*Corrección Taller 1 modelación lineal.*

*Cristian Gañan - Valentina Ruiz - María Isabel Vásquez - Tatiana Marín - Daniel Marín - Marlon Tejada*

Con el fin de iniciar la resolución del taller propuesto, se procedió a llevar los datos a R y a generar una variable combinada que permitiese trabajar con todas las variables dadas. Tal variable combinada fue decidida en conjunto y es:

Donde d2h es la variable combinada (independiente), DAP es diámetro a la altura del pecho (1.30 m) y Alt es la altura en metros del árbol.

Posterior a la definición de d2h, se realiza un análisis grafico como primera aproximación al comportamiento de los datos, dando como resultado la Figura 1.

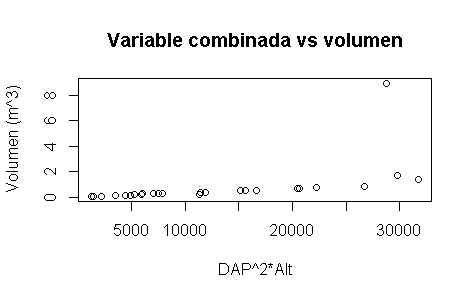


Figura 1. Grafica inicial.

Al observar la Figura 1 puede notarse que hay un dato extraño, que pareciese ser un outlier, lo cual será analizado con más herramientas en el desarrollo de este taller.

**Punto 1:** **Proponga, lógica y justificadamente, sin ignorar ninguna de las variables, en su forma más simple, un modelo de regresión lineal simple y resuélvalo de manera empírica, explicando sus procedimientos**

Como se menciono anteriormente, existe un dato atípico en el set de datos, el cual fue aislado para la generación del modelo empírico que se explicara a detalle a continuación.

Para la generación del modelo empírico, se escogieron dos puntos en la gráfica de datos (Figura 1), cuyas coordenadas se muestran en la Tabla 1, garantizando de forma visual que la línea que une tales puntos representara el comportamiento de la generalidad de los datos. Posterior a la determinación de los dos puntos, y de la línea que los une, se calculó la pendiente con el procedimiento matemático estándar (Ecuación 1)

|  |  |  |
| --- | --- | --- |
| Punto | Coordenada x (d2h) | Coordenada y (vol.) |
| Inicial | 1298.46624 | 0.048 |
| final | 31758.03892 | 1.300 |

Tabla 1. Coordenadas modelo empírico.

Ecuación 1. Pendiente de una recta.

resultando igual a 0.0000411. Luego, mediante la ecuación de la recta (*Ecuación 2*), se procedió a usar la pendiente encontrada e igualando x a 0 para hallar el intercepto de la ecuación, el cual es −0.0053.

*Ecuación 2. Ecuación de una Recta.*

Al analizar el intercepto hallado, se puede notar que este es muy cercano a 0, lo cual tiene mucho sentido al analizar que si la variable combinada, esta formada por un diámetro o una altura igual a 0, el árbol no podría tener volumen.

La ecuación hallada como tal para este modelo empírico es la Ecuación 3, y su correspondiente grafica es la Figura 2 :

Ecuación 3. Ecuación del modelo empírico.

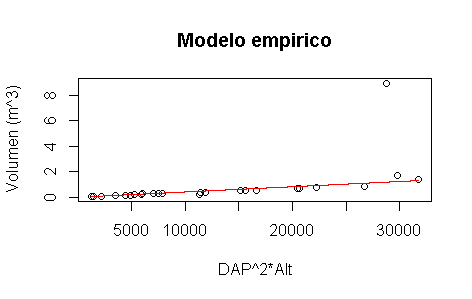


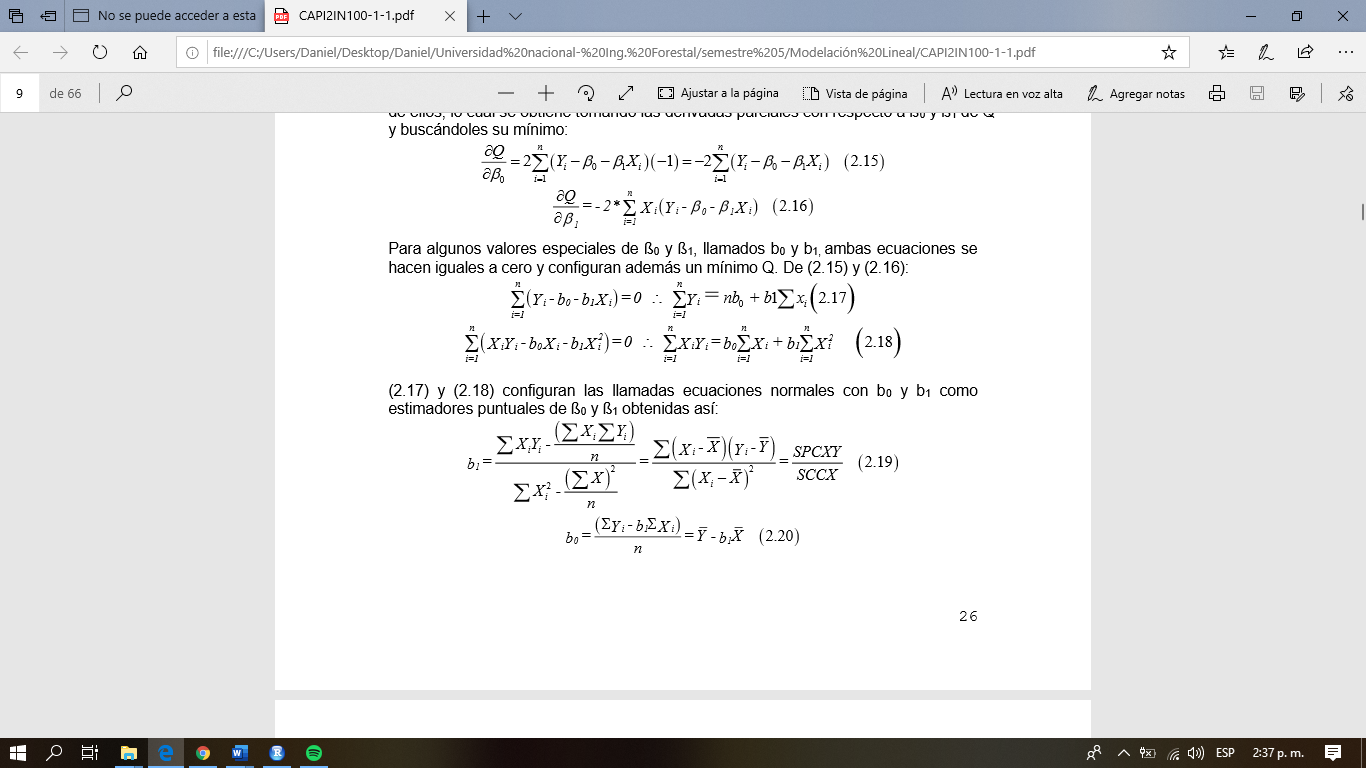
Figura 2. Modelo empírico.

**Punto 2**: **Modelación:**

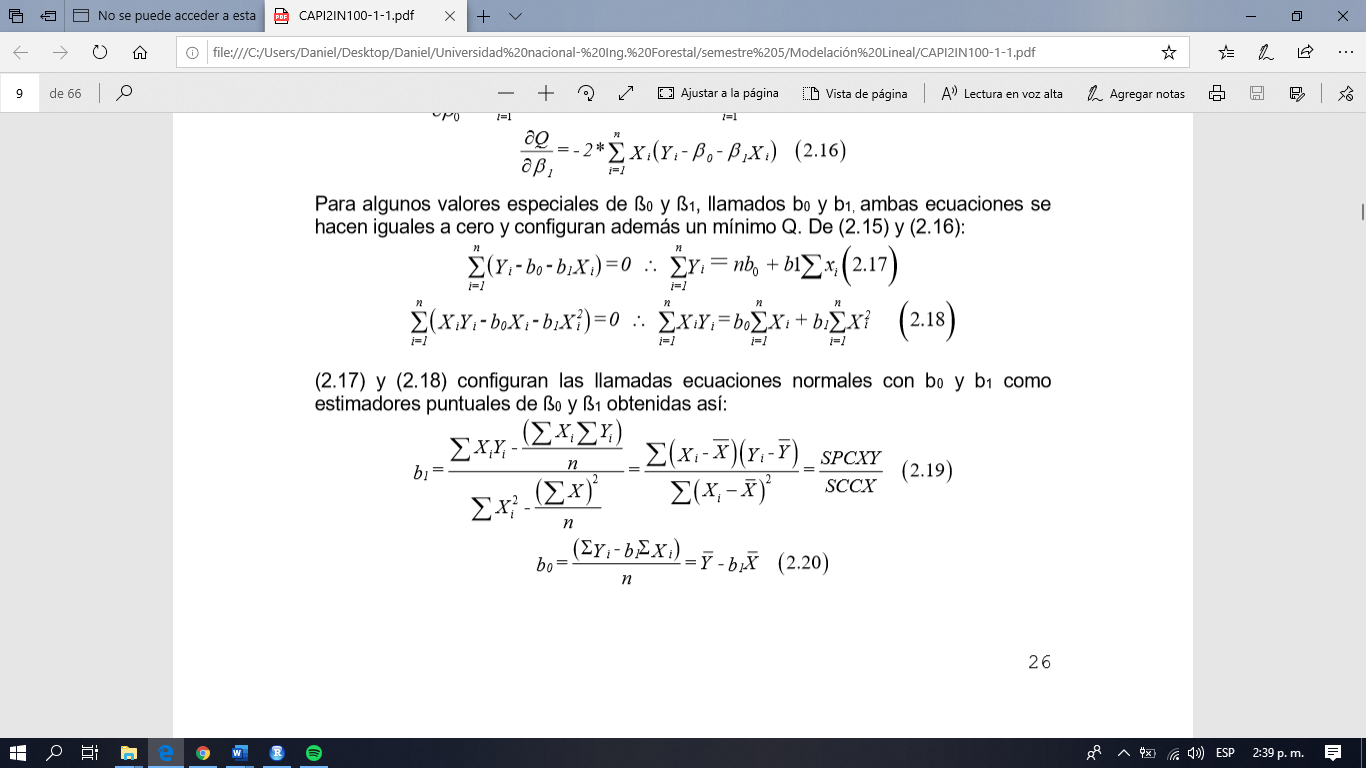
**2.1** **Encuentre sus parámetros de acuerdo con el método de los mínimos cuadrados**

Se encuentra el modelo siguiendo el método de los mínimos cuadrados, utilizando los procedimientos vistos en clase usando las sumas de la variable independiente (combinada), llamada x por facilidad, de la variable dependiente (volumen) llamada y, la suma de cuadrados de la variable d2h enunciada como sx2 y el número de datos llamado n.

