Aprendizaje Automático
Segundo Cuatrimestre de 2016

Regresión No Lineal y Regresión Logística

Clase dada en el pizarrón. Transparencias a modo de referencia. Bibliografía:

- James, Witten, Hastie & Tibshirani, "An Introduction to Statistical Learning", Springer, 2015. Secciones 4.3, 6.2 y 7.1.
- Bishop, "Pattern Recognition and Machine Learning", Springer, 2006. Secciones 1.1 y 3.1.
- S. Fortmann-Roe, "Understanding the Bias-Variance Tradeoff". Artículo online.

Gracias a Ramiro Gálvez por la ayuda y los materiales para esta clase.

DEPARTAMENTO DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

Regresión Lineal Simple
(repaso)

- Consiste en predecir una respuesta cuantitativa *Y* en base a una única variable predictora *X*, ajustando una **recta** a los datos.

$$Y \approx \beta_0 + \beta_1 \cdot X$$

Ordenada al origen
(*intercept*)

Pendiente
(*slope*)

- β_0 y β_1 son los coeficientes desconocidos que vamos a estimar, o ajustar en base a los datos de entrenamiento. Una vez estimados, los podemos usar para predecir:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$$

Valor predicho para *Y*
cuando *X*=*x*

Estimación de β_0

Estimación de β_1

Nueva instancia

Regresión de Polinomios

- También podemos ajustar un **polinomio de grado *M*** a los datos.

$$Y \approx \beta_0 + \beta_1\, X + \beta_2\, X^2 + \beta_3\, X^3 + \ldots + \beta_M\, X^M$$

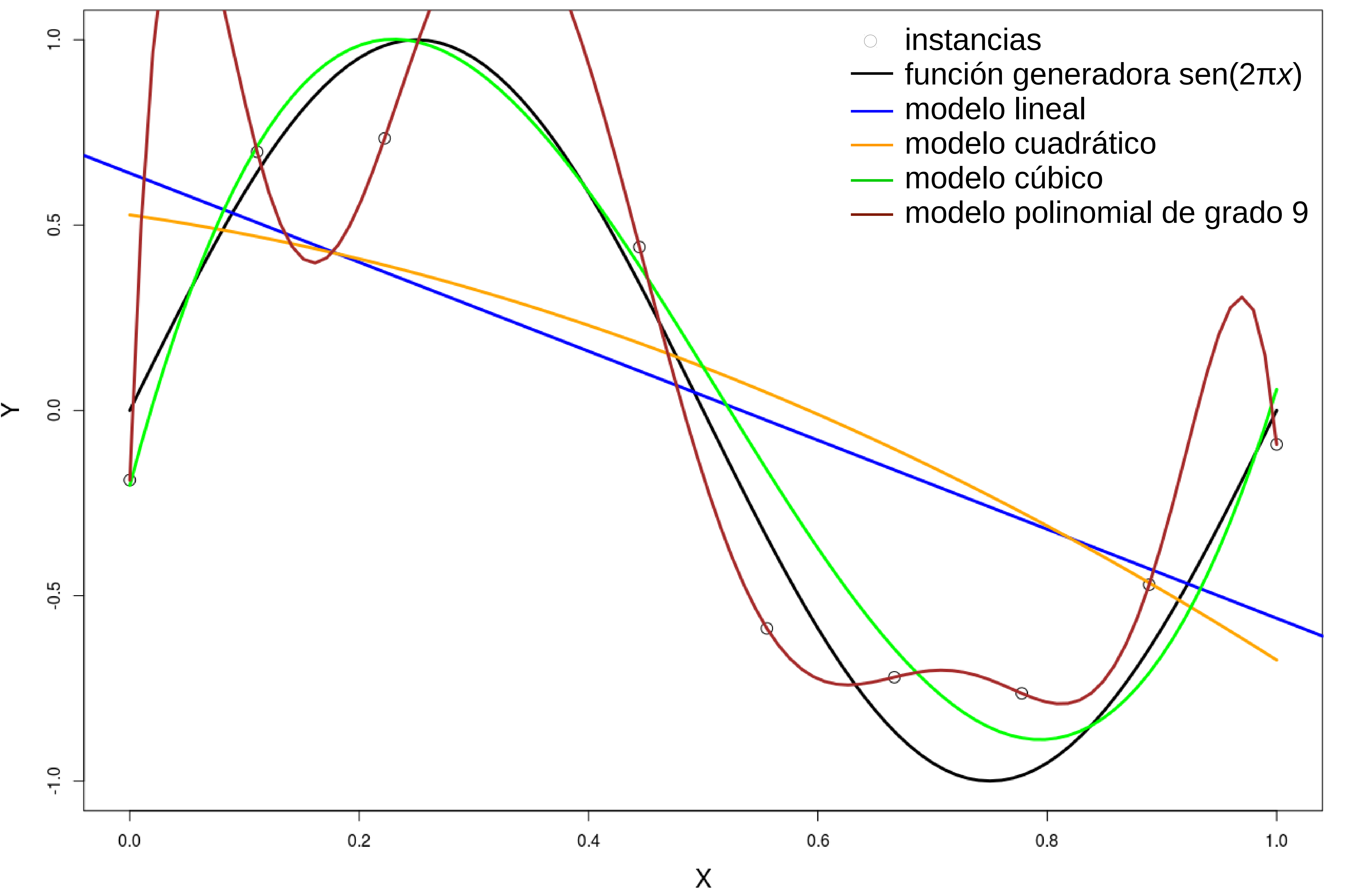
- Pese a su nombre, esto sigue siendo una regresión **lineal** de los coeficientes β_i con variables predictoras X, X^2, X^3, ..., X^M.

