Facultad de Ciencias Exactas, UNLP.
Estadística 2020, Dpto. De Matemática.

Estudio de modelo de regresión lineal simple aplicado a valores nutritivos de cereal

Micucci, Fernanda Daniela

ESTADÍSTICA

2020

Los cereales que se consumen en el desayuno y dietas nutricionales son productos ultra procesados y por ello, no todos ofrecen la misma cantidad de energía y puede que algunos sean escasos en buenos nutrientes y concentrados en calorías. En general, los que determinan su valor calórico son principalmente el azúcar, fibras, los hidratos de carbono, y, en pocos casos, las grasas, aunque también influyen el contenido de otras sustancias que se añaden como frutas, chocolate, frutos secos, miel, etc. Es decir, con la intención de obtener nuevas variedades, encontramos actualmente una gran influencia de estos agregados en los niveles calóricos.

Los cereales se introdujeron en la dieta hace unos 10.000 años, durante el desarrollo de la agricultura. El ser humano pasó de una alimentación basada en la caza y la recolección a una dieta con un alto contenido en granos y cereales. A raíz de las guerras mundiales se hizo evidente la necesidad de aumentar la producción agrícola, para satisfacer la creciente demanda de alimentos de la población. Se dio importancia a los tipos de cereales que se cultivan en la actualidad, los cuales fueron desplazando a los cultivos de legumbres. Los más empleados en la alimentación humana son el trigo, el arroz y el maíz.

Al ser un alimento seco es concentrado y por lo tanto alto en calorías. Por ello, es muy interesante hacer un análisis sobre cómo influyen sus componente en los niveles nutricionales.

El conjunto de datos se obtuvo gracias a las observaciones realizadas sobre las etiquetas nutricionales de las cajas de cereales que estaban vigentes en el mercado en el año 1990, con el objetivo de saber si los cereales ricos en fibra también son ricos en azúcar y calorías.

Para este trabajo sólo fueron consideradas la variable **Fibra** como variable independiente (el cual se considera predictor aleatorio) y como variable respuesta o dependiente a las **Calorías.** Se eligió una muestra de 36 cereales y se observaron sus niveles de fibra en gramos por porción , y calorías por porción (recordar que una caloría equivale a 4.19 J).

Objetivos del trabajo:

- Hacer un análisis descriptivo de las variables involucradas.
- Estudiar si es posible aplicar el método de regresión lineal simple, mediante la verificación de hipótesis y el análisis del conjunto de datos.
- Si se puede encontrar un modelo lineal adecuado, realizar los tests de hipótesis necesarios para inferir sobre los parámetros involucrados en el modelo, así como también intervalos de confianza y predicción para cada uno.
- Concluir, en base a la información recolectada, si el modelo propuesto finalmente es el más adecuado para predecir los niveles de calorías según los gramos de fibra.

Se realizaron análisis descriptivos para cada una de las variables involucradas:

Para la variable "Fibra":

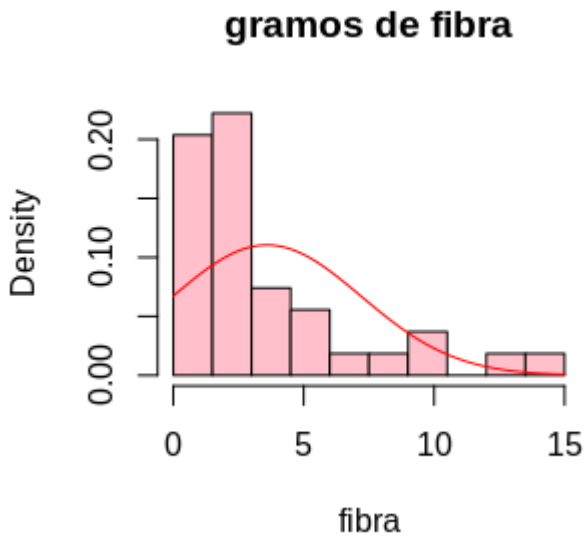
Con la intención de obtener información sobre la normalidad de la distribución, centralidad, dispersión, simetría y presencia de datos atípicos, se realizó un análisis descriptivo de las variables Fibra y Calorías que arrojaron los siguientes resultados:

Resumen de 5 números más la media y la desviación muestral (en gramos por porción):

Mínimo	Primer cuantil	Mediana	Tercer cuantil	Máximo	Media muestral	Desviación estándar muestral	Distancia intercuartil
0	1.0	3.0	4.25	14.0	3.59	3.61	3.25

Se calcularon las medias 0.1 y 0.2 -podadas:

***Media 0.1-podada=** 3.07
***Media 0.2-podada=** 2.78



El histograma junto con curva normal superpuesta

El gráfico muestra distribución de datos asimétrica (sesgado a derecha), lo cual se ve también reflejado en los valores de la media y mediana muestrales y las medias podadas, que no coinciden.

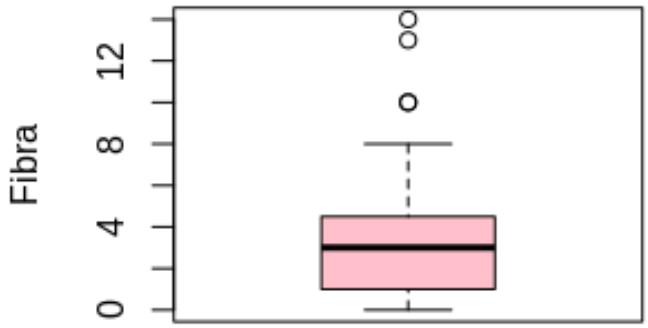


Diagrama de caja

En el diagrama de caja se observan datos atípicos y una caja estrecha que indica concentración del 50% de datos alrededor de la mediana muestral y asimetría en la distribución: los datos se encuentran mucho más concentrados entre el tercer cuantil y la mediana. El bigote inferior es mucho más pequeño que el superior, coincidiendo con la concentración de datos que muestra el histograma en el mismo intervalo.

