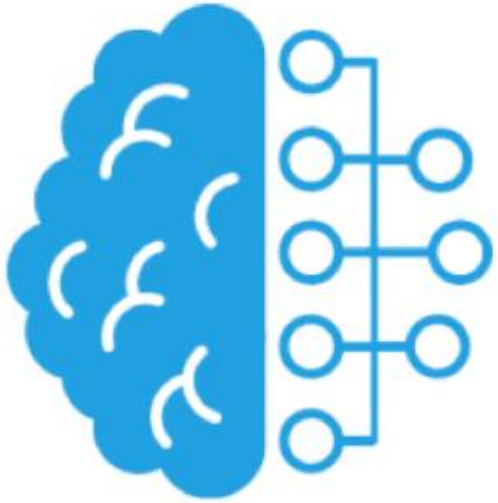
Aprendizaje de Máquina

Clase 6

Claudio Delrieux – DIEC - UNS
cad@uns.edu.ar



Regresión

En la Estadística, el análisis por regresión es la estimación de la relación entre una variable dependiente y una o más variables independientes. En este contexto, el foco es determinar relaciones causales entre observaciones.

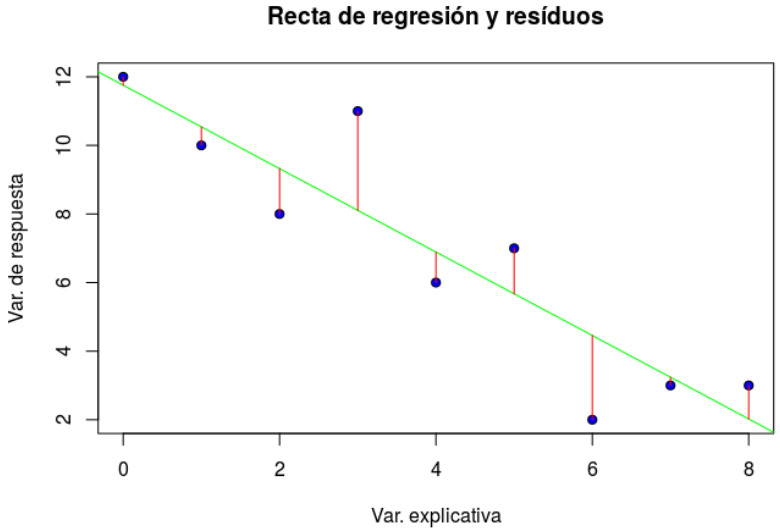
En nuestro contexto la regresión es el ajuste de un modelo supervisado para predecir un valor cuantitativo.

Definitivamente el modelo de regresión lineal univariada por cuadrados mínimos es el ejemplo mejor conocido.

Regresión Lineal

En la regresión lineal simple nuestro dataset es un conjunto de pares <x,y>, donde x es variable independiente o explicativa e y es la variable dependiente o de respuesta.

El modelo se interpreta gráficamente como la búsqueda de la recta de ajuste que minimice la suma de las distancias al cuadrado entre dichos pares y la recta.



Regresión Lineal

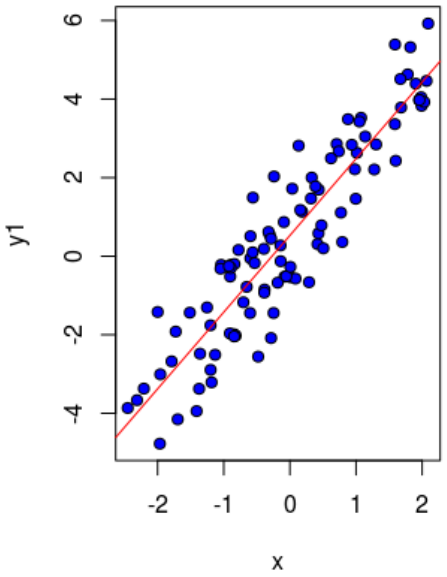
Los criterios de calidad de un modelo de regresión son varios, entre los cuales está principalmente el coeficiente de determinación R^2 (donde R es el índice de correlación de Pearson que establece un valor normalizado de relaciones entre productos de los momentos estadísticos de orden uno y dos).

