ESTADISTICA TECNICA

MODELOS LINEALES DE REGRESION

CLASE I: REGRESION LINEAL SIMPLE

CASO DE DISCUSION I

Problema RS7

Una consultora de RRHH desea evaluar si el promedio de las calificaciones de alumnos universitarios influye en el salario inicial que perciben (K$/año). Para ello efectúa una encuesta a 17 graduados de Ingeniería de la UBA.

PROMEDIO %	SALARIO k$
68	18,5
78	20
86	21,1
94	22,4
76	21,2
64	15
74	18
55	25
64	18,8
72	15,7
58	14,4
60	15,5
75	17,2
90	16,4
76	19
69	17,2
62	16,8

Objetivo: encontrar un modelo que se ajuste a los datos y poder efectuar predicciones de la variable explicada (y) en función de los valores de la variable explicativa (x), o bien dado el modelo por otra ciencia validarlo y estimar las constantes involucradas

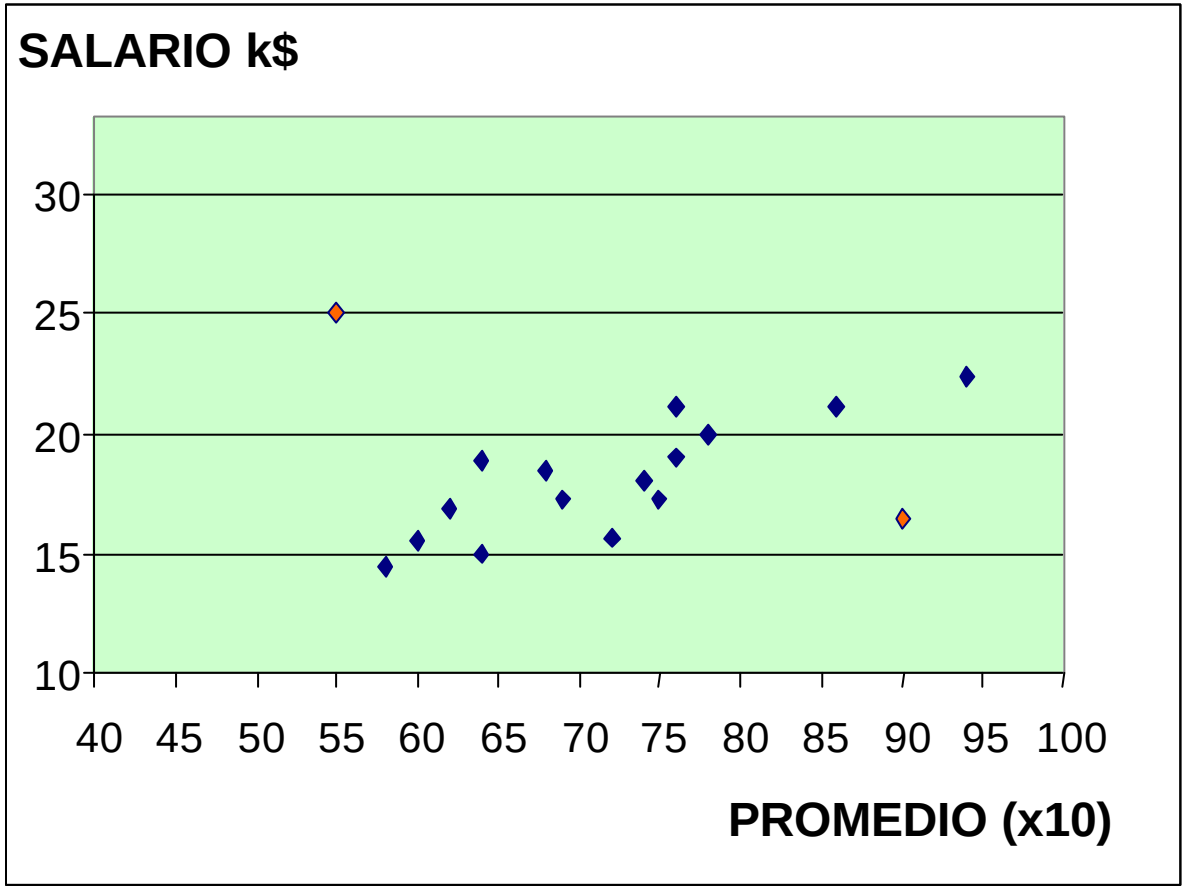
$$\tilde{y} = f(x) + \tilde{e}$$

$\tilde{y}$: Variable explicada o de respuesta. Se considera aleatoria.

x : Variable explicativa (o independiente). Se considera como no aleatoria.

