Cool Tools for PROC LOGISTIC

Paul D. Allison

Statistical Horizons LLC and the
University of Pennsylvania

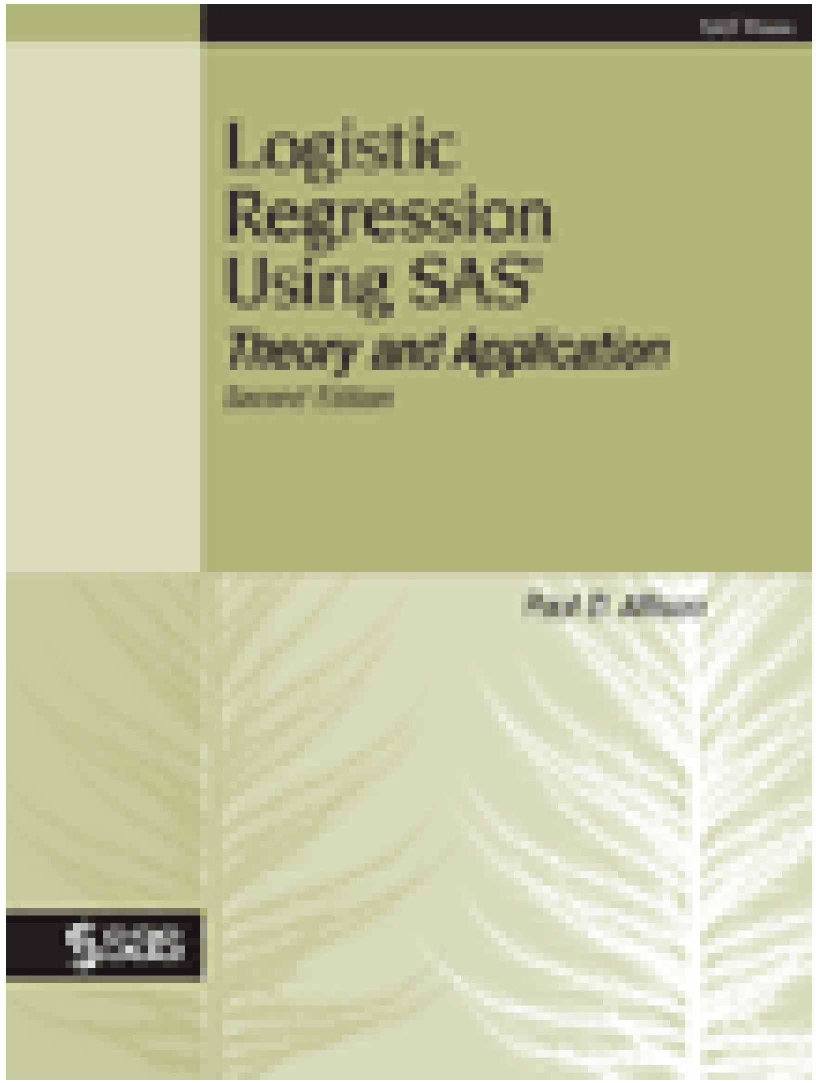
March 2013

www.StatisticalHorizons.com

New Features in LOGISTIC

- ODDSRATIO statement
- EFFECTPLOT statement
- ROC comparisons
- FIRTH option

Based on Recent Book



Also covered in 2-day course
Logistic Regression Using SAS
June 6-7, Philadelphia
Temple University Center City

Other SAS courses offered by Statistical Horizons:

- **Introduction to Structural Equation Modeling**
  Paul Allison, Instructor
  April 12-13, Washington, DC
- **Longitudinal Data Analysis Using SAS**
  Paul Allison, Instructor
  April 19-20, Philadelphia
- **Missing Data**
  Paul Allison, Instructor
  May 17-18, Los Angeles
- **Mediation and Moderation**
  Andrew Hayes & Kristopher Preacher
  July 15-19, Philadelphia
- **Event History & Survival Analysis**
  Paul Allison, Instructor
  July 15-19, Philadelphia

Get more info at StatisticalHorizons.com

Example Data Set

Data: 5960 loan customers

BAD	1=customer defaults, otherwise 0 (dependent variable)
LOAN	Amount of the loan
DEBTCON	1=debt consolidation, 0=home improvement
DELINQ	Number of delinquent trade lines
NINQ	Number of recent credit inquiries.
DEBTINC	Debt to income ratio

ODDSRATIO Statement

- Good for interactions
- Fit the following model:

```
PROC LOGISTIC DATA=my.credit DESC;
MODEL bad = loan debtcon delinq ninq
   debtinc debtcon*ninq ;
RUN;
```

Coefficients

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.8119	0.3434	286.4367	<.0001
loan	1	-0.00001	5.937E-6	3.6289	0.0568
debtcon	1	0.0831	0.1599	0.2705	0.6030
delinq	1	0.6480	0.0509	161.8830	<.0001
ninq	1	0.3475	0.0742	21.9398	<.0001
debtinc	1	0.0873	0.00848	105.8757	<.0001
debtcon*ninq	1	-0.2221	0.0822	7.3083	0.0069

Odds Ratios

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
LOAN	1.000	1.000	1.000
DELINQ	1.912	1.730	2.112
DEBTINC	1.091	1.073	1.109

Odds ratios not reported for the variables in the interaction.

