Presentacion

Table of contents

1 Introduccion

1.1 Formulacion del Problema de Investigacion

¿Cuáles son los componentes nutricionales que influyen significativamente en el contenido energético (kcal) de las preparaciones alimenticias y cómo se relacionan entre sí en los platos peruanos en el 2017?

2 Marco Teorico

2.1 Antecedentes

En 1950, siendo el Dr. Carlos Collazos Chiriboga jefe del Instituto Nacional de Nutrición, se establece convenio entre el Ministerio de Salud de Perú y el Servicio Cooperativo Interamericano de Salud Pública de los EE.UU., con la finalidad de fortalecer a los Institutos de Salud, dentro de los que se encontraba el Instituto Nacional de Nutrición, orientando sus actividades a la investigación; es así que durante su gestión, se realizan encuestas para evaluar la situación nutricional del país y analizar los alimentos que la población ingería, mediante análisis de sus componentes, que culminan con la publicación de la Tabla Peruana de Composición de Alimentos, en la revista Anales de la Facultad de Medicina de la Universidad Nacional Mayor de San Marcos en 1952, dicha tabla fue elaborada con muestras de alimentos obtenidos en las regiones del país donde se ejecutaron las encuestas dietéticas.

La elaboración de estas Tablas representa un trabajo analítico de la Institución, sostenido por más de sesenta años. Desde su primera publicación se han incorporado nuevos alimentos y las cifras anteriormente publicadas han sido sometidas a cuidadosa comprobación, puesto que entre las principales funciones del Centro Nacional de Alimentación y Nutrición (CENAN) se encuentra la investigación y el ofrecer herramientas para el estudio del estado nutricional de los habitantes del país.

Como producto del trabajo realizado, en octubre del 2009, el Instituto Nacional de Salud (INS) a través del CENAN, publicó la octava edición de las Tablas Peruanas de Composición de Alimentos, que actualiza parte de la información de los alimentos en relación a las versiones anteriores e incorpora datos de nutrientes importantes como zinc y fibra dietaria, así como de nuevos alimentos. Asimismo, el proceso de actualización de las Tablas Peruanas de Composición de Alimentos, está enfocado a evaluar el contenido de macro y micronutrientes de los alimentos de las diferentes zonas del país, dado que los datos existentes datan de ediciones anteriores y de otras tablas internacionales.

La presente actualización considera información sobre energía y 19 nutrientes de 928 alimentos consumidos en el Perú, distribuidos en 14 grupos. Pero, también se está incorporando el grupo de preparados, con información de 942 preparaciones; en este caso, los datos de composición

varían respecto a los demás, pues solo se informa lo que fue analizado para la encuesta de alimentos consumidos fuera del hogar, a cargo del INEI. Todos los datos referidos proceden de análisis químicos, propios o imputados, o son estimados por cálculo de acuerdo con las normas de compilación, para garantizar su confiabilidad.

3 Objetivos

3.1 Objetivo general

Analizar el componente energético de los platos peruanos en base a sus múltiples composiciones.

3.2 Objetivo Especifico

1. Indentificar cuales factores o componentes alimenticios influyen mas en el valor energetico-kcal (variable dependiente) proporcionada por el alimento.
2. Identificar si la cantidad de agua presente en los alimentos esta relacionada negativamente con el valor energetico, es decir, si alimentos con mayor contenido de agua tienden a tene menos calorias.
3. Evaluar como la categoria de alimentos influye en el valor energetico.

4 Metodologia

4.1 Formulación del Modelo de Regresión Lineal Múltiple

Se plantea un modelo de regresion lineal multiple para analizar la relacion entre el contenido energetico (ENERC) de las preparaciones peruanos y demás contenidos alimenticios. Este modelo permite explorar como diferentes variables independientes afectan de manera conjunta la variable dependiente, proporcionando una compresion entre los factores evaluados.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{20} X_{20} + \varepsilon$$

donde:

- Y: Variable dependiente cuantitativa continua, representa por el contenido energetico (ENERC) de las preparaciones peruanas.
- $X_1, X_2, \ldots, X_{19}$: Variables independientes cuantitativa continuas, representada por agua, proteinas, grasas totales, calcio, y demás componentes.

- X_{20}: Variable indendiente cualitatita representada por la categoria de preparaciones (Categoria)
- $\beta_0, \beta_1, \ldots, \beta_{20}$: Coeficientes parametrales
- ε: Error aleatorio

El planteamiento de este modelo se fundamenta en la necesidad de analizar las diferencias en el aporte energetico segun las caracteristicas especificas de las prepaciones peruanas. Ademas este planteamiento permitira no solo identificar los factores que influyen en el contenido energetico, sino tambien estimar su magnitud.

Ademas, el modelo asume los siguientes supuestos basicos:

1. **Linealidad**: existe una relación lineal entre las variables independientes y la variable dependiente.
2. **Independencias de los errores**: Los residuos no están correlacionados entre sí.
3. **Homocedasticidad}: La varianza de los residuos es constante en todos los niveles de las independientes.
4. **Normalidad de los errores**: Los residuos siguen una distribución normal.
5. Tamaño de muestra sea mucho mayor al numero de variables independientes.
6. **No multicolinealidad**: Las variables explicativas no están altamente correlacionadas entre sí.