Para la variable "Calorias":

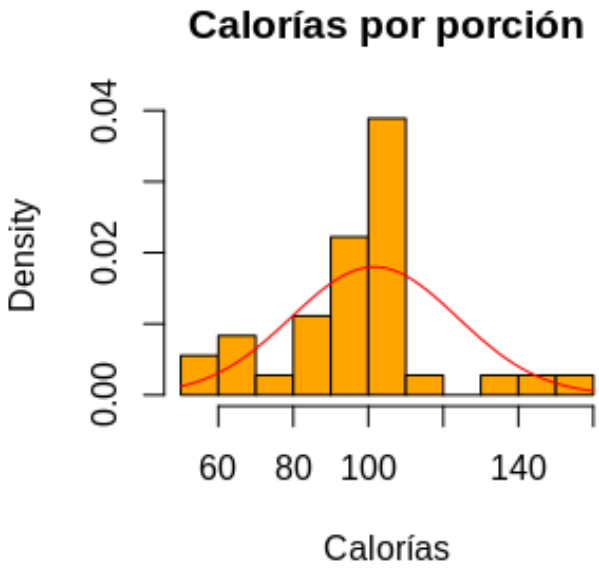
El resumen de 5 números más la media y la desviación muestral, arrojaron los siguientes valores (en calorías por porción):

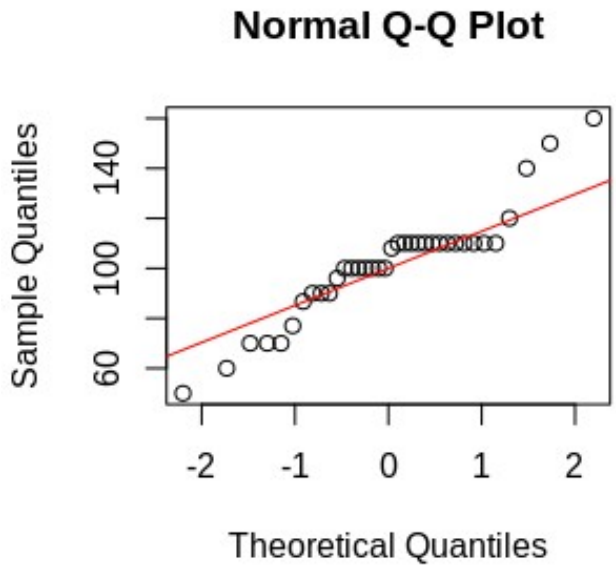
Mínimo	Primer cuantil	Mediana	Tercer cuantil	Máximo	Media muestral	Desviación estándar muestral	Distancia intercuartil
50.0	90.0	104.0	110.0	160.0	101.6	22.1639	20

Se calcularon las medias 0.1 y 0.2 -podadas:

***Media 0.1-podada=** 100.92
***Media 0.2-podada=** 103.36

El histograma junto con la curva normal superpuesta, y el ajuste normal de los datos correspondientes son los siguientes:





Las gráficas muestran distribución de los datos asimétrica, y el ajuste normal presenta colas a izquierda y a derecha. La media y mediana muestrales y las medias podadas no coinciden (ver valores en la tabla), y presentan una moda en el intervalo [100,110].

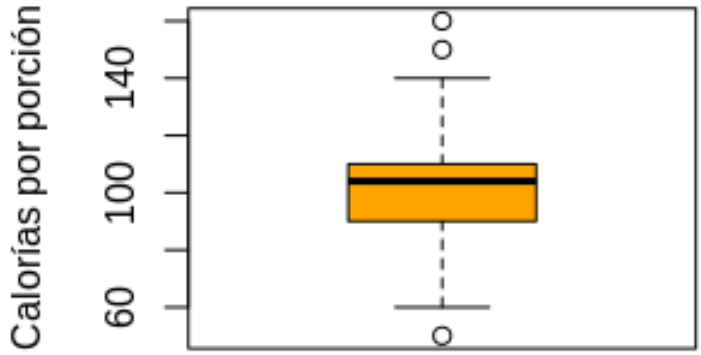


Diagrama de caja

Se observan datos atípicos y una caja pequeña que indica que el 50% de datos entre el primer y tercer cuantil tiene gran concentración, siendo más concentrados entre el tercer cuantil y la mediana muestral. Los bigotes superior e inferior tienen longitudes similares.

Modelo de regresión lineal simple:

El **modelo de regresión lineal simple** o **ajuste lineal simple**, es utilizado para aproximar la relación de dependencia lineal (si existe) entre una variable dependiente o respuesta *Y*, y la variable independiente X. La idea se basa en buscar un modelo que no será una caracterización exacta de la

relación entre las variables, pero que de la mejor aproximación lineal, de ser posible, que permita evaluar la validez de las inferencias realizadas.

En este trabajo, la que se considera variable independiente es un predictor aleatorio (variable aleatoria), por lo tanto se tomará el modelo probabilístico de la siguiente manera:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

donde dicha fórmula representa la relación lineal entre las variables X e Y, determinadas por coeficientes $\beta_0 y \beta_1$, y la variable aleatoria ϵ (independiente de X) que cumple $\epsilon \sim N(0, \sigma^2)$.

Luego, una forma equivalente para este modelo propuesto es $E(Y|X) = \beta_0 + \beta_1 X$.

Un método para estimar los parámetros $\beta_0 y \beta_1$ del modelo es el **método de cuadrados mínimos,** el cual consiste en ajustar una recta al conjunto de puntos estudiado, de modo que las diferencias entre ellos y los valores ajustados con la recta sean mínimas.
Si consideramos como estimador de Y a

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

donde $\hat{\beta}_0 y \hat{\beta}_1$ son estimadores de $\beta_0 y \beta_1$ respectivamente, de la forma:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Para esta muestra tomamos $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad \forall i \in \{1,..,36\}$

Para poder aceptar el modelo propuesto, deben analizarse ciertas hipótesis y tenerse en cuenta las consideraciones siguiente:

- Existencia de una relación lineal entre las variables Fibra y Calorías.
- El término de error ϵ debe tener distribución normal y, más específicamente, los errores individuales deben ser independientes entre sí y distribuirse normalmente (con media igual a cero).
- La varianza de los valores de la variable Calorías para cualquier valor x no debe depender de ese valor de x.
- Consideraciones para los residuos:
  - **i.** El gráfico de los residuos contra los valores del predictor x_i debe mostrarlos simétricos alrededor de la línea horizontal y=0 (homocedasticidad).
  - **ii.** El gráfico de los residuos contra los valores de la variable de respuesta y_i debe mostrar una disposición de los puntos que es simétrica alrededor de la línea horizontal y=0 .

iii. El gráfico de residuos contra los valores de predicción $\hat{y}_i$ debe producir una disposición de los puntos simétrica alrededor de la linea y=0.

iv. Los residuos deben tener una distribución normal.

➔ Análisis de la relación entre las variables Fibra y Calorías:

En este caso se tomó como variable respuesta Y a las calorías, y a la variable Fibra como la variable independiente X.

La recta de regresión lineal obtenida por el ajuste es: $\hat{Y} = 117.3635 - 4.3881\,X$

con $\hat{\beta}_0 = 117.3635$ y $\hat{\beta}_1 = -4.3881$

El coeficiente de determinación (que coincide con el coeficiente de correlación al cuadrado) es

$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2}$ con $\hat{y}_i$ el valor ajustado para cada x_i.

