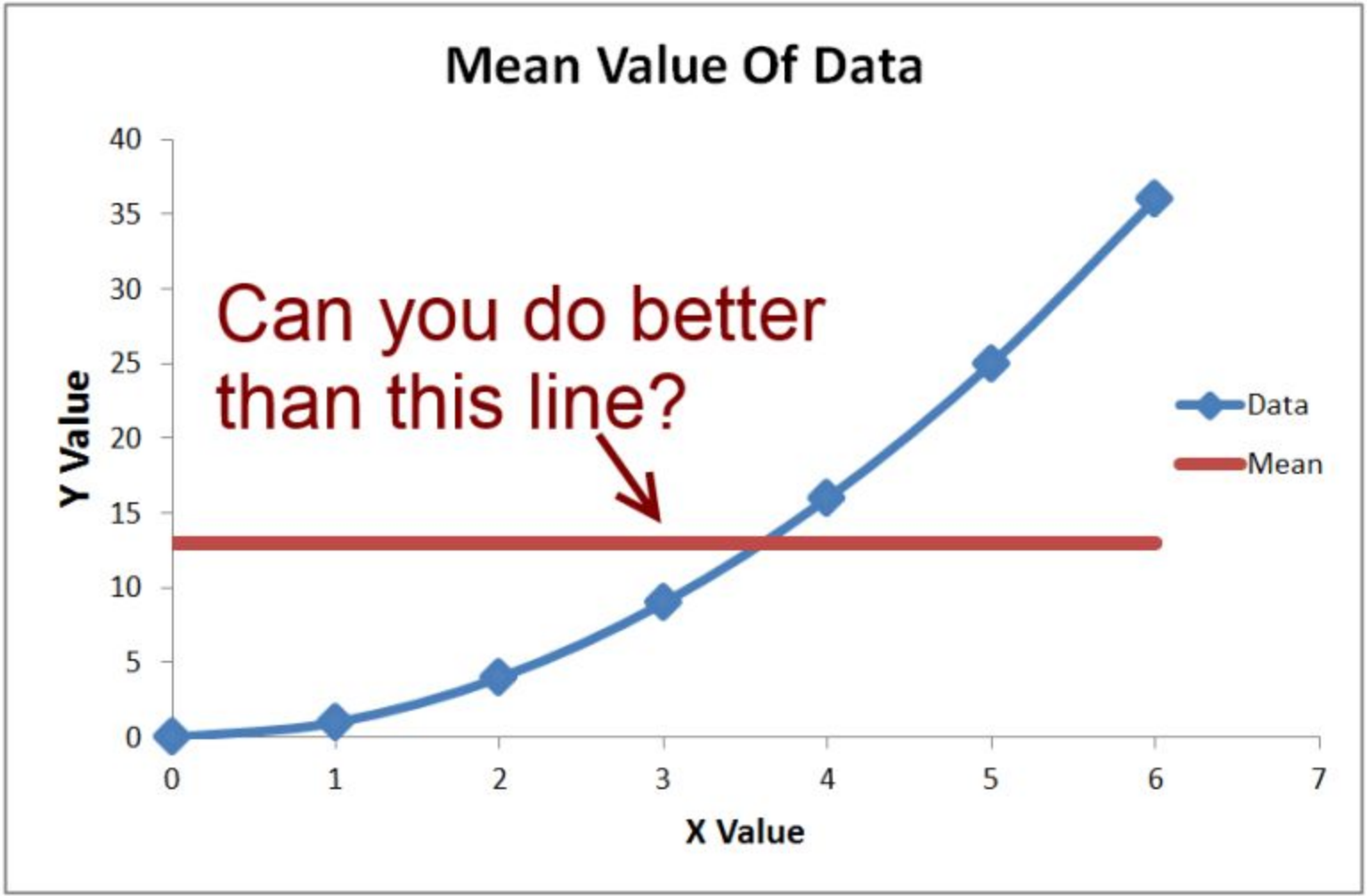
