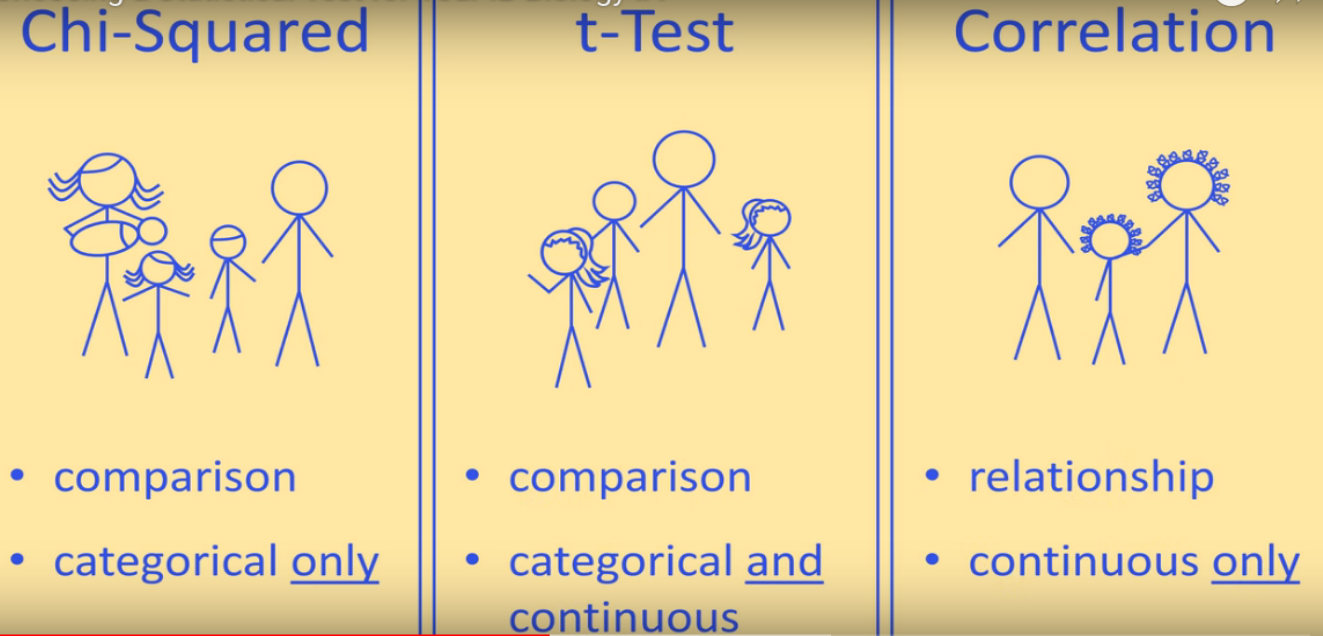
**Aprendizaje automático**

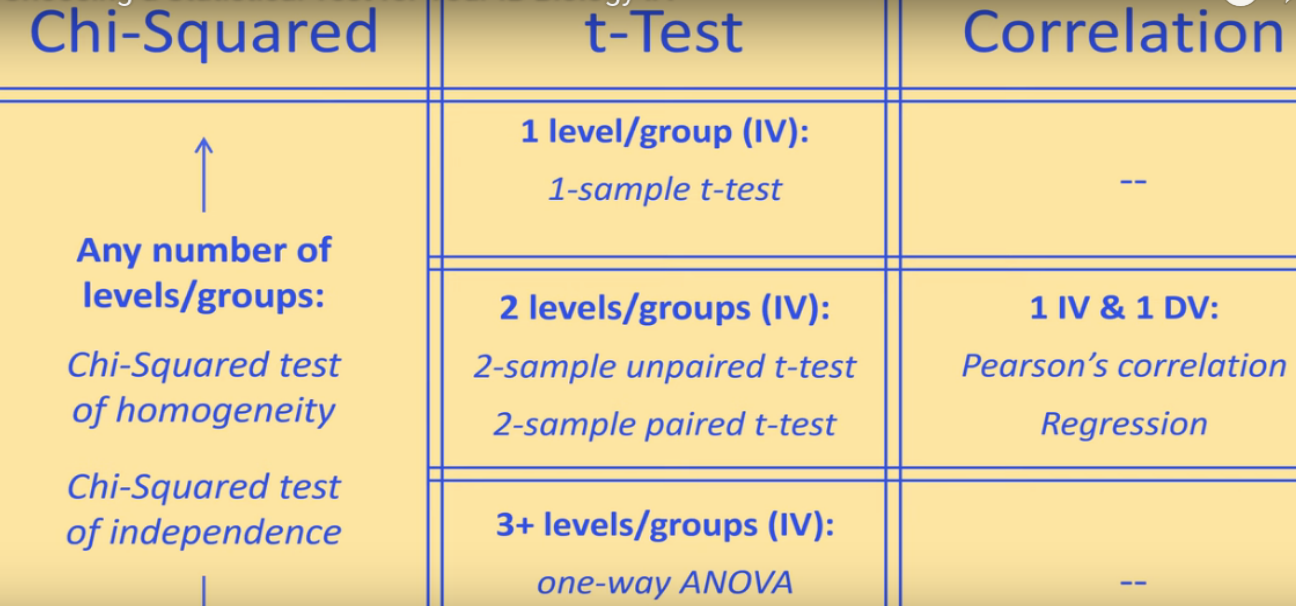
Translator: <https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/analytics-translator>

