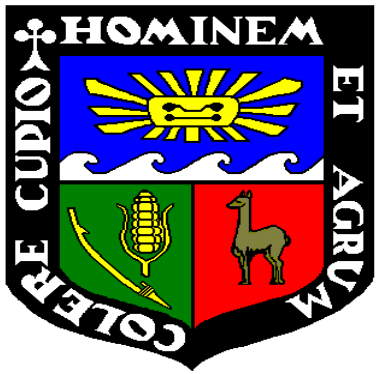
Técnicas Multivariadas Avanzadas

Métodos de remuestreo

Ms Carlos López de Castilla Vásquez

Universidad Nacional Agraria La Molina

2014-2



Introducción

- Los métodos de *remuestreo* constituyen una herramienta importante e indispensable dentro del Statistical Learning.
- Se discuten dos métodos de remuestreo:

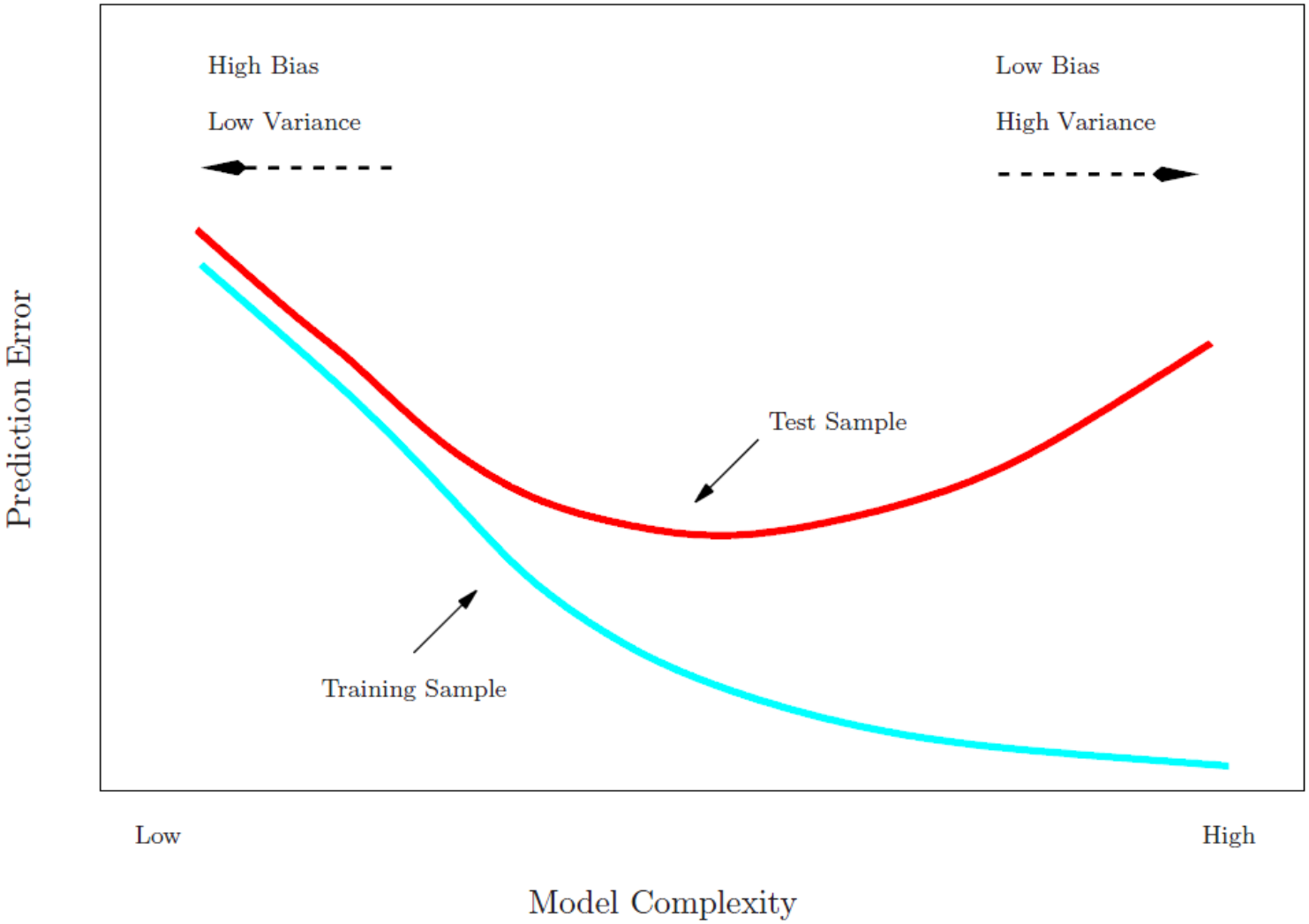
  validación cruzada y bootstrap.

