



K SCHOOL

Máster Data Science

Barcelona, 2019.
Henry Navarro





About me

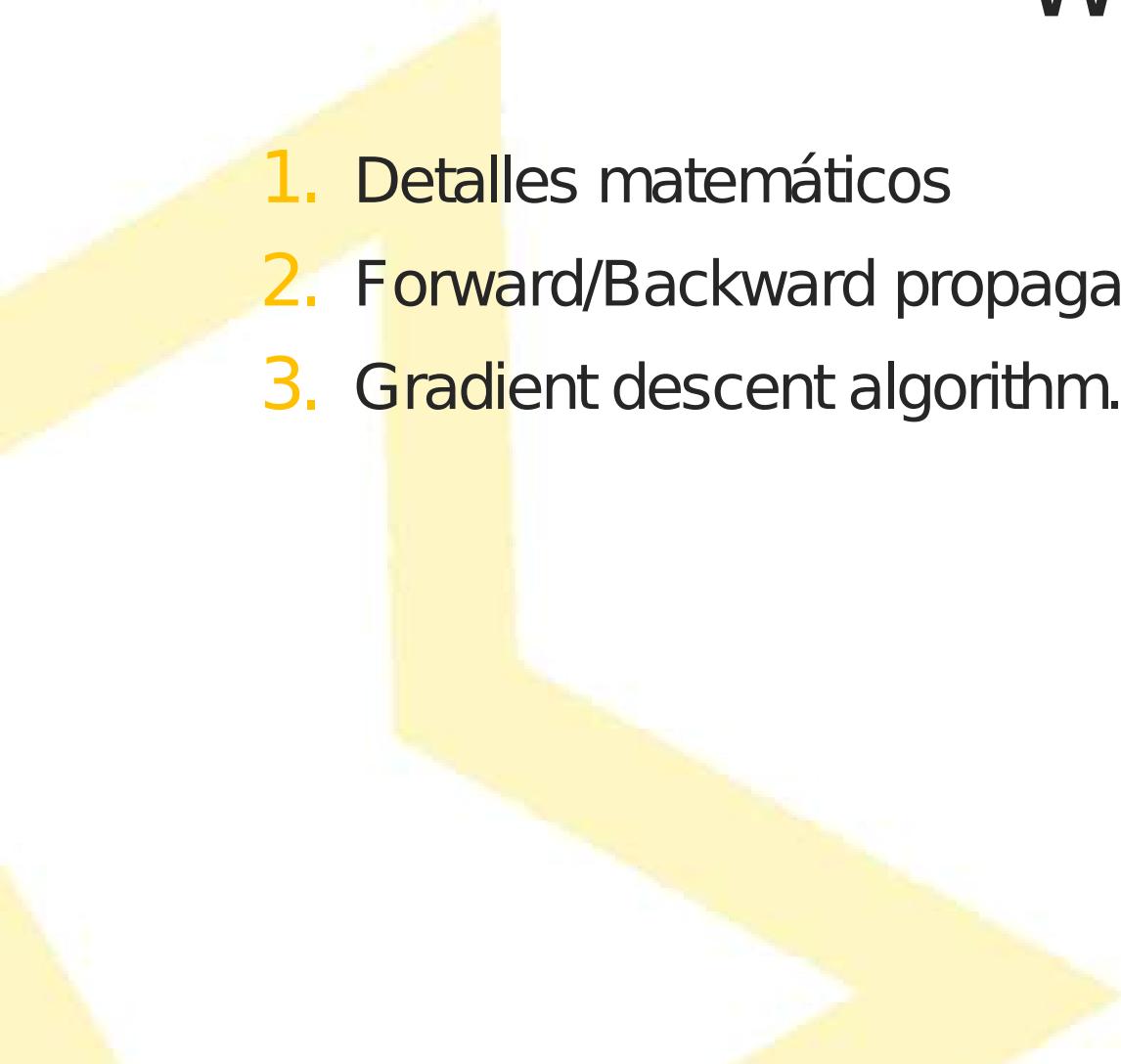
- LinkedIn: 
- Estudios:
 - Licenciado en Matemática – UCV.
 - Máster en Ingeniería Matemática – UC3M.
 - Estudios de doctorado en Machine Learning.
- Experiencia profesional:
 - Ahora: Lead Data Scientist – Altran Innovation.
 - Antes: Data Scientist – Equifax, Solutio, Enefgy.
 - Data Analyst – Ministerio de Justicia (Venezuela).
- Docencia:
 - Profesor Universidad Carlos III de Madrid – Grupo ML4DS & GISC. (<https://goo.gl/9eeHAz>)
 - Profesor Escuela de Organización Industrial.
 - Profesor Álgebra, Cálculo, Estadística en Universidad Central de Venezuela

Contenido

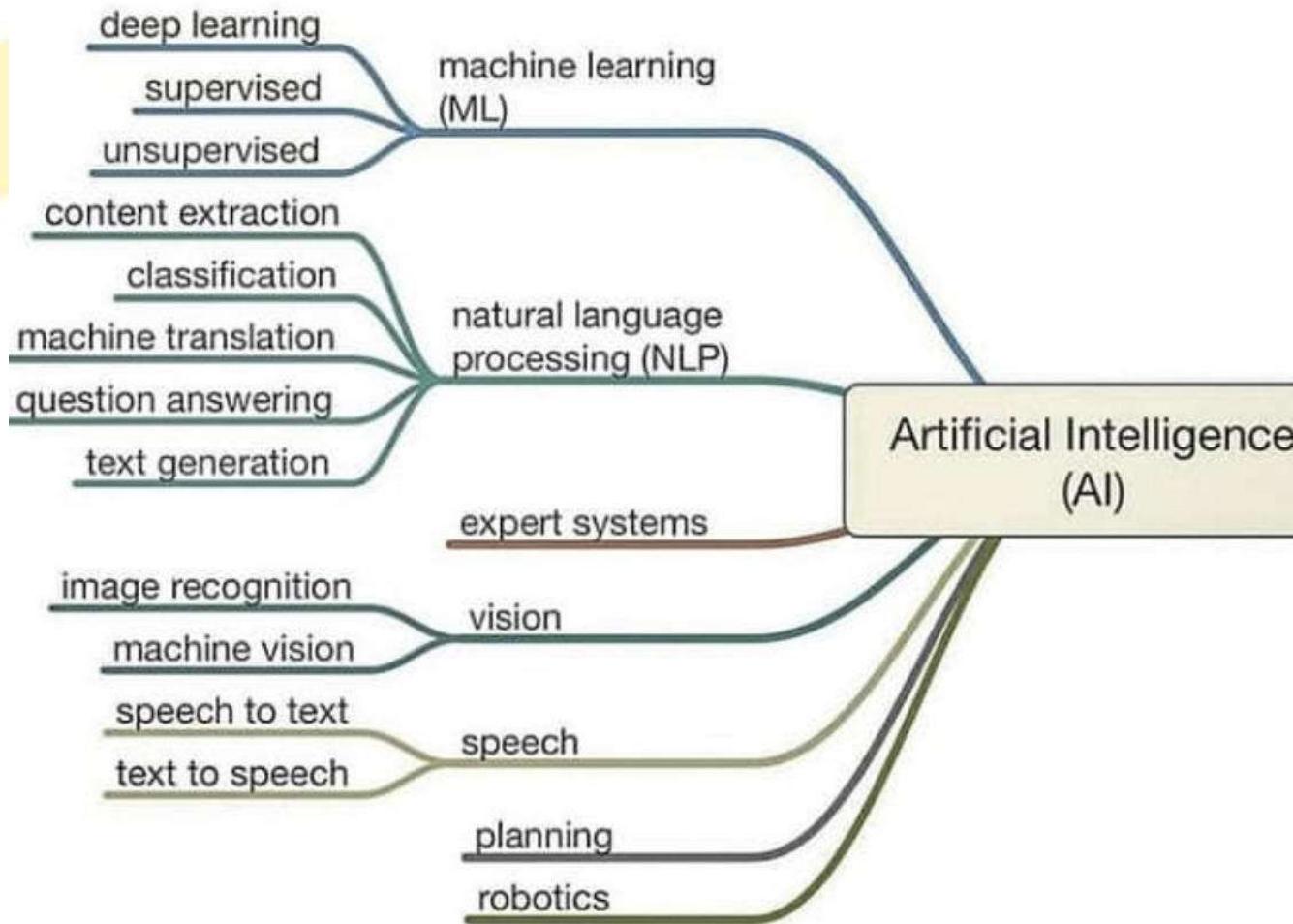
1. Conceptos previos:
 - Dataset de train y test.
 - Overfitting.
 - Matriz de confusión.
 - Cross - Validation
2. Linear regression.
3. Logistic regression.
4. Polynomial regression.
5. Nonlinear Models:
 - Decision trees.
 - Random forest.
 - K - nearest neighbors

1. Intro to Deep learning y computer vision
 - ¿cómo pasar de una imagen a una matriz?
 - Imagen en escala de grises.
 - Imagen en RGB.
 - Optical Character Recognition.
 - Proyecto de Machine Learning para un clasificador de imágenes (gatos) ☺.

