

La Regresión Lineal:

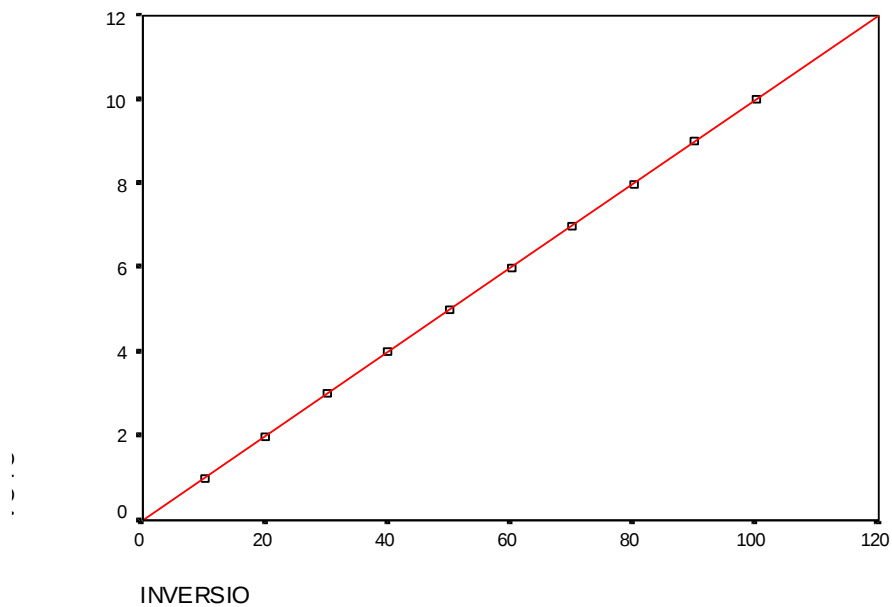
En la clase anterior estudiamos las correlaciones que es una medida de asociación entre variables intervalares, pero señalamos que no era posible predecir o establecer relaciones de causalidad usando solamente a las correlaciones como herramientas estadísticas. Para ello debemos usar un instrumento más complejo, las regresiones, especialmente las lineales que son las más comunes. En el análisis de regresiones buscamos construir un modelo para predecir los valores de la variable dependiente a partir de una (regresión univariada) o más variables independientes (regresión bivariada). En las regresiones tratamos de adecuar el modelo que mejor explica los datos mediante una línea que describa mejor los datos. En los modelos lineales esta será una línea recta. Esta línea recta se llama la línea de mejor ajuste y constituye la línea que hace mínima la suma de los cuadrados de las desviaciones respecto a ella de todos los puntos que corresponden a la información recogida.

La línea de mejor ajuste puede dibujarse si conocemos:

- a. La gradiente (slope) de la línea
- b. El punto donde la línea cruza el eje vertical del gráfico (conocido como el punto de intersección de la línea).

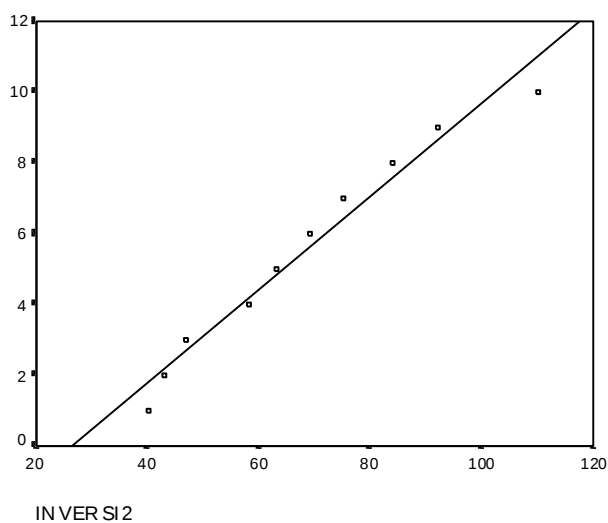
Así por ejemplo si tenemos la tabla:

Voto	Inversión
1	10
2	20
3	30
4	40
5	50
6	60
7	70
8	80
9	90
10	100



En este caso tenemos una correlación perfecta entre ambas variables donde sabemos que aumentaremos la votación por nuestro candidato cada vez que aumentemos en diez la inversión en publicidad. Entonces la ecuación es muy simple: $Y=10X$. Esta es la recta de regresión de la inversión sobre el voto. Ahora por costumbre a ese 10, generalmente desconocido se le llama b (beta). Por lo que la ecuación es $Y=bX$. Donde Y es un punto en la variable dependiente x un punto en la independiente y b el cambio producido en la variable dependiente por cada unidad de cambio en la variable independiente.

Pero estos casos de correlaciones tan perfectas son irreales y la gran mayoría de las veces el punto de intersección no es 0 y la gradiente no es tan clara.



En este caso debemos introducir el punto de intersección en la ecuación. Así la ecuación de la regresión es:

$$Y = a + bX$$

Donde nuevamente: Y es un punto en la variable dependiente que esperamos predecir, a es la intersección en Y o el punto donde la línea de la regresión cruza el eje Y; b es la gradiente (otra vez, la cantidad de cambio en Y por cada unidad de cambio en X; y X el punto o puntaje en la variable independiente).

Piensen en la gradiente de la línea de regresión como la medida de los efectos de la variable X en la variable Y. Si las variables poseen una asociación fuerte, entonces los cambios en el valor de X serán acompañados por cambios sustanciales en el valor de Y, y la gradiente tendrá un valor numéricamente alto. Cuanto más bajos sean los efectos de X sobre Y (menor asociación entre las variables) más bajo será el valor numérico de la gradiente. Si las dos variables no están relacionadas, la línea de regresión será paralela a la abscisa, y b será 0 (no gradiente).

Procedimiento:

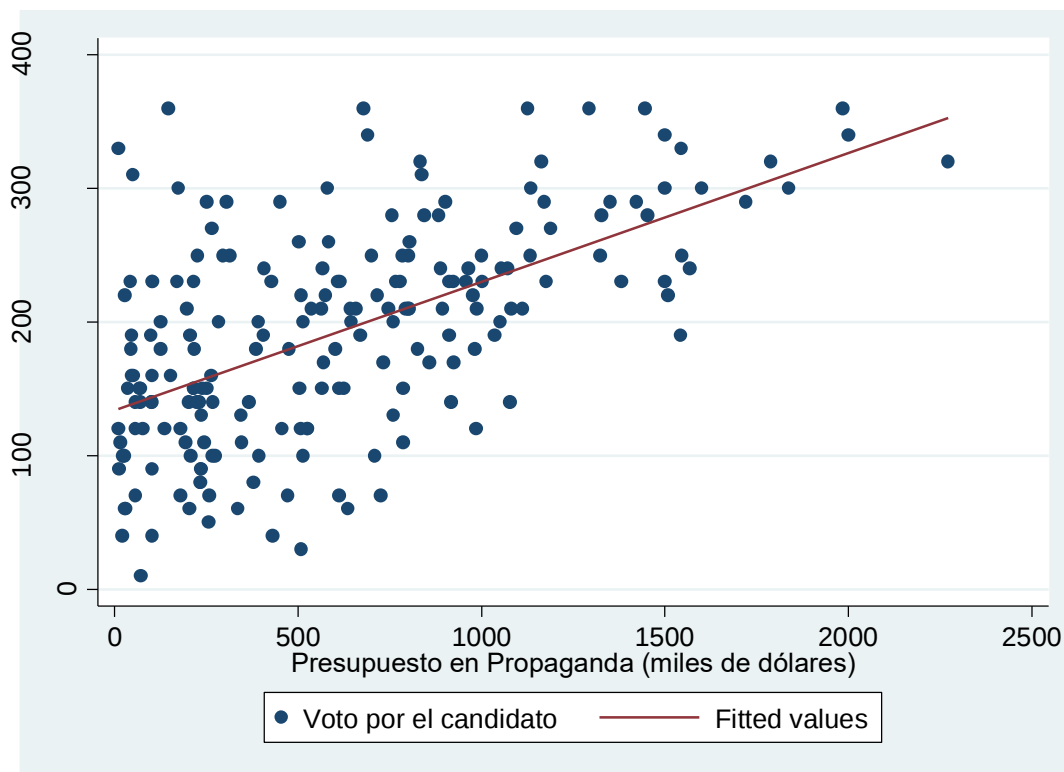
*Para realizar un grafico de la regresion usar los comandos:

twoway (scatter vary varx) (lfit vary varx)

En el ejemplo de clase: twoway (scatter voto Adverts) (lfit voto Adverts)

graph twoway (lfit voto Adverts) (scatter voto Adverts)

También se puede proceder a partir de Graph, luego easy graph y finalmente regresión fit



Para el modelo de regresión:

En la ventana statistics, escoger linear models and related y allí linear regresión. Debemos recordar que al buscar causalidad una de las variables sera dependiente (x) y una o varias independientes (y). Veamos un ejemplo: Como predecir la cantidad de votos por un candidato a partir de la inversión en publicidad.

Por los comandos reg vary varx
reg voto Adverts

Source	SS	df	MS	Number of obs =	200
-----+-----				F(1, 198) =	99.59
Model	433687.975	1	433687.975	Prob > F =	0.0000
Residual	862264.025	198	4354.86881	R-squared =	0.3346
-----+-----				Adj R-squared =	0.3313
Total	1295952	199	6512.32161	Root MSE =	65.991

voto	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
-----+-----					
Adverts	.0961245	.0096324	9.98	0.000	.0771293 .1151197
_cons	134.1399	7.536576	17.80	0.000	119.2776 149.0022

Considerando los valores obtenidos podemos decir que $b = 0.09612$. Si la unidad fuese 1 diríamos que a un dólar invertido la cantidad de votos aumenta en 0.096. Como la unidad está en miles ($0.096 * 1000 = 96$). Así por cada 1000 dólares aumentan 96 votos. Bastante mal, aunque la R cuadrado nos señala que la publicidad sólo explica 1 tercio de las intenciones de votos. La constante es a en la ecuación de la regresión por lo que sin gastar nada obtenemos 134 votos, lo que sería nuestra base.

El modelo no resulta tan bueno así que podemos introducir más variables buscando explicar un mayor porcentaje de variación. En este caso hablamos de una regresión múltiple. A la ecuación anterior agregamos las nuevas variables independientes:

$$Y = a + b_1X_1 + b_2X_2 + \dots + b_nX_n.$$

Donde b_1 es la gradiente parcial de la relación lineal entre la primera variable independiente e Y. Y b_2 es la gradiente parcial de la relación lineal entre la segunda variable independiente e Y. Así hasta completar el número de variables independientes.

Usando el mismo ejemplo agregamos como variables independientes el número de veces que el candidato sale en la prensa y el atractivo que posee.

reg voto Adverts airplay Attract

Source	SS	df	MS	Number of obs =	200
-----+-----				F(3, 196) =	129.50
Model	861377.22	3	287125.74	Prob > F =	0.0000
Residual	434574.78	196	2217.21826	R-squared =	0.6647
-----+-----				Adj R-squared =	0.6595
Total	1295952	199	6512.32161	Root MSE =	47.087

voto	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
-----+-----						
Adverts	.0848848	.006923	12.26	0.000	.0712316	.098538
airplay	3.367424	.2777709	12.12	0.000	2.81962	3.915227
Attract	11.08634	2.43785	4.55	0.000	6.278554	15.89412
_cons	-26.61294	17.35	-1.53	0.127	-60.8296	7.603716
-----+-----						