

Clase 2- Módulo 2: Introducción a la analítica

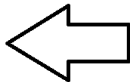
Mauricio Alejandro Mazo Lopera

Universidad Nacional de Colombia
Facultad de Ciencias
Escuela de Estadística
Medellín



UNIVERSIDAD
NACIONAL
DE COLOMBIA

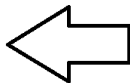
DATOS DE ENTRENAMIENTO



**¿ES MEJOR AJUSTAR UN
ÚNICO MODELO?**



SUBMUESTRAS



**¿AJUSTAR VARIOS
MODELOS CON
SUBMUESTRAS?**

ó

- Estos métodos pueden ser usados para evaluar el rendimiento de un modelo (“qué tan bien funciona”) o también para seleccionar el mejor modelo (“qué grado de flexibilidad puede tener para funcionar bien”).

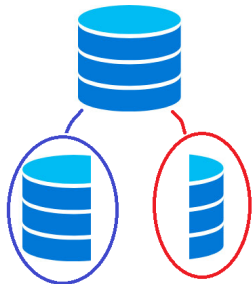
- Estos métodos pueden ser usados para evaluar el rendimiento de un modelo (“qué tan bien funciona”) o también para seleccionar el mejor modelo (“qué grado de flexibilidad puede tener para funcionar bien”).
- Dos de los métodos más utilizados son: **Validación cruzada** (cross-validation) y **bootstrap**.

- Estos métodos pueden ser usados para evaluar el rendimiento de un modelo (“qué tan bien funciona”) o también para seleccionar el mejor modelo (“qué grado de flexibilidad puede tener para funcionar bien”).
- Dos de los métodos más utilizados son: **Validación cruzada** (cross-validation) y **bootstrap**.
- Es importante tener en cuenta que existen dos tasas de error: **tasa de error del entrenamiento** (obtenida con los datos de entrenamiento) y **tasa de error de prueba** (obtenida con los datos de prueba).

Validación cruzada (Cross-Validation)

- **Método del conjunto de validación.**

CONJUNTO DE DATOS

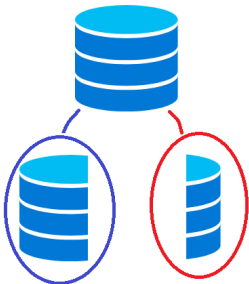


ENTRENAMIENTO VALIDACIÓN

Validación cruzada (Cross-Validation)

- Método del conjunto de validación.

CONJUNTO DE DATOS

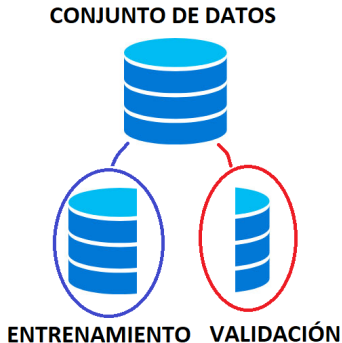


ENTRENAMIENTO VALIDACIÓN

Como no siempre se cuenta con datos de prueba (*set test*) se parte la base de datos en dos sub-bases. Una se usa para ajustar el modelo (**datos de entrenamiento**) y la otra para validar el modelo (**datos de validación**).

Validación cruzada (Cross-Validation)


- Método del conjunto de validación.




Como no siempre se cuenta con datos de prueba (*set test*) se parte la base de datos en dos sub-bases. Una se usa para ajustar el modelo (**datos de entrenamiento**) y la otra para validar el modelo (**datos de validación**). La tasa de error obtenida con el conjunto de validación (usualmente se usa el MSE cuando la variable respuesta es cuantitativa) es un estimador de la tasa de error de prueba, en inglés *test error rate*.

