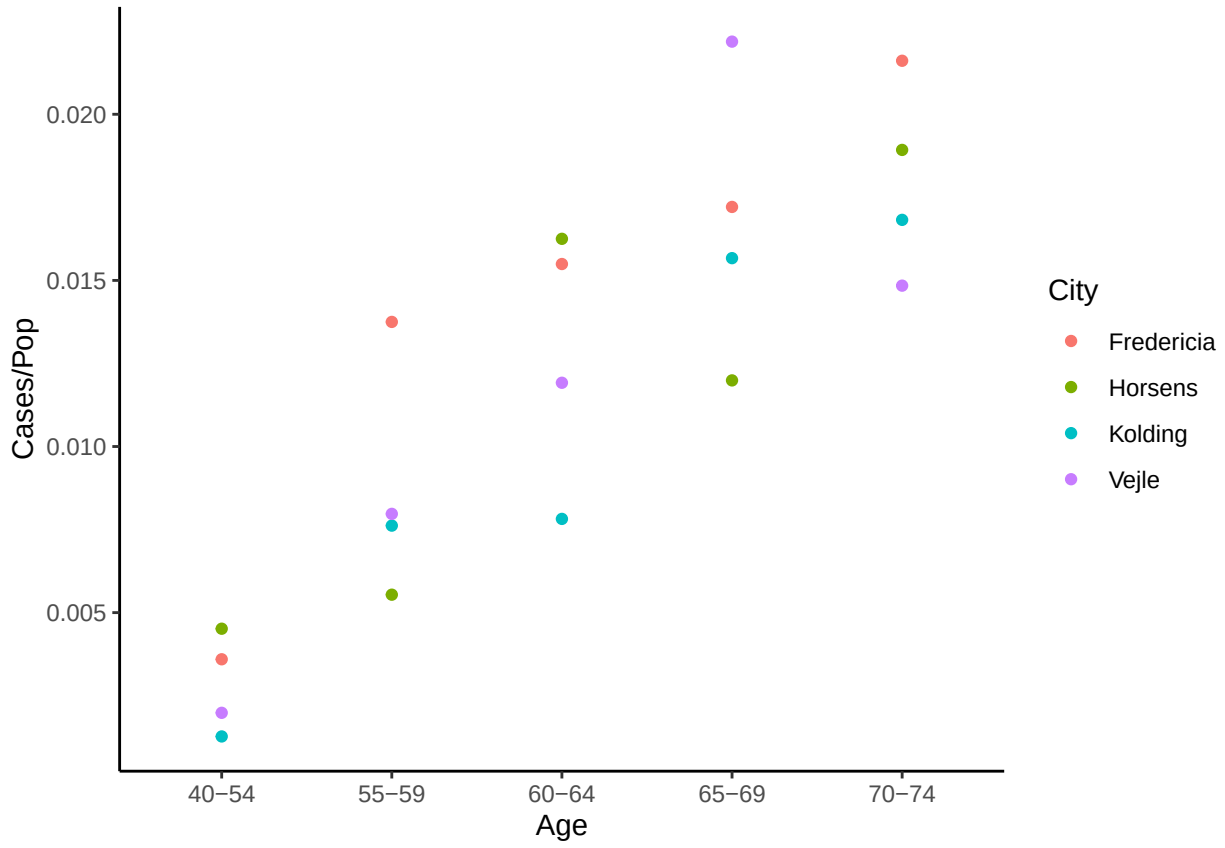
Tarea 1

Ejercicio 4

Analicemos los datos.



Podemos observar una fuerte tendencia en los datos respecto a la tasa de incidencia de los casos de cáncer de pulmón contra la edad, hay un claro incremento conforme avanzamos en los grupos de edad.

También podríamos argumentar que la distinción por ciudades no hace un gran cambio a simple vista. Probablemente la inclusión de esta variable cause ruido en nuestro modelo, pero no podemos descartar que afecte la ciudad para la tasa hasta hacer una prueba formal.

Modelos con Poisson

Para descartar este hecho, hemos hecho dos modelos y los contrastamos. En el primero consideraremos todas las interacciones entre la edad y la ciudad, y en el segundo únicamente a los grupos de edad.

En ambos casos los modelos fueron basados en la distribución Poisson con liga logarítmica.

Nuestro primer modelo con todas las interacciones entre ciudad y grupos de edad no cumple siquiera con los supuestos, mientras que el segundo únicamente considerando los grupos de edad sí los cumple.

