Pontificia Universidad
JAVERIANA
Bogotá

VIGILADA MINEDUCACIÓN

Regresión Lineal

Pontificia Universidad
JAVERIANA
Colombia

Regresión lineal simple

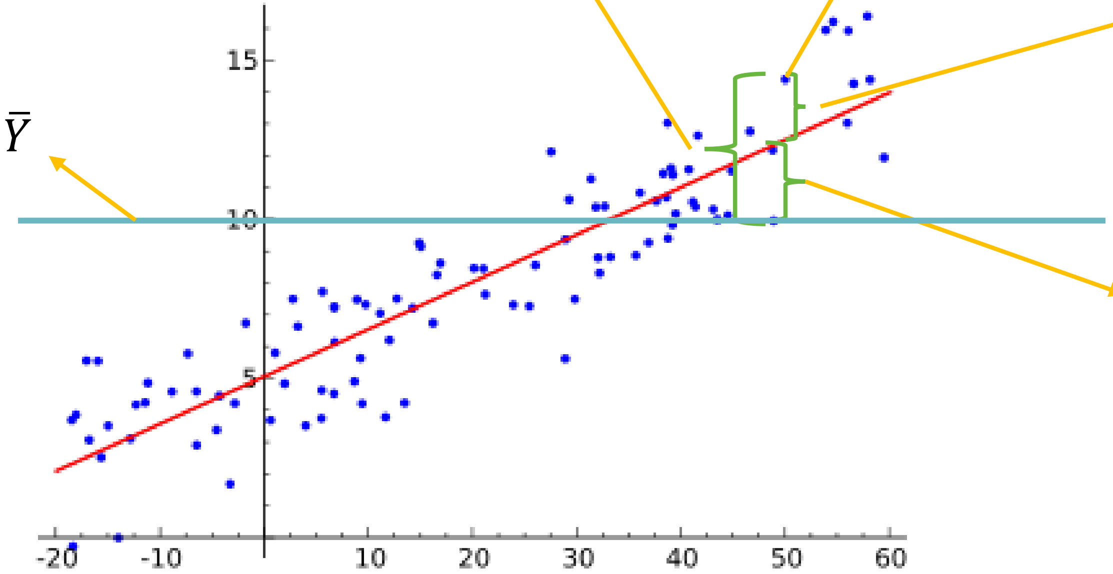
$$Y = b + mx$$

Distancia de la media a la observación $(Y_i - \bar{Y})$

Observación Y_i

Distancia de la predicción a la observación $(Y_i - \hat{Y}_i)$

$\bar{Y}$



Distancia de la Media a la predicción $(\hat{Y}_i - \bar{Y})$

Regresión lineal

Dos interpretaciones → Estadística clásica

Dos interpretaciones → Machine Learning

- El promedio de la variable dependiente (Y) bajo la condición de un valor de la variable independiente (X) - Factor explicativo.

- El mejor ajuste de los datos a una recta - Factor predictivo.

Interpretación de los coeficientes de la Regresión lineal

- El coeficiente m es el grado en que varía la variable dependiente (Y) por cada unidad de la variable independiente (X) - Razón de cambio.

- El coeficiente b es el de mejor ajuste a los datos o el resultado de Y (dependiente) cuando X es cero (independiente).