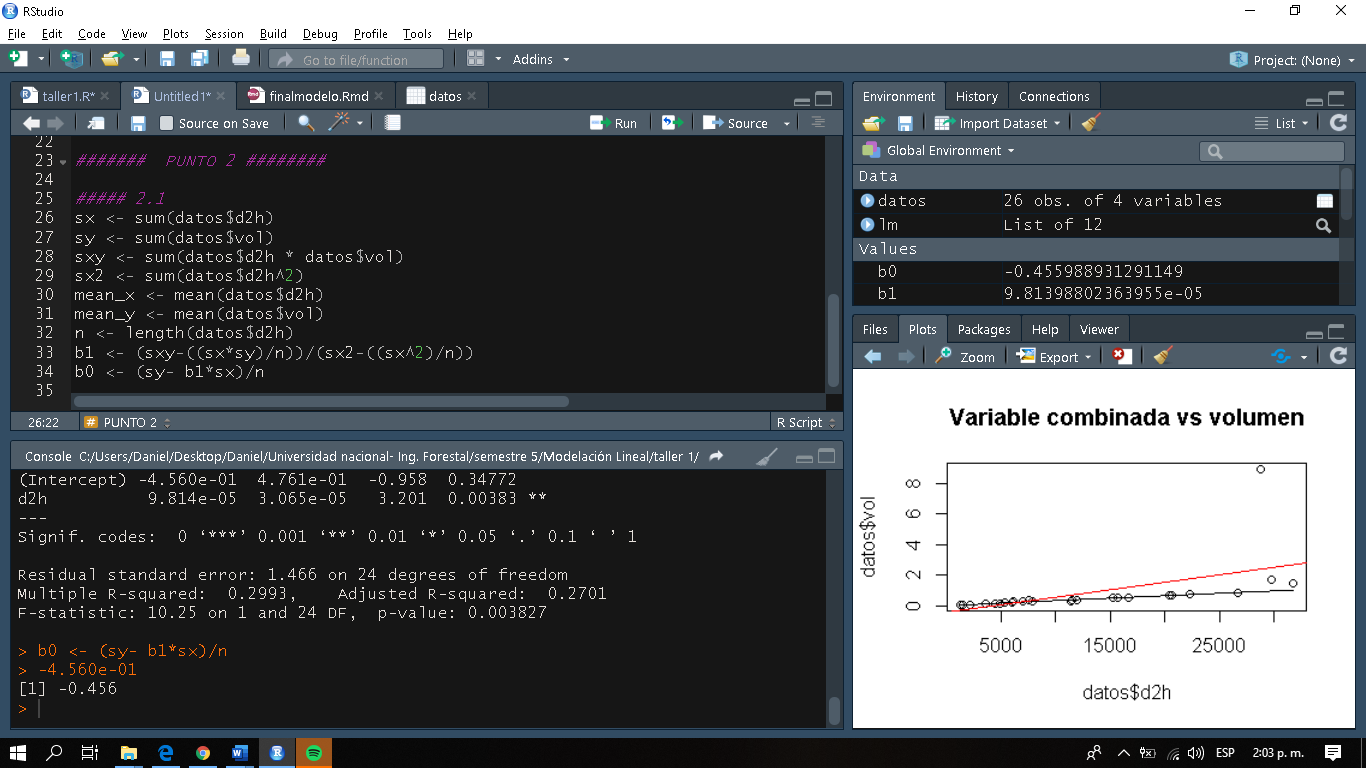
Para hallar b1 y b0 se usaron respectivamente la Ecuación 4 y Ecuación 5, aplicadas en el Script 1.



Ecuación 4. Estimador de b0.



Ecuación 5. estimador de b0



Script 1. estimación de b0 y b1.

Los resultados del Script 1 son (Tabla 2):

|  |  |
| --- | --- |
|  | Valor |
| sx | 321803 |
| sy | 19.726 |
| sxy | 468702.2 |
| sx^2 | 6271057143 |
| n | 26 |
| bo | -0.4559889 |
| b1 | 9.813988e-05 |

Tabla 2. estimación de b0 y b1

Con los estimadores de b0 y b1 se puede expresar el modelo como:

Y su grafica es (Figura 3):

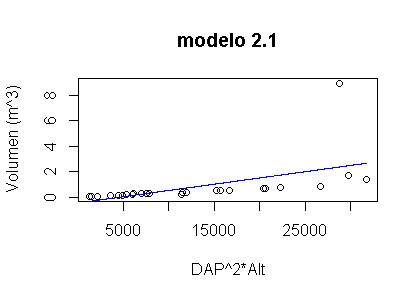


Figura 3. modelo 2.1

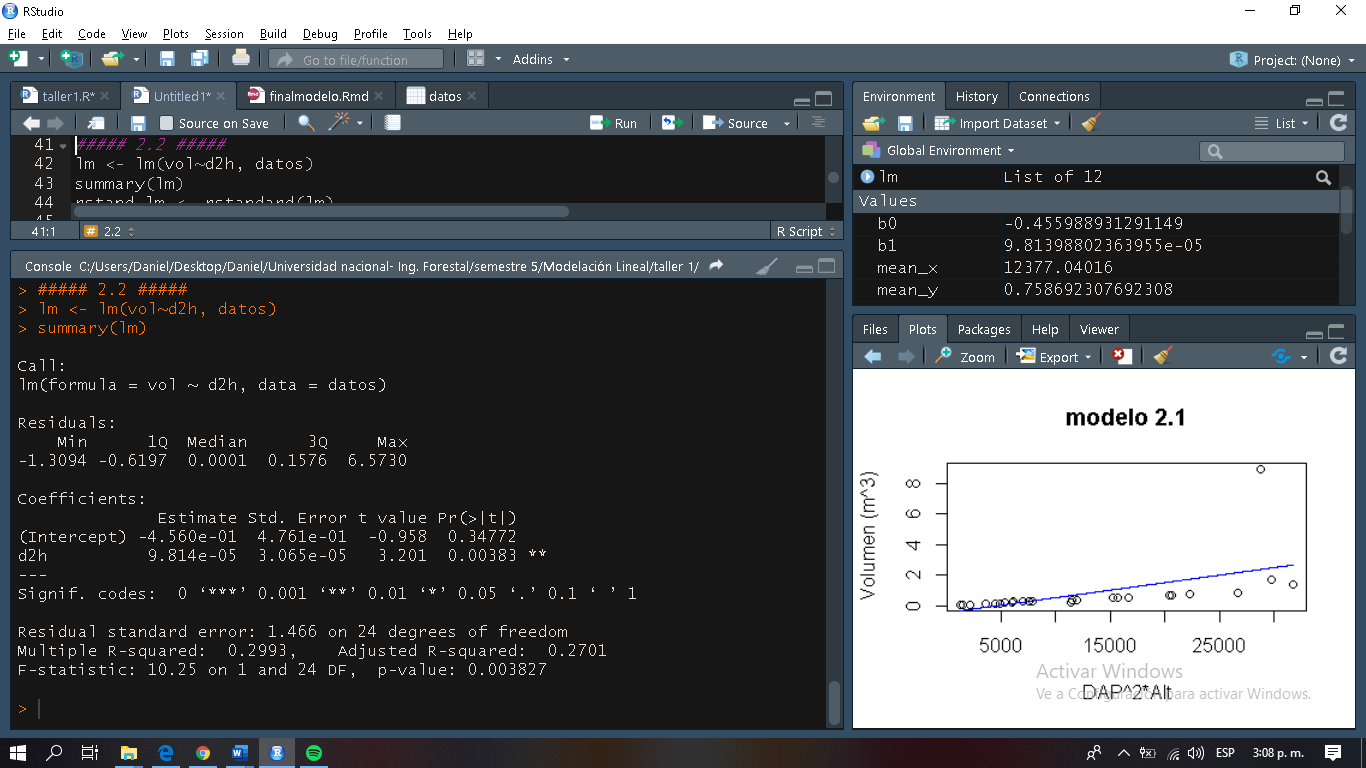
Al ver el modelo anterior, puede notarse que este se ve muy influenciado por el aparente outlier considerado con anterioridad, esto podría afectar de forma grave las predicciones realizadas con tal modelo.

**2.2 Ajuste por medio de R el modelo elegido en 1 y analícelo profusamente. (Debe incluir todas sus hipótesis y los intervalos de confianza para los parámetros del modelo ajustado)**

Con el fin de ajustar el modelo en R se utiliza el siguiente script (Script 2).

En el resultado del modelo (summary) se puede observar que tal modelo corresponde al generado por el método de los mínimos cuadrados, es significativo (p-value = 0.003827) y su intercepto también lo es (P\_value = 0.00383). El hecho de que el intercepto no haya sido significativo corresponde a lo explicado en el modelo empírico (si el diámetro o altura son 0, el volumen también deberá ser 0).