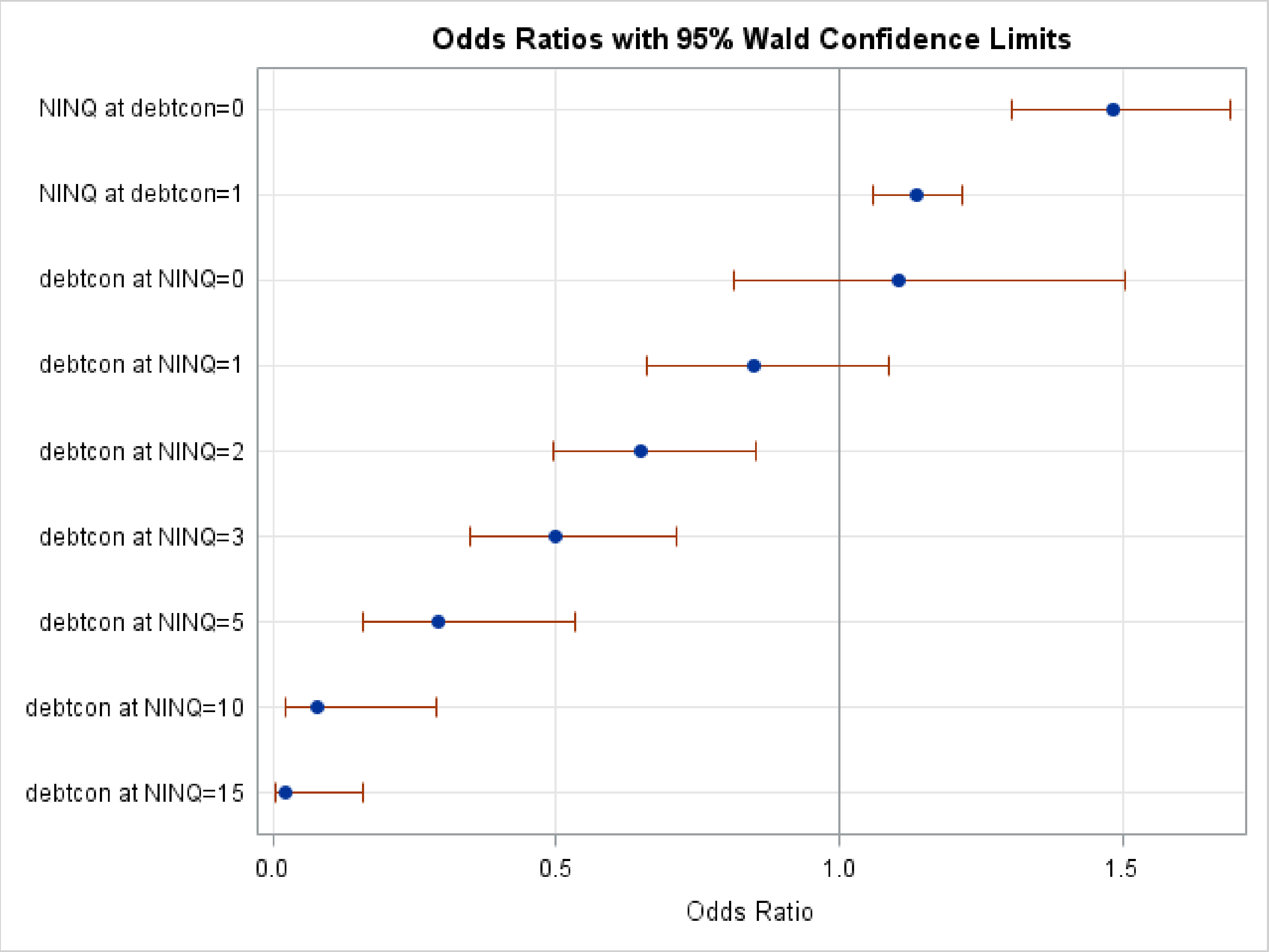
But now we can get them:

```
PROC LOGISTIC DATA=my.credit DESC;
MODEL bad = loan debtcon delinq ninq
   debtinc debtcon*ninq ;
ODDSRATIO ninq / AT(debtcon=0 1);
ODDSRATIO debtcon / AT(ninq=0 1 2 3 5 10 15);
RUN;
```

Odds Ratios for Interactions

Odds Ratio Estimates and Wald Confidence Intervals			
Label	Estimate	95% Confidence Limits	
NINQ at debtcon=0	1.484	1.303	1.690
NINQ at debtcon=1	1.136	1.061	1.218
debtcon at NINQ=0	1.106	0.813	1.506
debtcon at NINQ=1	0.847	0.660	1.088
debtcon at NINQ=2	0.649	0.495	0.851
debtcon at NINQ=3	0.497	0.348	0.711
debtcon at NINQ=5	0.292	0.159	0.535
debtcon at NINQ=10	0.077	0.021	0.287
debtcon at NINQ=15	0.020	0.003	0.157

