

Segunda práctica de REGRESIÓN.

DATOS: fichero “practica regresión 2.sf3”

1. Objetivo:

El objetivo de esta práctica es interpretar una regresión y realizar correctamente la diagnosis.

En la primera parte se partirá de la ecuación ajustada, se construirán intervalos de confianza, se interpretará en contraste t, el p-valor y R^2 .

En la segunda parte se realizará la diagnosis de los modelos mediante un gráfico de residuos frente a valores predichos.


Finalmente en la tercera parte se ajustará un modelo a datos que precisan una transformación.

Temas ya conocidos de prácticas anteriores:

- Estudiar si los datos son adecuados para analizarlos mediante regresiones.
- Ajustar la regresión y escribir la ecuación

2. Intervalos y contrastes.

Vamos a utilizar las dos primeras columnas del fichero de datos. Altura muestra la altura de 114 estudiantes de ingeniería. La variable Peso contiene el peso de los estudiantes.



	altura	peso	Peso cuerpo	peso cerebro	Col_5	Col_6	Col_7
1	180	68	3,385	44,5			
2	178	67	0,48	15,5			
3	192	85	1,35	8,1			
4	180	78	465	423			
5	162	48	36,33	119,5			
6	183	83	27,66	115			
7	168	59	14,83	98,2			
8	160	60	1,04	5,5			
9	182	68	4,19	58			
10	172	60	0,4245	6,4			

Pretendemos estudiar cómo depende el Peso (Y) de la Altura (X)

Si realizamos el gráfico de dispersión, vemos que cumple las hipótesis del modelo por lo que ajustamos la recta de regresión, tal como hicimos en la primera práctica, y obtenemos:

Regression Analysis - Linear model: Y = a + b*X

Dependent variable: peso
Independent variable: altura

Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	-90,2579	12,3654	-7,29923	0,0000
Slope	0,907651	0,0704586	12,882	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	6318,99	1	6318,99	165,95	0,0000
Residual	4264,77	112	38,0783		
Total (Corr.)	10583,8	113			

Correlation Coefficient = 0,772687

R-squared = 59,7046 percent

Standard Error of Est. = 6,17076

La ecuación de regresión será:

$$\text{Peso} = -90.3 + 0.9 \text{Altura}$$

Intervalo de confianza:

El intervalo de confianza cuando hay más de 30 datos y quiere construirse al 95% ($\alpha=0.05$) se calcula como:

$$\hat{\beta}_1 \pm 2SE(\hat{\beta}_1)$$

Donde $\hat{\beta}_1$ es el valor estimado del parámetro y $SE(\hat{\beta}_1)$ el error estándar al estimar dicho parámetro.

En nuestro caso el parámetro estimado vale 0.9 y su error estándar se encuentra en la segunda columna de la regresión (Marcado en rojo). Vale 0.07. El intervalo de confianza será por tanto:

$$0.9 - 2 \times 0.07 ; 0.9 + 2 \times 0.07$$

$$0.86 ; 1.04$$

Contraste t:

En teoría ya se ha visto la necesidad de decidir si una variable es o no significativa. Esta decisión se toma en función del valor del estadístico t.

Si $t > 2$ decimos que la variable es significativa.

Si $t < 2$ la variable no es significativa.

El valor del estadístico t se encuentra en la tercera columna de la tabla de resultados. Está marcado en azul. El valor de t para $\hat{\beta}_1$ es 12.88. Como es mayor que 2 podemos concluir que la altura es significativa, es decir que saber la altura de una persona aporta información sobre el peso de la misma.

p-valor:

El p-valor es análogo al contraste t. Si $p\text{-valor} < 0.05$ tenemos evidencia de que la variable x es significativa.

En el ejemplo, como el p-valor vale 0.0 el peso de una persona depende de la altura.

R²:

R² da información de cuánta variación de Y es explicada por X. En la regresión se ha marcado en rojo

R-squared = 59,7046 percent

Este resultado indica que el 59.70% de la variación del PESO entre personas se debe a su ALTURA. Queda un 40.3% debido a otras causas.

La ecuación de regresión se escribe poniendo bajo el parámetro estimado de la pendiente su valor t entre paréntesis.

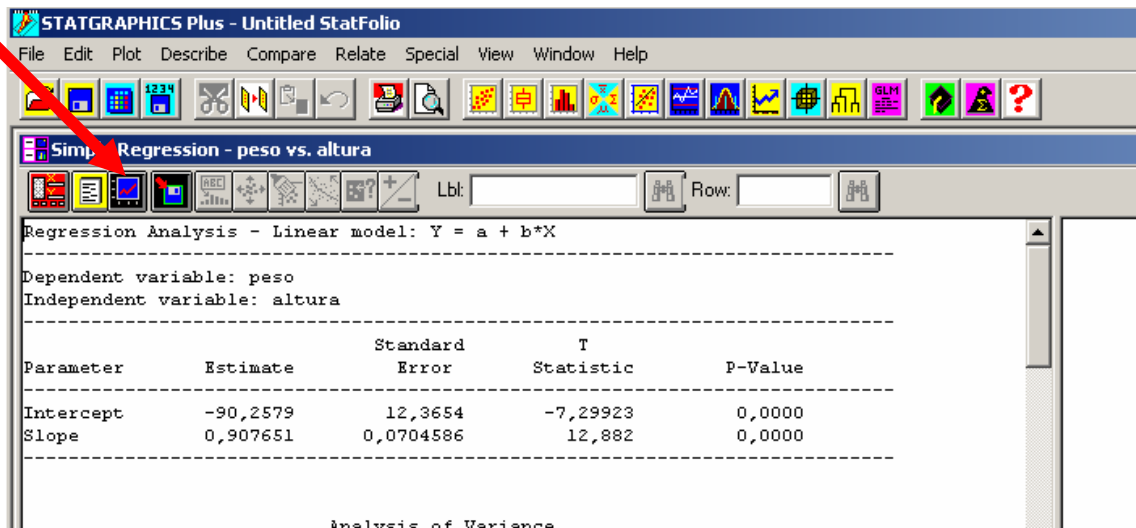
$$\text{Peso} = -90.3 + 0.9 \text{Altura}$$