- La idea es tomar *repetidamente* muestras a partir de la data de entrenamiento y estimar en cada una de estas muestras el modelo de interés para obtener información adicional.
- Validación cruzada: usado para estimar el error de prueba o elegir el grado apropiado de flexibilidad de un modelo.
- Bootstrap: usado para medir el comportamiento de las estimaciones a través del error estándar, sesgo, etc.

Error de entrenamiento y prueba

- Es importante recordar la diferencia entre el *error de prueba* y el *error de entrenamiento*.
- El *error de prueba* es el error promedio que resulta de aplicar un método para predecir el valor de la variable respuesta correspondiente a una nueva observación que no fue usada para estimar el modelo.
- El *error de entrenamiento* puede ser calculado aplicando el método a las observaciones usadas en la estimación del modelo.
- La tasa de error de entrenamiento es por lo general diferente de la tasa de error de prueba, en particular la primera *subestima* la segunda.

Error de entrenamiento y prueba



Estimación del error de predicción

- Para estimar el error de predicción se requiere un conjunto de prueba grande que por lo general no esta disponible.
- Algunos métodos realizan un ajuste al error de entrenamiento para estimar el de error de prueba que incluyen:

  el estadístico Cp, AIC y BIC.

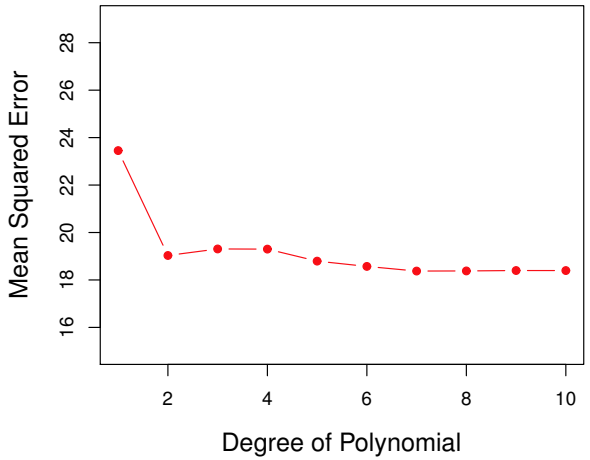
- Los métodos de remuestreo estiman el error de predicción aplicando el modelo estimado a un conjunto de datos que no participo del proceso de estimación y que fue seleccionado previamente de la data de entrenamiento.

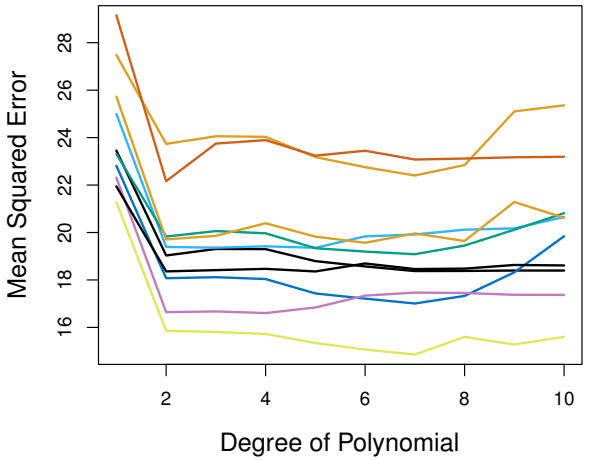
Conjunto de entrenamiento y validación

- Se divide aleatoriamente la muestra en dos partes: un *conjunto de entrenamiento* y un *conjunto de validación*.
- El modelo se estima usando el conjunto de entrenamiento. Luego el modelo es usado para predecir las observaciones dentro del conjunto de validación.
- El error obtenido en el conjunto de validación proporciona una estimación del error de prueba a través del *MSE* cuando la variable respuesta es cuantitativa y la *tasa de mala clasificación* si la variable respuesta es cualitativa.

Data Automóvil

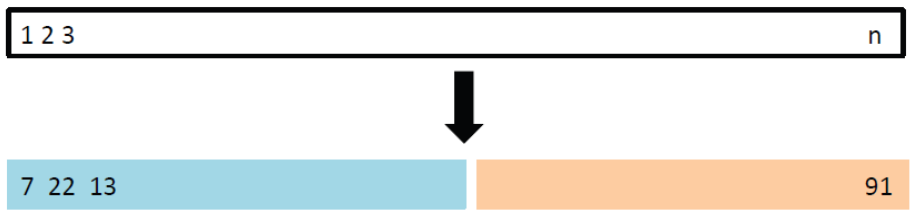
- Se divide aleatoriamente las 392 observaciones en un conjunto de entrenamiento y un conjunto de validación cada uno con 196 observaciones.
- Se estiman modelos de regresión lineal con diferente flexibilidad.





