laSalle
UNIVERSIDAD RAMON LLULL

MD004

Regresiones

Máster Universitario en Data Science

David Larrosa Camps

Ricard Sierra Calls

Xavier Vilasís

Míriam Calvo

laSalle
UNIVERSIDAD RAMON LLULL

1. Introducción - ¿Qué es la regresión y por qué es importante?

Definición: La regresión es una técnica que permite modelar la relación entre una variable dependiente ***Y*** y una o más variables independientes ***X*** , con el objetivo de predecir ***Y*** o entender cómo cambia en función de ***X***.

Caso Práctico

Queremos predecir el precio de una vivienda en función de características como el tamaño, la ubicación, el número de habitaciones, etc.

Preguntas

¿Cómo se relaciona una variable con otra?

¿Podemos predecir ***Y*** basado en ***X***?

1. Introducción – Contexto probabilístico: varianza y covarianza

Varianza (**Var(X)**)

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)^2$$

- Una alta varianza indica que los valores de ***X*** están dispersos.
- Una baja varianza indica que están agrupados alrededor de la media.

Covarianza (**Cov(X,Y)**)

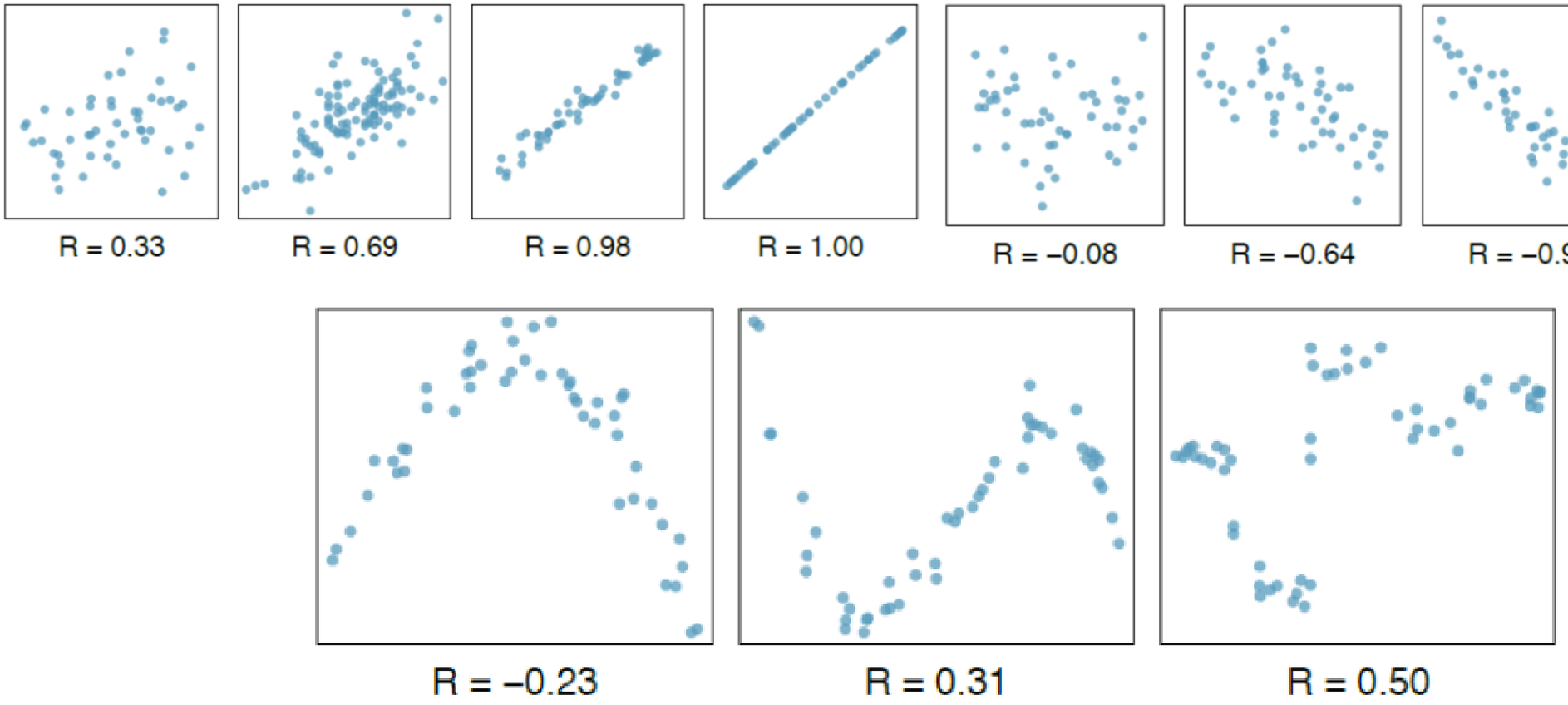
$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)(Y_i - \mu_Y)$$

- Si **Cov(X,Y)>0**: ***X*** e ***Y*** tienden a aumentar juntos.
- Si **Cov(X,Y)<0**: Cuando ***X*** aumenta, ***Y*** tiende a disminuir.
- Si **Cov(X,Y)=0**: No hay una relación lineal aparente.

Correlación (**ρ(X,Y)**)

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\cdot\mathrm{Var}(Y)}}$$

- **ρ=1**: Relación lineal perfecta positiva.
- **ρ=−1**: Relación lineal perfecta negativa.
- **ρ=0**: No hay relación lineal.



Existe relación, pero no es líneal

De la correlación a la regresión:

- La correlación mide la fuerza de la relación entre ***X*** e ***Y***, pero no describe la forma exacta de la relación ni permite predicciones.
- La regresión da un paso más al modelar ***Y*** en términos de ***X***, permitiendo hacer predicciones y estimar el impacto de ***X*** en ***Y***.

Modelo básico:

- Ecuación de la regresión lineal simple:

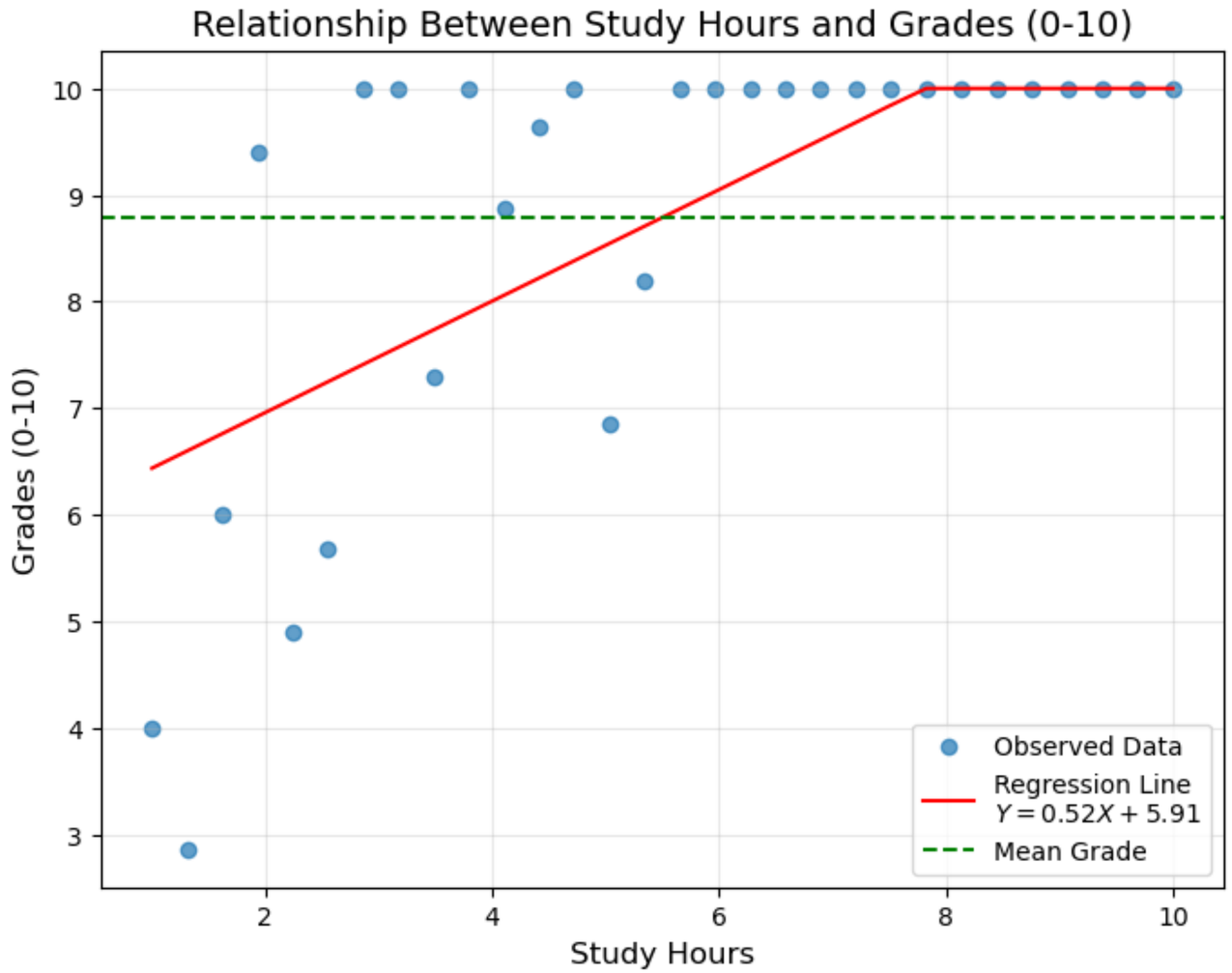
$$Y = \beta_0 + \beta_1 X + \epsilon$$

- β1 (pendiente) está relacionado con la covarianza entre ***X*** e ***Y***

1. Introducción – Ejemplo práctico y visualización

$$Y = 0'52 \cdot X + 5.91$$

	Study Hours (X)	Grades (Y)	Predicted Grades (Y_pred)
0	1.000000	3.990142	6.435777
1	1.310345	2.861069	6.598042
2	1.620690	5.994790	6.760306
3	1.931034	9.396676	6.922571
4	2.241379	4.900988	7.084835
5	2.551724	5.676899	7.247100
6	2.862069	10.000000	7.409364
7	3.172414	10.000000	7.571629
8	3.482759	7.298473	7.733893
9	3.793103	10.000000	7.896158
10	4.103448	8.868368	8.058422
11	4.413793	9.637293	8.220687
12	4.724138	10.000000	8.382951
13	5.034483	6.846366	8.545216
14	5.344828	8.187315	8.707480
15	5.655172	10.000000	8.869745
16	5.965517	10.000000	9.032009
17	6.275862	10.000000	9.194274
18	6.586207	10.000000	9.356538
19	6.896552	10.000000	9.518803
20	7.206897	10.000000	9.681067
21	7.517241	10.000000	9.843332
22	7.827586	10.000000	10.000000
23	8.137931	10.000000	10.000000
24	8.448276	10.000000	10.000000
25	8.758621	10.000000	10.000000
26	9.068966	10.000000	10.000000
27	9.379310	10.000000	10.000000
28	9.689655	10.000000	10.000000
29	10.000000	10.000000	10.000000



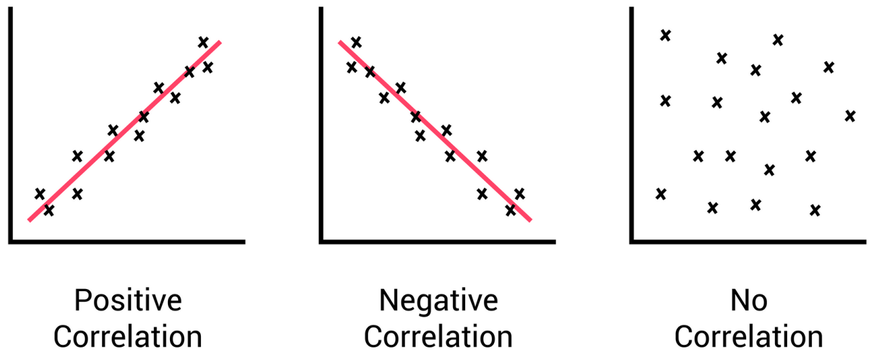
2. Regresiones lineales – Definición del Modelo

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- **Y**: Variable dependiente (respuesta o salida).
- **X**: Variable independiente (predictor o entrada).
- **β0**: Intercepto o valor de Y cuando X=0.
- **β1**: Pendiente, que mide el cambio esperado en Y por unidad de cambio en X.
- **ϵ**: Término de error (diferencia entre los valores observados y los predichos por el modelo).

Supuestos iniciales del modelo:

- Relación lineal entre X e Y.
- Los errores (ϵ) son independientes, tienen varianza constante (homocedasticidad) y siguen una distribución normal.



Intercepto (β0):

- Representa el valor esperado de Y cuando X=0.
  - Nota: En algunos casos X=0 puede no tener sentido práctico, pero el intercepto sigue siendo matemáticamente relevante.
- Ejemplo: Predicción del precio de una vivienda; β_0 sería el precio base sin superficie, lo cual puede no ser realista pero sigue siendo parte del modelo.

Pendiente (β1):

- Representa cuánto cambia Y por cada unidad adicional en X.
- Ejemplo: Si β1=2.5 en el caso de horas de estudio y calificaciones, entonces por cada hora extra de estudio se espera que las calificaciones aumenten en 2.5 puntos.

