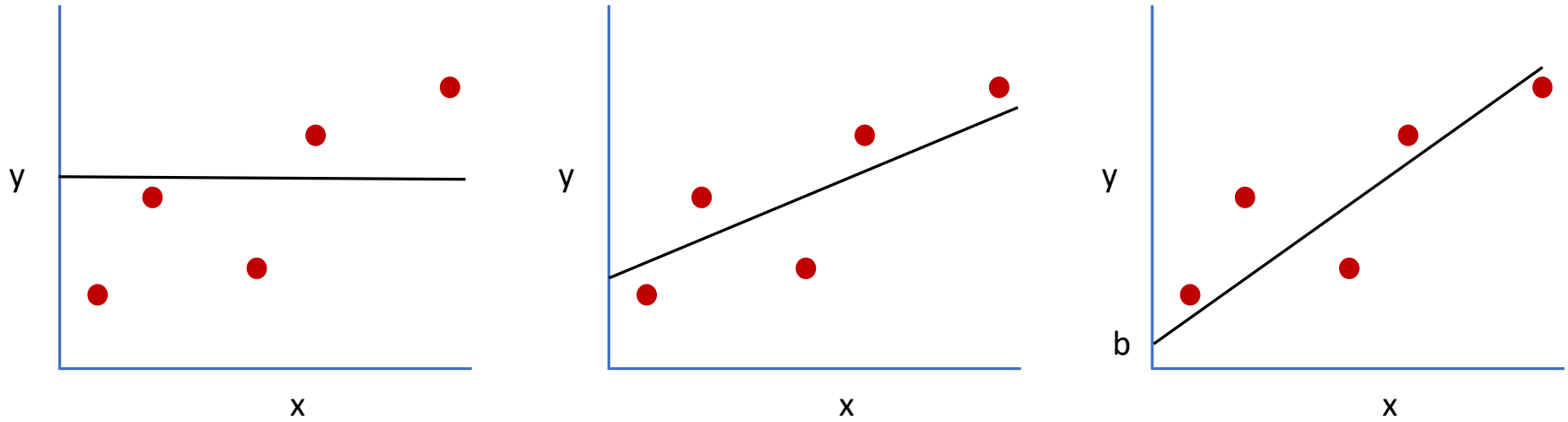
Linear
Regression

Linear Regression aka Least Squares

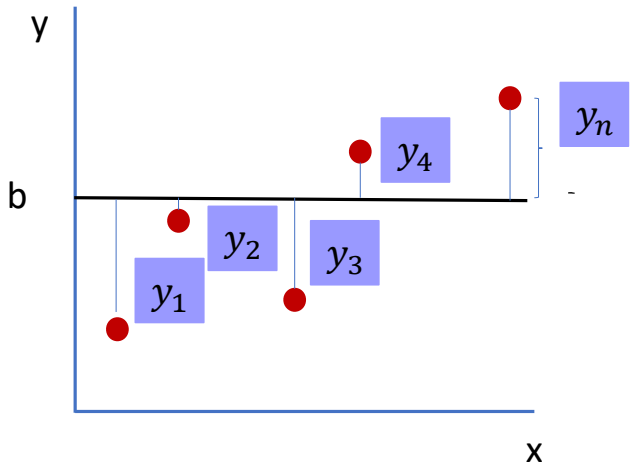
- Francis Galton's work

Line that Best Fits the Data



Sum of Squares

- The least-squares method is a statistical procedure to find the best fit for a set of data points by minimizing the sum of the offsets or residuals of points from the plotted curve.



$$SS1 = (b1 - y_1)^2 + (b1 - y_2)^2 + (b1 - y_3)^2 \dots (b1 - y_n)^2$$

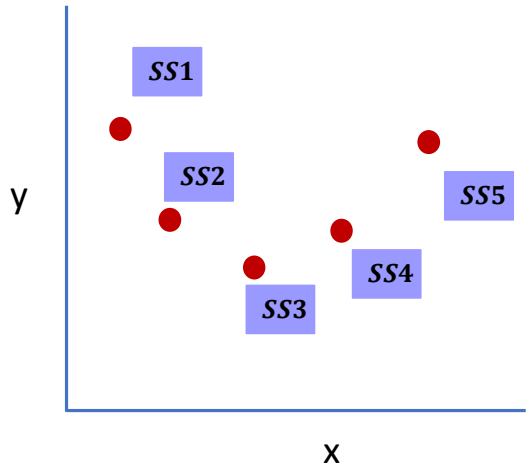
$$SS2 = (b2 - y_1)^2 + (b2 - y_2)^2 + (b2 - y_3)^2 \dots (b2 - y_n)^2$$

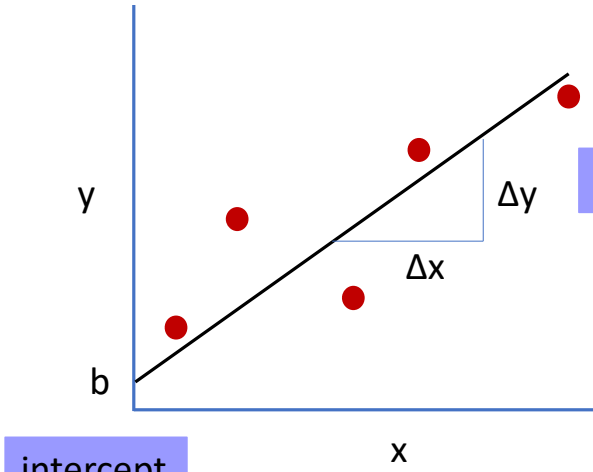
$$SS3 = (b3 - y_1)^2 + (b3 - y_2)^2 + (b3 - y_3)^2 \dots (b3 - y_n)^2$$

$$SS4 = (b4 - y_1)^2 + (b4 - y_2)^2 + (b4 - y_3)^2 \dots (b4 - y_n)^2$$

Also called as Residual Sum of Squares (RSS)