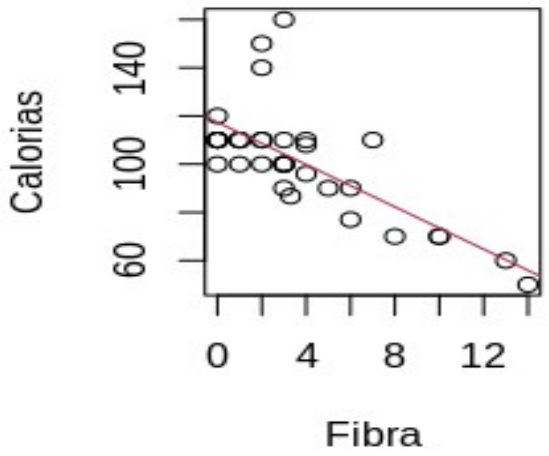
En este caso $R^2 = 0.5112$, es decir que en el modelo de regresión lineal, poco más del 51% de los datos de la variación de las $\hat{y}_i$ se explica por la variable x.

El coeficiente de correlación se estima mediante $Cor(X,Y) = \frac{\sum (x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{(\sum (x_i - \bar{x}_i)^2)(\sum (y_i - \bar{y}_i)^2)}}$

En este caso nos da Cor(X,Y)= -0.715. Este valor es próximo a -1, lo que nos dice que hay una posible relación lineal entre las variables, con pendiente negativa (es decir la variable Calorías disminuye a medida que la variable Fibra aumenta).

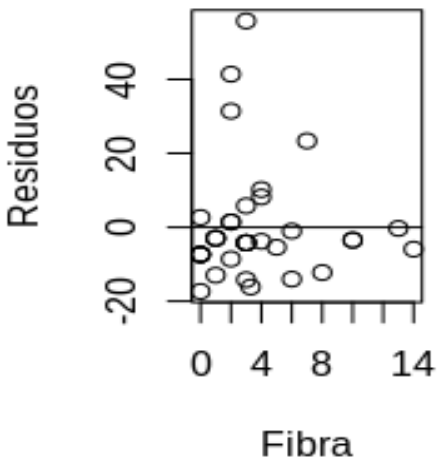
Se estimó la varianza del error σ^2 mediante: $\widehat{\sigma^2} = \frac{SSE}{n-2}$ donde $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

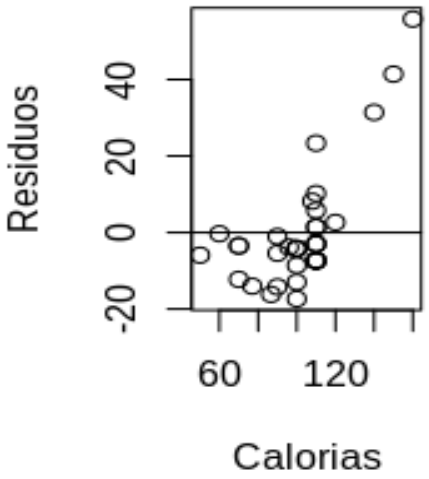
En este caso nos dio el valor de estimación $\widehat{\sigma^2} = 247.12$

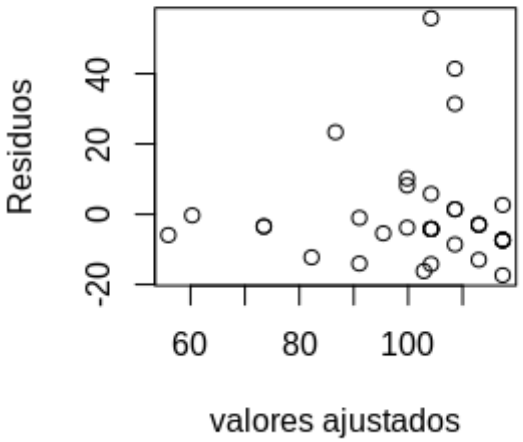


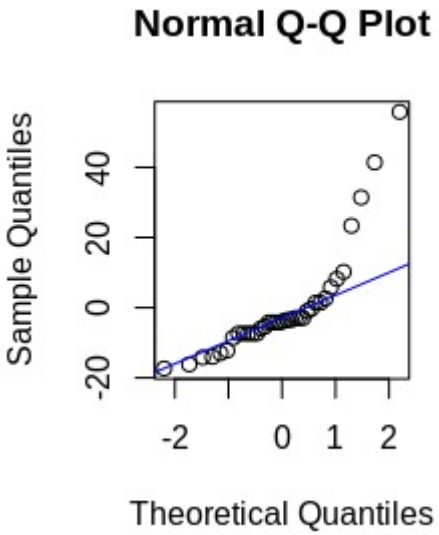
Gráfica de la recta ajustada con los puntos

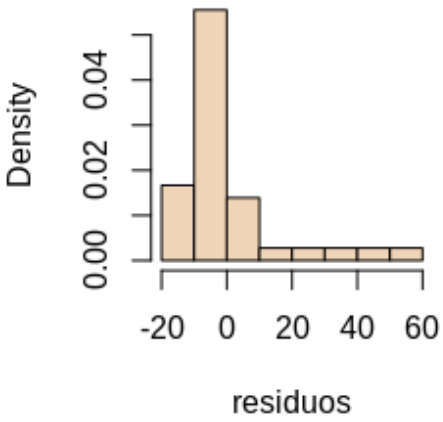
Se observa una posible relación lineal entre las variables, y datos que pueden ser outliers o puntos de influencia. Se realizó un análisis de residuos del ajuste, para estudiar si deben ser eliminados, como se presenta a continuación:

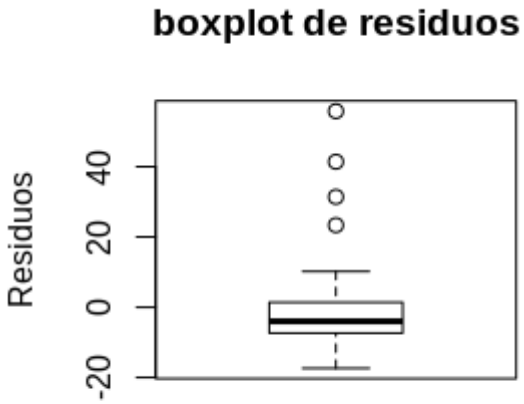












Las gráficas anteriores mostraron la dependencia de los residuos respecto a los valores ajustados, presentándose una gran variabilidad de los mismos en torno a la recta y=0 (lo cual indica la falta de homocedasticidad) en vez de mostrar simetría respecto a la recta y=0; y el gráfico del ajuste normal muestra la no normalidad de la distribución, viéndose una cola a derecha y la presencia de datos atípicos en el diagrama de caja.

