Diferencias en error de dos hipótesis (continuación)

- En algunos casos podemos estar interesados en la probabilidad de que alguna conjetura específica sea verdadera.

- Supongamos, por ejemplo, estamos interesados en verificar si "$error_{\mathrm{D}}(h_1) > error_{\mathrm{D}}(h_2)$"

- Supongamos que medimos los errores muestrales para h_1 y h_2 usando dos muestras independientes S_1 y S_2 de tamaño 100 y obtenemos que

$$error_{\mathrm{S}_1}(h_1) = 0.3 \text{ y } error_{\mathrm{S}_2}(h_2) = 0.2 \Rightarrow \hat{d} = 0.1.$$

- Para verificar lo anterior, testeamos

$$H_0: d \leq 0 \qquad \text{vs} \quad H_a: d > 0$$

- Luego aplicando el test asintótico,

se rechaza H_0 si $Z = {\hat{d}}/{\sigma_{\hat{d}}} > z_\alpha$.

- Por (#)

$$\sigma^2_{\hat{d}} = \frac{0.3 * 0.7}{100} + \frac{0.2 * 0.8}{100} = 0.037$$

- Entonces, si $\alpha = 0.05$

$$Z = \frac{0.1}{0.061} = 1.64 < z_{0.05} = 1.645$$

- Y por lo tanto no se rechaza H_0 a nivel asintótico 0.05

Comparando algoritmos de aprendizaje

- Frecuentemente estamos interesados en comparar la performance de dos algoritmos de aprendizaje L_A y L_B, en vez de dos hipótesis específicas.

- ¿Cuál es un test apropiado para comparar algoritmos de aprendizajes y cómo podemos determinar si la diferencia observada entre los algoritmos es estadísticamente significativa?

- Comencemos especificando el parámetro que se quiere estimar.

- Supongamos que queremos determinar si L_A o L_B es el mejor método de aprendizaje, en promedio, de una función objetivo f.

  - Una forma razonable de definir "promedio" es considerar la performance relativa de los dos algoritmos promediada sobre los conjuntos de entrenamiento de tamaño *n* que podrían ser obtenidos de la distribución subyacente de la instancia *V*.

  - En otras palabras, queremos estimar el valor esperado de la diferencia en sus errores

$$E_{S \subset D}\left[error_{\mathrm{D}}\left(\mathrm{L_A(S)}\right) - error_{\mathrm{D}}\left(\mathrm{L_B(S)}\right)\right] \quad (*)$$

donde:

- $L(S)$ denota la hipótesis de output del método aprendizaje *L* cuando se da la muestra *S* de entrenamiento
- el subíndice $S \subset D$ indica que el valor esperado se toma sobre las muestras *S* de acuerdo a la distribución subyacente de instancias *D*.

- Por supuesto, en la práctica sólo tenemos una muestra limitada D_0 de los datos cuando comparamos métodos de aprendizaje.

En estos casos, una aproximación obvia es estimar la cantidad anterior dividiendo a D_0 en el conjunto de entrenamiento S_0 y un subconjunto disjunto T_0.

- El conjunto de entrenamiento puede usarse para entrenar a L_A y L_B, y los datos de testeo pueden usarse para comparar la precisión de las dos hipótesis aprendida.

- Medimos la cantidad

$$error_{T_0}(\mathrm{L_A}(S_0)) - error_{T_0}(\mathrm{L_B}(S_0))$$

- Notemos que hay dos diferencias entre este estimador y la cantidad (*).

- Primero, estamos usando el $error_{T_0}(\mathrm{h})$ para aproximar el $error_{\mathrm{D}}(\mathrm{h})$

- Segundo, sólo estamos midiendo la diferencia en los errores de entrenamiento S_0 en vez de tomar el valor esperado de esta diferencia sobre todas las muestras *S* que pueden obtenerse en la distribución *D*.

- Una manera de mejorar el estimador anterior es repetir la partición de los datos D_0 en conjuntos disjuntos de entrenamiento y de testeo y tomar la media de los errores del conjunto de testeo para los distintos experimentos.

- Esto lleva al siguiente procedimiento para estimar la diferencia en los errores de los métodos de aprendizaje, basados en una muestra fija D_0 de los datos disponibles.

1. Particionar los datos disponibles D_0 en *k* conjuntos disjuntos $T_{1,\cdots,}T_k$ de igual tamaño, donde este tamaño es por lo menos 30

2. Para *i* desde *1* a *k*, hacer
   Usar T_i como conjunto de testeo, y los datos restantes como conjunto de entrenamiento S_i
   - $S_i \leftarrow \{D_0 - T_i\}$
   - $h_A \leftarrow \mathrm{L_A}(S_i)$
   - $h_B \leftarrow \mathrm{L_B}(S_i)$
   - $\delta_i \leftarrow \ error_{T_i}(h_A) - error_{T_i}(h_B)$

Devolver el valor $\bar{\delta}$, donde

$$\bar{\delta} = \frac{1}{k}\sum_{i=1}^{k} \delta_i$$

- La cantidad $\bar{\delta}$ puede usarse como estimador de la cantidad de la ecuación (*).
  Más precisamente, es un estimador de la cantidad

$$E_{S \subset D_0}\left[error_{\mathrm{D}}\left(\mathrm{L_A}(\mathrm{S})\right) - error_{\mathrm{D}}\left(\mathrm{L_B}(\mathrm{S})\right)\right]$$

  donde *S* representa una muestra aleatoria de tamaño $\left(k - 1/k\right)|D_0|$ tomada uniformemente de D_0.

- El intervalo de confianza asintótico de nivel α para la cantidad estimada en (*) está dada por

$$\left(\bar{\delta} - t_{\alpha,k-1} s_{\bar{\delta}}, \bar{\delta} + t_{\alpha,k-1} s_{\bar{\delta}}\right)$$

donde

$$s_{\bar{\delta}} = \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} \left(\delta_i - \bar{\delta}\right)^2}$$

- Notemos el procedimiento descripto aquí para comparar los métodos de aprendizaje involucra testear las dos hipótesis aprendidas en conjuntos de testeo idénticos.
- Esto difiere del método descripto anteriormente para comparar la hipótesis que había sido evaluado usando dos conjuntos de testeo independientes.

Clasificación paramétrica

- Supongamos que tenemos K clases mutuamente exclusivas $C_1, \cdots, C_K$ de modo que $C_i \cap C_j = \emptyset \quad \forall i \neq j$ y $\sum_{i=1}^{K} P(C_i) = 1$.

- Entonces, hemos visto que aplicando el teorema de Bayes, la probabilidad a posteriori en la clase C_i es:

$$P(C_i|\mathcal{X}) = \frac{P(\mathcal{X}|C_i)P(C_i)}{P(\mathcal{X})} = \frac{P(\mathcal{X}|C_i)P(C_i)}{\sum_{j=1}^{K} P(\mathcal{X}|C_j)P(C_j)}$$

- Y usando la función discriminante:

$$g_i(\mathcal{X}) = P(\mathcal{X}|C_i)P(C_i)$$

- O equivalentemente:

$$G_i(\mathcal{X}) = \ln g_i(\mathcal{X}) = \ln [P(\mathcal{X}|C_i)] + \ln[P(C_i)]$$

- Si suponemos que $P(\mathcal{X}|C_i)$ tiene distribución gaussiana, entonces:

$$G_i(x) = -\frac{1}{2} \ln (2\pi) - \ln (\sigma_i) - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \ln [P(C_i)] \quad (*)$$

Ejemplo

- Supongamos que una compañía de autos vende K modelos distintos, y supongamos que el único factor que influye en la decisión de compra de un cliente es su ganancia anual, el cual denotamos con x.

- Sea $P(C_i)$ es la proporción de clientes que compran el auto de la marca i.

- Si la ganancia anual de los clientes tiene una distribución gaussiana, entonces $P(x|C_i)$, la probabilidad de que un cliente que compró el auto de la marca i tenga un ingreso anual de x pesos, es $\mathcal{N}(\mu_i, \sigma_i^2)$, donde μ_i es la media del ingreso de tales clientes y σ_i^2 la varianza.

- Si no se conocen $P(C_i)$ y $P(x|C_i)$, se estiman a partir de la muestra y se reemplazan los estimadores para obtener la función discriminante.

