EL MODELO LINEAL

Clase 1

Regresión simple

Sergio Camiz

LIMA - Marzo-Mayo 2025

Asuntos de la clase 2

- El modelo lineal
- Análisis del modelo lineal
- Estimación de los parámetros
- La recta de los mínimos cuadrados
- Contribución y apalancamiento

El modelo lineal

En general, un *modelo matemático* se intenta

- para buscar si existe una relación de dependencia entre
- un carácter *respuesta* o *dependiente*, indicado por $\boldsymbol{\eta}$,
- y otros caracteres independientes llamados *explicativos* o *predictivos* $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_s$, así que se pueda escribir

$$\boldsymbol{\eta} = f(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_s | \beta_1, \beta_2, \ldots, \beta_p) \qquad (1)$$

- con $\beta_1, \beta_2, \ldots, \beta_p$ parámetros desconocidos que se necesita estimar.

Se quiere también evaluar la intensidad de la relación.

Se dice que $\boldsymbol{\eta}$ *sigue un modelo lineal en los parámetros* β_j *por respecto a* $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s$, si la (1) se puede escribir

$$\boldsymbol{\eta} = f(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s | \beta_1, \ldots, \beta_p) = \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s) = \sum_{j=1}^{p} \beta_j \boldsymbol{x}_j \qquad (2)$$

donde los $\boldsymbol{x}_j$ son funciones solo de los $\boldsymbol{z}_k$ sin parámetros desconocidos, mientras que los β_j, en principio desconocidos, aparecen linealmente en (2).

Para estimar el modelo, se elige un conjunto de unidades estadísticas donde se observan simultáneamente tanto los caracteres $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s$ como $\boldsymbol{\eta}$, los valores de los $\boldsymbol{z}_j$ formando una matriz $n \times s$ $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_s)$, mientras los de $\boldsymbol{\eta}$ forman un vector $\boldsymbol{y} = (y_1, \ldots, y_n)'$.

El modelo lineal se puede ver en cadauna de las fases que hemos descrito:

- función descriptiva y cognitiva, por tanto se emplea en la fase exploratoria, para estudiar relaciones entre caracteres;
- función confirmatoria, en cuanto permite de testar una relación entre caracteres de manera estadísticamente fiable y su inferencia, si los datos fueron muestrados correctamente;
- permite la modelización de gran parte de fenómenos naturales, económicos y humanos en general.

Por tanto intentaremos de distinguir muy claramente los diferentes contextos, tanto de su construcción como de su empleo.

Ejemplos

1. $\boldsymbol{\eta} = \beta_0 + \beta_1 \boldsymbol{z}_1 = \alpha + \beta \boldsymbol{x}$
   es un modelo lineal con $\boldsymbol{x}(\boldsymbol{z}_1) = \boldsymbol{z}_1$: los parámetros $(\beta_0, \beta_1) = (\alpha, \beta)$ entran linealmente en el modelo.

2. $\boldsymbol{\eta} = \beta_0 + \beta_1 \boldsymbol{z}_1 + \cdots + \beta_p \boldsymbol{z}_p = \alpha + \beta_1 \boldsymbol{x}_1 + \cdots + \beta_p \boldsymbol{x}_p$
   es un modelo lineal con $\boldsymbol{x}_j(\boldsymbol{z}_j) = \boldsymbol{z}_j$: los β_j entran linealmente en el modelo.

3. una relación polinomial
   $\boldsymbol{\eta} = \beta_0 + \beta_1 \boldsymbol{z}^1 + \beta_2 \boldsymbol{z}^2 + \cdots + \beta_p \boldsymbol{z}^p = \alpha + \beta_1 \boldsymbol{x}_1 + \cdots + \beta_p \boldsymbol{x}_p$
   es un modelo lineal con $\boldsymbol{x}_j(\boldsymbol{z}_j) = \boldsymbol{z}^j$: los β_j entran linealmente en el modelo.

4. $\boldsymbol{\eta} = \alpha + \beta \sin 2\pi \boldsymbol{z}$ es un modelo lineal, de tipo $\boldsymbol{z} = \alpha + \beta \boldsymbol{x}$ con $\boldsymbol{x}(\boldsymbol{z}) = \sin 2\pi \boldsymbol{z}$: los parámetros (α, β) entran linealmente en el modelo.