El método busca encontrar los valores de β0 y β1 que **minimizan la suma de los errores al cuadrado** (residuos)

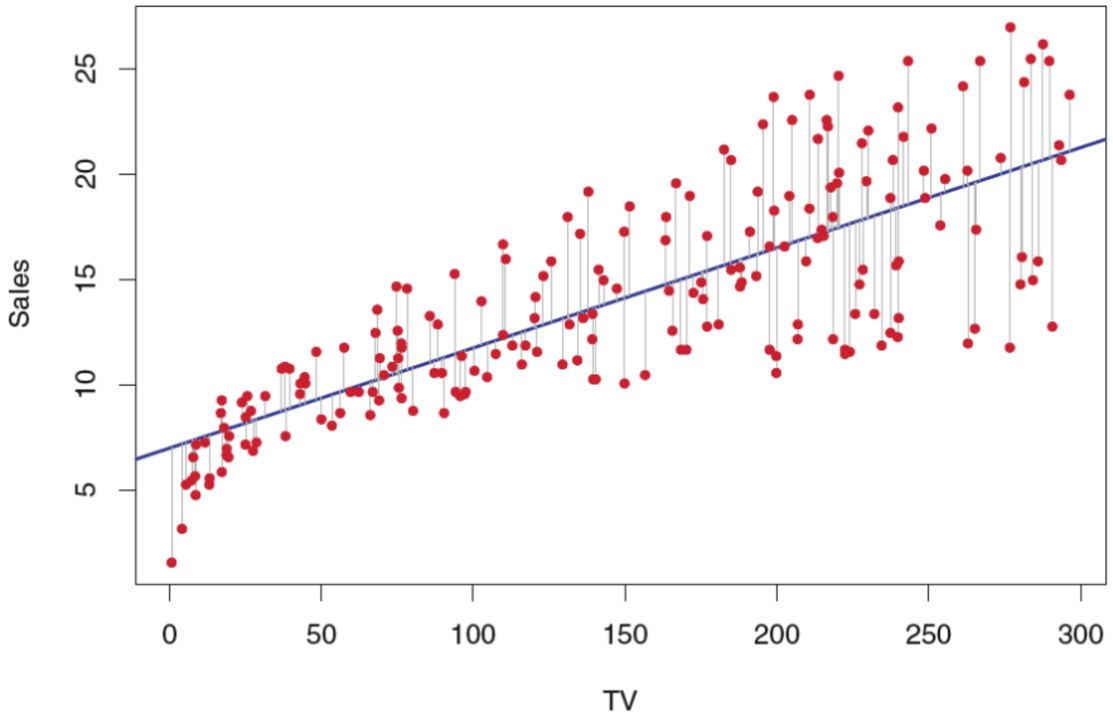
$$\text{Error total} = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

- Derivadas parciales:

$$\frac{\partial E}{\partial \beta_0} = 0, \quad \frac{\partial E}{\partial \beta_1} = 0$$

- Solución:

$$\beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}, \quad \beta_0 = \bar{Y} - \beta_1 \bar{X}$$



Definición de residuos (ϵi):

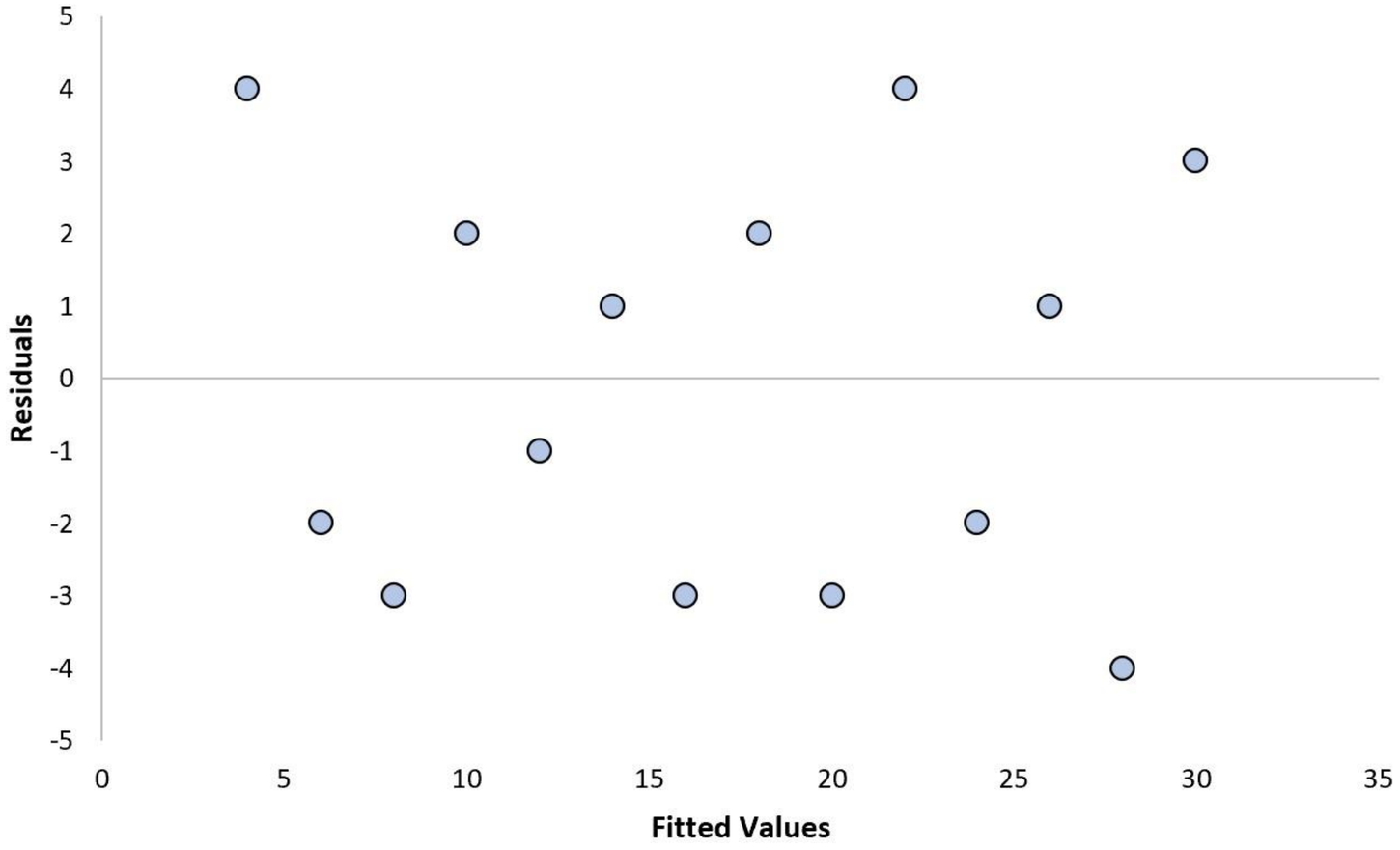
- Formula:

$$\epsilon_i = Y_i - \hat{Y}_i$$

  - Donde:
    - Y_i : Valor observado (real).
    - $\hat{Y}_i$: Valor predicho por el modelo.

