Análisis de regresión

24 de agosto
(semana 3)

Plan de trabajo

1. Introducción a la regresión lineal simple
2. Inferencia sobre los coeficientes de regresión

Nota:

- No olviden aceptar la invitación al CLASSROOM y revisar acceso al DRIVE. Pendiente de mi parte crear MOODLE.
- La primera entrega del trabajo final será para el domingo 4 de septiembre por CLASSROOM.
  - Llenen la lista que les compartí con los integrantes.
  - Si van a recoger sus datos, NO empiecen hasta que yo les dé el aval.
  - En el DRIVE, videos de repaso de estadística descriptiva.
- En la semana del 29 de agosto al 2 de septiembre no habrá clase. Usen ese tiempo para la 1era entrega y el taller 1.
- Impriman por favor material módulo 2.

Repaso de probabilidad

Repaso: operador covarianza

Teorema: propiedades de la covarianza

(1) $\operatorname{cov}(X,Y)=E[XY]-E[X]E[Y]$;

(2) $(Simetría)$ $\operatorname{cov}(X,Y)=\operatorname{cov}(Y,X)$;

(3) (*Positividad semidefinida*) $\operatorname{cov}(X,X)=Var(X)$;

(4) (*Linealidad* 1) $\operatorname{cov}(X,d)=0,\ \forall d\in\mathbb{R}$;

(5) (*Linealidad* 2) $\operatorname{cov}(aX+b,Y)=\operatorname{cov}(aX,Y)+\operatorname{cov}(b,Y)$

$=a\operatorname{cov}(X,Y),\ \forall a,b\in\mathbb{R}$;

(6) (*Linealidad* 3) $\operatorname{cov}(aX+b,cY+d)=ac\operatorname{cov}(X,Y)$,

$\forall a,b,c,d\in\mathbb{R}$.

(7) (*Independencia* $\Rightarrow \operatorname{cov}=0$) Si X,Y son independientes,

$\operatorname{cov}(X,Y)=0$.

Repaso: operador covarianza

Teorema: propiedades de la covarianza (II)

Para $X, Y, X_1, X_2, \ldots, X_n \in L^2(\Omega, \Im, p)$, se tiene

(1) Desigualdad de Cauchy-Schwarz:

$$\left|\operatorname{cov}(X,Y)\right| \leq \sqrt{Var(X)}\sqrt{Var(Y)}.$$

(2) $Var(X+Y) = Var(X) + Var(Y) + 2\operatorname{cov}(X,Y)$.

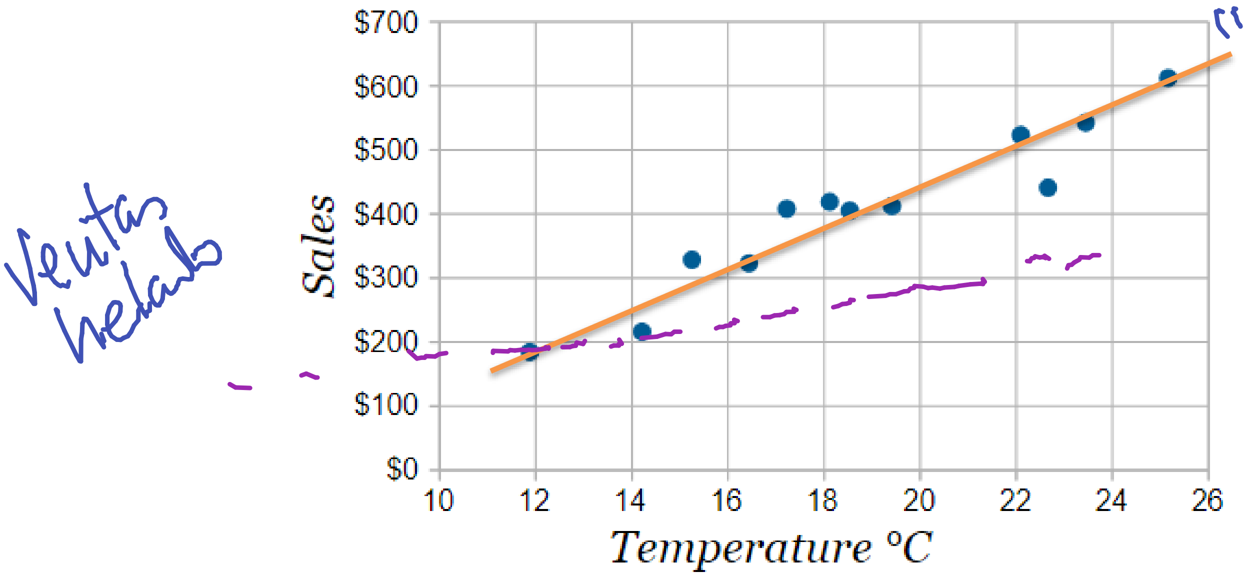
(3) $$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + \sum\sum_{i \neq j} \operatorname{cov}(X_i, X_j)$$

