Paper D10-2009

Ranking Predictors in Logistic Regression

Doug Thompson, Assurant Health, Milwaukee, WI

ABSTRACT

There is little consensus on how best to rank predictors in logistic regression. This paper describes and illustrates six possible methods for ranking predictors: 1) standardized coefficients, 2) p-values of Wald chi-square statistics, 3) a pseudo partial correlation metric for logistic regression, 4) adequacy, 5) c-statistics, and 6) information values. The methods are illustrated with examples using SAS PROC LOGISTIC and GENMOD. The interpretation of each ranking method is outlined, and upsides and downsides of each method are described.

INTRODUCTION

Logistic regression is a common method for modeling binary outcomes, e.g., buy/no buy, lapse/renew, obese/not obese. A frequently asked question is, which variables in the predictor set were most strongly associated with the outcome? To answer this question, it is necessary to rank the predictors by some measure of association with the outcome. Unlike linear regression, where predictors are often ranked by partial correlation, there is less consensus on how best to rank predictors in logistic regression. The purpose of this paper is to describe, illustrate and compare six options for ranking predictors in logistic regression: 1) standardized coefficients, 2) p-values of Wald chi-square statistics, 3) a pseudo partial correlation metric for logistic regression, 4) adequacy, 5) c-statistics, and 6) information values. These options are not exhaustive of all ranking methods – there are many other possibilities. These ones were chosen because the author has used them to rank predictors in logistic regression, or has seen others do so.

To illustrate the methods for ranking predictors in logistic regression, data from the National Health and Nutrition Examination Survey (NHANES), 2005-2006, was used. The data are available for free download from http://www.cdc.gov/nchs/nhanes.htm. Obesity (a binary outcome, obese vs. non-obese) was modeled as a function of demographics, nutrient intake and eating behavior (e.g., eating vs. skipping breakfast). The predictor set exhibits typical characteristics of data used in logistic regression analyses – the predictors include a mix of binary and continuous variables and the continuous variables differ in scale. Due to the scale differences, the predictors cannot be ranked in terms odds ratios or unstandardized coefficients (these would be a good option for ranking if all predictors were on the same scale). Importantly, prior to analysis, observations with any missing data were excluded. The methods described below only yield valid rankings when used with complete data, otherwise ranking may be impacted by using different subsets of observations in different analyses. In practice, one would need to carefully consider the method for handling missing data as well as other analytic issues (e.g., whether or not linear representation is appropriate for continuous variables, influence and residual diagnostics), but these are deemphasized in this paper because the sole purpose was to illustrate approaches for ranking

predictors. After excluding the observations with missing data, 3,842 observations remained in the sample.

RANKING METHODS

Six methods were used to rank predictors in the logistic regression model. The methods and their implementation in SAS are described in this section.

1. Standardized coefficients

Standardized coefficients are coefficients adjusted so that they may be interpreted as having the same, standardized scale and the magnitude of the coefficients can be directly compared (ranked). The greater the absolute value of the standardized coefficient, the greater the predicted change in the probability of the outcome given a 1-standard deviation change in the corresponding predictor variable, holding constant the other predictors in the model. A downside of this method is that standardized scale may not be as intuitive as the original scales. In PROC LOGISTIC, the standardized coefficient is defined as $\beta_i/(s/s_i)$, where β_i is the estimated unstandardized coefficient for predictor i, s_i is the sample standard deviation for predictor i, and $s = \pi/\sqrt{3}$. This definition enables the standardized logistic regression coefficients to have the same interpretation as standardized coefficients in linear regression (Menard, 2004). Standardized logistic regression coefficients can be computed in SAS by using the STB option in the MODEL statement of PROC LOGISTIC. Taking the absolute value of the standardized coefficients enables them to be ranked from highest to lowest, in order of strength of association with the outcome.

