

## **II Jornadas Internacionales de Estadística Aplicada**

### **5 y 6 De Diciembre de 2019**

### **REGRESION LINEAL SIMPLE: VARIABLE PESO Y ALTURA**

Autores: Anachuri, Rodrigo Enrique - Lauret, Jesús Abel Dario - Tapia Campos, Franco Israel.

Institución: Facultad de ingeniería, Universidad Nacional de Jujuy, San Salvador de Jujuy

*Datos de Contacto:israloa17@gmail.com-3884560527*

#### **RESUMEN**

En el presente informe de investigación se abordará el tema de regresión lineal simple en una muestra de 220 estudiantes que respondieron a una encuesta realizada por la cátedra de probabilidad y estadística de la FI- UNJu.

El objetivo fue encontrar, si existe una relación entre la altura y el peso de dichos estudiantes.

Se planteó como modelo matemático una regresión lineal simple, con su respectivo gráfico de dispersión, la cual se hizo a través del método de mínimos cuadrados; obtenido este modelo se realizaron inferencias sobre los coeficientes de regresión: hipótesis sobre la pendiente (verificar si era nula), predicción de valores particulares y para la respuesta media del peso.

#### **Palabras clave:**

Relación-Pendiente-Muestra de regresión lineal simple (MRLS)-Residuos- Centro de Gravedad Estadístico.

#### **INTRODUCCION**

A lo largo de los años se construyó la idea de que el peso y la estatura de una persona tienen algún tipo de relación entre sí, a partir de diversas investigaciones se obtuvieron modelos matemáticos que utilizan estas variables para poder generar conclusiones sobre el estado de salud de un individuo. Algunos de los más conocidos son, el IMC de Adolphe Quetelec, los gráficos de percentil desarrollados por la OMS y utilizados en pediatría, entre otros. Pero, ¿Es posible que exista entre el peso y la estatura de una persona saludable una relación lineal? Y si fuese así, ¿Podemos saber el grado de relación entre sí?

El concepto de análisis de regresión se refiere a encontrar la mejor relación entre, el peso representado por la variable dependiente Y, y la estatura como la variable independiente X, cuantificando la fuerza de esa relación, y empleando métodos que permitan predecir los valores de la respuesta Y, dados los valores del regresor X.

En este trabajo buscamos generar un modelo de regresión lineal simple para estimar el peso en función de la altura de los estudiantes de nuestra facultad utilizando datos muestrales.

El objeto de la regresión lineal es investigar la relación estadística que existe entre una variable dependiente (Y) y una independiente (X). Para poder realizar esta investigación, se debe postular una relación funcional entre las variables, la cual gráficamente representa una recta.

Cuanta mayor correlación haya entre dos variables, en la representación bidimensional, los valores estarán reunidos más próximos a la recta.

La ecuación de una recta destacamos que  $\beta_0$  es la ordenada al origen (Corte con el eje Y), y  $\beta_1$  la pendiente:  $Y = \beta_0 + \beta_1 x$ .

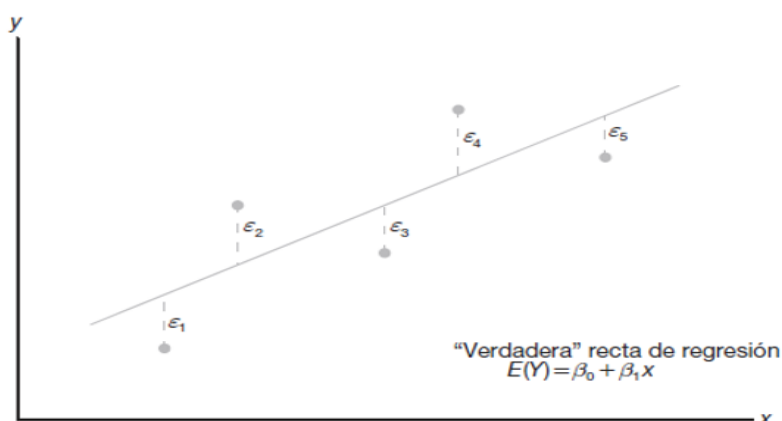


Gráfico: Observaciones individuales alrededor de la verdadera recta de regresión

## METODOLOGIA

En primer lugar, se recolectaron 220 datos a través de una encuesta elaborada a comienzos del ciclo lectivo, postada en el aula virtual de la cátedra Probabilidad y Estadística.

Los datos obtenidos fueron trasladados a una planilla Excel, donde se utilizaron sus herramientas para graficar y calcular.

Con el objeto de ganar experiencia con software estadístico también se hizo uso de Minitab.

## DESARROLLO

### INTERPRETACIÓN PROBABILISTICA DE LA REGRESION

Un diagrama de dispersión es una representación gráfica de los puntos de datos de una muestra en particular. Al escoger una muestra diferente, o aumentar la original, un diagrama de dispersión algo distinto se obtendría generalmente. Cada diagrama de dispersión resultaría en una recta de regresión diferente, aunque esperamos que las diferencias no sean significantes si las muestras se extraen de la misma población.

Las dispersiones de puntos alrededor de una recta de regresión indican que para un valor particular de x hay realmente varios valores de y distribuidos alrededor de la

recta. Esta idea de distribución nos conduce a la realización de que hay una conexión entre la recta de ajuste y probabilidad.

Modelo de Regresión Lineal Simple del Peso en función de la Altura de cada estudiante (Peso vs Altura).

