Ejercicios para practicar: modelo de conteo

El conjunto de datos "DatosGSS_poisson" contiene información sobre algunas cuestiones sociodemográficas de 2035 encuestados por el *GSSSurvey* (el equivalente estadounidense al CIS). Las variables contenidas en el fichero son:

Variable	Descripción	Codificación
ClaseSocial	Clase social autopercibida	
Felicidad	Respuesta a la pregunta: "¿Cómo de feliz se siente usted?"	Not too happy = No muy feliz Pretty happy = Bastante feliz Very happy = Muy feliz
Politica	Identificación política	
Tamano	Tamaño (en miles de habitantes) del municipio de residencia	
Region	Región de residencia	
Ingreso	Ingresos familiares anuales	0 = Ninguno; 1 = < 25.000$; 2 = ≥ 25.000$
Raza		
Genero		
Edad		
Hijos Var. Depend.	¿Cuántos hijos tiene?	
EstadoCivil		
Empleo	Ocupación actual	
Zodiaco	Signo del zodiaco (12 niveles)	

Los ejercicios consisten en lo siguiente (recuerda que debes llevarlos a cabo sobre un archivo rmd para posteriormente generar el html y subirlo al campus virtual):

1. Carga los datos en el entorno de *Rstudio* a través de la función *readRDS* (nota: como estás trabajando con datos "conocidos" no es necesario revisar el tipo de las variables, pues lo hiciste el primer día, pero se debe hacer siempre que el conjunto de datos es nuevo).
2. En estos ejercicios vamos a trabajar sobre la variable "Hijos". Indica de qué tipo es y obtén un gráfico apropiado para la misma. Comenta dicho gráfico.
3. Realiza una partición del conjunto de datos en entrenamiento (80%) y prueba (20%).
4. Utilizando los datos de la partición de entrenamiento, genera un primer modelo de regresión poisson para la variable dependiente "Hijos" con todas las variables independientes disponibles. ¿Cuántos parámetros lo componen? ¿Se podría saber si estos son significativos solo con la salida anterior?
5. Estudia la posible existencia de sobredispersión y saca una conclusión al respecto. Explica teóricamente por qué es importante verificarlo.
6. Construye ahora un modelo de conteo Binomial negativo para la misma variable dependiente con todas las variables independientes disponibles. Comenta las diferencias teóricas que existen entre este modelo y el anterior. A continuación, utilizando la herramienta que consideres oportuna, determina cuál de los dos modelos es preferible para estos datos.
7. Aplica el análisis de tipo II sobre el tipo de modelo determinado en el apartado anterior. A continuación, analiza los resultados y extrae las conclusiones pertinentes.
8. Construye un nuevo modelo (que llamaremos modelo2) que contenga únicamente las 3 variables más importantes según el análisis anterior. ¿Cuántos parámetros tiene? ¿Todas las variables del modelo son significativas ahora?
9. De nuevo sobre los datos de entrenamiento, construye 2 modelos del tipo que hayas determinado en el apartado 6 aplicando uno de los métodos de selección de variables estudiados y los 2 criterios de selección a partir de la función *step*. ¿Los modelos son diferentes? ¿Qué variables y cuántos parámetros contienen cada uno de estos modelos? (NOTA: sólo pido que generéis dos modelos para no alargar los ejercicios, pero lo recomendable sería generar siempre los 6).
10. Una vez generados los modelos (el manual con las 4 variables más importantes y los dos automáticos), debes determinar cuál es el mejor a partir de validación cruzada repetida (utiliza un bucle para simplificar esta tarea). Genera los *boxplots* correspondientes, así como los resúmenes y compara los modelos. ¿Cuál parece ser el mejor?
11. Interpreta todos los parámetros del modelo "ganador" (si lo necesitas usa un nivel de significación del 5%) y, usando el código que consideres, crea un ranking de las variables explicativas por importancia.
12. Evalúa dicho modelo tanto en los datos de entrenamiento como en los de prueba. ¿Qué puedes decir de la calidad del modelo?