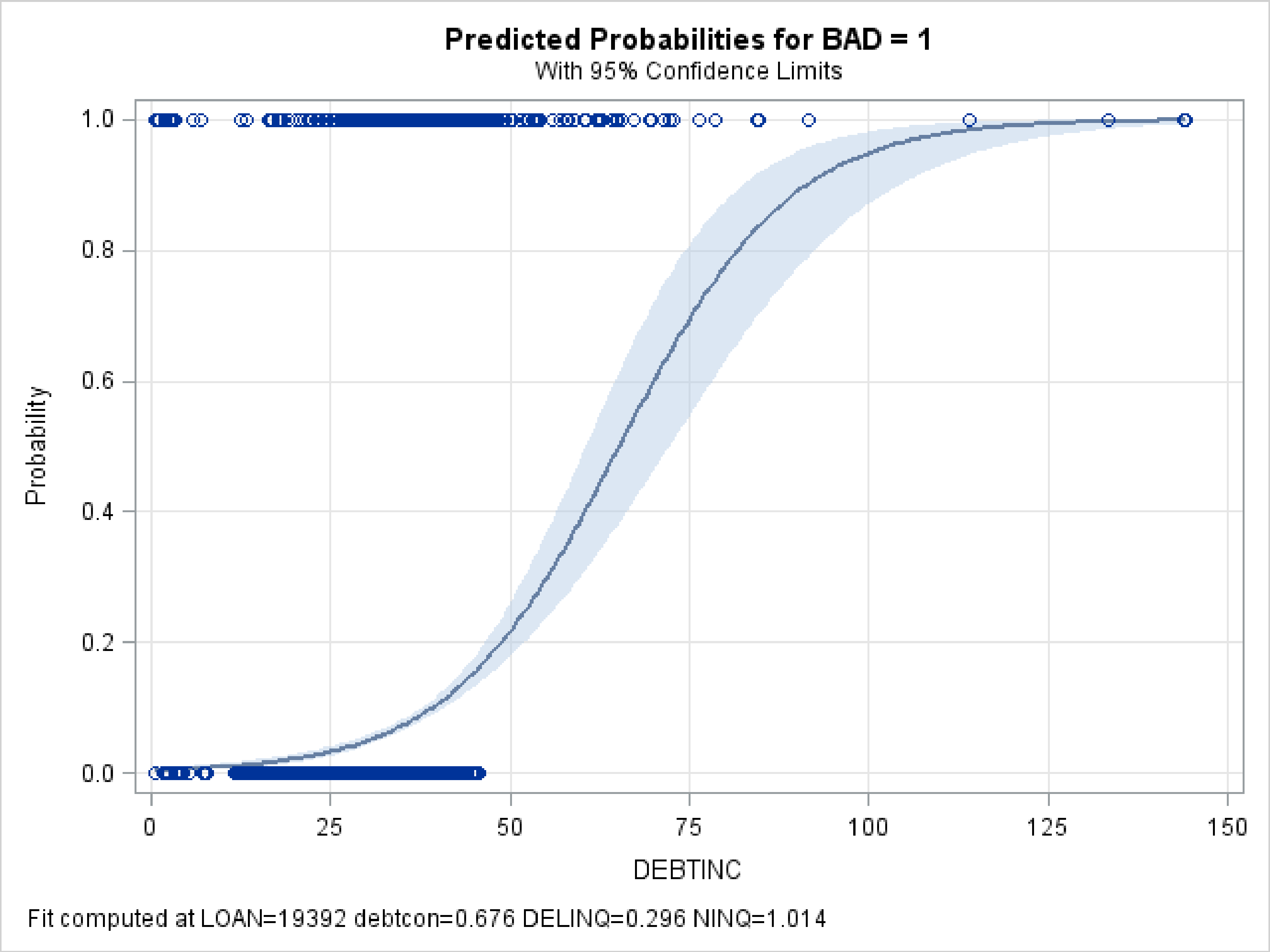
ODS Graphics



EFFECTPLOT Statement

The EFFECTPLOT statement offers many possibilities for plotting predicted values as a function of one or more predictors. Here's how to get a graph for one predictor, holding the others at their means (or reference category for CLASS variables).

```
ODS GRAPHICS ON;
PROC LOGISTIC DATA=my.credit DESC;
MODEL bad = loan debtcon delinq ninq
   debtinc;
EFFECTPLOT FIT(X=debtinc);
RUN;
ODS GRAPHICS OFF;
```



