UNIVERSIDAD
NACIONAL
DE COLOMBIA

# ANÁLISIS DE RELACIÓN ENTRE VARIABLES DEMOGRÁFICAS Y CONSUMO DE TABACO Y ALCOHOL UTILIZANDO LA ENCUESTA NACIONAL DE CONSUMO DE SUSTANCIAS PSICOACTIVAS EN POBLACIÓN GENERAL - ENCSPA- 2019


ALUMNOS:

Cardenas Rosas Maria Camila
Estadística
mcardenasro@unal.edu.co

Conde Hernandez Karen Lucianna
Matemáticas
kconde@unal.edu.co

Forero Laiton Angie Daniela
Matemáticas
anforerol@unal.edu.co


TEMA:
Análisis de Regresión
Regresión Lineal Múltiple

FECHA DE ENTREGA:
4 de junio del 2023

# 1. Descripción.

La base de datos utilizada contiene 2.884 registros de personas que residen en algunos de los municipios del país de entre 12 y 65 años. Se analizarán variables como edad, sexo, nivel educativo, edad en la cual fumaron tabaco por primera vez, frecuencia con que fuman tabaco, edad en la cual bebieron alcohol por primera vez, frecuencia con que beben alcohol, número de tragos habituales y estado de salud con el fin de dar explicación a la problemática de consumo de tabaco y su adicción en Colombia. La base de datos completa de la ENCSPA busca conocer la situación del consumo de drogas; además constituye el fundamento para el diseño de acciones de política públicas, planes programas y proyectos a nivel nacional. En esta parte vamos a seleccionar las variables apropiadas para nuestro modelo y verificaremos todo en el proceso.

## 2. Fase de identificación.

Nos interesa modelar la variable *CIGARR_HAB*. Esta es una variable cuantitativa de escala razón que indica el número aproximado de cigarrillos/tabaco que fuma diariamente el individuo, por tanto, para modelarla se realizó una búsqueda bibliográfica con el fin de tener presente de antemano que variables deberian estar presentes en el modelo.

En esta breve búsqueda, encontramos referencias que nos dicen que el consumo de drogas (legales e ilegales) en las sociedades occidentales es un fenómeno presente más que todo entre los jóvenes, especialmente varones. Nos dice que hay un periodo de riesgo de inicio que es la adolescencia mediana y tardía; y que el consumo precoz de tabaco se ha visto asociado al consumo habitual en la adultez. Esto quiere decir que se mantiene presente un comportamiento según el sexo del individuo y la edad en la que empieza el consumo.

También nos dice que, teniendo en cuenta que es la población joven la que está más vulnerable frente al consumo de drogas legales (entre estas el tabaco), es necesario implementar acciones preventivas y que estas se han mostrado mas efectivas al estar presentes en programas escolares. Esto nos incentivo a ver el comportamiento de la variables *NIV_ED*, que nos indica el nivel educativo de la persona, en nuestro modelo. ¿Puede el nivel educativo y/o tener acceso a estas campañas escolares influir en el consumo del individuo?

Realizaremos una exploración gráfica y calcularemos los coeficientes de correlación lineal, coeficientes de asociación y coeficientes de correlación parciales entre nuestra variable *CIGARR_HAB* y las demás variables para así determinar qué variables guardan relación con nuestra variable respuesta.

Como se tienen distintas variables cualitativas, se realizó el coeficiente de correlación de Pearson únicamente entre las variables cuantitativas, estos se pueden visualizar en la siguiente tabla.

| | *EDAD* | *F_PRIMVEZ* | *B_PRIMVEZ* | *CIGARR_HAB* |
|---|---|---|---|---|
| *EDAD* | 1.0000000 | 0.15538875 | 0.20993888 | 0.14004663 |
| *F_PRIMVEZ* | 0.1553887 | 1.00000000 | 0.27214714 | -0.06098092 |
| *B_PRIMVEZ* | 0.2099389 | 0.27214714 | 1.00000000 | -0.03913901 |
| *CIGARR_HAB* | 0.1400466 | -0.06098092 | -0.03913901 | 1.00000000 |

Observe que a primera vista ninguna de las variables guarda una relación lineal fuerte con la variable *CIGARR_HAB*, lo cual se interpreta como algo normal en un estudio social con tantos individuos, adicionalmente entre ellas (comparando dos a dos) tampoco parece haber una relación lineal fuerte, de hecho se denota una relación "fuerte" en comparación con las demás de las variables *F_PRIMVEZ* y *B_PRIMVEZ*, esto refuerza nuestra idea de que los consumos de estas sustancias están estrechamente relacionados, sin embargo, este proyecto se va a enfocar en la explicación del tabaquismo y los diversos factores que inciden en este de manera separada a otras sustancias, entre ellas el alcohol.

Por otro lado, para ver las demás relaciones con las demás variables cualitativas se realizó el coeficiente de Kendall y Spearman pues este nos puede ayudar a ver que tan fuerte es la relación entre una variable cuantitativa y otra cualitativa ordinal (este coeficiente no se puede hallar para las cualitativas nominales, luego por ejemplo se quita la variable *SEXO*).

Los coeficientes de correlación de Kendall se pueden ver a continuación:

| | *EDAD* | *NIVEL_EDUCA* | *F_FREQ* | *F_PRIMVEZ* | *B_FREQ* |
|---|---|---|---|---|---|
| *EDAD* | 1.00000000 | -0.18668632 | 0.0234766931 | 0.10257063 | 0.06069930 |
| *NIVEL_EDUCA* | -0.18668632 | 1.00000000 | -0.0455242817 | 0.08134748 | -0.03786476 |
| *F_FREQ* | 0.02347669 | -0.04552428 | 1.0000000000 | 0.02128704 | 0.54433610 |
| *F_PRIMVEZ* | 0.10257063 | 0.08134748 | 0.0212870398 | 1.00000000 | 0.02191846 |
| *B_FREQ* | 0.06069930 | -0.03786476 | 0.5443361026 | 0.02191846 | 1.00000000 |
| *B_PRIMVEZ* | 0.20488583 | -0.04479636 | 0.0274711905 | 0.33723855 | 0.04995623 |
| *TRAG_HAB* | -0.18410036 | 0.06485296 | -0.0294201336 | -0.05670929 | -0.07845710 |
| *CIGARR_HAB* | 0.14707941 | -0.07654853 | 0.0004754399 | -0.06588647 | 0.01169850 |
| *ESTADO_SALUD* | -0.14184396 | 0.17376309 | -0.0343157904 | 0.04172309 | -0.04135580 |

| | *B_PRIMVEZ* | *TRAG_HAB* | *CIGARR_HAB* | *ESTADO_SALUD* |
|---|---|---|---|---|
| *EDAD* | 0.20488583 | -0.18410036 | 0.1470794132 | -0.14184396 |
| *NIVEL_EDUCA* | -0.04479636 | 0.06485296 | -0.0765485257 | 0.17376309 |
| *F_FREQ* | 0.02747119 | -0.02942013 | 0.0004754399 | -0.03431579 |
| *F_PRIMVEZ* | 0.33723855 | -0.05670929 | -0.0658864665 | 0.04172309 |
| *B_FREQ* | 0.04995623 | -0.07845710 | 0.0116984966 | -0.04135580 |
| *B_PRIMVEZ* | 1.00000000 | -0.07687057 | -0.0130263667 | -0.02324881 |
| *TRAG_HAB* | -0.07687057 | 1.00000000 | 0.0124701909 | 0.09784867 |
| *CIGARR_HAB* | -0.01302637 | 0.01247019 | 1.0000000000 | -0.04361036 |
| *ESTADO_SALUD* | -0.02324881 | 0.09784867 | -0.0436103613 | 1.00000000 |

Véase que los coeficientes no tienden a ser fuertes con respecto a la variable que tenemos de estudio, ahora entre las variables escogidas con anterioridad ( *EDAD*, *SEXO*, *NIVEL_EDUCA*,*F_PRIMVEZ*) no se observan relaciones fuertes que indiquen multicolinealidad, nuevamente podemos ver que por ejemplo en comparación con los otros coeficientes la correlación entre *B_FREQ* y *F_FREQ* es bastante fuerte al igual que entre *F_PRIMVEZ* y *B_PRIMVEZ*.

Los coeficientes de correlación de Spearman son:

| | *EDAD* | *NIVEL_EDUCA* | *F_FREQ* | *F_PRIMVEZ* | *B_FREQ* |
|---|---|---|---|---|---|
| *EDAD* | 1.00000000 | -0.25779170 | 0.0293868286 | 0.14240146 | 0.07555476 |
| *NIVEL_EDUCA* | -0.25779170 | 1.00000000 | -0.0529923592 | 0.10614080 | -0.04367629 |
| *F_FREQ* | 0.02938683 | -0.05299236 | 1.0000000000 | 0.02580570 | 0.56395917 |
| *F_PRIMVEZ* | 0.14240146 | 0.10614080 | 0.0258057005 | 1.00000000 | 0.02634938 |
| *B_FREQ* | 0.07555476 | -0.04367629 | 0.5639591654 | 0.02634938 | 1.00000000 |
| *B_PRIMVEZ* | 0.28171521 | -0.05808383 | 0.0331928847 | 0.42637064 | 0.06041292 |
| *TRAG_HAB* | -0.24940155 | 0.08334496 | -0.0338730421 | -0.07347128 | -0.09074559 |
| *CIGARR_HAB* | 0.20831962 | -0.10165004 | 0.0006163199 | -0.08971267 | 0.01408445 |
| *ESTADO_SALUD* | -0.18207615 | 0.20517345 | -0.0378382196 | 0.05179489 | -0.04527824 |

| | *B_PRIMVEZ* | *TRAG_HAB* | *CIGARR_HAB* | *ESTADO_SALUD* |
|---|---|---|---|---|
| *EDAD* | 0.28171521 | -0.24940155 | 0.2083196194 | -0.18207615 |
| *NIVEL_EDUCA* | -0.05808383 | 0.08334496 | -0.1016500385 | 0.20517345 |
| *F_FREQ* | 0.03319288 | -0.03387304 | 0.0006163199 | -0.03783822 |
| *F_PRIMVEZ* | 0.42637064 | -0.07347128 | -0.0897126676 | 0.05179489 |
| *B_FREQ* | 0.06041292 | -0.09074559 | 0.0140844463 | -0.04527824 |
| *B_PRIMVEZ* | 1.00000000 | -0.09547500 | -0.0179249510 | -0.02817333 |
| *TRAG_HAB* | -0.09547500 | 1.00000000 | 0.0162557290 | 0.11642515 |
| *CIGARR_HAB* | -0.01792495 | 0.01625573 | 1.0000000000 | -0.05383476 |
| *ESTADO_SALUD* | -0.02817333 | 0.11642515 | -0.0538347561 | 1.00000000 |

Esta tabla tiene casi el mismo comportamiento que la de los coeficientes de Kendall, sin embargo, se podemos ver un aumento significativo en el valor absoluto de las correlaciones con las variables que queremos estudiar cómo posibles explicativas lo cual podría implicar que el modelo puede tener alguna función monótona como transformación de las explicativas, lo cual respalda la escogencia de las variables anteriormente mencionadas.

Ahora, no se debe olvidar que en la regresión múltiple se puede evaluar que tanto esta explicando una variable a otra descontando el efecto que ya explicaron las demás, por tanto se calculan los coeficientes de correlación parciales.

Los coeficientes de correlación parciales de Pearson se calculan nuevamente únicamente entre las variables cuantitativas.

| | *EDAD* | *F_PRIMVEZ* | *B_PRIMVEZ* | *CIGARR_HAB* |
|---|---|---|---|---|
| *EDAD* | 1.0000000 | 0.11385249 | 0.18208317 | 0.15830721 |
| *F_PRIMVEZ* | 0.1138525 | 1.00000000 | 0.24348929 | -0.06937117 |
| *B_PRIMVEZ* | 0.1820832 | 0.24348929 | 1.00000000 | -0.05161285 |
| *CIGARR_HAB* | 0.1583072 | -0.06937117 | -0.05161285 | 1.00000000 |

Así, se visualiza que el efecto de la explicación de las variables una vez ya contempladas las demás no es muy significativo, de hecho ninguna de las variables esta explicando de manera amplia a la variable de interés.