- Consideremos la muestra: $\mathcal{X} = \{(x_1, r_1), \cdots, (x_n, r_n)\}$ donde $x_j \in \mathbb{R}$ y $r_j \in \{0,1\}^K$ de modo que:

$$r_j^i = \begin{cases} 1 & \text{si } x_j \in C_i \\ 0 & \text{si } x_j \in C_k \quad k \neq i \end{cases}$$

1. Para cada clase separada, los estimadores de máxima verosimilitud de las medias y las varianzas son:

$$\hat{\mu}_i = \frac{\sum_{j=1}^n x_j r_j^i}{\sum_{j=1}^n r_j^i}$$

$$\hat{\sigma}_i{}^2 = \frac{\sum_{j=1}^n \left(x_j - \hat{\mu}_i\right)^2 r_j^i}{\sum_{j=1}^n r_j^i}$$

2. Y los estimadores de las probabilidades a priori son:

$$\hat{P}(C_i) = \frac{\sum_{j=1}^{n} r_j^i}{n}$$

3. Reemplazando en la ecuación (*) resulta:

$$G_i(x) = -\frac{1}{2} \ln (2\pi) - \ln (\hat{\sigma}_i) - \frac{(x - \hat{\mu}_i)^2}{2\hat{\sigma}_i^2} + \ln \left[\hat{P}(C_i)\right]$$

4. El primer término es una constante y puede ser eliminado porque es común para todas las G_i.

5. Si las probabilidades a priori son iguales, también se puede eliminar el último término.

6. Si además podemos asumir que las varianzas son iguales, podemos escribir:

$$G_i(x) = -(x - \hat{\mu}_i)^2$$

- Esta manera, podríamos asignar a x a la clase a la clase con la media más cercana:

Elegimos C_i si $|x - \hat{\mu}_i| = \min_k |x - \hat{\mu}_k|$

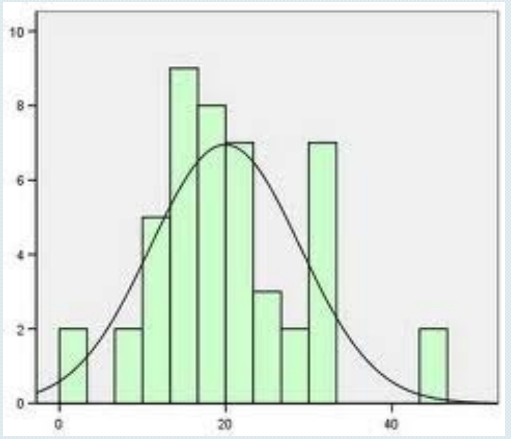
- Para dos clases adyacentes, el punto medio entre las dos medias es el umbral de decisión, ya que

$$G_l(x) = G_h(x) \Rightarrow (x - \hat{\mu}_l)^2 = (x - \hat{\mu}_h)^2$$

$$\Rightarrow x = \frac{\hat{\mu}_l + \hat{\mu}_h}{2}$$

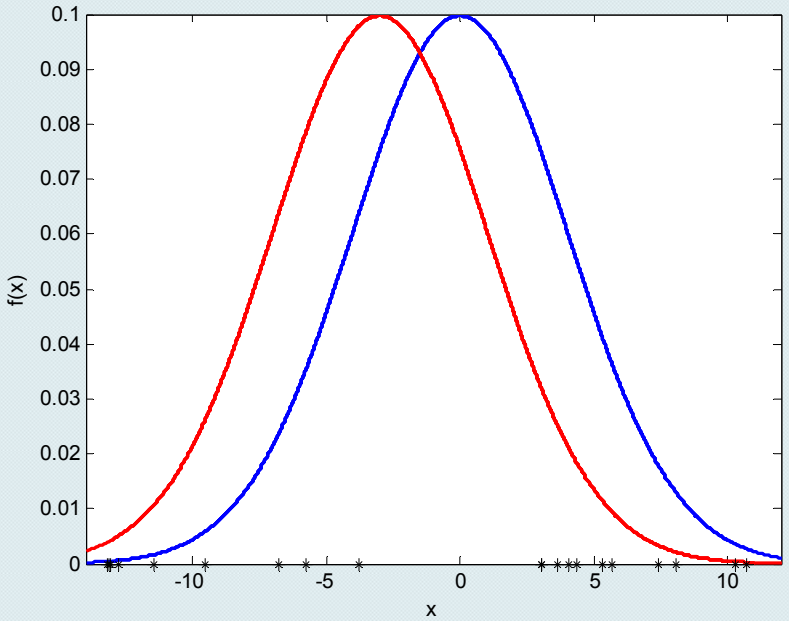
- Cuando $\mathcal{X}$ es continua, no necesariamente las $P(x|C_i)$ tienen una distribución gaussiana.
  - En el caso de que esto no ocurra, el algoritmo de clasificación que hemos dado es erróneo.

- Hay varias formas de verificar la normalidad
  - Por ej. en el caso de datos unidimensionales, se puede obtener el histograma y observar si:
    - tiene forma de campana
    - es unimodal
    - tiene simetría respecto al centro de los datos.



Algoritmo EM

Estimando la media de *k* distribuciones gaussianas



Instancias generadas por la mezcla de 2 distribuciones normales con la misma varianza
Las instancias son los puntos a lo largo del eje *x*

- Consideremos un problema en el cual los datos *D* es un conjunto de instancias generadas por una distribución de probabilidades que es una mezcla de *k* distribuciones normales distintas.

- Cada instancia es generada usando un proceso de dos pasos:

- Primero, se selecciona al azar una de las *k* distribuciones normales.

- Segundo, se genera al azar una única instancia x_i de acuerdo a la distribución elegida.

  Este proceso se repite para generar un conjunto de puntos como los mostrados en la figura.

- Para simplificar, consideremos el caso especial donde cada una de las *k* distribuciones normales tiene la misma varianza σ^2 conocida.

- La tarea de aprendizaje es dar como output una hipótesis $h = (\mu_1, \cdots, \mu_k)$ que describa la media de cada una de las *k* distribuciones.

- Nos gustaría encontrar la hipótesis de máxima verosimilitud para estas medias; esto es, una hipótesis *h* que maximice $P(D|h)$.

- Para calcular la hipótesis de máxima verosimilitud para la media de una única distribución normal dados como datos las instancias observadas $x_1, \cdots, x_n$, vimos que se obtiene como:

$$\mu_{ML} = \arg \min_{\mu} \sum_{i=1}^{n} (x_i - \mu)^2 \quad (*)$$

$$= \bar{x}$$

- Nuestro problema ahora, sin embargo, involucra una mezcla de *k* distribuciones normales distintas, y no podemos observar cuáles instancias fueron generadas por cuál distribución.

- En el ejemplo de la figura, podemos pensar que una visión completa de cada instancia es la 3-upla (x_i, z_{i1}, z_{i2}), donde x_i es el valor observado de la i-ésima instancia, z_{i1} y z_{i2} indica cuál de las dos distribuciones normales fueron usadas para generar el valor de x_i.

  - En particular,

$$z_{ij} = \begin{cases} 1 & \text{si } x_i \text{ fue generada por la } j\text{-ésima distribución normal} \\ 0 & \text{en otro caso} \end{cases}$$

- Si los valores de z_{i1} y z_{i2} fuesen observados, se puede usar la ecuación (*) para encontrar las medias, pero como esto no ocurre, usaremos el algoritmo EM.

- Aplicado al problema de la *k*-medias, el algoritmo busca la hipótesis de máxima verosimilitud re-estimando repetidamente los valores esperados de las variables z_{ij} dada su hipótesis actual $(\mu_1, \cdots, \mu_k)$, y luego recalcula la hipótesis de máxima verosimilitud usando estos valores esperados

- Para el problema particular de la figura, el algoritmo EM primero inicializa la hipótesis $h = (\mu_1, \mu_2)$, donde μ_1 y μ_2 son valores iniciales arbitrarios.

- Luego interativamente re-estima *h* repitiendo los siguientes dos pasos hasta que el proceso converja a un valor de *h*

Paso 1:
Calcular el valor esperado $E(z_{ij})$ para cada variable z_{ij}, suponiendo que la hipótesis actual $h = (\mu_1, \mu_2)$ ocurre

Paso 2:

Calcular una nueva hipótesis de máxima verosimilitud $h' = (\mu'_1, \mu'_2)$, asumiendo que el valor de cada variable z_{ij} es su valor esperado $E(z_{ij})$ calculado en el paso 1.

Luego, reemplazar la hipótesis $h = (\mu_1, \mu_2)$ por $h' = (\mu'_1, \mu'_2)$ e iterar

- Veamos cómo se implementan estos dos pasos en la práctica.

