DAT 246 - Modelación Estadística

Primer parcial con dos variables, buscar un ejemplo real para resolver un problema, grupos de a dos.

Hacer gráficas (con tres dimensiones) del modelo de regresión lineal. Dibujos a mano, no en computadora.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Universidad Mayor de San Andrés, facultad de ciencias puras y naturales.

Frecuencia: Cuatro horas por semana por cada materia.

Cuarto semestre, 2/2025.

Plan curricular por competencia.

Competencia de la asignatura.

Capacidad para integrar conocimientos adecuados en el modelo matemático, cálculo y diseño experimental, aplicando los modelos, métodos y técnicas relevantes en distintas áreas de la estadística matemática, participando en la creación de nuevas tecnologías que contribuyan al desarrollo de la sociedad de la información.

Contenido

Índice

1. Introducción a la modelación.

1.1. Concepto de modelo.

Es la representación de un sistema real de acuerdo al objetivo del problema.

Un modelo es una forma de ver, una abstracción de algo real.

Un modelo es la representación de la realidad, por lo tanto no incluye todos los aspectos del problema exacto.

Es una representación del sistema estudiado, para predecir y observar los posibles cambios.

Un modelo científico es una representación de algún objeto sujeto a investigación que pueden ser procesos, acontecimientos, sistemas, etc. Y que se utilizan con fines de predicción y control.

Un modelo científico tiene caracter explicativo mas que descriptivo.

Las ventajas de tener un modelo son:

- Cuando el sistema real no se puede manipular. Ejemplos: astronomia, modelo relacional (base de datos), planos de arquitectura
- Cuando el costo es muy alto. Ejemplos: Ventas, simuladores de conducción, mostrar al usuario sistemas informáticos (modelo entidad relación)

1.2. Clasificación de modelo.

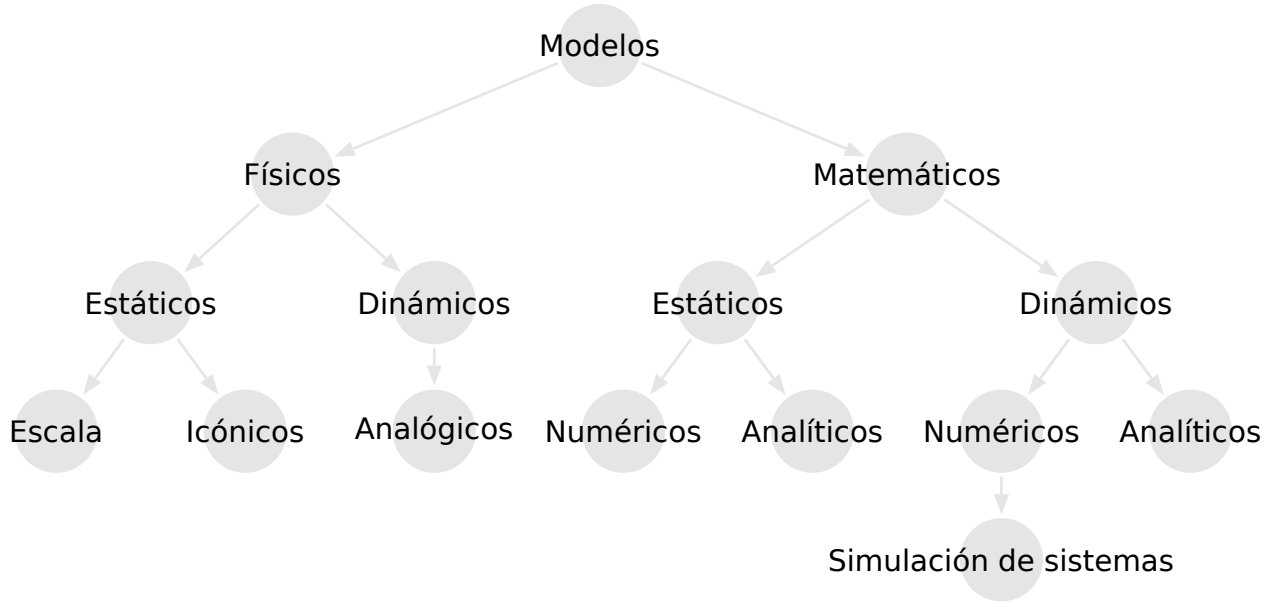
Clasificación segun el grado de abstracción

- Modelos icónicos: Representan de forma pictórica o visual ciertos aspectos, además se parece a lo que representa. Ejemplos: Fotografías, juguetes, esculturas, dibujos.

- Modelos analógicos: Utilizan una serie de propiedades para representar otro conjunto de propiedades que posee el sistema estudiado. Ejemplos: Estructuras geólogicas (se representa por colores de acuerdo a distintas propiedades), diagramas de flujo con programas.
- Modelos simbólicos: Se emplean para asignar las propiedades del sistema por medio de una ecuación matemática.

Clasificación de los modelos por el método de solución

Hay dos tipos de modelos, los físicos y los matemáticos, el físico tiene la parte estática y dinámica. El matemática también tiene la parte estático y dinámico. El estático de física se divide en la escala y los icónicos, la parte dinámica solo tiene los analógicos. El estático de matemática se divide en numérico y analitico, mientras que el dinámico tiene número y analítico, numérico tiene simulación de sistemas.



Simulación de sistemas

Es una técnica que utilizan los modelos matemáticos dinámicos.

Los modelos físicos

Los atributos de las entidades del sistema se representan mediante medidas físicas, y las actividades del sistema representan las leyes físicas.

Dentro de los modélos físicos estáticos, tenemos los modelos a escala, por ejemplo, un mapa, accidentes de tránsito, estudio de naves acuáticas.

Los modelos icónicos representan de manera visual/pictórica.

Los modelos matemáticos

Las entidades y los atributos de un sistema se representan mediante variables matemáticas. las actividades mediante actividades matemáticas que relacionan las variables. Dentro de ellas tenemos un modelo mateático estático.

El modelo matemático estático (no cambia en el tiempo): Despliega las relaciones entre los atributos del sistema cuando este está equilibrado, se puede resolver numéricamente y analíticamente. Es analítico si tiene solución, sino es numérico.

El modelo matemático dinámico significa que cambia a través del tiempo, permite deducir los cambios a los atributos del sistema en función del tiempo. De acuerdo a la complejidad, se puede resolver de manera analítica o numérica.

Proceso de modelación

Es las etapas del proceso de contrucción de un modelo.

Sistema real supuesto.
- Abstracción.
- Validación.

Modelo o modelo conceptual.
- Transformación.
- Validación.

modelo para ordenador.
- Implementación.
- Validación.

Cálculo.
- Experimento.
- Validación.

Sistema real supuesto.

Abstracción: Consisten en el aislamiento mental que permite separar las características escenciales y generales de otras propiedades secundarias.

Sea x el número de individuos de una especie e,suponemos que no existe otra especie que se alimente de los productos que e consume, también e no sirve de alimento a otra especie. Entonces la taza de crecimiento en función al tiempo $\frac{dx}{dt}$ donde x es proporcional a una variable a, entonces $\frac{dx}{dt} = ax$

1.3. Repaso de la recta.

La recta se define mediante ecuaciones:

$$ax + dy + c = 0$$
$$y - y_1 = m(x - x_1)$$
$$y = mx + b$$
$$\frac{y - y_1}{x - x_1} = \frac{y_2 - y_1}{x_2 - x_1}$$
$$\frac{x}{a} + \frac{y}{b} = 1$$

Ejemplo: Sean los puntos $p_1 = (2, 3)$ y $p_2 = (5, 7)$.

$$\frac{y - 3}{x - 2} = \frac{7 - 3}{5 - 2}$$
$$y = \frac{4}{3}(x - 2) + 3$$
$$y = \frac{4x}{3} + \frac{1}{3}$$
$$m = \tan\theta$$
$$\tan\theta = \frac{4}{3}$$
$$\arctan\left(\frac{4}{3}\right) = 0.927$$
$$\theta = 0.927 \cdot \frac{180}{\pi} = 53.18°.$$

Práctica nr4: graficar la recta de arriba.

