Regression Overview

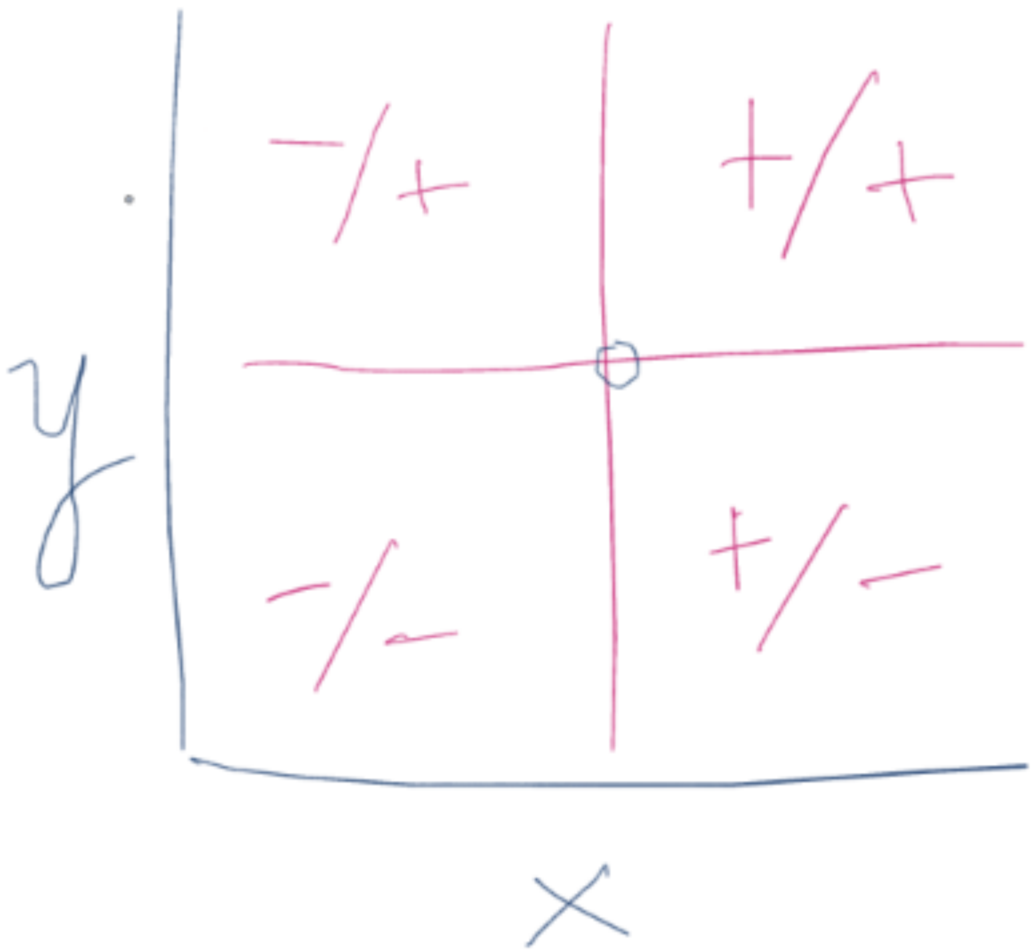
- **<u>Review:</u>** <u>Correlation:</u>

  - Measures the amount/degree of linear association between two **numerical** variables

  - Estimate the degree to which variables **covary**
    - With no attempt to interpret the causality of the association

  - <u>example:</u> arm length and leg length covary together (individuals with longer arms often have longer legs) but they are influenced by other underlying variables **not** each other (longer legs do not cause longer arms)

Correlation etc.



Regression:

- Statistics is about prediction
- Used to **predict** value of one numerical variable from the value of another
  - predicting dependent/response variable, Y from independent/ predictor X

- Linear regression assumes that the relationship between X and Y can be described by a line
  - Fits a straight line to a (messy) scatterplot

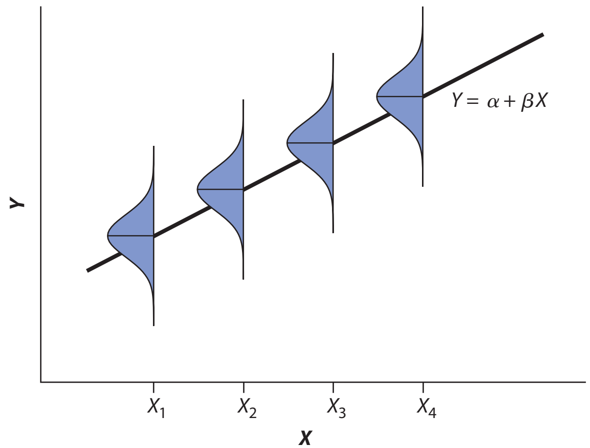
- Example: ambient temperature may effect growth rate of a plant species but the reverse is probably not true

