**2020/02/08**

**APRENDIZAJE SUPERVISADO**

**Mario Encinar**

**Introducción**

El aprendizaje supervisado está basado en técnicas de inferencia estadísitica. Técnicas que nos sirven para inferir relación etre variables.

Estadística se centra en explicación y Machine learning se centra en predecir.

Estadística basada en matemáticas, y Machine Learning en computación (algorithmos). Aunque en estadística también hay algorithmos.

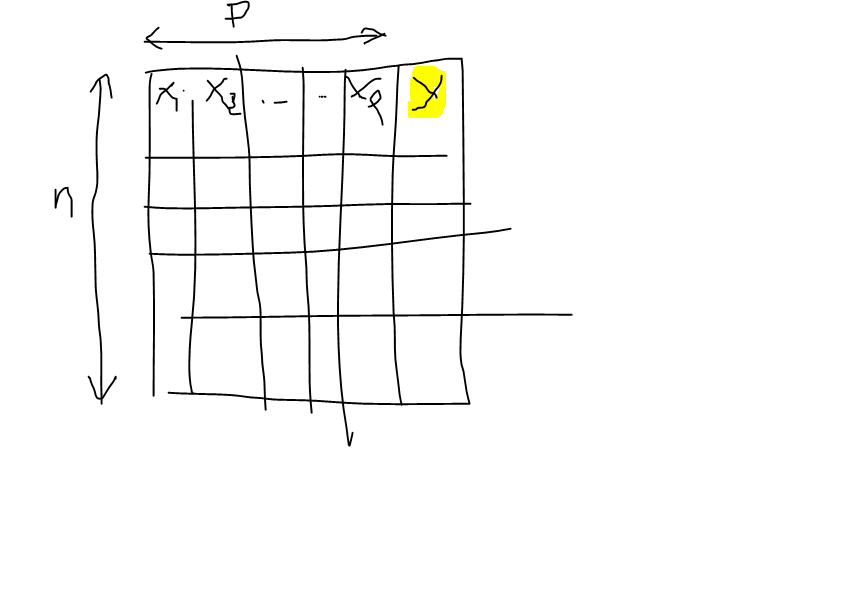
las varibales pueden ser continuas, discretas o categóricas. Inclusos variables continuas se pueden codificar como variables categóricas haciendo tramos.

Applied Machine Learning: es lo que realmente hacemos, es decir no necesitamos saber los algorithmos o las mates que calculan los modelos, sólo entenderlos.

**Aprendizaje Supervisado**

Dos tipos de problemas:

* **Regresión**: predice una variable continua. Nuestra “y” target es continua.
* **Clasificación**: predice una variable discreta. Nuestra “y” target es discreta. Si los valores de y son solo 0, 1 sería un problema de clasificación binaria. Si son más se llama clasificación multiclase, en este caso hay dos alternativas: romperlo en problemas binarios, o resolverlo como problema multiclase.



En aprendizaje supervisado tenemos un conjunto de variables de entrada xi y una variable objetivo **y** conocida.

xi: Variables de entrada (inputs, predictores, features, variables explicativas, variables independientes).

y: variable de salida, respuesta o variable objetivo.

**NOTAS**:

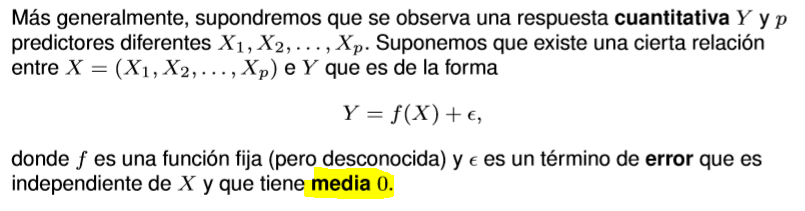
En aprendizaje **no supervisado** sólo tendríamos las variables xi, pero no la **y**.