5. al contrario, $\boldsymbol{\eta} = \frac{e^{\beta_1 \boldsymbol{z}_1} - e^{\beta_2 \boldsymbol{z}_2}}{\beta_2 - \beta_1}$ *no es* un modelo lineal, porque los parámetros ($\beta_1, \beta_2)$ no entran linealmente en el modelo.

6. $\boldsymbol{\xi} = \delta e^{\gamma \boldsymbol{z}}$ no es lineal en γ, pero si se toman los logaritmos resulta $log \boldsymbol{\xi} = log \delta + \gamma \boldsymbol{z}$ así que se lo devuelve, poniendo $\boldsymbol{\eta} = log \boldsymbol{\xi}, \beta_1 = log \delta, \beta_2 = \gamma, x_1(\boldsymbol{z}) = 1, x_2(\boldsymbol{z}) = \boldsymbol{z}$, etc.

Notaciones particulares para las varias somas:

$$S_x = \sum_i x_i \qquad S_{\dot{x}} = \sum_i (x_i - \bar{x})$$

$$S_{xx} = \sum_i x_i^2 \qquad S_{\dot{x}\dot{x}} = \sum_i (x_i - \bar{x})^2$$

y analogamente para y, luego

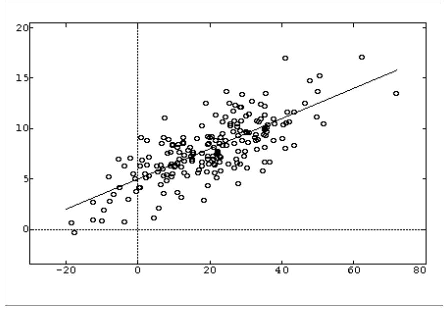
$$S_{xy} = \sum_i x_i y_i \qquad S_{\dot{x}\dot{y}} = \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

se observa también que

$$S_{\dot{x}\dot{y}} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - n\bar{x}\bar{y} =$$
$$= \sum_i x_i y_i - \sum_i x_i \bar{y} = \sum_i x_i (y_i - \bar{y}) = S_{x\dot{y}} = S_{\dot{x}y}$$

Análisis del modelo lineal

Veamos ahora más de cerca como se analiza el modelo lineal más sencillo, dado para $y = \alpha + \beta x$.



Esta elección de modelo tal vez se hace porque:

1. ya se sabe que la relación es lineal;
2. en la región de las elecciones usuales de x la relación lineal es muy buena aproximación;
3. se busca una relación funcional entre y y x y se utiliza un modelo lineal como primera etapa de investigación.

Problemas:

1. pasaje por el orígen;
2. linearidad: ¿es efectivamente linear la relación buscada?
3. distribución

Si se piensa que la relación entre y y x es lineal, exprimida para la función

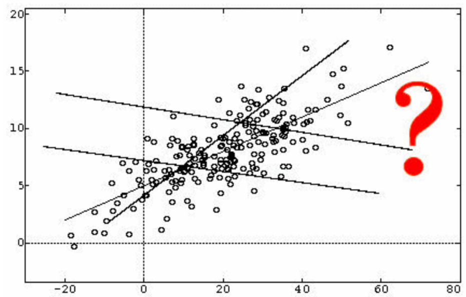
$$y = \alpha + \beta x \tag{3}$$

y se tienen más de dos observaciones con x_i diferente, está cierto que será muy difícil que todas se encuentren a lo largo de la recta misma, así que aplicando (3) a los valores x_i se resultará que:

$$y_i = \alpha + \beta x_i + \varepsilon_i,\ \forall i \in (1, n). \tag{4}$$

ε_i puede ser un error de medición, si suponemos que hay una relación funcional entre x y y, sino puede ser una parte explicada para otros caracteres o una variabilidad individual de las unidades estadísticas.