Para las cuantitativas y cualitativas hallamos los parciales de Kendall.

| | *EDAD* | *NIVEL_EDUCA* | *F_FREQ* | *F_PRIMVEZ* | *B_FREQ* |
|---|---|---|---|---|---|
| *EDAD* | 1.00000000 | -1.523131e-01 | -0.015561856 | 0.063507867 | 3.388467e-02 |
| *NIVEL_EDUCA* | -0.15231309 | 1.000000e+00 | -0.032490273 | 0.101858898 | 3.843996e-05 |
| *F_FREQ* | -0.01556186 | -3.249027e-02 | 1.000000000 | 0.016428106 | 5.428088e-01 |
| *F_PRIMVEZ* | 0.06350787 | 1.018589e-01 | 0.016428106 | 1.000000000 | -3.863631e-03 |
| *B_FREQ* | 0.03388467 | 3.843996e-05 | 0.542808770 | -0.003863631 | 1.000000e+00 |
| *B_PRIMVEZ* | 0.16215085 | -4.069308e-02 | -0.002590149 | 0.323584413 | 2.842619e-02 |
| *TRAG_HAB* | -0.15538028 | 2.453147e-02 | 0.016755798 | -0.030863835 | -6.317169e-02 |
| *CIGARR_HAB* | 0.14713987 | -4.199461e-02 | -0.006721068 | -0.064632024 | 8.516673e-03 |
| *ESTADO_SALUD* | -0.09773311 | 1.421365e-01 | -0.013020153 | 0.043802280 | -1.416834e-02 |

| | *B_PRIMVEZ* | *TRAG_HAB* | *CIGARR_HAB* | *ESTADO_SALUD* |
|---|---|---|---|---|
| *EDAD* | 0.162150852 | -0.15538028 | 0.147139866 | -0.09773311 |
| *NIVEL_EDUCA* | -0.040693078 | 0.02453147 | -0.041994614 | 0.14213647 |
| *F_FREQ* | -0.002590149 | 0.01675580 | -0.006721068 | -0.01302015 |
| *F_PRIMVEZ* | 0.323584413 | -0.03086383 | -0.064632024 | 0.04380228 |
| *B_FREQ* | 0.028426191 | -0.06317169 | 0.008516673 | -0.01416834 |
| *B_PRIMVEZ* | 1.000000000 | -0.02484032 | -0.019978288 | -0.00462973 |
| *TRAG_HAB* | -0.024840324 | 1.00000000 | 0.040072843 | 0.07047880 |
| *CIGARR_HAB* | -0.019978288 | 0.04007284 | 1.000000000 | -0.01545334 |
| *ESTADO_SALUD* | -0.004629730 | 0.07047880 | -0.015453337 | 1.00000000 |

Seguido de los coeficientes parciales de Spearmam.

| | *EDAD* | *NIVEL_EDUCA* | *F_FREQ* | *F_PRIMVEZ* | *B_FREQ* |
|---|---|---|---|---|---|
| *EDAD* | 1.00000000 | -0.20973565 | -0.021999196 | 0.08000762 | 0.037318112 |
| *NIVEL_EDUCA* | -0.20973565 | 1.00000000 | -0.040785091 | 0.14366096 | 0.006517840 |
| *F_FREQ* | -0.02199920 | -0.04078509 | 1.000000000 | 0.02282723 | 0.562439553 |
| *F_PRIMVEZ* | 0.08000762 | 0.14366096 | 0.022827235 | 1.00000000 | -0.010840374 |
| *B_FREQ* | 0.03731811 | 0.00651784 | 0.562439553 | -0.01084037 | 1.000000000 |
| *B_PRIMVEZ* | 0.21678588 | -0.04957865 | -0.005263367 | 0.40229145 | 0.032645184 |
| *TRAG_HAB* | -0.21340027 | 0.01576766 | 0.020725433 | -0.03270135 | -0.070439296 |
| *CIGARR_HAB* | 0.21695384 | -0.03521698 | -0.006796831 | -0.08957755 | 0.008376308 |
| *ESTADO_SALUD* | -0.11668118 | 0.15351578 | -0.014422816 | 0.05410290 | -0.011866907 |

| | *B_PRIMVEZ* | *TRAG_HAB* | *CIGARR_HAB* | *ESTADO_SALUD* |
|---|---|---|---|---|
| *EDAD* | 0.2167858778 | -0.213400268 | 0.216953837 | -0.1166811773 |
| *NIVEL_EDUCA* | -0.0495786450 | 0.015767658 | -0.035216976 | 0.1535157776 |
| *F_FREQ* | -0.0052633666 | 0.020725433 | -0.006796831 | -0.0144228161 |
| *F_PRIMVEZ* | 0.4022914545 | -0.032701349 | -0.089577554 | 0.0541028954 |
| *B_FREQ* | 0.0326451839 | -0.070439296 | 0.008376308 | -0.0118669071 |
| *B_PRIMVEZ* | 1.0000000000 | -0.005558271 | -0.035994809 | 0.0004137704 |
| *TRAG_HAB* | -0.0055582714 | 1.000000000 | 0.068954824 | 0.0734598119 |
| *CIGARR_HAB* | -0.0359948092 | 0.068954824 | 1.000000000 | -0.0067600241 |
| *ESTADO_SALUD* | 0.0004137704 | 0.073459812 | -0.006760024 | 1.0000000000 |

Realizando una revisión de las anteriores dos tablas se observa que las variables que conservan relaciones más fuertes ya sean negativas o positivas con la variable respuesta son las ya escogidas, por tanto no se agregan más variables, además con respecto a la variable *F_PRIMVEZ* se visualiza una relación mas fuerte monótona que lineal, sin embargo el cambio no es tan significativo, luego para ver si se necesita colocar alguna transformación de esta variable nos guiaremos del siguiente gráfico.

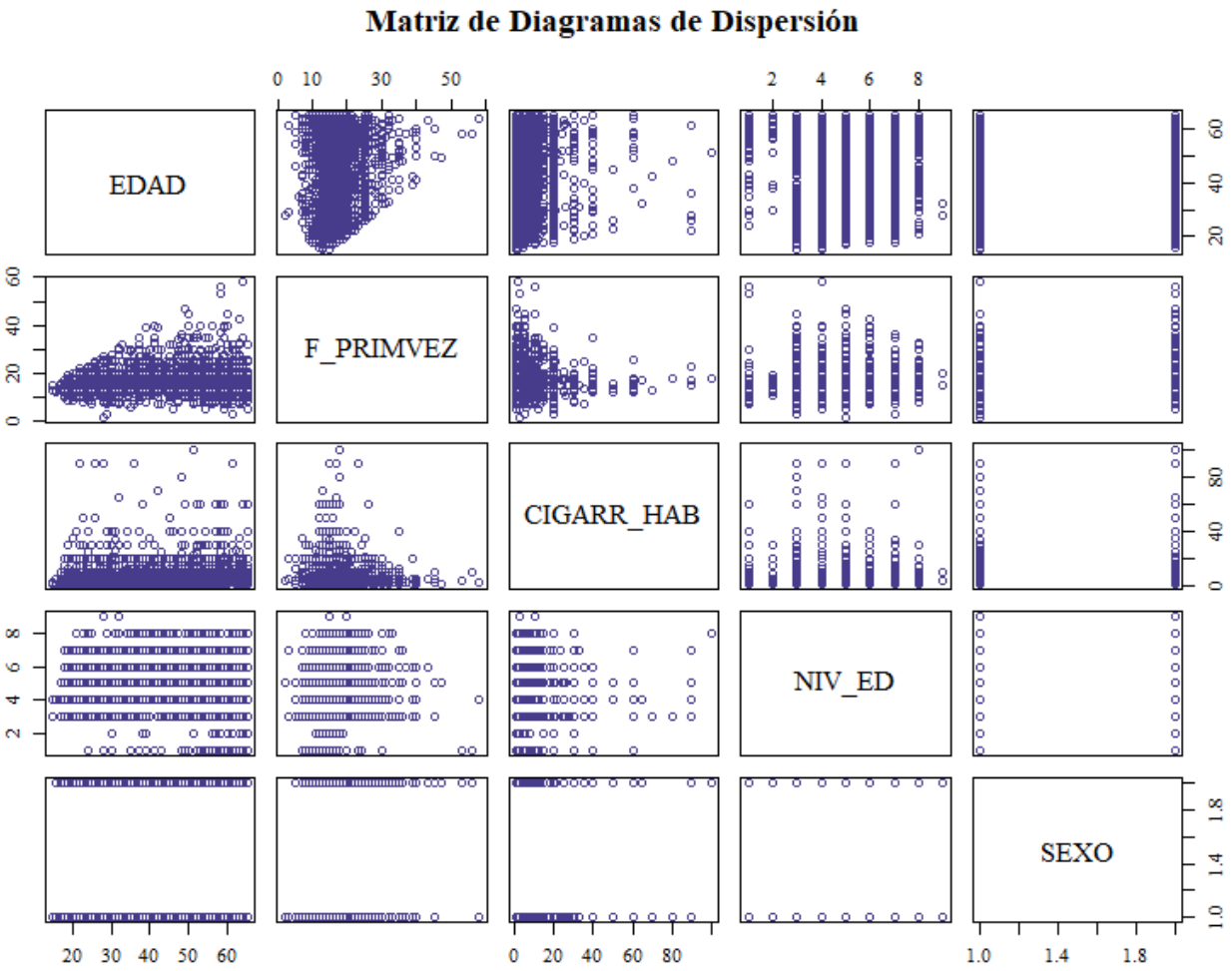



Figura 1.

En esta gráfica se puede ver que a pesar de que *F_PRIMVEZ* parece tener un comportamiento exponencial con respecto a la variable *CIGARR_HAB* el grueso de las observaciones no se encuentra en dicho comportamiento, por tanto escogemos colocar esta variable linealmente en el modelo, además no se observan relaciones a simple vista entre las otras variables escogidas, por tanto se sospecha la no presencia de multicolinealidad.

## 3. Fase de identificación/Estimación.

El modelo final construido es un modelo lineal con las variables explicativas dadas por la *EDAD*, *F_PRIMVEZ*, *NIV_ED* y *SEXO*. Ahora, este modelo se puede estructurar de la siguiente forma:

Sea

$$Y_k := \text{número de cigarrillos habituales que consume el individuo } k \text{ diariamente.}$$

Sean $x_{k1}$ y $x_{k2}$ las variables que guardan la edad del individuo $k$ y la edad de primera vez de consumo de tabaco del individuo respectivamente, por otro lado, como *NIV_ED* es una variable cualitativa que tiene $9$ categorías distintas surgen $8$ variables dummy que se agregan al modelo, las cuales se llaman $x_{k3}$, $x_{k4}, \cdots,$ $x_{k10}$ y que toman como referencia la categoría $1$, es decir, la categoría en donde el individuo no tienen ningún nivel educativo, por último para incluir el *SEXO* se agrega una variable dummy adicional que se llama $x_{k11}$ y que toma como referencia la categoría $1$, es decir, esta variable toma $0$ si el individuo es hombre y toma $1$ si el individuo es mujer.

Entonces, la estructura de nuestro modelo está dada por:

$$\begin{cases} Y_k = \mu_k + e_k \\ \mu_k = \beta_0 + \beta_1 x_{k1} + \beta_2 x_{k2} + \beta_3 x_{k3} + \beta_4 x_{k4} + \beta_5 x_{k5} + \beta_6 x_{k6} + \beta_7 x_{k7} + \beta_8 x_{k8} + \beta_9 x_{k9} + \beta_{10} x_{k10} + \beta_{11} x_{k11} \\ e_k \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

Así, al realizar el método de los mínimos cuadrados ordinarios para estimar el anterior modelo lineal se obtienen los siguientes resultados:

| | Estimate | Std. Error | t value | Pr(> \|t\|) |
|---|---|---|---|---|
| (Intercept) | 7.47138 | 1.29138 | 5.786 | 8.01e-09 *** |
| *EDAD* | 0.09401 | 0.01269 | 7.411 | 1.64e-13 *** |
| *F_PRIMVEZ* | -0.13953 | 0.03473 | -4.017 | 6.04e-05 *** |
| *NIV_ED2* | -3.50169 | 2.52311 | -1.388 | 0.1653 |
| *NIV_ED3* | -0.54806 | 1.08236 | -0.506 | 0.6126 |
| *NIV_ED4* | 0.27816 | 1.09378 | 0.254 | 0.7993 |
| *NIV_ED5* | -1.19907 | 1.08058 | -1.110 | 0.2672 |
| *NIV_ED6* | -1.80850 | 1.13561 | -1.593 | 0.1114 |
| *NIV_ED7* | -0.91565 | 1.12671 | -0.813 | 0.4165 |
| *NIV_ED8* | -2.08383 | 1.37538 | -1.515 | 0.1299 |
| *NIV_ED9* | -0.99405 | 6.19715 | -0.160 | 0.8726 |
| *SEXO2* | -0.71177 | 0.33972 | -2.095 | 0.0362 * |

| Residual standard error: 8.634 on 2872 degrees of freedom | |
|---|---|
| Multiple R-squared: 0.03476 | Adjusted R-squared: 0.03106 |
| F-statistic: 9.402 on 11 and 2872 DF | p-value: < 2.2e-16 |