Mean Squared Error (MSE)

$$MSE = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_i - \hat{f}(x_i))^2$$

n_{test}  Número de datos de test

\hat{f}  Modelo ajustado con los datos de entrenamiento

Ejemplo 1: Método del conjunto de validación

Considere la base de datos **Auto**:

```
require(ISLR)
dim(Auto)
names(Auto)
```

```
## [1] 392  9
```

```
## [1] "mpg"          "cylinders"    "displacement" "horsepower"   "weight"
## [6] "acceleration" "year"         "origin"       "name"
```

Ejemplo 1: Método del conjunto de validación

Considere la base de datos **Auto**:

```
require(ISLR)
dim(Auto)
names(Auto)
```

```
## [1] 392    9
```

```
## [1] "mpg"          "cylinders"    "displacement" "horsepower"   "weight"
## [6] "acceleration" "year"         "origin"       "name"
```

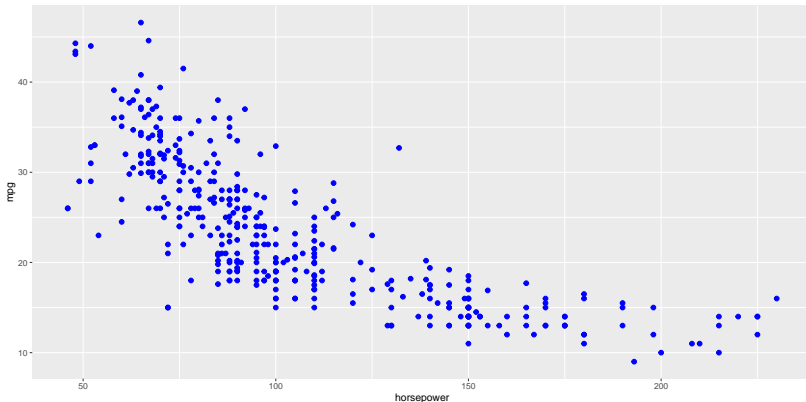
Seleccionemos los datos de prueba y los de entrenamiento

```
cedula<-33234 # Cambiando esta semilla cambian las submuestras
set.seed(cedula)
subset1<-sample((1:nrow(Auto)), 196)
Train1<-Auto[subset1,] # 196 datos de entrenamiento
Test1<-Auto[-subset1,] # 196 datos de prueba
```

Ejemplo 1: Método del conjunto de validación

Para relacionar mpg (millas por galón) con horsepower (caballos de fuerza), graficamos:

```
require("ggplot2")  
ggplot()+  
  geom_point(data=Auto, aes(horsepower, mpg), color="blue", size=2)
```



Ejemplo 1: Método del conjunto de validación

¿La relación entre estas dos variables es lineal, cuadrática, cúbica o de un grado más alto?

Ejemplo 1: Método del conjunto de validación

¿La relación entre estas dos variables es lineal, cuadrática, cúbica o de un grado más alto?

Para responder a esto, ajustemos varios modelos con distintos grados polinómicos para `horsepower` explicando `mpg`, utilizando los datos de entrenamiento y evaluando qué también funciona, calculando el MSE (Media Cuadrática del Error) con los datos de prueba.

Ejemplo 1: Método del conjunto de validación

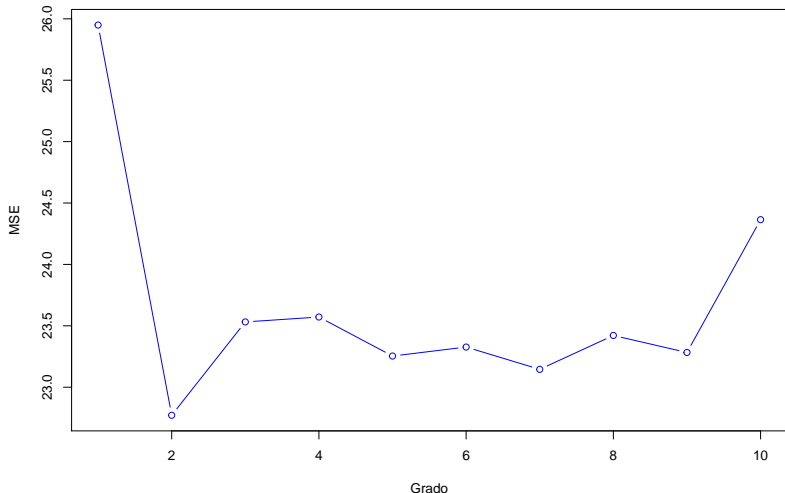
¿La relación entre estas dos variables es lineal, cuadrática, cúbica o de un grado más alto?

