Logistic Regression Training Material

Ou Zhang 11/20/2013

Logistic regression is part of a category of statistical models called generalized linear models. Logistic regression allows one to predict a discrete outcome, such as group membership, from a set of variables that may be continuous, discrete, dichotomous, or a mix of any of these. Generally, the dependent or response variable is dichotomous, such as presence/absence or success/failure. Discriminant analysis is also used to predict group membership with only two groups. However, discriminant analysis can only be used with continuous independent variables. Thus, in instances where the independent variables are a categorical, or a mix of continuous and categorical, logistic regression is preferred.

The Model:

The dependent variable in logistic regression is usually dichotomous, that is, the dependent variable can take the value 1 with a probability of success θ , or the value 0 with probability of failure $1-\theta$. This type of variable is called a Bernoulli (or binary) variable.

The relationship between the predictor and response variables is not a linear function in logistic regression, instead, the logistic regression function is used, which is the logit transformation of θ :

$$\theta = \frac{e^{(\alpha+\beta_1 x_1+..+\beta_i x_i)}}{1+e^{(\alpha+\beta_1 x_1+..+\beta_i x_i)}}$$

Where α = the constant of the equation and,

β = the coefficient of the predictor variables.

This code is used for screening clinical and non-clinical cases:

```
/*Logistical regression model check*/
data arthrits;
  input sex$ trtment$ improve$ count;
  _treat_=(trtment='Active');
  _sex_  =(sex='F');
  better =(improve='some');
/* some means improve, none means not improve*/
cards;
  F Active  none  6
  M Active  none  7
  F Active  some 21
  M Active  some  7
  F Placebo none 19
  M Placebo none 10
  F Placebo some 13
  M Placebo some  1
;

proc logistic data=arthrits descending;
  freq count;
  model better = _sex_ _treat_ / scale=none aggregate;
run;
```

- Use logistic regression to predict clinical cases.
- Create a screener test-only contains less amount of items.

Example:

Total-300 items, screener-50 items

Regardless what types of item response type, logistic regression can screen clinical and nonclinical cases.

Psychometrically speaking-using less subtest items will result in less cost, easier administration, and less possible distractions from multidimensionality, random bias.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	118.449	104.222
SC	120.880	111.514
-2 Log L	116.449	98.222

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	18.2272	2	0.0001
Score	16.7975	2	0.0002
Wald	14.0145	2	0.0009

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.9036	0.5982	10.1273	0.0015
sex	1	1.4685	0.5756	**6.5082**	**0.0107**
treat	1	1.7816	0.5188	**11.7949**	**0.0006**

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
sex	4.343	1.405	13.421
treat	5.939	2.149	16.417

Association of Predicted Probabilities and Observed Responses

Percent Concordant	61.7	Somers' D	0.480
Percent Discordant	13.8	Gamma	0.635
Percent Tied	24.5	Tau-a	0.243
Pairs	1764	c	0.740

Part of the default output from PROC LOGISTIC is a table that has entries including `percent concordant' and `percent discordant'. This implies the percent that would correctly be assigned, based on the results of the logistic regression.

- **Percent Concordant** - A pair of observations with different observed responses is said to be concordant if the observation with the lower ordered response value has a lower predicted mean score than the observation with the higher ordered response value.

  See **Pairs**, superscript z, for what defines a pair.
- **Percent Discordant** - If the observation with the lower ordered response value has a higher predicted mean score than the observation with the higher ordered response value, then the pair is discordant.
- **Percent Tied** - If a pair of observations with different responses is neither concordant nor discordant, it is a tie.

Solution to the problem of concordant and discordant in PROC LOGISTIC

The **CTABLE** option is used to ask for a classification table. You need the CTABLE option on the MODEL statement, which gives the proportion correctly classified, the sensitivity, the specificity, and other measures for each of a number of cutpoints of the predicted probability level. By default, it gives probability levels from 0 to 1 at intervals of .02, but if you just want a few, you can get them:

```
/*Logistical regression model check*/
data arthrits;
  input sex$ trtment$ improve$ count;
  _treat_=(trtment='Active');
  _sex_  =(sex='F');
  better =(improve='some');
/* some means improve, none means not improve*/
cards;
  F Active  none  6
  M Active  none  7
  F Active  some 21
  M Active  some  7
  F Placebo none 19
  M Placebo none 10
  F Placebo some 13
  M Placebo some  1
;

proc logistic data=arthrits descending;
  freq count;
  model better = _sex_ _treat_ / ctable pprob = (.25 .5 .75) scale=none
aggregate;
run;
```