**Identificación de variables:** Peso Neto en Kilogramos (Kg) de una muestra tomada de los estudiantes de la Facultad de Ingeniería de San Salvador de Jujuy que cursan la materia probabilidad y estadística, y la Altura misma correspondiente a cada uno medida en centímetros (cm). **(Datos reales). Tipo de Variables: Variables Cuantitativas Continuas, Escala de Medición: de Razón.**

**Objetivo:** Establecer un modelo de regresión lineal que permita estimar el Peso de cada estudiante en función de la Altura del mismo.

**Aclaración:** Para desarrollar nuestro trabajo tuvimos un problema debido al volumen de datos que estábamos por evaluar, entonces tomamos la decisión de agrupar dichos datos en nueve intervalos (de acuerdo a la regla de Sturges), Y a su vez trabajar con los centros de gravedad de cada intervalo, tomando como X (Variable Independiente) un valor promedio por cada intervalo y como Y (Variable Dependiente) un valor promedio de los datos que se encuentran en cada intervalo de la Altura. Decidimos trabajar con los centros de gravedad observando el problema de forma física y trabajar con un punto específico en vez de con un forma de datos muy variada. De tal manera concluimos que el punto generado por los promedios tanto en X como en Y no es otra que el centro de gravedad estadístico, De esta forma planteamos trabajar con una cantidad más reducida de datos, contando solo con las medias aritméticas de los intervalos.

#### **Puntos a tener en cuenta:**

Adjuntamos también un análisis de los datos muestrales separándolos en mujeres y varones.

#### **Observaciones:**

Cuando se obtienen pesos y/o alturas de sujetos, es sumamente importante pesarlos y medirlos de forma presencial en vez de pedirles que reporten sus propias mediciones. Se sabe que cuando la gente reporta su peso, generalmente da un peso más bajo que el real y una altura mayor. Entonces, ¿Cómo pueden verificar los investigadores que los pesos y alturas se obtuvieron por medio de mediciones reales y no por el reporte de los sujetos?

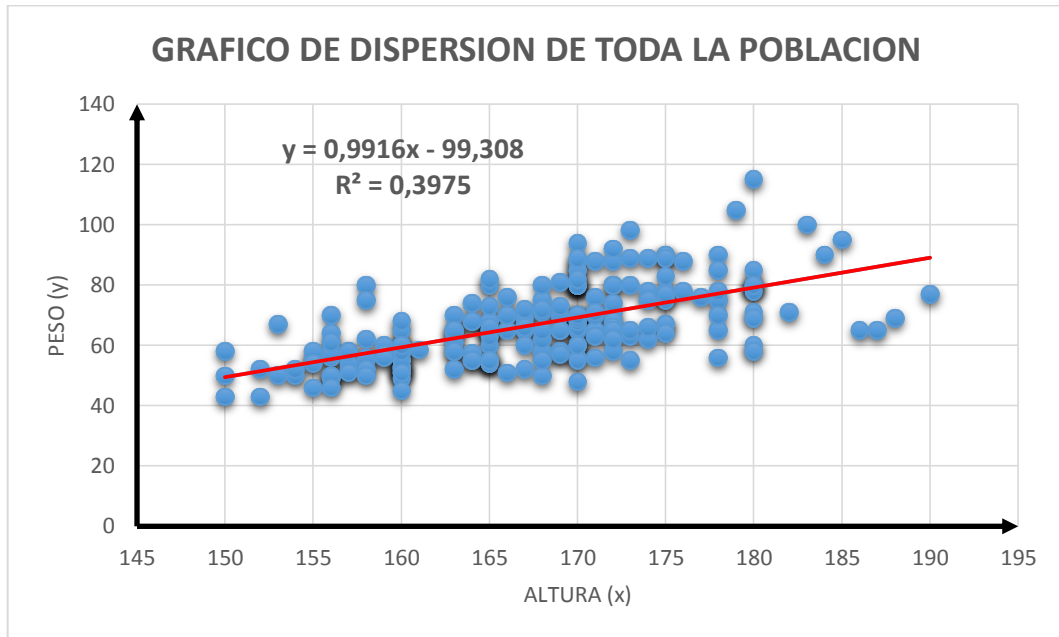
A modo de práctica para nosotros como estudiantes fue todo un reto encontrar un modelo que relacione las variables de la forma más adecuada para poder evidenciar dicha correlación con los 220 datos de la muestra.

Sin embargo, si quisiéramos ser más exactos tendríamos que analizar los últimos dígitos de los pesos. Cuando la gente reporta su peso, tiende a redondear la cifra, a menudo hacia el entero inferior. Los últimos dígitos de los pesos reportados suelen tener un número desproporcionado de ceros y cincos, en comparación con los últimos dígitos de los pesos obtenidos a través de un proceso de medición

