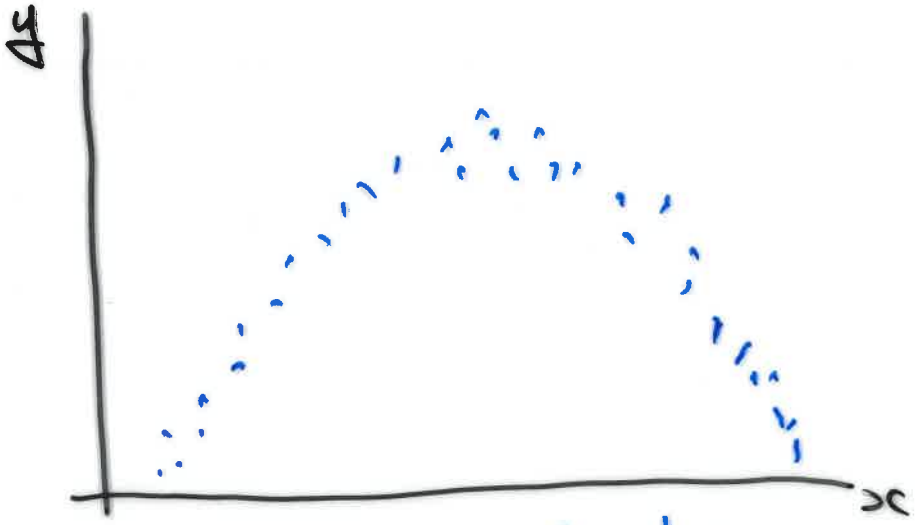
Non-linear data

Suppose we have points $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ whose plot looks like



we could try finding a quadratic equation

$$y = b_0 + b_1 x + b_2 x^2$$

which is a best fit, in the least squares sense, to the data.

To do this, we construct points

$(y_1, x_1, x_1^2), (y_2, x_2, x_2^2), \ldots \in \mathbb{R}^3$

Now find the hyperplane

$$y = b_0 + b_1 x + b_2 z$$

which is the least squares fit to the data.

This ensures that

$$y = b_0 + b_1 x + b_2 x^2$$

is the quadratic which best fits, in the least squares sense, the points in $\mathbb{R}^2$.

How good is a given least squares fit?

Data $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n) \in \mathbb{R}^2$

$p = 2$

Best fit $y = b_0 + b_1 x$ where

$$\left.\begin{aligned} \sum y_i &= n b_0 + b_1 \sum x_i \\ \sum x_i y_i &= b_0 \sum x_i + b_1 \sum x_i^2 \end{aligned}\right\} \text{normal equations}$$

