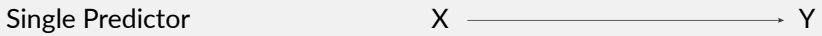
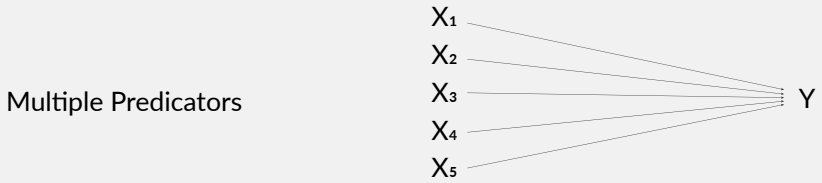
Linear Regression

- Regression is a statistical technique that relates a dependent variable to one or more independent (explanatory) variables.
- Many processes follow linearity, i.e., independent and dependent variables follow linear relations. These can be approximated by linear models, which are built using linear regression.
- **Simple linear regression** has a single feature (one independent variable) to model a linear relationship with a target (one dependent variable) by fitting the best straight line to describe the relationship.
- If there is more than one feature, then it becomes **multiple linear regression**.

Linear Regression



Multiple Linear Regression



Common Interview Questions:

1. What is regression, and when should it be used?
2. What are the assumptions associated with the linear regression model?
3. Why should the residuals be normally distributed?
4. How will you improve the accuracy of the linear model?
5. How will you check the performance of the linear regression model?
6. When would you prefer multiple linear regression to simple linear regression?
7. Why are residues important for linear regression models?
8. Give examples of problems where linear regression can be used.
9. Suppose the accuracy of your linear regression model is 60%. What steps will you take next?

Linearity

The relationship between independent and dependent variable is linear

Normality

Model residuals should follow a normal distribution.

Linear Regression Model: Assumptions

If these are not true, the model might generate incorrect results

Independence

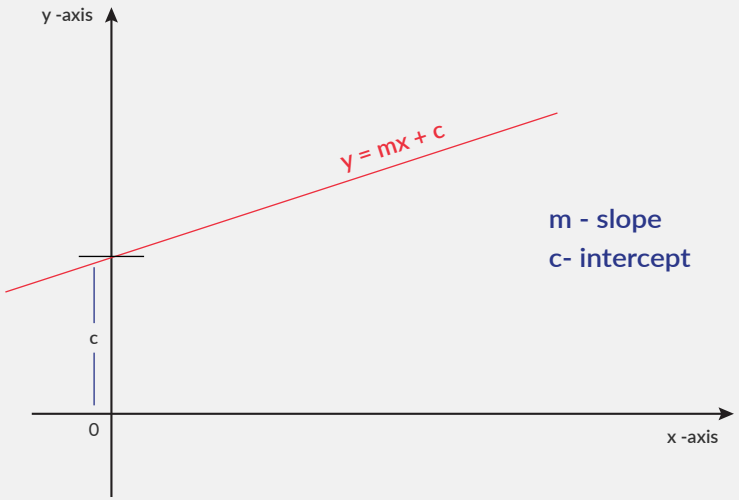
Each independent variable should be independent of other variables.

Homoscedasticity

The variance of the residual should be the same for any value of x.

Linear Regression

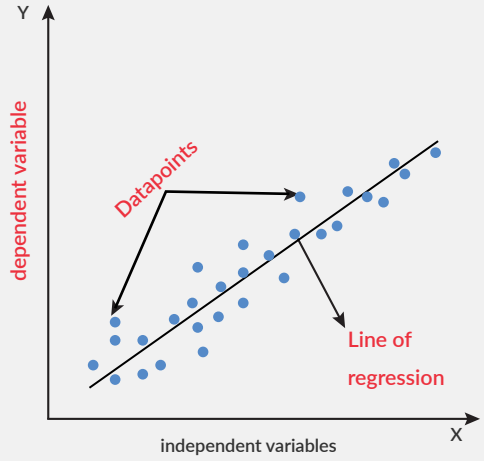
About Simple Linear Regression



Dependent variable (y): The variable that is being estimated and predicted, also known as a target

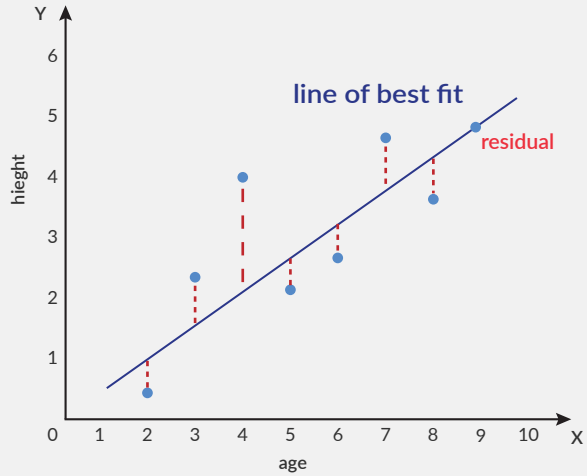
Independent variable (x): The input variable, also known as a predictor or feature

The linear regression model creates many lines and then selects the best-fit line as the model.