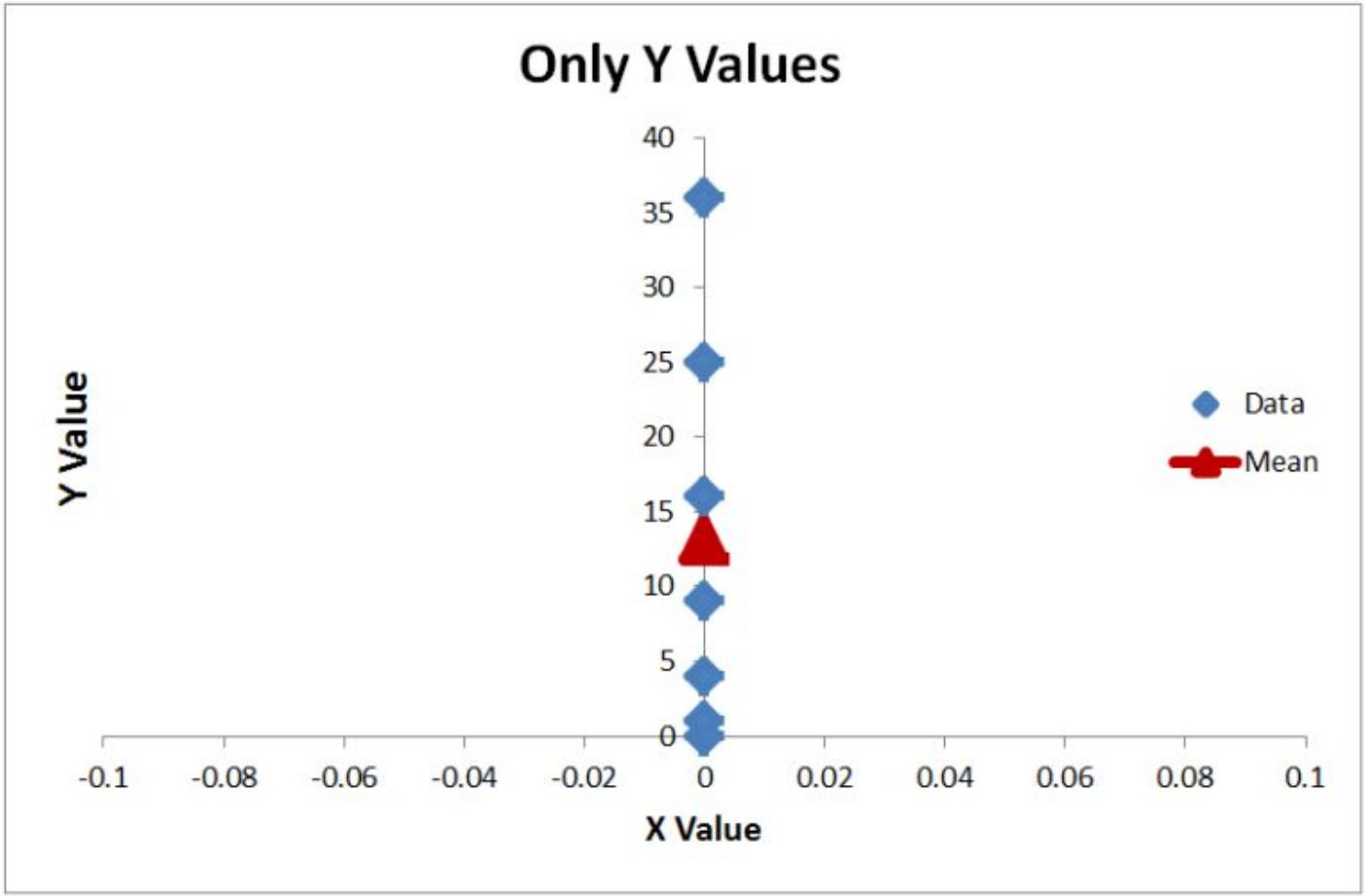
R^2 y R^2 negativo