Ahora, realizando bootstraping de los residuos se obtienen los resultados para las estimaciones que se siguen.

```
Number of bootstrap replications R = 20000
```

| | original | bootBias | bootSE | bootMed |
|---|---|---|---|---|
| (Intercept) | 7.471381 | -5.5851e-04 | 1.287130 | 7.425001 |
| EDAD | 0.094009 | -2.9667e-05 | 0.012611 | 0.094063 |
| F_PRIMVEZ | -0.139527 | 1.1244e-04 | 0.035052 | -0.140071 |
| NIV_ED2 | -3.501690 | 6.1757e-03 | 2.525969 | -3.782258 |
| NIV_ED3 | -0.548057 | 5.7544e-03 | 1.079932 | -0.474564 |
| NIV_ED4 | 0.278162 | 1.4263e-04 | 1.090264 | 0.344021 |
| NIV_ED5 | -1.199066 | -2.7149e-04 | 1.078716 | -1.129938 |
| NIV_ED6 | -1.808504 | -5.5805e-05 | 1.133304 | -1.739745 |
| NIV_ED7 | -0.915652 | -5.2515e-04 | 1.122630 | -0.848354 |
| NIV_ED8 | -2.083829 | -7.2174e-03 | 1.371941 | -2.071785 |
| NIV_ED9 | -0.994047 | 7.5001e-02 | 6.329986 | -2.380639 |
| SEXO2 | -0.711772 | -1.3412e-03 | 0.340141 | -0.715879 |

Adicionalmente, para la desviación estándar del modelo se obtiene:

| | R | original | bootBias | bootSE | bootMed |
|---|---|---|---|---|---|
| V1 | 20000 | 8.6335 | -0.012871 | 0.4351 | 8.6106 |

Es decir, las estimaciones de los parámetros del modelo planteado están dadas por:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $\hat{\beta}_0$ | 7.425001 | $\hat{\beta}_3$ | -3.782258 | $\hat{\beta}_6$ | -1.129938 | $\hat{\beta}_9$ | -2.071785 |
| $\hat{\beta}_1$ | 0.094063 | $\hat{\beta}_4$ | -0.474564 | $\hat{\beta}_7$ | -1.739745 | $\hat{\beta}_{10}$ | -2.380639 |
| $\hat{\beta}_2$ | -0.140071 | $\hat{\beta}_5$ | 0.344021 | $\hat{\beta}_8$ | -0.848354 | $\hat{\beta}_{11}$ | -0.715879 |

A continuación se muestran las interpretaciones de los coeficientes del modelo:

• El coeficiente $\hat{\beta}_0$ representa la cantidad promedio de cigarrillos que consume un individuo diariamente cuando la respuesta a las demás variables es $0$. Es un poco confuso dar una interpretación más allá de esto, pues sería el caso de un individuo de edad $0$ que fumó por primera vez a los $0$ años.(En particular su sexo es masculino y no tiene ningún nivel educativo).

• Por otro lado, $\hat{\beta}_1$ va a representar el cambio promedio, ocasionado por el cambio de un año de edad, en el total de cigarrillos diarios que consume un individuo cuando se mantienen todas las demás variables constantes.

• Análogamente, $\hat{\beta}_2$ va a representar el cambio promedio, ocasionado por el cambio de un año en la edad de primera vez que fumó, en el total de cigarrillos diarios que consume un individuo cuando se mantienen todas las demás variables constantes.

• $\hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6, \hat{\beta}_7, \hat{\beta}_8, \hat{\beta}_9$, y $\hat{\beta}_{10}$ representarán el aumento o disminución promedio en la cantidad de cigarrillos que consume diariamente un individuo que cuenta con un nivel educativo específico respecto a los que no tienen ningún nivel educativo, si se mantienen las demás variables constantes

• Además $\hat{\beta}_{11}$ representará la disminución promedio en la cantidad de cigarrillos que consume diariamente una mujer con respecto a la cantidad promedio de un hombre, cuando las demás variables se mantienen constantes.

• Finalmente $\hat{\sigma} = 8.610$ es la distancia promedio entre el número de cigarrillos que consume un individuo al día y la estimación dada por el modelo del número de cigarrillos que consume.

# 4. Fase de Validación.

- **Patrones inexplicados:** ¿Hay patrones no explicados en los residuales? ¿Qué evidencia estadística (gráficos o pruebas) tienen? ¿Hicieron algo durante el proceso de modelación para corregir problemas con ese supuesto? Si no se cumple el supuesto, ¿qué implicaciones tiene en los resultados y la utilidad del modelo?

  **Reporte:**

  Inicialmente se puede observar que no hay presencia de patrones no explicados en el modelo, pues para el modelo inicial se obtenían los siguientes gráficos del comportamiento de los residuales respecto a los valores ajustados del modelo y las diferentes variables explicativas:

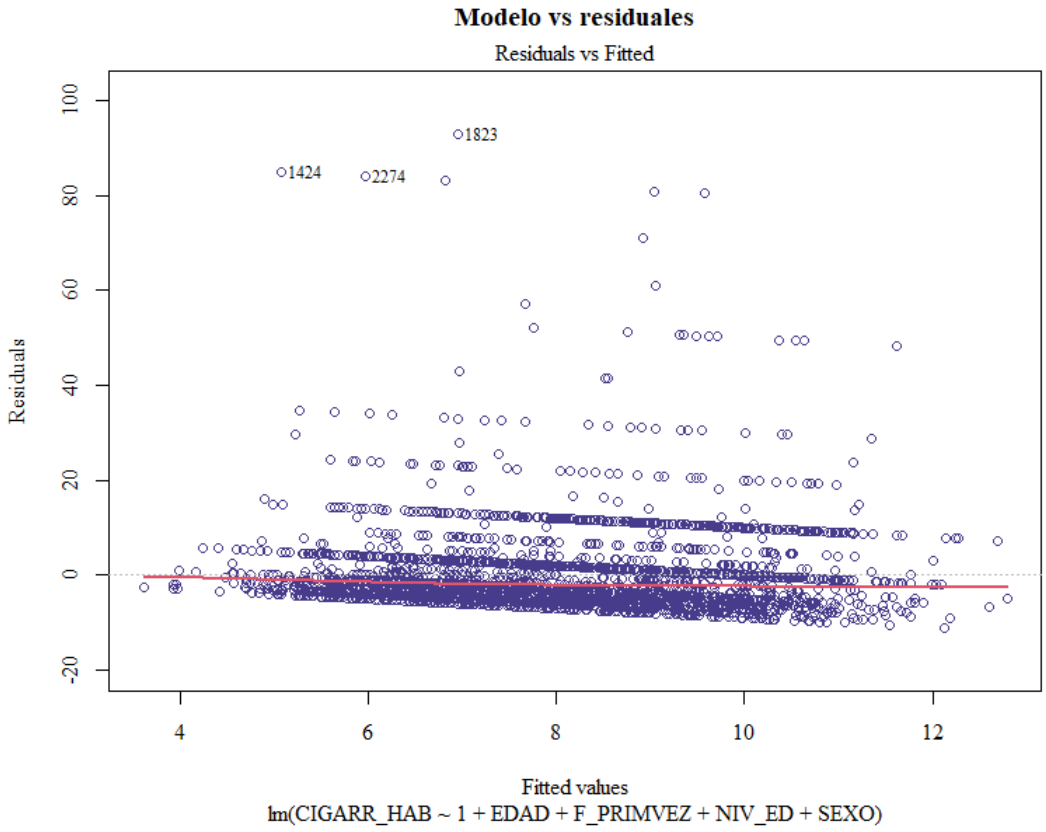
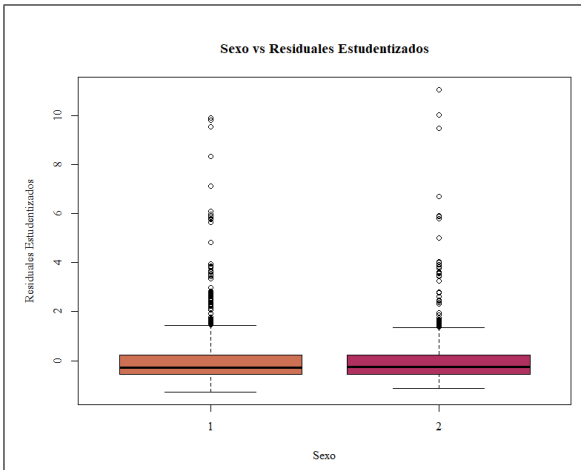
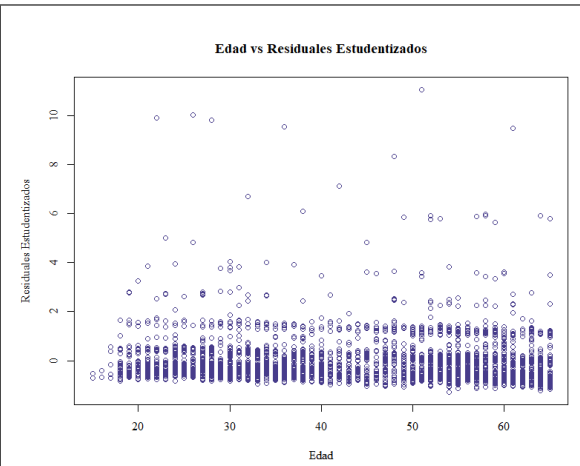
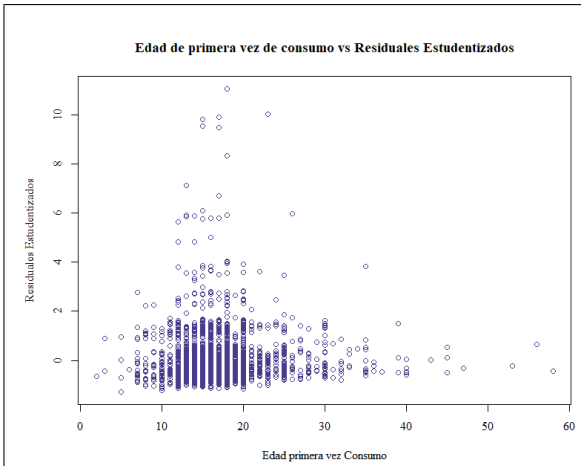
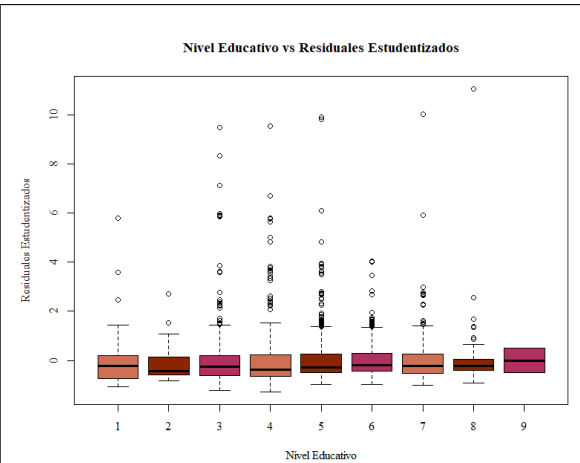



Figura 2.

En esta gráfica se puede ver que, aunque al parecer la variabilidad de los errores no es la misma para todos ellos, su media en efecto esta concentrada al rededor del 0 y que adicionalmente no se ven patrones con respecto a esta media, por tanto se puede concluir que no hay patrones no explicados en dicha gráfica.

Ahora, en la siguiente tabla se pueden ver las gráficas de los residuales estudentizados respecto a las variables explicativas, es así que tanto para las variables cuantitativas como para las cualitativas se puede observar que la media se encuentra aproximadamente al rededor del 0 y que no hay patrones no explicados al rededor de dicha media.

| Sexo - Residuales Estudentizados | Edad -Residuales Estudentizados |
|---|---|
|  |  |
| **Edad de F primera vez - Residuales Est** | **Nivel Educativo - Residuales Estudentizados** |
|  |  |



Por último, se puede ver un indicio de heteroscedasticidad reflejado sobre todo en la gráfica Edad de F primera vez- Residuales Est en donde se puede ver un comportamiento de campana para los residuales, es decir, al parecer la variabilidad en el centro de los datos es mayor que en los extremos para dicha variable, el supuesto de la homoscedasticidad se ilustrará más adelante.