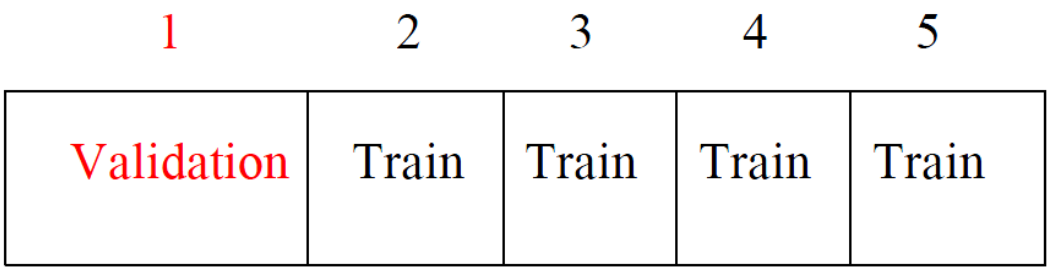
Metodología de validación

- La estimación del error de prueba puede ser altamente variable dependiendo precisamente de las observaciones incluidas en el conjunto de entrenamiento y de validación.
- En la metodología de validación solo un conjunto de observaciones, aquellas incluidas en el conjunto de entrenamiento, son usadas para estimar el modelo.
- Lo anterior sugiere que el error en el conjunto de validación puede tender a *sobreestimar* el error de prueba para el modelo estimado con la data completa.



Validación cruzada K-fold

- Se trata de una metodología ampliamente usada para estimar el error de prueba. Estas estimaciones pueden ser usadas para seleccionar el mejor modelo.
- La idea es dividir aleatoriamente la data en K partes de igual tamaño. Se deja de lado la parte k y se estima el modelo con las $K-1$ partes combinadas y luego se obtienen las predicciones para la parte k.
- Se repite el proceso para cada parte $k = 1, 2, \cdots, K$.



Validación cruzada K-fold

- Sean $C_1, C_2, \cdots, C_K$ las K partes donde C_k denota los índices de las observaciones en la parte k. Existen n_k observaciones en la parte k. Si n es múltiplo de K entonces $n_k = n/K$.
- Se calcula:

$$CV_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} MSE_k$$

donde $MSE_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$ y $\hat{y}_i$ es la estimación para la observación i que se encuentra en la parte k removida.

- Se obtiene *leave-one out cross-validation* (LOOCV) cuando $K = n$.

Validación cruzada K-fold

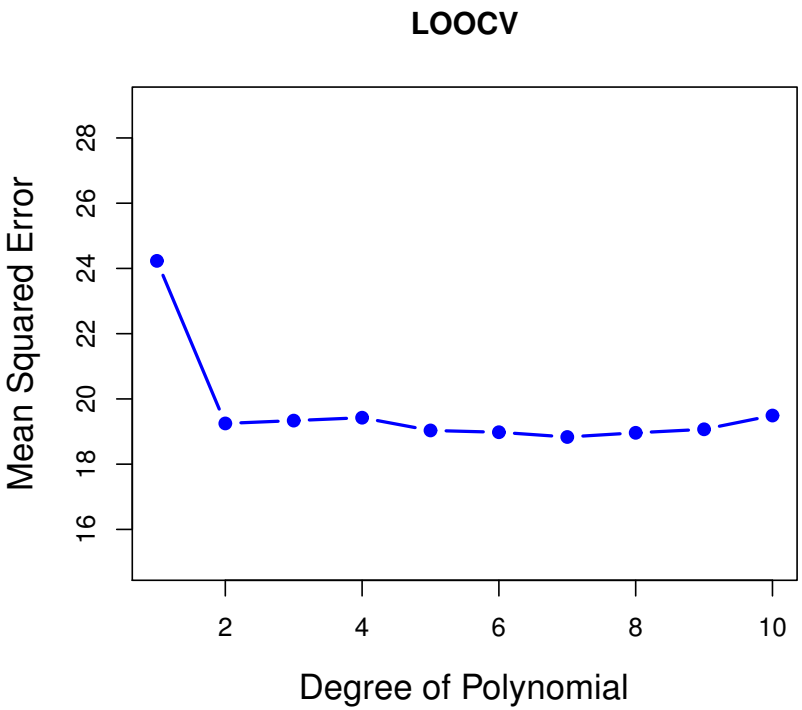
- La metodología LOOCV supone un importante esfuerzo computacional.
- Sin embargo se puede probar que:

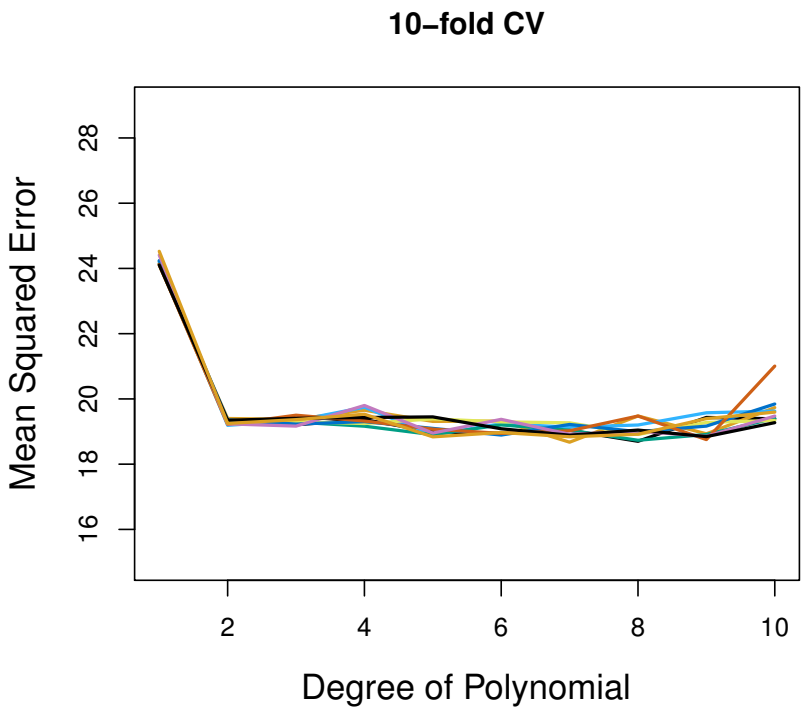
$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

donde $\hat{y}_i$ se obtiene con el modelo estimado usando todos los datos y h_i es el leverage.

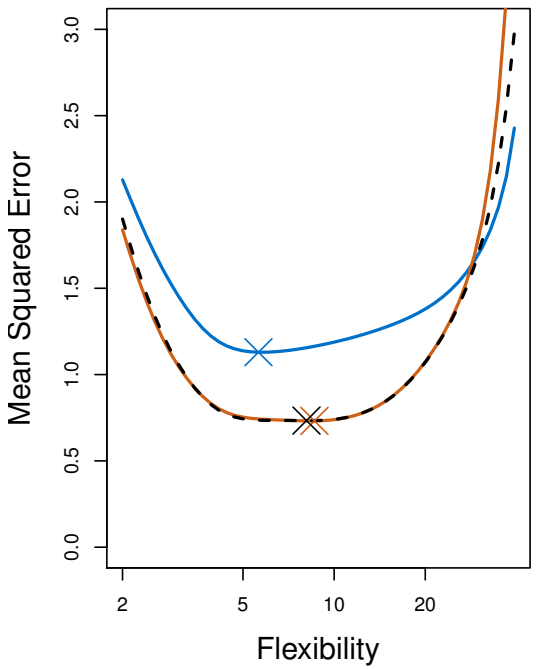
- Las estimaciones por LOOCV se encuentran altamente correlacionadas y podrían tener mucha variancia.
- Una mejor opción es usar $K = 5$ o $K = 10$.

