Introducción al Aprendizaje Estadístico - Actividades de Clase - Unidad 2
Lucía Babino - Paula Spano

Datos simulados de altura

Instrucciones: **En esta página web se podrá acceder a los datos simulados que se van a utilizar para resolver los puntos de esta clase. La idea es que cada grupo trabaje con sus propios datos.**

Ejercicio 1: Calentando motores

1. Descargar de la página dos conjuntos de datos con $n = 50$ y $n = 500$ observaciones, respectivamente, ingresando como nro de identificación los 5 últimos números del DNI de algún integrante del grupo y con todas las variables. Importar los archivos a R, cada uno en un data frame. (Nos concentraremos en el análisis de la altura de los individuos.)

2. Identificar el nombre de las columnas del data frame.

Para cada conjunto de datos:

3. Predecir la altura de un individuo, de la que no se tiene ninguna información. Ingresar los valores predichos en las columnas "n = 50" y "n = 500" de la siguiente planilla. **Importante:** utilice la coma (",") como separador decimal.

4. Realizar un histograma de las alturas de los individuos. Establecer 15 clases para la construcción de los mismos. ¿Cuántas modas se observan? ¿A qué se puede atribuir?

5. Predecir la altura de un individuo de género masculino (hijo) y comparar con la predicción anterior.

6. Predecir la altura de un hijo cuya madre es de contextura pequeña y comparar con el valor del ítem anterior.

Ejercicio 2: Teniendo en cuenta la altura de la mamá

Trabajar solamente con el conjunto de datos de tamaño 500.

7. Graficar la altura de la mamá (en el eje x) vs. la altura del individuo (eje y), utilizando un color distinto por cada género. ¿Qué se puede observar?

En adelante, trabajaremos sólo con los datos de los varones. Para ello, crear un data frame llamado `alturasdat500m` *que contenga los datos correspondientes a los varones.*

8. Indicar si hay alguna madre de un varón cuya altura sea 156 cm.

9. Predicir la altura de un varón cuya madre mide $x = 156$ (cm) calculando el *promedio local* centrado en 156 con ventana de ancho $h = 1$ (cm). Para ello:

   a) Indicar cuántos casos hay donde la madre registra una altura entre 155 y 157 cm., inclusive.

   b) Calcular el promedio de la altura de los varones cuyas madres registran una altura entre 155 y 157 cm.

   c) repetir con $h = 2$.

10. Ingresar las predicciones en la siguiente en las columnas "h = 1" y "h = 2" de la siguiente planilla. **Importante:** utilice la coma (",") como separador decimal.

Ejercicio 3: Implementando funciones

11. Implementar una función que en base a los datos de la altura de las madres (`x`) y de la altura de sus hijos (`y`), permita predecir la altura de un hijo cuya madre tiene altura igual a `x_nueva`, usando una ventana de tamaño `h`, mediante el cálculo del promedio local. Es decir, defina la función
    `pred_prom_loc(x, y, x_nueva, h)`

    (Notar que esta función se puede aplicar para cualquier conjunto de datos `(x, y)`)

12. Graficar la función predictora de la altura por promedios locales, `pred_prom_loc`, para `h=1` en base a los datos de los varones que guardó en `alturasdat500m`. Para ello, generar una grilla de 100 valores equidistantes entre 151 y 168 para `x_nueva` y evaluar la función en cada uno de esos puntos.

13. Repetir el ítem anterior usando `h = 2` y luego `h = 5`. Representar las tres funciones en un mismo gráfico utilizando un color diferente para cada valor de `h`. ¿Qué observa?

Ejercicio 4: Usando las funciones de R.

La función `ksmooth` *de R implementa el método de Nadaraya-Watson.*

14. Utilizar la función `ksmooth` de R para predecir la altura de un hijo (varón) cuya madre mide 156 cm, aplicando promedios locales con ventana `h = 2`. Verificar con el valor obtenido en el Ejercicio 2, ítem 9*c*.

    Hint1: recuerde que el método de promedios locales es un caso particular del de Nadaraya-Watson, ¿con qué núcleo?

    Hint2: como estamos prediciendo la altura de un varón, debemos utilizar el conjunto de datos `alturasdat500m` *definido en el Ejercicio 2.*

15. Repetir la predicción del ítem anterior pero utilizando el estimador de Nadaraya-Watson con núcleo normal.

16. Realizar un gráfico de dispersión de las alturas de los hijos vs. las alturas de las madres y superponer (con distinto color) las funciones predictoras de Nadaraya-Watson que utilizan el núcleo normal y el uniforme, en ambos casos con ancho de ventana $h = 2$.

Ejercicio 5: Vecinos más cercanos

17. Realizar la predicción para la altura de un hijo (varón) de una mamá que mide $x = 156$ cm calculando el promedio de los $k = 7$ vecinos más cercanos.

18. Utilizar la función `knn.reg`, incluída en la librería `FNN` de R, para predecir la altura de un hijo cuya madre mide 156 cm, utilizando los `k = 7` vecinos más cercanos. Comparar con el resultado obtenido en el ítem anterior.

Ejercicio 6: Selección del parámetro de suavizado

19. Utilizando el conjunto de datos `alturasdat500m` y la función `pred_prom_loc(x, y, x_nueva, h)` implementada en el **Ejercicio 3**, hallar la ventana óptima h_{opt} para el método de promedios locales mediante el criterio de Validación Cruzada - Leave One Out (LOOCV). Llevar a cabo la búsqueda para una grilla de valores de h entre 1.5 y 4 con paso 0.05. Para ello...

    a) Prediga la altura del primer individuo del conjunto `alturasdat500m` por el método de promedios locales con $h = 1{,}5$ utilizando todos los datos menos el de ese individuo. Calcule su error cuadrático de predicción, es decir $\{Y_1 - \widehat{r}^{(-1)}(X_1)\}^2$.

    b) Repita para el resto de los individuos y guarde dichos valores en un vector. Inspeccionar el primer elemento de ese vector para corroborar que dé igual al cálculo del ítem anterior.

    c) Calcule el error de validación cruzada para $h = 1{,}5$, es decir $CV(1{,}5)$.

    d) Implemente una función `loocv(h)` que calcule el error de validación cruzada $CV(h)$ para un h dado. Evalúe dicha función en 1,5 y corrobore con el resultado del ítem anterior.

    e) Calcule el error de validación cruzada para cada h en la grilla propuesta y guarde dichos valores en un vector. Inspeccione el primer elemento de dicho vector y corrobore con el resultado del ítem anterior.

    f) Halle el h_{opt} que minimiza $CV(h)$ en la grilla.

20. Realizar un gráfico de h vs. $CV(h)$. ¿Cuánto vale $CV(h_{\text{opt}})$?

21. Realice un gráfico de dispersión de las observaciones y superponga las funciones predictoras de promedios locales utilizando las ventanas del Ejercicio 3 (es decir $h = 1, 2$ *y* 5) junto con la ventana óptima hallada.