Mean Value Of Data
Can you do better than this line?
Y Value
X Value
0
5
10
15
20
25
30
35
40
0
1
2
3
4
5
6
7
Data
Mean

La línea roja es el valor que minimiza el error cuadrático de los puntos azules, asumiendo que sólo se conocen los valores de y para esos puntos.


Only Y Values
Y Value
40
35
30
25
20
15
10
5
0
-0.1
-0.08
-0.06
-0.04
-0.02
0
0.02
0.04
0.06
0.08
0.1
X Value
Data
Mean

Si quieren elegir un valor que minimice el error cuadrático, el valor elegido será la media (el triángulo rojo).

Otra forma de pensarlo: Si tomo todos los valores de y en un orden aleatorio, ustedes adivinan los valores de los siete puntos, sin darles ningún otro dato, ¿con qué estrategia tendrían el mínimo error cuadrático? La estrategia es adivinar la media de los puntos.

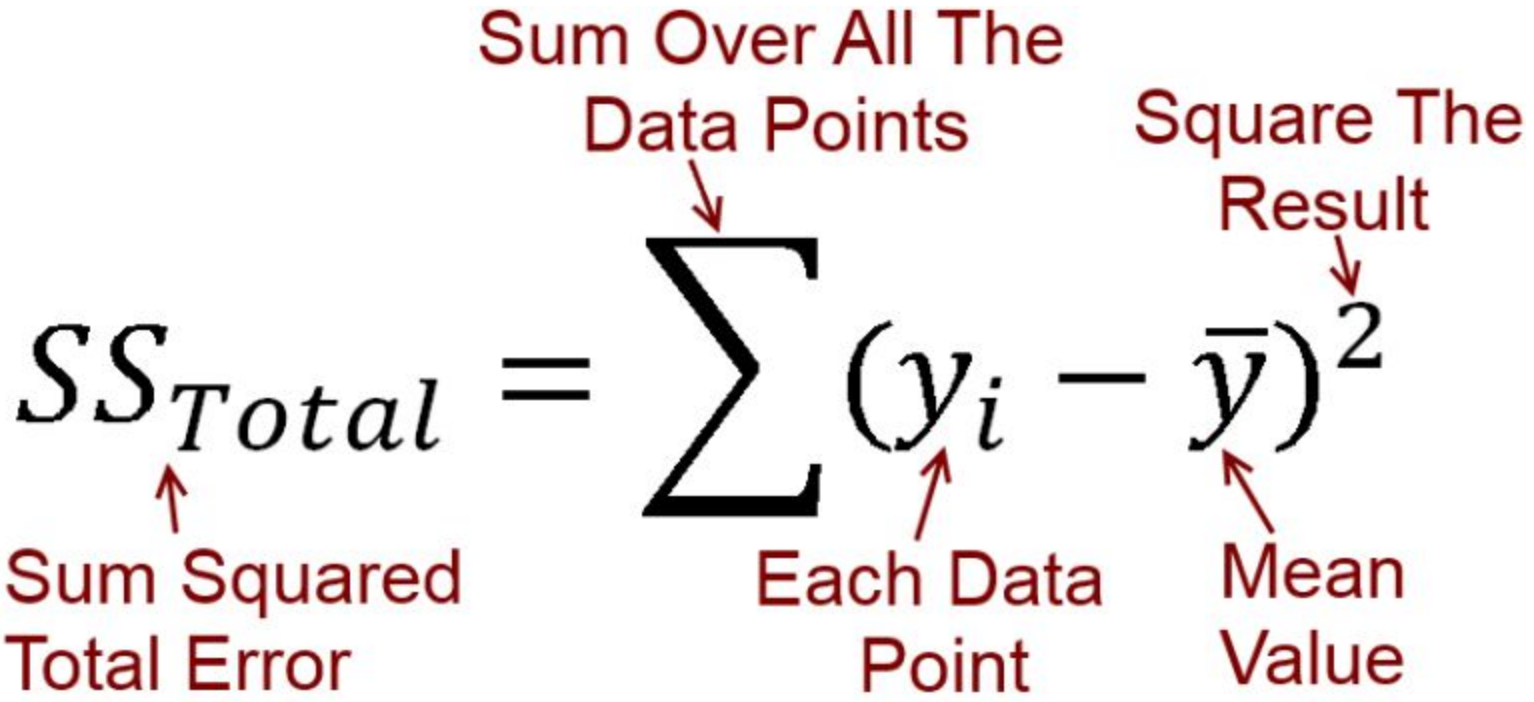
Teniendo la posibilidad de usar los valores de x de los puntos, podrían hacer una mejor aproximación que la media de los puntos? R^2 responde esta pregunta.

