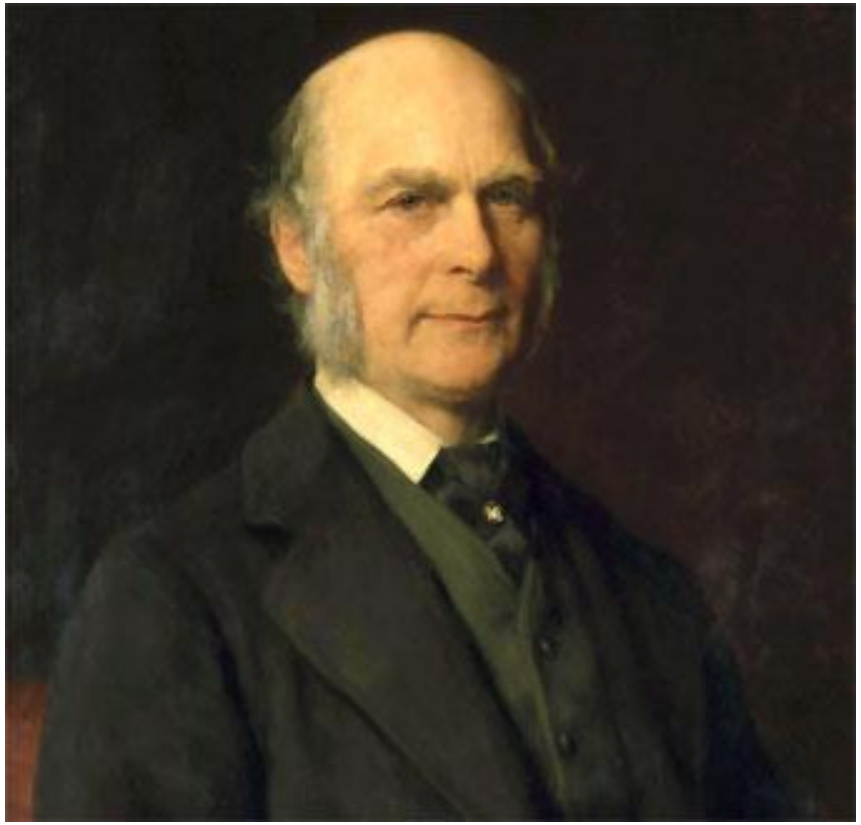
Introducción a la Regresión Lineal

Lectura Sugerida

- Capítulos 2 y 3 del libro “Introduction to Statistical Learning” de Gareth James

Historia

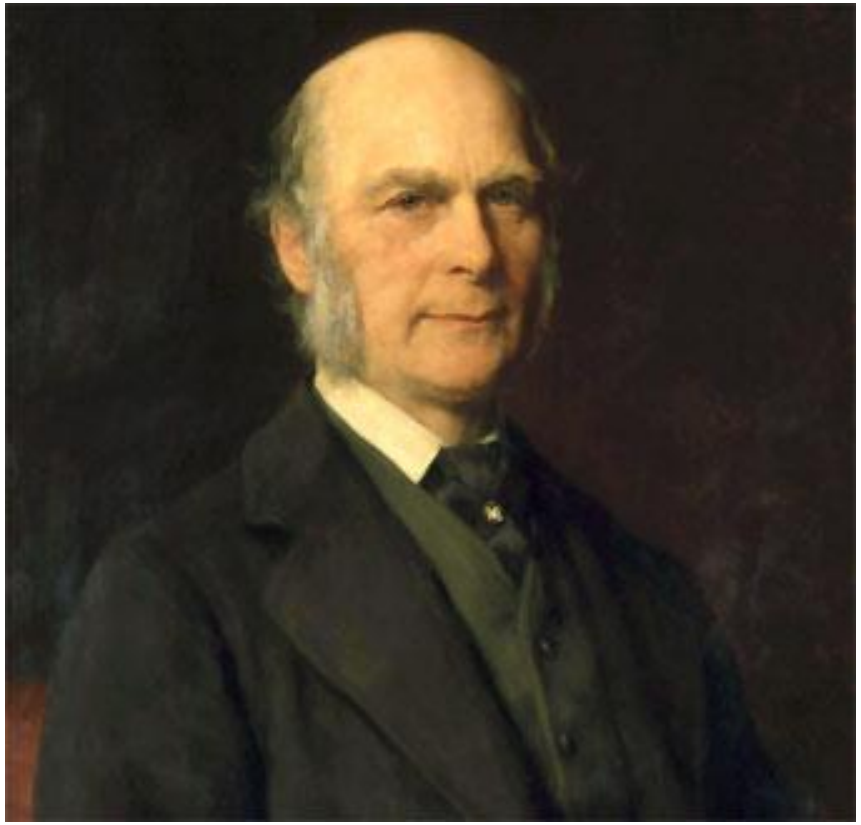
Todo comenzó en el siglo XIX con un tipo llamado Francis Galton. Galton estaba estudiando la relación entre los padres y sus hijos. En particular, investigó la relación entre las alturas de los padres y sus hijos.



Historia

Lo que descubrió fue que el hijo de cualquier hombre tendía a ser más o menos tan alto como su padre.