2. Regresión lineal y estimación de parámetros.

El modelo estadístico “regresión lineal” se puede escribir de dos formas:

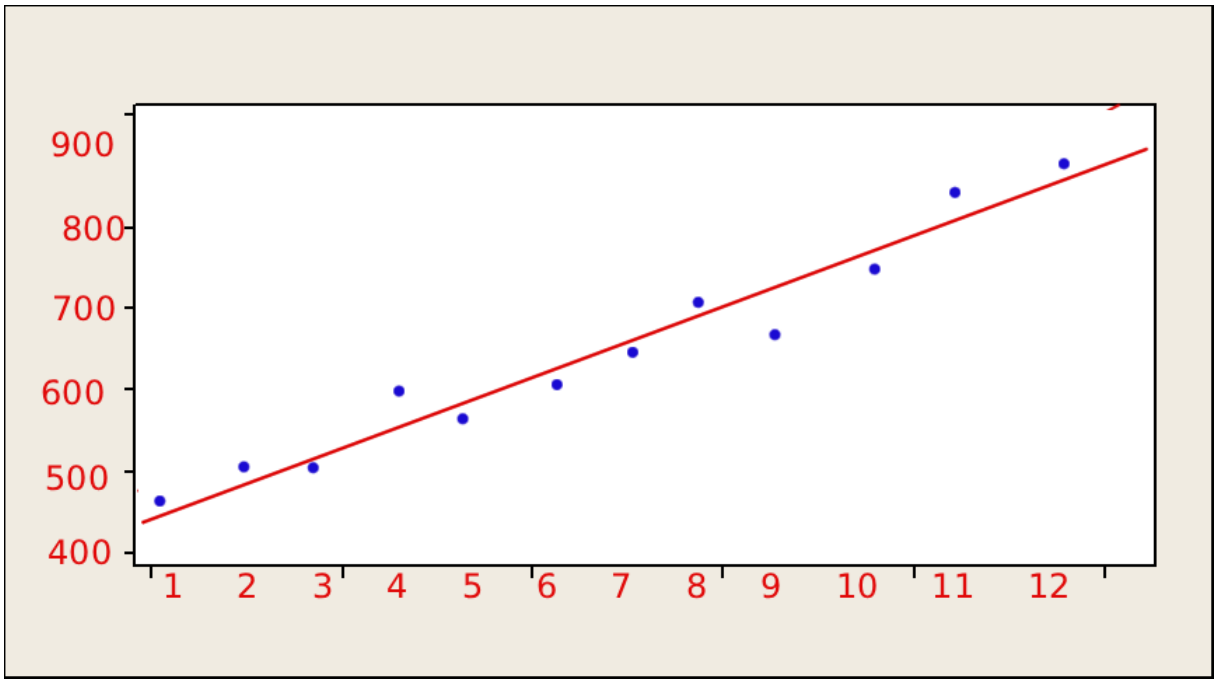
a) La variable respuesta o dependiente y se expresa como suma de $\beta_0 + \delta_1 x_i$ y un error aleatorio ε_i que tiene una distribución normal con una media cero y una varianza σ^2. Por lo tanto, el modelo se escribe $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i : \varepsilon \sim N\left(0, \sigma^2\right)$.

b) La variable respuesta o dependiente y tiene una distribución normal con media que cambia en función de la variable $|(x)$ pero con varianza constante. Entonces, el modelo se expresa como:

$$y_i \sim N\left(\mu_i, \sigma^2\right)$$
$$\mu_i = \beta_0 + \beta_i x_i$$
$$\sigma^2 = \text{Constante}$$

En cualquiera de las dos formas, el vector de parámetros del modelo es: $\theta = (\beta_0, \beta_1, \sigma)^T$.

Ejemplo:



$$\overline{x}_i = a + bt$$

$$\beta_0 + \beta_i T$$

$$b = 27,8$$

$$a = 452,6$$

$$\overline{x}_1 = 452,6 + 27,8T$$

$$\overline{x}_2 = 452,6 + 27,8(8)$$

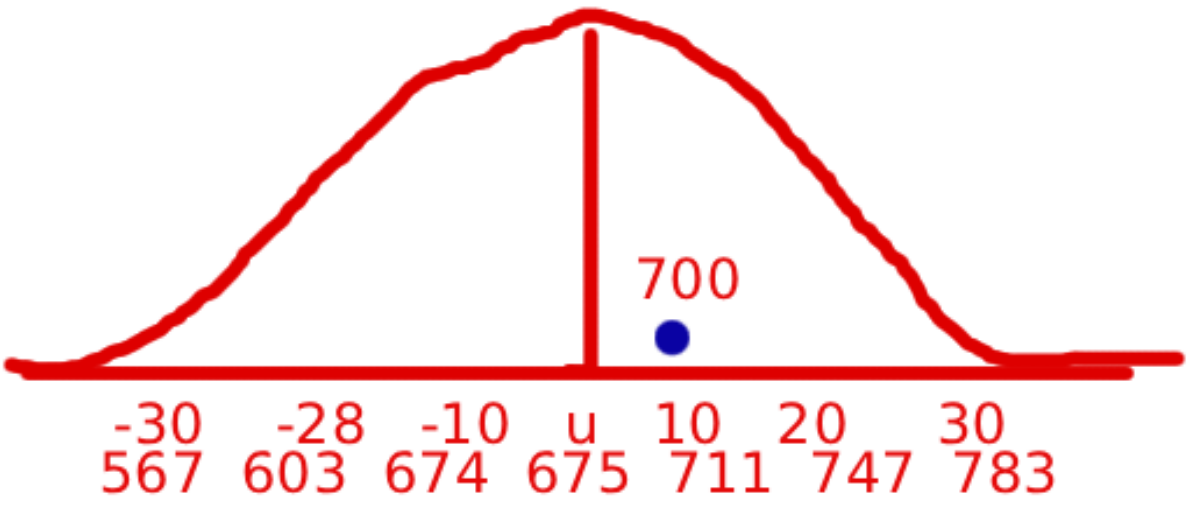
$$\overline{x}_3 = 675$$

De acuerdo con nuestras suposiciones sobre la curva normal de x_8, tiene el valor de 700, es un valor proveniente de una distribución normal, con media de 675.

$$x_8 \sim N(675, 36)$$

$$\sigma^2 = 36$$

$$\sigma = 6$$



Ejemplo:

Estud.	Estatura	Peso
	x	y
1	154	53
2	158	45
3	162	56
4	166	73
5	170	65
6	174	88
7	178	89
8	182	75
9	190	90

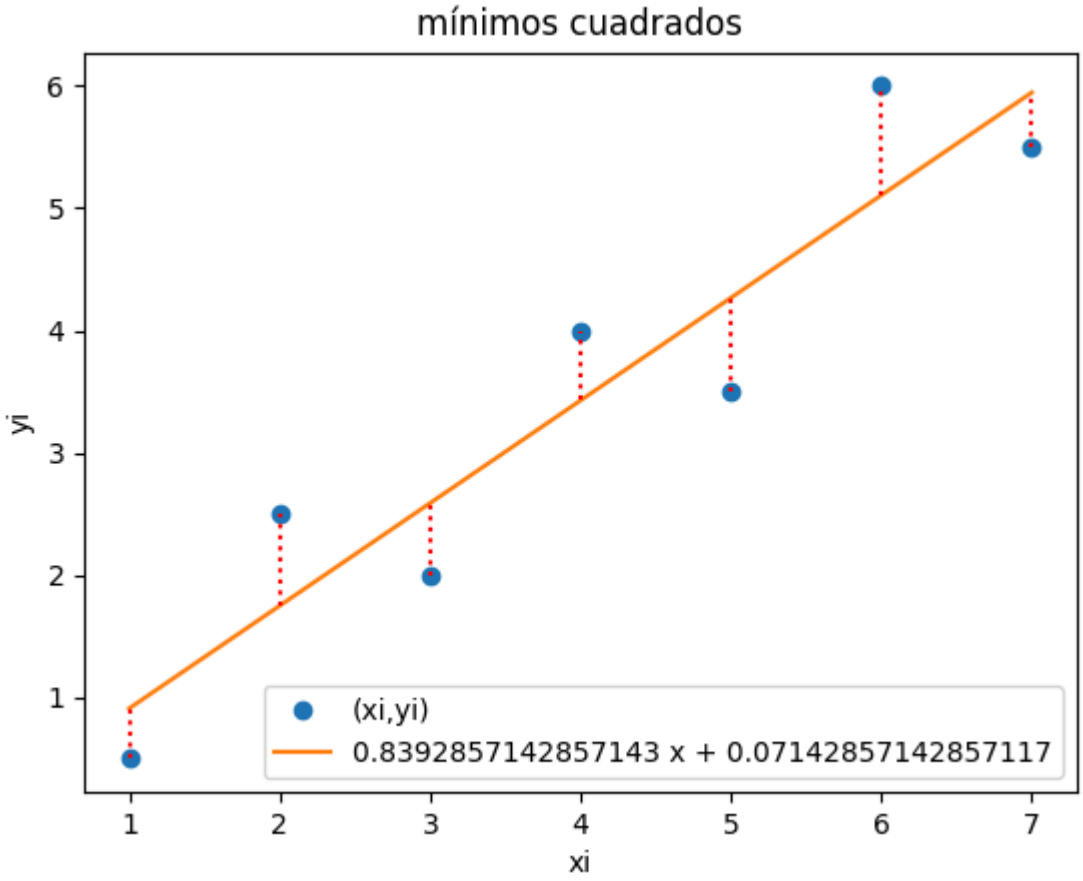
$$\text{Peso} \sim N\left(\mu_i, \sigma^2\right)$$
$$\mu_i = \beta_0 + \beta_1 \text{estatura}_i$$
$$\sigma^2 = \text{constante}$$

$$\widehat{\text{Peso}}_i \sim N\left(\hat{\mu}_i, \hat{\sigma}^2\right)$$
$$\hat{\mu}_i = -138.0091 + 1.2227 \text{estatura}_i$$

2.1. La recta de regresión.

Una recta de regresión es una línea recta que se ajusta a un conjunto de datos para mostrar la relación entre dos variables. En estadística, se utiliza conmunmente para predecir el valor de una variable en función de otra variable.

Sea una distribución bidimensional (x, y) y una serie de valores observados (x_i, y_i) con una correlación lineal que se representa en plano mediante una nube de puntos; se busca la recta que mejor se ajuste a la nube de puntos y esta recta recibe el nombre de recta de regresión o recta de regresión lineal. Para ello se utiliza el método de mínimos cuadrados que consiste en que la recta cumpla «la suma de los cuadrados de las distancias de todos los puntos a la recta sea mínimos».