$$= \sum_{i=1}^{n} Var(X_i) + 2\sum\sum_{i<j} \operatorname{cov}(X_i, X_j).$$

Regresión lineal simple

Coeficientes de regresión

- Estos coeficientes determinan la “mejor” relación lineal que puede ser descrita entre las dos variables



$$“y \approx \beta_0 + \beta_1 x”$$

Fuente: https://www.mathsisfun.com/data/scatter-xy-plots.html

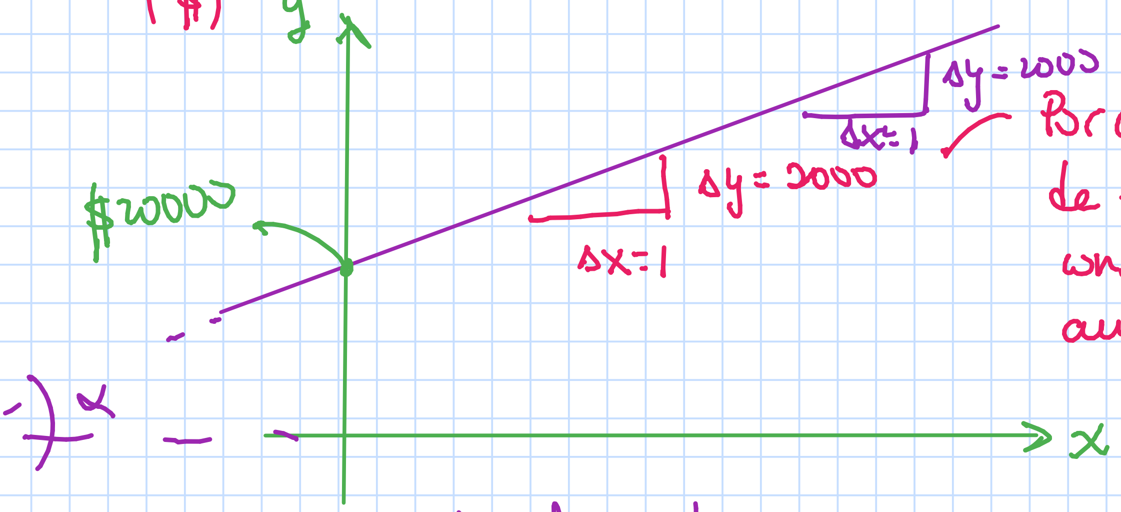
- Mucho más útil porque permite hacer predicciones de una variable en función de la otra (una vez conozcamos estimaciones de los parámetros).

- ¿Cómo encontrar los valores de los coeficientes?

$$y = \underbrace{20000}_{\text{Intercepto}} + \underbrace{2000}_{\text{pendiente}}\, x$$

costo recibo H_2O ($)

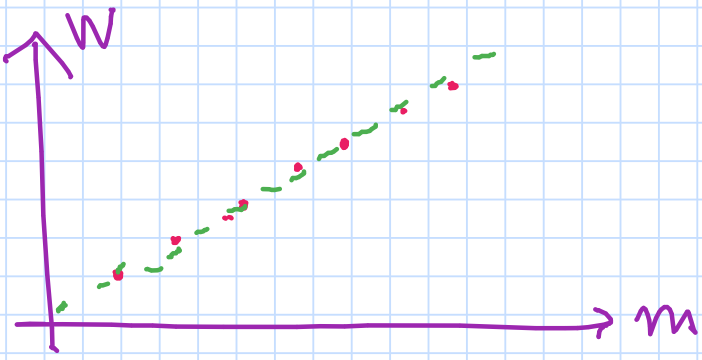
consumo (m^3)



Por cada incremento de $1m^3$ en el consumo, el costo aumentará $2000

Pendiente $= \tan\alpha$

$W = mg$



Modelo de los helados

$$\underbrace{Y}_{\text{Ventas por helados en un día}} = \boxed{\beta_0 + \beta_1 \tilde{X}} + \boxed{\varepsilon}$$