Least Sum of Squares





m = Δy/Δx

SS3 is the sweet spot

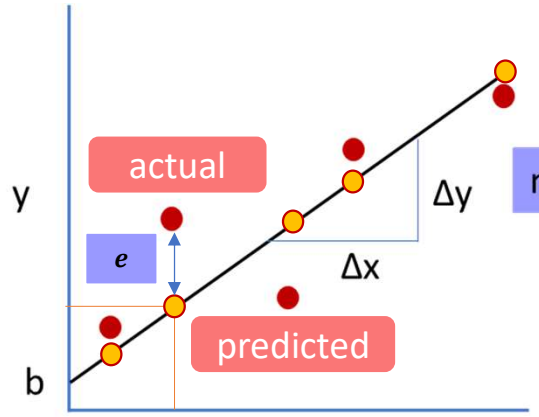
The line that produced ss3 is the line that best fits. The resulting line is as close as possible to all the data points.

Such a line has a slope **m** and intercept **b**
If the dependent variable is **y** and
independent variable is **x** then, using the straight line equation:

y = mx + b

Approximate linear regression equation

Error Term



error = actual – predicted
error = y – (mx + b)
y = mx + b + error

y = mx + b + e

Linear Regression Equation with Error

The objective of finding the best fit line is to minimize errors

Predicting Continuous Variable

$y = mx + b$

Approximate linear regression equation

- Using the above equation, if there is a new **x** we could easily get the value of **y**
- Note that in practice, **m** and **b** are unknown. Before we make predictions, we must use data to estimate the coefficients

Predicting Continuous Variables

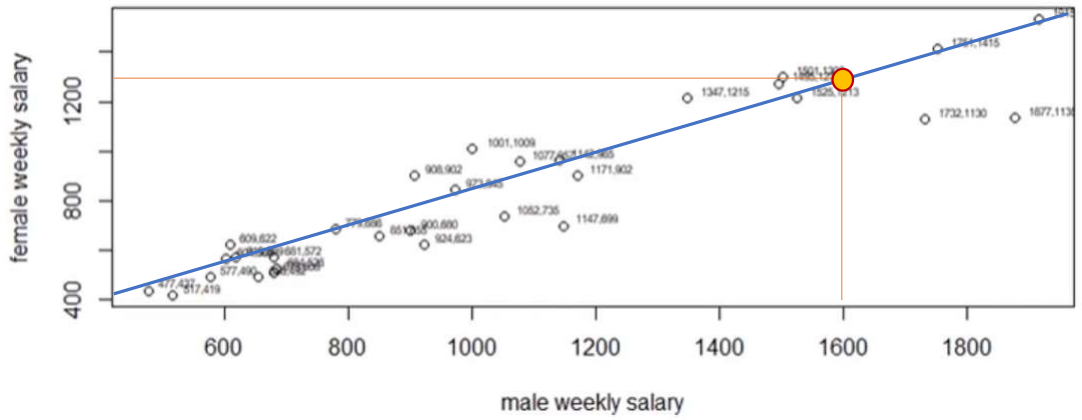
	Occupation	M_weekly	F_weekly
1	Social and community service managers	1142	965
2	Food service managers	820	680
3	Computer programmers	1501	1302
4	Food processing workers, all other	679	508
5	SOCIAL SERVICE	973	845
6	Health practitioner support technologists and technicians	652	633
7	Sales representatives, services, all other	1147	699
8	Postsecondary teachers	1405	1144
9	Pharmacists	2117	1811
10	Bailiffs, correctional officers, and jailers	779	686
11	Receptionists and information clerks	619	569
12	Chefs and head cooks	656	492
13	Nursing, psychiatric, and home health aides	526	457

actual

predicted

789.89

	M_weekly	F_weekly
Cement Industry	US$1600	?



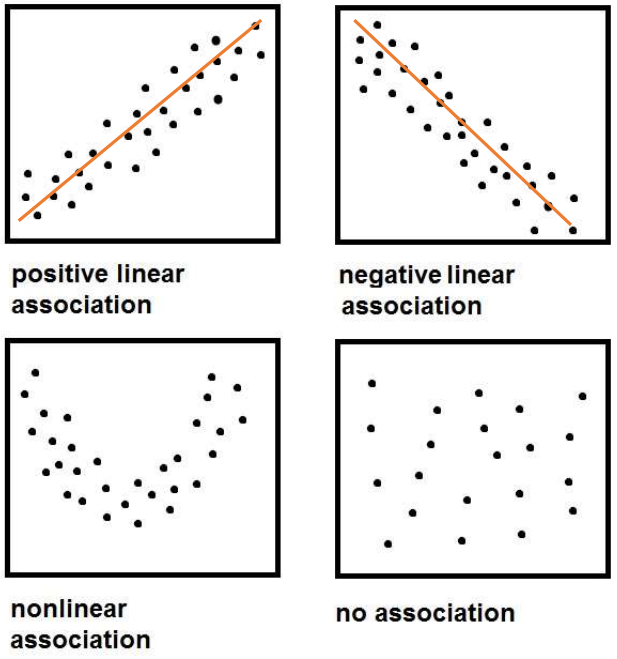
Assumptions/Requirements for Linear Model

- Simple linear regression cannot be used when the variables are non-quantitative in nature:
  - Binary
  - Categorical (more than two values)
- Linear relation between predictor and target — Scatter plot
- Very low or no multi-collinearity — Scatter plot
  - X terms should not be collinear
- No Heteroscedasticity – error should not be related — Regular output
- No auto-correlation between errors
- Normal distribution of errors/residuals — QQ plot
- All observations should be independent of each other

