Correlación y Regresión Lineal en JASP

Xochitl Alejandra Cárdenas Martínez

Lab 25, Facultad de Psicología, UNAM

Correlación y Regresión Lineal

¿Existe alguna relación entre el IQ y el nivel educativo de una persona? O ¿entre el nivel de estrés y las horas de sueño? Cuando tenemos dos (o más) variables cuantitativas con una distribución conocida podemos compararlas o relacionarlas. Dos variables están relacionadas si los pares de puntuaciones muestran un orden que se puede representar gráficamente con un diagrama de dispersión y numéricamente con un coeficiente de correlación (Witte & Witte, 2016).

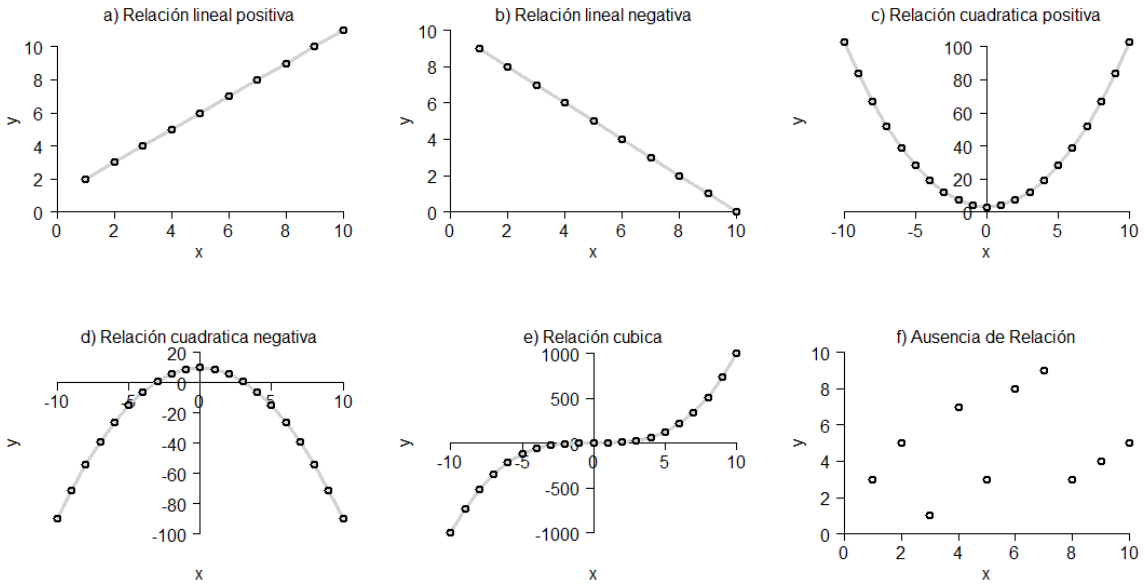
Correlación

Antes de realizar una regresión lineal, es esencial establecer si existe una relación entre las variables trabajadas, asimismo, con dos variables cuantitativas es posible identificar el tipo de relación: lineal, cuadrática, cubica. Para lo cual, es necesario iniciar reconociendo la dispersión de los datos en un **diagrama de dispersión**. Se ubica en un plano en el eje de las abscisas la variable *X* e *Y* en el de ordenadas, y cada par de puntuaciones (Xi, Yi) se representa con un punto.

La forma que se obtenga del conjunto de datos indica el tipo de relación. La figura 1 muestra 6 diagramas de dispersión con 3 posibles tipos. El diagrama a y b presentan relaciones lineales positiva y negativa, respectivamente; el diagrama c y d relaciones cuadráticas, igualmente positiva y negativa; el diagrama e, una relación cubica; y finalmente el diagrama f es un ejemplo de ausencia de relación. En el apéndice A se pueden revisar las funciones que generan estos conjuntos de datos y los datos.

Figura 1

Tipos de relaciones



Nota. Todas las gráficas fueron generadas por las ecuaciones del apéndice A en R Studio.

En esta ocasión nos centraremos en las relaciones lineales, como la que podemos observar en el diagrama a, donde puntuaciones bajas de X corresponden a puntuaciones bajas en Y, lo que indica una relación lineal positiva (como ocurriría con IQ y rendimiento escolar). Por otro lado, el diagrama b presenta una relación lineal negativa, donde puntuaciones bajas en X corresponden a puntuaciones altas en Y (por ejemplo, ansiedad y número de aciertos en un examen).

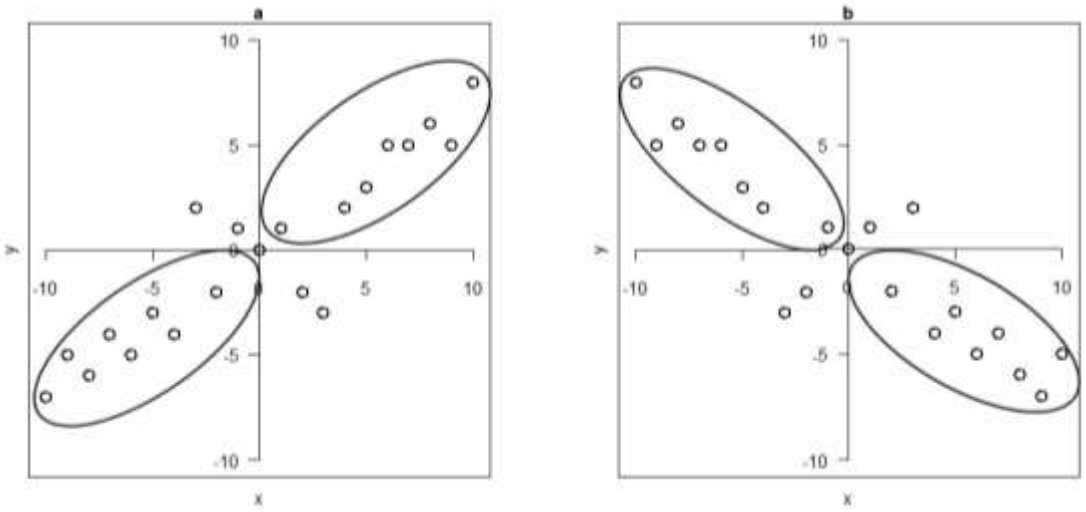
No obstante, para estudiar la relación entre dos variables y cumplir con el objetivo de valorar el impacto de la **variable independiente** sobre la **dependiente** y efectuar pronósticos sobre la dependiente es necesario cuantificar el grado de la relación, para lo cual, hace falta un índice numérico que lo precise. A estos índices se les conoce como **coeficientes de correlación**.

Contamos con diferentes estrategias para obtener estos coeficientes, siendo la primera y más sencilla calculando en el diagrama de dispersión la proporción de puntos

ubicados dentro de los cuadrantes que reflejan una relación lineal. Por ejemplo, en el diagrama a de la figura 2, 17 puntos se ubican dentro de los cuadrantes que reflejan una relación lineal positiva. Estos puntos se expresan en una proporción de 17/20 = 0.85. Y se hace lo mismo para el diagrama b. Esta estrategia arroja valores entre 0.5 (ausencia de relación) y 1 (relación perfecta). Sin embargo, la desventaja es que puede ofrecer valores de 1 a relaciones que no necesariamente son perfectas, y viceversa. Por lo tanto, una cuantificación del grado de relación requiere, no sólo el cuadrante donde se ubican los puntos, además, la ubicación exacta de estos.

Figura 2

Diagramas de dispersión para el cálculo de coeficientes



Esto es posible calculando las **puntuaciones diferenciales** (p y q). Estas representan la distancia a la media:

$$p = P - \bar{P},\ q = Q - \bar{Q} \quad (1)$$

Una puntuación diferencial positiva corresponde a una puntuación directa (P o Q) por arriba de la media; y una negativa a una directa por debajo de la media. Las puntuaciones diferenciales permiten ubicar los puntos en los cuadrantes adecuados y

establecer la distancia con respecto a la media, sin embargo, falla al tratar de expresar el grado de la relación.

