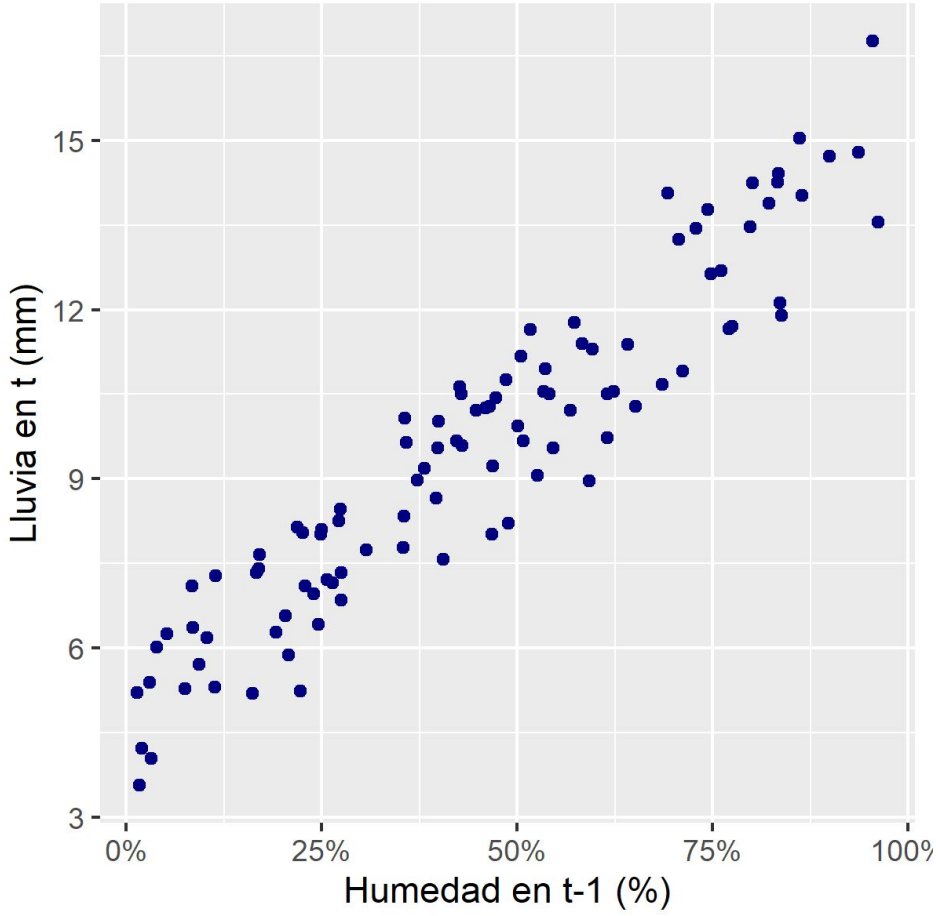
ITBA
Instituto Tecnológicc
de Buenos Aires

07/AGOSTO

APRENDIZAJE SUPERVISADO



Variable **target/respuesta (*y*)**

Variables **atributos/features/covariables (*X*)**

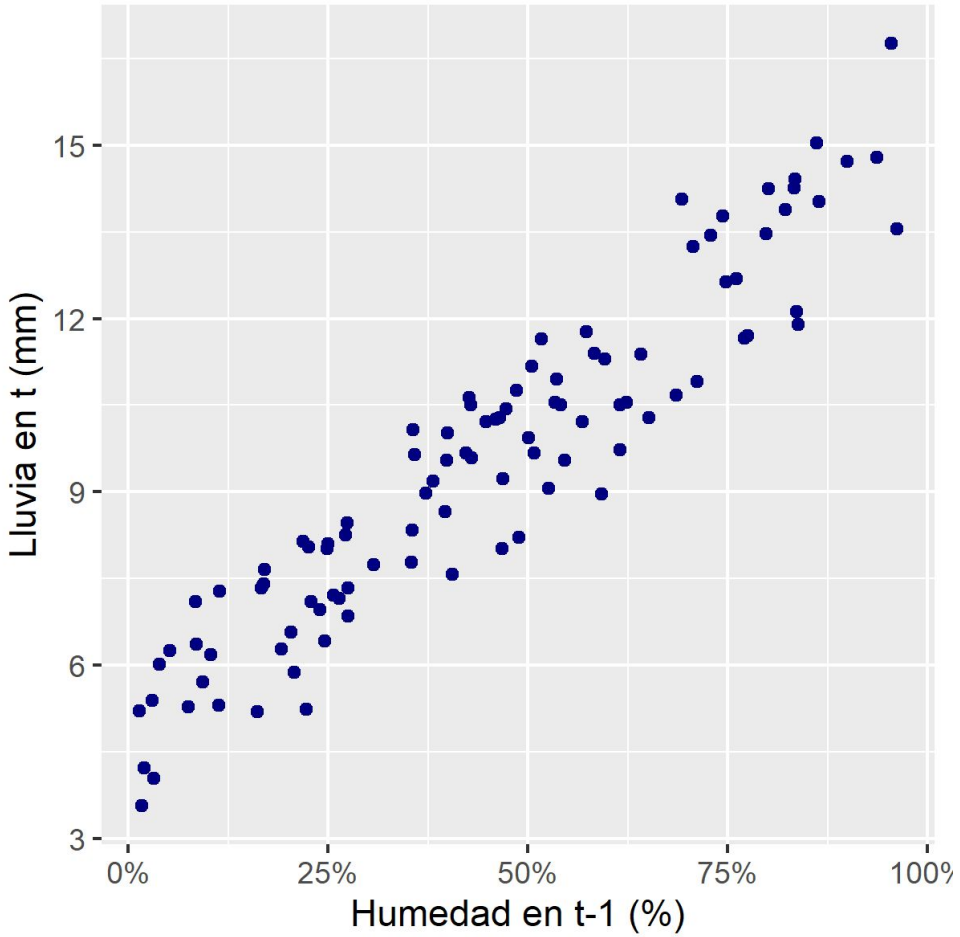
x1	x2	...	xp	y
0.127	0.159	...	0.675	0
0.706	0.073	...	0.681	1
0.541	0.118	...	0.864	0
...	...	...	...	...
0.114	0.449	...	0.484	0

n **observaciones/ejemplos/instancias**

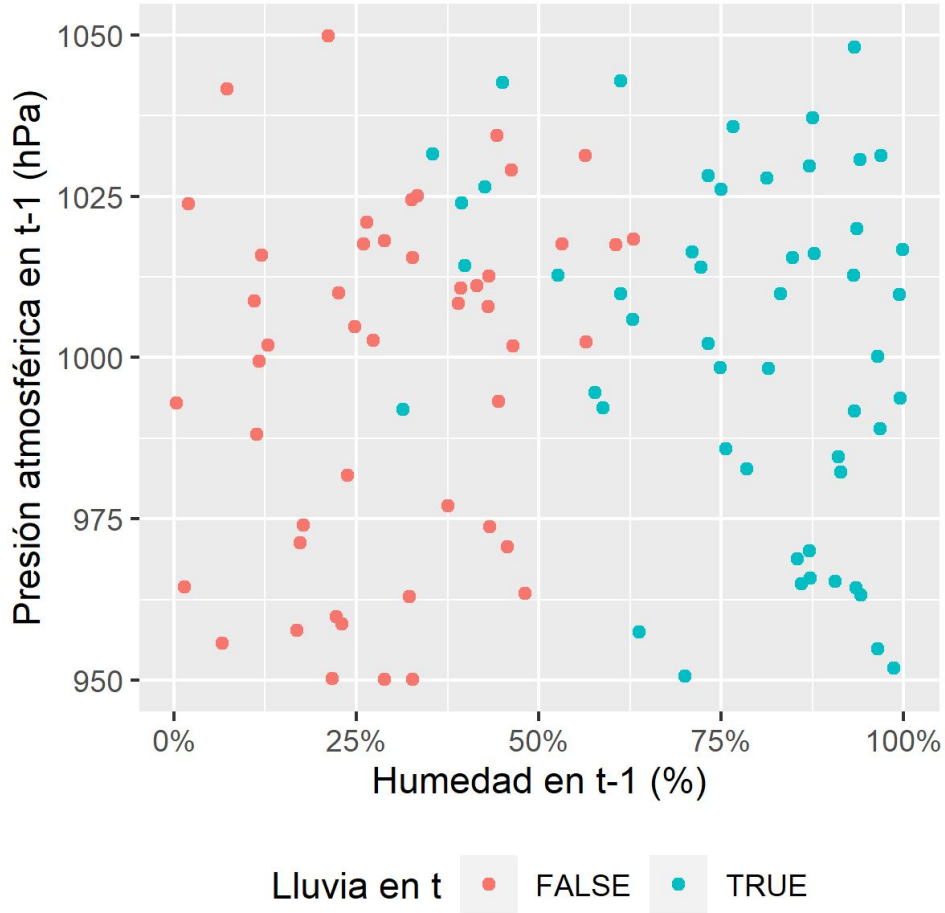
Objetivo:

estimar una función que relacione X con Y que sirva para estimar Y

Regresión
(target cuantitativo)

