Ejercicio 4

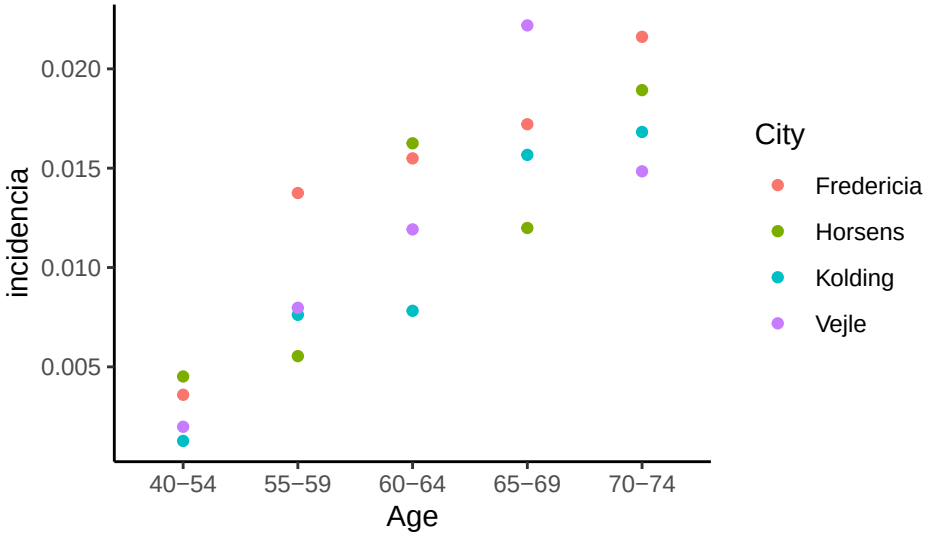
Equipo #

2024-03-31

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

```
##
## Call:
## glm(formula = Cases ~ offset(log) + Age * City, family = poisson(link = "log"),
##     data = data4)
```

```
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -5.62795    0.30151 -18.666  < 2e-16 ***
## Age55-59              1.34123    0.42640   3.145 0.001658 **
## Age60-64              1.46058    0.42640   3.425 0.000614 ***
## Age65-69              1.56578    0.43693   3.584 0.000339 ***
## Age70-74              1.79340    0.42640   4.206  2.6e-05 ***
## CityHorsens           0.22770    0.40967   0.556 0.578343
## CityKolding          -1.03837    0.58387  -1.778 0.075335 .
## CityVejle            -0.59463    0.53936  -1.102 0.270257
## Age55-59:CityHorsens -1.13671    0.65223  -1.743 0.081368 .
## Age60-64:CityHorsens -0.17991    0.57045  -0.315 0.752471
## Age65-69:CityHorsens -0.58918    0.60649  -0.971 0.331320
## Age70-74:CityHorsens -0.36029    0.58487  -0.616 0.537886
## Age55-59:CityKolding  0.44798    0.74620   0.600 0.548271
## Age60-64:CityKolding  0.35483    0.75807   0.468 0.639737
## Age65-69:CityKolding  0.94450    0.72926   1.295 0.195268
## Age70-74:CityKolding  0.78788    0.73684   1.069 0.284945
## Age55-59:CityVejle    0.04961    0.72434   0.068 0.945398
## Age60-64:CityVejle    0.33237    0.69413   0.479 0.632058
## Age65-69:CityVejle    0.84855    0.67995   1.248 0.212051
## Age70-74:CityVejle    0.21891    0.71191   0.307 0.758469
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance:  1.1543e+02  on 19  degrees of freedom
## Residual deviance: -6.6613e-16  on  0  degrees of freedom
## AIC: 121.47
##
## Number of Fisher Scoring iterations: 3
```

