LOGISTIC REGRESSION ASSUMPTIONS

- When the assumptions of logistic regression analysis are not met problem such as biased coefficient may lead to invalid statistic inference.

ASSUMPTION 1 → Must be Output Categorical Column.
ASSUMPTION 2 → Linearity of independent variable and log odds.
ASSUMPTION 3 → No strong influential outliers (Cook distance, Box plot)
ASSUMPTION 4 → Absence of Multicollinearity (Correlation & VIF)
ASSUMPTION 5 → Independence of Observations (Autocorrelation, Residual plot)
ASSUMPTION 6 → Sufficiently large sample size

ASSUMPTION 1 → Output is Categorical Column.

→ We can directly check datatype and distribution of the target column.

ASSUMPTION 2 → Linearity of independent variable and log odds.

→ Log of Odds solve the problem of symmetry.

- Suppose my team is bad
  Odd = 1/4 = 0.25, Worse = 1/8 = 0.125
  More worse = 1/16 = 0.06, Full worse = 0/anything = 0

Suppose my team is good.
Win = 5/3 = 1.7 improves 9/3 = 3
More improvement = 27/3 = 9.
that will go to = an ∞.

$$\text{Odds} = \frac{\text{something happening}}{\text{something not happening}} = \frac{P}{1-P}$$

So odds (win) will go from 1 to ∞
And odds (loss) will go from 0 to 1.

So, symmetry is not there here which is assymetry meaning it makes it difficult to compare Odd (win) and Odd (loose)
Suppose, if Odds are against 1 to 6, Odds = 1/6 = 0.17 but in favour, 6/1 = 6

So, logs of Odds will make it symmetry.
Log of odd against 1 to 6 = log (Odds) = log (1/6) = log (0.17) = -1.79
Log of odd against 6 to 1 = log (Odds) = log (6/1) = log (6) = 1.79
In short log make the midpoint to 0.

- There are 2 methods to check Assumption 2.
  i) Create a log variable of Independent variable and check p value.
  ii) Visualization

1) Box Tidewell Test

- Add log-transform interaction variable between continuous independent variable and run the model.
- Suppose we have two variables: Age and Fare.
  Transform them to natural log: Log(Age) and Log(Fare).
  Run the logistic model and check p value.

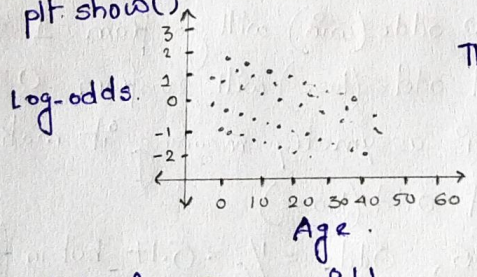
Age	- 0.051
Log(Age)	- 0.101
Fare	- 0.000
Log(Fare)	- 0.000

Log(Fare) < 0.001, statistically significant.
Log(Age) = 0.101, statistically insignificant.
This means there is non-linearity in Fare Feature is < 0.001, and assumption has been violated.
→ We can resolve this by including a polynomial term (eg Fare ^ 2) to account for non-linearity.

11) Visual Check

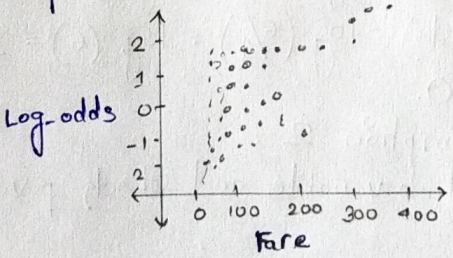
- Check scatter plot between Independent Variable and Log-odds(predicted)

eg, logit_result s = GLM (y, X, family = families.Binomial()).fit()
predicted = logit_results.predict(X)
# Get log odds value.
log_odds = np.log(predicted / (1-predicted))
# plot for Age variable.
plt.scatter(x = df_titanic['Age'].values, y= log_odds).
plt.show()



There is a linearity between Age & log-odds(predicted)

plot for Fare variable.
plt.scatter(x = df_titanic['Fare'].values, y=log_odds).
plt.show()



There is non linearity between log-odds and Fare.
So, we will use log(Fare) in and check linearity.

ASSUMPTION 3 – NO STRONG INFLUENTIAL OUTLIERS

- It distort accuracy of the model.