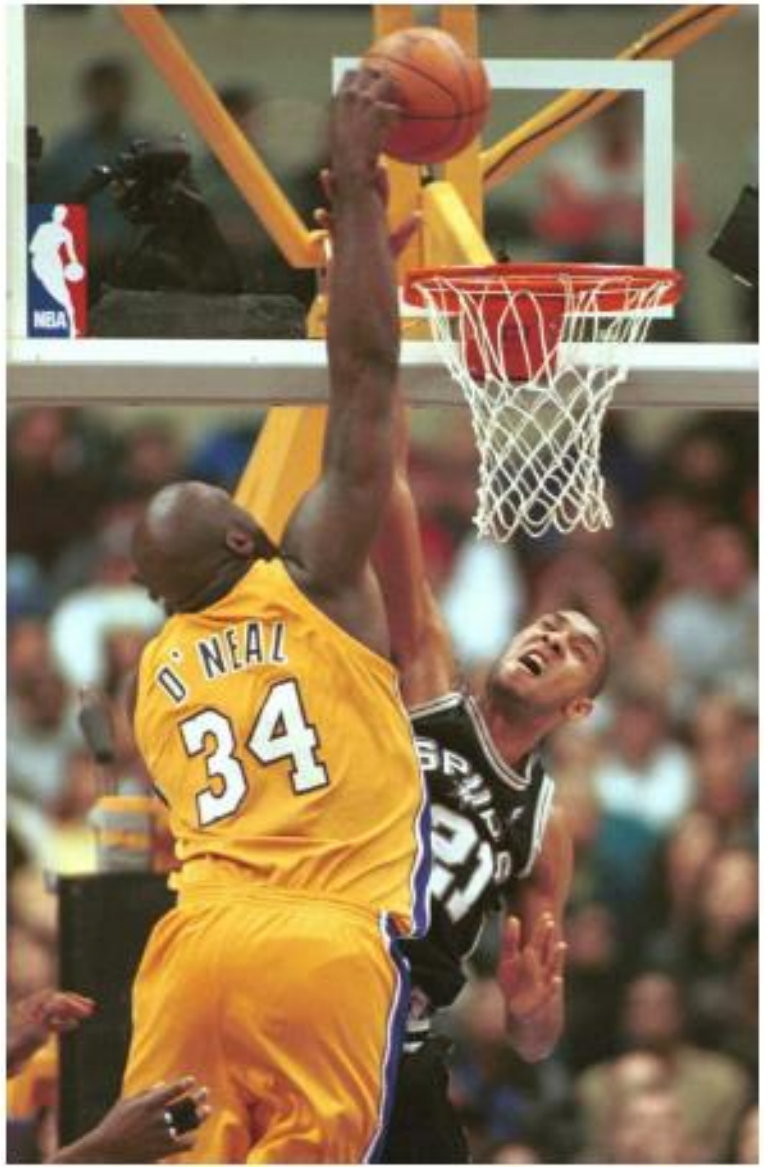
Sin embargo, el descubrimiento de Galton fue que la altura de un hijo tendía a estar más cerca de la estatura promedio general de todas las personas.



Ejemplo

Tomemos a Shaquille O'Neal como ejemplo. Shaq es realmente alto: 7 pies 1 pulgada (2,16 metros).

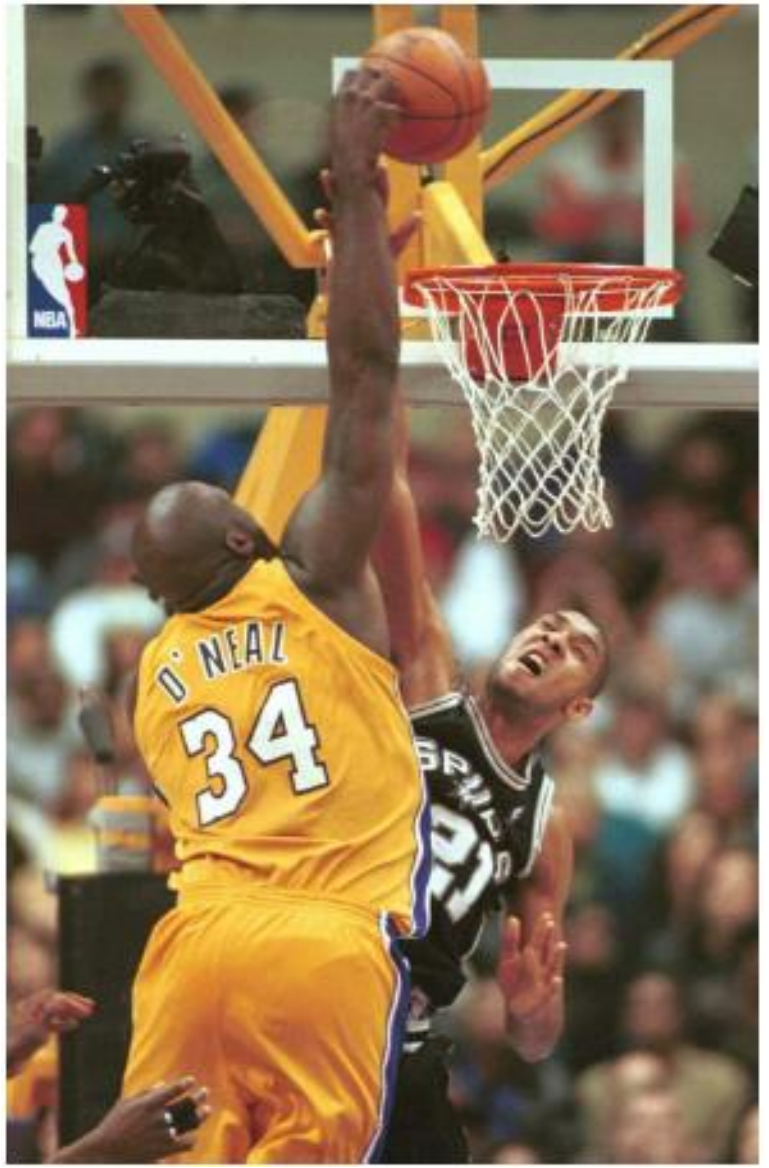
Si Shaq tiene un hijo, es probable que sea bastante alto también. Sin embargo, Shaq es una anomalía, tal que también hay una gran posibilidad de que su hijo no sea tan alto como Shaq