La recta de regresión de y sobre x está representado por la ecuación:

$$y = \overline{y} + \frac{S_{xy}}{S_x^2}(x - \overline{x})$$

Esta es la ecuación de la recta en forma punto-pendiente. La ecuación representa un modelo para encontrar la recta de regresión.

$$m = mxy = \frac{S_{xy}}{S_x^2}$$

$S(xy)$ es la covarianza, S es la muestra, S_x^2 es la varianza y σ es la población, S_x es la desviación estándar.

Encontrar el coeficiente re correlación r para saber si es una regresión lineal.

$$-1 \leq r \leq 1$$

$$r = \frac{S_{xy}}{S_x S_y}$$

r es para una variable, R (mayúscula) es para dos variables.

La varianza es una medida de dispersión que representa la variabilidad de una serie de datos con respecto a su media. Formalmente se calcula como la suma de los cuadrados de los residuos dividida por las observaciones totales. Es decir la varianza es el promedio de los cuadrados de las desviaciones medidas al rededor de la media.

	Varianza	Desviación estándar	Media
Población	$\sigma^2 = \frac{(\sum x_i - \mu)^2}{N} = \frac{\sum x_i}{N} - \mu^2$	$\sigma = \sqrt{\sigma^2}$	$\mu = \frac{\sum x_i}{N}$
Muestra	$S^2 = \frac{(\sum x_i - \overline{x})^2}{n-1} = \frac{\sum x_i^2 - \frac{\sum x_i^2}{n}}{n-1}$	$S = \sqrt{S^2}$	$\sum \frac{x_i}{n}$

La covarianza es una medida descriptiva que permite el tipo de asociación lineal entre dos variables, se obtiene mediante la ecuación:

$$\mathrm{Cov}(x, y) = S_{xy} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

$$S_{xy} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

$$S_{xy} \begin{cases} = 0 \text{ ; no existe relación entre las variables} \\ < 0 \text{ ;existe una relación inversa entre las variables} \\ > 0 \text{ hay una relación directa entre las variables} \end{cases}$$

Práctica 5, 6 y 7: Se tienen los datos de la muestra de un grupo de estudiantes de la variable estatura x y peso y. Encontrar la recta de regresión de y sobre x o modelo de regresión.

Estud.	Estatura	Peso
	x	y
1	175	88
2	182	82
3	187	89
4	190	90
5	162	56
6	154	53
7	189	68
8	158	45
9	155	58
10	168	75

Ejemplo (similar a la práctica):

Estud.	Estatura	Peso
	x	y
1	154	33
2	158	45
3	162	56

4	166	73
5	170	65
6	174	88
7	178	89
8	182	75
9	186	89
10	190	90

$$y = \overline{y} + \frac{S_{xy}}{S_x^2}(x - \overline{x})$$

Se necesita la media, $\overline{x} = 172$ $\overline{y} = 72.3$.

Estud.	Estatura	Peso	Varianza	Varianza	Covarian-za	Covarian-za	Covarian-za
	x	y	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$x_1 - \overline{x}$	$y_i - \overline{y}$	$(x_1 - \overline{x}) \cdot (y_i - \overline{y})$
1	154	33	324	372.49	-18	-19.3	347.4
2	158	45	196	745.29	-14	-27.3	382.2
3	162	56	100	265.69			163
4	166	73	36	0.49			-4.2
5	170	65	4	53.29			14.6
6	174	88	4	246.49			31.6
7	178	89	36	278.89			100.2
8	182	75	100	7.29			27
9	186	89	196	278.89		16.7	233.8
10	190	90	324	313.24	18	17.7	318.6
			1320	2562.8			1614

$$S_x^2 = \frac{1320}{9} = 146.67$$

$$S_x = \sqrt{146.67} = 12.11$$

$$S_y^2 = \frac{2562.5}{9} = 284.68$$

$$S_x = \sqrt{284.68} = 16.87$$

$$\nu = \frac{S_{xy}}{S_x S_y} = \frac{179.33}{12.11 \cdot 16.87} = 0.877$$

$$y = \overline{y} + \frac{S_{xy}}{S_x^2}(x - \overline{x}) = 72.3 + \frac{179.33}{146.77}(x - 172)$$

$$y = -138.01 + 1.12227x \blacksquare$$

Resolución de la práctica 5, 6 y 7:

Estud.	Estatura	Peso
	(x)	(y)
1	175	88
2	182	82
3	187	89
4	190	90
5	162	56
6	154	53
7	189	68
8	158	45
9	155	58
10	168	75

Se calcula la media de las variables:

$$\overline{x} = \frac{175 + 182 + 187 + 190 + 162 + 154 + 189 + 158 + 155 + 168}{10} = \frac{1720}{10} = 172$$

$$\overline{y} = \frac{88 + 82 + 89 + 90 + 56 + 53 + 68 + 45 + 58 + 75}{10} = \frac{704}{10} = 70.4$$

A continuación, se construye la tabla para calcular las varianzas y la covarianza:

Estud.	Estatura	Peso	Varianza	Varianza	Covarian-za	Covarian-za	Covarian-za
	x_i	y_i	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x}) \cdot (y_i - \overline{y})$)
1	175	88	9	309.76	3	17.6	52.8
2	182	82	100	134.56	10	11.6	116
3	187	89	225	345.96	15	18.6	279
4	190	90	324	384.16	18	19.6	352.8
5	162	56	100	207.36	−10	−14.4	144
6	154	53	324	302.76	−18	−17.4	313.2
7	189	68	289	5.76	17	−2.4	−40.8
8	158	45	196	645.16	−14	−25.4	355.6
9	155	58	289	153.76	−17	−12.4	210.8
10	168	75	16	21.16	−4	4.6	−18.4
			1872	2510.4			1755.2

Se calculan las varianzas y la covarianza:

$$S_x^2 = \frac{1872}{9} = 208$$

$$S_y^2 = \frac{2510.4}{9} = 278.933$$

$$S_x = \sqrt{208} \approx 14.422$$

$$S_y = \sqrt{278.933} \approx 16.701$$

$$S_{xy} = \frac{1755.2}{9} \approx 195.022$$

El coeficiente de correlación es:

$$\nu = \frac{S_{xy}}{S_x S_y} = \frac{195.022}{14.422 \cdot 16.701} \approx 0.810$$

La recta de regresión es:

$$y = \overline{y} + \frac{S_{xy}}{S_x^2}(x - \overline{x}) = 70.4 + \frac{195.022}{208}(x - 172)$$

$y = 70.4 + 0.9376(x - 172) = 70.4 + 0.9376x - 161.2672 \approx -90.867 + 0.9376x$ ■

Práctica 8, 9, 10: Hacer la gráfica de $y = -138.01 + 1.12227x$ y sacar la conclusión de la relación entre variables con:

$$S_{xy}\begin{cases} = 0 \text{ ; no existe relación entre las variables} \\ < 0 \text{ ;existe una relación inversa entre las variables} \\ > 0 \text{ hay una relación directa entre las variables} \end{cases}$$

Para el proyecto, usar 30 observaciones (datos).

$$y = \overline{y} + \frac{S_{xy}}{S_x^2}(x - \overline{x})$$

$$r = \frac{S_{xy}}{S_x S_y}$$

$$y = \overline{y} + \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}(x - \overline{x})$$

El análisis de regresión es una técnica estadística para modelar y investigar la relación entre dos o más variables. En general, hay una sola variable o respuesta y que se relaciona con k variables independientes o regresivas $x_1, x_2, ..., x_k$.

La variable de respuesta y es una variable aleatoria, en tanto que las variables regresivas $x_1, x_2, ..., x_k$ miden con error especial.

Las variables se llaman también variables matemáticas, y con frecuencia son controladas por el experimentador. La relación de y con $x_1, x_2, ..., x_k$ se caracteriza por medio de un modelo matemático llamado ecuación de regresión.

El valor esperado de y para cada valor de x es la esperanza de y dado x, $E(y/x) = \beta_0 + b_1 x$.

$y = \beta_0 + \beta_1 x + \varepsilon$.
ε : error.
y : variable de respuesta.
x : variable predictoria o matemática.

Donde el error ε es aleatorio.$\varepsilon \sim N\left(0, \sigma^2\right)$.

