3ª
Emisión

Módulo 03

TEMAS SELECTOS DE ESTADÍSTICA

Dr. Roberto Bárcenas Curtis

DGTIC
UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
Dirección General de Cómputo y de Tecnologías de información y Comunicación
Dirección de Docencia en TIC
50 AÑOS
Educación Continua
1971 - 2021

Presentación

Dada la base de conocimientos se consideran algunos tópicos especiales para complementar sus estudios de probabilidad y estadística en el campo de la Ciencia de Datos.

Objetivo

El participante aplicará los conceptos esenciales adquiridos para extenderlos hacia algunas herramientas estadísticas actuales.

Temas

5 TEMAS SELECTOS DE ESTADÍSTICA

- 5.1 Análisis de Regresión
- 5.2 Muestreo
- 5.3 Técnicas Bootstrap

Covarianza

Sean X y Y variables aleatorias cuyo valor esperado $E(X)$, $E(Y)$ respectivamente, es finito.

La covarianza entre X y Y, denotada como $Cov(X,Y)$, se define como

$$Cov(X,Y) = E\left[(X - E(X))(Y - E(Y))\right].$$

Como propiedades formales, la covarianza es simétrica y bilineal. Además, usando la linealidad y desarrollando la expresión, se obtiene la relación fundamental equivalente de la covarianza como

$$Cov(X,Y) = E(XY) - E(X)E(Y).$$

Propiedades

Considerando el valor α como constante, la covarianza tiene las siguientes propiedades.

- $Cov(X,Y) = E(XY) - E(X)E(Y)$.
- $Cov(X,Y) = Cov(Y,X)$.
- $Cov(X,X) = Var(X)$.
- $Cov(\alpha X,Y) = \alpha Cov(X,Y)$.
- $Cov(\alpha,Y) = 0$.
- $Cov(X,Y_1 + Y_2) = Cov(X,Y_1) + Cov(X,Y_2)$.

Coeficiente de correlación

El coeficiente de correlación de las variables aleatorias X y Y, denotado por $\rho(X,Y)$, se define como

$$\rho(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Se trata de la cuantificación de la relación lineal entre esas dos variables, la cual puede ser positiva o negativa.

Observaciones

Notemos que $\rho(X,Y)=0$ sucede, si y sólo si $Cov(X,Y)=0$. Esto es, si dos variables aleatorias son independientes, entonces, no poseen covarianza y, por lo tanto, tampoco estarán correlacionadas.

El hecho que dos variables sean independientes, implica que su coeficiente de correlación es cero. En cambio, el hecho de que dos variables aleatorias no estén correlacionadas i.e., su correlación es $\rho(X,Y)=0$, no quiere decir que sean independientes.

Tiene las siguientes propiedades:

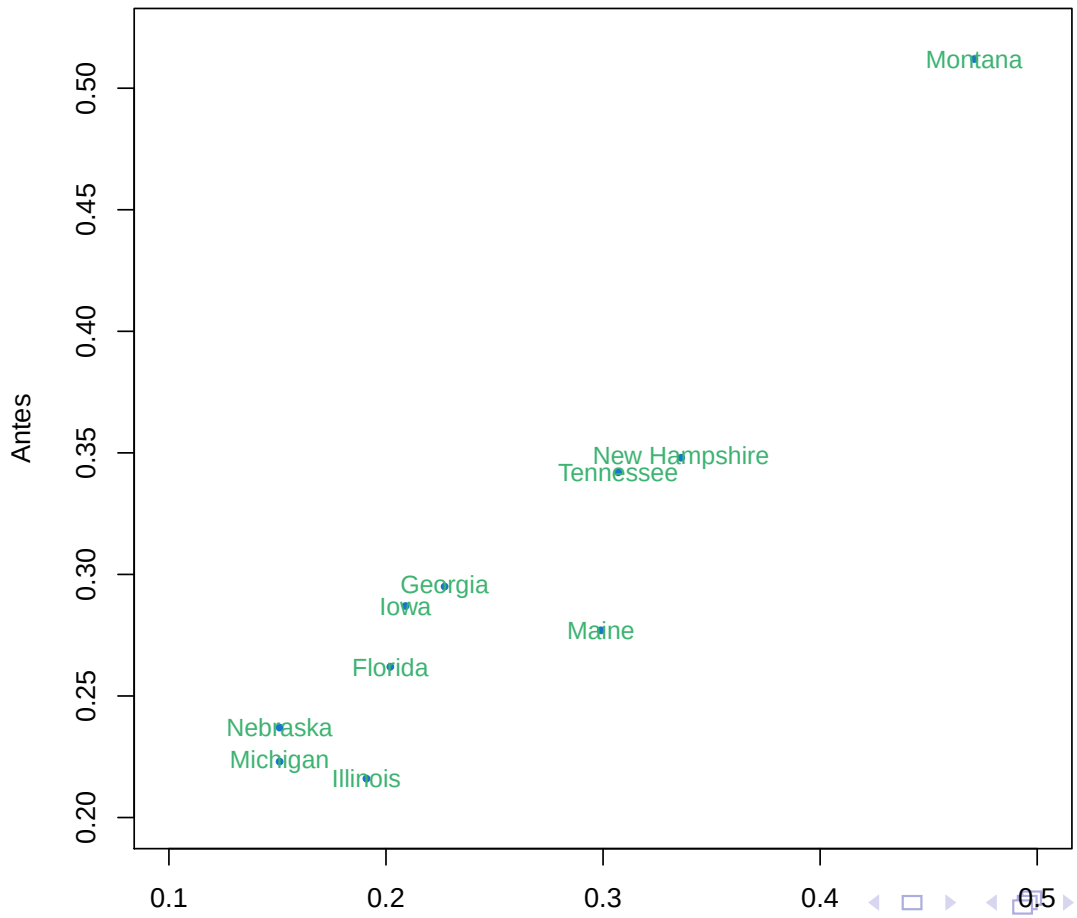
- $-1 \leq \rho(X,Y) \leq 1$.
- Si X y Y son independientes, entonces $\rho(X,Y) = 0$.

Entonces, cuando $\rho(X,Y) = 0$, se dice que las variables aleatorias X, Y son no correlacionadas.

En cambio, si $\rho(X,Y) = 1$, se dice las variables aleatorias están completamente correlacionadas positivamente, y viceversa, si $\rho(X,Y) = -1$, se dice que están perfectamente correlacionadas negativamente.

Coeficiente de Correlación

Hoskin et al. (1986) investigaron las incidencias de accidentes fatales de vehículos despues de que se incrementó la edad mínima que permitía a los jóvenes beber en 10 estados de Estados Unidos.



Correlación muestral

$$R = \begin{pmatrix} 1 & r_{12} & r_{13} & \dots & r_{1p} \\ r_{21} & 1 & r_{23} & \dots & r_{2p} \\ r_{31} & r_{32} & 1 & \dots & r_{3p} \\ \vdots & \vdots & \vdots & \dots & \dots \\ r_{p1} & r_{p2} & r_{p3} & \dots & 1 \end{pmatrix}$$

donde

$$r_{jk} = \frac{s_{jk}}{s_j s_k} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \cdot \sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}}$$