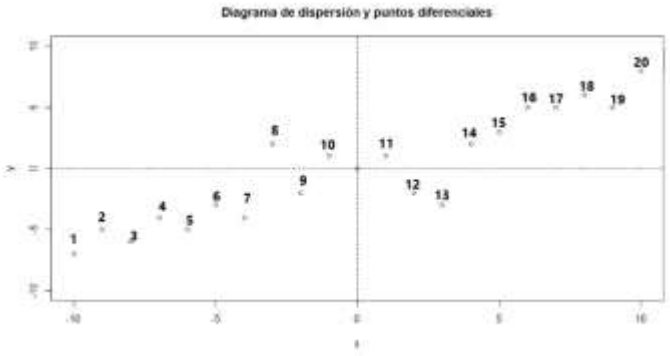
Tabla 1

Cálculo de puntuaciones diferenciales

Sujetos	P	Q	p	q
1	-10	-7	-10	-7.15
2	-9	-5	-9	-5.15
3	-8	-6	-8	-6.15
4	-7	-4	-7	-4.15
5	-6	-5	-6	-5.15
6	-5	-3	-5	-3.15
7	-4	-4	-4	-4.15
8	-3	2	-3	1.85
9	-2	-2	-2	-2.15
10	-1	1	-1	0.85
11	1	1	1	0.85
12	2	-2	2	-2.15
13	3	-3	3	-3.15
14	4	2	4	1.85
15	5	3	5	2.85
16	6	5	6	4.85
17	7	5	7	4.85
18	8	6	8	5.85

Figura 3

Puntuaciones diferenciales (Tabla 1)



19	9	5	9	4.85
20	10	8	10	7.85
Medias	0	-0.15		

Para lo cual, Karl Pearson propone utilizar las puntuaciones diferenciales y multiplicarlas por cada par (de cada sujeto): la relación lineal entre dos variables cuantitativas es tanto más intensa cuanto mayor es la suma de esos productos (en valor absoluto) (Pardo, 2014).

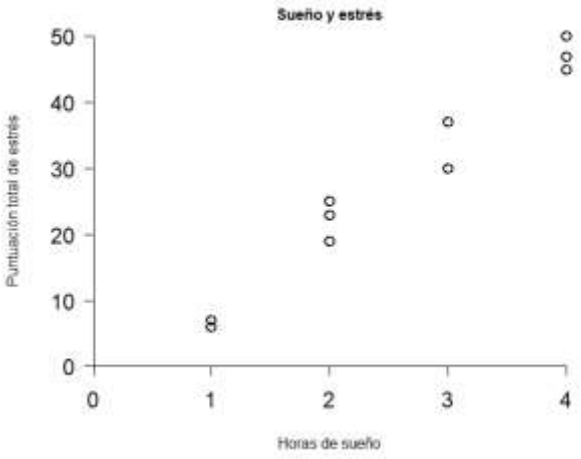
Tabla 2

Cálculo de suma de productos de puntuaciones diferenciales

Sujetos	X(Sueño)	Y(Estrés)	x	y	xy
1	1	6	-1.6	-22.9	36.64
2	1	7	-1.6	-21.9	35.04
3	2	19	-0.6	-9.9	5.94
4	2	25	-0.6	-3.9	2.34
5	2	23	-0.6	-5.9	3.54
6	3	37	0.4	8.1	3.24
7	3	30	0.4	1.1	0.44
8	4	45	1.4	16.1	22.54
9	4	47	1.4	18.1	25.34
10	4	50	1.4	21.1	29.54
Medias	2.6	28.9		Suma	164.6

Figura 4

Puntos sueño y estrés (Tabla 2)



De este modo, el producto de las diferencias expresa el grado de relación entre las variables, siendo una relación lineal perfecta cuando el resultante es el máximo y mínimo cuando hay ausencia de relación. No obstante, es necesario agregar un cálculo que permita incorporar la importancia del tamaño de la muestra con la que se está trabajando, para lo cual, la suma del producto de las diferencias se divide entre el número de puntuaciones sumadas (o el número de sujeto). De esta manera se obtiene un estadístico conocido como **covarianza.**

$$Cov_{XY} = S_{XY} = \frac{\sum_i x_i y_i}{n-1} \quad (2)$$

Aplicando la ecuación a los datos de la Tabla 2:

$$S_{XY} = \frac{164.6}{9} = 18.28$$

El grado de la relación es mayor cuando el valor de la covarianza es igualmente mayor, y el signo indica el sentido (positivo o negativo) de la relación. Sin embargo, este

cálculo no está exento de defectos. Sabemos que el valor mínimo de la covarianza es cero, indicando ausencia de relación, pero, no tiene un límite superior y depende del grado de dispersión de las variables. Por ejemplo, si tenemos una covarianza $S_{XY} = 10.5$ o $S_{XV} = 5.6$, podemos identificar que S_{XY} es mayor que S_{XV}, pero no podemos asegurar que la relación de la primera es alta. El valor máximo depende de las variables estudiadas y de la muestra utilizada. Esto complica su interpretación y la comparación entre diferentes muestras.

La solución a este problema es relativizar la covarianza[1] tomando como referencia su valor máximo. A este nuevo estadístico se le llama **coeficiente de correlación de Pearson.**

$$R_{XY} = \frac{S_{XY}}{S_X S_Y} \qquad o \tag{3}$$

$$R_{XY} = \frac{\sum_i Z_{x_i} Z_{y_i}}{n-1} \tag{4}$$

Esta ecuación se interpreta como el grado en que la covarianza alcanza su máximo, o, el equivalente a calcular la covarianza a partir de una **puntuación Z**[2]. Que es igualmente equivalente a la siguiente formula, cuya aplicación resulta más cómoda para cálculos a mano:

$$R_{XY} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{n\sum X^2{}_i - (\sum X_i)^2}\sqrt{n\sum Y^2{}_i - (\sum Y_i)^2}} \tag{5}$$

[1]Para obtener el denominador en la ecuación (1) se calcula la desviación estándar de las variables *X* e *Y*. Es decir:

$$S_X = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}} \tag{3.1}$$

que es equivalente a

$$S_X = \sqrt{\frac{\sum(x_i)^2}{n-1}} \tag{3.2}$$

siendo x_i la puntuación diferencial de variable X.

[2]Para calcular la puntuación Z de una natural, se utiliza la siguiente ecuación:

$$Z = \frac{X - \bar{X}}{S_X} \tag{4.1}$$

Tabla 3

Datos incluyendo puntuaciones naturales al cuadrado para ecuación (5)

Sujetos	X(Sueño)	Y(Estrés)	XY	X^2	Y^2
1	1	6	6	1	36
2	1	7	7	1	49
3	2	19	38	4	361
4	2	25	50	4	625
5	2	23	46	4	529
6	3	37	111	9	1369
7	3	30	90	9	900
8	4	45	180	16	2025
9	4	47	188	16	2209
10	4	50	200	16	2500
SUMATORIAS	26	289	916	80	10603
PRODUCTOS	ΣXΣY=	7514			
CUADRADOS	676	83521			

Aplicando la ecuación (5) a los datos de la tabla 3:

$$R_{XY} = \frac{10\,(916)-7514}{\sqrt{10(80)-676}\sqrt{10(10603)-83521}} = \frac{1646}{1669.72} = 0.985$$