See the last part of the output:

Output:

```
The LOGISTIC Procedure

                                Model Information

                    Data Set                      WORK.ARTHRITS
                    Response Variable             better
                    Number of Response Levels     2
                    Frequency Variable            count
                    Model                         binary logit
                    Optimization Technique        Fisher's scoring
```

Number of Observations Read	8
Number of Observations Used	8
Sum of Frequencies Read	84
Sum of Frequencies Used	84

Response Profile

Ordered Value	better	Total Frequency
1	1	42
2	0	42

Probability modeled is better=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	0.2776	1	0.2776	0.5983
Pearson	0.2637	1	0.2637	0.6076

Number of unique profiles: 4

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	118.449	104.222
SC	120.880	111.514
-2 Log L	116.449	98.222

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	18.2272	2	0.0001
Score	16.7975	2	0.0002
Wald	14.0145	2	0.0009

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.9036	0.5982	10.1273	0.0015

sex	1	1.4685	0.5756	6.5082	0.0107
treat	1	1.7816	0.5188	11.7949	0.0006

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
sex	4.343	1.405	13.421
treat	5.939	2.149	16.417

Association of Predicted Probabilities and Observed Responses

Percent Concordant	61.7	Somers' D	0.480
Percent Discordant	13.8	Gamma	0.635
Percent Tied	24.5	Tau-a	0.243
Pairs	1764	c	0.740

The LOGISTIC Procedure

Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.250	41	10	32	1	60.7	97.6	23.8	43.8	9.1
0.500	21	36	6	21	67.9	50.0	85.7	22.2	36.8
0.750	21	36	6	21	67.9	50.0	85.7	22.2	36.8

The likelihood ratio test

The LR test is performed by estimating two models and comparing the fit of one model to the fit of the other. Removing predictor variables from a model will almost always make the model fit less well (i.e., a model will have a lower log likelihood), but it is necessary to test whether the observed difference in model fit is statistically significant. The lr test does this by comparing the log likelihoods of the two models, if this difference is statistically significant, then the less restrictive model (the one with more variables) is said to fit the data significantly better than the more restrictive model. If one has the log likelihoods from the models, the lr test is fairly easy to calculate. The formula for the lr test statistic is:

$$\mathbf{LR = -2\ ln(L(m1)/L(m2)) = 2(ll(m2)-ll(m1))}$$

Where L(m*) denotes the likelihood of the respective model (either model 1 or model 2), and ll(m*) the natural log of the model's final likelihood (i.e., the log likelihood). Where m1 is the more restrictive model, and m2 is the less restrictive model.

The resulting test statistic is distributed chi-squared, with degrees of freedom equal to the number of parameters that are constrained (in the current example, the number of variables removed from the model, i.e. 2).

The Wald test

The Wald test approximates the LR test, but with the advantage that it only requires estimating one model. The Wald test works by testing the null hypothesis that a set of parameters is equal to some value. In the model being tested here, the null hypothesis is that the two coefficients of interest are simultaneously equal to zero. If the test fails to reject the null hypothesis, this suggests that removing the variables from the model will not substantially harm

the fit of that model, since a predictor with a coefficient that is very small relative to its standard error is generally not doing much to help predict the dependent variable.

Wald test tests how far the estimated parameters are from zero (or any other value under the null hypothesis) in standard errors, similar to the hypothesis tests typically printed in regression output. The difference is that the Wald test can be used to test multiple parameters simultaneously, while the tests typically printed in regression output only test one parameter at a time.

The Wald statistic，analogous to the t-test in linear regression, is used to assess the significance of coefficients. The Wald statistic is the ratio of the square of the regression coefficient to the square of the standard error of the coefficient and is asymptotically distributed as a chi-square distribution.