control alt i para crear el cuadro de ` ejecución en R ```{r} ```

Alt y menos te pone <-

Entrar para ver Google colab.

Interesante para grafico descriptivo  



**Supervisado:**

Tengo una variable dependiente y otras independientes.

Los modelos supervisados pueden ser de:

**a-Regresión** donde la variable dependiente es continua (ej:precio medio,etc)

**b-Clasificación** donde la variable dependiente es discreta.

La primera fase que suele hacer el Data engineering es sacar los datos de sistemas. La segunda es hacer un análisis exploratorio: Medias,varianzas,quantiles y depues gráficos por pares para ver si hay correlación entre variables. Después hace un análisis descriptivo. Y finalmente el prescriptivo(el resultado). Posteriormente sería la puesta en producción.

**Regresión**

Lo ideal sería encontrar una función que explique la variable dependiente perfectamente pero eso no es posible por 3 razones. Siempre hay ruido con los datos, normalmente hay más variables que explican el modelo de las que no disponemos y los datos generalmente son insuficientes para determinar la función con seguridad.

Existen diversos motivos por los que uno querría estimar dicha hipótesis, que

principalmente pueden clasificarse en relativos a la realización de predicciones y a

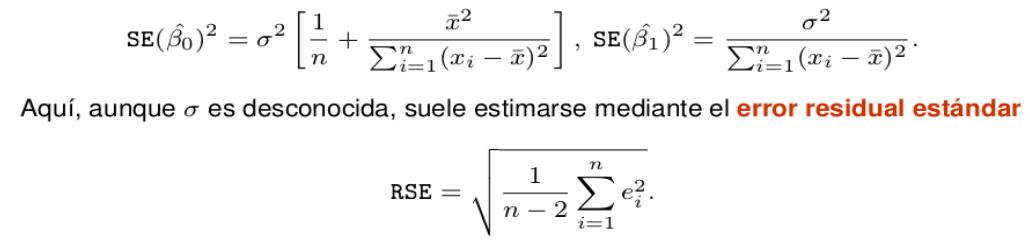
los relativos a inferencia de relaciones.

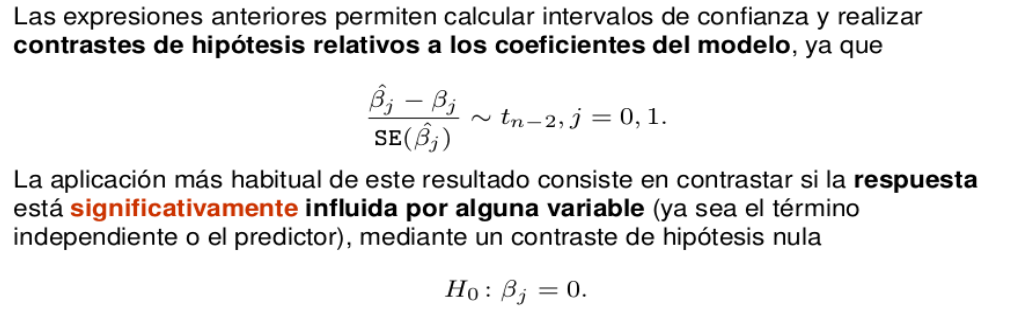
Y = f(X) + € tenemos una función fija desconocida y un término de error que es independiente de la función y tiene media.

