Aprendizaje Supervisado

Regresión Lineal

ML-6561

October 23, 2023

Contenido

Índice

¿Qué es el Aprendizaje Supervisado?

- El aprendizaje supervisado es un tipo de aprendizaje automático donde el algoritmo aprende a partir de **datos etiquetados** un modelo matemático que llamaremos f.
- **Cada instancia** de datos tiene una etiqueta, valor o clase conocida.

Objetivo del Aprendizaje Supervisado

Encontrar una función, denotada como f, que mapee los datos de entrada X a las etiquetas de salida Y.

$$Y = f(x) \text{ para } x \in X \tag{1}$$

o de igual forma

$$f : X \to Y \tag{2}$$

Donde:

- X representa los datos de entrada.
- Y son las etiquetas de salida o las respuestas deseadas.
- $f(\cdot)$ es la función que el modelo de aprendizaje supervisado busca aprender.

Definiciones sobre X y f

- Denominaremos a X como conjunto de datos de entrenamiento o datos de entrada

$$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})\}$$

- $y^{(i)}$ son las etiquetas de salida para cada $x^{(i)}$.
- Dada f, la **predicción** de $x^{(i)}$, $f(x^{(i)})$ la representaremos como $\hat{y}^{(i)}$.
- $Error_i(x^{(i)}) = y^{(i)} - \hat{y}^{(i)} = y^{(i)} - f(x^{(i)})$

Objetivo principal equivalente

El **objetivo** es encontrar una función f que minimice los errores entre las predicciones y las etiquetas reales.

Función de Coste en Aprendizaje Supervisado

Generalmente definida como $J(\theta)$ o $J_\theta(y^{(i)}, \hat{y}^{(i)})$ evalúa qué tan cerca están las predicciones del modelo $\hat{y}^{(i)}$ de las etiquetas reales $y^{(i)}$ en función de los parámetros del modelo θ.
Algunas funciones de coste comunes incluyen:

- **Clasificación**: Entropía Cruzada, Error Cuadrático Medio.
- **regresión**: Error Cuadrático Medio, Error Absoluto Medio.

Función de Coste, $J_\theta(y^{(i)}, \hat{y}^{(i)})$

El objetivo del Aprendizaje Supervisado es encontrar los parámetros (θ) que minimizan la función de coste.

Problema de Optimización

El objetivo es encontrar los parámetros θ que minimicen una función de costo $J(\theta)$:

$$\theta^* = \arg\min_{\theta} J(\theta) \tag{3}$$

Donde:

- $J(\theta)$ es la función de costo que mide la discrepancia entre las predicciones del modelo y las etiquetas reales.
- θ son los parámetros del modelo a aprender.
- θ^* representa los parámetros óptimos que minimizan $J(\theta)$.

La tarea es encontrar los parámetros que generen un modelo capaz de hacer predicciones precisas.

Error Cuadrático Medio (MSE)

El Error Cuadrático Medio (MSE) es una función de coste comúnmente utilizada en problemas de regresión y, en algunos casos, en problemas de clasificación. Se define como:

$$MSE = \frac{1}{m} \sum_{i=1}^{m} (y^{(i)} - \hat{y}^{(i)})^2 \tag{4}$$

Donde:

- m es el número de ejemplos en el conjunto de entrenamiento.
- $y^{(i)}$ es la etiqueta real del ejemplo i.
- $\hat{y}^{(i)}$ es la predicción del modelo para el ejemplo i.

El objetivo es minimizar el MSE ajustando los parámetros del modelo.

Conclusión

- El aprendizaje supervisado es fundamental en numerosos campos.
- Existen diversos modelos para abordar problemas de clasificación y regresión.
- La elección del modelo depende de la naturaleza del problema y los datos disponibles.

Tarea

Table: Datos de prueba

x	y
1	.7
2	3.1
3	2.6
4	3.1
5	5.1
6	7.3
7	8.3
8	9.2
9	5.4
10	10.9)