**Exploratorio**: sacar estadísticos básicos y hacer varios plots para explorar visualmente.

* Con variables continuas: *scatterplot*.
* Con continuas vs discreta, o discreta vs discreta: *histogramas.*

**Qué hace un científico de datos:**

|  |  |  |
| --- | --- | --- |
| DATA SCIENCE | Extracción y preprocesado:  SQL, HIVE, JOINTS | Data Engineer |
| Análisis descriptivo, exploratorio, diagnóstico. | Data Scientist |
| Análisis predictivo y prescriptivo. | Data Scientist |
| Producción | Data Engineer |



**Condiciones o hipotesis del modelo de regresion lineal:**

- Media 0: por qué media 0? porque si no el modelo estadístico de regresión no funciona.

- Distribución normal.

- Varianza constante.

**Función de Estimación**

Es la función que nos permitirá estimar los valores de Y^.

Y^ = h(X)

Cuando Y^ lleva el gorrito significa que es una estimación.

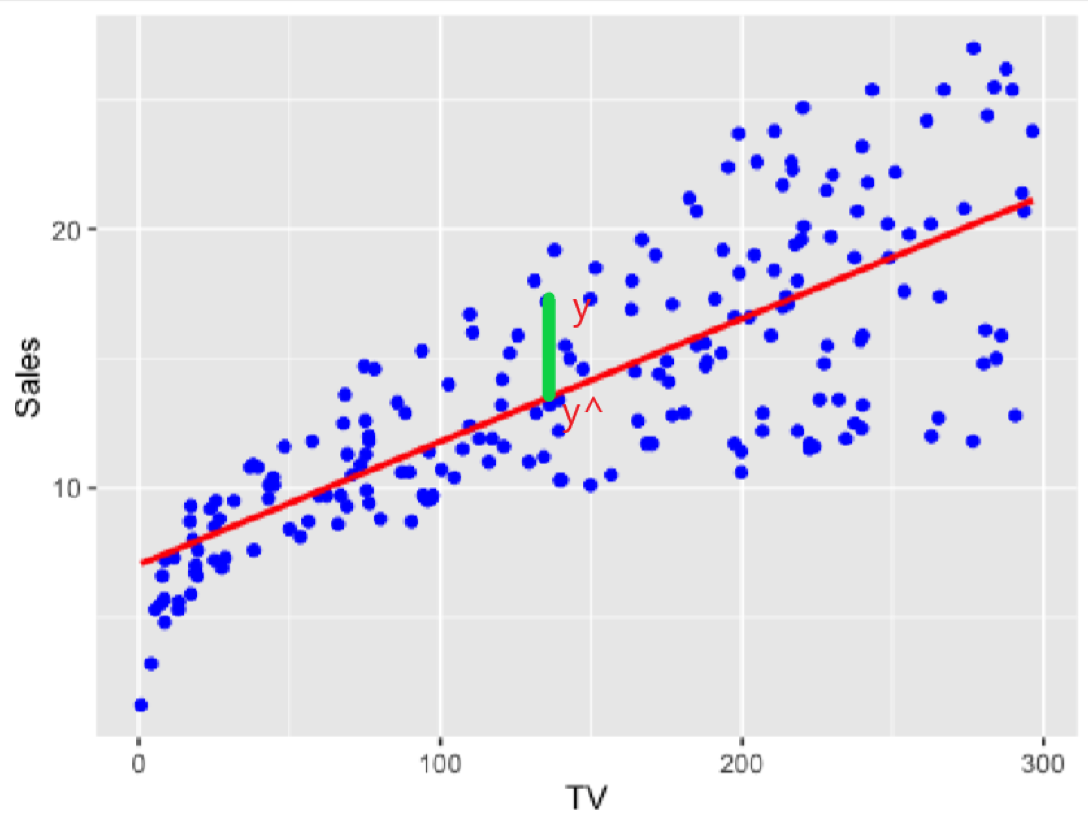
**Regresión Lineal Simple**

El modelo de regresión lineal es una de muchas funciones de estimación. Consiste en asumir que existe una relación lineal entre ambas variables.



