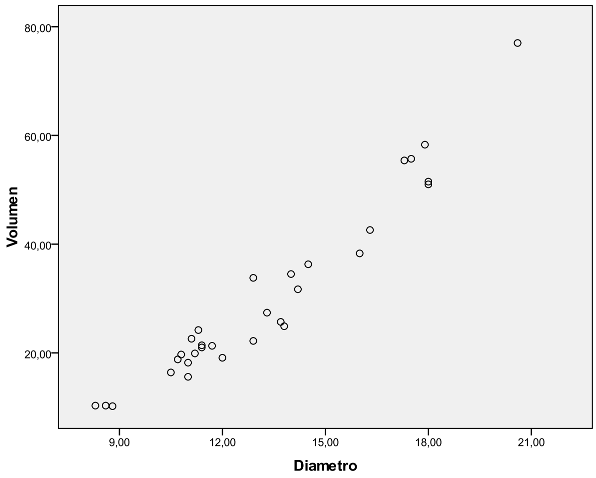
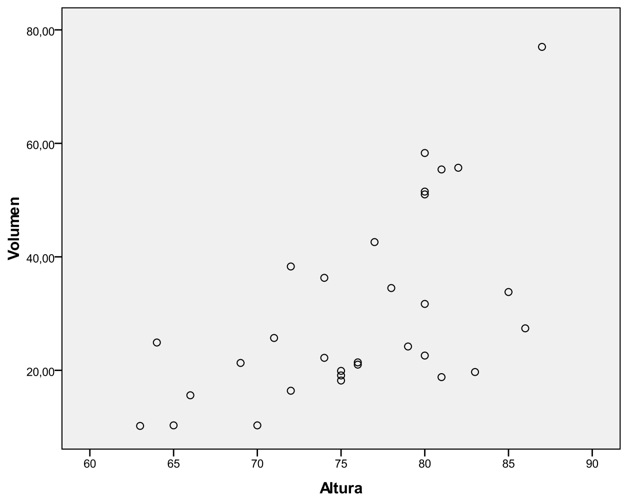
Resultados de la Práctica 3

Análisis descriptivo:





A simple vista, parece que la variable diámetro presenta una mejor relación lineal para explicar el volumen de un árbol (aunque la altura también podría contribuir.

Se decide hacer en primer lugar un análisis de regresión lineal simple para explicar el volumen por el diámetro.

Estadísticos descriptivos

	Media	Desviación típica	N
Volumen	30,1710	16,43785	31
Diametro	13,2484	3,13814	31

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,967[a]	,935	,933	4,25199

a. Variables predictoras: (Constante), Diametro

b. Variable dependiente: Volumen

El coeficiente de correlación es 0,967. Se estima que el 93,5 % de la variabilidad en el volumen se explica (en una relación lineal) por la variabilidad de los diámetros.

Modelo: regresión lineal simple

Y_x = volumen de un cerezo de diámetro x = **$\beta_0 + \beta_1 x + U_x$**
U_x = variaciones aleatorias (residuos) son variables aleatorias $N(0,\sigma)$ para todo x

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	7581,781	1	7581,781	419,360	,000[a]
	Residual	524,303	29	18,079		
	Total	8106,084	30			

a. Variables predictoras: (Constante), Diametro

b. Variable dependiente: Volumen

La tabla ANOVA, con un p-valor 0,000… nos indica que se rechaza H_0: el modelo no es explicativo. Hay evidencia estadística significativa a favor del modelo planteado.
La estimación de σ^2 (la varianza residual) es S_R^2 = 18 ,079

