Regresión lineal

Xochitl Cárdenas

Puedes encontrar esta presentación y materiales adicionales en mi perfil de Github: https://github.com/xochitlcardenas

Objetivos de aprendizaje

1. Definir regresión, línea de regresión y constante de regresión.
2. Especificar la relación entre la fuerza de la relación y la precisión de la predicción.
3. Construir la línea de regresión de mínimos cuadrados para predecir Y dado X, especificar lo que minimiza la línea de regresión de mínimos cuadrados.
4. Explicar qué se entiende por error estándar de estimación.
5. Especificar las condiciones para utilizar la regresión linear.
6. Reforzar lo aprendido con un ejercicio.

Regresión

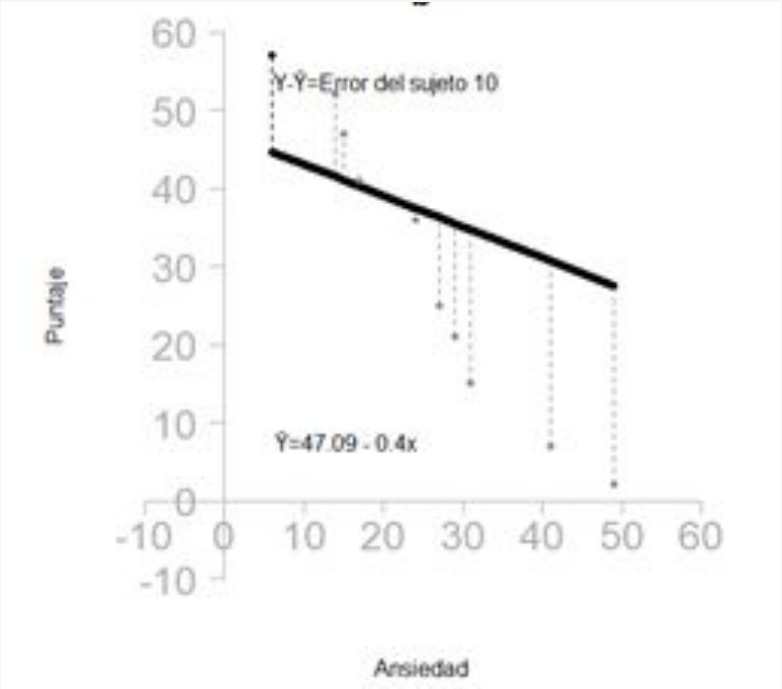
Correlación ≠ Regresión.

Regresión: utiliza la relación entre dos (o más) variables para hacer predicciones sobre una de ellas.

Paso siguiente una vez identificada la relación significativa.

1. Resumir los puntos en una recta
2. Valorar el grado de concentración o alejamiento de los puntos a partir de la recta.

Línea de regresión



Mínimos cuadrados ($\Sigma(Y-Y')^2$):

Construir una línea que minimiza los errores de predicción de acuerdo con los mínimos cuadrados.

Error de predicción (y) = Y - Y'

Y = valor real

Y' = valor predicho

Línea de regresión

$$Y' = \alpha + \beta X$$

β y α son valores constantes

Y’ = valor predicho de la variable Y (o dependiente).

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

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

X = valor de la variable independiente sobre el cual se hace la predicción de Y’

Línea de Regresión – β

$$\beta = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{SS_X}$$

$$SS_X = \Sigma X^2 - \frac{(\Sigma X)^2}{N}$$

SS_x = suma de los cuadrados de X

N = número de puntos

∑XY = suma de los productos de X y Y

Línea de Regresión – α

$$\alpha = \bar{Y} - \beta\bar{X}$$

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

$\bar{Y}$ = media de datos de la variable Y

$\bar{X}$ = media de datos de la variable X

Ejemplo:

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		

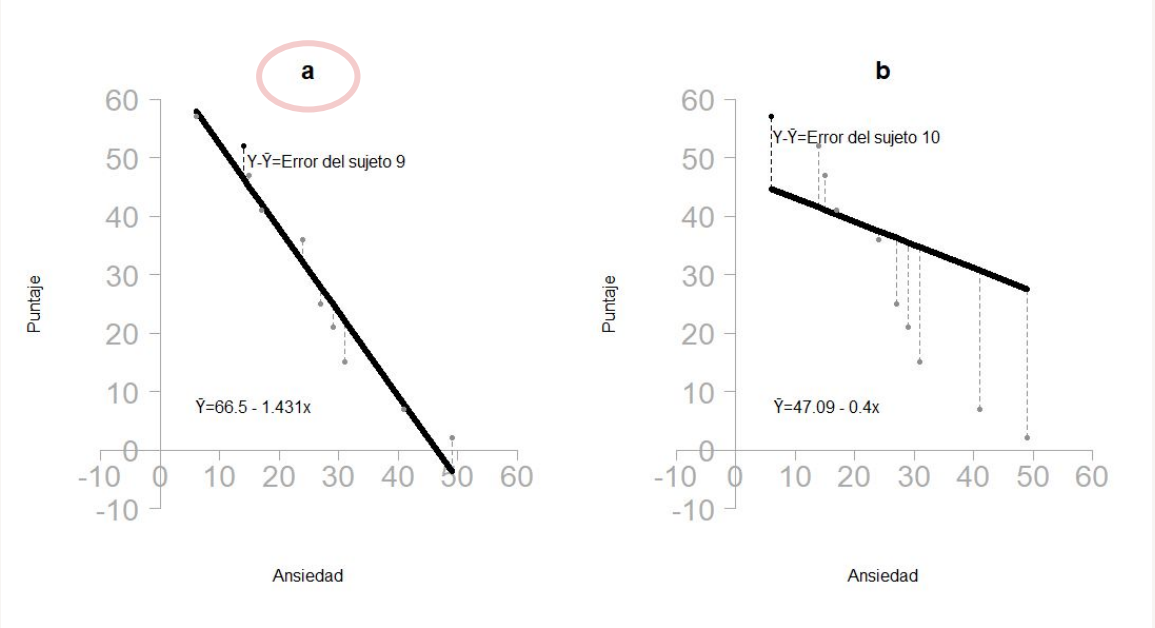
$$\beta = \frac{5470 - \frac{(253)(303)}{10}}{7935 - \frac{(253)^2}{10}} = -1.431$$