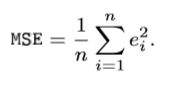
Beta0 es la ordenada en el origen y Beta1 la pendiente de la recta.

El error ei se puede calcular mediante Y menos el Y^ estimado.





Para elimnar los posibles valores negativos lo elevamos al cuadrado y dividimos por número de elementos y usamos el **MSE** (**Mean Squared Error**) o Error Cuadrático Medio para medir.



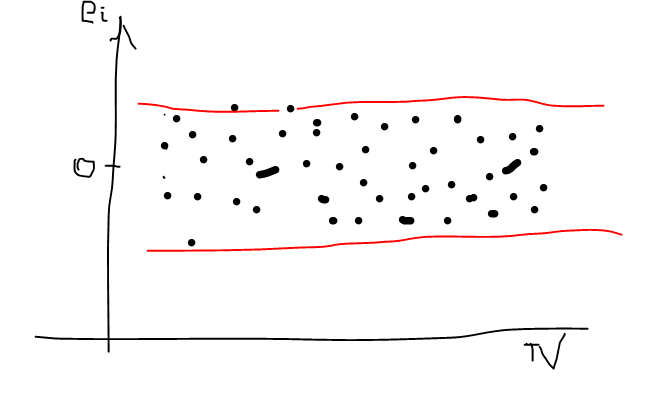
Nos gustaría encontrar el mínimo al cuadrado yi – yÎ, es decir del error. Cuando se quiere hayar el mínimo (o el máximo) de una función, la derivas y la igualas a 0. Así se calculan los **Mínimos Cuadrados Ordinarios (OLS – Ordinary Least Squares)**

la diapositiva 14 sólo explica como resolver matemáticamente los coef beta de la ecuación de la recta, que es la solución a la regresión lineal.

**Estimaciones del error**

Error Residual Estándar (**RSE** – Residual Standard Error). Los propios betas son estimaciones, no tienen solución única, no hay una recta única solución. Y ese error se puede calcular con el chorizote de formula de la página 15.

Si dibujas los residuos debería salir un scatter plot con media = 0 y varianza constante.



**Respuesta significativa**

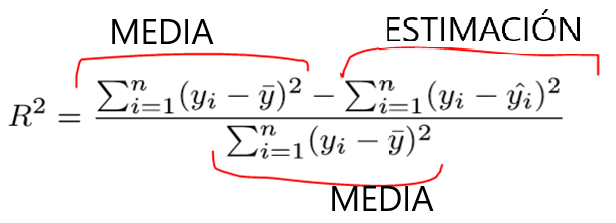
pag.16 se pueden chequear si las betas son más significativas o no, lanzando un test de hipótesis con beta = 0.

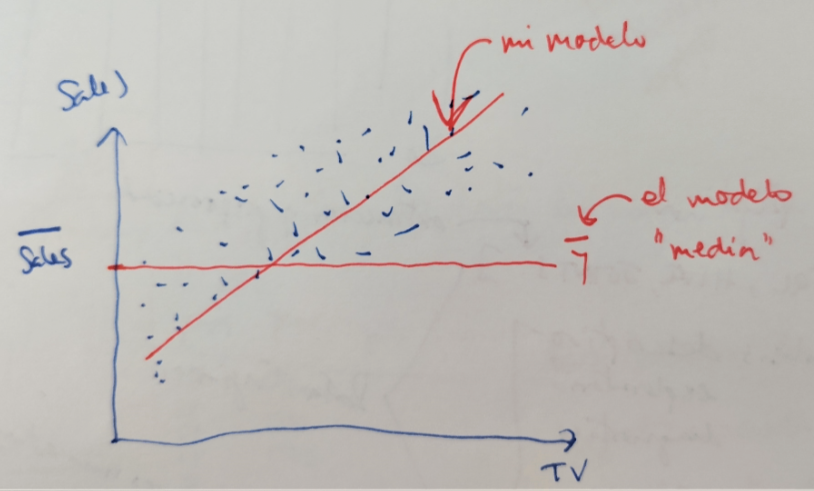
Hay una función de R que ya te hace los test de hipótesis automáticamente. Las librerías tipicas de python no, aunque hay otras que sí.