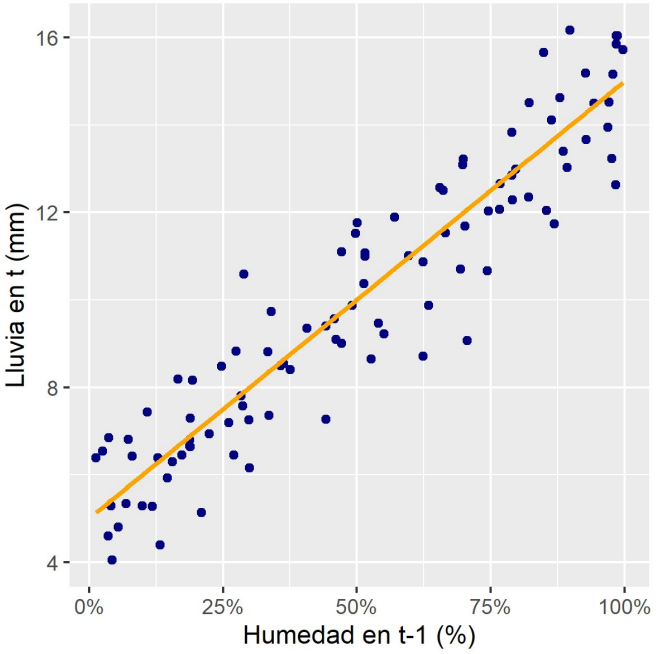


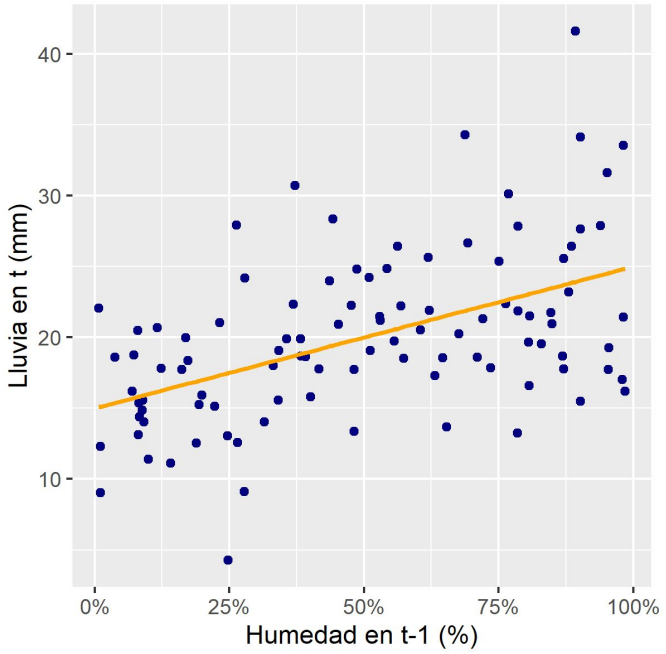
Clasificación
(target categórico)

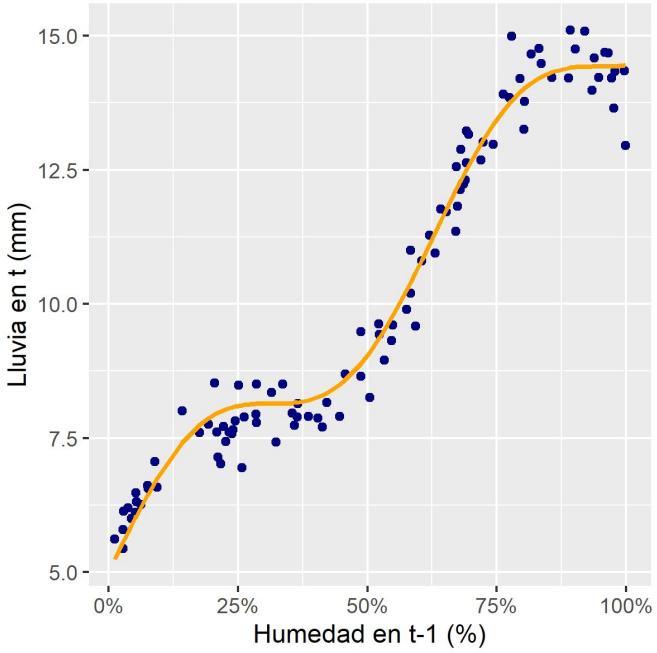


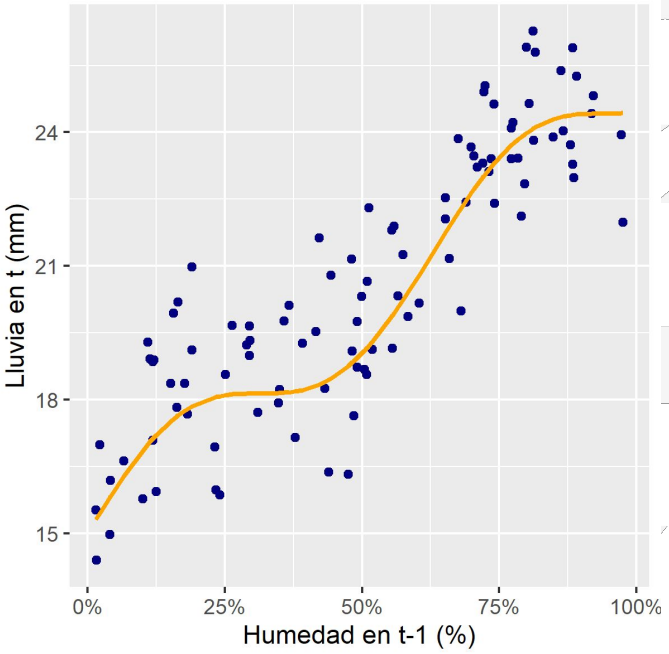
Asumimos que hay un ***Proceso Generador de Datos (DGP)***
(proceso con alguna estructura que genera Y según los valores de X)