En principio, se procedió a eliminar 11 datos atípicos con el fin de mejorar los coeficientes de correlación y determinación con respecto a los valores del primer ajuste, y que los gráficos de residuos cumplan con las hipótesis dadas (homocedasticidad y normalidad). Si la eliminación de outliers es efectiva, estos coeficientes tendrán valor absoluto próximo a 1 y además la recta del ajuste mostrará distancias más pequeñas con respecto a los puntos de la muestra (junto con un S^2 más chico) y puntos más alineados.
Los datos eliminados correspondientes se muestran en la siguiente tabla:

Fibra	Calorías
2	140
2	150
3	160
7	110
3.3	86.7
3	110
0	100
4	110
4	108
6	77
3	90

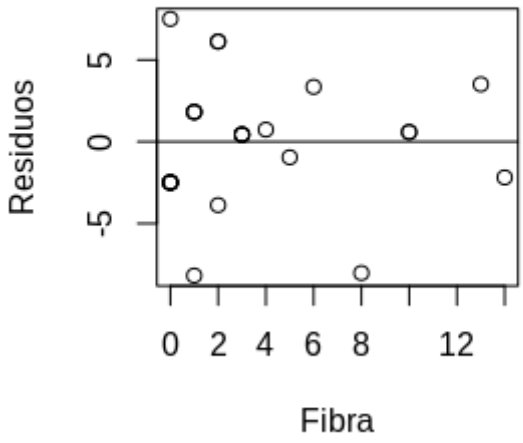
Los outliers encontrados tienen, en su mayoría, la particularidad de que presentan valores calóricos muy grandes, mientras que los gramos de fibra que les corresponden, salvo excepciones, no se alejan demasiado de la media muestral. Esto da un indicio,entre otras causas, de porqué se obtuvo una varianza del error grande que hizo que el ajuste no sea el óptimo.

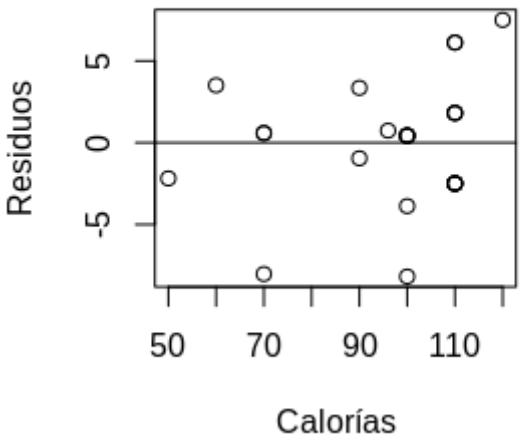
La recta de regresión obtenida es: $\hat{Y}_1 = 112.4928 - 4.3078\, X$

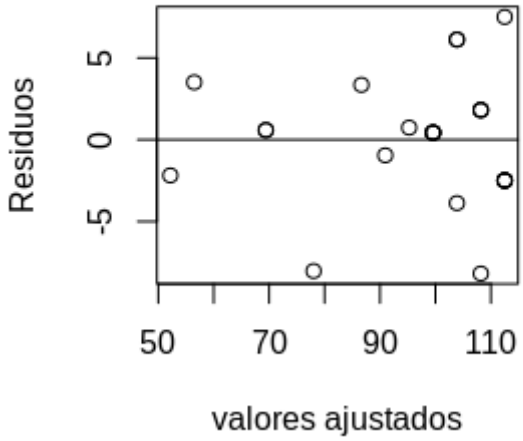
El coeficiente de determinación $R^2 = 0.957$

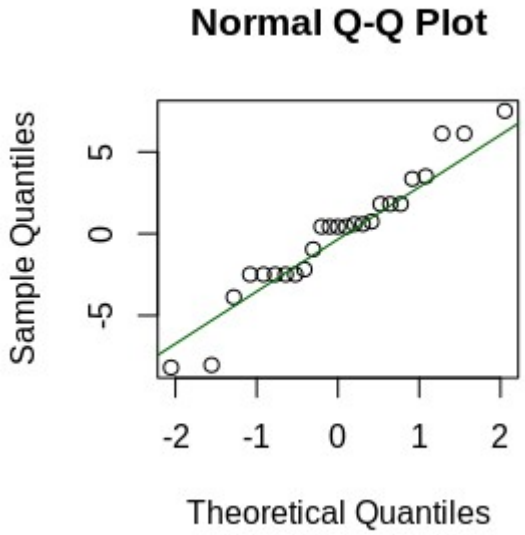
El coeficiente de correlación estimado $Cor(X, Y_1) = -0.978$

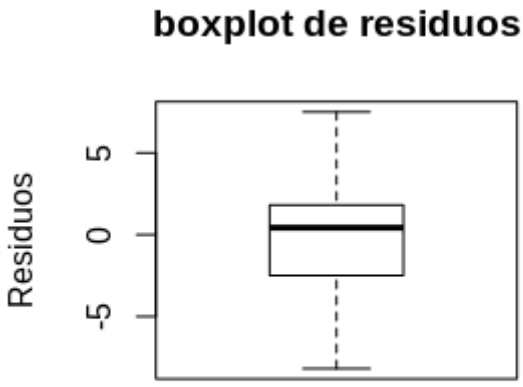
La varianza del error estimada es $\widehat{\sigma^2} = 15.2178$











Los residuos siguen presentando variabilidad respecto a la recta y=0, aunque mucho menor que en el caso del primer ajuste, por lo que hay una mejora en la homocedasticidad. Sin embargo el gráfico de ajuste normal sigue presentando la no normalidad de los residuos, los valores no se ven alineados con respecto a la recta del ajuste, lo cual se refleja también en el histograma y diagrama

de caja (ya que la distribución de los datos no es simétrica: en el caso del boxplot, por ejemplo, la caja muestra datos mucho más concentrados entre la mediana y el tercer cuantil).

El coeficiente de correlación está mas próximo a -1 y el de determinación más próximo a 1, por lo que estos valores también mejoraron. Pero al no cumplirse la hipótesis de normalidad de residuos, no es suficiente con quitar outliers para mejorar el ajuste y aceptar el modelo de regresión lineal.

Ante esta situación, se volvió a la muestra de 36 datos, y se continuó con la búsqueda de una transformación adecuada de la variable respuesta Y; se probó implementar tres de las más usuales que son tomar $\mathbf{log}(\boldsymbol{Y})$, $\sqrt{\boldsymbol{Y}}$ y $\boldsymbol{aY^{-1}}$ para a una constante no nula.