$\tilde{e}$: Perturbación o error. Es una variable aleatoria

PROMEDIO %	SALARIO k$
68	18,5
78	20
86	21,1
94	22,4
76	21,2
64	15
74	18
55	25
64	18,8
72	15,7
58	14,4
60	15,5
75	17,2
90	16,4
76	19
69	17,2
62	16,8



Hay dos datos sospechosos que por el momento excluiremos.

MODELOS LINEALES

$$\tilde{y} = f(x) + \tilde{e}$$

$$f(x) = b_0 + b_1 X$$

$$f(x) = b_0 + b_1 X^2$$

$$f(x) = b_0 + b_1 X^r$$

$$f(x) = b_0 + b_1 \operatorname{sen}(x)$$

$$f(x) = a X^b \qquad Y = a X^b$$

$$LnY = \ln a + b \ln X$$

La variable x puede estar bajo formas no lineales

El modelo debe ser lineal en los parámetros

MODELOS LINEALES

$$\tilde{y}_i = b_o + b_1 x_i + \tilde{e}_i$$

Supuestos del modelo:

$$E(\tilde{e}_i) = 0$$ Ausencia de vicio

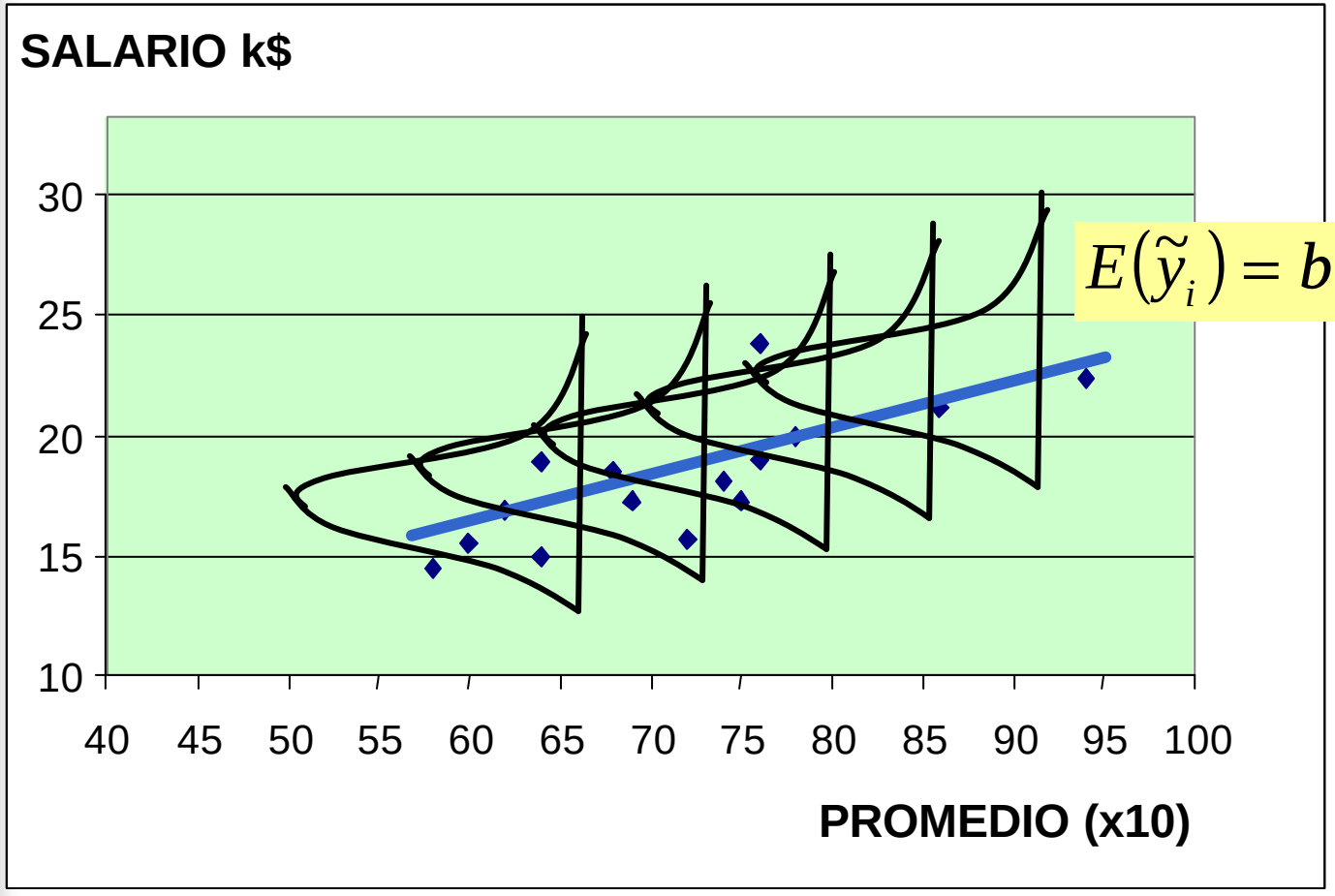
$$Var(\tilde{e}_i) = s^2 \quad \forall \quad i$$ Homocedasticidad

