Exploración de modelos

En este proyecto se realizó un algoritmo para encontrar el mejor modelo que represente adecuadamente nuestros datos. Se utilizaron regresión múltiple, regresión ponderada y modelos lineales generalizados (normal, gamma, gaussiana inversa), donde cada variable fue elevada a una potencia. Posteriormente, cada modelo fue evaluado utilizando la métrica AIC y se ordenaron de menor a mayor AIC para identificar el modelo con el valor mínimo y evaluar si cumplía con los supuestos.

La base de datos utilizada en este proyecto contiene información sobre 438 pacientes seleccionados de forma aleatoria. Con el modelo seleccionado, se busca analizar si existe una asociación entre la presión arterial sistólica (`bpsystol`) y el índice de masa corporal (`bmi`). En particular, se quiere observar si un índice de masa corporal elevado se asocia con una presión arterial sistólica alta.

Explicación del algoritmo

En el algoritmo empleado, definimos una función que generaba una **malla** (con todos los valores y combinaciones posibles) para **var_1** y **var_2**. Esta se creó a partir de una secuencia de [0, 5] con incrementos de 0.3. Luego, se construyó la columna **Num_Ponderada**, en la cual se definía qué variable iba a tomar el peso en nuestra regresión ponderada. El peso se especificaba en la columna **Ponderada**.

Para los modelos lineales generalizados, se utilizaron las columnas **Num_Ponderada** y **GLM**.

Primeros datos

var_1	var_2	Num_Ponderada	GLM	liga	Ponderada
0.0	0	1	0	identity	0
0.3	0	1	0	identity	0
0.6	0	1	0	identity	0
0.9	0	1	0	identity	0
1.2	0	1	0	identity	0
1.5	0	1	0	identity	0

Ejemplo de cómo se veía el data frame **malla**, con los primeros 10 elementos. Este data frame contenía: 157216 combinaciones.

Posteriormente, se realizó un filtrado a nuestra **malla** para saber qué tipo de modelo se iba a evaluar con cada fila. Este proceso ayudó a reducir el número de modelos a evaluar, ya que eliminaba elementos repetidos de nuestra **malla**. Los criterios para esta elección fueron los siguientes:

- Si las columnas **GLM** y **Ponderada** eran iguales a 0, entonces se aplicaba un modelo de regresión múltiple, donde:

$$\mathbb{E}[\text{bpsystol}] = \text{bmi}^{\text{var\_1}} + \text{age}^{\text{var\_2}} + \text{sex}$$

- Si la columna **Ponderada** era distinta de 0, se aplicaba regresión ponderada. Con **Num_Ponderada** se elegía a qué variable (**bmi** o **age**) se le asignaría el peso:

- Si **Num_Ponderada** era igual a 1, entonces:

$$\mathbb{E}[\text{bpsystol}] = \text{bmi}^{\text{var\_1}} + \text{age}^{\text{var\_2}} + \text{sex}, \quad \text{weights} = \frac{1}{\text{bmi}^{\text{Ponderada}}}$$

- En caso contrario:

$$\mathbb{E}[\text{bpsystol}] = \text{bmi}^{\text{var\_1}} + \text{age}^{\text{var\_2}} + \text{sex}, \quad \text{weights} = \frac{1}{\text{age}^{\text{Ponderada}}}$$

- Por último, si **GLM** era distinto de 0, se aplicaba un modelo GLM, donde:

$$g\left(\mathbb{E}[\text{bpsystol}]\right) = \text{bmi}^{\text{var\_1}} + \text{age}^{\text{var\_2}} + \text{sex}$$

Con todo esto, se evaluó un total de: 13005 modelos, de los cuales se evaluaron:

- 289 modelos de regresión múltiple
- 9248 modelos de regresión ponderada
- 3468 modelos GLM

Top 10 mejores modelos

	var_1	var_2	AIC	Num_Ponderada	Ponderada	GLM	liga
10069	1.2	4.2	3531.735	NA	NA	Gamma	identity
10070	1.5	4.2	3531.748	NA	NA	Gamma	identity
10053	1.5	3.9	3531.760	NA	NA	Gamma	identity
10052	1.2	3.9	3531.766	NA	NA	Gamma	identity
10086	1.2	4.5	3531.808	NA	NA	Gamma	identity
10087	1.5	4.5	3531.840	NA	NA	Gamma	identity