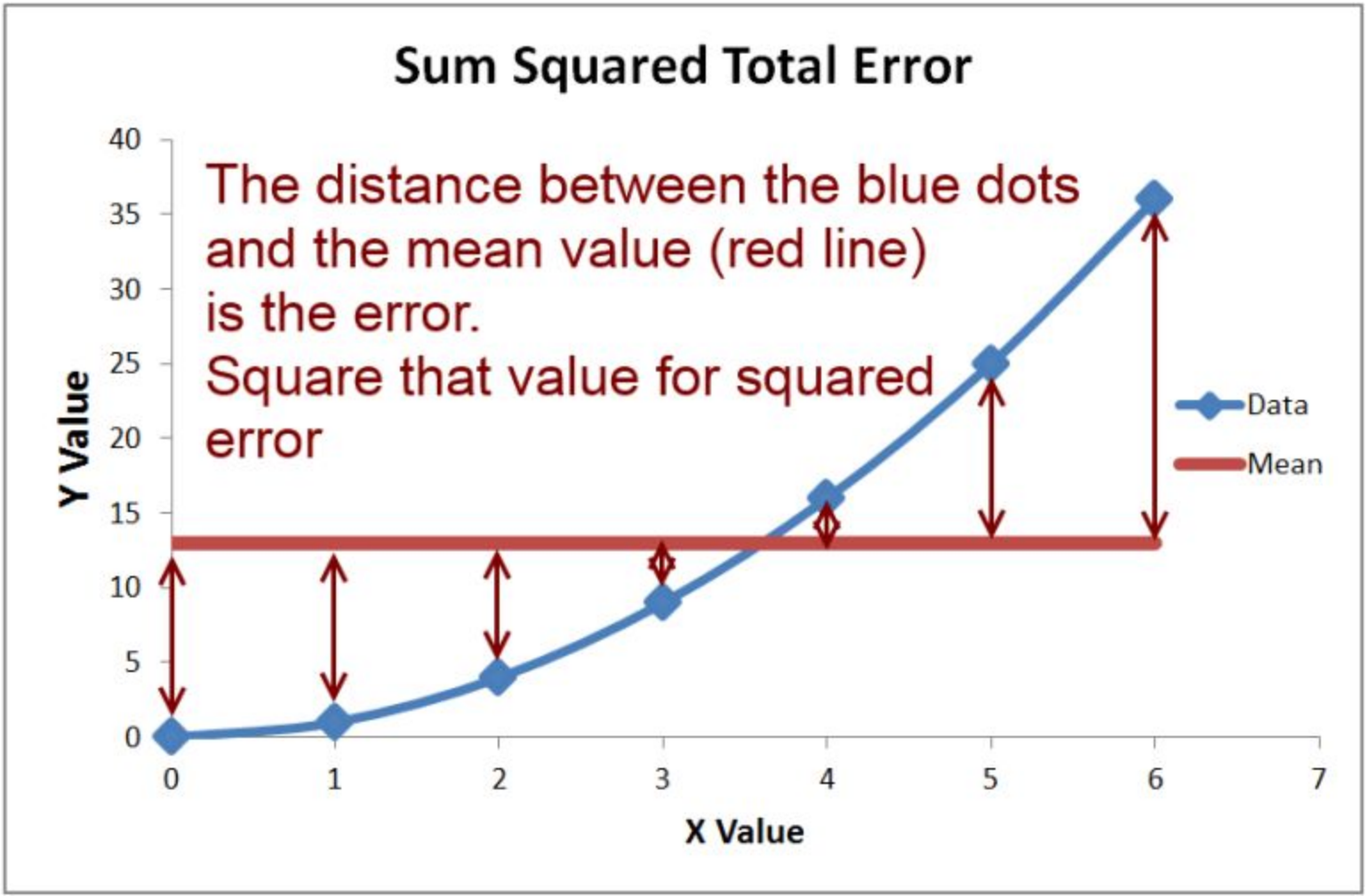
SS = summed squared error

Para calcular SS:

- Calcular el valor de la media
- Para cada punto, restar el valor de la media al valor del punto
- Calcular el cuadrado del resultado del paso anterior

Sumar los resultados del tercer paso, para todos los puntos. Ese valor es SS.

