Análisis y Visualización de Datos

Clase 4

Inferencia Bayesiana
Regresión Lineal

Docentes : Soledad Palacios(UNLP)

Milagro Teruel (UNC)

Nuestros golems raramente tienen forma física, pero a menudo están hechos de arcilla y viven *in silicio* como código de computadora -Richard McElreath

Filosofía bayesiana

La teoría de la probabilidad se ocupa del estudio de la incertidumbre, y del comportamiento de los fenómenos o experimentos aleatorios.
La probabilidad depende de dos elementos:

- el evento incierto
- las condiciones bajo las cuales es considerado

por lo que desde este punto de vista la probabilidad es siempre condicional. La estadística es una herramienta para la toma de decisiones bajo condiciones de incertidumbre.

Filosofía bayesiana

La razón de querer medir no es sólo para ser más precisos respecto a la intensidad de la incertidumbre sino también para combinar incertidumbres: en un problema típico de estadística encontramos combinadas la incertidumbre de los datos y la del parámetro.

Las discusiones en contra del enfoque bayesiano se centran en el punto de medir la incertidumbre sobre el parámetro probabilísticamente, esto es, darle tratamiento de variable aleatoria. Para la estadística frecuentista existe algo que llaman “el verdadero valor del parámetro” que consideran fijo y que “sólo Dios conoce” pero que resulta desconocido para nosotros los mortales.

Inferencia Bayesiana

La Estadística Bayesiana modela la incertidumbre que tenemos sobre θ probabilísticamente, esto es, consideramos al valor de θ como una variable (o vector) aleatoria (v.a.) con una distribución de probabilidad a priori (o inicial) p(θ). Se trata de una distribución basada en experiencia previa (experiencia de especialistas, datos históricos, etc.) antes de obtener datos muestrales nuevos.

Si no estamos seguros de la factibilidad de un evento, entonces es matemáticamente razonable asignar un valor entre 0 y 1, de acuerdo al grado de confianza que tenemos de que ocurra dicho evento. Los extremos son los casos especiales de absoluta certeza p(A=0)= falso y p(A=1) = verdadero

Teorema de Bayes

Luego procedemos a observar los nuevos datos (obtención de la muestra) x := (x 1 , . . . , x n) y combinamos esta información con la distribución a priori mediante la Regla de Bayes y obtenemos una distribución de probabilidad a posteriori (o final) :

$$p(\theta \mid \mathbf{x}) = \frac{p(\mathbf{x}, \theta)}{p(\mathbf{x})} = \frac{p(\mathbf{x} \mid \theta) p(\theta)}{\int_{\Theta} p(\mathbf{x} \mid \tilde{\theta}) p(\tilde{\theta})\, d\tilde{\theta}}$$

Teorema de Bayes

De manera general, el modelo especifica:

➔ ***p(datos|valores de parámetros y estructura del modelo)***

y usamos la regla de Bayes para convertir la expresión anterior a lo que nos interesa de verdad, que es, que tanta certidumbre tenemos del modelo condicional a los datos:

➔ ***p(valores de parámetros y estructura del modelo|datos)***

Teorema de Bayes

Ventaja

Podemos definir estimadores sin restricciones poco naturales o intuitivas como la insesgadez, consistencia, eficiencia, etc del enfoque frecuentista

Ejemplos

- La probabilidad a posteriori de que un coeficiente de regresión sea positivo (negativo, nulo,…)
- La probabilidad a posteriori de que un sujeto pertenezca a tal grupo
- La probabilidad a posteriori de que la hipótesis H sea cierta
- La probabilidad a posteriori de que cierto modelo estadístico sea el auténtico modelo de entre un conjunto de ellos. …

Bayesianos vs frecuentistas

En cambio, la inferencia “frecuentista” (o clásica) :

- Propone un modelo para los datos: y~ f(θ)
- La estimación de θ se basa en un estimador (una función de los datos $\hat{\theta} \cong \hat{\theta}$ (y))
- Para el contraste de cierta hipótesis nula se trabaja con la distribución del estimador condicionada a que la hipótesis sobre el parámetro es cierta: “asumiendo Ho en un proceso reiterado de muestreo cómo de frecuente sería obtener un resultado al menos tan extremo como el obtenido”. Si la respuesta a esta pregunta acaba siendo “muy infrecuente”; menos que el nivel de significación se rechazará la hipótesis.

Bayesianos vs frecuentistas

- Pero actuar con esa racionalidad no es, a menudo, fácil de explicar de forma intuitiva.

- En la Inferencia Clásica trata a los estimadores ($\hat{\theta}$) y a los propios datos como “variables aleatorias” mientras que los parámetros son rasgos fijos (aunque desconocidos) de la población de la que se obtiene la muestra “aleatoriamente”.