Linear Relationship Between Y and X

- The relationship between X variables (independent variables) and Y variable shall be linear
- If you fit a linear model to a non-linear, non-additive data set, the regression algorithm would fail to capture the trend mathematically, thus resulting in an inefficient model.
- Also, this will result in erroneous predictions on an unseen data set.
- How to check: Residuals vs Fitted Scatter Plot

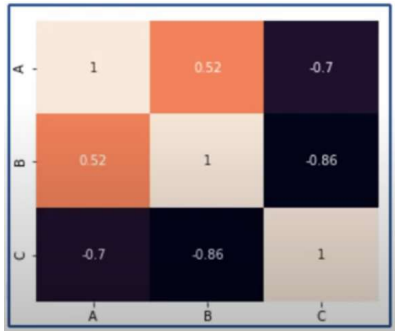
PS: Sometimes by applying log, exponential, polynomial terms or any other function we could convert non-linear to linear



No Multi-collinearity

- Multicollinearity is a condition when two independent variables are strongly correlated. The means, if one variable changes, the other variable is strongly affected. Example: **age** vs **years_of_experience** to target **salary**
- Using correlation graph or heat map, we can check correlation between variables
- Multicollinearity should be low or not present at all for linear regression to work.
- It becomes difficult to find out which variable is actually contributing to predict the response variable.

PS: You can remove one of the variable to fix it. In the graph you could remove either B or C as they show high collinearity



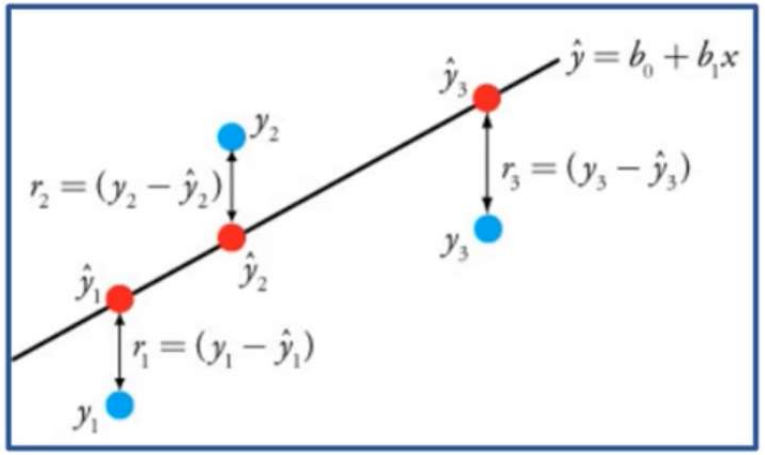
No Multi-collinearity

- **How to check:** You can use scatter plot to visualize correlation effect among variables. Also, you can also use VIF factor. VIF value <= 4 suggests no multicollinearity whereas a value of >= 10 implies serious multicollinearity. Above all, a correlation table should also solve the purpose.

VIF = Variance Inflation Factor

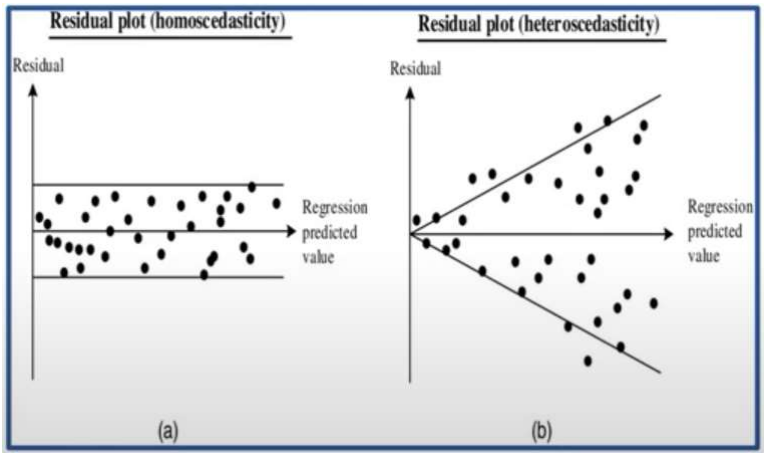
Mean of Residuals is Zero

- Residual is the difference between the observed value and estimated value.
- The mean of residuals is always zero.



HOMOSCEDASTICITY

- **Homo** means consistent and **Scedasticity** means variance
- If the variance in the residual error is consistent regardless of the dependent variable x, then it is homoscedastic
- Dependent variable have same variance across range of values of independent variable.

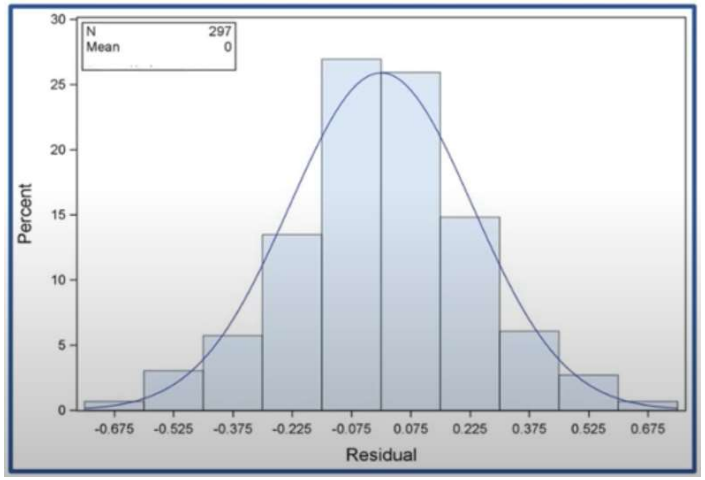


HOMOSCEDASTICITY

- **How to check:** You can look at residual vs fitted values plot. If heteroskedasticity exists, the plot would exhibit a funnel shape pattern (shown in next section).
- Also, you can use Breusch-Pagan / Cook – Weisberg test or White general test to detect this phenomenon.