Temp (°C)

comp. aleatoria

Brecha no explicada por la componente sistemática

Componente sistemática (μ)

Modelo de regresión lineal simple

$$\begin{cases} Y_k = \mu_k + \varepsilon_k \\ \mu_k = \beta_0 + \beta_1 x_k \\ \varepsilon_1, \varepsilon_2, \dots, \varepsilon_k \quad m.a. \\ E[\varepsilon_i] = 0 \\ Var[\varepsilon_i] = \sigma^2 \ \forall i \end{cases}$$

μ_k: comp. sistem. ε_k: comp. aleatoria

3 parámetros

$E[\varepsilon_i] = 0$ (S1)

$Var[\varepsilon_i] = \sigma^2 \ \forall i$ (S2. Homoscedasticidad)

$$E[Y_k] = E[\mu_k + \varepsilon_k] = E(\mu_k) + E[\cancel{\varepsilon_k}]^{\to 0}$$
$$= E[\beta_0 + \beta_1 x_k]$$
$$= \beta_0 + \beta_1 x_k$$

β_0: Ventas promedio o esperadas cuando $x = 0°C$

β_1: Por cada 1°C que aumente la temperatura, las ventas esperadas o promedio cambiarán β_1 unidades

Supuestos del modelo de regresión lineal simple

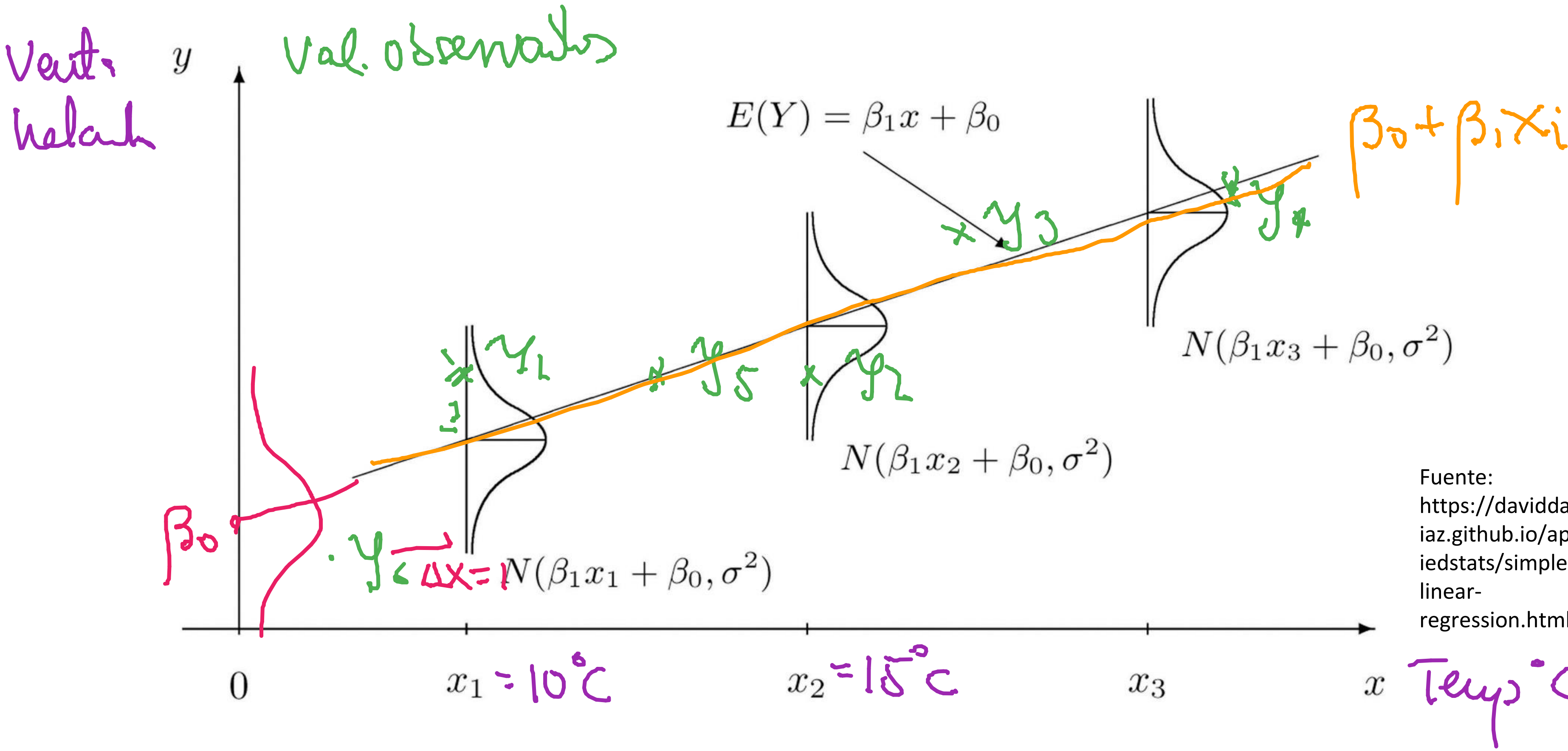
Supuestos del modelo de regresión lineal simple

