Regresión lineal y correlación

Objetivos

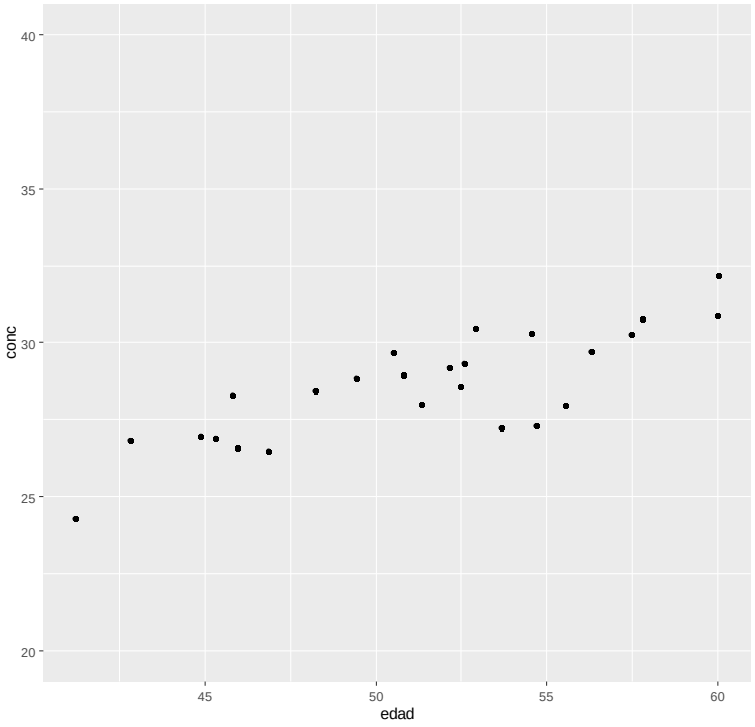
Estimar la relación (lineal) entre dos variables cuantitativas

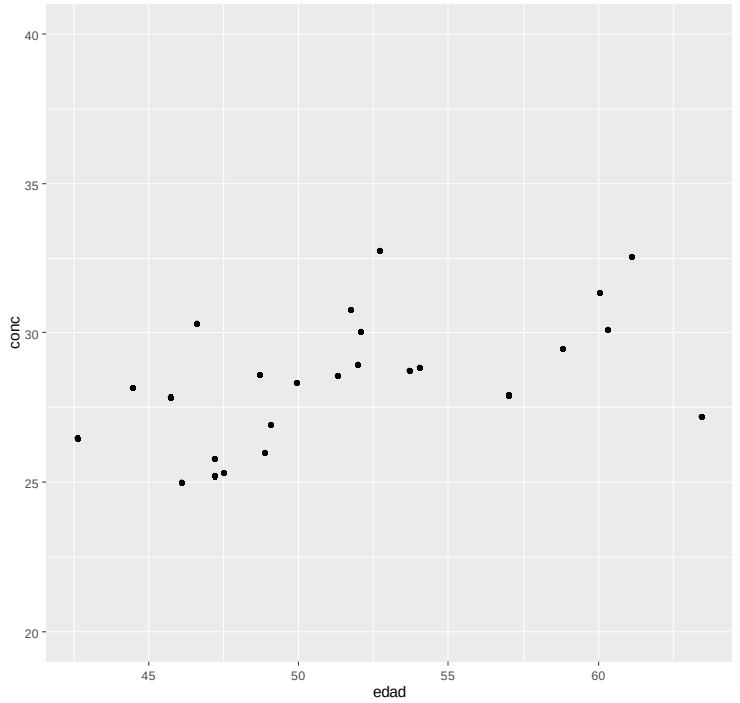
- Relación lineal
- Ajuste e interpretación de los parámetros
- Correlación lineal

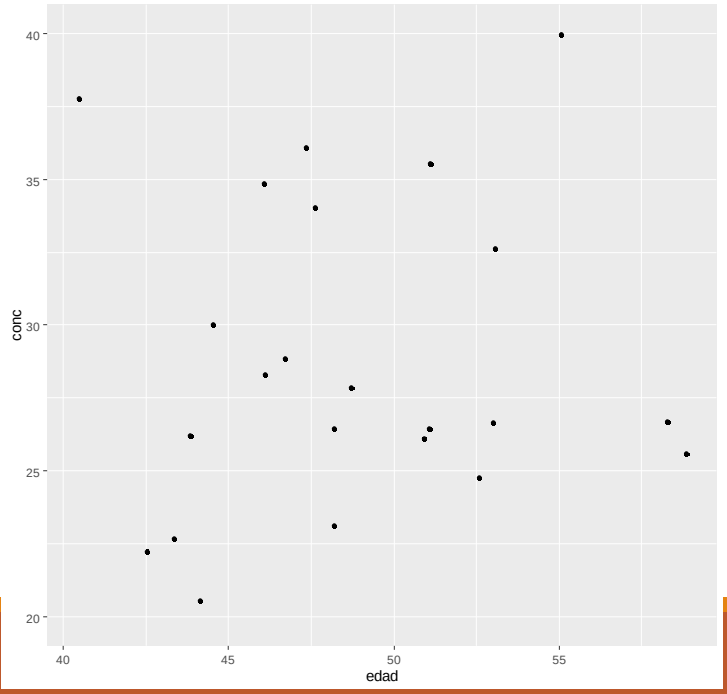
Aplicaciones

- Rectas de calibración
- Correlación entre parámetros clínicos
- Predicción de variables

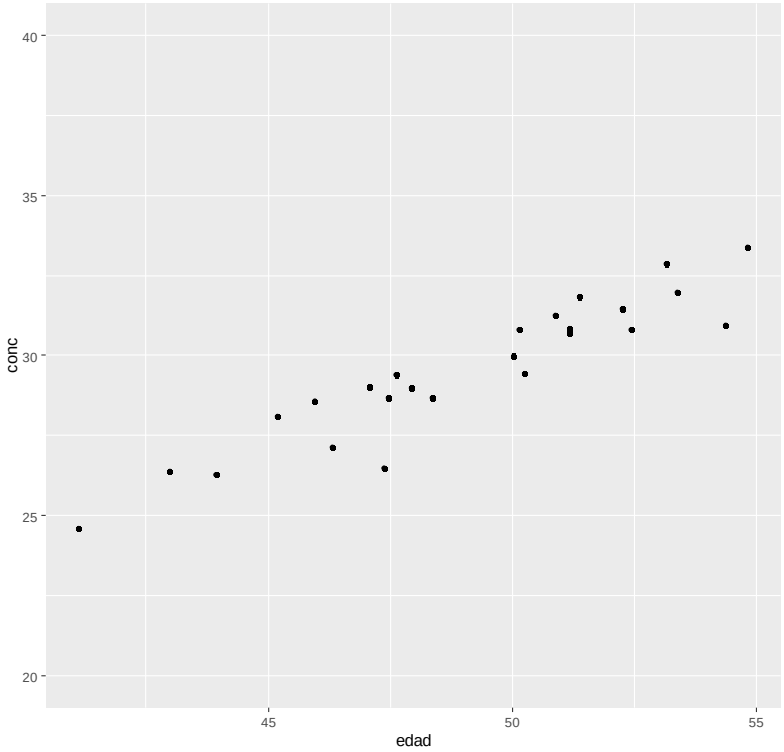
¿Existe relación entre la edad y la concentración de un metabolito?



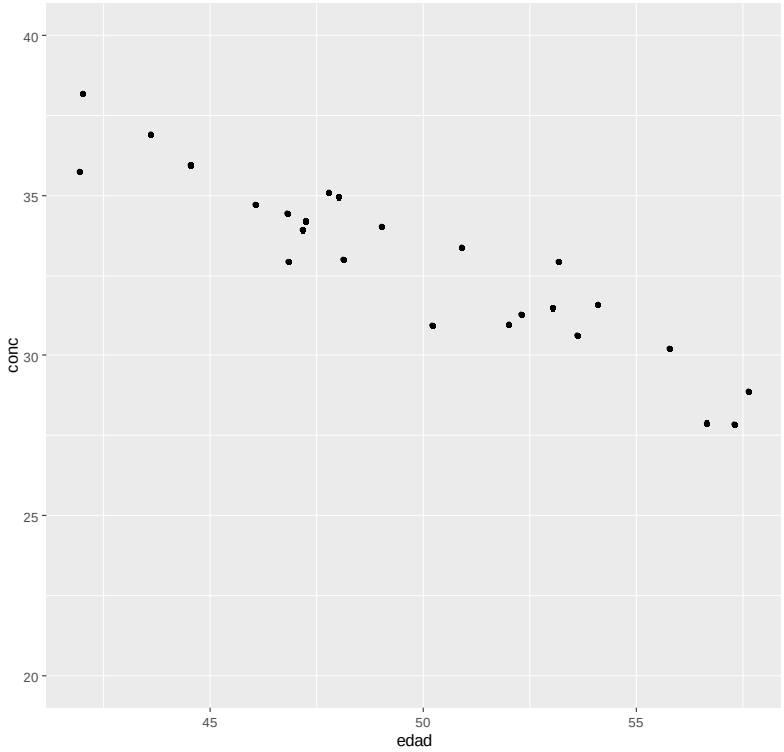




¿Existe relación entre la edad y la concentración de un metabolito?