$$\mathrm{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1\, x_i - \hat{\beta}_2\, x_i^2 - \ldots - \hat{\beta}_M\, x_i^M\right)^2$$

- Los coeficientes β_i se pueden estimar con mínimos cuadrados, en forma similar a lo visto la clase anterior.
- M es un hiperparámetro del modelo.

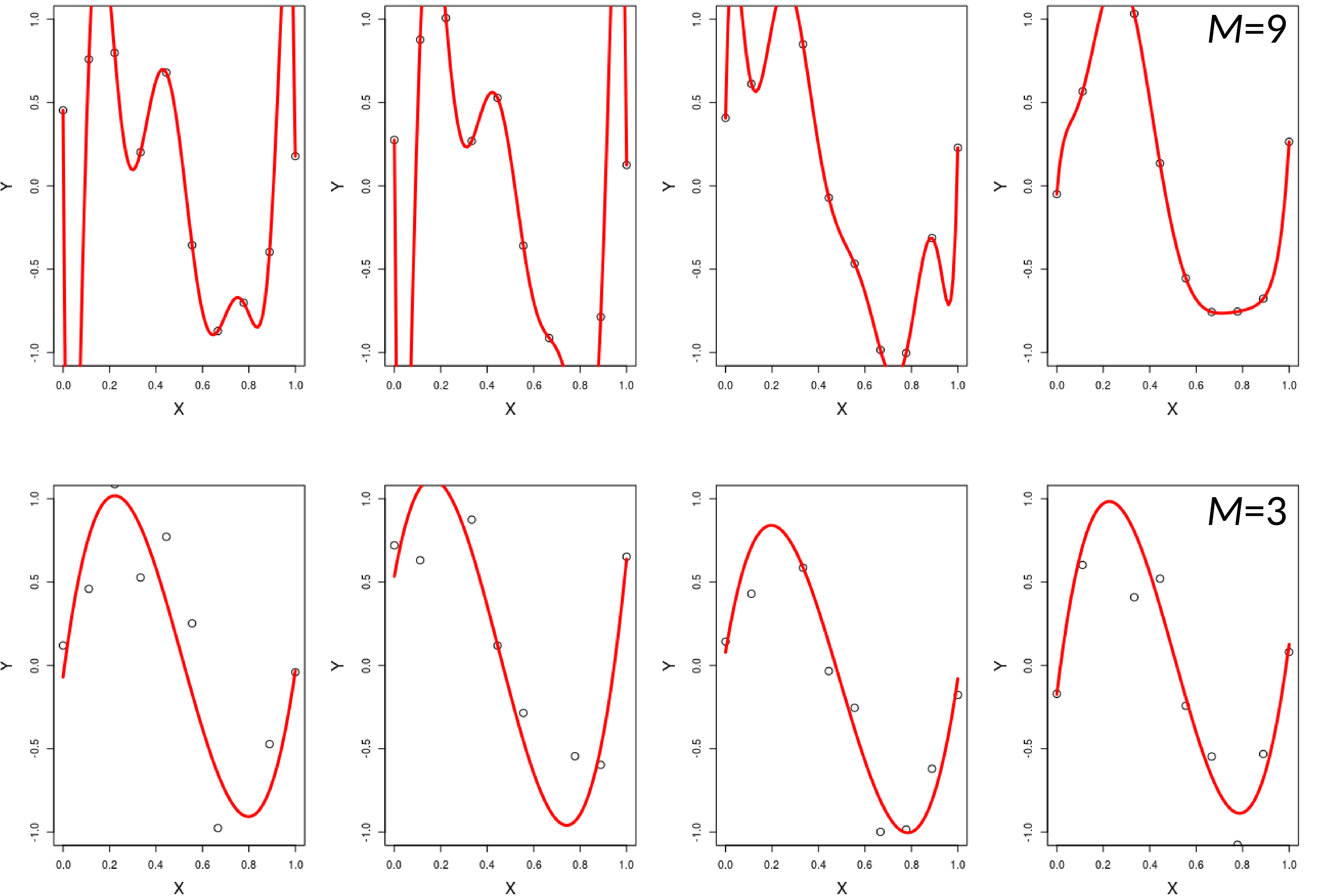
R Graphics: Device 2 (ACTIVE)

