

Universidad Autónoma de Zacatecas

Unidad Académica de Matemáticas



**Desarrollo de un Modelo Predictivo para Determinar  
Primas en Seguros de Gastos Médicos a través de  
Regresión Lineal**

Juan de Jesús Venegas Flores

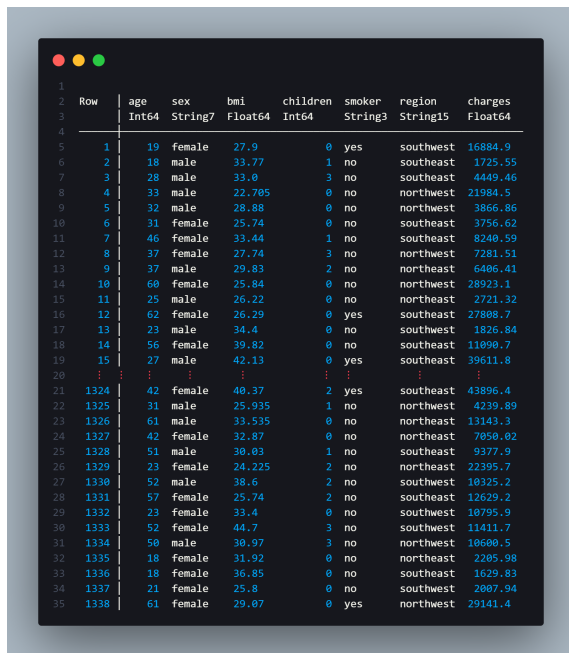
07/06/2024

# 1 Introducción

En este proyecto final para el curso de modelo lineal y no lineales trataremos de utilizar un modelo de regresion lineal multiple para poder predecir cuanto sera el costo de una prima de un seguro de gastos medicos dado las características de una persona

## 2 Obtencion de datos

los datos que se utilizaran se obtuvieron Kaggle y se cuentan factores personales (edad, género, IMC, tamaño de la familia, hábitos de fumar), factores geográficos y su impacto en los costos del seguro médico. se utilizara un modelo de regresion lineal multiple utilizando una seleccion de variables con una cantidad de registros de 1338.