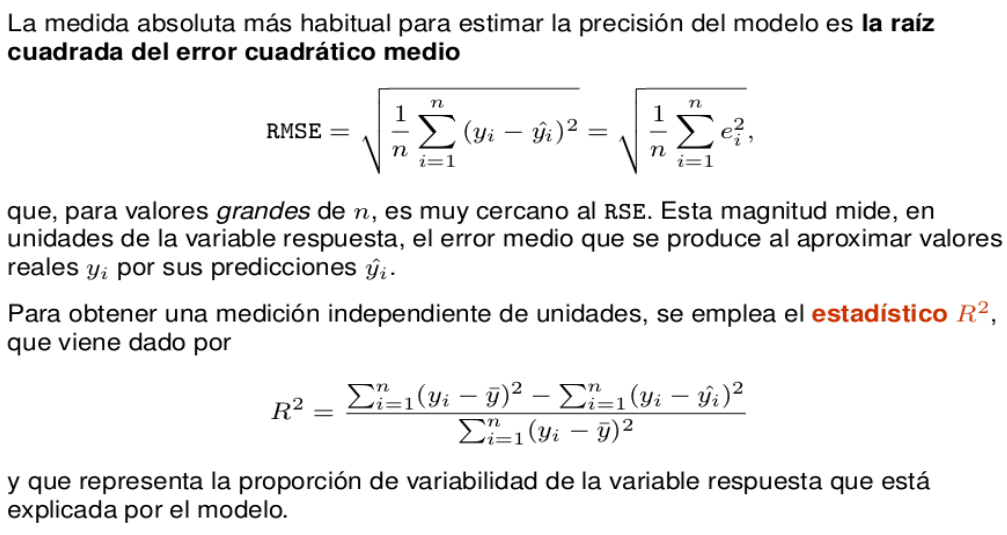
El objetivo es minimizar el error reducible de la estimación. E[(Y − Ŷ )^ 2 ] = E[(f (X) + **€** − h(X))^ 2 ]= E[(f (X) − h(X)) 2 ] + V(€)

Es razonable elegir β ˆ 0 y β ˆ 1 de forma que minimicen MSE

Para medir la precisión de β ˆ 0 y β ˆ 1 como estimaciones de los coeficientes reales utilizamos la **recta de regresión poblacional**. Donde además asumimos relación entre x e y, y que € se comporta como una normal de media cero y varianza constante







**R al cuadrado** explica la variabilidad entre mi modelo y la media de los datos. Cuanto más se acerque a 1 mejor será el modelo. Si es una regresión lineal simple coincide con el coeficiente de regresión de Pearson

Las variable dependiente se explica con las variables independientes y el ángulo mas el error. (ángulo de la recta sobre los puntos de la población- ventas=b + aX- el ángulo seria la derivada de ventas entre derivada de X)

Si el error(distancia entre punto y la recta) no tiene media cero y varianza constante(homocedastico) significa que el modelo necesita mas variables que son importantes y explican el modelo pero ahora no las tenemos.

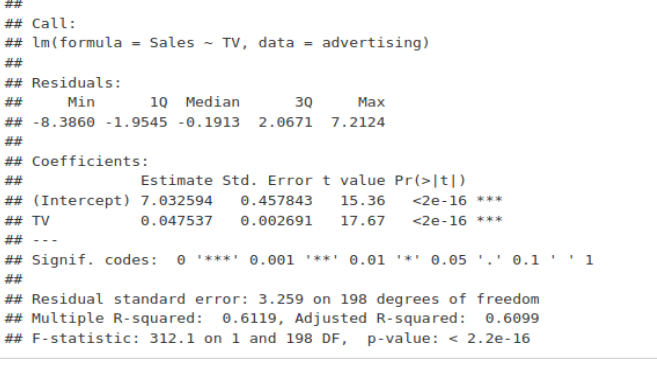
**OLS o mínimos cuadrados ordinarios** :Este es el objetivo de las regresiones lineales. Reducir la diferencia entre el punto y su estimación en la recta.

