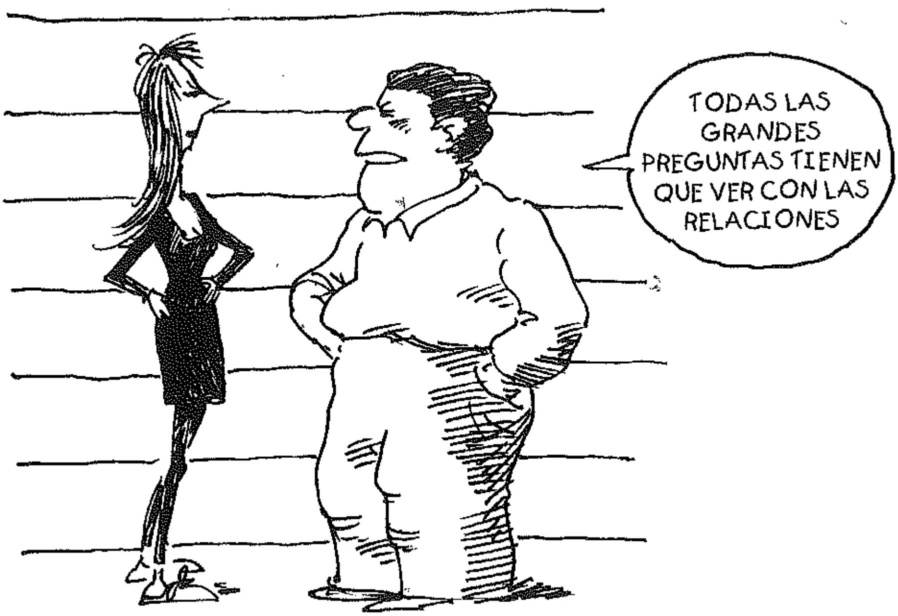
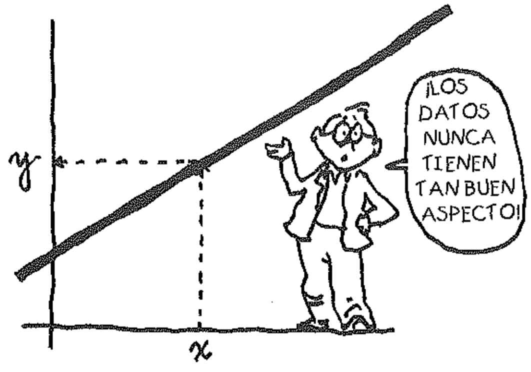
♦ Capítulo 11 ♦
REGRESIÓN

HASTA AHORA, HEMOS ESTUDIADO **UNA SOLA VARIABLE** CADA VEZ, TANTO SI SE TRATABA DE UNA POBLACIÓN DE PERSONAS A LAS QUE SE LES ADMINISTRABA UNA PÍLDORA, O DE UNA DE PEPINILLOS, COMO DE COCHES ACCIDENTADOS. EN ESTE CAPÍTULO, APRENDEREMOS A RELACIONAR **DOS VARIABLES**: DADOS LOS PESOS DE LOS 92 ESTUDIANTES DEL CAPÍTULO 2, NOS PREGUNTAREMOS QUÉ RELACIÓN TIENE EL PESO CON LA **ESTATURA** DE LOS ESTUDIANTES.

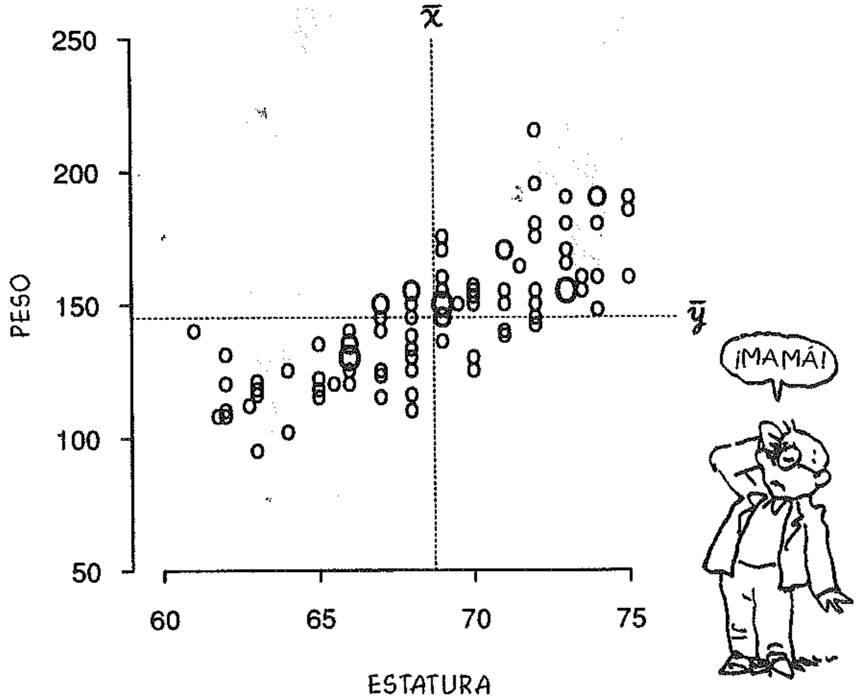


ÉSTE ES UN EJEMPLO DE UNA AMPLIA SERIE DE PREGUNTAS IMPORTANTES: ¿**PUEDE** PREDECIRSE **LA ESPERANZA** DE VIDA MIDIENDO LA **TENSIÓN ARTERIAL**? ¿**LAS NOTAS DE LA SELECTIVIDAD** PREDICEN EL **COMPORTAMIENTO ACADÉMICO EN LA UNIVERSIDAD**? ¿LEER LIBROS DE ESTADÍSTICA TE CONVIERTE EN **MEJOR PERSONA**?

SEGURAMENTE, EN CLASE DE MATEMÁTICAS HAS APRENDIDO A VER LAS RELACIONES REPRESENTADAS EN GRÁFICOS. DADA LA x PUEDES PREDECIR LA y. PERO, EN ESTADÍSTICA, ¡LAS COSAS NUNCA SON TAN SENCILLAS! SABEMOS (O CREEMOS SABER) QUE LA ESTATURA INFLUYE EN EL PESO, PERO NO SE TRATA DE LA ÚNICA INFLUENCIA. EXISTEN OTROS FACTORES COMO EL SEXO, LA EDAD, LA COMPLEXIÓN FÍSICA Y **LA VARIABLE ALEATORIA.**



EN ESTE CAPÍTULO ETIQUETAREMOS LOS DATOS RELATIVOS AL PESO CON LA y, Y LOS RELATIVOS A LA ESTATURA CON LA x. ASÍ (x_i, y_i) ES LA ESTATURA Y EL PESO DEL ESTUDIANTE i. REPRESENTAMOS LOS PUNTOS (x_i, y_i) EN UN DIAGRAMA BIDIMENSIONAL QUE RECIBE EL NOMBRE DE **GRÁFICO DE DISPERSIÓN DE PUNTOS.**



(ALGUNOS PUNTOS SON MÁS GRANDES PORQUE REPRESENTAN A DOS O TRES ESTUDIANTES DEL MISMO PESO Y ESTATURA.)

¿PODEMOS PREDECIR EL PESO y DE UN ESTUDIANTE A PARTIR DE SU ESTATURA x?