$$Y = f(X) + \epsilon$$





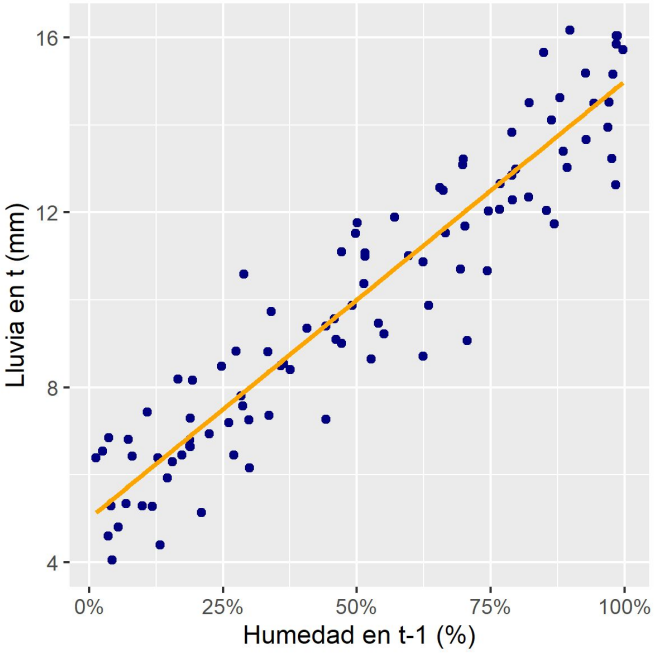


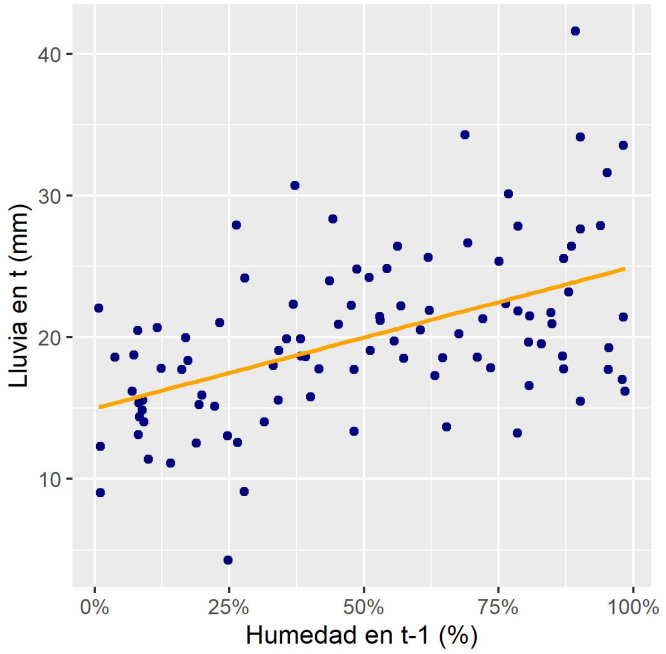


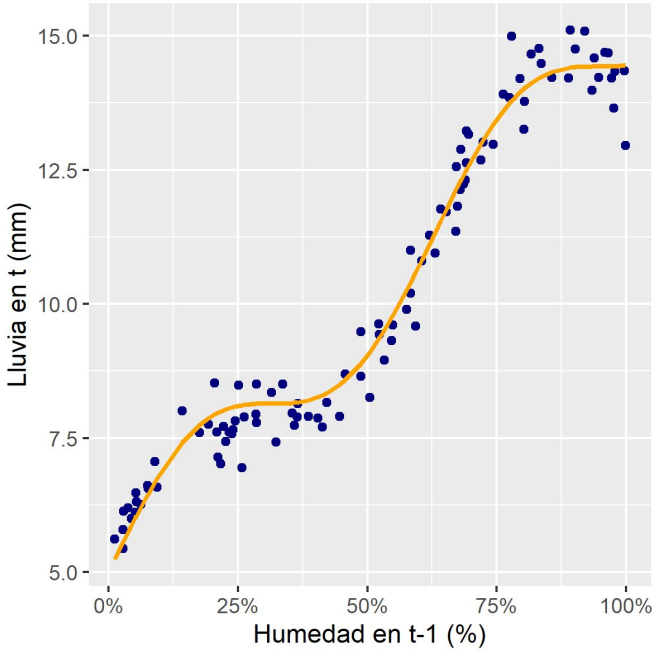
f(X) componente sistemático

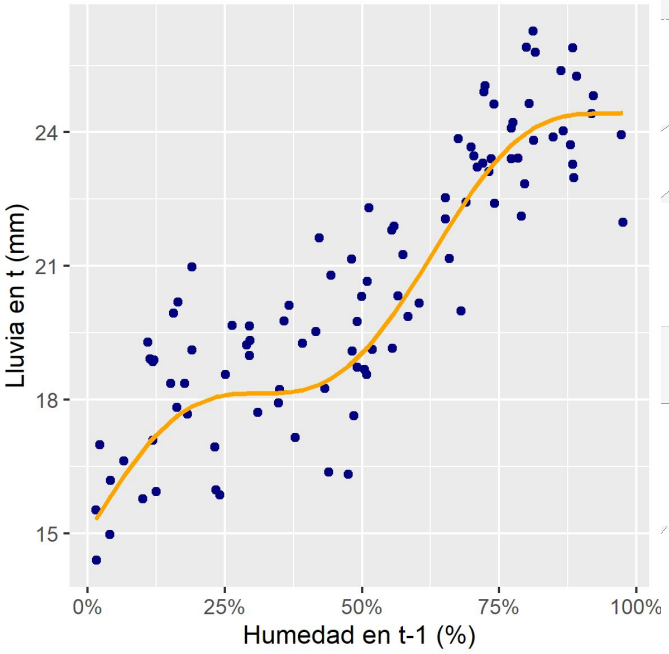
ϵ error aleatorio independiente de X

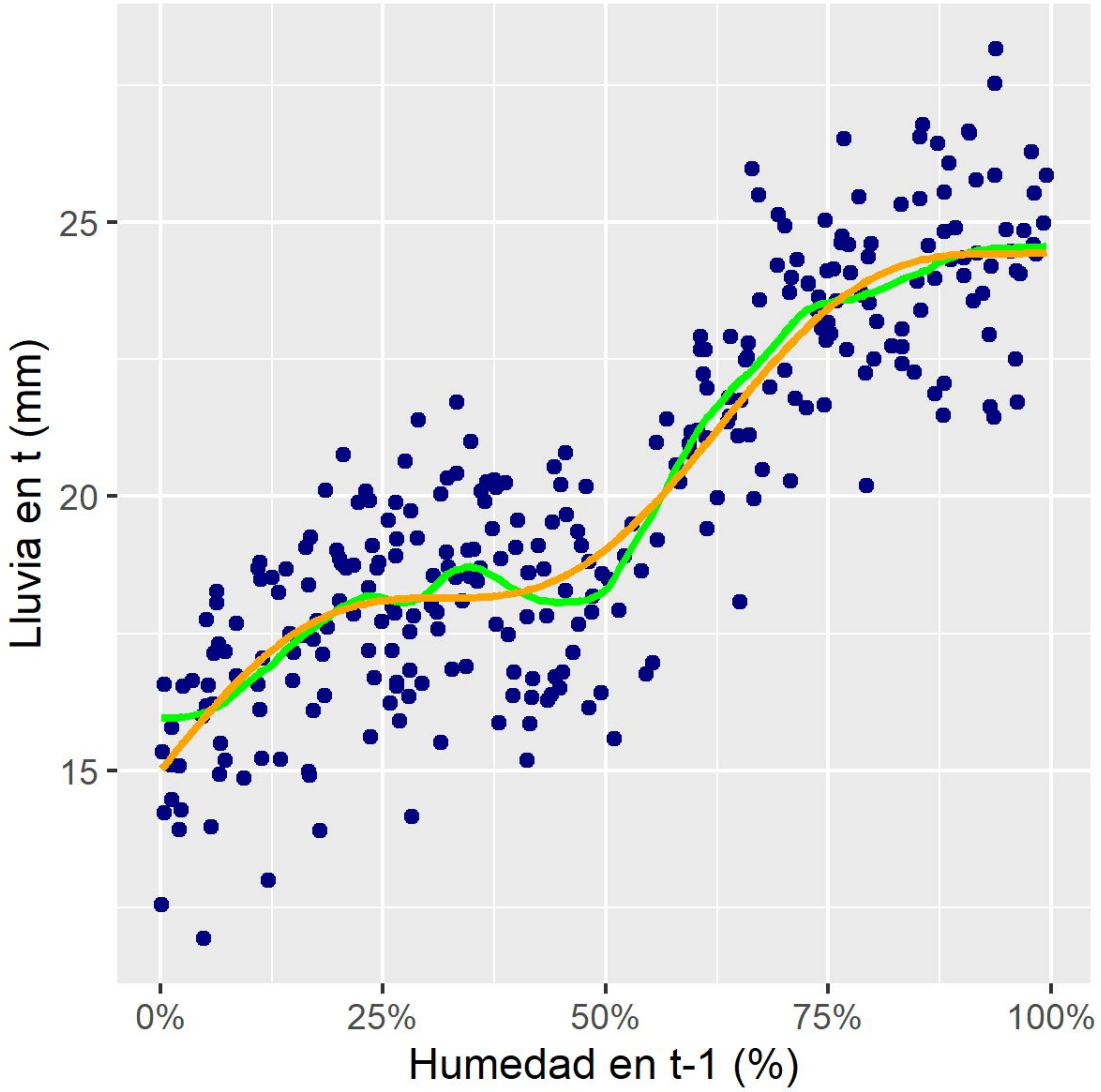
$$Y = f(X) + \epsilon$$











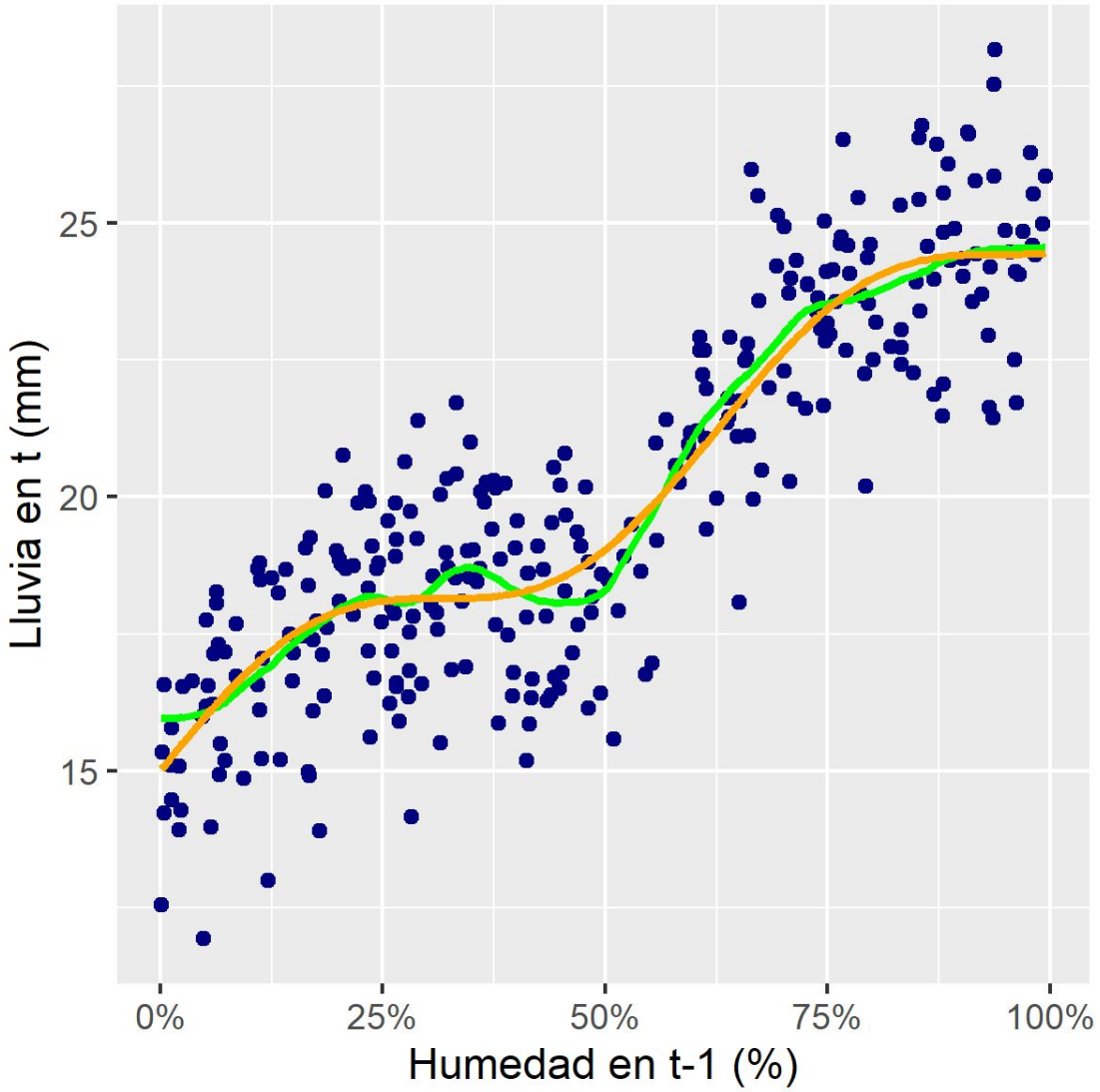
Aprendizaje supervisado

Aprender una función que se parezca a f(X), donde *Y* supervisa el proceso de aprendizaje

$$\hat{Y} = \hat{f}(X)$$

Predecir

Explicar



$$\hat{Y} = \hat{f}(X)$$

Predecir

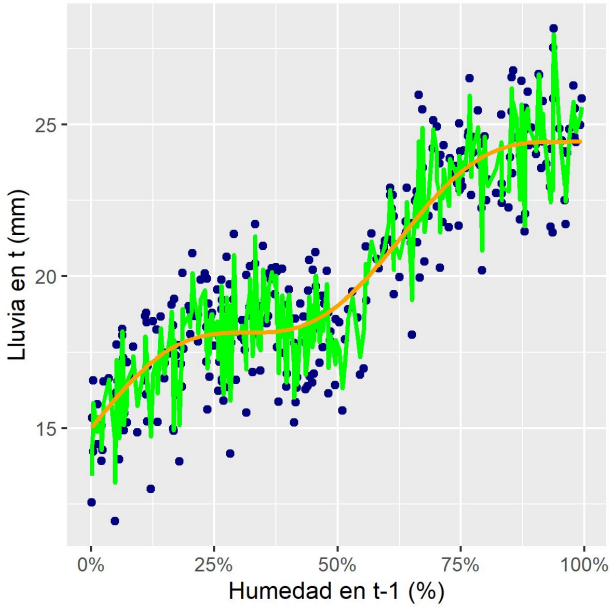
Estimaciones $\hat{Y}$ que se aproximen bien a Y en datos nuevos

