



**UNIVERSIDAD
CENTRAL**

Ingeniería de Sistemas
Data Analytics
Prof: Hugo Franco

**Sesión N° 14 | Correlación – Problemas
de Regresión**

Bogotá D.C., Abril 20, 2022

Identificación de relaciones
entre variables

Repaso: medidas de asociación lineal: covarianza

Datos cuantitativos

Covarianza: Es una medida de la **variabilidad conjunta** entre dos variables (x_1, x_2) o (x, y)

x	y
$x_{(1)}$	$y_{(1)}$
$x_{(2)}$	$y_{(2)}$
$\vdots$	$\vdots$
$x_{(n)}$	$y_{(n)}$

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- Si $\text{cov}(x, y)$ es positiva: la asociación entre x e y es directamente proporcional, es decir, cuando x aumenta y también aumenta; y viceversa.
- Si $\text{cov}(x, y)$ es negativa: la asociación entre x e y es inversamente proporcional, es decir, cuando x aumenta y disminuye; y viceversa.
- Si $\text{cov}(x, y)$ es cero: no hay asociación entre x e y .

Repaso: medidas de asociación lineal: correlación

Datos cuantitativos

Correlación: Cuantifica el grado de asociación entre dos variables. (x_1, x_2) o (x, y)

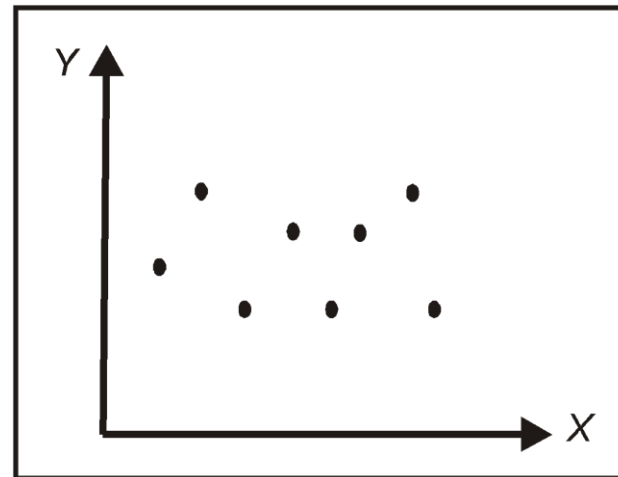
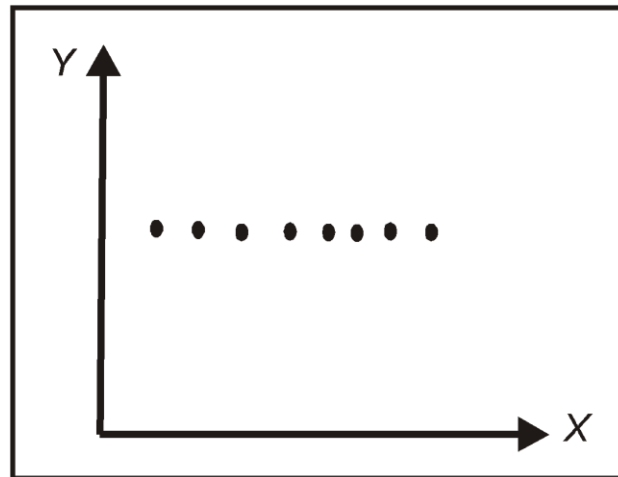
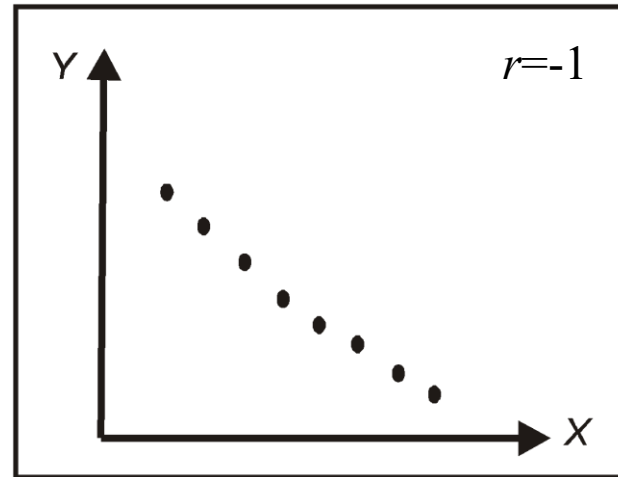
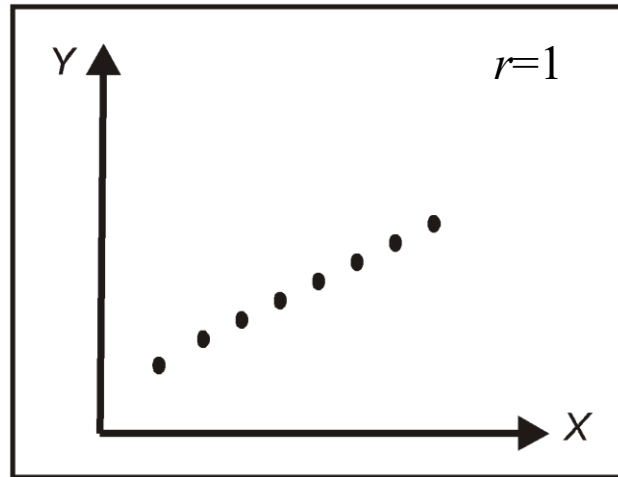
Coefficiente de Correlación de Pearson (r): mide el grado de asociación lineal entre dos variables cuantitativas