Modelo seleccionado:

Se seleccionó un modelo GLM con distribución Gamma y función de enlace identidad, el cual obtuvo un AIC de: 3531.7345378, donde:

$$\mathbb{E}[\text{bpsystol} \mid \text{bmi}, \text{age}, \text{sex}] = \beta_0 + \beta_1 x_1^{1.2} + \beta_2 x_2^{4.2} + \beta_3 x_1$$

Aquí:

- x_1 representa la variable **bmi**
- x_2 representa la variable **age**

Verificación de supuestos

Haciendo la verificación de supuestos mediante pruebas de hipótesis, concluimos lo siguiente:

- Nuestro modelo cumple con la linealidad en las tres variables, gracias al test de Ramsey (librería `car`). Con un p-valor de 0.9835715 verificamos la linealidad de **bmi**; con un p-valor de 0.9853097 la de **age**; y con un p-valor de 1 la de **sex**.
- También se cumple el supuesto de homocedasticidad. Con un p-valor de 0.0685565, se concluye que no existe evidencia suficiente para rechazar la hipótesis nula de homocedasticidad.

- Además, gracias al tamaño de nuestra muestra y a la prueba de Anderson-Darling, con un p-valor de 0.978872, podemos concluir que los residuos se distribuyen normalmente (por tamaño muestral y contraste de hipótesis).
- Por último, la prueba de Breusch-Godfrey arrojó un p-valor de 0.9623929, lo que indica que no hay evidencia suficiente para rechazar la hipótesis nula de no autocorrelación.

Por lo tanto, nuestro modelo cumple con todos los supuestos requeridos.

Poniendo a prueba nuestro modelo

¿Se puede afirmar que, para una persona de cierta edad y sexo, tener un índice de masa corporal alto se asocia con una presión arterial sistólica elevada?

Para responder esta pregunta, se plantea el siguiente razonamiento: supongamos que tenemos dos pacientes del mismo sexo C y misma edad K, pero con diferente **bmi**: uno con $x < y$. Entonces, lo que queremos contrastar es:

$$\mathbb{E}[\text{bpsystol} \mid x, K, C] < \mathbb{E}[\text{bpsystol} \mid y, K, C]$$

Dado el modelo:

$$\mathbb{E}[\text{bpsystol} \mid \text{bmi}, \text{age}, \text{sex}] = \beta_0 + \beta_1(\text{bmi})^{1.2} + \beta_2(\text{age})^{4.2} + \beta_3(\text{sex})$$

Se tiene:

$$\beta_0 + \beta_1 x^{1.2} + \beta_2 K^{4.2} + \beta_3 C < \beta_0 + \beta_1 y^{1.2} + \beta_2 K^{4.2} + \beta_3 C$$

Lo cual se reduce a:

$$\beta_1 x^{1.2} < \beta_1 y^{1.2} \Leftrightarrow 0 < \beta_1(y^{1.2} - x^{1.2})$$

Dado que $y > x$ y la diferencia está elevada a una potencia positiva, y considerando que la función de enlace es la identidad, esta afirmación se cumple **si y solo si** $\beta_1 > 0$.

Por lo tanto, la hipótesis a contrastar es:

$$H_0 : \beta_1 \leq 0 \quad \text{vs} \quad H_A : \beta_1 > 0$$

Usando la prueba de hipótesis en R, se obtuvo un p-valor de `2e-16` , por lo que existe suficiente evidencia estadística para rechazar H_0 y concluir que $\beta_1 > 0$. El valor estimado de β_1 fue de 0.5246, lo que implica que, en pacientes del mismo sexo y edad, un mayor índice de masa corporal (**bmi**) se asocia con una mayor presión arterial sistólica.