Generalizabilidad del modelo para distintos valores de M.



R Graphics: Device 2 (ACTIVE)

Sesgo vs. varianza



Sesgo vs. Varianza

datos observados

ruido en los datos

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

función objetivo

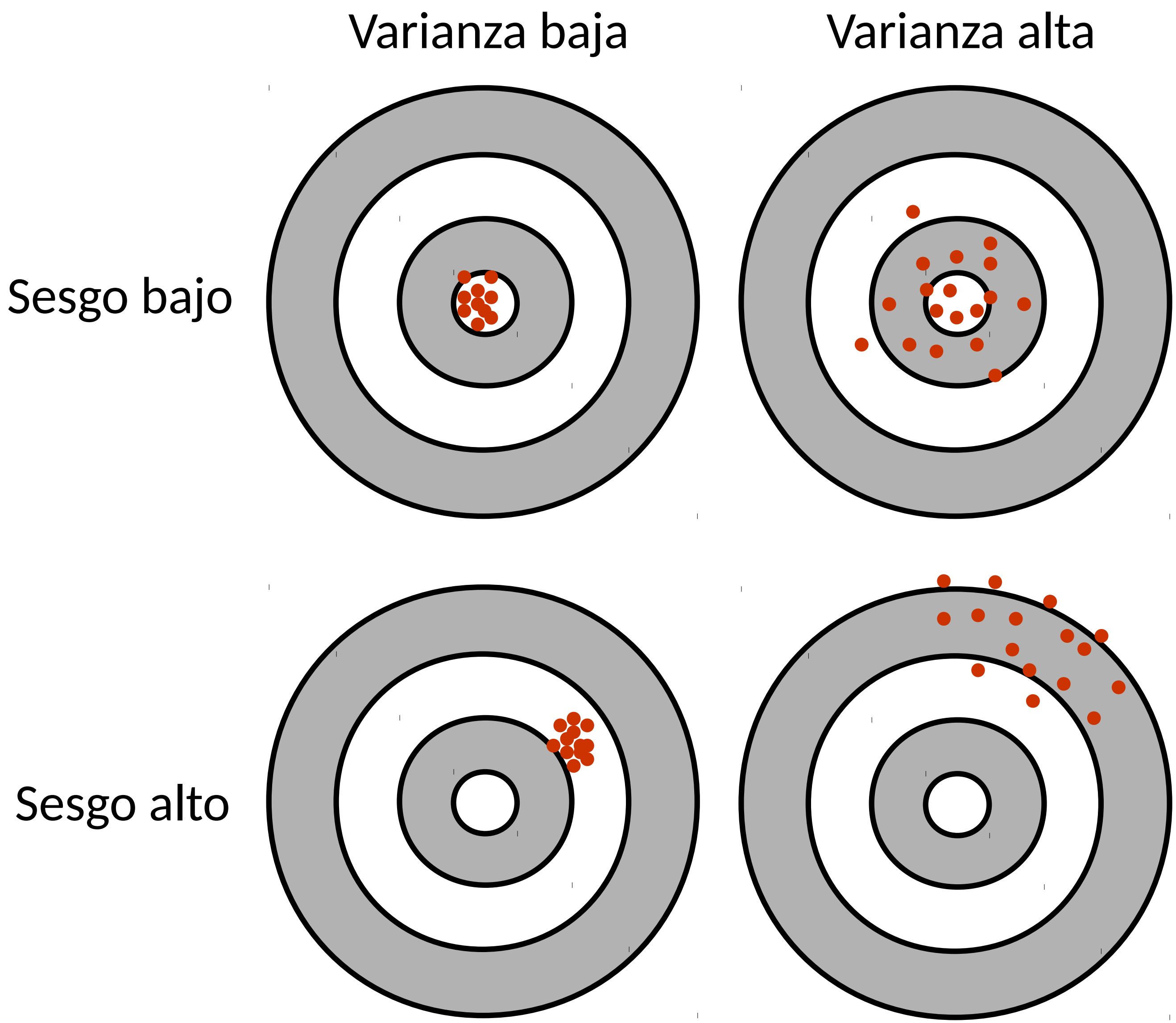
Error esperado del modelo:

$$\begin{aligned}
\mathbf{E}\left[y_0 - \hat{f}(x_0)\right]^2 &= \mathbf{E}\left[f(x_0) + \epsilon_0 - \hat{f}(x_0)\right]^2 \\
&= \mathbf{E}\left[f(x_0) - \hat{f}(x_0)\right]^2 + \mathrm{Var}(\epsilon) \\
&= \ldots
\end{aligned}$$

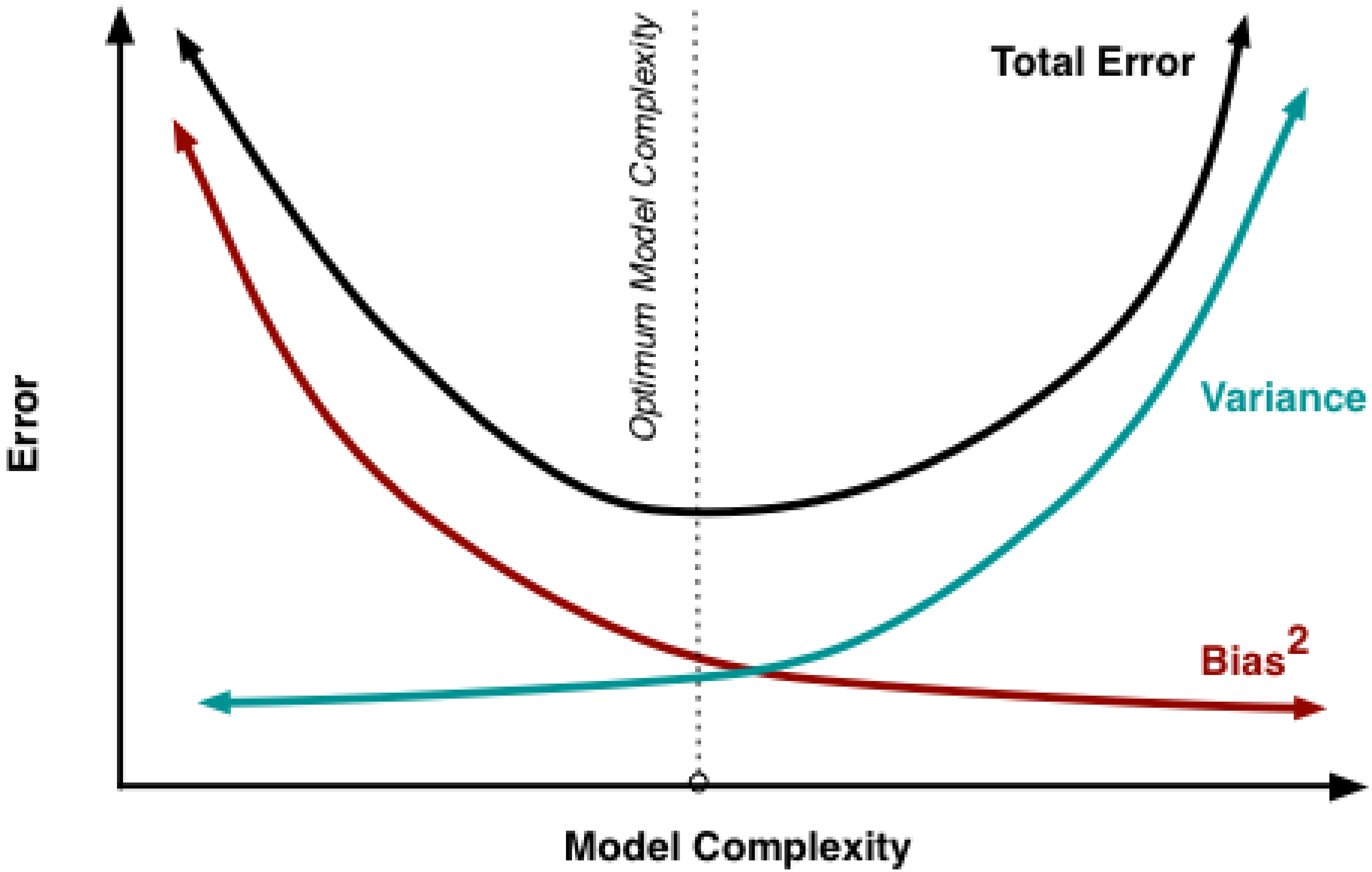
(pasos engorrosos aquí)

