Ciencia de datos, práctica 6

Juan Casado Ballesteros, Samuel García Gonzalez, Iván Anaya Martín

December 10, 2019

Abstract

TODO

Contents

1 Conceptos gnerales sobre visualización

2 Visualización de la regresión

Analizaremos ahora conceptos y soluciones específicas para la visualización de rectass de regresión. Lo más importante en estos diagramas no es la recta generada, pues si la correlación es baja esta puede no significar nada, si no la relación entre la recta y los datos. Deben poder visualizarse ambos elementos de forma adecuada pues es la relación entre ambos la que nos interesa, deseamos ver lo bien o mal que la recta se adapte a nuestros datos.

En el análisis de regresión compareremos como dos variables de nuestra muestra se relacionan entre ellas. Es recomendable por tanto, ya que este problema siempre estará en un dominio bidimensional, representarlo gráficamente. No siempre tenemos la oportunidad de visualizar de forma conveniente las relaciones entre nuestros datos, ya sea por la complejidad de la relación o por la anta dimensionalidad de las variables involocradas.

2.1 La importancia de la visualización

Como ejemplo ilustrativo de la importancia de la visualización ponemos el ya famoso ejemplo del libro de Edward Tufte, Data Analysis for Politics and Policy, Chapter 3: Two-Variable Linear Regression. En dicho ejemplo las rectas de regresión obtenidas son idénticas y se adaptan a ojos de la correlación de igual modo a los datos. Si no fuera por la visualización de ambos elementos en conjunto no podríamos conocer la gran disparidad entre cada una de las muestras.

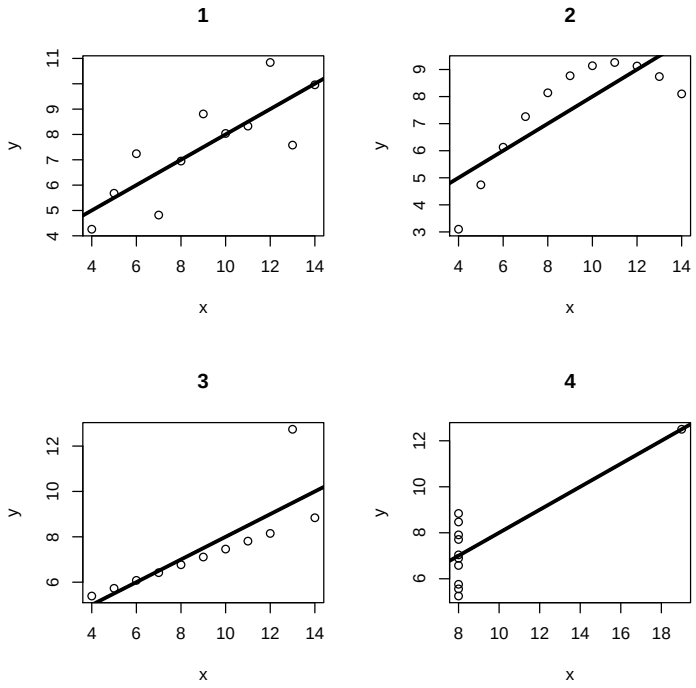
Si representáremos los datos solos sería difícil imaginar que tuvieran la misma regresión y si visualizáramos las regresiones solo sería difícil imaginar que podrían estar representando a datos tan dispares.

```
[1] "1: Correlacion cuadrada: 0.6665425 a: 3.000091 b: 0.5000909"

[1] "2: Correlacion cuadrada: 0.666242 a: 3.000909 b: 0.50"

[1] "3: Correlacion cuadrada: 0.666324 a: 3.002455 b: 0.4997273"

[1] "4: Correlacion cuadrada: 0.6667073 a: 3.001727 b: 0.4999091"
```



