CASO REGRESION LINEAL MULTIPLE

Tipo	Clase
Revisado	☑

1. **ANÁLISIS DE RELACIÓN ENTRE VARIABLES**

```
                  colesterol pa_sistolica    IMC   edad cant_cigarillos glucosa problemas_cardiacos
colesterol             1.000        0.032 -0.033 -0.019          -0.001  -0.017                0.418
pa_sistolica           0.032        1.000  0.000 -0.011          -0.041  -0.073                0.150
IMC                   -0.033        0.000  1.000 -0.054           0.040   0.055                0.313
edad                  -0.019       -0.011 -0.054  1.000           0.001  -0.041               -0.051
cant_cigarillos       -0.001       -0.041  0.040  0.001           1.000   0.011                0.659
glucosa               -0.017       -0.073  0.055 -0.041           0.011   1.000                0.294
problemas_cardiacos    0.418        0.150  0.313 -0.051           0.659   0.294                1.000
```

La tabla que has compartido muestra una matriz de correlación de Pearson entre varias variables que estás considerando para tu análisis de regresión lineal múltiple, donde el objetivo es predecir `problemas_cardiacos`.

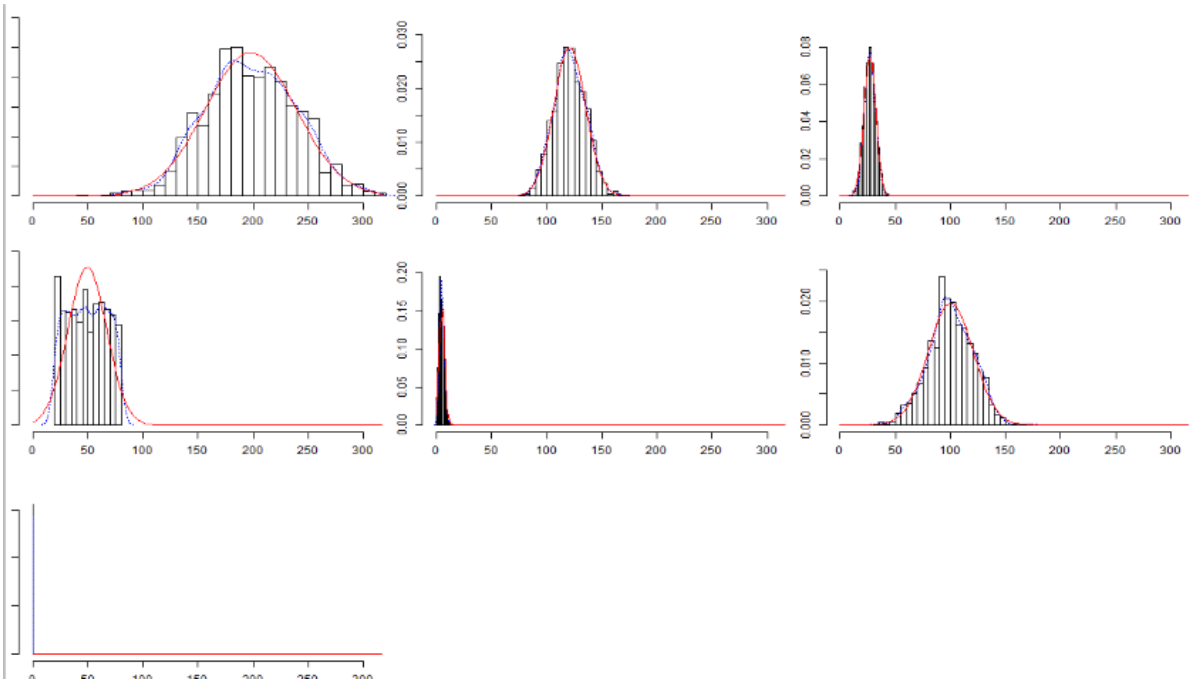
1. **Colesterol**: La correlación entre colesterol y problemas cardíacos es de 0.418, lo que indica una correlación moderada y positiva. Esto sugiere que, a medida que aumenta el colesterol, también es probable que aumente el riesgo de problemas cardíacos.
2. **Presión Arterial Sistólica (pa_sistolica)**: La correlación entre la presión sistólica y problemas cardíacos es de 0.150, lo que indica una correlación positiva, aunque baja. Esto sugiere que la presión arterial sistólica podría tener una relación positiva con los problemas cardíacos, pero no es un factor tan fuerte en comparación con otras variables.
3. **IMC (Índice de Masa Corporal)**: La correlación entre IMC y problemas cardíacos es de 0.313, lo que indica una correlación positiva moderada. Un

IMC más alto podría estar asociado con un aumento en el riesgo de problemas cardíacos.

