RELACIÓN DE PROBLEMAS: REGRESIÓN LOGÍSTICA Y MULTINOMIAL
ANÁLISIS ESTADÍSTICO MULTIVARIANTE
GRADO EN CIENCIA E INGENIERÍA DE DATOS

1. En la siguiente tabla se ha recopilado una serie de 20 datos que relacionan las horas de estudio de cada alumno y si han aprobado o suspendido un examen de estadística.

Horas de estudio	Aprobado (1 sí, 0 no)	Horas de estudio	Aprobado (1 sí, 0 no)
0.5	0	2.75	1
0.75	0	3	0
1	0	3.25	1
1.25	0	3.5	0
1.5	0	4	1
1.75	0	4.25	1
1.75	1	4.5	1
2	0	4.75	1
2.25	1	5	1
2.5	0	5.5	1

Se ha ajustado un modelo de regresión logística y los parámetros estimados han sido $\hat{\theta}_0 = -4.077$ y $\hat{\theta}_1 = 1.5046$.

   a) ¿Cómo se interpreta el valor de $\hat{\theta}_1$?

   b) A partir del modelo ajustado, obtener una predicción para la probabilidad de que un alumno apruebe si ha estudiado 2 horas. ¿Cuál sería dicha probabilidad si dedicara una hora más al estudio? ¿Cómo ha variado la razón de estas probabilidades?

2. En un estudio clínico se desea predecir la probabilidad de padecer una enfermedad coronaria (Y, con valores 1 sí, 0 no) a partir de las covariables siguientes: Nivel de colesterol (X_1, 1 alto, 0 bajo), Edad (X_2) y Resultado del electrocardiograma (X_3, 1 anormal, 0 normal). Para ello, se analizaron 750 casos y se propuso un modelo logístico para estimar el riesgo de padecer una enfermedad coronaria, obteniendo las siguientes estimaciones para los parámetros: $\hat{\theta}_0 = -3.912$, $\hat{\theta}_1 = 0.852$, $\hat{\theta}_2 = 0.025$ y $\hat{\theta}_3 = 0.441$.

   a) Interpretar el significado de los coeficientes $\hat{\theta}_1$, $\hat{\theta}_2$ y $\hat{\theta}_3$.

   b) Obtener una predicción para la probabilidad de padecer una enfermedad coronaria para una persona con 40 años, electrocardiograma normal y nivel de colesteroal bajo. ¿Cuál sería dicha probabilidad si tuviera un nivel de colesterol alto?

   c) Para una persona con 40 años y electrocardiograma normal, ¿cómo influye el nivel de colesterol en el riesgo de padecer una enfermedad coronaria?

3. Se desea evaluar la satisfacción con la enseñanza pública de 1,500 estudiantes mediante la variable *Satisfecho* (Y, con valores 1 sí, 0 no) y tres variables predictoras: Nacionalidad (España=1, Ecuador = 2, Colombia=3), Género (Hombre=0, Mujer=1) y Estudios (ESO=0, Primaria=1). Se ajusta el siguiente modelo logístico:

$$\log \frac{p}{1-p} = -0.877 - 0.052\text{Nacionalidad2} + 1.72\text{Nacionalidad3} + 0.256\text{Género} - 0.008\text{Estudios},$$

donde $p = Pr[Y = 1]$. Las variables Nacionalidad2 y Nacionalidad3 son variables dicotómicas ficticias (*dummy*) que toman el valor 1 si el valor de Nacionalidad se corresponde con su índice y valen cero en caso contrario. Por ejemplo, si Nacionalidad = 3, entonces Nacionalidad2 = 0 y Nacionalidad3 = 1.

   a) Predecir la probabilidad de que una alumna colombiana de primaria no esté satisfecha con la enseñanza pública. ¿Cuál sería dicha probabilidad si la alumna tuviera nacionalidad española y estudiara primaria? ¿Y si fuera un alumno de secundaria con nacionalidad española?

   b) Comparar el grado de satisfacción con la enseñanza pública de alumnos de primaria con nacionalidad española según el género.

4. Una determinada compañía desea mejorar el marketing de cinco variedades de cereales para el desayuno. Para ello planifica un estudio encuestando a 900 personas, registrando su edad, género y si tiene o no un estilo de vida activo. Cada participante degustó los cinco tipos de cereales y se le preguntó sobre su preferencia. En ta tabla adjunta se presentan las definiciones de las variables.

Variable	Categoría
Tipo de cereal preferido	1: Cebada
	2: Centeno
	3: Avena
	4: Esbelta
	5: Trigo
Edad	1: Menor de 30 años
	2: 31 a 50 años
	3: Más de 50 años
Género	1: Hombre
	2: Mujer
Estilo de vida	0: No activo
(realiza o no actividad física)	1: Activo

   a) Con el objetivo de explicar la preferencia del tipo de cereal en función de la edad, el género y el estilo de vida se desea ajustar un modelo logístico multinomial. ¿Qué variables ficticias (*dummy*) debemos crear para la formulación del modelo?

   b) Especificar el modelo tomando el trigo como la categoría de referencia para la variable respuesta ($g = 5$).

   c) ¿Cómo se interpretan las estimaciones de los coeficientes del modelo $\hat{\theta}_{ij}$ que verifican que $\exp(\hat{\theta}_{ij}) > 1$? ¿Y si $\exp(\hat{\theta}_{ij}) < 1$?