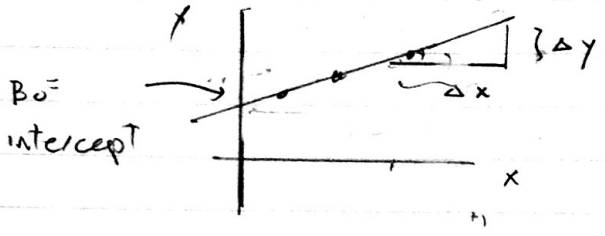
Simple Linear Regression

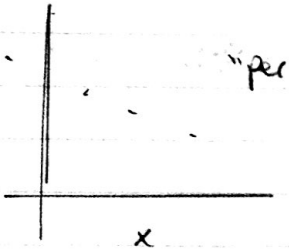
$|\text{cor}(\underset{\sim}{x}, \underset{\sim}{y})| = 1 \iff \underset{\sim}{y} = \beta_0 \underset{\sim}{1} + \beta_1 \underset{\sim}{x}$ some β_0, β_1

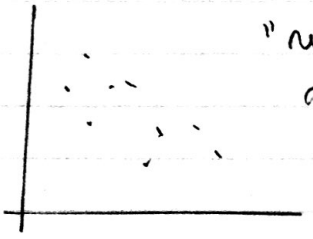


$$\frac{\Delta y}{\Delta x} = \frac{(\beta_0 + \beta_1(x+\Delta x)) - (\beta_0 + \beta_1 x)}{\Delta x}$$

$$= \beta_1 \Delta x / \Delta x = \beta_1 = \text{"slope"}$$

More typically, $|\text{cor}(\underset{\sim}{x}, \underset{\sim}{y})| < 1$





error, disturbance, deviation, ...

$\underset{\sim}{y} = \beta_0 \underset{\sim}{1} + \beta_1 \underset{\sim}{x}$

$y_i = \beta_0 + \beta_1 x_i$

$\underset{\sim}{y} = \beta_0 \underset{\sim}{1} + \beta_1 \underset{\sim}{x} + \underset{\sim}{e}$

$y_i = \beta_0 + \beta_1 x_i + e_i$

what values of β_0, β_1 provide the "best fit"?

OLS Regression line

$$RSS(\beta_0, \beta_1) = \sum (y_i - (\beta_0 + \beta_1 x_i))^2 = \| \underset{\sim}{y} - (\beta_0 \underset{\sim}{1} + \beta_1 \underset{\sim}{x}) \|^2$$

$$= \sum e_i^2 = \| \underset{\sim}{e} \|^2$$

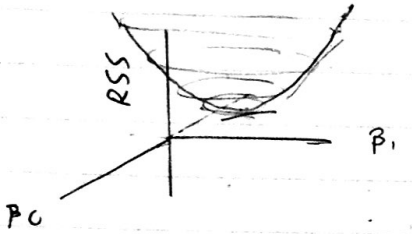
$e_i = y_i - (\beta_0 + \beta_1 x_i)$ = residual for the <u>particular</u> values of β_0, β_1

<u>Defn</u>: The OLS values of (β_0, β_1) are the values $(\hat{\beta}_0, \hat{\beta}_1)$ that minimize $RSS(\beta_0, \beta_1)$

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(\beta_0, \beta_1)}{\arg\min}\ RSS(\beta_0, \beta_1)$$

Optimization:

Method 1: geometry (will do calc.)
Method 2: calculus



* $RSS(\beta_0, \beta_1)$ is a quadratic (convex) function of (β_0, β_1)

* global minimized occurs where derivative (gradient) is zero.

$$\left.\frac{\partial RSE}{\partial \beta_0}\right| = -2 \sum (y_i - (\beta_0 + \beta_1 x_i)) = 0 \iff \sum \left(y_i - (\beta_0 + \beta_1 x_i)\right) = 0$$

$$\frac{\partial RSE}{\partial \beta_1} = -2 \sum x_i (y_i - (\beta_0 + \beta_1 x_i)) = 0 \iff \sum x_i (y_i - (\beta_0 + \beta_1 x_i)) = 0$$

The OLS values $(\hat{\beta}_0, \hat{\beta}_1)$ therefore satisfy...