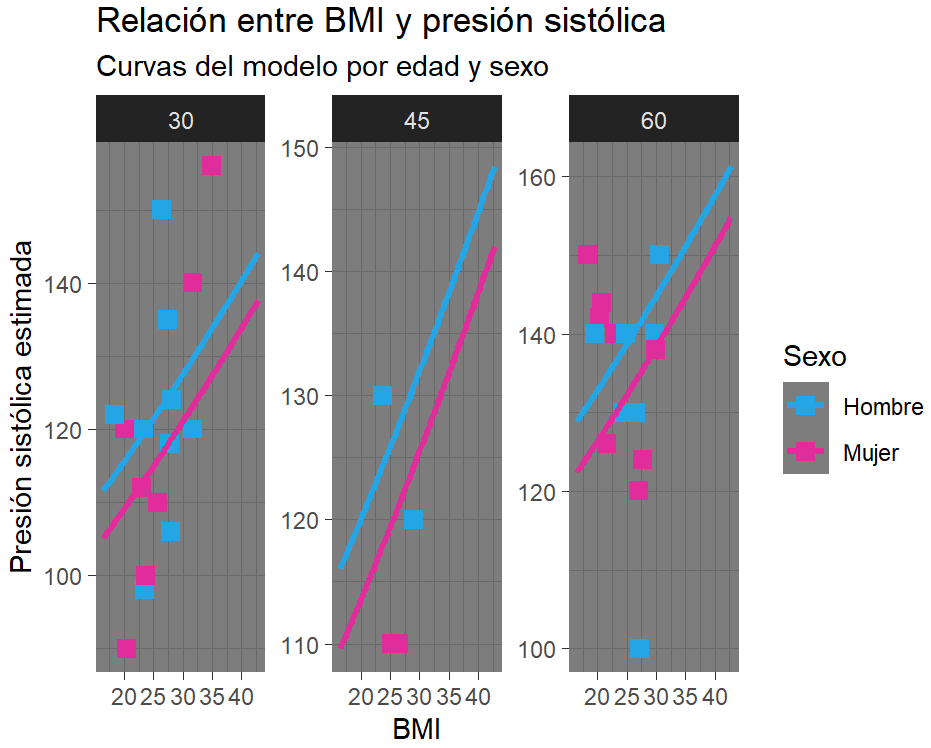
Este resultado sugiere que un aumento en el **bmi** podría tener un efecto negativo sobre la presión arterial, lo cual puede representar un riesgo para la salud cardiovascular. Por ello, se recomienda mantener un peso saludable como medida preventiva.

Gráfica

Para complementar la interpretación, se presenta la gráfica de la estimación puntual asociada a la relación entre **bpsystol** y **bmi**.

En esta visualización, se eligieron edades representativas de 30, 45 y 60 años (no se incluyó 65 años, ya que no había pacientes con esa edad en la muestra). Asimismo, se diferenciaron los grupos por sexo (hombres y mujeres), incorporando la predicción generada por nuestro modelo para cada combinación.

Esto nos permite observar cómo afecta el índice de masa corporal (**bmi**) a la presión arterial sistólica según edad y sexo.



A partir de la gráfica, y como conclusión general, se observa que efectivamente un menor **bmi** se asocia con una menor presión arterial sistólica, tal como fue respaldado por la prueba de hipótesis.

Este efecto no es invariante con respecto a la edad ni al sexo: aunque la tendencia general se mantiene, la magnitud del impacto varía entre grupos. Esto refuerza la importancia de mantener un peso saludable y una alimentación adecuada como medidas clave para preservar la salud cardiovascular.

Conclusión

En el modelo usado en el **Ejercicio 1**, se planteó lo siguiente:

$$\mathbb{E}[\ln(\text{bpsystol}) \mid \text{bmi},\ \text{age},\ \text{sex}] = \beta_0 + \beta_1 \ln(\text{bmi}) + \beta_2(\text{age})^3 + \beta_3(\text{sex})$$

Donde su AIC fue de -663.9559104, mientras que el GLM que elegimos obtuvo un AIC de: 3531.7345378. Estos AIC no son comparables directamente, ya que se basan en verosimilitudes distintas. Sin embargo, ambos modelos cumplen con los supuestos de regresión y permiten llegar a conclusiones sobre $\beta_1 > 0$. Por tanto, ambos son modelos válidos para explicar nuestros datos.

Aunque, como conclusión personal, me quedaría con el modelo **GLM**, ya que el modelo del ejercicio 1 transforma la variable dependiente a $\ln(\text{bpsystol})$ y también incluye $\ln(\text{bmi})$, lo que complica la interpretación de los resultados y dificulta explicar las escalas a personas que no conocen o no comprenden bien qué es un logaritmo. Por lo tanto, el modelo GLM sería una mejor opción.