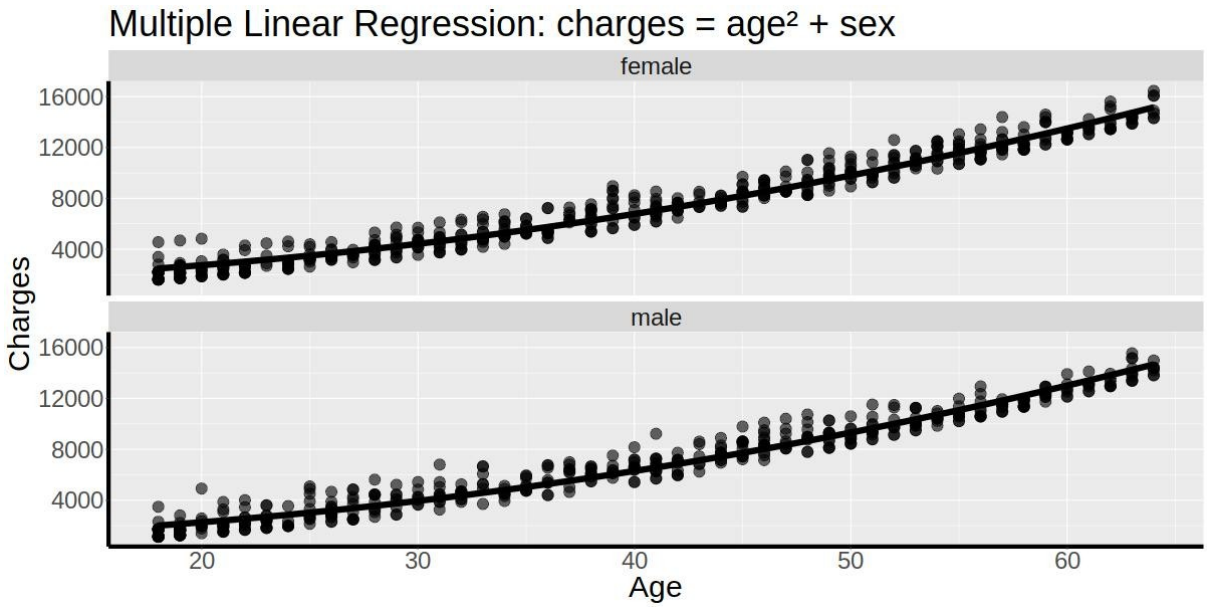
07 – Análisis de dataset Insurance

Autor: Peitsch, Pablo **mail:** preynosopeitsch@estudiantes.unsam.edu.ar
Github: @PPeitsch **Fecha: 2023-05-14**

Se utiliza el dataset de cargos de seguro médico en EEUU, de dominio público [1]. Se aplicó un modelo de MLR (regresión lineal múltiple) utilizando variables predictoras específicas y una transformación de variable. Se utilizaron la variable edad (age) y el sexo (sex) de los asegurados como variables explicativas. Se realizó una transformación cuadrática de la variable edad, representada en la figura como age^2, con el objetivo de identificar relaciones no lineales entre la edad y los cargos de seguro. Estas variables y su transformación se incorporaron en el modelo para investigar cómo influyen en la predicción de los cargos de seguro.



Coefficients				
	Estimate	Std. Error	t-value	Pr (>|t|)
charges	1402.5623	51.1853	27.402	<2e-16
I(age^2)	3.3591	0.0216	155.513	<2e-16
sexmale	-481.2816	48.6477	-9.893	<2e-16
Residual Standard Error:		755.1	Multiple R-squared:	0.962
Adjusted R-squared:		0.9619	p-value:	<2e-16

El modelo MLR se expresa como charges = b0 + b1 * age^2 + b2 * sex, donde b0, b1 y b2 son los parámetros estimados. El parámetro b0 representa el valor esperado de los cargos de seguro cuando la edad y sexo son iguales a cero, lo cual tiene sentido matemático pero no práctico; luego, b1 representa el cambio en los cargos de seguro asociado al incremento de la edad al cuadrado; por último, b2 refleja la diferencia de cargos de seguro entre ambos sexos. Si b1>0 indica que con el aumento de edad los cargos de seguro tienden a aumentar de manera cuadrática. Si b2>0 significa que los hombres tienen costos de seguro más altos en comparación con las mujeres, en cambio b2<0 implica lo opuesto, las mujeres tienen costos de seguro más altos que los hombres. Todos los coeficientes tienen un p-value muy cercano a cero, lo que indica que son estadísticamente significativos. El Adjusted R-squared muestra que aproximadamente el 96.19% de la variabilidad de los cargos de seguro se explica por las variables predictoras incluidas en el modelo, lo que indica un buen ajuste del modelo a los datos.

REFERENCIAS.

1. US Health Insurance Dataset, https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset