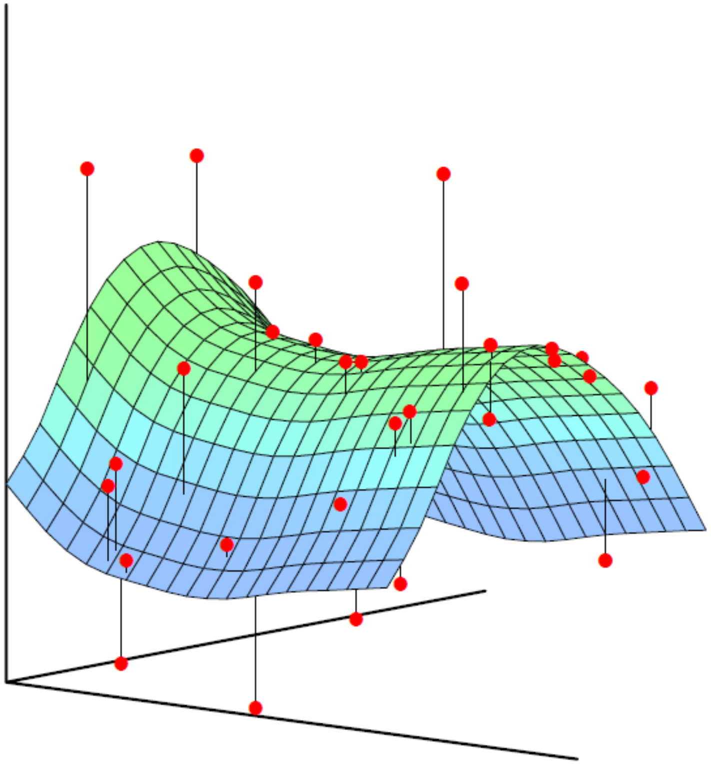
aprendizaje estadístico o automático

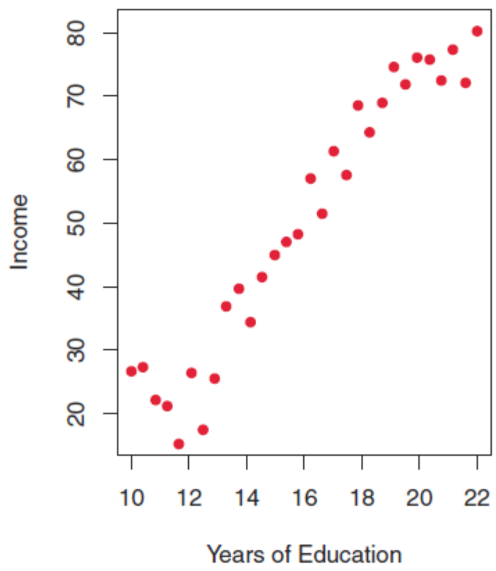


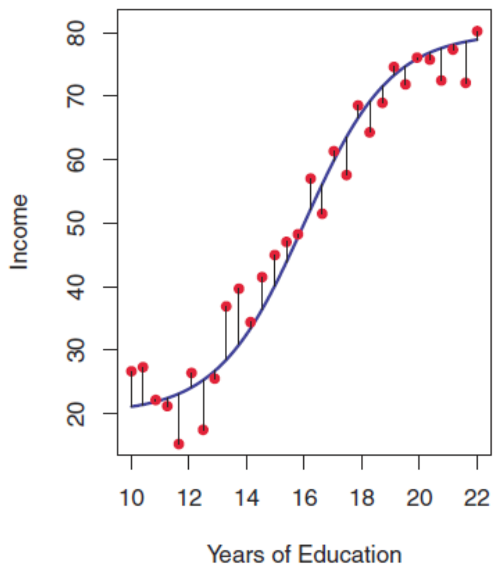
$$
\begin{aligned}
\mathrm{EPE}_k(x_0) &= \mathrm{E}[(Y-\hat{f}_k(x_0))^2|X=x_0] \\
&= \sigma^2 + [\mathrm{Bias}^2(\hat{f}_k(x_0)) + \mathrm{Var}_{\mathcal{T}}(\hat{f}_k(x_0))] \\
&= \sigma^2 + \Big[f(x_0) - \frac{1}{k}\sum_{\ell=1}^{k} f(x_{(\ell)})\Big]^2 + \frac{\sigma^2}{k}.
\end{aligned}
$$

¿Qué es el aprendizaje estadístico?

años de educación →**Ingresos**

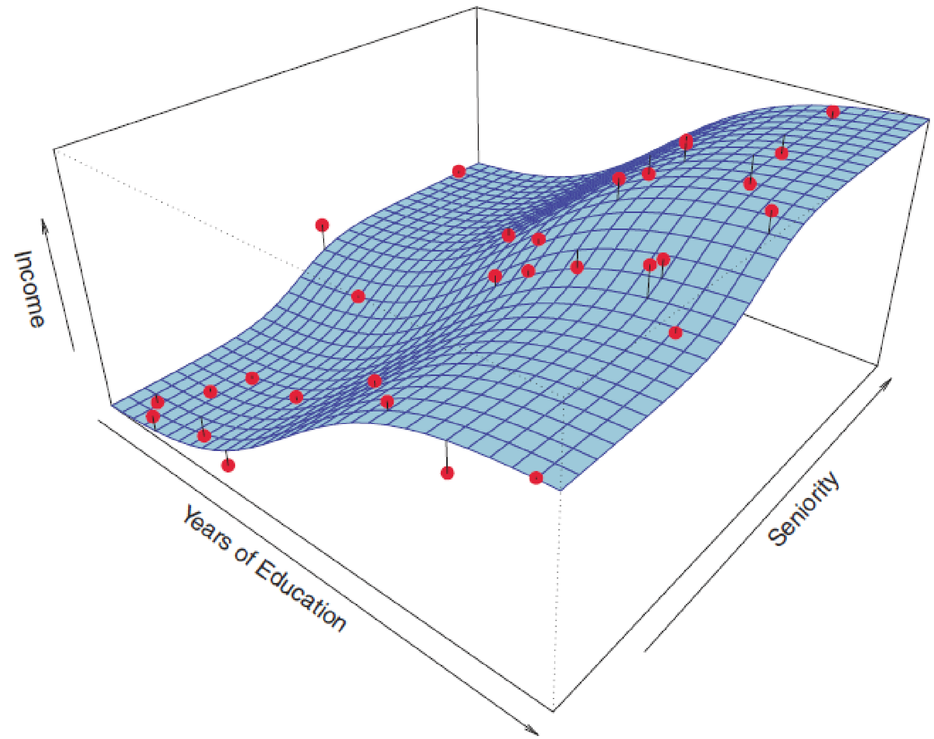
El gráfico de la izq sugiere que uno podría ser capaz de predecir los ingresos conciendo los años de educación. Sin embargo, la función ***f*** que relaciona la variable de entrada a la variable de salida es en general desconocida. En esta situación uno debe estimar ***f*** basado en los puntos observados.





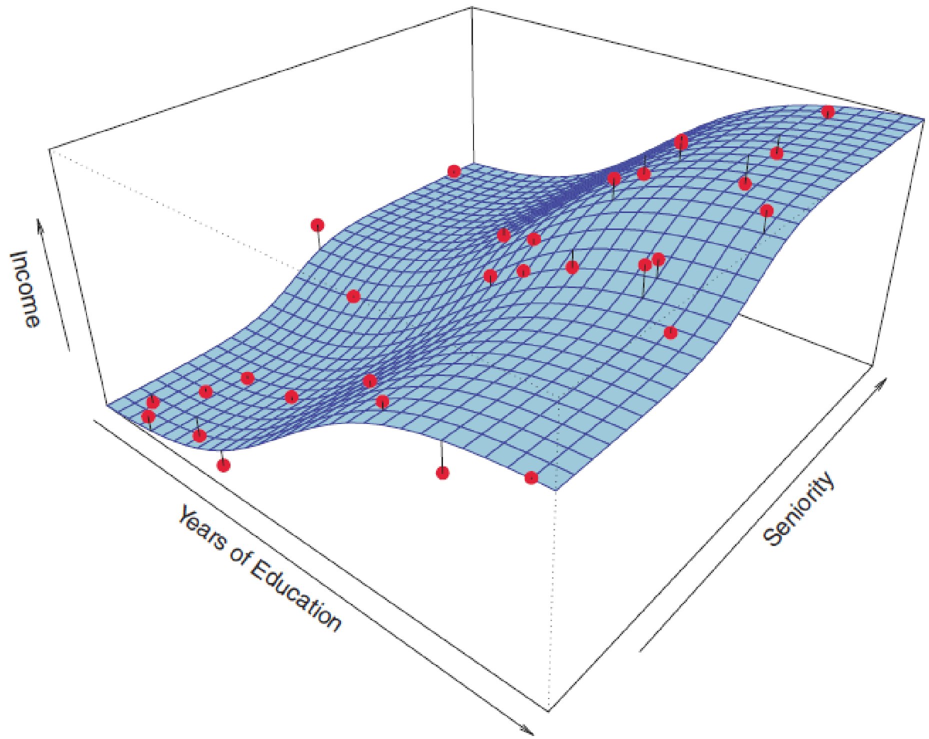
$$\boxed{Y = f(X) + \epsilon}$$

En general, la función *f* puede involucrar más de una variable de entrada. En el siguiente gráfico vemos el ingreso como una función de años de educación y antigüedad. Aquí ***f*** es una superficie curvada que debe ser estimada basado en los datos observados



En esencia, el **aprendizaje estadístico** se refiere a un conjunto de técnicas para estimar ***f*** a partir de datos conocidos, y al uso de herramientas (métricas) para evaluar las estimaciones obtenidas.

¿Porqué estimar *f*?



Hay dos razones principales por las que deseamis estimar ***f***:

→ ***Predicción***

→ ***Inferencia***

Predicción

Contamos con conjunto de entradas X y queremos conocer salida Y

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

pero resulta que Y no se puede obtener fácilmente. En esta configuración, dado que el término de error promedia a cero, podemos predecir Y usando

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

Donde *f^* representa nuestra estimación para ***f*** y por ende Y^ representa la predicción resultante para y . En esta configuración, *f^* a menudo se trata como una caja negra, en el sentido que uno **no suele estar preocupado con la forma exacta de f**, siempre que produce predicciones precisas para Y

La precisión de Y^ como predicción para Y depende de dos cantidades, que llamaremos el *error reducible* y el *error irreductible*. En general, f^ no será una estimación perfecta para f, y esta inexactitud introducirá algún error. Este error es reducible porque potencialmente podemos mejorar el precisión de f^ mediante el uso de la técnica estadística de aprendizaje más adecuada para estimar f.

$$
\begin{aligned}
E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\
&= \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}}
\end{aligned}
$$

El Aprendizaje Estadístico aborda las técnicas para estimar f con el objetivo de minimizar el error reducible.

Inferencia

A menudo nos interesa comprender la forma en que Y se ve afectada por los cambios en $X_1, . . . , X_p$. En esta situación, queremos estimar ***f***, pero nuestro objetivo no es, necesariamente, hacer predicciones para Y. En cambio, queremos entender la relación entre X e Y, o más específicamente, para entender *como cambia Y en función de $X_1, . . ., X_p$.*

· En este contexto nos interesa saber:

- ***¿Qué predictores están asociados con la respuesta?***

  A menudo se da el caso que solo una pequeña fracción de los predictores disponibles o están asociados con Y.