Para responder a esto, ajustemos varios modelos con distintos grados polinómicos para horsepower explicando mpg, utilizando los datos de entrenamiento y evaluando qué también funciona, calculando el MSE (Media Cuadrática del Error) con los datos de prueba.

```
MSE<-vector()
for (i in 1:10){
  Modelo1<-lm(mpg~poly(horsepower,i), data=Train1)
  Pred1<-predict(Modelo1,Test1)
  MSE[i]<-mean((Pred1-Test1$mpg)^2)
}
```

Ejemplo 1: Método del conjunto de validación

```
plot(1:10,MSE, xlab="Grado", ylab="MSE",type="b",col=4)
```



Método del conjunto de validación

De este ejemplo que acabamos de ver podemos concluir dos hechos:

De este ejemplo que acabamos de ver podemos concluir dos hecho:

- 1 El valor MSE que utilizamos para estimar la **tasa de error de prueba** depende de las submuestras o subconjuntos de datos que tomemos para el entrenamiento y la prueba. Si se cambia la semilla en `set.seed(cedula)` por otro valor, el gráfico del MSE en función del grado del polinomio también cambia.

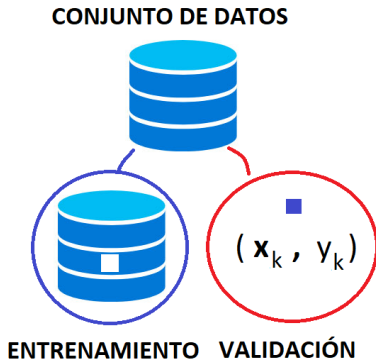
De este ejemplo que acabamos de ver podemos concluir dos hechos:

- 1 El valor MSE que utilizamos para estimar la **tasa de error de prueba** depende de las submuestras o subconjuntos de datos que tomemos para el entrenamiento y la prueba. Si se cambia la semilla en `set.seed(cedula)` por otro valor, el gráfico del MSE en función del grado del polinomio también cambia.
- 2 Si el conjunto de entrenamiento es pequeño, es decir, tiene pocos datos, entonces se corre el riesgo de ajustar un modelo que no apunte realmente a lo que debería. Técnicamente esto puede ocasionar que el MSE sea muy grande y se sobreestime la **tasa de error de prueba**.

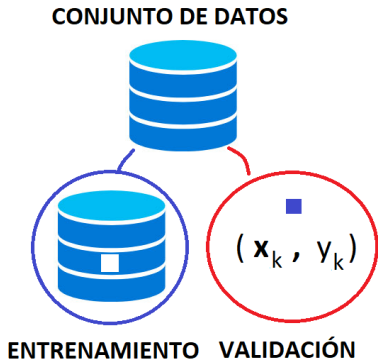
Ejemplo gráfico de sobreajuste:



Validación cruzada dejando uno afuera:



Validación cruzada dejando uno afuera:



Este tipo de validación que en inglés se escribe 'Leave-One-Out Cross-Validation' (LOOCV) consiste en dividir el conjunto de datos en dos subconjuntos: uno con un único dato (conjunto de prueba) y otro con todos los datos restantes (conjunto de entrenamiento). Este proceso se repite n veces, quitando en cada caso el k -ésimo dato y encontrando el $MSE_k = (y_k - \hat{y}_k)^2$, para $k = 1, 2, \dots, n$.

