# Departamento de Estadística y Matemáticas
# Facultad de Ciencias Económicas
# Estadística II
# Parcial IV

Nombre: _____ Cédula:_____

1. **(5 puntos)** El Ministerio de Salud de Colombia ha decidido realizar un estudio epidemiológico a nivel nacional para evaluar la prevalencia de enfermedades cardiovasculares en la población, y por ello, ha decidido recopilar la información de una muestra representativa personas de diferentes regiones, edades, género y condiciones socioeconómicas.

   La base de datos contiene los siguientes campos:

   - `Region`: Región del paciente, donde 'Andina', 'Caribe', 'Pacífica', 'Orinoquía' y 'Amazonía', representan la región de Colombia donde vive el paciente.
   - `Edad`: Edad del paciente en años.
   - `Altura`: Altura del paciente en centímetros.
   - `Peso`: Peso del paciente en kilogramos.
   - `Genero`: Género del paciente, donde 'Hombre' representa un paciente de género masculino y 'Mujer' representa un paciente de género femenino.
   - `Niv_Soc`: Nivel socioeconómico del paciente, donde 'Bajo', 'Medio' y 'Alto' son los niveles bajo los cuales se cataloga la capacidad adquisitiva y estado de la vivienda donde vive el paciente.
   - `Pre_S`: Presión arterial sistólica del paciente en mmHg.
   - `Pre_D`: Presión arterial diastólica del paciente en mmHg.
   - `Col`: Nivel de colesterol del paciente, donde 'normal' representa un nivel normal de colesterol, 'por encima de lo normal' representa un nivel de colesterol por encima de lo normal y 'muy por encima de lo normal' representa un nivel de colesterol muy por encima de lo normal.
   - `Col_sangre`: Nivel de colesterol del paciente en la sangre en mg/dL.
   - `Glu`: Nivel de glucosa del paciente, donde 'normal' representa un nivel normal de glucosa, 'por encima de lo normal' representa un nivel de glucosa por encima de lo normal y 'muy por encima de lo normal' representa un nivel de glucosa muy por encima de lo normal.
   - `Az_sangre`: Nivel de glucosa del paciente en forma numérica en mg/dL.
   - `Fumador`: Si el paciente es fumador o no, donde 'No' representa que el paciente no es fumador y 'Sí' representa que el paciente es fumador.
   - `Con_Alc`: Si el paciente consume alcohol o no, donde 'No' representa que el paciente no consume alcohol y 'Sí' representa que el paciente consume alcohol.
   - `Enf_Car`: Si el paciente tiene enfermedad cardiovascular o no, donde 'No' representa que el paciente no tiene enfermedad cardiovascular y 'Sí' representa que el paciente tiene enfermedad cardiovascular.
   - `IMC`: Índice de masa corporal del paciente, calculado como el peso en kilogramos dividido por el cuadrado de la altura en metros.
   - `Niv_Estres`: Nivel de estrés del paciente, donde 'Bajo' representa un nivel bajo de estrés, 'Moderado' representa un nivel moderado de estrés y 'Alto' representa un nivel alto de estrés.
   - `Con_Tabaco`: Frecuencia de consumo de tabaco del paciente, donde 'Nunca' representa que el paciente nunca consume tabaco, 'Ocasional' representa que el paciente consume tabaco ocasionalmente y 'Diario' representa que el paciente consume tabaco a diario.