**Estadístico R2**

pag.17 El estadístico “R2” es otra forma de medir el error. Mide cuanto se aleja mi modelo del modelo “tonto” que es la media, una recta horizontal centrado en la media de “y”.

Cuanto más se acerca a 1 más perfecto es mi modelo. Si vale 0 es que mi modelo es peor que la media!!! osea una mierda. Y si sale negativo es que mi modelo es peor que la media!! osea una puta mierda.





**NOTAS**:

Casualmente R2 mide exactamente lo mismo que el coeficiente de correlación de pierson, para Regresión Lineal Simple. **No sucede así para Regresión Lineal Múltiple**.

**Instrucciones para ajustar modelos lineales en R**

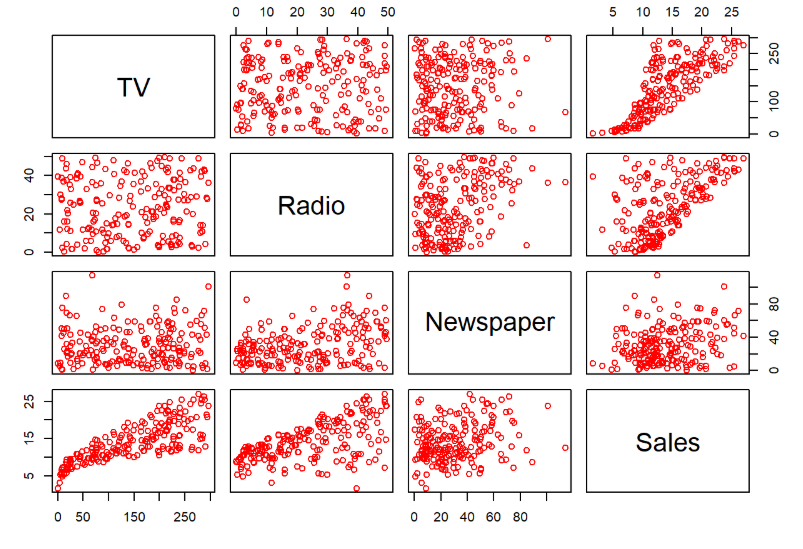
Explicado en fichero *Advertising\_Auto\_Ex.html*

**Explorar datos:**

**glimpse**(datos) <- del paquete dplyr. Es parecido al str() te da cierta información sobre el dataset

**summary**(datos)

**pairs**(datos) <- dibuja matriz de diagramas de dispersión por pares.



COLIANIDADES: correlaciones entre variables relativas. Es malo si vemos correlaciones en la variables cruzadas, es decir, entre Radio y TV o entre Radio y Newspapers. Darían resultados RAROS, habría que sacarlos del modelo. En la imagen de arriba vemos que NO HAY en este caso.

Es aceptable una correlación de hasta 0.8-0.7 entre variables relativas. Si es más quitamos una de ellas.

Lo bueno es que haya correlación entre las variables de entrada y la de salida.