Validación cruzada dejando uno afuera:

Cuando repetimos el proceso LOOCV n -veces (**¿Por qué?**) obtenemos los valores $MSE_1, MSE_2, \dots, MSE_n$ y el estimador para la **tasa de error de prueba** está dado por:

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^n MSE_i$$

Validación cruzada dejando uno afuera:

Cuando repetimos el proceso LOOCV n -veces (**¿Por qué?**) obtenemos los valores $MSE_1, MSE_2, \dots, MSE_n$ y el estimador para la **tasa de error de prueba** está dado por:

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^n MSE_i$$

Comparando este estimador con el obtenido por el método del conjunto de validación, anteriormente expuesto, vemos que :

- Tiene menos sesgo y, por tanto, no tiende a sobreestimar la **tasa de error de prueba**.

Validación cruzada dejando uno afuera:

Cuando repetimos el proceso LOOCV n -veces (**¿Por qué?**) obtenemos los valores $MSE_1, MSE_2, \dots, MSE_n$ y el estimador para la **tasa de error de prueba** está dado por:

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^n MSE_i$$

Comparando este estimador con el obtenido por el método del conjunto de validación, anteriormente expuesto, vemos que :

- Tiene menos sesgo y, por tanto, no tiende a sobreestimar la **tasa de error de prueba**.
- No hay aleatoriedad para la selección de submuestras de entrenamiento y prueba, llevando a resultados muy parecidos entre los distintos análisis de entrenamiento, a no ser que exista presencia de datos extremos.

Validación cruzada dejando uno afuera:

Cuando pensamos en las desventajas del método LOOCV, la primera que aparece es el “gasto” computacional debido a que hay que estimar n modelos y si n es grande el proceso requiere de mucho tiempo y recursos computacionales.

Validación cruzada dejando uno afuera:

Cuando pensamos en las desventajas del método LOOCV, la primera que aparece es el “gasto” computacional debido a que hay que estimar n modelos y si n es grande el proceso requiere de mucho tiempo y recursos computacionales.

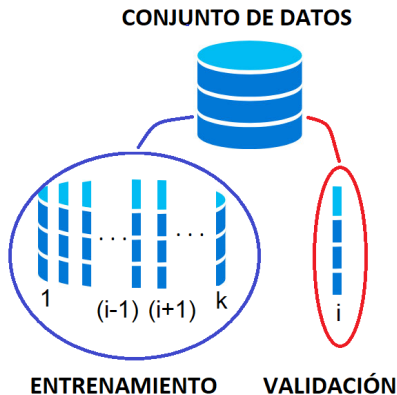
Sin embargo, cuando ajustamos un modelo de regresión lineal múltiple (**LA SIGUIENTE FÓRMULA SOLO FUNCIONA EN ESTE CASO, EN OTROS SE DEBEN AJUSTAR n MODELOS**), se puede probar que

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

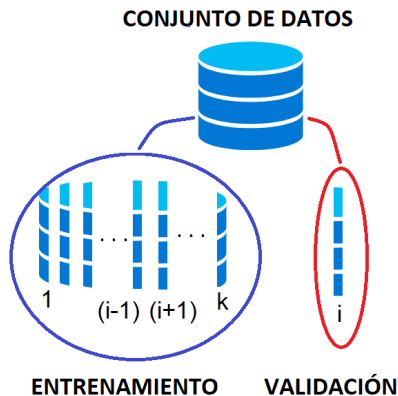
donde \hat{y}_i es el valor ajustado con el modelo de regresión lineal considerando todos los datos y h_i es el valor de influencia correspondiente al i -ésimo valor de la diagonal de la matriz “Hat”:

$$\mathbf{H} = \mathbf{X} \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top$$

Validación cruzada de k pliegues (k fold Cross-Validation)

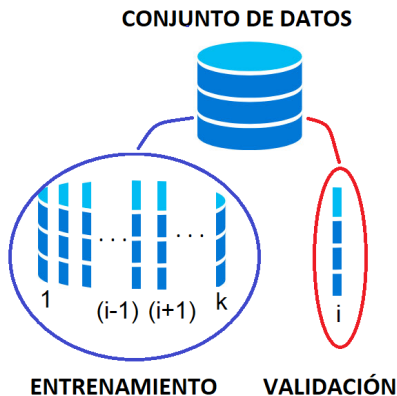


Validación cruzada de k pliegues (k fold Cross-Validation)



Los datos se dividen en k grupos o pliegues aproximadamente del mismo tamaño.