4. **Edad**: La correlación entre edad y problemas cardíacos es de -0.051, indicando una correlación muy baja y negativa. Esto sugiere que la edad, en este caso, no parece estar fuertemente relacionada con los problemas cardíacos en este conjunto de datos.
5. **Cantidad de Cigarillos (cant_cigarillos)**: La correlación entre el consumo de cigarrillos y problemas cardíacos es de 0.659, lo que indica una correlación fuerte y positiva. Esto sugiere que fumar cigarrillos es un factor importante y está fuertemente asociado con un mayor riesgo de problemas cardíacos.
6. **Glucosa**: La correlación entre glucosa y problemas cardíacos es de 0.294, lo que indica una correlación positiva moderada. Esto sugiere que niveles más altos de glucosa pueden estar asociados con un mayor riesgo de problemas cardíacos.

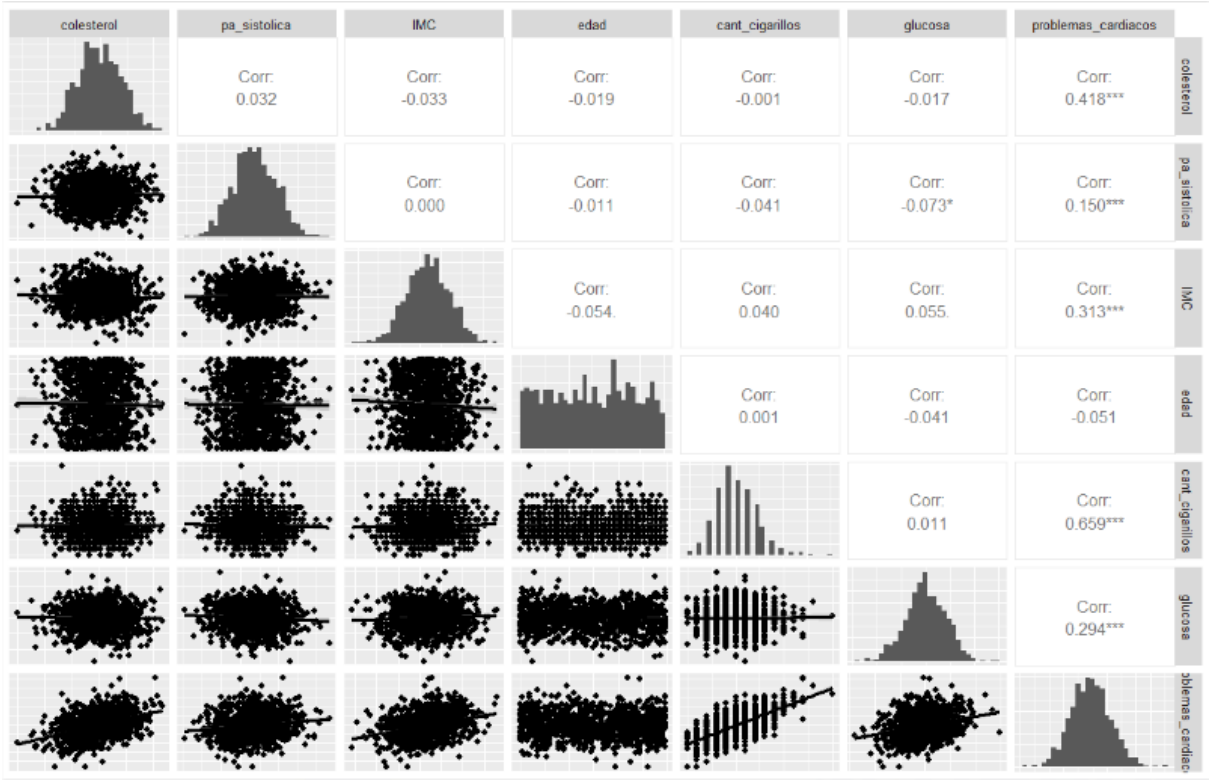
Resumen Interpretativo

En este conjunto de datos, las variables con mayor correlación positiva con `problemas_cardiacos` son **cant_cigarillos** (0.659) y **colesterol** (0.418). Estas variables podrían ser predictores importantes en tu modelo de regresión para predecir problemas cardíacos. Otros factores, como el IMC y la glucosa, también muestran una correlación positiva moderada y podrían contribuir al modelo, aunque no tan fuertemente como el consumo de cigarrillos.



