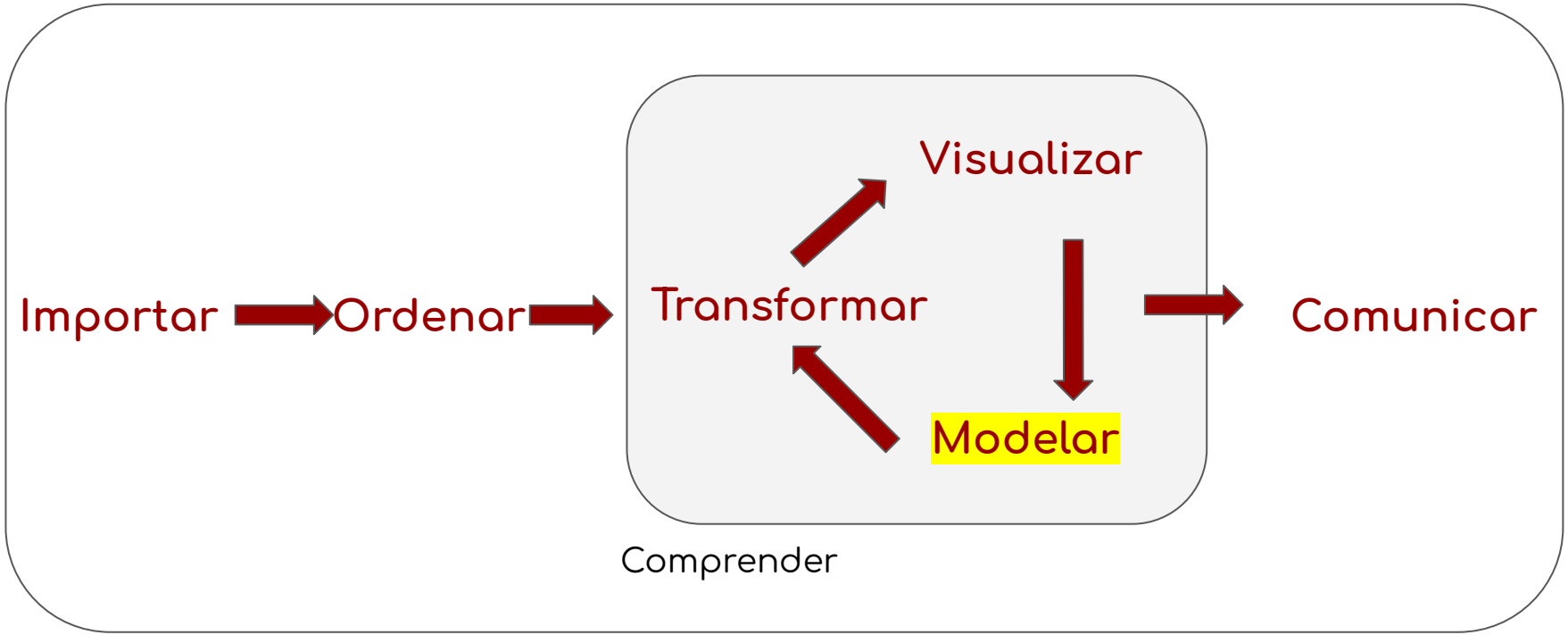
“R para Ciencia de Datos”

Construcción de Modelos

(Parte 1)

R-Ladies La Paz

MSc. Ing. Ruth Chirinos


Importar
Ordenar
Visualizar
Transformar
Modelar
Comunicar
Comprender
Programar

Tipos de Modelos

Predictivos

Descubrimiento de Datos

Supervisados
No Supervisados

¿Qué aprenderemos?

- Conceptos básicos de modelo
  - Modelos lineales
  - Aprenderás a interpretar ¿Qué es lo que el modelo dice de tus datos?
- Construcción de Modelos
  - Extraer patrones conocidos de tus datos.
  - Convertir el patrón en un modelo
- Muchos Modelos
  - Usar modelos simples para comprender datasets complejos.
  - Combinaremos herramientas de programación y modelado.