Relación directa



Relación inversa

Relación lineal entre dos variables cuantitativas

Dado un valor de $X=x_i$, el valor esperado de Y ($E(Y|X=x_i)=\mu_i$) es $a+b\cdot x_i$

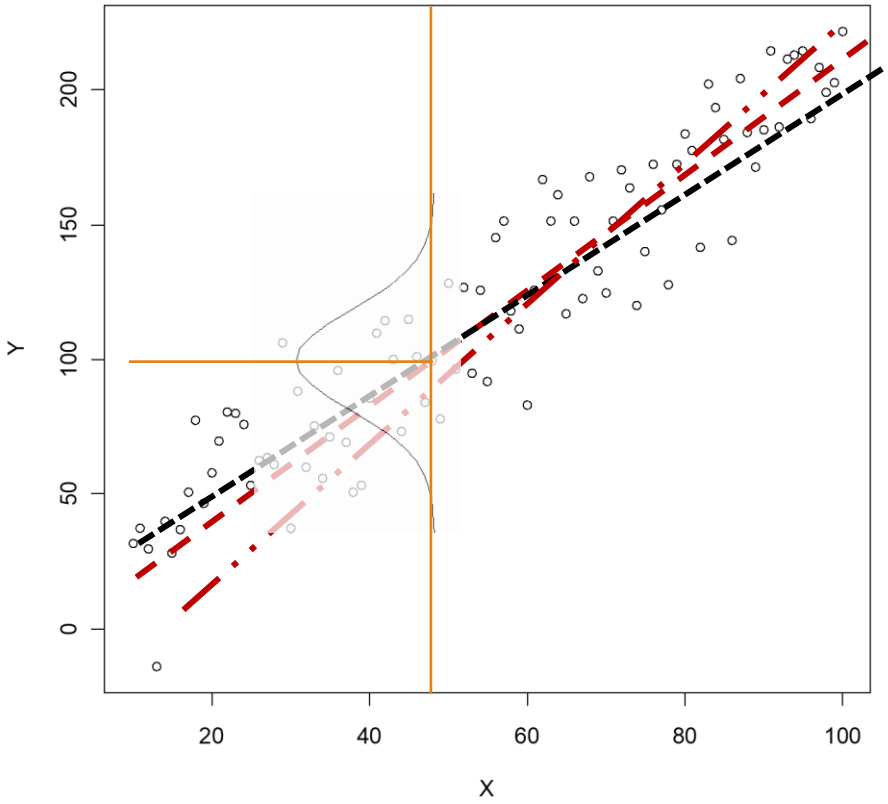
$$\mu_i = a + b \cdot x_i$$

Los valores observados se desvían del valor esperado siguiendo una $N(\mu_i, \sigma)$

$$y_i = \mu_i + \varepsilon_i$$

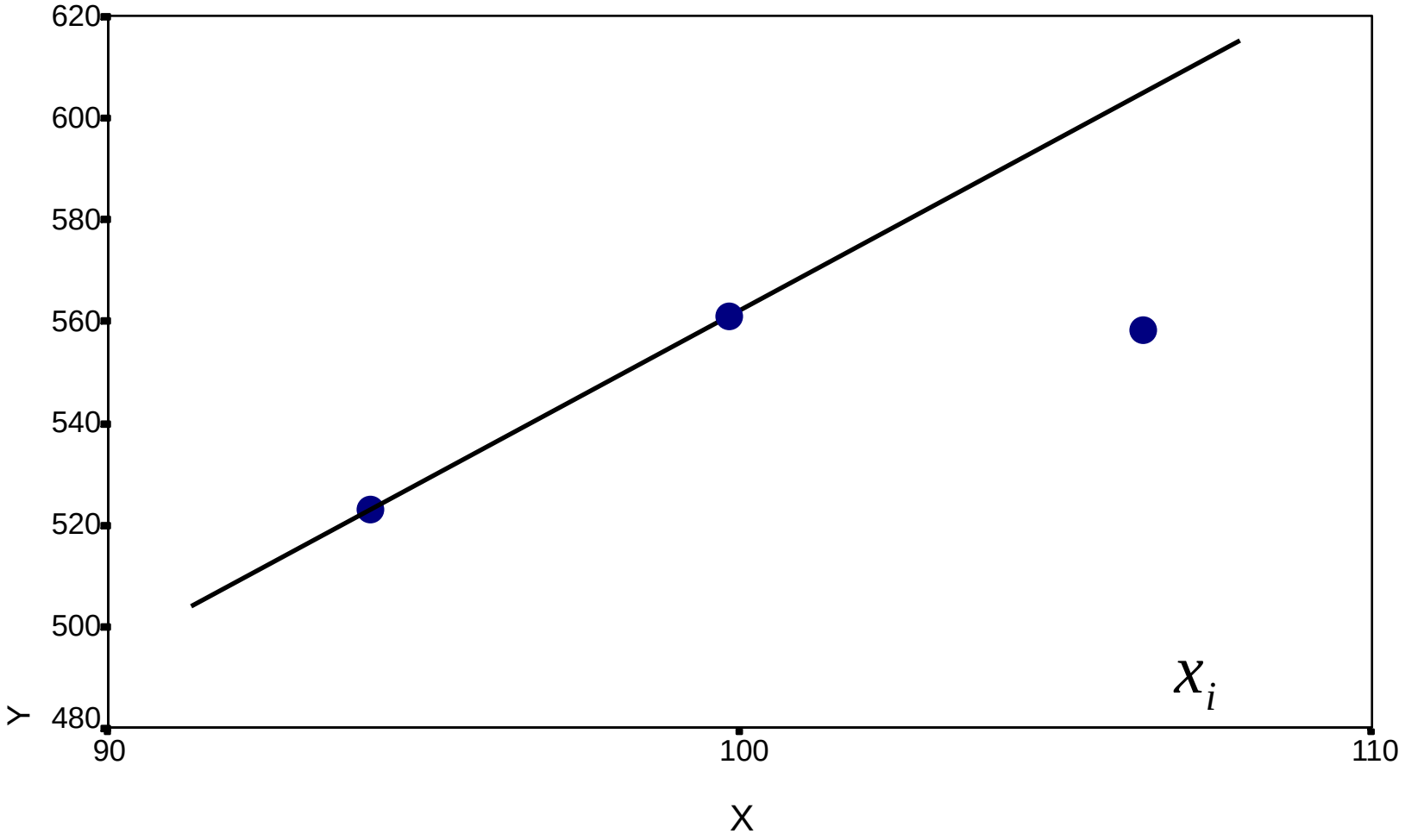
$$\varepsilon_i \rightarrow N(0, \sigma)$$