Entonces, el coeficiente de correlación de Pearson mide el grado de relación lineal, sólo, lineal. El signo indica la naturaleza (positiva o negativa) de esta, y el valor oscila entre -1 y 1, siendo estos los valores que indicarían una relación perfecta. La siguiente tabla

indica una forma rápida de interpretar el grado de relación a partir del valor obtenido por el coeficiente R_{XY}:

Tabla 4

Interpretación del coeficiente R_{XY}

Coeficiente de correlación de Pearson	
Valor	Interpretar como:
0	Relación nula
0 – 0.2	Relación muy baja
0.2 – 0.4	Relación baja
0.4 – 0.6	Relación moderada
0.6 – 0.8	Relación alta
0.8 - 1	Relación muy alta
1	Relación perfecta

No obstante, es necesario asegurar la fiabilidad del resultado del coeficiente y establecer que el valor muestral es mayor o parecido al que se esperaría por puro azar. Esto puede hacerse al probar la hipótesis nula de independencia lineal (H_0: $\rho_{XY} = 0$), dado que al rechazar la hipótesis es posible concluir que las variables X e Y no son linealmente independientes y existe un grado de relación. Este proceso se puede llevar a cabo al comparar el valor R obtenido contra un valor R que nos proporcionan tablas de valores críticos[3] para lo grados de libertad especificados por la misma y el nivel alfa (α) que escojamos (puede ir desde 0.10 hasta 0.001 e indica la probabilidad de rechazar la hipótesis nula cuando es verdadera).

[3]La tabla de valores críticos para coeficiente correlación de Pearson se encuentra en el apéndice B.

El primer paso es establecer las hipótesis básicas para la regla de decisión. Estas son: H_o o hipótesis nula que asume ausencia de relación; y H_1 o hipótesis de trabajo o alterna que indica relación entre las variables. Posteriormente, debemos establecer las condiciones para la regla de decisión, es decir, determinar cuándo elegir una hipótesis u otra.

Si $|r|$ > que el valor critico de R se rechaza H_o concluyendo relación lineal.

Si $|r| \leq$ que el valor critico de R se rechaza H_1 concluyendo ausencia de relación.

Ahora podemos pasar a hacer la comparación. Por ejemplo, retomando el ejemplo anterior, $R_{XY\ obt} = 0.985$, identificamos en la tabla de valores críticos para R de Pearson $R_{XY\ crit} = 0.63$ con $\alpha = 0.05$ y podemos apreciar que el valor obtenido es mayor que el valor crítico, por lo tanto, podemos concluir y asegurar que existe relación lineal entre las variables estudiadas.

En conclusión, para identificar el nivel de relación entre dos variables podemos hacer uso de un estadístico que cuantifica el grado de relación, y este puede interpretarse de acuerdo con 3 puntos: fuerza (valor arrojado por el coeficiente de correlación de Pearson), dirección (signo del coeficiente) y significancia (prueba de hipótesis nula). Un coeficiente de correlación alto no necesariamente implica causalidad. Dos variables pueden estar relacionadas sin que una cause la otra. Cuando existe efecto de causa, un cambio en una variable provoca un cambio en la otra, situación que no es necesariamente cierta cuando sólo se presenta relación. Confundir estos términos puede llevar a sentar relaciones ficticias, a estas se les conoce como relaciones espurias.

Análisis de Correlación en JASP

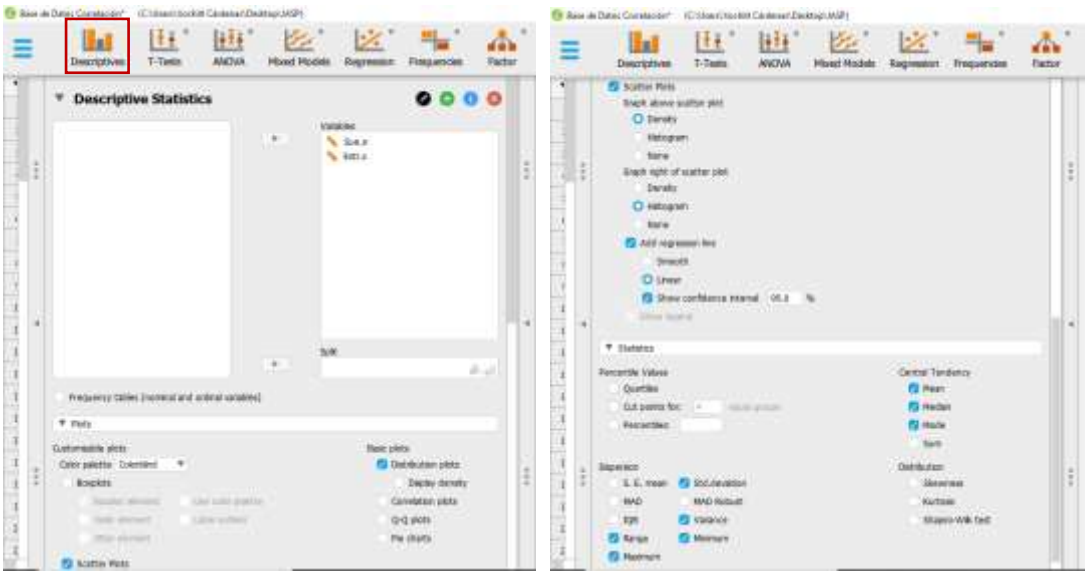
En mi opinión, aprender a realizar cualquier prueba estadística a mano y aplicando cada ecuación a lápiz y papel permite entender mejor la lógica de estas pruebas y facilita el entendimiento del resultado. Sin embargo, cuando tenemos bases de datos muy grandes, hacer estas pruebas a mano sería algo muy difícil y tardado, por no decir imposible. Para este tipo de situaciones (que suelen ser la mayoría) existen programas de análisis estadístico que generan el valor estadístico sólo con seleccionar algunos apartados. Uno de estos programas es JASP, un programa que permite realizar análisis frecuentistas y bayesianos. A continuación, revisaremos cómo podemos hacer un análisis de correlación usando este programa.

Análisis de correlación clásica (frecuentista).

Cada vez que buscamos analizar una base de datos, el primer paso siempre debe ser realizar análisis descriptivos para darnos una idea de la posición, centralización, dispersión y forma de los datos, así como identificar datos faltantes. En JASP podemos realizarlos desde ‘descriptives’. En la figura 5 se muestran todas las medidas que podemos ejecutar.

Figura 5

Estadísticas descriptivas en JASP



Nota. Captura de pantalla de JASP que muestra ventana de análisis descriptivos de variables del archivo Base de datos Correlación.csv

Para este ejercicio, usamos una base de datos (que se puede encontrar en mi perfil de Github https://github.com/xochitlcardenas) la cual contiene datos sobre una escala de sueño y una de estrés. Nuestro objetivo es identificar si estas dos variables (Sueño y Estrés) están relacionadas. Basta con enviar ambas variables a la caja 'Variables' y con las medidas seleccionadas que se observan en la figura 5 obtenemos los siguientes resultados:

Tabla 5

Estadística descriptiva

		Sueño	**Estrés**
Valid		100	100
Missing		0	0
Mean		50.540	52.170
Median		49.000	53.500
Mode	[a]	5.000	50.000
Std. Deviation		29.182	26.737
Variance		851.604	714.870
Range		98.000	98.000
Minimum		1.000	1.000
Maximum		99.000	99.000