① $\sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) = 0$

② $\sum x_i (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) = 0$

These are called the <u>normal equations</u> for simple linear regression.

Why "normal equations"? Let $\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ = "resid at OLS val." or just "residual"

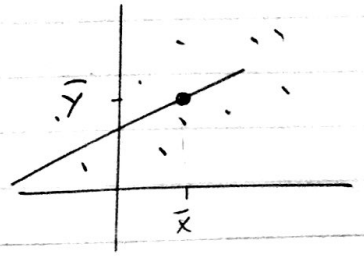
then ① $\Rightarrow$ $\sum \hat{e}_i = 0$ $\quad \hat{\underset{\sim}{e}} \cdot \underset{\sim}{1} = 0$

② $\Rightarrow$ $\sum x_i \hat{e}_i = 0$ $\quad \hat{\underset{\sim}{e}} \cdot \underset{\sim}{x} = 0$

$\Rightarrow \hat{\underset{\sim}{e}}$ is normal (orthogonal) to the vectors $\underset{\sim}{1}, \underset{\sim}{x}$.

Also, we see from ① that $\sum y_i = \sum (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$



regression line goes through the point $(\bar{x}, \bar{y})$

Solving the normal equations

① $\Sigma(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) = 0$

$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Plug this into ②: $\Sigma x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$

$\Sigma x_i (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) = 0$

$\Sigma x_i (y_i - \bar{y}) = \hat{\beta}_1 \Sigma x_i (x_i - \bar{x})$ ✱

Now note the following trick: $\Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i (y_i - \bar{y}) + \Sigma \bar{x}(y_i - \bar{y})$

$= \Sigma x_i (y_i - \bar{y}) + \bar{x}\,\Sigma(y_i - \bar{y})$

similarly, $\Sigma x_i (x_i - \bar{x}) = \Sigma (x_i - \bar{x})(x_i - \bar{x})$ $= \Sigma x_i (y_i - \bar{y})$

$= \Sigma (x_i - \bar{x})^2$

Let $SXX = \Sigma (x_i - \bar{x})^2$

$SXY = \Sigma (x_i - \bar{x})(y_i - \bar{y})$

Then ✱ says $SXY = \hat{\beta}_1 SXX$, $\hat{\beta}_1 = SXY / SXX$

Finally, the OLS values are: $\left[\begin{array}{l} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = SXY / SXX \end{array} \right]$

Relation to correlation

$\hat{\beta}_1 = \dfrac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}$ ← looks like covariance or correlation; ← looks like variance of x

$= \dfrac{SXY}{SXX} = \left(\dfrac{SYY}{SXX}\right)^{1/2} \dfrac{SXY}{SYY^{1/2}\, SXX^{1/2}}$

$= \left(\dfrac{S_Y}{S_X}\right) \times Cor(\underset{\sim}{x}, \underset{\sim}{y})$

What you need to find $\hat{\beta}_0, \hat{\beta}_1$: $\left[\begin{array}{l} \bar{x}, \bar{y} \\ s_x^2, s_y^2 \\ \text{Cor}(\underset{\sim}{x}, \underset{\sim}{y}) \end{array}\right.$

Regression Line

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{e}_i$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad = \text{"predicted" or "fitted" value, at } x.$$

$$= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x$$

$$= \bar{y} + \hat{\beta}_1 (x - \bar{x})$$

$$= \bar{y} + \frac{s_y}{s_x} \times \text{Cor}(\underset{\sim}{x}, \underset{\sim}{y}) \times (x - \bar{x})$$

Questions.

1) How does the line change if we change the location of the x's?

2) How does the line change if we change the scale of the x's?

3) What aspects of the regression line are invariant to a) location changes? b) scale changes? c) loc. + scale changes?