Bayesianos vs frecuentistas

	Bayesiana	**Clásica**
θ	Aleatorio (=desconocimiento)	Constante,aunque desconocida
$\hat{\theta}$	Fijo (datos)	Aleatorio
aleatoriedad	(conocimiento parcial) Subjetivo	Viene del muestreo
Distribución importante	A posteriori	D. muestral

Destripando la ecuación

Cada término del teorema de Bayes tiene un nombre específico:

p($\boldsymbol{\theta}$ | y)---------- a posteriori
p(y | $\boldsymbol{\theta}$)---------- likelihood (verosimilitud)
p(\theta)---------- a priori
p(y)---------- likelihood marginal

El a priori

El a priori es la forma de introducir conocimiento previo sobre los valores que pueden tomar los parámetros. A veces cuando no sabemos demasiado se suelen usar a prioris que asignan igual probabilidad a todos los valores de los parámetros, otras veces se puede elegir a prioris que restrinjan los valores de los parámetros a rangos razonables, algo que se conoce como regularización, por ejemplo solo valores positivos. Muchas veces contamos con información mucho más precisa como medidas experimentales previas o límites impuesto por alguna teoría.

likelihood

El likelihood es la forma de incluir nuestros datos en el análisis. Es una expresión matemática que especifica la plausibilidad de los datos. El likelihood es central tanto en estadística Bayesiana como en estadística no-Bayesiana. A medida que la cantidad de datos aumenta el likelihood tiene cada vez más peso en los resultados, esto explica el porqué a veces los resultados de la estadística Bayesiana y frecuentista coinciden cuando la muestra es grande.

El a posteriori

El a posteriori es la distribución de probabilidad para los parámetros. Es la consecuencia lógica de haber usado un conjunto de datos, un likelihood y un a priori. Se lo suele pensar como la versión actualizada del a priori. De hecho un a posteriori puede ser un a priori de un análisis a futuro.

El *a posteriori* representa todo lo que sabemos de un problema, dado un modelo y un conjunto de datos. Y por lo tanto todas las inferencias estadísticas pueden deducirse a partir de él. Típicamente esto toma la forma de integrales como la siguiente.

El a posteriori

Donde podemos ver el a posteriori como una especie de promedio pesado, de hecho es la esperanza de la funcion de verosimilitud, donde anotamos a cada valor que pueda tomar el parametro la probabilidad del mismo

$$\bar{\theta} = \int \theta \; p(\theta \mid y) d\theta$$

El likelihood marginal o evidencia

La likelihood marginal (también llamado evidencia) es la probabilidad de observar los datos promediado sobre todas las posibles hipótesis (o conjunto de parámetros) $\boldsymbol{\theta}$.

En general, la evidencia puede ser vista como una simple constante de normalización que en la mayoría de los problemas prácticos puede (y suele) omitirse sin pérdida de generalidad. Por lo que el teorema de Bayes suele aparecer escrito como:

$$p(\theta \mid y) \propto p(y \mid \theta)p(\theta)$$

Estimación Puntual Bayesiana

consistirá en un número que resuma adecuadamente la distribución a posteriori.

Pero ¿Cuál? ¿Su media , su moda, su mediana algún cuantil?

Porque elegir la beta como familia paramétrica

Hay varias razones para usar una distribución beta para este y otros problemas:

- La distribución beta varía entre 0 y 1, de igual forma que lo hace en nuestro modelo.
- Esta distribución combinada con la que elegiremos como *likelihood* (ver más adelante), nos permitirá resolver el problema de forma analítica.
- Es una distribución versátil para expresar distintas situaciones.

Ejemplo

Trataremos de determinar el grado en que una moneda está sesgada. En general cuando se habla de sesgo se hace referencia a la desviación de algún valor (por ejemplo, igual proporción de caras y cecas), pero aquí usaremos el término *sesgo* de forma más general. Diremos que el sesgo es un valor en el intervalo [0, 1], siendo 0 para una moneda que siempre cae ceca y 1 para una moneda que siempre cae cara y lo representaremos con la variable θ . A fin de cuantificar θ arrojaremos una moneda al aire repetidas veces, por practicidad arrojaremos la moneda de forma computacional (¡pero nada nos impide hacerlo manualmente!). Llevaremos registro del resultado en la variable y . Siendo y la cantidad de caras obtenidas en un experimento.

Ejemplo 1

1. Proponer una familia paramétrica para el experimento anterior,
2. proponer una distribución a priori para el parámetro del modelo, tomando especialmente en cuenta que no estamos seguros de que la moneda fue tomada al azar,
3. obtener la distribución predictiva a priori,
4. obtener la distribución a posteriori,
5. obtener la distribución predictiva a posteriori.

