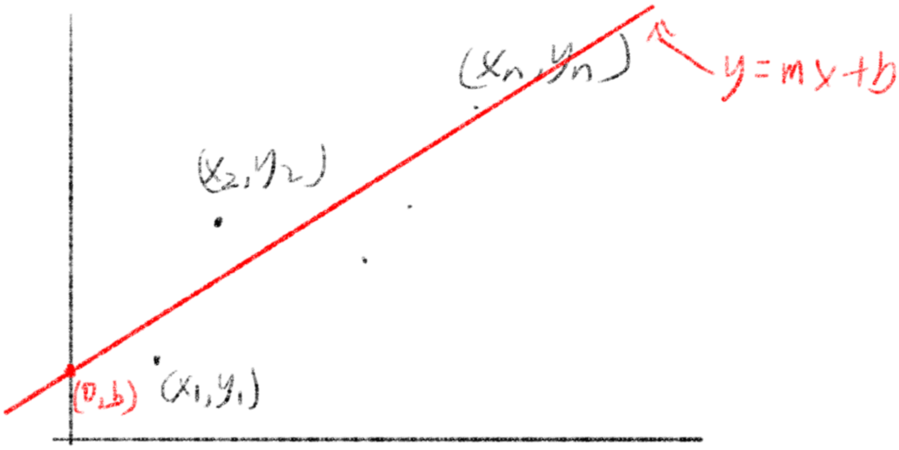
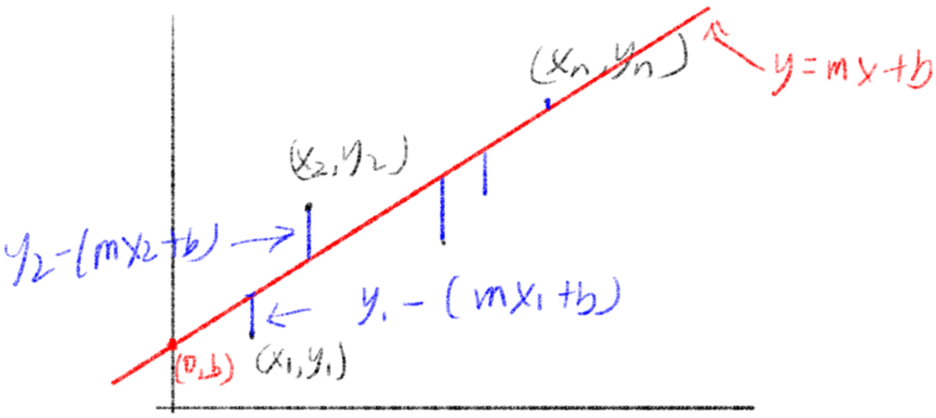
1. 直线与点的误差平方的表达式

假设坐标平面内有n点。我这里希望找到一条直线，最小化这些点到直线的平方误差。这是什么样的直线



我们要求的就是斜率m与y轴的截距b。如何确定m与b呢?

我们知道，每个点同直线的误差，也就是它到直线的垂直距离：



我们要做的，不是直接将这些误差加起来，而是将这些误差的平方加起来，然后最小化。这条线对应的平方误差等于所有这些平方误差之和：

$$(y_1-(m\cdot x_1+b))^2+(y_2-(m\cdot x_2+b))^2+...+(y_n-(m\cdot x_n+b))^2$$

接下来要做的是求出m和b使得整个误差最小。

2. 推导

2.1 化简

$$
\begin{aligned}
&(y_1-(m\cdot x_1+b))^2+(y_2-(m\cdot x_2+b))^2+\ldots+(y_n-(m\cdot x_n+b))^2\\
=&\quad y_1^2-2y_1(m\cdot x_1+b)+(m\cdot x_1+b)^2\\
&\quad+y_2^2-2y_2(m\cdot x_2+b)+(m\cdot x_2+b)^2\\
&\quad\vdots\\
&\quad+y_n^2-2y_2(m\cdot x_n+b)+(m\cdot x_n+b)^2\\
=&\quad y_1^2-2y_1mx_1-2y_1b+m^2x_1^2+2mx_1b+b^2\\
&\quad+y_2^2-2y_2mx_1-2y_2b+m^2x_2^2+2mx_2b+b^2\\
&\quad\vdots\\
&\quad+y_n^2-2y_nmx_n-2y_nb+m^2x_n^2+2mx_nb+b^2
\end{aligned}
$$