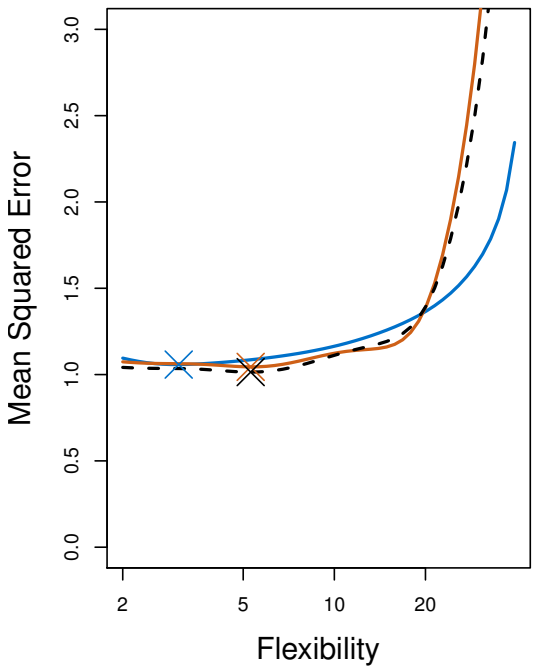
Validación cruzada K-fold

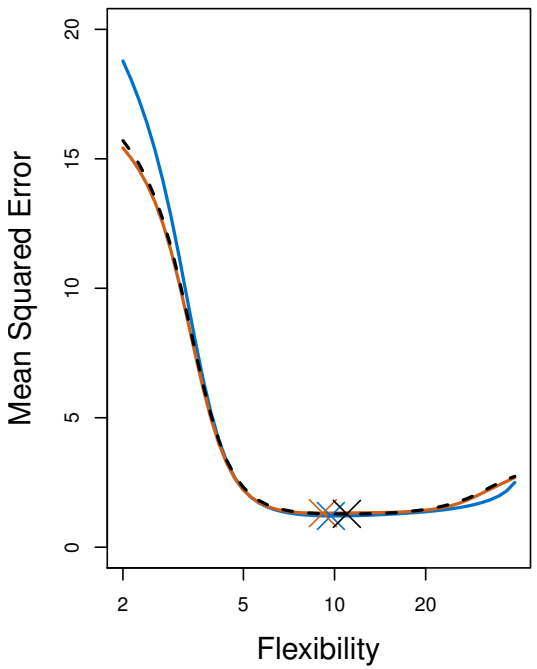




Validación cruzada K-fold







Validación cruzada para clasificación

- Sean $C_1, C_2, \cdots, C_K$ las K partes donde C_k denota los índices de las observaciones en la parte k.
- Se calcula:

$$CV_K = \sum_{k=1}^{K} \frac{n_k}{n} Err_k$$

donde $Err_k = \sum_{i \in C_k} I(y_i \neq \hat{y}_i)/n_k$.

- La estimación del error estándar es:

$$\hat{SE}(CV_K) = \sqrt{\frac{1}{K-1} \sum_{k=1}^{K} (Err_k - \overline{Err}_k)^2}$$

Introducción

- Se trata de una herramienta estadística poderosa y flexible que puede ser usada para cuantificar la incertidumbre asociada con un estimador o un método dentro del Statistical Learning.
- El uso del término deriva de la frase en "The Surprising Adventures of Baron Munchausen" de Erich Raspe:

> *The Baron had fallen to the bootom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by own bootstraps.*

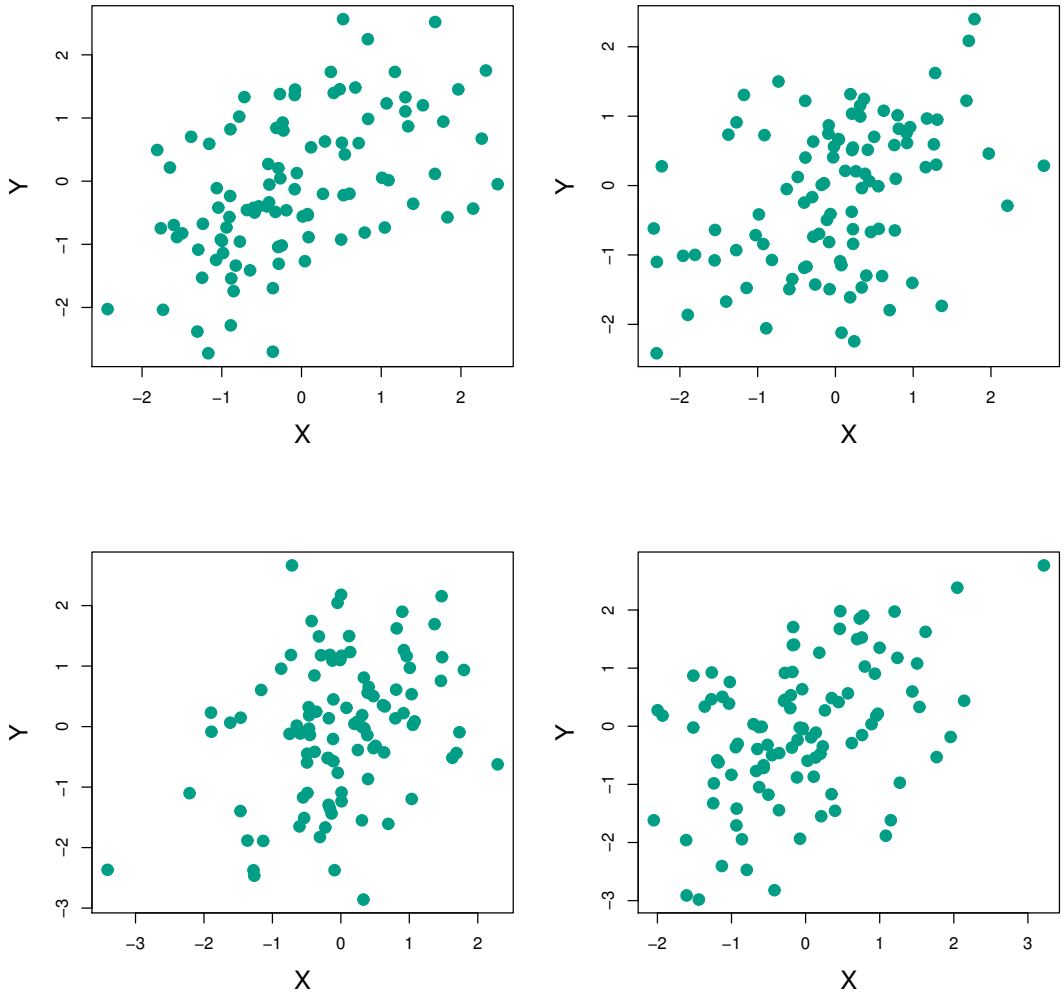
- No se trata del mismo término usado en las ciencias de la computación donde el término *boot* corresponde a realizar un conjunto de operaciones primarias esenciales.