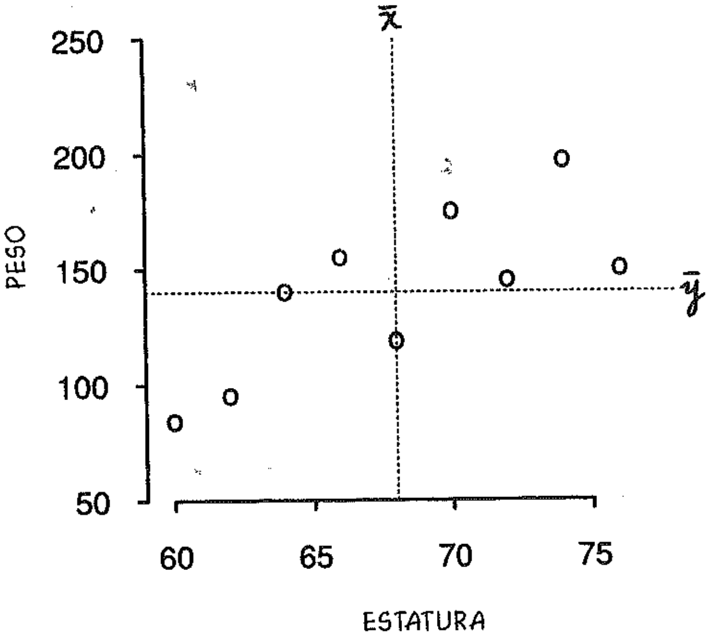
El análisis de regresión

AJUSTA UNA LÍNEA RECTA EN ESTE DESORDENADO GRÁFICO DE PUNTOS.
x RECIBE EL NOMBRE DE **VARIABLE INDEPENDIENTE** O **REGRESORA O PREDICTORA**. E y ES LA **VARIABLE DEPENDIENTE O RESPUESTA**. LA **RECTA DE REGRESIÓN O DE PREDICCIÓN** TIENE LA FORMA:

$$y = a + bx$$

PARA ILUSTRAR EL EJEMPLO DE AJUSTE DE LA RECTA, UTILIZAREMOS UN CONJUNTO MÁS REDUCIDO DE DATOS FICTICIOS CON SÓLO **NUEVE** PAREJAS DE PESOS Y ESTATURAS DE ESTUDIANTES.

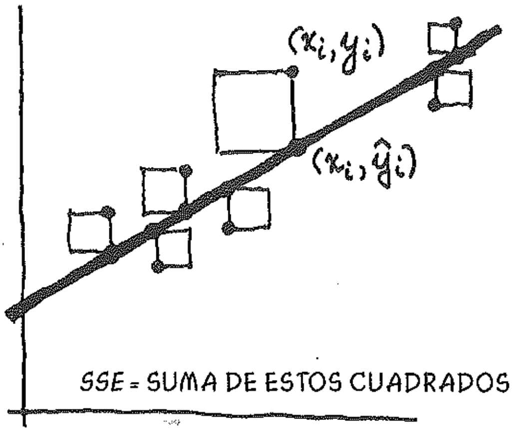
ESTATURA	PESO
60	84
62	95
64	140
66	155
68	119
70	175
72	145
74	197
76	150



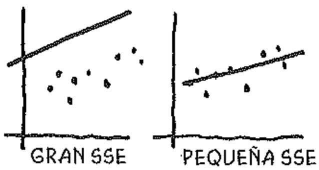
ENTONCES, ¿CÓMO PODEMOS CONSEGUIR LA MEJOR RECTA DE AJUSTE?

LA IDEA CONSISTE EN **MINIMIZAR** LA DISTANCIA TOTAL DE LOS VALORES *Y* A LA RECTA. IGUAL QUE CUANDO DEFINÍAMOS LA VARIANZA, BUSCAMOS LAS **DISTANCIAS AL CUADRADO** DE *y* CON LA RECTA Y LAS SUMAMOS PARA OBTENER LA **SUMA DE LOS ERRORES CUADRÁTICOS** (*SSE*):

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



ES UNA MEDIDA AGREGADA DE CUÁNTO PUEDEN DIFERIR LAS «PREDICCIONES y_i», LLAMADAS $\hat{y}_i$, CON RESPECTO A LOS VALORES REALES y_i.



La recta de regresión o recta de mínimos cuadrados

ES LA RECTA CON LA **MÍNIMA** *SSE*

NOTA HISTÓRICA: ¿POR QUÉ DENOMINAMOS ESTE PROCESO **ANÁLISIS DE REGRESIÓN**? A PRINCIPIOS DE SIGLO, EL ESTUDIOSO DE LA GENÉTICA **FRANCIS GALTON** DESCUBRIÓ UN FENÓMENO LLAMADO **REGRESIÓN A LA MEDIA**. BUSCANDO LEYES DE HERENCIA GENÉTICA, DESCUBRIÓ QUE LA ESTATURA DE LOS HIJOS SOLÍA SER UNA **REGRESIÓN** A LA ESTATURA MEDIA POBLACIONAL, EN COMPARACIÓN CON LA ESTATURA DE SUS PADRES. LOS PADRES ALTOS SOLÍAN TENER HIJOS ALGO MÁS BAJOS, Y VICEVERSA. GALTON DESARROLLÓ EL ANÁLISIS DE REGRESIÓN PARA ESTUDIAR ESTE FENÓMENO, AL QUE SE REFIRIÓ DE MANERA OPTIMISTA COMO «REGRESIÓN A LA MEDIOCRIDAD».

PARA NO ANDARNOS POR LAS RAMAS, PRESENTAMOS SIN MÁS EXPLICACIONES LA FÓRMULA DE LA REGRESIÓN LINEAL: ES LIADITA PERO CALCULABLE.

$$y = a + bx$$

DONDE

$$b = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Y

$$a = \bar{y} - b\bar{x}$$

AQUÍ $\bar{x}$ E $\bar{y}$ SON LAS MEDIAS DE $\{x_i\}$ Y $\{y_i\}$ RESPECTIVAMENTE.

COMO ESTAS EXPRESIONES VOLVERÁN A SALIR, LAS ABREVIAREMOS:

$$\left.\begin{aligned} ss_{xx} &= \sum_{i=1}^{n}(x_i-\bar{x})^2 \\ ss_{yy} &= \sum_{i=1}^{n}(y_i-\bar{y})^2 \end{aligned}\right\}$$

LA SUMA DE LOS CUADRADOS ALREDEDOR DE LA MEDIA MIDE LA DISPERSIÓN DE x_i Y DE y_i.

$$ss_{xy} = \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

EL PRODUCTO CRUZADO DETERMINA (CON ss_{xx}) EL COEFICIENTE b.

ESTE ES EL CÁLCULO TOTAL DE LOS VALORES FICTICIOS:

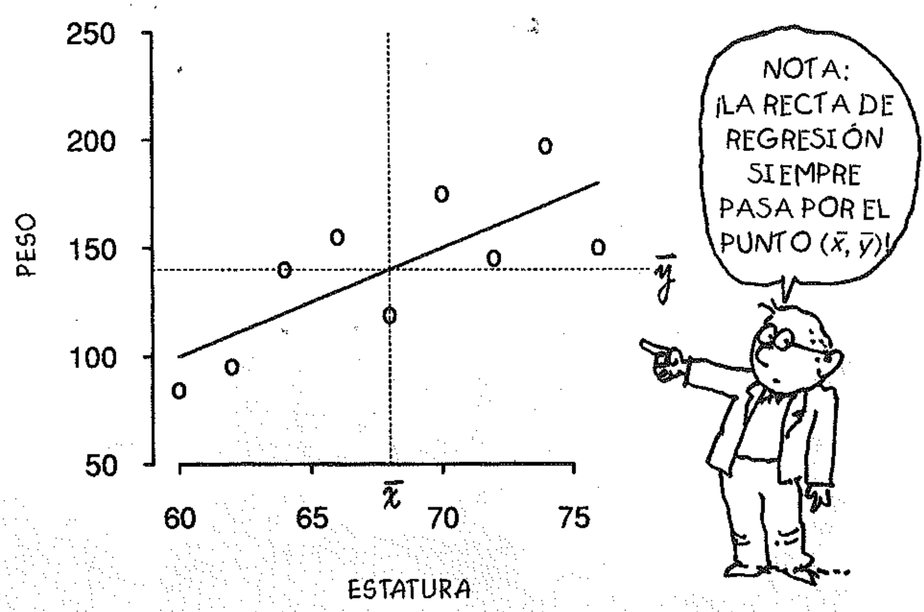
x_i	y_i	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
60	84	-8	-56	64	3136	448
62	95	-6	-45	36	2025	270
64	140	-4	0	16	0	0
66	155	-2	15	4	225	-30
68	119	0	-21	0	441	0
70	175	2	35	4	1225	70
72	145	4	5	16	25	20
74	197	6	57	36	3249	342
76	150	8	10	64	100	80
SUMA = 612	1.260			$SS_{xx}=240$	$SS_{yy}=10.426$	$SS_{xy}=1.200$