[a] More than one mode exists, only the first is reported

Figura 6

Distribution Plots

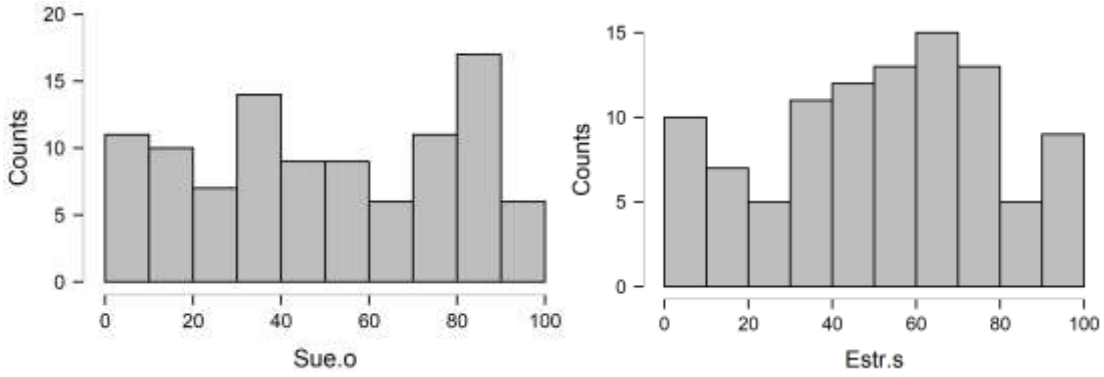
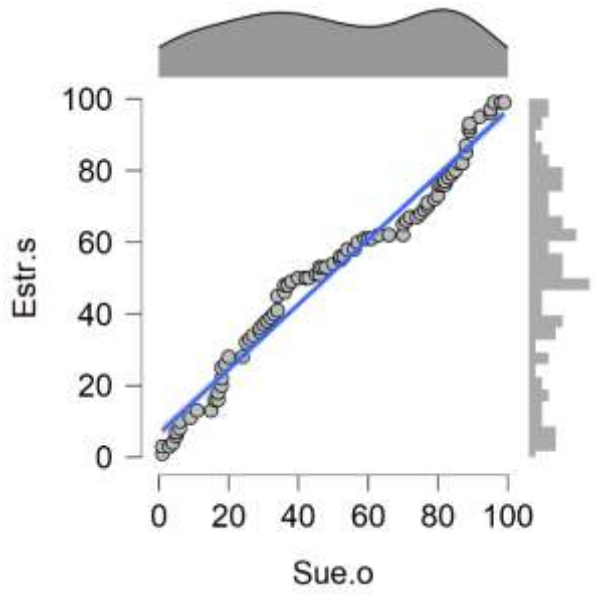


Figura 7

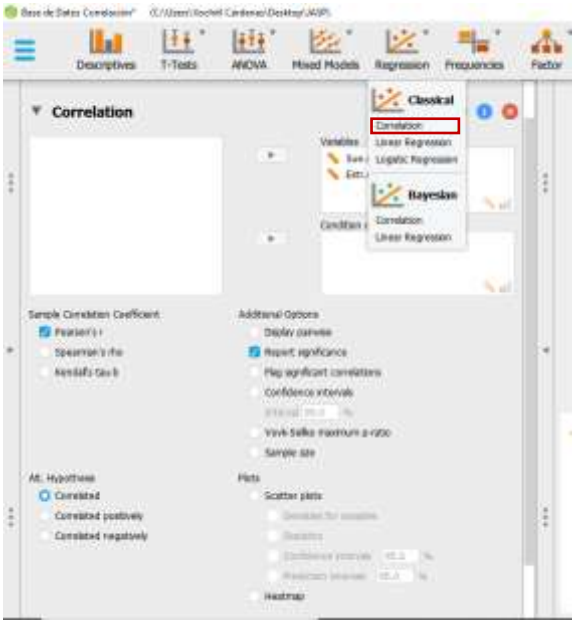
Scatter Plots



Estos datos tienen medias similares, y visualizando la gráfica de la figura 7, podemos observar que los datos trabajados reflejan una relación positiva. Sin embargo, como ya hemos aprendido, esta descripción gráfica no es suficiente para asegurar la relación y necesitamos comprobarlo con un estadístico. Esto es igualmente posible usando JASP, desde Regression > Classical > Correlation. Eligiendo estos apartados JASP nos muestra la ventana que se observa en la figura 8. Para este análisis sólo debemos colocar ambas variables involucradas en la caja ‘Variables’.

Figura 8

Correlación frecuentista en JASP



Nota. Captura de pantalla de JASP que muestra ventana de análisis de correlación de variables del archivo Base de datos Correlación.csv

Con las casillas que se muestran activadas en la figura 8 JASP arroja los siguientes resultados:

Tabla 6

Correlación de Pearson

Variable		Sueño	Estrés
1. Sueño	Pearson's r	—	
	p-value	—	
2. Estrés	Pearson's r	0.985	—
	p-value	< .001	—

El análisis evalúa la hipótesis nula (H0) que asume ausencia de asociación entre las dos variables. Estos resultados (además de confirmar que nuestro procedimiento a mano es correcto) se puede interpretar de la misma manera que el procedimiento manual (ver tabla 4). Y el valor p es equivalente a la significancia. Si $p < 0.05$, la relación es significativa, y podemos afirmar que las variables están relacionadas.

Análisis de correlación bayesiana.

Por otro lado, como se mencionó en párrafos anteriores, con JASP también podemos realizar estas pruebas estadísticas por factor de Bayes. Y no es diferente para una correlación. El procedimiento es el mismo que con la correlación clásica, primero debemos hacer un análisis descriptivo para conocer los datos que trabajamos y posteriormente, pasamos a Regression > Bayesian > Correlation.

Figura 9

Correlación bayesiana en JASP

Las casillas que se muestran activadas en la figura 9 bastan para un análisis de correlación bayesiano básico, si requerimos más datos podemos activar las casillas correspondientes, pero los siguientes resultados son suficientes:

Tabla 7

Bayesian Pearson Correlations

Variable		Sue.o	Estr.s
1. Sue.o	Pearson's r	—	
	BF_{10}	—	
2. Estr.s	Pearson's r	0.985 ***	—
	BF_{10}	8.199e +72	—

* $BF_{10} > 10$, ** $BF_{10} > 30$, *** $BF_{10} > 100$

Como podemos observar, el análisis clásico y el bayesiano reportan el coeficiente de correlación de Pearson, el cual se interpreta de la misma manera que en el análisis clásico (ver tabla 4). Y el factor de Bayes reporta mayor evidencia en favor de la hipótesis alterna (existencia de correlación) en tanto mayor sea su valor, como es para el caso de nuestro ejemplo. Finalmente, estos resultados pueden reportarse de la siguiente manera:

Usando una hipótesis alternativa unilateral, hubo una correlación positiva para el nivel de sueño con respecto al nivel de estrés (r = 0.985), esto fue acompañado por un factor de Bayes BF10 = 8.199e+72 que indica una probabilidad decisiva ("evidencia") de que esto ocurra bajo el H1 que H0.

Regresión lineal

Una correlación y una regresión involucran la relación entre dos variables, utilizando el mismo conjunto de datos (de los mismos sujetos o pareados). La diferencia es que la regresión utiliza la relación para hacer predicciones acerca de las variables. Se tiene una variable cuantitativa X, que denominaremos independiente, a partir de la cual es posible estudiar el comportamiento de una segunda variable Y (dependiente), es decir, buscamos estudiar el cambio de una por el cambio de la otra y hacer predicciones acerca de este.

Generar una predicción es fácil cuando la relación es perfecta, que sería el caso cuando todo el conjunto de datos cae sobre una misma línea recta, y sólo se necesita una ecuación de esta línea que permita generar las predicciones. Pero, difícilmente se encuentra una relación perfecta en el mundo real.

El primer paso que debemos llevar a cabo para hacer predicciones es formarnos una idea acerca del conjunto de datos con los que estamos trabajando. La manera más directa es mediante un diagrama de dispersión, donde la ubicación de cada punto (par de datos) indica el tipo de relación subyacente. Un ejemplo de esto se muestra en la figura 10 que corresponde a los puntos de los datos de la tabla 8.