We won't study...

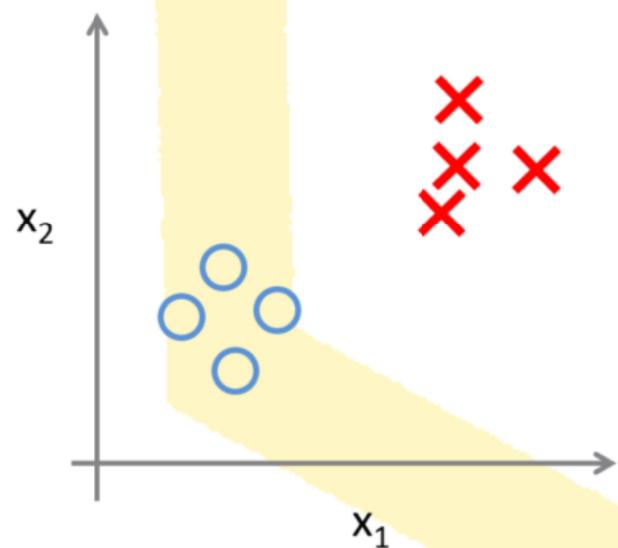
- 
1. Detalles matemáticos
 2. Forward/Backward propagation.
 3. Gradient descent algorithm.

Intro supervised learning

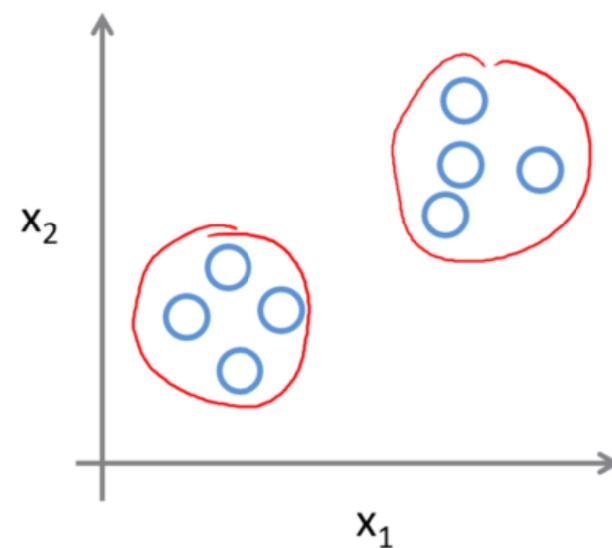


Supervised vs. Unsupervised learning

Supervised Learning



Unsupervised Learning



Variable respuesta/
objetivo/dependien-
te



supervised learning

y	x1	x2	x3	x4	x5	x6
0	186.176767	32.01013	7.389389	56.68727	171.3361	18.03844
1	159.659374	-95.06651	-83.210200	155.31220	-149.0119	-183.90314
0	44.307132	-167.88587	-90.023000	124.17956	170.8277	30.37569
0	129.380781	-83.71101	193.529927	193.97078	135.2245	-157.56599
0	-7.236501	150.92669	-75.665873	58.89800	-114.4337	-58.16047
0	-13.191041	51.07507	168.874093	-73.05704	-179.1995	-178.97354

unsupervised learning

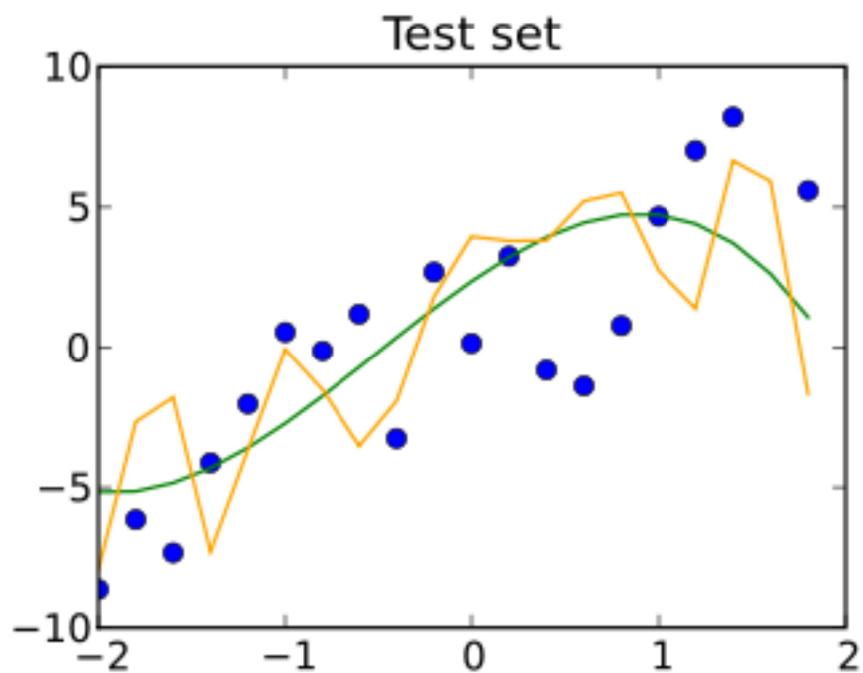
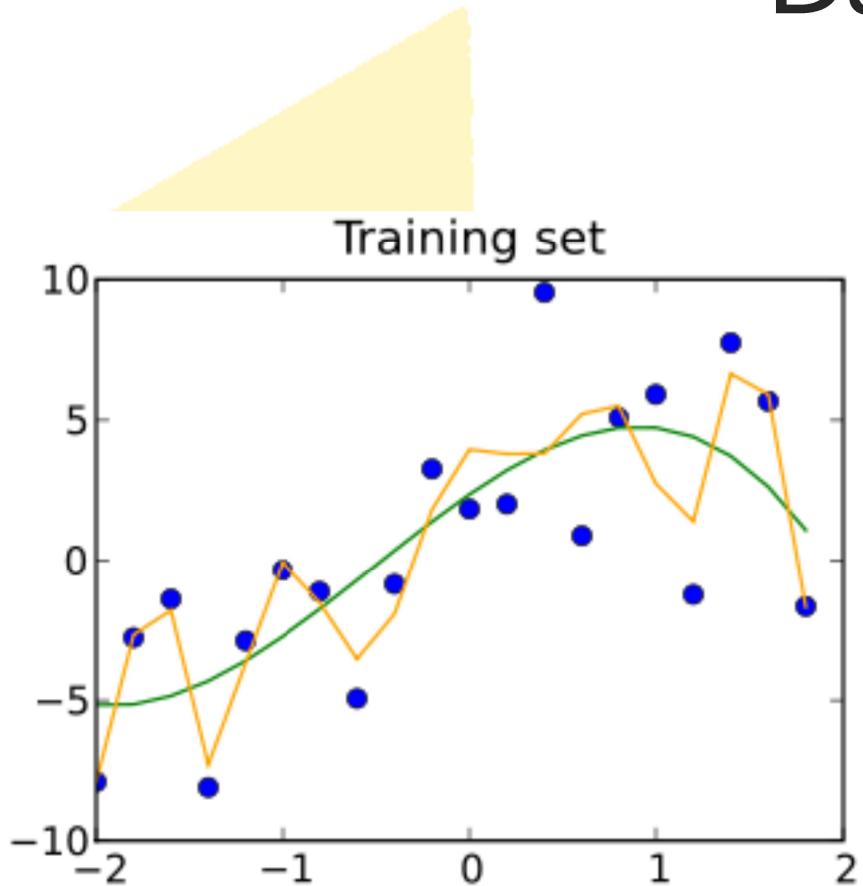
	x1	x2	x3	x4	x5	x6
186.176767	32.01013	7.389389	56.68727	171.3361	18.03844	
159.659374	-95.06651	-83.210200	155.31220	-149.0119	-183.90314	
44.307132	-167.88587	-90.023000	124.17956	170.8277	30.37569	
129.380781	-83.71101	193.529927	193.97078	135.2245	-157.56599	
-7.236501	150.92669	-75.665873	58.89800	-114.4337	-58.16047	
-13.191041	51.07507	168.874093	-73.05704	-179.1995	-178.97354	