No se realiza una nueva grafica dado que esta corresponderá con la gráfica del modelo de los mínimos cuadrados (Figura 3. modelo 2.1)



Script 2. Modelo en R.

Para corroborar que el modelo si ajuste de forma correcta, y que la observacion aparentemente remota observada con anterioridad lo sea en realidad se procede a hacer una grafica de residuales estandarizados (Figura 4) con el Script 3

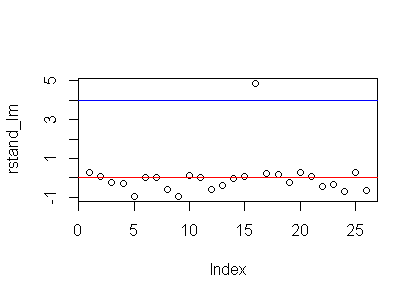
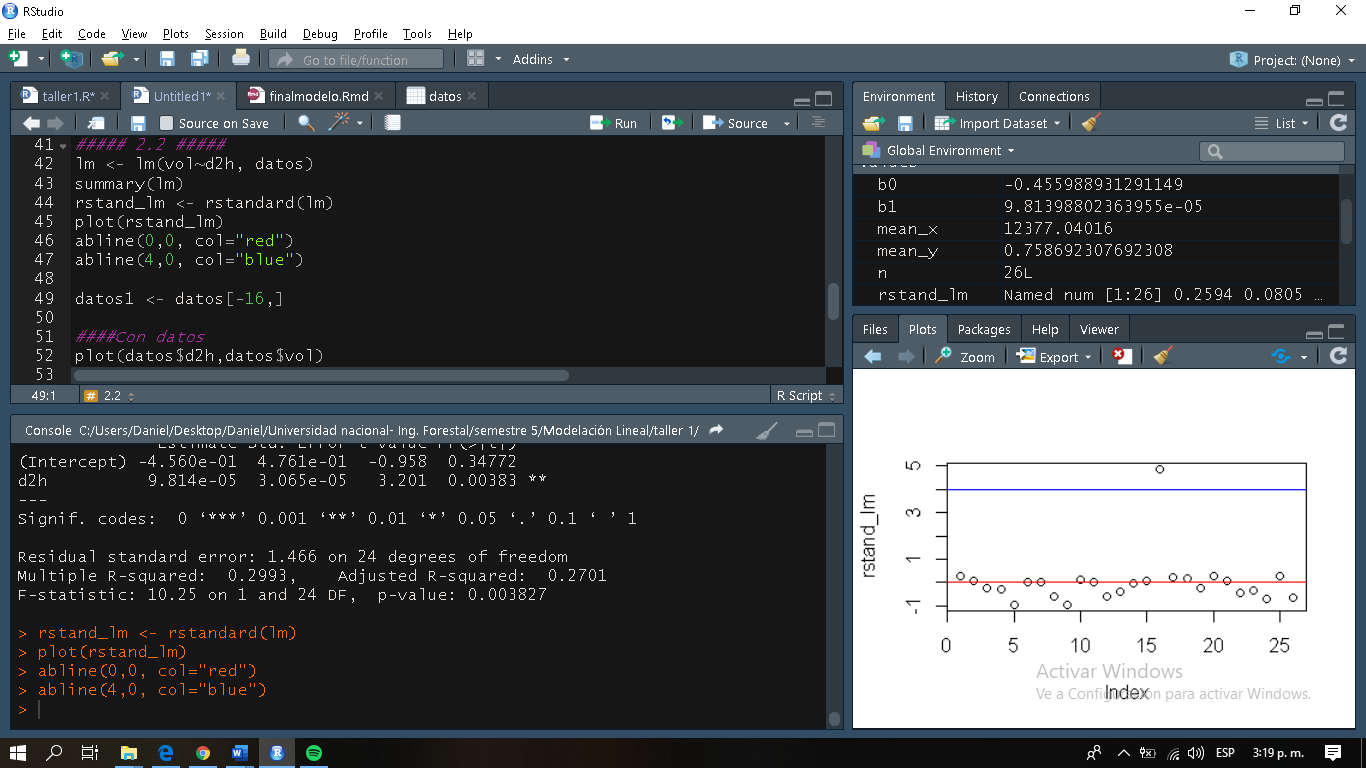


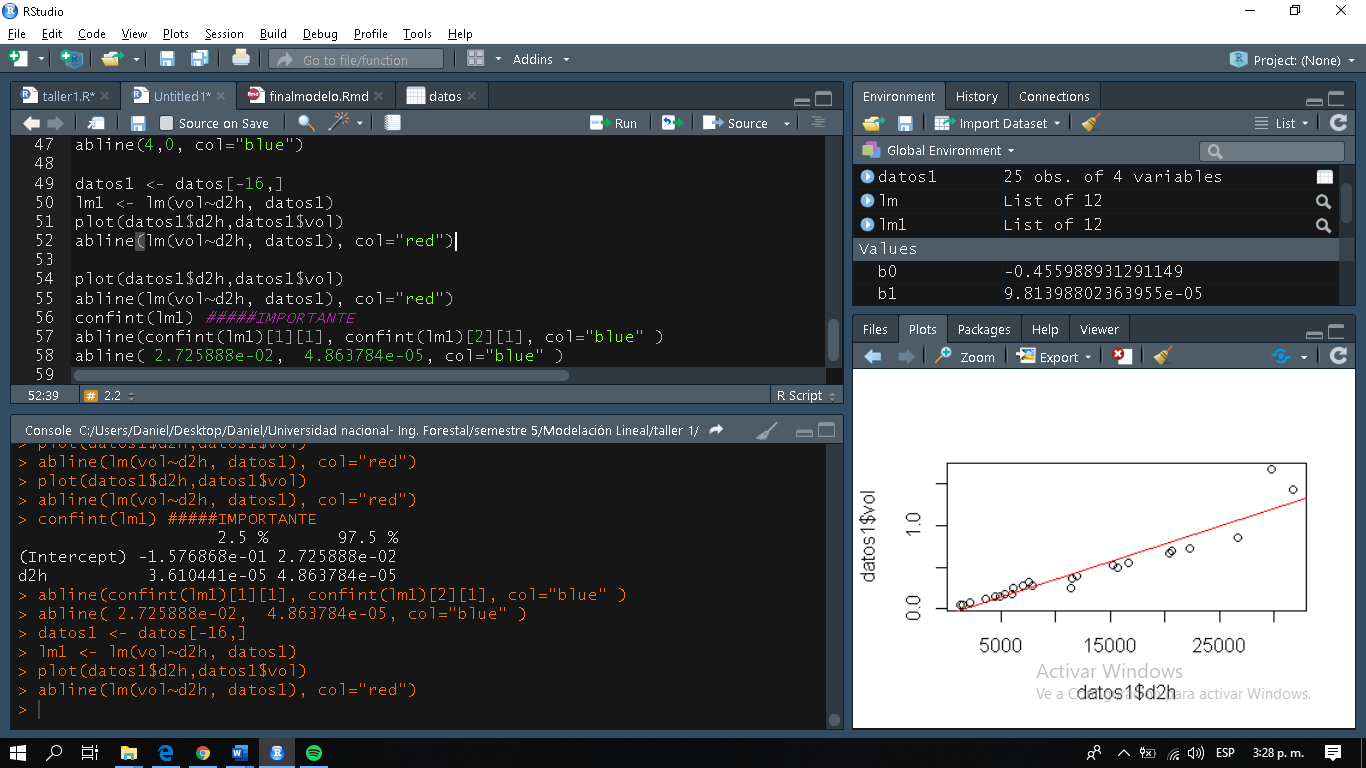
Figura . Residuales estandarizados.



Script . Gráfico de Residuales estandarizados.

En la Figura 4 puede evidenciarse que el dato que se pensaba como outlier efectivamente lo es, bajo el criterio de estar mas de 4 desviaciones estándar alejado de la media, por ello, se toma la decisión de dejarlo de lado y trabajar con un nuevo set de datos que no lo contenga. La exclusión de este dato se hace usando la función identify en la Figura 3, resultando ser el dato número 16.

Con el nuevo set de datos se procede a realizar un nuevo modelo, usando el siguiente código (Script 4) y resultando la gráfica (Figura 5) y el resumen de este nuevo modelo (Figura 6):



Script 4. Modelo lineal sin outlier.

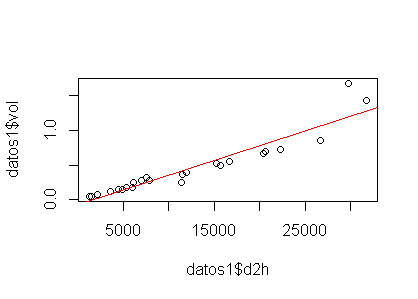


Figura . Modelo lineal sin outlier.