$$Cov(\tilde{e}_i ; \tilde{e}_j) = 0 \quad \forall \quad i \neq j$$ Ausencia de autocorrelación

De lo que se sigue:

$$E(\tilde{y}_i) = b_o + b_1 x_i \quad Var(\tilde{y}_i) = s^2 \quad \forall \quad i$$

RECTA DE REGRESION POBLACIONAL



$$E(\tilde{y}_i) = b_o + b_1 x = m_i$$

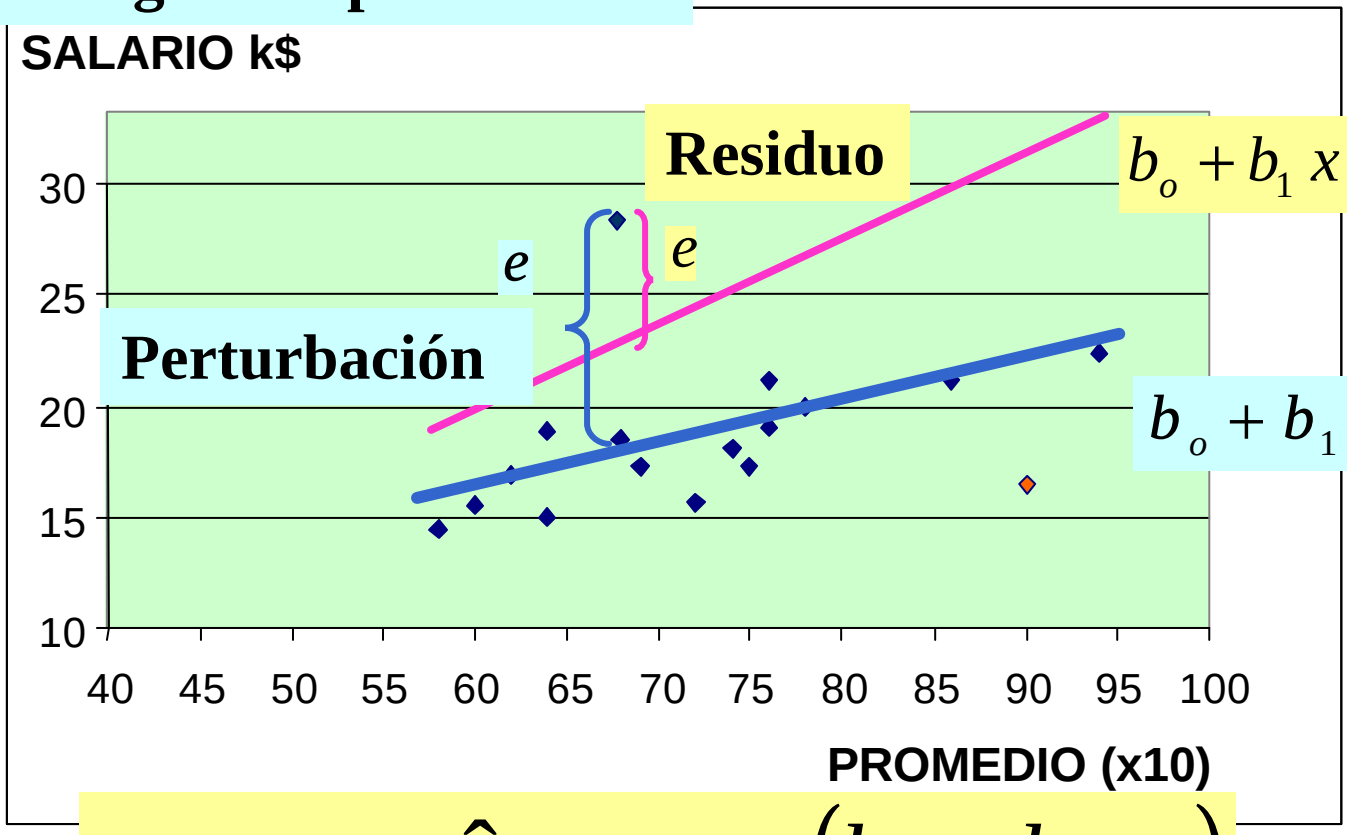
ESTIMACION DE LOS PARAMETROS

$$\tilde{y} = b_o + b_1 x + \tilde{e}$$

Recta de regresión poblacional

$$\hat{y} = b_o + b_1 x$$

Recta de regresión muestral



$$e_i = y_i - \hat{y}_i = y_i - (b_o + b_1 x_i)$$

ESTIMACION DE LOS PARAMETROS

$$e_i^2 = (y_i - \hat{y}_i)^2 = [y_i - (b_o + b_1 x_i)]^2$$

Aplicamos los mínimos cuadrados de Gauss:

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

MINIMIZAR

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

$$\frac{\partial Q}{\partial b_o} = 0 \qquad \frac{\partial Q}{\partial b_1} = 0$$

ESTIMACION DE LOS PARAMETROS

$$\frac{\partial Q}{\partial b_o} = 0 \qquad \frac{\partial Q}{\partial b_1} = 0$$

