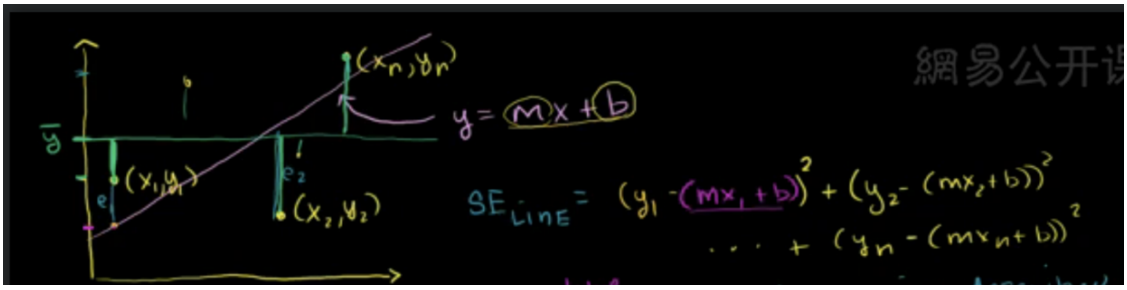
决定系数（coefficient determination）用来衡量回归的好坏,换句话说就是回归拟合的曲线它的拟合优度、

决定系数$R^2$是指y的总波动（variation）情况，可以被回归线描述的部分所占的比例。

y的总波动也就是y到均值的距离平方之和：

$$SE_{\bar{Y}} = (y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + ... + (y_n - \bar{y})^2$$