接下来化简。上面的表达式可以分组为：

$$
\begin{aligned}
&(y_1^2+y_2^2+\ldots+y_n^2)-2m(x_1y_1+x_2y_2+\ldots+x_ny_n)-2b(y_1+y_2+\ldots+y_n)\\
&+m^2(x_1^2+x_2^2+\ldots+x_n^2)+2mb(x_1+x_2+\ldots+x_n)+nb^2
\end{aligned}
$$

y平方的均值为：

$$\overline{y^2}=\frac{y_1^2+y_2^2+\ldots+y_n^2}{n}$$

也就是说：

$$y_1^2+y_2^2+\ldots+y_n^2=n\cdot\overline{y^2}$$

同理：

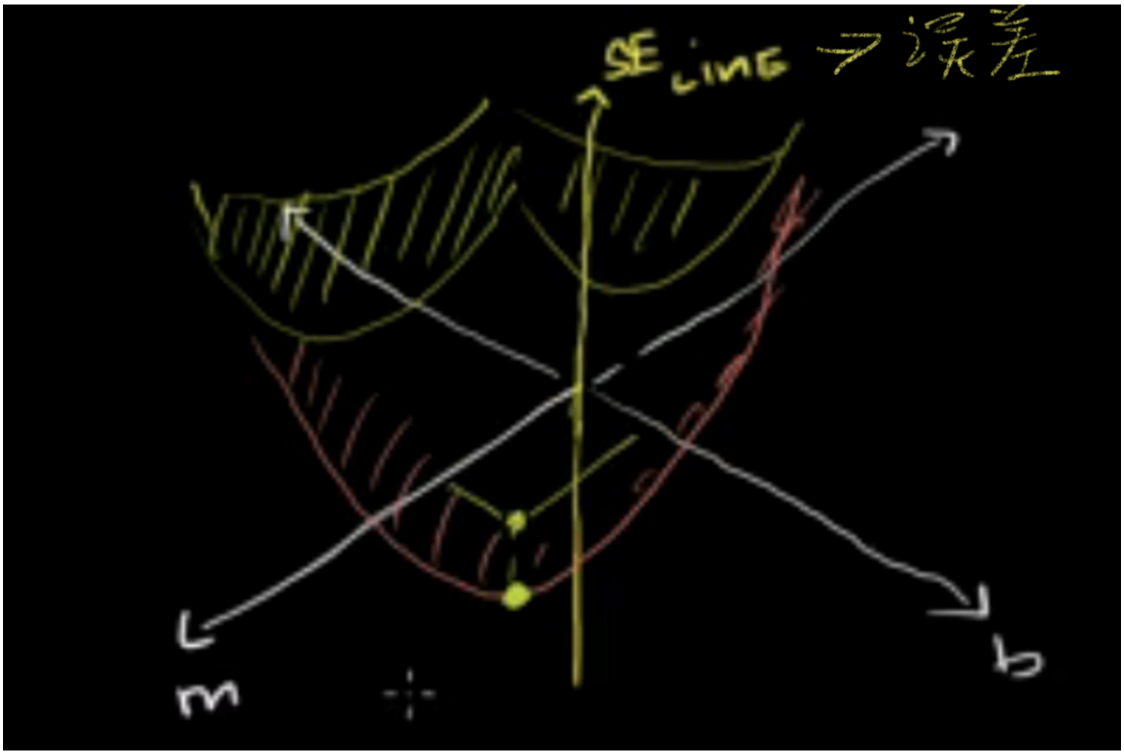
$$
\begin{aligned}
&x_1y_1+x_2y_2+\ldots+x_ny_n=n\cdot\overline{xy}\\
&y_1+y_2+\ldots+y_n=n\cdot\bar{y}\\
&x_1^2+x_2^2+\ldots+x_n^2=n\cdot\overline{x^2}\\
&x_1+x_2+\ldots+x_n=n\cdot\bar{x}
\end{aligned}
$$