Además de esto, el primer modelo generó una barbaridad de coeficientes. Deseamos poder considerar al segundo modelo en su lugar. En este caso usamos criterios como el AIC, BIC y una prueba de bondad de ajuste ya que estos modelos están anidados y queremos tomar al segundo modelo, que es una versión reducida del primero.

Primero los AIC y BIC:

```
c(AIC(fit.poi1), AIC(fit.poi2))
```

```
## [1] 121.4730 108.4512
```

```
c(BIC(fit.poi1), BIC(fit.poi2))
```

```
## [1] 141.3876 113.4299
```

Tenemos menores AIC y BIC para el segundo modelo.

Ahora la prueba de bondad de ajuste:

```
## 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
```

Como obtuvimos un p-value de 0.32, no tenemos evidencia en contra de que debemos tomar el modelo completo ante el reducido. Es decir, podemos quedarnos con el reducido.

Y esto hace mucho sentido ya que desde un inicio, en la gráfica pudimos observar que independientemente de la ciudad, las tasas de incidencia parecían ser muy iguales entre todas.

Modelo Binomial Negativo

Ahora que hemos seleccionado un buen modelo Poisson con liga logarítmica, también probaremos un ajuste con la distribución Binomial Negativa (liga logarítmica de igual manera).

Este nuevo modelo con la distribución Binomial Negativa también cumple con los supuestos.

Por lo que podemos comparar los AIC y BIC de este modelo con el Poisson anterior y decidirnos por alguno:

```
c(AIC(fit.nb), AIC(fit.poi2))
```

```
## [1] 110.4515 108.4512
```

```
c(BIC(fit.nb), BIC(fit.poi2))
```

```
## [1] 116.4259 113.4299
```

El modelo Poisson tiene menores índices, por lo que también lo seleccionamos por encima del binomial negativo.

Intervalos de Confianza: Grupos de Edad

Mostraremos intervalos de confianza al 95% para los diferentes grupos de edades:

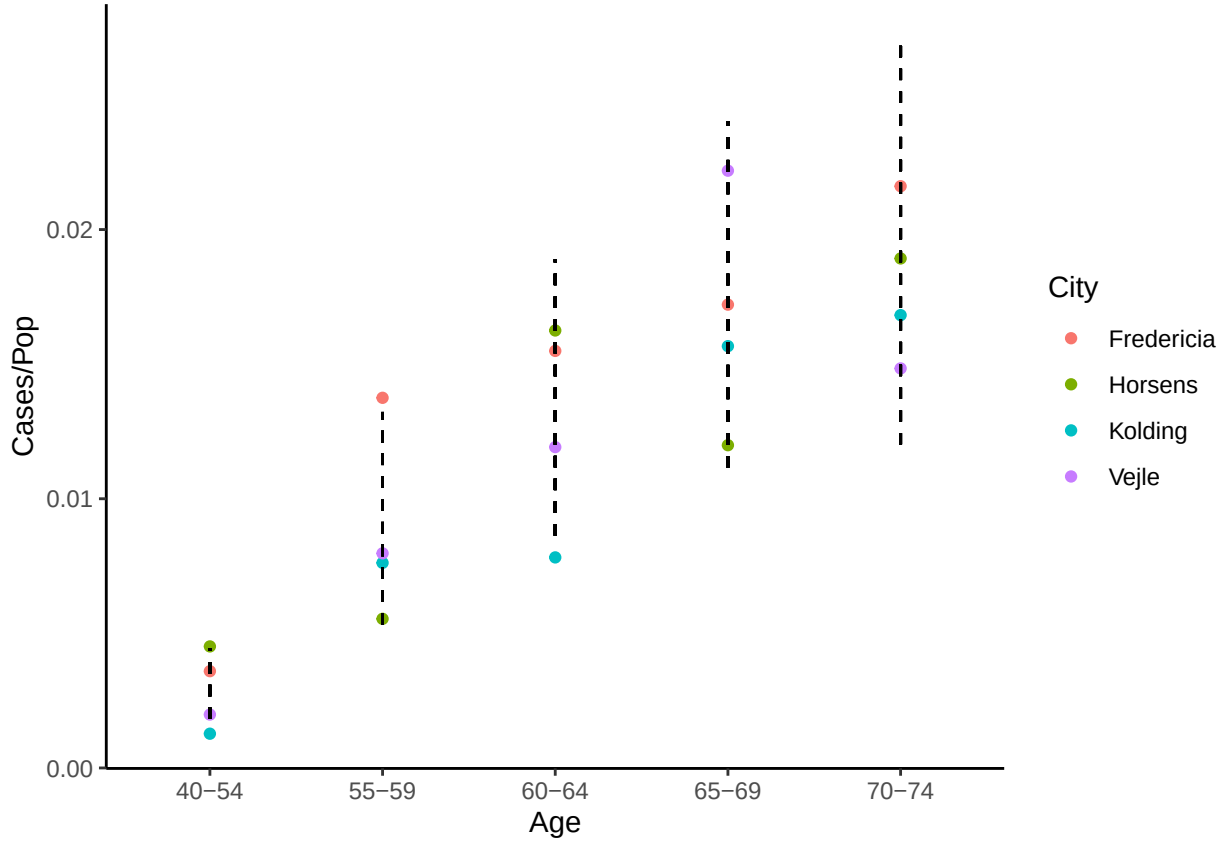
```
## Loading required package: mvtnorm
```

```
## Loading required package: survival
```

```
## Loading required package: TH.data
```

```
##
## Attaching package: 'TH.data'
```

```
## The following object is masked from 'package:MASS':
##
##     geyser
```