$$y_i \rightarrow N(\mu_i, \sigma)$$

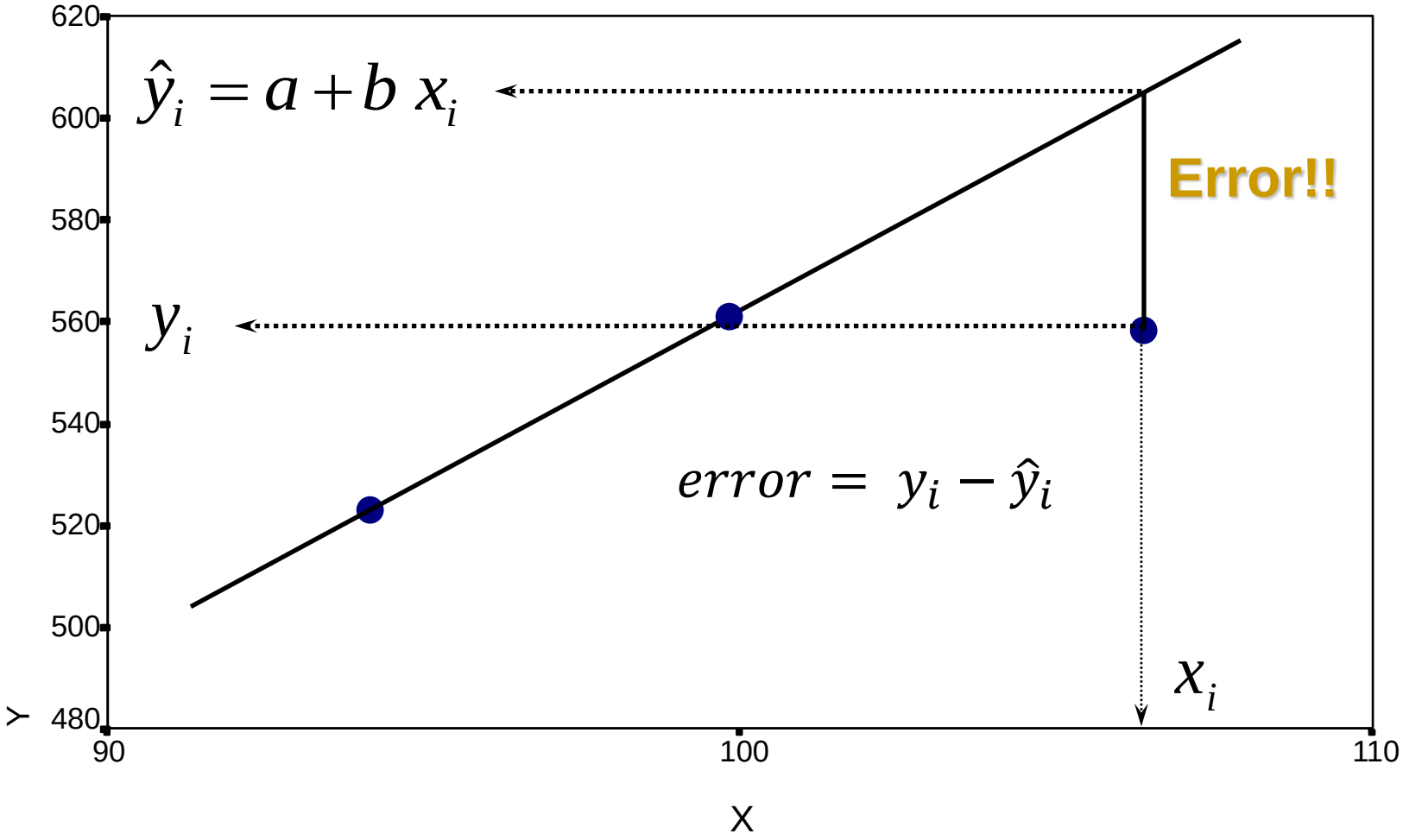


¿Cuál es la recta que mejor explica la relación entre X e Y?

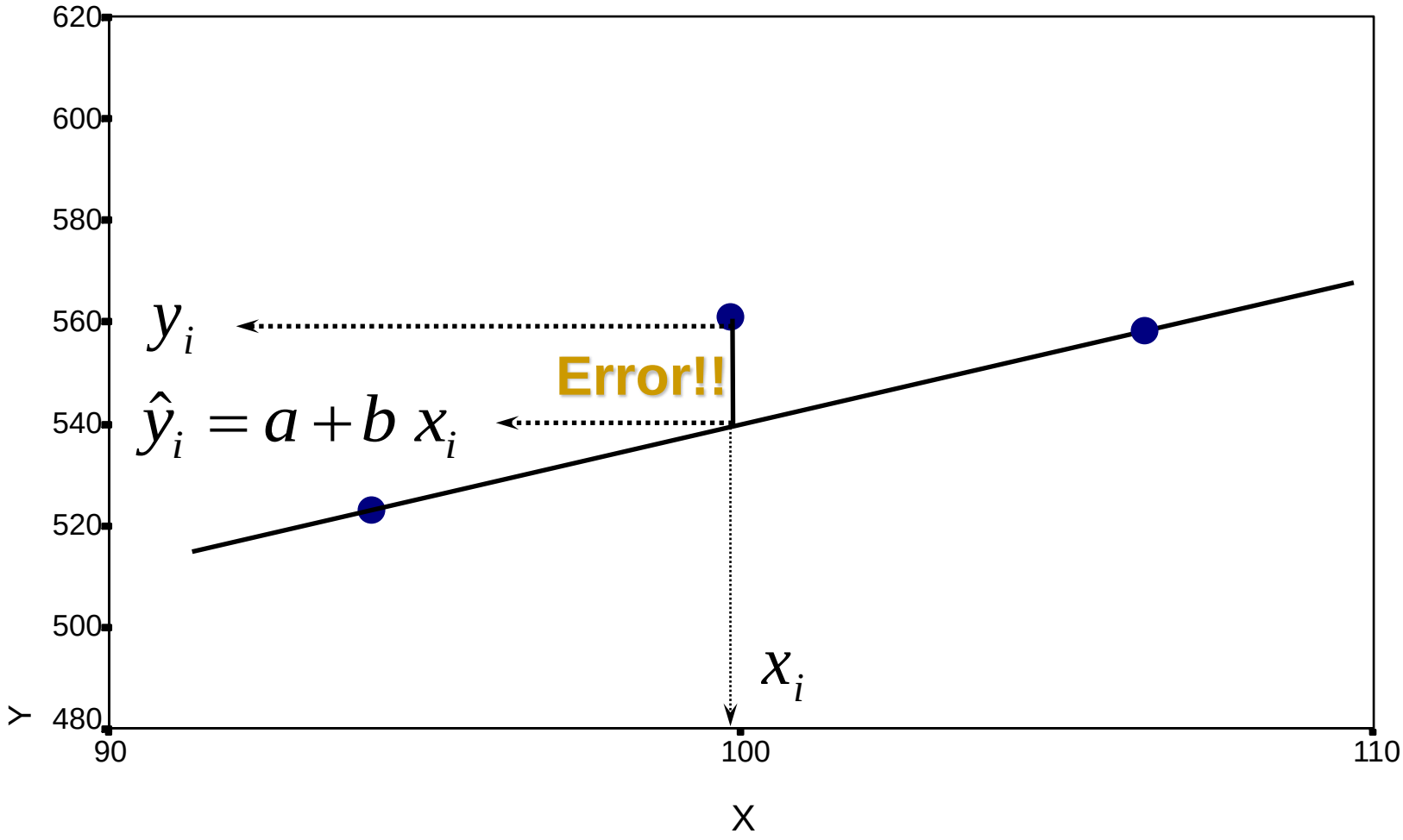
¿Cómo podemos ajustar una recta a unos puntos observados?



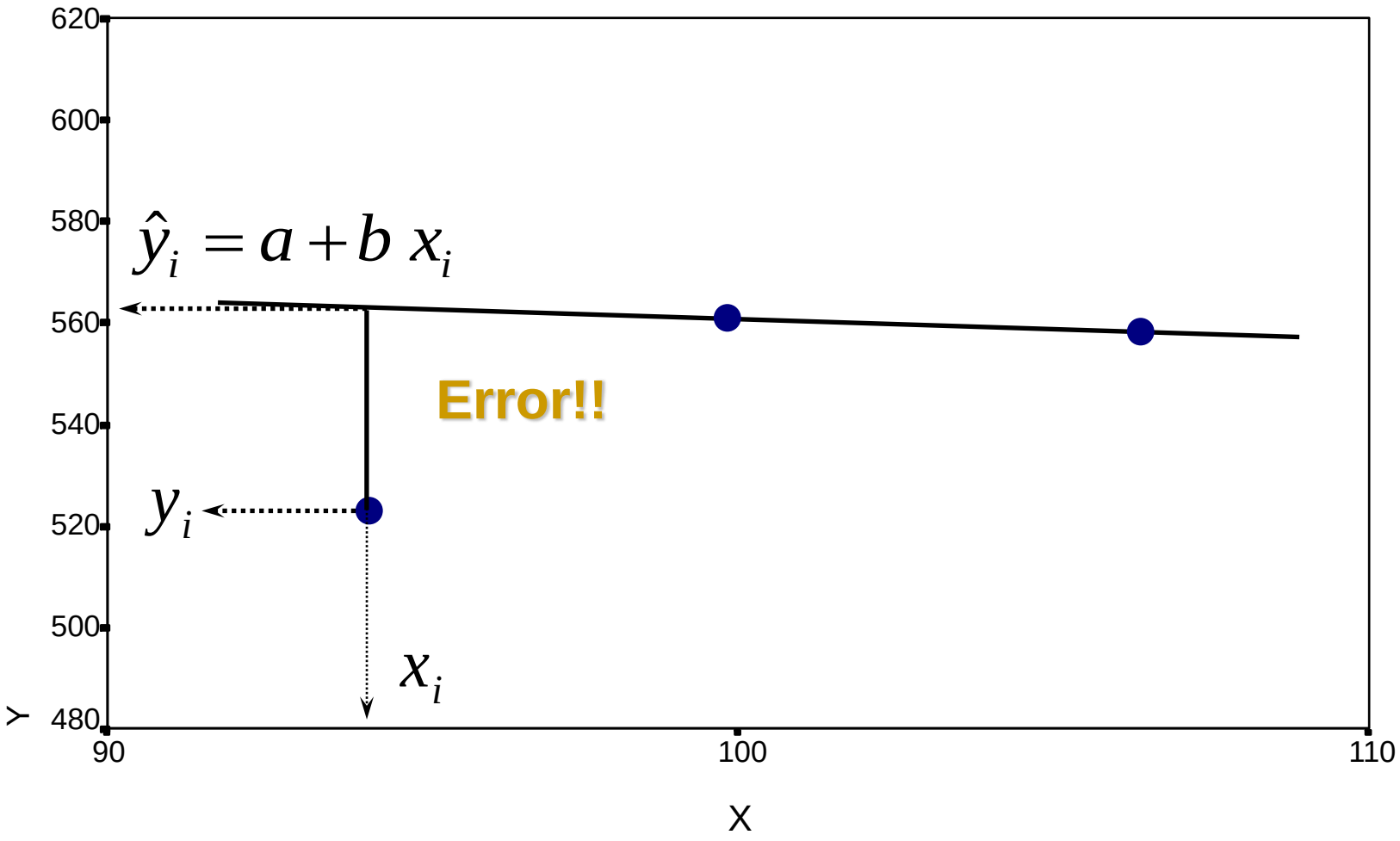
¿Cómo podemos ajustar una recta a unos puntos observados?