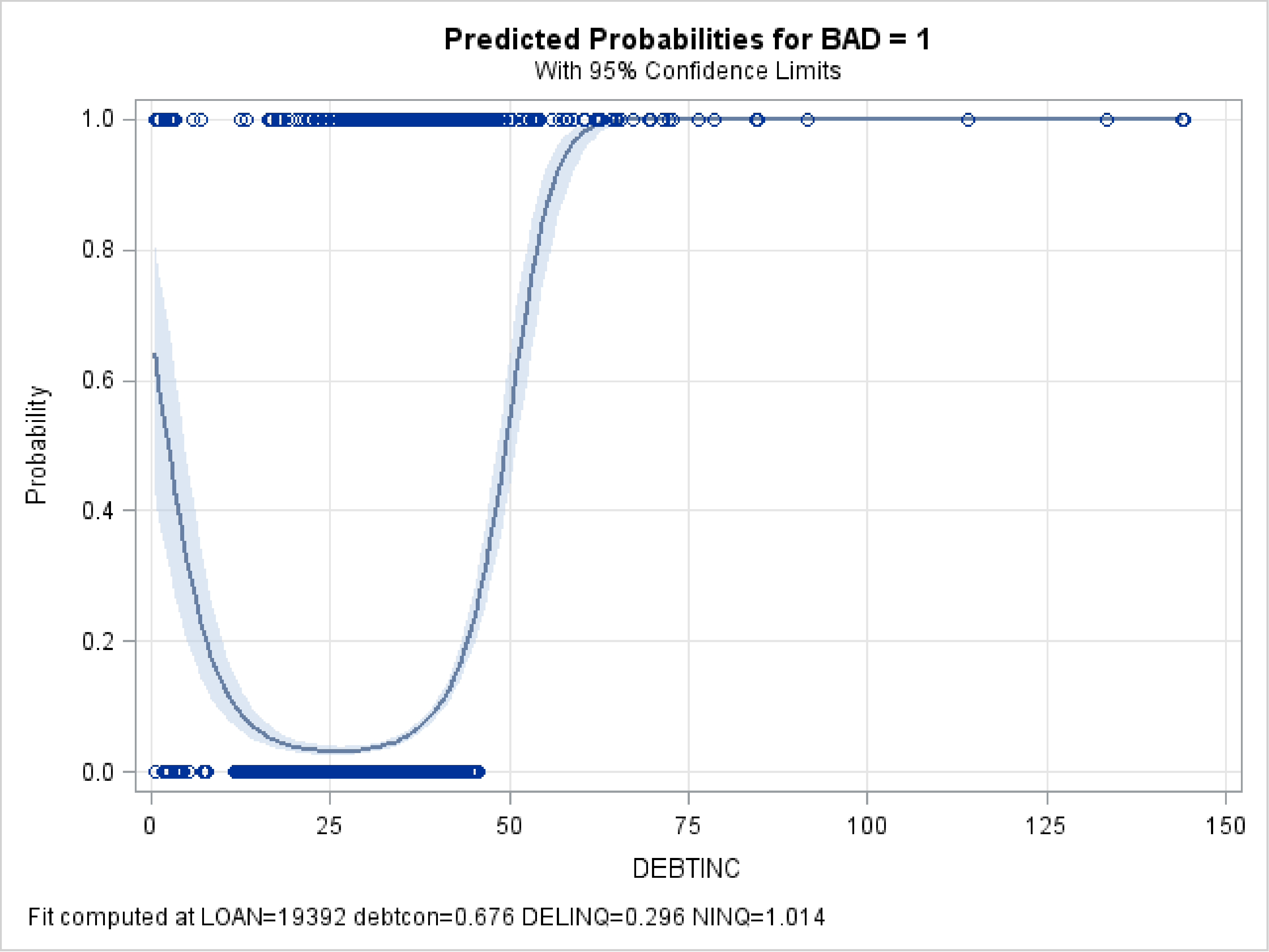
Predicted Probabilities for BAD = 1
With 95% Confidence Limits
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
25
50
75
100
125
150
DEBTINC
Fit computed at LOAN=19392 debtcon=0.676 DELINQ=0.296 NINQ=1.014

Polynomial Functions

Things get more interesting with polynomial functions:

```
ODS GRAPHICS ON;
PROC LOGISTIC DATA=my.credit DESC;
MODEL bad = loan debtcon delinq ninq
   debtinc debtinc*debtinc;
EFFECTPLOT FIT(X=debtinc);
RUN;
ODS GRAPHICS OFF;
```

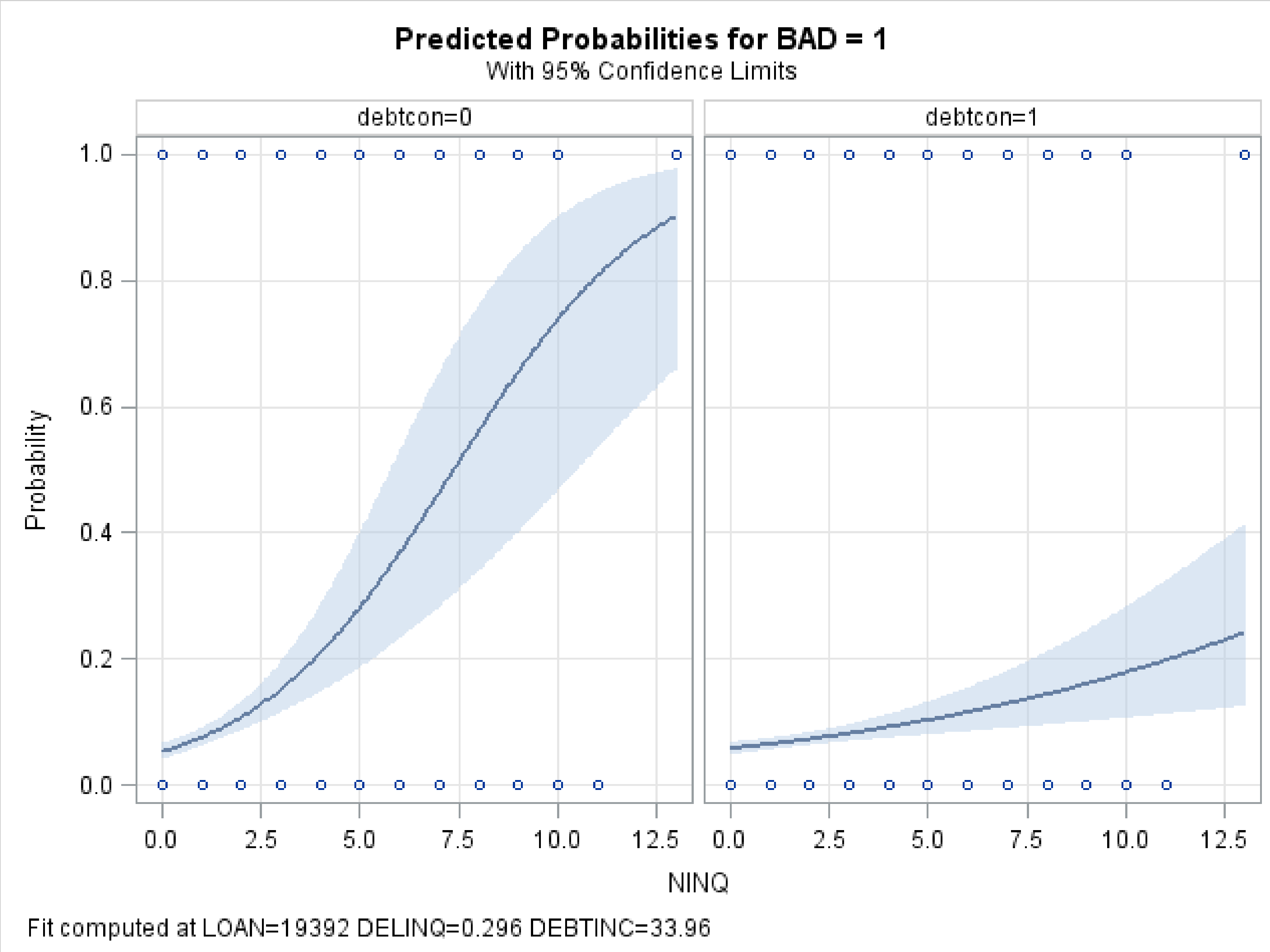
The squared term is highly significant.


