Intro. a la Ciencia de Datos
UNSAM

Entrega 6
Confounders

Julián Fraga
fragajulian96@gmail.com

Octubre 2022

1. Exploración inicial del dataset de Seguros Médicos

Se trabajó sobre el dataset de seguros médicos. Lo que se observa al plottear los datos es una división notable a simple vista en al menos tres grupos. Por los trabajos anteriores, se sabe que las condiciones que provocan esta diferenciación son el sobrepeso y si la persona fuma o no. A continuación se muestran los gráficos de gastos médicos según la edad diferenciando con color entre personas fumadoras y no fumadoras, y personas con y sin sobrepeso diferenciando (*figs.* 1 y 2).

Lo que se extrae de estos gráficos es, como se mencionó en los trabajos anteriores, que las personas fumadoras gastan más de base que las personas no fumadoras, y si tienen sobrepeso aún más. Sin embargo, a simple vista no se observa una tendencia diferente con la edad entre grupos, sino simplemente un *offset* entre ellos

Estos offset son lo que dificulta la tarea de realizar ajustes lineales y se discutirá en la sección siguiente.

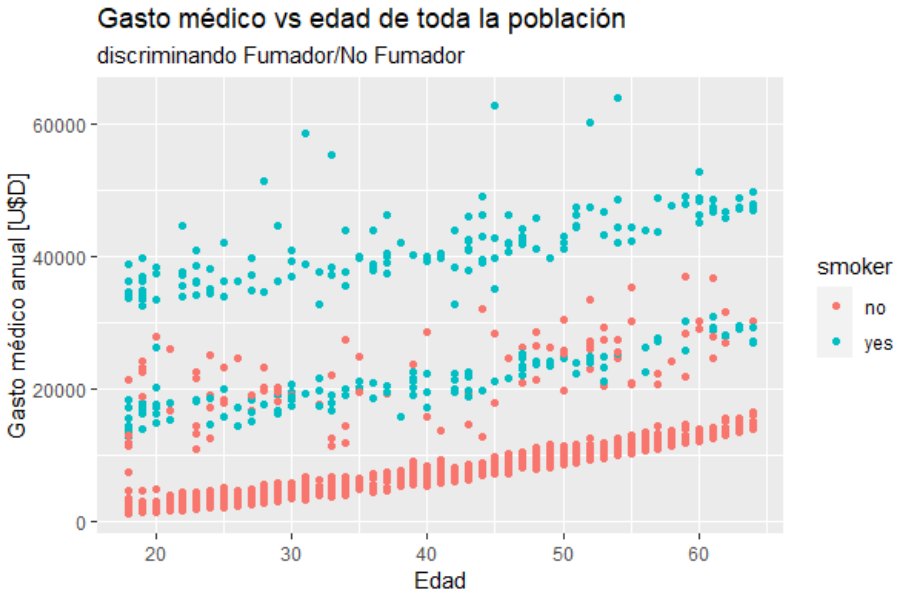

Figura 1: Gastos médicos por edad. Diferenciación por color de personas fumadoras

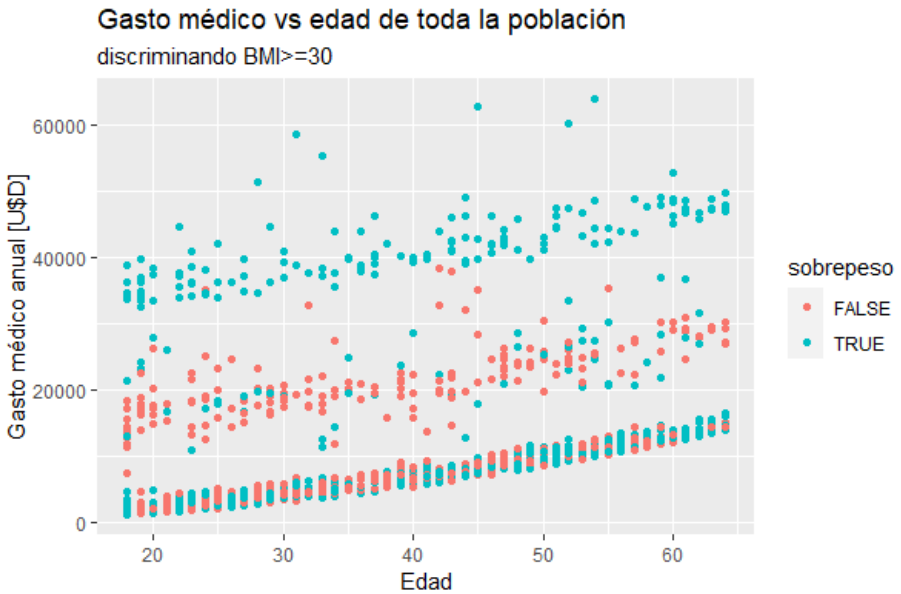

Figura 2: Gastos médicos por edad. Diferenciación por color de personas con BMI$\geq$ 30

2. Ajustes lineales

Al intentar realizar un ajuste lineal sobre todo el conjunto de datos, la aglomeración que se mencionó anteriormente representa el problema principal. Como se observa en el gráfico del ajuste (*fig.* 3), la gran mayoría de puntos ni siquiera toca al gráfico del ajuste.