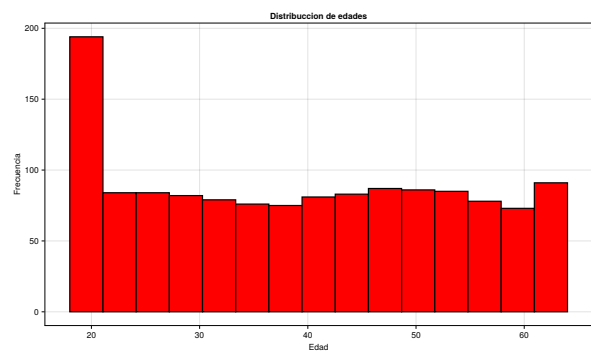
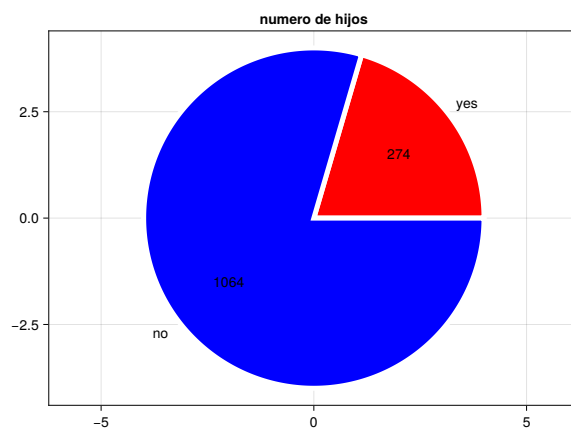
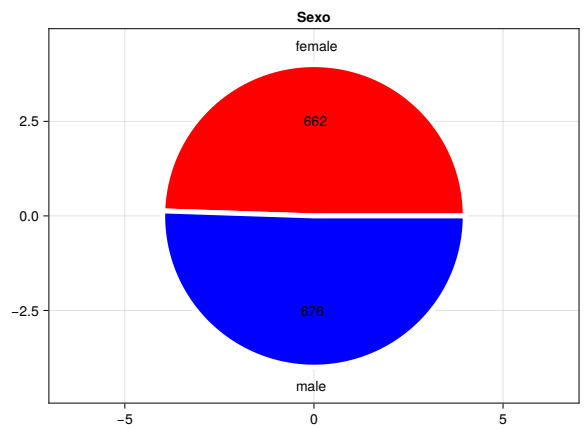
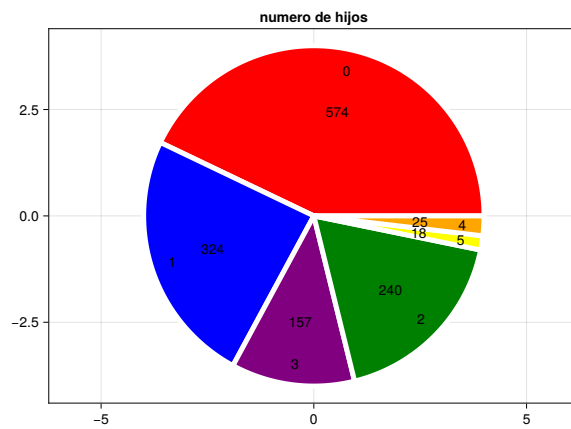
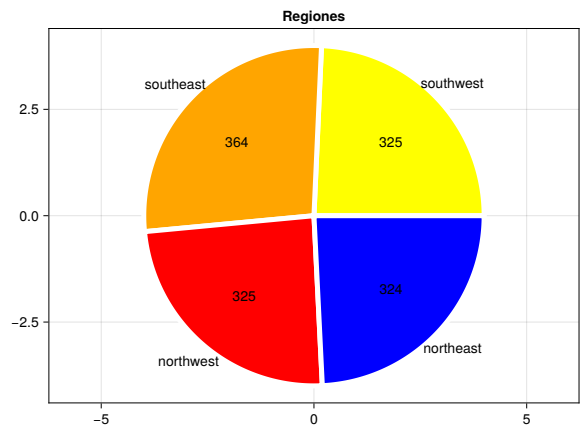
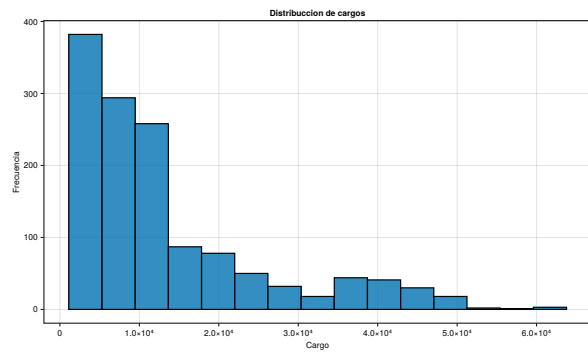
Row	age	sex	bmi	children	smoker	region	charges
	Int64	String7	Float64	Int64	String3	String15	Float64
1	19	female	27.9	0	yes	southwest	16884.9
2	18	male	33.77	1	no	southeast	1725.55
3	18	male	33.0	3	no	southeast	4449.46
4	33	male	22.705	0	no	northwest	21984.5
5	32	male	28.88	0	no	northwest	3866.86
6	31	female	25.74	0	no	southeast	3756.62
7	46	female	33.44	1	no	southeast	8240.59
8	37	female	27.74	3	no	northwest	7281.51
9	37	male	29.83	2	no	northwest	6406.41
10	60	female	25.84	0	no	northwest	28923.1
11	25	male	26.22	0	no	northwest	2721.32
12	62	female	26.29	0	yes	southeast	27808.7
13	23	male	34.4	0	no	southeast	1826.84
14	56	female	39.82	0	no	southeast	11090.7
15	27	male	42.13	0	yes	southeast	39611.8
1324	42	female	40.37	2	yes	southeast	43896.4
1325	31	male	25.935	1	no	northwest	4239.89
1326	61	male	33.535	0	no	northwest	13143.3
1327	42	female	32.87	0	no	northwest	7050.02
1328	51	male	30.03	1	no	southeast	9377.9
1329	23	female	24.225	2	no	northwest	22395.7
1330	52	male	38.6	2	no	southwest	10325.2
1331	57	female	25.74	2	no	southeast	12629.2
1332	23	female	33.4	0	no	southwest	10795.9
1333	52	female	44.7	3	no	southwest	11411.7
1334	50	male	30.97	3	no	northwest	10600.5
1335	18	female	31.92	0	no	northwest	2205.98
1336	18	female	36.85	0	no	southeast	1629.83
1337	21	female	25.8	0	no	southwest	2007.94
1338	61	female	29.07	0	yes	northwest	29141.4

La cantidad de variables que se tienen son 6, de las cuales 3 son variables categóricas.

Variable	Categorías
Género	Masculino, Femenino(Categorica)
Edad	Edad(Continua)
bmi	Indice de masa corporal(Continua)
children	Cantidad de hijos dependientes("Continua")
smoker	Si la persona fuma(Categorica)
Region	en que parte la region vive la persona(Categorica)
Charges	Recargos al costo de la prima(Continua)

antes de proponer un primero modelo de regresion haremos un pequeño analisis descriptivo de los datos, ver como se distribuyen, y como estan conformados

### 3 Analisis descriptivo de los datos



como primer modelo podemos considerar tanto la traslacion de las variables categoricas y aumento de pendiente en cada una de las variables continuas y nada de iteracciones, si consideramos todas las variable categoricas en el modelo serian  $4*2*2$  particiones de los datos con lo que que cada variable categorica tendria sus 4 parametros referente a las tres variables continuas que tenemos (bmi,edad,hijos) y su traslacion, por lo que en total tendríamos un total  $4*(5+1)$  parametros, el modelo tendria la siguiente forma