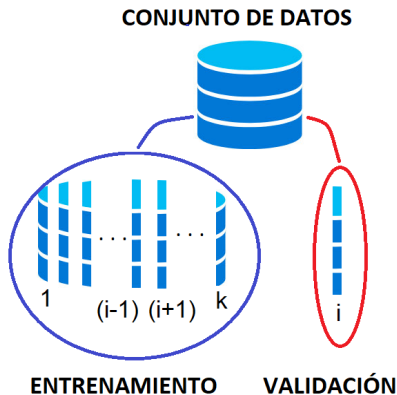
Validación cruzada de k pliegues (k fold Cross-Validation)



Los datos se dividen en k grupos o pliegues aproximadamente del mismo tamaño. Se selecciona el i -ésimo grupo ($i = 1, 2, \dots, k$) como conjunto de validación y se entrena el modelo con los datos de los $(k - 1)$ grupos restantes. Luego obtenemos el i -ésimo $MSE_{(k)i}$ y estimamos la **tasa de error de prueba** con:

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^k MSE_{(k)i}$$

Validación cruzada de k pliegues (k fold Cross-Validation)



Los datos se dividen en k grupos o pliegues aproximadamente del mismo tamaño. Se selecciona el i -ésimo grupo ($i = 1, 2, \dots, k$) como conjunto de validación y se entrena el modelo con los datos de los $(k - 1)$ grupos restantes. Luego obtenemos el i -ésimo $MSE_{(k)i}$ y estimamos la **tasa de error de prueba** con:

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^k MSE_{(k)i}$$

En la práctica se suele usar $k = 5$ o $k = 10$.

Compensación Sesgo-Varianza para k -fold CV:

Cuando tenemos una base de datos con n individuos y aplicamos:

Compensación Sesgo-Varianza para k -fold CV:

Cuando tenemos una base de datos con n individuos y aplicamos:

- **Método del conjunto de validación:** Obtenemos un único valor

$$MSE$$

- **LOOCV:** Obtenemos n valores de MSE

$$MSE_1, MSE_2, \dots, MSE_n \quad (\text{muy correlacionados positivamente})$$

- **k -fold:** Obtenemos k valores de MSE

$$MSE_{(k)1}, MSE_{(k)2}, \dots, MSE_{(k)k} \quad (\text{menos correlacionados})$$

Con respecto al sesgo:

- **Método del conjunto de validación:** Estima el error de prueba con un solo valor

$$MSE \longleftarrow \text{Tiene mucho sesgo}$$

- **LOOCV:** Estima el error de prueba con el promedio de n valores

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^n MSE_i \longleftarrow \text{Tiene poco sesgo}$$

- **k -fold:** Estima el error de prueba con el promedio de $k < n$ valores

$$CV_{(k)} = \frac{1}{k} \sum_{j=1}^k MSE_{(k)j} \longleftarrow \text{Tiene sesgo intermedio}$$

Con respecto a la varianza:

- **Método del conjunto de validación:** Tiene mucha varianza

$$\text{Var}(MSE)$$

- **LOOCV:** Tiene varianza intermedia, ya que los MSE_i ($i = 1, \dots, n$) son muy correlacionados y por tanto la covarianza es grande

$$\text{Var}(CV_{(n)}) = \text{Cov}(CV_{(n)}, CV_{(n)}) = \frac{1}{n^2} \sum_{i=1}^n \sum_{s=1}^n \underbrace{\text{Cov}(MSE_i, MSE_s)}_{\text{GRANDE para } i \neq s}$$