A ojo gracias a estos intervalos de confianza podemos suponer que en efecto, a mayor edad mayores tasas de incidencias. Pero haremos una prueba para verificar esto.

Contrastamos los dos primeros grupos de edad, 40-54 y 55-59, contra los dos últimos, 65-69 y 70-74, ya que a ojo es evidente que a gran escala sí podemos concluir que hay un aumento en la tasa de incidencia de casos de cáncer de pulmón.

Serán pruebas simultáneas siguiendo la lógica

$H_0 : \mu(Cases/Pop \mid AgeGroup_{70-74}) \leq \mu(Cases/Pop \mid AgeGroup_{40-54})$

Y más estrictamente contrastamos la siguiente hipótesis:

$H_0 : \beta_4 - \beta_0 \leq 0 \ \cap \ \beta_3 - \beta_0 \leq 0 \ \cap \ \beta_4 - \beta_1 \leq 0 \ \cap \beta_3 - \beta_1 \leq 0$

Es decir, las pruebas del primer grupo contra los últimos dos grupos de edad, y del segundo grupo contra los dos últimos grupos de edad.

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: glm(formula = Cases ~ Age + offset(logPop), family = poisson(link = "log"),
##     data = data)
```

```
##
## Linear Hypotheses:
##        Estimate Std. Error z value  Pr(>z)
## 1 <= 0   7.7095     0.3824  20.162 < 0.001 ***
## 2 <= 0   7.6125     0.3787  20.100 < 0.001 ***
## 3 <= 0   0.7649     0.2372   3.225 0.00208 **
## 4 <= 0   0.6679     0.2312   2.889 0.00684 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

En todas las pruebas obtuvimos un p-value menor al 0.05, por lo que en efecto tenemos evidencia en contra de que al aumentar la edad se reduce el riesgo.

Es decir, que lo que suponíamos se cumple. A mayor edad mayor riesgo de padecer cáncer de pulmón.

Modelo con Edad Continua

Por último consideramos un modelo suponiendo que la edad es una variable continua, no categórica. Para eso les dimos una edad puntual a cada observación de cada grupo. Ésta fue el punto medio de su grupo de edad (por ejemplo al Grupo 40-54 se le asignó la edad de 47 años)

Creamos 4 modelos, 2 con Poisson y liga logarítmica y 2 con Binomial Negativa y la misma función liga. Para cada distribución consideramos un modelo con sólo interacciones de la edad vista como variable continua, y para su segundo modelo respectivo uno donde se tome en cuenta la edad y la edad^2 como variables continuas.

Desafortunadamente sólo 2 de los 4 modelos cumplen con los supuestos, y coinciden en ser los dos modelos donde se toma en cuenta la influencia de la edad y la edad^2 (es decir tenemos un modelo Poisson y uno Binomial Negativo).