$$= \underbrace{\left(\mathbf{E}[\hat{f}(x_0)] - y_0\right)^2}_{\text{Sesgo}} + \underbrace{\mathbf{E}\left[\hat{f}(x_0) - \mathbf{E}\left[\hat{f}(x_0)\right]\right]^2}_{\text{Varianza}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Error no reducible}}$$

Sesgo vs. Varianza

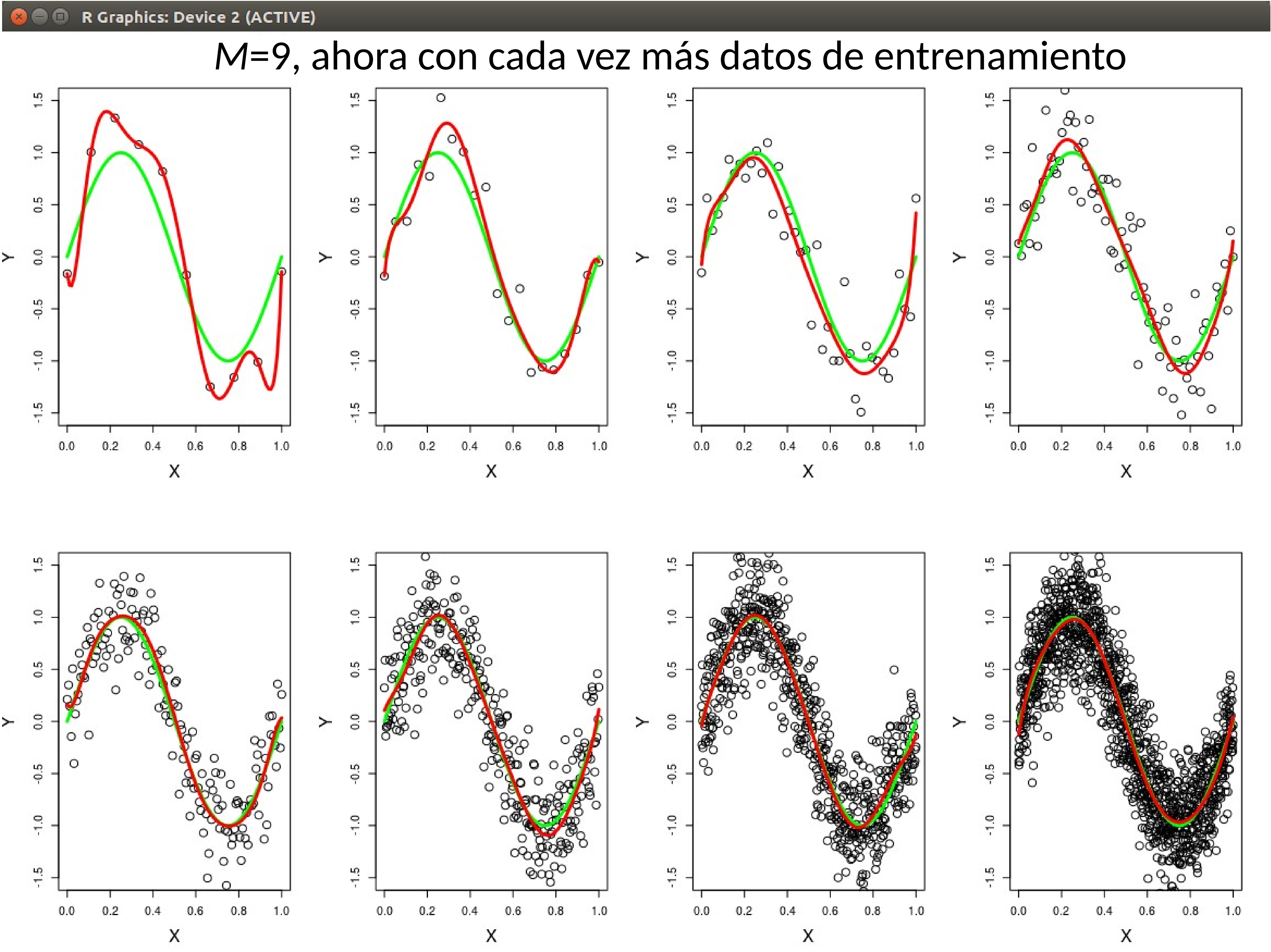


Sesgo vs. Varianza



http://scott.fortmann-roe.com/docs/BiasVariance.html

M=9, ahora con cada vez más datos de entrenamiento



Regularización

- Observaciones:
  - Si $\hat{\beta}_i$=0 para i>0, tenemos un modelo muy simple (constante).
  - A medida que crecen los $\hat{\beta}_i$, el modelo se hace más complejo.