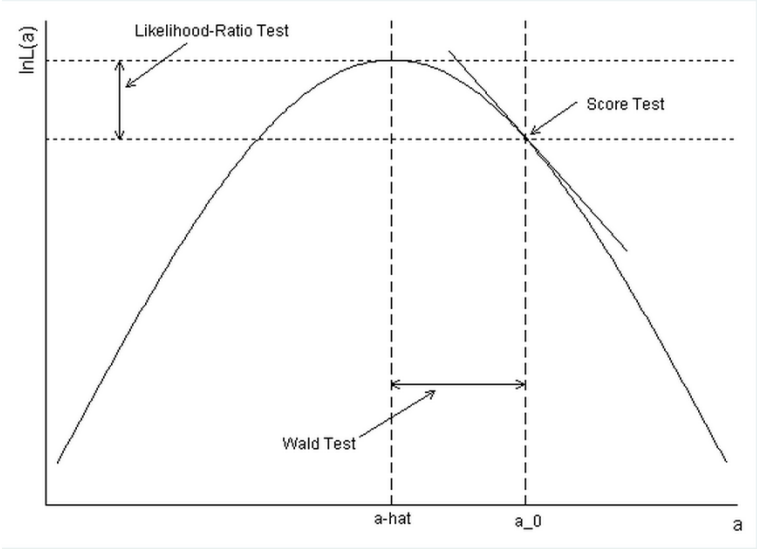


Figure based on a figure in Fox (1997, p. 570); used with authors permission.

Note: However, several authors have identified problems with the use of the Wald statistic. Menard (1995) warns that for large coefficients, standard error is inflated, lowering the Wald statistic (chi-square) value. Agresti (1996) states that the likelihood-ratio test is more reliable for small sample sizes than the Wald test.

Power of test to differentiate clinical and nonclinical cases.

- **Sensitivity**-how accurate to identify clinical case (type I error).

- **Specificity**- how accurate to identify non-clinical case (type II error)

* sensitivity and specificity must be both above **80%** (celf-5 requirement)

Example: 84 cases (42 clinical/non-clinical cases)

Reality	Clinical	Non-clinical
Clinical	X(33)	Y(9)
Non-clinical	W(9)	Z(33)

Positive predictive power= X×base rate /(X × base rate + (1-base rate) ×Y) (33/42)

Negative decision power= Z× (1-base rate) /(W×base rate + Z× (1-base rate)) (33/42)

The base rate here is 50% (42 clinical cases: 42 non-clinical cases)

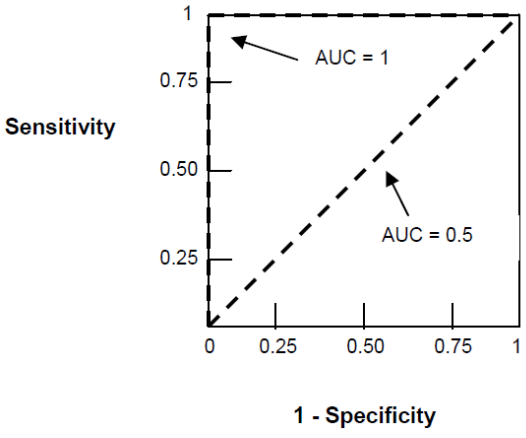
RECEIVER-OPERATOR CHARACTERISTIC (ROC) ANALYSIS

Receiver-Operator Characteristic (ROC) analysis is a method in which a continuous measure is evaluated as a diagnostic tool in predicting a dichotomous outcome. Some common examples include using chance of rain (continuous measure) to predict whether rain will actually occur (dichotomous outcome) and using a credit score (continuous measure) to predict whether a card holder will have a delinquent payment (dichotomous outcome).

For each of these various cutoff points, sensitivity and specificity can be calculated. After several cutoff points have been evaluated, we plot sensitivity versus 1 – specificity (the true positive percentage versus the false positive percentage). The resulting plot is called the **ROC curve**, and it characterizes the ability of the final model to "diagnose" groups.