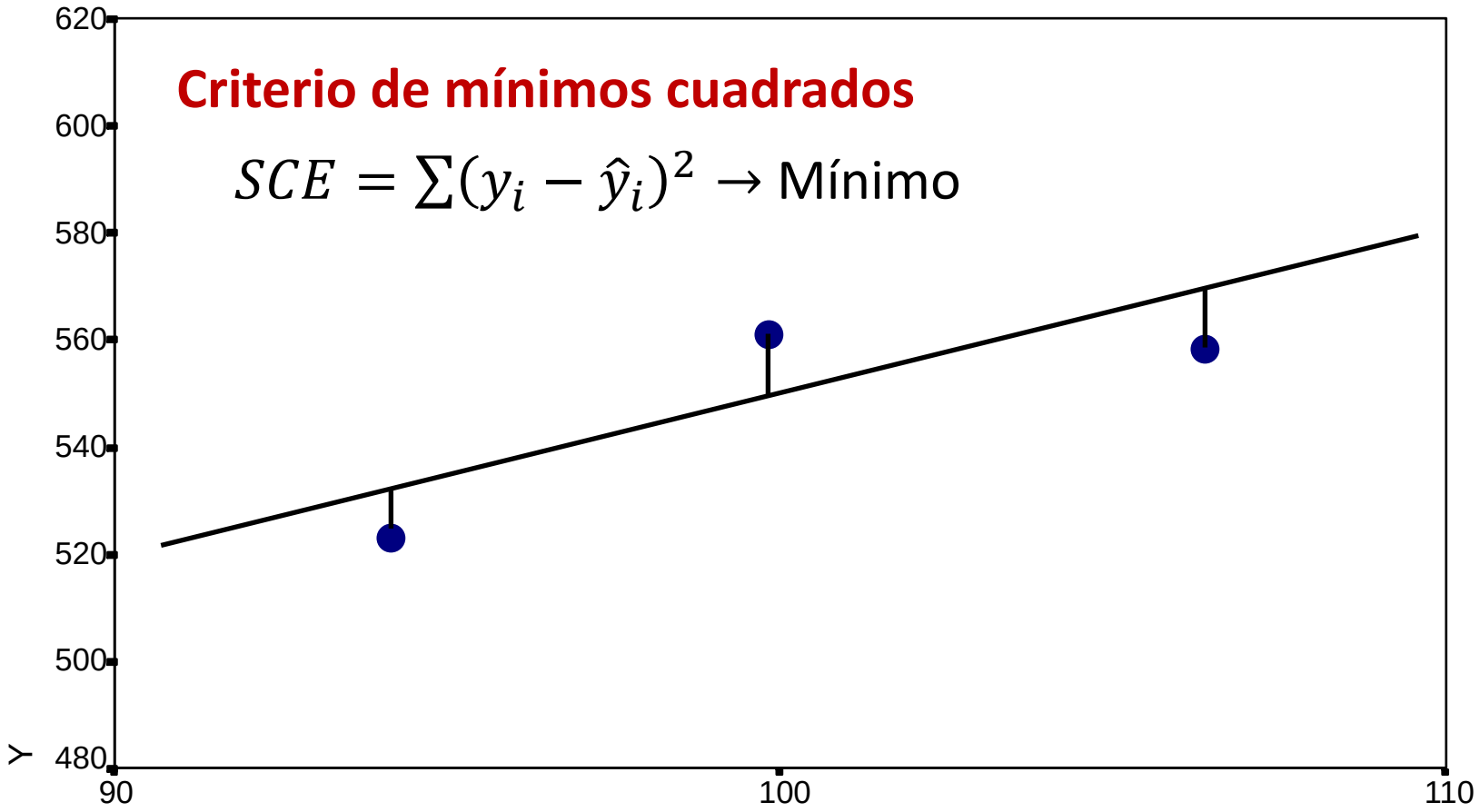
¿Cómo podemos ajustar una recta a unos puntos observados?



¿Cómo podemos ajustar una recta a unos puntos observados?



¿Cómo podemos ajustar una recta a unos puntos observados?



Criterio de mínimos cuadrados

$$SCE = \sum (y_i - \hat{y}_i)^2 = \sum \left(y_i - (a + b \cdot x_i)\right)^2 \rightarrow \text{Mínimo}$$

$$\frac{\partial SCE}{\partial a} = 0$$

$$\frac{\partial SCE}{\partial b} = 0$$

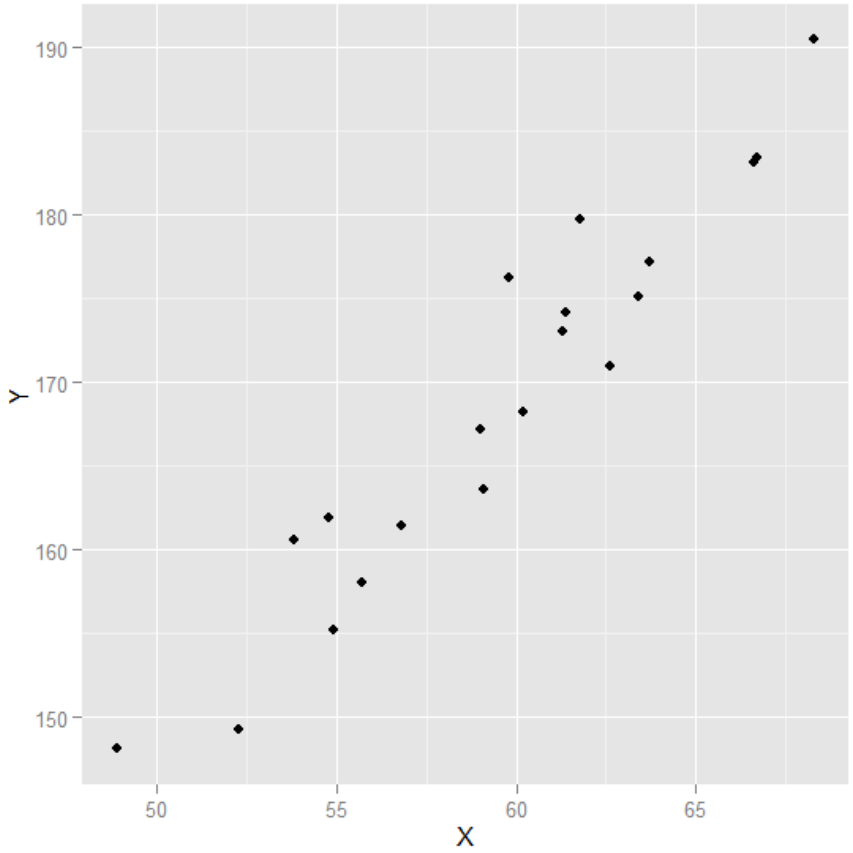
$$\hat{b} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2}$$

$$\hat{a} = \overline{Y} - \hat{b}\overline{X}$$

Ejemplo

```
> ggplot(data,aes(x=X,y=Y))+geom_point()
```

```
> data
      X     Y
1  63.4 175.1
2  54.8 161.9
3  53.8 160.6
4  60.2 168.2
5  61.8 179.8
6  68.3 190.5
7  59.0 167.2
8  61.4 174.2
9  66.6 183.2
10 54.9 155.2
11 66.7 183.4
12 59.8 176.3
13 56.8 161.4
14 62.6 171.0
15 48.9 148.1
16 63.7 177.2
17 61.3 173.1
18 52.3 149.3
19 55.7 158.0
20 59.1 163.6
```