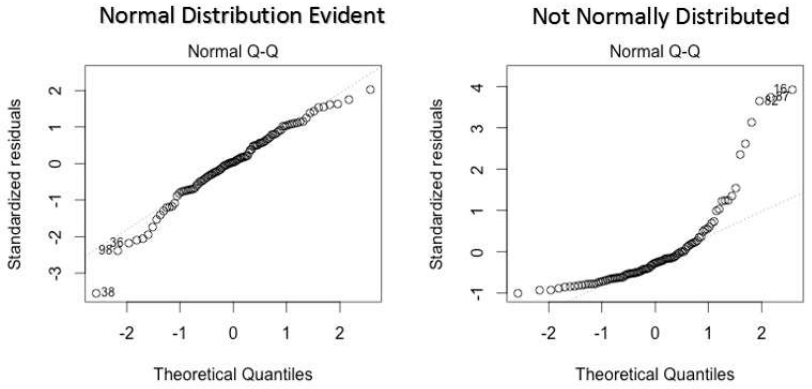
Residuals are Normally Distributed

- Normal distribution is a probability distribution that is symmetric about the mean
- Data near the mean are more frequent in occurrence than data far from the mean
- If the error terms are non- normally distributed, confidence intervals may become too wide or narrow. Once confidence interval becomes unstable, it leads to difficulty in estimating coefficients based on minimization of least squares.



Residuals are Normally Distributed

- **How to check:** You can look at QQ plot (shown below). You can also perform statistical tests of normality such as Kolmogorov-Smirnov test, Shapiro-Wilk test.



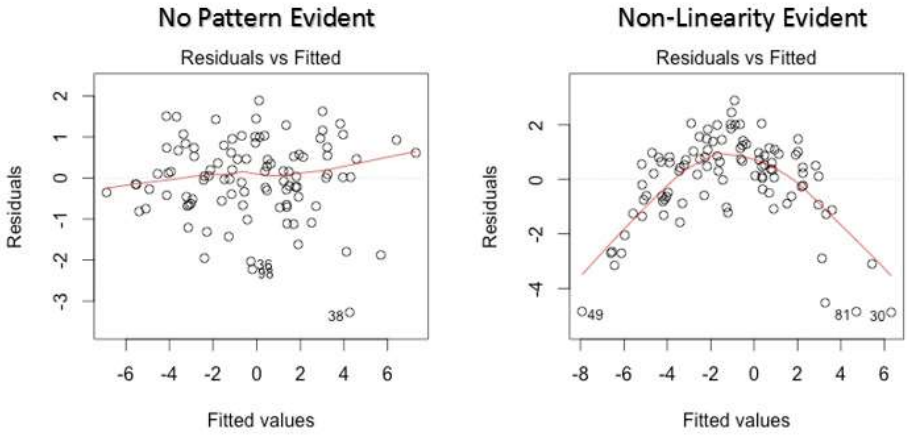
Uncorrelated X Variables and Residuals

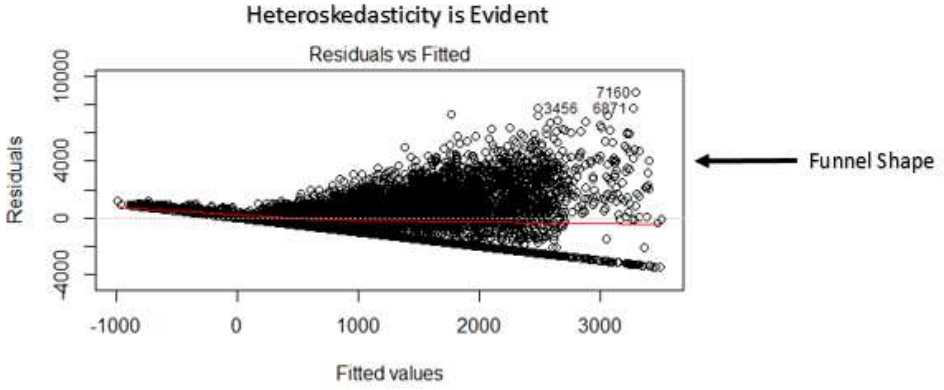
- The presence of correlation in error terms drastically reduces model's accuracy. This usually occurs in time series models where the next instant is dependent on previous instant.
- If the error terms are correlated, the estimated standard errors tend to underestimate the true standard error. If this happens, it causes confidence intervals and prediction intervals to be narrower.
- Also, lower standard errors would cause the associated p-values to be lower than actual. This will make us incorrectly conclude a parameter to be statistically significant.
- The correlation between X (independent variables) and residuals should be low or none.
- No auto-correlation (any relationship) between residuals and other independent variables, residuals should be as random as possible

Uncorrelated X Variables and Residuals

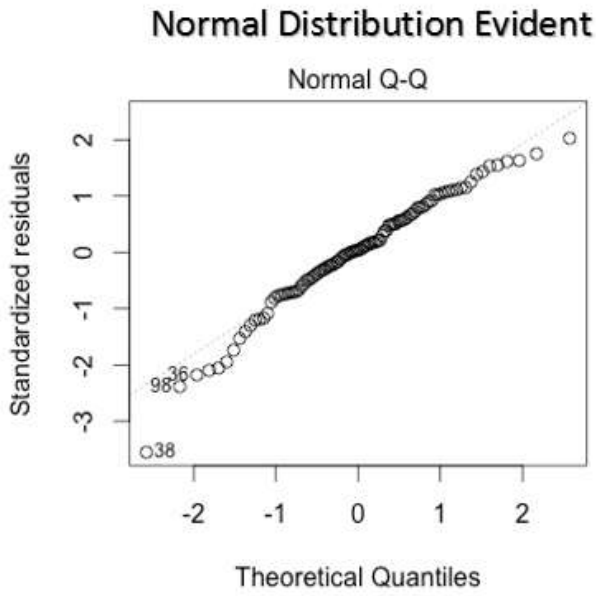
- **How to check:** Look for Durbin – Watson (DW) statistic. It must lie between 0 and 4. If DW = 2, implies no autocorrelation, $0 < DW < 2$ implies positive autocorrelation while $2 < DW < 4$ indicates negative autocorrelation.
- Also, you can see residual vs time plot and look for the seasonal or correlated pattern in residual values.

