Actividad 1 - Primeras Predicciones

ExactasPrograma - Datos

Invierno 2020

El principal objetivo de esta guía es seguir programando, pero hemos elegido en este nuevo taller una excusa con datos. Es una especie de ejemplo juguete, pero poderoso (¡el chiquitín!)

Para trabajar vamos a descargar los datos distintos para cada sala. Entrá a

https://pilas.exp.dc.uba.ar/datos/alturas/index.html

y descargá los que te tocan.

Entre las diferentes formas de levantar los datos, podes hacer esto: click derecho en el botón descargar, copiar link y eso pegarlo en el `pd.read_csv`.

1. Calentando motores

1. Levantar los datos del archivo CSV.
2. Identificar el nombre de las columnas (variables) del archivo de datos.
3. ¿Cuántos individuos conforman tu conjunto de observaciones?
4. Calcular el promedio de las alturas de los hijos. Este valor se podría usar para predecir la altura de un nuevo individuo, ¿no?
5. ¿Con qué valor se puede predecir la altura de un nuevo individuo masculino?
6. ¿Con qué valor se puede predecir la altura de un nuevo individuo masculino cuya madre es bajita?

2. Vamos ahora a considerar la altura de la mamá

7. Graficar altura de mamá (en el eje x) vs. altura del hijo (eje y), utilizando un color por cada género. ¿Qué se puede observar? Explorar `hue` para agrupar (según una nueva variable) y producir puntos de diferentes colores.

   En adelante, trabajaremos con los datos de los hijos (género masculino); volvamos a graficar la altura de mamá (en el eje x) vs. altura del hijo (eje y), cuando genero=="M".

8. Indicar si hay alguna madre de altura 156.5 cm con un hijo varón. ¿Cuántas son?
9. Vamos ahora a predecir la altura de un hijo correspondiente a una mamá que mide `x_nuevo`=156.5 cm haciendo *promedio móvil* (o promedio local) centrado en 156.5 cm con ventana de tamaño h =1 (cm).

   Atención: Ventana h=1 indica que hay que mirar h=1 a derecha y h=1 a izquierda.

   a) Indicar cuántos casos hay donde la madre registra una altura de 155.5 cm a 157.5 cm, inclusive.

   b) Calcular el promedio de la altura de los hijos cuyas madres registran una altura de 155.5 cm a 157.5 cm.

10. Promedio móvil centrado en 156.5 cm con ventana de tamaño $h = 2$ (cm).

    a) Indicar cuántos casos hay donde la madre registre una altura de 154.5 cm a 158.5 cm, inclusive.

    b) Calcular el promedio de la altura de los hijos cuyas madres registran una altura de 154.5 cm a 158.5 cm.

11. Repetir los ítems anteriores pero considerando ahora que la altura de la mamá es 159.5 cm. Es decir, calculamos los promedios en otro lado. Por eso hablamos de *promedios móviles*.

3. Implementando funciones

Hasta ahora hemos trabajado con dos posibles valores para la altura de la madre (156.5 cm y 159.5 cm) y dos posibles valores de ventana (h =1 y $h = 2$). Ahora vamos a implementar una función que permita predecir la altura de un hijo en función de la altura de la madre y el tamaño `h` de ventana elegida para hacer el promedio móvil.

12. Implementar una función que tenga por input una columna de un `dataframe` con los `X` (variable predictora - alturas de la madres), otra columna con sus correspondientes valores de `Y` (variable respuesta - altura de los hijos), un nuevo valor `x_nuevo` para la variable predictora (que sería la altura de la madre donde quiero predecir). Finalmente un valor h de `ventana` que vamos a utilizar a la hora de hacer promedios móviles; éste último (el promedio móvil) es el resultado (valor de retorno) de la función. Aprovechamos para mencionar que, en la jerga de pandas, una columna de un `dataframe` se la llama "serie" y es una estructura muy útil en sí misma.

    ```
    predigo_promedio_movil(X, Y, x_nuevo, ventana)
    ```

13. Graficar la función `predigo_promedio_movil`, utilizando como variable predictora la altura de las madres, como variable respuesta la altura, ambas correspondientes a los datos de género masculino, ventana $h = 1$, y con `x_nuevo` recorriendo una grilla sobre un intervalo que cubra todas las alturas observadas en las madres.

14. Repetir el ítem anterior usando ventana $h = 2$. Repetir usando ventana $h = 10$. Representar las tres funciones en un mismo gráfico utilizando un color diferente para cada valor de h.

15. Incluir en el gráfico anterior, con las tres curvas, el diagrama de dispersión de los datos utilizados.

4. Por si fuera poco

16. Calcular ahora predicciones para la altura del hijo conociendo la altura de la madre haciendo promedio de vecinos cercanos. Para ello, implementar una función que tenga por input una columna de un `dataframe` con los `X` (variable predictora), otra columna con sus correspondientes valores de `Y` (variable respuesta), un nuevo valor `x_nuevo` donde queremos predecir, y la cantidad k de vecinos más cercanos de `x_nuevo` que vamos a utilizar a la hora de hacer promedios.

    ```
    predigo_promedio_vecinos(X, Y, x_nuevo, k)
    ```

5. Cuadrados mínimos

Denotemos con (x_i, y_i), $i = 1, \ldots, n$, a los pares observados. La recta de cuadrados mínimos está dada por $y = m^* x + b^*$, donde (m^*, b^*) minimizan la función (de pérdida)

$$L(m, b) = \sum_{i=1}^{n} (y_i - (mx_i + b))^2$$

Se puede mostrar que esta función se minimiza en

$$b^* = \overline{y}_n - m^* \overline{x}_n \,, \quad m^* = \frac{\sum_{i=1}^n (x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sum_{i=1}^n (x_i - \overline{x}_n)^2}$$

$$\overline{x}_n = \frac{1}{n} \sum_{i=1}^n x_i \,, \quad \overline{y}_n = \frac{1}{n} \sum_{i=1}^n y_i$$

17. Implementar una función que tenga por entrada una columna de un `dataframe` con los `X` (variable predictora), otra columna con sus correspondientes valores de `Y` (variable respuesta) y devuelva la pendiente y la ordenada al origen de la recta de cuadrados mínimos: `cuadrados_minimos(X,Y)`.

18. Implementar una función que tenga por entrada un conjunto de valores `X`, sus correspondientes valores de `Y` y un nuevo valor `x_nuevo` donde queremos predecir la variable respuesta utilizando la recta de cuadrados mínimos construida con los datos `X` e `Y`: `predigo_con_cuadrados_minimos(X,Y, x_nuevo)`. Predecir la altura de un hijo varón cuya madre mide 156.5 cm utilizando la recta de mínimos cuadrados.

19. Graficar la altura de la mamá (en el eje x) vs. altura del hije (eje y), utilizando un color por cada género y agregar la recta de cuadrados mínimos para cada género.

    Para explorar...

```
from scipy import stats
slope_masculino, intercept, r, p, std_err = stats.linregress(df[filtro_masculino]['
  altura_madre'], df[filtro_masculino]['altura'])

print("El valor de la pendiente de la recta de minimos cuadadros para hijes  masculinos
  es",slope_masculino)
```

Si te quedas con ganas de más, en esta otra página podrás acceder a más datos como los que usaste en esta guía. A medida que cambia el tamaño n del conjunto de datos, se incluyen nuevos casos (se agregan más filas al archivo) o se excluyen si es lo reducimos.