Ejemplo

$$\hat{b} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2}$$

$$\hat{a} = \bar{Y} - \hat{b}\bar{X}$$

```
> data
      X     Y
1  67.3 151.4
2  55.9 128.2
3  67.2 142.9
4  62.8 145.4
5  50.6  99.0
6  60.2 132.8
7  58.2 126.5
8  55.0 113.2
9  55.2 125.6
10 51.8 107.0
11 56.9 121.5
12 66.8 152.6
13 61.5 125.3
14 64.7 128.0
15 63.1 140.1
16 69.4 153.6
17 69.5 140.6
18 69.1 145.8
19 54.4 110.6
20 65.5 145.9
```

$$\sum X_i Y_i = 202210.6$$

$$\sum X_i = 1191.1$$

$$\sum Y_i = 3377.1$$

$$\sum X_i^2 = 71432.85$$

$$n = 20$$

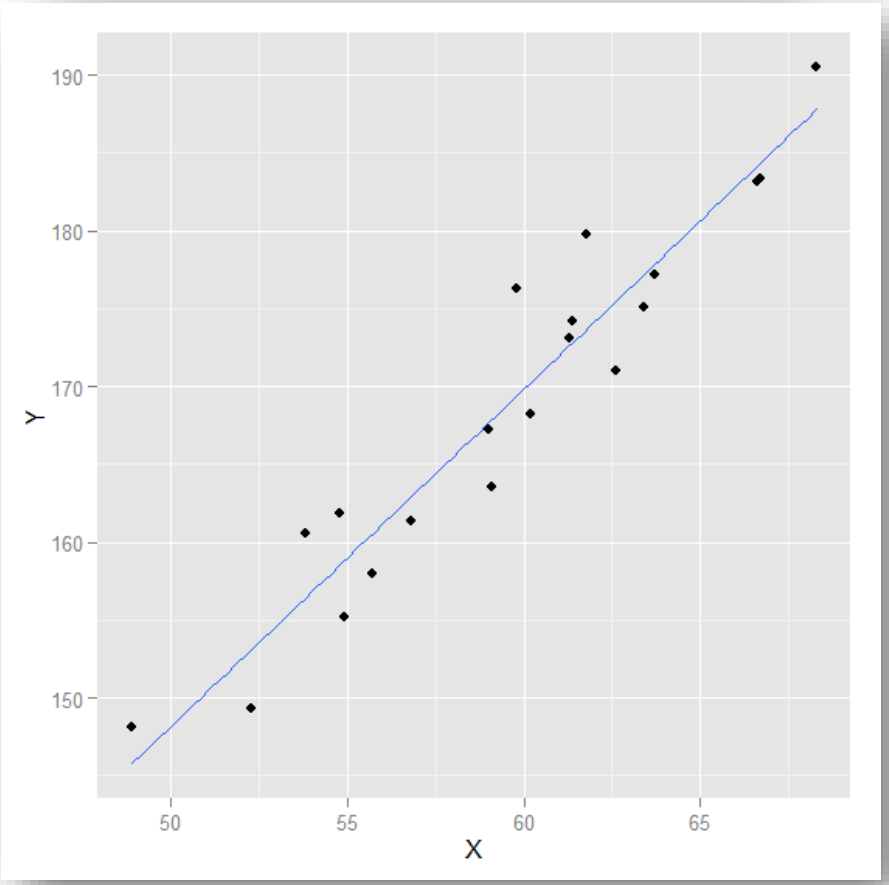
```
> lm(Y~X,data)

Call:
lm(formula = Y ~ X, data = data)

Coefficients:
(Intercept)            X
     39.960        2.164
```

$$\hat{b} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} = 2.164$$

$$\hat{a} = \bar{Y} - \hat{b}\,\bar{X} = 39.96$$

Interpretación de los parámetros

```
> ggplot(data,aes(x=X,y=Y))+
+ geom_point()+
+ geom_smooth(se=F,method=lm)
```



Cuando X vale 0, el valor esperado de Y es 39.86

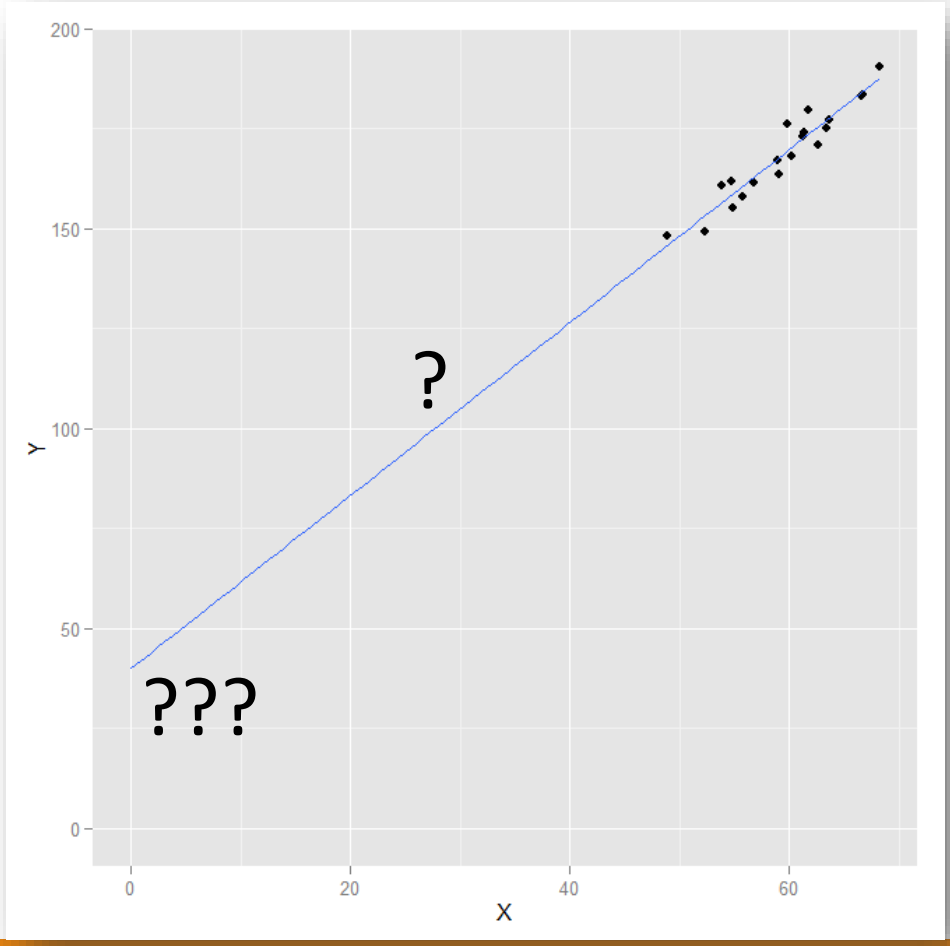
$Y=39.96+2.164*X$

Cuando X aumenta una unidad, el valor esperado de Y aumenta 2.164 unidades

IC de los parámetros

La ordenada en el origen suele tener un IC amplio

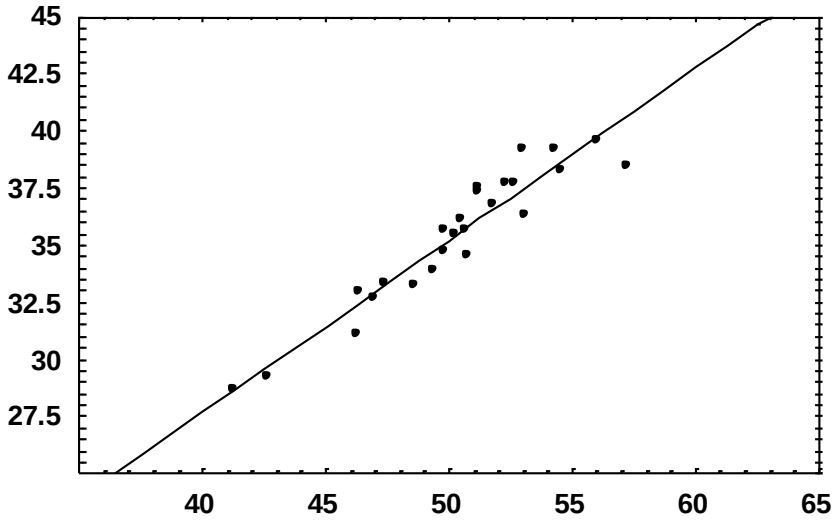
```
> confint(res)
                2.5 %    97.5 %
(Intercept) 20.568404 59.352513
X            1.839981  2.488944
```

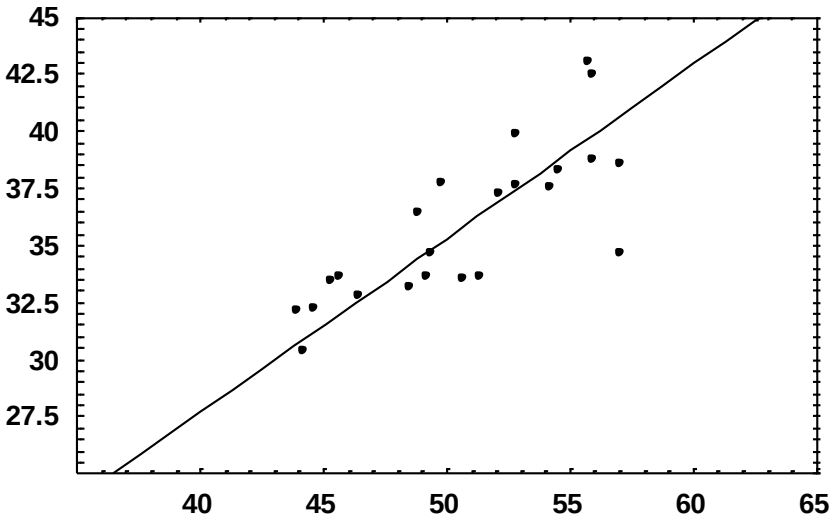


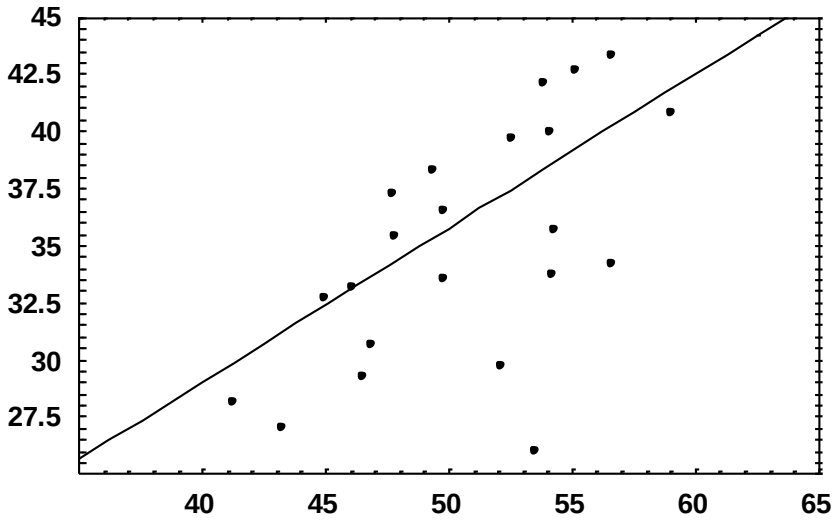
Una recta de regresión no debe utilizarse para extrapolar fuera del intervalo de los datos.