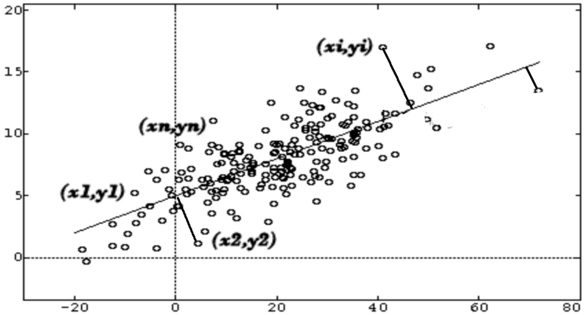
Estimación de los parámetros



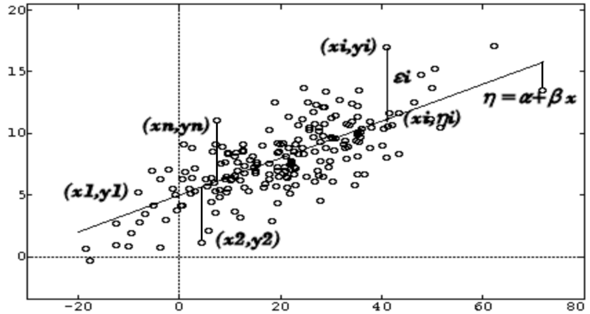
Ay que buscar un método que permita estimar los parámetros desconocidos α y β para encontrar la mejor recta en algún sentido.

Entonces se busca la recta que sea más cercana a los datos.

Una recta para un conjunto de puntos se puede *ajustar*, minimizando su distancia desde cada punto.



Pero nosotros queremos una recta correspondiente a un modelo, o sea que, dado x_i, el valor $\eta_{x_i} = \alpha + \beta x_i$ sobre la recta estime *a lo mejor* y_i.



Por cada observación (x_i, y_i), el punto estimado sobre la recta tiene como coordenadas $(x_i, \eta_{x_i} = \alpha + \beta x_i)$, así que, bajo (4) resulta

$$y_i - \eta_{x_i} = \varepsilon_i \tag{5}$$

correspondiente a la longitud del segmento vertical que une (x_i, y_i) con (x_i, η_{x_i}).

Hay tres posibles determinaciones de los errores:

1. $\Sigma_i\, \varepsilon_i$, pero las diferencias de signal no permiten de optimizar;
2. $\Sigma_i\, |\varepsilon_i|$, pero el algebra de los valores absolutos es difícil;
3. $\Sigma_i\, \varepsilon_i^2$, mejor, pero pesos grandes para errores grandes y reducidos para pequenos.

Para identificar a la recta de regresión vamos a buscar la que minimiza la suma de cuadrados de los errores. Dicha suma de cuadrados corresponde también a la calidad de la estimación.

$$SS_e = \sum_i \varepsilon_i^2 = \sum_i (y_i - \eta_{x_i})^2 = \sum_i (y_i - \alpha - \beta x_i)^2$$

así que entre la infinidad de rectas de regresión posibles, se elegirá la estimación de los parámetros

$$(\hat{\alpha}, \hat{\beta}) \text{ correspondiente a la recta } \hat{\eta} = \hat{\alpha} + \hat{\beta} x$$

que resulte la más cercana de los datos, o sea tal que

$$SS_e(\hat{\alpha}, \hat{\beta}) = \min_{(\alpha,\beta)} SS_e(\alpha, \beta) = \min_{(\alpha,\beta)} \sum_i (y_i - \alpha - \beta x_i)^2$$

El método empleado se llama *método de mínimos cuadrados* y se aplica al problema de optimización siguiente:

estimar $(\hat{\alpha}, \hat{\beta})$ *bajo la condición que* $SS_e(\hat{\alpha}, \hat{\beta}) = \text{mín}$
o sea estimar (α, β) tal que resulte

$$\begin{cases} \hat{\eta} = \hat{\alpha} + \hat{\beta}\, x \\ SS_e(\hat{\alpha}, \hat{\beta}) = \sum_i (y_i - \hat{\eta}_i)^2 = \\ \quad = \sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 = \\ \quad = \min_{(\alpha, \beta)} \sum_i (y_i - \alpha - \beta x_i)^2 = \\ \quad = \min_{(\alpha, \beta)} SS_e(\alpha, \beta) \end{cases} \tag{6}$$