Este tipo de gráfico permite visualizar la **distribución de cada variable**

Cada gráfico muestra la **distribución de frecuencias** de una variable en forma de histograma, con la línea roja sólida que representa el ajuste a una distribución normal y la línea azul punteada para otra posible distribución
Puedes observar que algunas variables siguen una **distribución aproximadamente normal** (campana), mientras que otras son muy **asimétricas** o tienen distribuciones **sesgadas** hacia un lado.



Este gráfico es una matriz de dispersión con histogramas y coeficientes de correlación, que permite observar visualmente las relaciones entre las variables y su distribución individual

Distribución de cada Variable

En la diagonal del gráfico (celdas de la esquina superior izquierda a la esquina inferior derecha) puedes ver histogramas que muestran la distribución de cada variable:

- **Colesterol, IMC, Glucosa y Cantidad de Cigarrillos**: Estas variables muestran distribuciones que parecen algo sesgadas, especialmente el colesterol y la cantidad de cigarrillos, lo que podría indicar una mayor frecuencia de valores en el extremo inferior.
- **Edad y Problemas Cardíacos**: La edad tiene una distribución más uniforme, mientras que los problemas cardíacos parecen estar más concentrados en ciertos valores.
- **Presión Sistólica**: Muestra una distribución aproximadamente simétrica, aunque hay una ligera concentración de valores en torno a un punto medio.

Relación entre las Variables

En la parte inferior izquierda de la matriz, se presentan gráficos de dispersión entre pares de variables. Aquí puedes ver la relación entre cada par de variables:

1. **Colesterol y Problemas Cardíacos**: Hay una ligera tendencia positiva, lo cual es consistente con la correlación positiva (0.418). Esto indica que los niveles más altos de colesterol tienden a asociarse con una mayor probabilidad de problemas cardíacos.
2. **Cantidad de Cigarrillos y Problemas Cardíacos**: La dispersión muestra una relación positiva fuerte, con una correlación de 0.659, lo que indica que el aumento en el consumo de cigarrillos está asociado con un mayor riesgo de problemas cardíacos.
3. **IMC y Problemas Cardíacos**: Existe una relación positiva moderada (0.313) en la dispersión, lo que sugiere que un IMC más alto también se asocia con un aumento en el riesgo de problemas cardíacos.
4. **Glucosa y Problemas Cardíacos**: Hay una relación positiva moderada (0.294). Esto indica que a mayor nivel de glucosa, puede haber una mayor probabilidad de problemas cardíacos.
5. **Presión Sistólica y Problemas Cardíacos**: Aunque la correlación es baja (0.150), hay una ligera tendencia positiva, que sugiere que la presión sistólica podría estar débilmente asociada con los problemas cardíacos.

2. GENERAR MODELO

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.104841 -0.020249 -0.000987  0.019504  0.104561