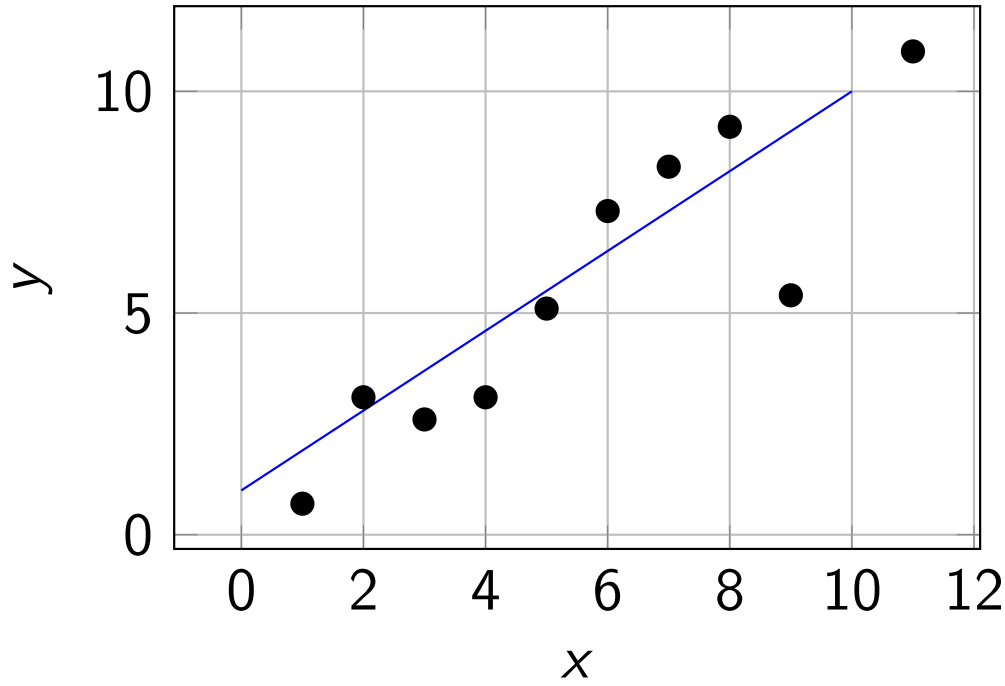


Figure: Calculen el MSE de los siguientes datos asumiendo $f(x) = 0.9 * x + 1$, identifiquen m, cómo saben si el modelo es "bueno"?

Índice

Predicción de sueldos

Salarios y otros datos para un grupo de 3000 trabajadores varones en la región del Atlántico Medio.

- **year**: Año en que fue preguntada la información
- **age**: Edad del trabajador
- **maritl**: Un factor con niveles '1. Never Married', '2. Married', '3. '3. Widowed', '4. Divorced' and '5. Separated' indicando estado civil.
- **race**: Un factor con niveles '1. White', '2. Black', '3. Asian' and '4. Other' indicando raza
- **education**: Un factor con niveles '1. ¡ HS Grad', '2. HS Grad', '3. Some College', '4. College Grad' and '5. Advanced Degree' indicando nivel educativo

Predicción de sueldos

- **region**: Región de USA (mid-atlantic solamente)
- **jobclass**: Un factor con niveles '1. Industrial' and '2. Information' indicando tipología de trabajo
- **health**: Un factor con niveles '1. <=Good' and '2. >=Very Good' indicando condición de salud del trabajador
- **health_ins**: Un factor con niveles '1. Yes' and '2. No' indicando si el trabajador tiene o no seguro médico
- **logwage**: Logaritmo de los salarios
- **wage**: Salario del trabajador.

Predicción de sueldos

	year	age	maritl	race	education	region	jobclass	health	health_ins	logwage	wage
0	2006	18	1. Never Married	1. White	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	2. No	4.318063	75.043154
1	2004	24	1. Never Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	4.255273	70.476020
2	2003	45	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	1. <=Good	1. Yes	4.875061	130.982177
3	2003	43	2. Married	3. Asian	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	5.041393	154.685293
4	2005	50	4. Divorced	1. White	2. HS Grad	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.318063	75.043154
...	...	...	...	...	...	...	...	...	...	...	...
2995	2008	44	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	5.041393	154.685293
2996	2007	30	2. Married	1. White	2. HS Grad	2. Middle Atlantic	1. Industrial	2. >=Very Good	2. No	4.602060	99.689464
2997	2005	27	2. Married	2. Black	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	2. No	4.193125	66.229408
2998	2005	27	1. Never Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	4.477121	87.981033
2999	2009	55	5. Separated	1. White	2. HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	1. Yes	4.505150	90.481913

Figure: Salarios y otros datos para un grupo de 3000 trabajadores varones en la región del Atlántico Medio. Disponible en ISLP.

Predicción de sueldos

Objetivo

Un startup que usa datos de Linkedin busca analizar estos datos para predecir dado un candidato el rango salarial que estaría dispuesto a aceptar.

1. Cuál es la entrada de nuestro modelo? X
2. Cuál es la salida de nuestro modelo? y
3. Cuáles son los mejores predictores de y?
4. Existe relación entre los predictores?