No tenemos
variable

Conceptos previos

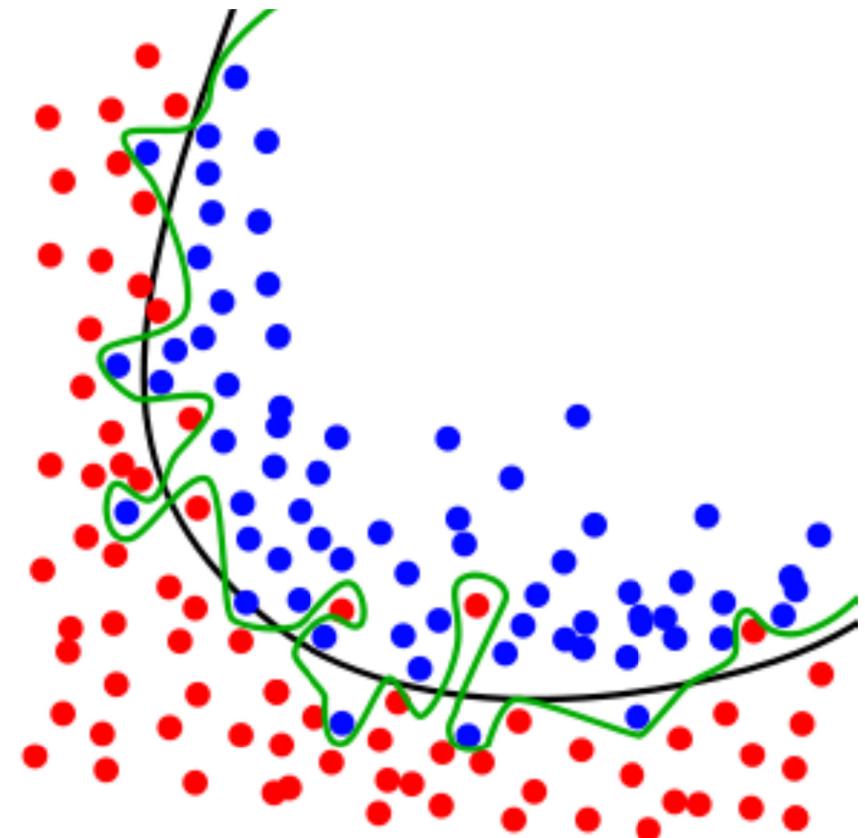
- Dataset de train y test.
- Overfitting.
- Matriz de confusión.
- Cross - Validation

Dataset de train y test.



Overfitting

- Es el efecto de sobreentrenar un algoritmo de aprendizaje con unos ciertos datos para los que se conoce el resultado deseado.
- El algoritmo de aprendizaje debe alcanzar un estado en el que será capaz de predecir el resultado en otros casos a partir de lo aprendido con los datos de entrenamiento, generalizando para poder resolver situaciones distintas a las acaecidas durante el entrenamiento.
- Cuando un sistema se entrena demasiado (se sobreentrena) o se entrena con datos extraños, el algoritmo de aprendizaje puede quedar ajustado a unas características muy específicas de los datos de entrenamiento que no tienen relación con la función objetivo.



Matriz de confusión

- Observado vs. Predicción

		Actual class	
		Cat	Non-cat
Predicted class	Cat	5 True Positives	2 False Positives
	Non-cat	3 False Negatives	17 True Negatives

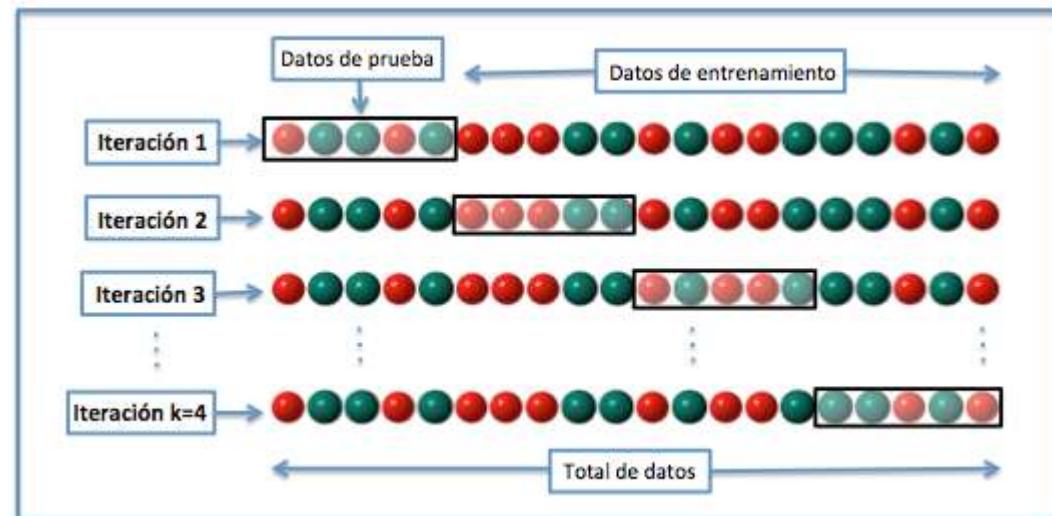
Matriz de confusión

- Observado vs. Predicción

		True condition	
		Condition positive	Condition negative
		Total population	
Predicted condition	Predicted condition positive	True positive, Power	False positive, Type I error
	Predicted condition negative	False negative, Type II error	True negative

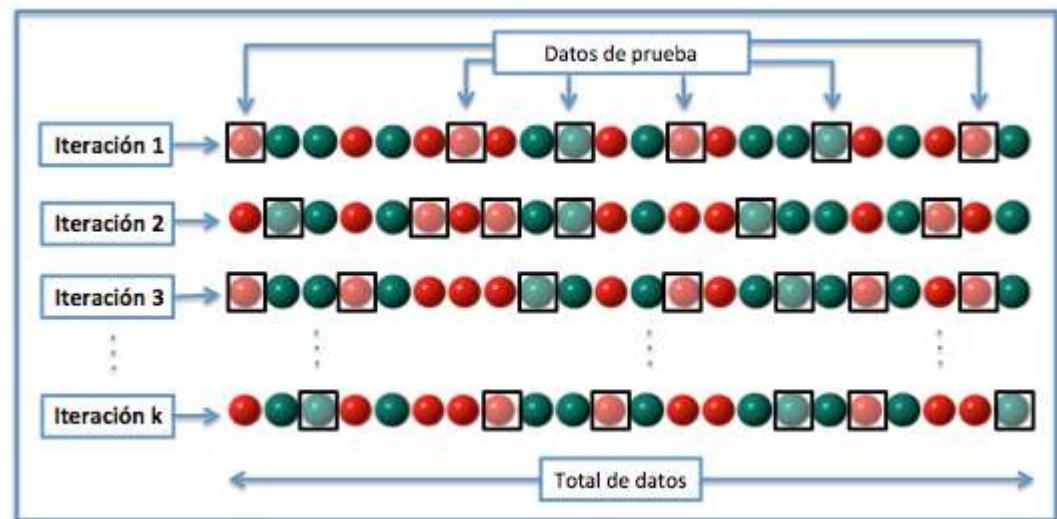
K-fold Cross Validation

- En la *K-fold cross-validation* los datos de muestra se dividen en K subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto ($K-1$) como datos de entrenamiento. El proceso de validación cruzada es repetido durante k iteraciones, con cada uno de los posibles subconjuntos de datos de prueba.
- Este método es muy preciso puesto que evaluamos a partir de K combinaciones de datos de entrenamiento y de prueba, pero aun así tiene una desventaja, y es que, a diferencia del método de retención, es lento desde el punto de vista computacional.