Los $\{\varepsilon\}$ (el conjunto de todos los errores) son variables aleatorias pero son no correlacionales

2.2. Mínimos cuadrados.

El procedimiento de optimización es el de mínimos cuadrados. Optimizar significa encontrar los valores máximos o mínimos. Funciona con algoritmos iterativos. Esto se estima β_0 y β_1 de manera que la suma de los cuadrados de las desviaciones entre las observaciones y la línea de regresión deben tener valores mínimos.

$$y_i = \beta_0 + \beta_i x_i + \varepsilon_i \quad ; i = 1, \ldots, n$$
$$\varepsilon_i = y_i - \beta_0 - \beta_i x_i$$

La suma de cuadrados de las desviaciones de las observaciones respecto a la línea de regresión verdadera es:

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_{\text{observado}_i} - y_{\text{modelo}_i}\right)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)$$

y son los valores observados, $\hat{y}$ son los valores del modelo de regresión lineal. La diferencia entre cada y y su respectivo $\hat{y}$ es el error.

$$L = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_i x_i)^2$$

Los estimadores de mínimos cuadrados de β_0 y β_1 es $\hat{\beta}_0$ y $\hat{\beta}_1$. Es el estimador de la recta del modelo.

$$\sum y_i - \sum \hat{\beta}_0 - \sum \hat{\beta}_1 x_i = 0$$
$$\sum y_i x_i - \sum \hat{\beta}_0 x_i - \sum \hat{\beta}_1 x_i^2 = 0$$

$$\sum y_i \sum x_i - n\hat{\beta}_0 \sum x_i - \hat{\beta}_1 \left(\sum x_i\right)^2 = 0$$
$$-n \sum y_i x_i + n\hat{\beta}_0 \sum x_i + n\hat{\beta}_1 \sum x_i^2 = 0$$

$$\sum y_i \sum x_i - n\sum y_i x_i = \hat{\beta}_1\left(\left(\sum x_i\right)^2 - n\sum x^2\right)$$

$$\hat{\beta}_1 = \frac{\sum y_i x_i - n\sum y_i \sum x_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$\hat{\beta}_1 = \frac{\sum y_i x_i - \frac{\sum y_i \sum x_i}{n}}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}} \blacksquare$$

$$n\hat{\beta}_0 = \sum y_i - \beta_1 \sum x_i$$

$$\hat{\beta}_0 = \frac{\sum y_i}{n} - \beta_1 \frac{\sum x_i}{n}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \blacksquare$$

Para el experimento del parcial 1, el muestreo de datos vale 20 puntos y los cálculos, 70. El tamaño del muestreo debe de ser 30.

Dados β_0 y β_1, se deben encontrar $\widehat{\beta_0}$ y $\widehat{\beta_1}$,si lo hacemos, hemos encontrado el modelo de regresión.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\hat{y}_i = \overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 x_i$$

$$\hat{y}_i = \overline{y} + \hat{\beta}_1 (x_i - \overline{x})$$

$$\hat{\beta}_1 = \frac{\frac{\sum y_i x_i}{n} - \frac{\sum y_i}{n}\frac{\sum x_i}{n}}{\frac{\sum x_i^2}{n} - \frac{\left(\sum x_i\right)^2}{n^2}}$$

$$\hat{\beta}_1 = \frac{\frac{\sum y_i x_i}{n} - \overline{yx}}{\frac{\sum x_i^2}{n} - \overline{x}^2}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2}$$

$$\hat{y} = \overline{y} + \frac{S_{xy}}{S_x^2}(x_i - \overline{x}) \blacksquare$$

2.3. Supuestos del modelo de regresión lineal.

Para que un modelo sea de regresión lineal, debe cumplir con:

1. Linealidad
2. Normalidad
3. Homocelasticidad
4. Independencia.

Linealidad

La relación entre las variables independientes y la variable dependiente debe ser lineal. Se puede medir con r.

Normalidad

Los errores de la regresión deben seguir una distribución normal. Si la normalidad no se cumple (si la suma de los errores no da cero), los intérvalos de confianza y las pruebas de hipótesis pueden verse afectados, lo que puede conducir a conclusiones erróneas.

Homocelasticidad

La varianza de los errores debe ser constante en todos los niveles de las variables predictorias. Caso contrario, se produce heterocelasticidad, lo que significa que la dispersión de los errores varía en diferentes rangos de las variables predictorias.

Independencia

Indica que los errores de la regresión no deben estar correlacionados entre sí. Los errores o residuos de un modelo de regresión deben distribuirse de manera aleatoria, con distribución aleatoria normal, media cero y varianza constante.

Calcular ε_i y σ^2.

Estud.	Estatura	Peso
	x	y
1	154	53
2	158	45
3	162	56
4	166	73
5	170	65
6	174	88
7	178	89
8	182	75
9	190	90

$$\text{Peso} \sim N\left(\mu_i, \sigma^2\right)$$
$$\mu_i = \beta_0 + \beta_1 \text{estatura}_i$$
$$\sigma^2 = \text{constante}$$

$$\widehat{\text{Peso}}_i \sim N\left(\hat{\mu}_i, \hat{\sigma}^2\right)$$
$$\hat{\mu}_i = -138.0091 + 1.2227 \text{estatura}_i$$
$$\hat{y}_i = -138.0091 + 1.2227 \text{estatura}_i$$

Estud.	Estatura	Peso	Regresión	Residuales	Varianza residual
	x	y	$\hat{y}$	$\hat{e}_i = y_i - \hat{y}_i$	$(x_i - \overline{x})^2$
1	154	53	50.29	2.71	7.34
2	158	45	55.18	−10.18	103.67

3	162	56	60.07	-4.07	15.56
4	166	73	64.96	8.04	64.64
5	170	65	69.85	-4.85	23.52
6	174	88	74.74	13.26	175.82
7	178	89	79.63	9.37	87.79
8	182	75	84.52	-9.52	90.63
9	186	89	89.42	-0.42	0.17
10	190	90	94.30	-4.30	18.49
				$\overline{x} \approx 0$	$\sum (x_i - \overline{x})^2 = 588.62$

$$S^2 = \frac{\sum (x_i - \overline{x})^2}{n-1}$$

$$\hat{S}^2 == \frac{588.62}{9} \approx 65.40$$

$$\hat{S} = \sqrt{65.40} \approx 0.087 \blacksquare$$

Como el error promedio da cero, es válido como regresión lineal.

2.4. Aproximación polinomial.

2.5. Regresión lineal múltiple.

2.6. Dilema del sesgo y varianza.

En estadística y aprendizaje automático, el dilema entre sesgo y varianza describe la relación entre la complejidad de un modelo, la exactitud de supervisiones y su capacidad para realizar predicciones sobre datos no-observados previamente que no se utilizaron para entrenar al modelo. El sesgo y la varianza dependen de la complejidad del modelo

La complejidad del modelo

La complejidad del modelo se refiere al tipo y número de parámetros, entidades e interacciones que el modelo utiliza para aprender de los datos.

Un modelo complejo puede capturar más matices y patrones en los datos, pero también puede ser más propenso al sobreajuste, lo que significa que aprende demasiado del ruido, que no generaliza nuevos datos.

Un modelo simple puede ser más robusto y eficiente, pero también puede ser más propenso a la inadaptación, por lo que pierde relaciones importantes.

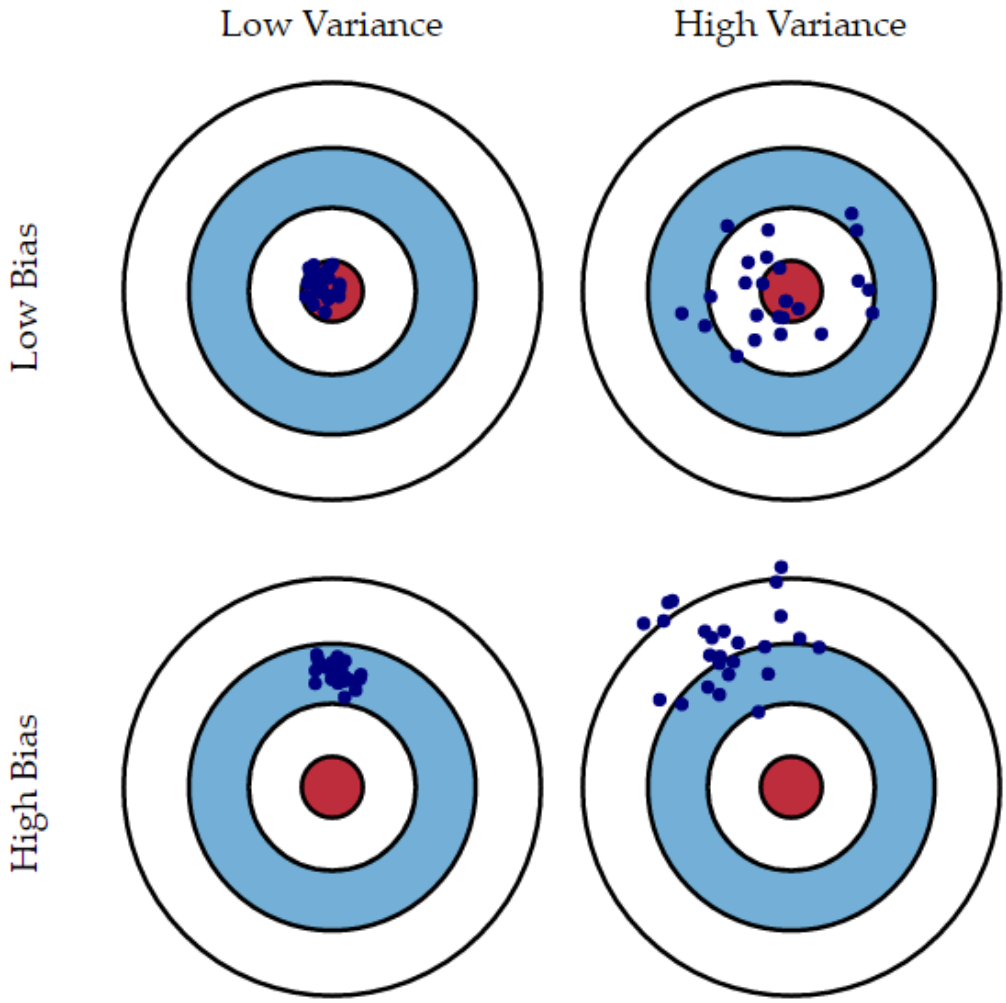
Comprender como las diferentes fuentes de error generan sesgo y varianza ayudará a mejorar el proceso de ajuste de datos.