Generación de hipótesis vs. confirmación de hipótesis

El modelado usa la inferencia para validar que una hipótesis es verdadera

Análisis de datos EDA (Review)

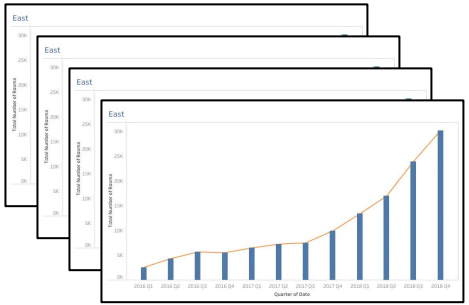
1. Genera preguntas acerca de tus datos.
2. Buscas respuestas visualizando, transformando y modelando tus datos.
3. Usas lo que has aprendido para refinar tus preguntas y/o generar nuevas interrogantes.

Modelado Confirmatorio

Ejecutar la inferencia correctamente

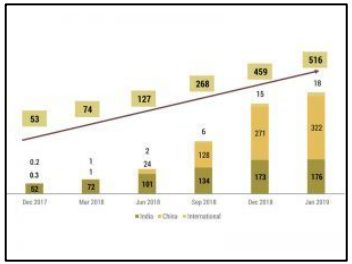
EXPLORACIÓN

1 **Observación**