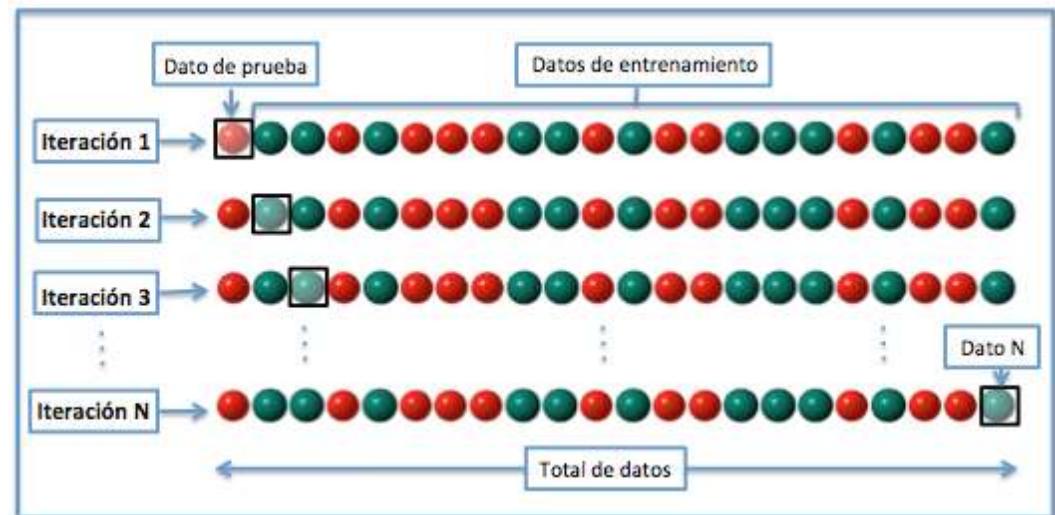
Random Cross Validation

- Este método consiste al dividir aleatoriamente el conjunto de datos de entrenamiento y el conjunto de datos de prueba. Para cada división la función de aproximación se ajusta a partir de los datos de entrenamiento y calcula los valores de salida para el conjunto de datos de prueba.
- La ventaja de este método es que la división de datos entrenamiento-prueba no depende del número de iteraciones. Pero, en cambio, con este método hay algunas muestras que quedan sin evaluar y otras que se evalúan más de una vez, es decir, los subconjuntos de prueba y entrenamiento se pueden solapar.