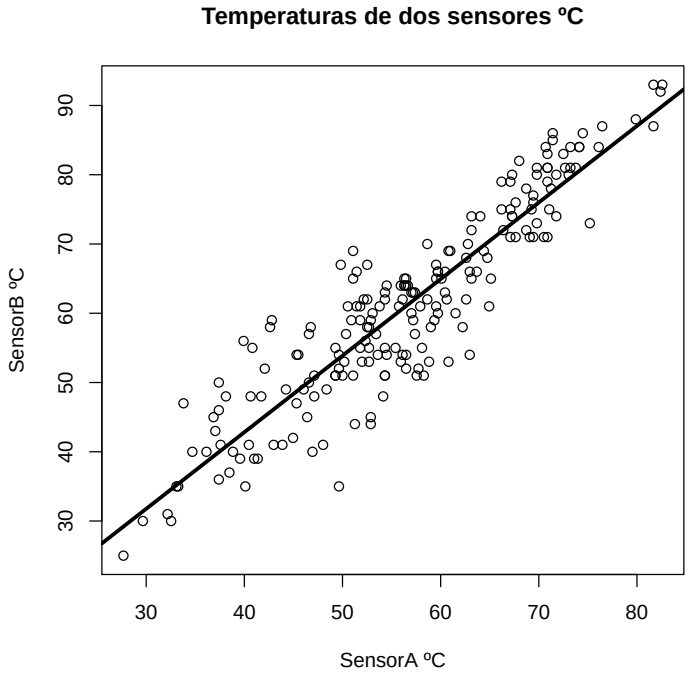
2.2 Cantidad de información mostrada

Debido a la simpolicidad de esta gráfica, solo se está mostrando una nube de puntos y la recta de regresión es tentador comenzar a añadir elementos adicionales. No obstante antes de hacerlo debemos tener en cuenta las siguientes consideraciones.

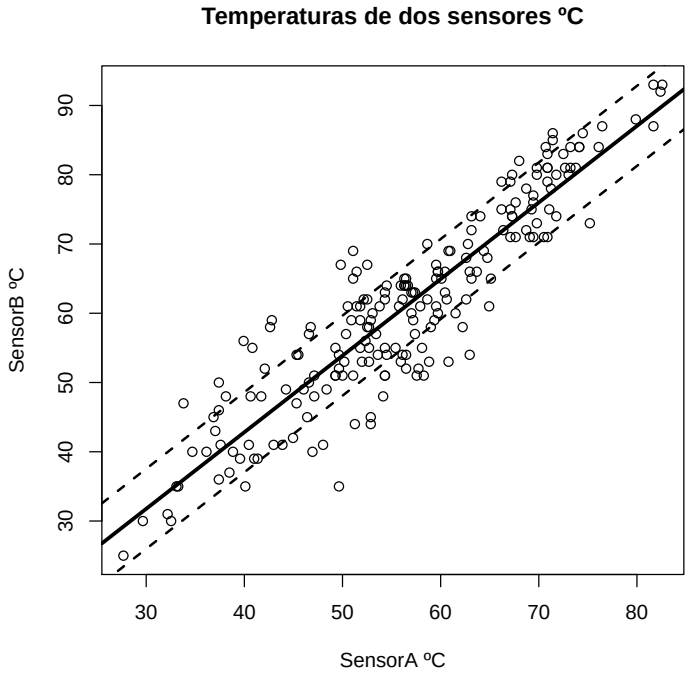
- Se debe mostrar la información mínima necesaria para mostrar aquello que deseamos. Si tenmos que mostrar una relación compleja entre los datos puede que sea necesario utilizar una representación compleja. No obstante si podemos hacerla simple será mejor pues podrá ser entendida por más gente de forma más rápida.
- Debemos de indicar el significado de cada elemento que añadamos a la gráfica cuando este no sea claro.
- Es de gran importancia indicar la margnitud de los datos tanto como el dato que se está representando en cada eje.

```
[1] "Correlacion cuadrada de Temperaturas: 0.835327071582904"
```

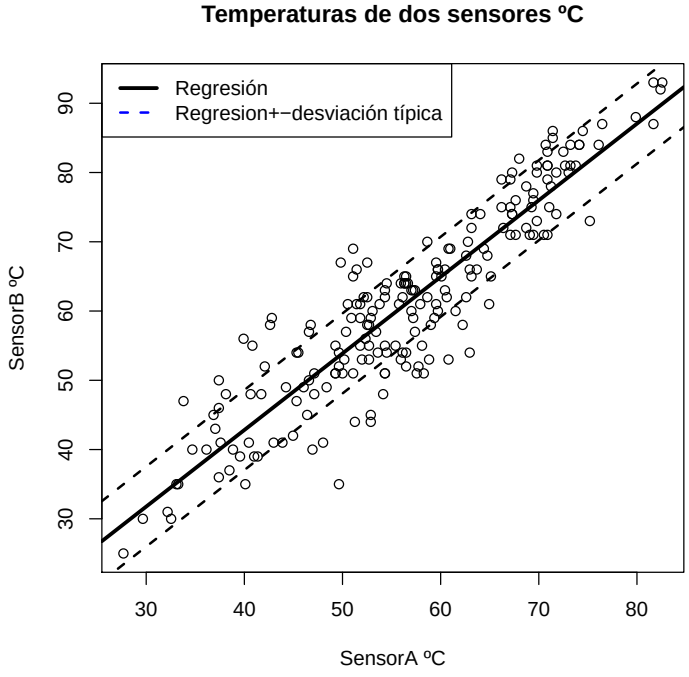
En este caso mostramos exclusivamente los datos jusnto a su recta de regresión. Obtenemos una representación muy clara y sencilla que nos muestra directamente la información que queríamos. La gráfica tiene un título y en cada eje mostramos la variable representada y las unidades de esta.



