Modelación lineal

Cristian Gañan - Valentina Ruiz - Maria Isabel Vasquez - Tatiana Marin - Daniel Marin - Marlon Tejada

Introducción

La regresión es una herramienta estadística que permite establecer una relación funcional entre una variable dependiente y otra(s) variables explicativas, a través de diferentes modelos que al cumplir una serie de condiciones permiten inferencias válidas con respecto a su comportamiento. Dentro de la regresión encontramos diferentes tipos de esta como: Regresión lineal (Que es la base de este trabajo), Regresión polinomial, Regresión segmentada,Regresión múltiple, Regresión no lineal, y más. A Continuación se encuentran diferentes ejercicios donde se puede observar como con un grupo determinado de datos de altura, diámetro y volumen se realizó una regresión lineal, donde primero se planteó un modelo empírico y luego se realizó un modelación técnica, se comenzó por encontrar los coeficientes de regresión (parámetros ß0 y ß1). haciendo esto se logró observar la tendencia de los datos, y se pudo encontrar la presencia de un dato que se denomina outlier. En el documento se encontrarán diferentes gráficas del modelo, y tablas donde se verá SSR, SSE, $SSTO$, $VALORP$, VARIANZA (σ^2). También una verificación de los datos, y la propuesta de diferentes modelos como el logarítmico, exponencial y potencia.

Modelación empírica

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.

Con el fin de visualizar el comportamiento de los datos, se procedió a realizar un gráfico de dispersión que permitió tomar decisiones con respecto al modelo empírico **figure 1**.

Para modelar las variabes propuestas `diámetro, altura y volumen` se escogió como variable dependiente el diámetro al cuadrado por la altura $(dap)^2 * h$, y como variable independiente el volumen; de esa manera se graficó el diagrama de dispersión con el fin de ver cómo era la distribución de los datos **figure 2**.

En la **figure 3** se puede notar la presencia de un dato *atípico* u *outlier*, así que se decidió por eliminarlo para el modelo empírico, teniendo en cuenta que luego se analizará este dato con todas las herramientas pertinentes. De igual manera, se escogieron arbitrariamente las coordenadas de los puntos que al unirlos describiesen el comportamiento general de los datos. Posterior a la determinación de los dos puntos, y de la linea que los une, se calculó la pendiente con el procedimiento matemático estandar $(m = Y_f - Y_i/X_f - X_i)$, resultando igual a 0.0000411 luego la ecuacion de la recta $Y - _i = m(X - X_i)$ y el intercepto que es igual a -0.0053. En el modelo, este intercepto no es significativo, ya que si en X el valor es cero, en Y también debe serlo, porque sin DAP y sin altura no se tiene volumen. Este valor de -0.0053 es netamente algebraico.

Modelación tecnica

Punto 1

Encuentre sus parámetros de acuerdo con el método de los mínimos cuadrados.

Los parámetros β_0 y β_1 se obtuvieron con el método de los mínimos cuadrados, en el cual para cada par cartesiano X, Y se consideran las desviaciones Qi de Yi observado con respecto a su valor calculado por la recta de regresión.