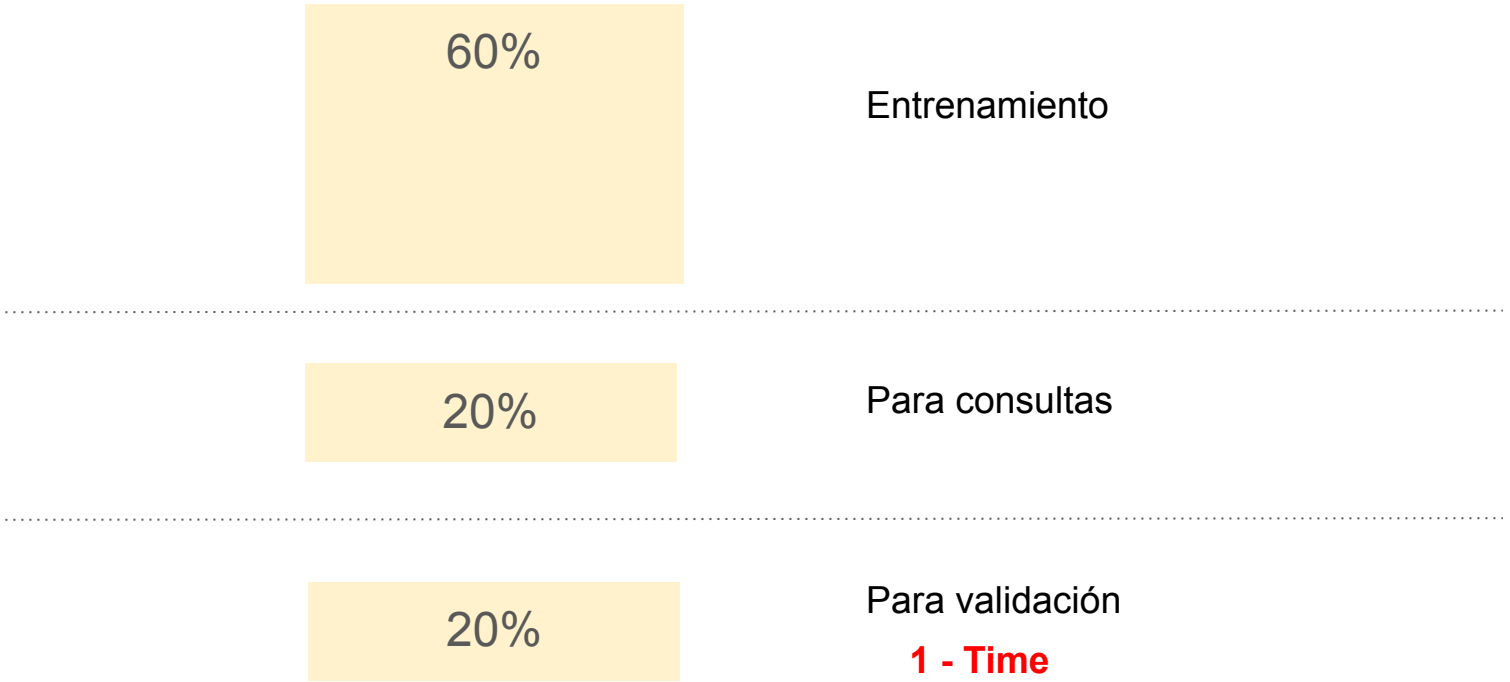
N - Times

2 **Confirmación**



1 - Time

Dividir tus datos en 3 partes



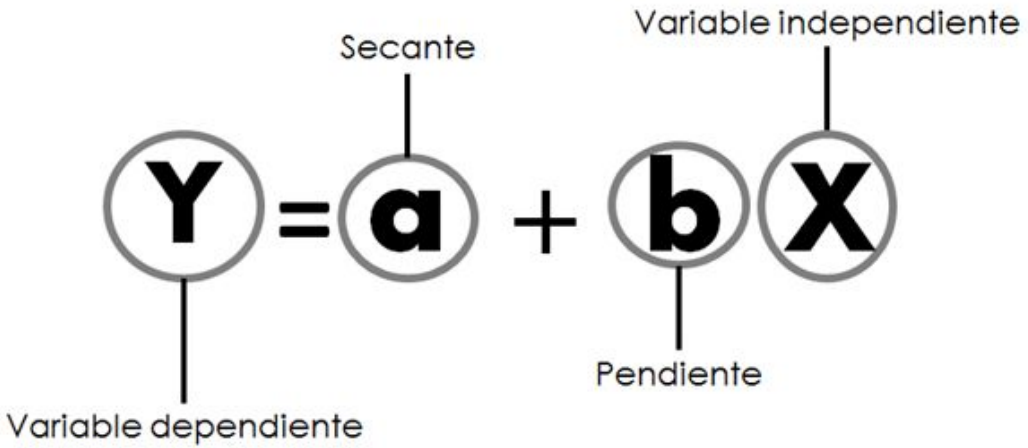
Conceptos básicos de Modelo

El objetivo de un modelo es:

Proveer un resumen de baja dimensión de un conjunto de datos

Partes de un Modelo

1. Elegir la familia de modelos, en este caso por ejemplo Regresión Lineal




```
Y =  1 +  5 * X
Y =  2 +  8 * X
Y =  5 +  7 * X
Y = 54 +  1 * X
Y =  6 + 77 * X
Y =  8 + 11 * X
...
...
...
...
```

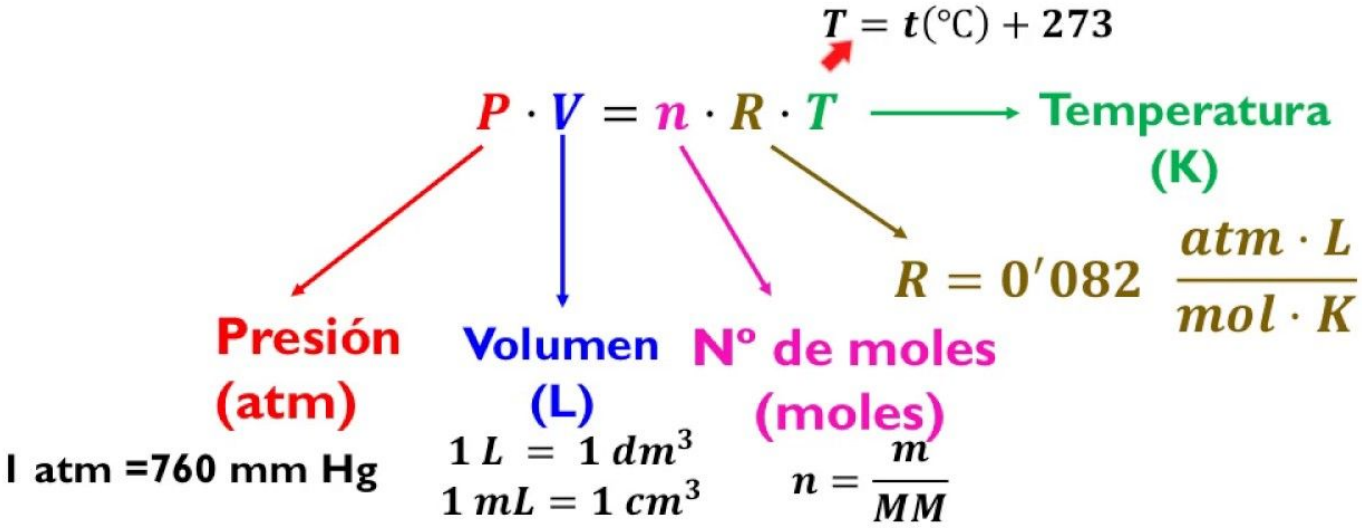
Partes de un Modelo (Cont.)

