Michal Chowaniak

CSC 423

**Logistic regression step-by-step**

1. Prepare boxplots for varialbes
2. Check for one independent binary variable
3. Number of observations
4. Check Samples size
5. Create dummy variables if necessary
6. Check for multicollinearity: Estimated Correlation Matrix
7. Check for highest influence variables
8. Check for significant predictors
9. Diagnostics and residual analysis:
   1. outliers (Pearson Residual or Deviance Residual abs val >= 3), influence diagnostics graph
   2. influential points: Threshold value |Dfbeta|2/sqrt(n) , influence diagnostics graph
10. Write full model
11. Split dataset into training and testing ( proc surveyselect)(use different seed value for each person in a group)
12. Build model with training data set only
13. Model selection method: Stepwise, Backwards
14. Model selection criteria: R2, AIC, SC, Goodness of Fit Test LR, Goodness of Fit Test p-value, predictors selected, standard error of predictors
15. Check for multicollinearity: Estimated Correlation Matrix
16. Diagnostics and residual analysis:
    1. outliers (Pearson Residual or Deviance Residual abs val >= 3), influence diagnostics graph
    2. influential points: Threshold value |Dfbeta|2/sqrt(n)
    3. influence diagnostics graph
17. Testing global Null Hypothesis
18. Write down the final model equation ( Analysis of Maximum Likelihood Estimates)
19. Analyze conditional effect of each variable, odds ration estimates
20. Using test dataset compute predictions (probability (phat), lower confidence interval (lcl), upper confidence interval (ucl)), predicted probability if Y
21. Generate classification table to identify cut off value, e.g. 0.5
22. Using cut-off value of eg. 0.5 classify predicted probability of Y into 1 (over 0.5) or 0 (less or equal 0.5)
23. Compare observed Y with predicted Y
24. Measure performance based on Observed vs Predicted Y: classification matrix
25. Find out TP, FP, TN, FN
26. Calculate sensitivity, specificity, accuracy, precision, F-metric
27. Based on above calculated numbers decide if this model is good or bad.