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.440e-01  1.210e-02 -20.165   <2e-16 ***
colesterol       6.993e-04  2.268e-05  30.833   <2e-16 ***
pa_sistolica     8.630e-04  6.501e-05  13.275   <2e-16 ***
IMC              3.752e-03  1.834e-04  20.461   <2e-16 ***
edad            -5.517e-05  5.346e-05  -1.032    0.302
cant_cigarillos  2.013e-02  4.265e-04  47.198   <2e-16 ***
glucosa          9.750e-04  4.642e-05  21.002   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02958 on 993 degrees of freedom
Multiple R-squared:  0.8108,    Adjusted R-squared:  0.8097
F-statistic: 709.4 on 6 and 993 DF,  p-value: < 2.2e-16
```

Residuos

- **Rango de Residuos**: Los valores de los residuos van desde -0.1048 hasta 0.1046 aproximadamente, con una mediana cercana a 0. Esto indica que, en promedio, el modelo no tiene un sesgo fuerte en los errores.
- **Interpretación**: Los residuos indican qué tan lejos están los valores predichos de los valores reales. Dado que los residuos son bastante pequeños, sugiere que el modelo ajusta razonablemente bien los datos.

Coeficientes

Cada coeficiente representa el cambio estimado en la variable dependiente `problemas_cardiacos` por cada unidad adicional de la variable independiente, manteniendo constantes las demás variables.

1. **(Intercepto)**: -0.244. Representa el valor esperado de `problemas_cardiacos` cuando todas las variables independientes son 0. Este valor puede no tener una interpretación práctica si no es realista que todas las variables sean 0 simultáneamente.
2. **Colesterol**: 0.000693. Un aumento de una unidad en el colesterol está asociado con un aumento de aproximadamente 0.000693 en `problemas_cardiacos`, manteniendo las demás variables constantes. Dado el valor de p (<2e-16), este coeficiente es estadísticamente significativo.

3. **PA Sistólica**: 0.000863. Aumento de una unidad en la presión sistólica está asociado con un incremento de 0.000863 en `problemas_cardiacos`, manteniendo constantes las demás variables. Este coeficiente también es estadísticamente significativo.
4. **IMC**: 0.003752. Aumento de una unidad en el IMC está asociado con un incremento de 0.003752 en `problemas_cardiacos`. Esto sugiere que el IMC tiene un efecto positivo y significativo en la variable de resultado.
5. **Edad**: -0.000055. Este coeficiente es negativo pero no es estadísticamente significativo (p-valor = 0.302), por lo que no hay evidencia suficiente para concluir que la edad tenga un efecto importante en `problemas_cardiacos` en este modelo.
6. **Cantidad de Cigarrillos**: 0.02013. Cada unidad adicional en la cantidad de cigarrillos está asociada con un aumento de 0.02013 en `problemas_cardiacos`. Dado el valor de p muy bajo (<2e-16), este es un predictor significativo y tiene un impacto positivo fuerte.
7. **Glucosa**: 0.00975. Un aumento de una unidad en la glucosa está asociado con un incremento de 0.00975 en `problemas_cardiacos`. También es estadísticamente significativo.

Evaluación del Modelo

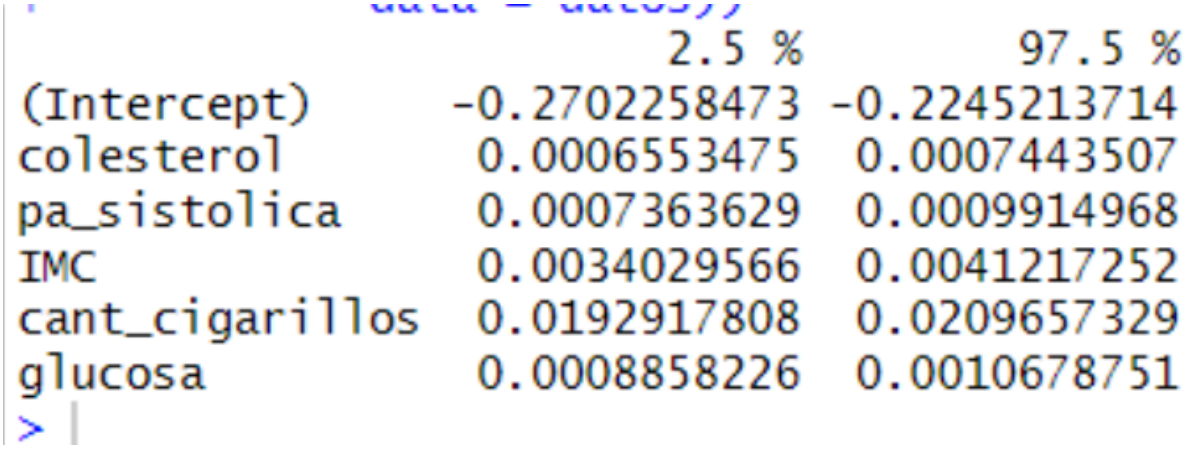
- **Error estándar residual**: 0.02958. Indica la desviación estándar de los residuos, es decir, qué tan lejos están los valores observados de los valores predichos. Un valor bajo sugiere un buen ajuste del modelo.
- **R-cuadrado (Multiple R-squared)**: 0.8108. Indica que el modelo explica aproximadamente el 81.08% de la variabilidad en `problemas_cardiacos`. Esto sugiere que el modelo tiene un buen nivel de ajuste, ya que explica una gran proporción de la variabilidad de los datos.
- **R-cuadrado ajustado**: 0.8097. Muy similar al R-cuadrado, lo cual indica que el modelo es estable y no se ve afectado significativamente por el número de variables.
- **Estadístico F**: 709.4 con un valor p < 2.2e-16. Esto indica que el modelo en su conjunto es estadísticamente significativo, es decir, al menos una de las variables independientes tiene un efecto significativo en la variable dependiente.

3. **INTERVALO DE CONFIANZA**