- ***¿Cuál es la relación entre la respuesta Y y cada predictor X_i?***

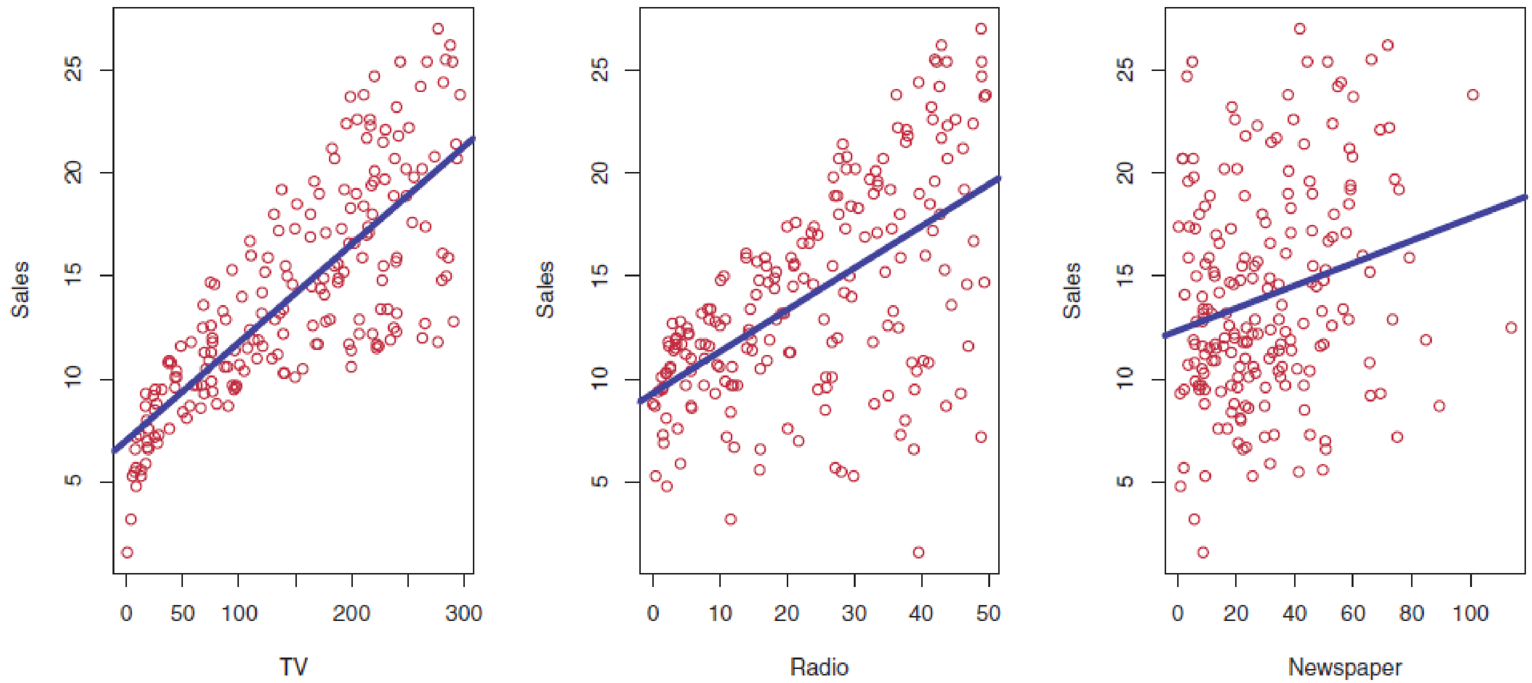
  Algunos X una correlación positiva con **Y** mientras que otros pueden tenerla opuesta. Dependiendo de la complejidad de ***f***, la relación entre la respuesta **Y** y un predictor también puede depender de los valores de los otros predictores.

- ***¿Se puede resumir adecuadamente la relación entre Y y cada predictor usando una recta, o la relación ec más complicada?***

  Históricamente,la mayoría de los métodos para estimar ***f*** han tomado una forma lineal . En algunas situaciones, tal suposición es razonable o incluso deseable.

Predicción/Inferencia

- *Modelo para incrementar las ventas*
- *¿Qué medios generan el mayor incrementoen las ventas?*
- *¿Cuánto aumento en las ventas se asocia con un aumento dado en TV?¿publicidad?*



¿Cómo estimamos f?

Siempre asumiremos que hemos observado un conjunto de n diferentes puntos de datos. Estas observaciones se llaman datos de **entrenamiento** porque usaremos estas observaciones de datos para entrenar o enseñar a nuestro método cómo estimar f.

- **Métodos paramétricos**
- **Métodos no paramétricos**

1. Primero, hacemos una suposición sobre el tipo de función, o forma .Por ejemplo, una suposición muy simple es que f es **lineal** en X:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

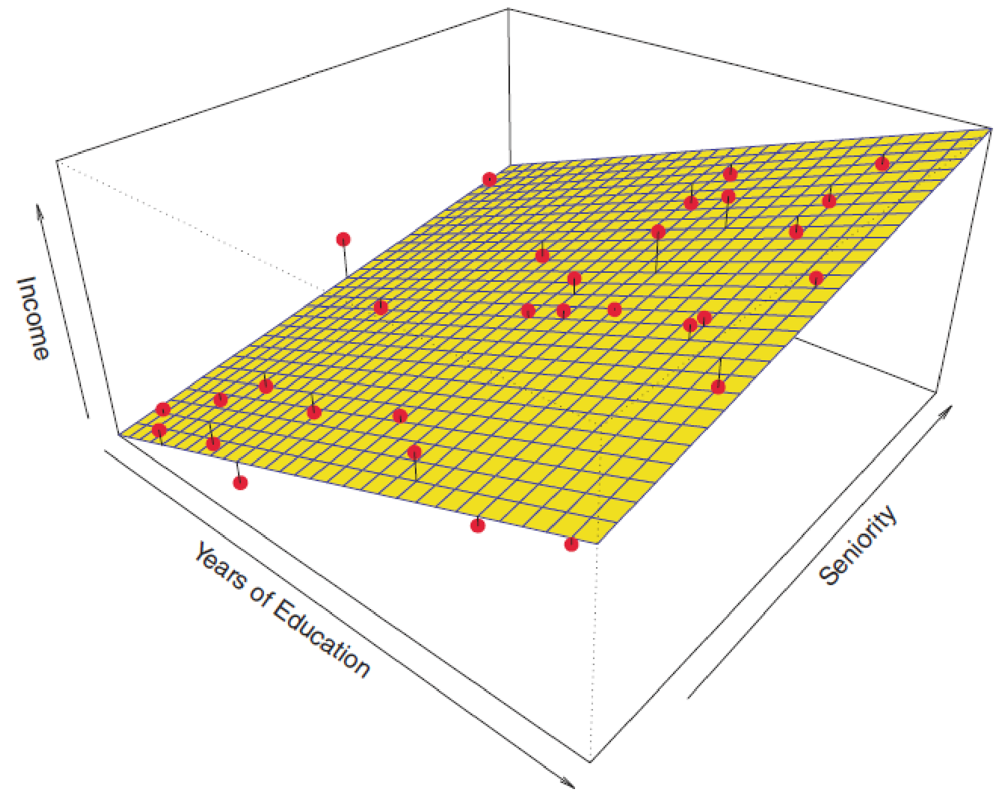
2. Después de que se haya seleccionado un modelo, necesitamos un procedimiento que use datos de entrenamiento para ajustar o entrenar el modelo. En el caso del modelo lineal necesitamos estimar los parámetros β_0, β_1,. . . , β_p. Es decir, encontrar valores de estos parámetros tales que

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

es más fácil estimar $\beta_0, \beta_1, \ldots, \beta_p$ en un modelo lineal , que ajustar una función completamente arbitraria ***f***.

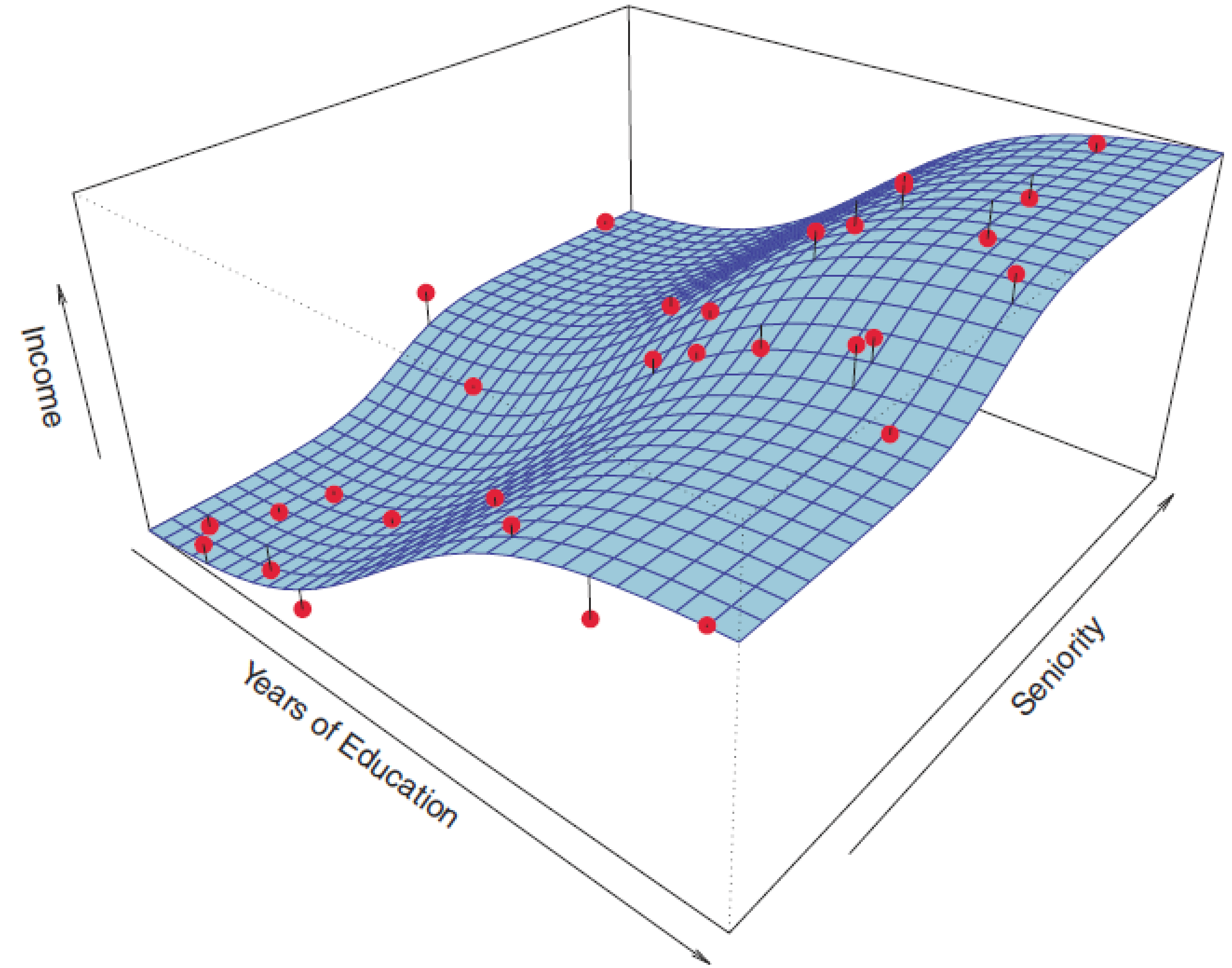
La desventaja potencial de un enfoque paramétrico es que el modelo que elijamos **generalmente** no coincidirá con la verdadera forma desconocida de f.

Podemos tratar de abordar este problema eligiendo modelos **flexibles** que pueden adaptarse a muchas formas funcionales posibles para f. Pero en general, ajustar un modelo más flexible requiere estimar un mayor cantidad de **parámetros**

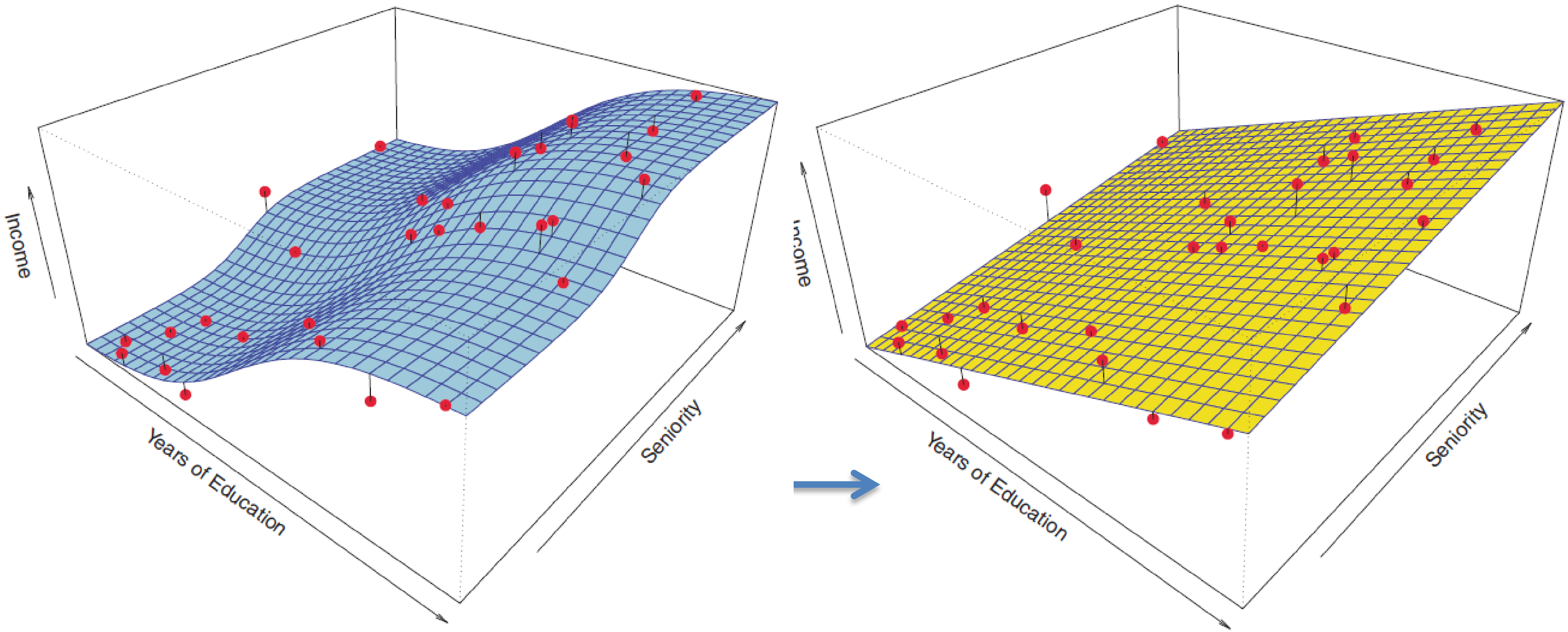


Veamos el modelo de ingresos extendido

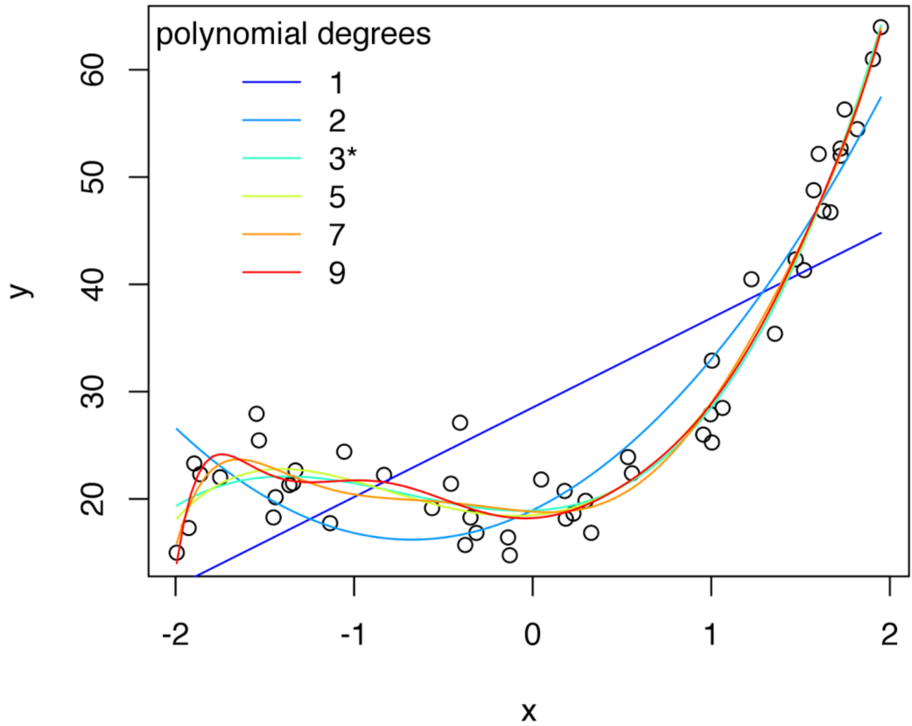
$$\texttt{income} \approx \beta_0 + \beta_1 \times \texttt{education} + \beta_2 \times \texttt{seniority}$$