- En primer lugar se usó $\log(Y)$ y como tanto R^2 como $Cor(X,\log(Y))$ daban valores pequeños, se continuó con quitar siete outliers, de modo que la recta estimada presente errores pequeños (que la varianza de los errores, $\widehat{\sigma^2}$, entre los valores ajustados y los de la muestra sean chicos) y que las estimaciones de los coeficientes tengan valor absoluto lo más próximo a 1 posibles. Otro punto importante a tener en cuenta fue que la hipótesis de normalidad de los residuos se cumpliera.

A continuación se presenta una tabla con los datos atípicos encontrados y los resultados del ajuste:

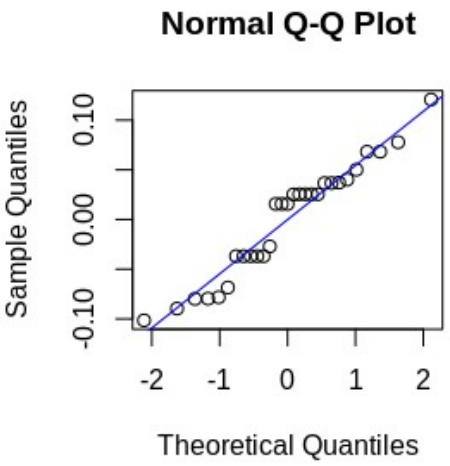
Fibra	Calorías	$Y_2=\log(Y)$
3	160	5.07
2	150	5.01
7	110	4.7
2	140	4.94
4	110	4.7
4	108	4.68
0	100	4.6

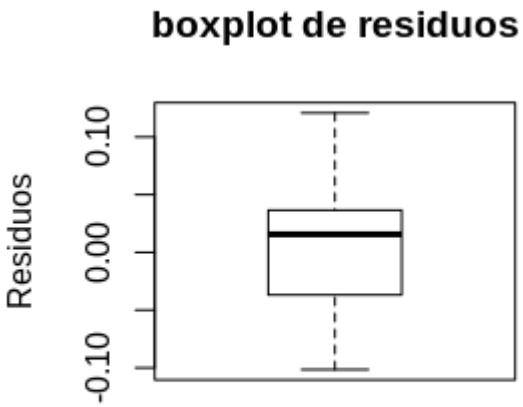
Recta ajustada para $Y_2=\log(Y)$: $\hat{Y}_2=4.73-0.05X$

Coeficiente de determinación estimado: $R^2=0.927$

Coeficiente de correlación estimado: $\hat{Cor}(X,Y_2)=-0.963$

Varianza del error: $\widehat{\sigma^2}=0.003$





Como se puede ver, si bien los coeficientes y la varianza mejoraron, la hipótesis de normalidad y simetría de los residuos sigue sin cumplirse, por lo que esta transformación fue descartada y, por lo tanto, también el modelo.

- Tomando ahora $Y_3=\sqrt{Y}$ se trabajó de igual manera, quitando los siguientes datos atípicos que fueron detectados con el mismo fin que para los casos anteriores:

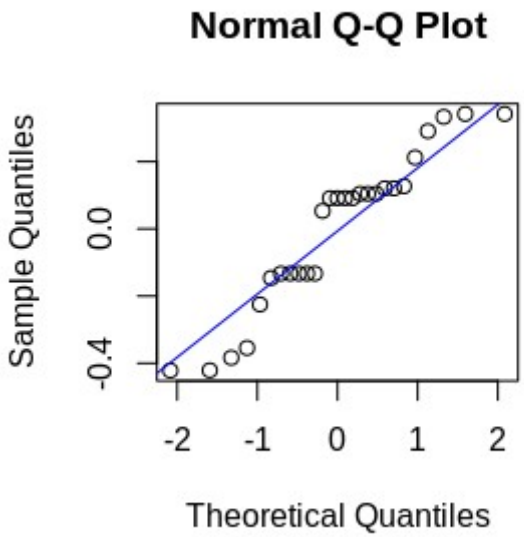
Fibra	Calorías	$Y_3=\sqrt{Y}$
2	140	11.83
2	150	12.25
3.3	86.7	9.31
3	110	10.48
0	100	10
3	160	12.65
7	110	10.48
4	110	10.48
4	108	10.39

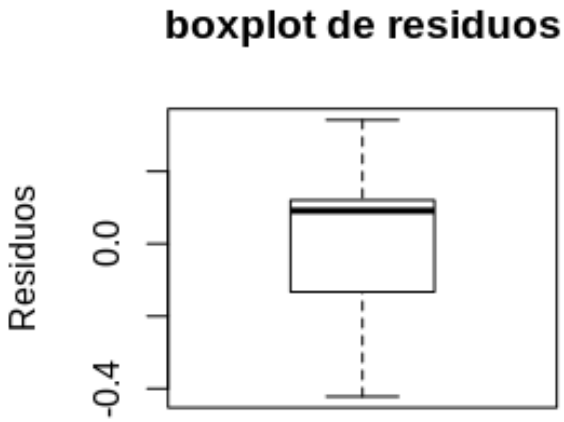
Recta ajustada para $Y_3=\sqrt{Y}$: $\hat{Y}_3=10.62-0.24\,X$

Coeficiente de determinación estimado: $R^2=0.946$

Varianza del error: $\widehat{\sigma^2}=0.055$

Coeficiente de correlación estimado: $\hat{Cor}(X,Y_3)=-0.972$





Esta transformación de la variable respuesta también fue descartada ya que, lejos de producir una mejora comparada con la transformación anterior, la normalidad de residuos no se presenta, y no tienen una distribución simétrica.

- Por último se presentará la transformación que sí fue aceptada, $W=100\,Y^{-1}$, con el siguiente procedimiento:

Se aplicó la transformación de la variable Y y en principio el modelo lineal quedó de la forma

$100\,Y^{-1}=\beta_0+\beta_1 X+\epsilon$ o equivalentemente $Y^{-1}=\frac{\beta_0}{100}+\frac{\beta_1}{100}X+\tilde{\epsilon}$ con $\tilde{\epsilon}=\frac{\epsilon}{100}$

donde la constante 100 se eligió para mejorar las escalas a la hora de hacer los análisis y gráficos.

Al igual que en los casos anteriores se realizó el ajuste de regresión y se obtuvieron las estimaciones de los coeficientes y varianza del error necesarios para estudiar qué tan bueno fue el mismo. Como estos datos y los gráficos de residuos no mejoraron lo suficiente, se comenzaron a eliminar outliers de modo que la recta de regresión estimada arroje diferencias pequeñas entre los valores ajustados y los puntos de la muestra.