$$b_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y}_i)(x_i - \bar{x}_i)}{\sum_{i=1}^{n} (x_i - \bar{x}_i)^2} = \frac{S_{xy}}{S_{xx}}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Formulas de trabajo

$$S_{xy} = \sum_{i=1}^{n} x_i \, y_i - n \, \bar{x} \, \bar{y}$$

$$S_{xx} = \sum_{i=1}^{n} x_i^{\,2} - n \, \bar{x}^2$$

$$S_{yy} = \sum_{i=1}^{n} y_i^{\,2} - n \, \bar{y}^2$$

Ojo con los errores de redondeo!!

USAR CALCULADORA EN MODO SD

$$S_{xy} = \sum_{i=1}^{n} x_i\, y_i - n\,\bar{x}\,\bar{y}$$

$$S_{xx} = \sum_{i=1}^{n} x_i^{\,2} - n\,\bar{x}^{2}$$

$$S_{yy} = \sum_{i=1}^{n} y_i^{\,2} - n\,\bar{y}^{2}$$

	x	y			
	PROMEDIO %	SALARIO k$	x*y	x2	y2
	68	18,5	1258	4624	342,25
	78	20	1560	6084	400
	86	21,1	1814,6	7396	445,21
	94	22,4	2105,6	8836	501,76
	76	21,2	1611,2	5776	449,44
	64	15	960	4096	225
	74	18	1332	5476	324
	64	18,8	1203,2	4096	353,44
	72	15,7	1130,4	5184	246,49
	58	14,4	835,2	3364	207,36
	60	15,5	930	3600	240,25
	75	17,2	1290	5625	295,84
	76	19	1444	5776	361
	69	17,2	1186,8	4761	295,84
	62	16,8	1041,6	3844	282,24
suma	1076	270,8	19702,6	78538	4970,12
promedio	71,73	18,05			
datos	15				