Podemos aumentar la complejidad de la representación añadiendo dos líneas paralelas a la recta de regresión que nos indiquen la desviación típica de esta. Este elemento adicional nos proporciona información adicional sobre la regresión que dependiendo del contexto puede ser necesaria si a partir de los datos y de la recta es difícil juzgar la calidad de esta. No obstante podemos ver como solo con haber añadido un elemento tan simple la representación se siente mucho más densa.



Como hemos indicado para este caso sería recomendable añadir una leyenda que nos indicara qué datos se estan representando de modo que la visualización obtenida sea más clara.



2.3 Paquetes de visualización

Para realizar las gráficas anteriores hemos utilizado la siguiente función que utiliza las funciones propias de R.

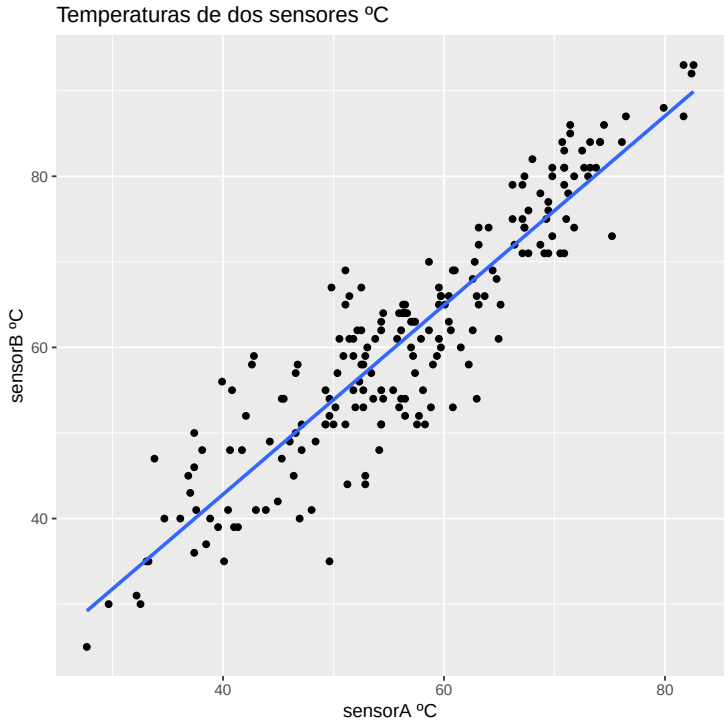
```
> regPlot

function (x, y, regresion, limit, title="", xlabel="", ylabel="") {
  plot(x, y, xlab=xlabel, ylab=ylabel, main=title)
  regUpLimit <- regresion
  regUpLimit$coefficients[1] = regUpLimit$coefficients[1] + limit
  regDownLimit <- regresion
  regDownLimit$coefficients[1] = regDownLimit$coefficients[1] - limit
  abline(regUpLimit, "gray", lty=2, lwd=2)
  abline(regresion, "black", lty=1, lwd=3)
  abline(regDownLimit, "gray", lty=2, lwd=2)
}
<bytecode: 0x7fa7563d2740>
```

Mediante plot visualizamos la nube de puntos y con abline dibujamos líneas rectas sobre la última gráfica creada.

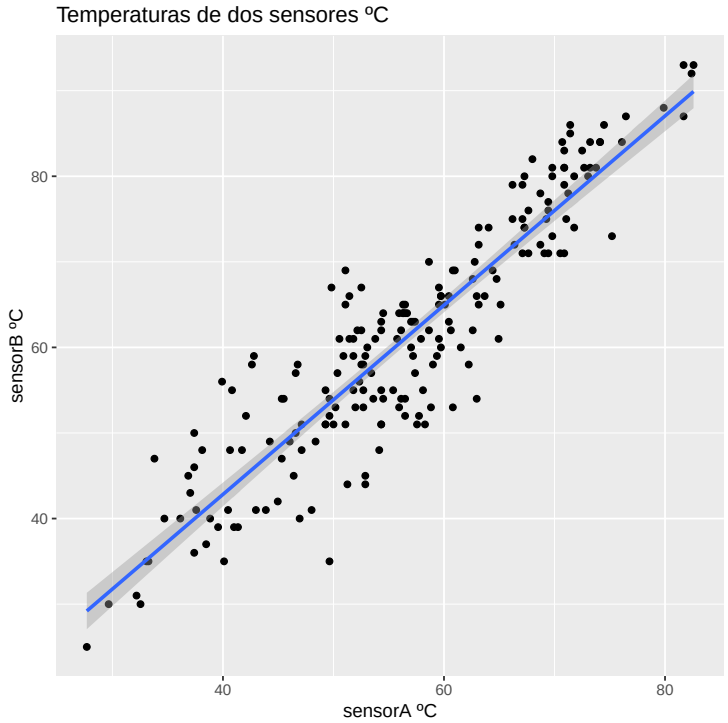
Existen una gran cantidad de paquetes que se pueden utilizar para visualizar datos. Uno de los más famosos es "ggPlot2". Esto se debe a que produce gráficas muy vistosas, altamente configurables todo con una sintaxis sencilla y fácil de aprender.

