Regresión lineal simple.

Mínimos cuadrados ordinarios.

Dr. Martín Lozano `https://mlozanoqf.github.io/`

20 de diciembre de 2025, 10:39 p.m.

	Fundamental	Intermedio	Especializado
Finanzas	×	✓	×
Estadística	×	✓	×
R	×	✓	×

1 Introducción.

- Se descargan precios históricos de 10 acciones y se calculan retornos mensuales. Con esos retornos se estiman medias, volatilidades y razón de Sharpe por activo.
- Se construye la frontera eficiente media–varianza y se comparan tres carteras. Finalmente, se evalúa la robustez de riesgo y rendimiento con un block bootstrap de 6 meses (500 réplicas).

2 Caso determinístico.

Planteamiento del problema.

Vamos a empezar con un ejemplo **determinístico** (sin incertidumbre): **distancia, velocidad y tiempo**.

- En física básica, si la **velocidad es constante** v, entonces la distancia recorrida d está **fijada** por:

$$d = v \cdot t$$

- Si conocemos v y t, entonces d queda **completamente determinada** (no hay "ruido", no hay error).
- En este caso, hablar de "estimación" no tiene sentido:
  no estamos infiriendo una relación a partir de datos con variabilidad, solo estamos re-expresando una identidad.

3 Datos de ejemplo

Supón que un objeto se mueve a velocidad constante:

$$v = 60 \text{ km/h}$$

y medimos distintos tiempos t. La distancia d resultante queda fijada por $d = 60t$.

Obs	t (h)	d (km)
1	0	0
2	1	60
3	2	120
4	3	180
5	4	240
6	5	300

4 Visualización (los puntos caen exactamente en una recta)

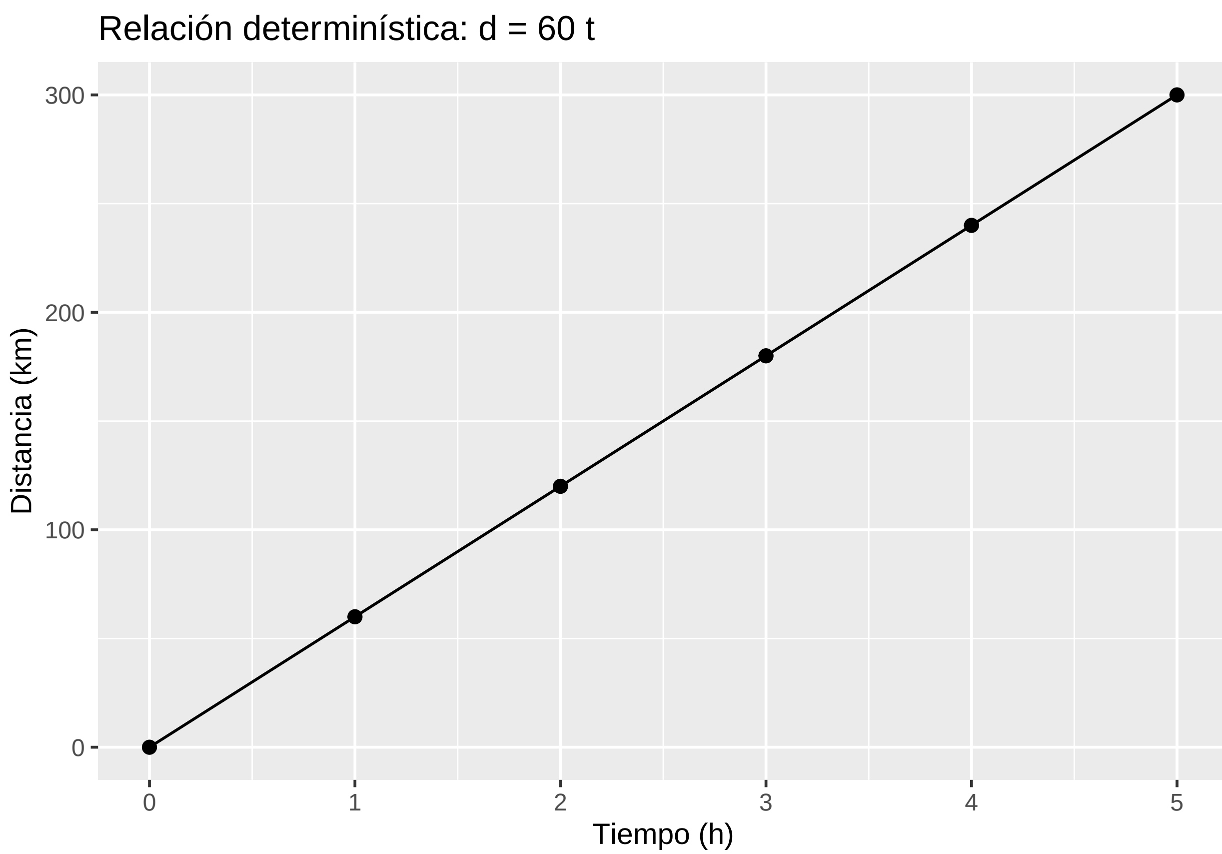
```r
library(ggplot2)

# Datos determinísticos: d = 60 * t
tabla <- data.frame(
  t = 0:5,
  d = 60 * (0:5)
)

tabla
```

```
##   t   d
## 1 0   0
## 2 1  60
## 3 2 120
## 4 3 180
## 5 4 240
## 6 5 300
```

```r
# Gráfico con ggplot: puntos + línea
ggplot(tabla, aes(x = t, y = d)) +
  geom_point(size = 2) +
  geom_line() +
  labs(
    title = "Relación determinística: d = 60 t",
    x = "Tiempo (h)",
    y = "Distancia (km)"
  )
```