Solución

Deben calcularse las derivadas parciales de

$$SS_e(\alpha, \beta) = \sum_i (y_i - \alpha - \beta x_i)^2$$

y igualarlas a cero:

$$\begin{cases} \frac{\partial SS_e(\alpha, \beta)}{\partial \alpha} = -2 \sum_i (y_i - \alpha - \beta x_i) = 0 \\ \frac{\partial SS_e(\alpha, \beta)}{\partial \beta} = -2 \sum_i (y_i - \alpha - \beta x_i)\, x_i = 0 \end{cases}$$

Desarrollando se consigue el sistema de *ecuaciones normales* en α, β:

$$\begin{cases} n\,\alpha + \sum_i x_i\, \beta = \sum_i y_i \\ \sum_i x_i\, \alpha + \sum_i x_i^2\, \beta = \sum_i x_i y_i \end{cases} \tag{7}$$

cuya solución resulta:

$$\begin{cases} \hat{\alpha} = \bar{y} - \hat{\beta}\, \bar{x} \\ \hat{\beta} = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \end{cases} \tag{8}$$

A $(\hat{\alpha}, \hat{\beta})$ se le llaman *estimadores de mínimos cuadrados.*
Siempre se puede hacer una regresión simple, a condicion que var$(\boldsymbol{x}) > 0$, o sea que haya por lo menos dos x diferentes.

Efectivamente, desde la primera ecuación: $n\alpha + \sum_i x_i \beta = \sum_i y_i$ se resulta $\alpha = \frac{\sum_i y_i - \sum_i x_i \beta}{n} = \bar{y} - \bar{x}\beta$.
Sustituyendo α en la segunda ecuación $\sum_i x_i \alpha + \sum_i x_i^2 \beta = \sum_i x_i y_i$ sigue $\sum_i x_i(\bar{y} - \bar{x}\beta) + \sum_i x_i^2 \beta = \sum_i x_i y_i$. Desarrollando

$$\sum_i x_i \bar{y} - \sum_i x_i \bar{x} \beta + \sum_i x_i^2 \beta = \sum_i x_i y_i$$

$$\sum_i x_i^2 \beta - \sum_i x_i \bar{x} \beta = \sum_i x_i y_i - \sum_i x_i \bar{y}$$

$$(\sum_i x_i^2 - \sum_i x_i \bar{x})\beta = \sum_i x_i y_i - \sum_i x_i \bar{y}$$

$$\beta = \frac{\sum_i x_i y_i - \sum_i x_i \bar{y}}{\sum_i x_i^2 - \sum_i x_i \bar{x}} = \frac{\sum_i x_i (y_i - \bar{y})}{\sum_i x_i (x_i - \bar{x})} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

```
n     <- dim(SudAmerica)[1]; n  # n es el número de observaciones
x     <- SudAmerica[,1]; x      # x es el caracter descriptivo
y     <- SudAmerica[,2]; y      # y es el caracter respuesta
xm    <- sum(x)/n; xm           # cálculo de xm, promedio de x
ym    <- sum(y)/n; ym           # cálculo de ym, promedio de y
ssx   <- sum(x^2); ssx          # ssx es la suma de los x cuadros
ssy   <- sum(y^2); ssy          # ssy es la suma de los y cuadros
sxy   <- sum(x*y); sxy          # sxy es la suma de los productos xy
ssxc  <- ssx-n*xm^2 ; ssxc      # ssxc es ssx centrado sobre el promedio
ssyc  <- ssy-n*ym^2 ; ssyc      # ssyc es ssy centrado sobre el promedio
sxyc  <- sxy-n*xm*ym; sxyc      # sxyc es sxy centrado sobre el promedio
varx  <- ssxc/n; varx           # varx es la varianza de x
vary  <- ssyc/n; vary           # vary es la varianza de y
covxy <- sxyc/n; covxy          # covxy es la covarianza de xy
bh    <- sxyc/ssxc ; bh         # bh es la estimación de beta
ah    <- ym - bh*xm ; ah        # ah es la estimación de alfa
plot(x,y)                       # el plot ordinario
text(x,y,lab=lab)               # las etiquetas
abline(ah,bh,col='red')         # la recta de regresión
```