Regresión lineal múltiple

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

Método del Gradiente

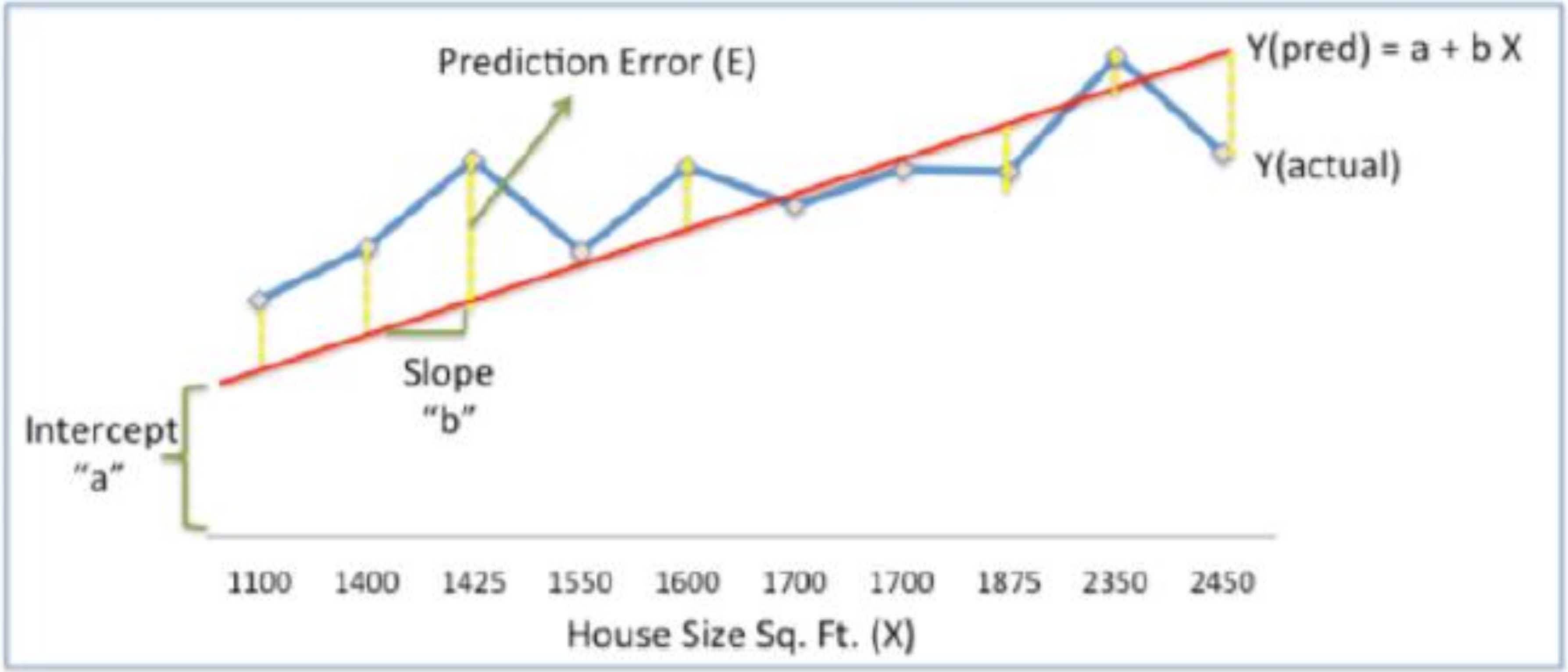
- Tomemos el ejemplo de pronosticar el precio de una casa a partir de su área en pies cuadrados

House Size sq.ft (X)	1400	1600	1700	1875	1100	1550	2350	2450	1425	1700
House Price$ (Y)	245,000	312,000	279,000	308,000	199,000	219,000	405,000	324,000	319,000	255,000

Dataset: Regresión Lineal con Optimización.xlsx

Dataset: Casas.csv

Método del Gradiente



- El problema consiste en hallar valores de a y b que minimizan el error de predicción total (SSE)