Income
Years of Education
Seniority

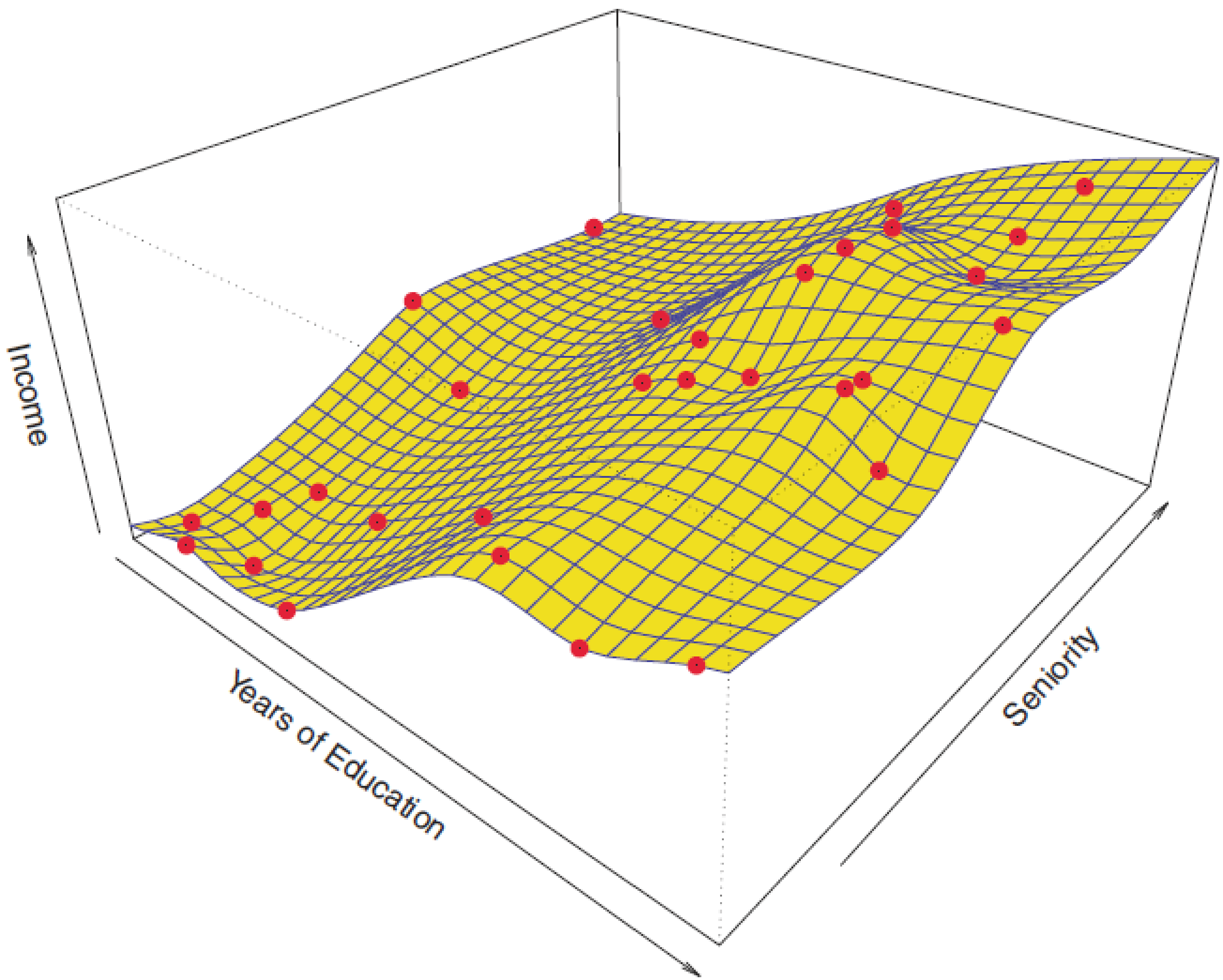
la verdadera f va a tener una tiene cierta curvatura que no es capturada por el modelo lineal



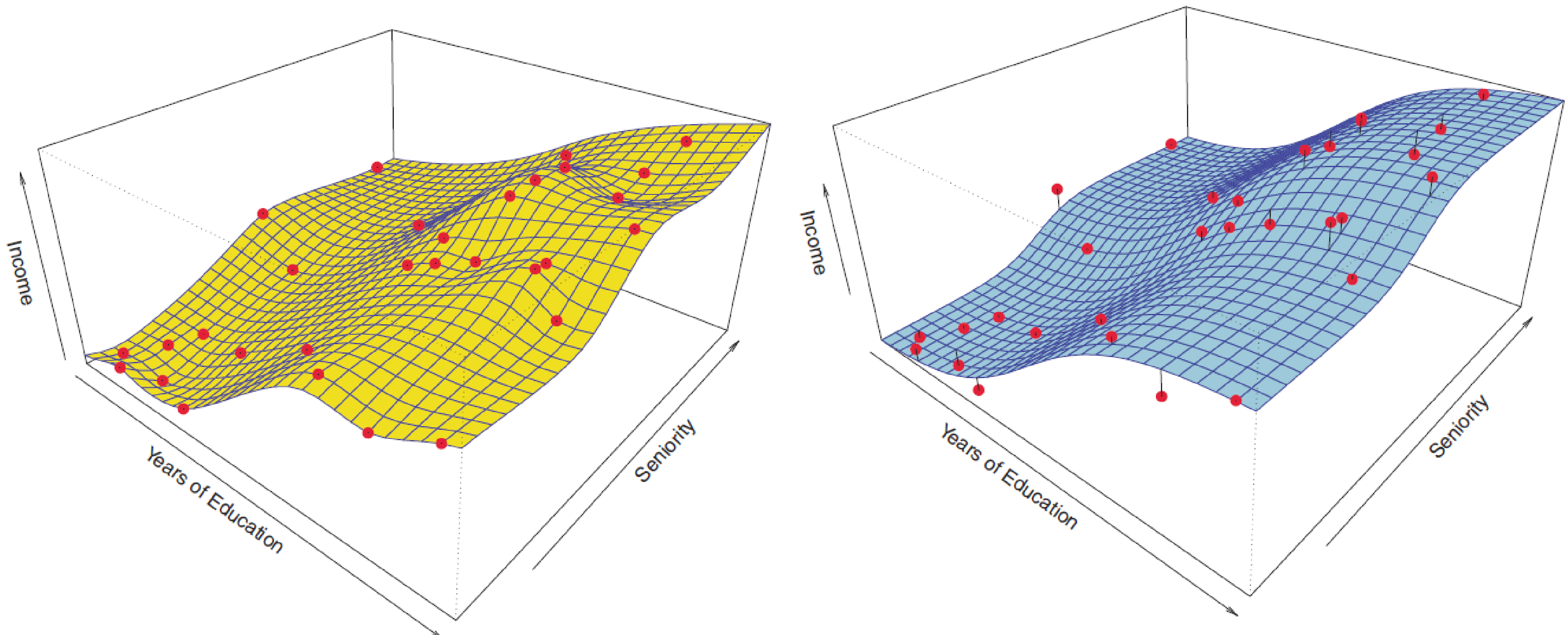
Los métodos no paramétricos no hacen suposiciones explícitas el tipo o la forma de ***f***. En cambio, buscan una estimación de f que **se acerque** tanto a la puntos de datos como sea posible sin ser demasiado ondulado o zigzagueante. Tales acercamientos puede tener una gran ventaja sobre los enfoques paramétricos: al evitar la suposición asunción de una forma funcional particular para f, tienen el potencial para adaptarse con mayor precisión a un rango más amplio de formas posibles para la desconocida **f**



