MATH 2022 Linear and Abstract Algebra

LECTURE 37 Tuesday 28/05/2019

An application to linear regression

Problem: Find or estimate a linear relationship

$$Y = aX + b$$

between random variables X and Y.

- producing a line "of best fit"
- navigating through values that fluctuate "randomly" or "unpredictably"
- making sense out of "uncertainty"
- maximizing intrinsic value of data points

An application to linear regression

Problem: Find or estimate a linear relationship

$$Y = aX + b$$

between random variables X and Y.

We sample n pairs of measurements:

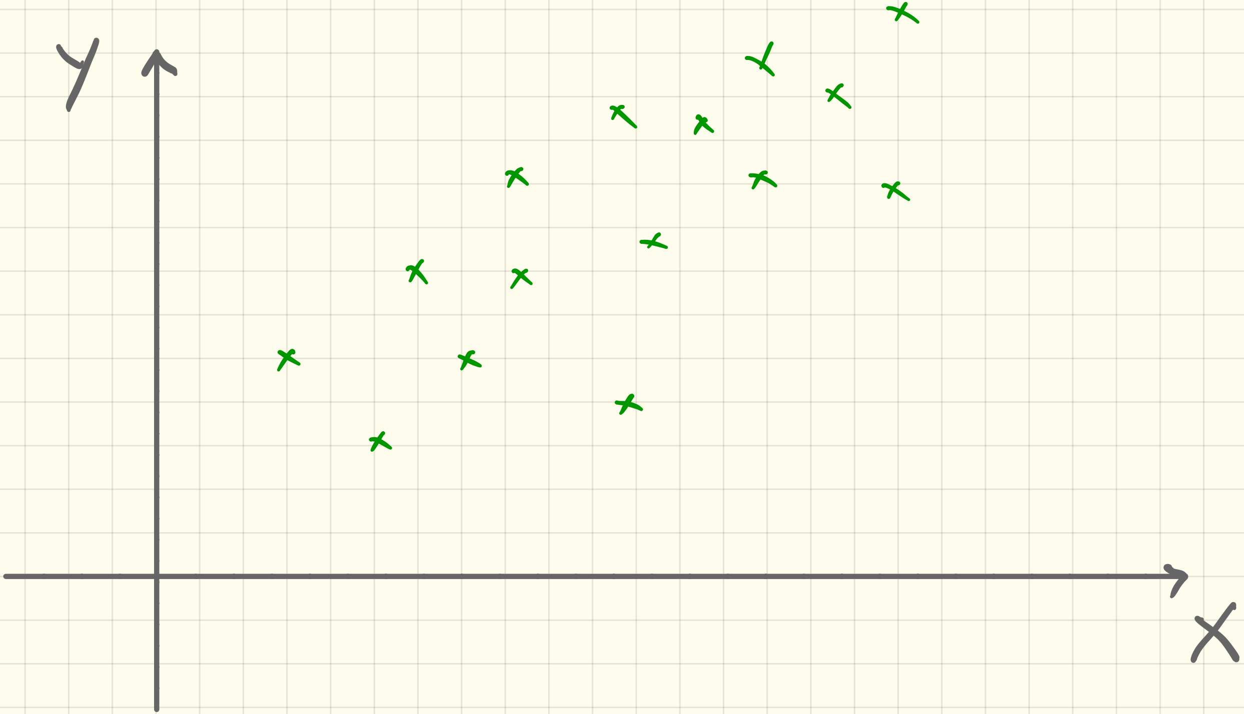
$$(x_1, y_1), \quad (x_2, y_2), \quad \ldots, \quad (x_n, y_n)$$

visualised as a scatter graph:

We sample n pairs of measurements :

$$(x_1, y_1) \; , \; (x_2, y_2) \; , \; \dots , \; (x_n, y_n)$$

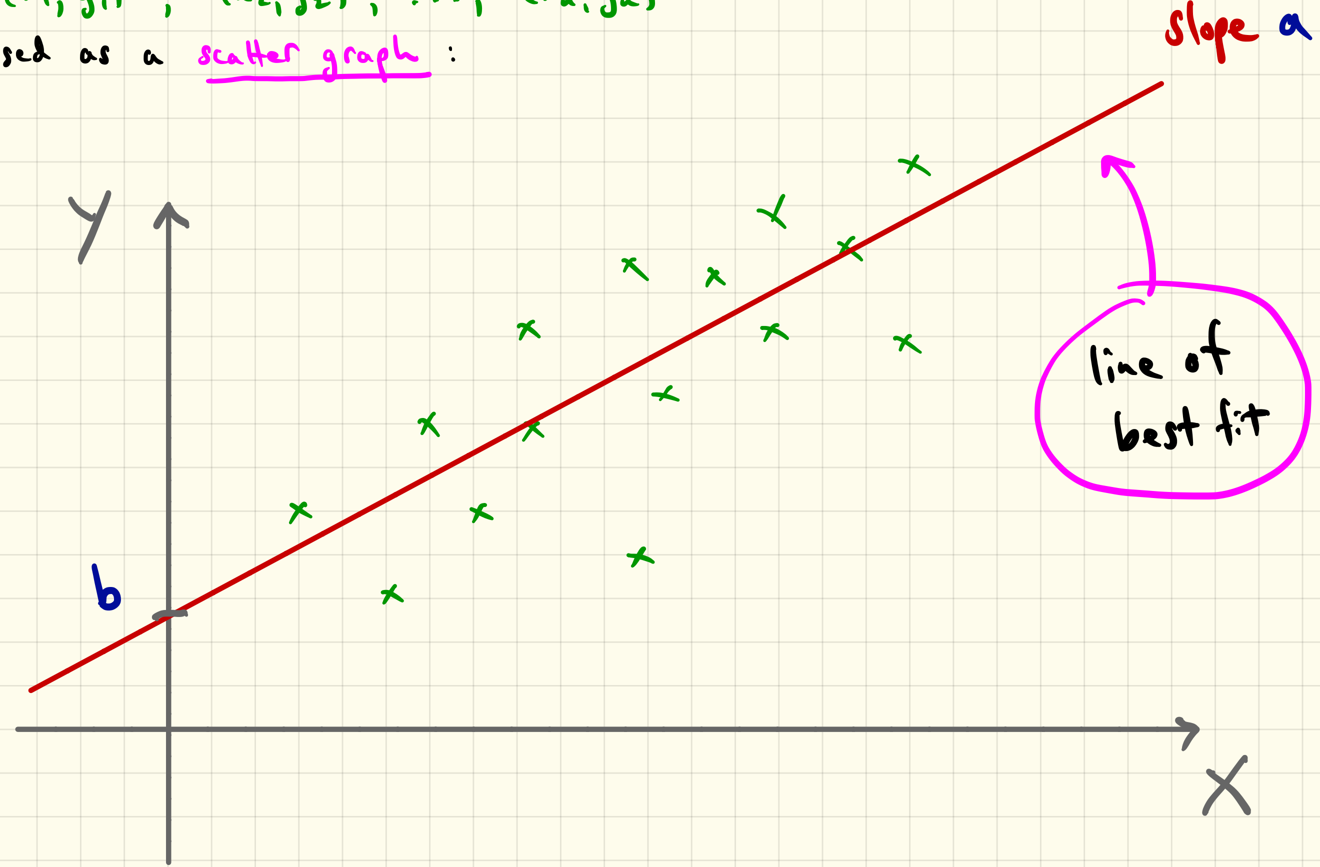
visualised as a **scatter graph** :



We sample n pairs of measurements:

$$(x_1, y_1), \quad (x_2, y_2), \quad \ldots, \quad (x_n, y_n)$$

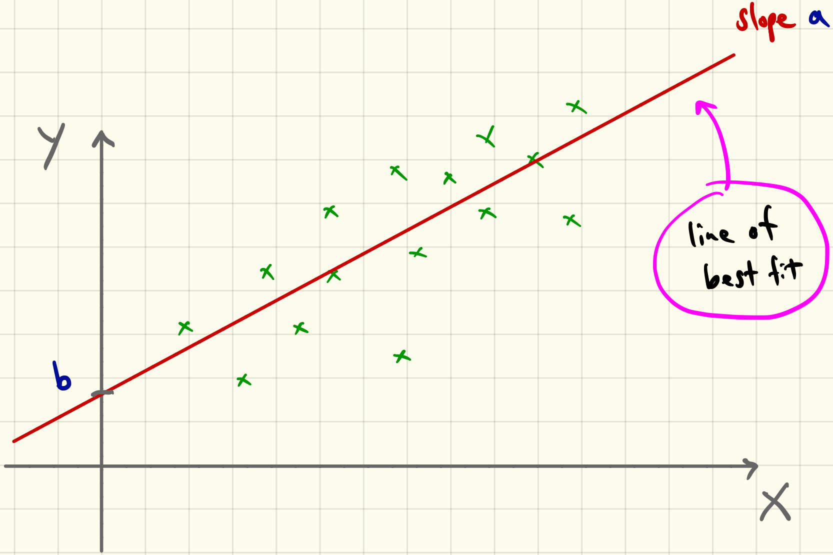
visualised as a scatter graph:



We sample n pairs of measurements :

$$(x_1, y_1) \ , \ (x_2, y_2) \ , \ \dots , \ (x_n, y_n)$$

visualised as a scatter graph :



Put

$$\underset{\sim}{x} = (x_1, \dots, x_n) \quad , \quad \underset{\sim}{y} = (y_1, \dots, y_n)$$

$$\underset{\sim}{1} = (1, 1, \dots, 1)$$

elements of $\mathbb{R}^n$

Put

$$\underset{\sim}{x} = (x_1, \dots, x_n) \quad , \quad \underset{\sim}{y} = (y_1, \dots, y_n)$$

$$\underset{\sim}{1} = (1, 1, \dots, 1) \qquad \text{elements of } \mathbb{R}^n$$

Then, using the usual inner product on $\mathbb{R}^n$:

$$\langle \underset{\sim}{x}, \underset{\sim}{1} \rangle = x_1 + \dots + x_n = \sum_{i=1}^{n} x_i \,,$$

$$\langle \underset{\sim}{y}, \underset{\sim}{1} \rangle = y_1 + \dots + y_n = \sum_{i=1}^{n} y_i \,,$$

$$\langle \underset{\sim}{x}, \underset{\sim}{y} \rangle = x_1 y_1 + \dots + x_n y_n = \sum_{i=1}^{n} x_i y_i \,.$$

Put

$$\mathcal{C} = \text{subspace of } \mathbb{R}^n \text{ spanned by } \underset{\sim}{1}$$

Put
$$\mathcal{C} = \text{subspace of } \mathbb{R}^n \text{ spanned by } \underset{\sim}{1}$$
$$= \{ \lambda \underset{\sim}{1} \mid \lambda \in \mathbb{R} \}$$
$$= \{ (\lambda, \lambda, \dots, \lambda) \mid \lambda \in \mathbb{R} \},$$
the subspace of "constant" vectors. Then

$$\boxed{\{ \underset{\sim}{1} \} \text{ is a basis for } \mathcal{C}.}$$

Put
$$\hat{\underset{\sim}{1}} = \frac{\underset{\sim}{1}}{\|\underset{\sim}{1}\|} = \frac{1}{\sqrt{n}} \underset{\sim}{1} = \left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}} \right).$$

Then
$$\boxed{\{ \hat{\underset{\sim}{1}} \} \text{ is an orthonormal basis for } \mathcal{C}.}$$

Observe that

$$\langle \underset{\sim}{y}, \hat{\underset{\sim}{1}} \rangle = \langle \underset{\sim}{y}, \frac{1}{\sqrt{n}} \underset{\sim}{1} \rangle = \frac{1}{\sqrt{n}} \langle \underset{\sim}{y}, \underset{\sim}{1} \rangle$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} y_i$$

so that

$$\langle \underset{\sim}{y}, \hat{\underset{\sim}{1}} \rangle \hat{\underset{\sim}{1}} = \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} y_i \right) \frac{1}{\sqrt{n}} \underset{\sim}{1}$$

$$= \left(\frac{1}{n} \sum_{i=1}^{n} y_i \right) \underset{\sim}{1} .$$

Hence

$$\boxed{\mathrm{proj}_{\ell} \, \underset{\sim}{y} = \bar{y} \, \underset{\sim}{1}}$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ = mean of $y_1, \ldots, y_n$.