- En el paso 1, $E(z_{ij})$ es la probabilidad de que la instancia x_i fuera generada por la *j*-ésima distribución normal.

$$E(z_{ij}) = \frac{P(x = x_i | \mu = \mu_j)}{\sum_{k=1}^{2} P(x = x_i | \mu = \mu_k)}$$

$$= \frac{e^{-\frac{1}{2\sigma^2}(x_i - \mu_j)^2}}{\sum_{k=1}^{2} e^{-\frac{1}{2\sigma^2}(x_i - \mu_k)^2}}$$

Así el primer paso se implementa sustituyendo los valores actuales (μ_1, μ_2) y el observado x_i en la expresión anterior.

- En el paso 2 usamos la $E(z_{ij})$ calculada en el paso 1 para obtener una nueva hipótesis de máxima verosimilitud $h' = (\mu'_1, \mu'_2)$, que está dado por:

$$\mu_j \leftarrow \frac{\sum_{i=1}^{n} E(z_{ij}) x_i}{\sum_{i=1}^{n} E(z_{ij})}$$

- Notemos que esta expresión es similar a la media muestral (*); es una media muestral pesada para μ_j, con cada instancia pesada por su esperanza $E(z_{ij})$ que fue generada por la *j*-ésima distribución normal.

- El algoritmo anterior ilustra la esencia del algoritmo EM: la hipótesis actual se usa para estimar las variables no observadas, y los valores esperados de las variables se usan entonces para calcular una hipótesis mejorada.

- Se puede probar en cada interación dentro del ciclo, el algoritmo EM incrementa la verosimilitud $P(D|h)$ salvo que sea máximo local.
  El algoritmo entonces converge a una hipótesis local de máxima verosimilitud para (μ_1, μ_2).

Generalización del algoritmo EM

- El algoritmo EM puede aplicarse en muchas situaciones donde se quiere estimar un conjunto de parámetros θ que escribe una distribución subyacente, dada solamente una parte observada del conjunto de datos producidos por esta distribución.

- En general,
  $\mathcal{X} = \{x_1, \cdots, x_n\}$: datos observados en un conjunto de n instancias independientes
  $\mathcal{Z} = \{z_1, \cdots, z_n\}$: datos no observados de estas mismas instancias
  $\mathcal{Y} = \mathcal{X} \cup \mathcal{Z}$: datos completos.

- Notemos que la no observada $\mathcal{Z}$ puede ser tratada una variable aleatoria cuya distribución depende de parámetros desconocidos θ y los datos observados $\mathcal{X}$.

- Análogamente, $\mathcal{Y}$ es una variable aleatoria porque está definida en términos de la variable aleatoria $\mathcal{Z}$.

- Usamos *h* para denotar los valores actuales de los parámetros θ de la hipótesis, y *h'* para denotar la hipótesis revisada que es estimada en cada iteración del algoritmo EM.

- El algoritmo EM busca la hipótesis de máxima verosimilitud *h'* buscando la *h'* que maximiza $E[\ln P(\mathcal{Y}|h')]$.

- Este valor esperado se toma sobre la distribución de $\mathcal{Y}$, la cual se determina por los parámetros desconocidos θ.
  Esta distribución se determinaría al conocer los valores de $\mathcal{X}$ y la distribución de $\mathcal{Z}$.

- En general no conocemos esta distribución porque está determinada por los parámetros θ que estamos tratando de estimar.

- Por lo tanto, el algoritmo EM usa la hipótesis actual *h* en lugar de los valores reales de θ para determinar dicha distribución.

- Definamos la función $Q(h'|h)$ que da a $E[\ln P(\mathcal{Y}|h')]$ como función de *h'*, bajo la suposición de que $\theta = h$ y dada la parte observada $\mathcal{X}$ de los datos completos $\mathcal{Y}$.

$$Q(h'|h) = E[\ln P(\mathcal{Y}|h')|h, \mathcal{X}]$$

Algoritmo EM

Se repiten los siguientes dos pasos hasta que converja:

Paso 1: paso (E) de estimación

Calcular $Q(h'|h)$ usando la hipótesis actual h y los datos observados X para estimar la distribución sobre Y

$$Q(h'|h) \leftarrow E[\ln P(Y|h')|h, X]$$

Paso 2: paso (M) de maximización

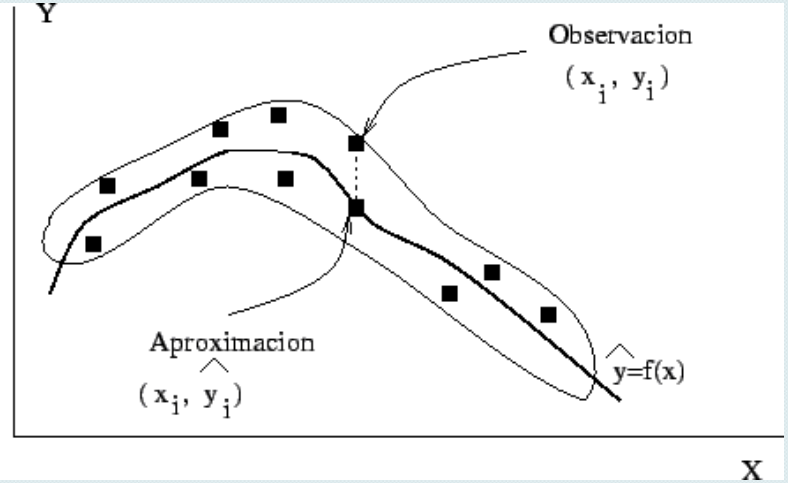
Reemplazar la hipótesis *h* por la hipótesis *h'* que maximiza Q

$$h \leftarrow \arg\max_{h'} Q\left(h' \middle| h\right)$$

- Cuando la función Q es continua, el algoritmo EM converge a la función de verosimilitud $P(\mathcal{Y}|h')$.

- Cuando la función de verosimilitud tiene un único máximo, EM convergerá al estimador que es el máximo global de la verosimilitud h'.
  En otro caso, sólo se garantiza el convergerá a un máximo global.

Mínimos cuadrados



- Supongamos que hemos medido un conjunto de pares de datos (x_i, y_i) en un experimento, por ejemplo, la posición de un móvil en ciertos instantes de tiempo.

- Queremos obtener una función $y = f(x)$ que se ajuste lo mejor posible a los valores experimentales.

- Se pueden ensayar muchas funciones, por ejemplo, rectas, polinomios, funciones potenciales o logarítmicas.

- Una vez establecido la función a ajustar se determinan sus parámetros, en el caso de un polinomio, serán los coeficientes del polinomio de modo que los datos experimentales se desvíen lo menos posible de la fórmula empírica.

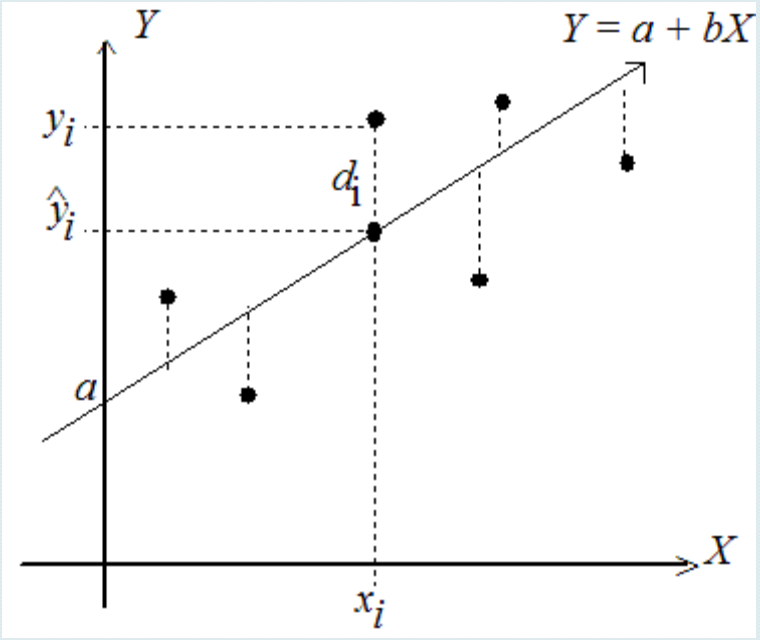
- Es decir, dado un conjunto de *n* datos $(x_1, y_1), \cdots, (x_n, y_n)$ se quiere aproximar la función f con el polinomio de grado m, $P(x) = a_0 + a_1 x + \cdots + a_m x^m$.
  Si

$$S = \sum_{i=1}^{n} [y_i - P(x_i)]^2$$

  los coeficientes $a_0, \cdots, a_m$ se obtienen de modo de minimizar *S*

Caso particular: recta de mínimos cuadrados

- Se deben determinar los coeficientes a y b de la ecuación de la recta $Y = a + bx$ que mejor "ajusten" a los n pares (x_i, y_i) observados.