Es de importancia recalcar que se realizo un procedimiento de Bootstraping de los residuales y de allí se obtuvieron nuevas estimaciones de los parámetros, estas aunque son cercanas a las anteriormente estimadas es preciso mostrar si afectan o no los comportamientos anteriores, para ello, con el fin de verificar que dicho remuestreo no afecta de ninguna manera este supuesto se realiza lo siguiente en $\mathcal{R}$ para hallar los nuevos residuales y residuales estudentizados debidos a este pequeño cambio en las estimaciones:

```
betahat<-summary(modelodef)$bootMed
Yhat<-X%*%betahat
Res<-db4$CIGARR_HAB-Yhat
Restudent<-numeric(a)
for(i in 1:length(Res)){ Restudent[i]<-(Res[i]/
(sigmahatnew*(1-Apalancamiento[i])))
#Vector de estudentizados}
```

Luego, la nueva gráfica de los residuales respecto al modelo es la siguiente:

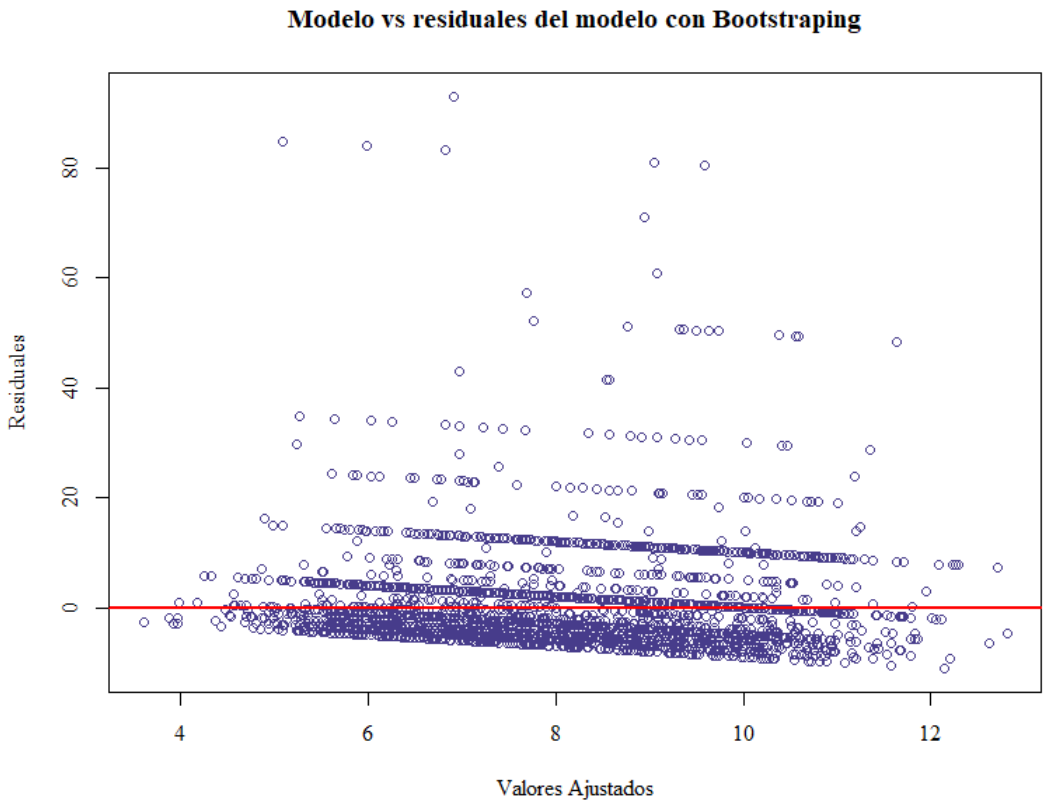
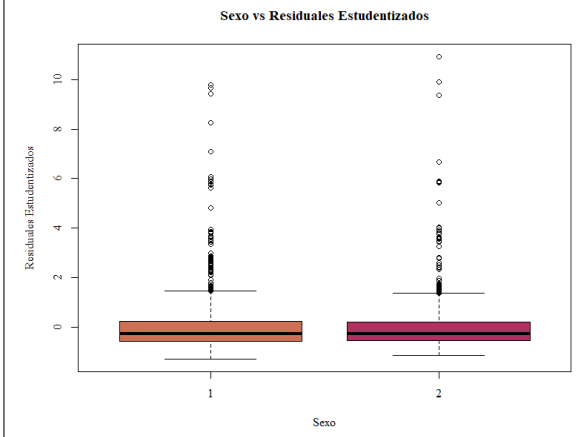
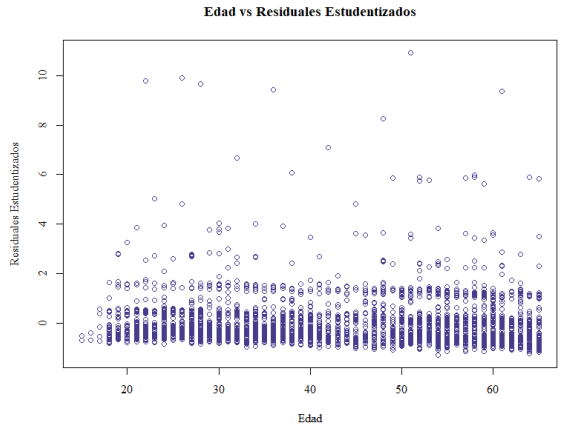
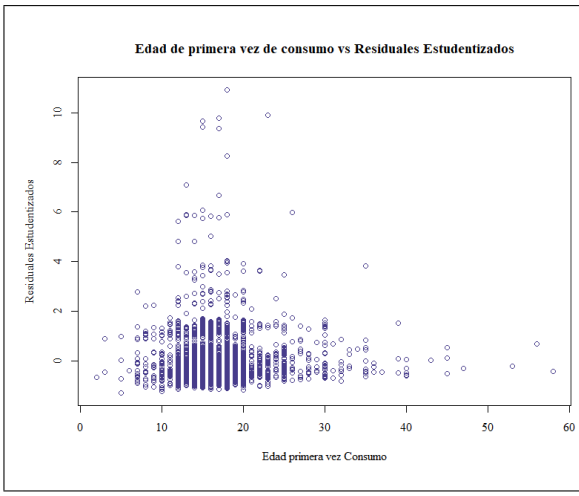
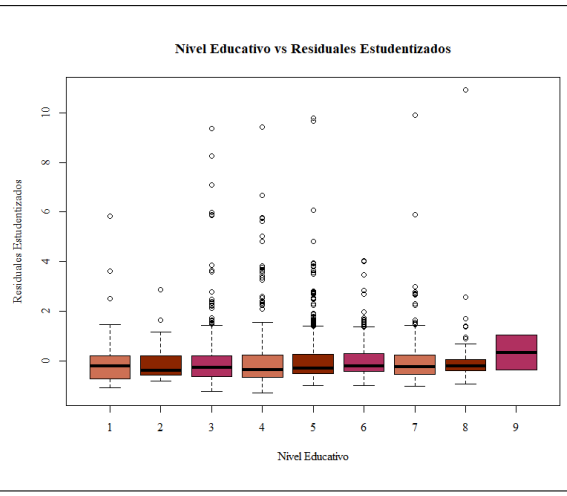



Figura 3.

Se pueden observar cambios bastante mínimos con respecto a la gráfica original, la media de dichos residuales sigue tendiendo a 0 y su comportamiento es prácticamente idéntico, de aquí podemos concluir dos cosas, en primer lugar como el bootstraping no cambia de manera sustancial los coeficientes del modelo inicial es lógico que no cambie de manera sustancial los residuales iniciales, por tanto aunque estos no son iguales si tienden a serlo y en segundo lugar al igual que en las gráficas mostradas anteriormente estas últimas no tienen patrones no explicados en los residuales.

Observe que, en la siguiente tabla se encuentra el comportamiento de los residuales estudentizados nuevos con respecto a las variables explicativas del modelo, se concluye que de igual manera se tiene un comportamiento casi idéntico al inicial y que por tanto el modelo final carece de patrones no explicados en los residuales.

| Sexo - Residuales Estudentizados | Edad -Residuales Estudentizados |
|---|---|
|  |  |
| **Edad de F primera vez - Residuales Est** | **Nivel Educativo - Residuales Estudentizados** |
|  |  |



Así, se termina de verificar que la diferencia entre los residuales resultantes del modelo sin bootstraping y los residuales resultantes de los coeficientes con bootstraping tiende a ser $0$.

- **Multicolinealidad:** ¿Hay problemas de multicolinealidad en el modelo? ¿Qué evidencia estadística (gráficos o pruebas) tienen? ¿Hicieron algo durante el proceso de modelación para corregir problemas con ese supuesto? Si no se cumple el supuesto, ¿qué implicaciones tiene en los resultados y la utilidad del modelo?

**Reporte:**

EL modelo no presenta problemas de multicolinealidad, para garantizar esto se calcularon los factores de inflación de varianza para cada una de las variables explicativas, obteniéndose así lo siguiente:

```
            GVIF      Df     GVIF^(1/(2*Df))
EDAD       1.210859    1          1.100390
F_PRIMVEZ  1.056654    1          1.027937
NIV_ED     1.173241    8          1.010036
SEXO       1.042087    1          1.020827
```

Se sabe con anterioridad que si para alguna de estas variables se tiene un `GVIF^(1/(2*Df))`

al cuadrado que es mayor a $5$, entonces hay multicolinealidad leve y si es mayor a $10$ entonces hay multicolinealidad severa, adicionalmente se sabe que `GVIF^(1/(2*Df))=VIF` para las variables cuantitativas.

Así, se observa que en las variables cualitativas *SEXO* y *NIV_ED* el factor `GVIF^(1/(2*Df))` es $1.020827$ y $1.010036$ respectivamente, luego sus cuadrados son $1.042088$ y $1.020173$ y por tanto estos son menores que $5$, por otro lado, para las variables cuantitativas *EDAD* y *F_PRIMVEZ* se tiene que `GVIF^(1/(2*Df))` es $1.100390$ y $1.027937$ respectivamente, de donde sus cuadrados son $1.210858$ y $1.056654$ de donde se tiene que estos son menores que $5$ y por tanto se descarta la multicolinealidad en el modelo.

Este supuesto también se puede garantizar calculando el determinante de la matriz en cuestión;

```
det(t(model.matrix(modelo1))%*%model.matrix(modelo1))
```

El cual es $1.285308^{32}$ lo cual descarta que esta matriz sea singular y por tanto la matriz $\boldsymbol{X}$ es de rango completo columna.

Adicionalmente, también se calculo el número de condición de la matriz, se sabe que este número es el cociente entre el valor propio mas grande y el valor propio más pequeño, es así que si tiende a infinito significa que la matriz tiende a ser singular, para este caso el número de condición da como resultado $3.381579^{6}$, sin embargo, aunque este número es grande creemos que puede ser por las unidades en las cuales se encuentran las observaciones y que por ello se puede estar alterando dicha medición. Por tanto, se concluye que el modelo carece de problemas de multicolinealidad y que no se necesita ninguna corrección para dicho supuesto.

- **Homoscedasticidad:** ¿Hay problemas de heteroscedasticidad en los residuales? ¿Qué evidencia estadística (gráficos o pruebas) tienen? ¿Hicieron algo durante el proceso de modelación para corregir problemas con ese supuesto? Si no se cumple el supuesto, ¿qué implicaciones tiene en los resultados y la utilidad del modelo?

  **Reporte:**

  No hay evidencia de heteroscedasticidad en el modelo, en un principio se tienen las siguientes gráficas de los residuales:

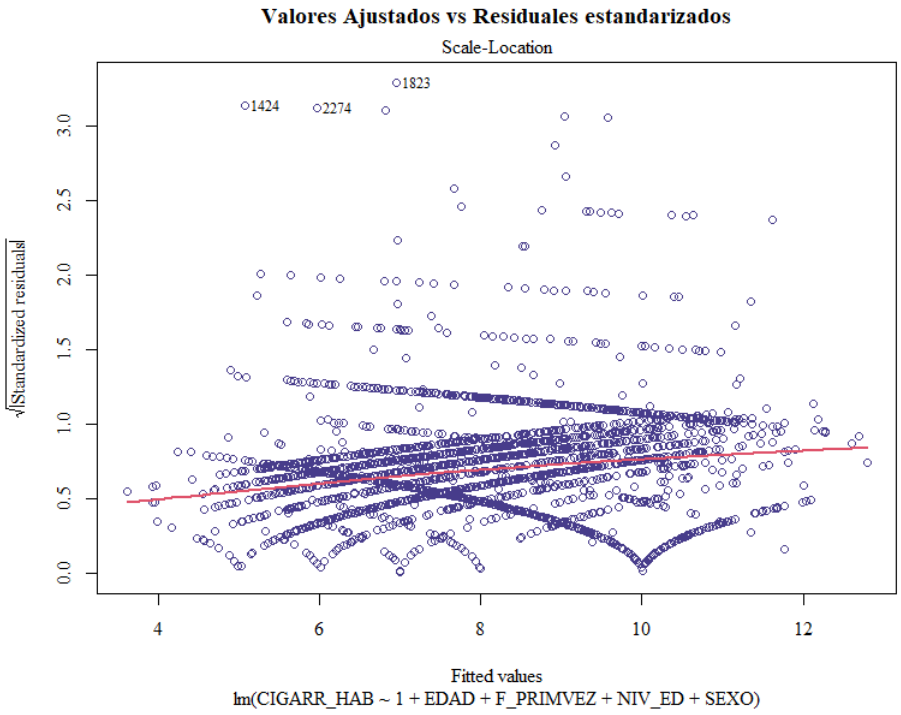



Figura 4.