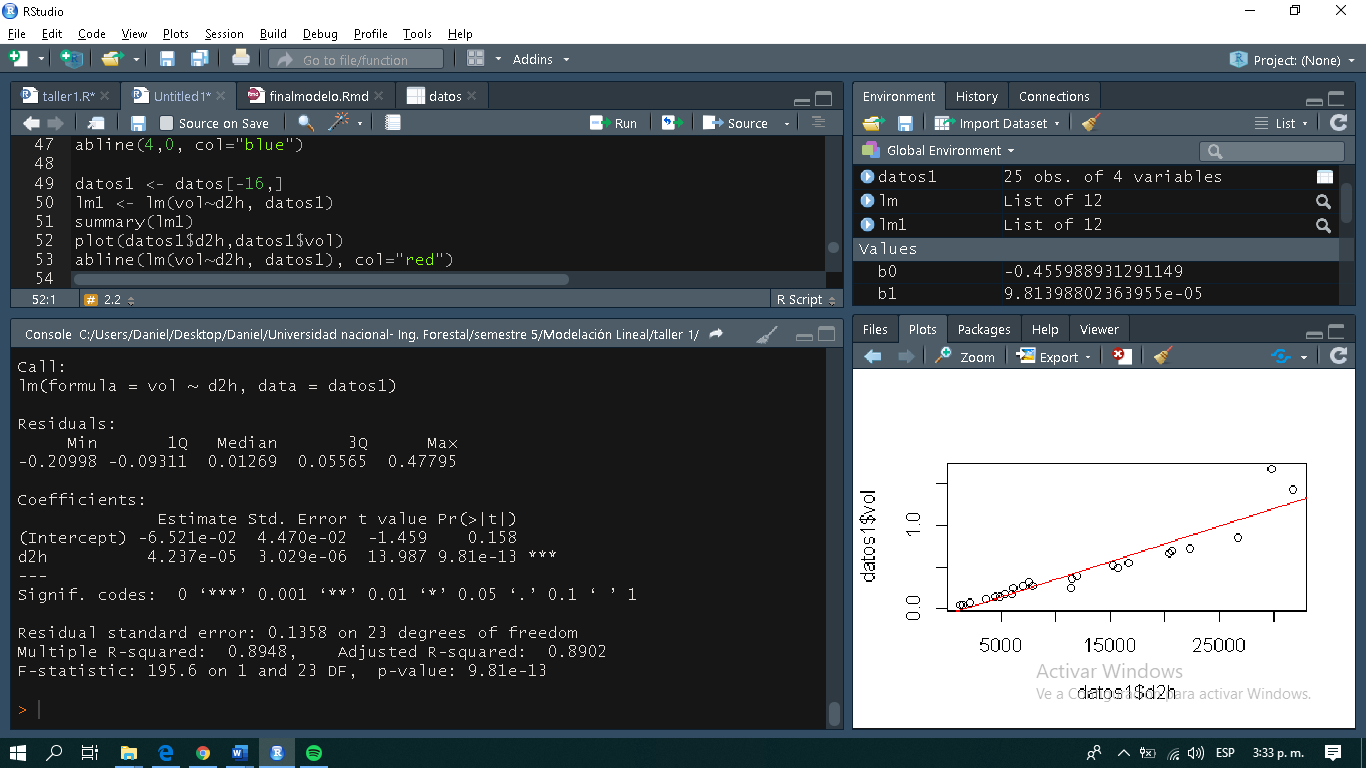
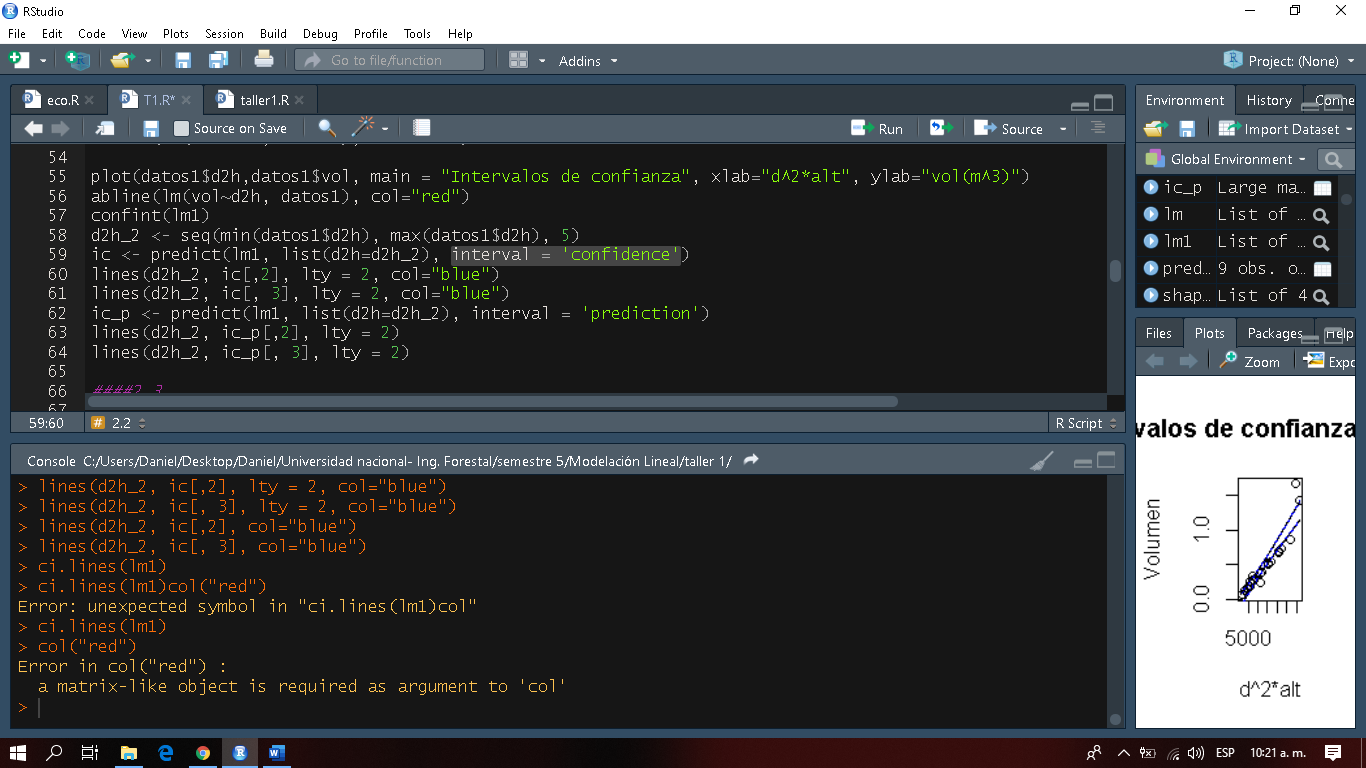


Figura . summary modelo lineal sin outlier.

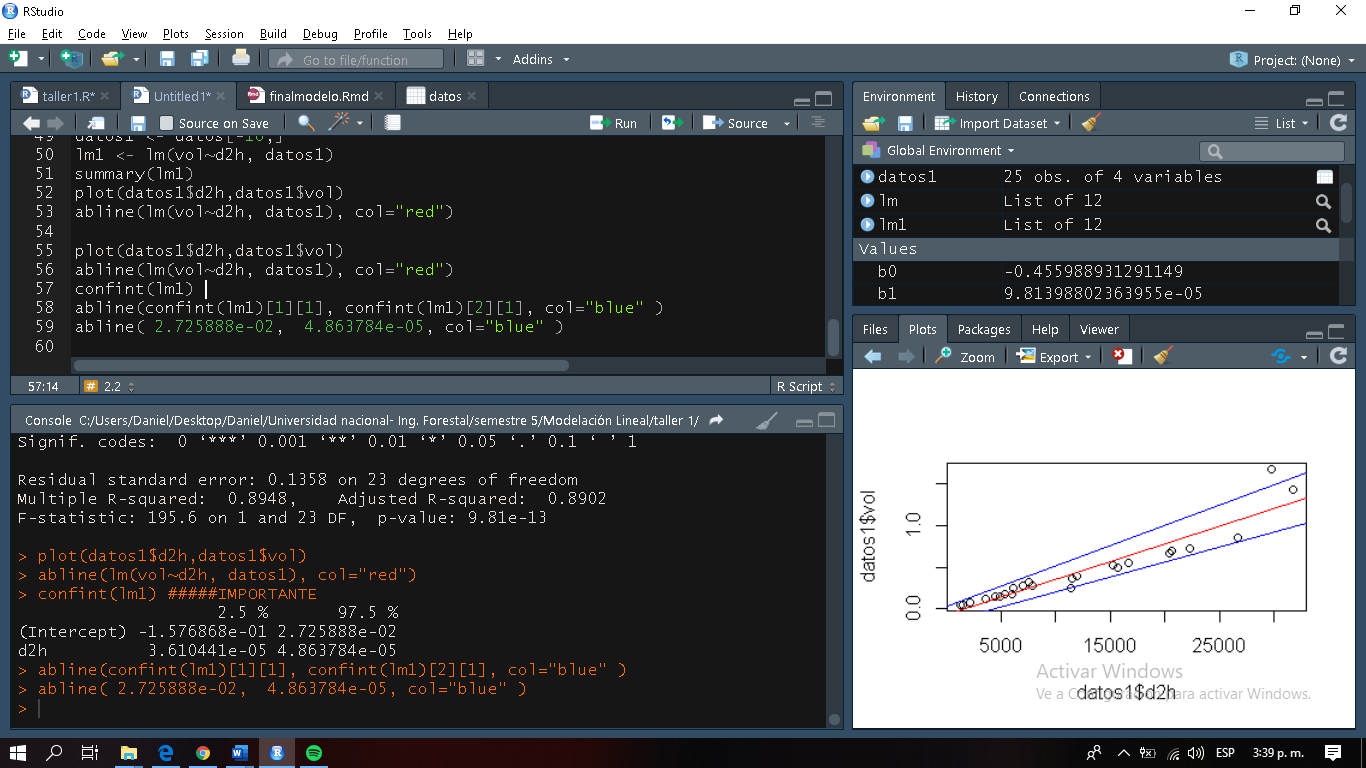
Puede verse que este nuevo modelo tiene una pendiente menor al anterior como consecuencia de la salida del outlier el cual tenia una palanca sobre los datos. Es importante resaltar que al igual que los anteriores modelos, el intercepto también es cercano a 0, y tanto el modelo como la pendiente resultan muy significativos.

Para generar los intervalos de confianza para este modelo (sin outlier), se utiliza el Script 5.



Script 5 Intervalos de confianza.

Cuyo resultado es:



Lo cual básicamente significa que el intercepto pertenece al intervalo de -1.576868e-01 a 2.725888e-02, lo cual incluye al 0 por las razones ya discutidas. En cuanto a la pendiente, esta se encuentra entre 3.610441e-05 y 4.863784e-05. Es importante resaltar que dichos intervalos son para la respuesta media, no para las predicciones. Los intervalos de confianza para la respuesta media se muestran en la Figura 7 de color azul, se incluye también en esta grafica los intervalos de confianza de predicción en negro.