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-36,943	3,365		-10,978	,000
	Diametro	5,066	,247	,967	20,478	,000

a. Variable dependiente: Volumen

Con los valores estimados de β_0 y β_1 y sus correspondientes errores típicos podríamos calcular intervalos de confianza (por ejemplo con confianza 0,95) para ambos parámetros según las fórmulas:

$$-36,943 +- t_{29,\,0,025}\ 3,365 \qquad 5.066 +- t_{29,\,0,025}\ 0,247$$

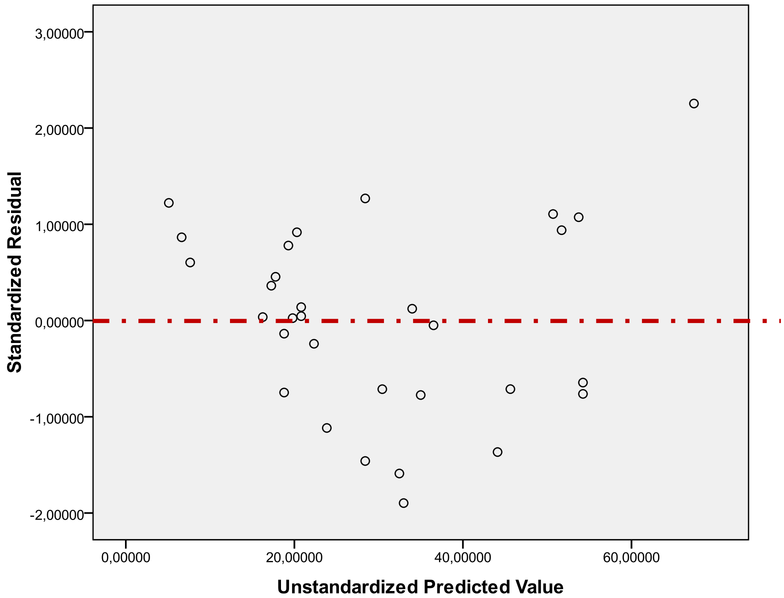
El valor de la t para el diámetro (o su p-valor = 0,000…) nos sirve para rechazar H_0: β_1 =0 (este contraste es equivalente al realizado antes con la tabla ANOVA.)

La recta de regresión estimada es: y_x = -36,943 + 5,066 x
Esta recta nos sirve para poder hacer predicciones puntuales y por intervalo del volumen de un cerezo dado su diámetro. Por ejemplo, para un cerezo con diámetro = 14 pulgadas tendríamos un volumen estimado de 33,98 pies cúbicos.
El intervalo de confianza para el volumen medio de los cerezos con 14 pulgadas de diámetro se obtendría con la fórmula:

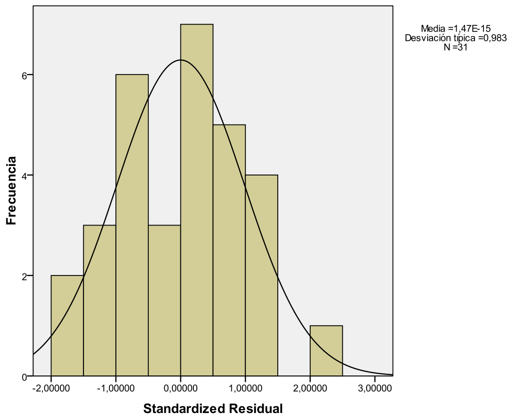
$$IC_{1-\alpha}(\text{estimación}) = \left(\hat{y}_0 \pm t_{n-2;\alpha/2}\, S_R \sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{n v_x}}\right)$$

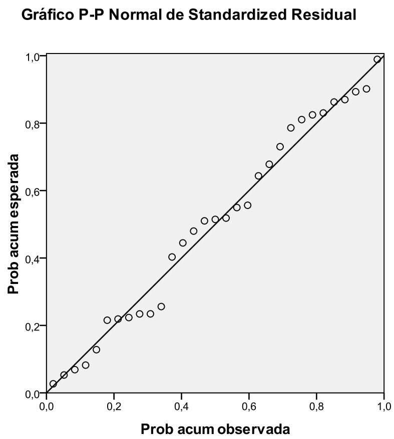
Y el intervalo para la predicción del volumen de un cerezo con 14 pulgadas de diámetro lo mismo pero sumando un 1 bajo la raíz cuadrada.