Hence

$$\boxed{\operatorname{proj}_{\ell} \underset{\sim}{y} = \bar{y}\, \underset{\sim}{1}}$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ = mean of $y_1, \ldots, y_n$.

Similarly

$$\boxed{\operatorname{proj}_{\ell} \underset{\sim}{x} = \bar{x}\, \underset{\sim}{1}}$$

where $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ = mean of $x_1, \ldots, x_n$.

Note that

$$\boxed{\sum_{i=1}^{n} (x_i - \bar{x}) = 0}$$

as expected, since $\langle \underset{\sim}{x} - \operatorname{proj}_{\ell} \underset{\sim}{x}, \underset{\sim}{1} \rangle = 0$.

The distance from $\underset{\sim}{x}$ to $\mathcal{b}$ is

$$\| \underset{\sim}{x} - \text{proj}_{\mathcal{b}} \underset{\sim}{x} \| = \| \underset{\sim}{x} - \bar{x} \underset{\sim}{1} \|$$

$$= \langle x - \bar{x} \underset{\sim}{1} , x - \bar{x} \underset{\sim}{1} \rangle^{1/2}$$

$$= \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= S_x$$

where

$$\boxed{S_x^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

the sum of the <u>squared deviations from the mean</u>.

The distance from $\underset{\sim}{x}$ to $\mathcal{L}$ is

$$\| \underset{\sim}{x} - \text{proj}_{\mathcal{L}} \underset{\sim}{x} \| = \| \underset{\sim}{x} - \bar{x} \underset{\sim}{1} \|$$

$$= \langle x - \bar{x} \underset{\sim}{1}, x - \bar{x} \underset{\sim}{1} \rangle^{1/2}$$

$$= \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= S_x$$

where $\boxed{S_x^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2}$

the sum of the squared deviations from the mean.

$\frac{S_x^2}{n}$ is known as the variance of $x_1, \ldots, x_n$

Similarly, $\| \underset{\sim}{y} - \text{proj}_{\mathcal{L}} \underset{\sim}{y} \| = S_y$

where

$$\boxed{S_y^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

the sum of the squared deviations from the mean and

$$\boxed{\frac{S_y^2}{n} \text{ is known as the } \underline{\text{variance}} \text{ of } y_1, \dots, y_n .}$$

Put

$$\boxed{\mathcal{L} = \text{subspace of } \mathbb{R}^n \text{ spanned by } \underset{\sim}{1} \text{ and } \underset{\sim}{x}}$$

so a basis for $\mathcal{L}$ is $\{ \underset{\sim}{1}, \underset{\sim}{x} \}$

assuming $\underset{\sim}{x}$ is not a constant vector!!

Put

$\boxed{\mathcal{L} = \text{subspace of } \mathbb{R}^n \text{ spanned by } \underset{\sim}{1} \text{ and } \underset{\sim}{x}}$

so a basis for $\mathcal{L}$ is $\{\underset{\sim}{1}, \underset{\sim}{x}\}$

assuming $\underset{\sim}{x}$ is not a constant vector!!

If $\underset{\sim}{y}$ is "near" $\mathcal{L}$ then

$$\boxed{\underset{\sim}{y} \approx a\underset{\sim}{x} + b\underset{\sim}{1}} \qquad \exists\, a, b \in \mathbb{R}$$

approximately equal to

which would be evidence of a linear relationship

$$\boxed{Y = aX + b}.$$

How close is $\underset{\sim}{y}$ to $\mathcal{L}$?

What is the "nearest point" (vector) in $\mathcal{L}$ to $\underset{\sim}{y}$?

We want $\text{proj}_{\mathcal{L}}\, \underset{\sim}{y}$

First apply Gram-Schmidt:

(1) Put $\underset{\sim}{b}_1 = \hat{\underset{\sim}{1}} = \frac{1}{\sqrt{n}}\, \underset{\sim}{1}$.

(2) Put

$$\underset{\sim}{b}_2 = \frac{\underset{\sim}{x} - \langle \underset{\sim}{x}, \underset{\sim}{b}_1 \rangle\, \underset{\sim}{b}_1}{\| \underset{\sim}{x} - \langle \underset{\sim}{x}, \underset{\sim}{b}_1 \rangle\, \underset{\sim}{b}_1 \|}$$

$$= \frac{\underset{\sim}{x} - \text{proj}_{\ell}\, \underset{\sim}{x}}{\| \underset{\sim}{x} - \text{proj}_{\ell}\, \underset{\sim}{x} \|}$$

$$= \frac{1}{S_x} \left(\underset{\sim}{x} - \bar{x}\, \underset{\sim}{1} \right)$$

How close is $\underset{\sim}{y}$ to $\mathcal{L}$?