Propiedades esperadas de los residuos:

- Media de los residuos debe ser aproximadamente 0 (E=0).
- No deben mostrar patrones sistemáticos (evaluado gráficamente).
- Varianza constante: Indica que el modelo es igualmente confiable en todo el rango de ***X***.



- Un gráfico de los residuos contra ***X*** o los valores predichos.
- Si el modelo es adecuado, los puntos deben estar distribuidos aleatoriamente alrededor de 0, sin patrones visibles.

- **Suma de los Cuadrados de los Residuos**
  - Mide la cantidad de variabilidad en ***Y*** no explicada por el modelo.

$$SS_{\text{res}} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

- **Suma Total de los Cuadrados**
  - Mide la variabilidad total en ***Y*** antes de aplicar el modelo.

$$SS_{\text{tot}} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- **Coeficiente de Determinación**
  - Indica la proporción de la variabilidad en ***Y*** explicada por ***X***
  - Valores cercanos a 1: Excelente ajuste.
  - Valores cercanos a 0: El modelo no explica la variabilidad de los datos.

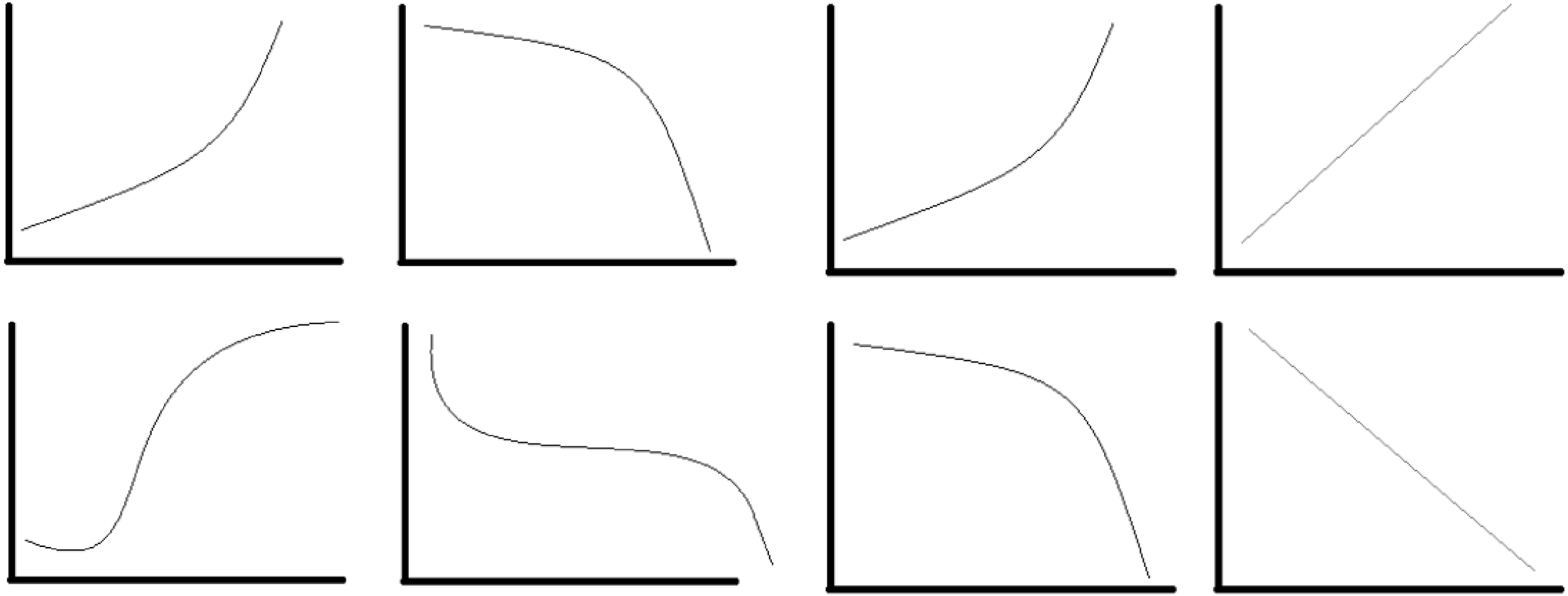
$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

- **Error Estándar de los Residuos**
  - Mide la dispersión de los puntos alrededor de la línea de regresión en las mismas unidades que ***Y***.
  - Valores más bajos indican un mejor ajuste.

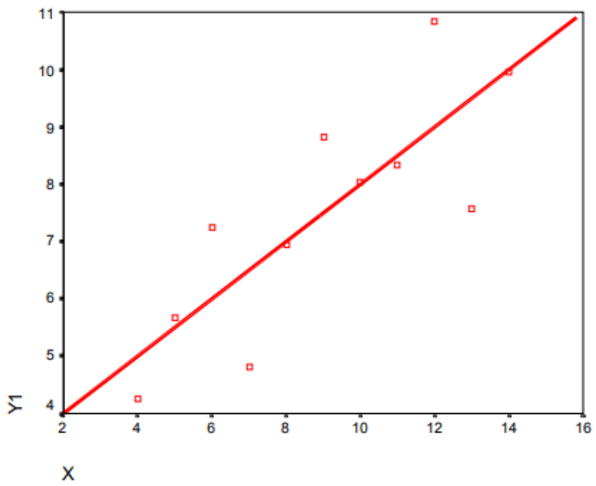
$$RSE = \sqrt{\frac{SS_{\text{res}}}{n-2}}$$

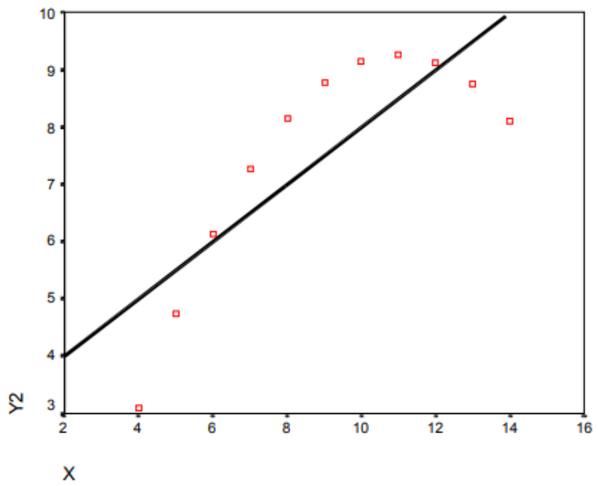
- **Estadístico *F***
  - Compara el modelo actual con uno sin predictores (modelo nulo) para evaluar si el modelo es estadísticamente significativo.

