



UNIVERSIDAD
NACIONAL
DE COLOMBIA



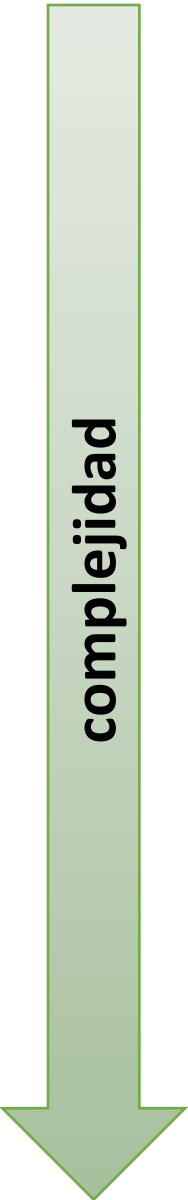
Universidad Nacional de Colombia

Sede Medellín

Regresión lineal y no lineal



Profesora: Patricia Jaramillo A. Ph.D



complejidad

NIVEL 1 variables

- características que se pueden medir o contar y ordenadas por observaciones

NIVEL 2 – gráficos y características

- . Herramientas: estadística descriptiva: media, dispersión, correlación etc...

NIVEL 3 – modelar datos

- Encontrar relaciones

NIVEL 4 – Reconocimiento de patrones

- reconocer patrones de los datos y crear herramientas predictivas: machine learning

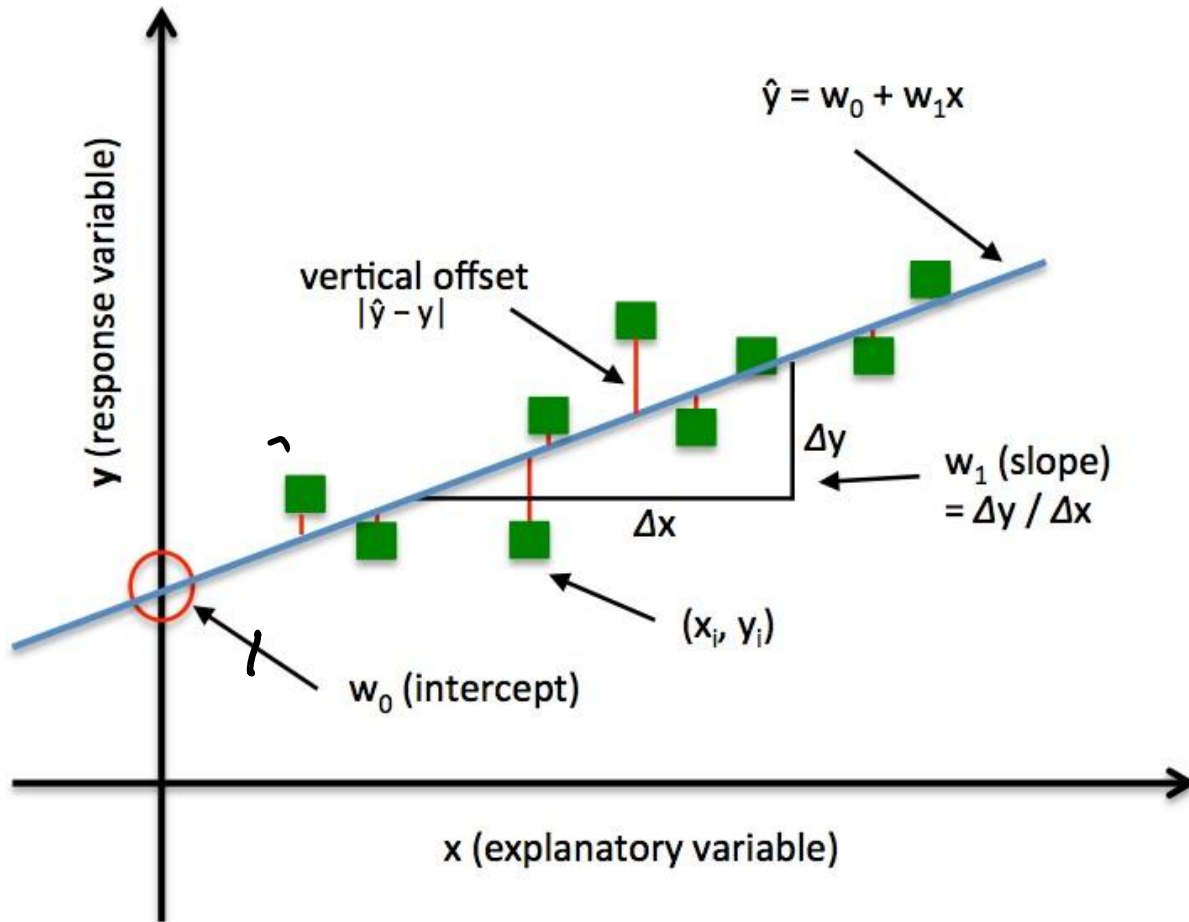
NIVEL 5 – Modelar el sistema completo

- **para la toma de decisiones.** Simulación, optimización

Análisis de regresión

- La regresión es el estudio de la **relación** entre dos o más variables
- Por ejemplo, si se requiere saber si hay relación entre algunas variables por ejemplo, publicidad y ventas, tiempo y costos u otras variables, de forma que, si se puede establecer una correlación, se hace posible encontrar un modelo que permita **realizar predicciones** o entender el comportamiento de dichas variables e incluir esta relación en un modelo más complejo.
- Una **serie de tiempo es un caso especial de regresión** que permiten un análisis especial en el que la variable independiente es el tiempo, y la dependiente la variable de interés. Sirven de apoyo para hacer proyecciones al futuro.