Predicción de sueldos

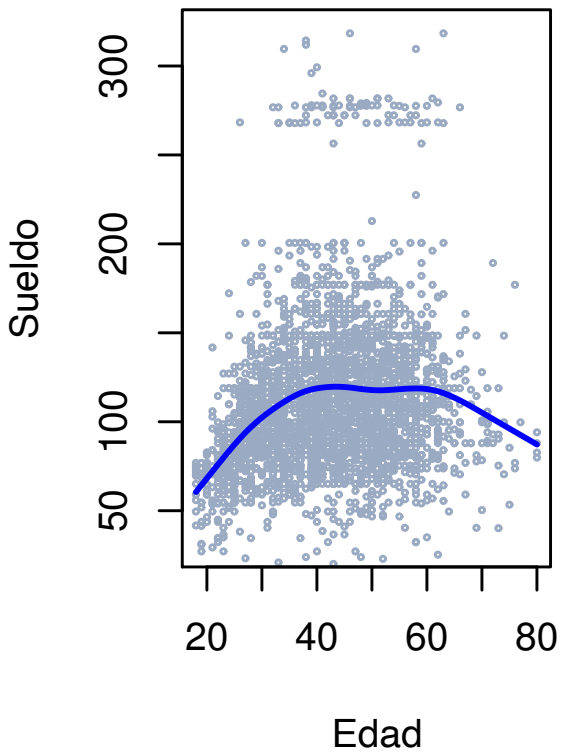
	year	age	maritl	race	education	region	jobclass	health	health_ins	logwage	wage
0	2006	18	1. Never Married	1. White	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	2. No	4.318063	75.043154
1	2004	24	1. Never Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	4.255273	70.476020
2	2003	45	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	1. <=Good	1. Yes	4.875061	130.982177
3	2003	43	2. Married	3. Asian	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	5.041393	154.685293
4	2005	50	4. Divorced	1. White	2. HS Grad	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.318063	75.043154
...	...	...	...	...	...	...	...	...	...	...	...
2995	2008	44	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	5.041393	154.685293
2996	2007	30	2. Married	1. White	2. HS Grad	2. Middle Atlantic	1. Industrial	2. >=Very Good	2. No	4.602060	99.689464
2997	2005	27	2. Married	2. Black	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	2. No	4.193125	66.229408
2998	2005	27	1. Never Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	4.477121	87.981033
2999	2009	55	5. Separated	1. White	2. HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	1. Yes	4.505150	90.481913

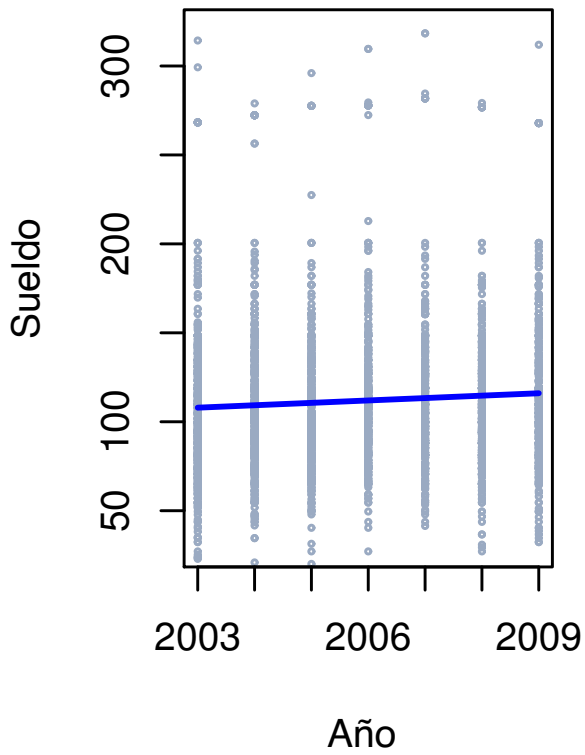
Figure: Salarios y otros datos para un grupo de 3000 trabajadores varones en la región del Atlántico Medio. Disponible en ISLP.

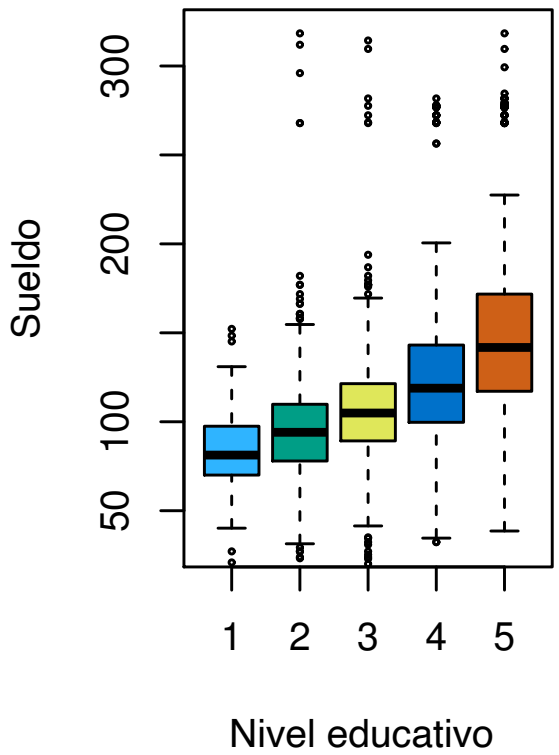
1. Cuál es la entrada de nuestro modelo? X
2. Cuál es la salida de nuestro modelo? y
3. Cuáles son los mejores predictores de y?
4. Existe relación entre los predictores?