Sxy	277,213
Sxx	1352,93
Syy	81,2773

b1	0,2049
bo	3,3553

R2	0,6988

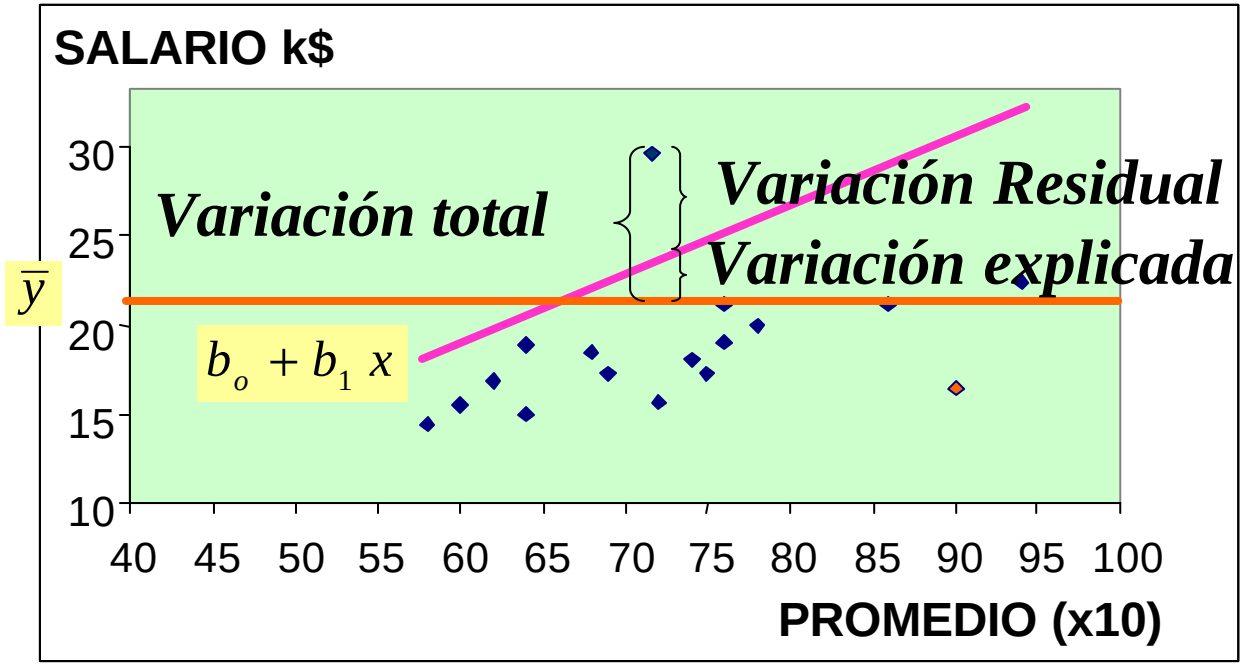
$$\hat{y} = 3.3553 + 0.2049x$$

VALIDACION DEL MODELO

PRIMER CRITERIO : coeficiente de determinación

$$R^2 = \frac{Variación\ explicada}{Variación\ total} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{Q}{S_{yy}}$$

Variación total = Variación explicada + Variación Residual



VALIDACION DEL MODELO

$R^2 \geq 0.8$ **Procesos físicos e industriales**

$R^2 \geq 0.7$ **Economía**

$R^2 \geq 0.5$ **Sociología**

Formula de trabajo

$$R^2 = \frac{S_{xy}^{\ 2}}{S_{xx}\, S_{yy}}$$

Para nuestro caso:

Sxy	277,213
Sxx	1352,93
Syy	81,2773

R2	0,6988

ESTIMACION DE LA VARIANZA RESIDUAL

$$Var(\tilde{e}_i) = s^2 \quad \forall \quad i$$

Se estima por:

$$S^2 = \frac{Q}{n-2}$$

$$siendo:$$

$$Q = S_{yy} - b_1 S_{xy}$$

$$\frac{n\ S^2}{s^2} \approx \quad c^2_{n-2}$$

Para nuestro caso:

Q	**24,4769**
s2	**1,88284**

VALIDACION DEL MODELO

SEGUNDO CRITERIO:

Prueba de significación del coeficiente de regresión

$$H_o : b_1 = 0$$

Si se rechaza Ho concluímos que x e y tienen un cierto grado de asociación LINEAL

Requerimos un supuesto teórico: $\tilde{e}_i \approx N(0; s^2)$

Se demuestra que : $b_1 \approx N\left(b_1; \frac{s^2}{S_{xx}}\right)$

Estadístico de prueba:

$$\frac{b_1 - b_1}{S_{b_1}} \approx t_{(n-2)} \qquad con \quad S_{b_1} = \frac{s}{\sqrt{S_{xx}}}$$