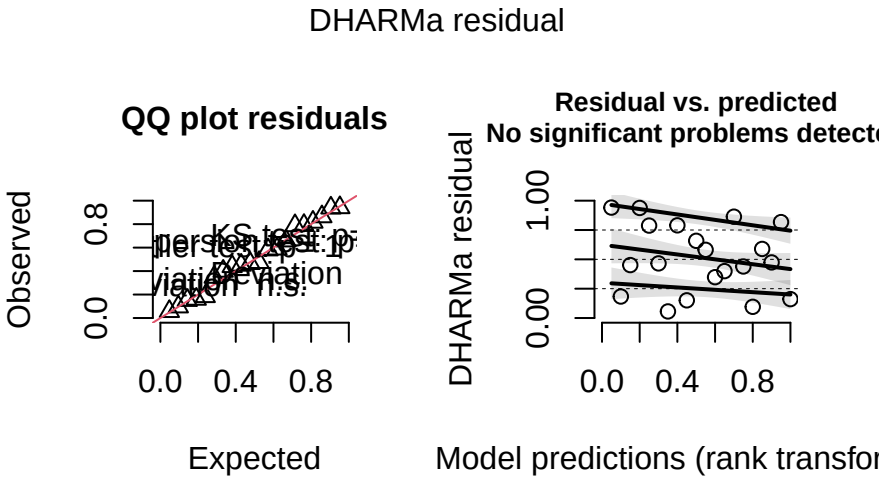
De las verificaciones de los supuestos para el primer modelo observamos que tenemos muchos problemas con los mismos del modelo por lo que optaremos por ajustar un segundo modelo, con la diferencia de que solo incluiremos a la covariable Age sin interacción.

```
##
## Call:
## glm(formula = Cases ~ offset(log) + Age, family = poisson(link = "log"),
##     data = data4)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.8623     0.1741 -33.676  < 2e-16 ***
## Age55-59      1.0823     0.2481   4.363 1.29e-05 ***
## Age60-64      1.5017     0.2314   6.489 8.65e-11 ***
## Age65-69      1.7503     0.2292   7.637 2.22e-14 ***
## Age70-74      1.8472     0.2352   7.855 4.00e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
```

```
##     Null deviance: 115.434  on 19  degrees of freedom
## Residual deviance:  16.978  on 15  degrees of freedom
## AIC: 108.45
##
## Number of Fisher Scoring iterations: 4
```

Veamos si este modelo para la regle de dedo:

```
## [1] 1.131886
```



Para este segundo modelo en el que solo se toma como covariable a la edad tuvimos un valor muy cercano a 1 con la regla del dedo, lo cual nos dice que es buen modelo, por lo que procedimos con la verificación de los supuestos del modelo con los cuales no parece tener ningún problema.

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

```
## Analysis of Deviance Table
##
## Model 1: Cases ~ offset(log) + Age * City
## Model 2: Cases ~ offset(log) + 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 con el anterior.

```
##
## Call:
## glm.nb(formula = Cases ~ offset(log) + Age, data = data4, link = "log",
##     init.theta = 152366.4999)
##
## 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(152366.5) 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:  152366
##           Std. Err.:  5232725
## Warning while fitting theta: iteration limit reached
##
##  2 x log-likelihood:  -98.451
```

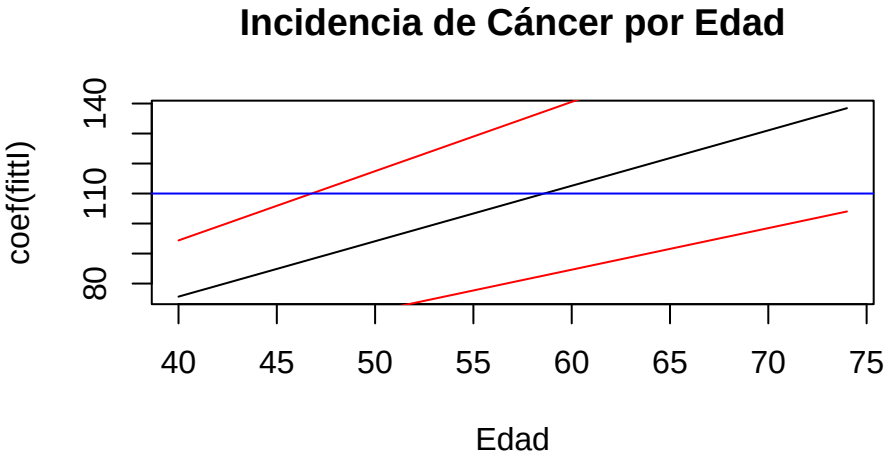
```
## [1] 108.4512 110.4515
```

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

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

Finalmente haremos intervalos de confianza simultáneos con el modelo Poisson 2 que para nosotros resulto ser el mejor de los 3.

Intervalos




```
## 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 si jugo un papel importante en el modelo y nos ayudo a ver de forma mas 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.