Ejemplo

- Suponga que se desea invertir una suma fija de dinero en dos activos financieros que brindan retornos de X y Y que se consideran cantidades aleatorias.
- Nos interesa conocer la fracción α de nuestro dinero a invertir en X y la fracción restante $(1-\alpha)$ en Y.
- Se desea escoger α para minimizar el riesgo total, o variancia, de nuestra inversión. En otras palabras se desea minimizar $\text{Var}(\alpha X + (1-\alpha)Y)$.
- Se puede probar que:

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}$$

Ejemplo



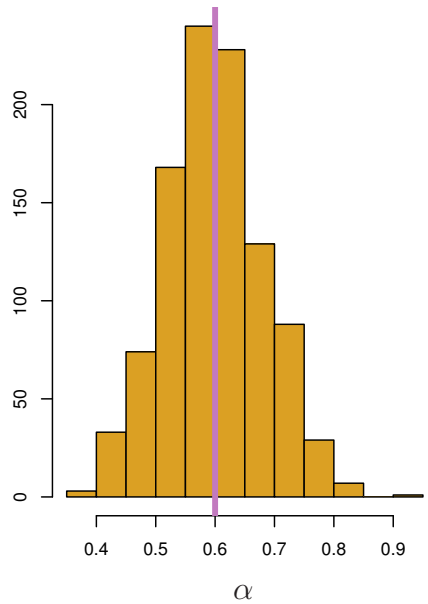
Ejemplo

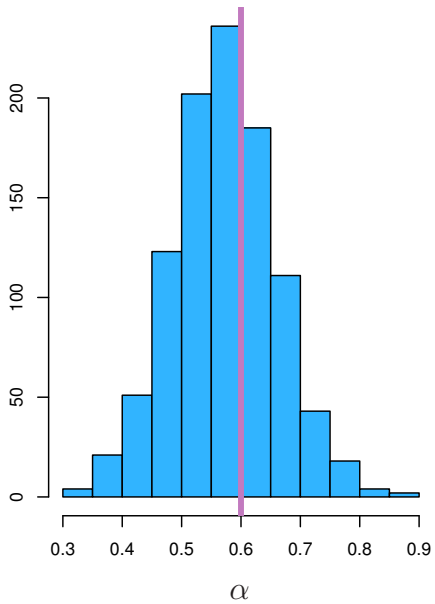
- Supongamos que $\sigma_X^2 = 1$, $\sigma_Y^2 = 1{,}25$ y $\sigma_{XY} = 0{,}5$ y por consiguiente el verdadero valor para $\alpha = 0{,}6$.
- Cada panel muestra 100 valores simulados para X y Y. Las estimaciones para α es 0.576, 0.532, 0.657 y 0.651.
- Para estimar la desviación estándar de $\hat{\alpha}$ se repite el proceso de simulación anterior 1000 veces para obtener $\hat{\alpha}_1, \hat{\alpha}_2, \cdots, \hat{\alpha}_{1000}$.
- La media sobre las 1000 estimaciones es:

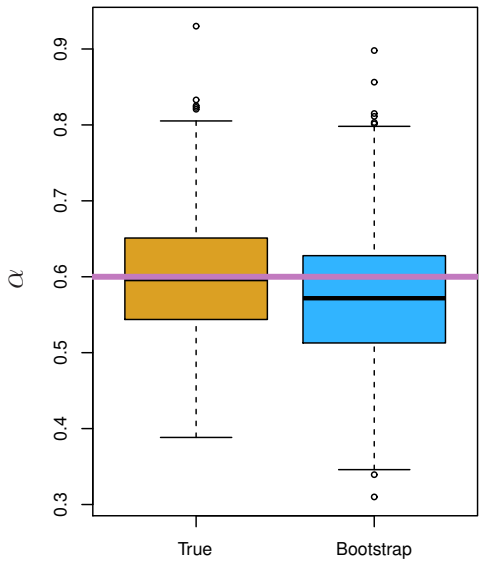
$$\overline{\alpha} = \frac{1}{1000} \sum_{r=1}^{1000} \hat{\alpha}_r = 0{,}5996$$

y su desviación estándar $SE(\hat{\alpha}) \approx 0{,}083$.