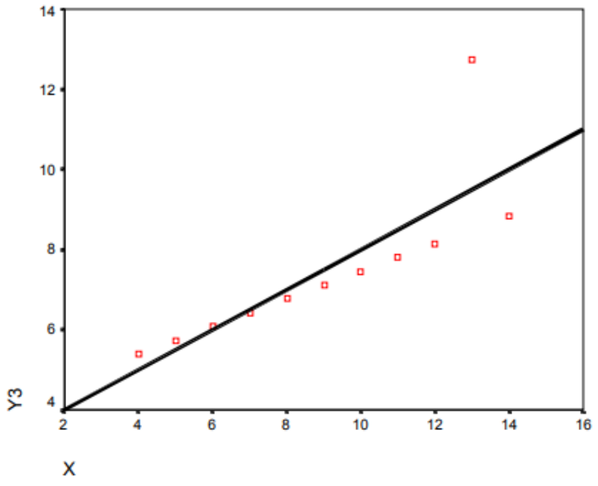
RELACIONES NO LINEALES

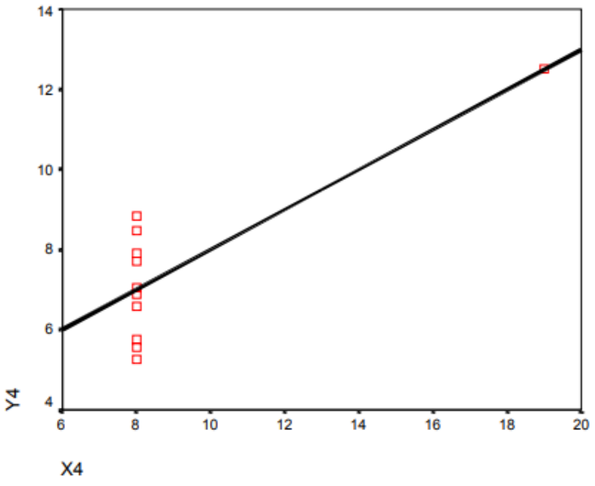


EFECTO PALANCA (OUTLIERS)



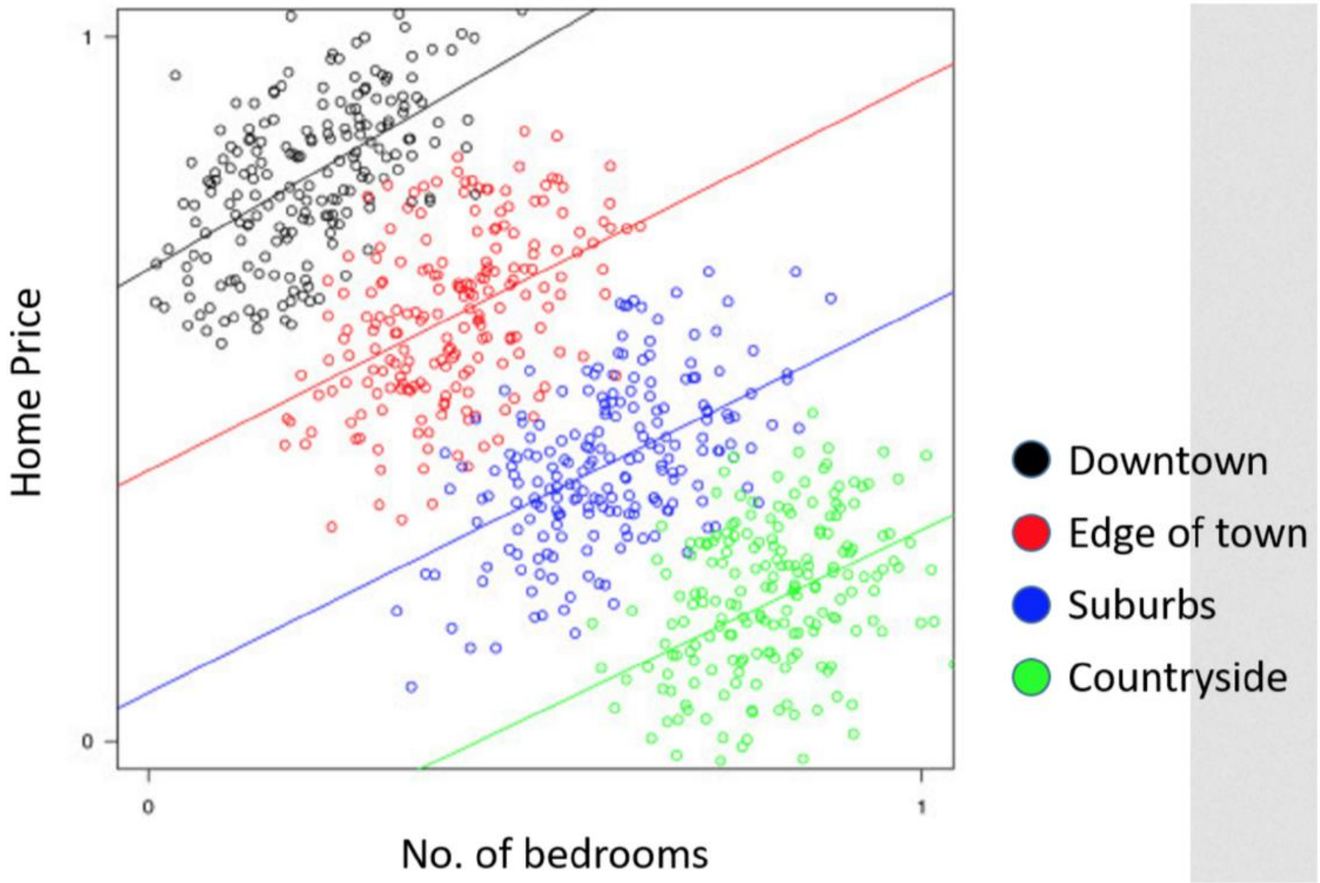






La influencia de un punto destacado puede afectar a todo el comportamiento general

PARADOJA DE SIMPSON



Fuente: Borgatti (2017)

Planteamiento del Ejemplo

- Contexto: Supongamos que estamos interesados en analizar la relación entre las horas de estudio (X) y las calificaciones de un examen (Y).
- Datos simulados:

Horas de Estudio (X)	Calificación Observada (Y)	Calificación Predicha ($\hat{Y}$)
1	3	2.9
2	4.5	4.3
3	5.5	5.7
4	7	7.1
5	8.5	8.5

  - Observamos una tendencia lineal aproximada en los datos: a más horas de estudio, mejores calificaciones.