El error de precicción estadísitca o cualquier algoritmo de aprendizaje automático se pueden dividir en tres partes

- Error de sesgo.
- Error de varianza.

- Error inverosimil.

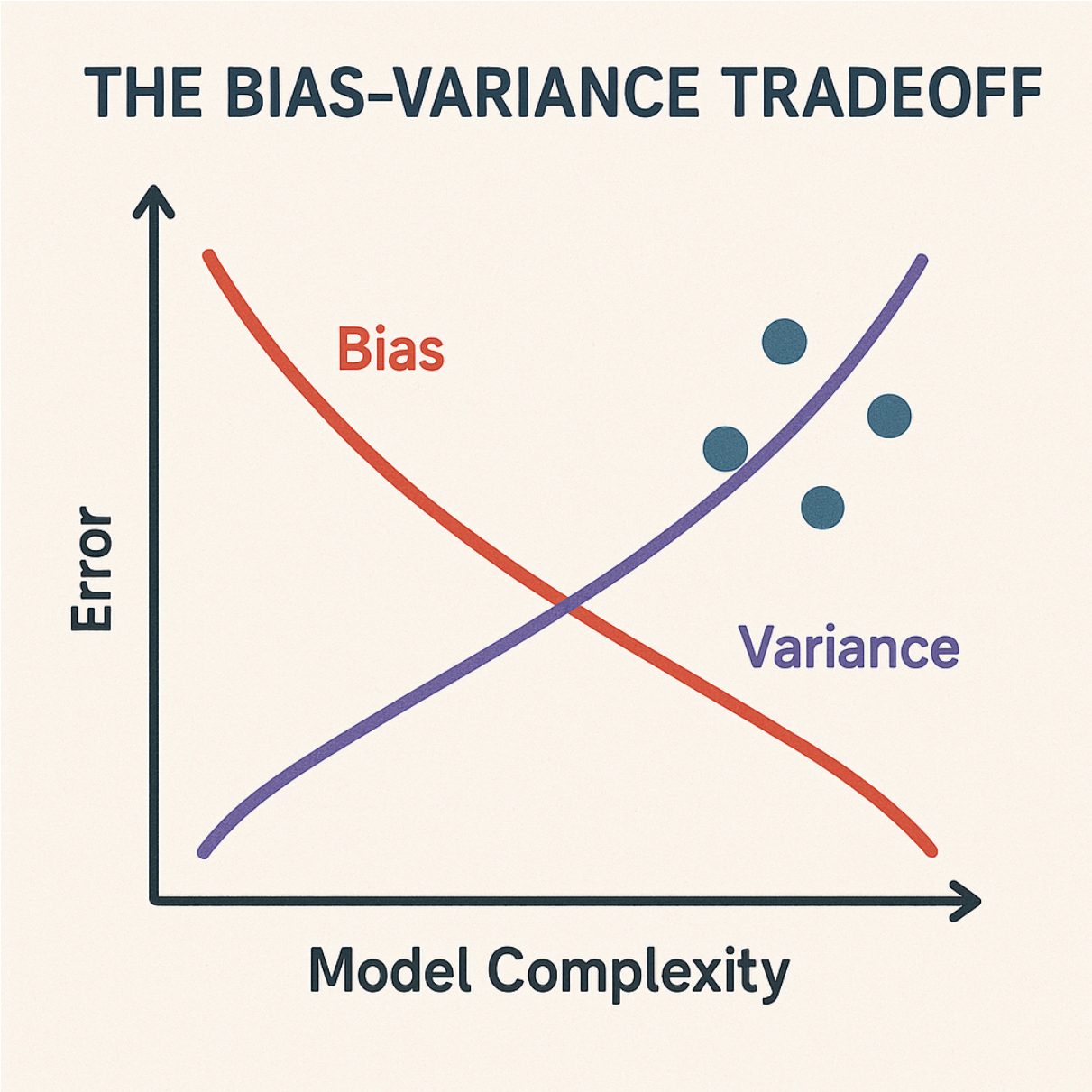
El error reducible no se puede reducir, independientemente de qué algoritmo se use. También se lo conoce como “ruido” y por lo general proviene de factores como variables desconocidos que influyen en el mapeo de las variables de entrada a la variable de salida.



El sesgo frente a la varianza se refiere a la presición frente a la consistencia de los modelos entrenados por un algoritmo.

- **Baja varianza - Alto sesgo**: Los algoritmos de baja varianza/alto sesgo tienden a ser menos complejos, con una estructura subyacente simple o rígida. Su fortaleza es la consistencia.
- **Bajo sesgo - Alta varianza**: Los algoritmos de bajo sesgo/alta varianza tienden a ser más complejos, con estructura subyacente flexible.

Error total. Comprender el sesgo y la varianza es fundamental para comprender el comportamiento de los modelos de predicción. El punto ideal para cualquier modelo es el nivel de complejidad en el que el aumento en el sesgo es equivalente a la reducción de la varianza.



El punto de equilibrio es cuando la función de la varianza intersecta con la del sesgo.

Por lo tanto, el dilema sesgo-varianza es: La idea inicial es reducir tanto el sesgo como la varianza, porque los dos son los componentes del error de predicción sobre los datos nuevos, pero esto no va a ser posible porque el dilema es: Si se quiere reducir el riesgo va a ser a costa de aumentar la varianza, y si se quiere reducir la varianza va a ser a costa de subir el sesgos.

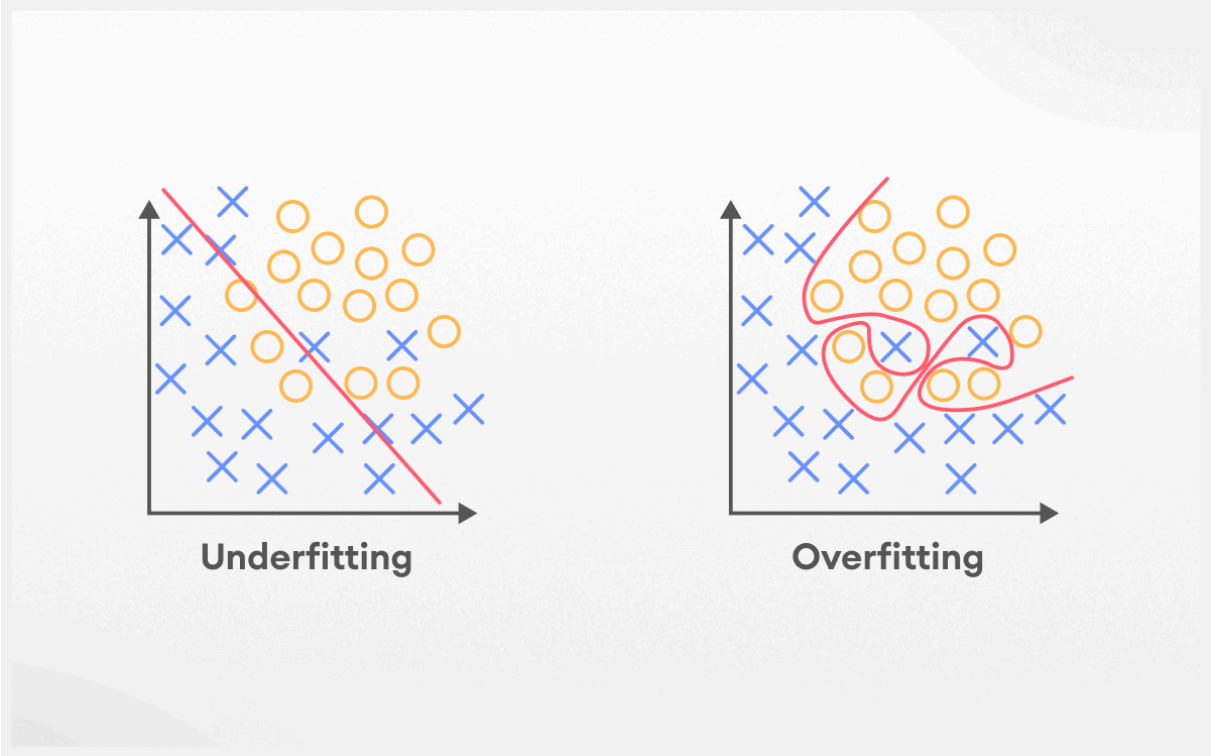
Sobreajuste (overfitting) y subajuste (underfitting)

Cuando se tiene un data-set o conjunto de datos, sirve para predecir o clasificar de acuerdo al problema, la idea es encontrar la presición con la implementación de un modelo con el conjunto de datos de entrenamiento y posteriormente con el conjunto de datos de pruebas.