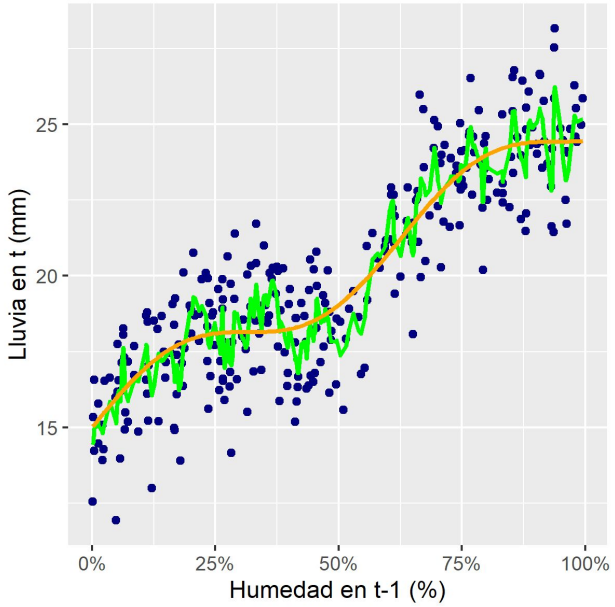
Explicar

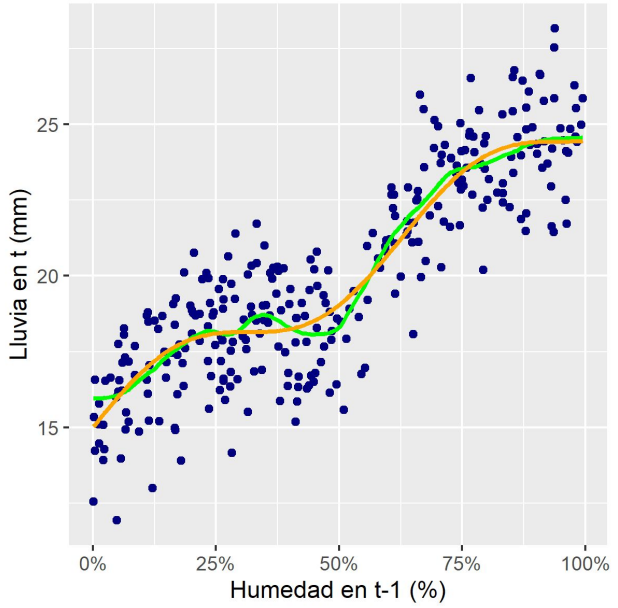
¿Qué variables son relevantes para explicar Y?
¿Qué forma tiene la relación entre X y Y?

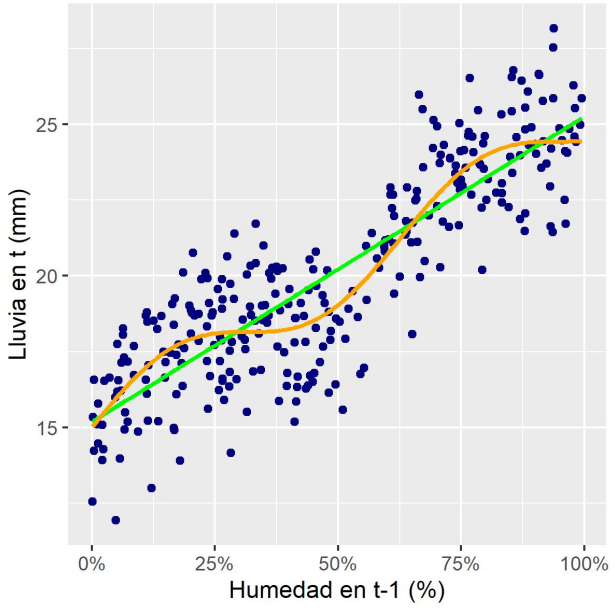
$$\hat{Y} = \hat{f}(X)$$

¿Cuál es un buen modelo?
¿Cómo podemos medir la bondad?