Predicted Probabilities for BAD = 1
With 95% Confidence Limits
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
25
50
75
100
125
150
DEBTINC
Fit computed at LOAN=19392 debtcon=0.676 DELINQ=0.296 NINQ=1.014

EFFECTPLOT with Interactions

EFFECTPLOT is also very useful for visualizing interactions:

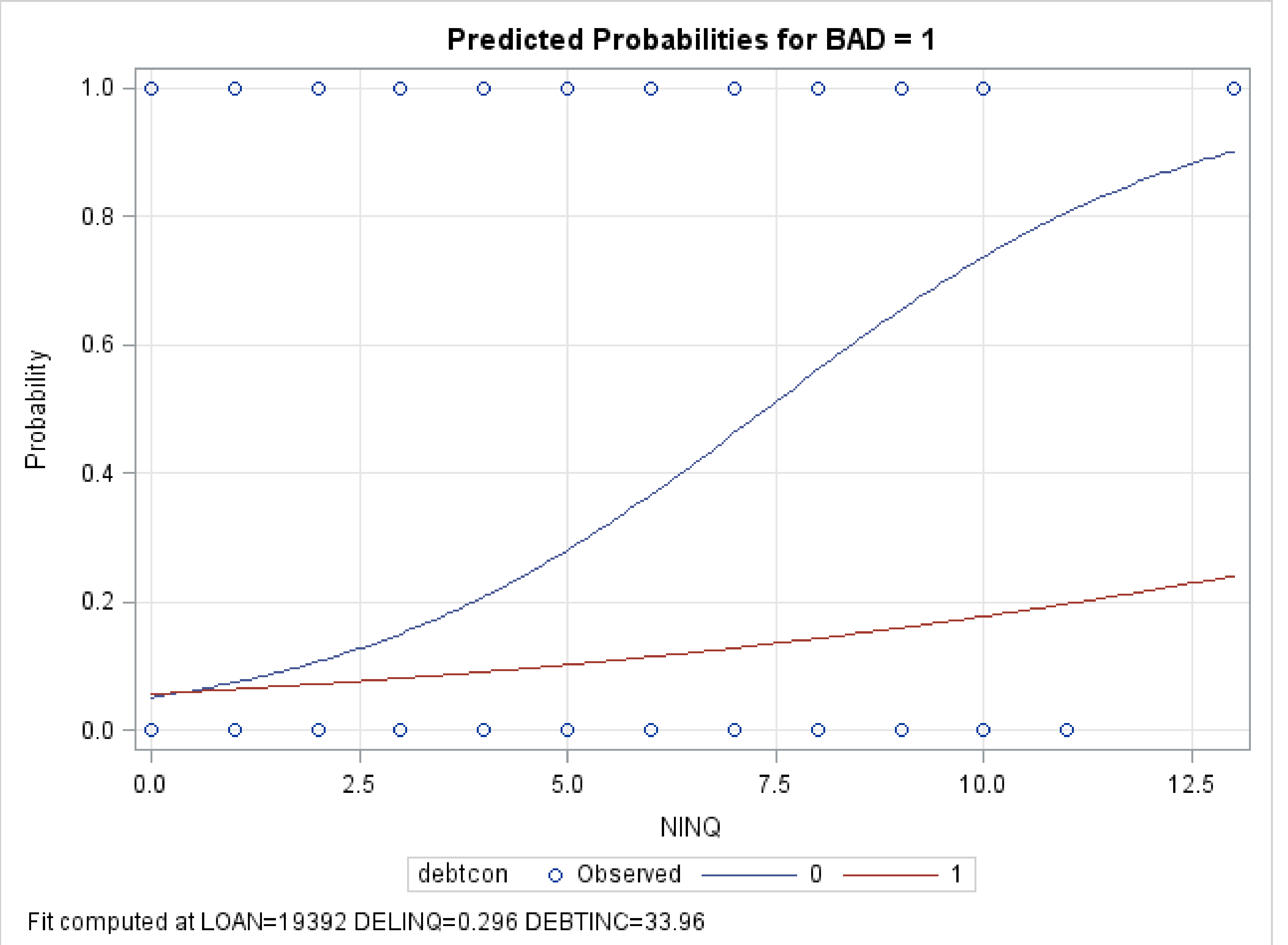
```
ODS GRAPHICS ON;
PROC LOGISTIC DATA=my.credit DESC;
MODEL bad = loan debtcon delinq ninq
   debtinc debtcon*ninq ;
EFFECTPLOT FIT(X=ninq) / AT(debtcon=0 1);
RUN;
ODS GRAPHICS OFF;
```