Ejemplo 1

En este caso podemos sospechar que la familia paramétrica es Bernoulli, ya que estamos en presencia de una experiencia con 2 posibilidades concretas y discretas

$$P = \{Ber(x \mid \theta) : \theta \in \Theta\}$$

Donde

$$Ber(x \mid \theta) = \theta^{x} (1 - \theta)^{1-x} (x)$$

Ejemplo 1

La distribución a priori juega un papel fundamental en el análisis bayesiano, ya que calcula el grado de conocimiento inicial que posee los parámetros en estudio.

En la distribución a priori, los parámetros son llamados hiperparametros, para diferenciarlos de los parámetros del modelo. Por ejemplo, utilizando una distribución Beta para modelar la distribución del parámetro, hay que tener en cuenta que éste es una variable aleatoria con distribución Beta y que alfa y beta son hiperparámetros de la distribución a priori

El a priori

El *a priori* lo modelaremos usando una distribución beta, que es una distribución muy usada en estadística Bayesiana. La *pdf* de esta distribución es:

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Constante de normalización de la Beta

El a priori

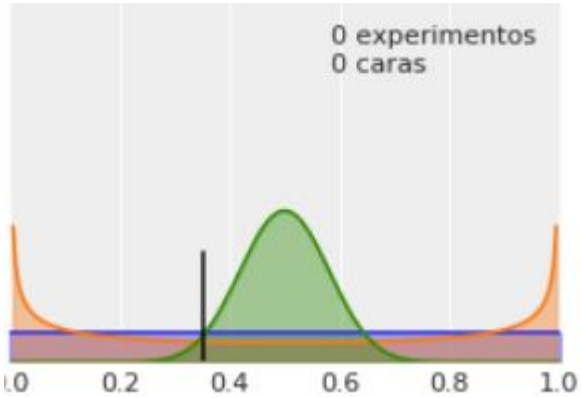
Hay varias razones para usar una distribución beta para este y otros problemas:

- La distribución beta varía entre 0 y 1, de igual forma que lo hace θ en nuestro modelo.
- Esta distribución combinada con la que elegiremos como *likelihood* (ver más adelante), nos permitirá resolver el problema de forma analítica.
- Es una distribución versátil para expresar distintas situaciones.

Calculando el a posteriori

Respecto al último punto, veamos un ejemplo. Supongamos que el experimento de la moneda es realizado por tres personas. Una de ellas dice no saber nada de la moneda por lo tanto *a priori* todos los valores de θ son igualmente probables. La segunda persona desconfía de la moneda, ya que sospecha que es una moneda trucada, por lo tanto considera que está sesgada, pero no sabe si hacia cara o hacia ceca. Por último, la tercer persona asegura que lo más probable es que θ tome un valor alrededor de 0.5 ya que según su experiencia así es como se comportan las monedas. Todas estas situaciones pueden ser modeladas por la distribución beta

