Curso

11

MACHINE LEARNING & BIG DATA

Regresión Lineal

Regresiones

JOSÉ NELSON ZEPEDA DOÑO

Cluster de Estudio: Advanced Analytics

Este material es el resumen de muchos autores que por medio de sus libros y documentos nos ofrecen fuentes riquísimas de conocimiento sobre los temas de Big Data y Machine Learning.
Algunas citas, figuras y tablas pueden ser encontradas de forma textual tal como lo indica el autor en su material original.
Especial mención para el autor del capítulo 18 de Análisis de Regresión Lineal para SPSS
http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/GuiaSPSS/18reglin.pdf

Nelson Zepeda
MIP • V 1.0
San Salvador El Salvador
Phone 503 79074137 • @nelsonzepeda733

Tabla de Contenido

Capítulo 1

Conceptos

Uno de los aspectos más relevantes de la Estadística es el análisis de la relación o dependencia entre variables.

Un análisis de regresión lineal es una técnica estadística utilizada para estudiar la relación entre variables ya que es uno de las aspectos más relevantes en el ámbito estadístico.

Una regresión lineal se adapta a una gran variedad de escenarios que incluyen ámbitos como la investigación social, economía, finanzas, medicina, en fin una gran cantidad de fenómenos en donde pueden intervenir dos o más variables, por ejemplo, en el contexto de la investigación de mercados, puede utilizarse para determinar en cual de diferentes medios de comunicación puede ser más eficaz invertir; o para predecir el número de ventas de un determinado producto, etc.

A manera de introducción y para poner una simiente del tema es necesario destacar lo siguiente:

Un análisis de regresión lineal es una técnica estadística que permite estudiar la relación entre una variable dependiente y unas o más variables independientes con el propósito de:

- Averiguar en qué medida la variable dependiente puede estar explicada por las variables independientes.
- Obtener predicciones de la variable independiente a partir de las variables independientes.

El procedimiento se resume en obtener la ecuación mínimo cuadrática que mejor expresa la relación entre las variables y estimar la calidad de la ecuación de regresión obtenida.

Tanto en el caso de dos variables como en el de más de dos variables, el análisis de regresión lineal puede utilizarse para explorar y cuantificar la relación entre la variable independiente y una o más variables independientes también conocidas como variables predictoras.

La Recta de Regresión

De todos es conocido que un diagrama de dispersión permite formarse una primera impresión muy rápida sobre el tipo de relación existente entre dos variables.

Es importante mencionar que aunque un diagrama de dispersión permite formarse tener una primera guía sobre el tipo de relación existente entre dos variables, utilizarlo como una forma de cuantificar esa relación es sumamente difícil ya por ser un elemento puramente grafico lo único que si podríamos afirmar de forma contundente es si la relación entre dos variables existe o es nula.

Supongamos que se dispone de una base de datos con una pequeña cantidad de observaciones relacionadas a 35 marcas de cerveza de las cuales se quiere estudiar la relación entre el grado de alcohol y su contenido calórico.

Las diferentes observaciones se muestran en la siguiente figura:

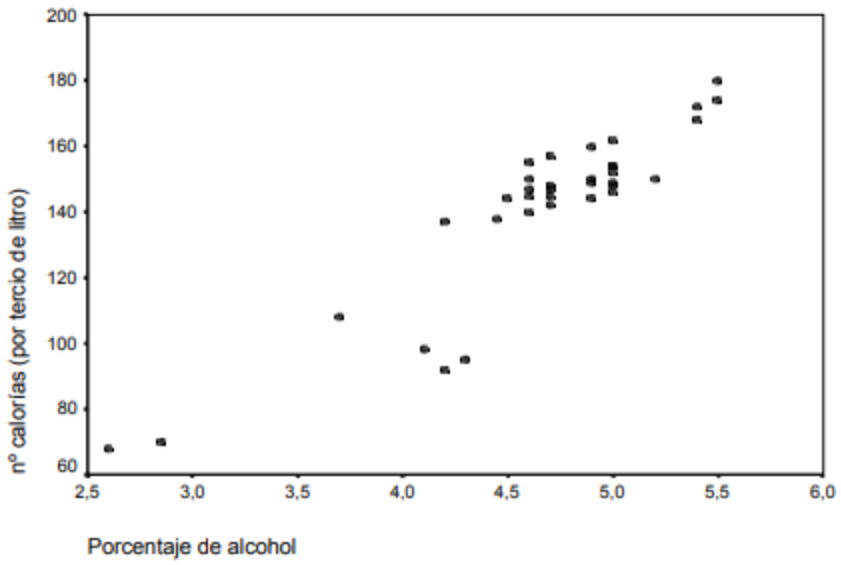


Fig 1-1 Observaciones 35 marcas de Cerveza

Sin entrar en muchos detalles se puede inferir que existe una relación positiva entre ambas variables: conforme aumenta el porcentaje de alcohol también aumentan las calorías, esto, aunque es correcto, es poco específico.