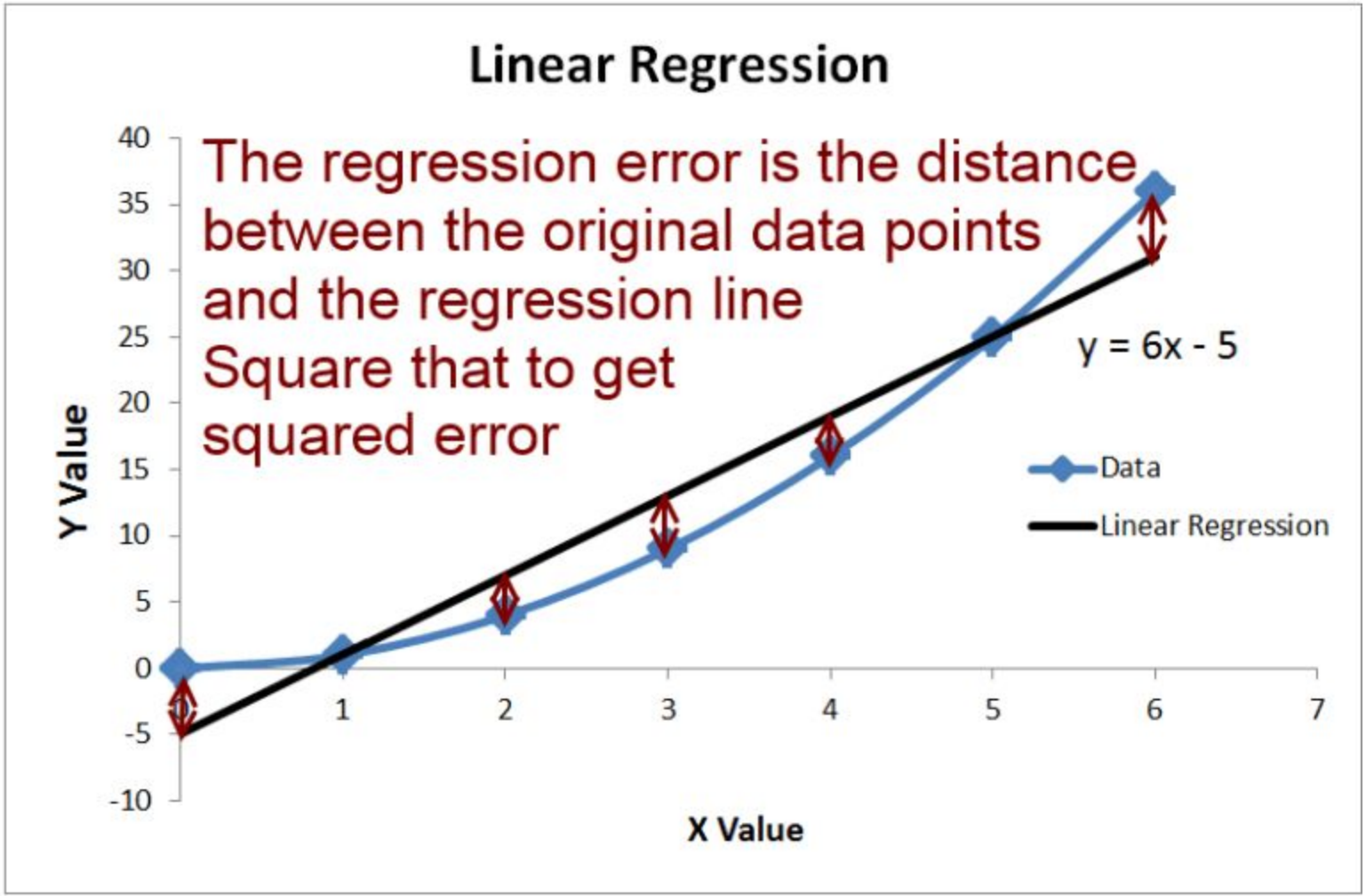

Sum Over All The Data Points
Square The Result
$SS_{Total} = \sum (y_i - \bar{y})^2$
Sum Squared Total Error
Each Data Point
Mean Value


Sum Squared Total Error
The distance between the blue dots and the mean value (red line) is the error.
Square that value for squared error
Y Value
X Value
Data
Mean
0
5
10
15
20
25
30
35
40
0
1
2
3
4
5
6
7

Ahora queremos calcular el error en los valores de la regresión respecto de los verdaderos valores.

Este es el error de la regresión.