es la correlación entre las variables x_j y x_k

Correlación entre variables

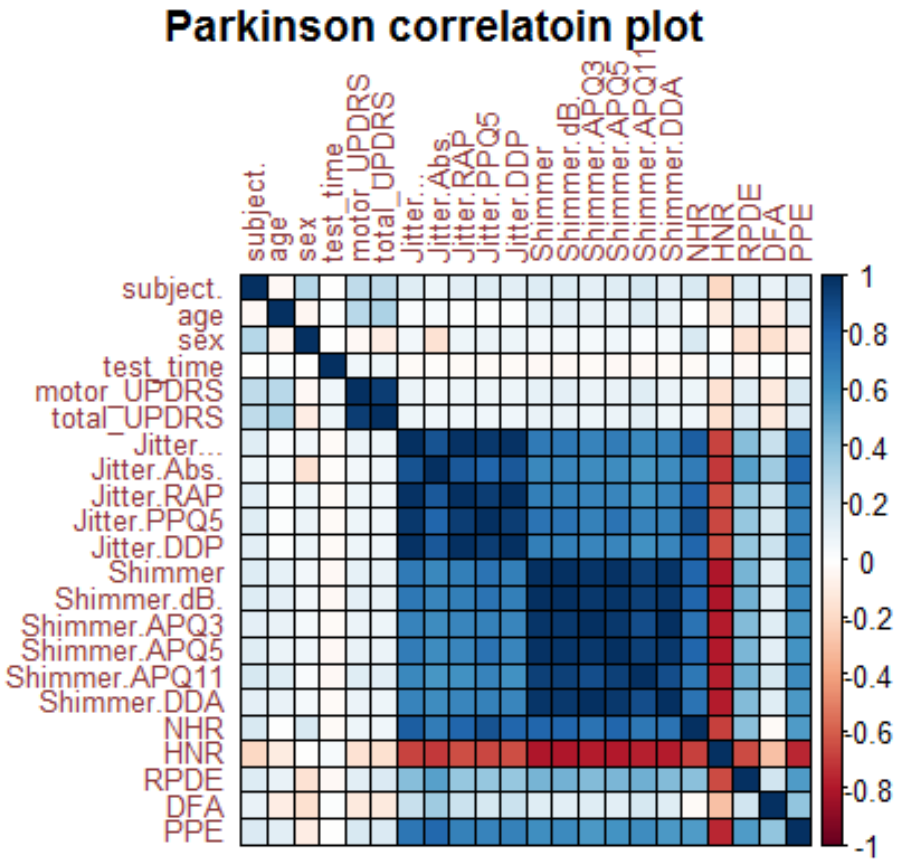


Figure 2: Gráfica de correlación

Regresión

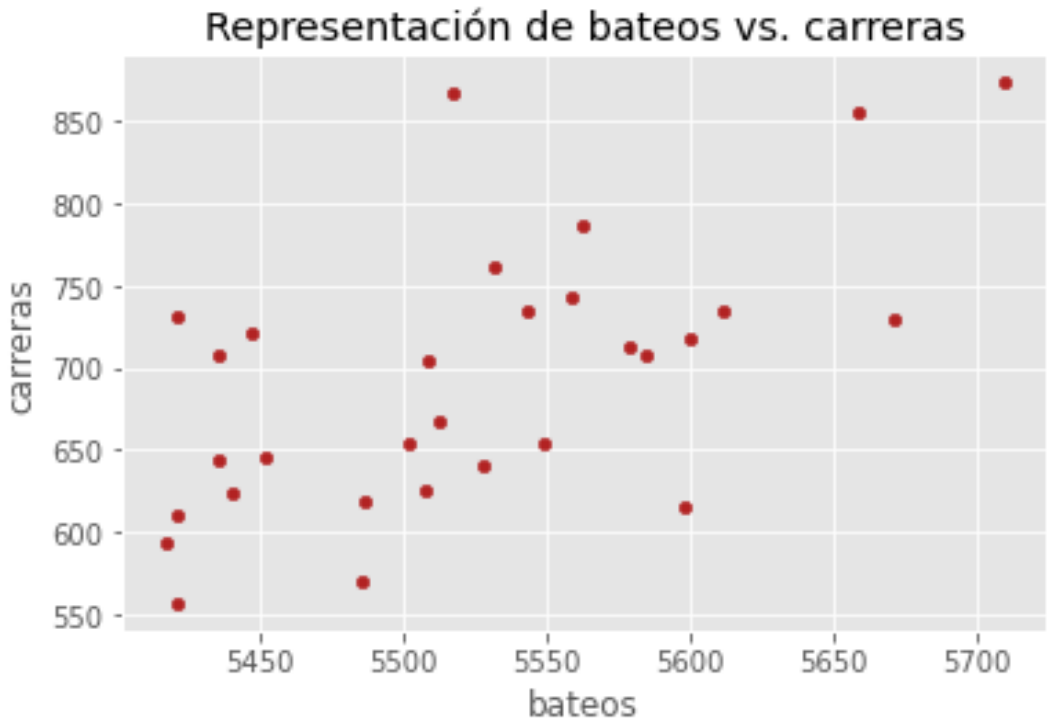
- Predecir la respuesta Y para un conjunto de covariables X dado.
- Modelar (de alguna forma) la relación entre la entrada X y la salida de un sistema Y.
- Evaluar y comparar el impacto de diferentes covariables X sobre la respuesta Y.

- **Independiente** X: Otros nombres alternativos para esta son variable explicativa, variable predictora y en ocasiones variable regresora.

- **Dependiente** Y: es llamada de diferentes maneras, algunas de ellas: variable respuesta, variable explicada o variable pronosticada.

Regresión

- ¿Es significativo el efecto que una variable (explicativa) X tiene sobre otra variable (respuesta) Y?
- ¿Es significativa la dependencia lineal entre esas dos variables?



Regresión lineal múltiple

El modelo de regresión lineal múltiple se representa como

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon, \tag{1}$$

donde $\beta_j,\ \ j = 0, \ldots, k$ son parámetros desconocidos; en particular β_0 es llamado intercepto.

Notar que usamos el índice k para indicar el número de variables predictoras, lo que significa que tenemos $k + 1$ parámetros de regresión (los coeficientes β), y en esta representación, ϵ es el término de error.

Interpretación

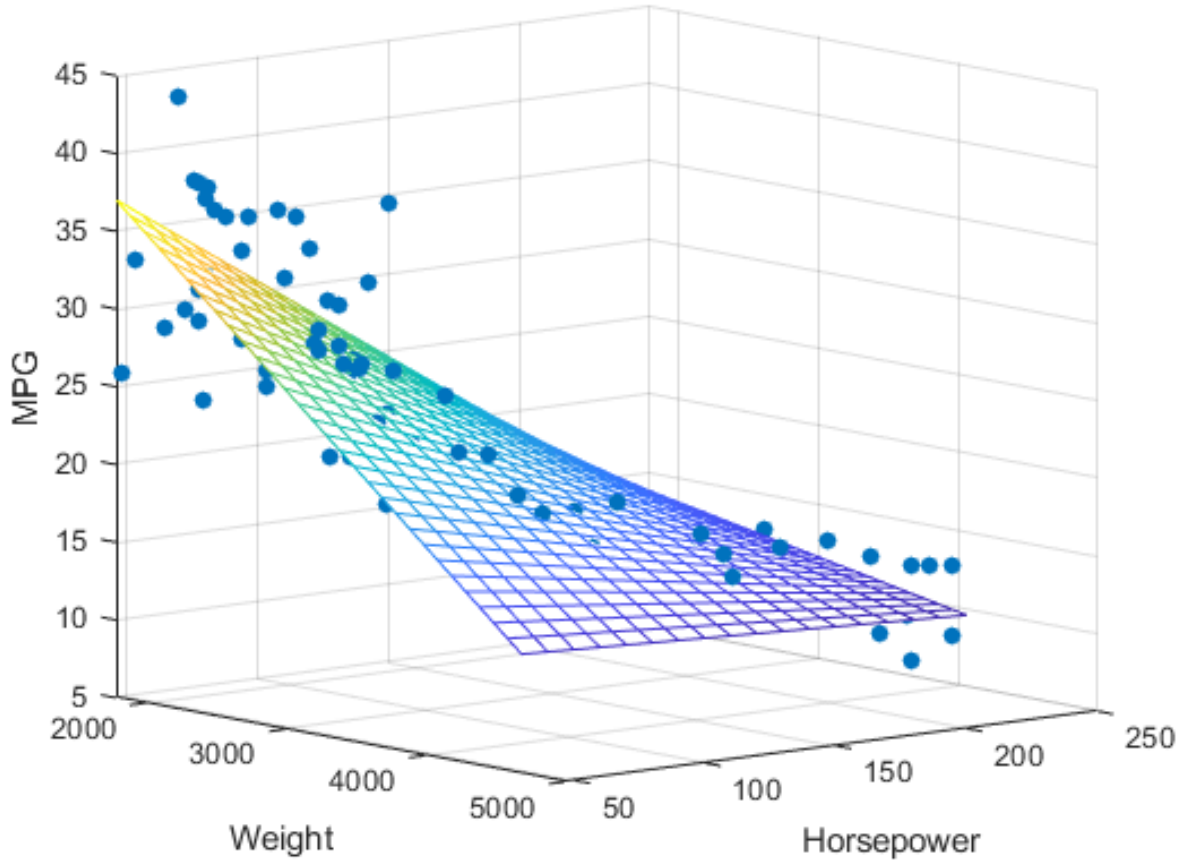
Suponiendo un modelo que para la observación i-ésima es

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \quad i = 1, \ldots, n \qquad (2)$$

Cada coeficiente β en (2), representa el cambio en la respuesta media, por unidad de aumento en la variable predictiva asociada cuando todos los otros predictores se mantienen constantes.

Por ejemplo, β_1 representa el cambio en la respuesta media por unidad de aumento en x_1, cuando x_2, x_3 están fijos. El término β_0 o intercepto, representa la respuesta media, cuando los predictores x_1, x_2, x_3, son todos cero (que puede tener o no un significado práctico).

Geométricamente



Notación matricial

En general, el modelo de regresión es de la forma:

$$y = X\beta + \epsilon, \tag{3}$$

donde Y un vector columna $n \times 1$, X es una matriz $n \times (k+1)$ dada por:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}.$$

En este caso, β es un vector columna $(k+1) \times 1$ que contiene los parámetros y ϵ también es un vector columna de errores de dimensión $n \times 1$. El modelo (3) extendido sería:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Podemos formular la función de regresión lineal en notación matricial:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$$

Es decir, en lugar de escribir las n ecuaciones, utilizando la notación matricial, nuestra función de regresión lineal simple se reduce a una expresión más manejable.

Supuestos

Las condiciones sobre ϵ_i y y_i se pueden expresar en términos del modelo descrito en (3):

1. $\mathbb{E}(\epsilon) = 0$, implica $\mathbb{E}(y) = X\beta$.
2. $Cov(\epsilon) = \sigma^2 I$, es decir, $Cov(y) = \sigma^2 I$, donde I denota a la matriz identidad, de dimensión conforme al producto resultante al obtener la covarianza. Notar que este supuesto, engloba las condiciones $\mathrm{Var}(\epsilon_i) = \sigma^2$ y $Cov(\epsilon_i, \epsilon_j) = 0$.

Dado que el modelo es desconocido, buscamos los parámetros óptimos β_0 y β_1 que mejor se ajusten a nuestros datos.

Estimación

La estimación de los parámetros se puede hacer:

- Por el método de mínimos cuadrados buscando minimizar la suma de cuadrados de las n diferencias entre los observados y y los predichos $\hat{y}$.
- Vía máxima verosimilitud, usando la distribución de las observaciones en una función objetivo.