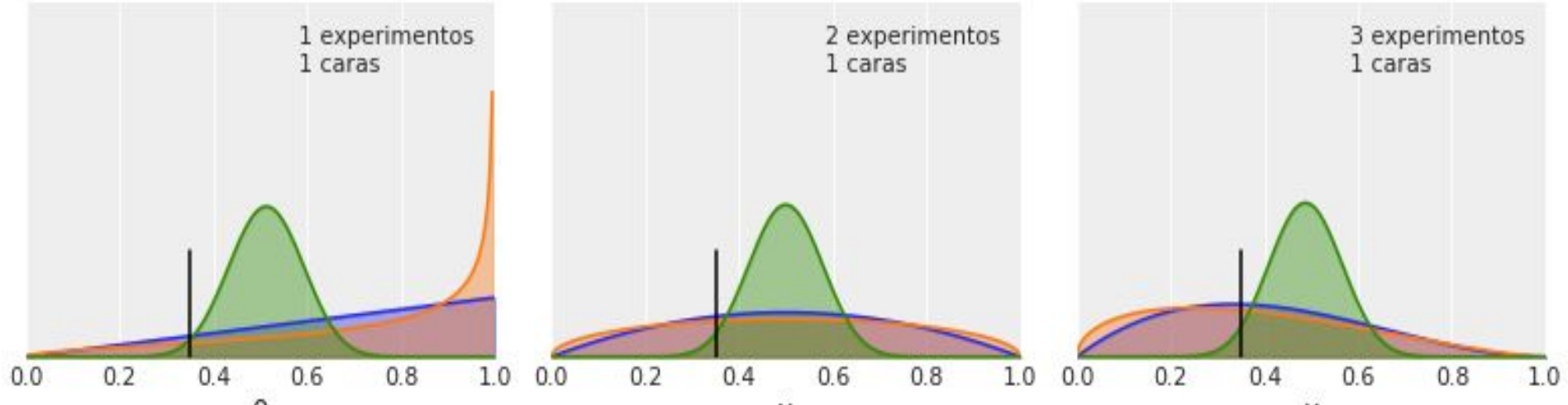
Calculando el a posteriori



- La gráfica en azul se corresponde con la probabilidad del que dice no saber nada de la moneda
- La gráfica en naranja de la persona que considera que está sesgada, pero no sabe si hacia cara o hacia ceca
- La gráfica en verde la persona que asegura que lo más probable es que θ tome un valor alrededor de 0.5

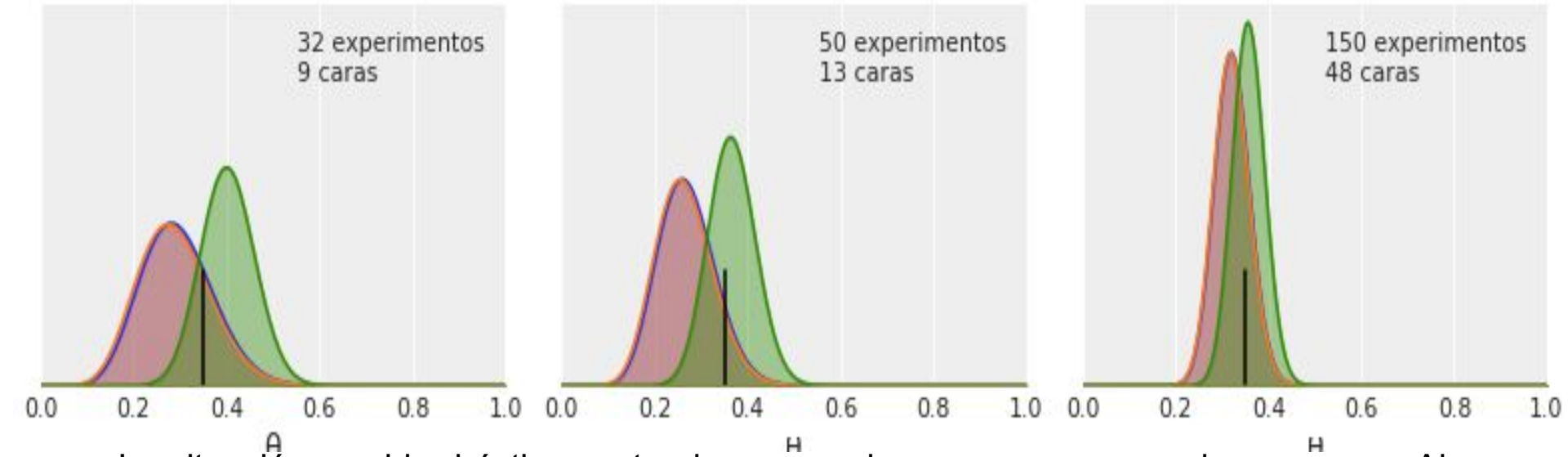
Partiendo de los *a priori* unifome (azul) o *sesgado* (anaranjado) y habiendo realizado un solo experimento y observado una sola cara, lo más razonable es pensar que estamos frente a una moneda con dos caras!

Calculando el a posteriori



- El valor más probable viene dado por la moda de la distribución (el *pico* de la distribución)
- La rapidez con la que los resultados convergen varía. En este ejemplo las curvas azul y anaranjada parecen converger con tan solo 8 experimentos, pero se necesitan más de 50 experimentos para que las tres curvas se muestren similares. Aún con 150 experimentos se observan ligeras diferencias.

Calculando el a posteriori



- La situación cambia drásticamente al ver por primera vez una moneda caer ceca. Ahora lo más probable (dado cualquiera de los tres *a prioris*) es inferir que θ=0.35. Los valores de θ exactamente 0 o 1 se vuelven imposibles.

Seguimos analizando las gráficas de θ

- El a priori no sesgado (verde) es más informativo que los otros dos (la distribución esta más concentrada), por ello se requiere de un número mas grande de experimentos para "moverlo".
- El a priori uniforme (azul) es lo que se conoce como no informativo. El resultado de un análisis Bayesiano usando un *a priori* no-informativos en general coinciden con los resultados de análisis frecuentistas (en este caso el valor esperado de $\theta = y/N$).

Seguimos analizando las gráficas de θ

- Dada una cantidad *suficiente* de datos los resultados tienden a converger sin importar el *a priori* usado.
- La rapidez con la que los resultados convergen varía. En este ejemplo las curvas azul y anaranjada parecen converger con tan solo 8 experimentos, pero se necesitan más de 50 experimentos para que las tres curvas se muestren similares. Aún con 150 experimentos se observan ligeras diferencias.