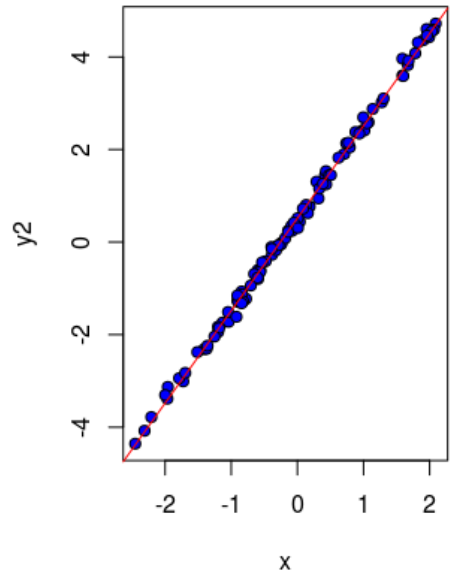
Para datasets muy alejados de distribuciones razonablemente ajustables a la normal, la correlación de Pearson es discutible (o directamente incorrecta) y se pueden utilizar estimadores no paramétricos como la correlación de Spearman (basada en rangos).

Regresión Lineal

El significado de R^2 es básicamente el poder explicativo del modelo de regresión lineal (a.k.a., cuánta de la variancia de la variable dependiente es explicada por la independiente).

R^2 necesita además una "interpretación" (p. ej., bajo, medio, alto) la cual es muy dependiente del contexto.





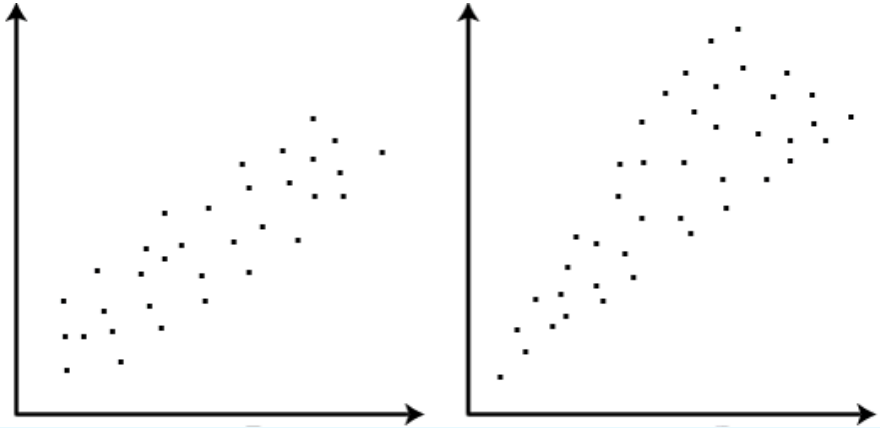
Regresión Lineal

Además del coeficiente de regresión (que indica un análogo a la "variancia" del modelo) se evalúan tambíen índices de error (análogos al "sesgo"), como por ejemplo el error absoluto medio, el RMSE, el SEE, etc.

La regresión lineal establece un único punto de equilibrio entre sesgo y variancia (de hecho, minimiza el RMSE), pero en regresión no lineal es posible plantear otras alternativas.

Regresión Lineal

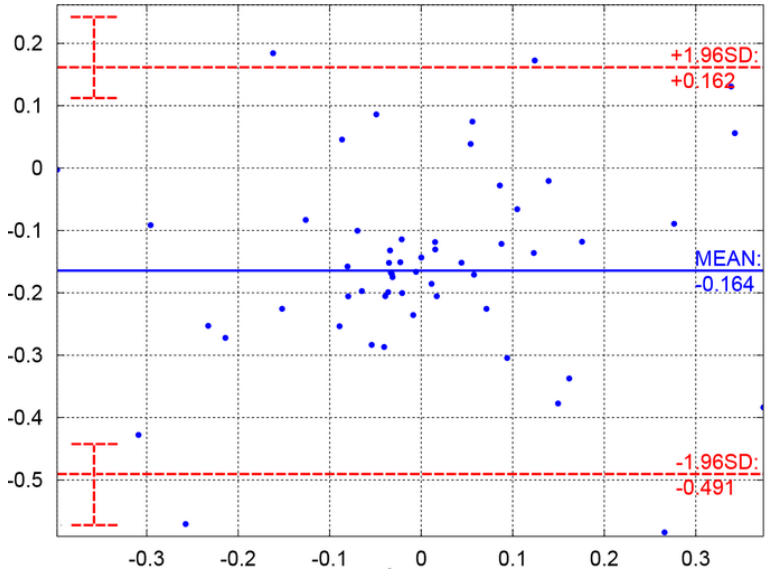
La regresión lineal también supone ciertas propiedades en los datos (que no siempre se observan ni se chequean). La más importante es la **homoscedasticidad** de los residuos (i.e., que los valores absolutos de los residuos sean independientes de los valores de la variable independiente). También, que los residuos sean no correlacionados (i.e., que la matriz de covariancia de los residuos sea diagonal).