Idealmente el error de la regresión es muy bajo, cercano a cero.

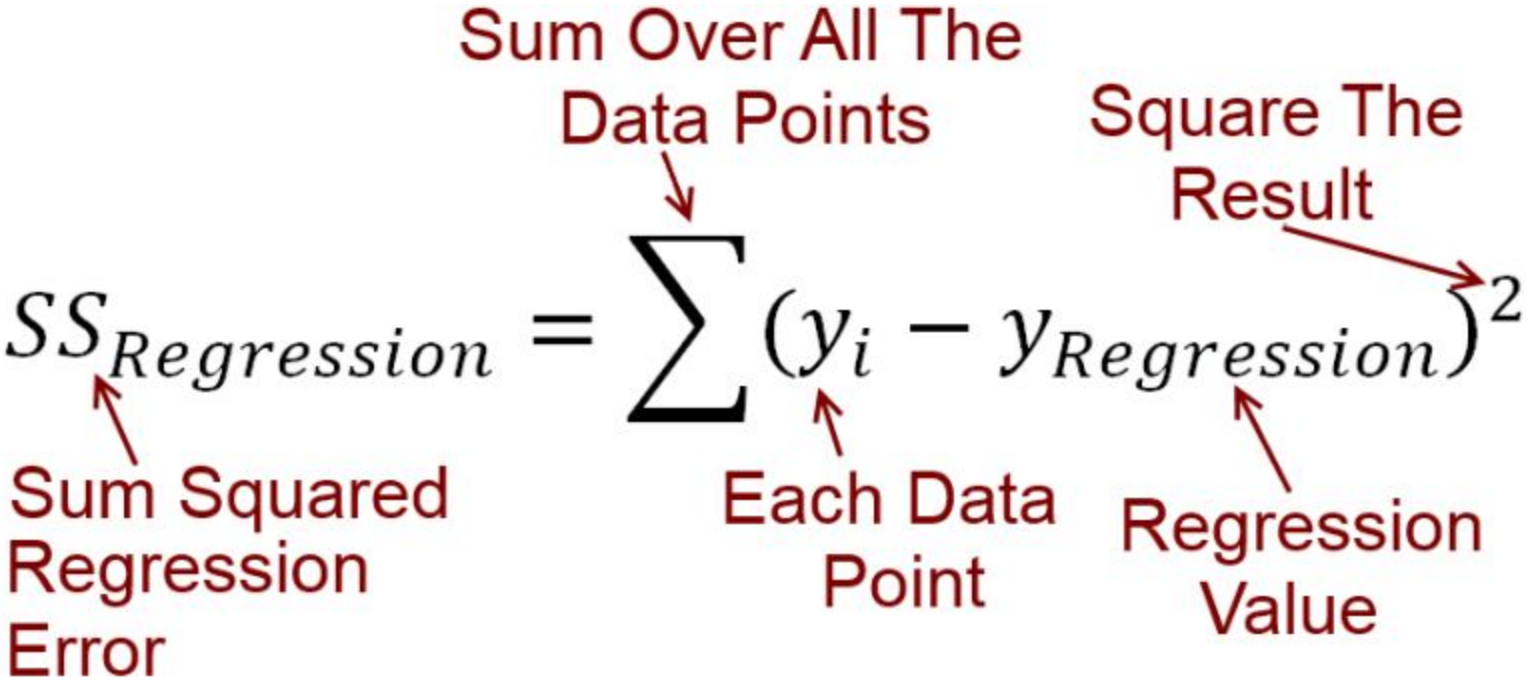

Linear Regression
The regression error is the distance between the original data points and the regression line
Square that to get squared error
y = 6x - 5
Data
Linear Regression
Y Value
X Value
40
35
30
25
20
15
10
5
0
-5
-10
1
2
3
4
5
6
7

El cociente entre el error de la regresión y el error total (respecto de la media) nos dice qué proporción del error total se mantiene en el modelo de regresión.

Restando este cociente de 1, se obtiene la proporción de error que es eliminada usando una regresión.

Esto es R^2:

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}}$$


Sum Over All The Data Points
Square The Result
$SS_{Regression} = \sum (y_i - y_{Regression})^2$
Sum Squared Regression Error
Each Data Point
Regression Value

Cuanto menor es el valor del error en la regresión respecto del error total, mayor es el valor de R^2

El máximo (y mejor) valor de R^2 es 1. Para obtener ese valor, el error de la regresión debe ser cero.

$$R^2 = 1 - \frac{0}{SS_{Total}} \rightarrow 1.0$$

R^2 puede ser menor que cero.