x	y
$x_{(1)}$	$y_{(1)}$
$x_{(2)}$	$y_{(2)}$
$\vdots$	$\vdots$
$x_{(n)}$	$y_{(n)}$

$$r = \frac{\text{cov}(x, y)}{s_x s_y} \dots \rightarrow r = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{(n-1) s_x s_y} \quad -1 \leq r \leq 1$$

- Si r es positivo: la asociación entre x e y es directamente proporcional, es decir, cuando x aumenta y también aumenta; y viceversa.
 - Si $r = 1$: la asociación lineal es perfecta.
- Si r es negativo: la asociación entre x e y es inversamente proporcional, es decir, cuando x aumenta y disminuye; y viceversa.
 - Si $r = -1$: la asociación lineal es perfecta.
- Si r es cero: no hay asociación entre x e y .

Representación gráfica y significado de la correlación



Problemas de regresión

Problemas de regresión

- La regresión busca relaciones cuantitativas entre variables continuas definidas en conjuntos de datos.
 - *Ejemplo:* Encontrar la dependencia entre la masa corporal y la altura de una población, en el estudio de los problemas de sobrepeso y obesidad
- Las medidas realizadas para determinar los valores de las variables en una *muestra representativa* de la población (un subconjunto de elementos o sujetos de la misma) se conocen como **observaciones**

Análisis de regresión

- Usualmente, en *análisis de regresión* se estudia algún fenómeno existente sobre el total de la población, por medio de un número representativo de observaciones. Cada observación tiene dos o más *características*.
- Siguiendo el supuesto de que (al menos) una de las características depende de las otras, se busca establecer una relación matemática (**función**) entre ellas que sea *consistente* con los datos.
 - Las características dependientes se denominan *variables dependientes*, *salidas o respuestas*.
 - Las características independientes se denominan *variables independientes*, *entradas o predictores*.
 - La función obtenida corresponde a una **parametrización** de un **modelo** propuesto (según el comportamiento observado en los datos), y asigna valores a las salidas que sigan dicho comportamiento, en función de los *predictores*.