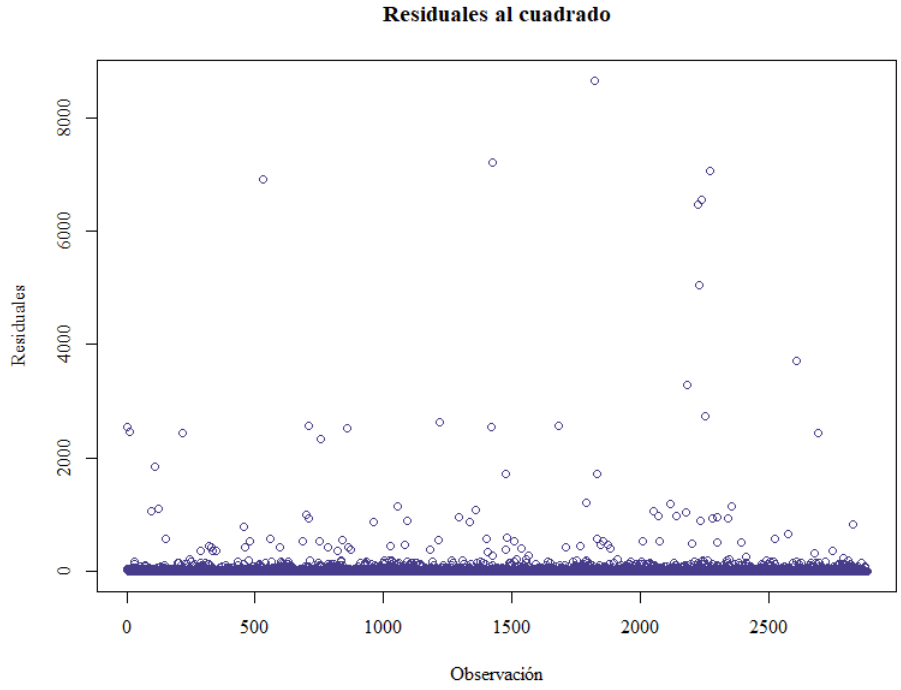



Figura 5.

Adicionalmente de estas, también se tienen las presentadas en el supuesto de multicolinealidad. De aquí pareciera que el supuesto de la homoscedasticidad no se cumple en el modelo inicial, por tanto se realizó la prueba bptest (Breusch-Pagan, White) para confirmar o no esta hipótesis, dicha prueba se realizo con una significancia del $5\,\%$ en donde el sistema de hipótesis tiene la siguiente estructura:

$$\begin{cases} H_0 : \text{Se tiene presencia de homocedasticidad.} \\ \textbf{vs} \\ H_1 : \text{No se tiende presencia de homoscedasticidad.} \end{cases}$$

Con ayuda del código `bptest(modelo1)` de $\mathcal{R}$ se realiza la prueba de hipótesis, el cual arroja lo siguiente:

```
        studentized Breusch-Pagan test
data:  modelo1
BP = 7.9065, df = 11, p-value = 0.7217
```

En donde, con un $5\,\%$ de significancia no se puede rechazar la hipótesis nula pues el $p - valor$ de esta prueba es bastante grande y por consiguiente no es menor a $0.05$. Es así que el supuesto de homoscedasticidad si se tiene para este modelo inicial y que por tanto se concluye que el comportamiento visualizado en los gráficos se debe a que el grueso de los datos en realidad tiene residuales que se comportan o que tienden a comportarse con la misma variabilidad pero hay algunos pocos que son los que más se visualizan que tienen variabilidades distintas.

Por otro lado, se realizo una técnica de remuestreo y gracias a las observaciones dichas en los anteriores supuestos se concluyó que los residuales generados por este son bastante parecidos a los anteriores, es decir, se puede intuir que realizar este proceso no afecta de ninguna manera que se siga cumpliendo la homoscedasticidad y que por tanto el modelo final también la posea. Para verificar las anteriores afirmaciones se realizo lo siguiente:

Primero, se realizaron las mismas gráficas que para el primer modelo pero con los nuevos residuales, estas gráficas son:

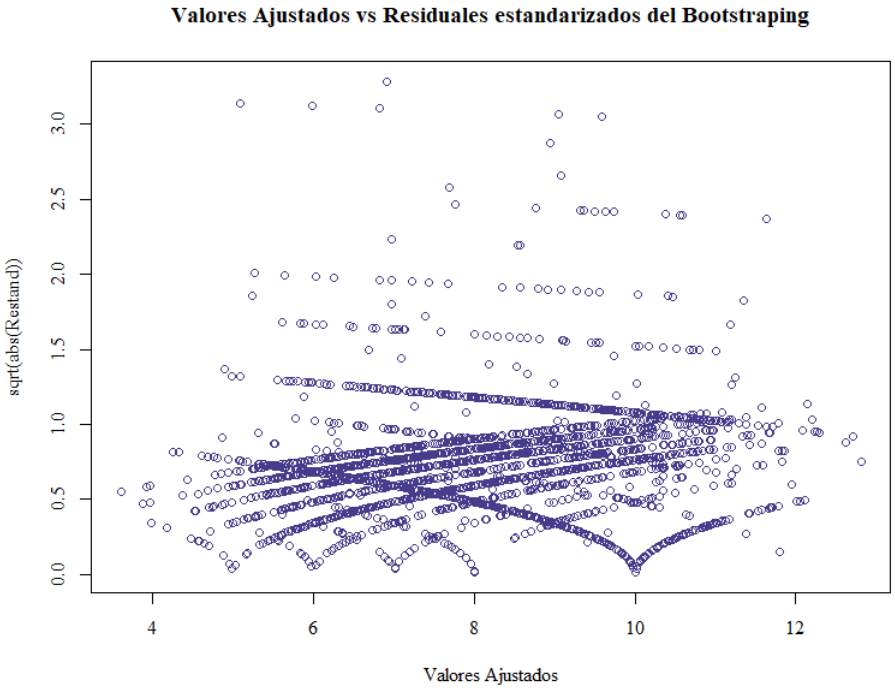



Figura 6.

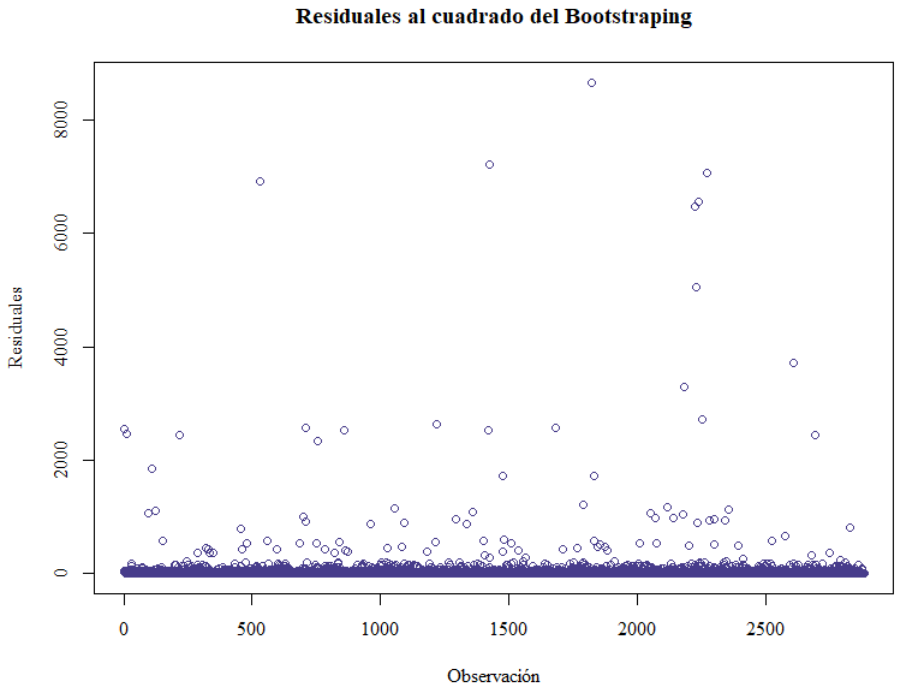



Figura 7.

En estas, se ven pequeños cambios pero el comportamiento es prácticamente el mismo que el ilustrado con anterioridad. En segundo lugar se realiza una prueba de hipótesis de Breusch-Pagan con el $5\,\%$ de significancia tomando un modelo de la siguiente forma:

$$r_i^2 = \delta_0 + \delta_1 x_{k1} + \delta_2 x_{k2} + \delta_3 x_{k3} + \delta_4 x_{k4} + \delta_5 x_{k5} + \delta_6 x_{k6} + \delta_7 x_{k7} + \delta_8 x_{k8} + \delta_9 x_{k9} + \delta_{10} x_{k10} + \delta_{11} x_{k11}$$

En donde $r_i$ es el residual debido a la observación $i$, así se tendrá homoscedasticidad si el p-valor que arroja la salida de $\mathcal{R}$ de significancia de la regresión de este modelo es mucho mayor que $0{,}05$, para ello nos ayudamos del siguiente código:

```
fit_var4<-lm((Res^2)~1+EDAD+F_PRIMVEZ+NIV_ED+SEXO,data=db4)
summary(fit_var4)
```

El cual arroja un $p-$valor de significancia de la regresión de 0.7227 que es mucho mayor a la significancia escogida y que por tanto no podrá rechazar el hecho de que el modelo es homoscedastico con el bootstraping realizado.

- **Normalidad:** ¿Hay problemas de no normalidad en los residuales? ¿Qué evidencia estadística (gráficos o pruebas) tienen? ¿Hicieron algo durante el proceso de modelación para corregir problemas con ese supuesto? Si no se cumple el supuesto, ¿qué implicaciones tiene en los resultados y la utilidad del modelo?

**Reporte:**

Si hay problemas de normalidad en los residuales, como se podrá ver en la siguiente parte del proyecto nuestro modelo tiene demasiados valores de alto leverage, por tanto para ver la normalidad o no normalidad de los residuales en un inicio se realizo un $QQ-plot$ que contribuyo

a observar que tanto se ajustaban dichos residuales a una distribución normal (observe que las pruebas de Shapiro y Jarque-Bera no son útiles aquí por la cantidad de leverage que se tiene en el modelo), por tanto se utilizó el envelope simulado, el cual nos da como resultado la siguiente gráfica:

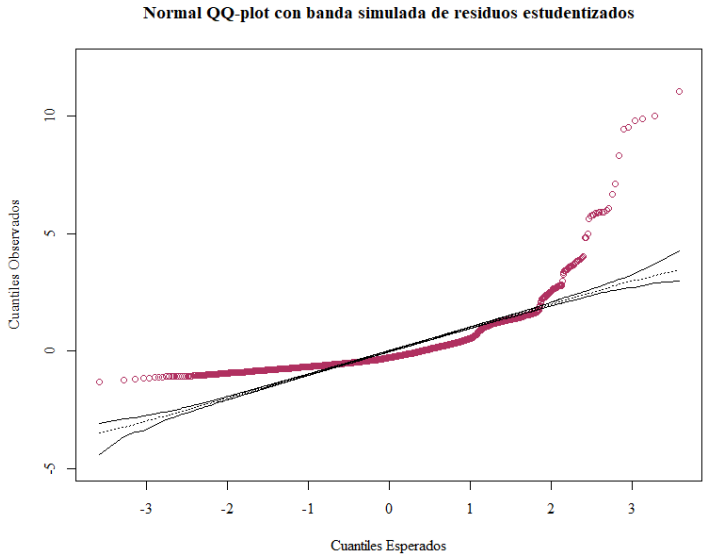



Figura 8.

Es evidente que los residuales no están teniendo un comportamiento normal, esto puede afectar la inferencia realizada sobre los diferentes coeficientes, por ejemplo la realización de intervalos de confianza e intervalos predictores podría verse entorpecida pues ya no se podría garantizar la normalidad de la cantidad pivote. Así, para rectificar esto se realizo Bootstraping sobre los residuales, no sobre las observaciones pues se quiere utilizar la información recolectada por el estudio de una forma fehaciente, ahora es importante tener en cuenta que esto no nos brinda una solución al problema de no normalidad, de hecho y como se podrá ver en el siguiente gráfico se puede ver que el nuevo modelo con las estimaciones derivadas del bootstraping sigue sin comportarse normalmente.

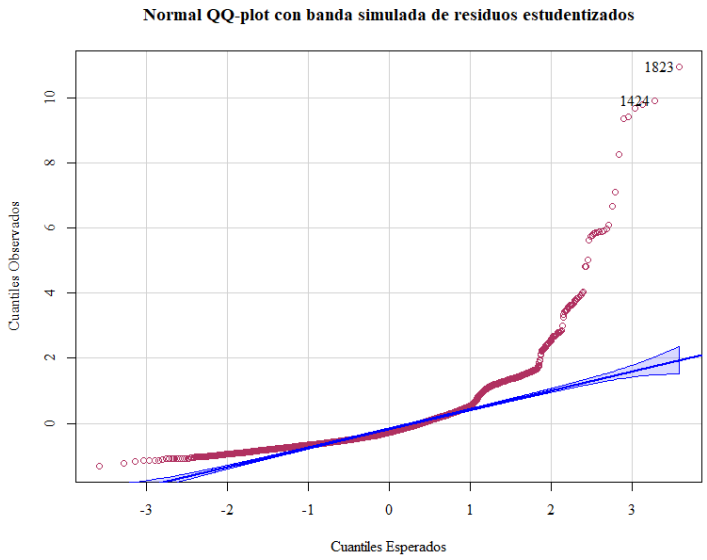



Figura 9.