Si la predicción es satifactoria, tendemos a aumentar la presición de la predicción con el conjunto de datos, ya sea aumentando o reduciendo la selección de las

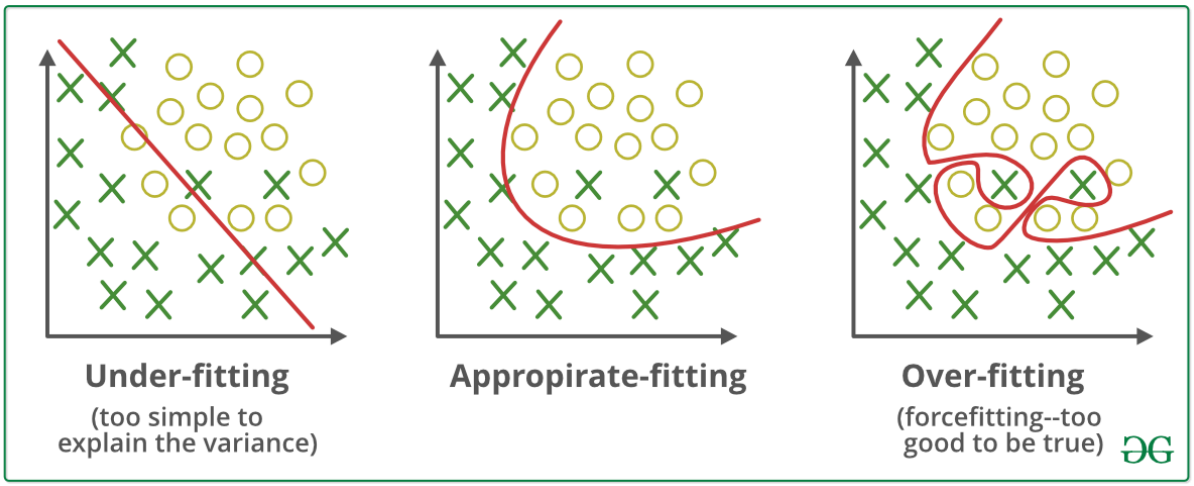
características o con la modificación de las condiciones del modelo de machine learning.

El pobre rendimiento del modelo puede ser porque es demasiado simple para describir el objetivo o por el contrario, que el modelo sea demasiado complejo para expresar el objetivo.



En el primer diagrama (underfitting) se muestra que la línea no cubre todos los puntos que se muestran en el gráfico, este modelo tiende a causar un ajuste insuficiente de los datos, se denomina también “alto sesgo”.

En el segundo diagrama (overfitting), la línea cubre todos los puntos del gráfico; en tal condición, se puede pensar que es algo bueno, ya que cubre todos los puntos, pero eso no es cierto en realidad, porque la línea en el gráfico cubre también los puntos que son ruido y los valores atípicos. Este modelo es responsable de predecir resultados deficientes debido a su complejidad. A esto también se lo denomina “alta varianza”.



Un modelo con un ajuste apropiado cubre la mayoría de los puntos, es decir, mantiene el equilibrio entre el sesgo y la varianza.

Subajuste

Se refiere a un modelo que no puede modelar los datos de entrenamiento o generalizar nuevos datos. Esto es porque el modelo es muy simple. El ajuste insuficiente destruye la presición del modelo en máquinas de aprendizaje o machine-learning. Su aparición simplemente significa que el modelo o el algoritmo no se ajusta a los datos suficientemente bien. Suele suceder cuando se tienen pocos datos para predecir el modelo o también cuando se intenta construir un modelo no lineal con datos lineales.

Sobreajuste

Se refiere a un modelo que modela los datos de entrenamiento demasiado bien. Esto ocurre cuando un modelo aprende el detalle incluyendo el ruido en los datos de entrenamiento. Esto significa que el ruido o las fluctuaciones aleatorias en los datos de entrenamiento son recogidos y aprendidos por el modelo.

Error cuadrático medio

$$\sum (y_i - \hat{y}_i)^2 = \sum \varepsilon_i^2$$

Se llama suma de errores al cuadrado, error cuadrático medio o media cuadrática.

$$MS_E = \frac{SS_E}{n-p}$$

Donde $n - p$ son los grados de libertad.

$$\hat{\sigma}^2 = MS_E$$

$$E(x) = \sum x_i P_i = \mu$$

$$\sigma^2 = \mathrm{Var}(x) = E\left((x-u)^2\right)$$

$$= \mathrm{Var}(x) = E\left(x^2\right) - E^2(x)$$

$$\mathrm{MSE}\left(\hat{\theta}\right) = E\left(\hat{\theta} - \theta\right)^2$$

$$\mathrm{MSE}\left(\hat{\theta}\right) = E\left(\hat{\theta} - E\left(\hat{\theta}\right) + E\left(\hat{\theta}\right) - \theta\right)^2$$

$$\mathrm{MSE}\left(\hat{\theta}\right) = \mathrm{Var}\hat{\theta} + \mathrm{Sesgo}\left(\hat{\theta}, \theta\right)^2$$

Otra alternativa de representar el error cuadrátivo medio es:

$$E\left(x^2\right) = \mathrm{Var}(x) + E^2(x)$$

Con un cambio de variable $x = \hat{\theta} - \theta$,

$$\mathrm{MSE}\left(\hat{\theta}\right) = E\left(\hat{\theta} - \theta\right)^2$$

$$\mathrm{MSE}\left(\hat{\theta}\right) = \mathrm{Var}\left(\hat{\theta} - \theta\right) + \mathrm{Sesgo}^2\left(\hat{\theta} - \theta\right)$$

El MSE mide la varianza y el sesgo.

Regresión lineal múltiple

Muchos problemas de regresión involucran más de una variable regresiva. Tales modelos se denominan regresión múltiple. En general, la variable dependiente o respuesta y puede relacionarse con k variables independientes. El modelo es $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon$. Donde $\beta_j = j = 0, 1, ..., k$ se denominan coeficientes de regresión.

Ecuación 1:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \blacksquare$$

Este modelo describe un hiperplano en el espacio k-dimensional de las variables regresoras x_j.

β_j representa el cambio esperado en la respuesta y. Y las variables restantes $x_i : i \neq j$ se mantienen restantes.

Los parámetros β_j se denominan algunas veces coeficientes de regresión parciales, porque ellos describen el efecto parcial de una variable independiente cuando las otras variables independientes en el modelo se mantienen constantes.

Los modelos más complejos en apariencia que la ecuación uno pueden ser analizados mediante técnicas de regresión lineal múltiple.

Ejemplos

Se tiene el modelo polinomial cubico en una variable independiente:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

$$\text{Si } x = x_1, x^2 = x_2, x^3 = x_3 :$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Los modelos que incluyen efectos de interacción también pueden analizarse por medio de métodos de regresión lineal múltiple.

Estimación de parámetros.

El método de mínimos cuadrados puede utilizarse para estimar los coeficientes de regresión lineal múltiple.

Se supone que se disponen $n < k$ observaciones y k_{ij} denota las observaciones i o el nivel de la variable x_j

Regresión lineal:

Datos	y	x_i
1	y_1	x_{11}
2	y_2	x_{21}
...	...	...
n	y_n	x_{n1}

Regresión lineal múltiple:

Datos	y	x_i	x_2	...	x_k
1	y_1	x_{11}	x_{12}	...	x_{ik}
2	y_2	x_{21}	x_{22}	...	x_{2k}
...	...	...	...	...	...
n	y_n	x_{n1}	x_{n2}	...	x_{nk}

Se supone que el término de error en el modelo tiene $E(\varepsilon) = 0$ y la varianza del error $V(E) = \sigma^2$.

El conjunto de variables del error son variables aleatorias no correlacionadas.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_k x_{ik} + \varepsilon_i$$

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \quad : 1 \leq i \leq n$$

La función de mínimos cuadrados es la sumatoria del error al cuadrado:

$$L = \sum \varepsilon_i^2$$

$$L = \sum \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

La función L se minimiza con respecto a $\beta_0, \beta_1, ..., \beta_k$

$$\frac{\partial L}{\partial \beta_0} \mid \hat{\beta}_0 \hat{\beta}_1, ..., \hat{\beta}_k = -2 \sum \left(y_i - \hat{\beta}_0 - \sum \hat{\beta}_j x_{ij} \right) = 0$$

$$\frac{\partial L}{\partial \beta_j} \mid \hat{\beta}_0 \hat{\beta}_1, ..., \hat{\beta}_k = -2 \sum \left(y_i - \hat{\beta}_0 - \sum \hat{\beta}_j x_{ij} \right) x_{ij} = 0 \quad : j = 1, 2, ..., k$$

$$\sum y_i - \sum \hat{\beta}_0 - \sum \hat{\beta}_1 x_{i1} - \sum \hat{\beta}_2 x_{i2} - ... - \sum \hat{\beta}_k x_{ik} = 0$$

Para $j = 1, \hat{\beta}_i$

$$\sum y_i x_{i1} - \sum \hat{\beta}_0 x_{i1} - \sum \hat{\beta}_1 x_{i1}^2 - \sum \hat{\beta}_2 x_{i2} - ... - \sum \hat{\beta}_k x_{i1} x_{ik} = 0$$

Para $j = 2, \hat{\beta}_2$

$$\sum y_i x_{i2} - \sum \hat{\beta}_0 x_{i2} - \sum \hat{\beta}_1 x_{i2} - \sum \hat{\beta}_2 x_{i2}^2 - ... - \sum \hat{\beta}_k x_{i2} x_{ik} = 0$$

...