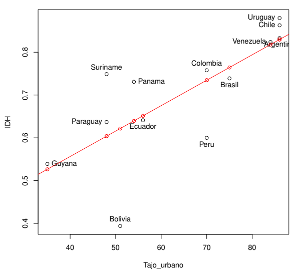

Para las variables se encuentran las estadísticas siguientes:

	x	y	xy
Mínimo	35.00000	0.3940000	
Máximo	86.00000	0.8800000	
Promedio	65.30769	0.7067692	
Total	849.00000	9.1880000	
Suma de cuadrados	59155.00000	6.7304880	622.094000
Cuadrados centrados	3708.76923	0.2366923	22.046923
Varianza covarianza	285.28994	0.0182071	1.695917
Desvío estándar	16.89053	0.1349337	

Las estimaciones de α y β son:

$$\alpha = 0{,}3185451, \ \beta = 0{,}005944539$$

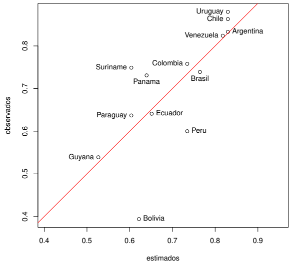
así que la recta resulta ser: $y = 0{,}31854511 + 0{,}005945x$

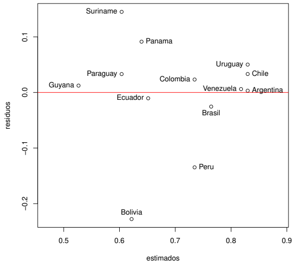


y se resultan las estimaciones η_i de los y_i bajo el modelo de regresión y los residuos correspondientes:

	Tajo_urbano	IDH	eta	residuos	%res
Argentina	86	0.833	0.8297755	0.003224541	0.003870997
Bolivia	51	0.394	0.6217166	-0.227716597	-0.577960906
Brasil	75	0.739	0.7643855	-0.025385531	-0.034351192
Chile	86	0.863	0.8297755	0.033224541	0.038498888
Colombia	70	0.758	0.7346628	0.023337163	0.030787815
Ecuador	56	0.641	0.6514393	-0.010439291	-0.016285946
Guyana	35	0.539	0.5266040	0.012396026	0.022998193
Panama	54	0.731	0.6395502	0.091449786	0.125102307
Paraguay	48	0.637	0.6038830	0.033117020	0.051989042
Peru	70	0.600	0.7346628	-0.134662837	-0.224438061
Suriname	48	0.749	0.6038830	0.145117020	0.193747690
Uruguay	86	0.880	0.8297755	0.050224541	0.057073342
Venezuela	84	0.824	0.8178864	0.006113618	0.007419440

Aquí los gráficos representando los valores estimados y observados de *IDH* y los residuos por respecto a los valores estimados.





La recta de los mínimos cuadrados

La estimación de $\eta_x = \alpha + \beta x$ es representada por la *recta de los mínimos cuadrados*

$$\hat{\eta} = \hat{\alpha} + \hat{\beta}\, x$$

donde

$$\begin{cases} \hat{\alpha} = \bar{y} - \bar{x}\,\hat{\beta} \\ \hat{\beta} = \dfrac{S_{\dot{x}\dot{y}}}{S_{\dot{x}\dot{x}}} \end{cases}$$

Propiedades

1. Si $x = \bar{x}$, resulta

$$\hat{\eta}_{\bar{x}} = \hat{\alpha} + \hat{\beta}\,\bar{x} = (\bar{y} - \hat{\beta}\,\bar{x}) + \hat{\beta}\,\bar{x} = \bar{y}$$

   La recta de los mínimos cuadrados pasa por el baricentro de los datos $(\bar{x}, \bar{y})$.