Regresión Lineal



■ Generalidades:

- En la **Regresión Lineal** el modelo a ajustar es una línea recta:

$$y = w_0 + w_1x$$

w_0 es la intercepción con el eje $x=0$, y w_1 es la pendiente de la recta ajustada.

El modelo se ajusta usando una **medida de error** sobre las predicciones que éste hace.

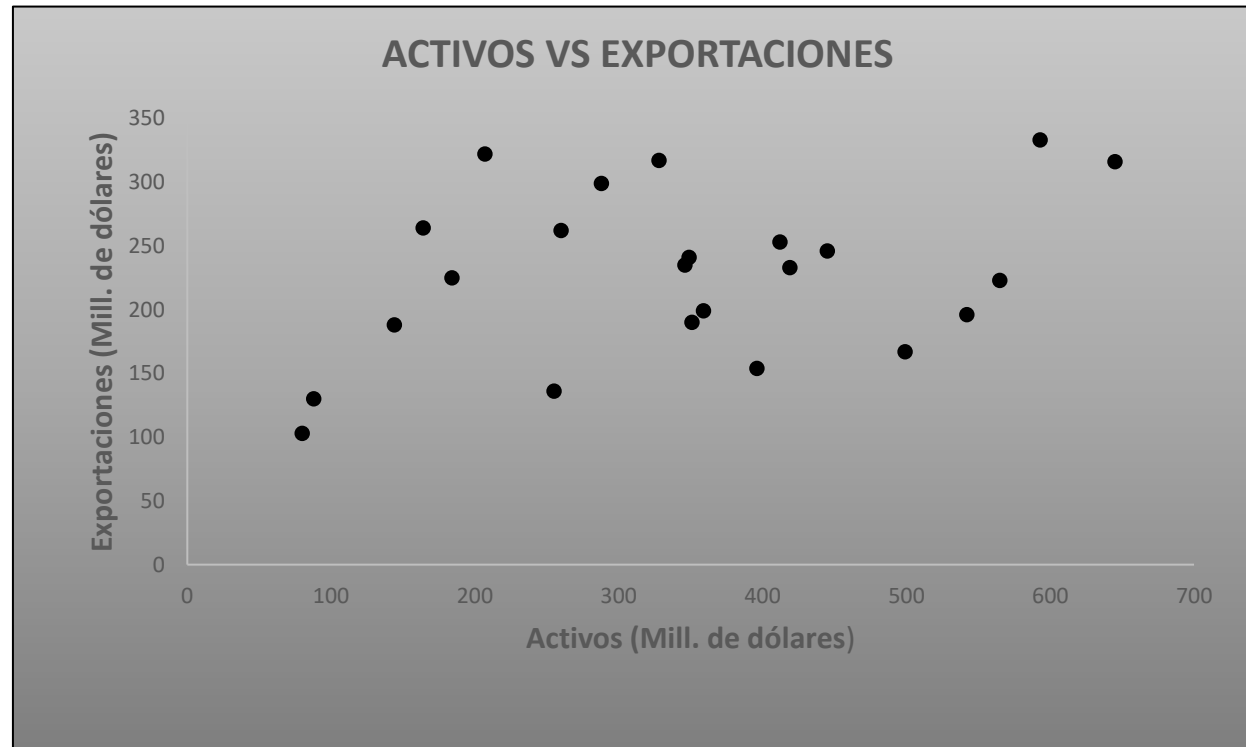
Puede haber múltiples líneas rectas dependiendo de los valores de intercepción y pendiente. Básicamente, lo que hace el algoritmo de regresión lineal es ajustar varias líneas y retornar la línea que produce el menor error.

Ejemplos

- ¿Qué tanto depende el salario de los empleados en una empresa de los años de experiencia, de los de educación y del género?
- ¿Qué tanto depende el precio actual de un stock de sus propios valores pasados y de los actuales y del pasado del Mercado ?
- ¿Qué tanto depende los niveles de ventas de una empresa de los precios de sus competidores?
- ¿Qué tanto depende el precio de venta de un inmueble del área, del sitio, del número de habitaciones, entre otras características?

Diagrama de dispersión

Importante inicialmente hacer diagrama de dispersión entre las variables de entrada y de salida para intuir que posible modelo es viable: lineal, exponencial, etc?



¿Cómo elegir cual variable es la dependiente y cuáles las independientes?

Una unión entre la lógica, la experiencia y experimentos controlados !

Ejemplos:

Se quieren explicar las ventas de una empresa únicamente en función de su inversión en publicidad.

- Variable independiente: Inversión en publicidad.
- Variable dependiente: Ventas.

Se desea evaluar que tanto influye la tasa de cambio del dólar en los Beneficios de una Empresa que importa materia prima

- Variable independiente: Tasa de cambio del dólar
- Variable dependiente: Beneficios.

Se desea evaluar el comportamiento de las ventas de una empresa en función del precio propio y del precio de la competencia.

- Variables independientes: Precio propio y precio de la competencia.
- Variable dependiente: Ventas

Causalidad?

- Una Análisis de regresión no significa necesariamente causalidad, a no ser que los datos hayan sido obtenidos en un experimento controlado.
- Ni siquiera se sabe cual es la dirección correcta de causalidad: X causa Y o Y causa X? o quizá haya una tercera variable Z que influye en X y Y pero que no fue observable?

Ecuación de regresión

- puede ser Lineal, Logarítmica, Exponencial, Polinómica, entre otras no lineales más complejas.
- La regresión puede ser:
 - **Simple**: si involucra únicamente dos variables (una independiente y una dependiente).
 - **Múltiple**: si involucra tres variables o más (una dependiente y dos o más independientes).