En ambos casos, se obtiene el estimador:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = (X^T X)^{-1} X y. \tag{4}$$

Predicción

El vector de valores ajustados $\hat{y}$ puede expresarse como

$$\hat{y} = X\hat{\beta} = X\left(X^T X\right)^{-1} y = Hy \tag{5}$$

donde se define la matriz $n \times n$, $H = X\left(X^T X\right)^{-1}$, llamada **hat matrix**.

Interpretación de H. Esta matriz es la proyección ortogonal de y en el espacio generado por las columnas de X. Mapea el vector de observados y al vector de ajustados $\hat{y}$ que están en el hiperplano de regresión.

Se define la suma de cuadrados de los residuales o suma de cuadrados del error, como

$$\begin{aligned} SCE &= \hat{\epsilon}^T \hat{\epsilon} = (y - X\hat{\beta})^T (y - X\hat{\beta}) \\ &= y^T (I - H)^T (I - H) y = y^T (I - H) y \\ &= y^T \left[I - X(X^T X)^{-1} X^T \right] y \end{aligned} \tag{6}$$

A través de la suma de cuadrados del error, podemos determinar un estimador insesgado para σ^2, basado en el estimador $\hat{\beta}$.

Para estimar σ^2 por su contraparte muestral, ses susituye β por $\hat{\beta}$,

$$
\begin{aligned}
s^2 \doteq \text{ECM} &= \frac{1}{n-p}\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2 \\
&= \frac{1}{n-(k+1)}(y - X\hat{\beta})^T(y - X\hat{\beta}) \\
&= \frac{y^Ty - \hat{\beta}X^Ty}{n-k-1} = \frac{SCE}{n-k-1}
\end{aligned}
\tag{7}
$$

El error cuadrático medio (ECM) o también denotado como s^2, es un estimador de la varianza de los errores σ^2. A la raíz cuadrada del error cuadrático medio, se le conoce como el error estándar de regresión o el error estándar residual y se trata del estimador de σ y

Análisis de varianza

Una descomposición importante es la siguiente:

$$\sum_{i=1}^{n}(y_i - \bar{y}) = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y})^2 \tag{8}$$

$$\text{SCT} = \quad \text{SCR} \quad + \quad \text{SCE}$$

- SCT: suma de cuadrados (total) respecto a la media.
- SCR: suma de cuadrados de la regresión.
- SCE: suma de cuadrados del error (o residuos).

La primera se interpreta como la variabilidad total, mientras que la segunda es la variabilidad explicada por la regresión y el último término, se conoce como variabilidad no explicada.

Análisis de varianza

La tabla de análisis de varianza (ANOVA) se utiliza para cuantificar la significancia de la regresión.

Fuente de Variación	Grados de libertad	Suma de Cuadrados	Media cuadrática	F	p-valor
Regresión	k	SCR	$CMR = SCR/k$	$F = \frac{CMR}{ECM}$	$P\left(F^{\alpha}_{k,n-k-1} > F\right)$
Error	$n-k-1$	SCE	$ECM = \frac{SCE}{n-k-1}$		
Total	$n-1$	SCT			

Coeficiente de determinación

El coeficiente de determinación (muestral) cuantifica el nivel del ajuste de la regresión.

Se calcula como

$$R^2 = \frac{SCR}{SCT} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{9}$$

Una medida alternativa, es el R^2 ajustado, el cual considera la situación donde se agregan más predictores:

$$R^2_{ajustado} = 1 - \left(\frac{n-1}{n-(k+1)}\right)(1-R^2). \tag{10}$$

Intervalos de confianza

Es posible mostrar que $\hat{\beta} \sim N(\beta, \sigma^2(X^T X)^{-1})$. De donde, si se reemplaza σ^2 por su estimador s^2, cada uno de los estadísticos

$$\frac{\hat{\beta} - \beta_j}{se(\hat{\beta}_j)}, \quad j = 1, \ldots, k, \tag{11}$$

tiene una distribución t de Student con $n - p$ grados de libertad, donde $se(\hat{\beta}_j)$ es el error estándar del coeficiente (ajustado) $\hat{\beta}_j$.

Por lo tanto, un intervalo de confianza al $100(1 - \alpha)\%$ para el coeficiente de regresión β_j, $j = 0, 1, \ldots, k$ es:

$$\hat{\beta}_j \pm t_{\alpha/2, n-p} \cdot se(\hat{\beta}_j) \tag{12}$$

Aprendizaje e hipótesis

Podría decirse que existen dos maneras de aprender de los datos sobre los parámetros de la población:

1. Estimación puntual + intervalos de confianza.
2. Prueba de hipótesis.

Hipótesis estadística

Es una aseveración sobre la población, usualmente basada en un parámetro.

Aprendizaje e hipótesis

Podría decirse que existen dos maneras de aprender de los datos sobre los parámetros de la población:

1. Estimación puntual + intervalos de confianza.
2. Prueba de hipótesis.

Hipótesis estadística

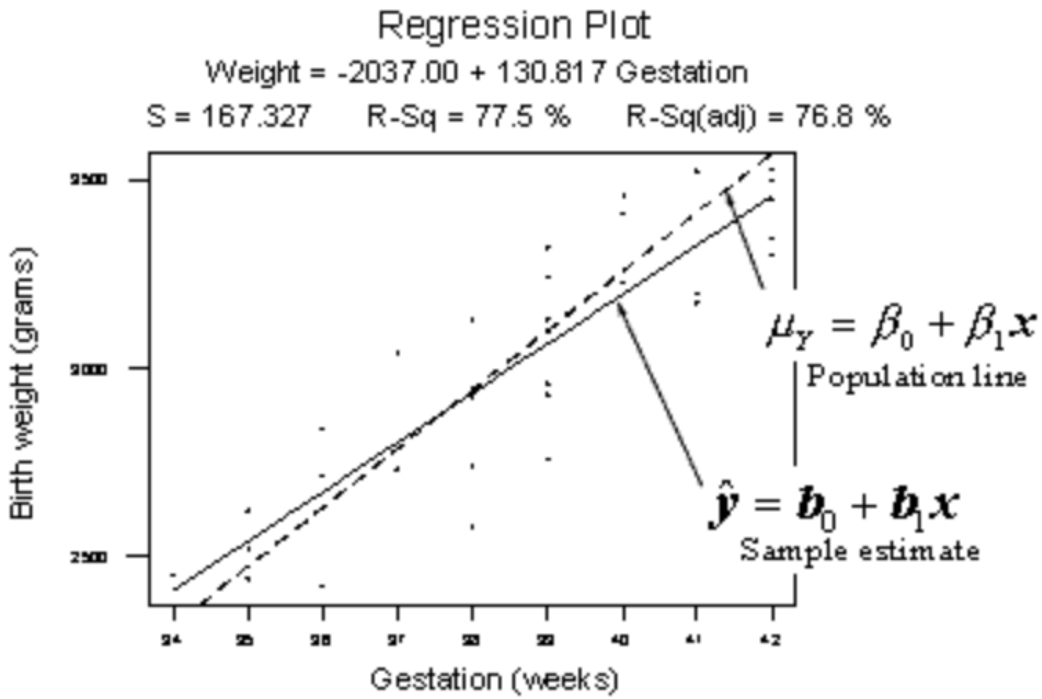
Es una aseveración sobre la población, usualmente basada en un parámetro.

Contraste de hipótesis: tratar de establecer conclusiones acerca del valor del parámetro.

$\rightarrow$ A través de la información (datos) de una muestra recabar evidencia para dar una resolución acerca de la hipótesis.

Ejemplos

- Pruebas de media (Z y t) y diferencia de medias.
- Pruebas de una proporción (p) y diferencia de proporciones.
- Pruebas acerca de la varianza y de cociente de varianzas.
- Igualdad de medias (ANOVA).
- Pruebas sobre los parámetros en Análisis de Regresión.



Prueba de hipótesis

Elementos

- Hipótesis nula H_0 vs. Hipótesis alternativa H_1.
- Estadística de prueba $T(x)$: es una función de la muestra. Bajo el supuesto H_0 (verdadera) tiene una distribución conocida.
- Nivel de significancia α: es la probabilidad de rechazar H_0 cuando esta es cierta.
- Región crítica C_α: permite establecer una zona de decisión. Si $T(x) \in C_\alpha$, implica rechazar H_0 con un nivel de significancia asociado α.

Tipos de error

- Error de tipo I: La hipótesis nula se rechaza siendo verdadera.
- Error de tipo II: La hipótesis nula no se rechaza y es falsa.

	H_0 falsa	H_0 cierta
Rechazar	✓	Error Tipo I
No rechazar	Error Tipo II	✓

Notar que $\alpha = P(\text{Error Tipo I})$.

Regla de decisión

En una prueba de hipótesis, tomar la decisión se reduce a determinar qué tan "probable" o "improbable" es la estadístico de prueba *observada*

Si es probable, no rechazamos la hipótesis nula. Si es poco probable, rechazamos la hipótesis nula a favor de no descartar la hipótesis alternativa.

Usualmente, hay dos formas de determinar si la prueba es probable o improbable dada la hipótesis (nula) inicial:

- Adoptar el enfoque del *valor crítico*.
- O bien, considerar el p-valor de la prueba.