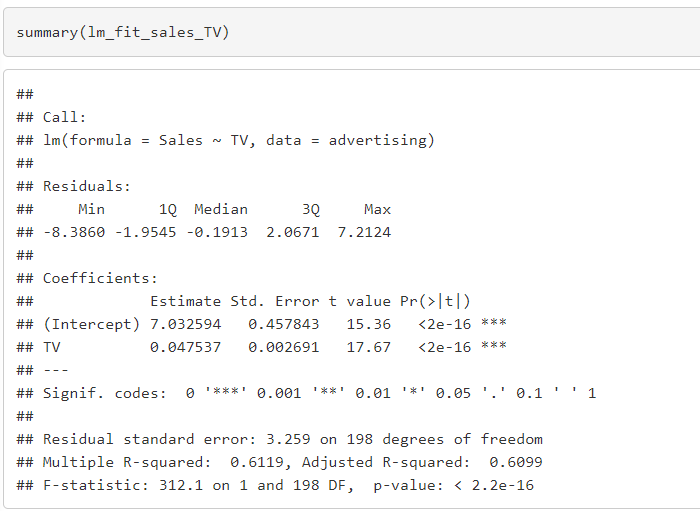
cor() saca una matriz de correlación.

**Ajuste del modelo lineal**

lm(formula, data, subset, weights, ...)

lm = lineal model

Si hago un **summary**() sobre el objeto que me devuelve lm()

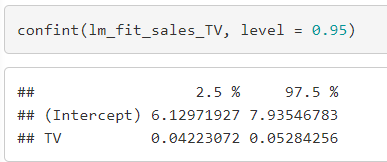


**Residuals:** atributo con información sobre todos los errores para cada muestra.

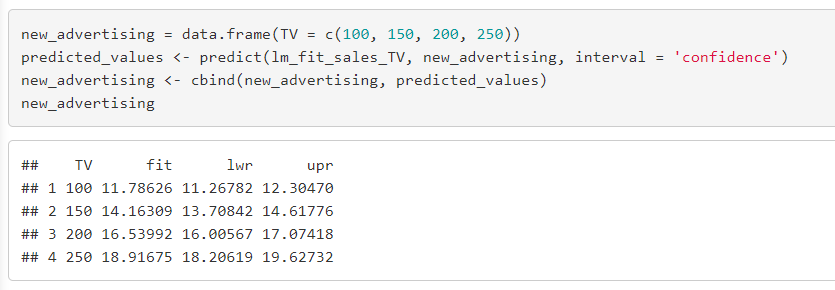
**Pr** = P valor. Cuanto más pequeño es más podemos rechazar la hipotesis nula (osea que beta sea = 0). Quiere decir que si Pr es casi 0, la variable elegida para construir el modelo es significativa, y el modelo es válido.

**Adjusted R-squared** es el que hay que mirar por defecto para medir el error. Cuando sólo tengo una variable es similar a R-squared. Cuando tengo multivariable no tiene por qué y debemos mirar el *“adjusted”*.