#### Análisis de la regresión

### Grafico del MRLS para el gráfico de dispersion utilizando Excel

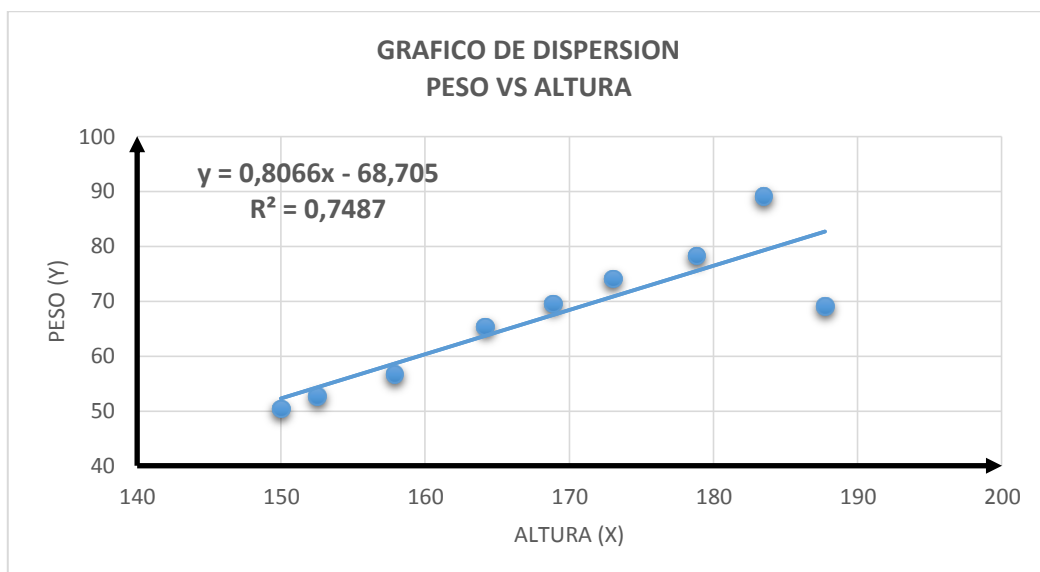
- Toda la muestra sin intervalos



Este gráfico de dispersión usando todos los puntos muestrales y cómo podemos observar los puntos están muy dispersos y el coeficiente de correlación es muy bajo, por ello decidimos trabajar por intervalos para obtener una gráfica más real.

### Grafico del MRLS para el gráfico de dispersion utilizando Excel

- Toda la muestra con intervalos



**¿Qué podemos observar con nuestra ecuación de regresión?**

Con la ecuación de regresión lineal simple estimada para las variables, independiente "estatura" y dependiente "peso", podemos observar que de acuerdo a la distribución t, muestran relación.

Esta relación que se ha estimado con un  $R=0,86$ , que indica dentro de toda una fuerte relación lineal positiva entre las variables peso y altura, por lo que determinamos una proporcionalidad directa entre X e Y.

Además, si consideramos el  $R^2=74,8\%$  podemos estimar que casi el 75% de las variaciones que caen en el peso se explicarían por las variaciones en la variable estatura.

### CÁLCULO E INTERPRETACIÓN DE LOS COEFICIENTES PARA LA RECTA DE REGRESIÓN A TRAVÉS DE LOS MÍNIMOS CUADRADOS:

Ecuaciones de Sumatorias para los diferentes términos de la tabla de Excel representan la sumatoria de cada término, de sus cuadrados y su producto. Éstos nos sirven para calcular la ecuación de la recta de regresión.

$\sum x_i =$	1516,5
$\sum y_i =$	604,87
$\sum x_i^2 =$	256982,8695
$\sum y_i^2 =$	41914,0973
$\sum y_i \cdot x_i =$	103092,173

$$S_{yy} - b_1 S_{xy}$$

- Cálculo del Término  $S_{xx}$ , Suma de los Cuadrados de X (Altura)

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}$$

$$S_{xx}=1452,43741$$

- Cálculo del Término  $S_{yy}$ , Suma de los Cuadrados de Y (Peso)

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n}$$

$$S_{yy}=1262,128756$$

- Suma de Productos Cruzados XY

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}$$

$$S_{xy}=1171,541668$$

### □ **Cálculo de los coeficientes**

**Pendiente  $b_1$ :** Me indica la variabilidad del Peso respecto de la Altura, como este término es positivo la variación es creciente (mayor altura implica un mayor peso). Este término también es llamado coeficiente de Regresión.

**Intercepto  $b_0$ :** Este término me dice el Peso correspondiente a la Altura 0 el cual es - 68,705 Kg, un valor no realista, pero que me sirve para construir la recta de regresión ajustada para estimar más valores.

**Fórmula de cálculo de la pendiente de la recta de regresión.**

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

También:

$$b_1 = S_{xy} / S_{xx}$$

$$b_1 = \mathbf{0,8066038877}$$

**Fórmula de cálculo de la ordenada al origen de la recta de regresión.**

$$b_0 = \frac{\sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i}{n} = \bar{y} - b_1 \bar{x}$$

$$b_0 = \mathbf{-68,70502571}$$

- Con los datos obtenidos podemos realizar la ecuación de la regresión lineal que estemos trabajando

La ecuación de regresión es

$$\mathbf{PESO (Y) = - 68,71 + 0,8066 ALTURA(x)}$$

### **¿Qué nos dice esta ecuación?**

Interpretación de la ecuación de la regresión lineal.

El  $b_0$  (-68,70) es el valor para la variable peso con una altura nula.

El  $b_1$  o pendiente (0.8066) es como crece el peso por cada cm de altura. En este caso se puede decir que por cada **1 cm aumenta 0.8066 kg**.

### □ **Cálculo de la varianza**

Variación no explicada: SCE = **317,1586914**

$$\text{Varianza}(S^2) = \sum_{i=1}^n \frac{(y_i - \bar{y})^2}{n - 2}$$

$$S^2 = \mathbf{45,30838448}$$

Indica un promedio de la varianza que tienen todos los puntos con respecto a la recta ajustada de regresión lineal.