$\bar{x}=68$ $\bar{y}=140$

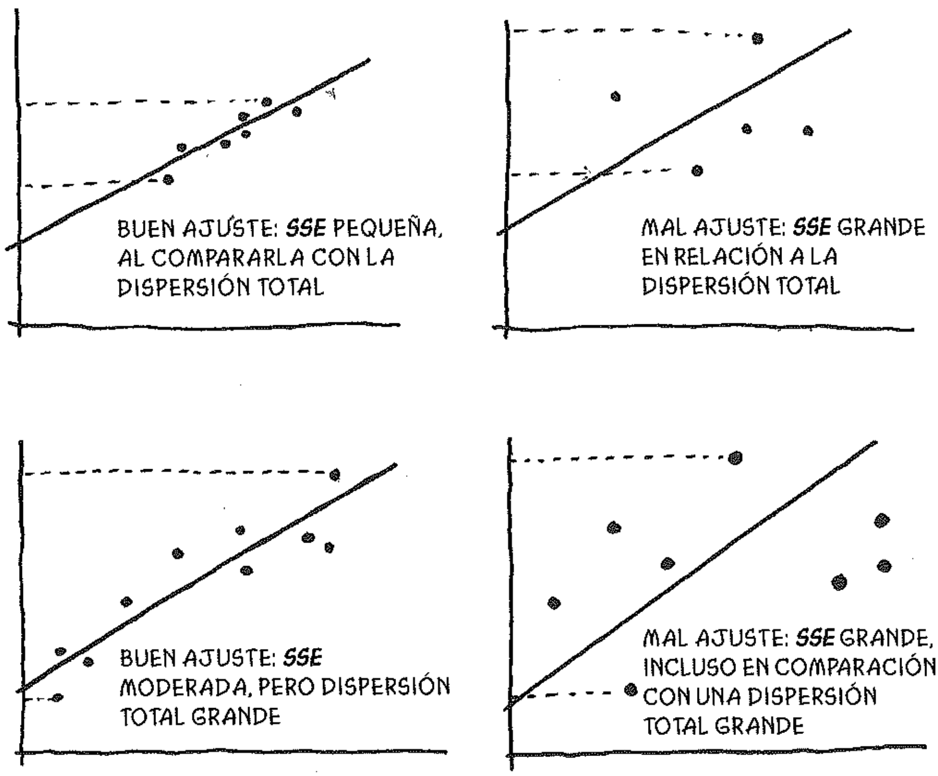
LO CUAL NOS DA VALORES PARA a Y b:

$$b=\frac{1.200}{240}=5 \qquad a=\bar{y}-b\bar{x}=140-5(68)=-200$$

ENTONCES $y=-200+5x$



COMO IMAGINAS, LA RESPUESTA A ESTA PREGUNTA DEPENDE DE LA FORMA EN QUE SE ESPARCEN LOS PUNTOS DE LOS DATOS. ES DECIR, ES LA MAGNITUD DE LA **SSE** RELATIVA A LA DISPERSIÓN TOTAL DE LOS DATOS. ALGUNOS EJEMPLOS:



VAMOS A CUANTIFICAR ESTO DESGLOSANDO LA VARIABILIDAD DE y. SEGUIREMOS COMO GUÍA EL DIBUJO DE LA DERECHA. TENEMOS

$$\hat{y}_i = a + bx_i$$

ENTONCES, $\hat{y}_i$ SON LOS PESOS PREDICHOS POR LA RECTA DE REGRESIÓN.

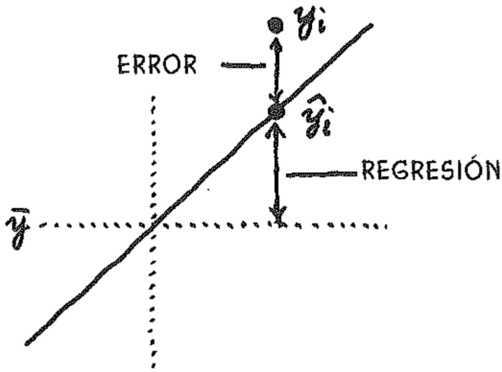


Tabla ANOVA

FUENTE DE VARIABILIDAD	SUMA DE CUADRADOS	VALOR DE LOS DATOS FICTICIOS
REGRESIÓN	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	6.000
ERROR	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	4.426
TOTAL	$SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$	10.426

(POR CIERTO, AUNQUE NO ES EVIDENTE QUE $SS_{yy} = SSR + SSE$, ES VERDAD) BUENO, DE TODOS MODOS, ASÍ ES COMO SE CALCULAN LAS SUMAS DE LA REGRESIÓN Y LOS ERRORES DE LOS CUADRADOS PARA EL CONJUNTO DE LOS DATOS REALES, CON $y = -200 + 5x$

			REGRESIÓN		ERROR	
x_i	y_i	$\hat{y}_i$	$(\hat{y}_i - \bar{y})$	$(\hat{y}_i - \bar{y})^2$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
60	84	100	-40	1.600	-16	256
62	95	110	-30	900	-15	225
64	140	120	-20	400	20	400
66	155	130	-10	100	25	625
68	119	140	0	0	-21	441
70	175	150	10	100	25	625
72	145	160	20	400	-15	225
74	197	170	30	900	27	729
76	150	180	40	1.600	-30	900
$\bar{x} = 68$	$\bar{y} = 140$			SSR = 6.000		SSE = 4.426

SSR MIDE LA VARIABILIDAD TOTAL DEBIDA A LA REGRESIÓN, O SEA, EXPLICADA POR LOS VALORES PREDICHOS DE *y*. YA NOS HEMOS ENCONTRADO CON ***SSE***. OBSERVA QUE:

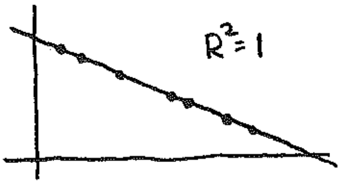
$$\frac{SSE}{SS_{yy}}$$

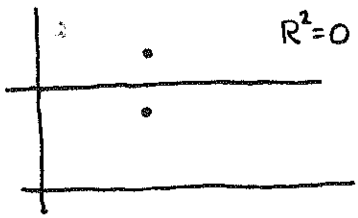
ES LA PROPORCIÓN DEL ERROR, RELATIVO A LA DISPERSIÓN TOTAL.