Pero, con la realización de este remuestreo ya se puede realizar inferencia y utilizar el modelo para dar explicación a diferentes preguntas, adicionalmente al tener un tamaño de muestra de $2.884$ individuos podemos asumir la normalidad asintoticamente y utilizar los $p-$valores de los coeficientes del modelo inicial.

- **Independencia de los errores:** (sólo si se sabe que las observaciones fueron reportadas en orden temporal) ¿Hay problemas de autocorrelación en los residuales? ¿Qué evidencia estadística (gráficos o pruebas) tienen? ¿Hicieron algo durante el proceso de modelación para corregir problemas con ese supuesto? Si no se cumple el supuesto, ¿qué implicaciones tiene en los resultados y la utilidad del modelo?

**Reporte:**

En este caso no se puede garantizar que las observaciones tales como fueron tomadas hallan sido registradas en un orden temporal, por tanto no se puede realizar gráficos de los residuales a través del tiempo ni la prueba de Durbin-Watson, pero se puede realizar la prueba de rachas de independencia.

Esta prueba de rachas tiene el siguiente sistema de hipótesis:

$$\begin{cases} H_0 : \text{Los datos provienen de una muestra aleatoria} \\ \textbf{vs} \\ H_1 : \text{Los datos no provienen de una muestra aleatoria} \end{cases}$$

Entonces, tomando una significancia del $5\,\%$ para realizar esta prueba, se obtiene que utilizando el código `x-factor(sign(stud_res))` seguido de `runs.test(x)` se obtiene la siguiente salida de $\mathcal{R}$:

```
                    Runs Test

data:  x
Standard Normal = 0.32093, p-value = 0.7483
alternative hypothesis: two.sided
```

De donde, se sigue que no se puede rechazar la hipótesis nula pues el $p-$valor no es menor a la significancia dada inicialmente, luego los residuales provienen de una muestra aleatoria que por tanto confirma su independencia.

Ahora, para los residuales resultantes del modelo con los coeficientes hallados por medio de remuestreo se espera que suceda lo mismo como ha venido ocurriendo a lo largo de la validación por su cercanía a los residuales originales.

Esto se puede garantizar utilizando el código `x-factor(sign(Restudent))` seguido de `runs.test(x)` en donde se obtiene la siguiente salida de $\mathcal{R}$:

```
                    Runs Test

data:  x
Standard Normal = 0.25704, p-value = 0.7971
```

```
alternative hypothesis: two.sided
```

Es decir, en los residuales del modelo con bootstraping se sigue manteniendo la independencia y por tanto se tiene que nuestros residuos son no correlacionados.

Así se concluye que el bootstraping no altera los supuestos que ya cumplía el modelo inicial, solo cambia de una manera mínima los coeficientes y nos brinda una manera de realizar inferencia de una manera más segura. Luego nuestro modelo final tiene la misma estructura que el modelo inicial pero tomamos los parámetros resultantes del remuestreo.

# 5. Observaciones de alta palanca, atípicas e influyentes.

- **Observaciones de Alto apalancamiento:**

Recuerde que una observación $i$ es de alta palanca si se tiene que $h_{ii} > \frac{2p}{n}$ o $h_{ii} > \frac{3p}{n}$, ahora, teniendo que en nuestro modelo $p = 12$ y $n = 2884$ entonces, toda aquella observación que tenga un $h_{ii}$ mayor a $8.32017^{-3}$ se considera de alto apalancamiento, esto se puede visualizar de una mejor manera en los siguientes gráficos:

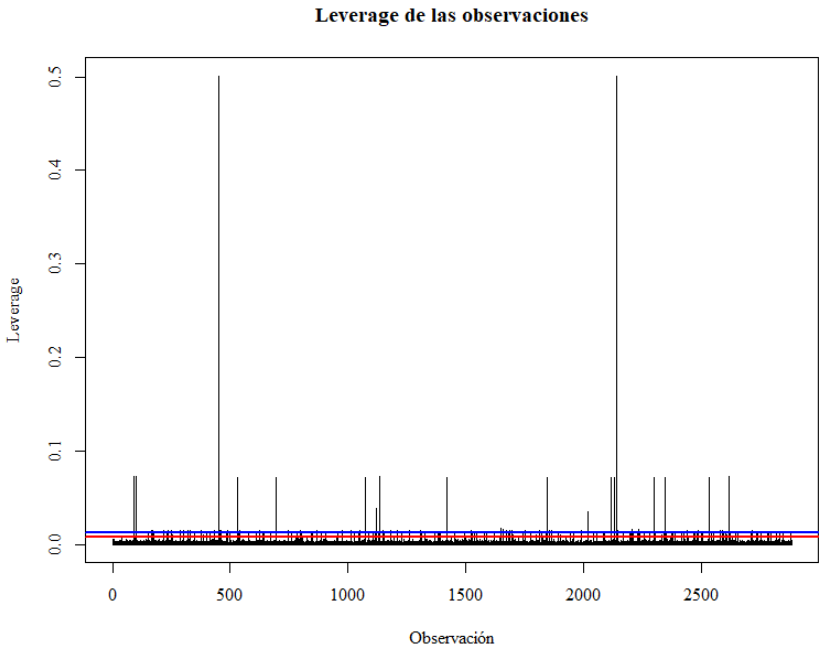



Figura 10.

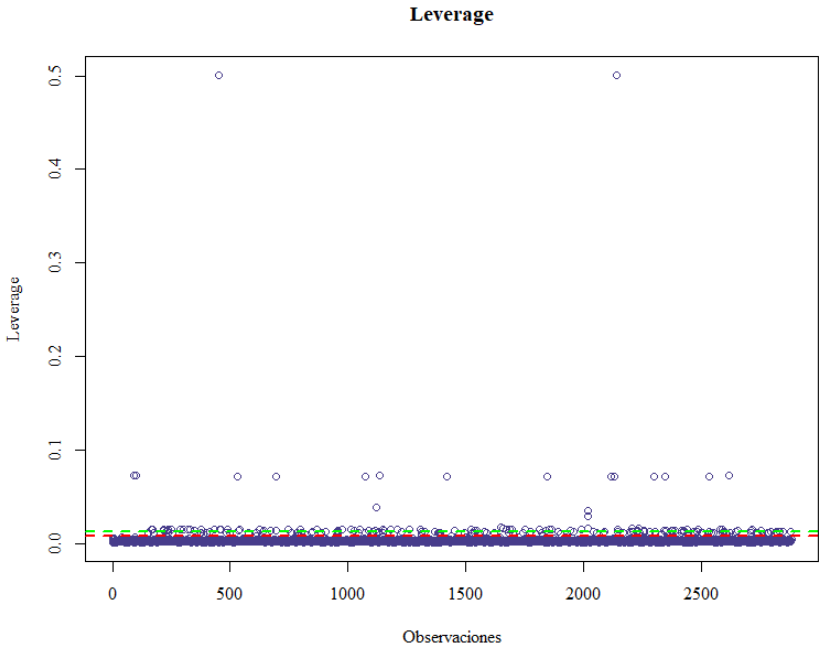



Figura 11.

Así, como se había mencionado anteriormente si se considera la primera cota se tienen bastantes

observaciones de alta palanca, en realidad si se toma de esta manera se tienen $388$ observaciones de alto apalancamiento lo cual teniendo en cuenta de $n = 2884$ representa aproximadamente al $13.5\,\%$ del total de observaciones y si se toma la segunda cota se tienen $115$.

Algunos de los datos con mayor apalancamiento se pueden ver en al siguiente tabla:

| Observación | 2142 | 451 | 2621 | 90 | 1135 | 101 |
|---|---|---|---|---|---|---|
| Leverage | 0.50042 | 0.50042 | 0.07280 | 0.07239 | 0.07237 | 0.07225 |
| Observación: | 1845 | 2348 | 533 | 2119 | 2534 | 694 |
| Leverage | 0.07220 | 0.0721 | 0.07211 | 0.07194 | 0.07183 | 0.07176 |
| Observación: | 2300 | 1076 | 1419 | 2130 | 1123 | 2019 |
| Leverage | 0.07167 | 0.07159 | 0.07159 | 0.07156 | 0.03903 | 0.03537 |

Esta situación se puede deber a las características de estas observaciones, en estudios sociales y específicamente en este estudio se pueden encontrar individuos con ciertas características especiales que no comparten muchos mas individuos y que por tanto a la hora de estimarsen tienen un peso bastante grande.

- **Observaciones atípicas:**

  Para realizar el estudio de observaciones atípicas se recuerda que una observación es atípica si el residual estudentizado relacionado a ella es mayor en valor absoluto que $3$, ahora en nuestro modelo inicial se tienen ciertos residuales estudentizados y en nuestro modelo final (con los coeficientes obtenidos del bootstraping) se tienen otros que ya se sabe que son cercanos a los iniciales. ¿Son necesariamente las observaciones atípicas del modelo inicial las mismas observaciones atípicas del modelo tomando los coeficientes del bootstraping? La cercanía de los valores y el comportamiento de los residuales en las gráficas nos hace presentir que si, sin embargo, se procede a verificar dicha pregunta.

  Primero se visualizan los valores atípicos del primer modelo, para ello, se realiza un gráfico de las observaciones vs el valor absoluto del residual de la observación, el cual se puede ver de la siguiente manera:

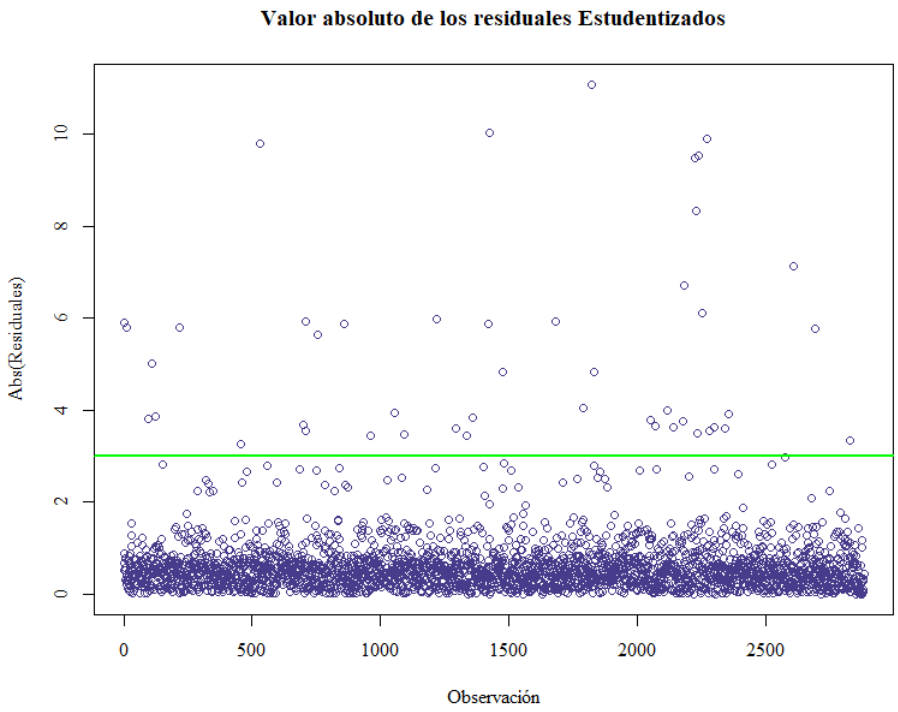



Figura 12.

Allí podemos ver que hay bastantes observaciones atípicas, en realidad hay un total de 46 observaciones, un número mucho menor al de observaciones de alta palanca pero que aun sigue siendo elevado. De hecho, se puede ver la misma atipicidad de algunos residuales con respecto a los demás en el siguiente gráfico *boxplot*.

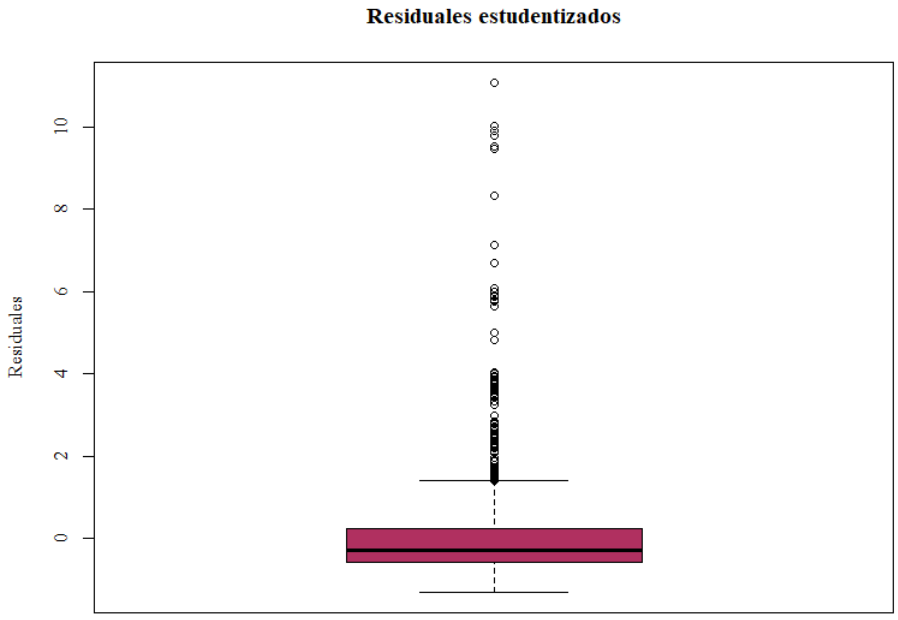



Figura 13.