2. Generar un modelo ajustado que sea lo más cercano a tus datos.

$$Y = 5 + 7 * X$$

Aquí tú tienes el **mejor** modelo!!

Ecuación de Gases “Ideales”



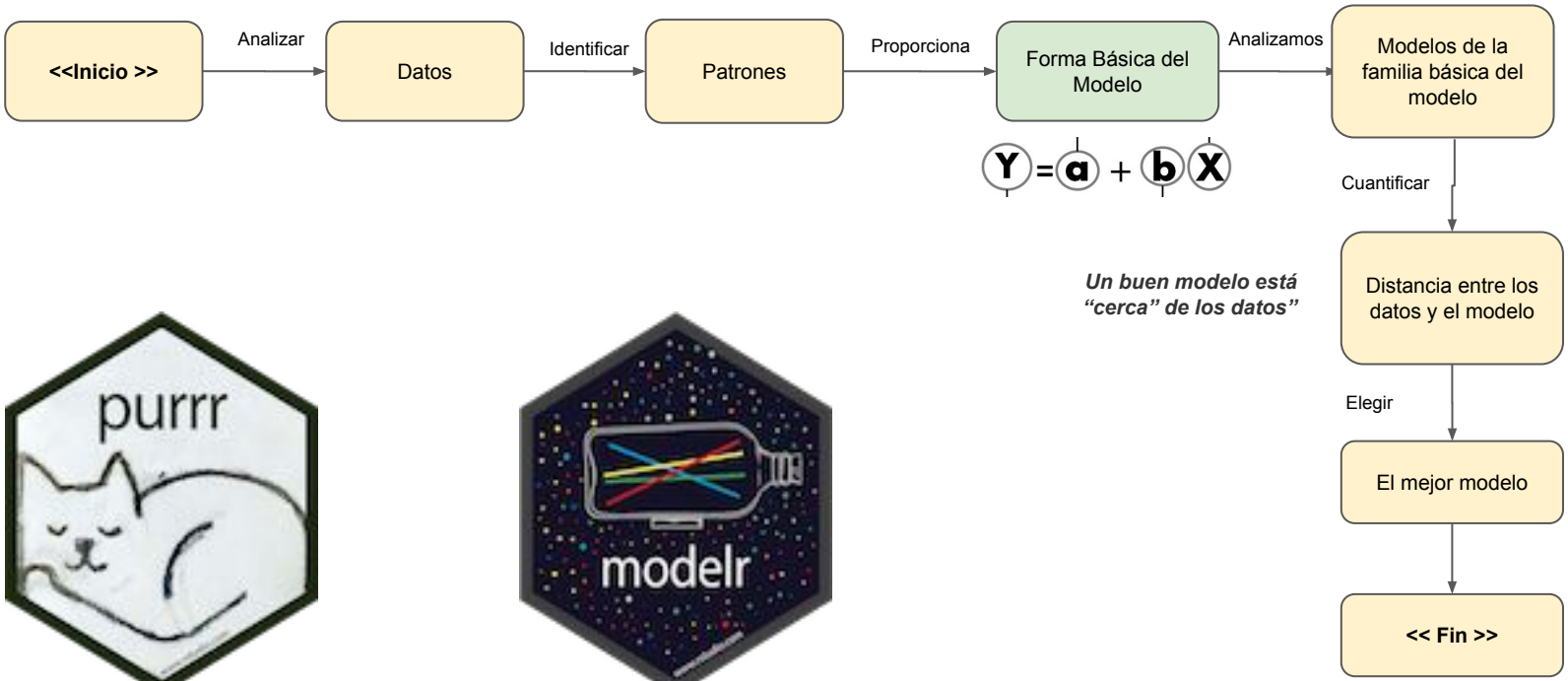
Y los gases reales?

Para tal modelo, no hay necesidad de preguntarse “¿Es el modelo verdadero?”. Si la “verdad” debe ser la “verdad completa”, la respuesta debe ser “No”. La única pregunta de interés es “¿Es el modelo esclarecedor y útil?”.

