PROBABILIDAD Y ESTADÍSTICA

Año de cursada: 1°

Clase N. º4: Análisis de regresión y correlación.

Contenido: Diagramas de dispersión. Ajuste de curvas. Método de los mínimos cuadrados.

1. **Presentación:**

Bienvenidos y bienvenidas a la clase n° 5 de Probabilidad y Estadística.

En las clases anteriores estuvimos trabajando acerca de los conceptos básicos de la Estadística, así como también hemos avanzado sobre varios aspectos referidos a la organización y presentación de información. Asimismo, en las últimas clases hemos abordado las medidas de resumen que de manera acotada podríamos pensar que son una forma de sintetizar la información, y es aquí donde aquellos valores característicos del conjunto en cuestión se convierten en un dato estadístico pues éste proviene de un proceso al cual se sometió la información inicial.

En esta clase vamos a trabajar sobre un análisis de dos variables en simultaneo, en Estadística se conoce como Análisis de Regresión y correlación.

El objetivo principal del Análisis de correlación y regresión es estimar el valor de una de las variables conociendo el valor de la otra, es decir, establecer una ecuación matemática que relacione estas variables de la distribución bidimensional.

2. **Desarrollo:**

Es común que las personas tomen decisiones personales y profesionales basadas en predicciones de sucesos futuros. Para hacer estos pronósticos, se

basan en la relación intuitiva y calculada entre lo que ya se sabe y lo que se debe estimar. Si los responsables de la toma de decisiones pueden determinar cómo lo conocido se relaciona con un evento futuro, pueden ayudar considerablemente al proceso de toma de decisiones.

Cualquier método estadístico que busque establecer una ecuación que permita estimar el valor desconocido de una variable a partir del valor conocido de una o más variables, se denomina **análisis de regresión**.

El término regresión fue utilizado por primera vez por el genetista y estadístico inglés Francis Galton (1822-1911). En 1877 Galton efectúo un estudio que demostró que la altura de los hijos de padres altos tendía a retroceder, o "regresar", hacia la talla media de la población. Regresión fue el nombre que le dio al proceso general de predecir una variable, (la talla de los niños) a partir de otra (la talla de los padres).

Hoy en día, esta tendencia de miembros de cualquier población que están en una posición extrema (arriba o debajo de la media poblacional) en un momento, y luego en una posición menos extrema en otro momento, (ya sea por sí o por medio de sus descendientes), se llama **efecto de regresión**.

En el análisis de regresión se desarrolla una ecuación de estimación, es decir, una fórmula matemática que relaciona las variables conocidas con las desconocidas. Para describir la forma de la relación que liga dos variables (x, llamada independiente; y llamada dependiente) se utiliza los llamados modelos de regresión. Esta se utiliza para predecir: se desarrolla un modelo que utiliza la variable independiente x, para obtener una mejor predicción de la otra variable dependiente y.

Luego de obtener el patrón de dicha relación, se aplica el análisis de correlación para determinar el grado de relación que hay entre las variables. Es decir, en contraste con el de regresión, se utiliza para medir la fuerza de

la asociación entre las variables. Por ejemplo: peso – estatura. Esta fuerza de correlación se mide a través de un indicador del grado de intensidad de la relación de la relación entre las dos variables que es independiente de sus escalas de medición, llamado coeficiente de correlación lineal o coeficiente de correlación de Pearson.

Ahora bien, vamos a definir los diagramas de dispersión, los cuales serán de gran utilidad para el trabajo en esta unidad.

Diagramas de dispersión

Un primer paso es recolectar datos que muestren los valores correspondientes de las variables en consideración. Por ejemplo, supongamos que X e Y denotan, respectivamente, la estatura y peso de los alumnos de una clase.