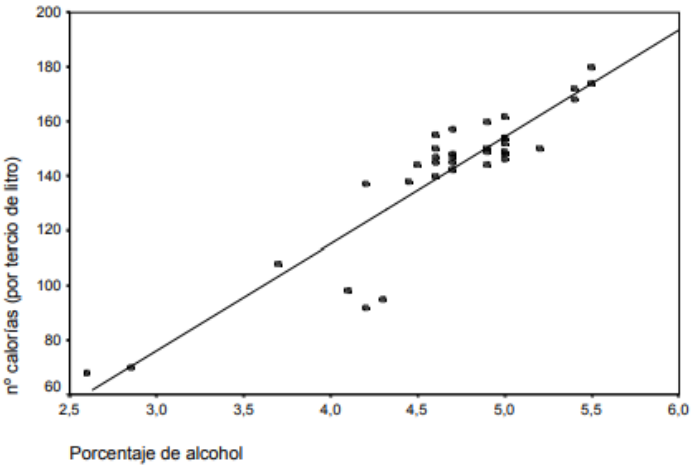
Para poder ser más detallistas se requiere establecer una función matemática simple, tal como una línea recta que se describe mediante la siguiente función:

$$Y_i = B_0 + B_1 X_i$$

El coeficiente B1 es la pendiente de la recta, es decir es el cambio medio que se produce en el número de calorías (Y) por cada unidad de cambio que se produce en la cantidad de alcohol (Xi).

El coeficiente B0 es el punto en el que la recta corta el eje vertical, es decir es el número de calorías que corresponde a una cerveza con porcentaje de alcohol cero.

Conociendo los valores de estos dos coeficientes es posible construir la representación de la recta y describir de manera más especifica la relación existente entre el contenido de alcohol y el número de calorías.



$$Y_i = -33{,}77 + 37{,}65\, X_i$$

nº de calorías = – 33,77 + 37,65 (% de alcohol)

Fig. 1-2 Representación de la función

La pendiente de la recta indica que, en promedio, a cada incremento de una unidad en el porcentaje de alcohol le corresponde un incremento de 37.65 calorías, el origen de la recta indica que una cerveza sin alcohol podría tener 33.77 calorías lo cual no parece algo posible, por otro lado, no podemos obviar que la nube de puntos no contiene cervezas con menos de 2% de alcohol, así, aunque el origen de la recta aporta información sobre lo que podría ocurrir si extrapolamos hacia abajo hasta llegar a cero grados de alcohol, ahora bien, al hacer esto, estaríamos efectuando pronósticos en un rango de valores que va más allá de los valores o rangos disponibles, lo cual es algo arriesgado en el contexto de las regresiones.

En una situación ideal, en la que todos los puntos de un diagrama de dispersión se encontraran en una línea recta, no haría falta preocuparse de encontrar la recta que mejor resume los puntos del diagrama, ya que simplemente uniendo un par de puntos se obtendría dicha recta. Ahora bien, en una situación real, es posible trazar muchas rectas y cada una de ellas se ajustara de manera diferente a la nube de puntos, por lo que ahora lo que se persigue es encontrar la recta capaz de convertirse en la mejor representante del conjunto de puntos.

Existen diferentes procedimientos para ajustar una función simple, cada uno de los cuales intenta minimizar una medida diferente del grado de ajuste. Tradicionalmente se elige aquella recta que hace mínima la suma de los cuadrados de las distancias verticales entre cada punto y la recta. Esto significa que de todas las rectas posibles, existe solamente una que consigue que las distancias verticales entre cada observación y la recta sean mínimas.

Como comentario adicional vale la pena indicar que las distancias se elevan al cuadrado porque, de lo contrario, al ser unas positivas y otras negativas, se anularían entre si al sumarlas.