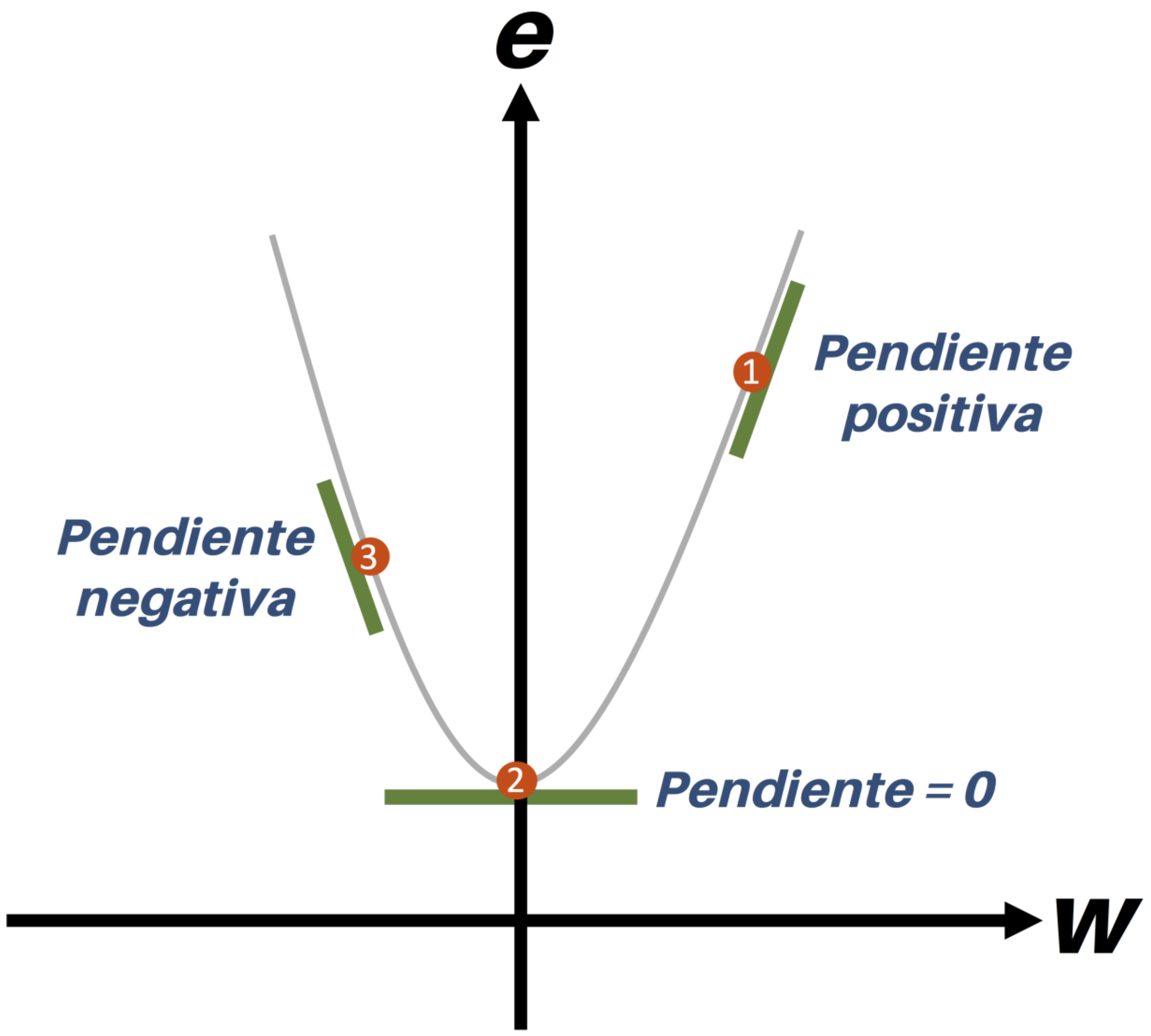
Método del Gradiente

$$SSE = \frac{1}{2} * \sum_{i=1}^{10} (Y_i - Ypred_i)^2$$

HOUSING DATA	
House Size (X)	House Price (Y)
1,100	1,99,000
1,400	2,45,000
1,425	3,19,000
1,550	2,40,000
1,600	3,12,000
1,700	2,79,000
1,700	3,10,000
1,875	3,08,000
2,350	4,05,000
2,450	3,24,000

1. Inicializar los valores a y b con valores aleatorios y calcular SSE
2. Calcular el gradiente: el cambio en SSE cuando los coeficientes a y b, son cambiados en un valor muy pequeño, de su valor inicial (ayuda a movernos en la dirección en la que SSE es minimizado).
3. Ajustar los coeficientes con los gradientes, para alcanzar los valores óptimos, cuando SSE sea minimizado
4. Usar los nuevos coeficientes para predicción y calcular el nuevo SSE
5. Repetir los pasos 2 y 3 hasta que los ajustes en los coeficientes no reduzca significativamente el error.

Optimización usando la derivada (gradiente) de una función.



Cómo hallar los coeficientes del modelo de regresión lineal

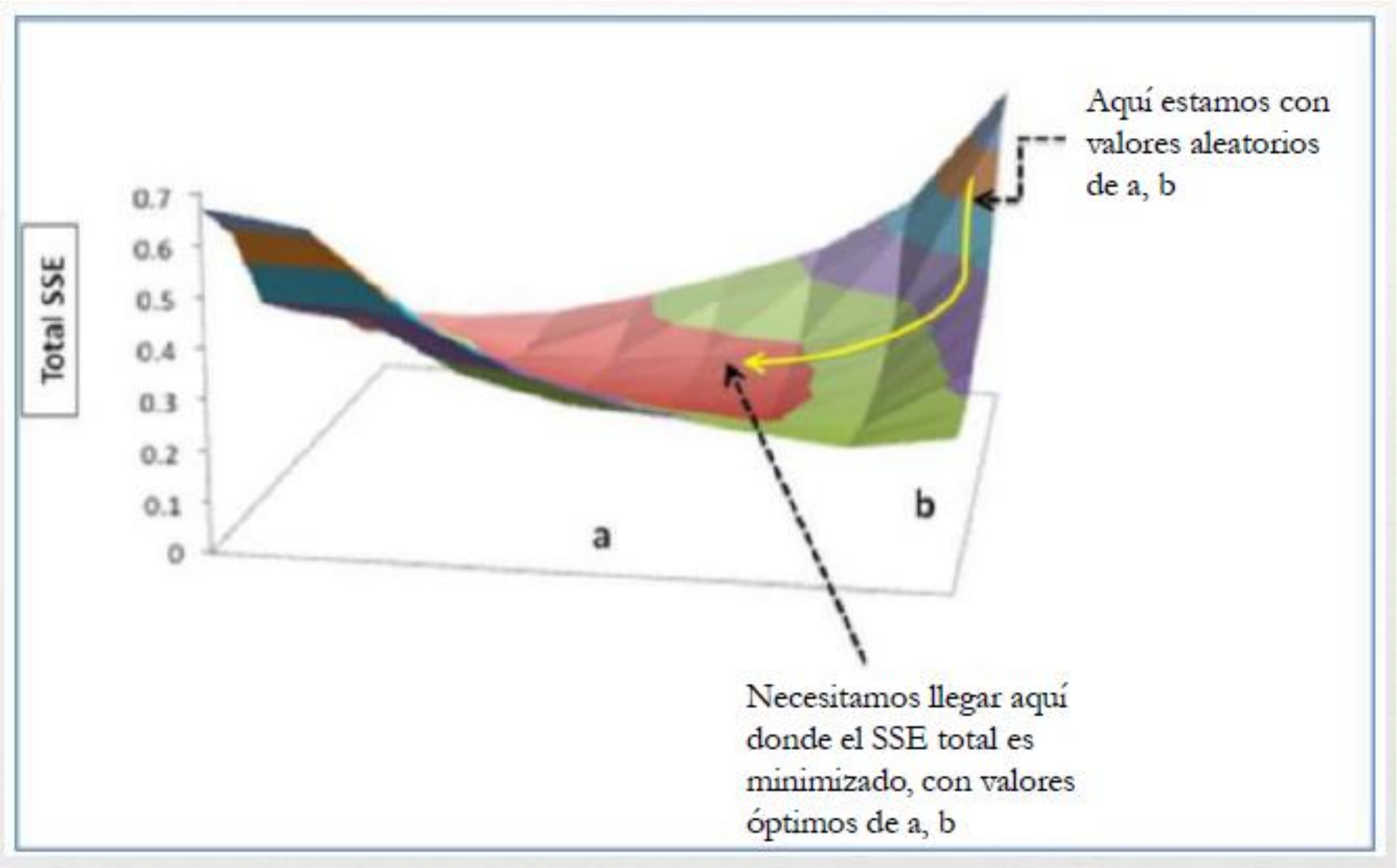
MÉTODO DEL GRADIENTE – Minimizar f(x)