4.2 Diccionario de Variables

Se utiliza una base de datos que recoge información sobre los componentes nutricionales de diversas preparaciones alimenticias peruanas. La base de datos contiene una muestra de 942 observaciones y 21 variables continuas.

```
# A tibble: 21 x 4
   Componente                  Identificador Unidad Comentario
   <chr>                       <chr>         <chr>  <chr>
 1 Agua                        WATER         g      H2O
 2 Energía                     ENERC         kcal   Calorias
 3 Proteínas                   PROCNT        g      <NA>
 4 Grasa total                 FAT           g      Lipidos
 5 Cenizas                     ASH           g      Minerales totales
 6 Carbohidratos totales       CHOCDF        g      Azucares totales
 7 Carbohidratos disponibles   CHOAVL        g      Azucares disponibles
 8 Fibra dietaria              FIBTG         g      Fibra alimentaria
 9 Calcio                      CA            mg     Mineral esencial
10 Fósforo                     P             mg     Mineral esencial
11 Hierro                      FE            mg     Fe
12 Zinc                        ZN            mg     Zn
```

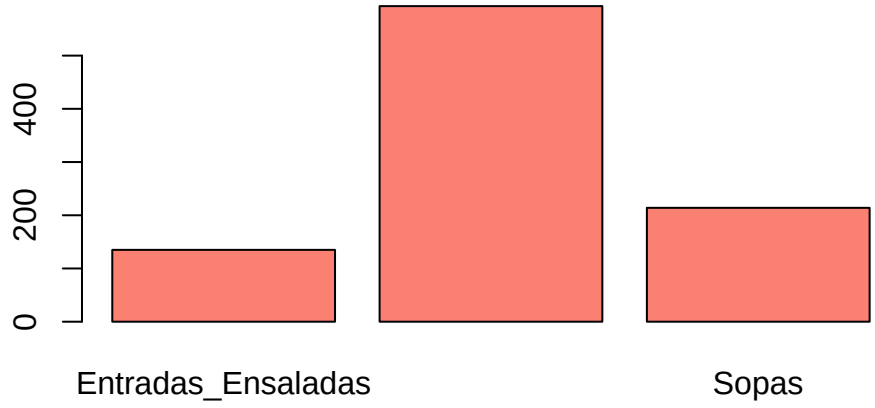
```
13 Potasio                          K            mg     K
14 Sodio                            Na           mg     Na
15 B caroteno equivalentes totales  CARTBQ        g     Precursos de Vitamina A
16 Vitamina A equivalentes totales  VITA          g     Retinol
17 Tiamina                          THIA         mg     Vitamina B1
18 Riboflavina                      RIBF         mg     Vitamina B2
19 Niacina                          NIA          mg     Vitamina B3
20 Vitamina C                       VITC         mg     Acido ascorbico
21 Ácido fólico                     Ácido fólico  g     Vitamina B9
```

4.3 Exploracion Inicial de Datos

4.3.1 Analisis de la variable independiente cualitativa

Se plantea agregar una variable categorica con niveles “Entradas y ensaladas”, “Sopas” y “Platos principales”, estas categorizan a las prepaciones peruanas.

```
Entradas_Ensaladas Platos_principales              Sopas
               135                593                214
```

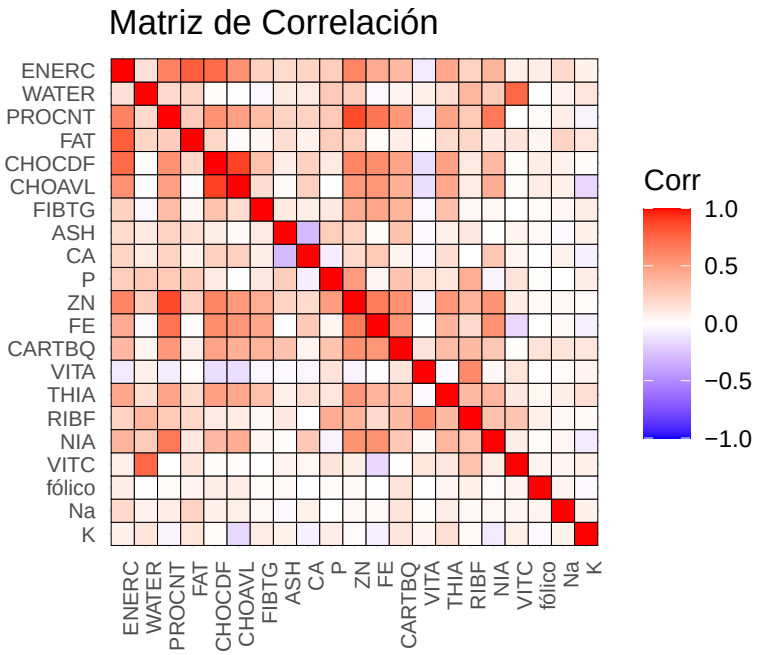


El gráfico muestra la frecuencia de preparaciones alimenticias por categoría. Se observa que los platos principales son la categoría más representada, seguidos por las sopas, mientras que las entradas y ensaladas tienen la menor frecuencia.

4.3.2 Analisis de la variables cuantitativa

4.3.2.1 Correlacion para las variables

La matriz de correlación muestra una relación positiva significativa entre variables como proteínas (PROCNT), grasas (FAT), y carbohidratos (CHOCDF) con el contenido energético (ENERC), lo que es coherente con su aporte calórico en preparaciones alimenticias. Además, se observan correlaciones fuertes entre algunas variables independientes, lo que sugiere posibles problemas de multicolinealidad que podrían influir en el modelo de regresión. Por otro lado, algunas variables como ciertos micronutrientes parecen tener poca o ninguna correlación con ENERC, indicando que su influencia podría ser limitada en este contexto.