Ahora, algunas de estas observaciones atípicas se pueden visualizar en la siguiente tabla:

| Observación | 1823 | 1424 | 2274 | 532 | 2239 | 2228 |
|---|---|---|---|---|---|---|
| Residual Estudentizado | 11.067639 | 10.026829 | 9.905654 | 9.801304 | 9.532462 | 9.470018 |

Lo primero que se puede visualizar comparando las tablas de observaciones atípicas y observaciones de alta palanca es que estos no coinciden necesariamente, lo cual es muy intuitivo pues se tienen muchas más observaciones del primer tipo, ahora la presencia de estos datos atípicos podría deberse a que a pesar de que el individuo relacionado a tales observaciones tenga características que tienen muchos más individuos de la muestra la cantidad habitual de cigarrillos que consume no es cercana a las de los individuos que integran el grupo a donde este individuo inicial pertenece.

Por último, para el modelo con coeficientes resultantes del bootstraping se sigue lo siguiente:

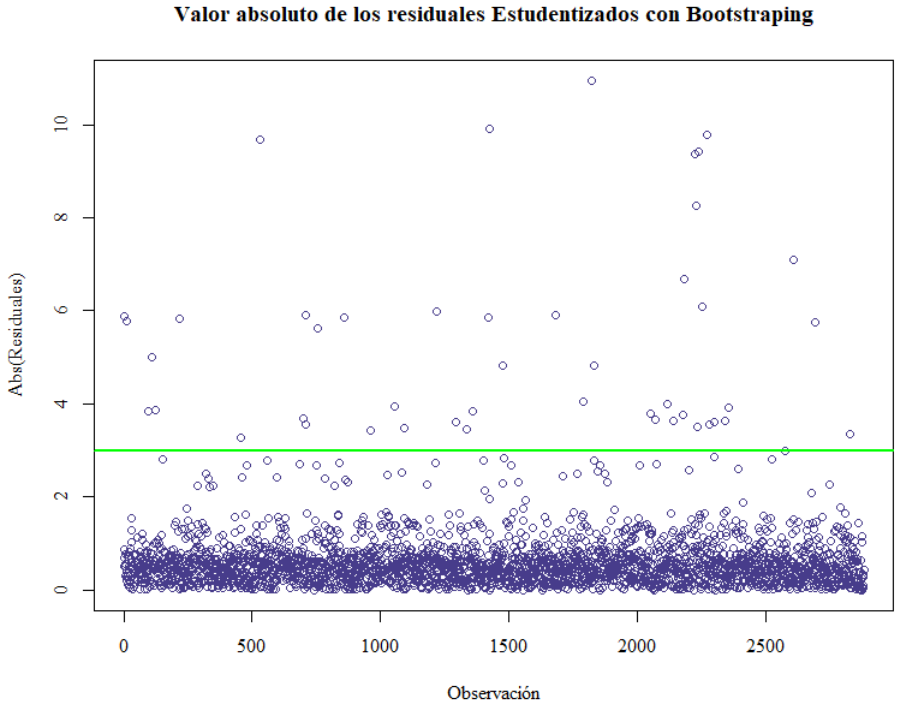



Figura 14.

Obteniendo así nuevamente 46 observaciones atípicas. Ahora el *boxplot* de los residuales es el siguiente.

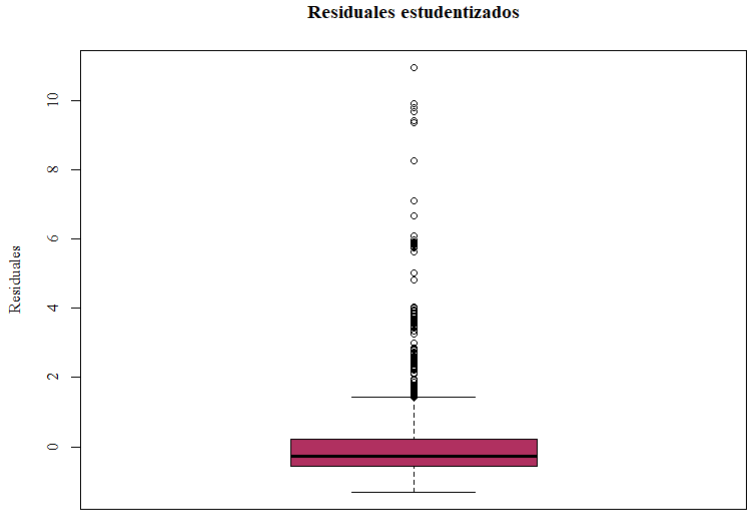



Figura 15.

En ambas gráficas se ven pequeños cambios, sin embargo, siguen la misma tendencia, por último los valores atípicos en este modelo son:

| Observación | 1823 | 1424 | 2274 | 532 | 2239 | 2228 |
|---|---|---|---|---|---|---|
| Residual Estudentizado | 10.939428 | 9.904402 | 9.775674 | 9.674316 | 9.419636 | 9.361561 |

De donde se concluye que la respuesta es que si, las observaciones atípicas del modelo con Bootstraping son las mismas que sin este método de remuestreo, sin embargo, revisando los valores podemos ver que en el modelo del bootstraping los residuales estandarizados de estas observaciones tienden a disminuir, lo cual nos hace creer que en un modelo con valores atípicos muy cercanos a 3 posiblemente con una cantidad suficiente de repeticiones en el remuestreo se va a tener que dicho valor no es atípico en el modelo con este modelo.

- **Observaciones influyentes:**

  Las observaciones influyentes se miden según la distancia de Cook, como criterio se escogió que una observación es influyente si su distancia de Cook es mayor a $\frac{4}{(n-p-2)}$, lo cual para nuestro modelo es $\frac{4}{(2870)} = 1.39372^{-3}$, luego los valores influyentes se pueden visualizar en la siguiente gráfica:

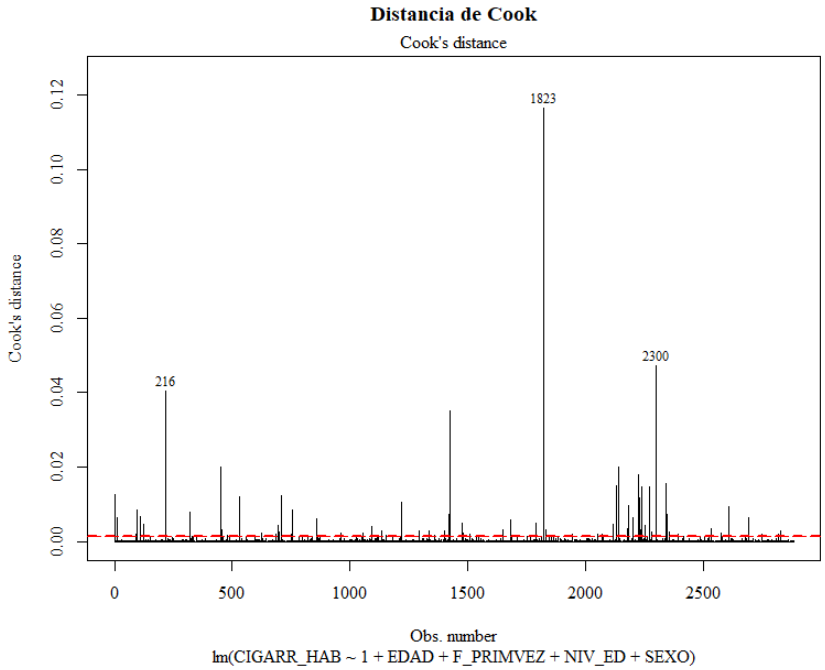



Figura 16.

En total, se tienen 86 observaciones cuya influencia se puede ver mejor en el siguiente gráfico:

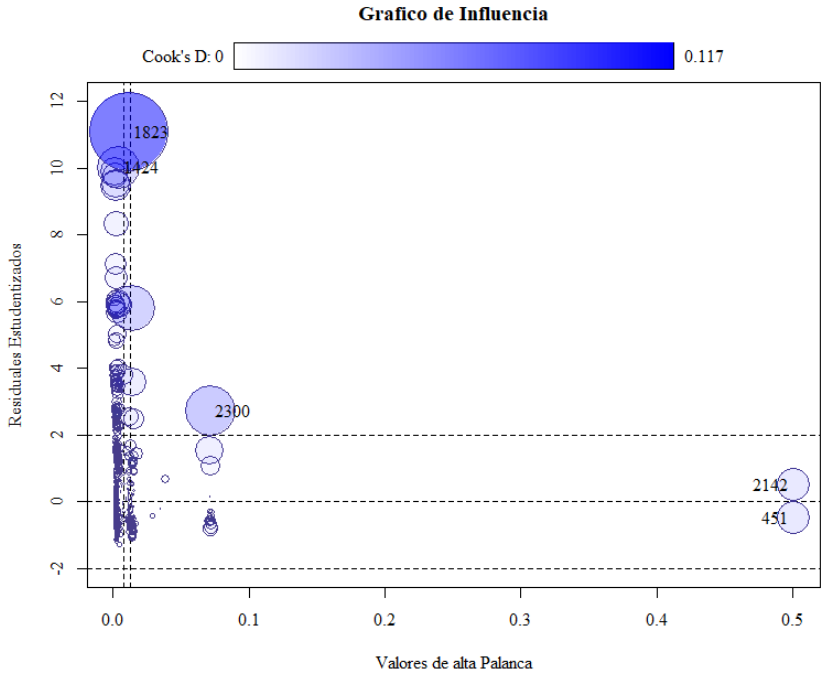



Figura 17.

En la siguiente tabla se pueden visualizar algunos de los valores influyentes con las distancias de Cook más grandes, es así que se puede ver que no necesariamente el valor atípico con un residual estudentizado más grande o el valor de alta palanca con el leverage más grande van a ser el valor con la distancia de Cook más grande.

| Observación | 4 | 12 | 90 | 95 | 110 | 122 |
|---|---|---|---|---|---|---|
| Distancia de Cook | 0.012551 | 0.006429 | 0.002010 | 0.00847 | 0.006812 | 0.004546 |

Ahora, para evaluar la verdadera influencia de estas observaciones en nuestro modelo se va a estimar un nuevo modelo sin dichos valores, de aquí obtenemos la siguiente salida de $\mathcal{R}$

```
             Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.282236   0.903069   4.742 2.22e-06 ***
EDAD         0.096349   0.008458  11.391  < 2e-16 ***
F_PRIMVEZ   -0.137072   0.023519  -5.828 6.25e-09 ***
NIV_ED2     -2.649183   2.644971  -1.002 0.316627
NIV_ED3      1.532357   0.780084   1.964 0.049589 *
NIV_ED4      2.120348   0.788215   2.690 0.007186 **
NIV_ED5      1.307659   0.779629   1.677 0.093599 .
NIV_ED6      0.869344   0.813512   1.069 0.285329
NIV_ED7      1.187368   0.808805   1.468 0.142203
NIV_ED8     -0.820537   0.971273  -0.845 0.398292
SEXO2       -0.819940   0.226781  -3.616 0.000305 ***
---
Residual standard error: 5.678 on 2787 degrees of freedom
Multiple R-squared:  0.06866,  Adjusted R-squared:  0.06532
F-statistic: 20.55 on 10 and 2787 DF,  p-value: < 2.2e-16
```

Observe que por ejemplo en el modelo original se tenia que $\hat{\beta}_4 = -0{,}474564$ y en este modelo sin los datos influyentes se tiene que $\hat{\beta}_4 = 1{,}532357$, adicionalmente este no es el único coeficiente con el que ocurre un cambio tan significativo como lo es el del signo, por tanto se concluye que estos datos son realmente influyentes en el modelo.

Una solución para dicho problema es estimar los coeficientes minimizando funciones que no sean tan intolerantes a los valores atípicos o identificar estas observaciones, sus características y buscar individuos que las cumplan para poder recolectar su información y así disminuir la influencia que estas observaciones tienen por si solas, es decir, es necesario ingresar nuevas observaciones que se encuentren"cerca" a las influyentes.