Aplicaciones del análisis de regresión

- El análisis de regresión permite responder si un fenómeno influye en el otro y cómo se relacionan las variables que los describen.
 - Tratándose de modelos que se ajustan a los datos y buscan descubrir las relaciones entre los mismos, el análisis de regresión es una aproximación *fenomenológica*.
- La regresión también es útil cuando desea pronosticar una respuesta utilizando un nuevo conjunto de predictores.
 - P.ej, predecir el consumo de electricidad de una vivienda durante un periodo próximo (hora/día/semana/mes) dada la temperatura exterior, la hora del día y el número de habitantes.
- Algunos modelos permiten predecir los valores futuros de una variable en función de sus valores previos y presente
 - P.ej. Predecir el consumo de electricidad de una vivienda, conjunto residencial, edificio o ciudad, en función de series históricas de su consumo. Estos modelos suelen denominarse “autorregresivos”
- La regresión se utiliza en muchos campos diferentes: economía, informática, ciencias sociales, etc. Su importancia aumenta cada día con la disponibilidad de grandes cantidades de datos y una mayor conciencia del valor práctico de los datos.

Regresión lineal

Regresión lineal

- La regresión lineal, una de las técnicas de regresión más importantes y más utilizadas, es a la vez uno de los métodos de regresión más sencillos.
 - Permite interpretar con facilidad para los resultados del proceso de regresión.
- Para obtener la regresión lineal de una variable dependiente y sobre el conjunto de variables independientes $\mathbf{x} = (x_1, \dots, x_r)$, donde r es el número de predictores, se asume la relación lineal:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_r x_r + \varepsilon.$$

esta es la **ecuación de regresión**, donde $\beta_0, \beta_1, \dots, \beta_r$ son los **coeficientes de regresión**, y ε es el **error del ajuste**, de tipo aleatorio (usualmente ruido blanco).

- Así, la regresión lineal calcula los **estimadores** de los coeficientes de regresión o **pesos**, denominados $b_0, b_1, \dots, b_r$. La **función de regresión estimada** se escribe como

$$f(\mathbf{x}) = b_0 + b_1 x_1 + \dots + b_r x_r$$

captando así la dependencia lineal entre las entradas y la salida (*predictores y respuesta*).

Estimación de la regresión lineal

- La **respuesta estimada**, $f(\mathbf{x}_i)$, para cada observación $i = 1, \dots, n$, debe ser tan cercana como sea posible a la **respuesta real** correspondiente, y_i
- Las diferencias $y_i - f(\mathbf{x}_i)$ para la i -ésima observación, $i = 1, \dots, n$, se conocen como **residuos**. Así pues, la regresión busca determinar los mejores **pesos** pronosticados, i.e., encontrar los *pesos correspondan a los residuos más pequeños* (o, dicho de otra forma, que los minimicen)
- Para obtener los mejores pesos, se suele **minimizar la Suma de los Residuos Cuadráticos** para todas las observaciones (*método de los mínimos cuadrados ordinario*), i.e. :

