Introducción a la analítica

Profesores César Augusto Gómez, Mauricio Alejandro Mazo y Juan Carlos Salazar

Quote of the Day

"Big data is at the foundation of all of the megatrends that are happening today, from social to mobile to the cloud to gaming." – Chris Lynch

Para el cliente no es posible incrementar directamente las ventas del producto, pero si puede controlar el gasto en publicidad en cada uno de los tres medios. Por lo tanto, si usted puede determinar que hay una asociación entre publicidad y ventas, le puede informar al cliente que ajuste los presupuestos de publicidad y así incrementar las ventas. El objetivo es contruir un modelo preciso que se pueda usar para predecir las ventas con base en los tres presupuestos de medios publicitarios.

Se muestran Ventas vs TV, Radio y Periódico, con una línea de regresión lineal púrpura de regresión lineal ajustada por separado para cada input ¿Se podrá predecir las ventas utilizando solo unos de estos tres inputs? o ¿Se podrá predecir las ventas utilizando estos tres?

Quizás se pueda hacer mejor usando un modelo:

$$Ventas \approx f(TV,\ RADIO,\ PERIODICO)$$

Aquí Ventas es la respuesta u objetivo que deseamos predecir (la Y). TV es una característica, input, o predictor; se denotará como X_1. De igual manera Radio como X_2 y periódico como X_3. Podemos referirnos al vector de entradas colectivamente como $X = (X_1, X_2, X_3)$ y escribir el modelo

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

donde ε captura errores de medición y otras discrepancias.

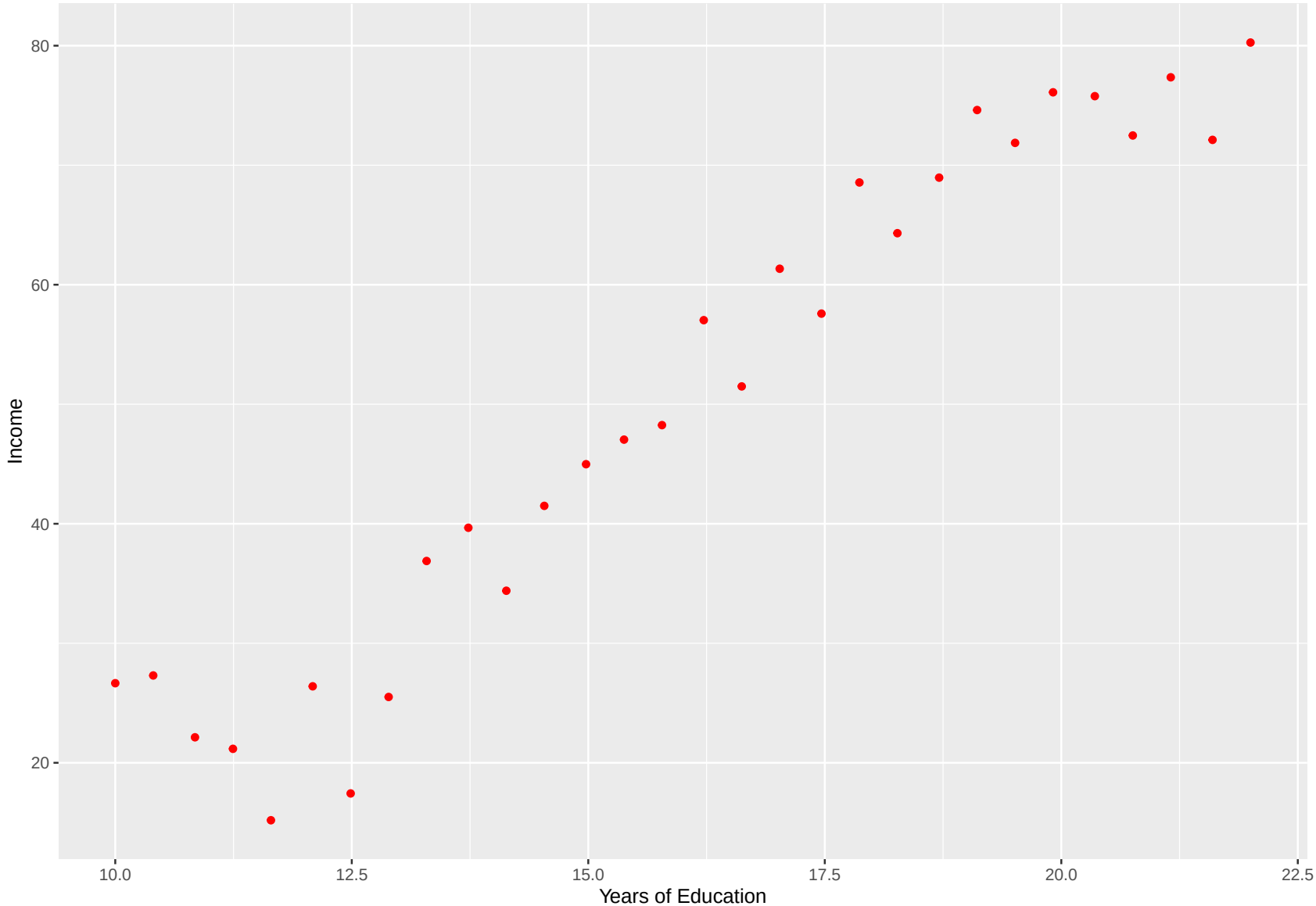
Más generalmente, suponga que se observa una respuesta cuantitativa Y y p predictores distintos, $X_1, \ldots, X_p$. Asuma que hay alguna relación entre Y y $X = (X_1, \ldots, X_p)$, la cual se puede escribir en una forma muy general como:

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

Aquí, f es una función fija pero desconocida de X y ε captura errores de medición y otras discrepancias, se conoce como término de error aleatorio, el cual es independiente de X y tiene media igual a cero.

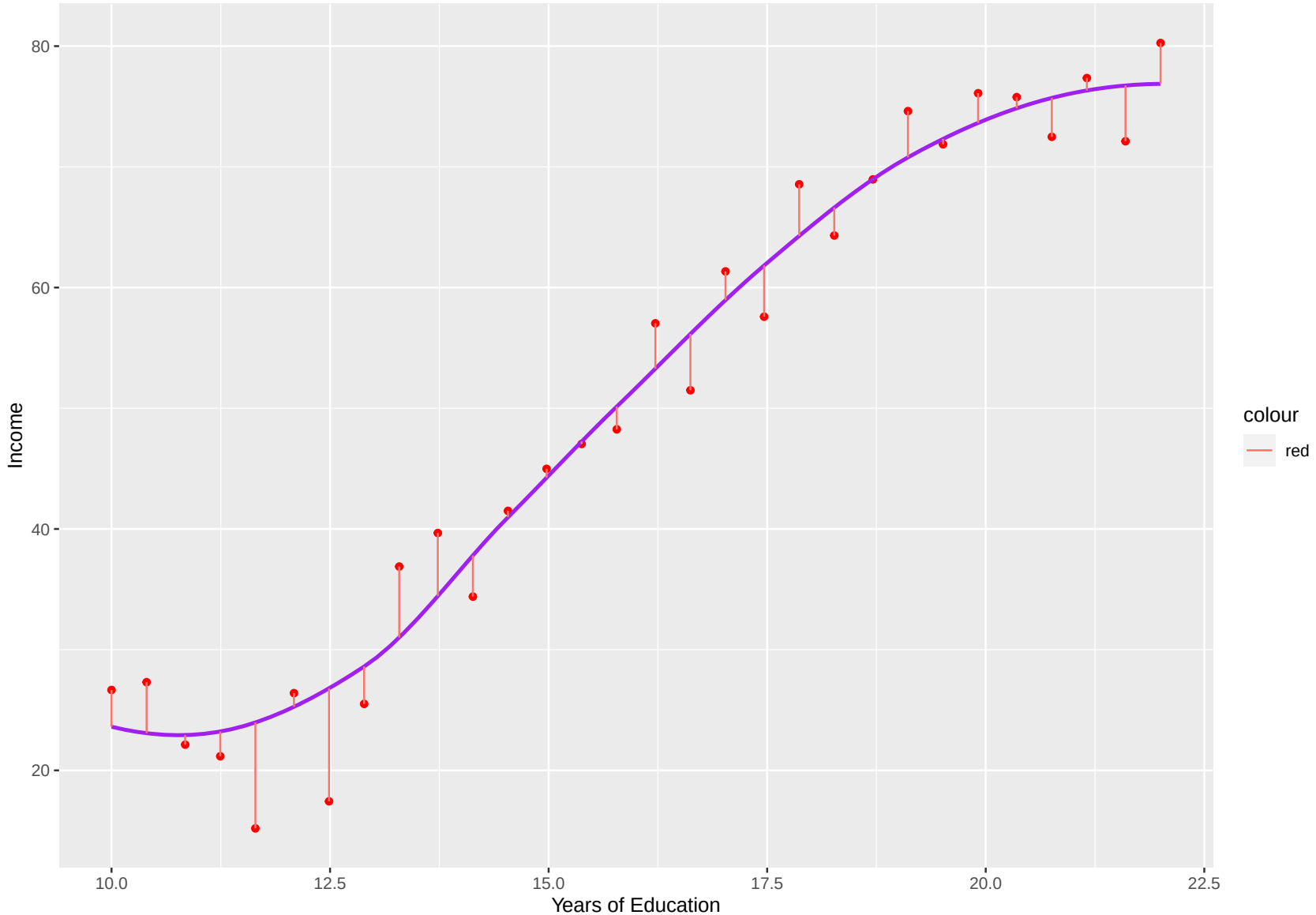
En esta formulación, f representa la información sistemática (componente sistemática) que X proporciona acerca de Y. Considere otro ejemplo relacionado con la base de datos llamada ingreso (income dataset), donde se tiene información de 30 personas respecto a ingreso(income, la y) y años de educación (la x):