Para $j = k, \hat{\beta}_k$

$$\sum y_i x_{ik} - \sum \hat{\beta}_0 x_{ik} - \sum \hat{\beta}_1 x_{ik} - \sum \hat{\beta}_2 x_{ik} - ... - \sum \hat{\beta}_k x_{ik}^2 = 0$$

$$n_0 \hat{\beta}_0 + \hat{\beta}_1 \sum x_{i1} + \hat{\beta}_2 \sum x_{12} + ... + \hat{\beta}_k \sum y_{ik} = \sum y_i$$

$$\hat{\beta}_0 \sum x_{i1} + \hat{\beta}_1 \sum x_{i1}^2 + \hat{\beta}_2 \sum x_{i1} x_{12} + ... + \hat{\beta}_k \sum x_{i1} x_{ik} = \sum y_i x_{i1}$$

...

$$\hat{\beta}_0 \sum x_{ik} + \hat{\beta}_1 \sum x(ik) + \hat{\beta}_2 \sum x_{ik} x_{12} + ... + \hat{\beta}_k \sum x_{ik}^2 = \sum y_i x_{ik}$$

$$p = k + 1$$

$$\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$$

La solución de las ecuaciones normales serán los estimadores de mínimos cuadrados de los coeficientes de regresión $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$

Práctica 11, 12, 13, 14: Encontrar $\beta_0, \beta_1, \beta_2$ de manera genérica

Todas estas ecuaciones se pueden resolver de manera matricial.

$$y_i = \beta_0 + \sum \beta_j x_{ij} + \varepsilon_i \quad : i = 1, ...n$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{bmatrix}$$

$$x = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1k} \\ x_{21} & x_{22} & ... & x_{2k} \\ ... & ... & ... & ... \\ x_{n1} & x_{n2} & ... & x_{nk} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_k \end{bmatrix}$$

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

$$y = x \cdot \beta + \varepsilon$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} . \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (y - x\beta)' - (y - x\beta)$$

$$= (y' - \beta' x')(y - x\beta)$$

$$= y'y - y'x\beta - \beta' x' y + \beta' x' x\beta L = y'y - 2\beta' y' y + \beta' x' x\beta$$

$$\frac{\partial L}{\partial \beta} = -2x'y + 2x'x\hat{\beta} = 0$$

$$\hat{\beta} = (x'x)^{-1} x'y$$

$$\boxed{\hat{\beta} = (x'x)^{-1} x'y \blacksquare}$$

Ejemplo

$$x = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \end{bmatrix}$$

$$x'x = \begin{bmatrix} 1 & 1 & 1 \\ x_{11} & x_{21} & x_{31} \\ x_{12} & x_{22} & x_{32} \end{bmatrix} \cdot \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \end{bmatrix} = \begin{bmatrix} 3 & \sum_{i=1}^{3} x_{i1} & \sum_{i=1}^{3} x_{i2} \\ \sum_{i=1}^{3} x_{i1} & \sum_{i=1}^{3} x_{i1}^2 & \sum_{i=1}^{3} x_{i2} \\ \sum_{i=1}^{3} x_{i1} & \sum_{i=1}^{3} x_{i2} x_{i1} & \sum_{i=1}^{3} x_{i2}^2 \end{bmatrix}$$

De manera general:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{3} x_{i2} & \dots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1} x_{i2} & \dots & \sum_{i=1}^{n} x_{i1} x_{ik} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik} x_{i1} & \sum_{i=1}^{n} x_{ik} x_{i2} & \dots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix} \cdot \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \dots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1} y_i \\ \dots \\ \sum_{i=1}^{n} x_{ik} y_i \end{bmatrix}$$

Para minimizar el error ε se usan mínimos cuadrados.

$x' \cdot x$ es una matriz simétrica de $p \cdot p$ y $x' \cdot y$ es $p \cdot 1$. Entonces $\hat{y} = x \cdot \hat{\beta}$

Ejemplo

Modelado de precios con regresión múltiple: Una tienda analiza las ventas y en función de:

- Gasto en publicidad (x_1).
- Peso del producto (x_2).

Mes	Ventas	Peso x_1	Precio x_2
1	500	10	20
2	550	15	18
3	600	20	17
4	700	25	16

Ajustar un modelo de regresión múltiple lineal para aumentar las ventas.

Ejemplo

En la clases de DAT 246 se realizó un estudio a doce estudiantes para ver cómo influyen faltar a clases y las notas semifinales en el exámen final. Se ajustará el modelo de regresión múltiple.

Estudiante	Examen final y	Nota semifinal x_1	Asistencia x_2
1	80	65	1
2	80	67	2
3	69	65	2
4	67	58	4
5	51	55	4
6	51	45	5
7	40	47	6
8	45	39	7
9	45	44	7
10	30	29	8
11	32	29	9
12	25	21	9

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \qquad i = 1, \ldots, 12$$

$$y = \begin{bmatrix} 80 \\ 80 \\ 69 \\ 67 \\ 51 \\ 51 \\ 40 \\ 45 \\ 45 \\ 30 \\ 32 \\ 25 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 65 & 1 \\ 1 & 67 & 2 \\ 1 & 65 & 2 \\ 1 & 58 & 4 \\ 1 & 55 & 4 \\ 1 & 45 & 5 \\ 1 & 47 & 6 \\ 1 & 39 & 7 \\ 1 & 44 & 7 \\ 1 & 29 & 8 \\ 1 & 29 & 9 \\ 1 & 21 & 9 \end{bmatrix}$$

$$\sum y_i = 615 \qquad \sum x_{i1} = 564 \qquad \sum x_{i2} = 64$$

$$\sum x_{i1}^2 = 29142 \qquad \sum x_{i1}x_{i2} = 2551 \qquad \sum x_{i2}^2 = 426$$

$$\sum x_{i1}y_i = 31969 \qquad \sum x_{i2}y_i = 2728$$

$$X'X = \begin{bmatrix} 12 & 564 & 64 \\ 564 & 29142 & 2551 \\ 64 & 2551 & 426 \end{bmatrix}$$

$$X'y = \begin{bmatrix} 615 \\ 31969 \\ 2728 \end{bmatrix}$$

$$\hat{\beta} = (X'X)^{-1}X'y = \begin{bmatrix} 47.736 \\ 0.505 \\ -3.793 \end{bmatrix}$$

$$\widehat{y_i} = 47.736 + 0.505x_{i1} - 3.793x_{i2}$$

Calculando los estimadores $\hat{y}_i$ para cada estudiante:

Usando el modelo ajustado:

$$\hat{y}_i = 47.736 + 0.505x_{i1} - 3.793x_{i2}$$

Calculamos $\hat{y}_i$ para cada estudiante:

- Estudiante 1: $x_1 = 65$, $x_2 = 1$

  $\hat{y}_1 = 47.736 + 0.505 \cdot 65 - 3.793 \cdot 1 = 47.736 + 32.825 - 3.793 = 76.768$

- Estudiante 2: $x_1 = 67$, $x_2 = 2$

  $\hat{y}_2 = 47.736 + 0.505 \cdot 67 - 3.793 \cdot 2 = 47.736 + 33.835 - 7.586 = 73.985$

- Estudiante 3: $x_1 = 65$, $x_2 = 2$

  $\hat{y}_3 = 47.736 + 0.505 \cdot 65 - 3.793 \cdot 2 = 47.736 + 32.825 - 7.586 = 72.975$

- Estudiante 4: $x_1 = 58$, $x_2 = 4$

  $\hat{y}_4 = 47.736 + 0.505 \cdot 58 - 3.793 \cdot 4 = 47.736 + 29.29 - 15.172 = 61.854$

- Estudiante 5: $x_1 = 55$, $x_2 = 4$

  $\hat{y}_5 = 47.736 + 0.505 \cdot 55 - 3.793 \cdot 4 = 47.736 + 27.775 - 15.172 = 60.339$

- Estudiante 6: $x_1 = 45$, $x_2 = 5$

  $\hat{y}_6 = 47.736 + 0.505 \cdot 45 - 3.793 \cdot 5 = 47.736 + 22.725 - 18.965 = 51.496$

- Estudiante 7: $x_1 = 47$, $x_2 = 6$

  $\hat{y}_7 = 47.736 + 0.505 \cdot 47 - 3.793 \cdot 6 = 47.736 + 23.735 - 22.758 = 48.713$

- Estudiante 8: $x_1 = 39$, $x_2 = 7$

  $\hat{y}_8 = 47.736 + 0.505 \cdot 39 - 3.793 \cdot 7 = 47.736 + 19.695 - 26.551 = 40.88$

- Estudiante 9: $x_1 = 44$, $x_2 = 7$

  $\hat{y}_9 = 47.736 + 0.505 \cdot 44 - 3.793 \cdot 7 = 47.736 + 22.22 - 26.551 = 43.405$

- Estudiante 10: $x_1 = 29$, $x_2 = 8$

  $\hat{y}_{10} = 47.736 + 0.505 \cdot 29 - 3.793 \cdot 8 = 47.736 + 14.645 - 30.344 = 32.037$

- Estudiante 11: $x_1 = 29$, $x_2 = 9$

  $\hat{y}_{11} = 47.736 + 0.505 \cdot 29 - 3.793 \cdot 9 = 47.736 + 14.645 - 34.137 = 28.244$

- Estudiante 12: $x_1 = 21$, $x_2 = 9$

  $\hat{y}_{12} = 47.736 + 0.505 \cdot 21 - 3.793 \cdot 9 = 47.736 + 10.605 - 34.137 = 24.204$

Estudiante	Examen final y	Nota semifinal x_1	Asistencia x_2	$\hat{y}_i$	$\varepsilon_i = y - \hat{y}_i$
1	80	65	1	76.768	3.22
2	80	67	2	73.985	6.00
3	69	65	2	72.975	−3.98
4	67	58	4	61.854	5.13
5	51	55	4	60.339	−9.34
6	51	45	5	51.496	−0.50
7	40	47	6	48.713	−8.72
8	45	39	7	40.88	4.4

