Análisis Avanzado de Datos.

Nicolás López

Primer semestre de 2023

Estimación de parámetros mediante MLE

Estimación de parámetros mediante MLE

Recordemos los 4 elementos fundamentales en el aprendizaje estadístico:

- Proceso generador P.
- Variable de entrada/covariable/input: X (uni/multivariada).
- Variable de salida/variable respuesta/output: Y (univariada usualmente).
- Observaciones/realizaciones/mediciones: $(x_1, y_1), \ldots, (x_n, y_n)$.

En la estimación del modelo logístico (paramétrico), se tenía que para cada observación (x_i, y_i):

$$P_{\theta=(\beta_0,\beta_1)} : y_i \sim Ber(\pi(x_i|\beta_0, \beta_1))$$

Donde

$$\pi(x_i|\beta_0, \beta_1) = E(Y|X = x_i) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_i))}$$

O equivalentemente

$$\text{logit}(\pi(x_i|\beta_0, \beta_1)) = \beta_0 + \beta_1 x_i$$

Para $i = 1, ..., n$.

El modelo RLS/RLM puede ser estimado bajo las mismas condiciones probabilísticas en lugar de MCO.

$$P_{\theta=(\beta_0,\beta_1,\sigma)} : y_i \sim N(\mu(x_i), \sigma)$$

Donde

$$\mu(x_i) = E(Y|X = x_i) = \beta_0 + \beta_1 x_i$$

Para $i = 1, ..., n$.

Pensemos en RLS/RLM como ejemplo. Dado P_θ, asumimos que nuestras observaciones siguen dicho modelo, es decir:

$$y_1 \sim N(\beta_0 + \beta_1 x_1, \sigma), ..., y_n \sim N(\beta_0 + \beta_1 x_n, \sigma)$$

Con lo cual la fdp que gobierna el proceso aleatorio generador del dato i-ésimo está dada por:

$$f(y_i|x_i, \theta = (\beta_0, \beta_1, \sigma)) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma}\right)^2$$

Para $i = 1, ..., n$. Similar para el caso discreto en la regresión logística.

Con lo cual, la fdp que gobiera el proceso generador de todos los n datos, bajo independencia (recuede que $P(A \cap B) = P(A) \times P(B)$ bajo independencia) es igual a

$$f(y_1, ..., y_n | x_1, ..., x_n, \theta = (\beta_0, \beta_1, \sigma)) = \Pi_i f(y_i | x_i, \theta = (\beta_0, \beta_1, \sigma))$$

Al observar esta fdp conjunta como función de los parámetros, obtenemos la denominada función de verosimilitud del conjunto de datos:

$$L(\theta = (\beta_0, \beta_1, \sigma) | (x_1, y_1), ..., (x_n, y_n)) = \Pi_i f(y_i | x_i, \theta = (\beta_0, \beta_1, \sigma))$$

Y el método de máxima verosimilutud maximiza dicha función:

$$\hat{\theta}_{MLE} = \text{argmax}_{\theta}(L(\theta|(x_1, y_1), ..., (x_n, y_n)))$$

Con lo cual se encuentra el $\theta \in \Theta$ más probable para los datos observados. Equivalentemente se puede maximizar la log-verosimilitud, al ser log una función monótona:

$$\hat{\theta}_{MLE} = \text{argmax}_{\theta}(l(\theta|(x_1, y_1), ..., (x_n, y_n)))$$

Con

$$l(\theta|(x_1, y_1), ..., (x_n, y_n)) = \sum_i \log\left(f(y_i|x_i, \theta = (\beta_0, \beta_1, \sigma))\right)$$

Para el caso de RLS/RLM se tienen las conocidas soluciones

$$(\hat{\beta}_0, \hat{\beta}_1)' = (X'X)^{-1}X'y$$

Y

$$\hat{\sigma} = \frac{1}{n}(\mathbf{y} - \widehat{\mathbf{y}})'(\mathbf{y} - \widehat{\mathbf{y}})$$

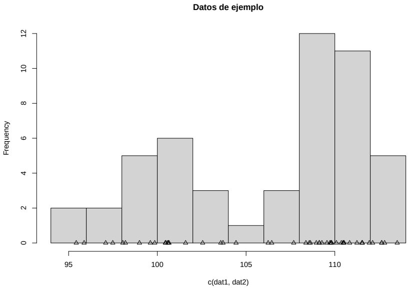
Note que para calcular esta verosimilitud, de manera tácita, asume que (x_i, y_i) con $i = 1, ...n$ son observados. El algoritmo *EM* (*Expectation Maximization*) es un acercamiento de estimación máximo verosimil en presencia de variables **latentes** (no observables).