2. P-value of the Wald chi-square test

Another approach is to rank predictors by the probability of the Wald chi-square test, H_0: $\beta_i = 0$; the null hypothesis is that there is no association between the predictor i and the outcome after taking into account the other predictors in the model. Small p-values indicate that the null hypothesis should be rejected, meaning that there is evidence of a non-zero association. This metric only indicates the strength of evidence that there is some association, not the magnitude of the association. Thus the ranking is interpreted as a ranking in terms of strength of evidence of non-zero association. Strengths of this method are ease of computation (this is a default in PROC LOGISTIC) and familiarity of many audiences with p-values. A downside is that it indicates only the strength of evidence that there is some effect, not the magnitude of the effect. In this paper, predictors were ranked by one minus the p- value of the Wald chi-square statistic, so that a ranking from largest to smallest would indicate descending strength of evidence of an association with the outcome.

3. Logistic pseudo partial correlation

In linear regression, partial correlation is the marginal contribution of a single predictor to reducing the unexplained variation in the outcome, given that all other predictors are already in the model (Neter, Kutner & Wasserman, 1990). In other words, partial correlation indicates the explanatory value attributable to a single predictor after taking into account all of the other predictors. In linear regression, this is expressed in terms of the reduction of sum of squared error attributable to an

individual predictor. Because estimation in logistic regression involves maximizing the likelihood function (not minimizing the sum of squared error), a different measure of partial correlation is needed. A partial correlation statistic for logistic regression has been proposed (Bhatti et al, 2006), based on the Wald chi-square statistic for individual coefficients and the log-likelihood of an intercept-only model. The pseudo partial correlation is defined as r = ± the square root of $((W_i\text{-}2K)/\text{-}2LL_{(0)})$, where W_i is the Wald chi-square statistic for predictor i, $(\beta_i/SE_{\beta i})^2$, $-2LL_{(0)}$ is -2 times the log likelihood of a model with only an intercept term, and K is the degrees of freedom for predictor i (K=1 in all of the examples presented in this paper). While this statistic has the same range as partial correlation in linear regression and there are some similarities in interpretation (e.g., the closer to 1 or -1, the stronger the marginal association between a predictor and the outcome, taking into account the other predictors), they are not the same, therefore the statistic is called "pseudo partial correlation" in this paper. The magnitude of the difference can be gauged by estimating a model of a continuous outcome (say, BMI) using maximum likelihood (PROC GENMOD with DIST=normal and LINK=identity), computing the pseudo partial correlation based on estimates from this model (i.e., Wald chi-squares and intercept-only model log likelihood), and then comparing the result with partial correlation estimates of the same model estimated via ordinary least squares (OLS, e.g., using PCORR2 in PROC REG). Differences may be substantial. In addition to the lack of identity with partial correlation based on OLS, another issue with this pseudo partial correlation is that the Wald chi-square statistic may be a poor estimator in small-to-medium size samples (Agresti, 2002). As noted by Harrell (2001), this problem might be overcome by using likelihood ratios rather than Wald chi-square statistics in the numerator, although that would involve more laborious computations, which may be why most published literature using this method has the Wald chi-square in the numerator. In this paper, logistic pseudo partial correlations were computed using the SAS code shown below.

```
proc logistic data=nhanes_0506 descending;
model obese =
breakfast
Energy
fat
protein
Sodium
Potassium
female
afam
college_grad
married
RIDAGEYR / stb;
ods output parameterestimates=parm;
run;

* Intercept-only model;
proc logistic data=nhanes_0506 descending;
model obese = ;
ods output fitstatistics=fits;
run;

data _null_;
set fits;
```

```
call symput('m2LL_int_only',InterceptOnly);
run;

data parm2;
set parm;
if WaldChiSq<(2*df) then r=0;
else r=sqrt( (WaldChiSq-2*df)/&m2LL_int_only );
run;
```

4. Adequacy

Adequacy is the proportion of the full model log-likelihood that is explainable by each predictor individually (Harrell, 2001). One can think about this in terms of the explanatory value of each predictor individually, relative to the entire set. This equates "explanatory value" with -2*log likelihood (-2LL). The full model (with all predictors) provides the maximal -2LL. -2LL can also be computed for each predictor individually (i.e., -2LL in a model including only a single predictor) and compared with the full model -2LL. The greater the ratio, the more of the total -2LL that could be explained if one had only the individual predictor available. An advantage of this metric is its interpretability – it indicates the proportion of the total explained variation in the outcome (expressed in terms of -2LL) that could be explained by a single predictor. Another strength relative to the logistic pseudo partial correlation is that it is entirely based on likelihood ratios, so it may be more reliable in small-to-medium samples. A downside is that adequacy can appear large, even if a predictor is weakly associated with the outcome – because adequacy is relative to the predictive value of the entire set of predictors, if the full model has a small -2LL, an individual predictor can have a large adequacy but small -2LL.