Cálculo de la Regresión

- Ecuación del modelo: Utilizamos la fórmula del modelo de regresión lineal simple:

$$Y = \beta_0 + \beta_1 X$$

- Cálculo de los coeficientes:
  - Pendiente (β1):

$$\beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

  - (Esto mide cómo cambia ***Y*** en promedio por cada unidad de ***X***)

- Intercepto (β0):
  - Donde $\bar{X}$ y $\bar{Y}$ son las medias de ***X*** y ***Y***, respectivamente.
  - β1 ≈ 0.8.
  - β0 ≈ 1.5.

- Ecuación resultante:

$$Y = 1.5 + 0.8X$$

Interpretación del Modelo

- Pendiente ($\beta 1=0.8$):
  - Por cada hora adicional de estudio, la calificación promedio aumenta en 0.8 puntos.
  - Esto sugiere una relación positiva moderada entre horas de estudio y calificaciones.

- Intercepto ($\beta 0=1.5$):
  - Si un estudiante no estudia ($X=0$), se espera que obtenga una calificación promedio de 1.5.
  - Aunque $X=0$ puede no ser realista en este caso, el intercepto sigue siendo útil para definir la ecuación del modelo.

Evaluación del Modelo

- Residuos (ϵi): Calculamos los residuos
  - Donde $Y_i = 1.5 + 0.8 * X_i$ son las calificaciones predichas por el modelo.
  - Ejemplo:
    - Para X=2 : Y = 1.5 + 0.8 (2) = 3.1.
    - Residuo: ϵ=3−3.1=−0.1

- Medida de ajuste (R2):
  - En este caso, R2 ≈ 0.85, lo que significa que el 85% de la variabilidad de las calificaciones es explicada por las horas de estudio.

Visualización

- Gráfico de Dispersión:
  - Muestra los puntos (X,Y) reales.
  - Añade la línea de regresión Y=1.5+0.8X en el gráfico.
  - Indica los residuos como líneas verticales entre los puntos observados y la línea ajustada.

- Residual Plot:
  - Grafica los residuos (ϵi) contra los valores predichos (Y_i).
  - Si los residuos están distribuidos aleatoriamente alrededor de 0, el modelo es adecuado.

4. Tabla de diagnósticos – Supuestos clave

Nombre del Supuesto	Descripción	Evaluación	Solución si no se cumple
Relación Lineal entre X y Y	Se asume que existe una relación lineal entre la variable independiente (X) y la dependiente (Y).	• **Gráfico de dispersión** (X vs. Y): Los puntos deben mostrar una tendencia aproximadamente lineal. • **Residual plot** (ϵi vs. X): Los residuos deben estar distribuidos aleatoriamente alrededor de 0, sin patrones no lineales.	• Transformar las variables (log(X), \sqrt{X}, etc.). • Usar un modelo no lineal o modelos polinomiales.
Independencia de los Residuos	• Los residuos (ϵi) deben ser independientes entre sí. • Este supuesto es especialmente relevante en datos de series temporales, donde las observaciones están correlacionadas en el tiempo.	• Test de **Durbin-Watson** (`lmtest::dwtest(model)`) • **Residual plot**: No debe haber patrones sistemáticos en los residuos.	• Considerar modelos de series temporales como ARIMA. • Incluir variables adicionales en el modelo para capturar la dependencia.
Normalidad de los Residuos	• Se asume que los residuos siguen una distribución normal (ϵi∼N(0,σ2)). • Este supuesto es importante para intervalos de confianza y pruebas de hipótesis.	• **Histograma de residuos**: Debe aproximarse a una distribución normal. • **Gráfico Q-Q** (quantile-quantile): Los puntos deben alinearse con la diagonal. • **Test de Shapiro-Wilk o Kolmogorov-Smirnov** (`shapiro.test(residuals(model))`)	• Transformar Y o los residuos (log(Y), \sqrt{Y}). • Usar un modelo robusto o técnicas no paramétricas.

4. Tabla de diagnósticos – Supuestos clave

Nombre del Supuesto	Descripción	Evaluación	Solución si no se cumple
Homocedasticidad	• La varianza de los residuos debe ser constante a lo largo del rango de X • Si la varianza cambia, se tiene heteroscedasticidad, lo que afecta la precisión de los coeficientes y las predicciones.	• **Residual plot**: Los residuos no deben formar un patrón de abanico. • **Test de Breusch-Pagan o White** (`lmtest::bptest(model)`)	• Transformar Y (log(Y), \sqrt{Y}). • Usar estimadores ponderados (Weighted Least Squares).
No Multicolinealidad (para Regresión Múltiple)	• Las variables independientes no deben estar altamente correlacionadas entre sí. • La multicolinealidad puede inflar los errores estándar y dificultar la interpretación de los coeficientes.	• **Matriz de correlación** entre variables independientes. • **Variance Inflation Factor** (VIF) (`car::vif(model)`)	• Eliminar variables redundantes. • Combinar variables correlacionadas en un único indicador.

4. Tabla de diagnósticos – Diagnóstico del modelo

Una vez identificados los posibles problemas, es importante realizar un diagnóstico detallado para confirmar si el modelo cumple con los supuestos.

Herramientas Gráficas

- Gráfico de Residuos: Puntos distribuidos aleatoriamente alrededor de 0 (Patrón esperado).
- Histograma y Q-Q Plot: Evalúan la normalidad de los residuos.
- Gráfico de Residuos Estandarizados: Residuos divididos por su desviación estándar, útiles para detectar outliers ($|\epsilon i|>2$).

Pruebas Estadísticas

- Durbin-Watson: Para detectar autocorrelación.
- Breusch-Pagan: Para identificar heteroscedasticidad.
- Shapiro-Wilk: Para evaluar la normalidad.

Intervalo de Confianza

Definición:

- Un intervalo de confianza (IC) es un rango de valores que tiene una probabilidad específica (1−α, típicamente 95%) de contener el valor verdadero de un parámetro.
- En regresión lineal, se usa para los coeficientes β0 y β1.

Fórmulas:

- Para el coeficiente β1 (el mismo para β0) : $$IC : \hat{\beta}_1 \pm t_{\alpha/2,df} \cdot SE(\hat{\beta}_1)$$

- Donde:
  - $\hat{\beta}_1$: Coeficiente estimado de la pendiente.
  - $t_{\alpha/2,df}$ Valor crítico de la distribución t con df=n−2.
  - $SE(\hat{\beta}_1)$: Error estándar del coeficiente.

Intervalo de Confianza

Interpretación:

- Un IC del 95% significa que, si se repitieran los experimentos muchas veces, el 95% de los intervalos calculados contendrían el valor verdadero del parámetro.
- Si un intervalo no incluye 0, el parámetro es significativo al nivel α.

Ejemplo:

- Para un modelo donde $\hat{\beta}_1 = 1.5$ y $SE(\hat{\beta}_1) = 0.2$, con $t_{0.025,28} \approx 2.048$

$$IC : 1.5 \pm 2.048 \cdot 0.2 = (1.09, 1.91)$$

- Interpretación: Con 95% de confianza, el verdadero efecto de ***X*** sobre ***Y*** está entre 1.09 y 1.91.