Ejemplo

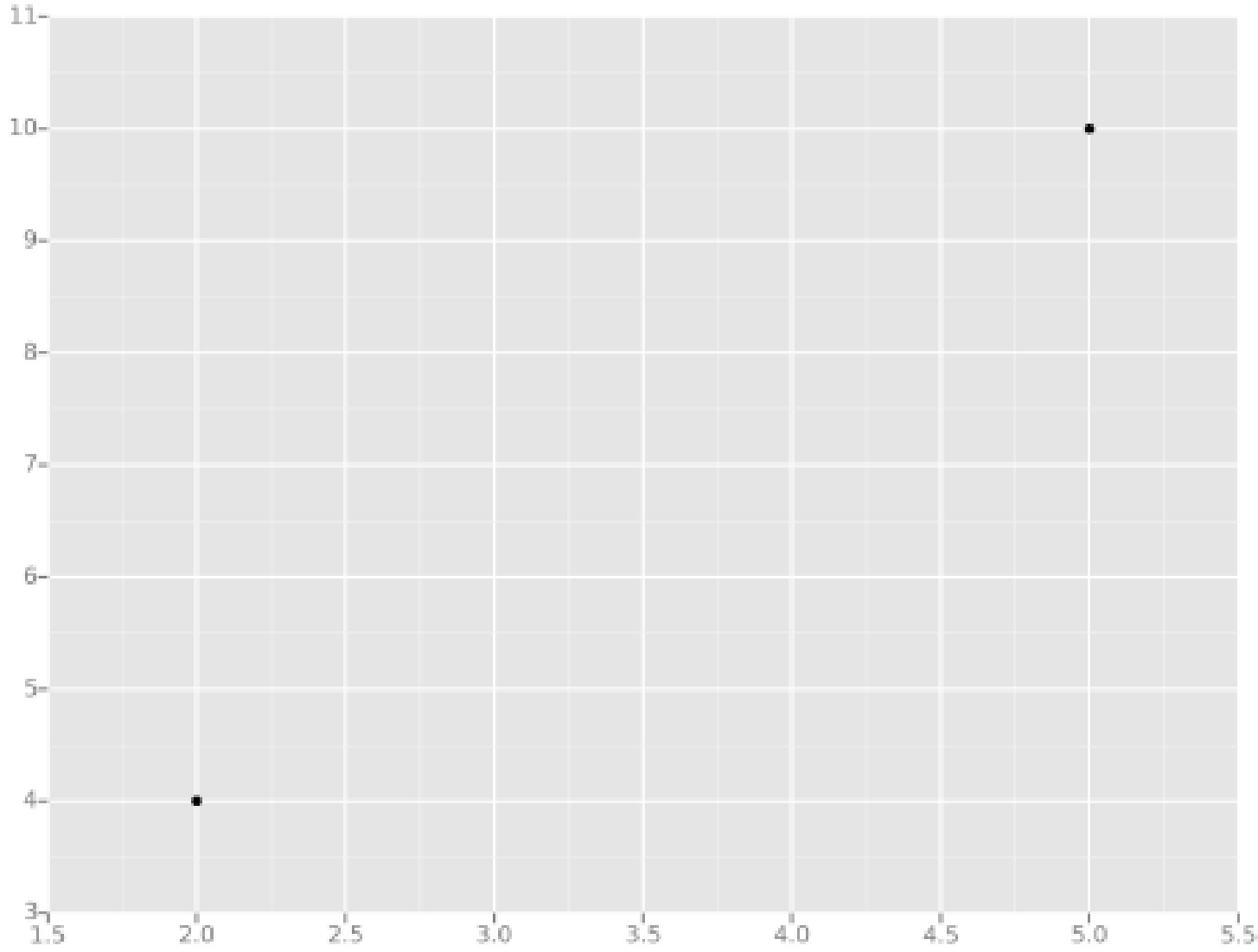
Resulta que este es el caso: el hijo de Shaq es bastante alto, 6 pies 7 pulgadas (2,0 metros), pero no tan alto como su padre.

Galton llamó a este fenómeno regresión, como en "La altura de un hijo de un padre tiende a retroceder (o deriva hacia) la altura media (promedio)".



