Confounders

Joaquin Saposnik

5/10/2022

Table of contents

Introducción

Se desea hacer un modelo lineal simple entre las variables del dataset de seguros registrando la amplitud del efecto en la forma de los coeficientes que acompañana a la o las variables independientes. Para eso se probarán y compararán distintos modelos identificando el mejor de ellos.

Limpieza y adecuación de los datos

El dataset provisto posee 7 columnas, con variables categóricas (sexo, fumador, región) y numéricas (edad, índice de masa corporal, hijes, gastos de seguro médico). Se verificó que la importación de los datos sea correcta y se eliminaron filas repetidas. Se agregó 1 nueva columna llamada "salud":

- **Salud:** considerando la obesidad de una persona como el BMI mayor o igual a 30 y si esta persona es fumadora, separa entre personas Obesas Fumadoras (OF), Obesas No Fumadoras (ONF), No Obesas Fumadoras (NOF) y No Obesas No Fumadoras (NONF).

Modelo lineal simple N° 1

Se realizaron dos variaciones de un modelo lineal simple (ver Figure 1) teniendo en cuenta los gastos de una persona en función de su edad. Se utilizaron las fórmulas:

- **Modelo 1.1:** $y \sim x_1$
- **Modelo 1.2:** $y \sim x_1 - 1$

Reemplazando: $y = charges$, $x_1 = age$.

Se observó que el Modelo 1.2 tiene mejor estimación que el Modelo 1.1 ya que las edades no comienzan en 0; y al suprimir la ordenada al origen del modelo, se obtiene una mejor tendencia. Sin embargo, ninguno de ambos modelos se ajusta debidamente a los datos ya que hay otros factores que influyen en los gastos de una persona como su estado de salud. Las estadísticas de los Modelos 1.1 (ver Table 1) y 1.2 (ver Table 2) demuestran que efectivamente no son ideales.

Table 1: Estadísticas del modelo lineal simple N° 1.1.

	charges					
Predictors	p	Statistic	Estimates	standardized std. Error	std. Error	std. Beta

	charges					
(Intercept)	**6.95e-04**	3.40	3190.02	0.03	938.40	0.00
age	**6.98e-29**	11.42	257.23	0.03	22.53	0.30
Observations	1337					
R^2 / R^2 adjusted	0.089 / 0.088					

Table 2: Estadísticas del modelo lineal simple N° 1.2.

	charges					
Predictors	p	Statistic	Estimates	standardized std. Error	std. Error	std. Beta
age	**5.68e-256**	43.21	329.33	0.03	7.62	0.30
Observations	1337					
R^2 / R^2 adjusted	0.583 / 0.583					

Modelo lineal simple N° 2

Para mejorar la estimación de los gastos, se procede a cambiar el modeloagregando la variable "salud" que posee 4 posibles estados de salud (ver Figure 2). Se utilizan los modelos:

- **Modelos 2.1:** $y \sim x_1 + x_2$
- **Modelos 2.2:** $y \sim x_1 + x_2 - 1$

Reemplazando: $y = charges$, $x_1 = age$, $x_2 = salud$.

Se observa que los nuevos modelos tienen mejores estimaciones para los distintos grupos de personas según su estado de salud, siendo mucho mejor el Modelo 2.2 según la comparación de las estadísticas de los modelos (Modelo 2.1, ver Table 3 y Modelo 2.2, ver Table 4).

Table 3: Estadísticas del modelo lineal simple N° 2.1.

	charges					
Predictors	p	Statistic	Estimates	standardized std. Error	std. Error	std. Beta
(Intercept)	**9.73e-114**	25.04	9523.57	0.01	380.29	0.56
age	**2.19e-151**	29.99	267.72	0.01	8.93	0.31
salud [linear]	**0.00e+00**	-82.82	-25286.22	0.03	305.33	-2.09
salud [quadratic]	**1.34e-165**	31.80	9851.18	0.03	309.77	0.81
salud [cubic]	**3.48e-06**	4.66	1466.13	0.03	314.62	0.12
Observations	1337					
R^2 / R^2 adjusted	0.858 / 0.858					