What is the "nearest point" (vector) in $\mathcal{L}$ to $\underset{\sim}{y}$?

We want

$\text{proj}_{\mathcal{L}}\, \underset{\sim}{y}$

First apply Gram-Schmidt :

(1) Put $\underset{\sim}{b}_1 = \hat{\underset{\sim}{1}} = \frac{1}{\sqrt{n}} \underset{\sim}{1}$.

(2) Put

$$\underset{\sim}{b}_2 = \frac{\underset{\sim}{x} - \langle \underset{\sim}{x}, \underset{\sim}{b}_1 \rangle \underset{\sim}{b}_1}{\| \underset{\sim}{x} - \langle \underset{\sim}{x}, \underset{\sim}{b}_1 \rangle \underset{\sim}{b}_1 \|}$$

$$= \frac{\underset{\sim}{x} - \text{proj}_{\ell}\, \underset{\sim}{x}}{\| \underset{\sim}{x} - \text{proj}_{\ell}\, \underset{\sim}{x} \|}$$

$$= \frac{1}{S_x} (\underset{\sim}{x} - \bar{x} \underset{\sim}{1})$$

Thus an orthonormal basis for $\mathcal{L}$ is

$$\{ \underset{\sim}{b}_1, \underset{\sim}{b}_2 \} = \left\{ \frac{1}{\sqrt{n}} \underset{\sim}{1} , \frac{1}{S_x} (\underset{\sim}{x} - \bar{x} \underset{\sim}{1}) \right\}$$

$$\boxed{\text{Thus an orthonormal basis for } \mathcal{L} \text{ is} \quad \{\underset{\sim}{b}_1, \underset{\sim}{b}_2\} = \left\{ \frac{1}{\sqrt{n}} \underset{\sim}{1}, \; \frac{1}{S_x}(\underset{\sim}{x} - \bar{x}\underset{\sim}{1}) \right\}}$$

We want

$$\boxed{\text{proj}_{\mathcal{L}} \underset{\sim}{y} = \langle \underset{\sim}{y}, \underset{\sim}{b}_1 \rangle \underset{\sim}{b}_1 + \langle \underset{\sim}{y}, \underset{\sim}{b}_2 \rangle \underset{\sim}{b}_2}$$

(1) $\langle \underset{\sim}{y}, \underset{\sim}{b}_1 \rangle \underset{\sim}{b}_1 = \text{proj}_{\ell} \underset{\sim}{y} = \bar{y} \underset{\sim}{1}$.

(2) $\langle \underset{\sim}{y}, \underset{\sim}{b}_2 \rangle = \left\langle \underset{\sim}{y}, \frac{1}{S_x}(\underset{\sim}{x} - \bar{x}\underset{\sim}{1}) \right\rangle$

$$= \frac{1}{S_x}\left(\langle \underset{\sim}{y}, \underset{\sim}{x} \rangle - \bar{x} \langle \underset{\sim}{y}, \underset{\sim}{1} \rangle \right)$$

(1) $\langle \underset{\sim}{y}, \underset{\sim}{b}_1 \rangle \underset{\sim}{b}_1 = \mathrm{proj}_{\ell}\, \underset{\sim}{y} = \bar{y}\, \underset{\sim}{1}$.