Predicted Probabilities for BAD = 1
With 95% Confidence Limits
debtcon=0
debtcon=1
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0.0
2.5
5.0
7.5
10.0
12.5
NINQ
Fit computed at LOAN=19392 DELINQ=0.296 DEBTINC=33.96

EFFECTPLOT with Interactions

If you want the two graphs on the same axes, use the SLICEFIT option:

```
ODS GRAPHICS ON;
PROC LOGISTIC DATA=my.credit DESC;
MODEL bad = loan debtcon delinq ninq
   debtinc debtcon*ninq ;
EFFECTPLOT SLICEFIT(X=ninq
    SLICEBY=debtcon=0 1); RUN;
ODS GRAPHICS OFF;
```


Predicted Probabilities for BAD = 1
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0.0
2.5
5.0
7.5
10.0
12.5
NINQ
debtcon
Observed
0
1
Fit computed at LOAN=19392 DELINQ=0.296 DEBTINC=33.96

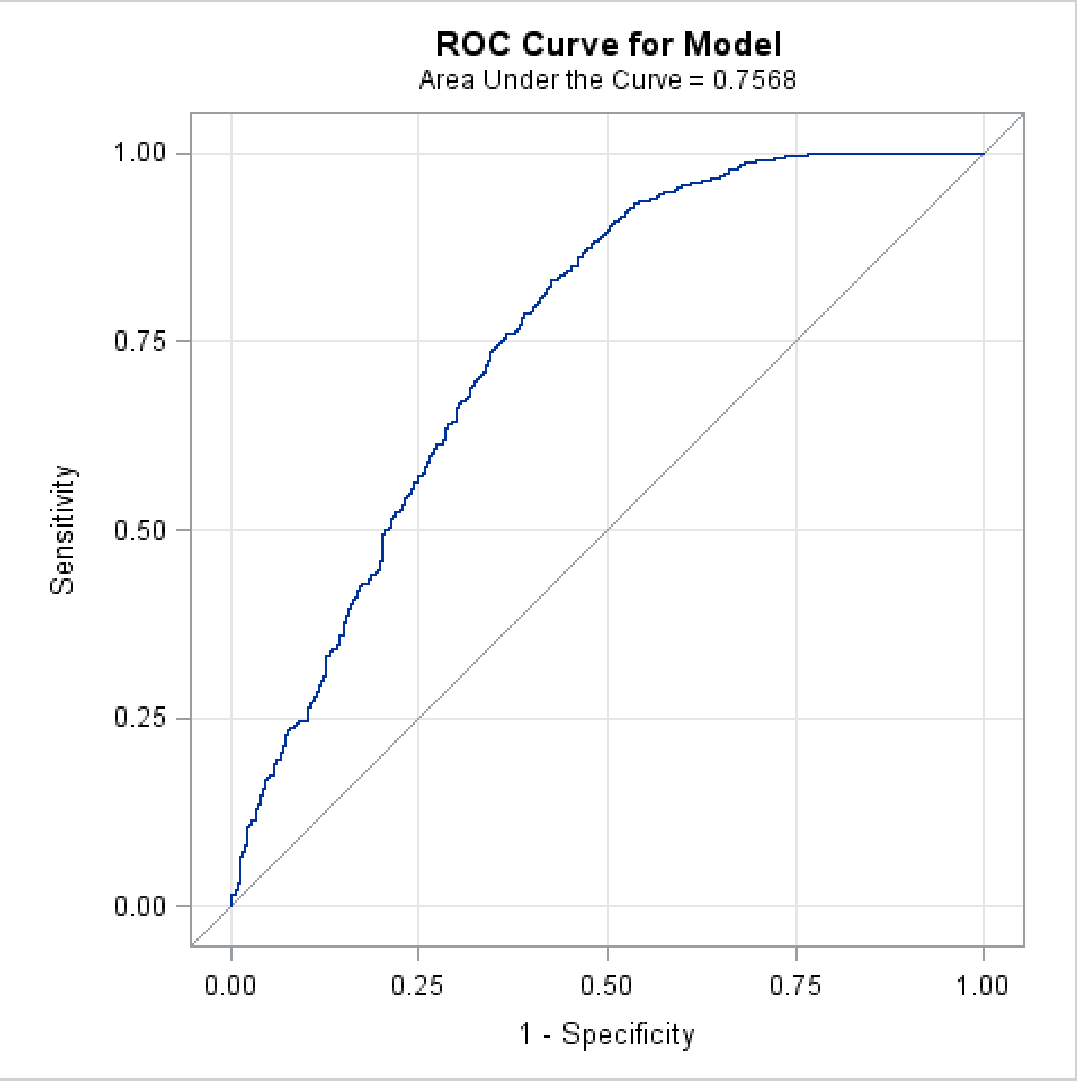
ROC Curves

The Receiver Operating Characteristic curve is a way of evaluating the predictive power of a model for a binary outcome. It's a graph of sensitivity vs. 1-specificity.