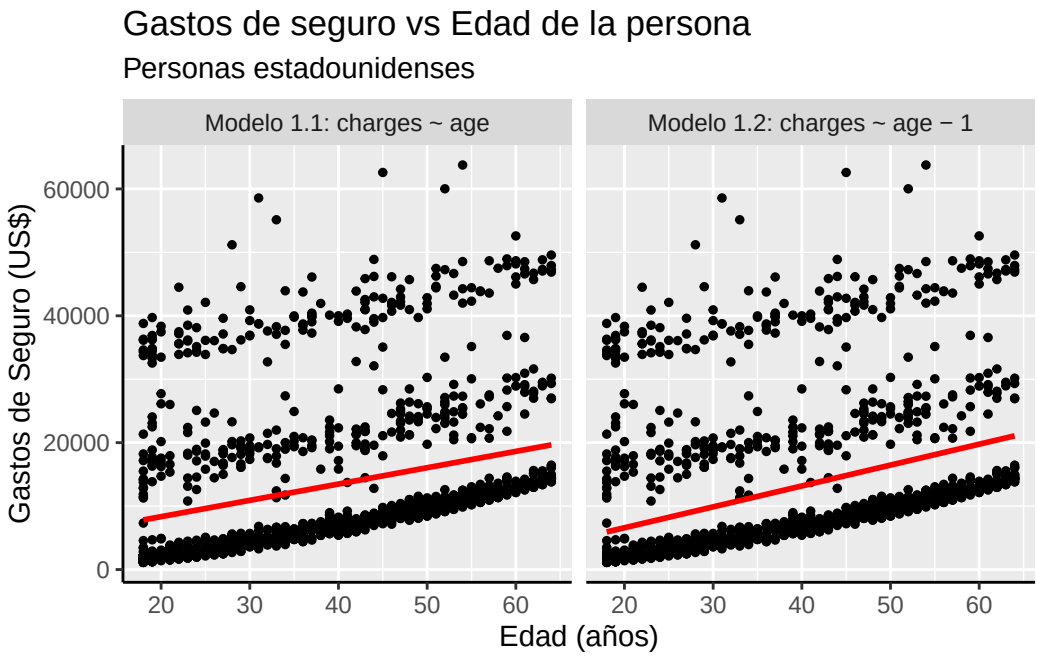

Figure 1: Modelo lineal simple N° 1 - Gastos vs Edad.

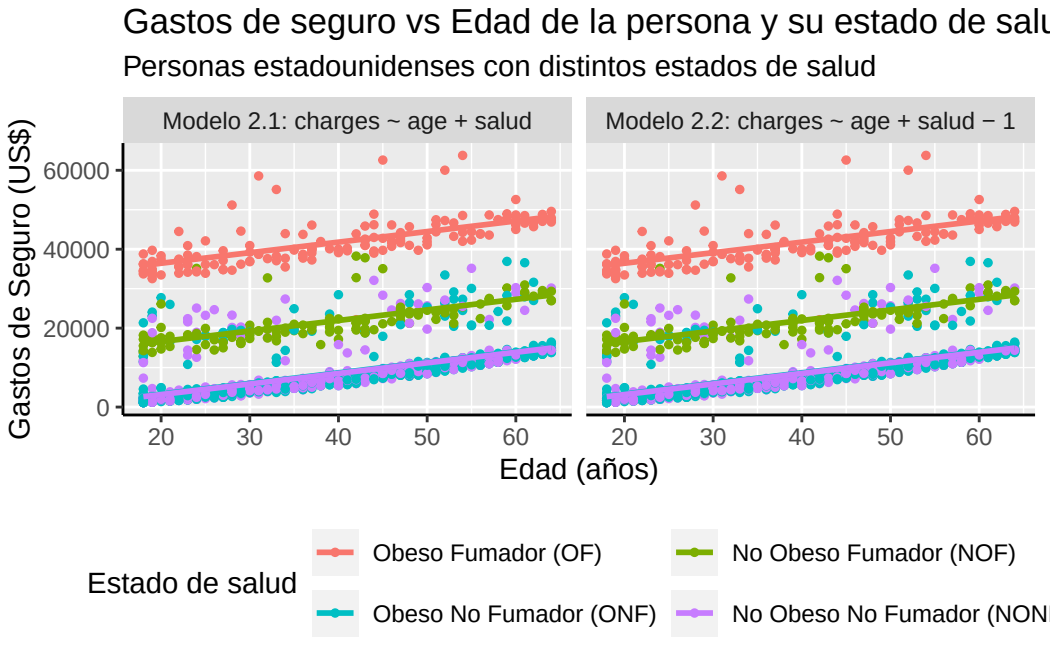

Figure 2: Modelo lineal simple N° 2 - Gastos vs Edad y Salud.

Table 4: Estadísticas del modelo lineal simple N° 2.2.

	charges					
Predictors	p	Statistic	Estimates	standardized std. Error	std. Error	std. Beta
age	**2.19e-151**	29.99	267.72	0.01	8.93	0.31
saludOF	**0.00e+00**	60.31	31083.83	0.03	515.40	2.34
saludNOF	**1.32e-87**	21.41	11235.67	0.03	524.89	0.70
saludONF	**7.98e-07**	-4.96	-2039.70	0.02	411.27	-0.40
saludNONF	**3.91e-08**	-5.53	-2185.51	0.02	395.41	-0.41
Observations	1337					
R^2 / R^2 adjusted	0.936 / 0.936					

Modelo lineal simple N° 3

Se analizó una última serie de modelos (ver Figure 3) para comprobar si pueden mejorar la estimación del Modelo 2.2 (ver Table 4). Se utilizaron las siguientes fórmulas:

- **Modelos 3.1:** $y \sim x_1 * x_2$
- **Modelos 3.2:** $y \sim x_1 * x_2 - 1$

Reemplazando: $y = charges$, $x_1 = age$, $x_2 = salud$.