No obstante esto también es uno de los problemas de esta librería. Por defecto las gráficas que obtendremos tendrán un fóndo grisáceo con una cuadrícula en blanco tal y como podemos ver en el ejemplo. Esto hace que la representaciín de los datos sea más difusa que la obtenida con las funciones propias de R que crean una visualización mucho más limpia.



Estas librería permite además añadir gran variedad de elementos de forma sencilla a las gráficas. Dichos elementos se ven siempre muy vistosos y bonitos pero pueden no tener gran significado para nuestro caso concreto. Debemos por tanto tener cuidado y no abusar de ellos.

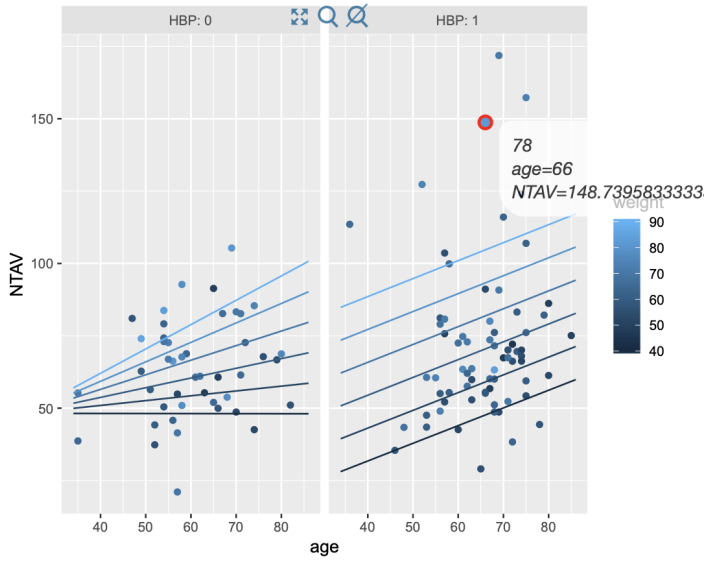
```
> ggplot(data = datos3, aes(x = datos3$Temperature_ElMonte,
+                           y = datos3$Temperature_Sandburg)) +
+   geom_point(color='black')+
+   geom_smooth(method="lm")+
+   labs(title = "Temperaturas de dos sensores ºC") +
+   xlab("sensorA ºC") + ylab("sensorB ºC")
```



2.3.1 Representaciones interavtivas

Una de las ventajas que tenmos con estas librerías es que podemos crear gráficas interactivas. Estas gráficas crean un archivo .html que podemos visualizar de forma interactiva en el navegador. Las gráficas interactivas nos permiten realizar representaciones más complejas de los datos así como representar más datos a la vez. Esto se debe a que cualdo pasemos el cursosr encima de los elementos se resaltarán los que están relacionados pudiendo ver simultáneamente estos con respecto a los otros.

Mostramos uan gráfica creada con "ggplot2" que se ha convertido en una gráfica interactiva con "ggiraphExtra".



En este caso mostramos una gráfica interactiva creada con ”plotly” la cual utiliza ”ggplot2” por debajo para crear sus gráficas. Esta librería tiene también una sintaxis sencilla y permite crear grácisas muy configurables. Como ventaja frente a ”ggplot2” debemos comentar que los estilos que se aplican por defecto a las gráficas son mucho más limpios y claros.

```
> datos3 %>%
+   plot_ly(x = ~Temperature_ElMonte, y = ~Temperature_Sandburg) %>%
+   add_markers (x = ~Temperature_ElMonte, y = ~Temperature_Sandburg,
+     name="temperaturas ºC") %>%
+   add_lines(x = ~Temperature_ElMonte, y = fitted(regresion3), name="regresion")%>%
+   layout(title = "Temperaturas de dos sensores ºC",
+     xaxis = list(title = "sensorA ºC"),
+     yaxis = list(title = "sensorB ºC"))
```