Regresión Lineal Simple

- Sebe utilizarse solamente si la nube de puntos del gráfico de dispersión lo sugiere, es decir, si la nube de puntos no se aleja mucho de una línea recta.
- La ecuación correspondiente es:
$$\hat{Y} = A + BX$$
- Donde A es el intercepto y B es la pendiente de la recta (equivale al cambio promedio en la variable Y debido a un cambio en una unidad en la variable X)

Error residual de observación i: diferencia entre los valores de Y modelo y Y observados:

$$e(i) = \text{Residual}(i) = Y \text{ observado}(i) - Y \text{ modelo } (i)$$

Se considera que la mejor ecuación de regresión es la que tiene menor suma del cuadrado de los errores residuales, pero suele usarse otros tipos de funciones. veamos

Caso: discriminación salarial, se cuenta con datos de salarios de una empresa y se desea analizar que tanto influye ciertas variables en el salario. Archivo “Discriminación salarial.csv”

	X1	X2	X3	X4	X5	Y
Empleado	genero (1 mujer)	Edad	Experiencia previa	experiencia en la empresa	Nivel educativo	Salario anual
1	1	39	5	12	4	57700
2	0	44	12	8	6	76400
3	0	24	0	2	4	44000
4	1	25	2	1	4	41600
5	0	56	5	25	8	163900
6	1	41	9	10	4	72700
7	1	33	6	2	6	60300
8	0	37	11	6	4	63500
9	1	51	12	16	6	131200