Si las variables dependientes tienen más de 0,7/0,8 de correlación hay que quitar alguna para que no me de datos raros.

Para seleccionar las variables mas importantes utilizo embede y brad. **Buscarlos Nos va a pasar documentación de selección de variables.**

**Dentro de coeficiente**

Error standard bajo vs estimate es bueno y que Pr sea bajo tambien. Las 3 estrellas implican que no podemos descartar la hipotesis nula que implica que la variable que estamos utilizando es significativa para la predicción. Lo contrario implicaria que mi modelo sería igual de bueno sin esa variable.



**Coefficients:** These are the weights that minimize the sum of the square of the errors.

* Std. Error is Residual Standard Error divided by the square root of the sum of the square of that particular x variable.
* t value: Estimate divided by Std. Error
* Pr(>|t|): Look up your t value in a T distribution table with the given degrees of freedom.

**Standard deviation** is the square root of variance. Standard Error is very similar. The only difference is that instead of dividing by n-1, you subtract n minus 1 + # of variables involved.

**Multiple R-Squared**. Also called the **coefficient of determination**, this is an oft-cited measurement of how well your model fits to the data.R-Squared subtracts the residual error from the variance in Y. The bigger the error, the worse the remaining variance will appear.

**Adjusted R-Squared.** Multiple R-Squared works great for simple linear (one variable) regression. However, in most cases, the model has multiple variables. The more variables you add, the more variance you’re going to explain. So you have to control for the extra variables. Adjusted R-Squared normalizes Multiple R-Squared by taking into account how many samples you have and how many variables you’re using.

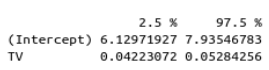
If you have 100 observations (n) and 5 variables, you’ll be dividing by 100-5-1 = 94. If you have 20 variables instead, you’re dividing by 100-20-1 = 79. As the denominator gets smaller, the results get larger: 99 /94 = 1.05; 79/94 = 1.25.

A larger normalizing value is going to make the Adjusted R-Squared worse since we’re subtracting its product from one.

**F-Statistic**. Including the t-tests, this is the second “test” that the summary function produces for lm models. The F-Statistic is a “global” test that checks if at least one of your coefficients are nonzero. Varianza de la regresion entre la varianza del error.

Se pueden obtener intervalos de confianza para los coeficientes.

**confint**(lm\_fit\_sales\_TV, **level = 0.95**)

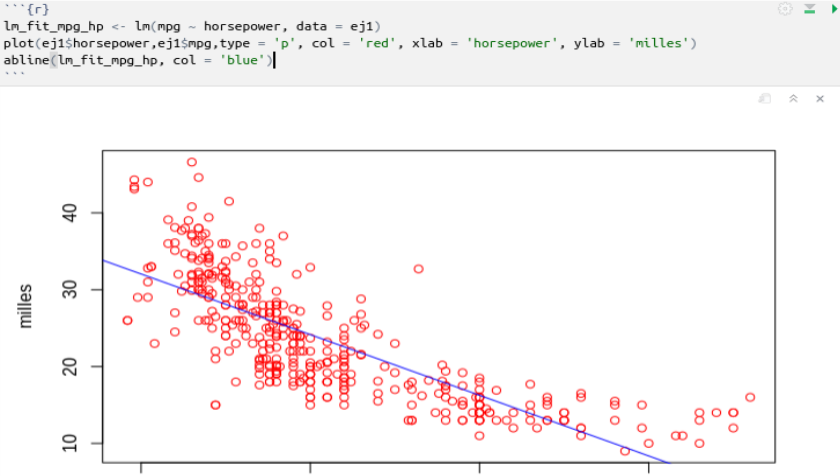


<https://www.youtube.com/watch?v=b_32VfHbdxc>

Bajar este libro en pdf . Gratis en internet.