- Es decir se van a buscar los valores de a y b que minimicen la distancia entre lo observado y_i y lo predicho $\hat{y}_i$.

- Las diferencias entre los valores observados y_i y los valores que predice el modelo $f(x_i) = \hat{y}_i$, se denominan residuos

$$r_i = y_i - \hat{y}_i = y_i - \hat{a} - \hat{b}x_i$$

- Para encontrar $\hat{a}$ y $\hat{b}$

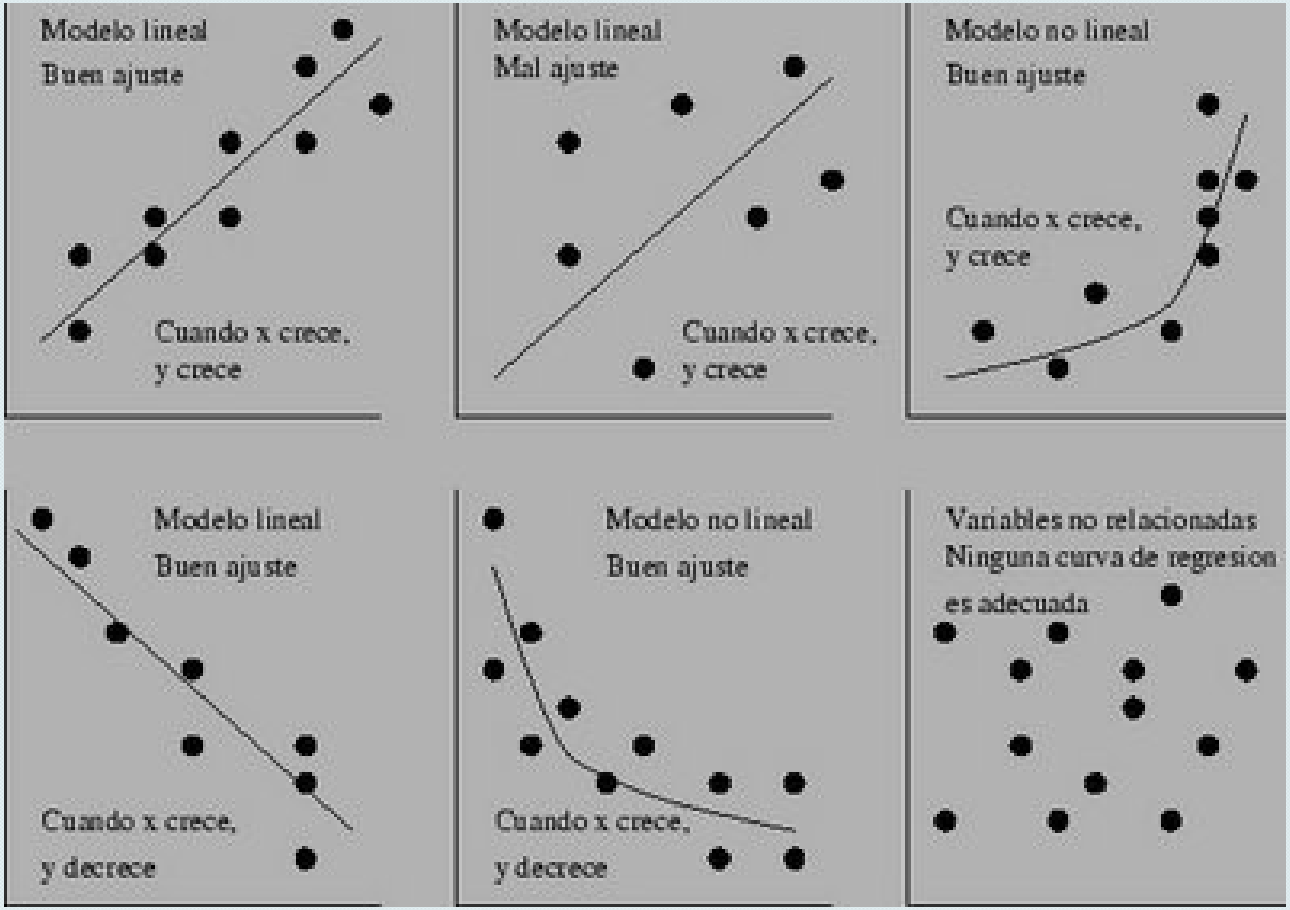
$$\left(\hat{a}, \hat{b}\right) = \arg \min_{(a,b)} \sum_{i=1}^{n} r_i^2$$

Y resolviendo las ecuaciones normales se obtiene que:

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

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

- ¿Es correcto ajustar una recta a los datos observados? Veamos algunos casos:



Una medida del ajuste para la regresión lineal

- Si el modelo ajusta “bien” los datos, los residuos son pequeños.
  Entonces se podría medir la bondad del ajuste con la suma $\sum_{i=1}^{n} r_i^2$.

- Sin embargo esta cantidad depende de las unidades en las que fueron medidos los y_i.

- Entonces, si $a \neq 0$, una medida de la bondad del ajuste es:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} r_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

- Se puede ver que $0 \leq R^2 \leq 1$, y que valores cercanos a 1 indican un buen ajuste.

- Si $a = 0$,

$$R^2 = 1 - \frac{\sum_{i=1}^{n} r_i^2}{\sum_{i=1}^{n} y_i^2}$$

Regresión no lineal

Para ciertas familias de funciones, se pueden hacer transformaciones sobre las variables de manera tal que se obtiene un modelo lineal.

En la siguiente tabla se presentan algunos casos frecuentes:

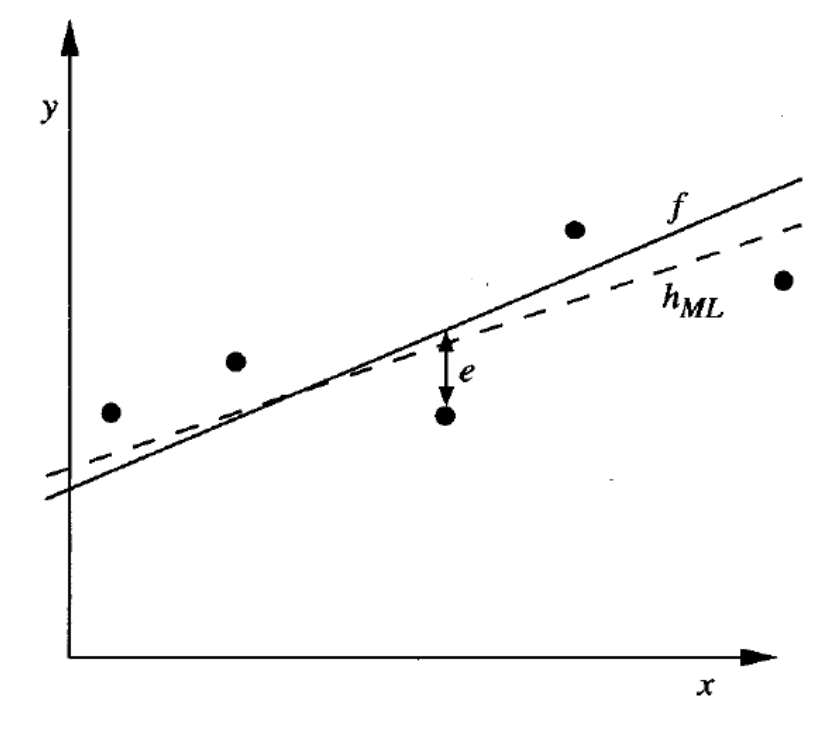
Familia	**Funciones**	**Transformaciones**	**Forma afín**
exponencial	$y = ae^{bx}$	$\tilde{y} = \ln(y)$	$\tilde{y} = \ln(a) + bx$
potencia	$y = ax^b$	$\tilde{y} = \ln(y) \quad \tilde{x} = \ln(x)$	$\tilde{y} = \ln(a) + b\tilde{x}$
inversa	$y = a + \frac{b}{x}$	$\tilde{x} = \frac{1}{x}$	$\tilde{y} = a + b\tilde{x}$
logística	$y = \frac{1}{1 + e^{-(ax+b)}}$	$\tilde{y} = \ln\left(\frac{y}{1-y}\right)$	$\tilde{y} = ax + b$

Máxima verosimilitud e hipótesis de error de mínimos cuadrados

- Vamos a considerar la tarea de aprender una función objetivo de valores continuos.