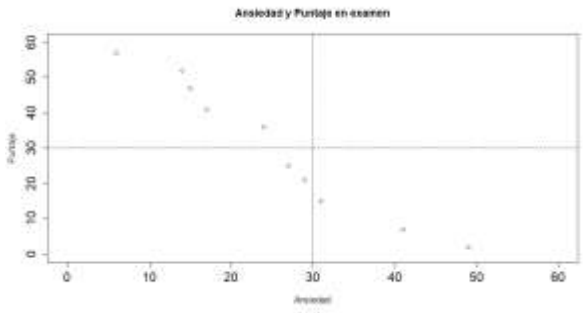
Tabla 8

Datos Ansiedad y Puntaje

Sujetos	X(Ansiedad)	Y(Puntaje)
1	49	2
2	41	7
3	31	15

Figura 10

Diagrama de dispersión para datos de Tabla 7



4	29	21
5	27	25
6	24	36
7	17	41
8	15	47
9	14	52
10	6	57

A primera vista podríamos decir que existe una relación lineal negativa entre las variables ‘ansiedad’ y ‘puntaje’. Sin embargo, no podemos cuantificar con precisión el grado o intensidad de la relación, ni podrías entender el cambio de una a partir del cambio de la otra.

Para describir una variable cuantitativa podemos hacer uso de tres propiedades de distribución: a) forma: determinando si la nube de puntos refleja una relación lineal, b) centro: resumir los puntos del diagrama en una recta y c) dispersión: valorar el grado de concentración o alejamiento de los puntos a partir de la recta. Para determinar si existe una relación entre las variables podemos valernos del coeficiente de correlación de Pearson (revisado en el apartado anterior), por lo tanto, en este apartado abordaremos las dos propiedades restantes.

Recta de regresión

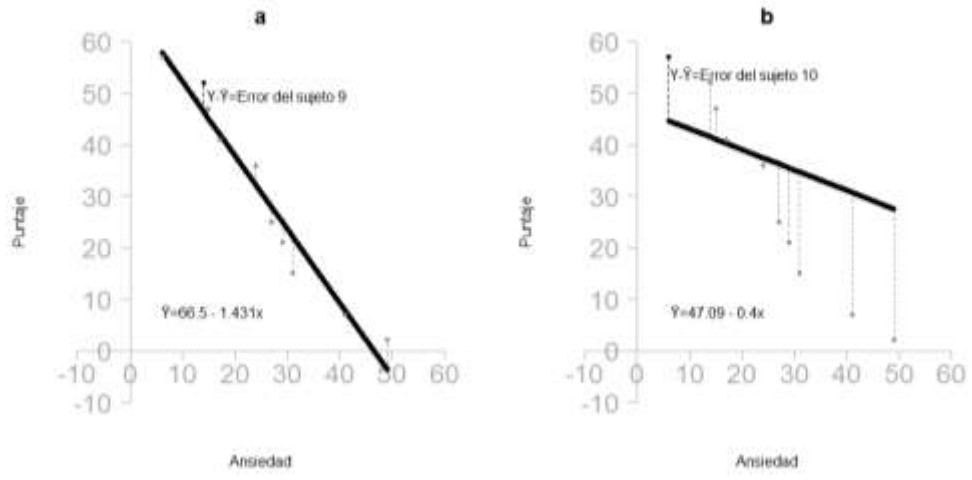
Una vez que identificamos la existencia de una relación lineal podemos pasar a establecer la recta que resume los puntos que representan los datos obtenidos de las variables estudiadas. La solución más utilizada es construir una línea que minimiza los

errores de predicción de acuerdo con un criterio de mínimos cuadrados. A recta resultante de esta estrategia se denomina **línea de regresión de mínimos cuadrados**.

En la figura 11.a podemos visualizar la línea de regresión para los datos de la tabla 8, donde la línea vertical punteada entre la recta y el punto (dato del sujeto 10) representa el error de predicción, siendo Y' = valor predicho de Y, y Y = valor real, por tanto, $Y - Y'$ es igual al error para cada punto. En primera instancia podríamos asumir que el error total de predicción sería la suma de cada una de las diferencias de los puntos trabajados, y podríamos construir la línea que minimice esta sumatoria. Sin embargo, por la naturaleza algebraica de estas operaciones, esta podría resultar en un total igual a cero dado que tendríamos cantidades negativas y positivas que se cancelarían unas a otras. Situación similar cuando se trabaja con el promedio. Por lo tanto, la solución inmediata para evitar un posible resultado de cero es elevar al cuadrado cada diferencia (Y-Y´), y podemos pasar a hacer la sumatoria total de estos cuadrados. Ahora, si minimizamos $\Sigma(Y-Y')^2$, podemos minimizar el error total de predicción, y sólo existe una línea de regresión que lo minimiza para cada relación de dos variables.

Figura 11

Líneas de regresión y predicción de error



Entonces, sabemos que el método ideal para encontrar la línea de regresión que mejor describa la relación es por mínimos cuadrados, sin embargo, surge la interrogante ¿cómo construimos la línea de regresión?

La ecuación de la recta de regresión de mínimos cuadrados para predecir Y dado X es:

$$Y' = \alpha + \beta X \quad (6)$$

Donde Y' = *valor estimado de Y*

β = *pendiente de la línea que minimiza el error de predicción*

α = *intercepto en el eje Y que minimiza el error de predicción*

α y β son **constantes de regresión** en la ecuación (6). Al igual que en el caso del coeficiente de correlación de Pearson, es posible encontrar los valores de estas constantes haciendo cálculos a mano.

Iniciando con el valor que corresponde al intercepto:

$$\alpha = \bar{Y} - \beta\bar{X} \quad (7)$$

Como bien podemos observar, para la constante α es necesario calcular la constante β, que se obtiene con la siguiente ecuación:

$$\beta = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{SS_X} \quad (8)$$

Donde SS_x = *suma de los cuadrados de X* = $\sum X^2 - \frac{(\sum X)^2}{N}$

N = número de puntos

$\sum XY$ = *suma de los productos de X y Y*

Utilizando los datos de la tabla 8, calcularemos el valor de β y α, para proceder a sustituirlos en la ecuación (6) y poder predecir el número de aciertos que alcanzaría un sujeto 11 con *X* valor en la escala de Ansiedad.

Personalmente, considero que el primer paso es calcular los cuadrados y productos que requiere la ecuación (8), especialmente para cálculos a mano, lo cual se puede visualizar en la tabla 9.

Tabla 9

Datos tabla 8 incluyendo cuadrados y sumatorias

Sujetos	Ansiedad (X)	Puntaje (Y)	X^2	XY
1	49	2	2401	98
2	41	7	1681	287
3	31	15	961	465
4	29	21	841	609
5	27	25	729	675
6	24	36	576	864
7	17	41	289	697
8	15	47	225	705
9	14	52	196	728
10	6	57	36	342
Σ	253	303	7935	5470
$(\Sigma X)^2$	64009			
μ	25.3	30.3		

Aplicando la ecuación (8):

$$\beta = \frac{5470 - \frac{(253)(303)}{10}}{7935 - \frac{(253)^2}{10}} = \frac{-2195.9}{1534.1} = -1.431$$

Sustituyendo el valor de β y las medias de las variables en la ecuación (7):

$\alpha = 30.3 - (-1.431)(25.3) = 66.50$

Sustituyendo β y α en la ecuación (6) que a su vez es la ecuación que genera la línea de regresión de la figura 6.a):

$Y' = 66.50 + (-1.431)X$

Finalmente, si quisiéramos conocer el número de acierto que obtendría un alumno cuya puntuación en la escala de ansiedad fue de 20:

$$Y' = 66.50 + (-1.431)20$$
$$= 66.50 - 28.62$$
$$= 37.88$$

Es importante puntualizar que sólo podemos hacer predicciones con base en valores de X que se encuentren dentro del rango de la variable con la que construimos la recta inicialmente.