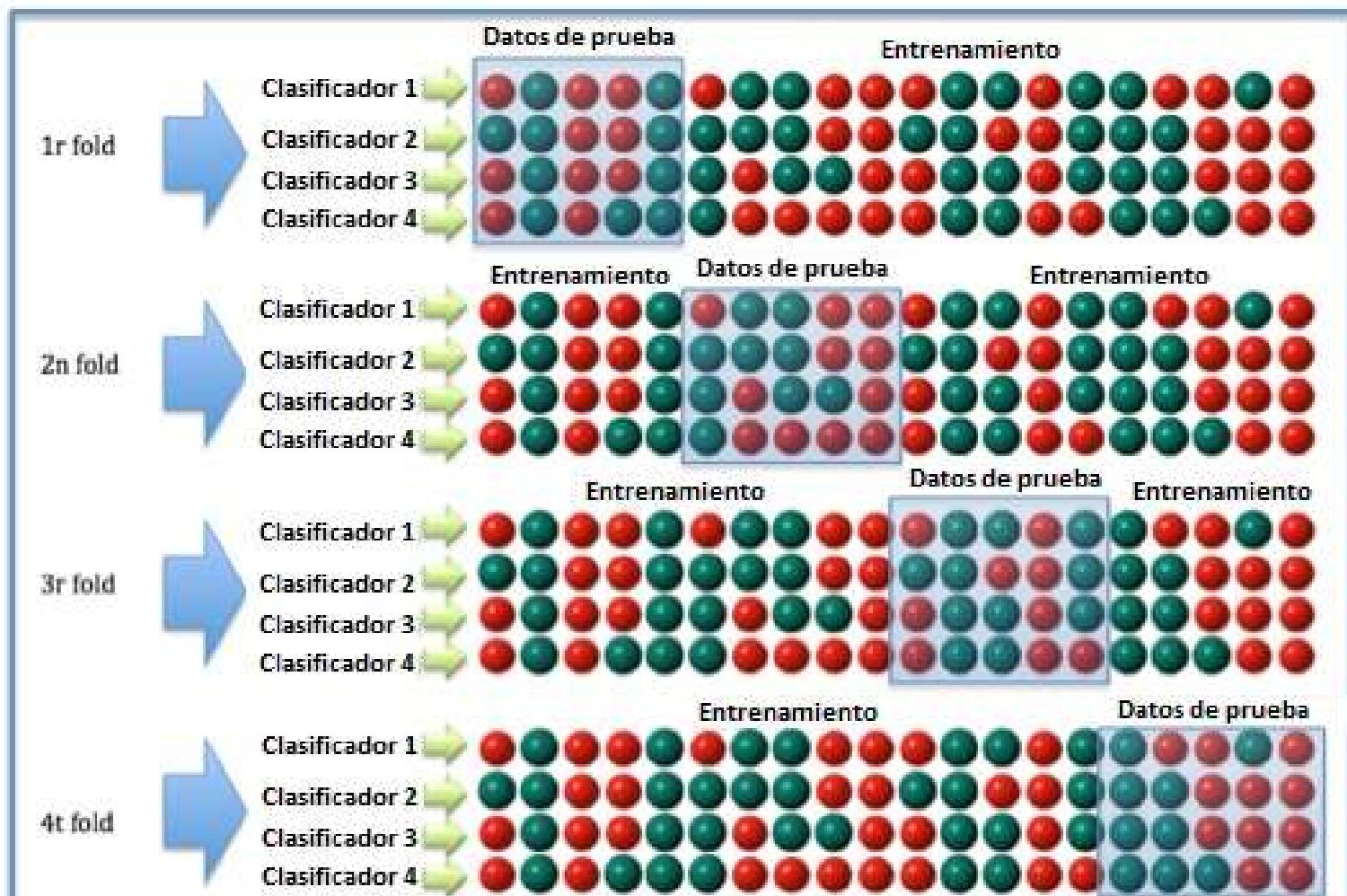


Leave-one-out cross-validation

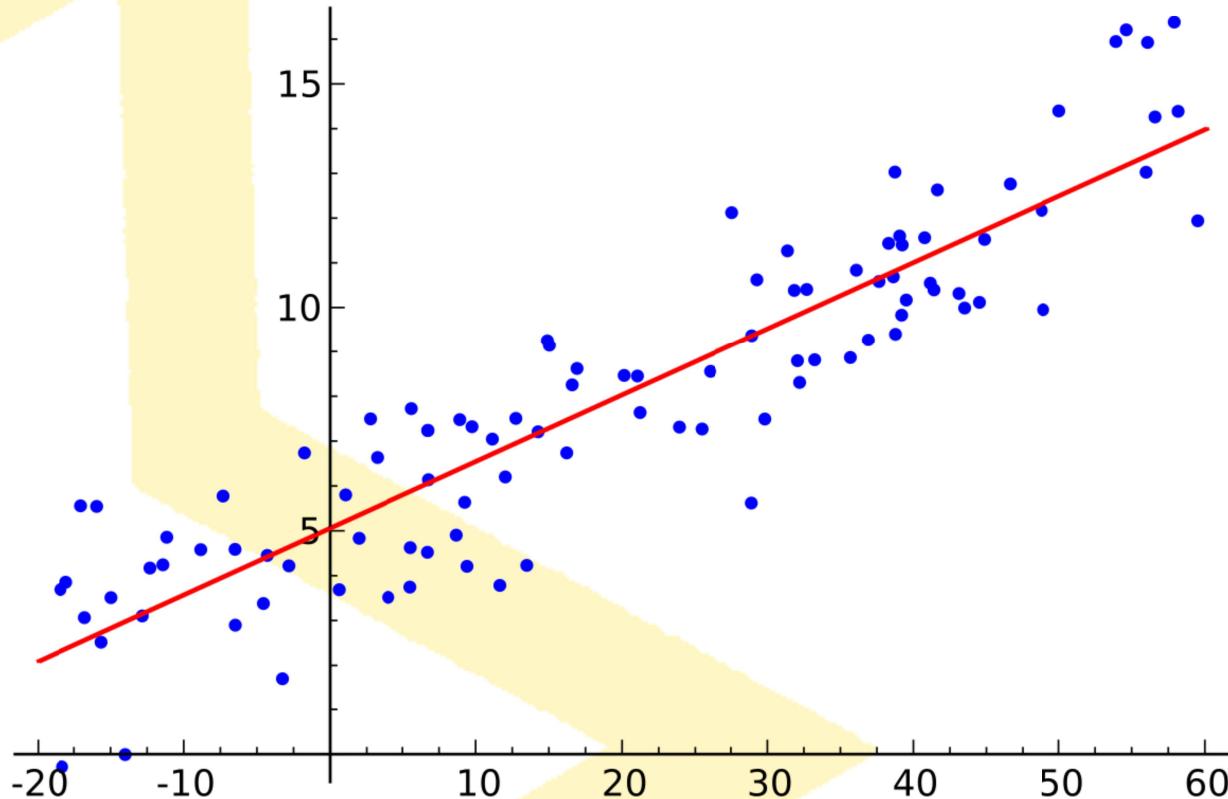
- La Leave-one-out cross-validation implica separar los datos de forma que para cada iteración tengamos una sola muestra para los datos de prueba y todo el resto conformando los datos de entrenamiento.
- La evaluación viene dada por el error, y en este tipo de validación cruzada el error es muy bajo, pero en cambio, a nivel computacional es muy costoso, puesto que se tienen que realizar un elevado número de iteraciones, tantas como N muestras tengamos y para cada una analizar los datos tanto de entrenamiento como de prueba.



Cross-Validation ejemplo

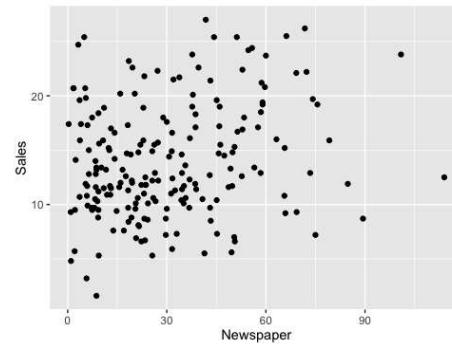
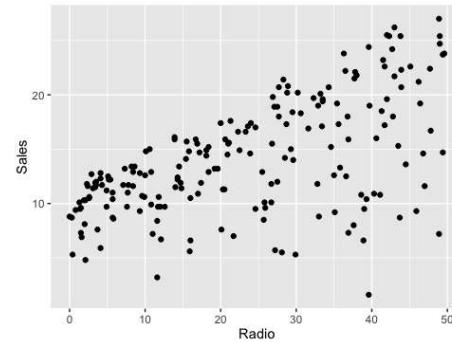
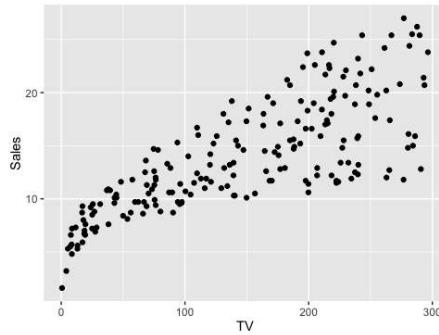


La famosa, simple pero poderosa
regresión lineal...



Simple Linear Regression

Ejemplo: Nuestro cliente es una compañía que desea mejorar las ventas de un determinado producto (en miles de unidades). Para conseguirlo, nos proporcionan un set de datos que contiene las ventas del producto en 200 mercados diferentes, junto con el presupuesto de publicidad en televisión, radio y periódicos en cada uno de tales mercados (en miles de dólares).



Naturalmente, nuestro cliente no tiene la capacidad de incrementar las ventas de su producto directamente, pero sí tiene la capacidad de decidir qué presupuesto dedicar a cada uno de los canales de publicidad.

Simple Linear Regression

- Sean un conjunto de datos, de las variables $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ un conjunto de datos de las variables El modelo tiene como objetivo hallar dos estimadores tales que hallar dos estimadores β_0, β_1 tal que

Donde para todo .

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Donde $\beta_i \in \mathbb{R}$ para todo $i = 1, 2$.
- Es necesario asumir que existe una relación lineal entre X e Y como sigue
- Es necesario asumir que existe una relación lineal entre X e Y como sigue

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Donde se considera el error, y se asume que es un vector con distribución

Donde ε se considera el error, y se asume que es un vector con distribución

- Aunque σ es desconocida, se suele estimar por el **error residual estándar**

- Aunque σ es desconocida, se suele estimar por el **error residual estándar**

$$RSE = \sqrt{\frac{1}{n-2} \sum_{i=1}^n e_i^2}$$

¿Cómo sabemos si nuestro modelo está bien ajustado?

- La medida absoluta más habitual para estimar la precisión del modelo es el error cuadrático medio

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- Sin embargo, se emplea un estadístico independiente de unidades

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{SSE}{SST}$$

- Representa la proporción de variabilidad de la variable respuesta que está explicada por el modelo.

- Es fácil ver, luego de varias operaciones matemáticas que $R^2 = \rho^2$, donde ρ es el coeficiente de correlación

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

¿Cómo sabemos si nuestro modelo está bien ajustado?

- La medida absoluta más habitual para estimar la precisión del modelo es el error cuadrático medio

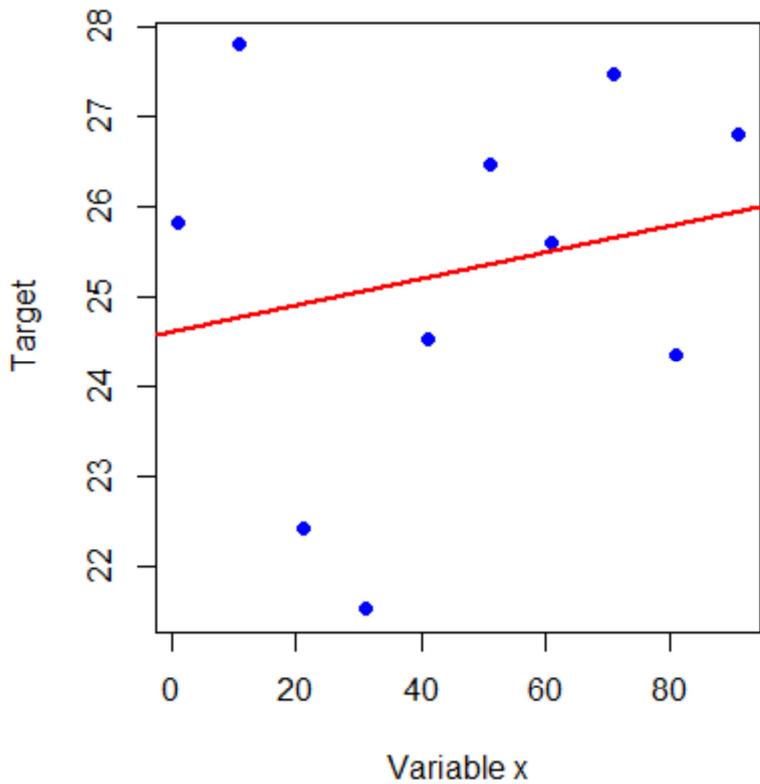
$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- Sin embargo, se emplea un estadístico independiente de unidades
- Sin embargo, se emplea un estadístico independiente de unidades
- Representa la proporción de variabilidad de la variable respuesta que está explicada por el modelo.
- Es fácil ver, luego de varias operaciones matemáticas que donde es el coeficiente de correlación.
- Es fácil ver, luego de varias operaciones matemáticas que $R^2 = \rho^2$, donde ρ es el coeficiente de correlación