- **k-fold:** Tiene poca varianza, ya que los $MSE_{(k)j}$ ($j = 1, \dots, k$) no son muy correlacionados

$$\text{Var}(CV_{(k)}) = \text{Cov}(CV_{(k)}, CV_{(k)}) = \frac{1}{k^2} \sum_{j=1}^k \sum_{t=1}^k \underbrace{\text{Cov}(MSE_{(k)j}, MSE_{(k)t})}_{\text{PEQUEÑO para } j \neq t}$$

**CJTO. DE
VALIDACIÓN**

VERSUS

LOOCV

VERSUS

***k*-fold**

← + SESGO
+ VARIANZA
+ RÁPIDO

← - SESGO
+ VARIANZA
- RÁPIDO

← + SESGO
- VARIANZA
± RÁPIDO

Validación cruzada en problemas de clasificación:

Hasta ahora vimos como aplicar validación cruzada en problemas con variable respuesta Y **cuantitativa** y el método se basó en el *MSE*.

Validación cruzada en problemas de clasificación:

Hasta ahora vimos como aplicar validación cruzada en problemas con variable respuesta Y **cuantitativa** y el método se basó en el MSE .

Sin embargo, también es posible realizar validación cruzada cuando Y es **cualitativa** y en lugar de usar el MSE estimamos el error de prueba con la proporción de errores en el conjunto de prueba

$$Err = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} I(y_i \neq \hat{y}_i)$$

y donde

$$I(y_i \neq \hat{y}_i) = \begin{cases} 1, & \text{si } y_i \neq \hat{y}_i \\ 0, & \text{si } y_i = \hat{y}_i \end{cases}$$

Validación cruzada en problemas de clasificación:

Los métodos de validación cruzada estarían dados entonces por:

- **LOOCV:**

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^n Err_i$$

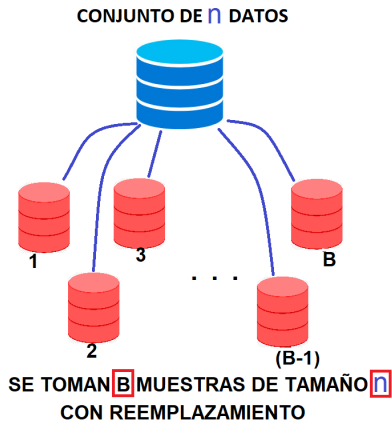
donde $Err_i = I(y_i \neq \hat{y}_i)$.

- **k-fold:**

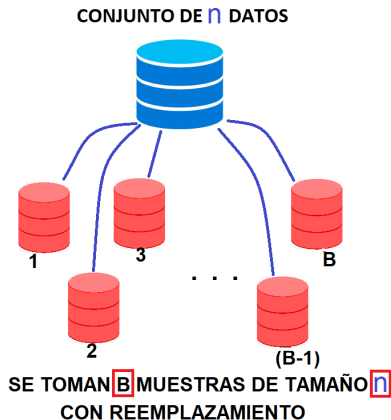
$$CV_{(k)} = \frac{1}{k} \sum_{j=1}^k Err_{(k)j}$$

donde $Err_{(k)j}$ es la proporción de errores en el j -ésimo pliegue.

El Bootstrap:



El Bootstrap:



Para una base de datos de n individuos, se hace un submuestreo con reemplazamiento para obtener B muestras cada una de tamaño n . En este caso, la misma observación puede estar dos o más veces en cada submuestra.

En cada caso se estiman los parámetros del modelo que se plantee, **por ejemplo**, si uno de ellos es α , el **objetivo** es estimar el error estándar del estimador $\hat{\alpha}$, dado por:

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

donde $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \dots, \hat{\alpha}^{*B}$ corresponden a los parámetros estimados con las B bases de datos obtenidas con Bootstrap.

CONJUNTO DE n DATOS