Predicción de sueldos

Análisis exploratorio de datos







- Hablemos del tercer gráfico

- El Análisis Exploratorio de Datos (AED) es una etapa crítica en el proceso de Machine Learning.
- Su objetivo principal es comprender los datos que se utilizarán para entrenar un modelo.
- El AED nos ayuda a identificar patrones, detectar valores atípicos y tomar decisiones informadas.
- Durante esta etapa, exploramos la estructura, calidad y distribución de los datos.

Predicción de sueldos

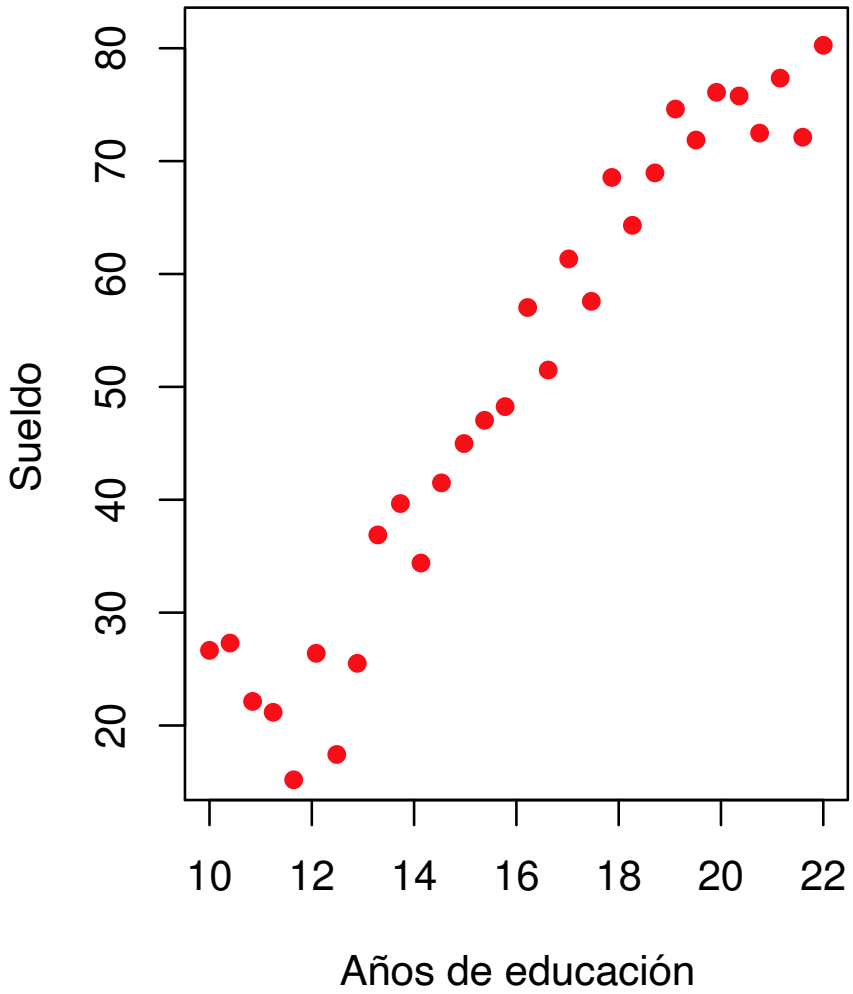
Análisis exploratorio de datos

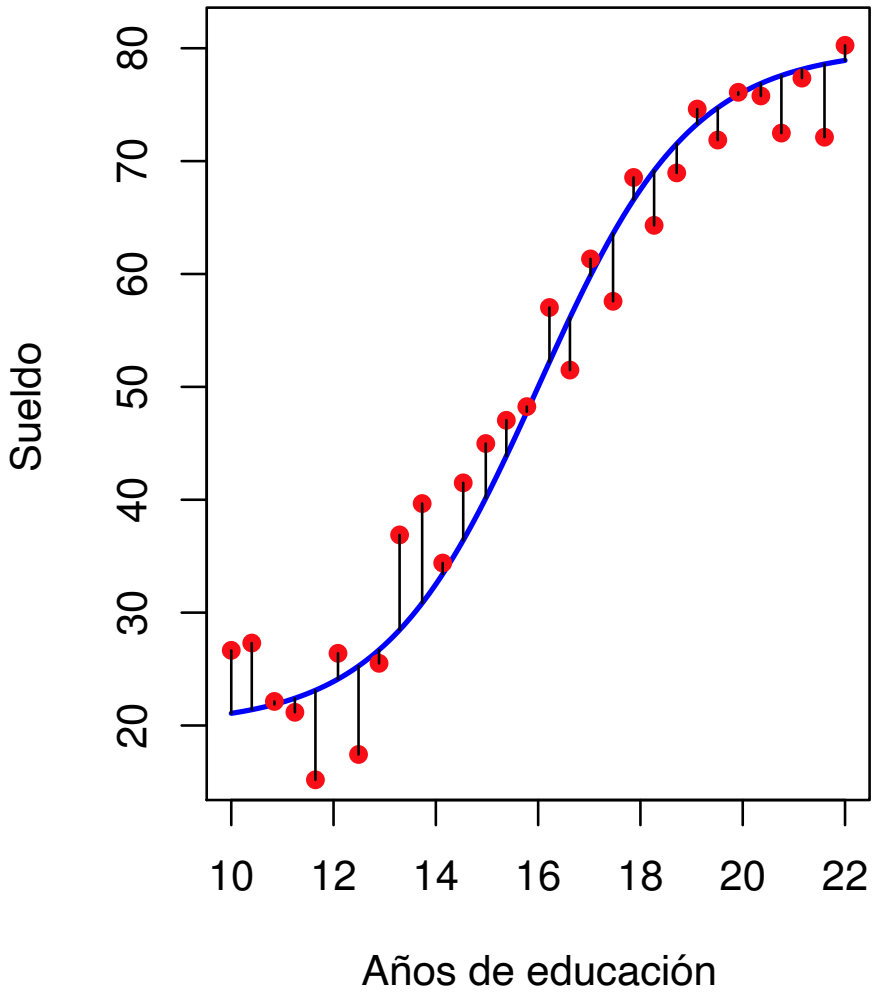
- Visualización de Datos: Utilizamos gráficos como histogramas, gráficos de dispersión y diagramas de caja para comprender la distribución y las relaciones entre las variables.
- Estadísticas Resumen: Calculamos estadísticas descriptivas como la media, la mediana y la desviación estándar para resumir las propiedades de los datos.
- Tratamiento de Datos Faltantes: Identificamos y manejamos valores faltantes o nulos en los datos.

- Detección de Valores Atípicos: Identificamos valores extremos que pueden afectar el rendimiento del modelo.
- Ingeniería de Características: Creamos nuevas características o transformamos las existentes para mejorar la representación de los datos.
- Análisis de Correlación: Evaluamos la correlación entre las variables para comprender las relaciones lineales.
- Análisis de Clusters: Aplicamos técnicas de clustering para identificar patrones y grupos en los datos.