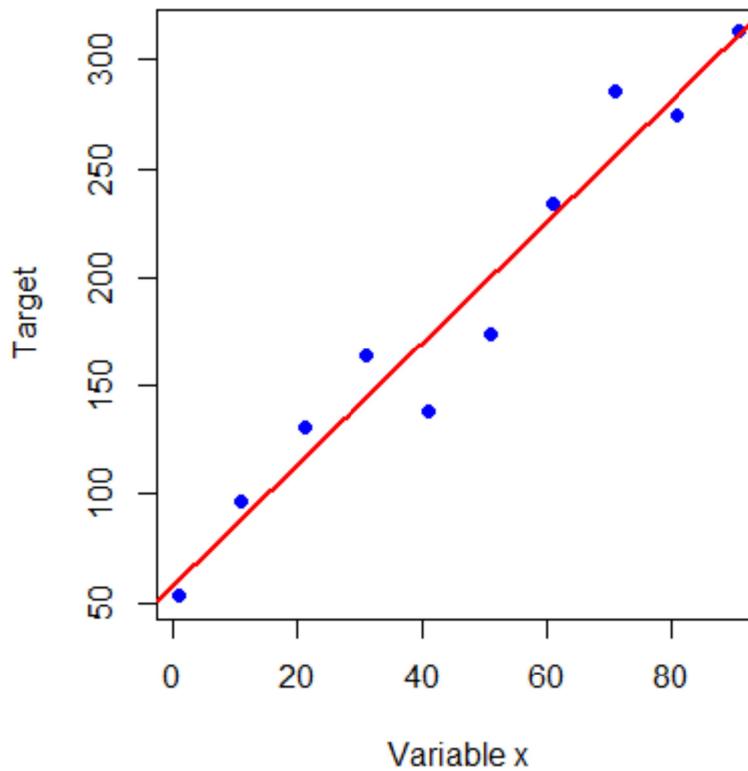
$$\rho^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

El Famoso²

LM with $R^2=0.074$



LM with $R^2=0.941$



Código para imagen anterior

```
rm(list=ls())
par(mfrow=c(1,2))

#Fake data
x<-seq(1,100,10)
y<-runif(10,20,30)

#Model
my_lm<-lm(y~x)

adj_rsquare<-round(summary(my_lm)$adj.r.squared,3)
#Plot
plot(x,y,col="blue",pch=16,ylab="Target",xlab="Variable x",
      main=paste0("LM with R^2=",adj_rsquare))
abline(my_lm,col="red",lwd=2)

#Fake data
y<-3*x+1+runif(10,0,100)

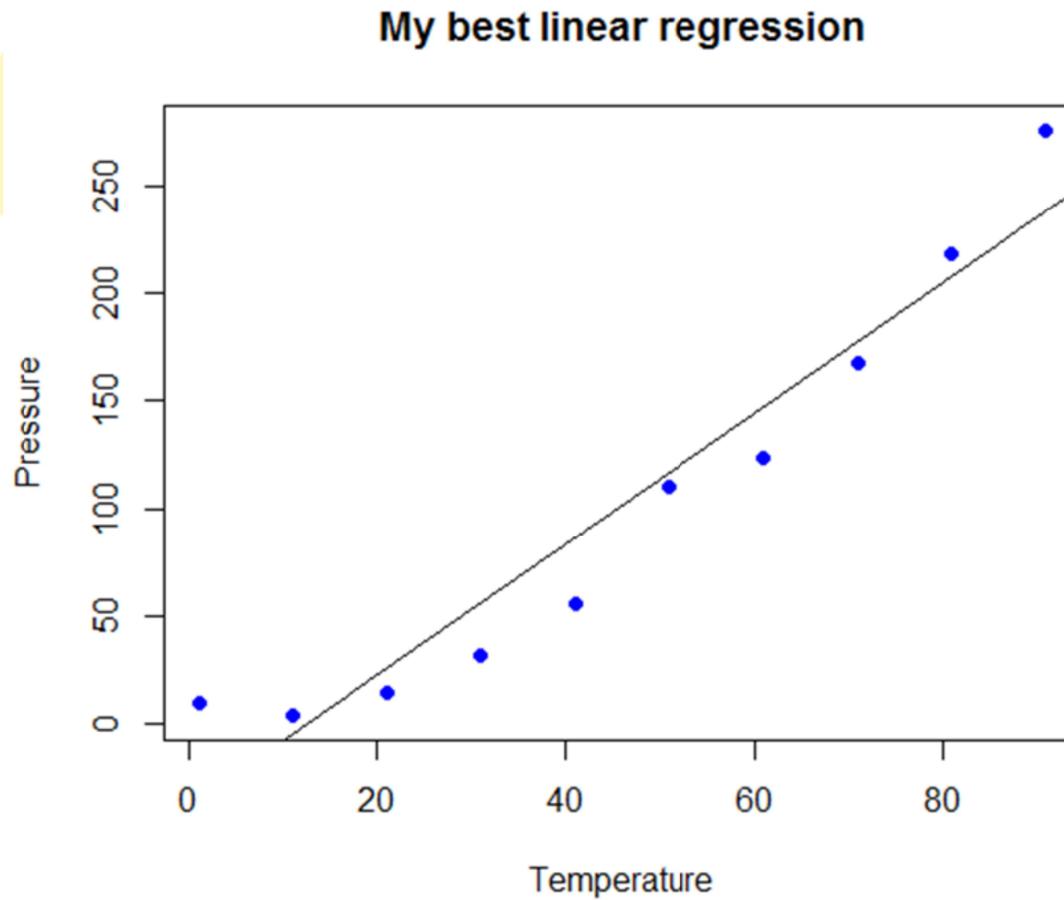
#Model
my_lm<-lm(y~x)

adj_rsquare<-round(summary(my_lm)$adj.r.squared,3)
#Plot
plot(x,y,col="blue",pch=16,ylab="Target",xlab="Variable x",
      main=paste0("LM with R^2=",adj_rsquare))
abline(my_lm,col="red",lwd=2)
```

Residuales

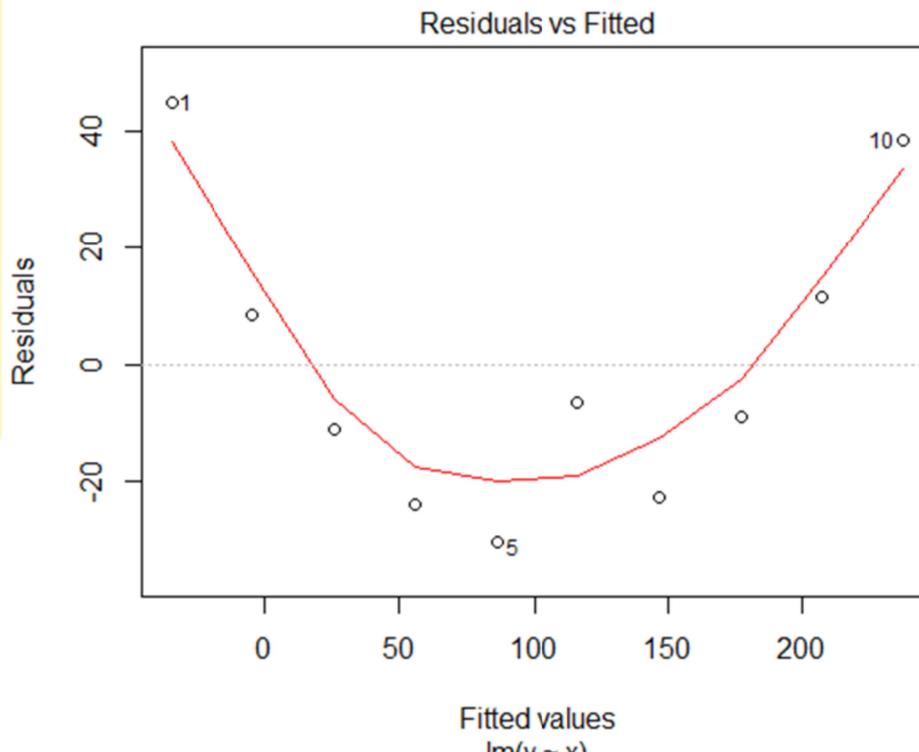
- Puedes tener un R^2 bastante bueno en tu modelo, pero no nos apresuremos a sacar conclusiones aquí.
- Veamos un ejemplo...