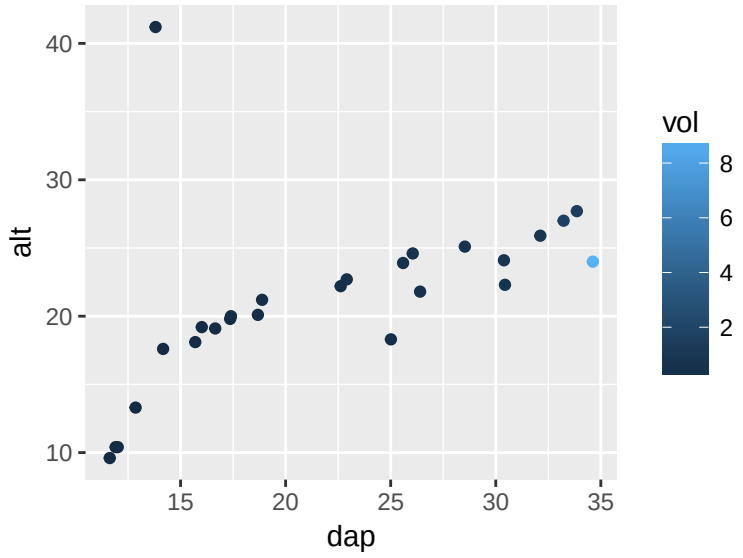

Figure 1: Datos en bruto

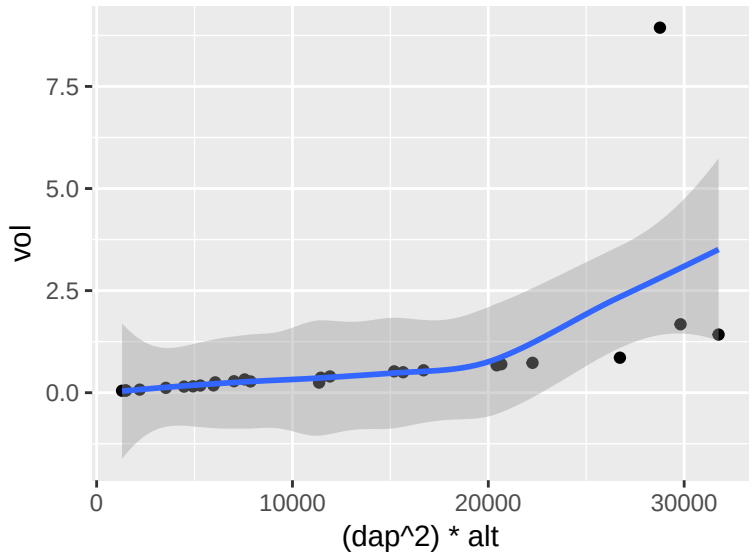

Figure 2: Nuevas variables

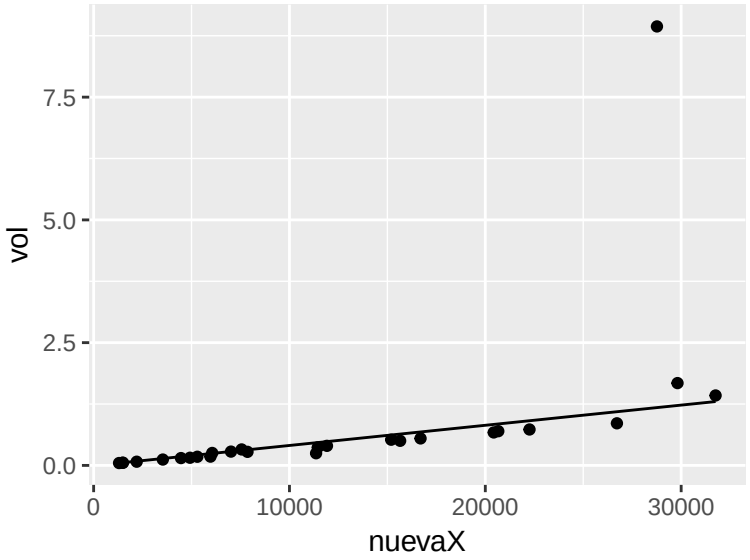

Figure 3: Modelo empirico

Coeficientes	Estimado	Str.error	P-value
Intercepto	$-4.560 * 10^{-01}$	$4.761 * 10^{-01}$	0.34772
Pendiente	$9.814 * 10^{-05}$	$3.065 * 10^{-05}$	0.00383

Table 1: Modelo ajustado con R

$b_i = \frac{\sum(X_i - \widehat{x}) * (Y_i - \widehat{Y})}{\sum(X_i - \widehat{X})^2} = 9.813988 * 10^{-5}$

$b_0 = \frac{(\sum(Y_i) - b_i \sum(X_i))}{n} = -0.4559889$

Se puede observar que la pendiente del modelo es mayor a la del modelo empírico, ya que en dicho modelo se decidió omitir el aparente outlier , y un intercepto que al igual que nuestro modelo empírico dio un valor negativo, aun siendo también un dato algebraico.