El coeficiente de determinación

ES LA PROPORCIÓN DE TODAS LAS SS_{yy} EXPLICABLES POR LA REGRESIÓN:

$$R^2 = \frac{SSR}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$



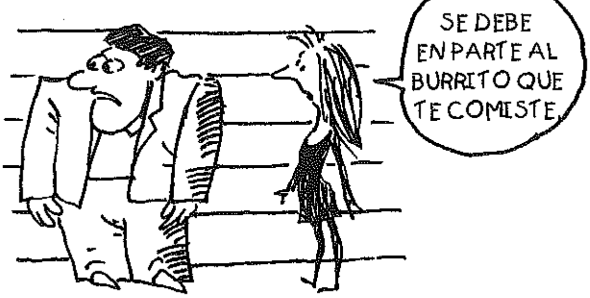


(PORQUE $SSR = SS_{yy} - SSE$). R^2 ES SIEMPRE MENOR QUE 1. CUANTO MÁS SE APROXIMA A 1, MÁS PRECISO ES EL AJUSTE DE LA CURVA. $R^2 = 1$ CORRESPONDE AL AJUSTE PERFECTO.

EL CÁLCULO DE R^2 DEL CONJUNTO DE DATOS FICTICIOS ES

$$R^2 = \frac{6.000}{10.426} = 0{,}58$$

LA VARIACIÓN DEL 58% EN EL PESO SE EXPLICA POR LA ESTATURA. EL 42% RESTANTE ES EL «ERROR».



POR OTRA PARTE, TENEMOS EL

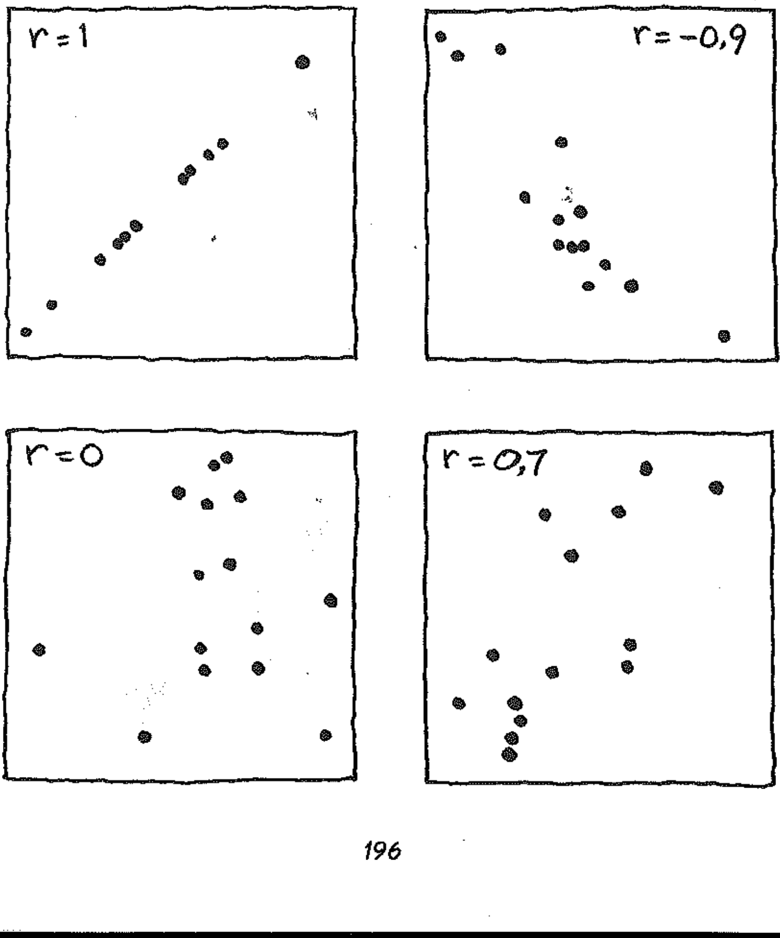
coeficiente de correlación

QUE ES LA RAÍZ CUADRADA DE R^2 CON EL SIGNO DE b.

$$r = (\text{SIGNO DE } b)\sqrt{R^2}$$

ENTONCES, r ES POSITIVA SI LA RECTA ES ASCENDENTE HACIA LA DERECHA, Y NEGATIVA SI LA RECTA TIENE FORMA DESCENDENTE HACIA LA DERECHA.

r MIDE LA PRECISIÓN DEL AJUSTE E INDICA SI AUMENTA LA x HACE SUBIR O HACE BAJAR LA y.



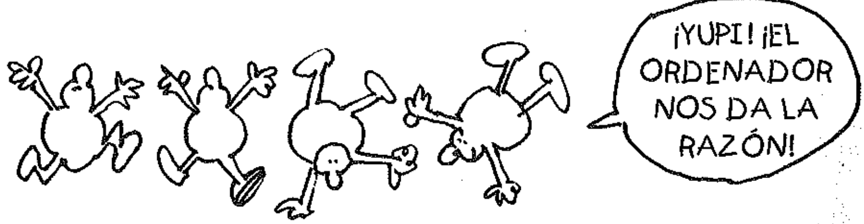
PERO SEAMOS SINCEROS: NADIE (BUENO, **CASI** NADIE) HACE YA ESTOS CÁLCULOS A MANO. CON EL ORDENADOR TODO ESTE TRABAJO PUEDE REALIZARSE ESCRIBIENDO **UNA SOLA LÍNEA DE CÓDIGO**...

EN EL SISTEMA DE SOFTWARE **MINITAB**, DISEÑADO EN EL ESTADO DE PENNSYLVANIA, EL ÚNICO COMANDO NECESARIO TIENE ESTE ASPECTO:

```
MTB > regress «PESO» on 1 independent variable «ESTATURA»
```

Y LOS RESULTADOS SON


```
The regression equation is

  PESO  - 200 + 5.00 ESTATURA

Predictor   Coef      Stdev   t-ratio     p
Constant   -200.0     110.7    -1.81    0.114
height      5.000     1.623     3.08    0.018

s = 25.15     R-sq = 57.5% R-sq(adj) = 51.5%


Analysis of Variance

SOURCE       DF     SS        MS        F        p
Regression   1    6000.0    6000.0    9.49    0.018
Error        7    4426.0     632.3
Total        8   10426.0
```



AHORA VAMOS A HACERLO CON LOS DATOS DE LOS 92 ESTUDIANTES:

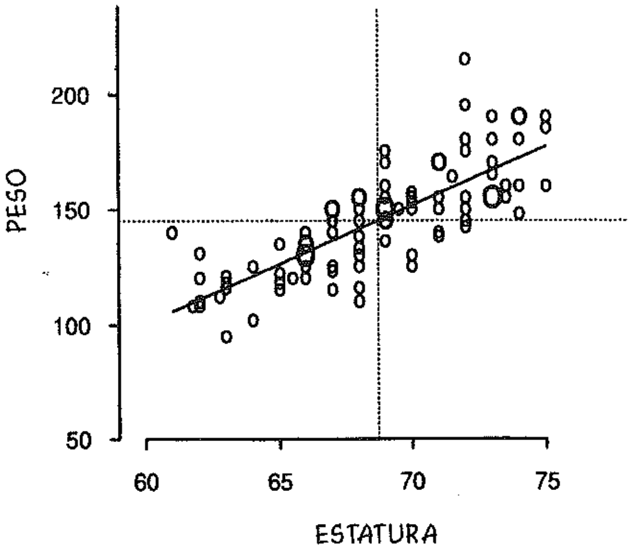
```
MTB > regress «PESO» on 1 independent variable «ESTATURA»
```