> Sensitivity = probability of predicting an event, given that the individual has an event.
>
> Sensitivity = probability of predicting a non-event, given that the individual does not have an event.

Both of these depend on the cut-point for determining whether a predicted probability is evaluated as an event prediction or a non-event prediction. Here's how to get the curve:

```
PROC LOGISTIC DATA=my.credit PLOTS(ONLY)=ROC;
MODEL bad = loan debtcon delinq ninq
   debtinc debtcon*ninq debtinc*debtinc;
RUN;
```



The area under the curve (the C statistic) is a summary measure of predictive power

ROCCONTRAST

With the ROC and ROCCONTRAST statements, we can get confidence intervals around the C statistic and compare different curves:

```
PROC LOGISTIC DATA=my.credit;
MODEL bad = loan debtcon delinq ninq
   debtinc debtcon*ninq debtinc*debtinc;
  ROC 'omit debtcon*ninq' loan debtcon delinq ninq
    debtinc*debtinc;
  ROC 'omit debtinc*debtinc' loan delinq ninq
    debtinc debtcon*ninq;
  ROC 'omit both' loan debtcon delinq ninq debtinc;
  ROCCONTRAST / ESTIMATE=ALLPAIRS;
RUN;
```

ROC Results 1

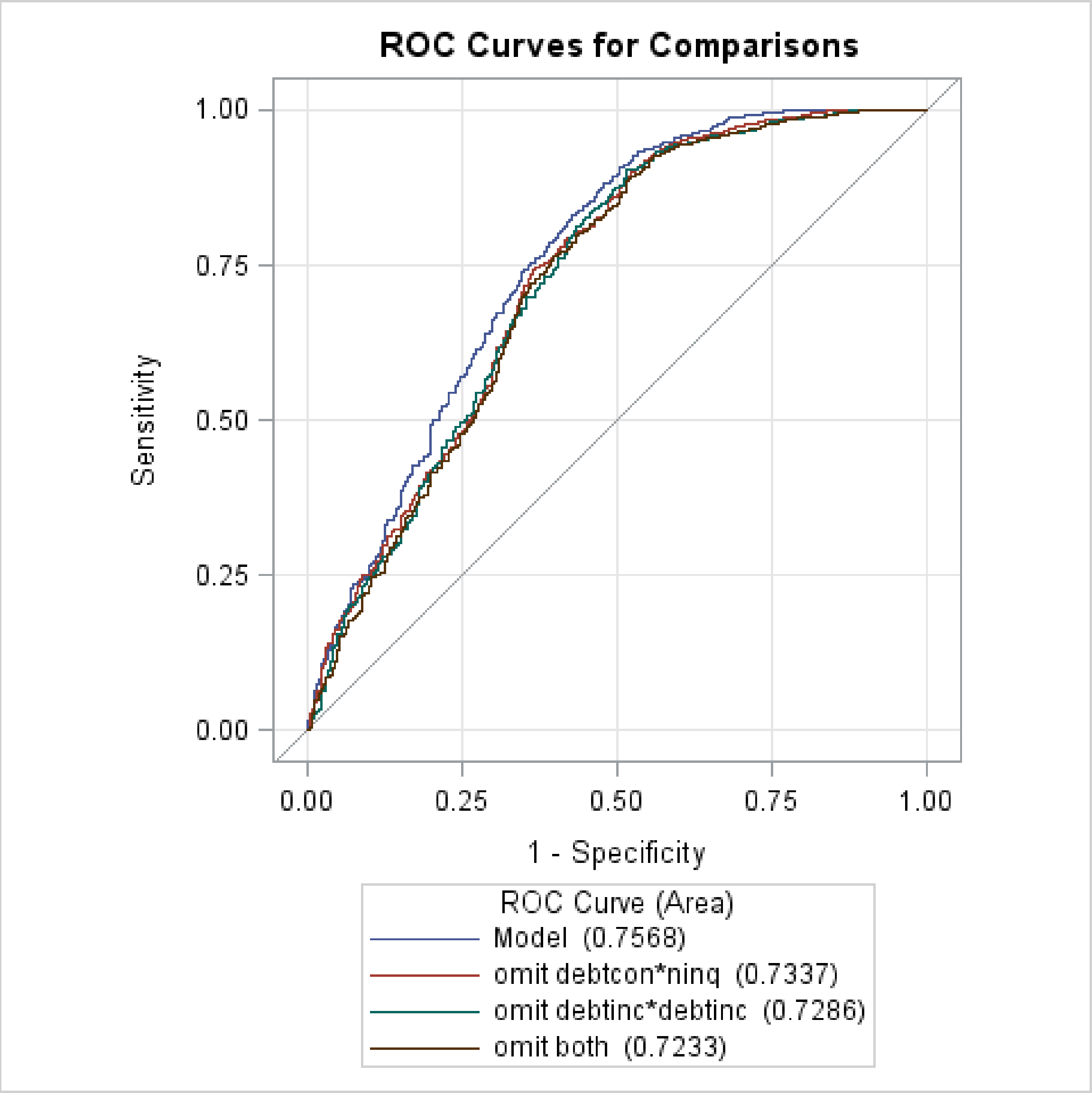
ROC Association Statistics							
ROC Model	Mann-Whitney				Somers' D (Gini)	Gamma	Tau-a
	Area	Standard Error	95% Wald Confidence Limits				
Model	0.7568	0.0152	0.7270	0.7865	0.5136	0.5136	0.0842
omit debtcon *ninq	0.7337	0.0156	0.7032	0.7643	0.4675	0.4675	0.0767
omit debtinc* debtinc	0.7286	0.0158	0.6976	0.7596	0.4571	0.4571	0.0750
omit both	0.7233	0.0160	0.6920	0.7546	0.4466	0.4466	0.0732