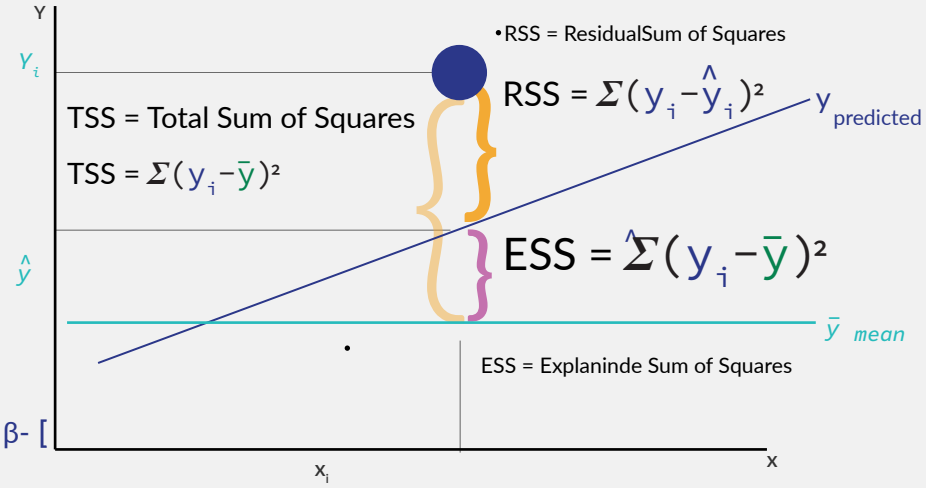


Residuals

Residual = Actual y value − Predicted y value



The residue tells us how much our model deviates from the actual values. We can study this deviation in the form of RSS, TSS and ESS.



Note: TSS = ESS + RSS

RSS is the amount of variation that can be removed using the model.

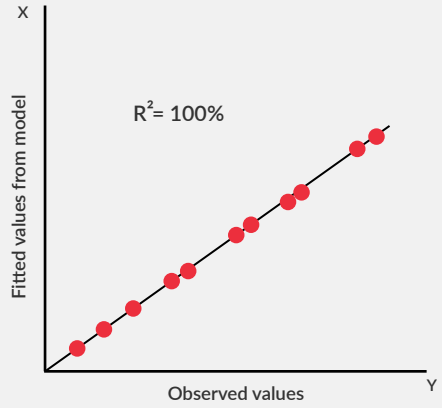
Linear Regression

Measures how much of the variability in the dependent variable can be explained by the model

R Squared (Coefficient of Determination/Goodness of Fit)

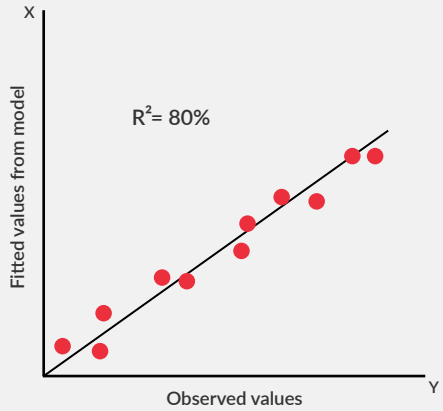
- Measures how much $R^2 = 1 - \frac{RSS}{TSS}$ the dependent variable can be explained by the model.
- Square of the correlation coefficient(R) and, therefore, referred to as R squared
- Can take a value between 0 and 1, where values closer to 0 represent a poor fit, and values closer to 1 represent an (almost) perfect fit

Great



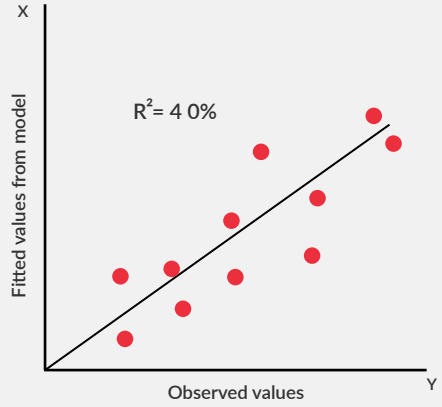
Fitted + observerd: Model explains all variance

Good



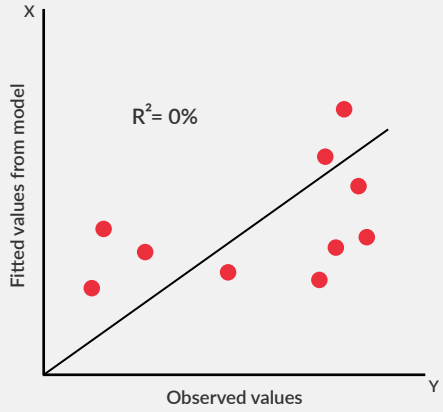
Model Explains bulk of variance

ok



Model explains 40% of variance, so is reasonable.

Inconsistent



Model fails to explain any variance