- Consideremos el siguiente problema:
  El aprendiz *L* considera un espacio de instancias *X* consistente en alguna clase de funciones reales definidas sobre *X* (i.e. cada $h \in H$ es una función de la forma $h: X \to \mathbb{R}$) y quiere aprender una función conocida $f: X \to \mathbb{R}$ en *H*.

- Se lo provee de un conjunto de *m* ejemplos de entrenamiento, donde el valor objetivo cada ejemplo está perturbado con un ruido aleatorio con distribución normal.
    - Cada ejemplo de entrenamiento es un par la forma $\langle x_i, d_i \rangle$ donde $d_i = f(x_i) + e_i$.
    - Se supone que los valores e_i son independientes con distribución normal con media cero.
- La tarea el aprendiz es dar como output una hipótesis de máxima verosimilitud, o equivalentemente, una hipótesis MAP suponiendo que todas las hipótesis son equiprobables a priori.



- Un ejemplo de tal problema es aprender una función lineal

- La figura ilustra una función objetivo lineal f graficada con una línea sólida, y con un de ejemplos de entrenamiento con ruido de esta función objetivo.

- La línea punteada corresponde a la hipótesis h_{ML} con error de mínimos cuadrados de entrenamiento, por lo tanto la hipótesis de máxima verosimilitud.

- Notemos que la hipótesis de máxima verosimilitud no es necesariamente idéntica a la hipótesis correcta, f, porque es inferida a partir de sólo una muestra limitada de ejemplos de entrenamiento con ruido.

Ejemplo

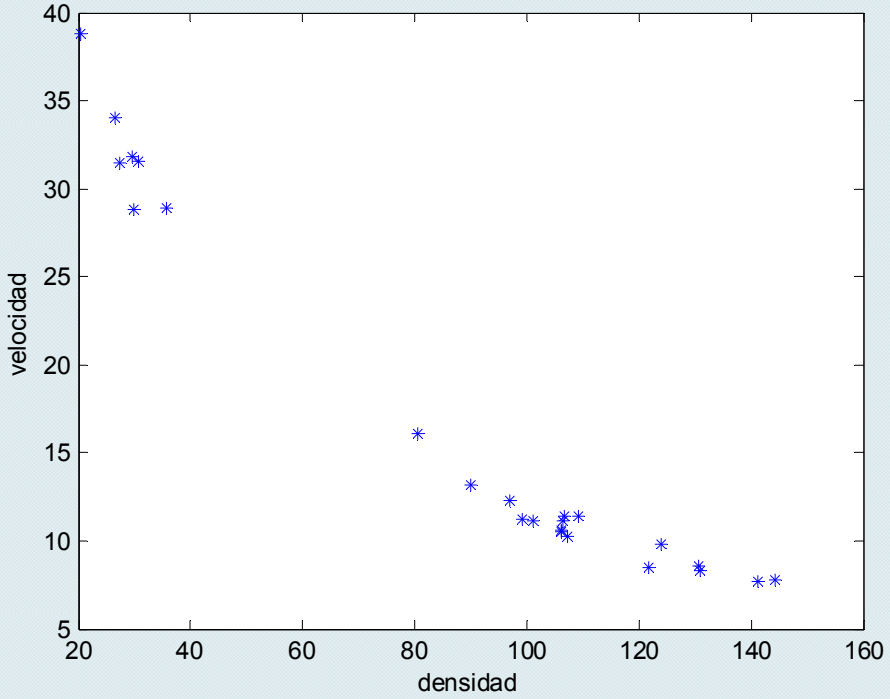
- Se sabe que a mayor cantidad de vehículos en las calles, más lenta se hace la velocidad del tráfico.
  Una compresión precisa de este problema permitiría planificar los sistemas de transporte.

- Los siguientes datos corresponden a la densidad de vehículos por kilómetro y la correspondiente velocidad en kilómetros por hora

densidad	velocidad	densidad	velocidad
20.4	38.8	29.5	31.8
27.4	31.5	30.8	31.6
106.2	10.6	26.5	34.0
80.4	16.1	35.7	28.9
141.3	7.7	30.0	28.8
130.9	8.3	106.2	10.5
121.7	8.5	97.0	12.3
106.5	11.1	90.1	13.2
130.5	8.6	106.7	11.4
101.1	11.1	99.3	11.2
123.9	9.8	107.2	10.3
144.2	7.8	109.1	11.4

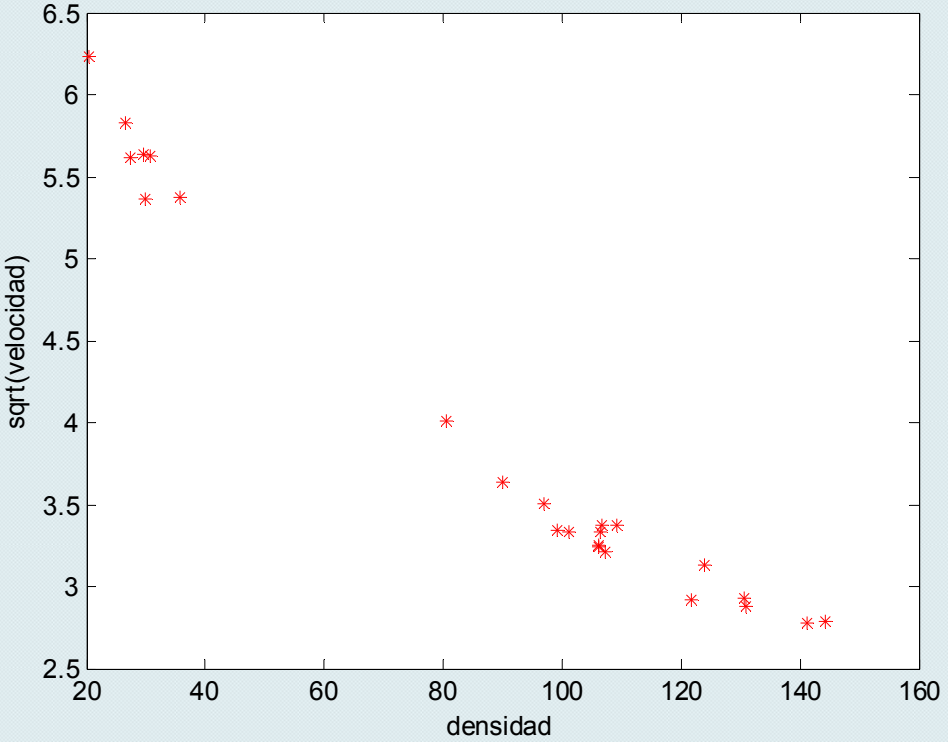
- Como la congestión afecta la velocidad estamos interesados en determinar el efecto de de la densidad en la velocidad.

- Grafiquemos *densidad* vs *velocidad*



- Se observa cierta “curvatura” en los datos, por lo cual no sería apropiado ajustar una recta

- Grafiquemos *densidad* vs $\sqrt{velocidad}$ (*)

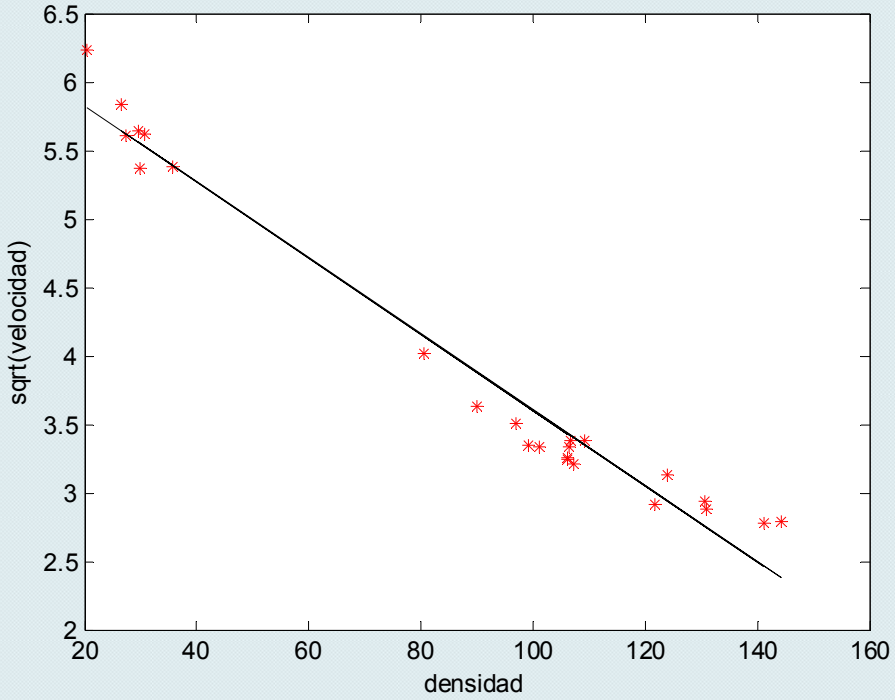


- Al ajustar la recta de mínimos cuadrados

$\sqrt{\text{velocidad}_i} = a + b * \text{densidad}_i + \varepsilon_i$ con $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ i.i.d.