Aprendizaje estadístico



El gráfico anterior sugiere que el ingreso se puede predecir con los años de educación, sin embargo la función f que conecta la variable input con el output es en general desconocida y debe ser estimada con los datos:

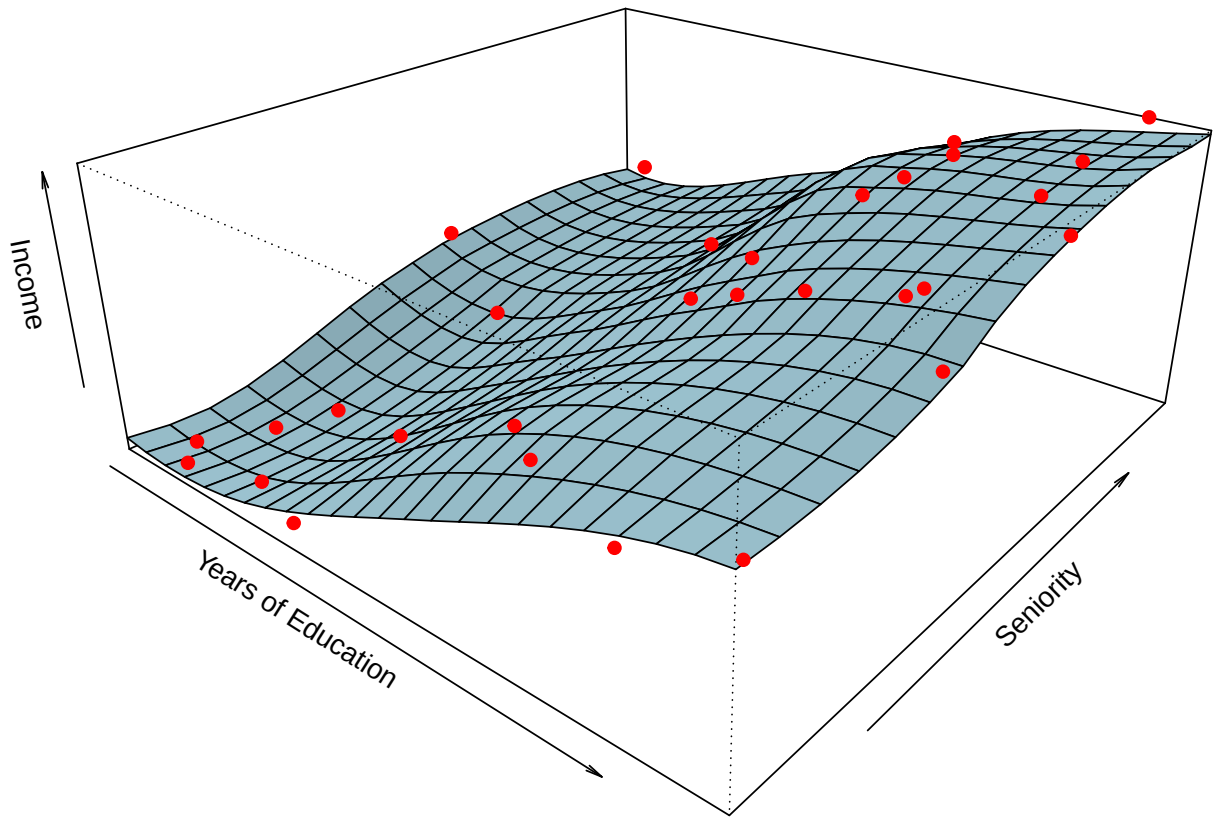
Aprendizaje estadístico



En el gráfico anterior, las líneas verticales representan los términos de error. Se observa que algunas de las 30 observaciones caen encima y algunas por debajo de la línea ajustada. Globalmente, los errores tienen media cero.

En general, la función f puede involucrar más de una variable input. En la siguiente figura aparece ingreso en función de años de educación y antigüedad (seniority). En este caso, f es una superficie bidimensional que debe ser también estimada usando los datos.

Aprendizaje estadístico



Esencialmente, AE se refiere a un conjunto de aproximaciones para estimar la función f. Pero, Porqué querer estimar a f? **Hay dos razones importantes para estimar a** f**: INFERENCIA Y PREDICCIÓN**.