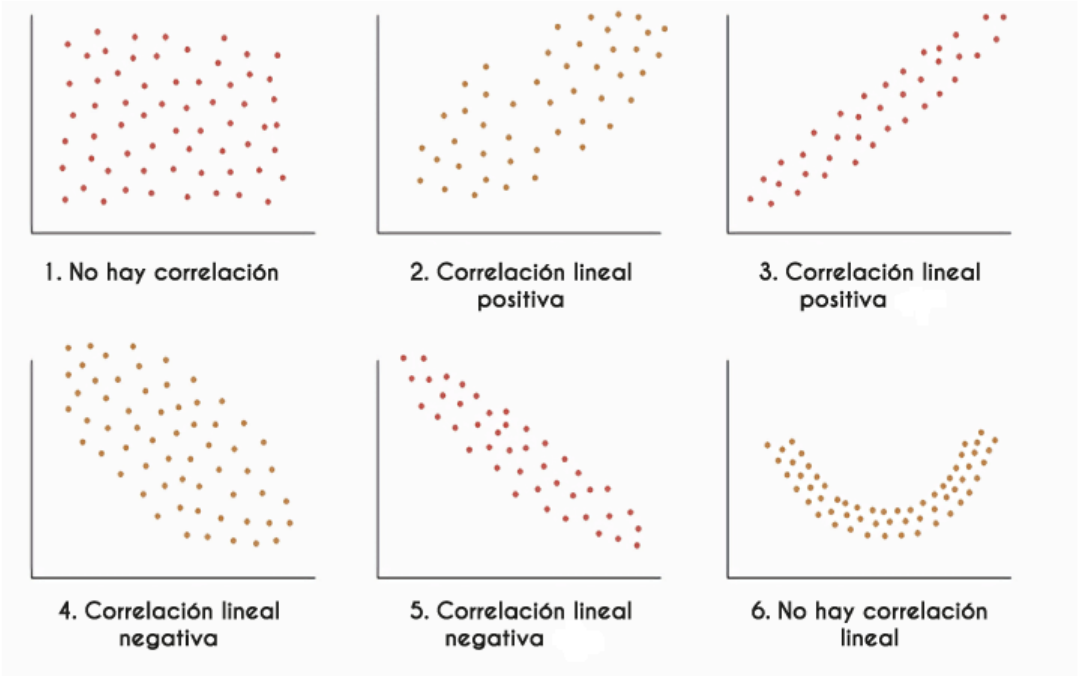
Recordemos que el peso de la persona depende de su estatura, podríamos decir que a mayor estatura del alumno mayor será su peso; X es una variable inde pendiente, mientras que Y es una variable dependiente. Así, una muestra de n individuos revelará las estaturas x_1; x_2; …; x_n y los correspondientes pesos y_1; y_2; …; y_n.

El siguiente paso es graficar los puntos (x_1; y_1); (x_2; y_2); … ; (x_n; y_n) en un sistema de coordenadas cartesianas ortogonal. El valor de la variable independiente se grafica con respecto al eje horizontal o eje de las abscisas (X) y el valor de la variable dependiente con respecto al eje vertical o eje de las ordenadas (Y). En el siguiente enlace podrán observar cómo se representan los puntos cuando construimos un diagrama de dispersión, y también un ejemplo del mismo.

Entonces, un **diagrama de dispersión** es una gráfica en el que se trazan cada uno de los puntos que representan un par de valores para las variables independiente y dependiente, observados en una muestra.

A partir del diagrama de dispersión es posible visualizar una curva suave que se aproxima a los datos. Tal curva se denomina curva de aproximación.

Si los datos parecen aproximarse bien a una línea recta, se dice que hay una relación lineal; en cambio, si los datos parecen ajustarse a una línea curva, se dice que existe una relación no lineal.



Obsérvese que en primera gráfica se indica que no existe relación, y ello se deba a que justamente la distribución de la gráfica de los puntos no refiere a ninguna curva en particular.