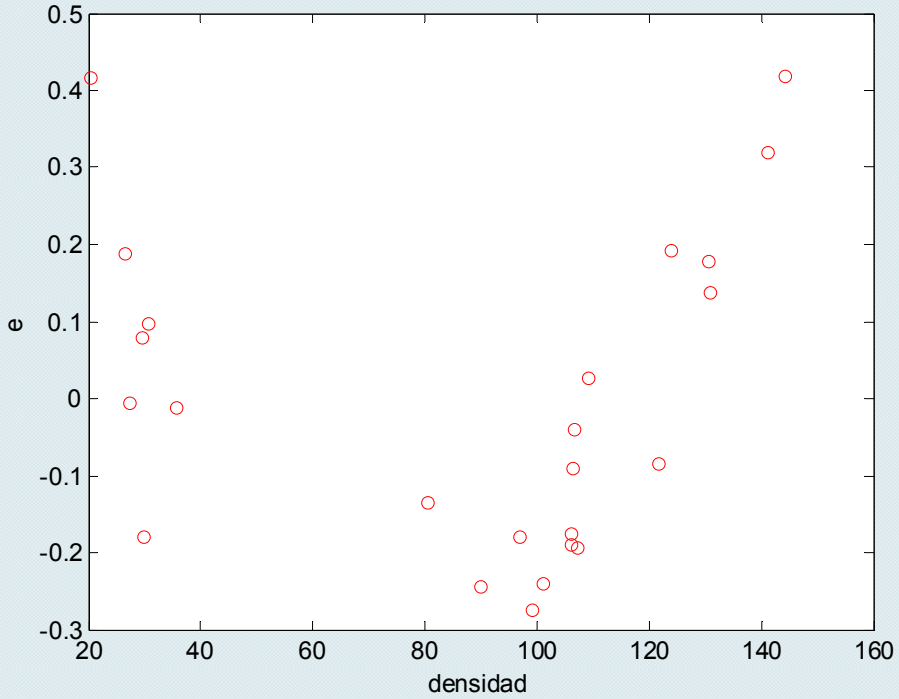
se obtiene:

$$\hat{a} = 6.3797 \qquad \hat{b} = -0.0278$$



$$R^2 = 0.9687$$

- Si hacemos un plot de *densidad* vs *residuos*



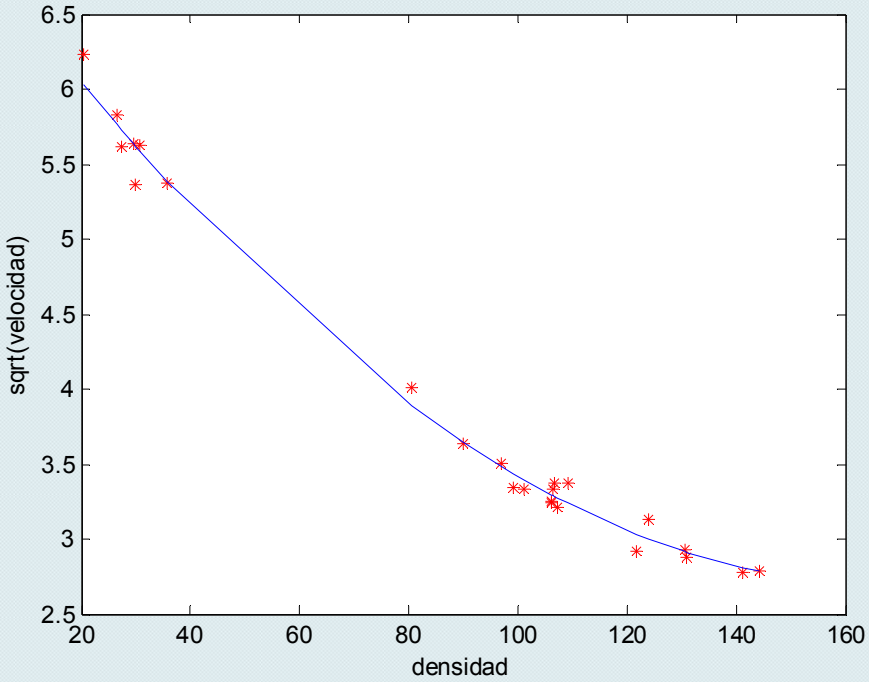
se observa cierto patrón en forma de “U”, lo cual indicaría que el ajuste anterior **no** es adecuado

- Como en el gráfico (*) se observa cierta curvatura, ajustemos

$$\sqrt{\text{velocidad}_i} = a + b * \text{densidad}_\text{i} + \text{c} * \ \text{densidad}_i{}^2 + \ \varepsilon_\text{i}$$

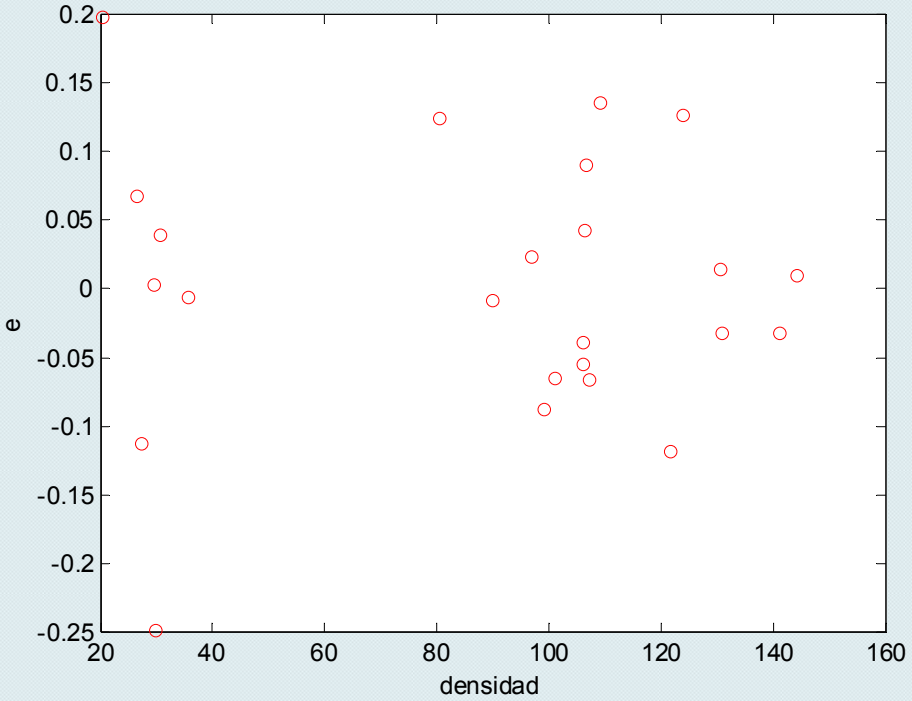
Se obtiene:

$$\hat{a} = 7.0026 \quad \hat{b} = -0.0507 \quad \hat{c} = 0.0001$$



$$R^2 = 0.9931$$

- Si hacemos un plot de *densidad* vs *residuos* para el ajuste de la función cuadrática, se obtiene



Se observa que los residuos tienen valores más pequeños y no presentan ningún patrón.

Métodos multivariados

- En muchas aplicaciones se efectúan varias mediciones de cada individuo o evento que genera un vector de observaciones.