Regression:

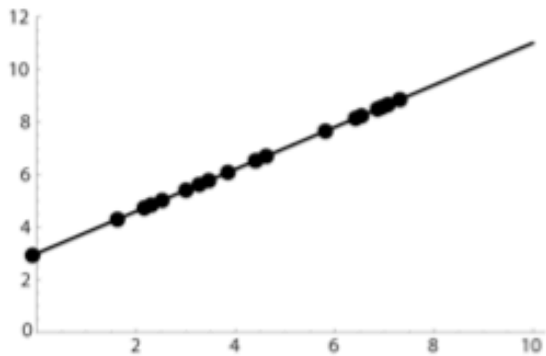
– Linear regression assumes that the relationship between X and Y can be described by a line
    - Fits a straight line to a (messy) scatterplot

– Homoscedasticity: Y is normally distributed with equal variance for all values of X

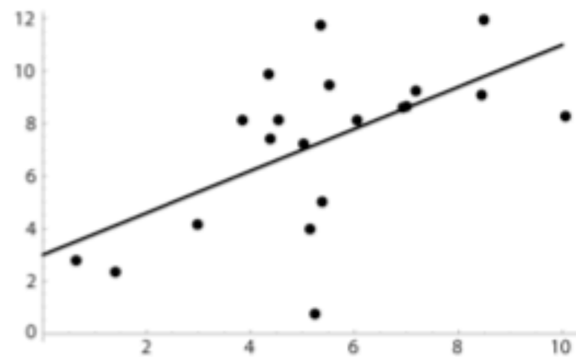
– Example: ambient temperature may effect growth rate of a plant specie but the reverse is probably not true