Punto 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).

En la **figure 4** se puede observar cómo se ajustó el modelo inicial, por medio de una regresión lineal simple en el cual se tiene como resultado la no significancia del parámetro b_0, y la significancia del parámetro b_1. En este modelo se utilizó un $\alpha = 0.95$.

En la **figure 4** se puede observar que la pendiente es significativa, además, esta aumentó en el modelo lineal. También se puede evidenciar que el intercepto con el eje Y no es significativo ya que no pasa por él, es decir, el intercepto en el eje Y no existe.

El resumen de los parámetros, el valor P y los intervalos de confianza se pueden ver a continuación **table 1** **table 2**.

Estimado	2.5%	97.5%
Intercepto	-1.438577	0.5265993252
Pendiente	$3.487133 * 10^{-05}$	0.0001614084

Table 2: I. confianza modelo con R

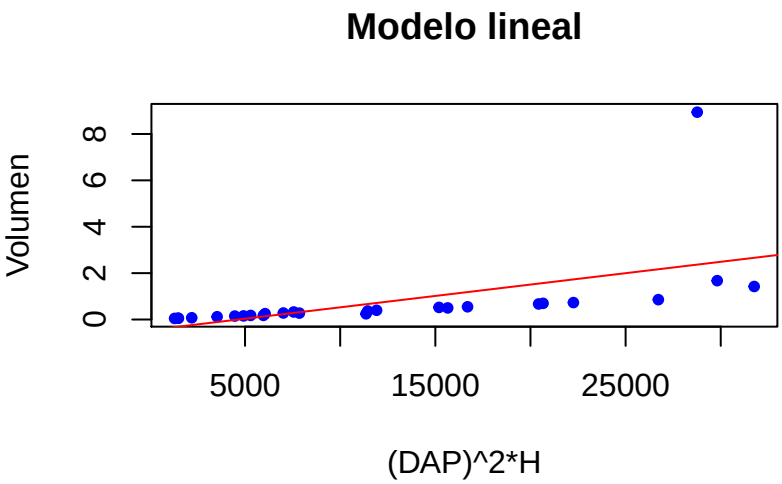

Figure 4: Modelo ajustado

Punto 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 $1.812077 * 10^{-17}$, el cual se aproxima mucho a cero, por tanto se 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 2.150165, y el MSE del modelo obtenido del `anova` es 2.1502. Estos dos valores son iguales, por tanto se cumple dicha propiedad.

Punto 4

Construya (y explique) una tabla de residuales.

Para este punto se consideraron los residuales normales, los estandarizados y los residuales predichos por el modelo. Los residuales se estandarizan para ver de una forma más sencillas cuáles son los los datos que más difieren de la media del modelo. En la **table 3** se puede notar claramente que el punto 16, a modo de ejemplo, tiene un residual mayor a uno, lo cual tiene bastante sentido puesto que este es el residual del outlier. Con respecto a los residuales normales y los predichos por el modelo, se puede decir que los datos no varían significativamente.