Los outliers fueron obteniéndose y eliminándose de a uno, y además se detectó un punto de alta influencia (correspondiente a los valores de 14g de Fibra y w=2 (para 50 calorías)) que se observaba alejado del resto de los valores y no parecía estar alineado; luego de dejarlo dentro y fuera del análisis los resultados de la regresión cambiaron, por lo que debió ser quitado porque modificaba sustancialmente la estimación). Su eliminación dejó ver otros datos atípicos que se encontraban enmascarados y que debieron ser eliminados con el mismo procedimiento.

Fibra	Calorías	$W=100\,Y^{-1}$
3	160	0.625
7	110	0.909
2	150	0.666
14	50	2
2	140	0.714
4	110	0.909
4	108	0.925

La tabla muestra los datos atípicos eliminados en el orden en el que fueron encontrados, teniendo en cuenta el punto de alta influencia marcado con color.

Las rectas ajustadas con y sin el punto de alta influencia son las siguientes:

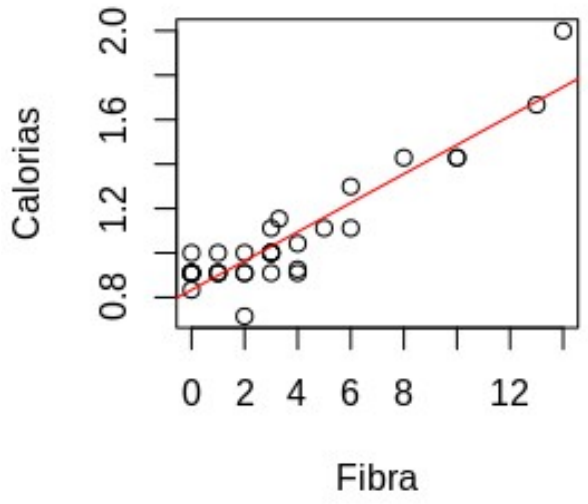
***Con el punto de alta influencia:**

Recta ajustada: $W=0.833+0.065\,X$

Coeficiente de determinación estimado: $R^2=0.85$

Coeficiente de correlación estimado: $\hat{Cor}(X,W)=0.92$

Varianza del error: $\hat{\sigma^2}=0.01$



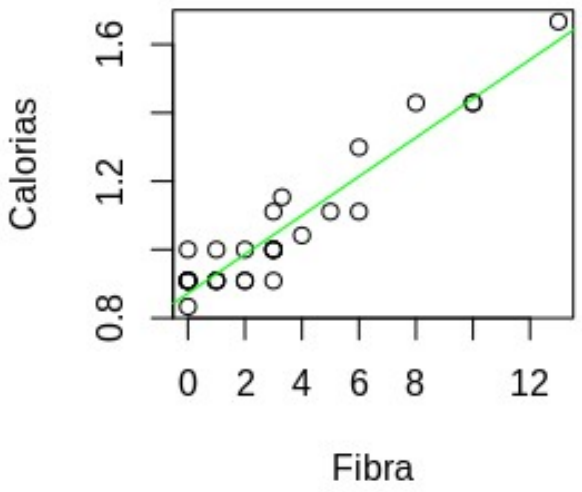
***Sin punto de alta influencia:**

Recta ajustada: $W = 0.8725 + 0.0568\,X$

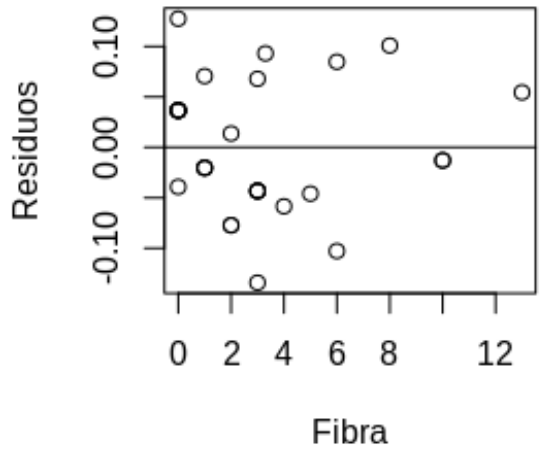
El coeficiente de determinación es: $R^2 = 0.899$

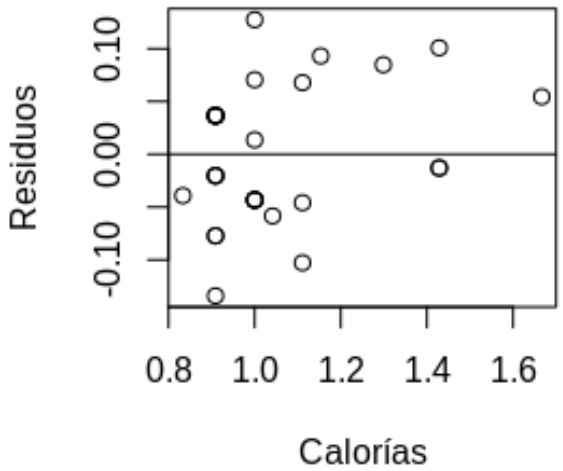
El coeficiente de correlación: $Cor(X, W) = 0.948$

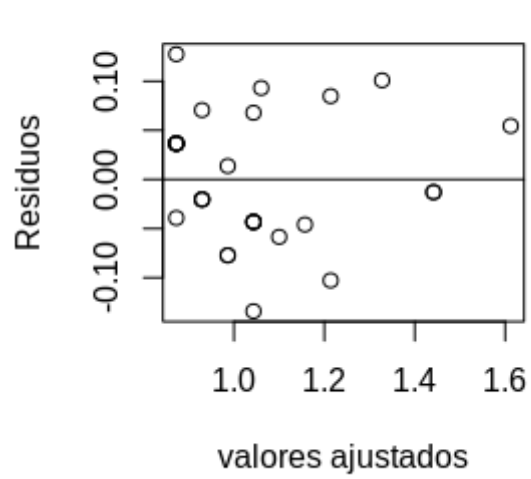
Varianza del error: $\widehat{\sigma^2} = 0.00438$