- La variable X (cualitativa o cuantitativa) se asume siempre condicionante, es decir, dado que $X = x$.
- La relación entre ambas variables es lineal, es decir:

$$Y_i \Big| \left\{ X = x_i \right\} = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- $\left\{ \varepsilon_i \right\}$ son una m.a. **generalmente** de una distr. $N\left(0, \sigma^2\right)$
  - × Las $\left\{ Y_i \right\}$ no son una m.a.
  - × $Y_i \Big| \left\{ X = x_i \right\} \sim N\left(\boxed{\beta_0 + \beta_1 x_i}, \sigma^2\right), \forall i.$

Supuestos del modelo de regresión simple (II)



Fuente: https://daviddalpiaz.github.io/appliedstats/simple-linear-regression.html

- ¿Por qué no usar los puntos correspondientes a un valor de *x* para estimar los parámetros de cada curva?

Supuestos del modelo de regresión lineal simple

- La lógica detrás de que la variable X no se asuma como aleatoria viene de asumir que esa variable es un factor controlado en un **estudio experimental**.
- Sin embargo, también puede asumirse como aleatoria y por ende, se habla de la distribución condicional una vez que se conoce X. Según esta lógica, aunque X no es controlable, sí es más fácil de conocer en un **estudio observacional**.
- En ambos casos, ojo con la **confusión de efectos**.

Fuente: https://researchhubs.com/post/ai/data-analysis-and-statistical-inference/observational-studies-and-experiments-sampling-and-source-bias.html

Estimación de los coeficientes de regresión

❖ **Parámetros:** β_0 : intercepto, β_1 : pendiente

Método de mínimos cuadrados ordinarios: Suponga que se tiene el siguiente modelo:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad 1 \le i \le n$$

Donde:

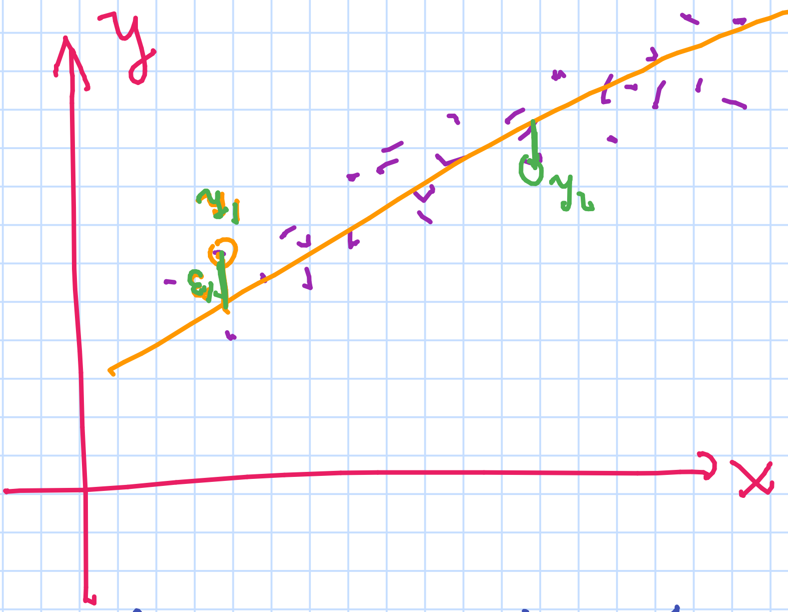
$$1.\ E(\varepsilon_i) = 0 \ \forall i$$

$$2.\ Var(\varepsilon_i) = \sigma^2 \ \forall i$$

$$3.\ Cov(\varepsilon_i, \varepsilon_j) = 0 \ \forall i \neq j$$

La estimación de mínimos cuadrados (dados los datos) es:

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = \arg\min_{\beta_0, \beta_1} Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$



$(\underbrace{y_1}_{\text{observé}} \underbrace{- \beta_0 - \beta_1 x_1}_{\text{recta del}})^2 +$

$(y_2 - \beta_0 - \beta_1 x_2)^2 +$

$(y_3 - \beta_0 - \beta_1 x_3)^2 \ldots$

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial Q}{\partial \beta_0} = \frac{\partial}{\partial \beta_0}\left\{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2\right\}$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_0}\left\{(y_i - \beta_0 - \beta_1 x_i)^2\right\}$$

$$= \sum_{i=1} -2(y_i - \beta_0 - \beta_1 x_i)$$

$$= -2\left[\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \beta_0 - \sum_{i=1}^{n} \beta_1 x_i\right]$$

$$= -2\left[n\bar{y} - n\beta_0 - n\beta_1 \bar{x}\right] = -2n\left[\bar{y} - \beta_0 - \beta_1 \bar{x}\right] \quad (1)$$

$$\frac{\partial Q}{\partial \beta_1} = \frac{\partial}{\partial \beta_1}\left\{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2\right\}$$

$$= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_1}\left\{(y_i - \beta_0 - \beta_1 x_i)^2\right\}$$

$$= \sum_{i=1} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) \quad (2)$$

$$= -2\left(\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \beta_0 x_i - \sum_{i=1}^{n} \beta_1 x_i^2\right)$$

$$= -2\left(\sum_{i=1}^{n} y_i x_i - \beta_0 n\bar{x} - \beta_1 \sum_{i=1}^{n} x_i^2\right) \quad (2)$$

Igualando a 0 (1) y (2):

$$\cancel{-2n}\left[\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}\right] = 0 \quad (1')$$

$$\cancel{-2}\left[\sum_{i=1}^{n} y_i x_i - \hat{\beta}_0 n\bar{x} - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2\right] = 0 \quad (2')$$

$\vdots$

Solución del problema de minimización (I)

1. Puntos críticos.

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -2\displaystyle\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \ (1) \\ \dfrac{\partial Q}{\partial \beta_1}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -2\displaystyle\sum_{i=1}^{n} x_i\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \ (2) \end{cases}$$

$$\begin{cases} \overline{y} - \hat{\beta}_0 - \hat{\beta}_1\overline{x} = 0 \therefore \hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} \ (1') \\ \displaystyle\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0\sum_{i=1}^{n} x_i - \hat{\beta}_1\sum_{i=1}^{n} x_i^2 = 0 \ (2') \end{cases}$$

Reemplazando $\hat{\beta}_0$ de (1') en (2') y despejando, se tiene que

$$\sum_{i=1}^{n} x_i y_i - \left(\overline{y} - \hat{\beta}_1\overline{x}\right)\sum_{i=1}^{n} x_i - \hat{\beta}_1\sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - n\overline{xy} - n\hat{\beta}_1\overline{x}^2 - \hat{\beta}_1\sum_{i=1}^{n} x_i^2 = 0 \therefore \hat{\beta}_1 = \frac{\displaystyle\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\displaystyle\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\displaystyle\sum_{i=1}^{n}\left(x_i - \overline{x}\right) y_i}{\displaystyle\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$$

Solución del problema de minimización (II)

2. Mínimo local (y global).

$$\begin{cases} \dfrac{\partial^2 Q}{\partial \beta_0^2}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -2\sum_{i=1}^{n}(-1) = 2n \\ \dfrac{\partial^2 Q}{\partial \beta_1^2}\left(\hat{\beta}_0, \hat{\beta}_1\right) = -2\sum_{i=1}^{n} x_i\left(-x_i\right) = 2\sum_{i=1}^{n} x_i^2 \\ \dfrac{\partial^2 Q}{\partial \beta_0 \partial \beta_1}\left(\hat{\beta}_0, \hat{\beta}_1\right) = 2\sum_{i=1}^{n} x_i \end{cases} \qquad \mathbf{J} = \begin{pmatrix} 2n & 2\sum_{i=1}^{n} x_i \\ 2\sum_{i=1}^{n} x_i & 2\sum_{i=1}^{n} x_i^2 \end{pmatrix}$$