Ejemplo

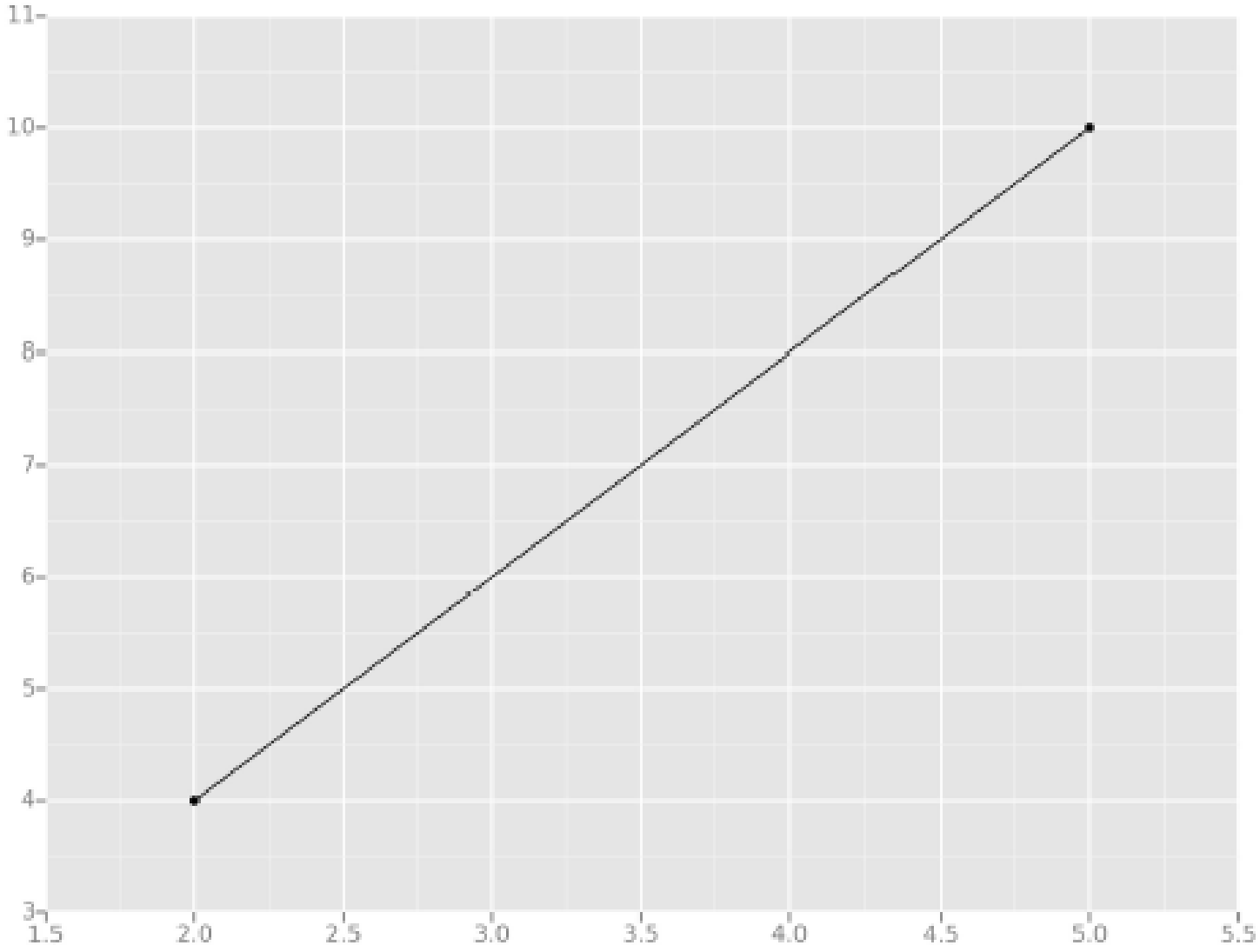
Tomemos el ejemplo más simple posible: calcular una regresión con solo 2 puntos de datos.



Ejemplo

Todo lo que intentamos hacer cuando calculamos nuestra línea de regresión es dibujar una línea lo más cercana posible a cada punto.

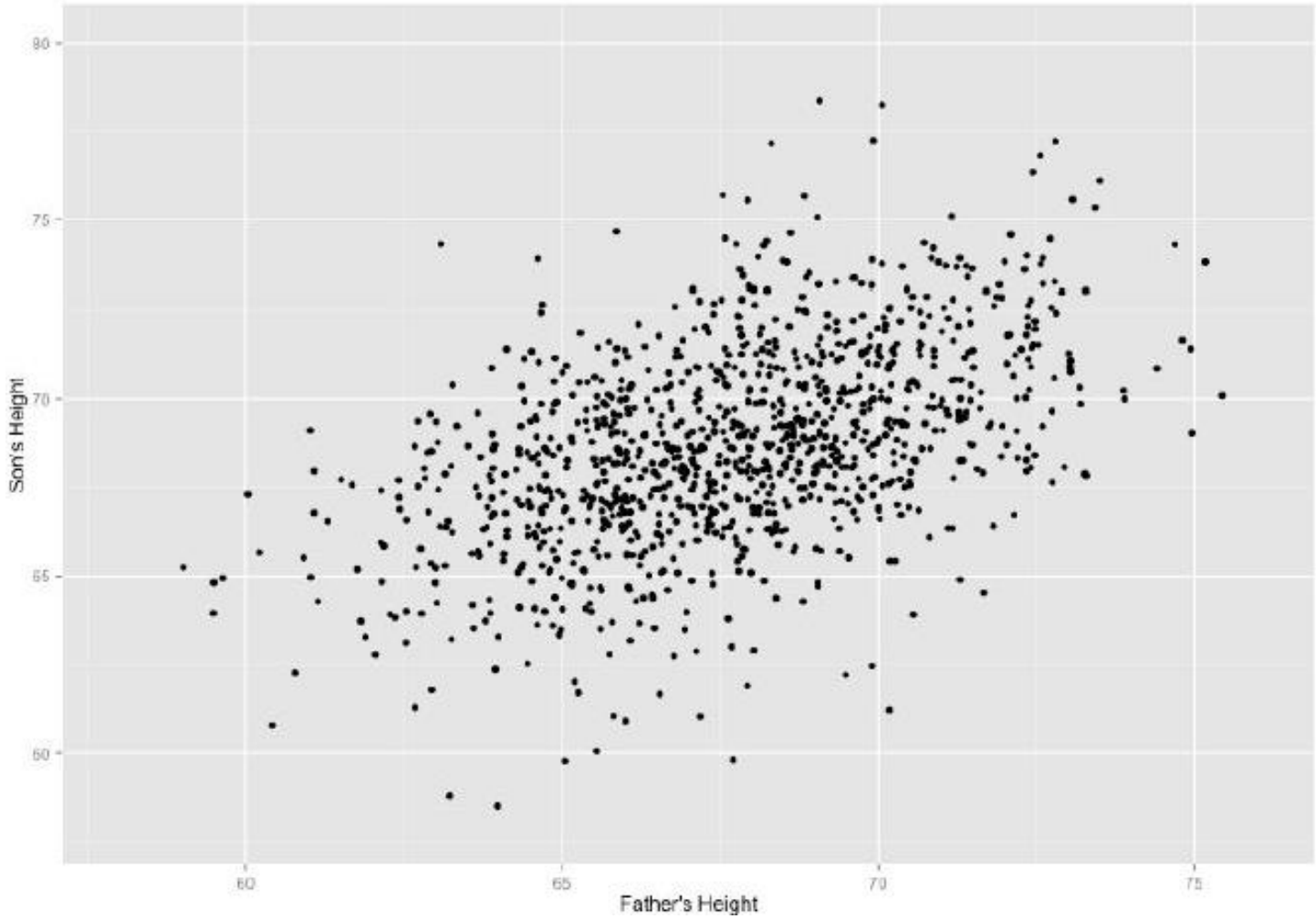
Para la regresión lineal clásica, o el "Método de mínimos cuadrados", solo se mide la cercanía en la dirección "arriba y abajo"



