Ejemplo modelado de regresión lineal con R (Tidyverse+Rmarkdown)

Jesús Manuel Nieto Carracedo (jesusmanuel.nieto@etani.es)

30/11/2020

Modelo predictivo basado en recta de regresión lineal

Tecnología empleada:

- Lenguajes de programación:
  - Análisis de datos: R
  - Maquetación del documento: RMarkdown
  - Maquetación de las fórmulas: Latex
- Librerías:
  - Agrupa a las siguientes: Tidyverse
  - Gráficos: ggplot2
  - Datos: ggplot2::mpg

Planteamiento:

Partimos del siguiente **dataframe** llamado **mpg** el cual contiene datos de economía de combustible de 1999 a 2008 para 38 modelos de vehículos más populares del mercado estadounidense. En ella queremos ver si hay relación directa entre el tamaño de un motor **displ** (cilindrada del motor en litros) y el número de millas de carretera por galón de combustible **hwy**. A priori, todo parece indicar que deber de existir. En este caso queremos estudiar **la variable dependiente displ(eje vertical) frente a la independiente hwy (eje horizontal)**

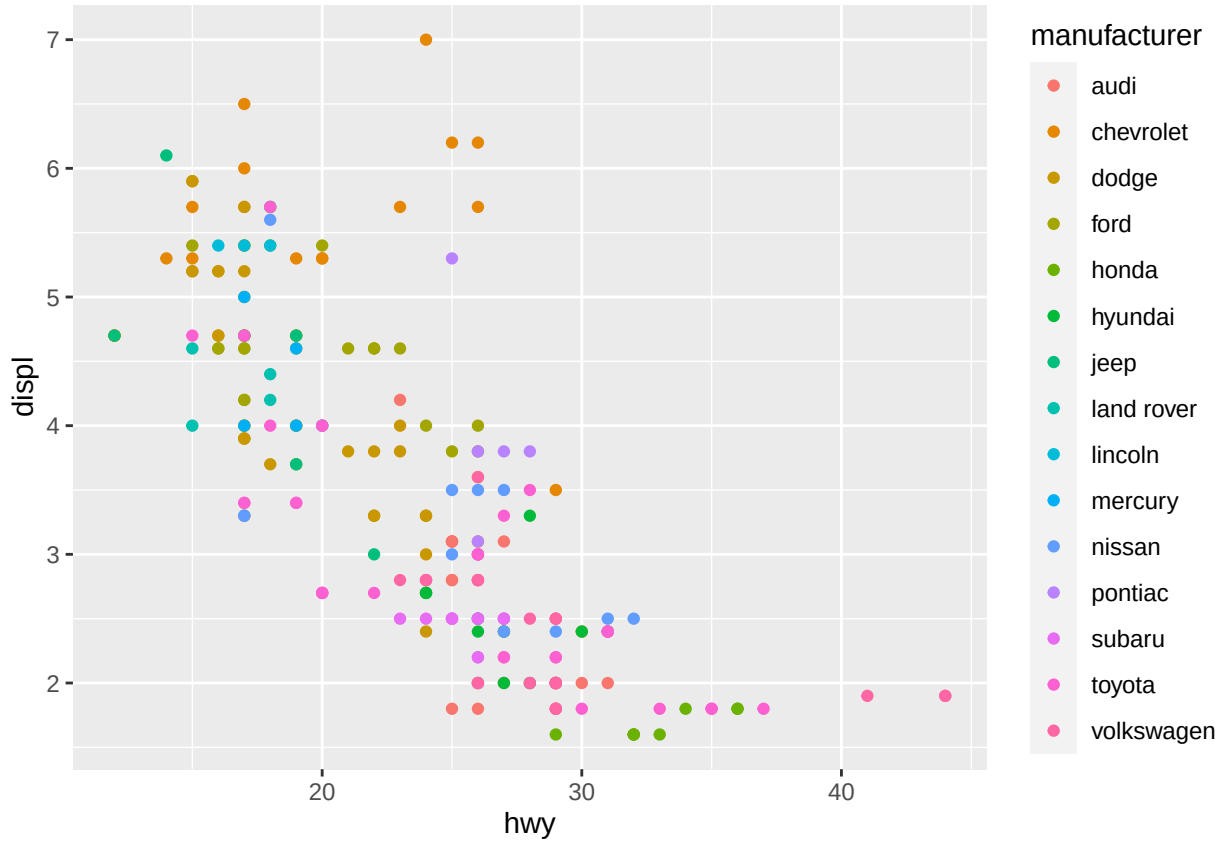
Para ello en esta tabla mostraremos **11** de **234** observaciones en total, para ver el formato de la misma:

Table 1: Dataframe mpg

manufacturer	model	displ	year	cyl	trans	drv	cty	hwy	fl	class
audi	a4	1.8	1999	4	auto(l5)	f	18	29	p	compact
audi	a4	1.8	1999	4	manual(m5)	f	21	29	p	compact
audi	a4	2.0	2008	4	manual(m6)	f	20	31	p	compact
audi	a4	2.0	2008	4	auto(av)	f	21	30	p	compact
audi	a4	2.8	1999	6	auto(l5)	f	16	26	p	compact
audi	a4	2.8	1999	6	manual(m5)	f	18	26	p	compact
audi	a4	3.1	2008	6	auto(av)	f	18	27	p	compact
audi	a4 quattro	1.8	1999	4	manual(m5)	4	18	26	p	compact
audi	a4 quattro	1.8	1999	4	auto(l5)	4	16	25	p	compact
audi	a4 quattro	2.0	2008	4	manual(m6)	4	20	28	p	compact
audi	a4 quattro	2.0	2008	4	auto(s6)	4	19	27	p	compact

Para este estudio, nos vamos a olvidar del modelo, tipo de coche, y de la marca, aunque debería ser relevante, ya que un motor de 2 litros de Dodge, es probable que consuma diferente a un 2 litros de Audi. Esta información es relevante, para anlizar los outliers, ya que habría que explorar el gráfico por marca y/o clase. Vamos a realizar un primer análisis exploratorio, mostrando una **nube de puntos**, un **diagrama de dispersión**, donde ya deberíamos poder intuir que aspecto tiene, y colocamos una tercera dimensión de variable, al introducir un color por marca, esto ya nos ayuda a ver que los outliers pertenecen a una marca casi en su totalidad, lo cual nos permitiría estudiar al grupo común por un lado y luego explicar cada grupo de outliers por separado.

Recordar que vamos a buscar la recta de regresión para la totalidad de la población:



Bueno, ya se van viendo detalles, se ve claramente un patrón, de relación indirecta o inversa, es decir, como era de esperar, aunque hay algunos outliers que tendrían que estudiarse, en este ejercicio me olvido de ellos, se observa que, a mayor cilindrada, menos millas por carretera.

Estadística descriptiva de variable displ:

Vamos a realizar ahora los cálculos necesarios para la recta de regresión:

- Media displ -> 3.4717949
- Desviación típica displ -> 1.291959

Estadística descriptiva de variable hwy

Vamos a realizar ahora los cálculos necesarios para la recta de regresión:

- Media hwy -> 23.4401709
- Desviación típica hwy -> 5.9546434

Desviación estándar conjunta $s_{x,y}$

Para calcular la desviación estándar conjunta $s_{x,y}$, haremos uso de la fórmula:

$$s_{x,y} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

- Desviación estándar conjunta $s_{x,y}$ -> -5.8679268

Coeficiente de correlación conjunta $r_{x,y}$

Ahora que tenemos la **media**, la **desviación estándar individual y conjunta**, vamos a calcular el **coeficiente de correlación lineal de Pearson:**, mediante la fórmula:

$r_{x,y} = \frac{s_{x,y}}{s_x * s_y}$.

- Coeficiente de correlación lineal de Pearson $r_{x,y}$ -> -0.7627464 Dado que es un valor negativo y cercano a -1, podríamos decir que tiene una relación **fuerte e inversa** al estar en el rango **[-1,1]** y por tanto si existe una relación lineal.

