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 8 que corresponde a los puntos de los datos de la tabla 7.

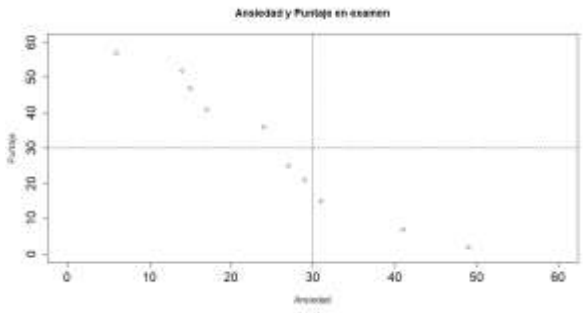
Tabla 7

Datos Ansiedad y Puntaje

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

Figura 8

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

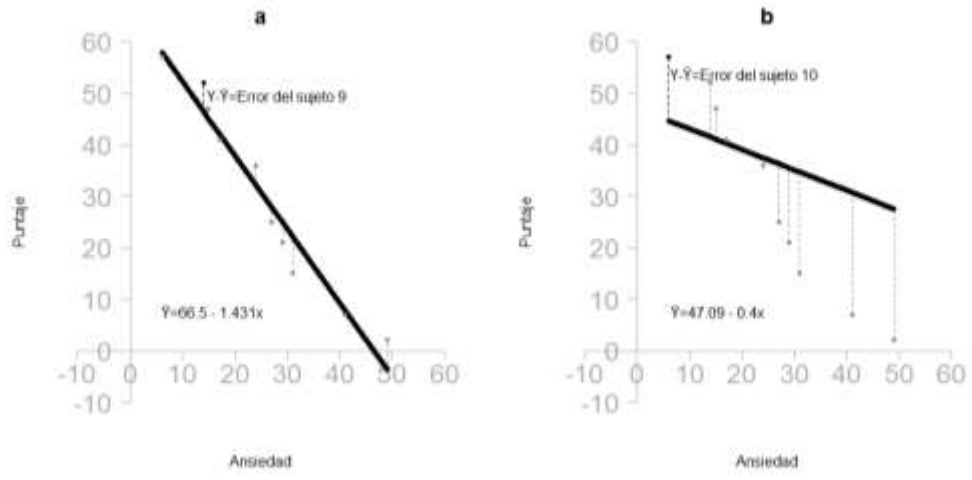
Una vez que identificamos la existencia de una relación lineal podemos pasar a establecer la recta que resume los puntos que representan los datos obtenidos de las variables estudiadas. La solución más utilizada es construir una línea que minimiza los

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 9.a podemos visualizar la línea de regresión para los datos de la tabla 7, 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 9

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 7, 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 6.

Tabla 8

Datos tabla 7 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 8 que se han transcrito a la siguiente tabla agregando la diferencia al cuadrado del valor de *Y*, y el valor predicho (*Y´):*

Tabla 9

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

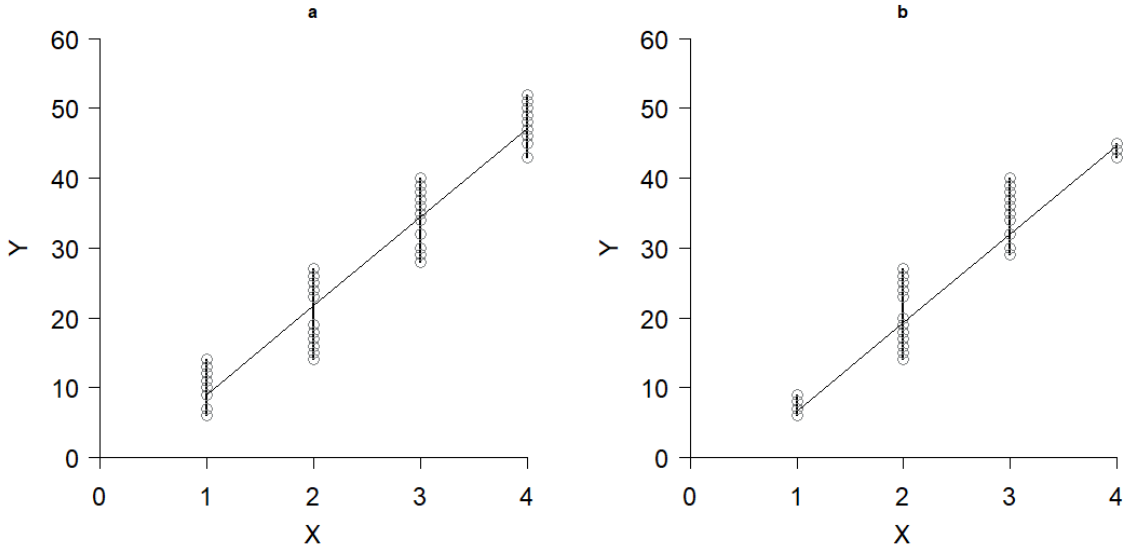
Pearson y la línea de regresión. Podemos definir esta nueva ecuación como ecuación **computacional** de error estándar de estimación al predecir *Y* dado *X*:

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

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 10.a podemos observar que se cumple esta suposición, situación contraria para la figura 10.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 10

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.