R.normales	R.estandarizados	R.precichos
0.36276959	0.25936674	-0.30876959
0.11473233	0.08052810	0.13826767
-0.30038411	-0.20895040	0.66738411
-0.41177703	-0.28644764	0.65877703
-1.30939552	-0.95647116	2.16639552
0.04937668	0.03466541	0.12962332
0.04866111	0.03406596	0.23233889
-0.87487907	-0.61817765	1.57187907
-1.23674120	-0.94452176	2.66074120
0.16714090	0.11793357	-0.01814090
0.03936222	0.02752110	0.28563778
-0.79398175	-0.59482437	2.46998175
-0.57898554	-0.40365372	1.07998554
-0.03914550	-0.02735194	0.31514550
0.12900863	0.09087753	0.02699137
6.57298884	4.87881209	2.36701116
0.31512549	0.22451029	-0.24012549
0.22817392	0.16158675	-0.10917392
-0.31529925	-0.21929304	0.71329925
0.36521099	0.26114624	-0.31121099
0.11234320	0.07903787	0.06365680
-0.63332108	-0.44233533	1.18232108
-0.51212171	-0.35681118	1.03512171
-0.99736982	-0.70956623	1.72836982
0.37655761	0.26951138	-0.32855761
-0.87804990	-0.61986568	1.54904990

Table 3: Residuales

Punto 5

***Use la función de la `library(MASS)`: `stdres(modelo)` y compórela con 'studres(modelo)*.**

La **table 4** muestra los residuales estudentizados vs los estandarizados.
Los residuales se estudentizan con el fin de buscar cuál es el residual que ocasiona un mayor efecto sobre el modelo, de igual forma, se estandarizan para acotarlos en intervalos entre cero y uno. En los datos, el residual que mayor importancia tiene en el modelo es el dato 16, como se puede ver en la tabla, puesto que el residual estudentizado es significativamente mayor a los demás.

Graficar y explicar:

Punto 1

Empíricamente el intervalo de confianza para todas las líneas de regresión.

En la **figure 5** se puede apreciar los intervalos de confianza empíricos, se trazaron sin tener muy en cuenta el outlier pues este no es de interés para este caso, lo que se quiso hacer fue hacer la tendencia con el grupo de datos que seguían el patrón lineal.

Punto 2

Lo obtenido si aplica la siguiente `ci.lines` después de la función `plot(x,y)` y relaciónelo con punto 1.

Los intervalos de confianza para el modelo se graficaron (**figure 6**) con una confianza del 95%. Como se puede observar en la gráfica, no todos los datos están contenidos en él, puesto que la palanca generada por el outlier es bastante fuerte. Tanto la línea de regresión como los intervalos están desplazados hacia arriba, lo que significa que se están sobreestimando los intervalos de confianza.

Estandarizados	Estudentizados
0.25936674	0.25426237
0.08052810	0.07884324
-0.20895040	-0.20473726
-0.28644764	-0.28089707
-0.95647116	-0.95470528
0.03466541	0.03393638
0.03406596	0.03334951
-0.61817765	-0.61003813
-0.94452176	-0.94231444
0.11793357	0.11548395
0.02752110	0.02694207
-0.59482437	-0.58664063
-0.40365372	-0.39650301
-0.02735194	-0.02677646
0.09087753	0.08897942
4.87881209	52.69054406
0.22451029	0.22001438
0.16158675	0.15827066
-0.21929304	-0.21489123
0.26114624	0.25601180
0.07903787	0.07738380
-0.44233533	-0.43479794
-0.35681118	-0.35022870
-0.70956623	-0.70202912
0.26951138	0.26423698
-0.61986568	-0.61173099