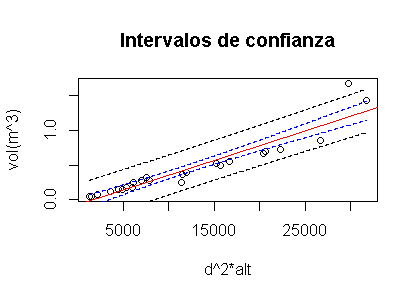


Figura intervalos de confianza.

**2.3 Verifique las propiedades de los residuales usando las formulas dadas para ellos**.

Para que un modelo sea adecuado, los residuales deben cumplir ciertas reglas, siendo:

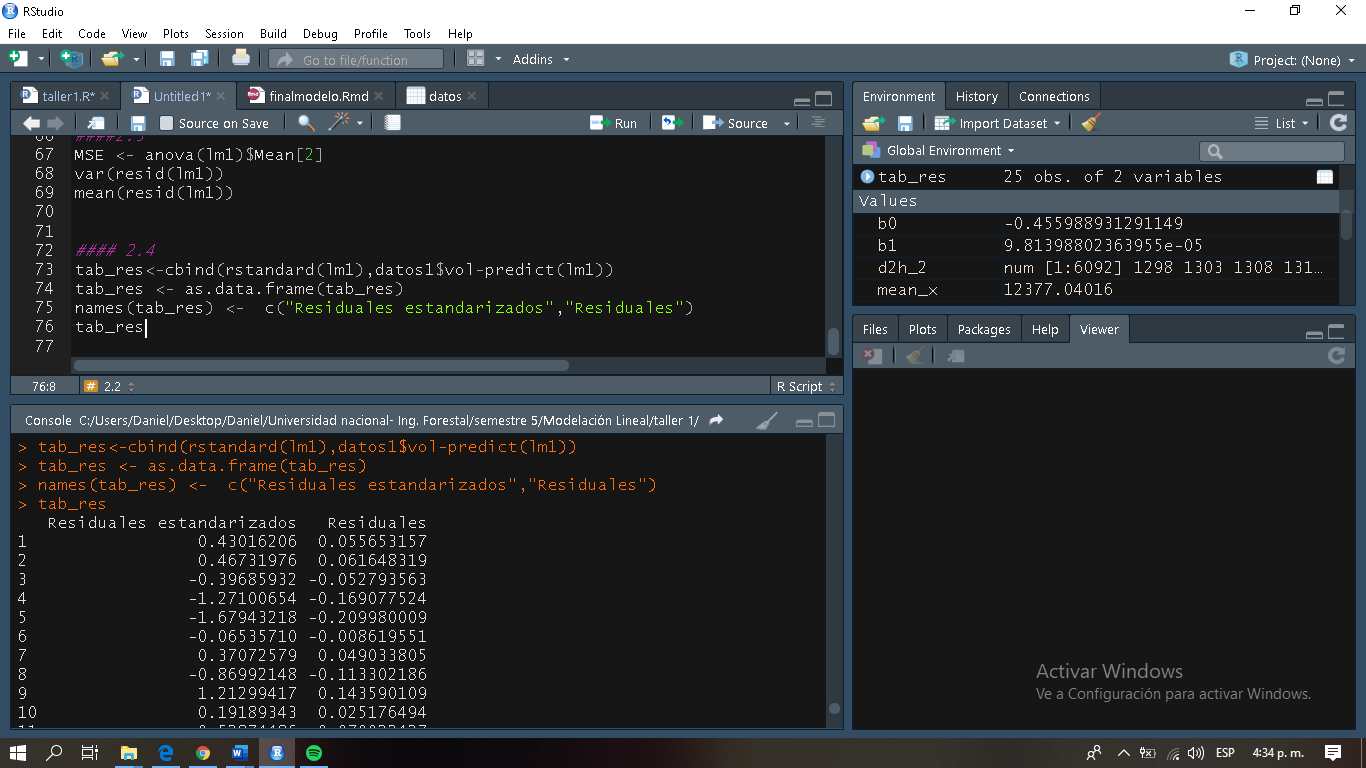
1) La primera propiedad dice que la media de los residuales es igual a cero. En el modelo se calculó la media de los residuales, la cual dio un valor de 1.806281e-18, el cual se aproxima mucho a cero, por tanto, se puede decir que cumple dicha propiedad.

2) La segunda propiedad dice que el modelo será apropiado, si el MSE es un estimador insesgado de la varianza de los términos del error: La varianza obtenida fue de 0.01843478 (), y el MSE del modelo obtenido del anova es 0.0184. Estos dos valores son iguales, por tanto, se cumple dicha propiedad.

**2.4 Construya (y explique) una tabla de residuales con la función**

**tabderes<-cbind(rstandard(modelo),(y-ypredicho))**

Para generar la tabla de residuales se utilizó el siguiente código (Script 6):



Script Tabla de residuales.

Y el resultado esta expresado en la Tabla 3.

|  |  |
| --- | --- |
| Residuales estandarizados | Residuales |
| 0,43016206 | 0,05565316 |
| 0,46731976 | 0,06164832 |
| -0,39685932 | -0,0527935 |
| -1,27100654 | -0,1690775 |
| -1,67943218 | -0,2099800 |
| -0,0653571 | -0,0086195 |
| 0,37072579 | 0,0490338 |
| -0,86992148 | -0,1133021 |
| 1,21299417 | 0,14359011 |
| 0,19189343 | 0,02517649 |
| 0,52874486 | 0,07002243 |
| 3,94297421 | 0,47794901 |
| -0,73156912 | -0,096931 |
| 0,06250385 | 0,0082827 |
| 0,09656354 | 0,01269099 |
| 0,36203926 | 0,0470166 |
| 0,26380823 | 0,03447929 |
| -0,31284034 | -0,0416170 |
| 0,43837042 | 0,05670722 |
| 0,12812414 | 0,01686098 |
| -0,70446893 | -0,0931135 |
| -0,41896919 | -0,0555614 |
| -1,13721375 | -0,1468658 |
| 0,45033443 | 0,05819648 |
| -0,99276893 | -0,1294458 |

Tabla . Residuales estandarizados y residuales.

Básicamente en la Tabla 3 se tienen dos columnas, en la primera pueden observarse los residuales estandarizados, los cuales son residuales que se transforman para hacer más fácil su interpretación.

El calculo de los residuales involucra los valores obtenidos a partir del modelo y aquellos observados (en este caso del set de datos), esto se realiza mediante la siguiente ecuación (Ecuación 6):

Ecuación 6 Residuales

La Ecuación 6 incluye el ei  el cual es el residual de la i-esima observación, hallada como la diferencia entre el valor observado o medido y el valor predicho por el modelo para esa misma observación.

Básicamente, el análisis de los residuales con respecto a los valores observados da una idea de cuan cercano es el modelo a los valores reales, es decir, es un acercamiento para observar que tan variables son los datos.

La primera columna de la tabla muestra los residuales estandarizados, los cuales son una transformación de los residuales a una población equiparable, pero con media 0 y varianza 1, lo cual permite facilitar la interpretación.

Con el fin de observar problemas en el ajuste del modelo se usarán los datos de residuales estandarizados hallados en la tabla para realizar una grafica de residuales vs volumen (Figura 8).

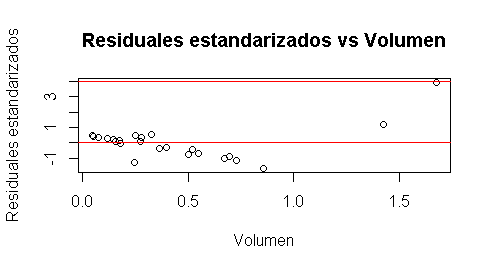
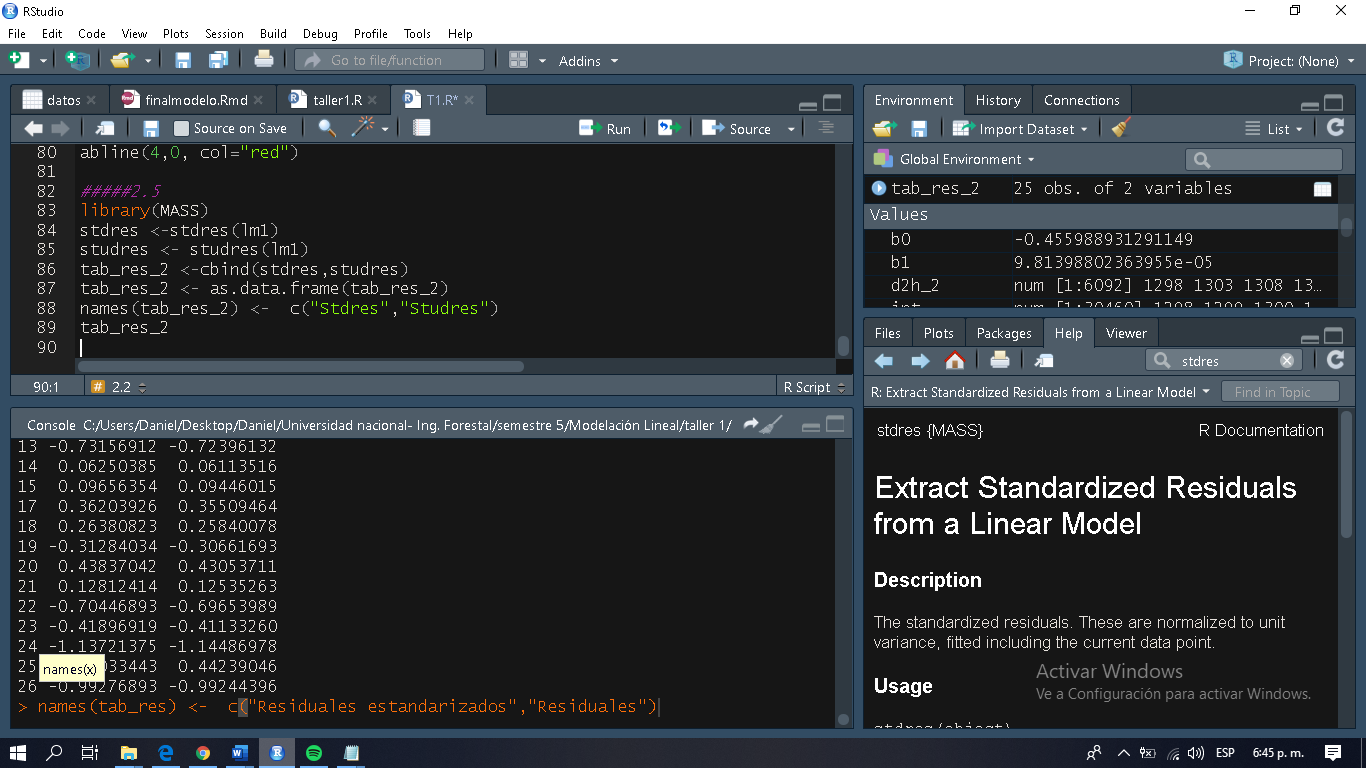