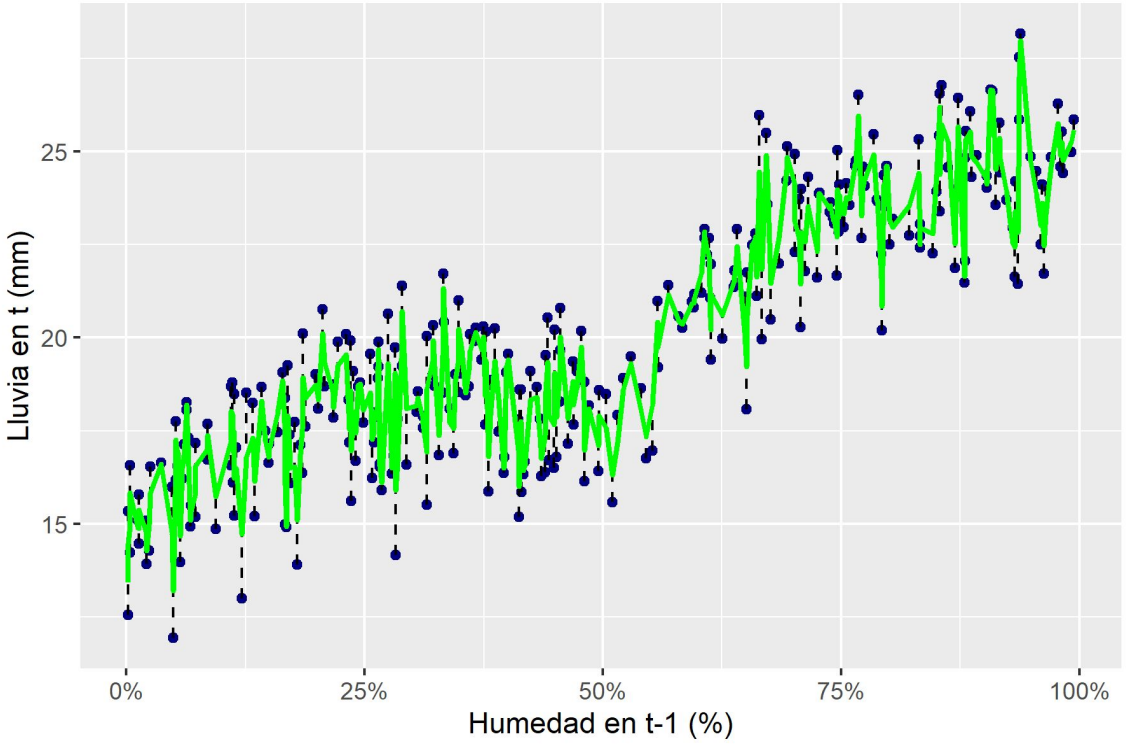


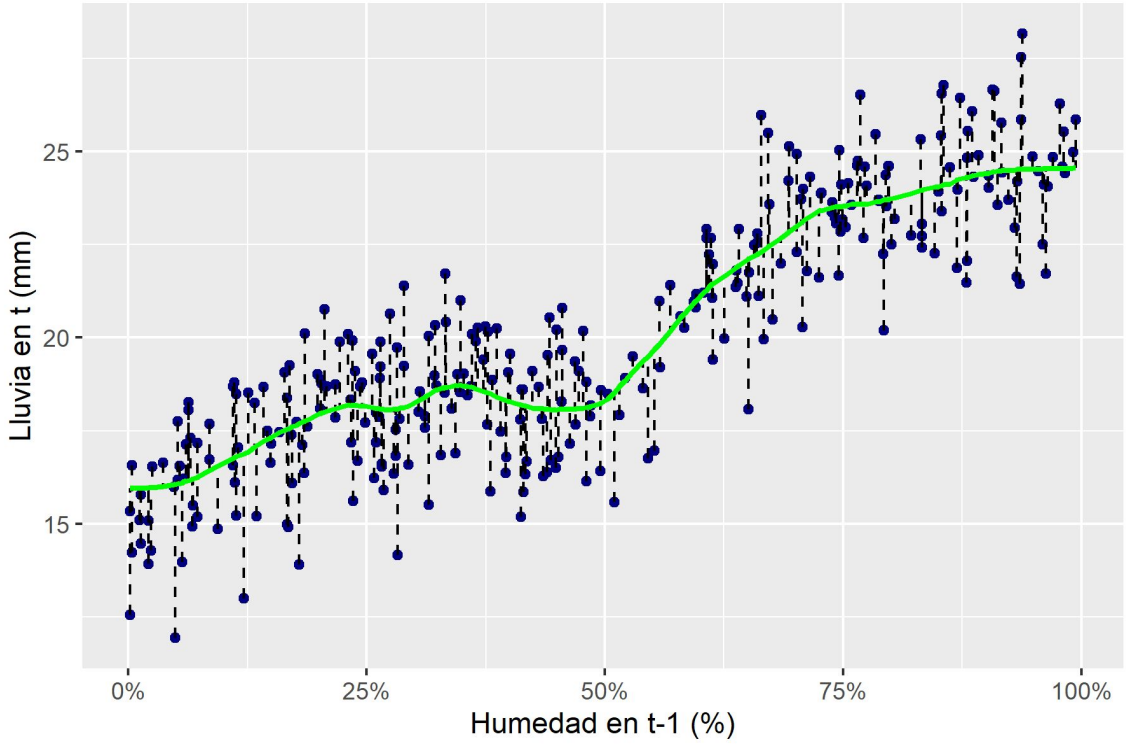




$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

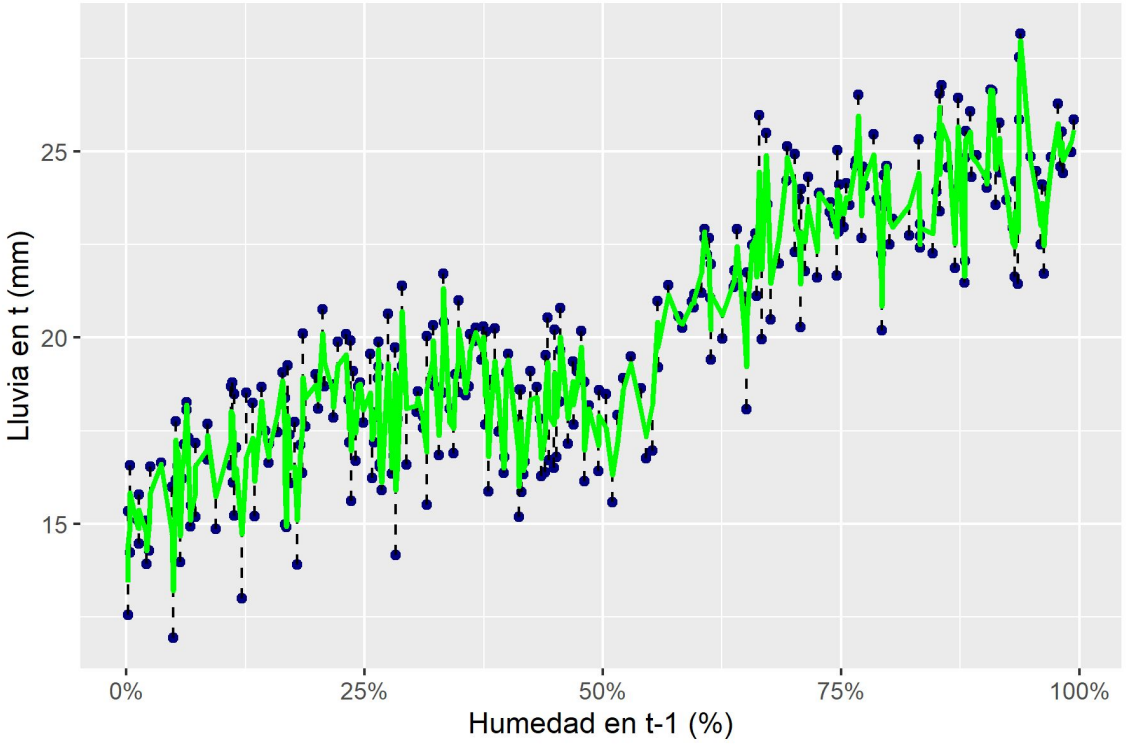
métrica de error empírica

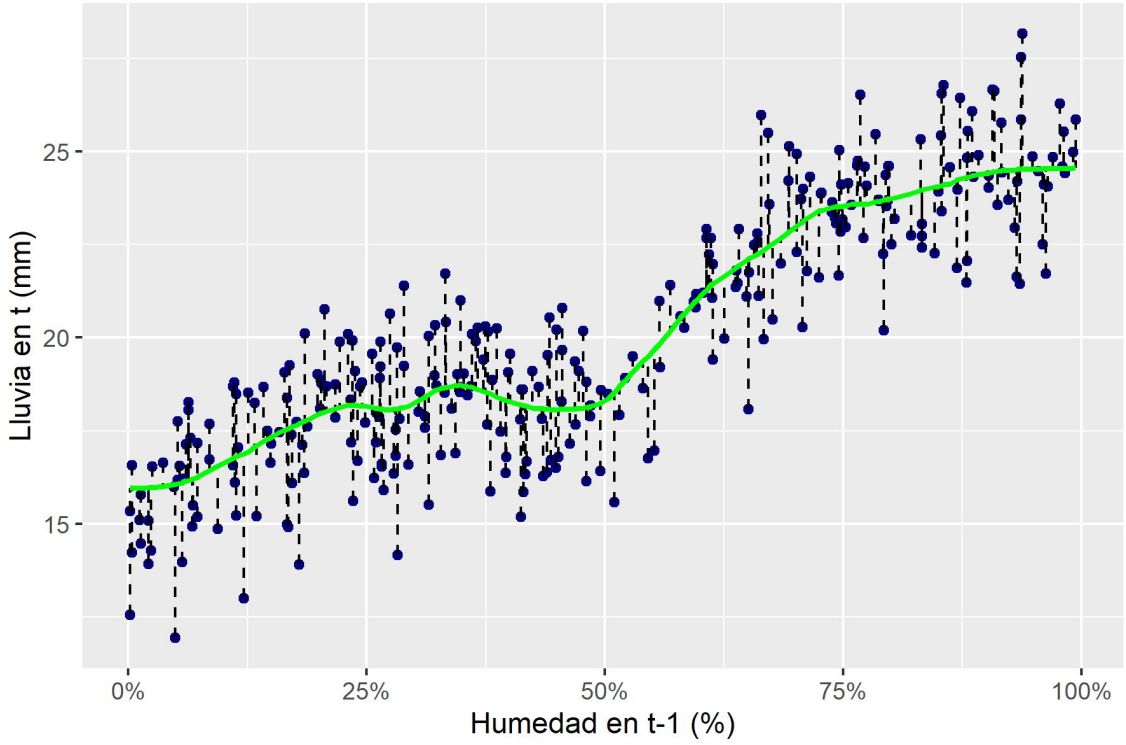




$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

métrica de error empírica





¿En qué datos nos interesa medir el error?

Datos

x1	x2	...	xp	y
0.127	0.159	...	0.675	0
0.706	0.073	...	0.681	1
0.541	0.118	...	0.864	0
...	...	...	...	...
0.114	0.449	...	0.484	0

Train

0.127	0.159	...	0.675	0
0.706	0.073	...	0.681	1
0.541	0.118	...	0.864	0
...	...	...	...	...

"Test" (datos no usados en train)

...	...	...	...	...
0.114	0.449	...	0.484	0

algoritmo de entrenamiento

$\hat{f}(x)$

algoritmo de predicción

$\hat{y}_{\text{test}}$

$\hat{y}_{\text{train}}$

$\text{MSE}_{\text{train}}$

MSE_{test}

Set de entrenamiento: para estimar f(X) (“el modelo”)

Set de test: para evaluar con qué precisión podemos predecir Y conociendo X. Simula ser un dataset con respuesta desconocida al momento de entrenar el modelo.

¡Queremos minimizar el error en test!

→ Indica que estamos estimando la parte sistemática del DGP bien

i.e. podemos **predecir *Y* con precisión en ejemplos en los que no lo sabemos de antemano** :)

IMPORTANTE: más adelante vamos a ver que, en la práctica, necesitamos por lo menos un tercer set adicional

La métrica de error empírica en test puede ser inexacta como medida general de error porque usa solo un conjunto de ejemplos del DGP.

Supongamos que podemos acceder a **muchas realizaciones del DGP**
→ podemos evaluar mejor el error estimando f(X) repetidamente en cada set de entrenamiento y promediando el MSE en todos los sets de test.
→ lo llamamos MSE, **el valor esperado del error de test**

$$MSE = E(Y - \hat{Y})^2 = E(Y - \hat{f}(X))^2$$

métrica de error teórica

$$MSE = E(Y - \hat{Y})^2 = E(Y - \hat{f}(X))^2$$

métrica de error teórica

$$MSE = [f(X) - \hat{f}(X)]^2 + \mathrm{Var}(\epsilon)$$

$$MSE = \underbrace{\mathrm{Var}(\hat{f}(X)) + \mathrm{Bias}(\hat{f}(X))^2}_{\textit{reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\textit{irreducible}}$$