Table 4: Residuales con ‘MASS‘

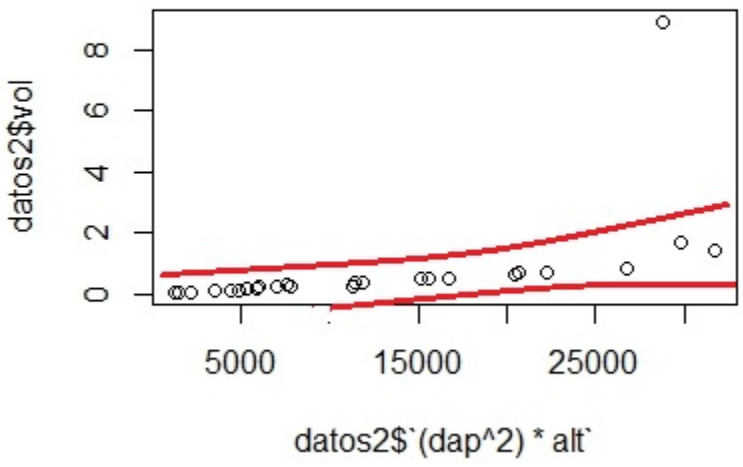

Figure 5: Intervalos empíricos

Resultado	lwr	upr
2.53123848	-0.7576955	5.820172
0.37912904	-2.7145991	3.472857
0.07744705	-3.0377030	3.192597
0.99721401	-2.0906530	4.085081
-0.31334595	-3.4738716	2.847180

Table 5: Intervalos de confianza

Resultado	lwr	upr
2.53123848	1.2435743	3.8189026
0.37912904	-0.2628566	1.0211147
0.07744705	-0.6608967	0.8157908
0.99721401	0.3840953	1.6103328
-0.31334595	-1.2243424	0.5976505

Table 6: Intervalos de predicción

Punto 3

Calcule con sus intervalos de confianza $E(y_h)$ $\widehat{y}_h$ $\widehat{y}h_n$ ***y escoja 5 valores para los cuales desee hacer alguna predicciones simples y de todas en conjunto. Explique cada uno de ellos.***

Se puede observar que la diferencia para los intervalos de confianza (**table 5**) entre el límite superior y el límite inferior no varían de manera significativa entre los cinco valores muestreados, dado que la diferencia neta se encuentra aproximadamente entre 1.82 y 2.58, sin embargo para los intervalos de confianza de las predicciones simples (**table 6**), se encontró que existe una mayor diferencia entre el límite superior e inferior, dado que se encuentran diferencias aproximadas de hasta 6.58 y es lógico pues con las predicciones el intervalo se agranda, es una distribución de un solo de dato. Caso contrario ocurre con los intervalos de confianza donde son distribuciones de datos en conjunto; es como querer hacer un modelo con una buena densidad de datos y con pocos, el modelo será más confiable cuando hay más información acerca del fenómeno que cuando no, pues se tendrá más parte de la población para analizar.

Transformaciónes

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):

Logaritmo: $Y = A + Bln(x)$ **Potencia:** $Y = Ax^B$ ó $ln(y) = ln(A) + Bln(x)$ **Exponencial:** $Y = Ae^B(Bx)$ ó $ln(y) = ln(A) + Bx$

En los tres modelos propuestos se encontró que cada uno de dichos modelos arroja valores estimados distintos al modelo lineal, dado que estos tienen las entradas de sus parámetros diferentes para calcular determinada variable dependiente, sin embargo se visualizó que el modelo que mejor se acopla para la estimación del volumen de los árboles es el modelo exponencial (**figure 9** y **table 9**), ya que nos muestra un intercepto más cercano a cero (-2.5208772), en cambio, el modelo logarítmico y potencial (**table 7**, **figure 7** y **table 8**, **figure 8** respectivamente) nos muestran interceptos de -6.1664 y -11.3084 respectivamente, además este modelo tiene una pendiente más baja (0.0001126), acercándose a la estimada por el modelo lineal, por el contrario los modelos logarítmico y potencial tienen pendientes de 1.7615 y 2.5899 respectivamente, también se tiene que el modelo exponencial tiene un $p - value$ del orden de 10^{-14}, a diferencia de los modelos logarítmico y potencial del orden de 10^{-2} y de 10^{-12} respectivamente.