Figura volumen vs residuales estandarizados

En la Figura 8 se puede apreciar que no existe ningún valor por fuera de las 4 desviaciones estándar, lo que no nos lleva a sospechar de un outlier (recordar que se retiró el dato 16 bajo este mismo criterio). En la gráfica se puede notar una tendencia sistemática a la disminución de los residuales estandarizados pasando de datos positivos a negativos, esto puede indicar ciertos problemas de ajuste del modelo con respecto a los datos, abriendo incluso a la posibilidad de que el modelo lineal no sea el más apto para describir o predecir datos en esta población, sin embargo, la tendencia gira alrededor de desviaciones cercanas a 0, por lo tanto puede no ser tan grave como para tener que cambiar de modelo.

**2.5 Use la función de la library(MASS): stdres(modelo) y compárela con studres(modelo)**

Para dar solución a este punto se utiliza el siguiente script (Script 7).



Script 7 Residuales estandarizados y estudentizados.

El resultado de correr el Script 7 esta expresado en la Tabla 4.

|  |  |
| --- | --- |
| Stdres | Studres |
| 0,43016206 | 0,42240941 |
| 0,46731976 | 0,45923318 |
| -0,3968593 | -0,3894718 |
| -1,2710065 | -1,2891669 |
| -1,6794321 | -1,7535515 |
| -0,0653571 | -0,0639264 |
| 0,37072579 | 0,36366515 |
| -0,8699214 | -0,8651519 |
| 1,21299417 | 1,22620111 |
| 0,19189343 | 0,1878259 |
| 0,52874486 | 0,5202945 |
| 3,94297421 | 6,77440329 |
| -0,7315691 | -0,7239613 |
| 0,06250385 | 0,06113516 |
| 0,09656354 | 0,09446015 |
| 0,36203926 | 0,35509464 |
| 0,26380823 | 0,25840078 |
| -0,3128403 | -0,3066169 |
| 0,43837042 | 0,43053711 |
| 0,12812414 | 0,12535263 |
| -0,7044689 | -0,6965398 |
| -0,4189691 | -0,411332 |
| -1,1372137 | -1,1448697 |
| 0,45033443 | 0,44239046 |
| -0,9927689 | -0,9924439 |

Tabla residuales estandarizados y estudentizados

El resultado de los comandos mostrados en el Script 7 son residuales estandarizados (los mismos hallados y usados para las gráficas con anterioridad) y los residuales estudentizados, los cuales no se han mencionado hasta el momento, y que son básicamente aquellos que tratan de mostrar que valores están muy alejados de los valores predichos, un proceso homologo al de los residuales estandarizados, con la diferencia de que estos son estandarizados dividiendo el i-ésimo termino por la desviación estándar de todos los residuos exceptuando al i-ésimo residual. Básicamente la función de los residuales estudentizados es mostrar outliers.

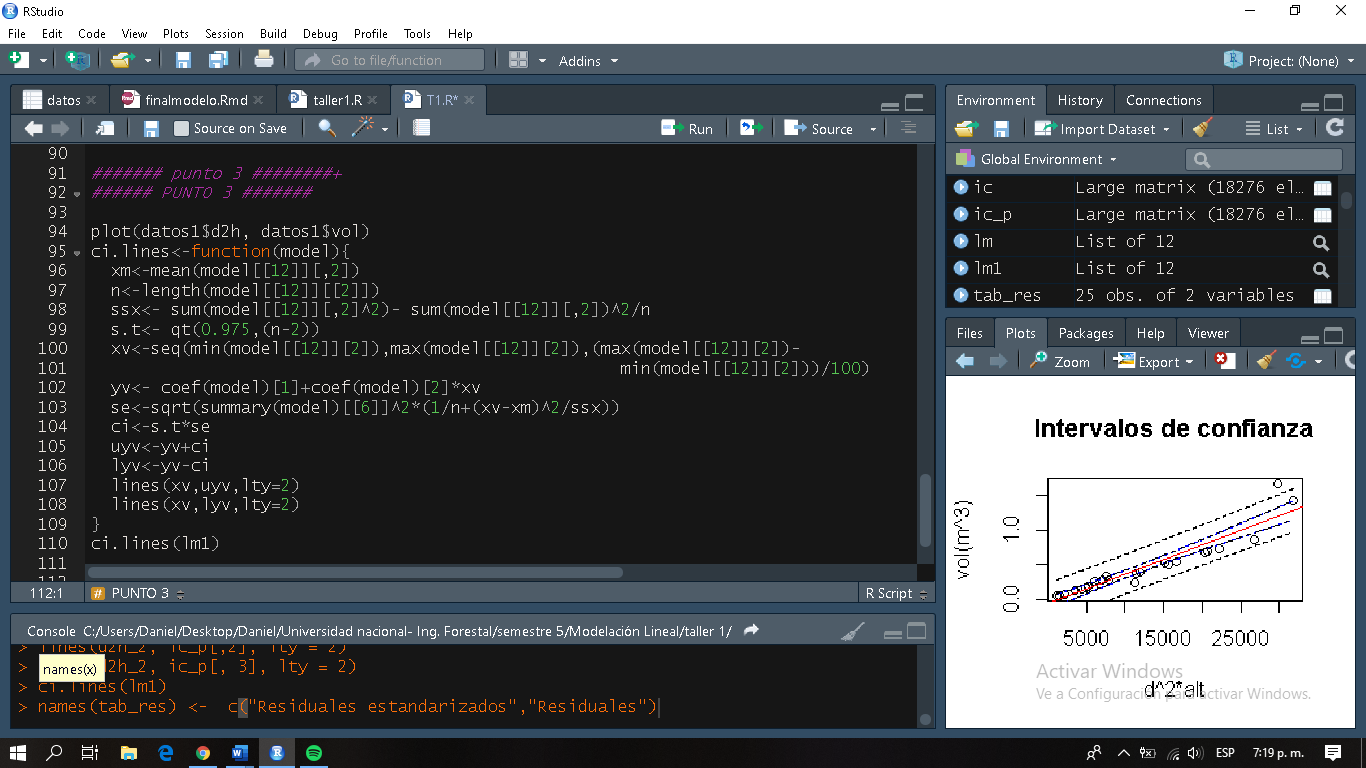
**Punto 3: Grafique y explique**

**3.1 Empíricamente el intervalo de confianza para todas las líneas de regresión**

Para el cálculo del intervalo de confianza y su correspondiente gráfica, se usó el mismo procedimiento mostrado anteriormente (Punto 2) en el Script 5, el cual consiste básicamente en el uso de la función predict incluyendo el argumento interval = 'confidence' , con el cual se crea una matriz de valores en la cual, las columnas 2 y 3 corresponden al valor inferior y superior del intervalo de confianza para cada uno de los datos ingresados en la predicción. La grafica generada a partir de las columnas 2 y 3 sobre el grafico original de dispersión de puntos es como tal el grafico de los intervalos de confianza (Figura 7 ).

**3.2 Lo obtenido si aplica la siguiente función después de la función *plot(x,y)* y relaciónelo con 3.1**

El uso del código mencionado en el taller se muestra a continuación (Script 8):



Script 8 intervalos de confianza

El cual genera como resultado la Figura 9

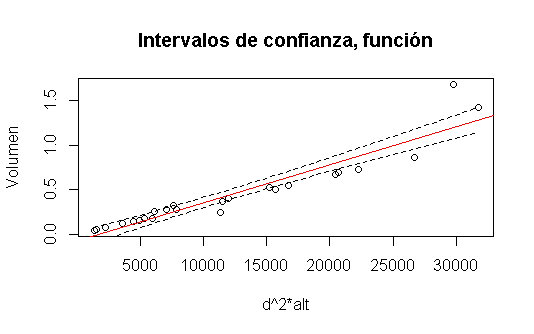


Figura 9 intervalos de confianza generados por función

Al superponer en R esta grafica de intervalos de confianza generado por la función, y la generada con el uso del argumento Inverval = “confidence” de la función de R predict se puede ver que estos están uno sobre el otro perfectamente, corroborando que ambas graficas son correctas e idénticas (Figura 10) . Tales gráficas, corresponden al intervalo de confianza para la media con el 95% de confianza, es decir, existe solo un 5% de probabilidad de que la línea real de la media de la población no se encuentre entre estas líneas.