Residuales



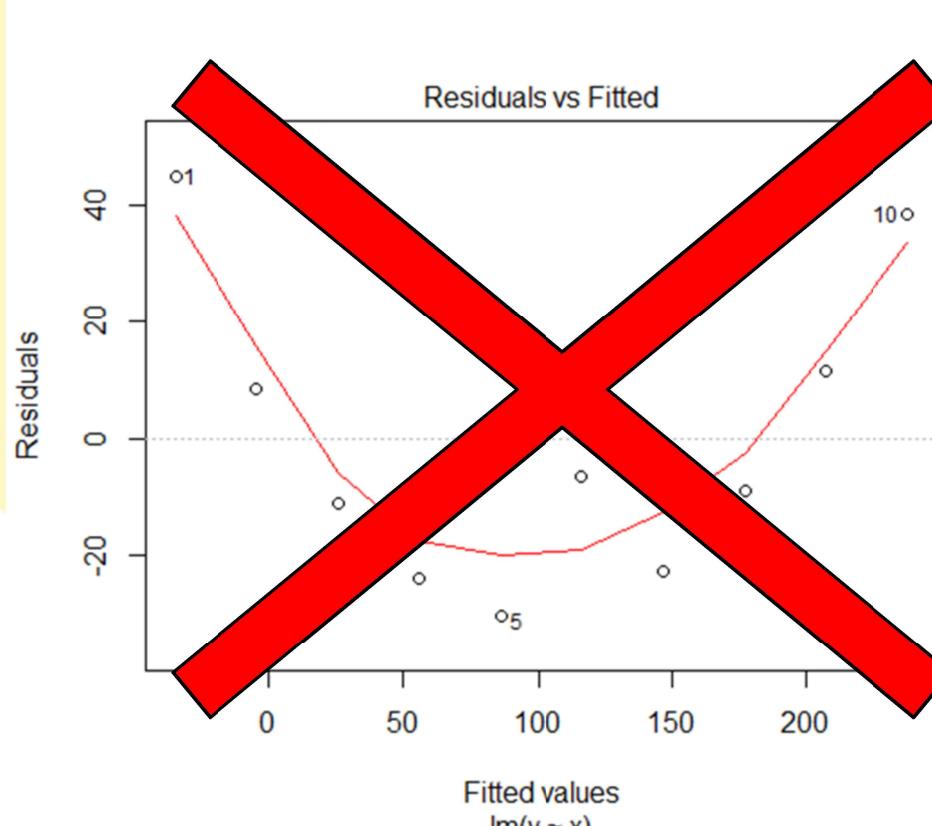
Residuales

- Idealmente, cuando grafiques los residuos, deberían parecer aleatorios. De lo contrario, significa que tal vez hay un patrón oculto que el modelo lineal no está considerando.



Residuales

- Idealmente, cuando grafiques los residuos, deberían parecer aleatorios. De lo contrario, significa que tal vez hay un patrón oculto que el modelo lineal no está considerando.



Código para imagen anterior

```
#Fake data
x<-seq(1,100,10)
y<-x^2/30
y[1]<-10
y[6]<-110

#Plot
plot(x,y,col="blue",pch=16,ylab="Pressure",
xlab="Temperature",main="My best linear regression")

#Model
my_lm<-lm(y~x)
abline(my_lm)
summary(my_lm)
plot(my_lm)
```

El Falso R^2

```
call:  
lm(formula = y ~ x)  
  
Residuals:  
    Min      1Q  Median      3Q     Max  
-34.239 -9.114   3.364  12.589  29.285  
  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  58.0034   12.5605   4.618  0.00171 **  
x             2.7879    0.2316  12.037 2.09e-06 ***  
---  
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1  
  
Residual standard error: 21.04 on 8 degrees of freedom  
Multiple R-squared:  0.9477,    Adjusted R-squared:  0.9411  
F-statistic: 144.9 on 1 and 8 DF,  p-value: 2.094e-06
```

- En el rectángulo rojo, observe que hay dos R^2 diferentes, uno múltiple y uno ajustado.
- Un problema con este R^2 es que no disminuye a medida que se agregan más variables al modelo. Puede seguir aumentando a medida que el modelo sea más complejo, incluso si estas variables no agregan ningún poder predictivo.

Multiple Linear Regression

- El modelo busca como objetivo hallar la siguiente ecuación

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Donde para todo $i = 1, \dots, n$.

- La bondad de ajuste se puede medir igual que en el caso univariante, mediante el R^2 . Sin embargo, se suele utilizar el R^2_a que tiene en cuenta el número de predictores empleados:

$$R_a^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1}$$

- Para contrastar si ha una relación lineal entre la variable respuesta y los predictores se realiza el contraste de hipótesis
- Para contrastar si la relación lineal entre la variable respuesta y los predictores se realiza el contraste de hipótesis

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_a: \text{al menos un } \beta_j \text{ es distinto}$$

- Se usa el estadístico
- Se usa el estadístico

$$\frac{n - p - 1}{F} \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Que sigue una distribución $F_{p, n - p - 1}$.

Que sigue una distribución $F(p, n - p - 1)$.

Selección de variables

Asumiendo que la relación entre x y y es aproximadamente lineal y que las predicciones obtenidas por el modelo de regresión lineal serían buenas si ésta no lo es. Esto es falso. Por esto, es necesario construir modelos que empleen pocas variables para que el error en las predicciones sea pequeño.

Por esta razón presentamos los siguientes estadísticos:

- , con .
- $Std. Error = \sqrt{MSE} = \sqrt{\frac{SSE}{n-p}}$, con $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$.
- , con
- F con el máximo valor de la función de verosimilitud para el modelo estimado (ver [esto](#) y [esto](#)) .
- $AIC = 2p - 2 \ln(L)$, con L el máximo valor de la función de verosimilitud para el modelo estimado (ver [esto](#) y [esto](#)) .
- $BIC = p \ln(n) - 2 \ln(L)$, con n tamaño de la muestra.
- .
- $t - statistic = \frac{\beta \text{ coeff}}{Std. Error}$.
- $\Pr(> |t|) = p - value$. Podemos interpretar el como: Un mayor indica que es menos probable que el coeficiente no sea igual a cero. Entonces, cuanto mayor sea , mejor. es la probabilidad de obtener un tan alto o más alto que el valor observado cuando la hipótesis nula (el coeficiente es igual a cero o que no hay relación) es verdadera. Entonces, si es bajo, los Coeficientes son significativos (significativamente diferentes de cero). Si es alto, los coeficientes no son significativos. Podemos interpretar el $t - statistic$ como: Un $t - statistic$ mayor indica que es menos probable que el coeficiente no sea igual a cero. Entonces, cuanto mayor sea $t - statistic$, mejor. $p - value$ es la probabilidad de obtener un $t - statistic$ tan alto o más alto que el valor observado cuando la hipótesis nula (el coeficiente β es igual a cero o que no hay relación) es verdadera. Entonces, si $\Pr(> |t|)$ es bajo, los coeficientes son significativos (significativamente diferentes de cero). Si $\Pr(> |t|)$ es alto, los coeficientes no son significativos.

Variance Inflation Factor

El factor de inflación de la varianza (VIF) es el cociente de la varianza en un modelo con múltiples términos por la varianza de un modelo con un solo término. Cuantifica la gravedad de la multicolinealidad en un modelo de regresión. Proporciona un índice que mide cuánto aumenta la varianza de un coeficiente de regresión estimado debido a la colinealidad.