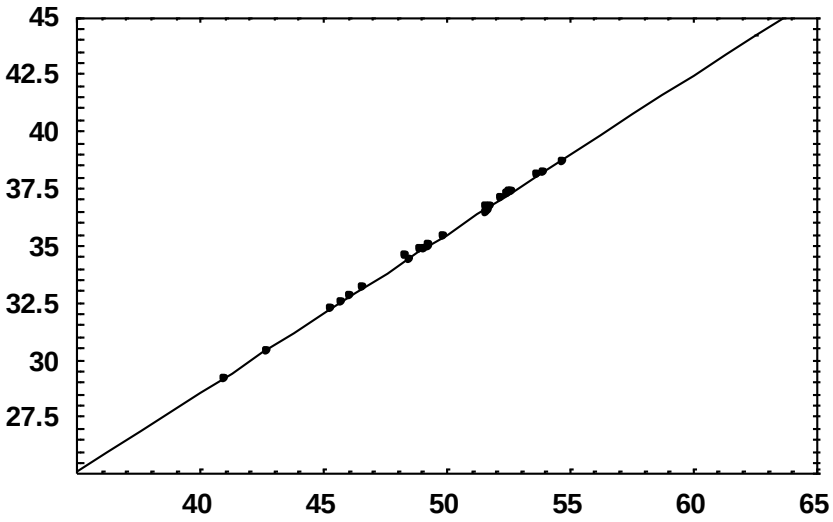
Coeficiente de correlación lineal

¿Hasta qué punto las variables X e Y están relacionadas linealmente?