Predicción de sueldos

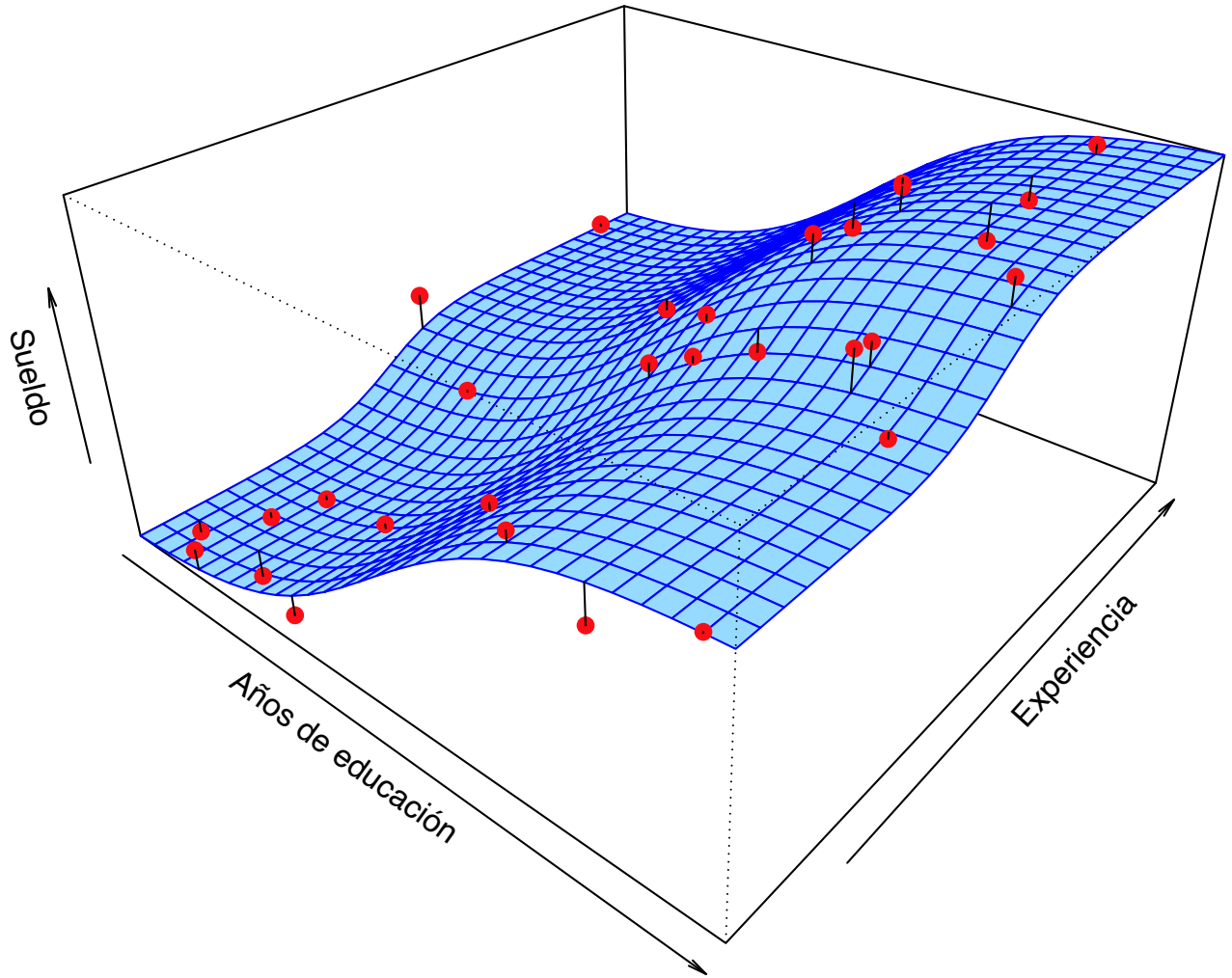
El modelo "verdadero"





Predicción de sueldos

Otro modelo ideal?



Predicción de sueldos

Qué haremos?