'An Introduction to Statistical Learning with Applications in R

Al hacer la linea de la regresión me daba error porque la primera variable tiene que ser la x y debe ir primero en el plot(horsepower debe ir primero en el ejemplo)



*Si tengo dos variables que no tienen relación lineal, puedo transformar las variables en logarítmicas que lo que hace es linealizar la relación.*

*En python se usa para estadística statsmodels*

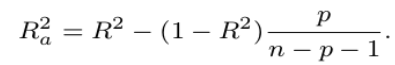
*Para machine learning con python se baja con scikit-learn y cuando se llama sklearn.*

**Modelo de regresion lineal multiple**

Se debe dar una relación lineal entre la variable respuesta y los predictores

La bondad de ajuste del modelo, al igual que en el caso univariante, se mide mediante

el RMSE , el estadístico R^2 o, aún mejor, mediante el R^2 ajustado



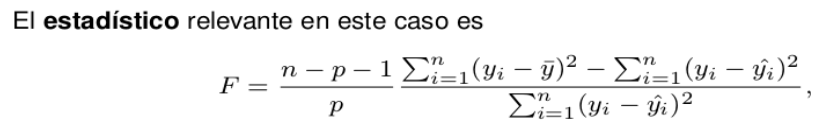
Para contrastar si hay una verdadera relación lineal entre la variable respuesta y los

predictores, se realiza el contraste de hipótesis nula

H 0 : β 1 = β 2 = · · · = β p = 0 cuya hipótesis alternativa es

H A : al menos un β j es distinto de 0.

que, bajo la hipótesis nula, sigue una distribución **F de Snedecor** de parámetros (p, n − p − 1). (asumimos n>p) siendo p el numero de variables y n el numero de observaciones.



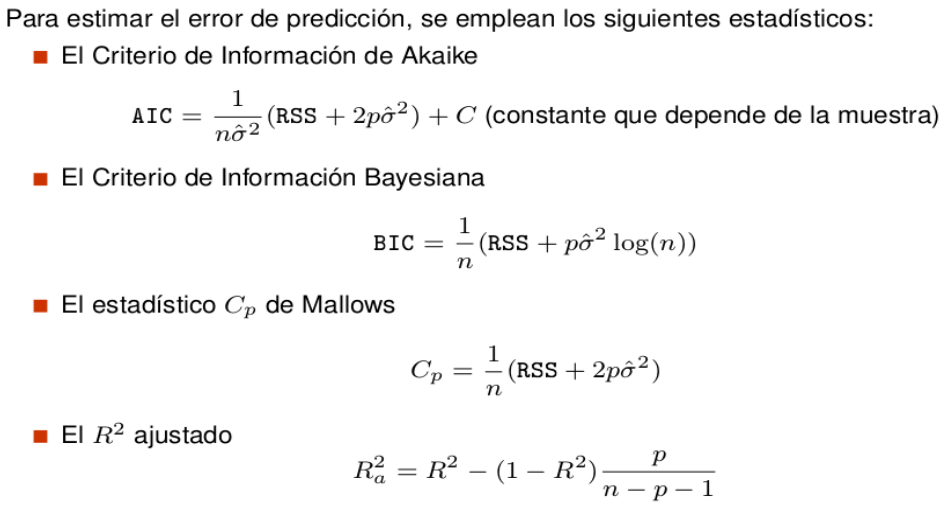
Una vez que se sabe que, al menos, hay un predictor que influye en la variable respuesta, es natural preguntarse cuáles son aquellos predictores influyentes para ajustar un modelo que sólo tenga en cuenta a éstos.

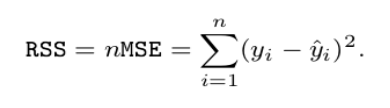
Podría parecer que tiene sentido realizar contrastes t individuales para cada predictor, pero esto puede llevar a errores si p es grande.

Idealmente, se deberían considerar todos los modelos lineales posibles de Y en función de subconjuntos de X 1 , X 2 , . . . , X p para elegir el mejor de ellos, pero esto no es viable. Dos consideraciones:

* Existen diversas alternativas para comparar la bondad de ajuste de dos modelos lineales y elegir el mejor ( AIC , BIC , Cp , R^2 , análisis de residuos.).

*El objetivo de la construcción del modelo debe ser minimizar cualquiera de los tres primeros o maximizar el cuarto.*





* También hay diferentes estrategias para guiar la selección de variables (forward selection, backward selection, mixed selection. . . ).

