# DETAILED DESCRIPTION OF LOGISTIC REGRESSION

1. Since, we have to divide the data into two parts, we will create a response variable called hits, where a rating of 6.0 or above is considered as a hit movie, and less than 6.0 is considered as flop.

The code used is as follows –

**proc** **import** datafile='D:\Users\asingh43\Desktop\BIA\CSM.xls' DBMS=xls out=CSM replace;

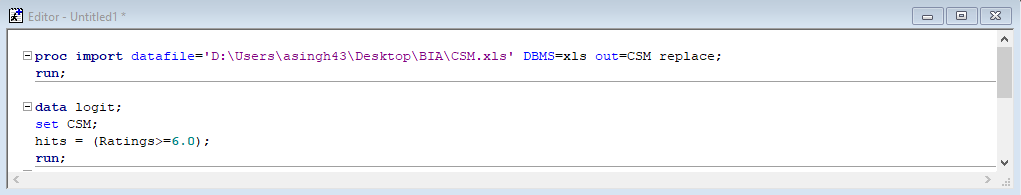
**run**;

**data** logit;

set CSM;

hits = (Ratings>**6.0**);

**run**;



1. We then run a logistic regression by exploring the relationship of hits with number of likes (Likes), number of comments (Comments) and number of screens (Screens).

The code is as follows -

**proc** **print** data = logit;

**run**;

**proc** **logistic** data = logit descending;

model hits = Likes Comments Screens;

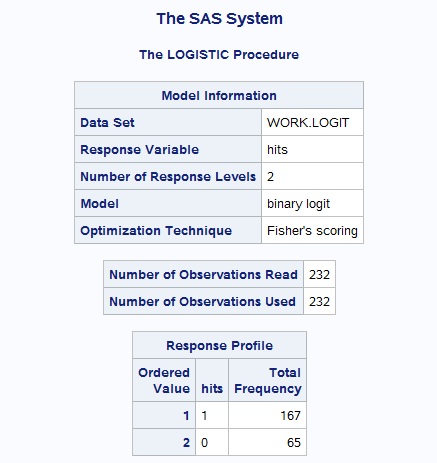
**run**;

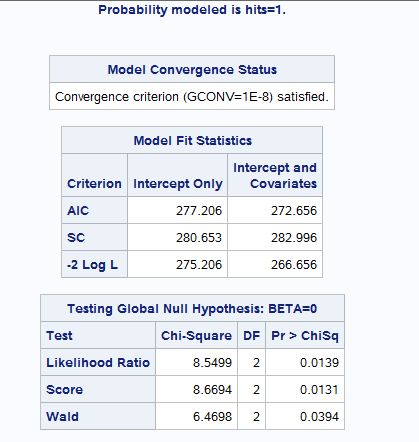
**proc** **logistic** data = logit descending;

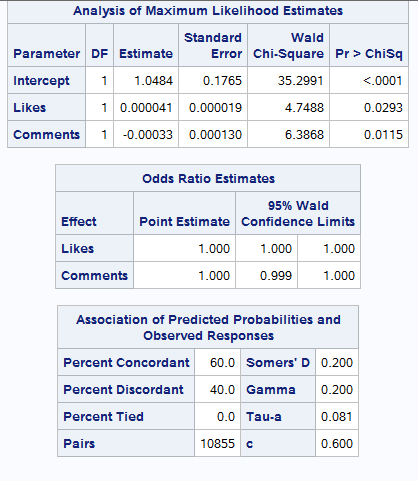
model hits = Likes Comments;

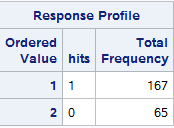
**run**;

Output for SAS Logistocs Regression will be –









* This is the frequency distribution of the response variable. Our response variable has 167 observations with a high write score and 65 with a low write score. It tells that how many number of data is with “Hit movie category” and “Flop movie category”.
* The Model Convergence Status (GCONV = 1E-8) is satisfied. It is very important that the model gets converged otherwise it is not a suitable model to proceed with.
* Likelihood Ratio – The LR Chi-Square Test shows that at least the one of the predictor’s regression coefficient is not equal to zero in the model. The LR-Chi Square Statistic can be calculated by subtracting -2Log L (Intercept only) with -2LogL (Intercepts and Covariates),  
  i.e. 275.206 – 266.656 = 8.5499
* Estimates – These are the binary logic regression estimates for the parameters in the model. The logistic model models the log odds of a positive response as a linear combination of the predictor variables.

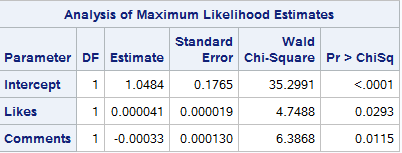
Log [p/ (1-p)] = b0 + b1\*Likes + b2\*Comments + b3\*Screens,

Where p is the probability that hits is 1.

Log [p/ (1-p)] = 0.9736 + 0.000041\*likes -0.00033\*Comments+ 0.000039\*Screens

The interpretation for this is 🡪 For one unit change in the predictor variable, the difference in the log-odds for a positive outcome is expected to change by the respective coefficient, given the other variables in the model are held constant.

Now, if we check the pr value for variable, we see that pr value for Screens is > 0.05 i.e. 0.7062 > 0.05 and hence **we will drop the variable** and run the model again as the model is not fit with the high pr value.

.

**Likes** - The difference in log-odds is expected to increase by 0.000041 unit, given the other variables in the model are held constant.

**Comments** -- The difference in log-odds is expected to decrease by 0.00033 unit, given the other variables in the model are held constant.

Now, if we see pr value is fine for this model.