Influencia y elección del a priori

- los *a priori* influencian los resultados de nuestros cálculos.
- a medida que aumentan los datos (como las tiradas de monedas) los resultados son cada vez menos sensibles al *a priori*. (el número de iteraciones necesarias dependerá del problema y del modelo)
- Hay quienes prefieren usar *a priori* no-informativos (también conocidos como *a priori* planos, vagos, o difusos)
- también es posible usar *a prioris informativos* (o *fuertes*). Hacer esto es razonable solo si contamos con información previa confiable. Si la información no viene por el *likelihood* (datos), entonces puede venir por el *a priori*. si contás con información confiable no hay razón para descartarla, menos si el *argumento* es algo relacionado con pretender ser *objetivo*
- Muchos modelos frecuentistas pueden ser aproximados a bayesianos con a prioris planos

Influencia y elección del a priori

- los *a priori* influencian los resultados de nuestros cálculos.
- a medida que aumentan los datos (como las tiradas de monedas) los resultados son cada vez menos sensibles al *a priori*. (el número de iteraciones necesarias dependerá del problema y del modelo)
- Hay quienes prefieren usar *a priori* no-informativos (también conocidos como *a priori* planos, vagos, o difusos)
- también es posible usar *a prioris informativos* (o *fuertes*). Hacer esto es razonable solo si contamos con información previa confiable. Si la información no viene por el *likelihood* (datos), entonces puede venir por el *a priori*. si contás con información confiable no hay razón para descartarla, menos si el *argumento* es algo relacionado con pretender ser *objetivo*
- Muchos modelos frecuentistas pueden ser aproximados a bayesianos con a prioris planos

Calculando el a posteriori

- los *a priori* influencian los resultados de nuestros cálculos.
- a medida que aumentan los datos (como las tiradas de monedas) los resultados son cada vez menos sensibles al *a priori*. (el número de iteraciones necesarias dependerá del problema y del modelo)
- Hay quienes prefieren usar *a priori* no-informativos (también conocidos como *a priori* planos, vagos, o difusos)
- también es posible usar *a prioris informativos* (o *fuertes*). Hacer esto es razonable solo si contamos con información previa confiable. Si la información no viene por el *likelihood* (datos), entonces puede venir por el *a priori*. si contás con información confiable no hay razón para descartarla, menos si el *argumento* es algo relacionado con pretender ser *objetivo*
- Muchos modelos frecuentistas pueden ser aproximados a bayesianos con a prioris planos

Familias conjugadas

Verosimilitud	A priori conjugada
Bernoulli	Beta
Binomial	Beta
Multinomial	Dirichlet
Binomial Negativa	Beta
Poisson	Gamma
Exponencial	Gamma
Gamma(χ^2)	Gamma
Normal μ	Normal
Normal σ^2	Gamma Inversa
Pareto α	Gamma
Pareto β	Pareto

Pasemos a los fierros

programación probabilística

Regresión Lineal

Cuando hemos detectado que entre dos o más variables hay una relación significativa una opción es intentar matematizar esa relación, crear una fórmula matemática que materialice, formalmente, esa relación y que permita calcular pronósticos de una o de varias variables a partir del conocimiento de valores de una o de varias variables evaluadas en un individuo concreto.

Regresión Lineal

La relación matemática determinística más simple entre dos variables ***x*** y ***y*** es una relación lineal

$$y=\beta_0 + \beta_1 x$$

Si las dos variables no están determinísticamente relacionadas, entonces con un valor fijo de x, el valor de la segunda variable es aleatorio.

Regresión Lineal

- X es denominada la variable pronosticadora o independiente
- Y es la variable dependiente o de respuesta
- Con x fija, la segunda variable será aleatoria
- Un primer paso en el análisis de regresión que implica dos variables es construir una gráfica de puntos de los datos observados. En una gráfica como esa, cada (x_i, y_i) está representado como un punto colocado en un sistema de coordenadas bidimensional.