(12,88)

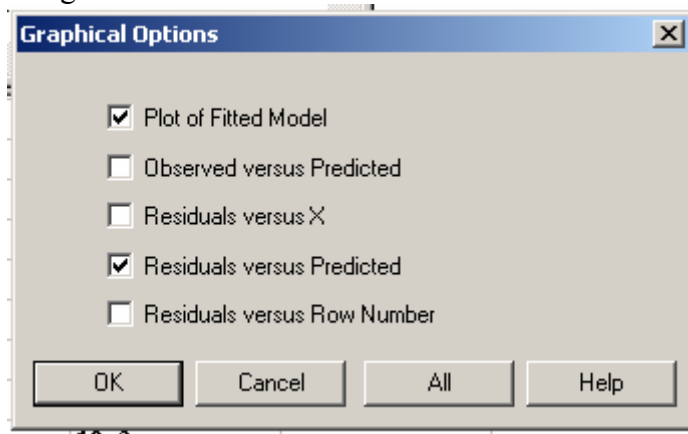
R²=59,7%

3. Diagnosis

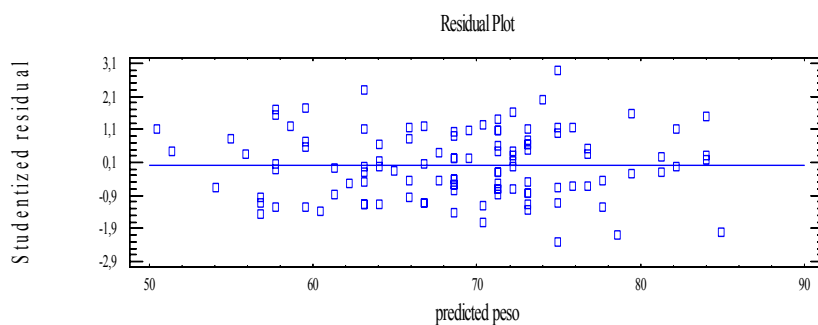
Una vez ajustado el modelo es preciso comprobar las hipótesis a posteriori. Este análisis se hace mediante el gráfico de Residuos frente a Valores Previstos.



Para hacerlo se pincha en el icono de gráficos de la pantalla del resultado. Se obtiene el siguiente menú:



Y se selecciona la opción Residuals versus Predicted, obteniéndose



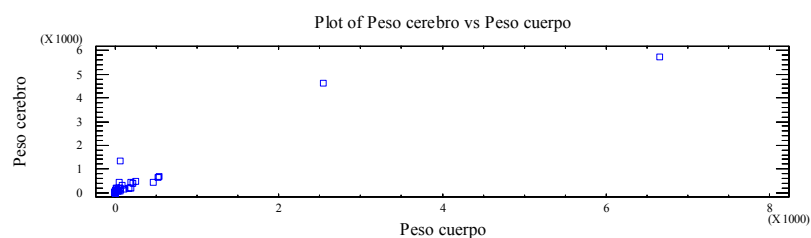
Si los residuos hubiesen mostrado estructura, habría que transformar los datos y volver a ajustar la recta de regresión.

4. Regresión con transformación.

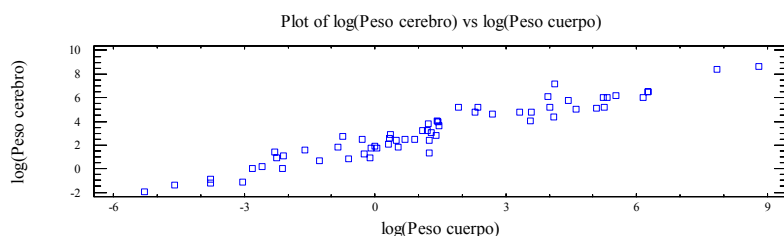
A continuación se va a ajustar un modelo de regresión a los datos de las columnas tercera y cuarta del fichero de datos. (Tomados de Weisberg)

Los datos representan el peso medio del cuerpo y del cerebro para un conjunto de mamíferos. Vamos a ver cómo influye el peso del cuerpo en el tamaño del cerebro.

Haciendo un gráfico de dispersión encontramos:



Como puede verse, los datos no cumplen las hipótesis. Vamos a transformar ambas variables a logaritmos.



Tras la transformación podemos ajustar la regresión.

Regression Analysis - Linear model: $Y = a + b \cdot X$

Dependent variable: log(Peso cerebro)

Independent variable: log(Peso cuerpo)

Standard Parameter	T Estimate	Error	Statistic	P-Value
Intercept	2,13481	0,0960442	22,2274	0,0000
Slope	0,751682	0,0284638	26,4084	0,0000

Correlation Coefficient = 0,959574

R-squared = 92,0782 percent

Standard Error of Est. = 0,694302

Con los datos proporcionados:

1. Escribir la ecuación de regresión.
2. Construir un intervalo de confianza para la pendiente.
3. ¿Es significativo el peso del cuerpo para determinar el peso del cerebro?
4. Cuantificar el efecto de un incremento del peso del cuerpo sobre el peso del cerebro.
(Atención a los logaritmos que implican relaciones porcentuales)
5. Diagnóstico del modelo. El gráfico de residuos ¿Es adecuado?
6. ¿Qué tamaño de cerebro previsto tendrá un mamífero de 80Kg?