Adequacy was computed using the %adequacy_genmod macro provided below. The macro can be used to compute adequacy for a variety of generalized linear models (logistic, linear regression, poisson, negative binomial, etc.) by changing the LINK and DIST options.

```
%macro adequacy_genmod(inset=,depvar=,indvars=,link=,dist=);

proc contents data=&inset(keep=&indvars) out=names(keep=name) noprint;
run;

data _null_;
set names end=eof;
retain num 0;
num+1;
if eof then call symput('numvar',num);
run;

* Compute log likelihood of model with all predictors;

ods noresults;
proc genmod data=&inset descending namelen=100;
model &depvar = &indvars / link=&link dist=&dist;
ods output Modelfit=fit;
run;
ods results;

data fit2;
```

```
set fit;
interceptandcovariates=-2*value;
keep interceptandcovariates;
if Criterion='Log Likelihood';
run;

ods noresults;
proc genmod data=&inset descending namelen=100;
model &depvar = / link=&link dist=&dist;
ods output Modelfit=fit;
run;
ods results;

data fit3;
set fit;
interceptonly=-2*value;
keep interceptonly;
if Criterion='Log Likelihood';
run;

data allfit;
merge fit2 fit3;
run;

data _fit_all(rename=(diff=all_lr));
attrib var length=$100.;
set allfit;
all=1;
diff=interceptonly-interceptandcovariates;
keep all diff;
run;

* 2. Compute log likelihood for models with one predictor only;

   %do j=1 %to &numvar;
   %let curr_var = %scan(&indvars,&j);

   ods noresults;
   proc genmod data=&inset descending namelen=100;
   model &depvar = &curr_var / link=&link dist=&dist;
   ods output Modelfit=fit;
   run;
   ods results;

   data fit2;
   set fit;
   interceptandcovariates=-2*value;
   keep interceptandcovariates;
   if Criterion='Log Likelihood';
   run;

   ods noresults;
   proc genmod data=&inset descending namelen=100;
   model &depvar = / link=&link dist=&dist;
   ods output Modelfit=fit;
   run;
   ods results;
```

```
   data fit3;
   set fit;
   interceptonly=-2*value;
   keep interceptonly;
   if Criterion='Log Likelihood';
   run;

   data allfit;
   merge fit2 fit3;
   run;

   data _fit;
   attrib var length=$100.;
   set allfit;
   var="&curr_var";
   diff=interceptonly-interceptandcovariates;
   keep var diff;
   run;

   %if &j=1 %then %do;
      data redfit;
        set _fit;
        run;
   %end;
   %else %do;
      data redfit;
        set redfit _fit;
        all=1;
        run;
   %end;
   %end;

* 3. Compute the ratio of log likelihoods;

proc sql;
create table _redfit as select * from
redfit a, _fit_all b
where a.all=b.all;
quit;

data _redfit;
set _redfit;
a_statistic=(diff/all_lr);
drop all;
run;

proc sort data=_redfit(rename=(var=variable));
by variable;
run;

%mend adequacy_genmod;

%let vars=breakfast
Energy
fat
Protein
```

```
Sodium
Potassium
female
afam
college_grad
married
RIDAGEYR;

%adequacy_genmod(inset=nhanes_0506,depvar=obese,indvars=&vars,link=logit,dist
=bin);
```

