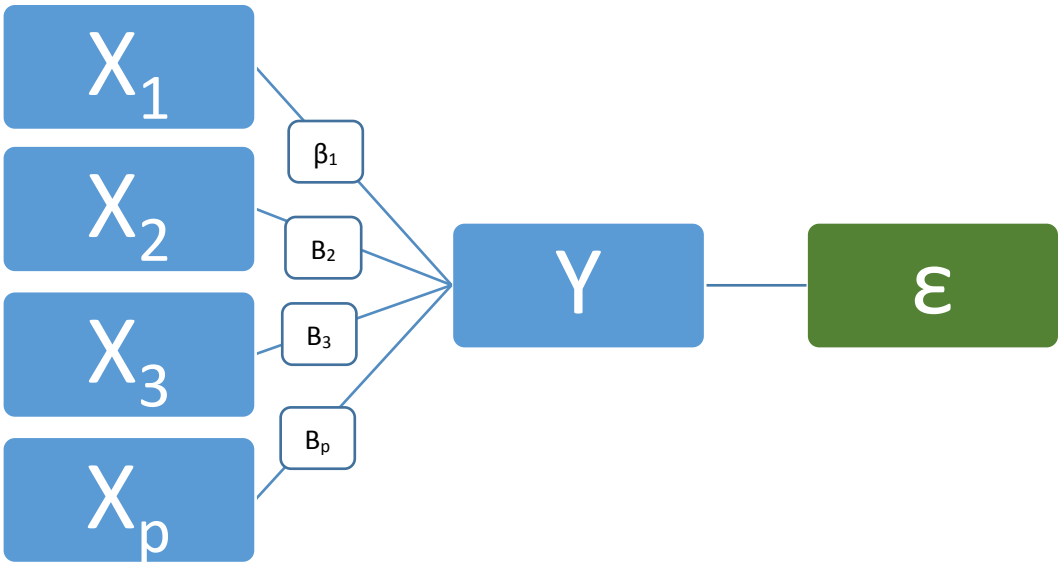
Análisis de regresión lineal múltiple

La regresión lineal múltiple es una extensión del modelo lineal simple. Como recordarás, en una regresión lineal simple se tiene una relación de proporcionalidad entre dos variables cuantitativas, una variable dependiente "Y" y una variable independiente "X". En el caso de la regresión múltiple se tiene también una variable dependiente "Y", pero más de una variable independiente "X".

El modelo de regresión lineal múltiple es el siguiente:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_p X_p \pm \varepsilon$$

En una forma gráfica se vería así:



Por ejemplo, imaginemos que tú trabajas para una organización que busca disminuir el consumo de cigarros en colegios. Para lograr este objetivo estás interesado en investigar acerca de aquellas variables que están relacionadas con el consumo. En este caso la variable dependiente será el número de cigarros fumados por cada alumno (Y) y las variables independientes serán aquellas que tú consideras están asociadas con fumar: Ansiedad rasgo (X_1), Consumo de otras sustancias (X_2), Desempeño académico (X_3), Porcentaje de amigos cercanos que fuman (X_4) Resiliencia (X_5), etc.

Cada una de estas variables tendrá un efecto diferenciado en la variable dependiente, algunas tendrán una relación positiva (por ejemplo Ansiedad rasgo "X_1") mientras que otras tendrán una relación inversa (por ejemplo Resiliencia "X_5"). Esto se observa utilizando el coeficiente de regresión (β), cada variable independiente X es multiplicada por una constante β diferente. Así $\beta_1 * X_1$, $B_2 * X_2$, etc. En el caso de relaciones positivas el β será positivo, pero en las inversas el β será negativo.

Un modelo de regresión múltiple nos permite comparar la importancia relativa de cada una de las variables en el modelo, algunas estarán más o menos relacionadas con la variable dependiente. Para determinar qué variable independiente está más vinculada a la variable dependiente puedes utilizar los coeficientes de regresión estandarizados (β estandarizado o tipificado): mientras más grande el valor

absoluto, mayor será la importancia de esa variable en particular. Recuerda que el β estandarizado es el parámetro adecuado para comparar la importancia relativa de las variables del modelo que planteas, el β no estandarizado no sirve para cumplir esta función.