Algoritmo EM - Mixtura de gaussianas

Algoritmo EM - Mixtura de gaussianas

Supongamos que contamos con una colección de observaciones univariadas $y_1 \ldots, y_n$ provenientes de la m.a. $Y_1, \ldots, Y_n$:

```
set.seed(100)
dat1 = rnorm(20,mean=100,sd =5)
dat2 = rnorm(30,mean=110,sd =2)
dat  = c(dat1,dat2)
hist(c(dat1,dat2),main="Datos de ejemplo")
points(c(dat1,dat2),rep(0,50),pch=2)
```



Asumiendo un modelo paramétrico, el normal parece apropiado, podemos encontrar el MLE de los parámetros fácilmente. Sin embargo la existencia bimodalidad implica la presencia de $\Delta_1, \ldots \Delta_n$ características no observables (observadas de manera tácita a través de las y's).

Tendríamos entonces que

$$Y_A \sim N(\mu_1, \sigma_1)$$

y

$$Y_B \sim N(\mu_2, \sigma_2)$$

Con lo cual

$$Y \sim (1 - \Delta)Y_A + \Delta Y_B$$

Con $P(\Delta = 1) = \pi$, resume el **proceso generador** de los datos $y_1, ..., y_n$. Si f_A es la densidad de la primera normal y f_B la de la segunda, se tiene:

$$f_Y(y) = (1 - \pi)f_A(y|\theta_1 = (\mu_1, \sigma_1)) + \pi f_B(y|\theta_2 = (\mu_2, \sigma_2))$$

Nos preguntamos por los estimadores MLE de los parámetros, que en este caso son $\theta = (\theta_1, \theta_2, \pi) = (\mu_1, \sigma_1, \mu_2, \sigma_2, \pi) \in \Theta$. La log-verosimilitud de la **mixtura de gaussianas** está dada por:

$$l(\theta|Y_T) = \sum_i \log((1-\pi) f_{A,\theta_1}(y_i) + \pi f_{B,\theta_2}(y_i))$$

Con $Y_T = (y_1, ..., y_n)$. La suma dentro del log dificulta la maximización directa sobre los 5 parámetros a ser encontrados. Asumamos Δ_i conocidos y reexpresemos la verosimilitud correspondientemente:

$$l_0(\theta|Y_T, \Delta_T) = \sum_i (1-\Delta_i) \log\left((1-\pi) f_{A,\theta_1}(y_i)\right) + \Delta_i \log\left(\pi f_{B,\theta_2}(y_i)\right)$$

Con $\Delta_T = (\Delta_1, ..., \Delta_n)$. Dado Δ_T, los MLE son la media y varianza de cada subpoblación y $\hat{\pi}$ es la proporción de $\Delta_i = 1$.

Sin embargo, como claramente los valores de Δ_i son desconocidos, estos son sustituidos en l_0 por su valor esperado

$$\gamma_i(\theta) = E(\Delta_i|\theta, Y_T) = 1 \times P(\Delta_i = 1|\theta, Y_T) + 0 \times P(\Delta_i = 0|\theta, Y_T)$$

$\gamma_i(\theta)$ es llamada la responsabilidad del modelo B ($Y_B \sim N(\mu_2, \sigma_2)$) para la observación i con $i = 1, .., n$.

El procedimiento de EM para dos mixturas gaussianas está dado por.