Letra	Variable	Significado
A	age	Edad del individuo
S	sex	Sexo del individuo
B	bmi	Índice de Masa Corporal (Body Mass Index)
C	children	Número de hijos dependientes
M	smoker	Fumador
R	region	Región geográfica (por ejemplo, nordeste, sureste, etc.)
Q	charges	Costos médicos (gastos de seguros)

$$Q = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 \mathbb{I}_{\{S=male\}} + \beta_5 A \mathbb{I}_{\{S=male\}} + \dots + \beta_{22} B \mathbb{I}_{\{R=southwest\}} + \beta_{23} C \mathbb{I}_{\{R=southwest\}}$$

se toma en cuenta como modelo base (donde todas las indicadoras valen cero) el caso donde el sexo es femenino, no fuma y se encuentra en la region noreste, en total son 24 parametros, trabajar con un modelo asi seria totalmente impractico si no fuera por los softwares que nos facilitan las cuentas, con este modelo completo haremos una discriminacion de variables para ver cuales son las que se deben tomar en cuenta en el modelo

utilizando el software Julia y R, podemos obtener el modelo con las variables relevantes utilizando el metodo backward y forward

el metodo de seleccion de variables considero suficientes 14 parametros, sin utilizar tantos parametros y solo suponiendo una traslacion para cada variable indicadora sin tener en cuenta un cambio de pendiente en cada variable continua se obtenia un  $R^2$  ajustado de .71 y utilizando esta seleccion se pudo obtener un  $R^2$  de casi .85, lo cual no hecha en vano esta seleccion de variables adicionales y un modelo mas complejo, algo que puede parecer un poco raro es que el coeficiente de  $\mathbb{I}_{\{M=yes\}}$  sea negativo y pense que estaba dectenado mal en mi modelo cual era la variable indicadora, pero lo que estaba pasando es que hay una correlacion muy fuerte entre este y la interaccion con el indice de masa corporal  $B \mathbb{I}_{\{M=yes\}}$  donde este si tiene un coeficiente positivo, haciendo que si la funcion indicadora sea uno cuando una persona es fumadora se tiene que  $\beta_8 \mathbb{I}_{\{M=yes\}} + \beta_{10} B \mathbb{I}_{\{M=yes\}} > 0$  donde ahora si tiene sentido, por que si una persona es fumadora es natural que a esta se le cobre mas, tambien se ve que hay muchas variables que bajo la prubea t no son signifcativas pero bajo la seleccion de variables estan si se deberian incluir

```

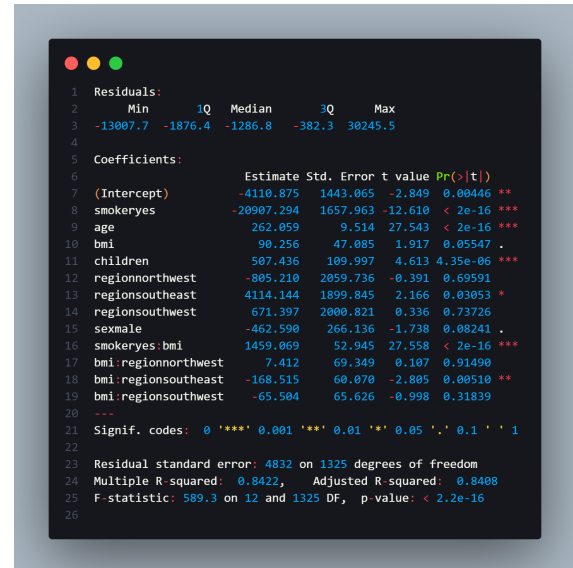
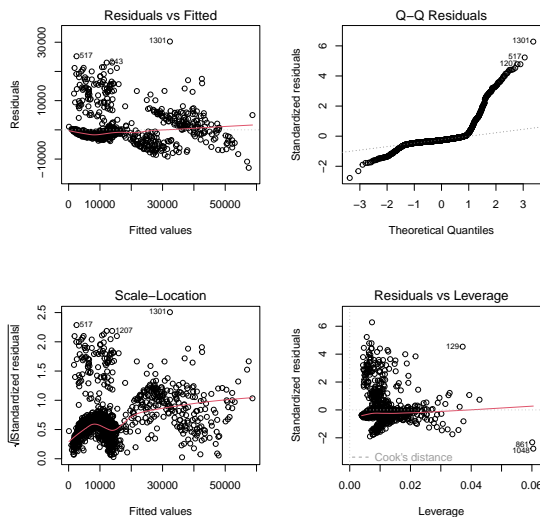
1 Call:
2 lm(formula = charges ~ smoker + age + bmi + children + region +
3 sex + smoker:bmi + bmi:region, data = Datos)
4
5
6 Residuals:
7      Min       1Q   Median       3Q      Max
8 -13007.7  -1876.4  -1286.8   -382.3   30245.5
9
10 Coefficients:
11             Estimate Std. Error t value Pr(>|t|)
12 (Intercept)   -4110.875    1443.065   -2.849   0.00446 **
13 smokeryes    -20907.294    1657.963  -12.610 < 2e-16 ***
14 age           262.059      9.514    27.543 < 2e-16 ***
15 bmi           90.256      47.085     1.917   0.05547 .
16 children     507.436     109.997     4.613 4.35e-06 ***
17 regionnorthwest -805.210     2059.736   -0.391   0.69591
18 regionsoutheast 4114.144     1899.845     2.166   0.03053 *
19 regionsouthwest 671.397     2000.821     0.336   0.73726
20 sexmale      -462.590     266.136   -1.738   0.08241 .
21 smokeryes:bmi 1459.069      52.945    27.558 < 2e-16 ***
22 bmi:regionnorthwest 7.412      69.349     0.107   0.91490
23 bmi:regionsoutheast -168.515     60.070   -2.805   0.00510 **
24 bmi:regionsouthwest -65.504      65.626   -0.998   0.31839
25 ---
26 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
27
28 Residual standard error: 4832 on 1325 degrees of freedom
29 Multiple R-squared:  0.8422,    Adjusted R-squared:  0.8408
30 F-statistic: 589.3 on 12 and 1325 Df, p-value: < 2.2e-16
31