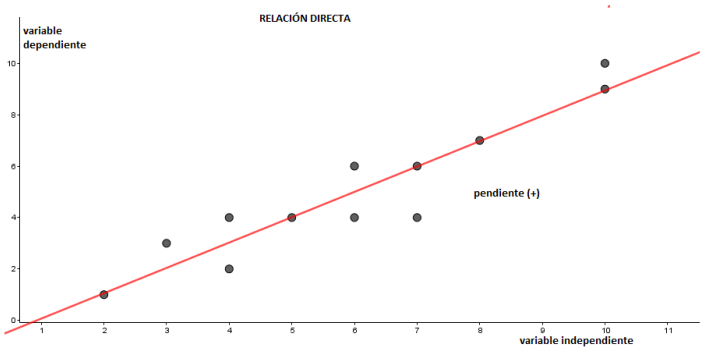
Regresión lineal simple

“Una técnica estadística que establece una ecuación para estimar el valor desconocido de una variable, a partir del valor conocido de otra variable, (en vez de valores de muchas otras variables) se denomina análisis de regresión simple.”

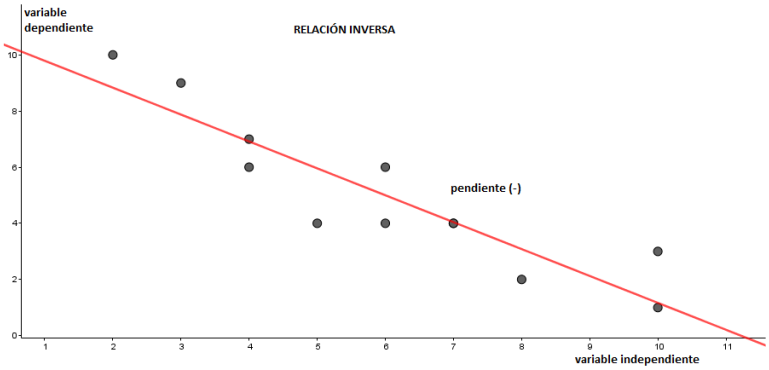
Por lo tanto, el análisis de regresión lineal simple, es el proceso general de predecir una variable (Y) a partir de otra (X).

Las relaciones entre las variables pueden ser directas o también inversas.

- Relación directa: la pendiente de esta línea es positiva, porque la variable Y crece a medida que la variable X también lo hace.



- Relación inversa: La pendiente de esta línea es negativa, porque a medida que aumenta el valor de la variable Y, el valor de la variable X disminuye.



Variable independiente (x)

En el análisis de regresión una variable cuyo valor se suponga conocido y que se utilice para explicar o predecir el valor de otra variable de interés se llama variable independiente; se simboliza con la letra X.

Otros nombres alternativos para la variable independiente (X), son variable explicativa, variable predictora y en ocasiones variable regresora.

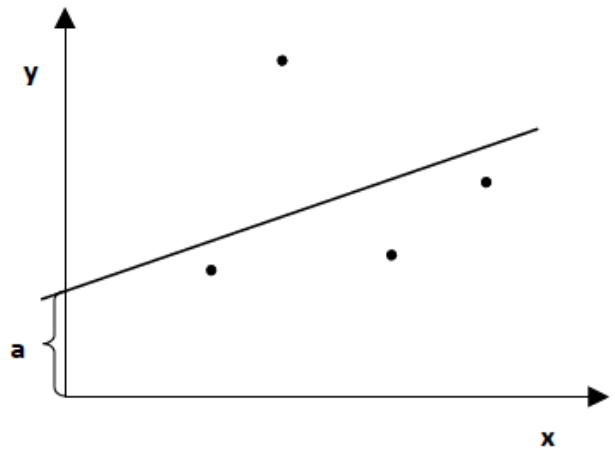
Variable dependiente (y)

En el análisis de regresión una variable cuyo valor se suponga desconocido y que se explique o prediga con ayuda de otra se llama variable dependiente y se simboliza con la letra Y.

La variable dependiente, al igual que la variable independiente es llamada de diferentes maneras algunas de ellas son: variable explicada o variable pronosticada.