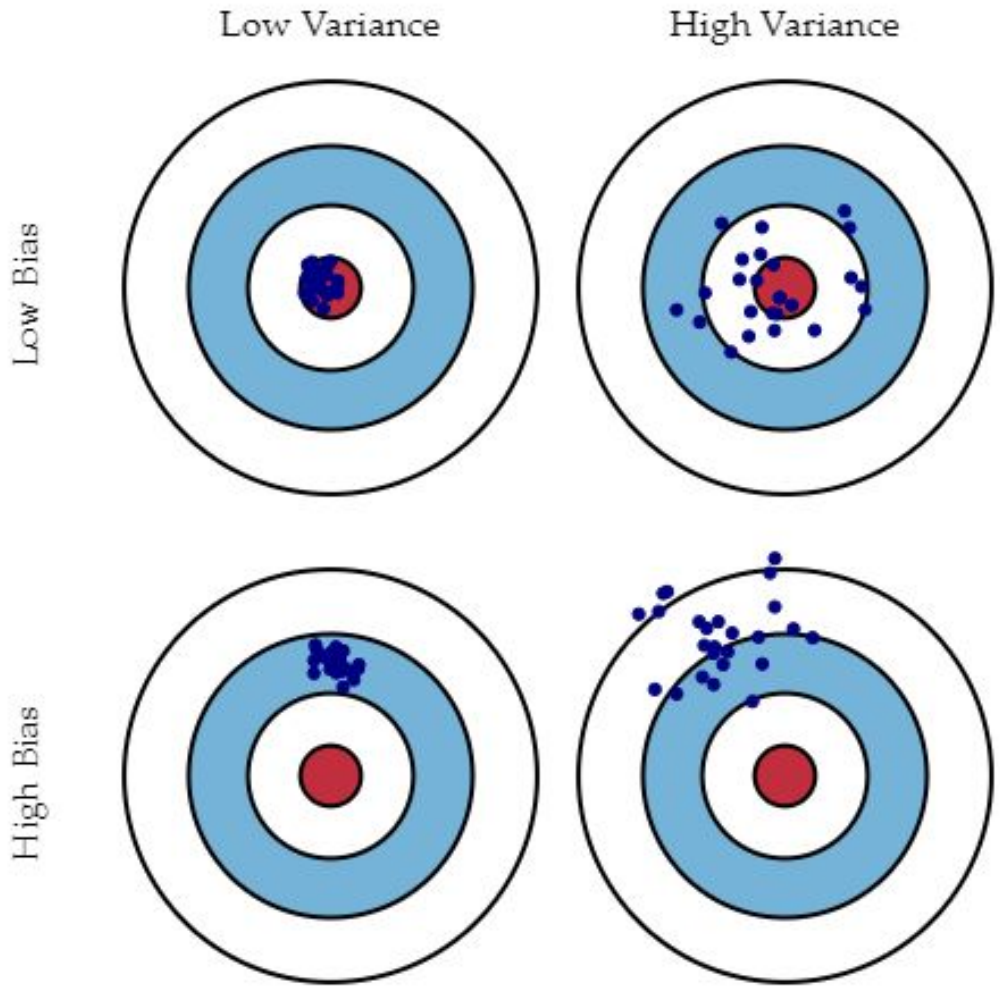
Sesgo

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Varianza

$$\text{Var}(\hat{\theta}) = E\left[(\hat{\theta} - E(\hat{\theta}))^2\right]$$

Son propiedades del estimador, ***no*** *de una estimación particular.*



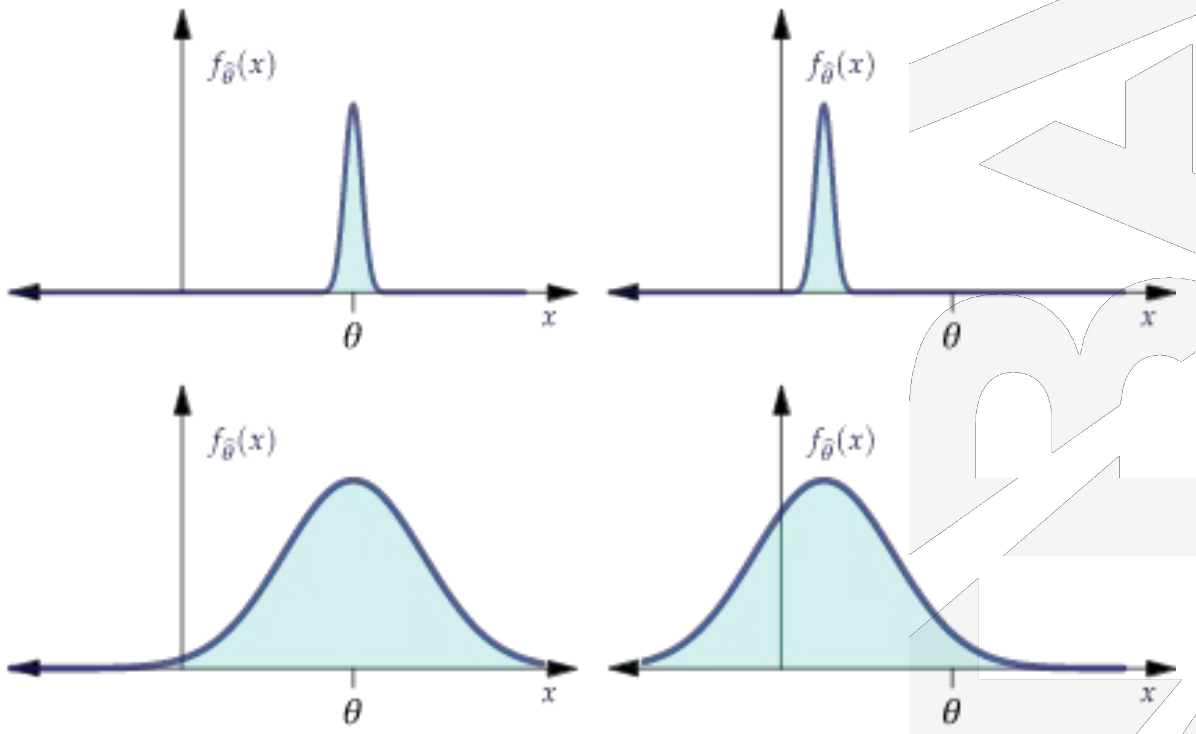
Fuente: https://scott.fortmann-roe.com/docs/BiasVariance.html

Sesgo

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Varianza

$$\text{Var}(\hat{\theta}) = E\left[(\hat{\theta} - E(\hat{\theta}))^2\right]$$



Fuente: https://es.mathigon.org/course/intro-statistics/point-estimation

$$MSE = \mathrm{Var}(\hat{f}(X)) + \mathrm{Bias}(\hat{f}(X))^2 + \mathrm{Var}(\epsilon)$$

Sesgo

El error de predicción que se comete cuando en promedio no podemos estimar correctamente f(X)

Varianza

Representa cuánto cambian los valores predichos cuando se ajusta el modelo con un conjunto diferente de ejemplos
Un modelo con varianza alta tenderá a tener un desempeño deficiente en otros sets de datos → es una fuente de error

$$MSE = \mathrm{Var}(\hat{f}(X)) + \mathrm{Bias}(\hat{f}(X))^2 + \mathrm{Var}(\epsilon)$$

El error irreducible

Representa **cuánto varía Y debido a factores distintos de X**

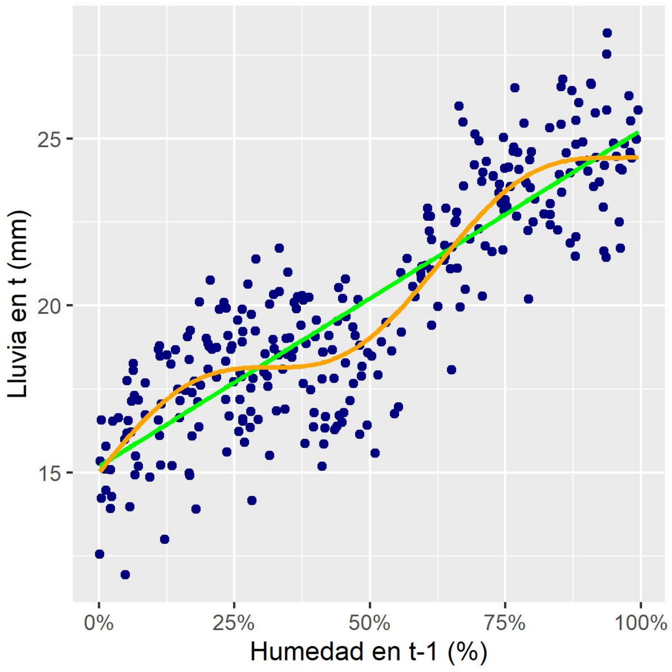
- Variables no medidas o inaccesibles que pueden ser útiles para predecir Y.
- Variabilidad intrínseca del fenómeno que no podemos explicar.

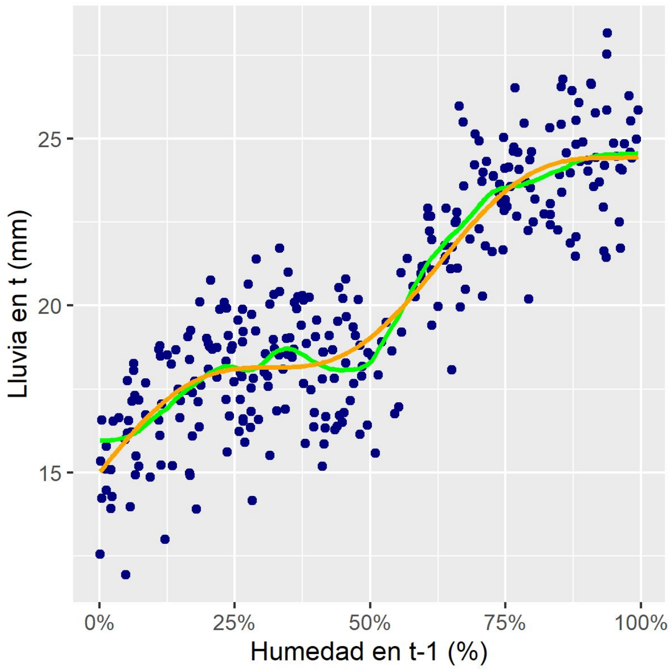
Si mucha variabilidad de Y no puede ser explicada por X
$\longrightarrow$ $\hat{Y}$ va a tender a diferir de Y
$\longrightarrow$ performance baja (MSE alto)

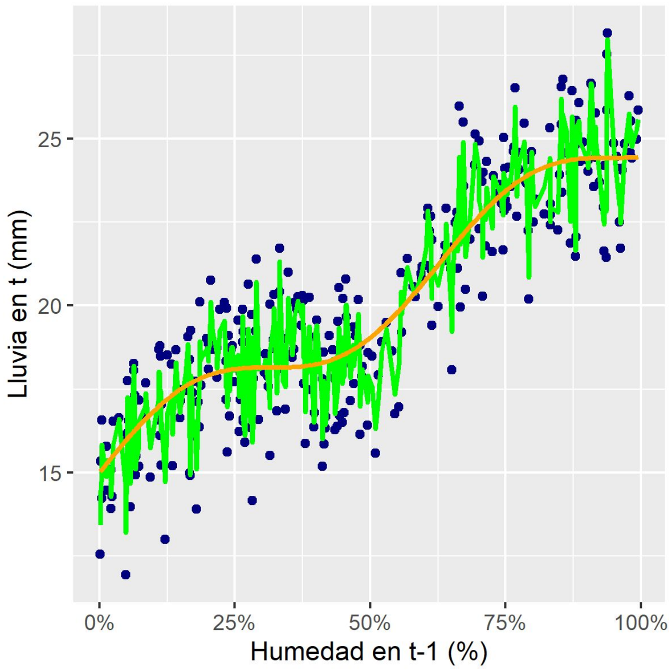
Y no hay nada que podamos hacer al respecto!
El error irreducible pone una **cota superior a la performance**