Y LOS RESULTADOS SON

```
The regression equation is
WEIGHT = - 205 + 5.09 HEIGHT

Predictor    Coef      Stdev    t-ratio     p
Constant   -204.74     29.16    -7.02     0.000
height       5.0918    0.4237   12.02     0.000

s = 14.79    R-sq = 61.6% R-sq(adj) = 61.2%


Analysis of Variance

SOURCE        DF       SS        MS         F         p
Regression     1      31592     31592     144.38    0.000
Error         90      19692       219
Total         91      51284
```

ESTE ES EL DIAGRAMA DE DISPERSIÓN DE PUNTOS CON LA RECTA DE REGRESIÓN AJUSTADA. EL COEFICIENTE DE CORRELACIÓN PARA ESTE CONJUNTO DE DATOS ES:

$$r = +\sqrt{0{,}616} = 0{,}78$$

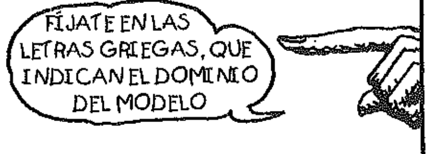


INFERENCIA ESTADÍSTICA

HASTA AHORA, HEMOS HECHO **ANÁLISIS DE DATOS** Y DESCRITO LA RELACIÓN LINEAL MÁS PRÓXIMA ENTRE LOS DATOS OBSERVADOS x E y. VAMOS A CAMBIAR NUESTRO PUNTO DE VISTA, RECORDEMOS A LOS 92 ESTUDIANTES COMO UNA **MUESTRA** POBLACIONAL DE TODOS LOS ESTUDIANTES. ¿QUÉ PODEMOS INFERIR?

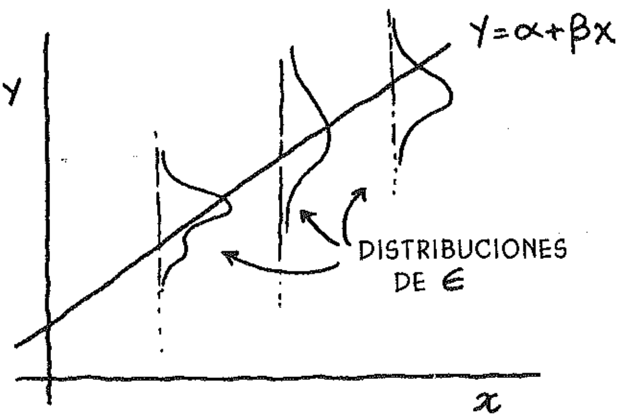
UN **MODELO DE REGRESIÓN** DEL TOTAL DE LA POBLACIÓN ES UNA RELACIÓN LINEAL

$$Y = \alpha + \beta x + \epsilon$$

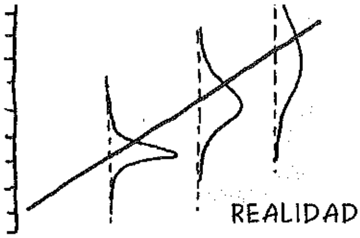


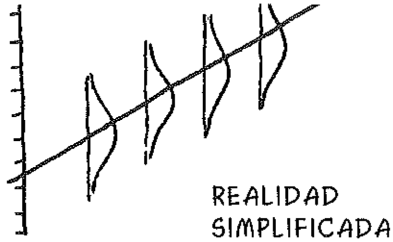
y ES LA VARIABLE ALEATORIA DEPENDIENTE; x ES LA VARIABLE INDEPENDIENTE (QUE PUEDE SER ALEATORIA O NO); α Y β SON LOS PARÁMETROS QUE QUEREMOS ESTIMAR; Y ϵ REPRESENTA LAS FLUCTUACIONES DEL ERROR ALEATORIO.

EN EL MODELO DE LA ESTATURA FRENTE AL PESO, x ES LA ESTATURA, α Y β SON LOS PARÁMETROS A ESTIMAR, Y PODEMOS CONSIDERAR ϵ COMO EL **COMPONENTE ALEATORIO** DE LOS PESOS y PARA CADA VALOR DE ESTATURA x.



DE HECHO, LA DISTRIBUCIÓN DE ϵ ES DIFERENTE PARA DISTINTOS VALORES DE x: LOS INDIVIDUOS QUE MIDEN 5 PIES (ALREDEDOR DE 1,52 METROS) VARÍAN MENOS EN EL PESO QUE LOS QUE MIDEN 6 PIES (ALREDEDOR DE 1,82 METROS). SIN EMBARGO, PODEMOS SIMPLIFICAR ESTA AFIRMACIÓN: SUPONGAMOS QUE PARA TODOS LOS VALORES DE x, LAS ϵ SON **INDEPENDIENTES, NORMALES Y TIENEN LA MISMA DESVIACIÓN TÍPICA** $\sigma = \sigma(\epsilon)$ Y **MEDIA** $\mu = 0$.



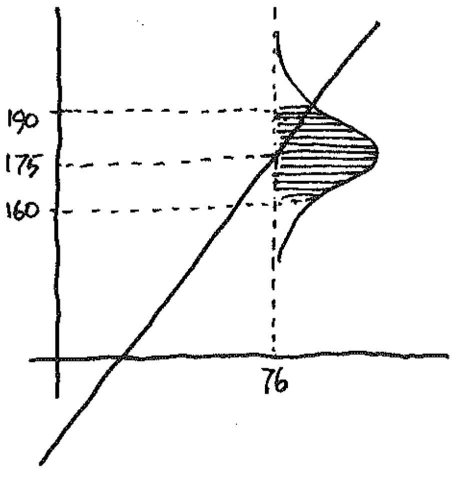


ASÍ QUE EL MODELO DE PESOS PUEDE SER

$$Y = -125 + 4x + \epsilon$$

ϵ ES NORMAL CON $\mu = 0$ Y $\sigma = 15$ LIBRAS (SUPONGAMOS). ENTONCES, DE ACUERDO CON ESTE MODELO, LOS ESTUDIANTES QUE TIENEN UNA ALTURA DE 6 PIES Y 4 PULGADAS (76 PULGADAS, O UNOS 193,4 CENTÍMETROS) TIENEN UNA DISTRIBUCIÓN DE

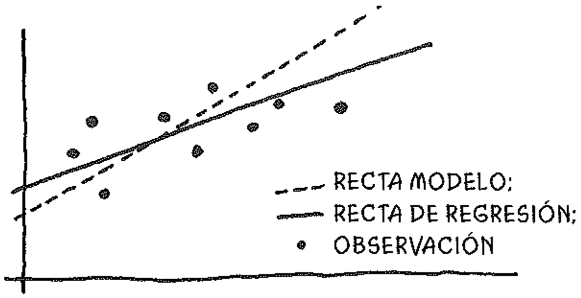
$$\begin{aligned} Y &= -125 + 4(76) + \epsilon \\ &= 175 + \epsilon \end{aligned}$$

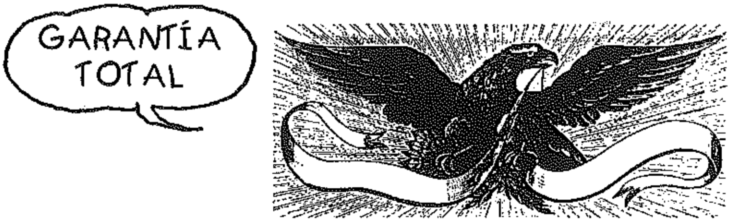


ASÍ QUE, PARA $x = 76$, Y ES NORMAL CON MEDIA 175 Y DESVIACIÓN TÍPICA DE 15 LIBRAS.

AHORA, DADO EL MODELO $Y = \alpha + \beta x + \epsilon$, QUEREMOS HACER LO MISMO QUE HEMOS HECHO EN ESTOS ÚLTIMOS CAPÍTULOS: TOMAR UNA **MUESTRA** Y UTILIZARLA PARA **ESTIMAR** α Y β.