4.4 Ajuste del modelo de regresión lineal múltiple inicial

El modelo ajustado inicial se incluyeron todas las variables cuantitativas como posibles predictores del contenido energético (ENERC) de las preparaciones alimenticias. El modolo incial permite obtener una visión inicial del impacto de cada predictor en la variable dependiente. Sin

embargo, dado el alto número de variables incluidas, es posible que algunas no sean significativas o que exista multicolinealidad, lo que podría afectar la estabilidad y la interpretabilidad del modelo. Por lo tanto, este ajuste inicial sirve como punto de partida para realizar análisis adicionales, como la selección de variables y la evaluación de supuestos, que refinen el modelo hacia uno más parsimonioso y representativo.

```
Call:
lm(formula = ENERC ~ ., data = data[1:21])

Residuals:
    Min      1Q  Median      3Q     Max
-38.248  -0.802   0.313   1.399  10.617

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4637410  0.3369815   1.376  0.16911
WATER       -0.0775532  0.0092506  -8.384  < 2e-16 ***
PROCNT       4.1368575  0.0875437  47.255  < 2e-16 ***
FAT          8.8056253  0.0250096 352.089  < 2e-16 ***
CHOCDF       4.0891945  0.0336739 121.435  < 2e-16 ***
CHOAVL      -0.1047057  0.0338953  -3.089  0.00207 **
FIBTG       -1.2472651  0.1072537 -11.629  < 2e-16 ***
ASH         -0.0992499  0.1286104  -0.772  0.44048
CA           0.0970507  0.0401938   2.415  0.01595 *
P           -0.0062928  0.0049305  -1.276  0.20217
ZN           0.0215841  0.0089192   2.420  0.01572 *
FE           1.7050067  0.4825123   3.534  0.00043 ***
CARTBQ      -1.1483989  0.2182882  -5.261 1.78e-07 ***
VITA        -0.0026061  0.0013072  -1.994  0.04649 *
THIA         7.5185664  3.6430466   2.064  0.03932 *
RIBF        -8.1129298  3.0670032  -2.645  0.00830 **
NIA          0.8432463  0.1909015   4.417 1.12e-05 ***
VITC         0.0869155  0.0433233   2.006  0.04513 *
fólico      -0.0266049  0.0418671  -0.635  0.52529
Na          -0.0001464  0.0003875  -0.378  0.70562
K           -0.0165456  0.0035195  -4.701 2.98e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.95 on 921 degrees of freedom
Multiple R-squared:  0.998, Adjusted R-squared:  0.998
F-statistic: 2.346e+04 on 20 and 921 DF,  p-value: < 2.2e-16
```

```
[1] "El R2 predictivo es igual a: 0.997218015950255"
```

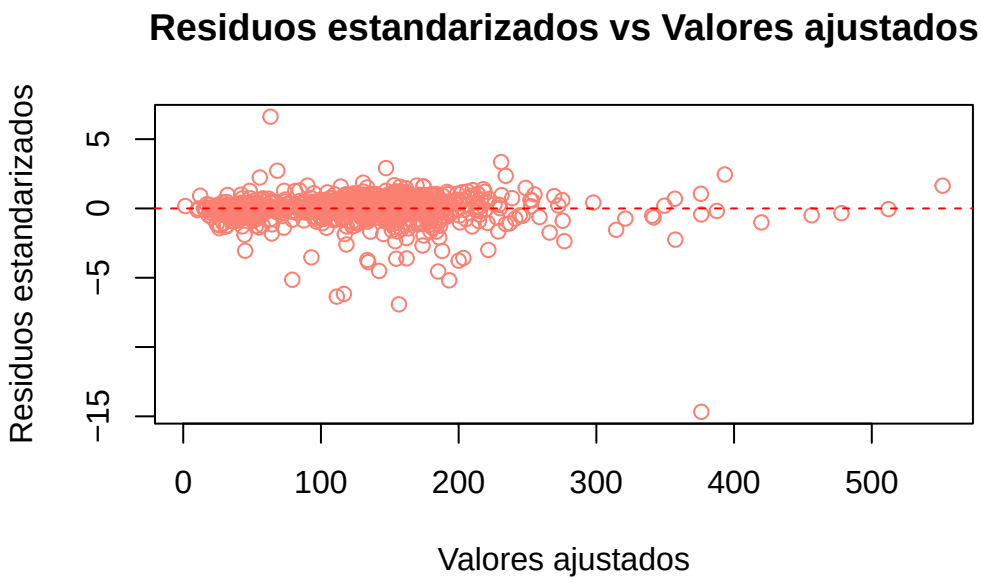
Con base en los resultados, donde tanto el R2 ajustado (0.998) como el R2 predictivo (0.9972) son extremadamente altos, se puede concluir que el modelo inicial tiene un excelente ajuste y una gran capacidad predictiva. Esto indica que las variables incluidas explican casi toda la variabilidad del contenido energético (ENERC) y que el modelo generaliza bien en datos no observados.

4.4.1 Analisis de Residuos para modelo inicial

4.4.1.1 Supuesto de homocedasticidad

La gráfica de residuos estandarizados frente a los valores ajustados permite evaluar el cumplimiento del supuesto de homocedasticidad en el modelo de regresión lineal múltiple. En este caso, los residuos parecen estar distribuidos de manera aleatoria alrededor de la línea horizontal (cero) sin mostrar un patrón claro o sistemático, lo cual sugiere que el supuesto de varianza constante de los errores se cumple razonablemente bien.

Sin embargo, se observa una ligera dispersión más amplia en los valores ajustados más grandes, lo que podría indicar una posible heterocedasticidad en esos puntos extremos. Si bien esto no parece ser un problema crítico, podría ser útil realizar pruebas estadísticas, como la prueba de White, para confirmar la homocedasticidad.