  - Regla del pulgar: valores altos de $\hat{\beta}_i$ llevan al sobreajuste.
- Regularización: Penalizar valores altos de $\hat{\beta}_i$

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 + \lambda \sum_{i=1}^{M} \hat{\beta}_i^{\,q}$$

- q, λ son hiperparámetros del modelo (por ejemplo, cuando q=2 la técnica se conoce como *"Ridge Regression"*).

Regresión de Funciones Base

- Generalización del problema de regresión lineal:

$$Y \approx \beta_0 + \beta_1\, \phi_1(\mathbf{X}) + \beta_2\, \phi_2(\mathbf{X}) + ... + \beta_M\, \phi_M(\mathbf{X})$$

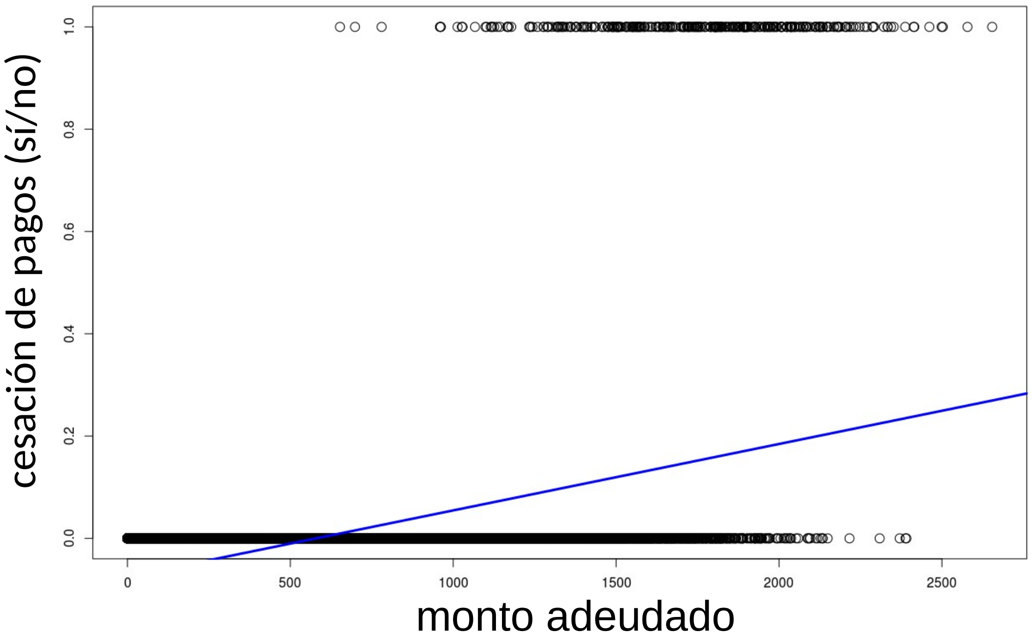
donde $\mathbf{X} = [X_1, X_2, ..., X_p]^{\mathrm{T}}$

- A las ϕ_i se las denomina **funciones base**.