Cook Distance →

- Cook's distance is the scale change in fitted values, which is useful for identifying in the X values.
- Cook's distance shows the influence of each observations on the fitted response value.
- An observations with Cook distance larger than three/four times the mean Cook's distance might be an outlier.

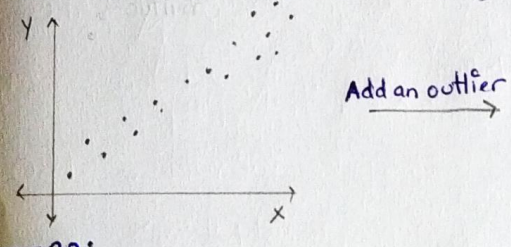
Definition →

- Each element in the Cook distance D is the normalized change in fitted response value due to deletion of an observation.

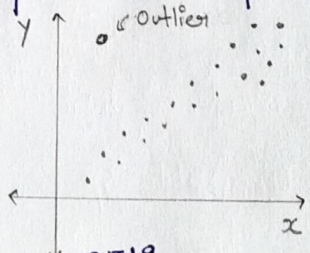
$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\,MSE}$$

$\hat{Y}_j$ → j^{th} fitted response value, $\hat{Y}_{j(i)}$ → j^{th} fitted response value, where the fit does not include observation j.

MSE → Mean Square Error, p → Number of coefficient in the regression mod



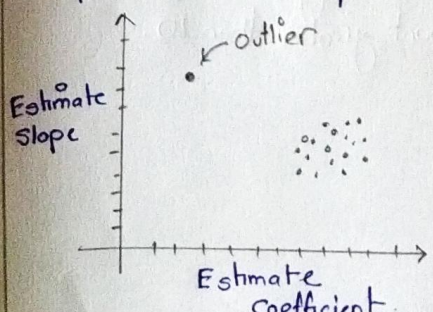
n=20;
Coefficient (b_0) = 1.732 [Intercept coefficient]
Slope Coefficient (b_1) = 5.117



n=19.
Intercept Coefficient (b_0) = 8.51
Slope Coefficient (b_1) = 3.32

This estimate change substantially, when there is an outlier.

So what we do is we remove one datapoint at a iteration and check the intercept coefficient and slope coefficient. Then plot the output value of Estimate Slope Coefficient (b_1) and Estimate Intercept (b_0).



→ So we use Cook Distance
→ Cut off is subject to visualization.
→ Note: Variable have high Cook Distance, does not immediately removed. check other dimension / angle of the variable

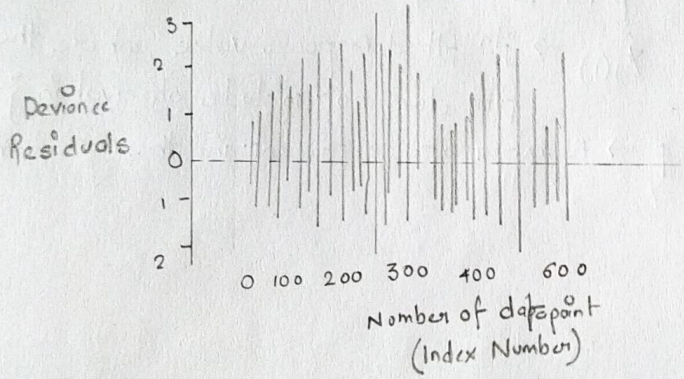
ASSUMPTION 4 - ABSENCE OF MULTICOLLINEARITY.

- Multicollinearity corresponds to a situation where the data contains highly correlated variable.
- Leads to reduce the precision of estimated coefficients.

1) Correlation Matrix → It check correlation between Independent variable.
   - If it exist between combination of more than two independent variable use VIF.
2) Variation Inflation Factor → VIF is equal to the ratio of the overall model variance to the variance of a model that include only one single independent variable.

ASSUMPTION 5 → INDEPENDENCE OF OBSERVATIONS (AUTO CORRELATION)

- We can also create Residual plot where we plot the deviance residual (Residual Deviance).



ASSUMPTION 6 → SUFFICIENTLY LARGE SAMPLE

- Atleast 10 observation of each Independent category.

Target Column → Survived

Independent Column → Gender (M, F), Trave (Alone, Family)

	Survived - Yes	Survived - No	
M	22	11	
F	36	2	So all categories (sub) have
Alone	13	17	count greater than 10.
Family	15	16	