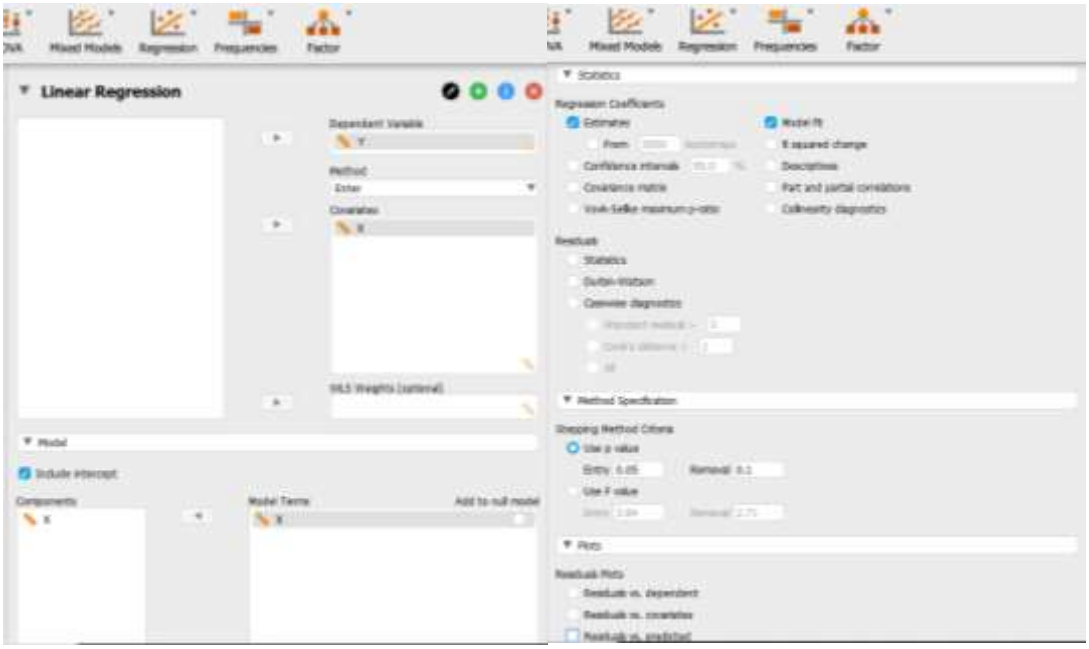
Nuevamente es recomendable iniciar con análisis descriptivos para darnos una idea de los datos con los que estamos trabajando.

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

Figura 11

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 10

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 11

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 12

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):

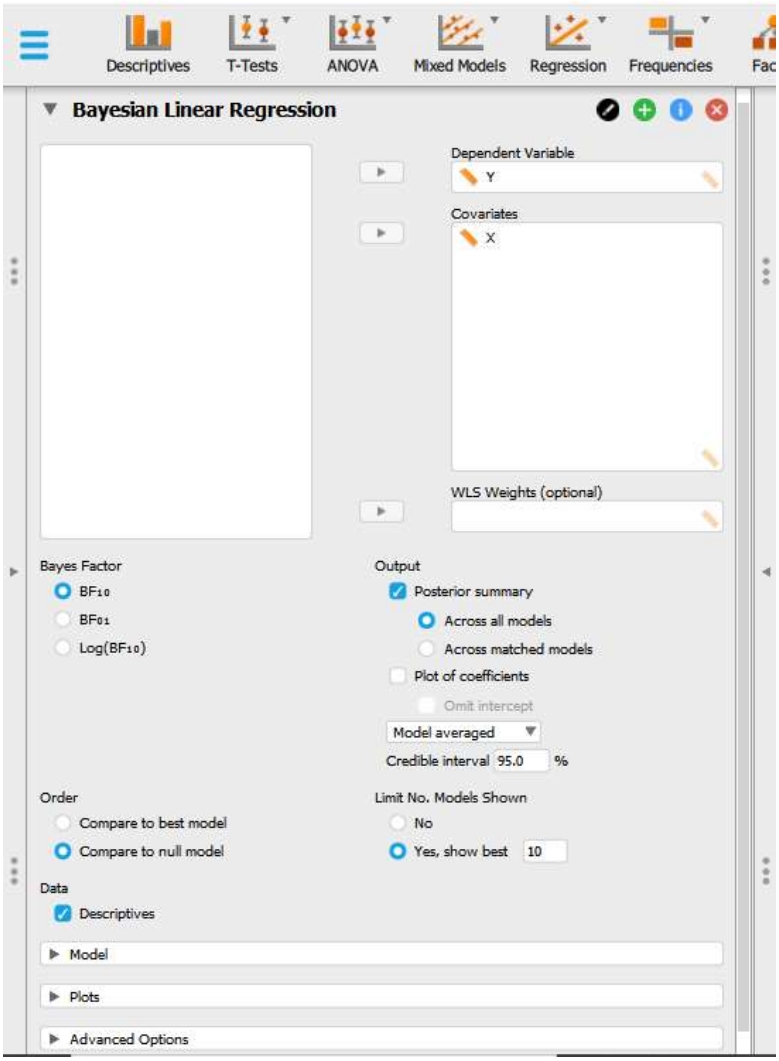
$$Y' = 4.13 + (0.902*X)$$

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:



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 13

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 14

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 14 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.

$$Y' = \beta_0 + \beta_1 * x_1 \quad (14)$$

Donde β_0 = intercepto

β_1 = pendiente

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

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/