Pero los enfoques no paramétricos sufren de desventajas, siendo una de ellas que requieren de **una gran cantidad de observaciones**

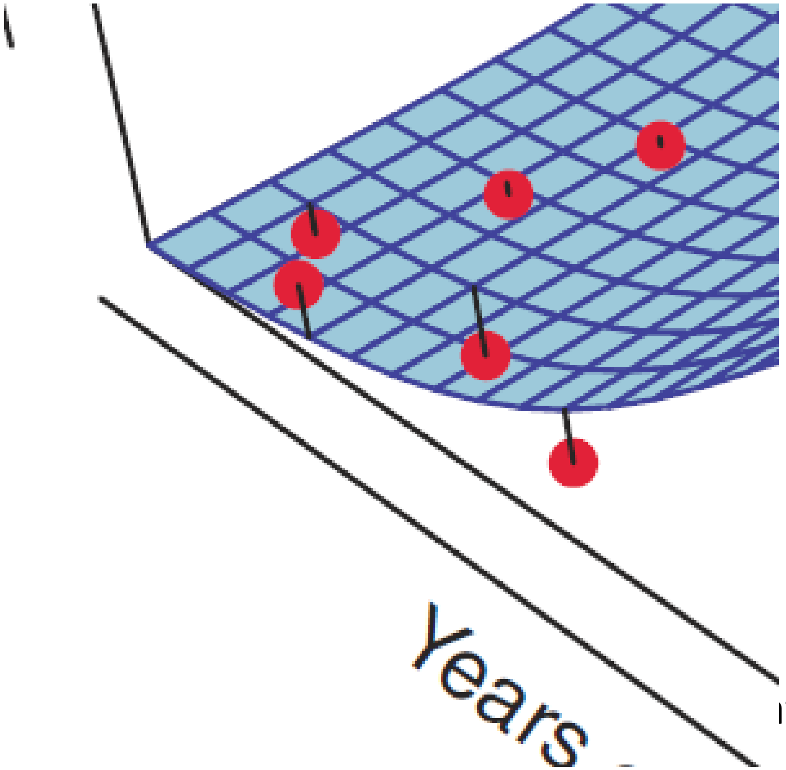

Income
Years of Education
Seniority

En este caso, el ajuste no paramétrico ha producido una notablemente precisa estimación de la verdadera f. ¡La estimación resultante se ajusta perfectamente a los datos observados! Sin embargo, el ajuste spline que se muestra en la Figura de la izq es mucho más variable que el verdadera función **f** de la derecha