Regresión Lineal

Existen otras maneras de representar el poder explicativo de un modelo de regresión, por ejemplo el diagrama Bland-Altman (diferencia vs. media, en el espacio de las variables o en el log-log), donde se observa más la concordancia entre pares de valores.

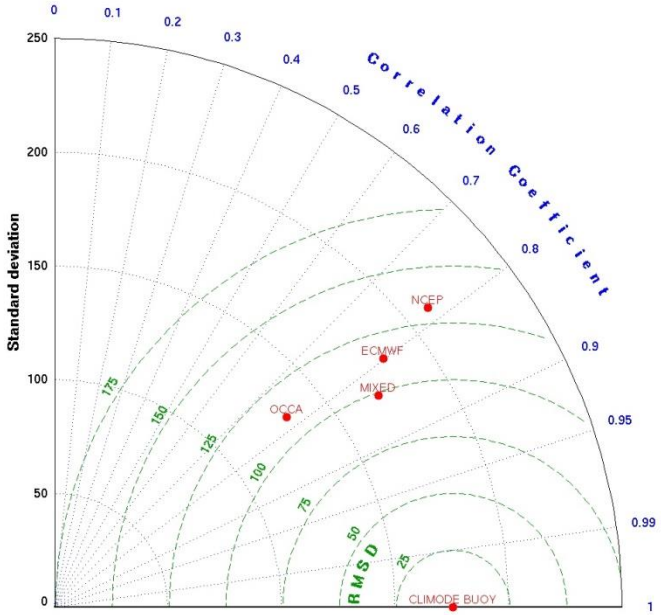
Por ejemplo, la correlación puede ser alta solo porque los valores extremos coincidan.



Regresión Lineal

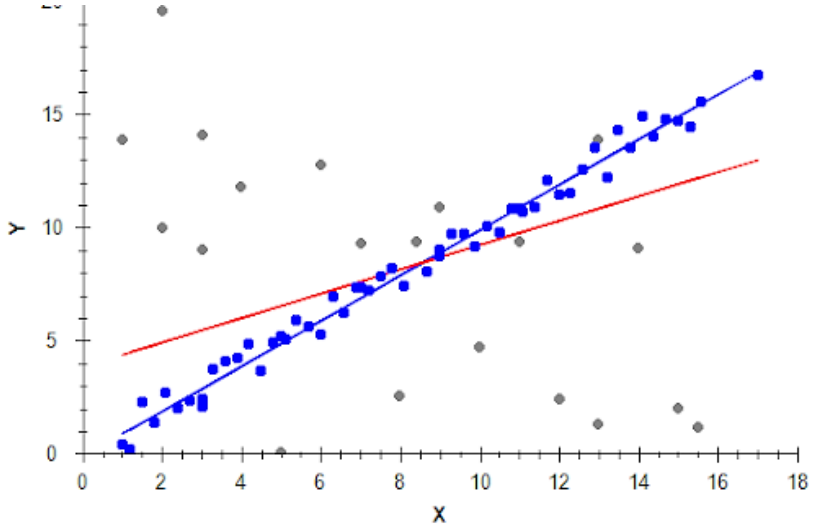
Hay aún técnicas gráficas menos conocidas pero muy poderosas para comparar modelos de regresión, como por ejemplo el diagrama de Taylor (inicialmente correlación vs. desvío estándar).

Ver un lindo tutorial y biblioteca en https://cdat-vcs.readthedocs.io/en/latest/notebooks/Taylor_Diagrams.html



Regresión Lineal

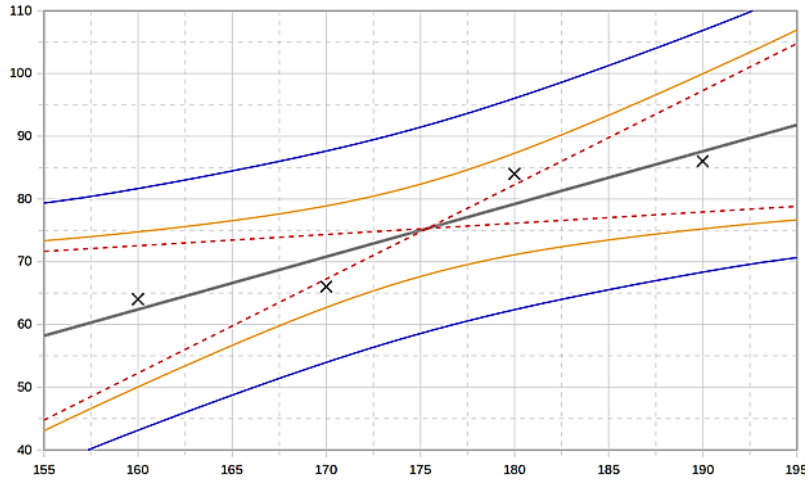
Otro tema importante relacionado con los modelos de regresión es controlar el efecto de los atípicos. En la imagen tenemos en rojo la recta de regresión, y en azul un posible modelo lineal alternativo que busca “consensuar” la mayor cantidad de puntos (haciendo que los puntos por fuera sean atípicos).