### ¿Cómo podemos saber si realmente es una regresión lineal?

Para saber si se trata realmente de una regresión lineal procedemos al cálculo de la hipótesis, donde planteamos como hipótesis nula ( $H_0$ ), que la pendiente sea igual a cero ( $\beta_1 = 0$ ) y como hipótesis alternativa ( $H_1$ ), que la pendiente sea distinta de cero ( $\beta_1 \neq 0$ ). Si se rechaza la hipótesis nula se puede decir que es una regresión lineal significativa.

#### ➤ Cálculo de Hipótesis para la pendiente de la recta ( $\beta_1$ )

ESTADÍSTICO DE PRUEBA T
2,36462
PRUEBA DE HIPÓTESIS

$H_0 =$	$\beta_1 = 0$
$H_1 =$	$\beta_1 \neq 0$
$\alpha = 0,05$	$\alpha/2 = 0,025$

TCAL
4,566881436

TA
2,36462

REGION DE RECHAZO		
Rechazar si $T_{cal} < -T_{\alpha/2, n-2}$	ó	$T_{cal} > T_{\alpha/2, n-2}$ DOS COLAS

Se rechaza la hipótesis nula por lo tanto se acepta la hipótesis alternativa, esto implica que **si existe relación lineal significativa** entre las variables

PERCENTILES	VALOR X
25	160
50	168
75	172
60	170

$y = 0,8066x - 68,705$	
Y =	60,351
Y =	66,8038
Y =	70,0302
Y =	68,417

También para realizar un mejor análisis realizamos el intervalo de la pendiente

$$b_1 - t_{\alpha/2} \frac{s}{\sqrt{S_{xx}}} < \beta_1 < b_1 + t_{\alpha/2} \frac{s}{\sqrt{S_{xx}}},$$

LÍMITE INFERIOR	< B1 <	LÍMITE SUPERIOR
0,3889640362	< B1 <	1,224243739

Con el intervalo de la pendiente también se puede realizar un análisis para poder saber si es una regresión lineal. Si el 0 no está incluido en el intervalo de la pendiente, eso nos dice que **si es una regresión lineal significativa**.

#### INTERVALO DE CONFIANZA

$$\hat{y}_0 - t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < \mu_{Y|X_0} < \hat{y}_0 + t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},$$

LI	<μ <sub>Y</sub>  X <sub>0</sub> <	LS
63,0746083		73,7593917

Podemos estimar que con un intervalo de confianza del 95%, la media poblacional del peso de los alumnos estará entre 63,0746 kg y 73,7594 kg, cuando la altura sea de 170 cm.

#### INTERVALO DE PREDICCIÓN

$$\hat{y}_0 - t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < y_0 < \hat{y}_0 + t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},$$

LÍMITE INFERIOR	<Y <sub>0</sub> <	LÍMITE SUPERIOR
51,62772643		85,20627357

#### □ Cálculo del coeficiente de Determinación:

Coeficiente de determinación:  $R^2 = 1 - \frac{SCE}{STCC} = 0,7487113023$

**Coeficiente de determinación  $R^2$ :** Es una medida de qué tan bien explica el modelo matemático la variabilidad del Peso en función de la Altura y está explicado en un



74,9%, esto quiere decir que es probable que la estimación es buena en ciertos puntos del modelo matemático, lo cual coincide con el Cálculo de Hipótesis.

Nota: STCC es la Suma Total de los cuadrados corregida: =  $\sum_{i=1}^n (y_i - \bar{y})^2$

r: Coeficiente de correlación muestra: =  $\frac{S_{xy}}{\sqrt{S_{xx} * S_{yy}}} = 0,8652810539$

Entonces podemos decir que existe una relación lineal bastante significativa entre nuestras variables PESO y ALTURA. Con esto también podemos confirmar se espera que la estimación de la regresión sea buena.

Fórmula de cálculo de la pendiente de la recta de regresión:

$$\hat{y} = b_0 + b_1 * x$$

$$\bar{Y} = 0,8066x - 68,705$$

### USO DE LA ECUACION DE REGRESION PARA HACER PREDICCIONES

Las ecuaciones de regresión a menudo se utilizan para predecir el valor de una variable, dado algún valor particular de la otra variable. Si la recta de regresión se ajusta bastante bien a los datos, entonces será sensato utilizar su ecuación para hacer predicciones, siempre y cuando no vayamos más allá del alcance de los valores disponibles.

Determinada la recta de regresión lineal podemos comenzar a pronosticar el posible valor de cualquier altura x que ingresemos en la ecuación, por ejemplo:

Para una Altura de 175 cm tenemos que:

$$\bar{Y} = 0,8066 * 175 - 68,705$$

$$\bar{Y} = 72,45$$

Esto quiere decir que según nuestro modelo matemático de estimación el peso más probable correspondiente a una altura de 175 cm sea 72,45 Kg.