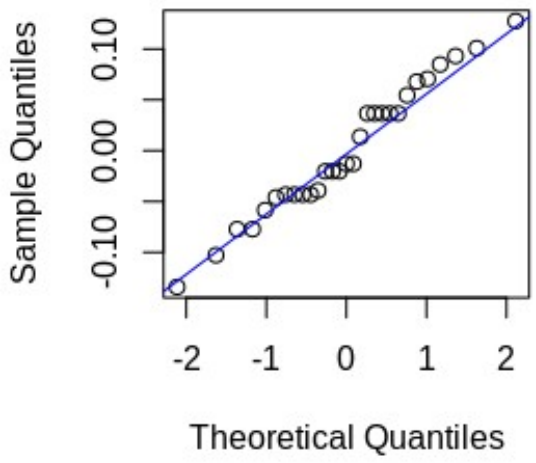


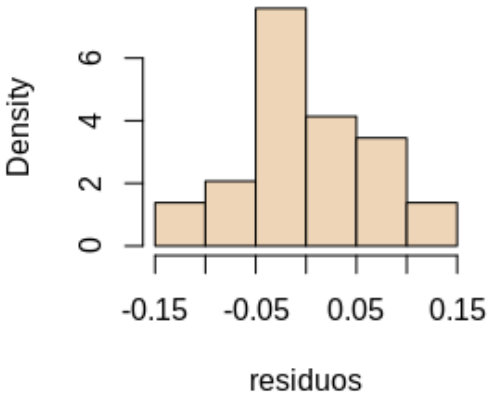
Con este último ajuste, las estimaciones de los coeficientes anteriores mejoraron y la varianza $\widehat{\sigma^2}$ disminuyó notoriamente, por lo tanto se puede proceder con el estudio de los residuos para saber si las hipótesis mencionadas se cumplen:

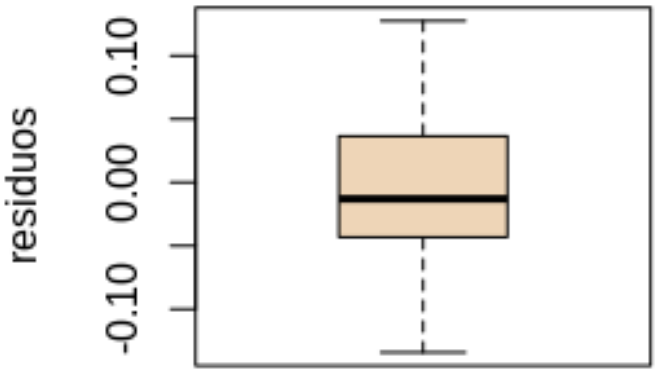












Se observan gráficos de residuos más simétricos respecto a la recta y=0, es decir que se cumplen las hipótesis de homocedasticidad, el ajuste con la recta normal y el histograma junto con el diagrama de caja indican que la distribución de residuos es razonablemente normal y la ausencia de outliers.

Por lo tanto, luego de estas modificaciones, se está en condiciones de decir que las hipótesis del modelo lineal se cumplen, y los datos se presentan más consistentes también con las hipótesis necesarias para poder continuar haciendo inferencias sobre los parámetros.

El resumen de 5 números más la media y desviación estándar muestral para la variable Fibra ahora son:

Mínimo	Primer cuantil	Mediana	Tercer cuantil	Máximo	Media muestral	Desviación estándar muestral	Distancia intercuartil
0	1.0	3.0	4.0	13.0	3.217	3.415	3.0

El siguiente paso es obtener intervalos de confianza para los parámetros $\beta_0 y \beta_1$ y la varianza de la variable respuesta:

Intervalos de confianza de nivel 0,95:

* Para la pendiente β_1 : $IC_{1-\alpha}(\beta_1)=[\hat{\beta}_1-t_{1-\frac{\alpha}{2},n-2}\,s\sqrt{c_{11}},\hat{\beta}_1+t_{1-\frac{\alpha}{2},n-2}\,s\sqrt{c_{11}}]$

con S un estimador insesgado de σ y $C_{11}=\frac{1}{S_{xx}}$ $S_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2$

que en este caso, para $\alpha=0.05$ y n= 29 nos queda:

$$IC_{0.95}(\beta_1)=[0.0566,0.0571]$$

***** Para la ordenada al origen β_0 **:** $IC_{1-\alpha}(\beta_0)=[\hat{\beta}_0-t_{1-\frac{\alpha}{2},n-2}\,s\sqrt{c_{00}},\hat{\beta}_0+t_{1-\frac{\alpha}{2},n-2}\,s\sqrt{c_{00}}]$

con $C_{00}=\frac{\sum x_i^2}{nS_{xx}}$

y para $\alpha=0.05$ y n=29 es:

$$IC_{0.95}(\beta_0)=[0.871,0.873]$$

***** Para la varianza del error σ^2: $IC_{1-\alpha}(\sigma^2)=[\frac{(n-2)s^2}{\chi^2_{1-\frac{\alpha}{2},n-2}},\frac{(n-2)s^2}{\chi^2_{\frac{\alpha}{2},n-2}}]$

para $\alpha=0.05$, n=29 y **s^2= 0.00438** nos queda:

$$IC_{1-\alpha}(\sigma^2)=[0.0043,0.0045]$$

Test de hipótesis para los parámetros:

En todos casos se utilizará el nivel de significación $\alpha=0.05$

*** Se quiere ver si la ordenada al origen β_0 es distinta del valor 0.8:**

Se planteó el siguiente test: $\begin{cases} H_0:\beta_0=0.8 \\ H_1:\beta_0\neq 0.8 \end{cases}$

El estadístico de prueba utilizado es $T=\frac{\hat{\beta}_0-0.8}{s\sqrt{c_{00}}}\sim t_{n-2}$ (si $\beta_0=0.8$ es verdadera)

con t_{n-2} distribución t-Student con n-2 grados de libertad.

La región de rechazo está dada por $C_{\beta_0}=\{t\in\mathbb{R}:|t|>t_{1-\frac{\alpha}{2},n-2}\}$

Para esta muestra tenemos: $t \sim t_{27}$, $C_{\beta_0} = \{t \in \mathbb{R} : |t| > 2.052\}$ y t = 4.262

se tiene un p-valor: $P(|T| > 4.262) = 2P(T > 4.262) = 2(1 - P(T \leq 4.262)) = 0.00022$

El estadístico de prueba está en la región de rechazo C_{β_0}
El p-valor es menor al nivel de significación $\alpha = 0.05$ por lo tanto podemos podemos rechazar la hipótesis de que $\beta_0 = 0.8$ con un nivel de significación 0.05

*** Para la pendiente se plantea el siguiente test con el fin de ver si β_1 es mayor que 0.05:**

$$\begin{cases} \boldsymbol{H_0 : \beta_1 = 0.05} \\ \boldsymbol{H_1 : \beta_1 > 0.05} \end{cases}$$

el siguiente estadístico de prueba utilizado es (si $\beta_1 = 0.05$ es verdadera)

$$T = \frac{\hat{\beta}_1 - 0.05}{s\sqrt{c_{11}}} \sim t_{n-2}$$

con t_{n-2} distribución t-Student con n-2 grados de libertad.