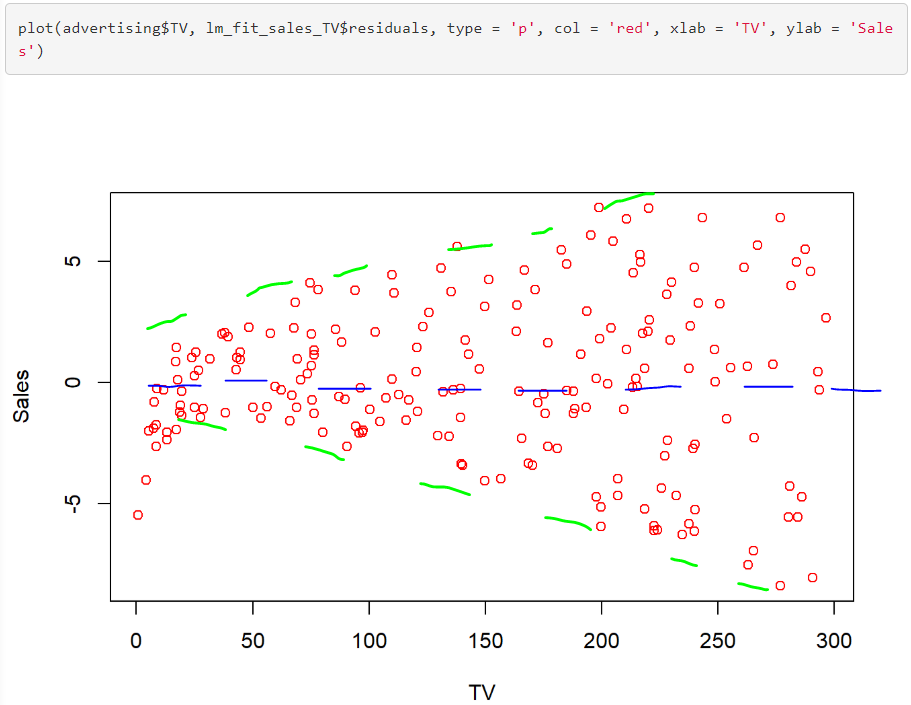
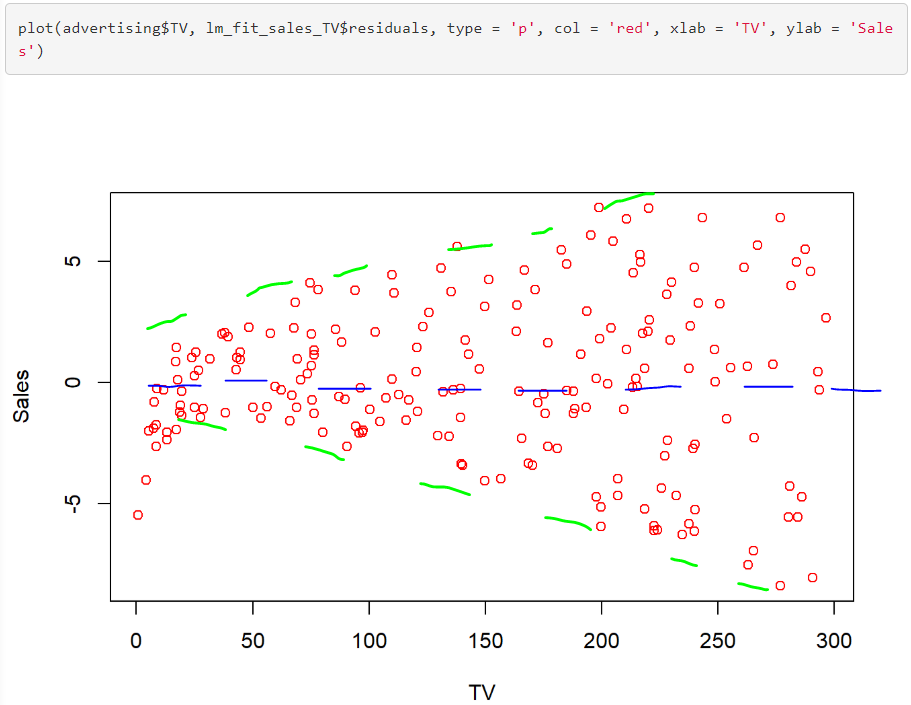
**confint**() <- Se pueden obtener **intervalos de confianza para los coeficientes:**