5 Parte 1 — Caso determinístico vs incertidumbre (un solo grupo)

5.1 Bloque 1.3 (nuevo) — Estimar una velocidad promedio con datos reales (un solo grupo)

Escenario (natural):

- Un equipo de entrenamiento universitario quiere tener una referencia simple: la **velocidad promedio en sprints** del equipo.
- No tienen un velocímetro confiable; lo que sí registran en cada sprint es:
  - t: tiempo (segundos),
  - d: distancia estimada (metros) medida con marcadores/GPS.
- Esos registros son imperfectos: hay error de medición, aceleración, fatiga, etc.

El objetivo es estimar una velocidad promedio v usando datos observados (t_i, d_i).

Una forma razonable de escribirlo es:

$$d_i = \beta_0 + \beta_1 t_i + u_i$$

donde: - β_1 se interpreta como **velocidad promedio** (m/s), - β_0 permite un sesgo sistemático (por ejemplo, error de inicio/fin), - u_i es variación no explicada.

5.1.1 Datos observados (tabla)

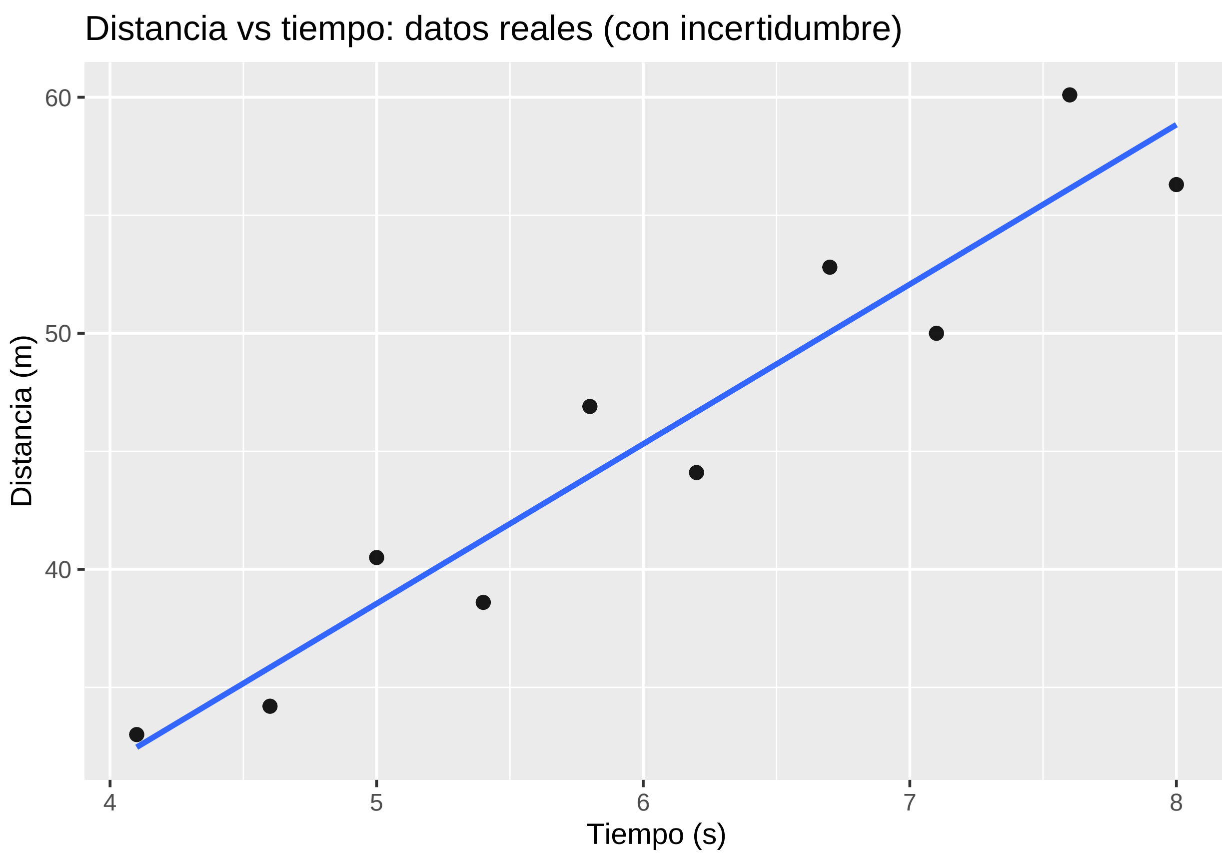
Obs	t (s)	d (m)
1	4.1	33.0
2	4.6	34.2
3	5.0	40.5
4	5.4	38.6
5	5.8	46.9
6	6.2	44.1
7	6.7	52.8
8	7.1	50.0
9	7.6	60.1
10	8.0	56.3