Realizando la Prueba white:

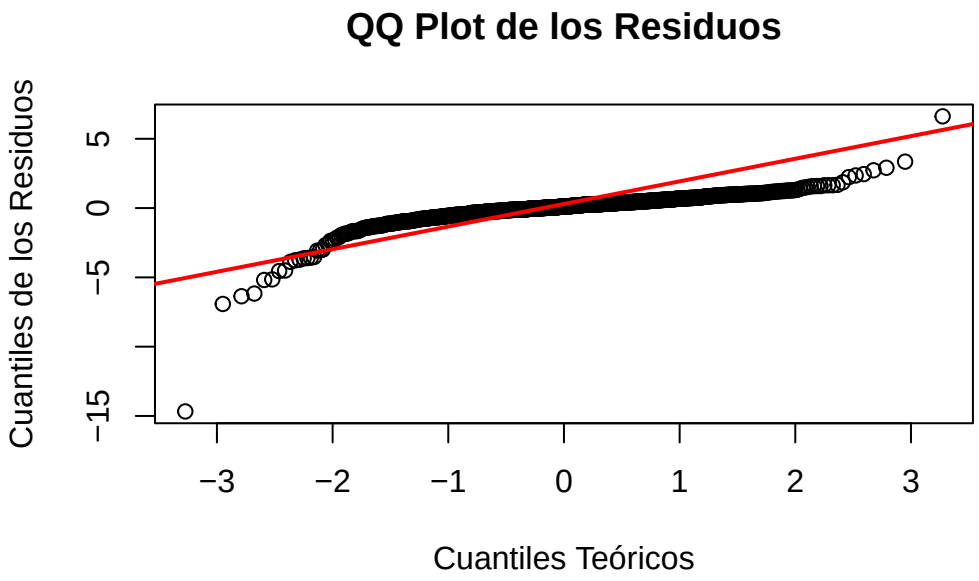
Dado que el p-valor es mayor a 0.05 (p>0.05), no se rechaza la hipótesis nula de homocedasticidad. Esto sugiere que no hay evidencia estadísticamente significativa para afirmar que el modelo presenta heterocedasticidad.

```
Estadístico de White: 26.54663
```

```
p-valor: 0.1156648
```

4.4.1.2 Supuesto de Normalidad

El gráfico Q-Q de los residuos sugiere que el supuesto de normalidad se cumple razonablemente bien, ya que la mayoría de los puntos siguen la línea diagonal. No obstante, se observan ligeras desviaciones en las colas, lo que podría indicar la presencia de valores atípicos que podrían ser analizados con mayor detalle para evaluar su impacto en el modelo



Realizando la prueba de normalidad: Dado que el p-valor es menor a 0.05 (p<0.05), se rechaza la hipótesis nula de que los residuos siguen una distribución normal.

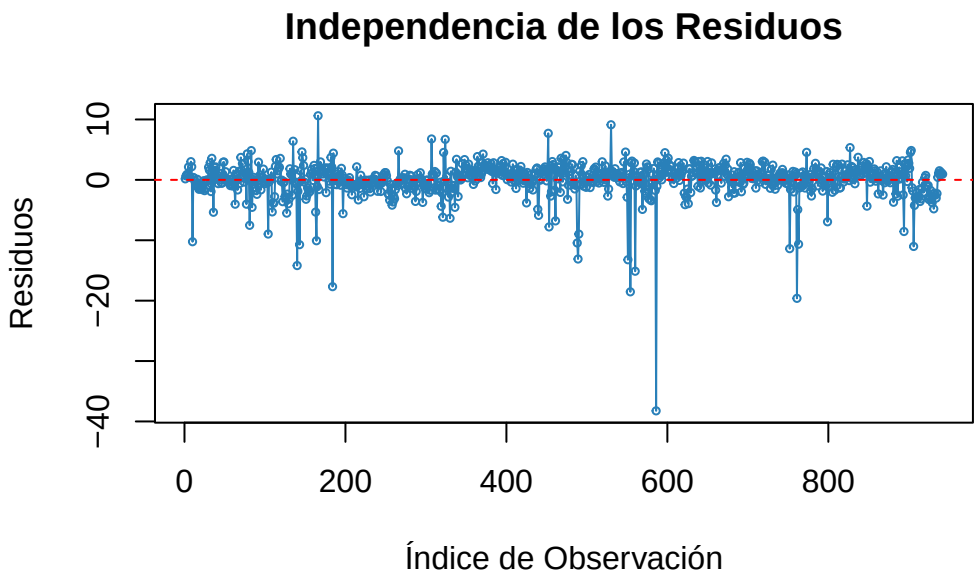
```
    Shapiro-Wilk normality test

data:  residuos
W = 0.73638, p-value < 2.2e-16
```

4.4.1.3 Supuesto de Independencia

La gráfica de residuos frente al índice de observación muestra que los residuos están distribuidos aleatoriamente alrededor de cero, sin patrones sistemáticos evidentes. Esto sugiere que el supuesto de independencia de los errores se cumple en el modelo, ya que no se observa autocorrelación o dependencia estructural relacionada con el orden de las observaciones.

Para confirmar, se aplicara una prueba formal de autocorrelación, como la prueba de Durbin-Watson.