VALIDACION DEL MODELO

Para nuestro caso:

$$H_o : b_1 = 0 \qquad H_1 : b_1 > 0$$

A UNA SOLA COLA

$$CR \qquad \frac{b_1}{S_{b_1}} \geq \qquad t_{(n-2)\quad 1-a}$$

sb1	**0,0373**

$$\frac{b_1}{S_{b_1}} = \frac{0.2049}{0.0373} = 5.4925$$

$$t_{13\quad 0.90} = 1.77$$

INFERENCIA

INFERENCIA SOBRE LA MEDIA

$$E(\tilde{y} / x_o) = b_0 + b_1 x_0$$

Estadístico

$$\hat{y} / x_o = b_0 + b_1 x_0$$

$$\approx N \left(b_o + b_1 x_o ; s \sqrt{\left(\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum (xi - \bar{x})^2} \right)} \right)$$

Entonces como se desconoce la varianza, la expresión del intervalo es :

$$b_0 + b_1 x_0 \pm t_{n\,1-\frac{a}{2}} \sqrt{s^2 \left(\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum (xi - \bar{x})^2} \right)}$$

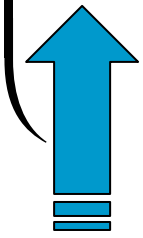
PREDICCION

PREDICCION SOBRE UN VALOR EN PARTICULAR

$$\tilde{y} / x_o = b_0 + b_1 x_0 + \tilde{e}$$

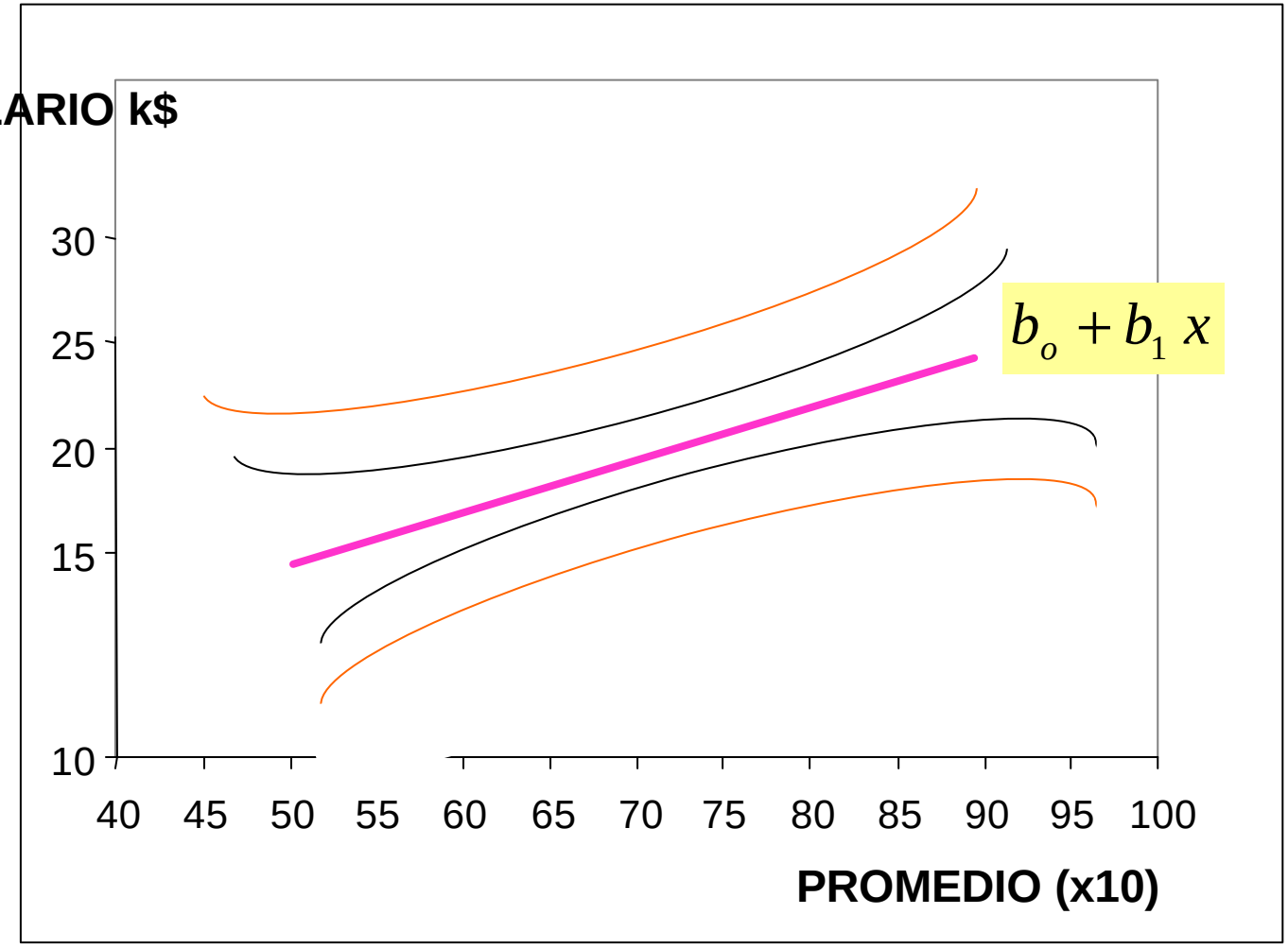
La expresión del intervalo es :