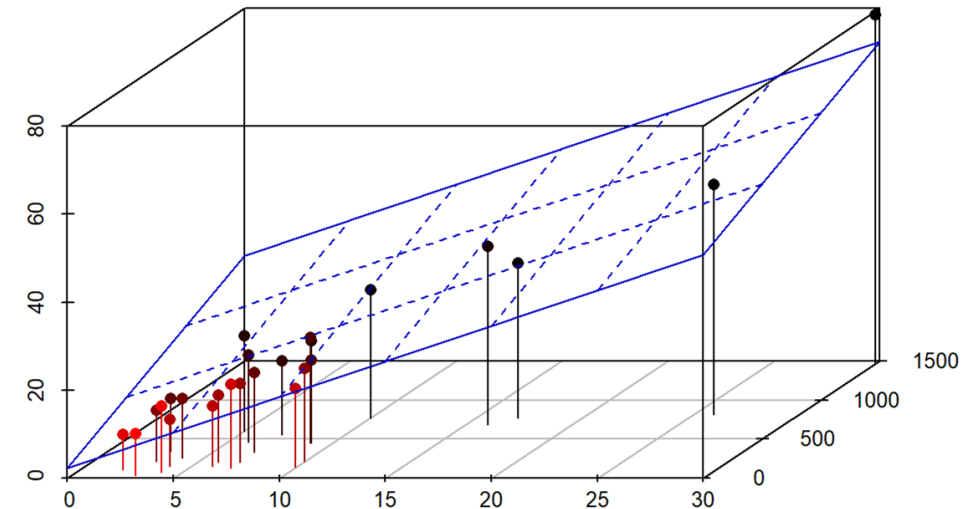
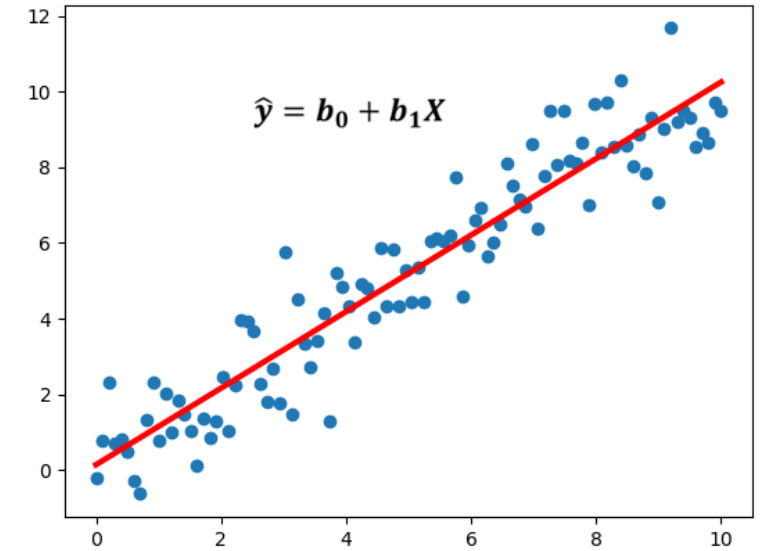
$$\text{SRC} = \sum_i (y_i - f(\mathbf{x}_i))^2$$

Coeficiente de determinación

- La variación en las respuestas reales y_i , $i = 1, \dots, n$, se produce en parte por la dependencia respecto de los predictores $\mathbf{x}_i$. Sin embargo, también hay una varianza adicional en la salida, inherente a la naturaleza de los datos (ruido, factores no contemplados, etc.)
- El coeficiente de determinación R^2 , correspondiente al cuadrado del coeficiente de correlación de Pearson, expresa la magnitud de la variación en y que puede explicarse por su dependencia respecto de $\mathbf{x}$ utilizando el modelo de regresión particular
 - Un R^2 más grande indica un mejor ajuste y significa que el modelo puede explicar mejor la variación de la salida para diferentes entradas
- El valor $R^2 = 1$ corresponde a un valor de SRC = 0, es decir, al ajuste perfecto, ya que los valores de y_i y $f(\mathbf{x}_i)$ se ajustan completamente entre sí para todo valor de i
 - Así, el coeficiente de determinación expresa el desempeño del proceso de regresión

Regresión lineal simple vs. regresión multilineal

- En la regresión lineal simple, solo existe una característica (variable independiente) y la regresión relaciona la salida con los valores del predictor (arriba), así $f(x) = b_0 + b_1x$
 - La línea roja corresponde a $f(x)$. El residuo es la suma cuadrática de la distancia de cada punto a la recta de regresión
- En la regresión lineal múltiple (abajo), existen varias características que describen cada observación, y la salida del modelo corresponde a: $f(\mathbf{x}) = b_0 + b_1x_1 + \dots + b_rx_r$
 - El hiperplano azul corresponde a la regresión