1. Escoger un punto inicial en R_n
2. Determinar una dirección de movimiento
3. Moverse en esa dirección de acuerdo a algún criterio determinando un nuevo punto
4. Verificar criterio de parada. Si no se satisface, volver a 2.

Preguntas:

- Qué dirección de movimiento escoger?
- Cómo determinar un punto para detenerse?

Método del Gradiente



Método del Gradiente – Dirección de movimiento

$$\vec{x}_{k+1} = \vec{x}_k - \alpha \nabla f(\vec{x}_k)$$

Análisis de Varianza

PRUEBAS DE HIPÓTESIS:

1. Prueba F de significación global

H_0: el modelo sin variables independientes se ajusta a los datos tan bien como nuestro modelo
H_1: nuestro modelo se ajusta a los datos mejor que el modelo con solo el intercepto

2. Pruebas t para significancia de cada coeficiente (una prueba por cada coeficiente)

H_0: La variable no tiene significado o no está asociada $(\beta = 0)$
H_1: La variable tiene significado y está asociada $(\beta \neq 0)$

Calidad de la Regresión Lineal (Métricas básicas de pérdida)

01 **S:** Variabilidad (error) de la predicción. Está en unidades de la variable.

02 **MSE =** S^2 (métrica clásica).

03 $\boldsymbol{R^2}$ **(r-sq)** = Porcentaje de la variabilidad que logra captar o recoger el conjunto de variables dependientes (comparada contra el promedio).

04 **Análisis de varianza** (prueba de hipótesis).

Coeficiente de determinación, R2

- El coeficiente de determinación se define por:

$$R^2 = \frac{SC\ Regresión}{SC\ Total} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Es decir, mide la proporción de la variación total de la variable dependiente que es explicada por la variación en las variables independientes

R2 AJUSTADO

- EL R2 múltiple mide la cantidad de variación de la variable dependiente que pueden explicar las variables independientes. Cuando se añaden predictores al modelo, el R2 múltiple siempre aumentará, ya que un predictor siempre explicará una parte de la varianza.
- El R2 ajustado controla este aumento y añade penalizaciones por el número de predictores en el modelo. Por lo tanto, muestra un equilibrio entre el modelo más moderado y el modelo que mejor se ajusta.