PREDICCIÓN. En muchas situaciones, un conjunto de inputs están disponibles, pero el output Y no se puede obtener de una manera fácil. En esta situación y teniendo en cuenta que el término de error es cero, se puede predecir a Y usando

$$\widehat{Y} = \widehat{f}\,(X)$$

Donde $\widehat{f}$ representa una estimación de f y $\widehat{Y}$ representa la predicción resultante para Y.

En este tipo de situaciones, f frecuentemente se trata como una *caja negra*, en el sentido de que no se está interesado en la forma exacta de f dado que produzca predicciones **precisas** y adecuadas de Y.

La **precisión** en la predicción de Y usando $\widehat{Y}$ depende de dos cantidades muy importantes que se conocen como **ERROR REDUCIBLE** y **ERROR IRREDUCIBLE**. En general $\widehat{f}$ no es un estimador perfecto de f y esta imprecisión introduce algo de error. **Este error es reducible ya que, potencialmente, se puede mejorar la precisión de $\widehat{f}$ usando una mejor técnica de AE para estimar a f.**

Sin embargo, **aún si fuera posible obtener un estimador perfecto para f, de tal manera que $\widehat{Y} = f(X)$ esta estimación todavía tendría algo de error en ella debido a que Y es función de ε el cual por definición no se puede predecir usando X.**

Por lo tanto, la variabilidad asociada con ε también afecta la precisión de las estimaciones. Este se conoce como el **error irreducible**, ya que sin importar que tan bien se estima a f, no se puede reducir el error introducido por ε (error aleatorio).

¿Porqué este error irreducible es mayor a cero?[1] La cantidad ε puede contener variables no medidas que pueden ser útiles en la predicción de Y; puesto que no se miden, f no las puede usar en la predicción, La cantidad ε también puede contener variación no medible (por ejemplo la propensión de una persona a enfermar)

[1] Es lo mismo que decir que porqué $\text{var}(\varepsilon) > 0$, es decir, no se admite la posibilidad de que sea igual a cero ya que en general puede contener variación no medible.

Considere un estimador $\widehat{f}$ y un input X con los que se obtiene la predicción $\widehat{Y} = \widehat{f}(X)$. Asuma de momento que $\widehat{f}$ y X son fijos. entonces, en este caso

$$E\left(Y - \widehat{Y}\right)^2 = E\left[f(X) + \varepsilon - \widehat{f}(X)\right]^2$$
$$= \underbrace{\left[f(X) - \widehat{f}(X)\right]^2}_{\text{Reducible}} + \underbrace{var(\varepsilon)}_{\text{Irreducible}}$$

(EJERCICIO). Aquí, $E\left(Y - \widehat{Y}\right)^2$ es el valor esperado de la diferencia al cuadrado entre el valor observado para Y y el valor predicho para Y, mientras que $var(\varepsilon)$ es la varianza del error.

Uno de los objetivos principales, será presentar técnicas para estimar a f que ayuden a minimizar el error reducible. Tenga en mente que el error irreducible siempre producirá una cota superior para la predicción $\widehat{Y}$ de Y.

Considere una compañía que quiere identificar individuos que respondan de manera positiva a una campaña por email con base en información demográfica. La compañía no está interesada en las relaciones entre los predictores de cada sujeto y su respuesta, solo quiere un modelo para predecir la respuesta usando los predictores. Este es un ejemplo de modelamiento para **Predicción**.

INFERENCIA. Frecuentemente interesa la manera en que Y se afecta con cambios en el input. En esta situación, se quiere estimar a f pero el objetivo no es hacer predicciones para Y. Lo que se quiere en cambio, es entender la relación entre X y Y, es decir, entender cómo Y cambia en función de X. Por lo tanto, f no puede tratarse como una caja negra ya que se necesita conocer su forma exacta. Se quiere dar respuesta a las siguientes preguntas:

- ¿Cuáles predictores están asociados con el respuesta?
- ¿Cuál es la relación entre la respuesta y cada predictor?
- ¿Puede la relación entre Y y cada predictor resumirse de manera adecuada usando una ecuación lineal o es esta relación más compleja?

Considere los datos de publicidad. Algunas preguntas de interés incluyen:

- ¿Cuál medio publicitario contribuye a vender?
- ¿ Cuál medio genera un mayor incremento en las ventas?
- ¿Qué tanto incremento en ventas está asociado con un incremento dado en cada medio?

Este es un problema de **Inferencia**.

Los modelos lineales permiten una inferencia simple e interpretable pero podrían no producir predicciones precisas en comparación a otros métodos. En contraste, algunos de los métodos altamente no lineales podrían producir mejores predicciones que un modelo lineal pero su interpretabilidad puede ser muy compleja y retadora.