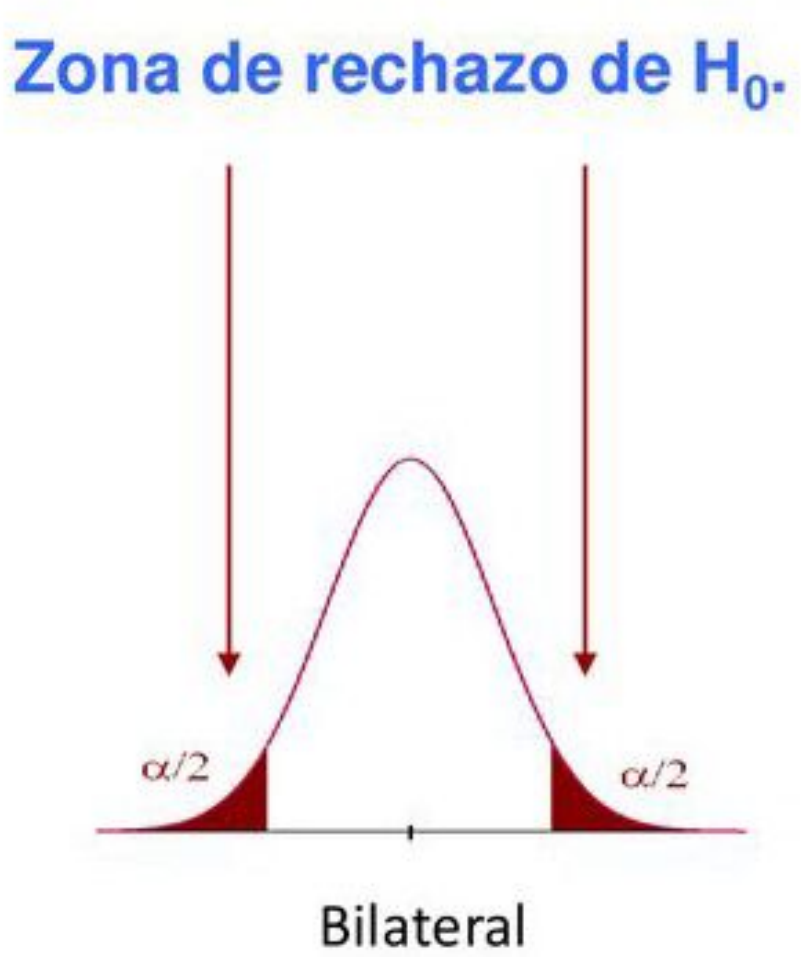
Tipos de hipótesis

- De dos colas

Nula
$H_0 : \theta = \theta_0$

vs.

Alternativa
$H_0 : \theta \neq \theta_0$



Tipos de hipótesis

- De cola derecha

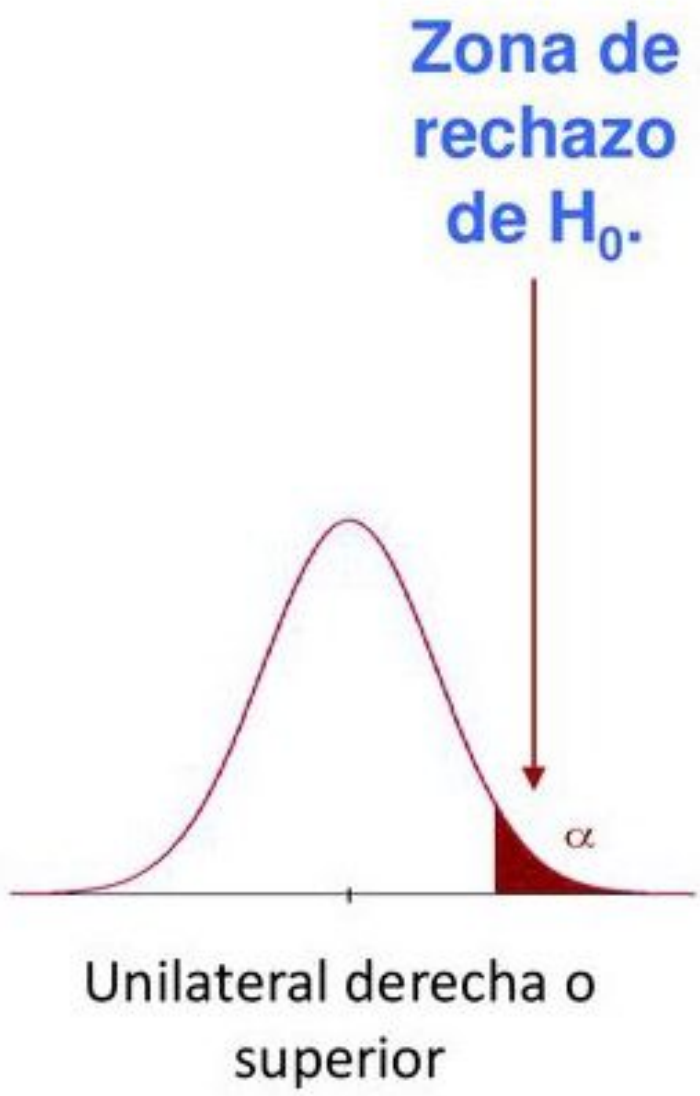
Nula

$H_0 : \theta = \theta_0$

vs.

Alternativa

$H_0 : \theta > \theta_0$



Tipos de hipótesis

- De cola izquierda

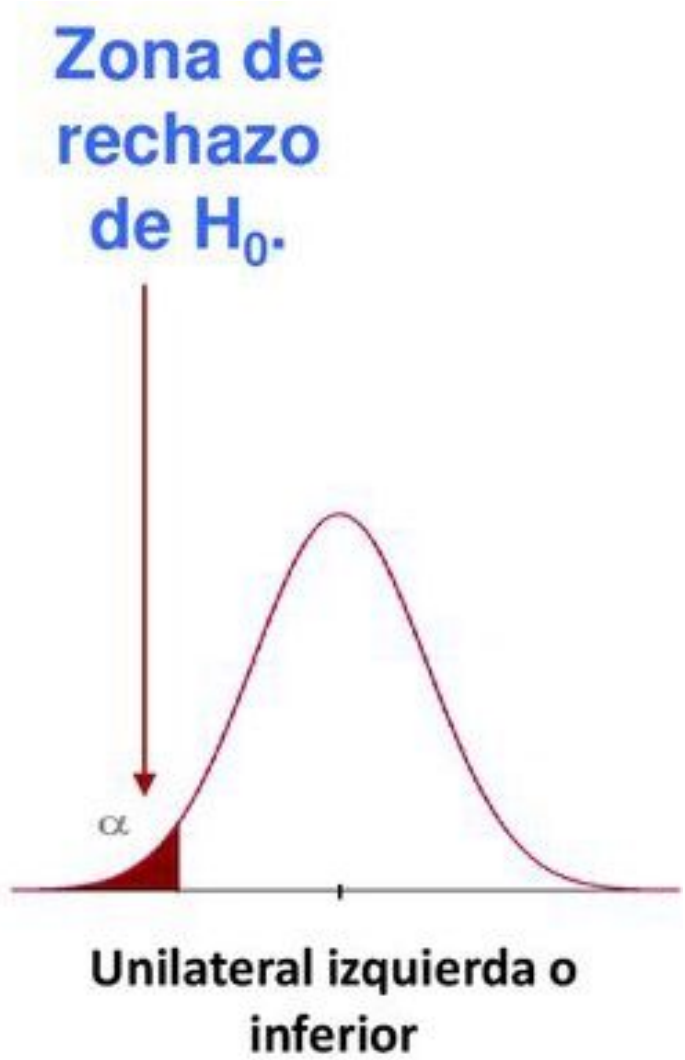
Nula

$H_0 : \theta = \theta_0$

vs.

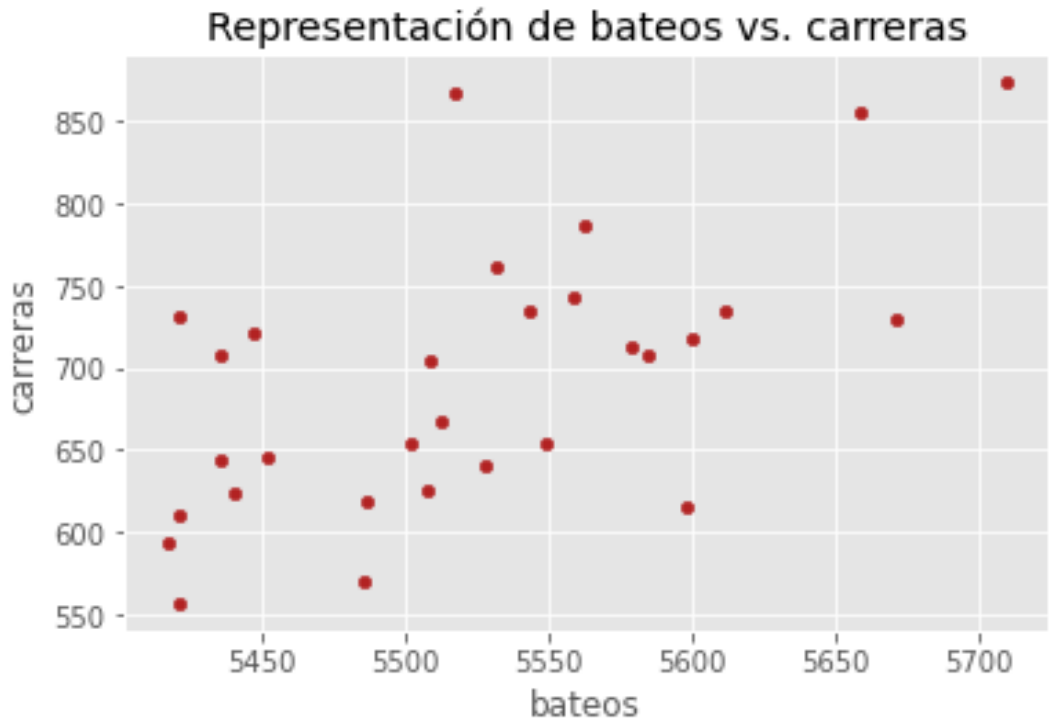
Alternativa

$H_0 : \theta < \theta_0$



Ejemplo

La siguiente gráfica representa la relación entre bateos y carreras en equipos de béisbol en Estados Unidos.