- **Regresión LAD (Bonus)**

  Para mitigar el peso que tiene en la estimación de los coeficientes de regresión las observaciones influyentes se realiza un modelo con regresión LAD, es decir, se minimiza la diferencia entre el valor observado y el valor estimado pero en valor absoluto, con ayuda de $\mathcal{R}$ se obtiene que los nuevos coeficientes del modelo de regresión están dados por:

| | Estimate | Std.Error | Z value | p-value |
|---|---|---|---|---|
| (Intercept) | 5.4146 | 1.4928 | 3.6272 | 0.0003 |
| *EDAD* | 0.1045 | 0.0147 | 7.1284 | 0.0000 |
| *F_PRIMVEZ* | -0.1498 | 0.0401 | -3.7319 | 0.0002 |
| *NIV_ED2* | -5.5122 | 2.9166 | -1.8899 | 0.0588 |
| *NIV_ED3* | -1.0801 | 1.2512 | -0.8633 | 0.3880 |
| *NIV_ED4* | -1.0488 | 1.2644 | -0.8295 | 0.4068 |
| *NIV_ED5* | -1.9582 | 1.2491 | -1.5677 | 0.1170 |
| *NIV_ED6* | -1.7213 | 1.3127 | -1.3112 | 0.1898 |
| *NIV_ED7* | -1.1254 | 1.3024 | -0.8641 | 0.3875 |
| *NIV_ED8* | -2.2962 | 1.5899 | -1.4442 | 0.1487 |
| *NIV_ED9* | 3.9059 | 7.1637 | 0.5452 | 0.1718 |

| Degrees of freedom: 2884 total; 2872 residual |
|---|
| Scale estimate: 7.056975 |
| Log-likelihood: -9518.902 on 13 degrees of freedom |

Luego, comparando este modelo con el modelo final y el modelo final sin las observaciones influyentes, se sigue que este tiende a parecerse más al modelo final pues la mayoría de los parámetros tienen el mismo signo que los parámetros de este modelo, sin embargo, se puede intuir que la influencia de estas observaciones que alteraban en un principio al modelo final se ve disminuida, pues por ejemplo en este modelo de regresión LAD se tiene que el coeficiente $\hat{\beta}_5$ tiene un signo negativo, esto es un cambio significativo con respecto a la estimación del valor que tenia en el modelo final y que en cambio se asemeja al signo que tiene este parámetro en el modelo final sin las observaciones influyentes, es así que este modelo logra mitigar de una mejor manera el efecto de las observaciones influyentes y arroja en la mayoría de los casos estimaciones que encuentran un equilibrio entre las iniciales y las obtenidas sin los valores influyentes, es decir, contempla estas observaciones sin necesidad de quitarlas de los datos, pero no se deja influenciar por ellas.

## 6. Uso del modelo.

- ¿El modelo de regresión parece aportar a la explicación de la variable dependiente? ¿Qué pruebas o estadísticas les ayudan a dar respuesta a eso?

  El modelo de regresión si parece aportar significativamente a la explicación de la variable respuesta escogida, como prueba de ello se puede hacer un test de significancia de la regresión con un $\alpha = 0.05$ en donde se evaluará el siguiente sistema de hipótesis:

$$\begin{cases} H_0 : \beta_i = 0 \text{ para todo i.} \\ \textbf{vs} \\ H_1 : \beta_i \neq 0 \text{ para algún i.} \end{cases}$$

  El $p - valor$ de dicho test se puede revisar en la salida que arroja $\mathcal{R}$ para el modelo, es decir, el $p-$valor es de $2{,}2e-16$ lo cual es claramente menor a $0.05$ y que por tanto nos confirma que la hipótesis nula de este test de significancia se rechaza.

  Ahora, también se puede realizar una prueba anova y comparar un modelo únicamente con el intercepto respecto a nuestro modelo, esto se realiza con el siguiente código de $\mathcal{R}$.

```
modelo3<-lm(CIGARR_HAB~1,data=db4)
anova(modelo3,modelo1,test="F")
```

  Estó nos arroja el mismo $p-$valor dado con anterioridad y por tanto se constata que en efecto la regresión tiene sentido y que nuestras variables explicativas están explicando a la variable *CIGARR_HAB*.

- ¿El o los coeficientes asociados a las variables cualitativas es (son) significativo(s)? ¿Qué pruebas o estadísticas les ayudan a dar respuesta a eso?

  En nuestro modelo se contemplaron dos variables cualitativas, la primera *NIV_ED* y la segunda *SEXO*, ahora con respecto a la primera tenemos $8$ coeficientes asociados a esta, los cuales son $\beta_3, \beta_4, \cdots, \beta_{10}$, como se nos esta pidiendo mostrar si dichos coeficientes son o no significativos, se nos esta pidiendo realizar $8$ pruebas de hipótesis con el $5\,\%$ de significancia que juzguen el sistema de hipótesis de la forma:

$$\begin{cases} H_0 : \beta_i = 0 \\ \textbf{vs} \\ H_1 : \beta_i \neq 0 \end{cases}$$

  Los $p-$valores de cada una de estas pruebas de hipótesis se pueden encontrar en la salida de $\mathcal{R}$ del modelo, por tanto se obtiene lo siguiente para el nivel educativo:

  • El p-valor para el coeficiente $\beta_3$ esta dado por $0.1653$ lo cual es mayor a la significancia escogida y por tanto el coeficiente no es significativo.
  • Ahora, para el coeficiente $\beta_4$ el $p-$valor esta dado por $0.6126$ lo cual también es mayor a la significancia escogida y por tanto el coeficiente no es significativo.
  • De la misma manera, el p-valor para el coeficiente $\beta_5$ esta dado por $0.7993$ lo cual es mayor a

la significancia escogida y se concluye que el coeficiente no es significativo.
• El p-valor para el coeficiente $\beta_6$ esta dado por $0.2672$ lo cual también es mayor a la significancia escogida y por tanto el coeficiente no es significativo.
• Para el coeficiente $\beta_7$ se tiene que su $p-$valor esta dado por $0.1114$ lo cual es mayor que la significancia escogida y por tanto el coeficiente no es significativo.
• De igual manera que en los anteriores coeficientes, el p-valor para el coeficiente $\beta_8$ esta dado por $0.4165$ este es mayor que la significancia escogida y por consiguiente el coeficiente no es significativo.
• El p-valor para el coeficiente $\beta_9$ esta dado por $0.1299$ lo cual es mayor que la significancia escogida y por tanto el coeficiente no es significativo.
• Finalmente, el p-valor para el coeficiente $\beta_{10}$ esta dado por $0.8726$ lo cual también es mayor a la significancia escogida y por tanto el coeficiente no es significativo.

Se concluye que ninguno de los coeficientes asociados a la variable cualitativa *NIV_ED* es significativo y que por tanto como tal esta variable no esta aportando en gran medida a la explicación de la variable respuesta.

Por otro lado, para la segunda variable cualitativa, se tiene que el único coeficiente asociado a esta es $\beta_{11}$ el cual continuando con el nivel de significancia del $5\,\%$ si es significativo pues su $p-$valor es de $0.0362$ lo cual es menor al $5\,\%$, por tanto la variable *SEXO* si esta aportando algo a la explicación del consumo de tabaco.

- ¿Valdría la pena colocar alguna interacción de segundo orden en el modelo? ¿Cuál(es) y por qué?

Para mirar si valdría la pena colocar alguna interacción de segundo orden en el modelo nos ayudamos del test RESET el cual se puede escribir de la siguiente manera:

$$\begin{cases} Y_k = \mu_k + e_k \\ \mu_k = \beta_0 + \beta_1 x_{k1} + \beta_2 x_{k2} + \cdots + \beta_{10} x_{k10} + \beta_{11} x_{k11} + \delta_1 \hat{y}^2 \\ e_k \sim \mathcal{N}(0, \sigma^2) \end{cases}$$

Entonces se formula la siguiente prueba de hipótesis con $5\,\%$ de significancia que juzga el siguiente sistema de hipótesis.

$$\begin{cases} H_0 : \delta_1 = 0 \\ \quad \textbf{vs} \\ H_1 : \delta_1 \neq 0 \end{cases}$$

Esta se puede realizar en $\mathcal{R}$ utilizando el código `resettest` y el tipo fitted o regressor, el primero tipo nos sirve para ver si hay alguna interacción que se omitió y que podría ser significativa a la hora de explicar el consumo del tabaco, el segundo tipo nos ayuda a entender si en realidad alguna de las variables que ingresamos en el modelo debería o seria de más utilidad con una potencia al cuadrado, así aplicando este código a nuestro modelo final se sigue que:

Para verificar si son necesarias las interacciones con el código dicho de antemano se obtiene la salida mostrada a continuación.

```
RESET test

data:  modelo1
RESET = 0.40975, df1 = 1, df2 = 2871, p-value = 0.5221
```

Entonces como el $p-$valor no es menor al nivel se significancia fijado no se puede rechazar la hipótesis nula y por tanto no hay interacciones que al menos con las variables escogidas como explicativas para el modelo valga la pena colocar.

Por otro lado, para las variables al cuadrado se obtiene que:

```
RESET test

data:  modelo1
RESET = 1.0611, df1 = 2, df2 = 2870, p-value = 0.3462
```

Nuevamente se obtiene que el $p-$valor es mayor a la significancia escogida por tanto no vale la pena colocar alguna de las variables explicativas al cuadrado en el modelo.

Es así que se concluye que en general no vale la pena colocar alguna interacción de segundo grado entre las variables que se escogieron para explicar el consumo de tabaco.

- **Conclusiones:**

  En un inicio se planteo como objetivo estudiar la relación entre la adicción al alcohol y/o el tabaco con sus consumos prematuros y el nivel educacional de los individuos. ¿Es el consumo temprano de estas sustancias psicoactivas un factor relevante en la dependencia al alcohol y el tabaco?¿Qué tanto puede influir el nivel educativo de un individuo en la adicción a estas dos sustancias?.

  En el presente trabajo podemos contestar estas preguntas únicamente para el consumo de tabaco pues es la variable de estudio que se escogió, entonces por lo dicho anteriormente, no se tiene evidencia de que el nivel educativo este influyendo o explicando el consumo de tabaco diario y por tanto su adicción, sin embargo, se tiene evidencia de que el consumo temprano de esta sustancia psicoactiva si influencia y explica la permanencia del consumo de un individuo y su consecuente adicción, lo cual nos hace concluir que la prevención del consumo de esta sustancia en edades tempranas es una buena vía de solución para el problema de salud publica que representa la adicción al tabaco, lo anteriormente dicho esta respaldado por el $p-$valor relacionado al coeficiente que acompaña a esta variable pues al ser tan pequeño $(6{,}04e-05)$ con un nivel de significancia del $5\,\%$ se obtiene que en efecto es significativo.

  Ahora, para contestar estas preguntas con respecto al alcohol es necesario realizar este mismo procedimiento realizado para el tabaco con la variable *TRAG_HAB* pues no podemos inferir que el comportamiento de los individuos que consumen estas dos sustancias es igual.

  Como segundo objetivo se planteo, deducir si hay o no incidencia del consumo de alcohol en el consumo del tabaco y viceversa, lo cual no es posible contestar con lo realizado en este trabajo pues se estudio como única variable al alcohol para poder explicarla de una manera separada, por tanto no podemos responder a ciencia cierta lo que se propuso, sin embargo, gracias a los coeficientes de correlación hallados en la fase de identificación se creería que si hay una

relación detrás del consumo de estas dos sustancias psicoactivas, para confirmarlo tendríamos que realizar un modelo de regresión lineal simple entre estas dos variables o realizar un modelo que incluya variables relacionadas con el consumo del alcohol.

Para el tercer objetivo se propuso analizar los factores que pueden incidir en el alcoholismo y tabaquismo de cierto individuo desde la perspectiva de género. ¿Realmente hay evidencia de que el sexo masculino tiene más predisposición a estas adicciones en la actualidad?. Esta pregunta se puede responder únicamente para el tabaco, en donde se observó que en efecto el sexo juega un papel importante a la hora de explicar el consumo del tabaco, de hecho la variable dummy que se le asocio escogió como categoría base a los hombres, es decir, que para las mujeres se tiene una disminución promedio de $0{,}7158$ cigarrillos por día, lo cual respalda la creencia de que los hombres suelen tener mas tendencias a la adicción a esta sustancia psicoactiva.

Por último el cuarto objetivo era observar la repercusión en el estado de salud de una persona ante el consumo de alguna de estas dos sustancias psicoactivas, esta pregunta no se puede contestar con el modelo planteado, primero porque no se añadió al estado de salud como variable explicativa del consumo de tabaco y segundo porque para observar la verdadera repercusión en el estado de salud de un individuo que consuma estas sustancias se debería realizar un modelo que en realidad escoja a dicha variable como la variable respuesta pues se estaría buscando explicar el daño o beneficio en cuestión de salud causado por estas sustancias, sin embargo, el estado de salud es una variable categórica y por tanto aun no se tienen las herramientas para realizar un modelo de esta manera.

## Referencias


[1] T. Hernández López, J. Roldán Fernández, A. Jiménez Frutos, C. Mora Rodríguez, D. Escarpa Sánchez-Garnica, and M. T. Pérez Álvarez, "La Edad de Inicio en el Consumo de Drogas, un Indicador de Consumo Problemático," *Psychosocial Intervention*, vol. 18, pp. 199 – 212, 12 2009.

[2] M. Morán and H. Shapiro, "Termodinámica," 2004.

[3] J. McGervey and M. Physics, *Introduction to Modern Physics*. Academic Press, 1971.