$$b_0 + b_1 x_0 \pm t_{n\,1-\frac{a}{2}} \sqrt{s^2 \left(1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum (xi - \bar{x})^2} \right)}$$



Refleja la varianza de la variable Y, además de la muestra.

INTERVALOS DE CONFIANZA Y PREDICCION



Para nuestro caso:

Pronosticar el salario que puede tener un alumno con promedio 9 con un NC de 90%

$$x_0 = 90 \qquad b_0 + b_1 x_0 = 21.7691$$

$$t_{n\,1-\frac{a}{2}} = 1.7709$$

$$\sqrt{s^2\left(1+\frac{1}{n}+\frac{(x_o-\bar{x})^2}{\sum(xi-\bar{x})^2}\right)} = 1.5725$$

$$21.7691 \pm 2.7848$$

Análisis de los datos arbitrariamente sospechosos:

(55; 25)

(90; 16.4)

Debemos construir los intervalos de predicción y verificar si los valores de los salarios están contenidos.

Para Xo= 55 se obtiene el intervalo [11.8823.; 17.3671]

Para Xo = 90 se obtiene el intervalo [19.0113; 24.5809]

Concluimos que los datos son estadísticamente sospechosos.

CASO DE DISCUSION I I

Un economista está investigando la relación entre la inflación y el desempleo en la Argentina durante la Convertibilidad, para ello adopta el modelo de Phillips

$$\tilde{y} = w + \frac{q}{x} + \tilde{e}$$

Donde Y es el índice de inflación , x la tasa de desempleo y θ y ω los parámetros.
Releva la siguiente información:

Año	1992	1993	1994	1995	1996	1997	1998	1999
Inflacion	10,9609	5,8033	3,5802	0,3584	-0,2470	-0,1400	0,0363	-2,3207
Desempleo	6,95	9,6	11,45	17,5	17,25	14,9	12,8	14,15

a) Estimar los parámetros del modelo y evaluarlo por los dos criterios básicos.
b)¿Cuál sería el índice de inflación que se tendría en un año en el que el desempleo fuese del 10%?. NC= 95%

Se resuelve haciendo el cambio de variable:

$$x' = \frac{1}{x}$$

La tabla de datos transformada es:

año	inflacion	desempleo	1/desempleo
1992	10,9609	6,9500	0,1439
1993	5,8033	9,6000	0,1042
1994	3,5802	11,4500	0,0873
1995	0,3584	17,5000	0,0571
1996	-0,2470	17,2500	0,0580
1997	-0,1400	14,9000	0,0671
1998	0,0363	12,8000	0,0781
1999	-2,3207	14,1500	0,0707

CASO DE DISCUSION I I I

En una gran carpintería se registraron los costos (hh) en función de la superficie de las piezas en m2.

A) Investigar la asociación lineal de las variables considerando intercepto nulo.

B) Pronostique el costo para una pieza de 1.2 m2 (NC= 90%)

Superficie	0.45	1.4	2.6	0.7	1.3
Costo	1.82	2.58	3.21	2.03	2.36

INTERCEPTO NULO

$$b_1 = \frac{\sum_{i=1}^{n} x_i (y_i - b_0)}{\sum_{i=1}^{n} x_i^{\,2}}$$

$$S^2 = \frac{Q}{n-1}$$

siendo :

$$Q = \sum (y_i - b_0)^2 - b_1 \sum x_i (y_i - b_0)$$

$$S^2{}_{b1} = \frac{S^2}{\sum x_i^{\,2}}$$

$$t_{n\,1-\frac{a}{2}}$$

$$S^2 \hat{y} = \frac{S^2 \, x_o^{\,2}}{\sum x_i^{\,2}}$$

$$ojo!\, en\ este\ caso\, n = n-1$$