### A continuación, se realizó una comparación de valores obtenidos con Minitab

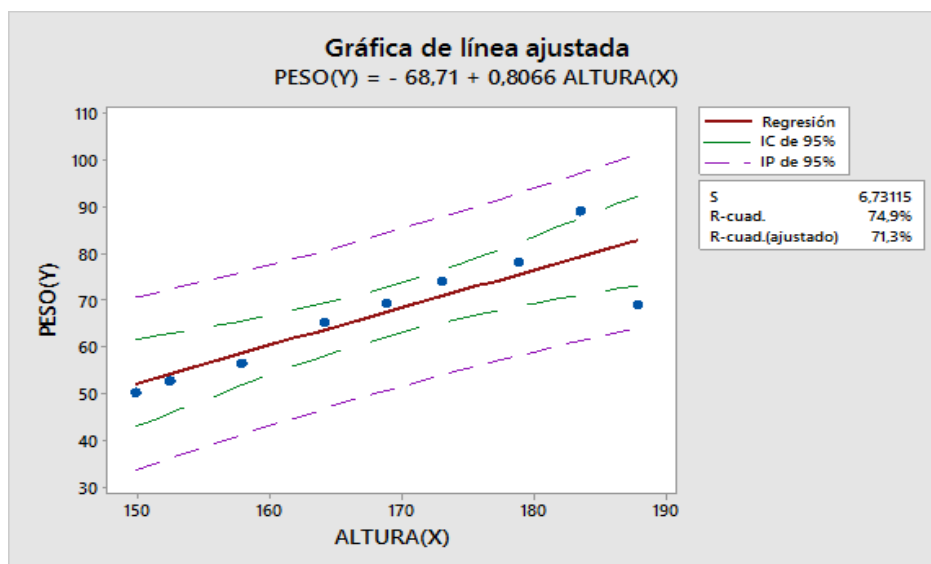
Análisis de regresión: PESO(Y) vs. ALTURA

La ecuación de regresión es: PESO(Y) = - 68,71 + 0,8066 ALTURA

S = 6,73115 R-cuad. = 74,9% R-cuad.(ajustado) = 71,3%

Análisis de Varianza

Fuente	GL	SC	MC	F	P
Regresión	1	944,97	944,970	20,86	0,003
Error	7	317,16	45,308		
Total	8	1262,13			



### Análisis de Residuales

Un residual es la diferencia entre el valor observado y el valor estimado por la línea de regresión, El residual puede ser considerado como el error aleatorio observado. También se acostumbra usar el Residual estandarizado, el cual se obtiene al dividir el residual entre la desviación estándar del residual, y el Residual estudentizado "deleted", que es similar al anterior, pero eliminando de los cálculos la observación cuyo residual se desea hallar.

En un análisis de residuales se puede detectar:

- Si efectivamente la relación entre las variables X e Y es lineal.
- Si hay normalidad de los residuos
- Si hay valores anormales en la distribución de residuos.
- Si hay varianza constante (propiedad de Homocedasticidad).
- Si hay independencia de los residuos.

**Plot de Normalidad:** Permite cotejar normalidad. Si los puntos están bien cerca de una línea recta se concluye, que hay normalidad.

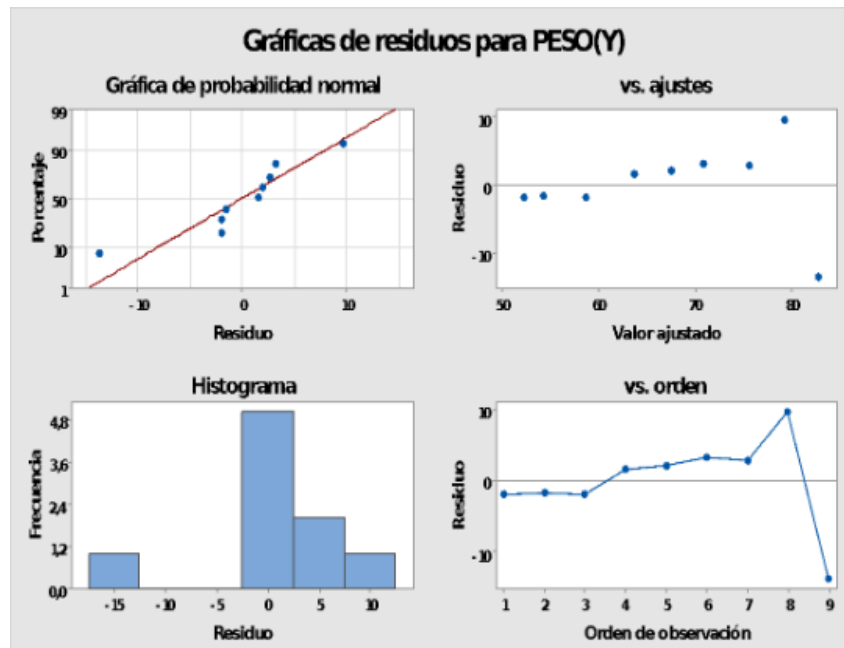
**Histograma de Residuales:** También permite cotejar normalidad. Cuando el histograma es simétrico, con un único pico en el centro, se concluye que hay normalidad.

**Plot de Residuales versus los valores predichos (FITS):** Se usa para detectar si hay datos anormales, cuando hay datos que caen bastantes alejados, tanto en el sentido vertical como horizontal. También permite detectar si la varianza de los residuos es constante con respecto a la variable de respuesta.

**Plot de Residuales versus el índice de la observación:** Es más específico para detectar que observación es un dato anormal. Si se usan residuales estandarizados,

entonces un dato con residual más allá de 2 ó -2 es considerado un "outlier" en el sentido vertical.

**Plot de Residuales versus la variable predictora:** Es usado para detectar datos anormales, así como si la varianza de los residuos es constante con respecto a la variable predictora.



## INTERPRETACION DE NUESTRAS GRAFICAS DE RESIDUOS

En el gráfico de la esquina superior izquierda se puede observar el gráfico de probabilidad normal se puede observar que no todos los puntos se encuentran sobre la recta por lo que se puede decir que no sigue una distribución normal. Si analizamos más detalladamente podemos observar una forma de S, lo que nos indicaría que tiene una distribución de cola larga con la mayor frecuencia de datos en el centro y con los mismos decreciendo hacia los extremos muy lentamente.