1. Tome valores iniciales para $\hat{\mu}_1, \hat{\sigma}_1, \hat{\mu}_2, \hat{\sigma}_2, \hat{\pi}$.
2. Paso E (*soft assignment*): Estime las responsabilidades $\gamma_i(\theta)$ para $i = 1, ..., n$:

$$\hat{\gamma}_i = \frac{\hat{\pi} f_{B,\hat{\theta}_2}(y_i)}{(1-\hat{\pi}) f_{A,\hat{\theta}_1}(y_i) + \hat{\pi} f_{B,\hat{\theta}_2}(y_i)}$$

3. Paso M: Estime los parámetros $\hat{\mu}_1, \hat{\sigma}_1, \hat{\mu}_2, \hat{\sigma}_2, \hat{\pi}$ dadas las responsabilidades.

$$\hat{\mu}_1 = \frac{\sum_i (1-\hat{\gamma}_i) y_i}{\sum_i (1-\hat{\gamma}_i)} \text{ y } \hat{\sigma}_1 = \frac{\sum_i (1-\hat{\gamma}_i)(y_i - \hat{\mu}_1)^2}{\sum_i (1-\hat{\gamma}_i)}$$

$$\hat{\mu}_2 = \frac{\sum_i \hat{\gamma}_i y_i}{\sum_i \hat{\gamma}_i} \text{ y } \hat{\sigma}_2 = \frac{\sum_i \hat{\gamma}_i (y_i - \hat{\mu}_1)^2}{\sum_i \hat{\gamma}_i}$$

$$\hat{\pi}_2 = \frac{\sum_i \hat{\gamma}_i y_i}{n}$$

4. Repita (2) y (3) hasta convergencia.

Nota en valores iniciales: $\hat{\mu}_1 = y_{rand}, \hat{\sigma}_1 = \hat{\sigma}, \hat{\mu}_2 = y_{rand}, \hat{\sigma}_2 = \hat{\sigma}, \hat{\pi} = 0.5$.

```r
library(mixtools)
set.seed(1)
init_mu = sample(length(dat),2)
gm      = normalmixEM(dat,k=2,lambda=c(0.5,0.5),
                              mu=c(dat[init_mu[1]],
                              dat[init_mu[2]]),
                              sigma=c(sd(dat),sd(dat)))
```

```
## number of iterations= 28
```

```r
round(gm$lambda,2)
```

```
## [1] 0.39 0.61
```

```r
round(gm$mu,2)
```

```
## [1] 100.20 110.26
```

```r
round(gm$sigma,2)
```

```
## [1] 2.74 1.67
```

```r
table(apply(gm$posterior,1,which.max),
      ifelse(1:50<20,"PI","PII"))
```

```
##
##      PI PII
##    1 19   0
##    2  0  31
```

Algoritmo EM - Caso general

Algoritmo EM - Caso general

El problema de mixtura anterior se simplificó al aumentarlo con v. latentes (Δ_i). Este es un caso particular, y podríamos tener otros escenarios bajo la misma caracterización probabilística:

A. Datos observados Z_T con log verosimilitud $l(\theta|Z_T)$.
B. Datos latentes o perdidos Z_M, con lo cual los datos completos son $T = (Z_T, Z_M)$, con log-verosimilitud $l_0(\theta|T)$

Para la mixtura gaussiana se tuvo $Z_T = Y_T$ y $Z_M = \Delta_T$.

El procedimiento de EM general está dado por.

1. Tome valores iniciales para $\hat{\theta}^{(0)}$.
2. Paso E: En el j-ésimo paso, calcule

$$Q(\theta|\hat{\theta}^{(j)}) = E(l_0(\theta|T)|(Z_T, \hat{\theta}^{(j)}))$$

3. Paso M: Encuentre la nueva estimación $\theta^{(j+1)}$ como:

$$\hat{\theta}^{(j+1)} = \text{argmax}_\theta Q(\theta|\hat{\theta}^{(j)})$$

4. Repita (2) y (3) hasta convergencia.

Note que $E(l_0(\theta|T)|(Z_T, \hat{\theta}^{(j)}))$ para la mixtura gaussiana no es más que $l_0(\theta|Y_T, \Delta_T)$ reemplazando Δ_T por $\hat{\gamma}_i(\hat{\theta})$ y el paso M no es más que las medias y varianzas estimadas.