Regresión Logística

Regresión Logística

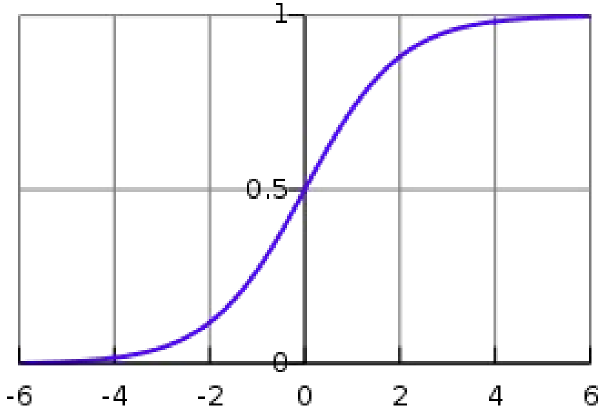
- La regresión lineal no es buena para modelar la probabilidad de ocurrencia de un evento:



- El modelo no parece ajustarse bien a los datos, y además puede arrojar predicciones negativas o mayores a 1.

Regresión Logística

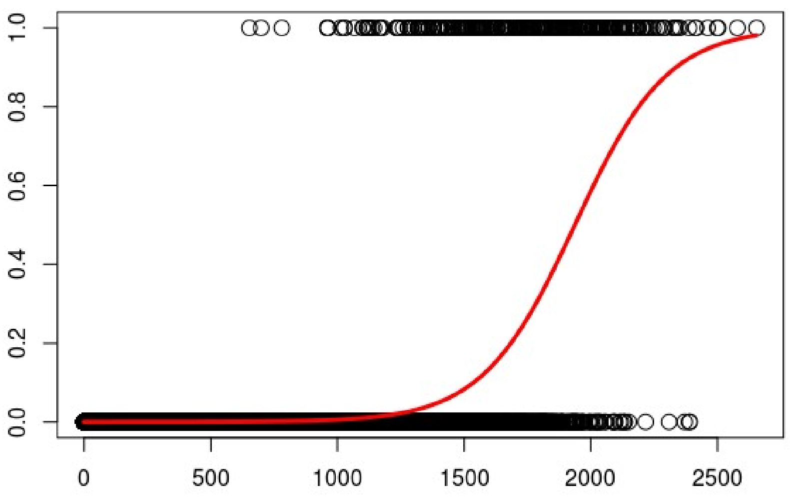
Función logística: $$f(t) = \frac{e^t}{1+e^t}$$



Usada para modelar, por ejemplo, el crecimiento de poblaciones biológicas y el desarrollo embrionario.

La **regresión logística** consiste en ajustar los coeficientes β_i a los datos de entrenamiento:

$$Y \approx \frac{e^{\beta_0+\beta_1 \cdot X}}{1+e^{\beta_0+\beta_1 \cdot X}}$$



Es una técnica útil para **clasificación**.

Regresión Logística

- Se extiende a múltiples variables predictoras:

$$Y \approx \frac{e^{\beta_0+\beta_1\cdot X_1+\beta_2\cdot X_2+\ldots+\beta_p\cdot X_p}}{1+e^{\beta_0+\beta_1\cdot X_1+\beta_2\cdot X_2+\ldots+\beta_p\cdot X_p}}$$

```
> modreg = glm(default=='Yes' ~ balance + student + income, family=binomial)
> summary(modreg)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
income       3.033e-06  8.203e-06   0.370  0.71152
```

```
> modreg = glm(default=='Yes' ~ balance + student, family=binomial)
> summary(modreg)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.075e+01  3.692e-01 -29.116  < 2e-16 ***
balance      5.738e-03  2.318e-04  24.750  < 2e-16 ***
studentYes  -7.149e-01  1.475e-01  -4.846 1.26e-06 ***
```