SE PUEDE DEMOSTRAR QUE LAS a Y b OBTENIDAS POR EL MÉTODO ANTERIOR DE MÍNIMOS CUADRADOS SON LOS **ESTIMADORES LINEALES** NO SESGADOS DE MENOR VARIANZA (SEA ESTO LO QUE SEA).





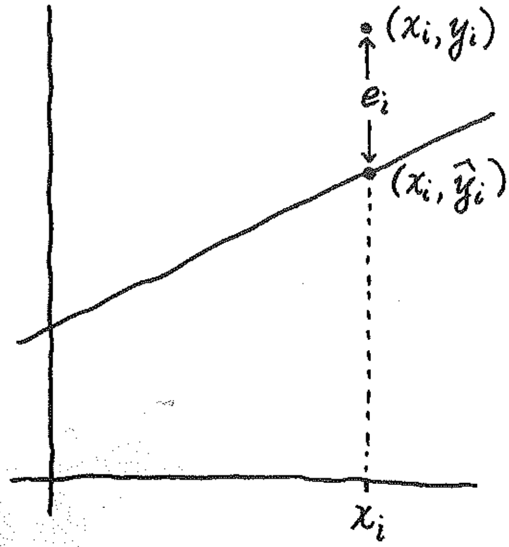
COMO SIEMPRE, **MUESTRAS** DIFERENTES PROPORCIONAN CONJUNTOS DE DATOS DIFERENTES, LO CUAL GENERA RECTAS DE REGRESIÓN DIFERENTES. ESTAS RECTAS SE **DISTRIBUYEN** ALREDEDOR DE $Y = \alpha + \beta x + \epsilon$. ENTONCES LA PREGUNTA ES: ¿CÓMO SE DISTRIBUYEN a Y b ALREDEDOR DE α Y β, RESPECTIVAMENTE, Y CÓMO CONSTRUIMOS LOS **INTERVALOS DE CONFIANZA** Y EL **CONTRASTE DE HIPÓTESIS**?

PARA CADA PUNTO (x_i, y_i) TENEMOS

$$y_i = a + bx_i + e_i$$

DONDE $e_i = y_i - \hat{y}_i$ ES LA DISTANCIA DE y_i HASTA LA RECTA DE REGRESIÓN. LOS e_i SON LOS **VALORES MUESTRALES DE** ϵ, Y NOS PROPORCIONAN UN ESTIMADOR s DE $\sigma(\epsilon)$:

$$s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$$

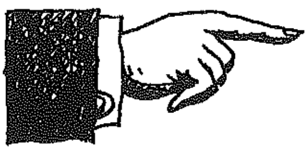


(¿POR QUÉ $n-2$ ES EL DENOMINADOR? POR QUÉ HEMOS UTILIZADO HASTA DOS GRADOS DE LIBERTAD PARA CALCULAR a Y b, DEJANDO $n-2$ PIEZAS INDEPENDIENTES DE INFORMACIÓN PARA ESTIMAR σ.)

AUNQUE NO RESULTE OBVIO, TAMBIÉN PODEMOS EXPRESAR s COMO:

$$s = \sqrt{\frac{SS_{yy} - bSS_{xy}}{n-2}}$$

UNA FÓRMULA QUE NOS PERMITE CALCULAR s DIRECTAMENTE A PARTIR DE LA ESTADÍSTICA MUESTRAL.



REPETIMOS, s ES UN ESTIMADOR **DEL GRADO DE DISPERSIÓN QUE TENDRÁN LOS PUNTOS ALREDEDOR DE LA RECTA.**

Intervalos de confianza

LOS INTERVALOS DE CONFIANZA DEL 95% PARA α Y β TIENEN ESTA YA CONOCIDA FORMA:

$$\beta = b \pm t_{0,025} SE(b)$$

$$\alpha = a \pm t_{0,025} SE(a)$$

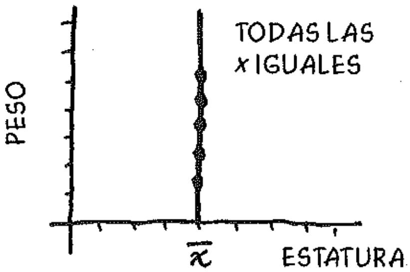
DONDE USAMOS LA DISTRIBUCIÓN t CON $n-2$ GRADOS DE LIBERTAD (POR LA MISMA RAZÓN QUE ANTES)

SIN EMBARGO, LOS ERRORES ESTÁNDAR NO NOS SUENAN PARA NADA. SON (SIN LA DERIVACIÓN):

$$SE(b) = \frac{s}{\sqrt{SS_{xx}}}$$

$$SE(a) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}}$$

¿QUÉ HA PASADO CON NUESTRO MARAVILLOSO $\frac{1}{\sqrt{n}}$? HA SIDO SUSTITUIDO POR SS_{xx}. AL IGUAL QUE n, SS_{xx} AUMENTA A MEDIDA QUE AÑADIMOS MÁS PUNTOS, PERO TAMBIÉN REFLEJA **LA DISPERSIÓN TOTAL DE LOS DATOS x**. POR EJEMPLO, SI **TODOS LOS ESTUDIANTES MUESTREADOS TUVIERAN LA MISMA ESTATURA**, NO TENDRÍAMOS **NINGUNA** JUSTIFICACIÓN PARA REPRESENTAR UNA CONCLUSIÓN SOBRE LA DEPENDENCIA DEL PESO CON RESPECTO A LA ESTATURA. SI ASÍ FUERA, $SS_{xx} = 0$, Y OBTENDRÍAMOS $b = \infty$ Y UNOS INTERVALOS DE CONFIANZA CON **AMPLITUD INFINITA**.



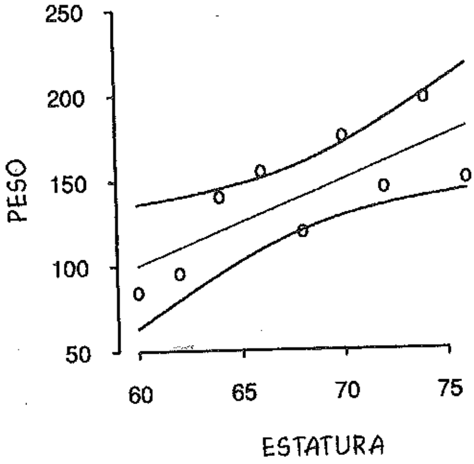
MÁS PREGUNTAS:

¿CON QUÉ PRECISIÓN PODEMOS INFERIR LA **RESPUESTA MEDIA** Y EN UN VALOR FIJO x_0? POR EJEMPLO, ¿CUÁL ES EL PESO MEDIO DE LOS ESTUDIANTES QUE MIDEN 76 PULGADAS? EL INTERVALO DE CONFIANZA DEL 95% PARA $Y = \alpha + \beta x_0$ ES:

$$\alpha + \beta x_0 = a + b x_0 \pm t_{0{,}025} SE(\hat{y})$$

DONDE

$$SE(\hat{y}) = s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$



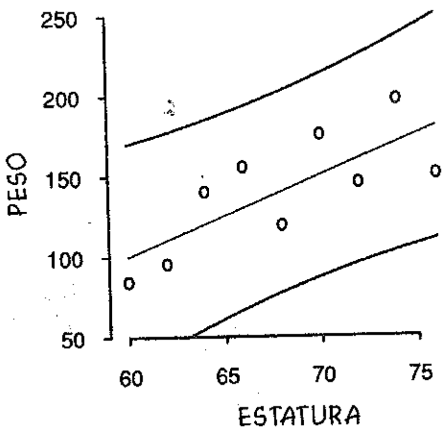
SUPONGAMOS QUE ENTRA UN NUEVO ESTUDIANTE QUE TIENE UNA ALTURA x_{nueva}. ¿CON QUÉ PRECISIÓN PODEMOS INFERIR y_{nuevo} SIN PESARLE?