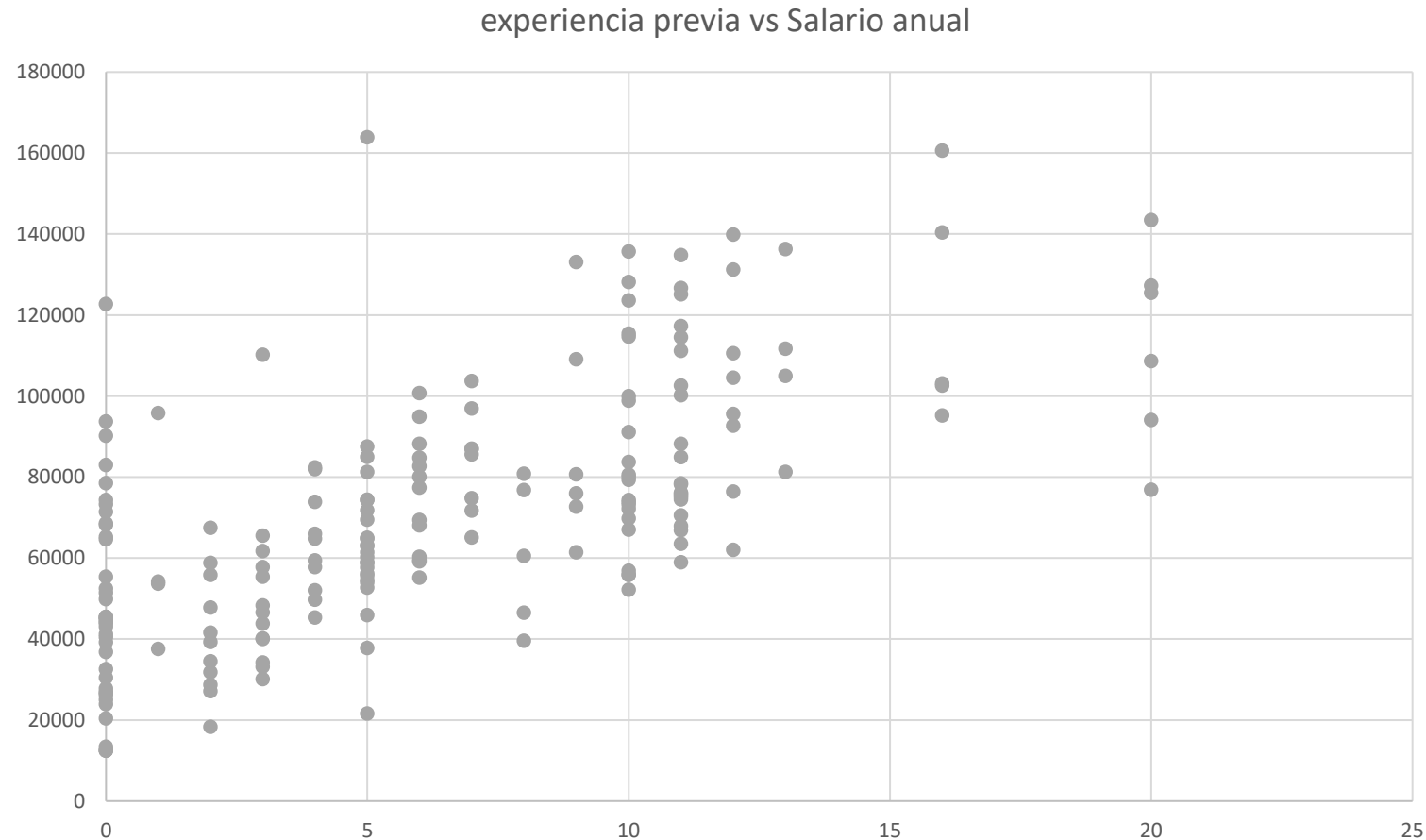


$Y=a+bX_3$ $a=1$ $b=2$

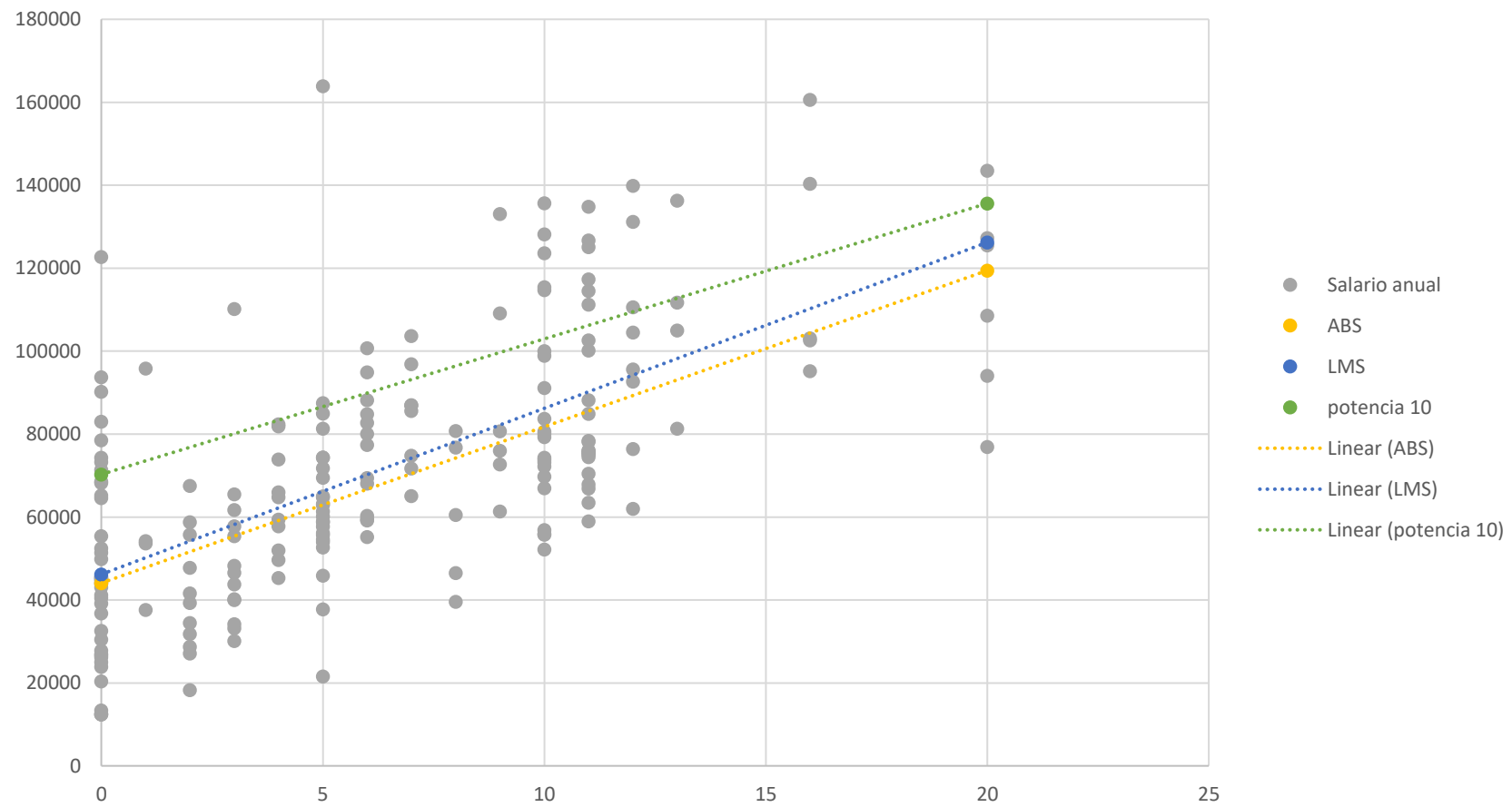
	X3	Y	Ymodel o A+bx3	error
Emple ado	Experienc ia previa	Salario anual		
1	5	57700		Ymodel o-y
2	12	76400		
3	0	44000		
4	2	41600		
5	5	163900		
6	9	72700		
7	6	60300		
8	11	62500		



Comencemos con algo muy sencillo Solo con X3= Experiencia previa y Y= salario anual



Ajustará bien una recta? ensayemos



Que diferencias observa entre los 3 ajustes?

Estadísticos de los errores

	ABS (o MAE)	LMS (o MSE)	LM- 10
Desviación estándar	15 238	14 406	15 152
max error	100 966	97 708	77 276
min error	0	386	876
rango	100 966	97 322	76 400

Que observa?

Usualmente, se considera que la mejor ecuación de regresión es menor suma del cuadrado de los errores residuales.

Error residual de observación i: diferencia entre los valores de Y modelo y Y observados: $e(i) = \text{Residual}(i) = Y \text{ observado}(i) - Y \text{ modelo}(i)$

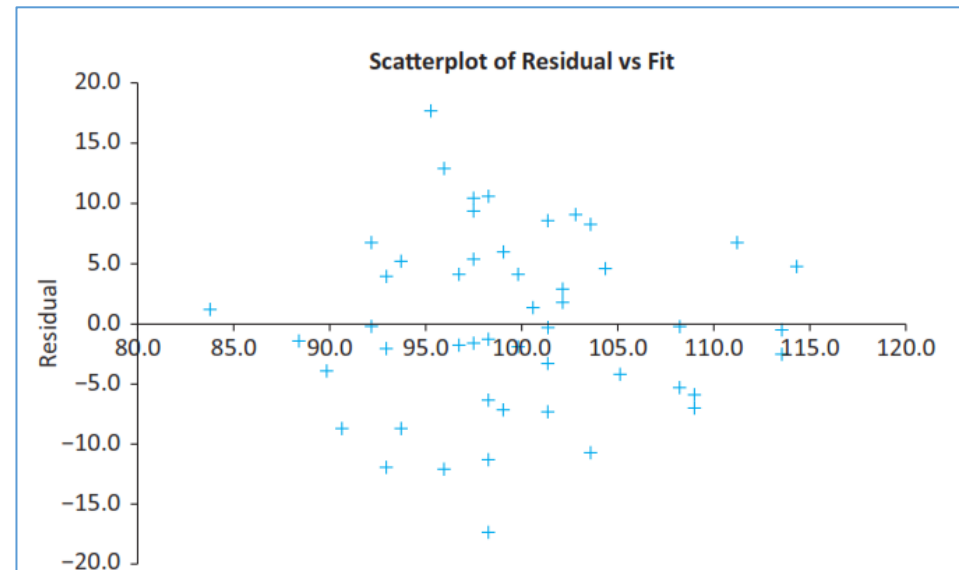
Por ahora, solo informemos que se obtiene (más adelante profundizaremos):

$$B = \rho_{XY} \frac{s_Y}{s_X} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$A = \bar{Y} - B\bar{X}$$