ROC Results 2

ROC Contrast Test Results			
Contrast	DF	Chi-Square	Pr > ChiSq
Reference = Model	3	24.4103	<.0001

ROC Contrast Estimation and Testing Results by Row						
Contrast	Estimate	Standard Error	95% Wald Confidence Limits		Chi-Square	Pr > ChiSq
Model - omit debtcon*ninq	0.0230	0.00774	0.00785	0.0382	8.8401	0.0029
Model - omit debtinc*debtinc	0.0282	0.00866	0.0112	0.0452	10.6136	0.0011
Model - omit both	0.0335	0.00914	0.0155	0.0514	13.3997	0.0003
omit debtcon*ninq - omit debtinc*debtinc	0.00519	0.00325	-0.00117	0.0116	2.5581	0.1097
omit debtcon*ninq - omit both	0.0104	0.00233	0.00585	0.0150	19.9831	<.0001
omit debtinc*debtinc - omit both	0.00524	0.00233	0.000666	0.00981	5.0423	0.0247

ROC Results


ROC Curves for Comparisons
1.00
0.75
0.50
0.25
0.00
Sensitivity
0.00
0.25
0.50
0.75
1.00
1 - Specificity
ROC Curve (Area)
Model (0.7568)
omit debtcon*ninq (0.7337)
omit debtinc*debtinc (0.7286)
omit both (0.7233)

FIRTH Option

A solution to the problem of quasi-complete separation—failure of the ML algorithm to converge because some coefficients are infinite.

Example:

```
PROC LOGISTIC DATA=my.credit ;
 CLASS derog /PARAM=GLM DESC;
 MODEL bad = derog;
RUN;
```

DEROG is the number of derogatory reports. This code produces the following in both the log and output windows.

```
WARNING: There is possibly a quasi-complete separation
of data points. The maximum likelihood estimate may not
exist.
WARNING: The LOGISTIC procedure continues in spite of
the above warning. Results shown are based on the last
maximum likelihood iteration. Validity of the model fit
is questionable.
```

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
DEROG 10 vs 0	<0.001	<0.001	>999.999
DEROG 9 vs 0	<0.001	<0.001	>999.999
DEROG 8 vs 0	<0.001	<0.001	>999.999
DEROG 7 vs 0	<0.001	<0.001	>999.999
DEROG 6 vs 0	0.100	0.034	0.293
DEROG 5 vs 0	0.228	0.083	0.632
DEROG 4 vs 0	0.056	0.021	0.150
DEROG 3 vs 0	0.070	0.039	0.126
DEROG 2 vs 0	0.190	0.138	0.262
DEROG 1 vs 0	0.315	0.255	0.387

Frequency

Table of BAD by DEROG

BAD	DEROG											
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
0	3773	266	78	15	5	8	5	0	0	0	0	4150
1	754	169	82	43	18	7	10	8	6	3	2	1102
Total	4527	435	160	58	23	15	15	8	6	3	2	5252

Why does this happen? Because for values of DEROG > 6, every individual had BAD=1.

Quasi-complete separation occurs when, for one or more categories of a CLASS variable, either everyone has the event or no one has the event.

Penalized Likelihood

An effective solution is to invoke penalized likelihood by the FIRTH option:

```
PROC LOGISTIC DATA=my.credit ;
  CLASS derog / PARAM=GLM DESC;
  MODEL bad = derog /FIRTH CLODDS=PL;
RUN;
```

The CLODDS options requests confidence intervals for the odds ratios based on the profile likelihood method.

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
DEROG 10 vs 0	1.0000	0.040	<0.001	0.492
DEROG 9 vs 0	1.0000	0.029	<0.001	0.295
DEROG 8 vs 0	1.0000	0.015	<0.001	0.130
DEROG 7 vs 0	1.0000	0.012	<0.001	0.094
DEROG 6 vs 0	1.0000	0.105	0.034	0.285
DEROG 5 vs 0	1.0000	0.227	0.084	0.625
DEROG 4 vs 0	1.0000	0.059	0.021	0.145
DEROG 3 vs 0	1.0000	0.071	0.038	0.125
DEROG 2 vs 0	1.0000	0.190	0.138	0.262
DEROG 1 vs 0	1.0000	0.314	0.256	0.387