La mala calidad del ajuste resulta aún más evidente viendo la diferencia entre los residuos del ajuste anterior y los del ajuste sólo sobre la población fumadora con BMI mayor o igual a 30 (*figs.* 4 y 5). Se observa que los residuos del ajuste sobre el grupo reducido de personas fumadoras y con sobrepeso están mucho menos dispersos alrededor del cero.

Los resultados completos, detallados para cada sub-categoría se muestran en la Tabla 1

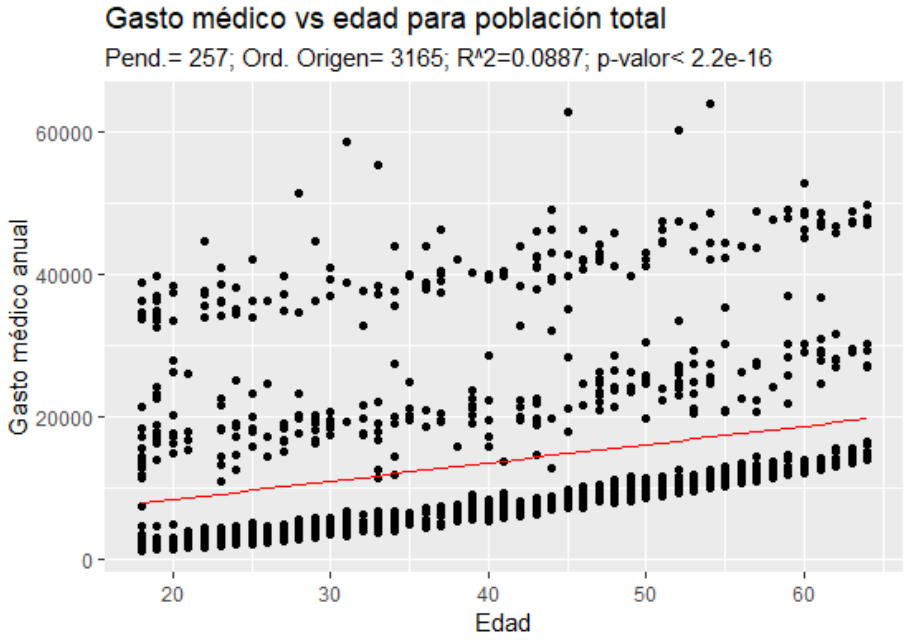

Figura 3: Ajuste lineal sobre gastos médicos por edad en la población total

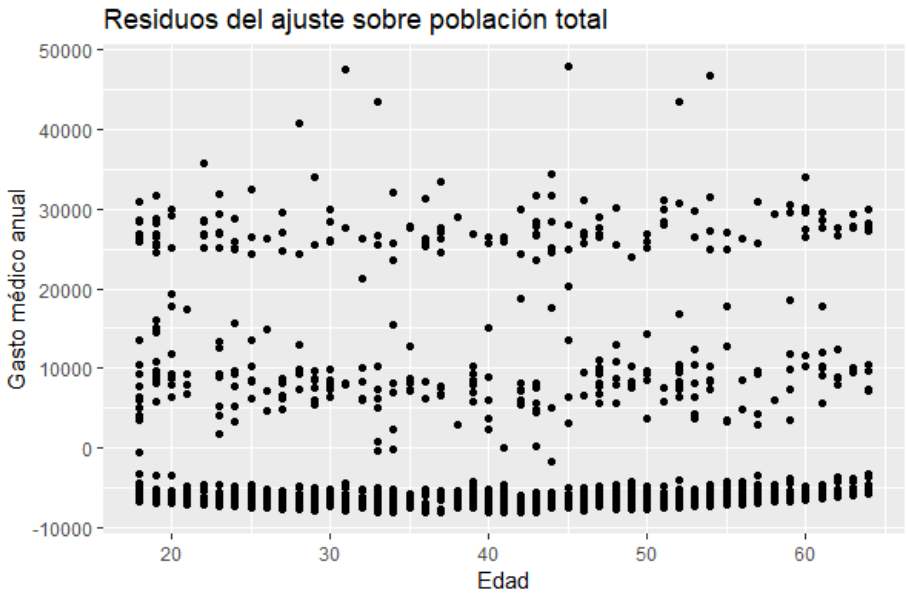

Figura 4: Residuos del ajuste lineal sobre toda la muestra

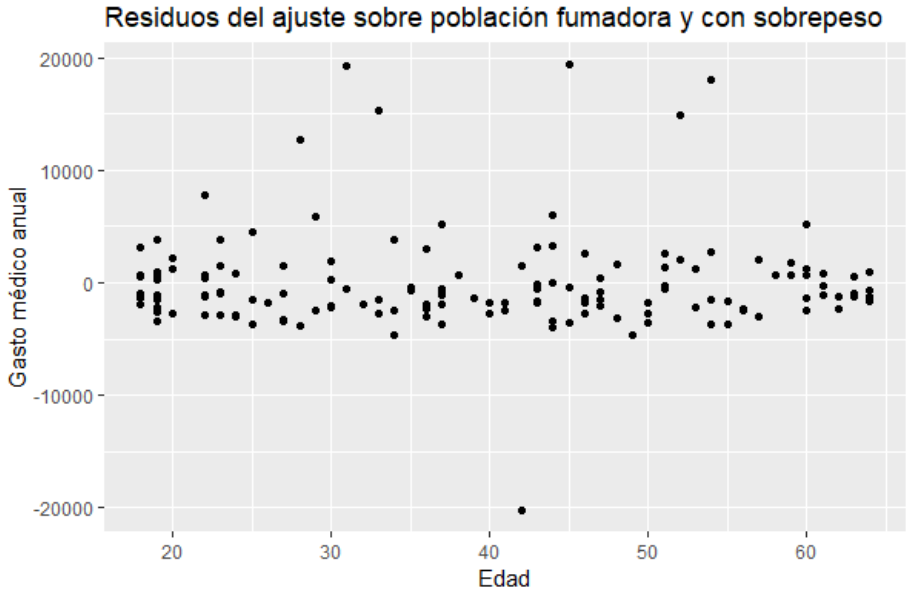

Figura 5: Residuos del ajuste lineal sobre el grupo de personas fumadoras con BMI mayor o igual a 30

Subgrupo	Ordenada al origen [$]	Desviación estándar relativa
Fumador BMI ≥ 30	30600	3,5 %
Fumador BMI< 30	11500	8,4 %
No Fumador BMI≥ 30	-2030	30 %
No Fumador BMI< 30	-2120	29 %

Tabla 1: Ordenadas al origen calculadas por ajuste lineal habiendo segmentado el dataset en cuatro grupos. Se observa una gran diferencia entre los valores medios obtenidos, con un promedio pesado de alrededor de 6300. Entre el máximo y el mínimo hay una diferencia de aproximadamente cinco veces el promedio.

Por otro lado, las pendientes calculadas diferenciando grupos no presentan una discrepancia tan significativa entre ellas. Los resultados se detallan en la Tabla 2.

Subgrupo	Pendiente [$/ años]	Desviación estándar relativa
Fumador BMI ≥ 30	281	9,3 %