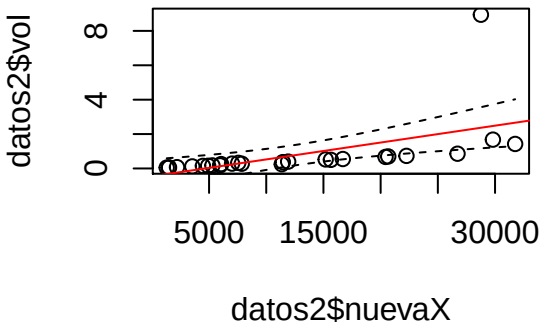

Figure 6: Intervalos de confianza

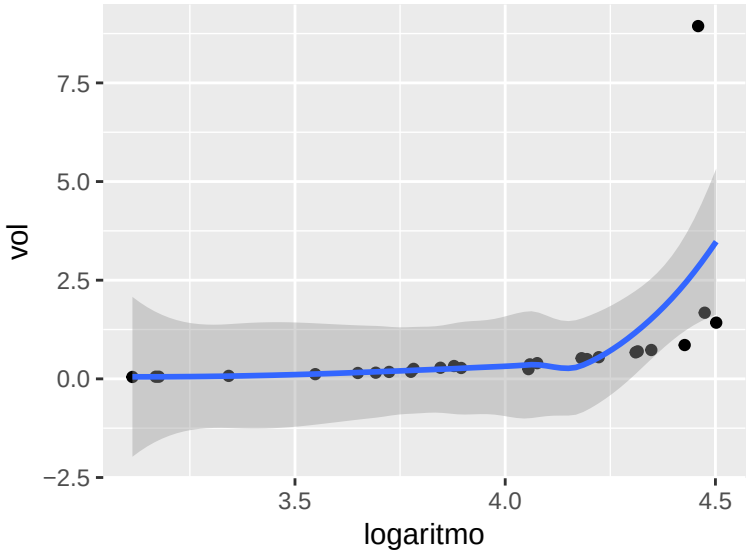

Figure 7: Transformación log

Coeficientes	Estimado	Str.error	P-value
Intercepto	-6.1664	2.9671	0.0486
Pendiente	1.7615	0.7506	0.0275

Table 7: Modelo ajustado con logaritmo

Coeficientes	Estimado	Str.error	P-value
Intercepto	-11.3084	0.8506	$1.46 * 10^{-12}$
Pendiente	2.5899	0.2152	$1.18 * 10^{-11}$