Reducción de dimensionalidad

Reducción de dimensionalidad

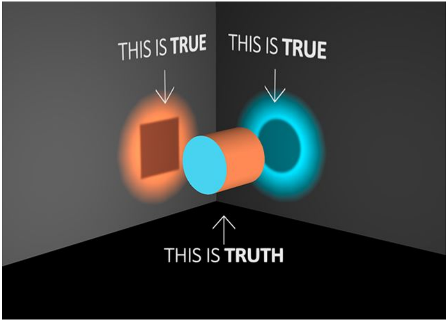


Figure 1: Tomado de este enlace

Estas figuras permiten entender la gran verdad a partir de verdades más simples. Note que estas verdades simples también son *latentes*.

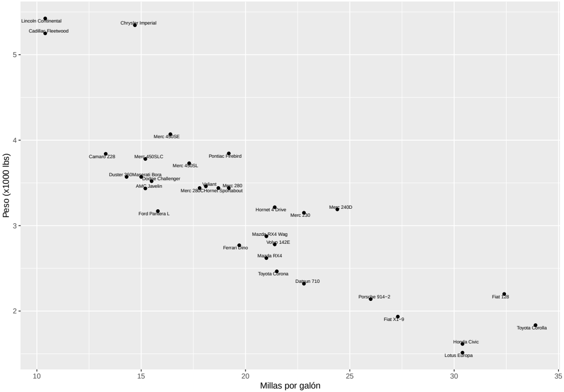
Para motivar los métodos de reducción de dimensionalidad que usaremos en esta sesión, usaremos el conjunto de datos multivariado mtcars. Estos datos describen 32 modelos de automóviles, tomados de una revista de automovilismo estadounidense (revista Motor Trend de 1974).

Para cada modelo de automóvil, se tienen 11 variables:

- **mpg** - Consumo de combustible (millas por galón EE.UU.).
- cyl - Número de cilindros.
- disp - Desplazamiento (cu.in.) el volumen combinado de los cilindros del motor.
- **hp** - Potencia bruta.
- drat - Relación del eje trasero: esto describe cómo un giro del eje de transmisión corresponde a un giro de las ruedas
- **wt** - Peso (1000 lbs)
- qsec - Tiempo de 1/4 de milla: la velocidad y aceleración de los autos .
- vs - Bloque del motor: esto indica si el motor del vehículo tiene forma de "V" o si es una forma recta más común.
- am - Transmisión: indica si la transmisión del automóvil es automática (0) o manual (1).
- gear - Número de marchas hacia adelante.
- carb - Número de carburadores.

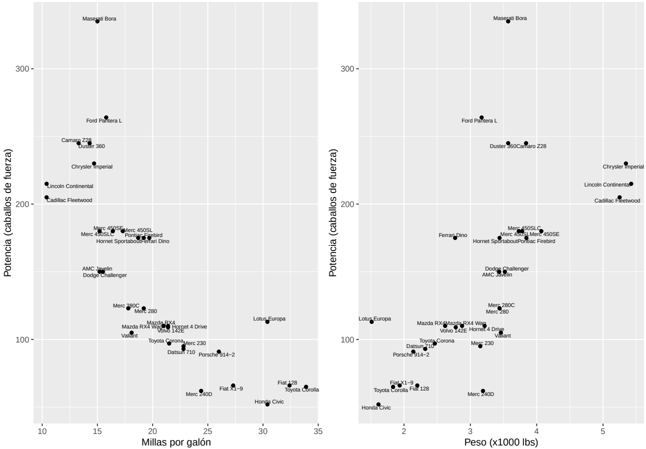
PCA

Si medimos dos variables podemos ver relaciones entre los vehículos



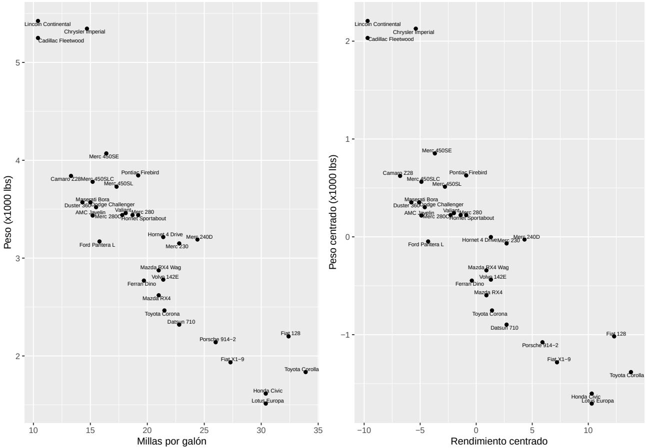
Fiat 128 y Toyota Corolla son los livianos de mayor rendimiento, mientras que Lincoln Continental y Cadillac Fleetwood son vehículos pesados de bajo rendimiento.