No todas las relaciones entre variables siguen una lógica lineal, para evaluar si el modelo planteado es adecuado se verifica el "Coeficiente de Correlación" (R), este nos permite evaluar la calidad de la capacidad predictiva del modelo que se ha planteado. Se interpreta como una correlación entre los valores predichos por el modelo y los valores reales de Y, mientras más alto, mejor ajuste. Otro parámetro importante en esta sección es el Coeficiente de determinación (R^2), éste nos permite explicar qué porcentaje de la varianza de Y es explicada por las distintas X. Imagina que en nuestro ejemplo anterior obtenemos un R^2 de 0.48, eso significa que 48% de la varianza del número de cigarros fumados por alumno se debe a las 5 variables que habíamos identificado.

En algunas situaciones deberás comparar distintos modelos lineales, para realizar esto debes revisar el Coeficiente de determinación ajustado (R^2A), esta versión del R^2 atenúa algunos efectos de "inflación" de los que se puede sufrir producto de incluir muchas variables independientes en el modelo.

En resumen, una regresión lineal múltiple te permite modelar la relación entre varias variables independientes (X) y una variable dependiente cuantitativa (Y). Algunas de estas relaciones serán positivas ($\beta > 0$), mientras que otras serán inversas ($\beta < 0$) además algunas serán más intensas que otras (valor absoluto del β estandarizado). La calidad y poder explicativo de los modelos planteados la puedes observar en el Coeficiente de correlación (R) y el Coeficiente de determinación (R^2) y, en caso requieras comparar el poder explicativo entre modelos es importante que utilices el Coeficiente de determinación ajustado (R^2A).

Ejercicio de Regresión Lineal Múltiple (Regresión lineal múltiple.sav)

El archivo **Regresión lineal múltiple.sav** contiene los datos de una investigación que busca explicar la claridad del autoconcepto *(Y)* en un grupo de adultos jóvenes limeños. La claridad del auto-concepto puede definirse como "la medida en que las creencias acerca de uno mismo se encuentran clara y confiablemente definidas, poseen coherencia interna y son temporalmente estables". Se busca explicar esta variable mediante otras variables psicológicas tales como la capacidad de insight (X1), la tendencia a la rumiación (X2) y la autoestima (X3). Además, se tomó nota de las variables sociodemográficas sexo (X4) y edad (X5).

Se selecciona de manera aleatoria a 50 participantes del estudio y se registra cada una de las variables en el programa estadístico SPSS. Trabajando en hojas de un archivo Excel:

- Suponga que se tiene como hipótesis de trabajo que *la capacidad de insight y la autoestima influyen directamente sobre la claridad del autoconcepto aunque la tendencia a la rumiación juega en contra de la claridad.* En este contexto ¿Cuál de los modelos siguientes representa todas las hipótesis? Justifique.

  Modelo 1: Claridad = ß0 + ß1×Insight + ß2×Rumiación + ß3×Autoestima + e, ß1>0, ß2<0, ß3>0
  Modelo 2: Claridad = ß0 + ß1×Insight + ß2×Rumiación + ß3×Autoestima+ e, ß1>0, ß2< 0, ß3≠0
  Modelo 3: Claridad = ß0 + ß1×Insight + ß2×Rumiación + ß3×Autoestima + e, ß1>0, ß2>0, ß3<0

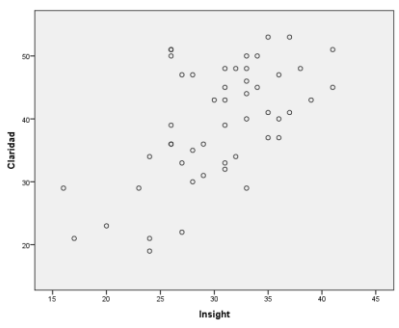
  El modelo 3 representa adecuadamente las hipótesis planteadas.

