Equation based learning

Regression Model

$Y =$ response or outcome variable
$X1, X2, X3, \ldots, Xp$ = explanatory or input variables

The general relationship approximated by:

$$Y = f(X_1, X_2, \ldots, X_p) + e$$

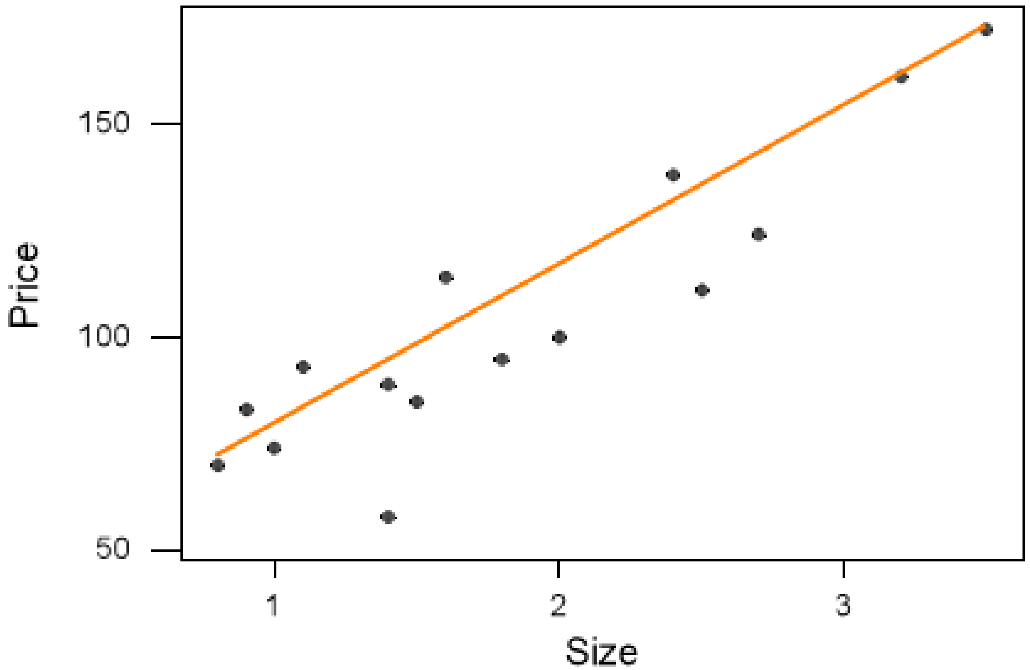
And a linear relationship is written

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p + e$$

Linear Prediction

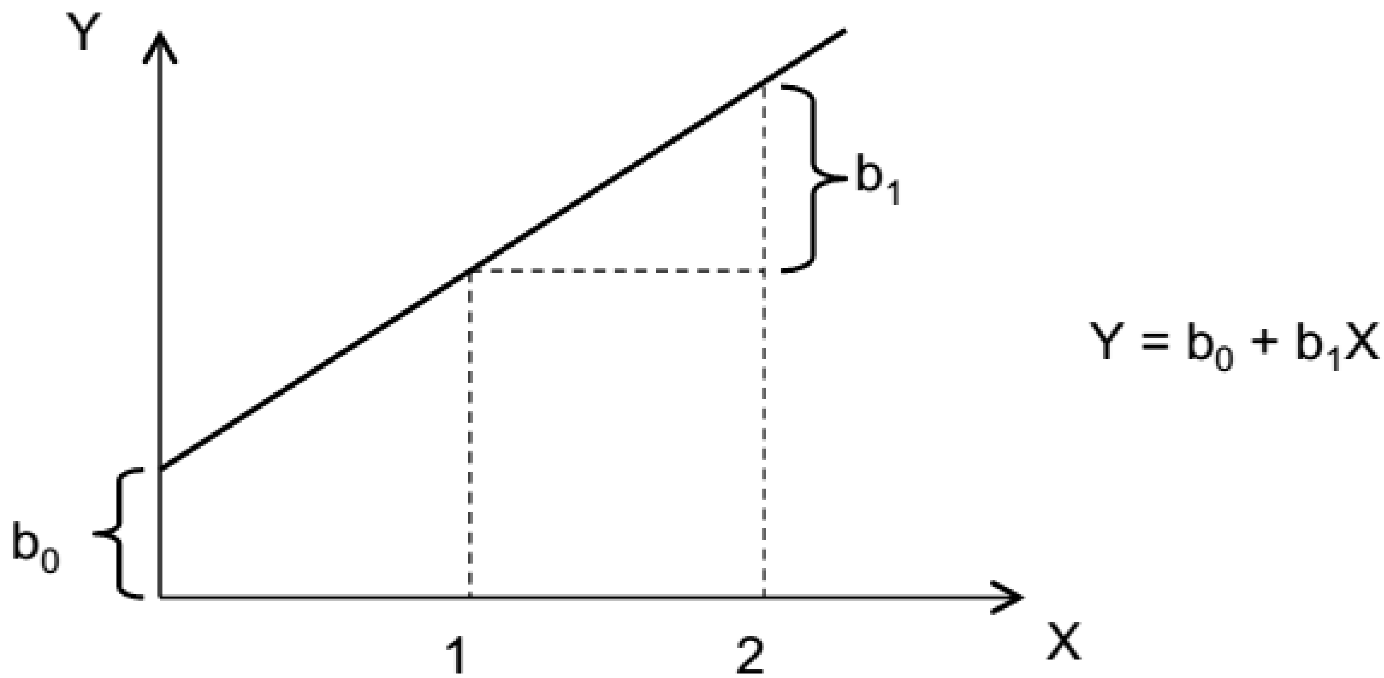
Appears to be a linear relationship between price and size:

As size goes up, price goes up.



The line shown was fit by the "eyeball" method.

Linear Prediction



Our "eyeball" line has $b_0 = 35$, $b_1 = 40$.

Linear Prediction

Can we do better than the eyeball method?

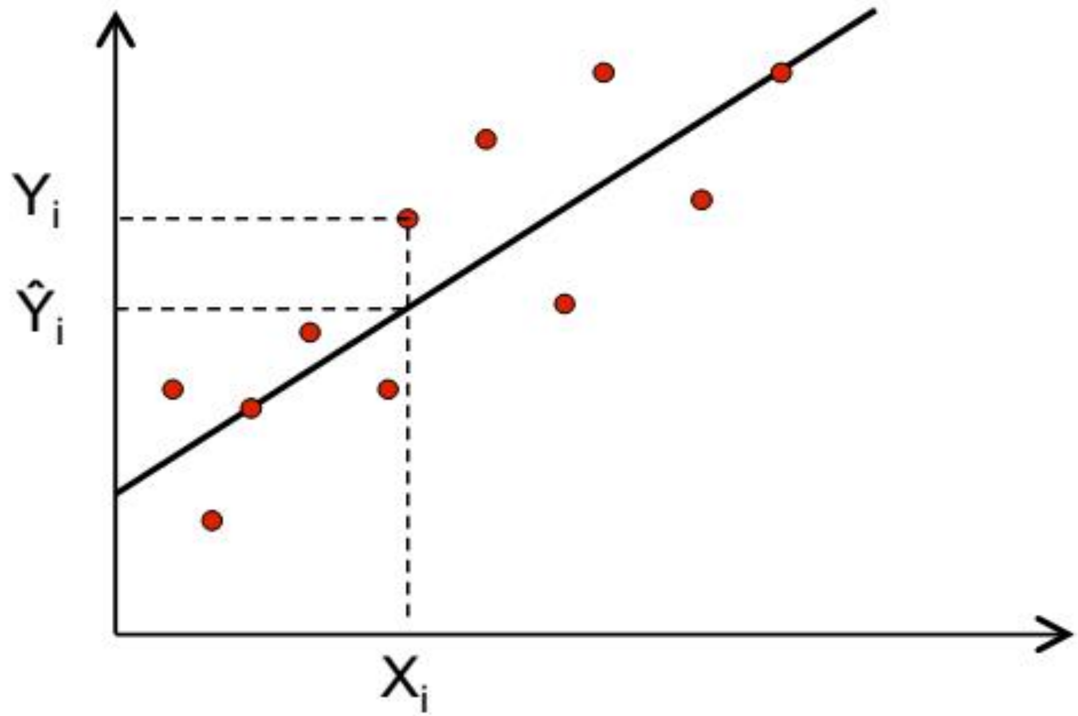
We desire a strategy for estimating the slope and intercept parameters in the model $\hat{Y} = b_0 + b_1 X$

A reasonable way to fit a line is to minimize the amount by which the fitted value differs from the actual value.

This amount is called the residual.

