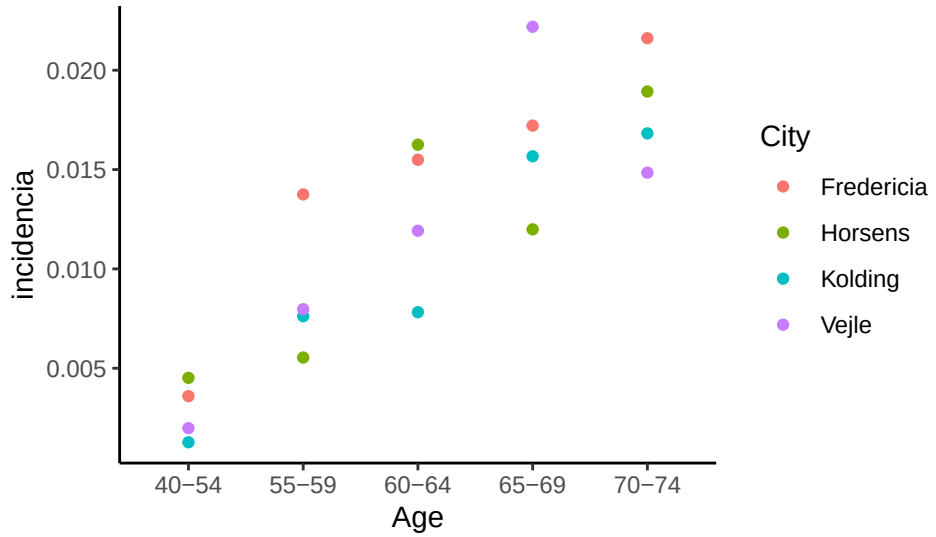
4. Modelos lineales generalizados para datos de conteos

La base de datos Preg4.csv contiene información sobre el número de casos de cáncer de pulmón (Cases) registrados entre 1968 y 1971 en cuatro ciudades de Dinamarca (City). En estos casos se registró también la edad de los pacientes (Age, variable categorizada en 5 grupos). El interés del análisis es estudiar si se puede indicar que a mayor edad existe mayor incidencia de cáncer de pulmón. Notemos que para realizar el análisis la variable de conteos Cases depende de forma inherente de la población de la ciudad (Pop), pues entre más grande la ciudad es mayor el número de casos que se pueden observar; de manera que el estudio se debe enfocar en las tasas de incidencia.

i) Gráfica de dispersión de grupos de edad e incidencia

Podemos apreciar de la siguiente Gráfica presentada que por cada grupo de edad la incidencia en cada ciudad va en aumento, por ejemplo en el grupo de edad de 40-54 la incidencia de cáncer esta por debajo de 0.005 pero conforme avanzan los grupos de edad los niveles aumentan para todas las ciudades.



ii) Distribución Poisson con liga logarítmica y un segundo modelo.

Como primer modelo consideraremos uno con distribución Poisson y función log, ademas de considerar las demás covariables de Age y City con su interacción. Aplicamos el código glm(formula = Cases $\sim$ offset(logPop) + Age * City, family = poisson(link = "log").

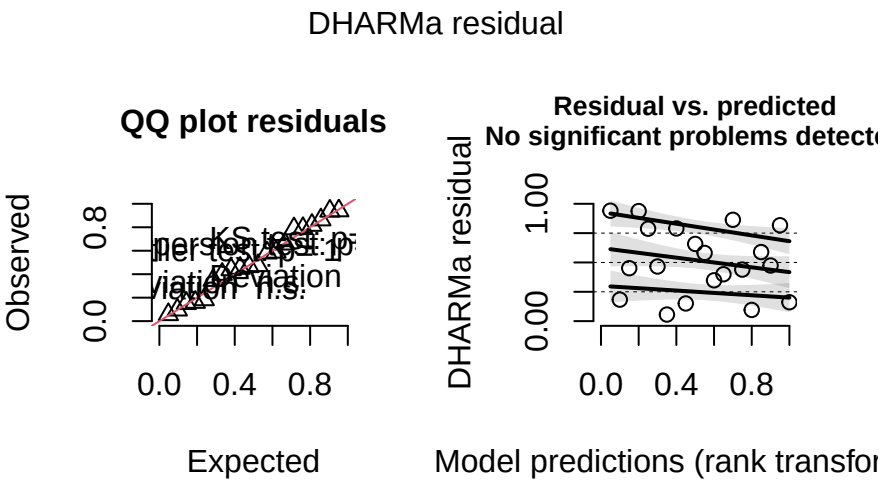
El AIC obtenido con este Modelo 1 es de 121.47 y de la regla del dedo para analizar si hay un problema por considerar el parametro de dispersion igual a 1 obtuvimos un valor de $-\infty$, lo cual nos dice que no es un bueno modelo, de todas formas se hizo la verificacion de los supuestos que por cuestión de espacio no se muestra, pero salio muy mal en estos por lo que se decidió no usarse.

De las verificaciones de los supuestos para el primer modelo observamos que tenemos muchos problemas por lo que optaremos por ajustar un segundo modelo, con la diferencia de que solo incluiremos a la covariable Age sin interacción. Usamos el código glm(formula = Cases $\sim$ offset(logPop) + Age, family = poisson(link = "log"), data = data4). Este Modelo 2 nos da un AIC de 108.45.

Vemos que este Modelo 2 tuvimos un valor de 1.131886, muy cercano a 1 con la regla de dedo para analizar si hay un problema por considerar el parametro de dispersion igual a 1, lo cual nos dice que es buen modelo, por lo que procedimos con la verificación de los supuestos del modelo de manera gráfica.