- La muestra puede ser vista como una matriz de datos

$$\boldsymbol{X} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix}$$

donde las *p* columnas corresponden a las *p* variables que denotan el resultado de las mediciones hechas en un individuo o evento.

- Las *n* filas corresponden a las observaciones independientes e idénticamente distribuidas, ejemplos, o instancias de *n* individuos o eventos.

- Uno de los objetivos es resumir esta gran cantidad de datos mediante poco parámetros.

- También podemos estar interesados en explorar, generando hipótesis sobre los datos.

- En algunas aplicaciones, nos podría interesar predecir el valor de una variable a partir de los valores de las otras variables.

- Si las predicciones son:
  - discretas es una clasificación multivariada
  - numéricas es un problema de regresión multivariada

Estimación de los parámetros

- El vector de medias $\boldsymbol{\mu}$ se define de modo que cada uno de sus elementos es la media de una columna de $\boldsymbol{X}$.

$$E(\boldsymbol{X}) = \boldsymbol{\mu} = \left[\mu_1, \cdots, \mu_p\right]^t$$

- La varianza de X_i se denota con σ_i^2 y la covarianza de dos variables X_i y X_j se define como:

$$\sigma_{ij} = \text{cov}\left(X_i, X_j\right) = E\left[\left(X_i - \mu_i\right)\left(X_j - \mu_j\right)\right] = E\left(X_i X_j\right) - \mu_i \mu_j$$

con $\sigma_{ij} = \sigma_{ji}$ y cuando $i = j$, $\sigma_{ii} = \sigma_i^2$

- Para las *p* variables, hay *p* varianzas y $p(p-1)\,/\,2$ covarianzas que se representan en una matriz de $p \times p$, llamada matriz de covarianza y está dada por:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1p} \\ \vdots & \ddots & \vdots \\ \sigma_{p1} & \cdots & \sigma_p^2 \end{bmatrix}$$

$$\Sigma \equiv \ \mathrm{cov}(\boldsymbol{X}) = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^t] = E(\boldsymbol{X}\boldsymbol{X}^t) - \boldsymbol{\mu}\boldsymbol{\mu}^t$$

- Si dos variables están relacionadas de una forma lineal, entonces la covarianza será positiva o negativa dependiendo de si la relación tiene una pendiente positiva o negativa.

- Pero el tamaño de la relación es difícil de interpretar porque depende de las unidades en las que se miden las variables.

- La correlación entre las variables X_i y X_j está normalizada entre -1 y 1 y se define como:

$$\text{corr}(X_i, X_j) = \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

- Si dos variables son independientes, entonces su covarianza, y por lo tanto su correlación es nula.

- Sin embargo, la recíproca **no** es cierta: las variables pueden ser dependientes y su correlación puede ser 0.

- Dada una muestra multivariada, se pueden calcular estimadores de estos parámetros.

- Los estimadores de máxima verosimilitud del vector de medias y de la matriz de covarianza están dados por:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{\boldsymbol{X}}$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{\boldsymbol{X}})(X_i - \bar{\boldsymbol{X}})^t$$

- La matriz de correlación muestral ***R*** tiene como elementos

$$r_{ij} = \frac{\hat{\sigma}_{ij}}{\hat{\sigma}_i \hat{\sigma}_j}$$

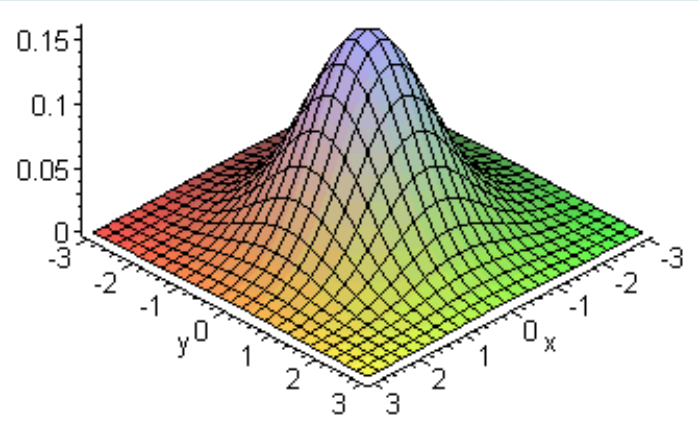
Estimación de valores faltantes

- Frecuentemente pueden faltar algunas observaciones de los valores de ciertas variables.

- La mejor estrategia es descartar estas observaciones todas juntas, pero generalmente la muestra no es muy grande y no se quiere perder la información que contiene el resto de las observaciones existentes.

- Para sustituir los valores faltantes, se hace lo siguiente:
  - Si la variable es numérica, se sustituye por el promedio de los datos disponibles en la muestra para esa variable.
  - Si la variable es discreta, se sustituye por el valor más frecuente que aparece en los datos.

Distribución normal multivariada



- En el caso multivariado, **X** es un vector de dimensión p con distribución normal $\mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$ si su función de densidad está dada por:

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^t \Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\right]$$

- En el caso de $p = 2$, como

$$\boldsymbol{\mu} = (\mu_1, \mu_2)^t \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

Entonces

$$f(X_1, X_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}(Z_1^2 - 2\rho Z_1 Z_2 + Z_2^2)\right]$$

donde

$$Z_i = \frac{X_i - \mu_i}{\sigma_i}$$

son las variables normalizadas.

- En este caso la distribución depende de 5 parámetros: μ_1, μ_2, σ_1, σ_2 y ρ.

Algunas propiedades:

- Si $\boldsymbol{X} = (X_1, \cdots, X_p) \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$, entonces $X_i \sim \mathcal{N}_1(\mu_i, \sigma_i^2)$ para $1 \leq i \leq p$.
- La recíproca **no** es cierta.

- Si las componentes de $\boldsymbol{X}$ son independientes, entonces $\text{cov}(X_i, X_j) = 0$ para $i \neq j$ y por lo tanto la matriz de covarianza Σ es diagonal.

  Además, la densidad conjunta de $\boldsymbol{X}$ es el producto de las densidades de las X_i's, es decir:

$$f(\mathbf{X}) = \prod_{i=1}^{p} f(X_i) = \frac{1}{(2\pi)^{p/2}\left(\prod_{i=1}^{p} \sigma_i\right)} \exp\left[-\frac{1}{2}\sum_{i=1}^{p}\left(\frac{X_i - \mu_i}{\sigma_i}\right)^2\right]$$

- Sea $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$ y sea $\boldsymbol{w} \in \mathbb{R}^p$, entonces:

$$\boldsymbol{w}^t \boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{w}^t \boldsymbol{\mu}, \boldsymbol{w}^t \Sigma \boldsymbol{w})$$

- Si $W \in \mathbb{R}^{p \times k}$ tal que $\text{rg}(W) = k < p$, entonces:

$$W^t \boldsymbol{X} \sim \mathcal{N}_k(W^t \boldsymbol{\mu}, W^t \Sigma \text{W})$$

Clasificación multivariada

- Cuando $\boldsymbol{x} \in \mathbb{R}^p$, si las densidades condicionales a las clases $f(\boldsymbol{x}|C_i)$ tiene densidad normal $\mathcal{N}_p(\boldsymbol{\mu_i}, \Sigma_\mathrm{i})$, tenemos:

$$f(\boldsymbol{x}|C_i) = \frac{1}{(2\pi)^{p/2}|\Sigma_\mathrm{i}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_i})^t \Sigma_\mathrm{i}{}^{-1}(\boldsymbol{x}-\boldsymbol{\mu_i})\right]$$

- Supongamos que queremos predecir el tipo de auto que va a comprar un cliente.

- Los distintos tipos de autos son las clases y $\boldsymbol{x}$ son los datos observados de los clientes, por ejemplo, la edad y el ingreso anual.

- $\boldsymbol{\mu_i}$: vector de medias de las edades y los ingresos de los clientes que compran el auto i

  Σ_i: matriz de covarianza tal que:

  σ_{i1}^2 y σ_{i2}^2: varianzas de las edades y los ingresos

  σ_{i12}: covarianza entre la edad y el ingreso

  del grupo de los clientes que compran el auto i.