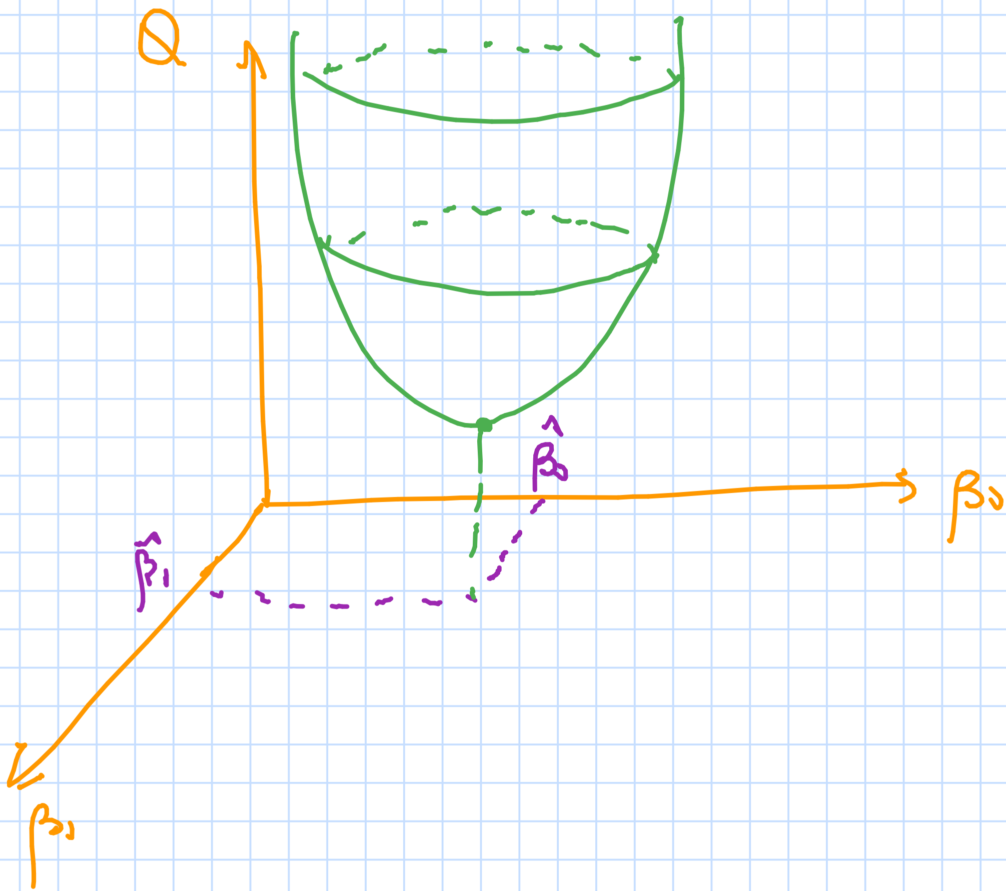
Luego, por el criterio de Sylvester, como

$$2n > 0 \text{ y } \det\left(\mathbf{J}\right) = 4n\sum_{i=1}^{n} x_i^2 - 4n^2\bar{x}^2 = 4n\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 > 0;$$

por ende, $\mathbf{J}$ es positiva definida, lo que implica que la solución es un mínimo local.

Como $\mathbf{J}$ no depende de $\hat{\beta}_0, \hat{\beta}_1$; quiere decir que Q es convexa y eso hace que la solución sea global.

No es difícil ver que $\lim_{\|\boldsymbol{\beta}\| \to \infty} Q\left(\beta_0, \beta_1\right) = \infty$; lo que completa la prueba.



Q
β_0
β_1
$\hat{\beta}_0$
$\hat{\beta}_1$

Una definición importante y un lema

Mejor estimador linealmente insesgado (BLUE)

Sean $Y_1, Y_2, \ldots, Y_n$ variables aleatorias involucradas en un proceso de estimación. Se dice que un estimador de la forma $\sum_{i=1}^{n} \phi_i Y_i$ con constantes ϕ_i no aleatorias y conocidas es un **BLUE** si

1) es un estimador insesgado,

2) es el estimador con menor varianza dentro de todos los estimadores lineales insesgados.

- Como siempre la definición no es constructiva. Usaremos el lema del máximo (Lema 11.2.7 de Casella & Berger) más adelante para verificar que un estimador es BLUE.