Ajuste de regresión

Modelo: $y = \beta_0 + \beta_1 x + \epsilon$.

- Output

Coefficients	Estimate	t	$P(T > \lvert t \rvert)$
(Intercept)	-2367.7 (1066.35)	-2.220	0.037^{*}
Gestation	0.55 (0.193)	2.862	0.009^{**}
$^{*}p - value < 0.05$; $^{**}p - value < 0.01$; $^{***}p - value < 0.001$			
$R^2 = 0.271$	Adj. $R^2 = 0.238$		
F-statistic: 8.191 on 1 and 28 DF, $p - value$: 0.00906^{**}			

Modelo ajustado

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = \text{-}2367.7 + 0.55x$$

Pasos

1. Especificar las hipótesis: nula (H_0) y alternativa (H_1).
2. Calcular la estadística de prueba, misma que bajo H_0 tiene una distribución conocida.
3. Para un nivel de significancia α preestablecido, determinar la región de rechazo.

Rechazar H_0 si:

- La estadística cae en la región de rechazo, o
- $p - valor \leq \alpha$.

4. Tomar una decisión con un nivel de significancia α.

Hipótesis sobre β's

1. $H_0 : \beta_1 = 0$ vs. $\quad H_1 : \beta_1 \neq 0$
2. Estadística de prueba $t = \dfrac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{n-2}$. En particular, tenemos que para β_1, $t^* = 2.862$.
3. Para un nivel de significancia α=0.05, la regla de decisión es:

Rechazar H_0 si

- Sucede que
$$|t^*| \geq t_{1-\alpha/2,n-2} = 2.05, \text{ o}$$
- $p - valor = P(t_{1-\alpha/2,n-2} > |t^*|) \leq \alpha = 0.5.$

4. **Conclusión**:Dado que $t^* = 2.862 > 2.05$ y el $p - valor = 0.009 < 0.05$, la decisión es **rechazar** la hipótesis nula H_0 con un nivel de significancia del 5%.

ANOVA: Significancia de la regresión

1. $H_0 : \beta_j = 0, \forall j$ vs. $H_1 : \beta_j \neq 0,$ para alguna j
2. Estadística de prueba

$$F = \frac{SCR}{SCE/(n-2)} =\sim F_{1,n-2}.$$

   En particular, tenemos que $F^* = 8.191$.
3. Para un nivel de significancia α=0.05, la regla de decisión es:

Rechazar H_0 si

- Sucede que

$$F^* > F^{1-\alpha}_{1,n-2} = 4.196, \text{ o}$$

- $p - valor = P(F^{1-\alpha > F^*}_{1,n-2}) \leq \alpha = 0.5.$

4. **Conclusión**: Como $F^* = 8.191 >$ 4.196 y además, p-valor $= 0.00906 < 0.05$. Se **rechaza** la hipótesis nula H_0 con un nivel de significancia del 5%.

Del ejemplo

Dada la información:

- $t^* = -2.220$
- p-valor$= 0.037$

Realizar la prueba de hipótesis para el coeficiente de la regresión β_0 que representa el intercepto.

Usar una significancia α =0.05.

Regresión Ridge

Consideremos el modelo estándar de regresión, $y = X\beta + e$, con $e \sim N_(0, \sigma^2 I)$ y consideremos el siguiente problema de estimación restringida:

$$\min_{\beta} \quad (y - X\beta)^T (y - X\beta) \qquad \text{s.a.} \quad ||\beta||^2 \leq c$$

El Lagrangiano es:

$$(y - X\beta)^T (y - X\beta) + \lambda(||\beta||^2 - c)$$

Derivando con respecto a β e igualando a cero, puede verse que:

$$\widehat{\beta}_R = (X^T X + \lambda I)^{-1} X^T y$$

es la solución del problema restringido (estimador Ridge).

La colinealidad ocasiona problemas numéricos e incremento en la varianza, así como en la magnitud de los estimadores de mínimos cuadrados.

Una vez que planteamos el problema de minimización restringida, sujeta a $||\beta||^2 \leq c$, es natural considerar, por ejemplo

$$\min_{\beta} \quad (y - X\beta)^T(y - X\beta) \qquad \text{s.a.} \quad ||\beta||_1^2 = \sum_{j=1}^{p} |\beta_j| \leq c$$

La solución a este problema da lugar a la Regresión Lasso.

Ridge y Lasso son casos especiales de "Regresión Penalizada" en donde minimizamos expresiones de la forma:

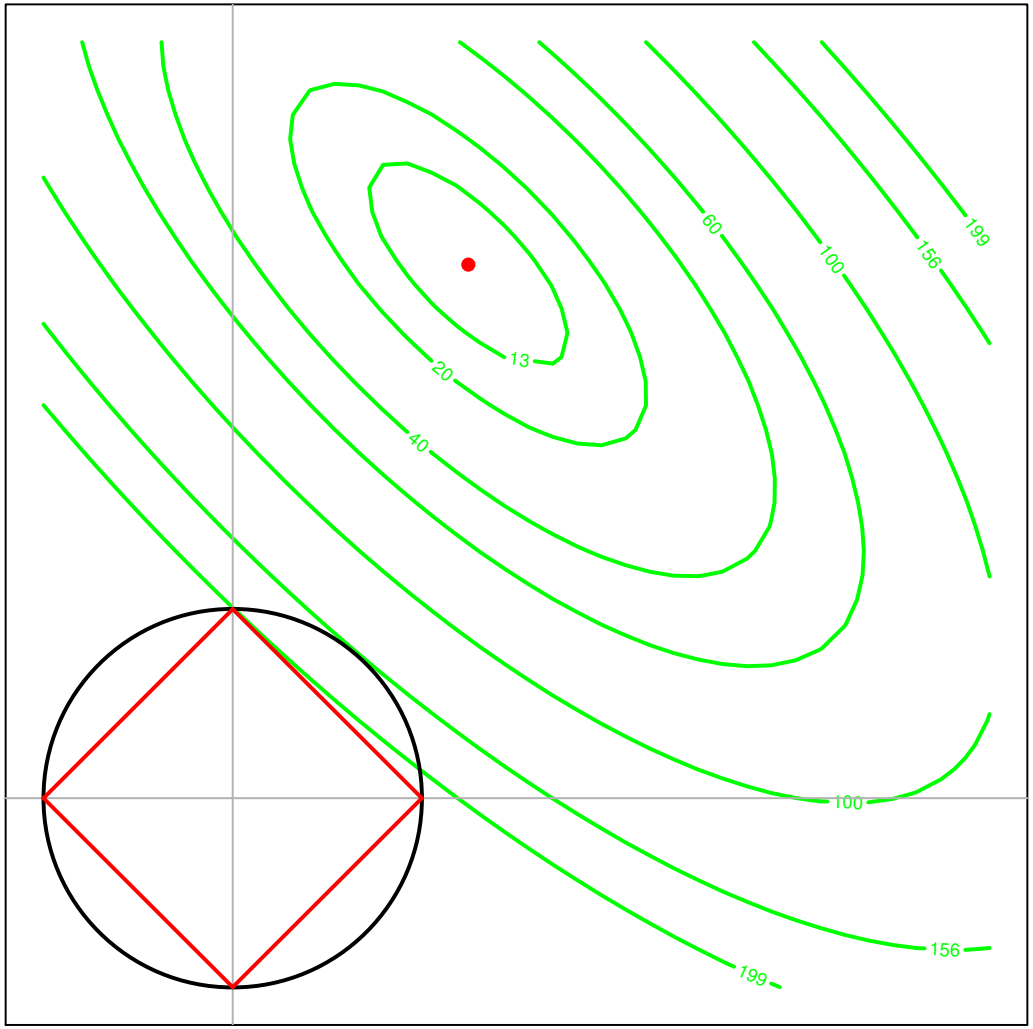
$$\text{Ajuste} + \text{Penalización}$$

(los ajustes de splines pueden verse como esta clase de problemas).

Una propiedad importante de Lasso es que, además de no permitir valores demasiado grandes (minimizando con ello problemas de colinealidad), también actúa como un procedimiento de selección de variables. La siguiente gráfica ilustra como Lasso elimina una variable y Ridge no.