5.1.2 Gráfico con ggplot (puntos + recta OLS)

```r
library(ggplot2)

tabla_sprint <- data.frame(
  t = c(4.1, 4.6, 5.0, 5.4, 5.8, 6.2, 6.7, 7.1, 7.6, 8.0),
  d = c(33.0, 34.2, 40.5, 38.6, 46.9, 44.1, 52.8, 50.0, 60.1, 56.3)
)

ggplot(tabla_sprint, aes(x = t, y = d)) +
  geom_point(size = 2, alpha = 0.9) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(
    title = "Distancia vs tiempo: datos reales (con incertidumbre)",
    x = "Tiempo (s)",
    y = "Distancia (m)"
  )
```




```r
m1 <- lm(d ~ t, data = tabla_sprint)
summary(m1)
```

```
##
## Call:
## lm(formula = d ~ t, data = tabla_sprint)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7507 -2.5575 -0.5537  2.5534  3.9681
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.7366     4.5561   1.040    0.329
## t             6.7625     0.7379   9.165 1.62e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.881 on 8 degrees of freedom
## Multiple R-squared:  0.913,  Adjusted R-squared:  0.9022
## F-statistic:    84 on 1 and 8 DF,  p-value: 1.621e-05
```

```r
b <- coef(m1)

beta0_hat <- b["(Intercept)"]
v_hat     <- b["t"]  # m/s

c(beta0_hat = beta0_hat, v_hat_mps = v_hat)
```

```
## beta0_hat.(Intercept)           v_hat_mps.t
##              4.736602              6.762545
```

```r
v_hat * 3.6
```

```
##        t
## 24.34516
```

6 Qué significa (y qué NO significa) decir que "el tiempo causa la distancia"

Con el modelo:

$$d_i = \beta_0 + \beta_1 t_i + u_i$$

es tentador decir: *"como t está a la derecha, el tiempo causa la distancia"*. Esa es una **mala interpretación típica**.

7 Qué sí es cierto (nivel físico / definicional)

En un sentido **físico**, en un movimiento hay una relación mecánica:

- la distancia recorrida cambia con el tiempo,
- pero la distancia no "aparece" porque el tiempo la empuje como un tratamiento.

En realidad, la relación subyacente es algo como:

$$d(t) = \int_0^t v(s)\,ds$$

Si la velocidad fuera constante v, entonces:

$$d = vt$$

Eso es una **identidad del sistema**, no un resultado "descubierto" por la regresión.

8 Qué está haciendo realmente la regresión aquí

Cuando ajustamos:

$$d_i = \beta_0 + \beta_1 t_i + u_i,$$

estamos diciendo:

- "con mediciones imperfectas, quiero una recta que resuma la tendencia promedio de d cuando t cambia".

MCO no está probando causalidad; está estimando una relación promedio que **describe** estos datos.

9 La confusión típica: "variable en X = causa"

Error típico: "Como t está en X, entonces t causa d."

Por qué es un error (en estadística aplicada):

- En regresión, poner una variable en X **no crea causalidad**.
- La causalidad requiere un diseño o supuestos adicionales (por ejemplo, intervención, aleatorización, controles adecuados, etc.).
- Con datos observacionales, muchas cosas pueden generar asociación sin causalidad directa.

10 ¿Entonces cuál sería una lectura correcta?

Una lectura correcta del coeficiente $\hat{\beta}_1$ en este contexto es:

- $\hat{\beta}_1$ aproxima el **cambio promedio esperado en distancia** (metros) por cada segundo adicional de tiempo, *en el rango de tiempos observado*.
- Si el modelo es razonable, $\hat{\beta}_1$ se interpreta como una estimación de **velocidad promedio**.

Es decir:

$$\hat{\beta}_1 \approx \text{velocidad promedio (m/s)}$$

11 Mensaje para decir en cámara (anti-malentendido)

- "La regresión no demuestra que t 'cause' d."
- "Lo que hace aquí es estimar una velocidad promedio a partir de mediciones con error."
- "La causalidad en regresión no viene por la posición izquierda/derecha; viene por el diseño y los supuestos."

12 Conclusión.

- Los rendimientos acumulados permiten comparar desempeño histórico más allá de un solo punto riesgo–retorno.
- El block bootstrap muestra qué tan estables son las carteras ante cambios plausibles del periodo.