Ejemplo

Ahora, ¿no sería genial si pudiéramos aplicar este mismo concepto a un gráfico con más de dos puntos de datos?

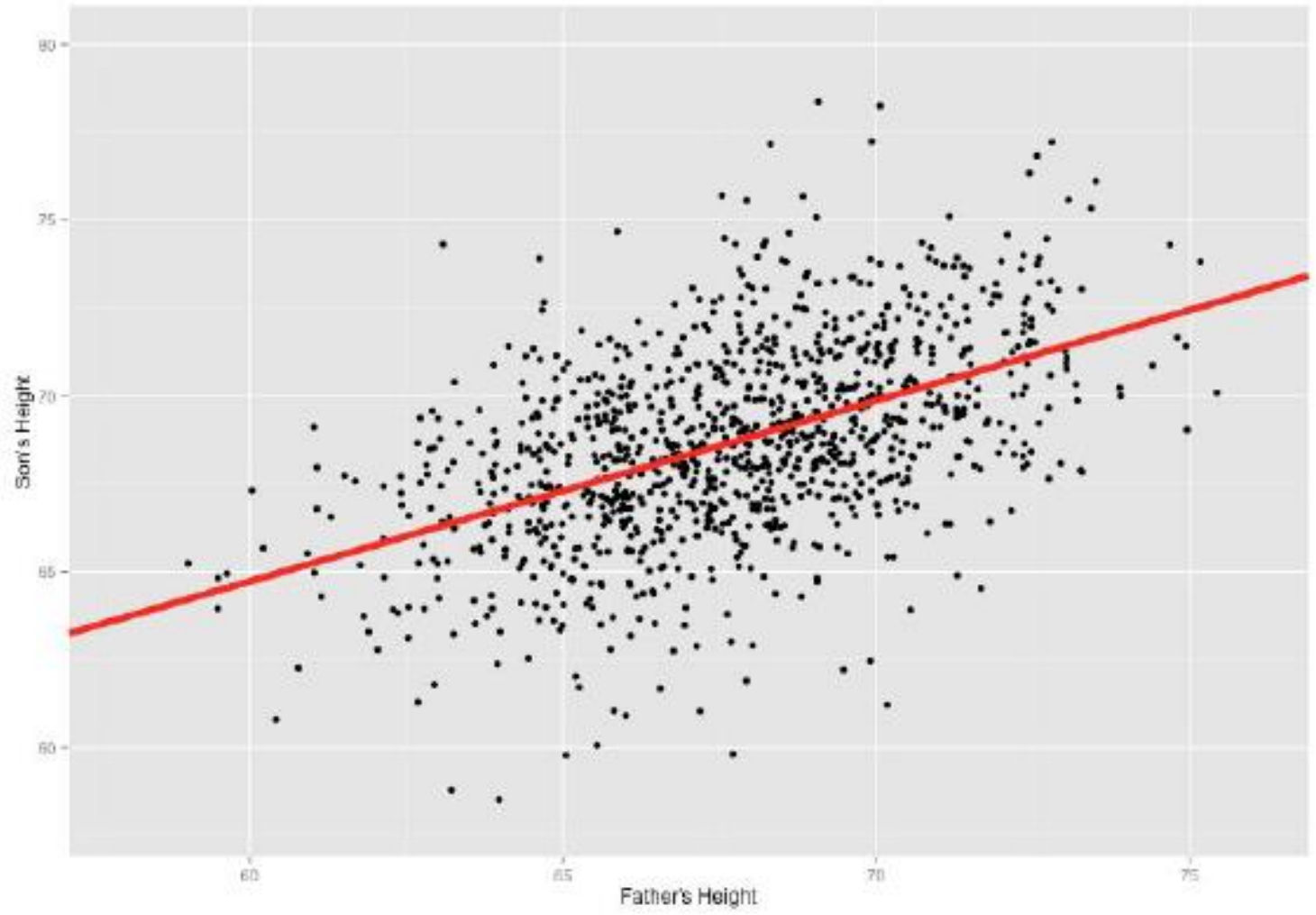
Al hacer esto, podríamos tomar múltiples hombres y las alturas de su hijos y hacer cosas como decirle a un hombre lo alto que esperamos que sea su hijo ... ¡incluso antes de que tenga un hijo!