Modelo de regresión

La naturaleza de la relación puede adoptar muchas formas, que van desde funciones matemáticas muy sencillas hasta las más complicadas. La relación más simple consiste en una línea recta o relación lineal. Observemos:



El modelo para recta lineal se representa como y = a + b x. Donde:

- a: es la ordenada al origen, o sea, la intersección de la recta con el eje y
- b: es la pendiente de la recta

En este modelo, la pendiente b de la recta representa el cambio en y cuando x cambia una unidad, es decir, representa la cantidad de cambio de y positivo o negativo) para un cambio unitario particular en x.

La intersección de la recta en a con el eje y, representa un factor constante que está incluido en la ecuación. Representa el valor de y cuando x es igual a cero.

Este modelo estadístico es solo una aproximación a la relación exacta entre las dos variables.

Estimación del modelo

Una vez confeccionado el diagrama de dispersión y observado que los puntos tienen una tendencia lineal, se tratará de deducir los parámetros a y b a partir de la distribución de los datos estadísticos de esa distribución de frecuencia conocida.

La técnica que nos permite obtener estos parámetros se denomina AJUSTAMIENTO.

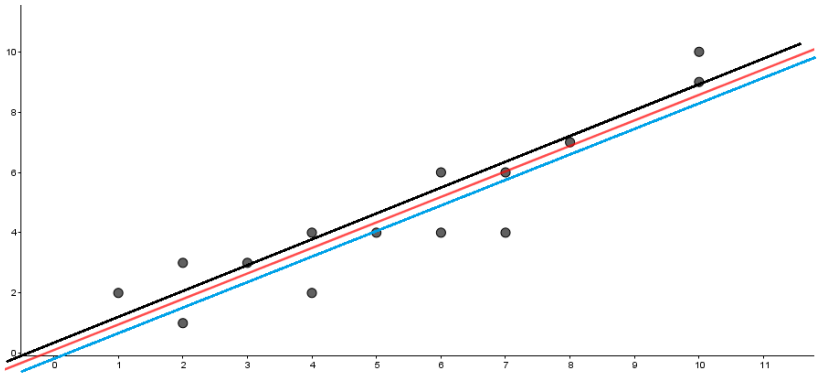
En general, las trayectorias lineales responden a la expresión **a + b x**, es decir que podemos escribir:

$$y_c = a + b\ x$$

y representa a todos y cada uno de los puntos de la recta.

Existen dos métodos para realizar al ajustamiento:

El método libre: en el que, luego de confeccionado el diagrama de dispersión, el observado analiza el comportamiento de los datos para determinar el tipo de curva que mejor ajusta a los mismos. Definido el tipo de curva, si por ejemplo esta es una recta se eligen dos puntos y se la traza.



El inconveniente es que diferentes observadores tendrán diferentes rectas, porque es difícil de que coincidan a la hora de elegir la recta que mejor muestre el comportamiento del conjunto de puntos.

El método de los mínimos cuadrados: es el más preciso, permite obtener la mejor recta de ajuste.

Método de mínimos cuadrados

El método que por lo común se utiliza para ajustar una línea a los datos muestrales indicados en el diagrama de dispersión, se llama método de mínimos cuadrados. La línea se deriva en forma tal que la suma de los cuadrados de las desviaciones verticales entre la línea y los puntos individuales de datos se reduce al mínimo.

El método de mínimos cuadrados sirve para determinar la recta que mejor se ajuste a los datos muestrales, y los supuestos de este método son:

- El error es cero.
- Los datos obtenidos de las muestras son estadísticamente independientes.
- La varianza del error es igual para todos los valores de X.

Una línea de regresión calculada a partir de los datos muestrales, por el método de mínimos cuadrados se llama línea de regresión estimada o línea de regresión muestral.

Este método consiste en hacer mínimo la sumatoria de las distancias al cuadrado, de cada valor observado y el calculado.

Convengamos que entre los valores observados y calculados (y_c o y_0) hay una diferencia o distancia. Tengamos en cuenta que es imposible que las observaciones estuvieran todos alineados. Normalmente eso no se da y debemos buscar aquella recta que deja los menores residuos posibles.