Un modelo Simple

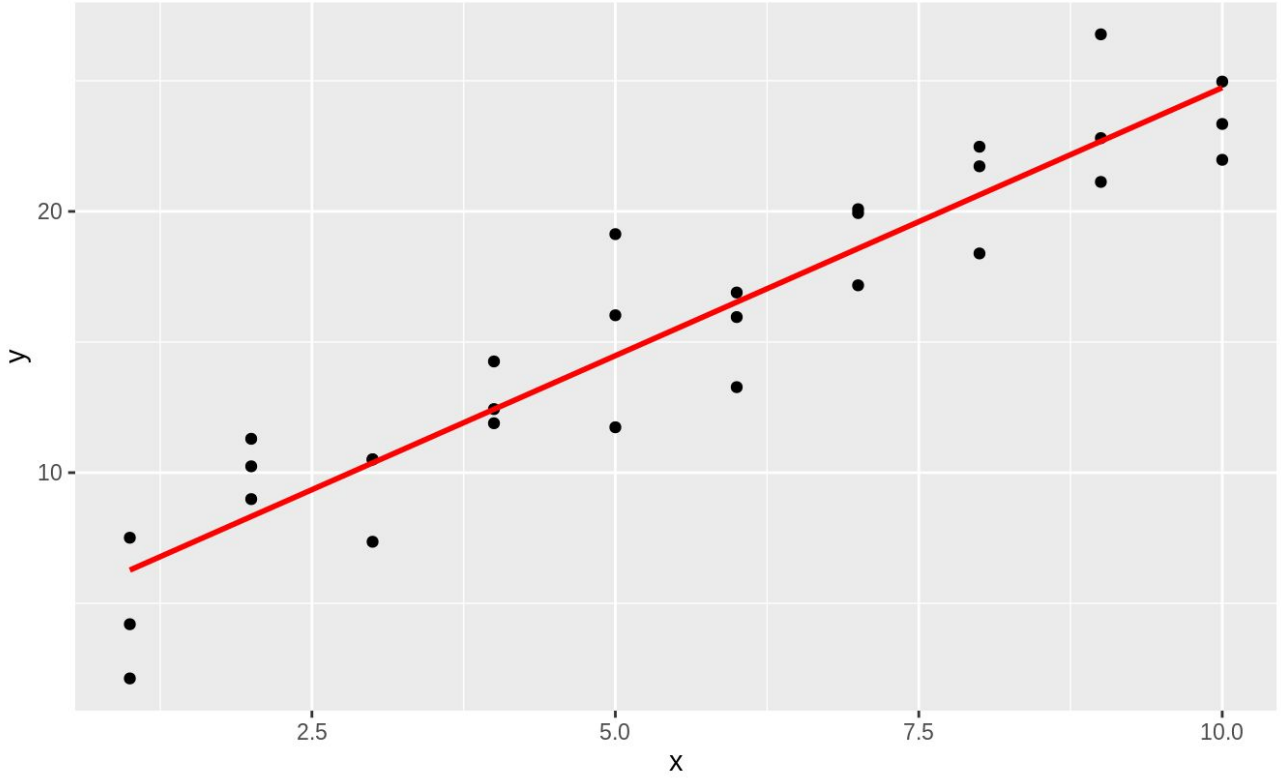
1. Básico. Entendiendo cómo es el proceso de construcción de un modelo
2. Utilizando optim()
3. Búsqueda en cuadrícula
4. Búsqueda de Newton - Raphson
5. Modelos lineales