```
Step:  AIC=-7035.36
problemas_cardiacos ~ colesterol + pa_sistolica + IMC + cant_cigarillos +
    glucosa

                Df Sum of Sq     RSS     AIC
<none>                       0.86970 -7035.4
+ edad           1   0.00093 0.86877 -7034.4
- pa_sistolica   1   0.15453 1.02424 -6873.8
- IMC            1   0.36926 1.23897 -6683.5
- glucosa        1   0.38803 1.25773 -6668.4
- colesterol     1   0.83329 1.70300 -6365.4
- cant_cigarillos 1   1.94871 2.81842 -5861.6

Call:
lm(formula = problemas_cardiacos ~ colesterol + pa_sistolica +
    IMC + cant_cigarillos + glucosa, data = datos)

Coefficients:
    (Intercept)       colesterol     pa_sistolica              IMC  cant_cigarillos          glucosa
     -0.2473736        0.0006998        0.0008639        0.0037623        0.0201288        0.0009768
```

Exclusión de Variables:

- Para cada variable existente, se muestra el AIC resultante si esa variable fuera excluida del modelo:
  - **"pa_sistolica"**: Si se elimina, el AIC aumenta a -6873.8, indicando una disminución en la calidad del modelo.
  - **"IMC"**: Eliminarlo aumenta el AIC a -6833.5.
  - **"glucosa"**: Eliminarlo lleva el AIC a -6688.4.
  - **"colesterol"**: Lleva el AIC a -6365.4.
  - **"cant_cigarillos"**: Aumenta el AIC a -5861.6.
- En general, la eliminación de cualquiera de las variables aumenta el AIC, lo cual sugiere que todas las variables actuales aportan información relevante al modelo.



1. En esta imagen se muestran los intervalos de confianza al 95% para los coeficientes de cada variable independiente en tu modelo de regresión lineal múltiple.
2. **Intercepto**:
    - El intervalo de confianza para el intercepto va de -0.2702 a -0.2245. Esto sugiere que, cuando todas las variables independientes están en cero, el valor esperado de la variable dependiente (problemas cardíacos) es negativo en este rango. Sin embargo, el intercepto no suele tener una interpretación práctica en modelos donde las variables independientes no son cero en condiciones reales.
3. **Colesterol**:
    - Intervalo: 0.0006553 a 0.0007444.
    - El intervalo de confianza es positivo y muy estrecho, lo cual indica que hay una relación positiva y consistente entre los niveles de colesterol y los problemas cardíacos en el modelo. Esto significa que, al aumentar el colesterol, el riesgo de problemas cardíacos también aumenta, aunque en una magnitud pequeña.
4. **Presión Arterial Sistólica (pa_sistolica)**:
    - Intervalo: 0.0007363 a 0.0009915.
    - También tiene un intervalo positivo, lo que indica una relación positiva entre la presión arterial sistólica y los problemas cardíacos. Esto significa que a mayor presión arterial, hay un mayor riesgo de problemas cardíacos en el modelo.
5. **Índice de Masa Corporal (IMC)**:

- Intervalo: 0.0034 a 0.0041.
- Este intervalo también es positivo, lo cual sugiere que el IMC tiene una influencia positiva en los problemas cardíacos. Es decir, un IMC más alto está asociado con un mayor riesgo de problemas cardíacos en el modelo.

6. **Cantidad de Cigarrillos (cant_cigarillos)**:
    - Intervalo: 0.0193 a 0.0210.
    - Este intervalo es positivo y relativamente más amplio que los otros, indicando que fumar está positivamente asociado con problemas cardíacos. A mayor cantidad de cigarrillos fumados, el riesgo de problemas cardíacos aumenta de manera notable.
7. **Glucosa**:
    - Intervalo: 0.0008858 a 0.0010679.
    - El intervalo de confianza es positivo y pequeño, lo cual sugiere que un nivel de glucosa más alto se asocia con un mayor riesgo de problemas cardíacos, aunque de manera leve en comparación con otras variables.

Conclusiones Generales

- Todos los intervalos de confianza de los coeficientes son positivos y no incluyen el cero, lo cual indica que todas las variables independientes en el modelo tienen un efecto significativo y positivo sobre la variable dependiente, "problemas cardíacos".