Ejemplo


0
50
100
150
200
0.4
0.5
0.6
0.7
0.8
0.9
α

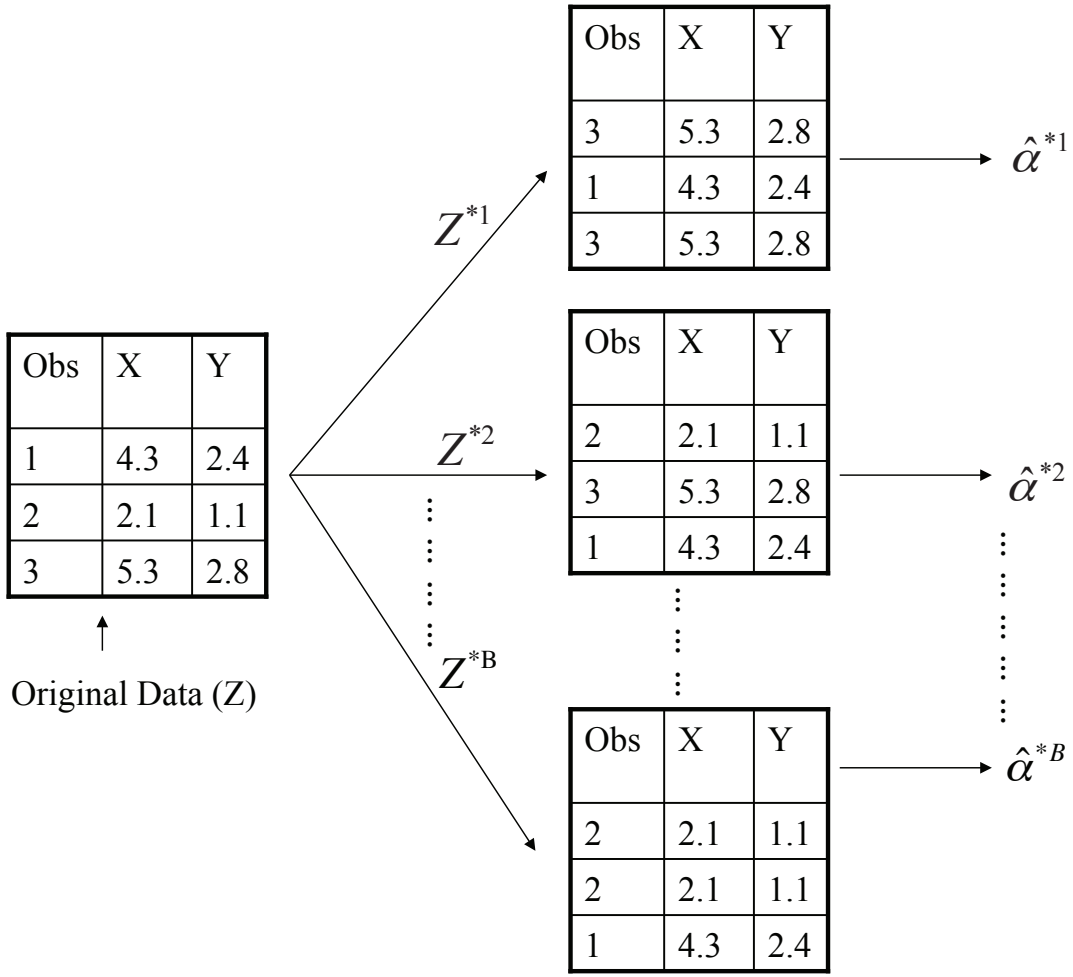

0
50
100
150
200
0.3
0.4
0.5
0.6
0.7
0.8
0.9
α


0.3
0.4
0.5
0.6
0.7
0.8
0.9
α
True
Bootstrap

Bootstrap en general

- El procedimiento anterior no puede ser aplicado por que en la vida real no es posible generar nuevas muestras de la población original.
- Sin embargo la metodología bootstrap permite *imitar* el proceso anterior para estimar la variabilidad de un estimador sin necesidad de generar muestras adicionales.
- Las muestras se obtienen *con reemplazo* a partir de la data original y tienen el mismo tamaño del conjunto original.
- Como resultado algunas observaciones pueden aparecer más de una vez en cada muestra bootstrap.