(2) $$\langle \underset{\sim}{y}, \underset{\sim}{b}_2 \rangle = \left\langle \underset{\sim}{y}, \frac{1}{S_x}(\underset{\sim}{x} - \bar{x}\,\underset{\sim}{1}) \right\rangle$$

$$= \frac{1}{S_x}\left(\langle \underset{\sim}{y}, \underset{\sim}{x} \rangle - \bar{x} \langle \underset{\sim}{y}, \underset{\sim}{1} \rangle \right)$$

$$= \frac{1}{S_x}\left(\sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i \right)$$

$$= \frac{1}{S_x}\left(\sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i - \bar{y} \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})}_{=0} \right)$$

from before

$$= \frac{1}{S_x} \sum_{i=1}^{n} (x_i y_i - \bar{x} y_i - \bar{y} x_i + \bar{x}\bar{y})$$

(2) $\langle \underset{\sim}{y}, \underset{\sim}{b}_2 \rangle = \left\langle \underset{\sim}{y}, \frac{1}{S_x}(\underset{\sim}{x} - \bar{x}\,\underset{\sim}{1}) \right\rangle$

$$= \frac{1}{S_x}\left(\langle \underset{\sim}{y}, \underset{\sim}{x} \rangle - \bar{x} \langle \underset{\sim}{y}, \underset{\sim}{1} \rangle \right)$$

$$= \frac{1}{S_x}\left(\sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i \right)$$

$$= \frac{1}{S_x}\left(\sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i - \bar{y} \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})}_{= 0} \right)$$

from before

$$= \frac{1}{S_x} \sum_{i=1}^{n} (x_i y_i - \bar{x} y_i - \bar{y} x_i + \bar{x}\bar{y})$$

$$= \frac{1}{S_x} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

(2) $\langle \underset{\sim}{y}, \underset{\sim}{b}_2 \rangle = \langle \underset{\sim}{y}, \frac{1}{S_x}(\underset{\sim}{x} - \bar{x}\,\underset{\sim}{1}) \rangle$

$$= \frac{1}{S_x}\left(\langle \underset{\sim}{y}, \underset{\sim}{x} \rangle - \bar{x} \langle \underset{\sim}{y}, \underset{\sim}{1} \rangle \right)$$

$$= \frac{1}{S_x}\left(\sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i \right)$$

$$= \frac{1}{S_x}\left(\sum_{i=1}^{n} x_i y_i - \bar{x} \sum_{i=1}^{n} y_i - \bar{y} \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})}_{= 0 \text{ from before}} \right)$$

$$= \frac{1}{S_x} \sum_{i=1}^{n} (x_i y_i - \bar{x} y_i - \bar{y} x_i + \bar{x}\bar{y})$$

$$= \frac{1}{S_x} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Thus

$$\langle \underset{\sim}{y}, \underset{\sim}{b}_2 \rangle = \frac{S_{xy}}{S_x}$$

where

$$\boxed{S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}$$

$\frac{S_{xy}}{n}$ is known as the covariance of X and Y

(2) Thus

$$\langle \underset{\sim}{y}, \underset{\sim}{b}_2 \rangle = \frac{S_{xy}}{S_x}$$

where $\boxed{S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}$

and

$$\langle \underset{\sim}{y}, \underset{\sim}{b}_2 \rangle \underset{\sim}{b}_2 = \frac{S_{xy}}{S_x} \frac{1}{S_x} (\underset{\sim}{x} - \bar{x} \underset{\sim}{1})$$

$$= \frac{S_{xy}}{S_x^2} (\underset{\sim}{x} - \bar{x} \underset{\sim}{1}).$$

Hence

$$\text{proj}_{\mathcal{L}} \underset{\sim}{y} = \bar{y} \underset{\sim}{1} + \frac{S_{xy}}{S_x^2} (\underset{\sim}{x} - \bar{x} \underset{\sim}{1})$$

$$= \frac{S_{xy}}{S_x^2} \underset{\sim}{x} + \left(\bar{y} - \frac{S_{xy}}{S_x^2} \bar{x}\right) \underset{\sim}{1}$$

Hence

$$\mathrm{proj}_{\mathcal{L}}\, \underset{\sim}{y} = \bar{y}\, \underset{\sim}{1} + \frac{S_{xy}}{S_x^2}\left(\underset{\sim}{x} - \bar{x}\, \underset{\sim}{1}\right)$$

$$= \underbrace{\frac{S_{xy}}{S_x^2}}_{a}\, \underset{\sim}{x} + \underbrace{\left(\bar{y} - \frac{S_{xy}}{S_x^2}\bar{x}\right)}_{b} \underset{\sim}{1}$$

We obtain the regression line

$$Y = aX + b$$

where

$$a = \frac{S_{xy}}{S_x^2}, \quad b = \bar{y} - \frac{S_{xy}}{S_x^2}\bar{x} = \bar{y} - a\bar{x}$$