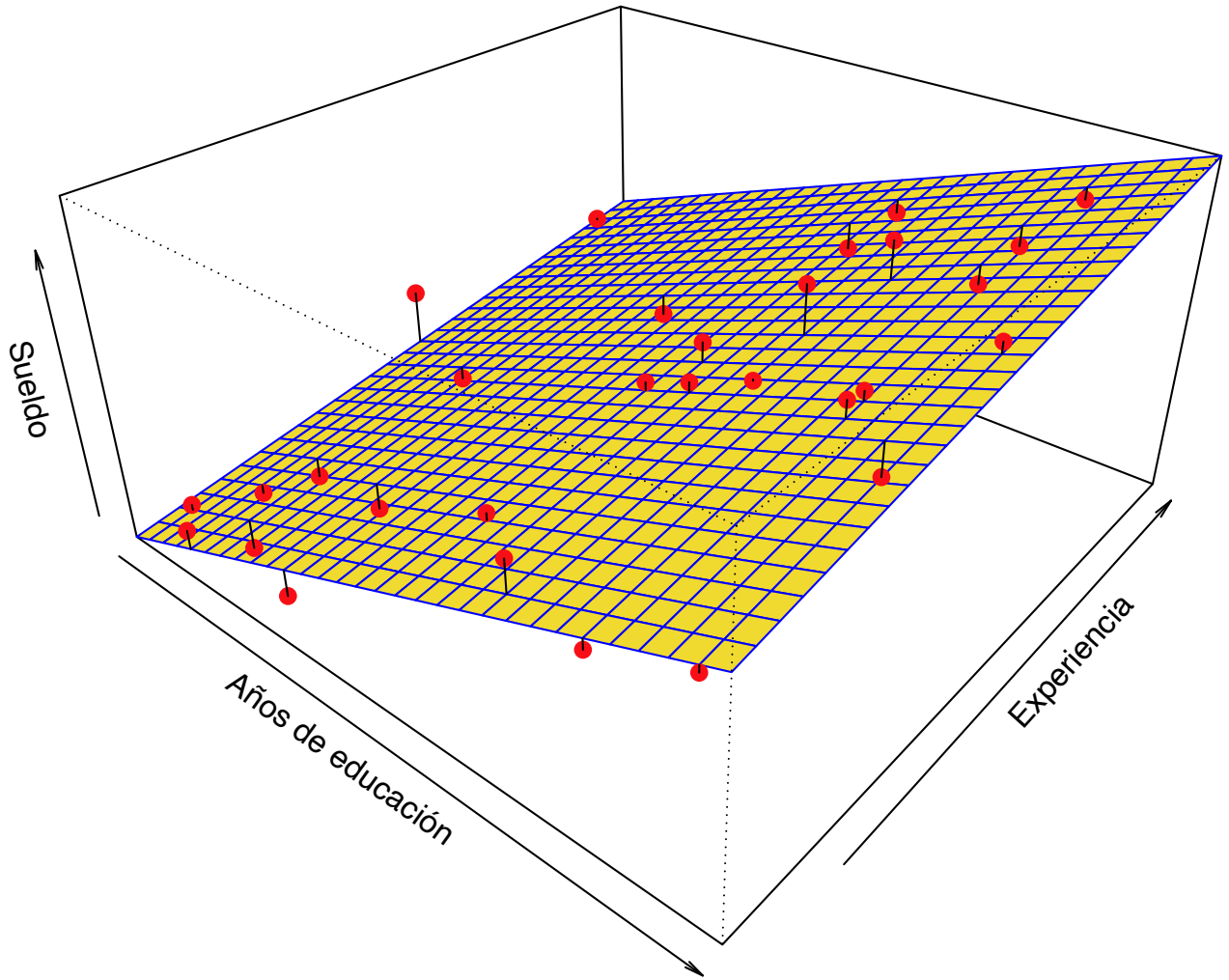
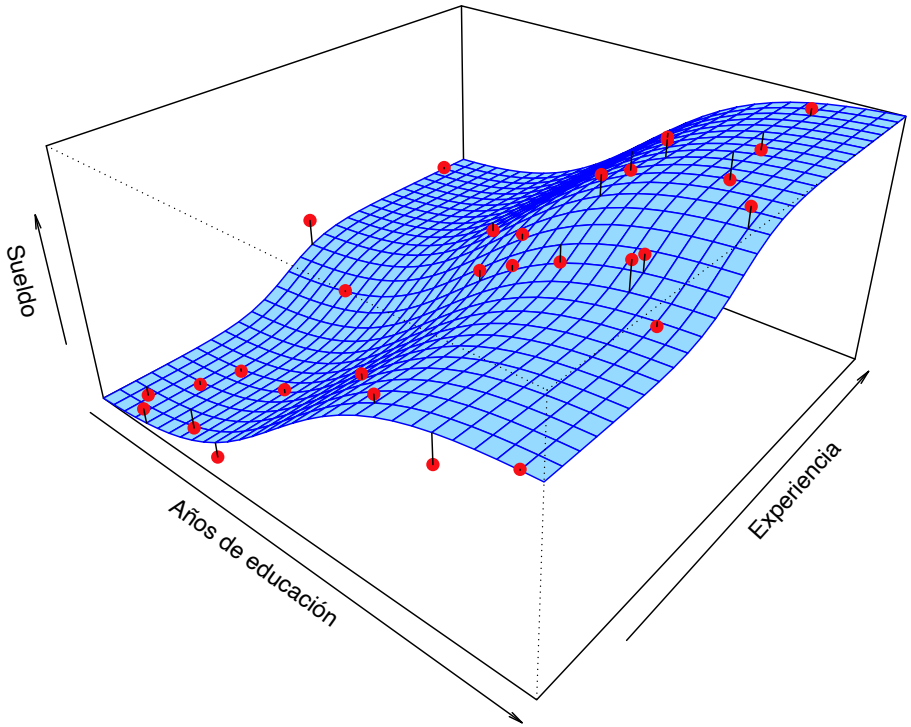