5. Concordance / c-statistic

Concordance (also known as the c-statistic) indicates a model's ability to differentiate between the outcome categories. In the example used in this paper, call obese observations "cases" and call the non-obese "non-cases". Assume that a separate model of obesity is constructed for each predictor and the score is the predicted probability of obesity based on the predictor alone. One interpretation of the c-statistic is the probability that a randomly chosen case has a higher score than a randomly chosen non-case. The greater the c-statistic, the more likely that cases have higher scores than non-cases; thus the greater the ability of a model to differentiate cases from non-cases. The c-statistic only indicates the probability of scores being higher or lower, not the magnitude of the score difference; it is possible for the c-statistic to be large while on average there is little difference in scores between the outcome categories. In other words, a large c-statistic points to a consistent (but not necessarily large) difference in scores. One way to rank predictors is to compute a separate model for each predictor, estimate the c-statistic for each model, then rank the predictors in terms of these c-statistics. A c-statistic of 0.5 indicates no ability to differentiate between the outcome categories, while scores greater or less than 0.5 indicate some ability to differentiate. A c-statistic of 0.45 indicates the same ability to differentiate between the outcome categories as a c-statistic of 0.55. For this reason, a larger c-statistic does not necessarily indicate greater predictive value than a smaller c-statistic – what matters is how far the c-statistic is from 0.5. So that c-statistics can be compared and ranked, it makes sense to express c-statistics in terms of the absolute value of their difference from 0.5. The greater the difference, the greater the usefulness of a predictor in differentiating between the outcome categories. An advantage of this metric is that the c-statistic has an intuitive interpretation and c-statistics are familiar in some areas of analysis, e.g., biostatistics and marketing analytics. A disadvantage is that the other predictors are not taken into account.

The c-statistic is produced by default in PROC LOGISTIC. To estimate the absolute value of the difference between the c-statistic and 0.5, for each predictor individually, the %single_c macro was used.

```
%macro single_c;
%let vars =
breakfast
Energy
fat
Protein
Sodium
Potassium
female
```

```
afam
college_grad
married
RIDAGEYR;

%do i=1 %to 11;
%let curr_var = %scan(&vars,&i);

ods listing close;
proc logistic data=nhanes_0506 descending;
model obese =
&curr_var;
ods output association=assoc;
run;
ods listing;

data _assoc(rename=(nvalue2=c));
length variable $100;
set assoc;
keep nvalue2 variable;
if Label2='c';
variable="&curr_var";
run;

%if &i=1 %then %do;
data uni_c;
set _assoc;
run;
%end;
%else %do;
data uni_c;
set uni_c _assoc;
run;
%end;
%end;
%mend single_c;
%single_c;
```

6. Information value

Information values are commonly used in data mining and marketing analytics. Information values provide a way of quantifying the amount of information about the outcome that one gains from a predictor. Larger information values indicate that a predictor is more informative about the outcome. One rule of thumb is that information values less than 0.02 indicate that a variable is not predictive; 0.02 to 0.1 indicate weak predictive power; 0.1 to 0.3 indicate medium predictive power; and 0.3+ indicates strong predictive power (Hababou et al, 2006). Like the c-statistic method described above, the information value method of ranking predictors is based on an analysis of each predictor in turn, without taking into account the other predictors. Information values have an important advantage when dealing with continuous predictors – to compute the information value, continuous predictors are broken into categories, and if there is a large difference within any category, the information value will be large. Thus information values are sensitive to differences in the outcome at any point along a continuous scale. There are various formulas for computing information values (e.g., Witten & Frank,

2005); this paper uses the definition described by Hababou et al, 2006, as implemented in the %info_values macro provided below.

```
%macro info_values;
%let vars =
breakfast
Energy
fat
Protein
Sodium
Potassium
female
afam
college_grad
married
RIDAGEYR;

%let binary =
1
0
0
0
0
0
1
1
1
1
0;

%do i=1 %to 11;
%let curr_var = %scan(&vars,&i);
%let curr_binary = %scan(&binary,&i);

%if &curr_binary=0 %then %do;
proc rank data=nhanes_0506 out=iv&i groups=10 ties=high;
var &curr_var;
ranks rnk_&curr_var;
run;
%end;
%if &curr_binary=1 %then %do;
data iv&i;
set nhanes_0506;
rnk_&curr_var=&curr_var;
run;
%end;

proc freq data=iv&i(where=(obese=1)) noprint;
tables rnk_&curr_var / out=out1_&i(keep=rnk_&curr_var percent
rename=(percent=percent1));
run;

proc freq data=iv&i(where=(obese=0)) noprint;
tables rnk_&curr_var / out=out0_&i(keep=rnk_&curr_var percent
rename=(percent=percent0));
run;
```

```
data out01_&i;
merge out0_&i out1_&i;
by rnk_&curr_var;
sub_iv=(percent0-percent1)*log(percent0/percent1);
run;

proc means data=out01_&i sum noprint;
var sub_iv;
output out=sum&i sum=infoval;
run;

data _sum&i;
length variable $100;
set sum&i;
variable="&curr_var";
drop _type_ _freq_;
run;

proc datasets library=work nolist;
delete iv&i out1_&i out0_&i out01_&i sum&i;
run;
quit;

%if &i=1 %then %do;
data iv_set;
set _sum&i;
run;
%end;
%else %do;
data iv_set;
set iv_set _sum&i;
run;
%end;
%end;
%mend info_values;
%info_values;
```