```

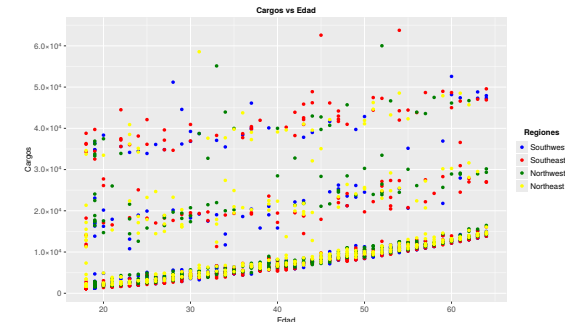
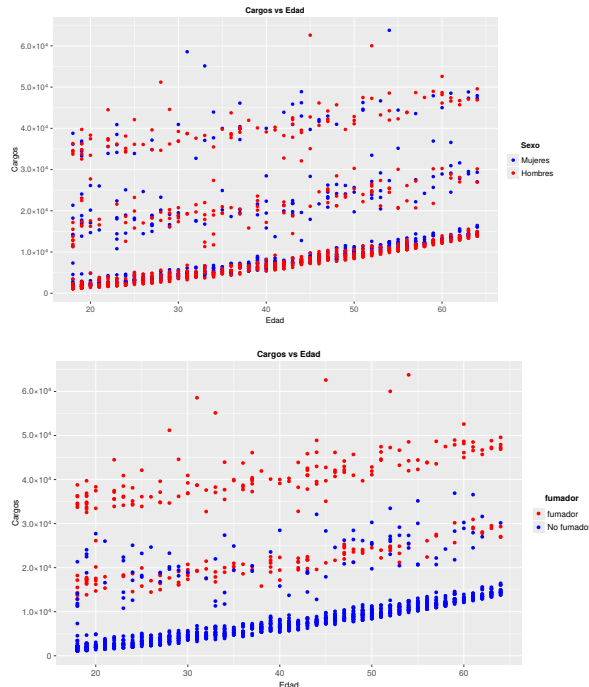
el modelo que se obtuvo fue el siguiente

$$Q = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_4 \mathbb{I}_{\{S=male\}} + \beta_8 \mathbb{I}_{\{M=yes\}} + \beta_{11} C \mathbb{I}_{\{M=yes\}} + \beta_{12} \mathbb{I}_{\{S=Nortwest\}} + \beta_{15} C \mathbb{I}_{\{S=Nortwest\}} + \beta_{16} \mathbb{I}_{\{S=southeast\}} + \beta_{18} C \mathbb{I}_{\{S=southeast\}} + \beta_{19} \mathbb{I}_{\{S=southwest\}} + \beta_{23} C \mathbb{I}_{\{S=southwest\}}$$

el problema fuerte ahora es al momento de analizar los residuos de este modelo



En los residuales se observa una cierta curvatura, lo que puede indicar que los datos no siguen una relación lineal. Además, parece haber una variabilidad significativa en los datos: algunos tienen una varianza pequeña, mientras que otros muestran una varianza considerablemente mayor. Los errores no parecen seguir una distribución normal y el modelo no pasa la prueba de normalidad. A pesar de que el modelo es significativo y tiene un  $R^2$  de 0.84, lo cual es bastante bueno, el principal problema radica en el incumplimiento de los supuestos de la regresión. En los residuales parece haber ciertos patrones que suelen indicar que los residuales sí dependen de los valores ajustados para encontrar por qué está sucediendo esto veamos las siguientes gráficas



Estas imágenes muestran que la única categoría que logra discriminar significativamente los datos es si la persona es fumadora o no. El género, es decir, si la persona es hombre o mujer, parece no tener un impacto considerable. En cuanto a la región, hay una tendencia que indica que las personas de la región sudoeste presentan valores bajos.

En los gráficos se comparan los cargos de la prima contra la edad, que considero como la principal variable continua involucrada, observando cómo se comportan diferentes categorías en relación con esta variable. Se ve claramente que la categoría más significativa es si la persona es fumadora o no. Los datos parecen estar segmentados en tres grupos, donde todos los datos en el grupo superior corresponden a fumadores y todos los datos en el grupo inferior corresponden a no fumadores. Esto sugiere que ser fumador es una característica muy discriminante en los datos.

Además, en el modelo anterior teníamos tres variables continuas (edad, índice de masa corporal (BMI), y número de hijos). Sin embargo, la cantidad de hijos varía entre 0 y 6, por lo que podríamos considerar esta variable como categórica. Aunque este cambio puede parecer complejo manualmente, con la ayuda del software es factible. Al hacer esto, podríamos simplificar el modelo a solo dos variables continuas, lo que facilitaría su visualización y análisis gráfico