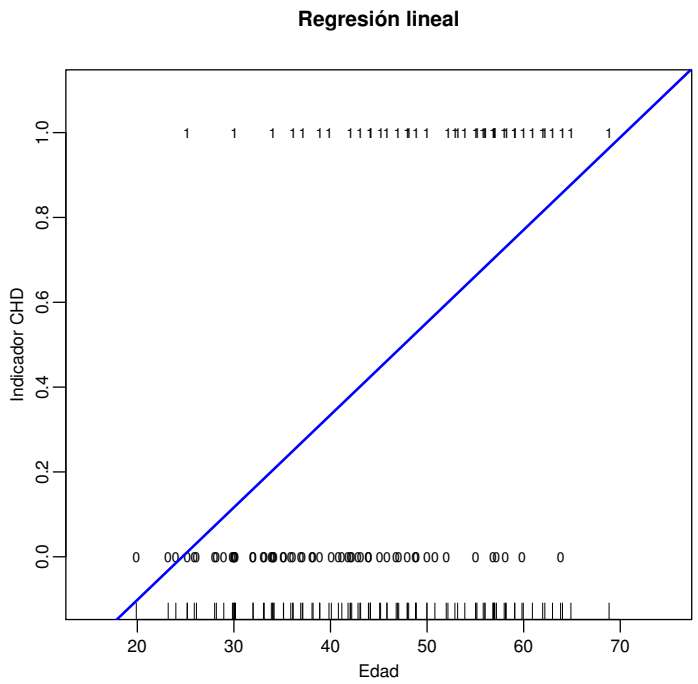
Restricciones Ridge y Lasso

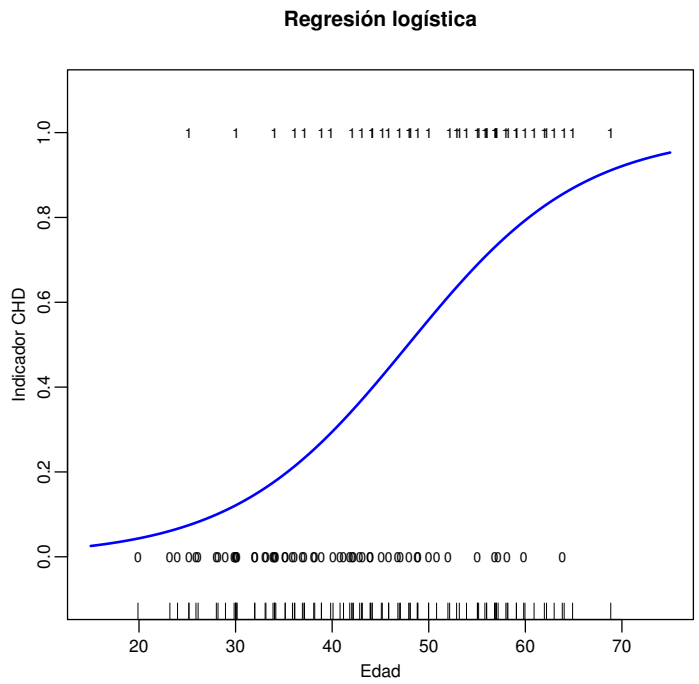
Minimización de SCE, sujeta a restricciones Ridge y Lasso



Regresión logística

Datos de edad (en años) e indicadores de presencia o ausencia de daño significativo en la coronaria (CHD).





Regresión logística

En regresión lineal modelamos el comportamiento medio de una variable de interés (variable de respuesta) como función de covariables

$$\mathrm{E}(y) = \beta_0 + \beta_1 z_1 + \cdots + \beta_k z_k$$

En regresión logística (esto es, cuando la respuesta es binaria) también se modela la media como función de covariables

$$\mathrm{E}(y) = h(\beta_0 + \beta_1 z_1 + \cdots + \beta_k z_k) = h(x^T \beta)$$

o, de forma equivalente

$$g(\ \mathrm{E}(y)\ ) = x^T \beta$$

Por ser y una variable binaria, entonces tenemos que sus dos posibles valores los toma con probabilidades:

$$P(y = 1) = p \qquad \text{y} \qquad P(y = 0) = 1 - p$$

Usamos que la media de una Bernoulli es $\mathrm{E(y)} = p$. Una forma muy usada (aparte de que es sensata) de modelar la dependencia de p sobre covariables es mediante la función logit:

$$\text{logit}(p) \;=\; \log\left(\frac{p}{1-p}\right) \;=\; x^T\beta$$

o (despejando para p), mediante la función logística:

$$p = E(y|x) = \frac{1}{1 + \exp(-x^T\beta)}$$

Este es el llamado Modelo de Regresión Logística.

Aprendizaje estadístico

La finalidad de un modelo de aprendizaje es predecir la variable respuesta en observaciones futuras o en observaciones (nuevas) que el modelo no ha “visto” antes.

- Conjunto de entrenamiento (training set): datos/observaciones con las que se entrena el modelo.
- Conjunto de validación (validation set): datos/observaciones del mismo tipo que las que forman el conjunto de entrenamiento pero que no se han empleado en la creación del modelo.
- Conjunto de prueba (test set): Son datos nuevos que el modelo no ha “visto”.

Errores

El error mostrado por defecto tras entrenar un modelo suele ser el error de entrenamiento, el error que comete el modelo al predecir las observaciones nuevas es el error de prueba o de generalización.

Para conseguir una estimación más certera del error, se tiene que recurrir a uno o varios conjuntos de prueba y emplear estrategias de validación basadas en remuestreo.

Técnicas de remuestreo

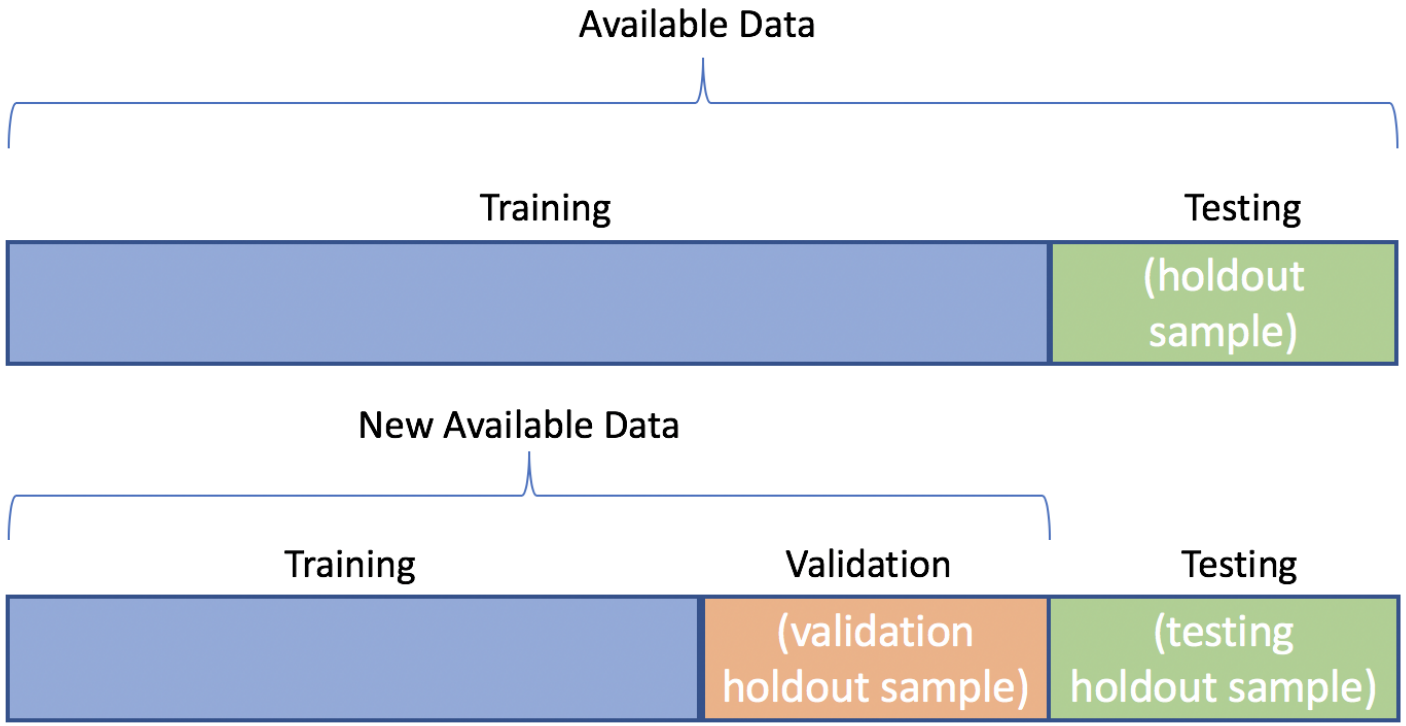
Son estrategias que permiten estimar la capacidad predictiva de los modelos cuando se aplican a nuevas observaciones, haciendo uso únicamente de los datos de entrenamiento.

Idea: el modelo se ajusta empleando un *subconjunto (*elegido aleatoriamente) de observaciones del conjunto de entrenamiento y se evalúa con las observaciones restantes. Este proceso se repite múltiples veces, los resultados se agregan y promedian.

Las repeticiones permiten compensar las posibles desviaciones que puedan surgir por la selección aleatorio de las observaciones.