**Por esto, se hace necesario construir modelos que empleen pocas variables para que el error en las predicciones sea pequeño y para que el modelo sea interpretable.**

Para comparar la bondad de diferentes modelos, o para estimar el error de las predicciones arrojadas por un modelo, caben dos alternativas:

* Ajustar el error de entrenamiento para estimar el verdadero error de predicción.
* Emplear un conjunto de validación, que no intervenga en el proceso de entrena-

miento del modelo, y calcular los errores de predicción sobre ese conjunto.

**Naturalmente, incluir más variables en el proceso de ajuste del modelo hará que el**

**RMSE o el R^2 (de entrenamiento) disminuyan, pero esto no quiere decir que el modelo**

**vaya a ofrecer mejores predicciones.**

**Elegir las mejores variables del modelo. (Variable selection procedure)**

1. **Subseting and best model election. Library(Leaps)--- subsets -regsubsets.**

[**https://www.youtube.com/watch?v=oevjpFdZqmk**](https://www.youtube.com/watch?v=oevjpFdZqmk)

1. **Stepwise selection** Revisa todos los posibles subset según resultados de AIC,R2 ajustado y BIC. Le dá un valor a cada modelo. <https://www.youtube.com/watch?v=ejR8LnQziPY>





Selecciono solo con el intercept sin otra variable predictora y después el sistema va haciendo combinaciones hasta que llega a none que sería la mejor combinación a partir de la cual el AIC no bajará.

Scope hay que ponerlo para decirle hasta que variable llegar. En este ejemplo serían todas. Es muy similar a foward selection.

1. **Backward selection.** [**https://www.youtube.com/watch?v=0aTtMJO-pE4**](https://www.youtube.com/watch?v=0aTtMJO-pE4)

Selecciono todas las variables y después el sistema va haciendo combinaciones hasta que llega a none que sería la mejor combinación.





1. **Forward selection.** [**https://www.youtube.com/watch?v=OYEII--K\_k4&t=303s**](https://www.youtube.com/watch?v=OYEII--K_k4&t=303s)

Selecciono solo con el intercept sin otra variable predictora y después el sistema va haciendo combinaciones hasta que llega a none que sería la mejor combinación. Scope hay que ponerlo para decirle hasta que variable llegar. En este ejemplo serían todas.





No dan los mismos resultados y un método puede darte mejor AIC( cuanto menor sea este número mejor explican las variables elegidas la variable independiente)

**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.

**Problemas potenciales:**

* **Relaciones no lineales**. En caso de que lo haya, pueden emplearse transformaciones no lineales de los predictores para mejorar el modelo.
* **Varianza no constante de los residuos**.Puede resolverse aplicando una transformación lineal a la variable respuesta.
* **Outliers.** Típicamente, eliminar outliers en el ajuste del modelo no cambia gran cosa el modelo final, pero puede mejorar drásticamente la bondad de ajuste.
* **High leverage points**. Es una oulier para las variables predictoras. La presencia de high leverage points tiene un impacto reseñable sobre el modelo de regresión lineal estimado, y suele ser recomendable eliminarlos.
* **Colinealidad.** variance inflation factor. Usualmente, un VIF > 5 sugiere que la variable en cuestión es colineal con el resto.

Ejemplo modelo regresion lineal: <https://www.youtube.com/watch?v=HP3RhjLhRjY>

**Remuestreo (sampling)** ver documentacion.

Se utiliza para estimar o para validar modelos.

* Para estimar: Jacknife o Bootstraps( se pueden repetir las observaciones en las distintas particiones.

a-Bootstraps: El objetivo, habitualmente, es conseguir estimaciones robustas o intervalos de confianza para parámetros cuando no se puede emplear la inferencia paramétrica clásica porque no se dispone de hipótesis sobre la población (o, disponiendo de ellas, su veracidad es dudosa) o porque el cálculo de las desviaciones estándar es muy complicado.

b- Jacknife: se emplea para estimar el sesgo y la varianza de un cierto estadístico a partir de una muestra, recalculando la estimación de tal estadístico dejando fuera una o varias observaciones de la muestra cada vez.

***En R sería.***

***Library(bootstrap)***

***X <- c(las observaciones)***

***Theta <- function(X) {mean(X)}***

***Jacknife(X,theta)***

* Para validar modelos: bootstrap y cross-validation

Es una variación del jacknife que se emplea para validar el ajuste de un modelo predictivo.