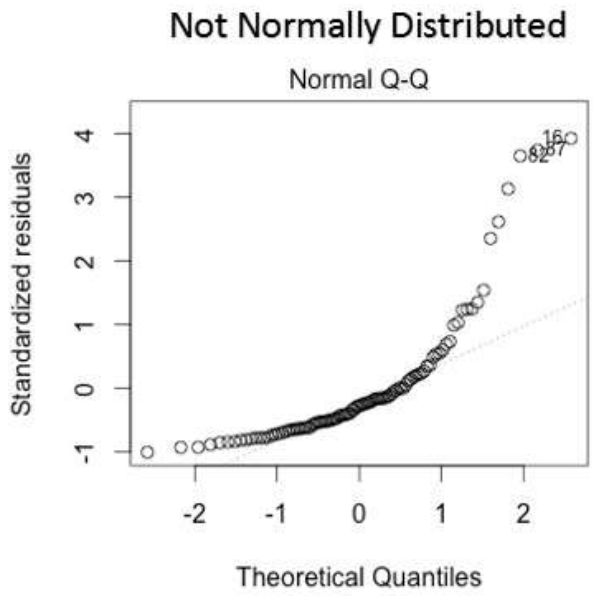
Residuals and Fitted Values



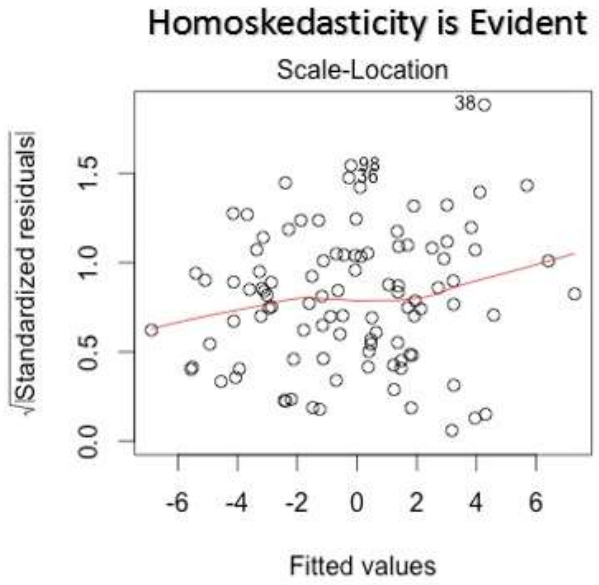


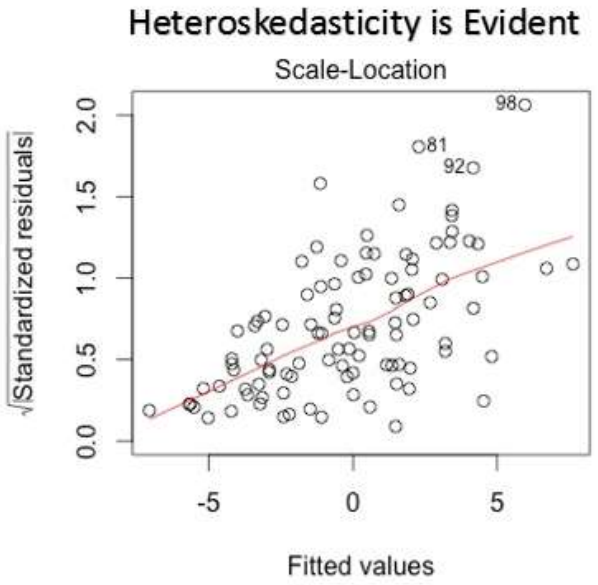
QQ Plots


Normal Distribution Evident
Normal Q-Q
Standardized residuals
Theoretical Quantiles
-3
-2
-1
0
1
2
98
36
38

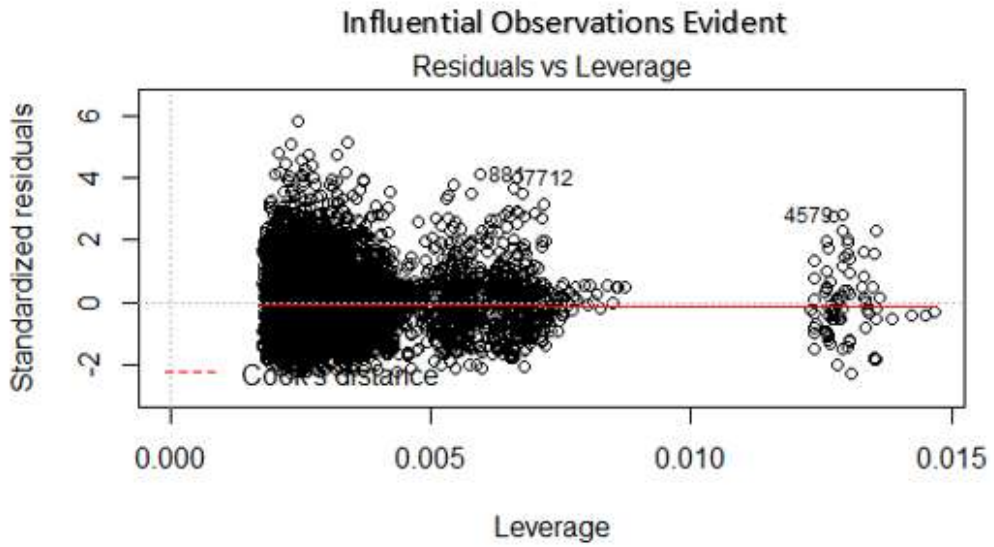

Not Normally Distributed
Normal Q-Q
Standardized residuals
Theoretical Quantiles
-1
0
1
2
3
4
16

Scale Location Plot





Residuals vs Leverage Plot



FAQ: What if Assumptions are Violated?