Pasos a seguir:

- Calcular para cada i :

$$X_i = \alpha_0 + \alpha_1 X_1 + \dots + \alpha_k X_k + \varepsilon$$

- Calcular el VIF_i :

$$VIF_i = \frac{1}{1 - R_i^2}$$

- Analizar la multicolinealidad:
- Analizar la multicolinealidad:

$$VIF_i(\hat{\beta}_i) > 5 \Rightarrow \text{colinealidad}$$

Resumen

Se dispone de un dataset X y una variable respuesta continua. Se asume:

- $Y = X^T \cdot \beta + \varepsilon$
- $\varepsilon \sim N(0, \sigma^2)$
- ε y y son independientes

En aprendizaje supervisado se suelen relajar estas hipótesis.

Mejor modelo:

- $R^2_{adj} \approx 1$.
- “bajo” AIC “bajo”
- “bajo” BIC “bajo”
- “alto” F “alto”
- Std. Error lo más cercano a cero
- Std. Error lo más cercano a cero
- (ver el contraste t-test).
- $t - statistic > 1.96$ (ver el contraste t-test).
- (nivel de significancia del test).
- $Pr(> |t|) < 0.05 = \alpha$ (nivel de significancia del test).
- bajo VIF_i bajo.

Variables categóricas

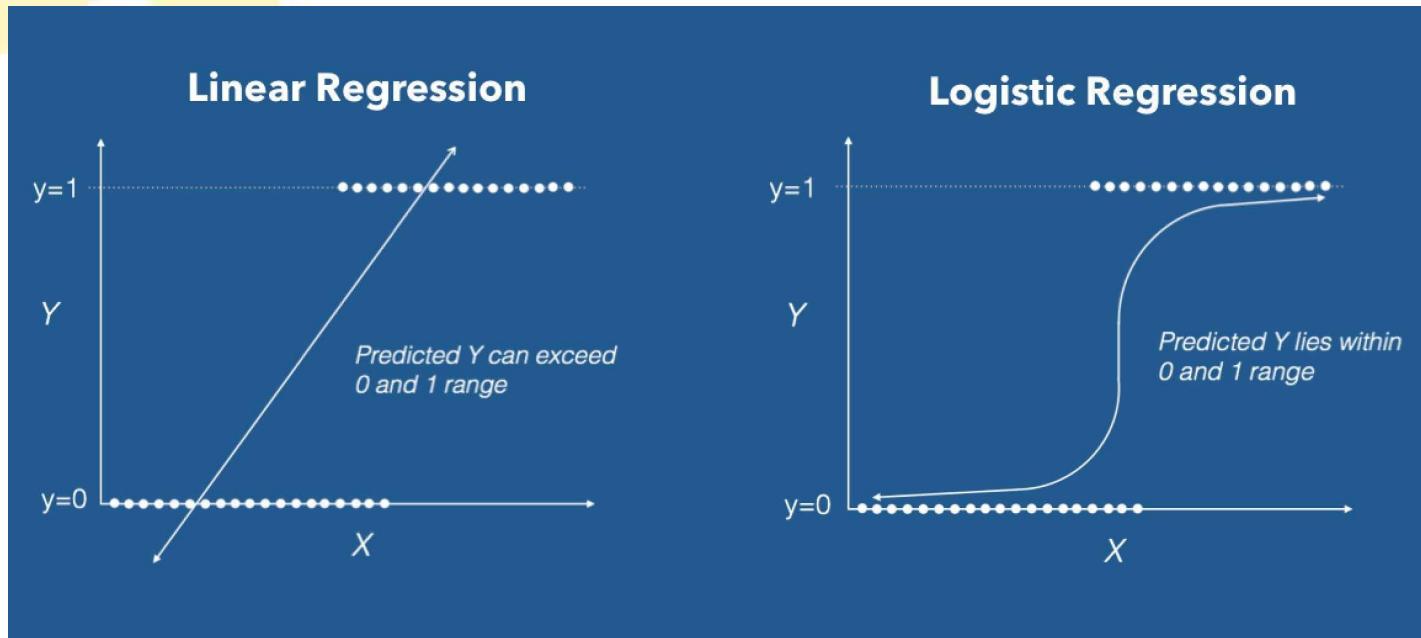
- Si una de las variables de nuestro set de datos es categórica, es necesario recodificarla en forma numérica para que el modelo de regresión lineal la acepte.
- En general, no tiene sentido codificar una variable categórica como una única variable numérica, y por esto se utilizan variables dummy: si la variable categórica X toma posibles valores x_1, \dots, x_k , se sustituye por las variables como sigue:

$$X_{(k)} = \begin{cases} 1 & \text{si } X = x_k \\ 0 & \text{otherwise} \end{cases}$$

Let's do it..



Logistic Regression, ¿verdadero o falso?



Logistic Regression

- Dado un problema de clasificación binaria, donde la variable objetivo toma valores se modeliza cada elemento como una y se establece la relación

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

- Nuestro objetivo es conseguir una función que modele la probabilidad de que un valor pertenezca a una de las dos categorías, cero o uno, partiendo de la ecuación para la regresión lineal.
- Nuestro objetivo es conseguir una función que modele la probabilidad de que un valor pertenezca a una de las dos categorías, cero o uno, partiendo de la ecuación para la regresión lineal.
- Si igualamos las probabilidades al modelo de regresión lineal (modelo de probabilidad lineal) y ajustamos
- Si igualamos las probabilidades al modelo de regresión lineal (modelo de probabilidad lineal) y ajustamos

podría ocurrir que y .

$$p(x) = \beta_0 + \beta_1 x$$

- Podría ocurrir que $p(x) > 1$ y $p(x) < 0$. Por eso la necesidad de una "función link". Realizando algunas operaciones matemáticas se obtiene (función logística o sigmoidal).
- Podría ocurrir que $p(x) > 1$ y $p(x) < 0$. Por eso la necesidad de una "función link". Realizando algunas operaciones matemáticas se obtiene (función logística o sigmoidal):

$$\frac{p}{1-p} = e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}$$

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}}$$

Selección de variables

- Supongamos que se tiene una variable dependiente, y , y tres variables predictoras X_1, X_2, X_3 . Ajustamos el modelo

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

my_model<-glm(y~x1+x2+x3,family="binomial"), en R

- Cuando se ejecuta la orden `anova(my_model)`, la función compara los siguientes modelos en orden secuencial
- Cuando se ejecuta la orden `anova(my_model)`, la función compara los siguientes modelos en orden secuencial

`glm(y~1,family="binomial")`, vs `glm(y~x1,family="binomial")`

`glm(y~x1,family="binomial")`, vs `glm(y~x1+x2,family="binomial")`

`glm(y~x1+x2,family="binomial")`, vs `glm(y~x1+x2+x3,family="binomial")`

- Por lo tanto, compara secuencialmente el modelo más pequeño con el siguiente modelo más complejo al agregar una variable en cada paso. Cada una de esas comparaciones se realiza mediante una prueba de razón de verosimilitud.
- Por lo tanto, compara secuencialmente el modelo más pequeño con el siguiente modelo más complejo al agregar una variable en cada paso. Cada una de esas comparaciones se realiza mediante una prueba de razón de verosimilitud.

Selección de variables

- También pueden compararse dos modelos diferentes. Utiliza el mismo test de razón de verosimilitud con el mismo estadístico (deviancia).

$$D = 2 \cdot (|LL_a - LL_0|) \sim \chi^2_{df_1 - df_2}$$

- Bajo la hipótesis nula: "El modelo que involucra más variables tiene al menos la misma verosimilitud (credibilidad o congruencia) que el modelo nulo".
- Se toma la decisión de eliminar la variable o no, de acuerdo al valor que tome el *p-value*.

Cutoff o umbral de decisión

- La función que hemos calculado devuelve una estimación de una probabilidad. Para poder utilizarla como clasificador, hay que definir un límite de decisión.
- En función de si el valor devuelto por la función es mayor o menor que ese límite, clasificaremos el evento.
- Tendremos entonces que

$$\hat{y} = \begin{cases} 1, & P(\hat{y} = 1 | x_1, \dots, x_n) \geq \varepsilon \\ 0, & P(\hat{y} = 1 | x_1, \dots, x_n) < \varepsilon \end{cases}$$