Table 8: Modelo ajustado con potencia

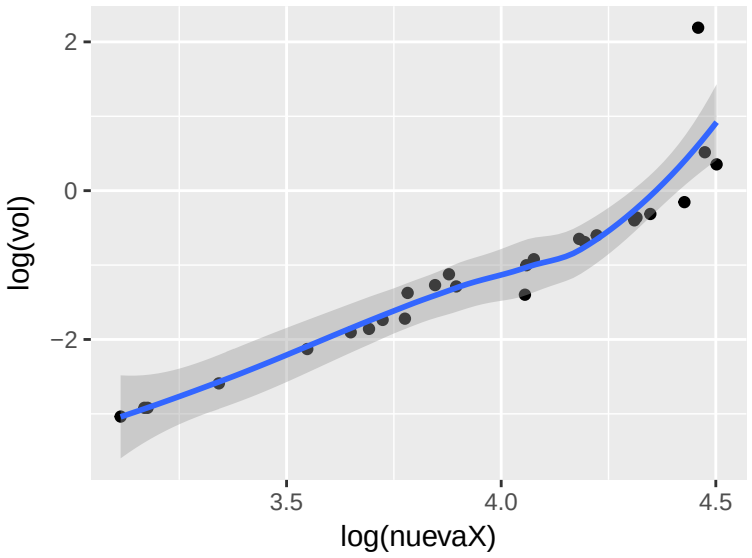

Figure 8: Transformación potencia

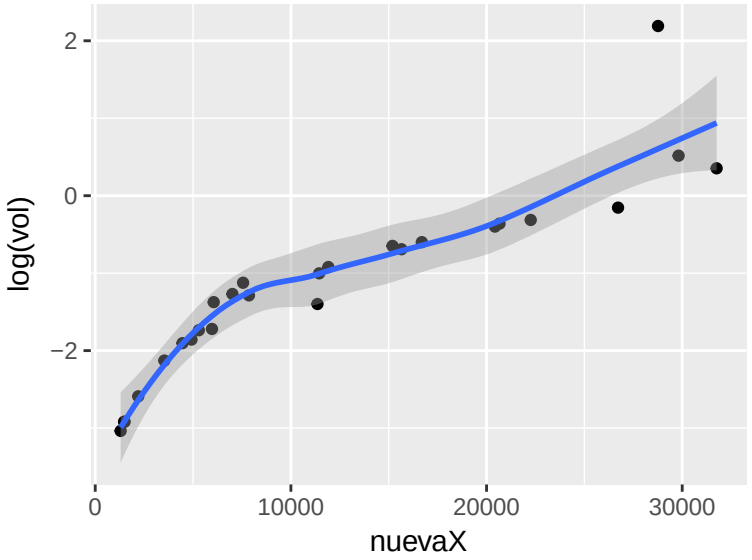

Figure 9: Transformación EXP

Coeficientes	Estimado	Str.error	P-value
Intercepto	-2.5208772	0.1568867	$2.41 * 10^{-14}$
Pendiente	0.0001126	0.0000101	$5.63 * 10^{-11}$

Table 9: Modelo ajustado con exponencial

Verificar

Punto 1

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

```
##
##  Shapiro-Wilk normality test
##
## data:  datos2$vol
## W = 0.37596, p-value = 1.759e-09
```

```
##
##  One Sample t-test
##
## data:  datos2$vol
## t = 2.254, df = 25, p-value = 0.03321
## alternative hypothesis: true mean is not equal to 1.812077e-17
## 95 percent confidence interval:
##  0.06546663 1.45191799
## sample estimates:
## mean of x
## 0.7586923
```

Al hacer una prueba de `Shapiro.test` a los residuales y a la variable independiente, se puede notar que en ambas distribuciones no hay normalidad, ahora tampoco se puede decir que no se distribuyan de forma igual solo se puede decir con un nivel $\alpha = 0.05$ que no se distribuyen normalmente.

Las medias de las dos distribuciones al hacer una prueba de hipótesis de donde $h_o : \widehat{Y}_{vol} = \widehat{Y}_{res}$ y $h_1 : \widehat{Y}_{vol} \neq \widehat{Y}_{res}$, se puede asegurar con un $\alpha = 0.05$ que no son iguales las medias.

La no normalidad de los datos no es preciso asegurar porque, tal vez una hipótesis que puede ser cierta es que además del dato atípico el numero de datos no sean suficientes para captar la distribución de los datos, quizás también, la trasformación de datos ayude al comportamiento normal, hay que recordar que esta prueba se hizo para datos en bruto es decir sin transformar.

Punto 2

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

```
##           SSE      SSR     SSTO
## [1,] 57.01006 22.03763 73.64153
```

En la **table 10** se observa el `anova` del modelo, al hacer las SS.. de este con las formulas se pudo notar algo: teniendo en cuenta que el b_0 no fue significativo no habría razón para ponerlo en el modelo, sin embargo, parece ser que el `R` ajusta el modelo con intercepto, porque si se hace con la fórmula y sin b_0 la SSE da distinto a arrojado en `R` ahora, dado esto, se procede a dejar el anova sin intercepto pues para este caso es lógico un intercepto de cero, las implicaciones de hacer esto se esperan que no sea significativas pues teniendo en cuenta que hay un intercepto negativo, no se puede asumir una X negativa pues no es lógico en crecimiento de árboles.