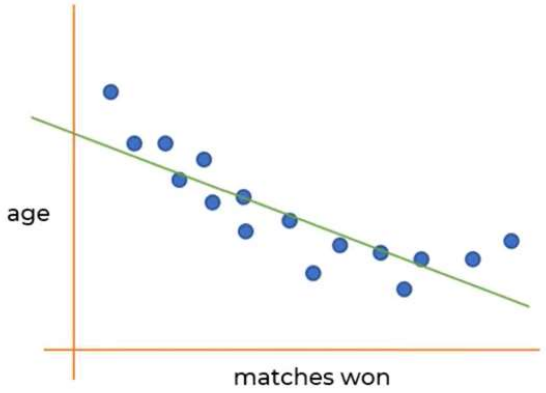
- You cannot rely on the coefficients **m** and **b**

Model Evaluation

- Machine learning model cannot have 100 per cent efficiency otherwise the model is known as a biased model.
- It is necessary to obtain the accuracy on training data, But it is also important to get a genuine and approximate result on unseen data otherwise Model is of no use.
- So to build and deploy a generalized model we require to Evaluate the model on different metrics which helps us to better optimize the performance, fine-tune it, and obtain a better result.

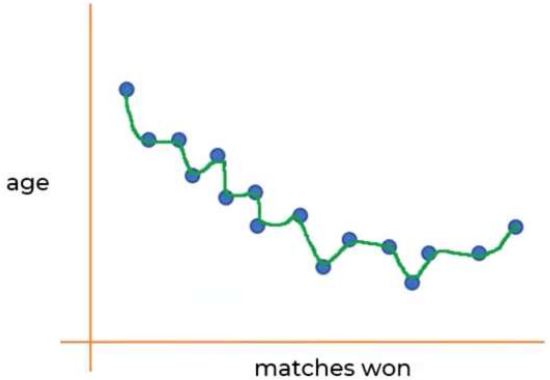
Overfitting and Underfitting

underfit



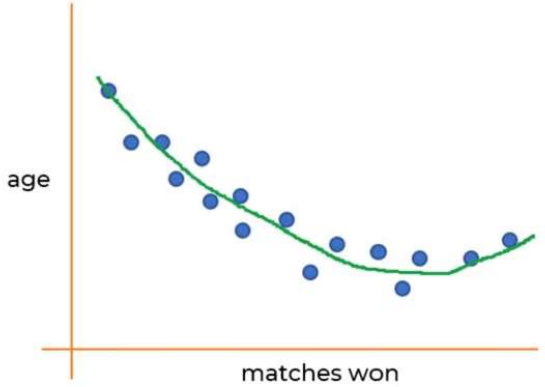
$$\text{match won} = \theta_0 + \theta_1 * age$$

overfit



$$\text{match won} = \theta_0 + \theta_1 * age \quad +\theta_2 * age^2 \quad +\theta_3 * age^3 \quad +\theta_4 * age^4$$

balanced fit



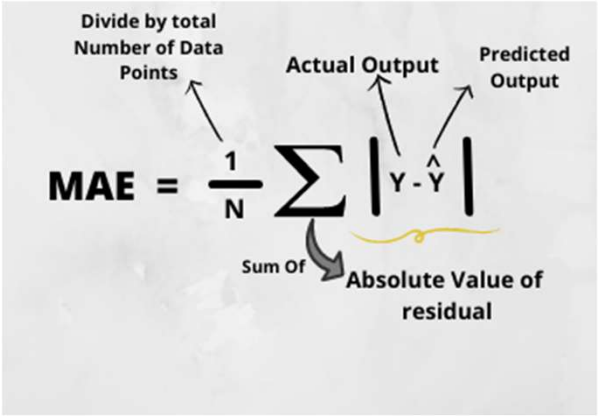
$$\text{match won} = \theta_0 + \theta_1 * age \quad +\theta_2 * age^2$$

Loss Functions

- MAE, MSE, RMSE are Loss Functions
- You would like to minimize them

Mean Absolute Error (MAE)

- MAE is a very simple metric which calculates the absolute difference between actual and predicted values.
- Advantages of MAE
  - The MAE you get is in the same unit as the output variable.
  - It is most Robust to outliers.
- Disadvantages of MAE
  - The graph of MAE is not differentiable so we have to apply various optimizers like Gradient descent which can be differentiable.




```
from sklearn.metrics import mean_absolute_error
```

Mean Squared Error (MSE)

- Mean squared error states that finding the squared difference between actual and predicted value.
- **Advantages of MSE**
  - The graph of MSE is differentiable, so you can easily use it as a loss function.
- **Disadvantages of MSE**
  - The value you get after calculating MSE is a squared unit of output.
  - If you have outliers in the dataset then it penalizes the outliers most and the calculated MSE is bigger. So, in short, It is not Robust to outliers which were an advantage in MAE.

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)}^{2}$$

The square of the difference between actual and predicted

```
from sklearn.metrics import mean_squared_error
```

Root Mean Squared Error (RMSE)

- It’s the square root valued of MSE
- **Advantages of RMSE**
  - The output value you get is in the same unit as the required output variable which makes interpretation of loss easy.
- **Disadvantages of RMSE**
  - It is not that robust to outliers as compared to MAE.
  - For performing RMSE we have to NumPy NumPy square root function over MSE.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)}$$

Residual Standard Error (RSE)

- Due to the presence of error terms, even if we know the true regression line (m, b known), we would not be able to perfectly predict Y from X.
- The RSE is an estimate of standard deviation of **e**
- It is the average amount that the response will deviate from the true regression line
- RSE provides an absolute measure of lack of fit of the model to the data

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

R-Square

- R2 squared is also known as Coefficient of Determination or sometimes also known as Goodness of fit.
- In contrast, MAE and MSE depend on the context as we have seen whereas the R2 score is independent of context.
- R2 score is a metric that tells the performance of your model
- As regression line moves towards perfection, R2 score move towards one. And the model performance improves.