flexibilidad

Underfitting ← - - - - - - → Overfitting







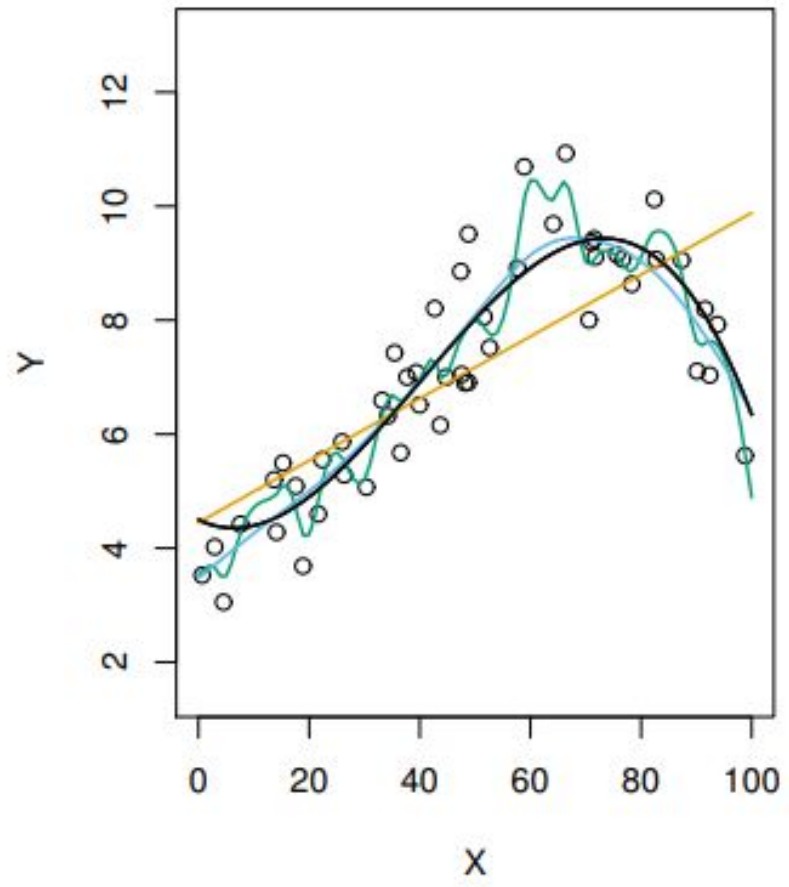
¿Cómo es el sesgo y la varianza en cada escenario?

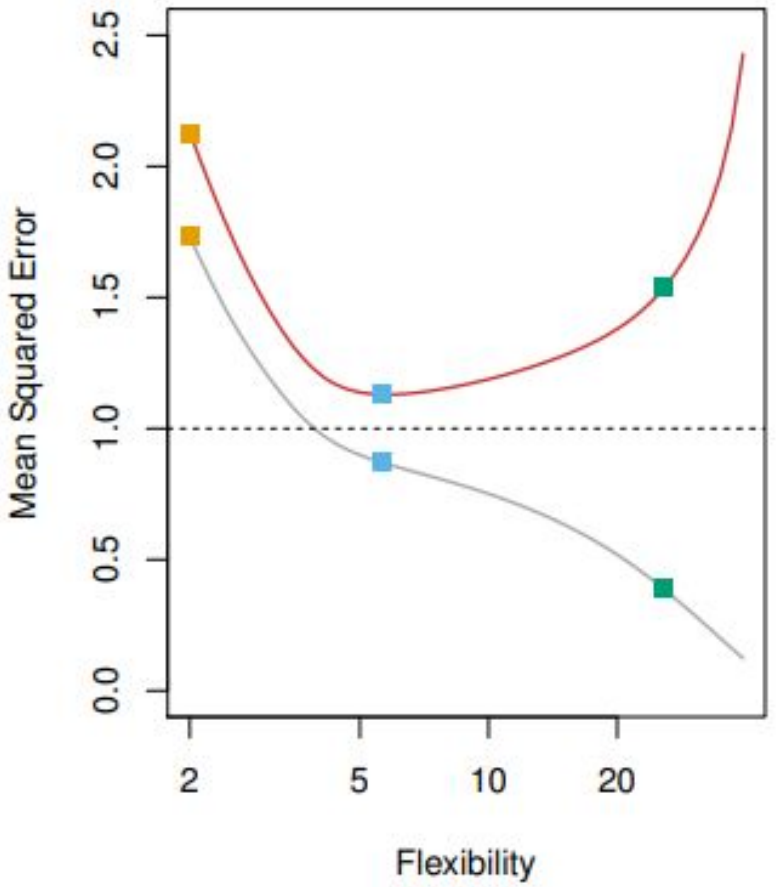
Idealmente, queremos un modelo con baja varianza y bajo sesgo, y nos gustaría que el error irreducible del DGP sea lo más bajo posible, dado X.

Sin embargo, existe un **trade-off entre el sesgo y la varianza** que hace que esto sea difícil de lograr.

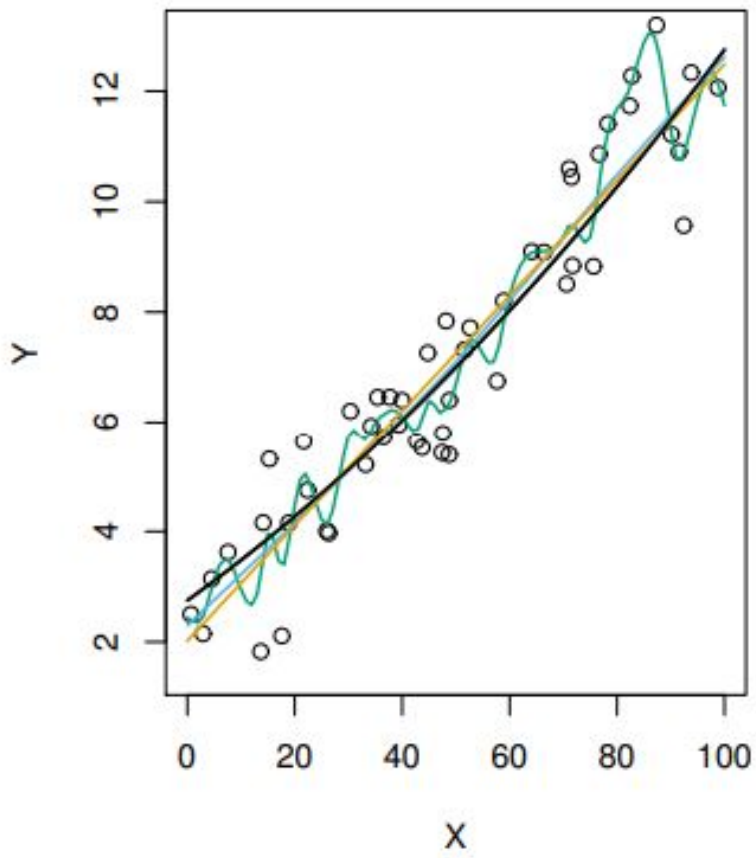
- demasiada flexibilidad → **overfitting** (alta varianza y bajo sesgo)
- muy poca flexibilidad → **underfitting** (baja varianza y alto sesgo)

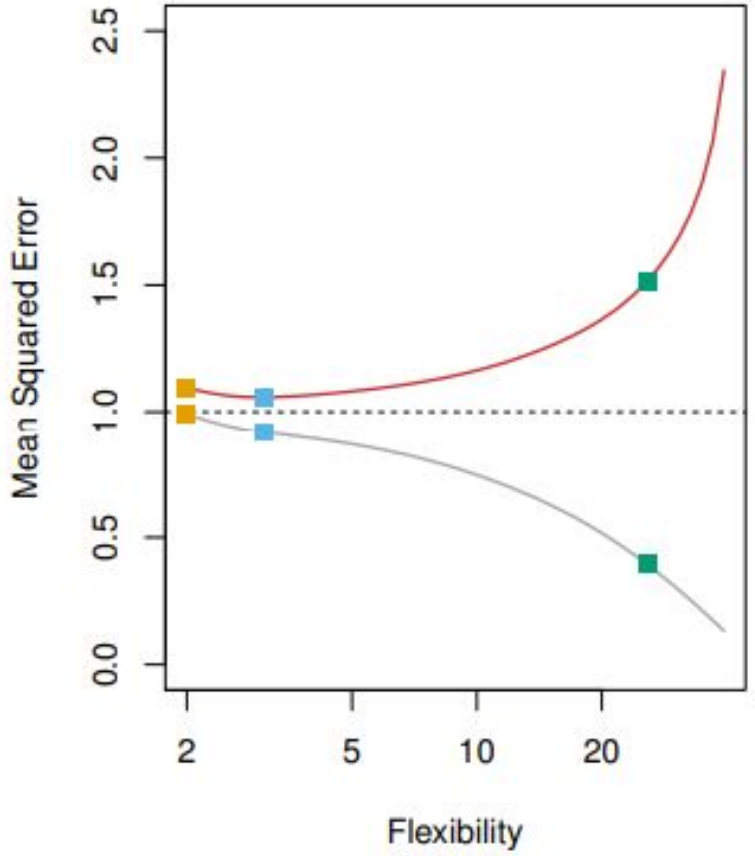
Debemos **calibrar cuidadosamente la flexibilidad** de los modelos para lograr una buena performance predictiva