Ejemplo

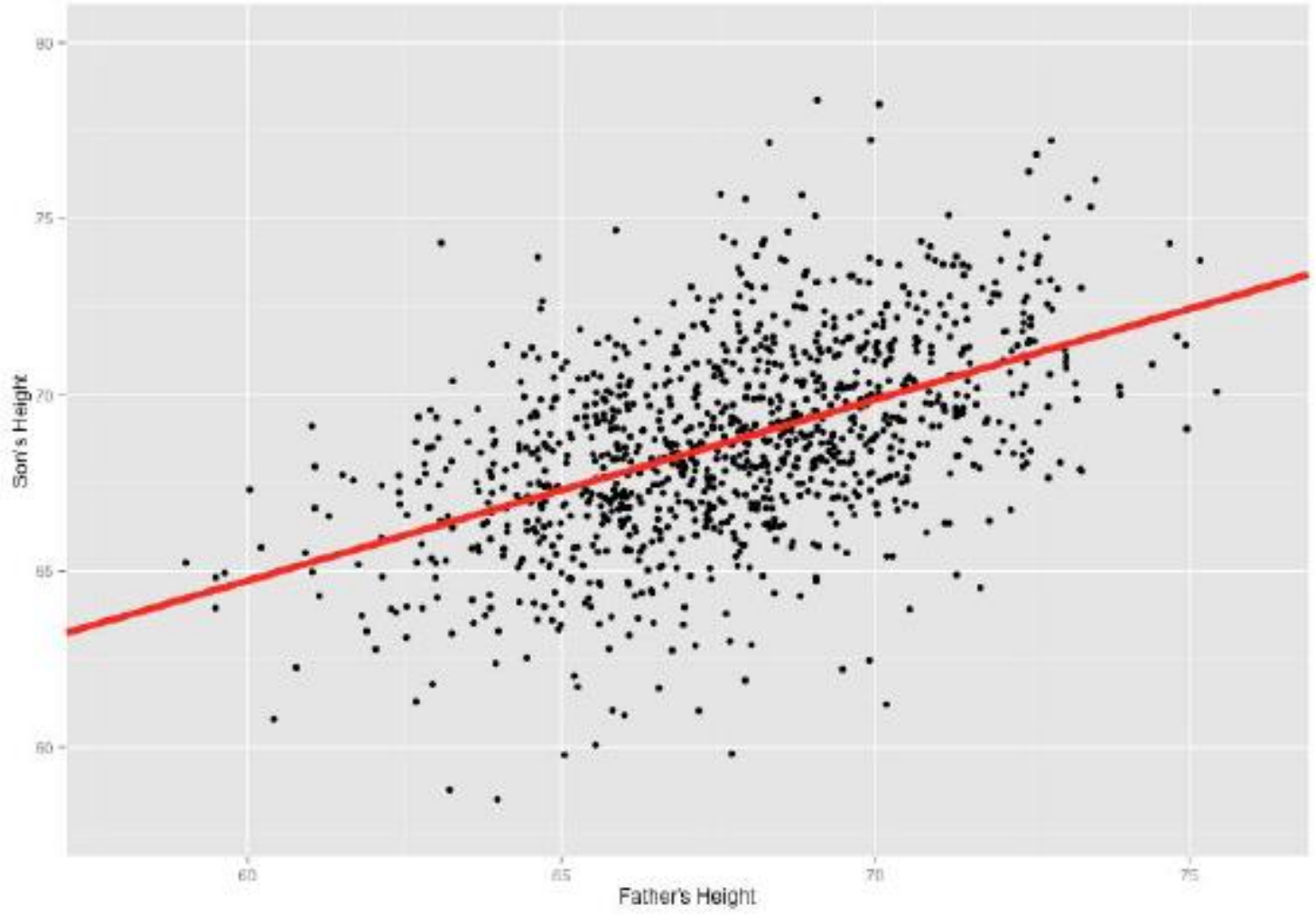
Nuestro objetivo con la regresión lineal es minimizar la distancia vertical entre todos los puntos de datos y nuestra línea.

Entonces, al determinar la mejor línea, intentamos minimizar la distancia entre todos los puntos y su distancia a nuestra línea.