¿Cómo estimar a f?. Asuma que se tiene un conjunto de datos conformado por n puntos diferentes, que se denominarán **Conjunto de entrenamiento** (Training Data) ya que se usarán para **entrenar** (o **enseñar**) el método para estimar a f.

Sea x_{ij} el valor de j-ésimo input o predictor, de la observación i, donde $i = 1, 2, \ldots, n$ y $j = 1, 2, \ldots, p$. Tambien, sea y_i la respuesta de la i-observación. Por lo tanto, el conjunto de entrenamiento es:

$$\{(x_1, y_1), \ldots, (x_n, y_n)\}$$

con

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$$

Se quiere aplicar un método estadístico (paramétrico o noparamétrico) a este conjunto de entrenamiento a fin de estimar la función f. Es decir, se quiere hallar $\widehat{f}$ tal que $Y = \widehat{f}$ dados (X, Y).

Métodos paramétricos. Estos métodos consideran una aproximación basada en un modelo de dos etapas o pasos.

- **Primero**, se hace un supuesto acerca de la forma funcional para f. Por ejemplo, un supuesto muy simple, es que f es lineal en X:

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \quad \text{(Modelo Lineal)}$$

Una vez se asume que f es lineal, el problema de estimar a f se simplifica enormemente ya que en vez de estimar toda la función p-dimensional solo se requiere estimar $p+1$ coeficientes $\beta_0,\ \beta_1, \ldots, \beta_p$.

- **Segundo**, después de que el modelo ha sido seleccionado, se necesita un procedimiento que use los datos de entrenamiento para ajustar o entrenar el modelo. Se quiere encontrar estimaciones de los coeficientes $\beta_0,\ \beta_1, \ldots, \beta_p$ tales que

$$Y \approx \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

El método más común es el de **Mínimos cuadrados ordinarios** (OLS), pero hay otros disponibles.

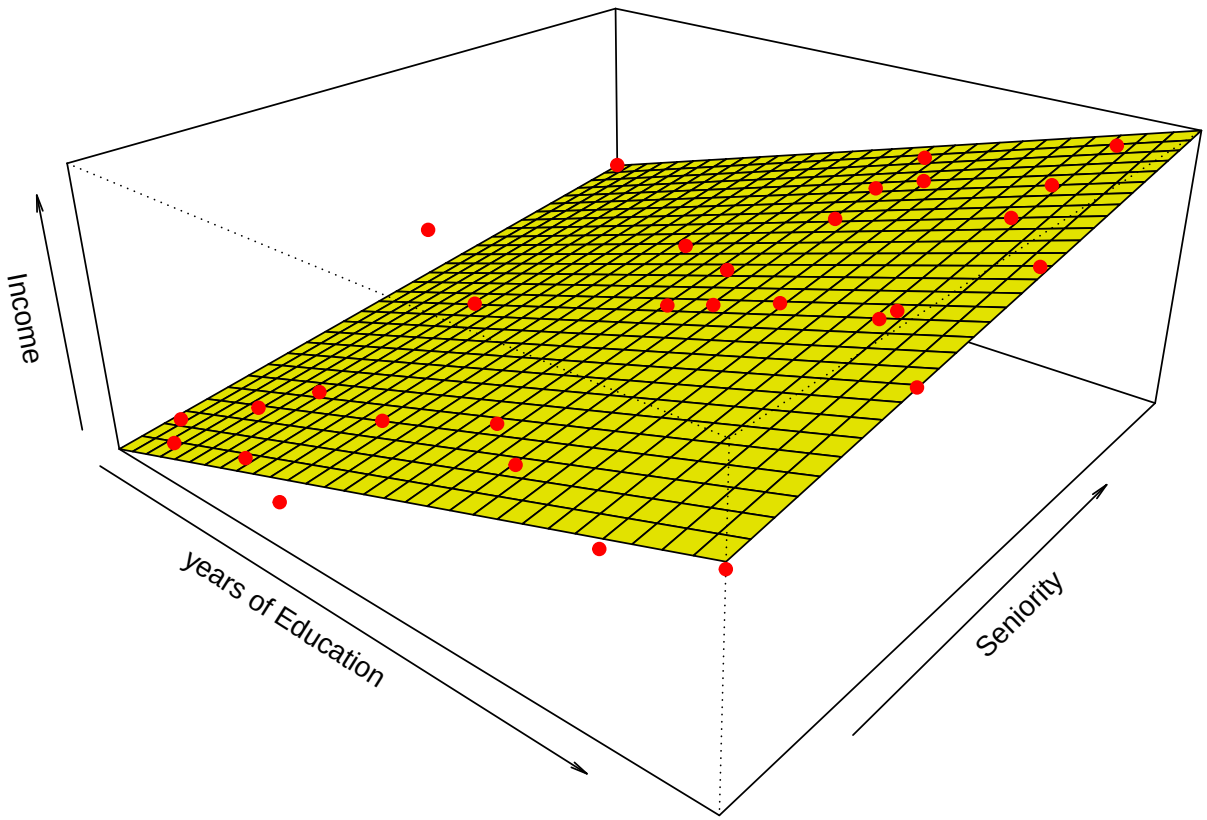
Este método es paramétrico en el sentido de que reduce el problema de estimar a f a uno de estimar un conjunto de parámetros $\beta_0,\ \beta_1, \ldots, \beta_p$.

