Transformations

Transformations

- One of the most important assumptions when fitting a linear regression model is that there exists a linear relationship between the independent and the dependent variables

Transformations

- One of the most important assumptions when fitting a linear regression model is that there exists a linear relationship between the independent and the dependent variables
- Often times the linear relationship is not clear

Transformations

- One of the most important assumptions when fitting a linear regression model is that there exists a linear relationship between the independent and the dependent variables
- Often times the linear relationship is not clear
- In that case, different transformations can be applied to the data to make the linear relationship clearer

Transformations

- $\hat{y}_i = b_0 + b_1 x_i$ - a unit increase in x is associated with an average of b_1 units increase in y
-
- $\log(\hat{y}_i) = b_0 + b_1 x_i$ - a unit increase in x is associated with an average of b_1 units increase in log(y)
- $\log(\hat{y}_i) = b_0 + b_1 \log(x_i)$- a k-fold increase in x is associated with k^b multiplicative increase in y
    - If x doubles, y changes by a multiplicative factor of 2^b

Transformations (Standardization)

- When modeling a multiple linear regression model, we might have to deal with independent variables that correspond to different units

Transformations (Standardization)

- When modeling a multiple linear regression model, we might have to deal with independent variables that correspond to different units
    - Age column on a scale of 20-70 vs. Salary column on a scale of 30,000-70,000

Transformations (Standardization)

- When modeling a multiple linear regression model, we might have to deal with independent variables that correspond to different units
    - Age column on a scale of 20-70 vs. Salary column on a scale of 30,000-70,000
- When we want to build a model where variables have different units, so a large value in one scale does not overpower a relatively smaller value in a different scale, we standardize our data

Transformations (Standardization)

- When modeling a multiple linear regression model, we might have to deal with independent variables that correspond to different units
    - Age column on a scale of 20-70 vs. Salary column on a scale of 30,000-70,000
- When we want to build a model where variables have different units, so a large value in one scale does not overpower a relatively smaller value in a different scale, we standardize our data
- Performed by converting data into Z-scores
    - mean = 0, sd = 1

Transformations (Standardization)

- $Z \sim N(0,1)$
- $z = (x - \text{mean}) / sd$
-
- Done separately for each attribute