Bondad de Ajuste

En los párrafos anteriores se explicaba la recta y su grado de importancia, ahora bien, es de suma importancia disponer de información precisa del grado en que la recta se ajusta a la nube de puntos.

Cualquiera que sea la nube de puntos, obtener la recta mínimo cuadrática no será ninguna inconveniente, pero necesitamos información adicional para determinar el grado de fidelidad con que esta recta describe el fenómeno.

Una medida de ajuste que ha recibido gran aceptación en el contexto del análisis de regresión es el coeficiente de regresión R^2: el cuadrado el coeficiente de correlación múltiple. Se trata de una medida estandarizada que toma valores entre 0 y 1 (0 cuando las variables son independientes y 1 cuando existe entre ellas una relación perfecta).

Este coeficiente posee una interpretación muy intuitiva: representa el grado de ganancia que podemos obtener al predecir una variable basándonos en el conocimiento que se tiene de otra u otras variables. Si queremos por ejemplo pronosticar el número de calorías de una cerveza sin el conocimiento de las otras variables, utilizaríamos la media del número de calorías. Pero si tenemos información sobre la variable y del grado de relación entre ambas, es posible mejorar el pronóstico.

Por ejemplo si en el ejercicio de la cerveza, el valor de R^2 fuera de 0.83, es correcto afirmar que si conocemos el porcentaje de alcohol de una cerveza, podemos mejorar en un 83% nuestros pronósticos sobre el número de calorías si, en lugar de utilizar como pronostico el número medio de calorías, basamos nuestra proyección en el porcentaje de alcohol.

Capítulo 2

Regresión Lineal Simple

Simple igual a una.

Un análisis de regresión lineal simple hace referencia a que solo se tendrá una sola variable independiente, en la práctica en los diversos escenarios y situaciones que se deben analizar generalmente intervienen 2 o más predictoras, sin embargo, comprender la regresión lineal simple es la base para poder avanzar al siguiente nivel.

Modelo de Regresión Lineal

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Donde

y = variable dependiente

β_0 = ordenada al origen

β_1 = pendiente

x = variable independiente

ε = Error aleatorio

La expresión $\beta_0 + \beta_1 x$ se denomina componente determinística del modelo de regresión lineal. La muestra de pares de datos se usará para estimar los parámetros $\beta_0 y \beta_1$ de la componente determinística.

La diferencia principal entre un modelo probabilístico y uno determinístico es la inclusión de un término de error aleatorio en el modelo probabilístico.

Los errores se denominan frecuentemente residuales. Podemos observar en la gráfica de regresión los errores indicados por segmentos verticales.

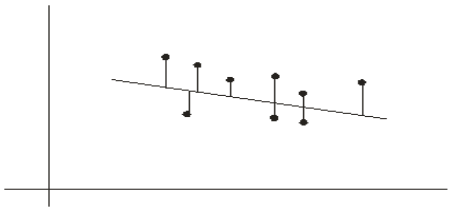


Fig 2-1: Representación gráfica del error.

Al usar el criterio de mínimos cuadrados para obtener la recta que mejor se ajuste a nuestros datos, podemos obtener el valor mínimo para la suma de cuadrados del error (SSE)

$$SSE = SS_y - b_1 SS_{xy}$$

A la varianza de los errores e se le llama varianza residual siendo denotada por s_e^2, se encuentra dividiendo SSE entre n-2

$$S_e^2 = \frac{SSE}{n-2}$$

La raíz cuadrada positiva de la varianza residual se llama error estándar de estimación y se denota por Se.

Pendiente de la recta[1]

La pendiente es la inclinación de la recta con respecto al eje de abscisas, se puede denotar por la letra m o bien en nuestro caso β_1, si la pendiente es mayor que cero, se dice que es creciente el ángulo que forma la recta con el eje es agudo.

Si la pendiente es menor que cero, la función es decreciente y el ángulo que forma la recta con el eje es obtuso.