Asumir una forma paramétrica para f simplifica el problema de estimar a f ya que en general es más sencillo estimar un conjunto de parámetros que una función completa. La desventaja potencial, es que el modelo que se asume usualmente no coincide con la forma desconocida de f. Si el modelo elegido está muy lejano de la f real la estimación será muy deficiente y pobre. Este problema se puede enfrentar escogiendo modelos flexibles que abarquen muchas posibles formas funcionales para f. Pero en general ajustar un modelo más flexible requiere estimar un mayor número de paramétros y esto puede llevar a un fenómeno conocido como **sobreajuste** u **overfitting** (se explica el ruido y no la señal, sigue muy de cerca el error).

La siguiente figura muestra un ejemplo de una aproximación paramétrica a los datos de ingreso. Se ajusta un modelo de la forma:

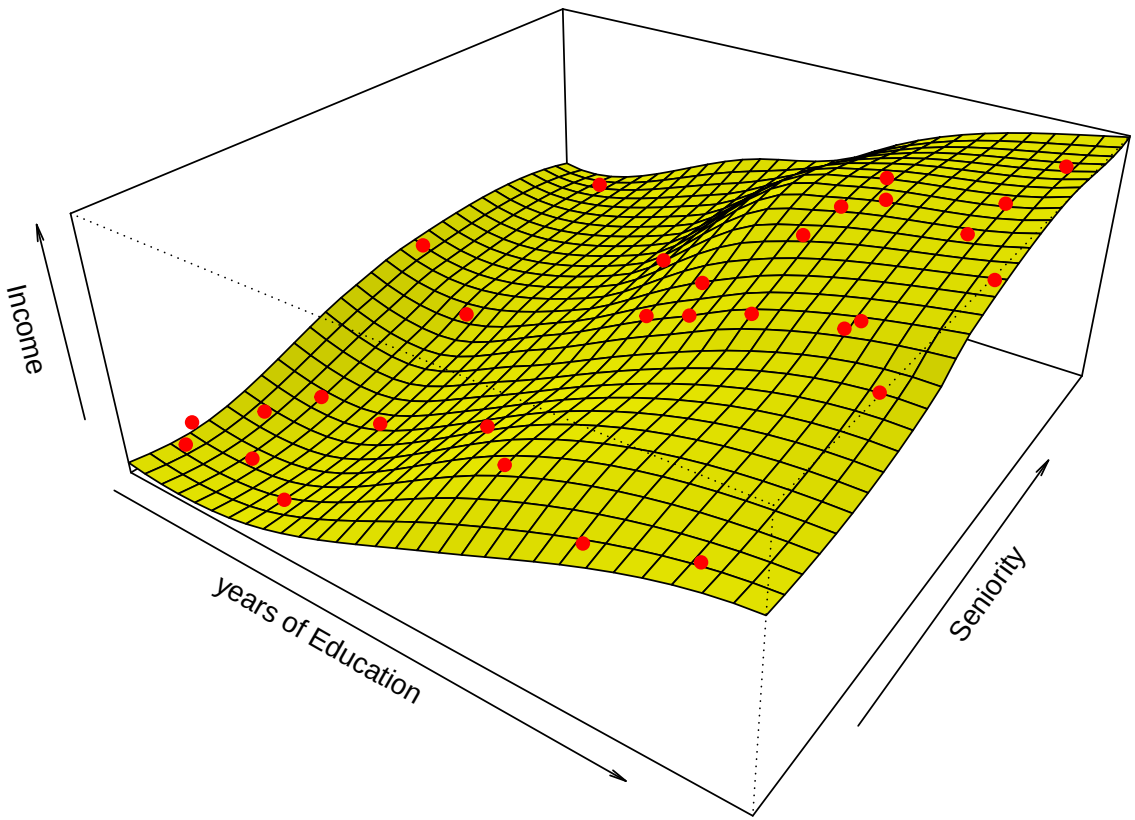
$$\mathit{Income} \approx \beta_0 + \beta_1 \times \mathit{Education} + \beta_2 \times \mathit{Seniority}$$

Puesto que se asume una relación lineal entre la respuesta y los dos inputs, el problema de ajuste se reduce a estimar el vector $(\beta_0,\ \beta_1,\ \beta_2)$ usando OLS. El ajuste no es del todo correcto ya que la superficie original tiene algo de curvatura que no la captura este plano. Sin embargo, parece hacer un trabajo razonable al capturar la asociación positiva entre años de educación y antigüedad, y entre antigüedad e ingreso.