Modelo de regresión lineal simple

Existen parámetros 0 , 1 y 2 de tal suerte que con cualquier valor fijo de la variable independiente x, la variable dependiente está relacionada con x por conducto de la ecuación de modelo

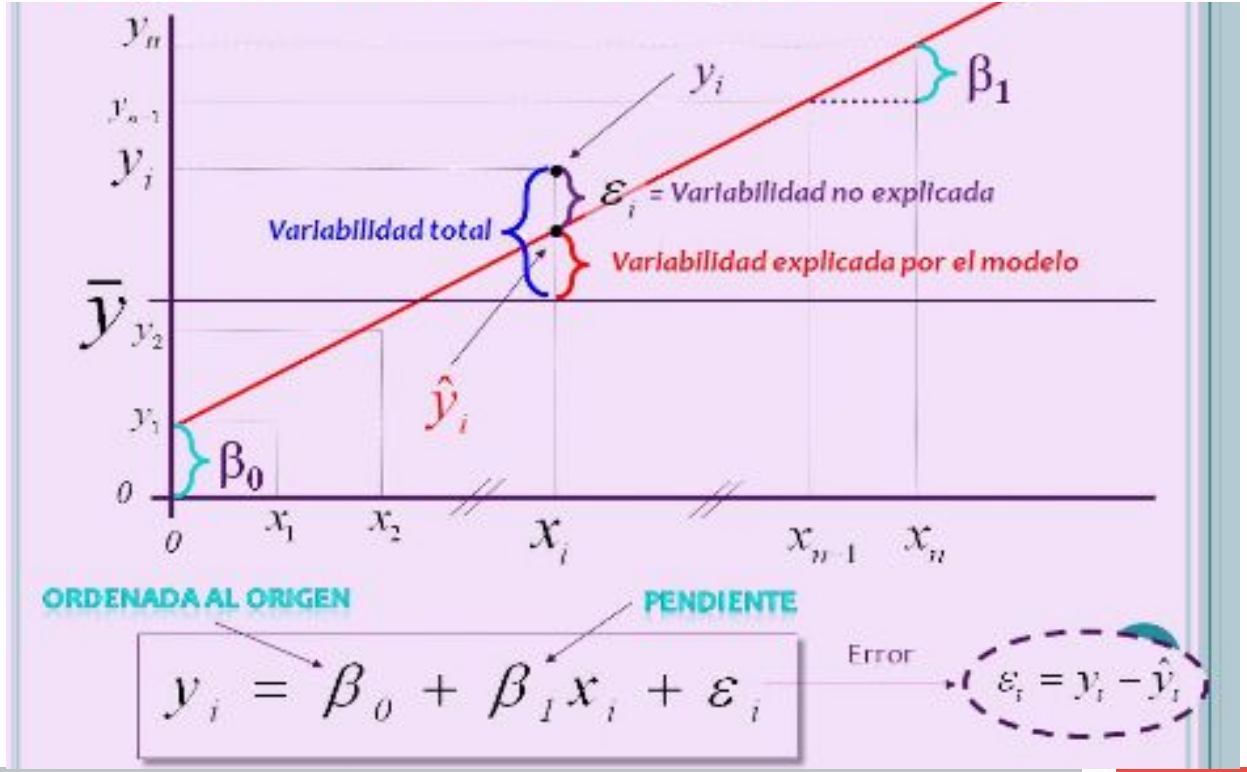
$$y=\beta_0 + \beta_1 x + \varepsilon$$

La cantidad en la ecuación de modelo es una variable aleatoria, que se supone está normalmente distribuida con $E(\varepsilon)= 0$ y $V(\varepsilon)=\sigma^2$

Término de error aleatorio

Sin , cualquier par observado (x, y) correspondería a un punto que queda exactamente sobre la línea $y=\beta_0 + \beta_1 x$, llamada línea de regresión (o de población) verdadera. La inclusión del término de error aleatorio permite que (x, y) quede o por encima de la línea de regresión verdadera (cuando $\varepsilon > 0$) o por debajo (cuando $\varepsilon < 0$).

Término de error aleatorio



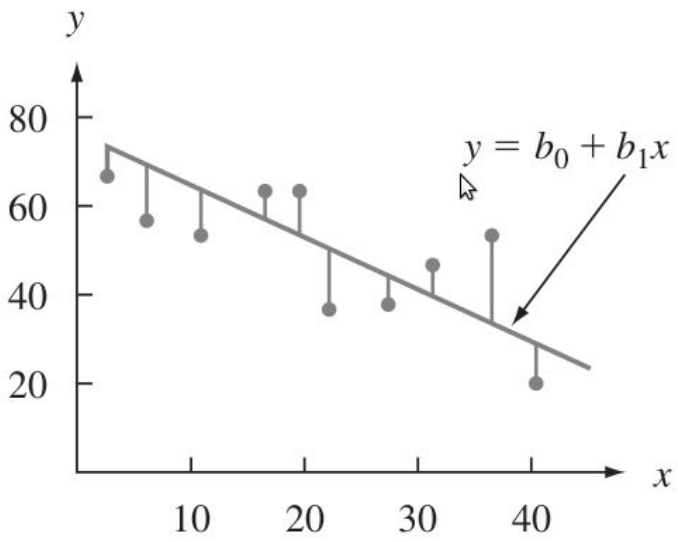
Término de error aleatorio

Un investigador casi nunca conocerá los valores de β_0, β_1 o σ^2 .
En cambio, estará disponible una muestra de datos compuesta de n pares observados $(x_1, y_1), \ldots, (x_n, y_n)$, con la cual los parámetros de modelo y la línea de regresión verdadera pueden ser estimados. Se supone que estas observaciones se obtuvieron independientemente una de otra.
Es decir, y_i es el valor observado de una variable aleatoria Y i , donde