Figura interesante de análisis: Valores ajustados vs errores

Para que la regresión sea confiable, lo ideal es que no se vea un patrón en los residuales sino que sea aleatorio



Error estándar de los residuales s_e

$$s_e = \sqrt{\frac{\sum e_i^2}{n - 2}}$$

Media de los residuales: 0.

Se usa $n-2$ porque hay que estimar 2 parámetros: A y B

Suponiendo normalidad en los residuales, puede decirse que 2/3 partes de los valores observados están a **un error estándar** del correspondiente Y ajustado.

Además, alrededor del 95% de los valores Y observados están dentro de 2 errores estándar de los valores Y ajustados correspondientes

- **Cómo saber si ese valor es pequeño o es un buen nivel de precisión?**

Se calcula la desviación estándar de la variable original Y (S_y)
(independiente, como si no tuviera influencia de ninguna variable)

Si $\underline{S_e} \ll \underline{S_y}$ (*mucho menor*) quiere decir que la regresión **si aporta significativamente** a la predicción de Y

R²: coeficiente de determinación

Expresa la fracción de variación de Y original explicada por la regresión:

Proporción de la variación total en los valores de la variable Y que se puede explicar mediante los cambios en los valores de x.

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

Si R² se acerca a 1, significa que los residuales son pequeños.

Si R²= 0,64 por ejemplo, significa que la regresión con X como única variable explicativa explica el 64% de la variación de Y.

El punto es que si una variable X puede explicar un gran porcentaje de la variable Y, las dos variables son muy correlacionadas (ya sea en dirección positiva o negativa)

R: Coeficiente de correlación entre Y observado y Y modelo

- En la regresión simple, la correlación entre Y_{obs} y Y_{modelo} es la misma entre Y_{obs} y X
- $R = \sqrt{R^2}$ con el signo de la pendiente
- Puede tener cualquier valor entre -1 y 1; pero su valor absoluto debe ser cercano a 1 para poder considerar un buen ajuste (preferiblemente superior a 0.8).

Regresión Múltiple Lineal

- Es de la forma:
- $\hat{y} = A + B_1X_1 + B_2X_2 + \dots + B_nX_n$
- Donde A es la intercepción y cada coeficiente B_i equivale al cambio promedio en la variable y debido a un cambio en una unidad en la variable X_i , mientras las demás variables independientes permanecen constantes.

Error estándar de los residuales en regresión múltiple

$$s_e = \sqrt{\frac{\sum e_i^2}{n - k - 1}}$$

Se k numero de variables independientes

los residuales tienen media 0.

Considerando que los residuales siguen una distribución normal: Puede decirse que 2/3 partes de los valores observados están a **un error estándar** del correspondiente Y ajustado.

R^2 y R

R^2 : El es porcentaje de variación de la variable dependiente, explicada por el conjunto de las variables explicativas

$R = \sqrt{R^2}$: correlación entre los valores Y_{obs} y Y_{model}

Usualmente R^2 mejora cuando se introduce un mayor numero de variables al análisis. Esto no es necesariamente bueno porque se puede estar inflando ilógicamente el modelos: debe ser monitoreado con lógica.

Variables *dummy*

- Permiten la inclusión de Variables categóricas (ejemplo mujer y hombre) en un análisis de regresión
- Se le puede asignar 2 valores: 0 y 1
 - =1 si la observacion esta en la categoria X,
 - 0, en caso contrario
- Si la variable tiene más observaciones, por ejemplo: niño, joven, adulto se usan 2 variables dummy. La restante se llama de referencia. Si las dos tienen el valor 0, ya se sabe que la de referencia =1
- Ejemplo: posible discriminación de genero en salarios. Ver **Archivo Discriminacionsalarial.csv. Hacer actividad**

Análisis simple genero-Salario

- Si la variable categórica tiene n posible valores, solo use n-1 dummies
- Por ejemplo si n=2 (hombre y mujer) solo es necesario usar 1 variable dummy. Si no es mujer=1, es hombre=0 (hombre seria la categoría de referencia)
- En el ejemplo de Salarios:

$$\text{Salario} = 78784,70 - 12874\text{Mujer} \quad (R^2=0,044 \text{ muy bajo!})$$

- Es decir, que si es hombre, gana en promedio \$78784,70, si es mujer \$65910
- Diferencia \$12874 en promedio.

