Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Guadalajara

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

M1.4 Regresión Lineal Múltiple

Samuel Padilla Esqueda | A01641383

Septiembre 23

Realizar las transformaciones adecuadas a las variables predictoras.

	Factor Coagulación	Índice pronóstico	Función de enzima	\
0	0.476744	0.593407	0.604167	
1	0.290698	0.560440	0.447917	
2	0.558140	0.538462	0.625000	
3	0.453488	0.714286	0.187500	
4	0.604651	0.626374	0.958333	

	Función de hígado	Edad	Género	Alcohol\n(moderado)	Alcohol\n(severo)
0	0.326855	0.500	0.0	1.0	0.0
1	0.169611	0.225	0.0	0.0	0.0
2	0.250883	0.625	0.0	0.0	0.0
3	0.224382	0.450	0.0	0.0	0.0
4	0.628975	0.375	0.0	0.0	1.0

Usamos la técnica de normalización de datos Min-Max para hacer que todos los datos se encuentren en la misma escala de entre 0 y 1, con esto evitamos que haya problemas relacionados a la escala al aplicar la regresión lineal múltiple.

Realizar el modelo de regresión con las variables significativas.

En Python utilice el código de regresión lineal múltiple que se nos proporciono en la clase de aprendizaje automático para crear el modelo. Se obtienen los siguientes coeficientes y valores.

Model coefficients: [-575.86374747 453.26085291 738.23244281 854.14304366 429.43164879 25.63892773 13.09258075 -41.26764482 195.70703222]

MSE: 26938.257179330714

MAE: 118.6071846460898

R^2: 0.775278095890658

Probar si se deben agregar interacciones o términos polinomiales.

Interpretar la tabla ANOVA, R2, R2 ajustada, p-values y FIV.

1. R^2: 0.775

Esto indica que el 77.5% de la variabilidad en la variable dependiente ("Sobrevivencia (días)") es explicada por las variables independientes en el modelo. Es un valor bastante alto, lo que sugiere que el modelo es bastante efectivo para explicar la variabilidad en la supervivencia.

2. R^2 Ajustada: 0.757

La R^2 ajustada es ligeramente más baja que la R^2, lo que es normal, ya que R^2 ajustada penaliza la inclusión de variables innecesarias. Una R^2 ajustada de 0.757 sigue indicando que el modelo explica bien la variabilidad, pero sugiere que algunas variables podrían no estar contribuyendo significativamente.

3. p-values

Los p-values indican la significancia estadística de cada coeficiente en el modelo. Un p-value menor a 0.05 sugiere que la variable es significativa.

- **Factor Coagulación:** p-value = 0.000 (Significativa)
- **Índice pronóstico:** p-value = 0.000 (Significativa)
- **Función de enzima:** p-value = 0.000 (Significativa)
- **Función de hígado:** p-value = 0.004 (Significativa)
- **Edad:** p-value = 0.659 (No significativa)
- **Género:** p-value = 0.702 (No significativa)
- **Alcohol (moderado):** p-value = 0.287 (No significativa)
- **Alcohol (severo):** p-value = 0.000 (Significativa)

Interpretación de p-values:

- Las variables "Factor Coagulación", "Índice pronóstico", "Función de enzima", "Función de hígado" y "Alcohol (severo)" son estadísticamente significativas en el modelo.
- Las variables "Edad", "Género" y "Alcohol (moderado)" no son estadísticamente significativas, lo que sugiere que podrían no tener un impacto importante en la supervivencia.

5. FIV

Los valores de FIV nos dicen cuánto la varianza de un coeficiente se incrementa debido a la colinealidad con otras variables.

- **Edad:** 1.020800 (Bajo)
- **Género:** 1.068036 (Bajo)
- **Alcohol (moderado):** 1.363438 (Bajo)
- **Alcohol (severo):** 1.444596 (Bajo)

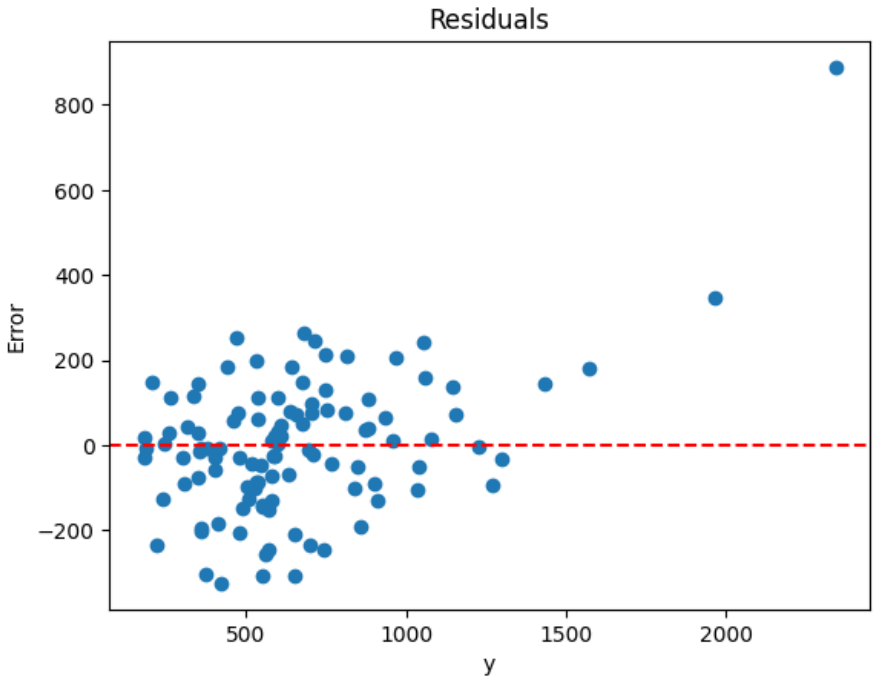
- **Factor Coagulación:** 1.577947 (Moderado)
- **Índice pronóstico:** 1.323686 (Moderado)
- **Función de enzima:** 1.629249 (Moderado)
- **Función de hígado:** 2.401849 (Moderado)

Interpretación de FIV:

Todos los valores de FIV son menores a 5, lo que sugiere que no hay un problema grave de multicolinealidad en el modelo. No es necesario eliminar ninguna variable en función del FIV.

Verificar el cumplimiento de los supuestos.

Homocedasticidad y linealidad



La homocedasticidad se cumple en el modelo de regresión lineal, ya que en la gráfica de homocedasticidad no se observa un patrón definido. Los puntos se distribuyen de manera aleatoria, esto es un buen indicativo de que el modelo es adecuado.

Autocorrelación

Utilizando la prueba Durbin-Watson se obtiene el valor 1.77, los resultados posibles de la prueba Durbin-Watson es de entre 0 y 4, un valor cercano a 0 indica

correlación positiva, un valor cercano a 4 indica correlación negativa y un valor cercano a 2 indica que no hay correlación. Podemos ver que nuestro valor es mas cercano a 0 y podemos interpretar que hay poco o nula autocorrelación entre nuestras variables.

Multicolinealidad

De acuerdo con el VIF, nuestros resultados se encuentran en un rango de $1 < VIF < 5$, esto indica que hay una correlación moderada, pero no preocupante. Este rango es generalmente aceptable y sugiere que no hay multicolinealidad significativa.