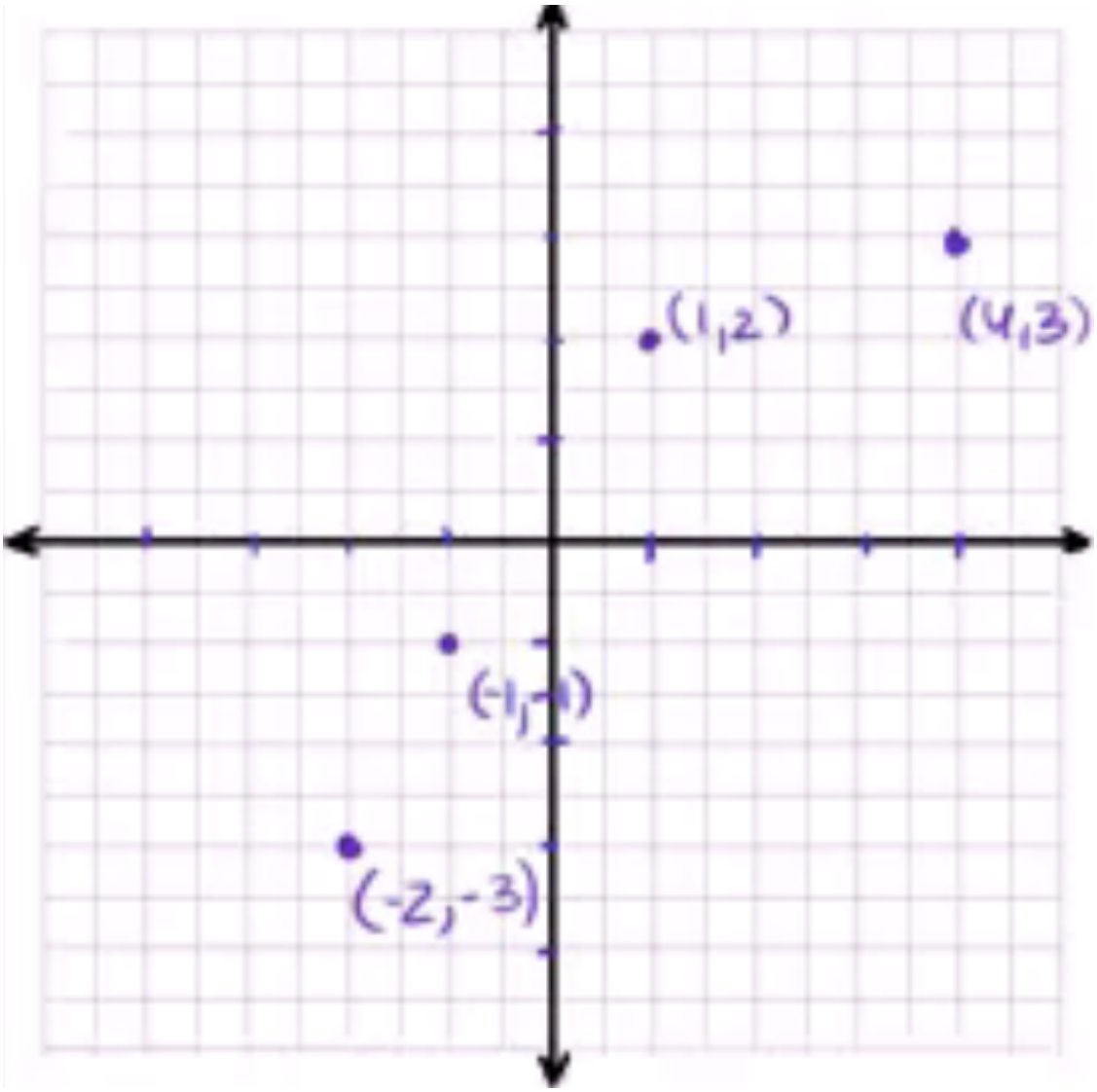
回归线的平方误差显示出总波动中有多少没有被回归线描述：




回归线的平方误差除以y的总波动就是没有被回归线描述的比例。总波动有多少百分比被直线描述则为：

$$1 - \frac{SE_{line}}{SE_{\bar{Y}}}$$

例题：在坐标轴中有如下四个点：




首先，求它的回归线。求得统计量：

$$\bar{x} = \frac{-2-1+1+4}{4} = \frac{1}{2}$$
$$\bar{y} = \frac{-3-1+2+3}{4} = \frac{1}{4}$$
$$\overline{xy} = \frac{6+1+2+12}{4} = \frac{21}{4}$$
$$\overline{x^2} = \frac{4+1+1+16}{4} = \frac{11}{2}$$

m为：

$$m = \frac{\frac{21}{4} - \frac{1}{2} \cdot \frac{1}{4}}{\frac{11}{2} - (\frac{1}{2})^2} = \frac{41}{42}$$

b为：

$$\frac{1}{4}-\frac{41}{42}\cdot\frac{1}{2}=-\frac{5}{21}$$

所以，回归线为：

$$y=\frac{41}{42}x-\frac{5}{21}$$

接下来，通过决定系数来判断回归线的拟合程度。

(1) 求总误差的平方。实际值减去预测值的平方和。对于(-2,-3)来说，它的实际值就是-3，而将-2代入回归线方程中，得到的就是预测值：

$$\frac{41}{42}\cdot(-2)-\frac{5}{21}=-2.1905$$

同理求得其它三个点的预测值：

| x | y | **预测值** |
| --- | --- | --- |
| -2 | -3 | -2.1905 |
| -1 | -1 | -1.2143 |
| 1 | 2 | 0.7381 |
| 4 | 3 | 3.6667 |

总误差平方：

$$(-3-(-2.1905))^2+(-1+1.2143)^2+(2-0.7381)^2+(3-3.6667)^2=2.738$$

(2) y离y均值的距离，也就是实际值减去y均值的平方和：

$$(-3-0.25)^2+(-1-0.25)^2+(2-0.25)^2+(3-0.35)^2=22.75$$

(3) 总波动除以离均值的平方和：也就是2.738÷22.75 = 0.12。也就是说有12%的总波动(variation)无法由x波动来解释。反过来，也就是说88%的y的总波动能被回归线或x波动解释。