Fitted value

$$\hat{y}_i = b_0 + b_1 x_i$$

Residual

$$e_i = y_i - \hat{y}_i$$

Sample mean

$$\bar{y} = \frac{1}{n} \sum y_i$$

$$SSTO = \sum (y_i - \bar{y})^2$$ total sum of squares

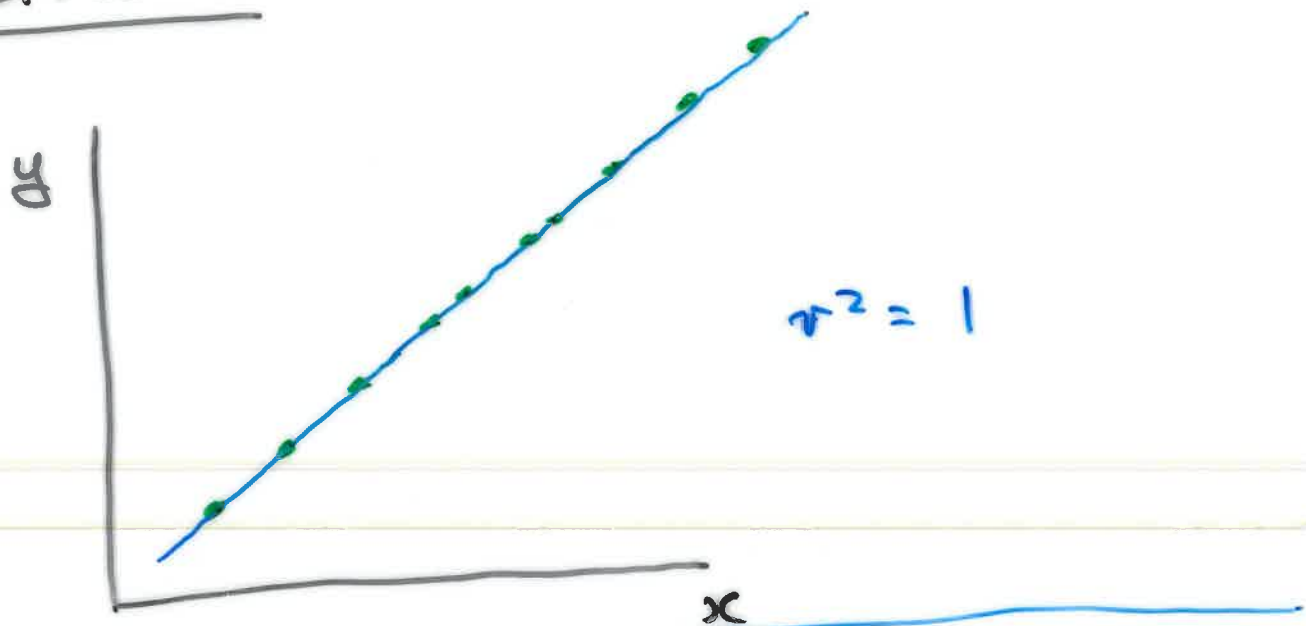
$$SSE = \sum (y_i - \hat{y}_i)^2$$ error sum of squares

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$ regression sum of squares

Defn coefficient of (simple) determination

$$r^2 := \frac{SSR}{SSTO}$$ = proportion of variation explained by regression

Illustration



Lemma i) $\sum e_i = 0$

ii) $\sum \hat{y}_i e_i = 0$

Proposition i) $SSTO = SSR + SSE$

ii) $0 \leq r^2 \leq 1$

Proof of Prop (i) ⇒ Prop (ii)

(i) implies $r^2 = \frac{SSR}{SSTO} = \frac{SSR}{SSR+SSE} = 1 - \frac{SSE}{SSTO}$.

But $0 \le SSE, SSTO, SSR$. By (i) $SSE \le SSTO$.

Thus $0 \le r^2 \le 1$.

Proof of Prop (i)

$$\sum (y_i - \bar{y})^2 = \sum \left\{ (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \right\}^2$$

$$= \sum (\hat{y}_i - \bar{y})^2 + (y_i - \hat{y}_i) + 2 \underbrace{\sum (\hat{y}_i - \bar{y})(y_i - \hat{y}_i)}_{A}$$

$$A = \sum \hat{y}_i (y_i - \hat{y}_i) - \bar{y}(y_i - \hat{y}_i)$$

$$= \sum \hat{y}_i e_i - \bar{y} \sum e_i$$

$$= 0 \quad \text{by Lemma}.$$

Hence $SSTO = SSR + SSE$.

∎

Proof of Lemma (i)

$$\sum e_i = \sum (y_i - b_0 - b_1 x_i)$$

$$= \sum y_i - nb_0 - b_1 \sum x_i$$

$$= 0 \text{ by first normal equation.}$$

Proof of Lemma (ii) use both normal equations.

Matrix Notation ($p \geq 2$)

$$B = (X^t X)^{-1} X^t Y$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1\,p-1} \\ 1 & x_{21} & & x_{2\,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & & x_{n\,p-1} \end{pmatrix}$$

$$SSTO = Y^t Y - n\bar{y}^2$$

$$SSR = B^T X^t Y - n\bar{y}^2$$

$$SSE = Y^t Y - B^t X^t Y$$

$R^2 = \frac{SSR}{SSTO}$. Again $0 \leq R^2 \leq 1$.