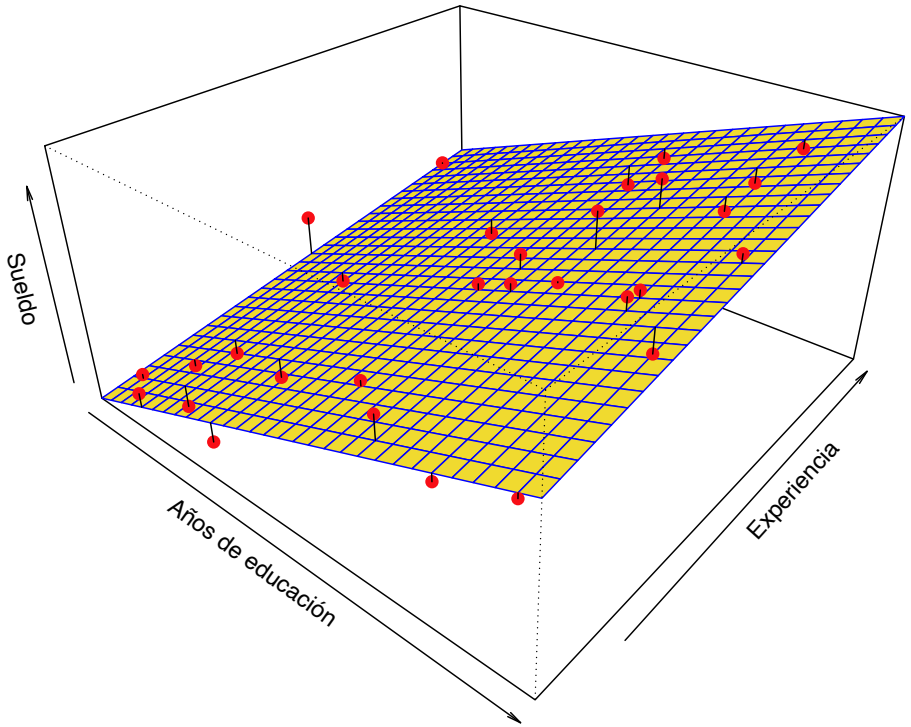