Cálculo de la recta de regresión

Para finalizar, ya tenemos todos los datos, para calcular la **recta de regresión**:

$y = \beta_0 + \beta_1 x$

Donde calcularemos el **intercepto** con:

$\beta_0 = \overline{y} - \beta_1 \overline{x}$

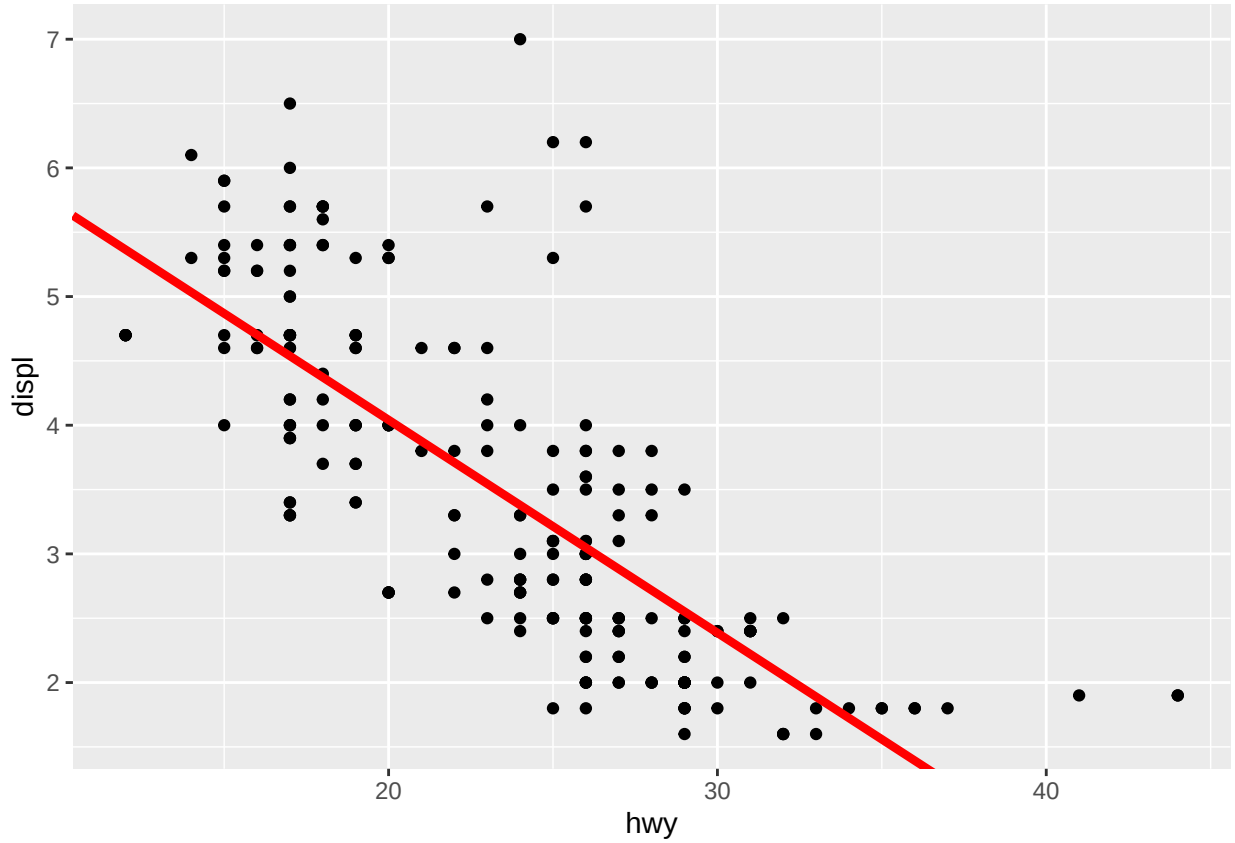
Y cuya **pendiente de la recta** será:

$\beta_1 = r_{x,y}\frac{s_y}{s_x}$

- Intercepto β_0 -> 7.3509213
- Pendiente de la recta β_1 -> -0.1654905
- y = 7.3509213 -0.1654905x

Representación gráfica del modelo

En la siguiente gráfica podremos observar como se define una recta que pintamos de color rojo, que indica la tendencia lineal inversa del modelo.



Nota final

El lenguaje de programación para estadística **R** dispone de una función llamada **lm()** la cual nos permite generar de forma *automágica* modelos re regresión lineal, pero dado que el objetivo del presente documento es realizar el cálculo del modelo de forma manual, no se ha hecho uso del mismo. Para más información puedes dirigirte a la url linear-regression-in-r/