Realizando la prueba Durbin-Watson

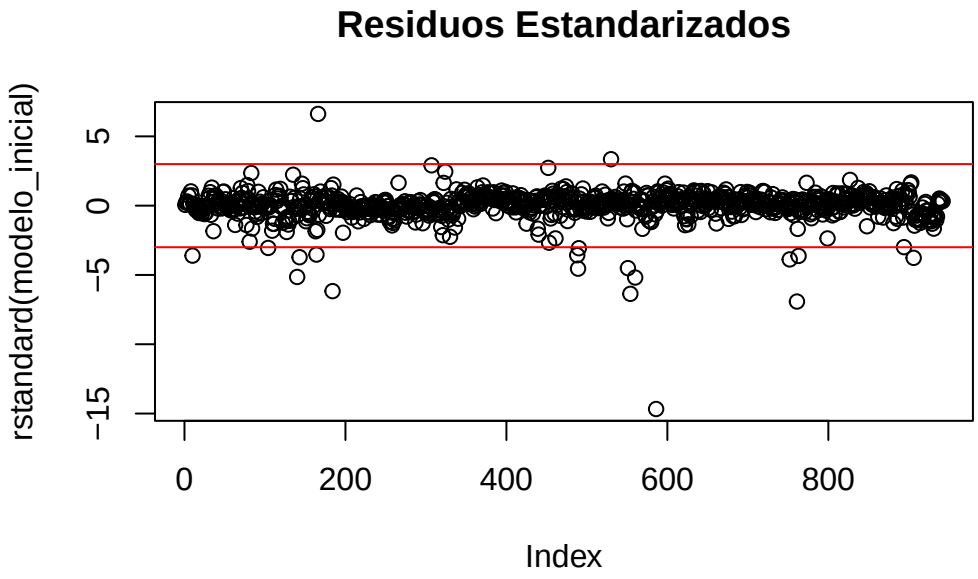
Dado que el p-valor es muy pequeño ($p<0.05$), rechazamos la hipótesis nula de que no existe autocorrelación en los residuos. Esto sugiere que hay autocorrelación positiva en los residuos, es decir, los residuos consecutivos están correlacionados.

```
    Durbin-Watson test

data:  modelo_inicial
DW = 1.6631, p-value = 3.947e-08
alternative hypothesis: true autocorrelation is greater than 0
```

4.4.1.4 Identificacion de Observaciones Atípicas

En la gráfica de residuos estandarizados se observa que la mayoría de los residuos se encuentran dentro del rango de -3 a 3, lo que indica un comportamiento adecuado del modelo para la mayoría de las observaciones. Sin embargo, se identifican algunas observaciones fuera de este rango, que podrían considerarse atípicas, ya que se alejan significativamente de los valores esperados según el modelo ajustado.



Analizamos los residuos estandarizados, cuyos valores absolutos son mayores a 3 y obtenemos las siguientes i-esimas observaciones

```
 10 104 140 143 164 166 184 488 489 490 530 551 554 560 586 752 761 763 906
 10 104 140 143 164 166 184 488 489 490 530 551 554 560 586 752 761 763 906
```

Con estos i-esimas observaciones, se procede a excluirlas de la data original

Y volvemos a plantear un modelo con todas la variables independientes cuantitativas pero con una data excluida de datos atipicos

```
Call:
lm(formula = ENERC ~ ., data = data[1:21])

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-13.2443  -0.8176   0.1823   1.0157   6.7971

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.2043112  0.2113738   0.967  0.33401
WATER       -0.0610165  0.0059137 -10.318  < 2e-16 ***
PROCNT       4.5381247  0.0583167  77.819  < 2e-16 ***
FAT          8.8157572  0.0157054 561.319  < 2e-16 ***
CHOCDF       4.0927500  0.0215627 189.807  < 2e-16 ***
CHOAVL      -0.0532022  0.0225067  -2.364  0.01830 *
FIBTG       -0.8737519  0.0698430 -12.510  < 2e-16 ***
ASH          0.5470103  0.0896679   6.100 1.57e-09 ***
CA           0.1186713  0.0253594   4.680 3.31e-06 ***
P           -0.0092804  0.0031063  -2.988  0.00289 **
ZN           0.0019411  0.0058210   0.333  0.73886
FE           1.7587536  0.3061081   5.746 1.25e-08 ***
CARTBQ      -2.3424971  0.1419207 -16.506  < 2e-16 ***
VITA        -0.0011298  0.0008548  -1.322  0.18661
THIA         5.0705357  2.4424834   2.076  0.03818 *
RIBF        -3.8658454  1.9733078  -1.959  0.05041 .
NIA          0.2036017  0.1238737   1.644  0.10060
VITC        -0.0136847  0.0274845  -0.498  0.61867
fólico      -0.0113448  0.0259763  -0.437  0.66241
Na          -0.0001288  0.0002483  -0.519  0.60420
K           -0.0349142  0.0041998  -8.313 3.40e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.826 on 903 degrees of freedom
Multiple R-squared:  0.9992,    Adjusted R-squared:  0.9992
F-statistic: 5.999e+04 on 20 and 903 DF,  p-value: < 2.2e-16

[1] "El modelo con  20  variables independientes tiene un R2 ajustado: 0.999231270436784"
```

Se visualiza que ante la exclusion de datos atipicos, el R2 ajustado (0.999231) mejora a comparacion del anterior (R2ajustado = 0.997218)

```
Estadístico de White: 19.02172

p-valor: 0.4554434
```

```
    Shapiro-Wilk normality test

data:  residuos
W = 0.92571, p-value < 2.2e-16
```

```
    Durbin-Watson test

data:  modelo_sin_outliers
DW = 1.7309, p-value = 8.34e-06
alternative hypothesis: true autocorrelation is greater than 0
```

4.4.1.5 Evaluar multicolinealidad

El análisis de multicolinealidad, realizado a través del Factor de Inflación de la Varianza (VIF), identificó que algunas variables presentan valores elevados de VIF, indicando una posible correlación alta entre ellas. En particular, las variables PROCNT (Proteínas), CHOCDF (Carbohidratos), CHOAVL (Carbohidratos disponibles) y ZN (Zinc) tienen un VIF mayor a 5, lo que sugiere problemas significativos de multicolinealidad según el criterio adoptado.

```
Model :
ENERC ~ WATER + PROCNT + FAT + CHOCDF + CHOAVL + FIBTG + ASH +
    CA + P + ZN + FE + CARTBQ + VITA + THIA + RIBF + NIA + VITC +
    fólico + Na + K
```