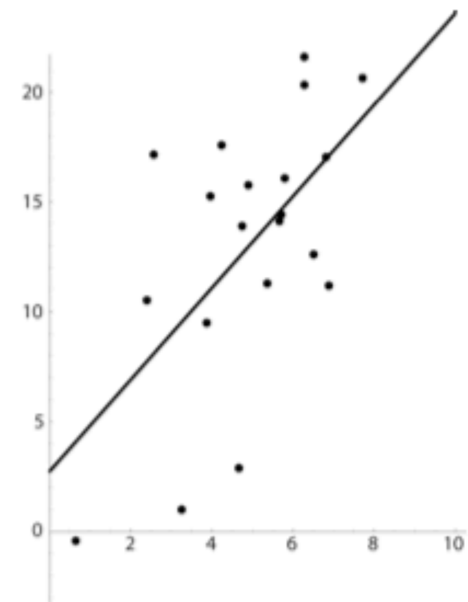
correlation vs regression



$r = 1$; $Y = 3 + 0.8X$



$r = 0.6$; $Y = 3 + 0.8X$



$r = 0.6$; $Y = 3 + 2X$

Different correlation;
same slope

Same correlation;
different slope

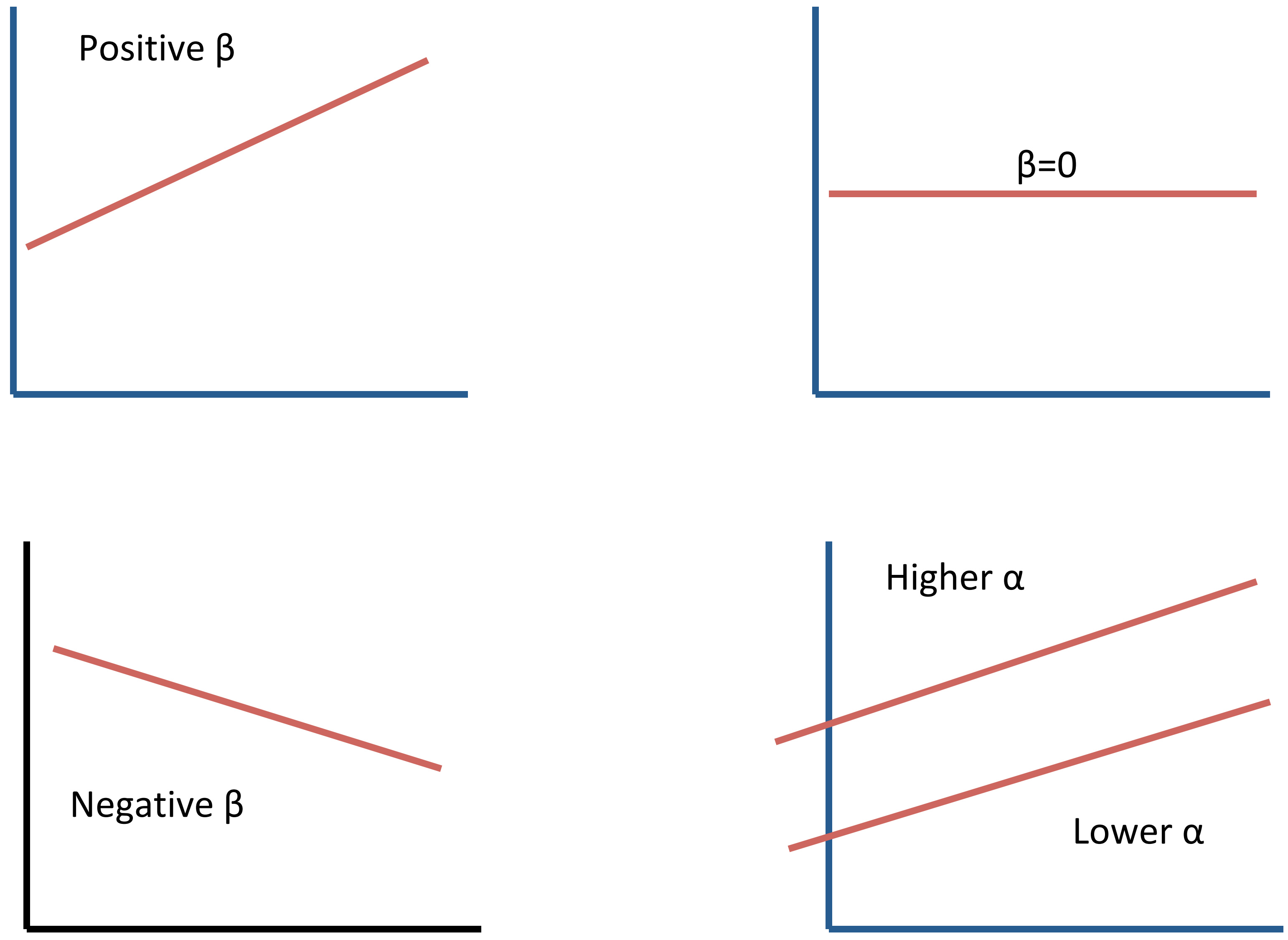
The parameters of linear regression

$$Y = \boxed{\alpha} + \beta X$$

intercept

slope

Regression Overview



Estimating a regression line

$$Y = a + bX$$

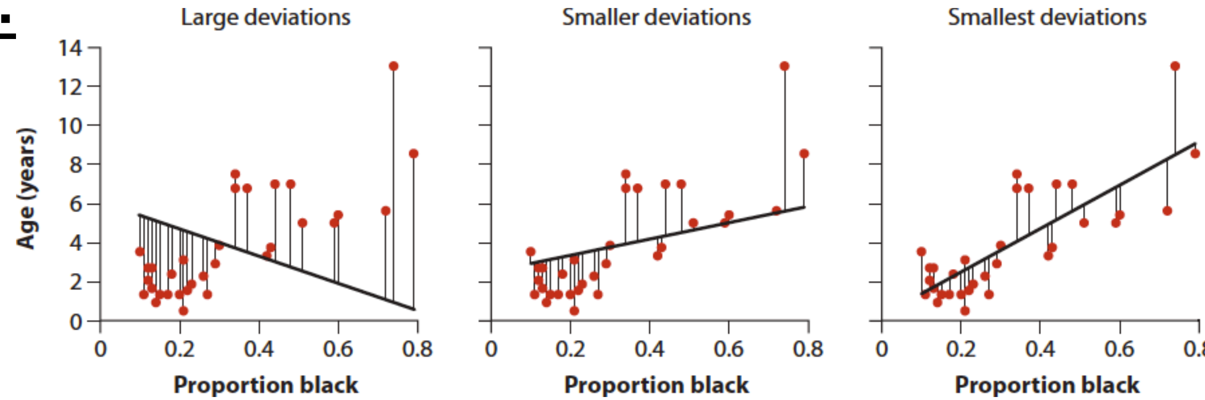
Least Squares:

- Best fitting line through a scatterplot
  - Line that minimized spread of y values

- Minimize $SS_{residuals}$ $\hat{y}_i$
  - Measurement of how much the line's predicted $\hat{y}_i$ deviate from actual data values

$$SS_{residual} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Least Squares:



- Best fitting line through a scatterplot
  - Line that minimized spread of y values
- Minimize $SS_{residuals}$
  - Measurement of how much the line's predicted y_i deviate from actual data values

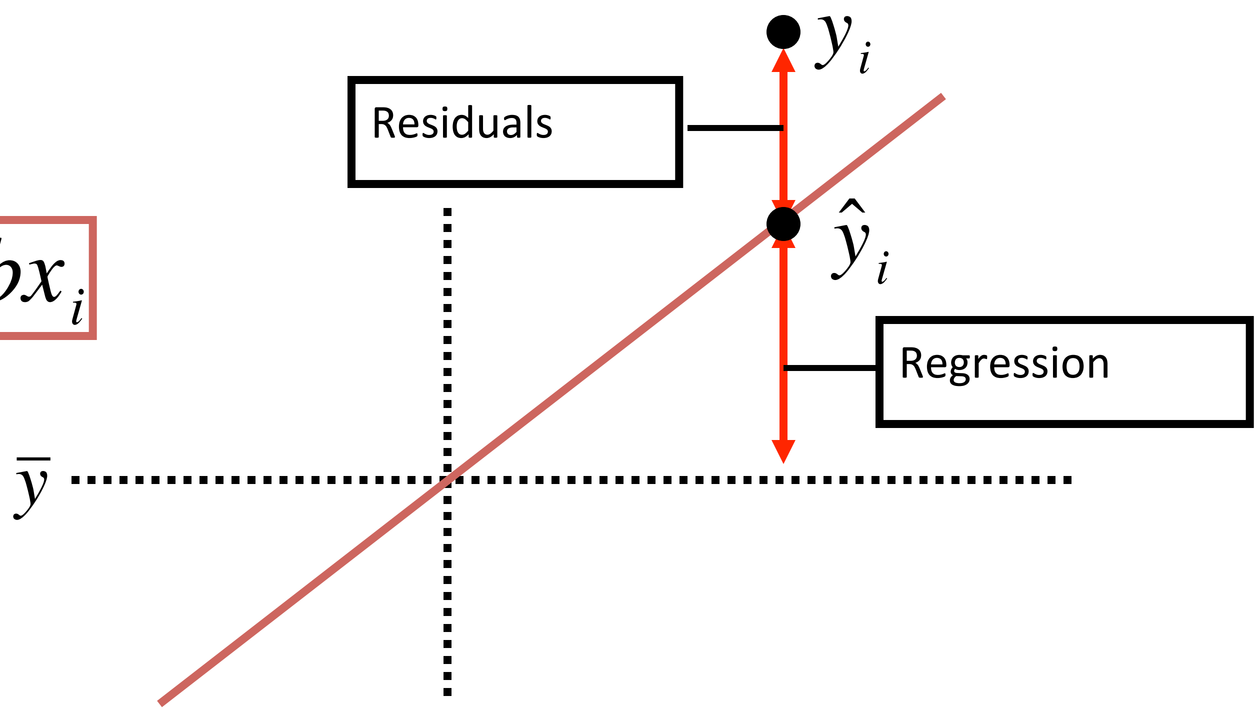
$$\boxed{SS_{residual} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Least Squares:

- What are the elements of this equation?

$$SS_{residual} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\boxed{\hat{y}_i = a + bx_i}$$



- <u>Residuals</u>:
  - Residuals measure the scatter of points above and below the least squares regression line

$$residual = Y_i - \hat{Y}_i$$

- $MS_{residual}$ is the variance of the residuals

$$MS_{residual} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}$$

Best estimate of slope:

$$b = \frac{\text{Sum of cross products}}{\text{Sum of squares of X}}$$

$$\boxed{b = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

Finding **a**:

$$\boxed{\bar{Y} = a + b\bar{X}}$$

OR

$$\boxed{a = \bar{Y} - b\bar{X}}$$