Análisis de los requisitos previos del modelo



El gráfico de dispersión de los residuos muestra una cierta curvatura (residuos por encima de cero en los extremos y por debajo en el centro) lo que puede hacer sospechar falta de linealidad. Se podría pensar en un modelo potencial. No parece que haya valores anómalos ni indicios claros de falta de igualdad de varianzas.





Los gráficos para la Normalidad también muestran algunas discrepancias posiblemente por la misma razón citada.

Realizaremos un segundo análisis con un modelo potencial (ajuste de la curva $y = kx^{\beta}$)

Es decir, ajustaremos una recta a Log (y) y Log (x) (transformación doble Log)

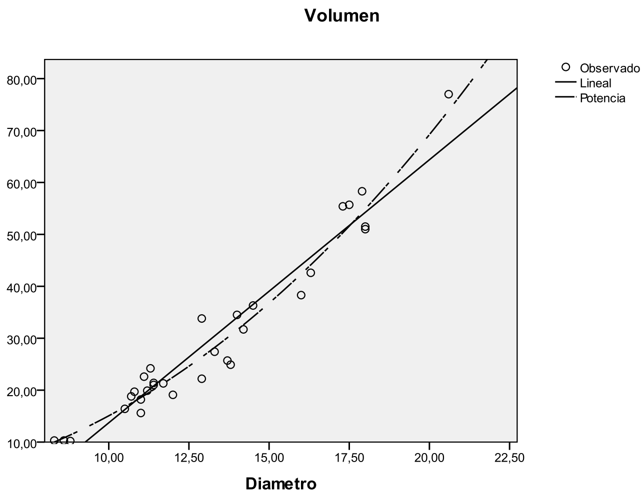
Resumen del modelo y estimaciones de los parámetros

Variable dependiente: Volumen

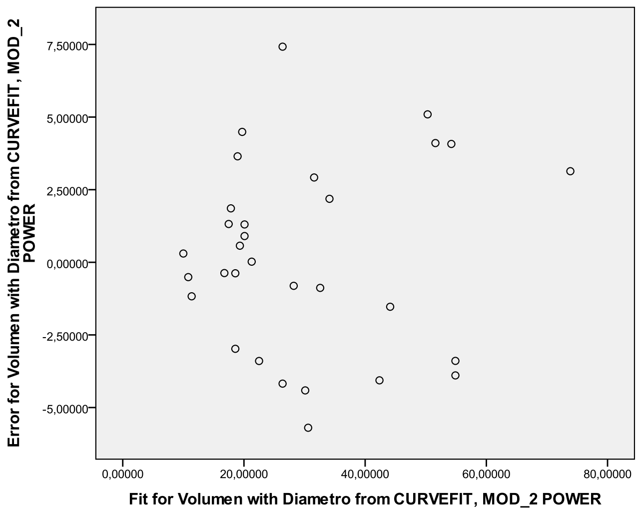
Ecuación	Resumen del modelo					Estimaciones de los parámetros	
	R cuadrado	F	gl1	gl2	Sig.	Constante	b1
Lineal	,935	419,360	1	29	,000	-36,943	5,066
Potencia	,954	599,717	1	29	,000	,095	2,200

La variable independiente es Diámetro.

Parece que el ajuste del modelo potencial es ligeramente mejor.



Veamos el aspecto de los residuos con este nuevo modelo:



Ahora los residuos (brutos) no presentan la curvatura que presentaban los anteriores.

Finalmente haremos un análisis de regresión lineal múltiple con dos variables explicativas (diámetro y altura)

Modelo: regresión lineal múltiple

$\mathbf{Y_{x1,x2}}$ = volumen de un cerezo de diámetro x_1 y altura x_2 = $\boldsymbol{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + U_{x1,x2}}$
$\mathbf{U_{x1,x2}}$ = variaciones aleatorias (residuos) son variables aleatorias $N(0,\sigma)$ para todos los valores de x_1 y x_2

Estadísticos descriptivos

	Media	Desviación típica	N
Volumen	30,1710	16,43785	31
Diámetro	13,2484	3,13814	31
Altura	76,00	6,372	31