Para fines prácticos, el mínimo R^2 que vamos a obtener es cero, pero sólo porque asumimos que si la regresión no mejora la estimación de la media, vamos a estimar con la media. Sin embargo, si la regresión es peor que usar la media, el valor de R^2 calculado es negativo.

Ejemplos:

Queremos predecir la cantidad de población de uno de los estados de USA.

La única información que tenemos es la población en cada uno de los 50 estados, pero no sabemos a qué estado corresponde cada unos de esos valores.

Debemos predecir la población (en millones) de cada uno de los estados en orden aleatorio.

Lo mejor que podemos hacer en este caso es tomar el valor de la media.

El error total (SS) es 2298.2

Si en lugar de usar la media usamos la mediana, el error total (SS) es 2247.2

Calculando R^2 obtenemos:

$$R^2 = 1 - \frac{2447.2}{2298.2} = -.0649$$

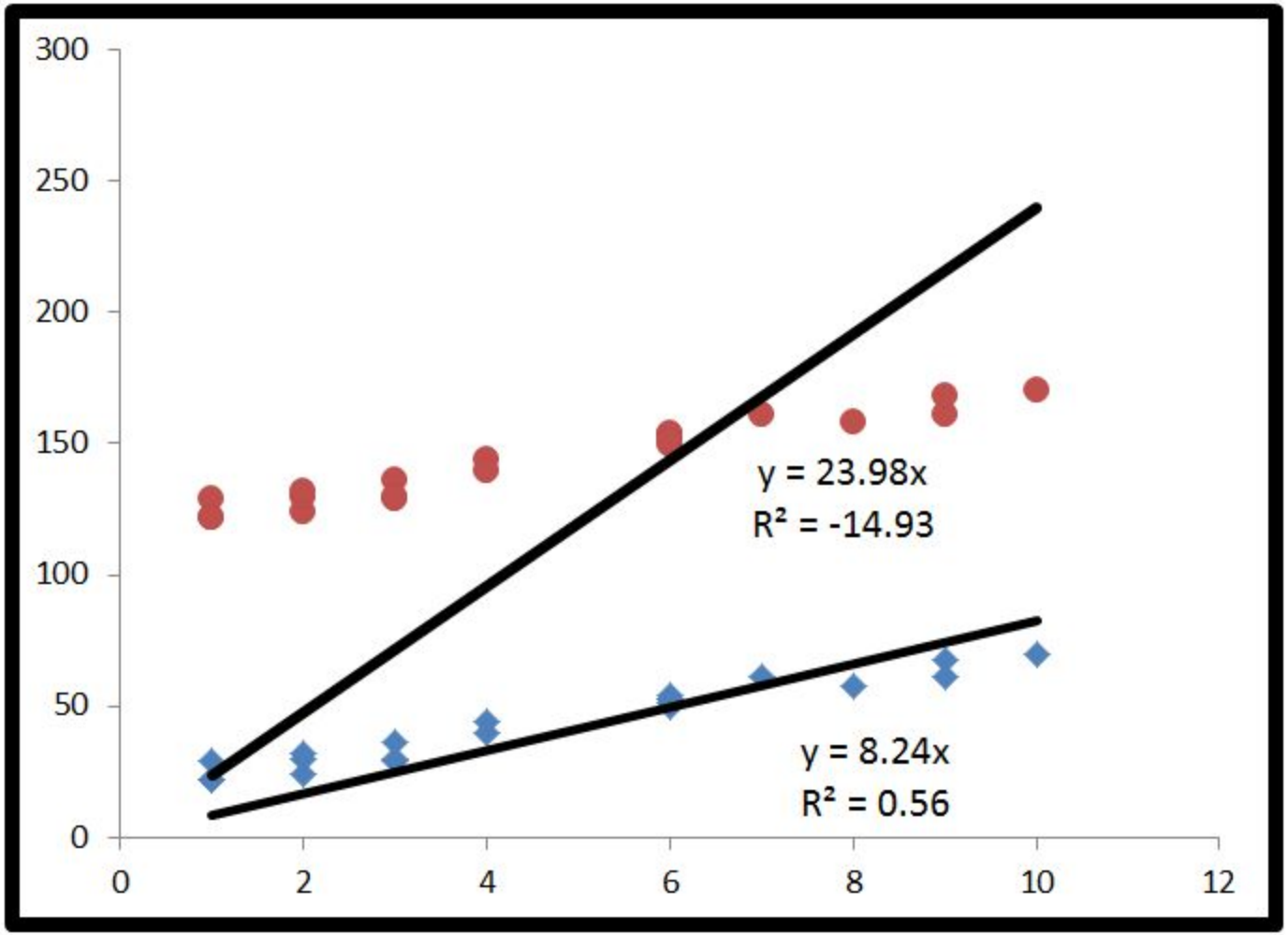
Ver que el valor que usa no coincide con el que pone en el enunciado.
Según el texto del enunciado, la estimación con la mediana es mejor que con la media, y eso daría un R^2 mayor que cero.