2. Se resulta
   $S_e = \Sigma_i(y_i - \hat{\eta}_i) = \Sigma_i(y_i - \hat{\alpha} - \hat{\beta}x_i) = \Sigma_i(y_i - \bar{y} + \hat{\beta}\bar{x} - \hat{\beta}x_i) = 0$
   La suma de los desvíos a los valores estimados es cero.

3. bajo esto, resulta $\bar{y}_i = \Sigma_i\, \hat{\eta}_{x_i} \,/\, n = \bar{\eta}$.
   El promedio de los valores estimados coincide con el promedio de los valores observados.

Los puntos sobre la recta correspondiente a los valores x_i tienen como coordenadas $(x_i, \hat{\eta}_{x_i})$, por lo que se tiene que

$$SS_e = SS_e(\hat{\alpha}, \hat{\beta}) = \sum_i (y_i - \hat{\eta}_{x_i})^2 = \min_{(\alpha, \beta)} SS_e(\alpha, \beta)$$

A SS_e se le llama *suma de los cuadrados de lo residuos.*

$$e_i = y_i - \hat{\eta}_{x_i} = y_i - \hat{\alpha} - \hat{\beta}\, x_i$$

A e_i se le llama *residuo* de y_i, cantidad residual resultante por la substitución del valor observado y_i con la estimación

$$\hat{\eta}_{x_i} = \hat{\alpha} + \hat{\beta}\, x_i$$

La recta de los mínimos cuadrados para el origen

Hay situaciones en las cuales, bajo como está planteado el problema, si $x = 0$ debería ser también $y = 0$, es decir la recta de regresión debería pasar por el origen.

En este caso se debería utilizar un modelo lineal *sin* α

$$y - \beta x = \varepsilon, \quad \forall x,$$

donde se resulta que

$$S_e = \sum_i (y_i - \hat{\beta} x_i) = 0 \text{ solo si } \bar{y} = \bar{x} = 0$$

En este caso se debería resolver el problema de optimización

$$SS_e(\beta) = \sum_i (y_i - \beta x_i)^2 = \text{mín}$$

Para esto hay que igualar a cero la derivada de SS_e con respecto de β

$$\frac{d\left(\sum_i (y_i - \beta x_i)^2\right)}{d\beta} = -2\sum_i (y_i - \beta x_i)\, x_i = 0$$

donde se consigue la ecuación normal en β: $S_{xx}\hat{\beta} = S_{xy}$ y su solución

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

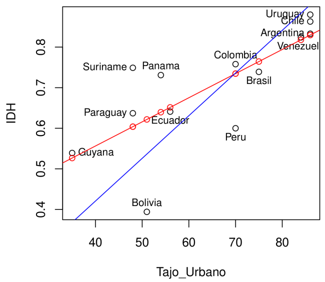
Pero esto implica que

$$S_e = \sum_i (y_i - \hat{\eta}_{x_i}) = \sum_i (y_i - \hat{\beta} x_i) = \sum_i \left(y_i - \frac{S_{xy}}{S_{xx}} x_i\right) =$$
$$= \sum_i y_i - \frac{S_{xy}}{S_{xx}} \sum_i x_i = n\bar{y} - n\hat{\beta}\bar{x} = n\bar{y} - \sum_i \hat{\eta}_{\bar{x}} = n(\bar{y} - \bar{\hat{\eta}})$$

En este caso, no se ha dicho que S$_e$ sea 0, ya que en el ejemplo se consigue:

```
> b0    <- sxy/ssx      ;b0    # = 0.01051634
> etaxm <- b0*xm        ;etaxm # = 0.6867978
> etam  <- mean(b0*x)   ;etam  # = 0.6867978
> ym                           # = 0.7067692
> ym-etam                      # = 0.01997144
> Se    <- sum(y-b0*x) ; Se    # = 0.2596287
> Se/n                  ;      # = 0.01997144
```


Por lo tanto se resulta no solo que la solución no es optima, ya que el origen es un vínculo para la recta, pero también no hay coincidencia entre el promedio $\bar{y}$ de los y_i observados y el promedio $\bar{\hat{\eta}}$ de su estimadores $\hat{\eta}_{x_j}$. De hecho, S_e es proporcional a este desvío.