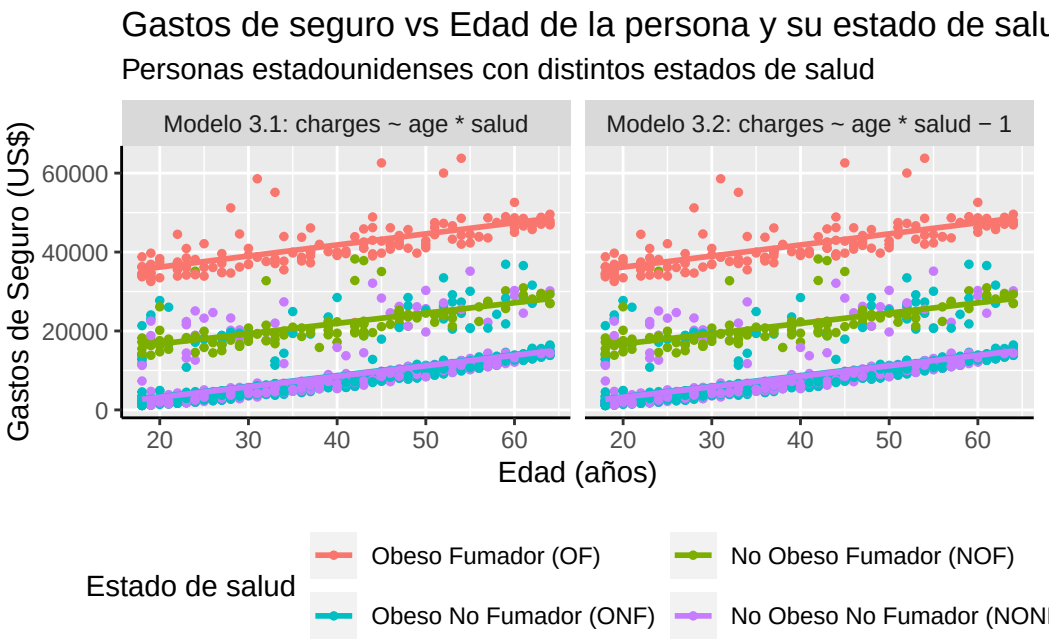

Figure 3: Modelo lineal simple N° 3 - Gastos vs Edad y Salud.

Al no poder visualizar un cambio significativo con respecto al gráfico del Modelo 2.2 (ver Figure 2), se procede a observar las estadísticas (Modelo 3.1 ver Table 5, y Modelo 3.2 ver Table 6.

Table 5: Estadísticas del modelo lineal simple N° 3.1.

	charges					
Predictors	p	Statistic	Estimates	standardized std. Error	std. Error	std. Beta
(Intercept)	**2.00e-82**	20.66	9481.83	0.01	459.05	0.56
age	**6.34e-106**	23.98	268.72	0.01	11.20	0.31
salud [linear]	**2.59e-134**	-27.78	-24943.04	0.03	897.83	-2.09
salud [quadratic]	**4.45e-24**	10.32	9476.10	0.03	918.11	0.81
salud [cubic]	6.05e-02	1.88	1762.24	0.03	937.94	0.12
age * salud [linear]	6.88e-01	-0.40	-8.75	0.03	21.77	-0.01
age * salud [quadratic]	6.66e-01	0.43	9.67	0.03	22.41	0.01
age * salud [cubic]	7.36e-01	-0.34	-7.75	0.03	23.03	-0.01
Observations	1337					
R^2 / R^2 adjusted	0.858 / 0.858					

Table 6: Estadísticas del modelo lineal simple N° 3.2.

	charges					
Predictors	p	Statistic	Estimates	standardized std. Error	std. Error	std. Beta
age	**6.34e-106**	23.98	268.72	0.01	11.20	0.31
saludOF	**7.27e-133**	27.59	30558.13	0.03	1107.51	2.34
saludNOF	**4.73e-21**	9.58	11503.36	0.03	1201.26	0.70
saludONF	**5.11e-04**	-3.48	-2015.80	0.02	578.67	-0.40
saludNONF	**4.88e-04**	-3.50	-2118.37	0.02	605.99	-0.41
age * salud L	6.88e-01	-0.40	-8.75	0.03	21.77	-0.01
age * salud Q	6.66e-01	0.43	9.67	0.03	22.41	0.01
age * salud C	7.36e-01	-0.34	-7.75	0.03	23.03	-0.01
Observations	1337					
R^2 / R^2 adjusted	0.936 / 0.935					

Puede verse que el Modelo 3.2 (ver Table 6) tiene muy buenas estimaciones con respecto al Modelo 3.1 (ver Table 5). A su vez, si comparamos los Modelos 2.2 (ver Table 4) y 3.2, se observa que son prácticamente idénticos siendo ligeramente mejor el Modelo 2.2.

Conclusiones

Se decide optar por usar el Modelo 2.2 (ver Figure 2 y Table 4) ya que tiene una mejor estimación para los gastos de las personas según su edad y estado de salud.