

Ejercicio 4

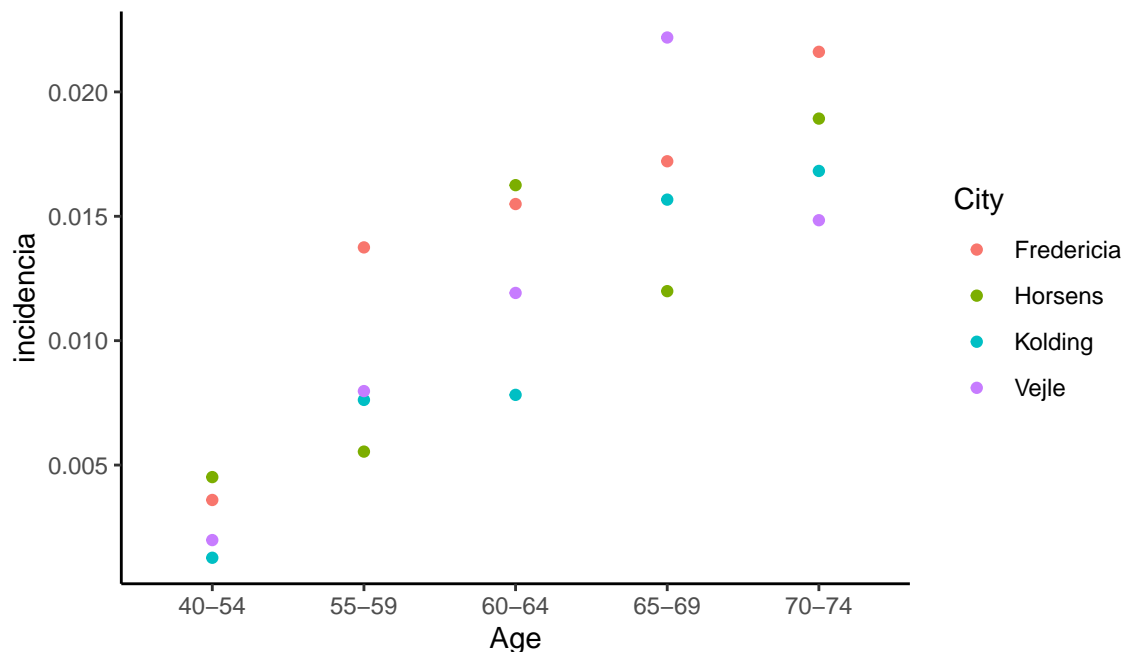
Equipo #

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, además de considerar las demás covariables de Age y City con su interacción

```
##
## Call:
## glm(formula = Cases ~ incidencia + Age * City, family = poisson(link = "log"),
##      data = data4)
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.72563    0.68127   4.001 6.31e-05 ***
## incidencia   -91.14105   138.05079  -0.660   0.5091
## Age55-59       0.92545    1.33746   0.692   0.4890
## Age60-64       1.08431    1.56838   0.691   0.4893
## Age65-69       1.14564    1.80111   0.636   0.5247
## Age70-74       1.64191    2.25757   0.727   0.4670
## CityHorsens     0.25086    0.38926   0.644   0.5193
## CityKolding    -1.22331    0.72501  -1.687   0.0915 .
## CityVejle      -0.93536    0.72268  -1.294   0.1956
## Age55-59:CityHorsens -1.60525    1.30074  -1.234   0.2172
## Age60-64:CityHorsens  0.12842    0.56588   0.227   0.8205
## Age65-69:CityHorsens -0.72673    0.93273  -0.779   0.4359
## Age70-74:CityHorsens -0.40843    0.57915  -0.705   0.4807
## Age55-59:CityKolding  0.34607    0.83602   0.414   0.6789
## Age60-64:CityKolding  0.07212    0.96506   0.075   0.9404
## Age65-69:CityKolding  1.17806    0.75562   1.559   0.1190
## Age70-74:CityKolding  0.58620    0.69637   0.842   0.3999
## Age55-59:CityVejle   -0.04318    0.62080  -0.070   0.9446
## Age60-64:CityVejle    0.51431    0.57777   0.890   0.3734
## Age65-69:CityVejle    1.72529    1.42564   1.210   0.2262
## Age70-74:CityVejle      NA         NA         NA         NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
## Null deviance:  1.7603e+01 on 19 degrees of freedom
## Residual deviance: -1.3323e-15 on  0 degrees of freedom
## AIC: 121.47
##
## Number of Fisher Scoring iterations: 3
##
## Call:
## glm(formula = Cases ~ incidencia + Age, family = poisson(link = "log"),
##      data = data4)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.9223    0.1883  10.207 < 2e-16 ***
## incidencia    64.8975   24.3660   2.663  0.00773 **
## Age55-59      -0.4294    0.2933  -1.464  0.14320
## Age60-64      -0.4053    0.3468  -1.169  0.24259
## Age65-69      -0.6185    0.4253  -1.454  0.14581
## Age70-74      -0.8045    0.4452  -1.807  0.07075 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for poisson family taken to be 1)
##
## Null deviance: 17.6029 on 19 degrees of freedom
## Residual deviance: 6.8785 on 14 degrees of freedom
## AIC: 100.35
##
## Number of Fisher Scoring iterations: 4
## Analysis of Deviance Table
##
## Model 1: Cases ~ incidencia + Age * City
## Model 2: Cases ~ incidencia + Age
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1 0 0.0000
## 2 14 6.8785 -14 -6.8785 0.9393
anova(fit_poisson2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: Cases
##
## Terms added sequentially (first to last)
##
##
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL 19 17.6029
## incidencia 1 6.7950 18 10.8079 0.009141 **
## Age 4 3.9295 14 6.8785 0.415636
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De lo que se obtuvo de los modelos diremos que si se podría utilizar el modelo 2 sin interacción de las covariables y dejando solo como covariable a la edad, sin embargo la variable edad no resulta significativa para ningún modelo, usando los AIC y la simplicidad de los modelos diremos que es mejor el segundo modelo.

iii) Ajuste un modelo binomial negativo. Comparación con el anterior.

```
fit_negbin <- glm.nb(Cases ~ incidencia + Age, data = data4, link = "log")
summary(fit_negbin)
```

```
##
## Call:
## glm.nb(formula = Cases ~ incidencia + Age, data = data4, link = "log",
## init.theta = 485152.1203)
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.9223 0.1883 10.207 < 2e-16 ***
```

```

## incidencia    64.8976    24.3663    2.663    0.00774 **
## Age55-59      -0.4294     0.2933   -1.464    0.14321
## Age60-64      -0.4053     0.3468   -1.169    0.24260
## Age65-69      -0.6185     0.4253   -1.454    0.14581
## Age70-74      -0.8045     0.4452   -1.807    0.07075 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(485152.1) family taken to be 1)
##
##      Null deviance: 17.6026  on 19  degrees of freedom
## Residual deviance:  6.8783  on 14  degrees of freedom
## AIC: 102.35
##
## Number of Fisher Scoring iterations: 1
##
##
##              Theta: 485152
##             Std. Err.: 15936689
## Warning while fitting theta: iteration limit reached
##
## 2 x log-likelihood: -88.352

```