Modelo de Parcial Integrador

Fundamentos de Ciencia de Datos

> **Descripción del dataset**
>
> El dataset `nacimientos.csv` contiene la siguiente información registrada en un efector de salud sobre un total de 345 madres que dieron a luz en el último año: peso del recién nacido (kg), semanas de gestación, número de partos previos atravesados y edad de la madre al momento del nacimiento. Por otro lado, se registró el peso (kg) y la altura (m) de las mujeres luego del parto y se las indagó acerca del número de cigarrillos que suelen consumir a diario, información que se encuentra en el dataset `datos_madres.xlsx`.

PARTE I

Ejercicio 1. Importe ambos datasets al entorno de trabajo y realice cualquier tarea de limpieza y adecuación de los mismos que considere necesaria para su posterior análisis.

Ejercicio 2. Considerando que el Índice de Masa Corporal (IMC) se define como **el peso de una persona en kilogramos dividido por el cuadrado de la estatura en metros**, represente gráficamente la distribución de dicha variable para las mujeres del dataset. En base al gráfico realizado, ¿cómo caracterizaría su distribución en relación a la simetría?

Ejercicio 3. Genere una variable categórica binaria realizando la *dicotomización* de la variable vinculada con el número de cigarrillos consumidos a diario, diferenciando aquellas mujeres que no fuman habitualmente (no consumen cigarrillos) de aquellas que sí lo hacen (consumen 1 o más cigarrillos diariamente). ¿Qué porcentaje de las mujeres del dataset son fumadoras?

Ejercicio 4. Represente gráficamente la distribución del peso de los recién nacidos en función del carácter o no de fumadora de la madre (variable generada en el ítem anterior). Comente brevemente lo observado.

Ejercicio 5. Realice un gráfico que le permita caracterizar el grado de asociación lineal que existe entre las diferentes variables cuantitativas estudiadas. ¿Cuáles son los pares de variables que presentan una asociación lineal más intensa? Justifique.

PARTE II

El objetivo principal de esta segunda parte es analizar si es factible ajustar un modelo de regresión lineal múltiple que permita predecir el peso de los recién nacidos en función de las semanas de gestación, el número de partos previos atravesados por la madre, la edad de la madre, su índice de masa corporal (IMC) y el carácter o no de fumadora.

a. Ajuste el modelo completo, incluyendo la totalidad de las variables predictoras de interés. ¿Cuál/es de las variables incluidas contribuye/n significativamente a explicar las diferencias en el peso promedio de los recién nacidos con un nivel de significación del 1%? Justifique.

b. Si a partir de su respuesta en el ítem anterior considera la posibilidad de ajustar un nuevo modelo que incluya menos variables predictoras que el modelo completo, realice el ajuste. De lo contrario, pase al ítem siguiente.

c. Escriba la ecuación del modelo ajustado en forma desarrollada e interprete el valor del coeficiente estimado por el modelo para la variable **semanas de gestación.**

d. Una médica quiere estimar el peso al nacer que tendrá un bebé que dará a luz una madre primeriza no fumadora de 25 años, que posee un índice de masa corporal de 19.5 kg/m^2, luego de un total de 38 semanas de gestación. Utilice el último modelo ajustado para informarle un peso estimado.

e. Represente gráficamente la distribución de los residuos del modelo y comente brevemente las características que observa.