Agregando Nuevas variables:

Dado que R^2 es muy bajo, agreguemos experiencia previa X_3 y en la empresa X_4

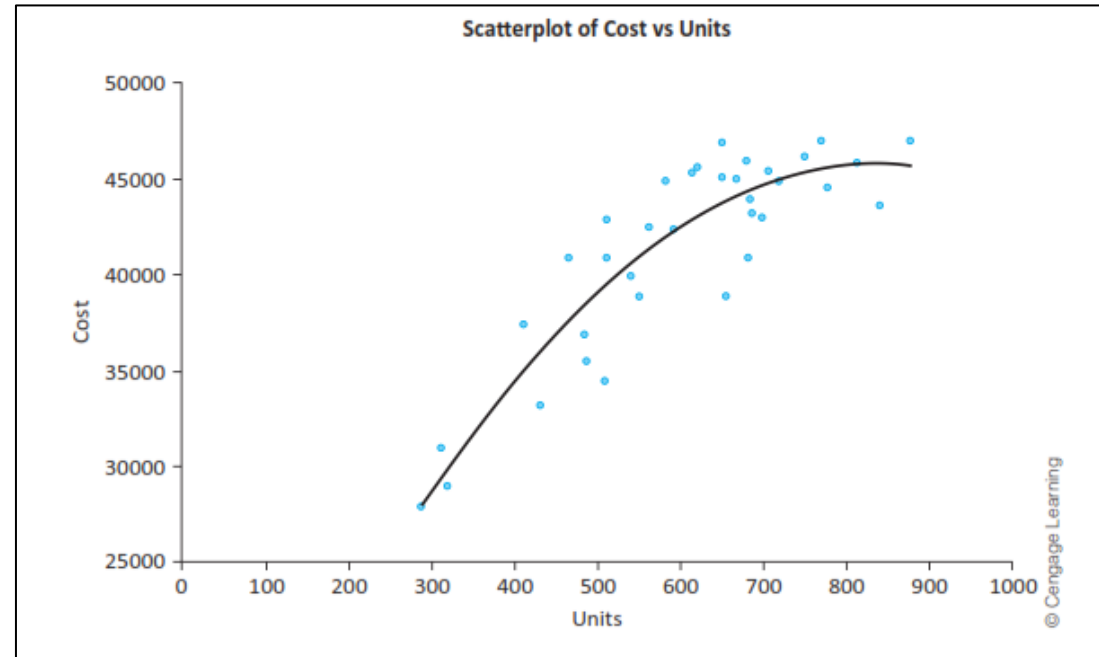
- continuar actividad....

Regresiones no lineales mediante linealización

- Algunas relaciones no lineales pueden linealizarse mediante transformaciones matemáticas para aplicarles el mismo tratamiento anterior.
- Se puede transformar la variable dependiente Y o una explicatoria X o ambas
- Por ejemplo:

Regresión	Forma funcional
Exponencial	$y = Ae^{Bx}$
Logarítmica	$y = A + B \ln x$
Polinomial de segundo grado o cuadrática	$y = Cx^2 + BX + A$

Ejemplo: costo vs unidades

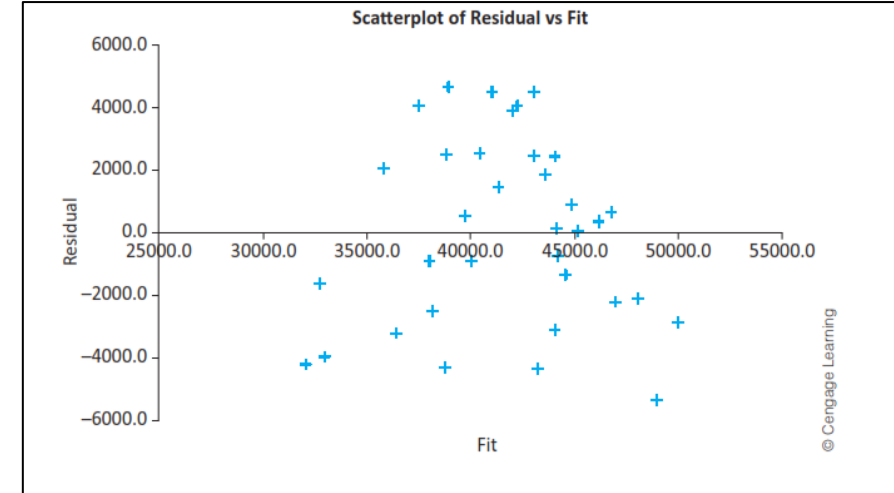


A partir de datos de unidades de energía y costos se hizo la siguiente Regresión lineal:

$$\text{costo} = 23651 + 30,53 \text{ Unidades} \quad (R^2 = 73.6\% \text{ relativamente baja})$$

Quizás la línea recta no es una buena representación. El dibujo sugiere economías de escala!

Además la gráfica de dispersión de Y_{ajus} vs residuales :



Sugiere un patrón, los residuos en el centro son positivos pero en la derecha e izquierda son negativos: esto sugiere cierta relación **no lineal (en este caso** parábola, es decir, una relación cuadrática con el cuadrado de unidades incluido en la ecuación)

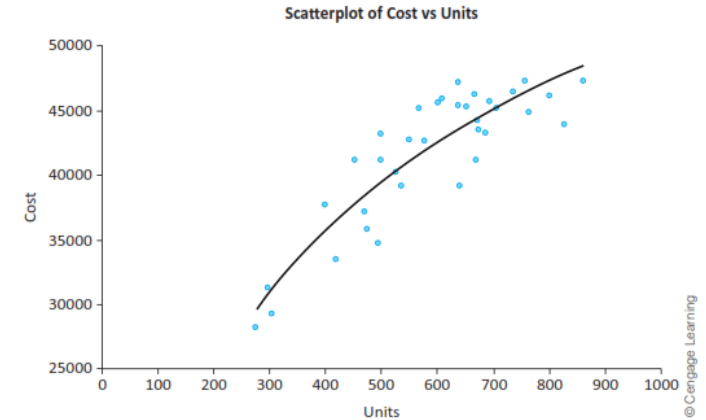
Se crea una nueva variable $(\text{Unidades})^2$ y se analiza una relación lineal múltiple, la cual correspondió a:

$$\text{Costo} = 5793 + 98,35\text{Unidades} - 0,06\text{Unidades}^2 \quad (R^2=82,2\%: \text{mucho mejor})$$

Probemos con otra función no lineal: $\log(X)$.