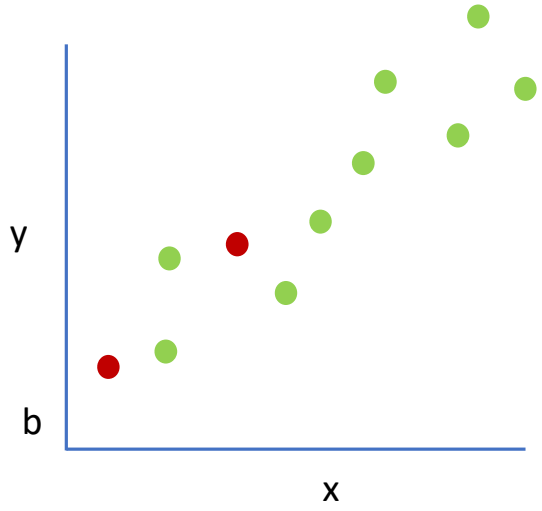
$$\text{R2 Squared} = 1 - \frac{SSr}{SSm}$$

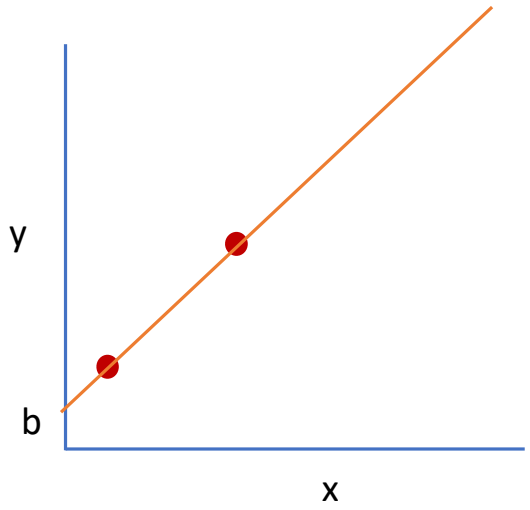
SSr = Squared sum error of regression line

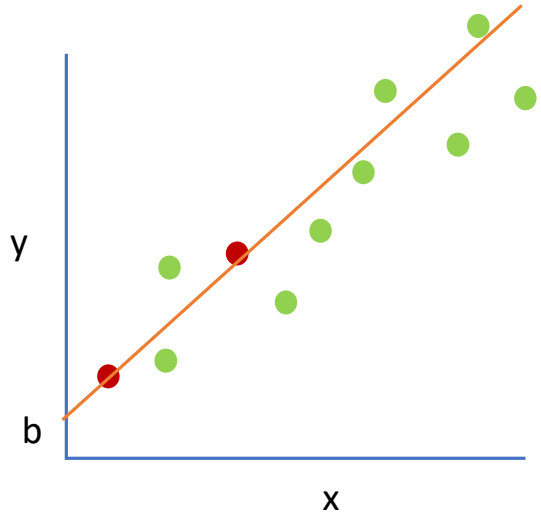
SSm = Squared sum error of mean line

```
from sklearn.metrics import r2_score
```

L1/L2 Regularization

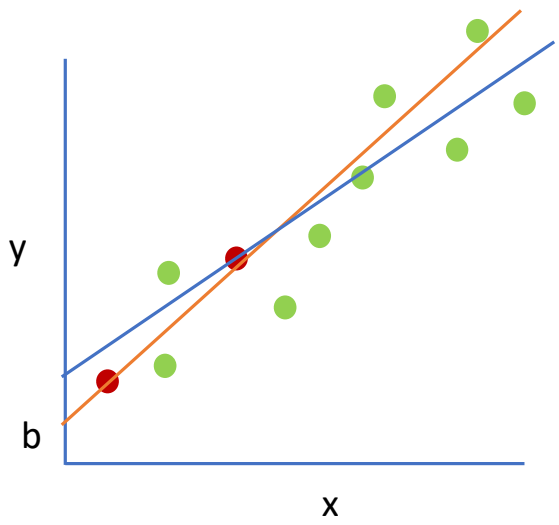






Bias is low
Variance is high for the test data
Accuracy reduces

L1/L2 Regularization



Slope changes

As slope changes, unit change in x result in appropriate changes in y

SSR + λ ($slope$) -> L1 Regularization -> Lasso Regression
SSR + λ $(slope)^2$ -> L2 Regularization -> Ridge Regression

By adding a little "penalty" the sum-of-squares line, we slightly increase the bias. There will be significant improvement in lowering the variance in the test data w.r.t to the new line.

Lasso & Ridge regression techniques are used to counter the overfitting which may result from the model complexity in simple linear regression.

L1 Regularization: Lasso Regression

Increases

$$mse = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - h_\theta(x_i)\right)^2 + \lambda \sum_{i=1}^{n} |\theta_i|$$

Slope

$$h_\theta(x_i) = \theta_0 + \theta_1 x_1 + \theta_2 x_2^2 + \theta_3 x_3^3$$

Lasso Regression

- “LASSO” denotes Least Absolute Shrinkage and Selection Operator
- Lasso regression follows the regularization technique to create prediction. It is given more priority over the other regression methods because it gives an accurate prediction. Lasso regression model uses shrinkage technique.
- In this technique, the data values are shrunk towards a central point similar to the concept of mean. The lasso regression algorithm suggests a simple, sparse models (i.e. models with fewer parameters), which is well-suited for models or data showing high levels of multicollinearity or when we would like to automate certain parts of model selection, like variable selection or parameter elimination using feature engineering.