Un subconjunto de la muestra de datos disponible (el conjunto de training) se emplea para ajustar un modelo y el resto de los datos (el conjunto de validación) se emplea para calcular valores predichos. La comparación de valores reales y valores predichos sirve para estimar el rendimiento real del modelo (error de test).

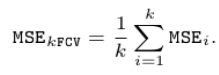
Cross validation tiene 2 problemas: 1 Las estimaciones del error de test pueden variar mucho dependiendo, precisamente, de qué observaciones caigan en el conjunto de training y cuáles en el de test. 2 Dado que sólo se emplea un subconjunto de los datos disponibles para entrenar el modelo, el error de test estimado será, probablemente, una sobre-estimación del verdadero error de test del modelo. Para solucionar parte del problema se puede usar el Leave one out cross validation system. Podemos incluir una sola observación (x i , y i ) en el conjunto de validación, y emplear el resto de observaciones como conjunto de training. Soluciona los problemas anteriores pero es muy caro computacionalmente.

Una generalización sencilla de la LGOCV que es menos costosa consiste en tomar, en cada iteración, un conjunto más grande de observaciones en lugar de una única observación en el conjunto de test.En k-fold cross-validation (en adelante, k FCV )

1 Se reparten los datos disponibles en k submuestras (folds) de tamaño comparable.

2 Para cada fold i, se entrena el modelo utilizando las k − 1 submuestras restantes y se valida con ese fold como conjunto de test, y se anota el error MSE i .

3 Se estima el error de test mediante



**Modelos lineales generalizados**

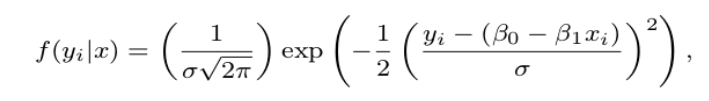
* La variable respuesta no tiene por qué ser continua.
* Los errores no tienen por qué seguir una distribución normal.
* No tiene por qué haber una relación lineal entre los predictores y la variable respuesta.

Los modelos lineales generalizados (GLM) sirven para extender el modelo de regresión lineal en todas estas direcciones de una forma unificada.Son los modelos utilizados en riesgos porque se pueden explicar.

Los modelos de machine learning son practicamente todos basados en random forest. No tienen parámetros. Todos son hiperparametros que decido yo antes. Numero de arboles,profundidad, numero preguntas, etc

Cuales son los parametros que mejor ajustan tu nube de puntos. Eso es la función de maxima verosimilitud. Intuitivamente elige los valores de los parámetros que hacen más probables los datos.

Lo que dice esta función es que si ocurre x, yi sigue una distribución normal. Dado una x todas las y que genere provendran de una normal.



Maximum likelihood: Valor óptimo para la media y para la desviación típica para una distribución dadas un conjunto de medidas observadas. <https://www.youtube.com/watch?v=XepXtl9YKwc>

The ordinary least squares, or OLS is a method for approximately determining the unknown parameters located in a linear regression model. This method is obtained by minimizing the total of squared vertical distances between the observed responses within the dataset and the responses predicted by the linear approximation.

The Maximum likelihood Estimation, or MLE, is a method used in estimating the parameters of a statistical model, and for fitting a statistical model to data. Using the maximum likelihood estimation, you can estimate the mean and variance of the height of your subjects. The MLE would set the mean and variance as parameters in determining the specific parametric values in a given model.

Importante. El modelo naive es el peor que puedo aceptar. Que es 1 menos el porcentaje de unos. Ej:titanic el 38% sobreviven. Por tanto el modelo Naive sería 100% - 38% = 62%

**Bernoulli y Binomial (ver videos documentos)**

Experimento con 2 posibles resultados.

Un ejemplo puede ser cuando tenemos una moneda y la lanzamos, la distintas probabilidades

P(X=1) = Theta

P(X=0)= 1-Theta

Si lo lanzo una vez la media es la probabilidad de que ocurra y la varianza es p\*(1-p) siendo las desviación típica la raíz cuadrada de p\*(1-p). P será ½

Si lo lanzo n veces será una binomial.