Linear Prediction

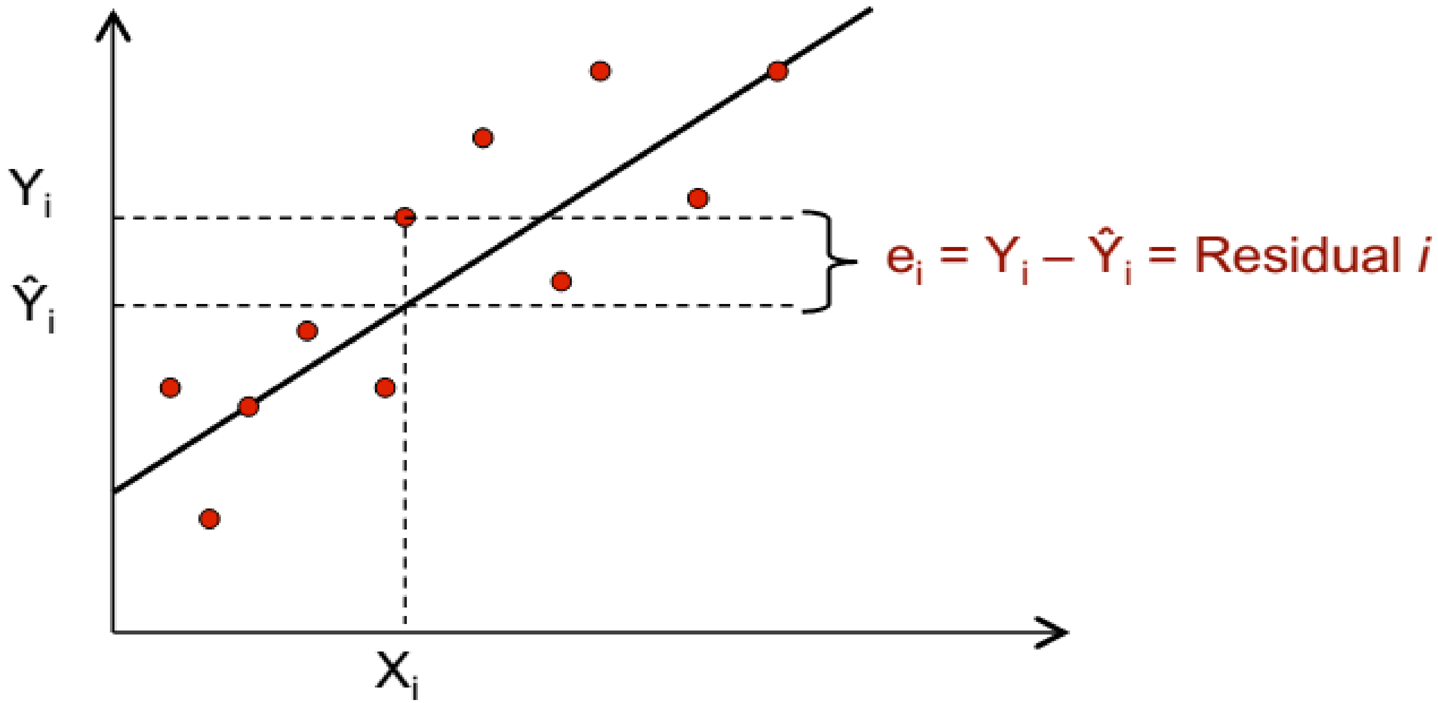
What is the "fitted value"?



The dots are the observed values and the line represents our fitted values given by $\hat{Y}_i = b_0 + b_1 X_1$.

Linear Prediction

What is the "residual"' for the ith observation'?



We can write $Y_i = \hat{Y}_i + (Y_i - \hat{Y}_i) = \hat{Y}_i + e_i$.

Least Squares

Ideally we want to minimize the size of all residuals:

- If they were all zero we would have a perfect line.
- Trade-off between moving closer to some points and at the same time moving away from other points.
- Minimize the "total" of residuals to get best fit.

Least Squares chooses b_0 and b_1 to minimize $\sum_{i=1}^{N} e_i^2$

$$\sum_{i=1}^{N} e_i^2 = e_1^2 + e_2^2 + \cdots + e_N^2 = (Y_1 - \hat{Y}_1)^2 + (Y_2 - \hat{Y}_2)^2 + \cdots + (Y_N - \hat{Y}_N)^2$$

$$= \sum_{i=1}^{n} (Y_i - [b_0 + b_1 X_i])^2$$

Objective function for Parameter Estimation

$$\text{Error}_{(\text{m,b})} = \frac{1}{N}\sum_{i=1}^{N}(y_i - (mx_i + b))^2$$