Error de estimación estándar

Retomando, la línea de regresión es el mejor predictor de la variable *Y*, sin embargo, a menos de que la relación sea perfecta (evento que difícilmente se hace presente), muchos de los puntos de *Y* caerán fuera de esta línea (errores de predicción), por lo cual, es necesario conocer la magnitud de estos errores. Si este error es grande, confiaríamos poco en la predicción; si es pequeño, podemos fiarnos y tomar decisiones con base en ella.

Para cuantificar el error podemos hacerlo al computar el error estándar de estimación que nos da una medida de la desviación promedio de los errores de predicción sobre la línea de regresión, es decir, es parecido a la desviación estándar. La ecuación para el error estándar de estimación para predecir *Y* dado *X* es:

$$S_{Y|X} = \sqrt{\frac{\sum(Y-Y')^2}{N-2}} \qquad (9)$$

Nuevamente, podemos aplicar esta ecuación a los datos de la tabla 9 que se han transcrito a la siguiente tabla agregando la diferencia al cuadrado del valor de *Y*, y el valor predicho (*Y´):*

Tabla 10

Datos ansiedad y puntaje, incluyendo diferencia al cuadrado (Y-Y´) ²

Sujetos	X(Ansiedad)	Y(Puntaje)	Y´	(Y-Y´)	(Y-Y´) ²
1	49	2	-3.619	5.619	31.57
2	41	7	7.829	-0.829	0.68
3	31	15	22.139	-7.139	50.96
4	29	21	25.001	-4.001	16
5	27	25	27.863	-2.863	8.19
6	24	36	32.156	3.844	14.77
7	17	41	42.173	-1.173	1.37
8	15	47	45.035	1.965	3.86
9	14	52	46.466	5.534	30.62
10	6	57	57.914	-0.914	0.83
				∑	158.90

Sustituyendo en la ecuación (9):

$$S_{Y|X} = \sqrt{\frac{158.90}{8}} = \sqrt{19.86} = 4.45$$

Si bien calcular *Y´* para cada valor *X* puede resultar laborioso, existe una alternativa que nos permite llegar al mismo resultado con datos que obtendríamos de calcular el coeficiente de

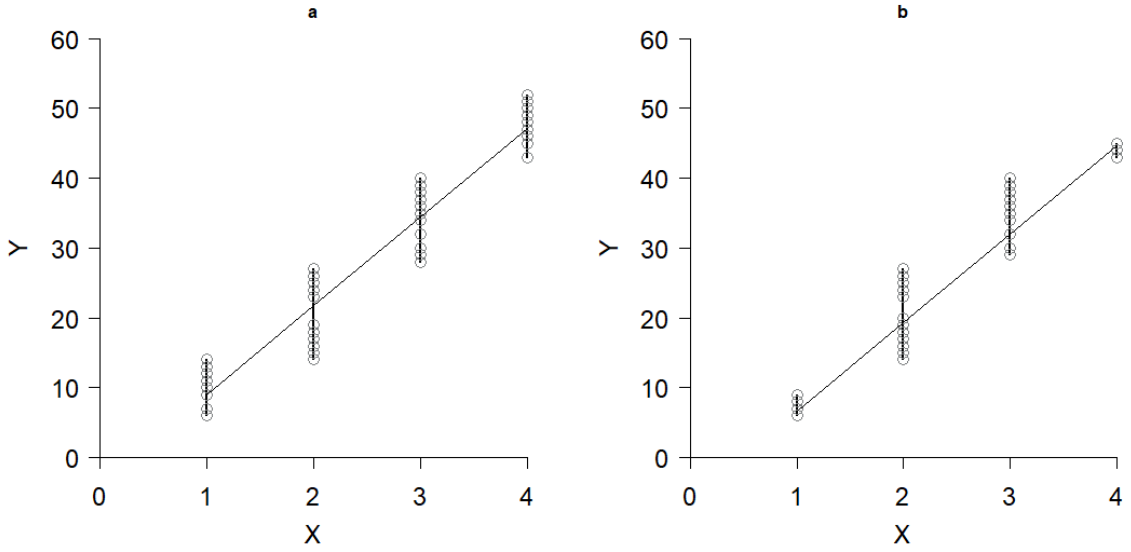
Pearson y la línea de regresión. Podemos definir esta nueva ecuación como ecuación **computacional** de error estándar de estimación al predecir *Y* dado *X*:

$$S_{Y|X} = \sqrt{\frac{SS_Y - \frac{[\sum XY - (\sum X)(\sum Y)/N]^2}{SS_X}}{N-2}} \quad (10)$$

Entonces, el error estándar para los datos de Ansiedad y Puntaje es $S_{Y|X} = 4.45$. Esta medida se calcula para todos los valores de *Y*, por lo que podemos asumir que la variabilidad de *Y* se mantiene constante a lo largo de todos los valores de *X*. A esta suposición se le conoce como **homocedasticidad**. En la figura 12.a podemos observar que se cumple esta suposición, situación contraria para la figura 12.b. El supuesto de homocedasticidad implica que, si dividiéramos las puntuaciones *X* en columnas, la variabilidad de *Y* no cambiaría de una columna a otra.

Figura 12

Homocedasticidad



Entonces, el error estándar de estimación nos permite cuantificar el error de las predicciones, y este es directamente proporcional a la variabilidad de los datos, e inversamente proporcional al tamaño de la muestra. Y ¿cómo se interpreta este valor?

Primero debemos calcular el intervalo de confianza. Podemos suponer que los puntos se distribuyen de manera normal en la line de regresión. Si esta suposición es cierta y válida, al construir dos líneas paralelas a la línea de regresión a las distancias $\pm 1_{S_{Y|X}}$, $\pm 2_{S_{Y|X}}$ y $\pm 3_{S_{Y|X}}$, encontraríamos que el 68% de los puntajes caen entre las líneas $\pm 1_{S_{Y|X}}$, el 95% en $\pm 2_{S_{Y|X}}$, y 99% en $\pm 3_{S_{Y|X}}$. Estos porcentajes representan la confianza sobre mi medición y el valor sobre el cual se va a calcular el intervalo de confianza.

Utilizando el error estándar de nuestro ejemplo y un porcentaje del 95%:

$$IC\ 95\%\ \bar{X} = \bar{X} \pm 2S_{Y|X} \tag{11}$$

$$IC\ 95\%\ \bar{X} = 25.3 \pm 2(4.45) = 16.4\ -\ 34.2$$

Y este intervalo se interpreta como que el valor medio de la variable de interés debe estar en un rango de 16.4 a 34.2 con una confianza del 95%, o, existe un 5% de probabilidad de que el valor medio de la variable se encuentra por abajo o por arriba de este intervalo. Y puede hacerse el mismo calculo para el resto de los porcentajes presentados.

Regresión lineal múltiple

Hasta el momento hemos hablado sobre relaciones que únicamente involucran dos variables cuantitativas, sin embargo, existen muchas otras que no se limitan a esta cantidad. Por ejemplo, cuando hablamos de puntaje en un examen y ansiedad, pueden existir muchas otras variables que afecten la calificación en la prueba, como la motivación, estímulos distractores, horas de estudio, entre otras. Al incluir otras variables predictoras importantes, la predicción sobre el puntaje puede ser más precisa.