Bootstrap en general



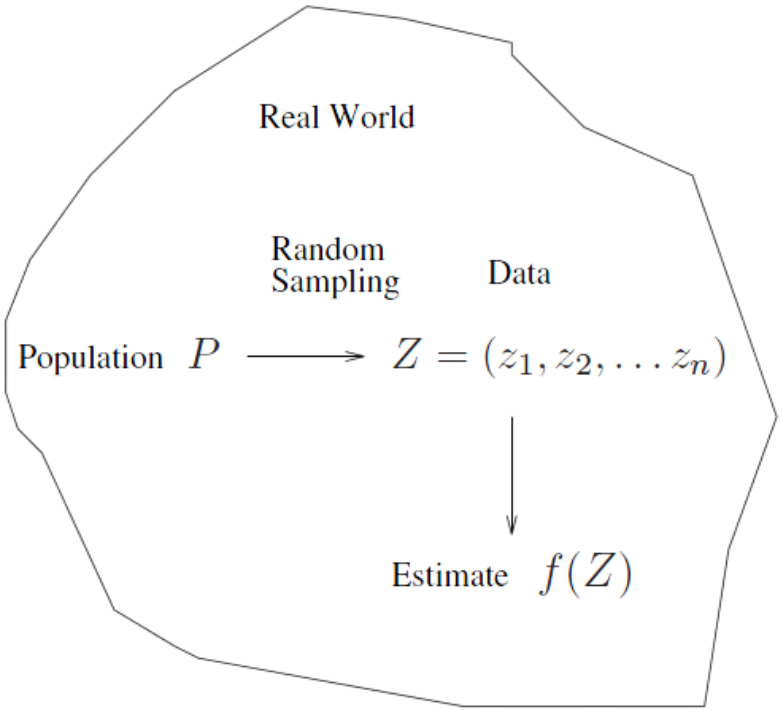
Bootstrap en general

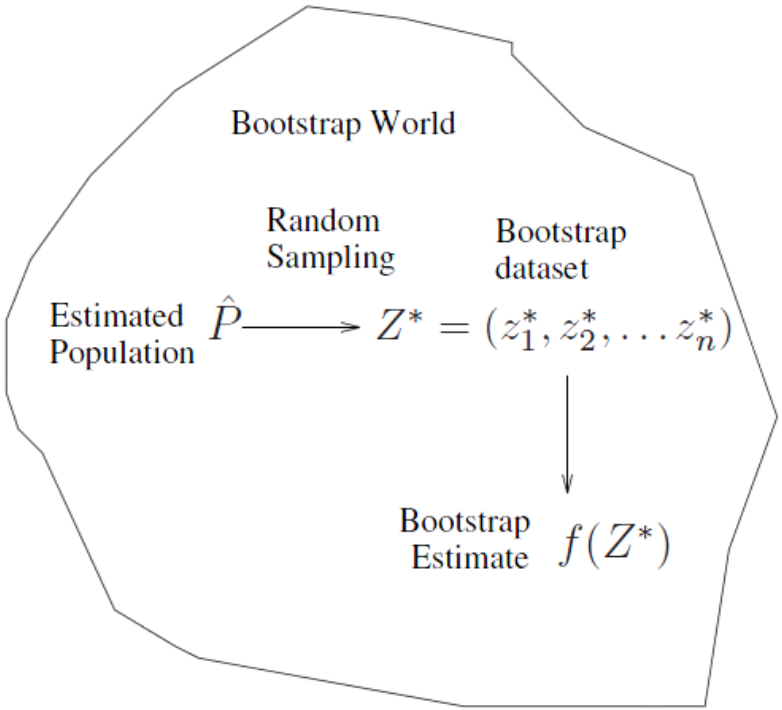
- Sea Z^{*1} la primera muestra bootstrap con la que se obtiene la estimación $\hat{\alpha}^{*1}$.
- El procedimiento anterior se repite B veces (100 o 1000) lo cual permite obtener: $Z^{*1}, Z^{*2}, \cdots, Z^{*B}$ y las estimaciones correspondientes: $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \cdots, \hat{\alpha}^{*B}$.
- Se estima el *error estándar de la estimación bootstrap* usando:

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}(\hat{\alpha}^{*r} - \bar{\hat{\alpha}}^{*})^2}$$

- Lo anterior es una estimación del error estándar de α estimado a partir del conjunto original de datos.

Bootstrap en general





Otros usos del bootstrap

- Las primeras aplicaciones del bootstrap estuvieron asociadas al cálculo del error estándar de un estimador.
- También proporcionan intervalos de confianza aproximados para un parámetro.
- En el histograma central los cuantiles 5 % y 95 % de los 100 valores son (0.43, 0.72) y representa un intervalo de confianza aproximado del 90 % para el verdadero α.
- El intervalo anterior es llamado *intervalo de confianza bootstrap por percentiles*. Es el método más sencillo para obtener un intervalo de confianza con esta metodología.

¿Es posible estimar el error de predicción con bootstrap?

- En validación cruzada cada una de K partes usada para el proceso de validación es diferente de las otras $K-1$ usadas para el entrenamiento. No existe superposición.
- Para estimar el error de predicción se podría pensar en usar cada muestra bootstrap como la muestra de entrenamiento y la muestra original como la muestra de validación.
- Pero cada muestra bootstrap tiene alta superposición con la data original. De esta forma podríamos estaríamos subestimando el verdadero error de predicción.
- Es peor aún considerar la muestra completa como muestra de entrenamiento y el conjunto bootstrap como muestra de validación. Es preferible usar validación cruzada.