Intervalo de Predicción

Definición:

- Un intervalo de predicción (IP) es un rango que contiene el valor de ***Y*** para una nueva observación ***X*** con una probabilidad específica (1−α, típicamente 95%).
- Refleja la incertidumbre no solo en el modelo, sino también en el término de error.

Fórmulas:

- Para un nuevo valor X0 :

$$IP : \hat{Y}_0 \pm t_{\alpha/2,df} \cdot \sqrt{SE(\hat{Y}_0)^2 + \sigma^2}$$

- Donde:
  - $\hat{Y}_0$: Valor predicho por el modelo para X0.
  - $SE(\hat{Y}_0)$: Error estándar de la predicción.
  - $\hat{\sigma^2}$: Varianza de los residuos.

Intervalo de Predicción

Diferencias con el IC:

- El intervalo de confianza aplica al valor promedio esperado de ***Y***, mientras que el intervalo de predicción aplica a valores individuales de ***Y***.
- Los IP son más amplios porque incluyen la incertidumbre de los residuos.

Ejemplo:

- Para un modelo donde $\hat{Y}_0 = 7$, $SE(\hat{Y}_0) = 0.5$, y $\sigma = 2$, con $t_{0.025,28} \approx 2.048$

$$IP : 7 \pm 2.048 \cdot \sqrt{0.5^2 + 2^2} = 7 \pm 2.048 \cdot 2.06 \approx (2.8, 11.2)$$

- Interpretación: Interpretación: Para un nuevo valor X0, esperamos que ***Y*** esté entre 2.8 y 11.2 con un 95% de confianza.

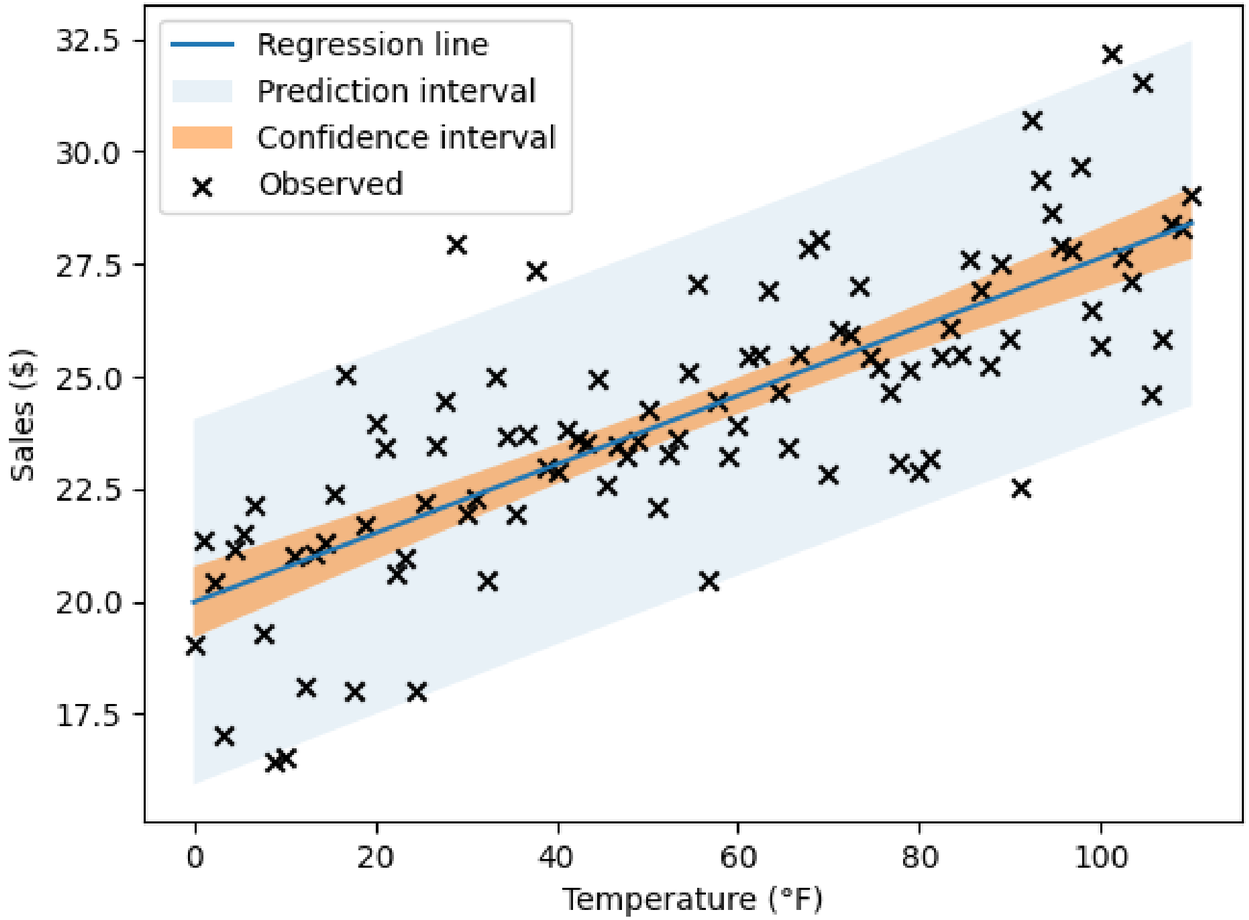
Gráfico de intervalos:

- Crea un gráfico que muestre:
  - La línea de regresión ajustada (***Y***).
  - Bandas para los intervalos de confianza del promedio (***Y***).
  - Bandas más amplias para los intervalos de predicción.
- **Herramientas sugeridas:**
  - En R: Usa ggplot2 para añadir bandas con geom_ribbon.
  - En Python: Usa matplotlib o seaborn con la función fill_between().

Interpretación gráfica:

- Muestra cómo la incertidumbre aumenta en los extremos del rango de ***X*** debido a la mayor influencia de los residuos.

5. Intervalos de confianza y predicción – Visualización y comparación



Definición:

- Es un modelo utilizado cuando la variable dependiente es categórica (generalmente binaria).
- En lugar de predecir un valor continuo (***Y***), la regresión logística estima la probabilidad de que Y pertenezca a una categoría determinada.

Ejemplo :

- Problema: ¿Cuál es la probabilidad de que un cliente compre un producto en función de su salario y edad?
- Variable dependiente (***Y***): Compra = 1, No compra = 0.
- Variable independiente (***X***): Salario del cliente.