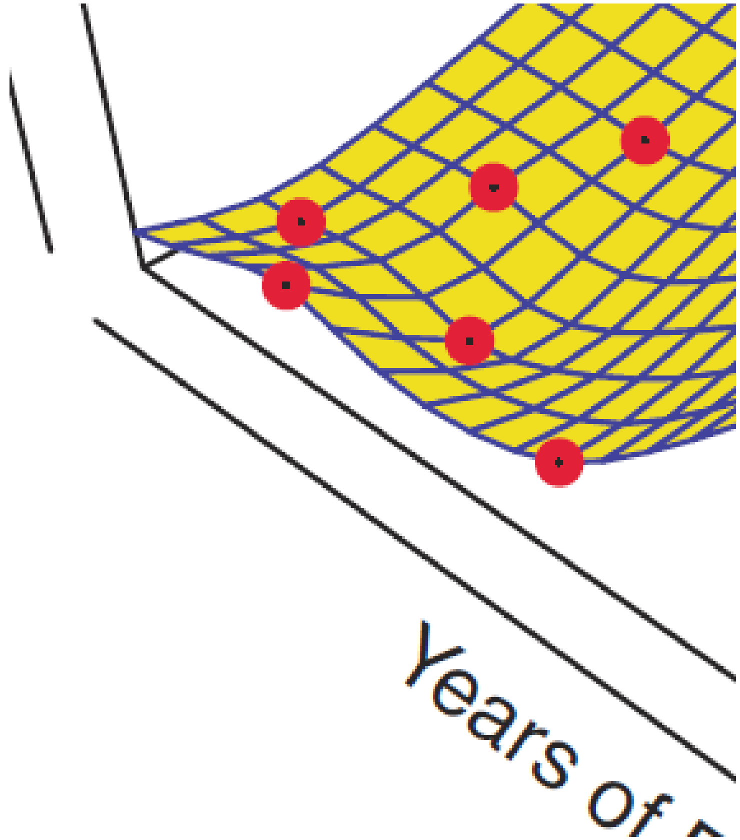


sobreajuste de los datos
OVERFITTING

Sobreajuste *(overfitting)*

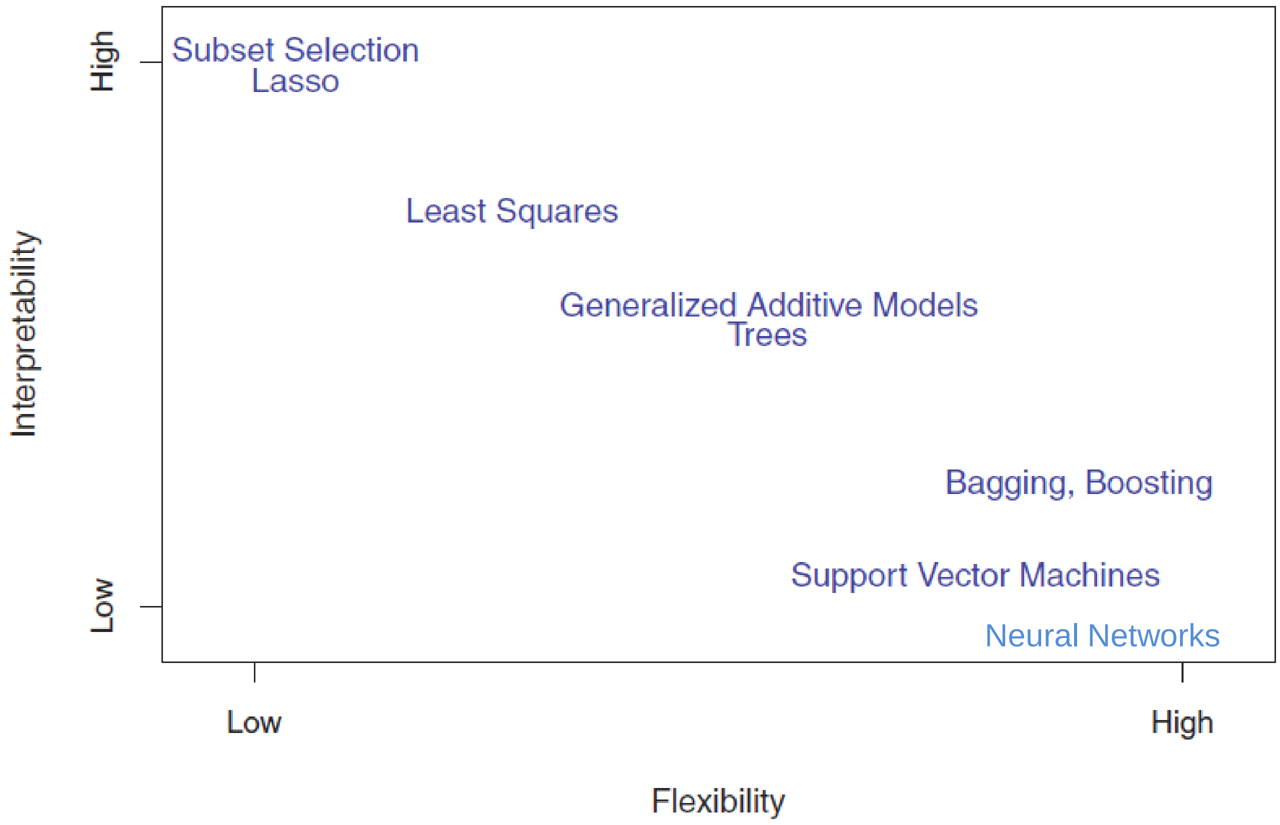


ta

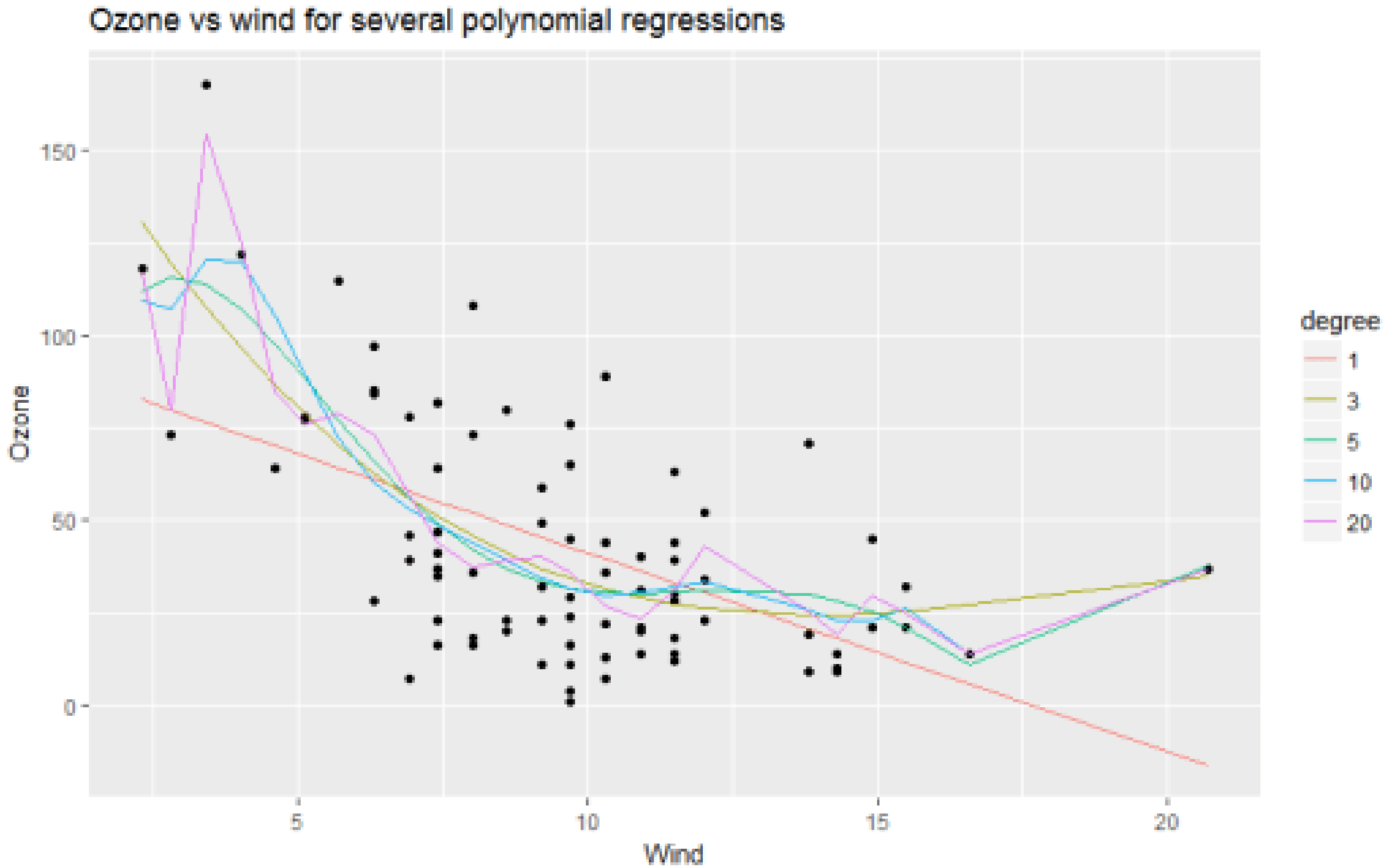


exactitud de la predicción vs interpretabilidad del modelo

Uno podría razonablemente hacer la siguiente pregunta: ¿por qué elegir usar un método más restrictivo en lugar de un enfoque muy flexible?



Inferencia → restrictivo
Predicción → flexible



Modelos de Machine Learning

Aprender de los datos

Grandes volumenes de datos *(capacidad computacional*

Predicción