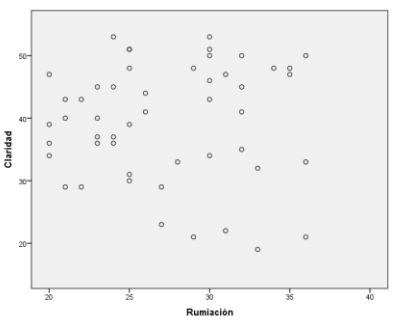
- Pida a SPSS diagramas de dispersión (*Gráficos ⇒ Cuadros de diálogo antiguos ⇒ Dispersión/puntos → Dispersión simple → Definir → Eje Y: Claridad (Y), Eje X: Insight (X1) /Rumiación (X2) / Edad (X3) →Aceptar*). ¿Hay evidencia gráfica a favor del modelo seleccionado en a)?

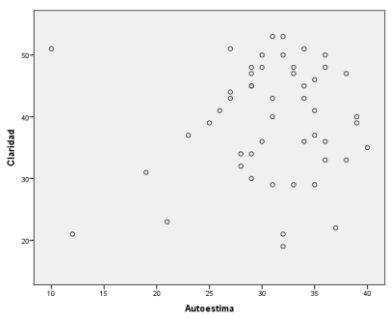
```
GRAPH
  /SCATTERPLOT(BIVAR)=X1 WITH Y
  /MISSING=LISTWISE.

GRAPH
  /SCATTERPLOT(BIVAR)=X2 WITH Y
  /MISSING=LISTWISE.

GRAPH
  /SCATTERPLOT(BIVAR)=X3 WITH Y
  /MISSING=LISTWISE.
```







A nivel gráfico hay evidencia solo para las hipótesis ß1>0 y ß2<0, más no para ß3>0.

- **(Uso de sintaxis de SPSS)** Aplique la secuencia de comandos SPSS: *Analizar ⇒ Regresión ⇒ Lineales → Dependientes: Claridad (Y), Independientes: Insight (X1), Rumiación (X2), Autoestima (X3)* →***Pegar***: SPSS en lugar de presentar resultados, abre una ventana Sintaxis 1 y muestra los comandos:

```
REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS R ANOVA
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT Y
 /METHOD=ENTER X1 X2 X3.
```

Estos comandos son *órdenes para ejecutar el análisis de regresión, pero **sin** realizar **inmediatamente** el análisis*. Para que SPSS *ejecute* los comandos: Sombrear o resaltar el texto desde REGRESSION hasta el punto final "**.**" y luego presionar el botón ▶ (Ejecutar la selección). SPSS procesa datos y muestra resultados. Ejecute la sintaxis generada, copie las tablas SPSS a la Hoja 1 de Excel y responda:

Observando el coeficiente **R** ¿cuán bien representa el modelo a los datos? Explique. Analice estadísticamente los datos y determine el porcentaje de varianza de Claridad que se debería a las v.i. del modelo completo. ¿Es significativo este porcentaje? ¿Es grande o pequeño? ¿Se podría rechazar la hipótesis nula H0: ß1= ß2= ß3=0 (o H0:R2=0)? ¿Cuánto valen las estimaciones de los parámetros ß1, ß2 y ß3? ¿Cuánto valen los errores estándar de estimación (E.E.) de ß1, ß2 y ß3? ¿Cuáles hipótesis de trabajo se comprueban? Justifique. Indique la o las tablas de resultados que ha usado.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	.589[a]	.347	.304	7.716

a. Variables predictoras: (Constante), Autoestima, Insight, Rumiación

El presente modelo de regresión múltiple representa a los datos entre moderada y altamente en un 58,9%.
El porcentaje de claridad que es producto de las v.i. del modelo completo es de 34,7%, que podría considerarse entre pequeño y moderado.

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	1453.764	3	484.588	8.139	.000[b]
	Residual	2738.656	46	59.536		
	Total	4192.420	49			

Es significativo (p =.00 < .05), por lo que podría rechazarse la hipótesis nula H0: ß1= ß2= ß3=0 (o H0:R2=0).