Otra forma de obtener un valor de R^2 negativo

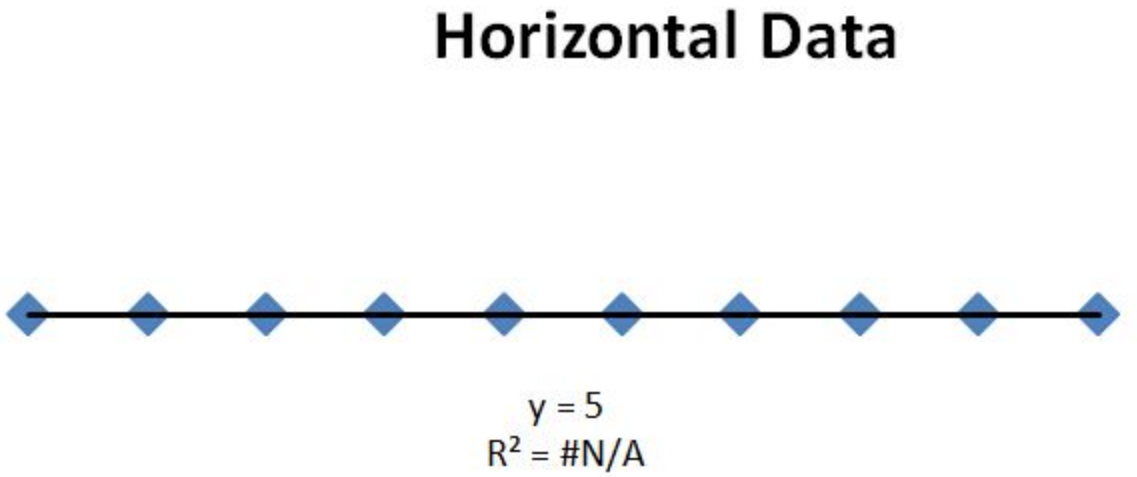
La forma más común de obtener un valor de R^2 negativo es forzar a que la línea de la regresión pase por un punto determinado, por ejemplo definiendo el intercepto.

Si el punto elegido es la media de x e y, la recta tendrá el mínimo SS y el máximo valor de R^2. Con esta elección es imposible obtener un valor de R^2 negativo.

En este ejemplo, el intercepto para ambas regresiones está seteado en cero. Para la regresión en los puntos rojos, el verdadero valor del intercepto es cercano a 120, setearlo en 0 aleja mucho la recta de los puntos rojos, y en este caso el SS aumenta y es mayor que el obtenido al usar la media, obteniendo así un valor de R^2 negativo.


300
250
200
150
100
50
0
0
2
4
6
8
10
12
y = 23.98x
R² = -14.93
y = 8.24x
R² = 0.56

Caso curioso



$$R^2 = 1 - \frac{0}{0} = undefined$$

Conclusiones

- Un valor de R2 de 1.0 es lo mejor que podemos conseguir. Indica que no hay ningún error en la regresión.
- Un valor de R2 de 0 significa que la regresión no es mejor que tomar el valor medio, es decir no estamos usando ninguna información de otras variables.
- Un valor negativo de R2, significa que estamos estimando peor que usando la media.

Referencia

http://www.fairlynerdy.com/what-is-r-squared/