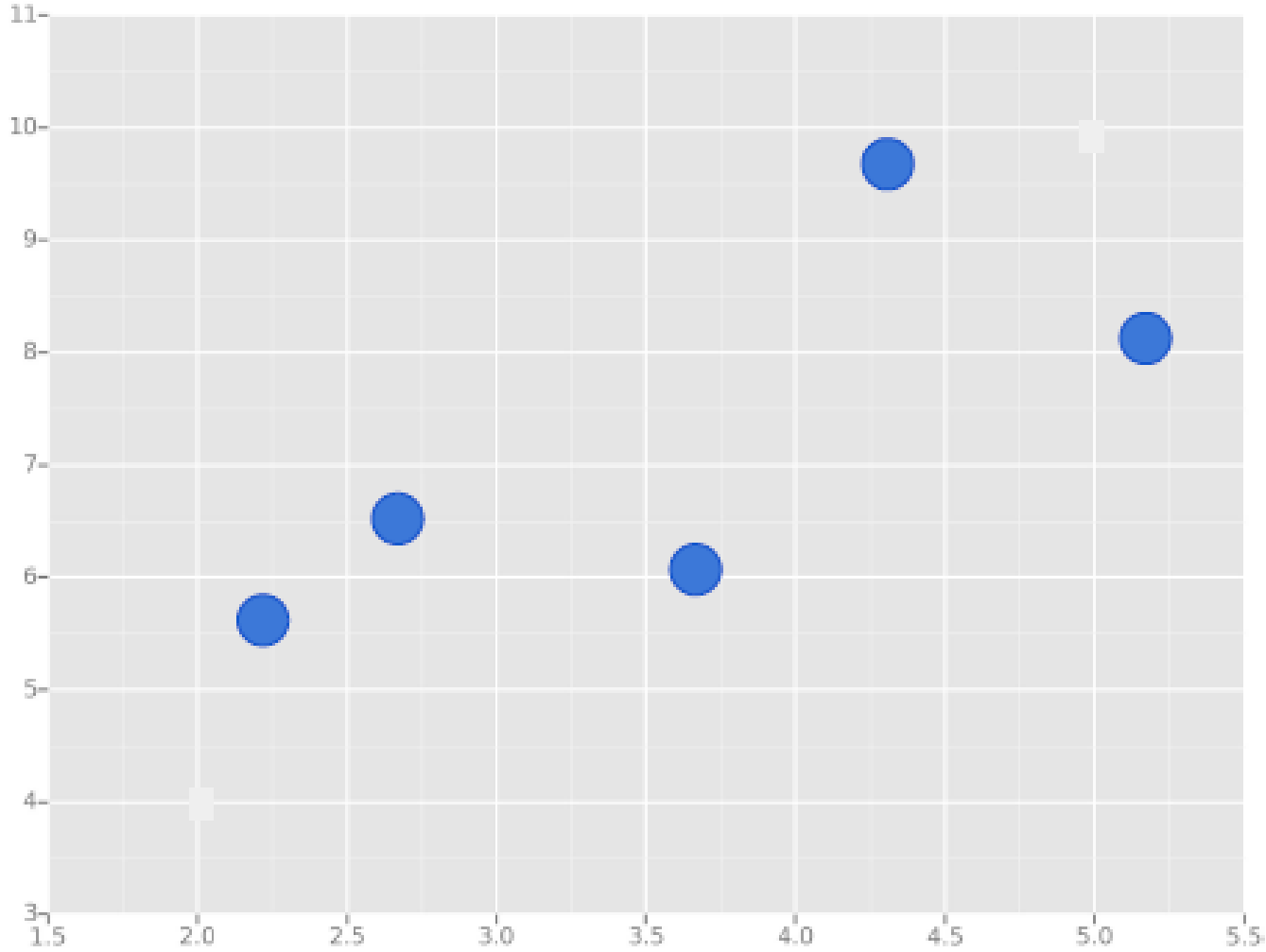
Ejemplo

Hay muchas formas diferentes de minimizar esto (suma de errores al cuadrados, suma de errores absolutos, etc.), pero todos estos métodos tienen el objetivo general de minimizar esta distancia.



