Ejercicios para practicar: logística binaria I

El conjunto "DatosGSS" contiene información sobre cuestiones sociodemográficas de 2035 encuestados por el *GSSSurvey* (un organismo de investigación sociológica). Las variables 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.	¿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 qmd 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*. Haz un *summary* de los datos para entender bien qué significan y verificar que las variables no tengan errores.
2. Revisa los niveles que toma la variable dependiente y haz las modificaciones necesarias.
3. Haz una partición entrenamiento-prueba de los datos.
4. Genera un primer modelo de regresión logística binario para la variable "Hijos" con todas las variables independientes disponibles y los datos de entrenamiento. ¿Cuántos parámetros tiene el modelo? ¿Todos los parámetros son significativos al 5%? Indica cuáles no lo son.
5. Con la información del modelo, responde a las siguientes preguntas justificando la respuesta:
    a. ¿El tamaño del municipio en el que reside el encuestado influye en el hecho de que tenga hijos?
    b. ¿Se puede afirmar que los encuestados independientes tienen la misma tendencia a tener hijos que los demócratas?
    c. ¿Se puede afirmar que los encuestados republicanos tienen la misma tendencia a tener hijos que los demócratas?
    d. ¿La felicidad de los encuestados influye en el hecho de que tengan hijos?
    e. ¿El género de los encuestados influye en el hecho de que tengan hijos?
6. 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.
7. Construye un nuevo modelo (que llamaremos modelo2) que contenga únicamente las 3 variables más importantes (usa la información del ejercicio anterior para saber cuáles son). ¿Cuántos parámetros tiene? ¿Este nuevo modelo tiene todos sus parámetros significativos?
8. Obtén los ODDS-ratio e interprétalos (haz una frase completa para cada uno de ellos que pudiera comprender alguien con escasos conocimientos de estadística). Recuerda que si algún parámetro no es significativo la frase correspondiente debe reflejar ese hecho.
9. Utilizando los datos de la partición de entrenamiento,
    a. Obtén la matriz de confusión para el punto de corte de 0.5, así como la tasa de acierto, el índice Kappa, la sensibilidad y la especificidad y explica qué significan.
    b. Si modificamos el punto de corte por la proporción de eventos (codificados como X1) que hay en los datos, ¿cómo se modifican las medidas del apartado anterior? ¿Tiene sentido?
    c. ¿Qué puedes decir sobre la calidad del modelo a partir de la información proporcionada por la curva ROC?
10. Repite el ejercicio anterior utilizando los datos de la partición de prueba, haciendo hincapié en la comparación con los resultados anteriores.
11. En líneas generales, ¿qué puedes comentar de la calidad del modelo y de su estabilidad? Nota: no necesitas ejecutar más código para responder esta pregunta.