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

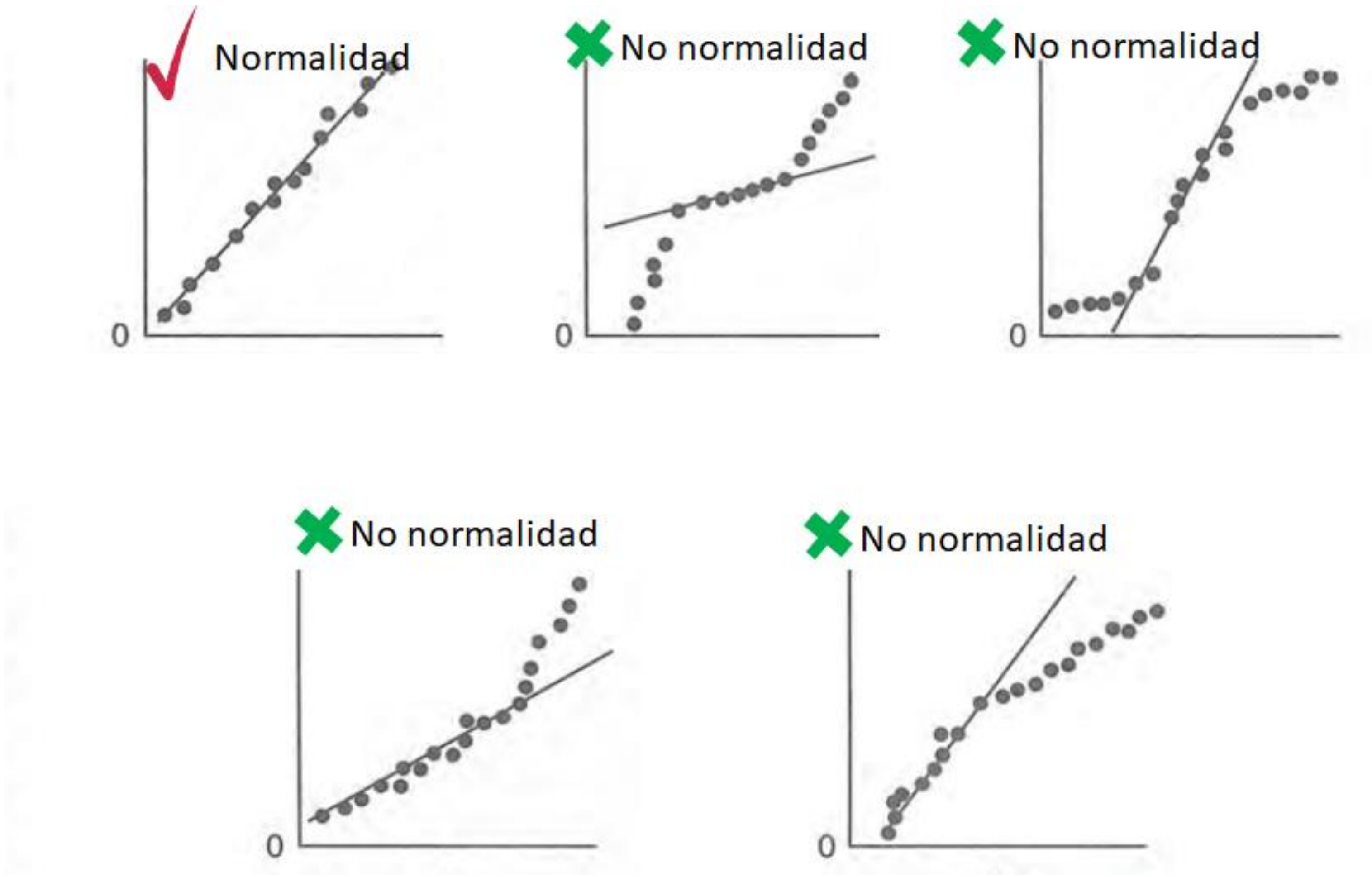
- Por lo general, una gran diferencia entre el R2 múltiple y el R2 ajustado indica que el modelo se ha ajustado en exceso (*overfitting*).

SUPUESTOS DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE

a) *Los residuales se distribuyen normalmente con media cero*
b) *La varianza de los residuos es constante (homocedasticidad)*
c) *Los residuos no están correlacionados*