Si midiéramos una tercera variable, podríamos cruzarla con las dos variables anteriores y visualizar nuevas posibles agrupaciones:

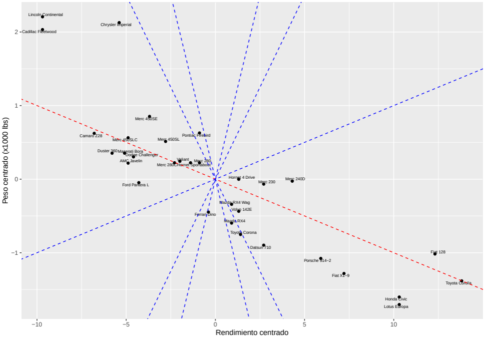


Ahora, Fiat 128 y Toyota Corolla, los livianos de mayor rendimiento, tienen menor potencia. Mientras que el Lincoln Continental el y Cadillac Fleetwood, los pesados de bajo rendimiento tienen una alta potencia.

El PCA asume los datos centrados, tomando como ejemplo las variables de rendimiento y peso:



Al gráfico centrado se añade una línea a través del origen. Esta se rota hasta encontrar aquella que se ajuste lo mejor posible a los datos:



En rojo observamos la línea que:

- Minimiza las distancias2 de cada punto a su proyección. O, de manera equivalente:
- Maximiza las distancias2 de cada proyección al origen (llamada suma de distancias proyectadas al cuadrado).

La línea roja corresponde al PC1: tiene una pendiente igual a -0.1 = -1/10: es decir, c.aumento en 10 unidades de rendimiento (centrado), disminuye una unidad de peso (centrado).

Formalmente, se tiene:

- x_i. i-ésimo punto en en espacio.
- $\langle x_i, w\rangle$ proyección de x_i en w.
- $w\langle x_i, w\rangle$ vector proyección de x_i en w.
- $\sum_i w\langle x_i, w\rangle = w(\sum_i x_i)w') = 0$. (vector 0).
- $\|x_i - w\langle x_i, w\rangle\|^2 = x_i'x_i + (w'x_i)^2$. Error de proyección de w para x_i.
- $ECM(w) = \sum(x_i'x_i + (w'x_i)^2)$. Error de proyección de w.

$$\text{argmin}_w ECM(w) = \text{argmax}_w \sum (w'x_i)^2 = \text{argmax}_w \sum \langle w, x_i\rangle$$

Es decir, buscamos el w que maximiza las proyecciones, como lo visualizamos anteriormente.

Note que

$$\frac{1}{n}\sum(w'x_i)^2 = \left(\frac{1}{n}\sum(w'x_i)\right)^2 + \mathrm{Var}(w'x_1, ..., w'x_n) = \mathrm{Var}(w'x_1, ..., w'x_n)$$

Y como queremos maximizar $\sum(w'x_i)^2$, esto es equivalente a maximizar $ Var(w'x_1,. . . , w'x_n)$, la varianza de las proyecciones.

De manera matricial se tiene la matriz diseño $X \in M(n, p)$ con los vectores de observaciones $x_i, i = 1, ..., n$ en filas.

$$\begin{aligned}\sigma^2(w) = \frac{1}{n}\sum(w'x_i)^2 &= \frac{1}{n}(Xw)'Xw \\ &= \frac{1}{n}w'X'Xw \\ &= w'\mathrm{Cov}(X)w\end{aligned}$$

Y el PC1 está dado por

$$\mathrm{argmax}_w \sigma^2(w) = \mathrm{argmax}_w w'\mathrm{Cov}(X)w$$

El cual tiene infinitas soluciones, por lo cual se restringe a $\|w\| = 1$, obteniendo así un problema de optimización con restricción (Lagrange).