En el gráfico contiguo se observa el valor ajustado versus los residuos, los cuales son las distancias verticales entre el punto y la recta de regresión lineal, concluyendo que sus picos son los valores más atípicos y por ende los más alejados de la recta de estimación.

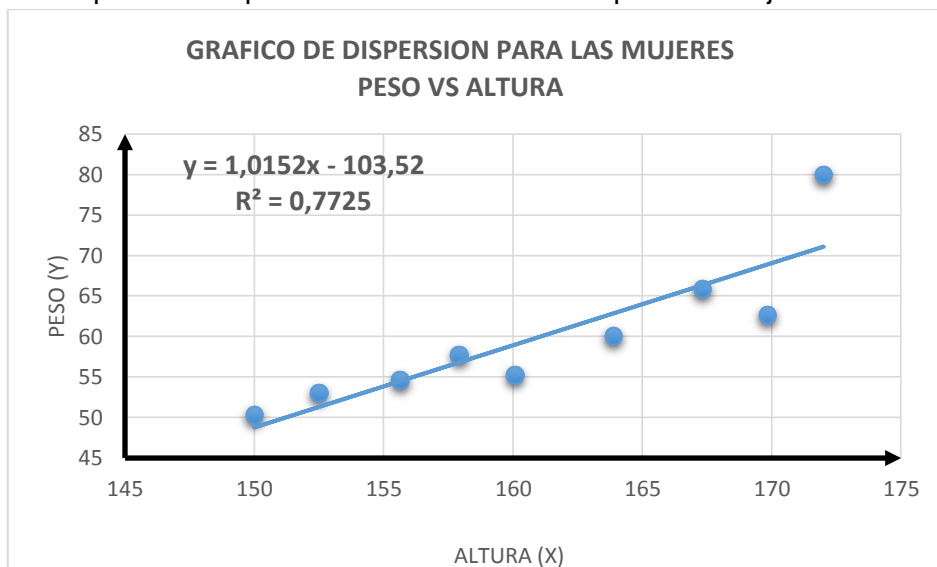
En el gráfico de orden se puede observar que los puntos siguen un cierto orden en la mayor parte del gráfico lo que explicaría un valor correlación elevado, debido al poco ajuste que se le deben hacer a los puntos para que se ajusten a la recta.

Con el gráfico de histograma no se puede realizar un gran análisis debido a los pocos datos que se evalúan, pero al ver esa distancia entre las columnas podemos decir que existe un dato atípico en la muestra.

Para poder tener un mejor análisis sobre el peso en función de altura podemos separar los datos en muestras de mujeres y varones.

### ANÁLISIS DE LOS DATOS CORRESPONDIENTES A LAS MUJERES

Gráfico de dispersión del peso en función de la altura para las mujeres



Como se puede observar a primera vista la pendiente cae en el intervalo de la pendiente realizada. También se puede observar que como la general también tiene relación una significativamente alta, por lo tanto la muestra de mujeres también presenta una correlación significativa

#### ➤ Análisis en Minitab

Análisis de regresión: Promedio de Altura (y) vs. Promedio de Peso(x)

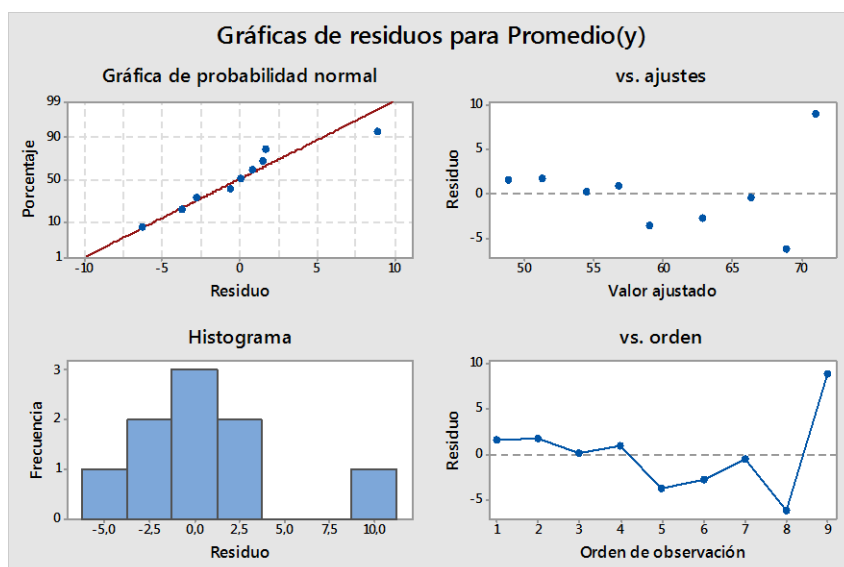
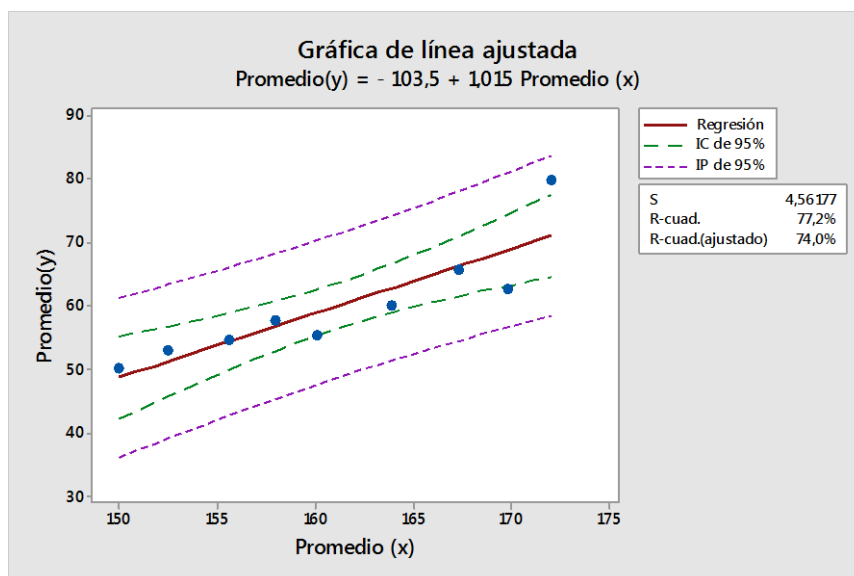
La ecuación de regresión es:

Promedio(y) = - 103,5 + 1,015 Promedio (x)

S = 4,56177 R-cuad. = 77,2% R-cuad.(ajustado) = 74,0%

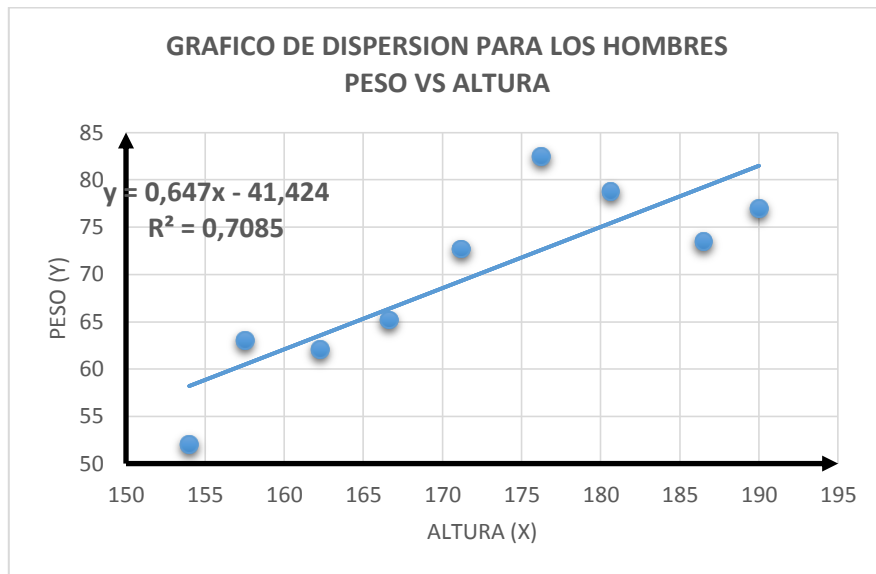
Análisis de Varianza

Fuente	GL	SC	MC	F	P
Regresión	1	494,623	494,623	23,77	0,002
Error	7	145,668	20,810		
Total	8	640,291			



## ANÁLISIS DE LOS DATOS CORRESPONDIENTES A LOS VARONES

Gráfico de dispersión del peso en función de la altura para los varones



En este análisis se puede ver que los varones tienen una correlación más baja que el de las mujeres. Esto se puede deber a que las mujeres tienen una mejor alimentación que los varones. También se puede tener en cuenta que las mujeres cuidan más su figura que los varones. En cambio, los hombres tienen una alimentación diversa y no tan buena.

#### □ **Análisis en Minitab**

Análisis de regresión: PESO vs. ALTURA(X)

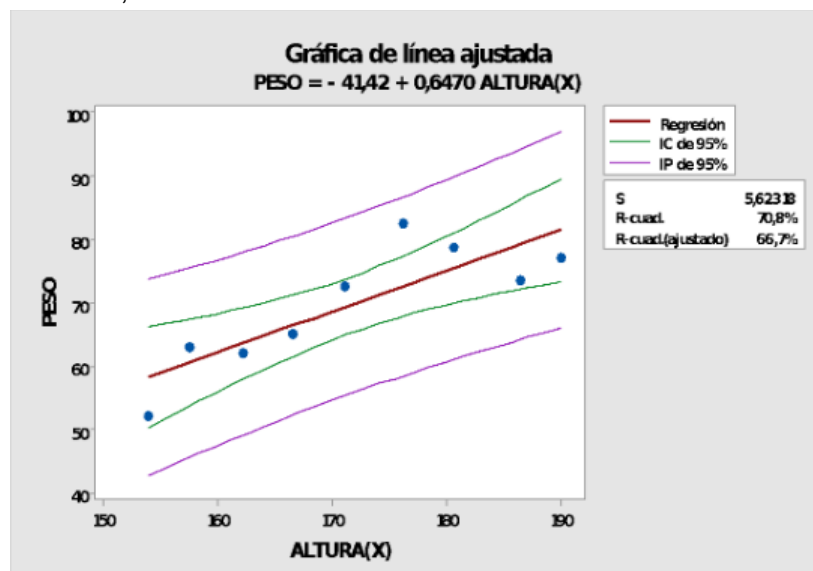
La ecuación de regresión es:

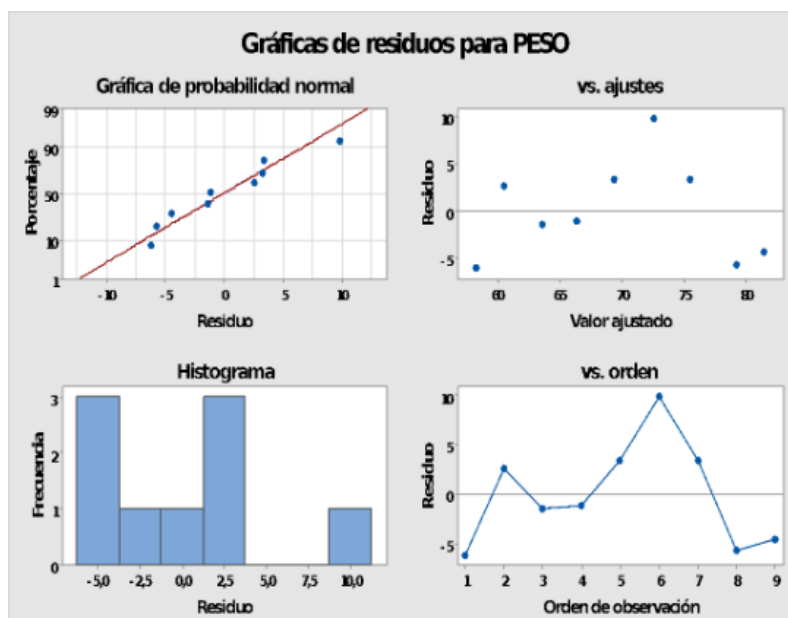
PESO = - 41,42 + 0,6470 ALTURA(X)

S = 5,62318 R-cuad. = 70,8% R-cuad.(ajustado) = 66,7%

Análisis de Varianza

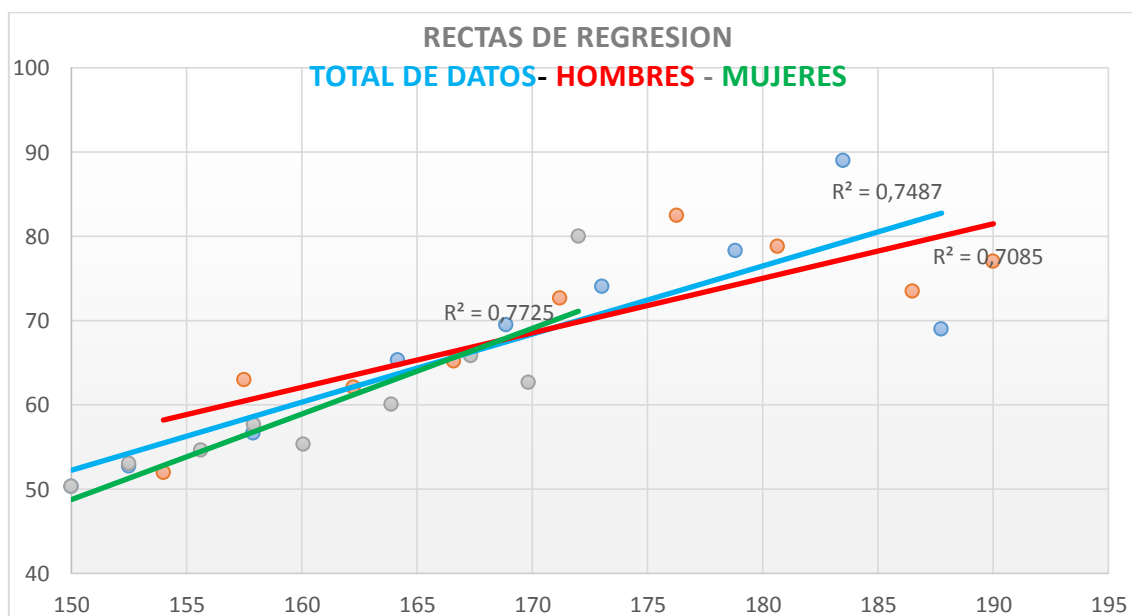
Fuente	GL	SC	MC	F	P
Regresión	1	537,917	537,917	17,01	0,004
Error	7	221,341	31,620		
Total	8	759,258			





## PARA FINALIZAR

Decidimos superponer las tres rectas de regresión lineal para analizar los coeficientes de determinación, y así ver si existe una relación entre las tres gráficas.



Como podemos observar los coeficientes de determinación no varían mucho, son similares. Las rectas se intersectan en un punto. Esto nos quiere decir que el análisis de los tres modos será similar.

## CONCLUSIONES:

Analizando todos los datos que se obtuvieron en la muestra se llegaron a las siguientes conclusiones:

- La relación peso-altura es buena según el coeficiente de correlación ( $r$ ) y de determinación.
- Se rechaza la hipótesis nula por lo tanto se acepta la hipótesis alternativa (que la pendiente sea distinta a cero).
- Los intervalos para los valores son bastante buenos para la relación peso-altura.
- Finalmente, después de habernos tomado el tiempo de analizar estos datos, podemos concluir y decir con cierto grado de seguridad que:

### **SI EXISTE RELACION LINEAL ENTRE PESO Y ALTURA**

#### **BIBLIOGRAFIA:**

##### **Libros:**

- + **Estadística para administración y economía, 10ª. Edición.** Anderson, David R. Dennis J. Sweeney y Thomas A. Williams **Ed. Cengage Learning.**
- + **Estadística. Novena edición.** Triola, Mario F. **Ed. Pearson educacion**
- + Apuntes de la cátedra de Probabilidad y Estadística Facultad de Ingeniería de la UNJU.

##### **Páginas:**

- + <https://support.minitab.com>