EL INTERVALO DE CONFIANZA DEL 95% DE y_{nuevo} PARA UN INDIVIDUO CON UNA x_{nueva} OBSERVADA ES

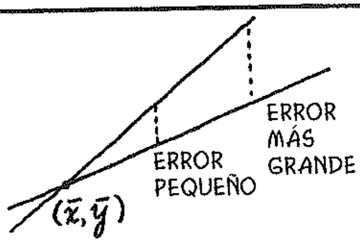
$$Y_{nuevo} = a + b x_{nueva} \pm t_{0{,}025} SE(Y_{nuevo})$$

DONDE

$$SE(Y_{nuevo}) = s\sqrt{1 + \frac{1}{n} + \frac{(x_{nueva} - \bar{x})^2}{SS_{xx}}}$$



AMBOS ERRORES ESTÁNDAR CONTIENEN UN TÉRMINO QUE CRECE A MEDIDA QUE EL VALOR x_0 O x_{nueva} SE ALEJA DEL VALOR MEDIO $\bar{x}$. ¿POR QUÉ EL ERROR SE ALEJA MÁS DE $\bar{x}$? PORQUE, SI DESPLAZAMOS LA RECTA DE REGRESIÓN, ¡SE QUEDA MUY ALEJADO DE LA MEDIA! (RECUERDA, LA RECTA SIEMPRE PASA POR $(\bar{x}, \bar{y})$.)



HAGAMOS LO MISMO CON LOS DATOS FICTICIOS: PARA EL **PESO MEDIO** CUANDO $x = 76$ PULGADAS, TENEMOS QUE $\beta = -200$ Y $\alpha = 5$. ENTONCES

$$Y = -200 + 5(76) \pm (2{,}365)(25{,}15)$$

$$= 180 \pm (2{,}365)(25{,}15)\sqrt{0{,}3777}$$

$$= 180 \pm 36{,}34 \text{ LIBRAS} = [144, 216]$$

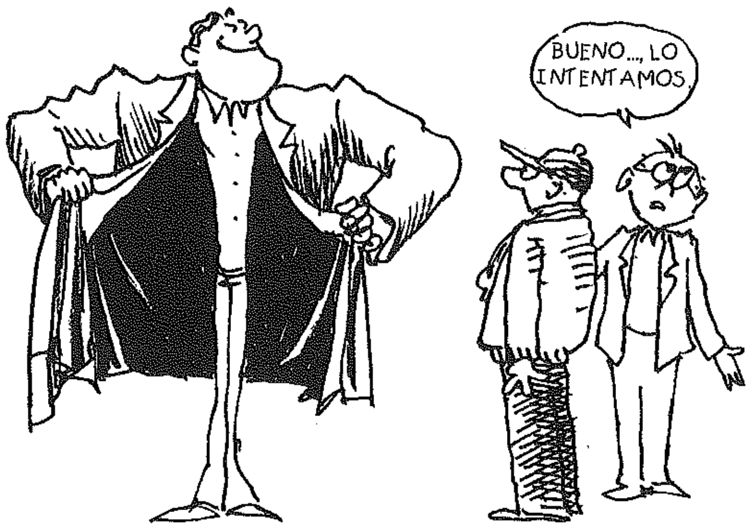
LA MEDIA ESTIMADA DE LOS ESTUDIANTES QUE MIDEN 6 PIES Y 4 PULGADAS ES DE 180 LIBRAS, Y TENEMOS UNA SEGURIDAD DEL 95% DE QUE ESTAMOS A MENOS DE 36 LIBRAS DE LA MEDIA REAL.

PARA UN **NUEVO** ESTUDIANTE QUE MIDA 76 PULGADAS, UTILIZAMOS NUESTRA MUESTRA FICTICIA DE NUEVE PUNTOS PARA INFERIR QUE

$$y_{nuevo} = -200 + 5(76) \pm (2{,}365)(25{,}15)\sqrt{1 + \frac{1}{9} + \frac{(76-68)^2}{240}}$$

$$= 180 \pm (2{,}365)(29{,}51)$$

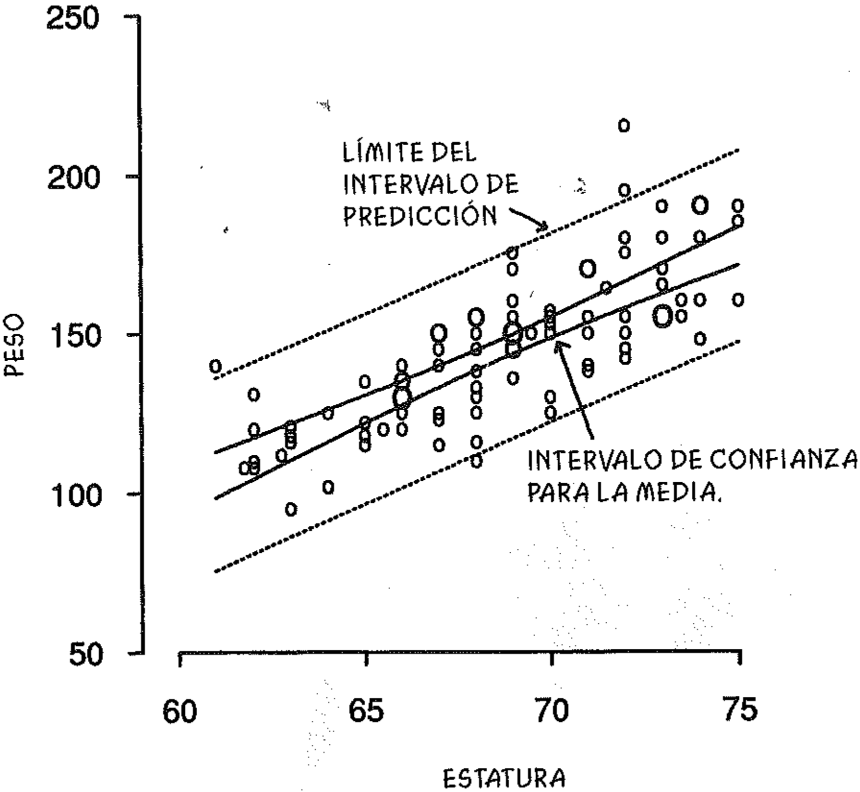
$$= 180 \pm 70 \text{ LIBRAS} = [110, 250]$$



LE DECIMOS AL ENTRENADOR DE FÚTBOL QUE ESTAMOS BASTANTE SEGUROS DE QUE EL NUEVO PESA ¡ENTRE 110 Y 250! (ENTRE 50 Y 115 KILOS)

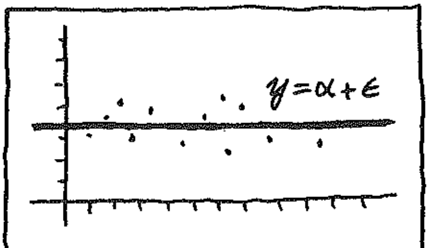
¡LOS INTERVALOS SON BASTANTE HORRIBLES! ¿CUÁL ES EL PROBLEMA? EN REALIDAD HAY DOS PROBLEMAS:

LOS ESTUDIANTES DE PENNSYLVANIA PRESENTAN MEJORES ESTIMACIONES.