Se tiene entonces

$$L(W,\lambda) = w'\mathrm{Cov}(X)w + \lambda(w'w - 1)$$

De dónde

$$\frac{\delta L}{\delta \lambda} = w'w' - 1$$
$$\frac{\delta L}{\delta w} = 2\mathrm{Cov}(X)w - 2\lambda w$$

Igualando a cero

$$w'w' = 1$$
$$\mathrm{Cov}(X)w = \lambda w$$

La definición de un vector propio de $\mathrm{Cov}(X)$. Al multiplicar por w' en la segunda ecuación vemos que:

$$w'\mathrm{Cov}(X)w = w'\lambda w = \lambda$$

Se selecciona la pareja (λ_1, w_1), con λ_1 el mayor valor propio y w_1 su correspondiente vector propio, para maximizar L. De manera iterativa se obtienen los demás componentes.

El ajuste desde R se realiza fácilmente y la interpretación de los componentes es clara, **note el parámetro de scale** en el ajuste

```
mtcars_sm <- mtcars %>% select(c(mpg,hp,wt))
pc_1      <- prcomp(mtcars_sm, scale = TRUE)
pc_1
```

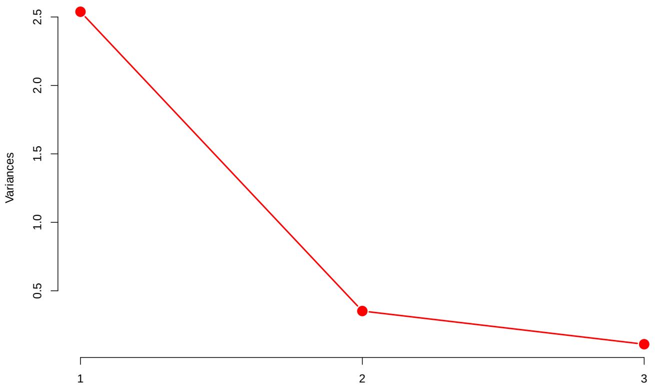
```
## Standard deviations (1, .., p=3):
## [1] 1.5931712 0.5933736 0.3312298
##
## Rotation (n x k) = (3 x 3):
##            PC1        PC2        PC3
## mpg -0.6032696  0.1634988 -0.7805984
## hp   0.5512426  0.7928131 -0.2599595
## wt   0.5763656 -0.5871248 -0.5684076
```

```
summary(pc_1)
```

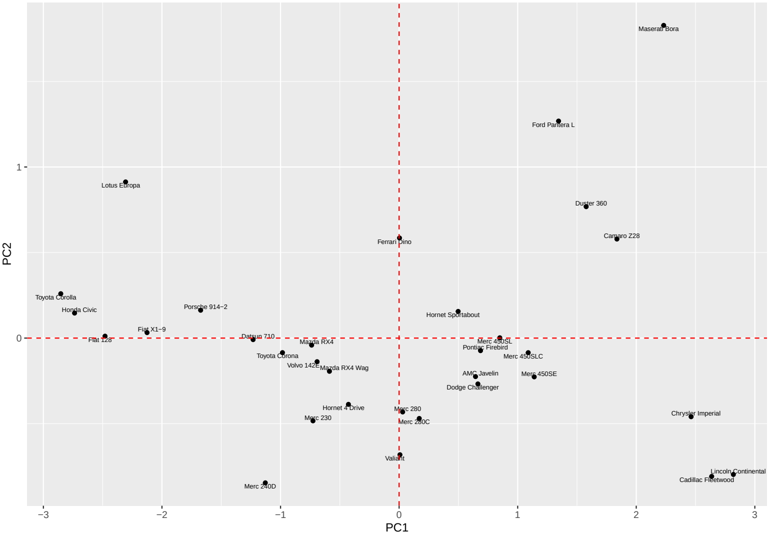
```
## Importance of components:
##                           PC1    PC2     PC3
## Standard deviation     1.5932 0.5934 0.33123
## Proportion of Variance 0.8461 0.1174 0.03657
## Cumulative Proportion  0.8461 0.9634 1.00000
```

Y finalmente se obtiene una representación gráfica de los valores propios.

```
screeplot(pc_1, col = "red", pch = 16,
type = "lines", cex = 2, lwd = 2, main = "")
```



Junto con el plano factorial correspondiente



Referencias

Referencias

1. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning. Springer.
2. Garet, Witten, Hastie, Tibshirani. Introduction to Statistical Learning with R.