Ejercicios para practicar: logística multinomial

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	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 Var. Depend.	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 r*md* 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). Formatea la variable dependiente si fuera necesario. Decide si quieres modificar la categoría de referencia o no.
2. Realiza una partición del conjunto de datos en entrenamiento (80%) y prueba (20%).
3. Utilizando los datos de la partición de entrenamiento, genera un primer modelo de regresión logística multinomial para la variable dependiente con todas las variables independientes disponibles. ¿Cuántos parámetros lo componen? ¿Cuántos son significativos al 5%?
4. 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.
5. Con la información de los ejercicios anteriores, responde a las siguientes preguntas, justificando tu respuesta (de nuevo, usa un nivel de significación del 5%):
    a. ¿El tamaño del municipio de residencia influye en la identificación política de los encuestados? De ser así, ¿influye de igual manera (es positivo/negativo/no influyente) en las 2 comparativas que contempla el modelo?
    b. ¿La edad influye en la identificación política de los encuestados? De ser así, ¿influye de igual manera (es positivo/negativo/no influyente) en las 2 comparativas que contempla el modelo?
    c. ¿El género influye en la identificación política de los encuestados? De ser así, ¿influye de igual manera (es positivo/negativo/no influyente) en las 2 comparativas que contempla el modelo?
6. Utilizando los datos de la partición de entrenamiento, construye 6 modelos de regresión logística multinomial para la variable dependiente "Politica" aplicando los 3 métodos de selección de variables estudiados y los 2 criterios de selección a partir de la función *step*. ¿Cuántos modelos diferentes se generan? ¿Cuántos parámetros tienen?
7. Una vez generados los modelos, se debe determinar cuál es el mejor de todos ellos, para lo cual debes aplicar validación cruzada repetida (utiliza un bucle para simplificar esta tarea).
   Genera los boxplots para las 3 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).
8. Para el mejor modelo obtén la matriz de confusión y sus medidas derivadas en el conjunto de datos de prueba. ¿Qué puedes decir de la calidad del modelo? Explica por qué aparecen varios valores de sensibilidad y especificidad.
9. ¿Qué puedes decir sobre la calidad del modelo a partir de la información proporcionada la curva Roc one-vs-all (tanto en entrenamiento como en prueba)? A la vista de los resultados, ¿podría decirse que el modelo funciona igual de bien a la hora de predecir los 3 niveles de la variable?
10. Para finalizar, realiza un análisis de tipo II sobre el modelo, saca las conclusiones oportunas e interpreta los *odds-ratio* asociados a dos variables (si es posible, una cuantitativa y una cualitativa). Recuerda que si algún parámetro no es significativo la frase correspondiente debe reflejar ese hecho.