Contraste de hipótesis

QUIEN SEA TOTALMENTE ESCÉPTICO PUEDE SUGERIR QUE NO EXISTE NINGUNA **RELACIÓN** ENTRE LA ESTATURA Y EL PESO. ESTO EQUIVALE A DECIR QUE $\beta = 0$.



x NO TIENE EFECTO SOBRE y

TOMAMOS ESTO COMO **HIPÓTESIS NULA.**

$$H_0 : \beta = 0$$

EN ESTE CASO, EL ESTADÍSTICO

$$t = \frac{b}{SE(b)}$$

TIENE DISTRIBUCIÓN t CON $n-2$ GRADOS DE LIBERTAD. COMO SIEMPRE, LA PRUEBA DE SIGNIFICACIÓN DEPENDE DE LA HIPÓTESIS ALTERNATIVA.

$t > t_\alpha$ PARA $H_a : \beta > 0$

$t < t_\alpha$ PARA $H_a : \beta < 0$

$|t| > |t_{\alpha/2}|$ PARA $H_a : \beta \neq 0$

PARA LOS DATOS DE PESO FICTICIOS, TENEMOS LA FIRME SOSPECHA DE QUE LA HIPÓTESIS ALTERNATIVA DEBERÍA SER

$$H_a : \beta > 0$$

LO PROBAMOS.

$$t_{OBS} = \frac{5}{SE(b)} = \frac{5}{1{,}62}$$

$$= 3{,}08$$

PARA 7 GRADOS DE LIBERTAD, $t_{0,05} = 1{,}895$. DADO QUE $t_{OBS} > t_{0,05}$ RECHAZAMOS LA HIPÓTESIS NULA AL NIVEL DE SIGNIFICACIÓN Y CONCLUIMOS AFIRMANDO QUE EXISTE UNA RELACIÓN POSITIVA ENTRE LA ESTATURA Y EL PESO.

Regresión lineal múltiple

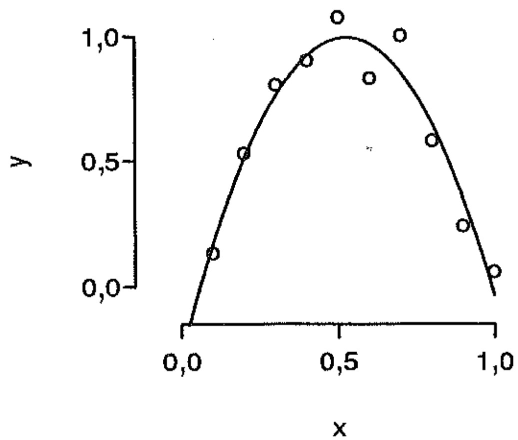
PODEMOS UTILIZAR LAS MISMAS IDEAS FUNDAMENTALES PARA ANALIZAR LA RELACIÓN ENTRE UNA VARIABLE DEPENDIENTE Y **DISTINTAS** VARIABLES INDEPENDIENTES:

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n + \epsilon$$

POR EJEMPLO, EL PESO ESTÁ DETERMINADO POR UNA SERIE DE FACTORES DIFERENTES A LA ESTATURA: LA EDAD, EL SEXO, LA DIETA, LA COMPLEXIÓN FÍSICA, ETC.

EL ÁLGEBRA MATRICIAL Y EL ORDENADOR SE COMPLEMENTAN PARA FACILITAR EL ANÁLISIS DE ESTOS PROBLEMAS.

Regresión no lineal



OBVIAMENTE, A VECES LOS DATOS DIBUJAN UNA **CURVA NO LINEAL**. LOS ESTADÍSTICOS TIENEN UN MONTÓN DE TRUCOS PARA UTILIZAR TÉCNICAS DE REGRESIÓN **LINEAL** PARA PROBLEMAS NO LINEALES. LO MÁS FÁCIL ES ESCRIBIR Y COMO UNA **POLINOMIAL**

$$Y = \alpha + \beta_1 x + \beta_2 x^2 + \epsilon$$

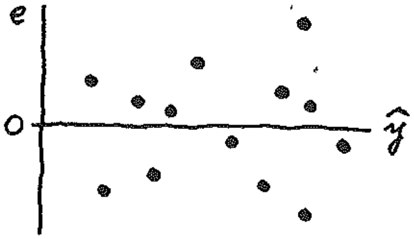
Y TRATAR A x Y A x^2 COMO VARIABLES INDEPENDIENTES EN UN MODELO LINEAL.

Diagnóstico de la regresión

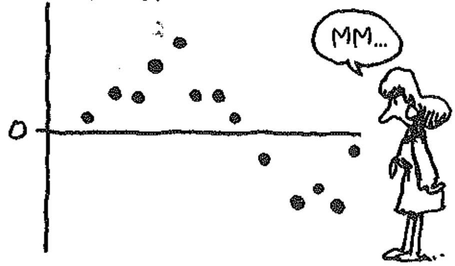
AJUSTAR UN MODELO COMPLEJO A LOS DATOS PUEDE OCULTAR MUCHAS DIFICULTADES. UTILIZAMOS LOS PROCEDIMIENTOS DE DIAGNÓSTICO DE LA REGRESIÓN PARA DESCUBRIR TODO TIPO DE SORPRESAS INDEFINIBLES Y DESAGRADABLES.

EL PROCEDIMIENTO MÁS SIMPLE CONSISTE EN REPRESENTAR EN UN DIAGRAMA DE PUNTOS LOS **RESIDUOS** e_i FRENTE A LA **PREDICCIÓN** $\hat{y}_i$. RECUERDA QUE ASUMIMOS QUE EL ERROR ϵ ES INDEPENDIENTE DE x.

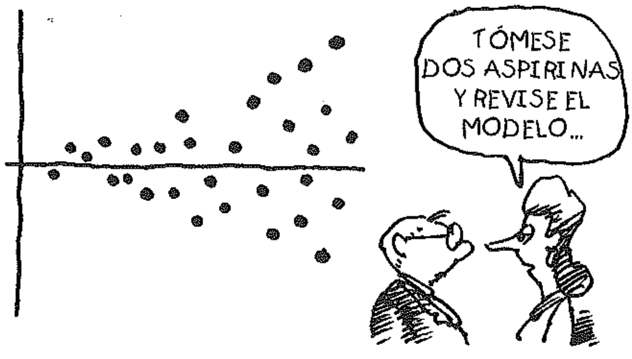
UNA DISPERSIÓN DE PUNTOS **ALEATORIA** INDICA QUE LAS PRESUNCIONES DEL MODELO SON PROBABLEMENTE CORRECTAS.



CUALQUIER **FORMA** QUE ADOPTE EL GRÁFICO INDICA PROBLEMAS CON LAS PREMISAS DEL MODELO.



UNA TÍPICA SORPRESA DESAGRADABLE (QUE SE DA EN LOS DATOS DE PESO/ESTATURA) ES QUE LOS ERRORES SON **HETEROCEDÁSTICOS**, ES DECIR, QUE LA DISPERSIÓN DE e AUMENTA A MEDIDA QUE AUMENTA y.



EN ESTE CAPÍTULO, HEMOS RESUMIDO LAS IDEAS FUNDAMENTALES Y LAS TÉCNICAS DEL ANÁLISIS DE REGRESIÓN, QUE ESTUDIA RELACIONES ENTRE VARIABLES. CON ESTO CONCLUIMOS NUESTRA DETALLADA DISCUSIÓN SOBRE LOS MÉTODOS BÁSICOS DE LA ESTADÍSTICA. EN NUESTRO ÚLTIMO CAPÍTULO HAREMOS UN REPASO RÁPIDO DE ALGUNOS PUNTOS QUE FALTAN.