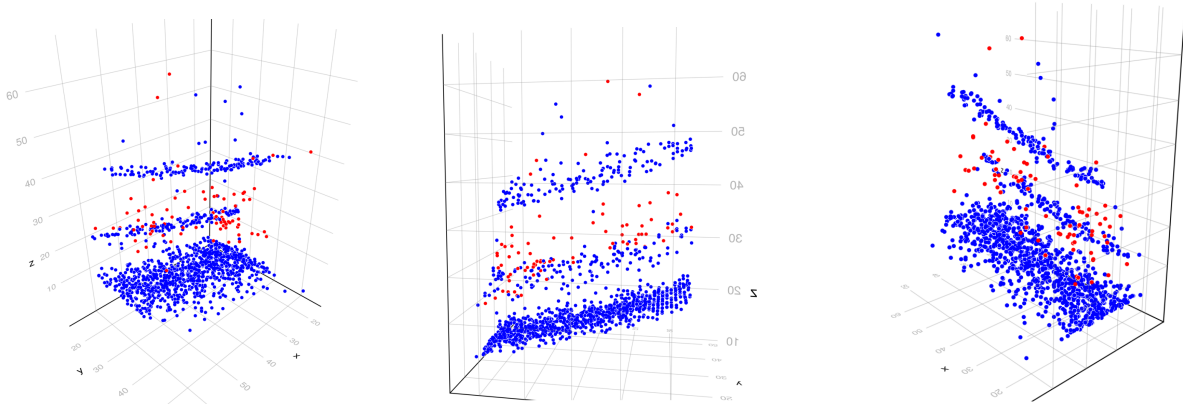


Figure 1: Datos influyentes bajo DFFITS

En la gráfica anterior, se observa que los datos están particionados en tres segmentos distintos (como ya se veía en las graficas anteriores) y el problema persiste: en cada segmento hay una varianza diferente. Para valores bajos de los ajustados, la varianza es considerablemente menor en comparación con los valores altos.

Otro aspecto que sospecho es incorrecto en el modelo y los datos es la posible omisión de una variable importante. Esto se evidencia en las imágenes anteriores, donde datos con características muy similares muestran grandes diferencias, sugiriendo que hay variables relevantes que fueron consideradas inicialmente para el cobro de cargos, pero que se omitieron en el modelo.

Para "comprobar" esta sospecha, propongo el siguiente enfoque: en la gráfica anterior se observa claramente que los datos se separan en tres grupos con diferentes traslaciones. Consideremos un modelo en el que simplemente asumimos un aumento en el intercepto, es decir, el siguiente modelo:

$$C = \beta_0 A + B\beta_1 + \beta_3 \mathbb{I}_{\{S=\text{male}\}} + \beta_4 \mathbb{I}_{\{R=\text{yes}\}} + \beta_5 \mathbb{I}_{\{C=1\}} + \dots + \beta_{11} \mathbb{I}_{\{C=6\}} + \beta_{12} \mathbb{I}_{\{R=\text{southwest}\}} + \dots + \beta_{15} \mathbb{I}_{\{R=\text{Northeast}\}} \quad (3)$$

Este modelo, al ajustar los parámetros para todos los datos, generaría 96 hiperplanos (en caso de que cada parámetro de cada categoría sea significativo). Si los datos realmente discriminan bien, se deberían generar tres grupos de planos en cada subgrupo de las acumulaciones observadas en la gráfica. parecerá un poco engorroso, pero es mas tardado escribir el modelo que pedirselo al softwar, veamos como se conglomeran estos 96 planos en la grafica

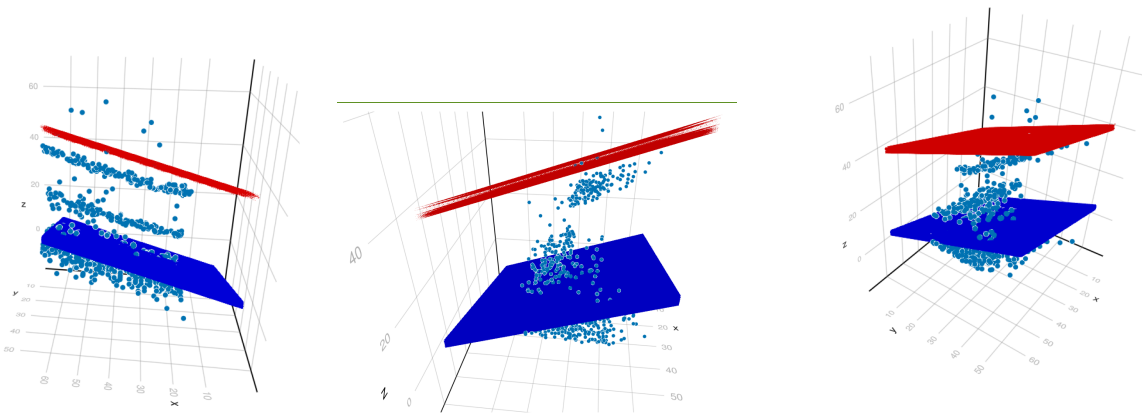


Figure 2: Diferentes planos para el modelo (3)

las conglomeraciones de los planos se hacen solamentente en dos partes y no las tres que esperaríamos, esto ayuda a ver por que los errores tienen ese comportamiento, los datos de "enmedio" no logran modelarse bien ya que no hay algún plano en ninguna de las categorías que lo modele bien, lo siguiente que intenté fue estabilizar la varianza pero me fue imposible intenté cada combinación de transformación sobre la respuesta y nada, los errores siguen teniendo ese problema, y además otro problema que no estamos considerando es que dentro de cada categoría se nota claramente que la varianza no es constante esta cambia para las tres conglomeraciones que se ven en el gráfico y esto no se considera en el modelo