Table 1:

	Dependent variable:	
	Cases	
	(1)	(2)
Age55-59	1.341***	1.082***
	(0.426)	(0.248)
Age60-64	1.461***	1.502***
	(0.426)	(0.231)
Age65-69	1.566***	1.750***
	(0.437)	(0.229)
Age70-74	1.793***	1.847***
	(0.426)	(0.235)
CityHorsens	0.228	
	(0.410)	
CityKolding	−1.038*	
	(0.584)	
CityVejle	−0.595	
	(0.539)	
Age55-59:CityHorsens	−1.137*	
	(0.652)	
Age60-64:CityHorsens	−0.180	
	(0.570)	
Age65-69:CityHorsens	−0.589	
	(0.606)	
Age70-74:CityHorsens	−0.360	
	(0.585)	
Age55-59:CityKolding	0.448	
	(0.746)	
Age60-64:CityKolding	0.355	
	(0.758)	
Age65-69:CityKolding	0.944	
	(0.729)	
Age70-74:CityKolding	0.788	
	(0.737)	
Age55-59:CityVejle	0.050	
	(0.724)	
Age60-64:CityVejle	0.332	
	(0.694)	
Age65-69:CityVejle	0.849	
	(0.680)	
Age70-74:CityVejle	0.219	
	(0.712)	
Constant	−5.628***	−5.862***
	(0.302)	(0.174)
Observations	20	20
Log Likelihood	−40.736	−49.226
Akaike Inf. Crit.	121.473	108.451
Note:	*p<0.1; **p<0.05; ***p<0.01	



Lo que sigue sera hacer una prueba anova en la que compararemos ambos modelos y decidir si se puede usar el segundo modelo.

```
## Analysis of Deviance Table
##
## Model 1: Cases ~ offset(logPop) + Age * City
## Model 2: Cases ~ offset(logPop) + Age
##   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 mayor que 0.05 no tenemos evidencia suficiente para rechazar la hipótesis nula, por lo que concluimos que no se tiene una mejora significativa tomando mas variables y su interacción.Adicional mente tenemos que el AIC del modelo con Age como única covariable es menor que el que incluye la interacción por lo que tenemos las herramientas suficientes para descartar dicho modelo.

iii) Modelo binomial negativo, comparación e intervalo de confianza simultáneo.

Planteando un modelo binomial negativo con el código glm.nb(Cases ~ offset(logPop)+ Age , data = data4, link = "log"), tenemos el resultado del siguiente Cuadro, con un AIC de 110.45.

Como ultimo criterio para escoger modelo final compararemos los resultados tanto del AIC como del BIC para los modelos poisson con covariable edad (Modelo 2) contra el binomial negativo (Modelo 3), donde es claro que el mejor fue el poisson.

```
## [1] "AIC:(2), AIC:(3)"
```

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

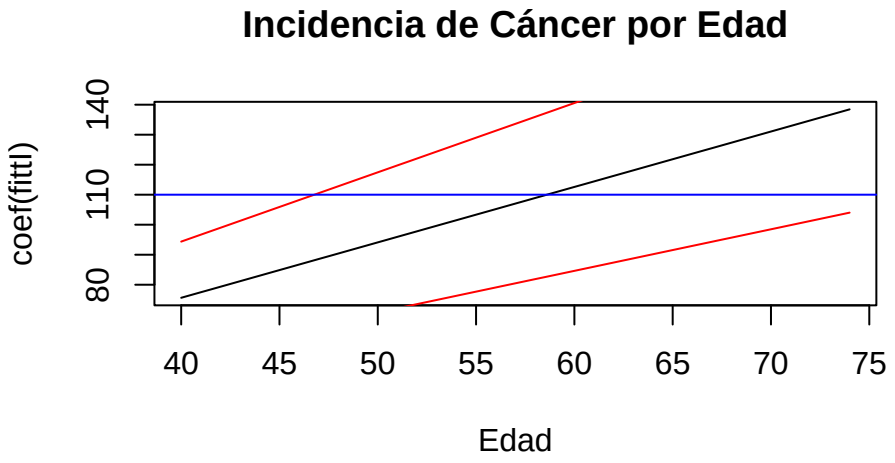
```
## [1] "BIC:(2), BIC:(3)"
```

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

Finalmente haremos intervalos de confianza simultáneos con el modelo Poisson con covariable edad, que para nosotros resultó ser el mejor de los 3.

Table 2:

	Dependent variable:
	Cases
Age55-59	1.082***
	(0.248)
Age60-64	1.502***
	(0.231)
Age65-69	1.750***
	(0.229)
Age70-74	1.847***
	(0.235)
Constant	−5.862***
	(0.174)
Observations	20
Log Likelihood	−50.226
θ	152,366.700 (s.e: 5,232,704.000)
Akaike Inf. Crit.	110.451
Note:	*p<0.1; **p<0.05; ***p<0.01




```
## integer(0)
```

De los resultados obtenidos en la gráfica con los intervalos podemos ver que aproximadamente a partir de los 57-58 años de edad la incidencia de cáncer de pulmón en las ciudades de Dinamarca va en aumento,cosa que ya se podía observar en la gráfica presentada en el primer punto del ejercicio.

Habiendo realizado todo el análisis, podemos decir que la edad juega un papel importante en el modelo y nos ayudó a ver de forma más clara y concisa que en efecto para las ciudades estudiadas por el equipo de investigadores en todas se incrementa la incidencia de cancer pulmonar conforme avanzan los años de edad.