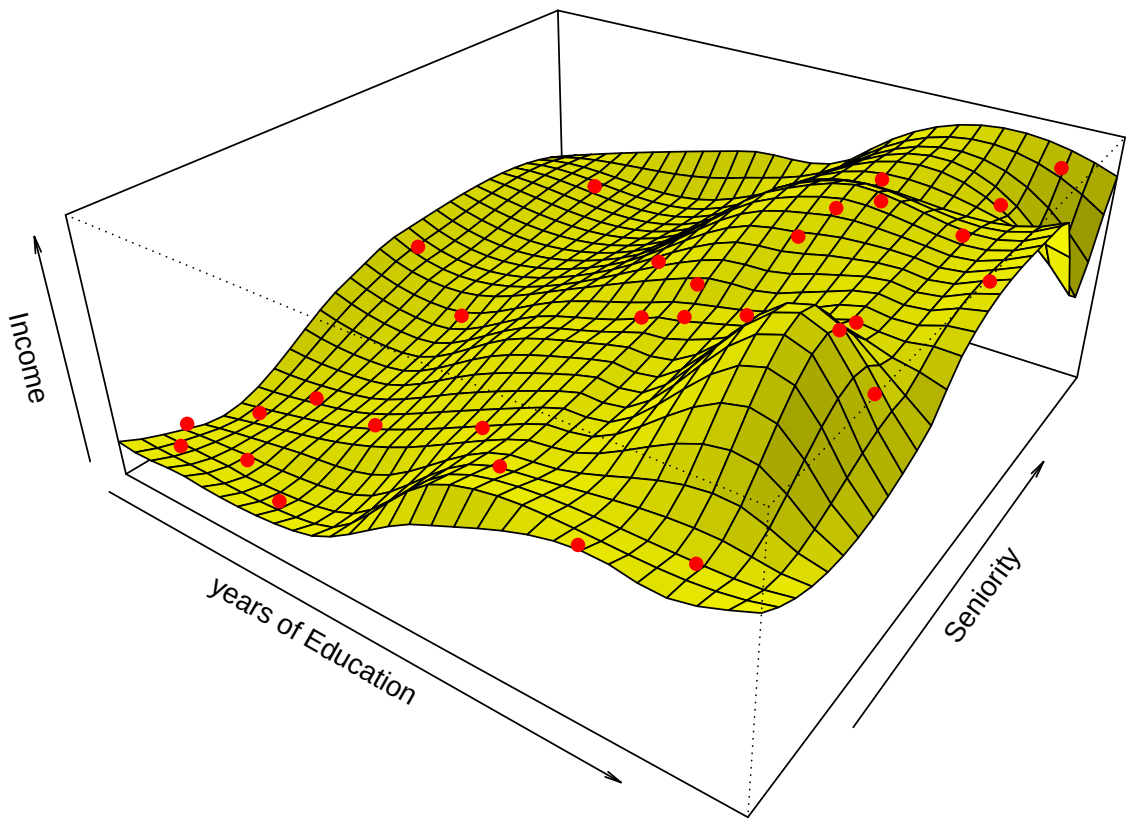

Income
years of Education
Seniority

Métodos noparamétricos. Estos métodos no asumen explícitamente una forma funcional para f. En vez de hacerlo, buscan una estimación de f que se acerque tanto a los datos como sea posible pero de una forma suave (no rugosa). Tienen una ventaja sobre los métodos paramétricos: **al evitar una forma funcional para f, tienen el potencial de ajustar un rango más amplio de posibles formas para f**. Sin embargo, tienen una desventaja: **puesto que no reducen la estimación de f a un número pequeño de parámetros, requieren usualmente muchos datos para estimar con precisión a f**. Ejemplo con datos de ingreso (Income dataset)

Con el nivel de suavizamiento (usando splines suavizados) parece obtenerse un ajuste razonablemente bueno:



El analista debe seleccionar el nivel se suavizamiento. Más adelante se discuten métodos para seleccionar el nivel de suavizamiento óptimo. Poco suavizamiento lucira así (explica la tendencia del ruido y no de la señal, rugosidad), este es un ejemplo de **overfitting** o **sobre-**



Aprendizaje estadístico: El intercambio entre precisión e interpretabilidad

Algunos métodos son menos flexibles o muy restrictivos en el sentido de que pueden producir solo un rango limitado de formas para f. Por ejemplo, la regresión lineal es un método relativamente inflexible ya que solo puede generar lineas o planos. Otros métodos como los splines[2], son considerablemente más flexibles ya que pueden generar un rango de formas más amplio para f que los modelos lineales.