Función logística:

- El modelo utiliza la función logística o sigmoide para transformar una combinación lineal de los predictores (***X***) en una probabilidad (p):

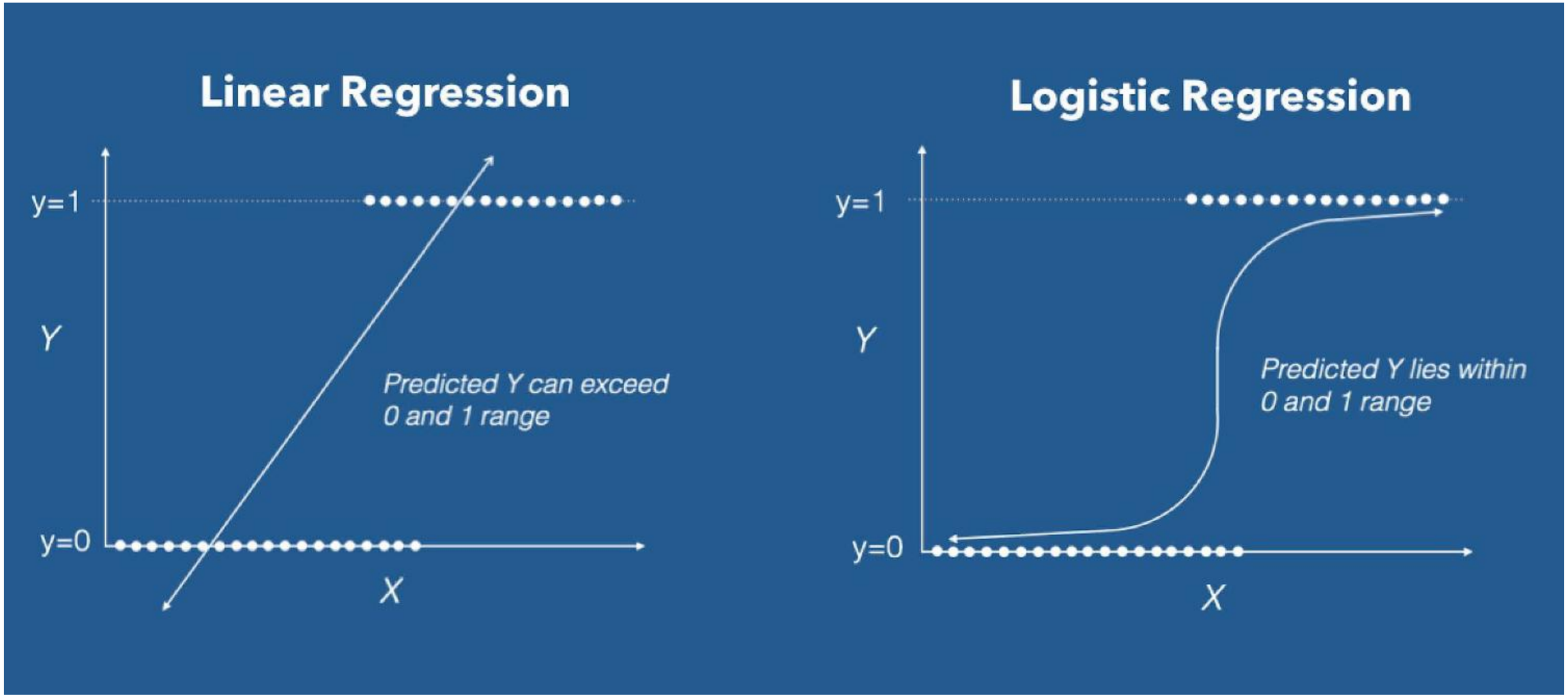
$$\hat{p} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

  - Donde:
    - $\hat{p}$: Probabilidad predicha de que ***Y***=1.
    - β_0: Intercepto.
    - β_1: Coeficiente de ***X***.

- La ecuación también se puede expresar en términos del logit (logaritmo de las probabilidades):

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \beta_0 + \beta_1 X$$

6. Regresión Logística – Ecuación del modelo



6. Regresión Logística – Interpretación de los coeficientes

¿Cómo interpretamos el coeficiente?:

- Indica el cambio en el logit (log-odds) por cada unidad de cambio en ***X***.
- En términos de odds ratio (razón de probabilidades):
  - Si $e^{\beta_1} > 1$, ***X*** aumenta la probabilidad de ***Y***=1.
  - Si $e^{\beta_1} < 1$, ***X*** reduce la probabilidad de ***Y***=1.

Ejemplo :

- Si β1=0.5, entonces e^0.5≈1.65.
  - Por cada unidad adicional en ***X***, las probabilidades de Y=1 aumentan en un 65%.

Método de estimación:

- La regresión logística no usa mínimos cuadrados; en su lugar, emplea el método de máxima verosimilitud (maximum likelihood) para ajustar los coeficientes.
- Se encuentra el conjunto de β0, β1 que maximiza la probabilidad de observar los datos dados.

Herramientas para ajustar el modelo:

- En R
  - Regresión Lineal: *model_lin <- lm(Y ~ X, data = datos)*
  - Regresión Logistica: *model_log <- glm(Y ~ X, family = binomial, data = datos)*

MATRIZ DE CONFUSIÓN

	1	0
1	TP	FP
0	FN	TN

- **TP (True Positive)** – Son los valores que el algoritmo clasifica como positivos y que realmente son positivos.
- **TN (True Negative)** – Son valores que el algoritmo clasifica como negativos (0 en este caso) y que realmente son negativos.
- **FP (False Positive)** – Falsos positivos, es decir, valores que el algoritmo clasifica como positivo cuando realmente son negativos.
- **FN (False Negative)** – Falsos negativos, es decir, valores que el algoritmo clasifica como negativo cuando realmente son positivos.

- **Accuracy:** La métrica accuracy representa el porcentaje total de valores correctamente clasificados, tanto positivos como negativos.
  - *Accuracy = (TP + TN) / (TP + TN + FP + FN)*
- **Sensibility (o Precision):** Es utilizada para poder saber qué porcentaje de valores que se han clasificado como positivos son realmente positivos.
  - *Sensibility= TP / (TP + FP)*
- **Specificity (o Recall):** Es utilizada para saber cuantos valores positivos son correctamente clasificados.
  - *Specificity = TP / (TP + FN)*
- **F1 Score:** Métrica muy utilizada en problemas en los que el conjunto de datos a analizar está desbalanceado.
  - *F1 = 2 * ((Specificity * Sensibility)/(Specificity + Sensibility))*

6. Regresión Logística – Ajuste del modelo

		Predicted Class: Positive	**Predicted Class: Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) Type II Error	**Sensitivity** $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	**Specificity** $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	**Negative Predictive Value** $\frac{TN}{(TN + FN)}$	**Accuracy** $\frac{TP + TN}{(TP + TN + FP + FN)}$

- **AUC**: El valor de esta métrica se encuentra en un rango entre 0 y 1, donde 0 es como si tuviéramos un modelo aleatorio y 1 un modelo predictivo perfecto (nunca sucede).