所以，表达式简化为了：

$$n \cdot y^2 - 2mn \cdot \overline{xy} - 2bn \cdot \bar{y} + m^2 n \cdot \overline{x^2} + 2mbn \cdot \bar{x} + nb^2$$

2.2 最小化m和b

这将用到三维微积分知识。上面的表达式在三维图像中就是一个曲面。



我们要求的最小误差也就是误差相对m以及b的偏导为0：

$$\frac{\partial SE}{\partial m} = \frac{\partial SE}{\partial m} = 0$$

> 偏导数其实和普通导数求法一样，只是除了求偏导的变量以外，其它都看作是常数。

对m求偏导时，唯一的变量是m，求导：

$$0 - 2n\overline{xy} + 2n\overline{x^2}m + 2bn\bar{x} + 0$$

令其最小化,也就是求偏导等于0：

$$\begin{aligned} &-2n\overline{xy} + 2n\overline{x^2}m + 2bn\bar{x} = 0 \\ &\Rightarrow -\overline{xy} + \overline{x^2}m + b\bar{x} = 0 \end{aligned}$$

同理，对b求偏导，并且最小化：

$$
\begin{aligned}
& -2n\bar{y} + 2mn\bar{x} + 2nb = 0 \\
& \Rightarrow -\bar{y} + m\bar{x} + b = 0
\end{aligned}
$$

这里相当于一元二次方程，未知数分别是m和b。将这两个表达式转化为mx+b = y的形式：

$$
\begin{aligned}
&(1) - \overline{xy} + \overline{x^2}m + b\bar{x} = 0 \\
&\qquad \Rightarrow \overline{x^2}m + b\bar{x} = \overline{xy} \\
&\qquad \Rightarrow m\frac{\overline{x^2}}{\bar{x}} + b = \frac{\overline{xy}}{\bar{x}} \\
&(2) - \bar{y} + m\bar{x} + b = 0 \\
&\qquad \Rightarrow m\bar{x} + b = \bar{y}
\end{aligned}
$$

> 也就是说，这个拟合的直线将包含两个点： $(\bar{x}, \bar{y}), (\frac{\overline{x^2}}{\bar{x}}, \frac{\overline{xy}}{\bar{x}})$

最后求得m和b：

$$
\begin{aligned}
& m(\bar{x} - \frac{\overline{x^2}}{\bar{x}}) = \bar{y} - \frac{\overline{xy}}{\bar{x}} \\
& \Rightarrow m = \frac{\bar{y} - \frac{\overline{xy}}{\bar{x}}}{\bar{x} - \frac{\overline{x^2}}{\bar{x}}} = \frac{\bar{x}\bar{y} - \overline{xy}}{(\bar{x})^2 - \overline{x^2}} \\
& b = \bar{y} - m\bar{x}
\end{aligned}
$$

此外，有的书上m写作：

$$
m = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - (\bar{x})^2}
$$

这是我们求得的m分子与分母同时乘以-1。两者是等价的。

例如：有三个点：(1,2),(2,1),(4,3)。求出最佳拟合直线

$$\bar{x} = \frac{1+2+4}{3} = \frac{7}{3}$$
$$\bar{y} = \frac{2+1+3}{3} = 2$$
$$\overline{xy} = \frac{1\cdot 2 + 2\cdot 1 + 4\cdot 3}{3} = \frac{16}{3}$$
$$\overline{x^2} = \frac{1+2^2+4^2}{3} = 7$$

则m等于：

$$m = \frac{\frac{16}{3} - 2\frac{7}{3}}{7 - (\frac{7}{3})^2} = \frac{3}{7}$$

b等于：

$$b = 2 - \frac{3}{7}\cdot\frac{7}{3} = 1$$

求得回归直线为：

$$y = \frac{3}{7}x + 1$$