**predict**() <- hace predicciones con datos nuevos basados en el modelo de regresión lineal simple.



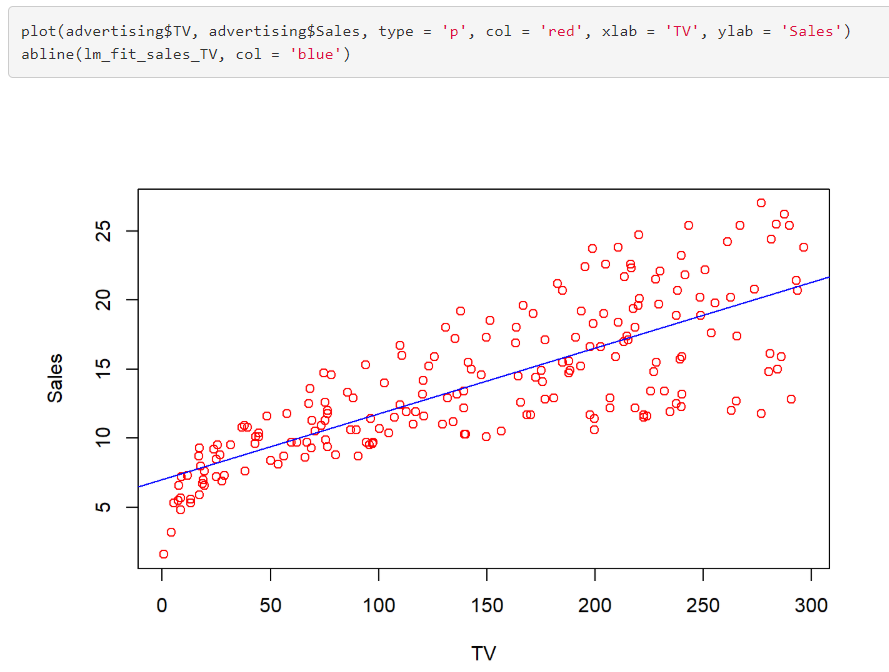
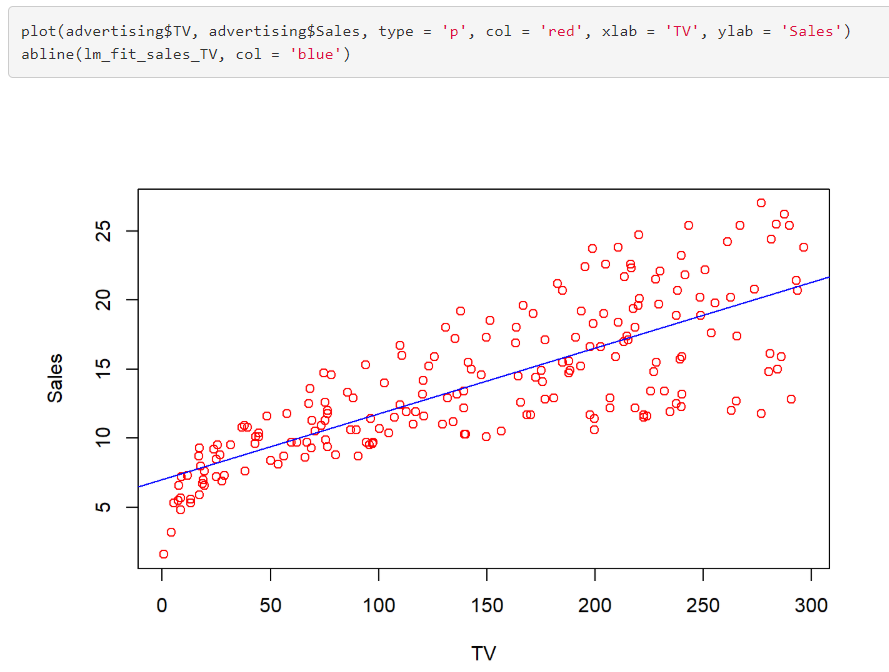
Podemos **representar** gráficamente los **residuos** **del** **modelo**:



En este caso la media se aproxima a cero, lo cual está bien.

Pero la variancia no es constante, se dispersa según crece TV. Esto significa que debemos añadir más información al modelo, tal vez alguna variable más.

Podemos **representar** gráficamente el **modelo** **sobre** **los** **datos** de entrenamiento:



Libro bueno y gratis en pdf sobre statistics in R:

an introduction to statistical learning with applications in r

Q-Q plot es sacar los cuantiles de una distribución y comprarlos con los cuantiles de otra distribución. Se usa para comparar distribuciones y ver si se ajustan a una distribución normal.

los logaritmos generalmente linearizan todas las relaciones que no son lineales. Se usan mucho para transformar las variables y ajustar una regresion lineal.