Esto se logra con funciones que “saturan” la contribución del error cuadrático de cada punto durante el cómputo del modelo.

Regresión Lineal

Una última cuestión antes de ir a modelos más generales está relacionada con la confiabilidad del modelo para hacer extrapolaciones (predicciones por fuera del rango de valores de la variable independiente). Un rápido análisis visual nos muestra que las rectas de ajuste posibles todas pasan por el centroide, y por lo tanto las de pendientes más extremas conforman asíntotas. Esto implica que la recta de ajuste tiene una "zona de confianza" que se vuelve más gruesa a medida que nos alejamos.



Regresión Lineal Múltiple

El modelo en este caso es una extensión natural del anterior, en vez de una sola variable independiente, tenemos varias:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

donde i representa cada uno de los datos, y p es la dimensionalidad del espacio de atributos. En este modelo la representación visual es un poco más difícil que con la RL simple, por lo que utilizamos trellis o reticulados de gráficos. En https://github.com/manlio99/Materia-de-aprendizaje/blob/master/5_DataMining/1_regression.ipynb mostramos un ejemplo de un regresor múltiple con un dataset conocido.

Regresión No Lineal

En principio, para la regresión no lineal aplican las mismas ideas, solamente que el modelo es una función no-lineal.

El ajuste del modelo tiene varios aspectos relativamente complejos relacionados con la convergencia y la estabilidad numérica, pero puede pensarse muy simplificadamente como una regresión lineal en un espacio donde la variable dependiente está “warpeada” (por la función inversa).

https://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html#sphx-glr-auto-examples-model-selection-plot-underfitting-overfitting-py

Regresión No Lineal

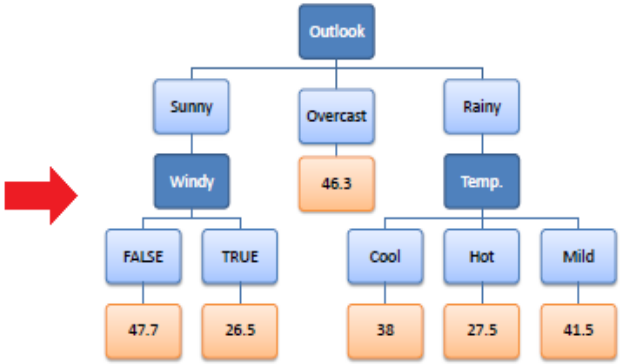
En la regresión no lineal, interviene además otro aspecto: la regularización, dado que también queremos controlar la complejidad del modelo final. Existen varios métodos:

- Lasso (least absolute shrinkage and selection operator)
- Ridge
- Elastic net
- RANSAC
- LARS
- Tweedie
- etc. etc.

Regresión no paramétrica

Son necesarios cuando las variables independientes no son numéricas. En este contexto básicamente reutilizamos algunos de los modelos de clasificación para tareas de regresión. Por ejemplo CART (classification and regression trees) es un modelo que permite generalizar los árboles de decisión para generar clasificaciones.

Predictors				Target
Outlook	Temp	Humidity	Windy	Hours Played
Rainy	Hot	High	False	25
Rainy	Hot	High	True	30
Overcast	Hot	High	False	48
Sunny	Mild	High	False	46
Sunny	Cool	Normal	False	62
Sunny	Cool	Normal	True	23
Overcast	Cool	Normal	True	43
Rainy	Mild	High	False	36
Rainy	Cool	Normal	False	38
Sunny	Mild	Normal	False	48
Rainy	Mild	Normal	True	48
Overcast	Mild	High	True	62
Overcast	Hot	Normal	False	44
Sunny	Mild	High	True	30



Regresión no paramétrica

Además de árboles (y por extensión random forests), muchos otros clasificadores generalizan a regresores:

- K-NN (se genera un valor como combinación ponderada de los valores de los K vecinos más cercanos).
- SVR (se utiliza la idea de máxima separación, en este caso como minimización del margen de error).
- Regresión Bayesiana (la variable dependiente tiene una densidad a-priori).
- Redes neuronales (se entrenan para predecir un valor de activación de salida a partir de las activaciones de las entradas).