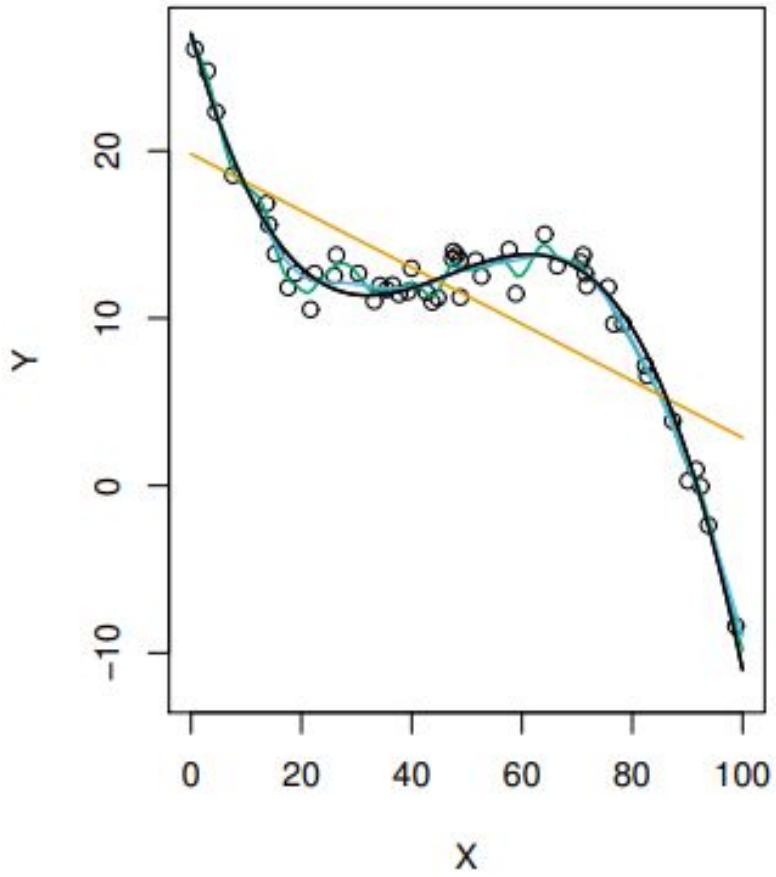


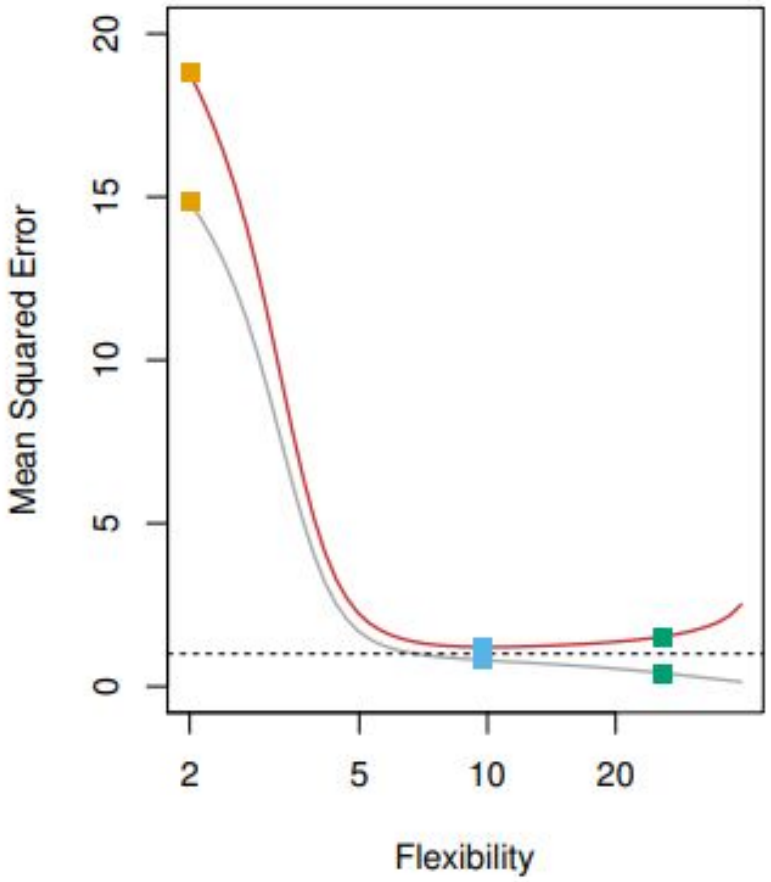
Fuente: https://www.statlearning.com/



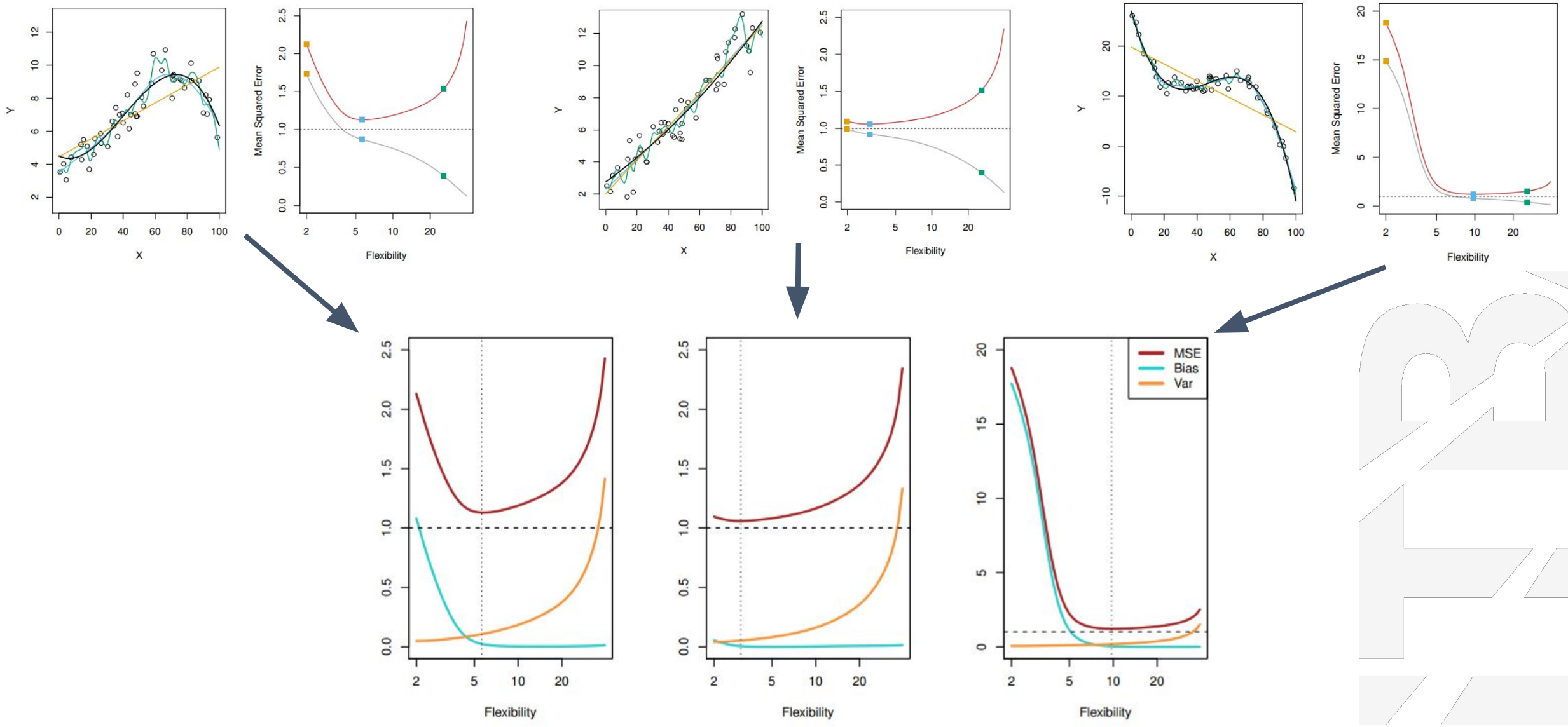


Fuente: https://www.statlearning.com/





Fuente: https://www.statlearning.com/



Fuente: https://www.statlearning.com/

Los tres problemas principales que causan una performance predictiva deficiente son el **error irreducible**, el **overfitting** y el **underfitting**.

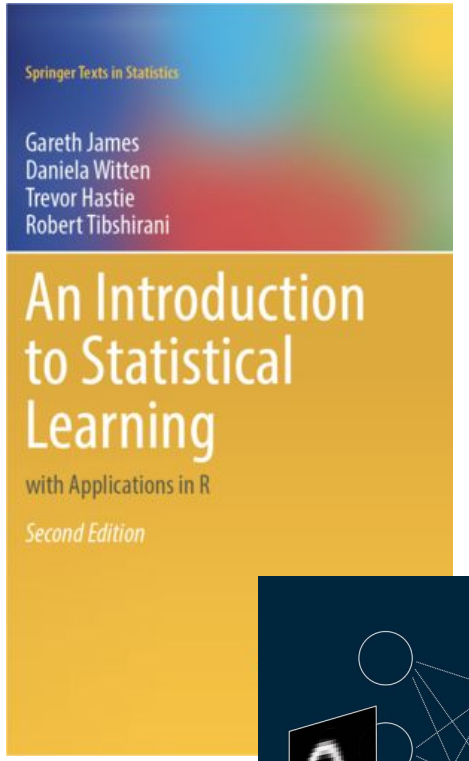
De estos, solo los dos últimos son controlables y ambos tienen su origen en el **equilibrio entre sesgo y varianza**.

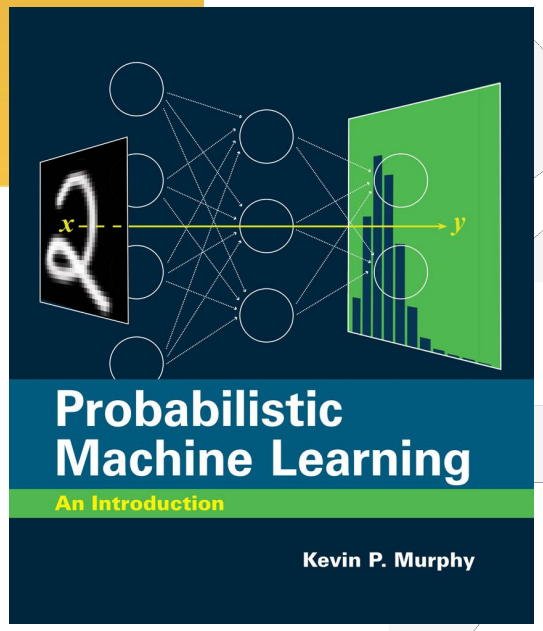
El tamaño y la forma de este trade-off se ven afectados por:

- la **flexibilidad** del método de estimación de f()
- la **complejidad** del DGP
- la **variabilidad irreducible** de la variable target

Lecturas recomendadas

- *An Introduction to Statistical Learning, 2nd Edition (cap. 2)*
- *Probabilistic Machine Learning: An Introduction (1.1 y 1.2)*





This Is The End
My only Friend
The End

(de la parte teórica) :D