**NOTAS**:

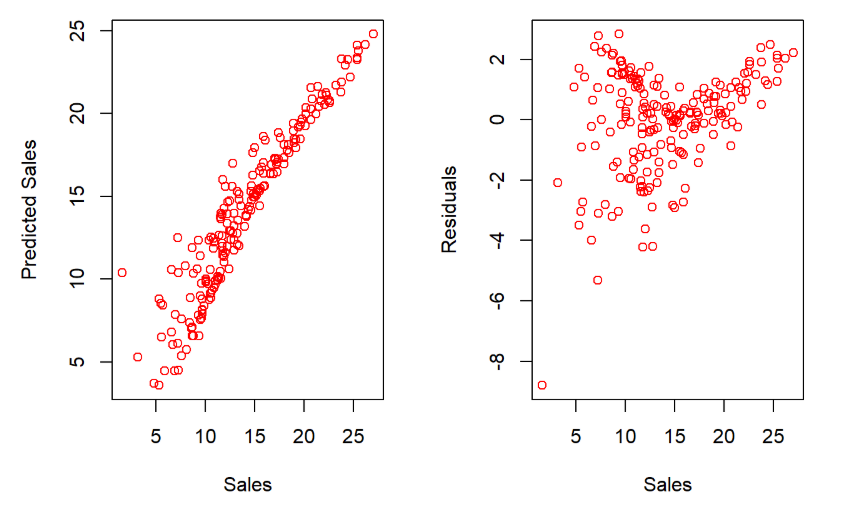
Para más detalle sobre cómo desarrollar un modelo de Regresión Lineal Simple mirar el archivo de ejercicios resuelto ***Advertising\_Auto\_RESUELTOS.html***

**Modelo Lineal Múltiple**

Se ajusta de manera muy similar el modelo simple pero se meten varias variables en lm().

el resultado no es una recta, es un plano (con dos variables) u otras cosas más complejas con más dimensiones, por eso no se suelen representar en el espacio real.

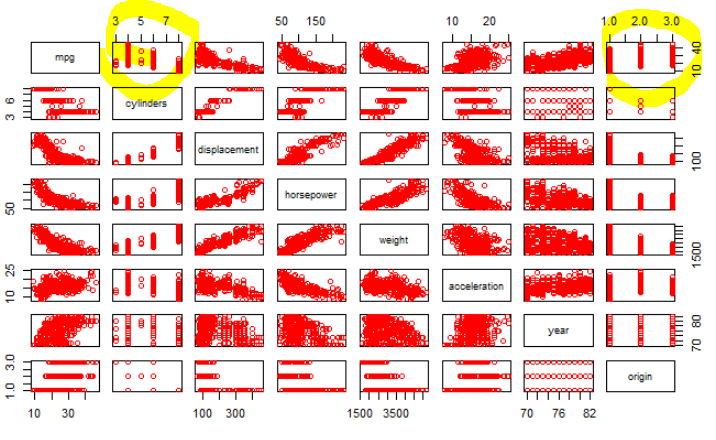
Lo que se representa es la variable objetivo frente a los valores fiteados (=valores predichos) y residuos frente a variable objetivo.



Saldrá una diagonal de puntos, se asemejará más a una recta cuanto más exacto sea el modelo.

Una **variable proxy** es una variable que está relacionada con la variable target. NO DEBEMOS NUNCA USARPROXIES PARA CALCULAR UN MODELO PREDICTIVO. Porque es como si estuvieran replicando la variable target y van a aumentar mucho la precisión del modelo ERRONEAMENTE.

OJO: Cuando una de las variables es discreta no es de fiar usar correlación para ver si la eliges o no. Mirar grafico abajo entre mpg vs cylinders o mpg vs origin



**DescTools**

libbrary con funcion **Desc**() que te saca mazo de información sobre un dataset. Es como un summary mejorado.

NOTA: hay que quitar variables que tienen mucha correlación con otras variables explicativas, porque pueden interferir en el resultado.