$$y_i = \beta_0 + \beta_1 x + \varepsilon$$

y las n desviaciones, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ son variables independientes. La independencia de $Y_1, Y_2, \ldots,$ Y n se desprende de la independencia de las ε_i .

Principio de los mínimos cuadrados



- Una regla fundamental: Cuanta mayor correlación haya entre dos variables, en la representación bidimensional, estructurada en forma de recta, los valores estarán reunidos más próximos a la recta.
- Si atendemos al números de variables independientes, distinguiremos dos tipos de Regresión: la Regresión simple y la Regresión múltiple.

Principio de los mínimos cuadrados

- El método minimiza la suma de las distancias verticales entre las respuestas observadas en la muestra y las respuestas del modelo
- Es el estimador de máxima verosimilitud siempre y cuando los regresionadores sean independientes entre si (en caso de regresión múltiple) y los errores sean homocedásticos.
- En esas condiciones es un estimador de varianza mínima e insesgado siempre que los errores tengan varianza finita

Principio de los mínimos cuadrados

La **suma de cuadrados del error** (o de forma equivalente, suma de cuadrados residuales) denotada por SCE, es

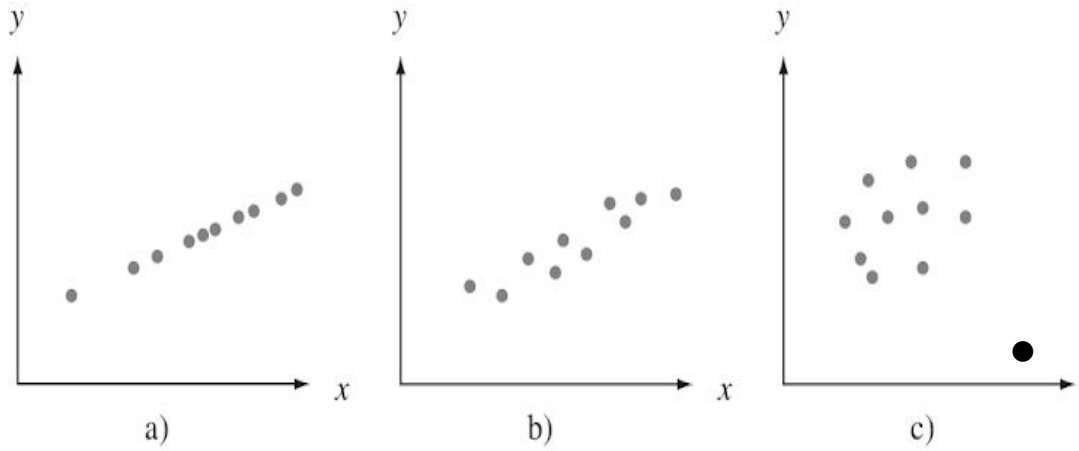
$$\text{SCE} = \sum (y_i - \hat{y}_i)^2 = \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

y la estimación de σ^2 es

$$\hat{\sigma}^2 = s^2 = \frac{\text{SCE}}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

El divisor n - 2 en s ^{2} es el número de grados de libertad (gl) asociado con la estimación. Perdimos dos gl por haber tenido que estimar beta cero y beta uno antes que s