Validación simple

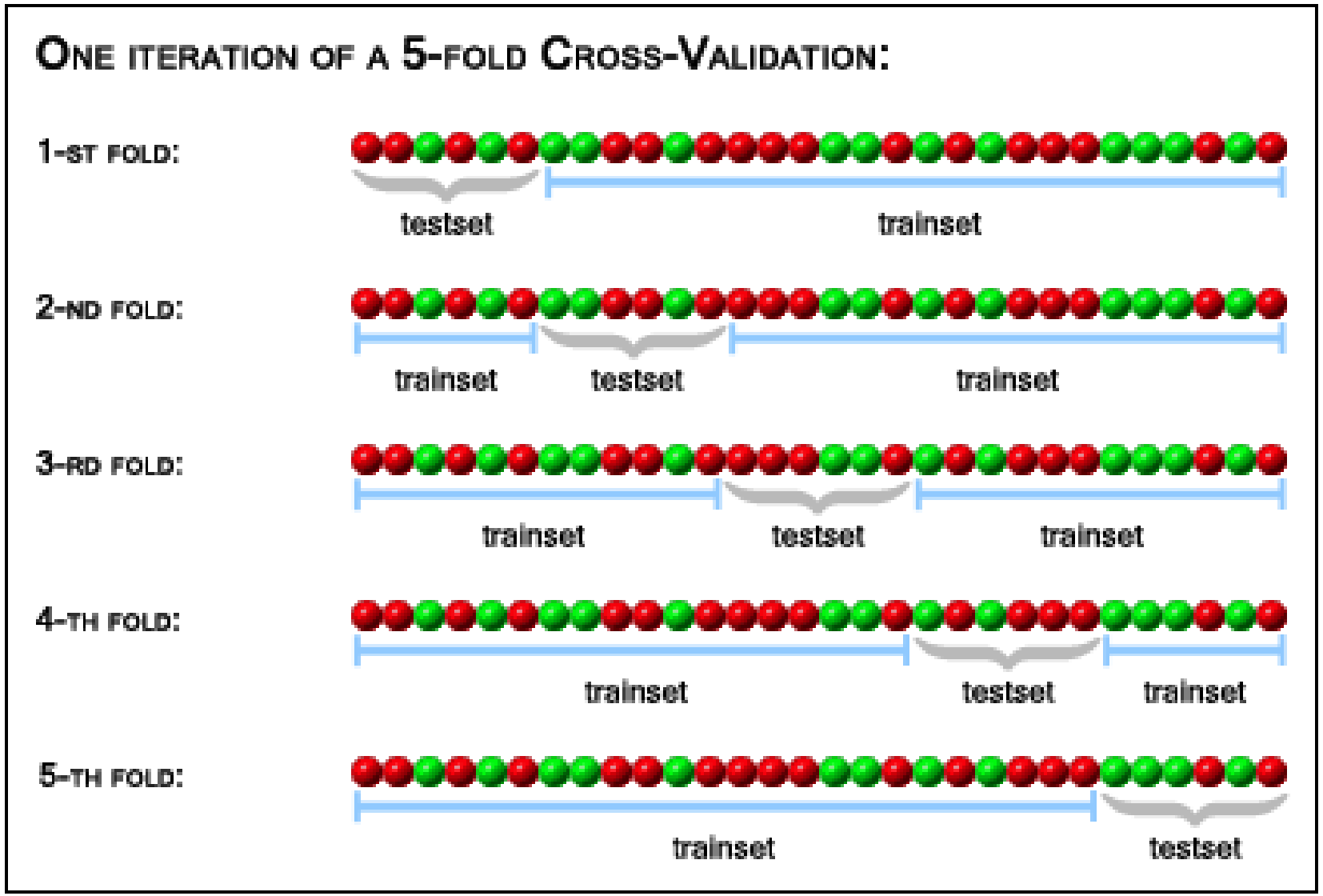
El método más sencillo de validación consiste en dividir aleatoriamente las observaciones disponibles en dos grupos, uno se emplea para entrenar al modelo y otro para probarlo.



K-Fold Cross-Validation

- Consiste en dividir los datos de forma aleatoria en k grupos de aproximadamente el mismo tamaño, k-1 grupos se emplean para entrenar el modelo y uno de los grupos se emplea como validación.
- Este proceso se repite k veces utilizando un grupo distinto como validación en cada iteración.
- El proceso genera k estimaciones del error cuyo promedio se puede emplear como desempeño del entrenamiento.

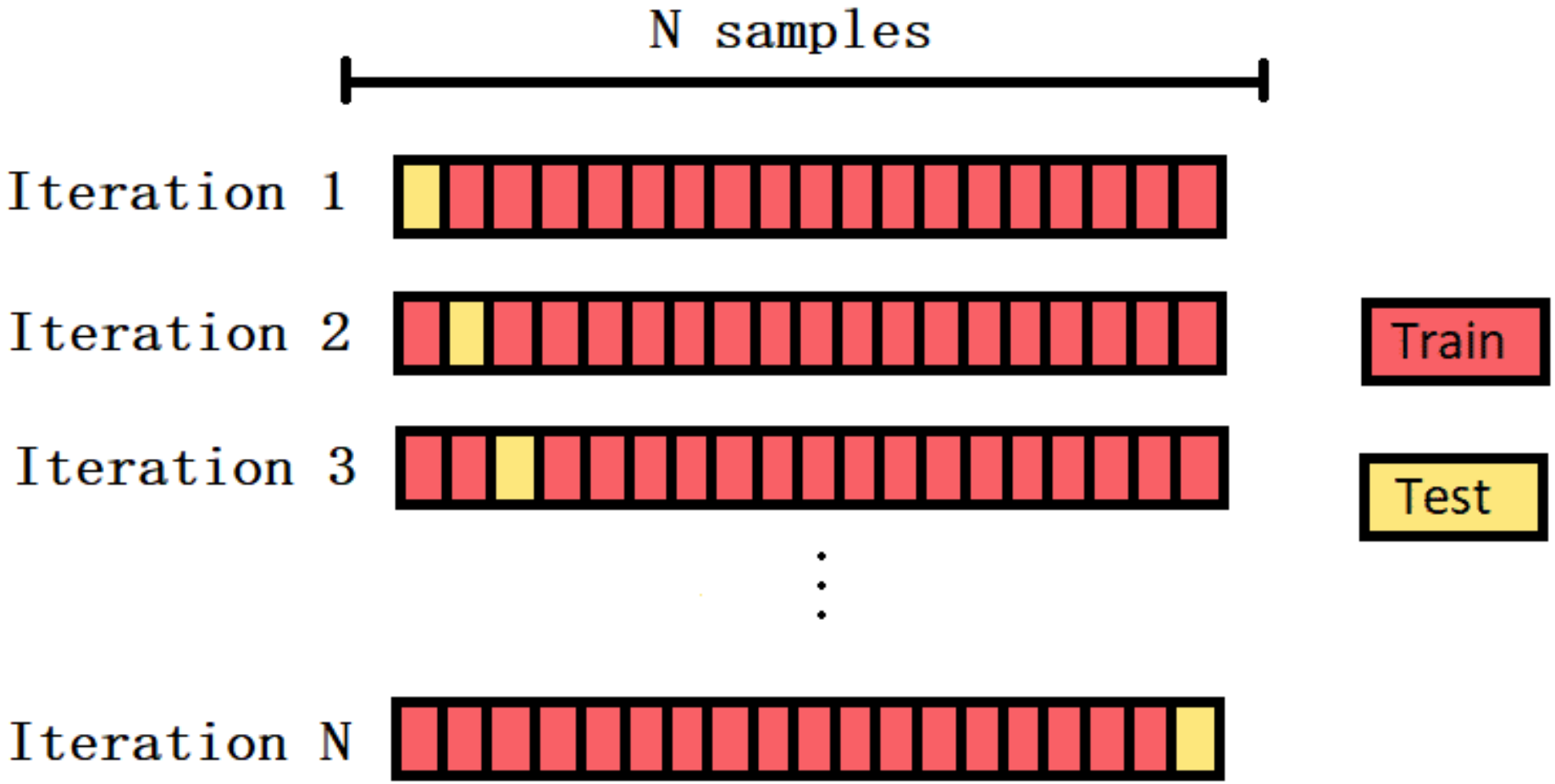
Ejemplo con k=5



Leave One Out Cross-Validation

- El método LOOCV en un método iterativo que se inicia empleando como conjunto de entrenamiento todas las observaciones disponibles excepto una, que se excluye para emplearla como validación.
- Si se emplea una única observación para calcular el error, este varía mucho dependiendo de qué observación se haya seleccionado. Para evitarlo, el proceso se repite tantas veces como observaciones disponibles, excluyendo en cada iteración una observación distinta, ajustando el modelo con el resto y calculando el error con dicha observación.