Fumador BMI< 30	260	9,2 %
No Fumador BMI≥ 30	267	5,2 %
No Fumador BMI< 30	265	5,6 %
Promedio Pesado	268	2,4 %

Tabla 2: Pendientes calculadas por ajuste lineal habiendo segmentado el dataset en cuatro grupos. Se observa una diferencia entre valores medios de no más de 71 $/años, un 25 % de variación respecto del promedio pesado.

3. Análisis de resultados y posible solución al problema

Los resultados muestran que efectivamente existe una relación lineal entre edad y costo de seguro médico para los distintos grupos, y que la tasa de crecimiento es estable de 268,0 $/año de vida $\pm$2,4 %. Sin embargo, es imposible estimar el gasto por edad si no se tiene en cuenta los factores de riesgo, los cuales desplazan la curva de gastos significativamente.

Una posible solución es considerar una nueva variable *riesgo* para eliminar el offset del análisis. Esta variable deberá ser una combinación lineal de las variables {*edad, BMI, Fumador*}, luego el gasto médico es directamente proporcional al riesgo. La ecuación relacionando las cuatro variables

se vería de esta forma:

$$gasto \sim riesgo = a \times edad + b \times BMI + c \times Fumador$$

Comentario personal

Por el momento no tengo las herramientas para ejecutar un ajuste así teniendo en cuenta que el BMI no afecta de manera lineal al riesgo, sino que se comporta más parecido a una variable dicotómica incluso siendo numérica continua. Lo mismo la condición de ser fumador, no sabría cómo operar con dos variables binarias en un ajuste lineal.

Esperaré a la devolución para poder implementarlo.