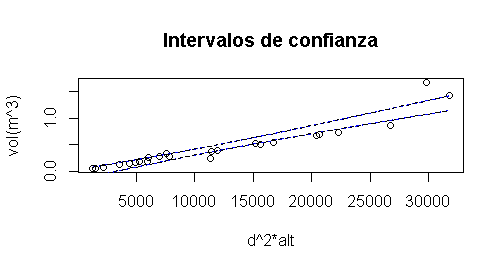


Figura 10. Intervalos de confianza superpuestos

**3.3 Calcule con sus intervalos de confianza  y escoja 5 valores para los cuales desee hacer algunas predicciones simples y de todas en conjunto.**

En las primeras dos lineas del Script 9 se muestra el procedimiento para la estimación de los volumenes esperados para cada valor de la variable independiente (d2h) al 95% de confianza, así como también sus respectivos limites de predicción y límites de confianza para las respuestas medias de la población de datos, las otras lineas restantes del Script 9**,** muestran el procedimiento para calcular el volumen a cinco valores escogidos de manera aleatoria que se encuentran en el dominio de la variable independiente (6000,4282,18756,23841 y 29783), con sus respectivos límites de predicción al 95%, estos valores no coinciden con los valores de “d2h” calculados a partir de los datos de diámetro y altura registrados, ya que, gracias a las características de los modelos de regresión, se pueden hacer predicciones de Y, siempre y cuando sean interpolaciones.

Se puede observar en la Tabla 5**,** que la gran mayoria de los volumenes de los 25 árboles no superan el metro cubico de volumen, tanto en los limites de predicción como en los límites de confianza, sin embargo hay tres árboles que si superan este rango, esto se debe a que estos tienen unas de las áreas basales más grandes y alturas mayores al resto; ya en la Tabla 6**,** se observa que cuatro de estos volumenes no superan el metro cúbico mostrando que hay concordancia de estos valores aleatorios con los demás valores mostrados en la Tabla 5**.**

Imagen que contiene monitor, negro, captura de pantalla, ordenador

Descripción generada automáticamente

Script .Confianza y predicción de los volumenes al 95%

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Árbol | Intervalos de predicción | | | Intervalos de confianza | | |
| Fit | lwr | upr | fit | lwr | upr |
| 1 | -0,00165316 | -0,29516072 | 0,2918544 | -0,00165316 | -0,0868502 | 0,08354388 |
| 2 | 0,19135168 | -0,09727365 | 0,479977 | 0,19135168 | 0,12489546 | 0,25780791 |
| 3 | 0,41979356 | 0,1333557 | 0,7062314 | 0,41979356 | 0,36359309 | 0,47599404 |
| 4 | 0,41607752 | 0,12963583 | 0,7025192 | 0,41607752 | 0,35985752 | 0,47229753 |
| 5 | 1,06698001 | 0,76551845 | 1,3684416 | 1,06698001 | 0,95747697 | 1,17648305 |
| 6 | 0,18761955 | -0,10107421 | 0,4763133 | 0,18761955 | 0,12086676 | 0,25437235 |
| 7 | 0,2319662 | -0,05598183 | 0,5199142 | 0,2319662 | 0,16851612 | 0,29541627 |
| 8 | 0,81030219 | 0,51844009 | 1,1021643 | 0,81030219 | 0,73095908 | 0,8896453 |
| 9 | 1,28040989 | 0,96766433 | 1,5931555 | 1,28040989 | 1,14285397 | 1,41796582 |
| 10 | 0,12382351 | -0,16619999 | 0,413847 | 0,12382351 | 0,05153608 | 0,19611094 |
| 11 | 0,25497757 | -0,03264168 | 0,5425968 | 0,25497757 | 0,19303665 | 0,3169185 |
| 12 | 1,19805099 | 0,88999384 | 1,5061081 | 1,19805099 | 1,07151619 | 1,32458579 |
| 13 | 0,597931 | 0,31044182 | 0,8854202 | 0,597931 | 0,53659689 | 0,65926511 |
| 14 | 0,2677173 | -0,0197371 | 0,5551717 | 0,2677173 | 0,20654642 | 0,32888819 |
| 15 | 0,14330901 | -0,14627638 | 0,4328944 | 0,14330901 | 0,07279984 | 0,21381818 |
| 17 | 0,0279834 | -0,26459884 | 0,3205656 | 0,0279834 | -0,0539691 | 0,10993589 |
| 18 | 0,08452071 | -0,2064713 | 0,3755127 | 0,08452071 | 0,0084405 | 0,16060092 |
| 19 | 0,43961707 | 0,15318183 | 0,7260523 | 0,43961707 | 0,38342998 | 0,49580415 |
| 20 | -0,00270722 | -0,29624884 | 0,2908344 | -0,00270722 | -0,08802153 | 0,0826071 |
| 21 | 0,15913902 | -0,1301111 | 0,4483891 | 0,15913902 | 0,09001971 | 0,22825833 |
| 22 | 0,64211357 | 0,35399112 | 0,930236 | 0,64211357 | 0,57787656 | 0,70635058 |
| 23 | 0,5785614 | 0,29130341 | 0,8658194 | 0,5785614 | 0,51832026 | 0,63880254 |
| 24 | 0,87786584 | 0,58392163 | 1,17181 | 0,87786584 | 0,79117648 | 0,9645552 |
| 25 | -0,01019648 | -0,30398239 | 0,2835894 | -0,01019648 | -0,09634755 | 0,07595459 |
| 26 | 0,80044587 | 0,50886014 | 1,0920316 | 0,80044587 | 0,72212546 | 0,87876628 |

Tabla .  Intervalos de predicción y confianzq para cada uno de los 25 árboles.

|  |  |  |  |
| --- | --- | --- | --- |
| d2h | fit | lower | upper |
| 6000 | 0,1890128 | -0,0996553 | 0,4776809 |
| 4282 | 0,1162192 | -0,1739828 | 0,4064213 |
| 18756 | 0,7294989 | 0,4396938 | 1,019304 |
| 23841 | 0,944956 | 0,6486252 | 1,2412869 |
| 29783 | 1,1967253 | 0,8887402 | 1,5047103 |

Tabla . Volumenes predichos de valores aleatorios

**Punto 4: Encuentre y trate de justificar un buen modelo con las 3 opciones siguientes, además del ya analizado, viendo el comportamiento teórico de ellos (y analizando la aptitud del modelo en toda su extensión.**

Para encontrar el mejor modelo de volumen~d2h, se procede a crear los tres modelos propuestos en el taller (logaritmico, potencial y exponencial), arrojando que todos sus parametros son significativos, sin embargo en ninguno de ellos arroja la normalidad en sus residuales, es por eso que se recurre al criterio del RSE y AIC para escoger el modelo que prediga de la mejor forma los volumenes (ver Tabla 7 ) , es así, que se encuentra que el modelo Potencial (Vol=-0,0053\*(d2h^0,0000411) es el que mejor predice estos volumenes porque tiene el menor AIC(-476,579), además al hacer el gráfico de los residuales estandarizados ( Figura 11), son los que menor rango de desviación estandar tiene (entre -2 y 1), minetras que las otras gráficas tienen un rango en las deviaciones estandar mucho más elevado (entre -2.5 y 1.3 para el logaritmico, y entre -1.5 y 3 para el exponencial), el segundo modelo que nos predice los volumenes es el logaritmico porque a pesar de que tenga el segundo menor AIC (-476,579), tiene el RSE má cercano a cero (1,623E-05). En la Figura 12 se muestran los diferentes modelos que permiten escoger al modelo potencial como el más adecuado para encontrar los volumenes predichos

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Modelo | Fórmula | F(p-value) | DF | R^2 ajust | Shapiro | RSE-ajustado | AIC |
| LINEAL | Vol=--0,0053+0,0000411\*(d2h) | 2,200E-16 | 23 | 1 | 0,0001 | 8,368E-17 | -1776,11 |
| LOGARITMICO | Vol=--0,0053+0,0000411\*ln(d2h) | 1,590E-07 | 23 | 0,8296 | 0,0008 | 1,623E-05 | -476,579 |
| POTENCIAL | Vol=-0,0053\*(d2h^0,0000411) | 1,591E-07 | 23 | 0,8296 | 0,0008 | 8,603E-02 | -738,565 |
| EXPONENCIAL | Vol=0,0053\*e^(0,0000411\*d2h) | 2,200E-16 | 23 | 0,9683 | 0,0746 | 7,043E-04 | -288,050 |

Tabla .Resumen modelos

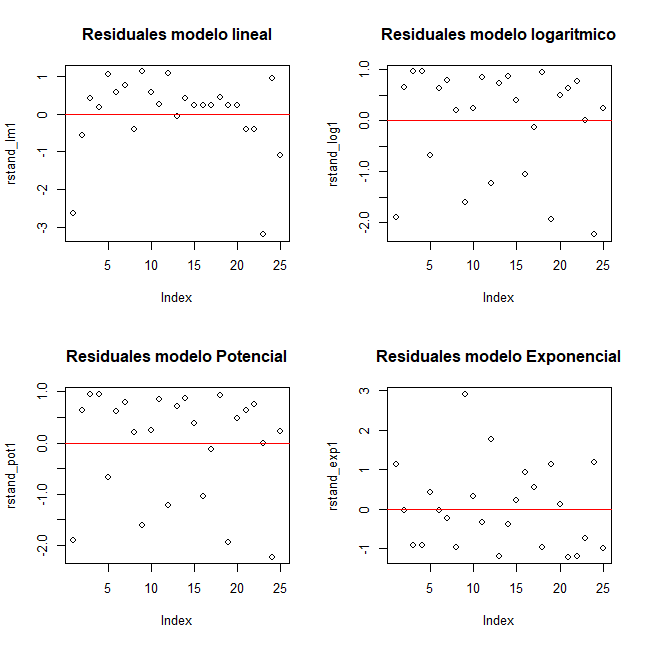


Figura .Distribución de los residuales de cada uno de los modelos

Imagen que contiene mapa, texto

Descripción generada automáticamente

Figura . Distribución de los volumenes en cada uno de los modelos

**Punto 5:** **Verifique para sus datos**

**5.1 Si la distribución de su variable dependiente coincide con la de los errores, excepto por la media**

Teniendo en cuenta que se usó el modelo lineal sin outliers para estimar los volumenes de cada uno de los árboles, se procede a relaizar el Script 10 para saber si los datos y los residuales se comportan de forma normal o no, encontrando que el p-value es menor a 0.05 (ver Tabla 8), indicandonos con una confianza del 95% de que los datos no se distribuyen de forma normal, lo anterior se corrobora al observar las densidades de la distribución de los residuales y los volumenes predichos por el modelo lineal (Figura 13), esta falta de normalidad puede que esté ocurriendo por los pocos datos con los que se están trabajando en este taller.