9	45	44	7	43.405	1.58
10	30	29	8	32.037	−2.04
11	32	29	9	28.244	3.75
12	25	21	9	24.204	0.79

2.7. Propiedades estadísticas del estimador.

Las propiedades estadísticas del estimador de mínimos cuadrados $\hat{\beta}$ pueden demostrarse como:

$$E\left(\hat{\beta}\right) = E\left((x' \cdot x)^{-1} x' \cdot y\right)$$
$$= E\left((x' \cdot x)^{-1} x' \cdot (x \cdot \beta + \varepsilon)\right)$$
$$= E\left((x' \cdot x)^{-1} x' \cdot x \cdot \beta + (x' \cdot x)^{-1} x' \cdot \varepsilon\right)$$
$$= I\beta$$
$$: E(\varepsilon) = 0, E\left(\hat{\beta}\right) = \beta$$

2.8. Propiedades de la varianza

$$\text{Cov}(x, y) = S_{xy} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$
$$S_{xy} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$
$$S_{xy} \begin{cases} = 0 \text{ ; no existe relación entre las variables} \\ < 0 \text{ ;existe una relación inversa entre las variables} \\ > 0 \text{ hay una relación directa entre las variables} \end{cases}$$

La propiedad de varianza de $\hat{\beta}$ se expresa mediante la covarianza:

$$\text{Cov}\left(\hat{\beta}\right) = E\left(\left(\hat{\beta} - E\left(\hat{\beta}\right)\right)\left(\hat{\beta} - E\left(\hat{\beta}\right)\right)'\right)$$

$$\text{Cov}\left(x_i, x_j\right) = \begin{bmatrix} \text{Var}(x_1) & \text{Cov}(x_1, x_2) & \ldots & \text{Cov}(x_1, x_n) \\ \text{Var}(x_2, x_1) & \text{Var}(x_2) & \ldots & \text{Cov}(x_2, x_n) \\ \ldots & \ldots & \ldots & \ldots \\ \text{Cov}\left(x_n, x_j\right) & \text{Cov}(x_1, x_2) & \ldots & \text{Var}(x_n) \end{bmatrix}$$

$$\text{Cov}\left(x_i, x_j\right) \text{ es simétrica}$$
$$\text{Var}(AX) = A \text{ Var } (X) A^T$$
$$\text{Cov}\left(\hat{\beta}\right) = \sigma^2 (x'x)'$$

Es necesario estimar la varianza σ^2 . Para desarrollar este estimador se considera la suma de los cuadrados de los residuos.

Suma de los errores al cuadrado:

$$\sum (y_i - \overline{y})^2 = \sum \varepsilon_i^2 = \varepsilon' \varepsilon = \mathrm{SS}_E$$

$$\hat{y} = x\hat{\beta}$$

$$\varepsilon = y - \hat{y} = y - x\hat{\beta}$$

$$\mathrm{SS}_E = \left(y - x\hat{\beta}\right)'\left(y - x\hat{\beta}\right)$$

$$= y'y - y'x\hat{\beta} - \hat{\beta}'x'y + \hat{\beta}'x'$$

$$= y'y - 2\hat{\beta}'x'y + \hat{\beta}'x'x\hat{\beta}$$

$$= y'y - 2\hat{\beta}'x'y + \beta'x'y$$

La media cuadrática es:

$$MS_E = \frac{SS_E}{n-p} = \sigma^2$$

$$y'y = \sum y^2$$

Ejemplo

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \quad i = 1, \ldots, 12$$

$$y = \begin{bmatrix} 80 \\ 80 \\ 69 \\ 67 \\ 51 \\ 51 \\ 40 \\ 45 \\ 45 \\ 30 \\ 32 \\ 25 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 65 & 1 \\ 1 & 67 & 2 \\ 1 & 65 & 2 \\ 1 & 58 & 4 \\ 1 & 55 & 4 \\ 1 & 45 & 5 \\ 1 & 47 & 6 \\ 1 & 39 & 7 \\ 1 & 44 & 7 \\ 1 & 29 & 8 \\ 1 & 29 & 9 \\ 1 & 21 & 9 \end{bmatrix}$$

$$\sum y_i = 615 \qquad \sum x_{i1} = 564 \qquad \sum x_{i2} = 64$$

$$\sum x_{i1}^2 = 29142 \qquad \sum x_{i1}x_{i2} = 2551 \qquad \sum x_{i2}^2 = 426$$

$$\sum x_{i1}y_i = 31969 \qquad \sum x_{i2}y_i = 2728$$

$$X'X = \begin{bmatrix} 12 & 564 & 64 \\ 564 & 29142 & 2551 \\ 64 & 2551 & 426 \end{bmatrix}$$

$$X'y = \begin{bmatrix} 615 \\ 31969 \\ 2728 \end{bmatrix}$$

$$\hat{\beta} = (X'X)^{-1}X'y = \begin{bmatrix} 47.736 \\ 0.505 \\ -3.793 \end{bmatrix}$$

$$\hat{y} = 47.736 + 0.05x_1 - 3.793x_2$$

Práctica 15, 16, 17: Sacar la inversa de:

$$\begin{bmatrix} 12 & 564 & 64 \\ 564 & 29142 & 2551 \\ 64 & 2551 & 426 \end{bmatrix}$$

Error estándar:

$$S_{y,1,2,3,\ldots,k} = \sqrt{\frac{\text{SS}_E}{n-(k+1)}}$$

Coeficiente de determinación múltiple.

Es el porcentaje de variación de la variable dependiente y explicada por el conjunto de variables independientes x_i

$$R^2 \qquad 0 \leq R^2 \leq 1$$

Sacar cero indica poca sucesión con el conjunto de varaibles independientes y la variable dependiente, y sacar uno indica una sucesión fuerte.

$$R^2 = \frac{\text{SCR}}{\text{STC}} = \frac{\sum(\hat{y}-\overline{y})^2}{\sum e_i^2 + \sum(\hat{y}-\overline{y})}$$

- SCR: Suma de cuadrados de regresión.
- STC: Suma de los cuadrados.

Ejemplo:

Estudiante	Examen final y_i	$\hat{y}_i$	$\varepsilon_i = y - \hat{y}_i$	e_i^2	$(\hat{y}-\overline{y})^2$
1	80	76.768	3.232	10.446	651.168
2	80	73.985	6.015	36.18	516.88
3	69	72.975	−3.975	15.801	471.976
4	67	61.854	5.146	26.481	112.445
5	51	60.339	−9.339	87.217	82.61
6	51	51.496	−0.496	0.246	0.061

7	40	48.713	−8.713	75.916	6.436
8	45	40.88	4.12	16.974	107.537
9	45	43.405	1.595	2.544	61.544
10	30	32.037	−2.037	4.149	369.139
11	32	28.244	3.756	14.108	529.276
12	25	24.204	0.796	0.634	731.486

El valor de R^2 es: 0.9258

El número de variables independientes x_i en una ecuación múltiple aumenta el coeficiente de determinación, hace que las predicciones sean más precisas.

R^2 aumenta solo debido al número total de variables independientes y no porque la variable independiente que se agrega sea un buen factor de predicción de la dependiente, por ejemplo, si se agrega una tercera variable x y que este fuera la distancia a la universidad, R^2 va a aumentar y esto no quiere decir que la distancia a la universidad sea un buen factor para predecir la nota del examen final, simplemente R^2 aumentará porque se ha añadido otra variable independiente.

Para ajustar el valor de R^2, se emplea un coeficiente ajustado de determinación múltiple, que es modificado para ajustar el número de variables y el tamaño de la muestra.

$$R^2 \text{ ajustado} = 1 - \left(1 - R^2\right)\frac{n-1}{n-k-1}$$

$$R^2 \text{ ajustado} = 1 - (1 - 0.9258)\frac{12-1}{12-2-1} = 0.9093$$

Tiene una exactitud de predicción del 90.93 %

Normalidad:

$$\varepsilon_i \sim \left(0, \sigma^2\right)$$

Linealidad:

$$E(\varepsilon_i) = 0$$

3. Máxima Verosimilitud.

3.1. Estimación de parámetros.

3.2. Estimación de máxima verosimilitud.

4. Estadística bayesiana.

4.1. Repaso del teorema de Bayes.

4.2. Descripción a priori y a posteriori.

4.3. Inferencia bayesiana.

4.4. Clasificador bayesiano ingenuo.

4.5. Estimación MAP (estimador máximo a posteriori).

4.6. Teoría de desición bayesiana.

4.7. Regresión lineal bayesiana.

5. Métodos basados en árboles.

5.1. Árboles de regresión.

5.2. Árboles de clasificación.

6. Bibliografía.

- Germán A., Carolin J., Donald. Bayesian Data Analysis.
- Murphy - Machine Learning: A Probabilistic Perspective.

Mas cualquier libro de estadística.

> Piensa como un bayesiano, preocúpate como un frecuentista.
>
> — Beadley Efron

Primer parcial.	30 puntos.
Segundo parcial.	25 puntos.
Examen Final.	30 puntos.
Prácticas y participación en clases.	15 puntos.
Asistencia.	5 puntos extra.

Para las prácticas en clase se hará una tarjeta de 12 cm × 10 cm con:

Dat 246 II/2025

Paterno, materno, nombres.				
Fecha.	Firma.			

Práctica 1: Plan de trabajo.

Práctica 2: Frase de motivación.

Práctica 3: Carátula donde cada uno se compromete a aprobar la materia.