Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Guadalajara

Inteligencia artificial avanzada para la ciencia de datos I

M1.4 Regresión Lineal Múltiple

Paola Félix Torres | A00227869

03/09/2024

Datos “Cirugía de Hígado y Sobrevivencia”

- **Realizar las transformaciones adecuadas a las variables predictoras.**

Factor Coagulación	Índice pronóstico	Función de enzima	Función de hígado	Edad	Género	Alcohol (moderado)	Alcohol (severo)	Sobrevivencia (días)
6.7	62	81	2.59	50	0	1	0	695
5.1	59	66	1.7	39	0	0	0	403
7.4	57	83	2.16	55	0	0	0	710
6.5	73	41	2.01	48	0	0	0	349
7.8	65	115	4.3	45	0	0	1	2343

Inicialmente, se contaba con datos en diferentes escalas. Como se puede ver en la tabla de arriba, tres de las variables predictoras tenían un rango entre 0 y 1, mientras que las demás tenían escalas variadas. Por lo tanto, era necesario normalizar estas variables para que estuvieran en un rango común. A continuación, se muestran las variables ya normalizadas.

Factor Coagulación	Índice pronóstico	Función de enzima	Función de hígado	Edad	Género	Alcohol (moderado)	Alcohol (severo)	Sobrevivencia (días)
0.476744186	0.593406593	0.604166667	0.326855124	0.5	0	1	0	695
0.290697674	0.56043956	0.447916667	0.169611307	0.225	0	0	0	403
0.558139535	0.538461538	0.625	0.250883392	0.625	0	0	0	710
0.453488372	0.714285714	0.1875	0.224381625	0.45	0	0	0	349
0.604651163	0.626373626	0.958333333	0.628975265	0.375	0	0	1	2343

- **Realizar el modelo de regresión con las variables significativas.**

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	8	10037033	1254629	42.69	0.000
Factor Coagulación	1	382582	382582	13.02	0.000
Índice pronóstico	1	1502570	1502570	51.13	0.000
Función de enzima	1	2276466	2276466	77.46	0.000
Función de hígado	1	259327	259327	8.82	0.004
Edad	1	5769	5769	0.20	0.659
Género	1	4327	4327	0.15	0.702
Alcohol Moderado	1	33713	33713	1.15	0.287
Alcohol Severo	1	448520	448520	15.26	0.000
Error	99	2909332	29387		
Total	107	12946365			

Al obtener el modelo de regresión, podemos identificar cuáles son las variables significativas analizando el análisis de varianza y fijándonos en el valor p. Los coeficientes con un valor p menor a 0.05 serán nuestras variables significativas. Esto significa que podemos realizar el modelo de regresión con las variables de Factor Coagulación, Índice Pronóstico, Función de Enzima, Función de Hígado y Alcohol Severo. A continuación se presenta la ecuación del modelo de regresión con las variables significativas:

Regression Equation

Sobrevivencia = -574.6 + 432 Factor Coagulación + 731 Índice pronóstico
+ 847.6 Función de enzima + 452 Función de hígado + 221.3 Alcohol Severo

- **Probar si se deben agregar interacciones o términos polinomiales.**

Al agregar interacciones y términos polinomiales, se elaboró un modelo de regresión con las variables significativas y se obtuvo una ecuación de regresión como la siguiente:

Regression Equation

Sobrevivencia = -461 + 648 Factor Coagulación + 1287 Índice pronóstico
- 138 Función de enzima + 160 Función de hígado - 133 Alcohol Severo
+ 29 Factor Coagulación ^2 - 342 Índice pronóstico ^2
+ 1061 Función de enzima ^2 + 491 Función de hígado ^2
- 689 FactorCoagulaciónxÍndicepronóst - 212 FunciónxdeenzimaxFunciónxdehígado
+ 853 AlcoholSeveroxFactorCoagulación

Sin embargo, con la ecuación por sí sola no podemos determinar si agregar interacciones o términos polinomiales es más óptimo que no agregarlos. A continuación, se presenta el resumen del modelo. Como se puede ver en la imagen de abajo, el modelo tiene una R cuadrada de 82.87%.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
152.773	82.87%	80.71%	71.63%

A continuación, tenemos el resumen del modelo antes de agregar las interacciones y términos polinomiales. Como se puede apreciar, la R cuadrada es de 77.18%. Por lo tanto, podemos concluir que agregar interacciones y términos polinomiales proporciona un modelo algo más preciso. Además, es posible que añadiendo más interacciones y términos polinomiales podamos lograr un modelo aún más preciso.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
170.178	77.18%	76.06%	73.20%

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

En la tabla ANOVA podemos obtener datos importantes sobre nuestro modelo. Por ejemplo, podemos ver los p-values de cada predictor. Un p-value menor a 0.05 indica que la variable es significativa en nuestro modelo. Observamos que algunas variables son significativas, mientras que otras no lo son tanto y podrían ser eliminadas de la regresión para mejorar las predicciones.