Imagen que contiene captura de pantalla, monitor, ordenador, portátil

Descripción generada automáticamente

Script . Distribución de volumenes y errores

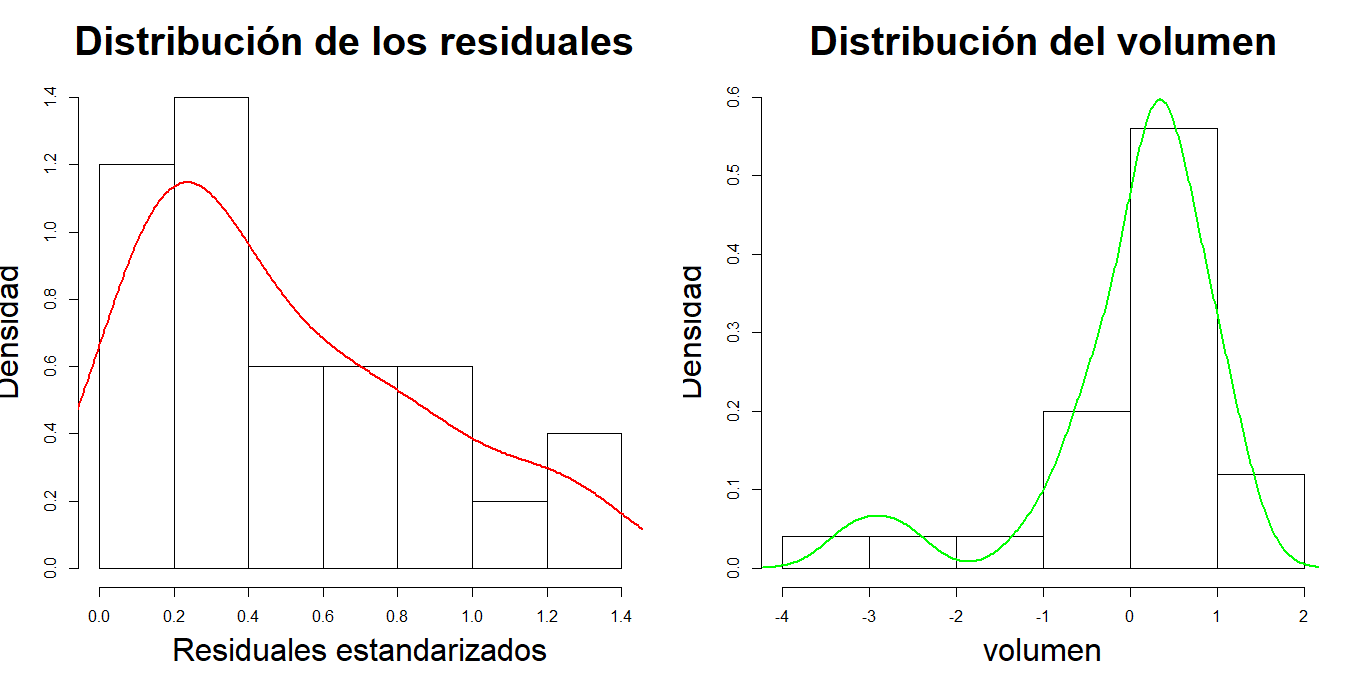


Figura .Densidades de residuales y volumen para el modelo lineal

|  |  |  |
| --- | --- | --- |
| Modelo lineal | Promedio | Shapiro(p-value) |
| Volumenes predichos | 0,4314 | 0,0200 |
| Residuales | 0,0147 | 0,0006 |

Tabla . Shapiro-wilk de los Volumenes predichos y residuales del modelo lineal

**5.2 El valor de las SSTO, SSE, SSR obtenidas con fórmulas y las que le arroja el programa y sus respectivas MS.**

Para hallar los valores de SSTO,SSE y SSE se procede a realizar el Script 11 en donde se calculan cada uno de los valores propuestos de forma manual, posteriormente en ese mismo Script 11 se le realiza el ANOVA al modelo lineal con el fin de que muestr los diferentes valores propuestos, al hacer una resumen( ver Tabla 9), encontramos que los valores calculados a través de los dos métodos no varían significativamente uno respesto al otro, dando a entender que estos valores son los correstos, en este mismo punto se calcula de forma manual y en R los MSE Y MSR del modelo lineal encontrando que no hay diferencias significativas ( ver Tabla 10).

Imagen que contiene monitor, captura de pantalla, ordenador, portátil

Descripción generada automáticamente

Script . Ssto, sse, ssr.

|  |  |  |  |
| --- | --- | --- | --- |
| Fórmula |  |  |  |
| Resultado manual | 3,606384 | 0,4239966 | 4,03038 |
| Valor arrojado en R | 3,606 | 0,424 | 4,03 |

Tabla . SSR, SSE Y SSTO del modelo lineal

|  |  |  |
| --- | --- | --- |
| Fórmula |  |  |
| Resultado manual | MSR=3,606384 | MSE=0,01843463478 |
| Valor arrojado en R | MSR= | MSE=0,018 |

Tabla . MSR Y MSE del modelo lineal

**5.**3 **Realice con ellas el anova respectivo y corrobórelo con R**

En el análisis de varianza, se tiene el siguiente juego de hipótesis:

Ecuación 7.Hipótesis nula

Ecuación 8.Hipótesis alternativa

Para este contraste, la hipótesis nula plantea que la pendiente de la línea de regresión es igual a cero y la alternativa plantea que dicha pendiente tiene un valor distinto a cero, es decir, en este análisis se evalúa si hay o no una regresión lineal entre d2h (tomada como independiente) y el volumen (tomada como dependiente).

El procedimiento para este contraste de hipótesis consta de calcular un F calculado (), y compararlo con un F tabulado , si este F calculado es mayor que el tabulado se concluye que si existe regresión. Al hacer los respectivos calculos manuealmente y en R (ver Script 12) encontramos que:

Al manejar un , para el F tabulado, encontramos que , al comparar ambos F, vemos que F calculado>, por lo que se rechaza la hipótesis nula y se concluye que hay regresión debido a que .

Por otra parte, esto también puede analizarse por medio del valor P, que para el F calculado sería:

P( 9.810436e-13

Como se puede ver, esto coincide con el F calculado y con el valor P del anova del modelo calculado en R (ver Script 12) ,y debido a que el valor P es menor que se concluye otra vez que se presenta regresión en el modelo lineal.

Imagen que contiene monitor, captura de pantalla, portátil, interior

Descripción generada automáticamente

Script 12. F calculado y tabulado.

**5.4** **Corrobore con sus datos las propiedades de la variable k que soporta las normalidades de bo y b1.**

Se procede a realizar el Script 13 en donde esa muestra paso a paso como llegar al valor de la variable K y a sus determinadas propiedades para corroborar las normalidades de bo y b1, encontrando que4,237113 E-05, lo cual es igual al valor de estimado para el coeficiente de regresión (=), por lo que se ve que es una combinación lineal de y como se distribuye normal, por esto, sigue la misma distribución. Aparte de esto, se puede verificar la normalidad de b1 con las tres propiedades de la variable K se cumplen dado que la primera dice que , y al hacer los respectivos calculos encontramos que -6.498331E-21, en la segunda propiedad nos dice que lo cual coincide con la sumatoria hecha en R que da exactamente 1, y la tercera propiedad dice que , comprobandose haciendos los respectivos calculos en R que efectivamente ambas operaciones dan el mismo resultado:

=

Imagen que contiene monitor, ordenador, captura de pantalla, interior

Descripción generada automáticamente

Imagen que contiene monitor, captura de pantalla, ordenador, interior

Descripción generada automáticamente

Script . Propiedades de la variable K

**5.5 Analice, con el uso de sus datos, las concepciones del** .

El coeficiente de determinación nos proporciona la variación de la variable dependiente Y (en este caso volumen), que es explicada por la variable independiente X (en este caso d2h), diciendonos que si la proporción es igual a cero, , significa que la variable independiente tiene nula capacidad predictiva de la variable dependiente, cuanto mayor sea la proporción, mejor será la predicción, si llegara a ser igual a 1, la variable independiente explicaría toda la variación de Y, dando a entender que las predicciones no tienen error.

Para este taller se muestra en la Tabla 11 los diferentes coeficientes de determinación propuestos por Kvålseth en 1985, que se hallaron con las fórmulas propuestas en el Script 14 para cada uno de los ocho coeficientes y cuyos resultados se pueden observar en Tabla 11, encontrando que tienen el mismo coeficiente de determinación(0.8947998) mientras que los otros tres restantes tienen un coeficiente de determinación mcho mayor(0.951744), dandonos a entender que hay una asociación lineal significativa entre el volumen y la variable combinada X.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | 3\_ |  |
|  |  |  |  |
| 0.8947998 | 0.8947998 | 0.8947998 | 0.8947998 |
|  |  |  |  |
|  | Entre |  |  |
| 0.8947998 | 0.9459386 | 0.951744 | 0.9511744 |

Tabla . Resumen de los erres

Imagen que contiene monitor, captura de pantalla, ordenador, portátil

Descripción generada automáticamente

Script . Coeficientes de determinación.