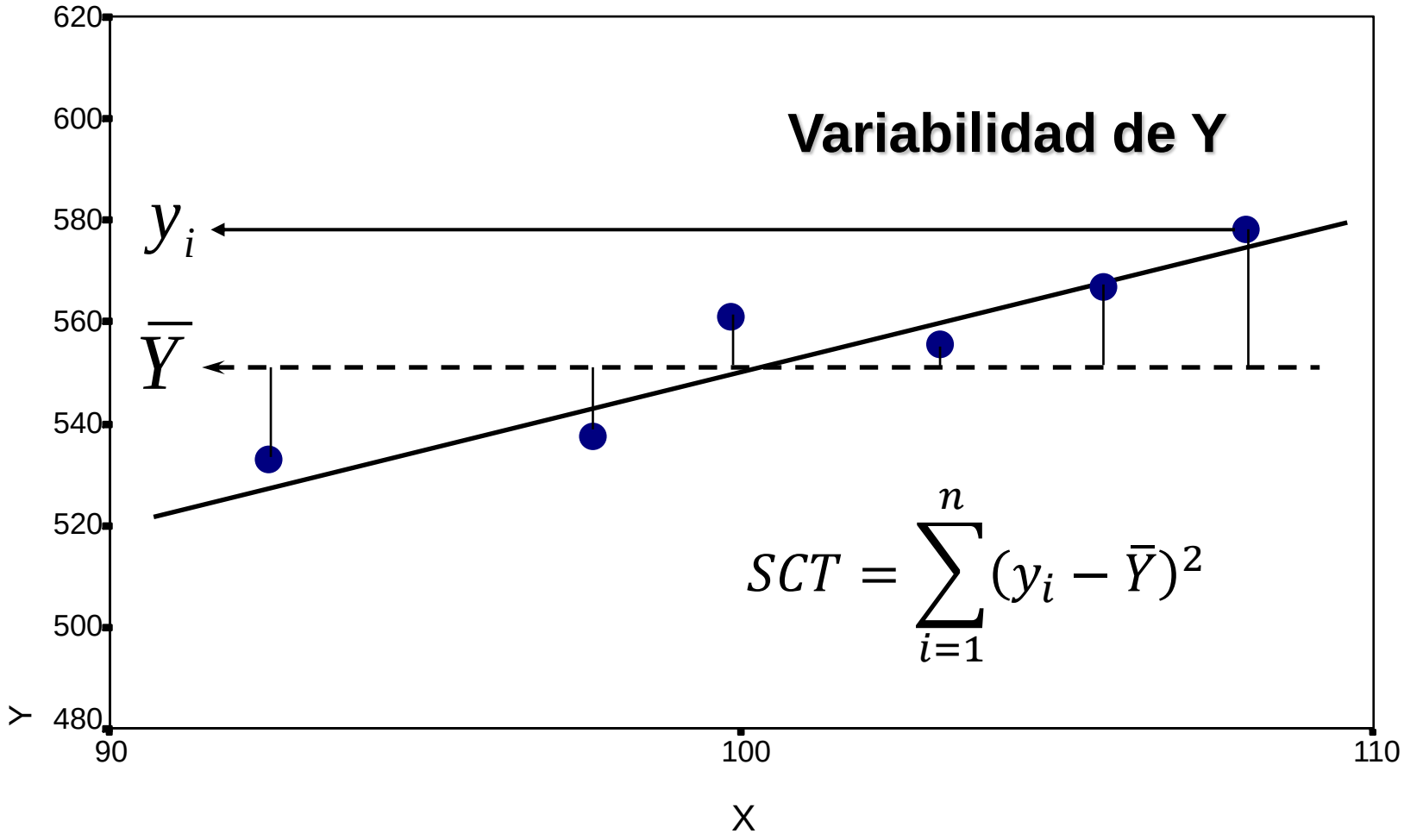




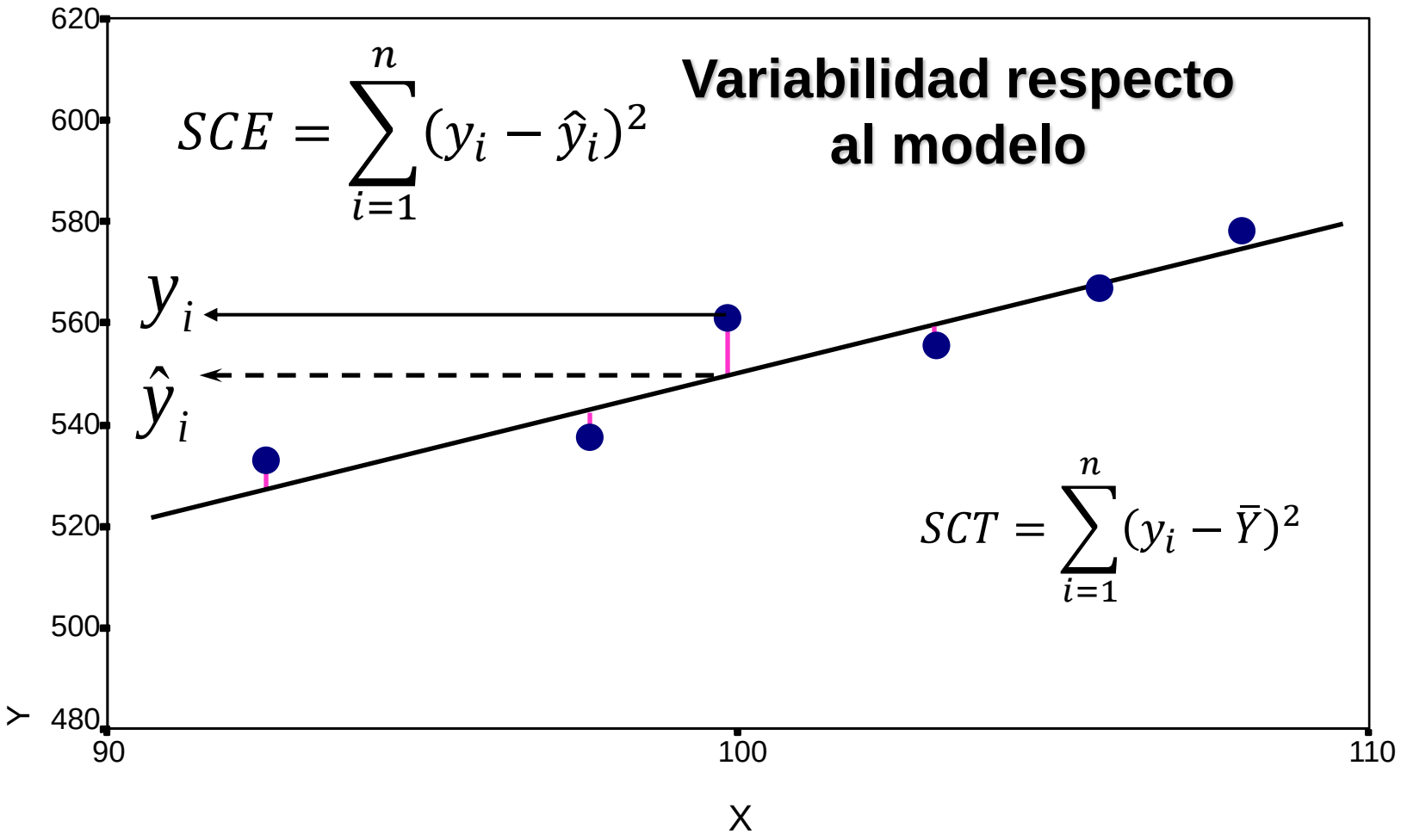




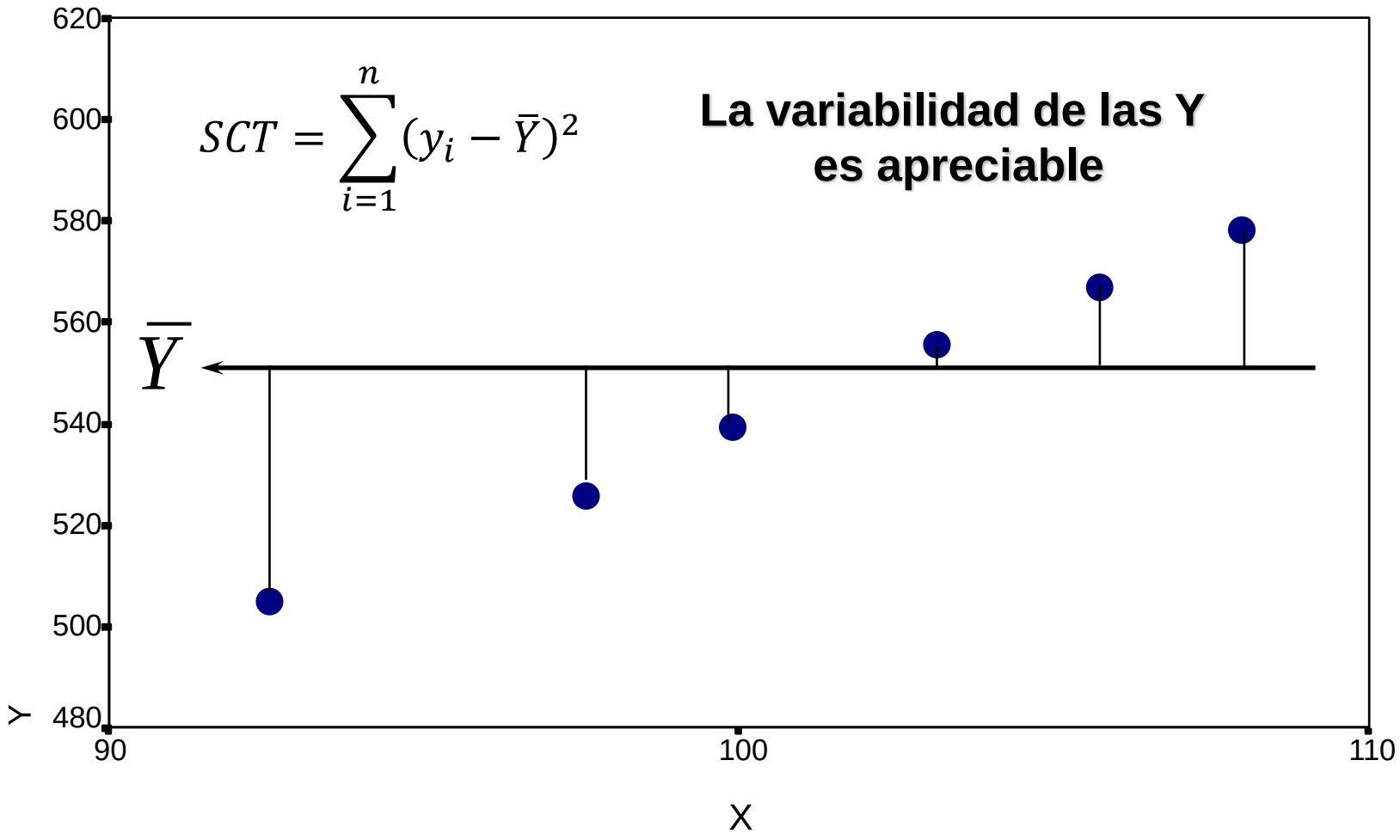
¿Cómo podemos medir el ajuste del modelo lineal?



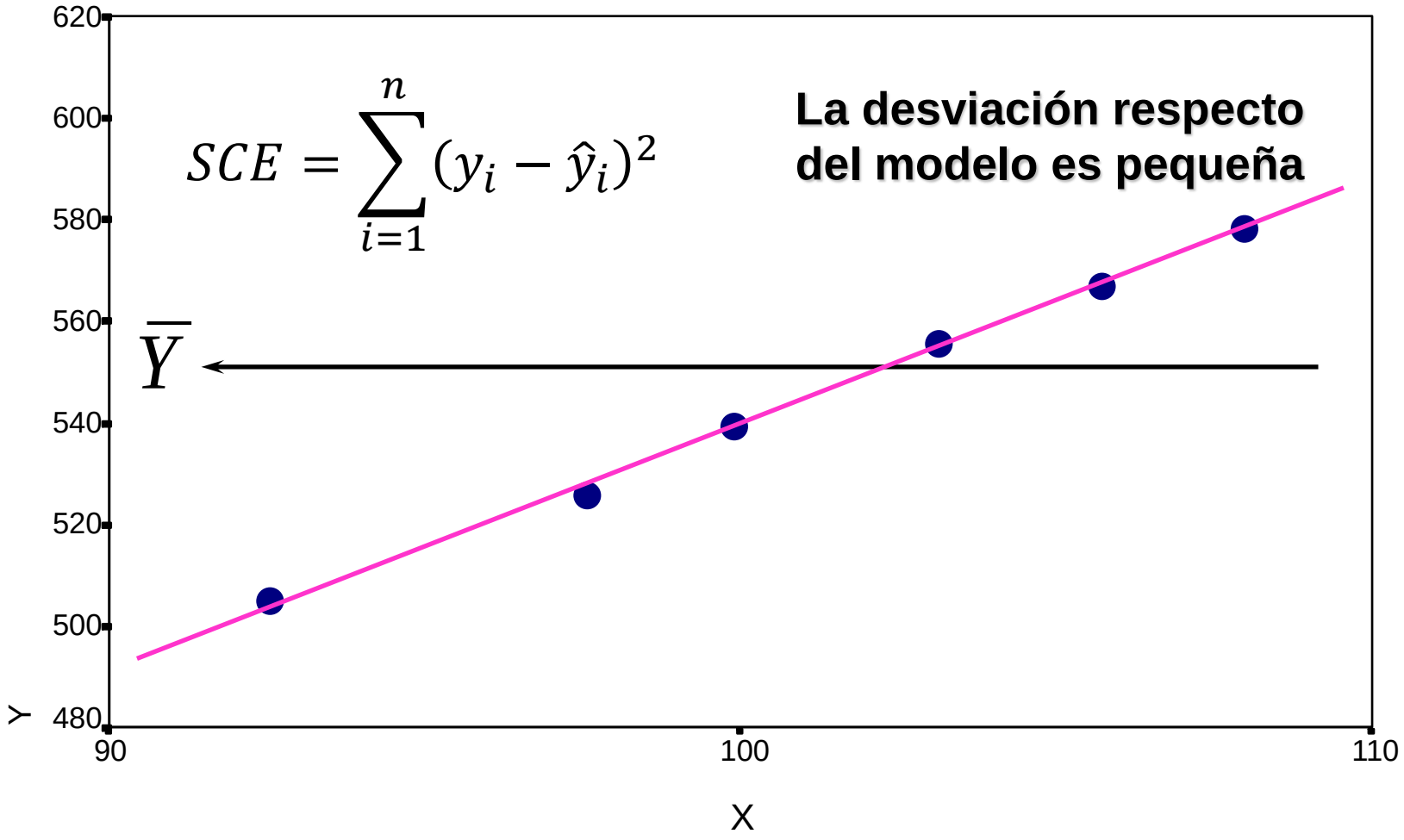
¿Cómo podemos medir el ajuste del modelo lineal?