Problema significativo VIF >5 o 10

```
   Variables Tolerance      VIF
1      WATER 0.3079051 3.247754
2     PROCNT 0.1483058 6.742827
3        FAT 0.6330520 1.579649
4     CHOCDF 0.1111956 8.993164
5     CHOAVL 0.1103071 9.065600
6      FIBTG 0.5111773 1.956268
7        ASH 0.7041518 1.420148
8         CA 0.7431498 1.345624
9          P 0.3477349 2.875754
10        ZN 0.1009800 9.902948
11        FE 0.2944803 3.395813
12    CARTBQ 0.4279118 2.336930
```

```
13      VITA 0.4423480 2.260664
14      THIA 0.4463545 2.240372
15      RIBF 0.2928809 3.414357
16       NIA 0.3102842 3.222852
17      VITC 0.3510244 2.848805
18    fólico 0.9326965 1.072160
19        Na 0.8699543 1.149486
20         K 0.5547061 1.802756
```

La multicolinealidad implica que estas variables están altamente correlacionadas con otras en el modelo, por ello se procedio a identificar aquellas variables con alto VIF: PROCNT, CHOCDF, CHOAVL, ZN

Despues de identifcar las variables con alto VIF, se analiza las correlaciones con las demás variables (Criterio de Correlacion >= 0.60)

- **Correlacion de PROCNT y las demas variables**

```
PROCNT     ZN     FE    NIA  ENERC CHOCDF CARTBQ CHOAVL   THIA  FIBTG      P
  1.00   0.87   0.69   0.69   0.63   0.56   0.56   0.48   0.47   0.35   0.29
  RIBF    FAT     CA  WATER    ASH     Na   VITA      K fólico   VITC
  0.29   0.27   0.25   0.22   0.21   0.10   0.07   0.04   0.02   0.01
```

Alto correlacion de PROCNT con variables independientes: ZN, FE, NIA

Correlacion con la variable dependiente: Alta

- **Correlacion de CHOCDF y las demas variables**

```
CHOCDF CHOAVL  ENERC     ZN     FE PROCNT   THIA CARTBQ    NIA  FIBTG     CA
  1.00   0.89   0.74   0.63   0.59   0.56   0.49   0.48   0.39   0.33   0.25
   FAT   VITA      P   RIBF fólico    ASH     Na      K  WATER   VITC
  0.21   0.14   0.11   0.11   0.10   0.08   0.07   0.04   0.03   0.01
```

Alto correlacion de CHOCDF con variables independientes: CHOAVL, ZN

Correlacion con la variable dependiente: Alta

- **Correlacion de CHOAVL y las demas variables**

```
CHOAVL CHOCDF  ENERC     FE     ZN PROCNT   THIA CARTBQ    NIA     CA      K
  1.00   0.89   0.55   0.55   0.53   0.48   0.46   0.44   0.43   0.25   0.19
 FIBTG   VITA   RIBF fólico     Na   VITC  WATER    FAT    ASH      P
  0.18   0.14   0.10   0.10   0.07   0.04   0.02   0.02   0.01   0.01
```

Alto correlacion de CHOAVL con variables independientes: CHOCDF

Correlacion con la variable dependiente: Baja

- **Correlacion de ZN y las demas variables**

```
   ZN PROCNT     FE  ENERC CHOCDF CARTBQ    NIA CHOAVL      P   THIA  FIBTG
 1.00   0.87   0.65   0.63   0.63   0.58   0.55   0.53   0.52   0.52   0.43
 RIBF  WATER    FAT    ASH     CA   VITC      K   VITA fólico     Na
 0.40   0.27   0.26   0.25   0.20   0.10   0.06   0.05   0.04   0.04
```

Alto correlacion de ZN con variables independientes: PROCNT, FE, CHOCDF

Correlacion con la variable dependiente: Alta

Como resultado, se tomó la decisión de eliminar aquellas variables que presentan redundancia y menor aporte al modelo con base en los criterios de correlación y VIF. Las acciones específicas fueron las siguientes:

- Eliminación de CHOAVL: Esta variable mostró una correlación baja con la variable dependiente (ENERC) y presentó alta colinealidad con CHOCDF (Carbohidratos). Por lo tanto, se consideró redundante y se eliminó del modelo
- Eliminación de ZN (Zinc): La variable ZN presentó una alta correlación con PROCNT (Proteínas). Debido a que PROCNT tiene un mayor impacto en el modelo (mayor correlación con ENERC), se eliminó ZN.
- Eliminación de NIA (Niacina): La variable NIA mostró una alta correlación con PROCNT, lo que generaba redundancia. Dado que PROCNT es más relevante para explicar la variabilidad de ENERC, se decidió eliminar NIA.