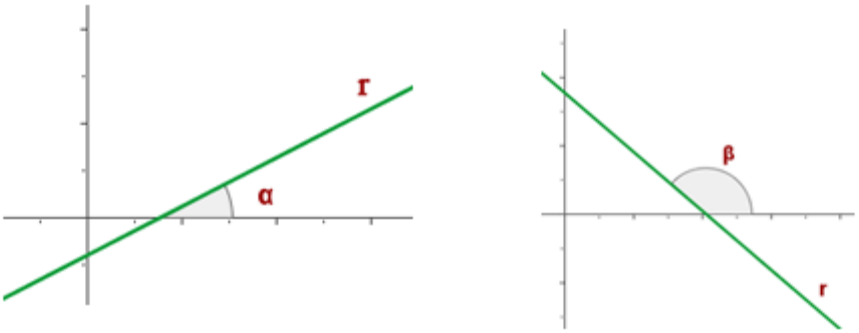


Fig 2-2 Tipos de Pendientes

Es posible calcular el valor de la pendiente si se conocen 2 puntos

[1] https://www.ditutor.com/funciones/pendiente-recta.html

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

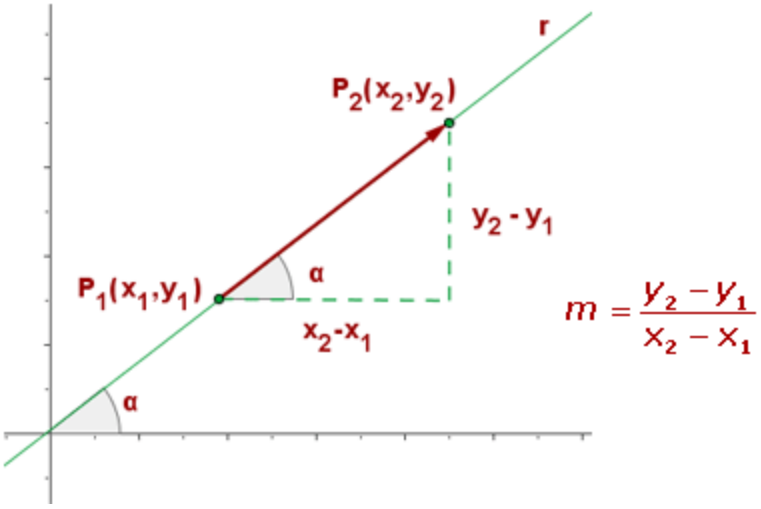


Fig 2-3 Calculo de la pendiente

Análisis de correlación

Establece si existe una relación entre las variables y responde a la pregunta,"¿Qué tan evidente es esta relación?".

La correlación es una prueba fácil y rápida para eliminar factores que no influyen en la predicción, para una respuesta dada.

Coeficiente de Correlación de Pearson

- Es una medida de la fuerza de la relación lineal entre dos variables x y y.
- Es un número entre -1 y 1
- Un valor positivo indica que cuando una variable aumenta, la otra variable aumenta
- Un valor negativo indica que cuando una variable aumenta, la otra disminuye
- Si las dos variables no están relacionadas, el coeficiente de correlación se aproxima a 0.

El coeficiente de correlación r se calcula mediante la siguiente fórmula:

$$r = \frac{SSxy}{\sqrt{SSx \cdot SSy}}$$

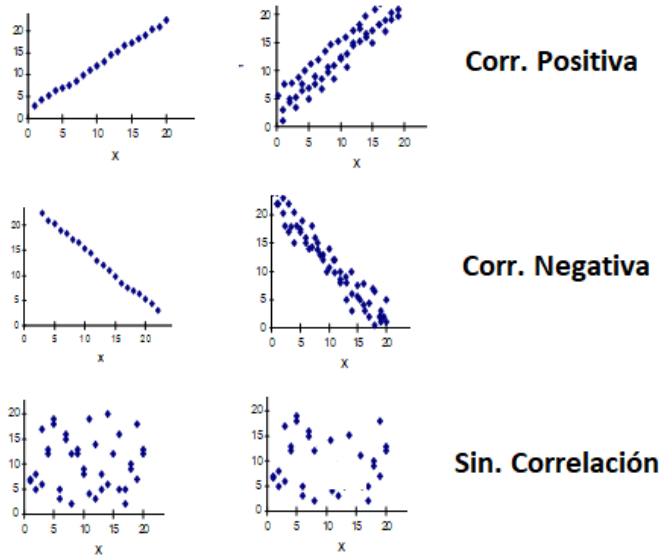


Fig 2-3 Tipos de Correlación

Bondad de Ajuste

Se refiere al coeficiente de correlación múltiple (R) y a su cuadrado, dado que solo intervienen 2 variables, el coeficiente de correlación múltiple viene siendo el valor absoluto del coeficiente de correlación de Pearson entre esas dos variables. Su cuadrado es el coeficiente de determinación.

$$R^2 = 1 - \frac{\text{Suma de cuadrados de los residuos}}{\text{Suma de cuadrados total}}$$

Los residuos corresponden a las diferencias existentes entre los datos reales de la variable dependiente (Y) y los pronósticos obtenidos con la recta.

