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 cuidades 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. Presente una gráfica de dispersión en donde en el eje `x` se incluyan los grupos de edad (ordenados de menor edad a mayor) y en el eje `y` la tasa de incidencia (`Cases/Pop`) por cada cruce `Age-City`, distinguiendo con un color la Ciudad. Describa lo que se observa.

ii. Como un primer modelo considere la distribución Poisson con liga logarítmica y las covariables `Age` y `City`, así como su interacción. Dado que las dos covariables son categóricas, este modelo con interacciones tiene muchos parámetros y es deseable trabajar con uno más simple. Para esto considere un segundo modelo donde sólo se usa como covariable a `Age`. Realice una prueba de hipótesis para argumentar si es posible considerar el segundo modelo [recuerde que dado que los modelos son anidados, podría usar la función `anova(mod1, mod2, test = "Chisq")`, también puede usar `multcomp`, pero hay muchos parámetros y podría ser tedioso]. Complemente su decisión con lo que se observa en la gráfica en i) y con medidas como **AIC** o **BIC**.

iii. Considerando el modelo seleccionado en ii), ajuste un modelo binomial negativo. Compare ambos modelos e indique cuál podría ser adecuado para realizar el análisis deseado. Con el modelo seleccionado, calcule intervalos de confianza simultáneos de las tasas de incidencia para cada grupo de edad, incluya estos en la gráfica presentada en i). Comente los resultados, en particular si se puede indicar que a mayor edad existe mayor incidencia de cáncer de pulmón.

iv. Los incisos anteriores usaron a la variable `Age` como categórica, sin embargo, eso dificulta un poco la interpretación, además de que por su naturaleza esa variable se podría haber registrado sin categorizar. Con los datos actuales, una aproximación sería usar el punto medio de cada intervalo de edad que define las categorías de `Age` y usar la `variable resultante` como una variable continua, llámela `Ageprima`. Ajuste modelos usando la distribución *Poisson* y *Binomial Negativa* con la covariable `Ageprima`, también considere la opción de incluir a `Ageprima2`. Entre esos 4 modelos indique cuál podría ser adecuado para realizar el análisis. Con ese modelo indique si a mayor edad existe mayor incidencia de cáncer de pulmón, por ejemplo, argumentando si la función es creciente considerando que el intervalo de edad que es de interés es entre 40 y 74 años. Presente una gráfica que complemente su análisis.