La región de rechazo está dada por $C_{\beta_1} = \{t \in \mathbb{R} : t > t_{1-\alpha, n-2}\}$

Para esta muestra tenemos: $t \sim t_{27}$ $C_{\beta_1} = \{t \in \mathbb{R} : t > 1.703\}$ y t = 1.883

se tiene un p-valor: $P(T > 1.883) = 1 - P(T \leq 1.883) = 0.035$

Para este test, el estadístico de prueba pertenece a la región de rechazo C_{β_1} y el p-valor obtenido es menor que el nivel de significación $\alpha = 0.05$ por lo que se puede rechazar la hipótesis nula $\beta_1 = 0.05$

***Intervalo de confianza para la media de la variable respuesta:**

Se calcularon intervalos de confianza para E(Y), de nivel 0.95 con $E(Y) = \beta_0 + \beta_1 \tilde{x}$ con $\tilde{x}$ un valor constante fijo.
Dado $\widehat{E(Y)} = \hat{\beta}_0 + \hat{\beta}_1 \tilde{x}$ el intervalo de confianza es:

$$IC_{1-\alpha} = [\hat{\beta}_0 + \hat{\beta}_1 \tilde{x} - t_{1-\alpha/2} S \sqrt{\frac{1}{n} + \frac{(\tilde{x} - \bar{x})^2}{S_{xx}}}, \hat{\beta}_0 + \hat{\beta}_1 \tilde{x} + t_{1-\alpha/2} S \sqrt{\frac{1}{n} + \frac{(\tilde{x} - \bar{x})^2}{S_{xx}}}]$$

con $\alpha = 0.05$, n=29 y $S_{xx} = 326.7$ y para distintos valores de $\tilde{x}$:

La longitud de los intervalos es $L=2\,t_{1-\alpha/2}\,S\sqrt{\frac{1}{n}+\frac{(\tilde{x}-\bar{x})^2}{S_{xx}}}$

- Para $\tilde{x}=\bar{x}$ =3.21 $\boxed{IC_{0.95}=[1.03, 1.08]}$ $L=0.05043$
- Para $\tilde{x}=primer\,cuartil$ =1.00 $\boxed{IC_{0.95}=[0.89, 0.95]}$ $L=0.06045$
- Para $\tilde{x}=tercer\,cuartil$ =4.00 $\boxed{IC_{0.95}=[1.07, 1.13]}$ $L=0.05179$

por lo que se puede observar que a medida que x se acerca al promedio, los intervalos de confianza tienen menor longitud.

***Intervalo de predicción para la variable respuesta:**

Se calcularon intervalos de predicción para la variable respuesta Y, de nivel 0.95 cuando $\tilde{x}$ toma un valor constante fijo:

$$IC_{1-\alpha}=[\hat{\beta}_0+\hat{\beta}_1\,\tilde{x}-t_{1-\alpha/2}\,S\sqrt{1+\frac{1}{n}+\frac{(\tilde{x}-\bar{x})^2}{S_{xx}}}\,,\hat{\beta}_0+\hat{\beta}_1\,\tilde{x}+t_{1-\alpha/2}\,S\sqrt{1+\frac{1}{n}+\frac{(\tilde{x}-\bar{x})^2}{S_{xx}}}]$$

con $\alpha=0.05$, n=29 y $S_{xx}=326.7$ y para distintos valores de $\tilde{x}$:

La longitud de los intervalos es $L=2\,t_{1-\alpha/2}\,S\sqrt{1+\frac{1}{n}+\frac{(\tilde{x}-\bar{x})^2}{S_{xx}}}$

- Para $\tilde{x}=\bar{x}$ =3.21 $\boxed{IC_{0.95}=[0.917, 1.194]}$ $L=0.1381$
- Para $\tilde{x}=primer\,cuartil$ =1.00 $\boxed{IC_{0.95}=[0.79, 1.07]}$ $L=0.2782$
- Para $\tilde{x}=tercer\,cuartil$ =4.00 $\boxed{IC_{0.95}=[0.96, 0.24]}$ $L=0.2765$

Se observa que el intervalo de predicción cuando x toma el valor del promedio, tiene una longitud menor que cuando toma los valores del primer y tercer cuartil, por lo que se puede decir que a medida que x se acerca a $\bar{x}$, los intervalos de predicción ganan menor longitud; y en comparación con los intervalos de confianza analizados para los mismos valores que puede tomar $\tilde{x}$, sus longitudes por lo general son mayores.

Concluyendo con un breve repaso...

Se ha partido de un conjunto de datos que estaba bastante alejado de cumplir las condiciones necesarias para poder establecer una relación lineal entre sus variables, pero que ha permitido trabajar sobre ellas intentando una serie de modificaciones y eliminación de datos atípicos e influyentes de manera hasta finalmente aceptar el modelo de regresión más adecuado para establecer un modelo lineal que mejor se ajusta a la hora de predecir las calorías de los cereales según los gramos de fibra.

El modelo que finalmente cumplió con las hipótesis es

$$Y^{-1} = \frac{\beta_0}{100} + \frac{\beta_1}{100} X + \frac{\epsilon}{100}$$

donde $\epsilon \sim N(0, \widehat{\sigma^2})$ con la transformación de la variable respuesta, con respecto a los datos originales, dada por $W = 100\, Y^{-1}$

Al eliminar outliers y un punto de influencia la muestra se redujo a 29 observaciones. Los parámetros pendiente y ordenada al origen son significativos y se pudieron establecer intervalos de predicción y de confianza tanto para estos como para la variable respuesta y su media.

Si bien el objetivo inicial de establecer una relación lineal entre las fibras y calorías de una porción de cereal no fue posible, se logró aceptar un modelo lineal anteriormente planteado (entre la variable X correspondiente a Fibra y la variable $100\, Y^{-1}$ con Y correspondiente a Calorías) con el cual , para los cereales que generalmente son consumidos en dietas nutricionales o simplemente como parte de una alimentación diaria es posible predecir, con algo de desarrollo y dando pie a posteriores análisis sobre el mismo producto, las calorías que tiene cada porción de cereal a partir de los valores nutritivos de otros nutrientes como, en el caso dado para este trabajo, la fibra.

Bibliografía:

- Wackerly,Mendenhall,Scheaffer,"Estadistica Matematica con Aplicaciones", séptima edición
- Tenko Raykov, George A.Marcoulides, "Basic Statistics- An introduction with R"
- https://www.asociacioncereales.es/asociacion/cereales-y-alimentacion/mitos-y-realidades/
- https://www.vitonica.com/alimentos/las-calorias-y-nutrientes-de-diferentes-cereales-de-desayuno
- https://nutricio.es/los-cereales-de-desayuno/
- https://es.wikipedia.org/wiki/Cereal#Caract%C3%ADsticas_nutritivas_y_efectos_sobre_la_salud
- https://vincentarelbundock.github.io/Rdatasets/datasets.html
- https://rpubs.com/Cristina_Gil/Regresion_Lineal_Simple