Se señaló en los párrafos iniciales del documento que R^2 expresa la proporción de varianza de la variable dependiente que esta explicada por la variable independiente, pero es muy importante resaltar que este indicador por sí solo no permite afirmar que las relaciones detectadas sean del tipo causal, solo es posible hablar de grado de relación.

Adicional a este indicador existe el de *R Cuadrado Corregida*, el cual es una correlación a las baja de R^2 que se basa en el numero e casos y de variables independientes.

$$R^2_{\text{corregida}} = R^2 - [p\,(1 - R^2)/(n - p - 1)]$$

En la ecuación previa, P se refiere al número de variables independientes.

Otro indicador de mucha popularidad es el que se conoce como *error típico de la estimación,* Se, y que es obtenido a partir de la desviación típica de los residuos, es decir, la desviación típica de las distancias existentes entre los datos reales y los datos pronosticados con nuestra ecuación.

$$\text{Error típico de estimación} = S_e = \sqrt{\sum (Y_i - \hat{Y}_i)^2 / (n-2)}$$

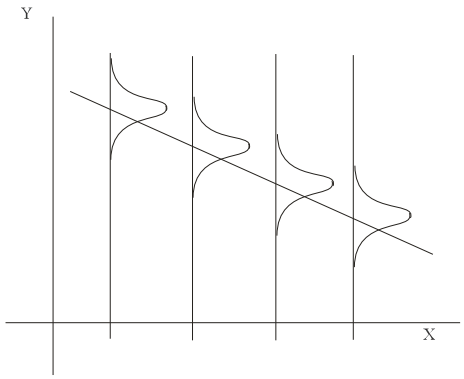
Estadístico F

El estadístico F fue creado por Ronald Fisher (1890-1962), matemático británico, cuyas teorías estadísticas hicieron mucho más precisos los experimentos científicos. Sus proyectos estadísticos, primero utilizados en biología, rápidamente cobraron importancia y fueron aplicados a la experimentación agrícola, médica e industrial. Fisher también contribuyó a clarificar las funciones que desempeñan la mutación y la selección natural en la genética, particularmente en la población humana.

El valor estadístico de prueba resultante se debe comparar con un valor tabular de F, que indicará el valor máximo del valor estadístico de prueba que ocurría si H0 fuera verdadera, a un nivel de significación seleccionado. Antes de proceder a efectuar este cálculo, se debe considerar las características de la distribución F

Supuestos para un Modelo de Regresión Lineal

1. Para cada valor de *x*, la variable aleatoria ε se distribuye normalmente.
2. Para cada valor de *x*, la media o valor esperado de ε es 0; esto es, $E(\varepsilon) = \mu_\varepsilon = 0$.
3. Para cada valor de *x,* la varianza de ε es la constante σ^2 (llamada varianza del error).
4. Los valores del término de error ε son independientes.
5. Para un valor fijo de *x*, la distribución muestral de *y* es normal, porque sus valores dependen de los de ε.



6. Para un valor fijo *x*, es posible predecir el valor de *y*.
7. Para un valor fijo *x*, es posible estimar el valor promedio de *y*

Capítulo 3

Regresión Lineal Multiple

Multidimensionalidad como respuesta.

Un regresión lineal múltiple implica una ecuación que ya no define una recta en el plano, sino un hiperplano en un espacio multidimensional.

El procedimiento de regresión lineal permite utilizar más de una variable independiente y, por tanto, permite llevar a cabo análisis de regresión múltiple.

En los capítulos anteriores se ha planteado la importancia de una representación gráfica, en el caso de una regresión lineal múltiple de 3 variables (2 independientes y 1 dependiente), se necesitan 3 ejes para poder representar el correspondiente diagrama de dispersión.