How well does the line fit?

$$\| \underset{\sim}{y} - \text{proj}_{\mathcal{L}} \underset{\sim}{y} \|^2 = \| \underset{\sim}{y} - (a\underset{\sim}{x} + b\underset{\sim}{1}) \|^2$$

$$= \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

$$= \sum_{i=1}^{n} (y_i - ax_i - \bar{y} + a\bar{x})^2$$

$$b = \bar{y} - a\bar{x}$$

$$= \sum_{i=1}^{n} \left((y_i - \bar{y}) - a(x_i - \bar{x}) \right)^2$$

$$= \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2a \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + a^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

How well does the line fit?

$$\| \underset{\sim}{y} - \text{proj}_{\mathcal{L}} \underset{\sim}{y} \|^2 = \| \underset{\sim}{y} - (a \underset{\sim}{x} + b \underset{\sim}{1}) \|^2$$

$$= \sum_{i=1}^{n} (y_i - a x_i - b)^2$$

$$= \sum_{i=1}^{n} (y_i - a x_i - \bar{y} + a\bar{x})^2$$

$$= \sum_{i=1}^{n} \left((y_i - \bar{y}) - a(x_i - \bar{x}) \right)^2$$

$$= \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2a \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + a^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$= S_y^2 - 2a S_{xy} + a^2 S_x^2$$

How well does the line fit?

$$\| \underset{\sim}{y} - \text{proj}_{\mathcal{L}} \underset{\sim}{y} \|^2 = \| \underset{\sim}{y} - (a \underset{\sim}{x} + b \underset{\sim}{1}) \|^2$$

$$= \sum_{i=1}^{n} (y_i - a x_i - b)^2$$

$$= \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2a \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + a^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$= S_y^2 - 2a S_{xy} + a^2 S_x^2$$

$$= S_y^2 - 2 \frac{S_{xy}}{S_x^2} S_{xy} + \left(\frac{S_{xy}}{S_x^2} \right)^2 S_x^2$$

$$= S_y^2 - 2 \frac{S_{xy}^2}{S_x^2} + \frac{S_{xy}^2}{S_x^2}$$

$$= S_y^2 - \frac{S_{xy}^2}{S_x^2}$$

How well does the line fit?

i.e. distance squared is

$$\| \underset{\sim}{y} - \mathrm{proj}_{\mathcal{L}}\, \underset{\sim}{y} \|^2 = s_y^2 - \frac{s_{xy}^2}{s_x^2}$$

Thus

$$\boxed{s_y^2 = (\text{distance})^2 + \frac{s_{xy}^2}{s_x^2}}$$

variation about $\bar{y}$

"error" of approximation

contribution to s_y^2 due to linear model

Thus

$$\left(\frac{S_{xy}^2}{S_x^2}\right) \Big/ S_y^2 = \frac{S_{xy}^2}{S_x^2 \, S_y^2}$$

is the proportion of the variation in y-values "explained" by using a linear model.

The closer this proportion is to 1, the better the line fits the data.

Put

$$\boxed{r = \frac{S_{xy}}{S_x S_y}}$$

called the correlation coefficient.

We have

$$s_y^2 = (\text{distance})^2 + \frac{s_{xy}^2}{s_x^2}$$

so

$$1 = \frac{(\text{distance})^2}{s_y^2} + \frac{s_{xy}^2}{s_x^2 s_y^2}$$

$$= \frac{(\text{distance})^2}{s_y^2} + r^2$$

so

$$\boxed{r^2 \leq 1 \quad \text{and} \quad -1 \leq r \leq 1}$$

We have

$$s_y^2 = (\text{distance})^2 + \frac{s_{xy}^2}{s_x^2}$$

so

$$1 = \frac{(\text{distance})^2}{s_y^2} + \frac{s_{xy}^2}{s_x^2 s_y^2}$$

$$= \frac{(\text{distance})^2}{s_y^2} + r^2$$

so $\boxed{r^2 \leq 1 \quad \text{and} \quad -1 \leq r \leq 1}$

If $|r| > 0.71$ then $r^2 > 0.5$, in which case we might regard the linear fit as reasonable, as it "explains" at least half of the variation.