L2 Regularization: Ridge Regression

$$mse = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - h_\theta(x_i)\right)^2 + \lambda \sum_{i=1}^{n} \theta_i^2$$

$$h_\theta(x_i) = \theta_0 + \theta_1 x_1 + \theta_2 x_2^2 + \theta_3 x_3^3$$

Ridge Regression

- Ridge Regression is another type of regression algorithm in data science and is usually considered when there is a high correlation between the independent variables or model parameters.
- As the value of correlation increases the least square estimates evaluates unbiased values. But if the collinearity in the dataset is very high, there can be some bias value.
- Therefore, we create a bias matrix in the equation of Ridge Regression algorithm. It is a useful regression method in which the model is less susceptible to overfitting and hence the model works well even if the dataset is very small.

Scikit Learn

- What is it?
- Official Documentation
- Installation
  - **pip install scikit-learn**

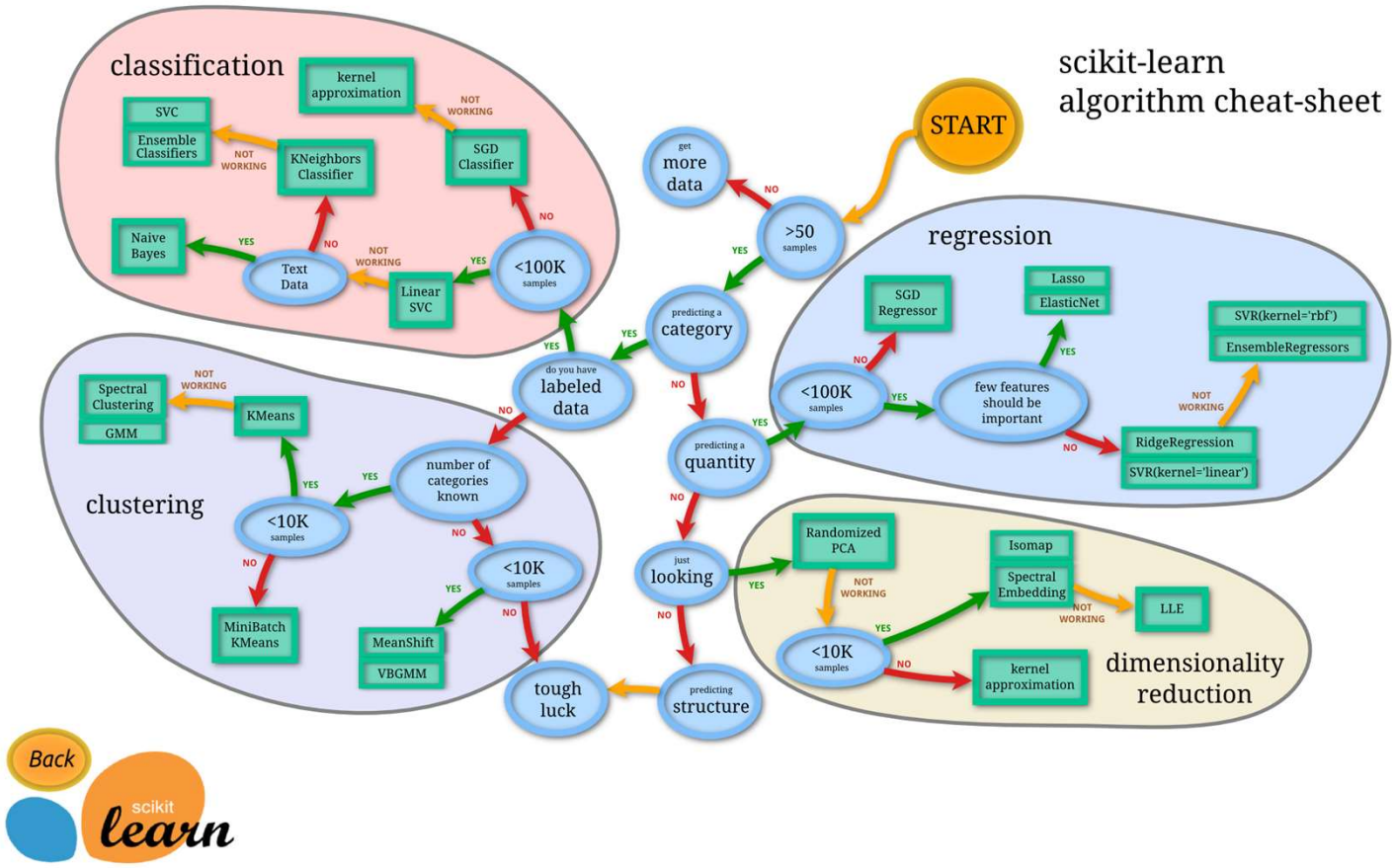
Process of using scikit-learn

- Every algorithm is exposed in scikit-learn via an “Estimator”
- You will have to import the model
  - Syntax: **from sklearn.family import Model**
  - Example: **from sklearn.linear_model import LinearRegression**

Estimator/Model itself

- Estimators can have several parameters
  - Set during instantiating
  - Possible values can seen in jupyter notebooks using shift + tab
- Split the data using **train_test_split()** method
- Fit the model using **model.fit()** method
  - Supply the x_train and y_train data
- Predict the values using **model.predict()** and **model.score()** for new set of data
- Methods such as **model.transform()** and **model.fit_transfor()** available for unsupervised data
- Evaluation of the model
  - Depends on the algorithm

Scikit-learn Cheat Sheet



Complete Flow: Demonstration

- Regular Linear Regression
- Lasso Regression (shown later)
- Ridge Regression (shown later)