Regresión polinomial

Regresión polinomial

- La **regresión polinomial** asume una *dependencia funcional polinomial* entre la salida y las entradas (y, en consecuencia, la función de regresión estimada del polinomio).
 - Además de los términos lineales como b_1x_1 , la función de regresión f puede incluir términos no lineales como $b_2x_1^2$, $b_3x_1^3$ o incluso términos cruzados como $b_4x_1x_2$, $b_5x_1^2x_2$, etc.
- El ejemplo más sencillo de regresión polinomial corresponde a una sola variable independiente con una la función de regresión estimada equivalente a un polinomio de grado 2: $f(x) = b_0 + b_1x + b_2x^2$.
 - Las incógnitas son los valores b_0 , b_1 y b_2 que minimizan la suma de residuos cuadráticos (SRC)

Solución al problema de regresión polinomial

La regresión polinomial se puede formular como un caso generalizado de la regresión lineal. P. ej.

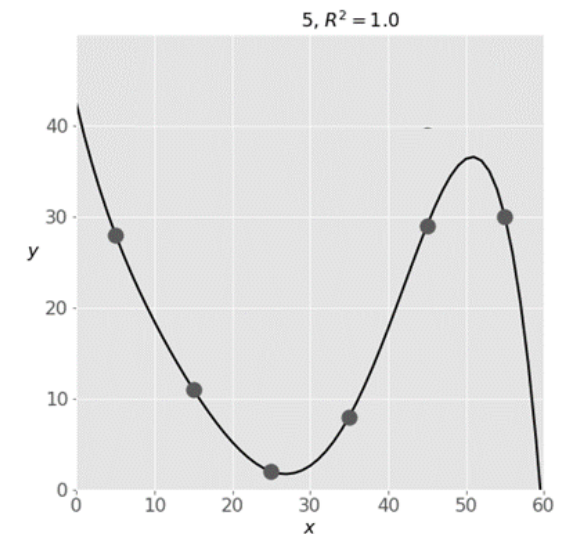
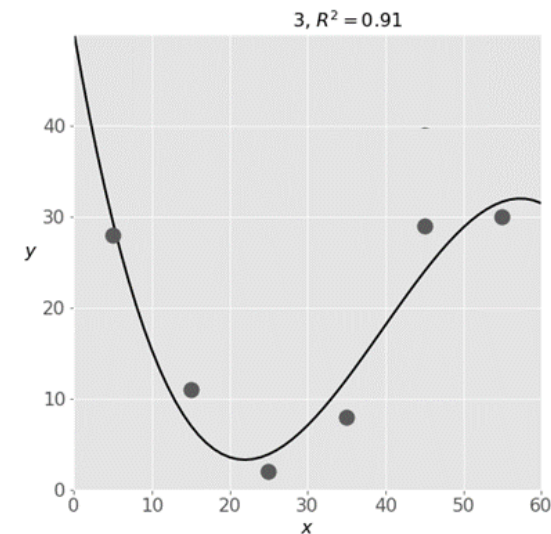
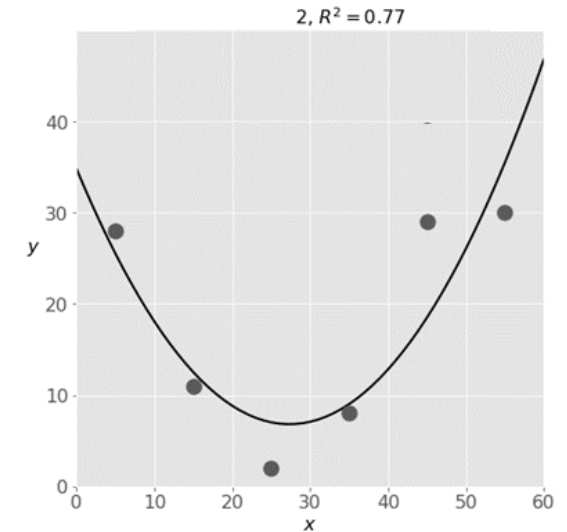
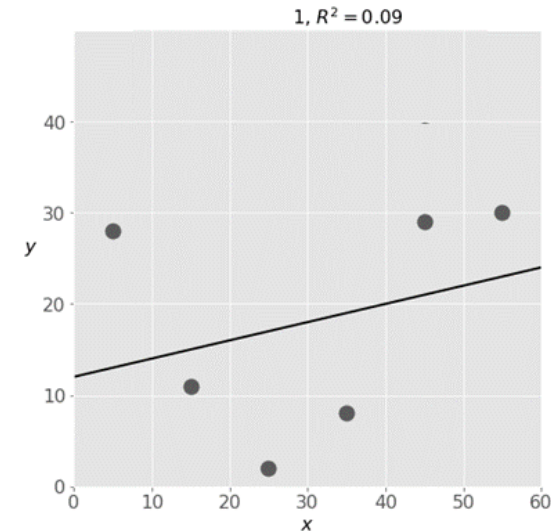
- La $f(x) = b_0 + b_1x + b_2x^2$ se puede expresar como una función de regresión lineal, i.e. $f(x_1, x_2) = b_0 + b_1x_1 + b_2x_2$ con incógnitas b_0, b_1 y b_2 . Así, el problema de regresión polinomial se puede solucionar como un problema lineal con el término x^2 considerado como una variable de entrada.
- En el caso de dos variables y polinomio de grado 2, la función de regresión se expresa como $f(x_1, x_2) = b_0 + b_1x_1 + b_2x_2 + b_3x_1^2 + b_4x_1x_2 + b_5x_2^2$ y la regresión lineal asociada se formula para para cinco entradas: $x_1, x_2, x_3 = x_1^2, x_4 = x_1x_2$ y $x_5 = x_2^2$, mientras que la solución de la regresión corresponde a seis pesos, b_0, b_1, b_2, b_3, b_4 y b_5 , que minimizan la SRC