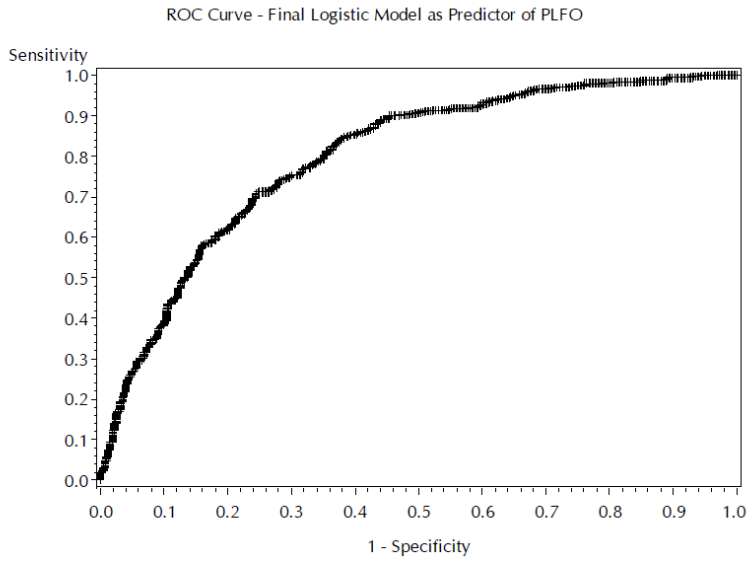
A summary measure of how well the final model predicts dichotomous group variable is **the area under the curve (AUC).** If it is possible to obtain a high true positive percentage without rapidly increasing the false positive percentage, the curve rises sharply, and the total area under the curve is close to 1. When AUC=0.5, the continuous predictor is non-informative (essentially like flipping a coin); when AUC=1, the predictor is a perfect diagnostic measure.

On the ROC curve, an AUC of 0.5 would be represented by the diagonal of the plot area starting at the origin, and an AUC of 1 would be represented by a step function of unity:



ROC data in PROC LOGISTIC is generated via the **outroc=** option in the model statement. SAS® basically runs through multiple cutoff points of the predictor and calculates the corresponding sensitivity and specificity, writing the resulting data to a specified dataset. To generate the ROC curve, we must plot the sensitivity versus 1 – specificity from this resulting dataset:

```
* Generating the ROC data;
proc logistic data=saslib.sad_PLFO;
where PLFO in (0,1);
class sexmf (ref='1') / param=ref;
model PLFO (event='1') = agebl_dec sexmf raceblack racehisp presentsmk
pastsmk
ihxad_fam ipet sxduyn sxasyn collgrad
sxasyn amdurbl_dec_male amdurbl_dec_female
/ outroc=rocdata;
run;
* Plotting the ROC data;
proc gplot data=rocdata;
title1 "ROC curve - Final model as a predictor for PLFO";
plot _sensit_*_1mspec_;
run;
quit;
```



The ROC analysis is also closely associated with a set of summary statistics that are part of the default SAS® output:

Association of Predicted Probabilities and Observed Responses

Percent Concordant	79.6	Somers' D	0.594
Percent Discordant	20.2	Gamma	0.595
Percent Tied	0.2	Tau-a	0.286
Pairs	243815	c	0.797

ROC curve (based on cut-off line)

Note: The more ROC curve above diagonal line, the better and more useful this ROC curve is!

Criteria:

1. Whether the LR model is significant?
2. Wald CHI-SQUARE and P-value check (e.g., gender, treatment)
3. Check "Percent Concordant" value (e.g., the higher the better, at least 80% baseline)

Cut score must be selected by each age group.

Stepwise Regression

Stepwise regression is used in the exploratory phase of research but it is not recommended for theory testing (Menard 1995). Theory testing is the testing of a-priori theories or hypotheses of the relationships between variables. Exploratory testing makes no a-priori assumptions regarding the relationships between the variables, thus the goal is to discover relationships. Stepwise regression involves starting with no variables in the model, testing the addition of each variable using a chosen model comparison criterion, adding the variable (if any) that improves the model the most, and repeating this process until none improves the model.