La regresión múltiple es una extensión de la regresión simple, y esto lleva a intuir que la ecuación simple debe sufrir alguna modificación para incluir una segunda variable. La forma general de la ecuación de regresión múltiple para dos variables predictoras es:

$$Y' = \beta_1 X_1 + \beta_2 X_2 + \alpha \quad (12)$$

Donde *Y´ = valor predicho de Y*

β_1 *= coeficiente de la primera variable predictora*

X_1 *= primera variable predictora*

β_2 *= coeficiente de la segunda variable predictora*

X_2 *= segunda variable predictora*

α *= constante predictora*

Esta ecuación sólo incluye un coeficiente para la para la variable X_2, por lo tanto, estos igualmente se calculan a partir del criterio de mínimos cuadrados.

Retomando, la precisión en la predicción puede incrementar al agregar variables, pero, caso similar ocurre para la variabilidad de *Y*. Por lo cual, también es importante computar la proporción de varianza, lo cual puede realizarse con la siguiente ecuación:

$$R^2 = \frac{r_{YX_1}{}^2 + r_{YX_2}{}^2 - 2 r_{YX_1} r_{YX_2} r_{X_1X_2}}{1 - r_{X_1X_2}{}^2} \quad (13)$$

Donde *R²* = *coeficiente múltiple de determinación*

r_{YX_1} = *correlación entre Y y la variable predictora* X_1

r_{YX_2} = *correlación entre Y y la variable predictora* X_2

$r_{X_1X_2}$ = *correlación entre las variables predictoras* X_1 *y* X_2

Finalmente, es importante puntualizar que agregar una variable no es un aumento inmediato en la precisión, al igual que la regresión simple, el aumento en la precisión de la predicción y la cantidad de varianza considerada depende de la fuerza de la relación entre la variable

que se predice y la variable predictora adicional y también en la fuerza de la relación entre las variables predictoras mismas.

En conclusión, la línea de regresión permite predecir valores de una variable *Y* dada una variable *X* que se relacionan de manera lineal, si no es así, la predicción será poco confiable y ausente de exactitud. Dicho lo anterior, podemos resumir que para llevar a cabo los análisis y aplicar los procedimientos anteriores, las variables estudiadas debe cumplir con ciertos requisitos, siendo estos los siguientes: 1) la relación entre las variables debe ser lineal, 2) si queremos hacer predicciones sobre un grupo diferente al que fue utilizado para calcular la línea de regresión, el grupo original debe ser representativo del grupo a predecir, 3) la línea de regresión es adecuada únicamente para el rango de la variable en la que está basada, hacer predicciones fuera de este rango resultaría inexacto.

Regresión lineal en JASP.

Al igual que para la correlación, es posible realizar un estadístico de regresión lineal en JASP, frecuentista o bayesiano.

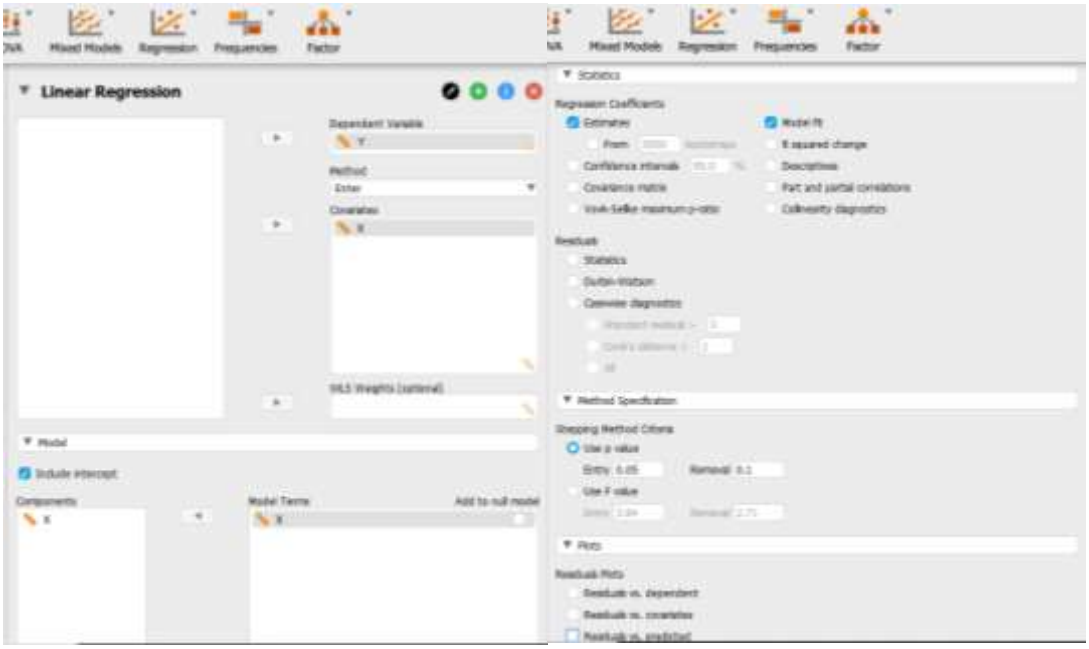
Nuevamente es recomendable iniciar con análisis descriptivos para darnos una idea de los datos con los que estamos trabajando.

Para este ejercicio, usamos una base de datos (que se puede encontrar en mi perfil de Github https://github.com/xochitlcardenas) la cual contiene datos sobre un estudio que mide los niveles de dos enzimas. Nuestro objetivo es identificar si el nivel de la enzima *A* puede predecir el nivel de la enzima *B*. Podemos pasar a realizar el análisis en Regression > Linear Regression.

La figura 13 muestra el panel ‘Linear Regression’, donde podemos indicar a JASP qué estadísticos o gráficas queremos realizar y bajo qué parámetros.

Figura 13

Panel de regresión lineal en JASP



Nota. Captura de pantalla de JASP que muestra ventana de regresión lineal.

Debemos mover la variable dependiente (Y o enzima B para el ejemplo) a la caja ‘Dependent variable’ y la independiente (X o variable A) a la caja ‘Covariantes’. JASP mostrara los resultados inmediatamente y en caso de activar casillas adicionales, estos cambios se visualizan con la misma rapidez.

Para un análisis de regresión lineal básico, es suficiente con tener las casillas activadas que se muestran en la figura anterior, y esta configuración arroja las siguientes tablas de resultados.

Tabla 11

Model Summary - Y

Model	R	R^2	Adjusted R^2	RMSE
H_0	0.000	0.000	0.000	26.768
H_1	0.991	0.982	0.982	3.618

La tabla 10 ‘Resumen del Modelo’ nos proporciona la información necesaria para determinar qué tan bien se ajusta el modelo de regresión a los datos. Debemos enforcanos en la fila que corresponde al modelo H1, donde R es el coeficiente de correlación múltiple, y dado que para una regresión lineal simple sólo tenemos una variable independiente, este coeficiente es simplemente el valor absoluto de una correlación de Pearson, que indica la fuerza de la correlación entre las variables estudiadas. Para el ejemplo de las enzimas, con un valor de R = 0.991 podemos afirmar que la correlación entre los niveles estudiados de las enzimas es muy alta.

R^2 es el valor que representa la proporción de varianza en la variable dependiente que puede ser explicada por la independiente. Para nuestro ejemplo, la enzima A explica 98.2 % de la variabilidad de la variable dependiente (enzima B). R^2 se basa en la muestra y es una estimación sesgada de la proporción de la varianza de la variable dependiente explicada por el modelo de regresión.

R^2 ajustada que corrige el sesgo de la R^2, ofreciendo un valor que sería el esperado en la población (grupo del cuál fue tomada la muestra).

Tabla 12

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
H_1	Regression	69653.248	1	69653.248	5321.617	< .001
	Residual	1282.696	98	13.089		
	Total	70935.944	99			

Note. The intercept model is omitted, as no meaningful information can be shown.

