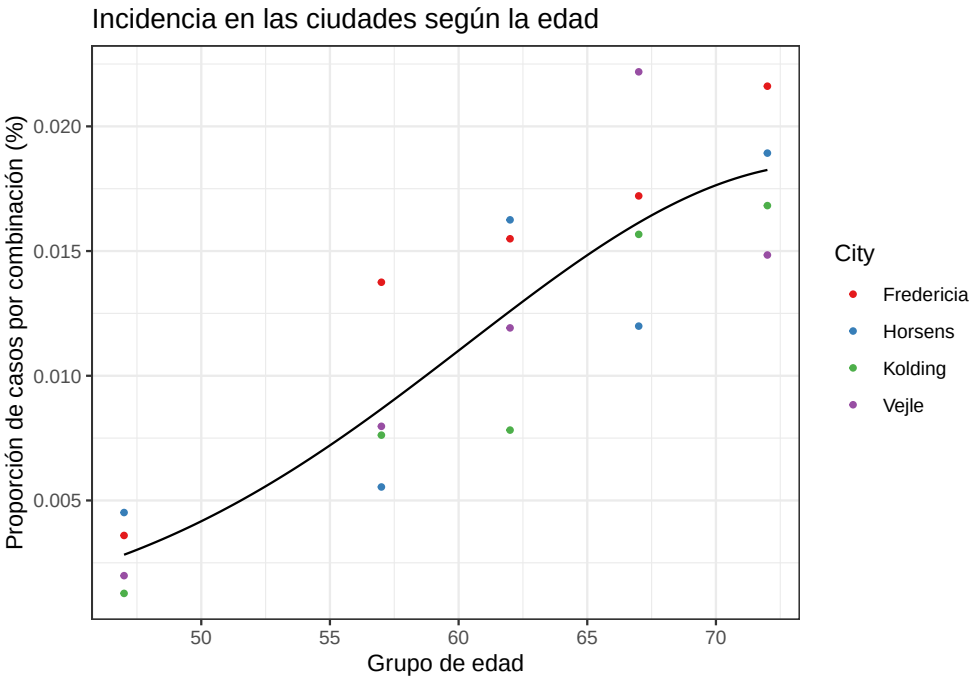

Figura 3: Crecimiento del modelo con la edad como variable continua