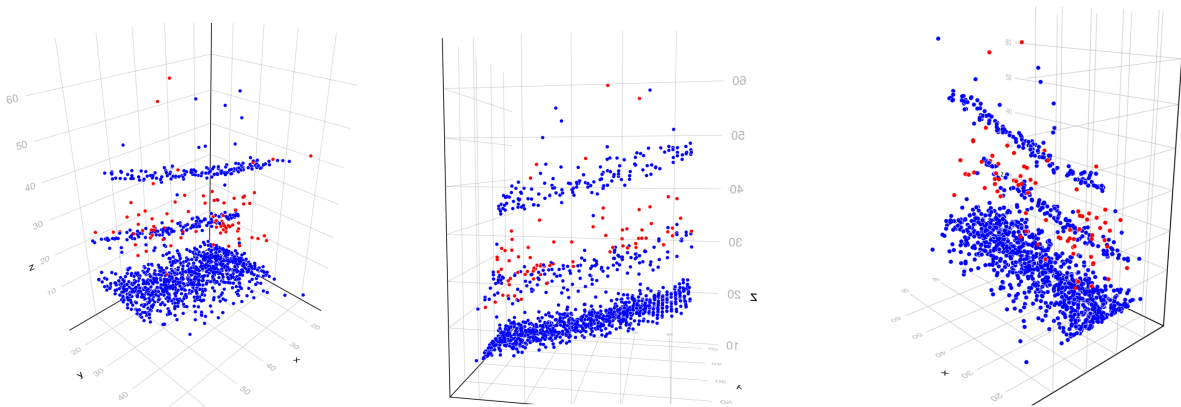


Figure 3: Datos influyentes bajo DFFITS

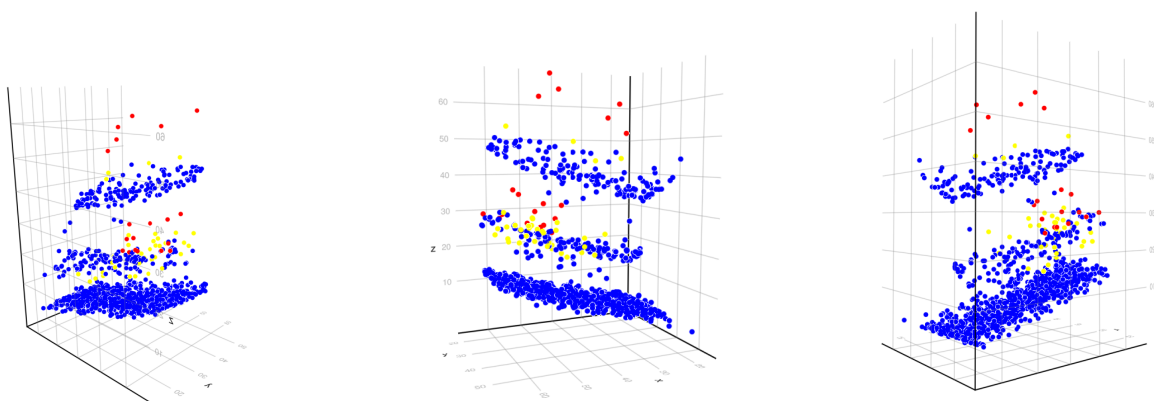


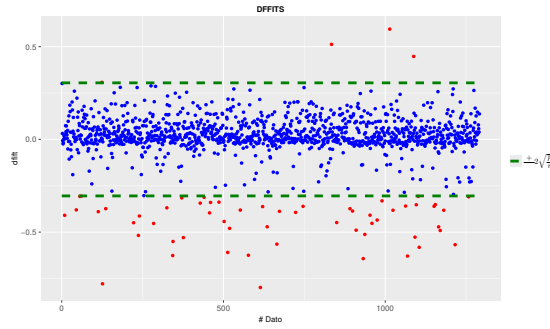
Figure 4: Datos con errores estandar altos

Al analizar el modelo para identificar datos influyentes y aquellos con errores estándar altos, encontramos que, como esperábamos, los puntos más problemáticos se sitúan en el centro del conjunto de datos, con algunos valores atípicos esparcidos. Observamos que los datos con errores estándar mayores a 2 (marcados en amarillo) y aquellos con errores estándar mayores a 3 (marcados en rojo) tienden a estar concentrados en esta región central, mientras que algunos puntos atípicos también muestran errores altos. Esto refuerza nuestra hipótesis inicial de que falta información a nuestro modelo

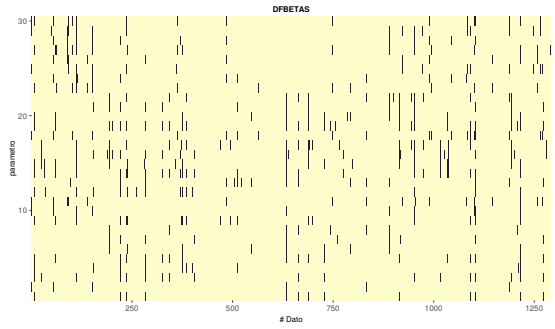
Regresando a nuestro modelo original, ahora vamos a suponer interacciones entre las variables para ver si podemos mejorarlo de alguna manera.