Vamos a dejar las fórmulas simplemente, puesto que para las finalidades de este curso estamos enfocados un poco mas en los procesos analíticos y no tanto así en los cálculos puesto que ello se realiza mediante el software.

La fórmula que nos permite obtener la pendiente de la recta de ajuste es:

$$\boldsymbol{b = \frac{N\sum y_i x_i - \sum y_i \sum x_i}{N\sum x_i^2 - (\sum x_i)^2}} \text{ ó } \boldsymbol{b = \frac{\sum y_i x_i - N\,.\bar{x}\,.\bar{y}}{\sum x_i^2 - N.\bar{x}^2}}$$

Para calcular el valor de a (ordenada al origen), que representa el punto en que la recta corta al eje de las Y, se emplea la siguiente fórmula:

$$\boldsymbol{a = \bar{y} - b\bar{x}}$$

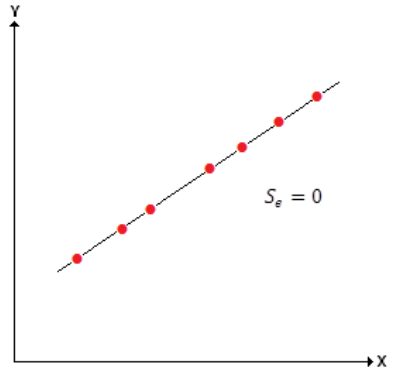
Error estándar de estimación

El siguiente proceso que se necesita en el análisis de la regresión lineal simple es cómo medir la confiabilidad de la ecuación de estimación que hemos desarrollado.

El error estándar de estimación mide la variabilidad o dispersión de los valores observados alrededor de la línea de regresión y se representa como Se. Su fórmula es la siguiente:

$$S_e = \sqrt{\frac{\sum y_i^2 - a.\sum y_i - b.\sum x_i y_i}{n-2}}$$

Cuanto mayor sea el error estándar de la estimación, más grande será la dispersión (o esparcimiento) de puntos alrededor de la línea de regresión. Por el contrario, si Se= 0, se espera que la ecuación de estimación sea un estimador "perfecto" de la variable dependiente, en este caso todos los puntos caerían directamente sobre la línea de regresión y no habría puntos dispersos, como se muestra en la siguiente figura:



El error estándar de estimación tiene la misma aplicación que de la desviación estándar que se vio en los temas anteriores. Esto es, suponiendo que los puntos observados tienen una distribución normal alrededor de la recta de regresión, podemos esperar que:

- 68% de los puntos están dentro de ± 1se
- 95.5% de los puntos están dentro de ± 2se
- 99.7% de los puntos están dentro de ± 3se

El error estándar de la estimación se mide a lo largo del eje "Y", y no perpendicularmente desde la recta de regresión.

Para la realización de un análisis de regresión lineal en InfoStat pueden recurrir al siguiente link. Cabe aclarar, que en la siguiente clase vamos a sumar otras herramientas que completarán el análisis que se lleva adelante en esta sección del curso.