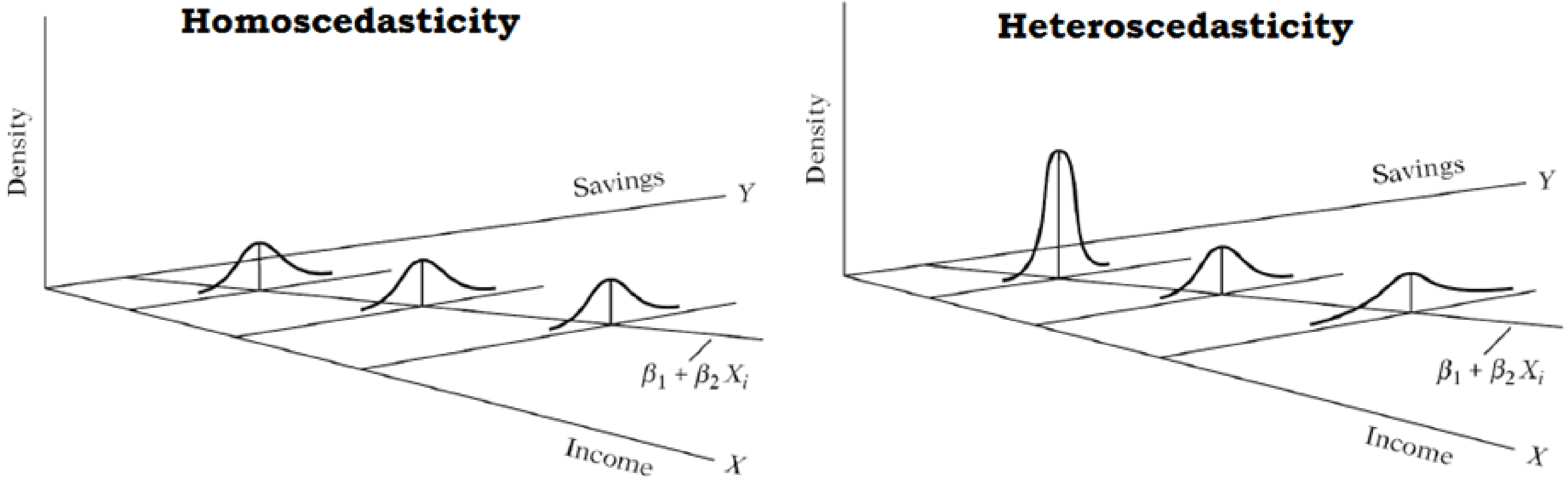
Prueba de normalidad de residuos

Q-Q Plot



Prueba de homocedasticidad de residuos

Prueba de Breusch–Pagan



H_0: los residuos tienen la varianza constante

H_1: los residuos NO tienen varianza constante

Prueba de autocorrelación de residuos

Prueba de Durbin-Watson

H_0: los residuos NO están autocorrelacionados

H_1: los residuos SI están autocorrelacionados

Tamaño de muestra

- Lo mejor es tener más de una medición por valor de la(s) variable(s) independiente(s) y garantizar la mayor cobertura posible de la misma en rango.

- **Radios empíricos:** 10 a 15 datos por variable independiente (Harrell); 5 mínimo, 20 recomendado (Hair); 50 a 1 si se hace stepwise.

Regresiones (Feature Selection)

Métodos de selección de variables en Regresión Múltiple

01 Modelo preespecificado

02 Pasos sucesivos (stepwise), adelante (forward) o atrás (backward)

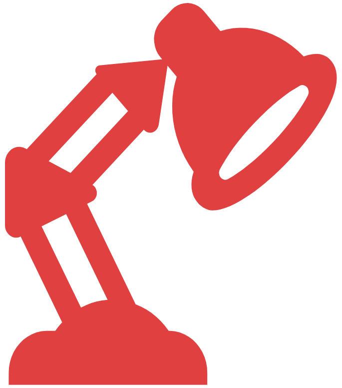
03 Mejores subgrupos (best subsets)

04 Regresión por componentes principales

05 Regresión de cresta (Ridge Regression)

06 Uso de Lasso

Entrenamiento y validación



En cualquiera de los casos anteriores, es buena práctica someter los modelos obtenidos a un proceso de entrenamiento y validación (partición, validación cruzada, etc.).

Suposiciones y diagnósticos de modelamiento (múltiples)

Otras variables: Dispersión de nueva variable vs. Modelo existente y Outliers

Relación funcional: Gráfico de predicciones estándar vs. Residuos estándar

Multicolinealidad: Correlaciones multivariadas y/o análisis factorial

Overfitting: Métricas penalizadas y tamaño de muestra

Calidad de la Regresión lineal (Métricas penalizadas de pérdida)

- **AIC:** Basado en teoría de la información. Penaliza por parámetros.
- **BIC:** Basado en teoría de la información. Penaliza por parámetros y ajusta por tamaño de muestra.
- **R2 ajustado:** Penaliza por el número de parámetros (a través de los grados de libertad).

AIC

- El **Criterio de Información de Akaike (AIC)** es un método matemático para evaluar hasta qué punto un modelo se ajusta a los datos a partir de los que se ha generado.
- El AIC se utiliza para comparar diferentes modelos posibles y determinar cuál es el que mejor se ajusta a los datos. El AIC se calcula a partir de:
  - el número de variables independientes utilizadas para construir el modelo.
  - la estimación de máxima verosimilitud del modelo (lo bien que el modelo reproduce los datos).

$$\mathrm{AIC} = 2\mathrm{K} - 2\ln(\mathrm{L})$$

- El modelo mejor ajustado según el AIC es el que explica la mayor cantidad de variación utilizando el menor número posible de variables independientes.
- Para comparar modelos utilizando el AIC, es necesario calcular el AIC de cada modelo. Si un modelo es más de 2 unidades AIC inferior a otro, entonces se considera significativamente mejor que ese modelo.

https://robjhyndman.com/hyndsight/lm_aic.html

Método Stepwise

Hace pruebas F de hipótesis parciales:

- ¿Dado el modelo actual de regresión, vale la pena incluir alguna variable? (Hacia adelante).
- ¿Dado el modelo actual de regresión, vale la pena excluir alguna variable? (Hacia atrás).