Contribución y apalancamiento

Vemos ahora como estudiar la contribución de las observaciones (x_i, y_i) a la determinación de la recta de regresión.
Como se sabe, esta pasa para el baricentro $(\bar{x}, \bar{y})$ de la nube de puntos y es enteramente determinata para su pendiente $\hat{\beta}$, cuya determinación se puede escribir

$$\hat{\beta} = \frac{\Sigma_i (y_i - \bar{y})(x_i - \bar{x})}{\Sigma_i\, (x_i - \bar{x})^2} = \sum_{i=1}^{n} \left(\frac{(x_i - \bar{x})^2}{\Sigma_l (x_l - \bar{x})^2} \right) \frac{y_i - \bar{y}}{x_i - \bar{x}} = \sum_i w_i b_i$$

Como $b_i = \frac{y_i - \bar{y}}{x_i - \bar{x}}$ es un pendiente, $\hat{\beta}$ es el promedio pesado de los pendientes de las rectas que unen cada (x_i, y_i) con $(\bar{x}, \bar{y})$ con peso w_i proporcional al cuadrado de su distancia de x_i de $\bar{x}$.

Se puede escribir también $\hat{\beta}$ como combinación lineal de los y_i; efectivamente

$$\hat{\beta} = \frac{\Sigma_i (y_i - \bar{y})(x_i - \bar{x})}{\Sigma_i\, (x_i - \bar{x})^2} = \frac{\Sigma_i (x_i - \bar{x}) y_i}{\Sigma_j (x_j - \bar{x})^2} = \sum_{i=1}^{n} c_i y_i \qquad (9)$$

con $c_i = \frac{x_i - \bar{x}}{\Sigma_j (x_j - \bar{x})^2}$. Pues $\hat{\eta}_i = \hat{\alpha} + \hat{\beta} x_i$, resulta

$$\hat{\eta}_i = \frac{1}{n} \sum_{j=1}^{n} y_j - (\sum_{j=1}^{n} c_j y_j)\bar{x} + (\sum_{j=1}^{n} c_j y_j) x_i = \sum_{j=1}^{n} h_{ij} y_j,$$

notando

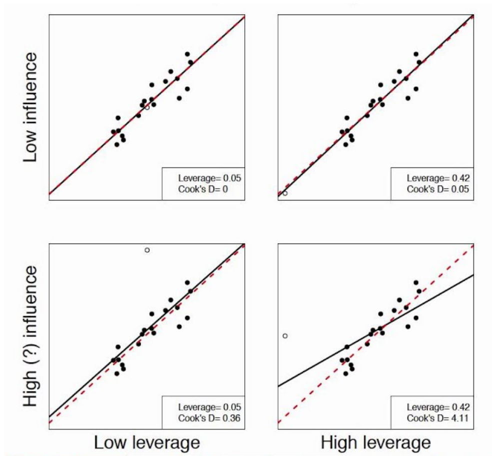
$$h_{ij} = \frac{1}{n} + c_j (x_i - \bar{x}) = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\Sigma_{l=1}^{n} (x_l - \bar{x})^2}.$$

La *matriz sombrero* (simétrica) $\hat{H} = (h_{ij})$ tiene valores que indican el impacto de cada observación en la estimación de todas las otras. Como los x_j son predefinidos, la regresión solo depende de los y. Resulta que más es lejos x_j de $\bar{x}$, más y_j influye sobre la determinación de $\hat{\eta}_i$.

Por cada observación, h_{ii} mida el impacto de y_i sobre $\hat{\eta}_i$. A

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma_{l=1}^{n} (x_l - \bar{x})^2} = \frac{1}{n} + w_i$$