[2] Ruppert, D., Wand, M.P., and Carroll. Semiparametric Regression. Cambridge Series in statistical and probabilistic mathematics. 2003

Aprendizaje estadístico: El intercambio entre precisión e interpretabilidad

Pero ¿Porqué seleccionar un método restrictivo en vez de un método bien flexible? Hay algunas razones por las que uno podría preferir un modelo restrictivo. **Si el interés es INFERENCIA, entonces los modelos más restrictivos son más interpretables**. Por su parte, **los modelos más flexibles, son en general más dificiles de interpretar y la forma en que los inputs se asocian con la respuesta no es clara**.

Aprendizaje estadístico: El intercambio entre precisión e interpretabilidad

El siguiente gráfico (James et al. 2014) muestra una ilustración del intercambio (tradeoff) entre flexibilidad e interpretabilidad[3]:

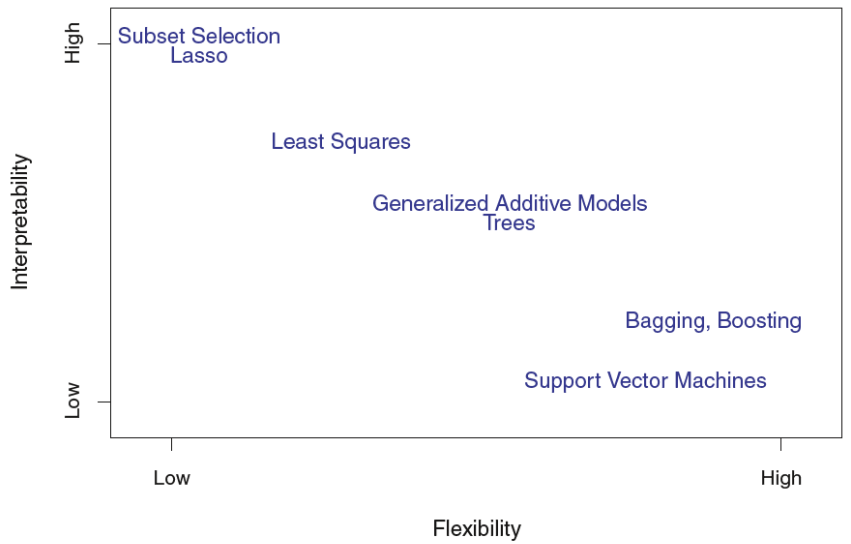


FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

Figura 1: Flexibilidad versus Interpretabilidad

[3]Taken from "An Introduction to Statistical Learning, with applications in R"

Aprendizaje estadístico: El intercambio entre precisión e interpretabilidad

Como una regla de dedo se puede usar la siguiente:

Cuando el objetivo es INFERENCIA es más ventajoso optar por métodos de AE simples y relativamente inflexibles. Si el objetivo es PREDICCIÓN, y la interpretabilidad no es de interés primario se opta por modelos de AE más flexibles pero prestando atención al potencial sobreajuste.

Aprendizaje supervisado versus no supervisado

Muchos problemas de AE se pueden clasificar en una de dos categorías: **Supervisados** y **No supervisados**. Muchos métodos clásicos de AE tales como regresión lineal y regresión logística, así como métodos modernos tales como GAM, boosting, y máquinas de soporte vectorial, operan en el dominio de **aprendizaje supervisado**. **Por cada observación del predictor x_i hay una medición de la respuesta asociada y_i, $i = 1, 2, \ldots, n$.**

Aprendizaje supervisado versus no supervisado

En contraste, en **el aprendizaje no supervisado se observa un vector de mediciones x_i pero no se observa una respuesta asociada y_i, es decir no hay una respuesta para supervisar el análisis**. Una herramienta que frecuentemente se usa en estas situaciones es el análisis de clusters (o agrupamientos, clustering). El objetivo del clustering es asegurar, con base en $x_1, x_2, \ldots, x_n$, si las observaciones se clasifican en grupos relativamente diferentes.

Aprendizaje supervisado versus no supervisado

Por ejemplo, en un estudio de segmentación de mercados se pueden observar múltiples características para consumidores potenciales, tales como código ZIP, ingreso familiar, y hábitos de compra. Se podría pensar que los consumidores se clasifican como grandes consumidores y pequeños consumidores. Si la información acerca de los patrones de gasto estuviera disponible (gran consumidor versus pequeño consumidor, $y = 0$ o $y = 1$), se podría llevar a cabo un análisis supervisado, pero si no está disponible, se puede tratar de agrupar los consumidores de acuerdo a las variables medidas a fin de identificar distintos grupos de consumidores.

Aprendizaje supervisado versus no supervisado