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	8.711	9.442		.923	.361
	Insight	.971	.199	.585	4.884	.000
	Rumiación	-.010	.231	-.005	-.044	.965
	Autoestima	.049	.185	.032	.265	.792

a. Variable dependiente: Claridad

Los valores de las estimaciones de los parámetros son: ß1=.971, ß2=-.10, ß3=.049.
Los valores de los E.E. son: ß1=.199, ß2=-.231, ß3=.185
Sólo se cumpliría la hipótesis El insight influye directa y proporcionalmente sobre la Claridad del Autoconcepto (H1:B1>0), ya que es la única que resulta significativa (p<.00)

La ventaja de generar la sintaxis es que deja un registro de lo que se hace durante una sesión de trabajo y se puede modificar para hacer el mismo tipo de análisis, pero con otras o con más variables: basta escribir los *nombres* SPSS de las variables en la sintaxis y luego ejecutarla.

- Entre los supuestos de análisis de regresión múltiple **uno es el de variables explicativas independientes o no excesivamente correlacionadas**. Cuando una v.i. correlaciona excesivamente con las demás v.i. de un modelo, una consecuencia es que su Tolerancia es muy baja y eso aumenta artificialmente la varianza del error o residuo. La inversa de la tolerancia se llama Factor de Inflación de Varianza (FIV donde FIV = 1/Tolerancia) y mide cuánto aumenta la varianza del residuo debido a una excesiva correlación de la respectiva v.i. con el resto. El FIV no debe ser alto y, en general, se pide que no pase de 4*. ¿Es alto el FIV de **Claridad cuando las v.i. son Insight, Rumiación y Autoestima**?

  *Se puede pedir a SPSS que muestre las Tolerancias y FIV de cada v.i. de un modelo. Para ello, hay que entrar a la opción **Estadísticos y marcar: Diagnósticos de colinealidad.**

```
REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS R ANOVA COLLIN TOL
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT Y
 /METHOD=ENTER X1 X2 X3.
```

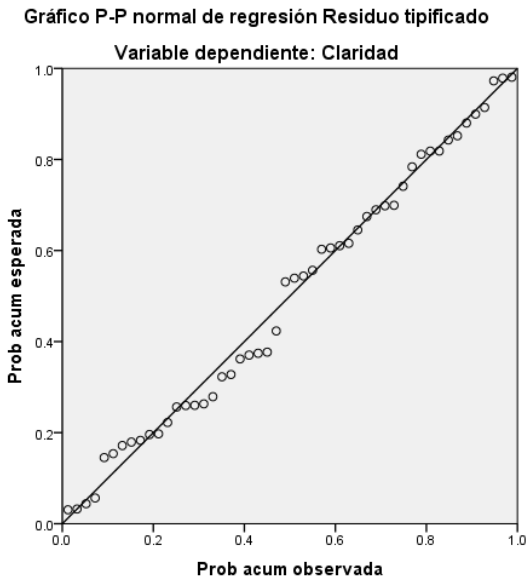
Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Estadísticos de colinealidad	
		B	Error típ.	Beta			Tolerancia	FIV
1	(Constante)	8.711	9.442		.923	.361		
	Insight	.971	.199	.585	4.884	.000	.990	1.010
	Rumiación	-.010	.231	-.005	-.044	.965	.960	1.041
	Autoestima	.049	.185	.032	.265	.792	.951	1.052

Dado que ninguna de las FIV es mayor a 4, no hay problemas de colinealidad.

- Otro de los supuestos de la regresión múltiple es el de la distribución normal de los residuos (errores). Una distribución no normal de los errores de las variables pueda afectar las relaciones y su significación. Nos enfocamos en los errores y no en la medición porque en las regresiones lineales pueden usarse también variables dicotómicas (p.e. sexo) y estas no siguen una distribución normal. Para ello, pida al SPSS un gráfico para conocer la distribución de los residuos siguiendo los siguientes pasos: *Analizar ⇒ Regresión ⇒ Lineales → Dependientes: Claridad (Y), Independientes: Insight (X1), Rumiación (X2), Autoestima → Gráficos: Gráfico de Prob. Normal → Aceptar.* Responda: ¿Tienen los residuos una distribución normal?

```
REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS R ANOVA
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT Y
 /METHOD=ENTER X1 X2 X3
 /RESIDUALS NORMPROB(ZRESID).
```



Los residuos tienen una distribución normal dado que los "o" caen sobre o cerca de la pendiente.