En la tabla también se encuentra el F-value. Cuanto mayor sea el F-value, mayor es el impacto de la variable sobre la respuesta. Por lo tanto, podemos concluir que Índice Pronóstico y Función de Enzima al Cuadrado tienen un fuerte impacto en el modelo.

Asimismo, el Adj SS (Suma de Cuadrados Ajustada) mide la cantidad de variación explicada por el modelo. Un valor alto de Adj SS, como se observa en la variable Índice Pronóstico, indica que es una variable significativa, como se mencionó anteriormente.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	12	10729112	894093	38.31	0.000
Factor Coagulación	1	41047	41047	1.76	0.188
Índice pronóstico	1	204515	204515	8.76	0.004
Función de enzima	1	5967	5967	0.26	0.614
Función de hígado	1	5425	5425	0.23	0.631
Alcohol Severo	1	33384	33384	1.43	0.235
Factor Coagulación ^2	1	124	124	0.01	0.942
Índice pronóstico ^2	1	26888	26888	1.15	0.286
Función de enzima ^2	1	224968	224968	9.64	0.003
Función de hígado ^2	1	25118	25118	1.08	0.302
FactorCoagulaciónxÍndicepronóst	1	37106	37106	1.59	0.210
FuncióndeenzimaxFunciónde hígado	1	2231	2231	0.10	0.758
AlcoholSeveroxFactorCoagulación	1	267116	267116	11.44	0.001
Error	95	2217253	23340		
Total	107	12946365			

En el resumen del modelo, podemos observar datos como R cuadrada y R cuadrada ajustada. Una R cuadrada de 82.87% indica que el 82.87% de la variabilidad de “Sobrevivencia” puede ser explicada por las variables predictoras del modelo. Por otro lado, una R cuadrada ajustada de 80.71% significa que, después de ajustar por el número de variables predictoras en el modelo, el 80.71% de la variabilidad de “Sobrevivencia” sigue siendo explicada por las variables predictoras.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
152.773	82.87%	80.71%	71.63%

VIF mide la multicolinealidad entre los predictores. Un VIF alto indica que la variable está altamente correlacionada con otras variables del modelo, lo que puede inflar los errores estándar y hacer que los coeficientes no sean confiables. En la tabla de coeficientes, podemos ver que variables como Factor Coagulación e Índice Pronóstico tienen un VIF alto, lo cual puede estar afectando la estimación de los coeficientes. Por otro lado, variables como Alcohol Severo tienen un VIF más bajo, lo que significa que no presentan problemas significativos de multicolinealidad.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-461	207	-2.22	0.029	
Factor Coagulación	648	488	1.33	0.188	30.01
Índice pronóstico	1287	435	2.96	0.004	29.57
Función de enzima	-138	274	-0.51	0.614	16.33
Función de hígado	160	332	0.48	0.631	16.00
Alcohol Severo	-133	111	-1.20	0.235	8.92
Factor Coagulación ^2	29	395	0.07	0.942	16.33
Índice pronóstico ^2	-342	318	-1.07	0.286	21.97
Función de enzima ^2	1061	342	3.10	0.003	27.46
Función de hígado ^2	491	473	1.04	0.302	24.07
FactorCoagulaciónxÍndicepronóst	-689	546	-1.26	0.210	22.04
FuncióndeenzimaxFunciódehígado	-212	684	-0.31	0.758	50.68
AlcoholSeveroxFactorCoagulación	853	252	3.38	0.001	9.99

- **Verificar el cumplimiento de los supuestos.**

Como se puede observar en las gráficas de abajo, tenemos cuatro gráficas que nos ayudarán a verificar el cumplimiento de los supuestos, que son normalidad, homocedasticidad e independencia. La normalidad se puede verificar a través del análisis de las gráficas de “Normal Probability Plot” e “Histogram”. En la primera, podemos observar que la mayoría de los puntos en el gráfico siguen la línea diagonal, mientras que en la segunda, vemos que la distribución es mayormente simétrica. Ambas observaciones nos llevan a la conclusión de que se sigue una distribución normal, cumpliendo así el supuesto de normalidad.

Por otro lado, la homocedasticidad se puede observar en la gráfica de “Versus Fits”, donde se nota una dispersión en los residuos, lo que es suficiente para afirmar que se cumple con este supuesto.

Finalmente, la independencia se puede observar en la gráfica de “Versus Order”, donde claramente no se observa un patrón en los residuos, lo que indica que no

existe autocorrelación y, por lo tanto, los residuos son independientes.