1. Consideremos los datos **warpbreaks** de datasets (hacer require(datasets)). En este conjunto de datos se registran los resultados de un experimento realizado para determinar el efecto del tipo de lana (A o B) y la tensión (baja, media o alta) en el número de roturas de deformación en la fabricación de telas. Se recopilaron datos de nueve telares para cada combinación de configuraciones. Las variables consideradas son **breaks** (número de roturas), **wool** (tipo de lana), y **tension** (tensión de la lana). La pregunta de interés es si resulta posible predecir el número de roturas en función del tipo de lana y de la tensión.

   a) Para cada una de las categorías de **tension**, realizar boxplots paralelos para el logaritmo de la variable **breaks** clasificando por tipo de **wool** (realizar los 6 boxplots paralelos, se sugiere hacer los de lana tipo A de un color y los de tipo B de otro para facilitar la visualización) .¿Qué sugiere? ¿Parece haber interacción?

   b) Realizar un scatterplot donde gráfica los promedios del logaritmo de **breaks** vs. **tension** para cada tipo de **wool**. Interprete. ¿Parece haber interacción?

   c) Ajustar un modelo de regresión para predecir el número de roturas usando como variables explicativas a **wool** y **tension**. ¿Incluye la interacción entre ambas? ¿Qué distribución propone para la variable de respuesta?

   d) A partir del ítem anterior: ¿cuánto valen las estimaciones de los coeficientes? ¿Cuáles de estos coeficientes son significativos?

   e) A partir del ajuste anterior, ¿Para que combinación de lana y tensión el modelo predice el mayor número de roturas? ¿Y el menor?

   f) Calcular un intervalo asintótico de nivel 0.95 para el valor esperado de roturas para una lana de tipo A y tensión L. Idem para una lana de tipo B y tensión H. ¿Son iguales las longitudes de los intervalos?

   g) ¿Cómo se calcula la deviance para la familia de distribución de la respuesta que propuso? ¿Cuánto vale en este caso?

   h) Realizar una tabla de ANOVA para el modelo ajustado. Interpretar.

2. *British doctor's smoking and coronary death*. Los datos de la siguiente tabla corresponden a un estudio muy famoso realizado por Sir Richard Doll y colegas. En 1951, los médicos británicos recibieron una breve encuesta acerca de si fumaban tabaco o no. Después de esa fecha se registró la causa de muerte y edad de cada uno de ellos. La

tabla muestra la información recabada 10 años después de la encuesta y el tamaño de la población en el correspondiente grupo etáreo.

Las preguntas de interés son: i) ¿es la tasa de muerte mayor para fumadores que para no fumadores? ii) Si es así, ¿por cuanto? iii) ¿Hay un efecto debido a la edad?

Age	Smoke	Deaths	Person-years
35-44	1	32	52407
45-54	1	104	43248
55-64	1	206	28612
65-74	1	186	12663
75-84	1	102	5317
35-44	0	2	18790
45-54	0	12	10673
55-64	0	28	5710
65-74	0	28	2585
75-84	0	31	1462

Distintos modelos pueden ajustarse a estos datos. Consideramos el modelo dado por:

$$\log(death_i) = \log(population_i) + \beta_1 + \beta_2 smoke_i + \beta_3 agecat_i + \beta_4 agesq_i + \beta_5 smoke * agecat_i$$

donde i representa al índice del i-ésimo grupo formado por la combinación de grupo etáreo y condición de fumador (1 a 5 para edades 35-44,..., 75-84) y grupos 6 a 10 para los correspondientes grupos etáreos en no fumadores, $death_i$ es el número esperado de muertes en el grupo i y $population_i$ es el número de de médicos en riesgo en el grupo i. Por otra parte, $smoke_i$ vale 1 en los fumadores y 0 en los no fumadores y $agecat_i$ toma los valores 1 a 5 para los grupos de edades de 35-44,..., 75-84, $agesq_i$ es su cuadrado y $smoke * agecat_i$ representa la interacción entre *smoke* y *agecat*, valiendo 1 a 5 en los fumadores y 0 en los no fumadores.