RESULTS

Table 1 provides a brief definition of each predictor in the model, along with coefficients and standard errors.

Table 1. Predictors, description, coefficients and standard error estimates.

Predictor label	Description	Coefficient	Standard error
Intercept	Model intercept term	-0.99676	0.17688
Energy	24-hour energy intake (kcal)	-0.00021	0.00009
Potassium	24-hour potassium intake (mg)	-0.00020	0.00005
Protein	24-hour protein intake (gm)	0.00353	0.00150
RIDAGEYR	Age (years)	0.00427	0.00216
Sodium	24-hour sodium intake (mg)	0.00011	0.00003
afam	African American (1=yes, 0=no)	0.52313	0.08195

breakfast	Ate breakfast (1=yes, 0=no)	-0.19279	0.09231
college_grad	College graduate (1=yes, 0=no)	-0.42860	0.09115
fat	24-hour fat intake (gm)	0.00353	0.00162
female	Female (1=yes, 0=no)	0.33356	0.07672
married	Married (1=yes, 0=no)	0.18392	0.07182

Table 2 shows the results for each ranking method. As described in Methods, the metrics were coded so that higher values point to greater strength of association with the outcome (or stronger evidence of non-zero association with the outcome, in the case of p-values).

Table 2. Results for each ranking method by predictor.

Predictor label	Standardized estimate (abs. val.)	1 minus p-value	Logistic regression pt. corr.	Adequacy	C-statistic (abs. val. of c - 0.5)	Information value
afam	0.1211	1.0000	0.0884	0.3430	0.0523	5.9929
college_grad	0.0956	1.0000	0.0637	0.1825	0.0357	3.2792
Potassium	0.1417	1.0000	0.0574	0.1268	0.0411	3.1325
female	0.0919	1.0000	0.0584	0.1125	0.0352	1.9855
Sodium	0.1146	0.9993	0.0439	0.0123	0.0180	1.6853
Protein	0.0876	0.9814	0.0267	0.0001	0.0148	2.0344
Energy	0.1197	0.9789	0.0259	0.0103	0.0082	0.6156
married	0.0505	0.9896	0.0303	0.0040	0.0066	0.0704
fat	0.0926	0.9707	0.0236	0.0058	0.0108	0.6469
breakfast	0.0409	0.9632	0.0218	0.0388	0.0160	0.6804
RIDAGEYR	0.0417	0.9516	0.0196	0.0020	0.0121	8.8320

Table 3 compares the methods in terms of ranks assigned to each predictor by each method. Predictors are ordered by the mean rank across methods.

Table 3. Predictor ranks by each ranking metric.

Predictor label	Standardized estimate (abs. val.)	1 minus p-value	Logistic regression pt. corr.	Adequacy	C-statistic (abs. val. of c - 0.5)	Information value	Mean rank	Min. rank	Max. rank
afam	2	1	1	1	1	2	1.3	1	2
college_grad	5	2	2	2	3	3	2.7	2	5
Potassium	1	4	4	3	2	4	3.1	1	4
female	7	3	3	4	4	6	4.3	3	7
Sodium	4	5	5	6	5	7	5.3	4	7
Protein	8	7	7	11	7	5	7.4	5	11
Energy	3	8	8	7	10	10	7.7	3	10
married	9	6	6	9	11	11	8.3	6	11
fat	6	9	9	8	9	9	8.4	6	9
breakfast	11	10	10	5	6	8	8.6	5	11
RIDAGEYR	10	11	11	10	8	1	8.9	1	11