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

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

Ejemplo:

Sujetos	X(Ansiedad)	Y(Puntaje)	Y′
1	49	2	-3.619
2	41	7	7.829
3	31	15	22.139
4	29	21	25.001
5	27	25	27.863
6	24	36	32.156
7	17	41	42.173
8	15	47	45.035
9	14	52	46.466
10	6	57	57.914



Error de estimación estándar

Muchos de los puntos caerán fuera de esta línea, por lo que es necesario cuantificar este error.

Si este error es grande, confiaríamos poco en la predicción; si es pequeño, podemos fiarnos y tomar decisiones con base en ella.

Error de estimación estándar: desviación promedio de los errores de predicción sobre la línea de regresión

Ecuación para el error estándar de estimación para predecir *Y* dado *X* es:

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

Error de estimación estándar

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
				Σ	158.90

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

¿Cómo interpretamos este valor?

Error de estimación estándar e intervalo de confianza

Este error se interpretar a partir del valor medio (media) de la variable de interes.

Porcentaje de confianza	Distancias
60%	±1
95%	±2
99%	±3

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

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

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.

Regresión Lineal

La línea de regresión permite predecir valores de una variable *Y* dada una variable *X* que se relacionan de manera lineal

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.

Ejercicio 1:

Un grupo de investigadores evaluó la existencia de una relación entre el estrés percibido y la memoria comparando la puntuación obtenida en una escala de estrés y la puntuación obtenida en una prueba de memoria de 10 sujetos. Obtuvieron un valor del coeficiente de Pearson de 0.88. Ahora quieren calcular la puntuación de memoria en un grupo diferente de sujetos a quienes ya se les ha aplicado la escala de estrés.

Valores del primer grupo

X (estrés)	Y (Memoria)
64	66
40	79
30	98
71	65
55	79
31	83
61	68
42	80
57	72
38	95

Valores del segundo grupo

X (estrés)	Y (Memoria)
39	?
56	?
32	?
74	?
54	?
34	?
65	?
43	?
59	?
36	?

X	Y	X^2	Y^2	XY
64	66	4096	4356	4224
40	79	1600	6241	3160
30	98	900	9604	2940
71	65	5041	4225	4615
55	79	3025	6241	4345
31	83	961	6889	2573
61	68	3721	4624	4148
42	80	1764	6400	3360
57	72	3249	5184	4104
38	95	1444	9025	3610
∑ = 489	∑ = 785	∑ = 25801	∑ = 62789	∑ = 37079
μ = 48.9	μ = 78.5			

$$SS_X = \Sigma X^2 - \frac{(\Sigma X)^2}{N} \qquad SS_X = 25801 - \frac{(489)^2}{10} = 1888.9$$

$$\beta = \frac{\Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{N}}{SS_X} \qquad \beta = \frac{37079 - \frac{(489)(785)}{10}}{1888.9} = -0.692$$

$$\alpha = \bar{Y} - \beta\bar{X} \quad \alpha = 78.5 - (-0.692 * 48.9) = 112.95$$

$$Y' = \alpha + \beta X \qquad Y' = 112.95 + (-0.692X)$$

X (estrés)	Y (Memoria)	Y’	Y-Y’	$(Y-Y')^2$
64	66	70.95	-2	4
40	79	82.95	-39	1521
30	98	87.95	-68	4624
71	65	67-45	6	36
55	79	75.45	-24	576
31	83	87.45	-52	2704
61	68	72.45	-7	49
42	80	81.95	-38	1444
57	72	74.45	-15	225
38	95	83.95	-57	3249
			Σ	14432

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

$$S_{Y||X} = \sqrt{\frac{265.18}{8}} = 5.757$$

$$IC\,95\%\bar{X} = 48.9 \pm (2 * 5.757) = 37.38 - 60.14$$

El valor medio de la variable de interés debe estar en un rango entre 37.38 y 60.14 con una confianza del 95%.

Valores del segundo grupo

X (estrés)	Y (Memoria)
39	86
56	74
32	91
74	62
54	76
34	89
65	68
43	83
59	72
36	88

$$Y' = 112.95 + (-0.692X)$$