Una idea que se me ocurrió fue analizar cuáles datos en el modelo eran influyentes y tenían errores estándar altos. Esto implica identificar los puntos de datos con un gran impacto en el modelo y altos niveles de error, para ver si al eliminarlos podríamos mejorar el rendimiento del modelo.

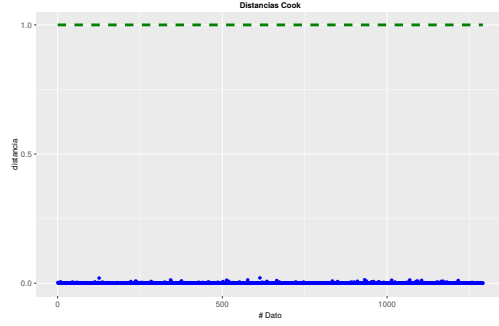
Este enfoque nos permitiría refinar el modelo eliminando posibles valores atípicos o puntos de datos problemáticos,



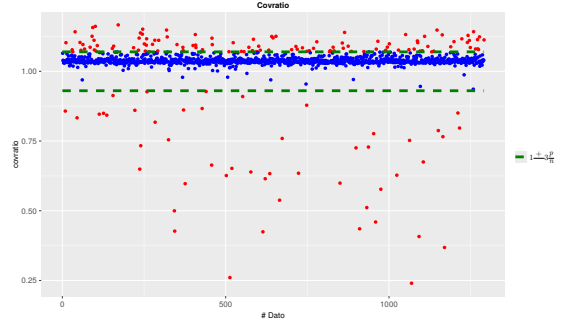
(a) Analisis DFFITS



(b) Analisis DFBETAS



(c) Analisis de Cook's Distance



(d) Analisis de covratio

Figure 5: Analisis de Influencia de datos en el modelo

Después de analizar Cook's distance, DFFITS y otros diagnósticos, parece que hay una discrepancia entre los resultados. Aunque Cook's distance no superó el umbral, los valores de DFFITS indicaron que muchos puntos de datos podrían tener una influencia considerable en los valores ajustados del modelo.

Por lo tanto, tome la decisión de eliminar los 76 puntos de datos que mostraron una alta influencia tanto en el análisis de Covratio como en DFFITS. Esta acción resultó en una mejora significativa en el modelo, aumentando el coeficiente de determinación ( $R^2$ ) de 0.84 a 0.95 y reduciendo ligeramente el error. Sin embargo, aún persisten problemas relacionados con los supuestos de los errores del modelo.

Esta discrepancia entre Cook's distance y DFFITS sugiere que, si bien los puntos de datos no tuvieron un impacto significativo en el ajuste global del modelo según Cook's distance, sí tuvieron un efecto notable en los valores ajustados del modelo según DFFITS. La eliminación de estos puntos de datos influyentes mejoró el ajuste general del modelo, lo que respalda la decisión de eliminarlos.

Aunque la eliminación de estos puntos de datos resultó en una mejora sustancial, es importante tener en cuenta que aún pueden existir problemas relacionados con los supuestos de los errores del modelo.