Un modelo Simple - Construcción



Visualizando modelos

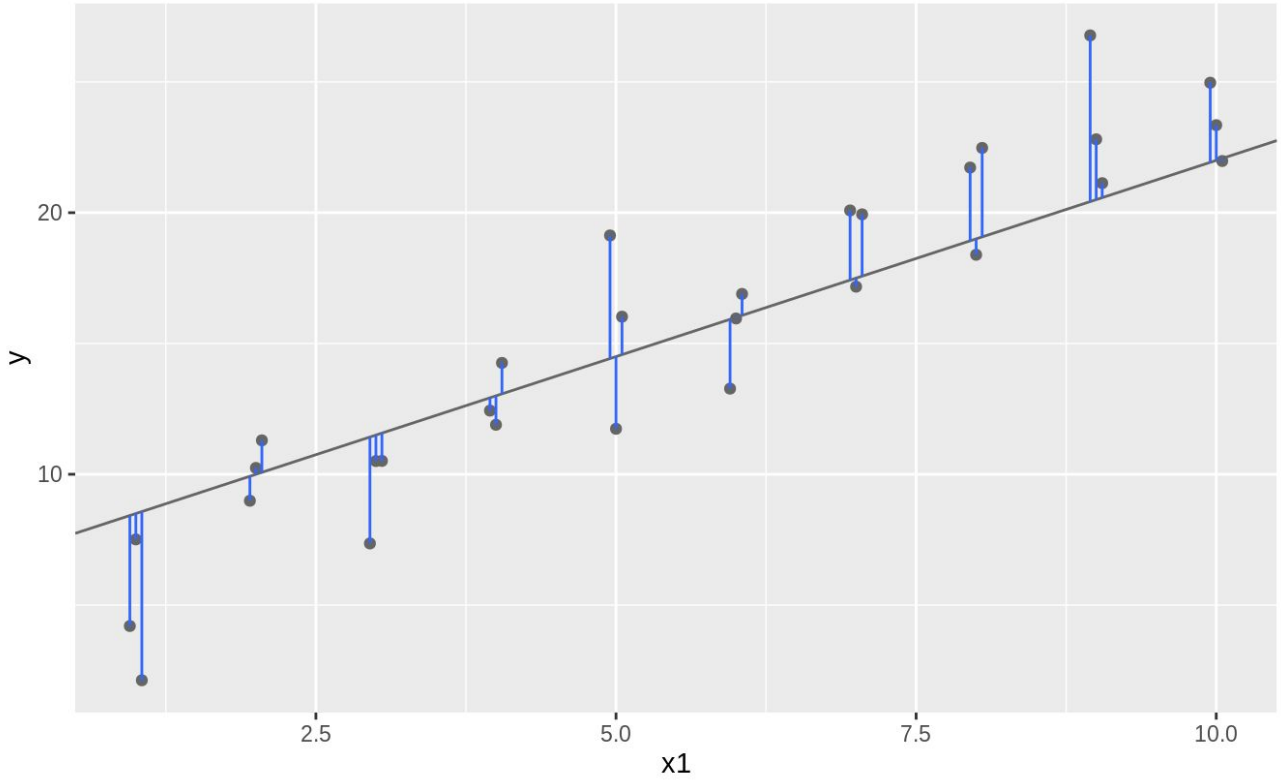
Predicciones



Las predicciones te informan de los patrones que el modelo captura

Visualizando modelos

Residuos

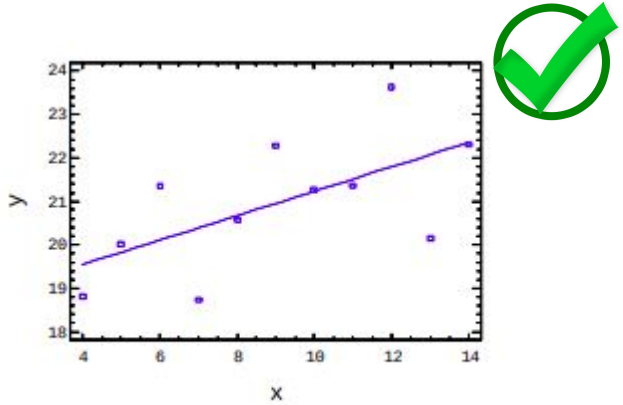


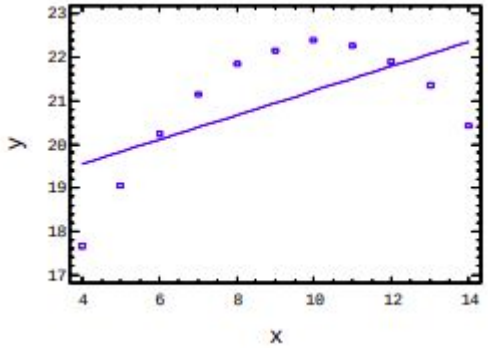
Distancia vertical entre cada punto y el modelo

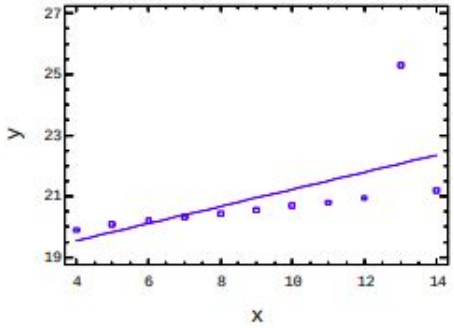
Los residuos te dicen lo que el modelo ignora