La regresión dio:

$$\text{Costo} = -63993 + 166.54 \log(\text{Unidades})$$



($R^2=79,8\%$ no es tan bueno como el ajuste cuadrático)

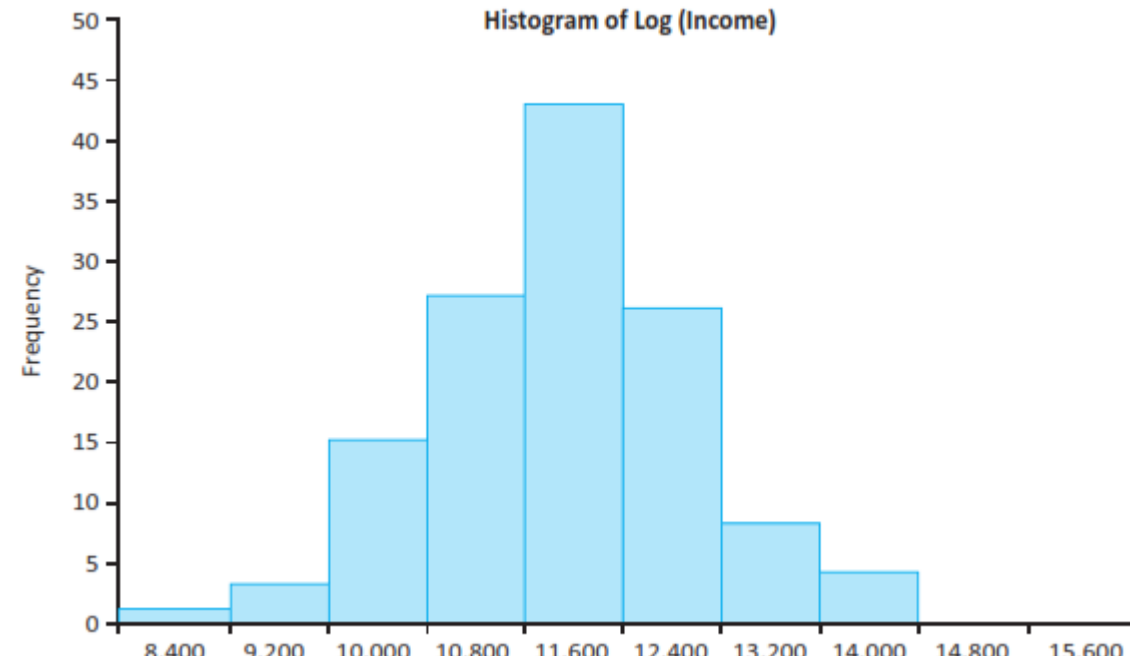
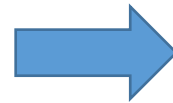
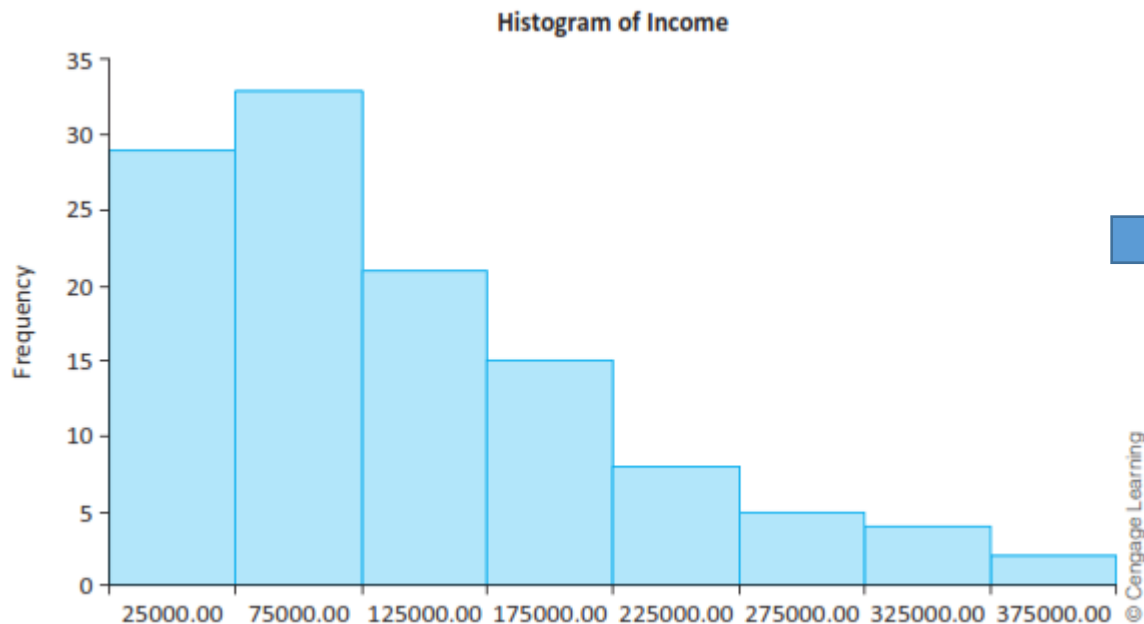
Sin embargo, la ventaja del Logarítmico es que la ecuación es más fácil de interpretar.

Con solo una variable explicativa, se puede interpretar su coeficiente como sigue:

Pero ojo: incrementar un 1 % de unidades de 700 A 707 (7 unidades) es mayor que de 600 a 606 (6 unidades) pero implica el mismo costo: **costo marginal decreciente**.