Figure: Figura

Predicción de sueldos

Comparando





Predicción de sueldos

Respuestas

	year	age	maritl	race	education	region	jobclass	health	health_ins	logwage	wage
0	2006	18	1. Never Married	1. White	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	2. No	4.318063	75.043154
1	2004	24	1. Never Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	4.255273	70.476020
2	2003	45	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	1. <=Good	1. Yes	4.875061	130.982177
3	2003	43	2. Married	3. Asian	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	5.041393	154.685293
4	2005	50	4. Divorced	1. White	2. HS Grad	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.318063	75.043154
...	...	...	...	...	...	...	...	...	...	...	...
2995	2008	44	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	5.041393	154.685293
2996	2007	30	2. Married	1. White	2. HS Grad	2. Middle Atlantic	1. Industrial	2. >=Very Good	2. No	4.602060	99.689464
2997	2005	27	2. Married	2. Black	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	2. No	4.193125	66.229408
2998	2005	27	1. Never Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	4.477121	87.981033
2999	2009	55	5. Separated	1. White	2. HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	1. Yes	4.505150	90.481913

Figure: Salarios y otros datos para un grupo de 3000 trabajadores varones en la región del Atlántico Medio. Disponible en ISLP.

1. Cuál es la entrada de nuestro modelo? Un subconjunto de las columnas disponibles. Cuál?, **no tengo ni idea**.
2. Cuál es la salida de nuestro modelo? Sueldo (o un intervalo de confianza del sueldo)
3. Cuáles son los mejores predictores de y? **no tengo ni idea**.
4. Existe relación entre los predictores? **no tengo ni idea**.

Regresión Lineal

La regresión lineal es un modelo estadístico que busca modelar la relación lineal entre una variable dependiente Y y una o más variables independientes X. En forma matricial, esto se representa como:

$$Y = X\beta + \varepsilon$$

donde:

- Y es el vector de variables dependientes.
- X es la matriz de variables independientes.
- β es el vector de coeficientes.
- ε es el vector de errores.

Mínimos cuadrados ordinarios (OLS)

Notación matricial

$$\min_{\beta} \|y - X\beta\|^2 \tag{5}$$

donde:

$\boldsymbol{X} \in \mathbb{R}^{?,?}$ Matriz de datos

$\boldsymbol{\beta} \in \mathbb{R}^{?,?}$ es el vector de coeficientes,

$y \in \mathbb{R}^{?,?}$ es el vector de observaciones$i$,

m es el número de predictores

n es el número de observaciones.

Tarea

Encontrar las dimensiones de X, β y y

Mínimos cuadrados ordinarios (OLS)

Notación vectorial

$$\min_{\beta} \sum_{i=1}^{?} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 \tag{6}$$

donde:

$\boldsymbol{x}_i \in \mathbb{R}^{?,?}$ es el vector de variables predictoras para la observación i,

$\boldsymbol{\beta} \in \mathbb{R}^{?,?}$ es el vector de coeficientes,

$y_i \in \mathbb{R}^{?,?}$ es el valor observado para la observación i,

m es el número de predictores

n es el número de observaciones.

Tarea

Encontrar las dimensiones de x_i, β y y_i

Ecuaciones normales

Para encontrar los coeficientes β que minimizan el error cuadrático, se pueden utilizar las ecuaciones normales:

$$X^T X \beta = X^T Y$$

La solución para β es:

$$\beta = (X^T X)^{-1} X^T Y$$

Tarea

1. Encontrar la expresión cerrada para 1 variable.
2. Intentar conseguir la solución general.
3. Investigar la definición de $X^T X$ para cualquier matrix X.