Práctico 6

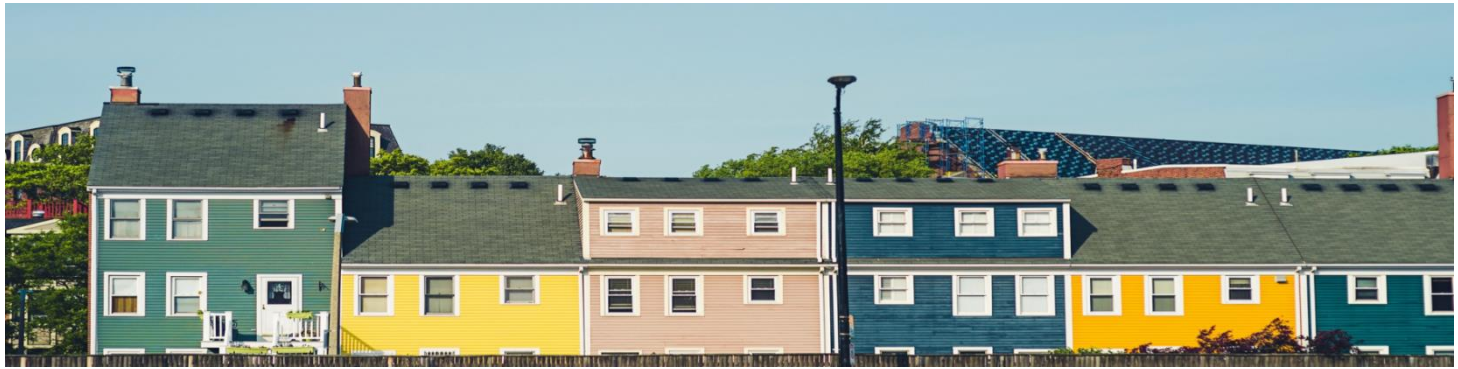
Elegir uno de entre los siguientes proyectos sugeridos y aplicar regresión para resolverlo. Cada proyecto tiene datasets de diferentes características y dificultad (indicada con el semaforito).

Para los más sencillos trabajar individualmente y aplicar más de un método de regresión, utilizar diferentes parámetros, comparar evaluaciones. Para los más complejos, trabajar en grupos de dos personas y aplicar un único modelo.

Si encuentran algún otro caso de regresión (que no esté totalmente resuelto) lo pueden proponer para negociarlo con la cátedra.

Ejercicio 6.1

Boston Housing Prices: El dataset contiene 506 observaciones de 14 variables y precios de venta de propiedades en la ciudad de Boston. El objetivo es conocer la influencia de cada variable y elaborar un modelo que prediga precios de venta. https://github.com/manlio99/Materia-de-aprendizaje/tree/master/3_MidtermProjects/ProjectBHP



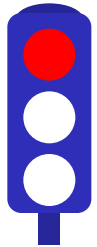
Ejercicio 6.2

Sueldos en la NBA: Tenemos dos archivos, uno con las estadísticas de juego de cada jugador, y otro con los sueldos por jugador en las temporadas 1985 a 2018. El objetivo es predecir el sueldo de un jugador a partir de sus estadísticas. https://github.com/manlio99/Materia-de-aprendizaje/tree/master/3_MidtermProjects/ProjectNBA

Ejercicio 6.3

Retrasos en los vuelos de enero: El dataset contiene 400.000+ registros, cada uno con 21 atributos, de vuelos realizados entre el 1/1/19 y el 31/1/19. El objetivo es predecir las demoras o retrasos de un vuelo en particular.

https://github.com/manlio99/Materia-de-aprendizaje/tree/master/3_MidtermProjects/ProjectFD



Ejercicio 6.4

IMDB Rating: El datast contiene 5000+ películas, con 28 atributos y su calificación. El objetivo es predecir el ranking de una película a partir de sus atributos. https://github.com/manlio99/Materia-de-aprendizaje/tree/master/3_MidtermProjects/ProjectIMDB

Ejercicio 6.5

Popularidad en Spotify: El dataset contiene un ranking de las 200 canciones de Spotify más escuchadas en 53 países durante 2017 y 2018. El objetivo es entender cuáles son los atributos que hacen que una canción sea popular. https://github.com/manlio99/Materia-de-aprendizaje/tree/master/3_MidtermProjects/ProjectTSS

Ejercicio 6.6

Ese vino lo vale?: El dataset contiene un scrappeo de 300K+ reviews de vinos del sitio Wine Enthusiast durante 2017. El objetivo es determinar cuáles tienen un review positivo, y si dicho review se correlaciona con el precio de venta. https://github.com/manlio99/Materia-de-aprendizaje/tree/master/3_MidtermProjects/ProjectWNF