- `Dieta`: Tipo de dieta del paciente, donde 'Saludable' representa una dieta saludable y 'No saludable' representa una dieta no saludable.
- `Act_Social`: Nivel de actividad social del paciente, donde 'Baja' representa un nivel bajo de actividad social, 'Moderada' representa un nivel moderado de actividad social y 'Alta' representa un nivel alto de actividad social.
- `Niv_Act`: Nivel de actividad física del paciente, donde 'Bajo' representa un nivel bajo de actividad física, 'Moderado' representa un nivel moderado de actividad física y 'Alto' representa un nivel alto de actividad física.
- `Con_Sal`: Cantidad en mg de sal del paciente.
- `Con_Azucar`: Cantidad en mg de consumo de azúcar del paciente.
- `Frec_Card`: Frecuencia cardíaca del paciente en latidos por minuto.
- `Con_FV`: Número de veces que consume de frutas y verduras al día.
- `Horas_Sueño`: Horas de sueño del paciente al día.
- `Horas_Ejercicio`: Horas de ejercicio que realiza el paciente al día.
- `Con_Agua`: Consumo de agua del paciente en litros diarios.

El objetivo del estudio es analizar los factores de riesgo asociados con las enfermedades cardiovasculares, como la edad, el género, el estilo de vida (fumador, obesidad), los niveles de presión arterial, colesterol y diabetes. Además, se busca evaluar la prevalencia de estas enfermedades en diferentes regiones y grupos socioeconómicos.

Basado en la información contenida en la base de datos suministrada, al crear una subpoblación de la base de datos mediante la aplicación de **al menos 2 condiconales** (**Aplique a la base de datos al menos 2 filtros, antes de trabajar con ella el parcial**) .

*a*) **(0.5 puntos)** Si tuviera que plantear una relación lineal entre la variable `Pre_D` y otra variable de la forma

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

cómo plantearía dicha relación desde sus racionalidad?. Seleccione **tres posibles variables explicativas** $X$ y explique el por qué estas variables podrían explicar el comportamiento de la variable respuesta $Y$.

*b*) **(0.5 puntos)** Basado en el planteamiento que realizó en el inciso anterior, justifique cómo esperaría usted que fuese el signo de los parámetros $\beta_0$ y $\beta_1$ del modelo, al usar de forma individual cada una de las **tres variables explicativas** $X$ propuestas.

*c*) **(0.5 puntos)** Realice el cálculo de los estimadores para los parámetros $\beta_0$ y $\beta_1$, para cada una de las **tres relaciones propuestas** e interprete éstos en el contexto de los datos. Los resultados obtenidos fueron consistentes con lo que esperaba en el inciso anterior?.

*d*) **(0.5 puntos)** Pruebe cada una de las **tres relaciones propuestas** la significancia estadística de los parámetros $\beta_0$ y $\beta_1$, empleando para ello un nivel de significancia del 5 %, e interprete en el contexto de los datos.

*e*) **(0.5 puntos)** Para cada una de las **tres relaciones propuestas**, construya intervalos de confianza para los parámetros $\beta_0$ y $\beta_1$, empleando para ello un nivel de confianza del 95 %, e interprete en el contexto de los datos.

*f*) **(0.5 puntos)** Dados los resultados obtenidos hasta el momento, explique cuál de las tres variable explicativa parece ser la más adecuada para describir la relación con la variable `Pre_D`?.

**NOTA**:A partir de este punto los análisis solo se tienen que realizar con la variable explicativa que haya considerado como la más adecuada para describir la relación con la variable respuesta `Pre_D`.

*g*) **(0.5 puntos)** Verifique si se cumplen o no los supuestos del modelo de regresión lineal planteado, e interprete los resultados.

*h*) **(0.5 puntos)** Pruebe la significancia estadística de la regresión lineal planteada, empleando para ello un nivel de significancia del 5 %, e interprete en el contexto de los datos.

*i*) **(0.5 puntos)** Realice el cálculo del coeficiente de determinación $R^2 = \frac{SCR}{SCT}$ asociado a la regresión lineal planteada, e interprete el resultado obtenido. Dicho resultado es consistente con lo que se concluyó en el inciso anterior?

*j*) **(0.5 puntos)** Seleccione tres valores para $x_0$ entre los posibles que considera que puede tomar la variable que escogió como $X$, y con éste, construya un intervalo de predicción del 95 % para la variable que escogió como $Y$ e interprete en el contexto de los datos.