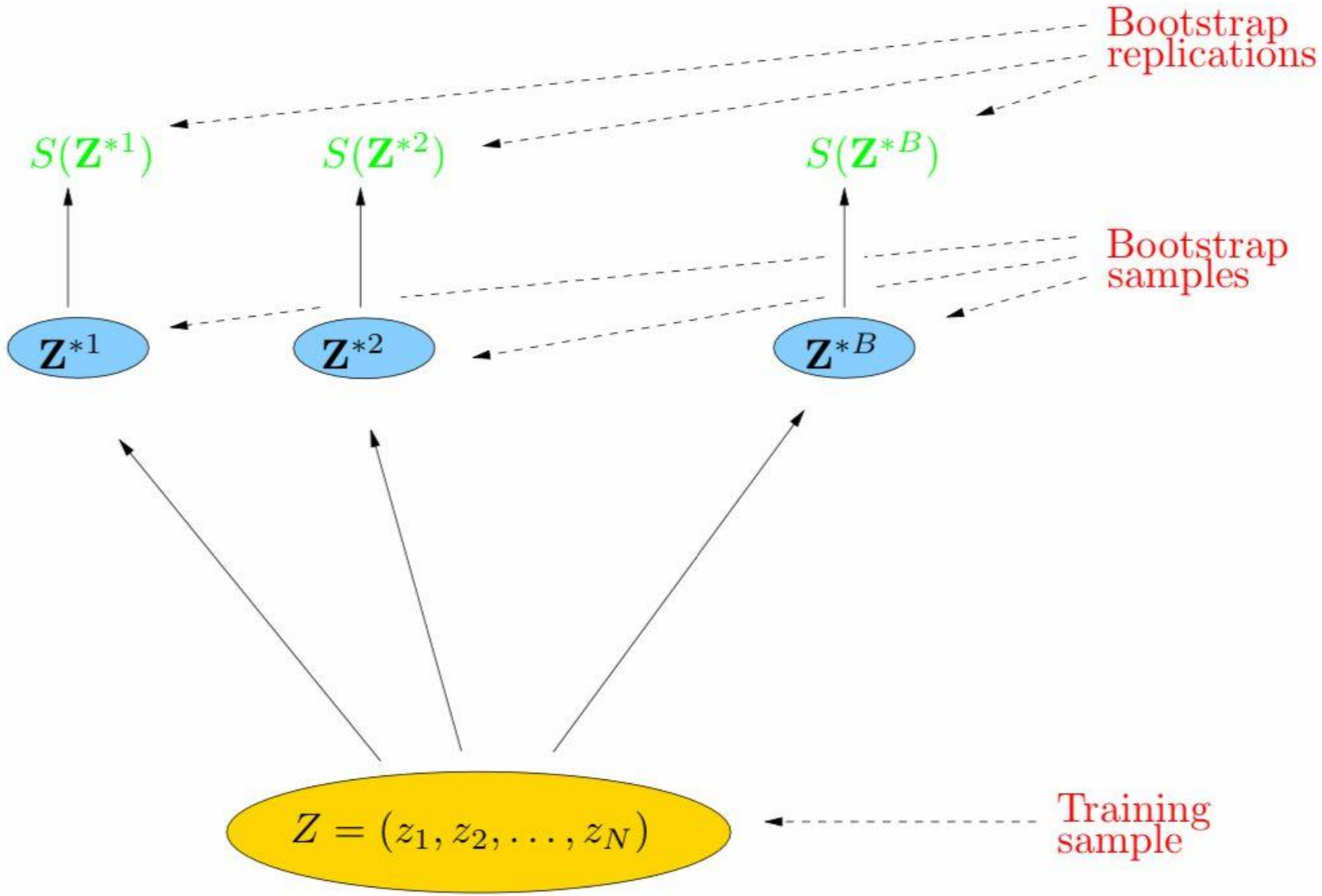
Leave-One-Out CV



Bootstrap

Consideremos una muestra $Z = (z_1, z_2, \ldots, z_N)$. La idea básica es extraer aleatoriamente muestras con reemplazo a partir de los datos de entrenamiento. Cada muestra es seleccionada del mismo tamaño que el conjunto de entrenamiento original. Esto se hace B veces (B = 100, por ejemplo), produciendo B conjuntos de datos Bootstrap.

A partir de las B muestras *Bootstrap* cualquier cantidad *S(Z)* se puede calcular como estimación de la misma, pero a partir de los datos *Z*.


Bootstrap replications
$S(\mathbf{Z}^{*1})$
$S(\mathbf{Z}^{*2})$
$S(\mathbf{Z}^{*B})$
Bootstrap samples
$\mathbf{Z}^{*1}$
$\mathbf{Z}^{*2}$
$\mathbf{Z}^{*B}$
$Z = (z_1, z_2, \ldots, z_N)$
Training sample

En aprendizaje

Una muestra bootstrap es una muestra obtenida a partir de la muestra original por muestreo aleatorio con reposición, y del mismo tamaño que la muestra original. Muestreo aleatorio con reposición (resampling with replacement) significa que, después de que una observación sea extraída, se vuelve a poner a disposición para las siguientes extracciones.

Como resultado de este tipo de muestreo, algunas observaciones aparecerán múltiples veces en la muestra bootstrap y otras ninguna. Las observaciones no seleccionadas reciben el nombre de out-of-bag (OOB). Por cada iteración de bootstrapping se genera una nueva muestra bootstrap, se ajusta el modelo con ella y se evalúa con las observaciones out-of-bag.

Procedimiento OOB

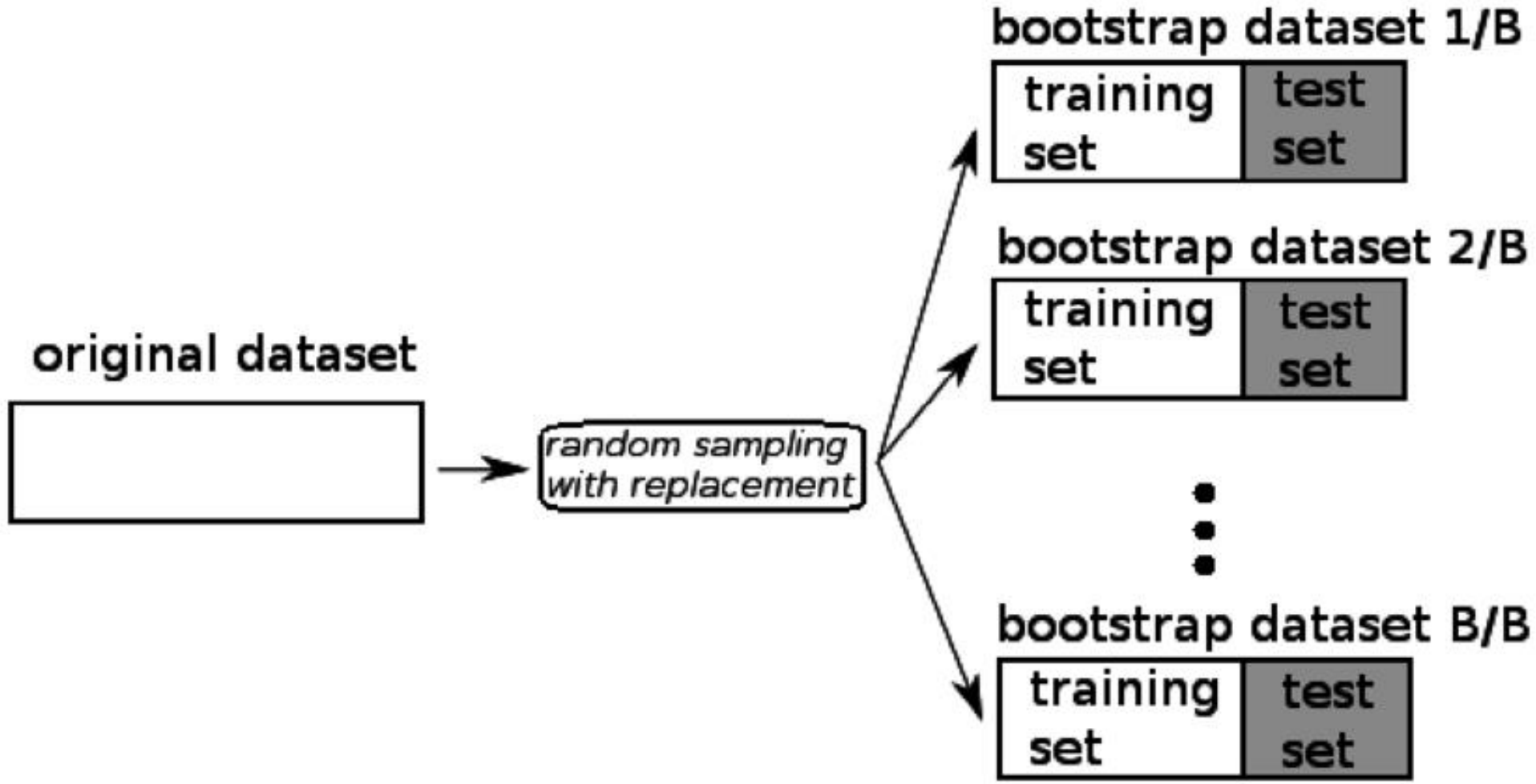
1. Obtener una nueva muestra del mismo tamaño que la muestra original mediante muestro aleatorio con reposición.

2. Ajustar el modelo empleando la nueva muestra generada en el paso 1.

3. Calcular el error del modelo empleando aquellas observaciones de la muestra original que no se han incluido en la nueva muestra. A este error se le conoce como error de validación.

4. Repetir el proceso B veces y calcular la media de los B errores de validación.

Bootstrap


original dataset
random sampling with replacement
bootstrap dataset 1/B
training set
test set
bootstrap dataset 2/B
training set
test set
bootstrap dataset B/B
training set
test set

Referencias

1. Bishop C. (2006) Pattern Recognition and Machine Learning. Wiley. Springer-Verlag New York
2. James G. et al. (2021). An Introduction to Statistical Learning 2nd Edition. Springer, New York, NY.
3. Hastie T., Tibshirani R. and Friedman J. (2009). The Elements of Statistical Learning 2nd Edition. Springer-Verlag.

Contacto

Roberto Bárcenas Curtis

Doctor en Ciencias con especialidad en Probabilidad y Estadística

rbarcenas@ciencias.unam.mx