A continuación mostramos sus AIC y BIC:

```
c(AIC(Poi), AIC(BN))
```

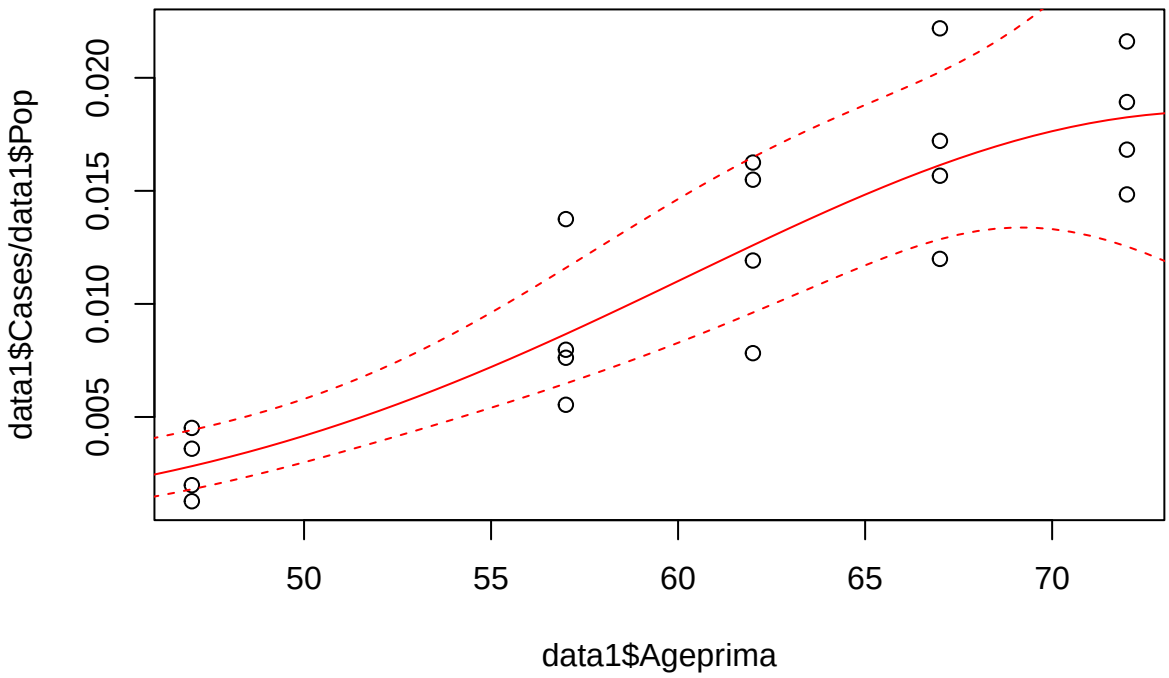
```
## [1] 104.5095 106.5098
```

```
c(BIC(Poi), BIC(BN))
```

```
## [1] 107.4967 110.4927
```

Por lo que consideraremos el Poisson.

Para finalizar este reporte, mostraremos intervalos de confianza continuos usando este último modelo tomando a la edad como variable continua para apoyar el hecho de que a mayor edad, mayor incidencia de cáncer de pulmón.



Para dar un poco más de rigurisidad veamos que esto tiene sentido, ya que nuestra variable edad sólo nos interesa en valores entre 40 y 74 años, es decir $40 \leq x \leq 74$, donde x representa la edad de algún paciente. Y si a mayor edad, mayor probabilidad de tener cáncer, eso nos implicaría que nuestra función de incidencia es creciente respecto a la edad. Lo que a su vez lo podemos traducir en que si tenemos una derivada positiva de esta función para el intervalo $[40,\ 74]$, entonces en efecto podemos concluir lo deseado.

Esto se cumple ya que, equivalentemente, la derivada de nuestra función liga se ve así:

$\frac{d}{dx}\mu(Cases/Pop \mid Age = x) = \frac{d}{dx}(\beta_0 + \beta_1 x + \beta_2 x^2) = \beta_1 + 2\beta_2 x$, y dado a que nuestros valores del modelo de β_1 y β_2 son respectivamente 0.3723 y -0.0025, la recta $\beta_1 + 2\beta_2 x$ es positiva en el intervalo $(-\infty, \frac{-\beta_1}{2\beta_2}) = (-\infty, 74.46)$.

En particular la derivada $\frac{d}{dx}\mu(Cases/Pop \mid Age = x) = \beta_1 + 2\beta_2 x$ es positiva en el intervalo $[40,\ 74]$.

Con lo que concluimos que en efecto a mayor edad, mayor indicencia en los casos de cáncer de pulmón, pues la función de tasa de incidencia es creciente en el intervalo deseado.

Ejercicio 5