Punto 3

Realice con ellas el `anova` ***respectivo y corrobórelo con*** `R`***.***

Coeficiente	Df	SS..	MS..	p-valvue
Pendiente	1	22.038	22.0376	0.003827
Residuales	24	51.604	2.1502	

Table 10: Anova

Variación	Df	SS..	MS..	p-valvue
Regresion	1	22.03763	22.03763	0.003827
Residuales	24	57.01006	2.375419	

Table 11: Anova con formula

Como se puede apreciar en las **table 10** y **table 11** se encuentra en anova hecho con `R` y con las formulas respectivamente, la diferencia nos es mucha para los parámetros SSE con `R` y SSE con fórmulas, la desigualdad radica en que con el anova del `R` se hizo con b_0 y el otro omite este valor.

Punto 4

Corrobore con sus datos las propiedades de la variable k que soporta las normalidades de b_0 ***y*** b_1***.***

1) $\sum_{i=1}^{n} k_i = 0$

```
## [1] -0.004557452
```

2) $\sum_{i=1}^{n}(k_i - x_i) = 1$

```
## [1] -30.40777
```

3) $\sum_{i=1}^{n} k_i^2 = \frac{1}{\sum(X_i - \widehat{X})^2}$

$\frac{1}{\sum_{i=1}^{n}(X_i - \widehat{X})^2}$

```
## [1] 4.370461e-10
```

$\sum_{i=1}^{n} k_i^2$

```
## [1] 7.399508e-06
```

Para la propiedad uno se puede decir que a pesar que no es cero, si se acerca mucho a este valor, era de esperase que esto fuera así pues estos supuestos se basan en la normalidad de los datos hay que recordar que los datos según la prueba de `Shapiro.test` rechazaba el supuesto de normalidad. sin embargo, se puede que siendo no muy estrictos esta propiedad se cumpliría.

Para la propiedad dos, el resultado es un término negativo que no se acerca a uno, talvez por la misma razón que los datos no se comportan normal, por ello es un resultado lógico.

En la propiedad tres pasa lo mismo que en la dos, y esto es quizás a falta de normalidad de los datos en estudio.

Punto 5

Analice, con el uso de sus datos, las concepciones del R^2.

Para R_1^2:

```
## [1] -0.2029911
```

Para R_2^2

```
## [1] 0.2031874
```

Para R_3^2

```
## [1] 9.63146e-09
```

Para R_4^2

```
## [1] 0.2992547
```

Para R_7^2

```
## [1] -103806044
```

Para R_8^2

```
## [1] 9.63146e-09
```

Para empezar, hay que decir que el valor del R^2 para la tabla de Snedecor y Cochran, 1970; Steel y Torrie, 1985, $24gl$,`variables independientes` 1 un nivel $\alpha = 0.05$ arrojó un resultado de 0.388,es decir, las variables tendrían un grado de asociación bajo.

Ahora para las distintas concepciones del R^2 hay que decir varias cosas: Según teóricamente los R_1^2 al R_6^2 tendrían que ser iguales para modelos lineales simples, en este caso el supuesto no se cumple pues los distintos R^2 dan un resultado diferente.

Los R_7^2 y R_8^2 son recomendados para modelos sin intercepto (como lo es este caso), además, es mayor a uno en modelos no lineales, este supuesto se cumple pues los resultados lo afirman es decir, podría suponerse la linealidad del modelo, sin embargo, no es suficiente solo con esto afirmarlo.

Para el R_1^2 según su resultado negativo se tendría una regresión inapropiada, sin embargo, este supuesto también apoya la idea, de la npresencia notoria de observaciones remotas y es lógico, hay que recordar que en los datos hay un dato que no se comporta de forma similar a la nube de puntos en conjunto.

En resumen, por los valores dados por los R^2 en conjunto, la relación de asociación es baja, pero la hay, esto lo confirma una prueba del `coeficiente de Pearson` donde se tiene que para esta prueba hay un `p-value` $< \alpha = 0.05$ es decir se tendría que rechazar hipótesis nula que apoya el supuesto de la no correlación lineal de las variables, entonces se puede asegurar con un $\alpha = 0.05$ que hay correlación entre las variables.