——Variables a eliminar: CHOAVL, ZN, NIA

Una vez eliminadas aquellas variables, procedo a plantear un modelo con las demas variables independientes, y vuelvo analizar el VIF que no supere mayor a 5

```
   Variables Tolerance      VIF
1      WATER 0.3180615 3.144046
2     PROCNT 0.3390076 2.949787
3        FAT 0.7044833 1.419480
4     CHOCDF 0.4758502 2.101501
5      FIBTG 0.6527149 1.532062
6        ASH 0.7079065 1.412616
7         CA 0.7492731 1.334627
8          P 0.6508000 1.536570
9         FE 0.3145320 3.179326
```

```
10    CARTBQ 0.4432458 2.256084
11      VITA 0.4735556 2.111685
12      THIA 0.5004610 1.998158
13      RIBF 0.3127723 3.197214
14      VITC 0.3569872 2.801221
15    fólico 0.9343183 1.070299
16        Na 0.8869766 1.127425
17         K 0.7496063 1.334034
```

Visualmente se observa que cumplen con VIF menor a 5; procedo a mostrar el modelo con las variables independientes restantes.

```
Call:
lm(formula = ENERC ~ ., data = data[1:18])

Residuals:
     Min       1Q   Median       3Q      Max
-13.3214  -0.7367   0.1867   1.0181   7.8916

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1530321  0.2104762   0.727    0.467
WATER       -0.0589343  0.0058346 -10.101  < 2e-16 ***
PROCNT       4.5922521  0.0386784 118.729  < 2e-16 ***
FAT          8.8263092  0.0149292 591.212  < 2e-16 ***
CHOCDF       4.0495815  0.0104523 387.433  < 2e-16 ***
FIBTG       -0.8485911  0.0619796 -13.691  < 2e-16 ***
ASH          0.5626593  0.0896776   6.274 5.44e-10 ***
CA           0.1222290  0.0253255   4.826 1.63e-06 ***
P           -0.0094649  0.0022769  -4.157 3.53e-05 ***
FE           1.8627977  0.2970109   6.272 5.52e-10 ***
CARTBQ      -2.4127291  0.1398307 -17.255  < 2e-16 ***
VITA        -0.0011860  0.0008285  -1.432    0.153
THIA         3.6826326  2.3130672   1.592    0.112
RIBF        -2.9658029  1.9148205  -1.549    0.122
VITC        -0.0215768  0.0273295  -0.790    0.430
fólico      -0.0115739  0.0260257  -0.445    0.657
Na          -0.0002160  0.0002466  -0.876    0.381
K           -0.0312416  0.0036228  -8.623  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.831 on 906 degrees of freedom
Multiple R-squared:  0.9992,    Adjusted R-squared:  0.9992
F-statistic: 7.019e+04 on 17 and 906 DF,  p-value: < 2.2e-16
```

```
[1] "El modelo con  17  variables independientes tiene un R2 ajustado: 0.999227004033299"
```

4.4.1.6 Modelo resultante

Hubo una reduccion de variables independientes despues de analizar la multicolinealidad entre variables, el modelo resultante es uno con las siguientes variables independientes cuantitativas:

```
 [1] "WATER"  "PROCNT" "FAT"    "CHOCDF" "FIBTG"  "ASH"    "CA"     "P"
 [9] "FE"     "CARTBQ" "VITA"   "THIA"   "RIBF"   "VITC"   "fólico" "Na"
[17] "K"
```

4.5 Metodos de Seleccion de Variables

4.5.1 Metodo del mejor subconjunto

El método del mejor subconjunto se utilizó para seleccionar las variables más relevantes en la explicación del contenido energético (ENERC) de los alimentos.

```
El mejor modelo tiene un R^2 ajustado de: 0.9992282
```

```
Variables seleccionadas:
```

```
 [1] "(Intercept)" "WATER"       "PROCNT"      "FAT"         "CHOCDF"
 [6] "FIBTG"       "ASH"         "CA"          "P"           "FE"
[11] "CARTBQ"      "VITA"        "THIA"        "RIBF"        "K"
```

```
Call:
lm(formula = mejor_formula, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-13.3857  -0.7351   0.1851   1.0102   7.9066
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1348463  0.2057820   0.655    0.512
WATER       -0.0624500  0.0038971 -16.025  < 2e-16 ***
PROCNT       4.5965660  0.0382568 120.150  < 2e-16 ***
FAT          8.8225614  0.0143835 613.382  < 2e-16 ***
CHOCDF       4.0482544  0.0102416 395.274  < 2e-16 ***
FIBTG       -0.8590773  0.0605023 -14.199  < 2e-16 ***
ASH          0.5676418  0.0893234   6.355 3.29e-10 ***
CA           0.1219208  0.0253010   4.819 1.69e-06 ***
P           -0.0089869  0.0022397  -4.012 6.50e-05 ***
FE           1.9376373  0.2886171   6.714 3.34e-11 ***
CARTBQ      -2.4436253  0.1363887 -17.917  < 2e-16 ***
VITA        -0.0012717  0.0008186  -1.554    0.121
THIA         3.4885812  2.3023193   1.515    0.130
RIBF        -2.9351656  1.8908866  -1.552    0.121
K           -0.0306783  0.0035880  -8.550  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.829 on 909 degrees of freedom
Multiple R-squared:  0.9992,    Adjusted R-squared:  0.9992
F-statistic: 8.535e+04 on 14 and 909 DF,  p-value: < 2.2e-16


[1] "El modelo escogio por el mejor Subconjunto tiene  14  variables independientes"


[1] "y un R2 ajustado: 0.999228162360057"
```