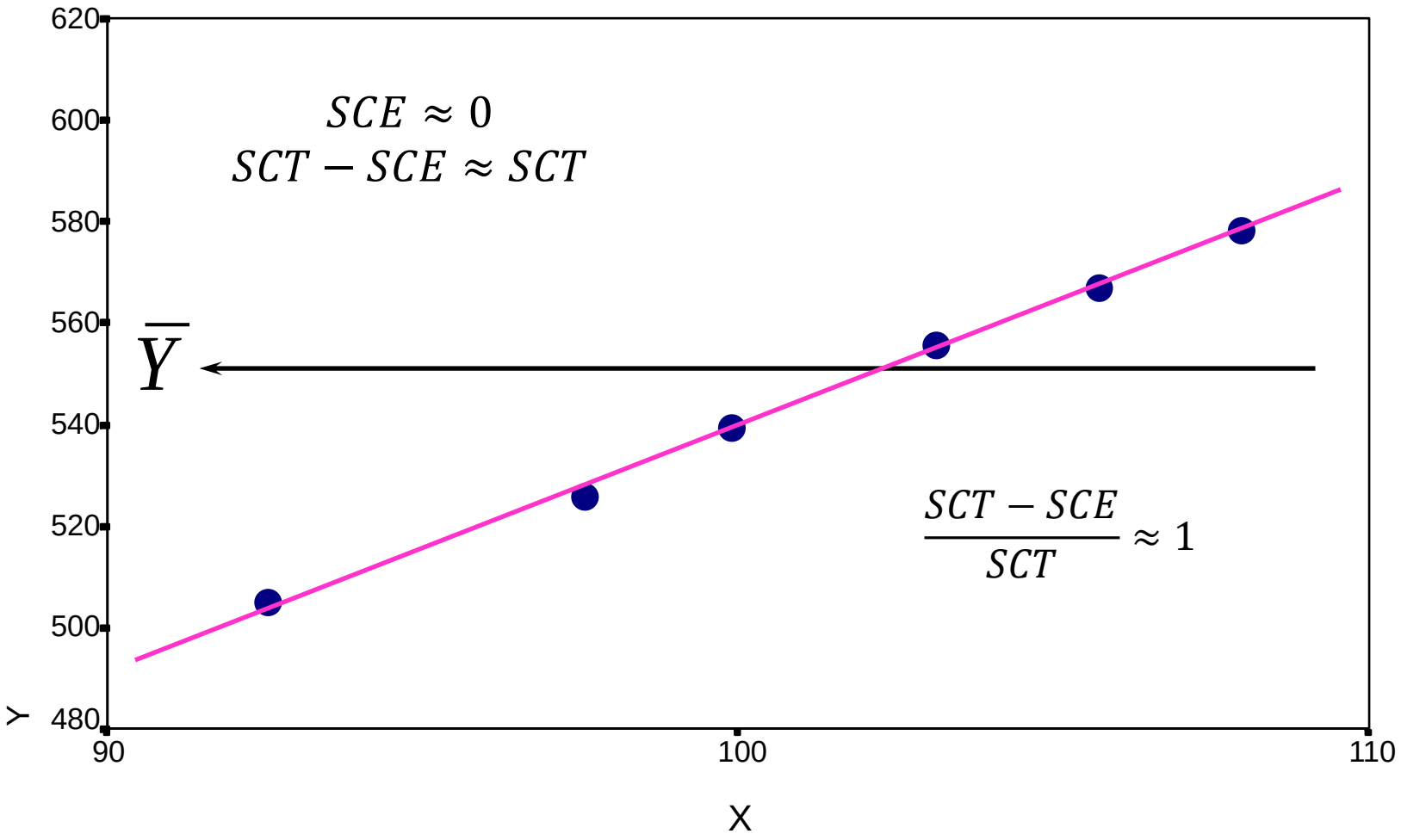
¿Cómo podemos medir el ajuste del modelo lineal?



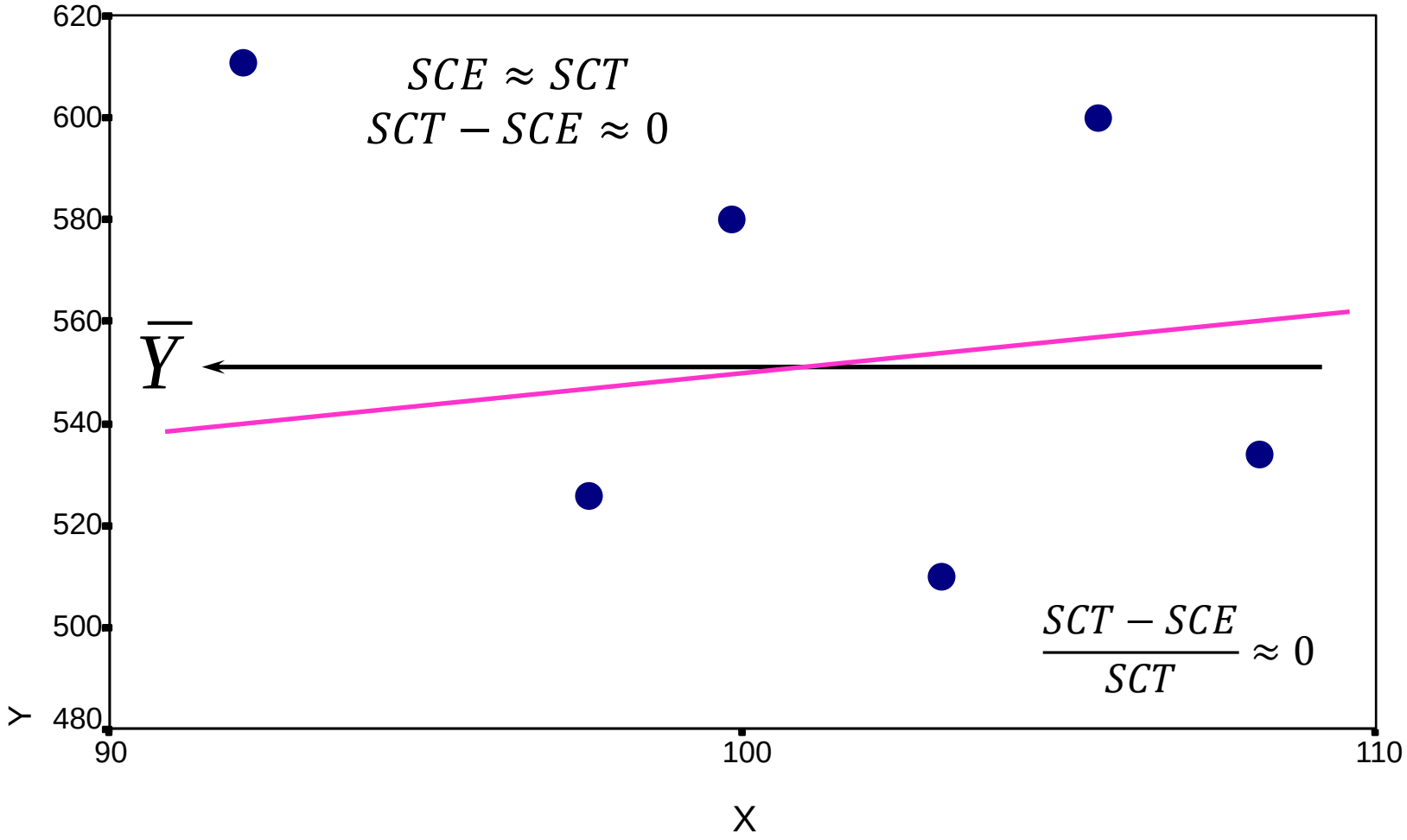
¿Cómo podemos medir el ajuste del modelo lineal?



¿Cómo podemos medir el ajuste del modelo lineal?



¿Cómo podemos medir el ajuste del modelo lineal?



Coeficiente de determinación

Para cualquier modelo de regresión, mide el ajuste del modelo a los datos

$$R^2 = \frac{SCT - SCE}{SCT}$$

$R^2 \to 0$ **Mal ajuste**

$R^2 \to 1$ **Buen ajuste**

Coeficiente de correlación lineal

$$r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}$$

$$S_{xy} = \sum_i X_i Y_i - n\bar{X}\bar{Y}$$

$$S_x^2 = \sum_i X_i^2 - n\bar{X}^2$$

$$S_y^2 = \sum_i Y_i^2 - n\bar{Y}^2$$

Interpretación de las sumas de cuadrados y de R^2

SCT: La suma de cuadrados totales cuantifica la variabilidad de la variable Y (variable dependiente). Si no disponemos de más información, los valores observados de Y se explican como una variabilidad alrededor de la media de las Y.

SCE: Si ajustamos un modelo, la suma de cuadrados del error indica la variabilidad de las observaciones respecto de la predicción del modelo. Si esta variabilidad es baja significa que el modelo explica las observaciones. **En este caso, la variabilidad inicial (SCT) se reduce ya que los valores de Y se explican a partir de su dependencia lineal con X. En el caso de que SCE sea alta, significa que la variabilidad inicial no se explica por el modelo**.

R^2: De acuerdo con su definición, representa el % de variabilidad inicial que queda explicada por el modelo.