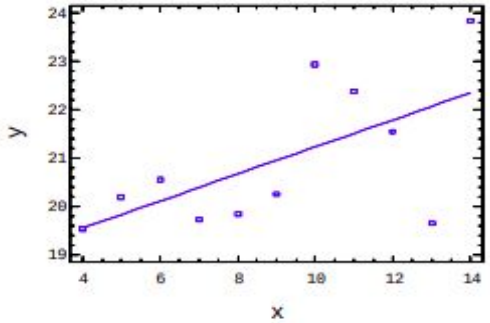
Visualizando modelos

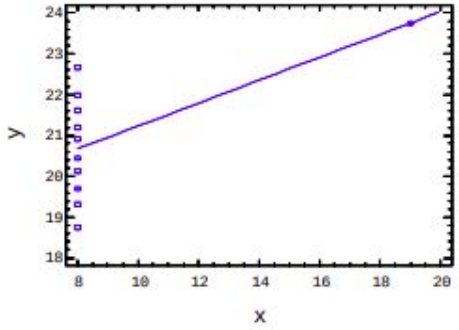
Residuos (Cont.)











Si el modelo es correcto, los residuos se aproximarían a los errores aleatorios **que hacen que la relación entre las variables X y la Y sea una relación estadística.** Si los residuos parecen comportarse de forma aleatoria, sugiere que el modelo se ajusta bien a los datos.

Fórmulas y Familias de Modelos

- La notación que se usa es “notación de Wilkinson-Rogers”
- Format Notation: https://faculty.chicagobooth.edu/richard.hahn/teaching/formulanotation.pdf

Fórmulas y Familias de Modelos

Variables Categóricas

```
Fórmula:

y ~ sexo
```

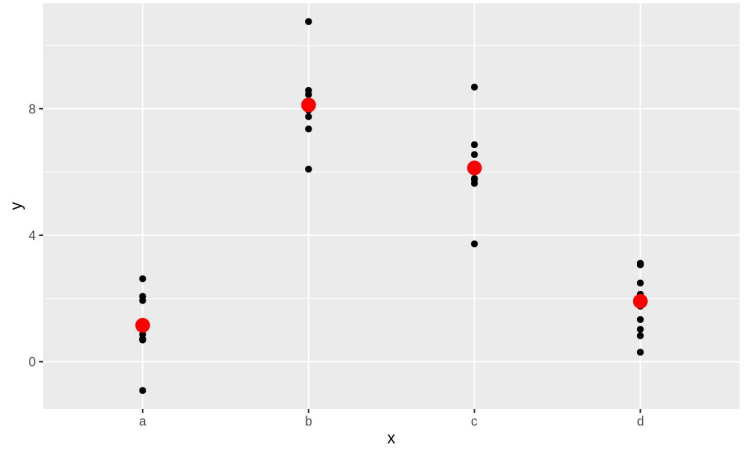
```
y = x_0 + x_1 * sexo
```

No se puede multiplicar porque sexo no es un número

```
y = x_0 + x_1 * sexo_hombre
```

Donde: sexo_hombre = 1 y 0 a sexo mujer

Un modelo con una variable `x` categórica va a predecir el valor medio para cada categoría.



Fórmulas y Familias de Modelos

Interacciones Continuas y Categóricas

```
mod1 <- lm(y ~ x1 + x2, data = sim3)
mod2 <- lm(y ~ x1 * x2, data = sim3)
```

- Tenemos dos predictores, por lo que necesitamos pasar ambas variables a `data_grid()`
- Generamos predicciones de ambos modelos simultáneamente, podemos usar `gather_predictions()` o `spread_predictions()`.

Fórmulas y Familias de Modelos

Interacciones Dos Variables continuas

```
mod1 <- lm(y ~ x1 + x2, data = sim3)
mod2 <- lm(y ~ x1 * x2, data = sim3)
```

- Tenemos dos predictores, por lo que necesitamos pasar ambas variables a `data_grid()`
- Generamos predicciones de ambos modelos simultáneamente, podemos usar `gather_predictions()` o `spread_predictions()`.

Referencias

- http://www.sthda.com/english/wiki/ggplot2-add-straight-lines-to-a-plot-horizontal-vertical-and-regression-lines
- http://www.learnbymarketing.com/tutorials/linear-regression-in-r/
- https://faculty.chicagobooth.edu/richard.hahn/teaching/formulanotation.pdf

Gracias!