La probabilidad de que P(X=K) = Theta^k\*(1-Theta)^(1-k) si lanzamos varias veces y generalizamos las función. K es el numero de vecesque acertamos. En el ejemplo sería que salga cara.

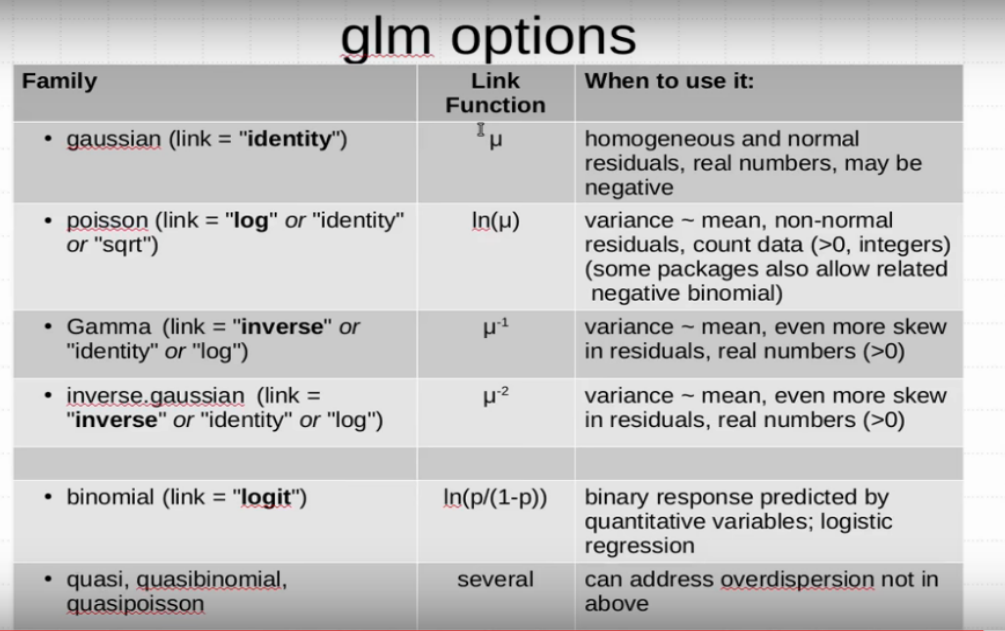
P(E *i=1 hasta N* xi =K)=(combinatoria de N K) de Theta^K \*(1-Theta)^(N-K)

La combinatoria de N a K es N factorial /K factorial\* (N-K) factorial

La media es n\*p y la varianza es es n\*p\*(1-p)

**Modelos lineales generalizados**

<https://www.youtube.com/watch?v=G5xIFdLL5Ic>



Ejemplo: <https://www.youtube.com/watch?v=kffIgjHxdpw>

**Modelos de Poisson:**

La distribución de Poisson es una distribución de probabilidad discreta que expresa, a

partir de una frecuencia de ocurrencia media, la probabilidad de que ocurra un

determinado número de eventos durante un cierto periodo de tiempo.

* Número de erratas por página que escribe una cierta persona.
* Número de e-mails recibidos al día por una cierta persona.
* Número de siniestros sufridos por un coche al año.
* Número de premios recibidos por estudiantes de un cierto instituto en un curso

Académico

**La distribución exponencial** es una distribución de probabilidad continua que sigue el

tiempo transcurrido hasta que sucede un determinado evento.

* Tiempo transcurrido en un call-center hasta que se recibe la primera llamada del

día.

* Cantidad de metros de hilo en una bobina que hay que desenrollar hasta
* encontrar un nudo.
* Tiempo que transcurre hasta que un cierto teléfono tiene su primera avería.

**Modelos gamma:**

La distribución Gamma es una distribución de probabilidad continua que suele

emplearse para modelizar variables positivas, asimétricas y con colas pesadas.

* Costes.
* Tiempos de espera.

seguros de no-vida

El modelado general es el siguiente:

Coste = Número de siniestros × Coste medio por siniestro.

* El número de siniestros se modela mediante un GLM de tipo Poisson.
* El coste medio por siniestro se modela mediante un GLM de tipo Gamma.
* La credibilidad que se otorga a cada una de las observaciones se pondera en

base a su exposición (la fracción del año que ha transcurrido desde que fue

contratada la póliza).