La segunda tabla que podemos obtener es la tabla ANOVA que nos ofrece información sobre si el resultado del modelo de regresión es estadísticamente significativo y si la predicción que ofrece sobre la variable dependiente es mejor que si se usara el valor

medio. Nuestra atención debe concentrarse en el estadístico – F, el cual debe estar por debajo de p = 0.05 para considerarse significativo. Para nuestro ejemplo, lo es.

Tabla 13

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
H_0	(Intercept)	51.062	2.677		19.076	< .001
H_1	(Intercept)	4.137	0.738		5.606	< .001
	X	0.902	0.012	0.991	72.949	< .001

Finalmente, la tabla 12 de coeficientes nos ofrece los valores de los coeficientes que forman la ecuación de la línea de regresión. Nuevamente debemos enfocarnos en la fila del modelo H1 (intercepto) y X (pendiente).

Si quisiéramos calcular el nivel de una enzima B sólo conociendo el valor de la enzima A deberíamos usar la siguiente sustitución de la ecuación (6):

$$Y' = 4.13 + (0.902 * X)$$

Resumiendo, podemos reportar los resultados obtenidos de la siguiente manera: Una regresión lineal estableció que el nivel de la enzima A podría predecir de manera estadísticamente significativa el nivel de la enzima B, $F(1, 98) = 5321.617$, $p<.001$, y la enzima A representó el 98% de la variabilidad explicada en el nivel de la enzima B. La ecuación de regresión fue: Enzima B prevista = 4.13 + 0.902(nivel de la enzima A).

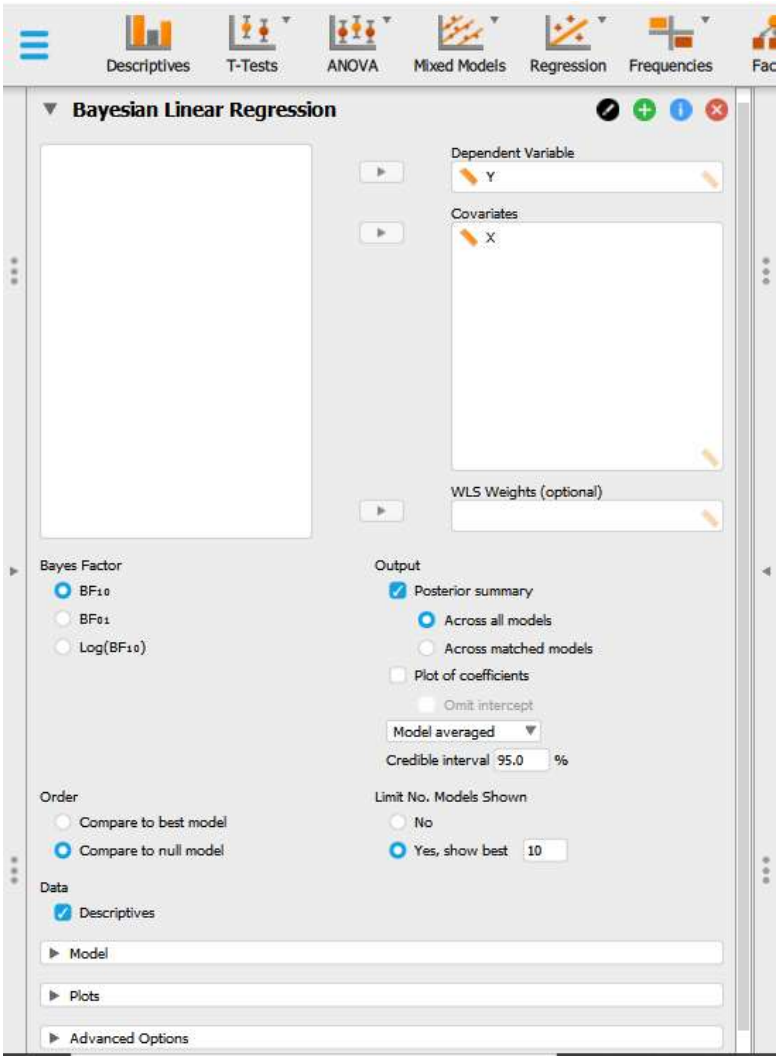
Regresión lineal Bayesiana en JASP.

Al igual que con el análisis de regresión frecuentista, debemos iniciar con un análisis descriptivo de los datos para darnos una idea acerca de las propiedades y naturaleza de estos.

Posteriormente, para el análisis de regresión por factor de Bayes nos dirigimos a Regression > Bayesian > Linear Regression. Esto nos mostrara el siguiente panel:

Figura 14

Regresión lineal bayesiana en JASP



Donde, al igual que con el análisis frecuentista, debemos posicionar la variable dependiente en la caja ‘Dependent Variable’ y la independiente en ‘Covariates’. Con las casillas activadas es suficiente para un análisis básico que nos ofrece las siguientes tablas de resultados:

Tabla 14

Model Comparison - Y

Models	P(M)	P(M|data)	BF_M	BF_{10}	R^2
Null model	0.500	1.013e -83	1.013e -83	1.000	0.000
X	0.500	1.000	∞	9.870e +82	0.982

Donde P(M) es la distribución de probabilidad prior asignada para cada modelo, a los cuales se les asigna probabilidades iguales. P(M) = 0.5. P(M|data) es la distribución de probabilidad posterior tomando en cuenta los datos, la cual pasa de 0.5 a 1 para el modelo que contiene la variable independiente X.

El valor BF10 sugiere que existe fuerte evidencia a favor del modelo alternativo (X) para contener la variable X. Y, el valor R^2 sugiere que por si sola la variable independiente representa una variación del 98.2% en el modelo.

Tabla 15

Posterior Summaries of Coefficients

								95% Credible Interval	
Coefficient	P(incl)	P(excl)	P(incl|data)	P(excl|data)	$BF_{inclusion}$	Mean	SD	Lower	Upper
Intercept	1.000	0.000	1.000	0.000	1.000	51.062	0.362	50.330	51.800
X	0.500	0.500	1.000	0.000	9.870e +82	0.901	0.012	0.876	0.926

La tabla 15 nos ofrece los valores de los coeficientes de la ecuación de la línea de regresión, enfocándonos principalmente en la media. La ecuación de regresión, ecuación (6), sufre una modificación al aplicar el análisis de regresión por factor de Bayes. El valor de X (valor de la variable independiente sobre la cual se hace la predicción del valor de la dependiente) se le resta la media de los valores de X.

$$Y' = \beta_0 + \beta_1 * x_1$$

(14)

Donde β_0 = intercepto

β_1 = pendiente

x_1 = diferencia del valor de X y la media de esta ($x_1 = X - \bar{X}$)

Por último, para reportar los resultados podemos decir:

Se llevó a cabo una regresión bayesiana simple utilizando el nivel de una enzima A como predictor del nivel de una enzima B. Se estableció un uniforme desinformado previo [P (M)] de 0,5 para cada modelo posible. Hubo evidencia sólida para un modelo de regresión que incluye la fuerza de la pierna derecha (BF10 9.87e+82) en comparación con el modelo nulo.

Esta es una condensación sobre correlaciones y regresiones, para un análisis más detallado se sugiere revisar las siguientes referencias:

Pagano, R. R. (2012). *Understanding Statistics in the Behavioral Sciences*. CENGAGE Learning.

Pardo, A. & San Martín, R. (2014). *Análisis de Datos en Ciencias Sociales y de la Salud, Vol I*. Universidad Autónoma de Madrid.

Pardo, A. & San Martín, R. (2014). *Análisis de Datos en Ciencias Sociales y de la Salud, Vol II*. Universidad Autónoma de Madrid.

Witte, R. S. & Witte, J. S. (2016). *Statistics*. WILEY.

Perfil de Github de The DOOM Lab https://github.com/doomlab

Página de JASP con materiales de aprendizaje https://jasp-stats.org/teaching-with-jasp/