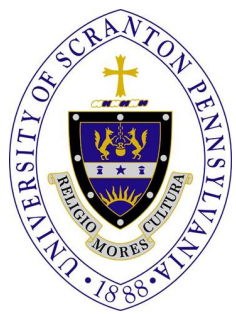
Classification: Logistic Regression
Part 3: Variable Selection and Using Logistic Regression for Profiling

Aram Balagyozyan

Department of Operations and Information Management
Kania School of Management
The
University of Scranton

January 13, 2020



1. Variable Selection
2. Logistic Regression for Profiling

Variable Selection Guidelines

- As it was the case in linear regression modeling, searching for alternative models is an important step in classification tasks.
- One option is to look for simpler models by trying to reduce the number of predictors used.
- We can also build more complex models that reflect interactions among predictors by creating and including new variables that are derived from the predictors.
- The choice among the set of alternative models is guided primarily by performance on the validation data. For models that perform roughly equally well, simpler models are generally preferred over more complex models.

Variable Selection Guidelines

- As in linear regression, in logistic regression we can use automated variable selection heuristics such as stepwise selection, forward selection, and backward elimination.
- In R, use function `step()` in the **stats** package or function `stepAIC()` in the **MASS** package for stepwise, forward, and backward elimination.
- If the dataset is not too large, we can even try an exhaustive search over all possible models (use R function `glmulti()` in package **glmulti**, although it can be slow).

Variable Selection Guidelines

- Note that performance on validation data may be overly optimistic when it comes to predicting performance on data that have not been exposed to the model at all.
- This is because when the validation data are used to select a final model among a set of model, we are selecting based on how well the model performs with those data and therefore may be incorporating some of the random idiosyncrasies of the validation data into the judgment about the best model. The model still may be the best for the validation data among those considered, but it will probably not do as well with the unseen data.
- Therefore, it is useful to evaluate the chosen model on a new test set to get a sense of how well it will perform on new data.

Logistic Regression for Profiling

- The presentation of logistic regression so far has been primarily from a data mining perspective where classification or ranking is the goal, and performance is evaluated by reviewing results with a validation sample.
- Sometimes, instead of trying to predict an important class, your goal may be to find important factors that explain the difference between records with different classes. This task is called predictor profiling.
- When the purpose of the analysis is profiling (identifying predictor profiles that distinguish the two classes, or explaining the differences between the classes in terms of predictor values), we are less interested in how well the model classifies new data than in how well the model fits the data it was trained on.
- For example, if we are interested in characterizing the average loan offer acceptor vs. nonacceptor in terms of income, education, and so on, we want to find a model that fits the data best.

Overall Strength-of-Fit in Logistic Regression

- As in multiple linear regression, we first evaluate the overall explanatory power of the model before looking at single predictors. We ask: Is this set of predictors better than a simple naive model for explaining the difference between classes?
- The deviance D is a statistic that measures overall goodness of fit. It is similar to the concept of sum of squared errors (SSE) in the case of least squares estimation (used in linear regression).
- We compare the deviance of our model, D (called Residual deviance in R), to the deviance of the naive (Null) model, D_0, with no explanatory (X) variables.
- Both measures are presented as part of the standard logistic regression output when you run `summary(logit.reg)`. In the interest of space, here I will pull out only these two deviance measures for the logistic regression we've run on the training set.

Overall Strength-of-Fit in Logistic Regression

```
logit.reg$deviance
```

```
## [1] 690.8251
```

```
logit.reg$null.deviance
```

```
## [1] 1915.103
```

- It appears our model provides a good overall fit, better than a model with no explanatory variables.

Impact of Single Predictors

- As in multiple linear regression, the output from a logistic regression procedure typically yields a coefficient table, where for each predictor X_i, we have an estimated coefficient b_i and an associated standard error.
- The associated p-value indicates the statistical significance of the predictor X_i, with very low p-values indicating a statistically significant relationship between the predictor and the outcome (given that the other predictors are accounted for), a relationship that is most likely not a result of chance.

Impact of Single Predictors

- Three important points to remember are:
  1. A statistically significant relationship is not necessarily a practically significant one, in which the predictor has great impact.
  2. A statistically significant predictor means that on average, a unit increase in that predictor is associated with a certain effect on the outcome (holding all other predictors constant). It does not, however, indicate predictive power. Statistical significance is of major importance in explanatory modeling, or profiling, but of secondary importance in predictive modeling (classification). In predictive modeling, statistically significant predictors might give hints as to more and less important predictors, but the eventual choice of predictors should be based on predictive measures, such as the validation set, confusion matrix (for classification), or the validation set lift chart (for ranking).
  3. Comparing the coefficient magnitudes, or equivalently the odds magnitudes, is meaningless unless all predictors have the same scale.