Ejercicios para practicar: logística ordinal

El conjunto de datos "DatosGSS" 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 Var. Depend.	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	¿El encuestado tiene hijos?	0 = No; 1 = Sí
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 ordinal "Felicidad". Verifica que el orden de los niveles de la variable están ordenados de menor a mayor nivel de felicidad y que dicha variable está bien formateada. De no ser así, soluciona los problemas que haya.
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 logística ordinal para la variable dependiente "Felicidad" con todas las variables independientes disponibles. ¿Cuántos parámetros β lo componen? ¿Cuántos son significativos al 5%?
5. Aplica el análisis de tipo II sobre el modelo anterior, explica en qué consiste este análisis y de qué sirve. A continuación, analiza los resultados y extrae las conclusiones pertinentes.
6. Con la información de los ejercicios anteriores (puede que necesites ejecutar algo más de código), responde a las siguientes preguntas, justificando tu respuesta (de nuevo, usa un nivel de significación del 5%):
    a. ¿La edad de los individuos influye en su nivel de felicidad? Si es así, construye una frase que permita cuantificar esa influencia.
    b. ¿La raza de los individuos influye en su nivel de felicidad? Si es así, construye una frase que permita cuantificar esa influencia.
    c. ¿El género de los individuos influye en su nivel de felicidad? Si es así, construye una frase que permita cuantificar esa influencia.
7. Construye un nuevo modelo (que llamaremos modelo2) que contenga únicamente aquellas variables significativas al 5%. ¿Cuántos parámetros tiene? ¿Todas las variables del modelo son significativas ahora?
8. De nuevo sobre los datos de entrenamiento, construye 2 modelos de regresión logística ordinal 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 los 6).
9. Una vez generados los modelos (el manual con las variables significativas al 5% 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 para las 4 medidas, así como los resúmenes y compara los modelos. ¿Cuál parece ser el mejor? Recuerda que si varios modelos son parecidos en cuanto a capacidad predictiva se debe escoger el más sencillo (el que tenga menos parámetros).
10. Para el mejor modelo obtén la matriz de confusión, la tasa de acierto y los índices Kappa en el conjunto de datos de prueba. ¿Qué puedes decir de la calidad del modelo? ¿Te sorprende algún resultado comparando con los resultados de validación cruzada?