Casos de orden superior se formulan y solucionan bajo la misma lógica, agregando tantas variables lineales como potencias y términos cruzados se hagan presentes.

Subajuste y sobreajuste (underfitting y overfitting) I

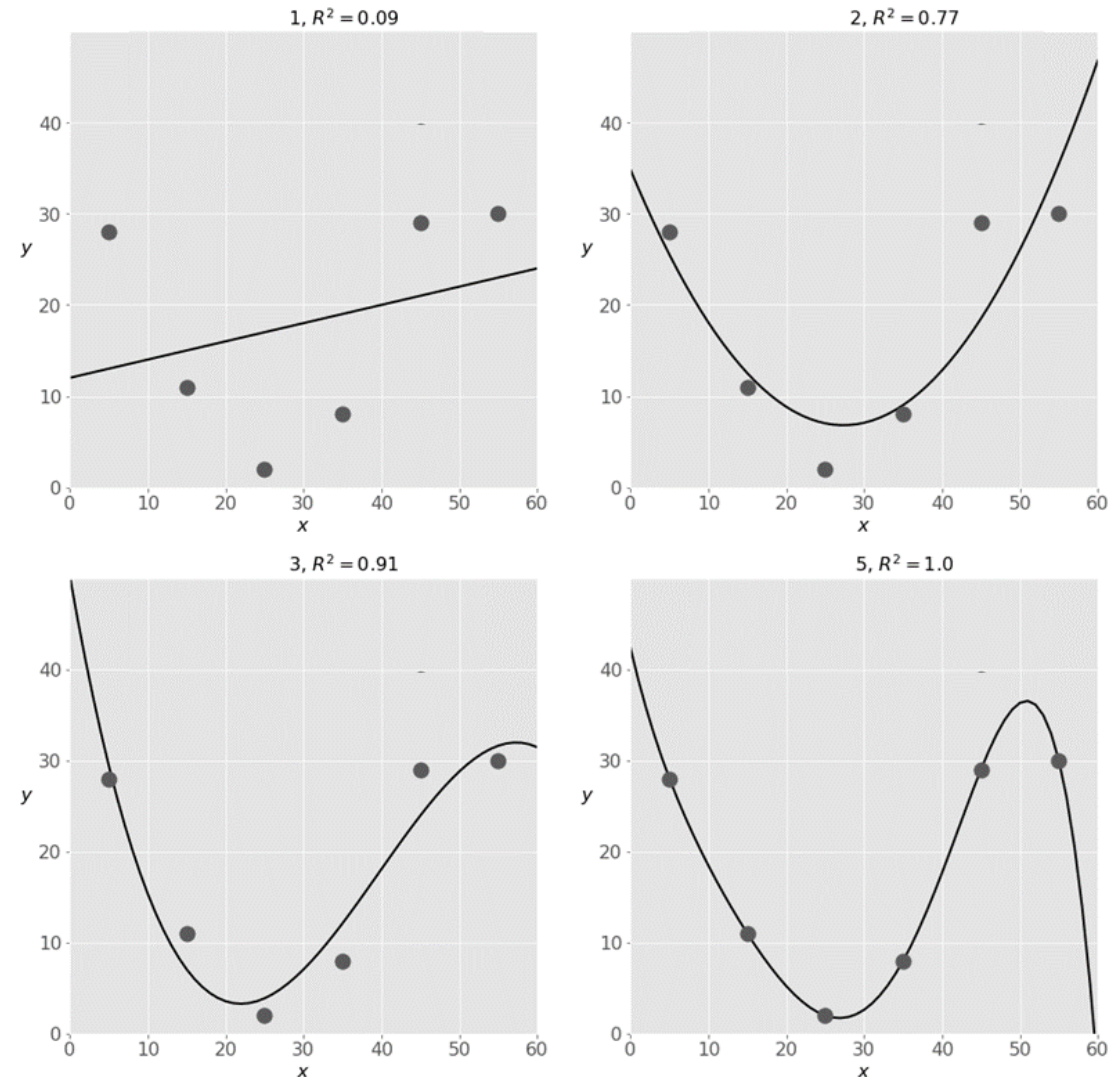
No existe una regla sencilla para determinar el grado ideal del polinomio de regresión. Sin embargo, se deben tener en cuenta dos problemas relacionados con dicha elección: subajuste y sobreajuste.

- El **subajuste** ocurre cuando un modelo no puede capturar con precisión las dependencias entre los datos, usualmente por su simplicidad.
 - A menudo produce un R^2 bajo con datos conocidos y capacidades de generalización deficientes cuando se aplica con datos nuevos.



Subajuste y sobreajuste (underfitting y overfitting) II

- El **sobreajuste** ocurre cuando un modelo “aprende” tanto las dependencias entre los datos como las fluctuaciones aleatorias.
- Así, el modelo “memoriza” los datos existentes. Los modelos complejos con muchas características o parámetros son propensos a sobreajustarse.
 - Cuando se aplican a datos conocidos, estos modelos suelen producir R^2 altos, pero no son capaces de generalizar y los valores de R^2 caen más bajos cuando se utilizan con datos nuevos (en producción).



Modelos autorregresivos

Modelos autorregresivos (introducción)

- Un modelo autorregresivo busca encontrar una función que describa la dinámica de una variable (usualmente dependiente del tiempo) para predecir nuevos valores a partir de valores previos.

$$y_t = \delta + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \epsilon_t$$

donde δ es una constante y ϵ es una función de ruido blanco (aleatorio, media cero, varianza constante).

- Los coeficientes ϕ_i establecen la contribución de muestras en valores (instantes) previos a la salida en el nuevo valor (pronosticado)
- De este modo, se pueden identificar comportamientos según registros históricos.
- Modelos multivariados pueden incorporar variables “explicativas” y/o introducir términos estacionales (periodicidad).

Su utilidad, y los detalles técnicos de la implementación de este tipo de modelos, está estrechamente ligada al análisis de series temporales