a) ¿Como se interpreta el término $\log(population_i)$ en este modelo? Note que no está acompañado por un coeficiente β en el modelo propuesto. ¿Cómo se incluye en la función `glm` de `R`?

b) Realizar un scatterplot de Edad vs. la tasa de muerte cada 100000 habitantes (Death/ Person-years * 100000). ¿Qué sugiere este gráfico?

c) Ajustar el modelo indicado, verificar la significación de los coeficientes y computar un intervalo de confianza de nivel aproximado 0.95 para cada uno de ellos.

d) En un modelo lineal generalizado en el que para cada observación se expresa

$$E(Y_i) = \mu_i = n_i e^{\mathbf{x}_i^t \boldsymbol{\beta}} \quad Y_i \sim Poisson(\mu_i)$$

para una covariable binaria x_j que corresponde a una variable indicadora que vale $x_j = 1$ si el factor está presente y $x_j = 0$ si está ausente, la **rate ratio** o tasa cociente, RR, de presencia vs ausencia se define como

$$RR = \frac{E(Y_i|presencia)}{E(Y_i|ausencia)} = e^{\beta_j}$$

a condición de que todas las otras variables se mantengan iguales.

En este sentido, el simple hecho de fumar afecta la tasa de muerte y tiene un efecto basal

$$e^{\beta_2},$$

al que se le debe sumar un efecto de la edad debido a la interacción.

Estimar esta cantidad e^{β_2} y hallar un intervalo de confianza de nivel aproximado 0.95 para la misma. Intepretar.

Observación: Dado que el modelo ajustado tiene una interacción ente *smoke* y el grupo etáreo, la RR depende de la edad y en este caso sería:

$$RR = \frac{E(Y_i|smoke = si)}{E(Y_i|smoke = no)} = e^{\beta_2 + \beta_5 agecat_i}$$

3. **Sobredispersión.** Cuando trabajamos en regresión Poisson la sobredispersión es siempre un problema potencial...

   Consideremos el conjunto de datos del archivo postdoc.txt (Allison, 1999). Los datos corresponden 557 hombres doctorados en bioquímica en 106 universidades norteamericanas entre 1950 y 1960. El siguiente cuadro describe las variables presentes.

Variable	**Nombre**
Edad a la que se doctoró	AGE
1 si está casado, 0 si no	MAR
Prestigio de la universidad donde hizo el Ph.D.	DOC
Medida de la selectividad de la institución de grado	UND
Número de citas a artículos publicados	CITS
Número de artículos publicados	ARTS

   Nos interesa estudiar la relación entre CITS y las demás variables.

a) Realice un ajuste para CITS considerando en su modelo las variables AGE, MAR, DOC, UND y una ordenada al origen. Analice la significación de cada una de las variables al 5 %.

b) Compare la deviance y el estadístico de Pearson con sus grados de libertad tomando el cociente. ¿Le parece que el ajuste es bueno? ¿Le parece que puede haber sobredispersión?

c) Haga un summary de los residuos deviance y de Pearson.

d) Lleve a cabo un nuevo ajuste en el que contemple un coeficiente de escala distinto de 1 mediante una distribución Binomial Negativa utilizando el comando glm.bn de la librería MASS. Analice la significación de cada una de las variables en este nuevo ajuste. Compare la deviance y el estadístico de Pearson con sus grados de libertad tomando el cociente. Compare todos estos resultados con los obtenidos en los ítems anteriores. ¿Le parece que el ajsute es mejor que el anterior?

e) Haga un summary de los residuos deviance y de Pearson de este nuevo modelo.

f) Con esta nueva perspectiva, mire los resultados del Ejercicio 1 y evalúe los resultados que obtuvo.... rehaga el ajuste si le parece necesario. Compare.