se le llama *apalancamiento* del punto (x_i, y_i) (*leverage* en inglés). Nótese que se tiene $\Sigma_i\, h_{ii} = 2$.



```
w   <- (x-xm)^2 / ssxc ; w      # w pesos de los pendientes
b   <- (y-ym)/(x-xm) ; b        # pendientes
bh2 <- t(w)%*%b ; bh2           # otra determinación de bh
c   <- (x-xm) / ssxc ; c        # coeficientes de beta según y
bh3 <-  t(c)%*%y ; bh3          # otra determinación de bh
Hs  <- matrix(0,n,n)            # definición de H sombrero
rownames(Hs)= lab               # inclusivo de su etiquetas
colnames(Hs)=lab
for (i in 1:n) {                # construcción de H sombrero
  for (j in 1:n) {
    Hs[i,j] <- 1 / n + c[j] * (x[i] - xm)
  }
} ; Hs
sum(Hs)                         # Hs es centrada
lev   <- diag(Hs); lev          # apalancamientos
lev   <- 1/n + w; lev
pesos <- cbind(x,w,b,y,c,lev)   # salida conjunta
rownames(pesos) <- lab ; pesos  # salida
```


	x	w	b	y	c	lev
Argentina	86	0.115448433	0.006100372	0.833	0.005579292	0.19237151
Bolivia	51	0.055196225	0.021860215	0.394	-0.003857801	0.13211930
Brasil	75	0.025329381	0.003325397	0.739	0.002613349	0.10225246
Chile	86	0.115448433	0.007550186	0.863	0.005579292	0.19237151
Colombia	70	0.005936673	0.010918033	0.758	0.001265193	0.08285975
Ecuador	56	0.023358999	0.007066116	0.641	-0.002509645	0.10028208
Guyana	35	0.247671439	0.005535533	0.539	-0.008171900	0.32459452
Panama	54	0.034476102	-0.002142857	0.731	-0.003048907	0.11139918
Paraguay	48	0.080769709	0.004031111	0.637	-0.004666694	0.15769279
Peru	70	0.005936673	-0.022754098	0.600	0.001265193	0.08285975
Suriname	48	0.080769709	-0.002440000	0.749	-0.004666694	0.15769279
Uruguay	86	0.115448433	0.008371747	0.880	0.005579292	0.19237151
Venezuela	84	0.094209789	0.006271605	0.824	0.005040030	0.17113287

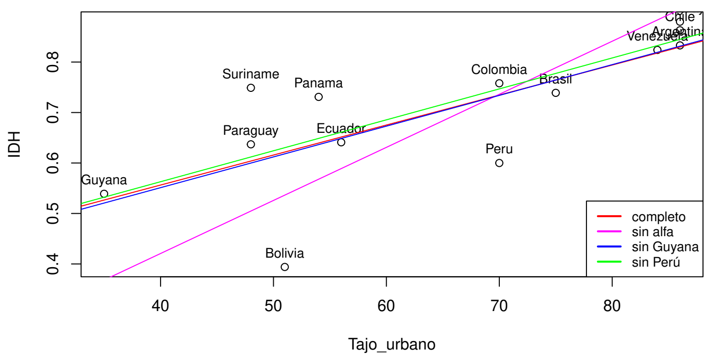
Se nota que la Guyana tiene el máximo peso y máximo apalancamiento y Perú el pendiente máximo.

Vemos como influyen sobre la regresión estos valores:

```
lm1 <- lm(IDH~Tajo_urbano,data=SudAmerica); lm1
lm2 <- lm(IDH~Tajo_urbano-1,data=SudAmerica); lm2
lm3 <- lm(IDH~Tajo_urbano,data=SudAmerica[c(1:6,8:13),]); lm3
lm4 <- lm(IDH~Tajo_urbano,data=SudAmerica[c(1:9,11:13),]); lm4 plot(
text(Tajo_urbano,IDH,labels = lab, cex = 0.8, pos=3)
  abline(lm1,col="red")                  # modelo completo
  abline(lm2,col="magenta")              # modelo sin alfa
  abline(lm3,col="blue")                 # modelo sin Guyana
  abline(lm4,col="green")                # modelo sin Perú
```

Efectivamente se resulta una variación de los parámetros:

	(Intercept)	Tajo_urbano
completo	0.318545	0.005945
sin alfa		0.01052
sin Guyana	0.307338	0.006095
sin Perú	0.31771	0.00613



Se resulta que la falta de intercepta influye mucho mientras, en este caso, ni fuerte apalancamientos ni fuerte pendiente influyen mucho.