Correlación simple

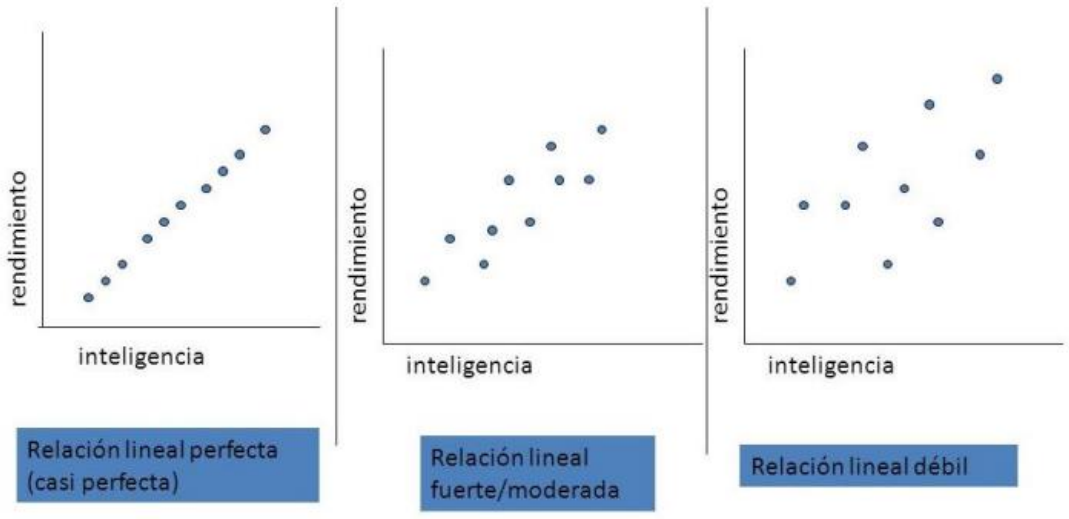
Mientras que el análisis de regresión simple establece una ecuación precisa que enlaza dos variables, el análisis de correlación es la herramienta estadística que podemos usar para describir el grado o fuerza en el que una variable esta linealmente relacionada con otra.

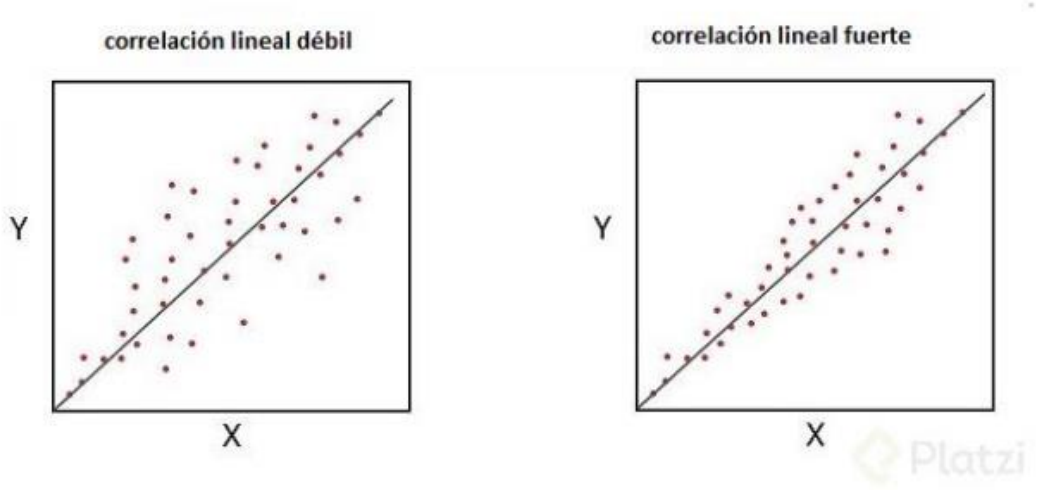
Dependiendo del tamaño de esta medida cuantitativa se puede decir, que tan cercanamente se mueven dos variables y, por lo tanto, con cuanta confiabilidad se puede estimar una variable con ayuda de la otra.

Una técnica estadística que establece un índice que proporciona, en un solo número, una medida de la fuerza de asociación entre dos variables de interés, se llama análisis de correlación simple.

El análisis de correlación es la herramienta estadística de que nos valemos para describir el grado de relación que hay entre dos variables. A menudo el análisis de correlación simple se utiliza junto con el análisis de regresión lineal simple para medir la eficacia con que la línea de regresión explica la variación de la variable dependiente, Y.

Diagramas de dispersión con correlación débil y fuerte.





Existen dos medidas para describir la correlación entre dos variables: el coeficiente de correlación y el coeficiente de determinación.