4.5.2 Metodo Backward

El método backward se utilizó para seleccionar las variables más relevantes en la explicación del contenido energético (ENERC) de los alimentos.

```
Call:
lm(formula = formula_final_back, data = data[1:18])

Residuals:
     Min       1Q   Median       3Q      Max
-13.3857  -0.7351   0.1851   1.0102   7.9066
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1348463  0.2057820   0.655    0.512
WATER       -0.0624500  0.0038971 -16.025  < 2e-16 ***
PROCNT       4.5965660  0.0382568 120.150  < 2e-16 ***
FAT          8.8225614  0.0143835 613.382  < 2e-16 ***
CHOCDF       4.0482544  0.0102416 395.274  < 2e-16 ***
FIBTG       -0.8590773  0.0605023 -14.199  < 2e-16 ***
ASH          0.5676418  0.0893234   6.355 3.29e-10 ***
CA           0.1219208  0.0253010   4.819 1.69e-06 ***
P           -0.0089869  0.0022397  -4.012 6.50e-05 ***
FE           1.9376373  0.2886171   6.714 3.34e-11 ***
CARTBQ      -2.4436253  0.1363887 -17.917  < 2e-16 ***
VITA        -0.0012717  0.0008186  -1.554    0.121
THIA         3.4885812  2.3023193   1.515    0.130
RIBF        -2.9351656  1.8908866  -1.552    0.121
K           -0.0306783  0.0035880  -8.550  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.829 on 909 degrees of freedom
Multiple R-squared:  0.9992,    Adjusted R-squared:  0.9992
F-statistic: 8.535e+04 on 14 and 909 DF,  p-value: < 2.2e-16


[1] "El modelo escogido por el metodo Backward tiene  14  variables independientes"


[1] "y un R2 ajustado: 0.999228162360057"
```

4.5.3 Metodo Forward

El método forward se utilizó para seleccionar las variables más relevantes en la explicación del contenido energético (ENERC) de los alimentos.

```
Call:
lm(formula = formula_final_for, data = data[1:18])

Residuals:
     Min       1Q   Median       3Q      Max
-13.3857  -0.7351   0.1851   1.0102   7.9066
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1348463  0.2057820   0.655    0.512
WATER       -0.0624500  0.0038971 -16.025  < 2e-16 ***
PROCNT       4.5965660  0.0382568 120.150  < 2e-16 ***
FAT          8.8225614  0.0143835 613.382  < 2e-16 ***
CHOCDF       4.0482544  0.0102416 395.274  < 2e-16 ***
FIBTG       -0.8590773  0.0605023 -14.199  < 2e-16 ***
ASH          0.5676418  0.0893234   6.355 3.29e-10 ***
CA           0.1219208  0.0253010   4.819 1.69e-06 ***
P           -0.0089869  0.0022397  -4.012 6.50e-05 ***
FE           1.9376373  0.2886171   6.714 3.34e-11 ***
CARTBQ      -2.4436253  0.1363887 -17.917  < 2e-16 ***
VITA        -0.0012717  0.0008186  -1.554    0.121
THIA         3.4885812  2.3023193   1.515    0.130
RIBF        -2.9351656  1.8908866  -1.552    0.121
K           -0.0306783  0.0035880  -8.550  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.829 on 909 degrees of freedom
Multiple R-squared:  0.9992,    Adjusted R-squared:  0.9992
F-statistic: 8.535e+04 on 14 and 909 DF,  p-value: < 2.2e-16


[1] "El modelo escogido por el metodo Forward tiene  14  variables independientes "


[1] "y un R2 ajustado: 0.999228162360057"
```

4.6 Evaluacion del modelo final

Tenemos como modelos resultantes:

1. Metodo del mejor subconjunto:

```
[1] "R2 ajustado: 0.999228162360057"
```

2. Metodo Backward

```
[1] "R2 ajustado: 0.999228162360057"
```

3. Metodo Forward

```
[1] "un R2 ajustado: 0.999228162360057"
```

En estos 3 metodos obtenemos un modelo con 14 variables independientes, todas ellas iguales para cada uno de los modelos; y incluso observando la significancia de las variables, se observa que la variable VITA, THIA , RIBF son menos significantes.

Se hizo la prueba quitando estas variables y se obtiene un modelo con

```
[1] "un R2 ajustado: 0.999218227551958"
```

Aunque estas variable individualmente parece poco significativa (su p-valor es alto), puede estar contribuyendo al modelo de forma conjunta con otras variables. Esto significa que su exclusión afecta la capacidad del modelo para explicar la variabilidad de la variable dependiente, por lo que se elige con estas variables independientes correspondientes al modelo final.

```
Model :
ENERC ~ WATER + PROCNT + FAT + CHOCDF + FIBTG + ASH + CA + P +
    FE + CARTBQ + VITA + THIA + RIBF + K
```

A un inicio hubieron 19 variables independientes cuantitativas y al final nos quedamos solo con 14 variables, asi optimizando la parsimonia y representatividad del modelo.

Visualizacion de los coeficientes parametrales del modelo final

```
 (Intercept)        WATER       PROCNT          FAT       CHOCDF        FIBTG
 0.134846287 -0.062450047  4.596565954  8.822561404  4.048254377 -0.859077315
         ASH           CA            P           FE       CARTBQ         VITA
 0.567641822  0.121920788 -0.008986903  1.937637348 -2.443625290 -0.001271725
        THIA         RIBF            K
 3.488581219 -2.935165593 -0.030678301
```

4.7 Conclusion

De los coeficientes parametrales podemos podemos responder los objetivos como:

- FAT (Grasas): Tiene un coeficiente alto (8.8226), lo que indica que un aumento en el contenido de grasa incrementa significativamente el valor energético.
- PROCNT (Proteínas) y CHOCDF (Carbohidratos) también tienen coeficientes positivos importantes (4.5967 y 4.0482, respectivamente), reflejando su aporte calorico o energetico.

- RIBF (Riboflavina): Tiene un coeficiente negativo significativo (-2.9317), aunque su impacto calórico directo es menor en comparación con macronutrientes como grasa o carbohidratos.
- El coeficiente de WATER es negativo (-0.0625), lo que confirma una relación inversa: los alimentos con mayor contenido de agua tienden a tener menos calorías. Esto es consistente con el hecho de que el agua no aporta calorías y diluye otros componentes energéticos.

Por lo tanto, grasas, proteínas y carbohidratos son los principales factores que aumentan el valor energético, mientras que agua y algunos micronutrientes (como riboflavina) lo disminuyen.