One clear result is that the ranking methods are not equivalent. The predictors were ranked differently by different methods -- in some cases the differences were substantial. For example, RIDAGEYR (age in years) was ranked last by two methods (p-value and pseudo partial correlation) but first by the information value method. The difference may be in part due to the fact that the p-value and pseudo partial correlation methods utilize the estimated effect of a predictor after taking into account the other predictors, while the information value method is based on an analysis of each predictor in turn without taking into account the other predictors. Another possible reason for the difference in rankings, as noted above, is that the information value method breaks continuous variables such as RIDAGEYR into categories prior to analysis, and this method is highly sensitive to differences in the outcome within any category. In the NHANES data used to illustrate the ranking methods, there was a large difference in obesity in one age range but minimal difference in other age ranges, leading to the high information value for RIDAGEYR. In contrast, RIDAGEYR was treated as continuous in the p-value and pseudo partial correlation methods, somewhat muting the estimated association between age and obesity.

The p-value and pseudo partial correlation methods yielded the same rankings, due to the fact that both methods were based on the magnitude of the Wald chi-square statistics. Although the methods did not differ in rankings, they have different interpretations.

With some exceptions, the c-statistic and information value methods yielded roughly similar rankings. This may be partly a function of the fact that both methods are based on analyses of individual predictors without taking into account the other predictors. Differences in rankings between the two methods may be partly due to the fact that the c-statistic method assumed linear association between continuous variables and the outcome, whereas the information value method broke continuous variables into categories.

CONCLUSION

This paper described and illustrated six possible methods for ranking predictors in logistic regression. The methods yielded different rankings, in some cases strikingly so. Differences in rankings may have been due to whether or not the other predictors were taken into account when ranking a predictor; whether predictors were treated as continuous or categorical; the statistics used as the basis of the rankings (e.g., Wald vs. likelihood ratio); and other factors.

One of the most important considerations when selecting a ranking method is whether the method yields a useful interpretation that can be readily communicated to the target audience. For example, adequacy has an interpretation that is relatively easy to understand and communicate – it is the proportion of the total explainable variation in the outcome (expressed as -2LL in the model with all predictors) that one could explain if one had only a single predictor available; in other words, it is the explanatory value of a single predictor relative to the entire set of predictors. In contrast, while information values have important strengths, they may be challenging to explain to some audiences.

Ranking predictors is the primary purpose of some analyses. For example, some analyses start with a large set of predictors and attempt to identify the top n (say, top 10) most strongly associated with an outcome. When ranking is the primary objective of an analysis, it may make most sense to use multiple

ranking methods, and if there are major differences, do further analyses to dig into the sources of the differences. However, when multiple ranking methods are used, it is advisable to choose a primary ranking method prior to analysis, in order to avoid the temptation to choose whichever ranking is most pleasing to the target audience.

REFERENCES

Agresti A. (2002). Categorical data analysis (2nd Ed.). Hoboken NJ: Wiley.

Bhatti IP, Lohano HD, Pirzado ZA & Jafri IA. (2006). A logistic regression analysis of ischemic heart disease risk. Journal of Applied Sciences, 6(4), 785-788.

Hababou M, Cheng AY, Falk R. (2004). Variable selection in the credit card industry. Paper presented to the Northeast SAS Users Group (NESUG).

Harrell F. (2001). Regression modeling strategies. New York: Springer.

Menard S. (2004). Six approaches to calculating standardized logistic regression coefficients. American Statistician, 58(3), 218-223.

Neter J, Wasserman W & Kutner MH. (1990). Applied linear statistical models (3rd Ed.). Boston MA: Irwin.

Witten IH & Frank E. (2005). Data mining (2nd Ed.). New York: Elsevier.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Doug Thompson
Assurant Health
500 West Michigan
Milwaukee, WI 53203
Work phone: (414) 299-7998
E-mail: Doug.Thompson@Assurant.com