Coeficiente de correlación

Cuando la correlación es del tipo lineal, el análisis se realiza a través del coeficiente correlación de Pearson. El coeficiente de correlación es una de las medidas con que puede describirse la eficacia con que una variable es explicada por otra, así pues, el signo de r indica la dirección de la relación entre las dos variables X y Y. Se simboliza con la letra r y se calcula de la siguiente manera:

$$r = \frac{N\sum y_i x_i - \sum y_i . \sum x_i}{\sqrt{[N\sum x_i^2 - (\sum x_i)^2].[N\sum y_i^2 - (\sum y_i)^2]}}$$

Si bien, no se realizan los cálculos a mano, y la manera en la que se obtiene usando InfoStat no contempla estos cálculos ponemos a su disposición la forma de hallarlo realizando los cálculos pertinentes.

El campo de variación de dicho coeficiente es de -1 a 1, es decir: $-1 < r < 1$

El siguiente esquema representa adecuadamente la intensidad y la dirección del coeficiente de correlación.

Por lo tanto, el coeficiente de correlación nos indica tres ideas fundamentales:

- La existencia o no de una relación entre las variables estudiadas
- La dirección de esta relación, si es que existe (ver gráfico)
- El grado o la intensidad de esta relación (ver gráfico)

Coeficiente muestral de determinación

La medida más importante de que también ajusta la línea de regresión estimada en los datos muestrales en los que está basada, es el coeficiente de determinación, este es igual a la proporción de la variación total de los valores de la variable dependiente, "Y", que puede explicarse por medio de la asociación de Y con X medida por la línea de regresión estimada. El coeficiente de determinación es la manera primaria de medir el grado, o fuerza, de la relación que existe entre dos variables, X y Y. El coeficiente de determinación se representa como r^2 , y mide exclusivamente la fuerza de una relación lineal entre dos variables. Indica la proporción de la varianza de y que queda explicada por conocimiento de x.

$$r^2 = \frac{variación\ explicada}{variación\ total}$$

El Cálculo del coeficiente de determinación se realiza elevando al cuadrado el coeficiente de correlación. Y el campo de variación es de 0 a 1, es decir: $0 < r^2 < 1$

Ahora bien, en InfoStat, podemos hacer el análisis de correlación entre diferentes variables sin mayores complicaciones, dejamos en el siguiente link un video para que pueda observar como se hace. Por lo general, realizaremos primeramente el análisis de regresión para, posteriormente, analizar aquellos modelos que representen una relación evidente entre variables.

3. Actividad integradora:

La actividad integradora de esta clase consiste estudiar la base de datos de la clase anterior "colesterol en niños" para identificar cuáles son las variables que se puede relacionar y qué tipo de relación existe entre ellas. Además, determinar las rectas de ajuste que relaciones todos los pares posibles de combinaciones entre las variables que se consideran en la base de datos. Con toda esta información, espera que elaboren un informe escrito que presente las principales conclusiones del estudio y las utilidades de los resultados encontrados.

4. Cierre:

Para resumir, en esta clase hemos avanzado sobre las herramientas básicas del Análisis de regresión y correlación, diagramas de dispersión y los errores de estimación. Hemos trabajado cerca de las herramientas teóricas y prácticas que utilizaremos en el desarrollo de esta unidad, lo cual nos permite analizar la relación entre variables.

Es importante, destacar que estudiar la relación entre dos variables para determinar un modelo de ajuste, en muchos casos resulta muy útil, dado que nos permite estimar el valor que podría sumir la variable para un caso determinado con cierto nivel de seguridad (en relación a la bondad del

modelo). Ello lo podemos hacer gracias a estas herramientas que acabamos de estudiar.

Por cualquier consulta, está habilitada la mensajería interna grupal para que podamos compartir y disipar las dudas que puedan surgir entre todos.

Saludos, los y las espero en la próxima clase.

5. Bibliografía:

- Ruiz Diaz, G. (2019). Estadística y Probabilidad. Notas de cátedra. Universidad Nacional de Formosa.
- García, J; López, N; Calvo, J. (2011). Estadísticas Básicas para Estudiantes de Ciencias. Facultad de Ciencias Físicas Universidad Complutense de Madrid. España.
- Wackerly, D; Mendenhall, W; Schaeffer, R. (2010). Estadística Matemática con Aplicaciones. 7ma Ed. Cengage Learning. Santa Fe, México.