http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_logistic_sect052.htm

This code is used for stepwise regression:

```
data Remission;
      input remiss cell smear infil li blast temp;
      label remiss='Complete Remission';
      datalines;
   1    .8    .83   .66   1.9   1.1      .996
   1    .9    .36   .32   1.4    .74     .992
   0    .8    .88   .7     .8    .176    .982
   0   1      .87   .87    .7   1.053    .986
   1    .9    .75   .68   1.3    .519    .98
   0   1      .65   .65    .6    .519    .982
   1    .95   .97   .92   1     1.23     .992
   0    .95   .87   .83   1.9   1.354   1.02
   0   1      .45   .45    .8    .322    .999
   0    .95   .36   .34    .5   0       1.038
   0    .85   .39   .33    .7    .279    .988
   0    .7    .76   .53   1.2    .146    .982
   0    .8    .46   .37    .4    .38    1.006
   0    .2    .39   .08    .8    .114    .99
   0   1      .9    .9    1.1   1.037    .99
   1   1      .84   .84   1.9   2.064   1.02
   0    .65   .42   .27    .5    .114   1.014
   0   1      .75   .75   1     1.322   1.004
   0    .5    .44   .22    .6    .114    .99
   1   1      .63   .63   1.1   1.072    .986
   0   1      .33   .33    .4    .176   1.01
   0    .9    .93   .84    .6   1.591   1.02
```

```
   1  1      .58   .58  1      .531  1.002
   0   .95   .32   .3   1.6    .886   .988
   1  1      .6    .6   1.7    .964   .99
   1  1      .69   .69   .9    .398   .986
   0  1      .73   .73   .7    .398   .986
   ;
title 'Stepwise Regression on Cancer Remission Data';
proc logistic data=Remission outest=betas covout;
      model remiss(event='1')=cell smear infil li blast temp
                  / selection=stepwise
/* A sig level of .3 is required to allow a variable into the model
(SLENTRY=.3)*/
                    slentry=0.3
/* A sig level of .35 is required to allow a variable stay in the model
(SLstay=.35)*/

                    slstay=0.35
                    details
                    lackfit; /*The Hosmer and Lemeshow goodness-of-fit test
for the final selected model
                              is requested by specifying the LACKFIT option*/
      output out=pred p=phat lower=lcl upper=ucl
             predprob=(individual crossvalidate);
run;
proc print data=betas;
      title2 'Parameter Estimates and Covariance Matrix';
run;
proc print data=pred;
      title2 'Predicted Probabilities and 95% Confidence Limits';
run;
```

Prior to the first step, the intercept-only model is fit and individual score statistics for the potential variables are evaluated. In the stepwise selection, an attempt is made to remove any insignificant variables from the model before adding a significant variable to the model. Each addition or deletion of a variable to or from a model is listed as a separate step in the displayed output, and at each step a new model is fitted.

Using "FIRTH" command in the PROC LOGISTIC

In logistic regression, when the outcome has low (or high) prevalence, or when there are several interacted categorical predictors, it can happen that for some combination of the predictors, all the observations have the same event status. A similar event occurs when continuous covariates predict the outcome too perfectly.

This phenomenon, known as "**separation**" (including complete and quasi-complete separation) will cause problems fitting the model. Sometimes the only symptom of separation will be extremely large standard errors, while at other times the software may report an error or a warning.

One approach to handling this sort of problem is **exact logistic regression**.

```
proc logistic data=dose descending;
model Deaths/Total = Dose;
exact Dose / estimate=both;
run;
```

But exact logistic regression is complex and may require prohibitive computational resources. Another option is to use a Bayesian approach. Here we show how to use a penalized likelihood method originally proposed by Firth (1993 Biometrika 80:27-38) and described fully in this setting by Georg Heinze (2002 Statistics in Medicine 21:2409-2419 and 2006 25:4216-4226).

In SAS, the corrected estimates can be found using the **firth** option to the model statement in **proc logistic**. We'll set up the problem in the simple setting of a 2x2 table with an empty cell. Here, we simply output three observations with three combinations of predictor and outcome, along with a weight variable which contains the case counts in each cell of the table

Firth logistic regression: (Example code)

```
%LET project=WISC5_Stdz_50;
%let lgvar=gt/* sldm*/;
%let cvars=wisc5_bdnb_s01-wisc5_bdnb_s16;
data logi; set WISC5(keep=stdz50 &lgvar &cvars);
   where (stdz50='Y' or &lgvar =1);
run;
proc logistic data=logi; /*plots=roc(id=obs) plot a ROC curve*/
   class &lgvar;
   model &lgvar.(event='1')  =&cvars /firth maxiter=150 rsquare lackfit
ctable;
   title1 "FIRTH Logistic Regression for Subtest &pre  &project";
   title2 "Predicting &lgvar = 1";
   title3 "***       Analyzed by &analyst    on &day       ***";
   ods output parameterestimates=firth;
run; title;
```