- Cuando definimos la función discriminante como

$$G_i(\boldsymbol{x}) = \ln [f(\boldsymbol{x}|C_i)] + \ln [P(C_i)]$$

y suponemos que $\boldsymbol{x}|C_i \sim \mathcal{N}_p(\boldsymbol{\mu_i}, \Sigma_i)$, tenemos

$$G_i(\boldsymbol{x}) = -\frac{p}{2} \ln (2\pi) - \frac{1}{2} \ln|\Sigma_i| - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu_i})^t \Sigma_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu_i}) + \ln [P(C_i)]$$

- Dada una muestra de entrenamiento para $K \geq 2$ clases, $\mathcal{X} = \{\boldsymbol{x}^j, \boldsymbol{r}^j\}$, donde

$$r_i^j = \begin{cases} 1 & \text{si } \boldsymbol{x^j} \in C_i \\ 0 & \text{sino} \end{cases}$$

los estimadores de máxima verosimilitud separados por cada clase son:

$$\hat{P}(C_i) = \frac{1}{n} \sum_{j=1}^{n} r_i^j$$

$$\boldsymbol{\hat{\mu}_i} = \frac{\sum_{j=1}^{n} r_i^j \, \boldsymbol{x^j}}{\sum_{j=1}^{n} r_i^j}$$

$$\widehat{\Sigma}_{\mathrm{i}} = \frac{\sum_{j=1}^{n} r_i^{\,j} \left(\boldsymbol{x^j} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{i}}\right)\left(\boldsymbol{x^j} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{i}}\right)^t}{\sum_{j=1}^{n} r_i^{\,j}}$$

- Reemplazando en la función discriminante e ignorando el término constante, resulta:

$$\begin{aligned} G_i(\boldsymbol{x}) &= -\frac{1}{2}\ln\left|\widehat{\Sigma}_{\mathrm{i}}\right| - \frac{1}{2}(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{i}})^t \widehat{\Sigma}_{\mathrm{i}}^{\,-1}(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{i}}) + \ln\left[\hat{P}(C_i)\right] \\ &= -\frac{1}{2}\ln\left|\widehat{\Sigma}_{\mathrm{i}}\right| - \frac{1}{2}\left(\boldsymbol{x}^t \widehat{\Sigma}_{\mathrm{i}}^{\,-1}\boldsymbol{x} - 2\boldsymbol{x}^t \widehat{\Sigma}_{\mathrm{i}}^{\,-1}\widehat{\boldsymbol{\mu}}_{\boldsymbol{i}} + \widehat{\boldsymbol{\mu}}_{\boldsymbol{i}}^{\,t}\widehat{\Sigma}_{\mathrm{i}}^{\,-1}\widehat{\boldsymbol{\mu}}_{\boldsymbol{i}}\right) + \ln\left[\hat{P}(C_i)\right] \end{aligned}$$

la cual define una función *discriminante cuadrática* que también puede escribirse como:

$$G_i(\boldsymbol{x}) = \boldsymbol{x}^t \boldsymbol{W}_i \boldsymbol{x} + \boldsymbol{w}_i^t \boldsymbol{x} + w_{i0}$$

donde:

$$\boldsymbol{W}_i = -\frac{1}{2} \hat{\Sigma}_i^{-1}$$

$$\boldsymbol{w}_i = \hat{\Sigma}_i^{-1} \hat{\boldsymbol{\mu}}_i$$

$$w_{i0} = -\frac{1}{2} \hat{\boldsymbol{\mu}}_i{}^t \hat{\Sigma}_i^{-1} \hat{\boldsymbol{\mu}}_i - \frac{1}{2} \ln\left|\hat{\Sigma}_i\right| + \ln\left[\hat{P}(C_i)\right]$$

- El número de parámetros para estimar es:
  - o Kp para las medias
  - o $\frac{Kp(p+1)}{2}$ para las covarianzas

- Cuando *p* es grande y las muestras son pequeñas, $\widehat{\Sigma}_i$ puede ser singular y no existir la inversa.
  $|\widehat{\Sigma}_i|$ puede ser no nulo pero muy pequeño, en cuyo caso será inestable; pequeños cambios en $\widehat{\Sigma}_i$ pueden producir grandes cambios en $\widehat{\Sigma}_i^{-1}$.

- Para solucionar este problema, una posibilidad es juntar los datos de todas las clases y estimar una matriz de covarianza común

$$\widehat{\Sigma} = \sum_{i=1}^{K} \hat{P}(C_i)\, \widehat{\Sigma}_i$$

- En este caso de matrices de covarianzas iguales, la función discriminante se reduce a:

$$G_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{i}})^t \widehat{\Sigma}^{-1}(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{i}}) + \ln\left[\hat{P}(C_i)\right]$$

- El número de parámetros para estimar es:
  - o Kp para las medias
  - o $\frac{p(p+1)}{2}$ para la covarianza compartida

- Se puede simplificar aún más suponiendo que todos los elementos fuera de la diagonal de la matriz de covarianza son nulos, es decir asumiendo la independencia de las variables.
- Así se obtiene el clasificador naive de Bayes, donde $\boldsymbol{x}|C_i$ tiene distribución normal univariada.

  $\widehat{\Sigma}$ y su inversa son matrices diagonales, por lo tanto obtenemos

$$G_i(\boldsymbol{x}) = -\frac{1}{2}\sum_{j=1}^{p}\left(\frac{x_{ij} - \hat{\mu}_{ij}}{\sigma_j}\right)^2 + \ln\left[\hat{P}(C_i)\right]$$

- El número de parámetros para estimar es:
  - o Kp para las medias
  - o p para la covarianza compartida

- Si se supone que todas las varianzas son iguales, la función discriminante se reduce a:

$$G_i(\boldsymbol{x}) = -\frac{\|\boldsymbol{x} - \hat{\boldsymbol{\mu}}_i\|^2}{2\hat{\sigma}^2} + \ln\left[\hat{P}(C_i)\right]$$

- Si todas las probabilidades a priori son iguales, tenemos que:

$$G_i(\boldsymbol{x}) = -\|\boldsymbol{x} - \hat{\boldsymbol{\mu}}_i\|^2$$

  Esto se llama el *clasificador de la media más cercana*

- Pero

$$\begin{aligned} G_i(\boldsymbol{x}) &= -\|\boldsymbol{x} - \hat{\boldsymbol{\mu}}_i\|^2 = -(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_i)^t(\boldsymbol{x} - \hat{\boldsymbol{\mu}}_i) \\ &= -\left(\boldsymbol{x}^t\boldsymbol{x} - 2\hat{\boldsymbol{\mu}}_i{}^t\boldsymbol{x} + \hat{\boldsymbol{\mu}}_i{}^t\hat{\boldsymbol{\mu}}_i\right) \end{aligned}$$

- Como $\boldsymbol{x}^t\boldsymbol{x}$ no depende de la clase, puede eliminarse y podemos escribir a la función como

$$G_i(\boldsymbol{x}) = \boldsymbol{w}_i^t\boldsymbol{x} + w_{i0}$$

donde

$$\boldsymbol{w}_i = \widehat{\boldsymbol{\mu}}_{\boldsymbol{i}}$$

$$w_{i0} = -\frac{1}{2}\|\widehat{\boldsymbol{\mu}}_{\boldsymbol{i}}\|^2$$

Características discretas

- En algunas aplicaciones, podemos tener atributos discretos toman uno de *n* valores distintos.
  - Por ejemplo, un atributo puede ser color (rojo, azul, verde, negro), y puede ser pixel (prendido, apagado).

- Sea $x_j \sim \mathcal{B}(p_{ij})$ una variable aleatoria Bernoulli tal que

$$p_{ij} \equiv P(x_j = 1 | C_i)$$

- Si las variables $x_1, \cdots, x_p$ son independientes entonces:

$$P(\boldsymbol{x}|C_i) = \prod_{j=1}^{p} {p_{ij}}^{x_j} \left(1 - p_{ij}\right)^{1-x_j}$$

- Este es otro ejemplo del clasificador naive de Bayes. La función discriminante es:

$$\begin{aligned} G_i(\boldsymbol{x}) &= \ln\left[P(\boldsymbol{x}|C_i)\right] + \ln\left[\hat{P}(C_i)\right] \\ &= \sum_{j=1}^{p} \left[x_j \ln p_{ij} + \left(1 - x_j\right) \ln\left(1 - p_{ij}\right)\right] + \ln\left[\hat{P}(C_i)\right] \end{aligned}$$

que es una función lineal.

- El estimador de p_{ij} es:

$$\hat{p}_{ij} = \frac{\sum_{l=1}^{n} x_j^l r_i^l}{\sum_{l=1}^{n} r_i^l}$$

- Esta aproximación se usa en la categorización de documentos, por ejemplo si hay que clasificar los reportes de noticias en varias categorías, como política, deportes, moda, etc.