Se puede hacer la transformación de $\text{Log}(Y)$

- Útil cuando la distribución of Y es *sesgada hacia la derecha*
- El efecto de la trasformacion es dispersar afuera los valores pequeños y comprimir los valores grandes, haciendo la distribución más simétrica.



Regresión exponencial

$$Y = ae^{bx}$$

$$\ln(Y) = \ln(ae^{bx})$$

$$\ln(Y) = \ln(a) + \ln(e^{bx})$$

$$\ln(Y) = \ln(a) + bx$$

$$Z = c + bx$$



Equivalente a una lineal

$$\text{Donde } Z = \ln Y \quad c = \ln(a)$$

Regresión exponencial múltiple

Ejemplo:

exponencial :

$$y = (e^{B_1x_1})(e^{B_2x_2})...(e^{B_nx_n})...$$

- Donde, $B_1 \dots B_n$ son constantes

Un ejemplo simple contagio

dia	contagiados reportados	ln(contagiados)
1	0	0
2	0	0
3	1	0
4	1	0
5	2	0.693147
6	2	0.693147
7	2	0.693147
8	2	0.693147
9	2	0.693147
10	3	1.098612
11	3	1.098612
12	3	1.098612
13	4	1.386294



$$\ln(\text{contagiados}) = \ln(a) + bt$$

Por regresión lineal

$$\ln(a) = -0.0681$$

$$b = 0.1$$

Que pasará el día 70?

Aproximadamente:

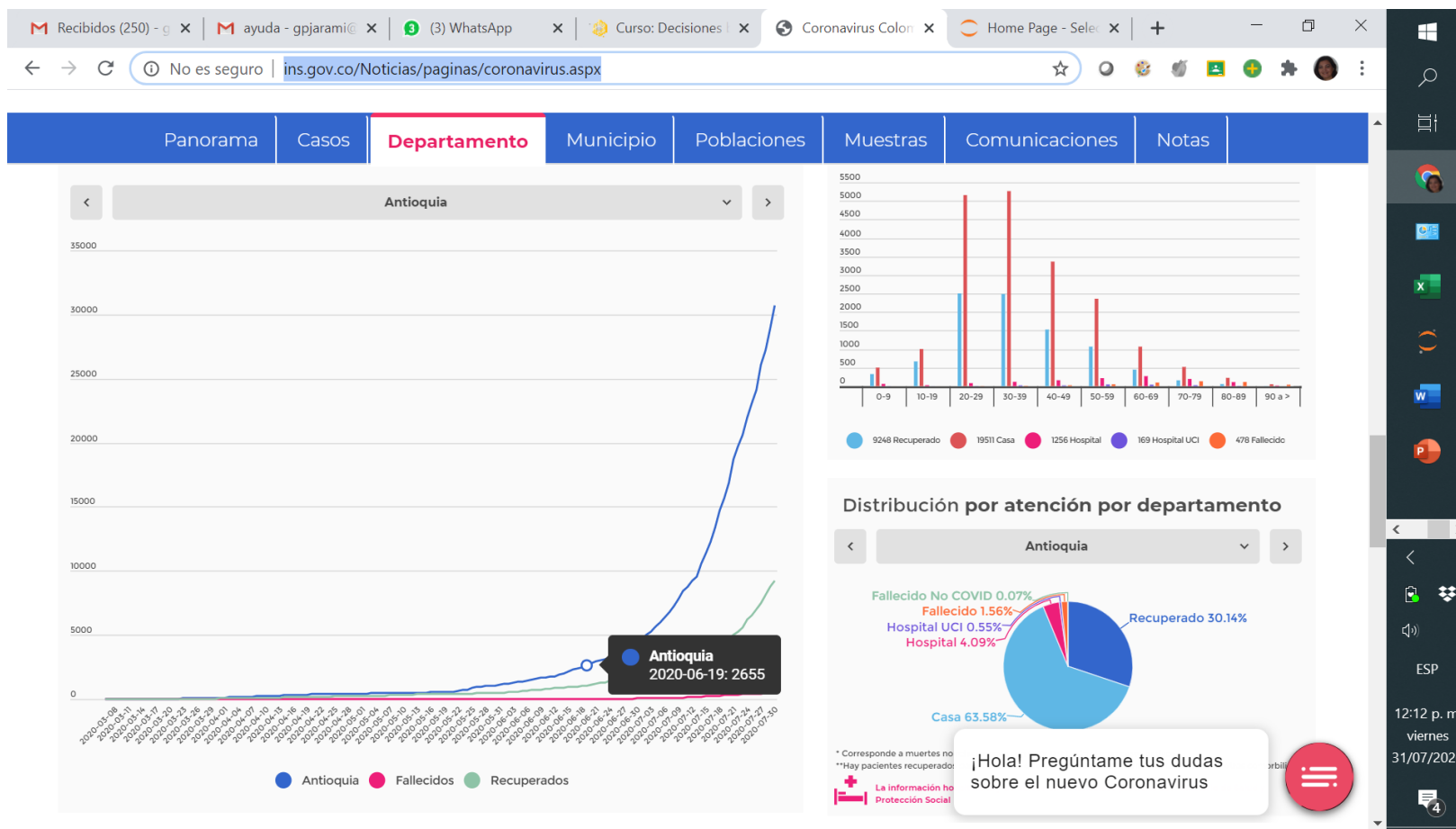
$$-0.0681 + 70 \cdot 0.1 = 7.09$$

$$Y = \text{Exp}(7.09) = 1202 \text{ contagiados}$$

Datos oficiales covid19 Colombia

<https://www.ins.gov.co/Noticias/paginas/coronavirus.aspx>

<https://www.datos.gov.co/Salud-y-Protecci-n-Social/Casos-positivos-de-COVID-19-en-Colombia/gt2j-8ykr/data>



How a virus with a reproduction number (R_0) of 2 spreads