Best Subsets

- Detecta el mejor modelo de cada posible tamaño k entre p predictores (k<p).

- **Escogencia del k final:** un punto de corte en el error.

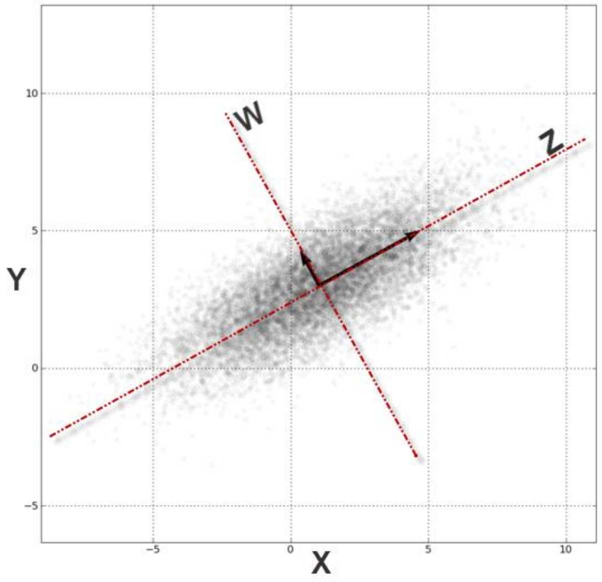
- Por origen usa una métrica llamada Cp de Mallows (de penalización) conectada con AIC.

- Da poder de decisión al modelador.

Regresión por Componentes Principales

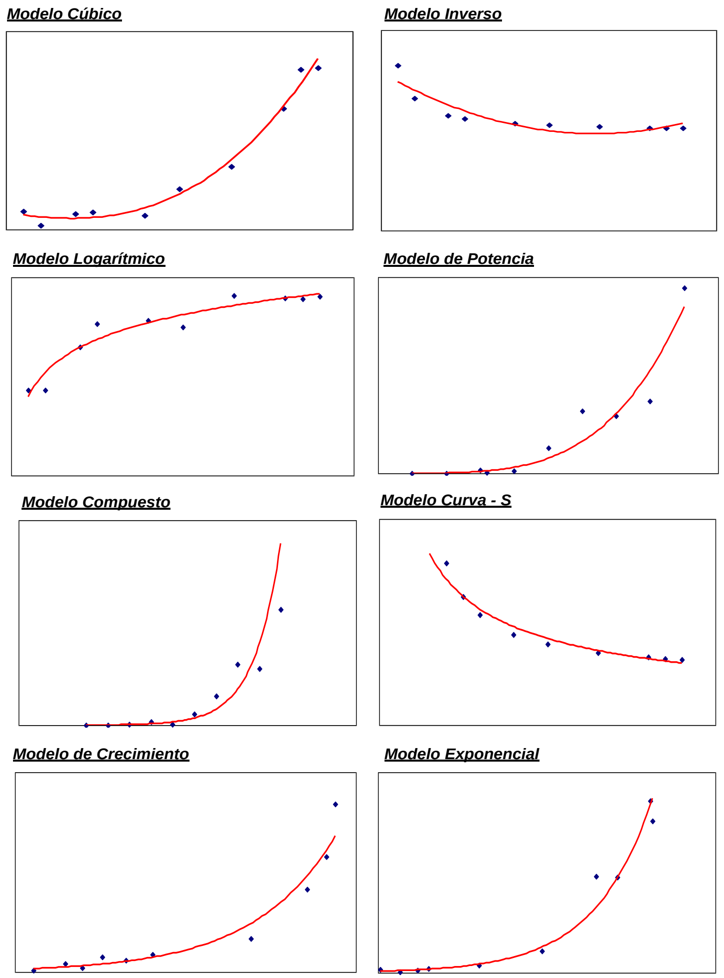
Realizar análisis factorial/componentes principales y usar dichos componentes como predictores.

Ayuda a reducir la complejidad del modelo y elimina la multicolinealidad.