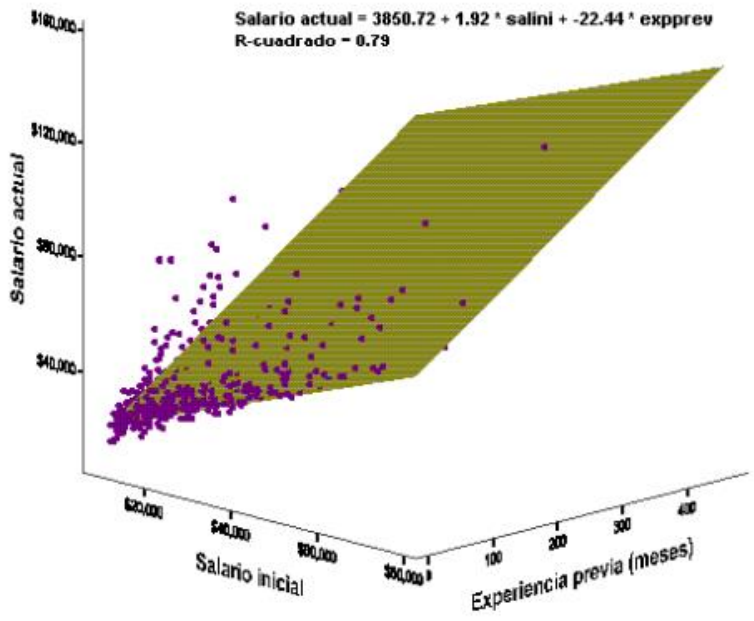


Fig 3-1 Diagrama de 3 dimensiones.

Ahora bien, si en lugar de 2 variables independientes tuviésemos 3, sería necesario un espacio de 4 dimensiones, y si tuviéramos 4 variables independientes, necesitaríamos un espacio de 5 dimensiones y así sucesivamente.

Por lo tanto, con más de una variable independiente, la representación gráfica de las relaciones presentes en un modelo de regresión es muy poco intuitiva, muy complicada y poco útil.

Es mucho más práctico partir de la ecuación del modelo de regresión lineal múltiple.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

En este modelo, la variable dependiente se interpreta como una combinación lineal de un conjunto de K variables independientes cada una de las cuales va acompañada de un coeficiente que indica el peso relativo de esa variable en la ecuación y además se pueden observar el componente constante y un componente aleatorio que recoge todo lo que las variables independientes no son capaces de explicar.

Al igual que en la regresión lineal simple, algunos supuestos son necesarios:

- Linealidad: la ecuación de regresión adopta una forma particular, en concreto, la variable dependiente es la suma de un conjunto de elementos:
    - el origen de la recta,
    - una combinación lineal de predictores y
    - los residuos.

  El incumplimiento del supuesto de linealidad suele denominarse error de especificación.

- Independencia: Los residuos son independientes entre sí, es decir, los residuos constituyen una variable aleatoria.

- Homocedestacidad: Para cada valor de la variable independiente, la varianza de los residuos es constante.

- Normalidad: Para cada valor de la variable independiente, los residuos se distribuyen normalmente con media cero.

- No-colinealidad: No existe relación lineal exacta entre ninguna de las variables independientes. El incumplimiento de este supuesto da origen a la colinealidad o multicolinealidad.

Adicional a todo lo abordado en el tema de regresiones lineales, es fundamental mencionar que existen muchos más tipos de regresiones[2]:

- Polynomial Regression

- Logistic Regression

[2] https://www.r-bloggers.com/15-types-of-regression-you-should-know/

- Quantile Regression
- Ridge Regression
- Lasso Regression
- ElasticNet Regression
- Principal Component Regression
- Partial Least Square Regression
- Support Vector Regression
- Ordinal Regression
- Poisson Regression
- Negative Binomial Regression
- Quasi-Poisson Regression
- Cox Regression

Algunos consejos básicos para poder seleccionar la estrategia de regresión son:

- Si la variable dependiente es continua y existe colinealidad o bien hay una gran cantidad de variables independientes, se pueden probar las regresiones de tipo PCR, PLS, Ridge, Lasso y Elastic. El modelo final se deberá seleccionar en función de R cuadrada, R Cuadrada Ajustada y si es posible calcular el AIC y BIC.
- Si tenemos un posible problema de sobre-entrenamiento, se puede utilizar Lasso, ridge y Elasticnet.
- Si la curva no describe una función lineal se puede evaluar una estrategia de polinomio.

Bibliografía

- Machine Learning for Beginners
  By Ken Richards, 2017

- R Data Analysis Cookbook
  by Kuntal Ganguly,2017

- Estadística Descriptiva: Series Temporales
  by Santiago de la Fuente Fernández