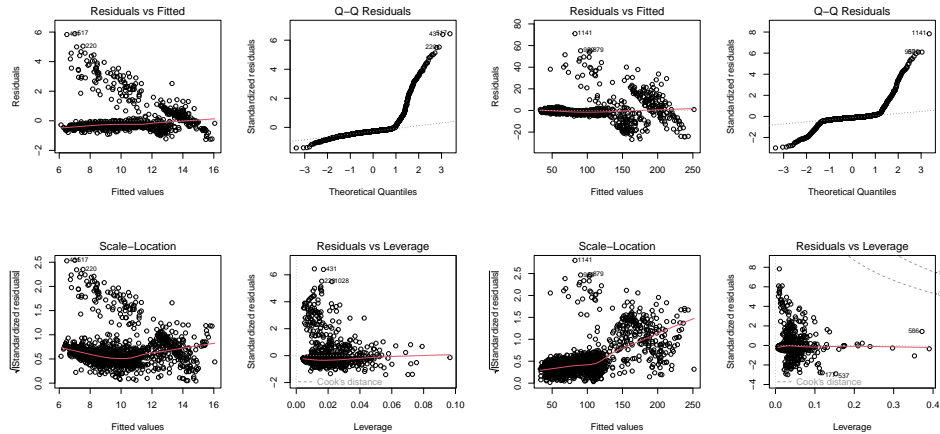


Figure 6: Errores antes y despues de eliminar los 76 datos influyentes



## 4 Resultados y Conclusiones

```
Call
lm(formula = charges ~ smoker + age + children + bmi + region +
sex + smoker:age + smoker:bmi + smoker:children + age:children +
age:sex + age:region + bmi:region + smoker:sex + age:bmi +
smoker:region + smoker:age:children + smoker:age:region,
data = Datos)

Residuals:
    Min       1Q   Median       3Q      Max
-0.45417 -0.18502  0.02003  0.33052  1.56647

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.559e+00  1.161e-03  56.841 < 2e-16 ***
smokeryes    4.559e+00  1.152e-02  39.480 < 2e-16 ***
age          4.579e-02  2.489e-03  18.362 < 2e-16 ***
children     2.897e-01  1.895e-02  15.288 < 2e-16 ***
bmi          2.327e-02  2.770e-03  8.398 < 2e-16 ***
regionnorthwest -7.161e-02  1.075e-03 -6.666 0.00000 ***
regionsouthwest -1.420e-01  1.030e-03 -1.378 0.18546
regionsouthwest 1.500e-01  1.020e-03  1.464 0.00229 ***
sexmale      -2.716e-01  1.820e-02 -14.924 < 2e-16 ***
smokeryes:age -1.821e-02  2.432e-03 -7.487 < 2e-16 ***
smokeryes:bmi  4.921e-02  2.080e-03  23.618 < 2e-16 ***
smokeryes:children -2.674e-01  4.132e-02 -6.470 2.51e-10 ***
age:children  -0.149e-01  8.360e-03 -17.802 < 2e-16 ***
age:sexmale   4.187e-03  8.893e-04  4.708 0.00007 ***
age:regionnorthwest 1.519e-01  1.441e-03  1.112 0.18225
age:regionnorthwest 2.224e-01  1.412e-03  1.575 0.00000 ***
age:regionnorthwest 5.672e-03  1.452e-03  3.899 0.00005 ***
bmi:regionnorthwest -1.288e-01  1.250e-03 -10.294 0.00000 ***
bmi:regionnorthwest -1.802e-02  2.450e-03 -7.352 0.00000 ***
bmi:regionnorthwest -5.679e-03  1.124e-03 -5.050 0.00000 ***
smokeryes:sexmale 2.171e-02  2.890e-02  0.750 0.00129 ***
age:bmi       4.908e-02  2.080e-03  23.618 < 2e-16 ***
smokeryes:regionnorthwest 3.360e-01  1.250e-03  2.688 0.00000 ***
smokeryes:regionnorthwest 1.638e-01  1.312e-03  1.248 0.00000 ***
smokeryes:age:children 3.763e-01  1.010e-03  3.725 0.00000 ***
smokeryes:age:regionnorthwest 1.461e-01  1.200e-03  1.217 0.00000 ***
smokeryes:age:regionnorthwest 6.827e-03  1.001e-03  6.827 0.00000 ***
smokeryes:age:regionnorthwest 1.446e-01  1.380e-03  1.044 0.29707

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.221 on 1262 degrees of freedom
Multiple R-squared:  0.944, Adjusted R-squared:  0.9428
F-statistic: 708 on 22 and 1262 Df, p-value < 2.2e-16
```

programando una pequeña calculadora se puede simplificar mucho el uso del modelo a simplemente meter los datos teniendo también en cuenta que no haya extrapolación, una cosa en la cual tuve problema por que se me dificultó poder obtener a  $X_0$  dado que estoy usando muchas variables categoricas, hubiera quedado mejor usar un intervalo de predicción pero bajo la violación de los supuestos esto no era confiable, me quede simplemente con una estimación puntual que a pesar de que no se cumplen los supuestos esta si la puedo hacer

Este es el resumen del modelo final. Después de aplicar la transformación de raíz cuadrada a la variable de respuesta, eliminar los 47 datos influyentes y considerar todas las interacciones posibles, escribir la fórmula completa del modelo sería poco práctico. El modelo se utilizará para intentar predecir el precio de la prima de seguro basado en las características de un sujeto. Dado lo complejo del modelo, sería más conveniente programar una pequeña calculadora que realice estas predicciones.

Observamos que, a pesar de los inconvenientes con los supuestos del modelo, hemos logrado un  $R^2$  mucho más alto y un error más pequeño. La mejora significativa en el  $R^2$  se debe principalmente a la eliminación de los 47 datos influyentes, lo que ha permitido un ajuste mucho más preciso del modelo.

```
function Calculadora_cobro_prima(edad, fumador, bmi, genero, hijos, Region)
# Convertir los valores booleanos a los formatos correctos para R
fumador_str = fumador ? "yes" : "no"

R"""
library(car)

valor <- predict(forward_selection,
  data.frame(
    age = $edad,
    bmi = $bmi,
    smoker = $fumador_str,
    sex = $genero,
    children = $hijos,
    region = $Region
  ))[1]

return (rccopy(R"valor"))^2
end

Calculadora_cobro_prima(
20,
false,
35,
"male",
0,
"southeast") 1579.1797277049552
```

En conclusión, hemos logrado desarrollar un modelo de regresión decente para estimar de manera puntual los costos de seguro médico. Sin embargo, como responsable de la creación del modelo, consideraría buscar más información para enriquecerlo aún más. Por ejemplo, al investigar en internet o mediante otros medios, podríamos obtener datos adicionales sobre el estado de salud de los asegurados, como la presencia de enfermedades crónicas en sus descendientes o si padecen alguna enfermedad específica. Incluir esta información adicional podría mejorar significativamente la capacidad predictiva del modelo, ya que nos permitiría capturar más aspectos relevantes que influyen en los costos de seguro médico.

## 5 Bibliografia

### References

- [1] Harsh Singh. (2022). Medical Insurance Payout Dataset. Kaggle. Recuperado de <https://www.kaggle.com/datasets/harshsingh2209/medical-insurance-payout>
- [2] JuliaPlots. (2022). Makie Package Documentation. Recuperado de <https://makie.juliaplots.org/stable/>
- [3] R Core Team. (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Recuperado de <https://www.R-project.org/>
- [4] Generalized Linear Models (GLM).jl. (2022). Recuperado de <https://github.com/JuliaStats/GLM.jl>