Suma total de los cuadrados

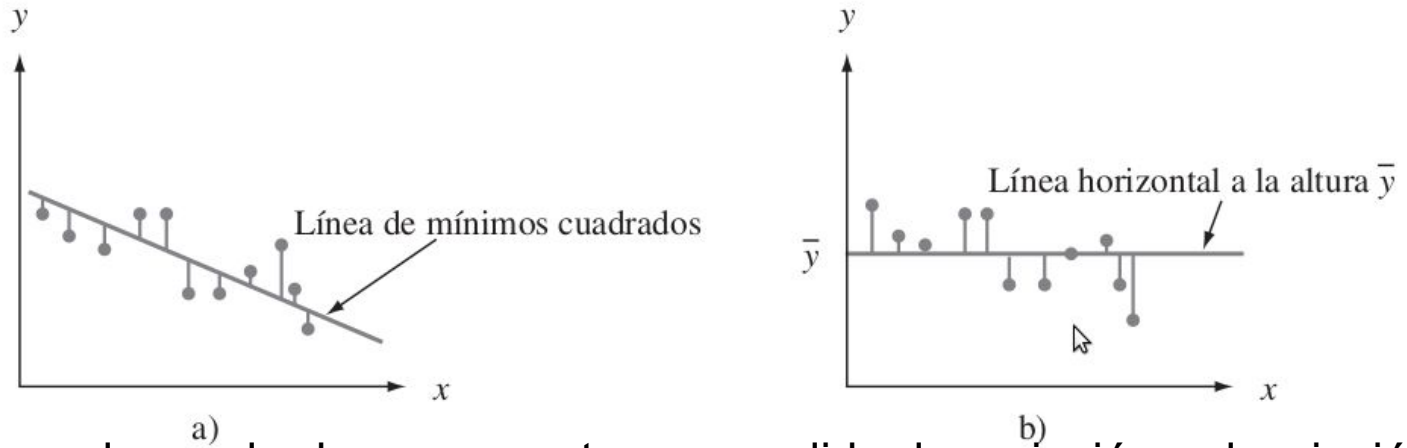


Que podemos decir de los RSS de cada uno de los gráficos?

- La suma de cuadrados SCE del error puede ser interpretada como una medida de cuánta variación de y permanece sin ser explicada por el modelo, es decir, cuánta no puede ser atribuida a una relación lineal.
- La suma total de los cuadrados da una medida cuantitativa de la cantidad de variación total en los valores y observados

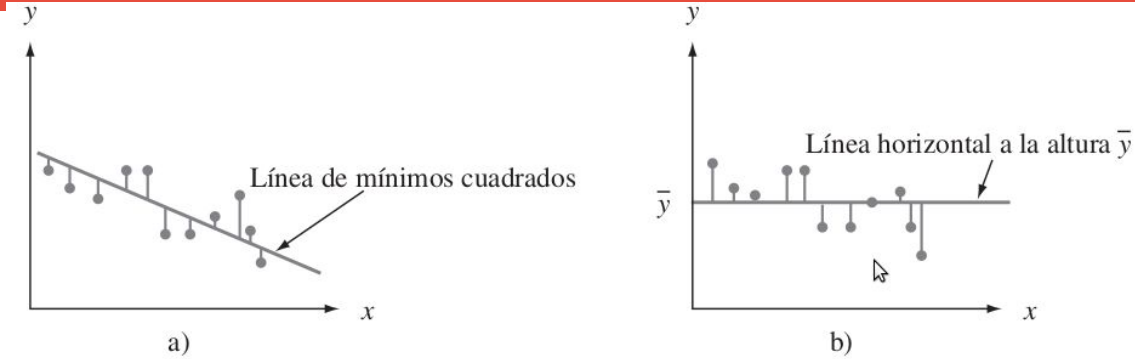
$$\text{STC} = S_{yy} = \sum(y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2/n$$

Suma total de los cuadrados



La suma de cuadrados representa una medida de variación o desviación con respecto a la media. Se calcula como una suma de los cuadrados de las diferencias con respecto a la media. El cálculo de la suma total de los cuadrados considera tanto la suma de los cuadrados de los factores como la de aleatoriedad o error.

Suma total de los cuadrados



La suma total de los cuadrados es la suma de las desviaciones al cuadrado con respecto a la media muestral de los valores y observados. Por consiguiente, se resta el mismo número y de cada y i presente en STC, mientras que SCE implica restar cada valor diferente pronosticado $\hat{y}_i$ de la y_i correspondiente observada. Así como SCE es la suma de desviaciones al cuadrado con respecto a la línea de cuadrados mínimos $y = \beta_0 + \beta_1 x$, STC es la suma de desviaciones al cuadrado con respecto a la línea horizontal a la altura y (en tal caso las desviaciones verticales son y_i - media(y))

Coeficiente de determinación

El coeficiente de determinación, denotado por r^2, está dado por

$$r^2 = 1 - \frac{\text{SCE}}{\text{STC}}$$

Se interpreta como la proporción de variación y observada que puede ser explicada por el modelo de regresión lineal simple (atribuida a una relación lineal aproximada entre y y x).

Coeficiente de determinación

- Mientras más alto es el valor de r^2, más exitoso es el modelo de regresión lineal simple
- Si r^2 es pequeño, un analista normalmente deseará buscar un modelo alternativo (como un modelo no lineal o un modelo de regresión múltiple que implique más de una sola variable independiente) que explique con más eficacia la variación de y.
- En definitiva buscaremos errores cuadráticos pequeños y valores de r^2 cercanos a 1