Regresión No-Lineal

Regresión no-lineal polinomial



Ideas fundamentales

Ventajas

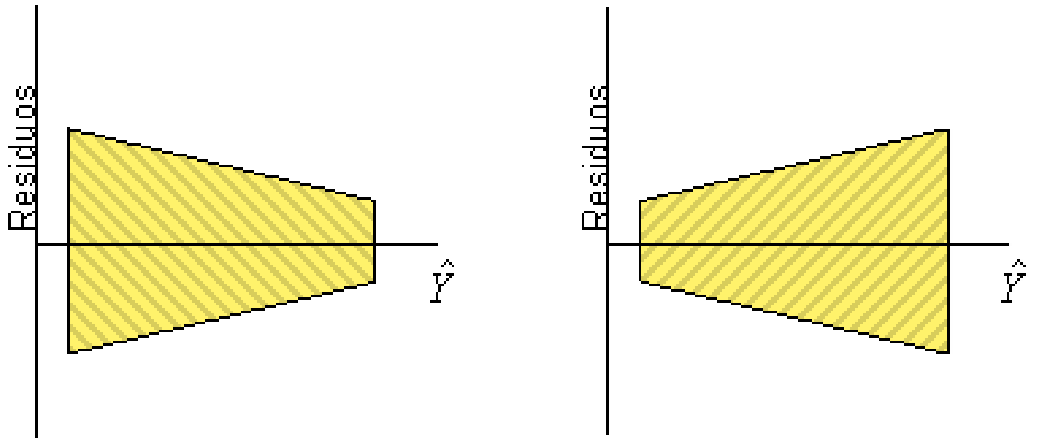
- Hacer transformaciones linealizables.
- Incluir grados polinomiales que puedan dar mejor ajuste.
- Coordinación con modelo preespecificado.

Desventajas

- Riesgo de overfitting.
- Basta con un polinomio grado 3 para la mayoría de los casos.
- Dificulta la comprensión/explicación.
- Riesgo de multicolinealidad

Transformación logarítmica

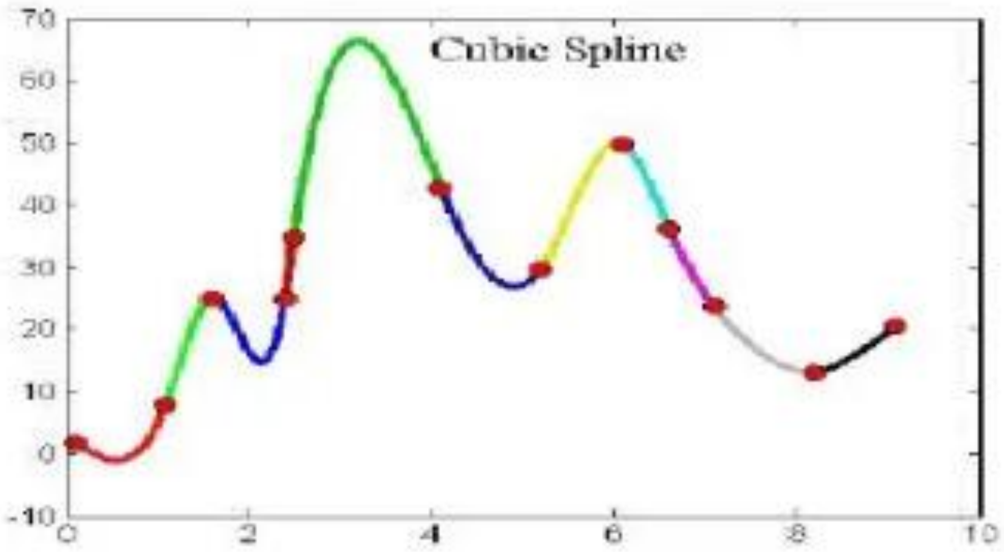
- El aumento de una variable X no siempre genera el mismo aumento en la Y.
- El clásico ejemplo es la satisfacción (o el rendimiento) en función de la cantidad de un satisfactor (insumo). Esto se conoce como la ley de los rendimientos marginales decrecientes y puede modelarse con un logaritmo.
- Puede detectarse con los residuos vs. predicciones.
- Es necesario preguntarse si tiene sentido en el negocio.



Regresión SPLINE

- Los splines son polinomios de orden K por segmentos. Los puntos de unión de los segmentos se llaman nudos.
- Se requiere que los valores de la función y las primeras K – 1 derivadas coincidan en los nudos, para que la spline sea una función continua con K – 1 derivadas continuas.

Ejemplo: **SPLINE CÚBICO**. Curva construida a partir de trozos de polinomios de grado 3 que se ensamblan perfectamente de forma que la curva que forman es continua hasta la segunda derivada. En este ejemplo se usan 10 nudos



Modelo Aditivo Generalizado (GAM)

Son una técnica versátil de modelización estadística utilizada para analizar relaciones complejas en los datos. A diferencia de los modelos lineales, los GAM pueden captar patrones no lineales combinando múltiples funciones suaves de variables predictoras. Los GAM son especialmente valiosos cuando se investigan dependencias complejas, lo que los convierte en una herramienta crucial para el análisis de datos y el modelado predictivo.

$$\mathrm{E}(Y|X_1, X_2, \ldots, X_p) = \alpha + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$$

$$Y = \alpha + \sum_{j=1}^{p} f_j(X_j) + \varepsilon$$

Para hallar las funciones se usa el *Backfitting Algorithm for Additive Models*. Hastie et al. Algorithm 9.1, Página 298