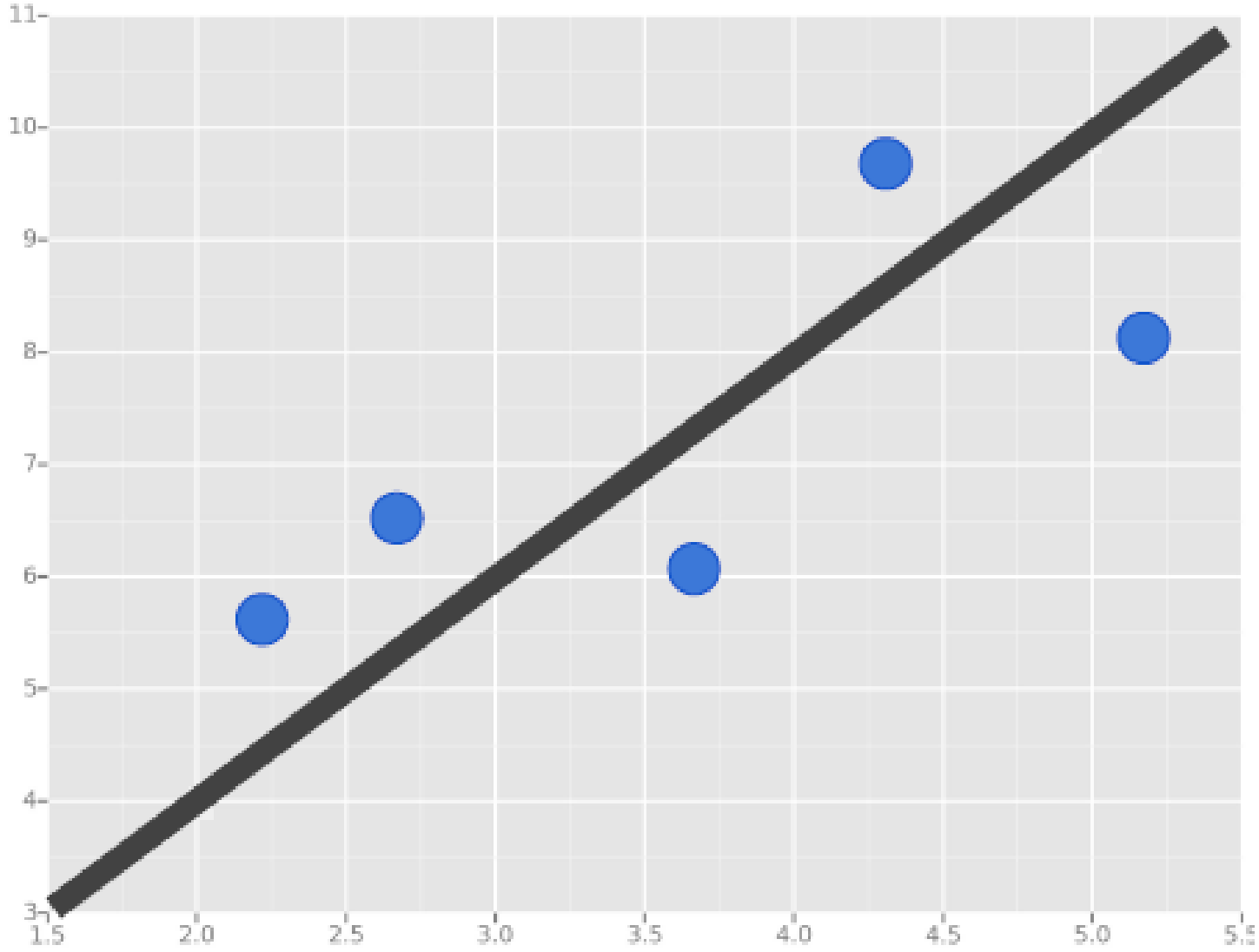
Ejemplo

Por ejemplo, uno de los métodos más populares es el método de mínimos cuadrados. Aquí tenemos puntos de datos azules a lo largo de un eje x e y.



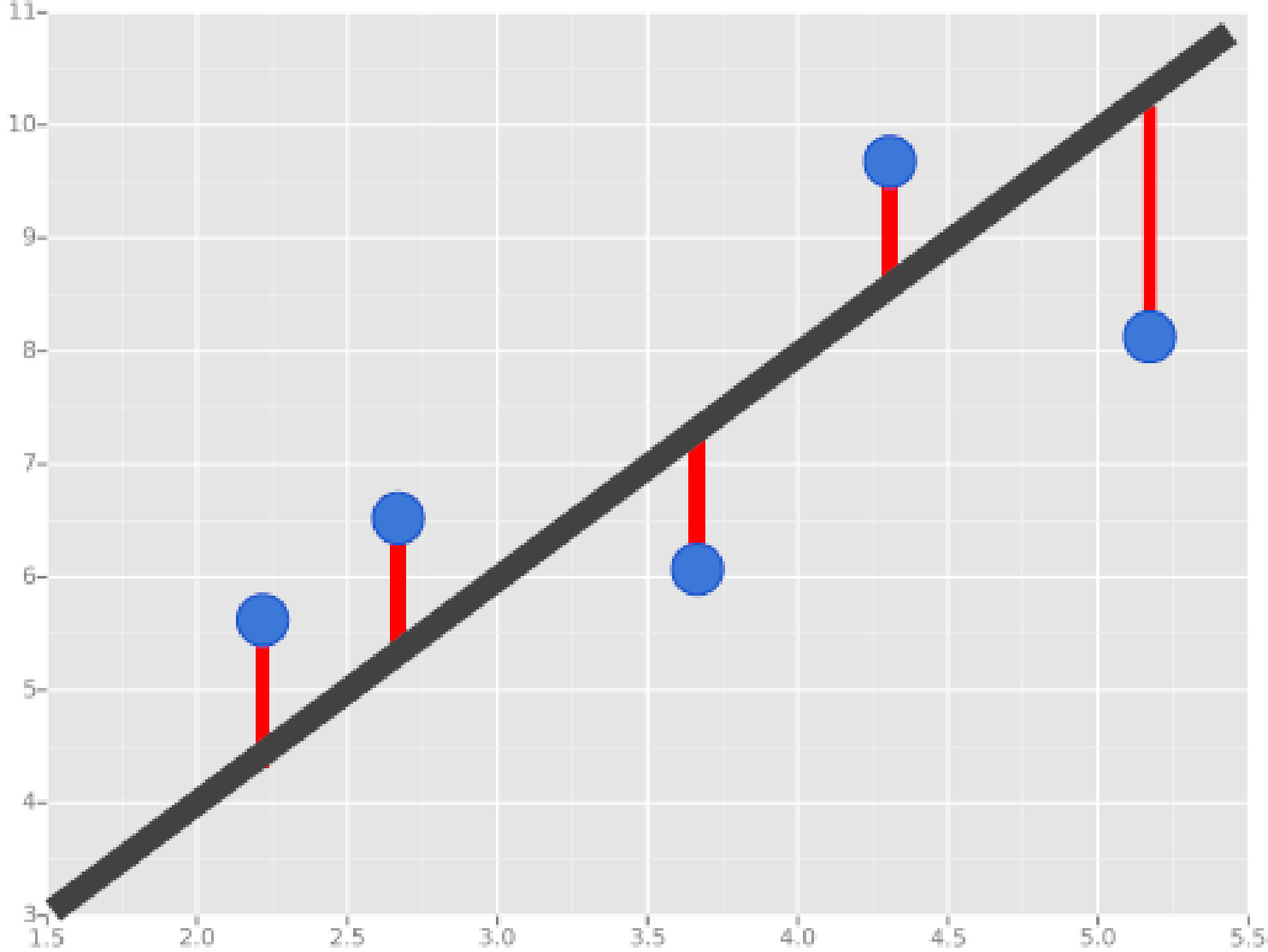
Ejemplo

Ahora queremos ajustar una línea de regresión lineal. La pregunta es, ¿cómo decidimos qué línea es la más adecuada?



Ejemplo

Utilizaremos el método de mínimos cuadrados, que se ajusta minimizando la suma de cuadrados de los residuos. Los residuos para una observación son la diferencia entre la observación (el valor y) y la línea ajustada.



Ejemplo con Python

- Ahora usaremos SciKit-Learn y Python para crear un modelo de regresión lineal. Luego resolverá un ejercicio propuesto y revisaremos las soluciones.