Correlaciones

		Volumen	Diametro	Altura
Correlación de Pearson	Volumen	1,000	,967	,598
	Diámetro	,967	1,000	,519
	Altura	,598	,519	1,000

La correlación entre la altura y el volumen es 0,598 bastante menor que la obtenida para el diámetro. Entre las variables explicativas (altura y diámetro) hay una correlación de 0,519

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,974[a]	,948	,944	3,88183

a. Variables predictoras: (Constante), Altura, Diámetro

b. Variable dependiente: Volumen

La variabilidad del volumen se explica en un 94,8% por el modelo. Sólo con el diámetro se explicaba un 93,5%.

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	7684,163	2	3842,081	254,972	,000[a]
	Residual	421,921	28	15,069		
	Total	8106,084	30			

a. Variables predictoras: (Constante), Altura, Diámetro

b. Variable dependiente: Volumen

La tabla ANOVA nos sirve para contrastar
H_0: el modelo no es explicativo en su conjunto.
El p-valor = 0,000… nos hace rechazar H_0
Hay evidencia estadística significativa (con cualquier a mayor o igual que 0,0005) de que el modelo es explicativo en su conjunto.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-57,988	8,638		-6,713	,000
	Diámetro	4,708	,264	,899	17,816	,000
	Altura	,339	,130	,132	2,607	,014

a. Variable dependiente: Volumen

La tabla de los coeficientes nos da el plano de regresión estimado:

$$\mathbf{Y_{x1,x2} = -57{,}988 + 4{,}708\, x_1 + 0{,}339\, x_2}$$

Lo que nos permite estimar el volumen de un cerezo con diámetro x_1 y altura x_2

Los errores típicos nos permiten calcular los intervalos de confianza para los 3 parámetros según la fórmula:

$$IC_{1-\alpha}(\beta_i) = (\hat{\beta}_i \pm t_{n-k-1;\alpha/2} S_R \sqrt{q_{i+1,i+1}})$$

(Lo incluido en el óvalo es el error típico. En este caso n=31, k=2)

Por ejemplo, el intervalo de confianza 0,95 para β_1 sería: 4,708 + - $t_{28,\ 0{,}025}$ 0,264 y el de β_2 sería: 0,339 + - $t_{28,\ 0{,}025}$ 0,130

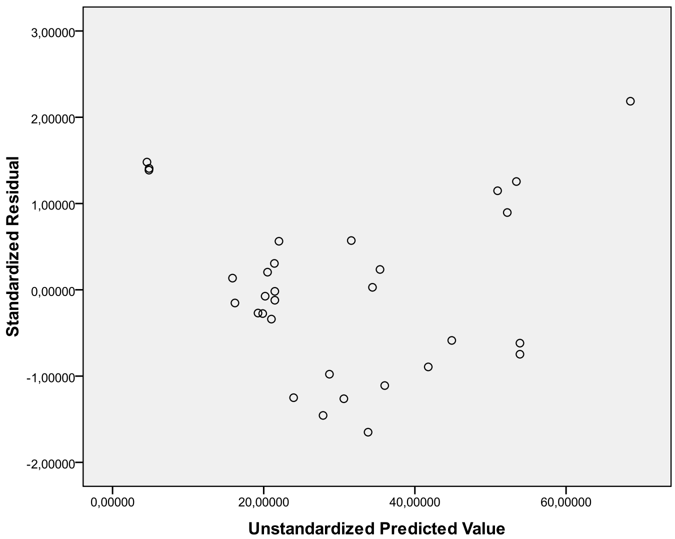
Los valores t y sus correspondientes p-valores nos permiten contratar las hipótesis:

H_0: β_1 =0 (el diámetro no influye)

Se rechaza con cualquier α razonable (p-valor 0,000..)

H_0: β_2 =0 (la altura no influye)

Se rechaza con α > 0,014 (por ejemplo 0,05), se acepta con α < 0,014 (por ejemplo 0,01)



El diagrama de dispersión de los residuos de este modelo presenta la misma curvatura que el de los de la regresión simple.